

# POLITECNICO DI TORINO

Corso di Laurea Magistrale  
in Ingegneria Gestionale

Tesi di Laurea Magistrale

Progettazione e sviluppo di un'architettura di Business Intelligence per  
l'analisi di Open Data con cruscotti dinamici



Relatore  
Prof.ssa Tania Cerquitelli

Candidato  
Daniele Ercoli

Correlatore  
Dott.ssa Evelina Di Corso

Anno Accademico 2018/2019



# Indice

<b>1. Introduzione .....</b>	<b>1</b>
<b>2. Business Intelligence .....</b>	<b>3</b>
2.1 Cosa si intende per Business Intelligence .....	3
2.2 Gli Open Data.....	16
2.3 Il caso di studio.....	23
<b>3. L'azienda ospitante .....</b>	<b>25</b>
3.1 Presentazione dell'azienda .....	25
3.2 Il prodotto di Business Analytics utilizzato.....	27
<b>4. Architettura analisi e risultati .....</b>	<b>41</b>
4.1 Descrizione di dettaglio del Dataset utilizzato.....	41
4.2 Caricamento dei dati sul Database.....	46
4.3 Tecniche di preparazione dei dati.....	63
4.4 Creazione Data Source e Dataset su KNOWAGE.....	74
4.5 Descrizione dei cruscotti (cockpit) realizzati.....	79
<b>5. Conclusioni .....</b>	<b>109</b>
<b>Riferimenti bibliografici e sitografia.....</b>	<b>110</b>



# 1. Introduzione

Gli *Open Data*, in italiano dati aperti, rappresentano grandi quantità di dati, non sempre facilmente accessibili ed utilizzabili, prodotti e diffusi da organizzazioni pubbliche e private. I dati *open*, strutturati sotto forma di *Database*, possono essere relativi a diverse tematiche come ad esempio inquinamento, ambiente, educazione, trasporti e disponibili in vari formati quali csv, xls, json e html.

Il tema d'interesse rappresenta la loro possibile utilizzazione per fini informativi e di analisi, quindi l'estrazione di conoscenza utile per la creazione di valore e per un impatto positivo su cittadini e imprese; a questo proposito, i dati aperti possono essere scaricati da vari portali governativi presenti sul web e utilizzati sia da enti pubblici che privati per scopi anche commerciali.

Tuttavia, a partire dalla selezione di un file di dati significativo in formato csv, lo scopo del lavoro consiste nel mostrare l'estrazione di conoscenza da una grande quantità di dati mediante la costruzione di cruscotti (*dashboard*) dinamici sul prodotto di *Business Analytics* aziendale; questo consente, inoltre, di operare analisi su più livelli e rispondere a domande di varia natura. Nel corso della trattazione vengono discussi ed illustrati tutti gli *step* intermedi necessari per il caricamento dei dati di partenza all'interno del prodotto aziendale con la conseguente creazione dei cruscotti; a questo proposito, mediante l'utilizzo di un DBMS (*Database Management System*), viene mostrata la progettazione di un'architettura di *Business Intelligence*, la memorizzazione e la gestione dei dati all'interno di un *Database* mediante il linguaggio di interrogazione SQL (*Structured Query Language*) e la creazione di una correlazione tra un prodotto di *Business Analytics* e un *Database*.

La descrizione del lavoro è articolata in cinque capitoli, organizzati come segue:

Nel **secondo capitolo** viene approfondito inizialmente tutto ciò che riguarda la *Business Intelligence*, l'ambito di sviluppo del presente lavoro, insieme ad una descrizione generale delle basi di dati e del linguaggio SQL corredate da opportuni esempi; subito dopo viene affrontato il discorso sugli *Open Data*, cosa sono ed i possibili vantaggi e rischi derivanti dalla loro utilizzazione.

Infine, viene presentato e argomentato il caso di studio con il file di dati scelto per l'analisi.

Il **terzo capitolo** offre una panoramica sull'azienda ospitante e sul settore aziendale di riferimento; qui, viene anche presentato il prodotto di *Business Analytics* utilizzato per l'analisi e vengono descritte, nel dettaglio, le funzionalità utilizzate.

Nel **quarto capitolo** viene offerta, inizialmente, una descrizione dettagliata del file di dati del caso di studio; subito dopo vengono illustrate e spiegate tutte le operazioni necessarie per il caricamento dei dati sul *Database*, quindi le tecniche di preparazione dei dati per una corretta gestione delle informazioni; viene poi mostrato il caricamento del *dataset* sul prodotto aziendale grazie alla creazione di una correlazione tra quest'ultimo e il *Database*, infine vengono descritti i cruscotti realizzati e le analisi effettuate illustrando tutti gli elementi definiti (grafici, tabelle, elementi html ecc.) arredati con opportuni commenti.

Il **quinto** ed ultimo **capitolo**, le conclusioni, vengono armonizzate con l'introduzione richiamando la metodologia seguita durante lo sviluppo del lavoro e mettendo in risalto i risultati raggiunti preposti all'inizio del lavoro.

## 2. Business Intelligence

### 2.1 Cosa si intende per Business Intelligence

Il termine *Business Intelligence* (BI) è stato coniato, per la prima volta, nel 1958 da Hans Peter Luhn, inventore e ricercatore tedesco, durante la sua esperienza lavorativa all'IBM.

Fin dall'origine questo termine comprende sia i sistemi di raccolta dei dati tradizionali volti ad analizzare il passato o il presente e a capirne i fenomeni, le cause dei problemi o i fattori delle performance ottenute, sia i sistemi volti a simulare scenari futuri.

Questi sistemi sono il risultato del mix tra differenti *software tool* e *software application* rivolti al *performance management*, all'ottimizzazione di decisioni operative e alle stime future.

In generale, la BI costituisce una disciplina di supporto alle decisioni strategiche aziendali riguardanti, ad esempio:

- L'evoluzione della domanda di mercato;
- L'individuazione di aree critiche;
- La definizione di strategie di successo;
- La minimizzazione dei costi e la massimizzazione dei profitti.

L'obiettivo principale consiste nella raccolta, pulizia e trasformazione di dati in informazioni e conoscenza significativi a diversi livelli di dettaglio e analisi.

Pertanto, disponendo di un'adeguata infrastruttura *hardware* e *software* di supporto possono essere oggetto di BI differenti ambiti applicativi di interesse tra cui: le industrie manifatturiere (gestione ordini e spedizioni), la distribuzione (gestione magazzino), i servizi finanziari (analisi acquisti), l'ambito assicurativo (analisi richieste di indennizzo e truffe), le telecomunicazioni (analisi chiamate e riconoscimento frodi) e i servizi pubblici (analisi dei risultati e delle *performance*).

Nello specifico, l'espressione *Business Intelligence* può far riferimento a:

- 1) L'insieme dei processi aziendali per la raccolta e l'analisi di dati e informazioni;
- 2) La tecnologia *software* utilizzata per realizzare questi processi;
- 3) Le informazioni e la conoscenza che si ottengono come *output* di tali processi.

Dal punto di vista dei processi aziendali, le organizzazioni si impegnano nel raccogliere e analizzare dati con l'obiettivo di estrapolare conoscenza, operare valutazioni e stime del mercato in cui competono e del contesto aziendale proprio.

Il fine ultimo è infatti quello di utilizzare la conoscenza, le informazioni e le valutazioni generate per incrementare il proprio vantaggio competitivo rispetto ai concorrenti così come per supportare le decisioni di chi è capo di ruoli direzionali.

Quindi la *Business Intelligence* consente di rafforzare e, allo stesso tempo, automatizzare alcuni processi aziendali come quelli di misurazione, controllo e analisi delle *performance* raggiunte mediante cruscotti dinamici, sistemi di *reporting* e *dashboard*.

Lo stesso vale per i processi che mirano a definire decisioni strategiche aziendali in condizioni di indeterminatezza supportati da sistemi di predizione e previsione.

Da un punto di vista tecnologico, la BI consente di trasformare i dati e le informazioni in conoscenza mediante l'utilizzo di strumenti *software*; la nascita di queste applicazioni si è resa necessaria nel momento in cui molte organizzazioni non riuscivano a sfruttare e ad utilizzare le grandi moli di dati a disposizione per scopi pratici e per assumere le giuste decisioni.

Tuttavia, mediante l'interazione dell'utente, un'applicazione *software* di *Business Intelligence* accede a grandi quantità di dati presenti generalmente su *database* e, attraverso l'esecuzione di *query* in linguaggio SQL, permette di recuperare informazioni nascoste; oppure, manipolando i dati, fornisce come *output* cruscotti dinamici o report contenenti grafici, tabelle, indicatori, e statistiche descrittive.

Questi grafici, tabelle e indicatori vengono impostati e profilati dall'utente ed il contenuto viene automaticamente aggiornato in seguito a ciascuna variazione dei dati presenti nel *database*.

Solitamente, per motivi tecnologici e di sicurezza, i dati su cui vengono effettuate le operazioni di manipolazione da parte dello strumento di BI non sono mai quelli presenti nel *database* originario ma quelli storicizzati in un *database* specifico a disposizione dell'applicazione.

Inoltre, la qualità, l'omogeneità, la completezza e l'affidabilità dei dati in *input* rappresenta la discriminante per poter effettuare analisi valide ed ottenere risultati e stime attendibili.

Tuttavia, i risultati possono essere di vitale importanza per dipendenti e manager in quanto danno loro la possibilità di migliorare l'efficienza operativa e i processi decisionali, identificare nuove opportunità di mercato e di business, individuare aree critiche.

Sempre più frequentemente, nelle abitudini di vita quotidiane, si ha a che fare con tecnologie o informazioni alla cui base è implementata un'applicazione di *Business Intelligence* che ne garantisce il funzionamento; il caso più frequente è quello della *home banking* in cui possono essere visualizzati, attraverso una semplice connessione Internet, dati e informazioni relativi al proprio conto corrente con l'ausilio interattivo di tabelle, grafici e indicatori.

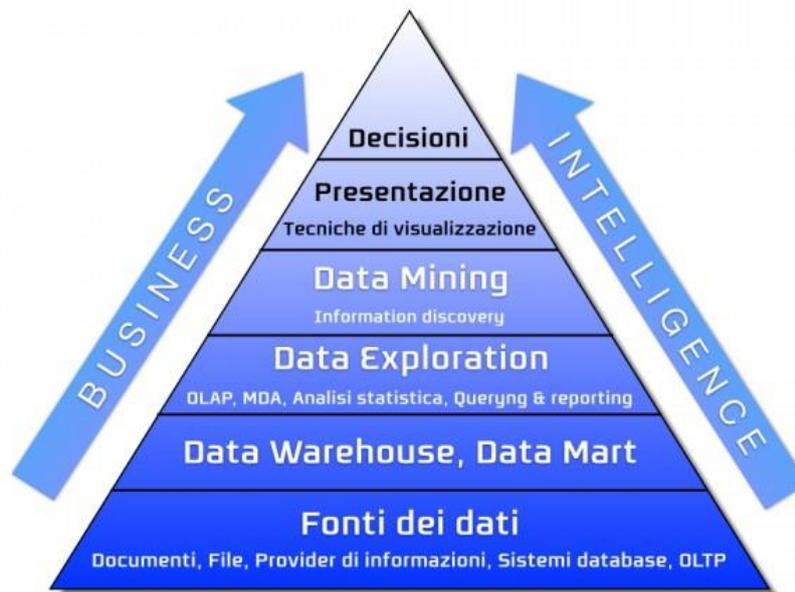
Un altro esempio con manifestazione frequente può essere quello relativo alle informazioni sul proprio traffico mobile ricevute via mail o via portale dal proprio gestore; infatti, anche in questo caso, le informazioni sul traffico mobile possono essere visualizzate mediante grafici e tabelle frutto di applicazioni BI.

Nella *Business Intelligence* esistono diversi strumenti *software* progettati per la manipolazione di grandi moli di dati da raccogliere, pulire, analizzare e trasformare.

Questi strumenti applicativi si suddividono nelle seguenti categorie:

- **Fogli di calcolo** (es. file Excel);
- **Data Warehousing**;
- **On Line Analytical Processing (OLAP) e Queryng**;
- **Data mining**;
- **Dashboard** (cruscotti dinamici).

Le categorie appena elencate rientrano nel vasto campo di attività e processi che racchiude la *Business Intelligence*, come mostra la **Figura 2.1**, tratta da un portale *online* di tecnologia<sup>[1]</sup>.



**Figura 2.1:** Piramide della *Business Intelligence*

La piramide della *Business Intelligence* presenta tutte le attività che racchiude questa disciplina ossia quelle di raccolta, analisi, trasformazione e visualizzazione dei dati; tuttavia, durante la trattazione, molte di queste attività saranno descritte con un caso pratico essendo state svolte per le analisi da compiere.

Secondo il giornalista ed esperto di tecnologia Ascenti<sup>[1]</sup>, la BI può essere concepita proprio come una piramide in cui, ad ogni attività successiva o livello, man mano che si sale, si ottengono dati e informazioni sempre più significativi e strutturati fino ad arrivare alla punta estrema della piramide: alla sua estremità infatti, si giunge ad un report o una *dashboard* in cui le informazioni, presentate interattivamente, fungono da supporto per le decisioni strategiche aziendali del *management*.

Alla base della piramide vi è la fonte dei dati che può essere sotto forma di: file, documenti o sistemi di *database*; i dati estratti dalla fonte originaria sono spesso destrutturati, frammentati e disomogenei.

Per ovviare a questo, i dati vengono dapprima manipolati tramite appositi strumenti di trasformazione dei dati di tipo ETL, che consiste in estrazione, trasformazione e caricamento dei dati, poi inseriti e archiviati in particolari *database* denominati *Data Warehouse* per migliorarne la struttura e l'organizzazione.

I *Data Warehouse*, che verranno descritti dettagliatamente nel seguito della trattazione, sono *database* che consentono di collezionare e aggregare enormi quantità di dati strutturati per poter operare analisi e report mediante tecniche specifiche.

I *Data Mart*, invece, rappresentano un sottoinsieme del *Data Warehouse*.

L'attività di *Data Exploration* comprende l'analisi di tipo OLAP e *Querying* che consente di processare velocemente, e in modalità piuttosto complesse, grandi quantità di dati selezionando solamente le informazioni ritenute più significative.

Queste tipologie di analisi sui dati vengono effettuate mediante l'utilizzo del linguaggio SQL, anche questo descritto in seguito, che risulta essere un linguaggio standardizzato utilizzato per eseguire operazioni sui *database* mediante l'utilizzo di costrutti denominati *query*.

Questo tipo di attività (*Data Exploration*) ha una valenza fondamentale per le organizzazioni in quanto risulta di grande utilità, ad esempio, per esaminare i risultati delle vendite, individuare nuove opportunità di business, verificare l'andamento di costi e ricavi, definire nuove strategie aziendali.

La successiva eventuale attività di *Data Mining* consiste, invece, nell'applicazione di sofisticati algoritmi, tecniche e metodologie con l'obiettivo di scoprire correlazioni tra più variabili o individuare, tra i dati, pattern significativi, potenzialmente utili e precedentemente ignoti; tra le tecniche maggiormente utilizzate, basate su particolari algoritmi, si ha il *clustering*, le *association rules*, e *classification*.

Utilizzando appositi strumenti applicativi, tutti i dati precedentemente manipolati e distillati con le attività descritte, vengono sintetizzati e presentati sotto forma di report, cruscotti dinamici e tecniche di visualizzazione.

Queste informazioni, messe a disposizione del *management* nel modo più immediato e intuitivo possibile, permettono di indirizzare le decisioni strategiche e operative del *decision maker*.

A conclusione di quanto esposto la BI è quindi un insieme di attività e processi che, all'interno di un'organizzazione, mira ad un miglioramento delle *performance* e al supporto dei processi gestionali e decisionali rendendoli *data driven* (cioè guidati dai dati).

Pertanto, tra i benefici che la *Business Intelligence* può generare, si ha:

- **Efficacia e rapidità del processo decisionale:** grazie alla velocità con cui un'applicazione di BI processa grandi moli di dati le decisioni, fondate su informazioni ed elementi certi, vengono assunte in modo quasi immediato;
- **Miglior gestione dei dati aziendali:** uno strumento di BI permette di migliorare l'utilizzo, la gestione e la sicurezza di grandi quantità di dati;
- **Efficienza operativa:** si può avere una miglior visione dei processi con la possibilità di ottimizzare specifiche fasi o interi processi produttivi;
- **Riduzione di spese inutili:** le informazioni e la conoscenza prodotte da un software di BI consentono di eliminare o ridurre sovrapposizioni o sprechi inutili.

### 2.1.1 Le basi di dati e il linguaggio SQL

All'interno delle organizzazioni, i dati, generalmente, vengono collezionati e aggregati all'interno di grandi magazzini virtuali denominati *Data Warehouse* e manipolati mediante il **linguaggio SQL**.

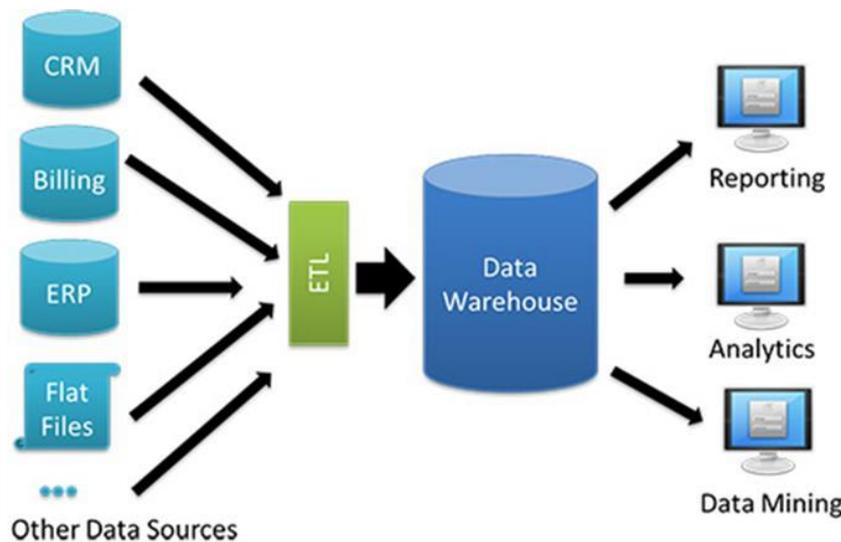
I *Data Warehouse* descrivono il processo di acquisizione, manipolazione e distribuzione dei dati e delle informazioni che fungono da supporto per le decisioni strategiche.

Infatti, questi dati, provenienti da fonti interne o esterne all'azienda, vengono prima manipolati mediante procedure di ETL e poi analizzati attraverso cruscotti dinamici, *query* complesse o tecniche di *Data Mining*.

Tuttavia, secondo lo studioso William H. Inmon, i dati immagazzinati all'interno di un *Data Warehouse* risultano essere:

- **Orientati ai soggetti di interesse:** in un DW (*Data Warehouse*) i dati vengono archiviati e organizzati in modo tale da facilitare la divulgazione di informazioni e della conoscenza mediante la lettura e l'elaborazione da parte degli utenti;
- **Integrati e consistenti:** uno dei requisiti fondamentali di un DW è l'integrazione e la consistenza dei dati raccolti; infatti, siccome i dati spesso provengono da fonti differenti questi devono essere resi uniformi e integrati tra loro mediante, ad esempio, l'utilizzo delle stesse unità di misura, o l'utilizzo di metodi di codifica identici;
- **Non volatili:** questa caratteristica indica che i dati contenuti nel DW non sono modificabili e possono essere acceduti in modalità di sola lettura;
- **Dipendenti dal tempo:** i dati contenuti in un DW sono relativi ad un orizzonte di tempo più esteso rispetto a quello di un sistema transazionale; ciò significa che in un sistema transazionale i dati risultano quasi sempre aggiornati e riferiti ad una situazione corrente al contrario del DW in cui l'orizzonte temporale dei dati è antecedente a quello in cui l'utente interroga il *database*.

La **Figura 2.2**, tratta da un portale online specializzato<sup>[2]</sup>, mostra gli elementi costitutivi di un *Data Warehouse*, a partire dalle sorgenti dati adattate mediante le procedure ETL (descritto in seguito con un caso pratico), la manipolazione dei dati nel *Data Warehouse* attraverso il linguaggio SQL (non visibile), fino ad arrivare alle analisi finali sui dati eseguite mediante strumenti specifici.



**Figura 2.2:** Elementi costitutivi di un *Data Warehouse*

In una base di dati o *Data Warehouse* il DBMS (*Database Management System*) è il sistema *software* che permette di memorizzare, gestire e interrogare collezioni di dati che siano grandi, condivise e persistenti, garantendo privacy ed affidabilità.

I dati, all'interno di un DBMS, vengono rappresentati per mezzo di un modello dei dati, ossia un insieme di concetti e costrutti che mirano ad organizzare i dati e la loro struttura in modo tale che risulti comprensibile ad un elaboratore.

Le caratteristiche che fanno di un modello dei dati un buon modello risultano essere:

- **Espressività:** i dati e le loro proprietà devono essere rappresentati in modo semplice e naturale;
- **Semplicità:** il modello deve essere basato su un numero minimo di meccanismi semplici da utilizzare e comprendere;
- **Realizzabilità:** la realizzazione del modello deve avvenire in modo efficiente su di un calcolatore.

Principalmente, esistono due tipologie di modello dei dati: il modello concettuale e il modello logico.

Mentre da una parte il modello concettuale descrive i concetti del mondo reale e viene utilizzato nelle primissime fasi di progettazione della base di dati, il modello logico è l'insieme di regole e costrutti utilizzati in un DBMS per descrivere la composizione ed il formato dei dati.

Nella realtà, esistono diversi tipi di modelli logici tra cui: il modello reticolare, gerarchico, quello relazionale e ad oggetti.

Il modello logico dei dati più diffuso e utilizzato nei DBMS è il modello relazionale, su cui si basano anche tutte le attività pratiche svolte per il lavoro di tesi in azienda.

Nel modello relazionale<sup>[3]</sup> tutti i dati vengono rappresentati in insiemi di record omogenei per mezzo di tabelle bidimensionali (dette anche relazioni) costituite da un certo numero di righe (tuple) e colonne (attributi e misure); inoltre, i dati vengono manipolati mediante il linguaggio SQL (*Structured Query Language*), uno specifico linguaggio di interrogazione (LI), componente essenziale delle basi di dati, che ne consente l'interrogazione e l'aggiornamento.

Ogni tabella, come quella mostrata in **Figura 2.3**, è un insieme di record in cui:

- Non è definito alcun ordinamento tra le n-uple (righe);
- Non è definito alcun ordinamento tra gli attributi e le misure (non è possibile individuarli tramite la posizione);
- Le n-uple (righe) sono tutte distinte tra loro (non esistono duplicati).

State	Mp	Mh	Commercial_name	mass_kg	Enedc_g_km	Wheel_base_mm	At1_mm	At2_mm	Fuel_type	Engine_capacity_m3
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	KADJAR	1448	132	2647	1562	1562	PETROL	1197
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	KADJAR	1448	132	2647	1562	1562	PETROL	1197
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	KADJAR	1395	130	2647	1562	1562	PETROL	1197
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CLIO	1165	85	2589	1509	1509	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461

**Figura 2.3:** Tabella Passenger\_cars

La tabella Passenger\_cars, mostrata in **Figura 2.3** e creata nel DBMS MySQL Workbench, è oggetto del caso di studio che verrà discusso nel seguito della trattazione; tuttavia, viene qui

mostrata per fornire un esempio di tabella nel modello relazionale e descriverne le caratteristiche di base:

- Le righe della tabella `Passenger_cars` rappresentano specifici record o tuple e descrivono le caratteristiche di un insieme di autovetture immatricolate in Europa tutte nel 2017.

Il loro numero totale è chiamato cardinalità ed in generale sono tutte diverse tra loro, ma in questo particolare caso, facendo le righe riferimento alle caratteristiche dell'autovettura, sono presenti duplicati;

- Ciascuna colonna della tabella (attributo o misura numerica) è identificata mediante un nome che è unico ed il quale descrive il ruolo del dominio.

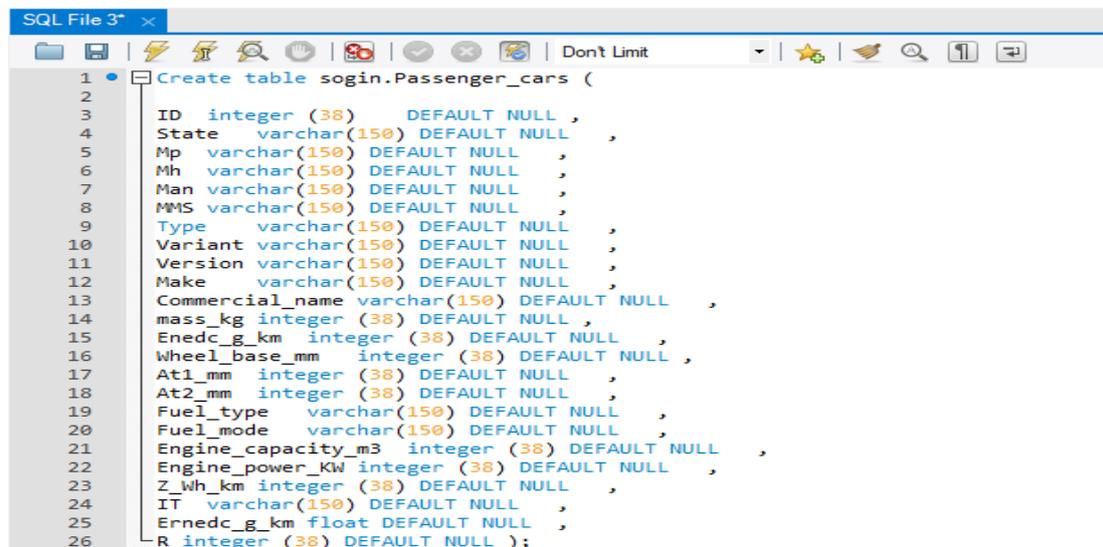
Tutti gli attributi e le misure visibili in tabella sono utilizzati come intestazione di ogni singola colonna (il cui ordine è irrilevante) ed il loro numero totale è detto grado.

In questo caso il grado risulta essere pari a 14;

- Il dominio, invece, indica l'insieme di valori che possono essere assunti da una colonna (attributo o misura) della tabella e possono essere numeri interi o decimali, stringhe, date e altro ancora.

Ad esempio, in riferimento alla tabella `Passenger_cars` in **Figura 2.3**, il dominio dell'attributo `State` (la prima colonna a partire da sinistra), di tipo stringa, è costituito da tutti gli stati europei in cui sono state immatricolate autovetture nel 2017.

Va aggiunto che, nel momento di creazione di una tabella e come mostrato (a titolo di esempio) in **Figura 2.4**, è necessario definire per ogni colonna il tipo di dato dei valori del dominio attraverso la dichiarazione di un tipo elementare (ad esempio dichiarando una stringa di dieci caratteri, numeri interi o decimali, date ecc.) oppure definendo tipologie di dato più complesse e articolate.



```
1 Create table sogin.Passenger_cars (
2
3 ID integer (38) DEFAULT NULL ,
4 State varchar(150) DEFAULT NULL ,
5 Mp varchar(150) DEFAULT NULL ,
6 Mh varchar(150) DEFAULT NULL ,
7 Man varchar(150) DEFAULT NULL ,
8 MMS varchar(150) DEFAULT NULL ,
9 Type varchar(150) DEFAULT NULL ,
10 Variant varchar(150) DEFAULT NULL ,
11 Version varchar(150) DEFAULT NULL ,
12 Make varchar(150) DEFAULT NULL ,
13 Commercial_name varchar(150) DEFAULT NULL ,
14 mass_kg integer (38) DEFAULT NULL ,
15 Enedc_g_km integer (38) DEFAULT NULL ,
16 Wheel_base_mm integer (38) DEFAULT NULL ,
17 At1_mm integer (38) DEFAULT NULL ,
18 At2_mm integer (38) DEFAULT NULL ,
19 Fuel_type varchar(150) DEFAULT NULL ,
20 Fuel_mode varchar(150) DEFAULT NULL ,
21 Engine_capacity_m3 integer (38) DEFAULT NULL ,
22 Engine_power_KW integer (38) DEFAULT NULL ,
23 Z_Wh_km integer (38) DEFAULT NULL ,
24 IT varchar(150) DEFAULT NULL ,
25 Ernedc_g_km float DEFAULT NULL ,
26 R integer (38) DEFAULT NULL );
```

**Figura 2.4:** Esempio sintassi nella creazione della tabella `Passenger_cars`

In questo caso, come visibile in **Figura 2.4**, per la maggior parte sono stati definiti tipi di dato stringhe (`varchar` di lunghezza pari a 150 caratteri), ma anche interi (`integer`) e decimali (`float`).

Inoltre, il DBMS (MySQL Workbench) controlla automaticamente che in ogni colonna della tabella vengano inseriti solo ed esclusivamente i valori di quel tipo di dato rispettando anche la lunghezza di cifre o caratteri.

Spesso può capitare che alcuni dati o informazioni non siano disponibili per specifici record (o tuple) della tabella.

E' il caso mostrato in **Figura 2.3** in cui, per le righe visualizzate, le ultime due colonne a destra della figura (identificate dalle misure `Z_Wh_km` e `Ernedc_g_km` ma non visibili per problemi di spazio) presentano solamente valori nulli, ossia valori speciali denominati NULL.

Il valore NULL può essere interpretato in diversi modi:

- **Valore sconosciuto o ignoto:** significa che esiste un valore del dominio di `Z_Wh_km` o `Ernedc_g_km` ma non è noto;
- **Valore inesistente o non definito:** significa che non esiste un valore del dominio;
- **Valore senza informazione:** non è noto se esista o meno un valore del dominio.

Tuttavia, i DBMS non distinguono le diverse tipologie di interpretazione del valore NULL né sono in grado di fornire le motivazioni alla base della presenza di un valore NULL.

Dal momento che i record contenuti in una tabella sono solitamente identificati in modo univoco da un attributo (ID, CF, ecc.) è prassi imporre restrizioni sulla presenza di valori nulli.

All'interno di una base dati, composta da un certo numero di tabelle, è importante distinguere:

- **Lo schema:** descrive la struttura dei dati ed è rappresentato dal nome di ogni tabella con le relative colonne;
- **L'istanza:** variabile nel tempo, è costituita dal contenuto o dalle righe di ciascuna tabella, ossia dai valori effettivi dei dati.  
Le righe della tabella `Passenger_cars`, visualizzate in **Figura 2.3** solo in parte, rappresentano l'istanza della base dati in cui è contenuta la tabella.

Detto ciò, il modello relazionale è un modello basato su valori: infatti, i riferimenti e le correlazioni tra i dati contenuti in tabelle diverse avvengono per mezzo dei valori comuni dei domini di specifici attributi o misure.

Questo è permesso grazie al vincolo d'integrità referenziale, fondamentale per garantire la correttezza dei riferimenti.

Tuttavia, la struttura basata su valori che possiede il modello relazionale, presenta diversi vantaggi:

- **Indipendenza dalle strutture fisiche**, che possono cambiare anche dinamicamente;
- **Si rappresenta solo ciò che è rilevante** dal punto di vista dell'applicazione;
- **I dati possono essere scambiati più facilmente da un sistema all'altro:** essendo le informazioni contenute nei valori risulta semplice trasferire basi di dati da un calcolatore ad un altro;
- **I valori consentono bi-direzionalità.**

Un DBMS deve prevenire l'immissione di dati e informazioni non corretti all'interno della base di dati; per questo si ha la possibilità di associare ad uno schema un insieme di vincoli.

Un'istanza (insieme di record o tuple) della base di dati viene detta legale quando soddisfa tutti i vincoli specificati nella creazione della tabella.

I vincoli, come riportato dettagliatamente anche da Golfarelli e Rizzi<sup>[4]</sup>, si suddividono in due categorie:

➤ **Vincoli intra-relazionali:** sono vincoli definiti sugli attributi e misure di una singola tabella e il cui soddisfacimento è definito rispetto a singole tabelle della base di dati, quindi sono vincoli associati alle tabelle in maniera indipendente l'una dall'altra e comprendono:

- **Vincoli di unicità:** grazie a questo tipo di vincolo non esistono due record con lo stesso valore per l'attributo su cui è definito un vincolo di unicità.

Un esempio è il vincolo di chiave, definito su uno o più attributi, che identifica in modo univoco le tuple di una tabella; ogni tabella ha almeno una chiave e non esistono limiti al numero di vincoli di chiave ammessi per una stessa tabella.

La chiave primaria, formata da uno o più attributi e scelta tra tutte le chiavi disponibili, non ammette valori nulli e identifica univocamente tutte le tuple della tabella; nella maggior parte dei casi è rappresentata da un attributo i cui valori risultano identificanti per l'organizzazione (targa, matricola, codice fiscale ecc.).

Nella tabella *Passenger\_cars* non è stata definita alcuna chiave primaria dal momento che, date le informazioni che rappresenta, sono ammessi duplicati;

- **Vincoli di tupla:** esprimono delle condizioni sul valore assunto da singole tuple, ed il cui soddisfacimento è indipendente dagli altri record della tabella.

Può essere, ad esempio, un'espressione booleana (and, or, not) di predicati semplici (es. confronto tra attributi) e può correlare diversi attributi;

- **Vincoli di dominio:** esprimono condizioni sul valore assunto da un singolo attributo di un record;

Anche questo può essere un'espressione booleana di predicati semplici.

➤ **Vincoli inter-relazionali:** sono vincoli che coinvolgono contemporaneamente più tabelle della base di dati e sono anche soprannominati:

- **Vincoli di integrità referenziale** (o di *foreign-key*): sono fondamentali per correlare correttamente le informazioni e i dati in tabelle diverse.

Tuttavia, permette di operare i riferimenti tra dati in tabelle diverse mediante valori comuni di uno o più attributi, nello specifico i valori delle chiavi primarie.

Esso identifica una o più colonne di una tabella (referenziante) che riferisce una o più colonne di un'altra tabella (referenziata); i valori presenti nelle colonne referenzianti devono essere presenti nelle colonne (referenziate) della tabella referenziata, ciò implica che un record nella tabella referenziante non può contenere valori che non esistono nella tabella referenziata (eccetto nel caso particolare di valori NULL).

Come riportato anche da Paraboschi insieme ad altri autori<sup>[5]</sup>, l'utilizzo di linguaggi per l'esecuzione delle operazioni di interrogazione e aggiornamento (inserimenti, modifiche, cancellazioni) sui dati e sullo schema è una componente fondamentale delle basi di dati e quindi di ciascun modello dei dati.

A questo proposito, il modello relazionale supporta linguaggi di interrogazione (LI) semplici e potenti caratterizzati da una forte base formale fondata sulla logica; i linguaggi di interrogazione (LI) permettono di reperire informazioni (sotto forma di dati) da un database, consentono un accesso semplice ed efficiente a grandi collezioni di dati e sono assai diversi dai linguaggi di programmazione, infatti non vengono utilizzati per calcoli troppo complessi.

Il linguaggio utilizzato per la memorizzazione, gestione, ed interrogazione delle basi di dati relazionali è il **linguaggio SQL (*Structured Query Language*)**.

L'SQL è un linguaggio dichiarativo che descrive cosa fare e non come fare, ponendosi ad un livello di astrazione superiore rispetto ai linguaggi di programmazione tradizionali.

Tuttavia, è un linguaggio che opera a livello di set, in quanto:

- Gli operatori operano su tabelle;
- Il risultato di una generica query è sempre una tabella.

Inoltre, mediante apposite istruzioni e una specifica sintassi, consente di:

- Definire e modificare lo schema di una base di dati relazionale;
- Leggere e manipolare i dati contenuti nelle tabelle;
- Definire lo schema di tabelle derivate;
- Definire i privilegi di accesso degli utenti;
- Gestire le transazioni.

Il linguaggio SQL può essere diviso in tre sotto-linguaggi principali:

- **DDL (*Data Definition Language*)**: è il linguaggio che consente di effettuare operazioni quali creazione, modifica, eliminazione degli oggetti (tabelle e viste) memorizzati nella base di dati;
- **DML (*Data Manipulation Language*)**: questo, invece, è il linguaggio che consente la lettura, l'inserimento, l'aggiornamento e l'eliminazione dei dati all'interno del database. Le interrogazioni, eseguite mediante il linguaggio SQL, sono rese possibili grazie a questo sotto-linguaggio;
- **DCL (*Data Control Language*)**: infine, questo linguaggio fornisce e revoca agli utenti i permessi necessari per poter utilizzare i comandi di DDL e DML su particolari istanze e oggetti del database.

Un'interrogazione SQL può essere espressa in diversi modi pur non cambiando il significato ed è l'utente che deve scegliere l'alternativa migliore ottimizzando il tempo di esecuzione, la velocità del DBMS, l'efficienza, e altri parametri.

Pertanto, l'interrogazione espressa nel linguaggio SQL è caratterizzata da una particolare sintassi di base che viene mostrata in **Figura 2.5** sottostante e ne vengono descritte, per completezza, le principali caratteristiche:

```
SELECT [DISTINCT] lista-select
FROM lista-from
[WHERE condizione]
[GROUP BY lista gruppo]
[HAVING qualificazione gruppo]
[ORDER BY AttrDiOrdinamento]
```

**Figura 2.5:** Sintassi di base di un'interrogazione SQL

- La clausola **SELECT** specifica quali attributi e misure (*lista-select*), selezionati dalle tabelle contenute in *lista-from*, faranno parte della tabella prodotta dalla *query*; se si desidera che la tabella risultato non contenga duplicati si deve inserire la parola chiave opzionale **DISTINCT**;
- La clausola **FROM** contiene i nomi delle tabelle (*lista-from*) a cui si vuole accedere e su cui si vuole calcolare la *query*; il risultato parziale dell'interrogazione, senza considerare la clausola *where*, consiste nel prodotto cartesiano delle tabelle presenti nella clausola *from*;
- La clausola **WHERE** contiene una o più condizioni ed espressioni booleane realizzate combinando gli operatori di confronto (<, <=, =, >, >=, >) con operatori logici (and, not, or).  
Le condizioni vengono applicate al risultato del prodotto cartesiano tra le tabelle presenti nella clausola *from*;
- Tenendo conto della clausola **FROM** e della clausola **WHERE**, la clausola **GROUP BY** consente di partizionare la tabella prodotta dall'interrogazione parziale in sottoinsiemi; il raggruppamento delle righe avviene in base ai valori comuni della lista di attributi specificata nella **GROUP BY**.  
Questo è un elemento di fondamentale importanza in quanto spesso sorge l'esigenza di applicare gli operatori aggregati (**SUM**, **COUNT**, **MIN**, **MAX**, **AVG**), specificati nella clausola *select*, solamente a sottoinsiemi di righe; le funzioni aggregate operano su un insieme di valori e producono, come risultato, un unico valore (aggregato);
- Si possono anche imporre condizioni di selezione sui sottoinsiemi di righe determinati dalla clausola **GROUP BY**; la condizione di filtro sui gruppi avviene attraverso la clausola **HAVING**.  
Ogni sottoinsieme di righe determinato dalla *group by* è oggetto nel calcolo del risultato dell'interrogazione se e solo se l'argomento della clausola **HAVING** risulta soddisfatto.  
Generalmente, dal momento che le condizioni sugli attributi vengono poste nella clausola *where*, nell'argomento della clausola **HAVING** vengono utilizzati solamente gli operatori aggregati;

- Infine, il comando **ORDER BY** permette di operare un ordinamento sulle righe del risultato di una *query*.

Infatti, la tabella risultato è formata da un insieme di tuple non ordinate e spesso nelle basi di dati vi è l'esigenza di costruire un ordine preciso sulle righe della tabella.

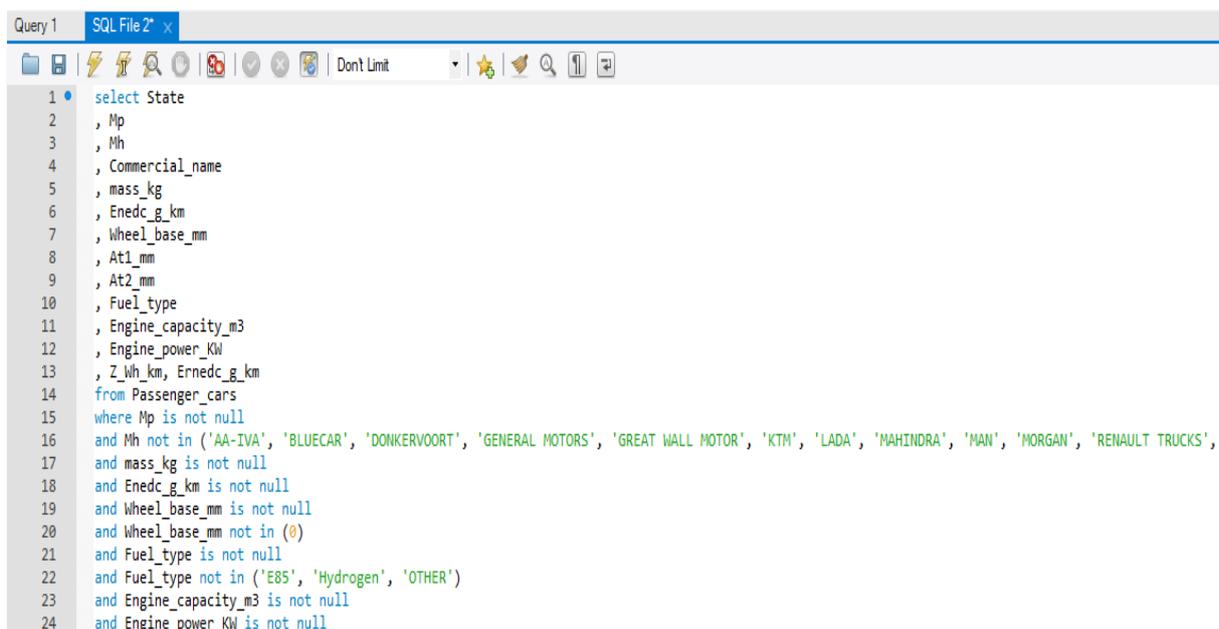
Le righe vengono ordinate in base al primo attributo presente nell'elenco e per righe che possiedono lo stesso valore si considerano, in sequenza, gli attributi successivi.

L'ordine degli attributi può essere ascendente o discendente a seconda che si usi il comando ASC o DESC (se è omissso, di default si assume il qualificatore ASC).

Quindi l'interrogazione SQL seleziona, tra le righe definite dal prodotto cartesiano delle tabelle elencate nella clausola FROM, solamente quelle che soddisfano le condizioni espresse nell'argomento della clausola WHERE.

Il risultato della *query* è ancora una tabella le cui colonne si ottengono dalla valutazione delle espressioni che appaiono nella clausola SELECT; se si desidera che nella tabella risultato siano presenti tutti gli attributi che caratterizzano le tabelle della clausola FROM si deve inserire, nella clausola SELECT, il carattere speciale asterisco (\*).

In **Figura 2.6** e in **Figura 2.7** viene mostrata, come esempio e a conclusione di quanto appena esposto, una *query* (ed il relativo risultato) eseguita per lo svolgimento del caso di studio.



```

1 select State
2 , Mp
3 , Mh
4 , Commercial_name
5 , mass_kg
6 , Enedc_g_km
7 , Wheel_base_mm
8 , At1_mm
9 , At2_mm
10 , Fuel_type
11 , Engine_capacity_m3
12 , Engine_power_KW
13 , Z_Wh_km, Ernedc_g_km
14 from Passenger_cars
15 where Mp is not null
16 and Mh not in ('AA-IVA', 'BLUECAR', 'DONKERVVOORT', 'GENERAL MOTORS', 'GREAT WALL MOTOR', 'KTH', 'LADA', 'MAHINDRA', 'MAN', 'MORGAN', 'RENAULT TRUCKS',
17 and mass_kg is not null
18 and Enedc_g_km is not null
19 and Wheel_base_mm is not null
20 and Wheel_base_mm not in (0)
21 and Fuel_type is not null
22 and Fuel_type not in ('E85', 'Hydrogen', 'OTHER')
23 and Engine_capacity_m3 is not null
24 and Engine_power_KW is not null

```

**Figura 2.6:** Esempio di una generica *query* SQL

State	Mp	Mh	Commercial_name	mass_kg	Eneqc_g_km	Wheel_base_mm	At1_mm	At2_mm	Fuel_type	Engine_capacity_m3	Engine_power_KW	Z_Vh_km	Ernedc_g_km
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	KADJAR	1448	132	2647	1562	1562	PETROL	1197	96	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	KADJAR	1448	132	2647	1562	1562	PETROL	1197	96	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	KADJAR	1395	130	2647	1562	1562	PETROL	1197	96	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CLIO	1165	85	2589	1509	1509	DIESEL	1461	55	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461	66	NULL	NULL

Figura 2.7: Esempio del risultato della generica query

Per quanto riguarda il linguaggio **DML (Data Manipulation Language)**, esso contiene i comandi e le istruzioni per modificare o aggiornare le istanze della base di dati.

Pertanto, le istruzioni di aggiornamento modificano lo stato della base dati, quindi è necessario verificare che siano rispettati i vincoli di integrità; ogni istruzione di aggiornamento può riguardare anche solo una tabella.

Tuttavia, l’inserimento, la cancellazione, e la modifica di tuple avviene mediante i seguenti comandi e un’opportuna sintassi:

- **INSERT** (inserimento di nuove tuple in una tabella): può essere inserita una singola tupla assegnando un valore costante agli attributi oppure possono essere inserite più tuple lette da altre tabelle mediante l’istruzione *select*;
- **DELETE** (cancellazione di tuple da una tabella): possono essere cancellate tutte le tuple di una tabella oppure solo quelle che soddisfano il predicato all’interno della clausola *where*;
- **UPDATE** (modifica del contenuto di tuple in una tabella): vengono aggiornati tutti i record della tabella che soddisfano il predicato nella clausola *where* e vengono modificati in base all’assegnazione “colonna = espressione” nella clausola *set*.

Per quanto riguarda, invece, il linguaggio **DDL (Data Definition Language)**, esso contiene i comandi che permettono di definire lo schema della base di dati, le tabelle derivate (es. viste) e strutture dati accessorie (es. indici).

Le operazioni più importanti, ossia quelle di creazione, modifica e cancellazione di una tabella avvengono mediante le seguenti istruzioni:

- **CREATE TABLE** (creazione di una tabella): è l’istruzione più importante e definisce una nuova tabella specificandone gli attributi, il loro dominio e un insieme, eventualmente vuoto, di vincoli;

- **ALTER TABLE** (modifica della struttura di una tabella): mediante questo comando è possibile modificare la struttura di una tabella aggiungendo o eliminando una colonna o un vincolo;
- **DROP TABLE** (cancellazione di una tabella): con la seguente istruzione tutte le righe della tabella vengono definitivamente eliminate, insieme alla tabella stessa, dalla base dati.

Tuttavia, si è voluta fornire un'impronta dettagliata sul modello relazionale e sul linguaggio SQL dal momento che tutte le attività svolte nel caso di studio, che sarà descritto nel seguito, sono state rese possibili grazie a prodotti *software* e applicazioni basate proprio sul modello relazionale e sull'utilizzo del linguaggio SQL.

## 2.2 Gli Open Data

Inizialmente, nel presente paragrafo, si descrivono i concetti di base degli *Open Data*, quindi i possibili vantaggi, rischi e problematiche derivanti dal loro utilizzo e diffusione; viene anche affrontato il tema relativo alla disciplina dell'*open government*, ossia dell'apertura della pubblica amministrazione nei confronti di cittadini e imprese.

Infine, vengono riportati i principali canali di diffusione di dati *open* visitati e i relativi *dataset* reputati potenzialmente utili e significativi per l'analisi e la costruzione di report (mediante il prodotto di *Business Analytics* aziendale).

### 2.2.1 Cosa sono gli Open Data

Il concetto di *Open Data* (in italiano dati aperti), letteralmente inteso come informazione e conoscenza pubblica facilmente accessibile ed utilizzabile, è adoperato ormai da molti anni.

Le eventuali restrizioni nell'utilizzo di questo tipo di dati sono l'obbligo di citarne la fonte e mantenere la banca dati sempre aperta.

L'inizio dell'uso di questa nozione risale all'anno 2009 quando alcuni governi come Regno Unito, Nuova Zelanda, Canada e Stati Uniti annunciarono diverse iniziative per favorire l'apertura dell'informazione e dei dati pubblici.

Ad oggi, non esiste un accordo condiviso su una definizione univoca di questo termine, al contrario, invece, del concetto di *open source* e del *software* libero, le cui definizioni formali sono state decise e condivise a livello internazionale.

A questo proposito, il progetto *Open Definition* di *Open Knowledge Foundation*<sup>[6]</sup> utilizza la seguente frase per spiegare cosa sono i dati *open*: ***“I dati aperti sono dati che possono essere liberamente utilizzati, riutilizzati e ridistribuiti da chiunque, soggetti eventualmente alla necessità di citarne la fonte e di dividerli con lo stesso tipo di licenza con cui sono stati originariamente rilasciati”***.

Le caratteristiche fondamentali che devono contraddistinguere i dati *open*, secondo la *Full Open Definition*<sup>[6]</sup>, risultano essere:

- ✓ **Disponibilità e accesso:** i dati devono essere disponibili ad un prezzo inferiore rispetto ad un ragionevole costo di riproduzione e l'accesso deve avvenire mediante un formato utile e modificabile;
- ✓ **Riutilizzo e redistribuzione:** i dati devono essere diffusi in condizioni tali da garantirne il riutilizzo, la redistribuzione e la possibilità di combinarli con altre basi di dati;

- ✓ **Partecipazione universale:** chiunque deve avere la possibilità di usare, riutilizzare, e ridistribuire i dati; non devono verificarsi discriminazioni né di ambito di iniziativa né contro soggetti o gruppi; questo significa che i dati possono essere utilizzati anche per scopi commerciali.

In generale, gli *Open Data* contengono informazioni e dati rappresentati in forma di *Database* e disponibili, solitamente, in diversi formati quali CSV, Excel, json, html; le tematiche di riferimento possono riguardare l'ambiente e l'inquinamento, i trend sulla popolazione, la geografia, il turismo, l'agricoltura così come dati medici, dati governativi e molto altro ancora. Il canale principale di diffusione dei *dati open* risulta, ovviamente, internet, anche se, la presenza di diversi fattori, che verranno discussi nel prossimo paragrafo, ne impediscono una vasta diffusione.

Tuttavia, questo tipo di dati sono accompagnati dalla locuzione “aperti” in quanto caratterizzati da una caratteristica essenziale: **l'interoperabilità**<sup>[6]</sup>.

L'interoperabilità rappresenta, sinteticamente, la capacità di sistemi e organizzazioni differenti di lavorare insieme, ossia, di inter-operare; in questo caso rappresenta l'abilità nel combinare una base di dati con altre.

La questione fondamentale di un insieme di dati accessibili ed utilizzabili consiste, quindi, nella possibilità di poterli mescolare con altri dati aperti.

In sostanza, l'interoperabilità dei dati consente di sviluppare migliori prodotti o servizi grazie alla possibilità di combinare insieme diverse basi di dati sfruttando il principale vantaggio di questi dati, l'apertura.

Esistono dei sostenitori degli *Open Data*, identificabili in un movimento specifico, i quali sostengono che i dati andrebbero considerati come beni comuni.

Alcune considerazioni, a sostegno di questo punto di vista, risultano essere le seguenti:

- I dati appartengono al genere umano, gli esempi sono i dati sulla scienza medica, metereologici e ambientali;
- I dati relativi alla pubblica amministrazione e finanziati da denaro pubblico devono essere divulgati pubblicamente per la conoscenza da parte della comunità contribuente;
- Le restrizioni sull'utilizzo dei dati e sulla loro diffusione rappresentano un freno allo sviluppo della società;
- Il tasso di scoperta e l'efficacia delle ricerche in campo scientifico sono favoriti da un migliore accesso ai dati.

## 2.2.2 Vantaggi e rischi nell'utilizzo di dati open

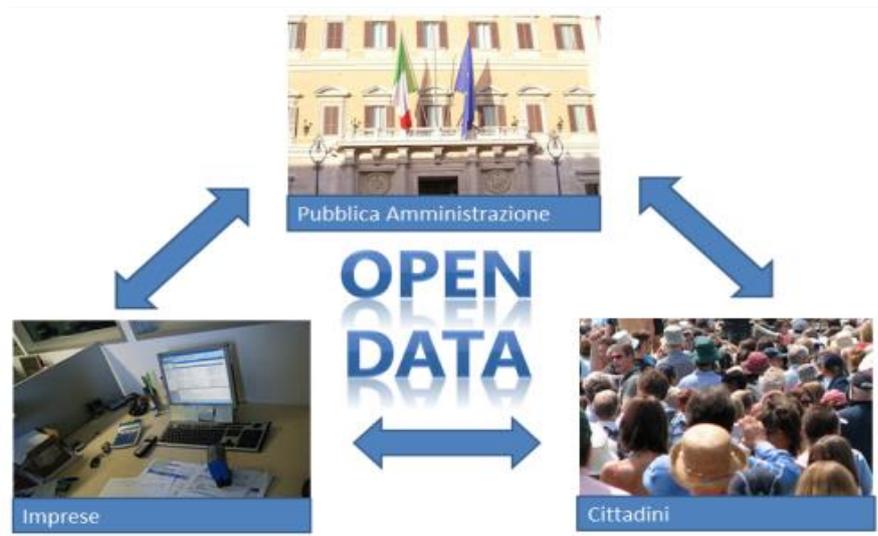
Gli *Open Data* costituiscono un'immane risorsa, spesso in larga parte inutilizzata.

Questo tipo di dati rimandano alla disciplina dell'*open government*, rappresentata graficamente in **Figura 2.8**, secondo cui, la pubblica amministrazione, in ottica di trasparenza verso i cittadini, dovrebbe rilasciare questi dati, appartenenti proprio alla comunità contribuente.

Infatti, molti organismi pubblici, così come organizzazioni private, nello svolgimento delle loro attività quotidiane, raccolgono e producono un'ampia gamma di dati, talvolta non disponibili in formati facilmente manipolabili e utilizzabili.

I dati governativi, per la maggior parte pubblici (per legge), giocano un ruolo significativo per la comunità e per la pubblica amministrazione stessa, sia per la quantità che per la qualità dei

dati raccolti; questi dati, infatti, possono essere scaricati e utilizzati liberamente da enti pubblici e privati, cittadini e aziende, anche con fini commerciali.



**Figura 2.8:** Rappresentazione grafica della disciplina dell'*open government*

Tuttavia, il grande vantaggio nell'apertura di grandi masse di dati consiste nell'estrazione di potenziale conoscenza e informazioni tali da creare valore e avere un impatto positivo per cittadini, imprese e pubblica amministrazione stessa.

Nello specifico, le grandi moli di dati rese pubbliche, che possono riguardare trasporti, informazioni su personale e stipendi, conti economici e spese, trend demografici, fenomeni ambientali, inquinamento, consentono di:

- Migliorare la vita e la conoscenza della collettività.
- Implementare app e servizi che riusino questi dati, su volontà di sviluppatori e aziende.

Questo permette senz'altro di aumentare l'efficienza e l'efficacia di governi e società e di definire delle *best practices* minimizzando, ove possibile, sprechi, costi inutili e malfunzionamenti.

A sostegno di quanto appena detto, vi sono molti esempi e casi reali in cui i dati pubblici hanno generato un impatto positivo e creato valore per la collettività<sup>[6]</sup>.

Uno di questi è il Finlandese "*tax tree*" (albero delle tasse) ed il Britannico "*where does my money go*" (dove vanno i miei soldi), che permettono di conoscere come il governo spende i soldi versati dai cittadini.

In Olanda, il servizio "*vervuilingsalarm.nl*", implementato grazie all'utilizzo di dati ambientali, consente di conoscere quando la qualità dell'aria del quartiere di residenza supera una certa soglia preimpostata.

Inoltre, grazie alla pubblicazione dei dati del sistema educativo da parte del ministero dell'istruzione olandese, si è avuta un'ottimizzazione del carico di lavoro e una riduzione dei costi.

In Danimarca, una sviluppatrice ha creato un'applicazione denominata "*findtoilet.dk*" che consente di conoscere la lista di tutti i bagni pubblici del paese con la relativa geo-localizzazione.

Un altro esempio importante può riguardare i dati sull'utilizzo dell'energia e sulle certificazioni energetiche degli edifici attraverso i quali si potrebbe aumentare l'efficienza della rete, ottimizzare i consumi e ridurre i costi.

A questo proposito, il sito danese "*husetsweb.dk*" permette di ottimizzare l'utilizzo di energia elettrica in casa offrendo la possibilità agli utenti di eseguire una pianificazione finanziaria e di servirsi dei migliori artigiani nel caso si necessiti di una riparazione.

Uno dei problemi nell'utilizzo di dati *open* riguarda restrizioni a monte quali *privacy*, *copyright*, brevetti, segreti aziendali, e altre forme di controllo che ne limitano la riproduzione e l'uso.

Un'altra difficoltà che si può incontrare riguarda il valore commerciale che gli stessi dati possono avere; infatti, essendo di proprietà di organizzazioni pubbliche e private, queste mostrano una scarsa propensione alla condivisione del proprio patrimonio informativo.

Pertanto, il rischio è che il controllo sui dati e il limite al loro utilizzo viene imposto tramite licenze apposite, diritti d'autore e diritti di riutilizzo che rendono difficile la creazione di valore. Può esistere anche il rischio che i dati, una volta pubblicati, necessitano di un'ulteriore autorizzazione per poter essere utilizzati<sup>[7]</sup>.

I dati aperti (governativi e non), come discusso, possiedono un enorme potenziale che può essere estratto e sfruttato a pieno solamente se la loro apertura risulta essere completa e non vi sono limitazioni di alcun tipo che ne ostacolano l'accesso e l'utilizzo.

### 2.2.3 Fruibilità e significatività degli Open Data trovati

Esistono moltissimi siti (governativi e non), nazionali e internazionali, che sono specializzati nella raccolta e diffusione di dati *open*, suddivisi per categoria o ambito.

Ovviamente, non sempre i dati *open* sono disponibili in formati che ne consentono l'accesso e l'analisi e solo una parte di questi risulta potenzialmente interessante, fruibile e significativa per l'effettuazione di analisi approfondite.

A tal proposito, sono stati visitati tutti i siti nazionali e internazionali di maggior spicco contenenti *Open Data* con l'obiettivo di selezionare, in un primo momento, tutti quei *dataset* (file di dati) con un elevato grado qualitativo di contenuto, fruibili, e riguardanti temi sensibili quali inquinamento, fenomeni ambientali, sicurezza, popolazione, trasparenza, crimini e altro ancora.

Quindi, prendendo in esame l'insieme dei *dataset* selezionati inizialmente, si è proceduto con l'identificazione e la scelta, tra tutti quanti, di un file di dati utilizzabile, potenzialmente interessante e relativo ad un tema tra quelli citati sopra (ambiente, inquinamento, popolazione, sicurezza ecc.).

Lo scopo ultimo è stato quello di operare analisi ed estrarre conoscenza dai dati di partenza attraverso la creazione di report dinamici su KNOWAGE<sup>[8]</sup>, il prodotto di *Business Analytics* aziendale.

Pertanto, nel seguito, vengono riportati i siti visitati di maggior rilevanza contenenti dati *open*, con i relativi *dataset* (file di dati) individuati (inizialmente) di maggior interesse e fruibilità.

Il requisito fondamentale nella selezione dei *dataset* consiste in un numero di record almeno pari a 500.000 ed un numero di colonne almeno pari a 20.

I file di dati, suddivisi per categoria o ambito, sono consultabili e scaricabili al *link* indicato.

[www.Dati.gov.it](http://www.Dati.gov.it): il portale governativo, gestito da Agid (agenzia per l'Italia digitale), racchiude tutti i dati rilasciati dagli enti pubblici nazionali e locali; in totale i file aperti sono circa 14.000 di cui 6.665 prodotti da comuni e province e suddivisi per categoria. La maggioranza si riferisce ai trend della popolazione e ai fenomeni ambientali. Nel sito in questione non è stato individuato alcun dataset potenzialmente analizzabile e significativo.

[www.data.europa.eu](http://www.data.europa.eu): il portale europeo dei dati, attivo dal 2012 e contenente 13880 file di dati, fornisce l'accesso ai *dataset* pubblicati dalle istituzioni appartenenti all'unione europea. Anche in questo caso non è stato selezionato alcun *dataset* fruibile e analizzabile.

[www.eea.europa.eu](http://www.eea.europa.eu): l'agenzia europea sull'ambiente, fondata nel 1994, fornisce dati e informazioni sull'ambiente riguardanti clima, natura, trasporti, energia, agricoltura, per tutti coloro che sono impegnati nello sviluppo e nella salvaguardia della politica ambientale. Sotto, sono descritti i file di dati selezionati.

## TOPIC: QUALITA' DELL'ARIA

### *“Air quality annual statistics calculated by the EEA”*

Il *dataset* contiene informazioni (annuali) sulla qualità dell'aria raccolte dagli stati membri dell'agenzia europea dell'ambiente; la copertura temporale dei dati è dinamica.

Gli attributi e le misure principali sono: paese, anno di riferimento, stazione di misurazione, longitudine e latitudine della stazione, tipo di inquinante, concentrazione dell'inquinante rilevato, unità di misura della concentrazione dell'inquinante, data di inizio e fine misurazione, tempo di misurazione totale, data e ora dell'ultimo update del dato.

[http://aidef.apps.eea.europa.eu/?source=%7B%22query%22%3A%7B%22match\\_all%22%3A%7B%7D%7D%2C%22display\\_type%22%3A%22tabular%22%7D](http://aidef.apps.eea.europa.eu/?source=%7B%22query%22%3A%7B%22match_all%22%3A%7B%7D%7D%2C%22display_type%22%3A%22tabular%22%7D)

**Righe: 1045954    Colonne: 19**

### *“Air quality zones information reported by countries”*

Il *dataset* contiene la maggior parte delle informazioni del *dataset* precedente e aggiunge: zona geografica del paese in analisi, popolazione residente nella zona interessata, superficie della zona in km quadrati, autorità di competenza e target di riferimento (vegetazione, salute ecc.).

[http://aideb.apps.eea.europa.eu/?source=%7B%22query%22%3A%7B%22match\\_all%22%3A%7B%7D%7D%2C%22display\\_type%22%3A%22tabular%22%2C%22from%22%3A80%7D](http://aideb.apps.eea.europa.eu/?source=%7B%22query%22%3A%7B%22match_all%22%3A%7B%7D%7D%2C%22display_type%22%3A%22tabular%22%2C%22from%22%3A80%7D)

**Righe: 44544    Colonne: 20**

## TOPIC: TRASPORTI E INQUINAMENTO

### *“Monitoring of CO2 emissions from passengers cars “*

Il *dataset* contiene informazioni (strutturali e di emissione) per ogni nuova auto immatricolata nel territorio europeo nel 2017.

Gli Attributi e le misure rilevanti sono: nome produttore, paese di riferimento, nome commerciale autovettura, emissioni di CO2 (g/km), peso, cilindrata, potenza, alimentazione, consumo di energia elettrica (Wh/km), riduzioni di CO2 (g/km) tramite tecnologie innovative.

<https://www.eea.europa.eu/data-and-maps/data/co2-cars-emission-15#tab-european-data>

**Righe: 1048576    Colonne: 33**

[www.data.ny.gov](http://www.data.ny.gov): Il sito in questione fornisce circa 1600 file di dati relativi allo stato di New York riguardanti svariati *topic* quali energia e ambiente, educazione, salute, sicurezza pubblica ecc.

I *dataset* di maggior interesse sono riportati sotto.

## TOPIC: TRASPORTI

### ***“Motor vehicles crashes”***

Il *dataset*, creato nel 2013, contiene informazioni sugli incidenti tra veicoli a motore; l’aggiornamento dei dati è frequente.

Le informazioni più rilevanti sono: anno, data e ora dell’incidente, condizioni di luce, meteo, descrizione incidente, numero veicoli coinvolti.

<https://data.ny.gov/Transportation/Motor-Vehicle-Crashes-Case-Information-Three-Year-/e8ky-4vqe>

**Righe: 896k Colonne: 18**

### ***“Annual average daily traffic volume counts”***

Il *dataset*, trovato su un altro sito web inerente lo stato di New York, contiene informazioni sul volume di traffico medio giornaliero annuale dei veicoli nello stato di New York.

Informazioni principali: contea, municipio, strada di provenienza e destinazione, anno di analisi, direzione del flusso di traffico, numero medio annuale giornaliero di veicoli, numero medio di auto, numero medio di camion ecc.

<https://data.buffalony.gov/Transportation/Annual-Average-Daily-Traffic-Volume-Counts/y93c-u65y>

**Righe: 32.3k Colonne: 14**

## TOPIC: ENERGIA E AMBIENTE

### ***“Spill incidents”***

Il *dataset* contiene informazioni sugli sversamenti di petrolio o altri materiali pericolosi e inquinanti in mare o nel terreno.

La data di creazione risale a fine 2015 e l’ultimo aggiornamento a maggio 2019; l’aggiornamento dei dati avviene ogni mese.

Le informazioni di maggior interesse sono: numero identificativo, nome programma del sito/struttura, indirizzo del luogo dello sversamento, città/villaggio dello sversamento, codice identificativo SWIS, regione dello sversamento, data, causa sversamento, fonte sversamento (industriale/commerciale/istituzionale), data chiusura, nome comune del materiale, famiglia del materiale, quantità del materiale sversato, unità della misura, quantità recuperata.

<https://data.ny.gov/Energy-Environment/Spill-Incidents/u44d-k5fk>

**Righe: 491k Colonne: 20**

### ***“Oil and gas summary production data”***

Il *dataset* contiene informazioni sulla produzione dalle risorse di petrolio e gas nel periodo 1967-1999.

Le principali informazioni sono: anno di produzione, anno inserimento dati, nome operatore, città e contea, nome formazione della produzione, numero di pozzi petroliferi attivi (nell’anno di analisi, città, contea, e gruppo di formazione), numero di pozzi inattivi, numero di risorse di gas attive e inattive (per le stesse dimensioni precedenti), volume di gas e petrolio prodotti, volume di gas tassabile, codice di acquisto dell’acquirente.

<https://data.ny.gov/Energy-Environment/Oil-and-Gas-Summary-Production-Data-1967-1999/8y5c-ebxg>

**Righe: 30.1k Colonne: 20**

### ***“Energy Efficiency completed projects”***

Il *dataset* contiene informazioni su progetti volti a migliorare l'efficienza energetica indirizzati a strutture governative e completati a partire dal 1987; il dataset viene aggiornato su base trimestrale.

Le principali informazioni sono: data progetto, nome e codice progetto, regione, contea, città, indirizzo, cliente, agenzia, categoria progetto, costo totale progetto, riduzione domanda (KW), riduzione energia elettrica(kWh), risparmi di gas, risparmi di petrolio, risparmi di carbone, energia rinnovabile (KW).

<https://data.ny.gov/Energy-Environment/Energy-Efficiency-Completed-Projects-Beginning-198/erjw-j2zx>

**Righe: 2363 Colonne: 31**

## TOPIC: TRASPARENZA

### *“Salary information for state authorities”*

Il *dataset* contiene informazioni sui salari per ciascun impiegato in autorità di stato.

Le informazioni di maggior interesse sono: nome pubblica autorità, data fine anno fiscale, nome e cognome impiegato, ruolo impiegato, gruppo di appartenenza (amministrativo, tecnico, ingegnere, manager), dipartimento, salario base annuale, salario reale, paga straordinario, bonus *performance*, paga extra, altro compenso, compenso totale (somma dei precedenti).

<https://data.ny.gov/Transparency/Salary-Information-for-State-Authorities/unag-2p27>

**Righe: 756k Colonne: 20**

## TOPIC: SICUREZZA PUBBLICA

### *“Index crimes by county and agency”*

Il *dataset* contiene informazioni sui crimini commessi e registrati da più di 500 commissari di polizia e dipartimenti degli sceriffi.

I dati partono dall'anno 1990 e vengono aggiornati su base annuale.

Le informazioni riportate sono: contea, dipartimento di polizia che ha registrato il crimine, anno, totale crimini commessi, numero omicidi, stupri, rapine, furti con scasso, furti di veicoli a motore, regione del crimine.

<https://data.ny.gov/Public-Safety/Index-Crimes-by-County-and-Agency-Beginning-1990/ca8h-8gjq>

**Righe: 19.3k Colonne: 15**

[www.kaggle.com](http://www.kaggle.com): Il sito contiene più di 19.000 *dataset* di ambiti diversi e relativi a differenti nazioni che per accedervi si è resa necessaria la registrazione tramite la mail personale.

I *dataset* di maggior interesse trovati sono descritti qui di seguito.

## TOPIC: CLIMA E AMBIENTE

### *“Air quality annual summary”*

Il *dataset* contiene moltissime informazioni sulla qualità dell'aria e come l'inquinamento colpisce la salute e l'ambiente; i dati vengono raccolti da agenzie specializzate che li inviano all'EPA (agenzia sulla protezione ambientale).

Le principali informazioni sono: codice stato, codice contea, latitudine, longitudine, sostanza misurata, metrica utilizzata, descrizione misurazione, anno, unità di misura, conteggio osservazioni, media annuale valore, deviazione standard valore, massimo valore annuale, secondo valor massimo annuale, data valore massimo misurato, nome sito monitoraggio, indirizzo sito, nome stato, nome contea, nome città, data ultimo aggiornamento dati.

<https://www.kaggle.com/epa/air-quality>

**Righe: 1048576 Colonne: 55**

## **TOPIC: POPOLAZIONE**

### ***“Suicide rates overview 1985 to 2016”***

Il *dataset* contiene varie informazioni sui suicidi avvenuti nel periodo 1985-2016 per consentire analisi di vario tipo e cercare di aumentarne la prevenzione.

Le informazioni di maggior rilievo sono: paese, anno evento, sesso, range di età, numero di suicidi, numerosità popolazione, HDI per anno, GDP.

<https://www.kaggle.com/russellyates88/suicide-rates-overview-1985-to-2016>

## **2.3 Il caso di studio**

**Il caso di studio** ed il file di dati scelto per l'analisi e la costruzione di cruscotti dinamici sul prodotto aziendale riguarda le nuove autovetture immatricolate nell'unione europea nel 2017.

Il titolo dell'insieme dei dati, ossia ***“Monitoring of CO2 emissions from passengers cars”*** fa esplicito riferimento al *topic* in questione: l'inquinamento, in termini di CO2, generato dalle automobili.

I dati, trovati al portale [www.eea.europa.eu](http://www.eea.europa.eu) e scaricati in formato CSV dal *link* <https://www.eea.europa.eu/data-and-maps/data/co2-cars-emission-15#tab-european-data>, contengono informazioni (strutturali e di emissione) per ogni nuova auto immatricolata in un paese dell'unione europea nel 2017.

Nello specifico, ogni riga del file rappresenta una nuova autovettura con le proprie specifiche caratteristiche, immatricolata in una determinata nazione.

Tra gli attributi e le misure più significative si ha: nazione di immatricolazione, gruppo automobilistico e casa costruttrice dell'automobile, tipo di alimentazione, peso (in kg), cilindrata (in cc) e potenza (in kw) dell'auto, emissioni di CO2 (in g/km), riduzione di emissioni di CO2 (in g/km) grazie all'utilizzo di nuove tecnologie.

Tuttavia, nel sito sono presenti più file con la stessa struttura relativi ad anni diversi che, purtroppo, sono risultati di difficile accesso e utilizzazione a causa del formato e altre ragioni ignote.

Sarebbe stato utile ed interessante unire i dati sulle autovetture inerenti ad anni diversi ma questo non è stato possibile.

Piuttosto, si è deciso di concentrare l'analisi sui dati in possesso, quindi sulle autovetture immatricolate in Europa nell'anno 2017.

Il file contiene circa 1 milione di record e 33 colonne, rispettando, di fatto, il requisito relativo al numero minimo di righe e colonne.

A questo proposito, infatti, si è ricevuto il suggerimento del supervisore aziendale di analizzare un file di dati con un numero minimo di 500 mila righe e 20 colonne in modo tale da avere la possibilità di effettuare analisi diversificate e più approfondite.

Va detto che, essendo ogni anno le autovetture immatricolate in Europa circa 15 milioni, si è operata l'assunzione che il file di dati analizzato costituisca un campione rappresentativo del totale.

Tra tutti i file di dati individuati nei vari portali contenenti dati *open*, quello sulle autovetture immatricolate in Europa nel 2017 risulta essere, per diverse ragioni, il più interessante e significativo da analizzare.

Questo è dovuto sia ad un **maggior numero di attributi e misure codificabili e analizzabili**, sia al rilievo e alla sensibilità che, attualmente, riveste il **tema dell'inquinamento e del surriscaldamento globale** correlato alle automobili e alla mobilità alternativa.

Attualmente, il tema delle emissioni inquinanti e del surriscaldamento globale trova sempre più frequentemente spazio sui principali strumenti di diffusione di informazioni (giornali, radio, televisioni ecc.) a causa delle condizioni peggiorative (scioglimento di ghiacciai, stravolgimenti climatici, desertificazione, innalzamento del livello dei mari) verso cui sta andando il nostro pianeta.

Il mondo dei mezzi di trasporto (compresi aerei e treni) contribuisce, in Europa, al 20% del totale delle emissioni prodotte.

Per emissioni prodotte, principalmente, si fa riferimento all'anidride carbonica (CO<sub>2</sub>) sprigionata dai processi, la quale è considerata un climalterante, che, se in eccesso, contribuisce al surriscaldamento e all'effetto serra del pianeta; tuttavia, vengono considerate emissioni anche il consumo dei freni, l'usura degli ingranaggi e lo sfregamento delle gomme sull'asfalto<sup>[9]</sup>.

La polvere prodotta dal consumo degli pneumatici viene sversata nei fiumi dalle piogge, e costituisce una delle cause primarie della presenza di microplastiche nei mari.

Per ridurre le emissioni serra e per salvaguardare l'integrità ambientale apportando benefici in termini di salute, innovazione e protezione per la collettività, l'Europa ha imposto una stretta sulle emissioni di CO<sub>2</sub> delle automobili<sup>[10]</sup>.

Infatti, dal 2021, le automobili, in media, potranno emettere 95 g/km e 59 g/km a partire dal 2030.

I suddetti limiti, imposti ai singoli costruttori, verranno calcolati prendendo in considerazione tutte le automobili prodotte; questo significa che ci saranno auto con valori maggiori del limite e auto con emissioni al di sotto del limite imposto.

L'Unione Europea prevede premi per i gruppi la cui produzione di auto elettriche e ibride supererà la soglia del 15% (entro il 2025) e del 30% sul totale (entro il 2030), ma, allo stesso tempo,

invierà sanzioni salatissime nei confronti delle case che non rispetteranno il limite imposto ad eccezione dei gruppi minori; i gruppi considerati minori sono quelli con una produzione annua inferiore alle 1000 unità i quali saranno esentati, e quelli con un numero inferiore alle 10.000 unità (es. Ferrari) i quali potranno usufruire di una deroga<sup>[11]</sup>.

L'obiettivo dell'Europa, in sostanza, è quello di creare un'economia sostenibile per quanto riguarda i trasporti favorendo una maggior pulizia dell'aria all'interno delle città.

Questo può offrire un impulso alle innovazioni nel settore automobilistico ed il cui successo si basa sulla crescita delle vendite di auto elettriche ed ibride.

A questo proposito, mediante il file di dati selezionato, è possibile eseguire analisi e produrre conoscenza sul livello di emissioni e riduzioni medie prodotte per singolo paese e produttore, valutare le case ed i paesi maggiormente impegnati nell'introduzione di veicoli migliori, più puliti e a basse emissioni grazie all'utilizzo di nuove tecnologie.

Inoltre, può essere analizzato il livello medio di emissioni prodotte anche sulla base di caratteristiche strutturali delle automobili (cilindrata, potenza, peso, alimentazione).

## 3. L'azienda ospitante

### 3.1 Presentazione dell'azienda

Tutto il lavoro di tesi in azienda, descritto e argomentato durante la presente trattazione, si è reso possibile grazie all'azienda ospitante: **Engineering Ingegneria Informatica S.p.A.**

Engineering Ingegneria Informatica S.p.A. è una società per azioni costituita nel 1980 che opera nel settore del software e servizi IT (*Information Technology*).

Si dice che alla sua nascita l'azienda ha raccolto la sfida olivettiana di essere un'industria informatica con testa e cuore italiani ma gambe e braccia multinazionali.

Ad oggi, la società è *global player* e specializzata nella trasformazione digitale, in particolar modo per i settori finanza, pubblica amministrazione (centrale, locale, sanità), industria e *utilities*.

Inoltre, la società è stata quotata presso la Borsa di Milano dal dicembre 2000 fino al 7 luglio 2016 con revoca dal giorno successivo in seguito al dell'OPA lanciata da MIC Bidco S.p.A.<sup>[12]</sup>.

Il gruppo *Engineering*, con all'attivo circa 11.000 dipendenti e 65 sedi sparse in tutto il mondo, concentra le proprie attività nella progettazione, sviluppo e gestione di soluzioni innovative per le aree di business in cui la digitalizzazione genera i maggiori cambiamenti.

Le principali aree in questione risultano essere:

- *Digital Finance*;
- *Smart Government & Health*;
- *Augmented city*;
- *Digital Industry*;
- *Smart Energy & Utilities*;
- *Digital Telco & Multimedia*.

L'azienda, con le sue attività, fornisce un importante contributo per la modernizzazione del mondo in cui si vive e si lavora, combinando competenze specialistiche nelle tecnologie di ultima frontiera, infrastrutture tecnologiche organizzate in un modello unico di *multicloud* ibrido e capacità di interpretazione dei nuovi modelli di *business*.

Tuttavia, con investimenti significativi in ricerca e sviluppo, Engineering svolge un ruolo di primo piano nella ricerca, contribuendo alla realizzazione di progetti nazionali e internazionali di primissimo piano<sup>[13]</sup>.

In Italia la sede principale è a Roma, sede in cui è stato svolto sia il tirocinio curricolare sia il lavoro di tesi in azienda nell'ambito *Data & Analytics*, costituito da un insieme di attività di analisi complesse ed estrazione di valore dai dati, eseguite mediante modelli statistici predittivi.

### 3.1.1 Il settore aziendale di riferimento

Il lavoro di tesi in azienda, così come il propedeutico tirocinio curricolare, si è svolto nel **settore aziendale** di *Business Intelligence*.

La motivazione principale dietro alla scelta di questo specifico ambito è da ricondurre ad uno spiccato interesse personale riguardo il processo di raccolta e analisi di grandi volumi di dati (*Big Data*) con l'obiettivo di estrarre informazioni nascoste e una conoscenza significativa.

In **Figura 3.1** viene mostrato e sintetizzato, graficamente, il processo di raccolta e analisi di grandi masse di dati.



**Figura 3.1** Raccolta e analisi di dati

La *Business Intelligence* può essere vista come una piramide le cui attività, dal basso verso l'alto, sono quelle di raccolta, analisi, trasformazione e visualizzazione dei dati.

Ad ogni attività, man mano che si sale lungo la piramide, si ottengono informazioni sempre più strutturate fino ad arrivare alla sommità in cui queste vengono visualizzate mediante *dashboard* o report.

Tuttavia, come già discusso in precedenza, le organizzazioni si impegnano nel raccogliere e analizzare dati con l'obiettivo di estrapolare conoscenza, operare valutazioni e stime del mercato in cui competono e del contesto aziendale proprio.

Il fine ultimo è proprio quello di utilizzare le informazioni e le valutazioni generate per incrementare il proprio vantaggio competitivo rispetto ai concorrenti così come per supportare le decisioni di chi è capo di ruoli direzionali.

Da un punto di vista tecnologico, come già detto, la *Business Intelligence* permette di modellare e trasformare i dati di partenza in conoscenza attraverso l'utilizzo di strumenti software.

Tuttavia, un'applicazione *software* o un prodotto di BI accede a grandi quantità di dati storicizzati in un database esterno oppure sotto forma di file (csv, xls ecc.) e, mediante l'interazione dell'utente, consente di definire cruscotti o report contenenti grafici, tabelle, indicatori e statistiche descrittive dei dati in input.

A questo proposito, KNOWAGE, descritto nel prossimo paragrafo, è il prodotto di *Business Analytics* aziendale con cui è stata effettuata la prototipazione delle analisi e la costruzione dei cruscotti dinamici (presentati nell'ultimo capitolo).



Figura 3.2: Logo Knowage

## 3.2 Il prodotto di Business Analytics utilizzato

In questo paragrafo viene presentato il **prodotto di Business Analytics** utilizzato presso l'azienda Engineering Ingegneria Informatica S.p.A.

Dapprima viene effettuata una panoramica generale di **KNOWAGE** e, successivamente, vengono spiegate, in dettaglio, le funzionalità utilizzate.

### 3.2.1 Panoramica

**KNOWAGE** è una suite di *Business Intelligence* professionale completamente *Open Source* utile per le analisi moderne di *business* concernenti i *Big Data*.

Lo strumento, sviluppato e gestito da Engineering Ingegneria Informatica S.p.A., rappresenta la nuova versione del progetto SpagoBI, caratterizzata da significativi aggiornamenti sia funzionali che tecnologici e dall'offerta di un modello rivisitato.

KNOWAGE rappresenta anche uno strumento di supporto per i processi decisionali di business sia a livello direzionale che operativo.

Questo grazie al fatto che, essendo una *suite* di BI, copre tutte le aree analitiche della *Business Intelligence*, offrendo un ausilio ad amministratori e sviluppatori nello svolgimento delle loro *task* quotidiane.

La *suite* è caratterizzata da diversi moduli, ognuno concepito per uno specifico dominio analitico.

I moduli possono essere utilizzati singolarmente come soluzione per un determinato tipo di lavoro, oppure possono anche essere combinati tra loro per assicurare un pieno soddisfacimento dei requisiti dell'utente<sup>[14]</sup>.

I suddetti moduli risultano essere:

- **Big Data (BD):** permette di analizzare dati storicizzati su *clusters* specifici o su database NoSQL;
- **Smart Intelligence (SI):** rappresenta la *Business Intelligence* incentrata su dati strutturati, ma più orientata verso una prototipazione *agile* e con capacità *self-management* per l'utente finale;

- **Enterprise Reporting (ER):** consente di produrre e distribuire report statici;
- **Location Intelligence (LI):** permette di correlare i dati e le misure di business a dati geografici e spaziali;
- **Performance Management (PM):** offre all'utente finale la possibilità di gestire i KPIs e di organizzare schede di valutazione bilanciata (*balanced scorecard*);
- **Predictive Analysis (PA):** è utile per formulare analisi più avanzate.

KNOWAGE è caratterizzato da una visione moderna sull'analisi dei dati, in quanto mette a disposizione strumenti che conferiscono una certa autonomia all'utente finale.

A tal proposito, quest'ultimo è in grado di definire diversi tipi di analisi ed esplorare il proprio spazio di dati combinandoli da svariate fonti.

KNOWAGE si basa principalmente su due modelli concettuali: il modello analitico e il modello di business.

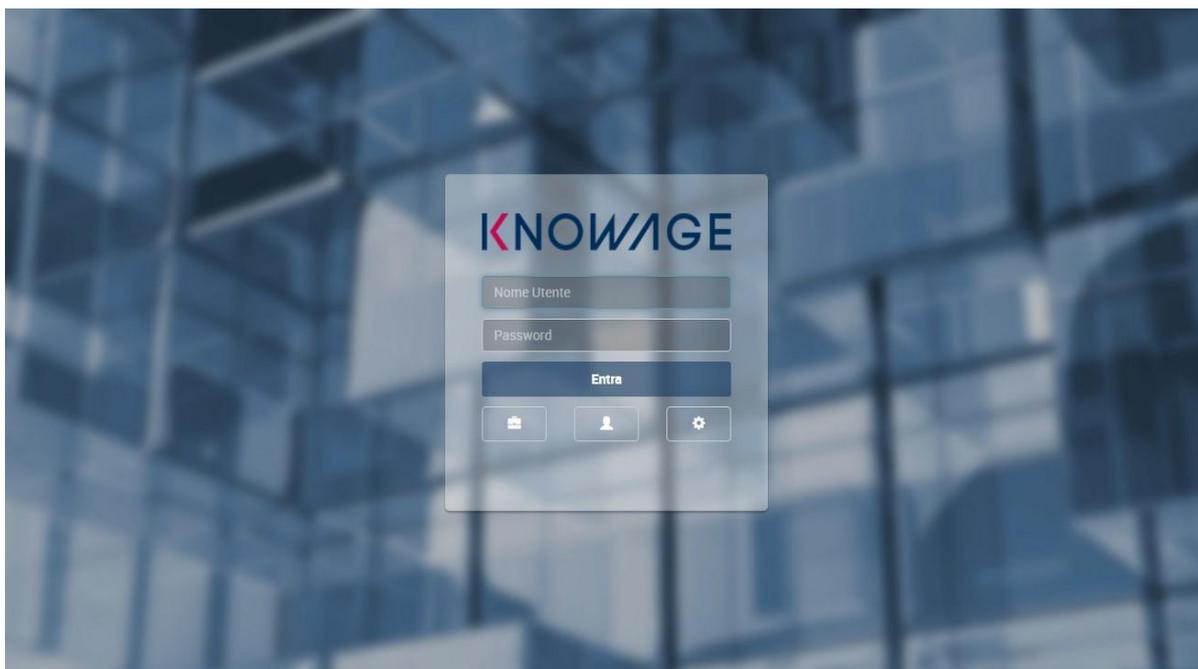
Il modello analitico costituisce il cuore della suite ed offre diverse soluzioni per le varie aree analitiche; nel seguito sono descritte le principali funzionalità offerte<sup>[14]</sup>:

- **Report:** permette di rappresentare dati strutturati;
- **Analisi OLAP:** consente la navigazione attraverso i dati;
- **Dashboard:** per il monitoraggio dei KPIs;
- **Modello KPI:** permette di costruire il proprio modello di monitoraggio prestazionale;
- **Cockpit:** consente di realizzare cruscotti dinamici e interattivi e rappresenta la funzionalità utilizzata per il lavoro di tesi in azienda;
- **Query by Example (Interrogazione libera dei dati):** fornisce la possibilità di comporre liberamente le proprie interrogazioni e generare report;
- **Data mining:** per scoprire patterns, informazioni e correlazioni tra i dati nascoste;
- **Processi ETL:** per la raccolta, trasformazione e caricamento dei dati provenienti da diverse sorgenti.

### 3.2.2 L'interfaccia utente e il modello comportamentale

Per poter accedere in un qualsiasi ambiente di KNOWAGE, con l'obiettivo di svolgere un determinato tipo di lavoro, sono necessari un *username* e una *password* forniti dall'amministratore.

In **Figura 3.3** è mostrata la schermata tipica di accesso.



**Figura 3.3:** Schermata di accesso principale a KNOWAGE

Tuttavia, le credenziali con cui si effettua l'accesso sono associate allo specifico ruolo dell'utente.

All'interno di KNOWAGE, i ruoli rappresentano una **categorizzazione di gruppi** di utenti differenti, con lo scopo di garantire, a ciascuno di questi, diversi diritti e criteri di visibilità su documenti e dati; quanto appena detto è riassunto nel modello comportamentale, che viene descritto nel seguito.

Quindi, una volta effettuato l'accesso, si visualizza la schermata in **Figura 3.4**.



**Figura 3.4:** Home page di KNOWAGE

Cliccando sul tasto in alto a sinistra l'utente può visualizzare la schermata del menu principale, raffigurata in **Figura 3.5**, che consente l'accesso a documenti, dati e a tutte le funzionalità per cui si ha il permesso di utilizzo.



**Figura 3.5:** Menu principale di KNOWAGE

Il menu principale si suddivide in due sottomenu:

- **Menu utente;**
- **Menu delle funzionalità di BI.**

A questo proposito, il menu utente è quello situato nel pannello di sinistra, mentre quello delle funzionalità di BI si trova sul pannello di destra.

Per quanto riguarda il menu utente, cliccando sull'icona in alto a sinistra che raffigura l'username (in questo caso tirocinio ADMIN), con il quale viene effettuato il *log-in*, si visualizzano una serie di piccole icone con extra funzionalità, descritte nella **Tabella 3.1**.

**Tabella 3.1:** Extra funzionalità di KNOWAGE

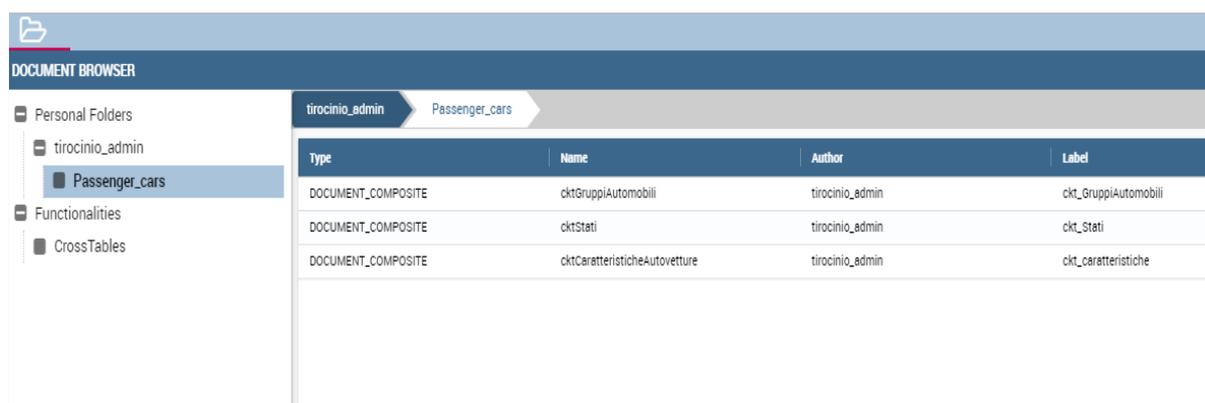
Icona	Nome	Descrizione
	Seleziona ruolo	Permette di selezionare il ruolo di autenticazione (disponibile se si è associati a più di un ruolo).
	Lingua	Consente di scegliere la lingua.
	Info	Fornisce informazioni generali sulla versione di KNOWAGE.
	Esci	Permette di effettuare l'uscita dall'ambiente.

Sempre in alto a sinistra, appena sotto l'icona che raffigura l'username, sono presenti due piccole icone che rappresentano, rispettivamente, il *Document browser* ed il *Workspace*.

Il *Document browser* consente di creare i documenti analitici e, al tempo stesso, di accedervi ed eseguirli.

La suddetta funzionalità è rappresentata in **Figura 3.6**, in cui sono presenti, sulla destra, i documenti analitici creati per le analisi sui dati delle autovetture e che saranno descritti nel quarto ed ultimo capitolo.

Tuttavia, è possibile accedere ed eseguire i documenti con la possibilità di effettuare modifiche così come di crearne di nuovi cliccando sul bottone rosso con il simbolo del più in alto a destra.



**Figura 3.6:** *Document browser*

Il *Workspace* invece, visibile in **Figura 3.7**, contiene le sezioni: *Recent*, *Data*, *Analysis*; nella prima sono presenti gli ultimi documenti sui cui si è lavorato mentre la sezione *Data* è

caratterizzata dai *dataset*, suddivisi in quattro categorie, utilizzati per le analisi effettuate sui documenti.

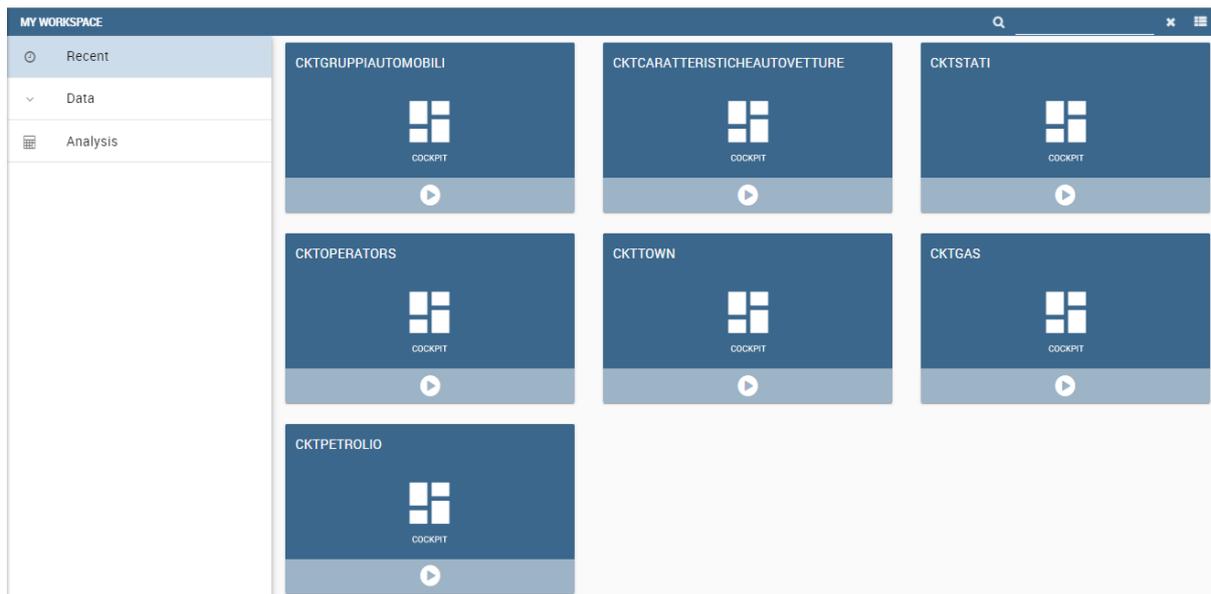


Figura 3.7: Workspace

Per quanto riguarda il menu delle funzionalità di BI, situato sul pannello di destra del menu principale, questo è costituito da diverse sezioni con specifiche funzionalità<sup>[14]</sup>:

- **Data Providers:** in questa sezione è possibile creare e gestire i *Data source* e i *Dataset*;
- **Profile Management:** qui si ha la possibilità di creare e gestire il profilo degli utenti, le autorizzazioni e i ruoli, così come di definire il modello analitico; inoltre si può configurare il menu e l'albero delle funzionalità;
- **Behavioural model:** in questo pannello si può gestire tutto ciò che riguarda il modello comportamentale, quindi accedere e modificare la configurazione dei vincoli e i driver analitici;
- **Catalogs:** in quest'area è possibile gestire differenti cataloghi, i quali possono variare da prodotto a prodotto;
- **Tools:** qui l'utente può accedere a diversi strumenti e opzioni di schedulazione;
- **Server Settings:** in questo pannello si ha accesso a tutte le opzioni di configurazione delle impostazioni del server;
- **Kpi Model:** qui l'utente ha la facoltà di definire KPIs, *scorecard*, e target;
- **Server Manager:** questo è un pacchetto opzionale che, se accessibile, rende possibile la gestione dei *template* e l'importazione o l'esportazione di documenti, KPIs, metadati, cataloghi, glossari e utenti.

**Il modello comportamentale (*Behavioural model*)<sup>[14]</sup>**, introdotto all'inizio di questo paragrafo, è un elemento di fondamentale importanza; infatti, prima di iniziare un qualsiasi nuovo progetto, è essenziale la creazione e la gestione del modello comportamentale.

In sostanza, il modello comportamentale decide **chi** può utilizzare la soluzione di *Business Intelligence*, **che cosa** è visibile agli utenti in termini di dati e documenti e **come** gli utenti lavorano con i loro documenti; pertanto, è basato su quattro concetti chiave:

- **Profilo utente:** definisce il ruolo e gli attributi dell'utente;
- **Diritti di visibilità:** definiscono i diritti dell'utente in termini di accessibilità su dati e documenti;
- **Driver analitici:** definiscono quali dati di uno specifico documento possono essere mostrati all'utente;
- **Impostazioni di ambiente:** definiscono come l'utente può eseguire i propri documenti.

A questo proposito, gli utenti di KNOWAGE sono definiti da:

- **Identità:** consiste in tutti quei dati utilizzati per identificare un utente, quali username e password;
- **Ruolo:** rappresenta una categorizzazione di gruppi di utenti, i quali possono corrispondere a specifiche posizioni all'interno dell'azienda; pertanto, diversi utenti possono avere lo stesso ruolo così come uno stesso utente può avere più ruoli;
- **Profilo:** consiste in un insieme di proprietà chiamate attributi che forniscono informazioni generali riguardo l'utente, come ad esempio sesso, età, l'unità organizzativa a cui appartiene ed altre aggiunte in modo facoltativo dall'amministratore.

In **Tabella 3.2** vengono riportati, per completezza, i diversi tipi di ruolo a cui può essere associato un utente e la relativa descrizione<sup>[14]</sup>.

**Tabella 3.2:** Tipi di ruolo

Tipo ruolo	Descrizione
ADMIN	Amministratore generale; può utilizzare tutte le funzionalità di KNOWAGE.
MODEL_ADMIN	Amministratore del modello; gestisce il modello comportamentale e le funzionalità associate.
DEV_ROLE	Sviluppatore; crea e modifica i dataset e i documenti analitici.
TEST_ROLE	Utente di test; verifica i documenti analitici.
USER	Utente finale; esegue i documenti a lui visibili e può definire report e analisi specifiche.

Mentre i primi quattro ruoli definiti sono di natura tecnica, il quinto ed ultimo ruolo descritto risulta coincidere con il reale utente finale.

KNOWAGE consente di definire più ruoli in base alle specifiche esigenze progettuali e di lavoro ed ogni ruolo deve necessariamente appartenere ad un unico tipo tra quelli appena descritti.

Quando un utente effettua l'accesso in un determinato ambiente di lavoro, il suo profilo ed il ruolo vengono automaticamente riconosciuti e autenticati; nel caso in cui l'utente è associato a più ruoli gli viene chiesto di selezionarne uno.

Mediante la sezione *Profile Management*, disponibile nel menu principale ma solo per gli utenti amministratori a causa dell'elevato livello di responsabilità, è possibile gestire il profilo degli utenti e i ruoli<sup>[14]</sup>.

Tuttavia, la sezione in questione contiene tre funzionalità specifiche:

- **Gestione attributi profilo:** serve per definire nuovi attributi di un profilo e per gestire quelli già esistenti;
- **Gestione ruoli:** permette di creare nuovi ruoli e gestire i permessi per ogni ruolo;
- **Gestione utenti:** consente di creare nuovi utenti, gestire le loro identità, assegnare dei valori ai loro attributi e associarli a uno o più ruoli.

Quindi, il modello comportamentale permette di regolare la visibilità e i diritti su documenti e dati in accordo ai ruoli e ai profili degli utenti; inoltre, all'interno di un progetto di BI, assicura diversi vantaggi, tra cui: riduzione del numero di documenti analitici da sviluppare e gestire, definizione univoca delle regole di visibilità su dati e documenti, crescita del progetto costante nel tempo, garanzia del rispetto delle regole di visibilità nel tempo con la possibilità di aggiungere più documenti analitici.

### 3.2.3 I cockpit (cruscotti dinamici)

La funzionalità utilizzata per la costruzione dei cruscotti dinamici, relativamente alle analisi sui dati delle autovetture, è quella dei cosiddetti *cockpit*.

I *cockpit* sono documenti o cruscotti dinamici che possono essere creati ed eseguiti su KNOWAGE sia da utenti tecnici che da utenti finali attraverso un'interfaccia intuitiva e interattiva aggiungendo diversi tipi di *widget* (testuali, diagrammi, grafici, tabelle ecc.); è possibile definire associazioni tra gli stessi *widget*, infatti, cliccando sui dati di un determinato *widget* tutti gli altri si aggiornano automaticamente in base al tipo di dato selezionato.

Un aspetto chiave consiste nel fatto che diversi *widget* di uno stesso *cockpit* possono basarsi su diversi *dataset* e anche su diversi *data sources*.

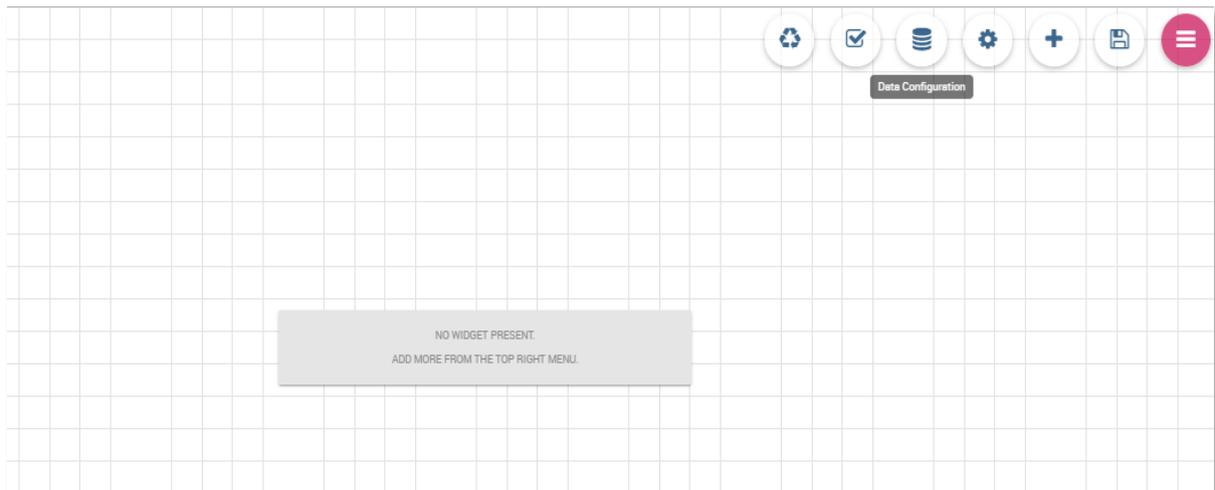
Inizialmente, nella creazione del primo *cockpit*, è necessario selezionare l'opzione *Functionalities Management* nella sezione *Profile Management* del menu principale (mostrato in **Figura 3.5**); in questo modo viene mostrato all'utente l'albero delle funzionalità (*functionalities tree*), che consente di organizzare i documenti analitici in modo gerarchico per facilitarne la ricerca, l'accessibilità e per gestire in modo efficiente la visibilità in base ai ruoli dell'utente.

Pertanto, è possibile inserire una nuova foglia nell'albero specificando nome, etichetta e permessi ai ruoli, cancellarne una già esistente a patto che non contenga uno o più documenti, oppure muovere su e giù un elemento lungo la gerarchia.

Fatto ciò, è possibile creare un nuovo *cockpit* secondo due differenti modalità, a seconda che si ricopra il ruolo di amministratore o il ruolo di utente finale.

Nel caso in cui si ricopra il ruolo di amministratore è necessario selezionare l'area *Document Browser* e cliccare sull'icona raffigurante il simbolo del più; viceversa, selezionando l'area *Workspace* e andando su *Analysis*, è necessario cliccare sempre sull'icona con il simbolo del più.

In entrambi i casi, l'interfaccia iniziale del cockpit appena creato, visibile in **Figura 3.8**, è una pagina vuota.



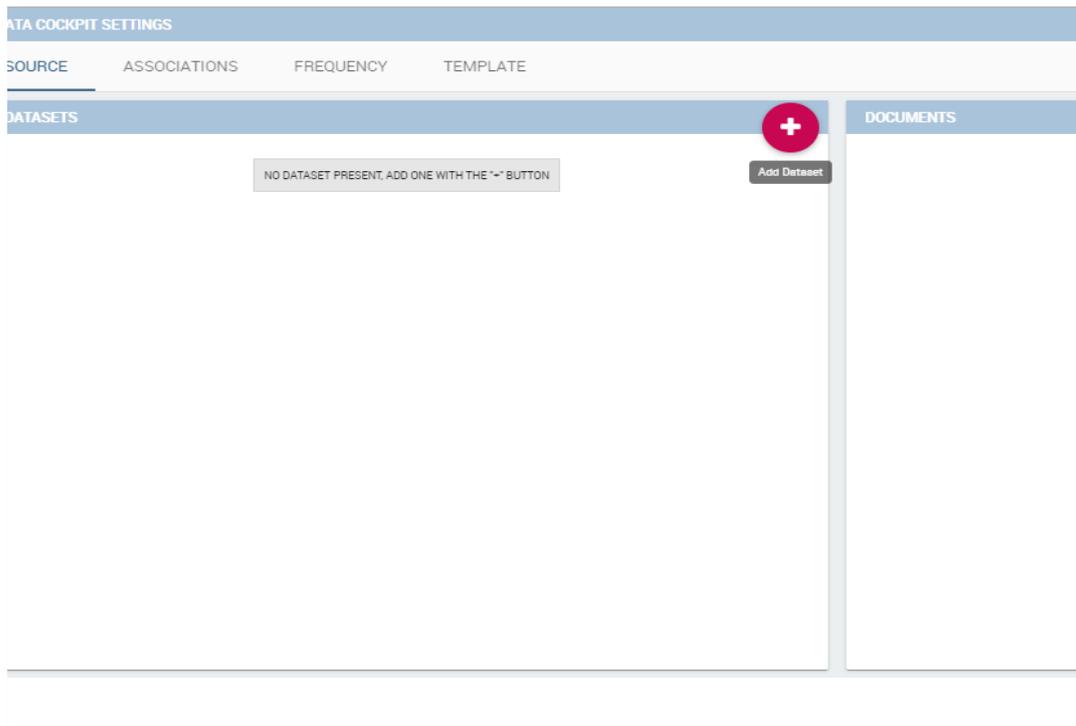
**Figura 3.8:** Interfaccia iniziale del *cockpit*

In alto a destra, cliccando sull'icona rossa e in ordine da sinistra verso destra, sono visualizzabili diverse opzioni, tra cui:

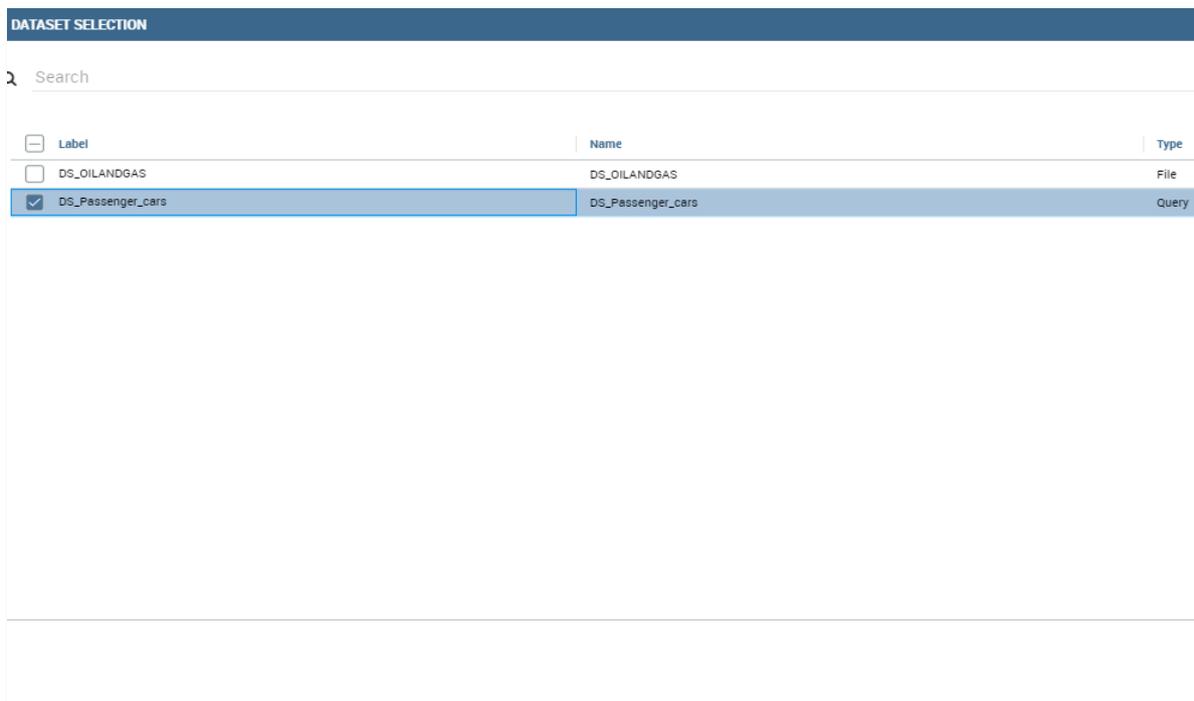
- **Clear Cache:** pulisce temporaneamente i dati;
- **Selections:** consente di aggiungere un *widget* per gestire le selezioni;
- **Data Configuration:** è l'opzione più significativa e permette di aggiungere e gestire uno o più *dataset* del cockpit e le relative, eventuali, associazioni;
- **General Configuration:** consente di gestire le opzioni generali del cockpit (nome, etichetta ecc.) e lo stile dei *widget* (titoli, bordi, intestazione colonne ecc.);
- **Add Widget:** dà la possibilità di aggiungere uno o più widget come testo, immagini, diagrammi, tabelle, documenti, html;
- **Save as:** effettua il salvataggio del documento.

Il primo step, dopo aver effettuato la creazione del *cockpit*, consiste nell'aggiunta di uno o più *dataset* al *cockpit* in questione selezionando l'opzione *Data Configuration*.

L'operazione appena descritta, prendendo a titolo di esempio il caso di studio, è visibile in **Figura 3.9** e in **Figura 3.10**.



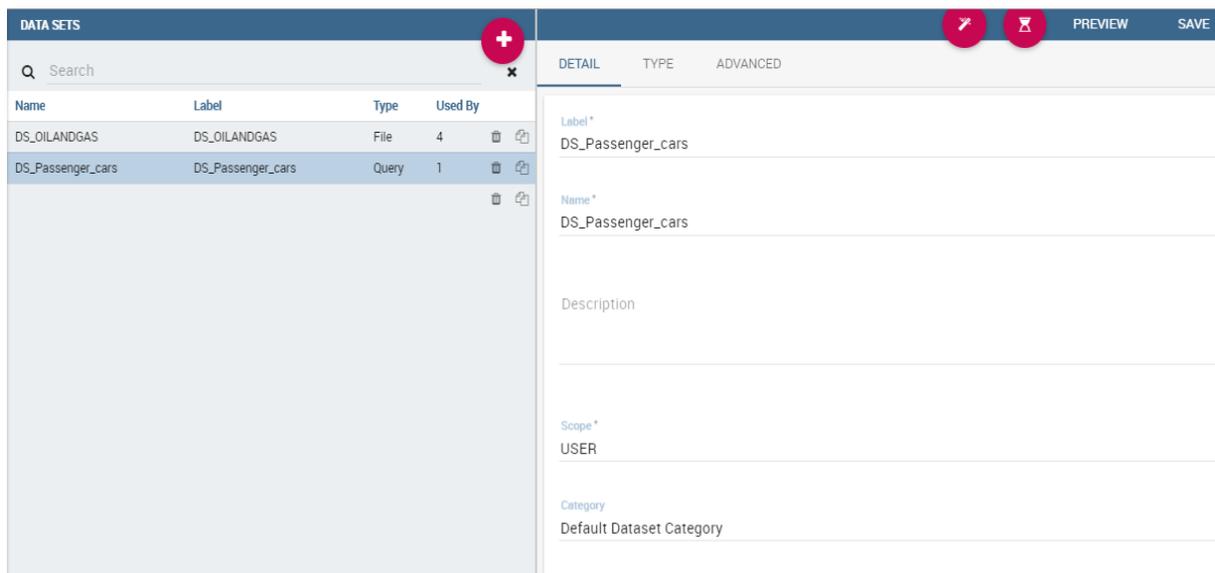
**Figura 3.9:** Aggiunta del dataset - 1



**Figura 3.10:** Aggiunta del dataset - 2

Tuttavia, per poter aggiungere uno o più *dataset* ad un determinato *cockpit* o documento, è necessario dapprima caricare il *dataset* su KNOWAGE selezionando l'opzione *Data set* nel pannello *Data Providers* del menu principale (rappresentato in **Figura 3.5**).

Quindi, la schermata che appare, visibile in **Figura 3.11**, consente di gestire i *dataset* già esistenti e di aggiungerne dei nuovi.



**Figura 3.11:** Interfaccia per la gestione *dataset*

Nel lato sinistro sono presenti tutti i *dataset* già esistenti e disponibili caratterizzati da nome, etichetta, tipologia e numero di documenti in cui il *dataset* in questione è in uso; inoltre, l'utente ha la possibilità, cliccando sulle due piccole icone sulla destra, di rimuovere o duplicare ciascun *dataset* presente.

Al contrario, cliccando sull'icona in alto a destra raffigurante il simbolo del più, è possibile aggiungere un nuovo *dataset* inserendo tutti i valori richiesti nei campi che appaiono sul lato destro della **Figura 3.11**.

Ogni pannello, presente sul lato destro, corrisponde ad uno specifico step del processo di definizione del nuovo *dataset*.

Tuttavia, nel primo pannello, denominato *DETAIL* e visibile in **Figura 3.11**, è necessario inserire etichetta, nome, un eventuale descrizione, l'ambito (*Scope*), e la categoria.

L'ambito consente di scegliere tra tre diverse opzioni a seconda dell'utilizzo che si intende fare del dataset: *user*, *technical*, *enterprise*.

Più in basso, anche se non visibile per problemi di spazio, è presente la sezione *OLDER DATASET VERSIONS* che riporta tutte le versioni modificate e salvate dello stesso *dataset* per favorirne l'archiviazione, l'accessibilità e l'utilizzo.

Nel secondo pannello, denominato *TYPE* e rappresentato in **Figura 3.12**, è necessario specificare il tipo di *dataset*, e, sulla base della tipologia scelta, scrivere, ad esempio, il codice della query in un apposito spazio, oppure caricare un file XLS.

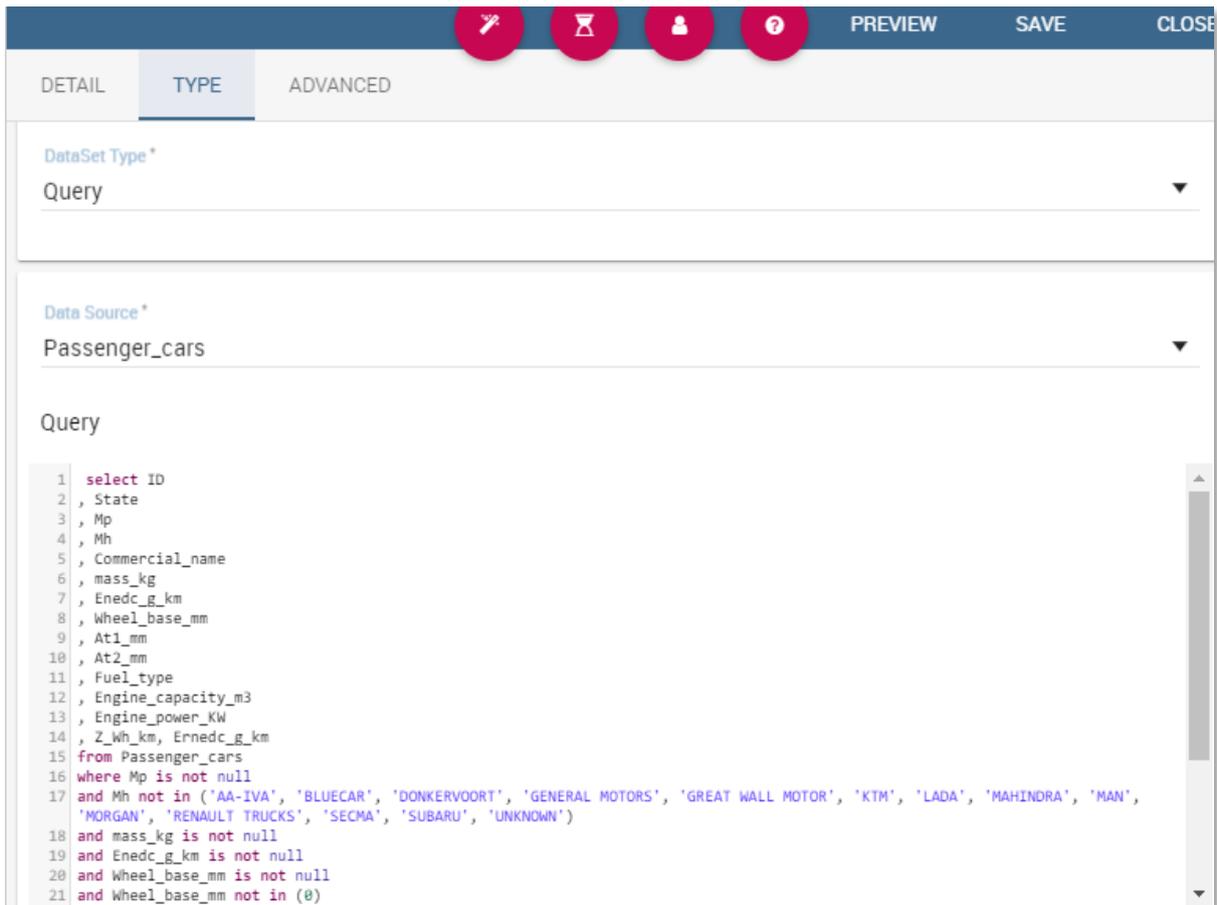


Figura 3.12: Pannello TYPE

Un *dataset* rappresenta la fonte dei dati di un documento analitico su cui costruire le analisi e KNOWAGE ne supporta diverse tipologie, tra cui<sup>[14]</sup>:

- **File;**
- **Query;**
- **Java Class;**
- **Script;**
- **Qbe (query by example);**
- **Custom;**
- **Flat;**
- **Ckan;**
- **Federated;**
- **REST;**
- **Big Data.**

Tutti i tipi di *dataset* appena citati condividono qualche operazione, mentre altre sono specifiche solo per alcune tipologie di dataset.e

Ad esempio, un *dataset* di tipo *File* richiede il caricamento di un file in formato XLS o CSV con la possibilità di verificare l’anteprima dei dati e definire i metadati (misura o attributo) per ciascuna colonna.

Al contrario, un *dataset* di tipo *Query*, utilizzato per la costruzione dei cruscotti e mostrato in **Figura 3.12**, richiede innanzitutto la definizione del *Data source* (mostrata nel quarto ed ultimo capitolo) per il collegamento tra KNOWAGE e il *Database* esterno con i dati, così come la scrittura della specifica *query* SQL nell'apposito spazio dedicato.

Le restanti tipologie di *dataset* non vengono spiegate per semplicità, dal momento che, quelle prese in considerazione, durante le attività in azienda, sono state quelle appena descritte.

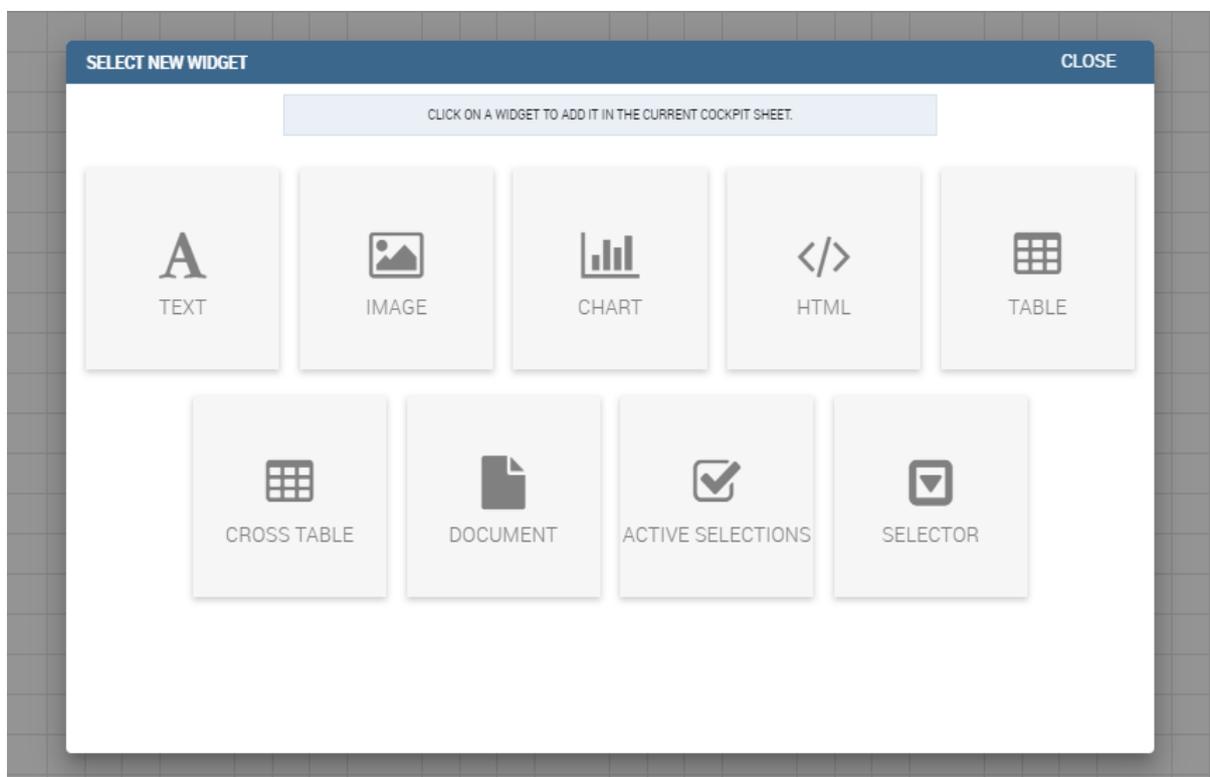
Nell'ultimo pannello, denominato *ADVANCED*, l'utente può applicare ai risultati del *dataset* la cosiddetta *PIVOT\_TRANSFORMATION* oppure selezionare l'opzione *Persist*.

Una volta terminate tutte queste operazioni è possibile accedere, cliccando sul bottone in alto a destra della **Figura 3.12**, ad una *preview* dei dati selezionati rappresentanti il *dataset*; questo risulta di fondamentale importanza in quanto, prima di associare il *dataset* al documento analitico, permette la verifica di eventuali errori nel codice della *query* e la successiva correzione.

Infine, è possibile procedere all'eventuale salvataggio dei dati selezionati.

Pertanto, concluso il processo di caricamento del *dataset* su KNOWAGE, quest'ultimo è possibile associarlo ad un generico *cockpit* (creato e raffigurato in **Figura 3.8**) selezionando, nell'interfaccia iniziale, l'opzione *Data Configuration* e procedere come mostrato nella **Figura 3.9** e nella **Figura 3.10**.

Quindi, cliccando sul bottone *Add Widget* nell'interfaccia iniziale del *cockpit*, è possibile aggiungere, al documento analitico, uno o più *widget* tra quelli presenti in **Figura 3.13**.



**Figura 3.13:** Aggiunta dei widget al *cockpit*

Un *widget* può contenere un carattere testuale, un'immagine, un grafico, un *html*, una tabella, una tabella incrociata, un documento o dei selettori:

- **Text widget:** consente di aggiungere un testo customizzato al *cockpit* grazie al pannello *text editor* con la possibilità di modificare a proprio piacimento lo stile (nel pannello *style* disponibile in tutti i *widget*) e di associare il *dataset* in modo tale da aggiungere uno o più valori al testo;
- **Image widget:** permette di aggiungere una o più immagini al documento selezionandole dalla galleria o caricandole dal pc;
- **Chart widget:** innanzitutto, consente di scegliere, nel pannello *Chart engine designer*, se inserire un diagramma a barre, un grafico a torta, un grafico a punti e altri ancora. Fatto ciò, è possibile specificare gli attributi da rappresentare e le misure su cui effettuare le operazioni di aggregazione (es. somma, media, conteggio, valore minimo, valore massimo), così come i filtri sui valori assunti dai dati selezionati;
- **HTML widget:** questo *widget* permette di definire un codice html e css customizzati in modo tale da aggiungere elementi dinamici al *cockpit*; nel menu a tendina presente all'interno del *widget* sulla destra, è possibile prendere i tag disponibili e copiarci, all'interno, il codice;
- **Table widget:** permette di definire una tabella contenente un massimo numero di righe per pagina selezionando i campi desiderati, eventualmente i campi calcolati, e scegliendo un attributo su cui effettuare la group by aggregando le misure desiderate; per ogni colonna è possibile scegliere il tipo di aggregazione e la misura unica su cui effettuare l'ordinamento (ascendente o discendente) delle righe. Nel pannello *Style* è possibile modificare lo stile delle righe, la griglia, i titoli delle colonne, il titolo del *widget*, i bordi;
- **Cross table widget:** simile al *widget* precedente, in questo caso è necessario inserire sotto la sezione colonne i campi che si vuole far apparire sulle colonne, lo stesso vale per le righe; nelle misure è necessario inserire campi numerici e aggregarli come si desidera. Nel pannello *Configuration* è possibile scegliere se far visualizzare i totali (su righe e/o colonne) così come le percentuali;
- **Document widget:** questo widget permette di aggiungere un documento esterno al cockpit riguardante report, grafici, mappe; nel pannello *place* è possibile caricare il documento, mentre nel pannello *Style* l'utente può modificarne lo stile;
- **Selector widget:** permette di aggiungere al *cockpit* un menu (a singolo valore o multivalore) a tendina o nella modalità *combobox* in cui poter selezionare uno o più valori; in questo modo i widget presenti si aggiornano automaticamente sui valori selezionati.

## 4. Architettura analisi e risultati

### 4.1 Descrizione di dettaglio del dataset utilizzato

In **Tabella 4.1** viene riportata una **descrizione** sintetica della **struttura dei dati** di partenza, fornendo una visione globale più chiara.

Tuttavia, da sinistra verso destra, si può osservare:

- Il nome dell'attributo o della misura;
- Una breve definizione in italiano;
- Il tipo di dato in questione.

**Tabella 4.1:** Struttura dei dati di partenza

<b>ID</b>	Identificativo	Integer
<b>State</b>	Stato	Varchar
<b>Mp</b>	Gruppo automobilistico	Varchar
<b>VFN</b>	Numero identificativo della famiglia del veicolo	Varchar
<b>Mh</b>	Brand/casa costruttrice	Varchar
<b>Man</b>	Altra definizione di casa costruttrice (ridondanza)	Varchar
<b>MMS</b>	Altra definizione di casa costruttrice (ridondanza)	Varchar
<b>TAN</b>	Numero approvazione	Varchar
<b>Type</b>	Tipologia	Varchar
<b>Variant</b>	Variante	Varchar
<b>Version</b>	Versione	Varchar
<b>Make</b>	Altra definizione di casa costruttrice (ridondanza)	Varchar
<b>Commercial_name</b>	Nome commerciale autovettura	Varchar
<b>Ct</b>	Categoria del tipo di veicolo approvato	Varchar
<b>Cr</b>	Categoria del tipo di veicolo registrato	Varchar

<b>mass (kg)</b>	Massa del veicolo	Integer
<b>Mt</b>	WLTP test	Integer
<b>Enedc (g/km)</b>	Emissioni CO2	Integer
<b>Ewltp (g/km)</b>	Riduzioni di emissioni CO2 (WLTP)	Integer
<b>Wheel_base (mm)</b>	Interasse	Integer
<b>At1 (mm)</b>	Ampiezza asse1	Integer
<b>At2 (mm)</b>	Ampiezza asse2	Integer
<b>Fue_type</b>	Tipo alimentazione	Varchar
<b>Fuel_mode</b>	Modalità alimentazione	Varchar
<b>Engine_capacity (cm3)</b>	Capacità motore	Integer
<b>Engine_power (KW)</b>	Potenza motore	Integer
<b>z (Wh/km)</b>	Consumo energia elettrica	Integer
<b>IT</b>	Tecnologie innovative	Varchar
<b>Ernedc (g/km)</b>	Riduzione emissioni grazie all'utilizzo di nuove tecnologie	Float
<b>Erwltp (g/km) (eliminato)</b>	Altra definizione di riduzione emissioni (ridondanza)	Float
<b>De (eliminato)</b>	Fattore di variazione	Float
<b>Vf (eliminato)</b>	Fattore di verifica	Integer
<b>R</b>	Nuove registrazioni	Integer

Preliminarmente al caricamento dei dati sul *Database* e alle tecniche di preparazione degli stessi dati, si è proceduto all'eliminazione di tutti quei campi non comprensibili, non codificabili e, di conseguenza, di sicuro inutilizzo.

Tra questi si ha: VFN, TAN, Ct, Cr, Mt, Ewltp, Erwltp, De, Vf.

Per avere una panoramica sul chiaro significato dei dati, nel seguito si ha una descrizione degli attributi e delle misure che caratterizzano il file dei dati (in formato CSV) sulle autovetture:

- **ID:** contiene un codice identificativo per ogni record (autovettura) del file;
- **State:** contiene la nazione di immatricolazione di ciascuna autovettura; ogni stato, nel file di partenza, è identificato da due soli caratteri.

Nel file sono presenti tutti i paesi dell'Unione Europea: AT (Austria), BE (Belgio), BG (Bulgaria), CY (Cipro), CZ (Repubblica ceca), DE (Germania),

DK (Danimarca), EE (Estonia), ES (Spagna), FI (Finlandia), FR (Francia), GB (Gran Bretagna), GR (Grecia), HR (Croazia), HU (Ungheria), IE (Irlanda), IT (Italia), LT (Lituania), LU (Lussemburgo), LV (Lettonia), MT (Malta), NL (Olanda), PL (Polonia), PT (Portogallo), RO (Romania), SE (Svezia), SI (Slovenia), SK (Repubblica slovacca);

- **Mp, Mh, Man, MMS, Make:** questi cinque attributi indicano il nome della casa automobilistica produttrice dell'auto secondo differenti denominazioni (come indicato nella tabella sopra).

Mp rappresenta il gruppo automobilistico a cui afferiscono una o più case costruttrici, mentre Mh identifica le diverse case costruttrici.

Man, MMS e Make rappresentano una ridondanza degli attributi Mp ed Mh ed inoltre contengono differenti denominazioni delle stesse case automobilistiche;

- **Type:** descrive la tipologia dell'automobile e contiene circa 1000 valori differenti costituiti da un insieme di lettere e numeri non codificabili in un dominio di valori con un senso logico.

Va detto che, a puro scopo informativo, le autovetture sono caratterizzate da tre volumi, detti anche vani: l'abitacolo (per conducente e passeggeri), la zona destinata ai bagagli e il vano del motore.

La tipologia di un'autovettura può essere: berlina (a tre volumi), due volumi, monovolume, station wagon, coupé, decappottabile, spyder (roadster) o veicolo multiuso (monovolume grande);

- **Variant:** contiene circa 5000 valori diversi costituiti da un insieme di lettere e numeri senza alcun significato preciso;
- **Version:** contiene circa 13000 valori differenti e identifica ciascuna autovettura mediante un codice alfanumerico (di difficile interpretazione);
- **Make:** l'attributo rappresenta il produttore dell'autovettura che è già specificato negli altri 4 campi descritti precedentemente;
- **Commercial\_name:** rappresenta il nome dell'autovettura assegnatole dalla propria casa automobilistica; contiene, in totale, 5000 diversi valori e nomenclature diverse per una stessa automobile;
- **Mass\_kg:** questa misura riporta il peso dell'automobile (in kilogrammi) che nel file dei dati varia dai 400 ai 3700 kg circa;
- **Enece\_g\_km:** questa misura descrive le emissioni di CO<sub>2</sub> per ogni autovettura, sulla base del rapporto g/km.

Le auto elettriche presentano un valore di emissione di CO<sub>2</sub> pari a zero.

I valori presenti in questa misura variano da 0 a 500 (g/km) circa e sono presenti 500 valori nulli;

- **Wheel\_base\_mm:** questo valore, misurato in millimetri, rappresenta l'interasse o passo delle autovetture, ossia la distanza che intercorre tra l'asse di una ruota anteriore e l'asse della ruota posteriore posta sullo stesso lato.

È un elemento che influisce sul comportamento dinamico di una vettura ma deve essere preso in considerazione anche con le masse (presenti nel veicolo) e la loro distribuzione.

Vantaggi di un passo corto: generalmente, accorciando il passo di un'autovettura, si ottiene sia una maggiore maneggevolezza /agilità nella guida (più velocità e facilità di gestione nei cambi di direzione), sia un maggiore trasferimento di peso (facilità di trasferimento della forza delle masse del mezzo).

Vantaggi di un passo lungo: al contrario, aumentando il passo di un'autovettura, si ha una maggiore stabilità, quindi miglior bilanciamento e minori

sbandate/ribaltamenti in curve effettuate ad alta velocità e in condizioni di accelerazioni e frenate violente.

Si ottiene, inoltre, anche una maggior abitabilità del veicolo, che si traduce in uno spazio maggiore per i passeggeri oppure in un incremento del numero dei passeggeri.

I valori di questa misura variano da 500mm a 5000mm circa ed i valori nulli sono un numero molto piccolo;

- **At1, At2:** questi due attributi descrivono l'ampiezza assale dello sterzo (in millimetri).

Dal momento che si tratta di una caratteristica strutturale molto specifica dell'autovettura i valori sono di difficile interpretazione.

Sono presenti molti valori nulli in entrambi gli attributi;

- **Fuel\_type:** questo attributo descrive l'alimentazione del motore.

L'impianto d'alimentazione è quella parte dei motori termici che serve per l'introduzione della carica fresca nel cilindro del motore; esso lavora accoppiato al sistema di distribuzione che comanda in funzione delle esigenze di motricità del veicolo.

L'attributo contiene 7 valori diversi corrispondenti a 7 diverse tipologie di alimentazione: benzina, diesel, ibrido con il motore termico a benzina o diesel, gas (GPL e metano), elettrico.

La quasi totalità delle auto risulta essere alimentata da benzina e diesel, circa 14500 auto risultano essere elettriche e ibride, e circa 4000 a gas (metano e GPL).

I valori nulli sono quasi del tutto assenti; nel seguito si ha una descrizione delle varie tipologie di alimentazione delle autovetture:

- Benzina: oggi in commercio si trovano le benzine verdi, ottenute grazie all'aggiunta di pigmenti colorati; il piombo è stato sostituito con altre sostanze.
  - Diesel: ha un costo inferiore alla benzina e dei consumi ridotti; l'acquisto di un'auto a diesel è più caro ed il motore non si presta a brevi tratti. Inoltre, le prestazioni dell'auto risultano inferiori rispetto ad un'auto a benzina, decisamente più scattante.
  - Gas metano: costa il 60% in meno rispetto alla benzina; le prestazioni non sono per nulla paragonabili alle auto benzina o diesel ed è sconsigliato installarlo su auto con cilindrata inferiore a 1400 cc.
  - Gas GPL (gas di petrolio liquefatto): il GPL non è una fonte rinnovabile (è un derivato del petrolio); ha un costo inferiore alla benzina di circa il 40% e fornisce prestazioni migliori rispetto al gas metano ma comunque inferiori rispetto agli altri tipi di alimentazione.
  - Ibrido a benzina/diesel: un'auto ibrida possiede due motori, uno elettrico ed uno termico alimentato a benzina o diesel.  
La presenza del motore elettrico consente la riduzione dei consumi e delle emissioni inquinanti rispetto alle auto tradizionali.
  - Elettrico: un'automobile con motore elettrico immagazzina l'energia chimica nelle batterie ricaricabili e la trasferisce al motore sotto forma di energia elettrica.
- **Fuel\_mode:** l'attributo, che presenta quattro differenti valori, ossia B, E, F, M, risulta di difficile interpretazione.

Infatti, molte ricerche effettuate sul significato dell'attributo non hanno prodotto alcun effetto.

Da notare che la quasi totalità dei record sono caratterizzati da un unico valore, la lettera M; i valori nulli, in questo caso, non sono presenti;

- **Engine\_capacity\_cm3:** descrive la capacità del motore o, più comunemente, **cilindrata** (in  $\text{cm}^3$ ).

Spesso si confonde questo termine con i cavalli (CV), e non è detto che due vetture di pari **cilindrata** abbiano anche pari potenza; infatti, il fatto che due automobili di pari **cilindrata** possano avere prestazioni differenti dipende proprio dai cavalli che i motori riescono ad esprimere.

A questo proposito, nei listini dei vari costruttori, si trovano spesso vetture dotate di motore con uguale cilindrata ma differente potenza.

La capacità del motore o cilindrata, misurata in centimetri cubici ( $\text{cm}^3$ ), indica il volume o lo spazio disponibile in ogni cilindro per l'entrata dell'aria e l'utilizzo del carburante per la combustione; pertanto indica lo spazio di uno dei più importanti elementi che permettono ad un motore a scoppio di funzionare.

Ovviamente, per ottenere la cilindrata totale di un'auto con motore a più cilindri è necessario calcolare quella unitaria riferita ad un singolo cilindro e moltiplicarla per il numero totale di cilindri.

In generale, la cilindrata di un'auto è utile a dare un'idea della dimensione e della potenza del motore anche se non esprime la potenza reale dello stesso, e rappresenta un parametro fondamentale per la valutazione delle prestazioni complessive dell'auto.

Inoltre, dal momento che la cilindrata influisce sulla potenza dell'auto deve essere adeguata al peso della stessa dal momento che il giusto rapporto tra questi due parametri favorisce la riduzione dei consumi.

Facendo un rapido esempio, tra un'auto diesel con un motore da 1.5 cc e 120 CV di potenza ed una con motore 1.9 cc e 100 CV non è detto che l'auto con 120 cavalli sia la più performante perché il motore 1.5 cc potrebbe essere sottodimensionato rispetto al peso dell'auto e quindi sia le prestazioni che i consumi ne risentirebbero notevolmente.

Tuttavia, negli ultimi tempi molti costruttori di auto stanno seguendo la strada del **downsizing**, spinti soprattutto dalle stringenti normative in termini di emissioni di CO<sub>2</sub>. Questo filone costruttivo prevede di dotare le vetture di motori sovralimentati con **cilindrata** ridotta in grado di fornire consumi minori e potenza maggiore rispetto ad un aspirato di pari **cilindrata**.

I valori del dominio oscillano tra 300 e 8000  $\text{cm}^3$  circa, con ben 300 diversi valori; circa 9000 record contengono il valore nullo;

- **Engine\_power\_KW:** questa misura, invece, descrive la potenza di cui dispone il motore (in Kilowatt).

Solitamente, la potenza di cui dispone il motore è descritta dal numero di cavalli (sigla CV: cavallo-vapore).

Tuttavia, il cavallo-vapore è una misura di consuetudine utilizzata per descrivere le capacità prestazionali dell'auto e la velocità massima.

Tuttavia, per conoscere la reale potenza, devono essere utilizzati i kilowatt.

In Inghilterra e negli Stati Uniti si utilizza l'HP (*Horse Power*) e l'equivalenza è: 1 CV = 0,98631.

Infine, 1 CV = 0,73549875 kW.

I valori del dominio sono compresi tra 5 e 1400 kw con 325 valori differenti e circa 22000 valori nulli;

- **Z\_Wh\_km**: questo campo descrive il consumo di energia elettrica al kilometro. Tale consumo viene misurato mediante il rapporto Wh/km ed è un valore presente solo per le auto elettriche e auto ibride con motore termico a benzina o diesel.  
I valori oscillano tra 1 e 400 Wh/km circa con 73 valori differenti; i record con un valore non nullo in tale campo sono circa 14000 (tutte le auto elettriche e ibride) mentre la restante parte dei record risulta ovviamente caratterizzata da un valore nullo;
- **IT**: rappresenta la tecnologia innovativa o il gruppo di tecnologie innovative utilizzate per l'autovettura. L'attributo contiene valori difficili da codificare presenti solo per una piccolissima quota dei record presenti;
- **Ernedc\_g\_km**: la misura descrive la riduzione di emissioni di CO2 (in g/km) dell'autovettura ottenuta grazie a tecnologie innovative.  
Circa 30000 record sono caratterizzati da un valore non nullo con valori compresi tra 1 e 3;
- **R**: descrive le nuove registrazioni per ogni autovettura.  
Pertanto, sia il significato della misura che i valori in questione risultano di difficile comprensione.

## 4.2 Caricamento dei dati sul Database

Nel presente paragrafo vengono descritte tutte le operazioni effettuate necessarie al **caricamento dei dati sul DB** in remoto.

Pertanto, viene spiegata la modalità di caricamento dei dati sul DB (in remoto) e viene motivato l'utilizzo di due *software* DBMS differenti; quindi, viene descritto ogni singolo *step* del caricamento dei dati sul DB in locale e, successivamente, in remoto.

### 4.2.1 Modalità di caricamento dati sul DB

Dopo aver eseguito un'analisi preliminare sui dati relativi alle autovetture si è proceduto con lo *step* successivo: l'inserimento dei dati del *file* (in formato CSV) all'interno di un Database in remoto.

Per fare questo si è resa necessaria l'installazione e l'utilizzo di 2 *software*: **MySQL Workbench** e **Oracle SQL Developer**.

Dapprima, su specifica indicazione della tutor aziendale, si è installato solamente il *software* MySQL Workbench ma si sono verificati diversi problemi nel caricamento dei dati; infatti, è risultato che il *software* non riusciva a gestire le righe con campi (o celle) vuoti relativi a tipi di dato numerici troncando completamente il record e, allo stesso modo, risultava molto lento nel caricamento dei record.

Ovviamente, nella creazione della tabella all'interno del DB, si è provato ad impostare un valore di default per i campi di tipo numerico, ma non si è avuto alcun riscontro positivo.

Quindi, su consiglio di un collega, si è proceduto all'installazione di Oracle SQL Developer ed in questo caso il caricamento dei dati del file (in formato CSV), all'interno di un DB creato in locale, è andato a buon fine in brevissimo tempo.

Però, essendo necessario caricare i dati nel DB in remoto come richiesto dall'azienda, e dato che questo non è possibile in SQL Developer, si è proceduto all'esportazione dei record della tabella in un file (di tipo SQL) per poi importarli nel DB in remoto in MySQL Workbench; in questo modo il caricamento dei dati è andato a buon fine.

L'altro metodo percorribile sarebbe stato quello di inserire manualmente il valore 0 sul *file* in formato CSV per i campi vuoti di tipo numerico; poi, mediante un'opportuna istruzione di aggiornamento (*update*) con il linguaggio SQL, sarebbe stato necessario convertire lo zero inserito con il valore *null*.

Ma il modo di operare scelto, appena descritto, ha permesso senz'altro di conoscere uno strumento in più (Oracle SQL Developer), di acquisire competenze nel suo utilizzo, così come la capacità di gestire dei dati esportandoli ed importandoli da un DB all'altro.

MySQL Workbench è un *software* DBMS ed uno strumento visuale per la progettazione e manutenzione di *Database* e per la gestione, modellazione dati tramite linguaggio SQL; la caratteristica distintiva consiste nella possibilità di accesso a *Database* remoti e nella loro gestione mediante diversi protocolli, tra cui SSH.

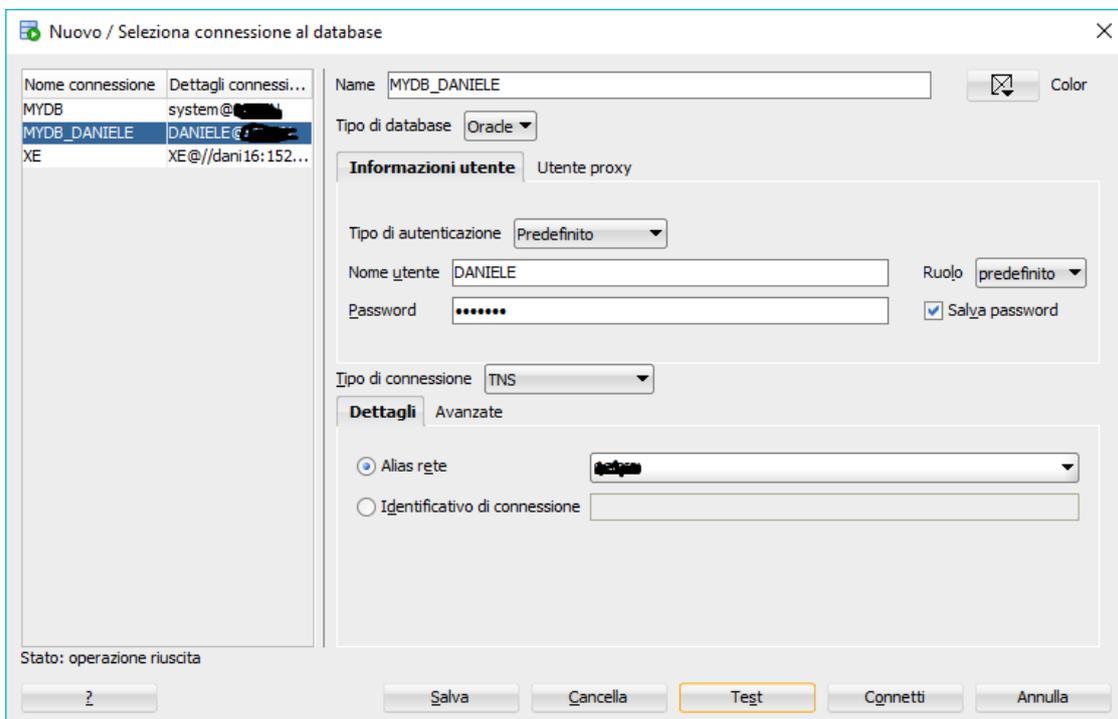
Oracle SQL Developer è un DBMS ed un ambiente di sviluppo integrato per operare nei *Database* Oracle con il linguaggio SQL; pertanto, anche questo è uno strumento visuale che permette di progettare *Database* così come memorizzare, gestire e interrogare grandi quantità di dati tramite linguaggio SQL.

Inoltre, supporta molti prodotti Oracle anche se, rispetto a MySQL Workbench, consente di lavorare solamente con *Database* locali o eventualmente remoti, ma specificatamente Oracle.

Nel seguito vengono riportate, in modo dettagliato, tutte le operazioni descritte sopra.

## 4.2.2 Caricamento dati sul DB (SQL Developer) in locale e successiva esportazione

Per prima cosa, all'interno di SQL Developer, come visibile in **Figura 4.1**, è stata creata una nuova connessione al *database* (in locale):

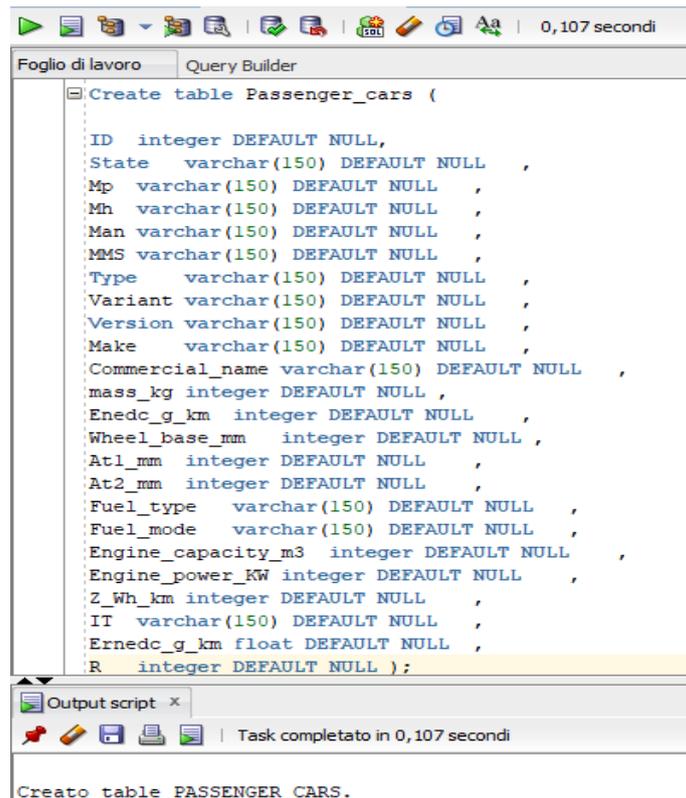


**Figura 4.1:** Creazione connessione al DB locale

Al nuovo *database* creato è stato assegnato il nome fittizio “MYDB\_DANIELE” e sono state inserite tutte le informazioni richieste nella figura sopra, di cui alcune non visualizzabili per privacy aziendale: tipo di *database*, nome utente, password, tipo di connessione e alias rete (nome fittizio definito nella procedura d’installazione del *software*).

Quindi, prima di procedere con il salvataggio, è stato effettuato il test della connessione al DB locale che ha dato esito positivo, come mostrato in basso a sinistra nella figura.

Subito dopo, all’interno del database “MYDB\_DANIELE”, è stata definita la tabella “Passenger\_cars” in cui caricare i dati, come mostrato in **Figura 4.2**.



**Figura 4.2:** Creazione tabella Passenger\_cars in locale

La creazione della tabella è andata a buon fine ed è stata inserita all’interno dello schema “MYDB\_DANIELE”.

Tuttavia, come si evince dalla struttura della tabella, mostrata in **Figura 4.3**, per ogni colonna caratterizzata dal tipo di dato “*varchar*”, nella definizione della tabella, si è scelto di inserire un limite massimo di 150 caratteri per cercare di limitare le perdite di dati.

Questo perché, inizialmente, si era inserito un limite massimo di caratteri pari a 50; in questo modo però, popolando la tabella, sono state caricate solo 113.000 tuple su un totale di 1048575 record.

Invece, per le colonne di tipo numerico, SQL Developer converte il tipo di dato “*integer*” in “*number*” con un limite massimo di cifre ammesse pari a 38.

Inoltre, per ogni colonna, è stato assegnato di default il valore “*null*”, così che, se per una qualsiasi colonna di una qualsiasi riga non si ha alcun valore, l’*editor* le assegna automaticamente il valore “*null*”.



Per effettuare un corretto popolamento della tabella è stato necessario caricare il *file*, originariamente in formato XLS, nel formato CSV.

Si noti che nell'opzione "delimitatore" si è dovuto selezionare il punto e virgola (;) per poter separare correttamente le colonne l'una dall'altra e garantire una corretta gestione dei dati.

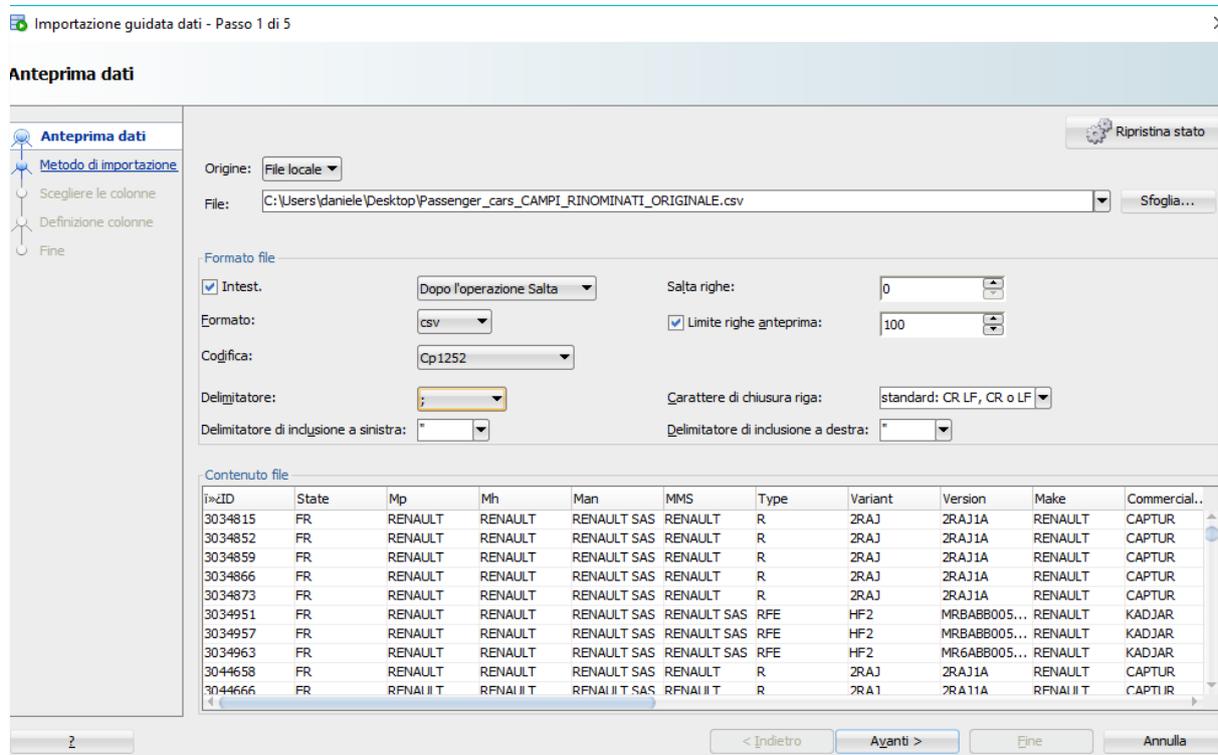


Figura 4.4: Importazione dati nel DB in locale – step 1

Tuttavia, per quanto riguarda il secondo *step*, quello relativo al metodo di importazione visibile in **Figura 4.5**, si è scelto di importare i dati mediante l'inserimento dello *script*.

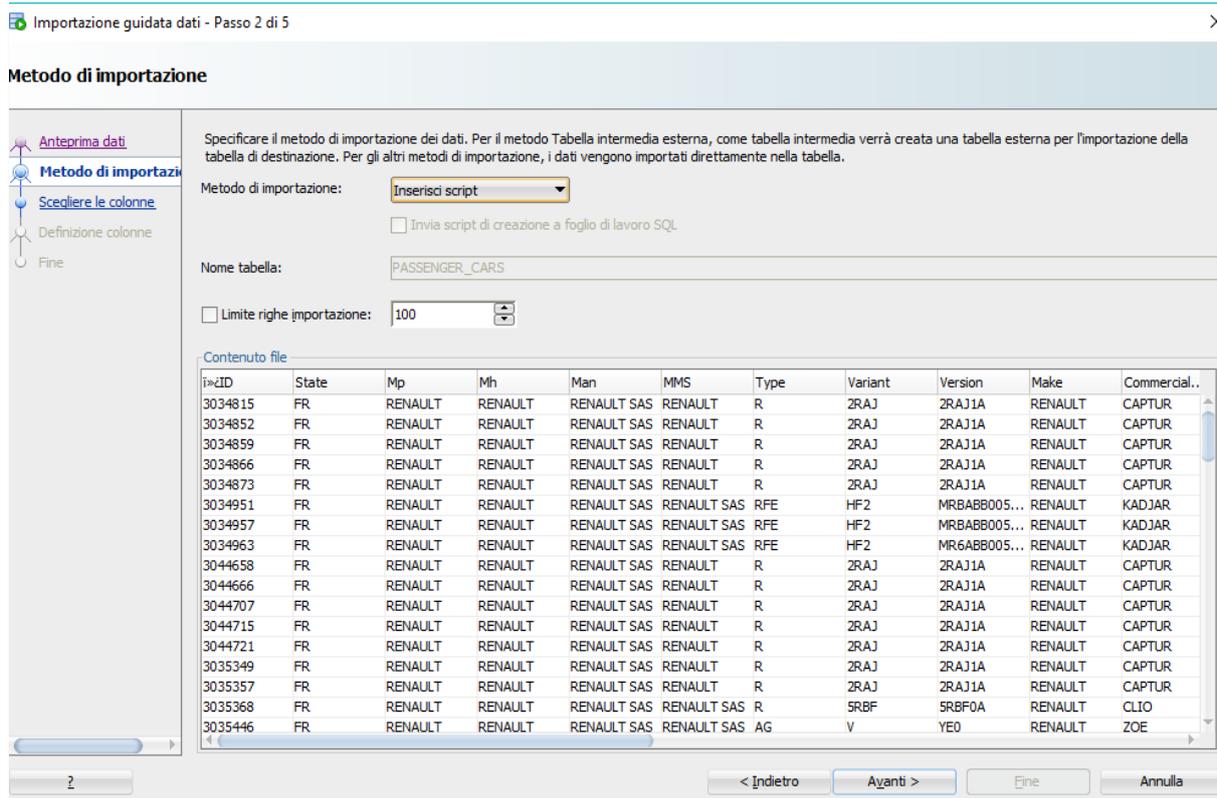


Figura 4.5: Importazione dati nel DB in locale – step 2

Quindi, secondo quanto riportato in **Figura 4.6**, per completezza sono state selezionate tutte le colonne, tra quelle disponibili nel *file* (CSV) in *input*.

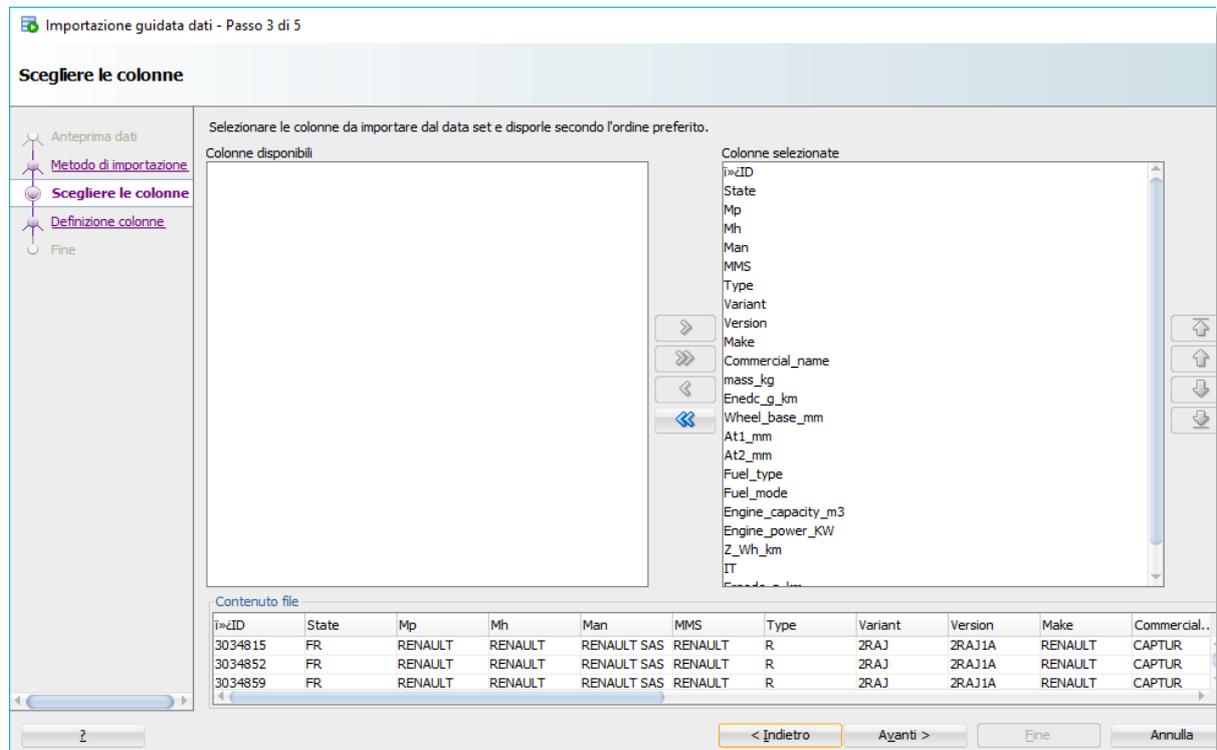
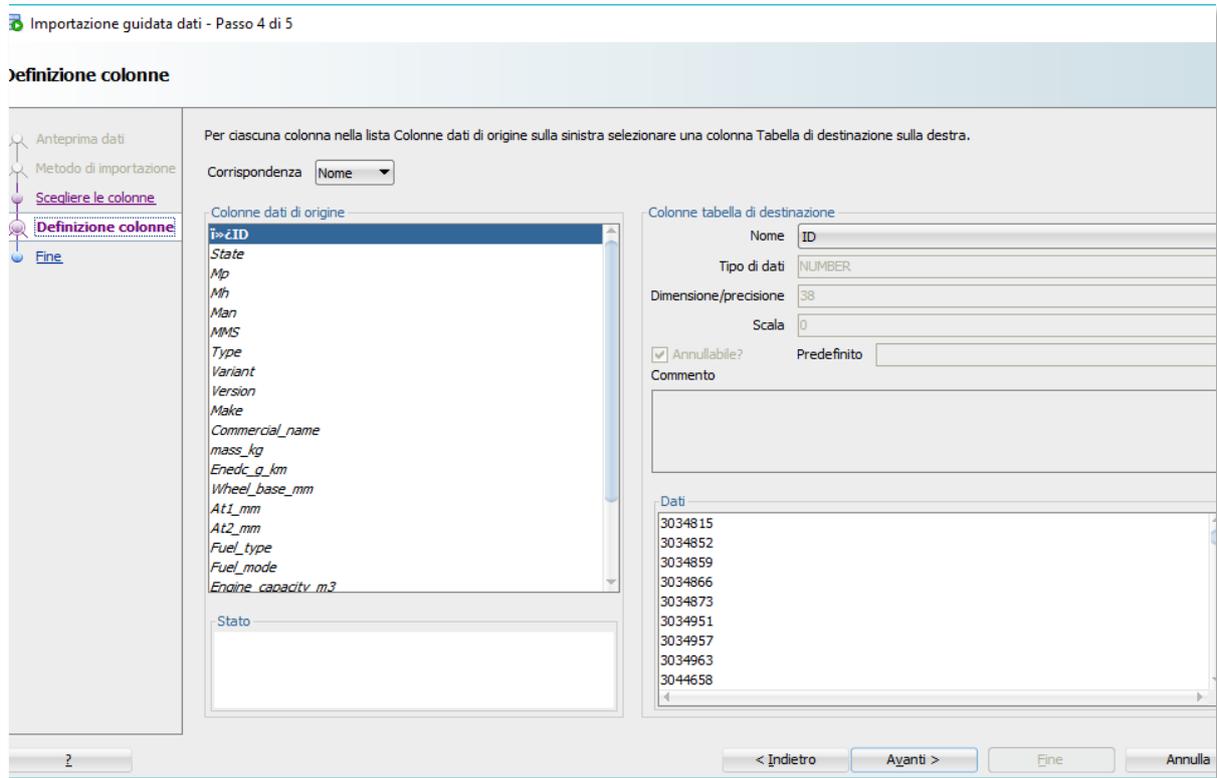


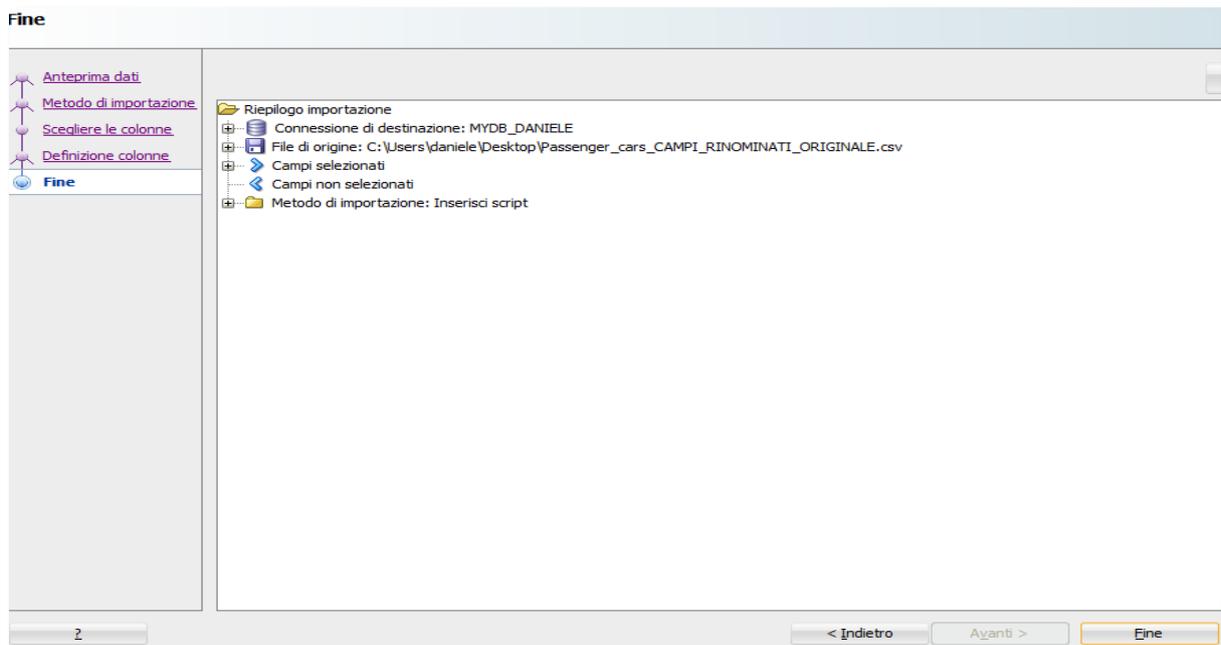
Figura 4.6: Importazione dati nel DB in locale – step 3

Il quarto *step*, invece, è visibile in **Figura 4.7**.



**Figura 4.7:** Importazione dati nel DB in locale – step 4

Infine, l'ultimo *step* del processo di importazione dei dati dalla sorgente alla tabella è schematizzato nella **Figura 4.8**.



**Figura 4.8:** Importazione dati nel DB in locale – step 5

Una volta terminato il popolamento della tabella in locale “Passenger\_cars”, come visibile in **Figura 4.9**, muovendo la *scrollbar* presente nell’*editor* è possibile leggere ed eventualmente eseguire operazioni su tutti i dati contenuti nella tabella.

ID	STATE	MP	MH	MAN	MMS	TYPE	VARIANT	VERSION	MAKE	COMMERCIAL_NAME	MASS_KG
1	3034815	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
2	3034852	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
3	3034859	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
4	3034866	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
5	3034873	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
6	3034951	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT SAS	RFE	HF2	MR6ABB005000	RENAULT KADJAR	1448
7	3034957	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT SAS	RFE	HF2	MR6ABB005000	RENAULT KADJAR	1448
8	3034963	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT SAS	RFE	HF2	MR6ABB005000	RENAULT KADJAR	1395
9	3044658	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
10	3044666	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
11	3044707	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
12	3044715	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
13	3044721	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
14	3035349	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
15	3035357	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
16	3035368	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT SAS	R	5RBF	5RBF0A	RENAULT CLIO	1165
17	3035446	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT SAS	AG	V	YE0	RENAULT ZOE	1555
18	3035452	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT SAS	AG	V	YE0	RENAULT ZOE	1555
19	3035522	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
20	3035529	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
21	3035536	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
22	3035544	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT SAS	R	2RAJ	2RAJ1A	RENAULT CAPTUR	1253
23	3035628	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT SAS	R	5RBF	5RBF0A	RENAULT CLIO	1165
24	3035636	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT SAS	R	5RBF	5RBF0A	RENAULT CLIO	1165
25	3035643	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT SAS	R	5RBF	5RBF0A	RENAULT CLIO	1165
26	3035650	FR	RENAULT	RENAULT	RENAULT SAS	RENAULT SAS	R	5RBF	5RBF0A	RENAULT CLIO	1165

**Figura 4.9:** Lettura dati nel DB in locale

La *query*, riportata in **Figura 4.10**, consente di ottenere il numero totale di record contenuti nella tabella “Passenger\_cars”.

```

select COUNT(*) from passenger_cars;

```

COUNT(*)
1048575

**Figura 4.2.10:** Query numero righe totali

Tuttavia, l'output della *query*, che rappresenta il numero totale di righe della tabella (in locale), coincide, come ci si aspettava, con le righe del *file* (CSV) in *input*: 1.048.575.

In **Figura 4.11** e in **Figura 4.12** sono riassunte alcune interessanti statistiche del *dataset* (raccolte attraverso SQL Developer al termine del caricamento dei dati sul DB).

Tra le più rilevanti, per ogni colonna, figurano:

- Numero di valori differenti (“NUM\_DISTINCT”).
- Valore più alto (“HIGH VALUE”).
- Numero di valori nulli (“NUM\_NULLS”).

COLUMN_NAME	NUM_DISTINCT	LOW_VALUE	HIGH_VALUE	DENSITY	NUM_NULLS	NUM_BUCKETS	LA
1 ID	1042752	C104	C40560385E	0,000000959000797888664	0		1 07-I
2 STATE	28 4154		534B	0,0357142857142857	0		1 07-I
3 MP	13 424D572047524F5550		56572047524F5550205043	0,0769230769230769	283394		1 07-I
4 MH	71 41412D495641		564F4C564F	0,0140845070422535	0		1 07-I
5 MAN	76 41412D495641		6475706C6963617465	0,0131578947368421	0		1 07-I
6 MMS	388 41412D495641		C384C2B9C382C2A04B4F4441204155544F204153	0,00257731958762886	58552		1 07-I
7 TYPE	1097 30		5A52	0,000911577028258888	377		1 07-I
8 VARIANT	4774 234244		6A35666B	0,000209467951403435	1777		1 07-I
9 VERSION	12995 2A		67	0,0000769526741054252	3893		1 07-I
10 MAKE	116 414C4641		C38EE280994D57	0,00862068965517241	1360		1 07-I
11 COMMERCIAL_NAME	5146 2D2D		756E6B6E6F776E	0,000194325689856199	260		1 07-I
12 MASS_KG	1656 C20515		C22556	0,000603864734299517	57		1 07-I
13 ENEDC_G_RM	332 80		C20628	0,00301204819277108	563		1 07-I
14 WHEEL_BASE_MM	457 80		C234	0,00218818380743982	36		1 07-I
15 AT1_MM	315 C20B33		C21D24	0,00317460317460317	2155		1 07-I
16 AT2_MM	323 C20239		C215	0,00309597523219814	49978		1 07-I
17 FUEL_TYPE	23 44494553454C		706574726F6C	0,0434782608695652	39		1 07-I
18 FUEL_MODE	4 42		4D		0,25	2	1 07-I
19 ENGINE_CAPACITY_M3	306 C20363		C2505E	0,00326797385620915	8745		1 07-I

**Figura 4.11:** Statistiche dataset - 1

COLUMN_NAME	NUM_DISTINCT	LOW_VALUE	HIGH_VALUE	DENSITY	NUM_NULLS	NUM_BUCKETS	LA
6 MMS	388 41412D495641		C384C2B9C382C2A04B4F4441204155544F204153	0,00257731958762886	58552		1 07-I
7 TYPE	1097 30		5A52	0,000911577028258888	377		1 07-I
8 VARIANT	4774 234244		6A35666B	0,000209467951403435	1777		1 07-I
9 VERSION	12995 2A		67	0,0000769526741054252	3893		1 07-I
10 MAKE	116 414C4641		C38EE280994D57	0,00862068965517241	1360		1 07-I
11 COMMERCIAL_NAME	5146 2D2D		756E6B6E6F776E	0,000194325689856199	260		1 07-I
12 MASS_KG	1656 C20515		C22556	0,000603864734299517	57		1 07-I
13 ENEDC_G_RM	332 80		C20628	0,00301204819277108	563		1 07-I
14 WHEEL_BASE_MM	457 80		C234	0,00218818380743982	36		1 07-I
15 AT1_MM	315 C20B33		C21D24	0,00317460317460317	2155		1 07-I
16 AT2_MM	323 C20239		C215	0,00309597523219814	49978		1 07-I
17 FUEL_TYPE	23 44494553454C		706574726F6C	0,0434782608695652	39		1 07-I
18 FUEL_MODE	4 42		4D		0,25	2	1 07-I
19 ENGINE_CAPACITY_M3	306 C20363		C2505E	0,00326797385620915	8745		1 07-I
20 ENGINE_POWER_KW	325 C106		C20F3E	0,00307692307692308	21801		1 07-I
21 Z_WH_RM	73 C102		C2050E	0,0136986301369863	1034981		1 07-I
22 IT	58 3139		6538203139	0,0172413793103448	1011694		1 07-I
23 ERNEDC_G_RM	16 C102		C120		0,0625	1018106	1 07-I
24 R	1004 C102		C3023D50	0,00099601593625498	0		1 07-I

**Figura 4.12:** Statistiche dataset - 2

Infine, nella **Figura 4.13**, **Figura 4.14** e **Figura 4.15**, viene descritta l'esportazione guidata di tutti i record della tabella (in locale) “*Passenger\_cars*” in un file di tipo SQL.

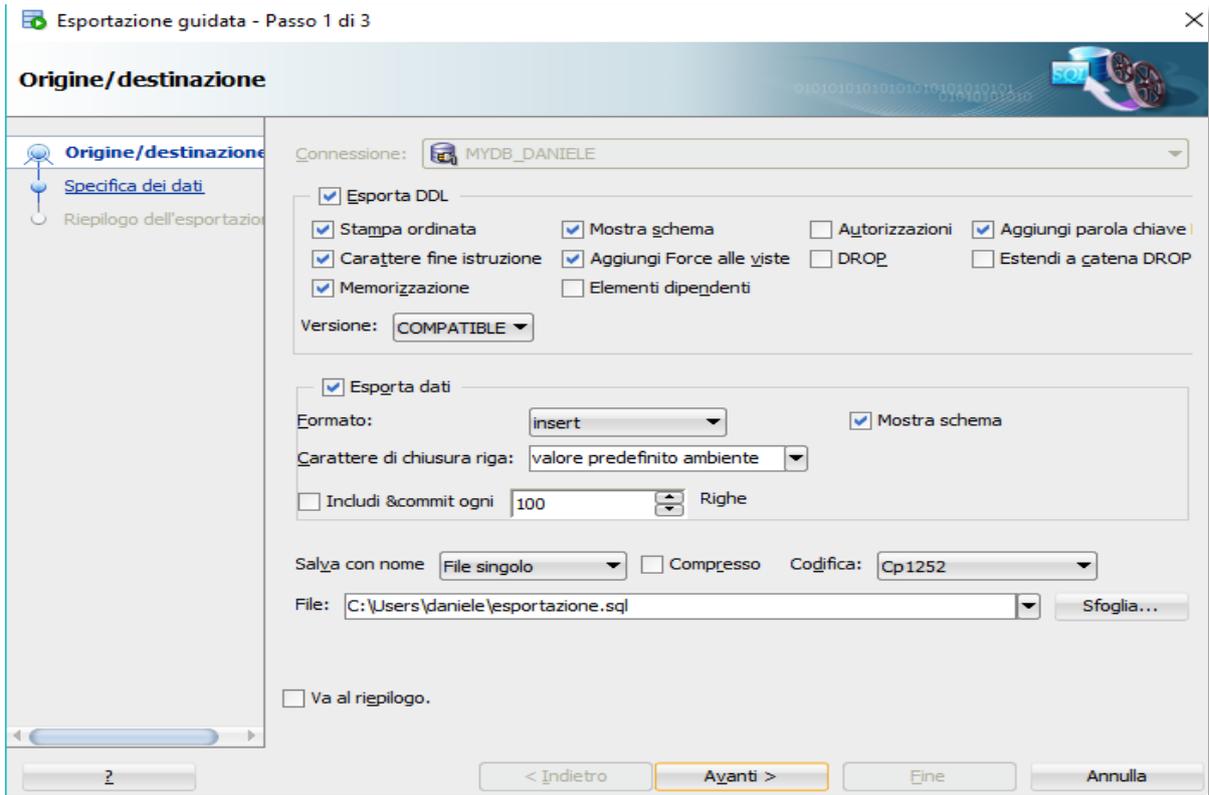


Figura 4.13: Esportazione guidata dei record – step 1

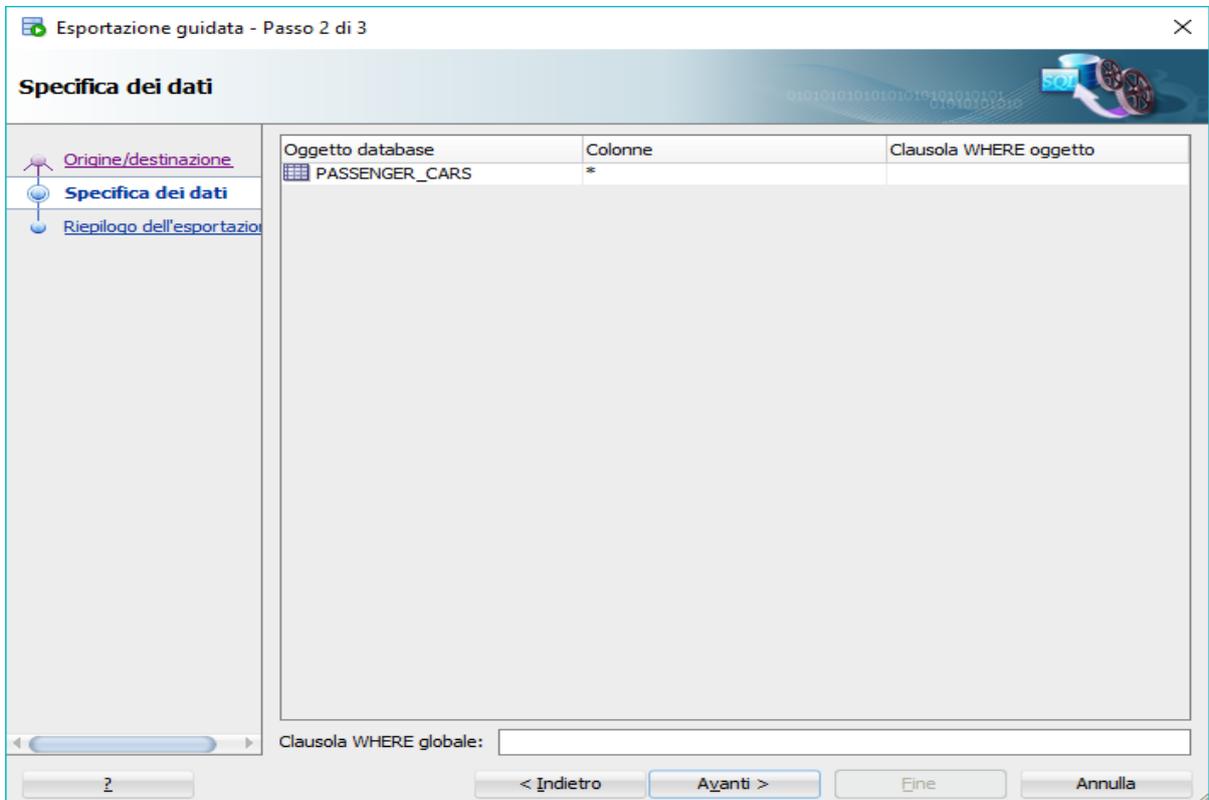


Figura 4.14: Esportazione guidata dei record – step 2

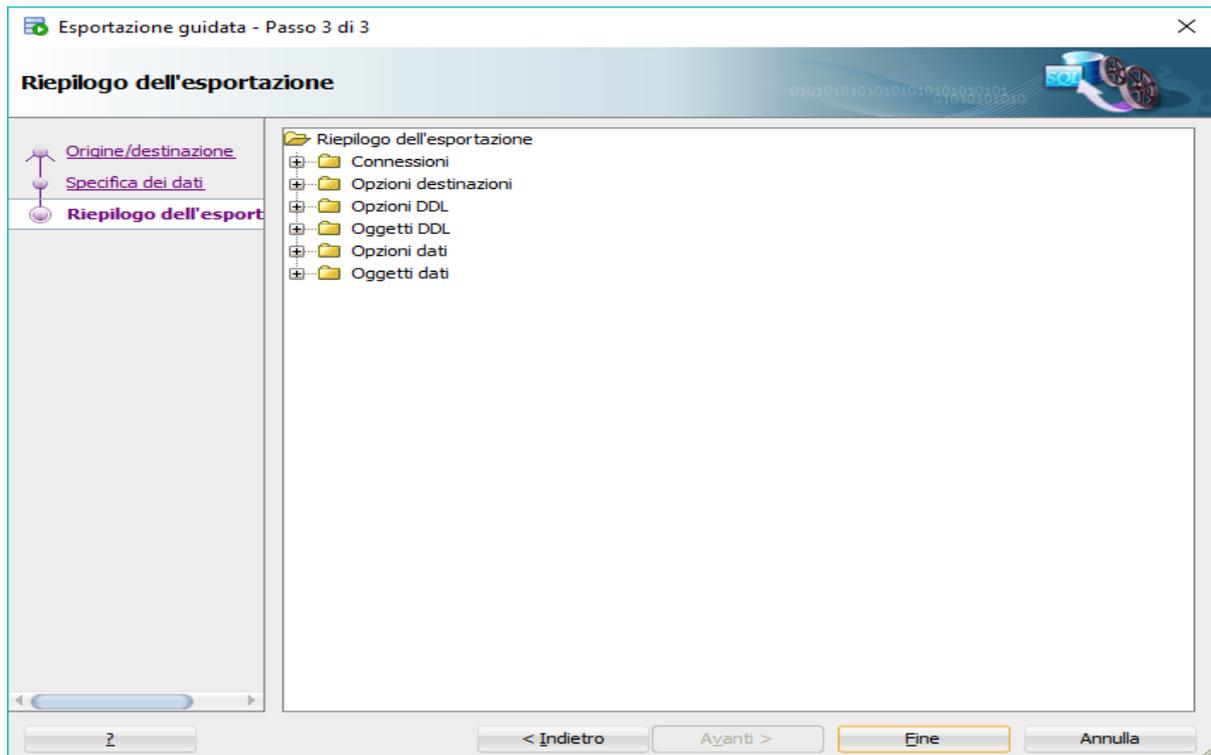


Figura 4.15: Esportazione guidata dei record – step 3

### 4.2.3 Caricamento dati sul DB (My SQL Workbench) in remoto

Conclusa l'esportazione dei record nel *file* di tipo SQL, per poter caricare correttamente tutti i dati all'interno del DB in remoto su MySQL Workbench, sono state effettuate, anche in questo caso, una serie di operazioni descritte nel seguito.

Inizialmente, come visibile in **Figura 4.16**, è stata definita la nuova connessione al **Database in remoto**:

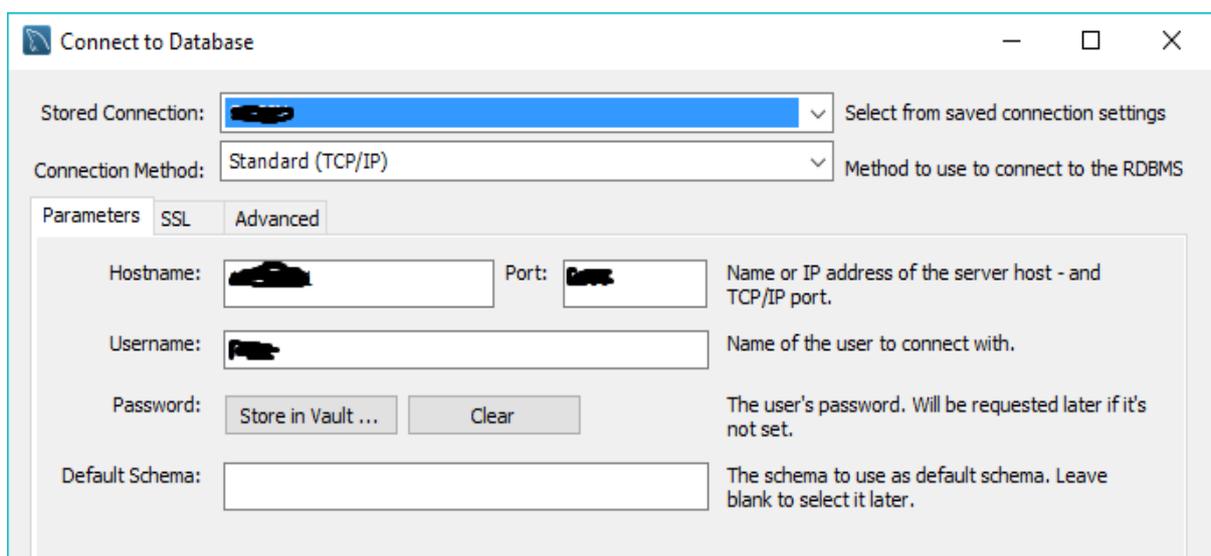


Figura 4.16: Creazione connessione al DB in remoto

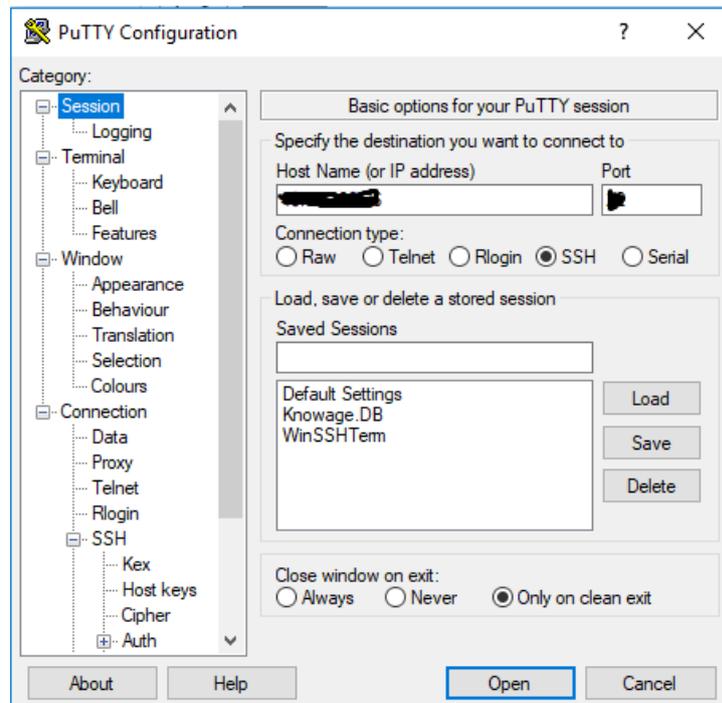
Su indicazione specifica, il nome del *Database* in remoto, già esistente e utilizzato da altri colleghi per progetti aziendali, è stato oscurato per la privacy aziendale; poi, sono stati definiti e oscurati, sempre su specifica indicazione, l'*host*, la porta, lo username e una password.

Quindi, per poter stabilire la connessione al *Database* in remoto, si è resa necessaria anche l'effettuazione di una procedura di connessione al *server* remoto mediante il *client* denominato "PuTTY".

"PuTTY" è un *client SSH, Telnet e rlogin* combinato con un emulatore di terminale per la gestione in remoto di sistemi informatici (es. computer, *server* ecc..).



- 1) Inizialmente, una volta aperta l'interfaccia del "PuTTY" rappresentata in **Figura 4.17**, si è inserito l'indirizzo IP fornito e, anche questo, è stato oscurato per la privacy:



**Figura 4.17:** Interfaccia iniziale PuTTY

- 2) Quindi, cliccando su "*Tunnels*", nella sezione *SSH*, si genera la schermata in **Figura 4.18** dove è necessario inserire porta e destinazione:

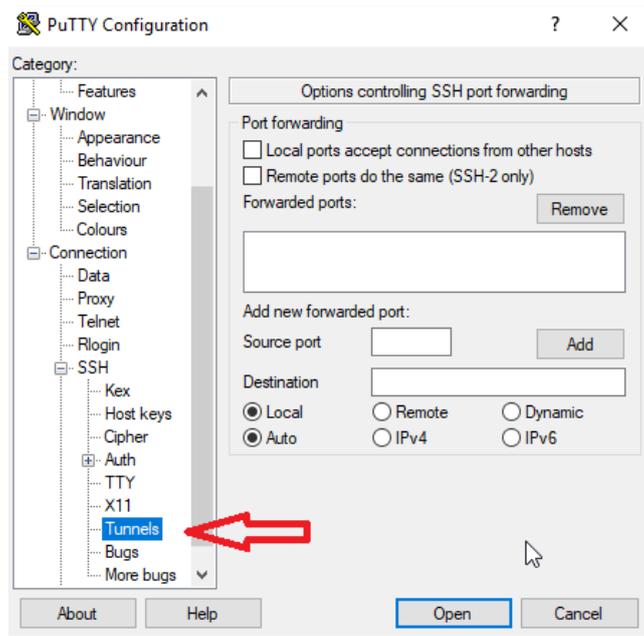


Figura 4.18: Inserimento porta e destinazione

- 3) Poi, cliccando sul tasto *Add* e sul tasto *Open*, il *client* “PuTTY” genera la console in cui è necessario inserire username e password forniti per stabilire la connessione:

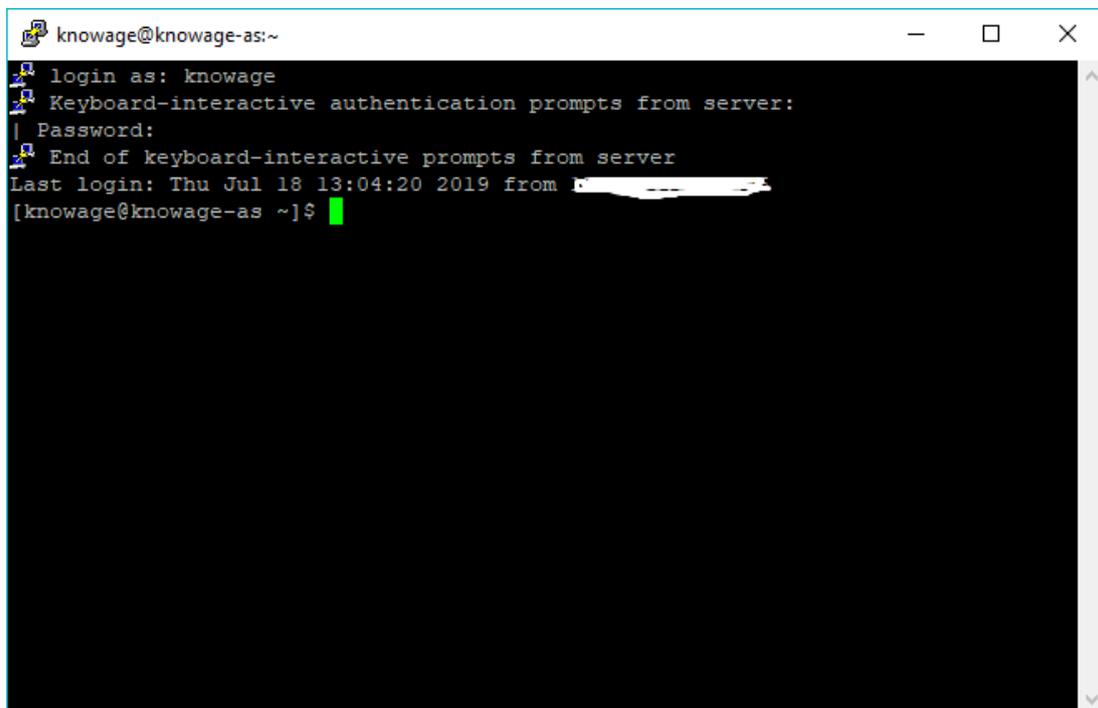
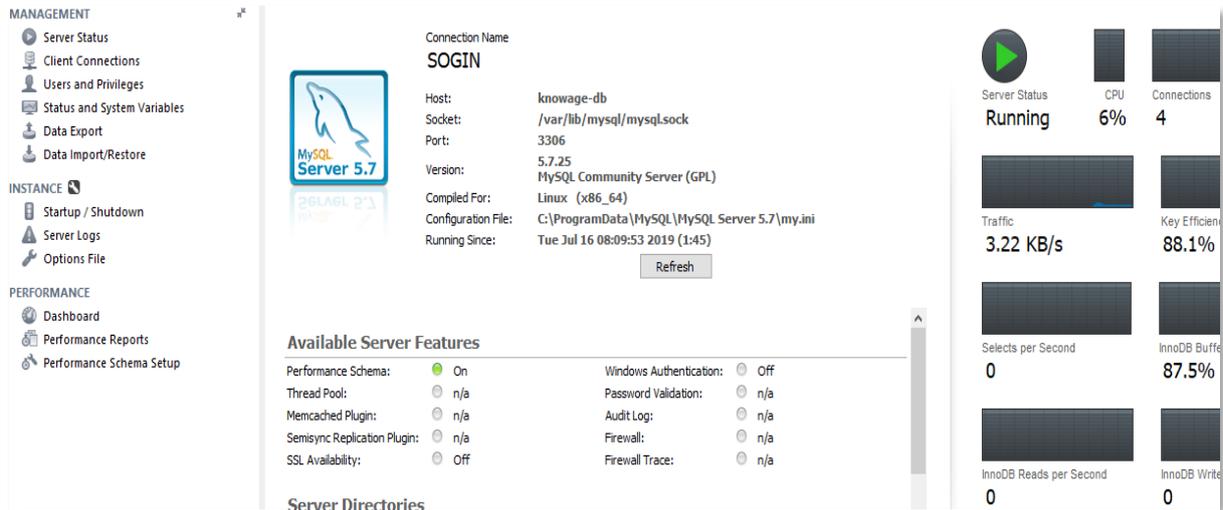


Figura 4.19: Console per inserimento username e password

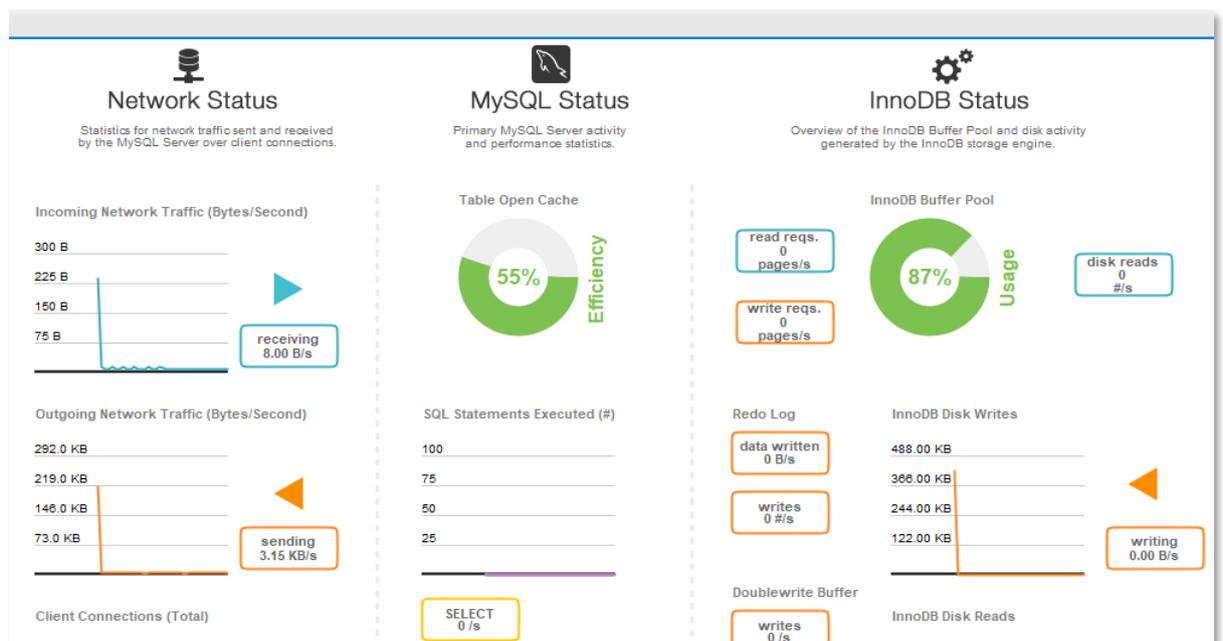
Va detto che la suddetta procedura si è ripetuta ogni qualvolta si è dovuta effettuare la connessione al DB in remoto per poter operare sulla tabella “Passenger\_cars” interagendo con il *software* MySQL.

Nella **Figura 4.20** è mostrato lo stato del server con una serie di informazioni specifiche, visualizzabile in qualsiasi momento tramite il *client* nella sezione *Management*, cliccando su *server status*.



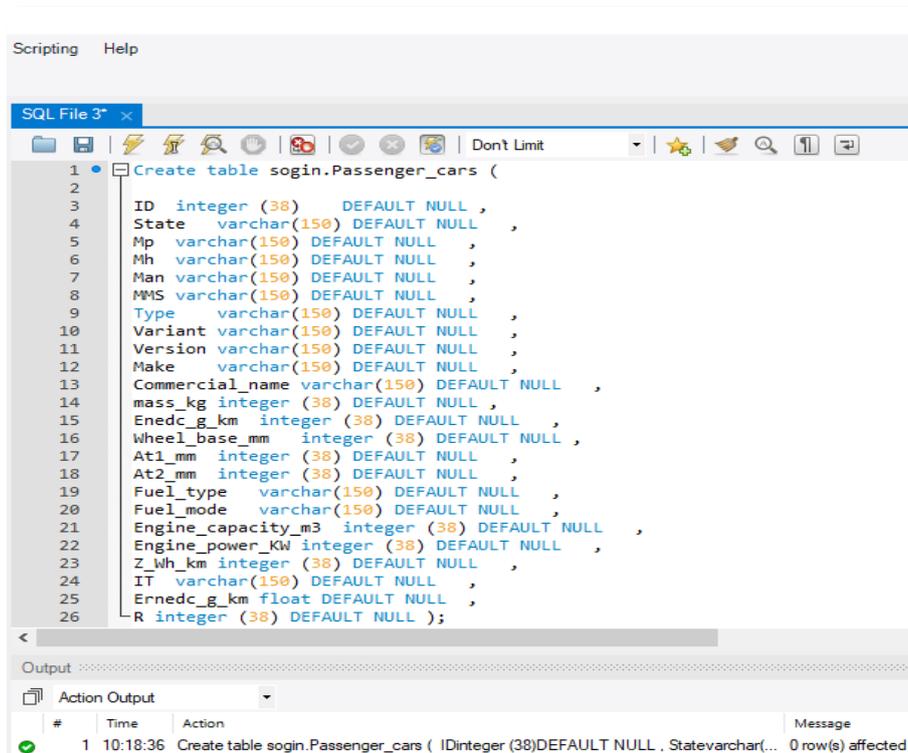
**Figura 4.20:** Stato Server

In **Figura 4.21**, nella sezione *Performance*, cliccando su *dashboard*, si possono visualizzare statistiche relative allo stato della rete, allo stato del *client*, con la relativa efficienza, utilizzazione (e molto altro) del DB:



**Figura 4.21:** Statistiche del DB

Fatto ciò, all'interno di MySQL, è stata definita la tabella "Passenger\_cars", visibile in **Figura 4.22**:



```
1 Create table sogin.Passenger_cars (  
2  
3 ID integer (38) DEFAULT NULL ,  
4 State varchar(150) DEFAULT NULL ,  
5 Mp varchar(150) DEFAULT NULL ,  
6 Mh varchar(150) DEFAULT NULL ,  
7 Man varchar(150) DEFAULT NULL ,  
8 MMS varchar(150) DEFAULT NULL ,  
9 Type varchar(150) DEFAULT NULL ,  
10 Variant varchar(150) DEFAULT NULL ,  
11 Version varchar(150) DEFAULT NULL ,  
12 Make varchar(150) DEFAULT NULL ,  
13 Commercial_name varchar(150) DEFAULT NULL ,  
14 mass_kg integer (38) DEFAULT NULL ,  
15 Enedc_g_km integer (38) DEFAULT NULL ,  
16 wheel_base_mm integer (38) DEFAULT NULL ,  
17 At1_mm integer (38) DEFAULT NULL ,  
18 At2_mm integer (38) DEFAULT NULL ,  
19 Fuel_type varchar(150) DEFAULT NULL ,  
20 Fuel_mode varchar(150) DEFAULT NULL ,  
21 Engine_capacity_m3 integer (38) DEFAULT NULL ,  
22 Engine_power_KW integer (38) DEFAULT NULL ,  
23 Z_Wh_km integer (38) DEFAULT NULL ,  
24 IT varchar(150) DEFAULT NULL ,  
25 Ernedc_g_km float DEFAULT NULL ,  
26 R integer (38) DEFAULT NULL );
```

#	Time	Action	Message
1	10:18:36	Create table sogin.Passenger_cars ( IDinteger (38)DEFAULT NULL , Statevarchar(...	0 row(s) affected

**Figura 4.22:** Definizione tabella Passenger\_cars

La suddetta tabella, una volta creata, si posiziona sotto alla connessione del DB nell'apposita sezione denominata *Tables*.

Per coerenza, anche in questo caso, è stato associato al tipo di dato "varchar" un limite massimo di 150 caratteri, al tipo di dato "integer" un numero massimo di 38 cifre ed è stato assegnato il valore di default "null" per tutti i campi.

Pertanto, in **Figura 4.23**, è descritta la struttura della tabella creata in MySQL:

Column	Type	Default Value	Nullable	Character Set	Collation	Privileges
ID	int(38)		YES			select,insert,update
State	varchar(150)		YES	latin1	latin1_swedish_d	select,insert,update
Mp	varchar(150)		YES	latin1	latin1_swedish_d	select,insert,update
Mh	varchar(150)		YES	latin1	latin1_swedish_d	select,insert,update
Man	varchar(150)		YES	latin1	latin1_swedish_d	select,insert,update
MMS	varchar(150)		YES	latin1	latin1_swedish_d	select,insert,update
Type	varchar(150)		YES	latin1	latin1_swedish_d	select,insert,update
Variant	varchar(150)		YES	latin1	latin1_swedish_d	select,insert,update
Version	varchar(150)		YES	latin1	latin1_swedish_d	select,insert,update
Make	varchar(150)		YES	latin1	latin1_swedish_d	select,insert,update
Commercial_name	varchar(150)		YES	latin1	latin1_swedish_d	select,insert,update
mass_kg	int(38)		YES			select,insert,update
Enedc_g_km	int(38)		YES			select,insert,update
Wheel_base_mm	int(38)		YES			select,insert,update
At1_mm	int(38)		YES			select,insert,update
At2_mm	int(38)		YES			select,insert,update
Fuel_type	varchar(150)		YES	latin1	latin1_swedish_d	select,insert,update
Fuel_mode	varchar(150)		YES	latin1	latin1_swedish_d	select,insert,update
Engine_capacity_m3	int(38)		YES			select,insert,update
Engine_power_KW	int(38)		YES			select,insert,update
Z_Wh_km	int(38)		YES			select,insert,update
IT	varchar(150)		YES	latin1	latin1_swedish_d	select,insert,update
Ernedc_g_km	float		YES			select,insert,update
R	int(38)		YES			select,insert,update

Figura 4.23: Struttura tabella creata in MySQL

A destra sono indicati i privilegi di cui l'utente può usufruire: selezione, inserimento e aggiornamento dei record della tabella.

Questo è di fondamentale importanza per l'esecuzione di *query* con l'inserimento di nuovi dati, l'aggiornamento di quelli presenti e la selezione.

Per concludere, si è proceduto con l'importazione dei record del *file* SQL all'interno della tabella "Passenger\_cars" creata nel DB in remoto in MySQL: in questo caso il caricamento dei dati è andato a buon fine.

In **Figura 4.24** e in **Figura 4.25** è mostrata la *query* che permette di selezionare tutti i record della tabella "Passenger\_cars", visualizzabili scorrendo la barra sulla destra in basso.

The screenshot displays a MySQL client window with the following components:

- Query Editor:** Contains the SQL query: `select * from Passenger_cars;`
- Result Grid:** Shows the first few rows of data from the 'Passenger\_cars' table. Columns include ID, State, Mp, Mh, Man, MMS, Type, Variant, Version, Make, Commercial\_name, mass\_kg, Enedc\_g\_km, Wheel\_base\_mm, and At1\_mm.
- Output Panel:** Shows the execution log with two entries:
 

#	Time	Action	Message	Duration / Fetch
1	10:54:51	select count(*) from Passenger_cars	1 row(s) returned	0.547 sec / 0.000 sec
2	10:56:05	select * from Passenger_cars	1048569 row(s) returned	0.156 sec / 522.094 sec

Figura 4.24: Selezione record – 1

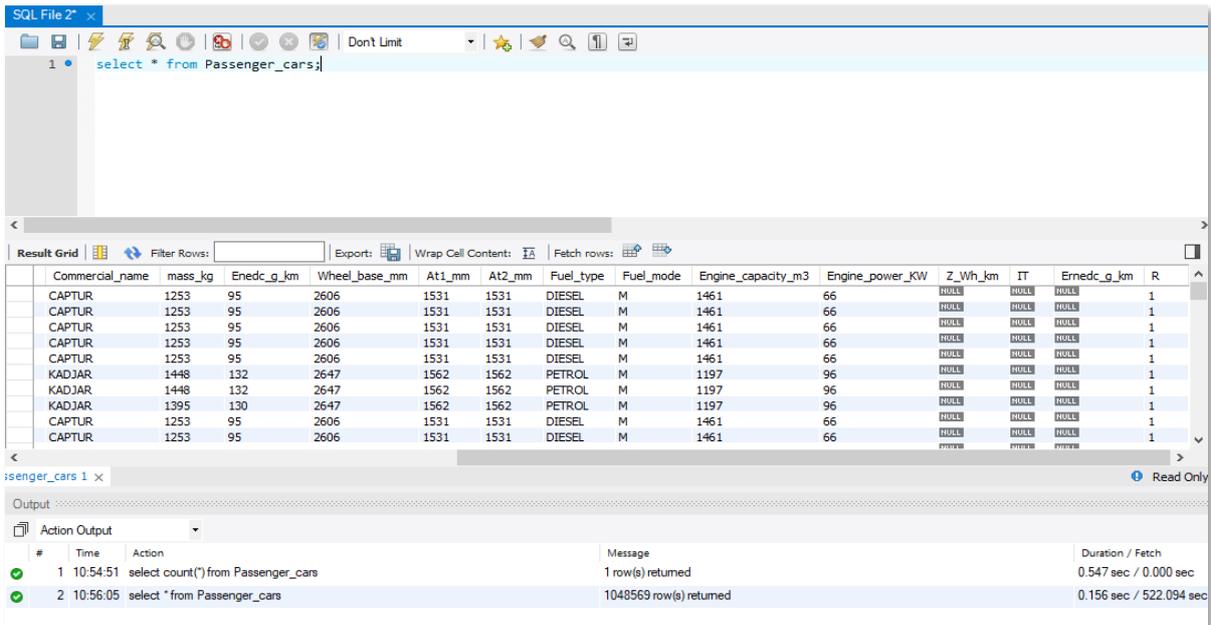


Figura 4.25: Selezione record - 2

Si noti che il tempo impiegato per l'esecuzione della query di cui sopra è stato di circa 520 secondi (approssimativamente 9 minuti); questo è visibile in basso a destra nelle due figure.

Nella **Figura 4.26** invece, viene riportata la *query* effettuata per il calcolo del numero di record totali contenuti nella tabella, che è proprio quello del *file* iniziale (circa 1048570 record).

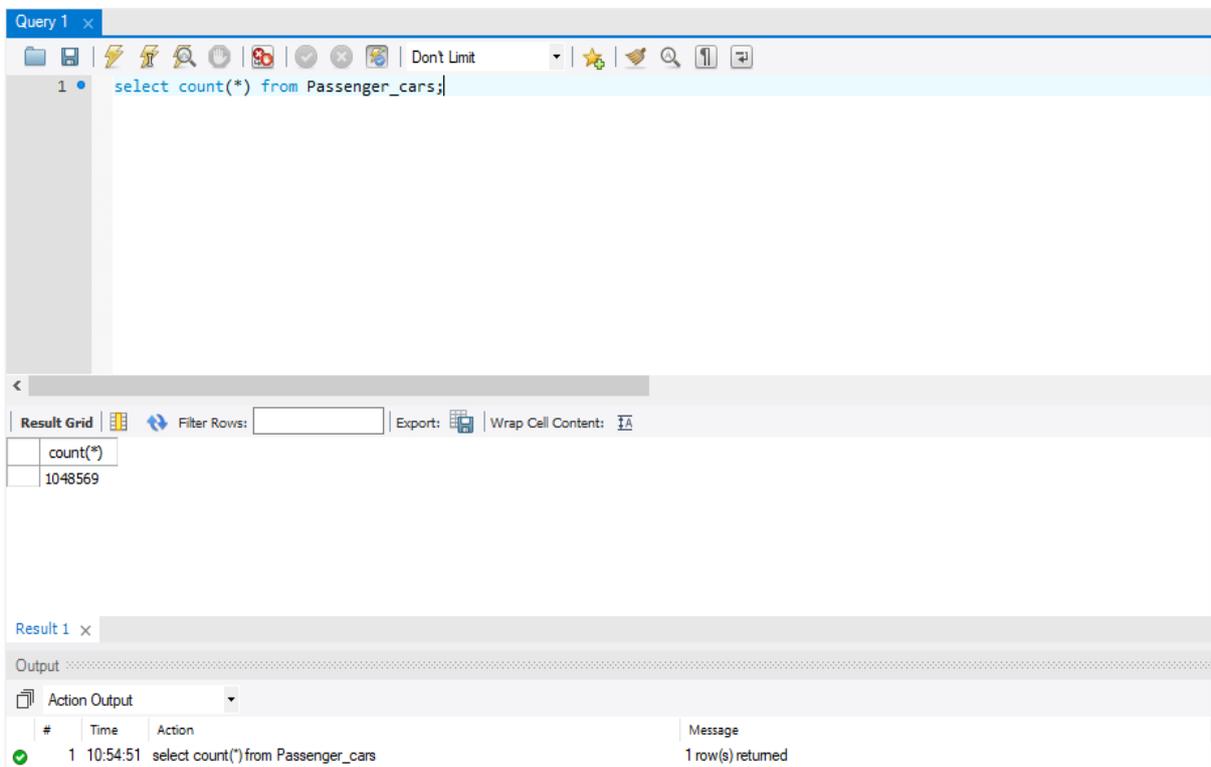
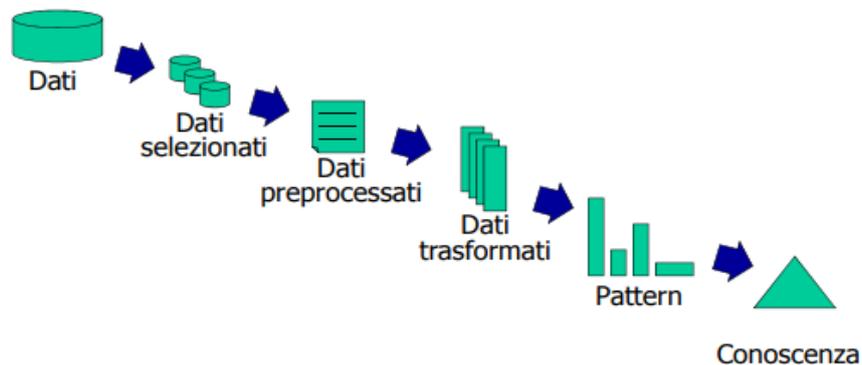


Figura 4.26: Query per righe totali

### 4.3 Tecniche di preparazione dei dati

In **Figura 4.27** viene mostrata la metodologia ed il percorso da seguire per poter estrarre conoscenza a partire da una grande mole di dati; in sintesi, la conoscenza può essere estrapolata dall'interpretazione e dalla valutazione dei cruscotti dinamici costruiti su KNOWAGE (*pattern*) definiti, a loro volta, solo dopo aver svolto opportune operazioni di *pre-processing* sui dati.



**Figura 4.27:** Percorso di estrazione della conoscenza dai dati

In generale, la procedura di ETL (*Extract, Transform, Load*), espressione inglese, si riferisce al processo di estrazione, trasformazione e caricamento dei dati in un sistema di sintesi. Mentre il processo di estrazione dei dati è stato descritto precedentemente, nel presente paragrafo vengono riportate, in dettaglio, tutte le operazioni di preparazione dei dati. L'ultimo *step*, quello del caricamento dei dati su KNOWAGE per l'analisi, è mostrato nel paragrafo successivo.

Il processo di preparazione dei dati, che ha lo scopo di consolidare i dati, cioè di renderli più adeguati alle analisi che devono essere svolte, consiste nella:

- Riduzione dei dati;
- Pulizia dei dati;
- Trasformazione dei dati.

Tuttavia, dopo aver effettuato una prima analisi del *dataset* generalmente si eseguono questa serie di operazioni che presuppongono la modifica dei dati e della loro struttura, come ad esempio l'eliminazione definitiva di colonne e record.

In casi come questo, in cui i dati sono storicizzati e gestiti mediante un DBMS (MySQL Workbench), e per l'analisi è necessario caricarli sul prodotto di *Business Analytics* (KNOWAGE) mediante opportune *query*, è considerata una *best practice* quella di lasciare immutata la struttura dei dati di partenza.

Quindi, per la creazione del *dataset* su KNOWAGE, si è deciso di selezionare e caricare solamente le colonne ed i record considerati utili e significativi ai fini dell'analisi.

Nel seguito sono descritte tutte le operazioni di riduzione, pulizia e trasformazione dei dati per la creazione del *dataset* sul prodotto di *Business Analytics*<sup>[15]</sup>.

Si noti che gran parte di queste operazioni non hanno determinato la modifica dei dati e della loro struttura ma sono state sintetizzate nella *query* per la creazione del *dataset* su KNOWAGE, che verrà mostrato nel prossimo paragrafo.

### 4.3.1 Riduzione dei dati

L'operazione di riduzione dei dati<sup>[15]</sup>, consiste, sostanzialmente, nella diminuzione del volume dei dati a disposizione, generando comunque risultati analitici simili, attraverso:

- **Feature Selection:** consiste nella selezione di un sottoinsieme degli attributi di partenza che conservi la distribuzione dei dati di partenza e li descriva in modo esaustivo e significativo;
- **Discretizzazione:** consiste nella ripartizione del dominio di un attributo continuo in un insieme di intervalli, riducendone, di fatto, la cardinalità. Possono essere creati N intervalli con la stessa ampiezza, oppure N intervalli con uguale cardinalità;
- **Campionamento:** consiste nella selezione di un sottoinsieme rappresentativo ed utile dei dati di partenza. Può essere eseguito un campionamento casuale (con o senza sostituzione) o stratificato.

Nella selezione dei dati per la creazione del *dataset* si è deciso di svolgere una *Feature Selection* e la discretizzazione di alcuni campi numerici:

- **Feature Selection:** consiste, come detto, nella selezione di un sottoinsieme degli attributi di partenza che permetta di descrivere in modo significativo i dati. Si è deciso di non includere, nella selezione dei dati della *query*, tutti quegli attributi o misure ritenuti irrilevanti (che non contengono informazioni ritenute utili per l'analisi), ridondanti (con informazioni contenute in un altro o più attributi), con un significato poco chiaro. I campi in questione risultano essere i seguenti:

- ID (irrilevante);
- Man (sintetizzato in Mh, informazione ridondante);
- MMS (sintetizzato in Mh, informazione ridondante);
- Type (significato poco chiaro);
- Variant (significato poco chiaro);
- Version (significato poco chiaro);
- Make (sintetizzato in Mh, informazione ridondante);
- IT (significato poco chiaro);
- R (irrilevante);
- Fuel\_mode (significato poco chiaro).

Al contrario, i campi ritenuti significativi ed utili per l'analisi, e selezionati per la creazione del *dataset*, risultano essere:

- **State;**
- **Mp;**
- **Mh;**
- **Commercial name;**
- **Mass\_kg;**

- **Enedc\_g\_km;**
- **Wheel\_base\_mm;**
- **At1\_mm;**
- **At2\_mm;**
- **Fuel\_type;**
- **Engine\_capacity\_m3;**
- **Engine\_power\_KW;**
- **Z\_Wh\_km;**
- **Ernedc\_g\_km.**

- Colonne dello schema di partenza: 24
- Colonne post Feature Selection: 14

➤ **Discretizzazione:** consiste nella suddivisione del dominio di un campo numerico in intervalli, riducendone, di fatto, la cardinalità. In questo caso, si è deciso di discretizzare le misure della potenza, della cilindrata, del peso e delle emissioni delle autovetture in relazione a possibili analisi da effettuare. Pertanto, sulla base di ricerche online e suggerimenti ricevuti da parte di persone competenti in materia, si è deciso di effettuare le seguenti suddivisioni:

- **Cilindrata (Engine\_capacity):** fino a 800 cc molto piccola, 801-1300 cc piccola, 1301-1800 cc media, 1801-2300 cc grande, oltre 2300 cc molto grande;
- **Potenza (Engine\_power):** fino a 55 kw molto bassa, 56-90 kw bassa, 91-125 kw media, 126-160 kw alta, oltre 160 kw molto alta;
- **Peso (mass):** fino a 1200 kg basso, 1201-2000 kg medio, oltre 2000 alto;
- **Emissioni CO2: (Enedc\_g\_km):** 1-110 g/km basse, 111-220 g/km medie, oltre 220 g/km alte.

### 4.3.2 Pulizia dei dati

L'operazione di pulizia dei dati<sup>[15]</sup> consiste nell'individuazione e successiva gestione risolutiva di:

- **Outliers o dati rumorosi:** rappresentano errori casuali o dati con una varianza significativa rispetto agli altri valori del dominio. Sono principalmente dovuti a problemi di *data entry*, malfunzionamenti di strumenti, limitazioni tecnologiche. Una volta riconosciuti si può optare per l'eliminazione o per l'aggiornamento.
- **Dati incompleti:** può succedere che ci siano mancanze di dati dovute al malfunzionamento di strumenti o alla scarsa importanza dell'informazione. E' possibile gestire il problema ignorando la tupla con l'informazione mancante, utilizzare un valore speciale (*null*) oppure utilizzare il valor medio dell'attributo.

In questo caso, per quanto riguarda gli *outliers* o dati rumorosi, si è deciso, nella *query*, sia di non selezionarli sia di aggiornarli, ove possibile, con i valori corretti.

Per quanto riguarda i dati incompleti invece, si è deciso, per ovvie ragioni, di utilizzare il valore speciale *null*.

Tuttavia, nel seguito, per ogni campo, viene descritta tutta la gestione dei valori nulli e dei dati rumorosi:

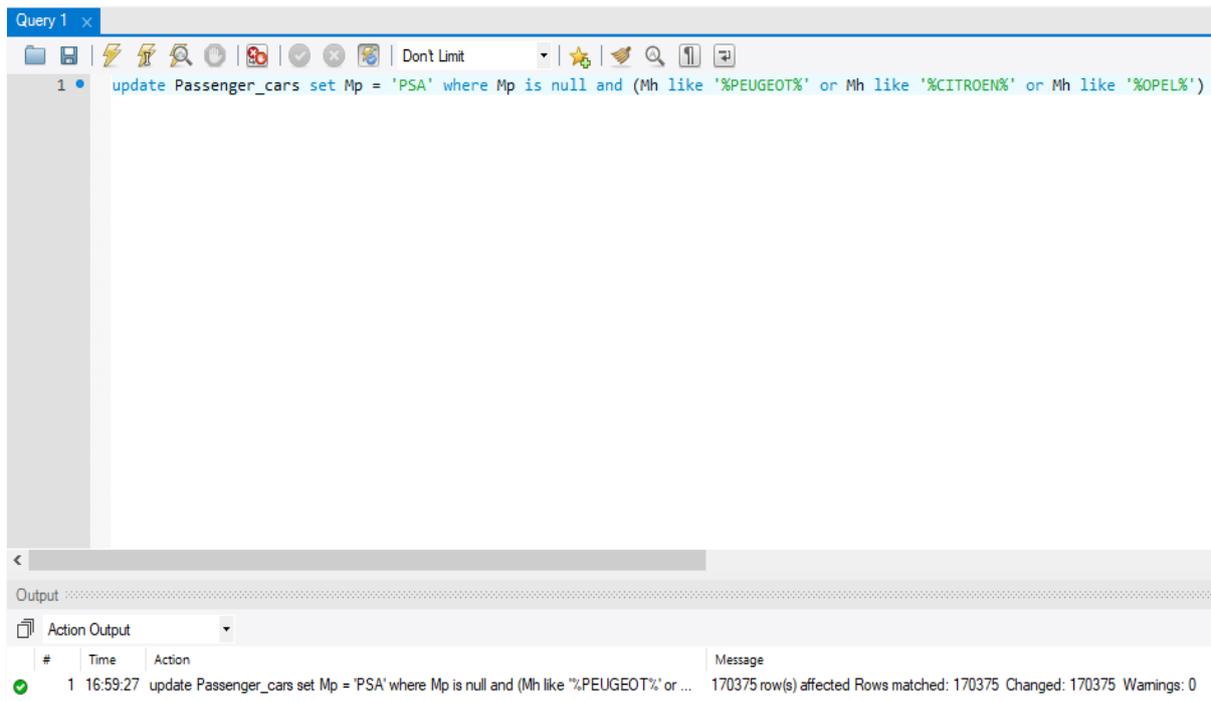
- **Gestione *outliers* (dati rumorosi) e valori nulli (dati incompleti):** per ciascun attributo e ciascuna misura, si è proceduto con la gestione dei valori considerati errori casuali (o con una varianza significativa rispetto agli altri valori del dominio, es. marchi di auto d'epoca, camion ecc.) e con la gestione dei valori nulli.

Nel seguito si ha una descrizione dettagliata in cui, per ogni campo, viene specificato quali dati si è deciso di aggiornare e quali dati si è deciso di escludere dalla selezione nella *query*.

- **State:** l'attributo non presenta *outliers* né valori nulli.
- **Mp (gruppo automobilistico a cui fanno riferimento diverse case costruttrici):** l'attributo non contiene *outliers* ma contiene circa 280.000 valori nulli; pertanto si è deciso di procedere con l'inserimento del gruppo automobilistico relativo alla casa costruttrice Mh di riferimento (effettuando le opportune ricerche). A titolo di esempio, appena sotto, viene mostrata qualche *query* ripetuta per ogni casa costruttrice (Mh) senza un gruppo di riferimento (Mp).

Nello specifico, la *query* mostrata in **Figura 4.28** mostra l'inserimento del gruppo automobilistico PSA in Mp per le case Citroen, Peugeot e Opel con eventuali nomenclature diverse presenti in Mh.

Va detto che è del tutto casuale che le tre case, controllate dallo stesso gruppo, abbiano tutte il valore nullo all'interno di Mp.

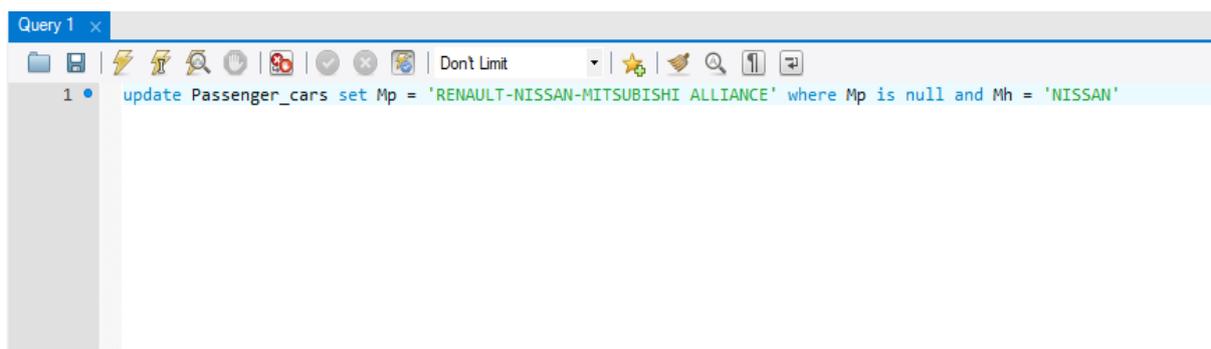


**Figura 4.28:** gestione valori nulli – Mp

Pertanto, in questo caso, sono state aggiornate circa 170.000 tuple delle 280.000 totali senza un gruppo automobilistico di riferimento.

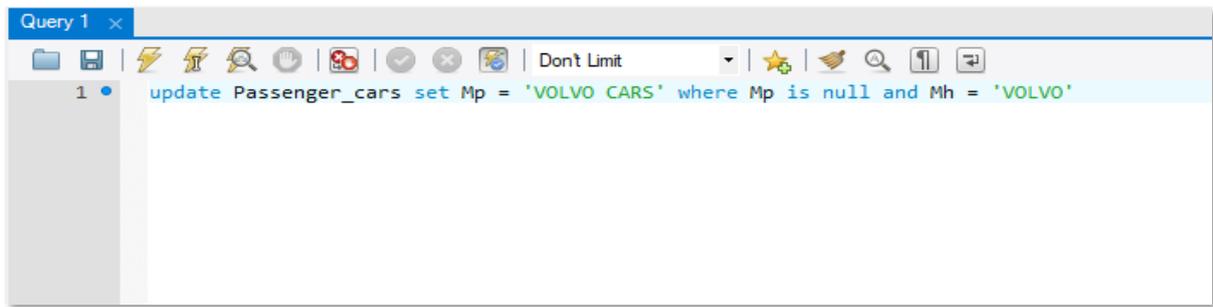
La *query* in **Figura 4.29**, invece, permette di inserire il gruppo citato (RENAULT- NISSAN-MITSUBISHI ALLIANCE) nel campo Mp per tutti i record della casa automobilistica il cui gruppo la controlla.

Tuttavia, sono state aggiornate circa 30.000 tuple.

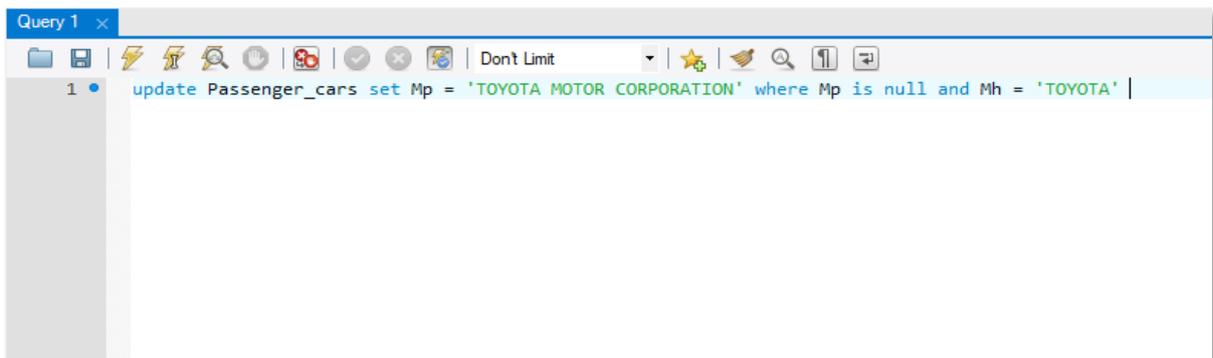


**Figura 4.29:** gestione valori nulli – Mp

La **Figura 4.30** e la **Figura 4.31** mostrano altre due *query* con l’inserimento del gruppo di riferimento (Mp) per la casa costruttrice in questione (Mh).



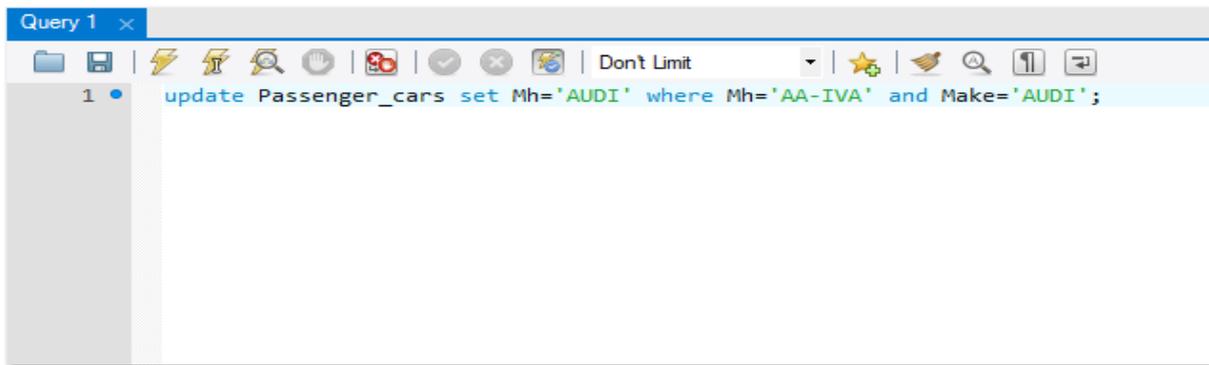
**Figura 4.30:** Gestione valori nulli – Mp



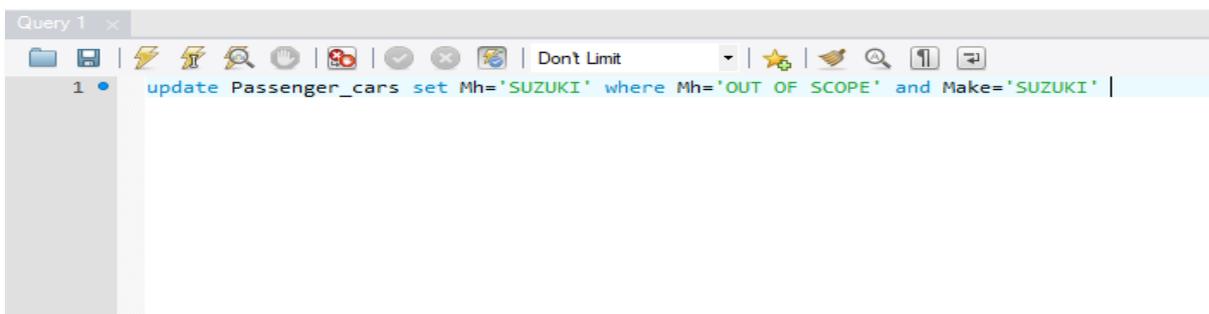
**Figura 4.31:** Gestione valori nulli – Mp

Infine, nella selezione dei dati, si è proceduto all'esclusione dei (pochi) valori nulli rimasti in Mp dal momento che non si è reso possibile aggiornare tutti quelli presenti.

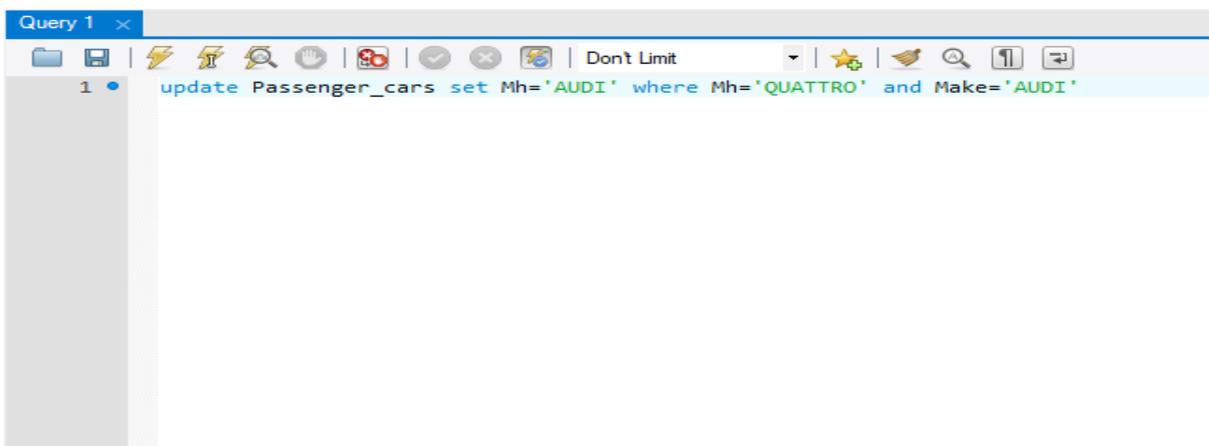
- **Mh (casa costruttrice dell'autovettura):** l'attributo non contiene valori nulli ma contiene dei valori considerati *outliers* tra cui AA-IVA, AA-NSS, BLUECAR, CATHERAM, DONKERVOORT, DUPLICATE, GREAT WALL MOTOR, KTM, LADA, MAHINDRA, MAN, MG MOTOR, MORGAN, OUT OF SCOPE, QUATTRO, RENAULT TRUCKS, SECMA, UNKNOWN. Sono stati aggiornati solo alcuni *outliers*, ove possibile, con la corretta casa costruttrice grazie alle informazioni presenti negli altri attributi ridondanti (Man, MMS, Make); nel seguito sono riportati, a titolo esemplificativo, alcuni esempi di *query* eseguite per l'aggiornamento dei dati.



**Figura 4.32:** Gestione outliers – Mh



**Figura 4.33:** gestione outliers – Mh

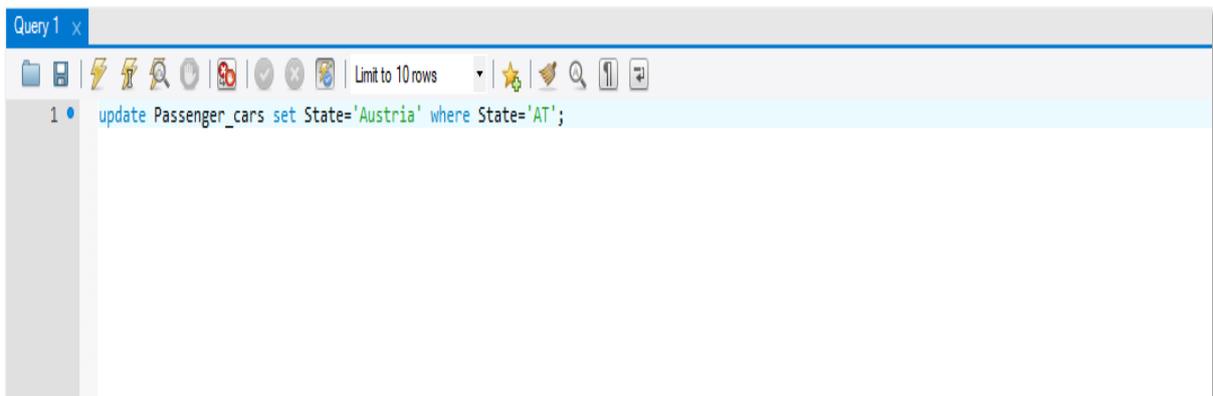


**Figura 4.34:** gestione outliers – Mh

Gli *outliers*, presenti in Mh, che si è deciso di escludere dalla selezione dei dati, poiché non del tutto aggiornabili o in contrasto con gli altri valori del dominio, risultano essere AA-IVA, BLUECAR, DONKERVOORT, GENERAL MOTORS, GREAT WALL MOTOR, KTM, LADA, MAHINDRA, MAN, MORGAN, RENAULT TRUCKS, SECMA, SUBARU, UNKNOWN.

- **Commercial name:** in questo caso non si è escluso alcun valore;
- **Mass kg:** nel peso dell'autovettura si è proceduto con la sola esclusione dei valori nulli;

- **Enedc g km:** anche qui si è proceduto con la sola esclusione dei valori nulli;
  - **Wheel base mm:** qui si è deciso di non selezionare tutti i record con il valore 0 e con il valore nullo;
  - **At1 mm e At2 mm:** si è proceduto all'esclusione dei valori nulli presenti nelle 2 misure;
  - **Fuel type:** si è deciso di non selezionare tutti i record con il valore nullo e con valori considerati *outliers* quali E85, Idrogeno, *other*;
  - **Engine capacity m3:** non si è escluso alcun valore (per le auto elettriche questo campo è nullo);
  - **Engine power KW:** si è deciso di non prendere in considerazione, nella selezione dei dati, tutti i record con il valore nullo;
  - **Z Wh km:** non sono presenti valori considerati *outliers* mentre i valori non nulli sono presenti solo quando l'autovettura in questione è elettrica;
  - **Ernedc g km:** gli *outliers* non sono presenti mentre i valori non nulli sono presenti solo al verificarsi di riduzioni di emissioni e, per questo motivo, i valori nulli non sono stati esclusi dalla selezione.
- **Ridenominazione dei valori:** per chiarezza e semplicità di lettura sono stati ridenominati i valori di alcuni attributi di tipo “*varchar*” per renderli più facilmente comprensibili, tra cui: stato (State), gruppo automobilistico (Mp), casa costruttrice (Mh), tipologia dell'alimentazione (Fuel\_type).
- **Attributo State:** sono stati ridenominati tutti gli stati, inserendo, al posto della sigla, il nome completo dello stato; la *query* in **Figura 4.35** è stata ripetuta per tutti gli stati.



**Figura 4.35:** Ridenominazione dei valori – State

- **Attributo Mh:** sono stati ridenominati (e riuniti con lo stesso nome) alcuni nomi delle case automobilistiche per evitare la presenza di differenti nomenclature dello stesso marchio.

Si noti che in ogni figura, in basso, è presente il numero di record aggiornati dall'istruzione specifica.

```
1 • update Passenger_cars set Mh = 'MERCEDES' where (Mh = 'DAIMLER AG' or Mh like '%MERCEDES%')
```

**Figura 4.36:** Ridenominazione dei valori – Mh

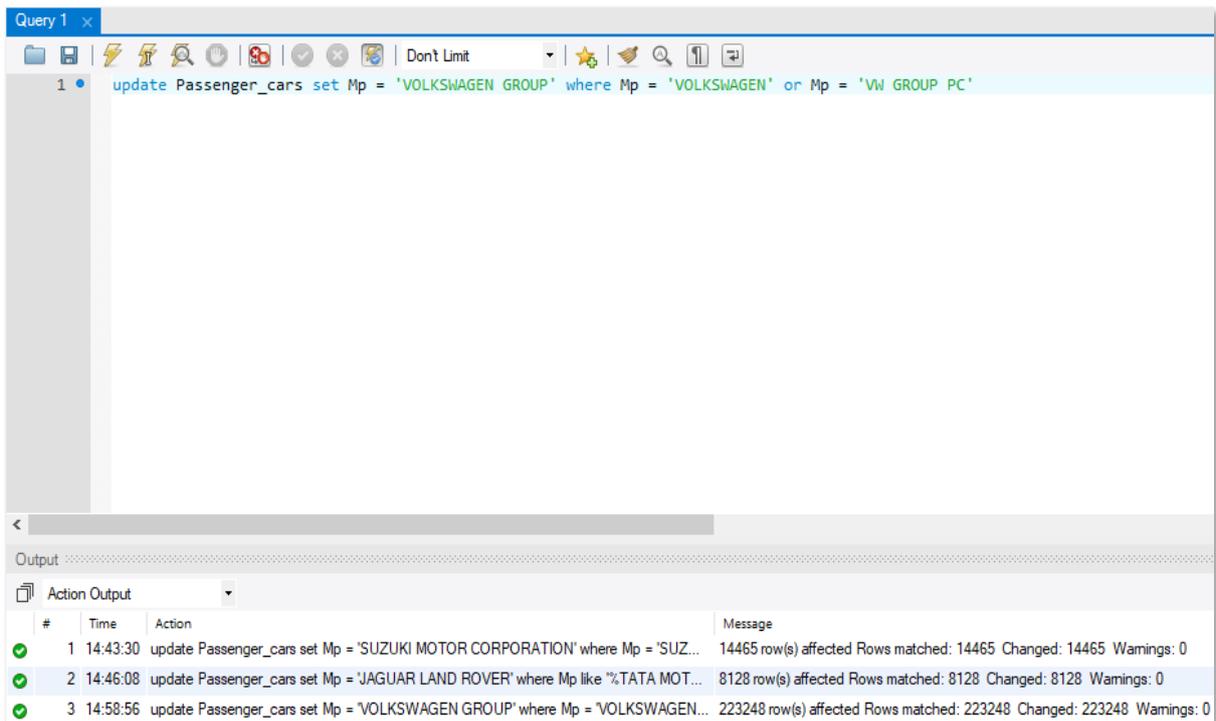
```
1 • update Passenger_cars set Mh = 'BMW' where Mh like '%BMW%'
```

**Figura 4.37:** Ridenominazione dei valori – Mh

- **Attributo Mp:** anche in questo caso, come nell'attributo Mh appena visto, si è proceduto alla ridenominazione (e definizione sotto lo stesso nome) dei nomi dei gruppi automobilistici per evitare la presenza di stessi gruppi con diverse nomenclature. A questo proposito, nelle figure sotto, sono mostrati esempi di *query* eseguite sui dati per la risoluzione del problema esposto.

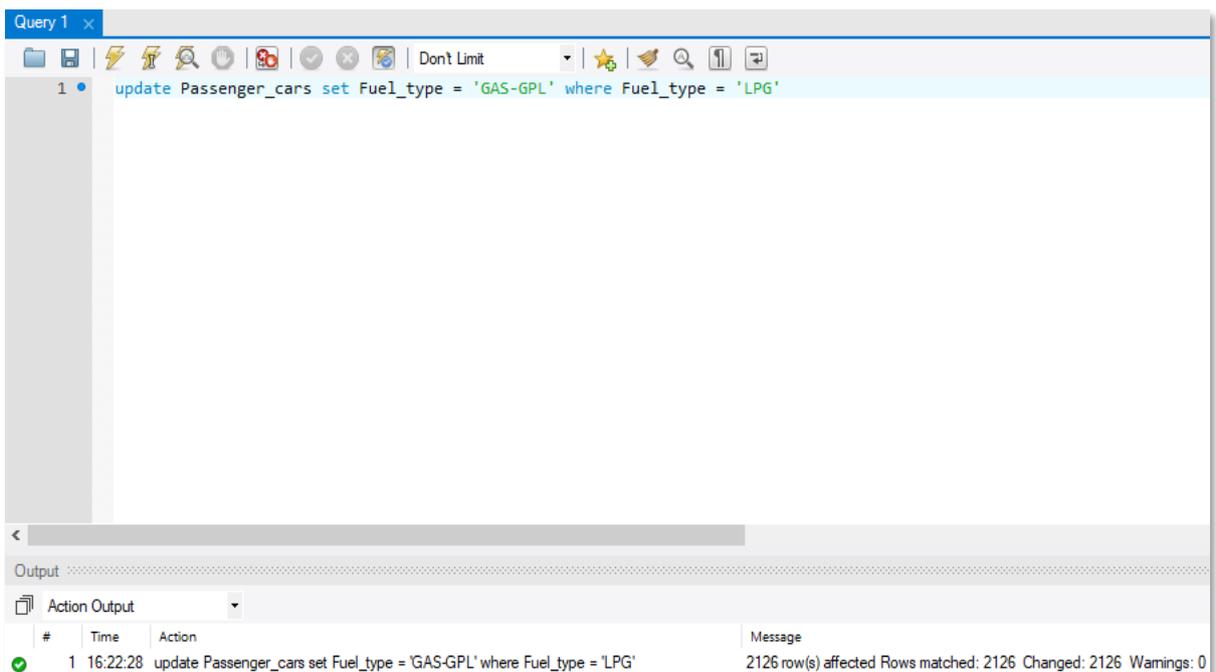
```
1 • update Passenger_cars set Mp = 'RENAULT-NISSAN-MITSUBISHI ALLIANCE' where (Mp = 'MITSUBISHI MOTORS' or Mp = 'RENAULT')
```

**Figura 4.38:** Ridenominazione dei valori – Mp

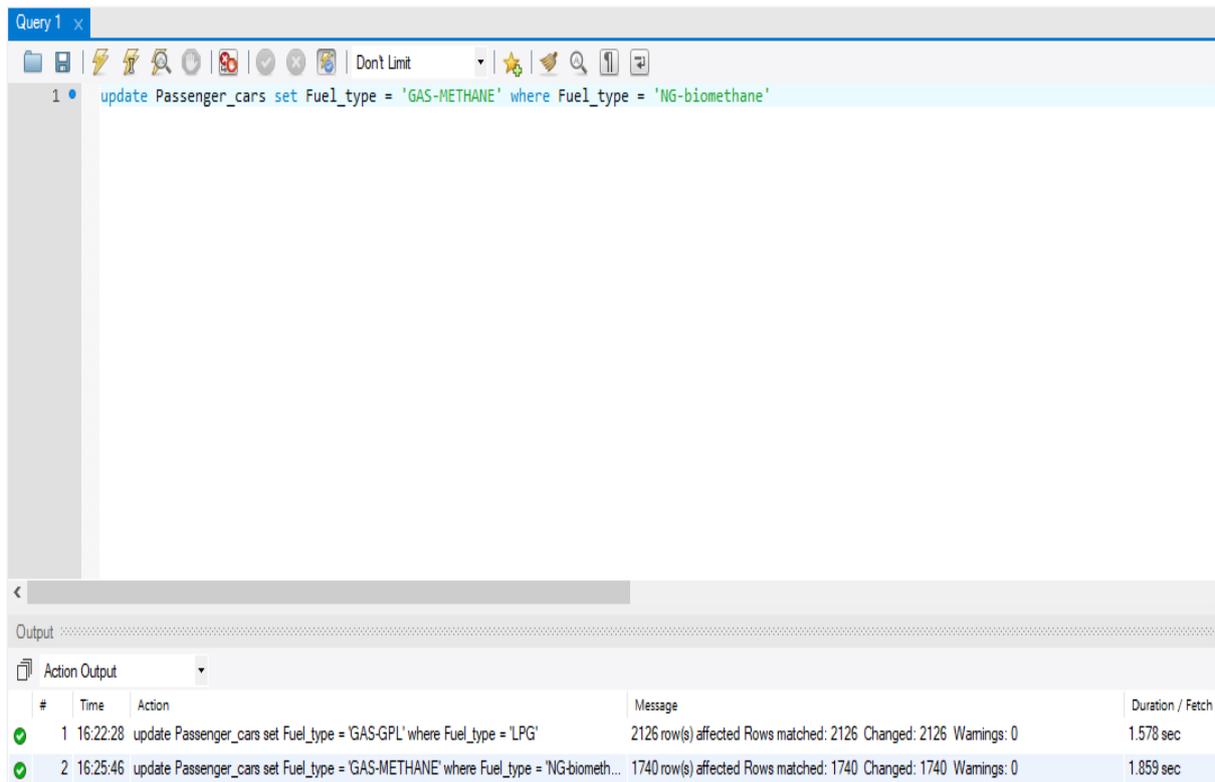


**Figura 4.39:** Ridenominazione dei valori – Mp

- **Attributo Fuel\_type:** sono stati ridenominati anche alcuni nomi relativi al tipo di alimentazione delle auto operando piccole variazioni nella nomenclatura. In basso un esempio di *query* che mostra l'aggiornamento da LPG a GAS-GPL e l'aggiornamento da NG-biomethane a GAS-METHANE.



**Figura 4.40:** Ridenominazione dei valori – Fuel\_type



**Figura 4.41:** Ridenominazione dei valori – Fuel\_type

### 4.3.3 Trasformazione dei dati

L'operazione di trasformazione dei dati<sup>[15]</sup> consiste nella modifica di una parte dei dati attraverso:

- **Normalizzazione:** consiste nel trasformare i valori di una misura in un dominio di valori (0, +1) oppure (-1, +1) utilizzando diverse tecniche;
- **Aggregazione:** consiste nel combinare insieme due o più attributi in un singolo attributo in modo tale da permettere una riduzione dei dati e di renderli più stabili (con meno variabilità).

Tuttavia, in questo caso, si è deciso di effettuare solamente un'aggregazione dei dati.

Come già mostrato nell'operazione di *Feature Selection* per la pulizia dei dati, gli attributi Man, MMS e Make, riguardanti i marchi delle autovetture, sono stati aggregati e sintetizzati in un singolo attributo che è Mh.

Tenendo conto di quanto esposto riguardo le operazioni di preparazione dei dati, inizialmente si è proceduto, per una semplice prova, alla selezione dei dati sul DBMS MySQL Workbench.

In **Figura 4.42** e in **Figura 4.43** è riportata la *query* effettuata e una parte del risultato restituito.

```

SQL File 2* x
select State
, Mp
, Mh
, Commercial_name
, mass_kg
, Enedc_g_km
, Wheel_base_mm
, At1_mm
, At2_mm
, Fuel_type
, Engine_capacity_m3
, Engine_power_KW
, Z_Wh_km, Ernedc_g_km
from Passenger_cars
where Mp is not null
and Mh not in ('AA-IVA', 'BLUECAR', 'DONKERVOORT', 'GENERAL MOTORS', 'GREAT WALL MOTOR', 'KTM', 'LADA', 'MAHINDRA', 'MAN', 'MORGAN', 'RENAULT TRUCKS', '
and mass_kg is not null
and Enedc_g_km is not null
and Wheel_base_mm is not null
and Wheel_base_mm not in (0)
and Fuel_type is not null
and Fuel_type not in ('E85', 'Hydrogen', 'OTHER')
and Engine_capacity_m3 is not null
and Engine_power_KW is not null

```

Figura 4.42: Query di prova

State	Mp	Mh	Commercial_name	mass_kg	Enedc_g_km	Wheel_base_mm	At1_mm	At2_mm	Fuel_type	Engine_capacity_m3
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	KADJAR	1448	132	2647	1562	1562	PETROL	1197
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	KADJAR	1448	132	2647	1562	1562	PETROL	1197
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	KADJAR	1395	130	2647	1562	1562	PETROL	1197
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CLIO	1165	85	2589	1509	1509	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461
Francia	RENAULT-NISSAN-MITSUBISHI ALLIANCE	RENAULT	CAPTUR	1253	95	2606	1531	1531	DIESEL	1461

Figura 4.43: Parte del risultato query di prova

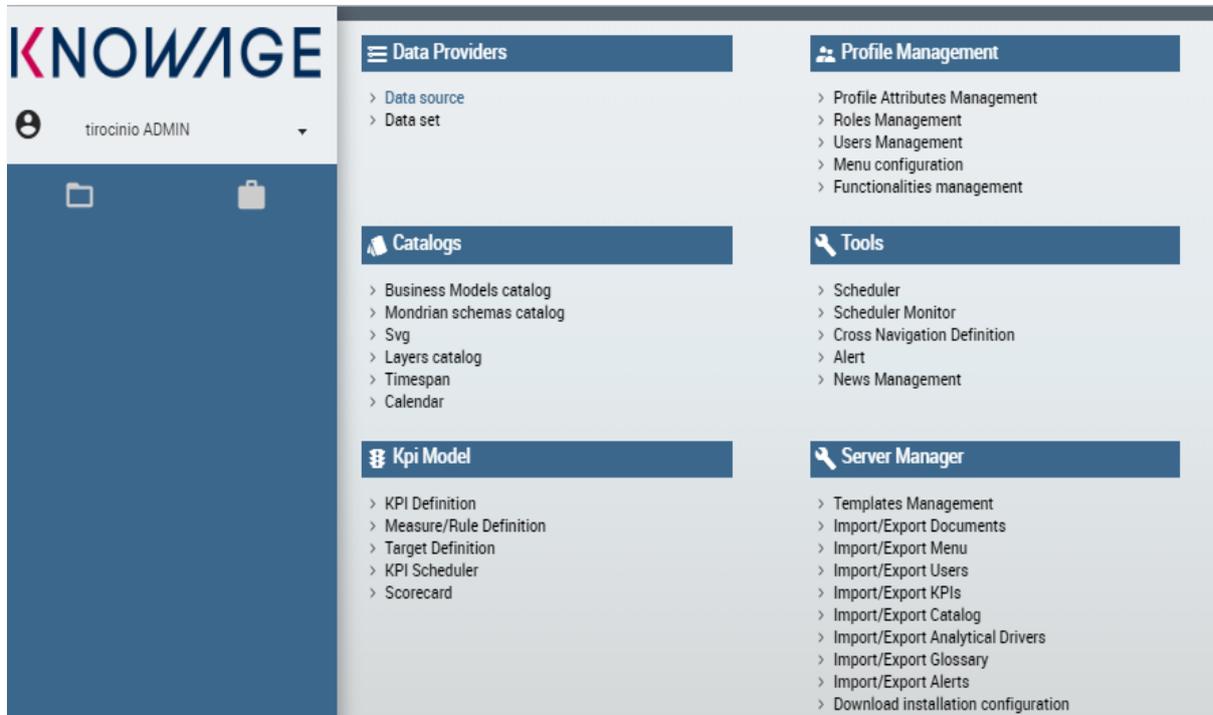
## 4.4 Creazione Data source e Dataset su KNOWAGE

Successivamente, la stessa *query* di cui sopra è stata eseguita su KNOWAGE per la creazione del *dataset* in modo tale da costruire i cruscotti evoluti e le dashboard; per fare questo è stato necessario caricare dapprima il *Data source*.

### 4.4.1 Caricamento Data source

Inizialmente, come mostrato in **Figura 4.44** e in **Figura 4.45**, cliccando su *Data source* nel menu principale di KNOWAGE, sotto la voce *Data Providers*, è stato possibile creare una connessione tra il prodotto di *Business Analytics* ed il DB in remoto.

Questo perché, come già discusso in precedenza, all'interno del DB in remoto è presente la tabella "Passenger\_cars" contenente tutti i dati necessari per la creazione dei cruscotti e dei relativi *widget*, i quali vengono selezionati mediante una specifica *query* per la creazione del *dataset*.



**Figura 4.44:** Creazione *Data source* - 1

Pertanto, come visibile in **Figura 4.45**, nell'aggiunta del nuovo *Data source* è stato necessario specificare alcune informazioni fondamentali quali l'etichetta, l'URL del DB, user, password e *driver* analitico.

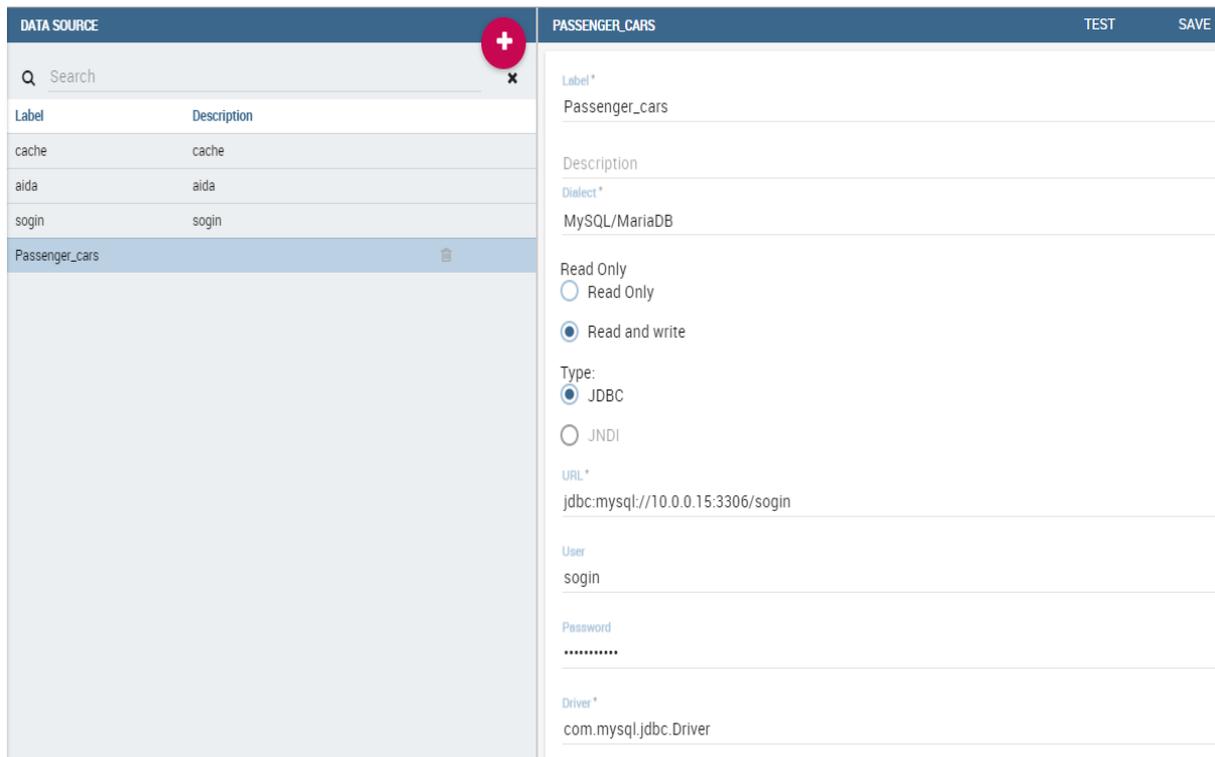


Figura 4.45: Creazione *Data source* - 2

## 4.4.2 Caricamento Dataset

Subito dopo la definizione del nuovo *Data source*, si è proceduto con la creazione del *Dataset*.

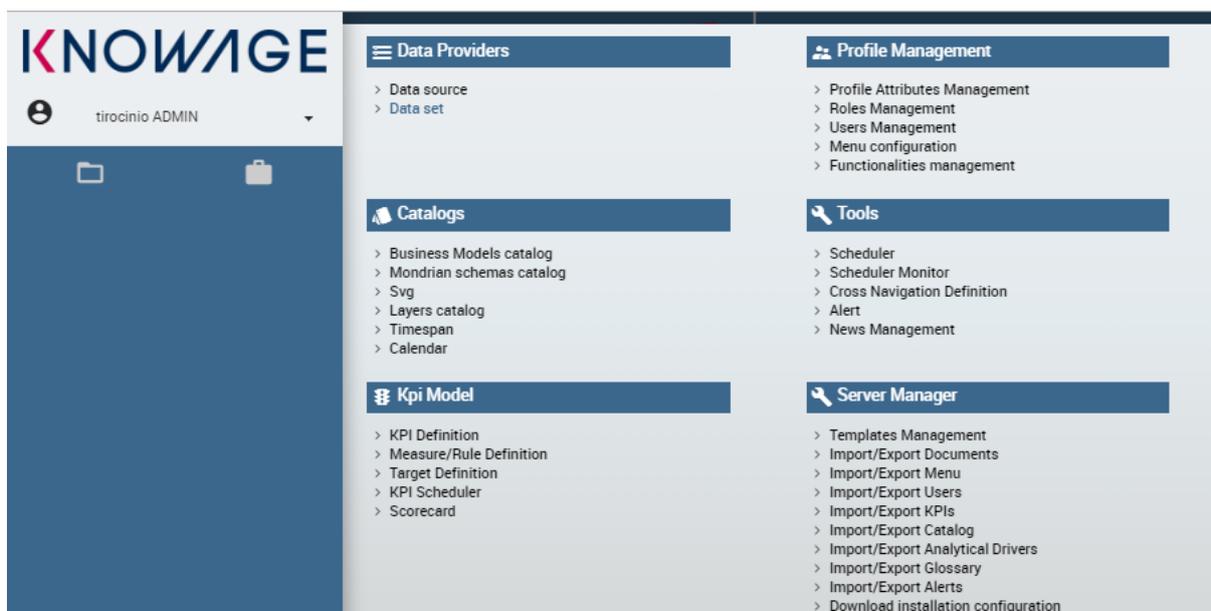
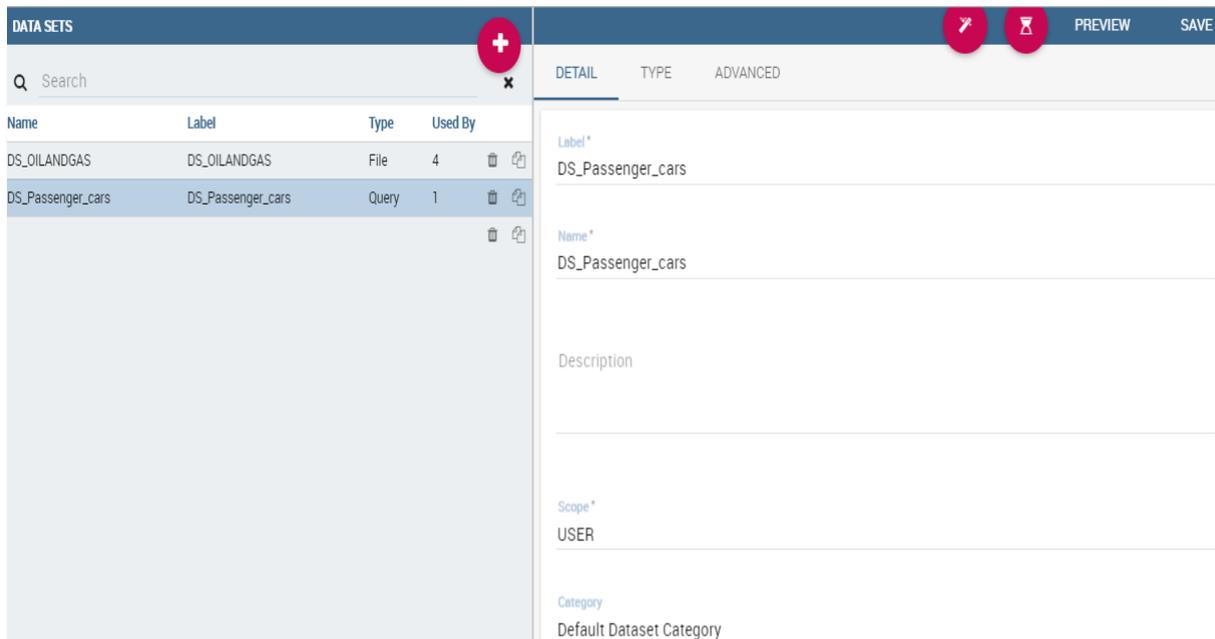


Figura 4.46: Creazione *Dataset* - 1

Dopo aver selezionato la voce *Data set* nella sezione *Data Providers*, come visibile in **Figura 4.47**, è stato necessario dapprima specificare tutta una serie di informazioni di base quali l'etichetta, il nome, l'ambito e la categoria.



**Figura 4.47:** Creazione *Dataset* - 2

Poi, nella voce *Type*, in **Figura 4.48**, è stato selezionato il tipo di *Dataset*, ovviamente *Query*. KNOWAGE supporta diverse tipologie di *Dataset*, tra cui, principalmente, *File*, *Query*, *Java Class*, *Script* e molti altri ancora, i quali rappresentano tutti la fonte dei dati primaria per il documento analitico che si vuole creare.

Mentre nella definizione di un *Dataset* di tipo *File* è necessario caricare un file XLS o CSV da cui vengono letti i dati, al contrario, in un *Dataset* di tipo *Query*, i dati vengono letti grazie ad un apposito codice SQL scritto all'interno di una specifica area.

In questo caso, la *Query*, visibile in **Figura 4.48** e successive due, è la stessa di quella mostrata e discussa alla fine del paragrafo precedente, effettuata per una semplice prova su MySQL e successivamente traslata su KNOWAGE per la creazione del *dataset* ma con l'aggiunta delle discretizzazioni (effettuate in seguito).

DETAIL TYPE ADVANCED

DataSet Type \*

Query

Data Source \*

Passenger\_cars

Query

```

1  select ID
2  , State
3  , Mp
4  , Mh
5  , Commercial_name
6  , mass_kg
7  , Enedc_g_km
8  , Wheel_base_mm
9  , At1_mm
10 , At2_mm
11 , Fuel_type
12 , Engine_capacity_m3
13 , Engine_power_KW
14 , Z_Wh_km, Ernedc_g_km
15 , case when Enedc_g_km between 0 and 110 then 'basse'
16     when Enedc_g_km between 111 and 220 then 'medie'
17     when Enedc_g_km > 220 then 'alte'
18     end as 'range_emissioni'
19 , case when Engine_capacity_m3 between 1 and 800 then 'molto piccola'
20     when Engine_capacity_m3 between 801 and 1300 then 'piccola'
21     when Engine_capacity_m3 between 1301 and 1800 then 'media'
22     when Engine_capacity_m3 between 1801 and 2300 then 'grande'

```

Figura 4.48: Creazione Dataset su KNOWAGE – 1

Query

```

22     when Engine_capacity_m3 between 1801 and 2300 then 'grande'
23     when Engine_capacity_m3 > 2301 then 'molto grande'
24     end as 'range_cilindrata'
25 , case when Engine_power_KW between 1 and 55 then 'molto bassa'
26     when Engine_power_KW between 56 and 90 then 'bassa'
27     when Engine_power_KW between 91 and 125 then 'media'
28     when Engine_power_KW between 126 and 160 then 'alta'
29     when Engine_power_KW > 160 then 'molto alta'
30     end as 'range_potenza'
31 , case when mass_kg between 1 and 1200 then 'basso'
32     when mass_kg between 1201 and 2000 then 'medio'
33     when mass_kg > 2000 then 'alto'
34     end as 'range_peso'
35
36 from Passenger_cars
37 where Mp is not null
38 and Mh not in ('AA-IVA', 'BLUECAR', 'DONKERVVOORT', 'GENERAL MOTORS', 'GREAT WALL MOTOR', 'KTM', 'LADA', 'MAHINDRA', 'MAN',
39 'MORGAN', 'RENAULT TRUCKS', 'SECMA', 'SUBARU', 'UNKNOWN')
40 and mass_kg is not null
41 and Enedc_g_km is not null
42 and Wheel_base_mm is not null
43 and Wheel_base_mm not in (0)
44

```

Figura 4.49: Creazione Dataset su KNOWAGE - 2

```

Query
32     when mass_kg between 1201 and 2000 then 'medio'
33     when mass_kg > 2000 then 'alto'
34     end as 'range_peso'
35
36 from Passenger_cars
37 where Mp is not null
38 and Mh not in ('AA-IVA', 'BLUECAR', 'DONKervoort', 'GENERAL MOTORS', 'GREAT WALL MOTOR', 'KTM', 'LADA', 'MAHINDRA', 'MAN',
39 'MORGAN', 'RENAULT TRUCKS', 'SECMA', 'SUBARU', 'UNKNOWN')
40 and mass_kg is not null
41 and Enedc_g_km is not null
42 and Wheel_base_mm is not null
43 and Wheel_base_mm not in (0)
44 and Fuel_type is not null
45 and Fuel_type not in ('E85', 'Hydrogen', 'OTHER')
46 and Engine_power_KW is not null
47 and ID not in (select ID
48                 from Passenger_cars
49                 where Fuel_type = 'ELECTRIC' and Z_Wh_km is null)
50 and ID not in (select ID
51                 from Passenger_cars
52                 where (Fuel_type = 'ELECTRIC/PETROL' or Fuel_type = 'ELECTRIC/DIESEL') and Z_Wh_km is null)

```

Figura 4.50: Creazione Dataset su KNOWAGE - 3

Al termine della definizione della *query* si è proceduto con la *Preview* del risultato, visibile in **Figura 4.51**, e con il conseguente salvataggio del *dataset*, da associare, quindi, a tutti i *cockpit* desiderati, come mostrato precedentemente.

Enedc_g_km	Wheel_base_mm	At1_mm	At2_mm	Fuel_type	Engine_capacity_m3	Engine_power_KW	Z_Wh_km	Enedc_g_km	range_emissioni	range_cilindrata	range_potenza	range_peso
95	2606	1531	1531	DIESEL	1461	66			basse	media	bassa	medio
95	2606	1531	1531	DIESEL	1461	66			basse	media	bassa	medio
95	2606	1531	1531	DIESEL	1461	66			basse	media	bassa	medio
95	2606	1531	1531	DIESEL	1461	66			basse	media	bassa	medio
95	2606	1531	1531	DIESEL	1461	66			basse	media	bassa	medio
132	2647	1562	1562	PETROL	1197	96			medie	piccola	media	medio
132	2647	1562	1562	PETROL	1197	96			medie	piccola	media	medio
130	2647	1562	1562	PETROL	1197	96			medie	piccola	media	medio
95	2606	1531	1531	DIESEL	1461	66			basse	media	bassa	medio
95	2606	1531	1531	DIESEL	1461	66			basse	media	bassa	medio
95	2606	1531	1531	DIESEL	1461	66			basse	media	bassa	medio
95	2606	1531	1531	DIESEL	1461	66			basse	media	bassa	medio
95	2606	1531	1531	DIESEL	1461	66			basse	media	bassa	medio

Figura 4.51: Parte della *Preview* del risultato della creazione del *Dataset* su KNOWAGE

## 4.5 Descrizione dei cruscotti (cockpit) realizzati

Dopo aver effettuato tutte le operazioni (descritte nei paragrafi precedenti) volte alla creazione del *dataset* all'interno del prodotto di BI, è stato possibile procedere con l'ideazione e la realizzazione dei **cruscotti dinamici**.

Pertanto, nel presente ed ultimo paragrafo del capitolo quattro vengono presentati e descritti i cruscotti dinamici (*cockpit*) realizzati su KNOWAGE per l'analisi sui dati delle autovetture.

A questo proposito, all'interno del prodotto di *Business Analytics* aziendale, si è proceduto alla creazione di tre differenti *cockpit* (documenti) contenenti tutte le tipologie di *widget* presentate nel capitolo tre: elementi testuali, HTML, diagrammi a barre, grafici a torta, tabelle e tabelle incrociate.

Ovviamente, per ogni *widget* inserito all'interno dei vari *sheet* (foglio) di ogni singolo documento (*cockpit*) si è opportunamente deciso di selezionare solo alcuni attributi e aggregare alcune misure, così come di definire dei filtri (condizioni) su specifici valori e scegliere lo stile desiderato del *widget* in termini di colori, grandezza e carattere del testo, bordi e altri piccoli dettagli.

I *widget* inseriti all'interno di ogni singolo *sheet* dei *cockpit* vengono mostrati nel corso della trattazione attraverso *screenshot* che non spiegano tutte le scelte operate in fase di prototipazione; per questo motivo, si cercano di spiegare, in modo generale, le operazioni eseguite per la realizzazione del singolo *widget*.

Durante la prototipazione dei *cockpit* l'analisi si è incentrata soprattutto sulle emissioni medie e sulle riduzioni medie di CO<sub>2</sub> (g/km) delle autovetture (grazie alle tecnologie innovative). Quindi, nei tre cockpit realizzati, si è deciso di analizzare sia le emissioni inquinanti che i risparmi di emissioni prendendo in considerazione i gruppi automobilistici, poi le nazioni, ed infine alcune caratteristiche strutturali delle automobili.

A questo proposito, si è scelto di effettuare un confronto tra le varie case costruttrici così come tra le singole nazioni in base alle attuali e sempre più stringenti restrizioni sulle emissioni ma anche per conoscere l'entità delle innovazioni tecnologiche volte a ridurre le emissioni inquinanti.

Infine, si è cercato di portare alla luce e di evidenziare le possibili correlazioni esistenti tra le diverse caratteristiche strutturali delle auto e le emissioni inquinanti di CO<sub>2</sub> (g/km).

Come già argomentato nel capitolo due, si è scelto di operare questo tipo di analisi anche per la rilevanza che, attualmente, rivestono l'inquinamento e l'effetto serra globale uniti al problema dei trasporti e della mobilità alternativa.

#### **4.5.1 Analisi per gruppi automobilistici (primo cockpit)**

In questo primo cruscotto, in cui sono stati presi in considerazione i gruppi automobilistici, si hanno ben cinque *sheet*, tra cui: ***overview***, **quota autovetture**, **emissioni di CO<sub>2</sub>**, **riduzioni di CO<sub>2</sub>**, **case costruttrici**.

Per quanto riguarda l'***overview***, si è voluto inserire alcuni *widget* con informazioni generali sui dati delle autovetture presi in analisi.

In **Figura 4.52**, utile solo a dare un'idea, è mostrato il *layout* dell'*overview* iniziale.

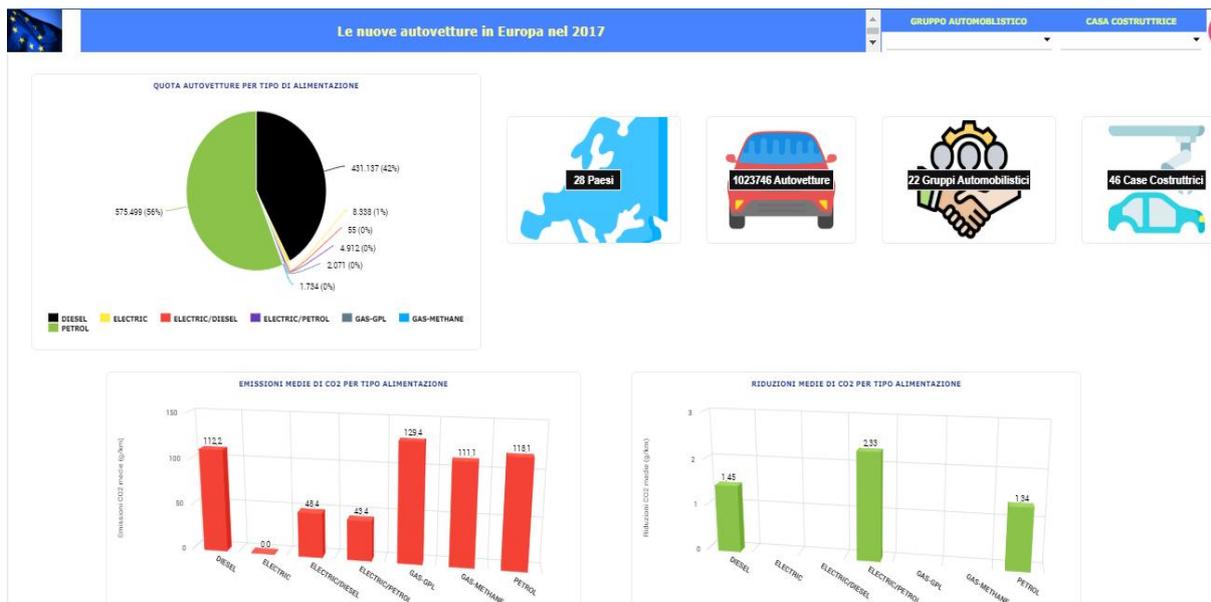


Figura 4.52: Layout overview

I quattro widget HTML, mostrati in **Figura 4.53** e successive tre figure, riportano, rispettivamente, il numero di nazioni di riferimento dei dati analizzati, il numero totale di autovetture del *dataset* (rappresentante il campione analizzato), il numero di case costruttrici ed i gruppi automobilistici a cui appartengono una o più case.



Figura 4.53: Paesi

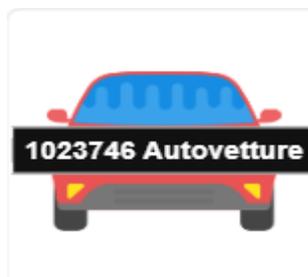


Figura 4.54: Automobili



Figura 4.55: Case



Figura 4.56: Gruppi

Il grafico a torta, presente in **Figura 4.57**, riporta la quota di automobili (in percentuale ed in valore assoluto) per ciascun tipo di alimentazione; come si può evincere guardando la figura, la quasi totalità (98%) delle autovetture europee di nuova immatricolazione risultano essere alimentate da benzina o diesel, mentre solo circa il 2% del totale comprende le auto ecologiche (elettriche, ibride e a gas).

Tali numeri sono certamente allarmanti considerando le maggiori emissioni inquinanti (in termini di particolato e altre sostanze nocive) prodotte da automobili a diesel o benzina rispetto a quelle alimentate a gas o elettriche.

Dal momento che i dati analizzati (circa un milione di record) rappresentano un campione, prendendo in considerazione dati ufficiali, secondo quanto riportato da una nota testata online<sup>[16]</sup>, le auto *green* (o ecologiche) immatricolate in Europa nel 2017 costituiscono circa il 6% delle auto totali vendute.

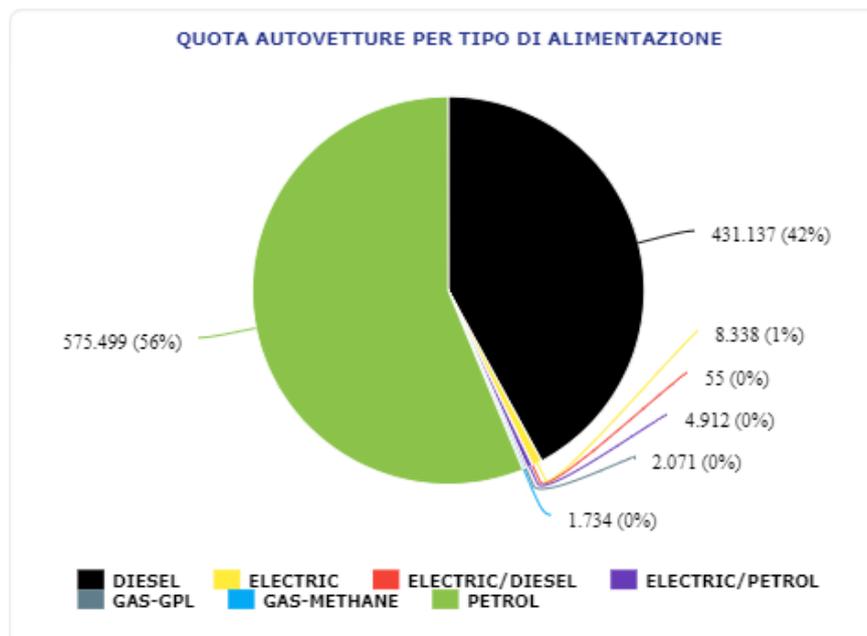


Figura 4.57: Quota autovetture per tipo alimentazione

Il diagramma a barre, visibile nell'*overview* in basso a sinistra e rappresentato in **Figura 4.58**, descrive le emissioni medie di CO<sub>2</sub> (g/km) per tipo di alimentazione; le maggiori emissioni medie di anidride carbonica provengono dalle auto alimentate a gas GPL con un valore di 129,4 g/km e a benzina con un valore di 118,1 g/km.

Al contrario, le minori emissioni di CO<sub>2</sub> provengono dalle auto elettriche con un valore pari a zero e dalle auto ibride con valori vicini ai 50 g/km.

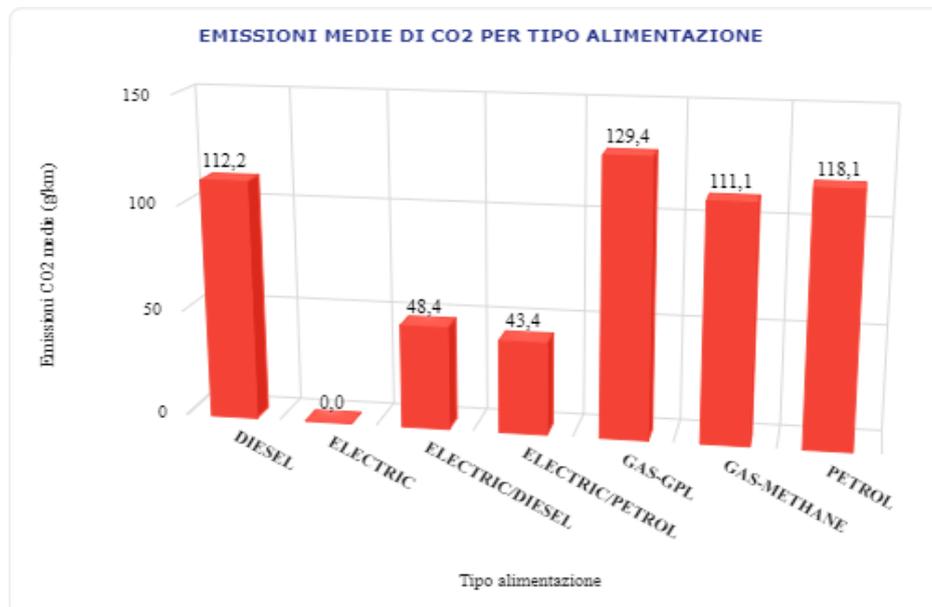
A questo proposito, il diesel, oltre a non contenere piombo, a differenza delle vetture a benzina dotate di catalizzatore emette emissioni molto più alte degli agenti inquinanti come ossidi di azoto, idrocarburi e particolato; l'introduzione dei convertitori catalitici ha comunque permesso, alle auto a benzina, di ridurre le emissioni di idrocarburi e ossidi di azoto.

Le automobili alimentate con gas GPL (gas di petrolio liquefatto) emettono una maggior quantità di CO<sub>2</sub> rispetto alle auto a diesel e a benzina favorendo il riscaldamento globale; al contrario, le loro emissioni in termini di particolato, ossidi di azoto e idrocarburi risultano di gran lunga inferiori sia al diesel che alla benzina.

Per questo motivo, l'uso di alimentazioni alternative come gas, auto ibride ed elettriche rappresentano l'unica soluzione per favorire un ambiente più pulito e un minor inquinamento globale.

Tuttavia, come visibile nel grafico, tranne per le automobili elettriche ed ibride, le altre tipologie di auto emettono emissioni medie ben al di sopra del limite dei 95 g/km imposto dall'Europa per il 2021.

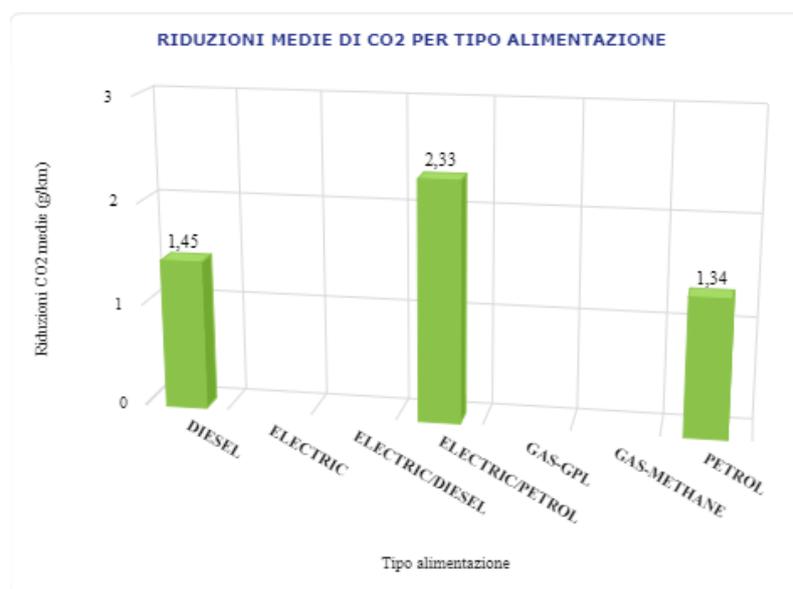
Questo è un risultato di notevole importanza in quanto presuppone ingenti investimenti in ricerca e innovazione da parte dei singoli gruppi automobilistici per la riduzione delle emissioni medie di CO<sub>2</sub> nel rispetto dei limiti imposti dall'Unione Europea.



**Figura 4.58:** Emissioni medie per tipo alimentazione

L'ultimo *widget* presente nell'*overview* e raffigurato in **Figura 4.59**, descrive la riduzione media di CO<sub>2</sub> (g/km) per tipo di alimentazione; in questo caso, la riduzione media si riferisce solamente alle autovetture per le quali le tecnologie innovative hanno portato ad una riduzione delle emissioni di CO<sub>2</sub>.

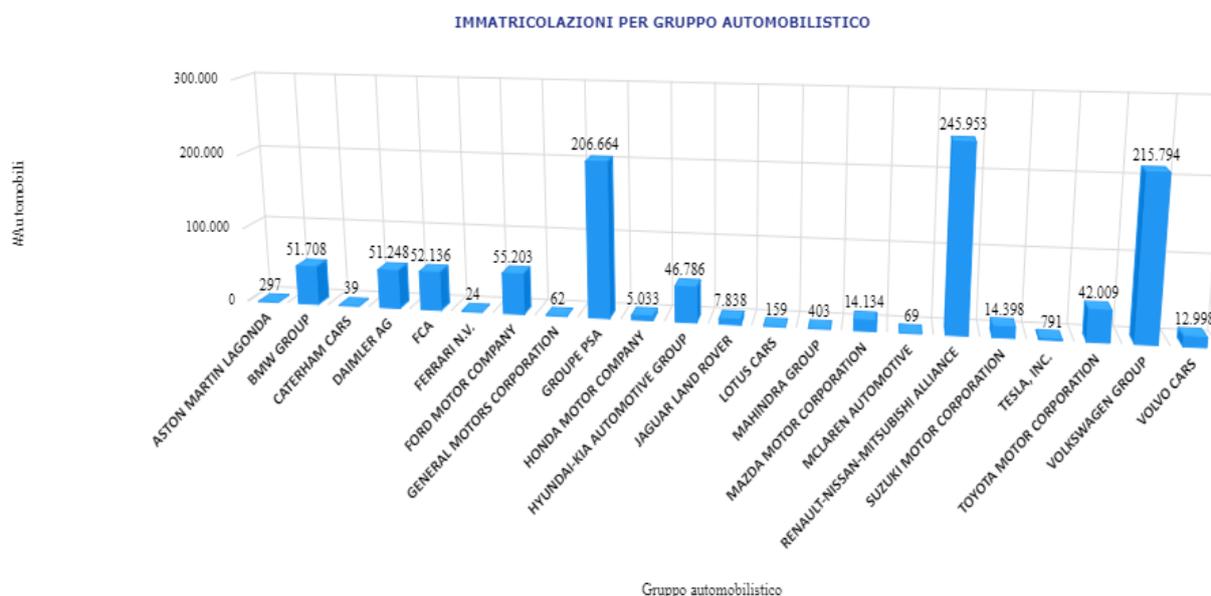
Alcune alimentazioni non presentano alcun valore, poiché, nel campione preso in considerazione, le autovetture relative non sono caratterizzate da alcuna riduzione di emissioni. Pertanto, le alimentazioni caratterizzate da riduzione di emissioni risultano essere diesel, benzina e le vetture elettriche con motore termico a benzina.



**Figura 4.59:** Riduzioni medie per tipo alimentazione

Per quanto riguarda il secondo *sheet*, denominato **quota autovetture**, in questo caso si è voluta dare un'idea dei gruppi automobilistici e le relative case con più vendite ma anche mettere in risalto la quota di autovetture per ciascun tipo di alimentazione e gruppo automobilistico.

Il diagramma a barre, mostrato in **Figura 4.60**, riporta il numero di immatricolazioni (nel campione) per i 22 gruppi automobilistici; ovviamente, non è importante il numero in sé quanto l'individuazione dei top gruppi in termini di vendite.



**Figura 4.60:** Immatricolazioni per ciascun gruppo automobilistico

A questo proposito, i primi tre gruppi in Europa, in termini di vendite, risultano essere il gruppo Renault, il gruppo Volkswagen ed il gruppo PSA; ad esempio, il gruppo Volkswagen, detentore di circa il 22% della quota di mercato nel 2017, è responsabile della vendita di quasi un'auto su quattro.

A seguire, gli altri gruppi presenti nella top 10 dei gruppi con più vendite in Europa risultano essere Hyundai-Kia, Bmw, Daimler AG, FCA, Ford, Toyota e Volvo.

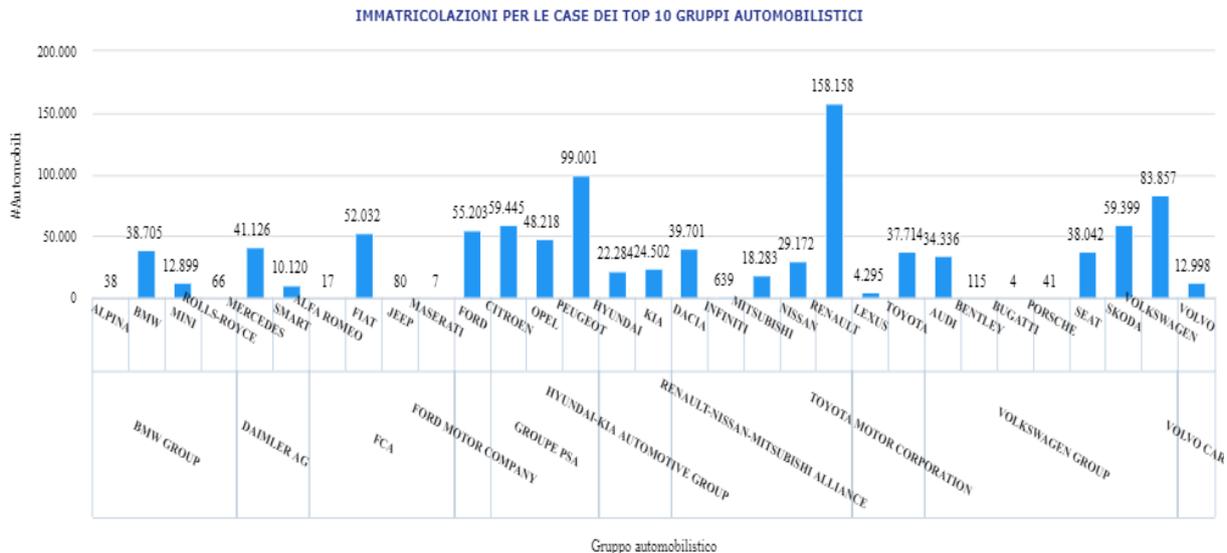
Tuttavia, sia nel 2017 che nel 2018, i principali gruppi come Volkswagen, PSA, Renault ma anche gruppi come Toyota, Hyundai-Kia e Volvo hanno visto crescere la loro quota di mercato in Europa.

Al contrario, Bmw, FCA, Ford e Daimler AG hanno visto diminuire, di qualche punto percentuale, la loro quota di mercato.

Secondo quanto sostenuto dal centro studi promotor guidato da Gian Primo Quagliano<sup>[17]</sup>, il mercato automobilistico mondiale è in forte crescita da diversi anni anche grazie alla forte innovazione tecnologica delle autovetture che attrae sempre più consumatori.

Va detto che, nel diagramma a barre appena descritto, dal momento che i dati analizzati fanno parte di un campione delle autovetture vendute totali, le posizioni occupate da alcuni gruppi automobilistici in termini di vendite totali non rispecchiano sempre la realtà ma possono cambiare di una o due posizioni.

La **Figura 4.61** mostra quali sono le singole case costruttrici che contribuiscono al maggior numero di vendite per i top 10 gruppi di automobili.



**Figura 4.61:** Immatricolazioni per ciascun gruppo e casa costruttrice

Ad esempio, per il gruppo PSA, secondo dati ufficiali<sup>[18]</sup>, le case con le maggior vendite risultano Peugeot e Citroen, entrambe con la quota di mercato crescita, nel 2018, di circa il 5%. Per il gruppo Renault-Nissan-Mitsubishi è senz'altro la Renault la casa più significativa in termini di vendite, anche se, nel 2018, ha perso circa il 4% di quota di mercato al contrario della Dacia le cui vendite hanno subito un incremento dell'11,7%.

Per quanto riguarda il gruppo Volkswagen, le case con il maggior peso in termini di vendite sono Volkswagen, Seat, Skoda e Audi; nel 2018, la crescita maggiore ha riguardato Seat con un valore di circa il 13%, quindi Volkswagen e Skoda con una crescita del 12%, mentre Audi ha subito una riduzione del 12%.

Per quanto riguarda, ad esempio, altri gruppi come FCA, Bmw e Daimler AG, le case più importanti in termini di vendite risultano la Fiat, la BMW e la Mercedes.

Nella *cross table*, rappresentata in **Figura 4.62**, si vuole dare un'idea della quota di auto per ciascun tipo di alimentazione e top 10 gruppi automobilistici.

IMMATRICOLAZIONI PER I TOP 10 GRUPPI AUTOMOBILISTICI E TIPO ALIMENTAZIONE								
↑ GRUPPI AUTOMOBILISTICI	↑ TIPO ALIMENTAZIONE							
	DIESEL	ELECTRIC	ELECTRIC/DIESEL	ELECTRIC/PETROL	GAS-GPL	GAS-METHANE	PETROL	TOTALE
BMW GROUP	33.876	188		2.482			15.162	51.708
DAIMLER AG	25.130	1.352		233		45	24.488	51.248
FCA	17.003				387	333	34.413	52.136
FORD MOTOR COMPANY	19.859						35.344	55.203
GROUPE PSA	89.572	528			633	90	115.841	206.664
HYUNDAI-KIA AUTOMOTIVE GROUP	14.815	395		208	111		31.257	46.786
RENAULT-NISSAN-MITSUBISHI ALLIANCE	124.861	4.551		707	935	2	114.897	245.953
TOYOTA MOTOR CORPORATION	1.965			3			40.041	42.009
VOLKSWAGEN GROUP	81.013	533	38	846		1.264	132.100	215.794
VOLVO CARS	9.936		17	433			2.612	12.998

**Figura 4.62:** Quota auto per top 10 gruppi e tipo alimentazione

Pertanto, tra i maggiori gruppi automobilistici in termini di vendite in Europa, la quota maggiore di auto elettriche vendute sul totale (circa 2,6%) spetta al gruppo Daimler AG con la Renault (circa 1,8%) subito dietro.

Considerando le auto ibride, nella fattispecie quelle con motore termico alimentato a benzina, è il gruppo Bmw a possedere la maggior quota di auto vendute sul totale (circa il 5%), subito dietro il gruppo Volvo (con il 3,3%).

Per quanto riguarda le autovetture a gas, sia quelle alimentate a GPL che quelle alimentate a metano, è il gruppo FCA a detenere la maggior quota sul totale, a seguire i gruppi Renault e Volkswagen.

In generale, tra i top 10 gruppi, quelli con il maggior numero di auto *green* (o ecologiche) vendute sul totale risultano essere Bmw (con circa il 5% sul totale), Volvo (con il 3,5%), Daimler AG (con il 3%) e Renault (con il 2,5%).

Nel terzo *sheet*, denominato **emissioni di CO<sub>2</sub>**, vengono analizzate più in dettaglio le emissioni di CO<sub>2</sub> (g/km) dei top 10 gruppi automobilistici.

Si è deciso di inserire una tabella, mostrata in **Figura 4.63**, per classificare in modo decrescente i top 10 gruppi automobilistici (definiti in precedenza) in base alle emissioni medie di CO<sub>2</sub> (g/km).

RANKING TOP 10 GRUPPI AUTOMOBILISTICI PER EMISSIONI MEDIE DI CO <sub>2</sub>				
Gruppo automobilistico	Emissioni CO <sub>2</sub> medie (g/km) ↓	Riduzioni CO <sub>2</sub> medie (g/km)	Peso/potenza medio	Cilindrata media (cc)
FORD MOTOR COMPANY	123,77	0,00	15,76	1.463,98
DAIMLER AG	123,11	1,37	15,18	1.810,19
VOLVO CARS	120,46	0,00	13,72	1.976,89
FCA	119,41	0,00	17,62	1.351,78
BMW GROUP	118,92	1,67	13,48	1.937,60
HYUNDAI-KIA AUTOMOTIVE GROUP	118,78	0,00	16,69	1.377,83
VOLKSWAGEN GROUP	115,64	1,00	15,11	1.462,54
RENAULT-NISSAN-MITSUBISHI ALLIANCE	109,69	0,00	17,48	1.345,28
GROUPE PSA	108,30	1,18	16,37	1.394,25
TOYOTA MOTOR CORPORATION	106,27	0,00	16,98	1.743,45

**Figura 4.63:** Ranking gruppi per emissioni medie

La prima colonna della tabella riguarda le emissioni medie di CO<sub>2</sub>, che vengono calcolate prendendo in considerazione le emissioni (g/km) di tutte le automobili che afferiscono al relativo gruppo automobilistico; quindi, possono esserci autovetture che inquinano di più ed altre che inquinano di meno.

Poi, per ognuno dei top 10 gruppi, viene mostrata anche la riduzione media di emissioni di CO<sub>2</sub>, il rapporto tra peso e potenza medi delle relative automobili e la cilindrata media delle autovetture di ciascun gruppo per evidenziare una possibile correlazione con le emissioni di CO<sub>2</sub>.

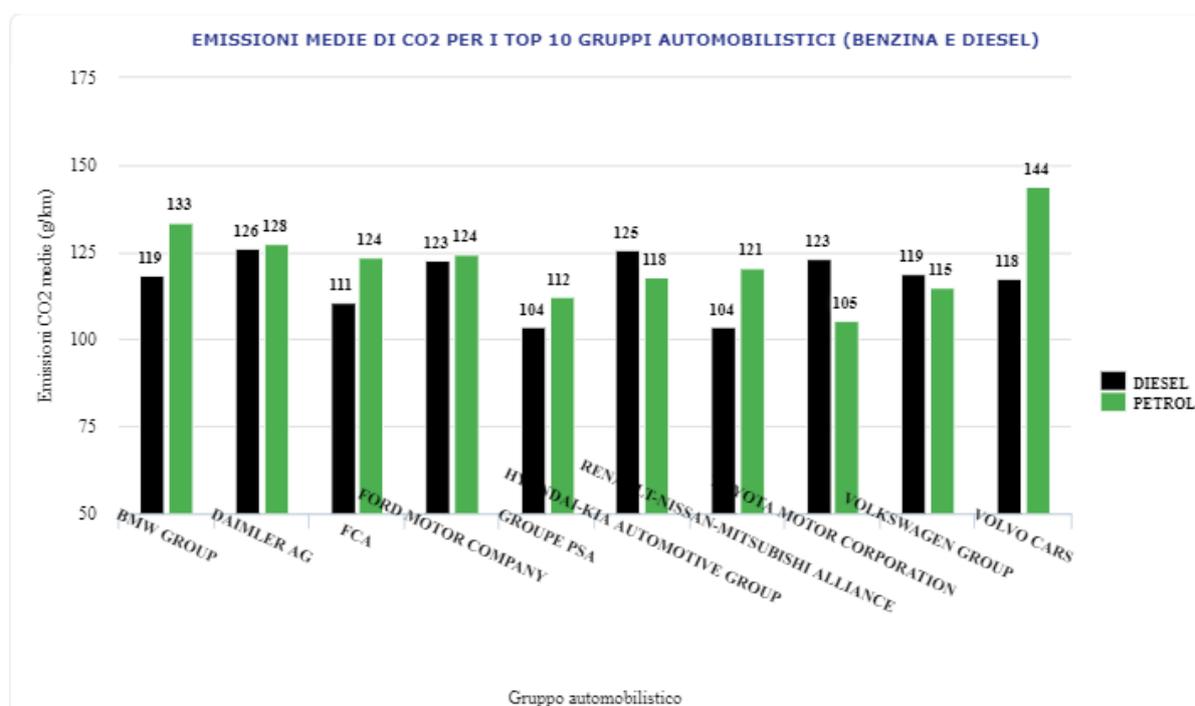
In tema di emissioni di CO<sub>2</sub>, il limite attuale, in vigore dal 2015, è pari a 130 g/km; al contrario, quello che entrerà in vigore dal 2021, sarà pari a 95 g/km e di 59 g/km quello in vigore dal 2030.

Prendendo come riferimento l'anno 2017, tutti i top costruttori, in termini di vendite di automobili in Europa, rispettano il limite di emissioni ancora in vigore dal 2015 (130 g/km) mentre risultano essere molto lontani dalla media di 95 g/km da rispettare a partire dal 2021; in questo senso, nell'anno 2018, secondo quanto riportato da una nota rivista online<sup>[19]</sup>, tra tutti i top 10 gruppi presenti in **Figura 4.63**, gli unici che hanno ridotto la propria media di emissioni di CO<sub>2</sub> sono stati Nissan e Toyota (rispettivamente -3.0 e -1.0 g/km) con Tesla l'unica al di

sotto del limite del 2021 (0 g/km), mentre tutti gli altri, specialmente quelli in testa come Volvo, Daimler AG, FCA, Ford e Bmw hanno avuto il maggior incremento di emissioni medie di CO<sub>2</sub>. Inoltre, per coloro che non rispetteranno i limiti imposti dall'Unione Europea scatteranno delle multe salatissime che possono essere evitate solamente mediante una vendita ingente di autovetture elettriche, ibride o a gas in modo tale da favorire l'abbassamento del livello medio di emissioni di CO<sub>2</sub>; a questo proposito, i gruppi automobilistici già da tempo e in modo massiccio investono e utilizzano tecnologie sempre più innovative nel segmento delle auto elettriche ed ibride favorendo una produzione su larga scala per cercare di ridurre le proprie emissioni medie di CO<sub>2</sub> in risposta ai vincoli dell'UE.

L'Unione Europea concederà vincoli meno stringenti, in tema di emissioni, alle case la cui produzione di auto elettriche e ibride supererà la soglia del 15% sul totale entro il 2025 e del 30% entro il 2030.

I diagrammi a barre, in **Figura 4.64** e in **Figura 4.65**, offrono un confronto riguardo le emissioni medie e per i top 10 gruppi, rispettivamente tra benzina e diesel e poi tra le alimentazioni alternative.



**Figura 4.64:** Emissioni medie per benzina e diesel

Per quanto riguarda il diesel, le emissioni medie di CO<sub>2</sub> più elevate appartengono al gruppo Daimler AG, mentre le auto a benzina più inquinanti, in termini di CO<sub>2</sub>, sono del gruppo Volvo. Al contrario, le emissioni medie più basse per diesel e benzina appartengono, rispettivamente, alle auto del gruppo PSA e Renault insieme e Toyota.

Non a caso, i gruppi che in media inquinano di più, considerando le auto a benzina e diesel, sono anche i gruppi che, in media, emettono più inquinanti di CO<sub>2</sub> posizionandosi nelle prime posizioni della speciale classifica mostrata prima in **Figura 4.63** (Daimler AG, Volvo, Ford, FCA, Bmw).

Al contrario, i gruppi con le minori emissioni medie di CO<sub>2</sub> per le auto alimentate a benzina e diesel sono anche quelli che, mediamente, in generale, inquinano di meno (PSA, Toyota, Renault).

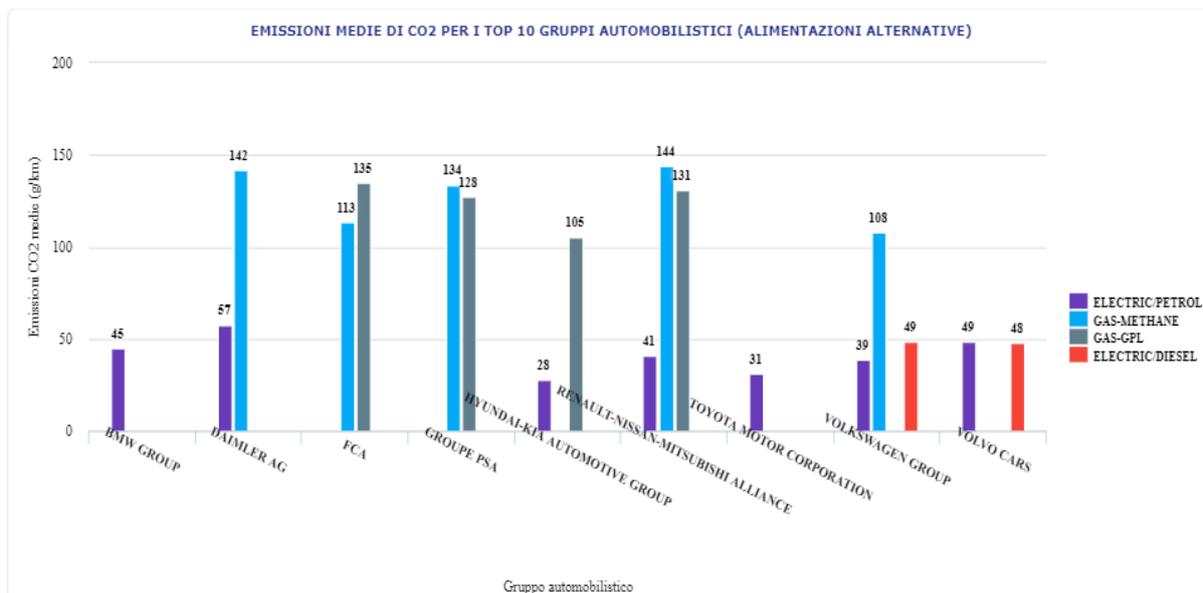


Figura 4.65: Emissioni medie per le alimentazioni alternative

Per quanto riguarda le alimentazioni alternative, il gruppo Renault è quello ad emettere maggiori inquinanti medi, in termini di CO2, considerando le auto a gas metano; al contrario, per le auto a gas GPL, è il gruppo FCA ad emettere, mediamente, maggior CO2.

Per le auto ibride con motore termico a benzina il gruppo che in media inquina di più, tra i top 10, è la Daimler AG.

Il quarto *sheet*, denominato **riduzioni di CO2**, è sviluppato allo stesso modo dello *sheet* appena descritto riguardante le emissioni di CO2, con l'attenzione in questo caso incentrata sulle riduzioni medie di CO2 da parte dei top 10 gruppi automobilistici.

Anche qui si è inserita una tabella, quella mostrata in **Figura 4.66**, per classificare in modo decrescente i top 10 gruppi automobilistici (definiti in precedenza) in base alle riduzioni medie di emissioni di CO2 (g/km).

Va detto che, in questo caso, per il calcolo delle riduzioni medie di emissioni, sono state considerate solamente le automobili per cui le nuove tecnologie hanno consentito di ridurre le emissioni.

Gruppo automobilistico	Riduzioni CO2 medie (g/km) ↓	Emissioni CO2 medie (g/km)	Peso/potenza medio	Cilindrata media (cc)
BMW GROUP	1,67	118,92	13,48	1.937,60
HONDA MOTOR COMPANY	1,40	127,95	12,62	1.500,93
DAIMLER AG	1,37	123,11	15,18	1.810,19
GROUP PSA	1,18	108,30	16,37	1.394,25
MAZDA MOTOR CORPORATION	1,09	135,91	13,67	1.909,91
VOLKSWAGEN GROUP	1,00	115,64	15,11	1.462,54
ASTON MARTIN LAGONDA	0,00	292,36	4,62	5.358,24
SUZUKI MOTOR CORPORATION	0,00	111,80	13,00	1.235,26
TOYOTA MOTOR CORPORATION	0,00	106,27	16,98	1.743,45
FERRARI N.V.	0,00	317,21	3,87	5.013,54

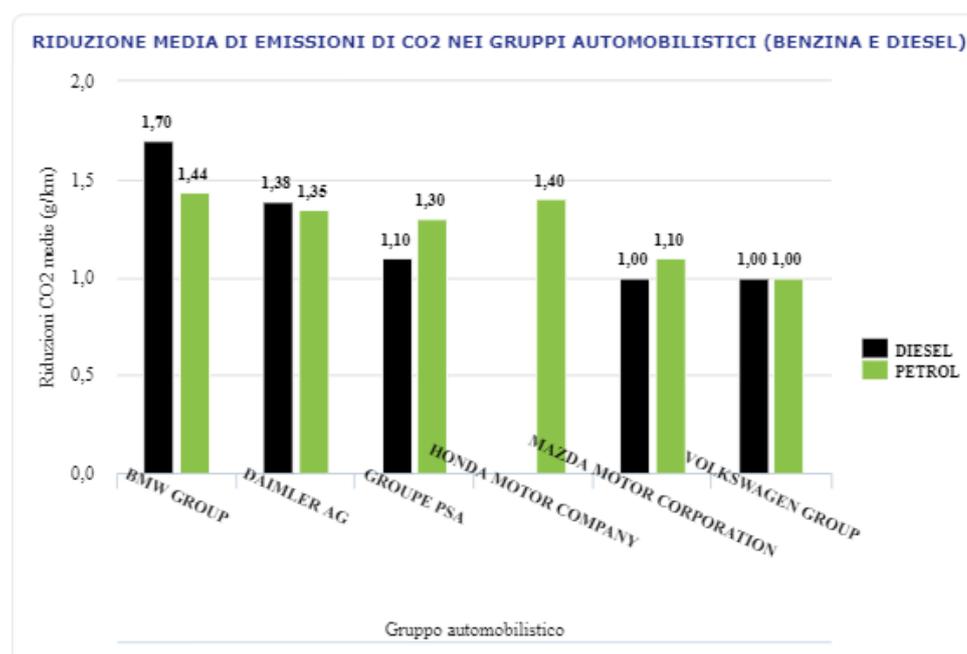
Figura 4.66: Ranking gruppi per riduzione media di emissioni

Tra i dati analizzati, figurano solamente sei gruppi nella riduzione delle emissioni delle automobili; tra questi, quelli maggiormente attivi in investimenti e utilizzo di tecnologie innovative per la riduzione dell'inquinamento (in termini di CO<sub>2</sub>) sembrano essere il gruppo Bmw, il gruppo giapponese Honda e il gruppo Daimler AG, seppur la riduzione media di CO<sub>2</sub> è comunque piccola.

Nella tabella, oltre alle riduzioni medie, sono presenti anche le emissioni medie per gruppo, il rapporto tra peso e potenza medi e la cilindrata media.

A questo proposito, tra queste misure e la riduzione di emissioni non sembra esserci alcuna correlazione.

La **Figura 4.67** e la **Figura 4.68** offrono un confronto sulle riduzioni medie di emissioni, tra tipologie di alimentazione simili e per i soli gruppi che figurano nella riduzione di emissioni.

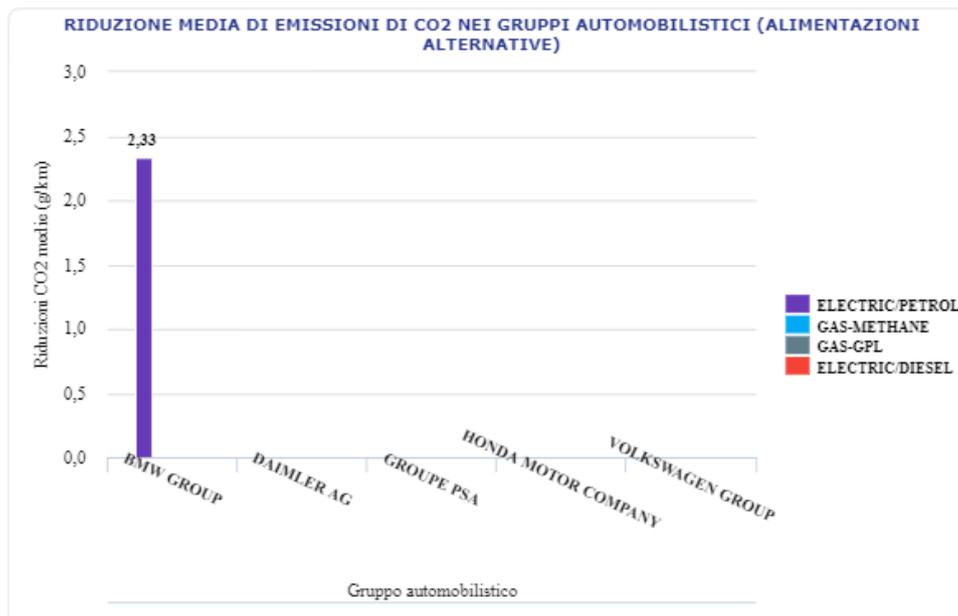


**Figura 4.67:** Riduzione media emissioni per benzina e diesel

Sia per le auto a diesel che per le auto a benzina, le maggiori riduzioni medie di emissioni appartengono al gruppo Bmw che è anche al primo posto nella classifica dei gruppi per riduzioni medie di emissioni, mostrata prima in **Figura 4.66**.

Al secondo posto, per le auto a benzina, vi è il gruppo Mazda, tra i meno attivi nel fronte riduzioni, mentre per le auto a diesel, al secondo posto, vi è il gruppo Daimler AG.

Per quanto concerne le alimentazioni alternative, come mostrato in **Figura 4.68**, non figurano riduzioni di emissioni se non per il gruppo Bmw, relativamente alle sole auto ibride con motore termico alimentato a benzina.



**Figura 4.68:** Riduzione media emissioni per le alimentazioni alternative

Nel quinto ed ultimo *sheet*, denominato **case costruttrici**, si è cercato di evidenziare, tramite alcuni *widget*, le maggiori emissioni e riduzioni di emissioni prendendo in considerazione, questa volta, tutte le case costruttrici afferenti a diversi gruppi automobilistici.

In **Figura 4.69** e in **Figura 4.70** è presente la stessa tabella esposta prima, in questo caso con le case costruttrici al posto dei gruppi, ordinate in modo decrescente sulla base alle emissioni medie di CO2 delle relative autovetture.

Casa costruttrice	Emissioni CO2 medie (g/km) ↓	Riduzioni CO2 medie (g/km)	Peso/potenza medio	Cilindrata media (cc)
BUGATTI	521,75	0,00	1,99	7.993,00
ROLLS-ROYCE	330,11	0,00	5,78	6.603,89
FERRARI	317,21	0,00	3,87	5.013,54
ASTON MARTIN	292,36	0,00	4,62	5.358,24
BENTLEY	277,47	0,00	6,36	4.792,10
MASERATI	262,71	0,00	10,04	2.984,71
CADILLAC	261,04	0,00	7,78	4.830,32
MCLAREN	252,77	0,00	3,33	3.861,17
CHEVROLET	245,38	0,00	6,67	5.267,15
LOTUS	203,81	0,00	5,52	2.628,81

Pagina 1 1 - 10 di 45 < >

**Figura 4.69:** Ranking case costruttrici per emissioni medie - 1

RANKING CASE COSTRUTTRICI PER EMISSIONI MEDIE DI CO2				
Casa costruttrice	Emissioni CO2 medie (g/km) ↓	Riduzioni CO2 medie (g/km)	Peso/potenza medio	Cilindrata media (cc)
PEUGEOT	104,59	0,00	16,08	1.407,69
RENAULT	104,58	0,00	18,11	1.329,30
TOYOTA	101,59	0,00	17,86	1.668,83
SMART	84,61	1,20	17,97	952,35
TESLA	0,00	0,00	23,21	0,00

Pagina 5    41 - 45 di 45

**Figura 4.70:** Ranking case costruttrici per emissioni medie - 2

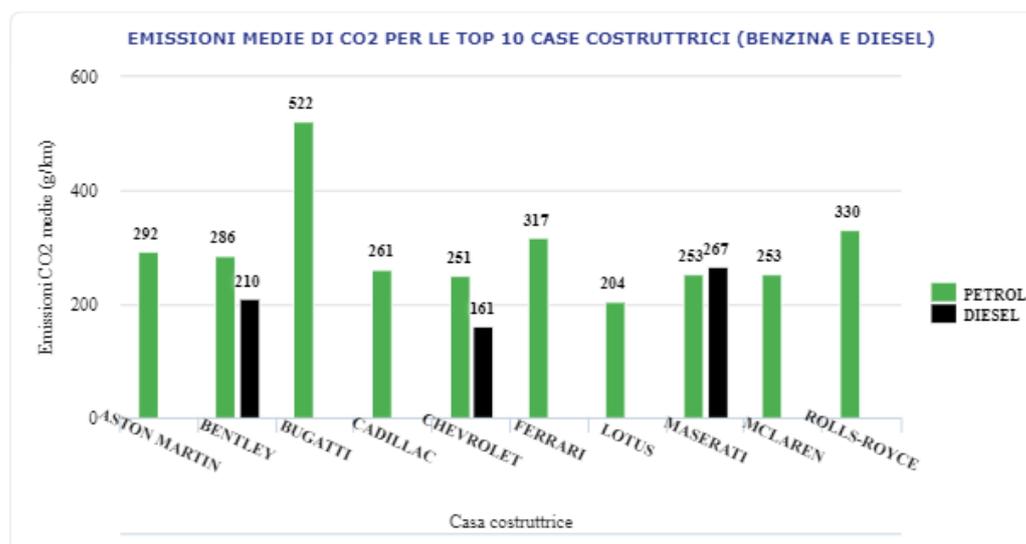
Tuttavia, soprattutto le automobili di alcuni brand sportivi e di lusso risultano essere in cima alla classifica delle case con le maggiori emissioni medie di CO<sub>2</sub>; tra queste ci sono la Bugatti, Rolls-Royce, Ferrari, Aston Martin, Bentley e molte altre.

Queste case costruttrici sembrano essere assai lontane dai limiti massimi fissati dall'Unione Europea per il 2021 e 2030 (rispettivamente, in media, 95 g/km e 59 g/km) e, per questo motivo, alle case più piccole, con meno di 1.000 unità annue prodotte, viene concessa l'esenzione dai limiti, mentre a quelle con meno di 10.000 unità (come la Ferrari) viene concessa la possibilità di chiedere una deroga<sup>[11]</sup>.

Infine, tra le case con le minori emissioni medie, mostrate in **Figura 4.70**, ai primi posti ci sono Tesla e Smart, le uniche pronte e sotto al limite imposto dall'UE per il 2021, con rispettivamente zero CO<sub>2</sub> emessa e 85 g/km medi.

Poi, tra le case con emissioni medie relativamente basse si ha Toyota, poco al di sopra del limite dei 95 g/km del 2021, e a seguire Renault e Peugeot.

La **Figura 4.71**, come in precedenza, offre un confronto tra le emissioni medie di CO<sub>2</sub> per le top 10 case (in termini di emissioni) e per tipologie di alimentazione simili.



**Figura 4.71:** Emissioni medie per benzina e diesel per le top 10 case

Come si evince dal grafico, le top 10 case costruttrici (in termini di inquinamento), costituite soprattutto da brand sportivi e di lusso, sono caratterizzate dalla prevalenza di autovetture a benzina piuttosto che a diesel; le uniche case con automobili a diesel, che figurano nel grafico, sono Bentley e Maserati con la seconda caratterizzata da emissioni medie maggiori.

Per quanto concerne le autovetture a benzina è ovviamente la Bugatti a inquinare di più, seguita da Rolls-Royce e Ferrari, così come mostrato nella tabella descritta prima.

Inoltre, non è stato inserito il confronto tra le alimentazioni alternative in quanto, per le top 10 case in questione, non sono presenti e non vengono prodotte, nel campione analizzato così come nella realtà, automobili ibride, a gas e puramente elettriche.

La **Figura 4.72** mostra il ranking delle case costruttrici ordinate in modo decrescente per riduzioni medie di emissioni.

RANKING CASE COSTRUTTRICI PER RIDUZIONE MEDIA DI EMISSIONI DI CO2				
Casa costruttrice	Riduzioni CO2 medie (g/km) ↓	Emissioni CO2 medie (g/km)	Peso/potenza medio	Cilindrata media (cc)
BMW	1,91	120,44	13,34	2.038,74
HONDA	1,40	127,95	12,62	1.500,93
MERCEDES	1,37	132,58	14,49	1.994,66
SMART	1,20	84,61	17,97	952,35
OPEL	1,18	120,18	16,64	1.397,36
MAZDA	1,09	135,91	13,67	1.909,91
MINI	1,06	113,13	13,95	1.608,18
VOLKSWAGEN	1,00	117,59	15,52	1.503,01
SKODA	1,00	111,99	14,83	1.377,11
LOTUS	0,00	203,81	5,52	2.628,81

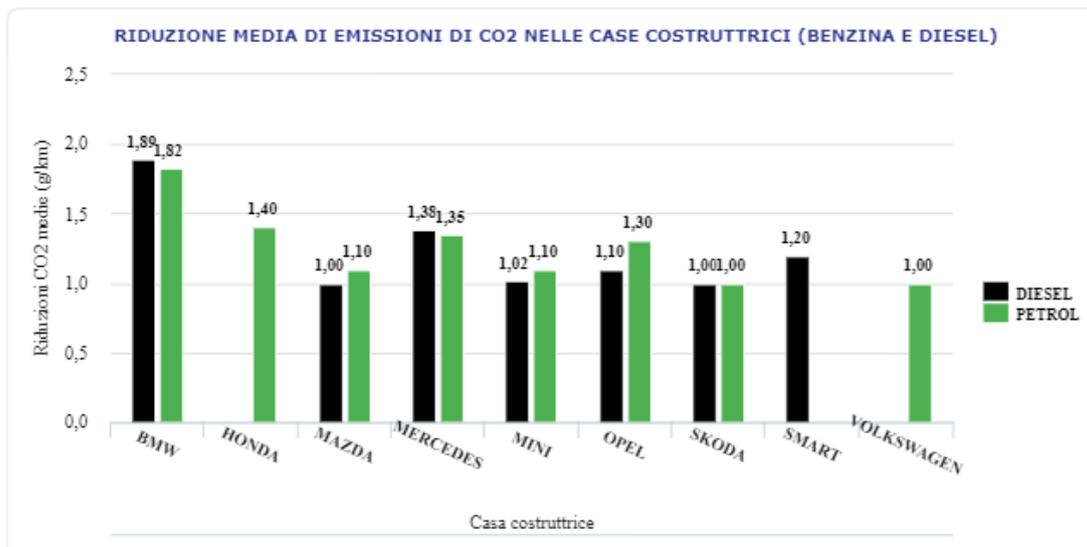
Pagina 1 ▼ 1 - 10 di 46 < >

**Figura 4.72:** Riduzione media di emissioni per le top 10 case

Le case maggiormente attive e impegnate negli investimenti di tecnologie innovative per la lotta all'inquinamento e con la maggior riduzione media di emissioni di CO2 risultano essere Bmw, come già visto in precedenza, quindi Honda e Mercedes.

Tuttavia, le case Bmw e Mercedes fanno parte entrambe di gruppi automobilistici con all'attivo le emissioni medie di CO2 tra le più elevate ma, secondo quanto riportato in tabella, risultano tra le più impegnate anche nella riduzione delle emissioni generate.

In **Figura 4.73** e in **Figura 4.74** viene mostrato un confronto tra le riduzioni medie di CO2 emessa per le varie case costruttrici e tipologie di alimentazione simili.

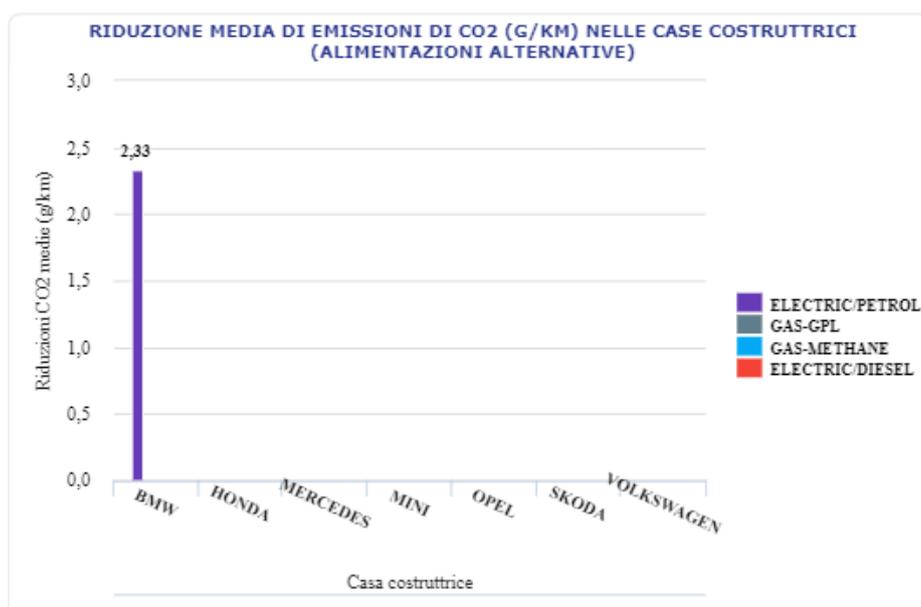


**Figura 4.73:** Riduzione media emissioni per benzina e diesel

Per quanto riguarda sia le auto a diesel che quelle a benzina la Bmw risulta la casa più attiva nella riduzione media di emissioni di CO<sub>2</sub>, occupando la prima posizione nel ranking generale in **Figura 4.72**.

Inoltre, per le auto a diesel, a seguire vi è la casa Mercedes, mentre per le autovetture a benzina la casa giapponese Honda.

Infine, in **Figura 4.74**, vengono mostrate le eventuali riduzioni medie per le varie case e per le diverse alimentazioni alternative.



**Figura 4.74:** Riduzione media emissioni per le alimentazioni alternative

Come visibile dal grafico, l'unica casa costruttrice ad effettuare riduzioni di emissioni per alimentazioni alternative è la Bmw, relativamente alle auto ibride con motore termico alimentato a benzina.

## 4.5.2 Analisi per nazioni (secondo cockpit)

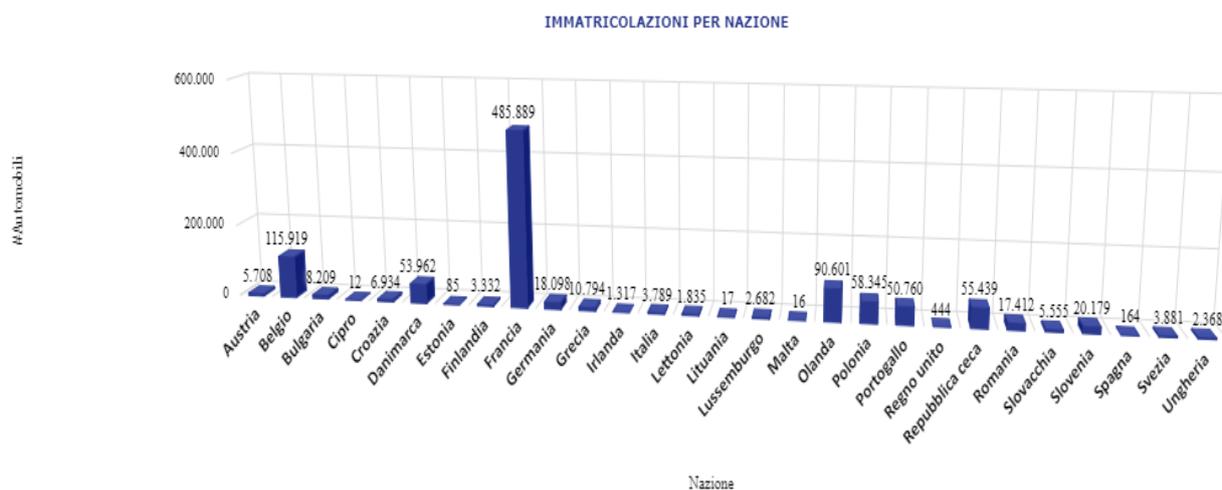
Nel secondo cruscotto si è deciso, per completezza, di realizzare le stesse analisi eseguite nel primo *cockpit* prendendo in considerazione, al posto dei gruppi automobilistici, i 28 paesi dell'Unione Europea.

Tuttavia, sono stati definiti quattro differenti *sheet*, tra cui: **overview**, **quota autovetture**, **emissioni di CO2**, **riduzioni di CO2**.

Per quanto riguarda l'**overview**, si è deciso di presentare le stesse informazioni mostrate nel primo *cockpit*, pertanto, in questo paragrafo, questo *sheet* non viene descritto.

Nel secondo *sheet*, denominato **quota autovetture**, si è voluta dare un'idea sulla quota annua delle nuove automobili per ciascun paese, considerando anche il peso dei top 10 gruppi in ciascuna nazione, così come sulla quota di autovetture per ciascun tipo di alimentazione e stato.

La **Figura 4.75** cerca di dare un'idea, per ciascun paese dell'UE, sulle automobili di nuova immatricolazione.



**Figura 4.75:** Nuove automobili per nazione

Tuttavia, nel campione di dati preso in considerazione, la quota di automobili nei vari paesi dell'Unione Europea non rispecchia il reale andamento del mercato.

Infatti, in questo caso, i paesi con il maggior numero di immatricolazioni risultano essere Francia, Belgio, Danimarca, Olanda, Polonia, Portogallo, Repubblica ceca, Germania e Slovenia.

Al contrario, nel mercato reale e secondo stime attendibili<sup>[18]</sup>, le cinque principali piazze del mercato europeo delle automobili, in termini di volume venduto, sono Spagna, Francia, Germania, Regno Unito e Italia.

Sia nel 2018 che nel 2017 la Germania si è mantenuta stabile mentre Francia e Spagna hanno mantenuto le proprie vendite in crescita, infine il Regno Unito ha registrato un forte calo in entrambi gli anni; nel 2017 l'Italia ha registrato una crescita quasi pari all'8% al contrario del 2018 dove ha registrato una diminuzione rispetto all'anno precedente.

Per quanto concerne i paesi più colpiti dalla crisi del 2007 e 2008, il mercato greco e quello Portoghese hanno evidenziato, nel 2017 e 2018, una forte crescita delle vendite delle automobili.

In **Figura 4.76** e in **Figura 4.77** è presente la *cross table* che mostra il volume di automobili vendute per ciascun paese e gruppo automobilistico appartenente ai top 10.

Prendendo in considerazione i principali stati per volume di autovetture vendute, in Germania le automobili più vendute sono dei gruppi Daimler AG e Volkswagen che insieme conquistano più della metà della quota di mercato locale, con le auto preferite in assoluto che sono la Volkswagen Golf e la Mercedes classe C; in Italia le maggiori vendite provengono dal gruppo di casa FCA, così come PSA e Volkswagen, che insieme raggiungono circa la metà della quota di mercato mentre l'auto più venduta agli Italiani risulta proprio la Fiat panda.

In Francia le automobili preferite appartengono al gruppo Renault-Nissan-Mitsubishi e al gruppo PSA con una quota di mercato complessiva superiore al 50%, mentre le automobili più vendute sono la Renault Clio, la Peugeot 208 e la Citroen C3, queste ultime due appartenenti proprio al gruppo PSA.

In paesi minori con un mercato in crescita come Grecia e Portogallo le automobili più vendute sono, rispettivamente, del gruppo Renault e Volkswagen per la Grecia e del gruppo Renault e PSA per il Portogallo; per quest'ultimo le auto preferite sono la Renault Clio e Megane e la Peugeot 208, mentre per la Grecia, a sorpresa, la Toyota Yaris.

Per i paesi nordici invece, in Finlandia le automobili più vendute appartengono al gruppo Volkswagen con la Skoda Octavia e la Volkswagen Golf nelle prime tre posizioni delle auto preferite; anche in Svezia le maggiori vendite provengono dal gruppo Volkswagen mentre l'auto preferita è la Volvo XC60.

Il gruppo Volkswagen spicca anche in paesi dell'est Europa come Croazia ed Estonia con la Skoda Octavia auto più venduta, ed in Romania dove l'auto più venduta è, però, la Dacia Logan<sup>[20]</sup>.

In generale, in base a quanto appena evidenziato, spesso le automobili più vendute appartengono ai marchi di casa.

IMMATRICOLAZIONI PER I TOP GRUPPI AUTOMOBILISTICI E NAZIONE										
NAZIONE	GRUPPI AUTOMOBILISTICI									
	BMW GROUP	DAIMLER AG	FCA	FORD MOTOR COMPANY	GROUPE PSA	HYUNDAI-KIA AUTOMOTIVE GROUP	RENAULT-NISSAN-MITSUBISHI ALLIANCE	TOYOTA MOTOR CORPORATION	VOLKSWAGEN GROUP	VOLVO CARS
Austria	369	555	2		332	57	2.790	311	1.105	74
Belgio	11.385		15.552	11.669	15.784	5.051	28.080		16.195	4.088
Bulgaria	82	589			1.218	1.289	4.234	2	452	12
Cipro									12	
Croazia	431		302		2.462	771	61		2.333	154
Danimarca	1.320			10.680	7.919	3.495	3.412		21.896	112
Estonia									85	
Finlandia	10				230		57		2.645	386
Francia	19.262	22.739	27.576	21.440	133.872	11.541	148.098	22.648	67.795	2.342
Germania	1.493	7.466	10		3.243	128	1.262	4	4.095	
Grecia	1.487		651		1.146	1.494	1.868		2.672	287
Irlanda	166				167	74	116		705	36
Italia	606	202	601		714	62	231	440	708	1
Lettonia			2		272	189	449		744	17
Lituania									17	
Lussemburgo	156	1.197			272	87	104	197	450	65
Malta									16	
Olanda	3.031	15.882	6.123		13.749	6.641	7.497	4.178	24.155	2.720
Polonia	3.338	1	829		5.134	7.166	4.982	8.661	21.949	1.396

**Figura 4.76:** Immatricolazioni per nazione e top 10 gruppi automobilistici - 1

Polonia	3.338	1	829		5.134	7.166		4.982	8.661	21.949	1.396
Portogallo	6.440	1.814	377		11.752	2.074		16.669	1.527	6.737	942
Regno unito										444	
Repubblica ceca	534		109	8.021	3.277	4.236		14.871		23.109	1
Romania	764		2	288	1.637	1.396		700	4.041	6.833	57
Slovacchia	391				980	113		59		3.351	100
Slovenia	211	803		3.105	1.659	728		10.174		3.204	13
Spagna										164	
Svezia					245	57		79		2.863	195
Ungheria	232				600	137		160		1.060	

**Figura 4.77:** Immatricolazioni per nazione e top 10 gruppi automobilistici - 2

In **Figura 4.78** e in **Figura 4.79** viene riportata la quantità di autovetture per nazione e tipo alimentazione.

A questo proposito, al 2017, i paesi più virtuosi e sensibili in tema di inquinamento, con la maggior quota di auto *green* sul totale, (elettriche, ibride e a gas) risultano essere: Finlandia (circa il 4%), Italia (circa il 3%), Olanda (circa il 2%), Portogallo (circa il 2,5%) e Svezia (circa il 7,5%) che rappresenta il paese più *green* ed ecologico nei trasporti e nella mobilità.

IMMATRICOLAZIONI PER NAZIONE E TIPO ALIMENTAZIONE								
↑ NAZIONI	↑ TIPO ALIMENTAZIONE							TOTALE
	DIESEL	ELECTRIC	ELECTRIC/DIESEL	ELECTRIC/PETROL	GAS-GPL	GAS-METHANE	PETROL	
Austria	2.980	45		17		18	2.648	5.708
Belgio	51.989	194	1	1.487	7	525	61.716	115.919
Bulgaria	4.480			7	261	44	3.417	8.209
Cipro							12	12
Croazia	4.021		8	2	33		2.870	6.934
Danimarca	13.800	67		58		9	40.028	53.962
Estonia	15						70	85
Finlandia	1.551		12	44		69	1.656	3.332
Francia	234.842	5.536		2.322	288	160	242.741	485.889
Germania	9.468	62			62	130	8.376	18.098
Grecia	3.692	2		23		49	7.028	10.794
Irlanda	866			7			444	1.317
Italia	2.291	7		7	77	32	1.375	3.789
Lettonia	856	2		4	13		960	1.835
Lituania	14						3	17
Lussemburgo	1.485					2	1.195	2.682
Malta	6						10	16
Olanda	15.219	1.498	7	174	17	249	73.437	90.601

**Figura 4.78:** Immatricolazioni per nazione e tipo alimentazione - 1

Olanda	15.219	1.498	7	174	17	249	73.437	90.601
Polonia	19.932			3	593	1	37.816	58.345
Portogallo	30.295	587	11	600	135	3	19.129	50.760
Regno unito	237						207	444
Repubblica ceca	11.854	48		71	376	312	42.778	55.439
Romania	6.763	16		28			10.605	17.412
Slovacchia	2.958				38	28	2.531	5.555
Slovenia	8.355	107			171	11	11.535	20.179
Spagna	84					8	72	164
Svezia	1.942	157	12	39		84	1.647	3.881
Ungheria	1.142	10	4	19			1.193	2.368

**Figura 4.79:** Immatricolazioni per nazione e tipo alimentazione - 2

Per quanto riguarda il terzo *sheet*, denominato **emissioni di CO2**, l'analisi si concentra sulle emissioni medie di CO2 per nazione, similmente a quanto fatto per i gruppi automobilistici.

In **Figura 4.80** e in **Figura 4.81** è mostrata la tabella che classifica, in modo decrescente, i paesi dell'UE per emissioni medie di CO2 (g/km).

RANKING NAZIONI PER EMISSIONI MEDIE DI CO2				
Nazione	Emissioni CO2 medie (g/km) ↓	Riduzioni CO2 medie (g/km)	Peso/potenza medio	Cilindrata media (cc)
Germania	141,42	1,42	13,79	2.004,56
Estonia	140,46	0,00	10,54	1.747,44
Lussemburgo	137,45	1,15	13,35	2.054,46
Slovacchia	131,67	1,03	14,45	1.816,12
Ungheria	131,10	0,00	13,54	1.862,13
Italia	130,48	2,01	15,91	1.818,18
Lettonia	130,16	1,40	14,86	1.683,94
Polonia	128,94	1,43	10,16	1.656,19
Svezia	127,77	1,00	12,44	1.924,51
Lituania	124,71	0,00	15,56	1.733,24

Pagina 1    1 - 10 di 28

**Figura 4.80:** Ranking nazioni per emissioni medie - 1

RANKING NAZIONI PER EMISSIONI MEDIE DI CO2				
Nazione	Emissioni CO2 medie (g/km) ↓	Riduzioni CO2 medie (g/km)	Peso/potenza medio	Cilindrata media (cc)
Cipro	115,25	0,00	12,85	1.245,25
Croazia	114,64	0,00	17,01	1.537,22
Olanda	112,09	1,37	15,80	1.427,65
Grecia	111,26	1,37	16,74	1.323,69
Francia	110,22	1,40	16,77	1.455,45
Danimarca	109,77	1,83	15,70	1.354,37
Malta	107,06	0,00	14,09	1.317,75
Portogallo	102,89	1,63	17,30	1.428,40

Pagina 3    21 - 28 di 28

**Figura 4.81:** Ranking nazioni per emissioni medie - 2

Tra tutti i paesi Europei con le maggiori emissioni medie spicca, come si evince dalla tabella, la Germania insieme a diverse nazioni dell'est Europa, tra cui: Estonia, Slovacchia, Ungheria, Lettonia, Polonia e Lituania.

A questo proposito, secondo Eco Expert<sup>[21]</sup>, i paesi più inquinanti sono proprio quelli dell'Europa dell'est, anche a causa della massiccia presenza di auto circolanti vecchie e di scarsa qualità.

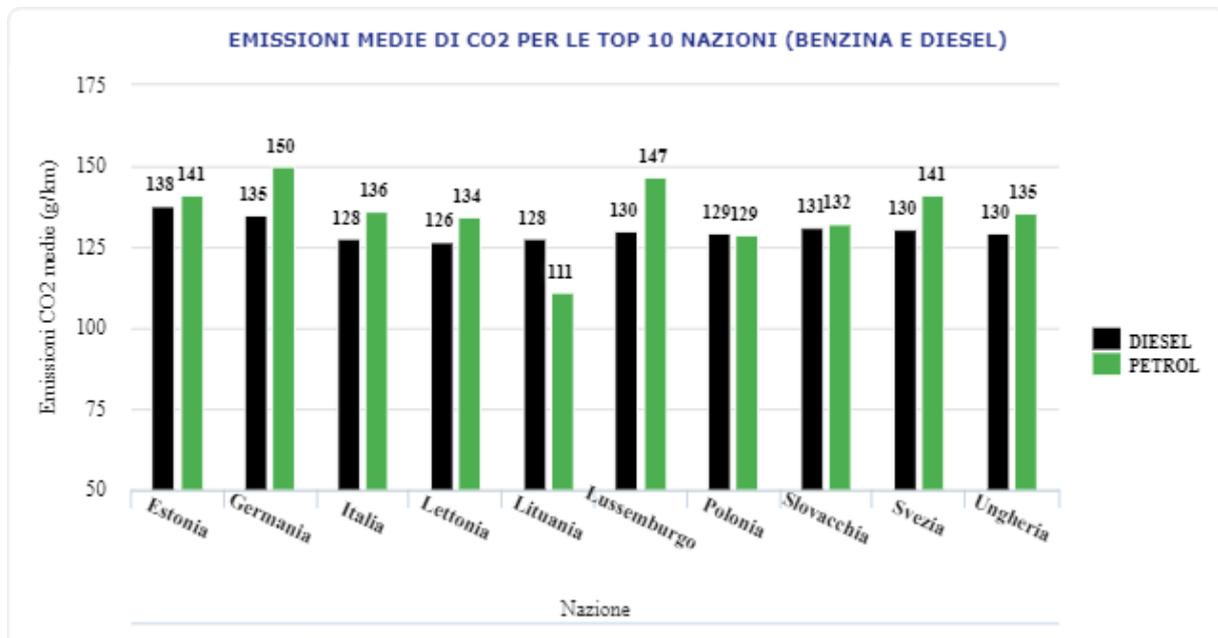
Pertanto, l'Europa dell'est viene anche definita "regione più sporca d'Europa".

Al contrario, secondo quanto riportato dai dati analizzati nel campione, i paesi più ecologici e con le minori emissioni medie di CO2 sono quelli nordici come la Danimarca, Malta, Olanda, Francia e Portogallo.

Infine, nella tabella, sono riportate anche le riduzioni medie di CO2 così come il rapporto tra peso e potenza medi delle autovetture di uno specifico paese e la cilindrata media; tuttavia, non sembra esserci alcuna correlazione tra queste misure e le emissioni medie di CO2.

In **Figura 4.82** e in **Figura 4.83** viene riportato, in aggiunta, un confronto tra le emissioni medie per paese e tipologie di alimentazione simili.

Tuttavia, in questa speciale analisi, sono state considerate le nazioni con le maggiori emissioni medie globali.

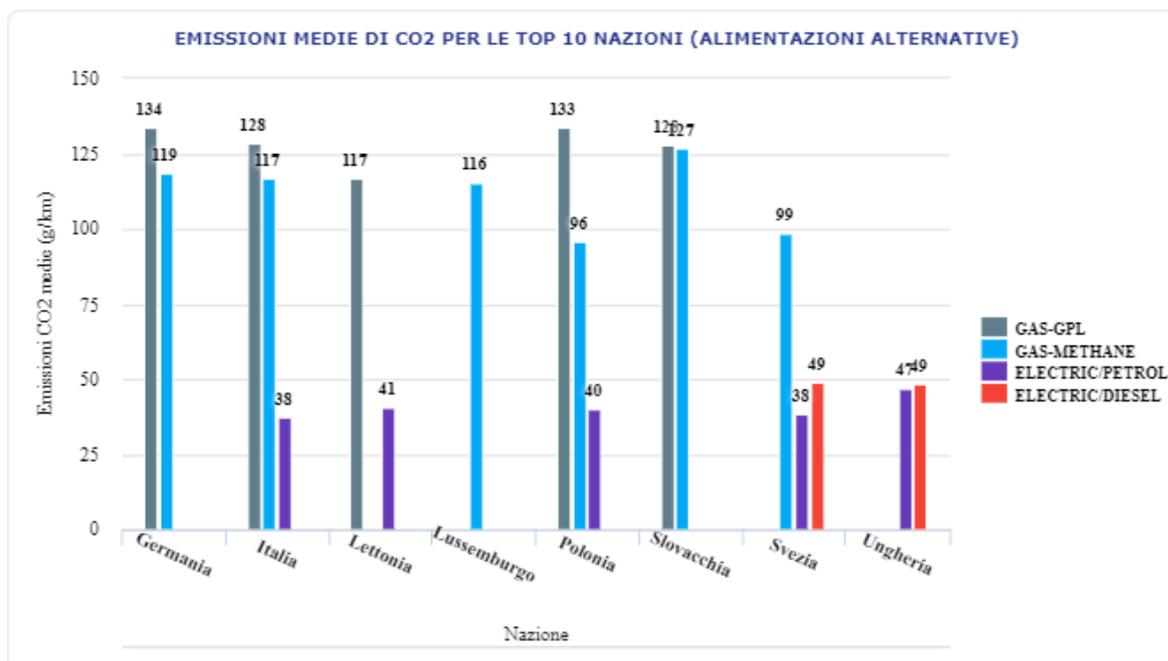


**Figura 4.82:** Emissioni medie per benzina e diesel nelle top 10 nazioni

Considerando le auto a diesel è l'Estonia lo stato ad inquinare di più in termini di CO2 media emessa allo scarico, seguita dalla Germania; quelle ad inquinare di meno sono, al contrario, Italia e Lettonia.

Inoltre, l'unico paese in cui le auto a diesel inquinano più di quelle a benzina è la Lituania.

Invece, per quanto riguarda le auto a benzina, è sempre la Germania ad essere in testa, seguita dal Lussemburgo e dall'Estonia.



**Figura 4.83** Emissioni medie per le alimentazioni alternative nelle top 10 nazioni

Per quanto riguarda le alimentazioni alternative, per le auto a gas-GPL, quelle ad emettere maggior CO2, è la Germania ad inquinare di più, di pochissimo sopra la Polonia; per le auto a gas metano è la Slovacchia ad emettere le maggiori emissioni inquinanti di CO2 seguita dalla Germania.

Per le auto con le minori emissioni, ossia quelle elettriche con motore termico, per quelle a benzina è la Lettonia ad emettere, in media, maggior emissioni, con l'Italia che è quella ad inquinare meno.

Infine, tra i paesi in cui figurano le auto ibride con motore termico a diesel, sia l'Ungheria che la Svezia emettono, in media, la stessa quantità di CO2.

L'ultimo *sheet*, denominato **riduzioni di CO2**, ricalca quello del primo *cockpit* e quello appena visto, dove, però, vengono analizzate le riduzioni medie di CO2 delle nuove automobili di ciascuna nazione.

In **Figura 4.84** viene mostrato il ranking delle nazioni prendendo in considerazione le riduzioni medie di CO2 delle autovetture per cui le tecnologie innovative hanno permesso di ottenere una riduzione delle emissioni.

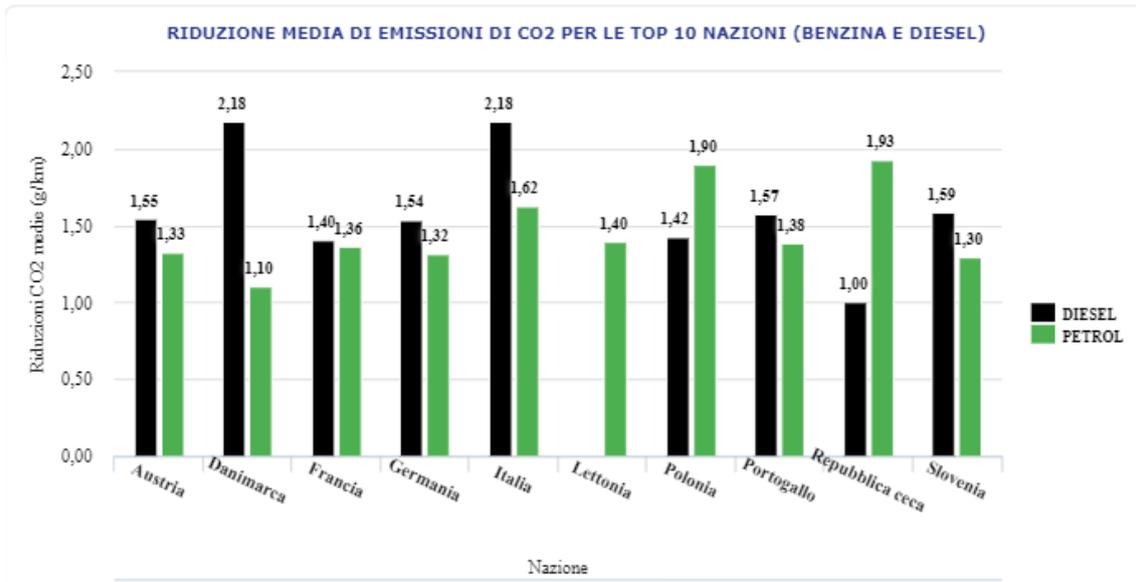
RANKING NAZIONI PER RIDUZIONE MEDIA DI EMISSIONI DI CO2				
Nazione	Riduzioni CO2 medie (g/km) ↓	Emissioni CO2 medie (g/km)	Peso/potenza medio	Cilindrata media (cc)
Italia	2,01	130,48	15,91	1.818,18
Danimarca	1,83	109,77	15,70	1.354,37
Repubblica ceca	1,74	122,55	16,17	1.380,59
Portogallo	1,63	102,89	17,30	1.428,40
Slovenia	1,54	118,22	16,41	1.390,89
Polonia	1,43	128,94	10,16	1.656,19
Austria	1,43	124,19	16,14	1.612,75
Germania	1,42	141,42	13,79	2.004,56
Francia	1,40	110,22	16,77	1.455,45
Lettonia	1,40	130,16	14,86	1.683,94

**Figura 4.84:** Riduzioni medie di emissione per le top 10 nazioni

Al primo posto tra le nazioni, per la riduzione media di emissioni di CO<sub>2</sub>, c'è l'Italia, seguita da Danimarca e Repubblica ceca.

Tra le top 10 nazioni per riduzione di emissioni c'è anche la Germania, che risulta prima per emissioni medie di CO<sub>2</sub>, come mostrato prima, così come Polonia e Lettonia.

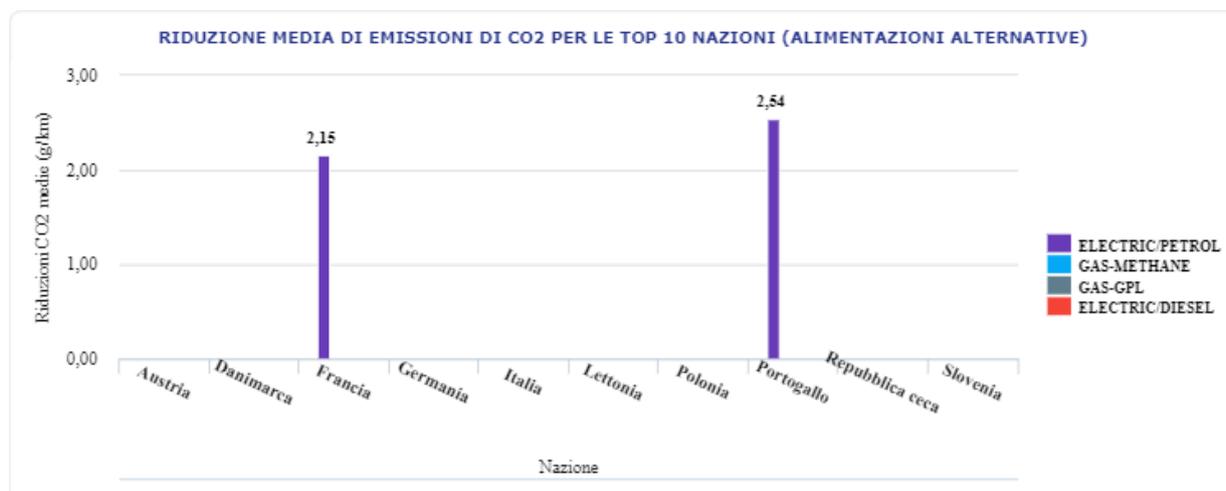
In **Figura 4.85** e in **Figura 4.86** sono mostrate le riduzioni medie di emissione per le top 10 nazioni e tipologie di alimentazione simili.



**Figura 4.85:** Riduzioni medie di emissione per benzina e diesel e le top 10 nazioni

Innanzitutto, per la prevalenza dei casi analizzati, vengono effettuate maggiori riduzioni di emissione per le auto a diesel piuttosto che per quelle a benzina.

Tuttavia, per quanto riguarda le auto a diesel, sia la Danimarca che l'Italia riducono, mediamente, il maggior quantitativo di CO<sub>2</sub> seguite dall'Austria; al contrario, per le auto a benzina, sono la Repubblica ceca, la Polonia e la Lettonia a ridurre maggiormente le emissioni di CO<sub>2</sub>.



**Figura 4.86:** Riduzioni medie di emissione per le alimentazioni alternative e le top 10 nazioni

Per quanto riguarda le alimentazioni alternative, le uniche nazioni per cui figurano autovetture con riduzioni di emissione sono Francia e Portogallo, le quali presentano entrambe elevate riduzioni medie di CO<sub>2</sub>, relativamente alle auto ibride con motore termico alimentato a benzina.

### 4.5.3 Analisi per caratteristiche strutturali delle autovetture (terzo cockpit)

Nel terzo ed ultimo *cockpit* realizzato, come già anticipato in precedenza, si è voluto comunque analizzare le emissioni medie di CO<sub>2</sub> delle autovetture sulla base di caratteristiche strutturali delle auto, tra cui; **cilindrata (in centimetri cubici cc)**, **potenza (in kilowatt KW)** e **peso (in kilogrammi kg)**.

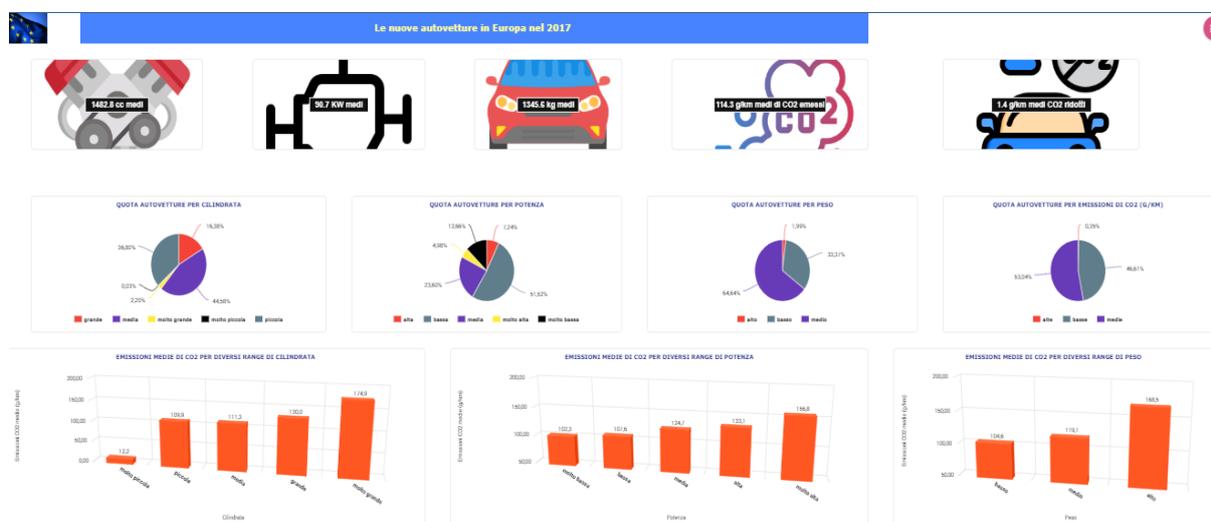
Inoltre, si è cercato di evidenziare e scoprire possibili correlazioni tra le suddette caratteristiche strutturali e le emissioni di CO<sub>2</sub> delle autovetture.

Tuttavia, come illustrato nel paragrafo 4.3, tutte le caratteristiche strutturali sopra riportate sono state discretizzate, ossia suddivise in più intervalli (es. basso, medio, alto) in modo tale da ridurre la cardinalità del dominio e rendere l'analisi più comprensibile, semplificata ed intuitiva.

Il terzo *cockpit*, qui descritto, è caratterizzato da tre *sheet*, tra cui: **overview**, **emissioni per caratteristiche auto**, **emissioni per caratteristiche e gruppi**.

Nel primo *sheet*, ossia l'**overview**, sono stati inseriti tutta una serie di *widget* con lo scopo di fornire informazioni molto generali relative al *topic* in questione, ossia le caratteristiche strutturali unite alle emissioni di CO<sub>2</sub> delle auto.

In **Figura 4.87** è mostrato, solamente per dare un'idea, il *layout* dell'**overview**.



**Figura 4.87:** *Layout overview*

I cinque *widget* HTML, presenti nell'*overview* in alto e in **Figura 4.88** e successive quattro, riportano, relativamente alle autovetture del campione, la cilindrata media, la potenza media, il peso medio, le emissioni medie e le riduzioni medie.



Figura 4.88 Cilindrata media

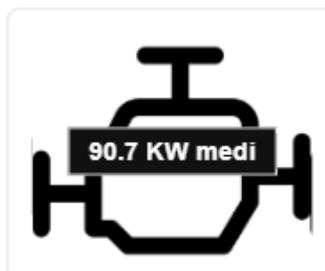


Figura 4.89 Potenza media

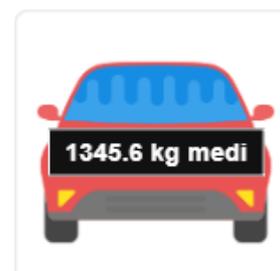


Figura 4.90 Peso medio



Figura 4.91 Emissioni medie



Figura 4.92 Riduzioni medie

I quattro grafici a torta, visibili in **Figura 4.93** e successive tre figure, riportano il volume di automobili per ciascuna delle caratteristiche strutturali e per le emissioni di CO<sub>2</sub>; cioè, viene riportata la quota di autovetture sul totale per ciascun *range* di cilindrata, potenza, peso ed emissioni.



Figura 4.93: Quota auto per cilindrata

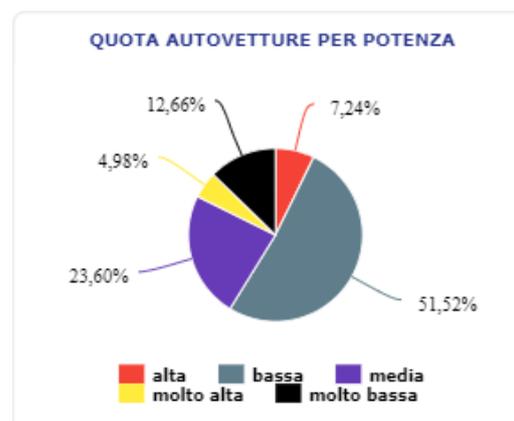


Figura 4.94: Quota auto per potenza

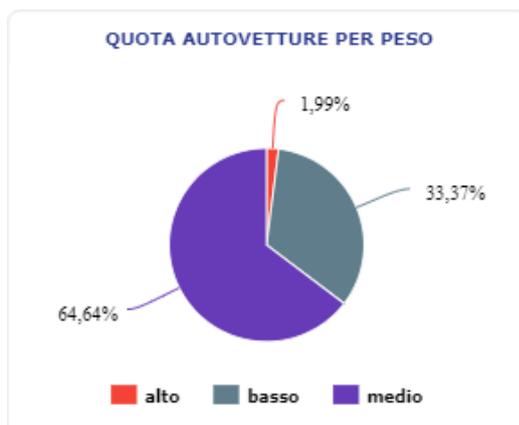


Figura 4.95: Quota auto per peso

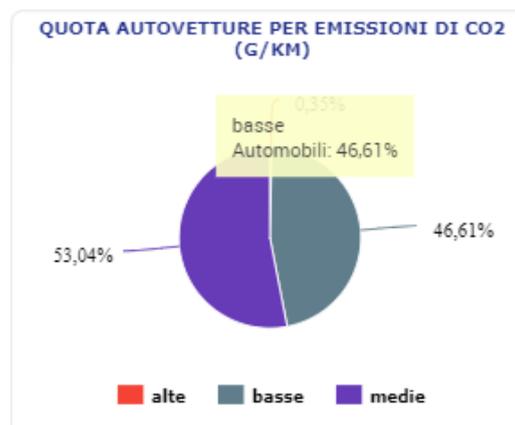


Figura 4.96: Quota auto per emissioni

Osservando i quattro grafici a torta, emerge che circa l'80% delle automobili sono di medio-piccola cilindrata (fino a 1800cc), il 75% di medio-bassa potenza (fino a 125 kw, circa 170 cavalli), circa il 98% di peso medio-basso (fino a 2000 kg, quindi utilitarie, berline e station wagon) e quasi l'intera totalità delle autovetture emette emissioni di CO2 considerate medio-basse (fino a 220 g/km).

Infine, nei diagrammi a barre riportati in **Figura 4.97** e successive due figure, vengono riportate le emissioni medie di CO2, rispettivamente, per cilindrata, potenza e peso.

Pertanto, osservando i grafici, è evidente che al crescere della cilindrata, della potenza e del peso delle automobili aumentano le emissioni medie di CO2.

Questo è un risultato molto interessante in quanto mette in relazione le emissioni di CO2 con alcune caratteristiche strutturali delle autovetture su cui è possibile intervenire per migliorare, qualitativamente, gli inquinanti emessi così come i consumi.

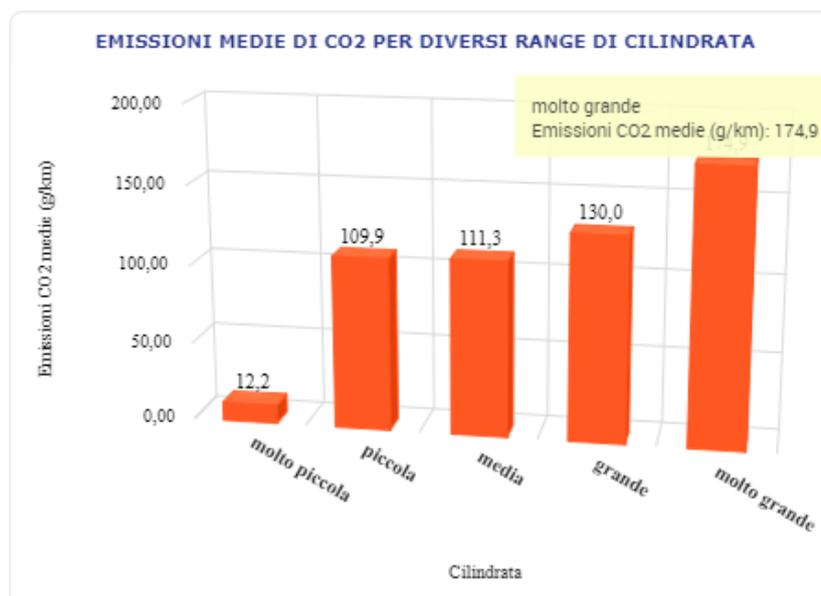
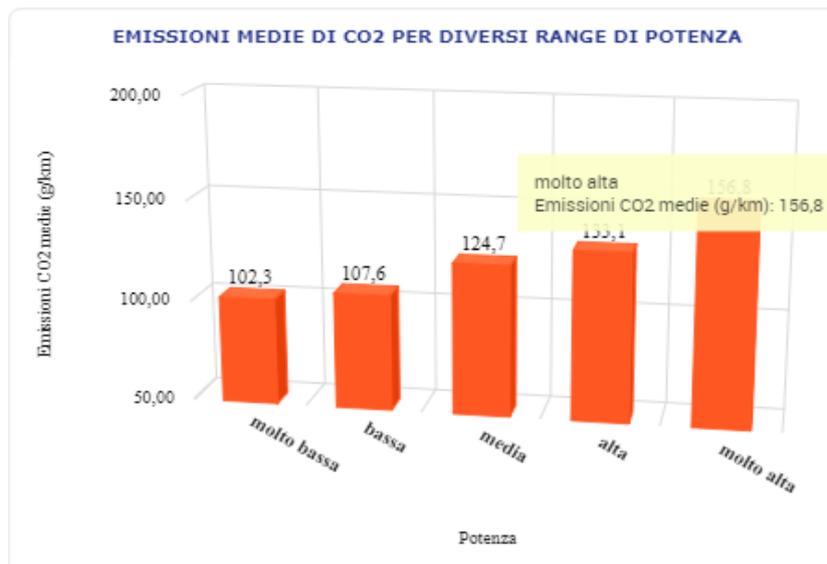
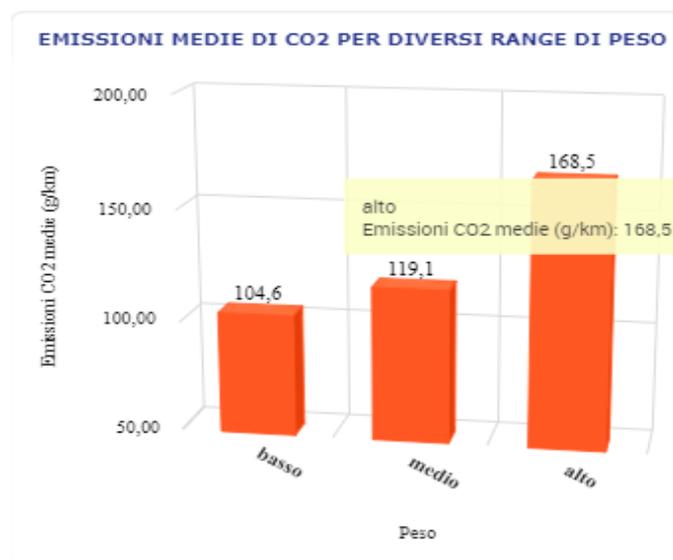


Figura 4.97: Emissioni medie per cilindrata



**Figura 4.98:** Emissioni medie per potenza



**Figura 4.99:** Emissioni medie per peso

Nel secondo *sheet*, denominato **emissioni per caratteristiche auto**, si è deciso di inserire delle *cross tables* che mettessero in relazione tra di loro le caratteristiche strutturali delle auto con le emissioni medie di CO2.

A questo proposito, come mostrato in **Figura 4.100** e successive tre figure, sono state inserite quattro *cross table* in cui, sia sulle righe che sulle colonne, sono presenti i diversi *range* delle caratteristiche prese in esame aggregando, su base media, le emissioni di CO2 delle autovetture, ponendo in questo modo l'attenzione sulla variazione delle emissioni.

EMISSIONI MEDIE DI CO2 PER CILINDRATA E POTENZA

	↑ POTENZA				
	molto bassa	bassa	media	alta	molto alta
↑ CILINDRATA	Emissioni medie CO2 (g/km)				
molto piccola	12,15	12,27	12,22		
piccola	106,60	109,70	117,30	131,35	
media	88,65	105,38	126,00	135,07	137,31
grande	88,99	117,83	126,80	130,67	144,61
molto grande	149,00	111,89	123,06	168,67	180,21

Figura 4.100: Emissioni medie per cilindrata e potenza

In questa prima *cross table* sono state messe in relazione la cilindrata e la potenza; in generale, emerge che, considerando una qualsiasi cilindrata, all’aumentare della potenza aumentano anche le emissioni medie di CO2 delle automobili.

Il viceversa, come si può osservare, non è sempre vero, dal momento che, per qualsiasi livello di potenza, all’aumentare della cilindrata le emissioni medie non hanno un andamento sempre crescente o decrescente ma aumentano e diminuiscono (tranne per il caso con potenza “molto alta”).

Tuttavia, potenza e cilindrata sono due parametri significativi per le *performance* dell’autovettura che devono essere ben bilanciati anche in considerazione dei consumi e delle emissioni inquinanti.

Inoltre, alle auto di piccola cilindrata corrisponde una potenza medio bassa.

EMISSIONI MEDIE DI CO2 PER PESO E CILINDRATA

	↑ PESO		
	basso	medio	alto
↑ CILINDRATA	Emissioni medie CO2 (g/km)	Emissioni medie CO2 (g/km)	Emissioni medie CO2 (g/km)
molto piccola		12,16	
piccola	106,43	118,95	
media	97,03	113,80	147,94
grande	153,07	128,31	154,68
molto grande	222,80	166,99	186,38

Figura 4.101: Emissioni medie per cilindrata e peso

In questo caso, mettendo in relazione il peso dell’autovettura con la cilindrata e osservando l’andamento delle emissioni medie, non emergono significative correlazioni tra le due caratteristiche strutturali.

In generale, però, si può osservare che per le autovetture di qualsiasi peso, all’aumentare della cilindrata aumentano anche le emissioni medie di CO2.

Inoltre, alle auto di piccola cilindrata, corrisponde un peso medio basso.

#### EMISSIONI MEDIE DI CO2 PER PESO E POTENZA

	↑ PESO		
	basso	medio	alto
↑ POTENZA	Emissioni medie CO2 (g/km)	Emissioni medie CO2 (g/km)	Emissioni medie CO2 (g/km)
molto bassa	101,29	46,71	
bassa	105,15	108,66	62,14
media	113,88	124,11	159,90
alta	131,21	130,83	173,84
molto alta	193,24	152,59	170,03

**Figura 4.102:** Emissioni medie per potenza e peso

In questa terza *cross table* sono state messe in relazione la potenza ed il peso delle automobili; in modo decisamente più significativo rispetto al caso precedente, per qualsiasi livello di peso, un aumento della potenza comporta maggiori emissioni medie di CO2. Al contrario, il viceversa non è vero.

#### EMISSIONI MEDIE DI CO2 PER PESO E ALIMENTAZIONE

	↑ PESO		
	basso	medio	alto
↑ TIPO ALIMENTAZIONE	Emissioni medie CO2 (g/km)	Emissioni medie CO2 (g/km)	Emissioni medie CO2 (g/km)
DIESEL	89,39	113,00	167,07
ELECTRIC/DIESEL	21,00	48,00	49,30
ELECTRIC/PETROL		40,74	58,10
GAS-GPL	107,58	135,65	
GAS-METHANE	82,23	117,52	214,00
PETROL	107,51	127,75	211,86

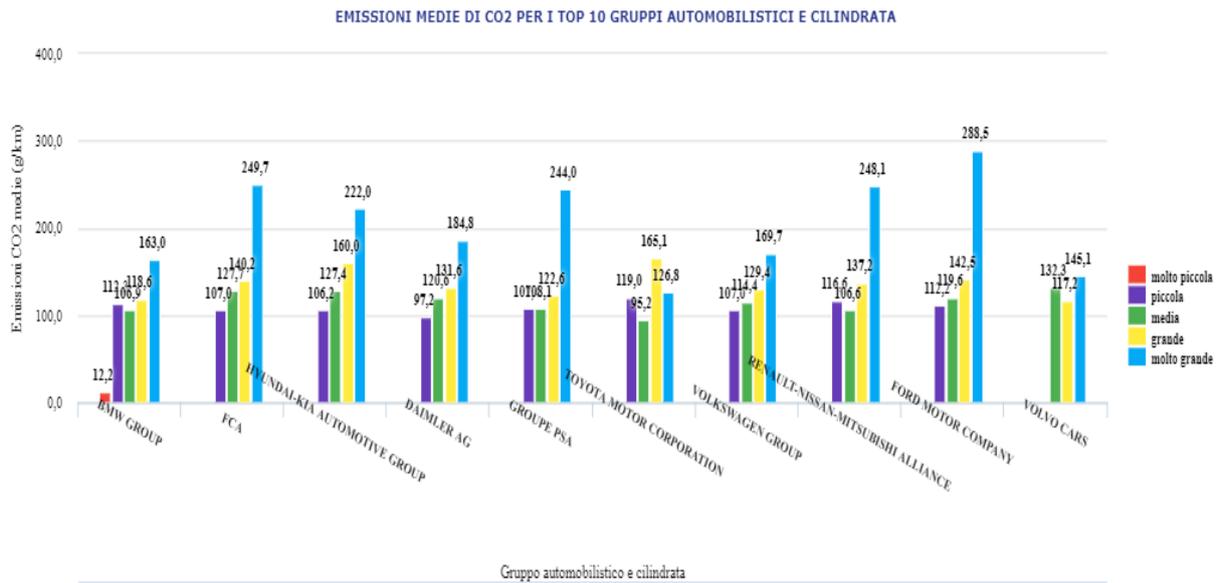
**Figura 4.103:** Emissioni medie per tipo alimentazione e peso

Nella quarta ed ultima *cross table* vengono messe in relazione il peso e l'alimentazione dell'automobile; tuttavia, per qualsiasi tipologia di alimentazione, all'aumento del peso corrisponde un aumento delle emissioni generate.

Inoltre, osservando la tabella, sono sempre evidenti le minori emissioni di CO2 generate dalle automobili ibride; infine, le auto ibride con motore a benzina hanno solo un peso medio alto, mentre alle auto a gas-GPL corrisponde solamente un peso medio basso.

Nel terzo ed ultimo *sheet*, denominato **emissioni per caratteristiche e gruppi**, sono stati inseriti dei diagrammi a barre per offrire un confronto tra le emissioni medie di CO2 dei top 10 gruppi automobilistici (definiti prima) distinguendole per le diverse caratteristiche strutturali delle auto (cilindrata, potenza e peso).

In **Figura 4.104** e successive due figure, sono mostrati i diagrammi a barre appena citati.



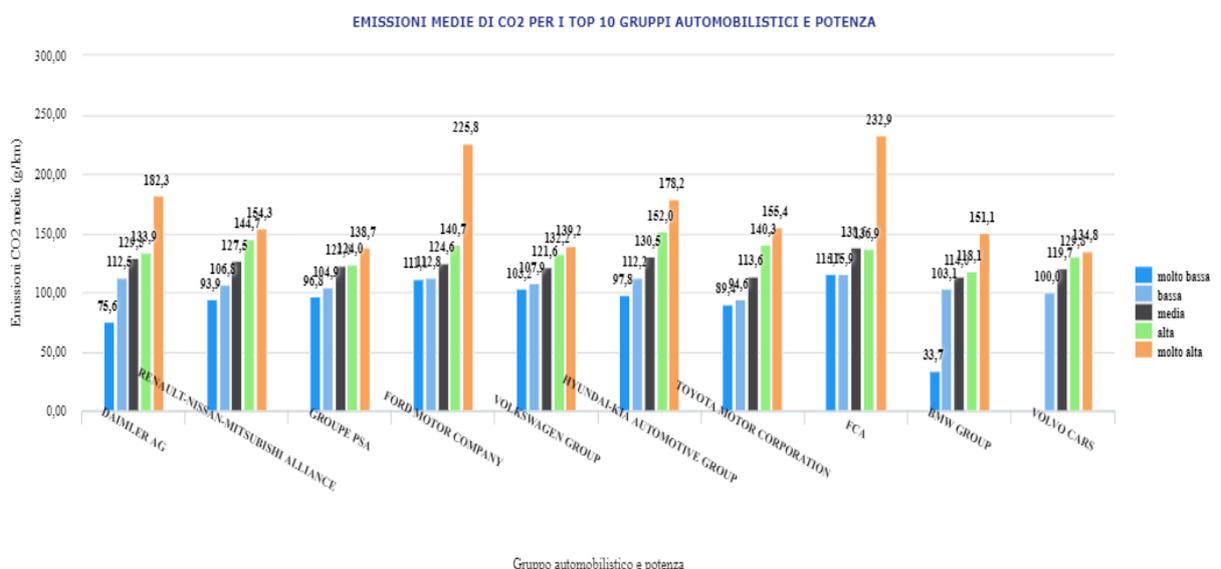
**Figura 4.104:** Emissioni medie per gruppo e cilindrata

In questo caso, in cui è stata considerata la cilindrata, in generale le emissioni medie aumentano all'aumentare della cilindrata delle automobili.

Tuttavia, tutti i top 10 gruppi, tranne Bmw, non hanno automobili di cilindrata molto piccola (fino a 800 cc) mentre Volvo neanche quelle di piccola cilindrata (fino a 1300 cc).

Pertanto, le emissioni medie aumentano all'aumentare della cilindrata tranne per alcuni gruppi in cui questo non è vero come Bmw, Toyota, Renault e Volvo.

Inoltre, le auto di cilindrata molto grande (oltre 2300 cc) che inquinano più di tutte appartengono al gruppo Ford seguita da Fca e Renault; per le auto di media cilindrata (1301-1800 cc) è il gruppo Volvo ad emettere, mediamente, maggior CO2, seguita da Fca e Daimler AG.



**Figura 4.105:** Emissioni medie per gruppo e potenza

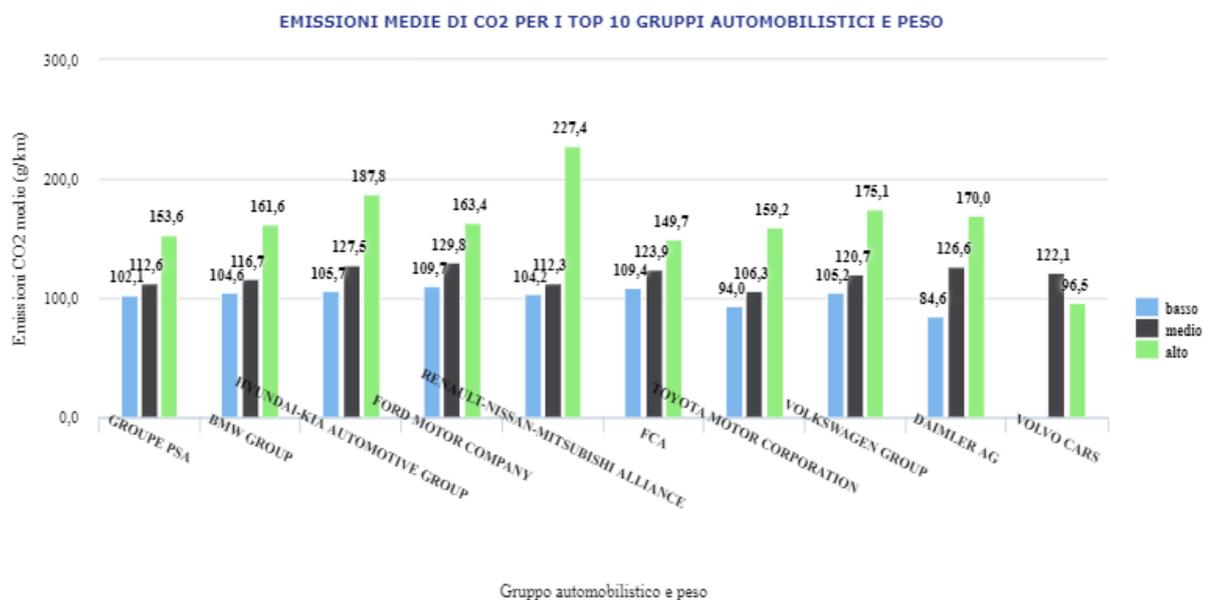
Anche in questo caso le emissioni medie aumentano, in generale, all'aumentare della potenza delle auto.

Qui, tutti i top 10 gruppi hanno autovetture per qualsiasi livello di potenza tranne Volvo che non possiede auto di potenza molto bassa.

Inoltre, per tutti i gruppi considerati, le emissioni medie aumentano sempre all'aumentare della potenza.

Per automobili di potenza molto alta (oltre 160 kw, cioè 218 cavalli) il gruppo FCA è quello che inquina di più, seguito da Ford e Daimler AG.

Per auto di potenza media (91-125 kw, cioè 124-170 cavalli) è sempre il gruppo FCA ad inquinare maggiormente, seguito da Hyundai-Kia e Daimler AG.



**Figura 4.106:** Emissioni medie per gruppo e peso

L'ultimo diagramma mette in relazione le emissioni medie con il peso delle autovetture; anche in questo caso, come nel precedente, per tutti i gruppi le emissioni medie aumentano sempre all'aumentare del peso tranne per Volvo in cui questo non accade.

Inoltre, tutti i gruppi possiedono autovetture per qualsiasi livello di peso considerato, tranne Volvo che non possiede automobili di basso peso.

Tuttavia, per le auto con un alto peso (oltre i 2000 kg) il gruppo Renault emette maggiori emissioni medie di CO2, seguito da Hyundai-Kia e Volkswagen.

Per le automobili con un peso medio (1201-2000 kg) inquinano maggiormente le auto del gruppo Ford, seguito da Hyundai-Kia e Daimler.

Infine, considerando un basso peso (fino a 1200 kg), è il gruppo Ford ad inquinare di più, quasi alla pari di FCA, seguiti da Hyundai-Kia.

## 5. Conclusioni

Nella prima parte della trattazione si è dedicato ampio spazio al vasto mondo della *Business Intelligence*, fornendo anche un quadro generale sulle basi di dati e sul linguaggio SQL.

Tuttavia, si è voluto mettere in risalto l'obiettivo principale della *Business Intelligence*: la raccolta, l'analisi, la pulizia e la trasformazione di dati di differenti ambiti applicativi in informazioni e conoscenza significativi; le organizzazioni e le aziende mettono in atto questo processo per supportare le decisioni di chi è a capo di ruoli direzionali, per migliorare i processi gestionali e decisionali rendendoli *data driven*, per incrementare il proprio vantaggio competitivo e per operare stime e valutazioni migliori del contesto competitivo in cui operano. Viene descritta l'esistenza di applicazioni *software* e prodotti di *Business Intelligence* che permettono di accedere a grandi quantità di dati, presenti generalmente su *Database*, con la possibilità di manipolarli e costruire report, *dashboard* e indicatori per la produzione di conoscenza e la definizione di analisi di vario tipo.

Pertanto, mediante il quadro generale sulle basi di dati e sul linguaggio SQL, si è voluta spiegare l'organizzazione e la struttura di grandi quantità di dati all'interno di *Database* e quelle che sono le principali caratteristiche del linguaggio che ne permette la memorizzazione, la gestione e l'interrogazione.

Successivamente, l'attenzione si è incentrata sulla grande risorsa inutilizzata dei dati *open*, prodotti ma non sempre diffusi dalle organizzazioni pubbliche e private, argomentando e illustrando come questi dati possono creare valore per cittadini e imprese e migliorarne la vita e la conoscenza.

Dopo aver descritto i principali portali governativi esistenti per la consultazione e la diffusione di *Open Data* è stato descritto e motivato il caso di studio ed il *file* di dati selezionato per dimostrare quanto appena esposto: l'estrazione di conoscenza da una grande quantità di dati mediante la costruzione di cruscotti all'interno di un prodotto di *Business Analytics*.

Una volta offerta una panoramica sul prodotto utilizzato in azienda e sulle funzionalità utilizzate sono state illustrate tutte le operazioni architetturali intermedie necessarie per il caricamento dei dati sul prodotto di BI e per la successiva costruzione dei cruscotti a partire da un *file* di dati grezzo.

Tra queste viene mostrata la progettazione di un'architettura di *Business Intelligence* tramite l'utilizzo di 2 *software* DBMS, quindi viene descritto l'intero processo di memorizzazione, gestione e interrogazione dei dati all'interno di un *Database* mediante opportune tecniche di *pre-processing* e sintassi del linguaggio SQL, infine è illustrata la connessione tra il prodotto di BI e un *Database* in remoto per il caricamento dei dati sull'ambiente dedicato.

La conoscenza estratta dai dati del file di partenza viene rappresentata per mezzo di cruscotti dinamici, ossia documenti interattivi costituiti da un insieme di grafici, tabelle, tabelle incrociate, elementi html, elementi testuali e altro ancora su uno o più *sheet*.

I risultati permettono di operare analisi diversificate e rispondere a domande di varia natura, ma l'elemento di fondamentale importanza che si intende mettere in risalto con il presente lavoro rappresenta la possibilità di eseguire lavori e analisi simili per diversi ambiti applicativi e per gli scopi desiderati.

## Riferimenti bibliografici e sitografia

- [1] Ascenti A., “*Business Intelligence: cosa è davvero*”, Computer World, Intelligent Retail, AI Driven Retail, 2015. <https://www.cwi.it/applicazioni-enterprise/business-intelligence-bi/business-intelligence-cosa-e-davvero-62>.
- [2] Walker J., “*Top 5 data warehouses on the market today*”, Monitis, a TeamViewer Company, 2018. <https://www.monitis.com/blog/top-5-data-warehouses-on-the-market-today/>.
- [3] Baralis E., Cerquitelli T., “*Introduzione alle basi di dati*”, Politecnico di Torino, 2013. [http://dbdmg.polito.it/wordpress/wp-content/uploads/2010/12/U1-L1-Ridotta\\_x2.pdf](http://dbdmg.polito.it/wordpress/wp-content/uploads/2010/12/U1-L1-Ridotta_x2.pdf).
- [4] Golfarelli M., Rizzi S., *Data Warehouse, teoria e pratica della progettazione*, seconda edizione, McGraw-Hill Education, 2006.
- [5] Paraboschi S., Atzeni P., Fraternali P., Ceri S., Torlone R., *Basi di dati*, quinta edizione, McGraw-Hill Education, 2018.
- [6] Dietrich D., Gray J., McNamara T., Poikola A., Pollock R., Tait J., and Zijlstra T., *The Open Data Handbook*, Open Knowledge Foundation. <http://opendatahandbook.org/guide/en/>.
- [7] Wikipedia, *Dati aperti*. [https://it.wikipedia.org/wiki/Dati\\_aperti](https://it.wikipedia.org/wiki/Dati_aperti).
- [8] The Open Source Suite for any modern Business Analytics, *KNOWAGE*. <https://www.knowage-suite.com/site/home/#nogo>.
- [9] Giliberto J., “*Diesel, auto e inquinamento: cosa bisogna sapere*”, *Il Sole24ORE*, 2019. [https://www.ilsole24ore.com/art/diesel-auto-e-inquinamento-cosa-bisogna-sapere-AEVypy7D?refresh\\_ce=1](https://www.ilsole24ore.com/art/diesel-auto-e-inquinamento-cosa-bisogna-sapere-AEVypy7D?refresh_ce=1).
- [10] Abu Eideh O., “*Limiti CO2, l’Europa ha deciso: dal 2030 emissioni medie di 59 g/km per le auto nuove*”, *La Stampa*, 2019. <https://www.lastampa.it/motori/ambiente/2019/03/30/news/limiti-co2-l-europa-ha-deciso-dal-2030-emissioni-medie-di-59-g-km-per-le-auto-nuove-1.33691643>.
- [11] Grassi M., “*Emissioni auto, l’Unione Europea fissa nuovi limiti*”, *Motor1*, 2017. <https://it.motor1.com/news/221620/lunione-europea-fissa-nuovi-limiti-per-le-emissioni-delle-auto/>.
- [12] <https://www.eng.it/>.
- [13] <https://www.linkedin.com/company/engineering-ingegneria-informatica-spa/about/>.
- [14] <https://knowage-suite.readthedocs.io/en/6.3/>.
- [15] Kumar V., Steinbach M., Tan. P., *Introduction to Data Mining*, Pearson College, 2005.

- [16] “Le auto verdi in Europa crescono del 39%”, *Alvolante*, 2018. <https://www.alvolante.it/news/auto-verdi-europa-vendite-2017--355856>.
- [17] Greco F., “Immatricolazioni, un 2017 d’oro per l’Europa: ecco i marchi più venduti”, *Il Sole 24 ORE*, 2018. [https://www.ilsole24ore.com/art/immatricolazioni-2017-d-oro-l-europa-ecco-marchi-piu-venduti-AEYLbgjD?refresh\\_ce=1](https://www.ilsole24ore.com/art/immatricolazioni-2017-d-oro-l-europa-ecco-marchi-piu-venduti-AEYLbgjD?refresh_ce=1).
- [18] Abu Eideh O., “Mercato auto 2018, ecco chi sale e chi scende in Europa”, *La Stampa*, 2019. [https://www.lastampa.it/motori/attualita/2019/01/23/news/mercato-auto-2018-ecco-chi-sale-e-chi-scende-in-europa-1.33672091?refresh\\_ce](https://www.lastampa.it/motori/attualita/2019/01/23/news/mercato-auto-2018-ecco-chi-sale-e-chi-scende-in-europa-1.33672091?refresh_ce).
- [19] Gemelli F., “Emissioni CO2 auto, chi è messo male e chi peggio”, *Motor1*, 2019. <https://it.motor1.com/news/343080/emissioni-co2-auto-a-che-punto-sono-i-costruttori/>.
- [20] Aiello V., “Mercato Europa, le auto più vendute paese per paese”, *Fanpage*, 2017. <https://motori.fanpage.it/mercato-europa-le-auto-piu-vendute-paese-per-paese/>.
- [21] “Inquinamento auto in Europa: l’Italia è 13°, Ecco la mappa 2018.” *Sicurato*, 2018. [https://www.sicurauto.it/news/inquinamento-auto-in-europa-litalia-e-13-ecco-la-mappa-2018/?refresh\\_ce=cp](https://www.sicurauto.it/news/inquinamento-auto-in-europa-litalia-e-13-ecco-la-mappa-2018/?refresh_ce=cp).

*A tutti voi che amorevolmente, emotivamente, economicamente e psicologicamente avete contribuito in modo speciale ed indimenticabile al raggiungimento di questo traguardo.*