Table of Contents

Table of Figures

1. Introduction

Road transportation, as a vital mode of travel, accounts for a significant 92% of passenger kilometres in Great Britain (Department for Transport, 2021). However, it also registers a concerning average of 90% fatalities per billion kilometres travelled during 2011-2020 (Statista, 2023). To address these safety concerns, the accident database comprising Accident, Vehicle, Casualty, and Lower Layer Super Output Area (LSOA) datasets is analysed. The analysis seeks to identify potential outliers, significant patterns of accidents occurrence in terms of hours and days, the impact on pedestrians, motorbikes, and other determinants of accident severity. Additionally, it will explore accident patterns within the Humberside region to gain insights into road safety. The study aims to accurately predict fatal injuries sustained in road traffic accidents, offering insights to inform and improve road safety measures implemented by the government.

2. Foundations: Literature Review and Data Preprocessing

2.1 Review of Related Literature

Traffic related accidents has remained a pressing public health concern. Studies have applied machine learning to uncover patterns in road safety data. Islam et al., (2021) used clustering with crash features to identify high-risk accident types. Association rule mining revealed connections between accident conditions and severity Verma & Rashmi, (2019). Despite progress, more research is needed to improve safety interventions using data-driven insights. This study aims to analyse British accident data to determine key factors influencing road safety outcomes.

2.2 Data Preprocessing

The data cleaning process involved accessing the database using the `sqlite3` library within a Jupyter Notebook environment. The extracted datasets for the year 2020 were converted into pandas DataFrames, comprising four tables: "Accident" (91,198 rows, 36 columns), "Vehicle" (167,374 rows, 28 columns), "Casualty" (115,583 rows, 19 columns), and "LSOA" (34,377 rows, 8 columns). Handling missing values was a critical aspect of the data cleaning process. For the "Accident" table, null values in the coordinate columns latitude were addressed by imputing the median values of similar rows based on matching "local_authority_ons_district" and "police_force" values.

Categorical columns containing -1, 99, or 999 were converted to NaNs. NaNs were replaced using the mode or median values, except for the "lsoa_of_accident_location," where NaNs were assigned the label "unknown." Additionally, columns related to driver and casualty ages with values of -1 were converted to NaNs and then filled with the mode for each respective column. By thoroughly handling missing values, the data cleaning process ensured that the final datasets were ready for in-depth analysis and machine learning.

3. Analysis and Result

3.1 Significant Hours and Days of Accidents

Analysis of the aggregated accident data in terms of hours and days reveals most accidents occur during the midday hours, until early evening around 7:00 PM. Additionally, there is a notable spike in accidents during the morning at 8:00 AM. The top three hours with the highest accident count are 17:00 (5:00 PM), 16:00 (4:00 PM), and 15:00 (3:00 PM) ranked in descending order.

Highest accident counts by days in descending order are Thursday, Wednesday, Tuesday, Monday, Saturday, and Sunday. The significant hours and days of accident occurs can be attributed to factors like:

1. Peak Traffic Hours: Corresponding to work, school, and activity schedules, mornings and evenings see heightened traffic volume.

2. Fatigue and Distraction: Evening accidents may relate to driver fatigue. Increasing weekday accidents suggest escalating stress affecting focus. Elevated Thursday and Friday rates could stem from weekend anticipation and social engagement.

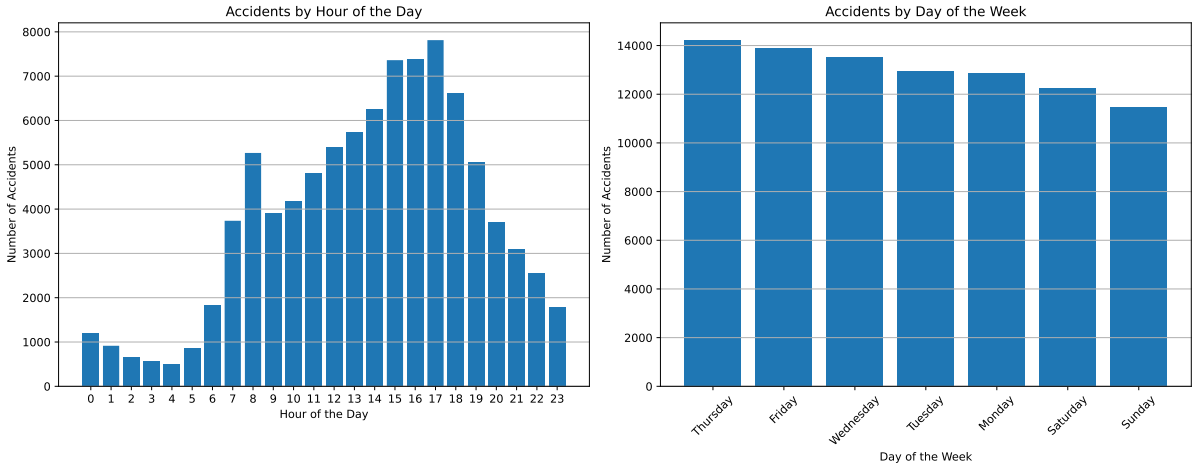

Figure 1: Accidents peak midday through early evenings and on Thursdays/Fridays, highlighting opportunities for targeted road safety interventions.

3.2 Analysis of Motorcycle Accidents by Hour and Day

Motorcycles under 500cc exhibit peak accidents from 3pm to 6pm, peaking at 5pm. Morning hours are calm, with accidents rising from 6am to 12pm. Fridays see the highest accidents, weekdays show gradual rise, and Sundays have the least. Likely causes include rush hour traffic, fatigue, and poor visibility. Mitigation involves improved road signage, rider awareness campaigns, and enforcing road safety regulations.

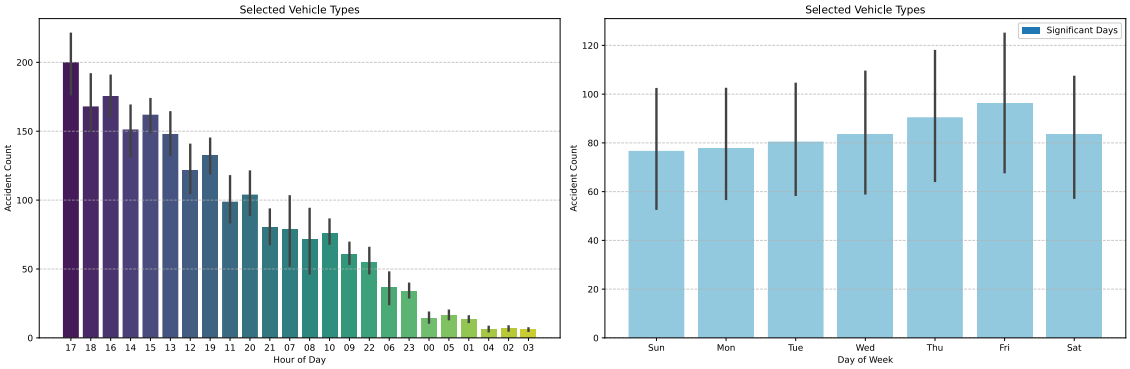

Figure 2: Motorcycle accidents peak afternoons to evenings, especially Fridays, while fewer occur mornings and Sundays, revealing opportunities for targeted prevention.

3.3 Significant Hours and Days for Pedestrian Accidents

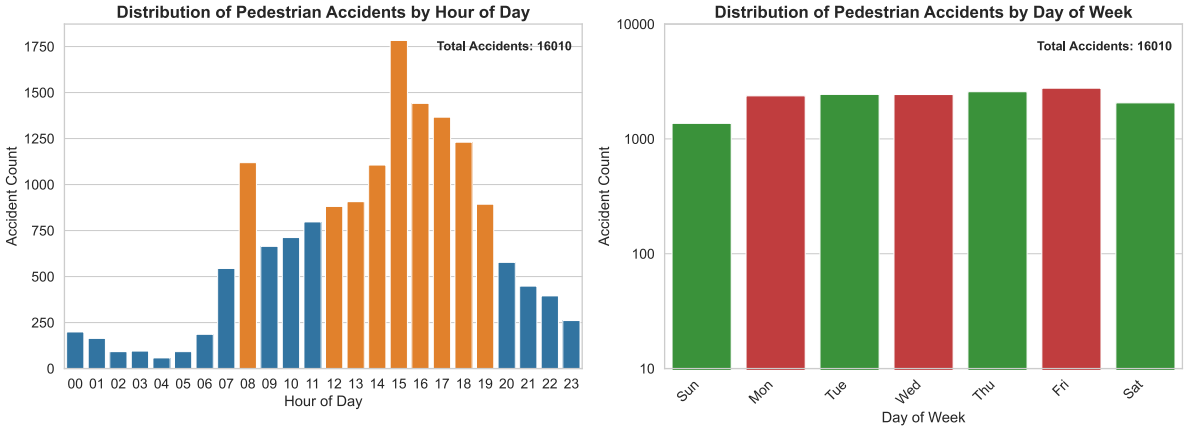

Figure 3: Peak pedestrian accidents observed from noon to early evening, with a notable spike at 8 AM during commuting hours. Accidents rise gradually from Sunday to Friday, then decrease on Saturdays.

Figure 3 shows peak accident times align with midday to early evening, consistently exceeding 1000 accidents per hour. Notably, 8 AM sees a substantial surge, coinciding with work, school, and business commute hours. In terms of daily occurrences, pedestrian accidents progressively rise from Sunday to Friday, experiencing a sharp decline on Saturdays. Pedestrian accidents can occur due to the following reasons:

a) Distraction and Inattention: High-end smartphones, EarPods and other devices used during walks does contribute to accidents due to unawareness of their surroundings.
b) Traffic Congestion and commuting rush: Increase in the volume of moment during peak hours increases the likelihood of accidents.
c) Fatigue, Stress and Weekend Factors: Pedestrian accidents rise Sunday to Friday indicating fatigue tied to work schedules, then fall on Saturdays with less work-related travel.

3.4 Exploring Impact of Selected Variables on Accident Severity using Apriori Algorithm

The process for exploring the Apriori algorithm on selected features entails the following steps:

Figure 4: The process for exploring Apriori algorithm.

The accident, vehicle and casualty tables were merged using the accident index column. Feature selection was executed on a few variables using random forest which considers interactions between features and captures nonlinear relationships. Five variables were selected then converted to dummy variable for the apriori execution and association generation using a minimum support of 0.2 and a minimum threshold of 0.8. The table shows the dominant features from the dataset are casualty, severity, and engine. The association rule mining table provide insights into the relationships between some of the antecedents and consequents in the dataset.

Table 1: Association Rule Metrics: Selected association rules with support, confidence, and lift metrics, highlighting relationships and strengths of association between antecedents and consequents.

Antecedents	Consequents	Antecedent Support	Consequent Support	Support	Confidence	Lift	Leverage	Conviction	Zhang's Metric
casualty_3	severity_3	0.813457	0.7692630	0.769263	0.945672	1.229321	0.143501	4.247081	1.000000
severity_3	casualty_3	0.769263	0.813457	0.769263	1.000000	1.229321	0.143501	inf	0.808467
engine_1597.0	severity_3	0.274896	0.769263	0.207532	0.754949	0.981392	-0.003935	0.941587	-0.02548

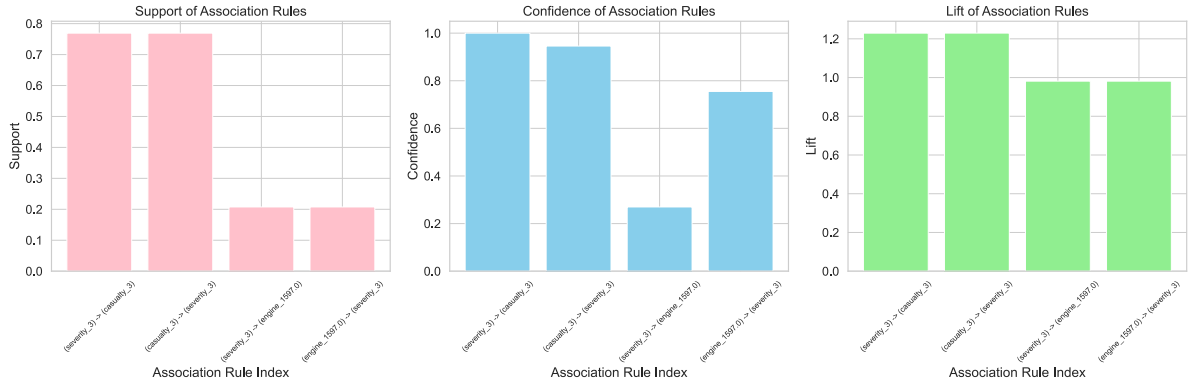

Figure 5: High support (casualty_3: 81%, severity_3: 77%) and strong predictive link (confidence: 0.945672) between variables. Lift value (1.229321) indicates robust association.

Figure 3 displays the outcomes of association rule mining for support, confidence, and lift metrics. The most notable comparison is between the casualty_3 and the severity_3. Subplot 1 highlights frequency of occurrence, casualty_3 of 81% and a severity_3 of 77% for the support, subplot 2 highlights the confidence value of 0.945672 indicate a strong predictive relationship between casualty_3 and severity_3. Subplot 3 indicates the strength with lift value of 1.229321 suggests strong relationship between both features.

3.5 Identification and Clustering of Accidents in the Humberside Region

To understand the accident distribution in the Humberside region, the dataset was filtered using the police force as 16 representing Humberside from the documentation. To reveal patterns and hotspots within the region, the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is deployed. According to Islam et al. (2021) this algorithm groups features that are near each other while identifying outliers or noise points.

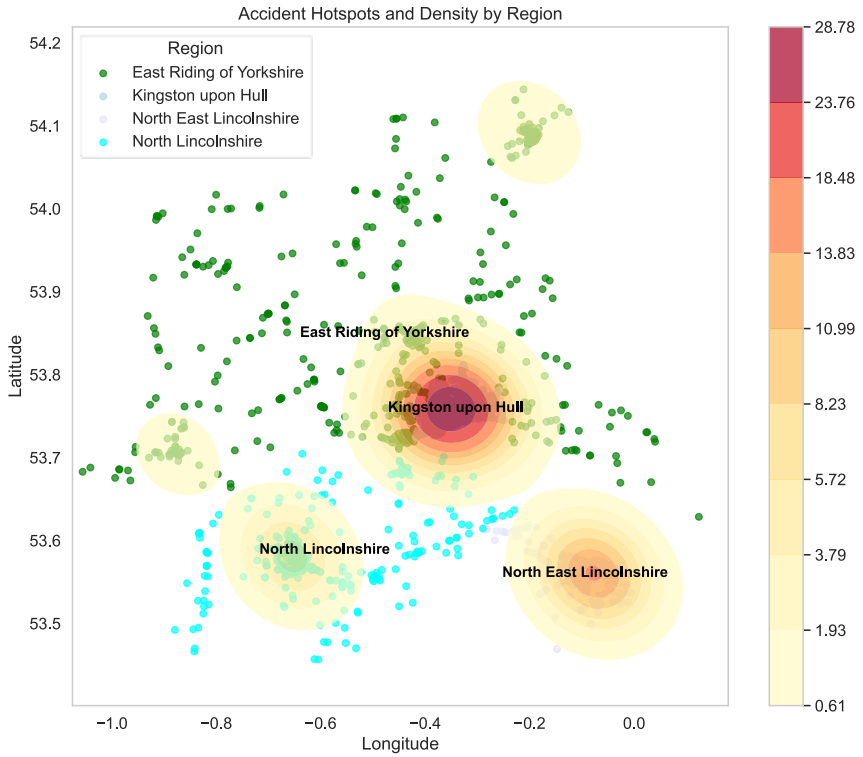

Figure 6: Accident Hotspots and Density: Spatial distribution of accident hotspots and density in the Humberside region. Clustering analysis using DBSCAN algorithm highlights high-density areas.

From **Figure 6**, the cluster density shows Kingston upon Hull has the highest density of accident, North and East Lincolnshire has strong clustering and the accident within the East Riding of Yorkshire is dispersed which might be due to the size of the area. This shows the highest area where accident occurs is the Kingston upon Hull which is also the most populated local authority in the region according to Hull Data Observatory (2021).

3.6 Identifying Unusual Entries through Outlier Detection

To identify unusual entries within the dataset the focus was on the geographical coordinates (longitude and latitude) of accidents and the age of drivers and vehicles involved.

a) **Age of Vehicles Analysis:** The Local Outlier Factor (LOF) method was deployed which evaluates how isolated certain points are from the local density deviation of a data point concerning its neighbours Abhaya (2022). The outliers were set at ages higher than 25 years which falls below the EURO 4 standard as minimum which for petrol cars according to expansion of Ultra Low Emission Zones of European emissions standards Evo Magazine (2023).

b) **Age of Drivers Analysis:** Driver's age was visualized using a histogram with a minimum value of 3 and a maximum value of 100 years. While the official age a person can start driving is 17 years, there are certain exemptions for 16 years olds according to (GOV.UK, 08/2023). All the age younger than 16 years are deemed as possible outliers.

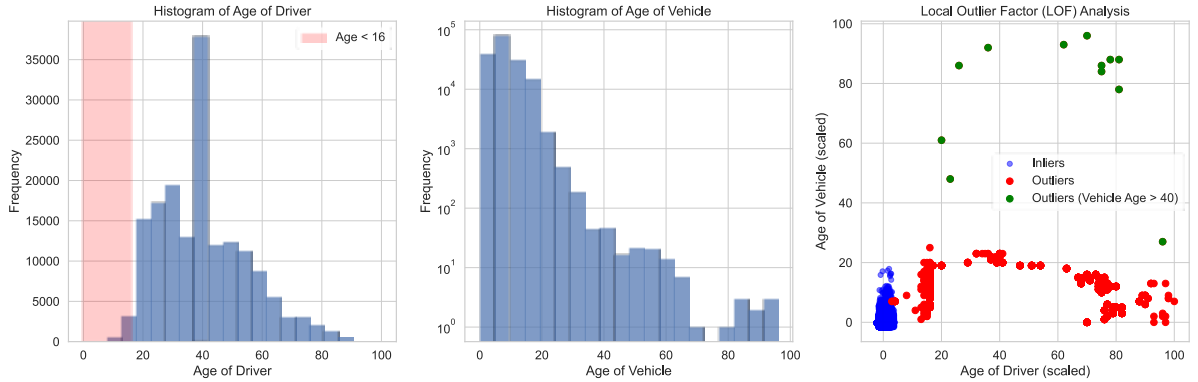

Figure 7: Geographical coordinates analysis for outlier detection based on ages above 25 years, in line with EURO 4 standard. Age analysis of driver ages with ages below 16 as potential outliers.

From **Figure 7**, the result shows outliers for the vehicles age from the third subplots where the ages are above 25years. Though it goes contrary to the European standard, there is still not sufficient information to get rid of the data points. The age of drivers less than 16 should be replaced by the median values of ages in the dataset.

c) **Geographical coordinates analysis:** The geographical coordinates were analysed for outlier detection using the latitude and longitude as features. The Isolation Forest algorithm, which isolates anomalies by creating isolation trees was used for this analysis Lesouple et al., (2021). Outlier flags were assigned to data points based on their predicted status as outliers or inliers. In the KMeans clustering analysis, 4 clusters were chosen using the elbow method and centroids were determined for each cluster.

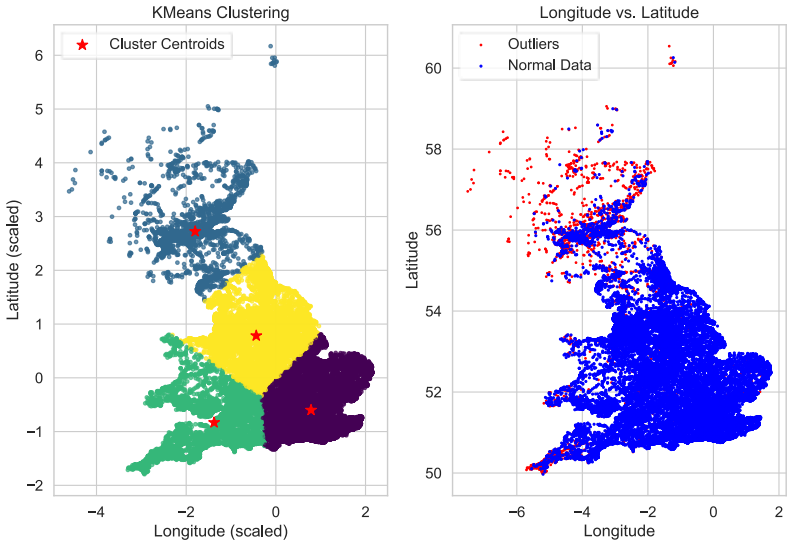

Figure 8: Geographical Coordinates Analysis: Utilizing Isolation Forest for outlier detection, latitude and longitude were analysed. Outliers were flagged based on anomalies, and KMeans clustering with 4 clusters identified centroids.

Figure 8 shows the outliers detected using the Isolation Forest matches with one of the clusters. While these are possible outliers, they are also geographical coordinates, so it is left in the data.

3.7 Developing a Classification Model for Predicting Fatal Injuries in Road Traffic Accidents

A highly efficient fatal accident prediction model is very important in reducing the tendency of accident fatalities in the UK. Using the supervised technique, the following steps were taken:

Figure 9: Steps for developing a predictive model.

a) **Dataset Preprocessing:** Preprocessing integrated tables, extracted the binary target, removed biases, balanced classes via under-sampling to enable unbiased modelling.
b) **Feature selection:** The SelectKBest identified the 15 most relevant features, objectively eliminating redundancy, before modelling.
c) **Baseline Model (Model Training)**: The decision tree classifier was used alongside a using 10-fold cross-validation on the training set to obtain a baseline model. It achieved an average accuracy of 0.731 across the folds.
d) **Algorithm Selection and Hyperparameter Tuning (Model Selection & Optimization):** Systematic comparison on a held-out set determined hyperparameter-tuned Random Forest as optimal, significantly outperforming Gradient Boosting and Voting Classifier with 85% accuracy.
e) **Test Set Evaluation (Model Evaluation):** The tuned Random Forest model demonstrated excellent predictive performance, achieving 0.852 accuracy, balanced 0.84-0.86 F1 scores, strong precision/recall exceeding 0.80, plus 671 true positives and 760 true negatives, confirming its capabilities for imbalanced classification.

f)

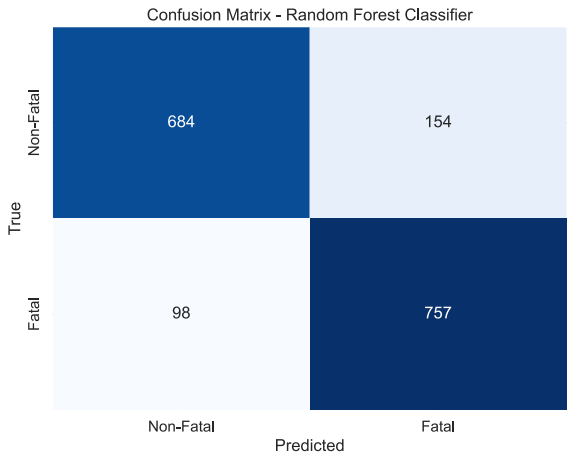

Figure 10: Confusion Matrix of the predictive model

In summary, the evaluation results validates that Random Forest model has achieved the capability to accurately flag fatalities from accident data needed to enable effective interventions and safety improvements.

4. Recommendations

From the study, key recommendations to government agencies for improving road safety are:

- Enhancing the infrastructure and traffic management during rush hours which will ease congestion and reduce risks.
- Increase in police presence and enforcement on high-accident days/times to deter violations.
- Massive public awareness campaigns on driver distraction, fatigue, and responsible conduct.
- Promote greater use of public transit to decrease traffic volume during peak hours.
- Incorporate safety considerations into urban planning through pedestrian/cyclist infrastructure.

5. Conclusion

This study provided valuable insights into road safety in Great Britain by exploring patterns and influential factors in traffic accident data. The hours and days feature revealed heightened accidents during rush hours and fatigue-related risks in the evenings and on Fridays. For motorcycles, afternoons and evenings posed greater threats, especially Fridays. Pedestrians faced risks in the midday through early evening commute periods, rising through the workweek.

The Apriori algorithm indicated strong associations between severity, and the number of casualties involved. DBSCAN revealed clusters pinpointing accident hotspots. Outlier detection flagged potential data anomalies but they do not have a significant impact on the study. A predictive model utilizing Random Forest achieved excellent performance in classifying fatal accidents. These findings and models provide actionable intelligence to guide policies and interventions for improving road safety.

Reference

Abhaya, A. and Patra, B.K., 2022. RDPOD: an unsupervised approach for outlier detection. Neural Computing and Applications, 34(2), pp.1065-1077.

Data Hull, (n.d.). How Has Hull Changed in 10 Years? [online] Available at: https://data.hull.gov.uk/how-has-hull-changed-in-10-years/ [Accessed 04 August 2023].

Department for Transport, (n.d.). Reported Road Casualties in Great Britain: Notes, Definitions, Symbols, and Conventions. [online] Available at: https://www.gov.uk/government/publications/road-accidents-and-safety-statistics-notes-and-definitions/reported-road-casualties-in-great-britain-notes-definitions-symbols-and-conventions [Accessed 04 August 2023].

Department for Transport, (n.d.). Transport Statistics Great Britain 2021. [online] Available at: https://www.gov.uk/government/statistics/transport-statistics-great-britain-2021/transport-statistics-great-britain-2021 [Accessed 04 August 2023].

Evo Magazine, 2023. ULEZ Explained: All You Need to Know About Ultra Low Emission Zones. [online] Available at: https://www.evo.co.uk/advice/202338/ulez-explained-all-you-need-to-know-about-ultra-low-emission-zones [Accessed 04 August 2023].

GOV.UK, 2023. Driving Lessons: Learning to Drive. [online] Available at: https://www.gov.uk/driving-lessons-learning-to-drive [Accessed August 2023].

Islam, M.R., Jenny, I.J., Nayon, M., Islam, M.R., Amiruzzaman, M. and Abdullah-Al-Wadud, M., 2021. Clustering Algorithms to Analyze the Road Traffic Crashes. In: 2021 International Conference on Science & Contemporary Technologies (ICSCT). 1-6. Dhaka, Bangladesh: IEEE. doi:10.1109/ICSCT53883.2021.9642542.

Lesouple, J., Baudoin, C., Spigai, M. and Tourneret, J.Y., 2021. Generalized isolation forest for anomaly detection. Pattern Recognition Letters, 149, pp.109-119.

Statista, (n.d.). Average Number of Fatalities According to Transport in the United Kingdom. [online] Available at: https://www.statista.com/statistics/300601/average-number-of-fatalities-according-to-transport-in-the-united-kingdom/ [Accessed 04 August 2023].

Verma, A. and Rashmi. 2019. Association between road accident causalities using apriori algorithm. Procedia computer science, 167, pp.715-724.