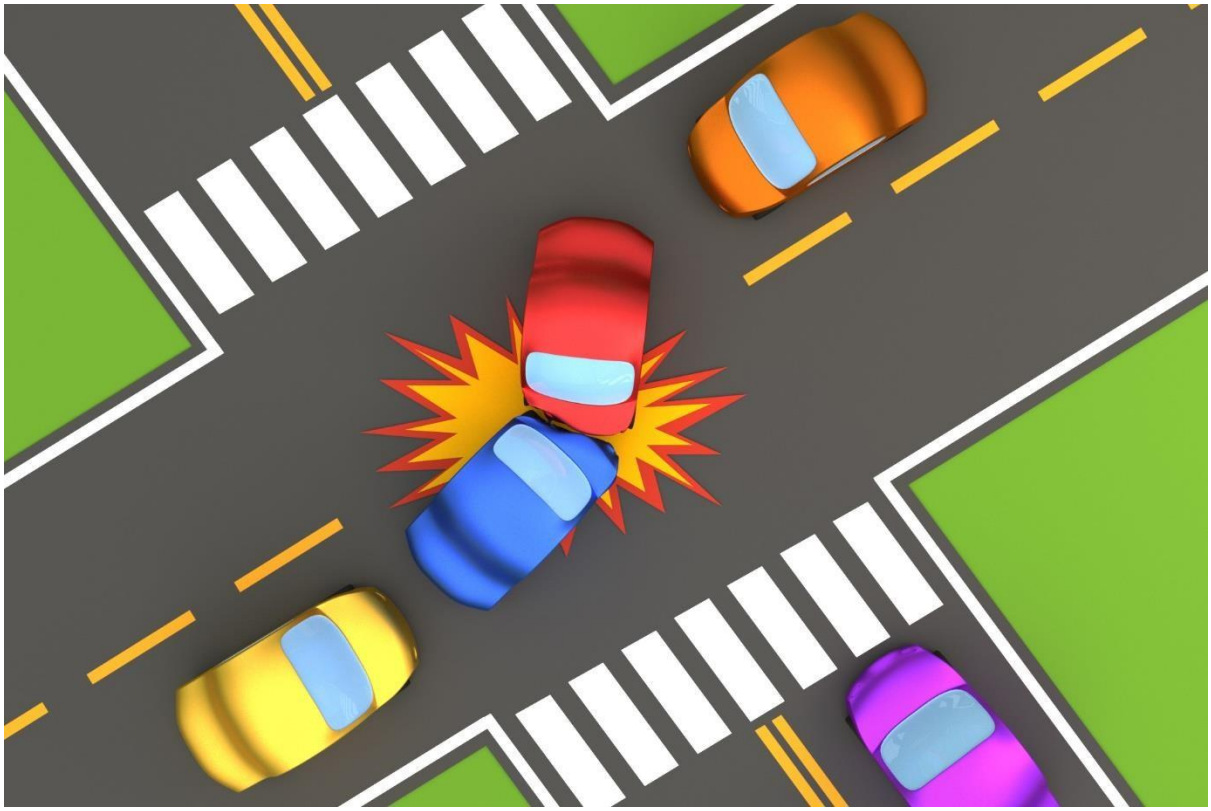


ROAD ACCIDENT DATA IN THE UNITED KIGDOM

A CASE STUDY OF YEAR 2020

OLIVIA .C. OKORO

1.0 INTRODUCTION

This study examines UK road safety statistics from 2020 to create a model that forecasts accidents and injuries. The aim of this project report is to provide insights that may be used to enhance road traffic safety procedures and avoid accidents. The dataset used for this analysis was extracted from a database containing various accident records within a range of years. It is divided into 4 parts/tables: The Vehicle, Casualty, Accident, and LSOA tables. We want to create a prediction model that may assist government organizations and other stakeholders in making educated judgments regarding road safety policies and programs using data cleansing, feature engineering, and machine learning approaches.

2.0 DATA OBSERVATIONS AND PREPROCESSING 2.1 DATA LOADING

The database file of the United Kingdom road traffic accident was loaded and its contents was checked after which four(4) tables Accident, Vehicle, Casualty and the LSOA table was extracted although, our focus will mostly be on the first three. We went further to extract that data for the year 2020.

2.2 DATA STRUCTURE

The data structure for the three (3) dataset for year 2020 was observed and the Accident data has 91,199 rows and 36 columns (with 14 rows and 4 columns containing Nan values), vehicle data has 167,375 rows and 28 columns (with no Nan values), and Casualty data has 115,5844 rows and 19 columns (with no Nan values). It is worthy of note that the Accident dataframe with primary key “accident index” is linked to both the Vehicle and Casualty data by their foreign keys “accident index” are linked while the Vehicle and Casualty data are linked by the “Vehicle reference.”

2.3 DATA PREPARATION, FEATURE ENGINEERING AND CLEANING

It was observed that the date and time columns was casted as a datatype object instead of as a datetime datatype. This was converted to its appropriate datatype; the date and time column was then merged to create the Time column after which the previous date and column was dropped. The columns that are of the object datatype was also converted to string to enable merging of the three dataframes.

The columns (longitude, latitude, Location Easting OSGR, and the Location Northing OSGR) were grouped according to the same local authority district and the mean values was calculated distinctly for each column and was used to replace the null values.

The null values in the time column were also replaced with the most occurring which is the mode of in the time column. Percentiles was also used to calculate the outliers in the age of the drivers and the age of the vehicle.

3.0 DATA ANALYSIS

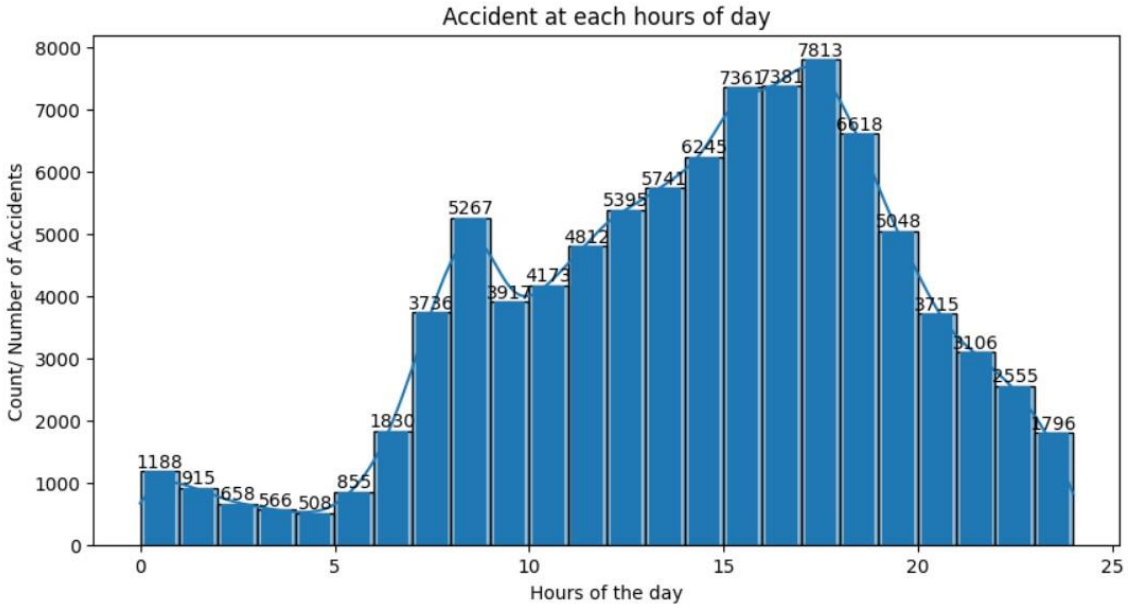


Fig 1

The graph above shows that the most accidents occurred by 6.pm (18.00 p.m.) with 7,813 counts followed by 5p.m with 7,381 counts of accidents. The least accidents occur around 5a.m in the morning with 508 counts of accidents.

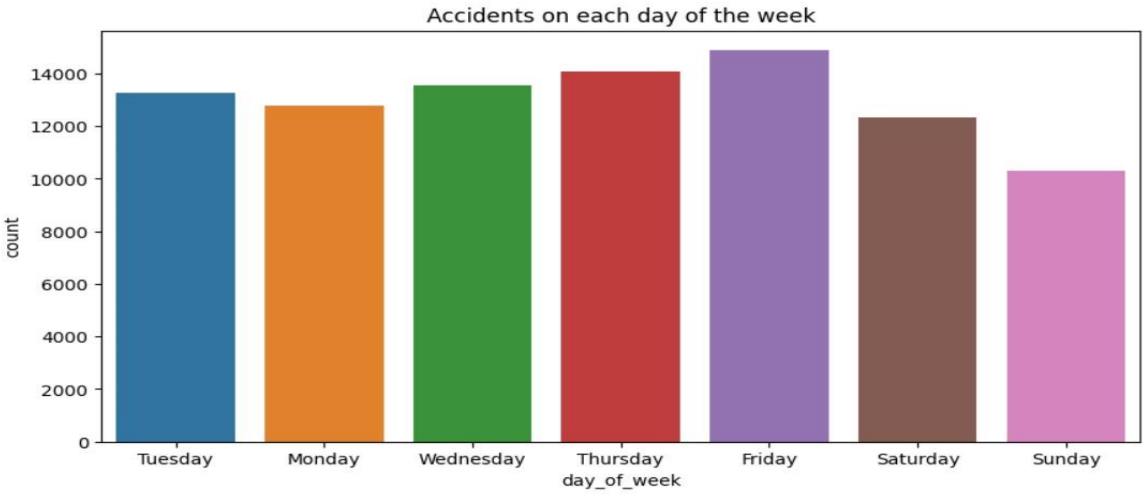


Fig 2

The diagram above depicts that the highest number of accidents over 14,000 occurs on the 6th of day of the week. The highest number of accidents occurs on Friday since the first day of the week is Sunday while the last day is Saturday.

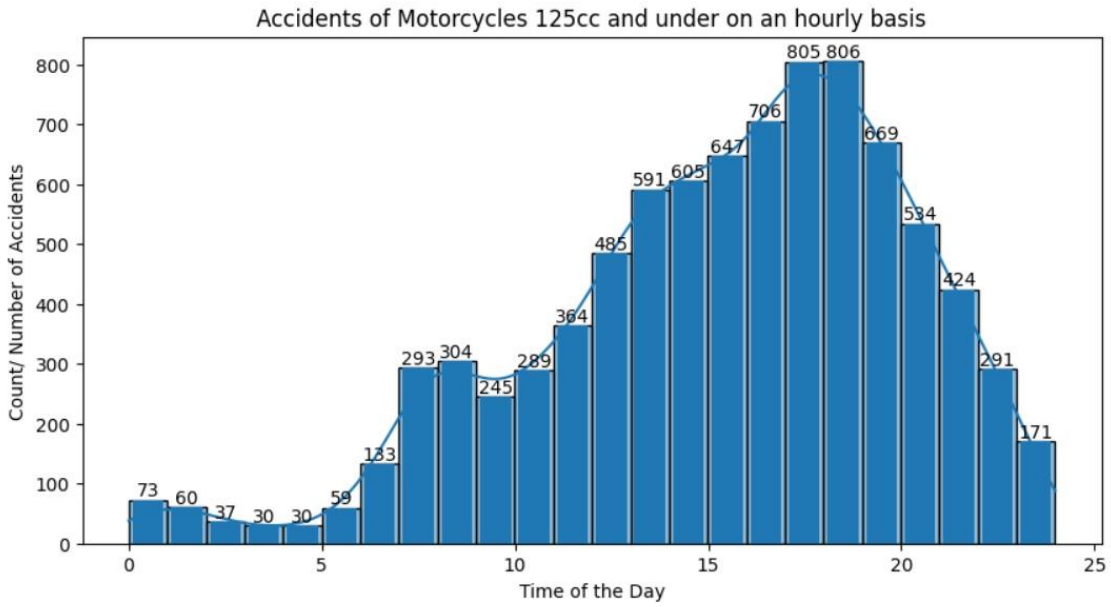


Fig 3

The highest count of motorcycle 125cc and under accidents occurs at 7p.m(19:00 hours) with a count of 806 followed by 6p.m with a count of 805. The least time for which accident occurs is at 4.am and 5am respectively with a count of 30 accidents each respectively.

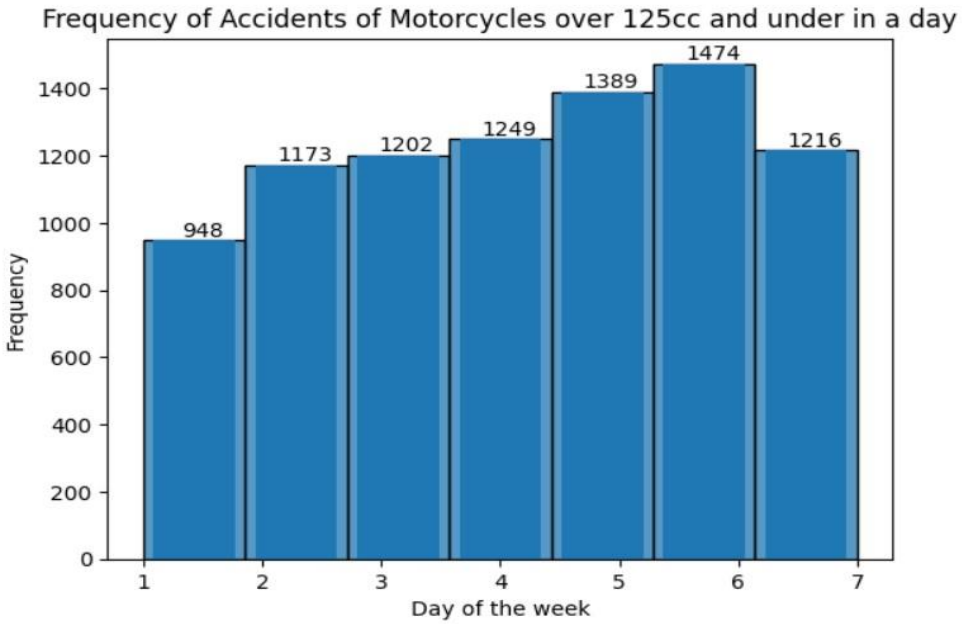


Fig 4

The diagram above depicts that the highest number of accidents 1,474 occurs on the 6th of day of the week (Friday) while the least days of accidents occurring with 948 accidents is on Sunday. Note the first day of the week is Sunday while the last day is Saturday.

The result shows that the p-value is 0.730 which is greater than 0.05 meaning that accidents in days of the week is Probably Gaussian therefore, there is no significant day of the week in which accident occurs and the stat is significant at 0.950.

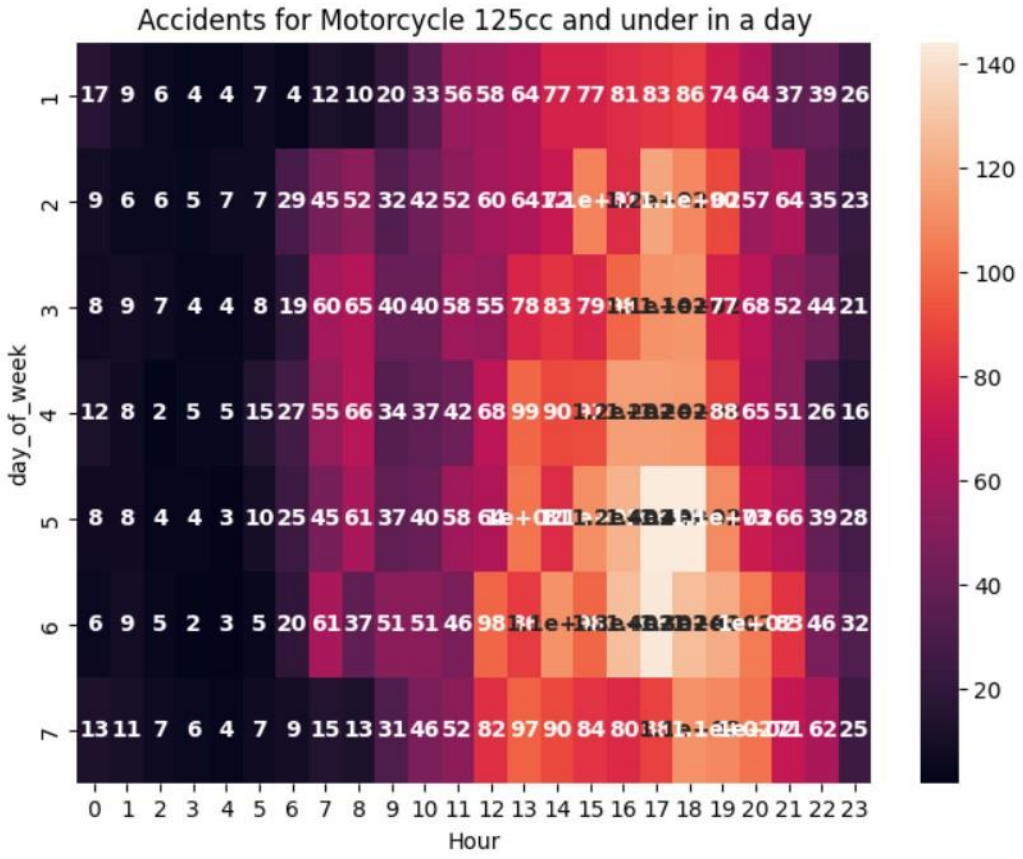


FIG 6

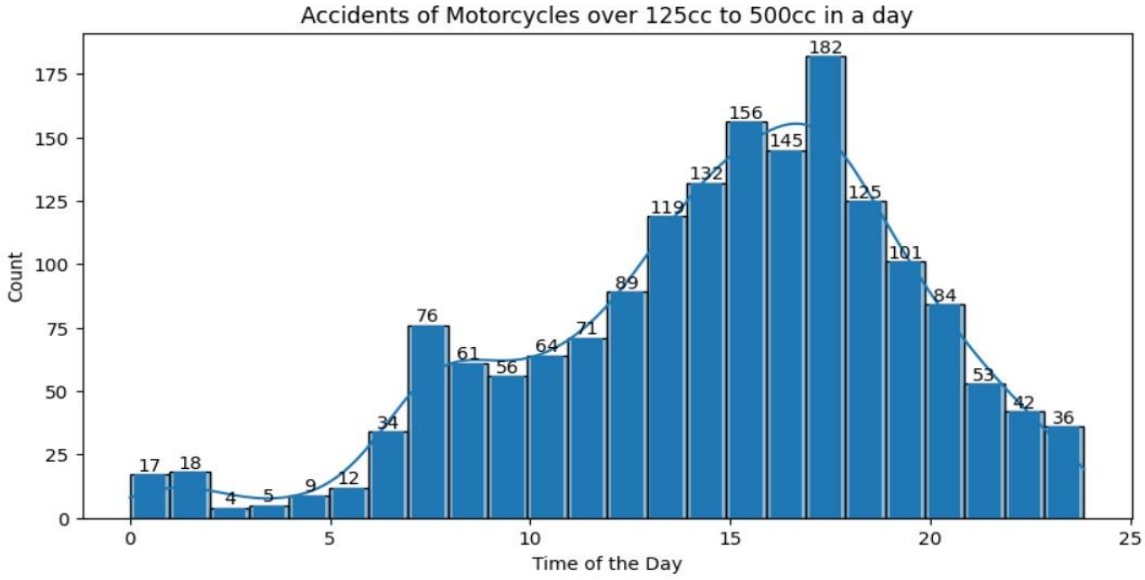


FIG 6

The highest count of motorcycle 125cc to over 500cc accidents occurs at 6p.m(18:00 hours) with a count of 182 accidents followed by 4p.m(16:00 hours) with a count of 156. The least time for which accident occurs is at 3a.m and 4a.m respectively with a count of 4 and 5 accidents each respectively. The Shapiro values shows that the p-value is 0.000 and is less than 0.050; therefore, accidents in hours of the day is Probably not Gaussian. What this means is that there are significant hours of the day in which accidents occurs.

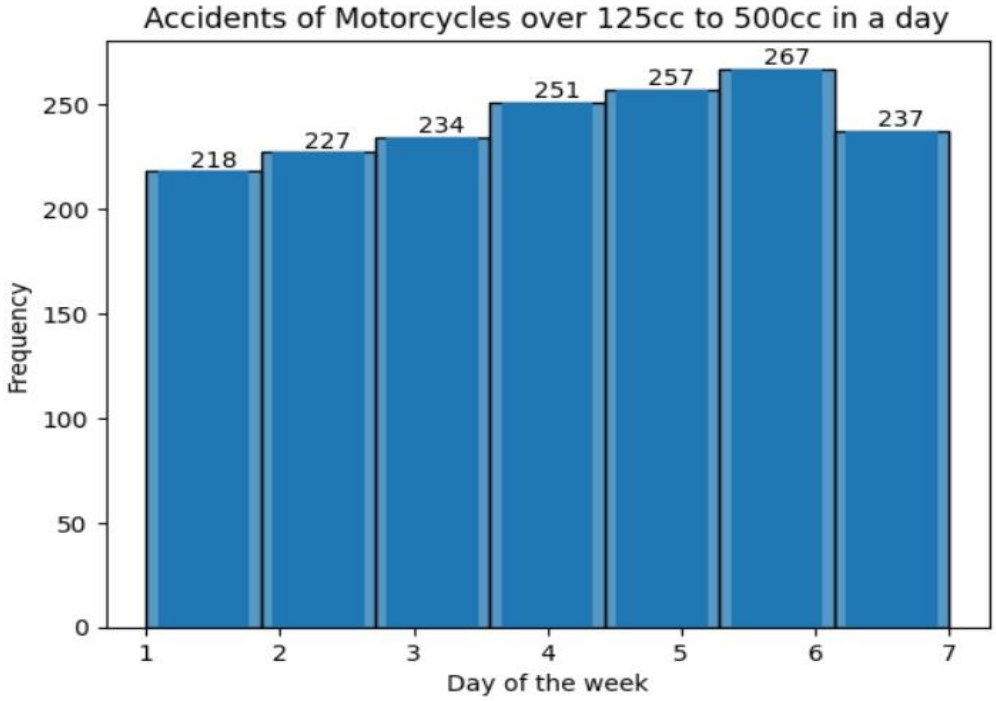


FIG 7

The diagram above depicts that the highest number of accidents 267 occurs on the 6th of day of the week (Friday) while the least day of accidents occurring with 218 accidents is on Sunday. Note the first day of the week is Sunday while the last day is Saturday. Also, the Shapiro value shows that the p-value is 0.730 which is greater than 0.05 means that accidents in days of the week is Probably Gaussian therefore, there is no significant day of the week in which accident occurs.

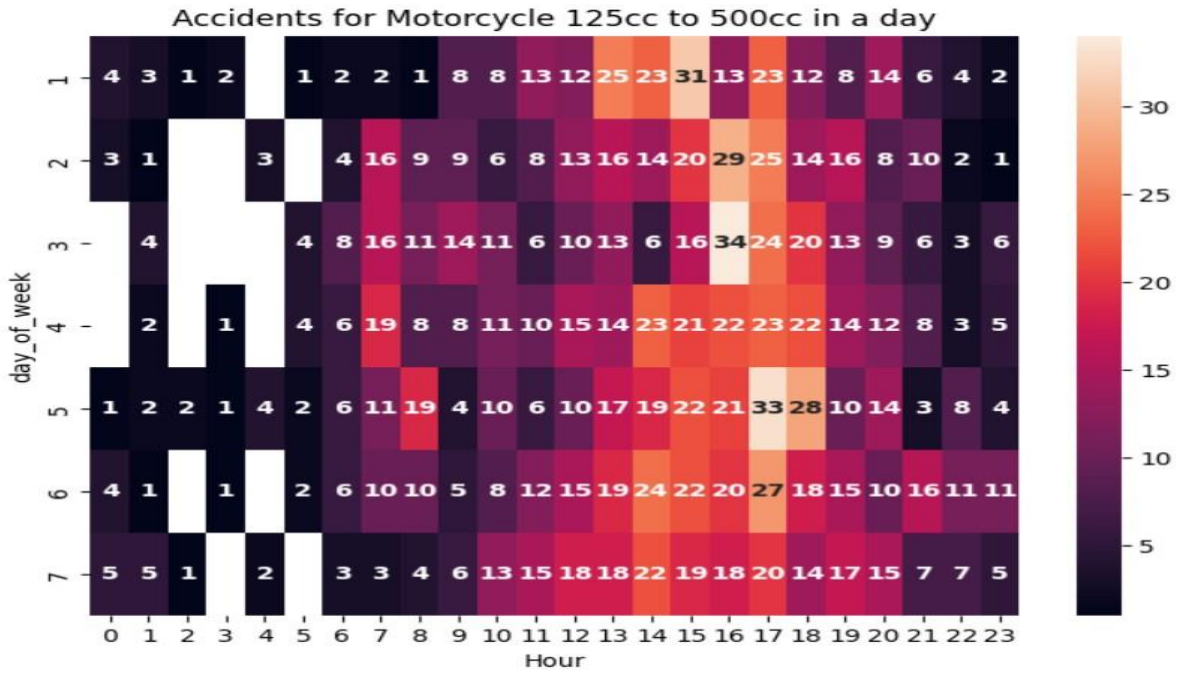


FIG 8

The above is the heatmap of the motorcycle 125 to 500cc.

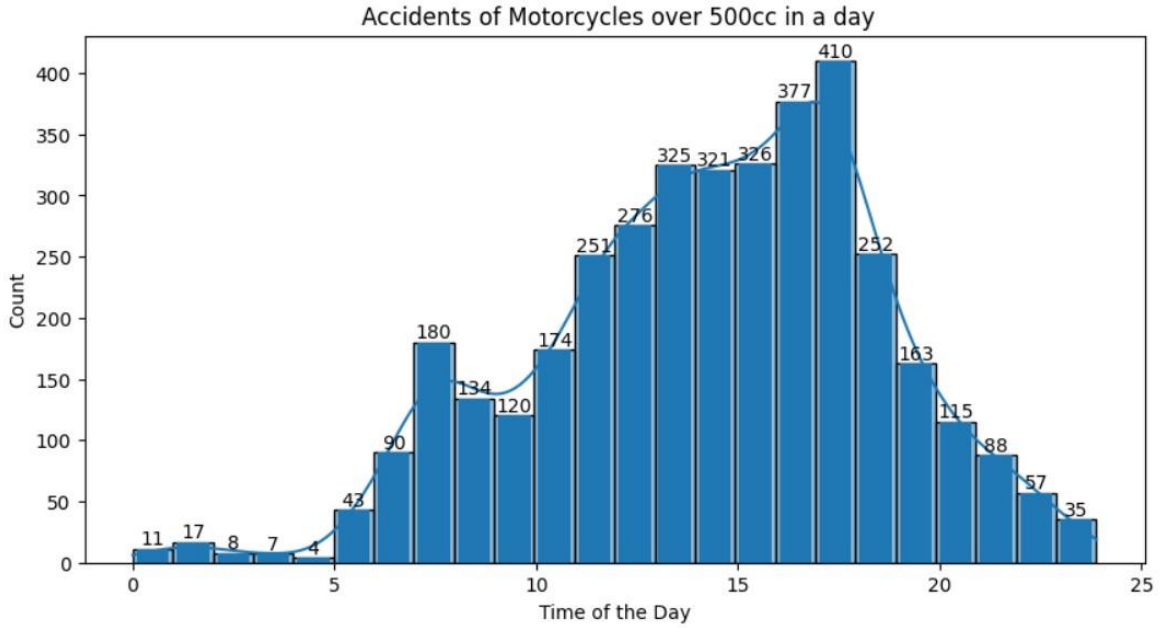


FIG 8

It can be seen that for motorcycle 500cc and over, accidents occur mostly at 6p.m(18:00 hours) with 410 accident counts while the least counts of accidents occur by 5a.m with accident counts of 4. From the Shapiro analysis, the p-value is 0.000 and is less than 0.05; therefore, accidents in hours of the day is

Probably not Gaussian. What this means is that there are significant hours of the day in which accidents occurs. with a stat value of 0.493.

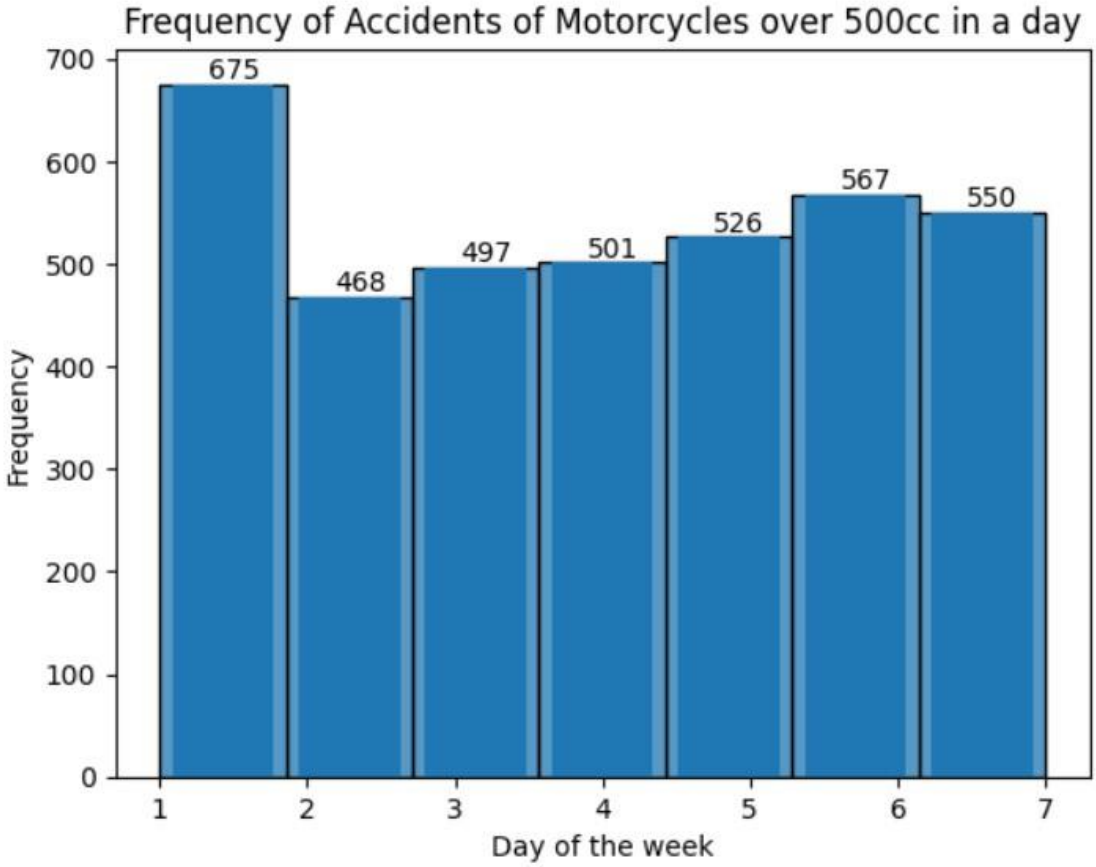


FIG 9

The day with the highest frequency of accident occurring is Sunday with accident frequency of 675 while the least accidents occurred on Monday with a count of 468. The Shapiro result shows that the p-value is 0.730 which is greater than 0.05 meaning that accidents in days of the week is Probably Gaussian therefore, there is no significant day of the week in which accident occurs. with a stat value that is 0.950 significant.

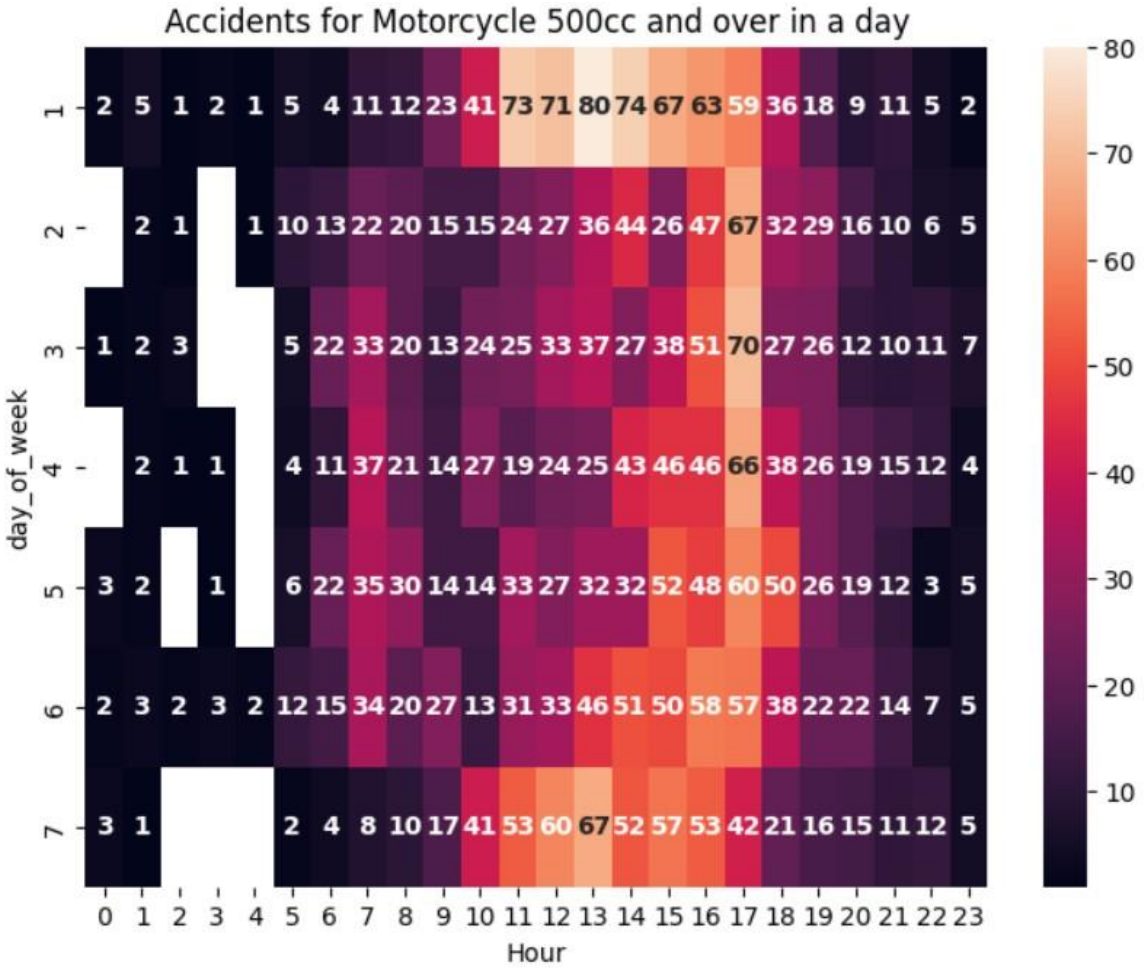


FIG 10: Heatmap of Accident for Motorcycle 500cc and over

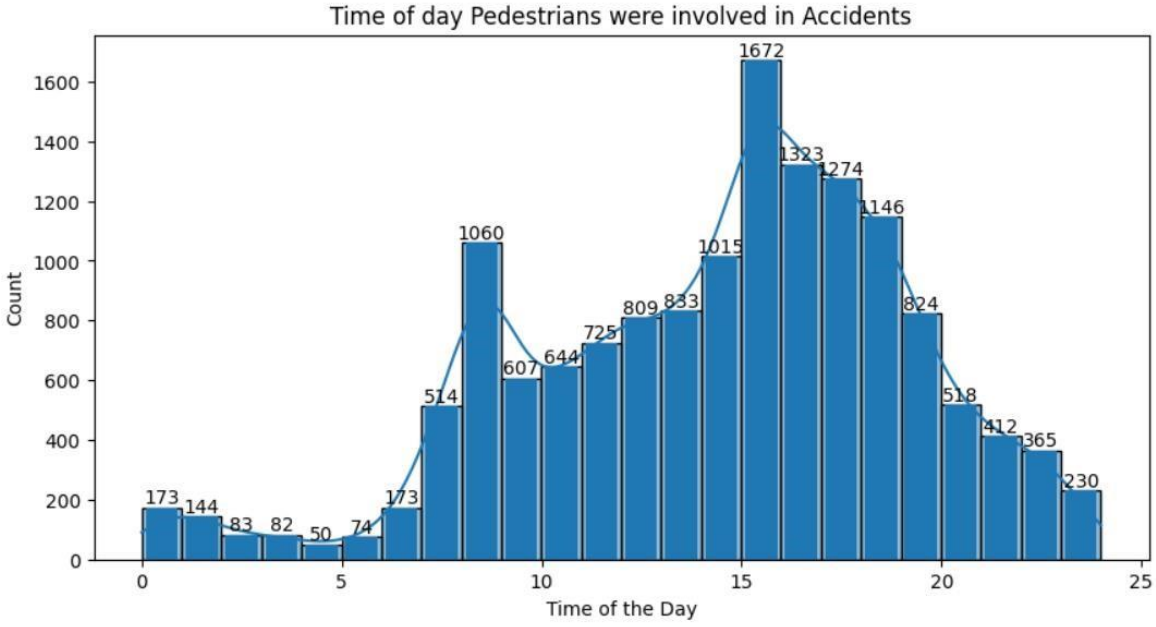


FIG 11

The count of accidents for which pedestrians were involved in was at its peak by 4p.m(16:00 hours) with 1,672 accident counts**.** This can be associated with

rush hour of the close of work or school runs and everyone scrambling to get to their destination on time without minding the road rules and regulations. The least accidents occurred at 5a.m with 50 accident counts.

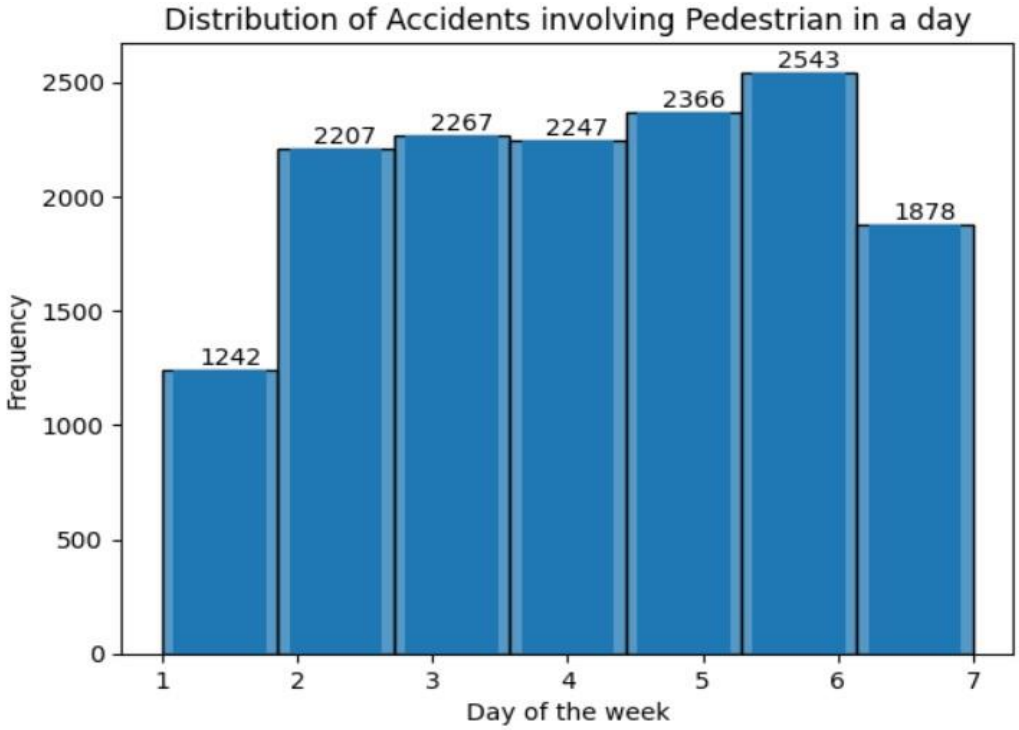


FIG 12

The above shows days of the week with the most frequency of pedestrians being involved in accidents is on Friday with 2,543 accident frequency followed by Thursday with 2,366. The least day is on Sunday with 1,242 accident frequency.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
0	(weather_1)	(severity_3)	0.775546	0.783484	0.603186	0.777757	0.992690	-0.004442	0.974230	-0.031765
1	(severity_3)	(weather_1)	0.783484	0.775546	0.603186	0.769877	0.992690	-0.004442	0.975365	-0.032891
2	(speed_30)	(severity_3)	0.573033	0.783484	0.459983	0.802717	1.024548	0.011021	1.097488	0.056116
3	(severity_3)	(speed_30)	0.783484	0.573033	0.459983	0.587099	1.024548	0.011021	1.034068	0.110660
4	(speed_30)	(weather_1)	0.573033	0.775546	0.450137	0.785534	1.012879	0.005723	1.046572	0.029780
5	(weather_1)	(speed_30)	0.775546	0.573033	0.450137	0.580413	1.012879	0.005723	1.017589	0.056649
6	(speed_30, weather_1)	(severity_3)	0.450137	0.783484	0.359697	0.799084	1.019911	0.007022	1.077643	0.035503
7	(speed_30, severity_3)	(weather_1)	0.459983	0.775546	0.359697	0.781979	1.008294	0.002959	1.029505	0.015233
8	(weather_1, severity_3)	(speed_30)	0.603186	0.573033	0.359697	0.596328	1.040653	0.014051	1.057709	0.098446
9	(speed_30)	(weather_1, severity_3)	0.573033	0.603186	0.359697	0.627708	1.040653	0.014051	1.065865	0.091493
10	(weather_1)	(speed_30, severity_3)	0.775546	0.459983	0.359697	0.463798	1.008294	0.002959	1.007115	0.036650
11	(severity_3)	(speed_30, weather_1)	0.783484	0.450137	0.359697	0.459099	1.019911	0.007022	1.016570	0.090164

FIG 13: The Apriori Table outcome

Rows 0 and 1: These rules show an association between severity_3 (Slight) and weather_1 (Fine without high winds). The confidence values are relatively high, indicating that when one of these variables occurs, there's a high likelihood of the other occurring. However, the lift values are close to 1, suggesting a weak

association. The negative leverage values indicate less co-occurrence than expected under independence.

Rows 2 and 3: These rules suggest a relationship between speed_30 (speed limit of 30) and severity_3 (Slight). The confidence values are relatively high, indicating that when one of these variables occurs, there's a significant likelihood of the other occurring. The lift values suggest a slight positive association. The positive leverage values indicate more co-occurrence than expected under independence.

Rows 4 and 5: These rules indicate a potential connection between speed_30 and weather_1(Fine without high winds). The confidence values suggest that when one variable occurs, the other is likely to occur as well. The lift values are close to 1, indicating a weak association.

Rows 6, 7, 8, 9, 10, and 11: These rules involve combinations of multiple variables. They suggest relationships between different combinations of speed_30, severity_3(Slight), and weather_1(Fine without high winds). Similar patterns in confidence, lift, leverage, and conviction values can be observed, indicating varying levels of association between these variables.

In summary, these association rules provide insights into relationships between different variables. High confidence values suggest strong relationships, while lift values above 1 suggest positive associations. Leverage and conviction values further provide information about the co-occurrence of variables.

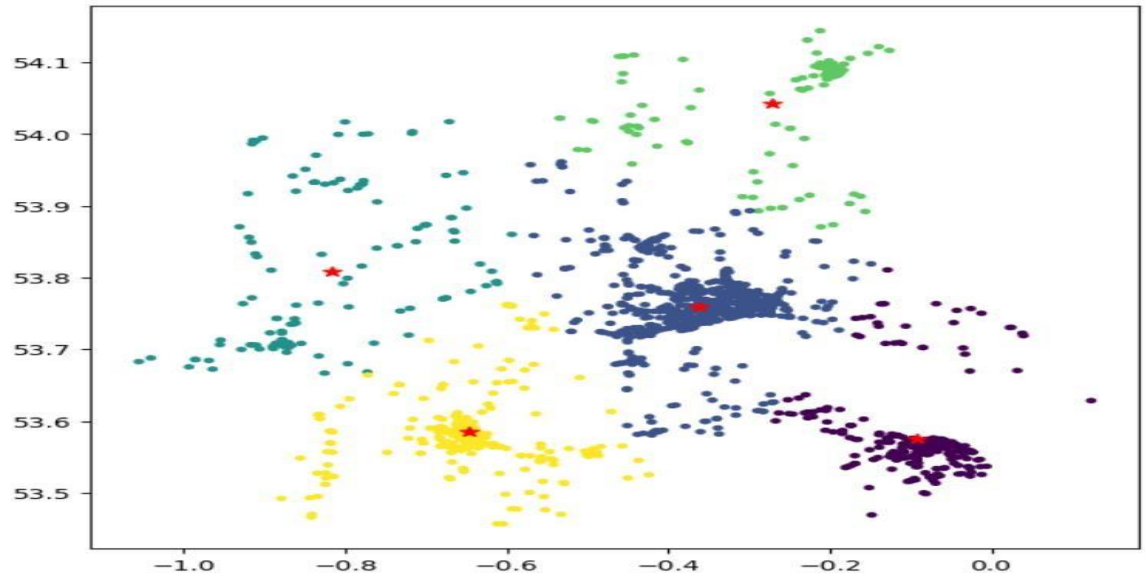


FIG 14 :
Snippet of the graph of Humber with a high inertia_ value of 7161.94 which

typically indicates that the data points are farther apart to the centroids of their respective clusters, suggesting that the clusters are not tightly packed.

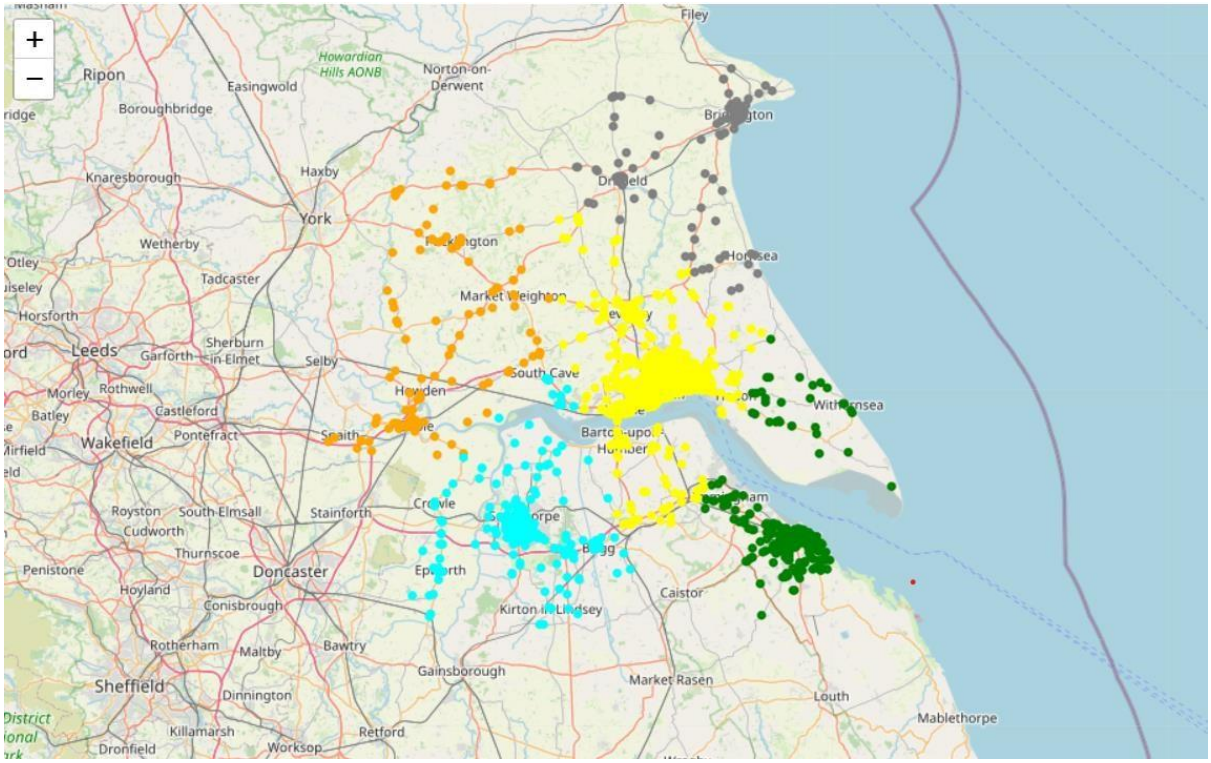
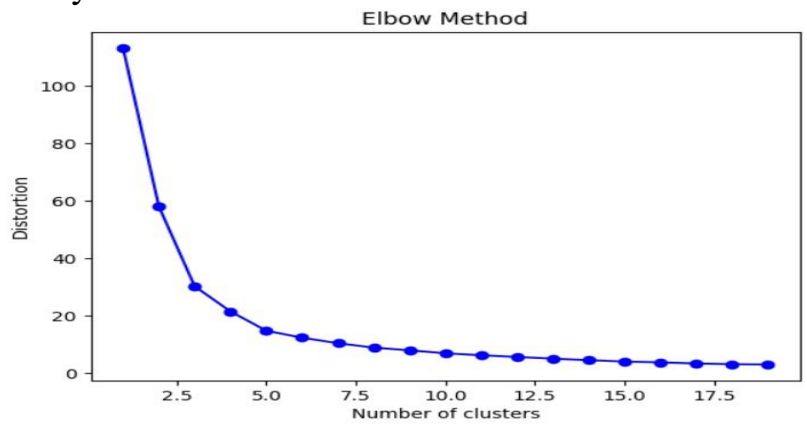


FIG 15:

The figure above depicts the map of Humberside and that there are cities that are highbrow for accidents occurrence. Scunthorpe in North Lincolnshire, Grimsby in Northeast Lincolnshire, Kingston upon hull, Hessle, Goole and Bridlington are urban areas with a high density of accidents. Accidents occurs mostly in Urban areas



	speed_limit	weather_conditions
0	30	1
1	30	1
2	30	1
3	30	1
4	30	1
...	...	...
1658	30	1
1659	20	1
1660	30	1
1661	30	1
1662	30	1

FIG 16: speed limit showing weather conditions

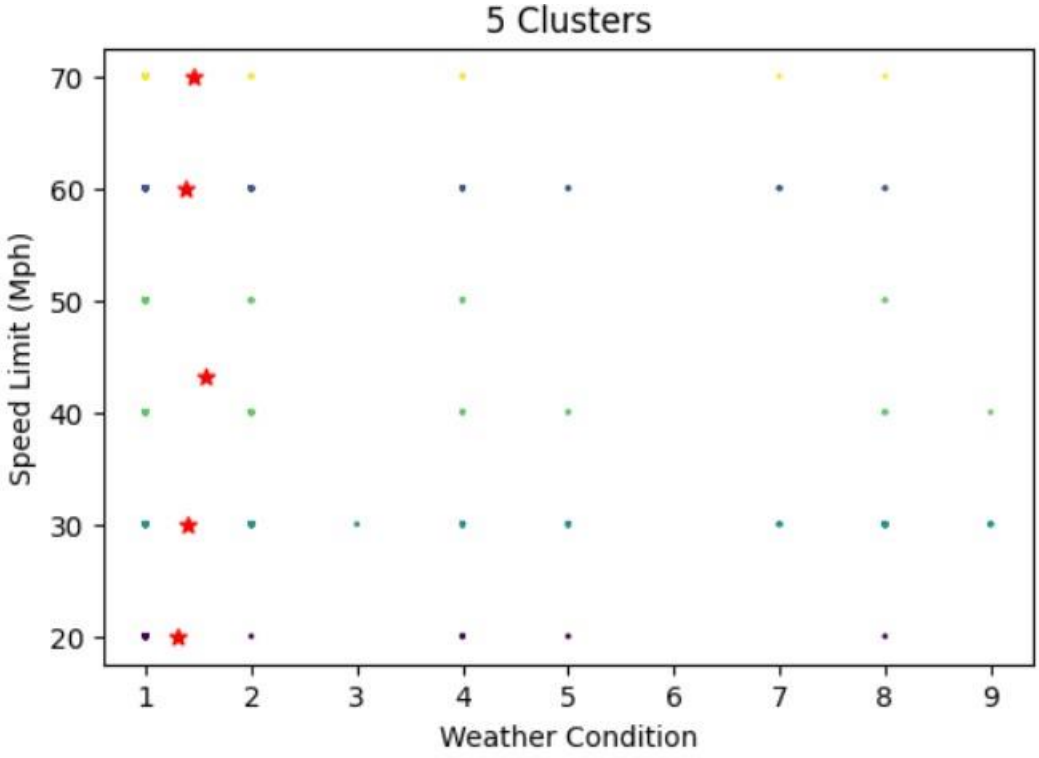


FIG 17

The graph above shows the weather conditions where speed limit is clustered. This means that there were more accidents when the weather condition was 1 (fine without high winds) and with speed limit between 30 to 70.

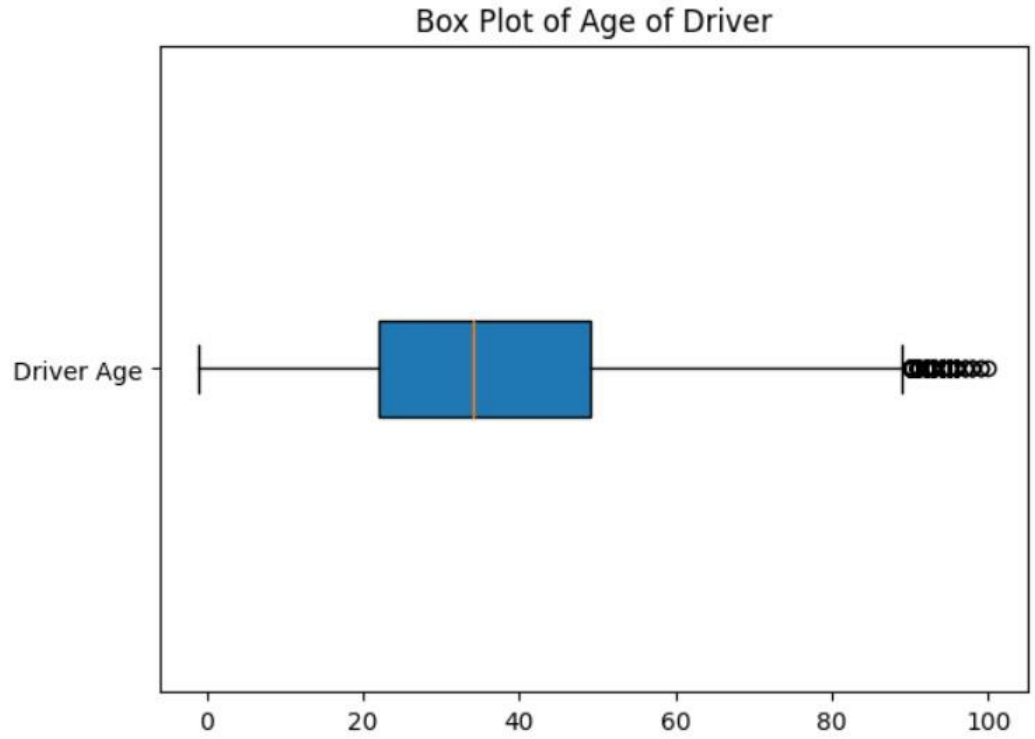


FIG 18: Graph of drivers age showing outliers

Grubbs(3.15) < Grubbs Critical(5.12), we accept the null hypothesis that there are no outliers

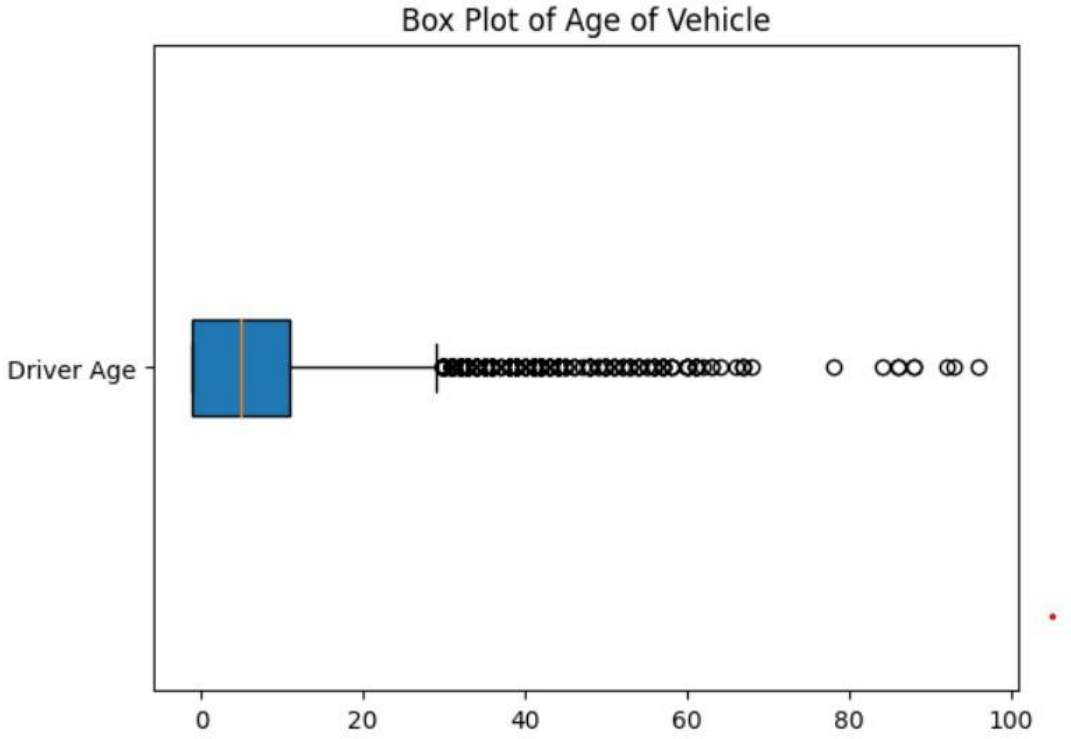


FIG 19: Graph of age of vehicle showing outliers

Grubbs(14.24) > Grubbs Critical(5.12), we reject the null hypothesis that there are no outlier and accept the alternate hypothesis that there are outliers in the data.

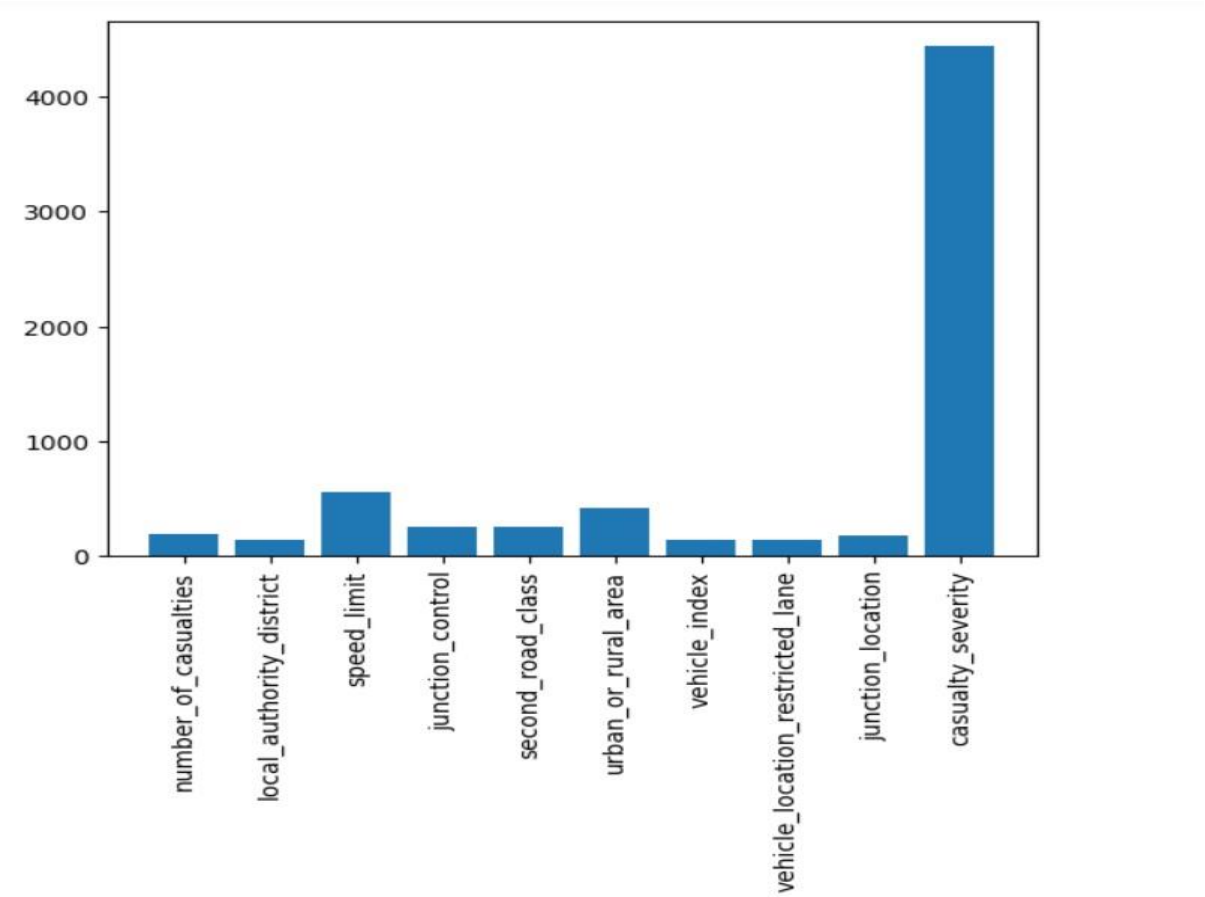


FIG 18

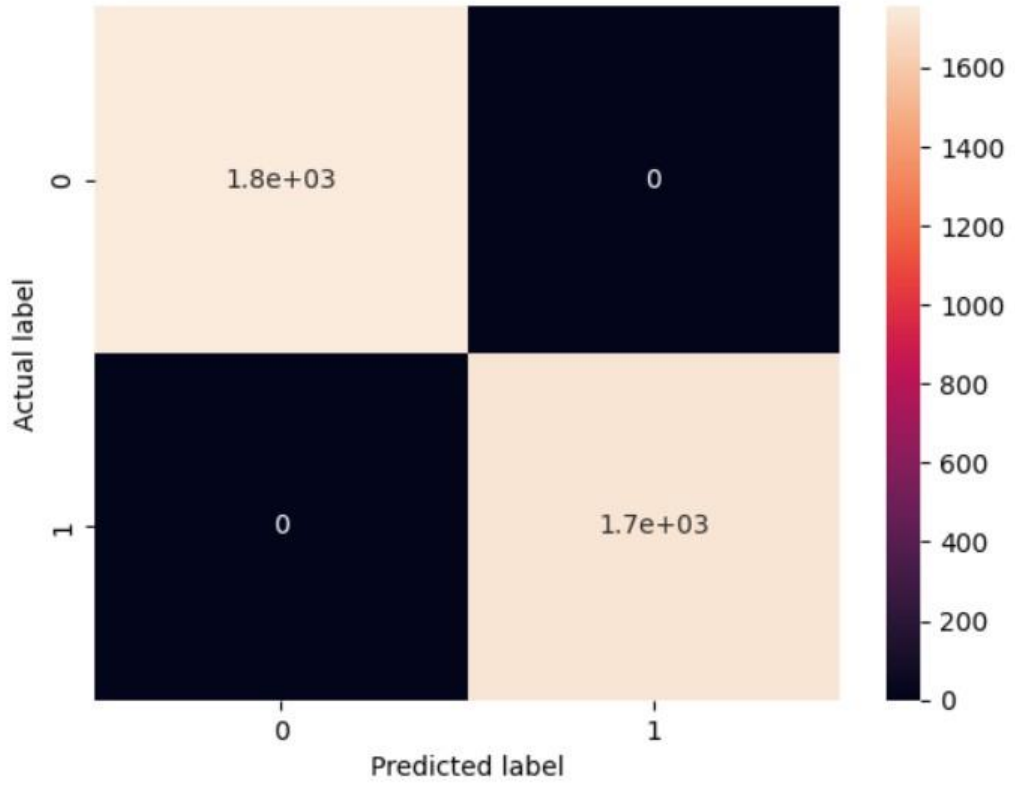
It can be seen that the feature with the best score is the casualty severity and is also the most useful in classifying whether accidents are fatal or not. it can also be seen that the speed limit and the settings of the location (whether it is a rural or urban area) and probably others.

4.0 PREDICTIONS

To carry out the predictions, accident data, casualty data and vehicles data were merged. Select k-best was used to select the best 10 features that are important in predicting outcomes, the merged data frame was split into 20% train and 80%

test data with x being the best features extracted and y (targeted column) being the casualty severity.
Multiple classification algorithms such as K-Neighbours, Decision tree, Logistic regression, Naïve bayes, Random Forest and Stacking classifier were trained using cross validation and balanced accuracy in which the results were displayed inbox plot as shown below.

FIG 19



The model has achieved high precision, recall, and F1-score for both classes, indicating strong performance in classifying both "False" (Non-fatal casualties) and "True" (Fatal casualties). The accuracy of 0.96 suggests that the model is performing well overall.

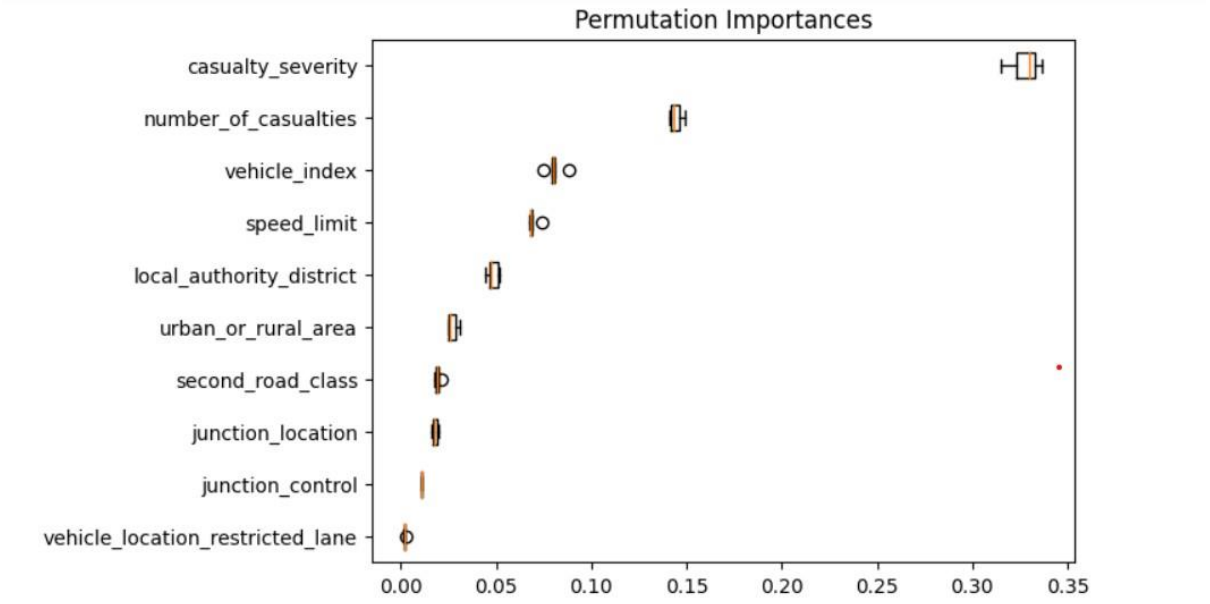


FIG 20

The graph above shows that casualty severity, number of casualties, and speed limit are significance features on road accidents.

5. RECOMMENDATIONS

1. It can be recommended that government authorities consider constructing new roads to alleviate traffic congestion in heavily populated areas and improve road visibility through initiatives such as enhanced road signs and markings.
2. Insights derived from factors like vehicle age, driver age, pedestrian movement, and accident locations hold significant value for policymakers. These insights can help lawmakers understand the underlying causes of accidents and enable them to revise policies to address these issues effectively.
3. To mitigate road traffic accidents, strategies such as notifying road users about accident-prone locations like hazardous intersections and implementing programs to enhance driver behaviour should be implemented.
4. Cities identified through the dot map distribution as having a higher concentration of accidents should be thoroughly investigated to determine infrastructure needs and potential solutions.
5. Government efforts should focus on educating and promoting safety awareness among all road users. Additionally, advocating for safety features in vehicles can contribute to accident reduction and prevention.