**Data-Driven Insights into Road Traffic Accidents and Public Safety: A Dual Analysis of Accident Data and Social Networks in Great Britain**

1. **Introduction**

This report provides data-driven insights into public safety and social behavior gained from a dual analysis on two distinct datasets. Firstly, we analyze the 2020 road traffic accident data for Great Britain to identify patterns and trends that may provide information on improvements towards road safety. The main objective in analyzing the above dataset is to identify significant hours and days when accidents are more likely to occur, to analyze and recognize specific vehicle types and pedestrian involvement, to understand the regional distribution of accidents, and to forecast any future occurrences.

Lastly, we perform a social network analysis on a separate dataset from Facebook to understand the community structures and relationships that might impact on public safety behavior and awareness. Through this comprehensive approach, we aim at deriving actionable insights that government agencies and public safety organizations can use to improve and enhance safety and community engage

1. **Data Preprocessing and Cleaning**

Several preprocessing steps were undertaken before conducting any analysis on the accident dataset to ensure data quality and relevance.

These are the various preprocessing steps that were undertaken:

**2.1.1 Filtering Data:** Data was filtered to mainly focus on collected data on road traffic accidents in the year 2020.

**2.1.2 Handling Missing Values:** Missing values were addressed to ensure analyses were based on complete records. The accident dataset initially contained approximately 10% missing values, particularly in weather conditions and road surface columns. To address this, we dropped most of the records that had missing values for all the analysis we did.

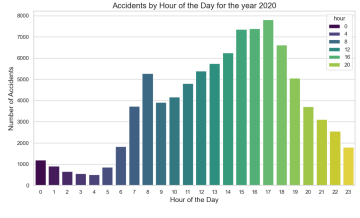
**2.1.3 Data Type Conversion:** We converted columns with dates and time information to proper datetime formats. For datetime columns, we converted the format to a standard YYYY-MM-DD HH:MM format using Python's pandas library to ensure accurate time-series analysis.

**Exploratory Data Analysis (EDA)**

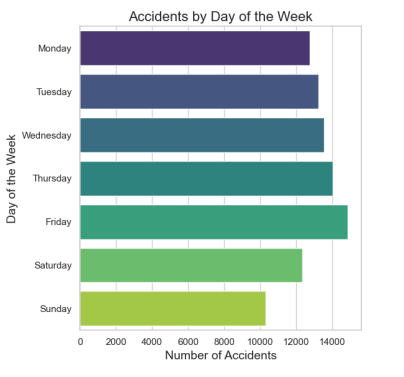
**3.1 Analysis on significant hours and Days of Accidents**

Our analysis revealed key insights into the days and hours at which accidents are most likely to happen. The figures below demonstrate and visualize the key insights into the overall accidents analysis.

**3.1.1 Overall Accidents Analysis**



***Figure 1: A Bar chart showing overall accidents by hour of the day for the year 2020.***

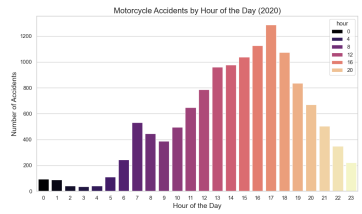


***Figure 2: A Bar chart showing overall accidents by day of the week for the year 2020.***

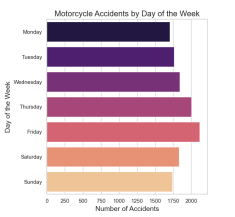
As illustrated in Figure 1, accidents are high during rush hours. Most accidents occurred during daytime hours, particularly between 8 AM and 6 PM, indicating a higher likelihood of occurrence during periods of high vehicle activity. This is consistent with the increased volume of vehicles on the road during commute times. For example, the peak at 5 PM coincides with the end of the workday when many individuals are traveling home. These findings suggest that implementing targeted safety measures, such as increased traffic enforcement or public awareness campaigns during these high-risk hours, could be beneficial in reducing accident rates. From Figure 2, Accidents were more evenly distributed across weekdays, with a slight increase on Mondays and Fridays, likely due to routine travel patterns.

**3.1.2 Motorcycle and Pedestrian Accidents**

Accidents involving motorcycles were more frequent during high-traffic periods around 11am to 9pm, with 5pm having the highest number of accidents likely to occur, highlighting the need for targeted safety interventions. Accidents were also distributed evenly across weekdays, with a slight increase on Fridays. Accidents involving motorcycles with larger engine sizes (over 500cc) were more frequent during high-traffic periods, highlighting the need for targeted safety interventions. When examining motorcycle accidents, we found that motorcycles with engine sizes over 500cc were involved in more severe accidents, particularly during evening hours (5 PM - 9 PM). This could be due to higher speeds achievable by these larger bikes, as well as lower visibility during twilight hours.

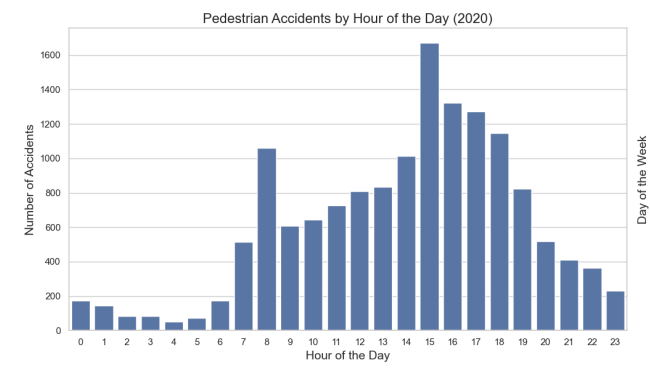


*FIGURE 3:* A Bar chart on *Motorcycle accidents by hour of the day (2020)*

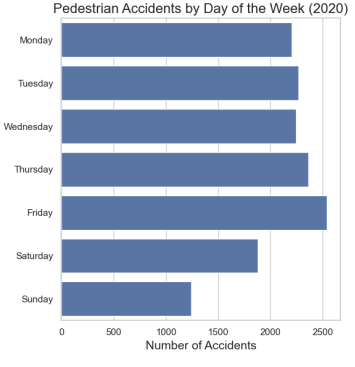


*FIGURE 4:* A Bar chart on *Motorcycle accidents by day of the week.*

Pedestrian accidents were more likely to occur during commuting hours, suggesting that safety measures like improved crossings and signals could reduce accidents. Pedestrian accidents predominantly occurred during commuting hours (8AM and 2 PM - 7 PM), when pedestrian activity is highest. For instance, in urban areas, crossings near schools and busy intersections showed a higher incidence of pedestrian accidents, highlighting the need for improved pedestrian infrastructure, such as traffic signals and raised crosswalks. The figures below visualize the insights gained from the analysis:



*Figure 5*: *A Bar chart showing pedestrians involved in accidents by hour of the day for the year 2020.*



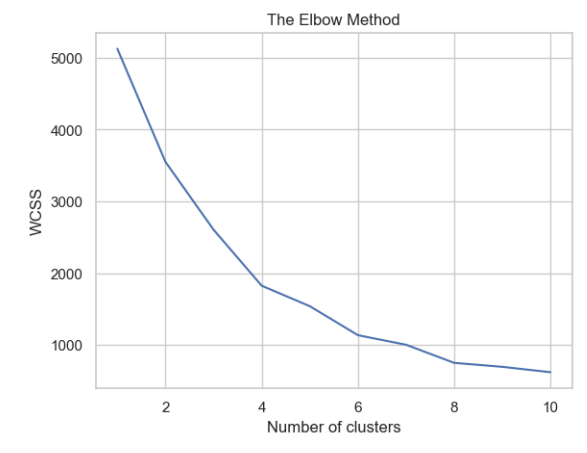
*Figure 6*: *A Bar chart showing pedestrians involved in accidents by day of the week for the year 2020.*

**3.2 Impact of Variables on Accident Severity**

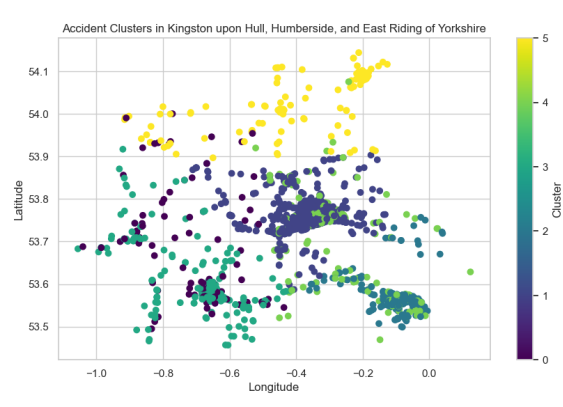
We used the apriori algorithm to explore associations between different variables and accident severity and to find frequent itemsets. This method was particularly useful in identifying combinations of variables, such as road type and speed limits, that contributed to severe accidents. In the report, the rule (light\_conditions) → (road\_type, urban\_or\_rural\_area, weather\_conditions) had a confidence of 100%, indicating that certain light conditions almost always occur with specific road types, urban/rural area classifications, and weather conditions. The rule (weather\_conditions) → (light\_conditions, accident\_severity) demonstrated a perfect association, indicating that specific weather conditions almost always lead to particular light conditions and accident severity levels. Since light conditions combined with urban or rural areas strongly predict accident severity, tailored safety guidelines for different areas under varying lighting conditions could be beneficial. The rule (road\_type, weather\_conditions) → (urban\_or\_rural\_area) shows how certain combinations of road types and weather conditions can reliably predict urban or rural area classifications, which can help in targeted interventions. Some rules, despite having high confidence, did not add much predictive value as their lift was close to 1, indicating that these associations might occur by chance. For instance, the rule (urban\_or\_rural\_area) → (road\_type, weather\_conditions) had a lift of 1.000, suggesting that while the association exists, it may not provide significant insight beyond what could be expected by chance. Rules with Zhang’s Metric of 1, such as (light\_conditions) → (road\_type, urban\_or\_rural\_area, weather\_conditions), were highly reliable, whereas those with a metric of 0 offered less significant insights.

**3.3 Clustering Analysis of Regional Accidents**

We performed clustering analysis using K-means to identify high-accident areas within Kingston upon Hull, Humberside, and the East Riding of Yorkshire. Clustering analysis revealed three high-accident clusters in Kingston upon Hull, particularly in areas with heavy traffic flow and multiple road intersections. One such cluster was identified near the A63, a major arterial road where several accidents occurred due to merging lanes and high-speed traffic. These clusters suggest the need for localized interventions, such as increased traffic monitoring or structural changes to problematic intersections. We derived that suggested targeted interventions could reduce accidents in these clusters.

****

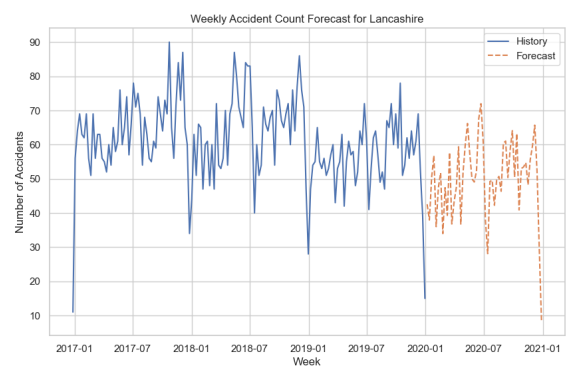
*Figure 7: A line plot demonstrating the elbow method for K-Means Clustering*

****

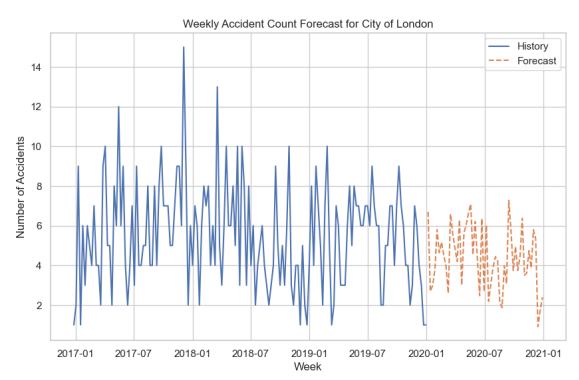
***Figure 7: A scattered plot demonstrating the accident clusters in Kingston upon Hull, Humberside and East Riding of Yorkshire.***

**3.4 Time Series Forecasting with SARIMA**

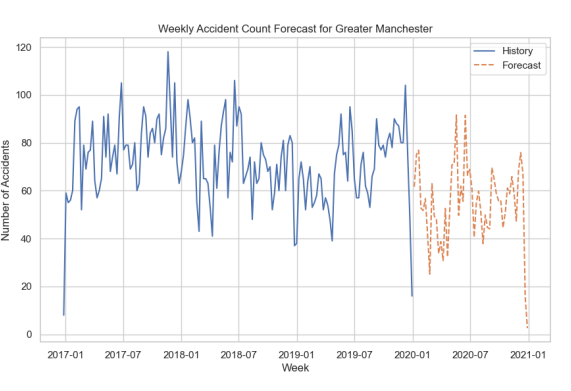
SARIMA models were employed to forecast future accident occurrences, capturing both trend and seasonal components to provide actionable insights for safety planning.



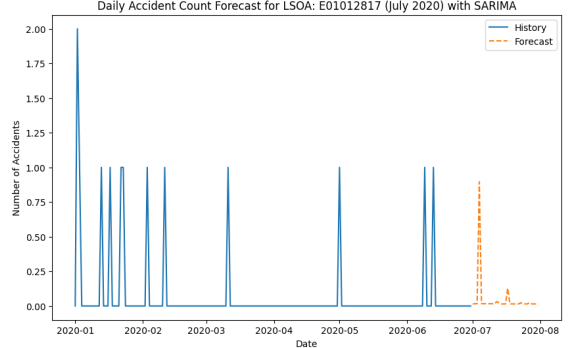
***FIGURE 9: Weekly Accident Count Forecast for Lancashire***

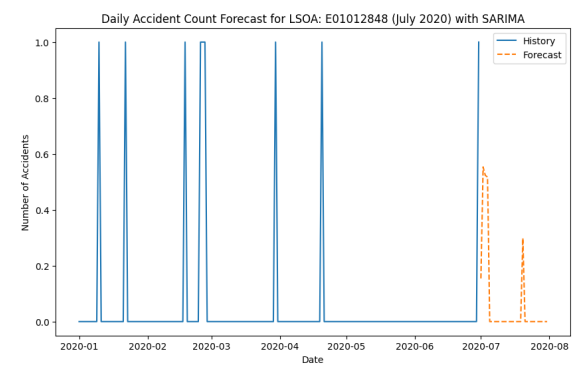
****

***FIGURE 10: Weekly Accident Count Forecast for London***

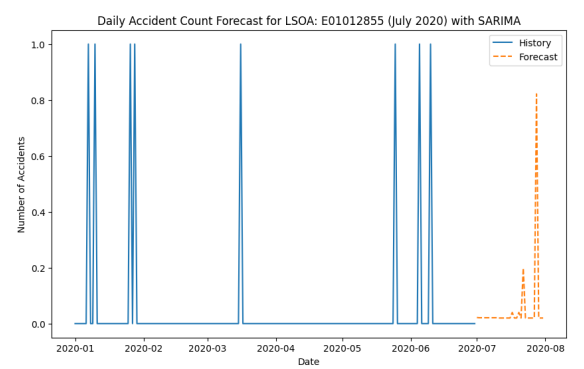
****

***FIGURE 11: Weekly Accident Count Forecast for Greater Manchester***

***FIGURE 12: Daily Accident Count Forecast for LSOA: E01012817 (July 2020) with SARIMA***

****

***FIGURE 13: Daily Accident Count Forecast for LSOA: E01012848 (July 2020) with SARIMA***

****

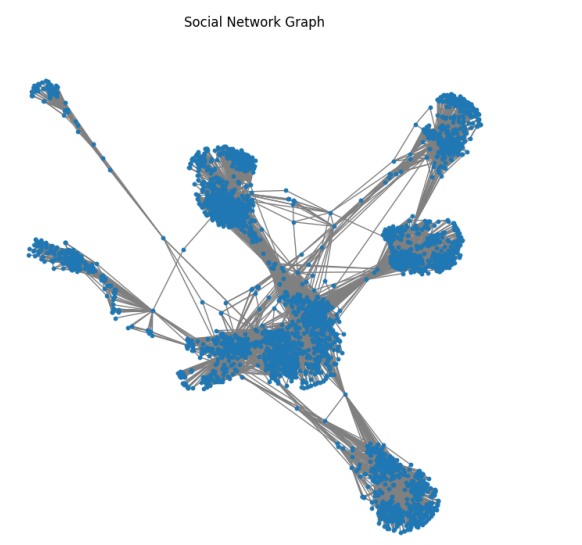
***FIGURE 13: Daily Accident Count Forecast for LSOA: E01012855 (July 2020) with SARIMA***

1. **Social Network Analysis**

We conducted a social network analysis, using a separate Facebook dataset, to understand community structures and relationships that could influence public safety awareness and behavior. This structure implies that any safety campaigns distributed via these networks would require strategic dissemination through key influencers.

**Network Construction and Basic Characteristics**

The social network was constructed with nodes representing Facebook users and edges representing friendships or connections. The network included a total of 4039 nodes and 88234 edges. Calculations showed a network density of 0.0108 and an average degree of 43.69, indicating a moderately connected network. The network density was calculated at 0.0108, indicating a moderately connected network. The average degree of 43.69 shows that most users had relatively few connections, suggesting a more localized and less influential network structure. This structure implies that any safety campaigns distributed via these networks would require strategic dissemination through key influencers."

****

*FIGURE 12: A Social Network Graph*

**Edge Centrality and Community Detection**

Analysis of edge centrality highlighted critical connections within the network, suggesting key influencers or nodes that play a significant role in maintaining network connectivity. This individual could be a valuable partner in promoting road safety initiatives through social media.

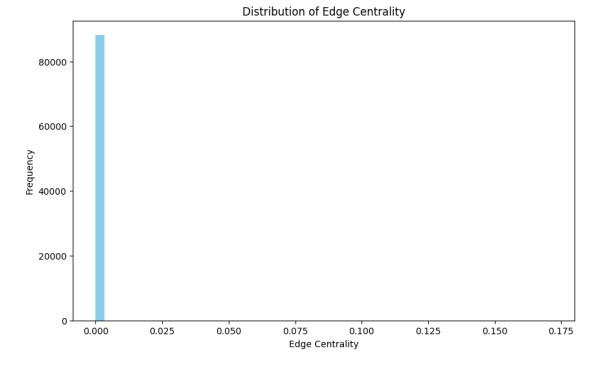
### **Edge Centrality Analysis**

**Max Edge Centrality**: **0.1715** – Represents the most influential connection in the network.

**Min Edge Centrality**: **1.23e-07** – Indicates the least influential connection.

**Mean Edge Centrality**: **4.18e-03** – The average level of influence across all connections.

High centrality edges are crucial for network connectivity and communication flow. Enhancing these connections can strengthen overall network stability. Low centrality edges have minimal impact and can be deprioritized in interventions.



*FIGURE 13: A bar chart demonstrating the distribution of Edge Centrality*

#### **Community Detection Analysis**

To gain insights into the community structures within the social network, we employed two distinct community detection algorithms: the **Louvain Method** and the **Label Propagation Method**. Both methods serve to uncover the underlying clusters or communities within the network, which are critical in understanding how information and behavior patterns, particularly those related to public safety, might propagate.

##### **Louvain Method**

The Louvain Method is a modularity-based community detection algorithm that optimizes the division of the network into communities by maximizing the density of edges within communities relative to edges between them. **Number of Communities Detected**: 16

**Community Sizes**:

The largest community contains **548 nodes**.

Other notable community sizes include **535, 435, and 423 nodes**.

The smallest community consists of **19 nodes**.

These results indicate that the Louvain Method identified several large, densely connected communities. The largest community, with 548 nodes, represents a significant subset of the network, suggesting that targeted interventions or communications within this group could reach a broad audience. The distribution of community sizes also suggests a hierarchy within the network, with a few dominant clusters and several smaller, but still significant, groups.

##### **Label Propagation Method**

The Label Propagation Method is a diffusion-based algorithm that spreads labels through the network until stable communities emerge.

**Number of Communities Detected**: 44

**Community Sizes**:

The largest community comprises **1,030 nodes**.

Other communities vary significantly in size, with some containing as few as **2 to 10 nodes**.

Notably, a community with **753 nodes** and another with **547 nodes** were also identified.

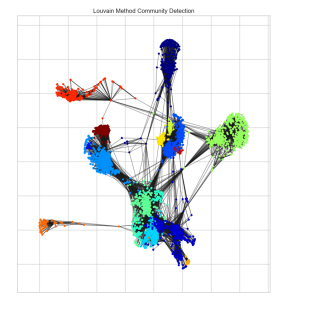
The Label Propagation Method revealed a much more fragmented network structure, identifying many small communities alongside a few large ones. This level of detail is particularly useful for pinpointing niche groups within the network that might require specialized communication strategies. The largest community, with 1,030 nodes, represents a broad segment of the network, but the presence of many small communities indicates potential areas where more personalized or localized interventions might be necessary.

##### **Comparison of Methods**

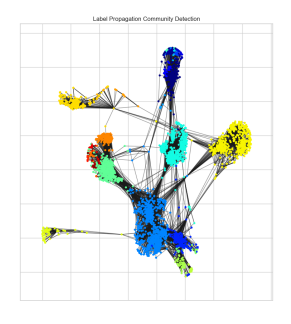
**Number of Communities**: The Louvain Method identified 16 communities, whereas the Label Propagation Method detected 44. This stark contrast highlights the different perspectives each method offers—the Louvain Method tends to group nodes into fewer, larger communities, while the Label Propagation Method finds a more fine-grained structure.

**Community Sizes**: The Louvain Method’s communities are generally larger and more cohesive, while the Label Propagation Method uncovers a more detailed and varied community structure with a mix of very large and very small groups.

These insights can be well understood from the following figures:



***Figure 13: A Louvain Method Community Detection***



*Figure 14: A Label Propagation Community Detection*

1. **Recommendations**

Based on the separate analyses of road traffic accidents and social network data, we propose the following recommendations:

For Road Safety:

Enhancing safety measures during peak hours is a major step to minimize road traffic accidents. There should be an implementation of targeted road safety measures during high-risk periods to reduce accident rates.

Another recommendation is targeted interventions in high-accident clusters. Government agencies should deploy specific safety measures in high-accident areas identified through clustering analysis.

Promotion of safety for motorcyclists and pedestrians are very key. There should be increased safety campaigns and infrastructure improvements for high-risk groups.

For Social Engagement and Awareness:

Government agencies should leverage social network insights by utilizing key influencers identified through social network analysis to disseminate safety messages and increase public awareness.

Community-Based Safety Initiatives is key. Government agencies should develop safety programs that leverage the community structures identified, fostering local engagement and support.

1. **Conclusion**

This report provides a comprehensive analysis of road traffic accidents in Great Britain for 2020, supplemented by a social network analysis to understand community behaviors. The findings highlight the importance of targeted safety measures, both in terms of direct interventions for road safety and leveraging social networks for public awareness and engagement. By combining these approaches, government agencies and public organizations can enhance road safety and foster a safer, more connected community environment.