Big Data Project Report

Introduction

This study analyses a database of UK road accident information to examine how to increase road safety. By analysing trends in road accidents, vehicle accidents, casualty involvement, the effect of factors on accident severity and others, I would look at the hidden dynamics of traffic accidents. The objective is to unearth insights that inform safety measures and forecast fatal injuries, resulting in safer roads for our communities. Techniques like association rule mining, classification model and clustering algorithms will be used to accomplish this. Finally, I would offer suggestions to government organisations on how to increase road safety.

Analysis

Question 1

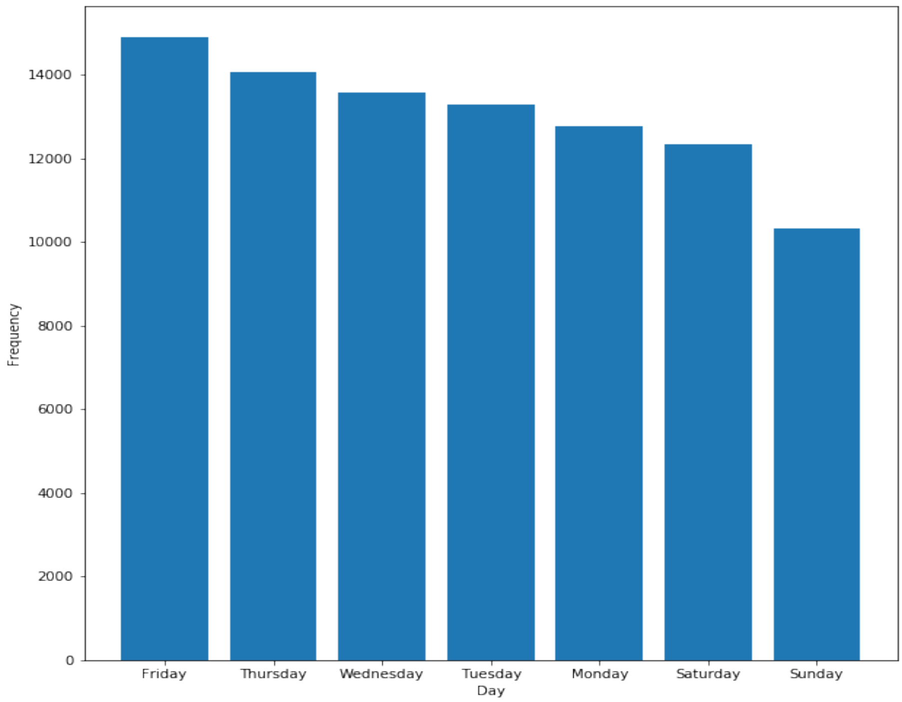

Figure 1a- significant days of the week of the day accidents occur

As seen in Figure 1a, accidents occur significantly more often on Fridays compared to other days of the week. This trend could be attributed to Friday being the end of the workweek, leading to an increased number of people on the road as they travel to their respective destinations for the weekend.

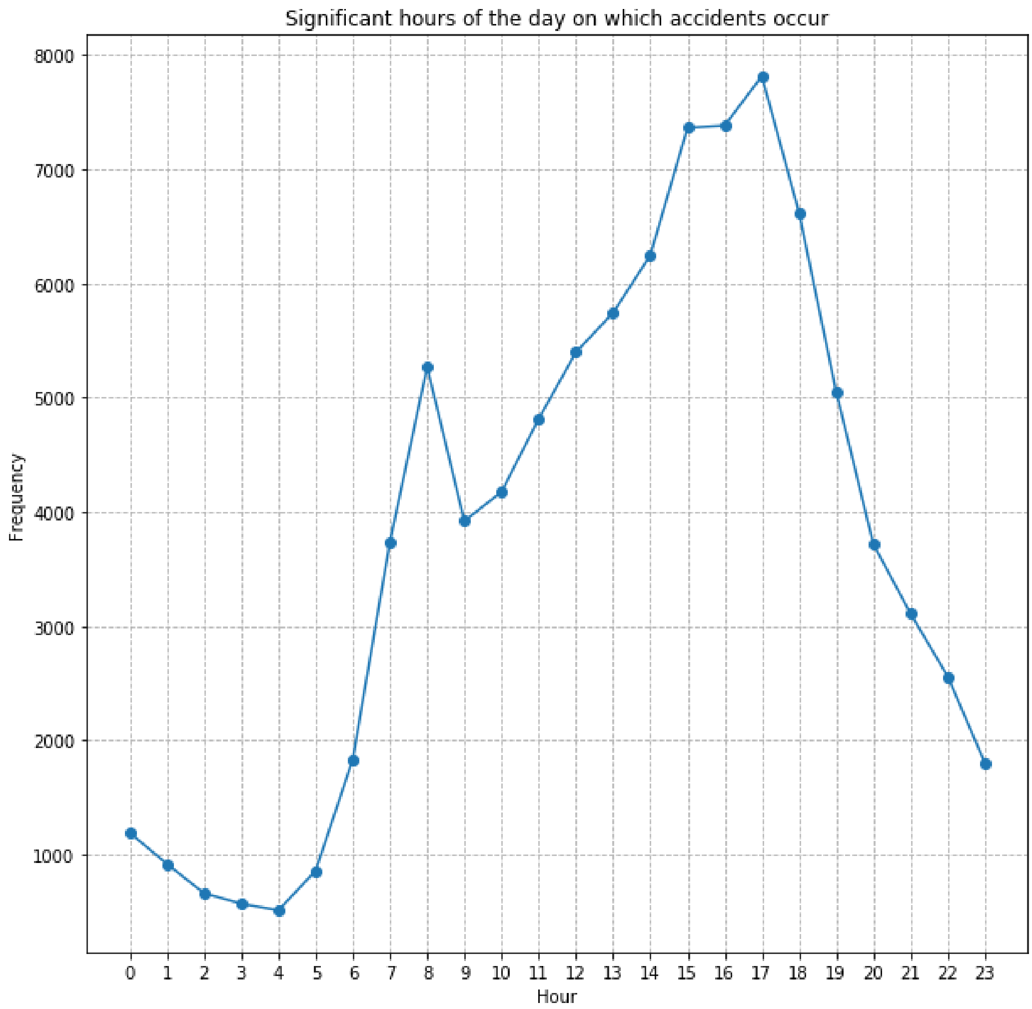

Figure 1b- Significant hours of the day on which accidents occur

As illustrated in Figure 1b, a significant number of accidents occur around 17:00, likely due to the increased rush-hour traffic as people leave work. Additionally, there is a sharp increase in accidents between 7:00 and 8:00, corresponding to the time when many people are on the road to commute to offices, school, or business.

Question 2

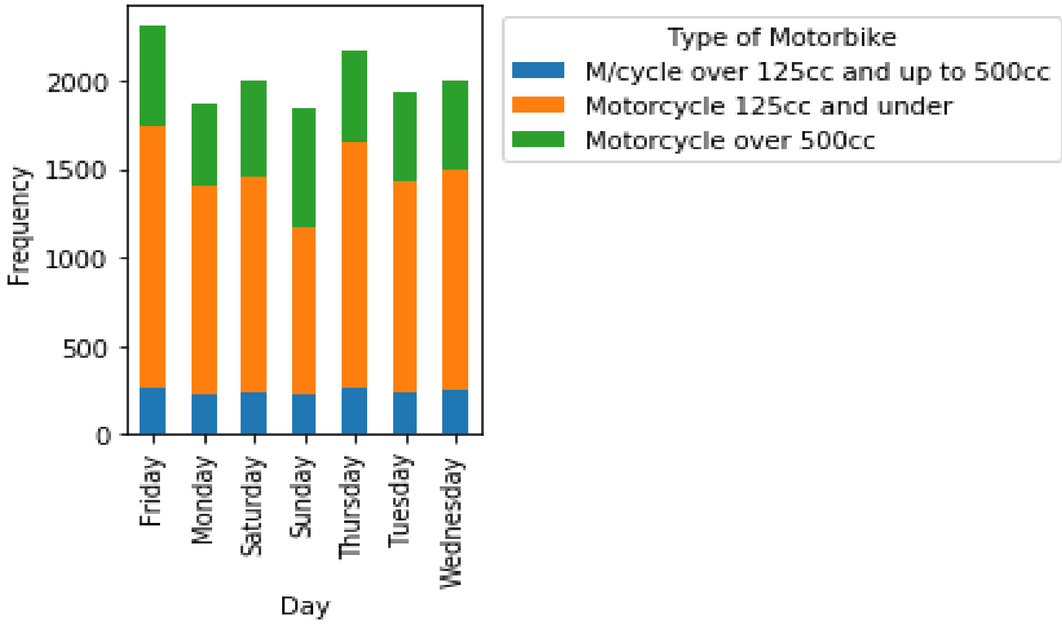

Figure 2a- significant days of the week accidents occur for motorbikes

As illustrated in Figure 2a, motorbike accidents occur most frequently on Fridays, with motorcycles 125cc and under being involved in the majority of these incidents during weekdays. Motorcycles 125cc and under are the least expensive type of motorbikes, making them a popular choice among riders. The spike in accidents on Fridays may be attributed to the end of the workweek, when roads are busier as people head to their various weekend destinations.

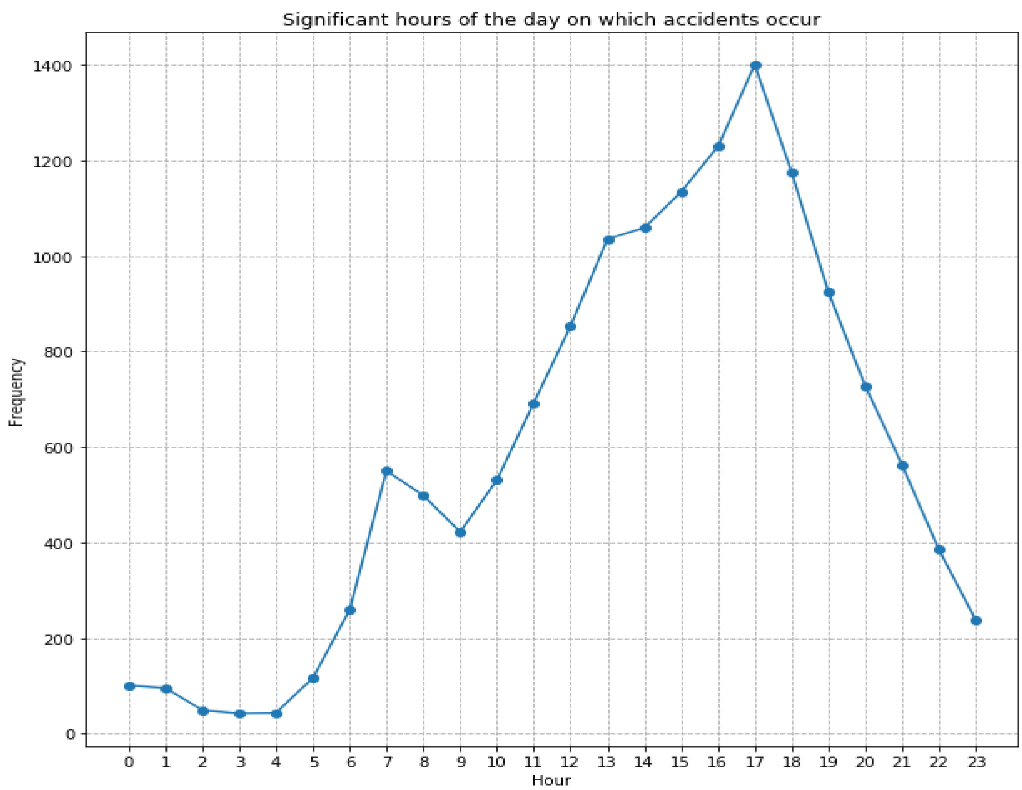

Figure 2b- Significant hours of the day on which motorbikes accidents occur

My analysis, as illustrated in Figure 2b, reveals that motorbike accidents occur more frequently around 17:00, a time that coincides with the closing of offices and businesses.

Question 3

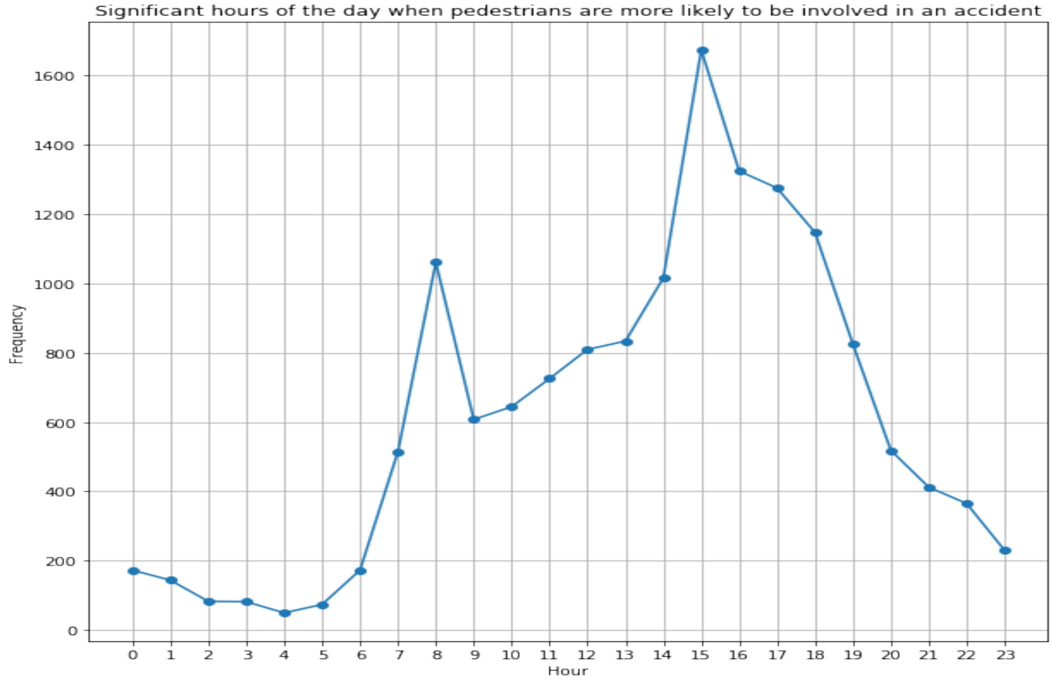

Figure 3a- Significant hours of the day when pedestrians are more likely to be involved in an accident

As illustrated in Figure 3a, 15:00 is a critical time for pedestrian accidents in comparison to other hours of the day. This time coincides with the common dismissal hour for many schools. Consequently, there is an increased number of children and teenagers on the roads, either walking home or waiting for public transportation. This surge in pedestrian activity could contribute to a higher likelihood of accidents involving this vulnerable age group.

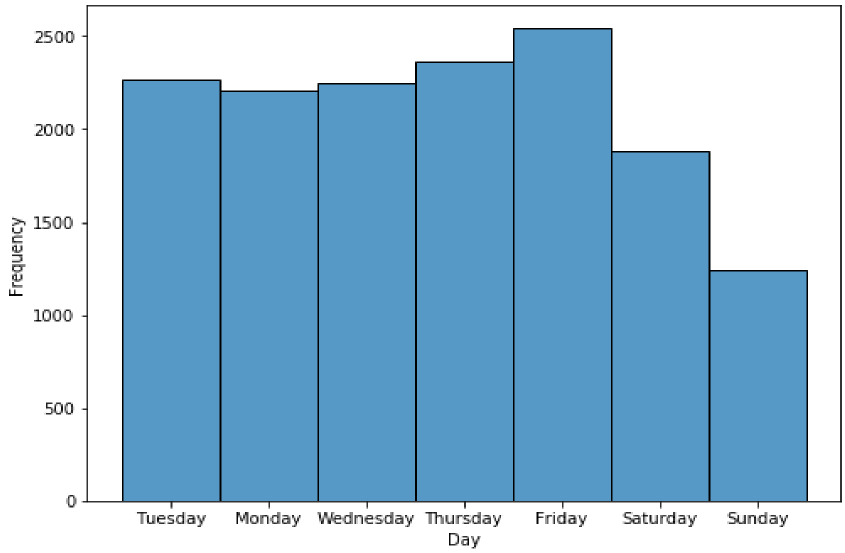

Figure 3b-Significant days of the week pedestrians more likely to be involved in accident

Figure 3b reveals that pedestrians are most likely to be involved in accidents on Fridays and Thursdays. This trend may be due to increased traffic from weekend commuting and shopping, as well as social activities that often escalate towards the week's end. The rise in alcohol consumption during these days may also impair judgment and reaction times for both drivers and pedestrians, contributing to the likelihood of accidents.

Question 4

The apriori algorithm is like a detective tool in data analysis, helping us spot common patterns or trends from heaps of data. When we talk about understanding the factors that might cause severe accidents, this tool helps us see how different elements, like weather or the type of road, often come together in certain accident scenarios. By using the apriori algorithm, we're essentially looking for clues or patterns that tell us more about what might make an accident severe.

<u>Data Exploratory and cleaning</u>

7 Features (accident severity, casualty class, vehicle type(motorbikes), weather conditions, road type, speed limit, road surface conditions, light conditions) were selected to investigate their impact on accident severity, based on their identification as contributory factors to accidents in the STAT20 document (page 113). These features were compiled into a Data Frame, df3, where an incorrect entry of -1 in the road surface conditions column was detected and replaced with the integer mean of that column.

<u>Analyses</u>

After cleaning the data, one-hot encoding was applied to each feature, and the encoded features were concatenated into a new data frame, df4. An apriori algorithm was then run on df4, using a 30% minimum support threshold to identify common feature combinations. Association rules were applied to uncover relationships between these features, focusing on their antecedents and consequents, along with metrics like lift.

The resulting data frame, 'data', was filtered to examine the connections between specific feature combinations (antecedents) and types of accident severity (consequents). The analysis revealed the most influential feature combinations affecting accident severity, identified by the highest lift values.

Table 4.1

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
6	(vehicle_type_3, casualty_class_1)	(severity_3)	0.489403	0.660765	0.369217	0.754423	1.141742	0.045836	1.381379	0.243137
56	(weather_conditions_3, casualty_class_1)	(severity_3)	0.489403	0.660765	0.369217	0.754423	1.141742	0.045836	1.381379	0.243137
60	(casualty_class_1, road_type_3)	(severity_3)	0.489403	0.660765	0.369217	0.754423	1.141742	0.045836	1.381379	0.243137
84	(vehicle_type_3, weather_conditions_3, casualty_class_1)	(severity_3)	0.489403	0.660765	0.369217	0.754423	1.141742	0.045836	1.381379	0.243137
96	(vehicle_type_3, casualty_class_1, road_type_3)	(severity_3)	0.489403	0.660765	0.369217	0.754423	1.141742	0.045836	1.381379	0.243137
200	(weather_conditions_3, casualty_class_1, road_type_3)	(severity_3)	0.489403	0.660765	0.369217	0.754423	1.141742	0.045836	1.381379	0.243137
229	(weather_conditions_3, vehicle_type_3, casualty_class_1, road_type_3)	(severity_3)	0.489403	0.660765	0.369217	0.754423	1.141742	0.045836	1.381379	0.243137

Table 4.1 offers insights into the relationships between various combinations of factors such as vehicle type, weather conditions, casualty class, and road type, and their association with a specific severity level in accidents. It uncovers patterns that could be crucial in formulating strategies to mitigate risks tied to these factors. From the table, we can observe that all possible combinations of features lead to a ***slight accident severity*** (severity_3).

Question 5

I use a SQL query retrieve accident data for the year 2020 from regions under Humberside which includes North Lincolnshire, Kingston upon Hull, North East Lincolnshire, and East Riding of Yorkshire. The result includes the latitude, longitude, local area code (lsoa01nm), police force, and accident year, which is then converted into a DataFrame as see in Table 5.1 below

	latitude	longitude	lsoa01nm	police_force	accident_year	location
0	53.744936	-0.393424	Kingston upon Hull 028E	16	2020	Kingston upon Hull
1	53.512895	-0.528743	North Lincolnshire 022C	16	2020	North Lincolnshire
2	53.791630	-0.324858	Kingston upon Hull 002E	16	2020	Kingston upon Hull
3	53.574501	-0.095008	North East Lincolnshire 003C	16	2020	North East Lincolnshire
4	53.767805	-0.327733	Kingston upon Hull 016D	16	2020	Kingston upon Hull
...	...	...	...	...	...	...
1658	53.566753	-0.651104	North Lincolnshire 017B	16	2020	North Lincolnshire
1659	53.839482	-0.424674	East Riding of Yorkshire 019D	16	2020	East Riding of Yorkshire
1660	53.782750	-0.308880	Kingston upon Hull 007C	16	2020	Kingston upon Hull
1661	53.569801	-0.703181	North Lincolnshire 005A	16	2020	North Lincolnshire
1662	53.742609	-0.342063	Kingston upon Hull 029C	16	2020	Kingston upon Hull

1663 rows × 6 columns

Table 5.1- accident location data frame

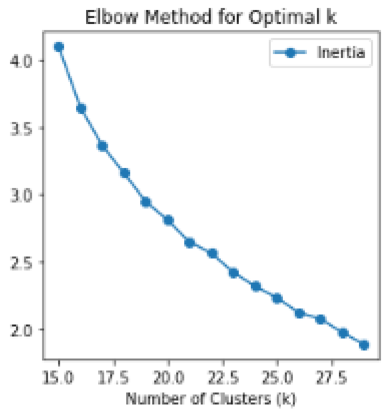

Figure 5a- Elbow Method for Optimal k

I clustered the geographic coordinates (latitude and longitude) using the K-Means algorithm, determining the optimal number of clusters (21 clusters) by analyzing the inertia of different cluster numbers, and performing the final clustering with the optimal number of clusters. Below in Figure 5b is the result of my lustering on a map

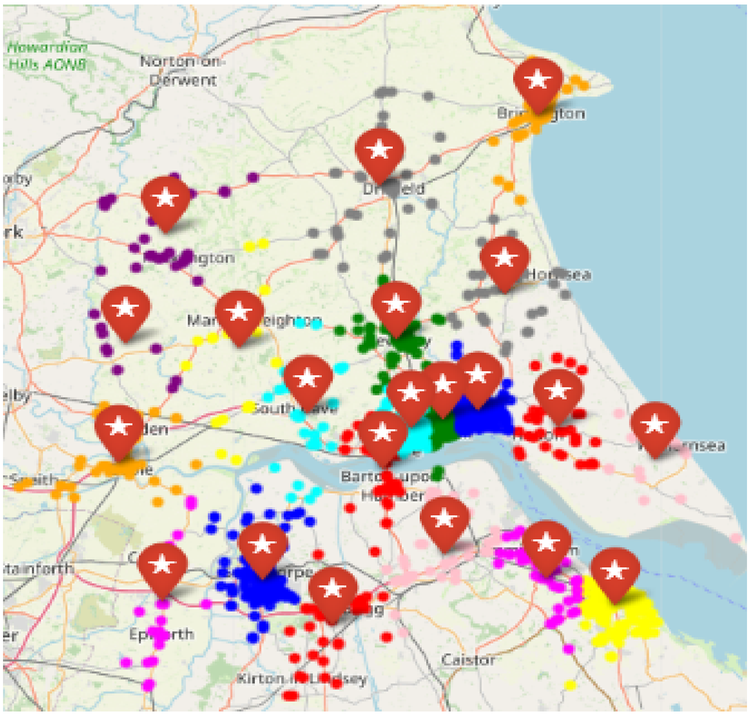

Figure 5b- clusters of accidents around the 4 regions in Humberside

A closer zoom into the map (Figure 5c) to ascertain the location where most clusters are been formed around Humberside shows we are seeing more accidents in Kingston Upon Hull as compared to other 3 locations. I noticed accidents clusters in Hull were around the city enter and major roads coming into Hull

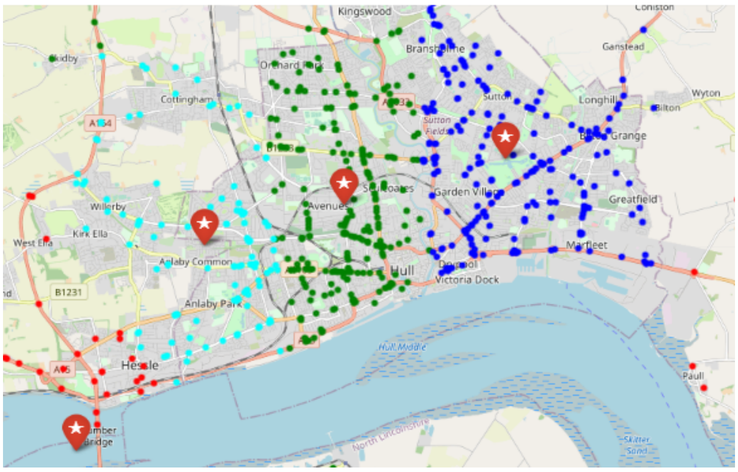

Figure 5c- Map of Hull

I also did a K-Mediods algorithm where I grouped the geographic locations (based on longitude and latitude) into 21 different clusters using a method called k-Medoids. Unlike k-Means, which finds the average point of each cluster, k-Medoids picks an actual location from the dataset to represent each cluster's center. This makes it more reliable when dealing with unusual or outlier data points. Here is the result of my clustering on the map below (Figure 5d)

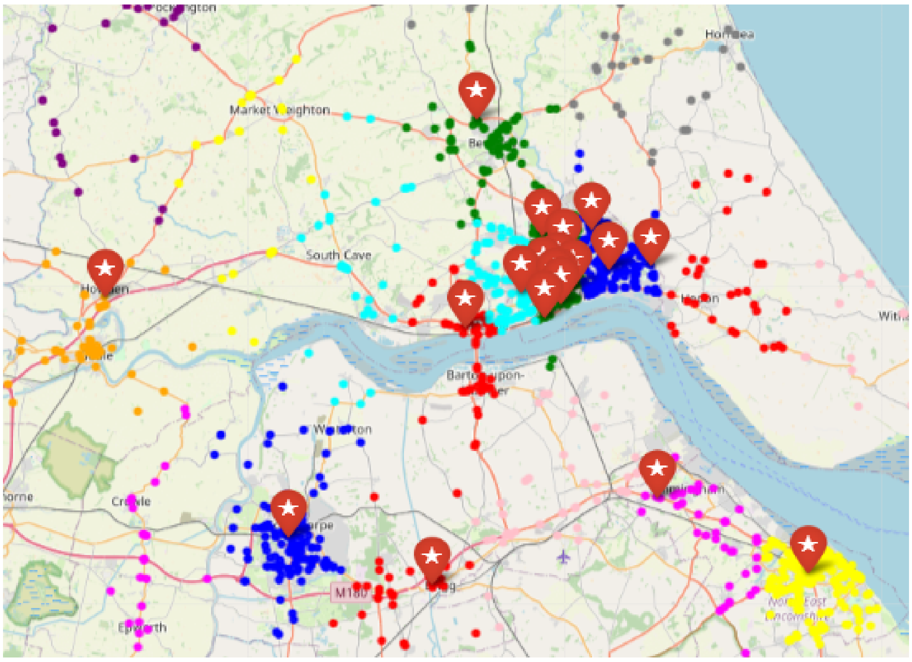

Figure 5d- Map of Humberside

A zoom into the map to ascertain where lots of accident clusters are been formed shows most accidents around Humberside region are clustered in Hull compared to other 3 locations as seen in Figure 5e. Most of the accidents in Hull are around the city center and major roads coming into Hull.

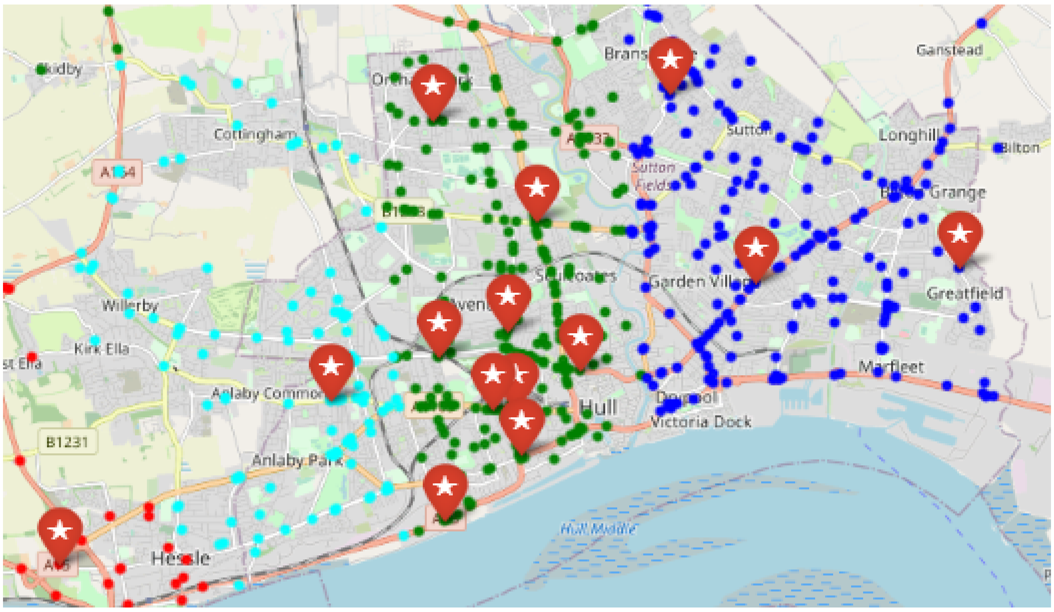

Figure 5e- Map of Kingstone Upon Hull

I then ran a silhouette scores algorithm to evaluate the 2 different clustering methods (K-Means and K-Medoids) for grouping geographic coordinates into 21 clusters. The result as seen in Figure 5f shows that K-Means was more effective in clustering the accident data around the Humberside region compared to K-Mediods

```
Silhouette Score for K-Means: 0.47941504274994245
Silhouette Score for K-Medoids: 0.32943194672736215
```

Figure 5f- silhouette score for the 2 algorithms

Question 6

A connection to the SQLite database containing accident data was established, and all accidents from 2020 were gathered into a data frame as seen in Table 6.1.

	accident_index	accident_year	accident_reference	location_easting_osgr	location_northing_osgr	longitude	latitude	police_force	accident_severity	nu
0	2020010219808	2020	010219808	521389.0	175144.0	-0.254001	51.462262	1	3	
1	2020010220496	2020	010220496	529337.0	176237.0	-0.139253	51.470327	1	3	
2	2020010228005	2020	010228005	526432.0	182761.0	-0.178719	51.529614	1	3	
3	2020010228006	2020	010228006	538676.0	184371.0	-0.001683	51.541210	1	2	
4	2020010228011	2020	010228011	529324.0	181286.0	-0.137592	51.515704	1	3	
...	...	...	...	...	...	...	...	...	...	
91194	2020991027064	2020	991027064	343034.0	731654.0	-2.926320	56.473539	99	2	
91195	2020991029573	2020	991029573	257963.0	658891.0	-4.267565	55.802353	99	3	
91196	2020991030297	2020	991030297	383664.0	810646.0	-2.271903	57.186317	99	2	
91197	2020991030900	2020	991030900	277161.0	674852.0	-3.968753	55.950940	99	3	
91198	2020991032575	2020	991032575	240402.0	681950.0	-4.561040	56.003843	99	3	

91199 rows × 36 columns

Table 6.1- accident table

I then identified rows where the 'junction_detail' was 0, set the corresponding 'junction_control' values to an empty string in line with a STATS20 document's instruction (Page 27, Note A). Incorrect entries (9 and 99) in the 'junction_control' and 'junction_detail' columns were replaced with NaN values.

I also dropped the 'second_road_number' column, as filling its many -1 entries (41% of its value entry) with the column's mean would create bias. All remaining NaN values in the data Frame were replaced with their respective column means to eliminate null values.

Finally, an Isolation Forest model was fitted to the numerical data to detect anomalies, and outliers were identified and printed as seen in Figure 6b. The Isolation Forest algorithm successfully detected -1 as unusual entries in 71% of the columns containing such values across the entire data frame.

```
['local_authority_district', 'speed_limit', 'second_road_class', 'pedestrian_crossing_human_control', 'pedestrian_crossing_phys
ical_facilities', 'road_surface_conditions', 'special_conditions_at_site', 'carriageway_hazards', 'trunk_road_flag', 'iforest_o
utlier']
```

Figure 4b- columns with outliers

All occurrences of -1 were replaced with the mean of each respective column, excluding the longitude and latitude columns, following the detection of -1 as an unusual entry by the Isolation Forest algorithm

This process was undertaken to ensure the integrity and accuracy of the dataset.

Predictions

The cleaned accident data frame from question 6 was used to continue my analysis as seen in Table 7.1

	accident_year	location_easting_osgr	location_northing_osgr	longitude	latitude	police_force	accident_severity	number_of_vehicles	number_of_casua
0	2020	521389.0	175144.0	-0.254001	51.462262	1	3	1	
1	2020	529337.0	176237.0	-0.139253	51.470327	1	3	1	
2	2020	526432.0	182761.0	-0.178719	51.529614	1	3	1	
3	2020	538676.0	184371.0	-0.001683	51.541210	1	2	1	
4	2020	529324.0	181286.0	-0.137592	51.515704	1	3	1	
...	...	...	...	...	...	...	...	...	
91194	2020	343034.0	731654.0	-2.926320	56.473539	99	2	2	
91195	2020	257963.0	658891.0	-4.267565	55.802353	99	3	1	
91196	2020	383664.0	810646.0	-2.271903	57.186317	99	2	2	
91197	2020	277161.0	674852.0	-3.968753	55.950940	99	3	2	
91198	2020	240402.0	681950.0	-4.561040	56.003843	99	3	1	

91199 rows × 27 columns

Table 7.1- Data Frame for analysis

A boolean label was created to mark accidents with fatal severity as true, and the occurrences of fatal and non-fatal accidents were counted. As seen in Figure 7a and 7b respectively.

```
0        False
1        False
2        False
3        False
4        False
         ...
91194    False
91195    False
91196    False
91197    False
91198    False
Name: accident_severity, Length: 91199, dtype: bool
```

Figure 7a- A boolean label showing Fatal accidents as true and non-Fatal as False

```
False    89808
True      1391
Name: accident_severity, dtype: int64
```

Figure 7b- counts of occurrences of fatal and non-fatal accidents before resampling

The dependent columns were dropped (accident_severity & did police officer attend scene of accident), leaving only the required features. The unbalanced distribution of fatal and non-fatal injuries was then resampled using Random Under Sampler, and the counts of fatal and non-fatal accidents were repeated after resampling as seen in Figure 7c.

```
True     1391
False    1391
Name: accident_severity, dtype: int64
```

Figure 7 c - counts of occurrences of fatal and non-fatal accidents after resampling

Using the K-Best method, we identified the most influential features affecting severity, as depicted in Figure 7d. These features have been compiled into a new dataframe.

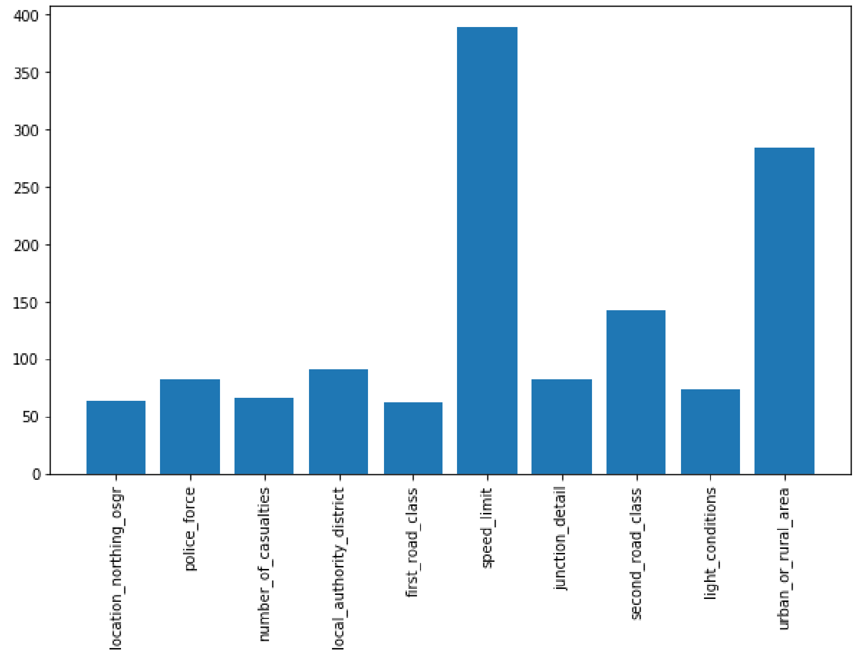

Figure 5d- Feature selection for accident severity

A decision tree classifier was trained using the feature selection data frame with a minimum of 100 samples per leaf. Features and labels were split into training and testing sets, and the tree model was fitted and tested, with a classification report being printed after 10-fold cross-validation.

	precision	recall	f1-score	support
False	0.71	0.60	0.65	417
True	0.65	0.75	0.70	418
accuracy			0.68	835
macro avg	0.68	0.68	0.67	835
weighted avg	0.68	0.68	0.67	835

Figure 7e- classification report of Decision Tree classifier

Several models, including a stacking classifier, were defined and evaluated using Repeated Stratified K-Fold cross-validation. Finally, the models' accuracies were compared and plotted in a boxplot, summarizing the analysis.

```
dtree:0.60 0.030
knn:0.63 0.036
bayes:0.67 0.026
rf:0.65 0.025
stacking:0.68 0.026
```

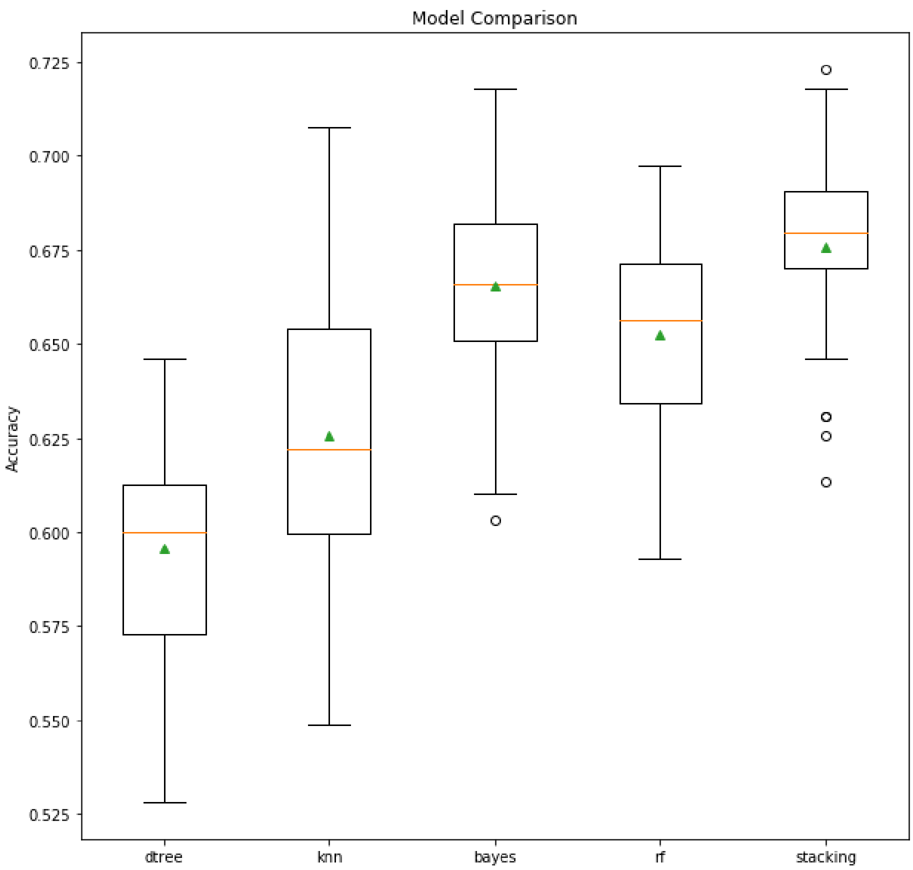

Figure 7f- Models accuracy

Overall, both the Bayes and the Stacking Classifier performed well in classifying fatal injuries, but the Stacking Classifier delivered the highest accuracy, making it potentially the most suitable choice for predicting fatal injuries sustained in road traffic accidents.

Recommendations to the Government

- **Monitor Peak Traffic Hours**: Perhaps an increase in traffic police presence during the busy times of 17:00 and between 7:00-8:00 would help regulate traffic flow. We all know rush hours can be chaotic.
- **Focus on Motorcyclists**: It's evident that motorbike accidents are a concern, especially with motorcycles 125cc and under. Targeted safety campaigns might make a real difference here.
- **Educate the Young Ones**: Kids and teenagers could benefit from learning about road safety in school. Let's help them understand how to stay safe on the road.
- **Enhance Safety in the region with the highest accident cluster (Hull)**: Though most accidents aren't severe, we can't be complacent. Improving infrastructure, signage, and lighting, along with enforcing speed limits around Hull's city center and major roads, could make those numbers even lower.
- **Stay Informed and Adapt**: Continuing to collect and analyze accident data helps us stay ahead of the curve. We can monitor trends and adapt safety measures accordingly. Let's keep our eyes on the road and hands on the wheel of safety planning!