UNIVERSITÀ POLITECNICA DELLE MARCHE
FACOLTÀ DI INGEGNERIA
Dipartimento di Ingegneria dell'Informazione
Corso di Laurea in Ingegneria Informatica e dell'Automazione



TESI DI LAUREA

Progettazione e realizzazione di una campagna di data analytics su dati relativi ai femminicidi in Italia

Design and implementation of a data analytics campaign based on data related to femicides in Italy

Relatore

Prof. Domenico Ursino

Correlatore

Dott. Michele Marchetti

Candidata

Sara Vaccaro

ANNO ACCADEMICO 2025-2026

A tutte le donne.

A quelle che hanno i miei stessi sogni,
le mie stesse speranze,
lo stesso desiderio di vivere e di essere libere,
ma non le stesse possibilità.

A quelle che hanno trovato la forza di alzarsi,
a quelle che stanno ancora cercando la propria voce
e a quelle che quella voce non hanno più potuto usarla.

Soprattutto, a quelle a cui sono state strappate le ali
prima ancora che potessero scoprire quanto lontano avrebbero potuto volare;
a quelle a cui è stato negato il tempo di diventare ciò che desideravano,
di vedere i propri sogni trasformarsi in realtà,
di continuare semplicemente a vivere.

Perché dietro ogni nome, ogni numero e ogni dato
c'era una donna, con una storia, dei sogni
e un futuro che meritava di esistere.

One day I am gonna grow wings

Radiohead, "Let Down"

Sommario

Il femminicidio costituisce un fenomeno di rilevanza sociale, la cui analisi richiede l'integrazione di informazioni relative agli eventi, alle persone coinvolte e al contesto territoriale. In questo lavoro è stato progettato e realizzato un modello di *Business Intelligence* per l'analisi dei dati relativi a tale fenomeno in Italia. Il *dataset* utilizzato è stato analizzato e sottoposto a operazioni di pulizia, trasformazione e riorganizzazione mediante Power Query, gestendo valori mancanti, nomenclature non uniformi e duplicazioni. I dati sono stati, inoltre, integrati con informazioni demografiche ISTAT per l'analisi della distribuzione territoriale in rapporto alla popolazione residente. È stato realizzato un modello relazionale composto da una Fact Table e da una Dimension Table, corredato da misure analitiche sviluppate in DAX. Sulla base del modello sono state progettate dashboard interattive in Power BI, dedicate all'analisi delle caratteristiche degli eventi, della distribuzione temporale e territoriale, dei profili delle vittime e degli autori e delle principali dinamiche relazionali e comportamentali. Le analisi consentono di evidenziare distribuzioni, ricorrenze e associazioni presenti nei dati, mantenendo distinta l'interpretazione descrittiva da eventuali valutazioni di tipo causale.

Keyword: Big Data, Business Intelligence, Data Analysis, Data Analytics, Femminicidio, Machine Learning, Power BI

Indice

Introduzione **1**

1 Introduzione alla Data Analytics **3**
1.1 Il dato: definizione, evoluzione e classificazione . . . 3
1.1.1 Le origini semantiche . . . 4
1.1.2 L'evoluzione quantitativa . . . 4
1.1.3 La digitalizzazione . . . 4
1.1.4 Tipologie e classificazione . . . 5
1.1.5 La piramide DIKW . . . 6
1.2 I Big Data: i nuovi paradigmi architetturali . . . 7
1.2.1 L'evoluzione della produzione massiva dei dati . . . 7
1.2.2 Il modello delle 5V . . . 8
1.2.3 Cambiamento di paradigma tecnologico, scientifico ed economico . . . 9
1.3 Dalla Data Analysis alla Data Analytics . . . 9
1.3.1 Categorie della Data Analytics . . . 9
1.3.2 Data Analytics e organizzazione aziendale . . . 11
1.4 Ciclo di vita della Big Data Analytics . . . 12
1.4.1 Fase 1: Business Case Evaluation . . . 12
1.4.2 Fase 2: Data Identification . . . 13
1.4.3 Fase 3: Data Acquisition and Filtering . . . 13
1.4.4 Fase 4: Data Extraction . . . 13
1.4.5 Fase 5: Data Validation and Cleansing . . . 13
1.4.6 Fase 6: Data Aggregation and Representation . . . 13
1.4.7 Fase 7: Data Analysis . . . 14
1.4.8 Fase 8: Data Visualization . . . 14
1.4.9 Fase 9: Utilization of Analysis Results . . . 14
1.5 Dai Big Data alla Data Science e al Machine Learning . . . 14
1.5.1 Dai Big Data alla Data Science, fino al Machine Learning . . . 15
1.5.2 Il ruolo del Machine Learning . . . 15

2 Introduzione a Power BI **16**
2.1 Cos'è Power BI . . . 16
2.1.1 La storia e l'evoluzione di Power BI . . . 16
2.2 Il posizionamento nel mercato . . . 17
2.2.1 Magic Quadrant . . . 17

2.2.2 Il confronto con i principali competitor . . . 19
2.3 L'architettura: integrazione tra Desktop, Service e Mobile . . . 21
2.3.1 Power BI Desktop . . . 21
2.3.2 Power BI Service . . . 21
2.3.3 Power BI Mobile . . . 22
2.3.4 Gateway dati locale . . . 23
2.3.5 Report Server . . . 23
2.3.6 Report Builder e report impaginati . . . 23
2.4 Il flusso di lavoro su Power BI . . . 23
2.4.1 Il processo ETL: Extract, Transform, Load . . . 23
2.4.2 Data Modeling . . . 25
2.4.3 Data Visualization e reportistica . . . 26
2.4.4 Automazione e aggiornamento dei dati . . . 29
2.4.5 Condivisione . . . 30
2.5 I linguaggi di programmazione in Power BI . . . 30
2.5.1 Il linguaggio SQL . . . 31
2.5.2 Il linguaggio M . . . 31
2.5.3 Il linguaggio DAX . . . 32
2.5.4 Il linguaggio Python . . . 32
2.5.5 Il linguaggio R . . . 32

3 Analisi dei dataset di partenza **33**
3.1 L'analisi data-driven del femminicidio . . . 33
3.2 Struttura dei dataset . . . 35
3.2.1 Feminicides in Italy: counterdata RAI from 2012 to 2018 . . . 35
3.2.2 Popolazione media per regione e ripartizione geografica, anni 1952-2014 37
3.3 Valutazione della qualità dei dati . . . 38
3.3.1 Criticità nella nomenclatura dei dati . . . 38
3.3.2 Duplicazione delle informazioni sulla nazionalità . . . 39
3.3.3 Incompletezza e valori nulli . . . 39
3.3.4 Struttura dei fattori comportamentali e psichiatrici . . . 39
3.3.5 Assenza di un identificativo univoco per la vittima . . . 40
3.3.6 Discontinuità nella copertura temporale . . . 40
3.4 Il processo di ETL . . . 40
3.4.1 Extract: acquisizione e importazione delle fonti . . . 40
3.4.2 Transform: pulizia e trasformazione dei dati . . . 41
3.4.3 Load: caricamento e modellazione dei dati . . . 43

4 Progettazione e realizzazione delle dashboard di base **45**
4.1 Misure DAX . . . 45
4.1.1 Analisi quantitativa e profilazione delle vittime . . . 46
4.1.2 Analisi temporale e spaziale . . . 48
4.1.3 Analisi relazionali e fattori criminologici . . . 49
4.1.4 Analisi comportamentale dell'autore del reato . . . 53
4.2 Analisi delle dashboard di base . . . 54
4.2.1 Definizione e scopo delle dashboard di base . . . 54
4.2.2 Struttura dell'analisi visiva . . . 54
4.2.3 Dashboard 1: "Modus operandi e dinamiche del delitto" . . . 55
4.2.4 Dashboard 2: "Analisi demografico-temporale" . . . 57
4.2.5 Dashboard 3: "Profilazione sociologica della vittima" . . . 59
4.2.6 Dashboard 4: "Profilazione sociologica dell'autore del reato" . . . 62

4.3 Considerazioni sulle dashboard di base 66
4.3.1 Sintesi del percorso di analisi 66
4.3.2 Principali evidenze emerse dall'analisi 66
4.3.3 Valore della soluzione di Business Intelligence 67

5 Progettazione e realizzazione delle dashboard avanzate 68
5.1 Introduzione all'analisi avanzata 68
5.2 Analisi dashboard avanzate 69
5.2.1 Struttura dell'analisi visiva 70
5.2.2 Dashboard 5: "Analisi multidimensionale delle dinamiche relazionali ed esecutive" 70
5.2.3 Dashboard 6: "Analisi comportamentale, vulnerabilità e fattori di rischio" 73
5.2.4 Dashboard 7: "Evoluzione temporale dei trend critici" 77
5.3 Considerazioni sulle dashboard avanzate 81
5.3.1 Sintesi del percorso di analisi 81
5.3.2 Principali evidenze emerse dall'analisi 82
5.3.3 Valore della soluzione di Business Intelligence 83

6 Discussione 84
6.1 Discussione delle principali evidenze emerse 84
6.1.1 La centralità della sfera intima 85
6.1.2 La prevalenza del contesto domestico 85
6.1.3 Il movente del possesso e la lettura dei fattori psicologici 86
6.1.4 Le dinamiche successive all'evento 88
6.1.5 La normalizzazione del dato territoriale 89
6.2 Il valore della soluzione di Business Intelligence 90
6.3 Limiti dello studio 91
6.3.1 Limiti legati alla natura del dataset 91
6.3.2 Limiti temporali 91
6.3.3 Limiti dell'interpretazione statistica 91
6.3.4 Limiti nell'interpretazione delle variabili psicologiche 91
6.4 Sviluppi futuri 92
6.4.1 Ampliamento e aggiornamento delle fonti 92
6.4.2 Analisi statistica 92
6.4.3 Machine Learning e analisi predittiva 92
6.4.4 Evoluzione verso un sistema di supporto alle decisioni 93

Conclusioni 94

Bibliografia 96

Sitografia 98

Ringraziamenti 100

Elenco delle figure

1.1 Relazione concettuale tra dato, elaborazione e informazione 3
1.2 Piramide DIKW 6
1.3 Il modello delle 5V di Douglas Laney 8
1.4 Le quattro categorie di analisi 10
1.5 Allineamento tra livelli organizzativi e paradigma DIKW 12
1.6 I nove stadi del ciclo di vita della Big Data Analytics 12
1.7 Le componenti interdisciplinari della Data Science 15

2.1 Il Magic Quadrant di Gartner 18
2.2 Power BI Desktop 21
2.3 Power BI Service 22
2.4 Power BI Mobile 22
2.5 Rappresentazione architetturale del flusso ETL 24
2.6 Data Modeling: lo Schema a stella 25
2.7 Pannello delle visualizzazioni native in Power BI 26

3.1 Esempio della nomenclatura originaria delle colonne del dataset 38
3.2 Esempio di dati mancanti nel dataset 39
3.3 Sequenza dei passaggi di trasformazione applicati in Power Query 41
3.4 Esempio di traduzione e standardizzazione delle intestazioni delle colonne . 42
3.5 Risultato della trasformazione Unpivot dei fattori comportamentali 43
3.6 Relazione tra la Fact Table e la Dimension Table demografica 44

4.1 Dashboard 1: "Modus operandi e dinamiche del delitto" 55
4.2 Dashboard 2: "Analisi demografico-temporale" 58
4.3 Dashboard 3: "Profilazione sociologica della vittima" 60
4.4 Dashboard 4: "Profilazione sociologica dell'autore del reato" 63

5.1 Dashboard 5: "Analisi multidimensionale delle dinamiche relazionali ed esecutive" 71
5.2 Dashboard 6: "Analisi comportamentale, vulnerabilità e fattori di rischio" . . 74
5.3 Dashboard 7: "Evoluzione temporale dei trend critici" 77

Elenco delle tabelle

1.1 Vantaggi e svantaggi delle tipologie di dato 6

2.1 Confronto tra Tableau, Power BI e Qlik Sense 20

Introduzione

Il fenomeno del femminicidio rappresenta una realtà sociale complessa, caratterizzata dalla presenza di molteplici dimensioni che riguardano le caratteristiche delle vittime e degli autori, il contesto in cui si verificano gli eventi, le dinamiche relazionali e gli aspetti territoriali e temporali. La disponibilità di dati relativi a tali dimensioni offre la possibilità di affiancare alla lettura qualitativa del fenomeno un approccio quantitativo, orientato all'esplorazione delle distribuzioni, delle ricorrenze e delle relazioni osservabili tra le variabili. In questo contesto, l'impiego di metodologie di *Data Analytics* e di strumenti di *Business Intelligence* può contribuire a organizzare fonti informative eterogenee e a trasformare i dati disponibili in rappresentazioni utili per l'analisi. Il problema non riguarda, infatti, esclusivamente la quantità di informazioni disponibili, ma anche la loro struttura, qualità e possibilità di integrazione.

Nel caso analizzato, il *dataset* principale presenta caratteristiche e livelli di completezza differenti tra le variabili, oltre ad alcune criticità nella nomenclatura, nella gestione dei valori mancanti, nella rappresentazione dei fattori comportamentali e nella granularità delle informazioni relative alle vittime. A queste informazioni vengono affiancati dati demografici di fonte ISTAT, necessari per contestualizzare la distribuzione territoriale degli eventi rispetto alla popolazione residente e costruire indicatori normalizzati. La presenza di fonti differenti rende, quindi, necessario un processo preliminare di acquisizione, verifica, pulizia e trasformazione dei dati, prima di poter procedere alla loro analisi.

A partire da queste considerazioni, il presente lavoro si colloca nell'ambito della *Data Analytics* applicata a un fenomeno sociale complesso e ha l'obiettivo di progettare e realizzare un ambiente di *Business Intelligence* dedicato all'analisi dei dati relativi ai femminicidi in Italia. L'utilizzo di Power BI permette, in particolare, di integrare le fasi di preparazione e modellazione dei dati con la definizione di misure analitiche e la realizzazione di visualizzazioni interattive, rendendo possibile un'esplorazione articolata del *dataset*.

L'obiettivo non è fornire una spiegazione causale del fenomeno né sostituire le competenze proprie degli ambiti specialistici coinvolti, ma mettere a disposizione uno strumento analitico capace di evidenziare configurazioni ricorrenti, distribuzioni e associazioni osservabili nei dati, fornendo una base per ulteriori approfondimenti.

Il lavoro prevede, in una prima fase, l'analisi delle caratteristiche delle fonti informative e la valutazione della loro qualità, con particolare attenzione alle criticità che possono influenzare le successive elaborazioni. Successivamente, le sorgenti vengono sottoposte a un processo di *Extract, Transform e Load* (ETL) mediante Power Query, attraverso il quale vengono uniformate le denominazioni, vengono gestiti i valori mancanti, vengono riorganizzate alcune variabili e vengono definite le strutture necessarie alla costruzione del modello.

Il *dataset* principale viene organizzato come Fact Table, mentre i dati demografici ISTAT vengono integrati mediante una Dimension Table, consentendo di collegare gli eventi alle informazioni relative alla popolazione residente e di calcolare indicatori territoriali normalizzati. Sul modello così ottenuto vengono definite misure in linguaggio DAX e vengono progettate dashboard interattive attraverso Power BI.

Il percorso analitico viene articolato in un primo livello dedicato alle dashboard di base, orientato alla descrizione delle principali caratteristiche degli eventi, delle vittime, degli autori, della distribuzione temporale e territoriale e delle dinamiche relazionali. A questo livello segue una seconda fase costituita dalle dashboard avanzate, nella quale le diverse dimensioni del *dataset* vengono messe in relazione per approfondire le configurazioni osservate e analizzare più specificamente le dinamiche tra autore e vittima, i fattori comportamentali e l'evoluzione temporale di alcune categorie. Le evidenze emerse vengono, infine, discusse alla luce dei limiti della fonte e della natura descrittiva delle analisi, distinguendo l'individuazione di associazioni osservate dall'interpretazione causale. Il lavoro si conclude con la definizione dei principali limiti metodologici e informativi e con l'individuazione di possibili sviluppi futuri.

La presente tesi è strutturata come di seguito specificato:

- Nel Capitolo 1 saranno introdotti i principali concetti alla base della *Data Analytics*, approfondendo l'evoluzione del dato, i Big Data, le diverse categorie di analisi e il ciclo di vita della *Big Data Analytics*, fino ad arrivare al ruolo della *Data Science* e del *Machine Learning*.

- Nel Capitolo 2 saranno presentati Microsoft Power BI e il suo ruolo nell'ambito della *Business Intelligence* e della *Data Analytics*. Saranno analizzati l'architettura della piattaforma, il flusso di lavoro, il processo ETL, il *Data Modeling*, la *Data Visualization* e i principali linguaggi utilizzabili.

- Nel Capitolo 3 saranno descritti i dati utilizzati per la costruzione del modello, costituiti dal *dataset* principale relativo agli eventi e dai dati demografici ISTAT. Saranno, inoltre, esaminate la qualità dei dati e le principali criticità individuate, per poi descrivere il processo di ETL e la successiva organizzazione del modello.

- Nel Capitolo 4 saranno illustrate la definizione delle misure DAX e la progettazione delle prime quattro dashboard, dedicate all'analisi delle caratteristiche degli eventi, della distribuzione territoriale e temporale e dei profili della vittima e dell'autore.

- Nel Capitolo 5 sarà introdotto il secondo livello di analisi, nel quale le variabili saranno messe in relazione per esaminare le dinamiche relazionali tra autore e vittima, i fattori comportamentali e l'evoluzione temporale delle principali configurazioni osservate.

- Nel Capitolo 6, infine, saranno interpretate le principali evidenze emerse dalle dashboard di base e dalle dashboard avanzate e sarà analizzato il contributo dell'ambiente di *Business Intelligence*. Saranno, inoltre, discussi i limiti dello studio e saranno delineati i possibili sviluppi futuri.

CAPITOLO 1

Introduzione alla Data Analytics

Il presente capitolo offre una panoramica sulla Data Analytics, esaminando il percorso che trasforma il dato da semplice elemento quantitativo a risorsa strategica fondamentale. La trattazione esplora i criteri di classificazione dei dati e il processo di raffinamento descritto dalla piramide DIKW, attraverso cui l'elemento grezzo acquisisce progressivamente valore semantico e decisionale. Viene, successivamente, analizzato il cambio di paradigma imposto dall'avvento dei Big Data, esplorando in dettaglio le nove fasi che compongono il ciclo di vita della Big Data Analytics. Infine, il percorso si concluderà illustrando il passaggio verso ecosistemi avanzati, dove la Data Science e il Machine Learning intervengono come motori analitici indispensabili per estrarre pattern nascosti e generare un reale vantaggio competitivo.

1.1 Il dato: definizione, evoluzione e classificazione

Il dato rappresenta un elemento grezzo, una rappresentazione oggettiva e non elaborata della realtà fenomenica. Esso costituisce l'unità di base dell'osservazione e può manifestarsi in diverse forme: valori numerici, stringhe testuali, flussi audio o contenuti visivi. Sebbene i termini "dato" e "informazione" siano frequentemente sovrapposti nel linguaggio colloquiale, a livello concettuale presentano una netta dicotomia, rintracciabile fin dalle rispettive radici etimologiche. Il termine "dato", dal participio latino *datum*, è traducibile come "ciò che è dato", ovvero assunto come fatto passivo e oggettivo. Per contro, la parola "informazione", dal latino *informatio*, esprime l'azione di "dare forma" o "rappresentare nella mente". L'informazione rappresenta, dunque, il prodotto di un processo di elaborazione; essa nasce nel momento in cui il dato grezzo viene contestualizzato, strutturato e dotato di una valenza semantica, come schematizzato nella Figura 1.1.



Figura 1.1: Relazione concettuale tra dato, elaborazione e informazione

1.1.1 Le origini semantiche

La concezione moderna del "dato" come misurazione quantitativa della realtà si discosta profondamente dalle sue origini storiche. La prima formalizzazione tecnica del concetto si deve al matematico greco Euclide, che utilizzò il termine per definire gli elementi noti (punti, angoli o linee) forniti come punto di partenza di un problema geometrico, in contrasto logico con le grandezze incognite da determinare. Come evidenziato dallo storico Daniel Rosenberg nel saggio *Data before the fact* (MIT Press, 2013), il primo utilizzo documentato della parola in lingua inglese risale al 1646, all'interno dell'espressione "heap of data" coniata dal teologo Henry Hammond. All'epoca, il vocabolo manteneva fedelmente la sua valenza puramente latina ed euclidea: il dato non era il risultato empirico di una ricerca, ma un dogma o un assioma teologico. Si configurava come un elemento assunto per vero, accettato preventivamente dai contendenti prima di avviare un dibattito intellettuale. Di conseguenza, il lemma denotava un presupposto logico all'interno di un ragionamento deduttivo, collocandosi in netta antitesi con il significato contemporaneo di scoperta derivante dall'osservazione.

1.1.2 L'evoluzione quantitativa

Sotto l'impulso dell'Illuminismo, parallelamente al consolidamento dell'epistemologia empirista e alla nascita delle scienze sociali, il termine subì un mutamento decisivo. I dati cessarono di essere proposizioni filosofiche concesse a priori e divennero frammenti di realtà misurabile: informazioni numeriche catalogate in tabelle di mortalità, registri doganali e inventari statali, offrendo così una rappresentazione empirica e quantitativa della società. William Petty, economista e scienziato inglese, nel suo trattato *Treatise of Naval Philosophy* del 1691 usò la parola "data" per riferirsi alle misure geometriche necessarie a calcolare la resistenza dei materiali nelle costruzioni navali. In seguito, la nascita della misurabilità del dato trovò un impulso decisivo nei resoconti fisici e astronomici pubblicati tra le pagine di *Philosophical Transactions*; all'interno di questa prima rivista scientifica della storia, i valori empirici rilevati vennero, infatti, codificati stabilmente come *data*.

In questa fase di profonda transizione semantica si inserisce una figura cruciale evidenziata dalla ricerca di Rosenberg: Joseph Priestley, filosofo ed educatore del Settecento. Priestley creò delle carte cronologiche strutturate in un formato che oggi chiameremmo "linea del tempo" (*timeline*). La vera svolta lessicale e concettuale risiede nel fatto che Priestley iniziò a definire gli eventi della storia non più come semplici "fatti" (*facts*), bensì proprio come "dati" (*data*). Il dato, dunque, smetteva di essere un'esclusiva del calcolo matematico puro per diventare un'unità informativa universale, codificabile e, soprattutto, rappresentabile visivamente. Questa evoluzione trovò la sua consacrazione ufficiale nel 1778, quando, per la prima volta in assoluto in un'opera enciclopedica, i redattori della *Encyclopaedia Britannica* misero nero su bianco che la parola *data* non apparteneva più esclusivamente alla matematica pura, ma si era radicata nella medicina e nella filosofia per indicare qualsiasi quantità numerica o fatto assunto come valido per poter effettuare un calcolo.

1.1.3 La digitalizzazione

Nel saggio *A Mathematical Theory of Communication* (1948) il matematico e ingegnere Claude Shannon, padre della Teoria dell'Informazione, sancisce il concetto di "Bit" (binary digit). Viene, così, introdotta per la prima volta un'unità di misura universale e rigorosa per l'informazione. Attraverso questa astrazione, il dato viene definitivamente slegato dal suo supporto fisico o analogico per trasformarsi in una sequenza astratta di stati logici, 0 e 1. Di conseguenza, il dato diventa un'entità autonoma, capace di essere trasferita da un sistema all'altro istantaneamente, senza più alcun vincolo a una forma materiale specifica.

La digitalizzazione ha rivoluzionato i paradigmi di archiviazione scientifica e aziendale, determinando il passaggio dai supporti fisici, come le schede perforate, alle memorie magnetiche. Il culmine di questo processo si raggiunge nel 1970, quando Edgar F. Codd pubblica l'articolo *A Relational Model of Data for Large Shared Data Banks*. In questo articolo, Codd teorizza un modello matematico basato sulla teoria degli insiemi, mostrando come i dati digitalizzati possano essere organizzati logicamente in tabelle. Viene, così, formalizzato il principio cruciale dell'indipendenza dei dati: la struttura logica dell'informazione viene, per la prima volta, del tutto separata dalla sua memorizzazione fisica sull'hardware. Questa intuizione ha consentito lo sviluppo dei moderni Database Management Systems (DBMS), sistemi in grado di gestire, aggregare e incrociare enormi volumi informativi in tempo reale. È grazie a questa evoluzione architetturale che l'archivio cessa di essere un semplice contenitore statico per trasformarsi in una risorsa dinamica e interrogabile, ponendo, di fatto, le fondamenta per la nascita della moderna Data Analytics.

1.1.4 Tipologie e classificazione

L'astrazione teorizzata da Codd e la successiva diffusione dei sistemi DBMS relazionali hanno tracciato definito le basi per l'organizzazione dell'informazione digitale. Nell'ambito della moderna Data Analytics, è diventato imprescindibile classificare i dati in tre macrocategorie fondamentali, distinte in base al loro livello di organizzazione formale e strutturale:

- *Dati strutturati:* rappresentano l'evoluzione diretta del modello relazionale. Si tratta di dati organizzati secondo uno schema logico rigido, predefinito e tipizzato, rappresentato attraverso tabelle composte da righe e colonne. Questa formattazione rende i dati strutturati facilmente archiviabili nei tradizionali RDBMS (Relational Database Management System) e nativamente predisposti per essere incrociati tramite linguaggi standardizzati come l'SQL (Structured Query Language).

- *Dati semi-strutturati:* costituiscono un formato informativo ibrido. Pur non possedendo un rigido schema tabellare, presentano una struttura interna auto-descrittiva basata su marcatori (tag) o coppie chiave-valore che gerarchizzano semanticamente l'informazione. Formati come l'XML (eXtensible Markup Language) o il JSON (JavaScript Object Notation) rientrano in questa categoria. Questi formati offrono una flessibilità architetturale che li ha resi lo standard per lo scambio di informazioni sul Web e per la comunicazione tramite API (Application Programming Interface).

- *Dati non strutturati:* comprendono tutte quelle informazioni prive di un modello di dati predefinito e di un'organizzazione formale. Rientrano in questa categoria i documenti di testo libero, le e-mail, le interazioni su social network e i file multimediali, come audio, video e immagini. Pur rappresentando la maggioranza del volume informativo prodotto a livello globale, la loro natura eterogenea ne rende complessa l'elaborazione, richiedendo architetture NoSQL e tecniche algoritmiche avanzate per estrarne valore analitico.

Al fine di fornire una panoramica chiara sulle implicazioni architetturali di ciascuna categoria, la Tabella 1.1 propone un confronto dei principali vantaggi e svantaggi associati ai tre tipi di dato.

Tipologia	Vantaggi	Svantaggi
Dati strutturati	interrogazione immediata e standardizzata (SQL); elevata integrità e consistenza informativa; ottimizzati per le analisi relazionali e storiche.	rigidità dello schema logico; complessità nell'integrare nuovi attributi; difficoltà nella scalabilità orizzontale dei sistemi.
Dati semi-strutturati	elevata flessibilità (assenza di schema rigido); standard di fatto per comunicazione Web e API; rappresentazione gerarchica naturale.	maggiore *overhead* di archiviazione a causa dei tag; minore efficienza nelle interrogazioni rispetto ai database relazionali.
Dati non strutturati	rappresentazione ricca e fedele della realtà; totale assenza di vincoli di formato; rappresentano il bacino informativo quantitativamente più ampio.	altissima complessità di estrazione del valore analitico; necessità di architetture specifiche (NoSQL) e algoritmi avanzati; difficoltà di validazione e pulizia.

Tabella 1.1: Vantaggi e svantaggi delle tipologie di dato

1.1.5 La piramide DIKW

Una volta definite le varie tipologie di dato, è necessario esaminare le relazioni gerarchiche relative alla transizione dal dato grezzo alla decisione strategica. Il quadro teorico di riferimento per descrivere questa evoluzione è rappresentato dalla piramide DIKW (Data, Information, Knowledge, Wisdom). Sebbene le origini dello schema facciano riferimento a una celebre suggestione poetica di T.S. Eliot del 1934, la sua formalizzazione scientifica si consolida negli anni Ottanta grazie alle formulazioni di Harlan Cleveland (1982), Milan Zeleny (1987) e Russell L. Ackoff (1988). Il modello viene rappresentato mediante una forma piramidale: una scelta che illustra visivamente un processo di raffinamento e distillazione progressiva, in cui l'ampia base dei dati grezzi si restringe e si eleva all'aumentare dei livelli di contestualizzazione, sintesi e valore semantico. Il modello, come mostrato in Figura 1.2, presenta quattro livelli.

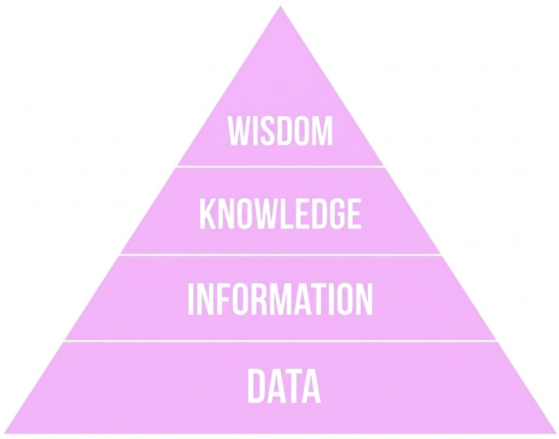


Figura 1.2: Piramide DIKW

- *Dato:* costituisce lo strato fondamentale alla base della piramide. Secondo la definizione "diaforica" del filosofo Luciano Floridi, un dato può essere inteso come una "mancanza

di uniformità" o una differenza nel tessuto della realtà. Si tratta di elementi che, in assenza di un'interpretazione esterna, si presentano come uno sfondo neutro privo di un significato autonomo.

- *Informazione:* rappresenta il livello in cui al dato grezzo viene attribuito un significato sulla base di un preciso contesto interpretativo. Riprendendo una celebre formulazione attribuita a Gregory Bateson, l'informazione è comunemente intesa come "una differenza che genera una differenza". Dunque i dati si trasformano in informazioni solo quando vengono riconosciuti e decodificati, modificando, così, la comprensione del ricevente o la struttura di un sistema.
- *Conoscenza:* associata a una competenza o capacità d'azione pratica (*know-how*), oppure definita in senso filosofico come una "credenza vera e giustificata" (*know-that*). Essa deriva dall'assimilazione, dall'organizzazione e dalla riflessione sulle informazioni, che si trasformano da elementi isolati a strutture cognitive integrate.
- *Saggezza:* costituisce il vertice della piramide. Essa trascende l'acquisizione di competenze per includere un uso critico, valutativo ed etico della conoscenza. Considerata quasi sempre una caratteristica prettamente umana, la saggezza implica la capacità di giudizio e la consapevolezza dei limiti della propria conoscenza, risultando difficilmente trasferibile a sistemi di Intelligenza Artificiale.

1.2 I Big Data: i nuovi paradigmi architetturali

Come evidenziato dal modello DIKW, la trasformazione del dato in saggezza decisionale richiede un costante processo di elaborazione. Tuttavia, la recente evoluzione tecnologica ha generato volumi informativi che hanno rapidamente saturato le capacità di archiviazione e analisi dei tradizionali sistemi relazionali, delineando il passaggio al paradigma dei Big Data. Il termine "Big Data" compare per la prima volta nel 1997 all'interno di un documento redatto da scienziati della NASA, utilizzato per descrivere le sfide computazionali e i limiti hardware riscontrati nella gestione della grafica computerizzata. Il concetto ha, poi, iniziato ad acquisire popolarità su scala globale intorno al 2008, quando la comunità informatica ha intuito che questa mole di dati avrebbe radicalmente trasformato le attività aziendali e la ricerca scientifica.

1.2.1 L'evoluzione della produzione massiva dei dati

Nel corso degli anni Novanta, l'architettura dei sistemi informativi poggiava su un modello di inserimento dati prevalentemente manuale. In tale contesto, la creazione del dato era un'azione intenzionale e limitata a operatori specializzati, i quali alimentavano i database relazionali. L'introduzione del Web 2.0 ha rivoluzionato il ruolo dell'utente: da produttore passivo di contenuti a creatore attivo di informazioni (Prosumer). La struttura della rete ha integrato lo strato della sensoristica IoT (Internet of Things): miliardi di oggetti interconnessi generano flussi continui di metriche ambientali e operative. Si è compiuta in questo modo una mutazione cruciale: dalla generazione manuale e attiva del dato si è passati a una produzione massiva e passiva, in cui la realtà fenomenica stessa viene tradotta in flussi algoritmici di dimensioni macroscopiche.

Fornire una definizione univoca di questo fenomeno si è rivelato complesso. I redattori dell'*Oxford English Dictionary* sono stati tra i primi a tentare una classificazione, definendoli come dati di dimensioni talmente vaste da presentare rilevanti criticità di gestione e manipolazione computazionale. Tuttavia, i Big Data non devono essere inquadrati come

una semplice grandezza quantitativa, bensì come un insieme eterogeneo di informazioni complesse che richiedono metodologie, architetture infrastrutturali e competenze analitiche specifiche per essere raccolte, elaborate e trasformate in valore. Ormai i dati sono universalmente riconosciuti come il petrolio di questo millennio: per tradursi in un reale vantaggio competitivo, richiedono complessi processi di raccolta, raffinamento e organizzazione.

1.2.2 Il modello delle 5V

Nel 2001, l'analista Douglas Laney ha formalizzato il fenomeno dei Big Data introducendo il celebre modello delle "3V". Tale paradigma è stato successivamente ampliato fino a comprendere cinque sezioni, note oggi come le "5V" dei Big Data: *Volume*, *Velocità*, *Varietà*, *Veridicità* e *Valore*, illustrate in Figura 1.3.

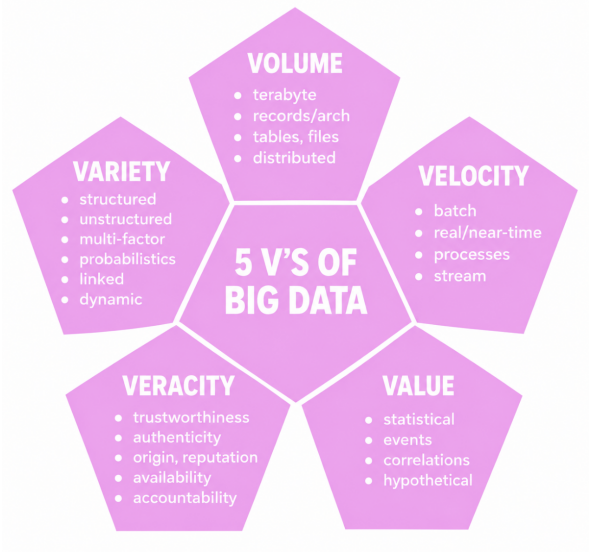


Figura 1.3: Il modello delle 5V di Douglas Laney

- *Volume:* rappresenta la quantità macroscopica di informazioni. L'unità di misura per i sistemi di *storage* si è spostata dai Gigabyte e Terabyte verso gli Zettabyte. Si stima che oggi, in un solo minuto, venga creato un volume di dati pari a tutte le informazioni prodotte dall'invenzione della scrittura fino al 2008. Questa crescita esponenziale ha reso obsoleti i server tradizionali, imponendo l'adozione di architetture *cloud* e sistemi di calcolo distribuito.
- *Velocità:* indica la rapidità con cui i dati vengono creati, raccolti e trasmessi. Anche questa metrica è soggetta a una tendenza esponenziale. Questo è evidente nel marketing digitale; le aste per gli annunci online si risolvono in meno di 10 millisecondi, lasso di tempo in cui gli algoritmi analizzano e trasferiscono i profili degli utenti generando milioni di previsioni al secondo. Inoltre, in ambito finanziario, l'*High-Frequency Trading* (HFT) opera ormai nell'ordine dei nanosecondi. Tale frequenza impone lo sviluppo di architetture in grado di elaborare flussi in tempo reale (*real-time analytics*), limitando la necessità di archiviazione statica.
- *Varietà:* riflette l'eterogeneità dei formati acquisiti. Come anticipato, sebbene esistano dati strutturati tabellari, attualmente circa l'80% del volume globale è composto da dati non strutturati: testi grezzi, immagini e video. I metodi statistici tradizionali e i DBMS relazionali risultano inefficaci di fronte a questa destrutturazione, richiedendo l'implementazione di nuovi algoritmi di apprendimento automatico (*Machine Learning*).

- *Veridicità:* si riferisce al grado di affidabilità, pulizia e qualità dell'informazione. Considerando l'acquisizione massiva e passiva da fonti eterogenee, i Big Data sono affetti da rumore di fondo, anomalie e dati mancanti. È, dunque, inevitabile dover sottoporre i *dataset* a rigorosi processi di *Data Cleansing* e validazione algoritmica.
- *Valore:* costituisce il fine ultimo dell'intero processo, ovvero la capacità di estrarre profitto e vantaggio competitivo da questi enormi set informativi. Il potenziale economico derivante dall'analisi predittiva sta spingendo ogni settore ad adottare processi di *Big Data Analytics*. Basti pensare che se nel 2009 l'intero mercato globale legato a questi progetti generava circa 100 milioni di euro, nel 2018 ha raggiunto quasi 42 miliardi, con proiezioni che stimano il superamento dei 100 miliardi di euro entro il 2030.

1.2.3 Cambiamento di paradigma tecnologico, scientifico ed economico

L'incremento informativo analizzato dal modello precedente, con particolare riferimento alle dimensioni del *Volume* e della *Varietà*, ha imposto una riorganizzazione radicale delle infrastrutture fisiche e logiche. Se inizialmente la gestione di questi carichi di lavoro è stata guidata dal naturale aumento della potenza di calcolo e delle capacità di archiviazione, come teorizzato dalla Legge di Moore, il raggiungimento di soglie critiche ha reso i tradizionali database locali insostenibili. Per le organizzazioni è diventato indispensabile esternalizzare l'archiviazione e l'analisi verso server remoti, decretando l'affermazione del paradigma del *Cloud Computing*.

È la *Varietà* dei dati ad aver stimolato l'adozione su larga scala dell'apprendimento statistico. La maggior parte dei moderni algoritmi di *Machine Learning* eccelle nella risoluzione di complessi problemi di classificazione su *dataset* eterogenei, "imparando" a riconoscere autonomamente *pattern* nel tempo. Le applicazioni pratiche sono innumerevoli: dai filtri che categorizzano le e-mail come posta indesiderata in base all'analisi del testo grezzo, fino alla ricerca medica, in cui i modelli predittivi supportano i processi decisionali classificando la natura benigna o maligna di un tumore intersecando dimensioni, colore e altre variabili cliniche. Questa natura transdisciplinare sancisce un definitivo cambio di paradigma tecnologico, scientifico ed economico per la nostra società.

1.3 Dalla Data Analysis alla Data Analytics

A causa della crescita dell'utilizzo dei Big Data, i termini *Data Analytics* e *Data Analysis* vengono spesso intercambiati. Purtroppo, nella lingua italiana questa distinzione fatica a emergere, poiché entrambe le espressioni vengono tradotte con la singola parola "analisi". La *Data Analytics* è una disciplina molto ampia, contenente tutti i processi riguardanti il ciclo di vita dei dati. D'altra parte, la *Data Analysis* è un singolo aspetto della *Data Analytics* e riguarda il processo di ispezione dei dati allo scopo di trasformarli in informazioni utili nel processo decisionale. L'obiettivo di quest'ultima è quello di trovare *trend*, relazioni, *pattern* e previsioni future.

1.3.1 Categorie della Data Analytics

Il ciclo di vita della *Big Data Analytics* comporta generalmente l'identificazione, l'acquisizione, la preparazione e l'analisi di grandi quantità di dati grezzi e non strutturati. Le tecniche di analisi vengono adottate da diverse organizzazioni con finalità differenti: negli ambienti orientati al *business* servono ad abbassare i costi operazionali e a facilitare il *decision-making* strategico; in ambito scientifico aiutano a identificare la causa di determinati fenomeni; negli enti pubblici permettono di rafforzare la qualità dei servizi abbassandone i costi. Il fine ultimo

è potenziare le decisioni fornendo loro un substrato scientifico basato sui fatti, superando i limiti della mera esperienza passata o dell'intuito.

A seconda della complessità, del valore generato e dei risultati prodotti, le elaborazioni adottano diverse tecniche e si distinguono in quattro categorie generali:

- *Analisi Descrittiva;*
- *Analisi Diagnostica;*
- *Analisi Predittiva;*
- *Analisi Prescrittiva.*

Come mostrato in Figura 1.4, queste quattro metodologie si posizionano lungo un duplice asse di crescita. Procedendo dal livello descrittivo di base verso il vertice prescrittivo, si osserva un incremento progressivo sia della complessità tecnica (*complexity*), che passa da un grado basso a uno elevato, sia del valore strategico generato per l'organizzazione (*value*). Dall'altra parte, l'orizzonte analitico parte da un'osservazione degli eventi passati (*hindsight*), passa attraverso una comprensione profonda delle dinamiche in atto (*insight*) e culmina, infine, nella capacità di visione e anticipazione futura (*foresight*).

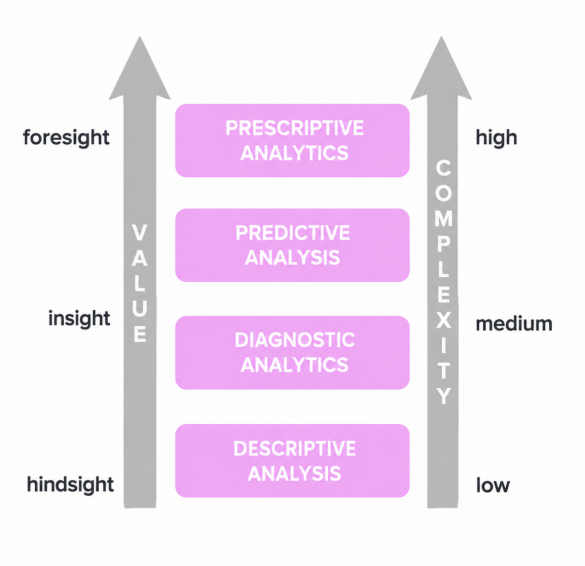


Figura 1.4: Le quattro categorie di analisi

Analisi Descrittiva

L'Analisi Descrittiva risponde a domande sugli eventi che sono già accaduti, contestualizzando i dati storici per generare informazioni chiare. Risponde a interrogativi di base, come calcolare il volume delle vendite, il numero di richieste di supporto suddivise per area geografica o le commissioni degli agenti. Si stima che l'80% dei risultati generati tramite la *Data Analytics* appartenga a questa tipologia. Pur fornendo il minor valore aggiunto, essa richiede un insieme di competenze relativamente semplice per essere eseguita. Le interrogazioni (*query*) vengono tipicamente eseguite su depositi di dati operazionali (come CRM o ERP) e i risultati assumono una natura statica, visualizzati tramite *dashboard* o griglie all'interno dei sistemi di *reporting*.

Analisi Diagnostica

L'Analisi Diagnostica si spinge oltre la semplice osservazione, con lo scopo di determinare la causa di un fenomeno accaduto nel passato. Essa cerca di correlare le informazioni per comprendere le ragioni sottostanti, rispondendo a domande sul perché le vendite siano calate in un dato trimestre o perché ci sia stato un incremento anomalo di richieste da una specifica area. Questo approccio fornisce più valore rispetto all'analisi descrittiva, ma richiede competenze nettamente più avanzate. Essa necessita, infatti, della raccolta di dati da sorgenti multiple, organizzate in strutture multi-dimensionali (sistemi OLAP) che permettono operazioni di approfondimento e aggregazione, come il *drill-down* e il *roll-up*. I risultati vengono fruiti tramite strumenti di visualizzazione interattiva che consentono agli utenti di identificare direttamente *trend* e *pattern* anomali.

Analisi Predittiva

L'Analisi Predittiva viene effettuata nel tentativo di determinare il risultato di un evento che potrebbe accadere nel futuro. Durante questa fase l'informazione viene potenziata con un significato volto a generare conoscenza, sfruttando la forza e l'intensità delle associazioni passate per costruire modelli in grado di effettuare predizioni. È di fondamentale importanza comprendere che tali modelli dipendono implicitamente dalle condizioni in cui sono avvenuti gli eventi storici: qualora queste condizioni dovessero mutare, i modelli predittivi devono essere obbligatoriamente aggiornati. Rispondendo a interrogativi di tipo *what-if*, questa analisi identifica rischi e opportunità su grossi *dataset* interni ed esterni. Pur basandosi su calcoli complessi, gli strumenti moderni cercano di astrarre le complicazioni statistiche sottostanti offrendo interfacce intuitive.

Analisi Prescrittiva

L'Analisi Prescrittiva si basa sui risultati dell'Analisi Predittiva e rappresenta il livello massimo di maturità analitica, avendo lo scopo di prescrivere le azioni che dovrebbero essere intraprese. L'obiettivo non è solo indicare l'azione migliore, ma spiegarne anche il motivo, fornendo risultati su cui si possono effettuare ragionamenti logici per comprendere lo scenario sottostante. Valutando vari esiti e simulando l'impatto di molteplici scenari, questo modello può essere usato per ottenere un vantaggio competitivo o per mitigare un forte rischio. A livello infrastrutturale, la complessità è massima: l'Analisi Prescrittiva incorpora regole di *business* intrecciando enormi quantità di dati interni (vendite storiche, soddisfazione del cliente, catalogo prodotti) con dati esterni (*social network*, previsioni meteorologiche, dati demografici governativi). Richiedendo *software* specializzati e le competenze più avanzate in assoluto, essa rappresenta la tipologia di analisi che può generare il maggior valore per l'organizzazione.

1.3.2 Data Analytics e organizzazione aziendale

Un'azienda opera tipicamente come un sistema stratificato, strutturato su tre livelli gerarchici interdipendenti: uno strato strategico di vertice, uno strato tattico o manageriale e uno strato operativo di base in cui vengono eseguiti i processi fondamentali e viene rilasciato valore ai clienti. Questi livelli operano a velocità differenti: il livello strategico si muove più lentamente per fornire stabilità e dettare la direzione generale, mentre il livello operativo è quello che muta più rapidamente. La connessione tra l'architettura di *business* e i Big Data si concretizza richiamando concettualmente il paradigma DIKW, come illustrato in Figura 1.5.

Da una prospettiva di monitoraggio (*bottom-up*), il livello operativo registra cosa sta succedendo nell'azienda generando i *Performance Indicator* (PI) e le relative metriche per

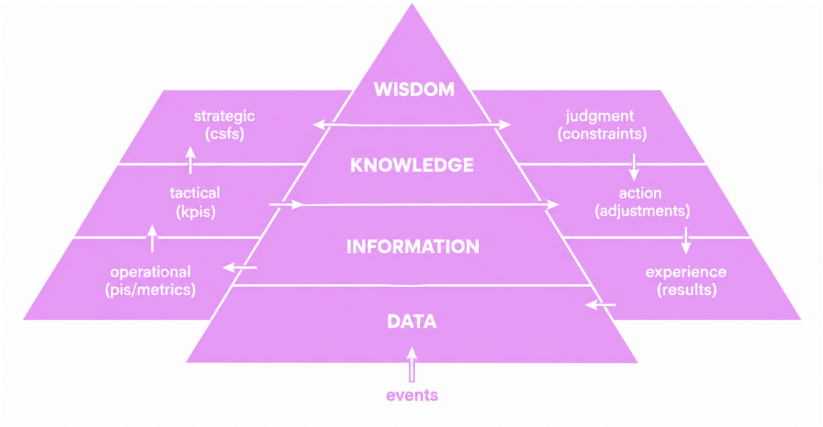


Figura 1.5: Allineamento tra livelli organizzativi e paradigma DIKW

servizi e processi. Queste misurazioni vengono successivamente aggregate al livello tattico sotto forma di *Key Performance Indicator* (KPI), strumenti utili a esaminare come viene condotto il *business*. Infine i KPI si interfacciano con i *Critical Success Factor* (CSF), fornendo le intuizioni necessarie per valutare il progresso fatto verso gli obiettivi e stabilire quali strategie debbano essere modificate o adottate per correggere o potenziare le *performance*.

Si innesca un vero e proprio ciclo virtuoso (*loop* di *feedback*): le decisioni e i giudizi del vertice strategico vengono comunicati come vincoli al livello tattico, il quale utilizza questa conoscenza per generare priorità e azioni.

1.4 Ciclo di vita della Big Data Analytics

La *Big Data Analytics* differisce dall'analisi dei dati tradizionale a causa delle specifiche caratteristiche di *Volume*, *Velocità* e *Varietà* dei dati processati. In generale, il ciclo di vita della *Big Data Analytics* può essere suddiviso in nove stadi sequenziali e iterativi, come illustrato nella Figura 1.6.

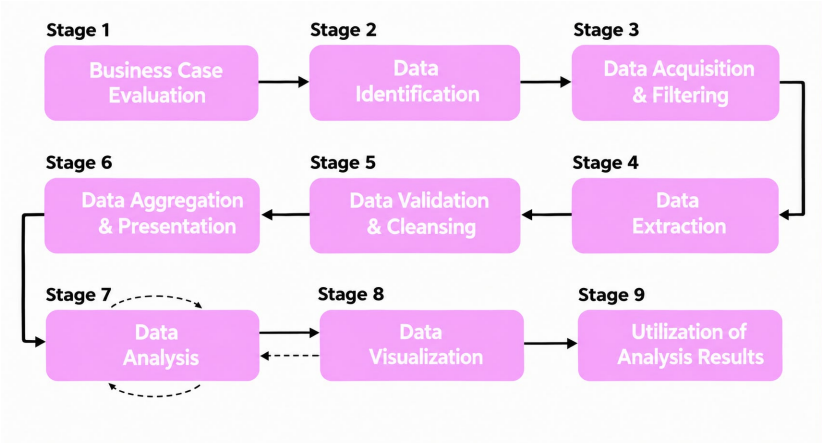


Figura 1.6: I nove stadi del ciclo di vita della Big Data Analytics

1.4.1 Fase 1: Business Case Evaluation

Ciascun progetto di analisi deve iniziare con un *business case* ben definito, creato, validato e approvato prima di procedere con le attività tecniche. Questa fase permette ai decisori

aziendali di comprendere le risorse necessarie e le sfide da affrontare. Durante questo stadio vengono identificati i KPI (che devono essere SMART: specifici, misurabili, raggiungibili, rilevanti e tempestivi) per definire i criteri di validazione dei risultati. Inoltre si determina se il problema di *business* da risolvere sia qualificabile come un reale problema di Big Data, presentando una o più caratteristiche di quest'ultimo. In questo stadio viene valutato anche il *budget* richiesto per *hardware*, *tool* e formazione, considerando che le iterazioni iniziali richiederanno investimenti anticipati maggiori.

1.4.2 Fase 2: Data Identification

Una volta definito il *business case*, si procede all'identificazione dei dati necessari. In caso di *dataset* esterni, viene compilata una lista di possibili *data provider* di terze parti, come *data market* o *dataset* pubblici. In alcuni casi, i dati esterni possono essere recuperati tramite *tool* automatici poiché immersi in *blog* o altri siti *web* basati sul contenuto.

1.4.3 Fase 3: Data Acquisition and Filtering

I dati vengono recuperati dalle sorgenti identificate tramite collezioni di file o integrazioni API. Essi subiscono un filtraggio per rimuovere il rumore, i dati classificati come corrotti e per ottimizzare lo spazio, mantenendo sempre una copia di *backup* compressa del *dataset* originale. Per quanto riguarda la persistenza, nell'analitica *batch* i dati vengono salvati su disco prima dell'analisi, mentre nell'analitica in tempo reale vengono prima analizzati e poi resi persistenti. A questi dati vengono automaticamente associati dei metadati comprensibili dalla macchina, fondamentali per preservare la qualità, l'accuratezza e la tracciabilità delle informazioni lungo tutto il ciclo di vita.

1.4.4 Fase 4: Data Extraction

Poiché i dati in ingresso, specialmente se di origine esterna, possono presentare formati incompatibili, questo stadio è dedicato alla loro estrazione e trasformazione nel formato nativo richiesto dalla soluzione di Big Data. La complessità dell'estrazione dipende fortemente dalle capacità di elaborazione del sistema sottostante.

1.4.5 Fase 5: Data Validation and Cleansing

Come visto nella sezione precedente, a differenza dei database tradizionali, nei Big Data i dati spesso arrivano non strutturati e privi di indicazioni di validità predefinite. Dati non validi possono inquinare l'analisi; pertanto, si stabiliscono regole di validazione complesse per ripulire il *dataset*. Questo processo può avvenire tramite operazioni ETL *offline* o tramite sistemi *in-memory* complessi. La ridondanza derivante da *dataset* differenti viene sfruttata per riempire dati mancanti o assemblare parametri di validità; inoltre, anche dati apparentemente anomali o non validi non vengono scartati a priori, in quanto potrebbero celare indicatori di nuovi *pattern* nascosti.

1.4.6 Fase 6: Data Aggregation and Representation

I dati sparsi su più sorgenti vengono integrati e riconciliati per ottenere una visione unificata, collegandoli attraverso campi comuni. Questa fase è complessa a causa di differenze riguardanti:

- *la struttura dei dati:* sebbene il formato dei dati possa essere lo stesso, il modello dei dati può essere differente.

- *la semantica:* un valore che viene etichettato in maniera differente in due diversi *dataset* può significare la stessa cosa.

Si stabilisce una struttura dati standardizzata capace di fungere da denominatore comune. A seconda della necessità di analisi, i dati possono essere memorizzati in più formati, con struttura a colonne a colonne o in BLOB.

1.4.7 Fase 7: Data Analysis

Questo è lo stadio dell'elaborazione vera e propria ed è intrinsecamente iterativo. Può variare da semplici aggregazioni fino a complesse tecniche statistiche e di *data mining* per la scoperta di anomalie. L'analisi si divide in due approcci principali:

- *confermativa:* un approccio deduttivo in cui si parte da una causa o un'ipotesi predeterminata; i dati vengono analizzati per confermare o respingere tale ipotesi, fornendo risposte definitive.
- *esplorativa:* un approccio induttivo strettamente associato al *data mining*, in cui non vi sono ipotesi iniziali; i dati vengono esplorati per sviluppare una comprensione e definire direzioni generali per la scoperta di *pattern*.

1.4.8 Fase 8: Data Visualization

L'analisi non ha alcun valore aziendale se i soli che possono interpretare i risultati sono gli analisti. Da tal ragione, durante questa fase, vengono utilizzate tecniche e strumenti di visualizzazione grafica per comunicare efficacemente gli *insight* agli utenti aziendali. È fondamentale fornire un metodo di *drill-down* per confrontare statistiche semplici e permettere agli utenti di comprendere come i risultati aggregati sono stati generati. Una buona visualizzazione abilita gli utenti a porre domande inesplorate e a fornire un riscontro fondamentale che può rimettere in circolo l'analisi verso lo Stadio 7.

1.4.9 Fase 9: Utilization of Analysis Results

L'ultimo stadio determina come e dove i dati elaborati verranno utilizzati operativamente. I risultati possono assumere la forma di modelli matematici o regole logiche e vengono impiegati per:

- *integrazione nei sistemi enterprise:* potenziamento automatico delle *performance* di sistema;
- *ottimizzazione dei business process:* affinamento dei processi produttivi e logistici sfruttando i *pattern* identificati;
- *sistemi dialerting:* creazione di avvisi e notifiche per informare gli utenti della necessità di intraprendere tempestive azioni correttive.

1.5 Dai Big Data alla Data Science e al Machine Learning

L'avvento di ecosistemi informativi complessi ha innescato una profonda evoluzione tecnologica e metodologica nella ricerca e nella teoria manageriale. Questo percorso, che parte dall'accumulo dell'informazione fino ad arrivare all'automazione decisionale, si articola in tre stadi fondamentali e strettamente interconnessi: i Big Data, la *Data Science* e il *Machine Learning*.

1.5.1 Dai Big Data alla Data Science, fino al Machine Learning

I Big Data costituiscono il fondamento infrastrutturale e la materia prima di questo nuovo paradigma. Superando i limiti dell'analisi *post hoc* tradizionale, basata su campioni ristretti e rilevazioni distanti nel tempo, i Big Data introducono due dimensioni cruciali per la ricerca: la *Data Scope* e la *Data Granularity*. La prima, legata al volume, consente di analizzare popolazioni olistiche nella loro interezza, monitorando, ad esempio, le comunicazioni *e-mail* o il traffico *web* di un'intera organizzazione. La seconda, legata alla varietà, permette misurazioni teoricamente più prossimali al fenomeno studiato, sostituendo i classici questionari retrospettivi con dati continui e in tempo reale, come i parametri biometrici o il riconoscimento facciale delle emozioni. Tuttavia, la disponibilità di immensi *dataset* non genera valore senza un impianto metodologico adeguato. In questa fase di transizione interviene la *Data Science*, la disciplina che fornisce il rigore scientifico e le architetture necessarie per gestire l'intero ciclo di vita dell'informazione. Essa affronta le sfide pratiche attraverso un vero e proprio insieme di strumenti operativi: dall'acquisizione tramite sensori e *web scraping*, fino all'archiviazione su *database* distribuiti (come Apache Hadoop) e all'elaborazione di dati non strutturati. La *Data Science* sposta, dunque, l'attenzione dalla sfida tecnologica della raccolta dati alla sfida analitica della loro interpretazione.

1.5.2 Il ruolo del Machine Learning

Il vertice di questa evoluzione è rappresentato dal *Machine Learning*, il motore analitico che permette alla *Data Science* di estrarre *pattern* nascosti e formulare previsioni sui Big Data. Quando il numero di variabili diventa troppo elevato per le inferenze statistiche e la ricerca della causalità si fa complessa, subentrano gli algoritmi di apprendimento automatico. La collocazione di tali algoritmi all'intersezione tra informatica e matematica, all'interno del più ampio ecosistema della *Data Science*, è schematizzata nella Figura 1.7.

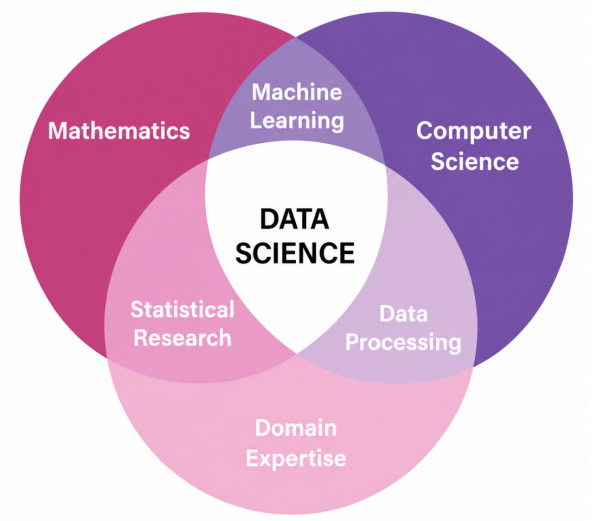


Figura 1.7: Le componenti interdisciplinari della Data Science

Modelli avanzati come le *Ridge Regression*, le *Lasso Regression*, i *Regression Trees*, le reti neurali *Deep Learning* e i modelli di *Topic Modeling* diventano strumenti indispensabili per processare enormi quantità di dati. Queste metodologie consentono non solo di affrontare problemi manageriali complessi, ma anche di esplorare nuove prospettive di analisi. Integrando strumenti come il *Text Mining*, risulta, ad esempio, possibile tracciare la genesi delle idee o utilizzare l'analisi del *Sentiment* per indagare in modo oggettivo le dinamiche decisionali del *management*.

CAPITOLO 2

Introduzione a Power BI

Il presente capitolo si propone di fornire una panoramica completa su Microsoft Power BI, definendone le caratteristiche all'interno del dominio della Business Intelligence e della Data Analytics. L'esposizione si articola seguendo un percorso logico che parte dall'evoluzione storica della piattaforma, per poi esaminarne il posizionamento competitivo attraverso l'analisi del Magic Quadrant di Gartner e il confronto diretto con i principali competitor del settore, quali Tableau e Qlik. Successivamente, verrà indagata l'architettura ibrida e distribuita della soluzione, esplorando l'integrazione tra i macro-componenti Desktop, Service e Mobile. Un'ampia sezione sarà dedicata al flusso di lavoro operativo: dai processi di estrazione, trasformazione e caricamento (ETL) gestiti da Power Query, passando per il Data Modeling ottimizzato dal motore in memoria VertiPaq, fino ad arrivare ai principi di Data Visualization e reportistica interattiva. Infine, la trattazione si concluderà esplorando i molteplici paradigmi di condivisione e i linguaggi di programmazione integrati nell'ecosistema, analizzando il ruolo specifico di SQL, M, DAX, Python ed R nelle diverse fasi di elaborazione e analisi del dato.

2.1 Cos'è Power BI

Power BI è la piattaforma di analisi aziendale di Microsoft che si inserisce nel dominio della *Business Intelligence* e della *Data Analytics*. Questo ambiente permette di trasformare enormi volumi di dati in report dettagliati immersivi e interattivi, indipendentemente dal fatto che i dati risiedano in un semplice foglio di calcolo, in un database relazionale o in un'architettura cloud complessa. Difatti, Power BI funge da ponte analitico, facilitando il ciclo di vita dell'informazione descritto nel capitolo precedente.

2.1.1 La storia e l'evoluzione di Power BI

La storia di Power BI inizia nei primi anni Dieci del Duemila, inizialmente sotto forma di estensione delle capacità analitiche di Microsoft Excel. Il progetto è nato dal team di sviluppo di SQL Server Reporting Services, sotto il nome di "Project Crescent". Reso disponibile al pubblico nel 2011, il prodotto mirava a superare i limiti di staticità e quelli computazionali dei tradizionali fogli di calcolo.

Nel 2013, la piattaforma è stata ufficialmente rinominata "Power BI for Office 365". In questa fase embrionale, l'ambiente di lavoro si basava essenzialmente sull'integrazione di tre *add-in* all'interno di Excel:

- *Power Query:* dedicato all'estrazione e alla trasformazione dei dati;

- *Power Pivot:* ottimizzato per la modellazione relazionale in memoria;
- *Power View:* focalizzato sulla *Data Visualization*.

Il punto di svolta, che ha consolidato la posizione del prodotto come leader nel settore della *Business Intelligence*, è avvenuto il 24 luglio 2015. In questa data, Microsoft ha rilasciato Power BI Desktop come applicazione *stand-alone*, slegandola definitivamente dai limiti di Excel. La nuova applicazione disponeva di una potente architettura in grado di supportare connettività aziendale avanzata, protocolli di sicurezza e una sintassi di calcolo dedicata, l'azienda ha trasformato il prodotto in una piattaforma autonoma. Da quel momento, l'ecosistema ha continuato a espandersi con aggiornamenti mensili, integrando funzionalità sempre più evolute per l'esplorazione spaziale, l'analisi predittiva e l'automazione supportata dal *Machine Learning*.

Il paradigma della Self-Service Business Intelligence

Storicamente, la *Business Intelligence* tradizionale era caratterizzata da un'architettura rigida e centralizzata, in cui ogni richiesta di esplorazione o modifica ai report doveva necessariamente transitare attraverso i dipartimenti IT. Power BI favorisce un cambio di paradigma radicale noto come *Self-Service Business Intelligence*. Questo modello facilita l'accesso all'informazione, consentendo agli analisti e agli utenti operativi di elaborare autonomamente vasti *dataset*. Senza la necessità di possedere competenze avanzate di ingegneria del software, l'utente finale è in grado di consultare basi di dati complesse, definire *KPI* su misura e costruire dashboard interattive.

La copertura del ciclo di vita della Big Data Analytics

Oltre a democratizzare l'accesso alle metriche, Power BI funge da ponte architetturale in grado di coprire quasi interamente il ciclo di vita della *Big Data Analytics* delineato nel capitolo precedente. Sebbene le fasi preliminari del ciclo analitico, in particolare la *Business Case Evaluation* e la *Data Identification*, rimangano attività di stretta competenza strategica e manageriale, Power BI si inserisce in continuità con questo processo, assorbendo interamente le successive fasi. Dalla *Data Acquisition and Extraction*, passando per la *Data Validation and Cleansing*, fino alla *Data Aggregation* e all'elaborazione matematica tramite codice DAX, la piattaforma centralizza l'intero flusso di lavoro. Infine, mediante i moduli di *Data Visualization*, lo strumento traduce queste elaborazioni in cruscotti interattivi, chiudendo il ciclo tecnico e restituendo risultati chiari per orientare le azioni correttive del management.

2.2 Il posizionamento nel mercato

Per comprendere appieno l'impatto di Power BI all'interno del mondo aziendale, è essenziale analizzarne il posizionamento competitivo. Nel settore della *Business Intelligence* e della *Data Analytics*, lo strumento di valutazione più autorevole a livello globale è il *Magic Quadrant* redatto annualmente dalla società di consulenza e ricerca Gartner.

2.2.1 Magic Quadrant

Il *Magic Quadrant* è una metodologia di ricerca proprietaria sviluppata e pubblicata da Gartner, ideata per fornire una visione globale e un'opinione autorevole sui principali fornitori di software all'interno di specifici mercati tecnologici. Nel contesto delle piattaforme di *Business Intelligence*, l'obiettivo del quadrante non è semplicemente quello di riflettere

i ricavi o la crescita finanziaria. Al contrario, esso valuta analiticamente i fornitori in base a una complessa serie di metriche ponderate che misurano le reali esigenze dei clienti, le dinamiche di mercato e la capacità di innovazione. Questo strumento aiuta i decisori aziendali a comprendere quali fornitori considerare per le proprie implementazioni.

Il *Magic Quadrant* valuta i fornitori di software basandosi su due assi fondamentali:

- *Completeness of Vision*: misura la capacità di innovazione, la strategia e la comprensione delle tendenze di mercato;
- *Ability to Execute*: valuta la solidità del prodotto, l'esperienza utente e la reattività del mercato.

I fornitori vengono conseguentemente classificati in quattro quadranti: *Niche Players*, *Visionaries*, *Challengers* e *Leaders*, come mostrato in Figura 2.1.



Figura 2.1: Il Magic Quadrant di Gartner

Nell'interpretare il grafico, Gartner precisa che il posizionamento nel quadrante dei *Leaders* non è l'unico indicatore di successo per un progetto. A seconda dei requisiti specifici dell'organizzazione, come ad esempio la necessità di un'interfaccia fortemente orientata all'esplorazione visiva o stringenti vincoli di costo, anche fornitori classificati come *Niche Players* o *Challengers* possono risultare la scelta ottimale.

Da numerosi anni consecutivi, Microsoft si posiziona stabilmente nel quadrante dei *Leaders*, collocandosi al vertice sia per completezza di visione sia per capacità di esecuzione. La sua posizione di mercato è riconducibile a diversi fattori:

- l'integrazione nativa con l'infrastruttura cloud di Azure e l'ecosistema Office 365;

- l'adozione di un modello di *Self-Service Business Intelligence* altamente scalabile;
- la continua implementazione di funzionalità avanzate basate su sistemi di Intelligenza Artificiale e sul *Machine Learning*.

In ambito manageriale, adottare Power BI significa implementare lo standard globale di mercato, assicurando all'organizzazione un vantaggio competitivo tangibile nella gestione e nell'interpretazione dei propri Big Data.

2.2.2 Il confronto con i principali competitor

Sebbene il *Magic Quadrant* di Gartner certifichi la leadership di Microsoft, un'analisi completa richiede di confrontare Power BI con i suoi principali rivali storici nel panorama della *Business Intelligence* ovvero Tableau e Qlink. Ciascuna di queste piattaforme è nata con filosofie progettuali profondamente diverse, che si riflettono ancora oggi nelle rispettive architetture.

Tableau

Tableau si è storicamente imposto nel mercato concentrandosi in modo quasi esclusivo sulla *Data Visualization* interattiva e sull'analisi visiva. Il suo punto di forza risiede nell'eccellenza estetica e nella flessibilità grafica, il che lo rende lo strumento adatto per lo storytelling analitico e per la creazione di dashboard dinamiche sviluppabili senza il supporto diretto dell'IT. L'architettura si basa sulla tecnologia proprietaria VizQL, che traduce le operazioni di *drag-and-drop* in interrogazioni ottimizzate verso il database. Integrando un motore di dati in memoria colonnare, la piattaforma garantisce prestazioni elevate anche su *dataset* complessi e multi-sorgente, abilitando operazioni agili di *Data Mashup* visuale. Poiché si focalizza sull'esplorazione visiva, Tableau viene tipicamente adottato per affiancare piattaforme aziendali preesistenti, sebbene presenti alcune limitazioni strutturali nella reportistica di produzione tradizionale e nella gestione centralizzata dei metadati.

Qlik

L'approccio di Qlik, in particolare attraverso la piattaforma QlikView, si fonda su un motore associativo interamente in memoria, progettato per consentire all'utente di esplorare i legami tra i dati in modo libero e non gerarchico, senza la necessità di predefinire percorsi rigidi di *query*. Questa architettura garantisce tempi di risposta estremamente rapidi e interfacce intuitive, il che lo rende lo strumento ideale per la prototipazione rapida e per lo sviluppo di applicazioni analitiche guidate direttamente dagli utenti finali. Il modello di mercato di Qlik segue tipicamente una strategia espansiva definita "land and expand"; questa prevede di partire da implementazioni dipartimentali per poi estendersi su scala aziendale. Nonostante l'elevata flessibilità, l'assenza di un livello semantico aziendale nativamente strutturato e la forte dipendenza dalle risorse RAM per supportare grandi quantità di utenti concorrenti rendono Qlik uno strumento che spesso viene affiancato ad altre piattaforme strutturate per coprire i requisiti di reportistica formale.

Power BI

Power BI, invece, ha conquistato la sua posizione dominante adottando un approccio sistemico e integrato. La piattaforma di Microsoft unisce eccellenti capacità visive a un motore di *Data Modeling* integrato, noto come VertiPaq. Unendo questa potenza computazionale

all'estesa infrastruttura cloud di Azure, Power BI supera il concetto di semplice creatore di grafici, proponendosi come un vero e proprio ambiente di analisi centralizzato.

La Tabella 2.1 propone un confronto tra i principali competitor, evidenziando come ciascuna piattaforma risponda a logiche progettuali e operative profondamente diverse.

Feature	**Tableau**	**Power BI**	**Qlik Sense**
Data Visualization	✓ *Best-In-Class*	! Ottima, ma meno flessibile	✓ Altamente interattiva
Facilità d'uso	! Richiede *Training*	✓ Intuitiva per utenti Microsoft	! Richiede comprensione del modello
Connettività	✓ Supporto più ampio	✓ Ottimizzata per l'ecosistema Microsoft	✓ Flessibile tramite motore associativo
Dati in *Real-Time*	✓ Sì, tramite *Live Connection*	✓ Sì, tramite *DirectQuery*	! Dipendente dai ricaricamenti in memoria
AI e automazione	! Funzionalità *AI* limitate	✓ Forti *insight* guidati dall'*AI*	✓ Supportato dal *Cognitive Engine*
Sicurezza aziendale	✓ Forti controlli basati sui ruoli	✓ Massima nel *Microsoft Ecosystem*	✓ *Governance* aziendale robusta
Costo	✗ Costoso	✓ Accessibile	! Fascia di prezzo intermedia
Ideale per	*Data Analyst*, esigenze visive complesse	Aziende Microsoft centriche, analisi con *AI*	Esplorazione dati non gerarchica, *Data Discovery*

Legenda dei simboli:
✓ *punto di forza*: pieno supporto o eccellenza di mercato;
! *attenzione*: supporto parziale o capacità intermedia;
✗ *svantaggio*: criticità o svantaggio competitivo.

Tabella 2.1: Confronto tra Tableau, Power BI e Qlik Sense

Mentre Tableau mantiene una forte specializzazione nella Data Visualization e Qlik Sense si distingue per la sua esplorazione associativa dei dati, Power BI emerge come la soluzione sistemica per le realtà aziendali. Il suo reale vantaggio competitivo risiede, infatti, nell'ottimale integrazione con l'infrastruttura Microsoft, nell'elevata accessibilità in termini di costi e nelle avanzate capacità di automazione guidate da sistemi di Intelligenza Artificiale. È proprio questa combinazione tra potenza del motore di calcolo e versatilità dell'ecosistema a giustificare l'adozione di Power BI come strumento utilizzato per la modellazione e l'analisi dei dataset demografici oggetto di questa trattazione.

2.3 L'architettura: integrazione tra Desktop, Service e Mobile

Da un punto di vista strutturale, Power BI opera come una piattaforma articolata piuttosto che come un singolo applicativo software. La soluzione proposta da Microsoft si struttura, infatti, come un'architettura ibrida e distribuita, in cui le fasi di sviluppo locale si integrano in tempo reale con le infrastrutture in cloud. L'architettura si articola in tre macro-componenti, ciascuna progettata per coprire una fase specifica del ciclo di vita del dato. Nelle sezioni successive verranno analizzati i principali strumenti che compongono l'ecosistema Power BI, con particolare riferimento a Power BI Desktop, Power BI Service e Power BI Mobile. Saranno, inoltre, esaminate alcune componenti a supporto della distribuzione e della gestione dei contenuti, tra cui il Gateway locale, Power BI Report Server e Power BI Report Builder, con particolare attenzione alla realizzazione e alla gestione dei report impaginati.

2.3.1 Power BI Desktop

Power BI Desktop rappresenta l'ambiente di sviluppo primario. È un'applicazione per sistemi operativi Windows dedicata esclusivamente all'analisi dei dati e alla creazione dei report. In questa sede, l'analista svolge le attività di connessione alle fonti, modellazione relazionale, scrittura del codice DAX e progettazione della *Data Visualization*. Un esempio dell'interfaccia di questo ambiente è illustrato in Figura 2.2.

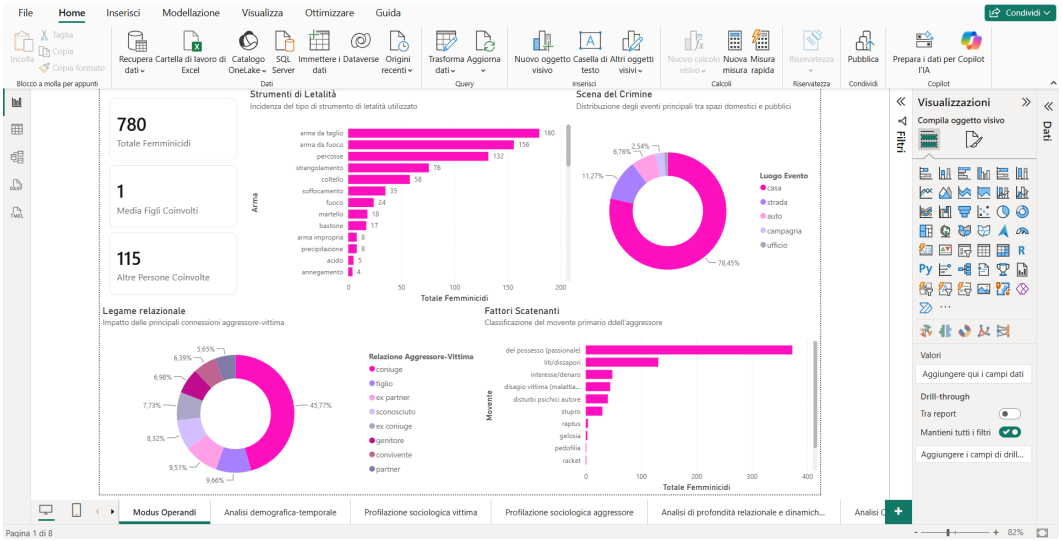


Figura 2.2: Power BI Desktop

2.3.2 Power BI Service

Power BI Service costituisce il nucleo cloud dell'architettura. Configurato come soluzione SaaS (*Software as a Service*), è la piattaforma dedicata alla pubblicazione, alla condivisione e alla collaborazione aziendale sui progetti precedentemente elaborati. Oltre a permettere la costruzione di dashboard, *Service* espande le funzionalità di base integrandosi con ecosistemi aziendali avanzati. Esso supporta l'analisi potenziata da sistemi di Intelligenza Artificiale tramite *Copilot* e la gestione centralizzata tramite Microsoft Purview. La schermata di navigazione è visibile in Figura 2.3.

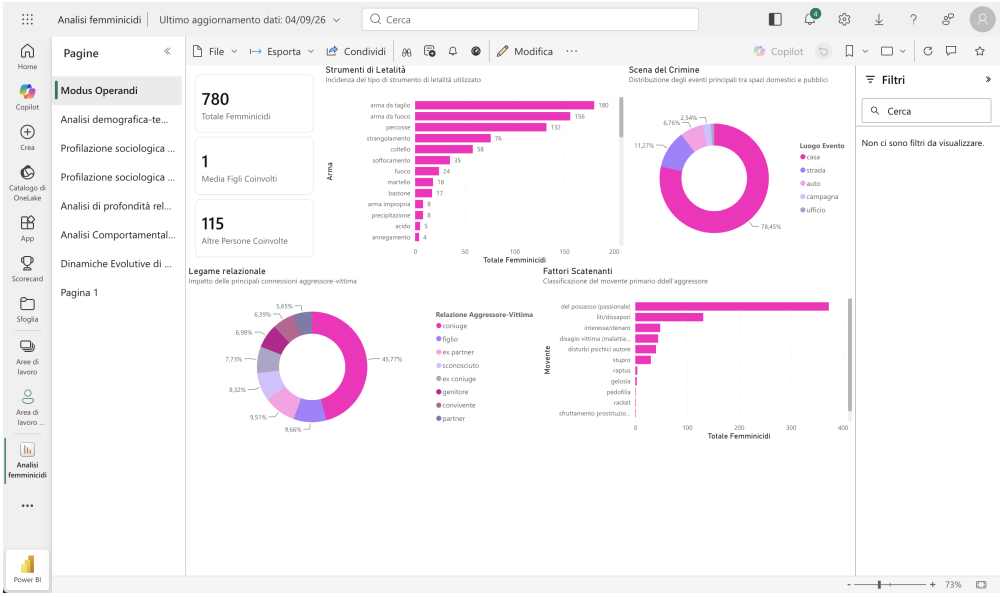


Figura 2.3: Power BI Service

2.3.3 Power BI Mobile

Power BI Mobile è l'interfaccia di fruizione finale, sviluppata sotto forma di applicazione nativa per dispositivi iOS, Android e Windows. Il suo scopo primario non è l'alterazione dei modelli analitici, bensì il consumo del dato. Permette al management di accedere in sicurezza ai propri cruscotti in mobilità, monitorando costantemente l'andamento dei *KPI* aziendali per garantire un supporto decisionale tempestivo. L'adattamento visivo di un report per gli schermi dei dispositivi mobili è presentato in Figura 2.4.

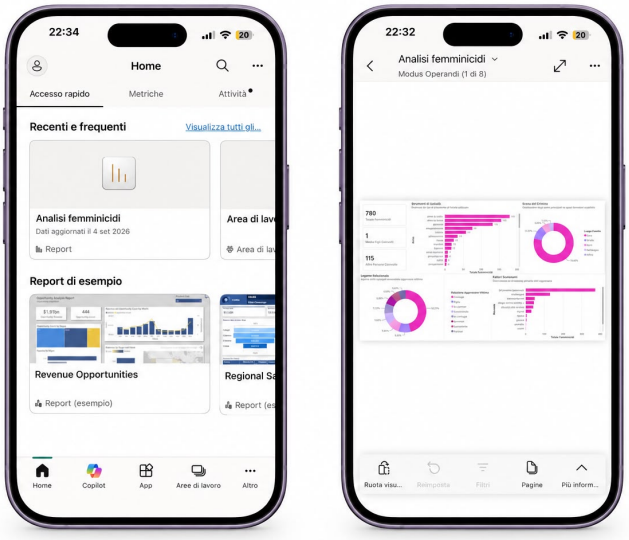


Figura 2.4: Power BI Mobile

2.3.4 Gateway dati locale

Quando le sorgenti dei dati risiedono all'interno della rete locale aziendale e non sono esposte pubblicamente su *Internet*, Microsoft prevede l'impiego dell'On-Premises Data Gateway. Questo componente agisce come un collegamento protetto da crittografia, garantendo un trasferimento dei dati rapido e sicuro tra la rete interna dell'organizzazione e i servizi cloud di Power BI. La sua implementazione consente di pianificare l'aggiornamento automatico dei *dataset* senza compromettere i rigidi protocolli di sicurezza e il firewall aziendale.

2.3.5 Report Server

Sebbene Power BI Service rappresenti lo standard per la distribuzione in cloud, alcune organizzazioni sono soggette a stringenti normative sulla residenza e sulla riservatezza dei dati che impediscono il trasferimento delle informazioni sensibili su server esterni. Per rispondere a questa limitazione, l'architettura include Power BI Report Server, un portale web locale dedicato alla pubblicazione e alla gestione locale dei cruscotti aziendali. Tale strumento replica le dinamiche di condivisione tipiche del cloud, garantendo però il controllo sull'infrastruttura. Infatti i *dataset*, le logiche di calcolo e i modelli relazionali rimangono rigorosamente isolati all'interno del perimetro di rete aziendale.

2.3.6 Report Builder e report impaginati

A complemento delle dashboard interattive sviluppate tramite la versione *Desktop*, l'ecosistema integra Power BI Report Builder, un applicativo *stand-alone* progettato specificamente per la creazione di report impaginati. A differenza dei cruscotti dinamici e interattivi, che sono orientati all'esplorazione nello schermo, i report impaginati sono strutturati per produrre documenti con impaginazione precisa, ottimizzati per l'esportazione in formati statici o per la stampa cartacea. Questo modulo risulta essenziale per la generazione di reportistica tabellare estesa, rendiconti demografici o documenti operativi che richiedono un *Layout* rigido e suddiviso su più pagine.

2.4 Il flusso di lavoro su Power BI

Una volta compresa l'architettura infrastrutturale della piattaforma, il flusso di lavoro operativo all'interno di Power BI segue rigorosamente il ciclo di vita dell'informazione. I dati grezzi, specialmente nel contesto dei Big Data, si presentano raramente in un formato pronto per l'analisi finale. Infatti, questi ultimi spesso risiedono in sistemi disconnessi, presentano valori mancanti o adottano formattazioni incoerenti. La vera potenza di uno strumento moderno di *Business Intelligence* risiede nella sua capacità di coordinare in modo fluido l'intero processo di acquisizione, strutturazione e restituzione visiva dell'informazione, garantendo l'affidabilità dell'output.

2.4.1 Il processo ETL: Extract, Transform, Load

L'ingestione e la strutturazione dei dati avvengono tramite il processo di *ETL* (*Extract, Transform, Load*), gestito interamente da Power Query. Questo motore di preparazione si occupa di orchestrare tre fasi logiche sequenziali: l'acquisizione delle fonti originarie, la loro normalizzazione e il caricamento finale nel modello dati per le elaborazioni analitiche. Il percorso logico del dato, dall'acquisizione delle fonti eterogenee fino alla sua preparazione

per i successivi livelli analitici, è rappresentato graficamente in Figura 2.5. Esso verrà illustrato in dettaglio nelle prossime sezioni.

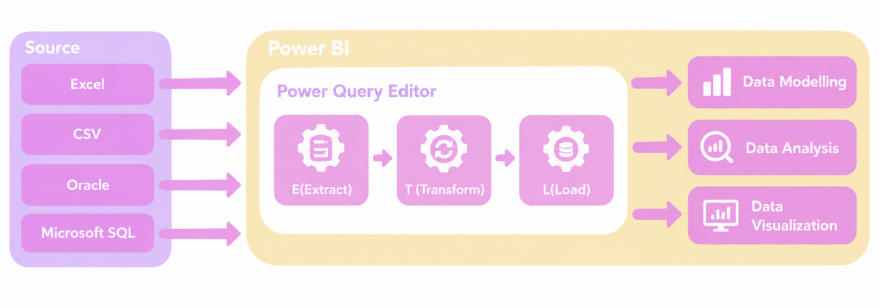


Figura 2.5: Rappresentazione architetturale del flusso ETL

Extract

Il processo di estrazione avvia l'intero ciclo analitico, stabilendo la connessione primaria tra Power BI e le fonti di dati esterne. Power Query offre un'ampia libreria di connettori nativi che permettono di stabilire connessioni sicure con un vasto ecosistema di fonti eterogenee. Queste spaziano dai semplici file locali non strutturati o semi-strutturati (come fogli di calcolo Excel, file di testo o CSV), fino ai database relazionali complessi (SQL Server, Oracle) e ai servizi cloud interrogati tramite *API* o flussi web diretti. L'obiettivo primario di questa fase è l'ingestione del dato grezzo, isolando logicamente il processo analitico dai sistemi transazionali originali.

Transform

Questa fase intermedia del processo di *ETL* svolge un ruolo fondamentale nel rendere omogenei e strutturati dati provenienti da archivi eterogenei, preparandoli alle successive fasi di analisi. Poiché le tabelle estratte dalle sorgenti primarie raramente possiedono un'impostazione nativa ottimizzata per l'analisi, l'ambiente di Power Query interviene per eseguire un rigoroso processo di *Data Cleansing*. Le routine di trasformazione permettono di risolvere le inconsistenze del *dataset*, di isolare o rimuovere i record anomali e duplicati, di trattare i valori mancanti e di allineare i tipi di dato per preservare la massima integrità referenziale. Inoltre avvengono operazioni avanzate come l'*Unpivoting* delle colonne, per riorganizzare la struttura dei *dataset*, e il *merging* di tabelle differenti, per consolidare dati frammentati. Ogni interazione grafica viene registrata in modo sequenziale nei "Passaggi applicati" e tradotta nel linguaggio funzionale *M*. Questo meccanismo assicura la tracciabilità, la reversibilità e l'automazione dell'intera *Data Pipeline*.

Load

L'ultima fase del ciclo prevede il trasferimento del *dataset* validato all'interno del motore denominato VertiPaq, che si trova nella memoria di Power BI. Durante la fase di *Load*, i dati subiscono una compressione colonnare, studiata per ottimizzare lo spazio in memoria e massimizzare la velocità di esecuzione delle interrogazioni successive. Seguendo le *best practice* dettate da Microsoft, l'analista può decidere se caricare fisicamente una tabella nel *Data Model* per renderla interrogabile tramite il linguaggio *DAX*, oppure se mantenerla esclusivamente come connessione di supporto per non gravare sul carico computazionale del report finale.

2.4.2 Data Modeling

Dopo aver importato le entità pulite, il flusso di lavoro si sposta sulla costruzione di una struttura logica che le metta in comunicazione; tale struttura prende il nome di *Data Modeling*. In questo ambiente l'analista non si limita a unire tabelle, ma definisce un ecosistema relazionale che garantisce la corretta propagazione dei filtri senza generare ambiguità. Una modellazione rigorosa risulta imprescindibile per assicurare che il calcolo di *KPI* complessi avvenga in modo rapido e matematicamente corretto.

Schema a stella

La *best practice* prevede l'adozione dello schema a stella. Questo paradigma architetturale classifica le tabelle in due macro-categorie:

- *Fact Table*: registrano gli eventi quantitativi, le misurazioni storiche e le tabelle dimensionali.
- *Dimension Table*: contengono gli attributi descrittivi, come l'anagrafica, le aree geografiche territoriali o il calendario.

Definendo le relazioni cardinali, tipicamente uno-a-molti (1:N), il modello assicura che il filtraggio applicato su una dimensione si propaghi univocamente verso la tabella dei fatti, isolando l'esatto sottoinsieme di record da analizzare. La Figura 2.6 illustra visivamente la struttura logica del modello, evidenziando la disposizione radiale delle tabelle dimensionali attorno al fulcro centrale dei dati quantitativi.

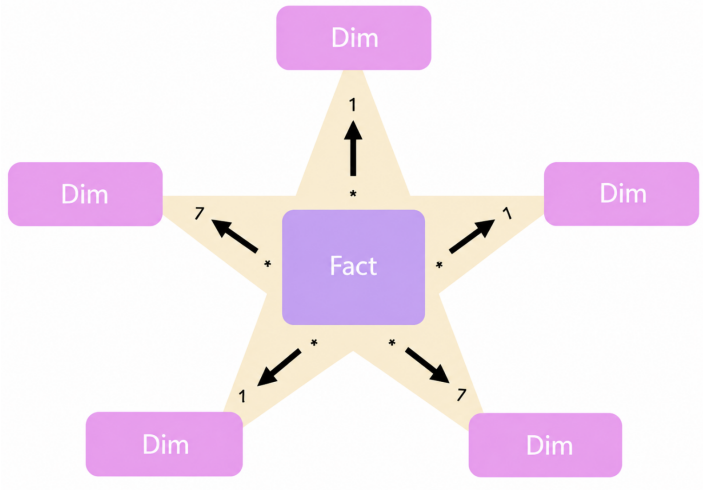


Figura 2.6: Data Modeling: lo Schema a stella

Il motore VertiPaq

L'intera infrastruttura logica si appoggia su VertiPaq, il motore di database in memoria integrato in Power BI. VertiPaq abbandona l'approccio basato su righe in favore di un'architettura di archiviazione colonnare. Durante la fase di elaborazione, il motore scansiona ciascuna colonna applicando algoritmi avanzati di codifica a dizionario e compressione del dato. Questa struttura riduce drasticamente lo spazio occupato dai *dataset* massivi e consente alla *CPU* di aggregare e filtrare milioni di record in frazioni di secondo, garantendo all'utente finale un'esplorazione analitica fluida e priva di latenza.

2.4.3 Data Visualization e reportistica

La *Data Visualization* consente di trasformare i dati del modello relazionale in rappresentazioni grafiche che ne facilitano la lettura e l'interpretazione. All'interno dell'ambiente Microsoft, questa attività si concretizza principalmente attraverso due strumenti distinti: i report e le dashboard. Il report è un documento analitico interattivo e multi-pagina, sviluppato tipicamente in Power BI Desktop, che permette un'esplorazione profonda dei fenomeni tramite matrici, grafici a dispersione o componenti cartografici avanzati, come Azure Maps. D'altra parte, la dashboard è una singola videata riepilogativa creata nell'ambiente cloud, progettata per consolidare le metriche chiave e le visualizzazioni più critiche provenienti da diversi report. Il suo obiettivo è fornire al management una sintesi esecutiva immediata e di altissimo livello.

In questo contesto, i dati numerici e le informazioni quantitative vengono tradotti in rappresentazioni grafiche in grado di far emergere *pattern*, tendenze e anomalie strutturali. Power BI eccelle in questa fase offrendo un motore di *rendering* visivo che trasforma l'informazione statica in report dinamici e dashboard interattive, ponendo l'utente al centro del processo di esplorazione dei Big Data.

L'interattività nella visualizzazione dei dati

A differenza della reportistica tradizionale o dei classici fogli di calcolo, l'ambiente visivo di Power BI si fonda sul paradigma dell'interattività continua. Ogni elemento grafico presente sulla videata non costituisce un'immagine isolata, ma è logicamente interconnesso al *dataset* sottostante. Attraverso meccanismi nativi come il *Cross-Filtering* e il *Cross-Highlighting*, la selezione di un singolo attributo ricalcola e filtra istantaneamente tutti gli altri oggetti visivi presenti nella medesima pagina. Inoltre, l'impiego di componenti quali le *KPI Card* permette di esporre in massima evidenza gli indicatori chiave di prestazione pre-calcolati, supportando le capacità di approfondimento gerarchico (*drill-down*) per passare agilmente da una visione macroscopica al dettaglio della singola osservazione.

Catalogo degli oggetti visivi di base

Oltre alle funzionalità interattive globali, l'efficacia di un report risiede nella corretta scelta delle rappresentazioni grafiche. Power BI offre una libreria nativa completa in cui ogni oggetto visivo è ottimizzato per una specifica funzione analitica, come illustrato nel pannello di selezione riportato in Figura 2.7.

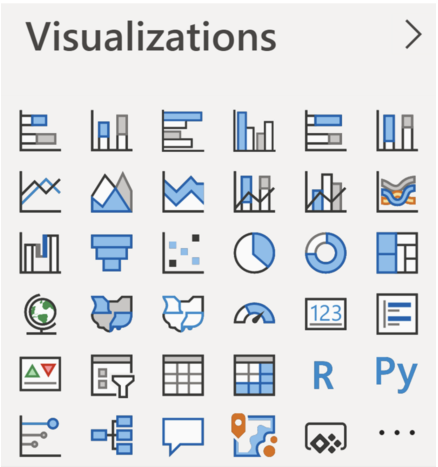


Figura 2.7: Pannello delle visualizzazioni native in Power BI

Tra le numerose opzioni messe a disposizione dalla piattaforma, le tipologie di visualizzazione più comuni per l'analisi dei dati includono:

- *Grafici a barre e a colonne:* rappresentano lo strumento per il confronto quantitativo tra categorie nominali. Sono ideali per evidenziare differenze volumetriche e generare classifiche.
- *Grafici a linee e ad area:* vengono impiegati per l'analisi delle serie storiche. Tracciando l'evoluzione di una o più metriche lungo un asse temporale continuo, permettono all'analista di individuare tendenze, ciclicità e stagionalità del fenomeno.
- *Grafici a torta e ad anello:* illustrano la composizione percentuale di un *dataset* rispetto al suo totale. Risultano efficaci quando il numero di dimensioni analizzate è ridotto, garantendo, così, un'immediata leggibilità delle proporzioni.
- *Grafici a dispersione:* mappano due variabili numeriche rispettivamente sull'asse orizzontale e verticale, consentendo di studiare la densità delle osservazioni e di rilevare eventuali correlazioni statistiche tra due indicatori.
- *Tabelle e matrici:* forniscono una visualizzazione analitica e granulare del dato. La matrice, in particolare, supporta l'esplorazione gerarchica a livelli sovrapposti, risultando essenziale per la consultazione dettagliata dei valori esatti.
- *Schede e indicatori KPI:* isolano e proiettano singoli valori di sintesi come totali complessivi, medie o scostamenti percentuali.

Il filtraggio dei dati

L'analisi interattiva dei *dataset* è supportata da diverse tipologie di filtri, che consentono di modificare dinamicamente i dati visualizzati. In base al livello di applicazione e alla modalità di interazione con il report, è possibile distinguere le seguenti categorie:

- *Filtri a livello di oggetto visivo:* agiscono esclusivamente su un singolo grafico o una singola matrice, limitando i dati elaborati per quella specifica rappresentazione senza alterare il resto della pagina.
- *Filtri a livello di pagina:* si applicano simultaneamente a tutti gli oggetti visivi presenti all'interno della medesima schermata del report, garantendo coerenza analitica d'insieme.
- *Filtri a livello di report:* rappresentano il livello gerarchico globale. Una volta impostati, condizionano l'intero documento e tutte le sue pagine, risultando ideali per restringere il perimetro assoluto dell'analisi.
- *Filtri dati (slicer):* a differenza delle tre categorie precedenti, solitamente gestite tramite un pannello a scomparsa dedicato agli sviluppatori, gli *slicer* sono veri e propri oggetti visivi collocati direttamente sulla pagina del report. Essi offrono all'utente finale un'interfaccia estremamente intuitiva per segmentare in autonomia i dati durante la consultazione.

Oltre ai filtri applicabili ai diversi livelli del report, Power BI mette a disposizione ulteriori modalità di filtraggio che consentono di definire condizioni più specifiche sui dati. Questi strumenti permettono di selezionare sottoinsiemi di dati sulla base di condizioni logiche, valori calcolati o intervalli temporali, rendendo possibile un'analisi più mirata e flessibile. Tra le principali tipologie rientrano:

- *Filtro avanzato:* consente di definire condizioni basate su operatori relazionali e logici (AND/OR), applicabili a campi testuali, numerici o temporali. È possibile, ad esempio, selezionare i valori che contengono una determinata stringa oppure quelli superiori o inferiori a una specifica soglia.
- *Filtro top N:* permette di selezionare un numero prefissato di elementi in base al valore di una determinata metrica. Può essere utilizzato per evidenziare i valori più elevati o più bassi all'interno di un insieme di dati.
- *Filtro di data relativa:* consente di filtrare dinamicamente i dati temporali sulla base di intervalli definiti rispetto alla data corrente. Questa modalità risulta particolarmente utile nei report soggetti ad aggiornamenti periodici.
- *Filtro sulle misure:* permette di applicare condizioni ai valori restituiti da misure DAX, anziché direttamente ai dati contenuti nei campi del *dataset*. In questo modo è possibile limitare la visualizzazione agli elementi che soddisfano specifiche condizioni basate su valori calcolati o soglie definite dall'utente.

Dashboard per il supporto alle decisioni

Le dashboard possono essere progettate in funzione del destinatario, del livello decisionale e delle finalità dell'analisi. In base al livello di dettaglio delle informazioni presentate e al tipo di attività supportata, è possibile distinguere tre principali categorie:

- *Dashboard strategiche:* sono destinate all'alta direzione e offrono una panoramica ad alto livello delle prestazioni generali dell'organizzazione. Sono caratterizzate da un'interfaccia tendenzialmente minimalista, focalizzata quasi esclusivamente su *KPI* sintetici e sul monitoraggio degli obiettivi a lungo termine. Consentono di esaminare i dati a diversi livelli di dettaglio per favorire una lettura istantanea dello stato di salute del *business*.
- *Dashboard analitiche:* sono progettate per i *Data Analyst* e per il management intermedio. Queste interfacce risultano estremamente interattive e ricche di profondità storica. Integrando l'intero catalogo dei filtri avanzati e le logiche di *drill-down*, consentono all'utente di indagare le relazioni complesse tra i dati e di condurre un'analisi approfondita sulle cause ultime che generano specifiche anomalie o andamenti demografici.
- *Dashboard operative:* vengono sviluppate per il monitoraggio continuo e tattico delle attività quotidiane. Si basano su dati aggiornati con altissima frequenza, spesso tramite connessioni in *real-time*, e sono utilizzate dai responsabili di reparto per controllare i processi in corso d'opera. Il loro scopo principale è evidenziare le criticità a brevissimo termine, sfruttando avvisi visivi per richiedere interventi immediati.

Ambiti di applicazione delle dashboard

Le dashboard trovano applicazione in numerosi contesti, adattandosi alle diverse esigenze di monitoraggio e analisi dei dati. Alcuni esempi di applicazione nei principali ambiti sono:

- *Sanità:* monitoraggio dell'occupazione dei posti letto per reparto, analisi dei tempi di attesa mediani nei presidi di pronto soccorso per turno e tracciamento dei tassi di somministrazione farmacologica.
- *Finanza:* confronto tra profitti e perdite rispetto al piano previsionale (*budgeting*), stato della liquidità di cassa residua e analisi delle code di revisione per il rilevamento delle frodi.

- *Vendite e CRM:* visualizzazione dello stato della pipeline commerciale, calcolo dei tassi di conversione (*Win Rate*) per segmento di clientela e misurazione dell'accuratezza delle previsioni di vendita.
- *Servizi digitali (SaaS):* tracciamento dei tassi di attivazione, analisi della *retention* per coorte e misurazione del livello di adozione di specifiche funzionalità software.
- *Logistica e pubblica amministrazione:* misurazione del tempo di ciclo per l'evasione dei permessi, monitoraggio della puntualità degli arrivi e ottimizzazione dell'indice di rotazione dell'inventario.
- *Ricerca statistica e demografica:* analisi geospaziale delle variazioni di densità abitativa, monitoraggio degli indici di invecchiamento, tracciamento dei flussi migratori e mappatura interattiva dell'incidenza di specifici fenomeni sociali o criminosi disaggregati a livello provinciale o regionale.

In tutti questi ambiti, le dashboard non si limitano alla semplice rappresentazione dei dati, ma possono supportare il processo decisionale fornendo una visione sintetica delle informazioni più rilevanti. L'integrazione di indicatori, filtri e rappresentazioni grafiche permette, infatti, di individuare più facilmente variazioni, anomalie e situazioni che richiedono attenzione, fornendo al decisore gli elementi necessari per valutare le azioni da intraprendere.

La visualizzazione dei dati geografici con Azure Maps

Quando l'analisi coinvolge *dataset* caratterizzati da una forte componente territoriale e demografica, le tradizionali rappresentazioni bidimensionali, come istogrammi o grafici a dispersione, risultano poco efficaci. Per ovviare a questo limite, la piattaforma integra Azure Maps, il sistema cartografico avanzato dell'ecosistema cloud di Microsoft. Tramite Azure Maps, l'analista ha la possibilità di aggregare milioni di coordinate geografiche utilizzando mappe di calore (*Heatmap*), di generare grafici a torta direttamente geolocalizzati sulla mappa o di analizzare flussi vettoriali in *real-time*.

La progettazione del report e la User Experience

L'efficacia analitica di una dashboard non dipende unicamente dalla mole di grafici esposti, ma anche da una rigorosa progettazione della *User Interface* e della *User Experience*. Un report professionale deve guidare il lettore attraverso uno storytelling sequenziale e ragionato. Per gestire basi di dati multi-dimensionali senza sovraccaricare cognitivamente l'utente, lo sviluppo in Power BI si avvale di architetture multi-pagina e strumenti di navigazione avanzati. L'impiego di menù a comparsa, gestiti tramite segnalibri, e di finestre di dettaglio dinamiche che appaiono al passaggio del cursore permette di mantenere le schermate pulite e prive di rumore visivo. Questa impostazione garantisce all'utente finale un ambiente di *Data Discovery* profondo, accessibile e strettamente focalizzato sull'estrazione del valore di *business*.

2.4.4 Automazione e aggiornamento dei dati

Affinché un sistema di *Business Intelligence* rimanga aggiornato nel tempo, il modello analitico non può limitarsi a rappresentare il fenomeno studiato in un determinato momento, ma deve essere aggiornato in funzione delle variazioni presenti nelle sorgenti di dati. Questo passaggio è gestito tramite i processi di *Data Refresh*. All'interno di Power BI Service, l'analista può pianificare finestre di aggiornamento automatico, permettendo alle dashboard di

riportare le informazioni più recenti. Qualora i dati originari risiedano in server locali e non direttamente in cloud, l'architettura Microsoft prevede l'implementazione di un *On-Premises Data Gateway*: un componente che consente una connessione sicura. Quest'ultimo permette al servizio web di interrogare in modo sicuro le basi di dati interne all'organizzazione, automatizzando il flusso *ETL*, con una conseguente riduzione di interventi manuali.

2.4.5 Condivisione

L'ultima fase del flusso di lavoro riguarda la distribuzione delle informazioni e dei risultati dell'analisi. Gli *insight* generati devono essere accessibili agli utenti autorizzati. Attraverso Power BI Service, i report vengono pubblicati in aree di lavoro collaborative (*workspace*). Questo stadio include la gestione dei permessi di accesso, come la Row-Level Security, che filtra dinamicamente le righe visualizzate permettendo a ciascun utente di accedere esclusivamente ai dati di propria competenza. Solo a questo punto i risultati vengono raccolti e distribuiti tramite le *App* aziendali per il consumo finale, anche tramite dispositivi mobile.

Le modalità di distribuzione

Il modello analitico sviluppato deve essere accessibile ai decisori in modo sicuro e scalabile. Il servizio cloud di Power BI offre diverse modalità di condivisione, ciascuna progettata per soddisfare specifici requisiti di collaborazione, sicurezza, numero e tipologia degli utenti destinatari. Tra le principali modalità di distribuzione dei contenuti rientrano:

- *Aree di lavoro:* rappresentano l'ambiente principale per la collaborazione. Invece di essere destinate al consumo finale, le aree di lavoro fungono da spazi di lavoro condivisi in cui i gruppi di lavoro possono creare e modificare *dataset*, report e dashboard, gestendo in modo centralizzato i ruoli e i permessi di modifica.
- *App di Power BI:* costituiscono una modalità strutturata per la distribuzione su larga scala all'interno dell'azienda. Consentono agli sviluppatori di raccogliere e distribuire molteplici report consolidati all'interno di un'unica interfaccia di navigazione, separando l'ambiente di sviluppo da quello destinato al management.
- *Condivisione diretta:* permette l'invio puntuale di un singolo cruscotto a utenti o gruppi specifici tramite *link* diretto. Questa modalità, pur essendo estremamente rapida, è solitamente impiegata per interazioni *ad hoc* e richiede l'impostazione attenta delle restrizioni di lettura per prevenire accessi non autorizzati ai dati.
- *Integrazione aziendale:* sfruttando l'integrazione con altri strumenti Microsoft, i report possono essere incorporati nativamente all'interno dei portali SharePoint aziendali o delle schede dei canali Microsoft Teams. Questo approccio permette di integrare la *Data Visualization* nei flussi di lavoro quotidiani, favorendo la collaborazione senza costringere l'utente a cambiare applicativo.
- *Pubblicazione sul web:* genera un codice *HTML* per incorporare i grafici interattivi all'interno di siti pubblici o *blog*. Poiché questa modalità non richiede l'autenticazione degli utenti, il suo utilizzo è destinato alla divulgazione di *open data*, come analisi demografiche pubbliche o statistiche governative.

2.5 I linguaggi di programmazione in Power BI

Nonostante Power BI offra un'interfaccia utente grafica intuitiva basata su interazioni *drag-and-drop*, la piattaforma supporta l'utilizzo di diversi linguaggi di programmazione,

ciascuno impiegato in specifiche fasi della gestione e dell'analisi dei dati. Questi linguaggi permettono di intervenire sulle diverse fasi del processo, dalla connessione alle fonti e dalla trasformazione dei dati fino alla loro elaborazione e analisi.

2.5.1 Il linguaggio SQL

Pur non essendo un linguaggio specifico dell'ecosistema Microsoft, lo *Structured Query Language* (SQL) svolge un ruolo importante nella connessione a database relazionali. In Power BI l'utilizzo di script SQL personalizzati consente di filtrare, aggregare e manipolare i dati direttamente nella sorgente, prima che vengano importati nel modello. In questo modo è possibile ridurre la quantità di dati trasferiti e, in determinate situazioni, migliorare le prestazioni del processo di caricamento e aggiornamento del *dataset*. L'utilizzo di query SQL può risultare particolarmente utile quando si adottano modalità di connessione come il *DirectQuery*, nelle quali le interrogazioni vengono eseguite direttamente sulla sorgente dei dati. A titolo di esempio, lo script SQL del Listato 2.1 mostra come estrarre e contare il numero di vittime per regione, considerando gli eventi registrati nel 2023.

```sql
SELECT
    Regione,
    COUNT(ID_Vittima) AS Totale_Vittime
FROM
    Dati_Femminicidi
WHERE
    Anno_Evento = 2023
GROUP BY
    Regione
ORDER BY
    Totale_Vittime DESC;
```

Listato 2.1: Query SQL: estrazione e conteggio delle vittime per regione nel 2023

2.5.2 Il linguaggio M

Il linguaggio *M*, noto anche come *Power Query Formula Language*, viene utilizzato da Power Query per definire le operazioni di estrazione e trasformazione dei dati. Le operazioni eseguite tramite l'interfaccia grafica di Power Query vengono tradotte automaticamente in istruzioni *M*, che possono essere successivamente visualizzate e modificate. L'utilizzo diretto del linguaggio permette di definire trasformazioni personalizzate e di gestire operazioni che richiedono una maggiore flessibilità rispetto a quelle disponibili attraverso la sola interfaccia grafica. A titolo di esempio, il codice in *M* del Listato **??** mostra come filtrare i dati relativi al 2023 e raggruppare le vittime per regione.

```
let
    Origine = Sql.Database("ServerIstat", "DB_Sicurezza"),
    Dati = Origine{[Schema="dbo",Item="Tabella_Femminicidi"]}[Data],
    FiltroAnno = Table.SelectRows(Dati, each [Anno_Evento] = 2023),
    Raggruppati = Table.Group(FiltroAnno, {"Regione"}, {{"
       Totale_Vittime", each Table.RowCount(_), Int64.Type}})
in
    Raggruppati
```

Listato 2.2: Script M: filtraggio e raggruppamento del numero di vittime registrate nel 2023

2.5.3 Il linguaggio DAX

Il linguaggio *DAX* (*Data Analysis Expressions*) viene utilizzato in Power BI per creare misure, colonne calcolate e altre espressioni necessarie all'analisi dei dati. A differenza di *SQL* e *M*, utilizzati principalmente per interrogare e trasformare i dati, *DAX* opera sul modello dati e permette di effettuare calcoli che tengono conto del contesto di filtro e delle relazioni tra le diverse tabelle. Le misure definite in *DAX* vengono, quindi, ricalcolate dinamicamente in base alle selezioni effettuate dall'utente nel report. A titolo di esempio, la misura DAX del Listato 2.3 calcola il numero di vittime registrate nel 2023 utilizzando la funzione `CALCULATE`.

```
Vittime_2023 =
CALCULATE(
    COUNT('Femminicidi'[ID_Vittima]),
    'Calendario'[Anno] = 2023
)
```

Listato 2.3: Misura DAX: calcolo del numero di vittime registrate nel 2023

2.5.4 Il linguaggio Python

L'integrazione di *Python* permette di estendere le funzionalità di Power BI introducendo strumenti per l'analisi dei dati e la *Data Science*. Il linguaggio può essere utilizzato sia nella preparazione dei dati, ad esempio attraverso librerie come `pandas`, sia per la creazione di rappresentazioni grafiche personalizzate. Inoltre, *Python* può essere impiegato per eseguire analisi statistiche e utilizzare modelli di *Machine Learning* sui dati disponibili. A titolo di esempio, il codice del Listato 2.4 utilizza la libreria *Pandas* per filtrare i dati relativi al 2023 e calcolare il numero di vittime per regione.

```
import pandas as pd
# Filtro per anno e conteggio delle vittime raggruppate per regione
df_2023 = dataset[dataset['Anno_Evento'] == 2023]
vittime_regione = df_2023.groupby('Regione')['ID_Vittima'].count()
print(vittime_regione)
```

Listato 2.4: Script Python: filtraggio e conteggio delle vittime per regione nel 2023

2.5.5 Il linguaggio R

Analogamente a *Python*, anche *R* può essere utilizzato in Power BI per eseguire analisi statistiche e creare rappresentazioni grafiche personalizzate. Il linguaggio è particolarmente diffuso nell'ambito dell'analisi statistica e permette di effettuare operazioni di esplorazione e modellazione dei dati. Attraverso gli script in *R* è, inoltre, possibile elaborare i dati e utilizzarli per creare *Visual* personalizzati all'interno dei report. A titolo di esempio, il codice del Listato 2.5 utilizza la libreria `dplyr` per filtrare i dati relativi al 2023 e calcolare il numero di vittime per regione.

```
library(dplyr)
# Filtraggio e conteggio delle vittime per regione
vittime_regione <- dataset %>%
  filter(Anno_Evento == 2023) %>%
  group_by(Regione) %>%
  summarise(Totale_Vittime = n())
```

Listato 2.5: Script R: filtraggio e conteggio delle vittime per regione nel 2023

CAPITOLO 3

Analisi dei dataset di partenza

Il presente capitolo descrive la struttura e le scelte metodologiche alla base del modello di Business Intelligence sviluppato per l'analisi del fenomeno del femminicidio in Italia. L'analisi di un fenomeno sociale, caratterizzato da una marcata complessità, richiede il superamento di una prospettiva narrativa in favore di un approccio data-driven. Questo metodo è orientato alla trasformazione dei dati in evidenze quantitative e all'individuazione di pattern e relazioni significative. Nelle sezioni successive vengono descritte le due principali sorgenti informative utilizzate per la costruzione dell'ambiente analitico. La prima è costituita dal dataset principale, estratto dal repository Zenodo, che raccoglie lo storico degli episodi oggetto dell'analisi. La seconda è rappresentata dalla tavola demografica ISTAT, integrata con l'obiettivo di rapportare gli eventi alla popolazione residente e consentire, quindi, la normalizzazione degli indicatori su base regionale. Particolare attenzione è dedicata alla valutazione della qualità dei dati, finalizzata a individuare le inconsistenze, le anomalie e le criticità presenti nei dati grezzi. A tale analisi segue la descrizione del processo di Extract, Transform e Load, implementato attraverso l'Editor di Power Query, mediante il quale le sorgenti informative sono state sottoposte a operazioni di pulizia, trasformazione e integrazione. Tali operazioni hanno permesso di uniformare i dati provenienti da fonti eterogenee e di organizzarli all'interno di una struttura informativa coerente, sulla quale sono state successivamente definite le misure in linguaggio DAX e realizzate le relative visualizzazioni interattive.

3.1 L'analisi data-driven del femminicidio

L'analisi dei fenomeni sociali complessi richiede un passaggio da una letteratura prevalentemente descrittiva a una valutazione dei diversi fattori che li caratterizzano. Il femminicidio non costituisce un evento criminale isolato o casuale, ma si inserisce all'interno di un quadro più ampio, definito dall'interazione di molteplici componenti. Analizzarne l'evoluzione nel tempo richiede, pertanto, strumenti in grado di valorizzare la componente quantitativa dei dati, consentendo di individuare *pattern* ricorrenti e possibili associazioni tra le diverse dimensioni del fenomeno. In questa prospettiva, la *Business Intelligence* offre un insieme di strumenti utili alla trasformazione dei dati grezzi in informazioni, funzionali all'analisi e al supporto di processi decisionali. Tale analisi prende forma dall'organizzazione di dati provenienti da fonti differenti, attraverso procedure di *Extract*, *Transform e Load* (ETL). La costruzione di un modello dati consente, quindi, di mettere in relazione le diverse caratteristiche degli eventi analizzati, rendendo possibile una lettura integrata delle dimensioni temporali, territoriali, anagrafiche e relazionali. L'elaborazione dei dati non si limita, dunque, alla produzione di statistiche descrittive, ma permette di costruire una rappresentazione multidimensionale del

fenomeno. Attraverso strumenti di analisi e visualizzazione interattiva, è possibile esplorare la distribuzione degli eventi, confrontarne l'andamento nel tempo e analizzarne la relazione con le principali caratteristiche delle vittime, degli autori del reato e dei contesti territoriali. In questo modo, l'analista può analizzare le diverse componenti del fenomeno sociale, fornendo una base empirica utile per il monitoraggio e lo sviluppo di strategie di prevenzione.

Le origini semantiche e il continuum della violenza

La comprensione di un fenomeno sociale complesso passa anche attraverso la sua definizione linguistica, poiché il linguaggio non si limita a descrivere la realtà, ma le conferisce forma, visibilità e legittimità strutturale. Il termine "femminicidio" trae origine dall'espressione anglosassone *femicide*, impiegata fino alla prima metà del Novecento per indicare in modo generico qualsiasi delitto in cui la vittima fosse di sesso femminile. La svolta semantica si deve alla sociologa e attivista Diana Russell che, con la pubblicazione del saggio *Femicide. The politics of woman killing* nel 1992, ha ridefinito la parola per descrivere un *continuum* di terrore anti-femminile. In questa accezione, il femminicidio non è un atto isolato, ma rappresenta l'estremità letale di una serie di abusi verbali, fisici e psicologici che possono culminare nella morte della donna.

L'evoluzione del concetto: da violenza domestica a violazione dei diritti

Un ulteriore contributo allo sviluppo del concetto è stato fornito dall'antropologa messicana Marcela Lagarde, che, all'inizio degli anni Duemila, contribuì ad ampliare il significato del termine *feminicidio*, inquadrando la violenza contro le donne non soltanto come esito di dinamiche individuali o relazionali, ma anche come fenomeno con una più ampia dimensione sociale e istituzionale. La sua riflessione si sviluppò anche in relazione ai numerosi casi di violenza e uccisione di donne verificatisi a Ciudad Juárez, in Messico, a partire dagli anni Novanta. Tale analisi contribuì a interpretare la violenza femminicida come una violazione dei diritti umani e a porre l'attenzione sul ruolo delle istituzioni nella prevenzione e nel contrasto del fenomeno. Questa prospettiva ha contribuito a superare una concezione più restrittiva della violenza contro le donne, a lungo associata prevalentemente alla *intimate partner violence*, ossia alla violenza esercitata all'interno di relazioni affettive o di coppia. Come evidenziato dalla sociologa Marinella Belluati, tale impostazione tendeva a concentrare l'attenzione sulle dinamiche di abuso circoscritte alla sfera privata e familiare. L'evoluzione del concetto di femminicidio ha progressivamente ampliato tale prospettiva, includendo forme di violenza che possono manifestarsi anche al di fuori della relazione di coppia e considerando il ruolo sociale e il contesto relazionale della vittima.

In Italia, il concetto si è progressivamente affermato nel dibattito pubblico e scientifico, assumendo una propria declinazione nel contesto nazionale. L'introduzione del termine nel linguaggio istituzionale si registra con l'inserimento nel dizionario Devoto-Oli del 2009, che definisce il femminicidio come una forma di violenza sistemica volta a perpetuare la subordinazione della donna all'interno di una sovrastruttura ideologica patriarcale. Oggi, il termine ha progressivamente esteso il proprio significato, superando una concezione circoscritta alla sola violenza all'interno delle relazioni intime. In questa prospettiva, il dizionario Treccani definisce il femminicidio come un' "uccisione diretta o provocata, eliminazione fisica o annientamento morale della donna e del suo ruolo sociale", evidenziando come il fenomeno non si esaurisca nella dimensione fisica, ma coinvolga anche quella morale e sociale.

Il femminicidio come fenomeno multidimensionale

Nel discorso pubblico e mediatico, il femminicidio viene frequentemente descritto e percepito come una "nuova emergenza". Tuttavia, l'analisi dei dati non conferma pienamente tale percezione: le statistiche storiche italiane, ad esempio quelle relative al periodo 2000-2012, mostrano una drammatica e costante invarianza numerica, dimostrando che non ci si trova di fronte a un picco anomalo o imprevisto, bensì a una tragica condizione di normalità. Inoltre, vi è la tendenza a circoscrivere l'evento all'ambito strettamente familiare, riconducendolo esclusivamente a dinamiche relazionali legate a fragilità individuali, gelosia o rifiuto. Questa interpretazione limita la comprensione del fenomeno e non ne considera la natura trasversale. Come teorizzato dall'antropologa messicana Marcela Lagarde, la categoria analitica del femminicidio deve includere l'insieme dei comportamenti discriminatori e violenti (fisici, psicologici, economici) tollerati a livello sociale.

Per comprendere il fenomeno nella sua complessità risulta, quindi, necessario affiancare alla lettura qualitativa un approccio quantitativo. L'estrazione, l'aggregazione e la modellazione dei dati consentono di individuare ricorrenze, distribuzioni e possibili associazioni tra le variabili, offrendo una rappresentazione più articolata del fenomeno. A partire da questa prospettiva, il lavoro prosegue con l'individuazione e la descrizione delle fonti utilizzate per la costruzione dell'analisi.

3.2 Struttura dei dataset

In questa sezione vengono descritti i due *dataset* utilizzati per la costruzione del modello dati e per la successiva analisi. Il primo rappresenta la fonte principale della ricerca e raccoglie lo storico degli eventi di femminicidio registrati sul territorio nazionale, che costituisce la base informativa per l'analisi del fenomeno. Il secondo, di natura istituzionale e demografica, fornisce i dati necessari a contestualizzare le occorrenze rilevate e a calcolare l'incidenza del fenomeno a livello regionale, consentendo di affiancare all'analisi dei valori assoluti una valutazione rapportata alla popolazione residente. Nel modello dati, il *dataset* ISTAT assume il ruolo di *Dimension Table* demografica, mentre il *dataset* principale relativo agli eventi di femminicidio costituisce la *Fact Table*. La relazione tra le due tabelle consente di affiancare alle informazioni relative ai singoli eventi il corrispondente dato sulla popolazione residente, creando la base per il calcolo di indicatori normalizzati.

3.2.1 Feminicides in Italy: counterdata RAI from 2012 to 2018

Il dataset principale utilizzato per l'analisi è stato acquisito attraverso *Zenodo*, un repository multidisciplinare ad accesso aperto dedicato alla condivisione e alla conservazione di dati e altri artefatti della ricerca scientifica. La risorsa, intitolata *Feminicides in Italy - counterdata RAI from 2012 to 2018*, raccoglie dati relativi a episodi di violenza di genere verificatisi in Italia nel periodo compreso tra il 2012 e il 2018. I dati sono stati originariamente raccolti presso la Direzione RAI Teche, la Direzione Centro Ricerche, Innovazione Tecnologica e Sperimentazione (RAI-CRITS) e il Dipartimento di Culture, Politica e Società dell'Università di Torino, per essere successivamente curati, anonimizzati e pubblicati in formato aperto.

Repository Zenodo

Il *dataset* rientra nella categoria dei *counterdata*, ossia dati raccolti e organizzati con l'obiettivo di integrare o mettere in discussione le statistiche ufficiali, soprattutto nei casi in cui queste risultino incomplete o non disponibili. Tale caratteristica risulta particolarmente rilevante nel contesto della presente analisi, poiché consente di utilizzare una base di dati

specificamente costruita per la raccolta e la sistematizzazione degli eventi considerati. Dal punto di vista della riproducibilità, la disponibilità del *dataset* in formato aperto permette di accedere alla medesima fonte utilizzata nel presente lavoro e di replicare le successive operazioni di trasformazione e modellazione dei dati. Inoltre, la risorsa è identificata dal *Digital Object Identifier* (DOI), che consente di riferirsi in modo univoco alla versione del *dataset* impiegata nell'analisi e ne garantisce la tracciabilità come risorsa scientifica. Il *dataset* è disponibile in formato CSV ed è accompagnato da un file JSON che ne descrive lo schema e i metadati.

Contenuto del dataset

Il file contiene 939 record, ciascuno corrispondente a un singolo evento, e 25 variabili descrittive, per una dimensione complessiva di circa 183 KB. La struttura dei dati è quella di una *Flat Table*, nella quale ogni riga rappresenta un evento e le colonne ne descrivono le principali caratteristiche. Per comprendere la struttura e il significato delle informazioni disponibili, sono riportati tutti i campi presenti nel *dataset* originale, accompagnati dalla relativa descrizione semantica.

- `id_event`: identificativo numerico univoco assegnato a ciascun reato;
- `victim_number`: numero progressivo che identifica la vittima all'interno dello stesso evento e consente di gestire correttamente i casi con vittime multiple;
- `anno`: riferimento temporale dell'evento;
- `ripartizione_geografica`: macro-area di aggregazione territoriale;
- `regione` e `cod_regione`: denominazione della regione in cui è avvenuto il fatto e relativo codice identificativo numerico standardizzato;
- `provincia`: livello di dettaglio geografico per la localizzazione dell'evento;
- `victim_gender`: genere della vittima;
- `victim_age_range`: classificazione dell'età della vittima raggruppata in fasce;
- `victim_nationality` e `victim_nationality_iso2`: nazionalità della vittima espressa sia in formato testuale esteso sia tramite lo standard internazionale *ISO 3166-1 alpha-2*;
- `victim_occupation`: condizione professionale della vittima;
- `victim_children`: indicatore numerico relativo alla presenza e al numero di figli della vittima;
- `attacker_victim_relationship`: natura del legame interpersonale preesistente tra l'autore del reato e la vittima;
- `attacker_age_range`: fascia d'età dell'autore del reato;
- `attacker_nationality` e `attacker_nationality_iso2`: nazionalità dell'autore del reato espressa sia in formato testuale esteso sia tramite lo standard internazionale *ISO 3166-1 alpha-2*;
- `attacker_occupation`: condizione professionale dell'autore del reato;

- `psichological_issues`: flag booleano che traccia l'eventuale presenza di pregressi disturbi di natura psichiatrica a carico dell'autore del reato;
- `suicide_attempt` e `suicide`: flag booleani che rilevano se, contestualmente all'evento, l'autore del reato ha tentato o portato a compimento il suicidio;
- `semantic_location`: tipologia di ambiente in cui si è consumato il delitto;
- `category`: classificazione giuridica del crimine;
- `weapon`: tipologia di arma o modalità utilizzata per commettere il delitto;
- `motive`: movente ipotizzato o accertato.

Questa classificazione costituisce il riferimento per le successive operazioni di trasformazione e modellazione dei dati.

3.2.2 Popolazione media per regione e ripartizione geografica, anni 1952-2014

L'analisi dei soli valori assoluti può restituire una rappresentazione distorta del fenomeno, risultando poco significativa quando si confrontano territori caratterizzati da popolazioni di dimensioni differenti. Una regione con una densità demografica elevata, infatti, può presentare un numero maggiore di eventi semplicemente in ragione della maggiore numerosità della popolazione residente, senza che ciò implichi necessariamente una maggiore incidenza del fenomeno. Per rendere i confronti territoriali più omogenei, il modello è stato integrato con una seconda fonte informativa di natura demografica, costituita dai dati sulla popolazione residente forniti dall'Istituto Nazionale di Statistica (ISTAT).

Fonte ISTAT

I dati utilizzati provengono dalla banca dati delle *Serie Storiche* di ISTAT e, nello specifico, dalle tavole relative alla popolazione residente a livello regionale. La fonte riporta i valori della popolazione residente distinti per territorio e anno, consentendo di associare a ciascuna regione il corrispondente valore demografico. L'integrazione di questa fonte con il *dataset* principale consente di rapportare il numero di eventi alla popolazione residente e di calcolare, attraverso misure *DAX*, indicatori normalizzati quali l'incidenza dei femminicidi ogni 100.000 abitanti. In questo modo, l'analisi non si limita alla distribuzione dei valori assoluti, ma permette di effettuare confronti territoriali considerando le differenze nella dimensione demografica delle regioni.

Contenuto del dataset

Il *dataset* ISTAT è organizzato in formato tabellare orizzontale (*Wide Format*) e distribuito su più fogli di calcolo. La struttura presenta una matrice bidimensionale nella quale le informazioni sono organizzate secondo tre elementi principali:

- *riferimento temporale:* la prima colonna, denominata `ANNI`, contiene gli anni di riferimento della serie storica, dal 1952 al 2014;
- *riferimento territoriale:* la prima riga contiene le intestazioni delle colonne, corrispondenti alle singole regioni italiane e alle province autonome, oltre alle eventuali aggregazioni territoriali;
- *valore della popolazione residente:* ogni cella della tabella, individuata dall'intersezione tra un determinato anno e un territorio, contiene il valore assoluto della popolazione residente corrispondente.

Selezione dei dati

La fonte ISTAT consultata presenta una serie storica della popolazione residente estesa su un arco temporale più ampio rispetto a quello necessario per l'analisi. Ai fini dell'analisi, dalla serie storica è stato estratto il dato relativo al 2014, mantenendo le informazioni riferite alle venti regioni italiane. La scelta di tale anno è stata effettuata in quanto il 2014 si colloca in posizione intermedia rispetto all'intervallo temporale considerato dal *dataset* principale, relativo agli eventi di femminicidio registrati tra il 2012 e il 2018, consentendo di utilizzare un valore demografico riferito a un anno intermedio del periodo analizzato. I dati selezionati sono stati, quindi, organizzati in una tabella di supporto contenente, per ciascuna regione, il corrispondente valore della popolazione residente. La tabella così ottenuta è stata integrata nel modello di Power BI come *Dimension Table* e collegata alla *Fact Table* degli eventi, fornendo il riferimento demografico necessario alle successive elaborazioni e al calcolo delle metriche di incidenza.

3.3 Valutazione della qualità dei dati

L'acquisizione di *dataset* di natura open source, pur garantendo accessibilità e trasparenza, può comportare la presenza di criticità strutturali e semantiche. Prima di definire l'architettura del modello relazionale, è stata condotta un'approfondita ispezione del tracciato `Dataset femminicidi in Italia.csv` al fine di misurarne la qualità dei dati. Infatti, i *dataset* reali possono presentare delle imperfezioni, delle incoerenze e lacune che, se non gestite in fase preventiva, rischiano di compromettere l'affidabilità delle dashboard finali e delle logiche decisionali che ne derivano. L'analisi della qualità dei dati è, quindi, un passaggio metodologico necessario per mappare le criticità strutturali e pianificare le successive operazioni di pulizia e trasformazione.

3.3.1 Criticità nella nomenclatura dei dati

Un primo elemento di criticità riguarda la nomenclatura delle colonne, che presenta una combinazione di termini in lingua italiana e inglese. Inoltre, essa adotta una convenzione tipica dei sistemi informatici, caratterizzata dall'utilizzo sistematico del carattere *underscore*, come nel campo `id_event`. Sebbene tale convenzione sia funzionale alla gestione dei dati in un ambiente informatico o di *database*, la presenza di denominazioni eterogenee e di termini tecnici può ridurre la leggibilità del *dataset* e rendere meno immediata l'interpretazione delle variabili. La criticità è visibile nella struttura originaria del *dataset*, come mostrato in Figura 3.1, nella quale sono riportate alcune delle intestazioni dei campi nella loro forma originale.

id_event	victim_number	anno	ripartizione_geografica	regione	cod_regione
1	1	2012	Nord-est	Emilia-Romagna	8
2	1	2012	Nord-ovest	Lombardia	3
3	1	2012	Sud	Puglia	16
4	1	2012	Sud	Puglia	16
5	1	2012	Sud	Campania	15

Figura 3.1: Esempio della nomenclatura originaria delle colonne del dataset

3.3.2 Duplicazione delle informazioni sulla nazionalità

Un'ulteriore criticità riguarda la ridondanza delle informazioni relative alla nazionalità delle vittime e degli aggressori. Nel *dataset* sono presenti, per entrambi i soggetti, sia una colonna contenente la nazionalità espressa in forma testuale, sia una corrispondente colonna contenente il relativo codice *ISO 3166-1 alpha-2*. Le colonne `attacker_nationality_iso2` e `victim_nationality_iso2` riportano in forma codificata un'informazione già presente nelle rispettive colonne di nazionalità testuale. La presenza di campi equivalenti aumenta la complessità della struttura dei dati senza introdurre informazioni aggiuntive rilevanti ai fini dell'analisi.

3.3.3 Incompletezza e valori nulli

L'ispezione della struttura originaria del *dataset* ha evidenziato la presenza di numerosi campi privi di valore, rappresentati da valori *null* o *NaN*. Questi valori risultano particolarmente frequenti nelle variabili di carattere sociodemografico e in quelle relative alle caratteristiche e alle dinamiche dell'evento. La presenza di informazioni mancanti costituisce una criticità ai fini dell'analisi, poiché l'assenza di una codifica uniforme per tali casistiche può compromettere la corretta interpretazione dei dati e la rappresentazione delle variabili nelle visualizzazioni Power BI. La presenza di valori mancanti nella struttura originaria del *dataset* è visibile nella Figura 3.2, che mostra alcuni esempi di campi a cui non sono associati valori.

attacker_age_range	attacker_nationality	attacker_occupation	psichological_issues	suicide_attempt
45-59	Italia	imbianchino	sì	sì
30-44	altro	operaio	no	no
18-29	Italia	studente	no	no
60+	Italia	disoccupato	no	no
45-59	Italia	custode	no	no
30-44	Italia	disoccupato	no	no
	altro	altro		
	altro	altro		

Figura 3.2: Esempio di dati mancanti nel dataset

3.3.4 Struttura dei fattori comportamentali e psichiatrici

Un'ulteriore criticità riguarda l'organizzazione delle informazioni relative ai fattori comportamentali e psichiatrici associati all'autore del reato. Nel *dataset* originario, tali informazioni sono distribuite orizzontalmente su tre variabili booleane distinte, `psichological_issues`, `suicide_attempt` e `suicide`, i cui valori sono espressi mediante le stringhe "sì" e "no". Questa organizzazione frammenta un'unica area tematica in più colonne e rende meno immediate l'aggregazione e la rappresentazione dei diversi fattori all'interno dell'analisi. In particolare, la presenza di variabili separate per ciascun fattore non consente di trattare direttamente tali informazioni come un'unica dimensione categoriale, rendendo più complesse le operazioni di raggruppamento e filtraggio necessarie per costruire le visualizzazioni.

3.3.5 Assenza di un identificativo univoco per la vittima

Nel *dataset* originario, l'identificativo primario `id_event` è associato al singolo evento criminoso, mentre la distinzione tra le eventuali vittime appartenenti al medesimo evento è affidata alla variabile progressiva `victim_number`. Tale struttura consente di identificare l'evento di riferimento e di distinguere le diverse vittime al suo interno, ma non fornisce un identificativo univoco che rappresenti direttamente la singola vittima. L'assenza di una chiave univoca semplice a livello di vittima rappresenta una criticità nella gestione della granularità dei dati, soprattutto nelle operazioni di aggregazione e conteggio effettuate durante l'analisi. La possibilità di identificare in modo univoco ciascuna vittima risulta utile per evitare ambiguità nel trattamento dei record e per garantire maggiore precisione nella costruzione delle misure in *DAX*.

3.3.6 Discontinuità nella copertura temporale

L'analisi della distribuzione temporale dei record ha evidenziato una discontinuità nella copertura relativa all'anno 2018. In particolare, il *dataset* presenta, per tale anno, un numero di osservazioni pari a 44, valore inferiore rispetto a quello degli anni precedenti rispetto ai valori riportati nelle statistiche disponibili relative al fenomeno del femminicidio in Italia. Ad esempio, ISTAT rileva 133 donne vittime di omicidio nel 2018, mentre l'analisi condotta dal Senato sui procedimenti giudiziari relativi al biennio 2017–2018 individua 115 potenziali femminicidi per il solo 2018. Tale differenza non consente, quindi, di interpretare il numero di record presenti nel dataset come una rappresentazione esaustiva degli eventi verificatisi nel 2018. La discontinuità osservata deve essere considerata come una possibile limitazione della copertura della fonte e rende opportuno trattare con cautela i confronti temporali che coinvolgono tale anno. La documentazione associata al *dataset* evidenzia, inoltre, i limiti legati alle modalità di raccolta e che la presenza di eventuali errori o incompletezze non possa essere esclusa.

3.4 Il processo di ETL

Le criticità individuate durante l'ispezione dei *dataset* hanno reso necessaria una fase di preparazione e trasformazione dei dati, realizzata attraverso l'editor di Power Query integrato in Power BI. Tale processo è stato strutturato secondo il paradigma di *Extract, Transform and Load*, articolato nelle fasi di estrazione, trasformazione e caricamento dei dati. L'obiettivo della procedura è stato quello di rendere le informazioni provenienti dalle diverse fonti coerenti dal punto di vista strutturale e semantico, intervenendo sulle criticità precedentemente individuate e predisponendo i dati per la successiva fase di modellazione e analisi. Le operazioni di trasformazione hanno riguardato, tra le altre, la gestione dei valori mancanti, la riorganizzazione delle variabili, l'omogeneizzazione della nomenclatura e la definizione di strutture adeguate alle successive operazioni di aggregazione e filtraggio. Il processo di ETL collega i *dataset* originari al modello dati utilizzato in Power BI. Nelle sezioni successive sono analizzate le singole fasi di *Extraction*, *Transformation* e *Load*, illustrando le principali operazioni effettuate e le relative motivazioni.

3.4.1 Extract: acquisizione e importazione delle fonti

La fase di *Extract* ha riguardato l'acquisizione dei dati dalle rispettive fonti e la loro importazione in Power Query. Il *dataset* principale, reperito tramite il repository *Zenodo*, è stato importato in formato CSV mantenendo la struttura originaria della fonte. Dopo l'importazione, Power Query ha riconosciuto preliminarmente la struttura tabellare e le

relative tipologie di dato, rendendo disponibili le informazioni per le successive operazioni di trasformazione. Per quanto riguarda il *dataset* demografico ISTAT, prima dell'importazione in Power BI sono stati individuati i dati relativi all'anno 2014 e alle venti regioni italiane, predisponendo una tabella contenente esclusivamente il territorio e il corrispondente valore della popolazione residente. Il *dataset* così predisposto è stato, quindi, importato in Power BI in una forma già coerente con l'utilizzo previsto all'interno del modello dei dati.

Le due fonti sono state, quindi, rese disponibili all'interno della *Data Pipeline*, con il *dataset* principale sottoposto alle successive operazioni di trasformazione in Power Query e il *dataset* demografico già predisposto per l'integrazione nel modello.

3.4.2 Transform: pulizia e trasformazione dei dati

La fase di trasformazione ha riguardato la pulizia, la riorganizzazione e la normalizzazione dei dati importati, con l'obiettivo di risolvere le criticità individuate durante la fase di ispezione qualitativa e predisporre le informazioni per la successiva modellazione. Le operazioni sono state realizzate all'interno di Power Query e registrate automaticamente in linguaggio *M*, consentendo di mantenere traccia sequenziale delle modifiche applicate al *dataset*. La Figura 3.3 mostra la sequenza delle operazioni di trasformazione applicate al *dataset* in Power Query.

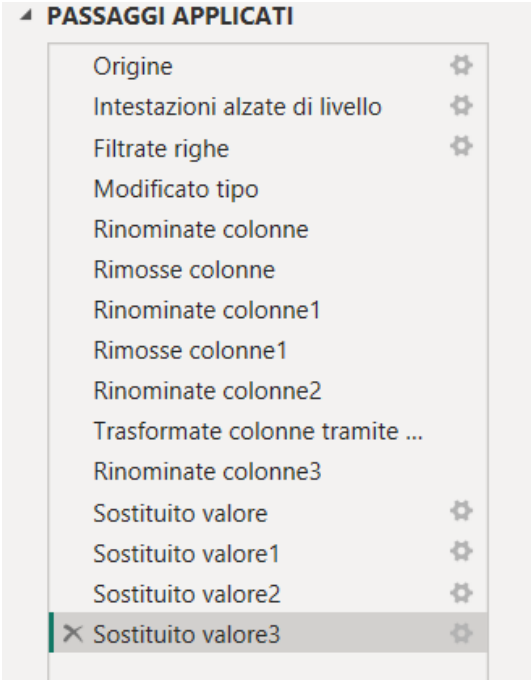


Figura 3.3: Sequenza dei passaggi di trasformazione applicati in Power Query

Traduzione e standardizzazione semantica

Una delle prime operazioni di trasformazione ha riguardato la nomenclatura delle colonne, originariamente caratterizzata dalla presenza di denominazioni in lingua inglese e da una notazione tecnica basata sull'utilizzo dell'*underscore*. Per uniformare lo schema e migliorarne la leggibilità all'interno della dashboard, le intestazioni delle variabili sono state tradotte in lingua italiana e rese maggiormente descrittive. Ad esempio, il campo `attacker_age_range` è stato rinominato in `Fascia Età Aggressore`, adottando una convenzione che facilita

l'interpretazione dei dati e delle successive visualizzazioni. La Figura 3.4 mostra un esempio della standardizzazione applicata alle intestazioni delle colonne.

ID Evento	Numero Vittime	Anno	Area Geografica
1	1	2012	Nord-est
1	1	2012	Nord-est
1	1	2012	Nord-est
2	1	2012	Nord-ovest
2	1	2012	Nord-ovest

Figura 3.4: Esempio di traduzione e standardizzazione delle intestazioni delle colonne

Esclusione dei record relativi al 2018

La seconda operazione di trasformazione ha riguardato la discontinuità temporale individuata nei dati relativi al 2018. Come evidenziato nella fase di analisi qualitativa, il *dataset* presentava per tale anno soltanto 44 record, un numero inferiore rispetto ai valori disponibili per gli anni precedenti e rispetto a quelli riportati da altre fonti per lo stesso periodo. Per evitare che tale discontinuità influenzasse la distribuzione temporale degli eventi e le successive analisi comparative tra gli anni, mediante Power Query sono stati filtrati ed esclusi tutti i record associati all'anno 2018. Il *dataset* risultante è stato, quindi, limitato agli anni 2012–2017, per i quali la copertura temporale è stata ritenuta più adeguata ai fini dell'analisi.

Rimozione delle variabili ridondanti relative alla nazionalità

La fase di trasformazione ha, inoltre, previsto la rimozione delle variabili contenenti i codici *ISO 3166-1 alpha-2* relativi alla nazionalità. In particolare, le colonne `attacker_nationali ty_iso2` e `victim_nationali ty_iso2` riportavano una codifica sintetica di informazioni già disponibili nelle corrispondenti variabili testuali relative alla nazionalità dell'aggressore e della vittima. Poiché tali attributi non aggiungevano informazioni rilevanti per le analisi previste, le due colonne sono state eliminate durante la fase di trasformazione. Questa operazione ha ridotto la ridondanza informativa e semplificato la struttura del *dataset*, mantenendo esclusivamente le variabili necessarie alla successiva fase di modellazione.

Gestione dei valori nulli

Un'ulteriore operazione di pulizia dei dati ha riguardato la gestione dei valori mancanti presenti nel *dataset*. I campi privi di informazione, identificati come valori *null*, sono stati uniformemente sostituiti con l'etichetta testuale "Dato non disponibile". Tale standardizzazione ha permesso di distinguere esplicitamente l'assenza di un'informazione da una categoria priva di valore, evitando la formazione di categorie vuote o non omogenee durante le operazioni di aggregazione e filtraggio in Power BI.

Trasformazione *Unpivot*

Le tre colonne originarie relative ai fattori comportamentali e psichiatrici, `psichological _issues`, `suicide_attempt` e `suicide`, sono state trasformate mediante un'operazione di *Unpivoting* all'interno di Power Query. La struttura originaria, nella quale ciascun fattore era rappresentato da una colonna distinta, è stata riorganizzata in una struttura verticale con

due nuove variabili: `Fattore Comportamentale`, contenente il nome del fattore considerato, ed `Esito`, contenente il relativo valore "sì" o "no". La trasformazione ha consentito di ricondurre le diverse variabili a un'unica struttura categoriale, rendendo più agevoli le operazioni di filtraggio, aggregazione e rappresentazione dei fattori all'interno del modello dei dati. La Figura 3.5 mostra il risultato della trasformazione, che riconduce i fattori originariamente distribuiti su colonne distinte alle due variabili.

Fattore Comportamentale	Esito
Suicidio Aggressore	no
Tentato Suicidio Aggressore	sì
Problemi Psicologici	sì

Figura 3.5: Risultato della trasformazione Unpivot dei fattori comportamentali

Creazione della chiave univoca della vittima

Per identificare univocamente le singole vittime anche quando più vittime sono associate allo stesso evento, è stata creata una nuova colonna personalizzata. La chiave è stata ottenuta concatenando l'identificativo dell'evento `ID Evento` con il progressivo della vittima `Numero Vittime`, separati dal carattere "-". La nuova variabile, denominata `ID Univoco Vittima`, consente, quindi, di distinguere ciascuna vittima all'interno dell'evento di appartenenza e costituisce un riferimento univoco per le successive operazioni di conteggio e aggregazione.

Mappatura dei codici territoriali

Per migliorare la leggibilità delle informazioni territoriali nelle dashboard, è stata effettuata una mappatura dei codici provinciali selezionati nel *dataset* principale. In particolare, sono stati associati alla relativa denominazione testuale i codici `001`, `015`, `048`, `058` e `063`, corrispondenti, rispettivamente, alle province di Torino, Milano, Firenze, Roma e Napoli. La trasformazione è stata applicata esclusivamente alle cinque province più colpite, per riportare nelle visualizzazioni le denominazioni estese anziché i codici numerici.

3.4.3 Load: caricamento e modellazione dei dati

La fase finale del processo di ETL, denominata *Load*, ha riguardato il caricamento nel modello dei dati di Power BI delle informazioni precedentemente estratte e trasformate. I dati elaborati sono stati caricati nel motore *in-memory* VertiPaq, che ne gestisce la memorizzazione e l'elaborazione all'interno del modello.

Modellazione relazionale

Durante il caricamento è stata definita la struttura relazionale tra le tabelle. La tabella contenente gli eventi di femminicidio assume il ruolo di *Fact Table*, mentre il *dataset* demografico ISTAT costituisce la *Dimension Table* territoriale. Le due tabelle sono state collegate attraverso il campo comune `Regione`, definendo una relazione con cardinalità molti-a-uno (*:1*), nella quale più record della *Fact Table* possono essere associati a un singolo valore della *Dimension Table*. La relazione consente di utilizzare le informazioni demografiche e territoriali per l'analisi degli eventi e permette la propagazione del contesto di filtro dalla *Dimension*

Table alla *Fact Table*. In questo modo, il modello risultante costituisce la base per la successiva definizione delle misure *DAX* e per la costruzione delle visualizzazioni e degli indicatori territoriali. La Figura 3.6 illustra la relazione tra le due tabelle.

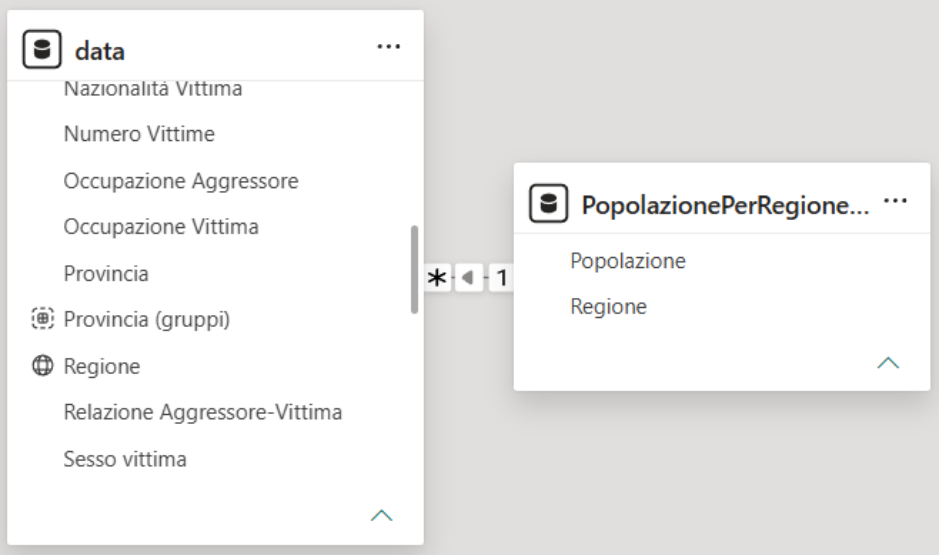


Figura 3.6: Relazione tra la Fact Table e la Dimension Table demografica

CAPITOLO 4

Progettazione e realizzazione delle dashboard di base

Il presente capitolo descrive il passaggio dal modello dei dati precedentemente strutturato alla realizzazione dell'ambiente di Business Intelligence, illustrando le principali fasi attraverso cui i dati vengono trasformati in indicatori e, successivamente, in rappresentazioni visive interattive. L'esposizione è articolata in tre parti. La prima è dedicata all'implementazione delle misure di calcolo mediante il linguaggio DAX (Data Analysis Expressions), utilizzato per definire indicatori dinamici in grado di aggiornarsi in funzione dei filtri e delle selezioni effettuate dall'utente. La seconda parte riguarda la fase di Data Visualization e presenta le quattro dashboard sviluppate, analizzandone struttura, obiettivi e principali visualizzazioni. Le dashboard seguono un percorso progressivo di esplorazione del fenomeno: dall'analisi delle modalità e delle dinamiche associate agli eventi alla loro distribuzione temporale e territoriale, fino alla caratterizzazione sociodemografica delle vittime e degli aggressori e al confronto tra le rispettive caratteristiche. La terza parte raccoglie le considerazioni conclusive sull'insieme delle dashboard e sulle principali evidenze emerse dall'esplorazione visiva. L'obiettivo non è soltanto sintetizzare i risultati restituiti dalle singole visualizzazioni, ma riflettere sulle possibilità interpretative dell'approccio adottato; distinguendo le distribuzioni e le associazioni osservabili nei dati dalle conclusioni che richiederebbero ulteriori analisi. In questo modo, il capitolo conclude il percorso di trasformazione del modello dei dati in uno strumento interattivo di analisi, mostrando come l'integrazione tra misure DAX e Data Visualization possa facilitare l'esplorazione multidimensionale del fenomeno.

4.1 Misure DAX

La transizione da un modello dei dati statico a un ambiente di *Business Intelligence* interattivo è stata resa possibile attraverso l'implementazione di misure esplicite in linguaggio *DAX* (*Data Analysis Expressions*). A differenza delle colonne calcolate, le misure vengono valutate dinamicamente al momento dell'interrogazione, sulla base del contesto di filtro definito dall'utente. Tale approccio consente di ottenere metriche in grado di aggiornarsi automaticamente in funzione delle selezioni effettuate all'interno del report e delle interazioni tra i diversi elementi visivi. L'utilizzo delle misure ha, inoltre, permesso di concentrare i calcoli analitici sulle sole interrogazioni necessarie, evitando di materializzare preventivamente risultati aggiuntivi all'interno del modello dati. In questo modo, le elaborazioni vengono eseguite dal motore *VertiPaq* in funzione del contesto di analisi, contribuendo a mantenere il modello più efficiente e flessibile.

Al fine di analizzare il fenomeno del femminicidio secondo differenti prospettive, le diciannove misure DAX implementate sono state organizzate in quattro macro-aree tematiche:

- analisi quantitativa e profilazione delle vittime;
- analisi temporale e spaziale;
- analisi relazionali e fattori criminologici;
- analisi comportamentale dell'autore del reato.

4.1.1 Analisi quantitativa e profilazione delle vittime

La prima macro-area è finalizzata a fornire una misura quantitativa della dimensione complessiva del fenomeno. Le misure considerate permettono di distinguere tra il numero degli eventi registrati e il numero effettivo delle vittime, introducendo, inoltre, una misura specifica per l'individuazione degli episodi caratterizzati dalla presenza di più vittime. Questa distinzione risulta particolarmente rilevante ai fini dell'analisi, poiché il numero degli eventi non coincide necessariamente con quello delle vittime: un singolo evento può, infatti, coinvolgere più persone. Le misure sono state, pertanto, progettate per restituire una rappresentazione coerente della fenomenologia sia in termini di episodi sia in termini di persone coinvolte. Nello specifico, la macro-area comprende le seguenti misure, ciascuna orientata a quantificare una diversa componente della dimensione complessiva del fenomeno:

- totale femminicidi;
- totale vittime complessivo;
- totale donne coinvolte;
- totale uomini coinvolti;
- totale persone trans coinvolte.

Totale femminicidi

La misura `Totale Femminicidi` determina il numero complessivo degli eventi presenti nel *dataset* attraverso un conteggio distinto dell'identificativo univoco associato a ciascun evento. L'utilizzo di `DISTINCTCOUNT` consente di evitare duplicazioni dovute alla presenza di più record riferiti allo stesso evento, garantendo che ciascun episodio venga conteggiato una sola volta. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.1.

```
1 Totale Femminicidi =
2 DISTINCTCOUNT(
3 'data'[ID Evento]
4 )
```

Listato 4.1: Misura DAX: conteggio univoco degli eventi di femminicidio

Totale vittime complessivo

La misura `Totale Vittime Complessivo` calcola il numero complessivo delle vittime attraverso il conteggio distinto della chiave univoca associata a ciascuna vittima. Tale identificativo, ottenuto durante la fase di trasformazione dei dati in Power Query, consente di distinguere le singole vittime anche quando più record appartengono al medesimo evento. La misura permette, quindi, di ottenere una quantificazione delle vittime indipendente dal numero complessivo degli eventi registrati e risulta particolarmente significativa nei casi in cui un singolo episodio coinvolga più persone. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.2.

```
Totale Vittime Complessivo =
DISTINCTCOUNT(
'data'[ID Univoco Vittima]
)
```

Listato 4.2: Misura DAX: conteggio complessivo delle vittime mediante identificativo univoco

Totale donne coinvolte

La misura `Totale Donne Coinvolte` determina il numero di vittime identificate come di sesso femminile, effettuando un conteggio distinto dell'identificativo univoco della vittima. La funzione `CALCULATE` applica al conteggio il filtro relativo al valore "femmina" presente nel campo `Sesso Vittima`. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.3.

```
Totale Donne Coinvolte =
CALCULATE(
DISTINCTCOUNT(
'data'[ID Univoco Vittima]),
'data'[Sesso Vittima] = "femmina")
```

Listato 4.3: Misura DAX: conteggio delle vittime di sesso femminile

Totale uomini coinvolti

La misura `Totale Uomini Coinvolti` segue la stessa logica della misura precedente e determina il numero di vittime identificate come di sesso maschile. Anche in questo caso, il conteggio viene effettuato sull'identificativo univoco della vittima e viene applicato un filtro specifico sul campo `Sesso Vittima`, considerando esclusivamente i record associati al valore "maschio". La relativa implementazione in linguaggio DAX è riportata nel Listato 4.4.

```
Totale Uomini Coinvolti =
CALCULATE(
DISTINCTCOUNT(
'data'[ID Univoco Vittima]),
'data'[Sesso Vittima] = "maschio")
```

Listato 4.4: Misura DAX: conteggio delle vittime di sesso maschile

Totale persone trans coinvolte

La misura `Totale Persone Trans Coinvolte` consente di quantificare le vittime per le quali il campo relativo al sesso o all'identità di genere riporta la categoria "persona trans". Anche in questo caso, il conteggio viene effettuato in maniera distinta sull'identificativo univoco della vittima, evitando che eventuali duplicazioni dei record producano una sovrastima del risultato. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.5.

```
Totale Persone Trans Coinvolte =
CALCULATE(
DISTINCTCOUNT(
'data'[ID Univoco Vittima]),
'data'[Sesso Vittima] = "persona_trans")
```

Listato 4.5: Misura DAX: conteggio delle persone trans coinvolte

4.1.2 Analisi temporale e spaziale

La seconda macro-area è finalizzata all'analisi della distribuzione temporale e territoriale del fenomeno. L'obiettivo è individuare l'andamento degli eventi nel corso degli anni, identificando i periodi caratterizzati dai valori più elevati, nonché quantificare la diffusione geografica attraverso il numero di province coinvolte. Un'ulteriore analisi è rappresentata dal calcolo dell'incidenza dei femminicidi rispetto alla popolazione residente, ottenuto sfruttando la relazione tra la tabella principale degli eventi e la *Dimension Table* contenente i dati demografici ISTAT.

Le misure appartenenti a questa macro-area consentono, pertanto, di combinare una lettura assoluta del fenomeno con una prospettiva temporale e territoriale normalizzata. In particolare, sono state implementate le seguenti metriche:

- media delle vittime per anno;
- anno caratterizzato dal maggior numero di eventi;
- numero di province coinvolte;
- incidenza dei femminicidi per 100.000 abitanti.

Media vittime per anno

La misura `Media Vittime per Anno` determina il numero medio di vittime registrate per ciascun anno considerato nel *dataset*. Il valore è ottenuto dividendo il totale dei femminicidi per il numero distinto degli anni presenti nel contesto di analisi. La misura consente, quindi, di sintetizzare l'andamento complessivo del fenomeno attraverso un indicatore medio, che viene ricalcolato dinamicamente in funzione dei filtri temporali applicati al report. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.6.

```
Media Femminicidi per Anno =
DIVIDE(
[Totale Femminicidi],
DISTINCTCOUNT(
'data'[Anno]
)
)
```

Listato 4.6: Misura DAX: calcolo della media delle vittime per anno

Anno picco storico

La misura `Anno Picco Storico` individua l'anno nel quale è stato registrato il numero più elevato di eventi. A tale scopo, la funzione `TOPN` ordina gli anni disponibili sulla base della misura `Totale Femminicidi` e seleziona il primo risultato. La funzione `MAXX` consente, quindi, di estrarre l'anno corrispondente, mentre `FORMAT` ne restituisce il valore nel formato appropriato per la visualizzazione. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.7.

```
Anno Picco Storico =
FORMAT(
MAXX(
TOPN(
1,
VALUES('data'[Anno]),
[Totale Femminicidi],
```

```
DESC
),
'data'[Anno]
),
"0"
)
```

Listato 4.7: Misura DAX: individuazione dell'anno con il maggior numero di eventi

Province coinvolte

La misura `Province Coinvolte` calcola il numero di province distinte nelle quali sono stati registrati gli eventi presenti nel *dataset*. Il conteggio viene effettuato attraverso la funzione `DISTINCTCOUNT`, che consente di considerare ciascuna provincia una sola volta anche in presenza di più eventi registrati nello stesso territorio. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.8.

```
Province Coinvolte =
DISTINCTCOUNT(
'data'[Provincia]
)
```

Listato 4.8: Misura DAX: conteggio delle province coinvolte

Incidenza per 100.000 abitanti

La misura `Incidenza per 100.000 Abitanti` consente di rapportare il numero di femminicidi alla popolazione residente, permettendo di confrontare il fenomeno indipendentemente dalla diversa consistenza demografica dei territori considerati. Il calcolo viene effettuato dividendo il numero complessivo degli eventi per la popolazione residente ottenuta dalla *Dimension Table* ISTAT e moltiplicando il risultato per 100.000. La misura restituisce, pertanto, il numero di femminicidi ogni 100.000 abitanti e costituisce un indicatore normalizzato della diffusione del fenomeno. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.9.

```
Incidenza per 100mila abitanti =
DIVIDE(
[Totale Femminicidi],
SUM(
'Tabella_ISTAT'[Popolazione]
)
) * 100000
```

Listato 4.9: Misura DAX: calcolo dell'incidenza dei femminicidi per 100.000 abitanti

4.1.3 Analisi relazionali e fattori criminologici

La terza macro-area approfondisce le caratteristiche relazionali e alcuni fattori associati alla dinamica degli eventi. L'analisi considera, in particolare, la relazione tra aggressore e vittima, il movente dichiarato e alcune modalità attraverso le quali si è verificato l'evento. La classificazione della relazione tra aggressore e vittima consente di distinguere gli episodi riconducibili alla sfera intima da quelli di natura extrafamiliare. A questa dimensione si affianca l'analisi di specifici fattori criminologici, quali il movente passionale e l'utilizzo di determinate tipologie di arma. Vengono, inoltre, considerate alcune misure relative alla

presenza di figli e agli eventi con più vittime. Le misure sviluppate per questa macro-area comprendono:

- casi riconducibili alla sfera intima;
- casi extrafamiliari;
- casi caratterizzati da un movente passionale;
- casi ad alta letalità;
- percentuale di femminicidi riconducibili alla sfera intima;
- media dei figli coinvolti;
- eventi multipli.

Casi sfera intima

La misura `Casi Sfera Intima` individua gli eventi nei quali la relazione tra aggressore e vittima rientra in una delle categorie riconducibili alla sfera intima o familiare. Il filtro viene applicato attraverso la funzione `CALCULATE`, che limita il conteggio della misura `Totale Femminicidi` alle relazioni specificate mediante l'operatore `IN`. In questo modo è possibile quantificare gli eventi associati a relazioni quali "coniuge", "partner", "ex partner", "amante", "convivente", o altri rapporti familiari. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.10.

```
Casi Sfera Intima =
CALCULATE(
[Totale Femminicidi],
'data'[Relazione Aggressore-Vittima]
IN {
"coniuge",
"partner",
"ex partner",
"ex coniuge",
"amante",
"ex amante",
"genitore",
"figlio",
"convivente"
}
)
```

Listato 4.10: Misura DAX: conteggio dei casi riconducibili alla sfera intima

Casi extrafamiliari

La misura `Casi Extrafamiliari` individua gli eventi nei quali la relazione tra aggressore e vittima non appartiene alle categorie definite per la sfera intima. La funzione `NOT` applicata all'operatore `IN` consente di escludere dal conteggio le relazioni precedentemente individuate, isolando, quindi, le restanti categorie presenti nel *dataset*. La misura permette, così, di analizzare separatamente gli eventi caratterizzati da una relazione extrafamiliare tra aggressore e vittima. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.11.

```
Casi Extrafamiliari =
CALCULATE(
[Totale Femminicidi],
NOT(
'data'[Relazione Aggressore-Vittima]
IN {
"coniuge",
"partner",
"ex partner",
"ex coniuge",
"amante",
"ex amante",
"genitore",
"figlio",
"convivente"
}
)
)
```

Listato 4.11: Misura DAX: conteggio dei casi con relazione extrafamiliare

Casi con movente passionale

La misura `Casi con Movente Passionale` quantifica gli eventi nei quali il campo relativo al movente riporta la categoria del "possesso (passionale)". Il filtro viene applicato alla misura `Totale Femminicidi` mediante la funzione `CALCULATE`, consentendo di isolare esclusivamente gli eventi associati alla specifica classificazione del movente. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.12.

```
Casi con Movente Passionale =
CALCULATE(
[Totale Femminicidi],
'data'[Movente] =
"del possesso (passionale)"
)
```

Listato 4.12: Misura DAX: conteggio dei casi con movente passionale

Casi ad alta letalità

La misura `Casi ad Alta Letalità` identifica gli eventi nei quali l'arma utilizzata appartiene alle categorie considerate nella definizione adottata per l'analisi, ossia "arma da fuoco", "arma da taglio" o "coltello". Il filtro viene applicato attraverso l'operatore `IN`, che consente di verificare l'appartenenza del valore a un insieme predefinito di categorie. La misura permette, quindi, di isolare una specifica configurazione degli eventi sulla base della modalità di aggressione registrata nel *dataset*. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.13.

```
Casi Alta Letalita =
CALCULATE(
[Totale Femminicidi],
'data'[Arma]
IN {
"arma da fuoco",
"arma da taglio",
"coltello"
```

```
}
)
```

Listato 4.13: Misura DAX: conteggio dei casi classificati ad alta letalità

Percentuale di femminicidi nella sfera intima

La misura `Percentuale Femminicidi Sfera Intima` esprime, in termini percentuali, la quota degli eventi riconducibili alla sfera intima rispetto al totale degli eventi registrati. Il valore viene calcolato mediante la funzione `DIVIDE`, che rapporta il numero dei casi appartenenti alla sfera intima al totale dei femminicidi. L'utilizzo del terzo argomento di `DIVIDE` consente, inoltre, di restituire il valore zero qualora il denominatore non sia disponibile o risulti nullo. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.14.

```
Percentuale Femminicidi Sfera Intima =
DIVIDE(
[Casi Sfera Intima],
[Totale Femminicidi],
0
)
```

Listato 4.14: Misura DAX: percentuale di femminicidi riconducibili alla sfera intima

Media figli coinvolti

La misura `Media Figli Coinvolti` calcola il numero medio di figli associato alle vittime presenti nel *dataset*. Il valore viene ottenuto attraverso la funzione `AVERAGE`, applicata al campo "Figli Vittima", e consente di sintetizzare in un unico indicatore la presenza media di figli tra le vittime considerate. Poiché si tratta di una misura, il risultato viene ricalcolato in funzione del contesto di filtro applicato al report. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.15.

```
Media Figli Coinvolti =
AVERAGE(
'data'[Figli Vittima]
)
```

Listato 4.15: Misura DAX: calcolo del numero medio di figli delle vittime

Eventi multipli

La misura `Eventi Multipli` consente di individuare gli eventi caratterizzati dalla presenza di più vittime. Per ciascun identificativo di evento vengono considerate le relative occorrenze e viene verificato se il numero di vittime distinte associate supera l'unità. La funzione `VALUES` genera l'insieme degli eventi distinti, mentre `FILTER` seleziona esclusivamente quelli per i quali il conteggio distinto delle vittime, effettuato mediante `DISTINCTCOUNT`, risulta superiore a uno. In questo modo, la misura restituisce il numero complessivo di eventi nei quali sono state registrate almeno due vittime. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.16.

```
Eventi Multipli =
COUNTROWS(
FILTER(
VALUES('data'[ID Evento]),
```

```
CALCULATE(
DISTINCTCOUNT(
'data'[ID Univoco Vittima]
)
) > 1
)
)
```

Listato 4.16: Misura DAX: individuazione degli eventi con più di una vittima

4.1.4 Analisi comportamentale dell'autore del reato

L'ultima macro-area approfondisce alcuni fattori comportamentali associati agli aggressori e le conseguenze successive all'evento. L'analisi si concentra, in particolare, sulla presenza di problematiche psicologiche e sui casi nei quali l'aggressore abbia commesso un suicidio successivamente al delitto. La disponibilità di tali informazioni è stata resa possibile dalla precedente trasformazione dei dati mediante operazioni di *Unpivot*, che hanno consentito di ricondurre specifici fattori comportamentali a una struttura analizzabile attraverso filtri e condizioni logiche. Le misure implementate sono:

- presenza di disturbi o problemi psicologici;
- numero di suicidi degli aggressori;
- percentuale di suicidi successivi al delitto.

Disturbi psicologici

La misura `Disturbi Psicologici` quantifica gli eventi nei quali è stata registrata la presenza di problemi psicologici associati all'aggressore. Il conteggio viene effettuato sugli identificativi distinti degli eventi, applicando contemporaneamente due condizioni: il campo `Fattore Comportamentale` deve assumere il valore "Problemi Psicologici" e il campo `Esito` deve riportare il valore "si". L'utilizzo di `DISTINCTCOUNT` permette di evitare che più informazioni riferite allo stesso evento determinino duplicazioni nel conteggio. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.17.

```
Disturbi Psicologici =
CALCULATE(
DISTINCTCOUNT(
'data'[ID Evento]
),
'data'[Fattore Comportamentale] =
"Problemi Psicologici",
'data'[Esito] = "si"
)
```

Listato 4.17: Misura DAX: conteggio degli eventi associati a problemi psicologici

Numero suicidi

La misura `Numero Suicidi` determina il numero di eventi nei quali è stato registrato il suicidio dell'aggressore. Anche in questo caso, il conteggio viene effettuato sugli identificativi distinti degli eventi e viene subordinato alla presenza della categoria "suicidio" nel campo "Fattore Comportamentale" e del valore "si" nel campo "Esito". La misura consente, pertanto, di isolare gli eventi nei quali il comportamento suicidario dell'aggressore costituisce

l'esito registrato nel *dataset*. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.18.

```
Numero Suicidi =
CALCULATE(
DISTINCTCOUNT(
'data'[ID Evento]
),
'data'[Fattore Comportamentale] =
"suicidio",
'data'[Esito] = "si"
)
```

Listato 4.18: Misura DAX: conteggio degli eventi con suicidio dell'aggressore

Percentuale suicidi post-delitto

La misura `Percentuale Suicidi Post-Delitto` esprime la quota di eventi nei quali è stato registrato il suicidio dell'aggressore rispetto al totale dei femminicidi considerati. Il valore viene calcolato attraverso il rapporto tra il numero di suicidi e il totale degli eventi, utilizzando la funzione `DIVIDE` per gestire in maniera sicura eventuali denominatori nulli. La misura permette, dunque, di trasformare il valore assoluto dei suicidi in un indicatore percentuale, facilitandone l'interpretazione all'interno del report. La relativa implementazione in linguaggio DAX è riportata nel Listato 4.19.

```
Percentuale Suicidi Post-Delitto =
DIVIDE(
[Numero Suicidi],
[Totale Femminicidi],
0
)
```

Listato 4.19: Misura DAX: percentuale di suicidi dell'aggressore successivi al delitto

4.2 Analisi delle dashboard di base

Dopo l'implementazione e la validazione delle misure DAX, ci si è concentrati sulla *Data Visualization*.

4.2.1 Definizione e scopo delle dashboard di base

Nel contesto della *Business Intelligence*, una dashboard di base costituisce il livello iniziale dell' *Exploratory Data Analysis*. Essa trasforma l'informazione elementare contenuta nelle tabelle in una rappresentazione sintetica e interrogabile, attraverso grafici e filtri interattivi. L'obiettivo di tali strumenti risiede nel costruire un percorso di lettura progressivo del fenomeno, nel quale ogni dashboard risponde a una specifica domanda analitica. Perciò, in questa fase non vengono introdotti modelli predittivi o algoritmi di classificazione; l'obiettivo è quello di individuare distribuzioni, frequenze, *pattern* e relazioni che emergono direttamente dai dati storici.

4.2.2 Struttura dell'analisi visiva

La struttura dell'analisi visiva è stata articolata nelle seguenti quattro dashboard:

- *Dashboard 1: "Modus operandi e dinamiche del delitto"*: analizza gli strumenti di letalità, la scena dell'evento, il legame relazionale tra vittima e aggressore e i principali fattori scatenanti presenti nel *dataset*.
- *Dashboard 2: "Analisi demografico-temporale"*: esamina la distribuzione geografica e temporale dei casi mediante una mappa regionale, filtri annuali e territoriali e una graduatoria delle province maggiormente rappresentate.
- *Dashboard 3: "Profilazione sociologica della vittima"*: descrive le persone coinvolte come vittime in termini di genere, età, condizione occupazionale, numero di figli e nazionalità.
- *Dashboard 4: "Profilazione sociologica dell'autore del reato"*: analizza il profilo sociologico dell'autore del reato secondo le medesime dimensioni considerate per la vittima, integrandole con informazioni relative alla presenza di disturbi psicologici e al suicidio dell'autore. Essa comprende, inoltre, un'analisi incrociata delle fasce d'età della vittima e dell'autore.

4.2.3 Dashboard 1: "Modus operandi e dinamiche del delitto"

La prima dashboard costituisce il punto di partenza dell'analisi, in quanto si concentra sulle principali caratteristiche e dinamiche associate agli episodi considerati. Essa ha l'obiettivo di individuare le modalità con cui gli eventi si manifestano e le principali circostanze che li caratterizzano. In particolare, l'analisi si concentra sugli strumenti di letalità impiegati, sui luoghi in cui si verificano gli eventi, sulla tipologia di legame relazionale tra vittima e autore e sui fattori individuati come movente principale. In questo modo è possibile evidenziare i *pattern* più ricorrenti all'interno del *dataset* e delineare le principali caratteristiche degli episodi analizzati, come illustrato nella Figura 4.1.

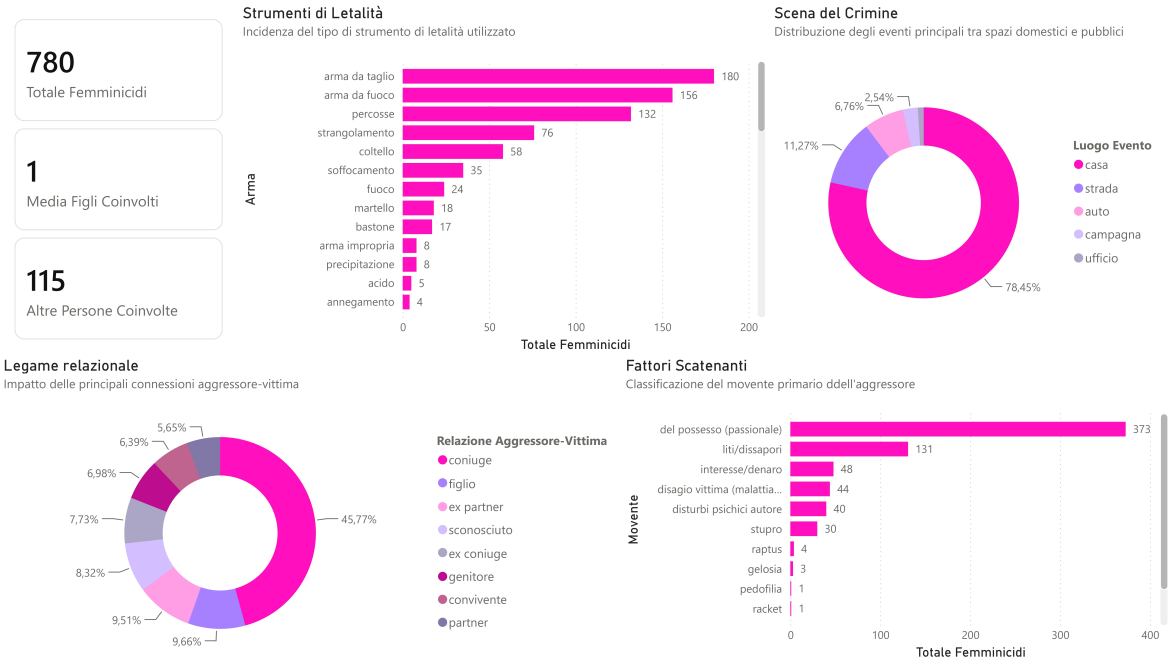


Figura 4.1: Dashboard 1: "Modus operandi e dinamiche del delitto"

I principali elementi di visualizzazione presenti nella dashboard, che verranno analizzati singolarmente nelle sezioni successive, sono:

- *indicatori KPI*;

- *grafici a barre orizzontali*;
- *grafici ad anello*.

Indicatori KPI

Nella parte sinistra della dashboard sono presenti tre *indicatori KPI*, che riportano alcune informazioni di sintesi relative ai dati analizzati e offrono un primo quadro generale del fenomeno prima di passare all'analisi dei grafici. Inoltre, l'applicazione di filtri temporali o di altri selettori modifica il numero di eventi considerati e, di conseguenza, il contenuto degli altri grafici. I tre indicatori considerati sono:

- *Totale Femminicidi (780)*: mostra il numero complessivo di eventi presenti nella vista corrente e fornisce il riferimento quantitativo per l'interpretazione delle altre visualizzazioni.
- *Media Figli Coinvolti (1)*: esprime il numero medio di figli associati agli eventi presenti nella vista, fornendo un'indicazione dell'impatto del fenomeno sul contesto familiare della vittima.
- *Altre Persone Coinvolte (115)*: indica il numero di ulteriori persone coinvolte negli eventi, secondo la classificazione presente nel *dataset*, evidenziando l'impatto del fenomeno oltre alla vittima principale.

Grafico a barre orizzontali: "Strumenti di Letalità"

Nella parte centrale superiore della dashboard è presente un grafico a barre orizzontali intitolato *Strumenti di Letalità*. La variabile riportata sull'asse verticale identifica lo strumento o la modalità di letalità, mentre l'asse orizzontale mostra il numero complessivo di femminicidi associato a ciascuna categoria. Il grafico permette, dunque, di confrontare la frequenza delle diverse modalità e di individuare quelle maggiormente ricorrenti. Le prime tre barre risultano più lunghe rispetto alle successive, evidenziando una maggiore frequenza di alcune modalità rispetto alle categorie meno rappresentate. In particolare, la categoria "arma da taglio" raggiunge 180 casi e risulta la più frequente, seguita da "arma da fuoco" e "percosse". Il grafico mostra, inoltre, la varietà degli strumenti e delle modalità presenti nel *dataset*, evidenziando come gli eventi analizzati non siano riconducibili a un'unica modalità di aggressione.

Grafico ad anello: "Scena del Crimine"

Nel quadrante superiore destro è collocato il grafico ad anello *Scena del Crimine*, che mostra la distribuzione percentuale degli eventi in funzione del luogo in cui sono stati registrati. In questo caso, l'obiettivo è evidenziare il peso relativo dei diversi contesti ambientali. Per rendere più leggibile la visualizzazione, il grafico è stato filtrato mantenendo i cinque risultati più rilevanti, corrispondenti alle categorie di luogo con il maggior numero di casi. Il filtro è stato applicato direttamente sulla visualizzazione, senza modificare il *dataset*. Tra le categorie selezionate emerge la voce "casa", che rappresenta il 78,45% degli eventi, mentre le restanti categorie presentano delle percentuali nettamente inferiori. Tale differenza, tra la prima modalità e le successive, rende evidente il peso del contesto domestico all'interno della casistica in analisi. Dal punto di vista metodologico, il grafico rappresenta percentuali sul totale dei record classificati per luogo e non costituisce una misura di rischio in senso stretto.

Grafico ad anello: "Legame Relazionale"

Nella parte inferiore sinistra la dashboard utilizza un secondo grafico ad anello per rappresentare il tipo di legame tra vittima e autore del reato. Questa visualizzazione permette di passare dalla dimensione materiale dell'evento a quella relazionale, mostrando quanto la distribuzione dei casi sia associata a specifiche tipologie di rapporto. Per rendere più leggibile la visualizzazione, anche per questa visualizzazione è stato applicato un filtro sulle categorie più rilevanti, mantenendo gli otto risultati con il maggior numero di casi. L'analisi dei dati restituiti dal grafico evidenzia la natura intrafamiliare e affettiva del fenomeno. La categoria "coniuge" rappresenta, da sola, la maggioranza relativa, attestandosi al 45,77% dei casi registrati. Se a questa categoria si sommano le altre figure appartenenti alla sfera intima, domestica e parentale - quali "figlio" (9,66%), "ex partner" (9,51%), "ex coniuge" (7,73%), "genitore" (6,98%), "convivente" (6,39%) e "partner" (5,65%) — emerge come la maggioranza degli eventi si consumi all'interno di relazioni familiari. Si può notare come tale grafico possa essere messo in relazione con il legame relazionale tra vittima e autore; la prevalenza del contesto familiare può essere osservata insieme al luogo della scena del crimine più frequentemente registrato. Tale aspetto verrà approfondito nel capitolo successivo attraverso l'analisi congiunta delle due dimensioni.

Grafico a barre: Fattori Scatenanti

Il quadrante inferiore destro contiene il grafico a barre *Fattori Scatenanti*, che ordina i moventi o i fattori scatenanti primari associati agli eventi. In questo caso l'asse verticale riporta le diverse categorie di movente, mentre l'asse orizzontale esprime il numero complessivo di casi. Il grafico deve essere interpretato come classificazione del movente primario registrato nel *dataset*: esso non equivale a una spiegazione esaustiva del comportamento dell'autore del reato e, dunque, non esclude la possibile compresenza di ulteriori fattori non rappresentati nella categoria primaria. La categoria del "possesso (passionale)", seguita da "liti/dissapori" e "interesse/denaro" presenta la barra più lunga e si distingue dalle successive. La categoria del "possesso (passionale)" rappresenta la barra più estesa in assoluto, registrando 373 casi e distaccandosi in modo netto dalle motivazioni successive. Questo risultato mette in luce un elemento intrinseco alla dinamica del fenomeno: l'ideologia del possesso e l'oggettificazione della vittime risultano quella maggiormente rappresentate nel *dataset*. Tale dato ridimensiona anche la narrazione del delitto generato da una temporanea perdita di controllo, dal momento che la categoria indicata come "raptus" risulta marginale, essendo associata a soli 4 eventi nell'intero *dataset*.

4.2.4 Dashboard 2: "Analisi demografico-temporale"

La seconda dashboard sposta l'attenzione dalla struttura interna dell'evento alla sua distribuzione nello spazio e nel tempo. L'obiettivo è quello di individuare come i casi si distribuiscano sul territorio e come il quadro complessivo possa cambiare restringendo il periodo o l'area geografica di interesse, come mostrato nella Figura 4.2.

I principali elementi di visualizzazione presenti nella dashboard, che verranno analizzati singolarmente nelle sezioni successive, sono:

- *indicatori KPI*;
- *mappa di Azure*;
- *filtro dei dati*;
- *grafico a barre orizzontali*.

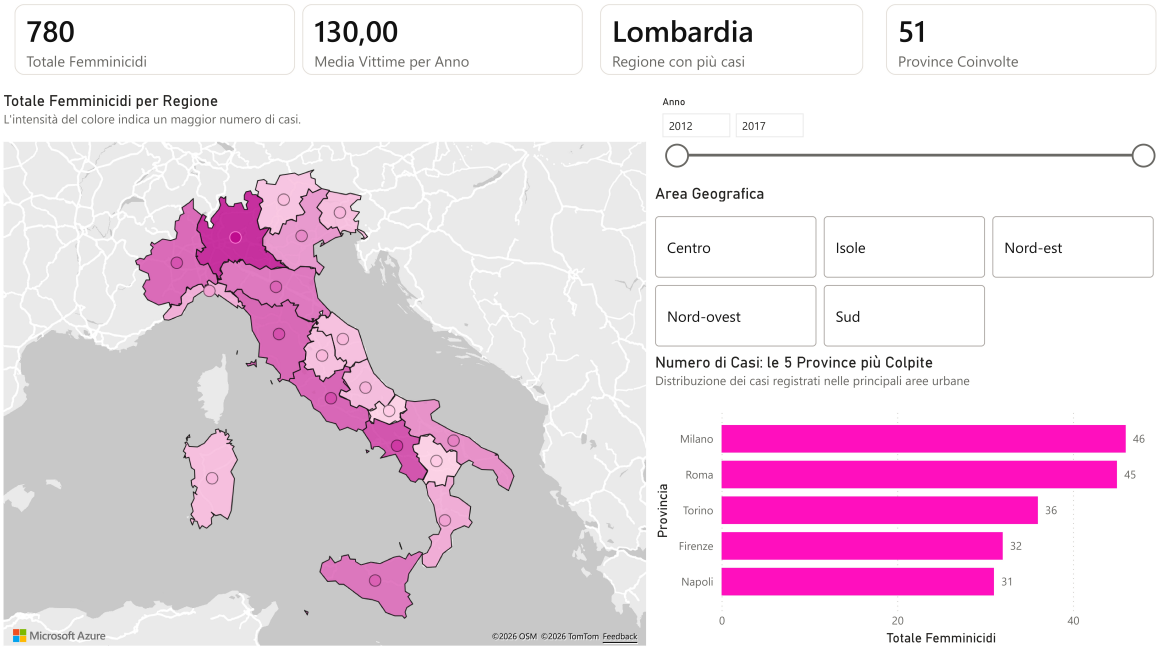


Figura 4.2: Dashboard 2: "Analisi demografico-temporale"

Indicatori KPI

Nella parte superiore della dashboard sono presenti tre *indicatori KPI*, che riportano alcune informazioni di sintesi relative ai dati analizzati e offrono un primo quadro generale del fenomeno prima di passare all'analisi dei grafici. Inoltre, l'applicazione di filtri temporali o di altri selettori modifica il numero di eventi considerati e, di conseguenza, il contenuto degli altri grafici. I tre indicatori KPI considerati sono:

- *Totale Femminicidi (780)*: mostra il numero complessivo di eventi presenti nella vista corrente e fornisce il riferimento quantitativo per l'interpretazione delle altre visualizzazioni.
- *Media Vittime per Anno (130,00)*: indica il numero medio di vittime per anno nel periodo visualizzato.
- *Regione con più casi (Lombardia)*: identifica la regione che presenta il maggior numero assoluto di casi nella vista corrente. Per ottenere questo KPI sono state considerate le diverse regioni presenti nel dataset e, successivamente, è stato applicato un filtro per selezionare la regione con il valore massimo del conteggio.
- *Province Coinvolte (51)*: mostra il numero di province presenti nel *dataset* analizzato.

Filtro dei dati

Nella porzione superiore destra dell'interfaccia sono collocati due strumenti interattivi utili per la segmentazione spaziotemporale del *dataset*. Il primo controllo è uno *slider* temporale continuo, il quale consente di circoscrivere l'indagine a un orizzonte temporale specifico impostando liberamente un anno di inizio e uno di fine. Al di sotto di quest'ultimo è implementato un selettore a pulsanti dedicato alla dimensione "Area Geografica", che divide il territorio nazionale in cinque macro-regioni: *Centro*, *Isole*, *Nord-est*, *Nord-ovest* e *Sud*. Essi non modificano un singolo grafico, ma il contesto dell'intera dashboard, permettendo di

osservare come cambiano KPI, mappa e graduatoria provinciale al variare del periodo. Di conseguenza, questi strumenti consentono di passare da una fotografia complessiva a una lettura condizionata da una dimensione precisa.

Mappa di Azure: "Totale Femminicidi per Regione"

La parte sinistra della dashboard è occupata dalla mappa di Azure *Totale Femminicidi per Regione*. Ogni regione viene rappresentata mediante una gradazione cromatica: a tonalità più intense corrispondono valori assoluti più elevati. La mappa mostra, dunque, la distribuzione spaziale del fenomeno sull'intero territorio nazionale. La principale informazione ottenibile è la presenza di una distribuzione non uniforme. Alcune regioni mostrano una colorazione più intensa e, quindi, una maggiore concentrazione di casi registrati, mentre altre presentano tonalità più chiare. La rappresentazione suggerisce una maggiore presenza di casi nelle regioni settentrionali e in alcune regioni del Centro, mentre le aree insulari e diverse regioni meridionali risultano complessivamente meno intense nella scala cromatica utilizzata. Il titolo e la legenda si riferiscono al *totale* dei casi e a valori assoluti. Non si tratta di un tasso normalizzato per popolazione residente; di conseguenza, una regione con più abitanti può presentare un numero maggiore di casi anche senza avere necessariamente un'incidenza pro-capite più elevata. Per questo motivo, la lettura della mappa deve essere considerata complementare, e non sostitutiva, rispetto alle successive analisi basate sull'incidenza. In questo modo, la dashboard permette, innanzitutto, di descrivere dove il fenomeno risulta maggiormente rappresentato nel *dataset*, senza attribuire ai valori osservati un significato diretto di rischio individuale. La dashboard consente un'analisi della distribuzione territoriale assoluta, lasciando ad analisi successive il confronto con denominatori demografici.

Grafico a barre orizzontali: "Numero di Casi: le 5 Province più Colpite

La sezione inferiore destra utilizza un grafico a barre orizzontali per rappresentare le cinque province con il maggior numero di casi registrati nella vista corrente. Per costruire questa visualizzazione è stato utilizzato il gruppo creato nella fase di trasformazione mediante il quale le province sono state organizzate in una selezione specifica contenente le cinque unità territoriali maggiormente rappresentate. Mentre la mappa di Azure consente, infatti, di individuare pattern spaziali generali, questo grafico permette di confrontare direttamente le province che occupano le prime posizioni. La presenza di Milano e Roma nella parte alta della graduatoria evidenzia, inoltre, il peso delle grandi aree urbane sul numero assoluto di casi registrati. Anche in questo caso il grafico visualizza il totale dei casi e non un tasso rapportato alla popolazione provinciale. La presenza di grandi centri urbani tra le prime posizioni può essere letta anche considerando la diversa dimensione demografica dei territori, che influenza necessariamente il numero assoluto di eventi registrati. Pertanto, il risultato evidenzia una maggiore numerosità dei casi nelle province considerate, ma non permette di stabilire, sulla sola base del conteggio, una maggiore esposizione individuale al fenomeno. Il confronto con la popolazione residente rappresenterebbe infatti un ulteriore livello di analisi, necessario per interpretare correttamente le differenze territoriali in termini relativi. L'elevato numero di eventi registrati in una grande provincia non è, quindi, da solo, sufficiente a stabilire un rischio individuale maggiore rispetto a una provincia meno popolata.

4.2.5 Dashboard 3: "Profilazione sociologica della vittima"

La terza dashboard modifica il livello di osservazione, passando dall'evento alla persona. L'obiettivo è descrivere le caratteristiche sociodemografiche delle vittime registrate e costruire una profilazione articolata che consideri genere, età, condizione occupazionale, genitorialità

e nazionalità. Mentre le dashboard dedicate al fenomeno nel suo complesso riportano 780 eventi, questa vista considera 895 persone coinvolte come vittime. Tale differenza è dovuta al fatto che, nel *dataset*, sono presenti vittime collaterali coinvolte nello stesso evento. La struttura complessiva della dashboard è illustrata nella Figura 4.3.



Figura 4.3: Dashboard 3: "Profilazione sociologica della vittima"

I principali elementi di visualizzazione presenti nella dashboard, che verranno analizzati singolarmente nelle sezioni successive, sono:

- *indicatori KPI*;
- *filtri dei dati*;
- *grafici a barre orizzontali*;
- *istogrammi a colonne*.

Indicatori KPI

Nella parte superiore della dashboard sono presenti quattro *indicatori KPI*, che riportano alcune informazioni di sintesi relative ai dati analizzati e offrono un primo quadro generale del fenomeno prima di passare all'analisi dei grafici. Inoltre, l'applicazione di filtri temporali o di altri selettori modifica il numero di eventi considerati e, di conseguenza, il contenuto degli altri grafici. Gli indicatori KPI sono:

- *Totale Donne Coinvolte (832)*: rappresenta il numero complessivo di donne registrate come vittime, comprendendo sia le vittime principali che quelle collaterali.
- *Totale Uomini Coinvolti (58)*: identifica il numero di uomini presenti come vittime collaterali. La sua presenza spiega perché il conteggio delle persone non coincida automaticamente con il numero di donne vittime degli eventi.
- *Totale Persone Trans Coinvolte (5)*: rappresenta il numero complessivo di persone trans registrate come vittime, comprendendo sia le vittime principali che quelle collaterali.

- *Totale Vittime Complessivo (895)*: sintetizza il numero complessivo delle vittime presenti nel *dataset*.

Filtro dei dati

Per garantire un'esplorazione interattiva e personalizzata del *dataset*, la sezione destra della dashboard ospita un pannello di controllo composto da tre filtri. L'interazione dell'utente con questi componenti modifica in tempo reale il contesto di filtro del modello dei dati, innescando il ricalcolo istantaneo delle misure DAX e l'aggiornamento simultaneo di tutte le visualizzazioni presenti nella pagina. Il primo strumento a disposizione è uno *slider* orizzontale che consente di restringere la finestra temporale di analisi, definendo l'anno iniziale e quello finale. Al di sotto di esso sono implementati due selettori a pulsanti. Il primo agisce sulla dimensione *Sesso vittima* e permette di isolare i dati relativi a "femmina", "maschio" o "persona_trans*". Il secondo controllo, dotato di una barra di scorrimento verticale, filtra il *dataset* in funzione della *Relazione Aggressore-Vittima*. Quest'ultimo strumento (che include opzioni come "affittuario", "amante" o "amico") consente di osservare in che modo i profili anagrafici e occupazionali varino al mutare del contesto relazionale in cui matura il delitto.

Grafico a barre orizzontali: "Profilo Occupazionale della Vittima"

Il grafico sulla sinistra classifica le vittime sulla base della condizione occupazionale o sociale registrata al momento dell'evento. La scelta delle barre orizzontali è adatta alla presenza di etichette categoriali diverse e permette di confrontare facilmente la lunghezza associata a ciascuna condizione. La visualizzazione mostra le condizioni occupazionali maggiormente rappresentate, concentrando l'attenzione sulle categorie che contribuiscono in misura più elevata alla distribuzione osservata. Per evitare che le numerose categorie occupazionali presenti nel *dataset* rendessero il grafico eccessivamente affollato, è stato applicato un filtro sui risultati con il maggior numero di casi. Questo strumento è stato applicato alla visualizzazione e non comporta la rimozione delle altre categorie dal modello dei dati. La distribuzione è caratterizzata dalla prevalenza della categoria "pensionata", seguita da "casalinga" e "extralegale". Tuttavia, la maggiore frequenza di una determinata categoria non permette, da sola, di affermare che essa costituisca un fattore causale o un indicatore di rischio; per sostenere una simile interpretazione sarebbe necessario disporre di un denominatore di riferimento e confrontare i casi osservati con la composizione della popolazione corrispondente. La dashboard utilizza, invece, la variabile occupazionale come elemento di profilazione descrittiva.

Istogramma a colonne: "Profilo Anagrafico della Vittima"

Al centro della dashboard, l'istogramma a colonne mostra la distribuzione delle vittime per fascia d'età. L'altezza delle colonne consente di confrontare direttamente il volume dei casi nelle diverse classi. Il grafico mostra che la distribuzione non è concentrata in una singola età, ma attraversa l'intero ciclo di vita, con un massimo nella fascia adulta compresa tra 30 e 44 anni. La lettura della distribuzione evidenzia, inoltre, una presenza significativa nelle fasce adulte, mentre le fasce più giovani e quelle più avanzate risultano rappresentate con frequenze differenti. L'organizzazione dei dati per intervalli consente di cogliere tali differenze e di confrontare la consistenza relativa dei diversi segmenti anagrafici. Il valore massimo osservato nella fascia 30–44 anni rappresenta, quindi, il principale elemento caratterizzante della distribuzione, senza, tuttavia, indicare una maggiore esposizione al fenomeno in termini di rischio individuale. Dal punto di vista dell'analisi descrittiva, la suddivisione per fasce d'età del *dataset* rende più immediata la lettura rispetto a una variabile di età continua. Essa

permette, infatti, di riconoscere quali segmenti anagrafici contribuiscano maggiormente al totale. Il grafico è, tuttavia, descrittivo e non consente di dedurre direttamente differenze di rischio tra fasce d'età senza un opportuno confronto con la popolazione di riferimento.

Istogramma a colonne: "Profilo di Genitorialità della Vittima"

Nella parte centrale inferiore è presente il grafico relativo al numero di figli associati a ciascuna vittima. L'asse orizzontale riporta il numero di figli, mentre l'asse verticale indica il numero complessivo di vittime corrispondenti a ciascuna categoria. La distribuzione mostra una forte concentrazione nella categoria "0 figli", che rappresenta la barra più alta. Seguono le categorie "1 figlio" e "2 figli", entrambe con volumi rilevanti ma inferiori al precedente valore osservato. Questa visualizzazione aggiunge una dimensione familiare alla profilazione della vittima. A differenza del KPI sulla media dei figli presente nella prima dashboard, il grafico non restituisce un unico valore riassuntivo, ma mostra l'intera distribuzione; esso permette, quindi, di distinguere tra l'assenza di figli, la presenza di uno o due figli e le situazioni meno frequenti caratterizzate da una numerosità familiare maggiore. La distribuzione assume particolare rilevanza nell'interpretazione del fenomeno poiché consente di evidenziare come, in una parte dei casi analizzati, la vittima fosse anche madre. La presenza di figli introduce, inoltre, una dimensione familiare che si estende oltre la persona direttamente coinvolta nell'evento. Il dato non viene utilizzato per attribuire al fenomeno ulteriori conseguenze non direttamente osservabili nel *dataset*, ma permette di mettere in luce una componente familiare della vicenda che rimarrebbe meno evidente attraverso i soli indicatori anagrafici e quantitativi.

Grafico a barre orizzontali: "Provenienza Geografica della Vittima"

Nel quadrante inferiore destro viene rappresentata la provenienza geografica mediante un grafico a barre orizzontali. Anche per questa visualizzazione è stato applicato un filtro sui risultati con il maggior numero di casi, così da mantenere nel grafico le provenienze geografiche che risultano maggiormente rappresentate. La visualizzazione consente inoltre di confrontare in modo immediato le diverse provenienze considerate, evidenziando la distanza tra la categoria maggiormente rappresentata e le successive. La categoria "Italia" presenta una barra più lunga rispetto a tutte le altre, seguita da "Romania" e, a distanza considerevole, da "Albania" e "Marocco". La prevalenza della categoria "Italia" deve essere interpretata tenendo conto della composizione complessiva del *dataset* e della popolazione di riferimento, poiché il semplice conteggio dei casi non consente di effettuare confronti sulla frequenza relativa del fenomeno tra persone appartenenti a differenti nazionalità. Il grafico assume, quindi, principalmente una funzione descrittiva, utile a delineare la composizione delle vittime registrate rispetto alla provenienza geografica.

4.2.6 Dashboard 4: "Profilazione sociologica dell'autore del reato"

La quarta dashboard conclude il percorso delle analisi di base concentrandosi sulle caratteristiche dell'autore del reato. La struttura riprende, intenzionalmente, quella della dashboard precedente, così da rendere possibile un confronto omogeneo tra i due profili. L'obiettivo è fornire una visione complessiva del profilo dell'autore, evidenziando le principali caratteristiche presenti nei dati disponibili. L'integrazione di queste diverse dimensioni permette, quindi, di affiancare alla precedente analisi delle vittime una lettura complementare del fenomeno. Accanto alle variabili anagrafiche, occupazionali e geografiche, questa pagina introduce, però, anche indicatori comportamentali specifici e una matrice che mette in relazione le fasce d'età

della vittima con quelle dell'aggressore. La struttura complessiva della dashboard è illustrata nella Figura 4.4.

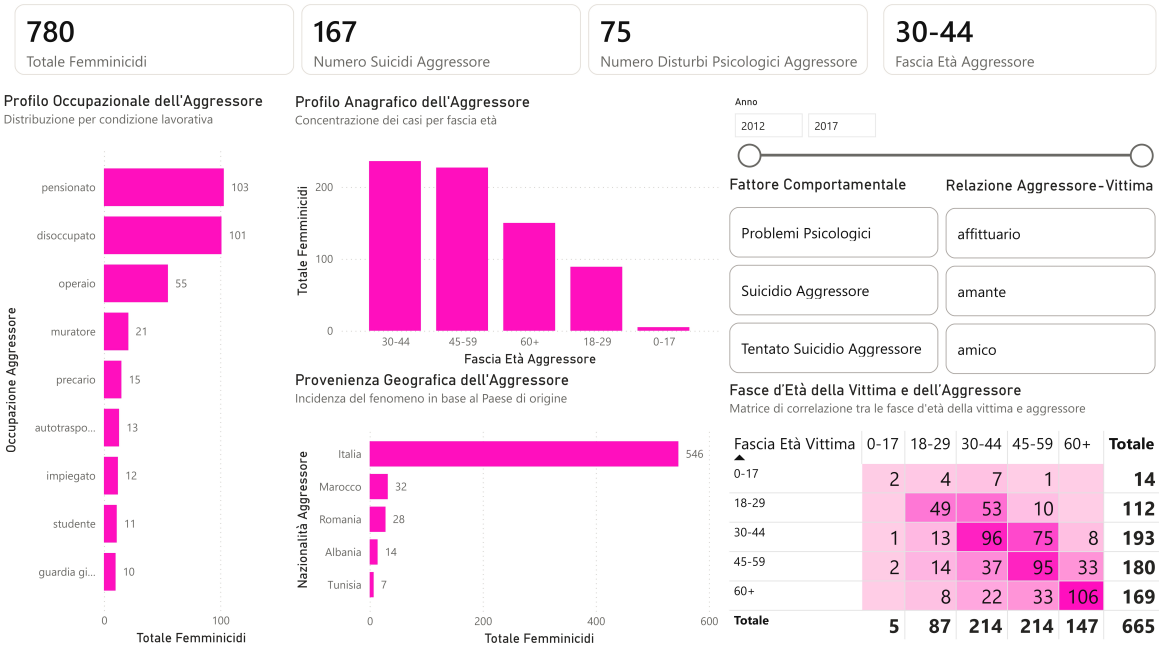


Figura 4.4: Dashboard 4: "Profilazione sociologica dell'autore del reato"

I principali elementi di visualizzazione presenti nella dashboard, che verranno analizzati singolarmente nelle sezioni successive, sono:

- *indicatori KPI*;
- *filtri dei dati*;
- *grafici a barre orizzontali*;
- *istogramma a colonne*;
- *matrice*.

Indicatori KPI

I tre KPI superiori sintetizzano dimensioni differenti del profilo dell'autore. Gli indicatori KPI considerati sono:

- *Totale Femminicidi (780)*: rappresenta il numero degli eventi compresi nella vista e costituisce il denominatore di riferimento per la lettura del resto della dashboard. La presenza dello stesso totale delle prime dashboard facilita il confronto con le analisi precedenti, mantenendo costante la numerosità degli eventi osservati.
- *Numero Suicidi Aggressore (167)*: quantifica gli eventi nei quali è registrato il suicidio dell'aggressore. Il dato permette di distinguere le dinamiche successive.
- *Fascia Età Aggressore (30-44)*: individua la fascia d'età maggiormente rappresentata tra gli autori nella vista corrente.

Filtro dei dati

Per garantire un'esplorazione interattiva e personalizzata del *dataset*, la sezione destra della dashboard ospita un pannello di controllo composto da tre filtri. L'interazione dell'utente con questi componenti modifica in tempo reale il contesto di filtro del modello dati, innescando il ricalcolo istantaneo delle misure DAX e l'aggiornamento simultaneo di tutte le visualizzazioni presenti nella pagina. Il primo strumento a disposizione è uno *slider* orizzontale che consente di restringere la finestra temporale di analisi, definendo l'anno iniziale e quello finale. Al di sotto di esso sono implementati due selettori a pulsanti. Il selettore relativo ai fattori comportamentali consente di isolare i casi classificati come *Problemi Psicologici*, *Suicidio Aggressore* o *Tentato Suicidio Aggressore*. Il secondo filtro, dotato di una barra di scorrimento verticale, filtra il *dataset* in funzione della *Relazione Aggressore-Vittima*, mantenendo la funzione della dashboard precedente. La combinazione di questi strumenti permette di osservare, ad esempio, come cambino età, occupazione o provenienza geografica quando vengono selezionati solo gli eventi che possiedono una determinata caratteristica comportamentale.

Grafico a barre orizzontali: "Profilo Occupazionale dell'Aggressore"

Il grafico posizionato sulla sinistra descrive la condizione occupazionale degli autori del reato. Anche in questo caso è stato applicato un filtro sui risultati con il maggior numero di casi, in modo da concentrare la visualizzazione sulle categorie occupazionali più rappresentate. La scelta consente di ridurre la frammentazione grafica, mantenendo comunque inalterate nel modello le categorie non visualizzate. Le barre mostrano una prevalenza delle categorie *pensionato* e *disoccupato*, seguite da *operaio*. Un elemento di interesse emerge dal confronto con il profilo occupazionale delle vittime, nel quale le categorie maggiormente rappresentate sono *pensionata* e *casalinga*, seguite da *extralegale*. Si osserva, quindi, una parziale ricorrenza di condizioni caratterizzate dall'assenza di un'occupazione lavorativa attiva tra le categorie più frequenti nei due profili, in particolare per quanto riguarda le condizioni di *pensionato/a* e, per gli aggressori, di *disoccupato*. Tale corrispondenza costituisce un elemento descrittivo di interesse all'interno della distribuzione osservata, ma non consente di stabilire l'esistenza di una relazione tra condizione occupazionale e fenomeno analizzato. Dal punto di vista interpretativo, la distribuzione occupazionale deve essere considerata in parallelo con quella delle vittime, ma senza stabilire automaticamente rapporti causali. Le frequenze osservate descrivono la struttura del campione analizzato; non indicano, da sole, che una condizione lavorativa determini un rischio maggiore, per il quale sarebbe necessario un confronto con la popolazione di riferimento.

Istogramma a colonne: "Profilo Anagrafico dell'Aggressore"

Il grafico centrale superiore rappresenta la distribuzione degli aggressori per fascia d'età e utilizza le stesse classi anagrafiche adottate per le vittime. La fascia "30-44" è la più rappresentata, con un valore molto vicino a quello della fascia "45-59". Un elemento di particolare interesse emerge dal confronto con il profilo anagrafico delle vittime: la graduatoria delle fasce d'età maggiormente rappresentate risulta, infatti, la medesima nei due profili. Anche per gli aggressori viene mantenuto lo stesso ordine delle fasce osservato nella distribuzione anagrafica delle vittime. Tale corrispondenza rappresenta un elemento descrittivo rilevante e suggerisce una distribuzione anagrafica dei due gruppi caratterizzata da una struttura analoga, senza tuttavia consentire, sulla base della sola distribuzione marginale, di stabilire una relazione tra l'età della vittima e quella dell'aggressore. L'adozione delle medesime classi anagrafiche utilizzate nella dashboard 3 è utile perché rende possibile un confronto

diretto. Non si tratta, quindi, soltanto di descrivere l'età degli aggressori, ma di predisporre le variabili necessarie alla successiva matrice incrociata presente nella parte inferiore destra.

Grafico a barre orizzontali: "Provenienza Geografica dell'Aggressore"

Il grafico sulla provenienza geografica completa il profilo dell'autore. Come per il corrispondente grafico della dashboard dedicata alle vittime, è stato applicato un filtro sui risultati con il maggior numero di casi. In questo modo la visualizzazione concentra l'attenzione sulle provenienze maggiormente rappresentate e consente un confronto più agevole tra le categorie principali. La categoria "Italia" presenta una frequenza nettamente superiore alle altre, con 546 casi, mentre seguono, con valori più ridotti, "Marocco" e "Romania". Un elemento di interesse emerge dal confronto con la provenienza geografica delle vittime, per le quali l'"Italia" rappresenta analogamente la categoria più frequente, con 637 casi. Inoltre, considerando le categorie maggiormente rappresentate, si osserva una sostanziale corrispondenza nella composizione delle prime posizioni: Italia, Romania, Albania e Marocco per le vittime e Italia, Marocco e Romania tra le principali provenienze degli aggressori. Sebbene l'ordine delle categorie non sia perfettamente coincidente e il numero di categorie visualizzate possa differire in funzione del filtro applicato, la presenza ricorrente delle stesse provenienze costituisce un elemento descrittivo di interesse nel confronto tra i due profili. L'interesse analitico di questa visualizzazione risiede nella possibilità di osservare la composizione geografica degli autori e di confrontarla con il profilo delle vittime. Anche qui, tuttavia, il confronto deve rimanere descrittivo: il maggior numero assoluto di record associati a una provenienza non costituisce una misura di rischio e non può essere interpretato senza tener conto delle dimensioni delle popolazioni di riferimento.

Matrice: "Correlazione delle Fasce d'Età" vittima-aggressore"

La visualizzazione più articolata della quarta dashboard è la matrice, o mappa di calore, collocata nel quadrante inferiore destro. Le colonne rappresentano le fasce d'età della vittima "0-17", "18-29", "30-44", "45-59", "60+", mentre le righe rappresentano le corrispondenti fasce d'età dell'aggressore. Ogni cella contiene il numero di eventi associato all'intersezione delle due categorie e il colore diventa progressivamente più intenso all'aumentare della frequenza. Questa struttura consente una lettura che non è possibile ottenere attraverso due istogrammi separati. I grafici anagrafici mostrano, infatti, la distribuzione marginale di vittime e aggressori, mentre la matrice osserva contemporaneamente le due variabili e permette di capire quali combinazioni siano più frequenti. Un primo elemento evidente è la presenza di valori elevati lungo la diagonale principale, cioè nei casi in cui vittima e aggressore appartengono alla stessa fascia d'età. In particolare, l'incrocio "30-44 vittima" e "30-44 aggressore" presenta 96 casi, mentre l'incrocio "60+ vittima" e "60+ aggressore" raggiunge 106 casi. La concentrazione della frequenza sulla diagonale costituisce un'evidenza descrittiva di una maggiore presenza di coppie appartenenti alla stessa classe anagrafica rispetto a molte delle combinazioni più distanti. Si può parlare, dunque, di associazione osservata tra fasce d'età piuttosto che di correlazione in senso statistico stretto: la mappa di calore non calcola, infatti, un coefficiente di correlazione, ma visualizza conteggi congiunti. La matrice permette, inoltre, di osservare gli incroci tra generazioni diverse. Ad esempio, nella fascia adulta compaiono anche valori rilevanti fuori dalla diagonale. Questi valori mostrano che la distribuzione non è esclusivamente diagonale: esistono anche combinazioni intergenerazionali che contribuiscono in modo significativo al totale. Nel complesso, la mappa aggiunge una dimensione relazionale all'analisi anagrafica. Non si limita a dire quale sia la fascia d'età più frequente per vittime e aggressori separatamente, ma permette di individuare le coppie di fasce maggiormente rappresentate e di riconoscere visivamente la concentrazione delle frequenze. Questa rap-

presentazione costituisce il punto di maggiore integrazione analitica della quarta dashboard, poiché mette in relazione due dimensioni precedentemente osservate in modo indipendente.

4.3 Considerazioni sulle dashboard di base

Considerate nel loro insieme, le quattro dashboard costruiscono un percorso di analisi che procede progressivamente dall'evento alla persona e, infine, al contesto in cui il fenomeno si manifesta.

4.3.1 Sintesi del percorso di analisi

La prima dashboard si concentra sulle caratteristiche dell'evento, cercando di descrivere *come* esso si manifesta attraverso le modalità esecutive, il luogo, il tipo di relazione e le circostanze associate. La seconda sposta l'attenzione sulle dimensioni spazio-temporali, permettendo di osservare *dove* e *quando* gli eventi si distribuiscono sul territorio nazionale. La terza sposta, invece, l'attenzione sulle vittime, delineandone un profilo anagrafico, sociodemografico e familiare. La quarta completa questo percorso concentrandosi sugli aggressori e mettendo in relazione alcune delle loro caratteristiche con quelle delle vittime. In questo modo, le dashboard non rappresentano quattro visualizzazioni isolate, ma diverse prospettive attraverso cui osservare uno stesso fenomeno.

Anche la scelta delle visualizzazioni risponde a questa logica. Le schede KPI sono state utilizzate per restituire valori sintetici; i grafici a barre permettono di confrontare con chiarezza categorie differenti; i grafici ad anello consentono di rappresentare la composizione percentuale di insiemi relativamente contenuti; la mappa di Azure introduce un confronto territoriale del fenomeno; infine, la mappa di calore permette di osservare simultaneamente due variabili categoriali e la loro distribuzione congiunta. La scelta della visualizzazione è stata orientata dalla necessità di rendere leggibile la specifica informazione contenuta nei dati. Lo sviluppo delle quattro dashboard ha rappresentato una verifica concreta dell'architettura logica definita nelle fasi precedenti del progetto. Il *dataset*, inizialmente costituito da numerose variabili eterogenee, è stato trasformato attraverso i processi di ETL e modellazione in una struttura interrogabile e visualizzabile in modo dinamico. L'impiego del motore *VertiPaq*, delle misure DAX e dei filtri interattivi ha permesso di passare da una semplice consultazione dei dati a un'esplorazione multidimensionale del fenomeno, nella quale le diverse dimensioni possono essere osservate singolarmente oppure messe in relazione tra loro.

4.3.2 Principali evidenze emerse dall'analisi

L'analisi visiva ha, inoltre, permesso di individuare alcuni elementi ricorrenti che risultano difficilmente percepibili attraverso la sola visione delle singole variabili. L'osservazione congiunta delle modalità dell'evento e del contesto relazionale mette in evidenza il peso della dimensione familiare e affettiva all'interno dei casi analizzati. Allo stesso modo, il confronto tra le caratteristiche delle vittime e quelle degli aggressori permette di cogliere alcune corrispondenze nelle distribuzioni. Ad esempio, le fasce d'età maggiormente rappresentate occupano posizioni analoghe nei due profili, mentre anche le provenienze geografiche più frequenti presentano alcune categorie ricorrenti. Si tratta di elementi che, considerati singolarmente, potrebbero apparire poco significativi, ma che acquistano maggiore interesse quando vengono osservati all'interno di una rappresentazione complessiva. Un'osservazione analoga riguarda il profilo occupazionale. La presenza tra le categorie, di condizioni quali "pensionata" e "casalinga" per le vittime e "pensionato" e "disoccupato" per gli aggressori, costituisce un elemento descrittivo che merita attenzione. La parziale sovrapposizione delle categorie

maggiormente rappresentate non permette, tuttavia, di stabilire un rapporto causale tra condizione occupazionale e fenomeno analizzato. Essa mostra, piuttosto, come la dimensione lavorativa e sociale possa essere utilizzata come una delle chiavi attraverso cui descrivere la popolazione presente nel dataset, lasciando eventuali interpretazioni di carattere esplicativo a studi che dispongano di ulteriori variabili e di adeguati denominatori di riferimento. Anche la dimensione territoriale assume un significato diverso quando viene osservata attraverso le dashboard. La distribuzione degli eventi non appare uniforme sul territorio nazionale e, in termini di valori assoluti, alcune aree e province presentano un numero maggiore di casi. La maggiore presenza di eventi nelle aree più popolose deve, tuttavia, essere interpretata tenendo distinta la frequenza assoluta dall'incidenza sulla popolazione: il numero di casi, da solo, non consente, infatti, di stabilire una maggiore esposizione al fenomeno. Questo aspetto evidenzia uno dei vantaggi dell'approccio adottato: la visualizzazione permette di individuare una distribuzione di dati e, allo stesso tempo, rende evidente la necessità di ulteriori livelli di analisi per interpretarla correttamente.

4.3.3 Valore della soluzione di Business Intelligence

Nel complesso, il principale risultato del lavoro consiste nella costruzione di uno strumento attraverso cui i dati possono essere interrogati da prospettive differenti. Le dashboard permettono di passare dalla semplice domanda *"quanti sono i casi?"* a interrogativi più articolati riguardanti la loro distribuzione nel tempo e nello spazio, le caratteristiche delle persone coinvolte e le relazioni tra le diverse variabili disponibili. È proprio nel passaggio dal singolo dato alla lettura congiunta di più dimensioni che la *Business Intelligence* mostra il proprio valore analitico. In questa prospettiva, il progetto evidenzia anche un limite importante: una dashboard può rendere visibili distribuzioni, ricorrenze e associazioni presenti nei dati, ma non può, da sola, spiegare le cause del fenomeno. Le frequenze osservate non devono, quindi, essere interpretate automaticamente come fattori di rischio o relazioni causali. Per giungere a conclusioni di questo tipo sarebbero necessarie ulteriori informazioni, confronti con la popolazione di riferimento e metodologie statistiche specificamente orientate all'analisi causale. La visualizzazione rappresenta, piuttosto, un punto di partenza: permette di individuare *pattern*, formulare domande e riconoscere quali aspetti meritino un approfondimento successivo. Da questo punto di vista, l'integrazione tra *Data Science*, *Data Visualization* e *Business Intelligence* assume un valore che va oltre la rappresentazione grafica. La trasformazione di dati storici attraverso ETL, modellazione relazionale e misure DAX consente, infatti, di costruire una lettura più accessibile, interattiva e multidimensionale di un fenomeno complesso. Le dashboard realizzate non sostituiscono l'analisi statistica o criminologica, ma possono costituirne un supporto, rendendo più immediata l'individuazione delle distribuzioni e delle relazioni che meritano di essere studiate. In questo senso, il lavoro mostra come l'organizzazione e la visualizzazione dei dati possano contribuire non solo a descrivere ciò che è già noto, ma anche a porre nuove domande sul fenomeno osservato.

CAPITOLO 5

Progettazione e realizzazione delle dashboard avanzate

Il presente capitolo descrive l'evoluzione del progetto di Business Intelligence verso una fase di analisi più approfondita, dedicata alla progettazione e alla realizzazione delle dashboard avanzate. Dopo aver fornito, nel capitolo precedente, una visione d'insieme del fenomeno attraverso un'analisi prevalentemente descrittiva, l'attenzione si concentra ora sull'esplorazione delle relazioni tra le diverse dimensioni presenti nel dataset e sulla loro distribuzione nel tempo e nello spazio. L'analisi si sviluppa attraverso tre nuove dashboard interattive, progettate con l'obiettivo di ampliare la lettura dei dati oltre la semplice rappresentazione dei valori aggregati. Attraverso visualizzazioni quali alberi di scomposizione, mappe ad albero, matrici, diagrammi di flusso, grafici a nastri e mappe territoriali, vengono analizzate congiuntamente le dinamiche relazionali, i fattori comportamentali, i moventi e l'evoluzione temporale e territoriale del fenomeno. L'obiettivo di questa fase è, quindi, quello di approfondire la distribuzione dei casi attraverso differenti prospettive di analisi, mettendo in evidenza ricorrenze e configurazioni che possono emergere dall'incrocio tra più variabili. Le dashboard avanzate non si limitano, pertanto, a rappresentare i dati, ma offrono un ambiente interattivo per esplorarne la struttura e supportare una lettura più articolata delle informazioni disponibili, fornendo elementi utili per ulteriori interpretazioni in ambito sociologico e criminologico.

5.1 Introduzione all'analisi avanzata

A seguito dell'implementazione delle dashboard di base, finalizzate principalmente alla descrizione quantitativa e all'esplorazione delle principali caratteristiche del fenomeno analizzato, il progetto è proseguito con la realizzazione di un ulteriore livello di analisi, definito *analisi avanzata*. Tale fase nasce dall'esigenza di superare una lettura dei dati esclusivamente descrittiva e di introdurre strumenti in grado di mettere in relazione tra loro differenti dimensioni del fenomeno, evidenziando ulteriori associazioni e ricorrenze significative. Nell'ambito della *Business Intelligence*, il concetto di *Advanced Analytics* indica un insieme di metodologie e strumenti finalizzati ad ampliare le possibilità di interpretazione dei dati oltre la semplice aggregazione e rappresentazione dei valori osservati. Se le dashboard di base consentono principalmente di rispondere alla domanda *"che cosa è successo?"*, l'analisi avanzata introduce una prospettiva maggiormente orientata alla comprensione delle relazioni tra le variabili e, conseguentemente, alla domanda *"perché è successo?"*. Tale passaggio non implica necessariamente l'individuazione di un rapporto causale tra le variabili considerate, ma consente di osservare come determinate caratteristiche del fenomeno tendano a presentarsi congiuntamente all'interno dei dati disponibili. Le dashboard avanzate sono state progettate con l'obiettivo di passare da una rappresentazione prevalentemente aggregata dei volumi a una

lettura multidimensionale del fenomeno. In questa prospettiva, l'attenzione non è rivolta esclusivamente alla quantità complessiva dei casi, ma alla modalità con cui tale quantità si distribuisce in funzione di variabili quali il legame tra autore del reato e vittima, il contesto relazionale, il movente, gli strumenti utilizzati, i fattori comportamentali, la scena del crimine e la visione temporale.

Un aspetto centrale di questa fase consiste, quindi, nella destrutturazione delle categorie aggregate. Attraverso visualizzazioni interattive e strumenti analitici avanzati è possibile partire da un valore complessivo e procedere, successivamente, verso livelli di dettaglio successivi, osservando come il fenomeno si articoli all'interno delle diverse categorie. In questo modo, l'utente non si limita a osservare il valore finale di una misura, ma può analizzare il percorso che conduce a tale valore.

A questo scopo sono state impiegate differenti tipologie di visualizzazione, selezionate in funzione della specifica domanda analitica. Tra queste rientrano gli alberi di scomposizione, le mappe ad albero, i grafici Sankey, i grafici a nastri e le rappresentazioni temporali e geografiche. Ciascuno strumento consente, infatti, di evidenziare una particolare dimensione del fenomeno. Gli alberi di scomposizione permettono di seguire gerarchie di analisi; le mappe ad albero consentono di confrontare il peso relativo dei dati gerarchicamente organizzati; i diagrammi Sankey rappresentano i flussi tra dimensioni differenti, mentre i grafici a nastri permettono di osservare la variazione nel tempo del posizionamento relativo delle categorie. L'analisi avanzata introduce, inoltre, una lettura maggiormente orientata all'individuazione di *pattern* ricorrenti. L'osservazione congiunta di più variabili permette di mettere in evidenza configurazioni che potrebbero non risultare immediatamente visibili attraverso l'analisi delle singole variabili considerate separatamente. Nel contesto del presente studio, questo approccio viene applicato, in particolare, all'analisi delle dinamiche relazionali, dei fattori comportamentali e dei fattori di rischio, nonché all'evoluzione temporale e territoriale del fenomeno.

Nel caso specifico dell'analisi del femminicidio, il passaggio alle dashboard avanzate permette, quindi, di affiancare alla precedente rendicontazione quantitativa una lettura più articolata delle dinamiche che caratterizzano gli episodi registrati. L'obiettivo è destrutturare le categorie aggregate per individuare i percorsi di letalità maggiormente ricorrenti, analizzare la distribuzione dei fattori comportamentali in relazione al legame tra autore del reato e vittima e osservare l'evoluzione temporale di specifici fattori critici. L'insieme delle visualizzazioni costituisce, in questo modo, uno strumento conoscitivo a supporto dell'interpretazione del fenomeno da una prospettiva sociologica e criminologica, mantenendo come riferimento i dati effettivamente presenti nel *dataset* analizzato.

5.2 Analisi dashboard avanzate

La fase di analisi avanzata è stata sviluppata attraverso la realizzazione di tre dashboard, progettate per analizzare differenti dimensioni del fenomeno complementari tra loro. La progettazione delle dashboard avanzate è stata orientata alla costruzione di viste multidimensionali nelle quali ciascuna visualizzazione svolge una specifica funzione analitica. A differenza delle dashboard precedenti, nelle quali le visualizzazioni erano principalmente finalizzati alla rappresentazione di indicatori descrittivi e alla quantificazione del fenomeno, in questa fase ogni elemento è stato scelto per favorire il confronto e l'interpretazione delle relazioni tra variabili.

Le tre dashboard condividono alcuni elementi comuni, quali gli indicatori KPI, utilizzati per fornire un'immediata sintesi quantitativa dell'ambito analizzato, ma si differenziano per le successive visualizzazioni. Sono stati, infatti, utilizzati strumenti differenti, tra cui grafici ad area, grafici a barre in pila al 100%, matrici di scomposizione, mappe ad albero,

grafici tornado, mappe geospaziali, grafici Sankey e grafici a nastri. La combinazione di tali visualizzazioni permette di alternare una lettura sintetica, fornita dagli indicatori di sintesi, a una lettura analitica, resa possibile dalla capacità di incrociare variabili e approfondire specifiche categorie.

5.2.1 Struttura dell'analisi visiva

La struttura dell'analisi visiva è stata suddivisa nelle seguenti tre dashboard:

- *Dashboard 5: "Analisi multidimensionale delle dinamiche relazionali ed esecutive"*: mostra l'interazione tra il contesto relazionale e le modalità di realizaione degli eventi. La vista mette in relazione diretta il legame tra autore del reato e vittima con gli strumenti di letalità impiegati e con il movente, permettendo di destrutturare il modus operandi attraverso un percorso di esplorazione gerarchico.
- *Dashboard 6: "Analisi comportamentale, vulnerabilità e fattori di rischio"*: consente di esplorare i fattori comportamentali e gli elementi di vulnerabilità clinica associati all'autore del reato. La pagina analizza la distribuzione di dinamiche critiche, come il suicidio post-delitto o la presenza di pregressi disturbi psicologici, incrociandole direttamente con il tipo di legame intercorso con la vittima e con le cause scatenanti.
- *Dashboard 7: "Evoluzione temporale dei trend critici"*: introduce una prospettiva prevalentemente longitudinale e territoriale, con l'obiettivo di osservare l'evoluzione di specifici fattori critici negli anni. La dashboard permette di monitorare la variazione dei moventi, degli strumenti di letalità e delle scene del crimine nel tempo, affiancando una mappatura geografica avanzata per localizzare la distribuzione territoriale dei picchi di rischio.

Questa articolazione consente di procedere progressivamente dall'analisi delle dinamiche relazionali ed esecutive alla caratterizzazione dei fattori comportamentali e, infine, all'osservazione della loro evoluzione nel tempo e nello spazio. Le tre prospettive risultano complementari e permettono di integrare livelli di analisi differenti, mantenendo una lettura coerente e multidimensionale del fenomeno considerato.

5.2.2 Dashboard 5: "Analisi multidimensionale delle dinamiche relazionali ed esecutive"

La prima dashboard avanzata è dedicata all'analisi multidimensionale delle dinamiche relazionali ed esecutive associate agli episodi analizzati. L'obiettivo è quello di osservare il fenomeno considerando simultaneamente la relazione tra autore del reato e vittima, il contesto relazionale, il movente e lo strumento attraverso il quale viene realizzato il crimine. La dashboard consente, quindi, di passare dal semplice conteggio dei casi alla caratterizzazione delle modalità con cui gli episodi si distribuiscono all'interno delle diverse configurazioni relazionali ed esecutive. In particolare, l'analisi prende in considerazione la distinzione tra sfera intima ed extrafamiliare, la presenza di un movente passionale, i casi caratterizzati da elevata letalità e la combinazione tra tipologia di legame e strumento utilizzato.

La dashboard è stata progettata per favorire il confronto tra le diverse categorie e per evidenziare come la distribuzione degli eventi cambi in funzione delle caratteristiche relazionali e delle modalità esecutive considerate. Le visualizzazioni presenti consentono, inoltre, di approfondire alcune configurazioni specifiche attraverso filtri e livelli di dettaglio differenti, mantenendo come riferimento il totale degli eventi analizzati. In questo modo, l'analisi non si limita a descrivere la frequenza delle singole categorie, ma permette di osservare le loro

associazioni all'interno della base di dati, senza attribuire automaticamente un significato causale alle relazioni individuate. La composizione complessiva della dashboard è riportata nella Figura 5.1.

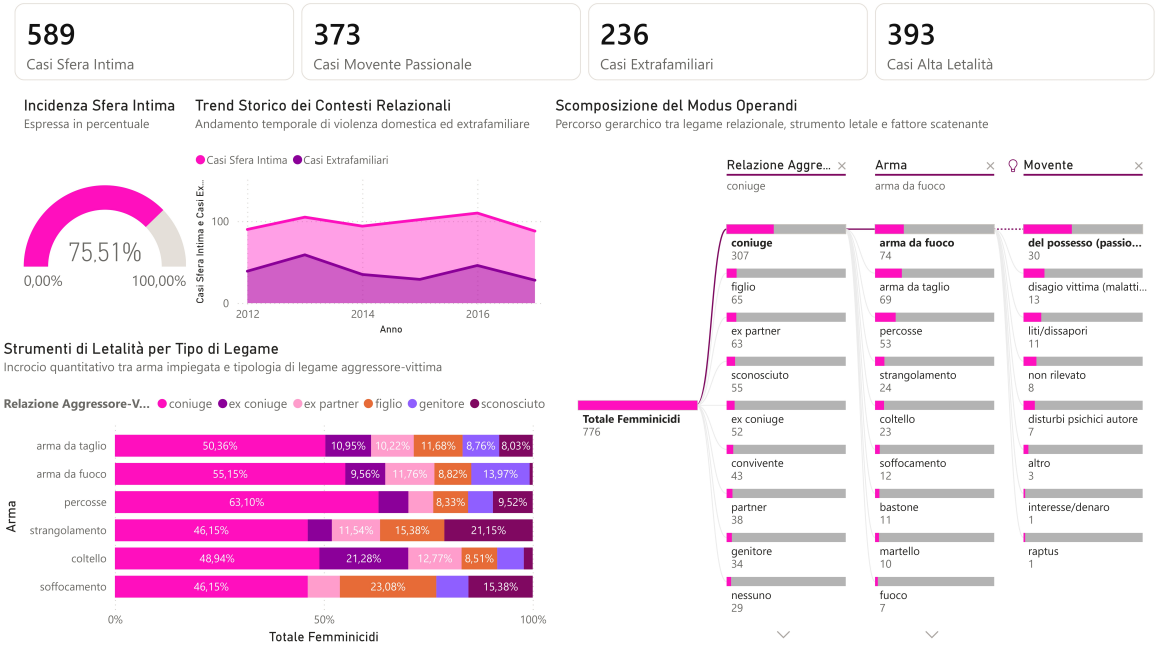


Figura 5.1: Dashboard 5: "Analisi multidimensionale delle dinamiche relazionali ed esecutive"

I principali elementi di visualizzazione presenti nella dashboard, che verranno analizzati singolarmente nelle sezioni successive, sono:

- *indicatori KPI*;
- *tachimetro*;
- *grafico ad area*;
- *grafico a barre in pila 100%*;
- *matrice di scomposizione*.

Indicatori KPI

La fascia superiore della dashboard presenta quattro indicatori KPI destinati a definire il perimetro quantitativo dell'analisi. I quattro indicatori KPI considerati sono:

- *Casi Sfera Intima (619)*: forniscono una misura della componente del fenomeno associata ai contesti relazionali più intimi della vittima, quali "coniuge", "partner", "ex partner", "ex coniuge", "amante", "ex amante", "genitore", "figlio" e "convivente".
- *Casi Movente Passionale (391)*: considera gli episodi caratterizzati dalla presenza di un movente del possesso.
- *Casi Extrafamiliari (247)*: misura i casi classificati come extrafamiliari.
- *Casi Alta Letalità (416)*: individua gli episodi caratterizzati da alta letalità, attraverso strumenti quali "arma da fuoco", "arma da taglio" e "coltello".

La funzione dei KPI non è soltanto quella di fornire quattro valori isolati, ma di costituire il riferimento quantitativo per la successiva interpretazione delle visualizzazioni. I valori riportati nella parte superiore consentono, infatti, di inquadrare le principali categorie sulle quali si concentra l'analisi successiva.

Tachimetro: "Incidenza Sfera Intima"

Accanto agli indicatori sintetici è stato inserito un grafico a tachimetro denominato *Incidenza Sfera Intima*. La visualizzazione ha lo scopo di rappresentare in termini percentuali il peso dei casi appartenenti alla sfera intima rispetto al totale degli eventi considerati. Il tachimetro restituisce un'incidenza pari al 75,51%; tale valore evidenzia come, nella vista analizzata, oltre tre quarti degli eventi siano associati a una relazione riconducibile alla sfera intima. In questa classificazione rientrano rapporti caratterizzati da una relazione affettiva, di coppia o familiare stretta. Il risultato mette, quindi, in evidenza la rilevanza delle relazioni personali strette all'interno degli eventi considerati: la maggior parte dei casi non si colloca, infatti, al di fuori di una relazione preesistente, ma presenta un legame tra vittima e autore del reato riconducibile alle categorie comprese nella sfera intima. Il dato percentuale consente, pertanto, di affiancare alla misura assoluta fornita dal KPI *Casi Sfera Intima* una lettura proporzionale del fenomeno, mostrando quanto questa componente incida sul totale degli eventi analizzati.

Grafico ad area: "Trend Storico dei Contesti Relazionali"

Nella parte centrale superiore della dashboard è presente un grafico ad area dedicato al *Trend Storico dei Contesti Relazionali*. La visualizzazione introduce la dimensione temporale nell'analisi e permette di osservare l'andamento dei casi nel periodo considerato, distinguendo tra due differenti contesti relazionali: la *Sfera Intima* e il contesto *Extrafamiliare*. L'asse temporale consente di seguire l'evoluzione degli episodi lungo gli anni compresi tra il 2012 e il 2017, mentre le aree rappresentate nel grafico permettono di confrontare visivamente il volume associato alle due categorie. La rappresentazione evidenzia una presenza costantemente superiore dei casi appartenenti alla sfera intima rispetto a quelli extrafamiliari nel periodo analizzato. L'area corrispondente alla sfera intima occupa, infatti, una porzione visivamente maggiore del grafico, mentre la componente extrafamiliare presenta valori inferiori. Il principale contributo analitico di questa visualizzazione consiste, quindi, nel mostrare che la differenza tra i due contesti non è limitata a un singolo anno, ma rimane osservabile lungo l'intervallo temporale rappresentato.

Grafico a barre in pila 100%: "Strumenti di Letalità per Tipo di Legame"

Nella parte inferiore sinistra della dashboard è presente un grafico a barre in pila normalizzato al 100%, denominato *Strumenti di Letalità per Tipo di Legame*. La visualizzazione è stata progettata per analizzare la relazione tra lo strumento o la modalità di letalità e la tipologia di legame esistente tra autore del reato e vittima. La normalizzazione al 100% fa sì che ciascuna barra rappresenti un'intera distribuzione percentuale, indipendentemente dal numero assoluto di casi associati alla relativa modalità di letalità. In questo modo, il confronto non riguarda direttamente la numerosità complessiva delle diverse categorie, ma la loro composizione interna, consentendo di osservare come si distribuiscono le diverse tipologie di legame all'interno di ciascuna modalità considerata. Sull'asse delle categorie sono riportate le differenti modalità di letalità presenti nel *dataset*, tra cui "arma da taglio", "arma da fuoco", "percosse", "strangolamento", "coltello" e "soffocamento". Ogni barra è, a sua volta, suddivisa in segmenti corrispondenti alle diverse tipologie di relazione tra

autore del reato e vittima, tra cui "coniuge", "ex coniuge", "ex partner", "figlio", "genitore" e "sconosciuto". L'osservazione della composizione percentuale evidenzia, ad esempio, una consistente presenza della categoria "coniuge" in corrispondenza di diverse modalità di letalità. Essa rappresenta il 63,10% della composizione relativa alla modalità "percosse" e il 55,15% di quella relativa all'"arma da fuoco". Questi valori devono essere interpretati come quote percentuali all'interno della specifica modalità di letalità e non come percentuali riferite all'intero insieme degli eventi analizzati. Il grafico permette, pertanto, di esaminare la composizione delle diverse modalità di letalità in funzione del tipo di legame tra autore del reato e vittima, mettendo in evidenza le differenze nella distribuzione relativa delle categorie relazionali. La scelta della normalizzazione consente, in particolare, di confrontare direttamente tali composizioni anche quando le modalità di letalità presentano numerosità complessive differenti.

Matrice di scomposizione: "Scomposizione del Modus Operandi"

La parte destra della dashboard è occupata dalla matrice di scomposizione, uno degli strumenti di analisi avanzata impiegati per approfondire la distribuzione del fenomeno attraverso una scomposizione progressiva della misura analizzata. La visualizzazione consente, infatti, di partire da un valore complessivo e di esaminarne, successivamente, la distribuzione rispetto a differenti dimensioni, costruendo un percorso gerarchico di approfondimento. Il punto di partenza dell'analisi è rappresentato dal nodo "Totale Femminicidi", corrispondente a 775 casi. A partire da questo valore complessivo, la matrice permette di procedere attraverso successive dimensioni di analisi, evidenziando come il totale si distribuisca al variare delle caratteristiche considerate.

La prima dimensione analizzata è la "Relazione Aggressore". All'interno di tale scomposizione, la voce "coniuge" è associata a 307 casi. A partire da questo sottoinsieme, l'analisi può essere ulteriormente articolata secondo la variabile "Arma". In corrispondenza della relazione "coniuge", risultano 74 casi associati all'"arma da fuoco" e 69 all'"arma da taglio". Il percorso di analisi può essere ulteriormente approfondito attraverso la variabile "Movente". Seguendo, ad esempio, il ramo relativo alla relazione "coniuge" e all'"arma da fuoco", il movente del "possesso (passionale)" risulta associato a 30 casi. La struttura della matrice consente, quindi, di seguire un percorso di approfondimento che parte dalla misura complessiva e procede attraverso dimensioni successive: dal totale dei casi alla relazione tra autore del reato e vittima, dalla relazione allo strumento utilizzato e, infine, dallo strumento al movente. La possibilità di modificare dinamicamente i livelli di scomposizione permette, inoltre, di esplorare differenti combinazioni di dimensioni e di individuare, all'interno del *dataset*, le distribuzioni maggiormente rappresentate lungo i diversi percorsi di analisi.

5.2.3 Dashboard 6: "Analisi comportamentale, vulnerabilità e fattori di rischio"

La seconda dashboard avanzata sposta l'attenzione dalle modalità esecutive alle caratteristiche comportamentali e ai fattori di rischio associati all'autore del reato. L'obiettivo è quello di analizzare la distribuzione di specifici fattori comportamentali in relazione al tipo di legame con la vittima e osservare la loro associazione con i diversi moventi. La dashboard consente, dunque, di integrare la dimensione relazionale con una serie di informazioni relative al comportamento dell'aggressore, considerando, in particolare, la presenza di suicidio successivo al delitto, il tentato suicidio e i problemi psicologici. Questa impostazione permette di esaminare come tali fattori si distribuiscano tra le diverse tipologie di relazione, mettendo in evidenza eventuali ricorrenze e differenze nella loro associazione con i moventi del delitto. La struttura complessiva della dashboard è riportata nella Figura 5.2.

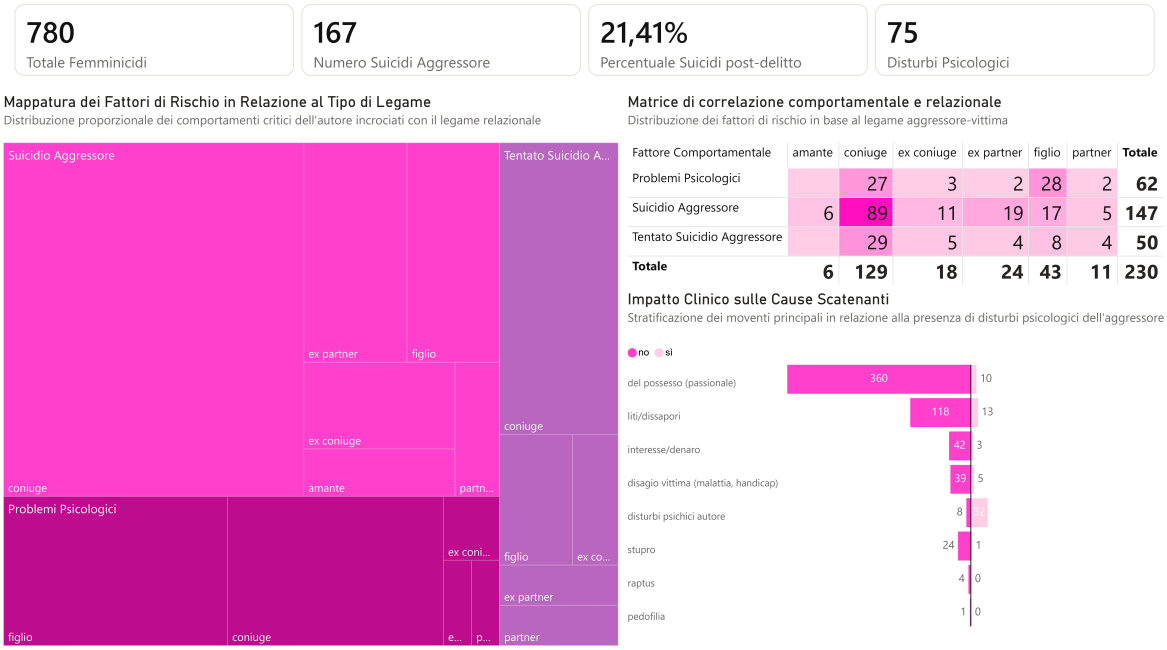


Figura 5.2: Dashboard 6: "Analisi comportamentale, vulnerabilità e fattori di rischio"

I principali elementi di visualizzazione presenti nella dashboard, che verranno analizzati singolarmente nelle sezioni successive, sono:

- *indicatori KPI*;
- *mappa ad albero*;
- *matrice*;
- *grafico tornado 3.2.1.*

Indicatori KPI

La parte superiore della dashboard presenta quattro indicatori KPI, finalizzati a fornire una sintesi quantitativa delle principali variabili considerate nell'analisi comportamentale. I quattro indicatori KPI considerati sono:

- *Totale Femminicidi (780)*: rappresenta il numero complessivo di eventi presenti nella vista corrente e costituisce il riferimento quantitativo generale per l'interpretazione degli altri indicatori.
- *Numero Suicidi Aggressore (167)*: indica il numero di casi nei quali, successivamente all'evento, si è registrato il suicidio dell'aggressore.
- *Percentuale Suicidi post-delitto (21,41%)*: esprime la quota percentuale degli eventi caratterizzati dal suicidio dell'aggressore rispetto al totale dei femminicidi considerati nella vista corrente.
- *Pregressi Disturbi Psicologici (75)*: indica il numero di casi nei quali, secondo le informazioni disponibili nel *dataset*, è registrata la presenza di pregressi disturbi psicologici dell'aggressore.

Nel loro insieme, i quattro indicatori consentono di delineare il perimetro quantitativo dell'analisi comportamentale. Il primo KPI fornisce il riferimento complessivo, mentre i successivi focalizzano l'attenzione su specifici fattori associati agli eventi analizzati. La presenza congiunta di valori assoluti e di un indicatore percentuale permette, inoltre, di distinguere la numerosità complessiva dei fenomeni considerati dalla loro incidenza relativa rispetto al totale degli eventi.

Mappa ad albero: "Mappatura dei Fattori di Rischio in Relazione al Tipo di Legame"

Nella parte sinistra della dashboard è collocata una mappa ad albero, denominata *Mappatura dei Fattori di Rischio in Relazione al Tipo di Legame*. La visualizzazione consente di rappresentare simultaneamente la distribuzione di alcuni fattori comportamentali e la loro articolazione in funzione del tipo di legame tra autore del reato e vittima.

La struttura della mappa è organizzata secondo una gerarchia a due livelli. Il primo livello è costituito dalle macro-categorie comportamentali considerate, ovvero "Suicidio Aggressore", "Problemi Psicologici" e "Tentato Suicidio Aggressore". Ciascuna di queste viene, successivamente, suddivisa in blocchi corrispondenti alle differenti tipologie di relazione tra autore del reato e vittima. La dimensione dei rettangoli rappresenta la numerosità relativa dei casi associati a ciascuna combinazione tra fattore comportamentale e tipo di legame: quanto maggiore è l'area del blocco, tanto maggiore è il numero di casi rappresentati. Tale caratteristica permette di individuare visivamente le combinazioni maggiormente rappresentate senza ricorrere alla lettura puntuale dei singoli valori.

Dalla distribuzione delle aree emerge una maggiore estensione complessiva della macrocategoria "Suicidio Aggressore" rispetto alle altre due dimensioni considerate. All'interno di quest'ultima, il blocco associato alla relazione "coniuge" presenta la maggiore numerosità seguito da "ex partner" e "figlio". Anche nelle sezioni relative a "Problemi Psicologici" e "Tentato Suicidio Aggressore", la voce "coniuge" risulta tra quelle caratterizzate dalla maggiore estensione visiva. La mappa ad albero consente, quindi, di mettere in relazione, all'interno di un'unica visualizzazione, i fattori comportamentali considerati e le diverse tipologie di legame, evidenziando le combinazioni che presentano una maggiore numerosità all'interno del *dataset*.

Matrice: "Matrice di correlazione comportamentale e relazionale"

Nella parte superiore destra della dashboard è presente una matrice che mette in relazione il `Fattore Comportamentale` con la `Relazione Aggressore-Vittima`. Le righe rappresentano i differenti fattori comportamentali considerati, mentre le colonne riportano le diverse tipologie di legame tra autore del reato e vittima. Ogni cella restituisce, quindi, la frequenza associata alla specifica combinazione tra un determinato fattore comportamentale e una determinata relazione. Per agevolare la lettura della matrice è stata applicata una formattazione condizionale mediante gradiente cromatico. L'intensità della colorazione varia in funzione della frequenza riportata nella singola cella, consentendo di individuare rapidamente le combinazioni caratterizzate dai valori più elevati e di distinguere visivamente le aree di maggiore concentrazione dei casi.

Tra le combinazioni maggiormente rappresentate si osserva l'incrocio tra "Suicidio Aggressore" e "coniuge", al quale corrispondono 89 casi. Il dato evidenzia che, all'interno della distribuzione osservata, questa specifica combinazione presenta una frequenza superiore rispetto alle altre combinazioni riportate nella matrice. La lettura congiunta delle righe e delle colonne permette, inoltre, di verificare come uno stesso fattore comportamentale possa distribuirsi tra differenti tipologie di relazione e, viceversa, come una medesima relazione possa essere associata a più fattori comportamentali.

La matrice consente, dunque, di introdurre una lettura più dettagliata della distribuzione dei fattori comportamentali. Ad esempio, la presenza di valori elevati in corrispondenza della relazione "coniuge" può essere confrontata con i valori osservati per "ex partner", "figlio", "genitore" o "sconosciuto", permettendo di individuare le combinazioni maggiormente rappresentate e quelle caratterizzate da frequenze inferiori. È importante sottolineare che la frequenza osservata in una cella descrive esclusivamente la numerosità della combinazione considerata e non implica, di per sé, l'esistenza di un rapporto causale tra il fattore comportamentale e il tipo di relazione. La matrice deve, pertanto, essere interpretata come uno strumento descrittivo, utile a evidenziare la distribuzione congiunta delle variabili presenti nel *dataset*.

Rispetto alla mappa ad albero, la matrice offre una modalità di lettura complementare. La prima privilegia la rappresentazione visiva dei pesi relativi attraverso l'estensione delle aree, rendendo immediatamente riconoscibili le combinazioni più numerose. La matrice, invece, permette di accedere direttamente al valore numerico associato a ciascuna combinazione, risultando particolarmente utile quando l'obiettivo è effettuare un confronto puntuale tra differenti incroci. L'integrazione delle due visualizzazioni consente, quindi, di combinare due livelli di lettura dell'analisi: la mappa ad albero favorisce l'individuazione immediata delle distribuzioni più rilevanti, mentre la matrice permette di verificarne con maggiore precisione la consistenza numerica. In questo modo, la dashboard non si limita a mostrare i singoli fattori comportamentali, ma ne approfondisce la distribuzione in relazione alle diverse tipologie di legame tra autore del reato e vittima.

Grafico tornado: "Impatto Clinico sulle Cause Scatenanti"

Nella parte inferiore destra della dashboard è presente un grafico a barre speculari, comunemente denominato *tornado chart*, dal titolo *Impatto Clinico sulle Cause Scatenanti*. La visualizzazione mette in relazione i principali moventi registrati nel dataset con la presenza o l'assenza di precedenti disturbi psicologici associati all'aggressore, consentendo di confrontare direttamente la distribuzione dei casi nelle due condizioni, indicate, rispettivamente, come "sì" e "no". La struttura speculare del grafico facilita il confronto tra le due distribuzioni: per ciascun movente; l'estensione delle barre permette di osservare la diversa numerosità dei casi in cui è presente o non è registrato un precedente disturbo psicologico. In questo modo, la visualizzazione consente di verificare se i singoli moventi risultino maggiormente rappresentati in una delle due condizioni.

Un dato particolarmente evidente riguarda il movente del "possesso (passionale)". Per questa tipologia sono registrati 360 casi nella condizione "no" e 10 nella condizione "sì". La distribuzione mostra, quindi, una marcata prevalenza dei casi nei quali non risultano registrati precedenti disturbi psicologici. Questo elemento è rilevante perché evidenzia come, almeno nella popolazione osservata, il movente del "possesso (passionale)" non sia, nella maggior parte dei casi, accompagnato dalla presenza di un disturbo psicologico precedentemente registrato nel *dataset*. L'osservazione della restante distribuzione consente, inoltre, di rilevare che la condizione "no" risulta prevalente anche per la maggior parte degli altri moventi rappresentati. Nel complesso il grafico non evidenzia una concentrazione generalizzata dei moventi nella presenza di precedenti disturbi psicologici; al contrario, i casi privi di un disturbo psicologico precedentemente registrato costituiscono, per diverse tipologie di movente, la componente numericamente più ampia. Tale evidenza contribuisce a fornire una lettura più articolata del rapporto tra dimensione comportamentale e movente, evitando di ricondurre automaticamente determinati moventi alla presenza di una condizione psicologica precedentemente documentata.

È, tuttavia, necessario interpretare con cautela questa distribuzione. La voce "no" indica che nel *dataset* non risulta registrato un precedente disturbo psicologico e non può essere

interpretata come dimostrazione dell'assenza di qualsiasi problematica clinica. Allo stesso modo, la presenza della voce "sì" segnala esclusivamente l'esistenza di un'informazione relativa a precedenti disturbi psicologici secondo le modalità di rilevazione adottate nella fonte dei dati.

5.2.4 Dashboard 7: "Evoluzione temporale dei trend critici"

La terza dashboard avanzata introduce una prospettiva prevalentemente temporale e territoriale. L'obiettivo è analizzare come alcuni dei principali fattori associati al fenomeno si distribuiscano nel tempo e nello spazio, osservando, contemporaneamente, l'evoluzione degli strumenti di letalità, la scena del crimine e i moventi. Questa dashboard si differenzia dalle precedenti anche per la presenza di un filtro temporale interattivo, attraverso il quale l'utente può modificare l'intervallo cronologico considerato e propagare tale selezione alle visualizzazioni presenti nella pagina.

La composizione complessiva della dashboard è riportata nella Figura 5.3.

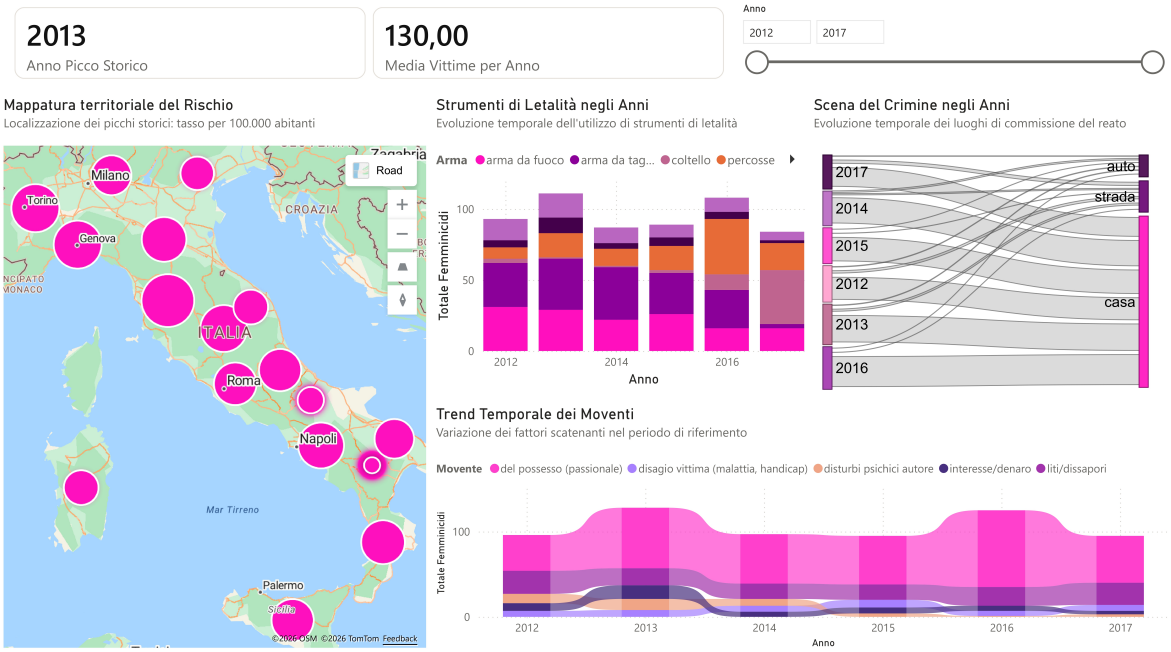


Figura 5.3: Dashboard 7: "Evoluzione temporale dei trend critici"

I principali elementi di visualizzazione presenti nella dashboard, che verranno analizzati singolarmente nelle sezioni successive, sono:

- *indicatori KPI*;
- *filtro dei dati*;
- *mappa di Azure*;
- *grafico colonne in pila*;
- *grafico Sankey*;
- *grafico a nastri*.

Indicatori KPI

Nella parte superiore della dashboard sono presenti due indicatori *KPI* finalizzati a fornire una sintesi dell'andamento temporale del fenomeno nel periodo considerato. I due indicatori KPI considerati sono:

- *Anno Picco Storico (2013)*: identifica l'anno in cui è stato registrato il valore massimo di vittime nel periodo temporale analizzato.
- *Media Vittime Annue (130)*: rappresenta il numero medio di vittime registrato annualmente nell'intervallo temporale considerato e fornisce un valore di riferimento per la lettura dell'andamento complessivo del fenomeno.

La lettura congiunta dei due indicatori permette, quindi, di affiancare all'individuazione dell'anno caratterizzato dal valore massimo una misura sintetica della numerosità media osservata nel periodo. Il primo KPI consente di localizzare temporalmente il picco, mentre il secondo fornisce un riferimento quantitativo rispetto al quale interpretare l'andamento delle vittime nei diversi anni considerati.

Filtro dei dati

Sul lato destro della dashboard è presente uno slider temporale che consente di selezionare l'intervallo di anni sul quale concentrare l'analisi. Il periodo disponibile comprende gli anni dal 2012 al 2017 e permette all'utente di restringere o ampliare dinamicamente l'intervallo temporale visualizzato. Il filtro non agisce su una singola visualizzazione, ma viene applicato alla vista complessiva della dashboard, propagando la selezione temporale ai diversi elementi presenti. La modifica dell'intervallo determina, pertanto, l'aggiornamento dei valori rappresentati nei grafici e degli indicatori dipendenti dal periodo selezionato. Questa funzionalità consente di confrontare differenti porzioni del periodo analizzato e di osservare come la distribuzione del fenomeno vari al variare dell'intervallo temporale considerato.

Mappa di Azure: "Mappatura Territoriale del Rischio"

Nel quadrante sinistro della dashboard è stata implementata una rappresentazione geospaziale avanzata mediante Azure Maps, denominata *Mappatura territoriale del Rischio*. La visualizzazione introduce una modalità di analisi territoriale complementare rispetto alle rappresentazioni precedentemente utilizzate, basate principalmente sulla distribuzione dei valori assoluti. In questo caso, il fenomeno viene rappresentato attraverso indicatori circolari, o *bubble layer*, posizionati in corrispondenza delle aree geografiche considerate. La dimensione delle bolle varia in funzione del valore della metrica associata alla specifica localizzazione, consentendo di cogliere immediatamente le differenze nella distribuzione territoriale del fenomeno.

L'elemento caratterizzante della visualizzazione è rappresentato dall'impiego di una misura normalizzata rispetto alla popolazione residente. Le dimensioni degli indicatori non esprimono, infatti, il semplice numero assoluto di eventi, ma sono associate all'incidenza del fenomeno, calcolata come tasso per 100.000 abitanti. La costruzione di tale indicatore è resa possibile dall'integrazione nel modello dei dati di una *Dimension Table* contenente i dati demografici ISTAT, collegata alla tabella dei fatti e utilizzata per rapportare il numero degli eventi alla popolazione del territorio di riferimento.

L'utilizzo dell'incidenza modifica, quindi, la prospettiva interpretativa rispetto a una rappresentazione basata esclusivamente sui valori assoluti. Un territorio caratterizzato da una popolazione numericamente elevata può presentare un numero complessivo di eventi

maggiore senza necessariamente registrare, in termini relativi, l'incidenza più elevata. Analogamente, territori con una popolazione inferiore possono assumere maggiore rilevanza nella rappresentazione quando il numero degli eventi viene rapportato alla rispettiva popolazione residente. Tale distinzione risulta particolarmente evidente nel confronto con le dashboard precedenti: la Lombardia, che nelle rappresentazioni basate sui valori assoluti risultava particolarmente evidente per la numerosità complessiva dei casi, può assumere una dimensione visiva differente quando la metrica considerata è normalizzata rispetto alla popolazione. Questo passaggio consente, pertanto, di distinguere due prospettive complementari: da un lato, la numerosità assoluta, utile per descrivere la dimensione complessiva del fenomeno in un territorio; dall'altro, l'incidenza, che permette di rapportare tale numerosità alla popolazione residente e di rendere più confrontabili territori caratterizzati da dimensioni demografiche differenti. L'incidenza non deve essere interpretata come una misura diretta e assoluta del "rischio" individuale, ma come un indicatore statistico della frequenza del fenomeno in rapporto alla popolazione considerata.

La mappa sfrutta, inoltre, le funzionalità interattive proprie dell'ambiente Power BI e di Azure Maps. Il posizionamento del cursore sulle singole bolle attiva il relativo *tooltip*, attraverso il quale vengono visualizzate informazioni aggiuntive associate all'area selezionata. Tra queste rientra, ad esempio, l'indicazione dell'"Anno Picco Storico" registrato per il territorio considerato. L'interazione permette, così, di affiancare alla lettura sintetica offerta dalla dimensione delle bolle un livello informativo più dettagliato, senza sovraccaricare la rappresentazione principale. L'introduzione di questa visualizzazione amplia, quindi, la prospettiva dell'analisi territoriale, passando dalla semplice individuazione dei territori con il maggior numero di casi a un confronto che tiene conto anche della dimensione demografica. La mappa permette, in questo modo, di osservare come la distribuzione del fenomeno cambi quando i valori vengono normalizzati rispetto alla popolazione residente, fornendo una lettura complementare a quella basata sui conteggi assoluti. L'integrazione tra componente geografica, indicatore normalizzato e funzionalità interattive rende, pertanto, la visualizzazione uno strumento utile per approfondire le differenze territoriali presenti nel *dataset*, mantenendo distinte la dimensione quantitativa complessiva e quella relativa alla popolazione di riferimento.

Grafico a colonne in pila: "Strumenti di Letalità negli Anni"

Nella parte centrale della dashboard è presente un grafico a colonne in pila dedicato all'analisi dell'evoluzione temporale degli strumenti e delle modalità di letalità. L'asse orizzontale rappresenta la dimensione temporale, mentre l'altezza complessiva di ciascuna colonna indica il numero totale di eventi associati all'anno considerato. Ogni colonna è suddivisa in segmenti corrispondenti alle differenti modalità di letalità registrate nel *dataset*, tra cui "arma da fuoco", "arma da taglio", "coltello" e "percosse".ù

La rappresentazione in pila consente di combinare due livelli di lettura all'interno della stessa visualizzazione. L'altezza complessiva delle colonne permette di confrontare la numerosità degli eventi nei diversi anni, mentre la suddivisione interna consente di osservare come tale numerosità sia composta rispetto alle differenti modalità di letalità. In questo modo, il grafico non permette soltanto di individuare gli anni caratterizzati dai valori complessivamente più elevati, ma anche di analizzare come varia nel tempo la distribuzione degli strumenti registrati.

Un primo elemento evidenziato dalla visualizzazione riguarda il 2013, che presenta la colonna complessivamente più alta e si conferma, quindi, come l'anno con la maggiore numerosità di eventi nel periodo rappresentato. L'osservazione della composizione delle singole colonne permette, inoltre, di individuare differenze nella distribuzione delle modalità di letalità tra le diverse annualità. In particolare, il segmento relativo all'"arma da fuoco" mostra

una riduzione della propria numerosità nel corso del periodo osservato. A partire dal valore registrato nel 2012, la componente associata a questa modalità si riduce progressivamente, raggiungendo valori più contenuti negli ultimi due anni della serie, il 2016 e il 2017. Tale andamento descrive una variazione nella composizione degli eventi registrati nel tempo, senza consentire, da solo, di individuare le cause alla base di tale variazione. Un andamento differente si osserva per la modalità "percosse", la cui rappresentazione mostra un incremento particolarmente evidente nel 2016. In corrispondenza di tale annualità, il relativo segmento assume una dimensione maggiore rispetto agli anni precedenti, contribuendo, in misura più consistente, alla composizione complessiva degli eventi. Il dato permette, quindi, di evidenziare una variazione temporale nella distribuzione delle modalità di letalità considerate. Particolarmente interessante risulta, infine, la distribuzione osservata nel 2017 per le modalità "coltello" e "arma da taglio". Nell'ultimo anno della serie, la componente relativa a "coltello" presenta un aumento evidente, accompagnato da una riduzione della componente associata ad "arma da taglio". Tuttavia, il grafico consente di evidenziare la variazione nella rappresentazione delle due modalità, ma non permette di stabilire autonomamente quale sia il fattore responsabile di tale cambiamento.

Grafico Sankey: "Scena del Crimine negli Anni"

Nella parte superiore destra della dashboard è collocato un diagramma di Sankey, denominato *Scena del Crimine negli Anni*, introdotto con l'obiettivo di analizzare congiuntamente la dimensione temporale e quella relativa al luogo in cui si verifica l'evento. La scelta di questa tipologia di visualizzazione consente, infatti, di rappresentare in modo immediato non soltanto la frequenza delle diverse scene del crimine, ma anche la loro distribuzione nel corso degli anni considerati.

I nodi collocati sul lato sinistro del diagramma rappresentano le annualità comprese tra il 2012 e il 2017, mentre quelli disposti sul lato destro identificano le differenti tipologie di scena del crimine presenti nel *dataset*, tra cui "casa", "strada" e "auto". I collegamenti tra i due insiemi di nodi rappresentano la relazione tra ciascun anno e la corrispondente scena dell'evento. Lo spessore dei flussi è proporzionale al numero di casi associati a ciascuna combinazione; quanto maggiore è lo spessore del collegamento, tanto più elevata risulta la frequenza osservata per quella specifica coppia anno-scena.

La visualizzazione permette, quindi, di osservare simultaneamente l'evoluzione temporale delle diverse scene del crimine e il loro peso relativo all'interno dei singoli anni. Un elemento particolarmente evidente è la presenza costante della voce "casa", che mantiene flussi consistenti lungo l'intero periodo analizzato. Tale risultato si pone in continuità con quanto osservato nelle dashboard precedenti, nelle quali l'ambiente domestico emerge ripetutamente come contesto maggiormente rappresentato di localizzazione degli eventi.

La sua ricorrenza nel tempo suggerisce, pertanto, che la prevalenza della scena domestica non sia riconducibile esclusivamente a una specifica annualità, ma costituisca una caratteristica ricorrente del *dataset* analizzato. Nel complesso, la visualizzazione rafforza, quindi, quanto già emerso dalle analisi precedenti, evidenziando la persistenza della scena domestica nel periodo considerato e consentendo di osservarne la distribuzione rispetto alle diverse annualità.

Grafico a nastri: "Trend Temporale dei Moventi"

Nella parte inferiore destra della dashboard è collocato un grafico a nastri, denominato *Trend Temporale dei Moventi*. La visualizzazione è stata scelta per analizzare l'evoluzione nel tempo della posizione occupata dai diversi moventi all'interno della graduatoria annuale, permettendo di osservare non soltanto la frequenza delle singole voci, ma anche le variazioni

del relativo posizionamento da un anno all'altro. A differenza di un tradizionale grafico a linee, il grafico a nastri pone l'attenzione sul *rank*, ossia sulla posizione relativa di ciascun movente rispetto agli altri presenti nella stessa annualità. L'andamento dei nastri consente, quindi, di seguire visivamente gli spostamenti nella graduatoria: un nastro che mantiene la propria posizione indica una relativa stabilità del posizionamento, mentre gli incroci e gli spostamenti tra i diversi nastri evidenziano variazioni nel ranking.

Nel periodo compreso tra il 2012 e il 2017, il movente del "possesso (passionale)" si colloca costantemente al primo posto della graduatoria rappresentata. Il relativo nastro mantiene, infatti, la posizione superiore lungo l'intero intervallo temporale, evidenziando una presenza quantitativamente maggiore rispetto agli altri moventi in ciascuna delle annualità considerate.

Anche il movente "lite" mostra un posizionamento particolarmente stabile, mantenendosi generalmente nelle posizioni più alte della graduatoria e, in particolare, in seconda posizione per la maggior parte del periodo osservato.

Diversamente, le altre tipologie di movente presentano una maggiore variabilità nel corso degli anni, alternandosi tra le posizioni successive della classifica. Gli spostamenti dei nastri rendono, pertanto, visibili le variazioni del ranking annuale, mostrando come, a fronte della sostanziale stabilità delle prime due posizioni, le restanti motivazioni assumano un peso relativo differente a seconda dell'anno considerato. Il grafico permette, così, di distinguere due livelli di lettura complementari: da un lato, la frequenza con cui i diversi moventi compaiono nei singoli anni e, dall'altro, la loro posizione relativa all'interno della graduatoria. La permanenza del "possesso (passionale)" al primo posto costituisce l'elemento di maggiore continuità evidenziato dalla visualizzazione, mentre l'alternanza delle altre voci mette in luce una distribuzione temporale più variabile.

Tali risultati descrivono il comportamento delle variabili all'interno del *dataset* analizzato e consentono di osservare l'evoluzione della loro posizione relativa, senza implicare necessariamente variazioni nelle cause sottostanti al fenomeno.

5.3 Considerazioni sulle dashboard avanzate

Considerate nel loro insieme, le tre dashboard avanzate definiscono un percorso di analisi volto ad ampliare e approfondire il livello di indagine sviluppato nelle precedenti fasi del progetto. Se le dashboard iniziali erano principalmente orientate alla descrizione quantitativa e alla rappresentazione delle principali caratteristiche del fenomeno, le dashboard avanzate introducono una prospettiva maggiormente esplorativa, finalizzata a mettere in relazione differenti dimensioni del *dataset* e a individuare ricorrenze e variazioni nella loro distribuzione.

5.3.1 Sintesi del percorso di analisi

Le tre dashboard sono state progettate secondo prospettive complementari, così da scomporre progressivamente la complessità del fenomeno e consentirne una lettura da differenti punti di vista.

La prima dashboard concentra l'attenzione sulle dinamiche relazionali ed esecutive, mettendo in relazione il legame tra autore del reato e vittima con le modalità di letalità e con il movente. Attraverso le diverse visualizzazioni è, quindi, possibile osservare come le tipologie di relazione si distribuiscano rispetto agli strumenti utilizzati e alle motivazioni registrate, nonché approfondire specifiche combinazioni mediante la scomposizione gerarchica dei dati.

La seconda dashboard sposta il livello di osservazione sulle caratteristiche comportamentali dell'autore del reato e sulla presenza, all'interno del *dataset*, di informazioni relative a precedenti disturbi psicologici o comportamenti suicidari. L'analisi considera tali fattori

in relazione al legame tra autore del reato e vittima e ai diversi moventi, permettendo di evidenziare le combinazioni maggiormente ricorrenti. In questo caso, l'obiettivo non è stabilire relazioni causali tra le variabili considerate, ma descriverne la distribuzione congiunta e mettere in evidenza le configurazioni che presentano una maggiore frequenza nei dati analizzati.

La terza dashboard introduce, infine, una prospettiva prevalentemente temporale e territoriale. Le visualizzazioni consentono di seguire l'andamento annuale delle principali modalità di letalità e dei moventi, osservandone sia la frequenza sia la posizione relativa nella graduatoria dei singoli anni. A questa dimensione viene affiancata l'analisi territoriale mediante la rappresentazione cartografica dell'incidenza, calcolata rispetto alla popolazione residente, così da integrare il semplice conteggio degli eventi con una misura normalizzata rispetto alla dimensione demografica delle aree considerate. Il percorso complessivo passa, quindi, dall'analisi delle relazioni tra le variabili alla loro distribuzione comportamentale e, infine, alla loro evoluzione temporale e territoriale.

5.3.2 Principali evidenze emerse dall'analisi

L'impiego delle visualizzazioni avanzate ha consentito di mettere in evidenza alcune ricorrenze che attraversano le diverse prospettive di analisi. Un primo elemento riguarda la rilevanza della sfera relazionale intima. Nella prima dashboard, la voce "coniuge" presenta una frequenza particolarmente elevata all'interno di diverse modalità di letalità, risultando una delle principali componenti delle distribuzioni osservate. Tale evidenza si inserisce in un quadro più ampio già emerso nelle dashboard precedenti, nel quale le relazioni di prossimità assumono un peso rilevante nella composizione dei casi analizzati.

Anche l'analisi dei fattori comportamentali evidenzia alcune configurazioni ricorrenti. In particolare, la voce "Suicidio Aggressore" presenta una frequenza elevata in corrispondenza della relazione "coniuge", mentre altre combinazioni risultano maggiormente distribuite tra le diverse tipologie di legame.

Questi risultati descrivono la frequenza con cui determinate caratteristiche compaiono congiuntamente nel *dataset*, ma non consentono, da soli, di stabilire un rapporto causale tra il comportamento dell'autore del reato e la tipologia di relazione.

Un'ulteriore evidenza riguarda la relazione tra i moventi e la presenza di precedenti disturbi psicologici registrati. Il movente del "possesso (passionale)" rappresenta la voce quantitativamente più frequente nella graduatoria dei moventi e, all'interno dei casi associati a tale motivazione, prevalgono quelli per i quali non risulta registrato nel *dataset* un precedente disturbo psicologico. Questo risultato deve essere interpretato in relazione alla natura della variabile considerata: la voce "no" indica l'assenza di un disturbo psicologico precedentemente registrato nella fonte dati e non consente di affermare che il soggetto non presentasse alcuna condizione psicologica. L'analisi evidenzia, pertanto, una distribuzione descrittiva delle informazioni disponibili, senza permettere di stabilire un'associazione causale tra condizioni psicologiche e movente.

La prospettiva temporale consente, inoltre, di individuare alcuni elementi di continuità nel periodo compreso tra il 2012 e il 2017. La scena "casa" mantiene una presenza costante e quantitativamente rilevante nel corso degli anni, confermando una caratteristica già osservata nelle precedenti dashboard. Tale evidenza risulta coerente con la significativa presenza di casi riconducibili alla sfera intima, che rappresentano circa il 75% dei casi analizzati, pur trattandosi di due dimensioni distinte che non possono essere considerate direttamente equivalenti.

Analogamente, il movente del "possesso (passionale)" mantiene la prima posizione nella graduatoria annuale dei moventi lungo l'intero periodo osservato, mentre le altre voci mostrano una maggiore variabilità nel loro posizionamento da un anno all'altro. Nel complesso,

le evidenze emerse non identificano singoli fattori come cause del fenomeno, ma permettono di riconoscere alcune configurazioni ricorrenti all'interno dei dati.

La possibilità di osservare contemporaneamente dimensioni relazionali, comportamentali, temporali e territoriali consente, dunque, di arricchire la lettura descrittiva iniziale, mettendo in luce aspetti che difficilmente sarebero emersi attraverso la sola analisi dei conteggi aggregati.

5.3.3 Valore della soluzione di Business Intelligence

La realizzazione delle dashboard avanzate evidenzia il contributo delle metodologie di *Business Intelligence* all'esplorazione di un fenomeno caratterizzato dalla presenza di numerose dimensioni interconnesse. Rispetto a una rappresentazione esclusivamente statistica dei dati, l'integrazione di strumenti quali gli alberi di scomposizione, le mappe ad albero, le matrici a doppia entrata, i diagrammi di flusso, i grafici a nastri e le rappresentazioni cartografiche consente di analizzare le informazioni secondo prospettive differenti e complementari. Il principale valore aggiunto della soluzione risiede nella possibilità di passare dalla semplice osservazione di un valore aggregato all'esplorazione delle modalità attraverso cui tale valore si distribuisce tra differenti combinazioni di variabili. Un determinato numero complessivo di eventi può, infatti, essere ulteriormente analizzato in funzione della relazione tra autore del reato e vittima, della modalità di letalità, del movente, dei fattori comportamentali, dell'anno o dell'area geografica considerata. In questo modo, il modello sviluppato non si limita a restituire una fotografia dei dati storici, ma costruisce un ambiente interattivo nel quale le diverse dimensioni possono essere analizzate congiuntamente. Le dashboard avanzate assumono, quindi, una funzione prevalentemente esplorativa e conoscitiva. La loro utilità risiede nella capacità di rendere più leggibili distribuzioni, ricorrenze e variazioni presenti nel *dataset*, fornendo una base strutturata per ulteriori approfondimenti.

Le evidenze ottenute devono essere interpretate considerando i limiti della fonte dati e la natura descrittiva delle analisi: le visualizzazioni consentono di individuare configurazioni ricorrenti e relazioni osservate, ma non sono sufficienti, da sole, a dimostrare rapporti causali o a identificare fattori determinanti del fenomeno. In questa prospettiva, la soluzione di *Business Intelligence* costituisce uno strumento di supporto all'analisi e all'interpretazione dei dati, capace di integrare le diverse dimensioni del fenomeno all'interno di un unico ambiente interattivo e di facilitare l'individuazione degli elementi che possono essere successivamente approfonditi mediante analisi statistiche o criminologiche più specifiche.

CAPITOLO 6

Discussione

Il presente capitolo conclude il percorso di progettazione e sviluppo dell'ambiente di Business Intelligence realizzato nel corso della tesi. Per fornire una sintesi critica del lavoro svolto, l'esposizione si articolerà in quattro sezioni fondamentali. Inizialmente, verranno discusse le principali evidenze emerse dall'analisi condotta attraverso le dashboard di base e le dashboard avanzate. Si ripercorreranno i principali pattern che attraversano i dati, analizzando la centralità della sfera intima, la prevalenza del contesto domestico, l'interazione tra il movente del possesso e i fattori psicologici, le dinamiche successive all'evento e la rilettura del rischio territoriale attraverso il calcolo dell'incidenza. Successivamente, l'attenzione si sposterà sul valore metodologico della soluzione proposta. Verrà analizzato il contributo offerto dall'adozione di un modello dimensionale interrogabile, capace di mettere in relazione dati di natura differente, superando i limiti dell'analisi puramente aggregata e permettendo un'interpretazione dinamica delle informazioni. Verranno, quindi, delineati i limiti del presente studio. Saranno affrontati i vincoli legati alla natura e alla qualità del dataset originario, i limiti temporali dell'analisi e le attenzioni necessarie nell'interpretazione statistica delle variabili puramente descrittive. A partire da tali limiti, verranno tracciate le prospettive di sviluppo futuro, con l'obiettivo di orientare l'attuale modello verso un sistema di supporto alle decisioni più avanzato.

6.1 Discussione delle principali evidenze emerse

Il lavoro ha previsto un percorso progressivo articolato in due livelli di analisi. Le dashboard di base sono state progettate principalmente con finalità descrittive e hanno consentito di delineare le caratteristiche generali del fenomeno attraverso indicatori quantitativi, distribuzioni demografiche, modalità esecutive, relazioni tra vittima e autore del reato e distribuzioni territoriali. Successivamente, le dashboard avanzate hanno approfondito alcune delle dimensioni emerse nella prima fase, introducendo visualizzazioni e modalità di interazione maggiormente orientate all'analisi delle relazioni tra variabili. Di conseguenza, le dashboard di base hanno fornito una prima rappresentazione delle distribuzioni principali, mentre le dashboard avanzate hanno consentito di approfondire alcune relazioni tra dimensioni differenti.

Le evidenze più rilevanti riguardano:

- la centralità della sfera intima;
- la prevalenza del contesto domestico;
- il movente del possesso e la lettura dei fattori psicologici;

- le dinamiche successive all'evento;
- la normalizzazione del dato territoriale.

6.1.1 La centralità della sfera intima

Una delle principali evidenze emerse dall'analisi riguarda la frequente presenza di un legame di prossimità tra autore del reato e vittima. Tale caratteristica emerge già nella prima dashboard *Modus operandi e dinamiche del delitto* attraverso il grafico *Legame relazionale*, dal quale si osserva che poco più del 90% dei legami considerati è riconducibile alla sfera intima. Questo primo risultato evidenzia il ruolo della dimensione relazionale all'interno del campione e costituisce il punto di partenza per l'approfondimento sviluppato nella dashboard avanzata *Analisi multidimensionale delle dinamiche relazionali ed esecutive*. In quest'ultima, la dimensione relazionale viene analizzata attraverso una misura specifica, rappresentata dal tachimetro *Incidenza Sfera Intima*. Il valore restituito, pari al 75,51%, corrisponde alla quota degli eventi considerati che presenta una relazione riconducibile alla sfera intima. La differenza rispetto alla percentuale osservata nel grafico della dashboard di base è dovuta alla diversa unità di analisi adottata dalle due visualizzazioni: il grafico *Legame relazionale* considera la distribuzione dei legami autore-vittima, mentre il grafico *Incidenza Sfera Intima* misura l'incidenza della sfera intima sul totale degli eventi. La distinzione tra i due indicatori consente, quindi, di leggere il medesimo fenomeno da prospettive complementari.

La prevalenza della sfera intima emerge anche dalla dimensione temporale. Difatti, nel grafico *Trend Storico dei Contesti Relazionali*, presente nella stessa dashboard avanzata, la componente relativa alla sfera intima risulta superiore rispetto a quella della sfera extrafamiliare in ciascuna delle annualità considerate. Tale dato evidenzia, dunque, che la dimensione relazionale intima costituisce una componente ricorrente lungo l'intero periodo analizzato e non risulta circoscritta a una specifica annualità. All'interno della sfera relazionale, nel grafico *Scomposizione del Modus Operandi*, la categoria "coniuge" rappresenta una delle componenti numericamente più rilevanti. Nella dashboard avanzata sono associati a tale categoria 307 eventi, mentre nella distribuzione considerata nella prima dashboard essa rappresenta il 45,77% degli eventi relativi alla variabile analizzata.

Nel complesso, l'analisi evidenzia, quindi, il contributo dell'approccio multidimensionale adottato. La distribuzione della variabile relativa al legame autore-vittima consente di individuare le categorie maggiormente rappresentate; l'analisi temporale permette di osservarne la ricorrenza nel periodo considerato; infine, l'incrocio con le modalità di letalità consente di approfondire come le diverse caratteristiche si combinino all'interno degli eventi analizzati. In questo modo, le dashboard avanzate non si limitano a mostrare le frequenze delle singole variabili, ma permettono di esaminare le loro distribuzioni congiunte, offrendo una lettura più articolata del fenomeno.

6.1.2 La prevalenza del contesto domestico

La dimensione relazionale si collega a un ulteriore aspetto da considerare nell'analisi della scena dell'evento. Nelle dashboard di base *Modus operandi e dinamiche del delitto*, la categoria "casa", rappresentata dal 78,45%, risulta la categoria più frequente nella distribuzione relativa al luogo in cui si verifica l'evento. La dashboard avanzata *Evoluzione temporale dei trend critici* ha permesso di approfondire tale caratteristica attraverso il diagramma di Sankey *Scena del Crimine negli Anni*. La visualizzazione mette in relazione l'anno dell'evento con la scena nella quale esso si verifica, consentendo di osservare la distribuzione dei flussi lungo l'intero periodo considerato, compreso tra il 2012 e il 2017. La presenza di un flusso rilevante verso la categoria "casa" nelle diverse annualità suggerisce che la presenza del contesto

domestico non sia limitata a uno specifico anno del campione. È, tuttavia, opportuno evitare di interpretare tale osservazione come una dimostrazione di stabilità statistica nel tempo, poiché il numero di osservazioni disponibili e la struttura del *dataset* non consentono di formulare conclusioni causali.

Il dato assume maggiore significato se considerato in relazione a quanto osservato nella precedente analisi della dimensione relazionale. Da un lato, infatti, la prevalenza della sfera intima evidenzia la frequente presenza di un legame di prossimità tra autore del reato e vittima; dall'altro, l'elevata frequenza della categoria "casa" individua proprio l'ambiente domestico come principale contesto di localizzazione degli eventi. Le due evidenze descrivono dimensioni differenti del fenomeno e non consentono, prese singolarmente, di stabilire una relazione causale tra legame relazionale e luogo dell'evento. La loro presenza congiunta, tuttavia, offre una chiave di lettura coerente del campione analizzato, nel quale risultano ricorrenti sia la dimensione relazionale sia la localizzazione degli eventi in ambito domestico. L'incrocio tra dimensione relazionale e dimensione spaziale consente, quindi, una lettura più completa dei risultati. Nel campione analizzato, una parte consistente degli eventi presenta contemporaneamente caratteristiche riconducibili alla prossimità relazionale e al contesto domestico. Questa convergenza non permette di affermare che l'una costituisca la causa dell'altra, ma evidenzia come le due dimensioni rappresentino, insieme, componenti ricorrenti nella descrizione degli eventi considerati. Considerate nel loro insieme, le dashboard consentono, pertanto, di passare dalla semplice osservazione della frequenza delle singole categorie a una rappresentazione più articolata delle caratteristiche che ricorrono all'interno del campione.

6.1.3 Il movente del possesso e la lettura dei fattori psicologici

Un'ulteriore evidenza riguarda la distribuzione dei moventi associati agli eventi. Nella dashboard di base *Modus operandi e dinamiche del delitto*, la categoria del "possesso (passionale)" presenta 373 eventi, risultando la categoria maggiormente rappresentata tra quelle disponibili. La sua prevalenza assume particolare rilevanza nella discussione del fenomeno, poiché richiama una dimensione relazionale nella quale la vittima viene percepita non semplicemente come soggetto della relazione, ma come oggetto sul quale l'autore del reato pretende di esercitare controllo.

Questa evidenza può essere messa in relazione con gli studi sociologici e criminologici sulla violenza di genere, nella quale il controllo, la subordinazione e l'asimmetria nelle relazioni sono frequentemente ricondotti anche alla persistenza di modelli culturali e sociali di matrice patriarcale. In tale prospettiva, la violenza non viene considerata esclusivamente come il risultato di una condizione individuale o di un impulso momentaneo, ma può essere interpretata anche all'interno di più ampi sistemi di rappresentazioni dei ruoli di genere e delle relazioni di potere.

Il dato esaminato nel *dataset* non consente, naturalmente, di dimostrare la presenza di una struttura patriarcale alla base dei singoli eventi, né di stabilire un rapporto causale tra tale struttura e il delitto; tuttavia, la frequenza della categoria del "possesso (passionale)" rende utile considerare anche la dimensione del controllo e della gestione della relazione.

La distribuzione temporale, analizzata nella dashboard avanzata *Evoluzione temporale dei trend critici* mediante il grafico a nastri *Trend Temporale dei Moventi*, consente di osservare l'andamento relativo delle principali categorie nel periodo considerato. In particolare, la categoria del "possesso (passionale)" si mantiene tra quelle più frequenti nelle diverse annualità analizzate. La presenza di questa categoria, nelle diverse annualità, ne evidenzia la ricorrenza nel periodo osservato all'interno dell'analisi, poiché non emerge come una caratteristica circoscritta a una singola annualità, ma come una componente ricorrente all'interno del campione. Anche in questo caso, il risultato deve essere interpretato come una descrizione della

distribuzione considerata e non come una misura dell'evoluzione generale del fenomeno nel tempo. L'interpretazione del movente diventa ancora più significativa quando viene affiancata all'analisi dei fattori psicologici registrati per l'autore del reato.

Un approfondimento particolarmente rilevante è stato ottenuto attraverso l'incrocio tra il movente e la presenza di disturbi psicologici. Nella dashboard avanzata *Analisi comportamentale, vulnerabilità e fattori di rischio*, il grafico a tornado *Impatto clinico sulle cause scatenanti* mostra 360 eventi relativi al movente del "possesso (passionale)" senza un disturbo psicologico registrato e 10 eventi nei quali tale informazione risulta presente. La differenza numerica tra le due categorie costituisce un ulteriore elemento da considerare rispetto a interpretazioni basate sul senso comune. Di fronte alla gravità degli eventi analizzati, può, infatti, risultare intuitivo ricondurre il comportamento dell'autore del reato a una condizione di instabilità o di disturbo psicologico, considerando la gravità dell'atto come conseguenza di un presunto problema individuale. La distribuzione osservata nel campione non conferma, tuttavia, questa associazione automatica: tra gli eventi classificati sotto il movente del "possesso (passionale)", la maggioranza non presenta una registrazione di disturbi psicologici nel campo considerato.

A questa interpretazione si affianca un'altra rappresentazione ricorrente nel senso comune, secondo cui un evento di tale gravità potrebbe essere attribuito a un'improvvisa perdita di controllo, spesso descritta attraverso il termine "raptus". Anche questo elemento trova nel *dataset* una presenza molto contenuta. La categoria "raptus" compare, infatti, in soli 4 eventi sui 780 complessivamente analizzati. Si tratta, quindi, di una classificazione marginale rispetto all'intero campione, che presenta, quindi, una frequenza limitata nella distribuzione dei moventi considerati.

Inoltre, il dato assume ulteriore significato se posto in relazione alla precedente osservazione relativa ai fattori psicologici. Sia l'associazione automatica tra violenza e disturbo psicologico, sia quella tra violenza e improvvisa perdita di controllo possono, infatti, condurre a un'analisi prevalentemente individuale dell'evento, concentrata sulle condizioni dell'autore e sulla presunta eccezionalità del comportamento. Le distribuzioni osservate nel campione invitano, invece, a una maggiore cautela: da un lato, la categoria del "possesso (passionale)" risulta prevalentemente associata all'assenza di un disturbo psicologico registrato; dall'altro, la categoria "raptus" riguarda soltanto una quota molto ridotta degli eventi analizzati. Questi risultati non consentono di escludere la presenza di condizioni psicologiche o dinamiche di perdita di controllo nei singoli eventi, ma suggeriscono che tali elementi non possano essere assunti automaticamente come spiegazione generale del fenomeno esaminato.

Tale evidenza risulta significativa perché consente di distinguere due piani interpretativi che nel dibattito sul fenomeno possono essere facilmente sovrapposti. Da un lato, possono esistere condizioni psicologiche individuali che contribuiscono alla complessità di specifici eventi; dall'altro, la presenza di un comportamento violento o caratterizzato da dinamiche di controllo non può essere automaticamente spiegata attraverso una condizione clinica o una perdita improvvisa di controllo. Nel campione analizzato, infatti, la categoria del "possesso (passionale)" risulta frequentemente associata all'assenza di un disturbo psicologico registrato, mentre la categoria "raptus" presenta una frequenza estremamente ridotta. Il dato rende, quindi, opportuno considerare che la violenza possa essere analizzata anche attraverso dimensioni relazionali, culturali e sociali, senza ricondurla, necessariamente, a una condizione clinica individuale.

Questa osservazione, inoltre, si collega alle precedenti analisi della sfera intima e del contesto domestico. La frequente presenza di relazioni di prossimità, la prevalenza della categoria "casa" come scena dell'evento e la rilevanza del movente del "possesso (passionale)" delineano, all'interno del campione, una ricorrenza di elementi riconducibili alla dimensione relazionale. Considerate congiuntamente, tali evidenze suggeriscono l'utilità di una lettura

che non si limiti alla dimensione individuale dell'autore, ma tenga conto anche delle dinamiche di relazione e di controllo all'interno delle quali gli eventi si collocano. Si tratta, tuttavia, di un'interpretazione coerente con i dati osservati e con gli studi richiamati nel lavoro, non di una dimostrazione causale del ruolo di tali dinamiche nei singoli eventi.

È importante sottolineare che l'assenza di una registrazione non equivale necessariamente all'assenza di una condizione clinica. La variabile rappresenta esclusivamente l'informazione disponibile nel *dataset* e può, pertanto, risentire della disponibilità, della completezza e delle modalità di raccolta delle informazioni. Di conseguenza, il risultato non permette di stabilire che gli autori classificati senza disturbi psicologici non presentassero alcuna condizione clinica, né consente di escludere condizioni non diagnosticate, non riportate o non disponibili nelle fonti utilizzate.

6.1.4 Le dinamiche successive all'evento

La dashboard avanzata *Analisi comportamentale, vulnerabilità e fattori di rischio* ha consentito di approfondire ulteriormente il fenomeno attraverso l'analisi dei comportamenti successivi all'evento. In particolare, la combinazione tra la mappa ad albero *Mappatura dei Fattori di Rischio in Relazione al Tipo di Legame*, relativa ai fattori comportamentali, e la matrice di relazione *Matrice di correlazione comportamentale e relazionale* ha permesso di osservare la distribuzione dei comportamenti successivi dell'autore in relazione al legame con la vittima. Nel campione, gli eventi associati al suicidio dell'autore sono 167, corrispondenti al 21,41% del totale degli eventi analizzati. La matrice evidenzia, inoltre, una concentrazione particolarmente rilevante nella categoria "coniuge", nella quale si registrano 89 eventi associati al suicidio dell'autore. Il valore risulta il più elevato tra le diverse categorie di relazione considerate, indicando che, all'interno del campione, il comportamento suicidario successivo all'evento presenta il valore più elevato negli eventi caratterizzati da una relazione coniugale.

Un ulteriore aspetto da considerare emerge considerando anche la categoria relativa al tentato suicidio. Anche in questo caso, la relazione "coniuge" presenta il valore più elevato, con 29 eventi. La presenza congiunta dei valori più alti per suicidio e tentato suicidio all'interno della stessa categoria relazionale introduce, quindi, un ulteriore elemento di riflessione. Non si tratta soltanto della maggiore frequenza di un singolo comportamento, ma della concentrazione, nella medesima tipologia di relazione, di due esiti differenti riconducibili alla dimensione comportamentale successiva all'evento.

Questa distribuzione può essere interpretata, con la necessaria cautela, considerando la particolare prossimità che caratterizza la relazione coniugale. Il rapporto tra coniugi implica, infatti, una dimensione affettiva, familiare e quotidiana particolarmente stretta, che può far sì che gli eventi caratterizzati da tale relazione siano associati a comportamenti successivi di diversa natura. La presenza dei valori più elevati per suicidio e tentato suicidio potrebbe, quindi, essere coerente con una maggiore concentrazione, nel campione, di dinamiche successive all'evento caratterizzate da una forte componente relazionale. Tale interpretazione non permette, tuttavia, di stabilire quali siano i meccanismi che determinano questi comportamenti, né di affermare che la relazione coniugale costituisca di per sé una causa del suicidio o del tentato suicidio.

Il risultato assume ulteriore significato se considerato alla luce delle evidenze già emerse nelle precedenti analisi. La categoria "coniuge" era, infatti, già risultata particolarmente rappresentata nella dimensione relazionale, con 307 eventi nella dashboard avanzata *Analisi multidimensionale delle dinamiche relazionali ed esecutive*, oltre a mostrare una presenza rilevante nelle combinazioni con differenti modalità di letalità. Il fatto che la stessa categoria presenti anche i valori più elevati per suicidio e tentato suicidio suggerisce che la dimensione coniugale costituisca, all'interno del campione, un elemento ricorrente in diverse variabili dell'analisi.

L'analisi congiunta di queste evidenze permette, perciò, di osservare la relazione coniugale non soltanto in termini di frequenza degli eventi, ma anche rispetto alle caratteristiche che accompagnano l'evento e alle dinamiche che possono manifestarsi successivamente.

In questo senso, l'analisi multidimensionale evidenzia come una stessa categoria relazionale possa comparire con frequenza elevata su più piani dell'analisi: nella distribuzione degli eventi, nelle modalità di letalità e nei comportamenti successivi dell'autore. La dashboard consente, così, di far emergere una ricorrenza che sarebbe meno immediata attraverso l'analisi separata delle singole variabili. Anche in questo caso, tuttavia, il risultato deve essere interpretato esclusivamente in termini descrittivi. L'elevata frequenza di suicidio e tentato suicidio nella categoria "coniuge" non consente di stabilire l'esistenza di un rapporto causale tra la tipologia di relazione e tali comportamenti. Inoltre, la frequenza osservata può essere influenzata dalla numerosità della stessa categoria all'interno del campione: una categoria più rappresentata può, naturalmente, presentare anche un numero assoluto maggiore di eventi associati.

Per valutare se la concentrazione osservata sia effettivamente superiore a quella attesa in funzione della numerosità delle diverse categorie relazionali sarebbe necessario ricorrere a misure relative e a specifiche metodologie statistiche. Nel complesso, questa analisi mostra, quindi, l'utilità dell'approccio multidimensionale adottato. La matrice non si limita a indicare quanti eventi di suicidio dell'autore del reato siano presenti nel campione, ma permette di osservare come tale comportamento, insieme al tentato suicidio, si distribuisca rispetto alla relazione con la vittima. L'emergere della categoria "coniuge", come quella con i valori più elevati per entrambi i comportamenti, rappresenta un'evidenza significativa per la descrizione del campione e costituisce, al tempo stesso, un possibile punto di partenza per approfondimenti successivi, nei quali l'informazione relazionale potrebbe essere integrata con variabili temporali, demografiche e comportamentali.

6.1.5 La normalizzazione del dato territoriale

Uno dei contributi metodologici più rilevanti dell'analisi riguarda la possibilità di affiancare ai valori assoluti una misura normalizzata rispetto alla popolazione residente. Nella dashboard di base *Analisi demografico-temporale e territoriale*, la distribuzione territoriale evidenzia un numero maggiore di eventi nei territori caratterizzati da una popolazione più numerosa. La Lombardia, ad esempio, presenta un numero elevato di eventi registrati. Analogamente, alcune delle province maggiormente rappresentate nel *dataset* corrispondono a importanti centri urbani, nei quali la maggiore numerosità della popolazione può contribuire a determinare valori assoluti più elevati. La valutazione dei soli valori assoluti presenta, tuttavia, un limite: territori caratterizzati da dimensioni demografiche differenti non possono essere confrontati direttamente attraverso il semplice numero di eventi. Un valore assoluto più elevato non indica necessariamente una maggiore incidenza del fenomeno rispetto alla popolazione residente, così come un numero inferiore di eventi non implica necessariamente una minore rilevanza se rapportato a una popolazione più contenuta.

Per questo motivo, nel modello è stata introdotta una *Dimension Table* alimentata attraverso i dati ISTAT, utilizzata per integrare la popolazione residente e calcolare l'incidenza degli eventi per 100.000 abitanti. L'integrazione dei dati ha permesso, quindi, di ampliare la prospettiva dell'analisi territoriale, passando da una rappresentazione basata sulla numerosità assoluta degli eventi a una misura che tiene conto della diversa dimensione demografica dei territori. Nella dashboard avanzata *Evoluzione temporale dei trend critici*, attraverso il grafico *Mappatura Territoriale del Rischio*, l'incidenza è stata rappresentata mediante una visualizzazione a bolle, consentendo di confrontare i territori non soltanto sulla base del numero complessivo di eventi, ma anche in relazione alla rispettiva popolazione residente.

Questa normalizzazione modifica la rappresentazione di alcune regioni. La Lombardia, che nella visualizzazione dei valori assoluti assume una posizione particolarmente rilevante, presenta, infatti, bolle di dimensioni inferiori nella mappa basata sull'incidenza. Tale riduzione è l'effetto diretto della normalizzazione rispetto alla popolazione residente; il numero assoluto viene, infatti, rapportato a una popolazione molto numerosa, producendo un valore di incidenza che può risultare relativamente più contenuto rispetto a quello di territori con una popolazione inferiore. La differenza tra le due rappresentazioni evidenzia quanto la scelta dell'indicatore possa influenzare la lettura del fenomeno territoriale.

L'introduzione dell'incidenza consente, in questo modo, di evitare che la sola dimensione demografica condizioni l'interpretazione della distribuzione geografica. Il confronto non si limita più a individuare i territori che presentano il maggior numero di eventi, ma permette di osservare anche quali territori presentino valori più elevati in rapporto alla propria popolazione residente. La normalizzazione permette, quindi, di rendere più significativo il confronto tra territori con diversa popolazione, senza trasformare l'indicatore in una misura causale o predittiva del fenomeno. Nel complesso, il passaggio dai valori assoluti all'incidenza evidenzia l'importanza dell'integrazione tra fonti informative differenti.

6.2 Il valore della soluzione di Business Intelligence

Il principale contributo metodologico del progetto risiede nella trasformazione di un insieme di dati eterogenei in un ambiente analitico strutturato, nel quale le diverse fasi di preparazione, modellazione, calcolo e visualizzazione risultano integrate all'interno di un unico ambiente. L'utilizzo di Power BI ha permesso, infatti, di definire un processo integrato che collega le diverse fasi, dall'acquisizione e trasformazione del dato fino alla sua rappresentazione e successiva esplorazione. La costruzione del modello dimensionale ha avuto un ruolo centrale in questo processo. La distinzione tra la *Fact Table*, contenente gli eventi analizzati, e la *Dimension Table* relativa ai dati demografici ISTAT ha consentito di integrare fonti differenti mantenendo una struttura coerente per le successive elaborazioni.

Questa organizzazione ha permesso di mettere in relazione informazioni di natura differente, rendendo possibile un'analisi territoriale più articolata attraverso il collegamento tra gli eventi e la popolazione residente. Anche l'impiego delle misure DAX ha contribuito a rendere il modello maggiormente dinamico e interrogabile. Indicatori come il numero complessivo degli eventi, il numero delle vittime, la percentuale di suicidio dell'autore e l'incidenza territoriale vengono calcolati direttamente sul modello e possono, quindi, aggiornarsi in funzione dei filtri e delle interazioni effettuate dall'utente. Tali misure non costituiscono, pertanto, valori staticamente associati alle singole visualizzazioni, ma rappresentano elementi del modello in grado di adattarsi alla selezione delle diverse dimensioni analizzate.

Dal punto di vista della *Data Visualization*, il progetto ha evidenziato, inoltre, come la scelta della rappresentazione possa influenzare il modo in cui l'informazione viene letta. Un semplice conteggio permette di evidenziare la numerosità di una categoria, mentre un'incidenza normalizzata consente di confrontare territori caratterizzati da dimensioni demografiche differenti; analogamente, una distribuzione univariata permette di osservare la frequenza di una singola variabile, mentre una matrice o un diagramma di Sankey consentono di esplorare le sue combinazioni con altre dimensioni. In questo senso, il contributo della soluzione di *Business Intelligence* emerge soprattutto nella possibilità di passare dalla semplice disponibilità dei dati alla loro interrogazione attraverso prospettive differenti. L'informazione non viene, dunque, considerata esclusivamente nella sua dimensione aggregata, ma può essere scomposta e ricombinata attraverso le diverse dimensioni presenti nel modello.

6.3 Limiti dello studio

Nonostante le evidenze emerse dall'analisi e il contributo offerto dal modello di *Business Intelligence*, i risultati devono essere interpretati alla luce di alcuni limiti metodologici e informativi. Tali limiti riguardano principalmente la natura e la struttura del *dataset*, l'estensione temporale del campione, il carattere prevalentemente descrittivo delle analisi e la qualità delle informazioni relative ad alcune variabili specifiche.

6.3.1 Limiti legati alla natura del dataset

Il primo limite riguarda la natura del *dataset* utilizzato. L'intera analisi dipende dalle variabili disponibili, dalla loro completezza e dalle modalità con cui gli eventi sono stati raccolti, classificati e registrati nella fonte originaria. La presenza di valori mancanti, categorie non uniformi o informazioni non disponibili può incidere sulla distribuzione delle variabili e, di conseguenza, sulla lettura dei risultati ottenuti. Nel corso della fase di preparazione dei dati, i valori nulli o mancanti sono stati ricondotti alla categoria "Dato non disponibile", così da distinguere esplicitamente l'assenza dell'informazione da una categoria effettivamente osservata. Questa scelta consente di preservare la presenza del valore mancante all'interno del modello, ma non permette, naturalmente, di ricostruire l'informazione originariamente non disponibile.

6.3.2 Limiti temporali

Un secondo limite riguarda l'orizzonte temporale e la dimensione del campione analizzato. L'analisi considera gli anni compresi tra il 2012 e il 2017, a seguito dell'esclusione degli eventi relativi al 2018 secondo i criteri definiti nella fase di preparazione dei dati. Il periodo disponibile permette di osservare alcuni *pattern* e distribuzioni tra le diverse annualità, ma non è sufficiente per descrivere l'evoluzione del fenomeno su un arco temporale più esteso.

6.3.3 Limiti dell'interpretazione statistica

Un ulteriore limite riguarda il carattere prevalentemente descrittivo delle analisi realizzate. Le dashboard consentono di individuare frequenze, distribuzioni, ricorrenze e combinazioni tra variabili, offrendo una sintesi articolata dei dati disponibili, ma la presenza di un'associazione o di una maggiore frequenza non è sufficiente, di per sé, a stabilire un rapporto causale. Ad esempio, una maggiore presenza di una determinata categoria occupazionale, di una specifica relazione personale o di una caratteristica psicologica all'interno del campione non implica che tale caratteristica costituisca un fattore di rischio per il verificarsi dell'evento. Per formulare un'interpretazione di questo tipo sarebbe necessario disporre di un appropriato gruppo di riferimento, di ulteriori variabili di controllo e di metodologie statistiche inferenziali in grado di valutare la significatività e l'eventuale associazione tra le variabili considerate.

6.3.4 Limiti nell'interpretazione delle variabili psicologiche

Particolare attenzione richiede, infine, l'interpretazione delle variabili relative alla presenza di disturbi psicologici e ai comportamenti dell'autore del reato. Si tratta di informazioni che possono presentare livelli di completezza differenti e la cui registrazione dipende dalla disponibilità dell'informazione nella fonte utilizzata. Di conseguenza, l'assenza di una determinata caratteristica nel *dataset* non equivale necessariamente alla sua assenza nella realtà, ma indica, più precisamente, che tale informazione non risulta registrata o disponibile

nel campo considerato. Questa distinzione assume particolare rilevanza nell'interpretazione dei risultati relativi al rapporto tra movente, disturbi psicologici e comportamenti successivi all'evento. Le frequenze osservate descrivono, infatti, la distribuzione delle informazioni presenti nel campione, ma non possono essere considerate una valutazione delle condizioni cliniche degli autori né permettono di stabilire un rapporto diretto tra una determinata condizione psicologica e il comportamento osservato. Per questo motivo, tali variabili sono state utilizzate nell'analisi come dimensioni descrittive del *dataset*, evitando di attribuire loro un significato diagnostico o causale che le informazioni disponibili non consentono di sostenere.

6.4 Sviluppi futuri

I risultati ottenuti aprono diverse possibilità di sviluppo del progetto, sia dal punto di vista informatico sia sotto il profilo analitico. Le principali prospettive riguardano l'ampliamento delle fonti, l'introduzione di metodologie statistiche e l'eventuale evoluzione del modello verso strumenti di analisi più avanzati.

6.4.1 Ampliamento e aggiornamento delle fonti

Un primo sviluppo potrebbe riguardare l'ampliamento del periodo temporale analizzato e l'integrazione di ulteriori fonti istituzionali. L'aggiornamento periodico dei dati permetterebbe di rendere il modello progressivamente aggiornabile, superando l'attuale impostazione orientata principalmente all'analisi di un campione storico. L'integrazione di ulteriori variabili territoriali, demografiche e socioeconomiche potrebbe contribuire, inoltre, a contestualizzare maggiormente le distribuzioni osservate, mantenendo sempre distinta l'osservazione di eventuali associazioni dall'interpretazione causale.

6.4.2 Analisi statistica

Un secondo sviluppo potrebbe consistere nell'affiancamento alle analisi descrittive di metodologie statistiche inferenziali. I test di associazione, modelli di regressione e altre tecniche appropriate potrebbero essere utilizzati per verificare statisticamente alcune delle relazioni individuate durante l'esplorazione delle dashboard. In questo modo, Power BI potrebbe continuare a svolgere una funzione di esplorazione e comunicazione dei dati, mentre le metodologie statistiche permetterebbero di sottoporre le evidenze emerse a ulteriori verifiche quantitative.

6.4.3 Machine Learning e analisi predittiva

Un'ulteriore prospettiva riguarda l'integrazione di tecniche di *Machine Learning*. Power BI può essere affiancato a linguaggi e ambienti dedicati all'analisi dei dati, come R e Python, rendendo possibile l'introduzione di tecniche quali *clustering* e classificazione. Il *clustering*, ad esempio, potrebbe consentire di individuare gruppi di eventi caratterizzati da combinazioni simili di variabili, mentre i modelli di classificazione potrebbero essere utilizzati per analizzare l'appartenenza a determinate categorie presenti nel *dataset*. L'introduzione di modelli predittivi richiederebbe, tuttavia, un'adeguata fase di validazione, selezione delle variabili e valutazione delle prestazioni, oltre alla gestione di eventuali problemi di sbilanciamento delle classi.

6.4.4 Evoluzione verso un sistema di supporto alle decisioni

Un ulteriore sviluppo potrebbe riguardare l'evoluzione dell'ambiente verso un sistema di supporto alle decisioni. In questa prospettiva, dashboard, indicatori territoriali, serie temporali, dati socio-demografici e risultati delle analisi statistiche potrebbero essere integrati in un ambiente aggiornabile e orientato all'esplorazione. L'obiettivo non sarebbe quello di sostituire il giudizio degli esperti o delle istituzioni, ma di fornire uno strumento aggiuntivo per organizzare e interpretare le informazioni disponibili, facilitando l'individuazione di configurazioni o territori che possano richiedere ulteriori approfondimenti.

Conclusioni

Nel presente lavoro di tesi abbiamo affrontato la progettazione e la realizzazione di un ambiente avanzato di *Business Intelligence* finalizzato all'analisi del fenomeno del femminicidio in Italia, mostrando come le metodologie di *Data Analytics* possano contribuire a trrasformare un insieme eterogeneo di dati grezzi in un supporto decisionale e conoscitivo concreto.

Nella prima fase del nostro percorso abbiamo esplorato i fondamenti teorici dell'elaborazione delle informazioni, analizzando il passaggio dai dati strutturati e non strutturati fino alla generazione di conoscenza secondo il paradigma DIKW, per poi calare tali concetti nelle architetture dei Big Data. Successivamente, abbiamo approfondito le caratteristiche e le potenzialità dell'ecosistema Microsoft Power BI, che è stato scelto come strumento principale per la sua capacità di coprire l'intero ciclo di vita del dato, dall'estrazione fino alla reportistica interattiva.

Una parte centrale del lavoro ha riguardato l'acquisizione di un *dataset* open source relativo ai femminicidi registrati in Italia tra il 2012 e il 2018, che abbiamo sottoposto a un processo di *Extract, Transform e Load* (ETL) tramite Power Query per risolverne le criticità strutturali, le ridondanze e i valori mancanti. Al fine di garantire un'interpretazione corretta delle dinamiche territoriali, abbiamo integrato i dati relativi agli eventi analizzati con una tabella dimensionale contenente i dati demografici ISTAT della popolazione residente, realizzando un modello di dati a stella ottimizzato dal motore in-memory VertiPaq.

Su questa base architetturale abbiamo sviluppato un set di metriche in linguaggio DAX e progettato un percorso esplorativo articolato in due livelli. Inizialmente, abbiamo realizzato delle dashboard di base con finalità puramente descrittive per quantificare le vittime, mappare gli scenari del crimine, profilare anagraficamente i soggetti coinvolti e individuare la preponderanza della sfera relazionale intima. Successivamente, per esaminare il fenomeno nelle sue diverse dimensioni e approfondire le dinamiche relazionali tra autore e vittima, abbiamo sviluppato delle dashboard avanzate impiegando strumenti visivi sofisticati, quali alberi di scomposizione, diagrammi di Sankey, mappe ad albero e visualizzazioni geografiche normalizzate con la mappa di Azure. L'analisi incrociata ci ha permesso di far emergere evidenze significative, come la stabilità temporale del contesto domestico quale scena primaria, la presenza rilevante dei gesti suicidari successivi al delitto all'interno delle relazioni coniugali e la frequente associazione del movente passionale alla mancata registrazione di disturbi psicologici, offrendo elementi utili per una lettura più articolata del fenomeno rispetto ad alcune rappresentazioni semplificate.

Pur riconoscendo i limiti intrinseci di un'analisi puramente descrittiva e le restrizioni legate alla completezza della fonte dei dati, il sistema implementato ha dimostrato il valore della *Data Visualization* nel facilitare l'identificazione di *pattern* spazio-temporali e

comportamentali.

A partire da questi risultati, i futuri sviluppi di questo lavoro si orientano verso un progressivo superamento dell'orizzonte puramente esplorativo, mirando all'ampliamento del periodo temporale analizzato e all'integrazione continua di nuove fonti istituzionali per consolidare le tendenze evidenziate. Un ulteriore sviluppo potrebbe consistere nell'affiancare alle attuali visualizzazioni descrittive l'applicazione di metodologie statistiche inferenziali, necessarie per validare quantitativamente la significatività delle relazioni osservate, come quelle tra dinamiche di letalità e profili occupazionali. In un'ottica di evoluzione tecnologica, l'integrazione nativa di Power BI con linguaggi di programmazione per la *Data Science*, quali R e Python, pone le basi per l'implementazione di algoritmi di *Machine Learning* e di analisi predittiva. L'impiego di tecniche di classificazione o *clustering* applicate ai dati storici consentirebbe, infatti, di individuare automaticamente ricorrenze, gruppi omogenei o possibili anomalie, evolvendo l'attuale sistema di analisi verso uno strumento di supporto alle decisioni. Una simile infrastruttura analitica non avrebbe lo scopo di sostituire l'interpretazione degli esperti, ma potrebbe fornire alle istituzioni e agli enti territoriali un ulteriore strumento a supporto della definizione e del monitoraggio delle politiche di prevenzione e della pianificazione di interventi mirati sul territorio.

Bibliografia

ABITEBOUL, S., BUNEMAN, P. e SUCIU, D. (2000), *Data on the Web: From Relations to Semistructured Data and XML*, Morgan Kaufmann.

ACKOFF, R. L. (1989), «From Data to Wisdom», *Journal of Applied Systems Analysis*.

BATESON, G. (1972), *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*, Chandler Publishing Company, San Francisco.

BELLUATI, M. (2013), «La cornice mediatica del femminicidio in Italia», *Problemi dell'informazione*.

BOURANY, T. (2019), «Les 5v du big data», .

CLEVELAND, H. (1982), «Information as a Resource», *The Futurist*.

CODD, E. F. (1970), «A Relational Model of Data for Large Shared Data Banks», *Communications of the ACM*.

COX, M. e ELLSWORTH, D. (1997), «Application-Controlled Demand Paging for Out-of-Core Visualization», Rap. Tecn. NAS-97-010, NASA Ames Research Center.

DEVOTO, G. e OLI, G. C. (2009), *Il Devoto-Oli. Vocabolario della lingua italiana*, Le Monnier, 2009 ed.

ELIOT, T. S. (1934), *The Rock: A Pageant Play*, Faber and Faber, London.

ELMASRI, R. e NAVATHE, S. B. (2015), *Fundamentals of Database Systems*, Pearson, 7th ed.

EUCLID (1993), *The Data of Euclid: Translated from the Text of Menge*, Union Square Press, Baltimore.

FLORIDI, L. (2010), *Information: A Very Short Introduction*, Oxford University Press, Oxford.

GANTZ, J. e REINSEL, D. (2012), «The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east», *IDC iView*.

GEORGE, G., HAAS, M. R. e PENTLAND, A. (2016a), «Big Data, Data Science, and Management Theory», *Academy of Management Journal*.

GEORGE, G., OSINGA, E. C., LAVIE, D. e SCOTT, B. A. (2016b), «Big Data and Data Science Methods for Management Research», *Academy of Management Journal*.

IANDIORIO, E. (2021), *Big data: cosa sono, come analizzarli e utilizzarli per fare marketing*, Flaccovio Dario.

LAGARDE, M. (2006), «Il femminicidio», in «Dal silenzio alla parola. La violenza contro le donne», Franco Angeli, Milano.

LANEY, D. (2001), «3-D Data Management: Controlling Data Volume, Velocity and Variety», Research Note 949, META Group.

MOORE, G. E. (1965), «Cramming More Components onto Integrated Circuits», *Electronics*, vol. 38.

PASQUALI, F. (2025), *Il fenomeno del femminicidio in Italia: tra dati e rappresentazioni*, Tesi di laurea magistrale, Università degli Studi di Padova, Dipartimento di Studi Linguistici e Letterari.

PETTY, W. (1691), *Treatise of Naval Philosophy*, London.

RIDI, R. (2019), *La piramide dell'informazione: una introduzione*, Editrice Bibliografica.

ROSENBERG, D. (2013), «Data before the fact», in «Raw Data is an Oxymoron», MIT Press.

RUSSELL, D. E. H. e RADFORD, J. (1992), *Femicide. The politics of woman killing*, Twayne Publishers, New York.

SALLAM, R. L., RICHARDSON, J., HAGERTY, J. e HOSTMANN, B. (2011), «Magic Quadrant for Business Intelligence Platforms», Research report, Gartner, Inc.

SENATO DELLA REPUBBLICA (2018), «Relazione della Commissione parlamentare di inchiesta sul femminicidio, nonché su ogni forma di violenza di genere», Rapporto istituzionale, Senato della Repubblica - XVIII Legislatura.

SHANNON, C. E. (1948), «A Mathematical Theory of Communication», *The Bell System Technical Journal*.

URSINO, D. (2026), «La Business Intelligence», Materiale didattico del corso di Project Management e Business Intelligence, Università Politecnica delle Marche, Anno Accademico 2025/2026.

VIOLI, P. (2015), «Femminicidio, una nuova emergenza?», .

ZELENY, M. (1987), «Management Support Systems: Towards Integrated Knowledge Management», *Human Systems Management*.

Sitografia

- ACM Digital Library, Association for Computing Machinery – dl.acm.org
- Analisi Predittiva in Power BI: cos'è e come funziona, Visualitics – visualitics.it/analisi-predittiva-power-bi/
- Business Intelligence Recap of 2010, Microsoft – www.microsoft.com/en-us/sql-server/blog/2010/12/28/business-intelligence-recap-of-2010/
- Cos'è Power BI?, Microsoft Learn – learn.microsoft.com/it-it/power-bi/fundamentals/power-bi-overview
- Encyclopaedia Britannica, voce "Data" – www.britannica.com/
- Femicides in Italy - counterdata RAI from 2012 to 2018, Zenodo – zenodo.org/records/17672139
- Gartner Magic Quadrant for Business Intelligence, Qlik – www.qlik.com/us/gartner-magic-quadrant-business-intelligence
- History of Philosophical Transactions, Royal Society – www.royalsociety.org/journals/publishing-activities/publishing350/history-philosophical-transactions/
- JSTOR Daily, approfondimento storico e etimologico sul dato – daily.jstor.org/data-not-just-another-four-letter-word/
- Microsoft Power BI: dai dati alle azioni, Beta 80 Group – news.beta80group.it/microsoft-power-bi-spiegato-semplice-dai-dati-alle-azioni/
- Oxford English Dictionary, voce "Big Data" – www.oed.com/
- Serie Storiche - Popolazione, ISTAT (Istituto Nazionale di Statistica) – seriestoriche.istat.it/#popolazione
- TechTarget, approfondimento storico e cronologia dei Big Data – www.techtarget.com/whatis/feature/A-history-and-timeline-of-big-data
- UTSA Math Research Wiki, portale di ricerca matematica – mathresearch.utsa.edu/wiki/index.php?title=Data

- Vocabolario Treccani online, voce "Femminicidio" – `www.treccani.it/vocabolario/femminicidio/`
- What powers Power BI in Office 365?, Microsoft – `www.microsoft.com/en-us/microsoft-365/blog/2013/07/08/what-powers-power-bi-in-office-365/`

Ringraziamenti

Desidero esprimere la mia più sincera gratitudine al mio relatore, il Prof. Domenico Ursino, per la dedizione, l'incredibile precisione e la costante puntualità con cui mi ha accompagnata, passo dopo passo, nella stesura di questo lavoro. Un sentito ringraziamento va, inoltre, al mio correlatore, il Dott. Michele Marchetti, per la disponibilità e il supporto che mi ha offerto durante questo percorso.

A mia madre e mio padre, per essere stati le mie fondamenta: grazie per i vostri sacrifici, per avermi sempre spronata a dare il massimo e per aver creduto in me, anche quando io per prima faticavo a farlo. Se oggi sono arrivata fin qui, è anche grazie alla forza che mi avete trasmesso.

A mio fratello, per le risate, i battibecchi e per esserci stato sempre, a modo tuo.

Ai miei nonni, le mie radici più salde e uno dei doni più preziosi che la vita potesse farmi. Grazie per l'amore incondizionato, per i vostri piatti deliziosi che sanno di casa e per l'affetto caloroso con cui mi avete avvolta a ogni mio ritorno.

Ai miei zii, per la presenza costante e rassicurante. Dalle infinite sfide ai giochi da tavolo, fino alle canzoni condivise a tutto volume, unite dagli stessi gusti musicali, abbiamo collezionato ricordi che porterò sempre con me. Grazie per esserci sempre stati, siete i miei fratelli maggiori.

A Davide, Marco e Tarek, i miei insostituibili compagni di (dis)avventure universitarie. Tra interminabili sessioni di studio su Discord, progetti di gruppo affrontati sempre insieme, risate a non finire e infinite pause su Brawl, abbiamo condiviso la vera essenza di questi anni. Senza la nostra squadra e il vostro supporto, la fatica sarebbe stata il doppio e le soddisfazioni la metà. Un grazie speciale va anche a tutti gli altri amici e compagni di corso incontrati in questi anni, per aver condiviso con me le difficoltà e i momenti più belli di questo percorso.

Infine, a Giacomo, l'amore della mia vita. Siamo cresciuti insieme, condividendo anni di cambiamenti, sogni, difficoltà e traguardi, e non riesco a immaginare questo percorso senza averti al mio fianco. Grazie per aver saputo accogliere e calmare ogni mio sbalzo d'umore durante le sessioni d'esame, con una pazienza disarmante, riuscendo sempre a trasformare la pesantezza di quelle giornate in momenti di assoluta serenità. Sei stato il mio porto sicuro nei momenti più difficili e la persona con cui ho potuto condividere ogni piccolo traguardo. Se oggi festeggio questo importante risultato, è anche merito tuo.