Accident 2020 Data Report

Introduction

Public safety is in danger due to traffic accidents, which are a severe problem. According to data gathered in the UK on average, five people die every day on the road in the UK and 84 are seriously injured (Brake, 2022). This is cause for concern, and analysis is needed to identify the trends in the incidents and derive actionable recommendations. This can aid the authority in enhancing safety measures and reducing future incidents.

We can apply and build remedies to prevent future accidents once we've located and recognised the reasons for this disaster. Important tables from the accident 2020 dataset in our database, include the accident, vehicle, casualty, and Lsoa tables, which contain significant data that were used for this analysis.

Hourly Analysis

The analysis of the data revealed that the biggest number of accidents occur during the day at 17:00 (5:00 PM). Figure 1 below illustrates that 862 accidents were reported in total during this hour, indicating a potential focus for concerns about traffic safety. The Top 20 accidents were chosen and displayed to provide significant insight. The accident rate was comparatively low between 8:00 and 11:00, indicating safer driving conditions during this time. Additionally, figure 1 reveals that accidents tend to increase right after work hours, possibly related to rush hour and exhaustion.

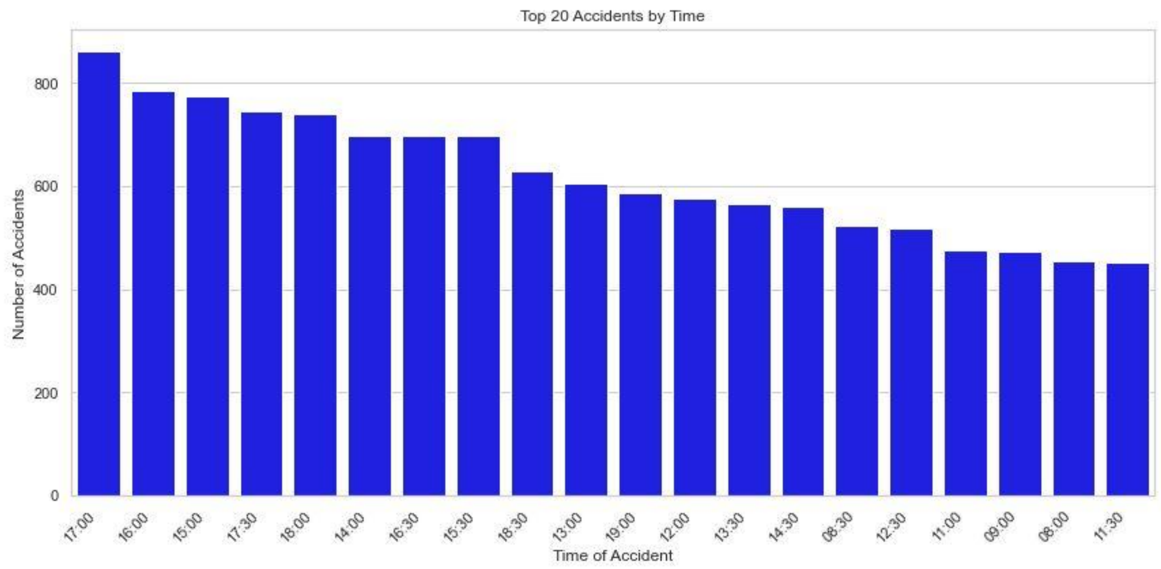


Figure 1

According to Figure 1, the number of accidents increases in the afternoon and evening, with few accidents occurring in the morning. The hours between 15:00 and 18:00 have the highest number of accidents, whereas the hours between 8:00 and 11:00 have the lowest number of accidents.

Weekly Analysis

The insights into the number of accidents that happen each week are displayed in Figure 2 below. According to the plot, Friday had the most accidents with a total of 14, 889. When compared to the weekdays, Saturday and Sunday have a greater average. These findings demonstrate that Friday presents significantly higher accident risks.

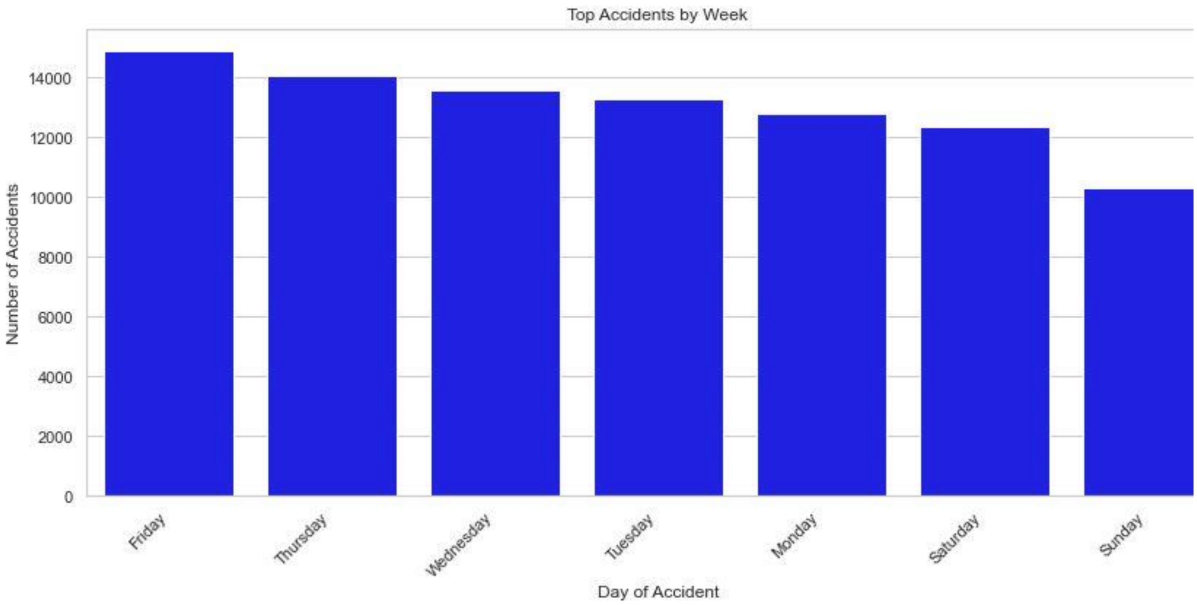


Figure 2

Motorbikes accident time of the day

Motorbike accidents are more common in the afternoon and evening. Figure 3 shows that the hours between 15:00 and 18:00 had the most motorbike accidents, with 17:00 having the highest number of motorbike accidents involving 142 victims. Furthermore, there are fewer motorbike accidents between the hours of 8:00 and 11:00.

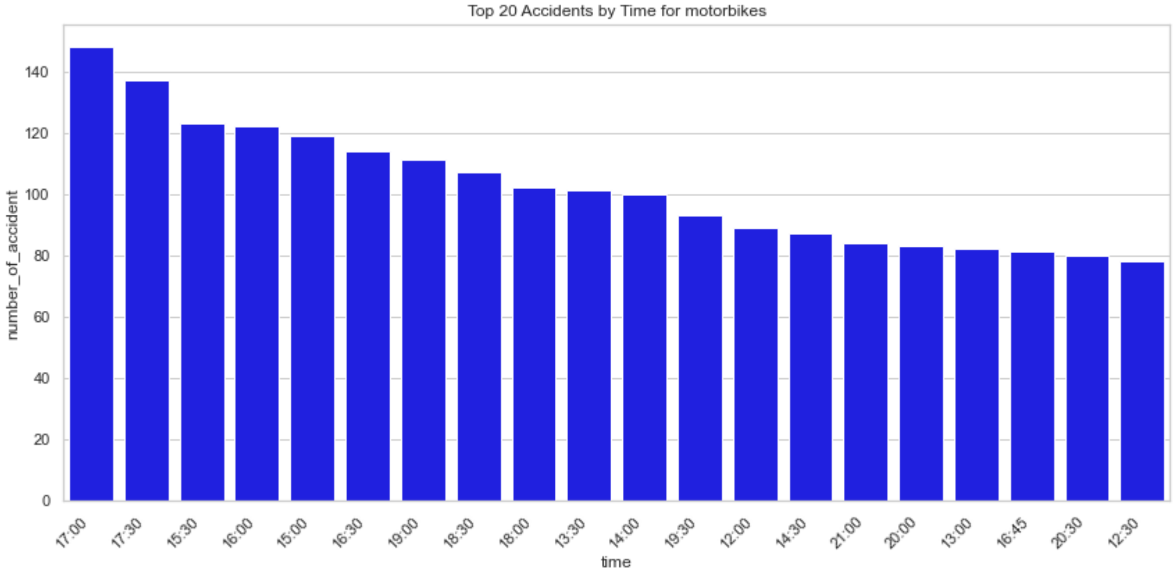


Figure 3

Motorbikes Accidents by Day of the Week

The plot below shows a significant increase in motorbike accidents on Fridays, with a total of 2308 accidents recorded; however, the records also show a relative decrease in motorbike accidents on Monday and Sunday, with 1868 and 1841 accidents, respectively. The plot shows that motorbikes have the highest number of accident on Friday with 2308 accident while the accident is low on Monday and Sunday respectively with 1868 and 1841.

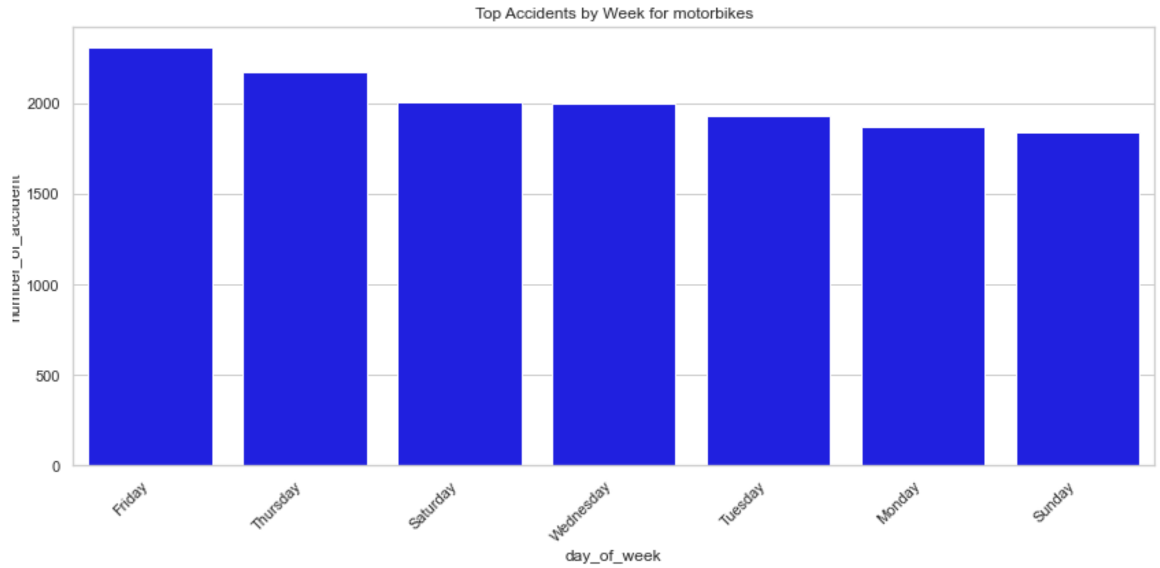
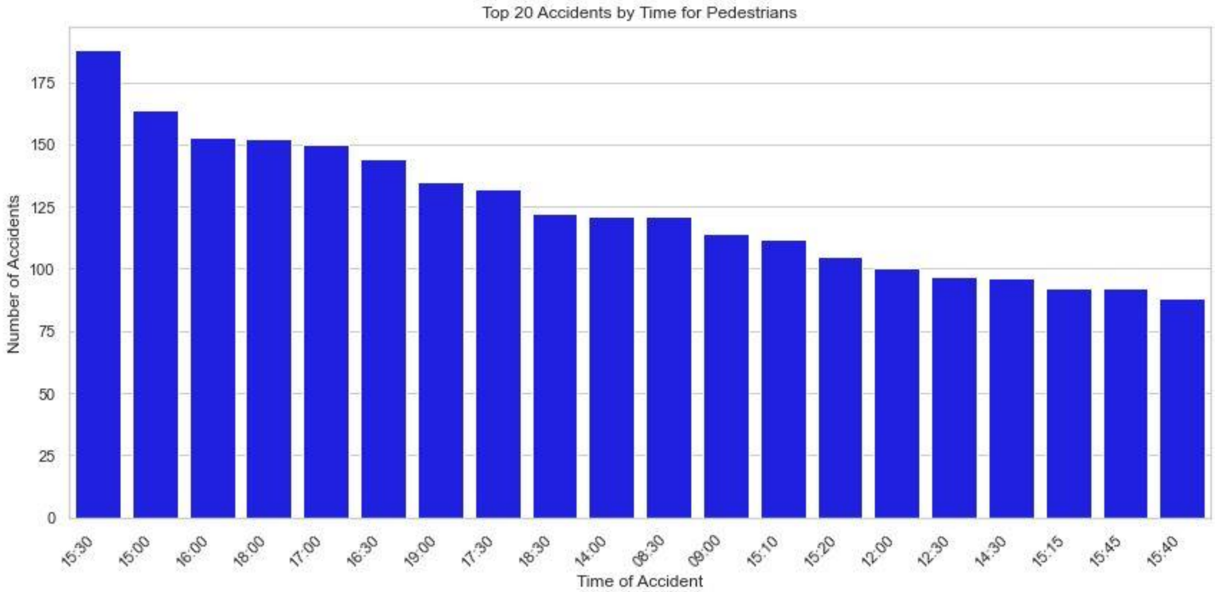


Figure 4

Pedestrian's accidents by hours of the day:

The bar graph below is labelled "Top 20 Pedestrian Accidents by Time." It shows how frequently accidents involving pedestrians occur throughout the day. The highest peak, with a total of 188 incidents, is at 15:30. According to the bar graph's findings, there is a significant drop in pedestrian accidents around 15:40, followed by a rise at 15:45. Additionally, there is a consistent increase in accidents between 16:00 and 18:00.



Pedestrians by days of the week:

According to figure 5 below, with a total of 2543 incidents, Friday has the most pedestrian accidents of any day of the week. Insights about Thursday and Tuesday are also included in Figure 5, which are close to Friday in terms of the number of accidents. Sunday has the lowest pedestrian accidents among the weekdays, which may indicate that it is a relatively secure day of the week.

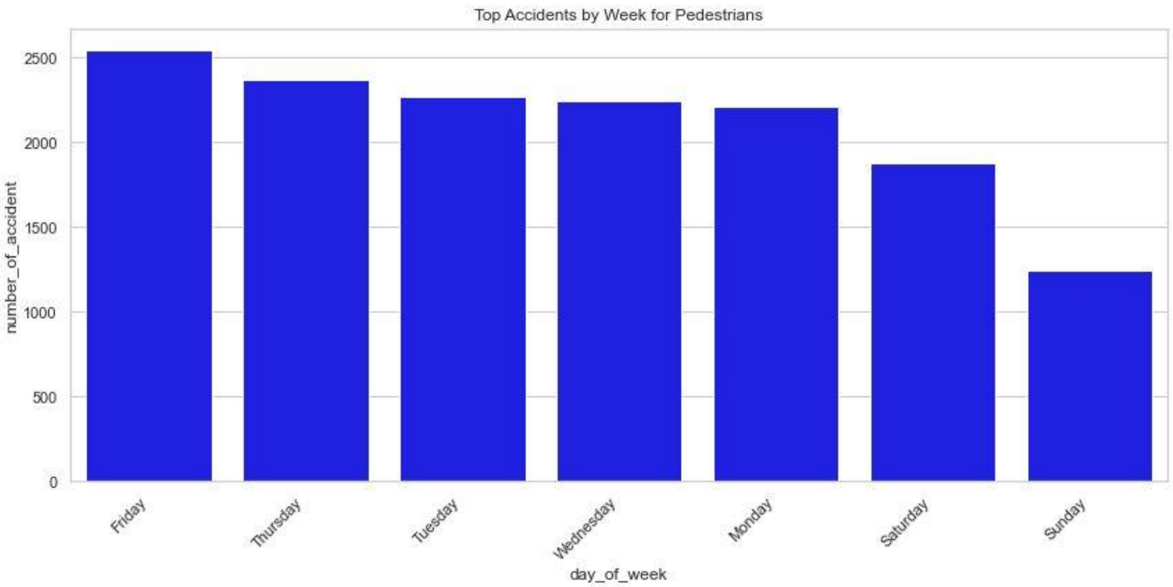


Figure 5

Accidents by Month

In 2020, January had the most accidents, according to our study, as seen by the bar graph below. The month of April in 2020 had the fewest accidents, with fluctuations in accident frequency from August to October.

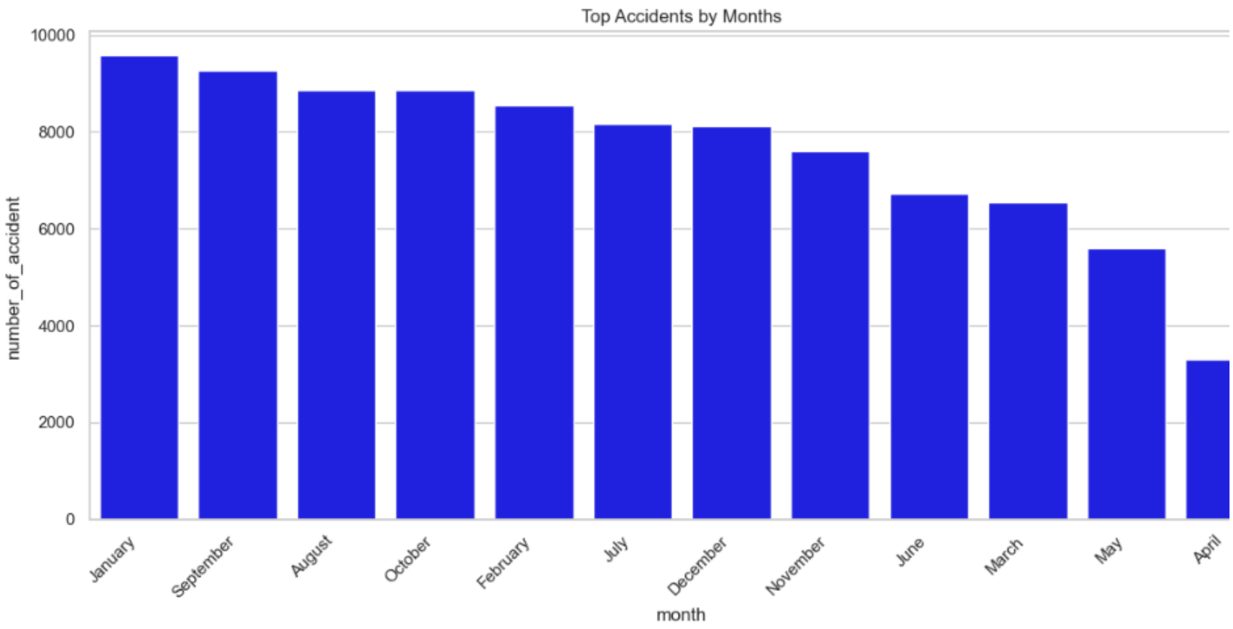
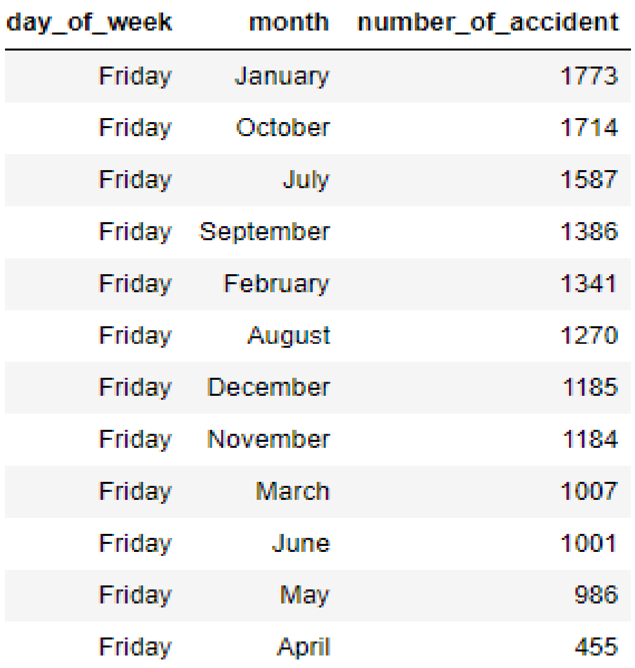


Figure 6

Exploring the Friday accidents

The Figure 7 below displays the total number of accidents that occur on Friday of each month in 2020 since our data has been revealing a high frequency of accidents on Friday.



day_of_week	month	number_of_accident
Friday	January	1773
Friday	October	1714
Friday	July	1587
Friday	September	1386
Friday	February	1341
Friday	August	1270
Friday	December	1185
Friday	November	1184
Friday	March	1007
Friday	June	1001
Friday	May	986
Friday	April	455

Figure 7

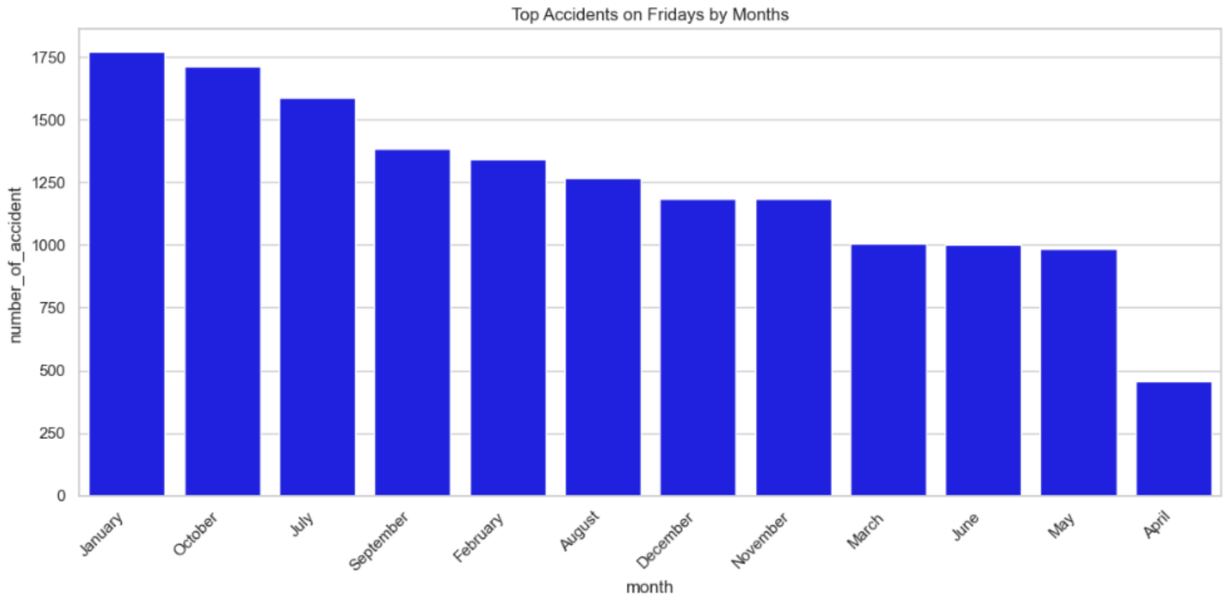


Figure 8

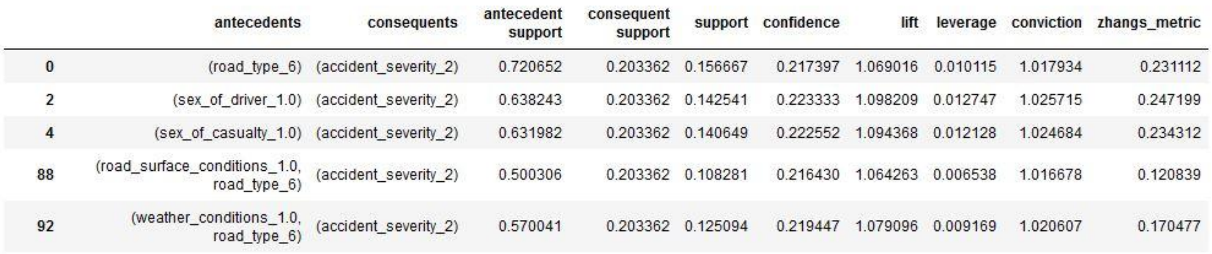
Figure 8 above displays a graph of accidents that occur on Fridays in every month in 2020, with January leading the way.

Apriori Algorithm

By selecting features from the accident 2020 data and analysing them using categorical variables, the Apriori algorithm was examined. The features where the data had an incorrect input (-1) were cleaned up so to retain the data's integrity and maintain the data's association, the incorrect input has been changed out for the mode with the highest incidence. Additionally, a dummies one hot encoding was used on the features. The analysis thus demonstrates how many factors are connected to various degrees of accident severity.

Accident severity 2, which denotes a slight accident, correlates with Road Type 6 with a lift value of 1.07, showing a very low impact on accident occurrences, and with a support value of 15.67%, denoting that there are 15.67% chances of the two features occurring simultaneously. Additionally, accident in weather condition 1 has a lift value of 1.07, which denotes that there is a negligible influence, a confidence value of 21.94%, which denotes a weak association between the two qualities, and a support value of 12.51, which denotes that 12.51 percent of respondents agreed with the rule.   The driver's gender (sex of driver 1) has a lift value of 1.10 from the algorithm, indicating a moderate impact on the risk of an accident occurring. The relationship has a confidence and support value of 22.3% and 14.25%, respectively.

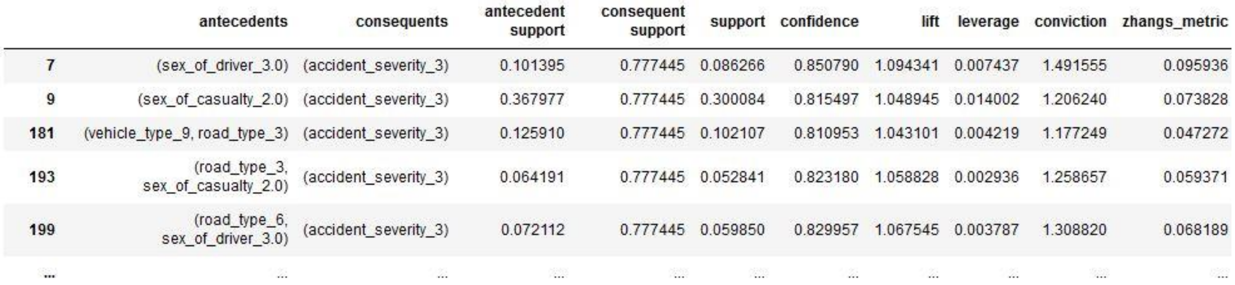
Table 1



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(road_type_6)	(accident_severity_2)	0.720652	0.203362	0.156667	0.217397	1.069016	0.010115	1.017934	0.231112
2	(sex_of_driver_1.0)	(accident_severity_2)	0.638243	0.203362	0.142541	0.223333	1.098209	0.012747	1.025715	0.247199
4	(sex_of_casualty_1.0)	(accident_severity_2)	0.631982	0.203362	0.140649	0.222552	1.094368	0.012128	1.024684	0.234312
88	(road_surface_conditions_1.0, road_type_6)	(accident_severity_2)	0.500306	0.203362	0.108281	0.216430	1.064263	0.006538	1.016678	0.120839
92	(weather_conditions_1.0, road_type_6)	(accident_severity_2)	0.570041	0.203362	0.125094	0.219447	1.079096	0.009169	1.020607	0.170477

A link between accident severity 3 (which indicates a serious accident) and sex_of_driver_3 (an unknown sex of the driver) has a lift value of 1.09, which indicates a moderate relationship. With a support level of 8.63% and a confidence level of 8.51%, this association has an 8.63% probability. Additionally, vehicle_type_9 and road_type_3 indicate a lift value of 1.04 that also points to a moderate association between the accident severity_3; the confidence value is 8.63%, and the support is 10.21%, both of which point to the accident pattern being 10.21%. From the Apriori Algorithm, the analysis association rules show insight into the potential factors contributing to accident severity.

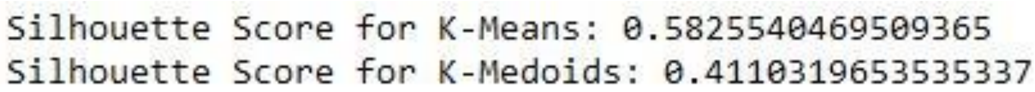
Table 2



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
7	(sex_of_driver_3.0)	(accident_severity_3)	0.101395	0.777445	0.086266	0.850790	1.094341	0.007437	1.491555	0.095936
9	(sex_of_casualty_2.0)	(accident_severity_3)	0.367977	0.777445	0.300084	0.815497	1.048945	0.014002	1.206240	0.073828
181	(vehicle_type_9, road_type_3)	(accident_severity_3)	0.125910	0.777445	0.102107	0.810953	1.043101	0.004219	1.177249	0.047272
193	(road_type_3, sex_of_casualty_2.0)	(accident_severity_3)	0.064191	0.777445	0.052841	0.823180	1.058828	0.002936	1.258657	0.059371
199	(road_type_6, sex_of_driver_3.0)	(accident_severity_3)	0.072112	0.777445	0.059850	0.829957	1.067545	0.003787	1.308820	0.068189
...	...	...	...	...	...	...	...	...	...	...

Clustering

Since the Humberside region was the focus of the clustering, the features data that were chosen were based on the data from that region, which includes Kingston upon Hull, North Lincolnshire, North East Lincolnshire, and East Riding of Yorkshire. To determine the specific position and geographic coordinates, latitude and longitude were employed. The grouping was done using the K-means and Kmedoids algorithms, although the results suggest that Kmeans performs better, as evidenced by the silhouette score below.



```
Silhouette Score for K-Means: 0.5825540469509365
Silhouette Score for K-Medoids: 0.4110319653535337
```

Figure 9

The map below illustrates the accident cluster locations in each region. Kingston upon Hull is represented by the blue data point, East Riding of Yorkshire by the data point, North Lincolnshire by the green cluster, and North East Lincolnshire by the data point. The intensity of this cluster reveals the variations in accident frequency throughout the Humberside area. Additionally, the map shows that there are more accidents in the Kingston upon Hull area.

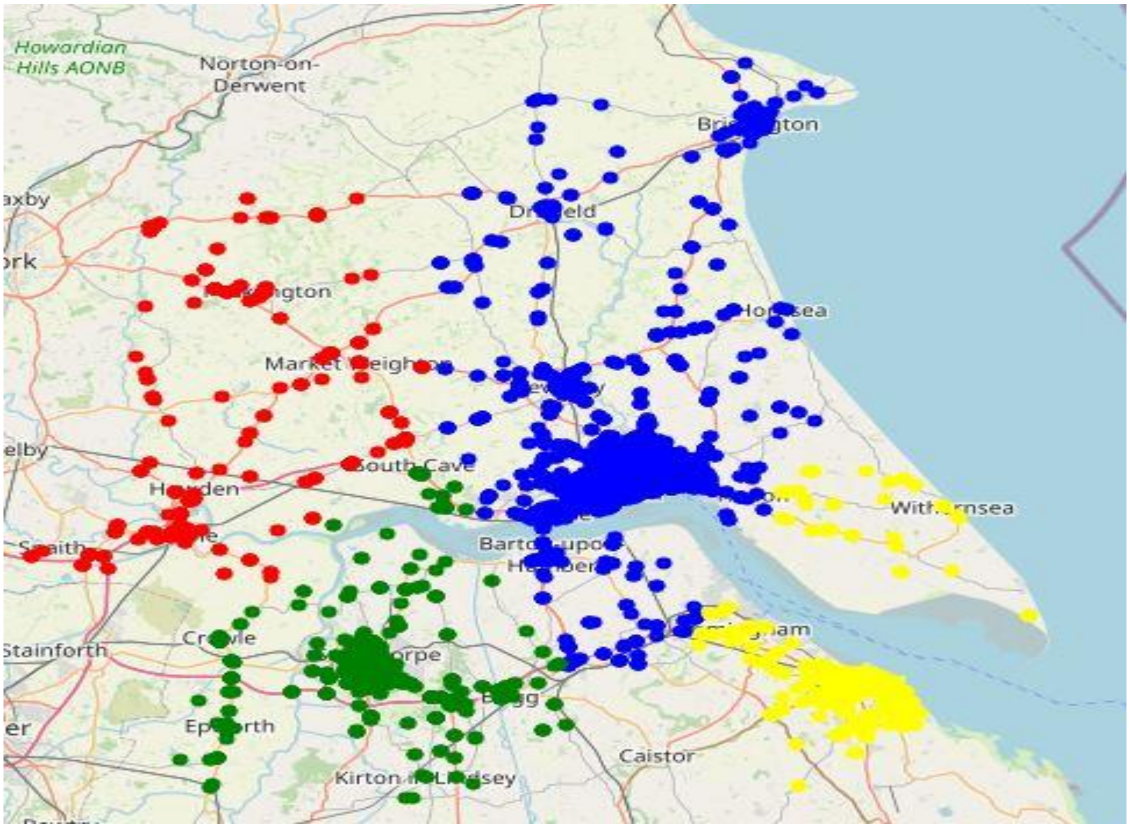


Figure 10

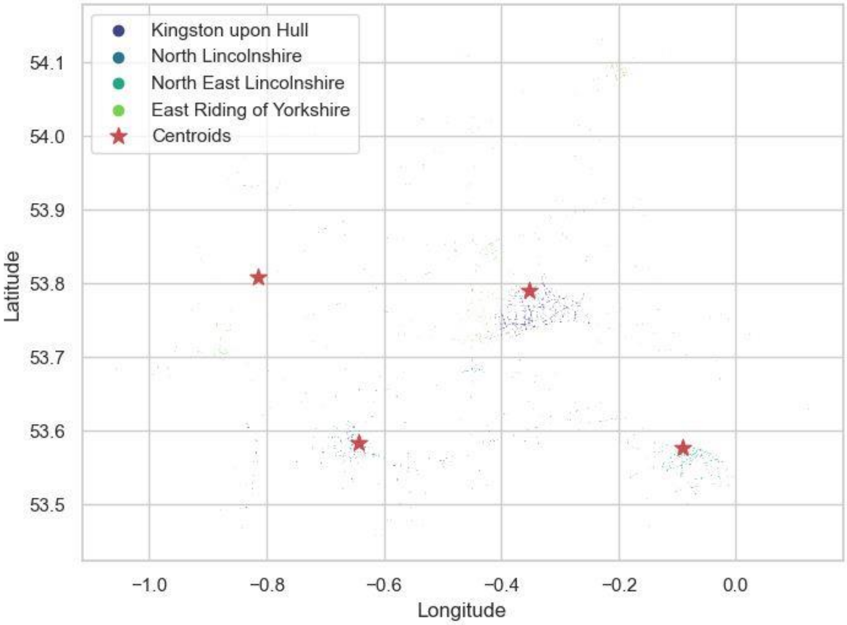


Figure 11 shows that the accident happened primarily in Hull, which appears to be because of Hull City Interchange. Major events have also occurred on all the main highways leading to the interchange. A concentrated number of points of incidents in the city of Hull indicates that it is particularly vulnerable to traffic accidents, which may stem from the interchange, which is a known accident hotspot.

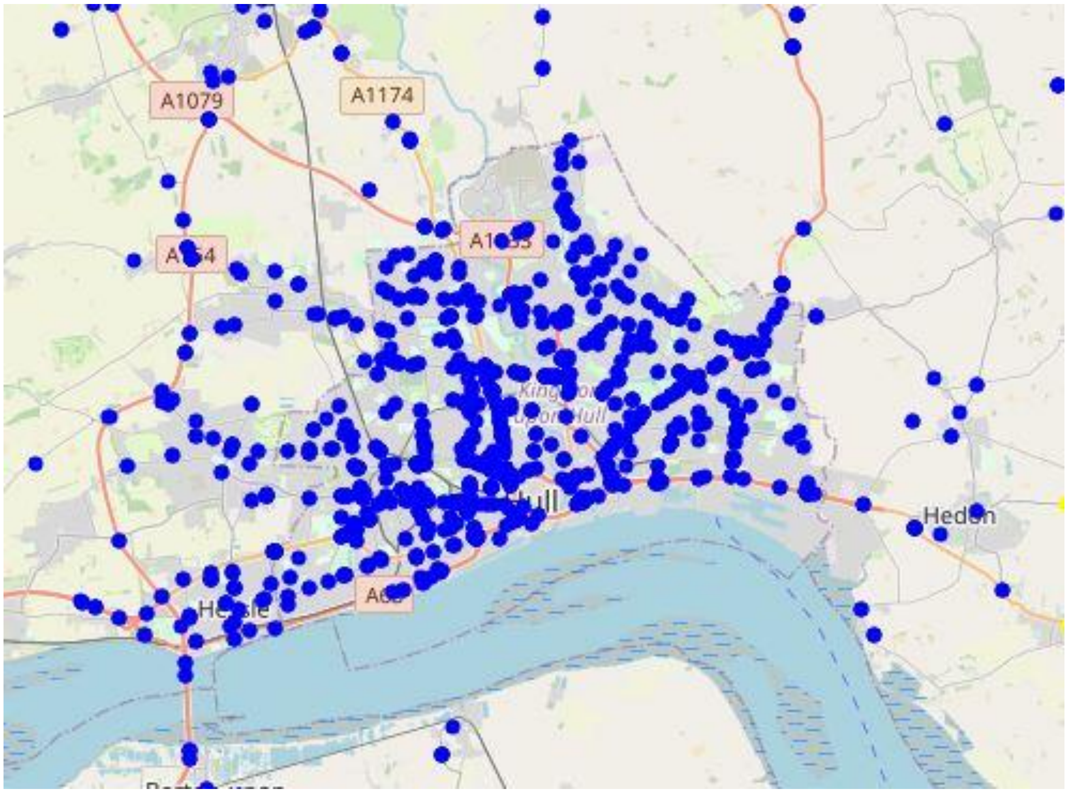


Figure 11

Data cleaning and Outlier detection

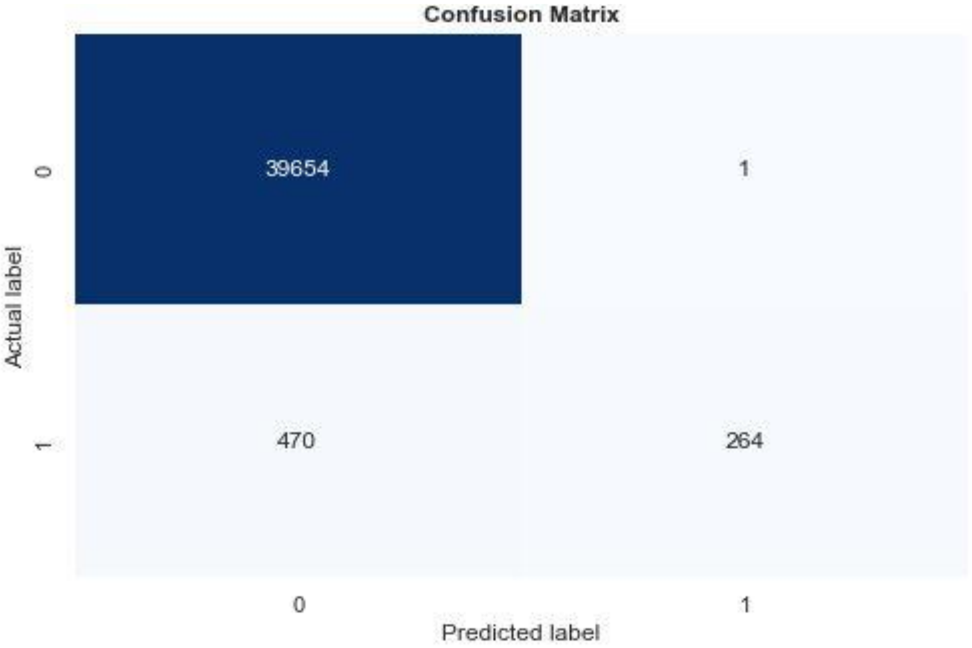
From the features selected in the accident data 2020, note that the features are accident severity, number of vehicles, number of casualties, road type, speed limit, road surface conditions, weather conditions, light conditions, vehicle type, age of vehicle, engine capacity cc, age of driver, sex of driver, pedestrian location, age of casualty, sex of casualty, casualty class, police force, lsoa01nm , urban or rural area. Thus, there are 20 columns and 201943 rows in the data. The dataset's 26219 rows have an outlier input of -1, which is present in all the dataset's columns, are found using the Isolation Forest model.

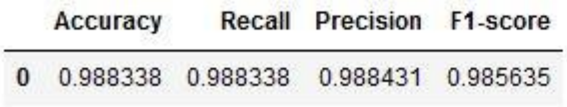
To avoid bias and preserve the data distribution, all the outlier inputs (-1) were replaced with null values, and the continuous variables with null values were replaced with the median of the data in the column, while the categorical variables with null values were replaced with the mode of the data in the column to maintain the data integrity and preserve the relationship between the data.

Machine Learning

Two models Random Forest classifier and Gradient boosting classifier were utilised to construct a model that successfully predicts the fatal injuries incurred in auto accidents. The two models were created to gain insight into which of them is performing best at the same time. The model was created both before and after the data had been balanced and the model's balance was achieved by utilising an under-sampler.

So the figure 9 below shows the confusion matrix of the predicted model using Random forest before the model was balanced by using Under Sampler



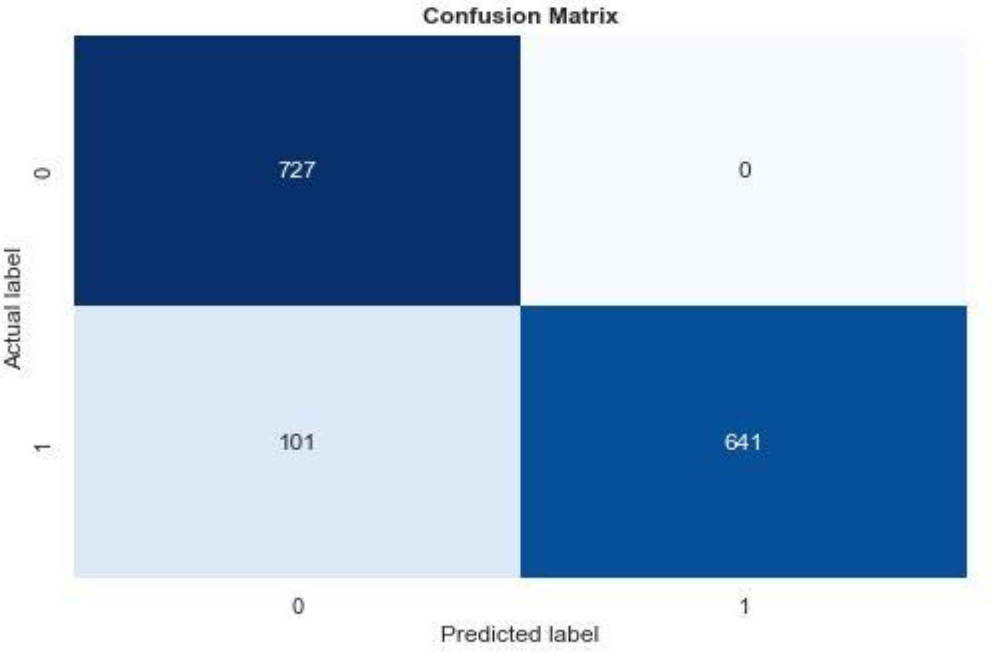


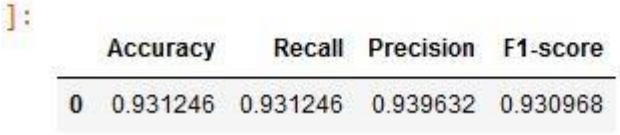
	Accuracy	Recall	Precision	F1-score
0	0.988338	0.988338	0.988431	0.985635

Figure 12

Accordingly, the confusion matrix in figure 12 above demonstrates that the model correctly predicted 264 True positives (TP), 470 False positives (FP), 1 True negative (TN), and 39654 True negatives (TN). The true positives are the cases where the model correctly predicted that the accident severity would result in the fatality, the false positives are the cases where the model incorrectly predicted that the accident severity would not result in the fatality, the false negatives are the cases where the model incorrectly predicted that the accident would not result in fatal outcomes but actually did, and the false positives are the cases where the model correctly predicted that the accident severity would result in fatalities but did not.

But after the Under Sampler was applied to the model, it performs better with the result below in Figure 13




]:

	Accuracy	Recall	Precision	F1-score
0	0.931246	0.931246	0.939632	0.930968

From figure 13, the Accuracy means that the model accurately predicted 93% accuracy rate in classifying whether the accident severity will be fatal or not, also the Recall indicate that the model predicted correctly 93% of the accident severity that are fatal and the precision shows that the model predicted correctly 93% which is 641 the accident that is really fatal in the accident data 2020

Recommendation

The analysis of the accident 2020 data has revealed a wealth of information and contributing factors that might serve as a guide and improve UK road safety. The depth of the understanding reveals several areas that require attention to increase traffic safety and keep the number of incidents to a bare minimum. The recommendation are:

Public Education and Training: Road users and the general public need to be informed of safe driving techniques, and the use of safety gear, such as seatbelts and helmets should be promoted. This awareness and campaigns need to be tailored down to different age groups of society.

Pedestrian safety measures: It is important to invest in infrastructure that will protect pedestrians, such as crosswalks, pedestrian bridges or overpasses. The placement and position of these structures in high-traffic and populated areas can aid in the safe passage of pedestrians.

Enforcement of road safety rules: Breaking traffic laws shouldn't be taken lightly, especially in high-risk situations like the late afternoon and nighttime. Anyone who violates traffic laws should face heavy penalties, which will send a strong message to the public, discouraging irresponsible driving and decreasing the risk of accidents.

Improvement in traffic management: At crucial roundabouts and interchanges where accidents are more likely to occur, like the Hull intercity interchange, traffic management devices like traffic lights should be built. Additionally, these technologies can be actively monitored to enhance and lower accidents caused by traffic jams. Additionally, to discourage speeding, patrols, roadblocks and the construction of speed monitoring equipment should be prioritised.

Visibility improvement: More caution lights that are sensitive to weather conditions should be installed due to the various weather conditions such as rain, fog, or snow that can have various effects on the visibility of the drivers, and road signs should be designed with reflective materials so that it is visible to drivers from a distance. This road's signs must also be maintained on a regular basis.

In conclusion, the government and stakeholders can collaborate to minimise the number of accidents and fatalities on the road by implementing recommendations and actions.

Reference

Brake (2022) UK collision and casualty statistics. Available at: https://www.brake.org.uk/get-involved/take-action/mybrake/knowledge-centre/uk-road-safety. [Accessed 07/08/2023]