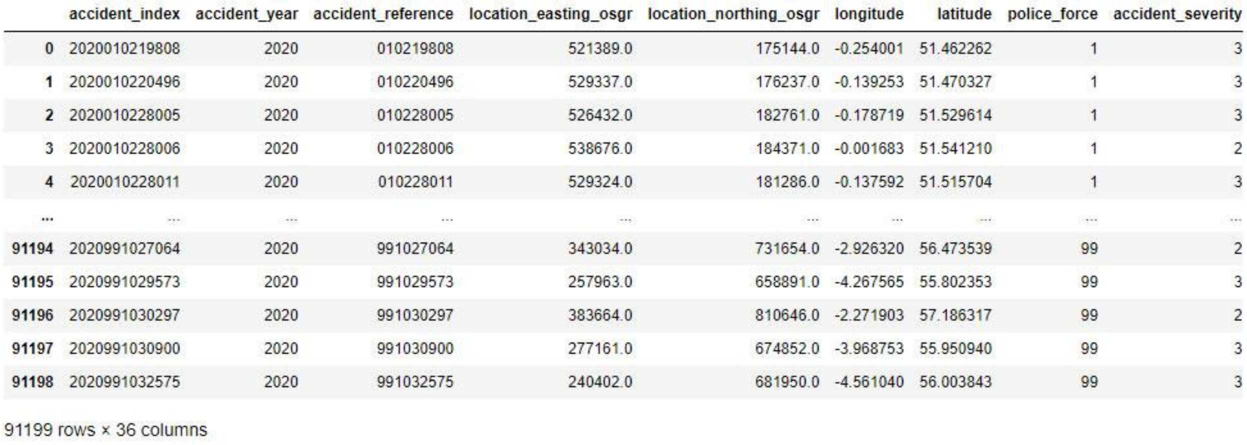
Big Data Project Report

This report examines a database of road traffic accidents in Great Britain in the year 2020 and investigates the relationships between specific factors and accident occurrence as well as accident severity. The insight gained from this analysis is subsequently utilized to provide recommendations to government agencies on enhancing road safety. Additionally, the findings are used to develop a model for predicting fatalities.

Overview of Data

The accident data was provided as an SQL relational database consisting of four tables: Accident, Vehicle, Casualty and LSOA. These tables were read individually into pandas dataframes, selecting only occurrences in the year 2020. An overview of the four dataframes is given below.

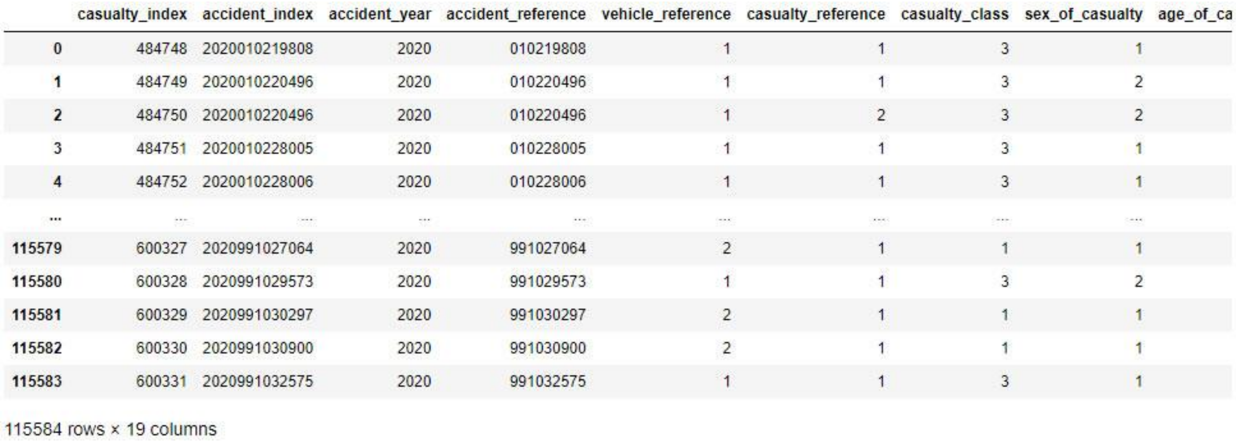
Accident table



	accident_index	accident_year	accident_reference	location_easting_osgr	location_northing_osgr	longitude	latitude	police_force	accident_severity
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

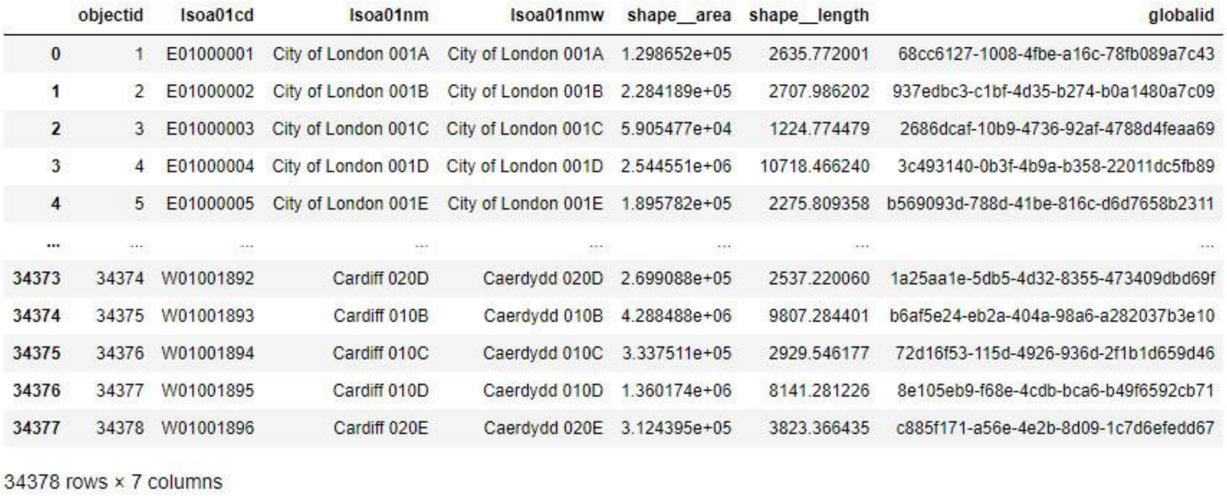
Casualty table



	casualty_index	accident_index	accident_year	accident_reference	vehicle_reference	casualty_reference	casualty_class	sex_of_casualty	age_of_ca
0	484748	2020010219808	2020	010219808	1	1	3	1	
1	484749	2020010220496	2020	010220496	1	1	3	2	
2	484750	2020010220496	2020	010220496	1	2	3	2	
3	484751	2020010228005	2020	010228005	1	1	3	1	
4	484752	2020010228006	2020	010228006	1	1	3	1	
...	...	...	...	...	...	...	...	...	
115579	600327	2020991027064	2020	991027064	2	1	1	1	
115580	600328	2020991029573	2020	991029573	1	1	3	2	
115581	600329	2020991030297	2020	991030297	2	1	1	1	
115582	600330	2020991030900	2020	991030900	2	1	1	1	
115583	600331	2020991032575	2020	991032575	1	1	3	1	

115584 rows × 19 columns

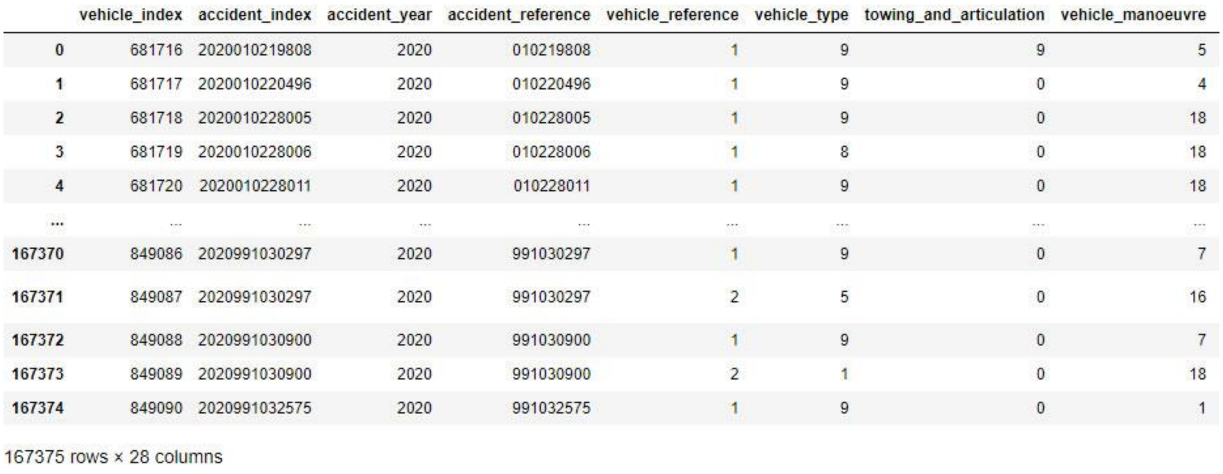
LSOA table



	objectid	lsoa01cd	lsoa01nm	lsoa01nmw	shape__area	shape__length	globalid
0	1	E01000001	City of London 001A	City of London 001A	1.298652e+05	2635.772001	68cc6127-1008-4fbe-a16c-78fb089a7c43
1	2	E01000002	City of London 001B	City of London 001B	2.284189e+05	2707.986202	937edbc3-c1bf-4d35-b274-b0a1480a7c09
2	3	E01000003	City of London 001C	City of London 001C	5.905477e+04	1224.774479	2686dcaf-10b9-4736-92af-4788d4feaa69
3	4	E01000004	City of London 001D	City of London 001D	2.544551e+06	10718.466240	3c493140-0b3f-4b9a-b358-22011dc5fb89
4	5	E01000005	City of London 001E	City of London 001E	1.895782e+05	2275.809358	b569093d-788d-41be-816c-d6d7658b2311
...	...	...	...	...	...	...	...
34373	34374	W01001892	Cardiff 020D	Caerdydd 020D	2.699088e+05	2537.220060	1a25aa1e-5db5-4d32-8355-473409dbd69f
34374	34375	W01001893	Cardiff 010B	Caerdydd 010B	4.288488e+06	9807.284401	b6af5e24-eb2a-404a-98a6-a282037b3e10
34375	34376	W01001894	Cardiff 010C	Caerdydd 010C	3.337511e+05	2929.546177	72d16f53-115d-4926-936d-2f1b1d659d46
34376	34377	W01001895	Cardiff 010D	Caerdydd 010D	1.360174e+06	8141.281226	8e105eb9-f68e-4cdb-bca6-b49f6592cb71
34377	34378	W01001896	Cardiff 020E	Caerdydd 020E	3.124395e+05	3823.366435	c885f171-a56e-4e2b-8d09-1c7d6efedd67

34378 rows × 7 columns

Vehicle table



	vehicle_index	accident_index	accident_year	accident_reference	vehicle_reference	vehicle_type	towing_and_articulation	vehicle_manoeuvre
0	681716	2020010219808	2020	010219808	1	9	9	5
1	681717	2020010220496	2020	010220496	1	9	0	4
2	681718	2020010228005	2020	010228005	1	9	0	18
3	681719	2020010228006	2020	010228006	1	8	0	18
4	681720	2020010228011	2020	010228011	1	9	0	18
...	...	...	...	...	...	...	...	...
167370	849086	2020991030297	2020	991030297	1	9	0	7
167371	849087	2020991030297	2020	991030297	2	5	0	16
167372	849088	2020991030900	2020	991030900	1	9	0	7
167373	849089	2020991030900	2020	991030900	2	1	0	18
167374	849090	2020991032575	2020	991032575	1	9	0	1

167375 rows × 28 columns

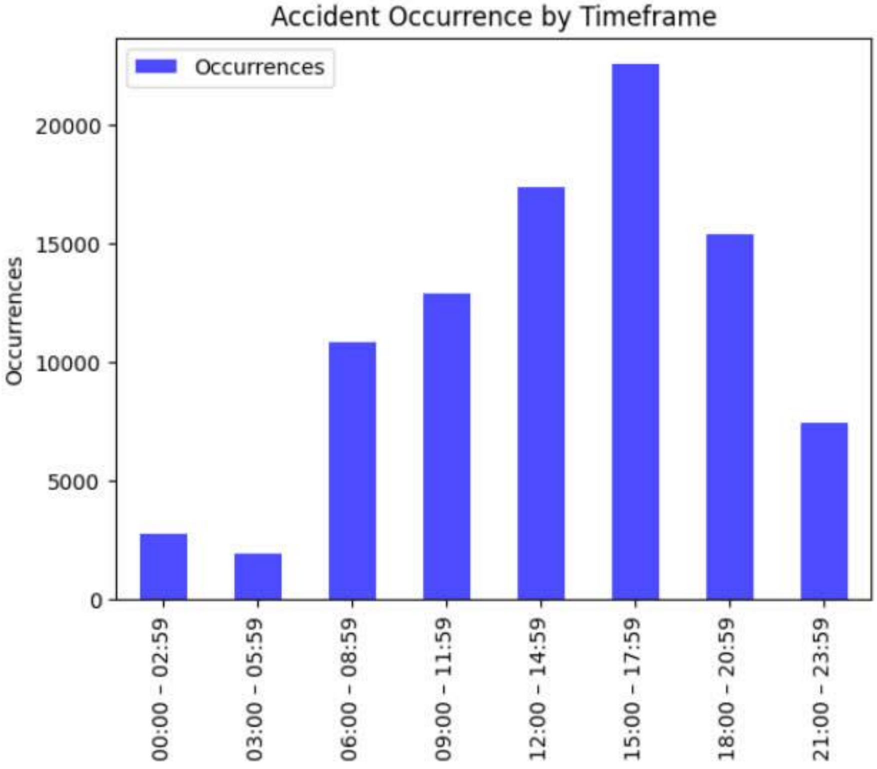
Data Cleaning

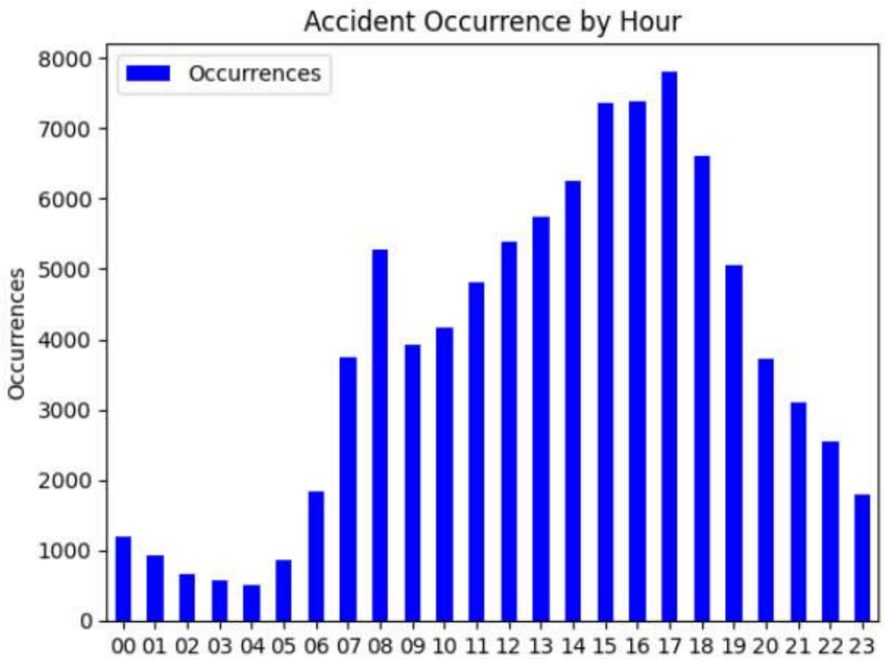
The dataframes were cleaned to address discrepancies and missing values. The cleaning primarily involved iterating through the dataframes, replacing erroneous and missing values with the modal values or random selections from a list of values associated with entries in other columns of the same row. These imputations considered factors such as geographical location, first road number, first road class, time of day, and vehicle type. Standard values provided in the supporting document were also utilized for certain imputations. In cases where insightful imputations could not be made, missing values were simply imputed with “Unknown” for columns containing string values and 0 for columns containing numerical values between 1 and 20. Step by step details of the cleaning done can be found in the python notebook file attached to this report.

Analysis of Accident Occurrence by Hours of the Day

To perform this analysis, the 24-hour day was split into 3-hour timeframes as illustrated below.

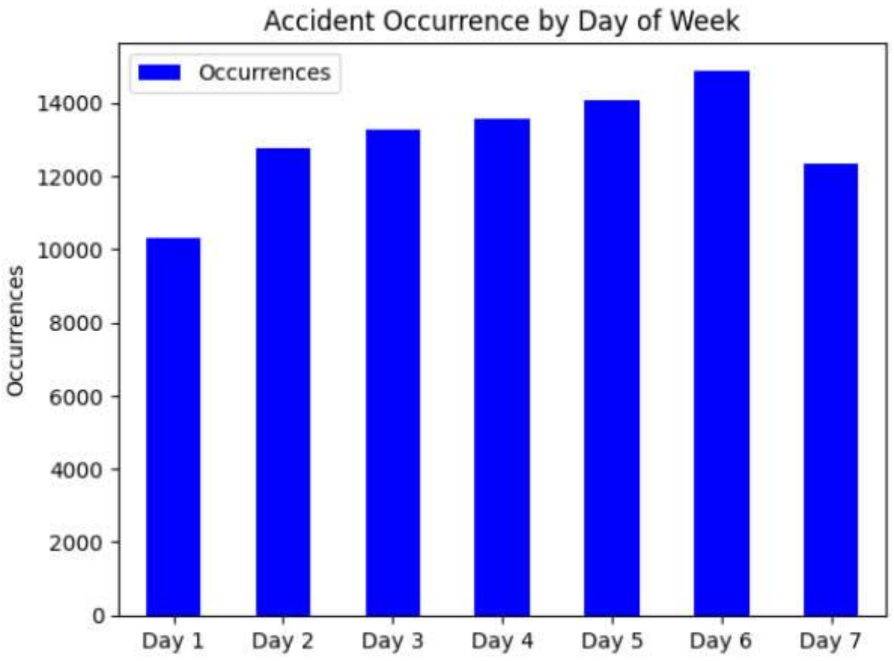
The accident count for each timeframe was obtained using SQL queries and it was determined that the period between 15:00 and 17:59 exhibited the highest frequency of accidents. Accident counts for the individual hours of the day were also obtained and it was determined that the 17th hour exhibited the highest frequency of accidents. The distribution of accidents across the hours of the day followed a sinusoidal pattern, with a gradual increase in the number of accidents from 6am up until the peak period after which the numbers begin to decline into the following day. This distribution is shown below.





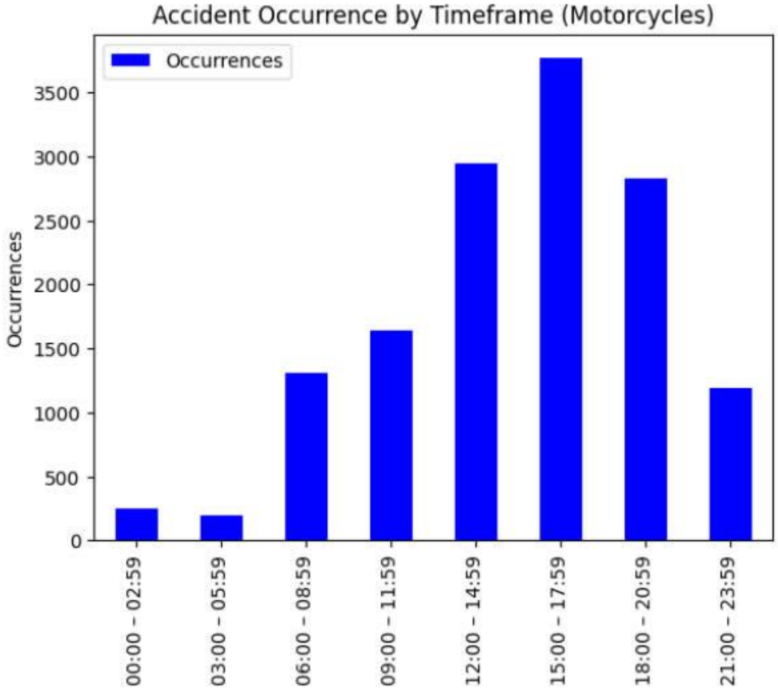
Analysis of Accident Occurrence by Days of the Week

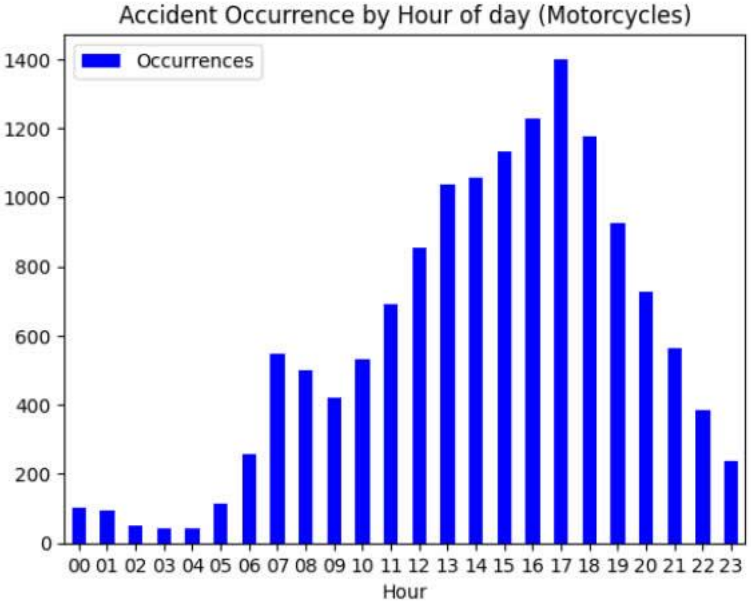
The accident counts for every day of the week was obtained using SQL queries. It was determined that day 6 had the highest number of accidents. Naturally, it is expected that every day of the week will have different value counts, with one day having the highest. The coefficient of variation for the accident counts across the week was computed and found to be 10.37% suggesting a balanced distribution (INSEE, 2016). This distribution is shown below.

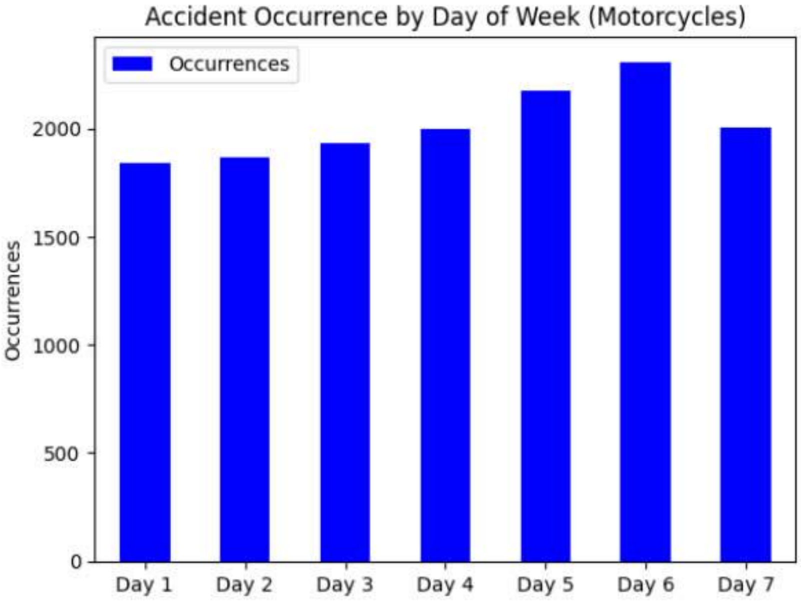


Motorcycles

The analyses done above were repeated for accidents involving specific motorcycle types. It was discovered that the highest number of accidents involving these motorcycle types occurred between 15:00 and 17:59. An hour-specific analysis showed that the 17th hour had the highest frequency of accidents for these motorcycle types. Day 6 had the highest number of accidents. The distribution of accidents across the days of the week for motorcycles was also balanced as the coefficient of variation was found to be just 7.71%. The distribution of accidents involving the three motorcycle types across the time quadrants, individual hours of the day and days of the week is given below.

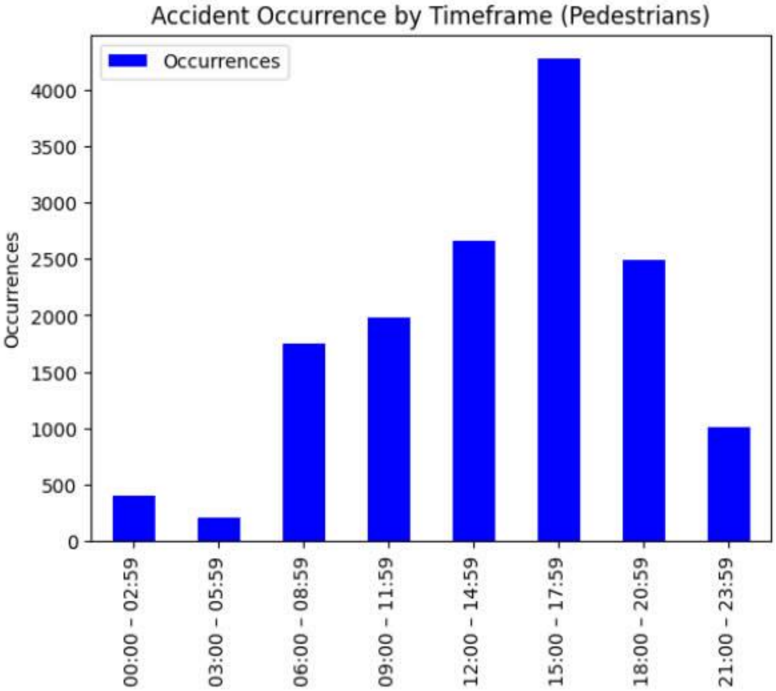


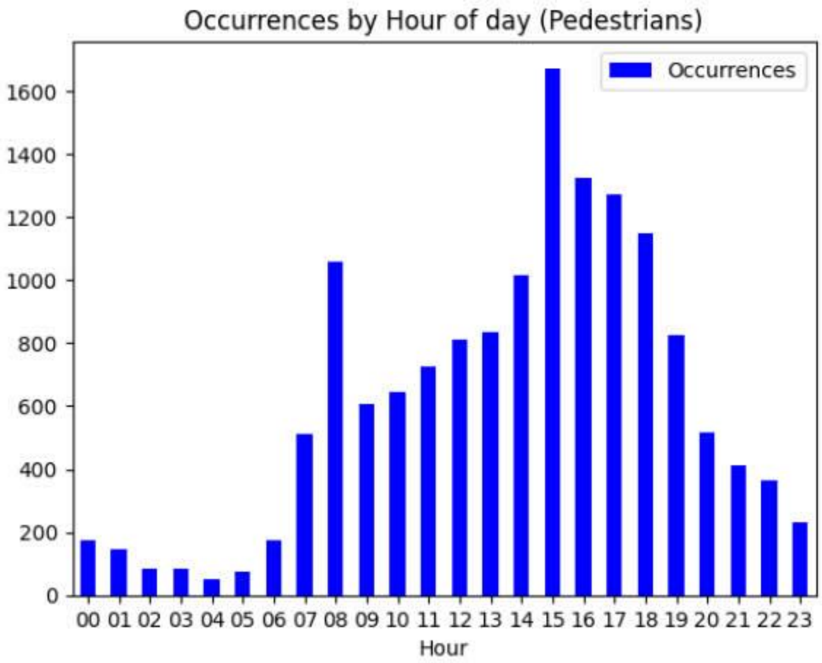


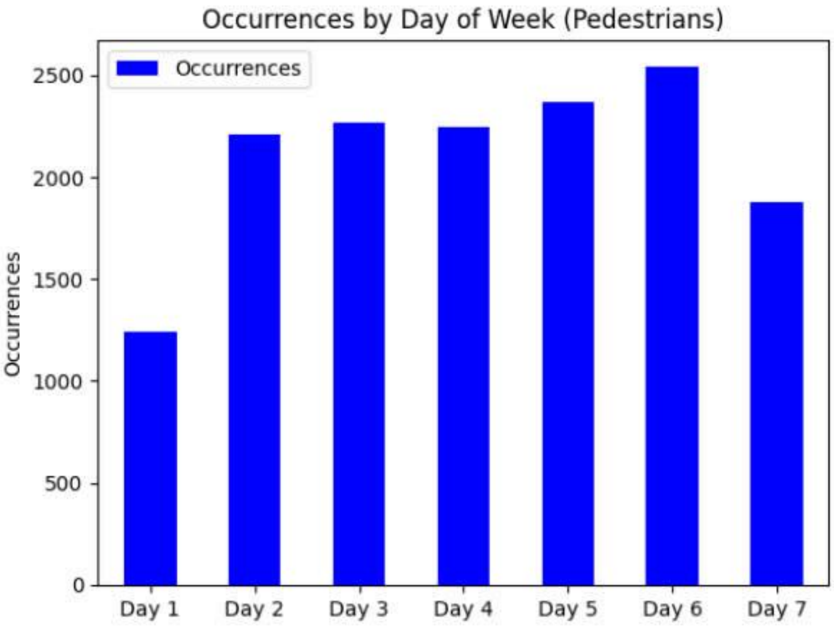


Pedestrians

The count of accidents involving pedestrians was obtained for the time quadrants and the individual hours of the day. The highest number of accidents occurred between 15:00 and 17:59. The specific hour with the most accidents was the 15th hour. Day 6 had the highest number of accidents and the coefficient of variation for accident counts across the week was 18.92%. Day 1 had a significantly lower accident count than other days, with an accident count of 1242 which was 41% less than the mean of daily accident counts.

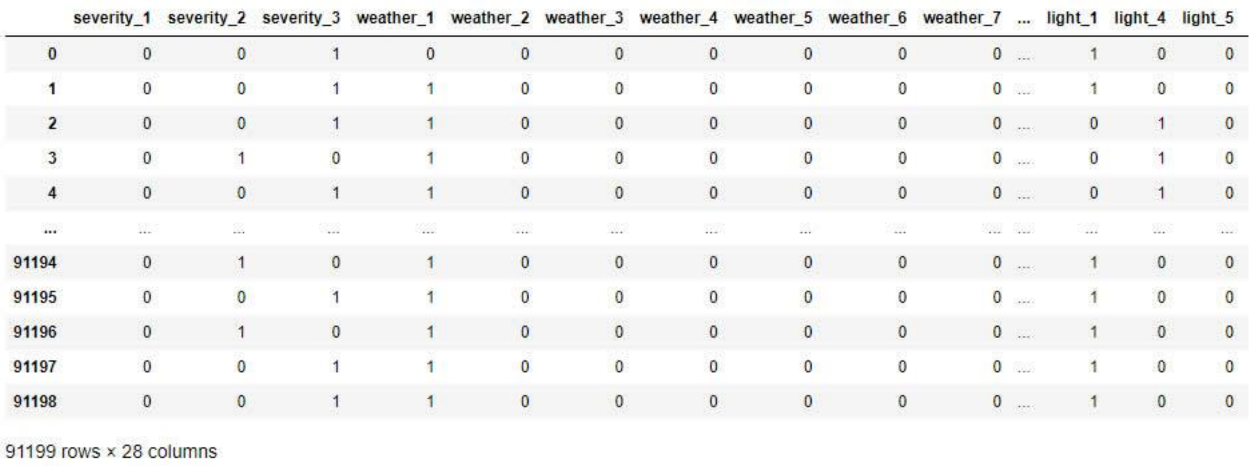






Apriori

The impacts of speed limit, light condition, weather condition, and road surface condition on accident severity were assessed utilizing the apriori algorithm. These variables were chosen based on their significant potential influence on accident severity, as established through general knowledge within the context of traffic safety. Dummies were created for these variables and concatenated into a single dataframe containing binary representations of the variables. An overview of this dataframe is given below.



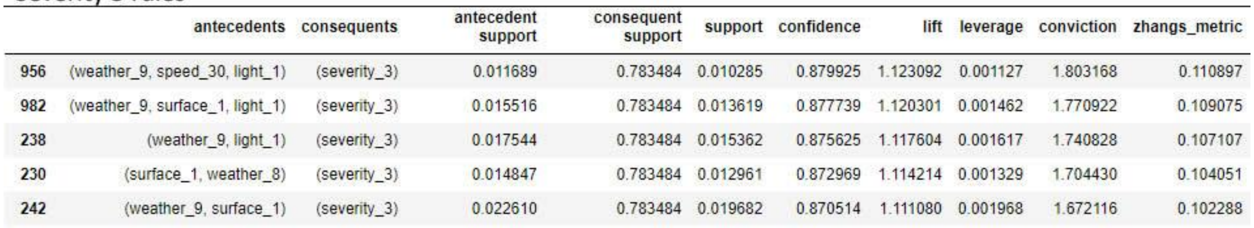
	severity_1	severity_2	severity_3	weather_1	weather_2	weather_3	weather_4	weather_5	weather_6	weather_7	...	light_1	light_4	light_5
0	0	0	1	0	0	0	0	0	0	0	...	1	0	0
1	0	0	1	1	0	0	0	0	0	0	...	1	0	0
2	0	0	1	1	0	0	0	0	0	0	...	0	1	0
3	0	1	0	1	0	0	0	0	0	0	...	0	1	0
4	0	0	1	1	0	0	0	0	0	0	...	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
91194	0	1	0	1	0	0	0	0	0	0	...	1	0	0
91195	0	0	1	1	0	0	0	0	0	0	...	1	0	0
91196	0	1	0	1	0	0	0	0	0	0	...	1	0	0
91197	0	0	1	1	0	0	0	0	0	0	...	1	0	0
91198	0	0	1	1	0	0	0	0	0	0	...	1	0	0

91199 rows × 28 columns

In addition to creating this dataframe, the values of support for accident severity 1, 2 and 3 were determined to be 0.0153, 0.2013 and 0.7835 respectively. The apriori algorithm was then applied to the dataframe created to mine frequent item sets based on the binary encoded data. Considering that the least frequent accident severity had a support of 0.0153, a minimum support of 0.1 was selected in order to include all severity types in the mining process. The result obtained was assigned to a variable which was then used to generate association rules with a minimum lift of 1. Generally, a lift value greater than 1 suggests that the relationship between the antecedent and the consequent is more significant than would be expected if the two sets were independent (Frontline Solvers, 2023). A minimum lift of 1 was therefore used to generate the association rules in order to eliminate insignificant relationships.

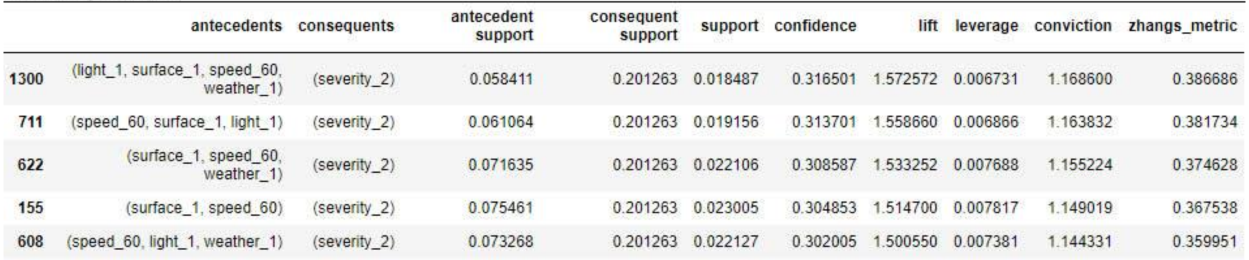
The association rules were filtered to produce rules in which the consequents comprised solely of severity levels 1, 2, and 3, respectively. This enabled a closer analysis of the associations specific to these severity levels. An overview of the top rules generated for each severity level is given below.

Severity 3 rules

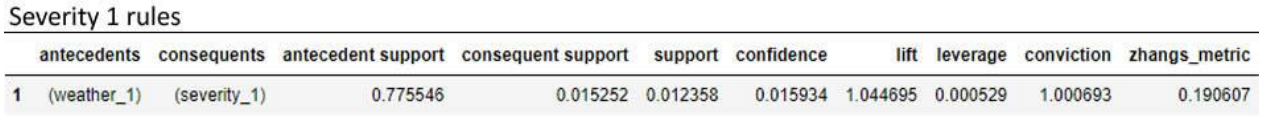


	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
956	(weather_9, speed_30, light_1)	(severity_3)	0.011689	0.783484	0.010285	0.879925	1.123092	0.001127	1.803168	0.110897
982	(weather_9, surface_1, light_1)	(severity_3)	0.015516	0.783484	0.013619	0.877739	1.120301	0.001462	1.770922	0.109075
238	(weather_9, light_1)	(severity_3)	0.017544	0.783484	0.015362	0.875625	1.117604	0.001617	1.740828	0.107107
230	(surface_1, weather_8)	(severity_3)	0.014847	0.783484	0.012961	0.872969	1.114214	0.001329	1.704430	0.104051
242	(weather_9, surface_1)	(severity_3)	0.022610	0.783484	0.019682	0.870514	1.111080	0.001968	1.672116	0.102288

Severity 2 rules



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
1300	(light_1, surface_1, speed_60, weather_1)	(severity_2)	0.058411	0.201263	0.018487	0.316501	1.572572	0.006731	1.168600	0.386686
711	(speed_60, surface_1, light_1)	(severity_2)	0.061064	0.201263	0.019156	0.313701	1.558660	0.006866	1.163832	0.381734
622	(surface_1, speed_60, weather_1)	(severity_2)	0.071635	0.201263	0.022106	0.308587	1.533252	0.007688	1.155224	0.374628
155	(surface_1, speed_60)	(severity_2)	0.075461	0.201263	0.023005	0.304853	1.514700	0.007817	1.149019	0.367538
608	(speed_60, light_1, weather_1)	(severity_2)	0.073268	0.201263	0.022127	0.302005	1.500550	0.007381	1.144331	0.359951



Severity 1 rules

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
1	(weather_1)	(severity_1)	0.775546	0.015252	0.012358	0.015934	1.044695	0.000529	1.000693	0.190607

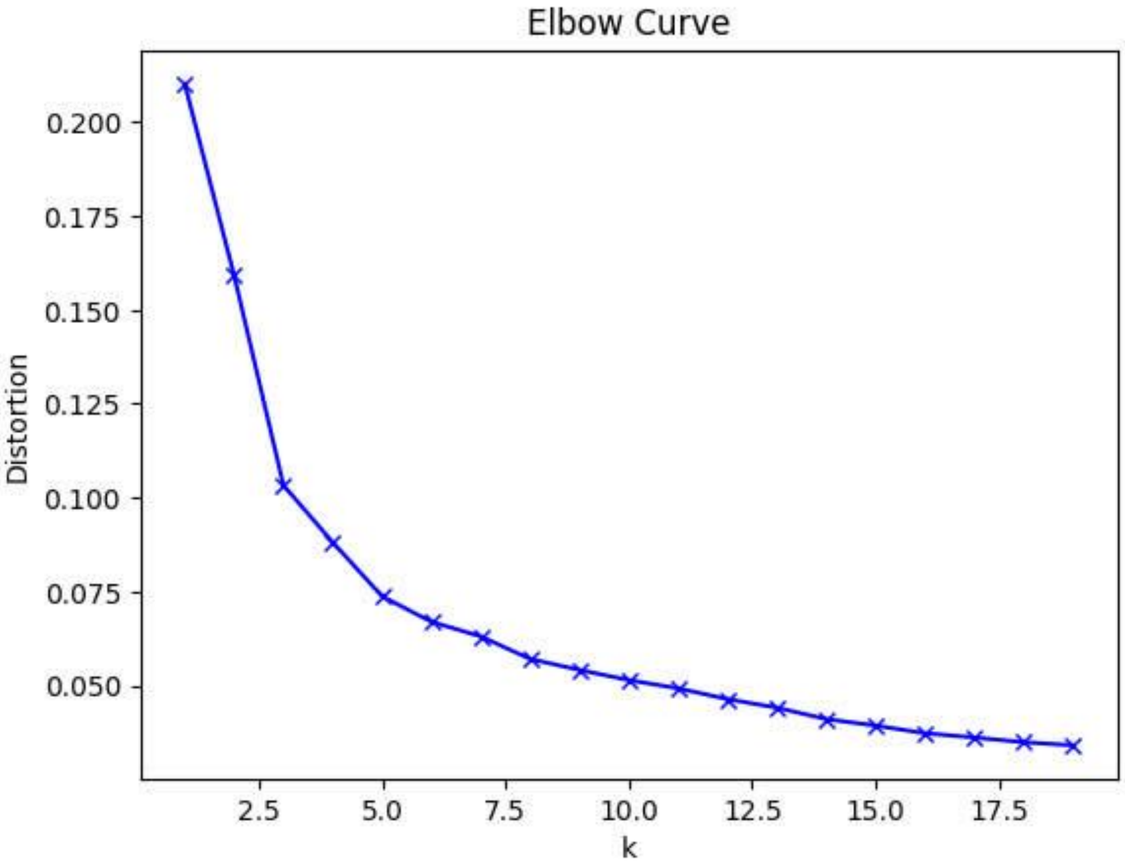
Severity 3, being the most frequent accident severity, was positively associated with 59 different combinations of the selected variables. The strongest relationship was between severity 3 and the combination of weather condition 9, speed limit 30 and light condition 1, with a lift of 1.1231.

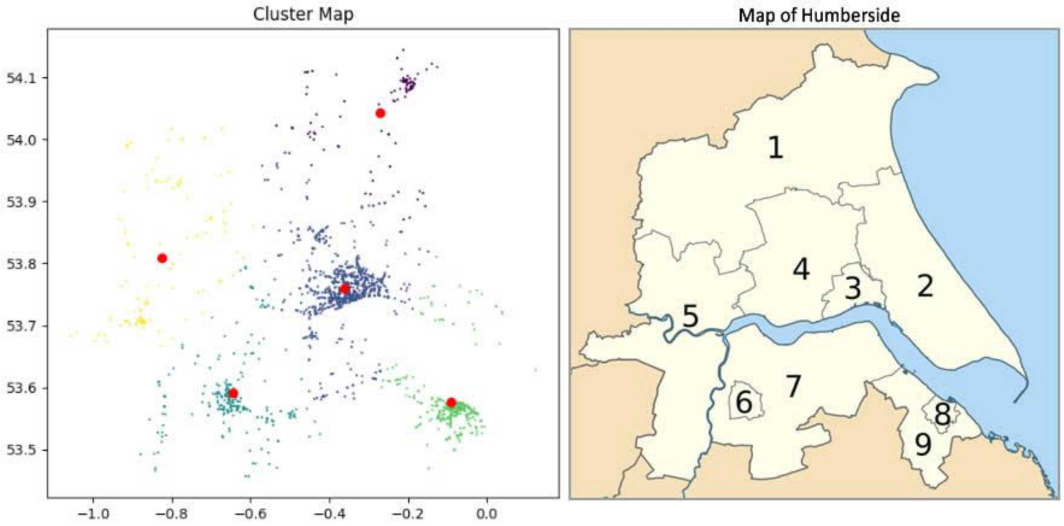
Severity 2 was positively associated with 35 different combinations of the selected variables. The most significant was the relationship between severity 2 and light condition 1, surface condition 1, speed limit 60 and weather condition 1, with a lift of 1.5726.

Severity 1 exhibited a positive association with weather condition 1. The lift of this association was 1.0447.

Clustering

Accidents in the Humberside region were located using police force 16 and passed into a dataframe. k-means clustering was done using the geographic coordinates in the dataframe. The number of clusters specified for this exercise was determined using the elbow method. The plots of the elbow curve and the clusters formed are given below.





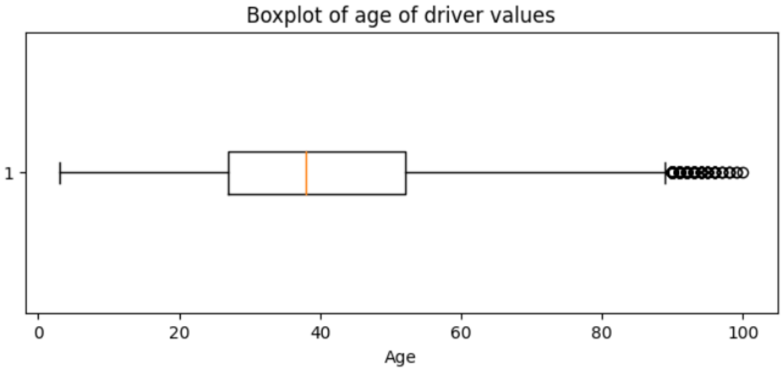
1. North Wolds, renamed East Yorkshire in 1981
2. Holderness
3. Kingston upon Hull
4. Beverley, renamed East Yorkshire Borough of Beverley in 1981
5. Boothferry
6. Scunthorpe
7. Glanford
8. Grimsby, renamed Great Grimsby in 1979
9. Cleethorpes

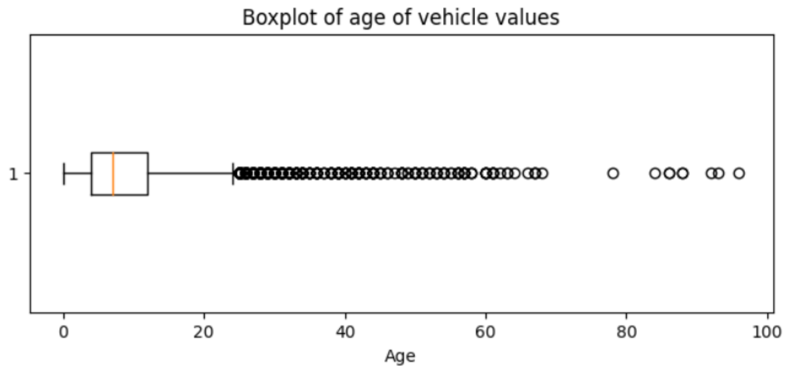
image source: wikipedia

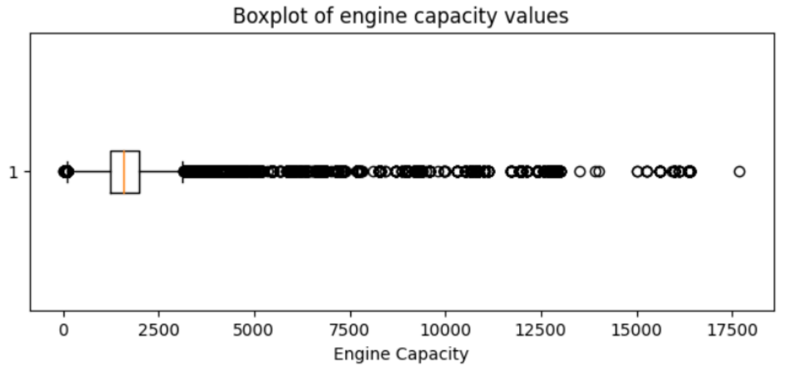
Some major accident hotspots (dense accumulation of points) were observed on the cluster map. These areas correspond to region 3 (Kingston Upon Hull), region 6 (Scunthorpe) and region 8 (Grimsby) on the map of Humberside. Smaller hotspots were also observed in the northeastern part of East Yorkshire, central Beverley and the eastern parts of Boothferry.

Outlier Detection

Age of driver, age of vehicle and engine capacity columns of the vehicle were investigated for outliers. The boxplots of values in these columns, showing potential outliers are given below.







Grubbs test did not identify any outliers in the age of driver column, however, drivers below 16 years of age were considered to be outliers based on an understanding of UK driving laws (UK GOV, 2023). To conduct a thorough analysis and provide meaningful recommendations, it is necessary to work with a dataframe consisting only of legally eligible drivers therefore drivers aged below 16 were removed from the vehicle table. The accident indexes for accidents involving these drivers were noted and used to drop them from the accident and casualty tables. The drivers above the upper age limit of a Multiple of interquartile range test (89.5 to 100) were accepted as rare but realistic entries.

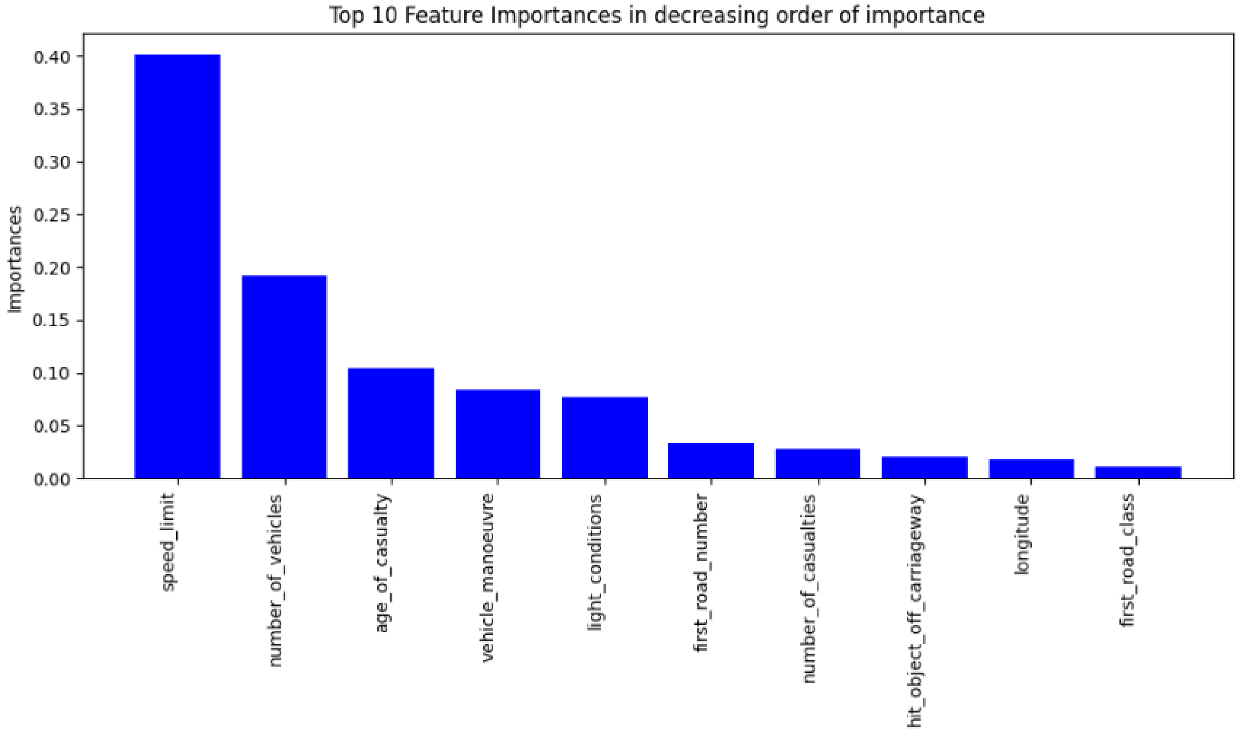
In a similar fashion, Grubbs test was performed on the age of vehicle values which deduced an upper age limit of 35years. The 143 vehicles aged between 36 and 96 years were however left in the table and not treated as outliers because in the legal sense, these vehicles could still be in use provided that they meet current road worthiness standards.

Grubbs test put the minimum engine capacity at 1998cc. This figure was easily disregarded as vehicle types like scooters and motorbikes have engine capacities that normally fall below that figure. The outliers in the engine capacity column were then investigated manually by sorting the table in ascending and descending orders of engine capacity and comparing the values with the other values in the table for the same vehicle types. One outlier was identified: a 15-year-old Seat Alhambra with an engine capacity of 7cc which was too low for a car. This engine capacity was imputed with 1984, the engine capacity of a 2005 Seat Alhambra obtained from an online resource (Cars Data, 2021).

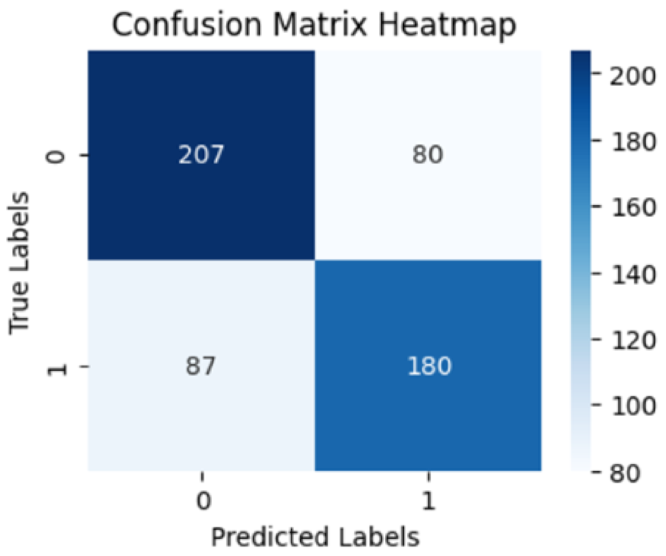
Classification

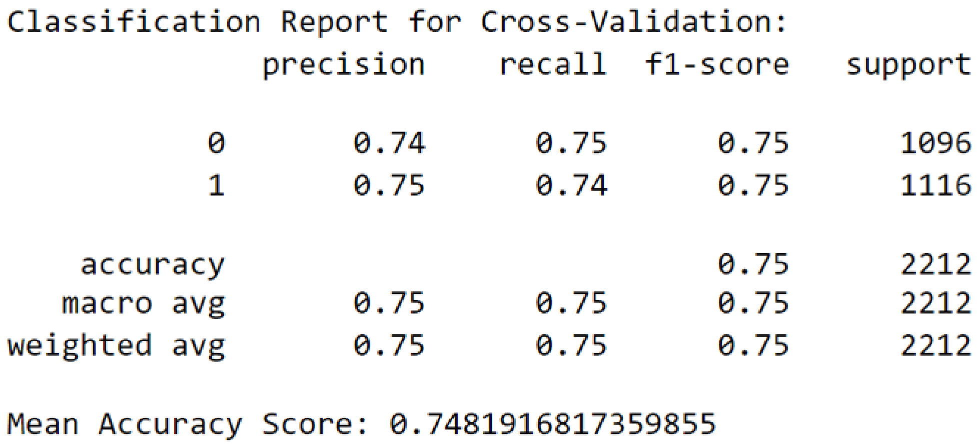
To develop a classification model to predict fatal injuries, the accident, casualty and vehicle tables were concatenated into a single dataframe using accident indexes, selecting only one row for accident indexes associated with multiple vehicles in the vehicle table and/or casualties in the casualty table. The accident severity column was modified to contain only two classes, representing fatal and non-fatal accidents. This was done by combining slight and serious accidents into a singular class to represent the non-fatal accidents.

Considering that fatal accidents in the dataset were significantly lower than non-fatal accidents, random undersampling was used to balance the data. Setting accident severity as the target variable, and other relevant columns as the input variables, the balanced data was split into training and test data using train-test-split. A decision tree classifier was trained using the training data and the performance of this model in classifying the test data was evaluated. Cross validation was used in this classification and the mean of the accuracies obtained from each fold was 71.52%. The feature importances of this classification were extracted and ranked. The top 10 features are shown below.



The classification was repeated using the top 10 features and a stacking classifier comprising of decision tree, gradient boosting, k-neighbors, naïve bayes and support vector. The final estimator for this stack was a logistic regression. Accuracy, precision, recall and f-1 scores were evaluated. The mean of the accuracy scores for each fold was 74.82% making the model consistent with industry standards and suitable for real world use (Kirsten Barkved, 2022). In addition to this, the precision, recall and f-1 scores were 0.75, 0.75 and 0.75 respectively. The report and confusion matrix for this classification can be found below.





```
Classification Report for Cross-Validation:
              precision    recall  f1-score   support

           0       0.74      0.75      0.75      1096
           1       0.75      0.74      0.75      1116

    accuracy                           0.75      2212
   macro avg       0.75      0.75      0.75      2212
weighted avg       0.75      0.75      0.75      2212

Mean Accuracy Score: 0.7481916817359855
```

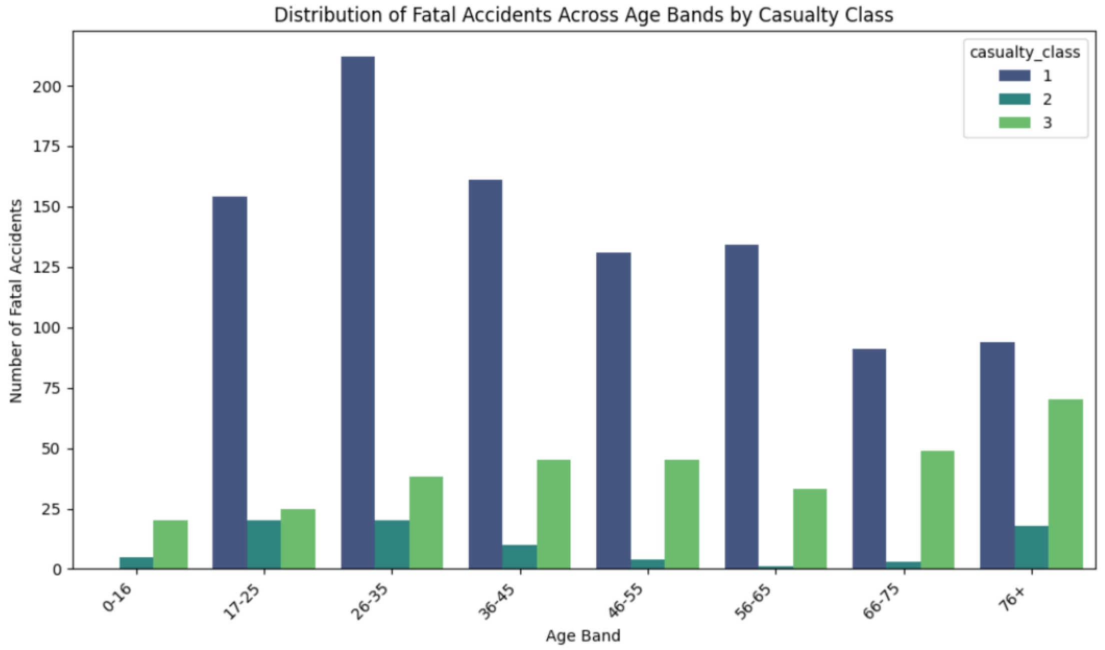
Recommendations

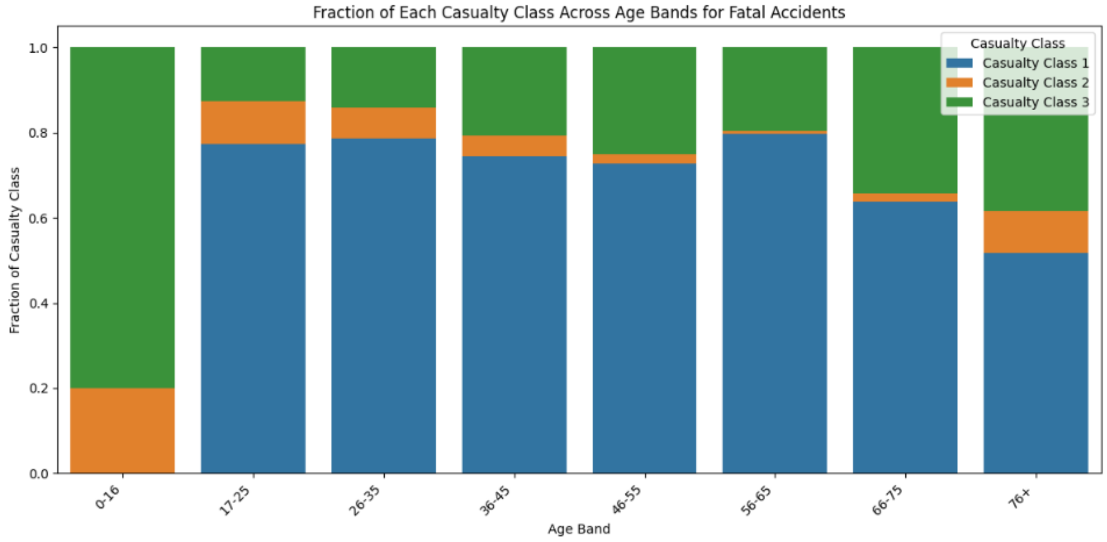
Speed control measures: The classification model demonstrated that speed limit had the most significant contribution to fatal accidents. Fatal accidents also occurred mostly in daylight, a condition that encourages speeding. The apriori rules showed a positive relationship between fatal accidents and clear weather conditions, which also encourage speeding. In addition, the apriori rules showed a strong correlation between higher speed limits and serious accidents. The Government should therefore implement speed control measures such as lower speed limits, road bumps and lane width restrictions.

Accident hotspot investigation: The clustering exercise revealed some accident hotspots. These areas should be investigated closely to determine and address specific causes of accidents. The classification model indicated a strong relationship between first road number and fatalities. Roads with prominent levels of fatalities should also be investigated closely.

Promotion of public transportation: Accident occurrence was observed to peak around and during business hours when traffic activity is high. Promoting the use of public transportation, along with improving its comfort and appeal, can help mitigate these accidents and enhance road safety.

Raising the standard of driver training: Analysis of the fatal accidents showed that the largest fractions of driver casualties fell into age bands below 65 years, suggesting a decline in driving skills, safety orientation and knowledge of traffic rules over the decades. The government should therefore intensify the current driver teaching and licensing process to improve overall driver competencies and reduce the chances of accident occurrence.





Bibliography

Cars Data (2023) *Seat Alhambra Engine capacity*
Available Online: https://www.cars-data.com/en/seat-alhambra/engine-capacity
[Accessed on 02 August 2023]

Frontline Solvers (2023) *Association Rules*
Available Online: https://www.solver.com/xlminer/help/association-rules
[Accessed on 02 August 2023]

INSEE (2016) *Coefficient of variation/CV*
Available Online: https://www.insee.fr/en/metadonnees/definition/c1366
[Accessed on 02 August 2023]

Kirsten Barkved (2022) *How To Know if Your Machine Learning Model Has Good Performance*
Available Online: https://www.obviously.ai/post/machine-learning-model-performance
[Accessed on 02 August 2023]

UK GOV (2023) *Driving lessons and learning to drive*
Available Online: https://www.gov.uk/driving-lessons-learning-to-drive
[Accessed on 02 August 2023]