
**ANALISI DEGLI INCIDENTI STRADALI NEL
REGNO UNITO AVVENUTI NEL PERIODO
2005-2017**

**Corso di Laurea Magistrale Ingegneria Informatica
e dell'Automazione**



Autori

Valerio Morelli

Federica Paganica

Federico Staffolani

Anno accademico 2023-2024

Indice

1 Introduzione al dataset	6
1.1 Descrizione	6
1.2 Struttura delle tabelle	6
1.3 Operazioni di ETL	9
2 Qlik	11
2.1 Introduzione	11
2.2 Caricamento dati	11
2.3 Data Analysis	12
2.3.1 Foglio 1 - Analisi temporale degli incidenti su base demografica	13
2.3.2 Foglio 2 - Analisi sul conducente	15
2.3.3 Foglio 3 - Analisi degli incidenti in base alle condizioni stradali	19
3 Tableau	22
3.1 Introduzione	22
3.2 Caricamento dati	22
3.3 Data Analysis	23
3.3.1 Dashboard 1 - Analisi degli incidenti in base alle condizioni ambientali	23
3.3.2 Dashboard 2 - Analisi degli aspetti relativi ai veicoli coinvolti in incidenti	25
3.3.3 Dashboard 3 - Predizione degli incidenti negli anni e distribuzione nei giorni settimanali	28
4 Power BI	34
4.1 Introduzione	34
4.2 Caricamento dati	34
4.3 Data Analysis	35
4.3.1 Dashboard 1 - Panoramica annuale della distribuzione degli incidenti su base temporale e spaziale	36
4.3.2 Dashboard 2 - Confronto della pericolosità stradale e cause di incidenti tra due distretti	39
4.3.3 Dashboard 3 - Analisi della pericolosità degli incidenti in relazione ai tipi di veicoli coinvolti	41
4.3.4 Dashboard 4 - Clustering dei tipi di veicoli coinvolti negli incidenti	44
4.4 Visualizzazioni installate dal marketplace	46
4.4.1 Waffle Chart (utilizzato nella dashboard 3)	46
4.4.2 Bullet Chart(utilizzato nella dashboard 3)	46
4.4.3 Timeline Slicer (utilizzato nella dashboard 4)	46

Elenco delle figure

1	Logo di Qlik	11
2	Creazione automatica delle associazioni in Qlik	12
3	Foglio 1 - Analisi temporale degli incidenti su base demografica	13
4	Foglio 1 - Analisi temporale degli incidenti e delle vittime	14
5	Foglio 1 - Analisi degli incidenti per genere e fascia d'età	15
6	Foglio 2 - Analisi sul conducente	16
7	Foglio 2 - Analisi degli incidenti in base alla deprivazione socio-economica nelle aree rurali	17
8	Foglio 2 - Analisi degli incidenti in base alla deprivazione socio-economica nelle piccole città	17
9	Foglio 2 - Analisi degli incidenti in base alla deprivazione socio-economica nelle aree urbane	18
10	Foglio 3 - Analisi degli incidenti in base alle condizioni stradali	19
11	Foglio 3 - Analisi degli incidenti filtrati in base al limite di velocità e al tipo di strada .	20
12	Foglio 3 - Analisi degli incidenti filtrati in base al limite di velocità	21
13	Logo di Tableau	22
14	Creazione delle associazioni in Tableau	23
15	Dashboard 1 - Analisi degli incidenti in base alle condizioni ambientali	24
16	Dashboard 1 filtrata per illuminazione della strada e severità degli incidenti	25
17	Dashboard 2 - Analisi degli aspetti relativi ai veicoli coinvolti in incidenti	26
18	Dashboard 2 filtrata per categoria di veicolo e per tipo di area	27
19	Dashboard 3 - Predizione degli incidenti negli anni e distribuzione nei giorni settimanali	28
20	Stima effettuata da Tableau dei valori del numero totale di vittime previsto per settembre 2021, con i rispettivi limiti superiori e inferiori	29
21	Andamento del numero totale di vittime di incidenti in UK negli anni 2021-2022, con enfasi sul mese di settembre 2021	30
22	Dashboard 3 filtrata per età delle vittime	33
23	Logo di Power BI	34
24	Importazione del dataset sugli incidenti in Power BI	35
25	Creazione delle associazioni tra i due dataset	35
26	Panoramica annuale della distribuzione degli incidenti su base temporale e spaziale . .	36
27	Andamento mensile della frequenza degli incidenti	37
28	Lunghezza delle giornate nel regno unito in funzione del mese dell'anno	37
29	Misura DAX rappresentante la percentuale di incidenti fatali rispetto al totale	38
30	Filtraggio per incidenti in orario notturno	38
31	Confronto della pericolosità stradale e cause di incidenti tra due distretti	39
32	Formula DAX per la classificazione della posizione di un distretto in funzione delle sue coordinate	40
33	Colonna calcolata per convertire dei limiti di velocità e gli anni dei veicoli.	40
34	Diagnosi dell'impatto dei limiti di velocità e sull'età dei veicoli sugli incidenti di due distretti	41
35	Riepilogo della pericolosità degli incidenti in relazione ai tipi di veicoli coinvolti	41
36	Dataset delle icone per le diverse tipologie di veicoli	42
37	Misura DAX per assegnare pesi diversi agli incidentati a seconda della gravità dell'incidente	42

38	Profilazione degli incidenti avvenuti in bicicletta	43
39	Profilazione degli incidenti in cui sono coinvolti i furgoni	44
40	Clustering dei tipi di veicoli coinvolti negli incidenti del 2012	45
41	Clustering dei tipi di veicoli coinvolti negli incidenti del 2016	45
42	Andamento nel tempo della frequenza di incidenti in base al tipo di veicolo	46
43	Waffle chart	47
44	Bullet chart	47
45	Timeline slicer	47

Elenco delle tabelle

1	Campi di Accident_Information.csv	7
2	Campi di Vehicle_Information.csv	8
3	Confronto tra stime e valori reali per il 2021 e 2022 (Valori Totali)	31
4	Confronto tra stime e valori reali per il 2021 e 2022 (Valori Medi)	32

1 Introduzione al dataset

Questa tesina fornisce un'analisi completa dei dati sugli incidenti stradali nel Regno Unito, concentrandosi sulle tendenze degli incidenti, sulla distribuzione delle vittime e sui modelli temporali. Il dataset analizzato comprende gli incidenti verificatisi nel periodo 2005-2016, ed offre preziose indicazioni per migliorare le misure di sicurezza stradale.

1.1 Descrizione

Il dataset preso in considerazione per questo progetto fornisce dettagliati dati sugli incidenti stradali e i veicoli coinvolti nel Regno Unito, coprendo il periodo dal 2005 al 2017.

Il governo del Regno Unito, infatti, raccoglie e pubblica, di solito su base annuale, informazioni dettagliate sugli incidenti stradali in tutto il paese. Queste informazioni provengono dal sito *Open Data* del governo britannico, pubblicate dal *Dipartimento dei Trasporti*.

La raccolta comprende informazioni riguardanti posizioni geografiche, condizioni meteorologiche, tipi di veicoli, numero di vittime e manovre dei veicoli coinvolti. È una risorsa completa e interessante per l'analisi e la ricerca nel campo della sicurezza stradale.

È possibile accedere a questo dataset tramite il seguente link: [https://www.kaggle.com/dataset/tciaras/uk-road-safety-accidents-and-vehicles](https://www.kaggle.com/dataset/tsiaras/uk-road-safety-accidents-and-vehicles).

1.2 Struttura delle tabelle

Il dataset è composto da due file .csv :

- **Accident_Information.csv** : ogni riga rappresenta un incidente stradale unico identificato dalla colonna **Accident_Index** , con varie proprietà legate all'incidente e con un intervallo temporale compreso tra il 2005 e il 2017.
- **Vehicle_Information.csv** : ogni riga rappresenta il coinvolgimento di un veicolo unico in un incidente stradale unico, con varie proprietà di veicoli e passeggeri e con un intervallo temporale compreso tra il 2005 e il 2017.

I due file possono essere collegati attraverso l'*identificatore univoco* dell'incidente stradale (colonna **Accident_Index**).

Si osserva, inoltre, che la sigla *LSOA*, presente in uno dei campi di **Accident_Information** , sta per "Lower Layer Super Output Area". Queste sono aree più piccole rispetto ai distretti, progettate dall'Ufficio Nazionale di Statistica del Regno Unito per consentire analisi più dettagliate a livello locale e per agevolare la raccolta di dati socio-economici e demografici. [1]

Oltre a ciò, è opportuno notare che l'*IMD Decile*, dove IMD sta per *Indices of Multiple Deprivation*, è uno strumento utilizzato per valutare il degrado in diverse aree geografiche, come quartieri o comunità. Nello specifico, i decili presenti in questo dataset vengono calcolati classificando i 32.844 LSOA in Inghilterra dal più deprivato al meno deprivato e suddividendoli in 10 gruppi uguali. Gli LSOA con valore 1 rientrano nel 10% più svantaggiato degli LSOA a livello nazionale e gli LSOA con valore 10 rientrano nel 10% meno svantaggiato degli LSOA a livello nazionale. Quindi, il campo **Driver_IMD_Decile** di **Vehicle_Information** riporta tale indice, relativo alla provenienza del conducente del veicolo che è stato coinvolto in un incidente. [2]

Di seguito sono forniti per esteso gli elenchi completi dei campi delle due tabelle.

Tabella 1: Campi di Accident_Information.csv

Campo	Descrizione
Accident_Index	Identificatore univoco dell'incidente stradale
1st_Road_Class	Classe della prima strada coinvolta
1st_Road_Number	Numero della prima strada coinvolta
2nd_Road_Class	Classe della seconda strada coinvolta
2nd_Road_Number	Numero della seconda strada coinvolta
Accident_Severity	Severità dell'incidente
Carriageway_Hazards	Pericoli sulla carreggiata
Date	Data dell'incidente
Day_of_Week	Giorno della settimana dell'incidente
Did_Police_Officer_Attend_Scene_of_Accident	Presenza di un ufficiale di polizia sulla scena
Junction_Control	Controllo dell'incrocio
Junction_Detail	Dettagli dell'incrocio
Latitude	Latitudine dell'incidente
Light_Conditions	Condizioni di illuminazione
Local_Authority_(District)	Autorità locale (distretto)
Local_Authority_(Highway)	Autorità locale (strada)
Location_Easting_OSGR	Coordinate est della posizione dell'incidente
Location_Northing_OSGR	Coordinate nord della posizione dell'incidente
Longitude	Longitudine dell'incidente
LSOA_of_Accident_Location	LSOA della posizione dell'incidente
Number_of_Casualties	Numero di vittime
Number_of_Vehicles	Numero di veicoli
Pedestrian_Crossing-Human_Control	Attraversamento pedonale - Controllo umano
Pedestrian_Crossing-Physical_Facilities	Attraversamento pedonale - Strutture fisiche
Police_Force	Forza di polizia
Road_Surface_Conditions	Condizioni della superficie stradale
Road_Type	Tipo di strada
Special_Conditions_at_Site	Condizioni speciali sul sito
Speed_limit	Limite di velocità
Time	Ora dell'incidente
Urban_or_Rural_Area	Area urbana o rurale
Weather_Conditions	Condizioni meteorologiche
Year	Anno dell'incidente
InScotland	Incidente in Scozia

Tabella 2: Campi di `Vehicle_Information.csv`

Campo	Descrizione
<code>Accident_Index</code>	Identificatore univoco dell'incidente stradale
<code>Age_Band_of_Driver</code>	Fascia di età del conducente
<code>Age_of_Vehicle</code>	Età del veicolo coinvolto
<code>Driver_Home_Area_Type</code>	Tipo di area di residenza del conducente
<code>Driver_IMD_Decile</code>	Decile di deprivazione del conducente
<code>Engine_Capacity_.CC.</code>	Cilindrata del motore
<code>Hit_Object_in_Carriageway</code>	Oggetto colpito sulla carreggiata
<code>Hit_Object_off_Carriageway</code>	Oggetto colpito fuori dalla carreggiata
<code>Journey_Purpose_of_Driver</code>	Scopo del viaggio del conducente
<code>Junction_Location</code>	Posizione dell'incrocio
<code>make</code>	Marca del veicolo
<code>model</code>	Modello del veicolo
<code>Propulsion_Code</code>	Codice di propulsione del veicolo
<code>Sex_of_Driver</code>	Sesso del conducente
<code>Skidding_and_Overturning</code>	Scivolamento e ribaltamento del veicolo
<code>Towing_and_Articulation</code>	Traino e articolazione del veicolo
<code>Vehicle_Leaving_Carriageway</code>	Veicolo che lascia la carreggiata
<code>Vehicle_Location.Restricted_Lane</code>	Posizione del veicolo - Corsia limitata
<code>Vehicle_Manoeuvre</code>	Manovra del veicolo
<code>Vehicle_Reference</code>	Riferimento al veicolo
<code>Vehicle_Type</code>	Tipo di veicolo
<code>Was_Vehicle_Left_Hand_Drive</code>	Veicolo a guida a sinistra
<code>X1st_Point_of_Impact</code>	Primo punto di impatto
<code>Year</code>	Anno dell'incidente

1.3 Operazioni di ETL

Il dataset in uso è particolarmente voluminoso sia dal punto di vista del numero di colonne che del numero di righe. Per questo motivo, prima di importarlo nei tre software per svolgere le analisi, è risultato conveniente rimuovere le informazioni di utilità minore.

In particolare, sono state rimosse le colonne `InScotland`, `Police_Force`, `LSOA_of_Accident_Location`, `Location_Easting_OSGR` e `Location_Northing_OSGR` dal dataset sugli incidenti e le colonne `Vehicle_Location.Restricted_Lane` e `Vehicle_Reference` dal dataset sui veicoli.

Successivamente ci si è accorti della presenza di veicoli nel secondo dataset relativi ad incidenti avvenuti nel 2004 (cosa che è stata possibile dedurre osservando campo `Accident_Index`), mancanti quindi di corrispettivo nel primo dataset riportante informazioni sugli incidenti a partire dall'anno 2005. Inoltre, ad un' analisi più approfondita si è notata la non corrispondenza di una minoranza di incidenti tra i due dataset. Quindi, per tali ragioni, a ciascuno dei due dataset sono state rimosse le righe riguardanti incidenti non registrati nell' altro dataset. Dunque sono state eseguite due operazioni di `inner join` su ciascun dataset.

Infine è stata creata una versione alternativa del dataset snellendone il volume verticale. Per far ciò è stato eseguito un ordinamento rispetto alla colonna `Date` e rimosse tutte le righe eccetto una ogni 100. In questo modo è stato possibile rimuovere in modo uniforme rispetto all'arco temporale il volume del dataset. Si noti però che tale versione ridotta è stata utilizzata per motivi di prestazioni, ma solo in quelle analisi che non coinvolgono l'estrazione di dati cumulativi come la somma del numero di incidenti. In tutti gli altri casi, per mantenere le analisi più fedeli possibili, si è utilizzato la versione senza rimozione di righe.

Tutte queste operazioni di *ETL* sono state eseguite utilizzando la libreria *Pandas* di *Python*.

```

1 import pandas as pd
2
3 # Caricamento del dataset sugli incidenti e rimozione delle colonne
4 # non necessarie
5 accident_df = pd.read_csv('Accident_Information.csv') \
6     .drop(columns=['InScotland', 'Police_Force', ,
7                   'LSOA_of_Accident_Location',
8                   'Location_Easting_OSGR', 'Location_Northing_OSGR'])
9
10 # Caricamento del dataset sui veicoli e rimozione delle colonne non
11 # necessarie
12 vehicle_df = pd.read_csv('Vehicle_Information.csv', encoding="ISO
13 -8859-1") \
14     .drop(columns=['Vehicle_Location.Restricted_Lane', ,
15               'Vehicle_Reference'])
16
17 # Estrazione di una data ogni 10 nel dataset degli incidenti
18 accident_df['Date'] = pd.to_datetime(accident_df['Date'])
19 accident_df.sort_values(by='Date', inplace=True)
20 accident_df = accident_df.iloc[::10, :]
21
22 # Rimozione degli incidenti senza informazioni sui veicoli tramite una
23 # inner join
24 # Nota: data la possibile presenza di piu' veicoli in un incidente,
25 # vengono rimossi i duplicati prima di eseguire il merge, in modo da
26 # non presentare duplicati nelle righe del primo dataset
27 incidents_in_veichle_df = vehicle_df[['Accident_Index']].
28     drop_duplicates(subset='Accident_Index')
29 accident_df = pd.merge(accident_df, incidents_in_veichle_df, on=
30     'Accident_Index', how='inner')
31
32 # Salvataggio del dataset degli incidenti ripulito in un file CSV
33 accident_df.to_csv('accident_etl.csv', index=False)
34
35 # Rimozione dei veicoli non associati ad alcun incidente nel primo
36 # dataset mediante una inner join
37 vehicle_df = pd.merge(accident_df[['Accident_Index']], vehicle_df, on=
38     'Accident_Index', how='inner')
39
40 # Salvataggio del dataset dei veicoli ripulito in un file CSV
41 vehicle_df.to_csv('vehicle_etl.csv', index=False)

```

Listing 1: Operazioni di ETL con Pandas

2 Qlik

Nella sezione seguente, esploreremo le analisi condotte utilizzando il primo dei tre software. Inizieremo delineando brevemente il processo di caricamento dei dati, sottolineando le difficoltà incontrate e le soluzioni adottate per affrontarle. Successivamente, presenteremo un elenco delle analisi eseguite e delle relative visualizzazioni, ognuna accompagnata dalle rispettive motivazioni e conclusioni ottenute.

2.1 Introduzione

Nell'ambito delle soluzioni di Business Intelligence, Qlik emerge come uno dei leader innovativi, offrendo una suite di strumenti software che facilitano la visualizzazione, l'analisi e l'interpretazione dei dati. Fondato nel 1993 in Svezia, Qlik ha ridefinito il modo in cui le organizzazioni interagiscono con i loro dati, spostando il focus dall'elaborazione tradizionale di report statici a un'esplorazione dinamica e interattiva delle informazioni.



Figura 1: Logo di Qlik

Qlik offre due prodotti distintivi che si rivolgono a esigenze specifiche nell'ambito della Data Analytics: *QlikView* e *Qlik Sense*. Entrambi i software si distinguono per il loro approccio unico nella gestione e visualizzazione delle informazioni, ma si differenziano sostanzialmente per il tipo di utenza a cui sono destinati e per le funzionalità offerte.

QlikView è progettato per facilitare la cosiddetta "*Guided Analytics*". Si tratta di uno strumento che permette alle organizzazioni di sviluppare applicazioni analitiche su misura che riflettono le esigenze specifiche dell'azienda. In questo contesto, l'utente finale si trova in una posizione di consultazione piuttosto che di creazione, non avendo la possibilità di modificare o estendere l'applicazione con nuovi elementi analitici. D'altra parte, *Qlik Sense* si posiziona come uno strumento di "*Self-service Analytics*", progettato per democratizzare l'accesso e l'analisi dei dati. Grazie a *Qlik Sense*, l'utente finale è messo in condizione di esplorare autonomamente i dati disponibili, creando report personalizzati e dashboard dinamiche.

Al contrario delle tradizionali piattaforme di BI, che si affidano a database relazionali e meccanismi di interrogazione basati su query, *Qlik Sense* si distingue per l'implementazione di un *Associative Engine*. Tale motore consente a *Qlik Sense* di superare le limitazioni imposte da SQL, il quale non è stato inizialmente ideato per supportare un'analisi interattiva dei dati. Grazie al suo motore associativo, *Qlik Sense* ha la capacità di eseguire calcoli e aggregazioni in tempo reale, aggiornando le analisi e evidenziando le connessioni tra i dati con un'efficienza notevole. Questo approccio non solo facilita una più profonda comprensione delle informazioni ma rende anche l'analisi dei dati accessibile e intuitiva per gli utenti finali, permettendo loro di esplorare liberamente le relazioni nascoste nei dati senza la necessità di formulare complesse query SQL.

2.2 Caricamento dati

Dopo aver effettuato le operazioni di ETL sui due dataset e aver creato una nuova applicazione su *Qlik Sense*, il primo passo da seguire consiste nel caricare i dati dalla loro sorgente.

Attraverso l'interfaccia di *gestione dati*, è stata selezionata l'opzione per l'importazione di dati da file locali consentendo il caricamento dei due dataset in formato *.csv*. La piattaforma ha facilitato il processo mediante una procedura guidata, che ha offerto la possibilità di visualizzare un'anteprima dei dati e controllarne la correttezza prima di integrarli nell'app. Inoltre, Qlik ha automaticamente creato l'associazione tra le tabelle utilizzando come chiave il campo **Accident Index**, come illustrato nella Figura 2.



Figura 2: Creazione automatica delle associazioni in Qlik

Successivamente al caricamento dei file, sono state confermate le impostazioni di importazione e avviato il processo di elaborazione, Qlik ha elaborato i dati con efficienza integrandoli nell'ambiente analitico. Tuttavia, è stata necessaria l'implementazione di manipolazioni aggiuntive per adattare i dati alle specifiche delle analisi da effettuare.

La prima manipolazione ha comportato la creazione di una nuova colonna calcolata denominata **Speed limit kmh**. Questa operazione è stata essenziale poiché il dataset originario presentava i limiti di velocità espressi in miglia orarie (mph) e si è reso necessario convertirli in chilometri orari (km/h) per allinearsi agli standard utilizzati nel contesto dell'analisi.

In aggiunta, è stato realizzato un secondo intervento sui dati attraverso la realizzazione di uno script specifico. Questo script carica i dati dal file *accident.etl.csv* e trasforma i valori presenti nel campo **Day of Week** in valori duali. Tale trasformazione genera per ogni giorno della settimana una coppia di valori che include sia la rappresentazione numerica (basata sull'ordine convenzionale dei giorni della settimana) sia quella testuale originale. Il risultato di questa operazione è stato assegnato a una nuova colonna chiamata **DayOfWeek**. Questa manipolazione ha garantito che i giorni della settimana fossero ordinati correttamente nelle visualizzazioni, facilitando così l'interpretazione dei pattern temporali all'interno dei dati analizzati.

I dati, ora disponibili per l'analisi, ci consentono di procedere alla realizzazione di fogli e dashboard interattivi.

2.3 Data Analysis

I dati, ora disponibili per l'analisi, ci consentono di procedere alla realizzazione di fogli interattivi mediante l'utilizzo del software Qlik. Nelle sezioni successive, si procederà con l'analisi dettagliata dei vari fogli, con l'obiettivo di esaminare i grafici presenti in ciascuno di essi. Nel dattaglio, sono stati realizzati tre fogli:

- Foglio 1 - Analisi temporale degli incidenti su base demografica
- Foglio 2 - Analisi sul conducente

- Foglio 3 - Analisi degli incidenti in base alle condizioni stradali

Ogni foglio è organizzato in modo da dividere nettamente le due funzionalità principali: la visualizzazione dei dati e l’interazione con essi. A partire da sinistra, sono collocati i grafici principali, ciascuno derivante da un diverso foglio di lavoro, supportati in alcuni casi da legende che chiariscono il significato dei colori impiegati per differenziare le varie categorie di dati. Sulla destra, invece, è situata l’area dedicata al filtraggio, dove gli utenti possono interagire con il foglio mediante strumenti di selezione in modo affinare la visualizzazione in base alle proprie necessità analitiche.

2.3.1 Foglio 1 - Analisi temporale degli incidenti su base demografica

Il primo foglio presentato è il risultato della prima analisi condotta utilizzando il software Qlik Sense, con l’obiettivo di familiarizzare con gli strumenti offerti dal tool e ottenere una panoramica generale degli incidenti. Le visualizzazioni realizzate, come riportato nella Figura 3, mostrano la distribuzione degli incidenti anno per anno, evidenziando la frequenza e le conseguenze di tali eventi, oltre a fornire una *ripartizione demografica* degli autisti coinvolti, suddivisa per genere e fascia di età.



Figura 3: Foglio 1 - Analisi temporale degli incidenti su base demografica

Procediamo ora con una descrizione dettagliata degli elementi costitutivi del foglio:

- La visualizzazione principale *in alto* è un grafico combinato verticale che mostra l’evoluzione annuale del numero di incidenti e vittime dal 2005 al 2016. Le linea sovrapposta rappresenta il numero totale di incidenti per anno, mentre le barre verticali indicano il numero di vittime correlate.
- Il primo elemento *in basso a sinistra* è un grafico a torta che mostra la distribuzione percentuale degli incidenti basata sul genere dei conducenti. Con una suddivisione chiara tra maschi e femmine, il grafico evidenzia la disparità di genere negli incidenti stradali.
- Il grafico *in basso a destra* è un diagramma Mekko, o grafico a barre marimekko, che combina due variabili: il genere e la fascia d’età dei conducenti. Ogni segmento del grafico rappresenta una fascia d’età specifica all’interno di ciascun genere, mostrando sia la percentuale che il numero assoluto di incidenti.

- Nella sezione *sulla destra*, è presente un KPI che fornisce un conteggio totale del numero di veicoli coinvolti negli incidenti durante il periodo in esame.
- Una componente essenziale di questo foglio è rappresentata dall'area dedicata ai filtri, situata nella parte *destra* dell'interfaccia. Questa sezione permette agli utenti di affinare e personalizzare le visualizzazioni selezionando specifici intervalli temporali, fasce di età o genere.

Dall'analisi visiva del grafico, riportato nella Figura 4 possiamo osservare:

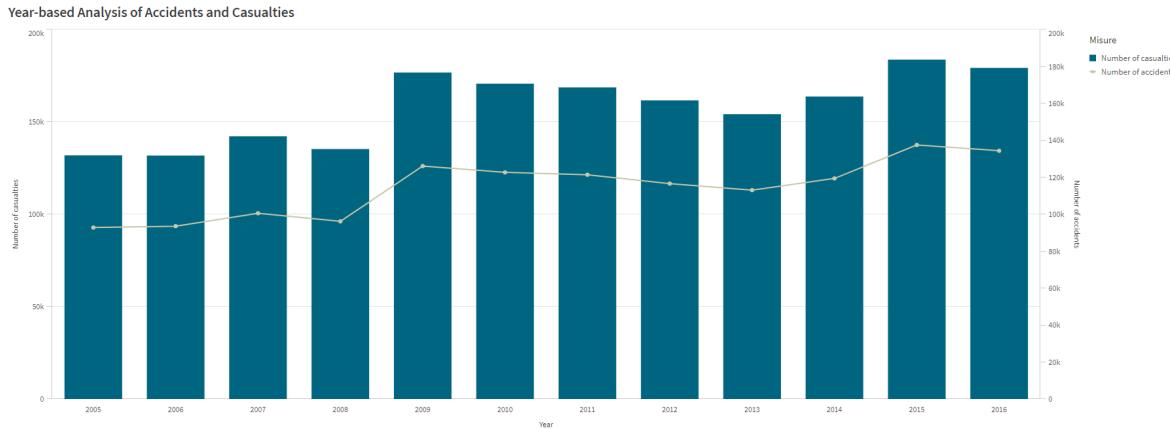


Figura 4: Foglio 1 - Analisi temporale degli incidenti e delle vittime

- Numero di incidenti (linea): la frequenza degli incidenti è relativamente costante fino al 2008. Nel 2009, si osserva un picco, con un numero di incidenti più alto rispetto agli anni precedenti. Successivamente, c'è una leggera diminuzione fino al 2013. Dal 2014 in poi, si nota un altro leggero aumento, con valori leggermente più alti rispetto al picco del 2009.
- Numero di vittime (barre blu): il numero di vittime sembra rimanere costante fino al 2008, con un aumento nel 2009. Dal 2009 al 2013, la tendenza è leggermente decrescente. Nel 2014 e 2015, c'è una risalita, e nel 2016 si stabilizza con valori simili a quelli dell'anno precedente.

La frequenza degli incidenti stradali mostra una certa costanza fino al 2008, suggerendo che le condizioni di guida, le regolamentazioni stradali, o altri fattori sono rimasti relativamente stabili in questo periodo. Il picco osservato nel 2009 potrebbe indicare un cambiamento nelle condizioni di guida (ad esempio un aumento del traffico veicolare o modifiche alle infrastrutture stradali) o una variazione nelle metodologie di raccolta dati. La leggera diminuzione osservata fino al 2015 potrebbe riflettere l'efficacia di misure di sicurezza stradale o campagne di prevenzione implementate in risposta all'aumento degli incidenti.

La costanza nel numero di vittime fino al 2008, seguita da un leggero aumento nel 2009, potrebbe suggerire che gli incidenti che hanno portato al picco di frequenza in quell'anno sono stati particolarmente gravi o che si è verificato un incremento generale nella gravità degli incidenti. La tendenza decrescente dal 2009 al 2013 potrebbe indicare un miglioramento nell'efficacia dei sistemi di sicurezza dei veicoli o nelle pratiche di primo soccorso, che hanno contribuito a ridurre il numero di vittime in caso di incidente.

Dall'analisi dei grafici nella Figura 5 possiamo notare che la maggior parte degli incidenti coinvolge conducenti di genere maschile (69.2%) rispetto a quelli di genere femminile (30.8%). Questo potrebbe

essere dovuto a una serie di fattori, come una maggiore presenza maschile sulla strada o differenze nel comportamento alla guida, tuttavia, per affermarlo con certezza sarebbero necessarie ulteriori informazioni.

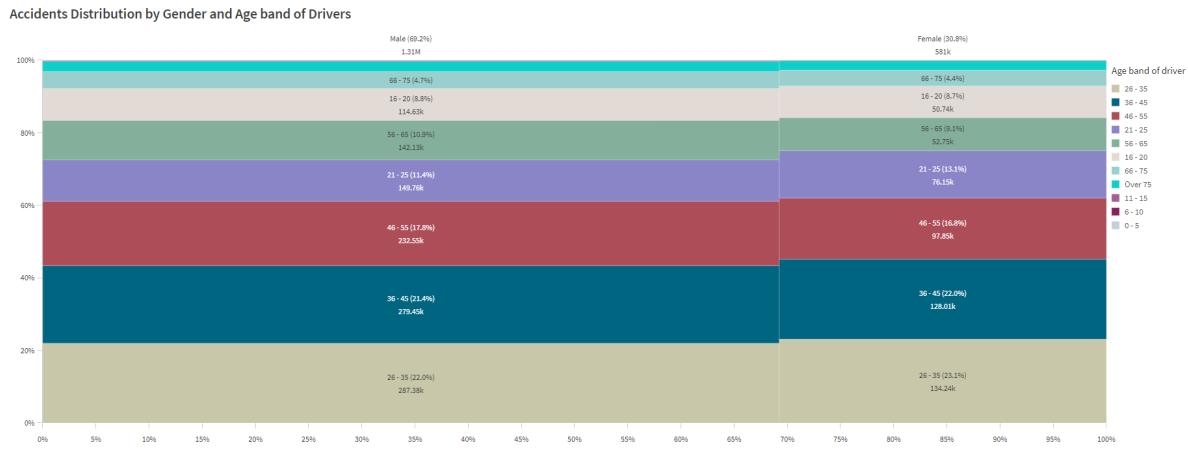


Figura 5: Foglio 1 - Analisi degli incidenti per genere e fascia d'età

Per quanto riguarda le fasce d'età, il gruppo più coinvolto in incidenti è quello tra i 26 e i 35 anni per entrambi i sessi, con il 22% degli incidenti per gli uomini e il 23.1% per le donne. Segue la fascia d'età 36-45 anni, con il 21.4% per gli uomini e il 22% per le donne. La fascia d'età 46-55 anni è la terza più coinvolta con il 17.8% degli incidenti per gli uomini e il 16.8% per le donne. Le fasce d'età più giovani (0-25 anni) e più anziane (oltre 65 anni) sembrano essere meno coinvolte negli incidenti, questo potrebbe essere dovuto a una minore presenza al volante o a una maggiore cautela nella guida. In tutte le fasce d'età, gli uomini sono coinvolti in un maggior numero di incidenti rispetto alle donne, ciò è coerente con il dato complessivo che mostra una percentuale maggiore di incidenti per i conducenti di genere maschile.

In conclusione, i dati suggeriscono che il *genere* e l'*età* del conducente possono essere fattori significativi nella distribuzione degli incidenti stradali: i conducenti maschi e quelli nelle fasce d'età medie sembrano essere i più coinvolti, suggerendo potenzialmente un focus per future iniziative di sicurezza stradale. Allo stesso tempo, l'analisi pone le basi per ulteriori indagini che potrebbero esplorare le cause sottostanti di tali tendenze, incluso il ruolo di fattori socio-economici e comportamentali che possono influenzare il rischio di incidenti. La comprensione di questi modelli è cruciale per lo sviluppo di strategie preventive mirate e per *migliorare la sicurezza* di tutti gli utenti della strada.

2.3.2 Foglio 2 - Analisi sul conducente

Il secondo foglio realizzato su Qlik presenta un'analisi dei dati relativi agli incidenti stradali, con l'obiettivo specifico di esaminare il ruolo del conducente. Questa analisi si concentra su tre aspetti principali: la gravità degli incidenti in relazione all'età del conducente, il numero di incidenti basato sulla deprivazione socio-economica del conducente, e la tipologia di area di residenza del conducente. Procediamo ora con una descrizione dettagliata degli elementi costitutivi del foglio, riportato nella Figura 6:

- la visualizzazione *in alto* è un grafico lineare ad area che mostra la distribuzione degli incidenti classificati come *lievi*, *gravi* e *mortal* tra le diverse fasce d'età dei conducenti;

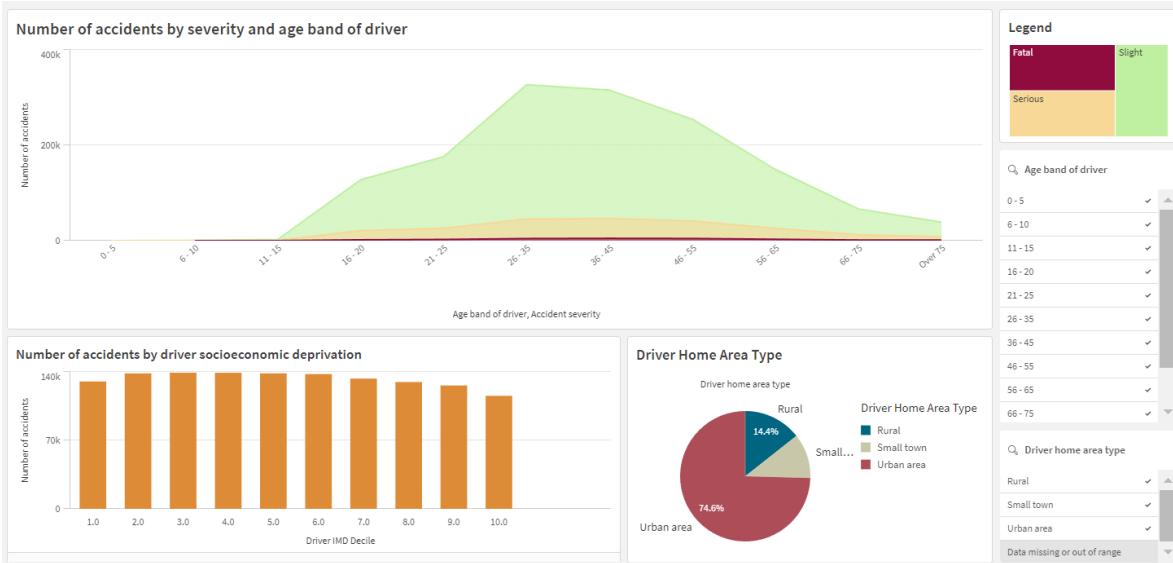


Figura 6: Foglio 2 - Analisi sul conducente

- il grafico a barre *in basso a sinistra* evidenzia il numero di incidenti divisi per decili di deprivazione socio-economica, con il decile 1 che rappresenta il più deprivato e il decile 10 il meno deprivato;
- il grafico a torta *in basso a destra* evidenzia come si distribuiscono i conducenti tra i diversi tipi di aree di residenza (aree urbane, rurali e piccole città);
- l'ultima componente di questo foglio è rappresentata dalla sezione dedicata ai filtri, situata nella parte *destra* dell'interfaccia, che permette agli utenti di personalizzare le visualizzazioni selezionando la gravità degli incidenti, il tipo di area di residenza e la fascia d'età.

La maggior parte degli incidenti, che sembrano essere di natura lieve, si verifica nella fascia d'età 26-35, con un picco significativo. Gli incidenti gravi e mortali sono molto meno frequenti e tendono a seguire un andamento simile, seppur a un livello molto inferiore. La maggior parte dei conducenti (74,6%) risiede in aree urbane, seguiti da quelli in aree rurali (14,4%) e in piccole città (11%). Questo indica che gli incidenti coinvolgono prevalentemente conducenti delle aree urbane, il che potrebbe essere dovuto a una maggiore densità di traffico, maggiore presenza di veicoli, o altri fattori di rischio associati alle aree urbane.

Dal grafico riguardante il numero di incidenti per deprivazione socio-economica del conducente si può notare che la distribuzione è relativamente uniforme, con un lieve aumento nei conducenti dei decili intermedi, da 2 a 8, che sono coinvolti in un numero maggiore di incidenti rispetto agli estremi della scala. Ciò potrebbe indicare che la deprivazione socio-economica non è un fattore distintivo nella frequenza degli incidenti, poiché non si osserva un incremento o decremento marcato in nessuna parte dello spettro socio-economico. Anche se da questa analisi iniziale non possiamo dedurre delle tendenze significative riguardo al legame tra la deprivazione socio-economica e il numero di incidenti stradali, applicando i filtri emergono, invece, dei risultati interessanti.

La Figura 7 mostra i dati filtrati per i conducenti residenti nelle aree rurali, infatti, notiamo una distribuzione diversa dalla visualizzazione che include tutte le aree geografiche: il decile 1, che rappresenta le aree con il livello più alto di deprivazione socioeconomica, ha il numero più basso di incidenti, indicando che i conducenti in queste aree sono coinvolti in meno incidenti stradali. Man mano

che ci si sposta verso i decili con un livello di deprivazione minore, il numero di incidenti aumenta, raggiungendo il massimo nel decile 7, per poi diminuire nei decili 8, 9 e 10.

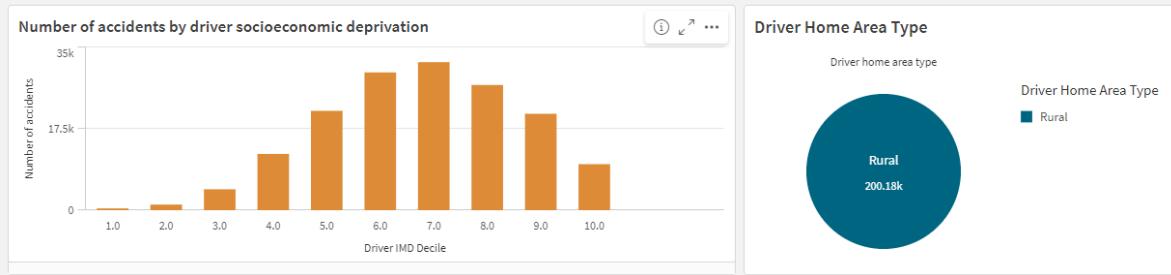


Figura 7: Foglio 2 - Analisi degli incidenti in base alla deprivazione socio-economica nelle aree rurali

Si può, quindi, dedurre che il *rischio di incidenti* nelle aree rurali è maggiore tra i conducenti di *livello socio-economico medio*. I conducenti appartenenti alle fasce medie di deprivazione potrebbero essere influenzati da una serie di fattori, ad esempio, potrebbero disporre di reddito sufficiente per possedere e mantenere un veicolo, ma non abbastanza per investire in veicoli con migliori caratteristiche di sicurezza. Al contrario, i conducenti più abbienti potrebbero permettersi veicoli più nuovi e sicuri, mentre quelli con minori risorse potrebbero non potersi permettere un veicolo affatto. I conducenti di livello socio-economico medio, inoltre, potrebbero essere dei pendolari e per raggiungere il posto di lavoro dover percorrere lunghe distanze, aumentando così la loro esposizione al rischio di incidenti. Infine, il pattern osservato potrebbe anche suggerire che nelle aree rurali ci sia una maggiore concentrazione di conducenti in specifiche fasce di reddito che coincidono proprio con i decili intermedi.

La Figura 8 mostra i dati filtrati per i conducenti residenti nelle piccole città, evidenziando una distribuzione del numero di incidenti legata alla deprivazione socioeconomica.

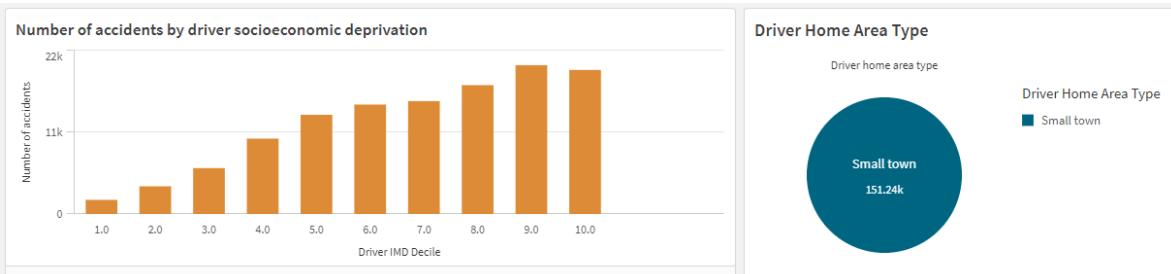


Figura 8: Foglio 2 - Analisi degli incidenti in base alla deprivazione socio-economica nelle piccole città

Il decile 1, che rappresenta le aree con il livello più alto di deprivazione socio-economica, presenta un numero di incidenti significativamente inferiore rispetto agli altri decili. Questo suggerisce che i conducenti nelle piccole città con un *alto livello di deprivazione* sono coinvolti in *meno* incidenti stradali. Il numero di incidenti aumenta progressivamente dal decile 2 al decile 10, che rappresenta le aree con il livello più basso di deprivazione socioeconomica, indicando che, nelle piccole città, i conducenti *meno deprivati* tendono a essere coinvolti in un numero *maggior* di incidenti, contrariamente a quanto si potrebbe ipotizzare.

Si osserva che i decili con una deprivazione socio-economica più alta (decili inferiori) hanno un minor numero di incidenti, il che potrebbe indicare una minore proprietà o utilizzo di veicoli in queste comunità. Invece, i conducenti con maggiori risorse economiche potrebbero tendere a viaggiare di più,

sia per motivi di lavoro che per svago, aumentando di conseguenza il loro tempo trascorso sulla strada e l'esposizione al rischio di incidenti. Potrebbe anche esserci una maggiore prevalenza di veicoli più potenti o di prestigio, che sono spesso guidati a velocità più elevate, potenzialmente contribuendo a un aumento degli incidenti. È possibile che nelle piccole città vi sia una minore presenza di misure di sicurezza stradale, come autovelox, che tendono a essere più diffusi in aree urbane densamente popolate, ciò potrebbe portare i conducenti a sottovalutare i rischi della guida ad alta velocità. Infine, il pattern osservato potrebbe anche suggerire che nelle piccole città ci sia una maggiore concentrazione di conducenti in specifiche fasce di reddito che corrispondono ai decili meno deprivati.

La figura 9 illustra i dati relativi agli incidenti stradali filtrati per i conducenti residenti nelle aree urbane.

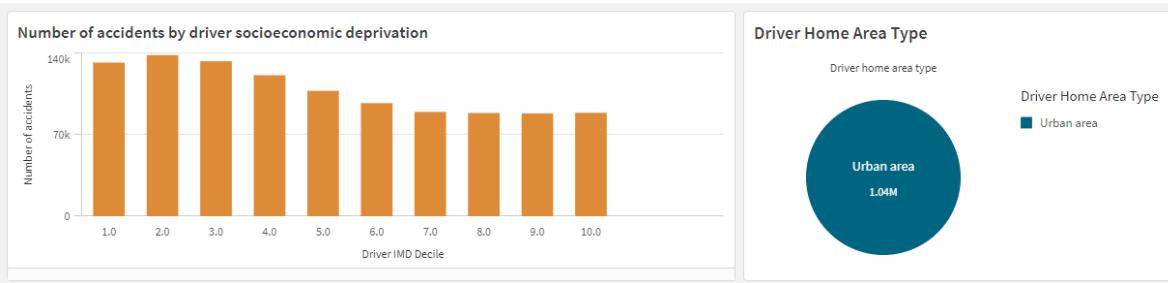


Figura 9: Foglio 2 - Analisi degli incidenti in base alla deprivazione socio-economica nelle aree urbane

A differenza della tendenza osservata per le piccole città, sembra esserci un *decremento graduale* nel numero di incidenti man mano che si passa dal decile 1 al decile 10. Questo suggerisce che, nelle aree urbane, vi è una relazione tra deprivazione socioeconomica e incidenza di incidenti stradali, infatti i conducenti nei decili di IMD *più bassi* (cioè più deprivati) sono coinvolti in un numero *maggior*e di incidenti rispetto a quelli nei decili più alti.

Le aree urbane sono caratterizzate da una maggiore densità di traffico, una diversa infrastruttura stradale e una più alta probabilità di interazioni complesse tra veicoli e pedoni. Queste zone urbane possono presentare una manutenzione stradale meno frequente, in particolare nelle aree con maggiori livelli di deprivazione socioeconomica, contribuendo così ad aumentare il rischio di incidenti. Le difficoltà economiche possono anche limitare la possibilità per i residenti di queste aree di accedere a veicoli più sicuri e a tecnologie di assistenza alla guida avanzate. Spesso, nei quartieri più deprivati si osserva una maggiore incidenza di fattori di stress psicosociale, che possono influenzare negativamente la concentrazione e la presa di decisioni durante la guida. La presenza di attività commerciali e di intrattenimento può altresì contribuire a un incremento di traffico e incidenti, specialmente durante le ore notturne. Inoltre, le aree con più alti livelli di deprivazione potrebbero essere meno equipaggiate con dispositivi di sicurezza stradale come semafori, passaggi pedonali e segnaletica adeguata, aumentando ulteriormente il rischio di incidenti. Questi fattori, insieme a una maggiore prevalenza di comportamenti di guida a rischio come la guida in stato di ebbrezza o l'eccesso di velocità, possono spiegare la tendenza osservata. In conclusione i conducenti che vivono in condizioni di maggiore deprivazione socio-economica potrebbero essere più vulnerabili agli incidenti stradali. Tali risultati forniscono un'importante base per la formulazione di *politiche di sicurezza stradale*, sottolineando l'urgenza di strategie d'intervento specifiche nelle aree urbane. È fondamentale implementare misure che non solo colmino le disparità socio-economiche, ma che migliorino complessivamente la sicurezza stradale, garantendo così una maggiore tutela a tutti gli utenti della strada.

2.3.3 Foglio 3 - Analisi degli incidenti in base alle condizioni stradali

Il terzo foglio realizzato su Qlik presenta un'analisi dei dati relativi agli incidenti stradali, con l'obiettivo di valutare l'impatto dei limiti di velocità sugli incidenti, determinare la correlazione tra tipo di strada e incidenti e esaminare la distribuzione delle vittime durante la settimana.

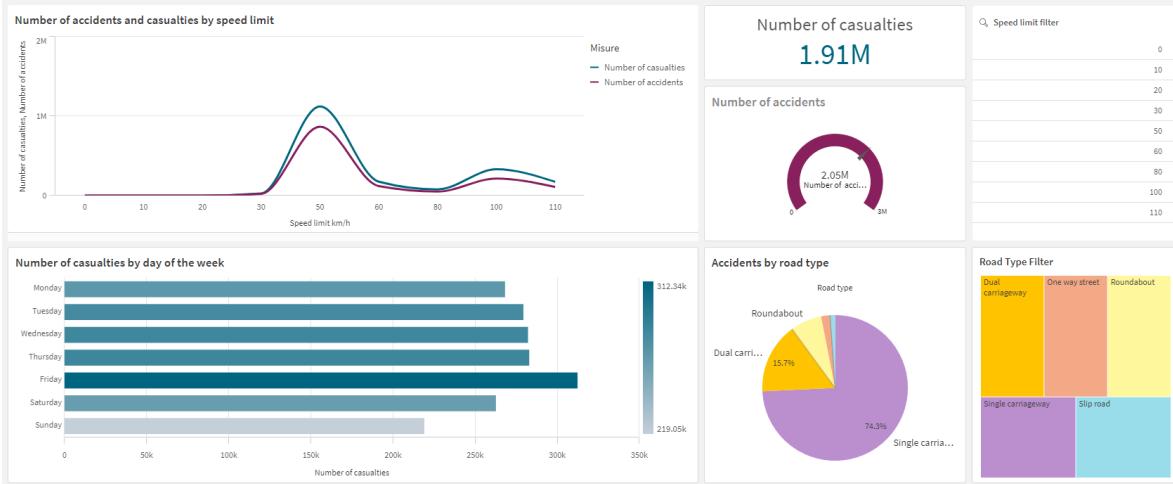


Figura 10: Foglio 3 - Analisi degli incidenti in base alle condizioni stradali

Procediamo ora con una descrizione dettagliata degli elementi costitutivi del foglio:

- la visualizzazione principale *in alto a sinistra* è un grafico lineare che mostra due linee, una che rappresenta il numero di vittime e l'altra il numero di incidenti, in relazione ai diversi limiti di velocità;
- il primo elemento *in basso a sinistra* è un grafico combinato orizzontale che raffigura la frequenza delle vittime per ogni giorno della settimana, con il sabato che mostra il numero più alto;
- la visualizzazione *in basso a destra* è un grafico a torta che mostra la distribuzione degli incidenti in base al tipo di strada, con una predominanza degli incidenti su strade a carreggiata unica;
- nella sezione *in alto sulla destra*, è presente un KPI che fornisce un conteggio totale del numero di vittime coinvolte negli incidenti e un misuratore che fornisce una rappresentazione visiva immediata del volume totale degli incidenti;
- una componente essenziale di questo foglio è rappresentata dall'area dedicata ai filtri, situata nella parte *destra* dell'interfaccia. Questa sezione permette agli utenti di affinare e personalizzare le visualizzazioni selezionando il limite di velocità e il tipo di strada.

Anche senza applicare alcun filtro, è possibile trarre alcune conclusioni diagnostiche osservando i grafici:

- *Concentrazione di incidenti*: le due linee del grafico lineare seguono un andamento simile, con un *picco* intorno al limite di *50 km/h*, e poi decrescono all'aumentare del limite di velocità. Quindi, il picco di incidenti e vittime coincide con i *limiti di velocità urbani*, che potrebbe essere dovuto a una maggiore densità di traffico o a comportamenti di guida meno prudenti in aree urbane.

- *Distribuzione temporale*: gli incidenti tendono ad aumentare verso il *fine settimana*, in particolare il venerdì, suggerendo che il comportamento di guida o le condizioni del traffico in quel giorno potrebbero contribuire ad un aumento del rischio.
- *Impatto del tipo di strada*: la maggioranza degli incidenti avviene su strade a *carreggiata singola*, il che potrebbe indicare che tali strade hanno maggiori rischi associati, forse a causa di limiti di velocità inadeguati, meno controlli o una progettazione stradale meno sicura.

La correlazione tra il limite di velocità e il tipo di strada è un'osservazione chiave per comprendere i pattern degli incidenti stradali. Generalmente, in molti contesti urbani, le strade a carreggiata singola hanno limiti di velocità inferiori per via della densità di traffico, della presenza di pedoni e della complessità dell'ambiente stradale. D'altra parte, le strade a doppia carreggiata, che possono includere autostrade e strade extraurbane, hanno di solito limiti di velocità più elevati.

Se applichiamo il filtro *Single carriageway* e selezioniamo il limite di velocità di *50 km/h*, possiamo vedere i risultati filtrati, come riportato nella Figura 11.

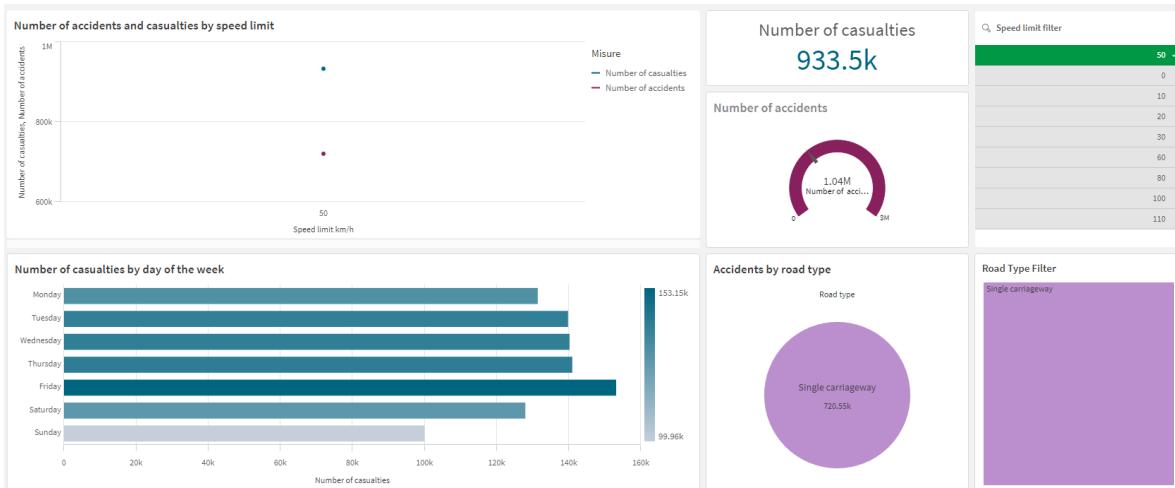


Figura 11: Foglio 3 - Analisi degli incidenti filtrati in base al limite di velocità e al tipo di strada

Il grafico relativo al numero di incidenti e vittime per limite di velocità mostra due punti in corrispondenza del limite di *50 km/h*, indicando il numero di incidenti e il numero di vittime su strade a carreggiata singola con tale limite di velocità. Questi punti suggeriscono che su strade a *carreggiata singola* con un limite di velocità di *50 km/h* c'è stato un grande numero di incidenti che hanno portato a un grande numero di vittime. Ciò può indicare che, nonostante il *basso limite di velocità*, queste strade presentano *rischi significativi*: potrebbero essere fattori come l'alta densità di traffico, la presenza di pedoni e ciclisti, scarsa visibilità o il non rispetto dei limiti di velocità che contribuiscono all'incidenza degli incidenti in queste aree. La differenza tra i giorni della settimana è meno marcatata rispetto al grafico precedente, ma sembra che il venerdì (e i giorni feriali in generale) abbiano un numero leggermente maggiore di vittime rispetto ai weekend.

Trattandosi di un trend costante nei dati, le autorità dovrebbero prendere in considerazione misure di mitigazione del rischio specifiche per queste strade, come l'installazione di dispositivi di controllo del traffico, l'incremento della segnaletica stradale, o la realizzazione di campagne di educazione per i conducenti e i pedoni.

Nella Figura 12, invece, il filtro applicato seleziona il limite di velocità di *110 km/h*.

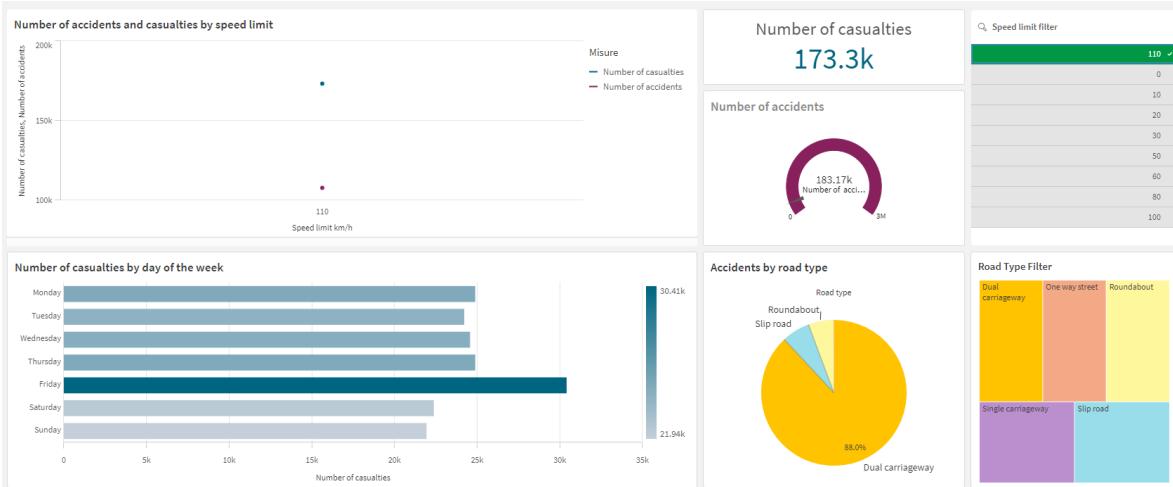


Figura 12: Foglio 3 - Analisi degli incidenti filtrati in base al limite di velocità

Dai grafici filtrati possiamo notare che:

- il grafico a linee mostra due punti in corrispondenza del limite di 110 km/h, indicando il numero di incidenti e il numero di vittime su strade con tale limite di velocità;
- la distribuzione delle vittime durante la settimana mostra che la maggior parte delle vittime si verifica durante il *fine settimana*, con il sabato come giorno con il maggior numero di vittime;
- la maggior parte degli incidenti (88.4%) avviene su strade a *doppia carreggiata*, che sono tipicamente strade ad alta velocità, come autostrade o superstrade.

A limiti di velocità elevati come 110 km/h, la maggior parte degli incidenti si verifica su strade a *doppia carreggiata* e ciò ha senso, poiché questi tipi di strade di solito permettono limiti di velocità più alti. Il fine settimana, in particolare il *sabato*, presenta il maggior numero di vittime e questo potrebbe essere dovuto all'aumento dei viaggi durante il weekend, quando le persone potrebbero viaggiare per piacere o per tornare a casa dopo una settimana di lavoro. Il numero di vittime è molto più elevato del numero di incidenti, suggerendo che gli incidenti a questa velocità hanno una *probabilità elevata* di risultare in *lesioni o decessi*.

Questi dati suggeriscono che le strade a doppia carreggiata con limiti di velocità di 110 km/h sono significativamente *pericolose*, soprattutto durante il fine settimana, e potrebbero beneficiare di interventi mirati a *migliorare la sicurezza*, come il potenziamento della segnaletica, l'impiego di pattuglie stradali, o la promozione di campagne di sensibilizzazione sulla sicurezza stradale.

3 Tableau

Esploriamo ora le analisi condotte utilizzando il secondo software, Tableau. Iniziamo descrivendo concisamente il software e la fase di caricamento dei dati. Successivamente, presentiamo un elenco delle analisi eseguite e delle relative visualizzazioni nella dashboard, ciascuna accompagnata da spiegazioni dettagliate sulle motivazioni e le conclusioni dedotte.

3.1 Introduzione

Tableau è un software di visualizzazione dei dati che offre una piattaforma versatile per esplorare, analizzare e comunicare informazioni attraverso grafici interattivi, dashboard e report. Nato con l'obiettivo di semplificare il processo di comprensione dei dati, Tableau consente agli utenti di tradurre facilmente informazioni complesse in rappresentazioni visive chiare e significative.



Figura 13: Logo di Tableau

Tra i pregi distintivi di Tableau, spicca la sua facilità d'uso, che permette anche a coloro che non sono esperti di analisi dati di creare visualizzazioni accattivanti. La piattaforma è nota per la sua versatilità, supportando diverse fonti di dati e offrendo integrazioni con numerosi strumenti e piattaforme.

Tableau Desktop è la versione dell'applicazione Tableau progettata per l'utilizzo individuale, offrendo agli utenti un ambiente locale per creare, analizzare e condividere visualizzazioni dei dati. Questo software è particolarmente apprezzato per la sua intuitiva interfaccia *drag-and-drop*, che consente agli utenti di trascinare e rilasciare facilmente le dimensioni e le misure per creare grafici e dashboard interattivi. Con Tableau Desktop, gli utenti possono connettersi a svariate fonti di dati, comprese basi di dati relazionali, fogli di calcolo, servizi cloud e molte altre.

Un suo punto di forza è senza dubbio la capacità di interazione in tempo reale con i dati, in modo da permettere agli utenti di esplorare dinamicamente le informazioni e ottenere insights immediati. In alternativa, è possibile effettuare un'estrazione grazie all' *Hyper Data Engine* di Tableau e poi operare su di essa, così da guadagnare in rapidità di esecuzione delle singole query.

3.2 Caricamento dati

La prima operazione da effettuare, come negli altri software, è il caricamento dei dati dalla sorgente. In Tableau, ciò è notevolmente semplificato grazie ad un meccanismo *drag and drop* con il quale si trascinano i file .csv nella schermata centrale dove poi saranno effettuati i *join* e rappresentate graficamente le relazioni tra le varie tabelle.

In questo caso, dunque, è stato effettuato un *join* tra i campi **Accident Index** di entrambe le tabelle, dopo averle collegate con una linea, semplicemente selezionando tali campi e l'operatore “=” dall'interfaccia fornita (Figura 14).

Una volta fatto ciò, l'interprete di Tableau si occupa di verificare che non ci siano errori di lettura dai file originali, così da poter procedere con la costruzione dei fogli e delle dashboard.

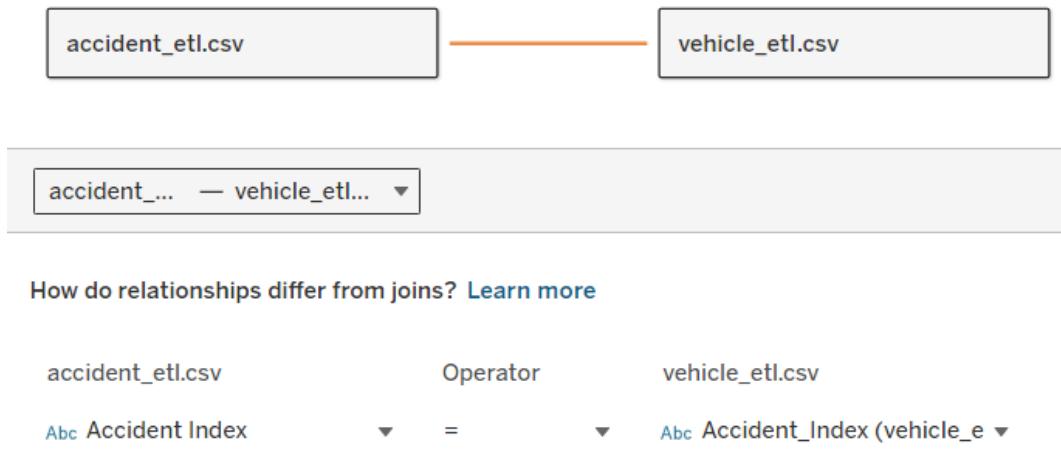


Figura 14: Creazione delle associazioni in Tableau

3.3 Data Analysis

Di seguito sono presentate le dashboard realizzate e saranno discusse le principali considerazioni effettuate su di esse. Nello specifico, le dashboard in questione sono tre, ognuna che ruota attorno ad una tematica specifica:

- Dashboard 1 - Analisi degli incidenti in base agli aspetti ambientali
- Dashboard 2 - Analisi degli aspetti relativi ai veicoli coinvolti in incidenti
- Dashboard 3 - Predizione degli incidenti negli anni e distribuzione nei giorni settimanali

In tutte le dashboard sono contenuti: al centro i grafici realizzati separatamente nei rispettivi fogli di lavoro, le legende relative ai colori utilizzati per differenziare le rappresentazioni nella parte sinistra e le finestre con i filtri interattivi selezionabili dall'utente nella parte destra.

3.3.1 Dashboard 1 - Analisi degli incidenti in base alle condizioni ambientali

La prima analisi effettuata con Tableau è di tipo diagnostico e nasce dall'esigenza di evidenziare l'impatto delle condizioni ambientali sul numero totale degli incidenti e, sulla base di ciò, individuarne le cause.

La dashboard, come riportato nella Figura 15, è divisa in tre fogli:

- *in alto* si trova un istogramma che mostra la correlazione tra numero di vittime e quantità di luce presente al momento dell'impatto, distribuite in base all'orario dell'incidente;
- *in basso a sinistra* è presente una rappresentazione a bolle della quantità di vittime in base alle condizioni della superficie stradale (asciutta, bagnata, ghiacciata, etc.) ed al tipo di strada (a singola corsia, a doppia corsia, etc.);
- *in basso a destra* si trova un grafico che divide gli incidenti in base alla zona (urbana o rurale) ed evidenzia la loro gravità.

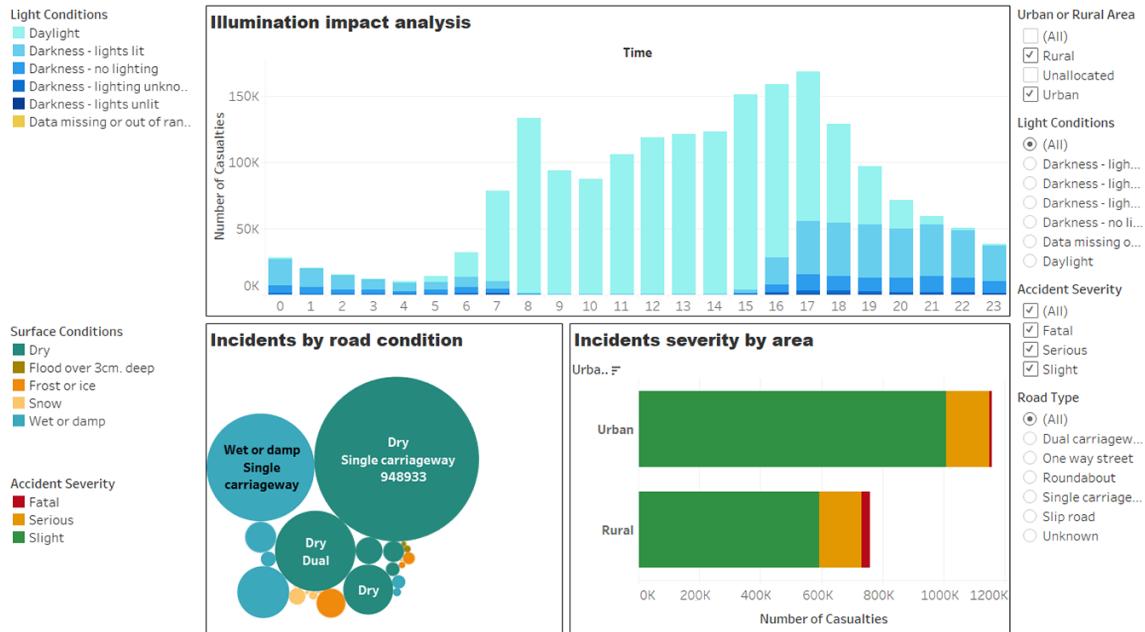


Figura 15: Dashboard 1 - Analisi degli incidenti in base alle condizioni ambientali

Anche senza applicare alcun filtro, è già possibile effettuare alcune considerazioni osservando i grafici.

Iniziando dal grafico in basso a destra è immediato notare come la maggior parte dei coinvolgimenti in incidenti avvenga nel contesto urbano, mentre in ambiente rurale, pur essendo stati rilevate meno vittime di incidenti, troviamo una percentuale maggiore di incidenti gravi o fatali.

Selezionando la gravità interessata dalla legenda a sinistra, è possibile, infatti, notare come le vittime di incidenti gravi in contesto urbano rilevate siano pari a 139.817 e pari a 139.474 fuori città. Questa differenza si fa ancora più netta se andiamo a considerare gli incidenti fatali, con valori rispettivamente di soltanto 9.913 in città a fronte dei 25.781 nelle zone rurali. Questo ci fa pensare che in città, a causa dei bassi limiti di velocità, è più probabile essere vittima di incidenti leggeri, dovuti anche agli alti volumi di traffico o alle distrazioni alla guida.

Non ci stupirà, inoltre, constatare, osservando l'istogramma in alto, come la maggior parte degli incidenti avvenga negli orari di punta del traffico, specialmente lavorativo. Possiamo trovare, infatti, valori complessivamente alti tra le 16 e le 18 di pomeriggio e alle 8 di mattina, rispettivamente orari di rientro e di partenza dal lavoro per molti inglesi.

Oltre a ciò, utilizzare i filtri ci consente di compiere analisi più dettagliate riguardo alle cause di incidenti. Nello specifico, i filtri presenti in tale dashboard sono:

- filtro sull'area: urbana o rurale;
- filtro sulle condizioni di illuminazione;
- filtro sulla severità dell'incidente (leggero, serio o fatale);
- filtro sulla tipologia di strada (a singola corsia, a due corsie, rotonda, etc).

Un aspetto importante da analizzare, infatti, è come l'*assenza di luce* (sia naturale che artificiale) e una *superficie stradale bagnata* abbiano una forte influenza sulla gravità degli incidenti.

Nella Figura 16 abbiamo, infatti, applicato due filtri alla dashboard: con primo sulla quantità di luce si selezionano soltanto gli incidenti avvenuti in luoghi privi di illuminazione e il secondo serve a mostrare soltanto quelli seri e fatali, rispettivamente in arancio e rosso.

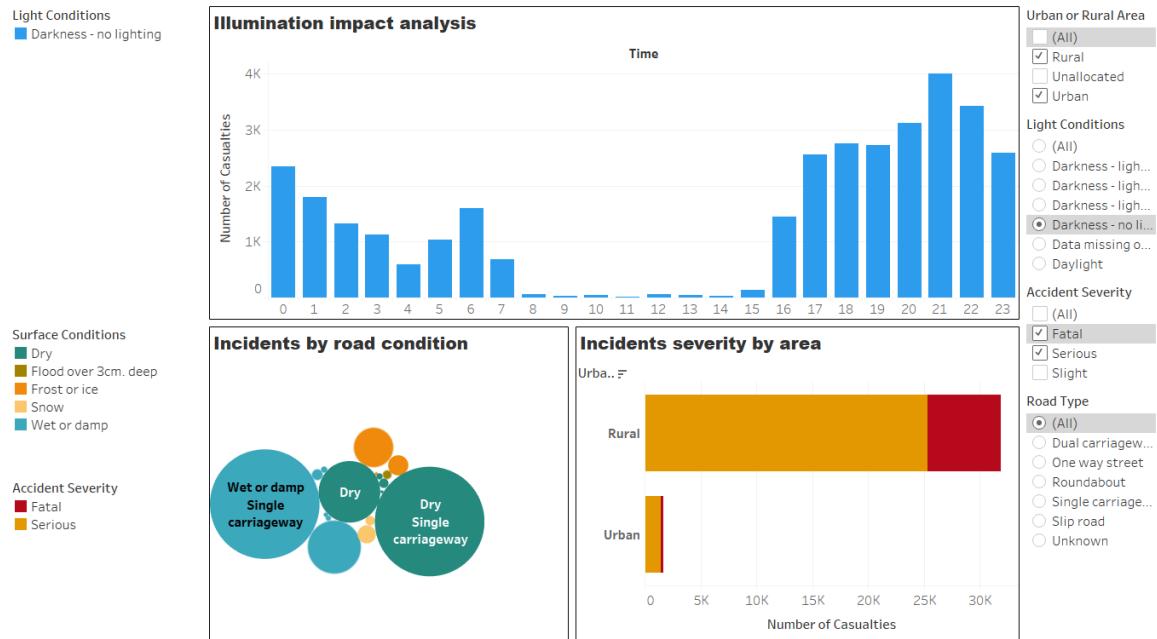


Figura 16: Dashboard 1 filtrata per illuminazione della strada e severità degli incidenti

Dunque, attraverso questa visualizzazione possiamo notare come la maggior parte degli incidenti senza illuminazione presente avvenga fuori dalle città, spesso illuminate in ogni angolo. Troviamo infatti un totale di 25.333 incidenti seri in contesti rurali e 6.552 fatali contro i rispettivi 1.452 e 241 delle città, nettamente inferiori.

Inoltre, andando ad osservare il grafico a bolle, differentemente rispetto al caso senza filtri, dove le strade asciutte a singola corsia comprendevano la maggior parte dei valori, qui abbiamo una sottile differenza tra queste e quelle bagnate della stessa tipologia, che addirittura superano di pochissimo quelle asciutte (12.072 contro 12.019). Un avvicinamento importante, se si rapporta ai rispettivi 72.114 e 163.908 per le strade a singola corsia bagnate e asciutte se si tengono in considerazione tutte le condizioni di luce.

Appare, quindi, evidente come in assenza di illuminazione, una strada bagnata, a causa degli spazi di frenata ridotti, renda più difficile evitare un sinistro, soprattutto in casi di frenate di emergenza dovute a imprevisti.

3.3.2 Dashboard 2 - Analisi degli aspetti relativi ai veicoli coinvolti in incidenti

La seconda dashboard realizzata con Tableau è anch'essa di tipo diagnostico e il suo scopo è quello di analizzare alcuni degli aspetti relativi ai veicoli coinvolti in incidenti, così da determinare quali possano essere stati i fattori scatenanti.

Tale dashboard, come riportato nella Figura 17, è divisa in tre fogli:

- *in alto* si trova una matrice di correlazione che incrocia le corrispondenze tra il primo punto di impatto del veicolo coinvolto in un incidente e la manovra che stava compiendo subito prima;

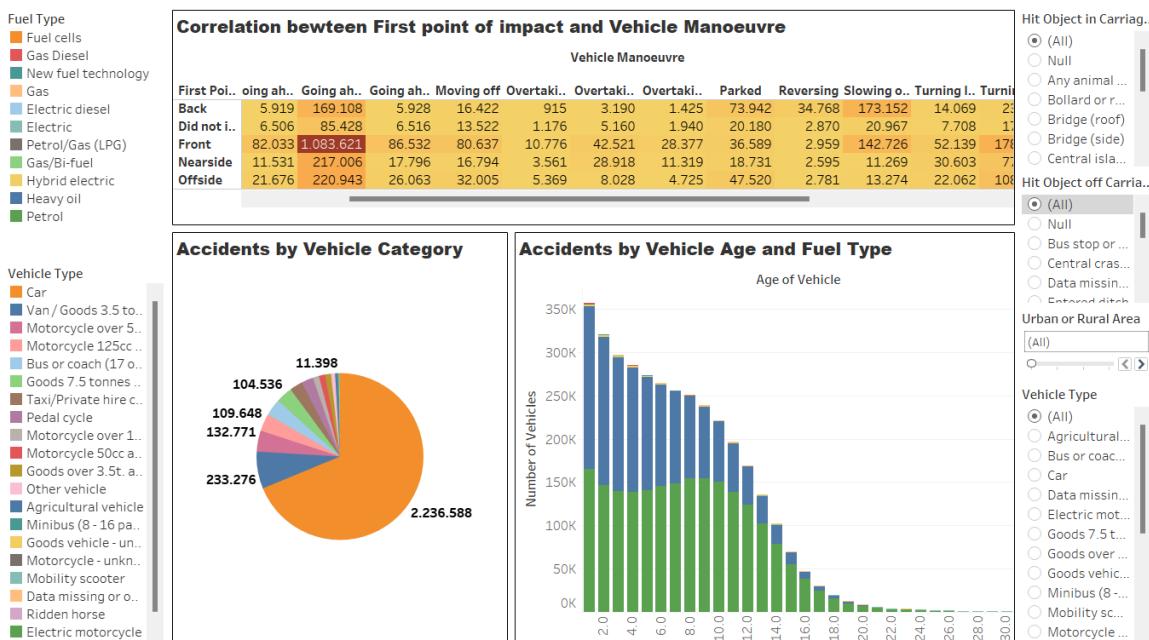


Figura 17: Dashboard 2 - Analisi degli aspetti relativi ai veicoli coinvolti in incidenti

- in basso a sinistra è presente una rappresentazione a torta che suddivide gli incidenti in base alla categoria dei veicoli coinvolti, ordinati in senso orario decrescente;
- in basso a destra si trova un istogramma che mostra la distribuzione degli incidenti in base all'età del veicolo coinvolto al momento della rilevazione, suddivisi per tipologia di carburante.

Con un primo sguardo, senza applicare i filtri, in basso a sinistra appare immediato come si distinguano gli incidenti compiuti con autovetture che ne costituiscono l'assoluta maggioranza (2.236.588 coinvolgimenti), superando di quasi dieci volte la seconda categoria costituita dai furgoni e camion con un peso inferiore a 3.5 tonnellate (233.276 coinvolgimenti).

Oltre a ciò, è possibile notare dall'istogramma come tra i veicoli più vecchi abbiamo una maggioranza di incidenti avvenuta con veicoli a benzina, che in passato costituivano la maggioranza delle vendite, per poi lasciare il primato ai veicoli a diesel con le ultime vendite. Si deve tenere, infatti, a mente che il periodo considerato nel dataset (2005-2017) rappresenta l'epoca del boom dei motori diesel ma, tra i veicoli più recenti (da 1 a 4 anni di vita), si può notare come inizino ad essere coinvolti anche veicoli ibridi ed elettrici, anticipando il trend di crescita delle loro vendite negli anni successivi.

Osservando invece la matrice, a colpo d'occhio si può osservare come la stragrande maggioranza degli incidenti sia avvenuta *marchiando in avanti*, con il picco che riguarda il *frontale* come primo punto di impatto (1.083.621 veicoli). Valori notevoli sono stati registrati anche tra i veicoli che stanno compiendo una frenata: rispettivamente 173.152 di quelli che hanno subito un impatto da dietro e 142.726 di quelli che hanno impattato frontalmente. Un altro picco è, inoltre, raggiunto dai veicoli fermi in coda che sono stati tamponati nel posteriore (236.264), probabilmente dovuto all'intenso traffico nelle aree urbane ed alle distrazioni durante le code nelle ore di punta.

È anche interessante notare come ci sia una *differenza importante* tra i veicoli coinvolti in un incidente durante una *svolta a destra* e quelli coinvolti durante una *svolta a sinistra*. Basti pensare

che, tra quelli che hanno impattato frontalmente, i veicoli che svoltavano a destra sono stati 178.063 mentre quelli a sinistra 52.139.

Questi numeri trovano una giustificazione nella regolamentazione delle strade prevista nel Regno Unito che, come tutti i paesi anglosassoni, prevede che la *circolazione per senso di marcia avvenga a sinistra* invece che a destra come avviene negli altri paesi, comportando quindi una peggiore visibilità e un maggiore pericolo nelle manovre di svolta a destra.

Come per la precedente dashboard, anche in questa è possibile applicare alcuni filtri per andare più nel dettaglio con le analisi. I filtri che si trovano in tale dashboard riguardano:

- filtro sull'oggetto colpito fuori dalla carreggiata (se presente, altrimenti sarà *null*);
- filtro sull'oggetto colpito fuori dalla carreggiata (se presente, altrimenti sarà *null*);
- filtro a scorrimento orizzontale sull'area (urbana o rurale);
- filtro sulla tipologia di veicolo.

Utilizzando i filtri, è, quindi, possibile indagare sulle cause degli incidenti attraverso i comportamenti dell'autista e sulle differenze tra i vari veicoli. Nella Figura 18, possiamo notare l'applicazione come filtro del veicolo la categoria "taxi/veicolo a noleggio privata" e la selezione esclusiva delle aree urbane attraverso il filtro a scorrimento.

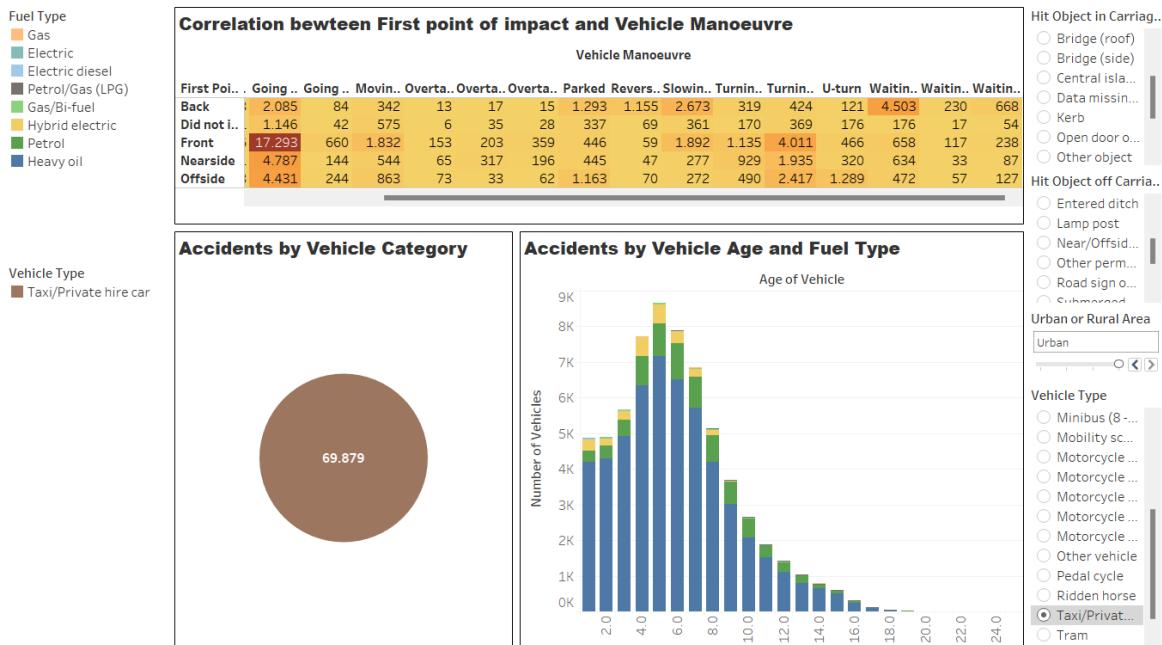


Figura 18: Dashboard 2 filtrata per categoria di veicolo e per tipo di area

Per prima cosa, è possibile visualizzare immediatamente dal grafico a torta che il numero totale di incidenti in cui è coinvolta questa categoria di veicoli si attesta a 69.879.

Dall'istogramma, inoltre, possiamo constatare che, a differenza della dashboard non filtrata, in questo caso *l'età media* dei taxi e dei veicoli a noleggio coinvolti in incidenti si *alza notevolmente*, con un picco che si attesta tra i 4 e i 6 anni. Questo probabilmente perché i veicoli appartenenti a questa categoria sono molto più sfruttati nel lungo termine da parte dei loro proprietari.

Un altro elemento degno di nota è la *preponderanza di veicoli diesel* rispetto a tutti gli altri, questo perché, rispetto alle corrispettive versioni a benzina sono caratterizzati da un maggiore rendimento che si traduce in un minor costo di carburante. Si può inoltre notare come in questo caso sia molto più tangibile la presenza dei veicoli ibridi, sia per le motivazioni sopra citate, sia per la presenza di incentivi statali previsti per l'acquisto di tali veicoli che, nelle zone maggiormente urbanizzate, hanno favorito una maggiore diffusione tra i taxi, anticipando il trend degli anni successivi.

3.3.3 Dashboard 3 - Predizione degli incidenti negli anni e distribuzione nei giorni settimanali

La terza e ultima dashboard realizzata sfrutta le funzionalità predittive di Tableau per provare ad individuare le *trendline* (linee di tendenza) e le stime negli anni successivi rispettivamente del numero di vittime totali e del numero medio di vittime per incidente. A scopo didattico, essendo il dataset in analisi limitato superiormente al 2016, abbiamo ritenuto opportuno confrontare le previsioni effettuate dal software con il dataset riportante i valori reali degli incidenti relativi agli anni 2021-2022.

La dashboard in questione, come riportato nella Figura 19, è divisa in due fogli:

- *in alto* troviamo il grafico che mostra l'andamento temporale, diviso per mesi, del numero totale e medio di vittime di incidenti per ogni anno incluso nella raccolta, con le rispettive previsioni e trendline;
- *in basso* è presente una suddivisione del numero totale di vittime in base al giorno della settimana e allo scopo del tragitto, ordinati in senso decrescente.

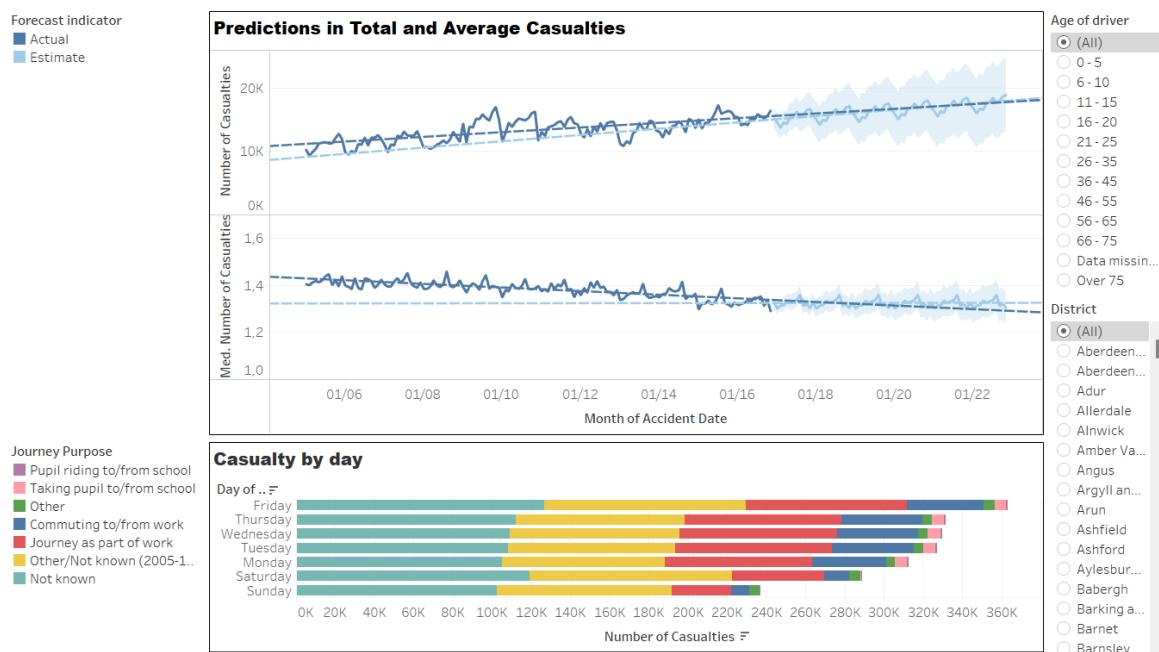


Figura 19: Dashboard 3 - Predizione degli incidenti negli anni e distribuzione nei giorni settimanali

Prima di iniziare a descrivere il primo grafico partendo dalle trendline, è necessario fare alcuni chiarimenti riguardo agli strumenti statistici utilizzati dal software per effettuare tali misurazioni, ossia il *p-value* e l'*r-square*.

- *p-value*: è una misura che ci aiuta a capire se i risultati di uno studio statistico sono significativi o se potrebbero essere dovuti al caso. In parole semplici, esso ci dice la probabilità di ottenere un risultato simile o più estremo rispetto a quello osservato, supponendo che l'ipotesi nulla sia vera, rappresentata da un'affermazione del tipo “non si riscontra alcun effetto” oppure “non ci sono differenze” ed è l'ipotesi che si cerca di contraddirsi. Dunque, più il valore del p-value è *piccolo*, più i dati osservati sono statisticamente *significativi*. Tableau, ad esempio, considera accettabile un p-value < 0.0001.
- *r-square*: è una misura di quanto bene le variazioni di una variabile dipendente siano spiegate dalle variazioni nelle variabili indipendenti in un modello di regressione. In altre parole, indica quanto della variazione nei dati della variabile dipendente viene spiegato dal modello. Di solito, più è *grande* il valore dell'r-square, più il modello ha un *alto potere predittivo*. Dunque, un r-square vicino a 0 suggerisce che il modello non spiega bene le variazioni nei dati e potrebbe non essere utile, al contrario, se più vicino a 1, indica che il modello spiega una grande parte della variazione nei dati.

Per quanto riguarda invece le trendline, esse possono essere generate tramite modello: lineare, logaritmico, esponenziale, di potenza e polinomiale. Nel primo grafico presente nella dashboard, abbiamo utilizzato il modello *lineare* per generare le trendline e sono stati ottenuti tali risultati:

- la trendline del numero totale di vittime ha come valori un p-value < 0.0001 e un r-square pari a circa 0,5 e pertanto abbastanza significativo;
- la trendline del numero medio di vittime ha come valori un p-value < 0.0001 e un r-square pari a circa 0,6 e pertanto anche più significativo della precedente.

Quindi, si può affermare che le due trendline analizzano efficacemente la tendenza negli anni del numero di vittime di incidenti per il periodo considerato.

Oltre a ciò, è opportuno parlare delle previsioni effettuate da Tableau su tali dati, evidenziate in celeste, a differenza dei valori attuali in blu scuro.

Prima di fare la predizione vera e propria, è stato scelto l'intervallo temporale su cui impostare il modello predittivo fornito da Tableau, scegliendo il range temporale partendo dal 2017 fino al 2022.

Una volta fatto ciò, si può notare, come riportato nella Figura 20, che, dopo aver cliccato su un qualsiasi punto dell'andamento della stima, il software riporta anche il *limite superiore e inferiore* dei valori stimati e va a rappresentare tale intervallo con la sfumatura celeste che si pone sullo sfondo.

Abc	Forecast indicator	Estimate
📅	Month of Date	09/21
#	Lower Prediction Interval ...	12.270
#	Number of Casualties	17.498
#	Upper Prediction Interval ...	22.727

Figura 20: Stima effettuata da Tableau dei valori del numero totale di vittime previsto per settembre 2021, con i rispettivi limiti superiori e inferiori

In questo caso, quindi, è possibile osservare come sia riportato 17.498 come valore previsto della somma del numero di vittime relativo a settembre 2021, mentre i limiti inferiore e superiore dell'intervallo calcolati sono rispettivamente 12.270 e 22.727.

Dunque, a scopo didattico, abbiamo potuto confrontare tali stime con i valori effettivi analizzando il dataset relativo al 2021-2022 con la creazione di un foglio per ottenerne una visualizzazione immediata (Figura 21).

Total and Average Casualties (2021-2022)

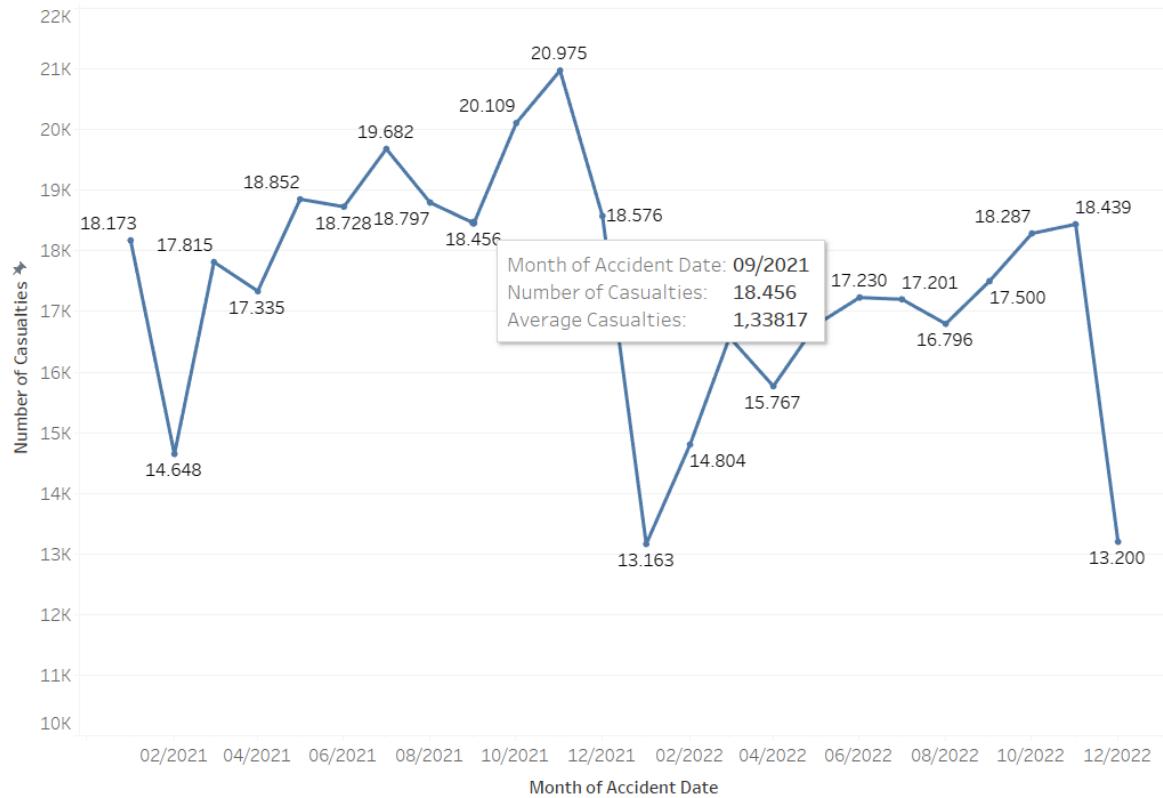


Figura 21: Andamento del numero totale di vittime di incidenti in UK negli anni 2021-2022, con enfasi sul mese di settembre 2021

Nel caso di settembre 2021, quindi, la stima effettuata da Tableau è leggermente ottimistica rispetto al dato reale, infatti la stima differisce di 958 vittime in meno rispetto al valore riportato nell'andamento reale di quell'anno.

Se si osserva però la stima del valore medio, coerentemente con le considerazioni fatte sui parametri statistici, si può notare come il valore medio di vittime per lo stesso mese sia di circa 1.338, che coincide quasi totalmente con il valore stimato (1.313).

Di seguito sono riportate le tabelle complete delle sue stime effettuate, da cui si può notare che, coerentemente con quanto riportato sopra, le stime migliori riguardano i valori medi piuttosto che quelli totali. Infine, osservando le tabelle si può anche affermare come in generale le stime siano state perlopiù *ottimistiche* nella maggior parte dei casi, avendo ottenuto molto spesso una differenza a favore del valore reale.

Tabella 3: Confronto tra stime e valori reali per il 2021 e 2022 (Valori Totali)

Mese e Anno	Stima	Valore Reale	Differenza
gennaio 2021	14553	18173	-2620
febbraio 2021	13728	14648	-920
marzo 2021	14555	17815	-3240
aprile 2021	14332	17335	-3003
maggio 2021	15433	18852	-3419
giugno 2021	15528	18728	-3200
luglio 2021	16239	19682	-3443
agosto 2021	15461	18797	-3336
settembre 2021	15734	18456	-2722
ottobre 2021	16458	20109	-2651
novembre 2021	16699	20975	-2276
dicembre 2021	15639	18576	-2937
gennaio 2022	14994	13163	1831
febbraio 2022	14169	14804	-635
marzo 2022	14996	16575	-1579
aprile 2022	14773	15767	-994
maggio 2022	15874	16775	-901
giugno 2022	15969	17230	-1261
luglio 2022	16680	17201	-521
agosto 2022	15902	16796	-894
settembre 2022	16175	17500	-1325
ottobre 2022	16899	18287	-1388
novembre 2022	17140	18439	-1299
dicembre 2022	16080	13200	2880

Tabella 4: Confronto tra stime e valori reali per il 2021 e 2022 (Valori Medi)

Mese e Anno	Stima	Valore Reale	Differenza
gennaio 2021	1,3017	1,3545	-0,0528
febbraio 2021	1,3145	1,3377	-0,0232
marzo 2021	1,3158	1,3494	-0,0336
aprile 2021	1,3350	1,3634	-0,0284
maggio 2021	1,3257	1,3650	-0,0393
giugno 2021	1,3319	1,3439	-0,0120
luglio 2021	1,3374	1,3764	-0,0390
agosto 2021	1,3607	1,4012	-0,0405
settembre 2021	1,3134	1,3382	-0,0248
ottobre 2021	1,3204	1,3556	-0,0352
novembre 2021	1,3041	1,3556	-0,0515
dicembre 2021	1,3277	1,3550	-0,0273
gennaio 2022	1,3017	1,3207	-0,0190
febbraio 2022	1,3145	1,3538	-0,0393
marzo 2022	1,3158	1,3431	-0,0273
aprile 2022	1,3350	1,3699	-0,0349
maggio 2022	1,3257	1,3559	-0,0302
giugno 2022	1,3319	1,3448	-0,0129
luglio 2022	1,3374	1,3594	-0,0220
agosto 2022	1,3607	1,3895	-0,0288
settembre 2022	1,3134	1,3503	-0,0369
ottobre 2022	1,3204	1,3512	-0,0308
novembre 2022	1,3041	1,3536	-0,0495
dicembre 2022	1,3470	1,3818	-0,0348

Tornando alla dashboard nel suo complesso, si può osservare come siano presenti due filtri: il primo per l'età delle persone coinvolte negli incidenti e l'altro per il distretto dove è avvenuto il sinistro.

Attraverso questi filtri si possono fare delle interessanti considerazioni se effettuate osservando anche il secondo grafico della dashboard; esso visualizza, infatti, il numero totale degli incidenti per giorno della settimana, suddividendoli con la legenda colorata in base allo scopo del tragitto.

A scopo dimostrativo, nella Figura 22 viene effettuato un filtraggio per età selezionando la fascia 46-55.

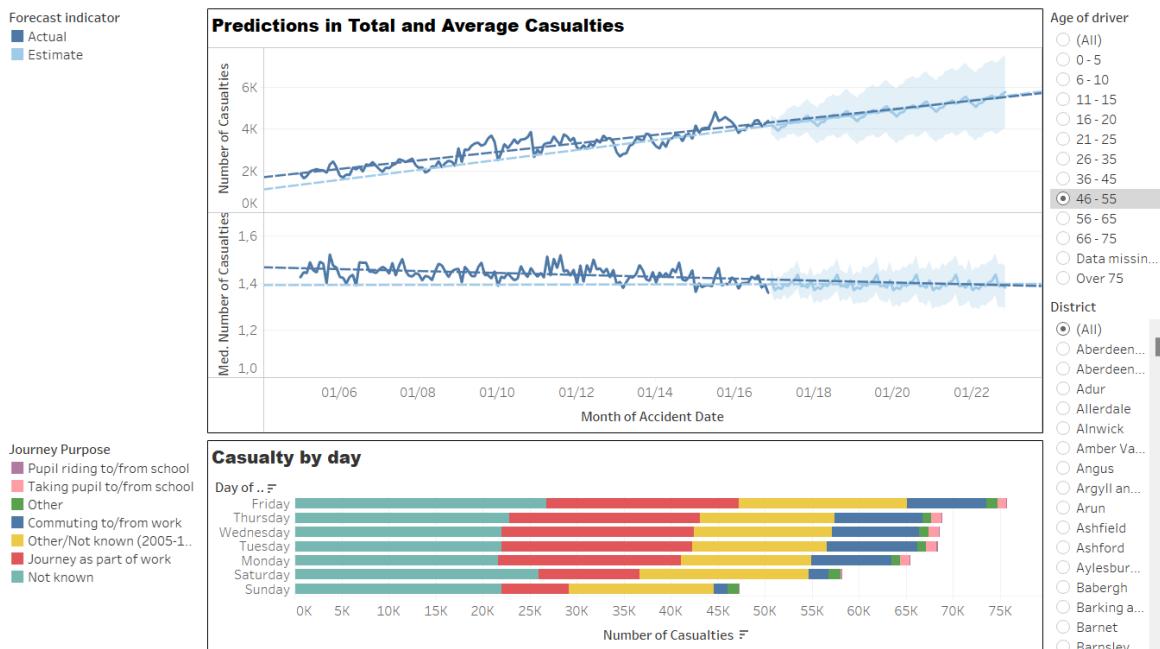


Figura 22: Dashboard 3 filtrata per età delle vittime

Contestualmente, si può notare come in tale fascia di età, pur rimanendo il venerdì il giorno caratterizzato dal maggior numero di incidenti, aumenti molto il numero di coloro che sono rimasti coinvolti coloro che viaggiavano per lavoro (evidenziati in rosso). Il valore più alto tra gli incidenti annoverabili in questa categoria si attesta, infatti, a 20.485 di venerdì e il valore più basso è 7.126 della domenica.

Infine, è possibile osservare come in questo caso la trendline del numero totale sia più ripida, evidenziando un maggiore aumento nel tempo delle vittime relative a questa fascia d'età, mentre, quella del valor medio, si mantiene più orizzontale rispetto al caso generale.

4 Power BI

Vediamo nel seguito le analisi svolte con il terzo ed ultimo software utilizzato. Innanzitutto descriviamo brevemente la fase di caricamento dei dati evidenziando i problemi riscontrati e le soluzioni adottate per fronteggiarli. Successivamente riportiamo un elenco di analisi svolte e delle loro rappresentazioni in dashboard, ciascuna correlata di motivazioni e conclusioni tratte.

4.1 Introduzione

Power BI è una soluzione di business intelligence sviluppata da Microsoft che offre una vasta gamma di strumenti per la visualizzazione e l'analisi dei dati. Una delle caratteristiche chiave di Power BI è la sua integrazione con altre applicazioni Microsoft come Excel e Azure, facilitando il lavoro per gli utenti già immersi nell'ecosistema Microsoft.



Figura 23: Logo di Power BI

Power BI utilizza un modello di dati colonnare altamente compresso per garantire prestazioni elevate durante la manipolazione di grandi quantità di dati. Offre strumenti di manipolazione dati intuitivi sia per la loro trasformazione che per la creazione di misure a partire da essi. In effetti, proprio per quanto riguarda il processo di *ETL*, Power BI offre, rispetto agli altri due software, un ambiente avanzato ricco di possibilità nella pulizia dei dati. L'*editor di query* Power Query, mette a disposizione dell'utente un ambiente grafico per la creazione di script M volti alla modifica dei dati prima della loro importazione nel modello. Anche per quanto riguarda la creazione di misure Power BI spicca rispetto agli altri due software con il suo linguaggio DAX, con oltre 250 funzioni per la creazione di formule. [3]

4.2 Caricamento dati

A seguito dell'installazione del software, è stato creato un nuovo progetto e come prima cosa sono stati importati i due dataset.

Per quanto riguarda il primo dataset, come è mostrato in figura 24, l'operazione di default sul riconoscimento del tipo di dato delle colonne eseguita da Power BI è stata subito rivista, questo perché non è risultata efficace per tutte le colonne. In particolare, il software faticava a riconoscere i campi di `latitude` e `longitude` come dati numerici decimali. Questo problema è stato risolto utilizzando il comodo strumento "Sostituisci valori" che ha consentito di sostituire tutti i "." separatori dei numeri decimali in ",", cosa non già permessa dal formato CSV con cui il dataset è codificato. A seguito di questa operazione, il tipo di dato è stato manualmente cambiato a numero decimale, e successivamente è stato possibile modificare le categorie delle due colonne come campi di latitudine e longitudine rispettivamente.

Un'altra operazione che è stata effettuata in questa fase riguarda l'unificazione delle colonne `Date` e `Time` nella nuova colonna `DateTime`. In questo modo, abilitando la *Funzionalità di Business Intelligence per le gerarchie temporali*, si riesce a lavorare molto efficientemente con esse in fase successiva.

Per quanto riguarda invece il secondo dataset, non sono state eseguite operazioni aggiuntive rispetto a

```

= let
    Origine = Csv.Document(File.Contents("accident_etl_10.csv"),
        [Delimiter=",", Columns=29, Encoding=1252, QuoteStyle=QuoteStyle.None]),

    #"Intestazioni alzate di livello" = Table.PromoteHeaders(
        Origine, [PromoteAllScalars=true]),

    #"Modificato tipo" = Table.TransformColumnTypes(
        #"Intestazioni alzate di livello", [{"Accident_Index", type text}, ...]),

    #"Sostituito valore" = Table.ReplaceValue(
        #"Modificato tipo", ".", ",",
        Replacer.ReplaceText, {"Latitude", "Longitude"}),

    #"Modificato tipo1" = Table.TransformColumnTypes(
        #"Sostituito valore", [{"Longitude", type number}, {"Latitude", type number}]),

    #"Data e ora unite inserite" = Table.AddColumn(
        #"Modificato tipo1", "DateTime", each [Date] & [Time], type datetime),
in
    #"Rinominate colonne"

```

Figura 24: Importazione del dataset sugli incidenti in Power BI

quelle già predisposte in automatico dal software.

Una volta importati i dati, nel pannello *Visualizzazione Modello* è stato possibile specificare la modalità di collegamento dei due dataset, come illustrato in figura 25. Fedelmente alle modifiche apportate nella fase di *ETL* iniziale, la relazione di associazione tra le due omonime colonne *Accident_Index* è stata specificata di tipo uno-molti. Questo perché ad ogni incidente possono essere associati più veicoli e ad ogni veicolo è associato un unico incidente.

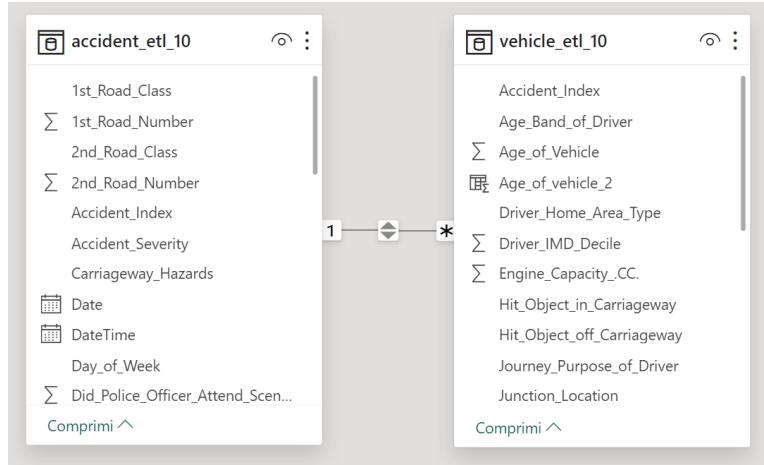


Figura 25: Creazione delle associazioni tra i due dataset

4.3 Data Analysis

Vediamo nel seguito le dashboard realizzate e le conclusioni che possono essere tratte dalle analisi svolte. Nel dettaglio, le dashboard realizzate sono quattro:

- Dashboard 1 - Panoramica annuale della distribuzione degli incidenti su base temporale e spaziale

- Dashboard 2 - Confronto della pericolosità stradale e cause di incidenti tra due distretti
- Dashboard 3 - Analisi della pericolosità degli incidenti in relazione ai tipi di veicoli coinvolti
- Dashboard 4 - Clustering dei tipi di veicoli coinvolti negli incidenti

4.3.1 Dashboard 1 - Panoramica annuale della distribuzione degli incidenti su base temporale e spaziale

La prima analisi realizzata dopo aver importato i dati è di tipo descrittivo, utile sia per familiarizzare con le funzionalità principali del software sia per avere una visione ad alto livello dell'intero dataset.

L'obiettivo di questa dashboard, mostrata in figura 26, è quello determinare l'andamento della frequenza degli incidenti, sia su base spaziale (come mostra la *Mappa colorata* a distretti sulla sinistra), sia su base temporale (come mostrano i due *Grafico ad aree* e *Istogramma a colonne raggruppate* sulla destra).

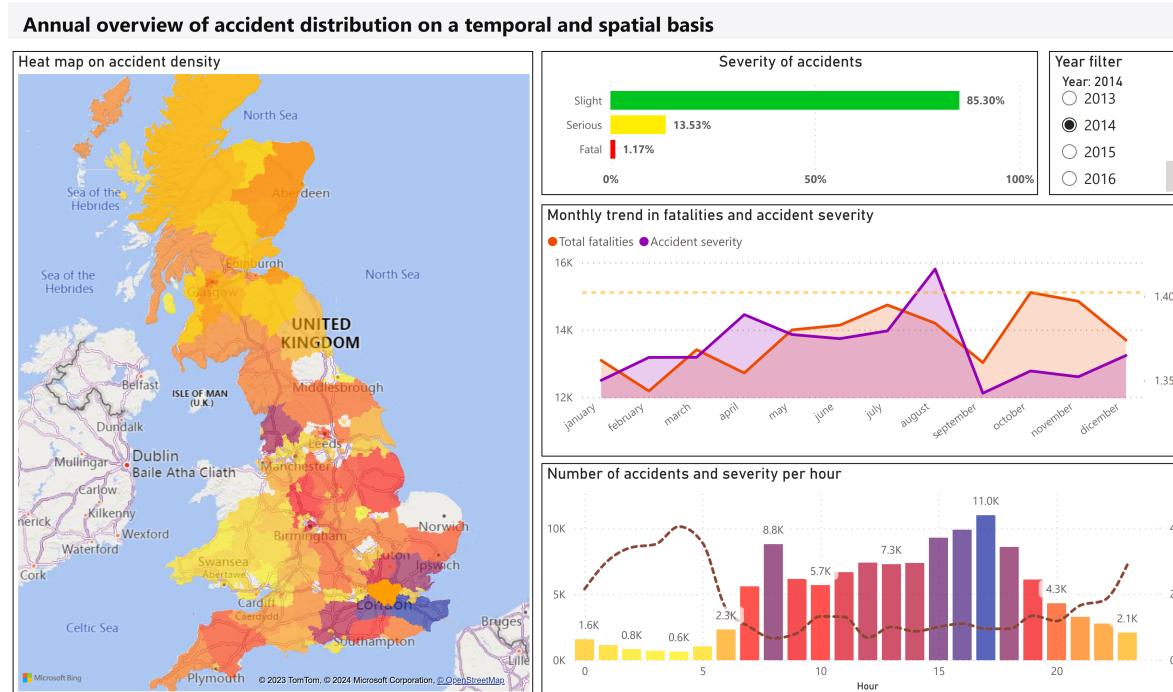


Figura 26: Panoramica annuale della distribuzione degli incidenti su base temporale e spaziale

Dalla *distribuzione spaziale* è possibile notare come la zona del Regno Unito più soggetta ad incidenti stradali è quella meridionale. Tale fenomeno può essere attribuito, innanzitutto, alla significativa densità demografica caratteristica di queste aree, con particolare riferimento ai dintorni della metropoli di Londra. La complessità della rete viaria e la presenza di numerosi veicoli e pedoni sono fattori che concorrono a creare un contesto in cui gli incidenti risultano più frequenti.

Si può però notare come nelle zone situate ad estremo nord, sebbene meno densamente popolate, il numero di incidenti non è poi così basso. Ciò lo si può ricondurre, come poi dimostreremo nel dettaglio nella seconda dashboard, alle condizioni meteorologiche di queste zone, caratterizzate da condizioni meteorologiche più avverse, quali il rischio di ghiacciamento e la presenza frequente di nebbia.

Rivolgendo ora l'attenzione alla *distribuzione temporale*, possiamo notare alcuni aspetti interessanti. Il grafico *Monthly trend in fatalities and accident severity* sulla destra della dashboard, mostrato senza

filtri in figura 27, mostra l'andamento della frequenza di incidenti e della loro gravità; calcolata come media aritmetica di morti per incidente.

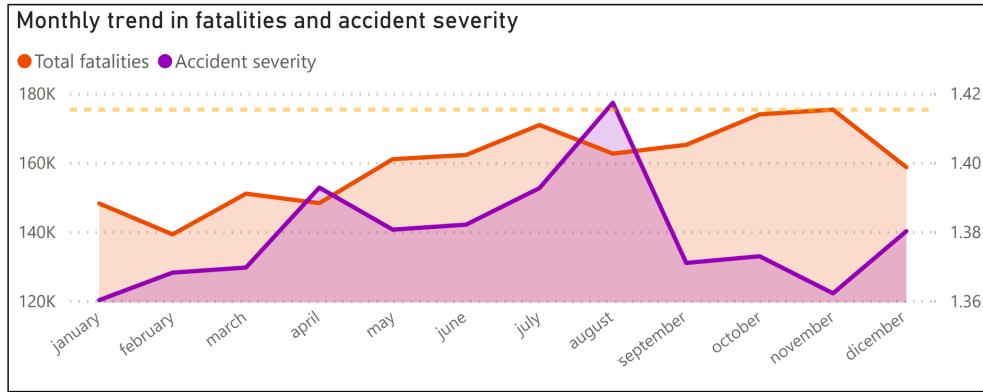


Figura 27: Andamento mensile della frequenza degli incidenti

Con 175.345 vittime, novembre ha registrato il numero più alto di vittime totali ed è stato superiore del 25,95% rispetto a febbraio, che ha registrato il numero più basso con 139.219 vittime totali. Novembre da solo, ha rappresentato il 9,15% del totale dei decessi nel decennio in esame. Infine, il totale dei decessi e la gravità degli incidenti divergono maggiormente quando il mese è novembre, quando la gravità raggiunge i suoi valori minimi.

In altre parole, possiamo affermare che si registra un *aumento della micro-incidentalità* nella stagione autunnale.

Come è possibile osservare in figura 28, una prima causa potrebbe essere ricondotta alla rilevante diversificazione della durata del tempo di luce al variare del periodo dell'anno, dovuta alla posizione geografica del Paese. Un giorno di giugno può contare fino 17 ore di luce, mentre nel periodo di

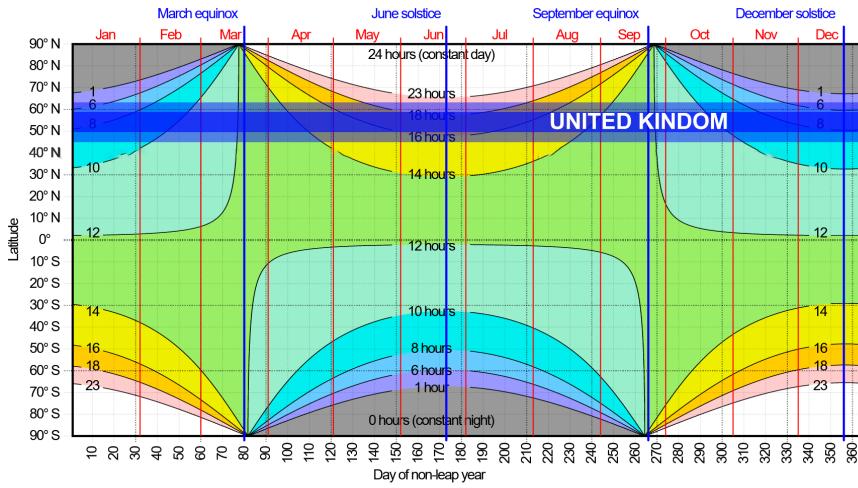


Figura 28: Lunghezza delle giornate nel regno unito in funzione del mese dell'anno

novembre, il tempo scende al di sotto delle 8 ore. Certamente questa non rappresenta una spiegazione esaustiva del grafico precedentemente illustrato; altri fattori come ad esempio un maggior utilizzo delle strade nel periodo estivo per viaggi e vacanze è certamente un altro aspetto da tenere in considerazione.

Anche dalla seconda distribuzione temporale, quella in funzione dell'orario giornaliero, è possibile notare alcuni aspetti interessanti. La linea al di sopra delle barre rappresenta la gravità degli incidenti avvenuti nelle diverse fasce orarie, espressa come percentuale di incidenti fatali sul totale e calcolata mediante la formula DAX mostrata in figura 29

```

1 Percentage_of_fatal =
2 CALCULATE(
3     COUNTA('accident_etl_10'[Accident_Severity]),
4     'accident_etl_10'[Accident_Severity] IN { "Fatal" }
5 ) / COUNTA('accident_etl_10'[Accident_Severity]) *100
6

```

Figura 29: Misura DAX rappresentante la percentuale di incidenti fatali rispetto al totale

Con 124.612, il 8,97% del totale, la 17° ora registra il maggior numero di vittime che è stato 17,3 volte più alto della 4° ora, dove trova il suo valore minimo (7.182). Inoltre, nella 3° ora, si registra il più alto tasso di gravità degli incidenti, 3.15 volte superiore alla media. *Le vittime e l'Indice di gravità totale sono correlati negativamente tra loro.*

Da queste informazioni è possibile dedurre alcuni aspetti. Innanzitutto, la correlazione negativa tra la frequenza e la fatalità degli incidenti non indica necessariamente la presenza di una *regressione* altrettanto rilevante, bensì una conseguenza di altri fattori che portano, per motivi potenzialmente differenti, ad una rilevante diminuzione del numero di incidenti nelle ore notturne, ma un sostanziale aumento della loro gravità.

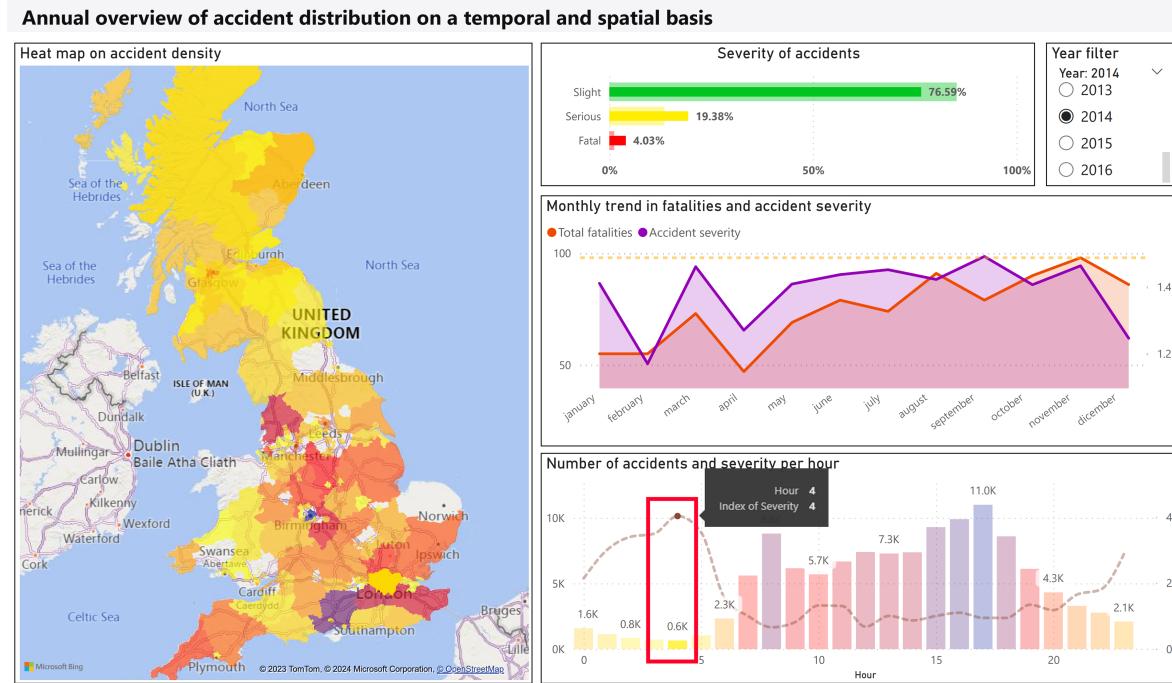


Figura 30: Filtraggio per incidenti in orario notturno

Le cause sono da ricercare nel minor volume di veicoli in circolazione in quella fascia oraria nel primo caso; e ad una minore visibilità, specialmente in condizioni di scarsa illuminazione stradale, e attenzione da parte dei guidatori nel secondo caso. Infine è possibile affermare che la scarsità di veicoli

circolanti può indurre i guidatori a raggiungere velocità elevate, e quindi aumentare la fatalità in caso di incidenti, come mostrato in figura 30.

Una soluzione a queste problematiche potrebbe consistere nel miglioramento dell'illuminazione stradale e nella promozione di iniziative per sensibilizzare sulla moderazione della velocità nei periodi notturni.

4.3.2 Dashboard 2 - Confronto della pericolosità stradale e cause di incidenti tra due distretti

Questa seconda analisi punta a mostrare le relazioni esistenti tra i limiti di velocità, l'età media dei veicoli e le condizioni avverse, e gli incidenti; eseguendo un confronto tra due distretti selezionati. Come è possibile osservare in figura 31, la dashboard è divisa orizzontalmente in due sezioni, ciascuna relativa ad un distretto. Per ogni distretto vengono mostrati diversi dati ottenuti come misure riepilogative.

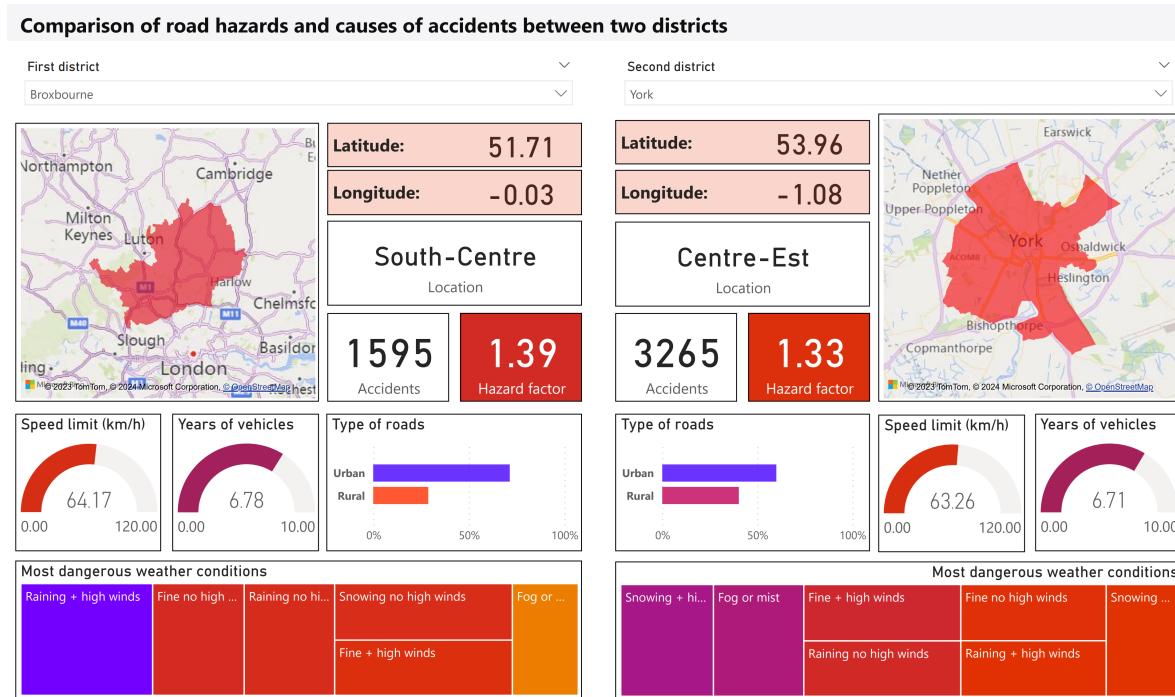


Figura 31: Confronto della pericolosità stradale e cause di incidenti tra due distretti

Prima di tutto, nella sezione più alta, trova posto un riepilogo della posizione del distretto nel regno unito. Come è possibile osservare dalla figura 32, per classificare la posizione di un distretto, è stata scritta una formula DAX che si occupa di ricavare i valori di latitudine e longitudine come valori mediani (e non medi) degli incidenti avvenuti in quella zona; e ritorna poi una delle nove classi a seconda della loro appartenenza a determinati range di valori.

Al di sotto troviamo un riepilogo del numero di incidenti e del fattore di pericolosità medio degli stessi; calcolato come media aritmetica del numero di morti. Nella fascia centrale invece, si mostrano due lancette riepilogative dei limiti di velocità medi e età media dei veicoli; nonché una comparazione del tipo di strade appartenenti al distretto in esame. Si noti come per analizzare i limiti di velocità, espressi in *mph* nel dataset, si sia ricorso ad una colonna calcolata con la formula DAX mostrata in figura 33.

```

Position_in_UK =
    VAR LAT = MEDIAN('accident_etl_10'[Latitude])
    VAR LON = MEDIAN('accident_etl_10'[Longitude])
    RETURN IF(
        LAT > 55.5,
        IF(
            LON > -3,
            "North-Est",
            IF(LON > -4.75, "North-Centre", "North-West")
        ),
        IF(
            LAT > 52.6,
            IF(
                LON > -1.17,
                "Centre-East",
                IF(LON > -2.35, "Centre", "Centre-West")
            ),
            IF(
                LON > 0,
                "South-East",
                IF(LON > -2, "South-Centre", "South-West")
            )
        )
    )
]

```

Figura 32: Formula DAX per la classificazione della posizione di un distretto in funzione delle sue coordinate

```

Speed_limit_kmh = accident_etl_10[Speed_limit] * 0.160934
Age_of_vehicle_2 = vehicle_etl_10[Age_of_Vehicle]/10

```

Figura 33: Colonna calcolata per convertire dei limiti di velocità e gli anni dei veicoli.

Chiude la dashboard una *Tree map*, rappresentante le principali condizioni metereologiche cause di incidenti.

Una volta compresa la struttura della dashboard, è possibile svolgerci due interessanti analisi diagnostiche ponendo a confronto distretti con valori diversi nelle due lancette. Se per esempio, come mostrato in figura 34, selezioniamo i due distretti *Hambleton* e *City of London*, notiamo subito la differenza sostanziale nel limite di velocità medio.

Ciò è certamente giustificato dalla conformazione del territorio, il primo composto dal 83.69% di strade extraurbane, mentre il secondo solo dall' 1,14%. A monte di tale osservazione, possiamo notare come il fattore di rischio sia 1,43 volte superiore e la frequenza degli incidenti 1,18 volte inferiore nel distretto di *Hambleton*. In particolare neve, pioggia e vento forte sono le condizioni meteo più pericolose, con un fattore di rischio che raggiunge valori di 2,18 Per quanto riguarda invece l'età media dei veicoli nelle due regioni, notiamo uno scarto di 13 mesi a favore di *City of London*. Infine il limite di velocità medio nel distretto di *Hambleton* è 1,88 volte superiore di quello del distretto prevalentemente urbano di *City of London*.

In sostanza, possiamo dedurre che limiti di velocità elevati e scarsità di centri urbani e quindi alta presenza di strade di scorrimento, siano benefici nella riduzione del numero di incidenti, ma siano anche fonte di incidenti più gravi. D'altro canto, un maggiore invecchiamento delle autovetture in circolazione, dovuto certamente ad un contesto economico più svantaggiato, comporta una minore sicurezza alla guida e quindi ad un aumento del fattore di rischio. Una soluzione a quest'ultimo problema potrebbe risiedere in una maggiore emissione di incentivi statali o regionali basati sul reddito per incentivare l'acquisto di nuove autovetture.

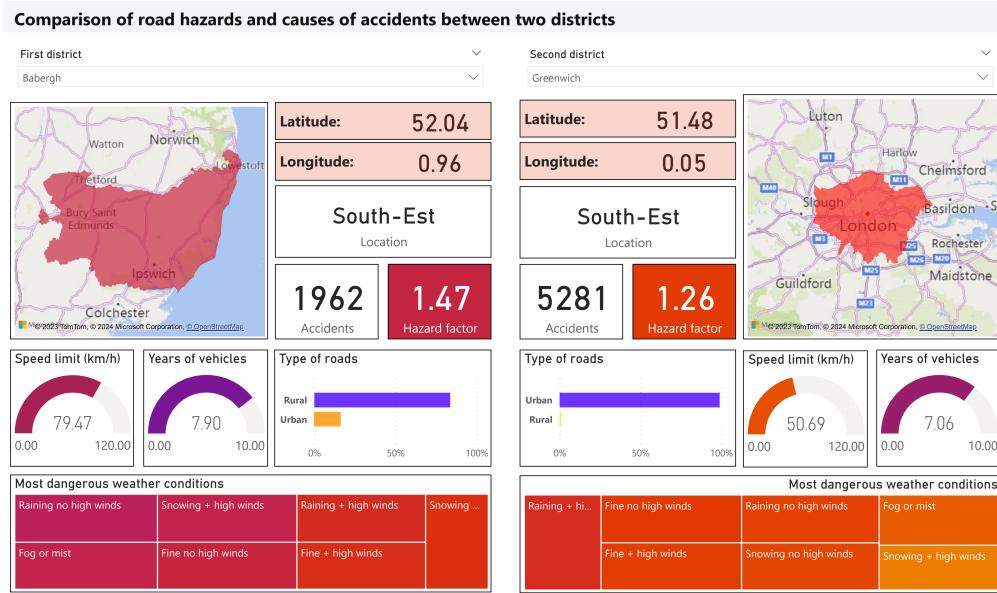


Figura 34: Diagnosi dell'impatto dei limiti di velocità e sull'età dei veicoli sugli incidenti di due distretti

4.3.3 Dashboard 3 - Analisi della pericolosità degli incidenti in relazione ai tipi di veicoli coinvolti

La terza dashboard realizzata permette di confrontare la pericolosità di incidenti avvenuti in base alla tipologia del veicolo coinvolto, ed è mostrata in figura 35.

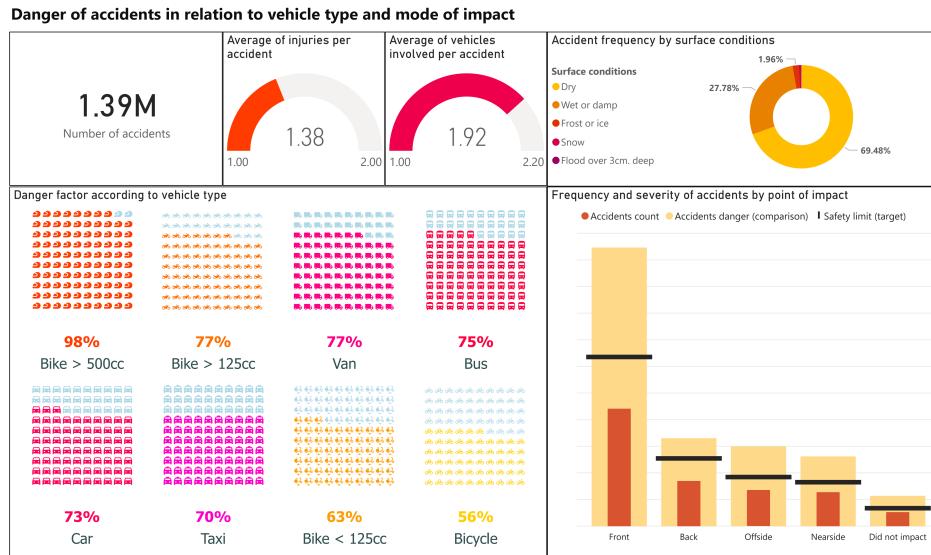


Figura 35: Riepilogo della pericolosità degli incidenti in relazione ai tipi di veicoli coinvolti

A tale fine la dashboard è stata divisa in 4 zone:

- Nella zona in basso a sinistra viene mostrato un *waffle chart* rappresentante la gravità degli incidenti avvenuti in relazione al tipo del veicolo coinvolto. Tale valore viene espresso come percentuale rispetto ad un valore limite. A ciascun veicolo è stata associata un'icona (in formato

vettoriale *SVG*) rappresentandone con chiarezza la tipologia. Per fare ciò è stato sufficiente creare la nuova tabella mostrata in figura 36; dove sono stati associati, ad ogni tipo di veicolo, un nome riepilogativo e un *path SVG* che ne rappresenta l'icona. Quest'ultimi sono stati estratti dalle icone del *material* di google. Successivamente, importata la tabella, è stata creata una associazione uno-molti con la tabella dei veicoli sotto la colonna **Vehicle_type**.

1	Vehicle_Type	Name	SVG_Path
2	Pedal cycle	Bicycle	M195.435-158.565q-81
3	Car	Car	M200.478-198.022v48.!
4	Tram	Tram	M154.022-262.63V-654
5	Motorcycle over 125cc and up to 500cc	Bike > 125cc	M158.398-198.565q-65
6	Motorcycle over 500cc	Bike > 500cc	M569.435-215.456H186
7	Taxi/Private hire car	Taxi	M240-200v40q0 17-11.
8	Bus or coach (17 or more pass seats)	Bus	M249-110.195q-15.391
9	Van / Goods 3.5 tonnes mgw or under	Van	M225.552-155.5q-49.59
10	Motorcycle 125cc and under	Bike < 125cc	M272.8-279.565q-48.47

Figura 36: Dataset delle icone per le diverse tipologie di veicoli

- Nella zona in alto a sinistra vengono riportati tre misure: il conteggio degli incidenti, la media di feriti coinvolti e la media dei veicoli coinvolti. Questi tre valori risultano particolarmente utili quando analizzati in presenza di un filtro su una tipologia veicolo, come si vedrà nel seguito.
- Nella zona in alto a destra viene mostrata la classificazione degli incidenti in base alle condizioni dell'asfalto.
- Infine, nella zona in basso a destra viene riportato un *bullet chart* che classifica la frequenza degli incidenti in base al tipo di manovra e alla loro pericolosità. Tale misura di pericolosità viene calcolata con la funzione DAX mostrata in figura 37, che assegna un diverso peso agli incidentati a seconda del danno fisico sofferto.

```

1 Casualties_weighted = SWITCH(
2     accident_etl_10[Accident_Severity],
3     "Slight", 1,
4     "Serious", 4,
5     "Fatal", 10,
6     0
7 ) * accident_etl_10[Number_of_Casualties]

```

Figura 37: Misura DAX per assegnare pesi diversi agli incidentati a seconda della gravità dell'incidente

Prendiamo ora in esame la figura 38 che filtra gli incidenti prendendo solo quelli avvenuti con biciclette.

Con 38.300 incidenti totali, di cui il 51,12% frontali, *la bicicletta* si giudica essere il veicolo con minore probabilità di incidenti senza coinvolgimento di un altro autoveicolo. Gli incidenti frontali hanno una percentuale di gravità media o alta 2,84 volte maggiore rispetto agli altri. L' 81,45% degli incidenti avviene su condizioni di strada asciutta, mentre il 18,01% avviene su strada bagnata o umida. Infine, si registra una media di 1,04 feriti e di 1,97 veicoli coinvolti per incidente.

Alla luce di questi dati, possiamo trarre delle conclusioni interessanti. Innanzitutto gli incidenti in bicicletta risaltano, come già intuibile dalle due lancette in figura 38, per avere un basso valore medio di feriti (molto vicino ad 1) ed un elevata media di coinvolgimento veicoli (prossima a 2). Ciò è presto spiegato dal fatto che gli incidenti con questo tipo di mezzo, salvo i casi di incidenti solitari (molto

Danger of accidents in relation to vehicle type and mode of impact



Figura 38: Profilazione degli incidenti avvenuti in bicicletta

rari come ci suggerisce la seconda lancetta), destano pericolo esclusivo per il conducente della stessa e mai per l'autovettura coinvolta. Dall'analisi sulle condizioni del terreno, notiamo che una buona parte di incidenti avviene su strada bagnata. Sebbene inferiore al 20%, questa tipologia di incidenti risulta essere decisamente più probabile delle altre, vista la scarsa aderenza al terreno delle ruote e alla spesso pesante diminuzione delle prestazioni dei freni delle biciclette.

Concludiamo questa analisi notando che, come è possibile osservare dal *waffle chart* in figura 38, la bicicletta si colloca all'ultimo posto della classifica di pericolosità dei veicoli. Quando messa a paragone con le moto a cilindrata superiore a 125cc, anch'essi veicoli a due ruote, notiamo una diversa collocazione nella graduatoria. Sebbene tutte e tre mostrano delle lancette molto simili, ovvero non causano il decesso del guidatore dell'altro veicolo e tendono a fare incidenti con uno ed un solo veicolo, i primi due hanno una frequenza di incidenti maggiore (di 2,38 volte) e una percentuale di fatalità maggiore (di 1,99 volte).

Una soluzione per ridurre questi incidenti, sebbene non già molto frequenti, potrebbe senz'altro consistere in un aumento di zone riservate ad uso ciclabile o ad una incentivazione nella loro predilezione rispetto alle strade. Ciò consegue dal fatto che la quasi totalità degli incidenti avviene per coinvolgimento con un autoveicolo. Si no

Terminiamo ora questa analisi osservando, in figura 39, cosa accade nel caso in cui sia il furgone l'oggetto del filtro.

Con 106.770 incidenti, *il furgone* è il veicolo che causa gli incidenti più coinvolgenti e pericolosi per gli altri. Dimostra infatti una media di 1,42 feriti e di 2,18 veicoli coinvolti per incidente. Registra inoltre un frequenza 3,32 volte maggiore di incidenti gravi in caso di impatto frontale rispetto agli altri tipi. Questi dati dimostrano come i furgoni, data la loro massa elevata, presentino un pericolo elevato per gli altri veicoli. Sebbene 10 volte meno frequenti di quelli causati dalle automobili, mostrano un coinvolgimento medio superiore a 2 veicoli per incidente. Alla luce di ciò, si potrebbe utilizzare questa

Danger of accidents in relation to vehicle type and mode of impact



Figura 39: Profilazione degli incidenti in cui sono coinvolti i furgoni

informazione per indagare sull' introduzione del divieto di transito per veicoli di massa superiore alle 3,5 tonnellate su strade dove la frequenza di questi incidenti particolarmente pericolosi per gli agenti della strada rimane alta.

4.3.4 Dashboard 4 - Clustering dei tipi di veicoli coinvolti negli incidenti

Analizziamo ora l'ultima dashboard, mostrata in figura 40, che fa uso dell'algoritmo *k-means* di clustering basato sul partizionamento offerto da Power BI. A seguito della creazione dello *scatter plot*, sono stati provati diversi valori di *k* per l'algoritmo *k-means*. La scelta è ricaduta su 3 in quanto è sembrata generare un buon compromesso tra profilazione assente ed eccessiva. Al fine di ottenere un'analisi corretta, sono stati esclusi mediante filtri, i veicoli senza motore come le biciclette e quelli a traino animale.

Il clustering così effettuato ha identificato 3 categorie con un simile valore di frequenza degli incidenti:

- Veicoli nuovi e semi-nuovi, di età inferiore a 8 anni e di bassa e media cilindrata. Fanno parte di questa categoria moto di ogni cilindrata, una buona parte di automobili e una piccola parte di veicoli pesanti.
- Veicoli vecchi, di età superiore a 8 anni e di bassa e media cilindrata. Fanno parte, in proporzioni leggermente maggiori, i veicoli della classe precedente.
- Veicoli nuovi, semi-nuovi e vecchi di cilindrata elevata. Fanno parte di questa categoria mezzi pesanti, quali furgoni, tram, minibus e bus, e una parte delle automobili.

Profiling of vehicles involved in accidents

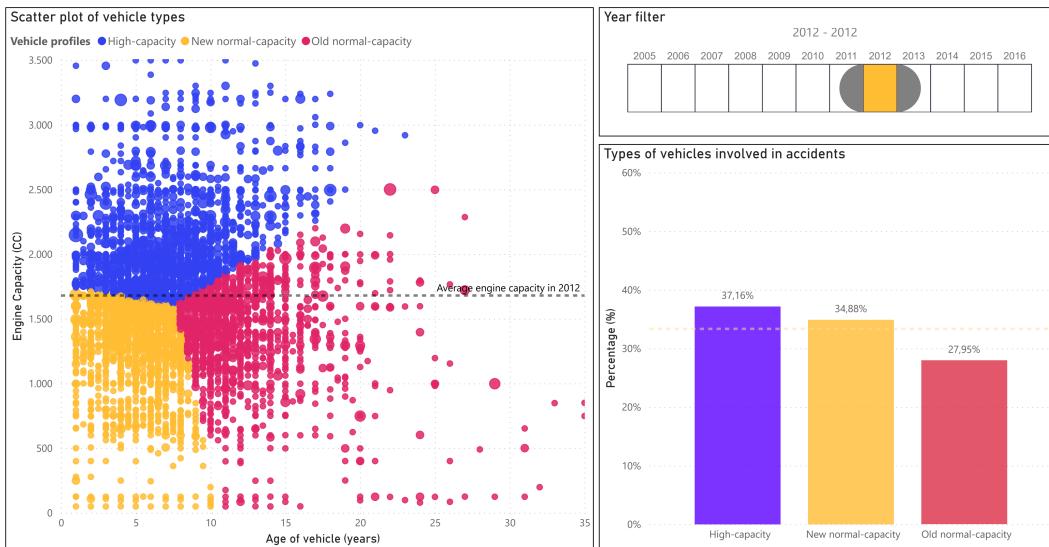


Figura 40: Clustering dei tipi di veicoli coinvolti negli incidenti del 2012

Nell'anno 2012, come illustrato in 40, la maggior parte (37,16%) di incidenti è stato causato da veicoli con motori ad *alta cilindrata*, mentre il 55,50% della restante parte (quella con veicoli a *media-bassa cilindrata*) è stata causata da veicoli con età inferiore agli 8 anni

Nell'anno 2016 invece, come illustrato in 41, la maggior parte (32,52%) di incidenti è stato causato da veicoli vecchi con motori a *bassa cilindrata*, mentre il 33,11% da veicoli *semi-nuovi ad alta cilindrata*).

Profiling of vehicles involved in accidents

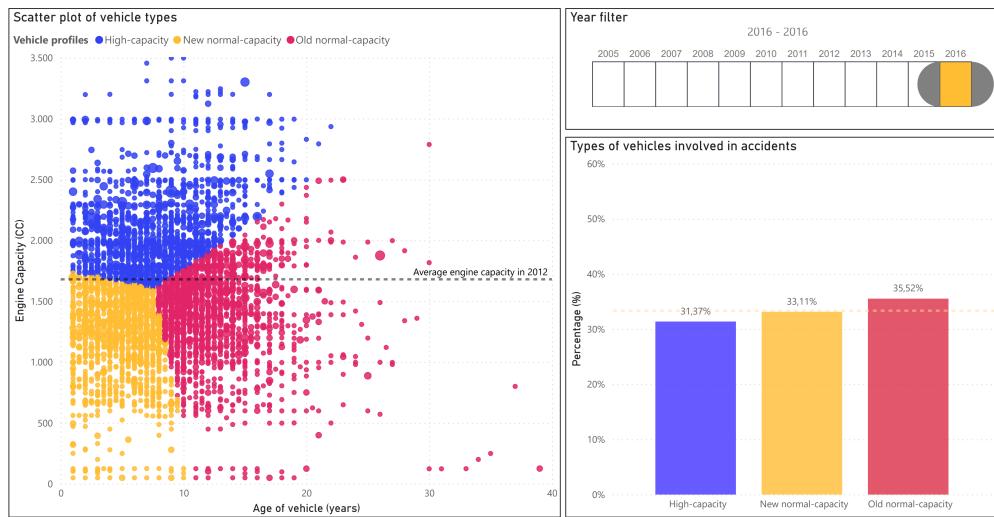


Figura 41: Clustering dei tipi di veicoli coinvolti negli incidenti del 2016

I dati presentati mostrano un aumento significativo del numero di incidenti avvenuti con mezzi a vecchi a bassa e media potenza, mentre sembrano rimanere pressoché costanti quelli avvenuti con veicoli nuovi e semi-nuovi. In effetti, provando a graficare la frequenza di incidenti in funzione del tempo delle tre diverse categorie, otteniamo il grafico mostrato in figura 42.

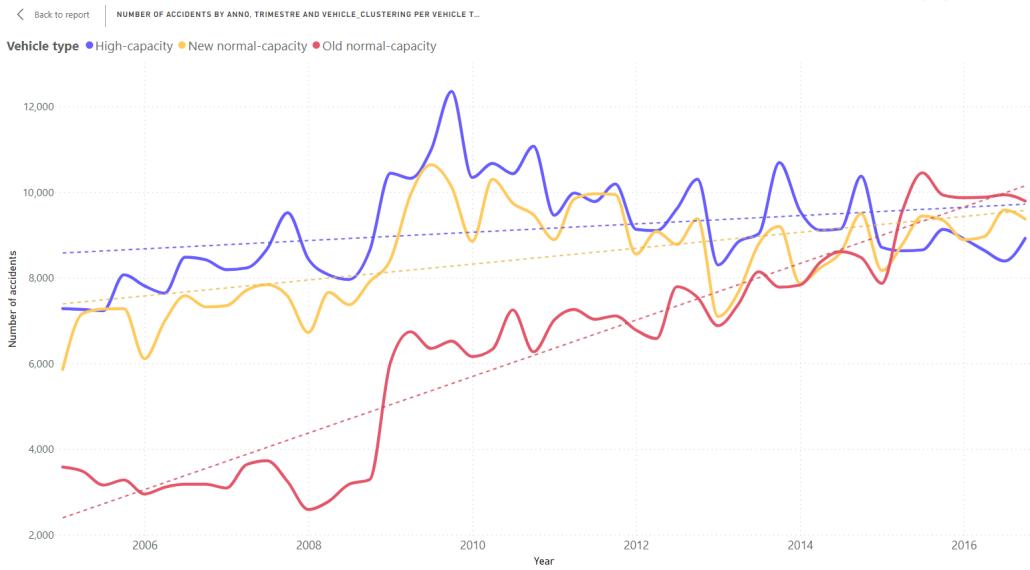


Figura 42: Andamento nel tempo della frequenza di incidenti in base al tipo di veicolo

Il grafico mostra chiaramente un aumento di incidenti avvenuti con mezzi vecchi a bassa e media potenza: rispetto al 2006, si registra un aumento del 335% degli stessi nel 2016. Risultano invece pressoché stabili le altre due tipologie di veicoli, con forse una leggera diminuzione degli incidenti avvenuti in veicoli ad alta cilindrata. Questo potrebbe essere dovuto in buona parte ad un avanzamento della tecnologia di sicurezza applicata in primis in veicoli pesanti destinati allo spostamento pubblico. L'avanzare del tempo, e quindi l'aumento dell'età dei veicoli, non sembra essere in perfetto equilibrio con l'acquisto di nuovi. Come già discusso in precedenza, un inevitabile decadimento delle prestazioni del veicolo comporta inevitabilmente un aumento della probabilità di incidenti.

4.4 Visualizzazioni installate dal marketplace

Nella realizzazione delle precedenti dashboard sono stati utilizzati le seguenti visualizzazioni gratuite e semi-gratuite installate tramite il marketplace di Power BI:

4.4.1 Waffle Chart (utilizzato nella dashboard 3)

Un *Waffle Chart* è un insieme di riquadri divisi per 10×10 celle; queste vengono riempite proporzionalmente al valore della categoria alla quale il riquadro corrisponde. Come mostra la figura 43, questa *visual* risulta molto espressiva nella classificazione di una serie di categorie mediante una misura.

4.4.2 Bullet Chart(utilizzato nella dashboard 3)

Questo comodo grafico a barre, mostrato in figura 44, consente di mostrare i valori di due serie di dati a confronto. Già nella sua versione gratuita il grafico si dimostra altamente personalizzabile nella scelta della modalità di rappresentazione della serie.

4.4.3 Timeline Slicer (utilizzato nella dashboard 4)

Questo filtro sviluppato dalla Microsoft stessa, permette di selezionare periodi temporali in modo estremamente flessibile in base alle esigenze. Come è possibile notare in figura 45, a differenza del più

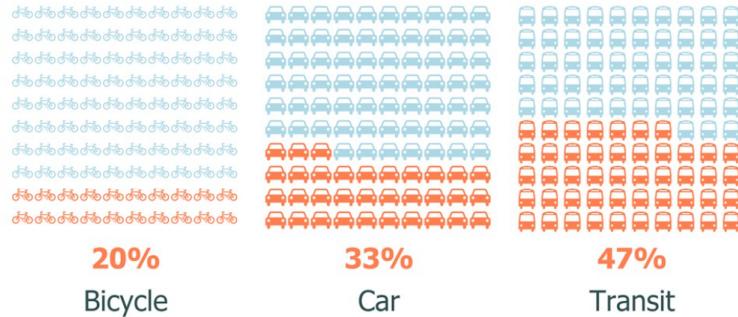


Figura 43: Waffle chart

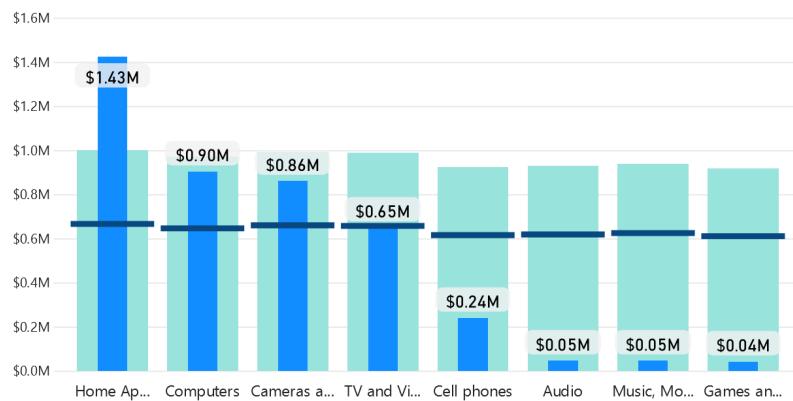


Figura 44: Bullet chart

basico filtro già offerto da Power BI, il *Timeline Slicer* ci permette di modificare la granularità del periodo temporale in esame.

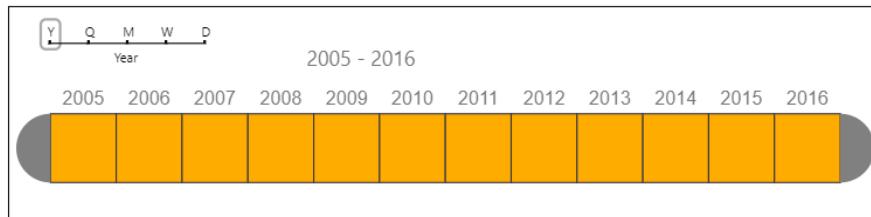


Figura 45: Timeline slicer

Siti web consultati

1. OCSI UK: Indices of Deprivation, LSOA. (2024) <https://ocsi.uk/indices-of-deprivation>
2. NCDR Reference Library: Indices of multiple deprivation (IMD) decile. (2024) https://data-england.nhs.uk/ncdr/data_element/indices-of-multiple-deprivation-imd-decile
3. DAX function reference. (2023, Ottobre 20). In Microsoft <https://learn.microsoft.com/en-us/dax/dax-function-reference>
4. Daytime. (2023, Settembre 8). In *Wikipedia*. <https://en.wikipedia.org/wiki/Daytime>
5. 2010 to 2015 government policy: road safety. (2015, maggio 8). In *Gov.uk*. <https://www.gov.uk/government/publications/2010-to-2015-government-policy-road-safety/2010-to-2015-government-policy-road-safety>