INTRODUCTION

Road traffic accidents are significant concerns globally due to the loss of lives, injuries, and economic impact. Various studies have been conducted to analyse and detect the severity of road traffic accidents and to identify key factors influencing accident severity to improve safety.

Yao (2023) analysed factors influencing the severity of road traffic accidents using logistic regression models. By studying over 2,000 accident cases, they established a model to understand the variables impacting accident severity, providing valuable insights for accident prevention measures.

This study analyses road traffic accidents in the United Kingdom for the year 2020. The study utilizes data from the UK government accident database for 2020, which includes four tables: accident, vehicle, casualty, and LSOA tables.

DATA CLEANING:

The following data cleaning procedures were carried out:

1. Replacing specific values with NaN:

Specific values (-1, 99) are replaced with NaN across three Data Frames (accident, vehicle, casualty) to mark them as missing data. Additionally, the value 9 is replaced with NaN in specific columns.

2. User-defined function for handling missing values:

A function named preprocess data takes three Data Frames as input and handles missing values.

3. Imputing missing values with SimpleImputer:

Utilizes SimpleImputer from the scikit-learn library to impute missing values using median and mode strategies for numeric and categorical columns, respectively.

4. Custom logic for handling specific columns:

addressing missing values in the 'age_of_driver' column within the vehicle_df DataFrame.

5. Forward filling based on a grouping:

- Addresses missing values in location-related columns of the accident_df DataFrame using forward fill based on the 'police_force' column.

EXPLORATORY DATA ANALYSIS

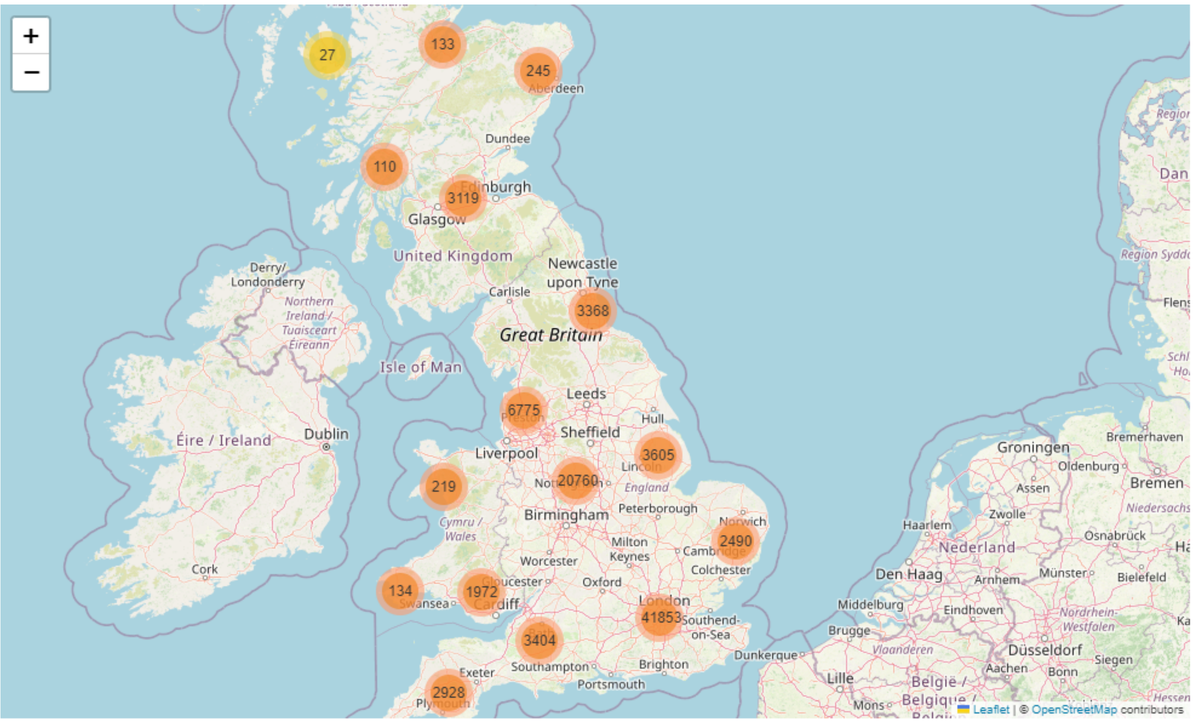


Figure 1:Accident hotspots in the United Kingdom

Figure 1 shows the main accident hotspots in the United Kingdom. The plot shows that London, Manchester Birmingham are places with high incidence of accidents.

For the age of the drivers involved in accidents, the predominant age age range is the 28-35as shown in figure 2

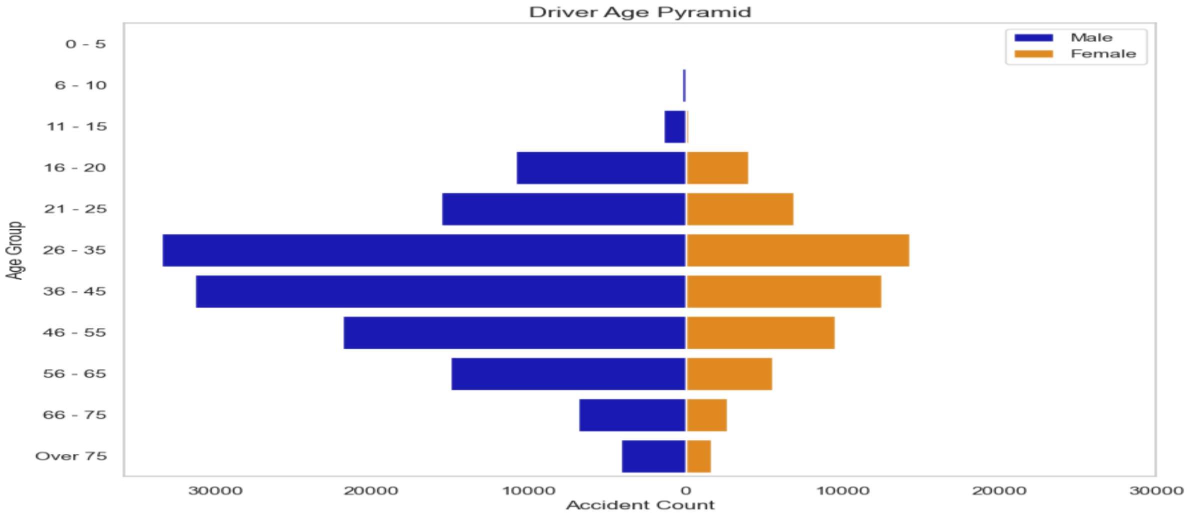


Figure 2:Age of Driver pyramid.

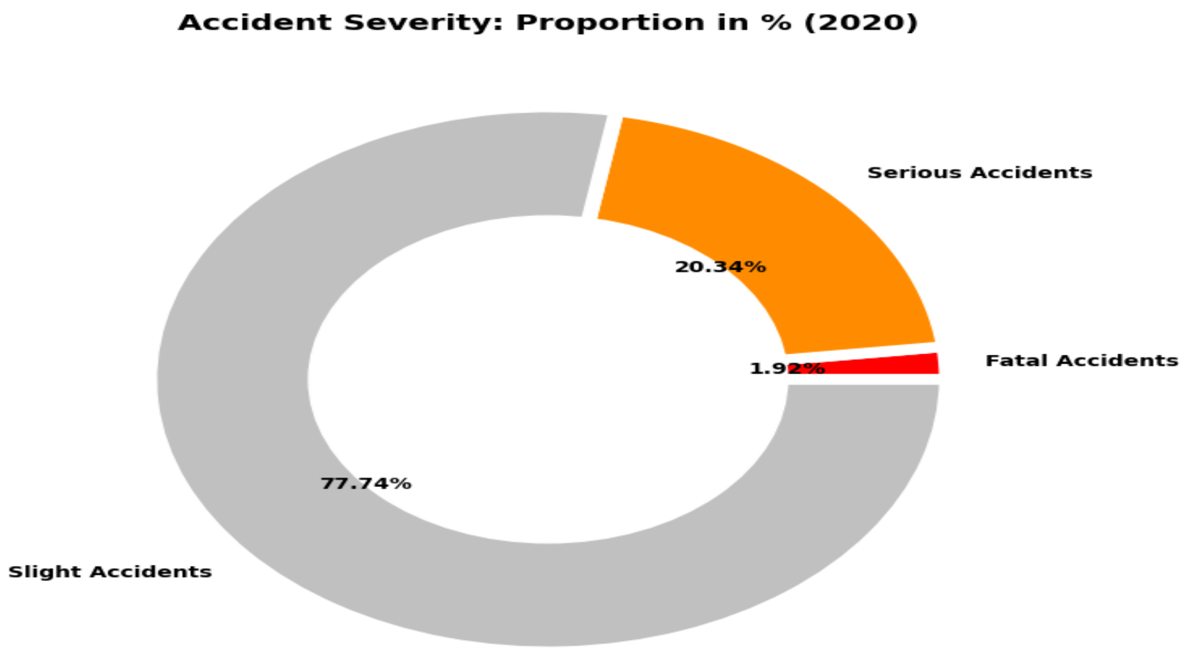


Figure 3: chart of accident severity.

The distribution of accident severity is highlighted in figure 3.

FINDINGS

Question 1.

Borucka et al., (2020) emphasised the impact of the time of the day on road traffic accidents, indicating that the hour of the day plays a crucial role in influencing the number of accidents.

The graph in Figure 4 illustrates the frequency of accidents in the United Kingdom at different times of the day for the year 2020. The highest frequency of accidents occurred between 3 pm and 6 pm, with 5 pm having the highest number of accidents, accounting for 8.6% of the total accidents. The lowest number of accidents occurred at 4 am. The number of accidents increased every hour after 9 am, reaching its peak at 5 in the evening. This is consistent with results from Mishra et al. (2010), which found that a significant proportion of accidents occurred between 3 to 7 p.m., and Meyyappan et al. (2018) reported that around 51.2% of road traffic accidents took place during busy traffic hours, particularly between 7 a.m. to 10 a.m. and 5 p.m. to 9 p.m.

The highest number of accidents occur on Fridays (14,889 or 16.3%), which is in line with studies by Pradhan et al. (2023), Widjajanti (2021), and Machetele and Yessoufou (2021), possibly due to end-of-week fatigue, increased social activities, and higher traffic volumes. Tuesdays to Thursdays also show high accident rates, requiring continuous vigilance throughout the workweek. Despite assumptions, Sunday has the lowest accident rate, and Saturday's rate is lower compared to weekdays.

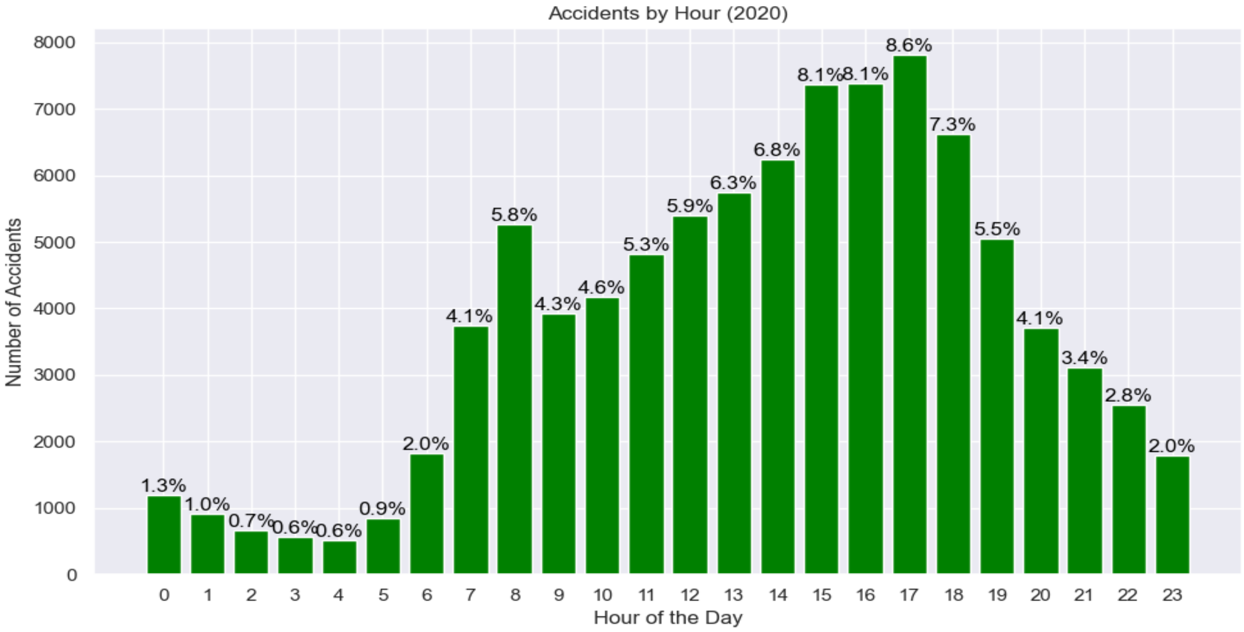


Figure 4:Number of accidents by hour of the day.

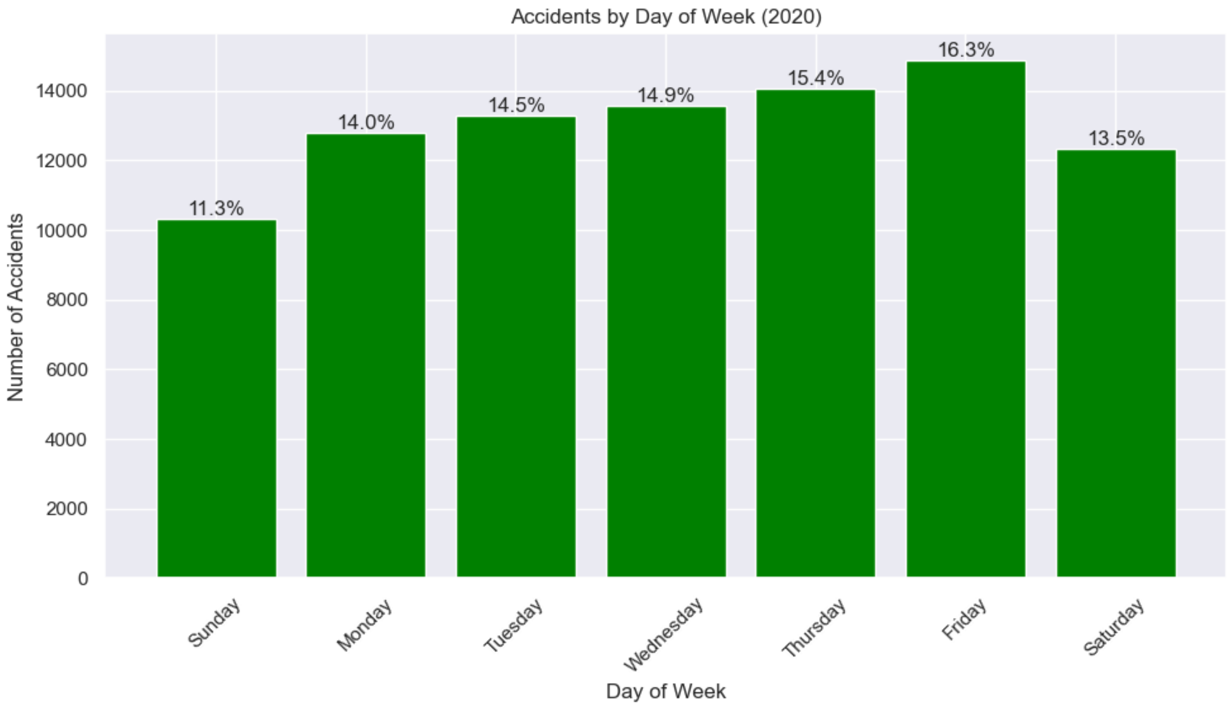


Figure 5:Number of accidents by day of week.

Question 2

Motorcycle riders make up a significant portion of road accident victims (Afkhaminia et al., 2018), highlighting the vulnerability of motorcyclists on the road.

Figure 6 highlights that motorbike riders are at the highest risk of accidents during specific hours of the day. The highest number of motorbike accidents occurs at 5 PM, accounting for 9.9% of the total accidents, likely coinciding with the evening rush hour. There is also a noticeable increase in accidents during the morning rush hour, peaking at 11 AM. Accidents significantly decrease after 7 PM, with the lowest number occurring between 12 AM and 5 AM.

Figure 6 and 7 shows that Fridays are the most dangerous days for motorcycle riders, particularly for those riding motorcycles in the 50cc-125cc and over 500cc categories.
This trend might be attributed to increased traffic volumes, end-of-week fatigue, and possibly more social activities.

Furthermore, the number of accidents steadily increases from Sunday through Friday, with the highest number of accidents occurring on Fridays, which accounts for 16.3% of all accidents.

These results align with the study by Rana et al. (2022) on motorcycle accidents in Punjab which showed that 54.6% of accidents occurred on weekdays (Monday to Friday).

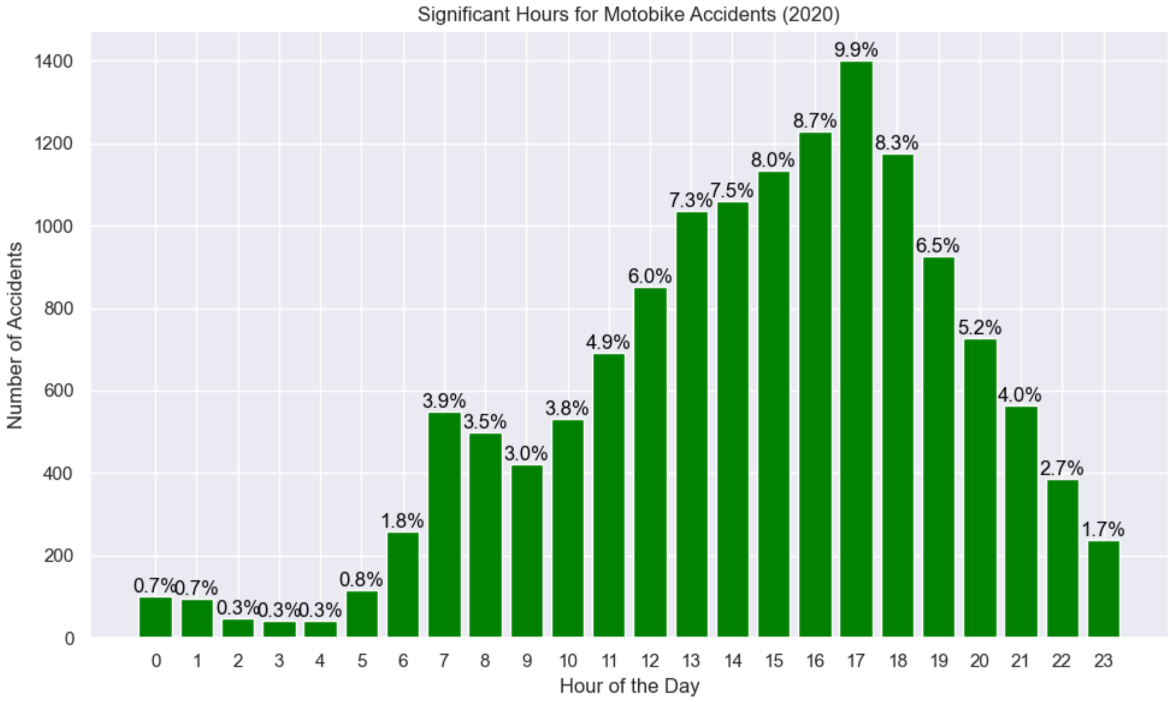


Figure 6: Significant hours of motorbike accidents.

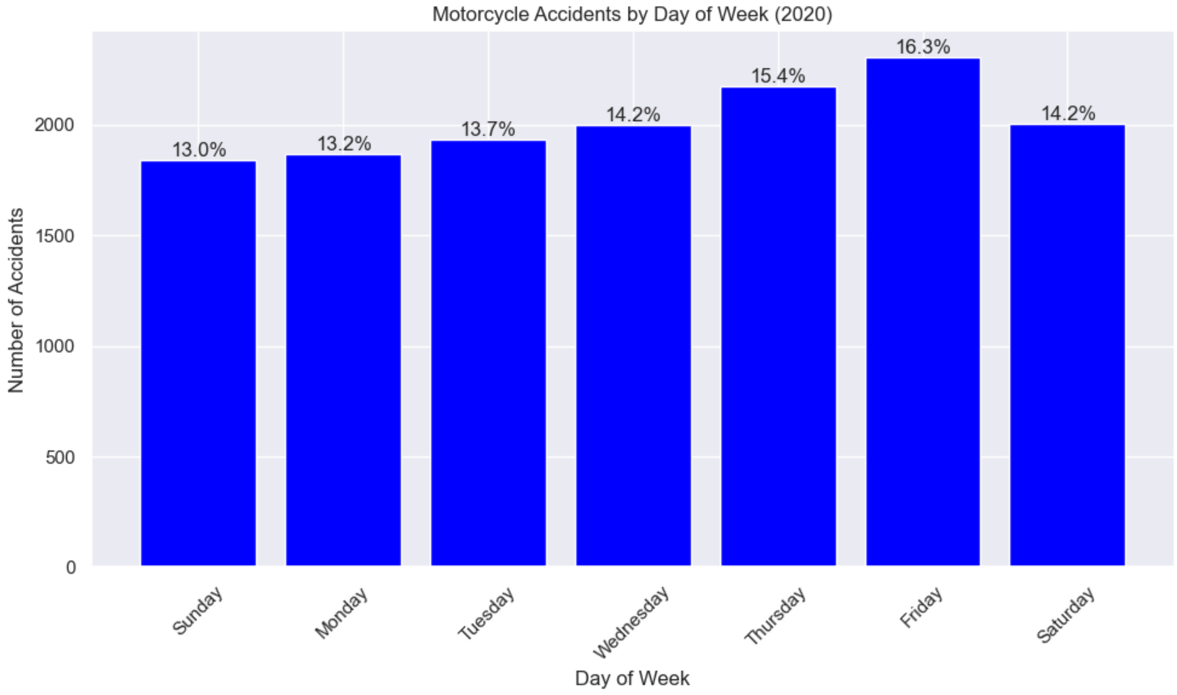


Figure 7: Motorcycle accidents by day of the week.

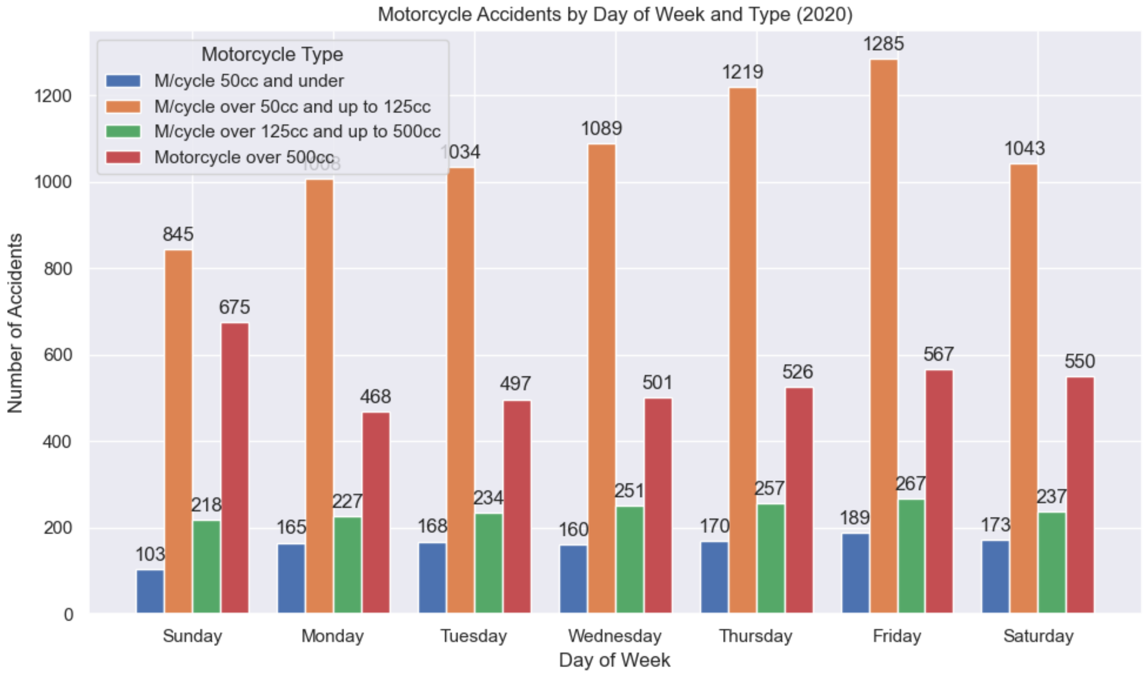


Figure 8:Motorcycle accidents day of week and type.

Question 3

Pedestrian-involved accidents are more common during rush hours when there is a high volume of pedestrian and vehicular activity on roads and streets. The highest number of accidents occurs in the late afternoon and early evening hours, particularly between 3 PM to 6 PM. Accidents involving pedestrians also occur during morning rush hours, with notable counts recorded between 7 AM to 9 AM. While the number of accidents decreases after 6 PM, significant incidents still occur during the evening and night hours, especially between 7 PM to 10 PM. The fewest incidents occur between 1 AM to 5 AM.

Friday has the highest number of accidents, with a total of 14,889 accidents, followed closely by Thursday with 14,056 accidents and Wednesday with 13,564 accidents. Tuesday recorded 13,267 accidents, making it the fourth highest, followed by Monday with 12,772 accidents. Saturday recorded 12,336 accidents, making it the second lowest. Sunday has the lowest number of accidents, with a total of 10,315.

Friday also has the highest number of pedestrian-involved accidents, followed by Thursday and Wednesday. Saturday and Sunday have relatively fewer pedestrian-involved accidents compared to weekdays, with Sunday having the lowest count. The pattern indicates that weekdays, especially towards the end of the week, might witness more pedestrian-related accidents, possibly due to factors such as increased traffic, rush hours, or pedestrian activity in urban areas.

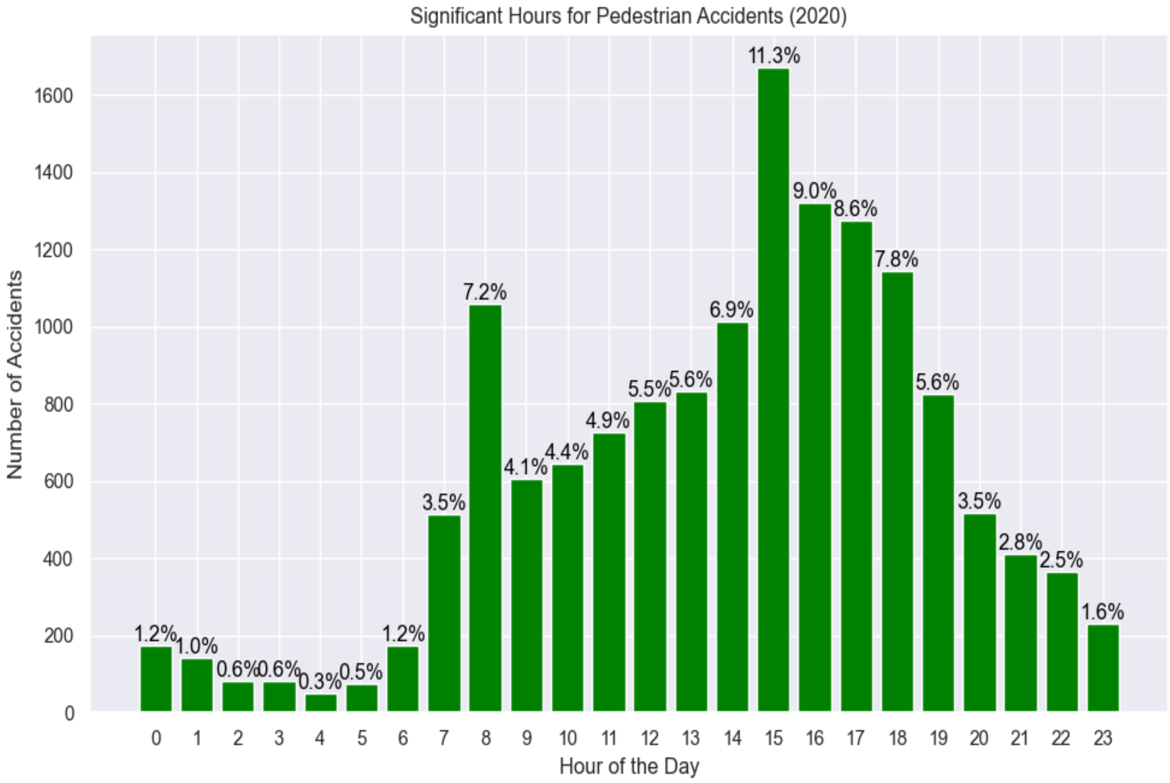


Figure 9:Pedestrian accident by hour

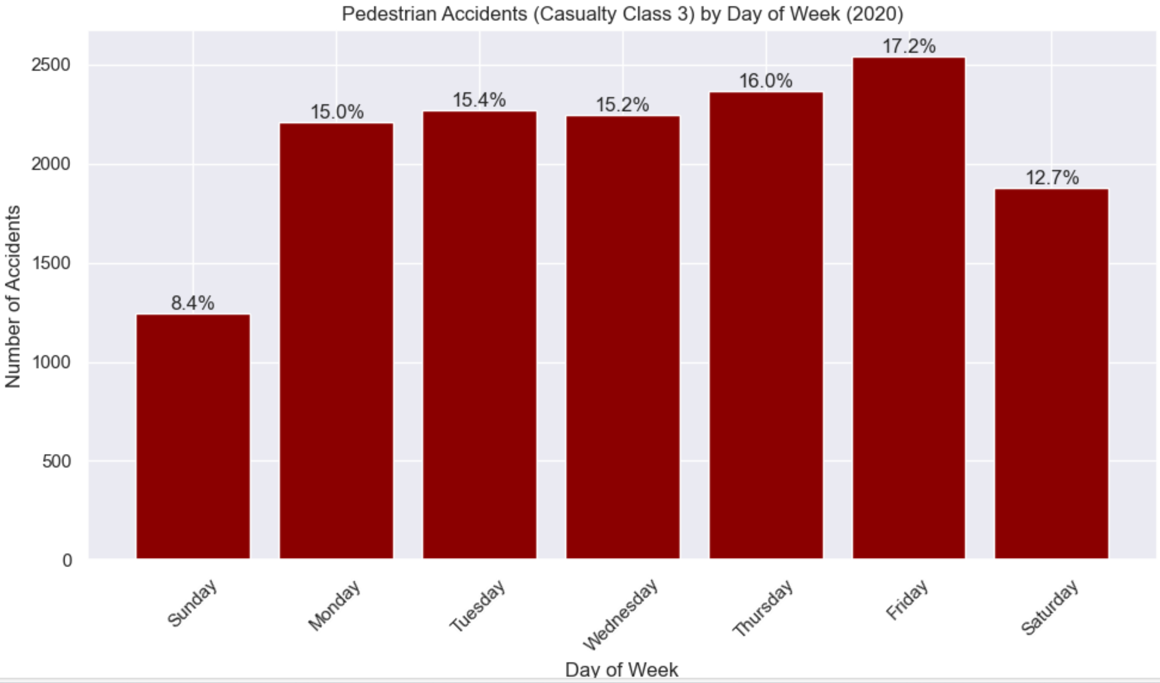


Figure 10:Pedestrian accident by day of week.

QUESTION FOUR

Rule, one states that 20% of accidents occur when streetlights are present and lit during darkness. With a confidence level of 98%, it means that under these conditions, 98% of the accidents are non-fatal. The lift of 1.004 indicates that the likelihood of a non-fatal accident slightly increases when streetlights are present and lit, but the effect is minor.

For the UK 2020 accident data, rule two states that 21% of accidents occur under unknown street lighting conditions. There is a 99% confidence in a high probability of non-fatal outcomes when the lighting conditions are unknown. The lift of 1.011 suggests a slightly higher likelihood of non-fatal accidents under these conditions, indicating a potential association between unknown lighting conditions and non-fatal outcomes.

This rule has a high support of 54%, meaning that more than half of the accidents occur in areas with a 30 km/h speed limit. The confidence is 99%, indicating a very high probability of non-fatal outcomes in such areas. The lift of 1.009 shows a slightly increased likelihood of non-fatal accidents at this speed limit, implying that lower speed limits might contribute to reducing the severity of accidents.

This condition is responsible for 12% of accidents. The 98% confidence level indicates a high likelihood of non-fatal outcomes during rain without high winds. The lift of 1.003 suggests a very minimal increase in the likelihood of non-fatal accidents under these weather conditions. This indicates that while rain affects accidents, it does not significantly increase the severity.

Only 2.1% of accidents involve unknown weather conditions. With 99% confidence, it is likely that most accidents under unknown weather conditions are non-fatal. The 1.015 lift suggests a slightly higher likelihood of non-fatal outcomes under these conditions, indicating that the unknown element might be related to less severe accidents. Only 2.1% of accidents involve unknown weather conditions. With 99% confidence, it is likely that most accidents under unknown weather conditions are non-fatal. The 1.015 lift suggests a slightly higher likelihood of non-fatal outcomes under these conditions, indicating that the unknown element might be related to less severe accidents.

Table 1:Apriori Analysis

Antecedent	Consequent	Support	Confidence	Lift
Darkness: street light present and lit	**Non-Fatal**	**20%**	**98%**	**1.004**
Darkness: street lighting unknown	**Non-Fatal**	**21%**	**99%**	**1.011**
Speed limit:30 km/h	**Non-Fatal**	**54%**	**0.99**	**1.009**
Raining without High winds	**Non-Fatal**	**12%**	**98%**	**1.003**
Unknown weather conditions	**Non-Fatal**	**2.1%**	**99%**	**1.015**

QUESTION 5

Clustering was perfomedon the dataset to determine accident clusters in the Humberside region.

Humberside region includes East Riding of Yorkshire, North Lincolnshire, Kingston Upon Hull, and Northeast Lincolnshire. The total number of accidents in the region was 1663. The highest percentage of accidents occurred in Kingston upon Hull (34.22%), followed by East Riding of Yorkshire (29.34%). North Lincolnshire and Northeast Lincolnshire had similar percentages of accidents at 18.28% and 18.16% respectively as shown in figure 8.

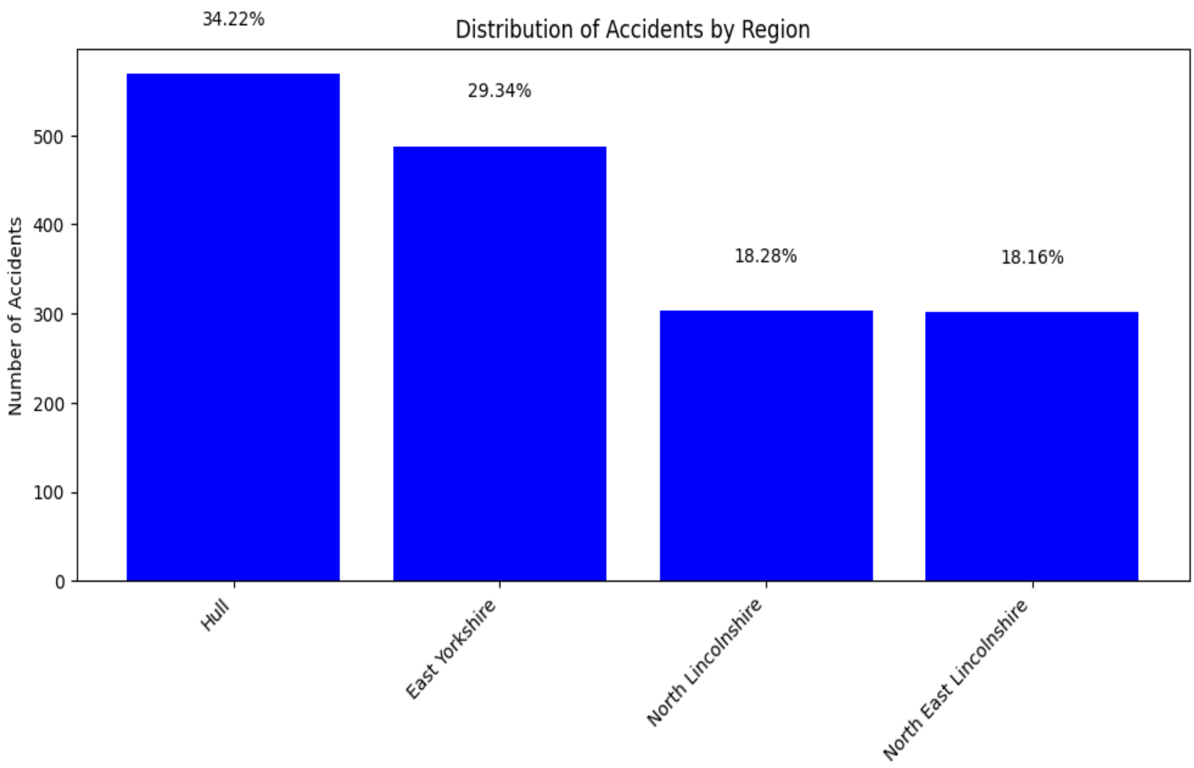


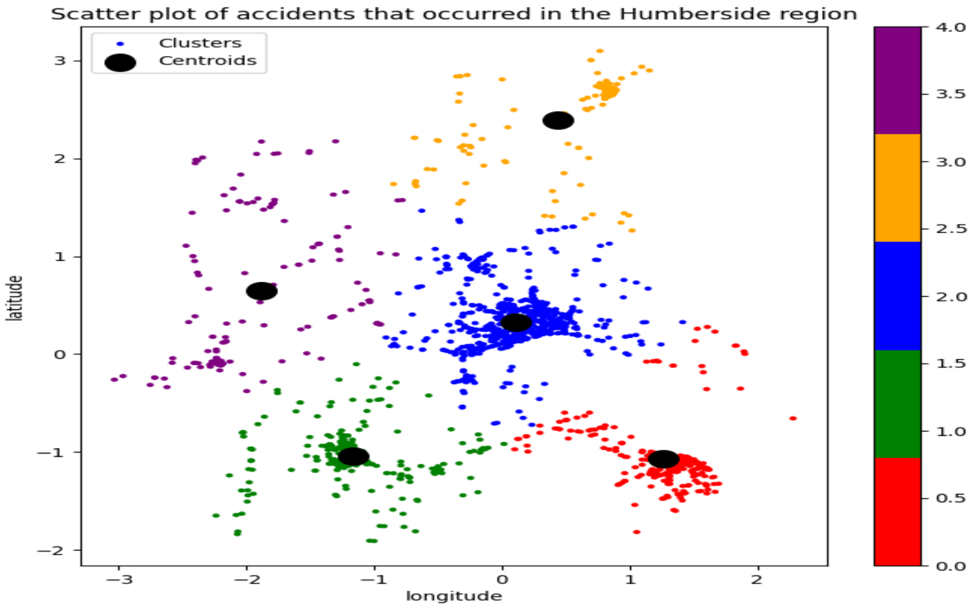
Figure 11:Distribution of accidents in the Humberside Region.

Table 2:Clustering Analysis

Clustering Algorithm	Silhouette Score
KMEANS	0.6223
DBSCAN	0.3376
KMEDOID	0..4342

KMeans (0.6223) outperforms both KMedoids and DBSCAN in terms of silhouette score, indicating well-defined clusters with minimal overlap. Table 2

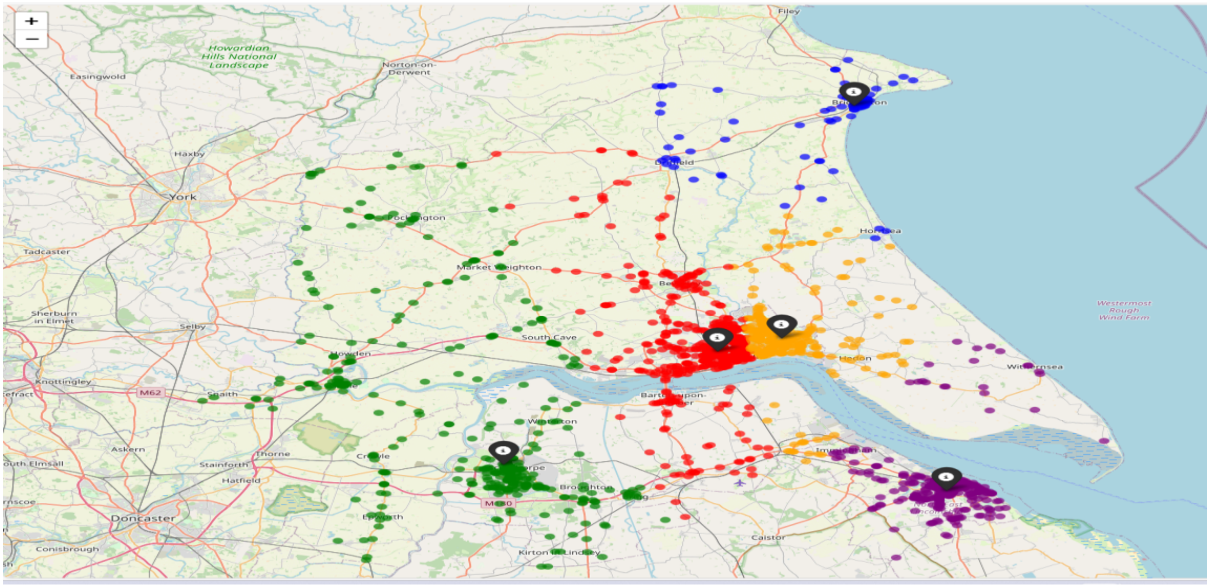
The map visualizes the clustering of road traffic accidents in the Humberside region. Each colored dot represents an accident, with different colors indicating distinct clusters. This type of visualization helps in identifying areas with high accident concentrations and understanding the spatial distribution of accidents.

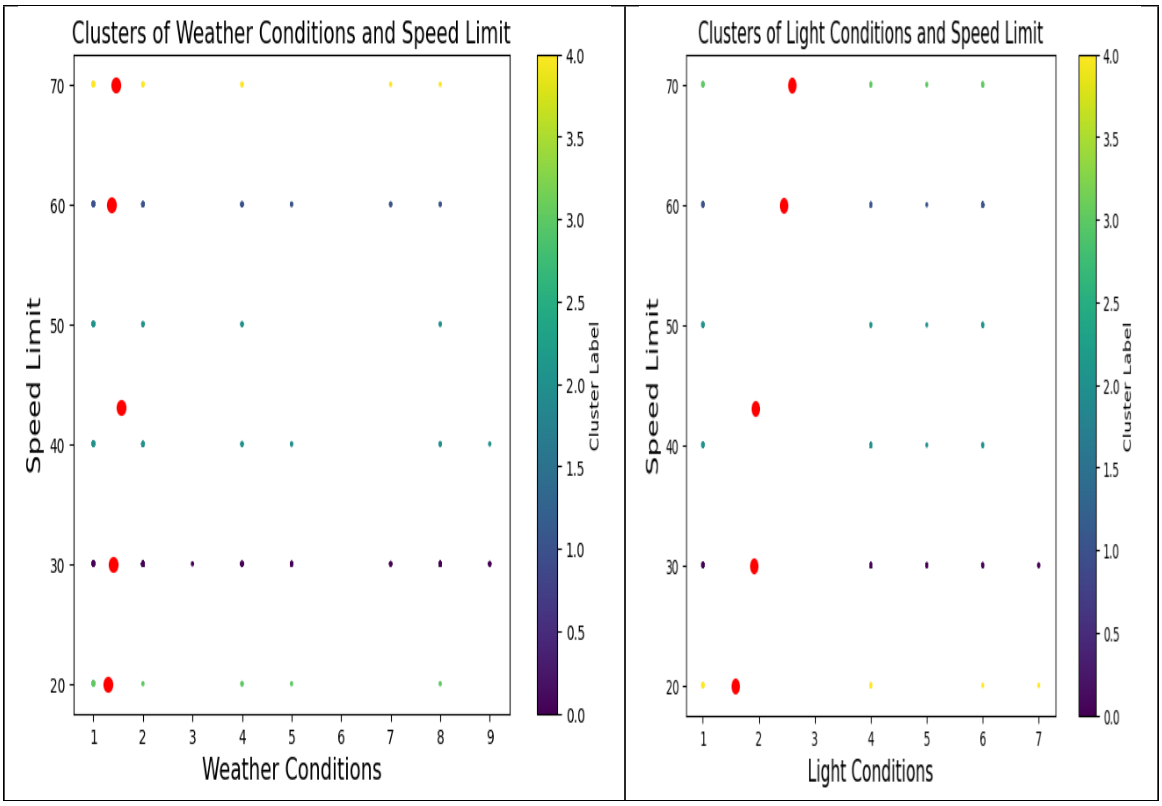


The area around Hull, indicated by a dense concentration of orange and red dots and blue dots on the scatter plots, shows a significant number of accidents. This suggests that Hull is a critical area for road safety interventions. The Scunthorpe and Grimsby areas, represented by green and purple dots on the map, also show high accident densities, indicating the need for targeted safety measures in these locations. Beverley and surrounding areas (blue dots) have a moderate concentration of accidents. This may be due to less traffic compared to urban centers but still represents areas with potential safety concerns.

High speed limits (60 to 70 km/h) generally show clustering in varied weather conditions but tend to have more concentrated clusters. Low speed limits (20 km/h) also exhibit distinct clustering patterns, potentially indicating areas with specific traffic control measures or residential zones. Intermediate speed limits (30 to 50 km/h) show more dispersed clustering, possibly reflecting diverse road types and environments.

The plot displays clusters labelled 0 to 4, each indicating different groupings of light conditions and speed limits. The color bar serves as a reference for cluster labels, with different colors representing different clusters. Some clusters are spread across multiple light conditions, indicating consistent speed limits regardless of lighting. The red cluster (Cluster 4) shows prominent points at higher speed limits, suggesting a focus area for potential safety interventions.

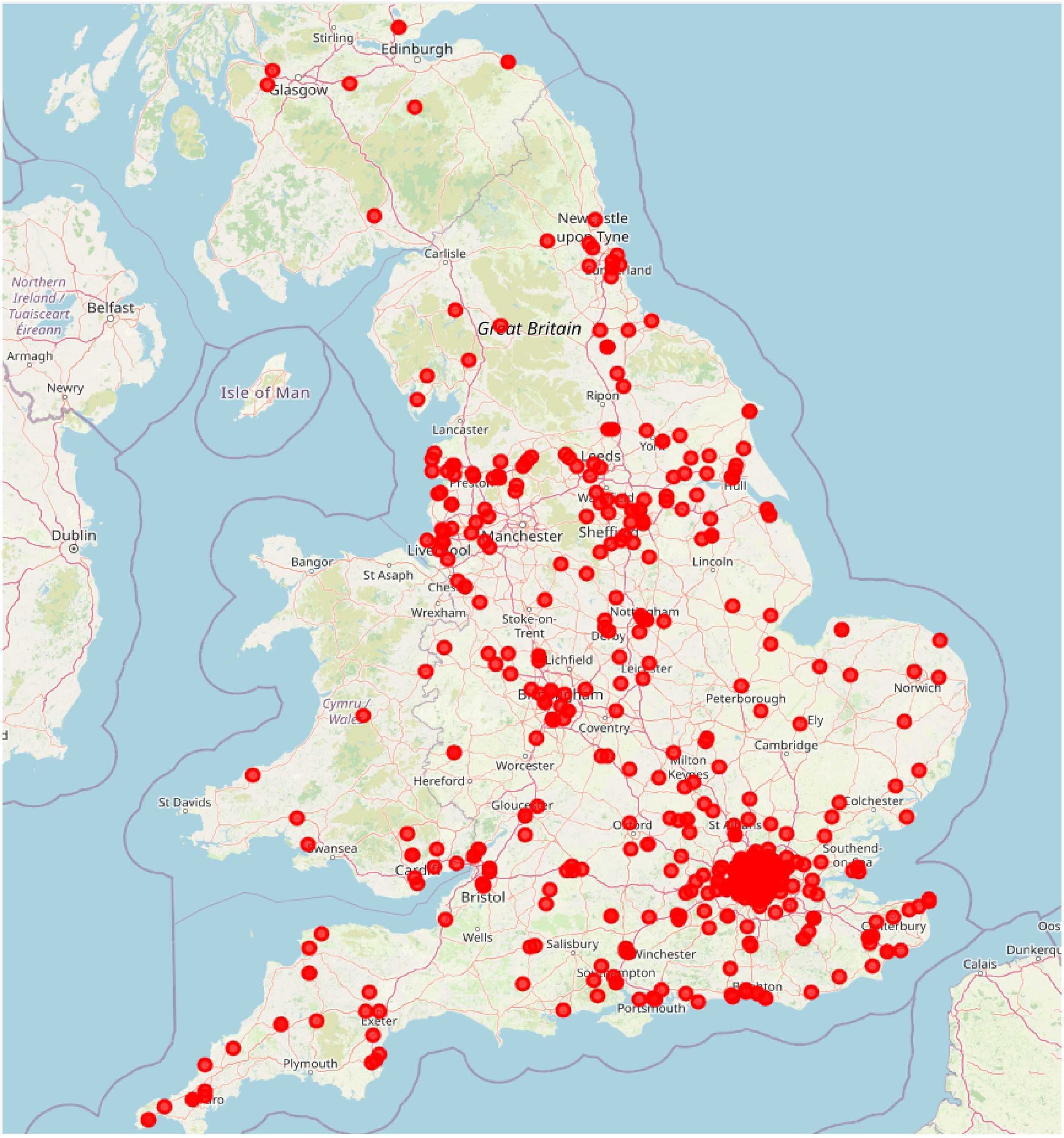


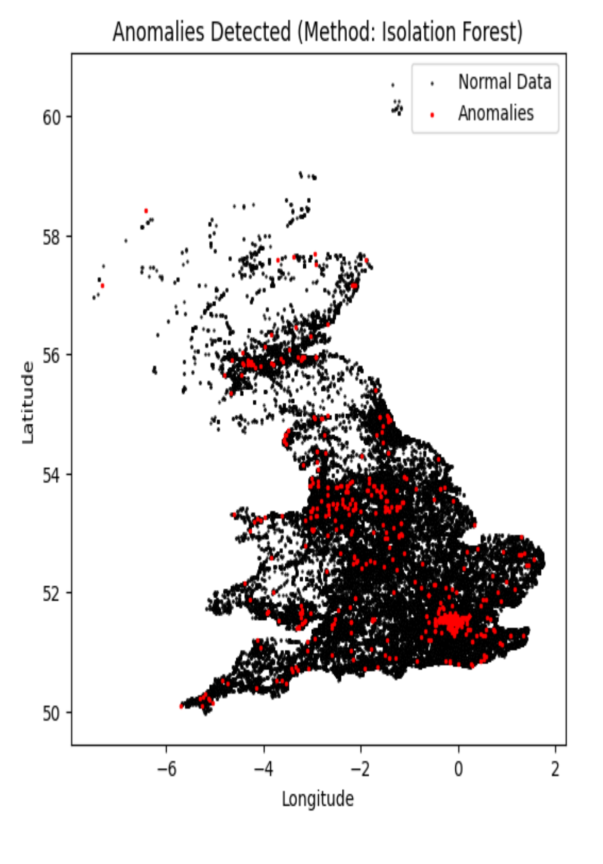


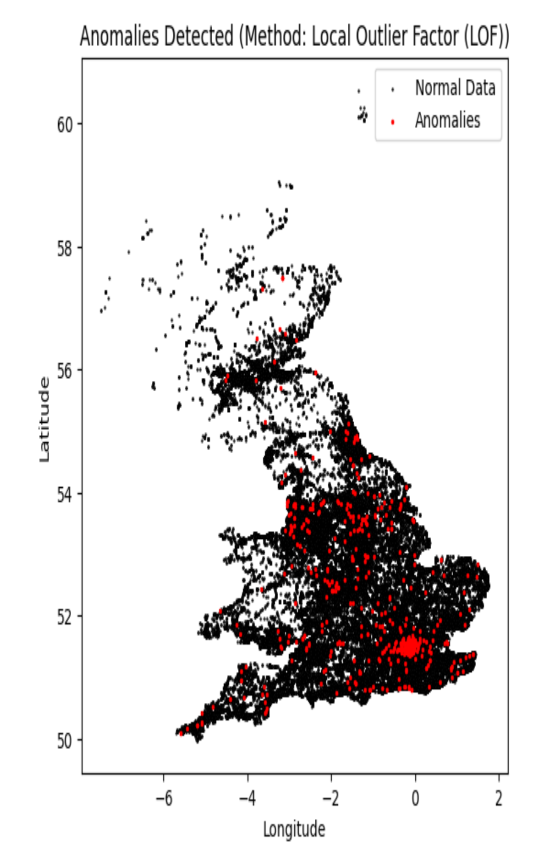
QUESTION SIX

The scatter plot and map visualize anomalies detected in the geographical distribution of road traffic accidents across the UK. Normal data points are shown in black, while anomalous data points are highlighted in red, identified using the Isolation Forest method and Local outlier factor.

most accidents occur in densely populated and urbanized areas like London, Birmingham, Manchester, and Liverpool. Anomalies are scattered throughout the map, with concentrations in central and southern England, as well as in less populated regions like parts of Wales and Scotland. Possible reasons for anomalies include geographical errors, unusual accident patterns, and data entry issues. Certain clusters of anomalies suggest specific areas where accident data significantly deviates from the norm, possibly due to changes in traffic patterns, road conditions, or external factors like weather.







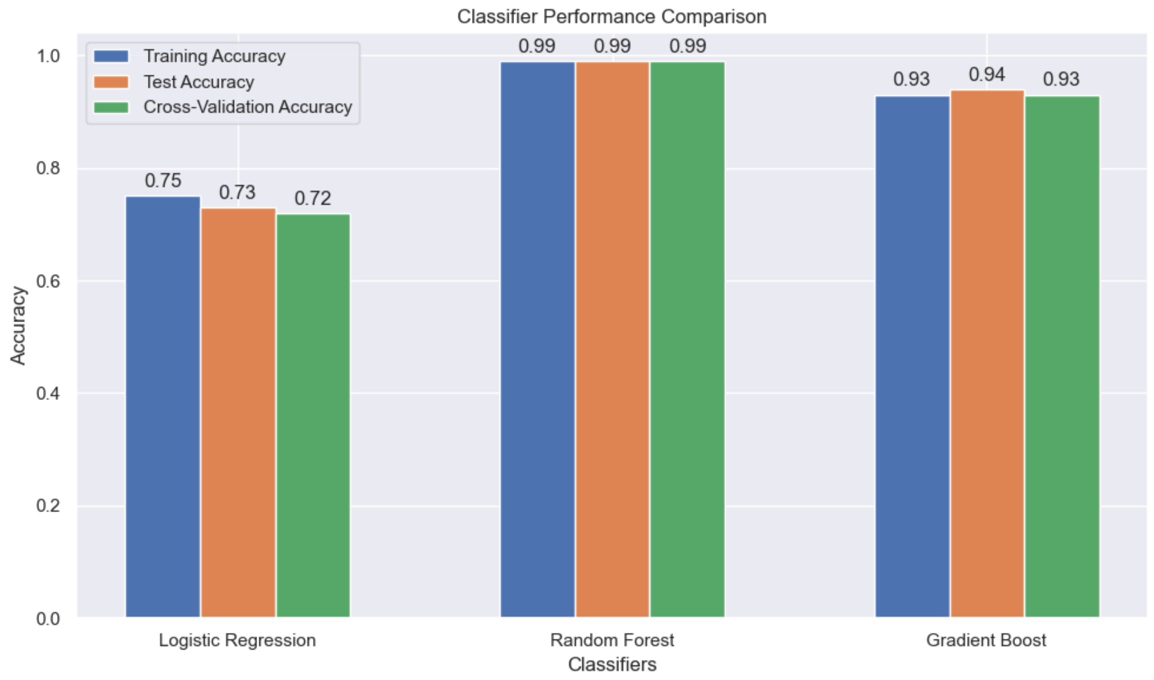
QUESTION SEVEN

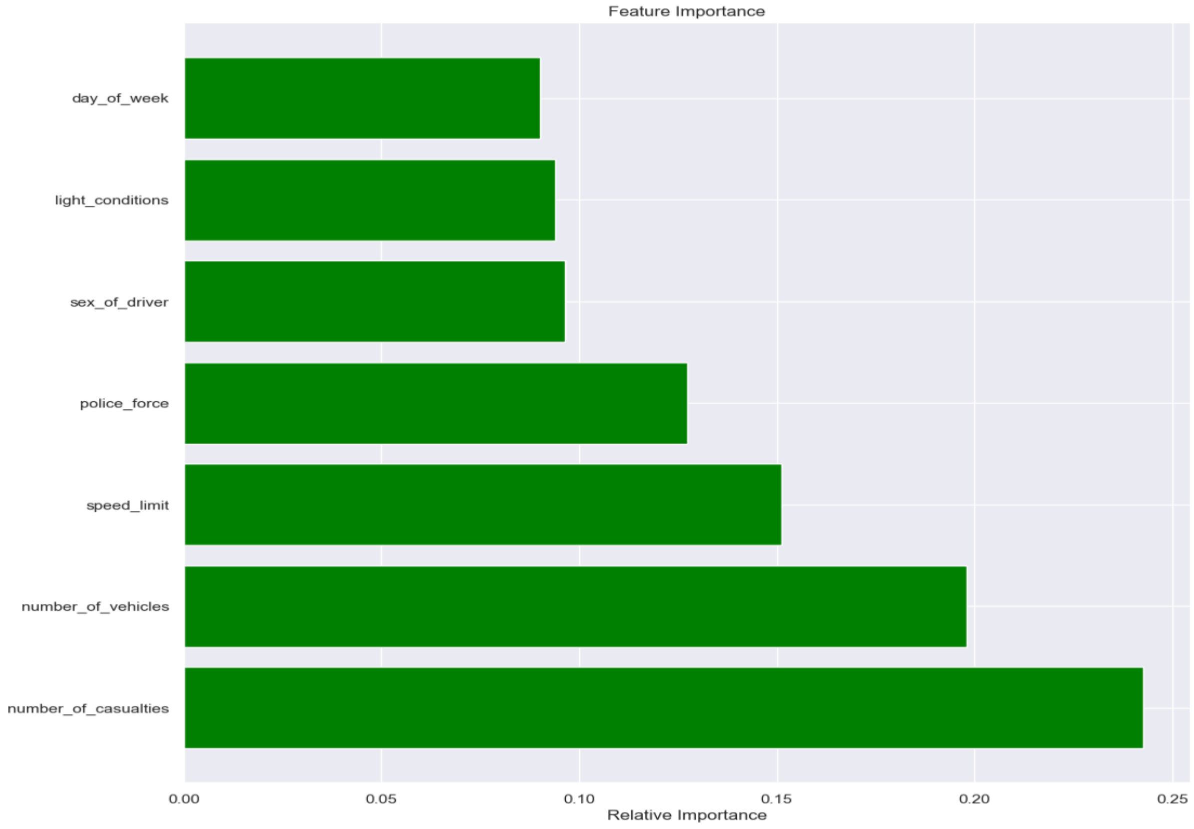
Table 3:Summary of Model Performance

CLASSIFIER	TRAINING ACCURACY	TEST ACCURACY	CROSS VALIDATION ACCURACY
Logistic Regression	**0.75**	**0.73**	**0.72**
Random Forest	**0.99**	**0.99**	**0.99**
Gradient Boost	**0.93**	**0.94**	**0.93**

- **logistic Regression**: Shows decent performance but has the lowest accuracy among the three classifiers. The close values for training, test, and cross-validation accuracies suggest that it is a stable model but may not be capturing the complexity of the data as well as the other models.
- **Random Forest**: Demonstrates excellent performance with high accuracies across all metrics. This model is very well-suited for the dataset and shows no signs of overfitting.
- **Gradient Boost**: Also shows strong performance, slightly lower than Random Forest but still very high. It generalizes well and is stable across different data splits.

Overall, **Random Forest** is the best-performing model in this comparison, followed closely by **Gradient Boost**. **Logistic Regression** performs adequately but is outperformed by the other two. The random forest feature selection is shown below.





RECOMMENDATIONS

- Accidents are most common during rush hours when traffic is heavy. Improved traffic management and road safety measures are crucial during peak hours (14:00 to 17:00) and morning rush hours (06:00 to 09:00).
- Raising awareness about road safety during these critical hours can help reduce accidents.
- Road safety campaigns and increased traffic enforcement are especially beneficial from Tuesday to Friday when accident rates are higher.
- Raising awareness about the risks and promoting safe driving practices could help reduce accidents on Fridays.
- Road safety initiatives could be focused on these peak hours to reduce accidents, such as increased patrolling, awareness campaigns, and encouraging safe riding practices.
- Motorbike riders should be particularly cautious during the identified high-risk hours, adjusting their riding habits to avoid potential accidents.

REFERENCES

Afkhaminia, F., Charati, J., Rahimi, E., & Nasab, N. (2018). Epidemiological study of the suburban accident mortalities recorded in Golestan, Iran in 2015. *Jorjani Biomedicine Journal, 6*(1), 67-73. https://doi.org/10.29252/jorjanibiomedj.6.1.67

Borucka, A., Kozłowski, E., Oleszczuk, P., & Świderski, A. (2020). Predictive analysis of the impact of the time of day on road accidents in Poland. *Open Engineering, 11*(1), 142-150. https://doi.org/10.1515/eng-2021-0017

Machetele, D., & Yessoufou, K. (2021). A decade long slowdown in road accidents and inherent consequences predicted for South Africa. https://doi.org/10.20944/preprints202103.0713.v1

Meyyappan, A., Subramani, P., & Kaliamoorthy, S. (2018). A comparative data analysis of 1835 road traffic accident victims. *Annals of Maxillofacial Surgery, 8*(2), 214. https://doi.org/10.4103/ams.ams_135_18

Mishra, B., Sinha, N., Sukhla, S., & Ak, S. (2010). Epidemiological study of road traffic accident cases from Western Nepal. *Indian Journal of Community Medicine, 35*(1), 115. https://doi.org/10.4103/0970-0218.62568

Pradhan, M., Upadhyay, H., Shrestha, A., & Pradhan, A. (2023). Road traffic accident among patients presenting to the emergency department of a tertiary care centre: A descriptive cross-sectional study. *Journal of Nepal Medical Association, 61*(258), 127-131. https://doi.org/10.31729/jnma.8032

Rana, A., Nasrullah, F., Hameedi, S., & Janjua, N. (2022). Motorcycle accidents in Punjab: A critical analysis. *Pakistan Armed Forces Medical Journal, 72*(2), 632-636. https://doi.org/10.51253/pafmj.v72i2.5797

Widjajanti, E. (2021). Karakteristik kecelakaan lalu lintas pada jalan tol Jagorawi km 19–km 40 kabupaten Bogor. *Borneo Engineering Journal Teknik Sipil, 5*(1), 76-88. https://doi.org/10.35334/be.v5i1.1648

Yao, W. (2023). Analysis of factors influencing the severity of road traffic accidents. https://doi.org/10.1117/12.2668530