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 implementazione di una campagna di data science sui dati relativi ad un'app per la vendita di prodotti vintage

Design and implementation of a data science campaign on data related to a vintage product sales app

Relatore

Prof. Domenico Ursino

Correlatore

Dott. Francesco Cauteruccio

Candidato

Alessandra D'Anna

ANNO ACCADEMICO 2022-2023

Fatto che si abbia il male, bisogna farlo tutto quanto.
È da pazzi sperare di fermarsi ad un punto qualunque del mostruoso!
Il delitto spinto agli estremi ha deliri di gioia.

Victor Hugo, "Notre-Dame de Paris"

Sommario

Nel corso degli ultimi anni, l'evoluzione rapida e la diffusione estesa della data analytics hanno offerto alle aziende l'opportunità di sfruttare in modo efficace le proprie informazioni, conseguendo un vantaggio significativo sulla concorrenza. La comprensione approfondita della distribuzione delle vendite, del comportamento dei clienti e dei trend dei prodotti, specialmente in relazione alle variazioni geografiche, riveste un ruolo cruciale, in particolare nel contesto dei siti di e-commerce. La presente tesi si concentra sull'analisi dei dati relativi agli utenti registrati su un sito di compravendita di capi d'abbigliamento, i cui dati sono stati successivamente pubblicati su Kaggle. La ricerca include un'analisi descrittiva dei dati e una fase di ETL. Successivamente, è stata condotta un'attenta attività di data visualization, che ha generato report scrupolosi riguardanti le diverse caratteristiche degli utenti. L'obiettivo è quello di fornire una panoramica completa delle abitudini degli utenti e ricavare informazioni dettagliate sulla tipologia di clientela cui il sito è rivolto.

Keyword: Data Analytics, Big Data, Kaggle, Fashion E-Commerce, Extract Trasform and Load, Data Visualization, Power BI

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

L'argomento centrale di questa tesi è la Data Analytics, scelta motivata dall'enorme rilevanza acquisita dai dati negli ultimi anni.

Con la diffusione di numerosi dispositivi tecnologici dotati di sensori, che generano e distribuiscono dati, e di connessione Internet, si è sviluppato l'Internet delle cose (IoT), creando una vasta rete di dati. I big data hanno influenzato profondamente il lavoro e la vita quotidiana, con impatti emotivi legati alle aspettative di un futuro più efficiente, sostenibile e prosperoso, ma anche alle preoccupazioni sulla privacy.

Le fonti di valore economico che derivano dallo sfruttamento dei dati sono molteplici; ad esempio, si potrebbero portare avanti campagne di marketing molto più accurate, basate su uno studio della clientela a cui si rivolgono. Inoltre, uno studio dei clienti, così come del mercato in cui ci si introduce o in cui si sta lavorando, permette di compiere scelte aziendali maggiormente ponderate ed informate che porterebbero, dunque, ad un miglioramento delle prestazioni. È possibile portare avanti uno studio incentrato sulle tendenze nei prodotti venduti dall'azienda, basandosi sulla regione/paese di vendita o sull'età o il genere dei clienti. Infine, sappiamo che, tramite l'analisi dei dati, è possibile migliorare le prestazioni aziendali, studiando le performance dei dipendenti, oltre che l'avanzamento dei processi interni.

Nel contesto aziendale è, dunque, lampante come l'adozione di soluzioni di data analytics possa garantire un vantaggio competitivo significativo.

Nel 2021 il mercato italiano era poco "data driven", dallo studio de *Osservatorio sulla Digital Innovation* emerge che.

> *Solo il 27% del campione può definirsi data science driven; il 14% è in una fase sperimentale; il 28% si colloca nel gruppo "primi passi", con in corso le prime sperimentazioni; il 16% è consapevole, ovvero sta valutando idee progettuali; il 15% è tradizionale.*

Mentre nel 2023 si è osservato che:

> *Rispetto al 2022, è possibile definire avanzate il 20% delle grandi imprese. Un anno fa era il 15%. Seguono le aziende definite focalizzate (12%): ben avviate sulla Data Science in alcune funzioni aziendali, ma con una scarsa attenzione alla valorizzazione complessiva del patrimonio informativo. D'altra parte, un terzo delle grandi aziende è immaturo o ai primi passi (32%). Si tratta soprattutto delle più piccole. Per queste aziende la priorità è il completo superamento dell'utilizzo di fogli elettronici e l'introduzione pervasiva di strumenti di Data Visualization e di Reporting avanzati. Il 13% delle aziende si sono concentrate sulla Data Science e hanno iniziato a sperimentare nell'ultimo anno. Il 23%*

hanno dato priorità ad una buona qualità dei dati e alla presenza di figure dedicate alla Data Governance.

Questa tesi si colloca proprio nel contesto della Data Science. In particolare, in essa verranno esaminati dati provenienti da Kaggle riguardanti un dataset specifico denominato "Fashion E-Commerce". Questo dataset fornisce informazioni sugli utenti di un sito di e-commerce di abbigliamento, focalizzandosi sulla provenienza, i movimenti e le abitudini degli utenti. La fase iniziale comprenderà la descrizione, l'esplorazione e la comprensione dei dati, seguite da un processo ETL (Estrazione, Trasformazione, Caricamento) per facilitare l'analisi.

Dopo l'attuazione delle opportune procedure di ETL, il risultato sarà trasferito su Power BI, piattaforma in cui si effettueranno le operazioni di analisi dei dati, con una particolare attenzione rivolta all'ambito della data visualization. Quest'ultima costituirà un elemento chiave nella presentazione delle operazioni di data analytics, consentendo una comprensione approfondita e intuitiva dei risultati ottenuti.

In Power BI, saranno creati quattro report distinti. Il primo report fornirà un'analisi dettagliata degli utenti iscritti al sito, suddivisi per paese di provenienza. Il secondo report presenterà uno studio analogo, ma con il focus sul genere degli utenti. I successivi due report, il terzo e il quarto, si concentreranno sull'analisi delle abitudini degli utenti, offrendo una panoramica completa delle tendenze e dei comportamenti registrati.

Queste operazioni mirano a ottenere una comprensione più approfondita dei comportamenti degli utenti nel sito e a identificare eventuali pattern legati al genere e al paese di provenienza. Nonostante la natura generica dello studio, sottolineiamo come tali analisi possono fornire informazioni cruciali per la gestione e l'ottimizzazione di un sito di e-commerce.

La presente tesi è composta da cinque capitoli, strutturati come di seguito specificato:

- Nel Capitolo 1 viene presentata una descrizione della Data Analytics; in particolare, ne riportiamo una definizione, le categorie in cui si divide, il concetto di Data Analytics e le differenze con la Data Analysis.
- Nel Capitolo 2 si introduce lo strumento utilizzato nello studio, ovvero PowerBI; in particolare, ne indichiamo le motivazioni dell'utilizzo, le componenti e le funzionalità.
- Nel Capitolo 3 presentiamo un'analisi descrittiva dei dati a disposizione e illustriamo l'attività di ETL su questi.
- Nel Capitolo 4 vengono riportate le analisi effettuate ed i risultati derivati da queste ultime.
- Nel Capitolo 5 si riporta una discussione in merito al lavoro svolto, concentrandoci sui punti di forza e di debolezza.
- Infine, la tesi termina con una serie di conclusioni che abbiamo potuto trarre dal lavoro svolto.

CAPITOLO 1

Introduzione alla Data Analytics

Il capitolo inziale vuole fornire una descrizione del concetto di Data Analytics, comprendendo anche nozioni sulla sua storia, la sua evoluzione, i suoi ambiti di utilizzo e le fasi che compongono il processo di Analisi in sè per sè.

1.1 Cos'è la Data Analytics

La Data Analytics, o analisi dei dati, è un campo cruciale nell'era digitale in cui viviamo. Si tratta di un processo attraverso il quale vengono raccolti, elaborati ed interpretati dati grezzi al fine di estrarre informazioni significative e prendere decisioni informate. Questa disciplina si avvale di una serie di metodi e strumenti, compresi algoritmi, software e tecniche statistiche, per analizzare grandi quantità di dati provenienti da varie fonti, come database aziendali, sensori IoT, social media e molto altro.

L'obiettivo della Data Analytics è scoprire tendenze, modelli nascosti e correlazioni nei dati, tutte nozioni che altrimenti sarebbero perse nella miriade di informazioni che vengono ricevute. Tutto ciò consente alle organizzazioni di prendere decisioni più accurate, migliorare le operazioni, ottimizzare i processi e anticipare le esigenze dei clienti. La Data Analytics è fondamentale in settori come il marketing, la salute, la finanza, la produzione e molti altri, contribuendo al progresso e alla competitività delle aziende e alla comprensione approfondita dei fenomeni nel mondo contemporaneo. Tale disciplina ha origini che risalgono a svariate decine di anni fa; tuttavia, essa trova uno sviluppo ed una crescita significative soltanto negli ultimi decenni. In particolare possiamo evidenziare le seguenti tappe:

- *Anni '60-'70:* a questo periodo risalgono i primi tentativi di utilizzo del computer per l'ananlisi dei dati. L'analisi era principalmente basata su metodi statistici tradizionali. L'uso di mainframe e computer per eseguire analisi statistiche ha rappresentato un passo significativo verso l'automazione del processo analitico.
- *Anni '80-'90:* durante questo periodo, l'aumento della potenza di calcolo dei computer ha reso possibile l'elaborazione di dati più complessi. Le organizzazioni hanno iniziato a raccogliere sempre più dati e a utilizzare sistemi di gestione dei database per archiviarli e recuperarli in modo più efficiente. Inoltre, sono emerse le prime applicazioni di data mining, che miravano a scoprire modelli e tendenze nei dati.

- *Anni 2000:* con la crescita di Internet, l'esplosione dei dati online ha portato all'accumulo di enormi quantità di dati. In risposta a questa sfida, molte aziende e organizzazioni hanno iniziato a sviluppare nuove tecniche di analisi dei cosiddetti big data. In questo periodo, il concetto di "Data Science" è emerso come una disciplina interdisciplinare che combina statistica, informatica e conoscenza di dominio per l'analisi dei dati.
- *Anni 2010 e oltre:* l'era moderna della Data Analytics è caratterizzata dalla crescente adozione di tecnologie di cloud computing, l'uso diffuso di strumenti e framework open source come Hadoop e Spark per l'elaborazione di big data, e lo sviluppo di algoritmi di machine learning avanzati. L'analisi dei dati è diventata una parte fondamentale delle operazioni aziendali in settori come il marketing, la sanità, le finanze, la scienza dei dati e molto altro.

1.1.1 Passi fondamentali

La Data Analytics è un processo complesso che coinvolge diverse fasi per estrarre informazioni significative dai dati. I passi fondamentali della Data Analytics includono:

1. *Definizione degli Obiettivi:* questa fase inizia con la definizione chiara degli obiettivi dell'analisi. Cosa si cerca di ottenere dai dati? Quali domande si vogliono risolvere? Gli obiettivi devono essere specifici, misurabili, realistici e temporizzati.
2. *Raccolta dei Dati:* in questa fase, vengono raccolti i dati necessari per l'analisi. I dati possono provenire da varie fonti, come database, sensori, registri, file CSV, API web e molto altro. È essenziale garantire che i dati siano accurati e completi.
3. *Pulizia dei Dati:* i dati spesso contengono errori, valori mancanti o informazioni non valide. Nella fase di pulizia dei dati, vengono identificati e corretti tali problemi. Questo passo è cruciale per garantire l'affidabilità dell'analisi.
4. *Esplorazione dei Dati:* qui, vengono eseguite analisi preliminari per comprendere meglio i dati. Questo può includere la creazione di grafici, il calcolo di statistiche di base e l'identificazione di tendenze o pattern nei dati.
5. *Preparazione dei Dati:* i dati vengono preparati per l'analisi vera e propria. Questo può comportare la selezione delle variabili rilevanti, l'ingegneria delle feature (creazione di nuove variabili basate sui dati esistenti) e la suddivisione dei dati in set di addestramento e test se si sta costruendo un modello predittivo.
6. *Analisi dei Dati:* in questa fase, vengono utilizzati metodi statistici, algoritmi di machine learning e altre tecniche per estrarre informazioni significative dai dati. Ciò può includere l'identificazione di correlazioni, la creazione di modelli predittivi, l'analisi delle serie temporali e molto altro.
7. *Visualizzazione dei Risultati:* i risultati dell'analisi vengono solitamente comunicati attraverso grafici e visualizzazioni. Questo aiuta a rendere tali risultati comprensibili e adatti alla presentazione a un pubblico non tecnico.
8. *Interpretazione e Comunicazione:* i risultati dell'analisi vengono interpretati in relazione agli obiettivi stabiliti nella Fase 1. Le conclusioni vengono, quindi, comunicate a chi deve prendere decisioni basate su tali risultati.
9. *Pianificazione delle Azioni:* sulla base delle conclusioni dell'analisi, vengono pianificate e implementate azioni o strategie. Questo è il passo finale, ma molto importante, poiché l'obiettivo dell'analisi è, spesso, quello di guidare azioni concrete.

10. *Monitoraggio e Ottimizzazione:* dopo aver implementato le azioni, è importante monitorare costantemente i risultati per valutare l'efficacia delle decisioni prese. Se necessario, è possibile apportare modifiche e ottimizzazioni.

Tali passaggi possono essere iterati più volte a seconda della complessità del progetto e dei nuovi dati che possono emergere durante il procedimento; inoltre non sono sequenziali.

1.1.2 Utilizzo della Data Analytics

La Data Analytics viene utilizzata in un vasto range di settori che spaziano da quello quello medico a quello industriale fino a quello economico. In quest'ultimo contesto alcune delle possibii applicazioni sono le seguenti: rilevamento delle anomalie, gestione dei dati dei clienti, gestione del rischio, rilevamento delle frodi, personalizzazione e customizzazione, realizzazione di ricerche di mercato

Le aziende ne fanno uso basandosi sui loro approcci decisionali e sul loro stato di sviluppo. In generale questi sei utilizzi si distribuiscono nelle quattro categorie, più una aggiuntiva, della Data Science (Figura 1.1)

Figura 1.1: Figura 1.1 Le varie tipologie di Data Analytics

- Analisi Descrittiva
- Analisi Diagnostica
- Analisi Predittiva
- Analisi Prescrittiva
- Forme di Analisi Avanzata

Nelle prossime sottosezioni descriveremo ciascuna di queste tipologie di analisi.

1.1.2.1 Analisi Descrittiva

L'analisi descrittiva (o descriptive analytics) è la forma più semplice e diffusa di Business Intelligence che, attraverso le KPI (Key Performance Indicator), consente di analizzare e comparare dati storici e attuali per comprendere meglio l'attuale andamento aziendale. In particolare, ci permette di trasformare dei dati grezzi in blocchi comprensibili.

Se si avesse la necessità di capire quanto un sito web o un sito di e-commerce stia crescendo in termini di interesse e traffico in relazione ad un'azione di marketing, questo sarebbe sicuramente il tipo di analisi che, in modo facile e rapido, potrebbe fornire le risposte cercate.

1.1.2.2 Analisi Diagnostica

L'analisi diagnostica (o diagnostic analytics) è una tipologia di data analytics che si concentra sull'identificazione delle cause che hanno comportato problemi, anomalie o situazioni indesiderate nei dati. Questo tipo di analisi è fondamentale per definire le strategie e le mosse correttive da adottare per migliorare le prestazione di un processo o di un sistema.

Supponiamlo di far parte del team di gestione di una società di e-commerce, la quale sta affrontando un periodo di calo delle vendite. In questa casistica l'analisi diagnostica fornisce lo strumento ideale per analizzare le eventuali cause di un calo delle vendite e individuare la soluzione per porvi rimedio

1.1.2.3 Analisi Predittiva

L'analisi predittiva (o predictive analytics) viene utilizzzata per prevedere i mutamenti futuri sulla base dei dati attuali. Grazie alle alte velocità di elaborazione e agli algoritmi complessi si possono elaborare modelli, anticipare l'insorgere di problematiche, prepararsi per tempo alle evoluzioni nei sistemi di acquisto e ai cambi di trend.

Si pensi, ad esempio, di voler lanciare un prodotto nel mercato; in questo caso serviranno delle analisi dei dati specifiche per individuare nuove opportunità e capire se l'idea può funzionare o meno. Determinando con anticipo vantaggi e svantaggi, è possibile ottimizzare la strategia e prendere le decisioni più efficaci per uscire sul mercato.

1.1.2.4 Analisi Prescrittiva

L'analisi prescrittiva (o prescriptive analytics) rappresenta uno dei livelli più avanzati nell'ambito dell'analisi dei dati ed è considerata un "miglioramento" rispetto all'analsi predittiva, questo in quanto, oltre alla previsione degli eventi futuri, è in grado di suggerire come muoversi per raggiungere un obiettivo o un risultato. Questi grandi passi in avanti richiedono una solida comprensione delle metodologie di modellazione, delle tecnologie di analisi dei dati e delle esigenze specifiche del dominio in cui viene applicata.

L'esempio pratico lo si trova nei suggerimenti di ottimizzazione proposti dai tool come Google Analytics, Search Console, o anche nelle campagne di sponsorizzazione di Google ADS e Facebook

1.1.2.5 Forme di analisi avanzate

Tra le varie forme di analisi avanzata annoveriamo Intelligenza Artificiale e Machine Learning, che forniscono strumenti in grado di contribuire significativamente all'elaborazione e all'interpretazione dei dati, rendendo il processo di analisi quasi completamente automatico.

Il Data Mining è il processo di trasformazione di dati non strutturati in informazioni utili, come schemi ricorrenti, correlazioni e anomalie.

Il Text Mining, utilizza l'elaborazione del linguaggio naturale (NLP) per l'estrazione di informazioni utili e strutturate da documenti testuali.

1.2 Il nuovo standard Aziendale

La Data Analytics è entrata a gamba tesa all'interno del mondo aziendale permettendo l'analisi dei dati coinvolti (da quelli ottenuti in tempo reale fino ad arrivare agli storici) consentendo un rilevamento di trend e metriche che, altrimenti, andrebbero persi all'interno della massa dei dati. In questo modo l'analisi dei dati permette di prendere decisioni più informate, migliorare l'andamento aziendale e rilevare i colli di bottiglia all'interno del processo. Questo tipo di studio dei dati permette, anche, alle aziende di risultare più competitive, caratteristica imprescindibile in un mercato ormai saturo di competitor. Da tale discipilina nasce dunque una figura chiave, il Data Analyst, ovvero colui che si occupa dell'estrazione dei dati, e non solo. In generale il Data Analyst dovrà seguire il corretto procedimento che la disciplina impone, quindi assicurarsi della correttezza dei dati estratti, creare delle dashboard comprensibili ed esporre al team il risultato del lavoro svolto.

1.2.1 Vantaggi della Data Analytics

Grazie all'inclusione della Data Analytics come parte integrante delle aziende abbiamo la possibilità di sfruttare tale strumento per ottimizzare al meglio il buisness stesso. Essa infatti comporta una serie di vantaggi elencati nelle seguenti sottosezioni.

1.2.1.1 Collaborazione

La collaborazione sinergica tra le competenze di Data Science, le esigenze della Linea di Business e le capacità dei Team IT rappresenta un pilastro fondamentale per ottimizzare l'efficienza e la produttività aziendale. Questa sinergia è cruciale in quanto le informazioni necessarie all'analisi provengono da diverse aree aziendali. In aggiunta, l'analisi dei dati stessi richiede una vasta gamma di competenze e conoscenze specializzate per garantire risultati accurati e affidabili. La collaborazione interfunzionale favorisce l'innovazione, consentendo l'integrazione di prospettive diverse e portando a soluzioni più complete e comprensibili per tutti i membri del team e per l'azienda nel suo insieme.

1.2.1.2 Identificazione anomalie

L'analisi di grandi volumi di dati di streaming sia all'interno dei sistemi di core business che nelle aree periferiche permette di trovare anomalie, prendere decisioni e agire nel punto di impatto. Con volumi sempre maggiori di dati, essere in grado di analizzare, filtrare, riassumere e ottenere informazioni in tempo reale consente di individuare le anomalie prima che diventino un problema più grande. Con l'espansione continua dei dati disponibili, sia internamente che esternamente all'azienda, la sfida principale diventa quella di essere in grado di analizzare, filtrare, riassumere e ottenere informazioni in tempo reale. Questo è cruciale per individuare le anomalie prima che diventino problemi più gravi. Il monitoraggio dei dati in tempo reale è quindi l'attività più utile in relazione alle anomalie, permettendoci di rilevarle quanto prima.

1.2.1.3 Monitorazione e gestione dati

L'implementazione di soluzioni di monitoraggio dei dati sta diventando sempre più cruciale per le aziende che desiderano sfruttare appieno il potenziale dell'analisi dei dati. Mentre molte aziende si stanno impegnando per rendere operativa l'analitica, la capacità di monitorare i dati in modo continuo e automatizzato rappresenta un passo fondamentale per garantire risultati affidabili e tempestivi. Il monitoraggio dei dati non si limita solo

alla loro raccolta e alla loro archiviazione, ma coinvolge anche il rilevamento di tendenze, comportamenti anomali o pattern significativi all'interno dei dati stessi. Questo processo consente alle aziende di comprendere meglio il proprio ambiente operativo e di agire in risposta a cambiamenti o a problemi. Essere in grado di conservare, aggiornare e distribuire automaticamente nuovi modelli analitici è essenziale per garantire che le decisioni aziendali siano sempre basate sui dati più recenti. L'automazione viene in soccorso consentendo di implementare rapidamente i modelli nelle diverse aree aziendali senza interruzioni o ritardi significativi

1.2.1.4 Utilizzo dell'Intelligence

L'Intelligenza Artificiale (IA) e l'apprendimento automatico (Machine Learning) stanno rivoluzionando il campo della Data Analytics, portando a nuove prospettive e capacità sorprendenti. L'integrazione dell'intelligenza nell'analisi dei dati offre alle aziende una serie di opportunità senza precedenti per ottenere insight profondi e per prendere decisioni più accurate. L'IA consente di affrontare sfide complesse che erano difficili da gestire con approcci tradizionali. Grazie all'apprendimento automatico, i sistemi possono riconoscere modelli nei dati che potrebbero essere difficilmente individuabili da un essere umano o da un software convenzionale. Questo significa che le aziende possono trarre vantaggio da analisi più sofisticate e predittive. Un esempio tangibile è l'utilizzo dell'IA nel settore del marketing. Le aziende possono utilizzare algoritmi di Machine Learning per analizzare i dati dei clienti e identificare i segmenti di pubblico più promettenti, personalizzando le offerte e le strategie di marketing in modo più efficace. Ciò porta a una maggiore fedeltà dei clienti e a una crescita delle entrate.

1.3 Big Data Analytics

La Big Data Analytics è attualmente la tipologia di analisi dei dati più richiesta ed utilizzata, soprattutto in ambito aziendale. Per introdurre questo argomento dobbiamo spiegare nel dettaglio cosa intendiamo con Big Data ed il loro utilizzo nell'attuale quotidianità Malgrado non esista una netta separazione tra "Big Data" ed altre tipologie di data, ci affidiamo alla descrizione fornitaci da Teradata nel 2011 che afferma:

> Un sistema di big data eccede i sistemi hardware e software comunemente usati per catturare, gestire ed elaborare i dati in un lasso di tempo ragionevole per una popolazione di utenti anche massiva

Possiamo, quindi, affermare che si faccia riferimento ad una raccolta informatica di dati estesa in termini di volume, velocità di elaborazione e varietà di dati contenuti al suo interno. Essi richiedono metodi specifici di estrazione ed elaborazione di tale contenuto. Tutta questa risorsa di informazoni può provenire dalle più disparate fonti come, ad esempio, tracce GPS dei nostri cellulari, transazionI fatte sul web o, semplicemente, post pubblicati. Proprio a causa di questa loro natura i big data sono stati al centro di un'enorme bufera mediatica avente come tema cenTrale, quello della privacy. La loro definizione non è sufficiente per esprimere l'impatto sociale, economico e tecnologico che la loro introduzione ha significato. L'evoluzione tecnologica derivata è tuttora in atto.

1.3.1 Storia dei Big Data

Il concetto di Big Data nasce all'incirca verso gli anni 90 e il più grande contributo al loro utilizzo e alla loro comprensione fu dato da Doug Laney nel 2001, il quale formulò la teoria delle 3V:

- Velocità
- Varietà
- Volume

A tale regola succedette quella delle 5V, coniata del 2012 ad opera dell'analista di ricerca Gartner, il quale ampia il precedente modello grazie a due attributi aggiuntivi:

- Veridicità
- Valore

Nel seguito esaminiamo più in dettaglio queste cinque proprietà dei big data.

1.3.1.1 Velocità (Velocity)

La V della velocità riguarda la rapidità con cui i dati vengono generati, raccolti e devono essere elaborati. Con la crescente connettività, i sensori, i dispositivi IoT (Internet delle cose) e i social media, i dati possono essere generati in tempo reale o a una velocità molto elevata. Ciò pone sfide particolari per l'elaborazione e l'analisi dei dati in tempo reale o quasi in tempo reale per estrarre informazioni significative e prendere decisioni rapide.

1.3.1.2 Varietà(Variety)

La V della varietà si riferisce alla diversità dei tipi di dati che possono essere inclusi nei big data. Questi dati possono essere strutturati, semi-strutturati o non strutturati, e possono provenire da una varietà di fonti, come testi, immagini, audio, video, dati geospaziali, dati transazionali e altro ancora. La gestione e l'analisi di questa varietà di dati richiedono spesso l'uso di diverse tecnologie e approcci, compresi i motori di ricerca, l'elaborazione del linguaggio naturale, l'analisi delle immagini e molto altro.

1.3.1.3 Volume(Volume)

Questa V si riferisce alla quantità di dati generati, raccolti e archiviati. Nei big data, il volume dei dati è enorme e supera di gran lunga la capacità dei sistemi tradizionali di gestione dei dati. I dati possono essere strutturati (come i database), semi-strutturati (come file XML o JSON) e non strutturati (come testi, immagini o video). La gestione di grandi volumi di dati richiede, spesso, infrastrutture di archiviazione ed elaborazione altamente scalabili.

1.3.1.4 Veridicità(Veracity)

Questa V si concentra sulla qualità e l'affidabilità dei dati. Spesso i dati possono essere incompleti, erronei o inaffidabili, il che può influenzare negativamente l'accuratezza delle analisi. Mantenere la veridicità dei dati è cruciale per ottenere risultati affidabili.

1.3.1.5 Valore(Value)

Questa V rappresenta l'obiettivo finale dell'analisi dei big data, ovvero l'estrazione di valore dai dati. Gli sforzi di raccolta e analisi dei dati devono essere mirati a generare informazioni utili e insight che possano guidare decisioni aziendali, ottimizzare processi o migliorare il rendimento.

1.3.2 Big Data Analytics

Una volta parlato dei Big Data possiamo definire il concetto del Big Data Analytics come:

> Processo che racchiude l'analisi e la raccolta dei Big Data per ottenere informazioni utili sul business sfruttando al massimo i dati stessi. è introdotto in quanto i volumi di dati che arrivano in input diventano talemente voluminosi che necessitano di una maggiore capacità di calcolo.

Le tecnologie di analisi dei dati sono, infatti, determinanti per analizzare questa risorsa fondamentale per le organizzazioni; in questa situazione emergono due discipline che vanno a spiccare in quanto propedeutiche alla Big Data Analytics: Business Intelligence, ovvero la BI, che si occupa soprattutto delle analisi descrittive, e la Business Analytics, finalizzata alla realizzazione di analisi predittive e prescrittive. In altri termini, la big data analytics rappresenta una tipologia di advanced analytics che impiega applicazioni complesse, come modelli predittivi, algoritmi statistici e analisi what-if sfruttando sempre più spesso la capacità di elaborazione del cloud computing, che ha consentito di superare un limite oggettivo che, per molti anni, ne ha limitato la diffusione su larga scala. L'analisi dei Big Data porta con sé la responsabilità di gestire e proteggere i dati in modo etico e conforme alle normative sulla privacy. La capacità di generare valore informativo attraverso l'analisi dei dati può garantire alle organizzazioni una serie di importanti benefici, a condizione che le attività di big data analytics vengano implementate in maniera consapevole nei processi aziendali.

Tra i principi che una corretta implementazione delle attività di Big Data Analytics permettono di garantire, annoveriamo i seguenti

- *Olistica del business*: la Big Data Analytics può aiutare le organizzazioni a ottenere una visione ricca dei dati, coerente e completa del business. Dashboard analitici di facile utilizzo e applicazioni aziendali aumentano il processo decisionale guidato dai dati e permettono agli utenti non tecnici di operare sulla base di informazioni accurate e tempestive, invece che sull'istinto.
- *Time-to-action più veloce*: le aziende hanno bisogno di big data analytics per permettere a tutti al loro interno di prevedere situazioni e opportunità, porre domande pertinenti e tempestive ed ottenere le risposte di cui hanno bisogno per intraprendere azioni decisive. Queste ultime possono anche essere automatizzate per garantire una risposta rapida.
- *Visibilità nell'ignoto*: per scoprire tendenze e modelli non visti o nascosti in grandi e complessi insiemi di dati, le aziende dovrebbero usare la Big Data Analytics. Questo permetterà di identificare più velocemente le opportunità strategiche o i rischi per l'organizzazione.
- *Data Discovery self-service*: la big data analytics può permettere agli utenti di esplorare i dati e ottenere risposte senza la necessità di una modellazione dei dati specializzata e approfondita. Ciò riduce la dipendenza dall'IT e accelera il processo decisionale.

1.3.2.1 Casi di Utilizzo

I casi di utilizzo della Big Data Analytics, secondo quanto dichiarato a Forrester dalle aziende utenti, rientrano in tre gruppi:

1. *Efficienza e rischi operativi*: gran parte degli esempi di Big Data Analytics realizzati o pianificati per esserlo a breve riguarda la riduzione del rischio nelle analisi finanziarie.

Altri ambiti dove contano efficienza e risk reduction sono l'asset management (con una punta nell'analisi delle frodi), la gestione del personale e la supply chain, dove emergono le applicazioni di big data per la manutenzione preventiva. Un approccio globale a questi problemi deve considerare la condivisione dei dati e lo scambio di idee con i business partner, nonché il tracciamento dei risultati avuti dalle azioni intraprese in seguito a dette analisi, in modo da avviare un ciclo virtuoso.

2. *Sicurezza e performance applicative*: Predictive analytics e analisi dei big data sul funzionamento dell'IT servono a prevenire problemi nell'erogazione dei servizi e a monitorare gli eventi per potervi rispondere in tempo reale. I modelli d'analisi, che vanno discussi con i responsabili della sicurezza e dei servizi, si servono dei data-log generati da server e dispositivi di rete. Essi sono, infatti, utili per valutare i livelli prestazionali, trovare i colli di bottiglia e quant'altro.

3. *Conoscenza e servizio ai clienti*: soluzioni e applicazioni per la Big Data Analysis sono utilizzati per progetti di marketing e vendite, per lo sviluppo dei prodotti, ma anche per l'ottimizzazione della digital experience.

1.3.2.2 Questione etica

Come avevamo introdotto nella definizione di Big Data, il loro utilizzo ha generato un dissenso comune sfociato in un'ampia discussione sul diritto alla privacy e su come questo non debba essere violato in nome di un guadagno disinteressato da parte delle stesse aziende. I concetti principali attorno ai quali tale polemica ruota sono: la volontà di conoscere quali informazioni di interesse vengono ottenute dai comportamenti degli utenti, chi sono i destinatari di queste informazioni ed in che modo queste vengono utilizzate.

La rapida modernizzazione della raccolta dei dati ha lasciato i clienti all'oscuro di ciò che le aziende stanno facendo alle loro spalle. Un esempio reale di ciè si è avuto quando Mark Zuckerberg ha dato alla Cambridge Analytic accesso non autorizzato ai dati personali degli utenti di Facebook. Facebook e molte altre aziende possono cercare di minare la privacy dei loro utenti. Questa debacle ha causato molti dibattiti su come i big data potrebbero essere utilizzati come arma per massimizzare i profitti aziendali.

Inoltre, quando le aziende acquisiscono informazioni sui loro clienti, sono in grado di utilizzare tali informazioni per manipolare i clienti o diffondere disinformazione. Con il livello avanzato di profilazione che gli analisti possono fare con i big data, è facile vedere cosa vogliono i clienti e utilizzare tali informazioni a vantaggio dell'azienda. Manipolare i clienti con i dati non è una novità. Infatti, le aziende stanno usando le strategie sneaky di vendita per ingannare i consumatori per molto tempo. Tuttavia, i big data fungono da booster per strategie di marketing più manipolative, consentendo alle aziende di analizzare i dati dei clienti a velocità più elevate e con maggiore efficienza rispetto a prima.

1.4 Data Analytics o Data Analysis?

Volendo trattare le differenze tra queste due branche della ben più vasta Data Science, dobbiamo introdurre il concetto di Data Analysis

> L'analisi dei dati consiste nel pulire, manipolare, modellare e interrogare i dati per scoprire informazioni rilevanti. È una parte vitale dell'analisi dei dati. Ci aiuta a identificare le soluzioni fornendo informazioni.

I due termini vengono erroneamente utilizzati come sinonimi mentre la Data Analysis rappresenta un passaggio propedeutco alla Data Analytics.

Nella Figura 1.2 riportiamo gli skill che caratterozzano la Data Science e la Data Analytics.

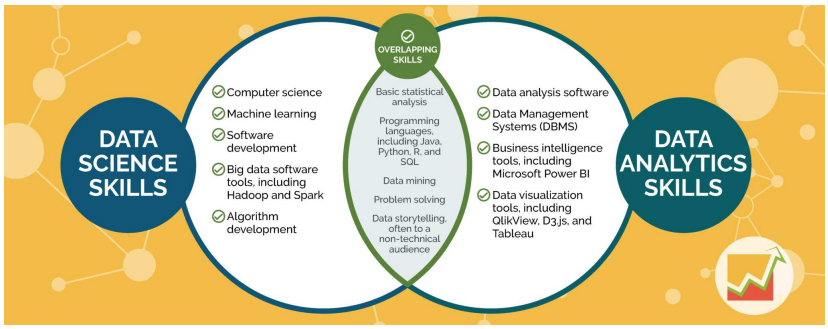

Figura 1.2: Figura 1.2 Skill che caratterizzano la Data Science e la Data Analytics

1.4.0.1 Data Analysis tipologie

Esistono molteplici tecniche per la Data Analysis. Di seguito riportiamo le più note:

- *Data Mining*: è un metodo che serve per evidenziare possibili Pattern all'interno di dataset estesi.
- *Statistica Descrittiva*: questa tecnica è utilizzata per riassumere e descrivere i dati attraverso misure centrali (media, mediana, moda), misure di dispersione (deviazione standard, intervallo interquartile) e rappresentazioni grafiche, come istogrammi, grafici a barre e diagrammi a dispersione.
- *Analisi di Regressione*: questa tecnica è utilizzata per studiare le relazioni tra variabili. La regressione lineare è la forma più comune, ma ci sono anche altri tipi di regressione, come la regressione logistica, per dati categorici, e la regressione polinomiale, per relazioni non lineari.
- *PCA (Analisi Componenti Principali)*: utilizzata per ridurre la dimensionalità dei dati mantenendo, al contempo, le informazioni più significative. PCA trasforma i dati in un nuovo sistema di coordinate basato sulle componenti principali.
- *Analisi dei Sentimenti*: utilizzata per estrarre informazioni sugli atteggiamenti, le opinioni e le emozioni dai testi; spesso è adottato nel monitoraggio dei social media e nelle recensioni dei prodotti.
- *Apprendimento Automatico*: comprende un'ampia gamma di tecniche che consentono al computer di apprendere da dati passati per fare previsioni o prendere decisioni. Queste tecniche includono l'apprendimento supervisionato, non supervisionato e profondo.

1.4.0.2 Differenze e Similitudini

L'origine dell'interscambiabilità dei due termini è da ricercare all'interno delle loro somiglianze, ed è proprio per questo che metteremo in evidenza, anche e sopratutto, le loro divergenze.

1. Focus e obiettivo:
    - Data Analysis
    - Data Analytics

2. Processo:
    - *Data Analysis*: l'analisi dei dati si concentra principalmente sulla comprensione dei dati grezzi, compresi il loro significato, la loro struttura e le relazioni tra di essi, mentre l'obiettivo si focalizza sull'interpretazione dei dati e sulle conclusioni che da essi si possono trarre.
    - *Data Analytics*: l'analisi dei dati si concentra sulla generazione di informazioni significative e sull'estrazione di valore dai dati attraverso l'applicazione di tecniche quantitative e statistiche avanzate, in uqesto ambito l'obiettivo fondamentale è rappresentato dalla capacità di prendere decisioni informate.
3. Finalità:
    - *Data Analysis*: solitamente si riferisce alle attività di esplorazione, pulizia, trasformazione e interpretazione dei dati, spesso utilizzando metodi statistici e matematici.
    - *Data Analytics*: include un processo più ampio che può comprendere anche la raccolta, la preparazione e l'analisi dei dati nonché la presentazione dei risultati.
4. Strumenti e tecniche:
    - *Data Analysis*: l'analisi dei dati può coinvolgere strumenti di visualizzazione dei dati, analisi descrittiva e tecniche qualitative per esplorare e rappresentare i dati in modo chiaro e comprensibile.
    - *Data Analytics*: la Data Analytics coinvolge spesso l'uso di strumenti software specializzati per l'analisi statistica, la modellazione predittiva e l'elaborazione avanzata dei dati. Le tecniche possono includere regressione, clustering, analisi delle serie temporali, machine learning e altre metodologie quantitative.
5. Coinvolgimento Aziendale:
    - *Data Analysis*: spesso coinvolge scienziati dei dati o analisti dei dati che si concentrano sulla comprensione dei dati stessi.
    - *Data Analytics*: coinvolge spesso professionisti aziendali che utilizzano i risultati dell'analisi per prendere decisioni operative o strategiche.

CAPITOLO 2

Introduzione a Power BI

Questo secondo capitolo vuole introdurre lo strumento utilizzato all'interno del progetto che verrà esposto, ovvero Power BI. Di questo si andranno a descrivere le caratteristiche, in cosa si differenzia da altri tool simili e quali azioni permette di compiere.

2.1 Power BI come strumento

Nel contesto dell'analisi dei dati, i ricercatori e i professionisti si affidano a una vasta gamma di strumenti e tecnologie per esplorare, interpretare e sfruttare il potenziale nascosto nei dati. Questi strumenti costituiscono il pilastro su cui si basa l'intera pratica della Data Analytics. Ciascuno strumento svolge un ruolo cruciale nel processo di estrazione di valore dai dati, fornendo la capacità di trasformare dati grezzi in intuizioni preziose e decisioni informate.

Come menzionato in precedenza, la Data Analytics è un campo in rapida evoluzione che richiede la convergenza di competenze tecniche e strumenti avanzati. In questo contesto, gli strumenti diventano fondamentali per automatizzare compiti ripetitivi, analizzare dati su vasta scala e creare visualizzazioni significative. Sia che si tratti di eseguire analisi statistiche complesse, di costruire modelli di machine learning sofisticati o, semplicemente, di esplorare e visualizzare dati, esistono strumenti specializzati progettati per soddisfare queste esigenze specifiche.

Riportiamo di seguito gli strumenti propedeutici ad una corretta analisi dei dati

- *Strumenti di Manipolazione dei Dati*: essi consentono di raccogliere, pulire e preparare i dati per l'analisi.
- *Strumenti di Visualizzazione dei Dati*: cruciali per rendere i dati comprensibili e comunicare risultati in modo efficace.
- *Strumenti di Anlisi Statistica*: essenziali per l'elaborazione e l'interpretazione statistica dei dati.
- *Strumenti di Machine Learning*: consentono l'automazione delle previsioni e delle decisioni basate sui dati.

Tra le categorie riportate, l'attenzione verrà focalizzata su quella della Visualizzazione dei Dati, le cui funzionalità principali comprendono la scoperta di approfondimenti di

Business Intelligence (BI), la trasmissione di risultati complessi e dettagliati a terzi, l'analisi e la comprensione di tendenze nascoste nei dati ed il rapido confronto di cifre in grandi insiemi di dati.

2.1.1 Strumenti di Visualizzazione dei Dati

Nella presente analisi si inizierà con il porre l'accento sulle caratteristiche che accomunano i vari strumenti di visualizzazione, e che trovaimo di seguito enumerati.

1. *API di Importazione dei Dati*: un'ampia gamma di strumenti,dallo strumento di raccolta dati a quello di visualizzazione, include un'API (Application Programming Interface) che consente l'importazione di dati. L'utilizzo delle API accelera notevolmente il processo di visualizzazione, evitando la necessità di scaricare manualmente i dati, caricarli sullo strumento e formattarli.

2. *Modelli di Grafico*: i modelli di grafico sono configurazioni predefinite che consentono di creare rapidamente visualizzazioni accattivanti. Il loro principale vantaggio risiede nella semplicità di utilizzo, poiché è sufficiente inserire i dati nel grafico stesso. Inoltre, la maggior parte dei modelli di grafico offre opzioni di personalizzazione per colori, caratteri ed intestazioni.

3. *Grafici Interattivi*: i grafici interattivi reagiscono alle interazioni dell'utente permettendo, così, di evidenziare facilmente cifre chiave, tendenze o variabili, senza la necessità di creare visualizzazioni separate per ciascuna analisi.

4. *Storia della Versione*: la cronologia delle versioni consente di visualizzare e ripristinare versioni precedenti, fornendo la possibilità di correggere eventuali errori evitando una possibile perdita dei dati.

5. *Ottimizzazione Mobile*: l'ottimizzazione mobile permette di modificare la presentazione delle visualizzazioni per adattarle ai dispositivi mobili.

Esploreremo, ora, alcuni dei più diffusi e potenti strumenti di visualizzazione utilizzati nell'ambito della Data Analytics. Questi saranno descritti nelle prossime sottosezioni.

2.1.1.1 Power BI

Power BI è un'applicazione di Business Intelligence sviluppata da Microsoft che ha guadagnato notevole riconoscimento nel mondo aziendale. Questa potente piattaforma offre una suite completa di strumenti progettati per l'analisi dei dati e la creazione di report informativi.

L'ecosistema di Power BI opera sinergicamente per soddisfare le esigenze della Business Intelligence, consentendo agli utenti di creare pannelli di controllo informativi e report personalizzati. Un ulteriore servizio, denominato Power BI Embedded, è stato introdotto su Microsoft Azure nel 2016, offrendo ulteriori possibilità di personalizzazione ed integrazione.

La sua genesi è legata ai componenti aggiuntivi di Microsoft Excel, come Power Query, Power Pivot e Power View, che hanno fornito una solida base per la sua evoluzione.

Power BI è divenuto un punto di riferimento nel contesto aziendale, utilizzato per il monitoraggio delle prestazioni, l'individuazione di tendenze, la presa di decisioni basate sui dati e la condivisione delle informazioni tra i membri del team. La sua versatilità, oltre che la sua capacità di tradurre complessi dati in visualizzazioni chiare ed informative, lo hanno reso uno strumento imprescindibile per l'analisi dei dati e la comunicazione visiva delle informazioni, caratteristica che conferisce ad esso un vasto utilizzo anche tra gli utenti meno pratici.

2.1.1.2 Google Chart

Google Chart è una libreria JavaScript potente e altamente flessibile sviluppata da Google per agevolare la creazione di visualizzazioni di dati interattive e grafici per le applicazioni web. Questa libreria offre un'ampia gamma di strumenti e opzioni che consentono agli sviluppatori di combinare facilmente capacità di visualizzazione avanzate all'interno dei loro siti web o applicazioni.

Una delle caratteristiche distintive di Google Chart è la sua facilità d'utilizzo. Gli sviluppatori possono sfruttare API semplici ed intuitive per popolare i grafici con dati dinamici provenienti da fonti come database, servizi web o dati in tempo reale. Questa libreria è anche altamente personalizzabile, offrendo agli sviluppatori il controllo completo della presentazione dei dati.

Un altro punto di forza di Google Chart è la sua integrazione fluida con altre tecnologie web di Google, tra cui Google Sheets e Google Maps, semplificando notevolmente l'importazione e la condivisione di dati tra queste applicazioni.

Nel contesto dello sviluppo web, Google Chart è largamente adottato per migliorare l'esperienza utente, rendere i dati più comprensibili e prendere decisioni basate sui dati.

2.1.1.3 Tableau

Tableau è una rinomata piattaforma di visualizzazione dei dati e di Business Intelligence che consente alle organizzazioni di esplorare, analizzare e comunicare in modo efficace i propri dati.

La forza di Tableau risiede nella sua facilità d'uso e nell'approccio "drag-and-drop". Gli utenti, anche se privi di competenze tecniche avanzate, possono facilmente importare dati da una varietà di fonti, come database, fogli di calcolo e servizi cloud. Una volta importati i dati, gli utenti possono generare rapidamente grafici, tabelle e mappe interattive per esplorare e analizzare le informazioni.

Tableau offre una vasta selezione di strumenti di analisi dei dati, tra cui la capacità di combinare dati da fonti diverse, quella di creare calcoli personalizzati e quella di utilizzare funzioni statistiche avanzate per rivelare tendenze e modelli nascosti nei dati. In aggiunta, Tableau offre la possibilità di connettersi ai dati in modo immediato, permettendo agli utenti di monitorare in tempo reale le prestazioni aziendali.

Tableau è ampiamente utilizzato in molteplici settori, tra cui aziende, istituzioni governative, istituti accademici e organizzazioni senza scopo di lucro, per supportare le decisioni basate sui dati, identificare opportunità di miglioramento e comunicare in modo efficace le informazioni critiche.

2.1.1.4 Zoho Analytics

Zoho Analytics è una soluzione avanzata di Business Intelligence (BI) e di analisi dei dati sviluppata da Zoho Corporation. Questa piattaforma offre alle organizzazioni uno strumento completo per acquisire, elaborare, analizzare e visualizzare dati provenienti da diverse fonti.

Zoho Analytics presenta la capacità di connettersi a una vasta gamma di fonti di dati, tra cui database, file CSV, servizi cloud e applicazioni aziendali. Gli utenti possono importare dati in modo intuitivo e flessibile e, quindi, elaborarli.

La piattaforma offre strumenti di analisi avanzati, inclusa la creazione di calcoli personalizzati, l'elaborazione delle query e la modellazione dei dati. Ciò consente ad essi di esplorare i dati in profondità, identificare tendenze, individuare correlazioni e ottenere insight significativi.

La piattaforma è progettata per utenti non tecnici, il che significa che anche coloro che non hanno una formazione specifica in analisi dei dati possono utilizzarla in modo efficace, grazie agli strumenti drag-and-drop e alle interfacce intuitive.

Zoho Analytics supporta anche la condivisione e la collaborazione dei dati e dei report all'interno dell'organizzazione, consentendo ai team di lavorare insieme in modo efficace. Inoltre, offre opzioni flessibili di distribuzione, tra cui la pubblicazione su web, la condivisione via e-mail e l'integrazione con altre applicazioni aziendali.

2.1.1.5 DataWrapper

Datawrapper è una potente piattaforma online di visualizzazione dei dati progettata per semplificare la trasformazione dei dati in visualizzazioni informative. Questo strumento è ampiamente utilizzato da giornalisti, ricercatori, analisti e professionisti del settore per comunicare in modo efficace informazioni complesse attraverso grafici, mappe e diagrammi.

Anche questa piattaforma è accessibile a utenti di tutti i livelli di competenza tecnica, consentendo loro di importare facilmente dati da fogli di calcolo o altre fonti e, quindi, di creare rapidamente visualizzazioni personalizzate.

Una caratteristica interessante di Datawrapper è la possibilità di incorporare facilmente le visualizzazioni dei dati all'interno di siti web, blog o articoli. La piattaforma genera automaticamente il codice embedded, semplificando il processo di condivisione delle visualizzazioni con il pubblico desiderato.

Datawrapper è ampiamente utilizzato nell'ambito della comunicazione visiva delle informazioni. Aiuta a rendere i dati più accessibili e comprensibili per il pubblico, migliorando la qualità delle presentazioni e facilitando la narrazione dei dati.

2.1.1.6 Qlik Sens

Qlik Sense è una piattaforma di Business Intelligence (BI) sviluppata da Qlik, progettata per aiutare le organizzazioni a trarre insight significativi dai propri dati. Questa piattaforma offre un approccio innovativo alla visualizzazione e all'analisi dei dati, consentendo agli utenti di esplorare e interagire con i dati in modo del tutto intuitivo e dinamico.

Una delle particolarità distintive di Qlik Sense è la sua tecnologia di associazione dei dati. A differenza di molte altre soluzioni di BI, Qlik Sense non richiede la creazione di modelli dati rigidi o di cubi predefiniti, ossia strutture di dati multidimensionali predefinite con categorie e aggregazioni fisse. Invece, gli utenti possono caricare dati grezzi da varie fonti e Qlik Sense permette di generare automaticamente un modello associativo, consentendo agli utenti di navigare liberamente attraverso i dati senza restrizioni predefinite.

Oltre alla semplice visualizzazione dei dati, Qlik Sense offre potenti funzionalità di analisi e ricerca; in particolare, è possibile eseguire ricerche associative per scoprire relazioni nascoste nei dati. La funzione "Associative Insights" di Qlik Sense aiuta a identificare automaticamente correlazioni nei dati, contribuendo a migliorare la comprensione dei pattern e delle tendenze.

Qlik Sense è utilizzato in una vasta gamma di settori e ambiti aziendali, tra cui vendite, marketing, finanza, risorse umane e operazioni, per supportare la presa di decisioni basate sui dati e migliorare le prestazioni aziendali complessive.

2.1.1.7 Looker

Looker è una piattaforma di Business Intelligence (BI) basata su cloud che offre agli utenti la possibilità di esplorare, analizzare e visualizzare i dati in modo efficace. Questa soluzione, acquisita da Google Cloud, è progettata per aiutare le organizzazioni a prendere decisioni basate sui dati e migliorare le operazioni aziendali.

Una delle caratteristiche chiave di Looker è il suo approccio basato su modelli dei dati. Gli utenti possono creare modelli dei dati personalizzati che rappresentano la struttura e la logica dei dati aziendali, consentendo una visione coerente e coesa dei dati in tutta l'organizzazione. Questi modelli possono essere condivisi e utilizzati per creare report, dashboard e visualizzazioni dei dati.

Looker offre un ambiente di sviluppo visuale e interattivo che consente agli utenti, anche senza una conoscenza avanzata della programmazione, di creare facilmente visualizzazioni dei dati personalizzate. La piattaforma supporta anche il linguaggio di interrogazione SQL, consentendo agli utenti più tecnici di eseguire query complesse sui dati.

La piattaforma promuove la condivisione e la collaborazione dei dati all'interno dell'organizzazione attraverso la possibilità di distribuire report e dashboard in modo sicuro e di incorporarli in altre applicazioni aziendali. Inoltre, Looker offre funzionalità avanzate di analisi dei dati, inclusa l'integrazione con strumenti di analisi statistica e la possibilità di creare calcoli personalizzati.

Looker è utilizzato in molteplici settori, tra cui vendite, marketing, finanza e operazioni, per ottenere insight dai dati e prendere decisioni basate sui dati in tempo reale.

2.1.1.8 Domo

Domo è una soluzione di Business Intelligence (BI) e di gestione dei dati basata su cloud progettata per operare in ambito aziendale. Questa piattaforma offre una serie di strumenti per la visualizzazione dei dati, l'analisi, la collaborazione e la condivisione delle informazioni.

Una delle caratteristiche principali di Domo è la sua capacità di aggregare e consolidare dati provenienti da diverse fonti in un unico ambiente centralizzato. Gli utenti possono connettersi a una vasta gamma di sorgenti dei dati, compresi database, applicazioni cloud, fogli di calcolo e altro ancora, per creare una vista unificata dei dati aziendali. Questo approccio aiuta a eliminare i silos di dati e a fornire una visione completa dell'andamento aziendale.

Domo offre un'ampia varietà di visualizzazioni dei dati e supporta anche l'approccio "drag and drop" per la creazione rapida di report personalizzati.

Un aspetto notevole di Domo è la sua capacità di condividere e collaborare sui dati in tempo reale. Così come con Looker, gli utenti possono distribuire report e dashboard in modo sicuro e consentire la collaborazione all'interno dell'organizzazione, facilitando la condivisione delle informazioni e il processo decisionale collaborativo.

Domo è utilizzato in una vasta gamma di settori e funzioni aziendali, tra cui vendite, marketing, finanza, risorse umane e operazioni, per supportare la gestione delle prestazioni, il monitoraggio delle metriche chiave e l'ottimizzazione delle operazioni aziendali.

2.1.1.9 Google Analytics

Google Analytics è una delle soluzioni di analisi web più ampiamente utilizzate e conosciute attualmente disponibili. Sviluppata da Google, questa potente piattaforma offre strumenti essenziali per la misurazione, l'analisi e la comprensione del traffico web e del comportamento degli utenti su siti web e app.

La chiave della grande diffusione di Google Analytics è la sua capacità di fornire una visione dettagliata delle attività degli utenti online. Il codice di tracciamento di Google Analytics è facilmente integrabile nel sito web o nell'app, consentendo al sistema di raccogliere dati su visite, visualizzazioni di pagina, tassi di conversione, tempi di permanenza e molto altro ancora. Questi dati vengono, quindi, elaborati e presentati in forma di report e dashboard, offrendo una panoramica completa delle prestazioni digitali.

Google Analytics fornisce informazioni preziose su diverse metriche, come la provenienza del traffico (da motori di ricerca, social media o siti di riferimento), il comportamento degli utenti (le pagine visitate, il tempo trascorso sul sito, le azioni compiute) e il monitoraggio delle conversioni (acquisti, iscrizioni, download, etc.). Questi dati aiutano i proprietari di siti web e le aziende a prendere decisioni informate per ottimizzare le loro strategie online.

Un altro punto di forza di Google Analytics è la sua ampia integrazione con altre soluzioni di Google, come Google Ads, Google Tag Manager e Google Data Studio.

Inoltre, Google Analytics fornisce una vasta disponibilità di funzionalità, da quelle gratuite e accessibili a tutti, agli strumenti avanzati disponibili su Google Analytics 360 per aziende di grandi dimensioni. Questa flessibilità rende Google Analytics un'opzione adatta a una vasta gamma di utenti, dalle piccole imprese agli editori online e alle aziende di e-commerce.

2.1.2 Motivazioni dell'utilizzo di Power BI

Nei capitoli successivi verrà analizzato un progetto di Data Analytics realizzato tramite l'utilizzo di Power BI. La scelta di utilizzare questo strumento è stata determinata da diversi fattori che lo rendono considerabilmente più conveniente rispetto ad altri.

Abbiamo precedentemente menzionato la sua facilità d'utilizzo; infatti, Power BI offre un'interfaccia utente intuitiva e strumenti di trascinamento e rilascio. Queste funzionalità rendono la creazione di visualizzazioni dei dati e report accessibile non solo agli utenti tecnici, ma anche a coloro che si stanno avvicinando per la prima volta agli strumenti di Business Intelligence. Tuttavia, anche altri strumenti, come Tableau, godono di questa proprietà. La distinzione di Power BI risiede nella sua capacità di offrire funzionalità avanzate senza alcuna limitazione di potenza, permettendo agli utenti esperti di sfruttare appieno queste caratteristiche per estendere la loro analisi dei dati.

In aggiunta a ciò, Power BI consente di modellare i dati in quanto offre strumenti appositi, tra i quali annoveriamo la capacità di creare relazioni complesse tra tabelle, definire misure personalizzate e utilizzare DAX (Data Analysis Expressions) per calcoli avanzati. Queste funzionalità sono fondamentali per l'analisi approfondita dei dati. In aggiunta a ciò Power BI garantisce anche la possibilità di integrare codice SQL e legarsi a database relazionali, caratteristica che ritroviamo anche in altri competitor (come Looker).

L'analisi è poi facilitata sia dall'integrazione nativa con altre applicazioni Microsoft (come Excel, Azure, SharePoint e Dynamics 365), che semplifica notevolmente il flusso di lavoro e la gestione dati, sia dal fatto che Power BI supporta la connessione con un'ampia gamma di fonti; questa flessibilità nella gestione delle fonti dati consente di lavorare con dati provenienti da molteplici origini

2.1.3 Power BI nell'ambito lavorativo

Power BI è spesso annoverato tra i tool migliori da utilizzare nell'ambito della Data Visualization e, questa nomea ha permesso un largo sfruttamento anche in ambito aziendale (come espresso nel sottoparagrafo dedicato). Soffermandoci sui settori in cui Power BI trova applicazione, questi includono:

1. *Settore finanziario*: Power BI viene impiegato per l'analisi dei dati finanziari, la gestione del rischio, la generazione di report relativi ai portafogli e la valutazione delle performance degli investimenti da parte delle istituzioni finanziarie.

2. *Vendite e marketing*: Power BI è un alleato prezioso per il monitoraggio delle metriche di vendita, l'analisi del comportamento dei clienti, la valutazione dell'efficacia delle campagne di marketing e l'ottimizzazione delle strategie di vendita.

3. *Sanità*: Power BI svolge un ruolo chiave nell'analisi dei dati clinici, nella gestione delle prestazioni degli ospedali, nella valutazione delle cure dei pazienti e nella generazione di report sulle informazioni sanitarie.

4. *Produzione e logistica*:le aziende manifatturiere fanno affidamento su Power BI per monitorare le catene di approvvigionamento, ottimizzare la produzione, gestire l'inventario e analizzare i dati relativi alla produzione.

5. *Risorse umane*: il settore delle risorse umane utilizza Power BI per supervisionare le performance dei dipendenti e valutarne il livello di soddisfazione per analizzare i dati relativi al personale e per pianificare le risorse umane.

6. *Istruzione*: nel campo dell'istruzione, Power BI è uno strumento cruciale per l'analisi dei dati degli studenti, la valutazione delle prestazioni scolastiche, il monitoraggio delle tendenze nell'istruzione e la generazione di report utili per le decisioni accademiche.

7. *Governo*: le organizzazioni governative sfruttano Power BI per l'analisi dei dati dei cittadini, la generazione di report sui servizi pubblici, la valutazione delle politiche pubbliche e il monitoraggio delle metriche governative.

8. *E-commerce*: nel settore dell'e-commerce, Power BI è sfruttato per l'analisi delle metriche di vendita online, il monitoraggio del comportamento degli acquirenti e l'ottimizzazione delle strategie di vendita online.

9. *Media e intrattenimento*: le aziende di media e intrattenimento si avvalgono di Power BI per analizzare i dati di audience, valutare le performance dei contenuti e monitorare le metriche pubblicitarie.

10. *Tecnologia e IT*: nel campo della tecnologia, Power BI è uno strumento essenziale per l'analisi dei dati IT, il monitoraggio delle prestazioni dei sistemi, la generazione di report sulla sicurezza informatica e la gestione delle risorse IT.

2.2 Componenti di Power BI

2.2.1 Architettura

Power BI è costituito da un insieme di componenti chiave che operano sinergicamente per fornire una piattaforma completa di Business Intelligence. In Figura 2.1 viene mostrata l'architettura di Power BI

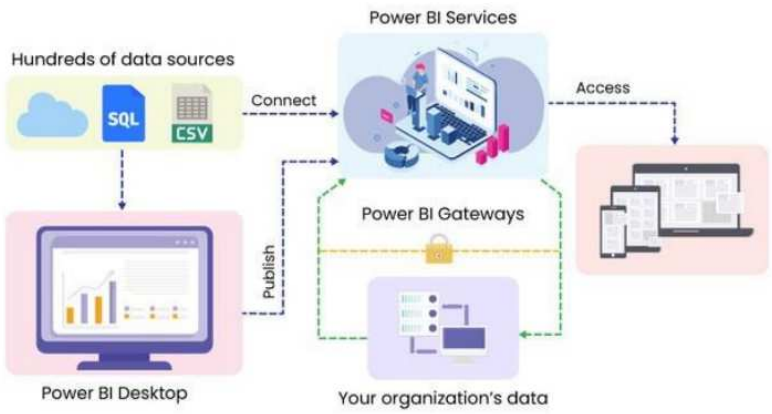

Figura 2.1: Architettura Power BI

I principali componenti di Power BI includono:

1. *Power BI Desktop*: ovvero l'applicazione Desktop di Power BI che consente la generazione di oggetti visivi e raccolte. Essa agevola il processo di raccolta di dati, in quanto consente di connettersi a più sorgenti e combinarle, così come il processo di trasformazione dei dati. Infine, è grazie a questo servizio che è possibile implementare la condivisione dei report.

2. *Power BI Service*: rappresenta il SaaS (Software as a Service) basato sui servizi online, ed è una piattafroma su Cloud ideata da Microsoft.

3. *Power BI Mobile*: è l'applicazione mobile di Power BI che permette agli utenti di usufruire dei dati in mobilità; disponibile sia per Android che per iOS.

4. *Power BI Gateway*: è un componente sviluppato da Microsoft che consente di connettere Power BI a diverse fonti di dati locali o basate su cloud. Esso è fondamentale quando si desidera ottenere l'accesso a dati aziendali che risiedono all'interno della rete locale o su altre piattaforme cloud e di renderli disponibili per la creazione di report, dashboard e analisi all'interno di Power BI.

5. *Power BI Embedded*: tale componente offre una capacità di integrazione avanzata, permettendo di incorporare report e visualizzazioni di Power BI direttamente nelle applicazioni o nei siti web.

6. *Power BI Dataflows*: un componente cardine per il corretto utilizzo di Power BI è la capacità di eseguire l'elaborazione dei dati in modo scalabile e ripetibile prima che gli stessi vengano importati in Power BI. In questo modo, la gestione e l'elaborazione dei dati mostrano un miglioramento esponenziale.

7. *Power Query*: è uno strumento di trasformazione dei dati utilizzato sia in Power BI Desktop che in Excel per importare, trasformare e combinare dati da diverse fonti.

2.2.2 Flusso di azioni

Nel contesto dell'impiego di Power BI, emergono passaggi fondamentali e ampiamente diffusi che consentono agli utenti di lavorare in modo efficiente con i dati, dall'acquisizione iniziale fino alla loro rappresentazione visuale. Questo flusso di lavoro comprende una serie di fasi, tra cui l'acquisizione dei dati, la progettazione e la creazione del modello di dati, la generazione di modelli visivi per illustrare le informazioni, la creazione di report che organizzano gli elementi visivi su una o più pagine, e la condivisione dei report con altri utenti tramite il servizio Power BI.

Nel prosieguo, ci concentreremo sull'analisi dettagliata della progettazione e creazione del modello di dati, sulla generazione di modelli visivi e sulla stesura dei report. Questi aspetti saranno approfonditi nei sottoparagrafi successivi.

2.2.2.1 Progettazione e creazione del modello dei dati

Tale passaggio, previa acquisizione ed estrazione dei dati interessati, consiste nel generare un modello che possa rappresentare correttamente la struttura e la logica dei dati. È un passaggio cruciale in quanto consente la creazione di relazioni tra le tabelle, di definire misure personalizzate (tramite formule DAX) e di preparare i dati per l'analisi.

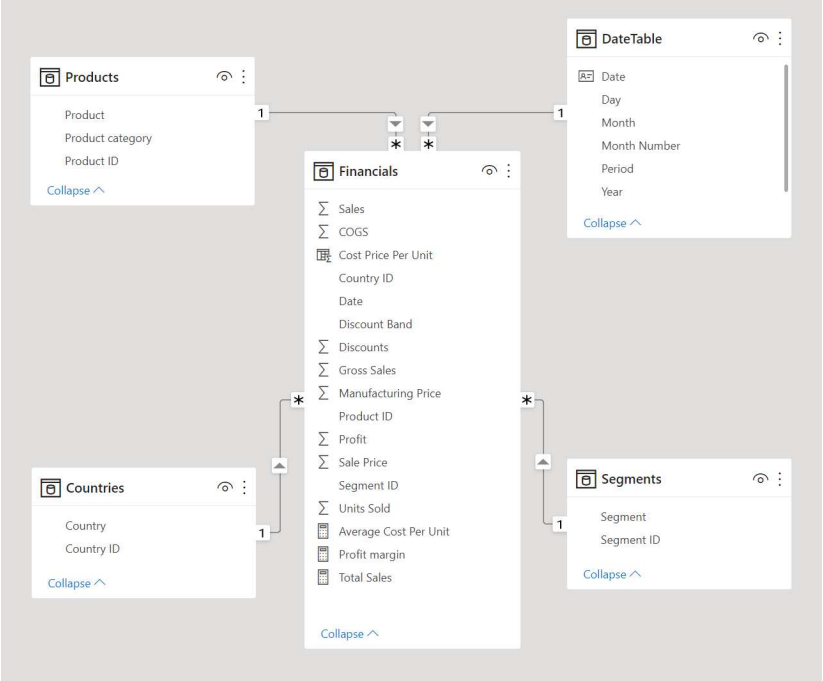

Figura 2.2: Esempio di un modello dati

2.2.2.2 Creazione dei modelli visivi

Questa fase implica la trasformazione dei dati dal modello dati in visualizzazioni significative e comprensibili. Power BI mette a disposizione un'ampia selezione di opzioni, tra cui grafici a barre, grafici a torta, mappe e diagrammi a dispersione. Ogni grafico generato può essere ulteriormente personalizzato, dalla scelta dei colori ai titoli.

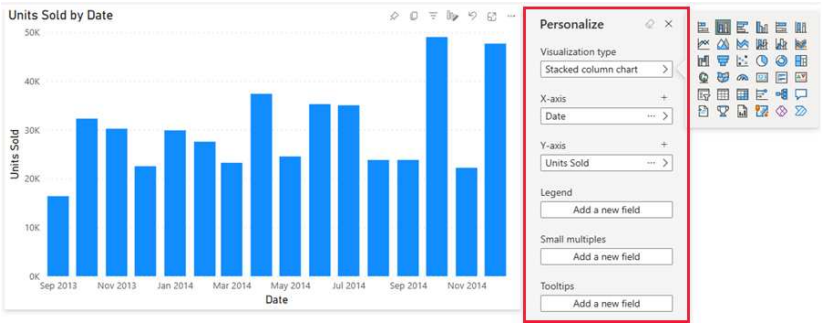

Figura 2.3: Esempio di un modello visivo

2.2.2.3 Creazione report e dashboard

Una volta generate le visualizzazioni necessarie, esse possono essere incorporate all'interno di un report informativo o di una dashboard interattiva. Tali strumenti permettono una fruizione chiara dei dati analizzati e delle conclusioni raggiunte, permettendo agli utenti di applicare filtri e condurre un'analisi dettagliata.

Figura 2.4: Esempio di dashboard

2.3 Funzionalità di Power BI

2.3.1 Uso dei linguaggi di porgrammazione

Power BI è progettato per essere altamente interoperabile con i linguaggi di programmazione, consentendo agli sviluppatori di estendere le funzionalità della piattaforma ed integrarla. Alcuni dei linguaggi utilizzabili sono:

- *Data Analysis eXpressions (DAX)*: è un linguaggio utilizzato all'interno di Power BI per consentire agli utenti di creare misure personalizzate e condurre analisi avanzate sui dati. Questo linguaggio include una libreria con oltre 200 funzioni, operatori e costrutti che offrono un'ampia flessibilità. Grazie a DAX, è possibile eseguire calcoli basati sulle colonne, manipolare testi e condurre calcoli temporali.

- *Python*: Power BI offre una completa integrazione con Python, consentendo agli utenti di incorporare script Python direttamente nei loro report e modelli dati. Questa sinergia tra Power BI e Python permette di sfruttare le librerie e le capacità di analisi avanzate di Python all'interno dell'ecosistema di Power BI. Gli utenti possono utilizzare Python per eseguire analisi statistiche complesse, creare visualizzazioni personalizzate o implementare algoritmi di Machine Learning per ottenere insight più approfonditi dai dati.

- *R*: l'integrazione di R in Power BI rappresenta una risorsa fondamentale, in quanto consente agli sviluppatori di accedere ad un'ampia gamma di analisi statistiche e di data science avanzate. Questo collegamento tra Power BI e R consente di sfruttare le librerie di R per eseguire analisi complesse dei dati, creare visualizzazioni personalizzate e sviluppare modelli predittivi basati su algoritmi di Machine Learning. L'uso di R in Power BI semplifica il lavoro con analisi avanzate, come l'analisi delle serie temporali, la segmentazione dei clienti o la previsione delle vendite.

- *M*: è un linguaggio di programmazione utilizzato all'interno di PowerQuery. Il suo utilizzo è cruciale per la preparazione dei dati prima dell'importazione nei modelli dati e nella creazione dei report. La potenza di M risiede nella sua capacità di eseguire trasformazioni avanzate sui dati, come l'unione di diverse fonti dati, la rimozione dei duplicati, la pulizia dei dati sporchi e la creazione di colonne personalizzate.

- *SQL*: l'utilizzo di SQL in Power BI rappresenta una solida integrazione che consente agli utenti di connettersi direttamente ai loro database relazionali e acquisire dati in modo efficiente. Grazie a questa funzionalità, è possibile scrivere query SQL personalizzate per estrarre dati specifici da sorgenti di dati, come database SQL Server, MySQL, Oracle e molti altri. Lo sfruttamento di SQL semplifica notevolmente il processo di acquisizione

e trasformazione dei dati, consentendo agli utenti di creare modelli dati complessi e report avanzati basati su dati aggiornati. Inoltre, SQL può essere integrato nelle formule DAX per effettuare ricerche e calcoli avanzati su dati provenienti da fonti SQL.

- *JavaScript*: la piattaforma Power BI consente agli sviluppatori di incorporare script JavaScript nei loro progetti, consentendo la creazione di visualizzazioni personalizzate, l'aggiunta di funzionalità interattive e l'integrazione di report in applicazioni web esterne. L'uso di JavaScript nella piattaforma è, quindi, particolarmente vantaggioso per gli sviluppatori e gli utenti che cercano di adattarla alle loro esigenze specifiche e di migliorare l'esperienza di visualizzazione e condivisione dei dati aziendali.

- *API e SDK*: l'integrazione delle API e degli SDK (Software Development Kit) in Power BI rappresenta un metodo aggiuntivo, messo a disposizione degli sviluppatori, per personalizzare ed estendere le funzionalità della piattaforma, adattandole alle specifiche esigenze aziendali. Le API di Power BI consentono l'automatizzazione dei flussi di lavoro, la gestione di report e dashboard e l'interazione con i dati in modo personalizzato. Questa opportunità apre le porte a una serie di possibilità, tra cui l'automazione nella creazione di report, l'aggiornamento dei dati in tempo reale, l'accesso ai metadati dei report e la gestione degli utenti e delle autorizzazioni. Gli SDK semplificano l'integrazione di Power BI in applicazioni personalizzate, indipendentemente dalla piattaforma su cui sono basate. Grazie agli SDK, gli sviluppatori possono creare applicazioni personalizzate basate su Power BI, semplificando il processo di integrazione delle funzionalità di reporting e analisi nei propri strumenti aziendali.

CAPITOLO 3

Descrizione dei dati a disposizione ed attività di ETL

In questo capitolo analizzeremo le varie tipologie di dati per poi proseguire con una descrizione del dataset Fashion E-commerce, fino alla presentazione delle attività preliminari svolte sui dati

3.1 Tipologie di dati

I dati costituiscono le fondamenta di un qualsiasi progetto di Data Analysis e possono provenire dalle più disparate fonti, come sistemi interni di un'organizzazione, sensori, social media, applicazioni mobile o sondaggi; possono, inoltre, differenziarsi in base alla loro rappresentazione, nel formato e nel tipo.

Le soluzioni di Big Data coinvolgono l'elaborazione di una varietà di tipi di dati, tra cui:

- *dati strutturati*;
- *dati non strutturati*;
- *dati semi-strutturati*.

Oltre a queste tre tipologie, i metadati svolgono un ruolo cruciale nell'ambito dell'organizzazione dei dati e sono spesso definiti come 'dati sui dati'. Essi sono essenziali per le soluzioni di Big Data e sono comunemente conosciuti come formati dei dati.

Analizzeremo nel dettaglio le diverse tipologie di dati nelle prossime sottosezioni.

3.1.1 Dati strutturati

I dati strutturati (in inglese, structured data) sono dati che utilizzano un formato predefinito e previsto. Essi vengono formattati in una struttura impostata prima che essi vengano inseriti nell'archivio di dati, che viene spesso definito schema-on-write; essi risultano, dunque, altamente organizzati e seguono uno schema ben definito. Questo modello predeterminato agevola l'inserimento, l'esecuzione di query e l'analisi, oltre che l'identificazione di pattern, da parte degli algoritmi di apprendimento.

Il miglior esempio di dati strutturati è il database relazionale; in esso i dati sono stati formattati in campi definiti con precisione, come numeri di carta di credito o indirizzi, per poter essere facilmente interrogati tramite SQL.

Alcuni degli usi più comuni nel business includono moduli BRM, transazioni online, dati di azioni, dati di monitoraggio della rete aziendale e moduli Web.

L'utilizzo dei dati strutturati offre una serie di vantaggi significativi, ovvero:

- *Facilità di gestione ed organizzazione*: i dati strutturati, che sono organizzati in tabelle con chiare colonne e righe, semplificano notevolmente la ricerca, l'analisi e l'elaborazione dei dati.
- *Query efficienti*: grazie alla loro struttura ben definita, i dati strutturati si prestano all'esecuzione di query SQL e all'analisi statistica; tali dari sono, quindi, ideali per applicazioni come database e sistemi di gestione dell'informazione.
- *Affidabilità*: i dati strutturati sono generalmente più coerenti, in quanto seguono uno schema definito, il che contribuisce alla loro affidabilità.

Tuttavia, è importante notare che l'utilizzo di dati strutturati comporta anche alcuni svantaggi, tra cui:

- *Limitazione della rappresentazione*: questi dati non sono adatti per rappresentare informazioni complesse o non numeriche, come il testo libero o le immagini, limitando la loro capacità di catturare dati eterogenei.
- *Mancanza di flessibilità*: qualsiasi modifica all'organizzazione dei dati richiede spesso cambiamenti significativi nello schema del database, il che può risultare oneroso e complesso.
- *Inadeguatezza per dati non standard*: i dati strutturati non si adattano bene a dati in continua evoluzione o a dati che non seguono uno standard predefinito, limitando la loro applicabilità in contesti dinamici.

3.1.2 Dati non strutturati

I dati non strutturati (in inglese unstructured data) sono dati che mancano di una definizione. Vengono archiviati nel loro formato nativo e non elaborati fino a quando non vengono utilizzati; tale formato è noto come schema-on-read.

Questi dati mancano di una struttura organizzata e si possono presentare in numerosi formati, tra cui e-mail, post sui social media, presentazioni, chat, dati dei sensori IoT ed immagini satellitari. Data la vasta gamma di formati che comprendono i dati non strutturati, spesso questo tipo di dati costituisce circa l'80% dei dati di un'organizzazione.

I dati non strutturati vengono generalmente memorizzati in DBMS NoSQL. NoSQL sta per "Not Only SQL", e indica che il database può gestire una gamma più ampia di dati con funzionalità superiori a quelle dei database SQL. I database NoSQL non hanno uno schema o una struttura tabulare; si tratta di una raccolta di dati raggruppati insieme.

Volendo nominare i vantaggi derivanti dall'uso dei dati non strutturati troviamo:

- *Ricchezza delle informazioni*: i dati non strutturati abbracciano una vasta gamma di formati, che li rendono in grado di contenere informazioni dettagliate ed eterogenee. Questa caratteristica li rende particolarmente preziosi per l'acquisizione di informazioni dettagliate su una varietà di soggetti.
- *Flessibilità*: i dati non strutturati sono altamente flessibili poiché possono rappresentare dati in molti modi differenti e possono adattarsi agilmente a nuovi contenuti o contesti. Questa flessibilità li rende ideali per situazioni in cui i dati possono evolvere nel tempo o non essere completamente prevedibili.

- *Adattabilità per l'apprendimento automatico*: questi dati costituiscono una risorsa preziosa nell'addestramento di modelli di Machine Learning, soprattutto in ambiti come l'elaborazione del linguaggio naturale e la visione artificiale, dove la comprensione di dati non strutturati è essenziale.

Anche i dati non strutturati comportano una serie di svantaggi, come:

- *Complessità*: l'elaborazione dei dati non strutturati spesso richiede l'applicazione di tecniche avanzate e può essere computazionalmente intensiva, aumentando i requisiti di risorse e tempo necessari per l'analisi.
- *Difficoltà nell'organizzazione*: la mancanza di una struttura predefinita nei dati non strutturati può rendere complessa l'organizzazione e la ricerca di informazioni specifiche, richiedendo strumenti sofisticati e competenze specializzate.
- *Interpretazione soggettiva*: l'interpretazione dei dati non strutturati è spesso soggettiva, il che può comportare ambiguità o errori. L'interpretazione dei contenuti, come il testo o le immagini, può variare da individuo a individuo.

Nella Figura 3.1 viene fornito un confronto esaustivo tra i dati strutturati e quelli non strutturati per evidenziarne le caratteristiche chiave e le implicazioni nell'ambito dell'analisi dei dati.

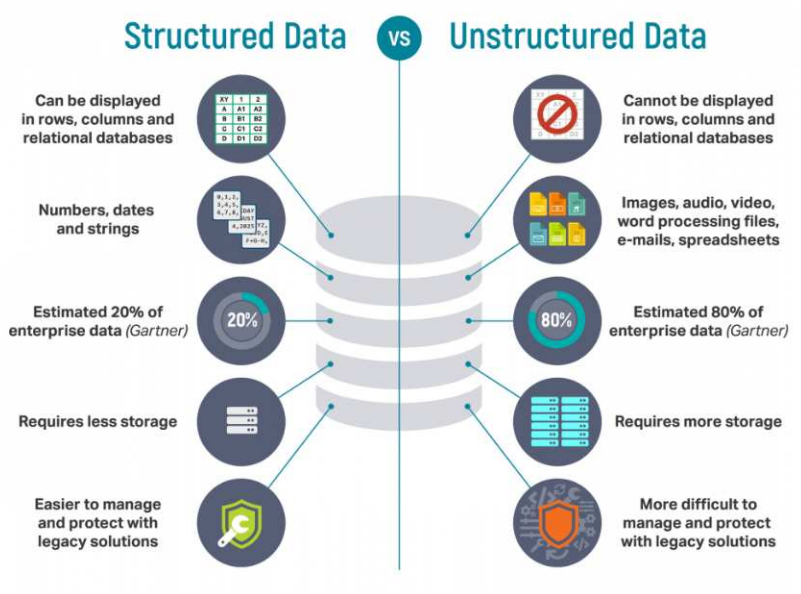

Figura 3.1: Confronto tra dati strutturati e non strutturati

3.1.3 Dati semi-strutturati

I dati semi-strutturati (in inglese semi-structured data) sono una forma di dato strutturato che non è conforme alla struttura formale dei modelli di dati associati con le basi di dati relazionali o altre forme di tabelle dati, ma che hanno anche metadati che identificano determinate caratteristiche, utilizzando etichette o tag in modo da separare elementi semantici e rafforzare le gerarchie di record e campi all'interno del dato. I metadati contengono informazioni sufficienti per consentire ai dati di essere catalogati, cercati e analizzati in modo più efficiente rispetto ai dati strettamente non strutturati. I dati semistrutturati rappresentano una

categoria di dati che si colloca a metà tra i dati strutturati ed quelli non strutturati; pertanto, non seguono uno schema rigido ma mantengono una certa forma di struttura.

Tale tipologia di dati risulta fondamentale per le applicazioni di Business Intelligence, l'automazione dei processi aziendali e l'analisi dei trend. Inoltre, essi giocano un ruolo essenziale nello sviluppo di tecnologie come l'Intelligenza Artificiale e il Machine Learning

Un buon esempio di dati semi-strutturati rispetto a dati strutturati sarebbe un file delimitato da tabulazioni contenente i dati dei clienti rispetto a un database contenente tabelle CRM (Customer Relationship Management). Notiamo come i dati semi-strutturati presentino una maggiore gerarchia rispetto ai dati non strutturati, in quanto il file delimitato da tabulazioni è più specifico di un elenco di commenti da Facebook di un cliente.

Nel panorama dell'analisi dei dati, l'utilizzo dei dati semi-strutturati offre una serie di vantaggi distintivi, tra cui:

- *Combinazione di struttura e flessibilità*: questi dati presentano un equilibrio unico tra una certa struttura e una notevole flessibilità. Possono rappresentare dati con un certo grado di organizzazione, ma, al contempo, accogliere elementi non strutturati o variabili, offrendo, così, una soluzione flessibile per la rappresentazione dei dati.

- *Adattabilità*: i dati semi-strutturati sono particolarmente idonei per scenari in cui i dati possono subire modifiche nel tempo o non seguire uno schema completamente prevedibile. La loro capacità di adattarsi alle fluttuazioni nei dati è un attributo prezioso.

- *Ampie possibilità di applicazione*: questi dati trovano applicazione in una varietà di contesti, tra cui l'archiviazione di dati in database NoSQL e formati come XML e JSON. La loro versatilità li rende adatti per molteplici casi d'uso.

Tuttavia, l'utilizzo di dati semi-strutturati comporta alcune sfide:

- *Complessità nelle elaborazioni*: l'elaborazione dei dati semi-strutturati può richiedere l'impiego di strumenti specifici e metodologie più complesse rispetto ai dati completamente strutturati. Questa complessità può richiedere risorse aggiuntive e competenze specializzate.

- *Rischio di ambiguità*: la flessibilità intrinseca dei dati semi-strutturati può portare a interpretazioni ambigue o inconsistenze nei dati, in quanto l'organizzazione non è rigidamente definita. Questa ambiguità può influire sulla precisione dell'analisi e richiede un'attenzione particolare nella gestione e nell'interpretazione dei dati.

3.1.4 Metadati

I metadati, spesso definiti come dati sui dati, sono informazioni di cui bisogna dotare il documento informatico, in quanto forniscono un contesto e una spiegazione sui dati, aiutando a comprenderne l'origine, la struttura, il significato e altri dettagli rilevanti. In particolare, vengono utilizzati per poter creare, gestire e conservare nel tempo, in maniera corretta, un documento informatico, in quanto esso necessita di essere posto in relazione ad un insieme di informazioni.

La centralità di tali dati nell'ambito della gestione delle risorse informative ha portato ad una distinzione di massima dei metadati in tre grandi categorie:

1. *Metadati descrittivi*: funzionali all'identificazione e al recupero dei documenti stessi; essi, inoltre, risultano costituiti da descrizioni normalizzate.

2. *Metadati amministrativi e gestionali*: utili alla gestione dei metadati stessi all'interno dell'archivio; e ne comprendono anche informazioni di natura tecnica.

3. *Metadati strutturali*: comprendono le informazioni necessarie a descrivere l'articolazione interna e le relazioni fra le parti che compongono gli oggetti digitali.

3.2 Fashion E-Commerce

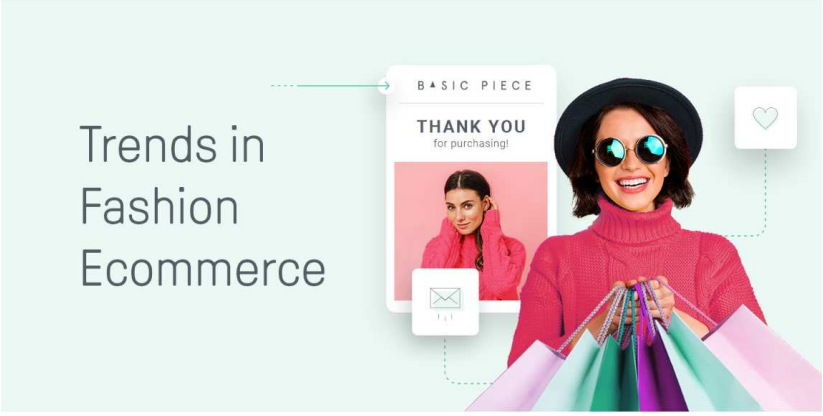

Figura 3.2: Immagine del dataset

I dati analizzati in questa tesi provengono da Kaggle, definito come:

> *Kaggle is a data science competition platform and online community of data scientists and machine learning practitioners under Google LLC. Kaggle enables users to find and publish datasets, explore and build models in a web-based data science environment, work with other data scientists and machine learning engineers, and enter competitions to solve data science challenges.*

Nell'ambito dell'ampia gamma di dataset ospitati sulla piattaforma Kaggle, il focus verrà posto su uno specifico noto come "Fashion E-Commerce User Data". Questo particolare dataset è stato creato per fornire informazioni dettagliate sugli utenti registrati su un sito web, specializzato nella compravendita di capi d'abbigliamento. L'obiettivo del nostro studio è esaminare attentamente i dati presenti in esso per estrarre informazioni di rilevanza e condurre analisi dettagliate, al fine di approfondire la comprensione di aspetti specifici legati al comportamento degli utenti all'interno di questo contesto.

Il dataset è formato da due tabelle denominate `ds1` e `ds2`, scaricabili in formato CSV (Comma- Separeted Values), i cui parametri verranno analizzati nei seguenti sottoparagrafi

3.2.1 ds1

Nella tabella iniziale, sono presentate le caratteristiche di ciascun utente registrato sul sito web oggetto di studio, fornendo un punto di partenza essenziale per l'analisi condotta.

Questa particolare tabella risulta formata da 98913 righe e 24 colonne, le quali contano indicatori come:

1. *identifierHash*: identificatore Hash univoco per ogni accesso;

2. *type*: tipologia di utente;

3. *country*: paese di provenienza dell'utente;
4. *language*: lingua di preferenza dell'utente;
5. *socialNbFollowers*: numero dei followers sui social media dell'utente;
6. *socialNbFollows*: numero di account che l'utente segue sui social media;
7. *socialProductLiked*: numero dei prodotti online piaciuti all'utente;
8. *productListed*: numero dei prodotti messi nella lista dei desideri dall'utente;
9. *productSold*: numero di prodotti venduti dall'utente;
10. *productPassRate*: livello di qualità dei prodotti dell'utente;
11. *productWished*: numero dei prodotti desiderati dall'utente;
12. *productBought*: numero dei prodotti acquistati dall'utente;
13. *gender*: sesso dell'utente;
14. *civilityGenderId*: id corrispondente al sesso dell'utente;
15. *civilityTitle*: titolo di civiltà dell'utente, ovvero "miss", "mrs" e "mr";
16. *hasAnyApp*: indica se l'utente possiede una qualunque app per telefono;
17. *hasAndoirdApp*: indica se l'utente possiede una qualsiasi app Android;
18. *hasIosApp*: indica se l'utente ha una qualsiasi app IOS;
19. *hasProfilePicture*: indica se l'utente ha un'immagine del profilo;
20. *daysSinceLastLogin*: numero di giorni passati dall'ultimo login dell'utente;
21. *seniority*: anzianità dell'utente espressa in giorni;
22. *seniorityAsMonths*: anzianità dell'utente espressa in mesi;
23. *seniorityAsYears*: anzianità dell'utente espressa in anni;
24. *countryCode*: codice identificativo del paese di provenienza.

3.2.2 ds2

Nella seconda tabella, si pone l'attenzione sui paesi di origine degli utenti, con l'obiettivo di mettere in evidenza le abitudini e le caratteristiche distintive di ciascun gruppo di utenti in base alla loro provenienza. Questa tabella è stata progettata per approfondire l'analisi delle differenze culturali e comportamentali tra gli utenti registrati, contribuendo così a una comprensione più approfondita del contesto e delle dinamiche del sito web in esame.

Questa particolare tabella presenta 62 righe e 32 colonne, le quali contano parametri come:

1. *country*: nome del paese;
2. *buyers*: numero totale di acquirenti da quel paese;
3. *topbuyers*: numero dei migliori acquirenti di quel paese;

4. *topbuyerratio*: rapporto di migliori acquirenti di quel paese;
5. *femalebuyers*: numero di acquirenti donne in quel paese;
6. *malebuyers*: numero di acquirenti uomini di quel paese;
7. *topfemalebuyers*: numero delle migliori acquirenti donne di quel paese;
8. *topmalebuyers*: numero dei migliori acquirenti uomini di quel paese;
9. *femalebuyersratio*: rapporto di acquirenti donne di quel paese;
10. *topfemalebuyersratio*: rapporto di migliori acquirenti donne di quel paese;
11. *boughtperwishlistratio*: rapporto dei prodotti acquistati in quel paese provenienti dalle wishlist;
12. *boughtperlikeratio*: rapporto di prodotti acquistati rispetto a quelli piaciuti in quel paese;
13. *topboughtperwishlistratio*: rapporto di prodotti acquistati rispetto ai prodotti inseriti nelle wishlist in quel paese dai migliori acquirenti;
14. *topboughtperlikeratio*: rapporto di prodotti aquistati ripetto ai prodotti piaciuti in quel paese dai migliori acquirenti;
15. *totalproductsbought*: numero totale di prodotti acquistati in quel paese;
16. *totalproductswished*: numero totale di prodotti inseriti nelle wishlist in quel paese;
17. *totalproductsliked*: numero totale di prodotti piaciuti in quel paese;
18. *toptotalproductsbought*: numero totale di prodotti acquistati dai migliori acquirenti in quel paese;
19. *toptotalproductswished*: numero totale di prodotti inseriti nelle wishlist dei migliori acquirenti di quel paese;
20. *toptotalproductsliked*: numero totale di prodotti piaciuti ai migliori acquirenti di quel paese;
21. *meanproductsbought*: numero medio di prodotti acquistati per acquirente nel paese;
22. *meanproductswished*: numero medio di prodotti inseriti nelle wishlist per gli acquirenti nel paese;
23. *meanproductsliked*: numero medio di prodotti piaciuti agli acquirenti in quel paese;
24. *topmeanproductsbought*: numero medio di prodotti acquistati per acquirenti per i migliori acquirenti in quel paese;
25. *topmeanproductswished*: numero medio di prodotti inseriti nella wishlist per i principali acquirenti nel paese;
26. *topmeanproductsliked*: numero medio di prodotti piaciuti per acquirente per i principali acquirenti nel paese;
27. *meanofflinedays*: numero medio di giorni offline per gli acquirenti di quel paese;

28. *topmeanofflinedays*: numero medio di giorni offline per i principali acquirenti di quel paese;
29. *meanfollowers*: numero medio di follower sui social media per gli acquirenti nel paese;
30. *meanfollowing*: numero medio di persone seguite dagli acquirenti di quel paese;
31. *topmeanfollowers*: numero medio di follower sui social media per i principali acquirenti nel paese;
32. *topmeanfollowing*: numero medio di account seguiti sui social media dai principale acquirenti di quel paese.

3.3 Attività di ETL

Il termine ETL (Extract, Transform, and Load) fa riferimento al processo fondamentale di acquisizione, trasformazione e caricamento dei dati all'interno di un sistema informativo. Queste operazioni sono cruciali per la gestione efficiente dei dati, specialmente quando si tratta di dati provenienti da diverse fonti. L'ETL svolge un ruolo critico nell'organizzazione e nella preparazione dei dati per le analisi future.

Le tre fasi principali dell'ETL, cioè l'estrazione, la trasformazione ed il caricamento, rappresentano i pilastri di questo processo. Nel contesto di questo studio, esamineremo dettagliatamente come queste specifiche fasi sono state implementate nei prossimi sottoparagrafi, e sarà fondamentale evidenziare le metodologie e le procedure specifiche utilizzate.

3.3.1 Estrazione

L'estrazione dei dati rappresenta la prima fase fondamentale del processo di ETL. Durante questa fase l'attenzione è rivolta all'acquisizione di dati da una vasta gamma di fonti, che possono includere database relazionali, file strutturati o non strutturati, servizi web, applicazioni aziendali e molto altro. L'obiettivo principale è raccogliere dati grezzi in modo efficace ed efficiente, tenendo conto delle esigenze specifiche del progetto. Questa operazione può coinvolgere sia l'estrazione di dati completi da una fonte che la selezione di dati specifici in base a criteri definiti. L'efficacia dell'estrazione dei dati è fondamentale, in quanto determina la quantità e la qualità dei dati disponibili per le fasi successive di trasformazione e caricamento. L'automazione e la pianificazione delle operazioni di estrazione sono spesso utilizzate per garantire una raccolta dati efficiente e regolare, consentendo di mantenere aggiornate le informazioni in un repository centralizzato per l'analisi e l'elaborazione.

L'estrazione può essere svolta in diversi modi; in particolare, è possibile:

- scrivere programmi appositi;
- utilizzare uno dei numerosi tool di ETL presenti sul mercato;
- utilizzare entrambe le soluzioni.

L'approccio specifico all'estrazione dei dati dipenderà dalla natura delle fonti e dagli strumenti disponibili. È essenziale pianificare attentamente il processo di estrazione per garantire che i dati siano acquisiti in modo accurato, completo e coerente, e che siano pronti per le fasi successive di trasformazione e caricamento. La scelta delle tecniche e degli strumenti appropriati è fondamentale per il successo del progetto di ETL.

3.3.1.1 Come è stato implementato

Power BI Desktop si distingue per la sua capacità di semplificare la fase di estrazione in quanto offre diverse opzioni per acquisire i dati. Inizialmente, dopo aver importato il dataset *Fashion E-Commerce* da Kaggle, è possibile effettuare il recupero dei dati direttamente all'apertura dell'applicazione, visualizzando quanto riportato in Figura 3.3.

Figura 3.3: Schermata di Power BI per la selezione delle sorgenti di dati

In alternativa, è possibile farlo utilizzando la barra superiore di Power BI, come illustrato nella Figura 3.4.

Figura 3.4: Barra superiore di Power BI per il recupero dei dati da diverse sorgenti

In entrambi i casi, tale processo ci porterà alla schermata di *Recupero Dati*, come mostrato nella Figura 3.5.

Questa schermata offre la flessibilità di selezionare la fonte di estrazione dai vari tipi di sorgenti compatibili con Power BI. Una volta scelta la tipologia, nel nostro caso due file CSV, è possibile procedere con il comando *Connetti* per stabilire una connessione tra la fonte dati e Power BI. Questo approccio semplificato all'estrazione dei dati è particolarmente rilevante per la fase iniziale del processo ETL e può migliorare significativamente l'efficienza nel recupero dei dati da diverse fonti, fornendo una solida base per ulteriori trasformazioni e analisi.

3.3.2 Trasformazione

La fase di trasformazione nel processo di ETL è cruciale per la preparazione e l'elaborazione dei dati estratti prima che vengano caricati nel repository centrale. Durante questa fase, i dati vengono sottoposti a una serie di trasformazioni e manipolazioni che mirano a renderli coerenti, puliti e pronti per l'analisi. Queste trasformazioni possono includere la rimozione di dati duplicati, la gestione di valori mancanti, la standardizzazione dei formati dei dati, la creazione di nuove variabili derivanti da calcoli o logiche personalizzate, e la normalizzazione dei dati per garantirne la consistenza.

La trasformazione dei dati è spesso guidata da regole e criteri definiti all'interno del processo di ETL. Gli strumenti di ETL consentono di automatizzare molte di queste trasformazioni, garantendo la coerenza dei dati e risparmiando tempo ed errori potenziali. Inoltre, è possibile applicare logiche aziendali specifiche durante la fase di trasformazione per garantire che i dati siano adatti al contesto dell'analisi.

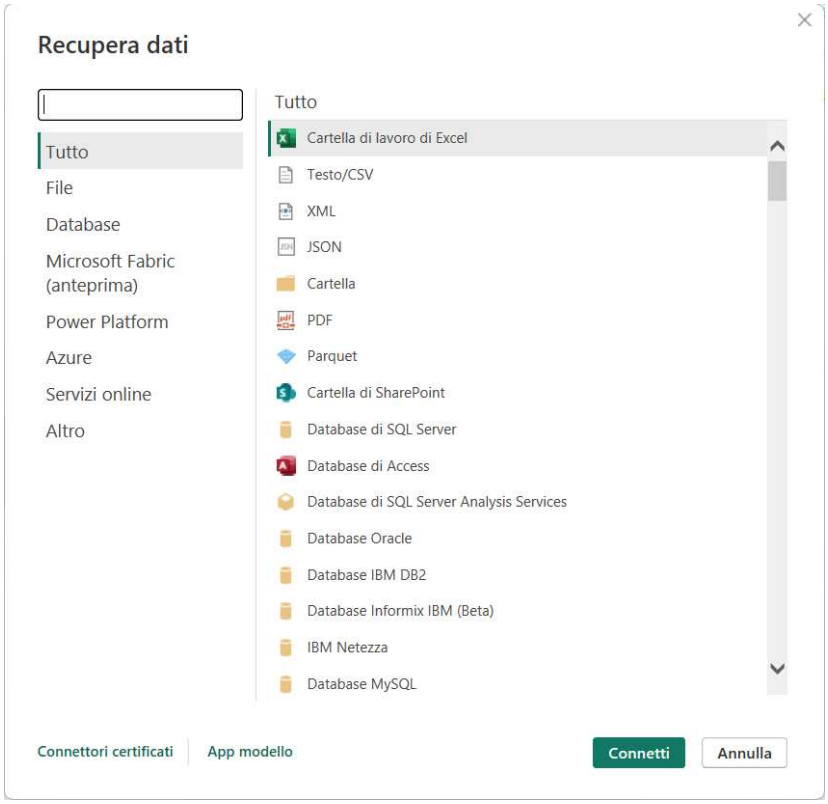

Figura 3.5: Schermata di Power BI per il recupero dei dati

3.3.2.1 Come è stato implementato

Durante il processo di trasformazione, ci dedichiamo principalmente sulla pulizia e sulla formattazione dei dati per renderli adatti all'analisi. Iniziamo verificando che tutti gli attributi, ovvero le colonne del dataset, siano espressi nel formato di dati corretto. Nel dataset esaminato, non si riscontrano problematiche di questo genere; pertanto, otteniamo quanto raffigurato in Figura 3.6.

Tuttavia, è emerso che i nomi dei paesi di provenienza degli utenti erano riportati in francese, a causa delle origini del dataset. La gestione di un'operazione complessa, come la traduzione, può essere fatta in Power BI utilizzando script come M o DAX. Per semplicità, abbiamo scelto di utilizzare uno script Python, per risolvere questa problematica all'esterno di Power BI, con l'ausilio di uno snippet denominato *My Memory*. Tale codice è illustrato nelle Figure 3.7, 3.8 e 3.9.

Con l'ausilio del codice Python ci occupiamo di tradurre gli elementi della colonna *country* dalla lingua originale (francese) in inglese. Successivamente, tutti i dati, compresi quelli tradotti, sono stati salvati in un nuovo file CSV chiamato `ds2_translated`. Dopo aver generato il nuovo file, sono state nuovamente applicate le procedure adottate nella fase di estrazione.

Ciononostante, una volta importato il file, ci si è accorti che alcuni paesi erano stati riportati in modo errato, come *Italy* segnato come *Observations of Italy*, *Austria* come *Observations of Austria*, *Ireland* come *Statement in intervention submitted by Ireland* ed *Estonia* come *Reply of Estonia*. Il risultato finale di una delle tabelle è riportato in Figura 3.10.

hasProfilePicture	daysSinceLastLogin	seniority	seniorityAsMonths	seniorityAsYears	countryCode
TRUE	709	3205	106,83	8,9	us
TRUE	709	3205	106,83	8,9	de
TRUE	689	3205	106,83	8,9	se
TRUE	709	3205	106,83	8,9	tr
TRUE	709	3205	106,83	8,9	fr
TRUE	709	3205	106,83	8,9	gb
TRUE	591	3205	106,83	8,9	gb
TRUE	709	3205	106,83	8,9	it
TRUE	701	3205	106,83	8,9	it
TRUE	703	3205	106,83	8,9	fr
TRUE	709	3205	106,83	8,9	us
TRUE	709	3205	106,83	8,9	es
TRUE	709	3205	106,83	8,9	us
TRUE	709	3205	106,83	8,9	gb
TRUE	709	3205	106,83	8,9	us
TRUE	709	3205	106,83	8,9	es
TRUE	558	3205	106,83	8,9	us
TRUE	42	3205	106,83	8,9	dk
TRUE	32	3205	106,83	8,9	us
TRUE	276	3205	106,83	8,9	fr
TRUE	708	3205	106,83	8,9	gb
TRUE	709	3205	106,83	8,9	us
TRUE	423	3205	106,83	8,9	de
TRUE	709	3205	106,83	8,9	fr
TRUE	603	3205	106,83	8,9	us
TRUE	709	3205	106,83	8,9	fr
TRUE	709	3205	106,83	8,9	re

Figura 3.6: Risultato dell'importazione dei dati

```python
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
import json
import requests

import csv

class SentimentManager:
    _email = 'aledanna3@gmail.com'
    _from_lang = 'fr'
    _to_lang = 'en'
    _headers = {
        'User-Agent': 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.11 (KHTML, like Gecko)'
                      ' Chrome/23.0.1271.64 Safari/537.11',
        'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q=0.8',
        'Accept-Charset': 'ISO-8859-1,utf-8;q=0.7,*;q=0.3',
        'Accept-Encoding': 'utf-8',
        'Accept-Language': 'en-US,en;q=0.8',
        'Connection': 'keep-alive'
    }
    @classmethod
    def extract_sentiment(cls, text):
        api_url = f'http://mymemory.translated.net/api/get?q={text}' \
                  f'&langpair={cls._from_lang}|{cls._to_lang}&de={cls._email}'
        response = requests.get(api_url, headers=cls._headers)
        response_json = json.loads(response.text)
        translation = response_json["responseData"]["translatedText"]

        return translation
```

Figura 3.7: Prima parte del codice Python per la traduzione della colonna "country"

```python
if __name__ == "__main__":
    country_name = []
    with open(r"C:\Users\aleda\OneDrive\Allegati\OneDrive\Desktop\TIROCINIO\ds1.csv", newline='', encoding="utf-8") as d1:
        reader = csv.DictReader(d1)
        for riga in reader:
            country_name.append(riga["country"])
        print(country_name)

    translated_countries = []
    for country in country_name:
        translated_country = SentimentManager.extract_sentiment(country)
        translated_countries.append(translated_country)

    # Leggi tutti i dati dal file originale e memorizzali in una lista di dizionari
    with open(r"C:\Users\aleda\OneDrive\Allegati\OneDrive\Desktop\TIROCINIO\ds1.csv", newline='', encoding="utf-8") as d1:
        reader = csv.DictReader(d1)
        data_list = [row for row in reader]

    # Verifica se il numero di traduzioni corrisponde al numero di righe nel file
    if len(translated_countries) == len(data_list):
        # Aggiungi le traduzioni alla lista dei dati
        for i, row in enumerate(data_list):
            row["country"] = translated_countries[i]

        # Scrivi tutti i dati, inclusi i dati tradotti, nel nuovo file CSV
        with open(r"C:\Users\aleda\OneDrive\Allegati\OneDrive\Desktop\TIROCINIO\ds1_translated.csv", "w", newline='', encoding="utf-8") as d1_translated:
            fieldnames = reader.fieldnames
            writer = csv.DictWriter(d1_translated, fieldnames=fieldnames)
            writer.writeheader()
            for row in data_list:
```

Figura 3.8: Seconda parte del codice Python per la traduzione della colonna "country"

```python
        # Scrivi tutti i dati, inclusi i dati tradotti, nel nuovo file CSV
        with open(r"C:\Users\aleda\OneDrive\Allegati\OneDrive\Desktop\TIROCINIO\ds1_translated.csv", "w", newline='', encoding="utf-8") as d1_translated:
            fieldnames = reader.fieldnames
            writer = csv.DictWriter(d1_translated, fieldnames=fieldnames)
            writer.writeheader()
            for row in data_list:
                writer.writerow(row)
    else:
        print("Errore: il numero di traduzioni non corrisponde al numero di righe nel file.")
```

Figura 3.9: Terza parte del codice Python per la traduzione della colonna "country"

country	buyers	topbuyers	topbuyerratio	femalebuyers
France	1251	53	4,2	851
United Kingdom	792	38	4,8	560
The United States of America	912	31	3,4	700
Germany	578	29	5	409
Italy	400	21	5,3	283
Spain	255	21	8,2	189
Netherlands	144	15	10,4	118
Sweden	151	11	7,3	113
Finland	64	10	15,6	53
Denmark	157	9	5,7	127
Australia	126	9	7,1	92
Belgium	90	7	7,8	73
Austria	49	6	12,2	40
Bulgaria	9	4	44,4	7
Canada	65	3	4,6	45
Romania	28	3	10,7	22
Hong Kong	28	3	10,7	22
Luxembourg	13	3	23,1	10
Portugal	18	2	11,1	12
China	13	2	15,4	9
Hungary	7	2	28,6	7
Slovenia	2	2	100	2

Figura 3.10: Risultato dato dall'importazione della tabella tradotta

3.3.3 Caricamento

La fase di caricamento (Load) è l'ultimo passo cruciale nel processo di ETL e rappresenta il momento in cui i dati trasformati vengono inseriti in un repository centrale o in un data warehouse. Questa fase è fondamentale per rendere i dati pronti per l'analisi e l'elaborazione, mettendoli a disposizione degli utenti finali o dei sistemi di Business Intelligence. Per procedere ad una fase di caricamento corretta è fondamentale operare un'esaustiva trasformazione, in modo da non dover avere a che fare con dati mancanti o incompatibili con il formato previsto.

Inoltre, è comune pianificare il caricamento in modo che i dati siano disponibili per l'analisi in tempo reale o in batch, a seconda delle esigenze del progetto. Il monitoraggio delle prestazioni garantisce che il caricamento sia efficiente e che i dati siano sempre aggiornati.

3.3.3.1 Come è stato implementato

Nel contesto del progetto descritto, l'ultima fase del processo di elaborazione dei dati comporta il caricamento di questi ultimi dall'editor Power Query a Power BI. Durante questa fase, i dati estratti e trasformati vengono utilizzati per creare le dashboard, alimentando i grafici e le visualizzazioni che le compongono. Per completare questa operazione, è necessario fare clic sull'opzione *Chiudi e Applica*, come illustrato nella Figura 3.11, in modo da consolidare le modifiche apportate ai dati e di sincronizzarli con l'interfaccia di Power BI.

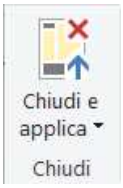

Figura 3.11: Opzione "Chiudi e Applica"

CAPITOLO 4

Anlisi effettuate e risultati derivati

Questo quarto capitolo vuole esporre il lavoro svolto sui dati provenienti dal dataset "Fashion E-Commerce" e riportare i risultati derivati dall'analisi.

4.1 Data Visualization

La visualizzazione dei dati è il processo di traduzione di dati complessi in elementi visivi, come grafici e diagrammi, per rendere le informazioni maggiormente accessibili. Questa pratica è essenziale nell'analisi dei dati, in quanto garantisce una comunicazione delle informazioni in modo efficace ad un pubblico ampio.

Poiché la cultura umana è principalmente orientata alla percezione visiva, la visualizzazione dei dati è un mezzo efficace per attirare e mantenere l'attenzione. L'uso di rappresentazioni visive facilita il rilevamento di tendenze, valori anomali e relazioni complesse nei dati. Tuttavia, nonostante possa sembrare un processo semplice, è cruciale riconoscere l'importanza delle scelte stilistiche nella visualizzazione dei dati, poiché decisioni errate portano ad un'errata interpretazione dei dati.

In particolare, le principali problematiche che possono emergere includono la presentazione di informazioni parziali o inaccurate, la possibile mancanza di evidenza delle correlazioni reali, che potrebbero essere scambiate per casualità, e la possibilità di trascurare i messaggi chiave nel processo di rappresentazione dei dati.

La visualizzazione dei dati può essere utilizzata per vari scopi, come l'illustrazione dei risultati degli esami scolastici da parte degli insegnanti, l'analisi dei progressi nell'Intelligenza Artificiale da parte degli esperti informatici o la condivisione di informazioni aziendali da parte dei dirigenti. Harvard Business Review classifica la visualizzazione dei dati in quattro categorie principali:

- *Generazione di idee*: la visualizzazione dei dati è comunemente usata per stimolare la generazione di idee da parte del team. Essa trova un utilizzo frequente durante le sessioni di brainstorming o Design Thinking all'inizio di un progetto.
- *Illustrazione di idee*: la visualizzazione dei dati è una pratica diffusa in svariati ambiti, con l'obiettivo di agevolare la comunicazione tra individui in relazione a compiti specifici. Un esempio emblematico è il diagramma di Gantt, spesso utilizzato dai project manager per rappresentare in modo chiaro il flusso di lavoro.

- *Scoperta visiva e studi quotidiani*: la scoperta visiva e la visualizzazione dei dati quotidiana sono più strettamente allineate con i team di dati. Mentre la scoperta visuale aiuta gli analisti dei dati, i data scientist e altri professionisti dei dati a identificare modelli e tendenze all'interno di un set di dati, la visualizzazione dati quotidiana supporta la narrazione successiva alla scoperta di una nuova intuizione.

4.1.1 Data Visualization in Power BI

È stato precedentemente esposto quanto Power BI sia, in effetti, uno strumento fondamentale nell'ambito della Data Visualization in quanto sufficientemente intuitivo e, al contempo, potente per poter esprimere in maniera efficace i risultati derivati da un'analisi dei dati.

Innanzitutto, sappiamo che gli strumenti che ci vengono messi a disposizione sono:

- fogli di lavoro;
- grafici;
- filtri;
- formule DAX.

Il tutto verrà esposto in maniera più dettagliata nelle prossime sottosezioni.

4.1.1.1 Fogli di lavoro

Il foglio di lavoro è rappresentato in Power BI dalla prima schermata che incontriamo una volta avviata l'applicazione; esso è una parte fondamentale dell'ambiente di sviluppo ed analisi. In particolare, il termine "foglio di lavoro" in Power BI fa riferimento alle "pagine" o "schede" all'interno di un report. Ogni pagina che costituisce un report può essere formata da numerose visualizzazioni, grafici, tabelle ed altri elementi visivi per rappresentare i dati in modo efficace. Gli utenti possono navigare tra le pagine di un report usando le schede nella parte inferiore della finestra di Power BI. Questa navigazione è utile per esplorare diverse visualizzazioni e informazioni all'interno di un report. Altra caratteristica fondamentale dei fogli di lavoro è la possibilità, tramite questi, di visualizzare gli stessi dati in diversi formati. Nella parte all'estrema sinistra del foglio troviamo tre icone che ci permettono di osservare le pagine tramite tre diverse tipologie di visualizzazione, ovvero:

- *Report*: questo tipo di rappresentazione ci permette di visualizzare il report concentrandoci sulla parte grafica; quindi i dati vengono rappresentati tramite grafici.
- *Tabella*: in questo caso possiamo visualizzare le tabelle che contengono i dati rappresentati nel report stesso.
- *Modello*: quest'ultima tipologia di visualizzazione ci permette di notare come le varie tabelle del report siano tra loro interconnesse, evidenziandone i rapporti.

4.1.1.2 Grafici

I grafici sono gli strumenti visivi che vengono utilizzati in tutti i tool di Data Visualization per rappresentare efficacemente i dati oggetto di studio. Ne esistono di diverse tipologie e la combinazione di queste ci permette di ottenere, come risultato finale, una dashboard.

In Figura 4.1 vediamo riportati tutte le varie tipologie di grafico disponibili su Power BI, tra queste ne citiamo alcune utilizzate nella campagna di data analytics oggetto della presente tesi:

1. *Grafico a barre*: mostra dati categorici o numerici in forma di barre orizzontali o verticali. È adatto per il confronto tra categorie o l'illustrazione di tendenze nel tempo.
2. *Grafico a dispersione*: detto anche scatter plot, permette di visualizzare la relazione tra due variabili numeriche, consentendo di individuare correlazioni o modelli nei dati.
3. *Grafico a torta*: rappresenta la suddivisione percentuale di una variabile in parti. È efficace per mostrare la composizione di un insieme di dati.
4. *Grafico a linee*: collega i punti dati con linee; e è ideale per mostrare tendenze temporali o dati in continua evoluzione.
5. *Grafico ad aree*: questo tipo di grafico visualizza i dati come un'area sottesa da una curva; è ideale per mostrare le tendenze temporali e le variazioni tra categorie.
6. *Grafico a nastro*: questo grafico mostra il flusso dei dati o delle transizioni tra diverse categorie o periodi. È utile per evidenziare le connessioni o le interazioni tra i dati.
7. *Mappa*: questo grafico visualizza i dati su una mappa geografica, evidenziando posizioni o distribuzioni geografiche dei dati.
8. *Istogramma*: questo tipo di grafico visualizza la distribuzione di una variabile numerica in forma di colonne verticali. È utile per esaminare la frequenza di valori all'interno di intervalli specifici.
9. *Schede*: le schede consentono di organizzare i contenuti in diversi pannelli o schede all'interno di una dashboard, permettendo all'utente di passare tra diverse visualizzazioni o aspetti dei dati.

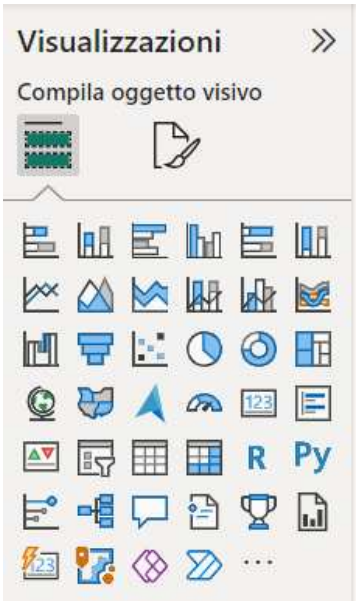

Figura 4.1: Tipologie di grafici disponibili in Power BI

4.1.1.3 Filtri

I filtri in Power BI rappresentano una delle funzionalità fondamentali per controllare quali dati vengono analizzati in un report o in una visualizzazione specifica. Essi consentono

di concentrarsi su segmenti specifici dei dati o di escludere dati indesiderati per analizzare i restanti in modo più mirato. Di seguito riportiamo una panoramica parziale dei filtri in Power BI.

1. *Filtro pagina*: possiamo applicare filtri specifici a una singola pagina del nostro report. Ad esempio, se abbiamo una pagina che mostra dati per un mese specifico, possiamo applicare un filtro di pagina per selezionare quel mese e visualizzare solo i dati corrispondenti.
2. *Filtro visuale*: un filtro visuale è specifico per una visualizzazione particolare all'interno di una pagina. Ad esempio, possiamo aggiungere un filtro visuale a un grafico per consentire agli utenti di selezionare una categoria specifica da esaminare.
3. *Filtro globale*: un filtro globale può essere applicato a più pagine di un report e influisce su tutte le visualizzazioni su quelle pagine. È utile quando desideriamo applicare lo stesso filtro a diverse parti del nostro report.
4. *Filtro a scorrimento*: questo tipo di filtro consente di selezionare un intervallo di valori in una colonna numerica, ad esempio per filtrare date in un intervallo di tempo specifico.
5. *Filtro con misure*: possiamo creare filtri basati su misure personalizzate che abbiamo definito in DAX. Questo è utile per consentire agli utenti di eseguire analisi più avanzate sui dati.
6. *Filtro incrociato*: quando si applica un filtro a una visualizzazione, tale filtro può influenzare anche altre visualizzazioni correlate. Ad esempio, la selezione di un punto su un grafico a dispersione può filtrare i dati in un grafico a barre correlato.

4.1.1.4 Formule DAX

Le formule DAX rappresentano un insieme di funzioni ed operatori utilizzati in Power BI per eseguire calcoli avanzati e ottenere una comprensione più approfondita dei dati. DAX sta per "Data Analysis Expressions" ed è il linguaggio di formula utilizzato in Power BI. Si tratta di un linguaggio funzionale, ovvero, un linguaggio in cui tutto il codice è racchiuso all'interno di funzioni.

DAX offre una vasta libreria di oltre 200 funzioni, operatori e costrutti, che forniscono una flessibilità significativa, in modo da poter essere in grado di creare misure di calcolo per soddisfare una vasta gamma di esigenze di analisi dei dati. Le funzioni in DAX possono essere annidate l'una all'interno dell'altra, supportare istruzioni condizionali e fare riferimento a valori. L'esecuzione delle formule DAX inizia dall'interno e procede verso l'esterno.

Le misure DAX rappresentano calcoli dinamici che si aggiornano costantemente in base alle azioni dell'utente nei report, consentendo l'esplorazione dinamica dei dati. Queste misure possono essere create nelle visualizzazioni dei report, nuove o esistenti, o nelle viste dei dati. Sono identificate da un'icona a forma di calcolatrice nell'elenco dei campi, e si presentano come colonne calcolate e tabelle calcolate, ampliando ulteriormente le opzioni per l'analisi e la visualizzazione dei dati. Gli elementi chiave necessari per creare una misura calcolata includono il nome della misura e almeno una funzione o espressione associata.

4.2 Metodologie di analisi

Tra le numerose opzioni per la visualizzazione dei dati, abbiamo utilizzato Power BI. Questo strumento consente di ottenere insight significativi grazie alle sue capacità di elaborazione dei dati, offrendo un equilibrio ottimale tra facilità d'uso e prestazioni.

Il nostro studio si è concentrato sul dataset *Fashion E-Commerce*, da cui abbiamo sviluppato quattro report distinti. Ogni report è stato progettato per esaminare dati da prospettive diverse, come analisi geografiche, suddivisione per genere ed analisi delle abitudini degli utenti.

4.2.1 Report 1

Il primo report, illustrato nella Figura 4.2, è incentrato sull'analisi condotta in base all'origine geografica degli utenti. Tale report è composto da quattro grafici distinti:

1. *Istogramma a colonne in pila*: visualizza la distribuzione degli utenti iscritti al sito per paese, distinguendo in base al loro titolo di cortesia ("Mr," "Miss," "Mrs"). Su tale grafico è stato applicato un filtro che permette di escludere i paesi in cui la voce `buyers` ha valore zero, ovvero i paesi privi di acquirenti.
2. *Mappa*: mostra la quantità di vendite riportate per ciascun paese. Anche in questo caso è stato applicato un filtro che ci permette di evidenziare solo i paesi in cui il totale delle vendite risulta essere superiore a cinquanta.
3. *Grafico a linee*: confronta il numero di acquirenti donne (rappresentati dalla linea rosa) e uomini (rappresentati dalla linea blu) suddivisi per paese di provenienza.
4. *Schede*: forniscono il conteggio totale degli acquirenti (scheda superiore) e il totale degli utenti iscritti al sito (scheda inferiore).

Ai grafici elencati si aggiunge una descrizione, che riportiamo:

> Abbiamo notato una forte correlazione positiva tra il numero totale di "Acquirenti Donne" e "Acquirenti Uomini".
>
> In Francia, abbiamo osservato che il 21,56% degli acquirenti sono donne.
>
> Negli Stati Uniti, abbiamo notato la maggiore differenza tra il numero di acquirenti uomini e donne. Le donne avevano un vantaggio di 488 utenti rispetto agli uomini.
>
> In tutti i 200 paesi considerati nei dati, abbiamo notato che il numero di utenti con il titolo "Miss" variava da 1 a 140, mentre il titolo "Mr" aveva valori compresi tra 1 e 7050. Il titolo "Mrs" aveva valori compresi tra 1 e 17996.

Tale descrizione è stata realizzata automaticamente dalla funzionalità "Narrazione Intelligente" di Power BI, la quale aggiunge un riepilogo dei dati proposti nel report in cui è inserito.

4.2.2 Report 2

Nel secondo report conduciamo un'analisi basata sulla suddivisione degli utenti iscritti al sito in base al genere (quindi uomini o donne). Ogni grafico presenta una breve descrizione, posizionata strategicamente al di sotto o di fianco a esso, per evitare confusione sull'effettiva analisi proposta.

Questo secondo report è illustrato nella Figura 4.3 e si compone di otto grafici distinti:

1. *Due diagrammi a torta*, che rappresentano uno studio sulla percentuale di utenti, distinti per genere, che possiedono o non possiedono applicazioni installate sui propri telefoni. Gli utenti di sesso femminile sono rappresentati dal grafico rosa, mentre quelli di sesso maschile sono rappresentati dal grafico blu.

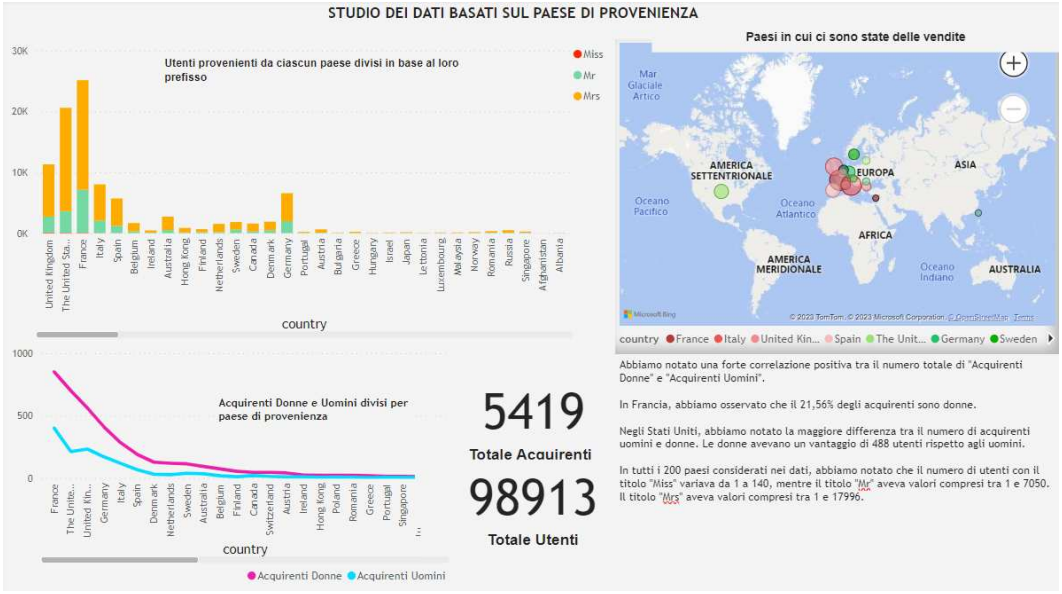

Figura 4.2: Primo report creato

2. *Un grafico a nastri*, che mostra la quantità di "mi piace" assegnati ai diversi prodotti social da parte di utenti uomini e donne, suddivisi per paese di provenienza.

3. *Tre schede*, che riportano, dall'alto verso il basso, il numero totale di utenti inattivi e la media dei giorni di inattività per uomini e donne.

4. *Due grafici ad aree in pila*, i quali rappresentano, da destra verso sinistra, il numero di follower e il numero di "follows" da parte di utenti uomini e donne, suddivisi per paese di provenienza.

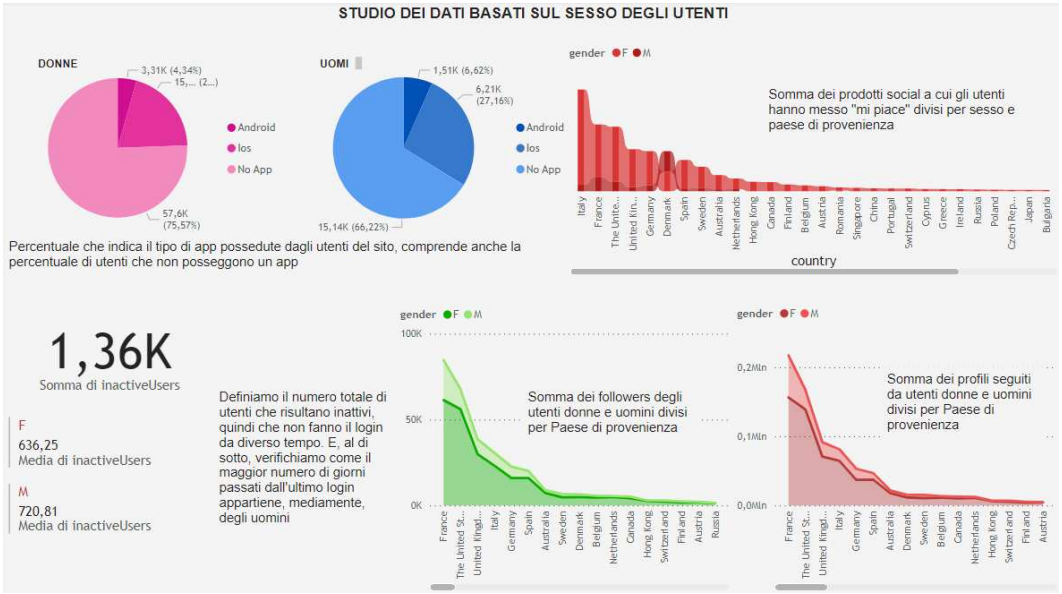

Figura 4.3: Secondo report creato

4.2.3 Report 3

Nel terzo report presentiamo quattro grafici di dispersione (scatter) progettati per visualizzare le abitudini comportamentali degli utenti iscritti al sito, suddivisi in base al loro genere e ai paesi di provenienza:

1. *Vendite*: rappresenta il trend delle diverse vendite, distinguendo la tipologia di utente che le effettua e il paese di provenienza.
2. *Acquisti*: mostra il trend degli acquisti, differenziando la tipologia di utente che li effettua e il paese di provenienza.
3. *Prodotti desiderati*: visualizza il trend dei prodotti desiderati, in particolare quanti di essi vengono inseriti nelle liste dei desideri da parte dei diversi utenti, suddivisi per paese di provenienza e genere.
4. *Prodotti listati*: illustra il trend dei prodotti, in particolare quanti di essi vengono elencati da parte degli utenti, differenziati per paese di provenienza e genere.

Il report descritto può essere visualizzato in Figura 4.4.

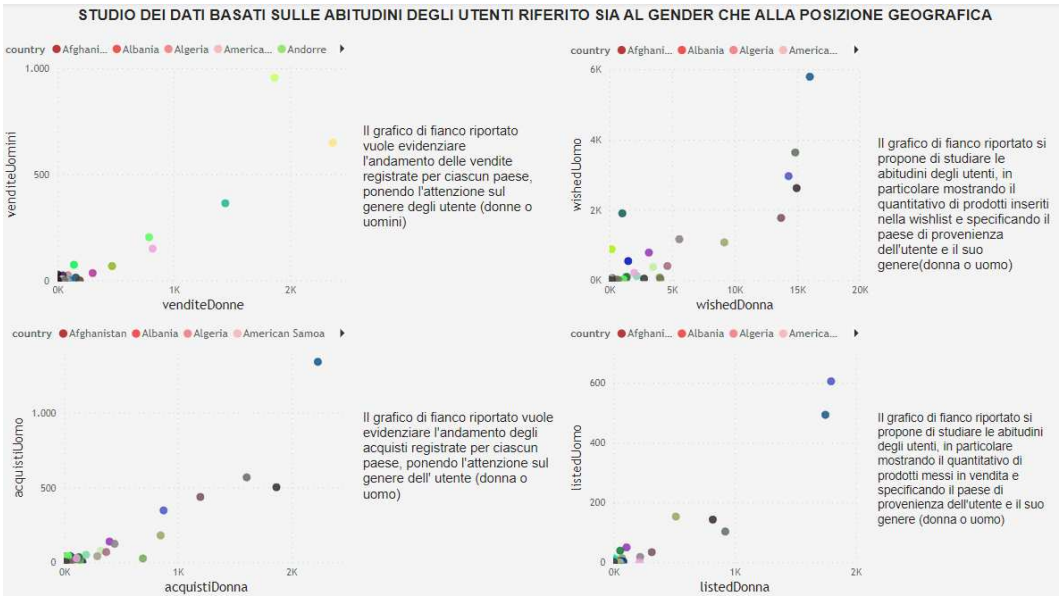

Figura 4.4: Terzo report creato

4.2.4 Report 4

Con questa ultimo report miriamo a confrontare valori correlati tra loro. In particolare, troviamo tre istogrammi a colonne in pila che, partendo dall'alto verso il basso, ci definiscono i seguenti confronti:

- *Acquisti e vendite*: dove è stato applicato un filtro per visualizzare i paesi in cui il valore `Acquisti-Vendite` risulti essere maggiore di 20 e minore di -5.
- *Prodotti desiderati e prodotti effettivamente acquistati*: dove è stato applicato un filtro per poter vedere i paesi in cui il valore `wished-acquisti` presenta un valore maggiore di 300 o minore di -1.

- *Prodotti messi in vendita dagli utenti e prodotti che sono stati venduti*: anche su quest'ultimo grafico è stato applicato un filtro per mostrare i paesi in cui il valore di `listed-Sold` è maggiore di 5 e minore di -10.

Questi confronti consentono di esaminare le relazioni tra diversi aspetti del comportamento degli utenti e di ottenere insight significativi sulle loro abitudini e preferenze. Il tutto può essere osservato in Figura 4.5.

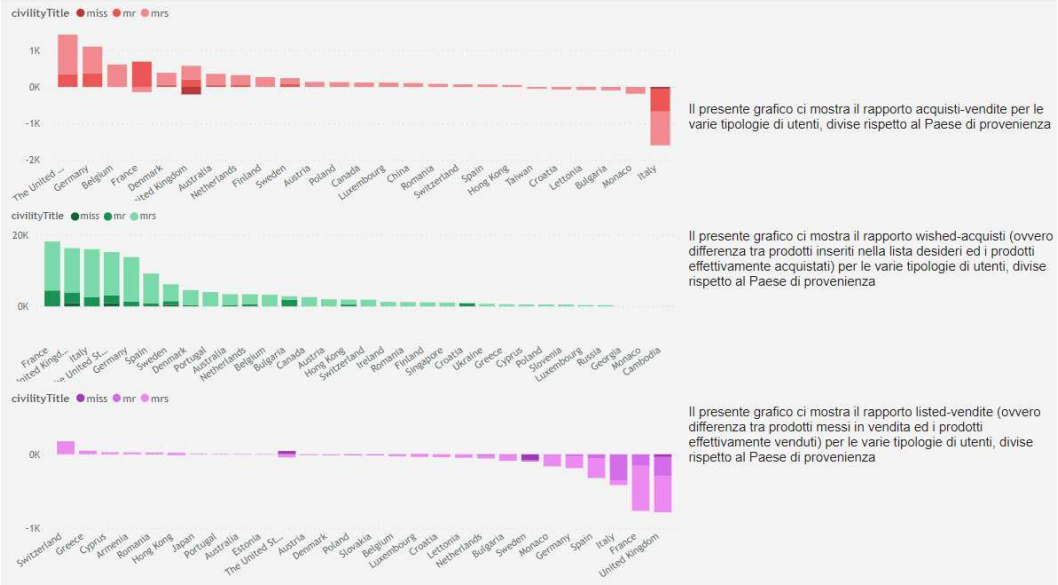

Figura 4.5: Quarto report creato

4.3 Risultati derivati

È stata fornita una dettagliata descrizione di ciascun report e del loro contenuto. In ognuno di essi, sono presenti grafici che ci consentono di identificare aspetti distintivi riguardanti il dataset.

Nel primo report è evidente che il maggior numero di utenti proviene dalla Francia, il che risulta coerente con le origini francesi del sito. Questo fatto è stato evidenziato dal processo di trasformazione dei dati, ma diventa ancora più evidente a causa dell'istogramma, il quale rende chiara questa informazione anche per gli utenti esterni. L'istogramma è visibile in Figura 4.6.

In particolare, quando si esamina la distribuzione di utenti donne e uomini, si può notare che gli utenti con il titolo "Miss" costituiscono una piccola parte del totale degli utenti, mentre gli utenti con titolo "Mrs" ne rappresentano la maggioranza. Questo suggerisce che il dataset proviene da un sito web il cui pubblico principale è formato da donne sposate; pertanto, si rivolge ad un pubblico più maturo. Tale informazione assume particolare valore dal momento che non sono offerte informazioni sull'età degli utenti in nessuna tabella di partenza del dataset.

Dalla mappa, emergono ulteriori informazioni, in particolare che la maggior parte delle vendite avvengono in Europa; si riconoscono Italia e Francia come paesi leader per il numero di vendite.

Infine, osserviamo che, nella maggior parte dei paesi, il numero di acquirenti donne supera quello degli acquirenti uomini.

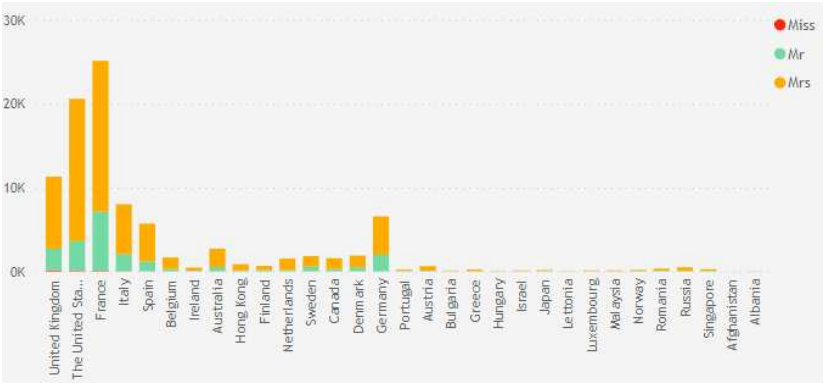

Figura 4.6: Istogramma presente nel primo report

Nel secondo report analizziamo gli utenti iscritti al sito distinguendoli per genere (uomini e donne). Approfondiamo le differenze tra queste due categorie in vari aspetti, tra cui il numero di follower e "follow", la presenza di applicazioni installate sui loro telefoni, il periodo di inattività, nonché la quantità di "like" assegnati ai vari prodotti visualizzati sul sito.

È interessante notare che il trend dei follower e dei "follow" sia generalmente proporzionato, tranne in corrispondenza dell'Italia, dove si osserva un picco significativo di "follow"; il dettaglio dei grafici ad aree che lo mostra è riportato in Figura 4.7.

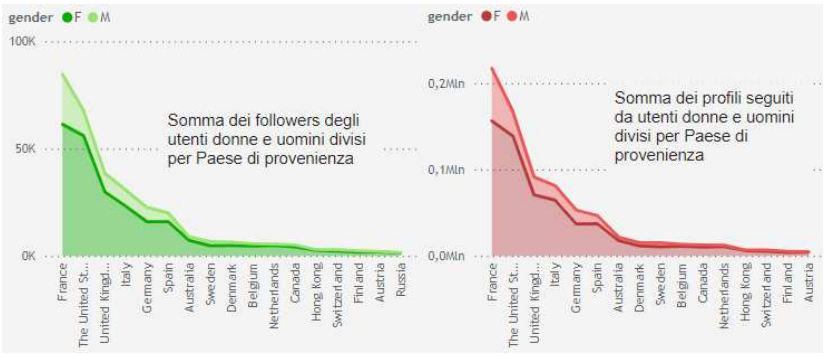

Figura 4.7: Grafici ad aree presenti nel secondo report

Successivamente, ci concentriamo sul grafico a nastro, che rappresenta il numero di "mi piace" assegnati dai vari utenti a prodotti specifici, suddivisi per sesso e paese di provenienza. In questi confronti tra utenti uomini e donne, è risultato evidente come, da un punto di vista dell'interazione sui social network, gli utenti femminili risultino più attivi rispetto agli utenti maschili. Tuttavia, nel grafico a nastro considerato, osserviamo come, in corrispondenza della Danimarca, si ha un'inversione di questo trend in quanto gli uomini sembrano essere più attivi in termini di "Mi piace"; tale dettaglio è raffigurato in Figura 4.8.

Nelle schede riportate, esaminiamo il numero di utenti inattivi e la media di giorni di inattività per donne e uomini. Notiamo che, in media, gli uomini presentano circa un centinaio di giorni di inattività in più rispetto alle donne. Questo dettaglio è conforme con la maggiore attività di utenti donne.

Nel terzo report, esaminiamo nel dettaglio i grafici a dispersione (o scatter) e cosa questi dovrebbero mostrarci. In particolare, essi evidenziano che i paesi più attivi risultano essere l'Italia e la Francia.

Nel quarto report, confrontiamo dati correlati attraverso tre istogrammi disposti uno sotto l'altro. Il primo istogramma mira a esaminare la quantità di prodotti acquistati rispetto a

Figura 4.8: Grafico a nastro presente nel secondo report

quelli venduti, aggregando questi movimenti per utenti e suddividendoli in base al titolo ("Mr", "Miss" e "Mrs") e ai paesi di provenienza.

Il secondo istogramma opera in modo simile, ma si concentra sul rapporto tra i prodotti acquistati e quelli inseriti nelle liste dei desideri. Curiosamente, notiamo che non ci sono totali negativi; ciò suggerisce che, in generale, quando un prodotto viene acquistato, proviene dalla lista dei desideri del cliente.

Il terzo grafico risulta essere il più intrigante, poiché fornisce informazioni chiave per la comprensione del dataset e dei dati in esso contenuti. Questo grafico esamina il rapporto tra i prodotti venduti e quelli messi in vendita dagli utenti. Qui notiamo che, in generale, i totali calcolati sono negativi, indicando che i prodotti venduti non compaiono nelle liste degli utenti. Questa scoperta ha generato due possibili spiegazioni: la prima potrebbe suggerire la presenza di dati corrotti o non corretti nel dataset originale, mentre la seconda ipotizza che nel dataset venga indicato solo il tipo di capo messo in vendita (ad esempio, "pantaloni Nike") senza specificare la quantità disponibile di quell'oggetto. Si può visualizzare tale grafico nella Figura 4.9.

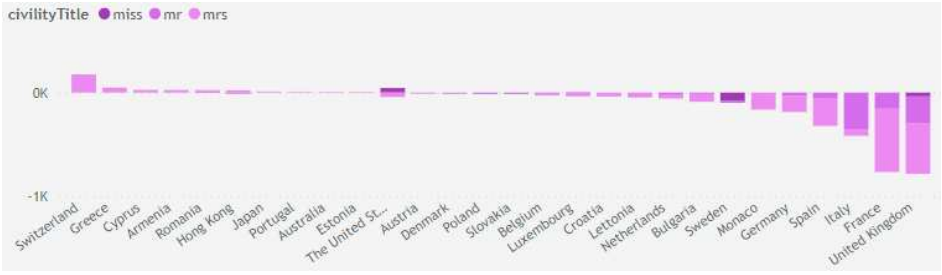

Figura 4.9: Terzo istogramma nel quarto report

CAPITOLO 5

Discussione in merito al lavoro svolto

In questo capitolo riportiamo le considerazioni maturate riguardo il progetto svolto. In particolare vogliamo evidenziarne i punti di forza e di debolezza; a ciò si aggiunge una riflessione nei confronti della Data Analytics e ai suoi utilizzi nel mondo attuale.

5.1 Punti di merito del lavoro svolto

In questa sezione vogliamo evidenziare i punti di merito del lavoro svolto. In particolare, questi verranno esposti in maniera dettagliata nelle seguenti sottosezioni.

5.1.1 Fruibilità

La fruibilità è un elemento cruciale in ogni progetto di Data Analytics, essenziale per ottimizzare l'efficacia e la comprensione da parte degli utenti. Lo studio proposto si impegna a garantire questa caratteristica attraverso un'interfaccia utente intuitiva. Questa interfaccia semplifica l'accesso ai dati, agli strumenti e fornisce visualizzazioni chiare ed informative, facilitando, così, una rapida comprensione dei risultati ottenuti. Ciò rende il nostro progetto non solo un'opportunità per i proprietari o gestori di siti di e-commerce, ma anche un punto di partenza stimolante per avviare una propria campagna di Data Analytics, coinvolgendo attivamente gli specialisti di marketing.

5.1.2 Scalabilità

Un altro punto di forza del progetto presentato è la sua scalabilità, ovvero la capacità del sistema di crescere e gestire un aumento del volume dei dati, degli utenti o delle richieste senza perdita di prestazione.

Una progettazione scalabile è cruciale per affrontare l'evoluzione delle esigenze aziendali. Oltre alla capacità del progetto di essere ripreso ed adattato, quest'ultimo può anche evolversi verso l'ottenimento di una maggiore complessità. Questo può essere ottenuto integrando più sorgenti, realizzando nuovi tipi di dashboard interattive, aggregando metodi predittivi di machine learning o, anche, realizzando un bot Telegram aziendale che, a partire da semplici parole chiave, sia in grado di fornire dashboard da visualizzare.

5.1.3 Power BI

Un ulteriore nota positiva propria del progetto è l'utilizzo di Power BI come tool di Data Analytics. Tale strumento è stato ampiamente descritto nei precedenti capitoli ma ribadiamo la sua versatilità, adattabilità, e potenzialità. Lo poniamo tra i punti di forza in quanto è utilizzato in molte realtà aziendali ed è utilizzabile anche da utenti con poca esperienza nel settore.

5.1.4 Innovazione

L'innovazione è annoverata tra i punti di merito, in quanto l'argomento stesso della tesi è da considerarsi come innovativo. La Data Analytics è un settore che sta crescendo esponenzialmente in questi anni, acquistando sempre maggiore interesse, grazie ai numerosi vantaggi derivati dal suo utilizzo.

5.2 Punti di vulnerabilità

In questa sezione proponiamo i punti di vulnerabilità che riguardano il lavoro svolto.

5.2.1 Qualità dei dati

Dall'analisi condotta nel più recente report. È emerso un interessante punto di riflessione riguardante alcuni dati che mostrano un comportamento apparentemente non congruente da parte degli utenti. Nello specifico, si è notato che il numero di prodotti effettivamente venduti, nella maggior parte dei casi, superava il numero di prodotti messi in vendita. Tale fenomeno potrebbe derivare da diverse cause. Inizialmente, i dati raccolti dal sito potrebbero essere soggetti a corruzione o inesattezze, oppure potrebbe esserci una limitazione nell'informazione fornita dal sito stesso, che potrebbe riportare solo la tipologia di prodotto messa in vendita senza considerare le relative quantità. Un'altra spiegazione potrebbe risiedere nel fatto che i dati analizzati si riferiscono a un intervallo temporale specifico, trascurando eventi precedenti e, quindi, sottostimando il numero reale di prodotti disponibili. Di conseguenza, è plausibile che il numero effettivo di prodotti messi in vendita sia superiore rispetto a quanto indicato nei dati oggetto di analisi.

5.2.2 Quantità dei dati

Una limitazione evidente del dataset selezionato è la sua mancanza di diversità nei dati. Attualmente, le uniche fonti di informazioni disponibili sono rappresentate dalle tabelle `ds1` e `ds2`, focalizzate sugli utenti e sui paesi di provenienza. Tuttavia, risulterebbe alquanto interessante esplorare ulteriormente lo studio includendo dati dettagliati sugli effettivi prodotti venduti e acquistati. Tale analisi potrebbe fornire una panoramica più approfondita, consentendo di identificare variazioni nei trend a livello internazionale, comprendendo aspetti come marchi predominanti, preferenze di colore e capi d'abbigliamento di moda. Introdurre questi elementi nei dati forniti permetterebbe di arricchire notevolmente la comprensione dei comportamenti degli utenti e delle dinamiche di mercato, contribuendo, così, a una visione più completa e approfondita del contesto in esame.

5.3 Applicazioni in ambito pratico

L'ambito della Data Analytics, è attualmente in rapida evoluzione, e riveste un ruolo centrale nella società contemporanea, influenzando ogni settore e guidando decisioni a livello

individuale e aziendale. Con l'ampia quantità di dati generati quotidianamente da fonti come i social media, le transazioni online, i sensori e i dispositivi connessi, l'analisi dei dati è diventata cruciale per estrarre significato da questa vasta mole di dati.

Le organizzazioni stanno impiegando un numero sempre maggiore di professionisti specializzati in Data Science per tradurre dati grezzi in insight strategici. Questa pratica non si limita alla retrospettiva, ma include anche la previsione e la prescrizione, consentendo alle aziende di anticipare tendenze, ottimizzare processi e prendere decisioni informate. L'evoluzione della Data Analytics è strettamente legata alla crescita di tecnologie come l'Intelligenza Artificiale e il Machine Learning, che permettono di analizzare dati complessi in modi sempre più sofisticati.

Inoltre, la sicurezza dei dati e la privacy emergono come sfide cruciali, richiedendo un approccio equilibrato per sfruttare i vantaggi dell'analisi dei dati senza compromettere la protezione delle informazioni sensibili. La Data Analytics sta trasformando radicalmente la nostra capacità di comprendere il mondo, guidare l'innovazione e prendere decisioni più informate.

In particolare, l'ampia diffusione di strumenti di Data Analytics ha trovato grande fortuna nel mondo post-COVID, come indicato da un articolo riportato da Innovation Post, che afferma:

> *La crescita è trainata soprattutto dalla componente software, che registra un incremento del 17%, con punte di oltre il 30% per le piattaforme di Data governance, Data science e AI, e dai servizi di consulenza e personalizzazione tecnologica, mentre la spesa in risorse infrastrutturali aumenta meno della media del mercato.*

Negli ultimi tre anni, oltre la metà delle grandi imprese ha intrapreso almeno una sperimentazione nell'ambito dell'Advanced Analytics. Tuttavia, durante la fase di implementazione, sono emerse alcune sfide significative, tra cui la scarsa qualità e integrazione dei dati, la parziale carenza di competenze interne, la difficoltà nel valutare i benefici di singoli progetti e la complessità nel coinvolgere gli utenti aziendali.

In un contesto come quello italiano, queste difficoltà rappresentano un ostacolo significativo, specialmente considerando che il tessuto aziendale è prevalentemente composto da piccole e medie imprese, molte delle quali non possono essere effettivamente classificate come "data driven". Nel contesto qui descritto, per "data driven" intendiamo un'organizzazione che orienta le proprie decisioni e strategie basandosi su analisi e interpretazioni dei dati, anziché su intuizioni o esperienze passate. In tale contesto, un'azienda data driven raccoglie, analizza e attivamente utilizza dati provenienti da varie fonti per guidare le decisioni, ottimizzare le operazioni e migliorare le prestazioni complessive. La necessità di superare queste sfide diventa, quindi, cruciale per promuovere l'adozione più diffusa di pratiche aziendali orientate ai dati, specialmente tra le imprese di dimensioni più contenute.

Conclusioni ed uno sguardo al futuro

Nel corso della presente tesi di laurea, sono state illustrate le metodologie, gli strumenti e le fasi coinvolte nella realizzazione di una campagna di Data Analytics, con particolare focus sul dataset "Fashion E-Commerce" ottenuto dal sito Kaggle. Il sito in questione è stato denominato "vintage" nel titolo della tesi, poiché la seniority media degli utenti registrati si attesta attorno agli 8,5 anni.

La campagna di Data Analytics è stata condotta attraverso una serie di passaggi ben definiti. Innanzitutto, è stata condotta un'analisi esplorativa dei dati, al fine di acquisire una comprensione approfondita degli stessi. Successivamente, è stata avviata la fase di ETL in cui i dati, una volta ottenuti da Kaggle, vengono inseriti in PowerQuery. Una volta inseriti nell'editor, essi vengono standardizzati, puliti e modellati per far sì che soddisfino opportuni requisiti, come l'integrità, la coerenza, l'accessibilità e la conformità alle regole di business o agli standard dell'industria. Una volta completato tale processo, essi vengono caricati all'interno di PowerBI, dove vengono visualizzati e filtrati per poi essere apposti in specifici report. In tutto, dallo studio dei suddetti dati, sono stati generati quattro report distinti.

Il primo report analizza i dati suddividendoli in base alla provenienza geografica degli utenti, mentre il secondo si focalizza sull'analisi dei dati in relazione al genere degli utenti. Il terzo ed il quarto approfondiscono le abitudini degli utenti nel sito, concentrandosi su acquisti, vendite, liste desideri e liste di prodotti in vendita, categorizzati in base al paese di provenienza, al sesso e al titolo ("Mr","Mrs","Miss").

Tutte le attività e le fasi descritte sono applicabili anche a progetti più ampi destinati a supportare aziende di varie dimensioni. Lo studio delle abitudini degli acquirenti consente alle aziende di avere una comprensione più profonda della popolarità dei propri servizi o prodotti, facilitando la presa di decisioni informate e consapevoli per il futuro aziendale. Questo approccio permette anche l'implementazione di analisi predittive e prescrittive a diversi livelli.

Un'analisi del genere potrebbe stimolare il panorama aziendale italiano, costituito principalmente da imprese di piccole e medie dimensioni, ad adottare una mentalità più "data-driven". Una delle sfide affrontate da molte di queste aziende è la necessità di adattarsi rapidamente a un mercato in costante cambiamento. L'analisi dei dati fornisce un vantaggio competitivo significativo in questo contesto, apportando i seguenti vantaggi:

- *Miglioramento delle strategie di marketing e vendita*: attraverso l'analisi dei dati dei clienti, è possibile identificare trend di acquisto, preferenze e comportamenti, consentendo una personalizzazione più efficace delle campagne pubblicitarie. Ciò non solo aumenta l'efficacia delle iniziative di marketing, ma contribuisce anche a una maggiore soddisfazione del cliente.

- *Previsione della domanda di ottimizzazione e delle scorte*: le aziende possono sfruttare la data analytics per prevedere la domanda dei propri prodotti o servizi. Questa capacità di previsione consente di ottimizzare le scorte, riducendo i costi legati a eccessi di magazzino o perdite dovute a scorte insufficienti. Una gestione accurata delle scorte è particolarmente cruciale per le PMI che devono operare in modo efficiente con risorse limitate.

- *Personalizzazione dell'esperienza del cliente*: l'analisi dei dati consente alle aziende di offrire un'esperienza del cliente altamente personalizzata. Dalla personalizzazione dei suggerimenti di prodotti alla creazione di offerte speciali basate sul comportamento passato del cliente, le aziende possono costruire relazioni più profonde e durature con la propria clientela.

In conclusione, la data analytics si configura non solo come uno strumento per comprendere il passato e il presente, ma soprattutto come un motore di innovazione cruciale per il futuro delle aziende italiane. L'adozione di queste pratiche non solo offre una panoramica dettagliata delle dinamiche aziendali attuali, ma costituisce anche un veicolo di trasformazione culturale. In particolare, uno studio come quello riportato si distingue per la sua focalizzazione chiara sull'orientamento al cliente.

L'analisi dei dati mira a cogliere le esigenze e le preferenze dei clienti, fungendo da catalizzatore per un cambiamento culturale significativo. Questo approccio strategico non solo stimola le PMI a diventare più agili e competitive, ma soprattutto le orienta in modo deciso verso la creazione di valore per il cliente. Attraverso l'impiego di pratiche analitiche avanzate, il presente studio rappresenta un esempio tangibile di come le imprese possano mettere al centro delle proprie strategie il soddisfacimento delle esigenze dei clienti, contribuendo, così, a plasmare un futuro aziendale più sostenibile e orientato al successo a lungo termine.

Bibliografia

ASPIN, A. (2014), *High impact data visualization with power View, Power Map, and Power BI*.

ASPIN, A. (2016), *Pro Power BI Desktop*.

CHEN, H., CHIANGAN, R. H. L. e STOREY, V. C. (2012), *Business intelligence and analytics: From big data to big impact*.

ERL, T., KHATTAK, W. e BUHLER, P. (2016), *Big data fundamentals : concepts, drivers techniques*.

FERRARI, A. e RUSSO, M. (2015), *The Definitive Guide to DAX: Business Intelligence with Microsoft Excel, SQL Server Analysis Services, and Power BI*.

FERRARI, A. e RUSSO, M. (2016), *Introducing Microsoft Power BI*, Microsoft Press.

FERRARI, A. e RUSSO, M. (2017), *Analyzing Data with Power BI and Power Pivot for Excel*.

FINLAY, S. (2014), *Predictive analytics, data mining and big data: Myths, misconceptions and methods*, Springer.

LACHEV, T. e PRINCE, E. (2017), *Applied Microsoft Power BI (2nd Edition): Bring Your Data to Life!*

MACHIRAJU, S. e GAURAV, S. (2018), *Power BI Data Analysis and Visualization*.

SERVICES, E. (2015), , *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*.

Siti web consultati

- Microsoft Power BI – `www.powerbi.microsoft.com/it-it/`
- Kaggle – `www.kaggle.com/`
- Wikipedia – `www.wikipedia.org`
- Osservatori – `www.osservatori.net/it/home`

- Tableau – `www.tableau.com/`
- Talend – `www.talend.com`
- IBM – `www.ibm.com/it-it`

Ringraziamenti

Vorrei dedicare queste ultime righe per ringraziare le numerose persone che mi hanno affiancato e supportato per la durata di questo percorso.

Prima di tutto, la mia famiglia: mamma la tua gentilezza e la tua dolcezza mi hanno sempre guidato e motivato; grazie perchè sei la mia roccia. Papà, spero un giorno di avere la tua stessa risolutezza e spensieratezza; so che avrai sempre una parola dolce per tirarmi su il morale. A mio fratello Lorenzo, ispirazione di una vita e sostegno morale, mi hai insegnato tutto quello che so e se sono come sono lo devo a te. A Laura che più che una cugina è la sorella che non ho mai avuto; averti così vicina è una gioia infinita.

Ai miei nonni Rita e Antonio, per la gioia e l'amore che emanano; senza di voi il Natale non sarebbe così divertente. A zio Daniele, che ci ha sempre mostrato come essere creativi nella vita.

A nonno Aldo, nonna Ilde, zia Livia e zia Sidonia, perchè siete sempre con noi.

Ai miei amici di sempre: Elettra, Vittoria, Christian, Giada, Matteo e Federico.

Al gruppo Rustell che mi è stato affianco dal primo giorno di università.

A quelli del pullman che mi hanno accompagnato in questi tre anni passati e che saranno al mio fianco per i successivi: Gians, Walterone, Laura, Luca ,Valeria, Edoardo ed, infine, Alessio, colui senza il quale non sarei qui.

Alle coinquiline storiche Giulia, Giada e Gaia, mi avete fatto passare due anni incredibili; grazie per le risate, i film e le mille bottiglie di vino rigorosamente finite alle 21.

A piazza stamira, la mia nuova casa popolata da strane creature, grazie a Laura, Stefano, Dora, Francesco, Angelo e Andrea.

Ai migliori amici che si possano avere, che si sono dovuti sorbire mille scleri, pianti e sfoghi; grazie dal profondo del cuore a Matilda, Irene e Mirco. Vi voglio un bene dell'anima.

Infine un ringraziamento speciale al Professor Domenico Ursino per avermi guidato in questo percorso con una disponibilità ed una gentilezza fuori dal comune, e al Professor Francesco Cauteruccio, che mi ha affiancato per tutto il periodo estivo con un'incredibile professionalità, disponiblità e simpatia.