Exploring the Social and Environmental Factors Contributing to Single-Vehicle Accidents in the UK

1. Introduction

Vehicle accidents are a big global public health challenge, claiming approximately 1.19 million lives annually and causing 20 to 50 million injuries (World Health Organization, 2023). They rank among the leading causes of death worldwide particularly for young people aged 5 to 29 years. (World Health Organization, 2023) and account for 13% of all external causes of death in the UK (RAC Foundation Briefing Note for the UN Decade of Action for Road Safety, 2017). Beyond the human impact, the economic costs are large, estimated to account for 3% of most countries' GDPs (World Health Organization, 2023).

The causes of vehicle accidents involve a mix of demographic, environmental, and situational factors. (GOV.UK, 2024). Efforts to address road safety require a comprehensive understanding of these contributing factors. Studies suggest that targeted interventions, such as enforcing speed limits, improving road design, and implementing educational campaigns, could significantly reduce accident rates and save lives (ITF, 2020).

2. Objectives

The main objective of this project is to preprocess data and perform exploratory data analysis techniques to discover meaningful insights into single-vehicle accidents across the UK. By analysing a rich dataset pertaining to accidents, vehicles, and drivers, this project aims to:

1. Investigate Relationships:

- Explore correlations between different factors
- Examine how vehicle and driver attributes contribute to accident causation.

2. Discover Patterns and Trends:

- Examine temporal patterns in road traffic accidents, including seasonal and hourly trends, to identify high-risk periods.
- Analyse spatial distributions to highlight accident-prone regions and their characteristics, such as urban versus rural disparities and road types.

3. Highlight Risk Factors:

- Identify specific conditions (e.g., light conditions, road types, and junction details) that are most associated with severe accidents.
- Contrast urban and rural accident statistics to understand the impact of environmental and infrastructural differences.

3. Methodology

3.1 Data Source

The dataset used in this project was taken from Kaggle: (UK Road Safety: Traffic Accidents and Vehicles) and originates from the UK government's collection of traffic accident records. This dataset is very detailed as it has around 2 million records ranging from 2005 to 2017 and encompasses 58 attributes. Its combination of descriptive, numeric, geographical and temporal attributes allows for the exploration of various factors contributing to accidents, such as social influences, situational conditions, and vehicle characteristics. By using this dataset, valuable insights can be gained looking into accident causation and identify potential areas for intervention to improve road safety.

3.2 <u>Data Preprocessing</u>

Data preprocessing is an important step needed for ensuring the dataset was ready for exploratory data analysis (EDA) and gaining insights. Below is a summary of the key steps, the methods used, and the rationale behind each process. Visualisations are referenced to illustrate the impact of preprocessing.

<u>Removal of Irrelevant Columns</u>: Using the *pd.read_csv()* and *drop()* functions, non-essential columns such as "Did A Police Officer Attend" and "Driver Home Area" were removed from the datasets. These attributes were deemed less relevant to analysing accident causation. This step reduced the dataset's complexity, focusing on core attributes that contribute directly to understanding traffic accidents.

<u>Filtering by Year</u>: The dataset initially spanned 2005 to 2017 but was filtered using the *isin()* method to retain only the years 2015 and 2016, which were the latest available detailed records for accidents and vehicles. This ensures the analysis reflects recent trends and actionable insights for current road safety strategies.

<u>Eliminating Duplicate and Multi-Vehicle Records:</u> Duplicate rows in the Vehicle_Information.csv were identified and removed using the *drop_duplicates()* function to ensure data integrity. And rows with duplicate Accident_Index values were filtered out to focus on single-vehicle accidents, using the *duplicated()* function with *keep=False*. This allows for an in-depth analysis of factors surrounding individual vehicles.

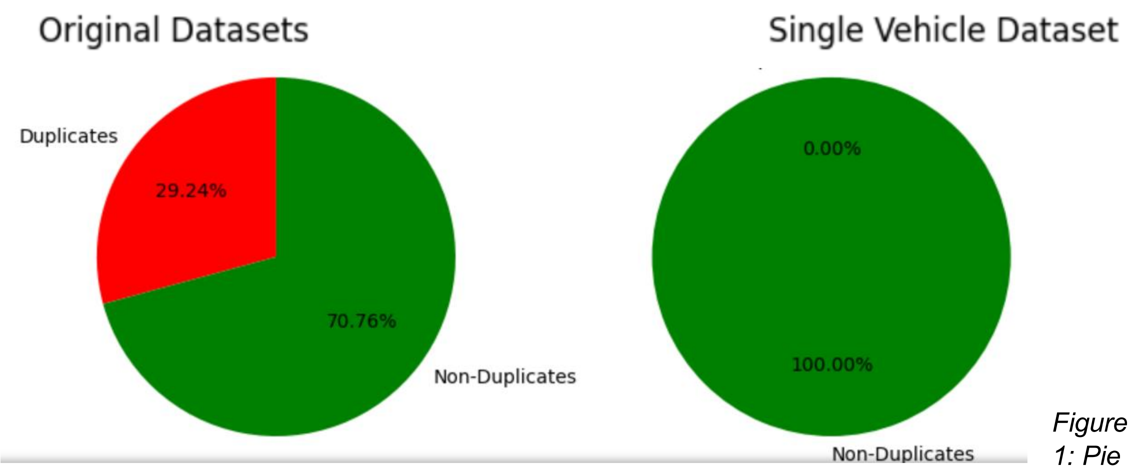

Figure 1: Pie charts comparing the percentage of duplicate rows in the original datasets and the pre-processed dataset.

Data Integration: The datasets were merged using the *pd.merge()* function on the Accident_Index attribute. This integration combined accident, vehicle, and driver data into a single dataset. This step streamlined data retrieval and analysis, ensuring all relevant information was accessible in one file.

Handling Missing Values: Instead of removing rows with missing values, placeholder strings such as “Data missing or out of range” or “Unknown” are used to preserve potentially valuable information in other attributes. The unique terms used for missing values were identified for conditional filtering during specific analyses, allowing flexibility in handling incomplete data.

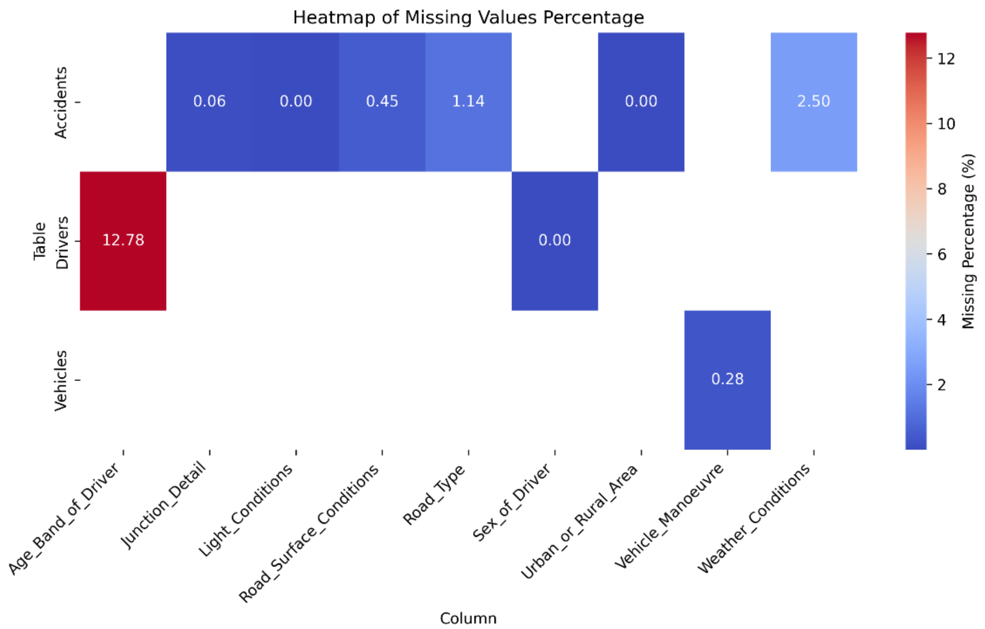

Figure 2: Heatmap illustrates the percentage of missing values for each attribute.

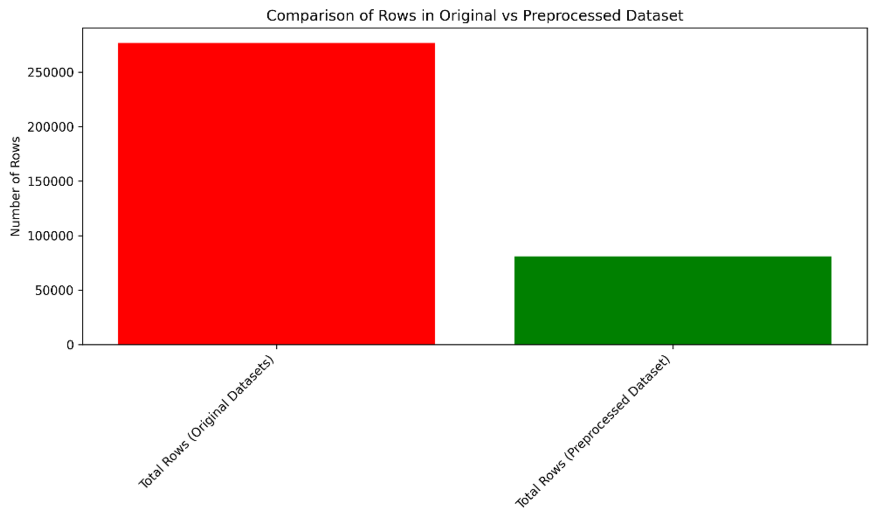

Figure 3: Bar chart comparing the total number of rows in the original and pre-processed datasets

3.3 <u>Database Creation and Implementation</u>

The database schema was designed to be able to store information regarding single vehicle accidents in the UK, noting details based on three entities: accidents, vehicles, and drivers while ensuring it is scalable and maintains data integrity. The schema consists of three tables: **Accidents, Vehicles and Drivers.**

The **Accidents table** is the foundational reference table for the database, as it stores important details about each accident. The Primary Key (PK) is *Accident_Index* which is a unique identifier for each accident. As a primary key, it is both *unique* and *not null,* ensuring there are no duplicate or missing values. This table includes the attributes with details such as the day of the week, date, time, year, location (longitude and latitude), environmental conditions (light, weather, and road surface), road details (type and speed limit), and the severity of the accident.

The **Vehicles table** stores detailed information about vehicles involved in accidents. The Primary Key is *Vehicle_ID:* An auto-incremented unique identifier for each vehicle. This ensures that every vehicle has a distinct identity. The foreign key is *Accident_Index which r*eferences the *Accident_Index* in the Accidents table. This establishes a one-to-many relationship, allowing multiple vehicles to be linked to a single accident if I was to scale my analysis using this database schema. The attributes include vehicle-specific details include the make, age, engine capacity, and the manoeuvre at the time of the accident. The *Vehicle_ID* is referenced as a foreign key in the Drivers table, establishing a one-to-one relationship between vehicles and drivers.

The **Drivers table** stores information about the drivers involved in accidents. The Primary Key (PK) is *Driver_ID*: An auto-incremented unique identifier for each driver, ensuring that every record in this table is distinct. The foreign key is *Vehicle_ID* which references the *Vehicle_ID* in the Vehicles table. This enforces a one-to-one relationship, where each driver is associated with exactly one vehicle as this is what I am going to focus on in my analysis. The attributes are Driver-specific details like, the age band and sex of the driver.

The design rationale is as follows:

<u>Maintain Data Integrity:</u>

- All primary keys (Accident_Index, Vehicle_ID, Driver_ID) are unique and not null to ensure that each record is uniquely identifiable and avoid invalid or duplicate data.
- Constraints on Year (2015 or 2016) and Speed_limit (>0) to ensure entries are consistent and valid.

<u>Data Types:</u>

- **TEXT:** Selected for descriptive and categorical data, such as accident details (Day_of_Week, Weather_Conditions).
- **DATE:** Used for Date to allow for time-series analysis.

- **INTEGER and REAL:** Chosen for numerical data (Year, Speed_limit, Longitude, Latitude) for correct and efficient storage and to do calculations.

Scalability:

- The one-to-many relationship between Accidents and Vehicles allows for the addition of multi-vehicle accident data in the future.
- The one-to-one relationship between Vehicles and Drivers ensures simplicity and consistency when analysing individual drivers.

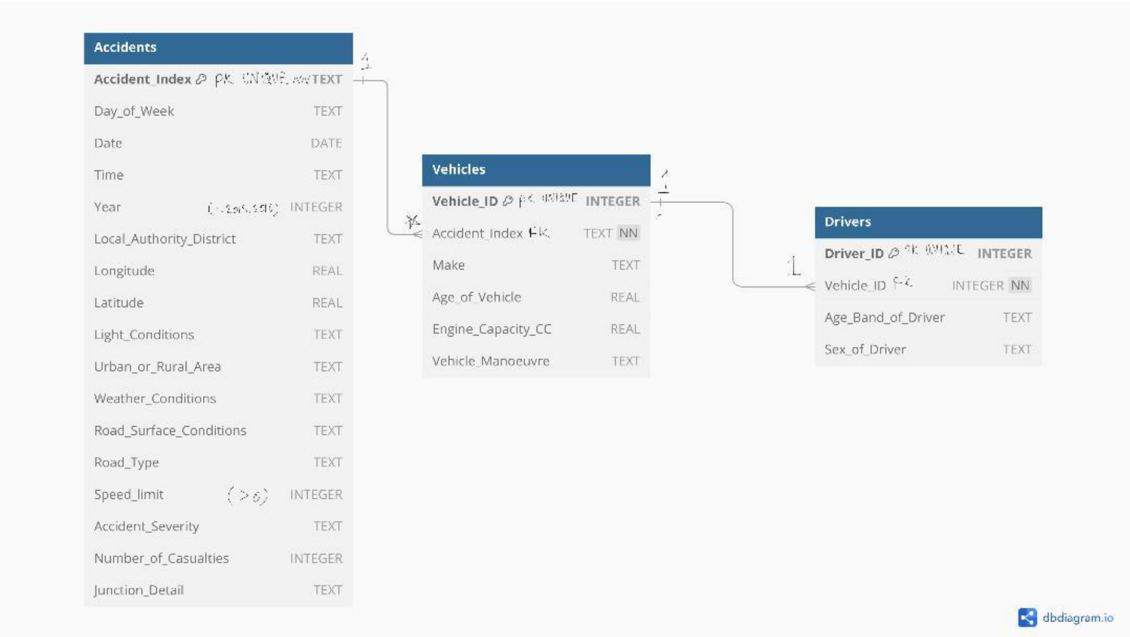

Figure 4: Relational Database Schema

Indexes were also created in this database to ensure query performance is fast and it allows for efficient data retrieval for key analytical attributes. By indexing attributes, the database can quickly locate relevant records without looking through entire tables.

The physical database was implemented using SQLite, and data was populated via Python’s sqlite3 library. This schema allows for good analysis of road accident data while maintaining flexibility and scalability for future extensions.

3.4 EDA

The EDA process for this project seeks to uncover meaningful insights, identify patterns, and establish relationships within the UK traffic accident dataset. By categorising the analysis into distinct types, the process ensures a structured approach to identifying factors contributing to accidents.

Category-based Analysis: Grouping data into categories allows for the comparison of attributes across distinct groups, helping to highlight disparities, patterns, or trends within different contexts. This approach is particularly useful for understanding categorical variables like accident severity or road types.

Relationship Analysis: This type of analysis seeks to reveal relationships between attributes, identifying how various factors like weather, road type, or driver demographics interact and contribute to accidents

Situational and Demographic Analysis: This analysis aims to investigate the role of situational and demographic factors in accidents, such as weather, light conditions, or driver age, to understand how these contribute to accident risk. These insights can support potential targeted interventions.

Temporal Trend Analysis: Analysing temporal trends provides insights into how accident-related metrics evolve over time. By studying these changes, you can uncover patterns related to peak accident periods.

4. Discussion of Key Findings

Temporal Trend Analysis:

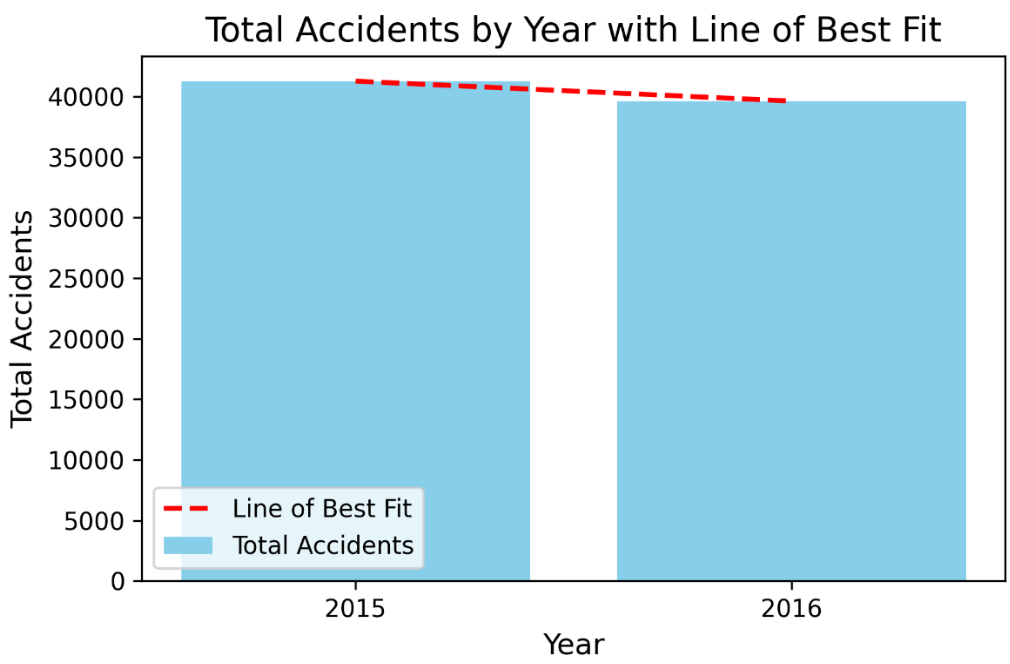

Figure 5: Bar Chart showing Total Accidents by Year with Line of Best Fit

The graph of total accidents per year, as visualised in Figure 5, shows a slight decline in the number of recorded accidents from 2015 to 2016. In 2015, there were 41,266 accidents, which decreased to 39,622 accidents in 2016. While the line of best fit highlights this downward trend, the drop is not substantial and appears stagnant. Further exploration into the underlying causes is necessary to identify areas for impactful intervention.

The graph of accidents by month, visualised in Figure 6, highlights fluctuations in the number of accidents throughout 2015 and 2016. It shows a noticeable increase in accidents during the winter months which can potentially mean that this coincides with bad weather conditions. This seasonal pattern expresses the influence of environmental and situational factors on accident rates.

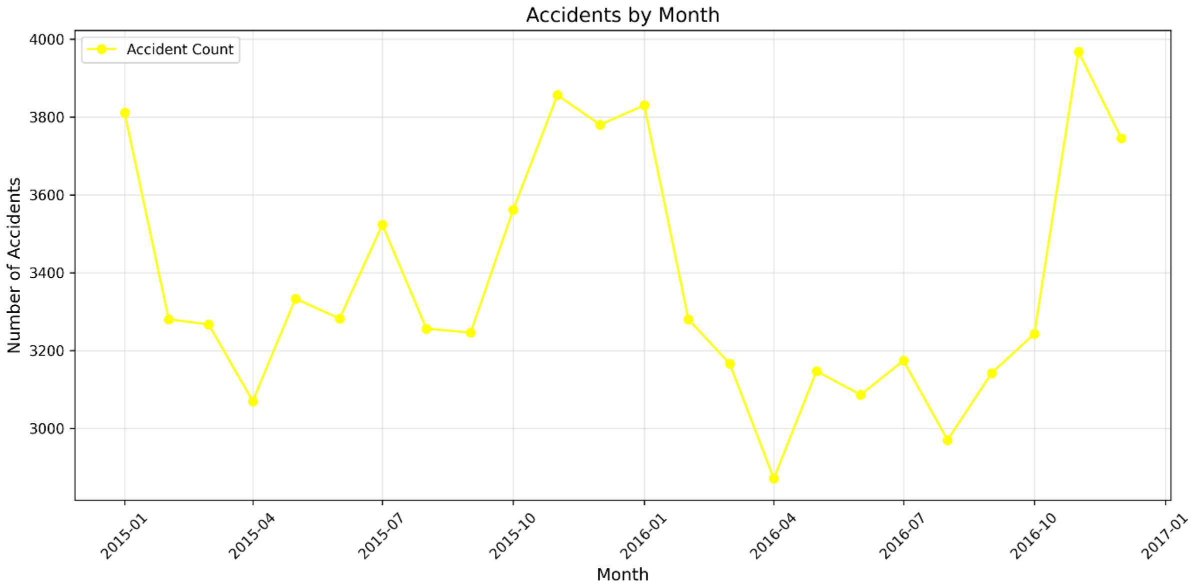

Figure 6: Line graph of Number of Accidents by Month

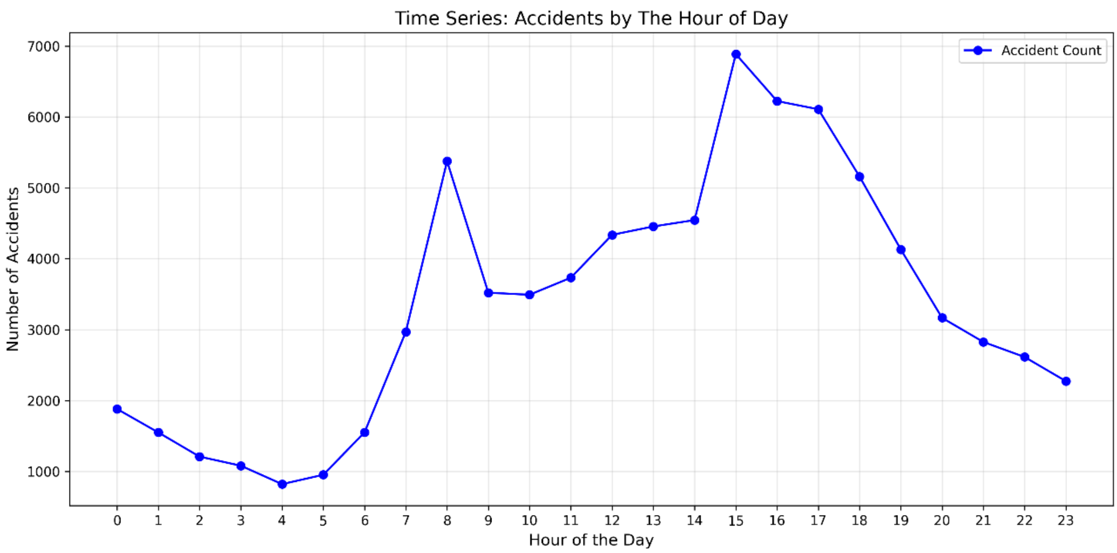

Figure 7: Line graph of The Number of Accidents by The Hour of Day

The analysis of accidents by hour of the day, show in Figure 7, reveals clear patterns of peak accident times. Accidents are most frequent during late afternoon and early evening hours which likely coincide with increased traffic during school arrival, school dismissal and rush hour. A smaller peak is also observed around 8 AM, aligning with morning rush hour. These findings highlight the temporal variability in accident occurrences and suggest the need for targeted traffic management during peak hours to improve road safety.

Relationship and Situational Analysis:

The scatter plot that shows accidents by geographical location, shown in Figure 8, highlights the distribution of accidents across the UK in urban and rural areas. There is a large difference in accident density between urban and rural areas. Urban areas account for 52,446 accidents, nearly double the 28,439 accidents recorded in rural areas, despite rural areas occupying a larger geographic space, urban areas are clear hotspots for accidents. The dense clustering of urban accident points aligns with high-population regions such as London and Birmingham, which emphasise the impact of traffic congestion and higher vehicle density in city environments. These findings suggest targeted road safety initiatives in urban zones to reduce accident risks.

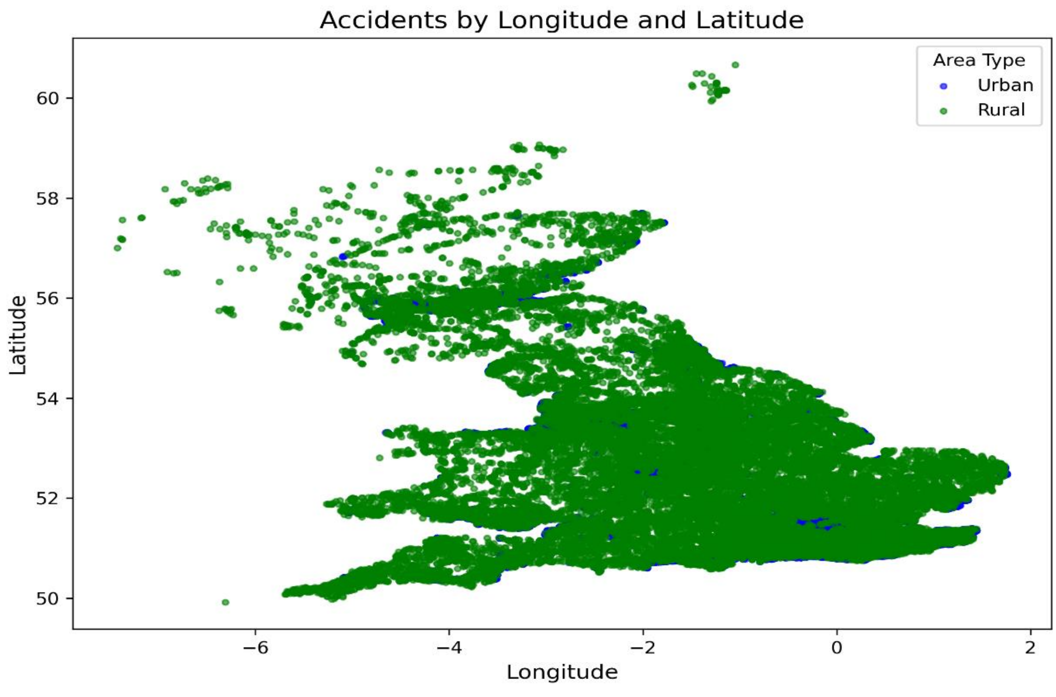

Figure 8: Scatter Plot showing Accidents by Longitude and Latitude

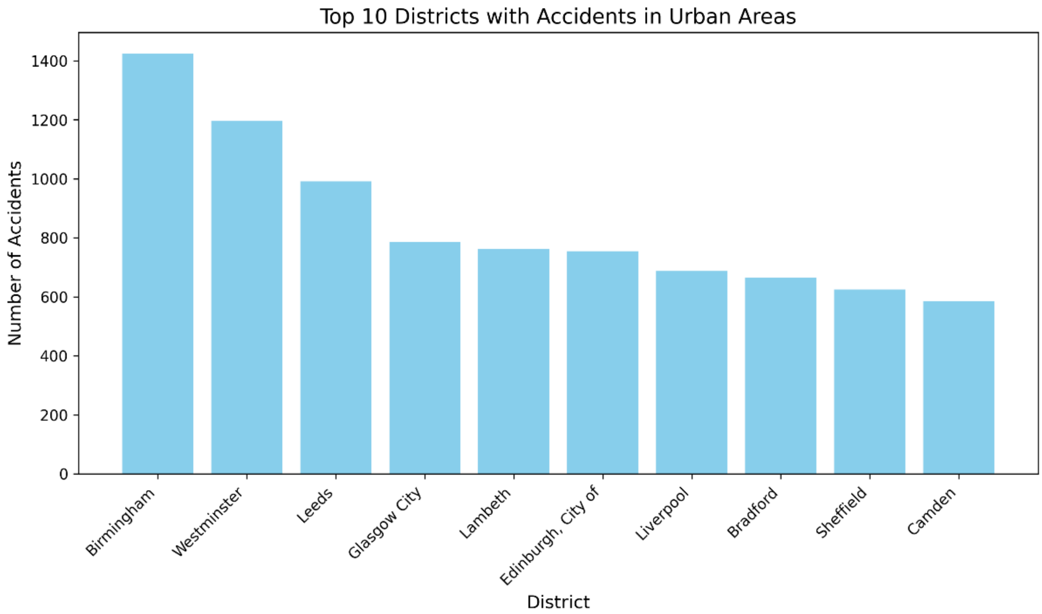

Figure 9: Bar Chart showing Top 1o Districts with Accidents in Urban Areas

Figure 10 explores the relationship between the age of vehicles, the total number of casualties, and accident counts. The line graph shows that newer vehicles are associated with the highest number of accidents and casualties. This may reflect their higher presence on the road. However, Older Vehicles show a steep decline in both accidents and casualties. This trend may be attributed to their limited numbers on the road or reduced usage. This finding suggests that vehicle age alone is not a definitive factor in accident causation. Instead, other factors, such as driver behaviour, road conditions, may have a greater influence. The visual comparison of casualties and accident counts highlights the need to analyse broader contextual factors.

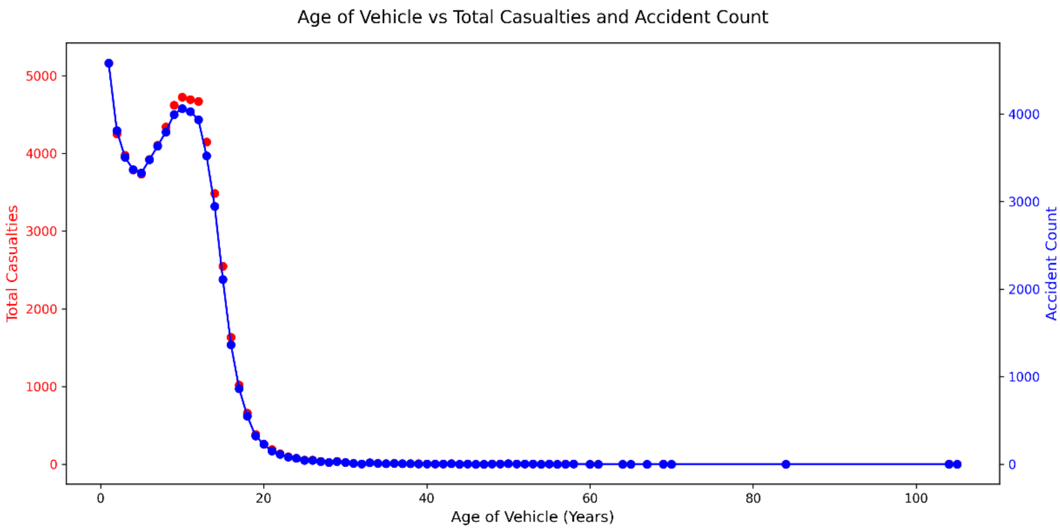

Figure 10: Line Graph showing Age of Vehicle vs Total Casualties and Accident Count

Categorical and Situational Analysis:

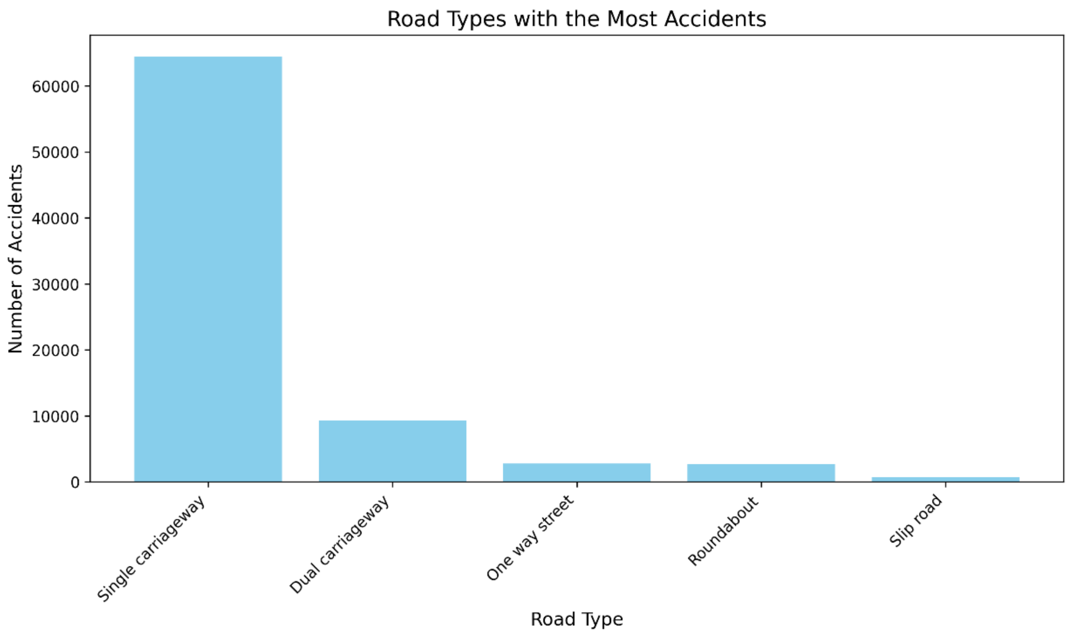

Figure 11: Bar Chart showing Road Types with the Most Accidents

Figure 11 illustrates the distribution of accidents across different road types. The Single Carriageway is the road type with the most accidents, accounting for an overwhelming 64,482 incidents, which significantly surpasses the second-highest road type, Dual Carriageway, with 9,307 incidents.

In terms of accident severity, Single Carriageways also dominate with the highest counts in the accident severity of 'fatal'. This stark difference underscores the need for targeted safety measures on single carriageways, likely due to their higher prevalence.

Figure 12 compares accident trends by Junction Detail, Light Conditions, Road Surface Conditions, and Speed Limit for single carriageways (the road type with the highest accidents) and all road types. Across all accidents and single carriageways, the 30-mph speed limit sees the highest accident count, followed by 60 mph roads. This suggests a need for enhanced safety measures at these speed limits. Most accidents occur on dry roads, indicating that weather conditions are not the sole contributor to accidents. However, wet or damp roads rank second, emphasising the need to consider weather-related risks. T or staggered junctions and areas not at junctions are common hotspots for accidents. This highlights the significance of driver behaviour and visibility at such points.

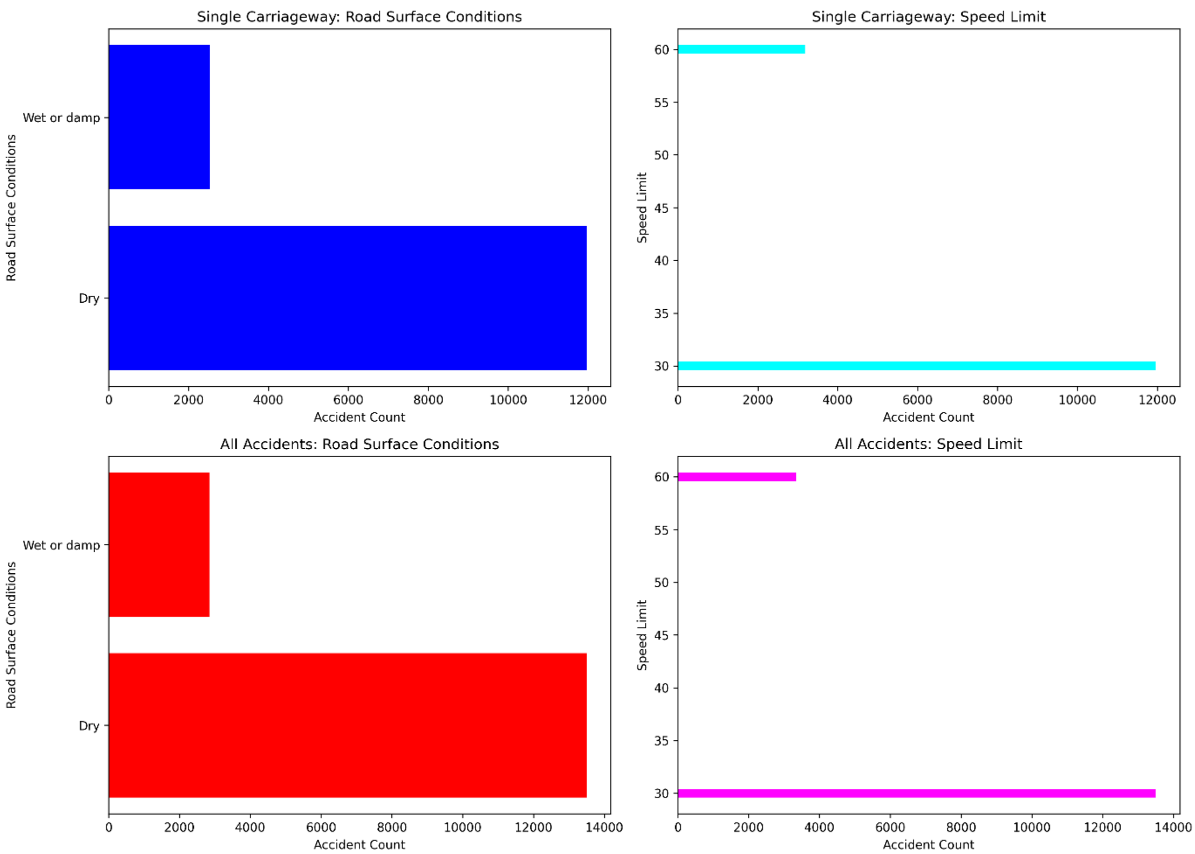

Figure 12.1: Comparison of accident trends by Road Surface Conditions, Speed Limit for single carriageways and all road types

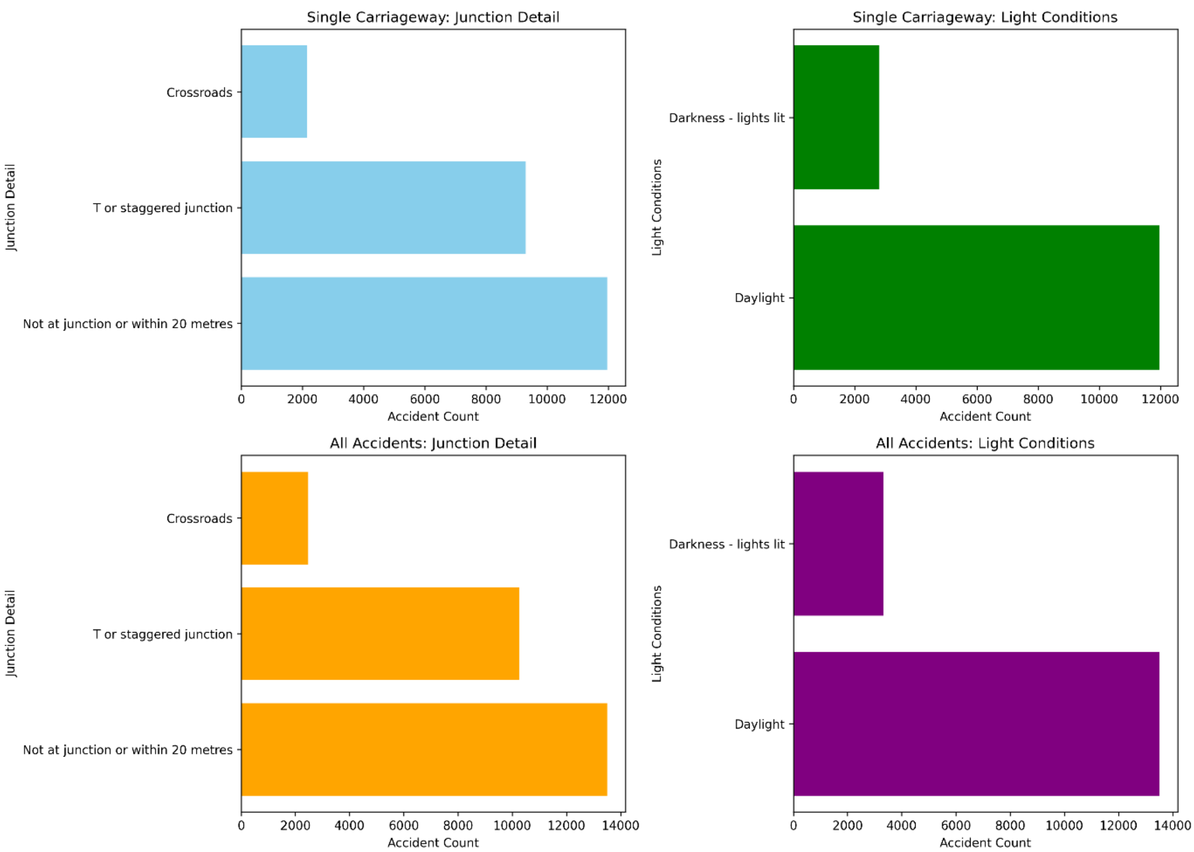

Figure 12.2: compares accident trends by Junction Detail, Light Conditions for single carriageways (the road type with the highest accidents) and all road types

Figure 13 highlights the relationship between the average speed limit and the average number of casualties across different road types. Slip Roads and Dual Carriageways have the highest average speed limits, reflecting their design for faster travel. However, these road types are also associated with a higher average number of casualties, suggesting the increased risk associated with high-speed travel. This relationship shows the direct relationship between higher speed limits and casualty rates, emphasising the critical role of speed management in reducing road traffic accidents and casualties.

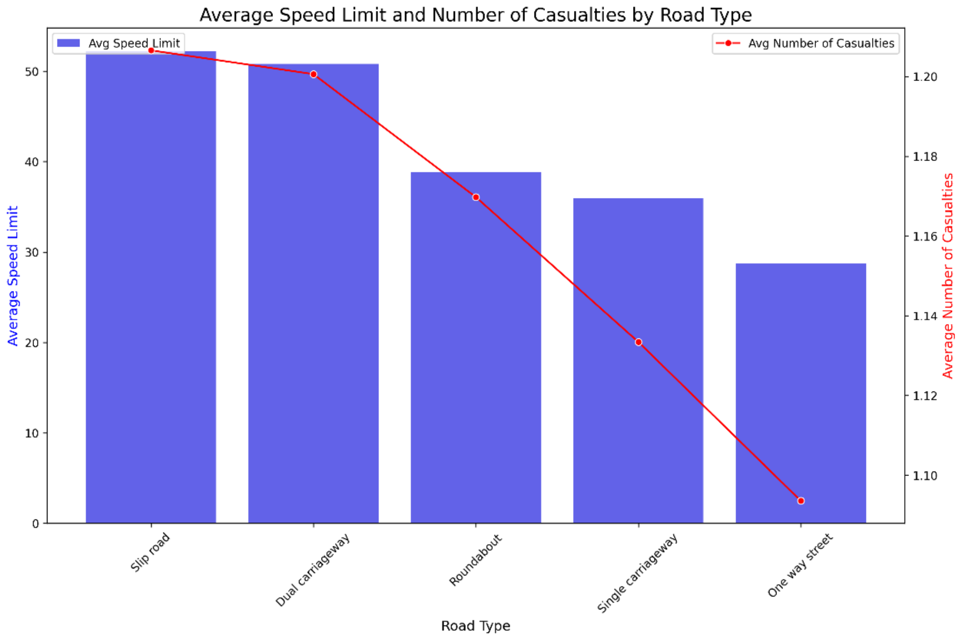

Figure 13: Shows Average Speed Limit and Number of Casualties by Road Type

Relationship and Demographic Analysis:

Figure 14 is a grouped bar cart showing the distribution of accidents by driver's sex and age band. Across all age bands, male drivers are involved in significantly more accidents than female drivers. This finding may highlight behavioural or situational differences between genders, which could inform targeted safety interventions. The 26–35 age group has the highest accident count, followed closely by the 36–45 age group. This trend suggests that drivers in these age ranges may face specific challenges, such as higher exposure due to frequent driving. This analysis shows the importance of tailored safety measures, particularly focusing on male drivers in the high-risk age groups, to mitigate accident rates.

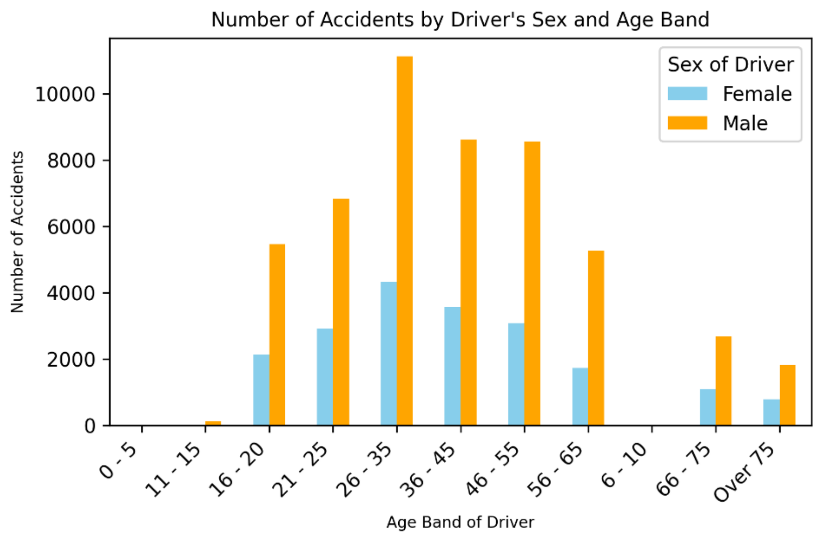

Figure 14: Grouped Bar Chart showing No. of Accidents by Driver's Sex and Age Band

5. Conclusion and Recommendations

The analysis of UK singular-vehicle accidents revealed important insights into accident trends and factors contributing to accidents. Urban areas emerged as hotspots for accidents, likely due to high traffic density. Single carriageways accounted for most accidents, with specific road types like dual carriageways and slip roads showing a correlation between higher speed limits and increased casualties. Additionally, demographic analysis identified younger age groups, particularly those aged 26-35, as more likely to be involved in accidents, and male drivers were disproportionately represented across all age bands. Environmental factors like wet road surfaces and poor lighting also played a notable role in accident severity.

To address these findings and improve road safety in the UK, the following recommendations are proposed:

1. Implement Targeted Educational Campaigns:

Focus on demographic groups at higher risk, such as young male drivers, to emphasise safe driving practices and the dangers of reckless driving.

2. Reduce Speed Limits:

Lower speed limits on high-risk road types like dual carriageways and slip roads, especially in areas with high accident and casualty rates.

3. Enhance Junction Safety:

Improve the design and visibility of junctions, particularly T or staggered junctions, where a significant proportion of accidents occur.

4. Focus on Urban Areas:

Prioritise road safety initiatives in urban areas with high accident densities, particularly in major cities.

5. **Promote Infrastructure Enhancements:**

Investments should go towards better lighting and improved road surface conditions, especially in areas prone to accidents during poor weather conditions or at night.

References

GOV.UK. (2024). *Guide to contributory factors for reported road casualties Great Britain*. [online] Available at: https://www.gov.uk/government/publications/guide-to-contributory-factors-for-reported-road-casualty-statistics/guide-to-contributory-factors-for-reported-road-casualties-great-britain.

ITF (2020). *Road Safety Annual Report 2020*. [online] ITF. Available at: https://www.itf-oecd.org/road-safety-annual-report-2020.

Mortality statistics and road traffic accidents in the UK An RAC Foundation Briefing Note for the UN Decade of Action for Road Safety. (2017). Available at: https://www.racfoundation.org/wp-content/uploads/2017/11/road-accident-casualty-comparisons-box-110511.pdf.

World Health Organization (2023). *Road traffic injuries*. [online] World Health Organization. Available at: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

Link to Kaggle dataset: UK Road Safety: Traffic Accidents and Vehicles

Link to Google Colab Notebook:
https://colab.research.google.com/drive/1thLJCays5e00O0JdGwtaDKlKuQs8sRaz?usp=sharing