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 tramite PowerBI di una campagna di data analytics sulle gare di Formula 1

Design and implementation using PowerBI of a data analytics campaign on Formula 1 races

Relatore

Prof. Domenico Ursino

Candidato

Luca D'Ambrosio

Correlatore

Michele Marchetti

ANNO ACCADEMICO 2023-2024

Meglio zoppicare sulla strada giusta
che correre su quella sbagliata

Sant'Agostino

Sommario

Negli ultimi anni, grazie al progresso tecnologico, i dati hanno assunto un ruolo fondamentale nel settore della Formula 1. Essi sono diventati la base per lo sviluppo di nuove strategie, per l'ottimizzazione delle monoposto e il miglioramento di ogni aspetto delle gare. Nella presente tesi vengono analizzati i dati relativi alle gare di Formula 1. In una prima fase, viene effettuata una pulizia dei dati; in particolare, viene effettuato il download del dataset dalla piattaforma *Kaggle*. In seguito, vengono modificati vari campi di determinate tabelle del dataset, in modo da condurre analisi più accurate. A seguire, le tabelle vengono caricate sulla piattaforma *Power BI*. Successivamente, vengono sviluppate varie dashboard, utilizzando sia gli oggetti visivi predefiniti che quelli importati. Quindi vengono analizzate le dashboard relative: al confronto tra i costruttori e i piloti e al confronto tra due determinati piloti; all'andamento del passo di gara; allo sviluppo delle posizioni, dei sorpassi in gara, dei pit-stop e dell'affidabilità. L'obiettivo finale è quello di estrapolare le informazioni e i pattern sull'evoluzione della Formula 1 nel corso del tempo.

Keyword: Data Analytics, Power BI, Big Data, Extract Trasform and Load, Structured Query Language, Data Warehouse, Data Analysis, Python.

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Negli ultimi anni, la Data Analytics è diventata uno strumento fondamentale in numerosi settori, specialmente nella Formula 1.

In questo sport è importante ricordare che ogni millesimo di secondo può essere cruciale; di conseguenza, è necessario analizzare gli aspetti essenziali di questo settore per poter migliorare le prestazioni in pista.

Le scuderie di Formula 1 possono utilizzare la Data Analytics per ottimizzare strategie di gara, per sviluppare vetture più performanti, per comprendere al meglio le prestazioni dei piloti e delle monoposto in diverse condizioni e in diversi circuiti ed, infine, per elaborare previsioni mirate a supportare le decisioni strategiche in tempi estremamente brevi.

La presente tesi ha l'obiettivo di analizzare in modo approfondito gli aspetti caratteristici e fondamentali delle gare di Formula 1, attraverso l'utilizzo di Power BI. Quest'ultimo metterà in risalto il valore della Business Intelligence, capace di arricchire le conoscenze degli utenti, consentendo loro di accedere ad un livello di comprensione più ampio dei gran premi e delle prestazioni dei piloti.

Power BI, infatti, ci permette di rappresentare enormi e complessi dataset in dashboard interattive e intuitive da comprendere e rappresentare; di conseguenza, è possibile estrapolare informazioni fondate su dati concreti.

Nello specifico, il lavoro svolto si concentrerà sull'elaborazione e sull'analisi dettagliata di dashboard relative all'andamento del passo di gara, allo sviluppo di posizioni e sorpassi nel corso di un gran premio, di un'intera stagione o di più stagioni.

Inoltre, verrà esaminato il confronto sia tra i vari piloti, che tra i piloti e le scuderie relativamente alle vittorie conquistate, per cercare di comprendere se vi sono dei collegamenti sui risultati ottenuti. Successivamente ci si focalizzerà sulla progressione dei pit-stop ponendo l'attenzione su determinati circuiti e determinati piloti. Infine, sarà approfondita l'evoluzione dell'affidabilità nel tempo, concentrandosi sulle cause maggiori che provocano guasti alle monoposto, mettendo in luce i principali fattori che ne influenzano le prestazioni.

Questo lavoro si colloca in uno scenario più ampio relativo all'utilizzo della Data Analytics nello sport, dimostrando che gli strumenti di visualizzazione possono essere utilizzati sia a livello professionale, in questo caso per le scuderie, che da appassionati e da analisti, per comprendere tutti gli aspetti di questo settore che risulta essere uno dei più complessi.

In particolare, esso risulta utile anche per chi desidera entrare a far parte di questo mondo, specialmente in ambito lavorativo, e ha difficoltà nell'approccio iniziale per approfondire le proprie conoscenze in merito. In una prima fase, verranno eseguite le attività di ETL. In particolare, come primo passo, si effettuerà il download del Formula 1 dataset dalla piattaforma *Kaggle*; in seguito, grazie all'utilizzo degli strumenti messi a disposizione da

Power BI e grazie all'utilizzo di script sviluppati in linguaggio *Python*, saranno modificati vari campi di determinate tabelle, in modo da condurre analisi più accurate. Successivamente, come ultima attività, le tabelle saranno caricate su *Power BI*.

Una volta terminate le attività di *ETL*, ci si dedicherà allo sviluppo delle dashboard attraverso l'utilizzo degli oggetti visivi di base e altri oggetti visivi importati relativi a *Power BI*, che rappresentaranno a livello visivo i dati all'interno del dataset.

Successivamente verranno analizzate le dashboard, con lo scopo di acquisire informazioni essenziali relative all'evoluzione nel corso degli anni del passo di gara, dei pit-stop, delle posizioni e dei sorpassi in gara, dell'affidabilità, delle vittorie, confrontando i piloti e i costruttori. Infine verrà esaminato il confronto delle prestazioni tra due determinati piloti.

Questa tesi è strutturata come di seguito specificato:

- Nel Capitolo 1 sarà introdotto il mondo della Data Analytics partendo dalla sua definizione ed evoluzione, proseguendo con le tecniche e gli strumenti ad esso correlati, per poi concludere con un confronto con la Data Analysis.
- Nel Capitolo 2 verrà presentata la piattaforma Power BI, analizzando il suo scopo, le sue componenti e i settori di applicazione.
- Nel Capitolo 3 verranno descritti inizialmente le varie tipologie di dati, proseguendo con la descrizione dei dati presenti nel dataset utilizzato.
- Nel Capitolo 4 verranno analizzate le fasi di ETL sul dataset della Formula 1 scelto.
- Nel Capitolo 5 saranno descritte le attività di analisi di base. Quindi, le correlazioni tra le vittorie dei piloti e dei costruttori, e il confronto delle prestazioni tra due determinati piloti, in un determinato anno.
- Nel Capitolo 6 verranno illustrate le attività di analisi avanzate relative allo sviluppo del passo di gara, dello sviluppo delle posizioni e dei sorpassi in gara, dei pit-stop e dell'affidabilità.

CAPITOLO 1

Introduzione alla Data Analytics

In questo primo capitolo si affronteranno diversi temi relativi alla Data Analytics partendo dalla sua definizione ed evoluzione storica fino al suo utilizzo nel contesto moderno e nei vari settori, concentrandoci anche sugli strumenti e i software più comunemente adottati. Si proseguirà con un approfondimento sul termine "Big Data", per poi concludere analizzando le differenze tra Data Analytics e Data Analysis

1.1 La Data Analytics

La Data Analytics è una disciplina che si occupa di estrarre, elaborare ed analizzare dati grezzi, con lo scopo di ricavarne informazioni utili per supportare decisioni strategiche. Questo processo, infatti, consente di identificare e affermare supposizioni, modelli, correlazioni e divergenze tra i diversi dati.

L'importanza e l'influenza della Data Analytics accresce e si espande progressivamente in diversi settori. Ciò deriva dal fatto che le aziende hanno urgenza di ottimizzare i loro processi, in quanto si è arrivati ad un punto in cui il tempo che si impiega a prendere una decisione è determinante. Di conseguenza, per creare nuove strategie e nuovi modelli nei vari ambiti del mercato, è cruciale capire effettivamente cosa esprimono i dati raccolti.

1.1.1 Storia ed evoluzione della Data Analytics

Percorrere l'evoluzione della Data Analytics è estremamente importante. Ci permette, infatti, di entrare nel merito dello sviluppo tecnologico e ci aiuta a comprendere e contestualizzare le soluzioni adottate contemporaneamente alla nascita di nuove esigenze. Si possono suddividere le diverse fasi della Data Analytics nelle seguenti fasce di tempo:

- 1950-1970: a partire dalla metà del XX secolo, è avvenuta una svolta significativa per quanto concerne l'analisi dei dati. In passato, la raccolta dei dati e la loro relativa analisi erano totalmente manuali, di tipo statistico e riguardava il tracciamento delle operazioni economiche e la gestione dell'inventario. Con l'utilizzo dei computer, questo processo è stato rivoluzionato sia a livello di tempi che di quantità, nonostante i limiti imposti dalla potenza di calcolo. L'analisi veniva effettuata in modo semplice, ma era pur sempre efficace ed efficiente.
- 1970-1980: durante questo periodo, ci furono altre innovazioni, come lo sviluppo dei database relazionali e del linguaggio SQL (*Structured Query Language*), che hanno fornito

maggiore sviluppo e supporto circa l'archiviazione e l'analisi dei dati. Contestualmente, l'incremento della potenza di calcolo dei computer ha contribuito a rendere più rapidi ed efficienti i processi di analisi. In parallelo anche la Business Intelligence progrediva, poichè le aziende erano capaci di sfruttare i dati in loro possesso in modo più accurato; ciò permetteva loro di attuare tecniche di mercato più accurate. A discapito di ciò, è sorto il problema degli elevati costi per lo spazio di archiviazione dei dati.

- 1980-1990: nel corso di questi anni si registrò una crescita considerevole nella quantità di dati raccolti. Ciò portò alla nascita dei Data Warahouse, progettati per convertire i dati generati dai sistemi operativi in risorse utili al processo decisionale. Molte aziende iniziarono a seguire l'esempio di quelle che già utilizzavano la Business Intelligence, attirate dai risultati concreti e dai vantaggi ottenuti da chi ne faceva uso. Questo portò non solo a un aumento delle analisi possibili sui dati, ma anche a un'intensificazione della concorrenza tra le imprese, spingendole a sfruttare sempre di più le tecnologie per ottenere un vantaggio competitivo.

- 1990-2010: in questa fase c'è stato un grande progresso per quanto riguarda la Data Analytics. Anzitutto, furono sviluppati i primi software di Business Intelligence, i quali fornivano informazioni basate sui dati storici e permettevano alle aziende di realizzare report e analisi descrittive, anche se inizialmente le capacità predittive erano ridotte. Inoltre, l'avvento dei personal computer, permise a molte altre persone di accedere e analizzare i dati e anche alle aziende di raccoglierne in quantità maggiori e svolgere analisi approfondite. Durante lo stesso periodo, ha guadagnato importanza un nuovo processo di evoluzione dei data warehouse, ovvero il data mining, che aveva l'obiettivo di scoprire pattern nascosti all'interno di grandi set di dati. Nei primi anni del 2000, il data mining si è espanso maggiormente anche grazie alla diffusione di Internet, al cui interno ci fu un notevole incremento della quantità di dati disponibili. Questo fenomeno segnò l'inizio dell'era dei "Big Data", un'esplosione nei volumi, nella velocità e nella varietà di dati che ha creato nuove sfide e opportunità. In aggiunta, nuovi strumenti e tecnologie, come Hadoop, hanno reso possibile la gestione di questi enormi set di dati, consentendo alle aziende di ottenere analisi più complesse e predittive.

- 2010-oggi: in questi ultimi anni, la diffusione del cloud computing ha reso accessibili a un pubblico più ampio, potenti risorse di calcolo e archiviazione, semplificando ulteriormente l'accesso all'analisi dei dati e l'elaborazione di grandi quantità di dati. Inoltre, le aziende, per effetto del costante sviluppo di strumenti e piattaforme, ovvero con l'emergere del machine learning e dell'Intelligenza Artificiale, hanno la possibilità di prendere decisioni in tempo reale analizzando enormi quantità di dati, il che consente ad esse di reagire prontamente alle variazioni del loro ambito. Da questo aspetto è possibile dedurre che la Data Analytics svolge un ruolo centrale nelle grandi aziende, orientandole verso il raggiungimento del successo e definendone le iniziative strategiche.

1.1.2 Utilizzo della Data Analytics

Come affermato precedentemente, la Data Analytics, nel corso del tempo, si è espansa notevolmente in vari settori. Tra i più importanti ritroviamo i seguenti:

- *Economico-finanziario*: nel corso di questi ultimi anni i settori economico e finanziario stanno attraversando una significativa trasformazione per via della costante evoluzione tecnologica. Difatti, banche ed aziende ad oggi utilizzano la Data Analytics analizzando dati storici e tendenze di mercato e monitorando transazioni in tempo reale, sia per

stimare i rischi futuri con maggiore accuratezza tra cui possibili insolvenze, sia per attuare strategie contro le perdite. In aggiunta, anche nel mondo degli investimenti, l'analisi dei dati è fondamentale per aumentare i propri ricavi, individuando opportunità e ottimizzando le performance del portafoglio. Banche e aziende attualmente sono anche capaci di proporre, tramite comportamenti di spesa e fasce di reddito, servizi personalizzati aumentando così la fidelizzazione dei propri clienti.

- *Sanitario*: anche in questo settore la Data Analytics, con il passare degli anni, si è espansa maggiormente dato che genera notevoli vantaggi. Innanzitutto, analizzando i dati relativi alle cartelle cliniche dei pazienti, la Data Analytics si è rivelata di grande aiuto a tutti gli operatori sanitari supportandoli nel prendere decisoni cliniche, quali prescrivere farmaci, attuare trattamenti e terapie mediche specifiche per ogni paziente, ottimizzando i risultati e riducendo gli effetti collaterali. In aggiunta la Data Analytics è determinante anche nella prevenzione poiché un'analisi dei dati continua permette di rilevare precocemente patologie, disfunzioni e fattori di rischio per la salute stessa.

  Infine, la Data Analytics è divenuta fondamentale anche per la ricerca e lo sviluppo, con l'obiettivo di favorire l'innovazione in campo medico, nella scoperta di nuovi farmaci e terapie.

- *Sportivo*: la Data Analytics, in modo progressivo, ha influenzato notevolmente anche il mondo dello sport, a partire dalla Formula 1 sino ad arrivare al calcio. Il suo ruolo in questo settore è simile a quello dei settori precedenti. In tutte le attività sportive lo scopo principale è quello di migliorare le prestazioni; pertanto, soprattutto negli sport di squadra, l'analisi dei dati può aiutare sia il coach nella ricerca di nuovi schemi e strategie e nell'analisi dell'avversario, sia i preparatori atletici, svolgendo un ruolo di prevenzione contro infortuni e analisi biomeccaniche. Anche nel motorsport i dati vengono analizzati per scoprire punti di forza e di debolezza di una vettura, per analizzare circuiti e per sviluppare nuove componenti o migliorare le strategie. Inoltre, svolgendo anche il marketing un ruolo importantissimo in questo contesto, l'analisi dei dati supporta decisoni economiche strategiche per quanto concerne lo sviluppo di prodotti e servizi e per migliorare la visibilità dei vari sponsor e brand. Infine, piattaforme digitali, archivi, sensori e dispositivi indossabili sono i mezzi attraverso cui sono facilmente individuabili questi dati.

- *Trasporti e logistica*: in questo settore la Data Analytics si è rivelata fondamentale poiché, grazie ad essa, vi sono stati ottimi margini di miglioramento. Il settore dei trasporti si trova ad affrontare varie difficoltà, tra cui carenze di autisti e limitazioni del carico di ogni mezzo. Di fronte a questi vincoli, l'analisi dei dati ha ottimizzato i relativi trasporti, riducendo i costi e migliorandone l'efficienza. Inoltre, queste analisi hanno migliorato anche la programmazione dei tragitti: ogni consegna ha una sua precisa necessità a livello di tempi e, proprio per questo, è necessario avere infomazioni riguardo traffico e mobilità in tempo reale. Necessariamente, però, si presentano delle limitazioni in base alla tipologia di oggetto che deve essere trasportato; alcuni prodotti potrebbero essere maggiormente soggetti a rotture o guasti; ecco perché attuare tecniche di manutenzione preventiva incrociando i dati riduce notevolmente il rischio di tali inconvienienti.

1.1.3 Strumenti e software

Si dispone di una serie di tool per la Data Analytics, i quali ci consentono di estrarre informazioni utili per condurre ed elaborare nuove strategie in ogni ambito, usando tecniche di visualizzazione dati, analisi statistica, machine learning. Nel corso degli anni sono stati

sviluppati vari tipi di questi software; tra i più importanti vi sono Microsoft Power BI, che approfondiremo nel Capitolo 2, Tableau e Qlik.

Tableau mantiene ancora la sua importanza essendo tra le prime scelte per le aziende; un punto di forza di questo tool sta nel fatto che riesce a gestire enormi quantità di dati ed è capace di condurre analisi in tempo reale, nonostante la semplicità del suo utilizzo. Qlik è un fondamentale strumento di supporto per le aziende. Le sue applicazioni permettono di condurre analisi per interpretare il comportamento dei clienti, attuare nuove strategie di guadagno ed aiutano persino a gestire rischi e opportunità.

Oltre ai tool elencati si può condurre l'analisi dei dati anche con il supporto di linguaggi di programmazione, tra i quali:

- *Python*; è attualmente il linguaggio più utilizzato per l'analisi dei dati; esso fornisce, infatti, diverse librerie, come pandas, NumPy e Matplotlib, che facilitano la manipolazione e la visualizzazione di dati complessi. L'utilizzo maggiore di questo linguaggio include processi di ETL (*Extract, Transform e Load*), visualizzazione di dati, web scraping, machine learning.
- SQL (*Structured Query Language*); il linguaggio principale per interfacciarsi con i database relazionali. Lo scopo principale di questo linguaggio e delle sue varianti è quello di gestire, interrogare e modificare i dati presenti nei database in maniera efficiente.
- *R*; è utilizzato principalmente per effettuare analisi statistiche; esso, grazie alla sua capacità di supportare vari tipi di dati e alla sua efficacia nella loro visualizzazione è altamente raccomandato tra i professionisti del settore.

1.2 Tecniche di Data Analytics

Nell'ambito della Data Analytics è importante essere a conoscenza delle diverse tecniche, ognuna delle quali si riferisce ad una esigenza specifica. Nel nostro caso tratteremo quattro tipi di analisi, ovvero analisi descrittiva, analisi diagnostica, analisi predittiva e analisi prescrittiva. Come mostrato in Figura 1.1, si può anticipare che la complessità di elaborazione dei dati e il valore dei risultati forniti aumentano mano a mano che si passa dall'analisi descrittiva a quella prescrittiva.

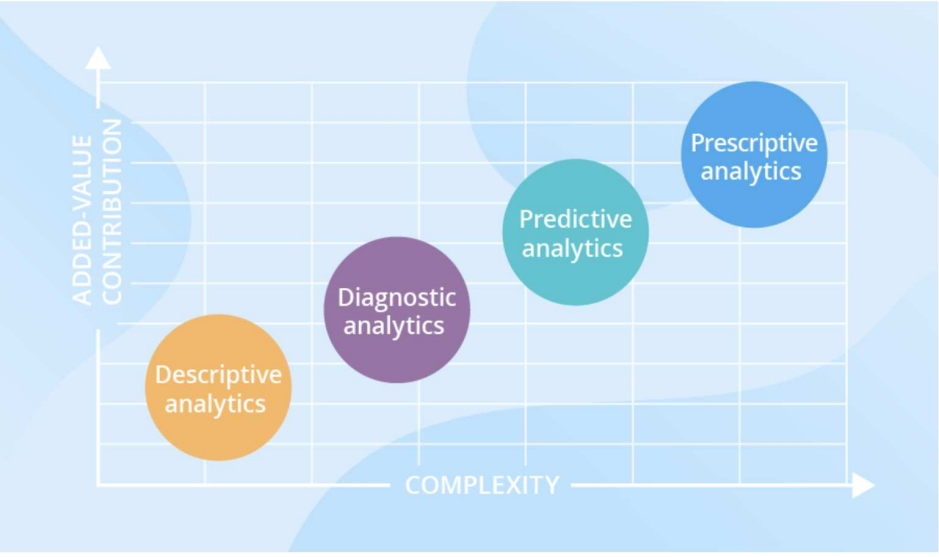

Figura 1.1: Tecniche di Data Analytics. Fonte ScienceSoft

1.2.1 Analisi descrittiva

L'analisi descrittiva si basa sulla domanda "cosa è successo?", dato che ha l'obiettivo di analizzare dati grezzi provenienti da diverse fonti per offrire informazioni sul passato, fornendoci una visione statica di quest'ultimo. Questo tipo di analisi è raccomandato per le aziende soltanto se rappresenta una fase introduttiva di elaborazione dei dati, nella quale viene elaborato un quadro generale dei dati passati con lo scopo di effettuare analisi avanzate, cioè quella predittiva, che analizzeremo in seguito.

Per chiarire meglio questo concetto, può essere utile fare riferimento a un esempio pratico. Un'azienda che produce una serie di prodotti, alla fine del mese, analizzando i dati sui ricavi, avrà una visione più chiara su quale prodotto concentrarsi e investire maggiori risorse. Sono, infatti, previste 5 fasi nell'analisi descrittiva, ovvero:

1 *Quantificare gli obiettivi e le metriche di business*: questa prima fase traduce gli obiettivi aziendali in metriche misurabili, come ricavi, vendite e riduzione dei costi.

2 *Identificare i dati rilevanti*: dopo aver definito le metriche, è necessario raccogliere i dati appropriati, che spesso provengono da fonti diverse, interne ed esterne all'azienda.

3 *Estrarre e organizzare i dati*: i dati devono essere controllati, puliti e formattati correttamente per evitare errori; questo processo, infatti, diventa più complesso con l'aumento della loro quantità e varietà.

4 *Analizzare i dati*: nella fase di analisi, i dati vengono combinati, riassunti e confrontati utilizzando tecniche di analisi descrittiva per descrivere la metrica di riferimento.

5 *Presentare i dati*: le metriche vengono presentate in report o dashboard per essere comprese dai decision maker.

1.2.2 Analisi diagnostica

Questo tipo di analisi confronta dati passati con altri dati in modo da capire cosa si sia verificato, e quindi scoprirne le cause. Proprio per questo risponde a domande del tipo, come mai è accaduto? Pertanto, l'analisi diagnostica viene ritenuta come il passo successivo all'analisi descrittiva. Quest'analisi si concentra fortemente a fornire una visione di un problema specifico; a volte, però, questo può essere controproducente nel caso in cui i dati raccolti si focalizzano soltanto su un singolo problema che potrebbe portare a un grande dispendio di tempo. L'analisi diagnostica viene applicata in diversi campi, come, ad esempio, nella Business Intelligence, nella logistica, nella finanza, nel marketing, e sopratutto nella sanità. In quest'ultimo ambito, essa gioca un ruolo cruciale poiché, analizzando i dati dei pazienti, dal loro stile di vita fino ad arrivare ai protocolli di trattamento, medici e ricercatori possono risalire alle cause di determinate malattie, e quindi identificare modelli dei fattori di rischio che contribuiscono alla diagnosi di determinate malattie, migliorando trattamenti e cure.

Purtroppo, anche l'analisi diagnostica deve essere supportata da altri tipi di analisi in quanto ha carenze nell'analisi in tempo reale e nell'analisi predittiva; inoltre, la sua efficacia è decisamente limitata riguardo alla qualità e disponibilità dei dati.

1.2.3 Analisi predittiva

Il fine di questo tipo di analisi è quello di svolgere previsioni future basandosi sull'utilizzo dei dati per supportare decisioni future. Normalmente quest'analisi è interconnessa sia a quella descrittiva che a quella diagnostica per identificare cluster ed eccezioni, in modo

da condurre analisi più accurate sul futuro. Quest'ultimo può essere sia immediato, come ad esempio anticipare un aumento improvviso della domanda di un prodotto entro una settimana, ma anche più lontano, come calcolare l'andamento delle vendite per i prossimi cinque anni.

L'analisi predittiva si avvale di diversi strumenti; in particolare spicca l'analisi regressiva, una tecnica di analisi statistica che permette di identificare la relazione tra due o più variabili. Queste relazioni, che vengono descritte tramite equazioni matematiche, hanno lo scopo di riuscire a condurre analisi sul futuro anche se una delle variabili dovesse cambiare. Poi vi sono gli alberi decisionali, utilizzati prevalentemente per prevedere e comprendere le decisioni di una persona. Un albero decisionale è formato da rami, che rappresentano le opzioni possibili, e da foglie, che rappresentano l'esito della decisione.

Infine, abbiamo le reti neurali in situazioni di relazioni estremamente complesse, in cui non esistono equazioni matematiche note per analizzare i dati e il set di dati è non lineare; questo strumento è di supporto ai due precedenti per confermare i risultati.

1.2.4 Analisi prescrittiva

L'analisi prescrittiva è una tecnica che utilizza i dati per supportare aziende e organizzazioni a prendere decisioni per ottimizzare i risultati e mitigare i rischi, in modo da far accadare ciò che hanno pianificato.

I metodi per eseguire questo tipo di analisi sono diversi; vi sono, infatti, modelli probabilistici, Machine Learning/Data Mining, programmazione matematica, evolutionary computation, simulazioni, Logic-Based Model.

Infine, l'analisi prescrittiva viene utilizzata in vari campi. In primo luogo si ritrova nelle decisioni di investimento; infatti, nel caso in cui bisogna prendere decisioni fondamentali, il supporto di quest'analisi aiuta a comprendere se l'investimento può generare utili o meno. In secondo luogo, è presente anche nel campo delle vendite tramite il lead scoring, una tecnica che assegna un punteggio alle azioni compiute dai potenziali clienti in modo da stabilire la probabilità che diventino clienti effettivi.

Indubbiamente è utilizzata nei social media in tutte le sezioni "For You" dedicate alle nostre preferenze e create in base ai dati raccolti, tramite l'interazione tra noi e le nostre app.

1.3 Big Data

Nel corso di questi ultimi anni possiamo affermare di essere circondati dai Big Data, poiché li ritroviamo in ogni ambito della nostra vita. Questo fenomeno, nato nei primi anni del 2000, è aumentato a dismisura con il passare degli anni. Nonostante questo, ancora non vi è una definizione univoca del termine "Big Data"; potremmo riassumerla dicendo che sono enormi quantità di dati eterogenei, i quali non posso essere analizzati semplicemente con dei fogli di calcolo. Ecco perchè possiamo individuarli tramite alcune caratteristiche fondamentali, che prendono il nome di 5V. Esse sono le seguenti:

- *Volume*: per volume si intende enormi quantità di dati, dell'ordine terabyte. Queste quantità di dati solitamente vengono generati da macchine, come, ad esempio, sensori o DCS (Distributed Control System), oppure anche da social media, pagine Web o app mobili.

- *Velocità*: questo termine indica la rapidità con cui i dati vengono generati, ricevuti ed elaborati. A seguito dell'espansione dell'Internet of Things e dei sensori, i dati vengono generati con velocità estremamente elevate, a volte in tempo reale o poco meno. Di conseguenza, essi devono anche essere maneggiati con altrettanta efficienza; spesso,

infatti, vengono inviati direttamente sulla memoria, invece che salvarli su disco, in modo da analizzarli subito.

- *Varietà*: la varietà indica che i dati, essendo eterogenei e quindi di diverso formato, provengono da fonti diverse e sono anche strutturati diversamente. Generalmente i dati erano principalmente strutturati e facilmente adattabili a database relazionali. Ciò nonostante, con l'avvento dei Big Data, i dati comprendono anche formati non strutturati e semi-strutturati, come testo, audio, video e dati provenienti da sensori. Questi dati necessitano di un'elaborazione iniziale per estrarre informazioni utili, poiché non sono facilmente riconducibili a schemi fissi.

- *Veridicità*: i dati, specialmente quando sono in quantità elevata, possono presentare errori o essere rumorosi e disordinati; ecco perché bisogna valutare quanto affidamento si può fare su di essi. Ma se si lavorasse su set di dati più piccoli potrebbero mancare diverse informazioni; per questo motivo la veridicità dei dati è cruciale, poiché maggiore è la loro veridicità e più sono affidabili. Basta pensare, infatti, alle informazioni presenti sul Web da parte di concorrenti che tentano in modo sleale di fare marketing.

- *Valore*: ovviamente ogni dato ha un suo valore intrinseco, fondamentale per condurre analisi e prendere decisioni. Le aziende in possesso dei Big Data analizzandoli sono capaci di otterene un vantaggio non indifferente rispetto alle aziende che non ne possiedono. Il valore può essere interno, per ottimizzare i processi operativi, oppure esterno, per avere indicazioni utili sui clienti, in modo da incrementare il loro coinvolgimento.

Per sfruttare e comprendere al meglio i Big Data nel proprio settore, c'è bisogno di tre attività chiave:

1 *Integrazione*: i big data aggregano informazioni provenienti da numerose fonti e applicazioni differenti. I metodi tradizionali di integrazione dei dati, come estrazione, trasformazione e caricamento (ETL) non sono sufficienti per gestire set di dati di dimensioni così grandi; per questo motivo c'è bisogno di tecniche e strategie avanzate. Difatti, è di fondamentale importanza che i dati vengano raccolti, elaborati, formattati e disponibili in un formato che consenta ad analisti aziendali di poterli elaborare senza troppe difficoltà.

2 *Gestione*: per archiviare i Big Data è necessario un enorme spazio di archiviazione. Negli ultimi anni, infatti, hanno acquisito grande importanza i data lake, poiché essi soddisfano le attuali esigenze di elaborazione e permettono di espandere le risorse quando necessario.

3 *Analisi*: l'ultimo step, ma altrettando fondamentale, è l'analisi dei Big Data. Di grande importanza è il modo in cui vengono visualizzati questi dati, ragione per cui è necessario creare dashboard, per avere una visione più chiara, e per poterla spiegare ai membri dell'azienda in modo che essi possano comprendere in quali ambiti agire maggiormente.

I Big Data sono ormai un elemento centrale in una vasta gamma di ambiti, dalla sanità all'automotive, dal commercio al marketing. Nel marketing, i Big Data hanno sviluppato diverse opportunità per le aziende; ad esempio, con il location based-marketing, sfruttando la geolocalizzazione, è possibile inviare al cliente determinate pubblicità. Inoltre, un altro grande utilizzo riguarda la riduzione dei costi di produzione nelle aziende manifatturiere; infatti, attraverso la raccolta dei dati tramite sensori è possibile monitorare la condizione operativa dei macchinari. Infine, essi consentono anche di prevedere le frodi, monitorando la nostra situazione bancaria, inviando notifiche per ogni transazione ed operazione bancaria.

1.4 Confronto tra Data Analytics e Data Analysis

Spesso i termini Data Analytics e Data Analysis sono considerati uguali e vengono ritenuti intercambiabili; essi, invece, presentano delle differenze sostanziali. La Data Analysis è un sottoinsieme della Data Analytics; si concentra, infatti, specificamente sull'analisi di un dataset per trovare informazioni utili, individuare pattern o rispondere a specifiche domande. La Data Analytics non si ferma lì ma comprende l'intero processo di lavoro, fino a supportare la strategia aziendale e il processo decisionale. L'obiettivo della Data Analytics, diversamente dalla Data Analysis, è di rispondere a domande specifiche sui dati già disponibili, organizzandoli e interrogandoli per estrarne significati e relazioni. In aggiunta, essa presenta differenze anche nel processo poiché si concentra soltanto sul dataset stesso, comprendendo soltanto pulizia, trasformazione e modellazione dei dati. Basta, infatti, pensare a quanto una dashboard interattiva trasmetta informazioni più approfondite e predittive, rispetto ad un report che mostra solo informazioni in tempo reale.

È possibile concludere dicendo che, nonostante la Data Analytics e la Data Analysis siano entrambe utili per analizzare dati, la Data Analytics consente un'analisi più approfondita e, di conseguenza, viene maggiormente consigliata.

CAPITOLO 2

Introduzione a Power BI

In questo secondo capitolo, verranno approfonditi gli aspetti relativi a Power BI, partendo dalla sua definizione e dai vantaggi, per poi analizzare la sua architettura. Infine, si proseguirà con un approfondimento sui principali settori di utilizzo.

2.1 Cos'è Power BI

Viviamo in un mondo in cui i dati acquisiscono un ruolo sempre più rilevante. Nelle aziende, sfruttare questi dati è di fondamentale importanza per riuscire ad essere competitivi nel proprio settore. Per questo motivo, è necessario avere a disposizione diversi strumenti in grado di raccogliere, analizzare e trasformare grandi quantità di informazioni in insight utili.

Power BI si inserisce in questo contesto come una soluzione di Business Intelligence versatile e accessibile, progettata per supportare aziende di qualsiasi dimensione nel processo decisionale. *Power BI* non è solo uno strumento tecnico, ma rappresenta una piattaforma strategica, capace di connettere dati eterogenei e di renderli disponibili in una forma comprensibile e visivamente intuitiva. La sua flessibilità lo rende adatto in molti settori.

2.1.1 Definizione e scopo

Power BI è una piattaforma sviluppata da Microsoft, che include vari servizi software che cooperano per trasformare grandi quantità di dati provenienti da fonti eterogenee, in informazioni da visualizzare ed analizzare.

Lo scopo principale di *Power BI* è dare alle aziende la possibilità di creare report e dashboard interattivi, per analizzare qualsiasi tipo di dati, in grado di supportare le decisioni da intraprendere. Attraverso il suo utilizzo, infatti, è possibile indentificare informazioni nascoste, che siano in grado di apportare un vantaggio all'azienda in qualsiasi ambito, soprattutto per tenere sotto controllo *KPI* (*Key Performance Indicators*) e altri indicatori cruciali per valutare il proprio andamento aziendale.

2.1.2 Vantaggi e benefici

La scelta di utilizzare *Power BI* è giustificata dalle seguenti motivazioni:

- *Versatilità*: poiché è capace di connettersi a una vasta gamma di sorgenti di dati, tra cui file *Excel*, database aziendali, servizi cloud e fonti dati di terze parti.

- *Facilità di utilizzo*: grazie alla sua interfaccia user-friendly (Figura 2.1) e ai vari oggetti che mette a disposizione per generare visualizzazioni semplici da impostare, questo strumento è efficace e, al contempo, semplice da utilizzare. Ciò implica che non c'è bisogno di avere grandi competenze, persino a livello di programmazione, per creare report relativamente semplici, favorendo, al tempo stesso, un'ottima comprensione dei dati.

  La facilità di utilizzo riguarda anche l'utente finale, il quale può accedere direttamente da cloud o applicazione mobile. Inoltre, le informazioni possono essere analizzate in modo interattivo, passando rapidamente da informazioni generali a informazioni dettagliate con pochi click.

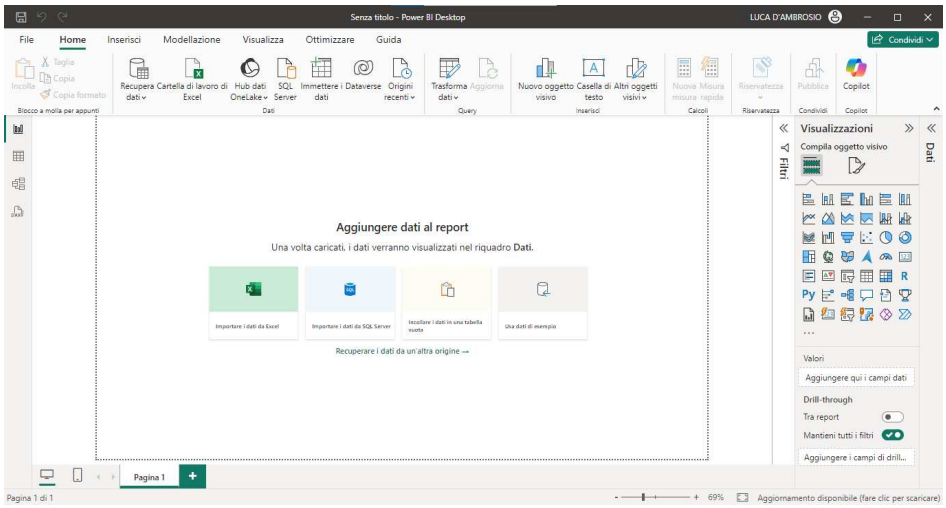

Figura 2.1: Interfaccia di Power BI

- *Economico*: la versione di *Power BI* Desktop è gratuita, mentre altre versioni, tra cui *Power BI* Service, hanno abbonamenti mensili a costi accessibili.

- *Sviluppo costante*: è il motivo per il quale *Power BI* continua ad essere una delle prime scelte. Microsoft dedica vari investimenti a questo strumento, fornendo nuovi oggetti di visualizzazione, cercando di migliorare le prestazioni, fornendo nuove soluzioni di integrazione dati e caratteristiche avanzate di Intelligenza Artificiale. Ciò garantisce che la piattaforma rimanga al passo con il nascere di nuove necessità aziendali.

- *Supporto nel campo della BI*: *Power BI* è fondamentale quando è il momento di attuare decisioni. Esso offre un notevole supporto tramite report e dashboard, permettendo a tutti i membri del team di scegliere una soluzione rispetto ad un'altra.

- *Sicurezza*: un vantaggio estremamente importante è la sicurezza, specialmente nei tempi odierni, in cui gli attacchi hacker sono aumentati notevolmente. Power BI offre un set completo di funzionalità di sicurezza, con lo scopo di proteggere i dati sensibili e rispettare i requisiti di conformità. Nell'archiviazione dei dati Power BI utilizza due repository; i dati caricati dagli utenti vengono generalmente trasferiti nell'Archiviazione BLOB di Azure, mentre tutti i metadati, compresi gli elementi per il sistema, sono archiviati nel database SQL di Azure.

- *Comunità e supporto*: *Power BI*, essendo sviluppato da Microsoft, è affiancato da un importante sistema di supporto, integrato da una vasta comunità attiva. Difatti, è alta la disponibità di tutorial e risposte specifiche a determinate problematiche.

2.1.3 Differenze rispetto ad altri strumenti di BI

Nel contesto della *Business Intelligence* (BI), esistono diverse piattaforme, oltre a *Power BI*, tra cui *Qlik*. Una delle principali differenze tra queste due soluzioni riguarda la gestione delle licenze. *Qlik* offre due tipi di piani, ovvero:

1 *Business*: prevede una tariffa mensile fissa, che consente di sfruttare appieno le potenzialità del software.

2 *Enterprise*: la tariffa è personalizzata e viene concordata direttamente con il reparto commerciale di *Qlik*, in base alle specifiche esigenze aziendali. Questo piano è pensato per le grandi imprese e richiede il supporto di un team tecnico per gestire accessi, pubblicazioni e aggiornamenti.

In confronto, Power BI si distingue per la possibilità di accedere a un piano gratuito, diversamente da *Qlik*.

Vi sono differenze anche nel *front-end* (*Interfaccia*), poiché, per sviluppare analisi più approfondite o complesse, *Qlik* mette a disposizione la *Set Analysis*, invece, *Power BI* propone il linguaggio *DAX*.

Non mancano, inoltre, differenze anche nel *back-end* (*Modello dati*). Qlik permette di svolgere operazioni di *ETL* in due modalità. La prima modalità è basata sull'*Editor caricamento dati* ed è ideata per gli sviluppatori per svolgere diversi passaggi tramite script, personalizzando operazioni complesse sui dati.

L'altra modalità è la *Gestione dati*, in cui possono essere condotte operazioni più basilari con dati meno complessi; essa è ideata per tutti gli utenti che non hanno competenze con i linguaggi di programmazione.

Power BI mette a disposizione il DAX per effettuare operazioni su tabelle e dati importati; inoltre, vi è l'opzione *Visualizzazione modello*, per individuare e modificare relazioni tra le varie tabelle.

Passando, ora, alle differenze tra *Tableau* e *Power BI*, una delle prime considerazioni riguarda l'interfaccia utente: Power BI ne offre una più intuitiva e semplice.

Tuttavia, un vantaggio significativo di Tableau è la sua maggiore compatibilità con i sistemi operativi. Tableau, infatti, è supportato sia su *Windows* che su *macOS*, mentre Power BI è disponibile esclusivamente per *Windows*.

Le differenze nei linguaggi di programmazione si riscontrano principalmente tra il *DAX*, di Power BI e il *Tableau Calculation Language*, di Tableau. Tuttavia, entrambi supportano *Python*, *R* ed *SQL*.

Infine, si può confrontare *Looker* con *Power BI*. Il punto di forza di Power BI rispetto a Looker è che offre una versione gratuita. Tuttavia, entrambi offrono diversi piani di abbonamenti.

Per quanto concerne l'integrazione e la compatibilità dei dati, Power BI, essendo un prodotto Microsoft, è conforme a tecnologie come *Azure, Excel, SharePoint*, ed offre anche connettività con *AWS*, *IBM* e le altre piattaforme più comuni per adattarsi a vari contesti di dati.

Looker, invece, oltre a fornire una ricca varietà di connettori, è progettato anche per adattarsi a fonti di dati meno comuni.

Un'altra differenza riguarda l'archiviazione dei dati; essi, una volta estratti, vengono memorizzati nella cache per un periodo limitato di tempo, per velocizzare tempi di recupero. Inoltre, Looker non include uno strumento dedicato alle attività di *ETL*.

Anche nella visualizzazione dei dati Power BI è più efficiente. La sua interfaccia user-friendly e gli oggetti visivi intuitivi lo rendono adatto anche a persone che non dispongono di grandi conoscenze. Al contrario, Looker è adeguato per persone esperte del settore.

2.2 Architettura e componenti

Per sfruttare *Power BI* al meglio, bisogna conoscere la sua architettura e individuare le varie componenti che costituiscono il software. Tali componenti vengono illustrate nelle prossime sottosezioni.

2.2.1 Power BI Desktop

Power BI Desktop è uno strumento che consente all'utente, non obbligatoriamente con esperienze pregresse di programmazione, di creare dashboard e report, importando, trasformando e visualizzando dati. Questa versione offre vari oggetti visivi di base, con la possibilità di importarne altri, creati da Microsoft o da altri membri della comunità.

Le principali funzionalità di questa piattaforma sono:

- *Connettersi ai dati*: la prima operazione da eseguire in Power BI consiste nel recuperare i dati, sia connettendosi a diverse piattaforme sia importandoli da dataset precedentemente scaricati.
- *Trasformare e pulire i dati*: una volta che i dati sono stati importati nel software, è necessario esaminarli per verificare i tipi di dato, assicurandosi che l'importazione sia avvenuta correttamente. In caso contrario, si dovranno eseguire altre operazioni che approfondiremo quando analizzeremo la componente Power Query.
- *Creare report e dashboard e condividerli*: i dati importati nel software saranno correlati tra loro. Per mostrare queste relazioni è necessario creare degli oggetti visivi che le rappresentino, rendendo le informazioni comprensibili a chi ne ha bisogno. L'insieme di questi oggetti visivi crea un *report* o *dashboard*, che, nella maggior parte dei casi, è interattivo. Questi strumenti sono frequentemente utilizzati nelle aziende per supportare le decisioni. Inoltre, è possibile condividere i report con altri utenti.

2.2.2 Power BI Service

Qualora un utente abbia la necessità di connettersi, visualizzare e analizzare dati in breve tempo, può utilizzare la componente *Power BI Service*, una piattaforma basata su *cloud* in cui sono presenti report e dashboard condivisi da altri utenti. Nel caso in cui gli utenti appartengano alla stessa organizzazione e *report* e *dashboard* siano resi pubblici, *Power BI* farà in modo che essi siano visibili.

Questa componente include altri strumenti, tra cui il più importante è *Workspace*.

Il concetto di *Workspace* in Power BI Service ricopre un ruolo fondamentale, poiché è lo strumento principale attraverso cui l'utente organizza i propri contenuti.

Questi ultimi avranno un *workspace* dedicato in cui essere salvati, e l'accesso sarà consentito a determinati utenti invece che ad altri. Per questo motivo, Power BI Service permette di assegnare accessi e ruoli agli utenti esterni, in base ai vari *workspace*.

Ci saranno utenti, ad esempio, che avranno la possibilità di visualizzare i contenuti, altri che potranno creare e modificare gli elementi, e altri ancora che avranno la possibilità di gestire i permessi al dataset, effettuare aggiornamenti e pubblicare nel *workspace*.

2.2.3 Power BI Mobile

Un altro strumento di *Power BI* che consente di sfruttare appieno le potenzialità di questo *software* è *Power BI Mobile*. Questa applicazione è disponibile per dispositivi Android, iOS e Microsoft. Grazie ad essa, è possibile accedere a dashboard e report in qualsiasi momento,

interagendo con essi sia in locale che nel cloud. Inoltre, è possibile impostare avvisi nel caso in cui i dati all'interno di un riquadro superino i limiti definiti.

2.2.4 Power Query e Power Pivot

Una delle ultime componenti che tratteremo, ma che è una tra le prime quando si inizia a lavorare con i dati è, *Power Query*. Questo strumento offre un'interfaccia grafica intuitiva per connettersi a diverse fonti dati e ha lo scopo di trasformare e preparare i dati. Esso consente, infatti, di eseguire attività di estrazione, trasformazione e caricamento (*ETL*) attraverso un editor dedicato. Essendo *Power Query* integrato in vari prodotti e servizi, la destinazione finale dei dati dipende dall'ambiente in cui esso viene utilizzato.

È importante notare che viene speso circa l'80% del tempo per la preparazione dei dati; di conseguenza dati non preparati ritardano il lavoro di analisi e il processo decisionale, motivo per cui è fondamentale l'uso di *Power Query*.

Power Pivot è uno strumento avanzato di *Excel* che merita di essere menzionato, poiché si integra perfettamente con *Power Query*. Mentre Power Query rappresenta lo strumento consigliato per l'importazione e la trasformazione dei dati, *Power Pivot* è progettato per la modellazione dei dati già importati. L'utilizzo combinato di *Power Query* e *Power Pivot* consente di modellare i dati in *Excel* in modo efficiente, offrendo funzionalità avanzate per esplorare e visualizzare i dati attraverso tabelle pivot, grafici pivot e report in *Power BI*.

2.3 Settori di applicazione

La versatilità di *Power BI* lo rende una delle soluzioni più adottate in diversi settori aziendali. In questa sezione approfondiremo i principali ambiti di applicazione di *Power BI*, evidenziando come esso abbia contribuito a migliorare l'efficienza operativa e come esso abbia ottimizzato i processi decisionali per raggiungere obiettivi specifici. In particolare, *Power BI* trova impiego nei seguenti settori:

- *Sanità*: in questo settore, esso viene utilizzato per analizzare i dati sulla salute dei pazienti, migliorando la prevenzione e l'efficacia dei trattamenti medici. Oltre ai pazienti, anche l'efficienza delle strutture ospedaliere può essere monitorata e ottimizzata.

- *Economia e finanza*: l'utilizzo di *Power BI* si concentra prevalentemente nel settore economico-finanziario, poiché, grazie ai vari *report* e *dashboard* sviluppati, possono essere condotte analisi su tutti gli ambiti di questi settori, dagli investimenti alle vendite, passando per l'analisi dei clienti fino allo sviluppo di nuove strategie di marketing.

- *Logistica e trasporti*: il traffico e le limitazioni dei mezzi rappresentano una sfida costante in questo settore. L'uso di *Power BI* risulta fondamentale per monitorare questi aspetti in modo da supportare l'adozione di nuove soluzioni.

- *Produzione industriale*: *Power BI* è ampiamente utilizzato nell'ambito della produzione industriale per ottimizzare i processi e migliorare l'efficienza operativa. Tra le applicazioni principali, lo strumento consente di analizzare l'utilizzo attuale e storico dei centri di lavoro, analizzare gli ordini di produzione completati e calcolare i tempi medi di produzione, identificando colli di bottiglia e pianificando la capacità produttiva futura. Inoltre, *Power BI* aiuta a ridurre gli sprechi monitorando gli scarti di produzione e le variazioni di consumo, fornendo, così, insight utili per migliorare la gestione complessiva dei processi industriali.

CAPITOLO 3

Descrizione dei dati di partenza

In questo capitolo esamineremo tre categorie di dati, per poi analizzare le tabelle presenti nel dataset di Formula 1 scelto.

3.1 Tipi di dati

Nel contesto dell'analisi dei dati è fondamentale conoscere le diverse tipologie di dati per selezionare gli strumenti e le tecniche più adeguati alla loro gestione e analisi. I dati, in base alle fonti di acquisizione, come sensori, database relazionali, web, social media e app per dispositivi mobili, possono essere classificati in tre categorie principali:

1 *dati strutturati*;

2 *dati semi-strutturati*;

3 *dati non strutturati*.

Analizzeremo queste tre categorie nei prossimi sottoparagrafi.

3.1.1 Dati strutturati

I dati strutturati sono informazioni descritte da un formato standardizzato che consente un accesso efficiente sia da parte di utenti che di software. Solitamente, i dati strutturati sono organizzati in tabelle, in cui righe e colonne specificano in modo preciso gli attributi, ogni colonna esprime una categoria di dati e ogni riga rappresenta una collezione di dati connessi. Questo tipo di dati viene frequentemente inserito, analizzato e manipolato in fogli di calcolo e database relazionali, come *MySQL, Oracle Database o Microsoft SQL Server*.

I dati strutturati presentano le seguenti caratteristiche e vantaggi:

- *Scalabilità*: i dati possono essere scalati utilizzando algoritmi. Lo spazio di archiviazione e la capacità computazionale possono essere incrementati man mano che cresce il volume dei dati. Infatti, i sistemi di dati strutturati, come i database relazionali, sono progettati per gestire grandi volumi di dati, arrivando fino a diversi TB (*terabyte*).
- *Capacità relazionali*: i dati strutturati consentono relazioni tra varie tabelle o entità. Queste relazioni rendono i grandi dataset interconnessi tra loro; inoltre, tramite il linguaggio

SQL, gli utenti possono recuperare in modo facile ed efficiente le informazioni specifiche ed eseguire operazioni complesse sui dati.

- *Integrità dei dati*: grazie al loro formato predefinito, i dati strutturati seguono regole precise che garantiscono accuratezza, validità e affidabilità. La presenza di procedure di convalida dei dati riduce significativamente la possibilità di errori e ridondanza, rendendo i dati strutturati affidabili e semplici da elaborare.
- *Analisi efficiente dei dati*: i dati strutturati, grazie alla loro struttura, si integrano agevolmente con strumenti di analisi e Business Intelligence, permettendo alle aziende di effettuare analisi statistica, modellazione predittiva, nonché creare report per estrarre informazioni utili, semplificando i processi decisionali.

Tuttavia, vi sono anche degli svantaggi, ovvero:

- *Flessibilità limitata*: a causa della loro struttura predefinita, alcuni dati, come immagini, video e linguaggio naturale, potrebbero non essere compatibili con i formati strutturati.
- *Problemi di manutenzione e standardizzazione*: con l'aumento della dimensione e della complessità dei dati, lo schema predefinito deve essere costantemente aggiornato per integrare nuovi tipi di dati o relazioni, rallentando la gestione dei dati, soprattutto nelle elaborazioni in tempo reale.
- *Aumento dei costi*: questi database, aumentando progressivamente, possono richiedere molte risorse, specialmente per le grandi aziende che devono archiviare grandi quantità di dati. Inoltre, con la necessità di personale specializzato, il costo complessivo aumenta notevolmente.

I dati strutturati, grazie al loro formato predefinito, sono semplici da analizzare, motivo per cui vengono utilizzati in vari settori, tra cui:

- *Transazioni finanziarie*: grazie a questi dati le aziende riescono a tenere traccia facilmente di entrate e uscite, creando report velocemente.
- *Gestione delle relazioni con i clienti*: i dati strutturati vengono utilizzati per registrare le informazioni dei clienti; infatti, aiutano le aziende a individuare tendenze e strategie di marketing per essi.
- *Cartelle cliniche e dati medici*: i dati delle cartelle cliniche dei pazienti sono dati strutturati. Ciò permette agli operatori sanitari di accedere velocemente a questi dati, garantire diagnosi accurate e monitorare i risultati dei trattamenti.
- *Ottimizzazione dei motori di ricerca*: i motori di ricerca sfruttano il markup dei dati strutturati per interpretare più accuratamente il contenuto delle pagine web. Grazie ad essi, i siti web possono migliorare la loro visibilità e la visualizzazione delle informazioni.

3.1.2 Dati semi-strutturati

I dati semi-strutturati sono un formato di dati che non segue la rigida struttura dei tradizionali database relazionali, ma include comunque elementi organizzativi, come tag o marcatori, che ne semplificano l'analisi; esempi di dati semi-strutturati sono quelli relativi a *XML, JSON*, metadati e-mail e dati raccolti da sensori. Questo tipo di dati consente un parsing e un'interpretazione più semplice rispetto ai dati non strutturati, con la possibilità di rappresentare le informazioni in modo diverso. I dati semi-strutturati sono un'ottima soluzione per i sistemi che richiedono flessibilità e scalabilità; essi presentano, infatti, diverse caratteristiche e vantaggi, ovvero:

- *Schema libero*: questi dati non presentano uno schema predefinito, motivo per cui sono ideali per applicazioni dove i formati dei dati variano spesso.
- *Autodescrittivi*: nella maggior parte dei casi, i dati semi-strutturati sono autodescrittivi. Infatti, gli elementi che li compongono includono metadati che ne specificano il significato o la struttura. Un esempio è un documento *XML* che utilizza tag per identificare i tipi di dati contenuti, rendendoli più facili da interpretare anche in assenza di uno schema fisso.
- *Supporto per tipi di dati complessi*: essi, rappresentando dati più complessi rispetto ai dati strutturati, come, ad esempio, array e oggetti nidificati, sono funzionali per applicazioni che gestiscono diversi tipi di formati e relazioni di dati.
- *Facilità di integrazione con il web e servizi cloud*: molti siti web e applicazioni basati sul cloud, come *API* e *database NoSQL*, utilizzano dati semi-strutturati, come *JSON* e *XML*, per la condivisione delle informazioni.
- *Scalabilità*: i dati semi-strutturati sono capaci di gestire enormi quantità di dati eterogenei senza sacrificare le prestazioni. Ciò li rende particolarmente vantaggiosi nei contesti di Big Data, dove è essenziale archiviare ed elaborare volumi crescenti di informazioni in modo efficiente.

Tuttavia, in questa tipologia di dati sono presenti anche alcuni svantaggi, ovvero:

- *Complessità nelle interrogazioni*: i dati semi-strutturati, non avendo uno schema fisso, rendono più complessa l'esecuzione di query avanzate. Inoltre, i linguaggi di interrogazione specifici per questo tipo di dati, come *XPath* per *XML* e *JSONPath* per *JSON*, sono meno sviluppati del linguaggio *SQL*.
- *Aumento delle spese di elaborazione*: l'elaborazione dei dati semi-strutturati richiede un sostanziale impiego di risorse computazionali. La necessità di interpretare e gestire strutture dati flessibili implica un aumento del carico di elaborazione, che può rallentare le prestazioni delle applicazioni, specialmente in contesti ad alto volume o in tempo reale.
- *Problemi di convalida dei dati*: a causa della mancanza di una struttura predefinita, aumentano le probabilità di incongruenze, errori e duplicazioni.

3.1.3 Dati non strutturati

A differenza dei dati strutturati, che seguono una definizione chiara, i dati non strutturati ne sono privi. Anziché utilizzare campi predefiniti in un formato specifico, i dati non strutturati possono assumere qualsiasi forma e dimensione. Questi dati possono comprendere testo libero, e-mail, video, immagini, file audio e post sui social media. I dati non strutturati si distinguono per la loro varietà, la loro velocità e il loro volume, caratteristiche che ritroviamo nei *Big Data*. Ovviamente anche i dati non strutturati presentano sia vantaggi che svantaggi. Tra i vantaggi abbiamo:

- *Flessibilità e variazione delle informazioni*: poiché provengono da varie fonti e hanno molteplici formati, essi possono portare a novità e deduzioni che nei dati strutturati non scopriremmo.
- *Soluzioni innovative*: le aziende, attraverso l'analisi dei dati non strutturati, potrebbero identificare nuove tendenze e pattern, che possono portare a nuove idee per il marketing.

- *Applicazioni in Intelligenza Artificiale e Machine Learning*: per allenare gli algoritmi di *Intelligenza Artificiale* e *Machine Learning* i dati non strutturati sono fondamentali; contribuendo a migliorare processi automatizzati nelle aziende.

Gli svantaggi presenti nei dati non strutturati sono i seguenti:

- *Difficoltà di gestione e analisi*: a causa della mancanza di una struttura fissa, la gestione e l'analisi di questi dati sono molto complessi. Infatti, essi richiedono strumenti e competenze avanzate. Inoltre, i tempi di analisi dei dati non strutturati sono elevati.
- *Costi di archiviazione*: poiché la maggior parte dei dati è non strutturata, le aziende devono sostenere costi significativi per archiviarli.

3.2 Formula 1 dataset

Lo studio condotto in questa tesi deriva da dati provenienti da Kaggle, una piattaforma che ospita una vasta gamma di dataset per ogni ambito, in cui persone che praticano *data science* e *machine learning* collaborano per risolvere avvincenti competizioni. Il dataset ottenuto da Kaggle, prende il nome di "Formula 1 World Championship (1950-2024)".

Esso contiene dati relativi a circuiti, piloti, costruttori e gare di Formula 1. L'obiettivo è quello di analizzare questi dati per individuare il cambiamento e lo sviluppo di questo sport, analizzando i vari elementi che lo compongono.

Il dataset è formato da 14 tabelle, che possono essere scaricate in formato *CSV* (Comma-Separated Values). Nelle prossime sottosezioni illustreremo più in dettaglio tali tabelle.

3.2.1 Circuiti

Per quanto riguarda i circuiti è presente una tabella chiamata *circuits*, nella quale sono presenti varie informazioni relative a tutti i circuiti presenti in tutto il mondo. Questa tabella è composta da 77 righe e 9 colonne, in cui sono presenti i seguenti parametri:

1. *circuitId*: un numero intero univoco associato a ciascun circuito;
2. *circuitRef*: un codice abbreviato che identifica univocamente ciascun circuito;
3. *name*: il nome completo del circuito;
4. *location*: il nome della città in cui è presente il circuito;
5. *country*: il nome della nazione che ospita il circuito;
6. *lat*: la posizione geografica, espressa in gradi, del circuito rispetto all'equatore;
7. *lng*: la posizione geografica, espressa in gradi, del circuito rispetto al meridiano di Greenwich;
8. *alt*: l'altezza espressa in metri del circuito rispetto al livello del mare;
9. *url*: un link del circuito su *Wikipedia*.

3.2.2 Costruttori

Per i costruttori sono presenti 3 tabelle, ovvero: *constructors* con informazioni generali su ciascuna scuderia, *constructor_results*, in cui sono presenti i risultati di ogni scuderia per ogni gara e, infine, *constructor_standings*, nella quale sono presenti informazioni sulla classifica costruttori per ogni stagione.

Per quanto riguarda la tabella *constructors*, formata da 212 righe e 5 colonne, vi sono i seguenti campi:

1. *constructorId*: un numero intero univoco per ogni scuderia;
2. *constructorRef*: un codice di rifermento univoco per ciascuna scuderia;
3. *name*: il nome proprio della scuderia;
4. *nationality*: la nazionalità della scuderia;
5. *url*: un link della scuderia su *Wikipedia*.

La tabella *constructor_results* formata da 12.505 righe e da 5 colonne presenta i seguenti campi:

1. *constructorResultId*: un numero intero univoco per identificare un determinato risultato di una scuderia;
2. *raceId*: un numero intero univoco per ciascuna gara;
3. *constructorId*: un numero intero univoco per ogni scuderia;
4. *points*: i punti guadagnati da una scuderia in una determinata gara;
5. *status*: indica lo stato di una monoposto a fine gara.

Infine, vi è la tabella *constructor_standings* formata da 13.271 righe e 7 colonne. Essa contiene i seguenti campi:

1. *constructorStandingId*: un identificativo univoco usato per distinguere le singole voci relative ai costruttori in diverse gare;
2. *raceId*: un numero intero univoco per ciascuna gara;
3. *constructorId*: un numero intero univoco per ogni scuderia;
4. *points*: somma dei punti guadagnati nelle gare disputate da una scuderia fino a quel punto del campionato;
5. *position*: un numero per identificare la posizione in classifica di una scuderia;
6. *positionText*: un campo di testo per identificare la posizione in classifica di una scuderia;
7. *wins*: rappresenta il numero totale di vittorie ottenute da una scuderia fino a quel punto del campionato.

3.2.3 Piloti

Per i piloti sono presenti 2 tabelle, *drivers* e *driver_standings*, nelle quali sono presenti dati anagrafici dei piloti e informazioni relative alla classifica piloti. I campi della tabella *drivers*, che è composta da 859 righe e 9 colonne, sono i seguenti:

1. *driverId*: un numero univoco per ciascun pilota;
2. *driverRef*: un identificativo di riferimento per ogni pilota;
3. *number*: un numero assegnato al pilota, utilizzato per identificarlo ufficialmente durante le gare;
4. *code*: un'abbreviazione univoca per ogni pilota;
5. *forename*: il nome proprio del pilota;
6. *surname*: il cognome del pilota;
7. *dob*: sta per "date of birth" e indica la data di nascita del pilota;
8. *nationality*: la nazionalità del pilota;
9. *url*: un link del pilota su *Wikipedia*.

La tabella *driver_standings* formata da 34.595 righe e 7 colonne, ha i seguenti campi:

1. *driverStandingsId*: un identificativo univoco usato per distinguere le singole voci relative ai piloti in diverse gare;
2. *raceId*: un numero intero univoco per ciascuna gara;
3. *driverId*: numero univoco per ciascun pilota;
4. *points*: la somma dei punti guadagnati nelle gare disputate da un pilota fino a quel punto del campionato;
5. *position*: un numero per identificare la posizione in classifica di un pilota in una determinata fase del campionato;
6. *positionText*: un campo di testo per identificare la posizione in classifica di un pilota in una determinata fase del campionato;
7. *wins*: il numero di vittorie conquistate da un pilota in una determinata fase del campionato.

3.2.4 Gare

In questo ultimo sottoparagrafo vengono approfonditi i dati relativi alle gare, che sono presenti in 8 tabelle.

La prima tabella è *lap_times*; essa è formata da 575.029 righe e 6 colonne. I campi di questa tabella sono i seguenti:

1. *raceId*: un numero intero univoco per ciascuna gara;
2. *driverId*: un numero univoco per ciascun pilota;
3. *lap*: il numero del giro di una gara;

4. *position*: la posizione del pilota in gara durante un determinato giro;
5. *time*: il tempo impegato per percorrere un determinato giro in formato minuti, secondi, millisecondi;
6. *milliseconds*: il tempo impegato per percorrere un determinato giro, tutto in formato millisecondi.

Un'altra tabella è *pit_stops*; essa è formata da 10.990 righe e 7 colonne, e contiene i dati relativi ai pit-stop, che sono i seguenti:

1. *raceId*: un numero intero univoco per ciascuna gara;
2. *driverId*: un numero univoco per ciascun pilota;
3. *stop*: il numero di pit-stop effettuati in una determinata fase della gara;
4. *lap*: il numero del giro di una gara;
5. *time*: l'orario in cui è stato effettuato il pit-stop;
6. *duration*: il tempo impiegato per effetturare il pit-stop in formato secondi e millisecondi;
7. *milliseconds*: il tempo impiegato per effetturare il pit-stop in formato millisecondi.

In seguito, vi è la tabella *qualifying*, la quale è composta da 10.254 righe e 9 colonne. I campi di questa tabella sono i seguenti:

1. *qualifyId*: un codice univoco per identificare una determinata qualifica;
2. *raceId*: un numero intero univoco per ciascuna gara;
3. *driverId*: un numero univoco per ciascun pilota;
4. *constructorId*: un numero intero univoco per ogni scuderia;
5. *number*: il numero assegnato al pilota, utilizzato per identificarlo ufficialmente durante le gare;
6. *position*: la posizione consquistata in qualifica;
7. *q1*: il tempo sul giro della prima fase della qualifica;
8. *q2*: il tempo sul giro della seconda fase della qualifica;
9. *q3*: il tempo sul giro della fase finale della qualifica.

Dopodiché, c'è la tabella *races*. Essa è composta da 1.125 righe e 18 colonne, che sono le seguenti:

1. *raceId*: un numero intero univoco per ciascuna gara;
2. *year*: un anno;
3. *round*: il numero del gran premio in una determinata stagione;
4. *circuitId*: un numero intero univoco associato a ciascun circuito;
5. *name*: il nome completo del circuito;

6. *date*: la data in cui si è disputato il gran premio;
7. *time*: l'orario in cui è iniziato il gran premio;
8. *url*: un link del gran premio su *Wikipedia*;
9. *fp1_date*: la data delle prime prove libere;
10. *fp1_time*: l'orario in cui iniziano le prime prove libere;
11. *fp2_date*: la data delle seconde prove libere;
12. *fp2_time*: l'orario in cui iniziano le seconde prove libere;
13. *fp3_date*: la data delle terze ed ultime prove libere;
14. *fp3_time*: l'orario in cui iniziano le terze prove libere;
15. *quali_date*: la data in cui si svolgono le qualifiche;
16. *qualy_time*: l'orario in cui iniziano le qualifiche;
17. *sprint_date*: la data in cui si svolge la gara sprint;
18. *sprint_time*: l'orario in cui inizia la gara sprint.

Una tabella molto importante per le analisi che sono state condotte è la tabella *results*; essa è composta da 26.519 righe e 17 colonne, e presenta i seguenti campi:

1. *resultId*: un codice univoco per identificare il risultato di un determinato pilota in una determinata gara;
2. *raceId*: un numero intero univoco per ciascuna gara;
3. *driverId*: un numero univoco per ciascun pilota;
4. *constructorId*: un numero intero univoco per ogni scuderia;
5. *number*: il numero assegnato al pilota, utilizzato per identificarlo ufficialmente durante le gare;
6. *grid*: la posizione del pilota nella griglia di partenza;
7. *positionOrder*: la posizione del pilota a fine gara;
8. *positionText*: la posizione del pilota a fine gara in formato stringa;
9. *points*: i punti guadagnati dal pilota in una determinata gara;
10. *laps*: il numero di giri di una gara percorsi dal pilota;
11. *time*: il tempo impegato dal pilota per completare tutta la gara;
12. *milliseconds*: il tempo impegato dal pilota per completare tutta la gara in millisecondi;
13. *fastestLap*: il giro in cui è stato registrato il giro più veloce personale dal pilota in tutta la gara;
14. *rank*: la posizione in classifica del giro veloce del pilota;

15. *fastestLapTime*: il tempo del giro più veloce del pilota in formato minuti, secondi e millisecondi;
16. *fastestLapSpeed*: la massima velocità raggiunta dal pilota in tutta la gara;
17. *statusId*: un codice univoco per identificare lo stato di una monoposto a fine gara.

In seguito, vi è la tabella *season* formata da 75 righe e 2 colonne, ovvero:

1. *year*: anno per ogni stagione di Formula 1;
2. *url*: link *Wikipedia* a ciascuna stagione.

La penultima tabella di questo dataset è *sprint_results*, che è formata da 300 righe e 15 colonne, ovvero:

1. *resultId*: un codice univoco per identificare una gara sprint;
2. *raceId*: un numero intero univoco per ogni gara sprint;
3. *driverId*: un numero univoco per ciascun pilota;
4. *constructorId*: un numero intero univoco per ogni scuderia;
5. *number*: un numero assegnato al pilota, utilizzato per identificarlo ufficialmente durante le gare;
6. *grid*: la posizione in griglia di partenza;
7. *positionOrder*: la posizione del pilota a fine gara sprint;
8. *positionText*: la posizione del pilota a fine gara sprint in formato stringa;
9. *points*: i punti guadagnati dal pilota in una determinata gara sprint;
10. *laps*: il numero del giro di una gara sprint percorsi dal pilota;
11. *time*: il tempo impegato dal pilota per completare tutta la gara sprint;
12. *milliseconds*: il tempo impegato dal pilota per completare tutta la gara sprint in millisecondi;
13. *fastestLap*: il giro della gara in cui è stato registrato il giro più veloce personale del pilota in tutta la gara sprint;
14. *fastestLapTime*: il tempo del giro veloce del pilota in formato minuti, secondi e millisecondi;
15. *statusId*: un codice univoco per identificare lo stato di una monoposto a fine gara sprint.

Infine, l'ultima tabella è *status*, la quale è formata da 139 righe e 2 colonne, ovvero:

1. *statusId*: un codice univoco per identificare lo stato di una monoposto a fine gara;
2. *status*: indica lo stato di una monoposto a fine gara.

CAPITOLO 4

Descrizione delle attività di ETL

In questo capitolo verranno approfonditi il concetto di ETL e le attività di ETL svolte sul dataset di Formula 1.

4.1 Premessa

Una delle attività da condurre, ogni volta che si vuole effettuare l'analisi dei dati, è l'*Extract-Transform-Load* (*ETL*). Essa è fondamentale poiché si occupa di estrarre i dati da diverse fonti, trasformarli e integrarli, nella maggior parte dei casi, in un *data warehouse*.

Questa attività, anche se non è visibile nel risultato finale di un processo di analisi dei dati, richiede un investimento significativo in termini di tempo, risorse e complessità, contribuendo a condurre un'analisi accurata e affidabile.

Le tre fasi dell'*ETL* sono: Estrazione, Trasformazione e Caricamento. Esse verranno illustrate in dettaglio nelle prossime sezioni.

4.2 Estrazione

La prima fase dell'*ETL* è l'estrazione. Essa, oltre ad essere la fase iniziale, concettualmente è anche la più semplice; si tratta concretamente di estrarre diverse tipologie di dati da varie sorgenti, ad esempio, database relazionali, fogli di calcolo e siti web.

Inoltre, ciascuna sorgente di dati, possedendo un insieme unico di caratteristiche, deve essere gestita per estrarre efficacemente i dati necessari al processo di ETL.

È fondamentale che il processo riesca a collegare e integrare sistemi che operino su piattaforme diverse in modo da garantire il successo del data warehousing. Poiché questa fase viene eseguita con periodicità, si consiglia di eseguirla durante i periodi di inattività del sistema sorgente, in modo da non sovraccaricarlo con altre attività.

Tra le diverse soluzioni di estrazione dei dati, una delle più consigliate è quella di estrarre un'istantanea dei dati e confrontarla con l'istantanea precedente, così da individuare inserimenti, eliminazioni e aggiornamenti. Questo approccio permette di evitare la rielaborazione dei dati rimasti invariati.

4.2.1 Attività di estrazione del dataset

L'estrazione dei dati in Power BI è molto semplice da eseguire grazie alla sua interfaccia user-friendly. Dopo aver effettuato il download del *Formula 1 dataset* da Kaggle, esso è stato importato su Power BI. Quest'ultimo passaggio può essere eseguito seguendo tre approcci.

Il primo metodo consiste nell'utilizzo della barra degli strumenti, che Power BI mette a disposizione nell'interfaccia iniziale, come riportato in Figura 4.1.

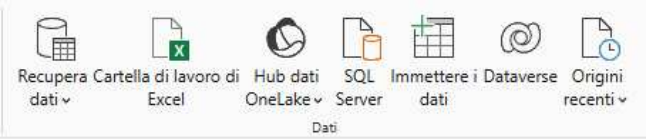

Figura 4.1: Interfaccia di Power BI per il recupero dei dati

Altrimenti, si potrà utilizzare la schermata che successivamente ospiterà la *dashboard*, come riportato in Figura 4.2.

Anch'essa è visibile nell'interfaccia iniziale, dopo la creazione di un nuovo progetto.

Figura 4.2: Ulteriore interfaccia di Power BI per il recupero dei dati

Infine vi è il terzo approccio, che consiste nel cliccare nella scritta in verde "*Recuperare i dati da un'altra origine*", visibile in Figura 4.2; dopodiché si aprirà un'ulteriore schermata, mostrata in Figura 4.3.

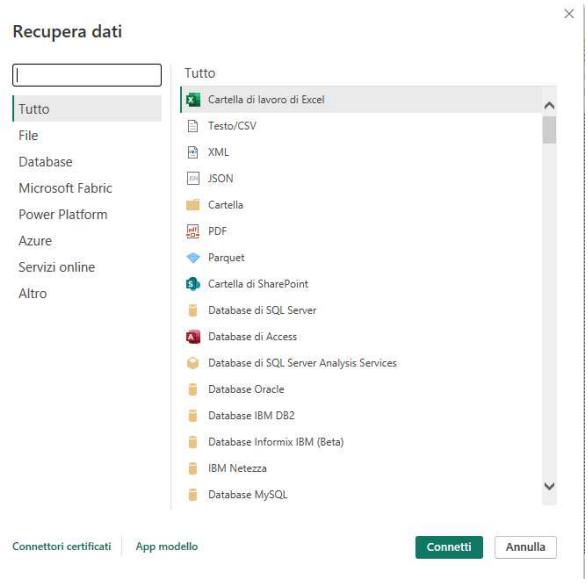

Figura 4.3: Interfaccia di Power BI per il recupero dei dati da molteplici sorgenti

4.3 Trasformazione

La fase successiva è quella di Trasformazione. Durante questa fase, i dati precedentemente estratti vengono puliti e conformati per ottenere informazioni corrette e complete. L'importanza di questa fase è cruciale, in quanto una corretta trasformazione dei dati è essenziale per garantire che dalle analisi successive si riescano a estrarre informazioni significative.

Le attività di trasformazione spesso riguardano: la mappatura diretta, la manipolazione dei singoli campi, la gestione delle stringhe di testo, la determinazione dei valori nulli, la conversione dei dati e le operazioni sulle date.

4.3.1 Attività di trasformazione relative al dataset

In questa fase, ci si è concentrati sulla pulizia e nella modifica di questi dati, in modo da condurre in seguito delle analisi significative.

La prima tabella da cui si è partiti è *results*, la quale inizialmente presentava il campo *positionText* in formato stringa, ma il campo era comunque un numero intero.

Poiché è già presente la colonna *position* all'interno di questa tabella come campo numerico, per evitare ripetizioni si è deciso di sostituirlo con il campo testo.

Questo processo è stato eseguito con l'ausilio del codice Python riportato, Listato 4.1. Inoltre, è stato modificato il campo *position* dove era presente il valore "\N", poiché risulta problematico per le analisi che si dovranno effettuare e anche per convertirlo a numero intero. La procedura seguita è quella riportata nelle Figure 4.4, 4.5 e 4.6.

```python
import pandas as pd


from num2words import num2words


def replace_numbers_with_strings(input_file, output_file, column_name):
    try:
        # Legge il file CSV specificando il delimitatore ';'
        df = pd.read_csv(input_file, delimiter=';')
        # Stampa i nomi delle colonne per il debug
        print("Colonne nel file:", df.columns)

        # Verifica se la colonna esiste
        if column_name not in df.columns:

            print(f"Errore: La colonna '{column_name}' non esiste nel file CSV.")
            return


        # Funzione per sostituire i numeri con le parole e gestire 'R' come 'ritirato'
        def number_to_word(x):
            if pd.notnull(x):
                if x == "R":
                    return "ritirato"
                try:

                    # Converte il numero in parola (in italiano)
                    return num2words(int(x), lang='it')
                except ValueError:

                    return str(x)

            return ''
```

```
35
36
37
38          # Applica la funzione di conversione
39          df[column_name] = df[column_name].apply(number_to_word)
40
41
42
43          # Salva il DataFrame aggiornato in un nuovo file CSV
44          df.to_csv(output_file, index=False, sep=';')  # Usa il punto e virgola
        come delimitatore
45
46
47          print(f"File modificato salvato come '{output_file}'.")
48
49
50      except Exception as e:
51          print(f"Errore durante l'elaborazione: {e}")
52
53
54  input_file = r"C:\Users\lucadambro\Downloads\results.csv"
55  # Percorso del file di input
56
57
58
59  output_file = r"C:\Users\lucadambro\Downloads\results_updated_new.csv"  #
        Percorso del file di output
60
61
62  column_name = 'positionText'  # Nome della colonna da modificare
63
64
65
66  replace_numbers_with_strings(input_file, output_file, column_name)
```

Listing 4.1: Codice Python per la conversione dei numeri in parole

Figura 4.4: Componente dell'interfaccia di Power BI per sostituire i valori

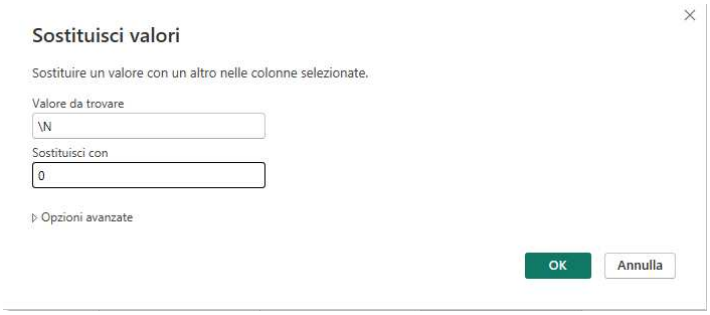

Figura 4.5: Interfaccia per sostituire i valori

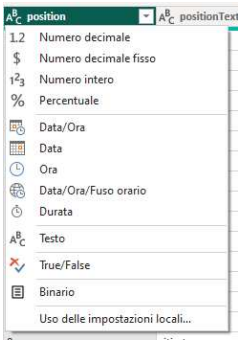

Figura 4.6: Interfaccia per convertire il tipo di valore

Il risultato di queste modifiche può essere visualizzato concretamente nelle Figure 4.7 e 4.8.

	grid	position	positionText	positionOrder	points
1	22	1	1	1	10
2	3	5	2	2	8
3	7	7	3	3	6
4	5	11	4	4	5
5	23	3	5	5	4
6	8	13	6	6	3
7	14	17	7	7	2
8	1	15	8	8	1
9	4	2	\N	R	9	0
10	12	18	\N	R	10	0
11	18	19	\N	R	11	0
12	6	20	\N	R	12	0
13	2	4	\N	R	13	0
14	9	8	\N	R	14	0
15	11	6	\N	R	15	0
16	20	22	\N	R	16	0
17	10	14	\N	R	17	0
18	16	12	\N	R	18	0
19	19	21	\N	R	19	0
20	15	9	\N	R	20	0
21	21	16	\N	R	21	0
22	17	10	\N	D	22	0
23	1	2	1	1	1	10

Figura 4.7: Tabella results iniziale

	number	grid	position	positionText	positionOrder
1	1	22	1	1	uno	1
2	2	3	5	2	due	2
3	3	7	7	3	tre	3
4	4	5	11	4	quattro	4
5	1	23	3	5	cinque	5
6	3	8	13	6	sei	6
7	5	14	17	7	sette	7
8	6	1	15	8	otto	8
9	2	4	2	0	ritirato	9
10	7	12	18	0	ritirato	10
11	8	18	19	0	ritirato	11
12	4	6	20	0	ritirato	12
13	6	2	4	0	ritirato	13
14	9	9	8	0	ritirato	14
15	7	11	6	0	ritirato	15
16	10	20	22	0	ritirato	16
17	9	10	14	0	ritirato	17
18	11	16	12	0	ritirato	18
19	8	19	21	0	ritirato	19
20	5	15	9	0	ritirato	20
21	10	21	16	0	ritirato	21

Figura 4.8: Tabella results finale

La stessa procedura è stata eseguita per la tabella *sprint_results*. Si possono visualizzare le

modifiche nelle Figure 4.9 e 4.10.

	constructorId	number	grid	position	positionText	
1	830	9	33	2	1	
2	1	131	44	1	2	
3	822	131	77	3	3	
4	844	6	16	4	4	
5	846	1	4	6	5	
6	817	1	3	7	6	
7	4	214	14	11	7	
8	20	117	5	10	8	
9	847	3	63	8	9	
10	839	214	31	13	10	
11	832	6	55	9	11	
12	842	213	10	12	12	
13	8	51	7	17	13	
14	840	117	18	15	14	
15	841	51	99	14	15	
16	852	213	22	16	16	
17	849	3	6	18	17	
18	854	210	47	19	18	
19	853	210	9	20	19	
20	815	9	11	5	\N	
21	822	131	77	1	1	
22	830	9	33	3	2	
23	817	1	3	5	3	

Figura 4.9: Tabella sprint iniziale

	position	positionText	positionOrder	points	laps
1	2	1	uno	1	3
2	1	2	due	2	2
3	3	3	tre	3	1
4	4	4	quattro	4	0
5	6	5	cinque	5	0
6	7	6	sei	6	0
7	11	7	sette	7	0
8	10	8	otto	8	0
9	8	9	nove	9	0
10	13	10	dieci	10	0
11	9	11	undici	11	0
12	12	12	dodici	12	0
13	17	13	tredici	13	0
14	15	14	quattordici	14	0
15	14	15	quindici	15	0
16	16	16	sedici	16	0
17	18	17	diciassette	17	0
18	19	18	diciotto	18	0
19	20	19	diciannove	19	0
20	5	0	ritirato	20	0
21	1	1	uno	1	3
22	3	2	due	2	2
23	5	3	tre	3	1

Figura 4.10: Tabella sprint finale

Un'ulteriore modifica è avvenuta nella tabella *circuits*, poiché, essendo i numeri decimali definiti con il punto anziché con la virgola, quando la tabella veniva importata in *Power BI*, i campi di latitudine e longitudine venivano interpretati come numeri interi. Nel Listato 4.2 è riportato il codice Python utilizzato per risolvere questa problematica. Inoltre, possiamo visualizzare le modifiche apportate nelle Figure 4.11 e 4.12

```python
import pandas as pd

def replace_decimal_and_save_with_semicolon(file_path, output_file):
    try:
        # Legge il file CSV in un DataFrame
```

```
        df = pd.read_csv(file_path)

        # Controlla se le colonne "lat" e "lng" esistono nella tabella
        missing_columns = [col for col in ['lat', 'lng'] if col not in
    df.columns]
        if missing_columns:
            print(f"Errore: le seguenti colonne non esistono nel file
    '{file_path}': {', '.join(missing_columns)}")
            return

        # Sostituisce il punto decimale con la virgola nelle colonne 'lat' e
    'lng'
        for col in ['lat', 'lng']:
            df[col] = df[col].apply(lambda x: str(x).replace('.', ',') if
    pd.notnull(x) else x)

        # Salva il risultato in un nuovo file con separatore punto e virgola
        df.to_csv(output_file, sep=';', index=False)
        print(f"Operazione completata. File salvato come '{output_file}'.")

    except FileNotFoundError:
        print(f"Errore: file '{file_path}' non trovato.")
    except pd.errors.EmptyDataError:
        print(f"Errore: file '{file_path}' vuoto o non valido.")
    except Exception as e:
        print(f" errore verificato: {e}")


file_path = r"C:\Users\lucadambro\Downloads\circuits_updated2.csv"        #
    Percorso del file di input

output_file = r"C:\Users\lucadambro\Downloads\circuits_update2_updated2.csv"  #
    Percorso del file di output

replace_decimal_and_save_with_semicolon(file_path, output_file)
```

Listing 4.2: Sostituzione del punto con la virgola nelle colonne lat e lng

	name	location	country	lat	lng	
1	ert Park Grand Prix Circuit	Melbourne	Australia	-378497	144968	
2	ang International Circuit	Kuala Lumpur	Malaysia	276083	101738	
3	rain International Circuit	Sakhir	Bahrain	260325	505106	
4	uit de Barcelona-Catalunya	Montmeló	Spain	4157	226111	
5	nbul Park	Istanbul	Turkey	409517	29405	
6	uit de Monaco	Monte-Carlo	Monaco	437347	742056	
7	uit Gilles Villeneuve	Montreal	Canada	455	-735228	
8	uit de Nevers Magny-Cours	Magny Cours	France	468642	316361	
9	erstone Circuit	Silverstone	UK	520786	-101694	
10	kenheimring	Hockenheim	Germany	493278	856583	
11	garoring	Budapest	Hungary	475789	192486	
12	encia Street Circuit	Valencia	Spain	394589	-331667	
13	uit de Spa-Francorchamps	Spa	Belgium	504372	597139	
14	odromo Nazionale di Monza	Monza	Italy	456156	928111	
15	ina Bay Street Circuit	Marina Bay	Singapore	12914	103864	
16	Speedway	Oyama	Japan	353717	138927	
17	nghai International Circuit	Shanghai	China	313389	12122	
18	ódromo José Carlos Pace	São Paulo	Brazil	-237036	-466997	
19	anapolis Motor Speedway	Indianapolis	USA	39795	-862347	
20	burgring	Nürburg	Germany	503356	69475	
21	odromo Enzo e Dino Ferrari	Imola	Italy	443439	117167	
22	uka Circuit	Suzuka	Japan	348431	136541	
23	Vegas Strip Street Circuit	Las Vegas	United States	361147	-115173	

Figura 4.11: Tabella circuits iniziale

	me	location	country	lat	lng	
1	Park Grand Prix Circuit	Melbourne	Australia	-37,8497	144,968	
2	International Circuit	Kuala Lumpur	Malaysia	27,6083	101,738	
3	International Circuit	Sakhir	Bahrain	26,0325	50,5106	
4	de Barcelona-Catalunya	Montmeló	Spain	41,57	22,6111	
5	l Park	Istanbul	Turkey	40,9517	29,405	
6	de Monaco	Monte-Carlo	Monaco	43,7347	74,2056	
7	Gilles Villeneuve	Montreal	Canada	45,5	-73,5228	
8	de Nevers Magny-Cours	Magny Cours	France	46,8642	31,6361	
9	one Circuit	Silverstone	UK	52,0786	-101,694	
10	heimring	Hockenheim	Germany	49,3278	85,6583	
11	oring	Budapest	Hungary	47,5789	19,2486	
12	a Street Circuit	Valencia	Spain	39,4589	-0,331667	
13	de Spa-Francorchamps	Spa	Belgium	50,4372	59,7139	
14	omo Nazionale di Monza	Monza	Italy	45,6156	92,8111	
15	Bay Street Circuit	Marina Bay	Singapore	12,914	103,864	
16	edway	Oyama	Japan	35,3717	138,927	
17	ai International Circuit	Shanghai	China	31,3389	121,22	
18	omo José Carlos Pace	São Paulo	Brazil	-23,7036	-46,6997	
19	polis Motor Speedway	Indianapolis	USA	39,795	-86,2347	
20	gring	Nürburg	Germany	50,3356	69,475	
21	omo Enzo e Dino Ferrari	Imola	Italy	44,3439	117,167	
22	Circuit	Suzuka	Japan	34,8431	136,541	
23	as Strip Street Circuit	Las Vegas	United States	36,1147	-115,173	

Figura 4.12: Tabella circuits finale

4.4 Caricamento

La terza ed ultima fase è quella del *caricamento*. In questa attività i dati vengono caricati in una struttura multidimensionale o in un data warehouse in modo da condurre le relative analisi. Questo processo include sia il caricamento delle tabelle delle dimensioni, le quali contengono informazioni descrittive, che delle tabelle dei fatti, ad esempio quelle che contengono i dati quantitativi o numerici.

Nella fase di caricamento, vi sono due principali opzioni, il *caricamento "massivo"* dei dati, nel quale vengono utilizzati strumenti specifici del sistema di gestione dei database per inserire grandi quantità di dati in modo efficiente. L'altra scelta è *l'inserimento dei dati come sequenza di righe*, in cui ogni riga viene processata singolarmente. Il metodo di *caricamento "massivo"* è fortemente consigliato, in quanto permette di caricare enormi quantità di dati in breve tempo.

Al contrario, nel secondo metodo, i semplici comandi *SQL* non sono adatti a gestire grandi volumi di dati, poiché la tecnica tradizionale *open-loop-fetch* che inserisce un record per volta, è estremamente lenta e inefficiente.

Inoltre, per quanto concerne la distinzione tra dati nuovi e dati già esistenti, questa problematica è facilmente gestibile, dato che molti *DBMS* forniscono strumenti dichiarativi, come il comando *MERGE* di *Oracle*, che combina operazioni di inserimento e aggiornamento in un'unica istruzione.

4.4.1 Attività di caricamento del dataset

Infine, per trasferire i dati da *Power Query* a *Power BI*, e per poter iniziare a condurre delle analisi sui dati, con la conseguente creazione delle relative dashboard, basta cliccare sull'icona *Chiudi e applica* (Figura 4.13).

In questo modo tutte le operazioni eseguite per la trasformazione dei dati verranno applicate, e i dati verranno caricati definitivamente su *Power BI*.

Figura 4.13: Chiudi e applica

CAPITOLO 5

Progettazione e implementazione delle analisi di base

In questo capitolo illustreremo due dashboard sviluppate con Power BI, descrivendo gli oggetti visivi utilizzati e analizzando i risultati ottenuti.

5.1 Premessa

Attraverso l'utilizzo di *Power BI*, dei suoi oggetti visivi mostrati in Figura 5.1 e delle misure sviluppate in linguaggio *DAX*, sono state create le due dashboard che approfondiremo nelle prossime sezioni.

In aggiunta, per migliorare le analisi eseguite, sono stati importati altri oggetti visivi oltre a quelli di base, ovvero il *Selection slicer* e la *Timeline 2.4.0*.

Figura 5.1: Oggetti visivi di Power BI

5.2 Dashboard "Piloti-Costruttori"

Nella dashboard "Piloti-Costruttori" verranno analizzati aspetti riguardanti le vittorie ottenute dai piloti e dai relativi costruttori, cercando di correlare i risultati per ottenere una visione complessiva delle performance nelle varie stagioni di campionato di Formula 1. In

Figura 5.2 è riportata la dashboard corrispondente, applicando il filtro temporale per l'intero arco di tempo considerato.

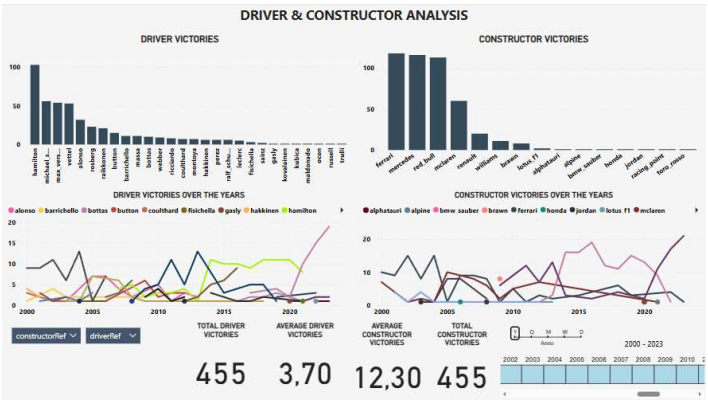

Figura 5.2: Dashboard "Piloti-Costruttori" filtrata dal 2000 al 2023

5.2.1 Oggetti visivi e misure utilizzate

Gli oggetti visivi utilizzati che compongono la dashboard sono i seguenti:

- *Istogramma a colonne raggruppate*: questo oggetto è fondamentale per visualizzare e confrontare il numero di vittorie in tutto o in un determinato arco di tempo; è previsto un istogramma per i piloti e uno per i costruttori.
- *Scheda*: essa è stata utilizzata per visualizzare il numero totale o parziale di vittorie, sia dei piloti che dei costruttori. Grazie alla configurazione dinamica della scheda, è possibile visualizzare alternativamente le vittorie di un singolo pilota, di tutti i piloti, di un singolo costruttore o di tutti i costruttori, a seconda delle selezioni effettuate tramite i filtri temporali.
- *Grafico a linee*: esso è importante per visualizzare e confrontare l'andamento delle vittorie dei piloti e dei costruttori nel tempo.
- *Timeline 2.4.0*: essa è stata utilizzata per filtrare gli anni su cui focalizzarsi;
- *Selection slicer*: questo oggetto è stato utilizzato per filtrare il nome dei costruttori e dei piloti.

Per quanto concerne le misure, esse sono le seguenti:

- *Total driver victories*: essa ha lo scopo di calcolare la somma di tutte le vittorie dei piloti in Formula 1, opportunamente filtrate in base all'arco di tempo.
- *Total constructor victories*: essa è utilizzata per calcolare la somma di tutte le vittorie dei costruttori in Formula 1, opportunamente filtrate in base all'arco di tempo.
- *Average driver victories*: essa è stata utilizzata per calcolare la media teorica di vittorie per ciascun pilota. Essa è stata sviluppata dividendo il numero di gare per il numero di piloti, in base all'arco di tempo selezionato.
- *Average constructor victories*: essa è stata utilizzata per calcolare la media teorica di vittorie per ciascuna scuderia. Essa è stata sviluppata dividendo il numero di gare per il numero di scuderie, in base all'arco di tempo selezionato.

5.2.2 Analisi e risultati

Le analisi precedentemente svolte evidenziano come la storia della Formula 1 sia caratterizzata da diverse fasi, spesso marcate dai cambi di regolamento. Queste modifiche hanno un impatto significativo, determinando periodi di dominio di alcune scuderie rispetto alle altre. Di conseguenza, anche le prestazioni dei piloti risentono di queste evoluzioni, influenzando il loro successo in pista.

Nelle prossime sezioni analizzeremo alcuni periodi chiave della Formula 1.

5.2.2.1 Anni 2000-2004

In Figura 5.3 possiamo notare come, in questo periodo, la scuderia Ferrari abbia dominato, conquistando 57 vittorie su 85 gare disputate, con una media di 11,4 Gran Premi vinti ogni anno, al di sopra della media teorica di vittorie di 6,54. Questa superiorità si riflette anche nelle prestazioni dei suoi piloti, Michael Schumacher e Rubens Barrichello.

Il pilota tedesco ha ottenuto 48 vittorie su 85 gare, superando di gran lunga la media teorica di vittorie di 1,85 successi per stagione. Barrichello, invece, con 9 vittorie, non è riuscito a mantenere la media teorica di vittorie per 0,05. Queste analisi evidenziano come Schumacher abbia prevalso persino sul compagno di scuderia.

Infine, il resto delle vittorie sono state distribuite tra diverse scuderie, in ordine decrescente di successi: McLaren, Williams, Renault e Jordan. Per quanto riguarda i piloti, i principali vincitori, dopo Michael Schumacher, sono stati Coulthard, Häkkinen, Ralf Schumacher, Montoya e Räikkönen, anch'essi in ordine di numero di vittorie decrescente. Inoltre, nessuna di queste scuderie è riuscita a mantenere la media teorica.

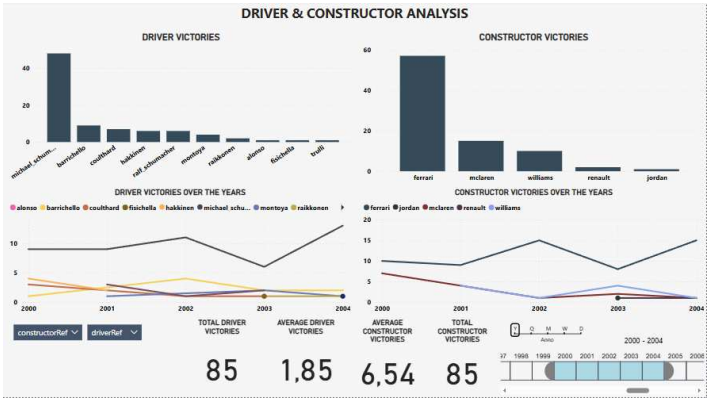

Figura 5.3: Dashboard "Piloti-Costruttori" filtrata dal 2000 al 2004

5.2.2.2 Anni 2005-2009

In Figura 5.4 si evidenzia che tra il 2005 e il 2009 non vi è stato un vero e proprio dominio da parte di una singola scuderia.

Le scuderie Ferrari, McLaren, Renault e Brawn, quest'ultima solo nell'ultimo anno, hanno gareggiato alternandosi nella conquista del campionato. Di conseguenza, essi hanno dato vita ad una competizione equilibrata. La Ferrari, seguita dalla McLaren, sono state le uniche due scuderie sopra la media teorica di vittorie stagionali, del valore di 4,68.

Relativamente ai piloti, coloro che hanno vinto maggiormente sono stati Alonso, Räikkönen, Hamilton, Button e Massa, di cui i primi quattro hanno conquistato il primo posto nel campionato dei piloti. In particolare, spicca Fernando Alonso, unico pilota ad aver superato

la media di 1,89 vittorie. Inoltre, sia Alonso che la scuderia Renault hanno vinto entrambi i rispettivi campionati del mondo per due anni consecutivi.

Per una maggiore chiarezza, nella dashboard sono stati filtrati i piloti e le scuderie con il maggior numero di vittorie.

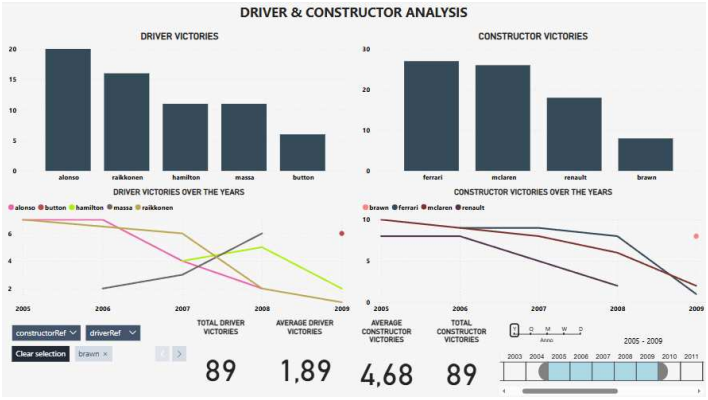

Figura 5.4: Dashboard "Piloti-Costruttori" filtrata dal 2005 al 2009

5.2.2.3 Anni 2010-2013

In Figura 5.5 notiamo come, durante questo periodo, ha dominato la scuderia Red Bull, ottenendo il titolo costruttori per quattro anni consecutivi. La scuderia, infatti, ha collezionato 41 vittorie su 77 gare disputate, superando abbondantemente la media teorica di vittorie stagionali del valore di 5,13. Le altre scuderie ad aver conquistato il maggior numero di vittorie sono state McLaren, Ferrari e Mercedes; esse però, non sono riuscite a raggiungere la soglia minima di vittorie stagionali e il titolo costruttori.

Per quanto concerne i piloti, Sebastian Vettel, il primo della scuderia austriaca, ha vinto quattro volte il campionato piloti, dominando anche sul compagno di squadra Mark Webber. Il quattro volte campione del mondo ha vinto 34 gran premi su 77, superando abbondantemente la media teorica di vittorie stagionali del valore di 1,83.

Gli altri piloti con il maggior numero di vittorie conquistate sono stati Alonso, Hamilton, Button e Webber, di cui, esclusivamente quest'ultimo, non è riuscito a raggiungere la media teorica di vittorie.

Anche in questa dashboard, per una maggiore chiarezza, sono stati filtrati i piloti e le scuderie con il maggior numero di vittorie.

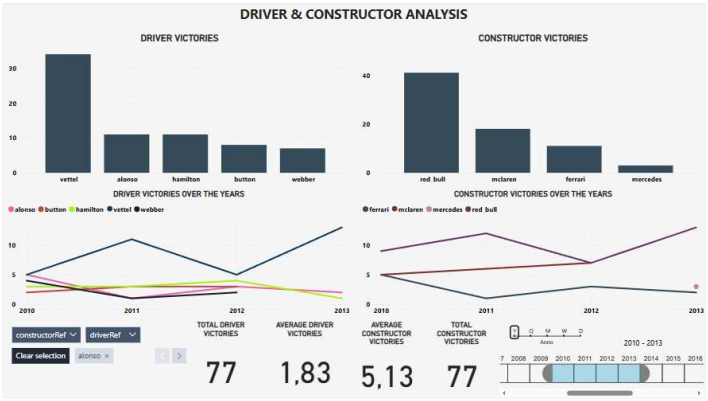

Figura 5.5: Dashboard "Piloti-Costruttori" filtrata dal 2010 al 2013

5.2.2.4 Anni 2014-2021

In Figura 5.6 emerge che, durante l'era "turbo-ibrida" della Formula 1, la Mercedes ha dominato incontrastata, ad eccezione dell'ultimo anno. La scuderia tedesca ha conquistato ben 8 titoli costruttori consecutivi e 111 vittorie su 160 gare disputate, con una media di vittorie per stagione nettamente superiore a quella degli avversari.

Le restanti vittorie sono state ottenute principalmente da Ferrari e Red Bull. Quest'ultima si è rivelata particolarmente competitiva nel 2021, anno in cui Max Verstappen ha interrotto il dominio di Hamilton. La Ferrari, invece, ha avuto stagioni molto altalenanti, con un minimo di zero vittorie e un massimo di sei, al di sotto della media teorica di vittorie stagionali.

Dal punto di vista individuale, Lewis Hamilton ha registrato numeri impressionanti, con 81 vittorie su 160 gare, mantenendo una media stagionale di circa 10 vittorie su 20 gare. Le restanti vittorie sono state distribuite tra cinque piloti, ovvero: Max Verstappen, soprattutto nel 2021; Nico Rosberg nei primi tre anni dell'era "turbo-ibrida" prima di ritirarsi; Valtteri Bottas, soltanto dopo essere arrivato in Mercedes al posto di Rosberg; infine, Daniel Ricciardo e Sebastian Vettel, tutti con numeri inferiori alla media di Hamilton.

In questa dashboard per una maggiore chiarezza, sono stati filtrati i piloti con il maggior numero di vittorie.

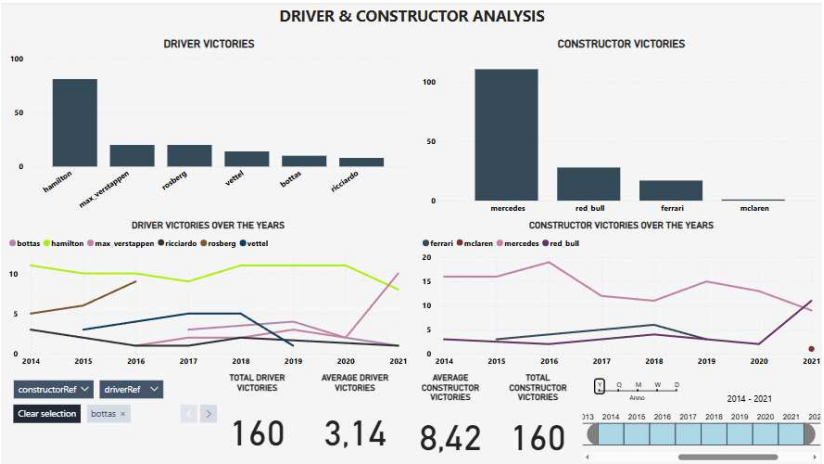

Figura 5.6: Dashboard "Piloti-Costruttori" filtrata dal 2014 al 2021

5.2.2.5 Anni 2022-2023

In Figura 5.7 viene illustrato come, nei primi due anni del nuovo regolamento, la situazione in Formula 1 si sia completamente ribaltata. La Red Bull ha riconquistato il dominio, con Max Verstappen e Sergio Pérez alla guida della monoposto. La scuderia austriaca ha conquistato 38 vittorie su 44 gare disputate, vincendo il campionato costruttori in entrambe le stagioni con diverse gare d'anticipo e superando abbondantemente la media teorica di vittorie stagionali.

Al contrario, la Ferrari ha ottenuto un totale di sole cinque vittorie. La Mercedes, invece, dopo anni di supremazia, è caduta in una profonda crisi, vincendo esclusivamente una gara in due stagioni. Di conseguenza, nessuna delle due scuderie è riuscita a raggiungere la media teorica di vittorie annuali del valore di 4,40.

Per quanto riguarda i piloti, Max Verstappen ha dominato entrambi gli anni stabilendo nuovi record. Ha conquistato 34 vittorie su 44 gare, sorpassando il compagno di squadra Sergio Pérez, che si è fermato a soli cinque successi, appena sopra la media teorica di vittorie del valore di 1,76. Le restanti vittorie, sono state conquistate da George Russel, alla guida

della Mercedes, e da Charles Leclerc e da Carlos Sainz, quest'ultimo con una vittoria in ciascuna delle due stagioni.

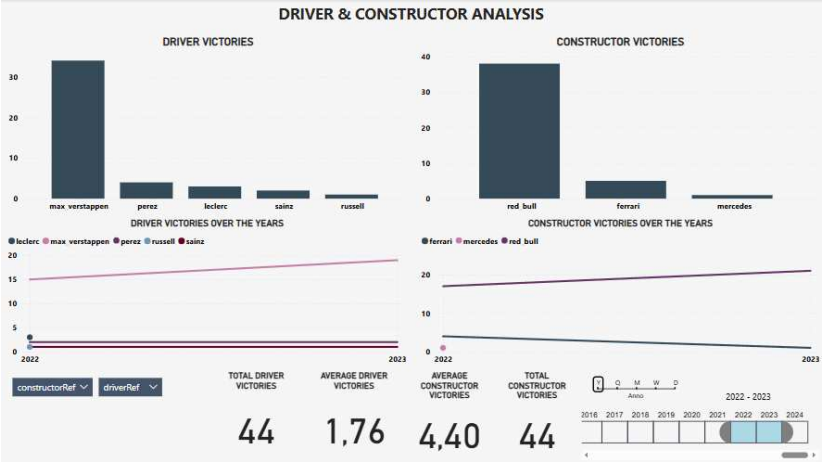

Figura 5.7: Dashboard "Piloti-Costruttori" filtrata dal 2022 al 2023

Possiamo, quindi, concludere affermando che, nella Formula 1, si instaurano dei cicli di dominio tra le scuderie, prevalentemente con i cambi di regolamento, in cui nuove vetture devono essere progettate secondo le nuove direttive. Inoltre, durante questi periodi, il pilota alla guida della vettura appartenente alla scuderia dominante riesce anch'esso a conquistare il titolo piloti e a dominare sugli avversari.

5.3 Dashboard "Confronto piloti"

Lo scopo della dashboard "Confronto piloti" è analizzare e mettere a confronto le prestazioni dei piloti, con un focus specifico sulla stagione 2021 e sui piloti Lewis Hamilton e Max Verstappen. Viene preso in esame questo anno in quanto è stato unico nella storia della Formula 1. I due piloti si sono contesi il campionato in ogni gran premio, dando vita a una delle rivalità più intense di sempre. La loro lotta si è protratta per l'intero campionato, culminando in un finale straordinario, in cui Hamilton e Verstappen si sono presentati all'ultima gara con lo stesso numero di punti in classifica.

A causa delle limitazioni di spazio, il report è stato suddiviso in due dashboard per garantire una visualizzazione ottimale delle informazioni. Nella parte sinistra della dashboard sono riportate le analisi relative al pilota Mercedes, mentre nella parte destra quelle del pilota Red Bull, entrambe illustrate utilizzando i rispettivi colori rappresentativi. Verranno analizzati diversi aspetti, ovvero:

1. *Passo gara*;
2. *Tempo in qualifica*;
3. *Posizione di partenza media*;
4. *Posizione di arrivo media*;
5. *Punteggio*;
6. *Posizione finale*;

5.3.1 Oggetti visivi e misure utilizzate

Gli oggetti visivi utilizzati che compongono la dashboard sono i seguenti:

- *Grafico a linee*: questo oggetto è stato utilizzato per rappresentare l'andamento del passo gara, i tempi di qualifica nel corso del campionato, la posizione finale di ogni gara e i punti conquistati per ogni gara.
- *Scheda*: essa è utilizzata per visualizzare la posizione di partenza media e di arrivo media, durante la stagione di Formula 1.
- *Timeline 2.4.0*: essa è utilizzata per filtrare l'anno considerato.

Le misure utilizzate per condurre le analisi sono le seguenti:

- *Average start position*: questa misura ha lo scopo di calcolare la posizione media in griglia di partenza, per essere poi filtrata per un determinato anno e pilota.
- *Average finish position*: questa misura è stata utilizzata per calcolare la posizione media finale, per essere poi filtrata per un determinato anno e pilota.
- *Race pace*: questa misura calcola il passo gara escludendo i giri in cui viene effettuato il pit-stop, in modo che risulti un passo gara reale. Anch'esso viene filtrato per un determinato anno e per un particolare pilota.

5.3.2 Analisi e risultati

Dalle analisi precedentemente condotte, mostrate nelle Figure 5.8 e 5.9, emerge che, nonostante Max Verstappen partisse mediamente da una posizione migliore rispetto a Lewis Hamilton, terminava le gare in una posizione finale più arretrata rispetto al suo diretto rivale. Tuttavia, è riuscito comunque a guadagnare più punti e a vincere il Campionato Mondiale di Formula 1 nel 2021, ottenendo complessivamente 11 vittorie contro le 9 del pilota Mercedes.

Questo risultato è dovuto al fatto che la media della sua posizione finale è influenzata negativamente da tre risultati particolarmente sfavorevoli, due diciottesimi posti e un ventesimo posto. Al contrario, Hamilton non ha mai concluso una gara in posizioni così arretrate.

Quest'analisi evidenzia come il pilota della Red Bull fosse generalmente più veloce in qualifica, avendo ottenuto una posizione di partenza migliore rispetto al pilota Mercedes in sei occasioni. Invece, entrambi i piloti hanno registrato il giro più veloce della gara in sei occasioni ciascuno. Nonostante ciò, Hamilton si è dimostrato superiore nel passo gara, risultando più veloce in due gare in più rispetto a Verstappen. Questo gli ha permesso di recuperare posizioni nel corso delle gare, sebbene fosse penalizzato da partenze più arretrate, e quindi costretto ad effettuare più sorpassi in pista, perdendo decimi di secondo preziosi per una prestazione senza svantaggi.

L'andamento delle posizioni finali riflette questa tendenza: Verstappen, pur partendo in media davanti al suo rivale, non sempre riusciva a mantenere il vantaggio, mentre Hamilton, grazie a un passo gara mediamente migliore, era in grado di rimontare.

Possiamo quindi concludere che la posizione di partenza ha un'importanza relativa e non assicura la vittoria, poiché il passo gara è una componente fondamentale per raggiungere la vittoria a fine gara.

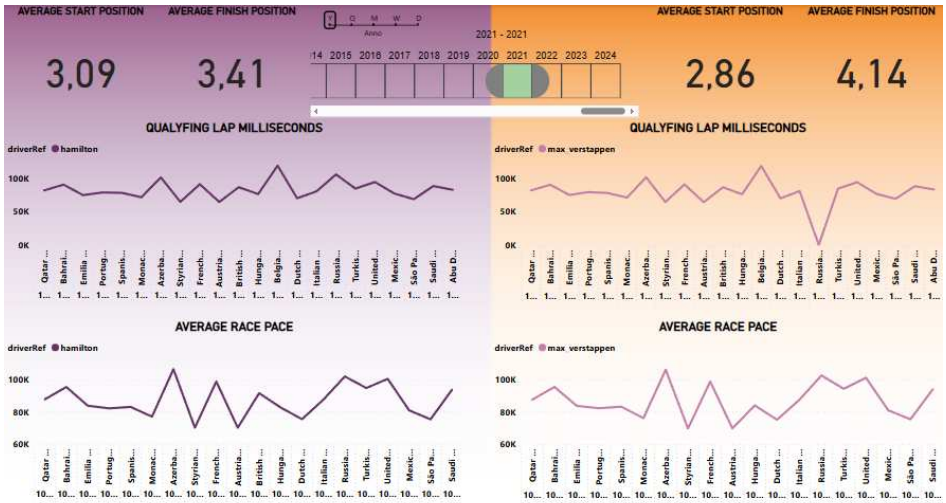

Figura 5.8: Dashboard "Confronto piloti" (prima parte)

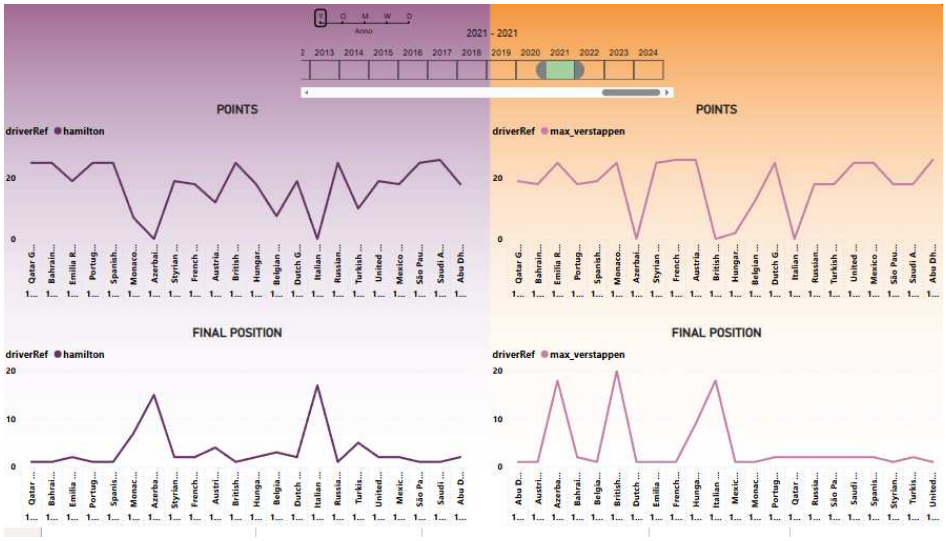

Figura 5.9: Dashboard "Confronto piloti" (seconda parte)e

CAPITOLO 6

Progettazione e implementazione delle analisi avanzate

In questo capitolo illustreremo le dashboard relative all'evoluzione del passo gara, delle posizioni in gara, dei pit-stop e dell'affidabilità; verranno analizzati anche gli oggetti visivi e le misure utilizzate.

6.1 Dashboard "Passo gara"

In questa dashboard verranno analizzati gli aspettivi relativi all'evoluzione del passo gara, della velocità massima raggiunta in media e del tempo medio per percorrere un giro in pista. Tutti questi aspetti saranno poi filtrati per arrivare a dei risultati evidenti. In Figura 6.1 è riportata la dashboard senza aver applicato i filtri disponibili.

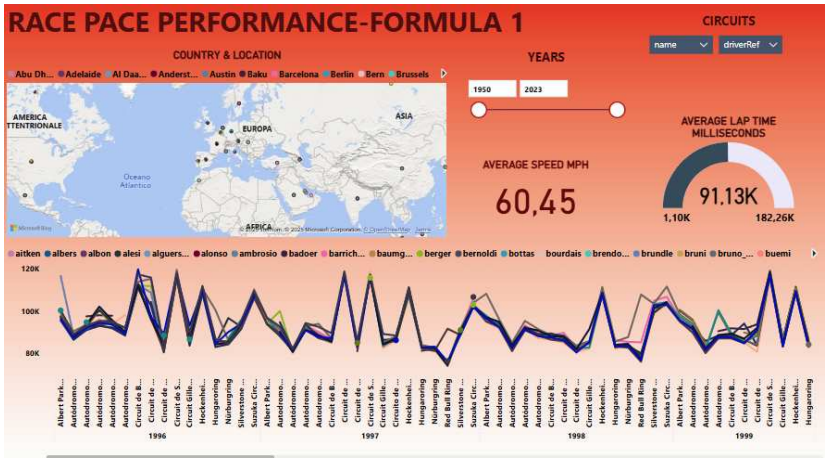

Figura 6.1: Dashboard "Passo gara" filtrata dal 2000 al 2023

6.1.1 Oggetti visivi e misure utilizzate

Gli oggetti visivi utilizzati che compongono la dashboard sono i seguenti:

- *Mappa*: essa ha la funzione di mostrare i diversi circuiti presenti nei vari paesi.
- *Filtro dei dati*: questo oggetto è stato utilizzato per filtrare gli anni, il nome dei circuiti e il nome dei piloti.

- *Scheda*: essa ha lo scopo di mostrare la media della velocità massima raggiunta dai piloti, rappresentata in miglia per ora.
- *Misuratore*: questo dispositivo è stato necessario per mostrare il tempo medio per compiere un giro in pista.
- *Grafico a linee*: questo strumento è stato utilizzato per rappresentare l'andamento del passo gara nei diversi anni e circuiti dei vari piloti.

6.1.2 Analisi e risultati

Per condurre le analisi mostrate in Figura 6.2, è stato scelto Hamilton come pilota, poiché, avendo sempre corso in una scuderia competitiva ed essendo uno dei piloti più longevi, ci permette di ottenere delle medie relative al passo gara molto affidabili. Per quanto riguarda il circuito, è stato selezionato quello del Belgio, Spa-Francorchamps, considerato come uno dei tracciati più completi in tutte le specifiche di questo sport.

Esso è soprannominato "Università della Formula 1", in quanto presenta delle curve lente e delle curve veloci, dei lunghi rettilinei, delle brusche frenate e vari dislivelli di altitudine. Dall'analisi emerge che lo sviluppo costante della Formula 1 non equivale necessariamente a un miglioramento delle prestazioni. Possiamo notare, infatti, che la media del passo gara nell'anno 2007, del valore di 109.681 millisecondi, è stata la più bassa e, di conseguenza, la migliore fino al 2016. Soltanto nel 2017 essa è migliorata di 619 millesimi di secondo, abbassandosi a 109.062 millisecondi.

La media ha continuato a migliorare fino al 2019, diminuendo di altri 605 millisecondi, per poi peggiorare nel 2020, pur restando al di sotto di 8 decimi di secondo rispetto al 2007. Infine, nel 2023 il passo gara è nuovamente peggiorato a 112.710 millisecondi.

Possiamo quindi concludere che, in Formula 1, la presenza di molte variabili e sfide tecniche, che cambiano ogni stagione, non garantisce un miglioramento costante delle prestazioni anno per anno.

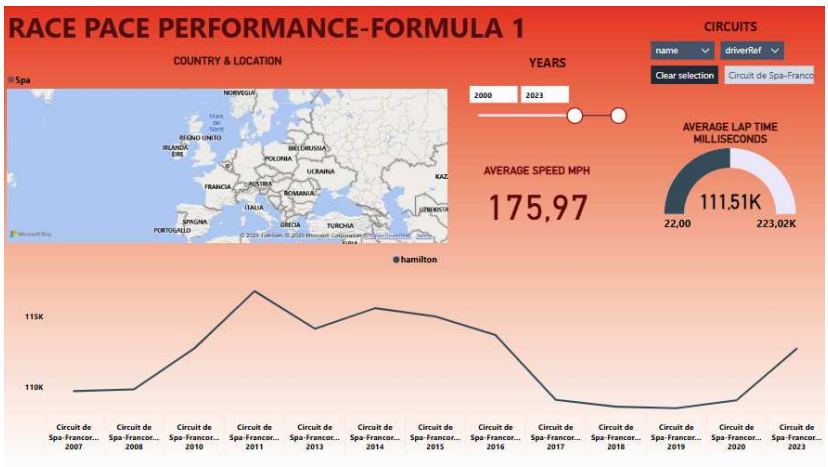

Figura 6.2: Dashboard "Passo gara" filtrata dal 2008 al 2023

6.2 Dashboard "Posizioni"

Lo scopo della dashboard "Posizioni" è analizzare le variazioni delle posizioni in pista durante una singola gara, un'intera stagione o più stagioni, includendo anche il numero di sorpassi effettuati. In Figura 6.3 è mostrata la dashboard filtrata per un determinato pilota, per

una determinata gara e per un determinato anno, in modo tale da visualizzare correttamente la funzione di ogni componente.

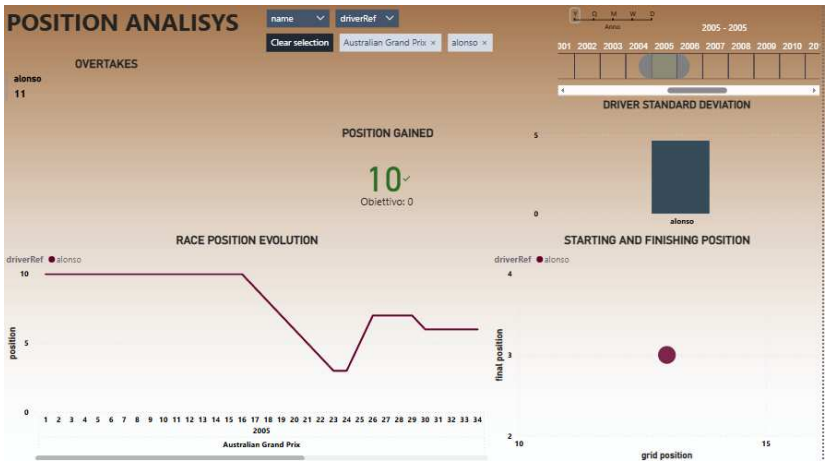

Figura 6.3: Dashboard "Posizioni" filtrata per un determinato pilota

6.2.1 Oggetti visivi e misure utilizzate

Gli oggetti visivi utilizzati che compongono la dashboard sono i seguenti:

- *Scheda con più righe*: questo oggetto è stato utilizzato per visualizzare il numero di sorpassi effettuati dai piloti.
- *Indicatore KPI*: esso ha lo scopo di visualizzare il numero di posizioni guadagnate o perse dal pilota.
- *Istogramma a colonne raggruppate*: questo strumento ha il fine di visualizzare la deviazione standard delle posizioni del pilota.
- *Selection Slicer*: questo oggetto è importante per filtrare il gran premio e il pilota.
- *Timeline 2.4.0*: questo filtro viene utilizzato per selezionare gli anni desiderati.
- *Grafico a linee*: questo oggetto è stato utilizzato per visualizzare l'andamento delle posizioni in gara di uno o più piloti nei vari gran premi dei diversi anni.
- *Grafico a dispersione*: esso ha lo scopo di rappresentare la posizione iniziale e finale di una o più gare evidenziando, in base alla grandezza del cerchio, se il pilota ha perso o guadagnato delle posizioni durante il gran premio.

Per quanto concerne le misure, esse sono le seguenti:

- *Driver Standard Deviation*: questa misura calcola la deviazione standard delle posizioni dei piloti tramite la funzione *STDEV.P* in linguaggio *DAX*, prendendo come argomento i valori della colonna *position* della tabella *lap_times*.
- *Posizione guadagnata*: per sviluppare questa misura, in primo luogo è stata generata una nuova colonna, nella quale veniva effettuata la differenza tra i valori di due colonne esistenti, relative alla posizione iniziale e a quella finale. Successivamente, è stata sviluppata la misura sommando i valori della nuova colonna, per poi essere filtrati in base al pilota al momento della visualizzazione.

- *Sorpassi effettuati*: questa misura scorre i valori della tabella *lap_times*, confrontando la posizione del pilota nel giro corrente con quella del giro precedente. Nel caso in cui il pilota abbia migliorato la sua posizione in pista, il conteggio viene incrementato di 1, restituendo, così, il numero totale di sorpassi effettuati, filtrando il pilota e la gara. Questa misura è mostrata nel Listato 6.1.

```
SorpassiEffettuati_misura =
SUMX( 'lap_times',
    VAR PosizionePrecedente =
        CALCULATE(
            MAX('lap_times'[position]),
            FILTER( 'lap_times',
                    'lap_times'[driverId] =
    EARLIER('lap_times'[driverId]) &&
                     'lap_times'[raceId] = EARLIER('lap_times'[raceId])
    &&
                     'lap_times'[lap] = EARLIER('lap_times'[lap]) - 1
            ) )
    RETURN IF(PosizionePrecedente > 'lap_times'[position], 1, 0) )
```

Listing 6.1: Codice DAX per il calcolo dei sorpassi effettuati

6.2.2 Analisi e risultati

In questa dashboard è interessante analizzare l'evoluzione della Formula 1 nell'ambito dei sorpassi e della variazione delle posizioni durante le gare. Per illustrare questo tema, nelle Figure 6.4 e 6.5 sono riportati i dati relativi ai piloti, la cui scelta è stata determinata in base agli anni della loro carriera in formula 1. Per quanto concerne i sorpassi, è fondamentale notare che piloti come Mika Häkkinen e Damon Hill, che hanno gareggiato negli anni '90, nella loro intera carriera hanno compiuto un numero di sorpassi inferiori a quelli effettuati da Max Verstappen e Charles Leclerc, piloti con carriere più recenti. Altri come Michael Schumacher, Lewis Hamilton e Sebastian Vettel, con carriere più longeve, hanno effettuato un numero di sorpassi maggiore rispetto ai loro rivali nella fase centrale di questi anni.

Possiamo, quindi, affermare che, grazie allo sviluppo costante della Formula 1, il numero dei sorpassi effettuati sia aumentato, garantendo sfide più avvincenti.

Figura 6.4: Grafico dei sorpassi effettuati filtrato per determinati piloti

Anche in questo oggetto visivo riscontriamo quanto detto precedentemente. Il riscontro maggiore lo ritroviamo paragonando il pilota più recente, Charles Leclerc, e il meno recente, Hakkinen. Il pilota monegasco, infatti, ha una variazione della sua posizione in gara maggiore rispetto al pilota finlandese, nonostante quest'ultimo abbia gareggiato per 11 anni in Formula 1, contro le 7 stagioni di Leclerc.

Figura 6.5: Grafico delle variazioni delle posizioni filtrato per determinati piloti

6.3 Dashboard "Pit-Stop"

In Figura 6.6 è rappresentata la dashboard relativa ai pit-stop. Quest'ultimo è un momento cruciale all'interno di una gara in quanto può stravolgere la posizione dei piloti. Il tempo per eseguire un pit-stop comprende: l'entrata nella pit-lane, l'arrivo nel proprio box, il cambio gomme, la ripartenza dal box e l'uscita dalla pit-lane. In questa sezione illustreremo l'evoluzione dei pit-stop attraverso le varie analisi effettuate.

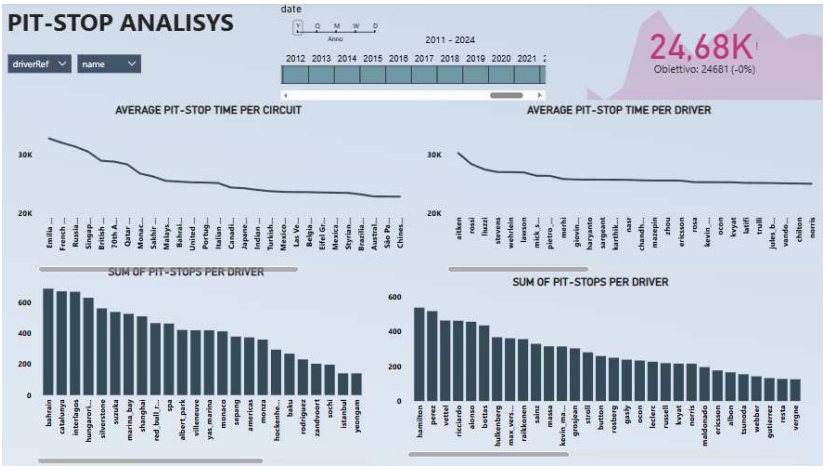

Figura 6.6: Dashboard "Pit-Stop" filtrata dal 2011 ad oggi

6.3.1 Oggetti visivi e misure utilizzate

Gli oggetti visivi utilizzati che compongono la dashboard sono i seguenti:

- *Grafico a linee*: questo oggetto ha lo scopo di rappresentare il tempo medio per effettuare un pit-stop in un determinato circuito e per un determinato pilota, riuscendo a visualizzare le differenze tra diversi circuiti o piloti.

- *Istogramma a colonne raggruppate*: esso è stato utilizzato per rappresentare il conteggio dei pit stop effettuati in un determinato circuito in uno specifico arco di tempo e per visualizzare il conteggio dei pit-stop effettuati da determinati piloti in un determinato arco di tempo, potendo confrontare diversi circuiti o piloti.
- *Indicatore KPI*: questo oggetto è stato utilizzato per visualizzare il tempo medio per effettuare un pit-stop in base agli anni selezionati, con lo scopo di visualizzare lo scostamento in percentuale dal tempo medio generale.
- *Timeline 2.4.0*: questo filtro ha lo scopo di selezionare gli anni desiderati.
- *Selection Slicer*: questo oggetto è stato utilizzato per filtrare il gran premio e il pilota.

Per quanto concerne le misure, esse sono le seguenti:

- *Media Pit-stop*: questa misura è stata utilizzata per calcolare il tempo medio per effettuare un pit-stop. Essa è stata calcolata tramite l'utilizzo della funzione *AVERAGE* in linguaggio *DAX*, selezionando come argomento la colonna *duration* della tabella *pit_stop*.
- *Numero pit-stop*: questa misura è stata utilizzata per calcolare il totale dei pit-stop effettuati, filtrato in seguito tramite la selezione dei circuiti che si vogliono analizzare. La misura è stata calcolata attraverso la funzione *COUNT* in linguaggio *DAX*, selezionando come argomento la colonna *stop* della tabella *pit_stop*.

6.3.2 Analisi e risultati

In Figura 6.7 dalle analisi condotte emerge che il numero di pit-stop effettuati in un determinato circuito non è strettamente correlato al tempo impiegato mediamente per eseguirli.

Sebbene un numero maggiore di pit-stop aumenti la probabilità di inconvenienti, non vi è una relazione diretta tra la quantità di fermate e il tempo medio di esecuzione. Possiamo, quindi, dedurre che il tempo medio per effettuare un pit-stop è influenzato dalla lunghezza della pit-lane. Essa, infatti, varia in ogni circuito e può variare anche nel corso degli anni. Ad esempio, il circuito di Imola, situato in Emilia-Romagna, ha un tempo medio del valore di 32.680 millisecondi, maggiore rispetto a tutti gli altri circuiti, nonostante dal 2011 vi siano stati disputati soltanto 4 gran premi. Relativamente al circuito di Baku situato in Azerbaijan, si registra il tempo mediamente più basso per effettuare il pit-stop, del valore di 21.321 millisecondi.

Per quanto concerne il numero di pit-stop effettuati, troviamo al primo posto il circuito della Catalunya mentre all'ultimo posto il circuito Paul Ricard situato in Francia.

Inoltre, emerge che il circuito della Catalunya primeggia rispetto ai circuiti in cui sono stati disputati lo stesso numero di gran premi, ovvero 14. Ricordiamo il circuito di Yas Marina, di Silverstone, di Monza e di Singapore.

In conclusione, possiamo affermare che, in base al circuto, vi siano differenti strategie adottate per massimizzare il risultato.

Relativamente ai piloti, riscontriamo quanto affermato per i circuiti. Infatti, emerge che il pilota ad aver effettuato il numero maggiore di pit-stop, ovvero 369, dal 2011 al 2024 (in corso), è Sergio Perez. Quest'ultimo si classifica tra i piloti con il tempo medio di pit-stop più basso, del valore di 24.548 millisecondi, appena inferiore rispetto al tempo medio generale del valore di 24.681 millisecondi.

Inoltre, possiamo osservare come il pilota Pascal Wehrlein, nonostante abbia effettuato soltanto 60 pit-stop nel corso dei suoi due anni di carriera in Formula 1, abbia registrato il tempo medio più alto, del valore di 27.792 millisecondi.

Pertanto, possiamo affermare che un elevato numero di pit-stop effettuati non implica obbligatoriamente un tempo mediamente più alto. Difatti, si può notare come i piloti che hanno iniziato a gareggiare negli ultimi due anni non abbiano sempre una media più bassa dei piloti che hanno concluso la loro carriera in Formula 1 diversi anni fa. Ad esempio, Lawson ha una media di 25.529 millisecondi, maggiore rispetto alla media di Kubica, del valore di 25.163 millisecondi.

Da questa informazione potrebbe emergere che non vi sia stato un miglioramento nel corso del tempo; tuttavia bisogna tener conto del fatto che dal 2022 le monoposto di Formula 1, a causa del cambio del regolamento tecnico, montano degli pneumatici più grandi di quelli precedenti.

Di conseguenza, essendo aumentato anche il peso, il montaggio nel corso del pit-stop risulta più difficile per i meccanici.

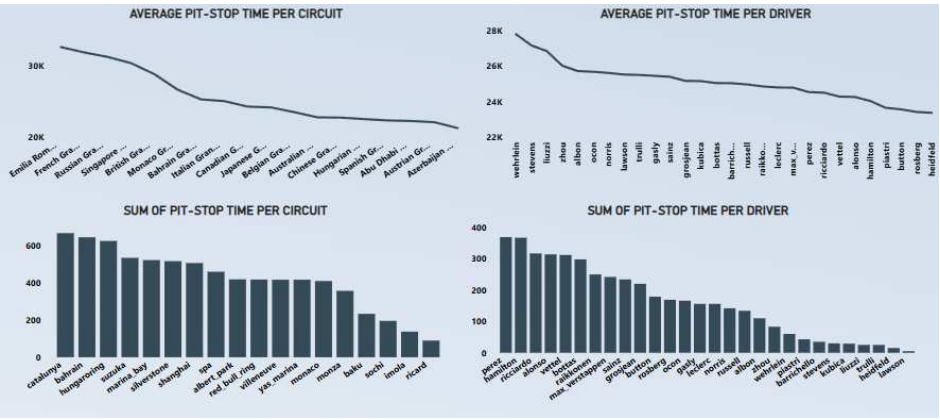

Figura 6.7: Dashboard "Pit-Stop" filtrata dal 2011 fino ad oggi, per i principali circuiti i principali piloti

6.4 Dashboard "Affidabilità"

La dashboard *Affidabilità* ha lo scopo di evidenziare l'evoluzione di un aspetto di questo sport che spesso passa in secondo piano ma che, negli ultimi anni, ha acquisito sempre più importanza, ovvero la sicurezza. L'obiettivo, infatti, è di migliorarla per l'incolumità dei piloti, per assicurare che le monoposto riescano a concludere quante più gare possibili e, infine, per ridurre i costi di riparazione o di cambio delle componenti dell'auto.

6.4.1 Oggetti visivi e misure utilizzate

Gli oggetti visivi utilizzati che compongono la dashboard sono i seguenti:

- *Scheda*: essa è stata utilizzata per visualizzare il numero totale di problemi tecnici relativamente agli anni selezionati.
- *Grafico ad aree*: questo oggetto ha lo scopo di visualizzare l'andamento della somma dei problemi tecnici per ogni anno, in base a quelli selezionati.
- *Istogramma a colonne raggruppate*: questo strumento è stato utilizzato per rappresentare il numero totale di problemi tecnici relativamente al circuito e al costruttore suddivisi per cause.

- *Timeline 2.4.0*: questo filtro è stato utilizzato per selezionare gli anni desiderati.
- *Selection Slicer*: questo oggetto è stato utilizzato per filtrare il gran premio, il costruttore e il tipo di problema tecnico.
- *Mappa colorata*: questa mappa è stata utilizzata per mostrare i paesi in cui accade maggiormente un determinato problema tecnico; per evidenziare questo aspetto, le nazioni vengono colorate in base alla legenda presente negli istogrammi, che illustra il numero e il tipo di problemi tecnici. Sono previsti un istogramma per i costruttori e uno per i circuiti.

6.4.2 Analisi e risultati

Dalle analisi condotte sulla dashboard mostrata in Figura 6.8 emergono diverse informazioni. Dal grafico ad aree risulta che, nel corso degli anni, l'affidabilità in Formula 1 è migliorata notevolmente, nonostante qualche piccolo rialzo di problemi tecnici negli anni intermedi del periodo considerato. Possiamo, quindi, affermare che i problemi tecnici dal 2000 al 2023 sono diminuiti in maniera considerevole; ciò implica delle monoposto più affidabili e una maggiore sicurezza per i piloti.

Dall'istogramma relativo ai circuiti emerge che il problema tecnico relativo al motore è quello verificatosi più spesso in Formula 1 e nella maggior parte dei circuiti. Di conseguenza, si evince che il cuore della monoposto è la parte maggiormente soggetta a guasti, in quanto si cerca di raggiungere prestazioni ottimali e costanti.

Immediatamente dopo troviamo il problema tecnico della rottura del cambio, strettamente correlato al motore; a seguire altri guasti spesso ricorrenti sono quelli relativi alla batteria e ai freni.

Possiamo, inoltre, affermare che il circuito Villenueve in Canada sia quello in cui si è verificato il numero maggiore di problemi tecnici; a seguire il circuito dell'Australia, della Spagna, di Monza e della Malesia.

Questi ultimi sono i circuiti in cui sono state disputate un numero maggiore di gare, oltre al fatto che le loro caratteristiche mettono a dura prova la monoposto. Ad esempio, il circuito di Monza è quello nel quale viene tenuta la velocità media più alta; ciò implica che il motore viene portato all'estremo delle proprie condizioni, e per tale ragione il suo guasto è il più ricorrente.

Il circuito cittadino di Monaco presenta numerose curve pericolose; di conseguenza, a causa dei continui cambi marcia, la rottura del cambio è il guasto più ricorrente.

Sempre nei circuiti cittadini, come, ad esempio, quello di Monaco, del Canada e di Singapore, è di rilievo anche il guasto all'ala frontale della monoposto: a causa della presenza di mura ai lati della pista e di una carreggiata più stretta, la collisione delle monoposto si verifica frequentemente.

Per quanto concerne i costruttori, parallelamente ai circuiti, il problema tecnico più rilevante è la rottura del motore.

Dal grafico emerge che la McLaren è la scuderia che dal 2000 in poi ha registrato in totale il maggior numero di problemi tecnici più significativi, seguita dalla Williams e dalla Red Bull. Quest'ultima si trova al terzo posto nonostante abbia iniziato a gareggiare in Formula 1 dal 2005.

La Ferrari, invece, risulta essere una delle scuderie più affidabili, in quanto ha registrato circa la metà dei guasti della McLaren, nonostante entrambe abbiano gareggiato in tutte le stagioni dal 2000.

Anche la Mercedes risulta essere molto affidabile, in quanto, in 14 anni, ha registrato un numero molto basso rispetto alle scuderie rivali.

La Renault, invece, non risulta essere una delle più affidabili, in quanto si posiziona al quarto posto, avendo gareggiato in Formula 1 un numero ridotto di anni rispetto alle rivali.

Le altre scuderie che registrano un numero basso di problemi tecnici non obbligatoriamente risultano essere affidabili, in quanto hanno gareggiato in Formula 1 per pochi anni.

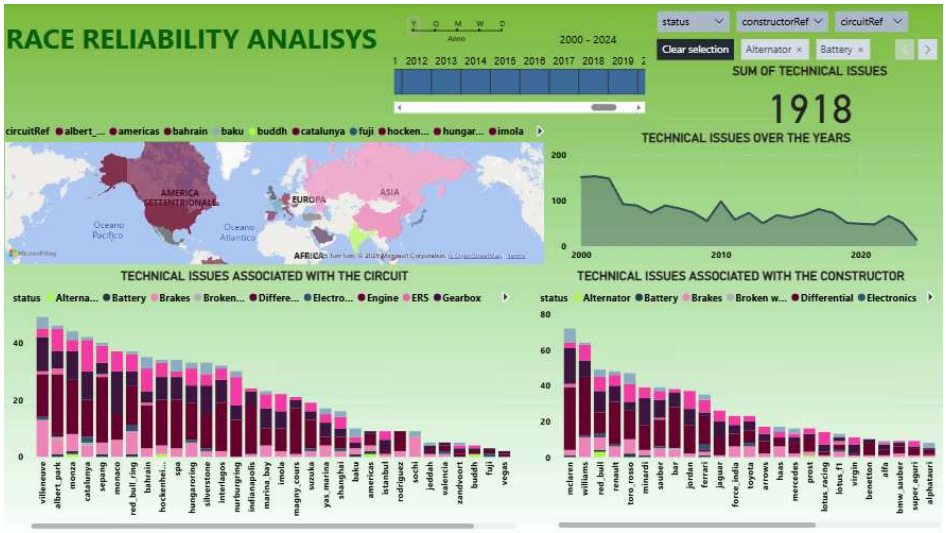

Figura 6.8: Dashboard "Affidabilità" filtrata dal 2000 fino al 2024 in corso e strutturata rispetto ai problemi tecnici

Conclusione e sviluppi futuri

La presente tesi ha analizzato gli aspetti fondamentali delle gare di Formula 1 in determinati archi di tempo, con lo scopo di ricavare delle informazioni utili e di riuscire a rappresentare in modo intuitivo e semplice enormi quantità di dati sullo sport più tecnico e complesso al mondo.

Le prime attività svolte sul relativo dataset sono state quelle di ETL. In una prima fase, ci si è concentrati sull'estrazione di questi dati, che si è svolta effettuando il download del dataset di Formula 1 sulla piattaforma Kaggle. In seguito, analizzando il dataset, sono stati modificati alcuni campi di determinate colonne, in modo da eseguire delle analisi più accurate. Ciò è stato sviluppato tramite l'ausilio del linguaggio Python e degli strumenti di trasformazione dei dati messi a disposizione dalla piattaforma di Business Intelligence Power BI. L'ultima attività svolta è stata il caricamento di questo dataset all'interno della piattaforma.

Per quanto concerne la attività di analisi dei dati, sono state sviluppate varie dashboard, ciascuna delle quali per rappresentare un determinato aspetto della Formula 1. La prima dashboard progettata è stata quella relativa al confronto delle vittorie tra piloti e costruttori, cercando di estrapolare eventuali tendenze relativamente ad esse.

In seguito, è stato elaborato un ulteriore report relativamente al confronto di prestazione tra due determinati piloti. Dopodiché, sono state sviluppate dashboard avanzate per approfondire gli aspetti tecnici di questo sport, ovvero l'evoluzione del passo di gara con la possibilità di filtrare i piloti; i circuiti e vari archi di tempo; l'andamento delle posizioni e dei sorpassi durante una gara, una stagione o più stagioni, con lo scopo di comprendere l'evoluzione della competitività e le aspettative nel corso degli anni, relativamente ai diversi circuiti e piloti.

Successivamente, si è affrontato il tema dei pit-stop, analizzando l'andamento dei tempi e la frequenza con cui vengono effettuati, anche in questo caso relativamente ai circuiti e ai piloti nel corso degli anni. Infine, l'ultima dashboard progettata è stata quella dedicata all'affidabilità, con lo scopo di analizzare la frequenza e le cause dei guasti tecnici nel corso degli anni, relativamente ai circuiti e ai construttori.

Un possibile sviluppo futuro di questo progetto potrebbe riguardare l'estensione di questo dataset, incrementando le informazioni disponibili quali i tipi di mescole utilizzati nelle ruote, le condizioni meteorologiche ed eventuali penalità ai danni dei piloti e costruttori, con lo scopo di approfondire ulteriormente le strategie e le cause dei risultati.

Bibliografia

ALBRECHT, A. e NAUMANN, F. (2008), «Managing ETL Processes.», *NTII*, vol. 8 (2008), p. 12–15.

ANDERSEN, O., THOMSEN, C. e TORP, K. (2018), «SimpleETL: ETL processing by simple specifications», in «Proceedings of the 20th International Workshop on Design, Optimization, Languages and Analytical Processing of Big Data co-located with 10th EDBT/ICDT Joint Conference», CEUR Workshop Proceedings.

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

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

BANSAL, A. (2023), «POWER BI SEMANTIC MODELS TO ENHANCE DATA ANALYTICS AND DECISION-MAKING.», *International Journal of Management (IJM)*, vol. 14 (5), p. 136–142, URL `https://mylib.in/index.php/IJM/article/view/IJM_14_05_012`.

BANY MOHAMMAD, A., AL-OKAILY, M., AL-MAJALI, M. e MASA'DEH, R. (2022), «Business intelligence and analytics (BIA) usage in the banking industry sector: an application of the TOE framework», *Journal of Open Innovation: Technology, Market, and Complexity*, vol. 8 (4), p. 189.

BATISTIČ, S. e VAN DER LAKEN, P. (2019), «History, evolution and future of big data and analytics: a bibliometric analysis of its relationship to performance in organizations», *British Journal of Management*, vol. 30 (2), p. 229–251.

BIAGINI, G. e GABBI, E. (2024), *L'educazione attraverso i dati: Dall'alfabetizzazione critica all'educational data mining*, Pensa MultiMedia.

BLUMBERG, R. e ATRE, S. (2003), «The problem with unstructured data», *Dm Review*, vol. 13 (42-49), p. 62.

CHEN, Y., WANG, W., LIU, Z. e LIN, X. (2009), «Keyword search on structured and semi-structured data», in «Proceedings of the 2009 ACM SIGMOD International Conference on Management of data», p. 1005–1010.

CLARK, D. (2017), *Beginning Power BI: A Practical Guide to Self-Service Data Analytics with Excel and Power BI Desktop*, Apress, second ed.

COOPER, A. (2012), «A brief history of analytics», *JISC Cetis Analytics Series*, vol. 1 (9), p. 1–21.

DAS, T. K. e KUMAR, P. M. (2013), «Big data analytics: A framework for unstructured data analysis», *International journal of engineering and technology*, vol. 5 (1), p. 153–156.

DIOUF, P. S., BOLY, A. e NDIAYE, S. (2018), «Variety of data in the ETL processes in the cloud: State of the art», in «2018 IEEE International Conference on Innovative Research and Development (ICIRD)», p. 1–5.

DOMINGUEZ, X., PRADO, A., ARBOLEYA, P. e TERZIJA, V. (2023), «Evolution of knowledge mining from data in power systems: The Big Data Analytics breakthrough», *Electric Power Systems Research*, vol. 218, p. 109 193.

EBERENDU, A. C. e OTHERS (2016), «Unstructured Data: an overview of the data of Big Data», *International Journal of Computer Trends and Technology*, vol. 38 (1), p. 46–50.

ECKERSON, W. W. (2007), «Predictive analytics», *Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report*, vol. 1, p. 1–36.

EL-SAPPAGH, S. H. A., HENDAWI, A. M. A. e EL BASTAWISSY, A. H. (2011), «A proposed model for data warehouse ETL processes», *Journal of King Saud University-Computer and Information Sciences*, vol. 23 (2), p. 91–104.

EMANI, C. K., CULLOT, N. e NICOLLE, C. (2015), «Understandable big data: a survey», *Computer science review*, vol. 17, p. 70–81.

ENTWISTLE, N. (2003), «Concepts and conceptual frameworks underpinning the ETL project», *Occasional report*, vol. 3, p. 3–4.

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

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*, Microsoft Press.

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

FRAZZETTO, D., NIELSEN, T. D., PEDERSEN, T. B. e ŠIKŠNYS, L. (2019), «Prescriptive analytics: a survey of emerging trends and technologies», *The VLDB Journal*, vol. 28, p. 575–595.

GALETSI, P., KATSALIAKI, K. e KUMAR, S. (2020), «Big data analytics in health sector: Theoretical framework, techniques and prospects», *International Journal of Information Management*, vol. 50, p. 206–216.

GANDOMI, A. e HAIDER, M. (2015), «Big data concepts, methods, and analytics», *International Journal of Information Management*, vol. 35 (2), p. 137–144, URL `https://www.sciencedirect.com/science/article/pii/S0268401214001066`.

JUN, T., KAI, C., YU, F. e GANG, T. (2009), «The Research & Application of ETL Tool in Business Intelligence Project», in «2009 International Forum on Information Technology and Applications», vol. 2, p. 620–623.

KAUR, H. e PHUTELA, A. (2018), «Commentary upon descriptive data analytics», in «2018 2nd International Conference on Inventive Systems and Control (ICISC)», p. 678–683.

KIMBALL, R. e CASERTA, J. (2004), *The Data Warehouse ETL Toolkit: Pratical Techniques for Extracting, Cleaning, Conforming and Delivering Data*, Wiley Publishing, Inc.

KRISHNAN, V. (2017), «Research data analysis with power bi», .

KUMAR, V. e GARG, M. (2018), «Predictive analytics: a review of trends and techniques», *International Journal of Computer Applications*, vol. 182 (1), p. 31–37.

LEPENIOTI, K., BOUSDEKIS, A., APOSTOLOU, D. e MENTZAS, G. (2020), «Prescriptive analytics: Literature review and research challenges», *International Journal of Information Management*, vol. 50, p. 57–70.

MIHIRANGA, N. (2022), *Power BI Data Modeling: Build Interactive Visualizations, Learn DAX, Power Query, and Develop BI Models*, BPB Publications.

PICCIANO, A. G. (2012), «The evolution of big data and learning analytics in American higher education.», *Journal of asynchronous learning networks*, vol. 16 (3), p. 9–20.

REZZANI, A. (2017), *Big Data Analytics: Il manuale del data scientist*, Apogeo Education.

SAPSFORD, R. e JUPP, V. (2006), *Data Collection and Analysis*, SAGE Publications Ltd, second ed.

SHARMA, A. K., SHARMA, D. M., PUROHIT, N., ROUT, S. K. e SHARMA, S. A. (2022), «Analytics techniques: descriptive analytics, predictive analytics, and prescriptive analytics», *Decision intelligence analytics and the implementation of strategic business management*, p. 1–14.

SIMITSIS, A., SKIADOPOULOS, S. e VASSILIADIS, P. (2023), «The History, Present, and Future of ETL Technology», in «DOLAP», p. 3–12.

SIMITSIS, A. e VASSILIADIS, P. (2003), «A Methodology for the Conceptual Modeling of ETL Processes.», in «CAiSE workshops», vol. 75.

VASSILIADIS, P. e SIMITSIS, A. (2009), «Extraction, Transformation, and Loading.», *Encyclopedia of Database Systems*, vol. 10, p. 14.

WEBB, C. e OTHERS (2014), *Power query for power BI and Excel*, Apress.

WOLNIAK, R. e GREBSKI, W. (2023), «The concept of diagnostic analytics», *Silesian University of Technology Scientific Papers. Organization and Management Series*, vol. 175, p. 650–669.

Sitografia

- ADHOX – www.adhox.it
- Aeonvis – www.aeonvis.com
- Agenda Digitale – www.agendadigitale.eu
- Alibaba Cloud – www.alibabacloud.com
- AWS – aws.amazon.com
- BMC – www.bmc.com
- Brain on strategy – www.brainonstrategy.com
- Cornell Univrsity – www.cornell.edu
- Coupler – www.blog.coupler.io
- Coursera – www.coursera.org
- Data Flair – www.data-flair.training
- DataMasters – www.datamasters.it
- Datamaze – www.datamaze.it
- DataDeep – datadeep.it
- Data Skills – www.dataskills.it
- Dataversity – www.dataversity.net
- DEV4SIDE – www.dev4side.com
- dgroove – www.dgroove.it
- Digital Coach – www.digital-coach.com
- Ecoh Media – tableau.ecohmedia.com
- Facile BI – www.facilebi.it
- GEEKS ACADEMY – www.geeksacademy.it

- Google Cloud – `cloud.google.com`
- Google Search Central – `developers.google.com`
- Harvard Business Review – `www.hbr.org`
- Harvard Business School Online – `online.hbs.edu`
- IBM – `www.ibm.com`
- IDEASCALE – `www.ideascale.com`
- igmGuru – `www.igmguru.com`
- Informa TechTarget – `www.techtarget.com`
- Innowise – `www.innowise.com`
- Inside – `www.inside.agency`
- IntelliPaat – `www.intellipaat.com`
- InterSystems – `www.intersystems.com`
- KINEXON SPORTS – `kinexon-sports.com`
- knowi – `www.knowi.com`
- Kuma Vision – `www.kumavision.com`
- Marketing Freaks – `www.themarketingfreaks.com`
- Medium – `www.medium.com`
- Microsoft – `www.microsoft.com`
- Microsoft Learn – `www.learn.microsoft.com`
- NEXTRE – `www.nextre.it`
- Noble desktop – `www.nobledesktop.com`
- Opta – `www.opta.it`
- ORACLE – `www.oracle.com`
- ORBYTA technologies – `technologies.orbyta.it`
- Osservatori Digital Innovation – `blog.osservatori.net`
- Partitaiva – `www.partitaiva.it`
- phData – `www.phdata.io`
- phoenixNAP – `www.phoenixnap.it`
- PLOS – `journals.plos.org`
- PREDIK: Data-Driven – `www.predikdata.com`
- PwC – `www.pwc.nl`

- ScienceSoft – `www.scnsoft.com`
- SELDA – `www.selda.net`
- Slideshare – `www.slideshare.net`
- SMART EDGE – `www.studyblog.smart-edge.in`
- Splunk – `www.splunk.com`
- Spotfire – `www.spotfire.com`
- Talend: A Qlik Company –`www.talend.com`
- The Information Lab – `www.theinformationlab.it`
- TheKnowledgeAcademy – `www.theknowledgeacademy.com`
- Wikipedia – `www.wikipedia.org`
- ZeroUno – `www.zerounoweb.it`

Ringraziamenti

Sono molte le persone che vorrei ringraziare.

Innanzitutto ringrazio i miei genitori e mia sorella Giulia, che mi hanno sempre supportato e sostenuto in questi anni, e che hanno creduto in me anche nei momenti più duri.

Vorrei ringraziare la mia fidanzata Anna Giulia, per essermi stata vicina ogni giorno, specialmente in quelli più difficili.

Ringrazio la mia comunità neocatecumenale della parrocchia di Sant'Eufemia per avermi supportato con la preghiera e con la comunione fraterna, ed insieme a loro i catechisti, i quali hanno sempre avuto una parola per me.

Vorrei ringraziere tutti miei amici Blem, Bomber di razza, Cordons, Piccio, Chiara, Morgan, Fraffo, Chiara Ciaff, e i miei coinquilini Puttu, Fabio e Marco, che mi hanno accompagnato in questo percorso.

Vorrei ringraziare anche tutti i miei familiari che sono in cielo e hanno intercesso per me.

Vorrei ringraziare il Prof. Ursino, che mi ha accompagnato nel lavoro di tesi seguendomi costantemente con estrema disponibilità, e l'Ing. Michele Marcehtti, che mi ha aiutato nella parte implementativa del lavoro.

Infine, il ringraziamento più grande va Dio, che mi ha donato tutto quello scritto sopra, per aver fatto con me questa storia nella quale mai mi ha abbandonato e che ha permesso questo traguardo nella mia vita sostenendomi sempre.