BIG DATA PROJECT

Name of student: Kelvin Idogun

Student number: 202250185

Introduction

The data set for this assignment has been transformed into an SQLite database containing tables representing information on the accident, vehicle, casualty, and the lower layer super output layer (Lsoa). A few key questions about the data set have been put forward which the analysis aims to address, chief amongst them being the sort of road safety measures that could be put in place by the government to significantly reduce the occurrence of future accidents. This report details a robust analysis of the data and gives insights into factors that lead to fatalities when an accident occurs.

Analysis

Data cleaning I:

The pandas profiling module was used to get a quick overview of the data and this exposed a key issue straight away. According to the Road Safety open dataset guide, all numbers represented in each column had specific meanings. Inputs represented as -1 indicated a missing value and the data set had large numbers of -1 values across all three data frames. In order to retain as much data as possible, it was decided to not drop these records from the data set instead replacing them with values. There are multiple methods of doing this such as using the median or mode data points, but those methods posed the risk of introducing bias into the dataset. It was decided to use an experimental method from the scikit learn library called the iterative imputer which tries to predict the missing values by utilizing the data available in other features in an iterated round-robin fashion (Learn, 2023). Applying this experimental method required ascertaining the structure of missingness from the data, if the data were missing at random, another method such as the KNN imputation could have been adopted but it was discovered that there was a structure to the missing data. The lsoa_of_accident location column which was of type object had a few -1 values as well, these were replaced with the unknown string, and the location_easting_osgr', 'location_northing_osgr', 'longitude', 'latitude' columns that had to do with coordinates had 14 missing values for each of the columns, these rows had to be dropped.

Having completed the first stage of the cleaning, the code labels from the road safety open dataset were mapped to the respective codes for each column in all three tables. This allowed for better visualizations of the analysis to the questions posed.

Discussing the 'When'

Significant hours of the day, and days of the week, on which accidents occur.

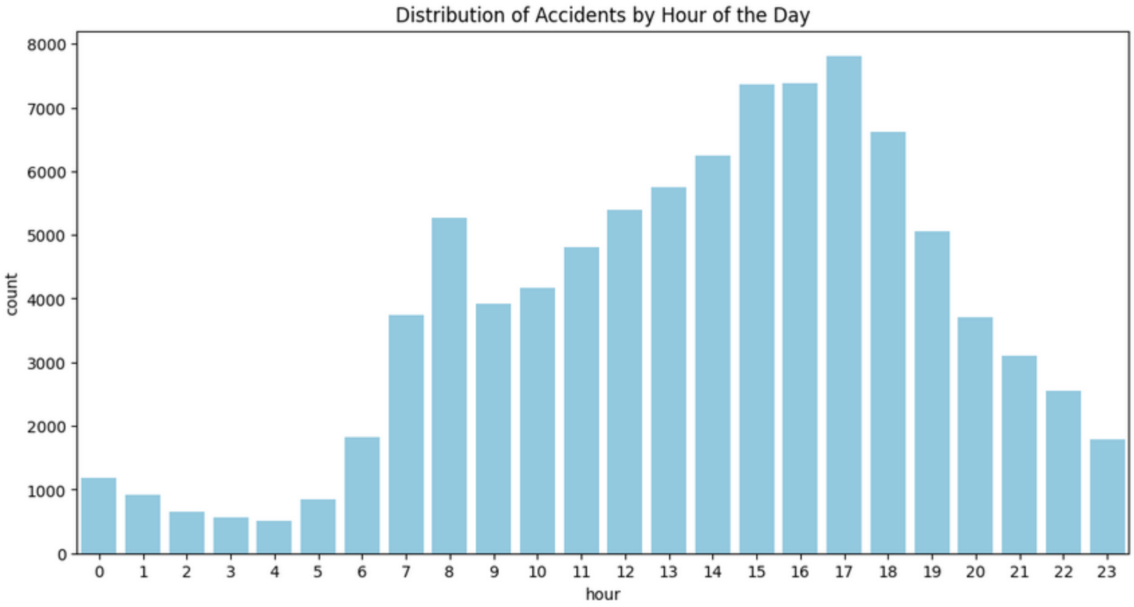


Fig 1: Number of accidents by the hour.

Figure 1 above shows the distribution of accidents over a 24-hour period. As seen from the chart, there seems to be a peak in accidents towards the evening time of the day. Visualizing the days these occur would make it clearer.

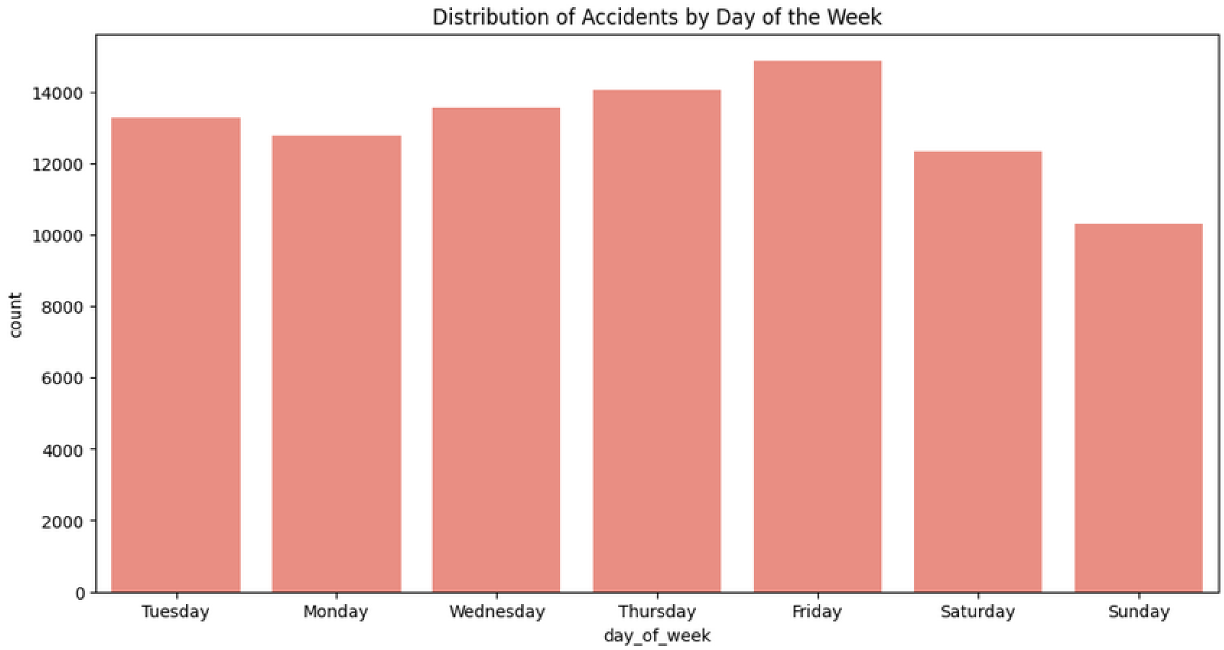


Fig 2: Number of accidents per day

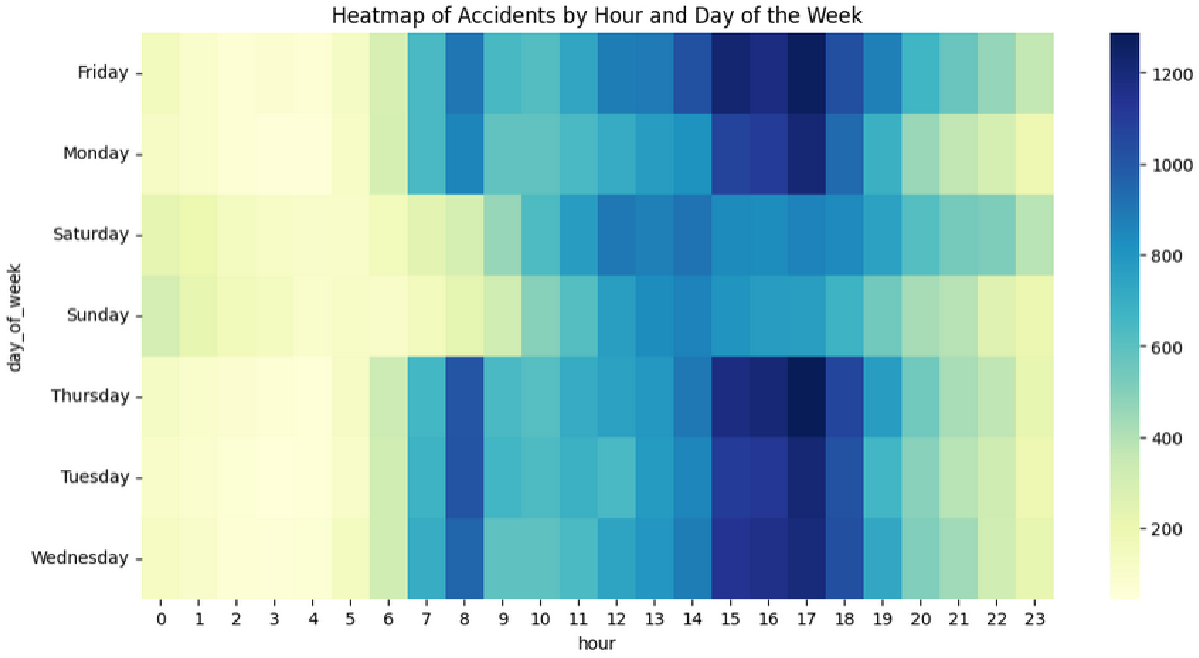


Fig 3: Accident per hour and day

Figures 2 and 3 clearly show that these accidents occur mostly during the weekdays and as seen from the heatmap in Figure 3, these peak periods during the weekdays which are often towards the close of business could explain that trend as workers tend to close at those periods and the roads would typically be congested.

With Motorbikes in focus, what significant hours and days do accidents occur?

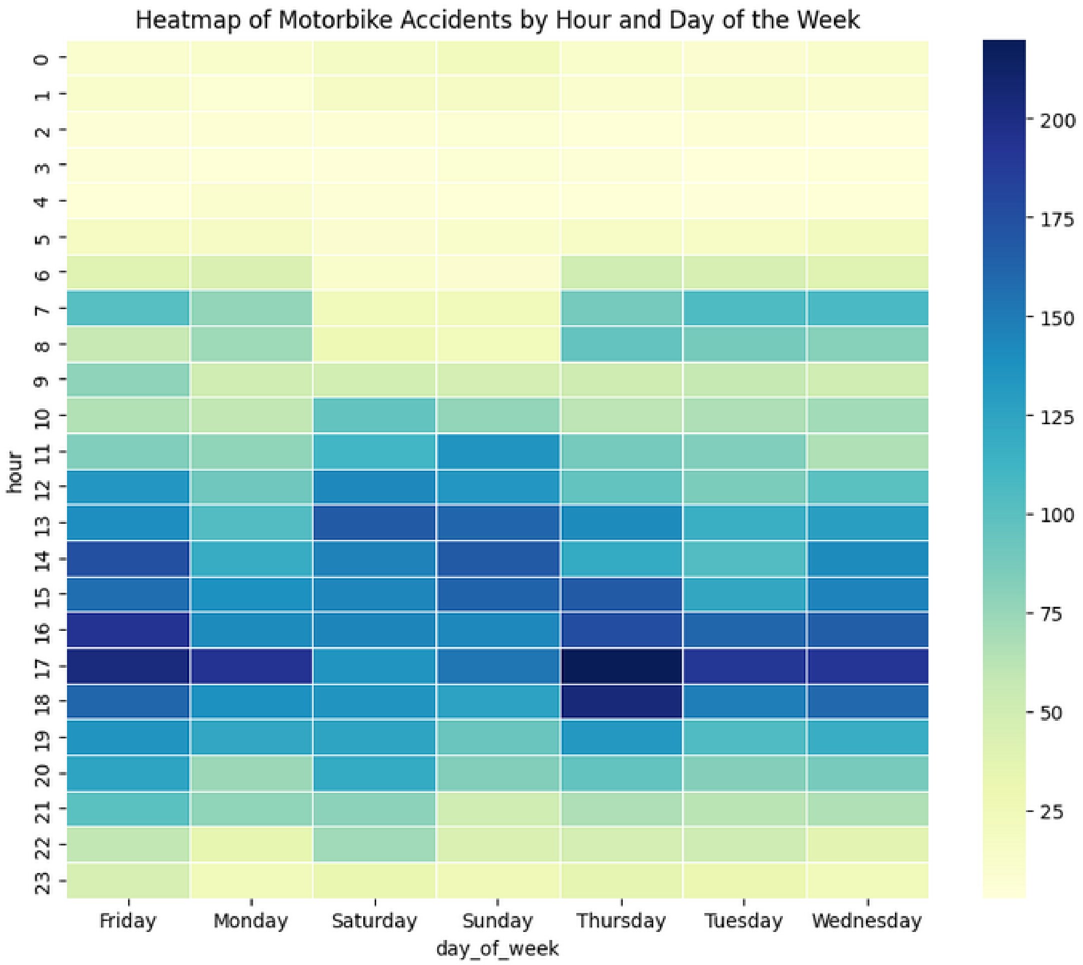


Fig 4: Motorbike Accident Heatmap

As seen from Figure 4, It was clear that a significant number of motorcycle accidents also occurred during the weekdays as well with peak numbers experienced on Mondays, Thursdays, and Fridays at the close of business (17:00 Hours). This could also be explained by the rush hour period experienced by workers eager to return home for the weekend. Also, Mondays are notoriously hectic days of the week, and it could be inferred that worker frustrations are usually transformed into road rage making them more impatient and increasing the likelihood of an accident occurring. Fatigue could also be a reason, according to (NHTSA, 2023), Drowsy driving crashes also occur late in the afternoons.

For pedestrians involved in accidents, are there significant hours of the day, and days of the week, on which they are more likely to be involved?

Using the same approach employed for the motorcycle data, the casualty type feature of the casualty table was filtered and joined to the accident table on the foreign key. The result is shown in figure 5 below.

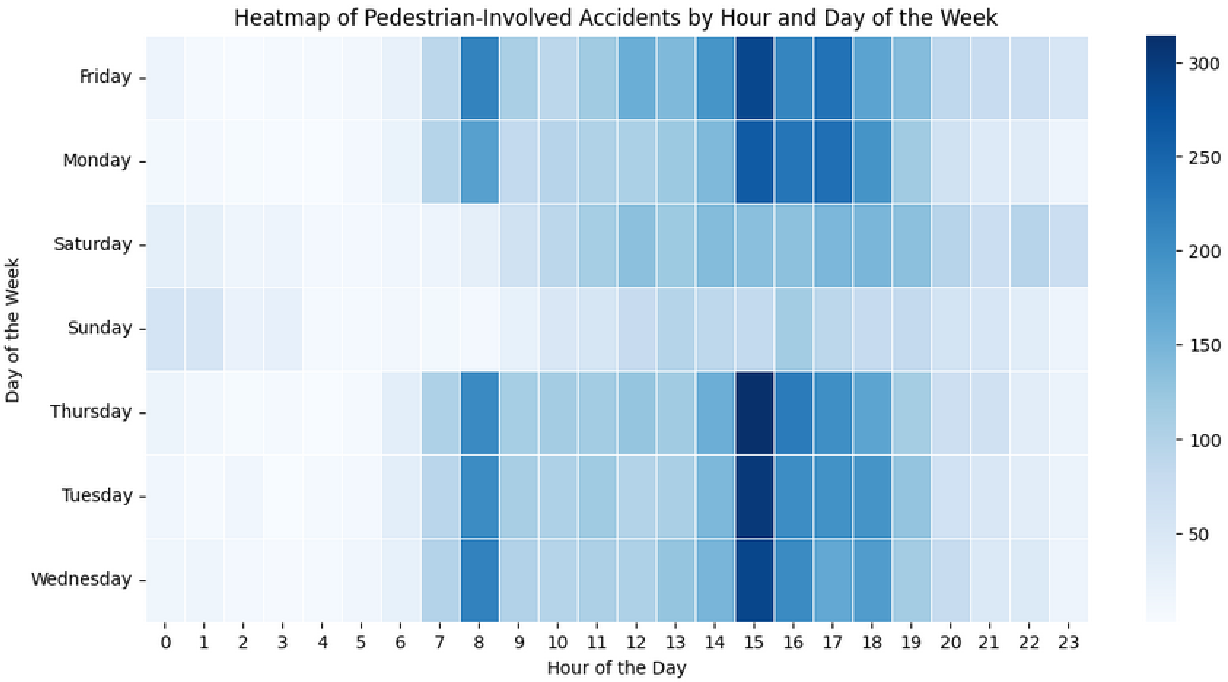


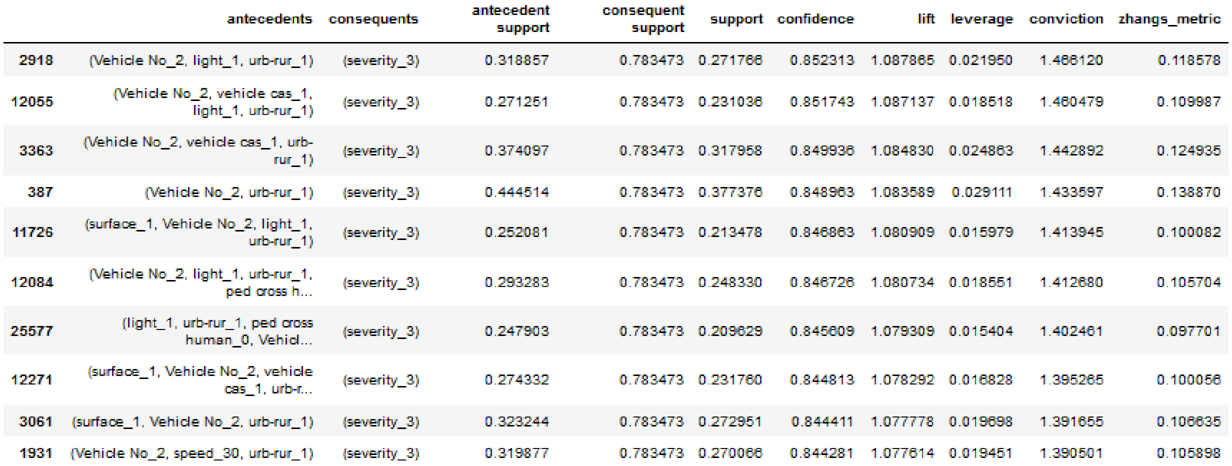
Fig 5: Pedestrian accident heatmap

The visualization shows pedestrian accidents between the range of 200-300 occurring at 3:00 pm during the weekdays and much less during the weekend as fewer people are active and about during those periods.

Discussing the conditions accidents happen.

Using the apriori algorithm, explore the impact of selected variables on accident severity.

To understand the conditions under which accidents occurred, the apriori algorithm was applied to the data set which is used to understand association patterns between features in a data set. The mlxtend library was used which helps determine important features like the Antecedents, consequents (A sort of cause-and-effect measurement), support, confidence, lift, leverage, conviction, and Zhang metrics. To achieve this, selected columns in the accident data frame without the labels where one hot encoded because the binary nature of one got encoded data, 1's and 0's is a more suitable format for the Apriori algorithm. These one-hot encoded data were concatenated and passed through the algorithm with a minimum support of 0.2. In this case, the aim was to find the top 10 rules that had a consequent of the one hot encoded accident severity columns with a lift greater than 1, confidence greater than 0.2 and support greater than 0.2. Figure 6 below shows a snapshot of the findings.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
2918	(Vehicle No_2, light_1, urb-rur_1)	(severity_3)	0.318857	0.783473	0.271766	0.852313	1.087865	0.021950	1.466120	0.118578
12055	(Vehicle No_2, vehicle cas_1, light_1, urb-rur_1)	(severity_3)	0.271251	0.783473	0.231036	0.851743	1.087137	0.018518	1.460479	0.109987
3363	(Vehicle No_2, vehicle cas_1, urb-rur_1)	(severity_3)	0.374097	0.783473	0.317958	0.849936	1.084830	0.024863	1.442892	0.124935
387	(Vehicle No_2, urb-rur_1)	(severity_3)	0.444514	0.783473	0.377376	0.848963	1.083589	0.029111	1.433597	0.138870
11726	(surface_1, Vehicle No_2, light_1, urb-rur_1)	(severity_3)	0.252081	0.783473	0.213478	0.846863	1.080909	0.015979	1.413945	0.100082
12084	(Vehicle No_2, light_1, urb-rur_1, ped cross h...	(severity_3)	0.293283	0.783473	0.248330	0.846726	1.080734	0.018551	1.412680	0.105704
25577	(light_1, urb-rur_1, ped cross human_0, Vehicl...	(severity_3)	0.247903	0.783473	0.209629	0.845609	1.079309	0.015404	1.402461	0.097701
12271	(surface_1, Vehicle No_2, vehicle cas_1, urb-r...	(severity_3)	0.274332	0.783473	0.231760	0.844813	1.078292	0.016828	1.395265	0.100056
3061	(surface_1, Vehicle No_2, urb-rur_1)	(severity_3)	0.323244	0.783473	0.272951	0.844411	1.077778	0.019698	1.391655	0.106635
1931	(Vehicle No_2, speed_30, urb-rur_1)	(severity_3)	0.319877	0.783473	0.270066	0.844281	1.077614	0.019451	1.390501	0.105898

Fig 6: Association Patterns

This same approach was repeated using features from the Vehicle and casualty data frames with the accident severity in focus. It was observed that features from the accident data frame had a much higher direct causal effect on the accident severity. Features like the number of vehicles involved in an accident, the lighting conditions, and the area had good values of confidence and lift but were not very much supported throughout the dataset. Please see the code analysis for a more detailed outlook.

Discussing the where

Identifying the accidents in Kingston upon Hull, Humberside, and the East Riding of Yorkshire region using clustering analysis.

The DBSCAN algorithm was used for this analysis, the Lsoa data frame which was largely untouched so far was queried using SQL commands, joining the data frame with the accident data table on the lsoa of accident location which on close inspection, had similar inputs with the lsoa01nm column of the LSOA table. Only data that had the same lsoa names for both tables of the specific locations were pulled together to form a new data frame containing the longitudes and latitudes of the accidents in the regions and their respective lsoa name. According to (Wikipedia, 2023), Humberside was abolished on 1 April 1996 and split up into Kingston upon hull, North Lincolnshire, Northeast Lincolnshire, and East Riding of the Yorkshire. This informed the locations chosen during filtering. The epsilon value for the dbscan was chosen by computing the nearest neighbor for each point and plotting the distance before selecting the elbow point of the graph.

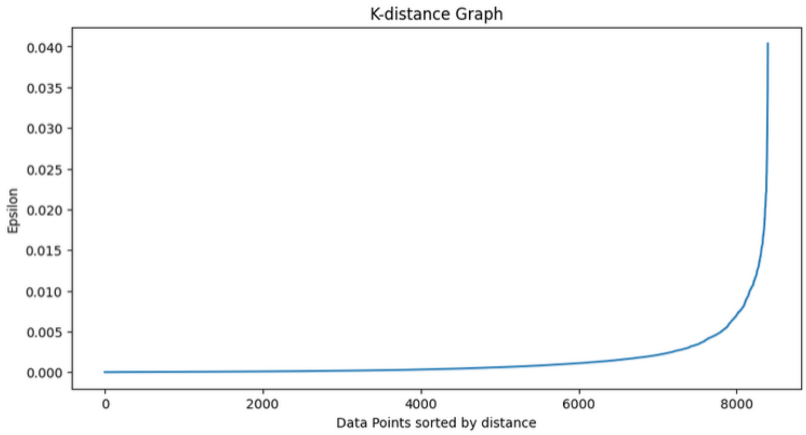


Fig 7: Optimal Epsilon value.

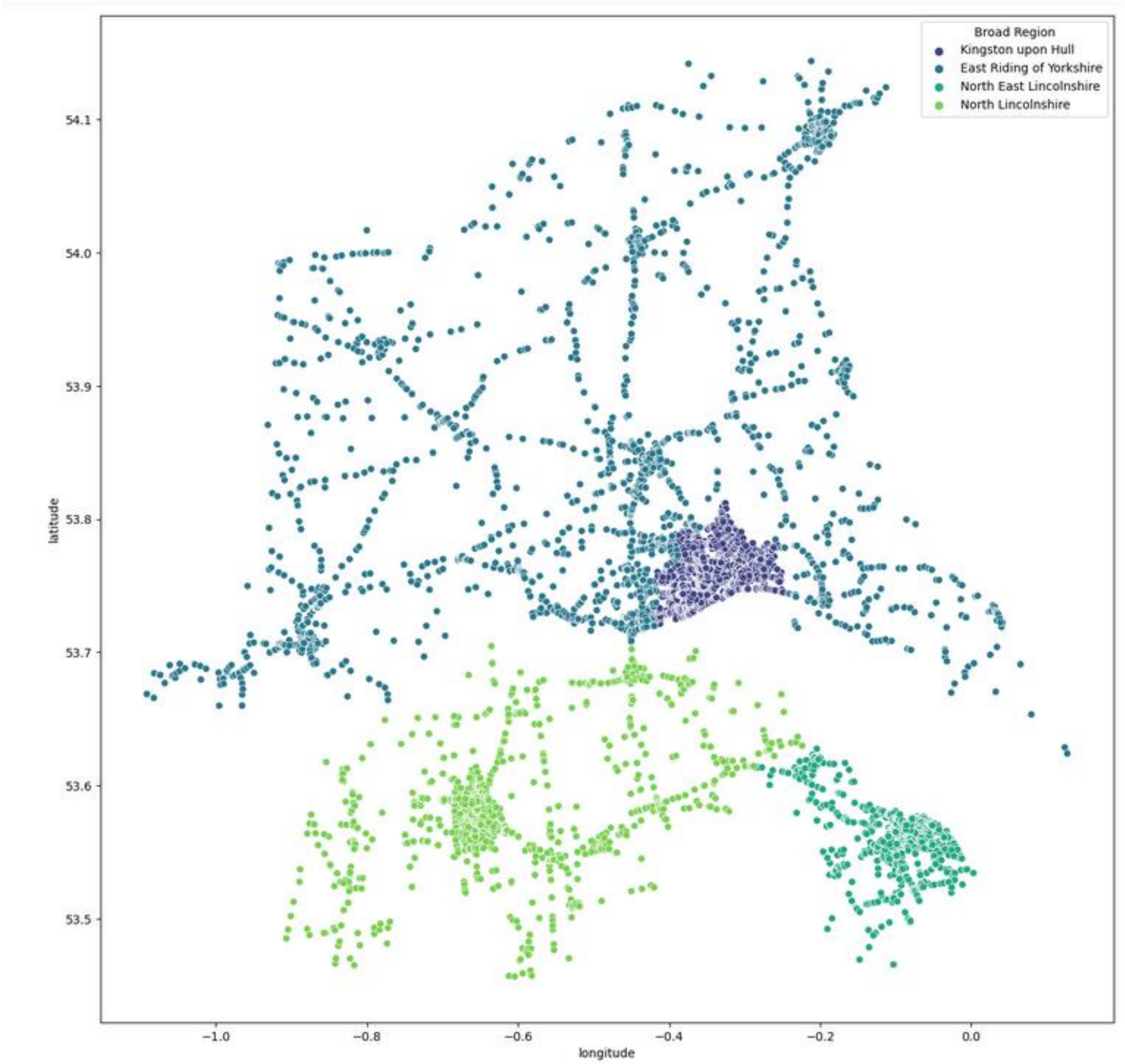
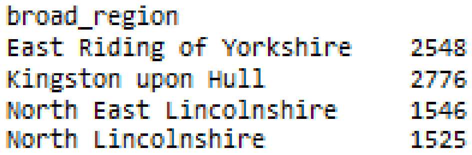


Fig 8: Cluster Map

The cluster map plotted afterward revealed the distribution of accidents across the regions with Kingston upon Hull region having the largest number of accidents.



```
broad_region
East Riding of Yorkshire    2548
Kingston upon Hull          2776
North East Lincolnshire     1546
North Lincolnshire          1525
```

Fig 9: Accident spread

Data Cleaning II

Using the Multiple IQR method for outlier detection, several outliers were detected. However, on closer inspection, most of them proved to be actual values. The columns were mostly heavily skewed to a particular input, and these were the ones highlighted by the Multiple IQR method. A few columns which were likely candidates for real outliers were made the focus of this analysis.

Starting with the age of the driver column in the vehicle data frame.

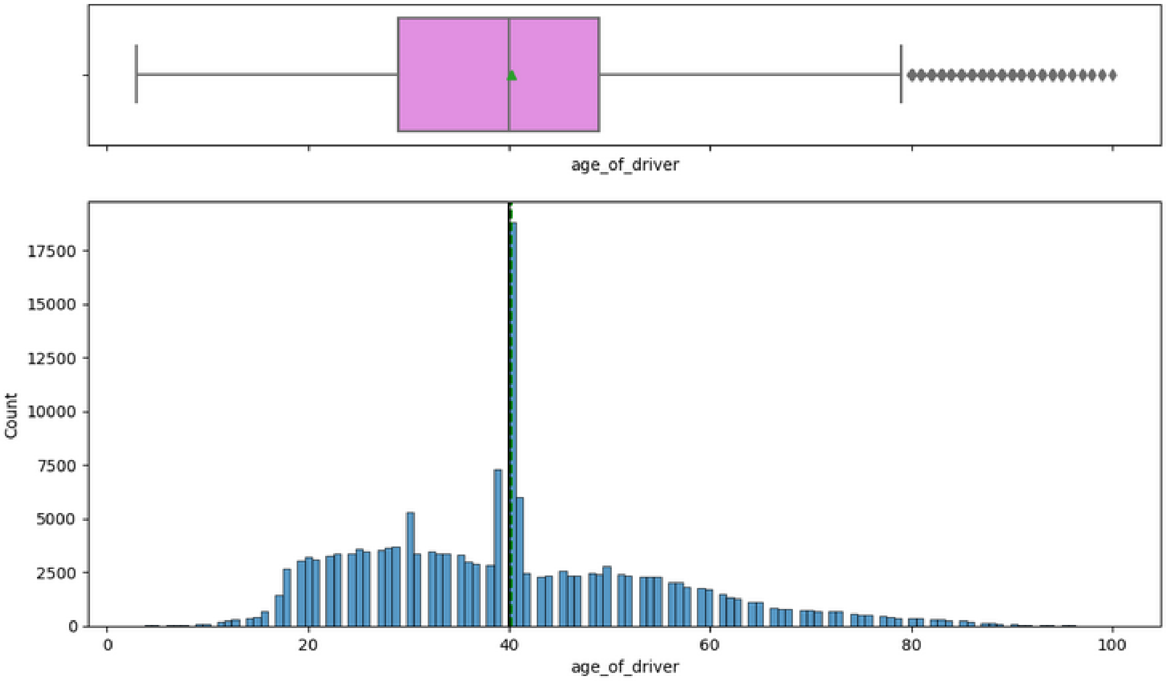


Fig 9: Age of Driver Box-Hist Plot

As seen from Figure 9, there were some unrealistic ages discovered after plotting the data. According to the UK government (Govt, 2023), the minimum age allowed by law to drive was 16 years, the plot showed otherwise. The IQR test seemed to only consider ages above 80 as outliers but there is no limit to the age a person can drive and this was made evident by the late Queen of England who often liked to drive herself in her jaguar, so these were very likely occurrences. For this reason, it was decided to replace rows that contained ages less than the government-approved age of 16 with 16. The other column of focus was the age of the vehicle column.

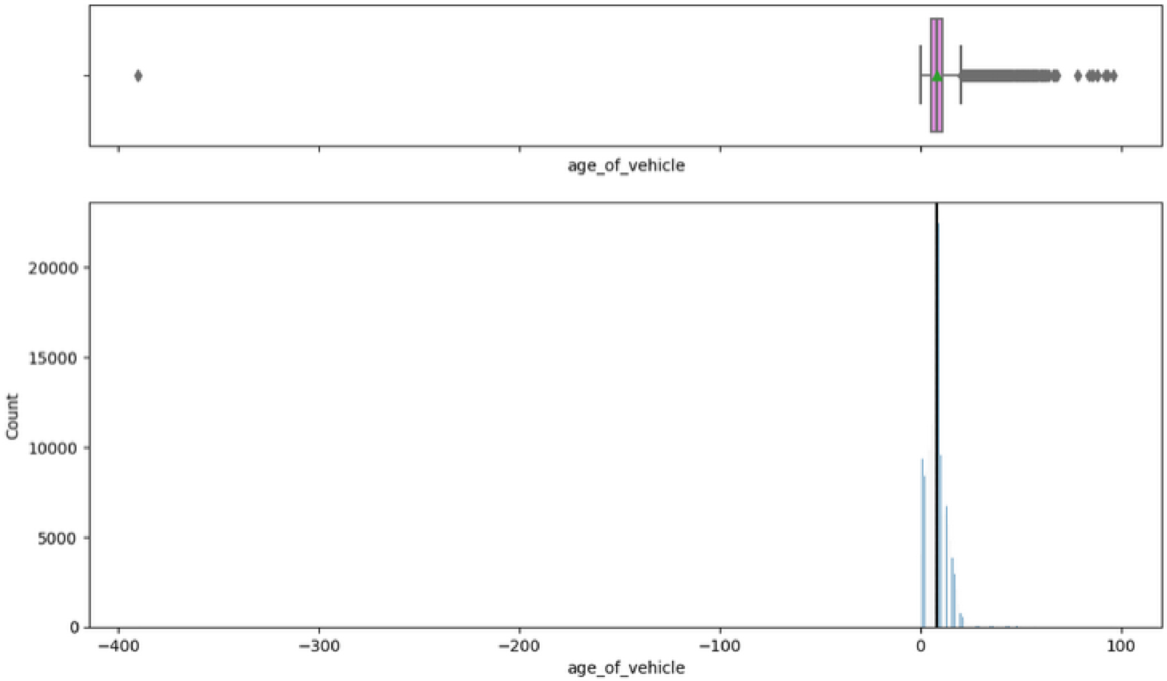


Fig 10: Age of Vehicle

It was obvious that there was an incorrect entry from this plot, there was a negative value included, Also the average age for a car at scrappage is 13.9 years (SMMT, 2023). It was decided to replace all inputs below 0 and above 14 with the boundary ages. A similar analysis was carried out for the engine capacity column.

Fatal Accident Severity Prediction Model

It was decided to apply tree-based models to the data set due to their ability to handle non-linear relationships. In this case, Random Forest and gradient-boosting models were employed. After merging the accident, vehicle, and casualty data frames on the foreign key, the features required for the model were selected and saved to a variable name. It was discovered that the accident severity feature was imbalanced and the smote resampling technique was used to correct it. After using a 70:30 train test split, the feature importance was plotted to show which features had more impact on the prediction model.

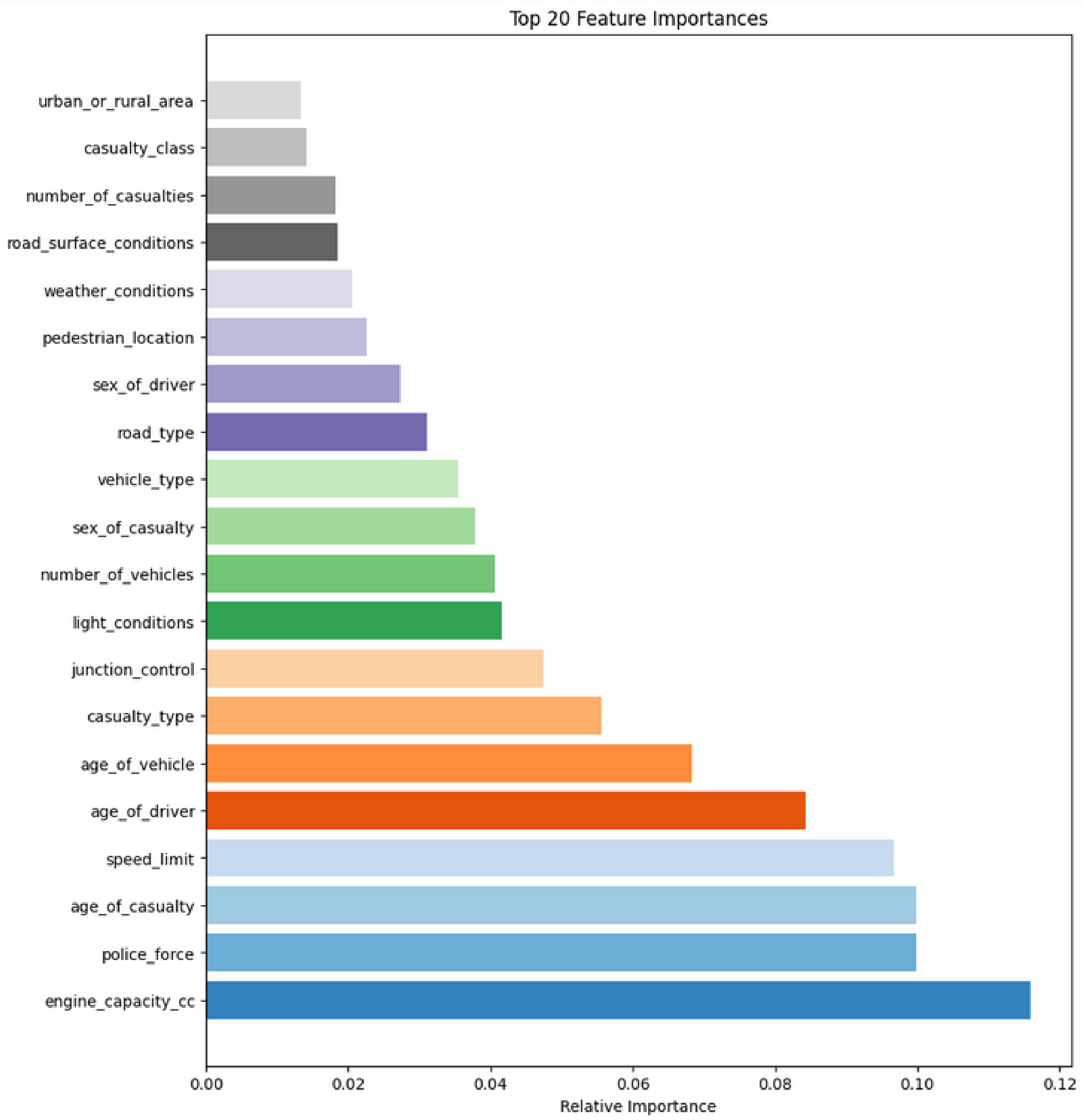
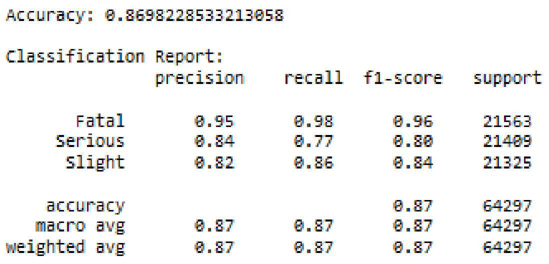


Fig 11: Random Forest feature importance

Model results



```
Accuracy: 0.8698228533213058

Classification Report:
               precision    recall  f1-score   support

       Fatal       0.95      0.98      0.96     21563
     Serious       0.84      0.77      0.80     21409
      Slight       0.82      0.86      0.84     21325

    accuracy                           0.87     64297
   macro avg       0.87      0.87      0.87     64297
weighted avg       0.87      0.87      0.87     64297
```

Fig 12: Random Forest classification report.

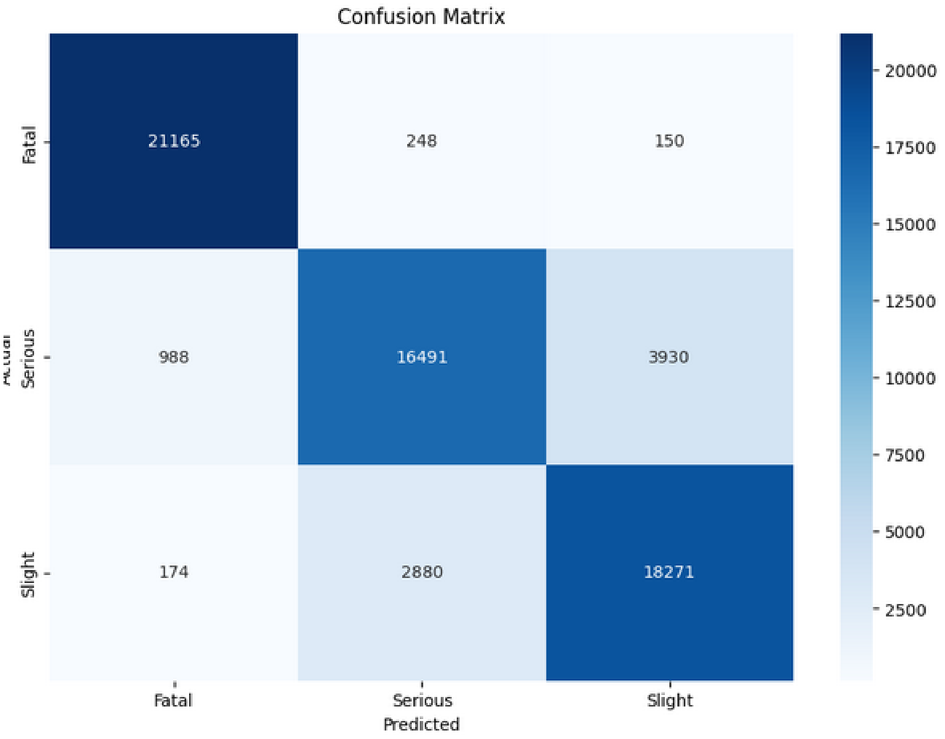
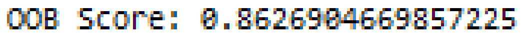


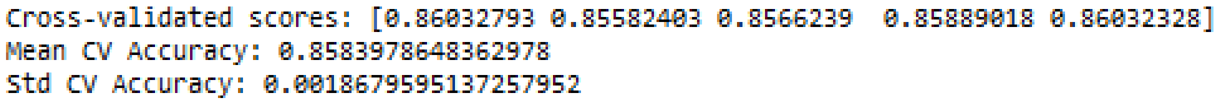
Fig 13: Random Forest confusion matrix

As seen from the classification report produced, the model performed quite well with good values of precision and recall for all targets predicted and an 87% prediction accuracy. The model was also evaluated for overfitting using the cross-validation technique and the OOB score, an inbuilt overfitting checker in the random forest model. The values produced were in the range of those produced in the classification report.



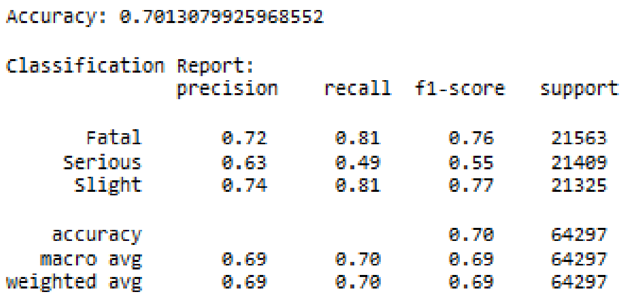
```
OOB Score: 0.8626904669857225
```

,



```
Cross-validated scores: [0.86032793 0.85582403 0.8566239  0.85889018 0.86032328]
Mean CV Accuracy: 0.8583978648362978
Std CV Accuracy: 0.0018679595137257952
```

The other model employed was the gradient boosting model which did not perform as well as the random forest as seen in the classification report in fig 14 below.



```
Accuracy: 0.7013079925968552

Classification Report:
              precision    recall  f1-score   support

       Fatal       0.72      0.81      0.76     21563
     Serious       0.63      0.49      0.55     21409
      Slight       0.74      0.81      0.77     21325

    accuracy                           0.70     64297
   macro avg       0.69      0.70      0.69     64297
weighted avg       0.69      0.70      0.69     64297
```

Fig 14: Gradient boosting classification report.

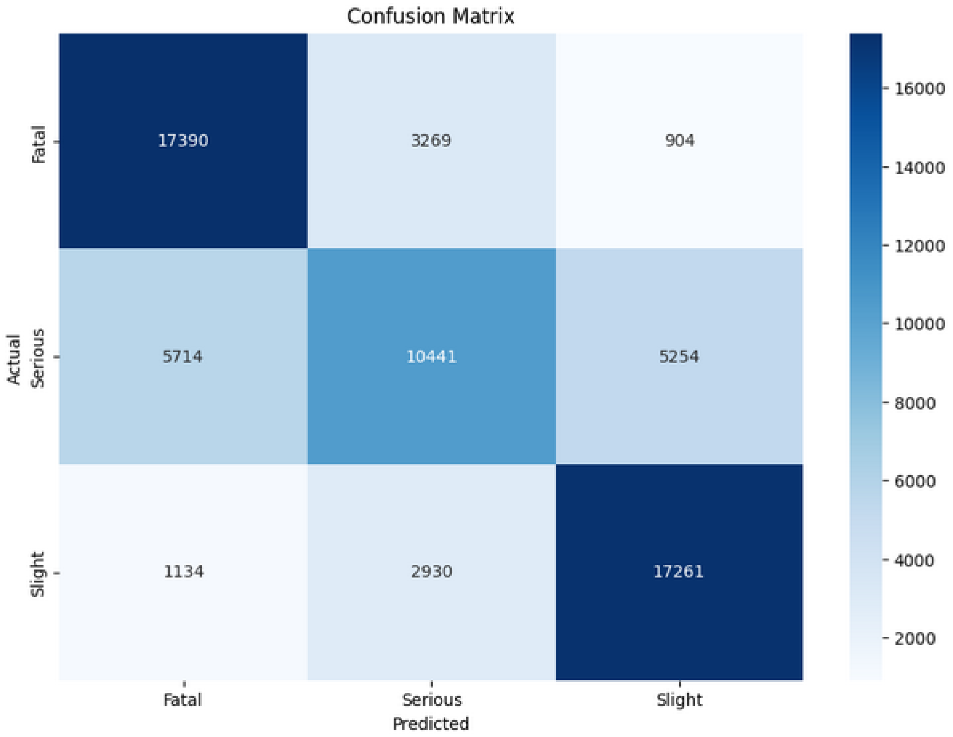


Fig 15: Gradient boosting confusion matrix.

The confusion matrix simply represents the number of True positives, false positives, true negatives, and false negatives predicted for each target category.

Recommendations:

Based on the feature importance:

- Vehicles with larger or smaller engines have different risk profiles. Regulations around licensing or additional training for driving certain types of vehicles could be implemented.
- Targeted safety campaigns or training programs for the age groups frequently involved in accidents could help in reducing the numbers.
- In relation to the age of the vehicles, regular vehicle inspections and encouraging timely maintenance can help reduce accidents caused by vehicle mishaps.
- Ensuring consistency in the effectiveness of police forces in enforcing traffic rules and safety measures across all regions.

References

Govt, U., 2023. *Learning to Drive.* [Online]
Available at: https://www.gov.uk/driving-lessons-learning-to-drive
[Accessed 3 August 2023].

Learn, S., 2023. *Imputation of missing values.* [Online]
Available at: https://scikit-learn.org/stable/modules/impute.html#iterative-imputer
[Accessed 12 August 2023].

NHTSA, 2023. *Drowsy Driving.* [Online]
Available at: https://www.nhtsa.gov/risky-driving/drowsy-driving
[Accessed 12 August 2023].

SMMT, 2023. *Average Vehicle Age.* [Online]
Available at: https://www.smmt.co.uk/industry-topics/sustainability/average-vehicle-age/
[Accessed 6 August 2023].

Wikipedia, 2023. *Humberside.* [Online]
Available at: https://en.wikipedia.org/wiki/Humberside
[Accessed 12 August 2023].