NAME: KEMI AMUNI

COURSE: BIG DATA & DATA MINING 771762_B22_T3A

TASK: REPORT ON THE ANALYSIS OF ROAD TRAFFIC ACCIDENTS DATA FOR 2020 IN GREAT BRITAIN

INTRODUCTION

The accident dataset contains information about road traffic accidents in Great Britain, including information about the time, location, and nature of the accidents. The tables in the dataset include accident, vehicle, casualty, and LSOA in stats19 form. As a Data Scientist, the purpose of this analysis is to identify relevant time, day and location elements, analyze factors that influence accidents and their severity, and create a predictive model to estimate the chance of fatal accidents. The main goal of this analysis is to increase road safety and suggest insights that will help government agencies formulate effective road safety measures by utilizing data science methodologies and modelling.

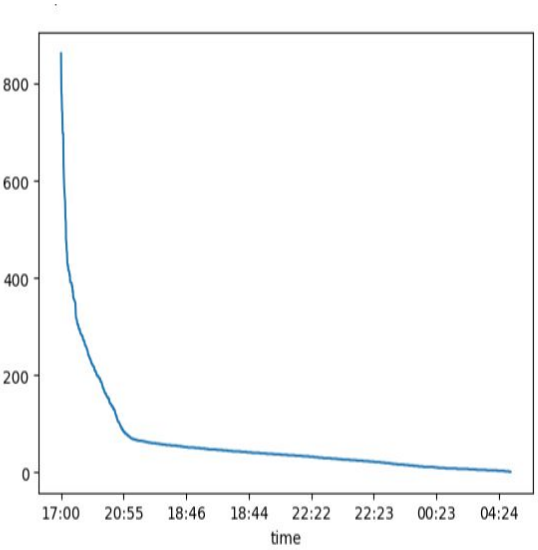
METHODOOGY

- Data cleaning and Preprocessing: Missing values was checked before running clustering in order to determine the distribution of accidents in the region.
- Time Analysis: To prioritize specific safety actions during accident peak hours, data visualization techniques are used to identify the significant hours and days of the week by which accidents occur.
- Motorcycle Analysis: Different investigations are carried out for motorbike accidents based on engine capacity (125cc or under, over 125cc up to 500cc, 500cc and over, and unknown cc) in order to examine accident trends for different motorcycle categories.
- Pedestrian Analysis: The significant hours of day and days of the week when accidents involving pedestrians occur were examined.
- Modelling: Apriori algorithm was applied on the dataset to explore the impact of selected variables on accident severity using one-hot encoding and association rules. Clustering was carried out to show the distribution of accidents in the region and Elbow method was explored to show how good the number of clusters are. The clusters were compared between 2, 4, and 2 clusters was adopted for this study using kmeans. Outlier detections was carried out to detect unusual entries using local outlier factor(LOF) and isolation forest. Gradient boosting classifier was also explored for predictions of fatal injuries sustained in the road traffic accidents.

ANALYSIS AND VISUALIZATIONS

Significant hourly occurrences of accidents

This study shows that accidents occur during the rush hour of 8am when the road is very busy with lots of commuters going to resume at their offices and students are also going to school. Later on the accident trends moves from 12noon which is lunch hour and 18:00 peaking at 17:00 hour when workers are returning back to their different homes. Fig.1 shows that the significant hours of the day on which accidents occur peaking at 17:00 with 862 accident index in the road traffic accidents data for 2020. Fig 3 shows the accident hour of the day and the accident severities while fig 4 shows the number of fatal accidents by hours of the day. This shows that fatal accidents occur starting from 8am morning rush hour and peaking at 18:00 when commuters are returning back



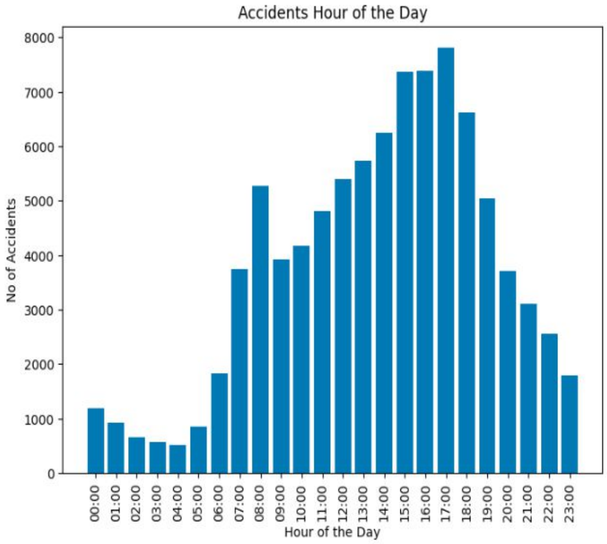
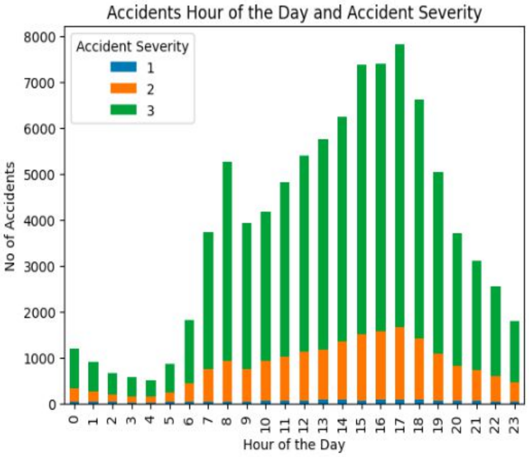


Fig 1. Significant accident hours and acident index

Fig 2. Significant accident hours and number of accidents



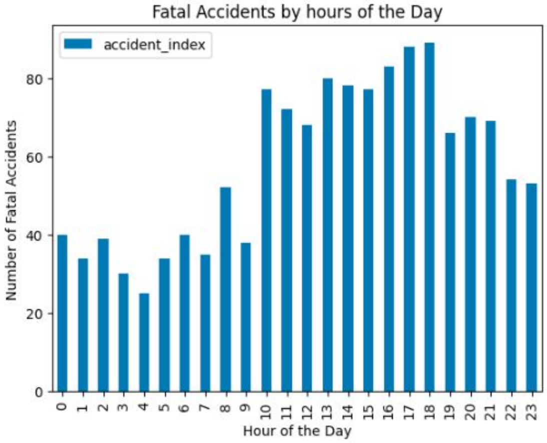


Fig.3 Accidents hour of the day and accident severity

Fig.4 Fatal accidents by hours of the day

Hourly accidents data for Motorcycle 125cc and under shows that

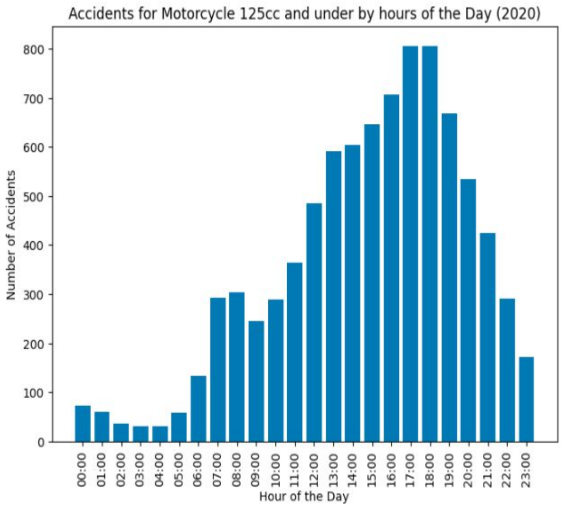


Fig.5 Accidents hours of the day for motorcycle 125cc and under

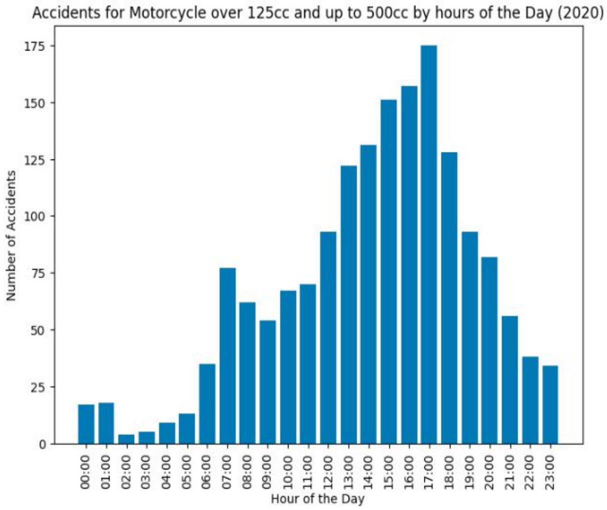


Fig.6 Accidents hours for motorcycle over 125cc up to 500cc

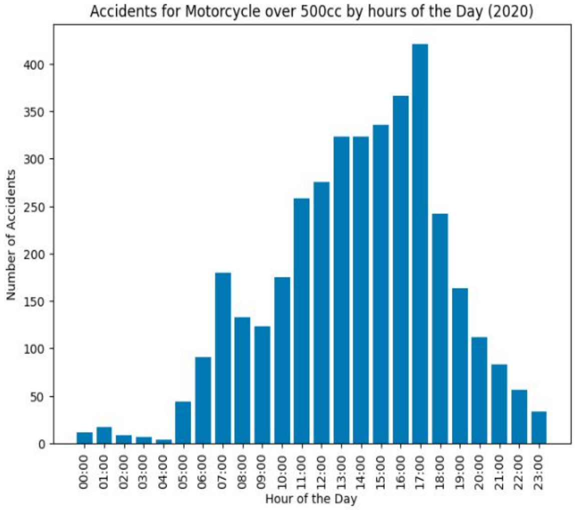


Fig.7 Accidents hours for motorcycle over 500cc

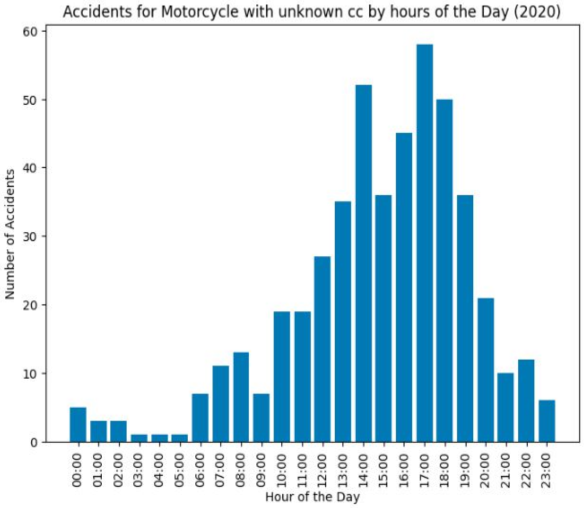
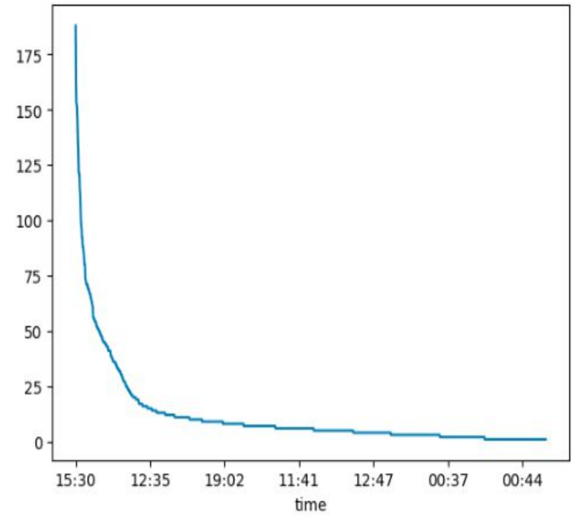


Fig.8 Accidents hours for motorcycle unknown cc

The significant hours of the day on which pedestrians with casualty class 3 are more likely to be involved in accidents are shown below. Fig 9 shows the Pedestrian accidents hours by accident index while fig 10 shows the number of pedestrian casualties by Age groups peaking at age group 10 -14 which suggests children of secondary school age group. The significant hours of the day on which accidents occur for "Pedestrian" is peaking at "15:30" giving a total of "188" accidents in the road traffic accidents data for 2020. This time also suggests school closing hours.



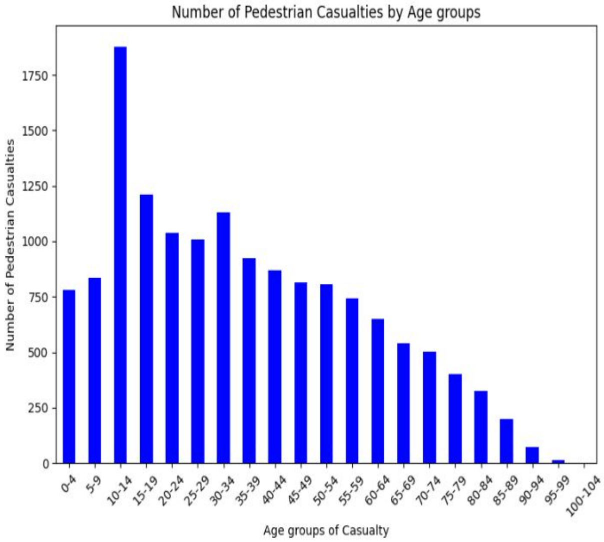
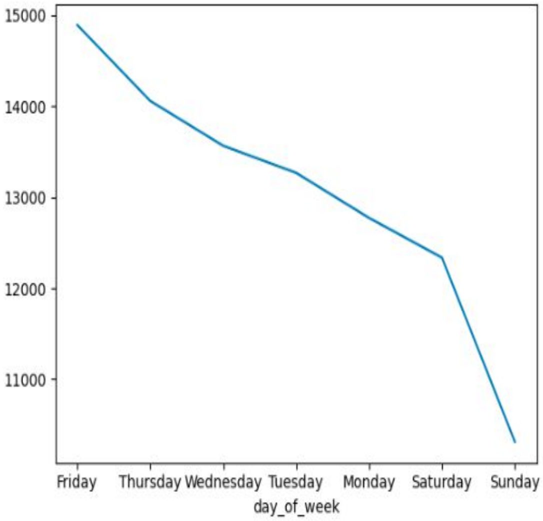


Fig.9 Accidents hours for motorcycle over 500cc

Fig.10 Accidents hours for motorcycle unknown cc

Significant daily occurrences of accidents

From the investigation on the accident datasets, fig 11 shows that the significant days of the week on which accidents occur and peaking at "Friday" giving a total of "14889" accidents in the road traffic accidents data for 2020. Fig 12 shows that the variations of the significant days of the week on which accidents occur for "Pedestrian" peaking at "Friday" which correlates with the weekend rush by commuters in the accidents in the road traffic accidents data for 2020.



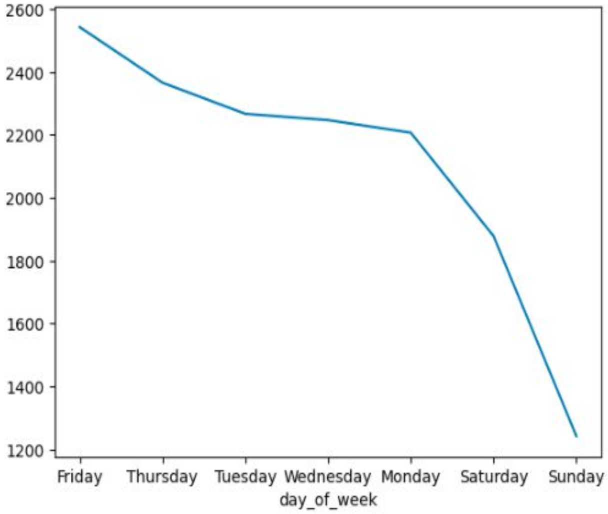


Fig.11 Significant accidents days of the week

Fig.12 Accidents hours for motorcycle unknown cc

Fig.13 shows variations of the significant accidents days of the week for motorcycle 125cc and under peaking on Friday, fig.14 accidents days of the week for over motorcycle 125cc up to 500cc peaking on Friday. Fig 15 shows the variations in the significant accidents days of the week for motorcycle over 500cc peaking on Sunday while Fig.16 shows the variations of accidents days for motorcycle unknown cc also peaking on Friday.

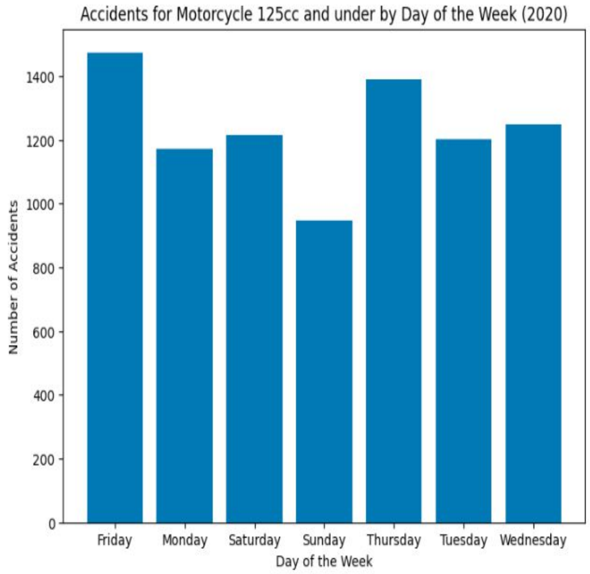


Fig.13 Significant accidents days of the week for 125cc and under

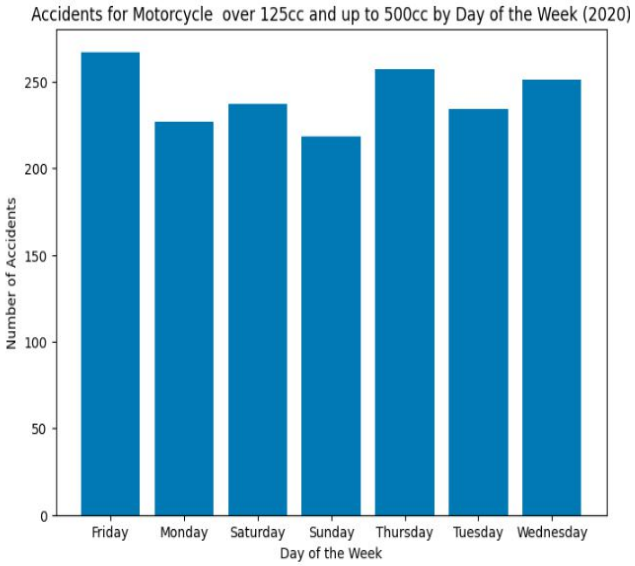


Fig.14 Accidents days of the week for over 125cc up to 500cc.

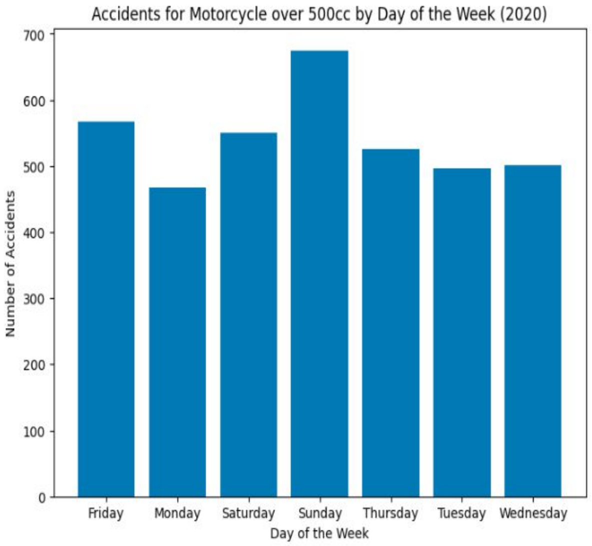


Fig.15 Significant accidents days of the week for over 500cc

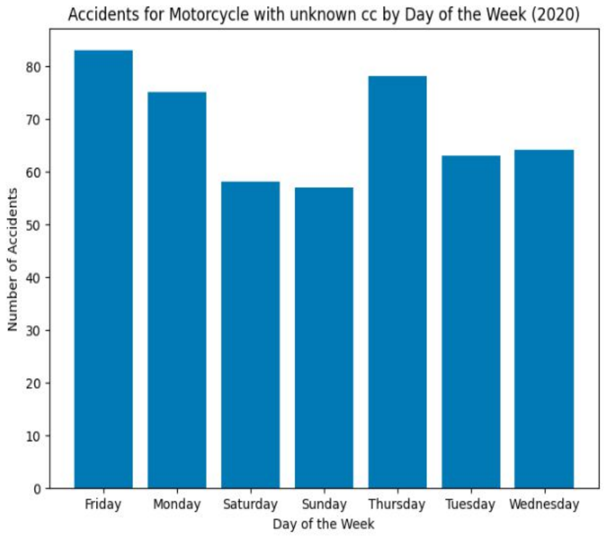
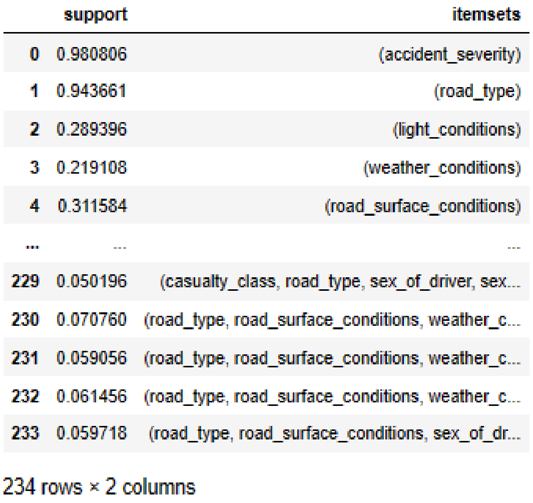


Fig.16 Accidents days for motorcycle unknown cc

The impact of selected variables on accident` severity using the Apriori Algorithm

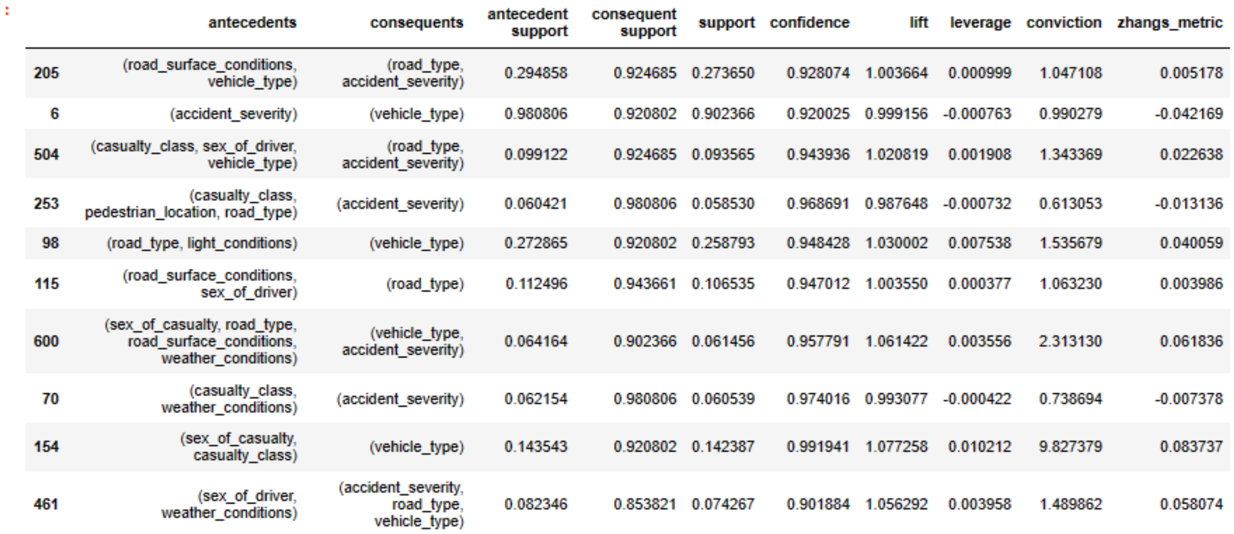
Tab1 indicates possible correlations among variables involved in accidents by highlighting the combos that commonly occur together in the accident dataset. A support value of 0.980806 indicates that accident severity data is included in almost all accident data. Their significance in comprehending accident trends is indicated by their high support values of 0.943661 for road type accordingly. Association rule in Tab2 shows that “casualty class” and weather conditions are associated with the accident severity from the accident data with a consequent support of 0.94 and confidence of 0.97.



	support	itemsets
0	0.980806	(accident_severity)
1	0.943661	(road_type)
2	0.289396	(light_conditions)
3	0.219108	(weather_conditions)
4	0.311584	(road_surface_conditions)
...	...	...
229	0.050196	(casualty_class, road_type, sex_of_driver, sex...
230	0.070760	(road_type, road_surface_conditions, weather_c...
231	0.059056	(road_type, road_surface_conditions, weather_c...
232	0.061456	(road_type, road_surface_conditions, weather_c...
233	0.059718	(road_type, road_surface_conditions, sex_of_dr...

234 rows × 2 columns

Tab 1 Significant accidents days of the week for over 500cc



:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
205	(road_surface_conditions, vehicle_type)	(road_type, accident_severity)	0.294858	0.924685	0.273650	0.928074	1.003664	0.000999	1.047108	0.005178
6	(accident_severity)	(vehicle_type)	0.980806	0.920802	0.902366	0.920025	0.999156	-0.000763	0.990279	-0.042169
504	(casualty_class, sex_of_driver, vehicle_type)	(road_type, accident_severity)	0.099122	0.924685	0.093565	0.943936	1.020819	0.001908	1.343369	0.022638
253	(casualty_class, pedestrian_location, road_type)	(accident_severity)	0.060421	0.980806	0.058530	0.968691	0.987648	-0.000732	0.613053	-0.013136
98	(road_type, light_conditions)	(vehicle_type)	0.272865	0.920802	0.258793	0.948428	1.030002	0.007538	1.535679	0.040059
115	(road_surface_conditions, sex_of_driver)	(road_type)	0.112496	0.943661	0.106535	0.947012	1.003550	0.000377	1.063230	0.003986
600	(sex_of_casualty, road_type, road_surface_conditions, weather_conditions)	(vehicle_type, accident_severity)	0.064164	0.902366	0.061456	0.957791	1.061422	0.003556	2.313130	0.061836
70	(casualty_class, weather_conditions)	(accident_severity)	0.062154	0.980806	0.060539	0.974016	0.993077	-0.000422	0.738694	-0.007378
154	(sex_of_casualty, casualty_class)	(vehicle_type)	0.143543	0.920802	0.142387	0.991941	1.077258	0.010212	9.827379	0.083737
461	(sex_of_driver, weather_conditions)	(accident_severity, road_type, vehicle_type)	0.082346	0.853821	0.074267	0.901884	1.056292	0.003958	1.489862	0.058074

Tab 2 Significant accidents days of the week for over 500cc

Clustering revelations about the distributions of accidents in the regions of Kingston upon Hull, East Riding of Yorkshire and Humberside.

Several clustering of accidents in the Humberside region are visible in the scatter plot. Such clust ers can be found in many different places, such as cities like Kingston upon Hull. 2 and 4 number s of were compared using the elbow method in order to determine how well the clusters are. 2 numbers of clusters generated Kmeans inertia of 58.095291259563695 while 4 numbers generate d Kmeans inertia of 21.518076924980857 (fig 17 and fig 18) and 2 numbers of clusters was adopted for the purpose of this study. Fig 19 shows scatter plot of speed limit vs weather conditions while tab3 shows clustering of speed limit and weather conditions which means that most accident happens under fine weather conditions without high winds. In fig 20 and tab 4 using accident severity and road surface conditions, it can be seen that most accidents occur under dry road surface conditions with good light conditions.

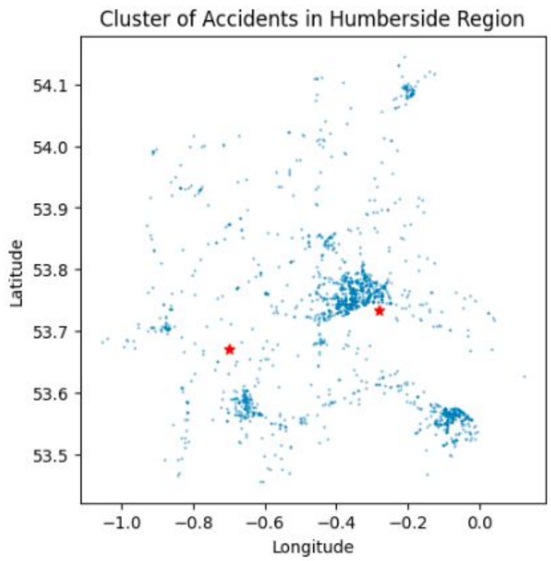


Fig.17 Accident clusters in Humberside using 2 clusters

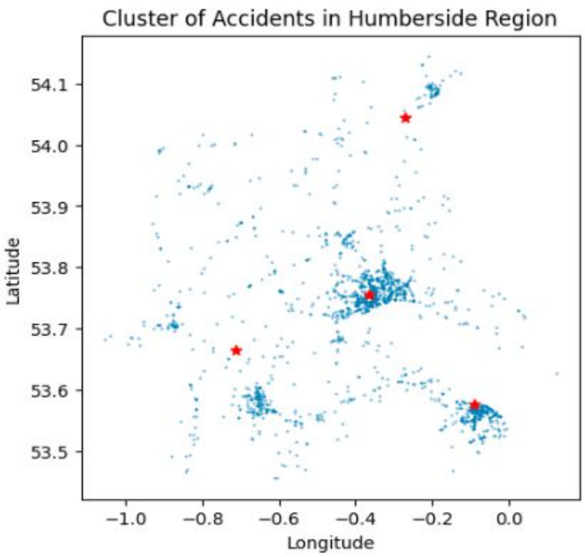
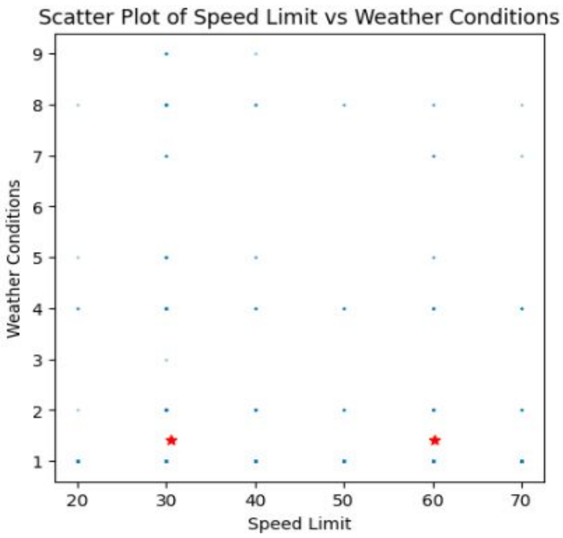
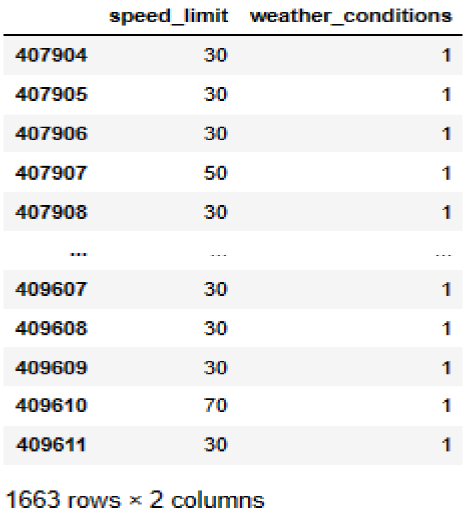


Fig.18 Accidents clusters in Humberside using 4 clusters



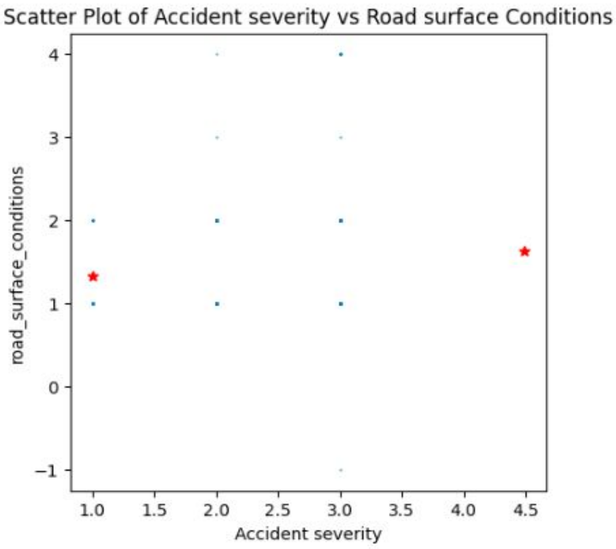


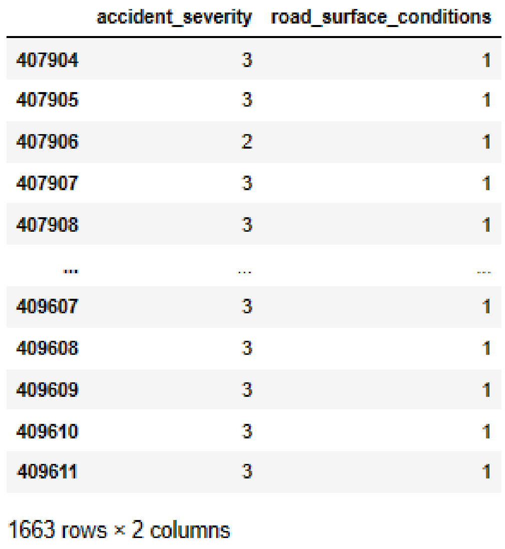
	speed_limit	weather_conditions
407904	30	1
407905	30	1
407906	30	1
407907	50	1
407908	30	1
...	...	...
409607	30	1
409608	30	1
409609	30	1
409610	70	1
409611	30	1

1663 rows × 2 columns

Fig.19 Scatter plot of speed limit vs weather conditions

tab 3 clustering of speed limit and weather conditions



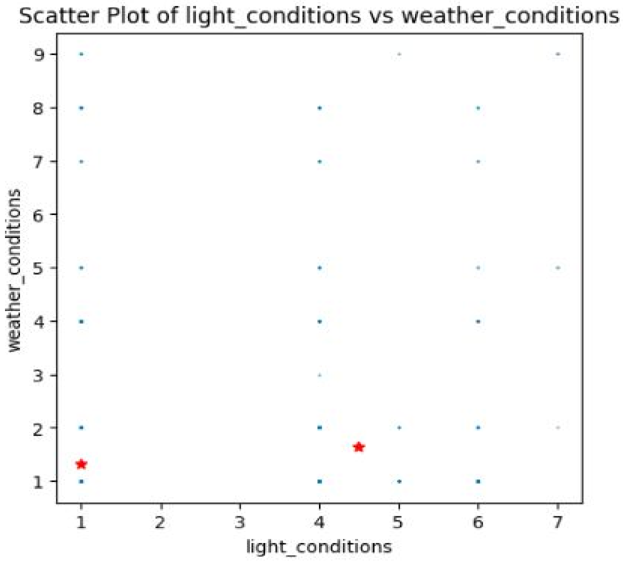


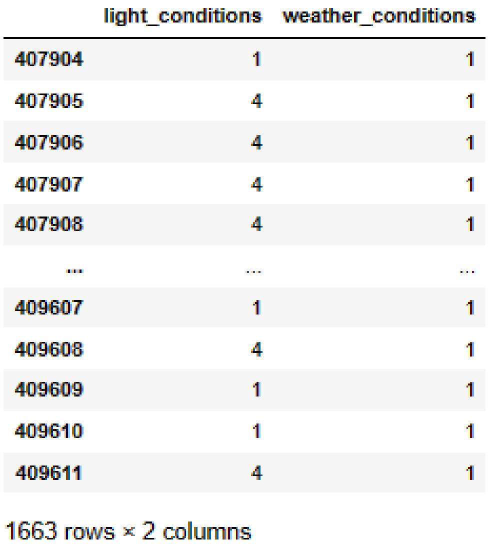
	accident_severity	road_surface_conditions
407904	3	1
407905	3	1
407906	2	1
407907	3	1
407908	3	1
...	...	...
409607	3	1
409608	3	1
409609	3	1
409610	3	1
409611	3	1

1663 rows × 2 columns

Fig.20 Scatter plot of accident severity vs road surface conditions

tab 4. Cluster of accident severity and road surface conditions





	light_conditions	weather_conditions
407904	1	1
407905	4	1
407906	4	1
407907	4	1
407908	4	1
...	...	...
409607	1	1
409608	4	1
409609	1	1
409610	1	1
409611	4	1

1663 rows × 2 columns

Fig.21 Scatter plot of light conditions vs weather conditions

tab 4. Cluster of light conditions and weather conditions

Using Outlier detection methods to identify unusual entries in the data set

Outliers was carried to detect the unusual entries in the accident dataset. The first unusual entry in the accident dataset is the age of driver <1 and the age of vehicles <. Figure 24 shows the scatter plot of accidents in Humberside region. Fig 25 and 26 shows the plot of accidents in Humberside region using the LOF and isolation forest outliers respectively to capture the unusual entries in the accident data while fig 27 shows accident plot for the entire UK.

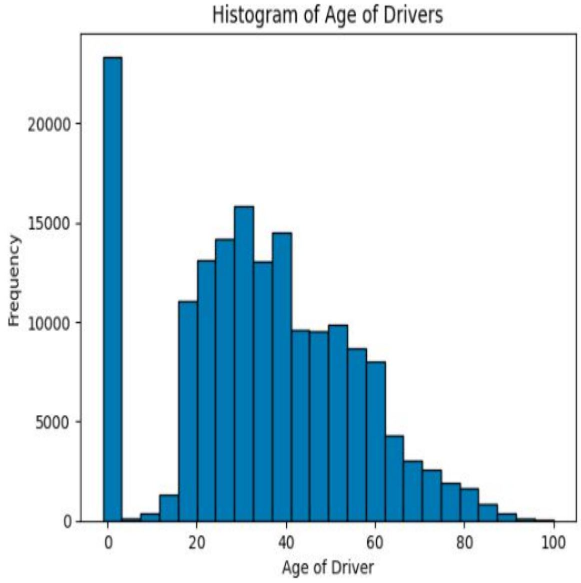


Fig.22 Histogram of age of drivers

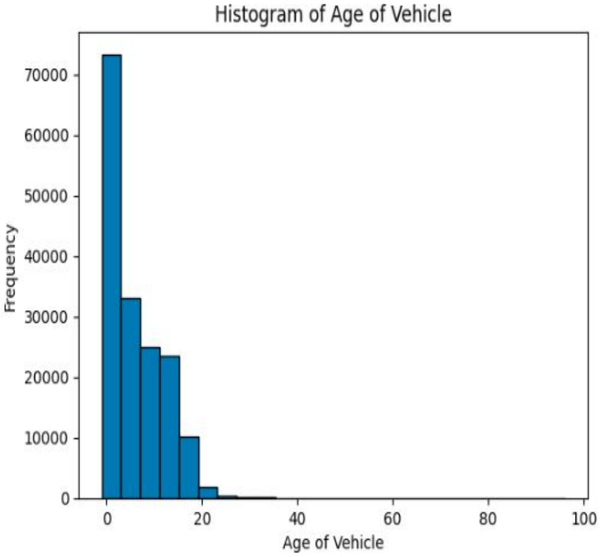


Fig 23 Histogram of age of vehicles

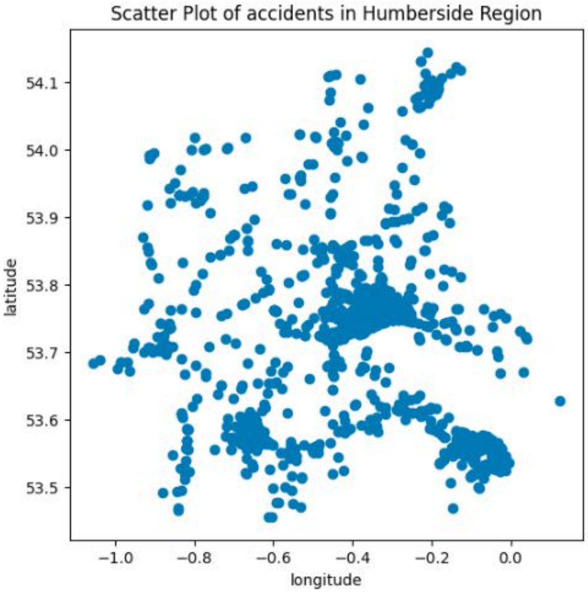


Fig.24 Scatter plot of accidents in Humberside region

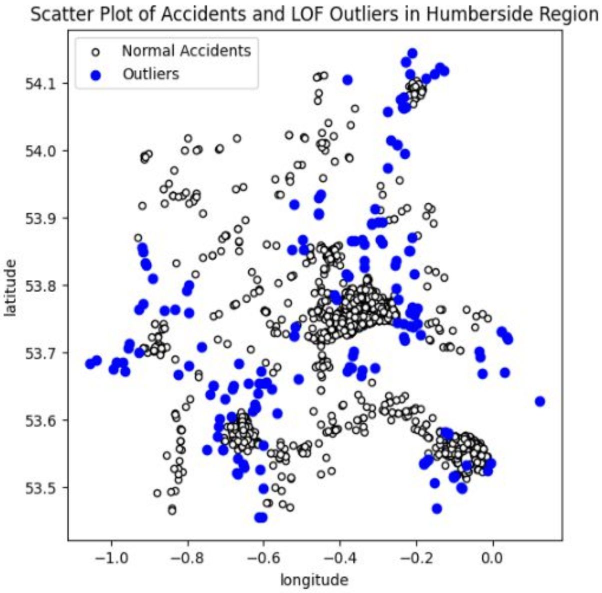


Fig 25 Scatter plot using LOF outliers in Humberside region

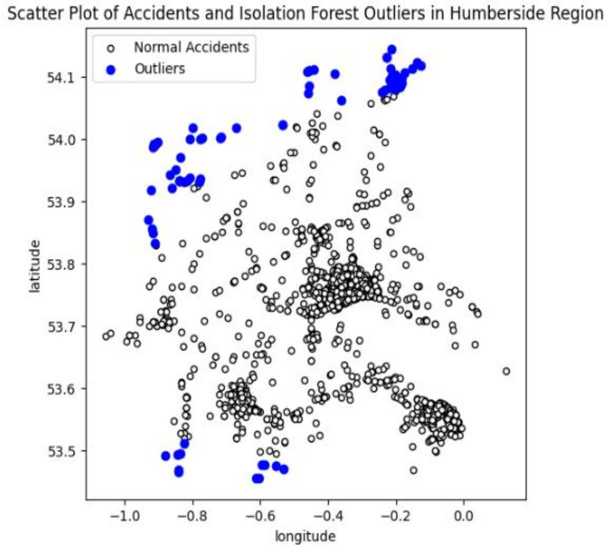


Fig.26 Scatter plot using isolation forest outliers in Humberside region

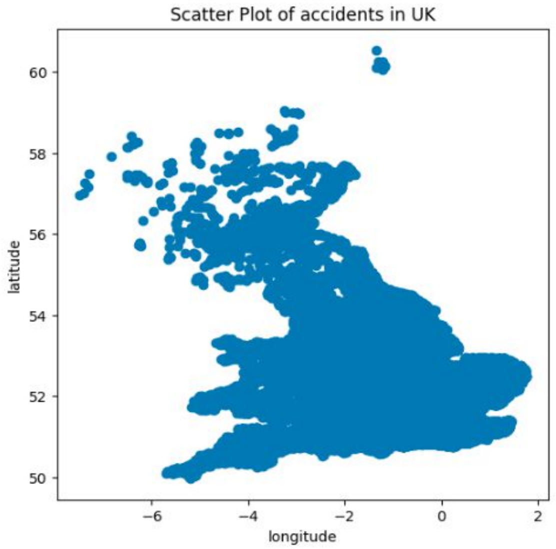
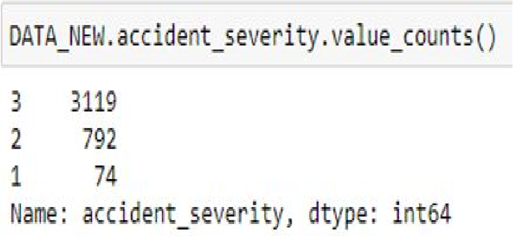


Fig 27 Scatter plot of accidents in the entire UK

PREDICTIONS

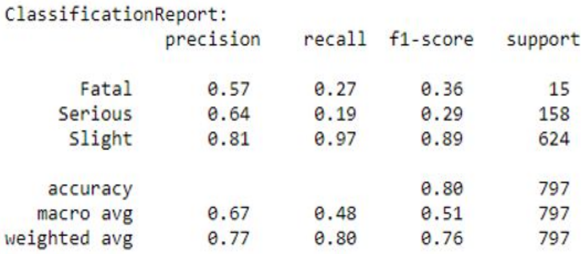
Classification model using the provided data that accurately predicts fatal injuries sustained in road traffic accidents, with the aim of informing and improving road safety measures. 20% of the accident severity data was used for testing, Gradient boosting classifier was explored and results were generated.



```
DATA_NEW.accident_severity.value_counts()

3    3119
2     792
1      74
Name: accident_severity, dtype: int64
```

Tab 5. Accident severity counts



```
ClassificationReport:
               precision    recall  f1-score   support

       Fatal       0.57      0.27      0.36        15
     Serious       0.64      0.19      0.29       158
      Slight       0.81      0.97      0.89       624

    accuracy                           0.80       797
   macro avg       0.67      0.48      0.51       797
weighted avg       0.77      0.80      0.76       797
```

Tab 6. Classification report of the accident severity

Out of the total number of fatal accident predicted, the model is able to make 57% fatal accidents, 64% serious accidents and 81% slight accidents predictions correctly. Recall shows that out the fatal accident in the test data, the model is able to recall 27% correctly, 19% of serious accidents and 97% of slight accidents were also recalled, with an overall accuracy of 80% on the test dataset.

RECOMMENDATIONS

Based on my analysis and predictions, these are my recommendations to government agencies on measures to improve safety

- Targeted Enforcement by increasing number of traffic police during periods when accidents are most likely to occur, such as 5pm on Fridays when accidents are more common.
- The government should ensure that speed limit is enforced especially in school locations considering this study that shows that most accidents occur at 15:30 and casualty age of 10-14 which implies school closing hours to prevent drivers from endangering their lives and the life of pedestrians.
- Pedestrian Safety Measures including well-lit pedestrian crossings and awareness-raising initiatives to raise pedestrian awareness during times that have been identified as being high-risk.
- Concentrate on Contributing Factors: Based on Apriori analysis, address contributing factors like road type that have been linked to accident severity. The government should invest in road infrastructure like speed breakers on high density accident locations in order to fix the issues of road accidents arising from the road type.
- The government should also ensure that road signage is more visible and clear especially in areas where the speed limit is 30 in order to save lives

Governmental agencies can act proactively to enhance traffic safety, reduce accidents, and most importantly save lives**.**

References

Department for Transport (2021). Reported road casualties Great Britain, provisional results: 2020. Retrieved from https://www.gov.uk/government/statistics/reported-road-casualties-great-britain-provisional-results-2020/reported-road-casualties-great-britain-provisional-results-2020

GOV.UK. (2020b). Apply for your first provisional driving licence. Retrieved from https://www.gov.uk/apply-first-provisional-driving-licence