Introduction

Accidents are the result of energy dissipation; they are one of the most terrifying events that may occur; they are a global issue, and road accidents are one of the leading causes of unnatural deaths, impairments, and property damage. One strategy to reduce road traffic accidents is to look into the details of traffic incidents that have been documented and to figure out what caused them. To increase road safety, determine where most traffic accidents occur and identify peak time slots. To qualify the likelihood of traffic accidents, a model has been completed and tweaked. But first, there's the data analysis. This report has been divided into categories depending on the most important characteristics versus the severity of the accident. This is done in the hopes that some elements will have a greater impact on the severity of a road traffic collision than others. The Lasso approach was used to choose the features that would be used to predict our model. In general, the statistics indicated that the majority of accidents happened on Fridays between 4 and 5 p.m., with November having the largest number of incidents in 2019. My data is comprised of 117536 different records, with repeated casualties and vehicles.

Data Cleaning

For the time column, it was backfilled based on the sorted data column to put seasons and periods into consideration. The date and time column were further split to give us weeks as well as months for better data analysis.
By grouping depending on the location of the police force and filling them with the mean and median values, the Latitude, Longitude, Location easting, and Location northing was updated.

Data Analysis

Months

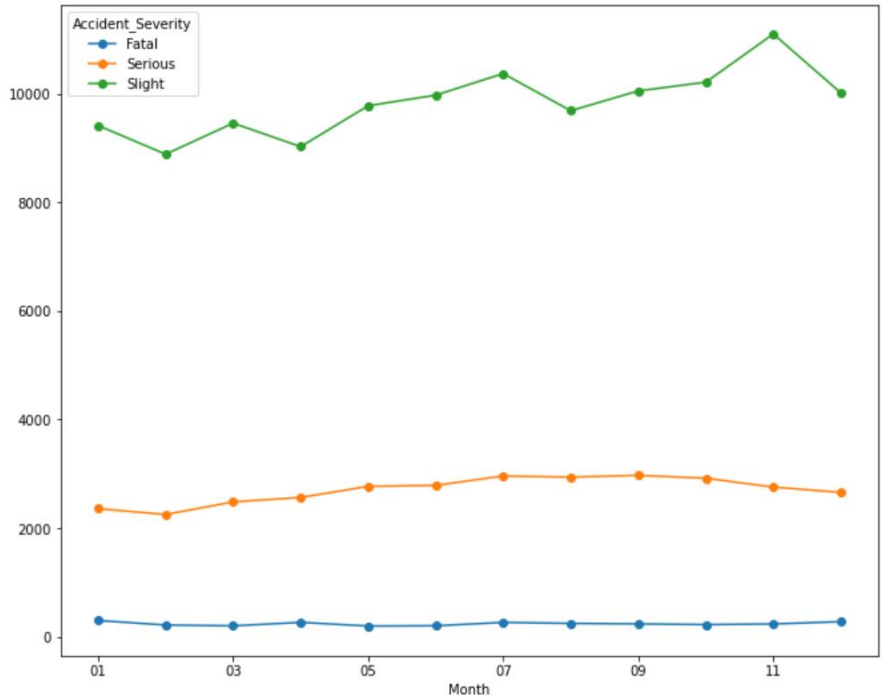

Figure 1 Accident Severity based on the months

This graph shows that for the severity of each month. Fatal accidents titled as accident Severity 1(Fatal) are almost static throughout the year as it is appearing to almost be in a straight line in this line plot, although January has the highest number of causalities by fatalities. The Severity 2(Serious), peaks in the month of September as well as the severity 3(Slight) which peaks in the month of November.Dissecting further based on the severity, we realise that in the months November, September and January which recorded the highest occurring accidents based on the severity: slight, serious and fatal respectively.

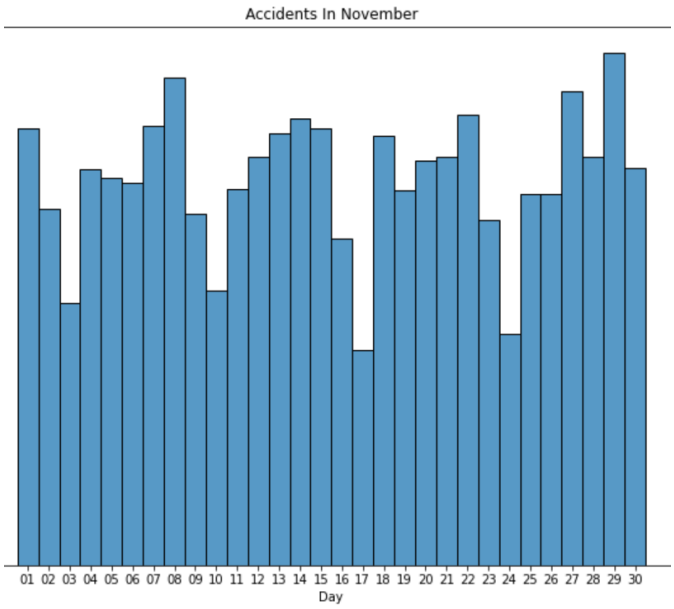

Figure 2 Accidents with severity slight based on the months of November

In November, Accidents with severity slight occured mostly at 5pm, with a large percentage occuring on Friday. Peak accident days, can be attributed to events such as blackfriday was celebrated on the 29th of November, so also thanksgiving shopping on the 27th of November.

In September, on the 20th, 15th and 21st the most accidents occurred.

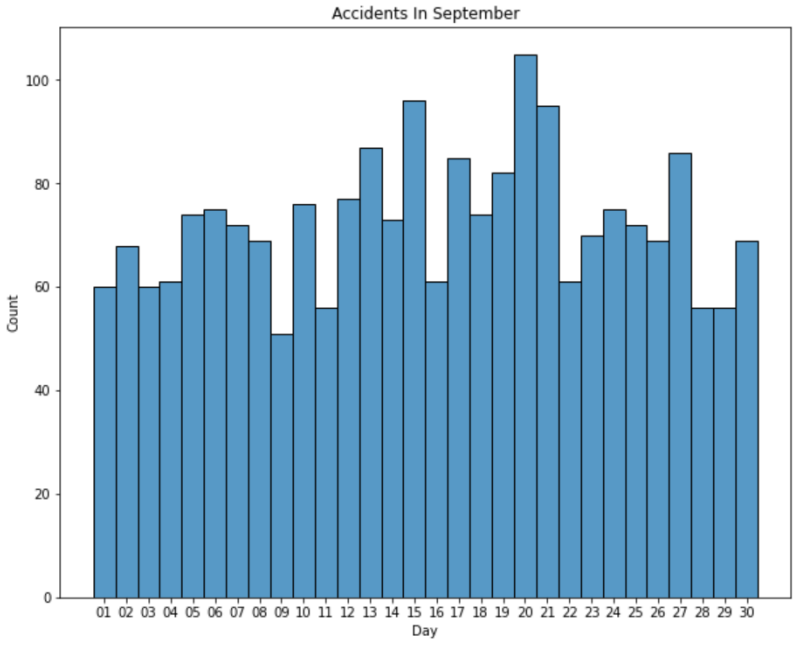

Figure 3 Accidents with severity serious based on the months of September

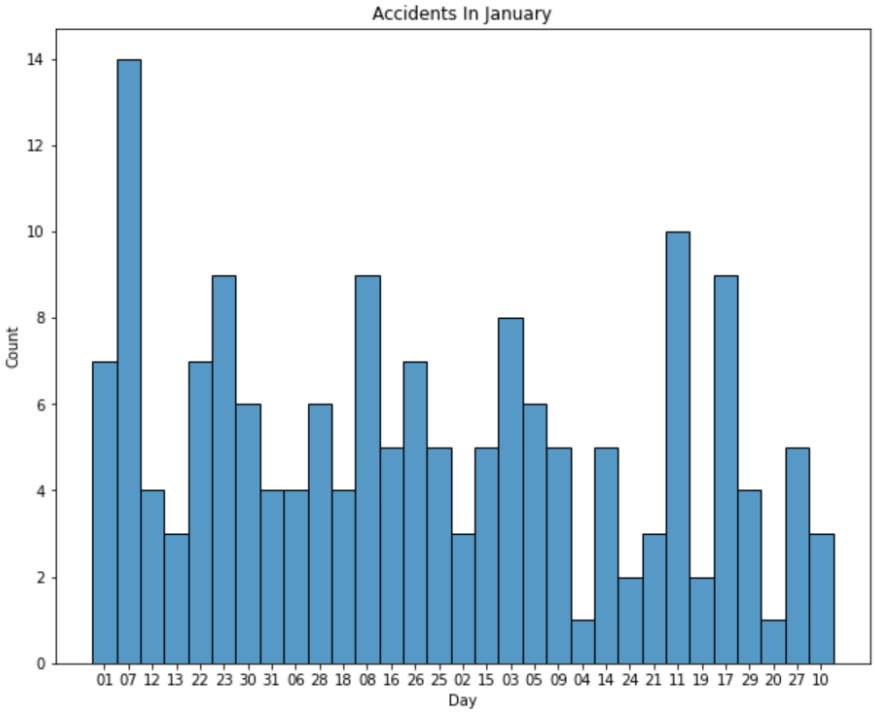

Figure 4 Accidents with Fatal slight based on the months of January

While In January, on the 7th day as described in the chart, a bank holiday in the United Kingdom giving people the opportunity to commute, this slightly explains the amount of accidents.

Location

Location was also a huge determining factor for accidents; the metropolitan area of London recorded the highest count of accidents in 2019.

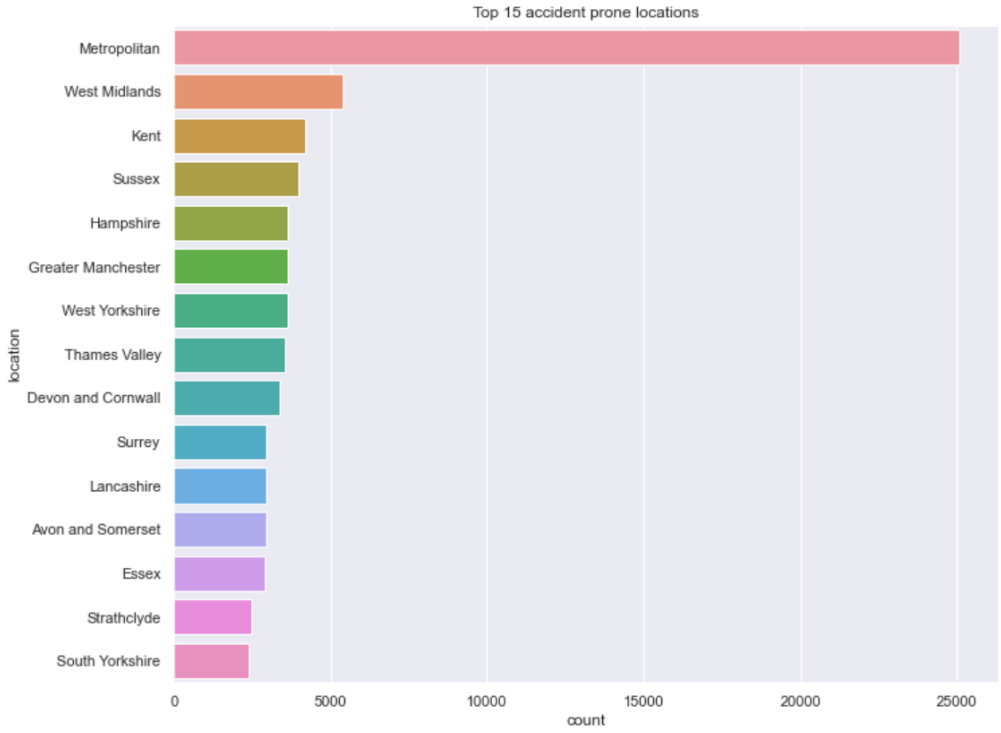

Figure 5 top 15 accident prone locations

The top 15 accident-prone locales are depicted in this graph. In 2019, incidents in the London Metropolitan region, excluding the city of London, accounted for 21% of all accidents. This is a large number, but further study shows that it is one of the most densely populated locations in the European Union as of September 2019, which explains the high rate of accidents.

This is a KNN Cluster of the Latititude and Longitude of my dataset using the elbow method, it pinpointed 4 major areas as best fit for the data where accidents clustered.

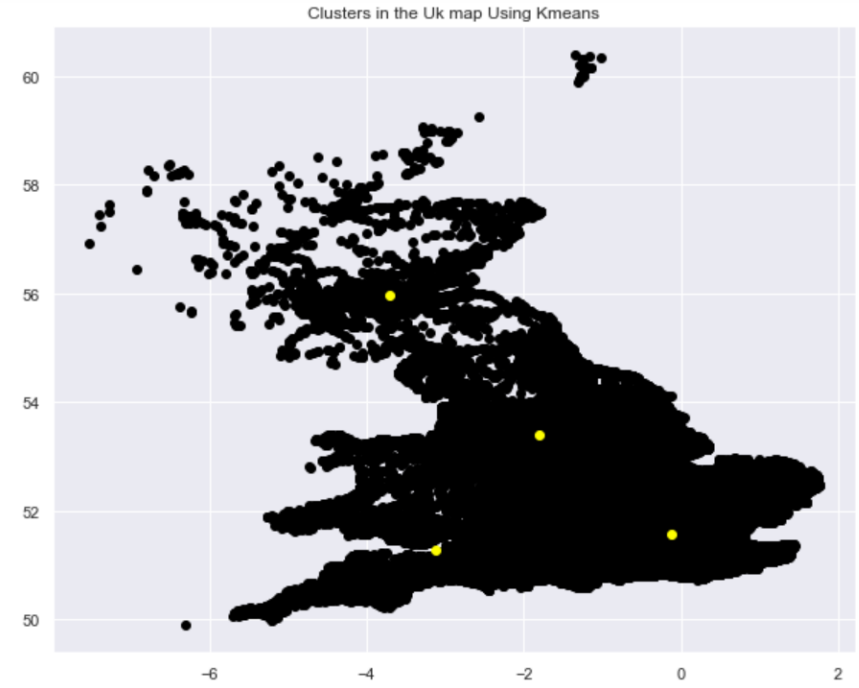

Figure 6 A KNN cluster showing the locales with the highest clusters

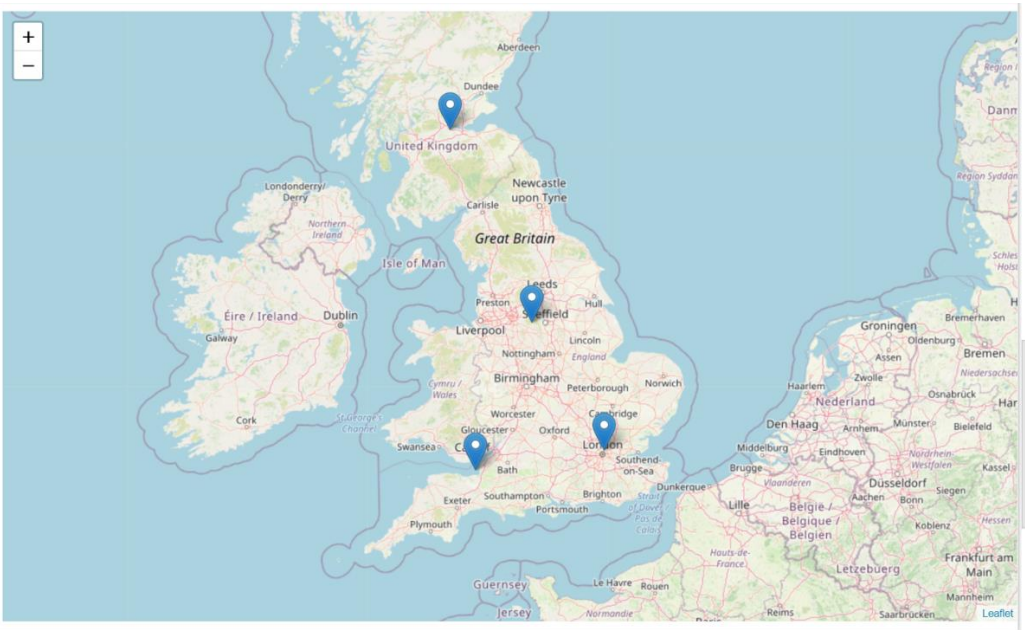

Figure 7 The four Centroid areas obtained from KNN as represented below using folium.

Pedestrains

Despite their perceived insignificance in every community, pedestrians are an important segment when assessing the effects of accidents. The majority of pedestrian accidents were inferred to occur on Fridays at peak hours of 8 to 9 am and 3 to 4 pm, respectively. Males appeared to be the most impacted, with a majority of those affected being between the ages of 11 and 15 (students), 26 to 35 (working class).

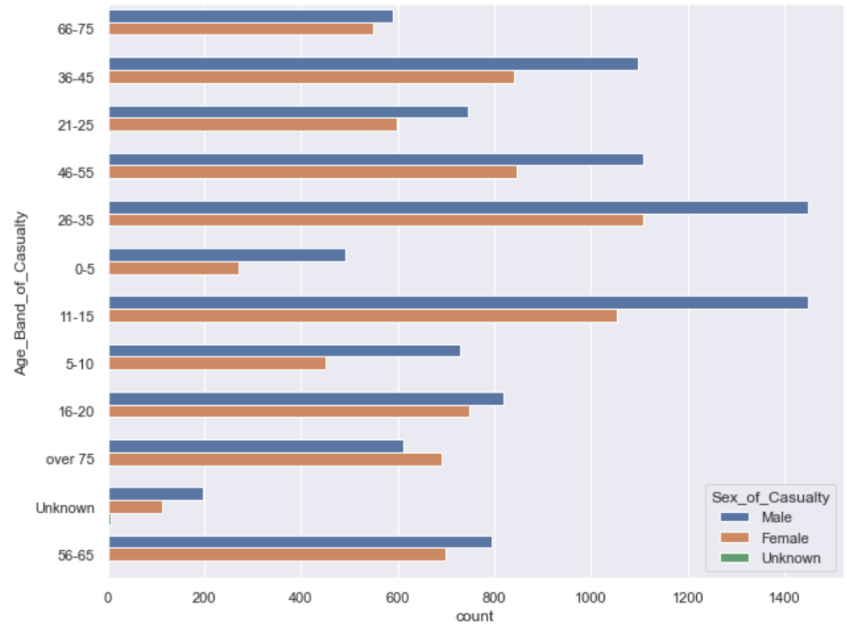

Figure 8 Pedestrian accidents against age and gender

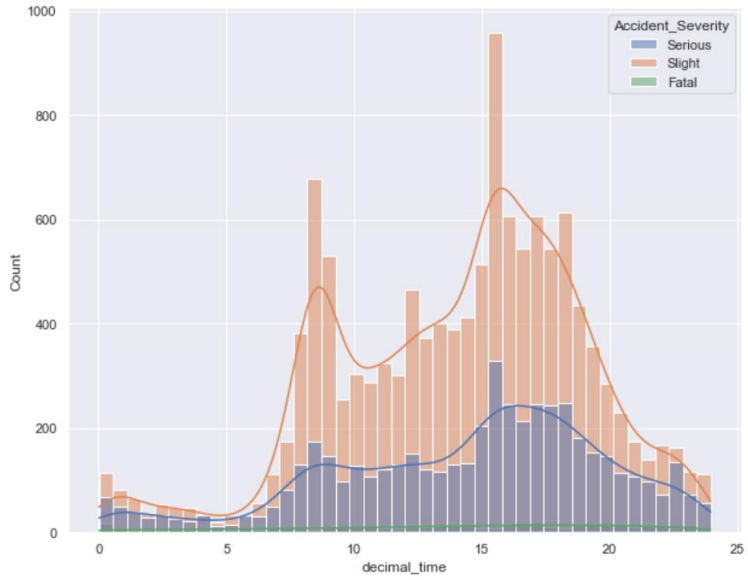

Figure 9 pedestrian accidents based on the time

Motorcycles

Motorcycles had a significant impact on accidents in 2019, accounting for 14317 accidents. Motorcycles with engines smaller than 125cc were the most likely culprits, owing to their tiny size, although motorcycles with engines greater than 500cc had the greatest fatalities; speed is more likely to blame in this situation. The 125cc has a top speed of around 70mph, while the 500cc has a top speed of over 100mph.

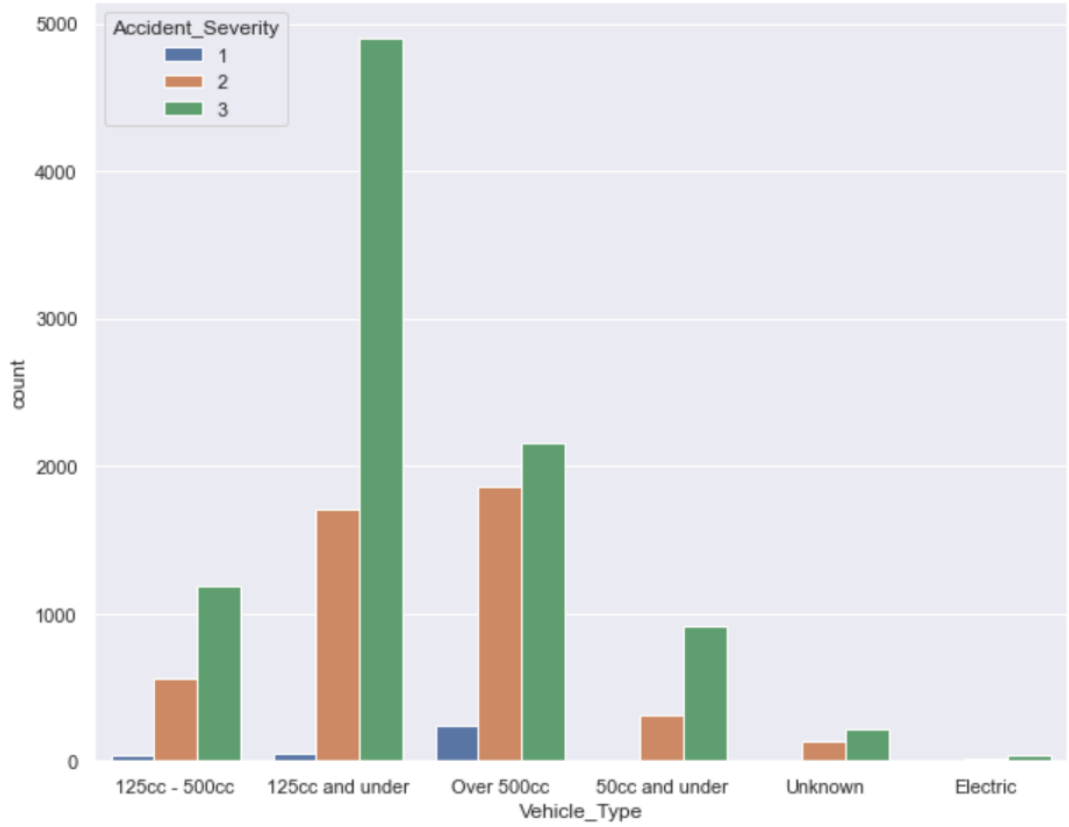

Figure 10 Accidents based on Motorcycle types

Daylight Savings

For a comprehensive assessment, it is only appropriate to inquire if there is any difference in the number of casualties on the road during daylight savings time against normal time. This graph depicts the accident chart for the weeks of 2019. According to the accident chart, week 47 is in November during the Thanksgiving holiday, and week 37 is in September as previously in fig ().They explain the high peaks depicted in the graph.

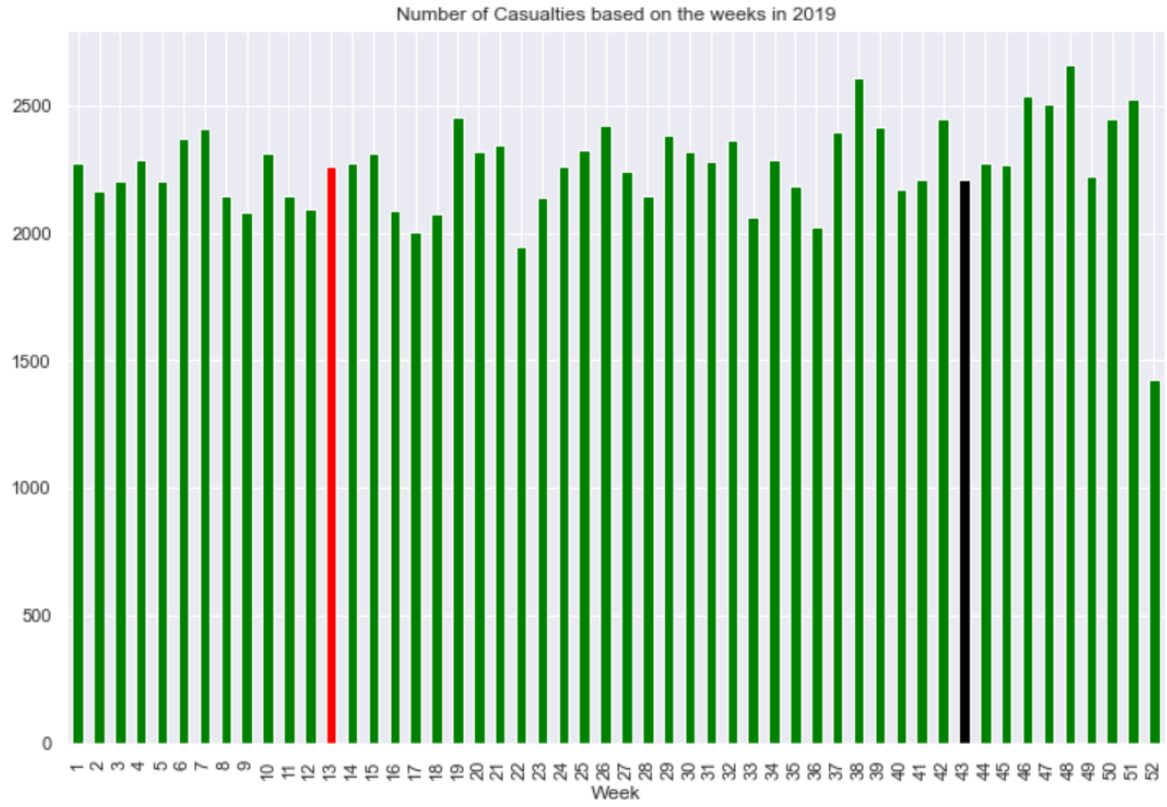

Figure 11 Number of Casualties in the 52 weeks

To determine the impact of daylight savings time on the number of fatalities in 2019, a Ttest was conducted. There were 4831 casualties a week before and after daylight savings began (Week 13) (M = 1.28, SD = 0.76) compared to 4714 casualties a week before and after daylight savings ended (Week 43) (M = 1.3, SD = 0.75), with a T-value of (0.0186), pvalue of (0.0186). (0.98). These findings reveal that there is no substantial difference in the number of casualties between the week daylight began and the week it ended, despite the fact that

the number of casualties in the week daylight ended was somewhat greater.

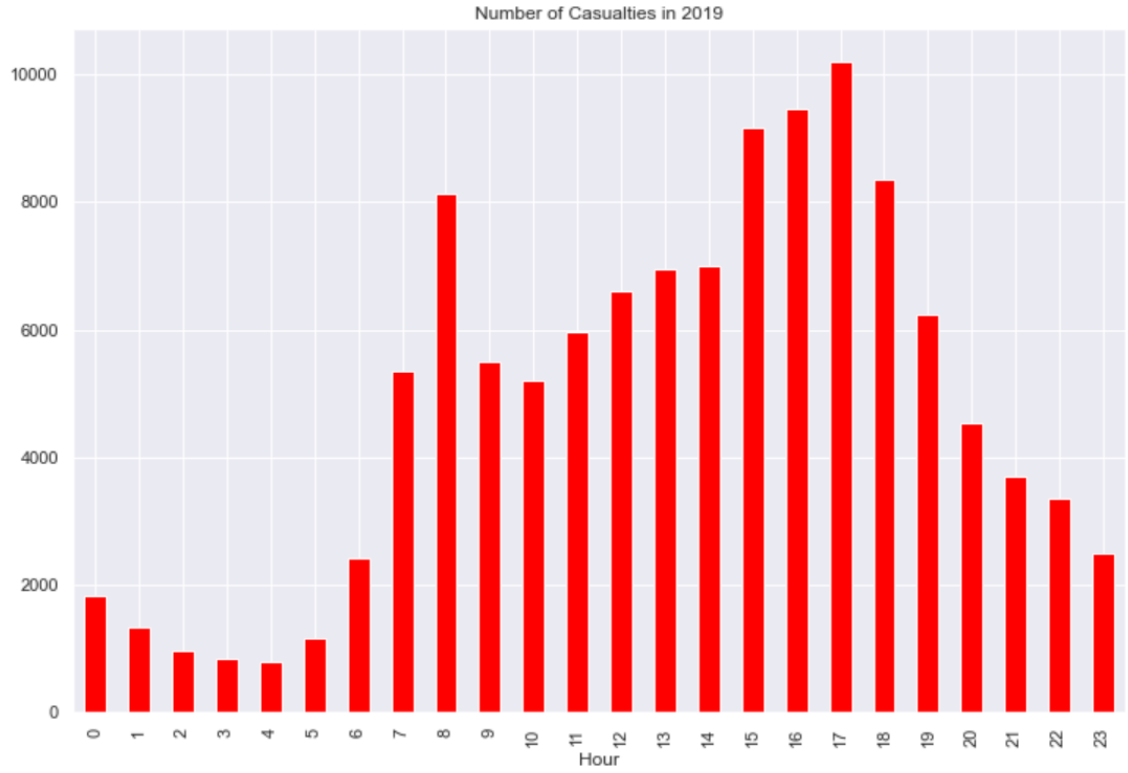

Figure 12 Time Frame of accidents in 2019

Then I tried to figure out how sunrise and sunset affected the number of casualties. There is a difference in the time the sun rises at Standard time (7 am–8 am) and during daylight savings (6 am–7 am), as well as the time the sun sets at Standard time (4 pm–5 pm) and during daylight savings (6 am–7 am) (7 pm-8 pm). The T-value of both proportions = (-0.4463), pvalue= (0.66). we reject the alternative hypothesis that says the mean values differ, although the number of casualty ratio differences is roughly (2:1).

Gender

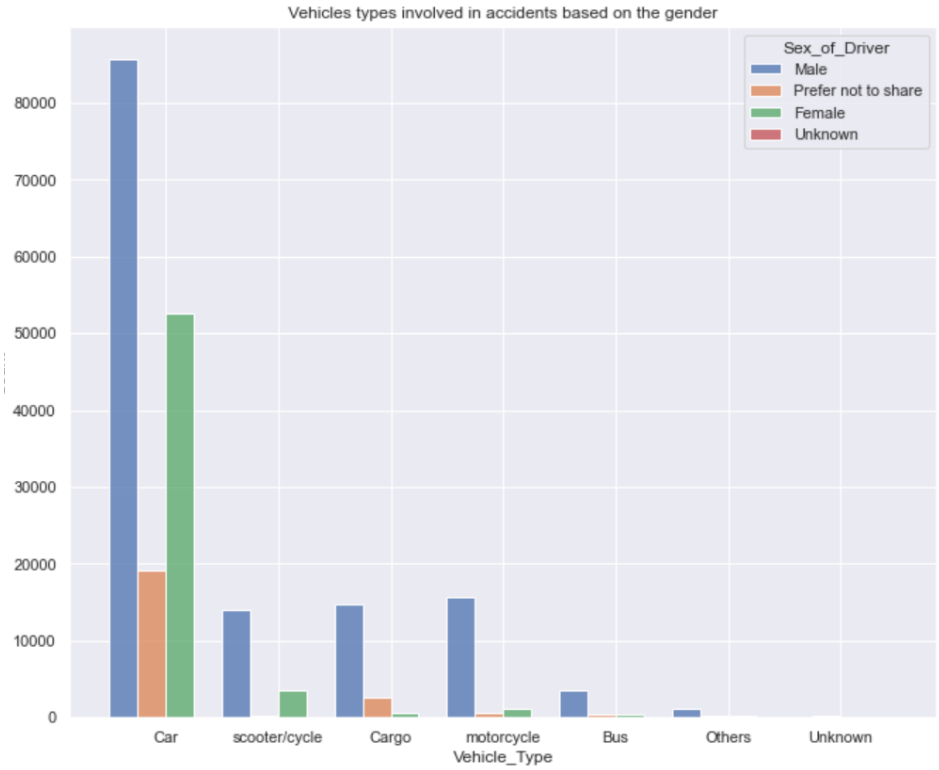

Figure 13 Vehicle types involved in accidents based on the gender

From the graph, it shows that males are the major road users, few females and/or are road users. I was then interested if there were relationships between the engine capacity, Age of vehicle, propulsion type, and if the vehicle was a left-hand driver or not on the number of accidents. At 70% threshold, I realized that vehicles with engine capacity of 0-2000 run on petrol, it then follows that the driver is left-handed. This was derived at 50% support, 96% Confidence plus 4.35 conviction.

Weather Conditions, Road Conditions and Light Conditions.

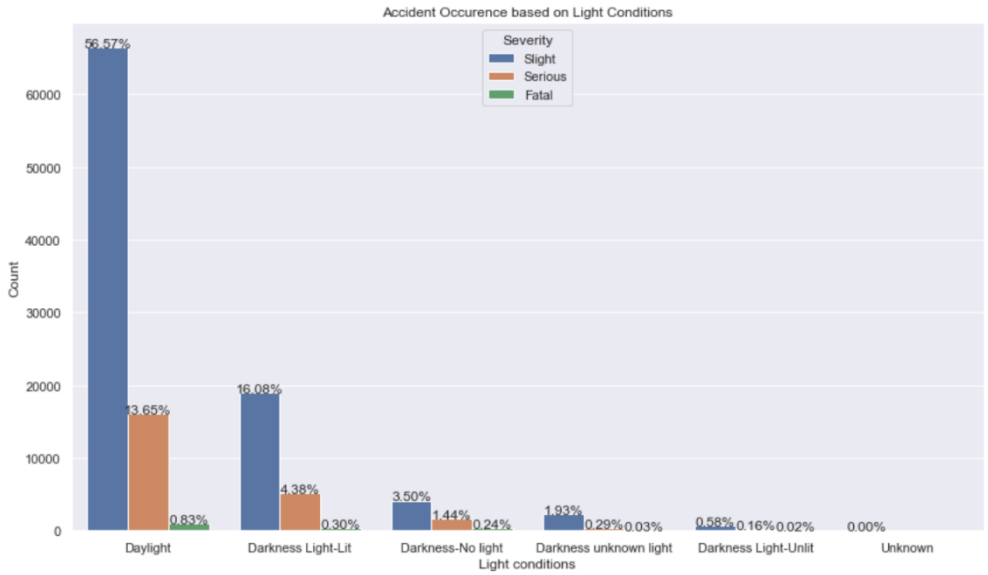

Figure 14 Accident Count based on light conditions

71 % of accidents occurred when there was daylight, 23% of accidents happened during the dark with the presence of a form of light, and 5.9% accidents took place in the absence of any light whatsoever.

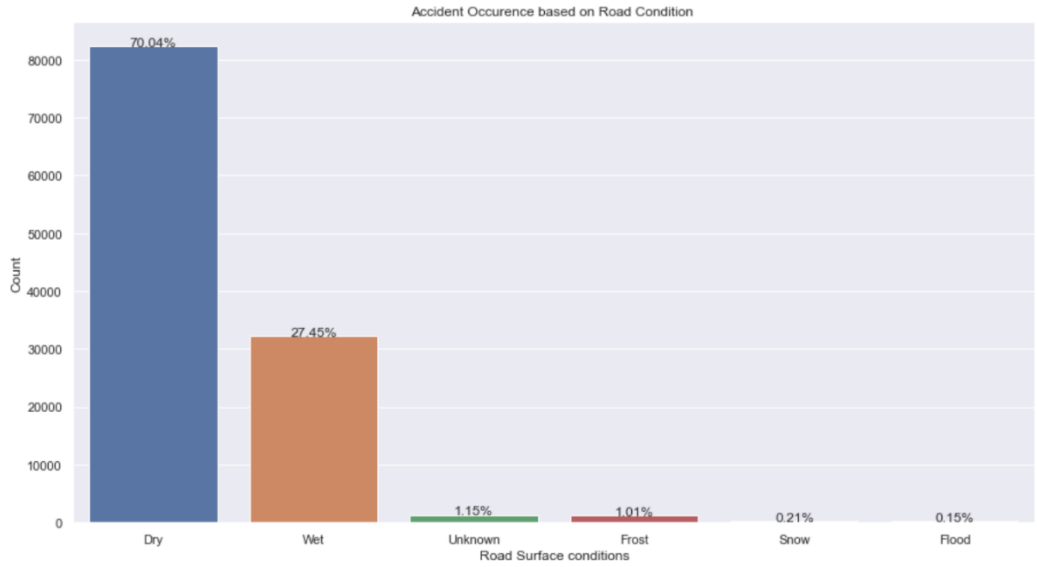

Figure 15 Accidents based on Road Occurrence

70% of accidents happened when it wasn't raining or snowing, most accidents will be said to have occurred with the road in good condition.

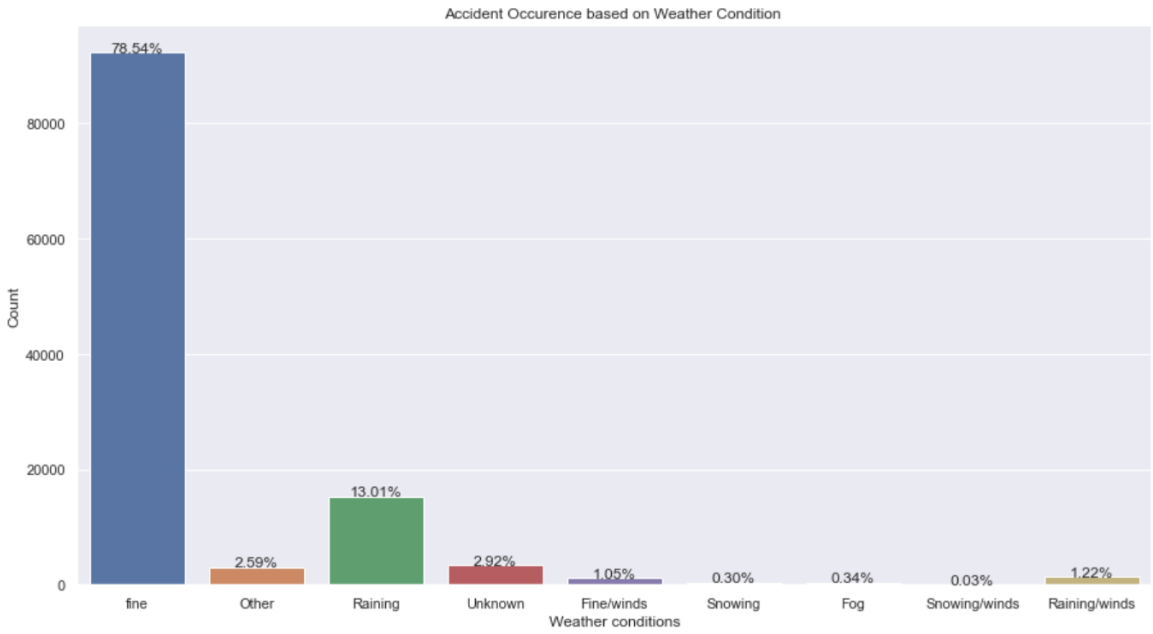

Figure 16 Accidents in percentages based on the Weather Condition count

This shows that 78%accidents occurred when the weather was fine, this means the weather didn't influence accidents greatly.

Relationship between Features Using Apriori

An Apriori to check the relationship between light conditions, weather conditions, road conditions, and the accident Severity revealed that at a 50% threshold, slight accident's happened mostly during the day when the road is relatively dry. This is supported by 40%, 94% confidence, and 3.92 conviction.

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(Road_Dry)	(weather_fine)	0.700390	0.785427	0.661848	0.944972	1.203130	0.111743	3.899300
(Severity_Slight, Road_Dry)	(weather_fine)	0.552069	0.785427	0.520011	0.941931	1.199259	0.086401	3.695113
(Road_Dry, light_Daylight)	(weather_fine)	0.545220	0.785427	0.517339	0.948863	1.208085	0.089108	4.196049
(Severity_Slight, Road_Dry, light_Daylight)	(weather_fine)	0.432931	0.785427	0.409253	0.945308	1.203559	0.069217	3.923293

Figure 17 A table depicting the relationship between the road conditions and the severity

The reason for a journey was also concerned to know if it had any impact on the accident's severity,

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(Sex_Female)	(Journey_purpose_extra-curricular)	0.264765	0.733288	0.202687	0.765534	1.043975	0.008538	1.137532
(Journey_purpose_extra-curricular)	(Sex_Female)	0.733288	0.264765	0.202687	0.276408	1.043975	0.008538	1.016091
(Sex_Female)	(Severity_Slight)	0.264765	0.798930	0.219065	0.827392	1.035625	0.007536	1.164895
(Severity_Slight)	(Sex_Female)	0.798930	0.264765	0.219065	0.274198	1.035625	0.007536	1.012996
(Journey_purpose_extra-curricular)	(Severity_Slight)	0.733288	0.798930	0.587436	0.801099	1.002715	0.001591	1.010907
(Severity_Slight)	(Journey_purpose_extra-curricular)	0.798930	0.733288	0.587436	0.735279	1.002715	0.001591	1.007522

Figure 18 This shows the relationship between the gender, severity, purpose of the journey

At 20% threshold, women engage in more extracurricular activities with a 20% support and 70 % confidence. Also, when there's the severity is slight, there's a 79% chance that it's a female driver. 58% support says a lot of slight accidents happen during extracurricular activities.

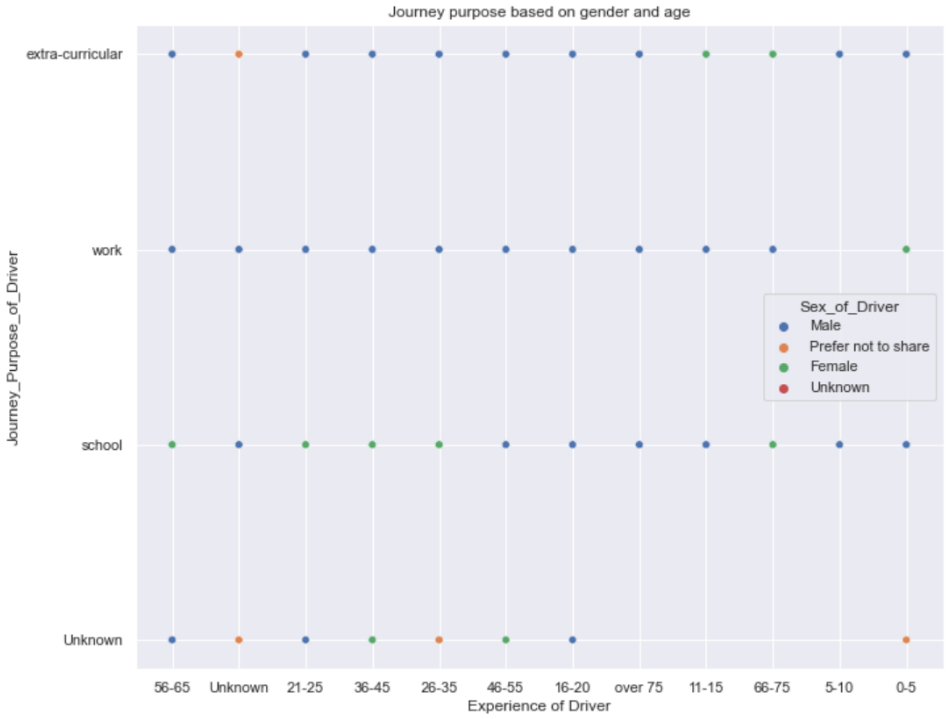

Figure 19 A scatter plot showing the relationship between the gender, experience of drivers based on their gender

Model Building

Data Preparation

The casualty, vehicles and accident data was merged on the columns that had similarly, after merging a total of 109518 entries with 80 columns. Accident severity “Fatal” was merged into “Serious” because it had very few entries, the accident severity “serious” had 6666 entries while severity “slight” had 25985 entries. The data was then balanced using resampling to have a total of 51970 entries with 80 columns. Entries such as Location, date and time were dropped to reduce dimensionality. The lasso feature selection approach was utilized, which is a logistic regression embedding method, to choose the most significant features for my model. Lasso is a combination of the terms "least absolute shrinkage" and "selection operator." It shrinks the regression coefficients, lowering some of them to zero, to regularise model parameters. It picks features that haven't been shrieked to zero after it's been applied, reducing prediction mistakes that are frequent in statistical approaches. The features were then fit into the model and it generated a total of 21 features;

'Vehicle_Reference', 'Casualty_Reference', 'Casualty_Class',

'Sex_of_Casualty', 'Age_of_Casualty', 'Car_Passenger', 'Casualty_Type',

'Vehicle_Manoeuvre', 'Skidding_and_Overturning',

'Vehicle_Leaving_Carriageway', '1st_Point_of_Impact', 'Sex_of_Driver',

'Age_of_Driver', 'Age_Band_of_Driver', 'Engine_Capacity_(CC)',

'Police_Force', 'Number_of_Casualties', 'Speed_limit', '2nd_Road_Class',

'Light_Conditions', 'Did_Police_Officer_Attend_Scene_of_Accident'.

I wanted to see if I could minimize dimensionality using PCA after implementing Lasso, Because PCA is particularly sensitive to variable variation, I then standardized the variables. The hypothesis that the variables were uncorrelated within the population was tested using Bartlett's test of sphericity,

where,

H0: There is no correlation between any of the variables.

H1: indicates that at least one variable is linked.

A p-value of (0.0) was achieved, which is less than 0.05. We then reject the null hypothesis and conclude that at least one pair of variables is correlated, implying that PCA should be used.

Although PCA is recommended, the Kaiser-Meyer-Olkin (KMO) test applied to our dataset is 0.5943 (considering 4dps), which is lower than 0.7. It is then concluded that PCA isn't expected to produce any significant decrease in dimension or extraction of relevant components on this dataset, hence it wasn't used for the model building.

Algorithms Tested

Using a 75/25 split, I implemented five machine learning algorithms on my 21 selected features to predict the possibility of a slight or serious accident to achieve the following results as represented in the tables below

Table 1: Model accuracy

Model	Accuracy	Precision	Recall	F1-Score
Naïve Bayes	0.62	0.63	0.62	0.61
Decision Tree	0.89	0.90	0.89	0.89
Random Forest	0.84	0.84	0.84	0.84
Logistic Regression	0.66	0.67	0.66	0.66
K-Nearest Neighbors	0.75	0.76	0.75	0.75

Table 2: Comparison of accuracy

Model	Train Accuracy Score	Test Accuracy Score
Naïve Bayes	0.62	0.62
Decision Tree	0.99	0.89
Random Forest	0.92	0.84
Logistic Regression	0.66	0.66

K-Nearest Neighbors	0.84	0.75

All the models tested were competitive although they appears to be some overfitting in the Decision Tree, Random Forest and K-Nearest Neighbors, but because the data is balanced the accuracy score is equivalent to its recall value.

Conclusion

In conclusion, there is no specific justification for picking a particular algorithm, therefore the choice of algorithms for model implementation can include Logistic Regression, Naive Bayes classifier, Decision Tree, K-Nearest Neighbors, and Random Forest classifier. The dataset showed to be quite detailed in the areas where the models suggested whether an accident may be serious or not. Each model's accuracy is calculated and compared to the accuracy of the other models. When compared to the other models, the model trained using Decision Tree, Random Forest, and K-Nearest Neighbors has demonstrated good performance with an accuracy of 89%, 84%, and 75% in the prediction of accidents.

Recommendations.

From the Analysis of the data,

- Areas like the metropolitan areas should decrease the speed limits because of the population.
- Bikers should be made to use helmets to decrease fatalities.
- According to the analysis, there were a lot of pedestrian accidents at 8:00 a.m., 4:00 p.m., 5:00 p.m., and 3:00 p.m. This could be attributed to student movement to and from school. It will be necessary to teach students in schools the necessity of following traffic regulations. Additionally, at peak hours of pedestrian accidents, traffic wardens should be deployed to pedestrian accident-prone areas.
- Because pedestrians, motorcycles, and other types of vehicles have a high accident rate, traffic wardens will need to be more visible on Fridays to help monitor traffic.

References

- Akash Dubey(Feb 4, 2019) *Feature Selection Using Regularisation* Available Online: https://towardsdatascience.com/feature-selection-using-regularisation-a3678b71e499 [Accessed 20/04/ 2022].
- *CFI Education Inc.(2022)* [Accessed 20/04/ 2022].
- *https://corporatefinanceinstitute.com/resources/knowledge/other/lasso/* [Accessed 20/04/ 2022].
- R. Singh(2019). *UK accident Data.* Available online: https://www.kaggle.com/code/ambaniverma/uk-traffic-accidents [Accessed 15/03/ 2022]
- Lavanya R (2019)