Big Data and Data Mining Project (771762)

AYOMIDE OLAREWAJU

[202122882]

INTRODUCTION

Road traffic fatalities and the resulting impairments, loss of life, and property damage are becoming a major public health concern in many countries. According to the World Health Organization (WHO, 2018), 1.3 million people die every year as a result of traffic accidents. Non-fatal injuries affect between 20 and 50 million more people, with many of them resulting in disability as a result of their injury.

DATA SOURCE

Our data was sourced from the UK Government's Data Repository, it consists of the accident, vehicle and casualty files with the 3 datasets been linked by the 'Accident Index column'.

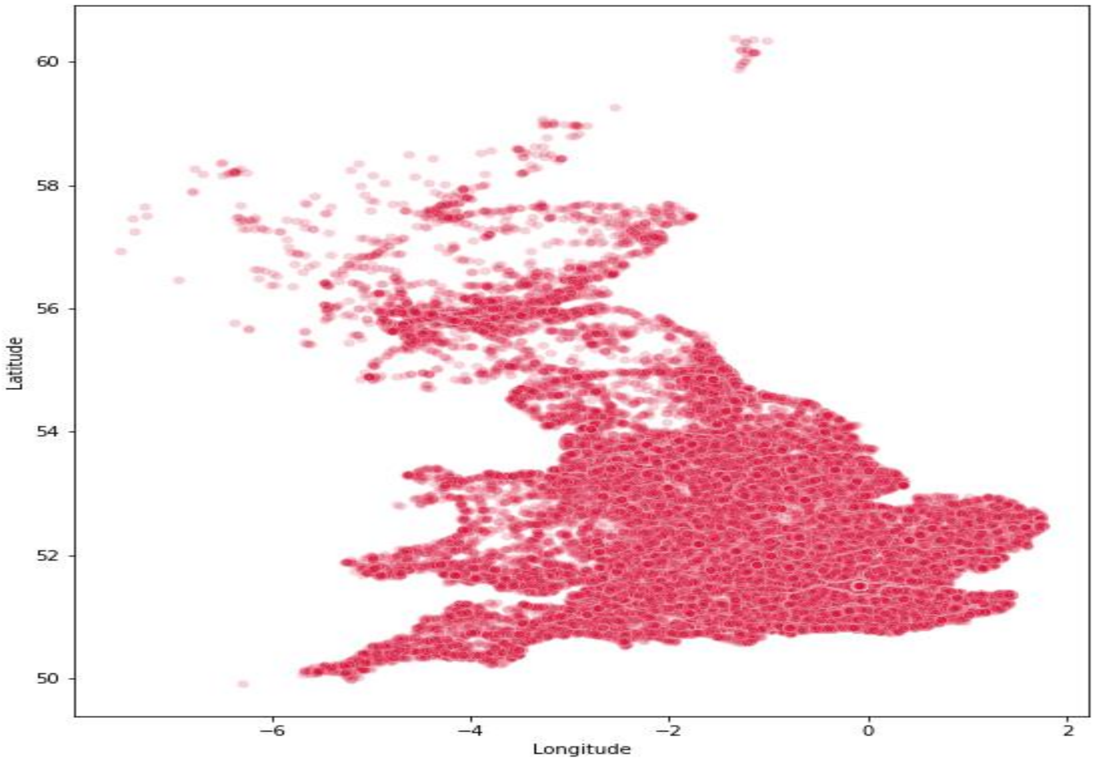


Figure 1 | Distribution of Accident Data

Figure 1 shows a general macroscopic understanding of the distribution of the entire accident data according to the latitude and longitude information our dataset

Using the above dataset from 2019, the project will use data science methodology to gain an understanding of the critical factors that lead to occurrence of accidents, develop insights and prevention mechanisms for Traffic Accidents and Road Safety, and use the enormous power of machine learning to build models that can predict the severity of accidents and prevent them before they happen.

METHODOLOGY

In order to have a great accuracy in our prediction, our dataset will undergo pre-processing which includes data cleaning, & data normalization, feature selection to ensure only important features are fed to the model, and lastly predicting accident severity using machine modelling.

Data Cleaning

There were 117,536 rows and 32 columns in the accident dataset, 216,381 rows and 23 columns in the Vehicle dataset, and 153,158 rows and 16 columns in the Casualty dataset.

The date, time, local authority, and location of each accident were all categorical in the accident dataset, aside from the index number. The date and time were correctly formatted. Missing (NaN) values were detected as shown in figure 3 and replaced using mean in the case of latitude & longitude and median in the case of time, while forward fill was employed for the locations. Duplicates were also found and eliminated. Outliers were handled using quantiles and the interquartile range (IQR) for some features, such as the Engine capacity column, which had a skewed distribution (Figure 4&5).

The dataset had some missing or out-of-range data points represented by -1 (figure 6) which were cleaned using the mean, mode, and median procedures. For both the vehicle and casualty datasets, the cleaning process was carried out in the same way. For features such has Junction control which had over 44% as negative figure, they were not cleaned using the mode and standard deviation procedures, rather they were labelled as Data missing or out of range and further converted to a positive binary figure during model training.

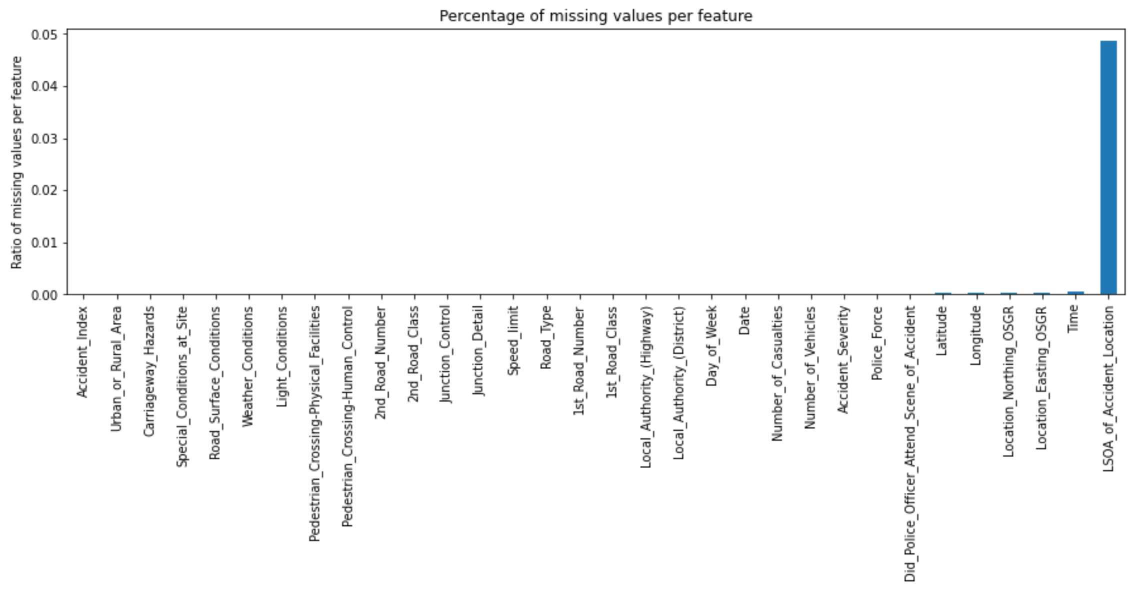


Figure 3 | Percentage of missing values per feature

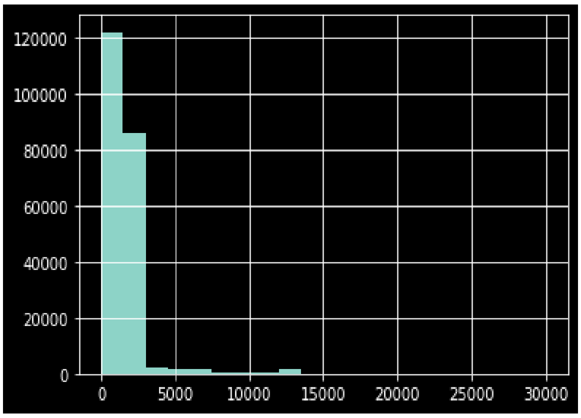


Figure 4 |Skewed Engine capacity distribution

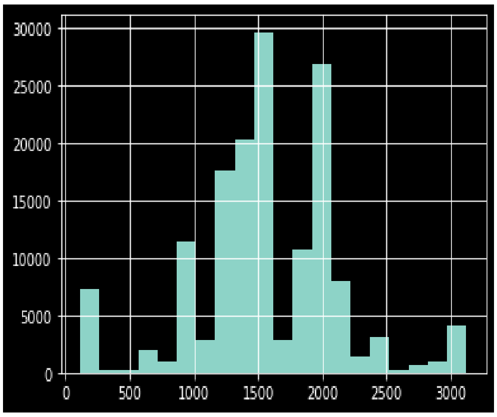


Figure 5| Normalized Engine capacity distribution

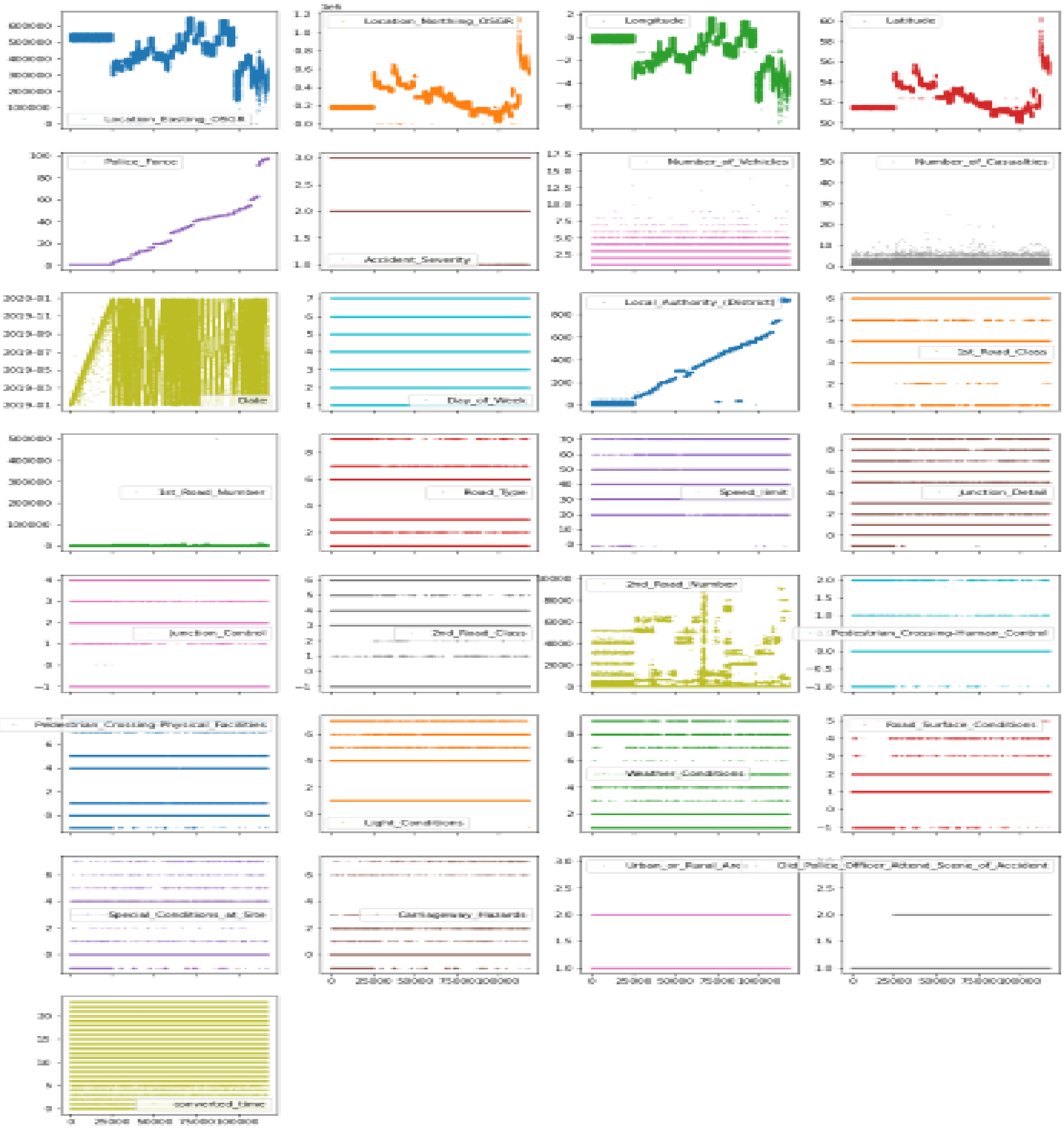


Figure 6 | Plot showing all features in accident dataset

DATA EXPLORATION

The severity of an accident is determined by the severity of the injured casualty: Slight, Serious, and Fatal are the severity levels with slight injuries been sustained in about 78% of the accident data.

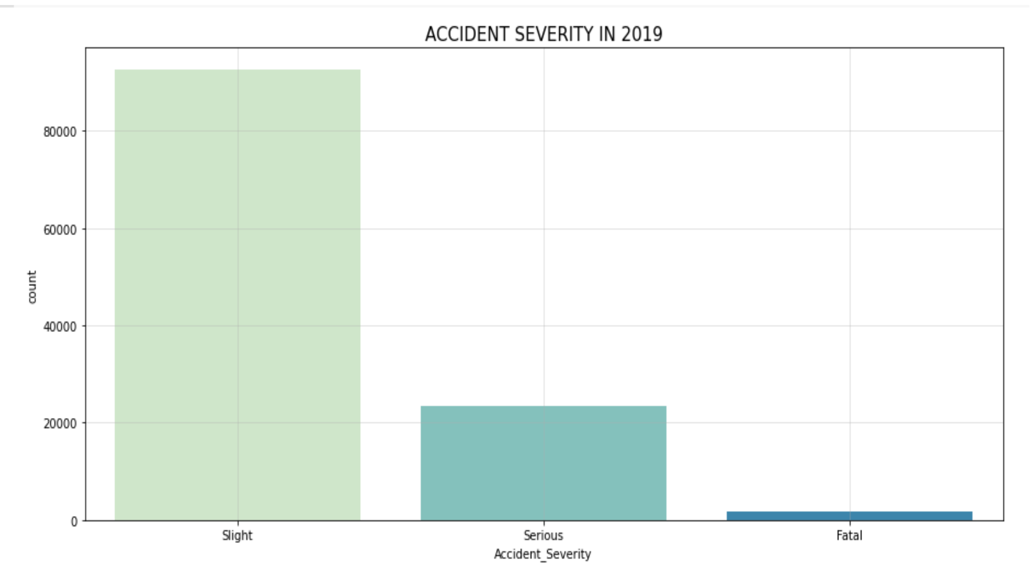
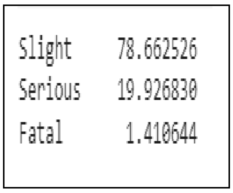


Figure 7| Accident severity count



```
Slight     78.662526
Serious    19.926830
Fatal       1.410644
```

Significant hours of the day and week on which accidents occur

Accidents occurred at different locations in the UK with more concentration on Latitude 51.5 and longitude -0.089, a location in the city of London as shown in figure 10. Most incidents happened between 15:00 (3pm) and 17:00 (5pm) most weekdays, which is the closing period for most workers. As expected, there are more accidents on weekdays, maybe due to more people going to work, resulting in more drivers on the road, with a peak on Friday evenings, the end of the weekdays, and lower incidences on weekends.

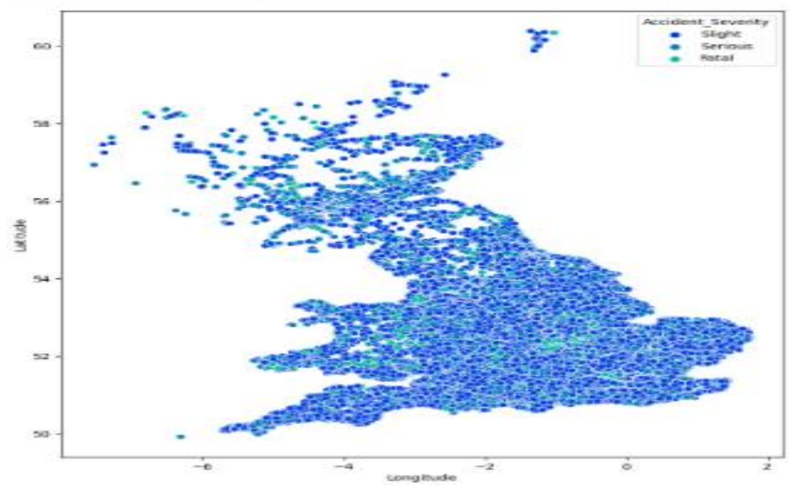


Figure 8 | Accident severity distribution plot

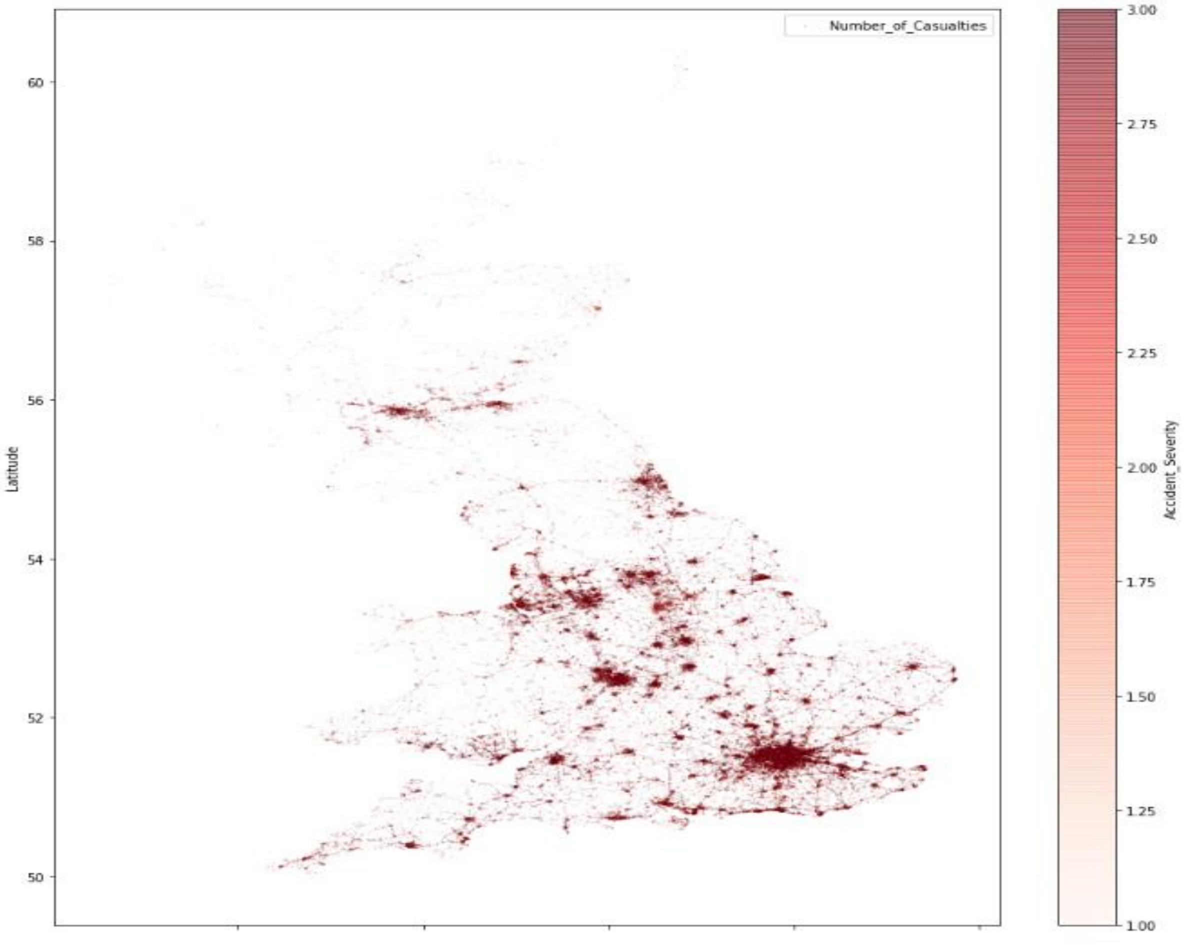
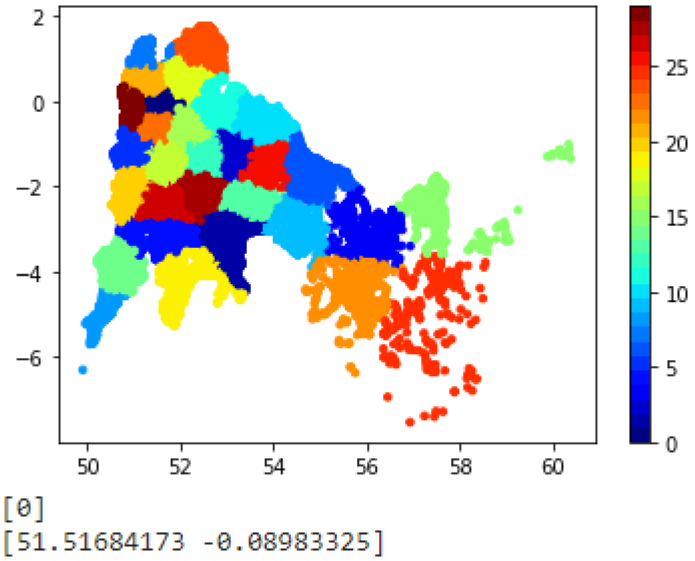


Figure 9 | Plot showing concentration of accident



```
[0]
[51.51684173 -0.08983325]
```

Figure 10 | Plot stating specific latitude and longitude of frequent accident.

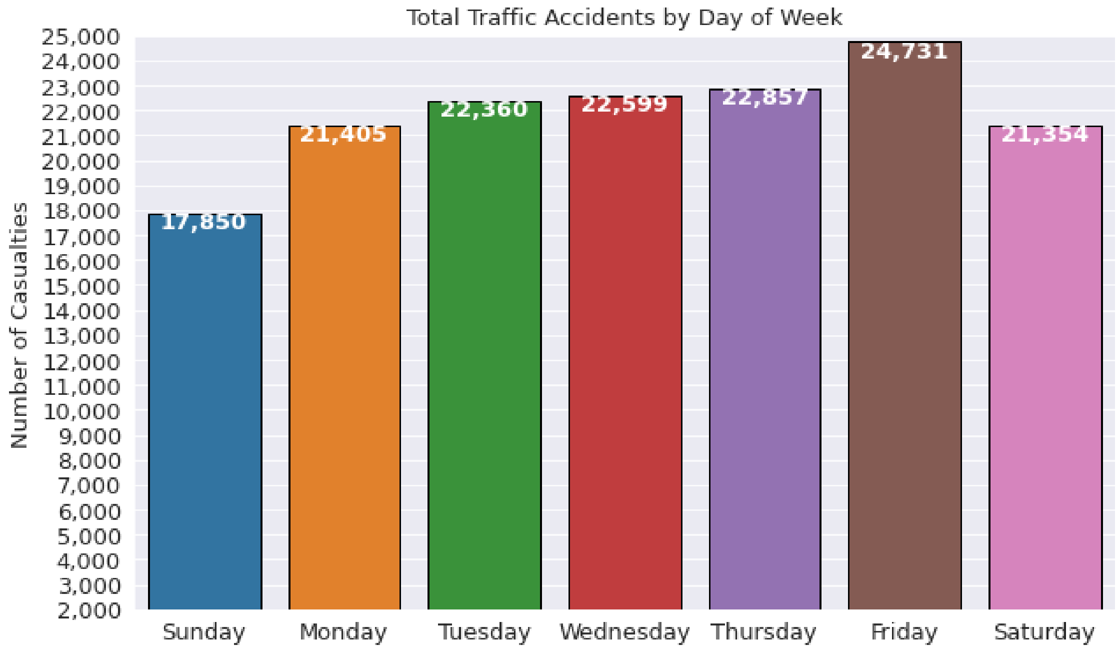


Figure 11 | Traffic accidents by Day of Week

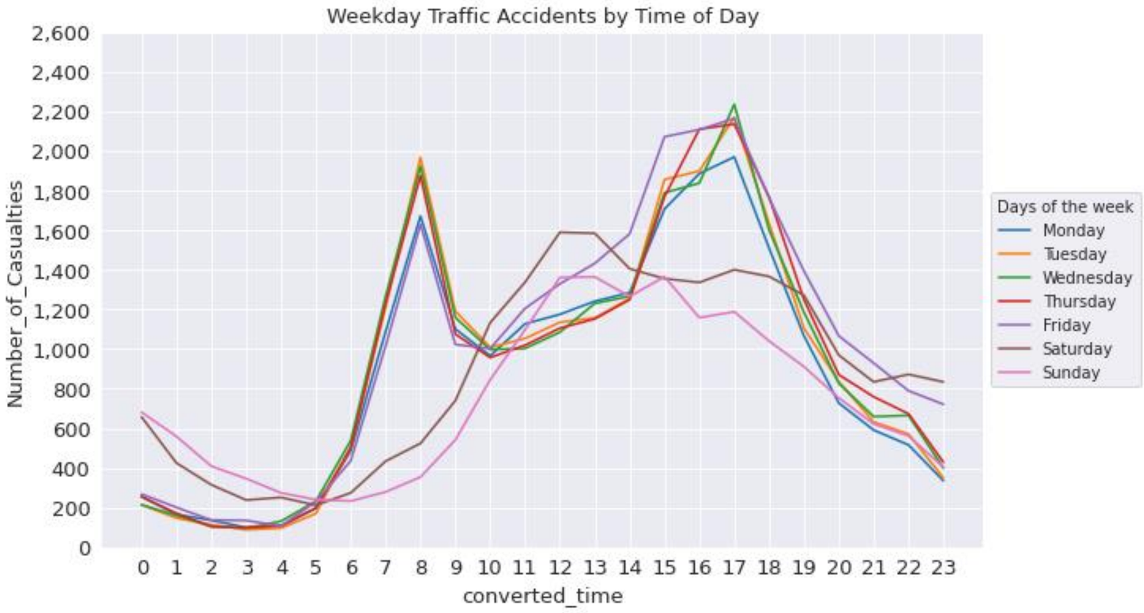


Figure 12 | Traffic accident by time of day

Significant hours and days of the week for motorcycle accidents

As depicted in Figure 13, vehicle type bikers (pedal, motorcycle) come in second as the most frequent mode of transportation involved in accidents and casualties, but additional investigation reveals an increase in motorcycle accidents on Fridays at 17:00.

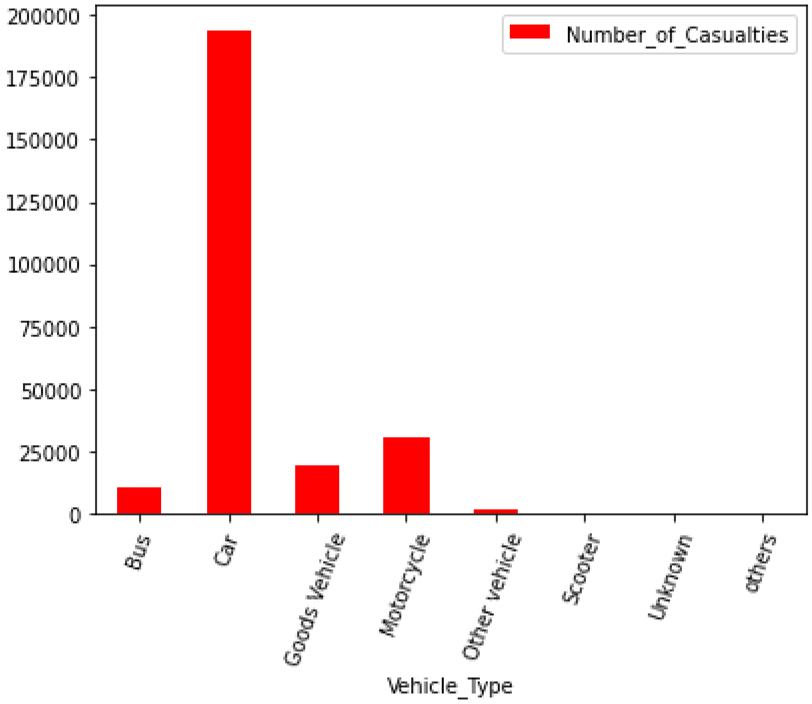


Figure 13| Plot showing Motorcycle as the second Vehicle type known for Accident Severity

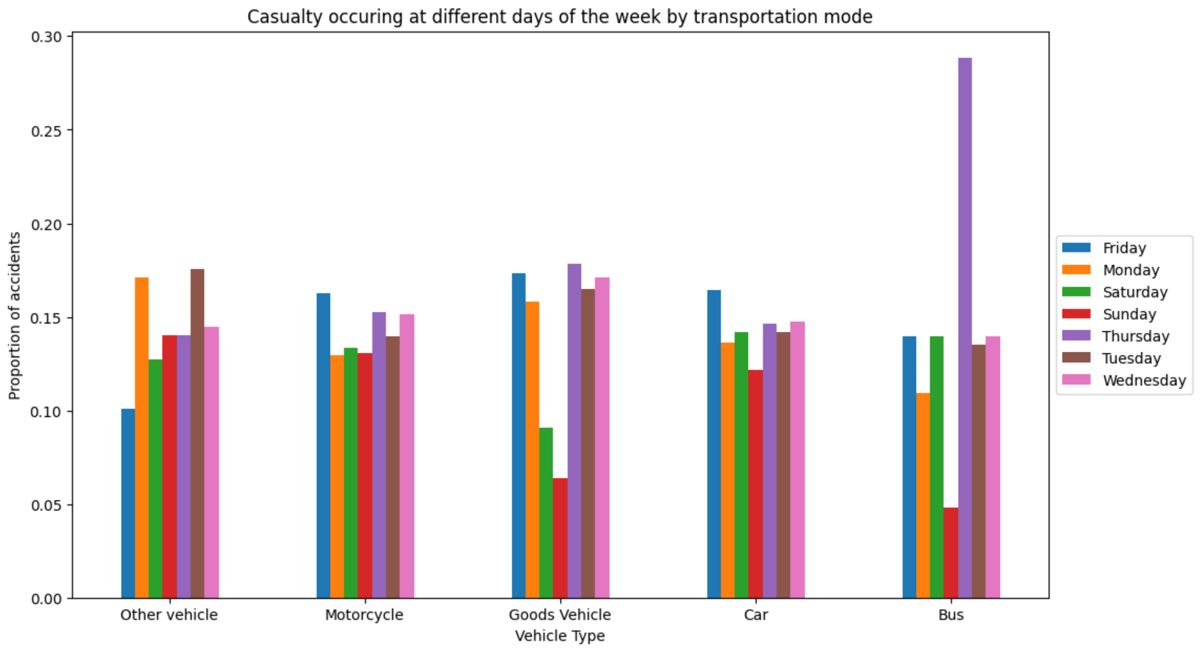


Figure 14| Plot showing frequency of accidents per vehicle type within the week

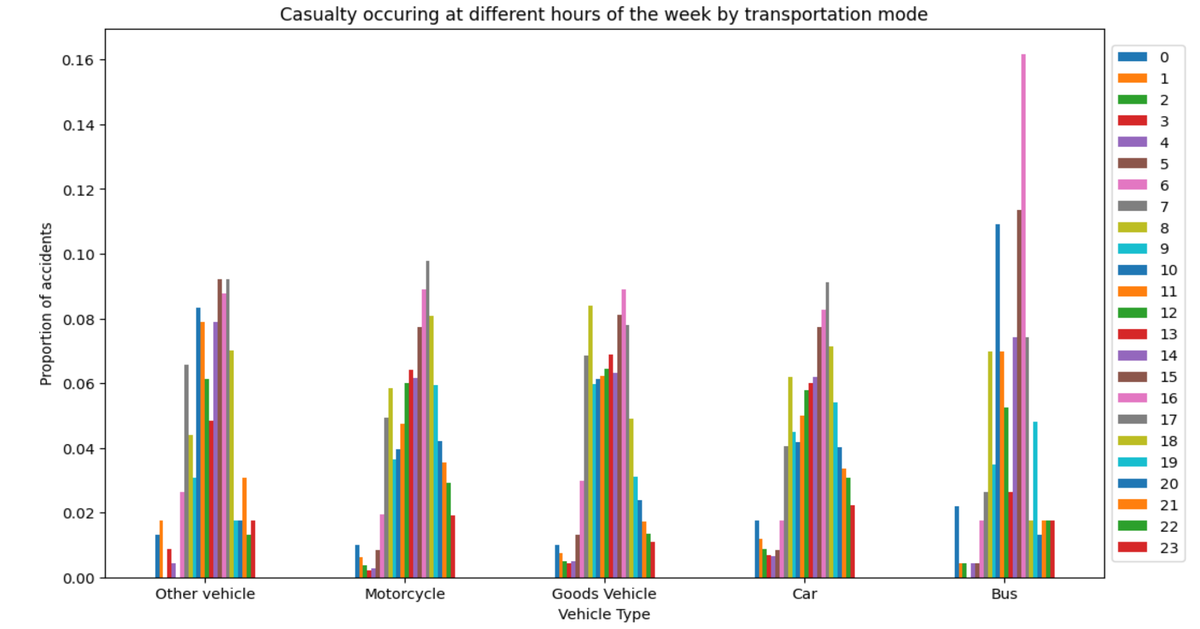


Figure 15 | Plot showing hourly accident per vehicle type

Significant hours of the day and week notorious for pedestrian accidents

Within the casualty class, the driver/rider had a greater percentage of casualties than passengers and pedestrians; however, as seen in the plots below, pedestrians were more prone to accidents on Fridays at 15:00, while passengers were more likely to be casualties on Saturdays.

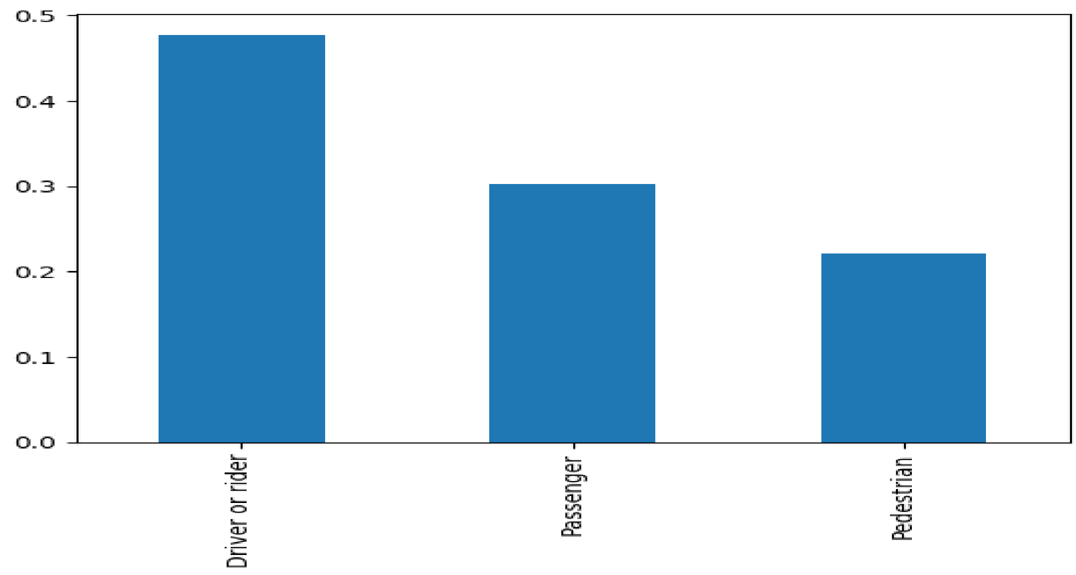


Figure 16 | Plot showing the proportion of accident by casualty class

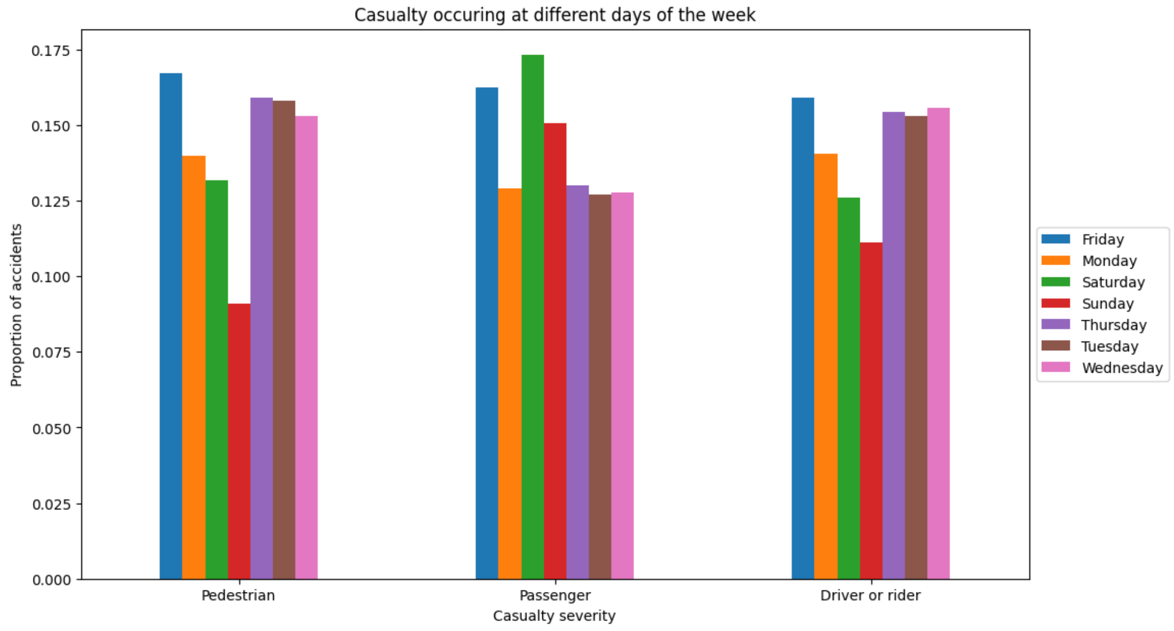


Figure 17 | Plot showing cumulative proportion of accidents during the week by Casualty class

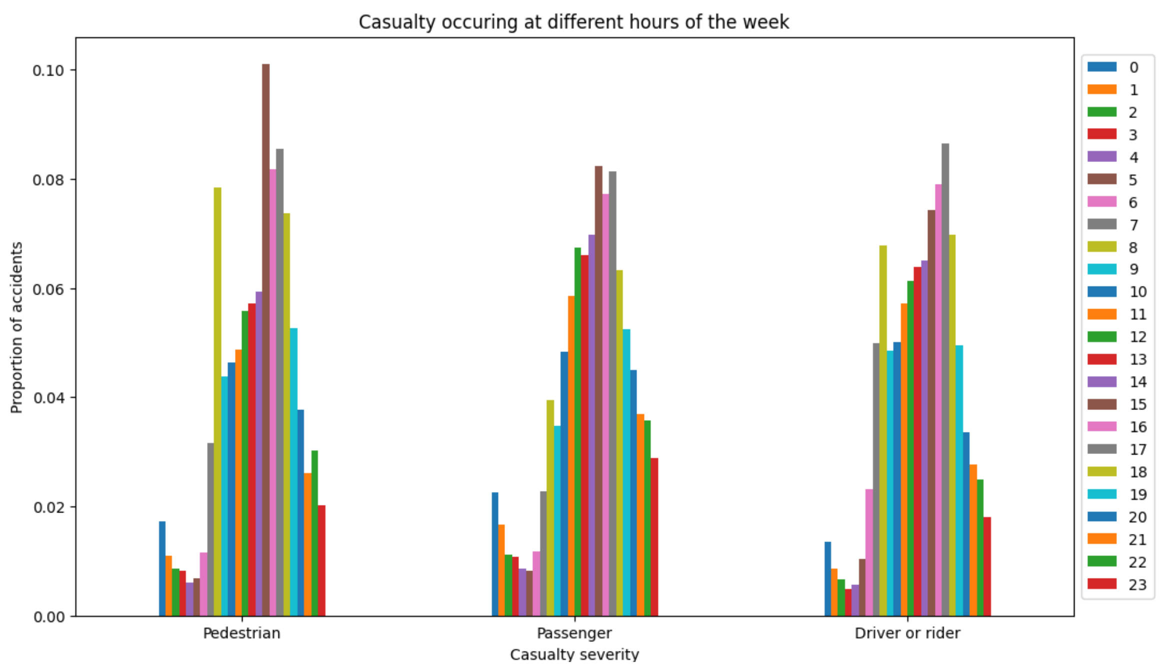


Figure 18 | Plot showing cumulative proportion of accidents per hour by Casualty class

Impact of daylight savings on road traffic accidents

The sun rises and sets one hour later than usual during daylight savings time (DST), resulting in darker mornings and brighter evenings. According to my findings, the likelihood of a road traffic collision is highest in the late afternoon and evening hours (15:00 to 17:00 hours), which may be attributed to the interplay of decreasing lighting conditions and other risk factors before DST. There were 3,502 accidents the week before DST began, and 1,306 after DST ended. Pedestrians were the greatest beneficiaries of the DST, as shown in Figure 19 (0.0 signifies not a pedestrian, while other variables can be categorised as pedestrian). However, fatalities for pedestrians were substantially lower when the DST stopped. Using the casualty type as a filter as shown in figure 20, it was discovered that pedestrians had more serious accidents during the DST while they had more slight accidents when DST ended.

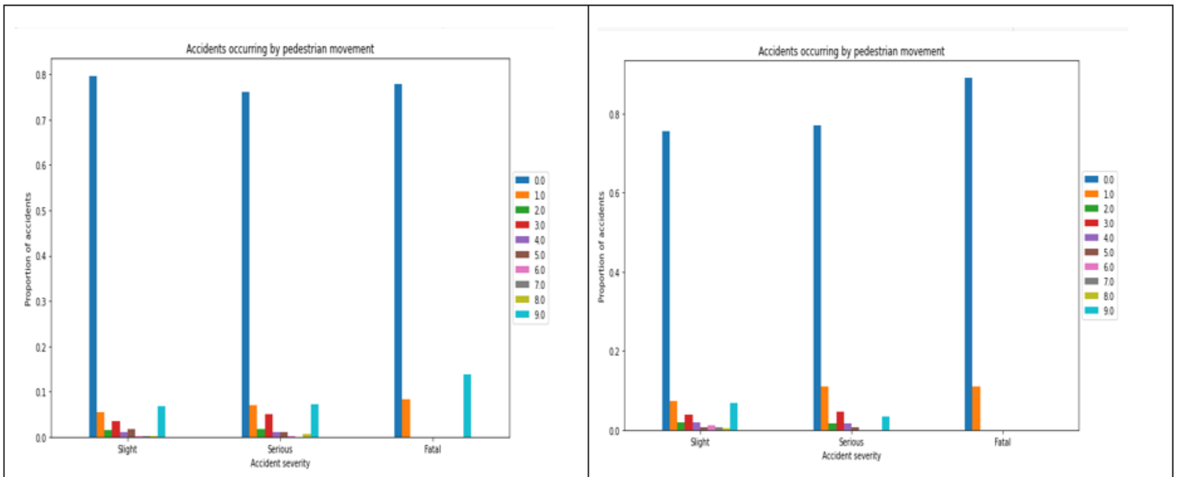


Figure 19 | Daylight Saving conditioned on Pedestrian movement

Daylight Ending conditioned on Pedestrian Movement

Impact of Sunrise and sunset on road traffic accident

Drivers are big casualties on DST and non-DST days, according to the graph below. This shows that light is rarely the direct cause of a traffic accident; rather, light and darkness tend to exacerbate the effects of more direct causes. We can say that poor lighting conditions degrade driver performance because of slower visual reaction times and the inability to process critical information like critical stopping distances, but we can also say that collisions are caused by driver error, which can occur in both ambient and dark conditions but is amplified in the latter(Plainis et al., 2006). Light can also interact with environmental factors like weather, heightening the risk of a collision.

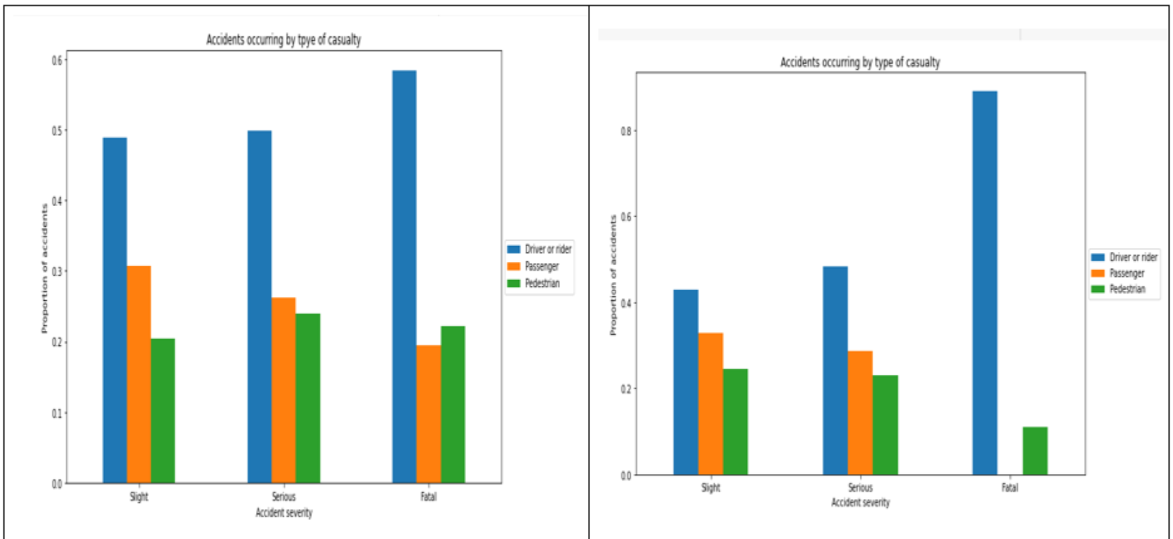


Figure 20 | Daylight Saving conditioned on Casualty Type

Daylight Ending conditioned on Casualty Type

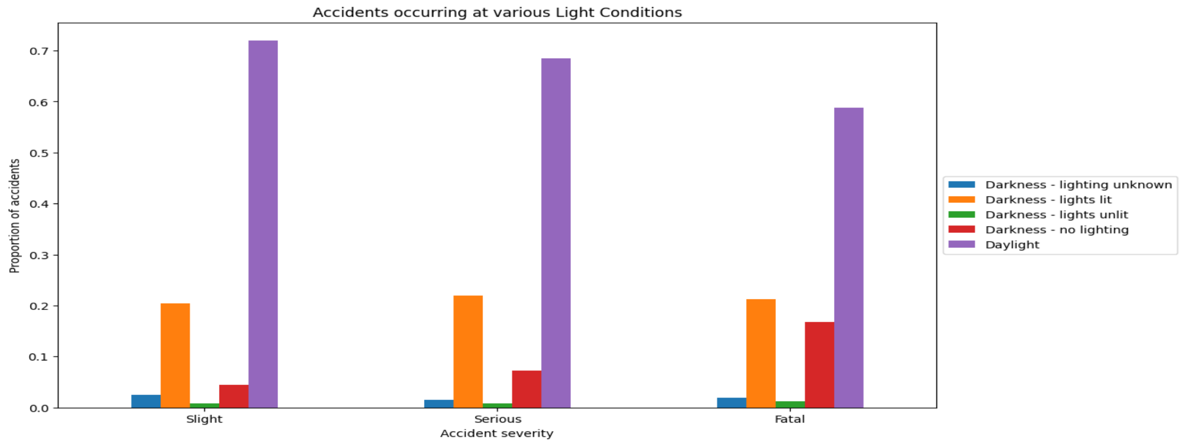


Figure 21 | Impact of Light conditions on Accident & Severity.

Vehicle related impact on traffic accidents

Cars below 13 years were mostly involved in accidents with a peak on cars used for 8years. Right-hand-drive petrol cars are also more likely to be involved in traffic accidents, particularly at uncontrolled intersections or places. During the year, cars with engine capacities of 1500-2000cc were involved in more incidents as shown in Figure 5 with most of these accidents caused by drivers whom we cannot identify the purpose of their journey.

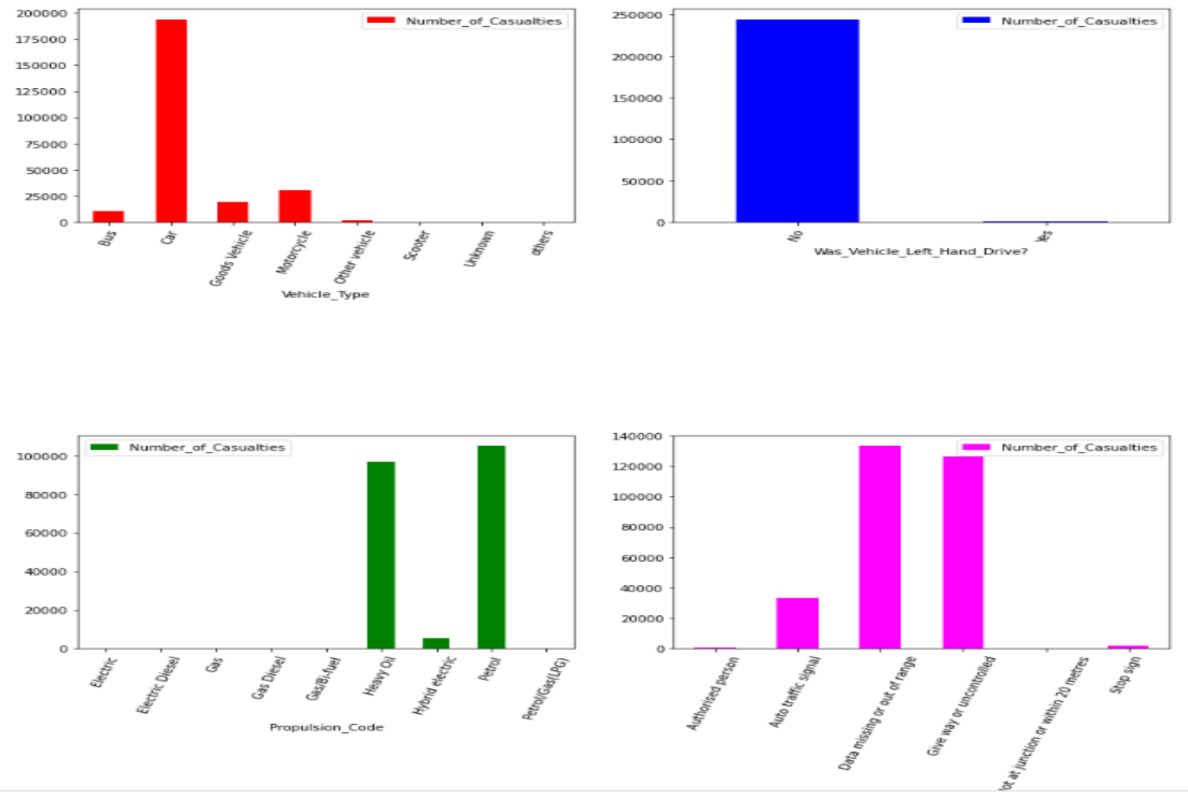
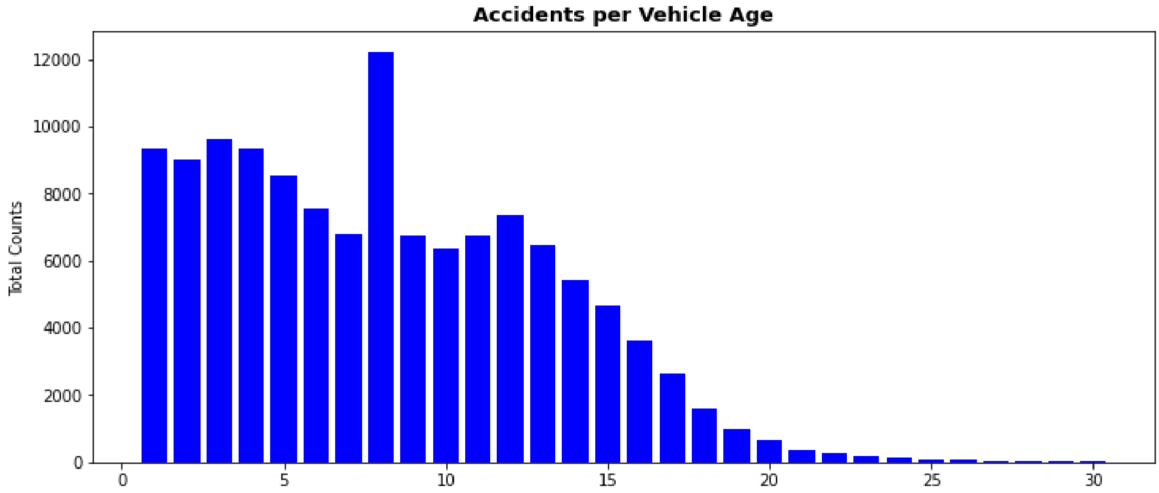


Figure 22| Plots showing impacts of specific vehicle features on accidents.



Conditions that generate more road traffic accidents

A variety of factors influence road traffic accidents. The majority of incidents occur on class A highways and single carriageways, increased speed limits resulted in more fatal accidents as shown in figure 26 although speed limit of 30mph was responsible for a greater percentage of the accident which is usually slightly severe, but contrary to popular belief, most accidents occur during daylight hours and in good weather. We can simply infer that on days when the weather is nice, people go out more during the day, resulting in a higher population on the road and more accidents while bad weathers discourages people from going out and driving carefully when they do. Traffic accidents were more in urban areas, which are often highly inhabited, than in rural areas as well as locations which were not a junction.

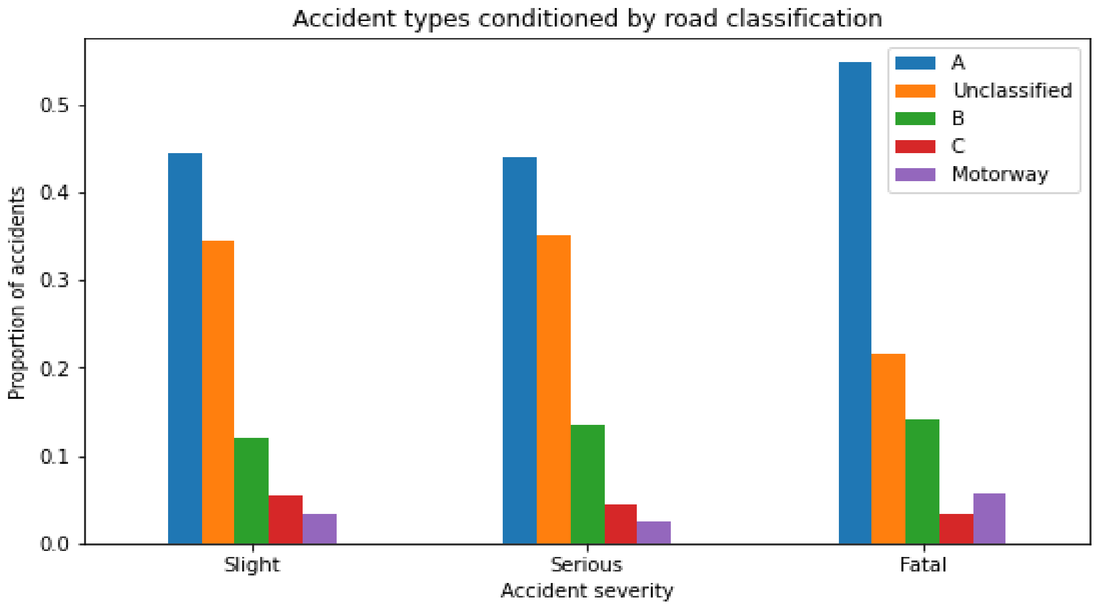


Figure 23 | Accident severity by road classification

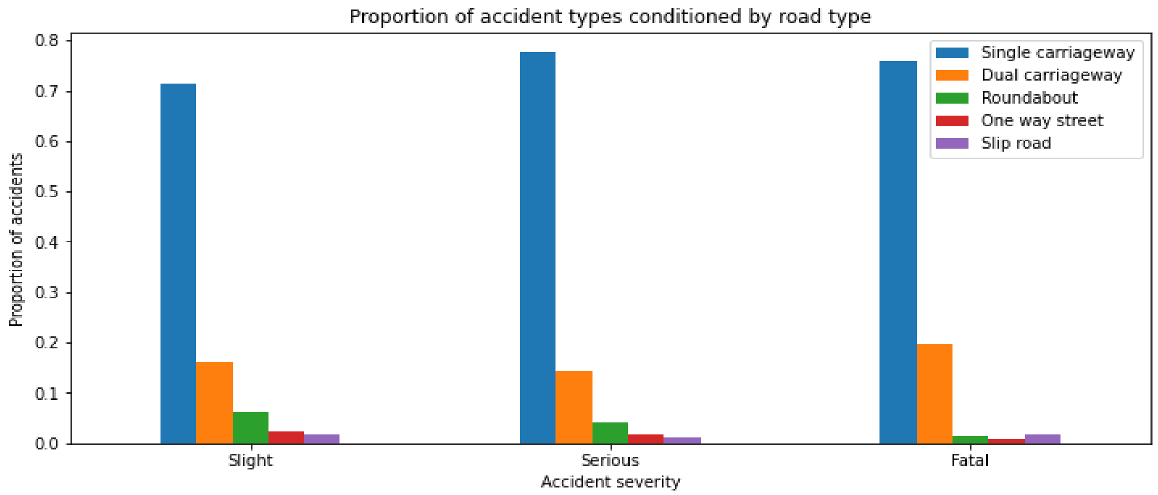


Figure 24 | Accident severity by Road Type

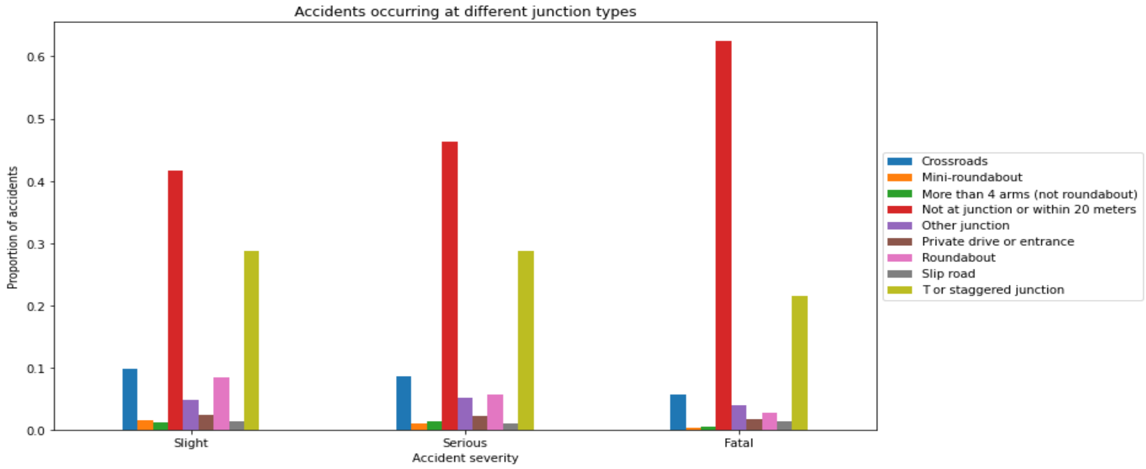


Figure 25 | Accident Severity by Junction Types

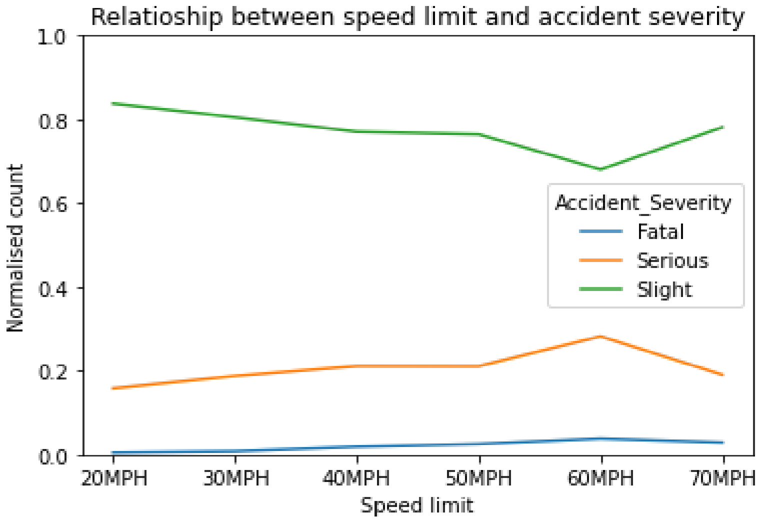


Figure 26 | Accident severity and speed limit

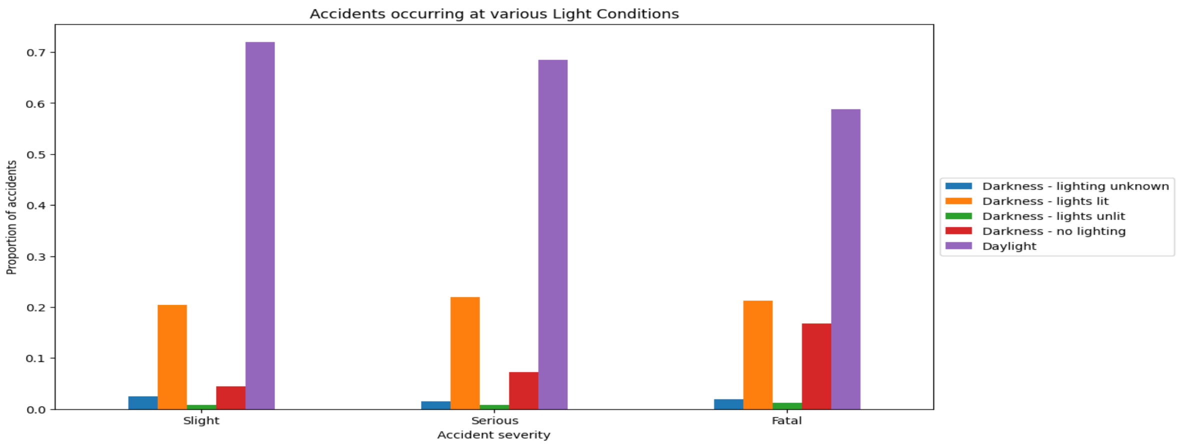


Figure 27 | Impact of Light conditions on accident

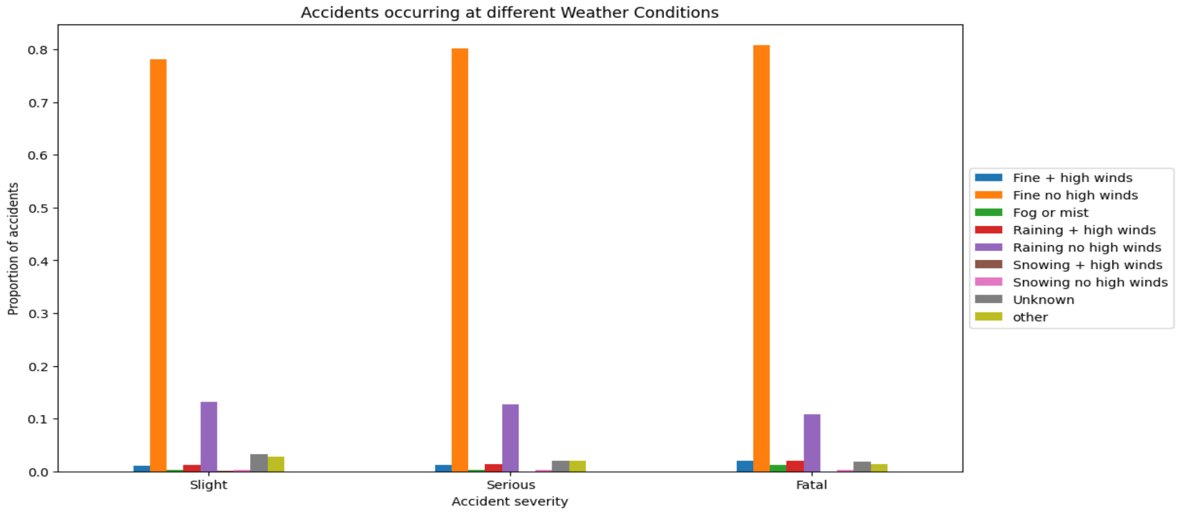


Figure 28 | Impact of Weather conditions on accident

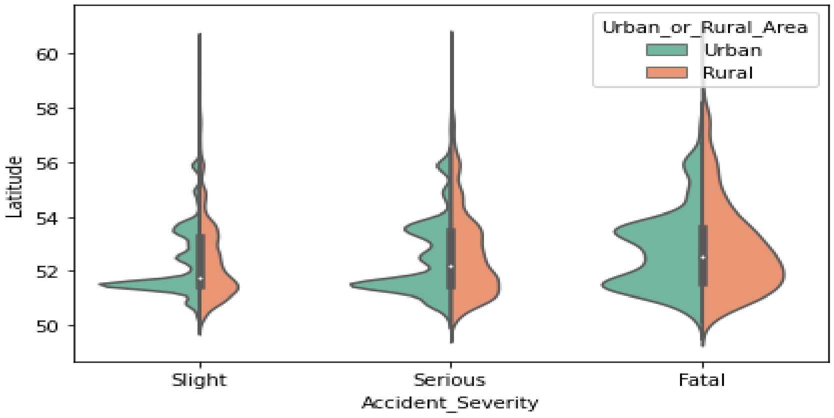


Figure 29 | Concentration of accident by Area

Impact of Driver related variables on accidents

Drivers between the ages of 26 and 45 were the most prevalent in accidents, with a focus on those aged 37. The mean age was similarly in the 40s, and the majority of the incidents could be traced back to drivers whose destination was unknown.

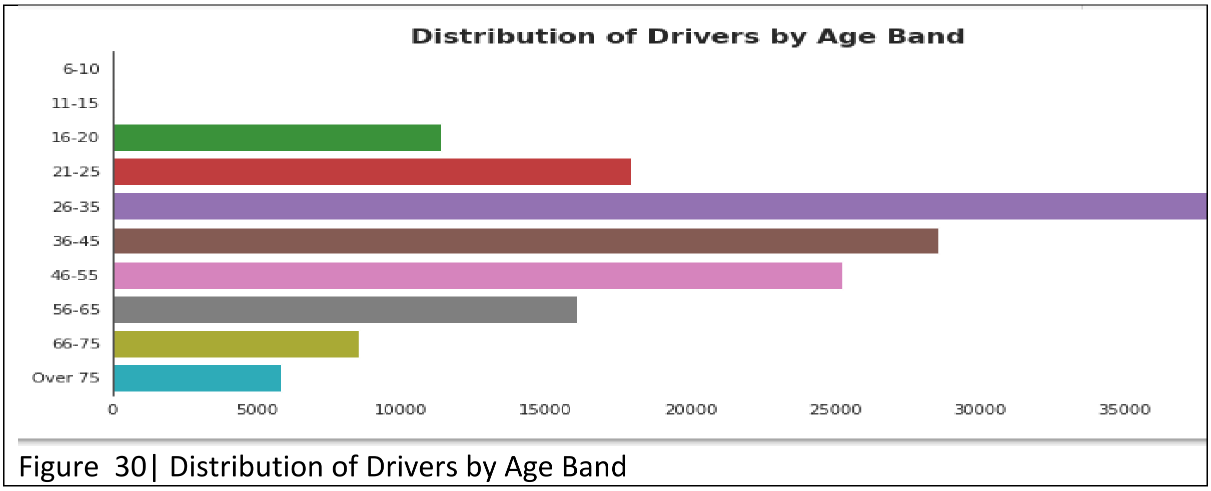


Figure 30| Distribution of Drivers by Age Band

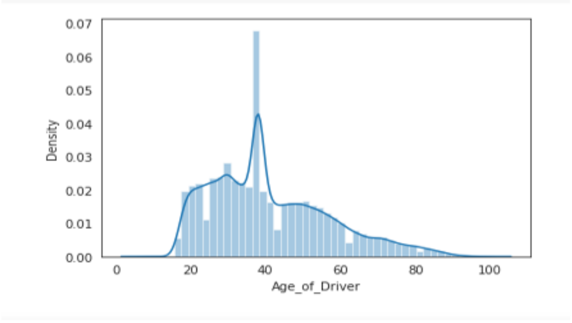
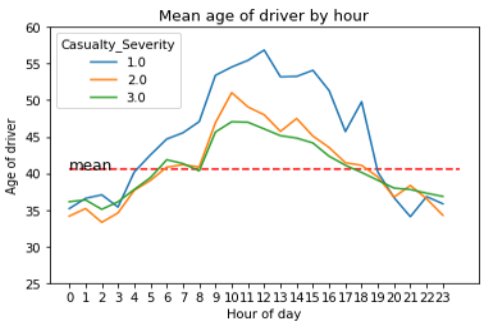


Figure 31 | Age of Drivers in Road Accident



Mean Age of Driver in Road Accident

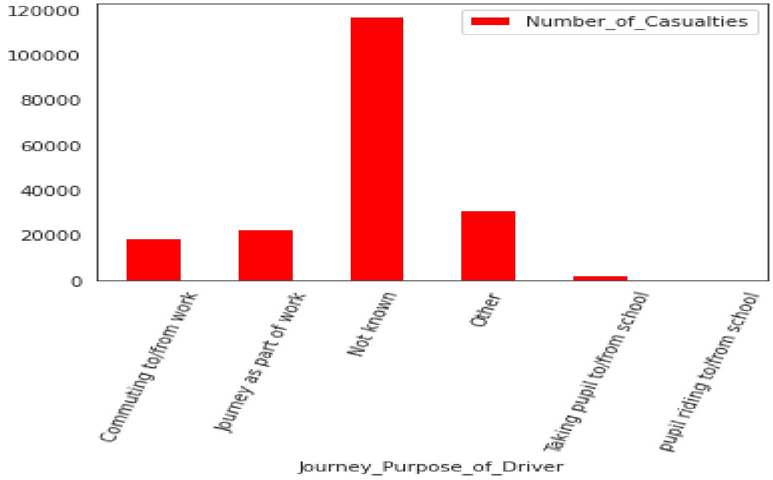


Figure 32| Plot showing Journey purpose of Driver

MODELLING AND ACCIDENT PREDICTION

Data Correlation & Feature selection

Our dataset comprises 64 features that are almost independent of one another, indicating that it is extremely complex and that not all elements are significant for accuracy prediction.

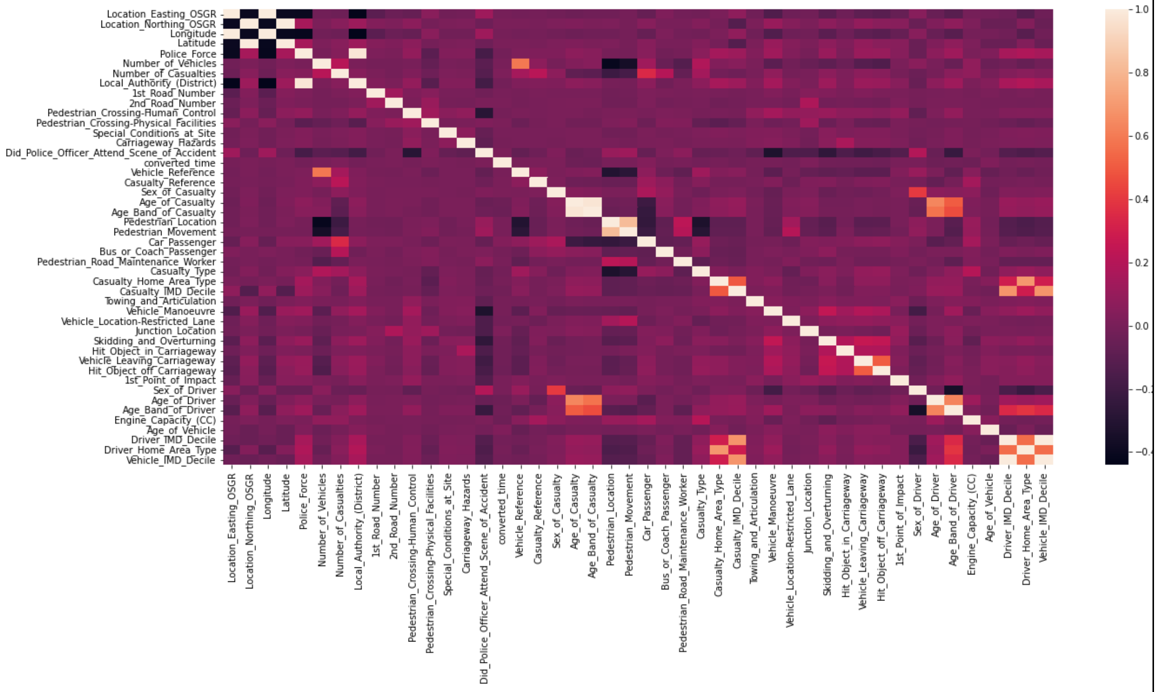


Figure 33 | Correlation Map of features in the dataset

The principal component analysis (PCA) was used, however it did not reduce columns as predicted, so the Random Forests (RFs) method was used to pick features based on their importance index

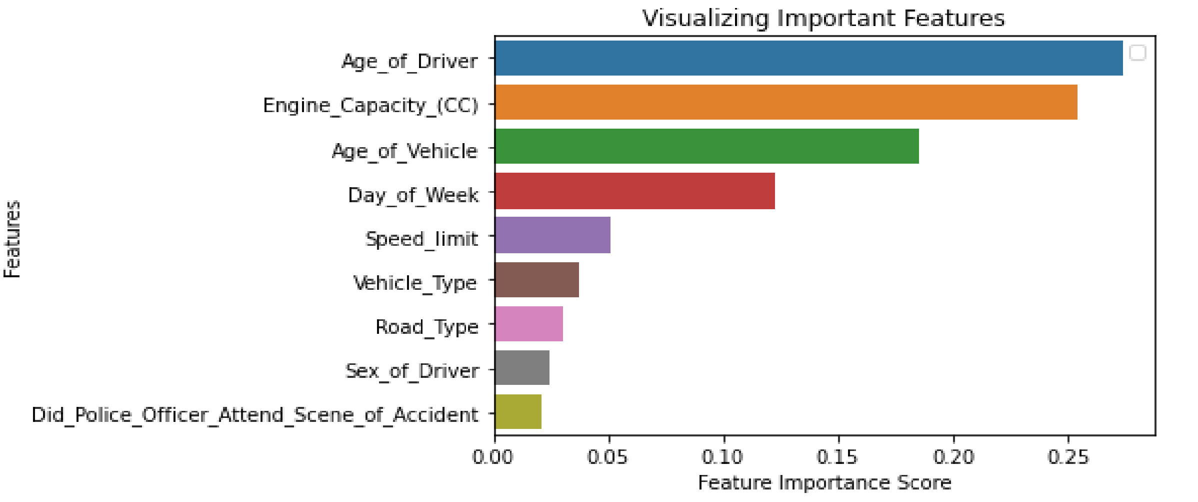


Figure 34| Plot of Feature Importance

Oversampling technique known as SMOTE was used to balance the dataset's imbalanced class Accident Severity, which raised the number of minority class samples (serious, fatal) while random undersampling was employed to diminish the majority until the dataset was balanced.

On the unbalanced, oversampled, and undersampled training data, four algorithms were trained (Decision tree, KNN, Logistic regression, and Random forest). TPR, FPR, TNR, precision, recall, overall accuracy, and the confusion matrix were generated for each method, as were the performance metrics. The models of each training set were stacked for better accuracy.

TABLE 1 | RESULT

MODELS	UNBALANCED		OVERSAMPLED		UNDERSAMPLED	
	DEFAULT	STACKED	DEFAULT	STACKED	DEFAULT	STACKED
RF	78.05	78.1	81.03	79.0	34.17	89.7
LR	76.75	77.1	53.26	52.8	21.1	84.7
DT	70.27	69.9	77.5	75.2	49.6	83.1
KNN	73.87	74.1	76.5	72.9	23.3	83.7
STACKED	-	78.7	-	81.3	-	90.5

As shown in Table 1, The best classifiers in predicting accidents is the undersampled stacked model with 90.5%, however, considering accuracy is not the only performance metrics, the random oversampled-based models showed better predictive capabilities in detecting the minority class in the confusion matrix than the algorithms trained by the unbalanced and undersampled dataset.

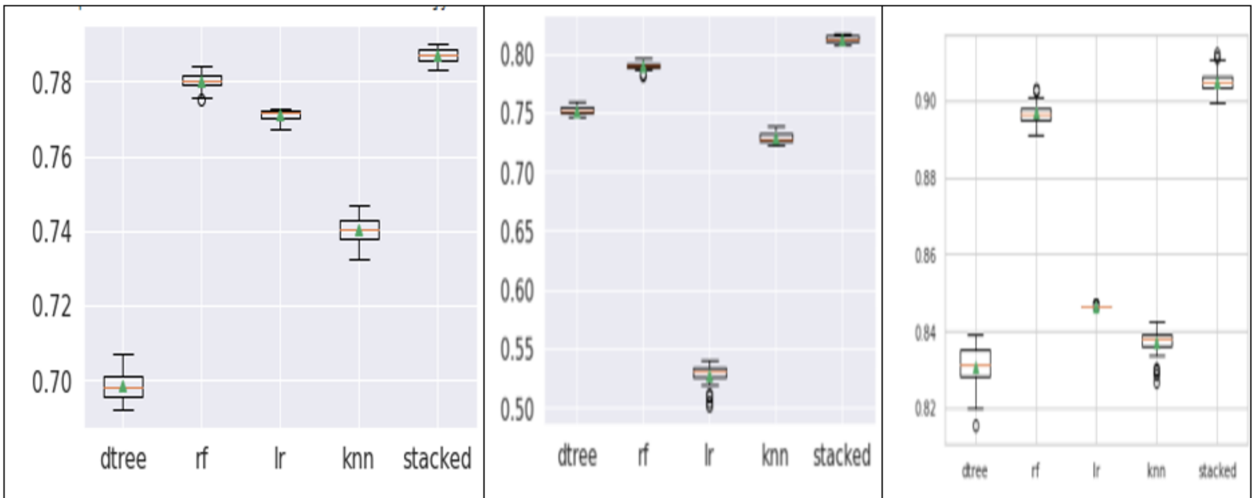


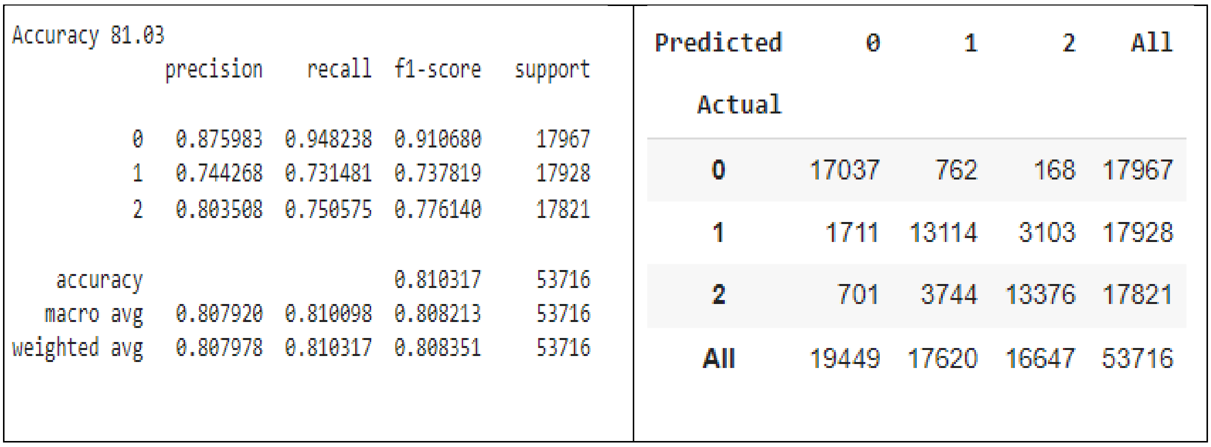
Figure 34 | Box plot of imbalanced dataset

Box plot of oversampled dataset

Box plot of undersampled dataset

The random forest classifier for the oversampled dataset as shown in Table 2 had a precision of 81% and recall of 94% for the minority class (Fatal) which implies the model returned relevant results labelled correctly.

Table 2 | Confusion matrix



```
Accuracy 81.03
              precision    recall  f1-score   support

           0   0.875983  0.948238  0.910680     17967
           1   0.744268  0.731481  0.737819     17928
           2   0.803508  0.750575  0.776140     17821

    accuracy                       0.810317     53716
   macro avg   0.807920  0.810098  0.808213     53716
weighted avg   0.807978  0.810317  0.808351     53716
```

Predicted	0	1	2	All
Actual				
0	17037	762	168	17967
1	1711	13114	3103	17928
2	701	3744	13376	17821
All	19449	17620	16647	53716

COMPARISM TO GOVERNMENT MODEL

Government model gave 91.8% prediction which is better than my stacked model of 90.5% using the under sampled dataset.

CONCLUSION

Because traffic accident severity prediction is vital for accident management, we can conclude that machine learning is a potential method for predicting and ultimately avoiding road traffic accident severity. However, to accurately predict higher-severity incidents, balanced dataset is required.

RECOMMENDATION

- Traffic congestion in urban areas should be reduced through a variety of methods and initiatives, including improved bus service, workplace parking fees, and existing rail networks.
- Speed Limit: Installing speed and highway cameras to catch vehicles exceeding speed limits, particularly on single carriageways, which are infamous for accidents, and enforcing speeding fines.
- Safety Education: Proper road safety education is critical, so it’s important to promote knowledge about road safety protocols and best practises.
- Reporting Channels: Information on how to report dangerous drivers as well as penalties for reckless driving should be readily available.
- Proper traffic management: The majority of accidents seemed to have happened at uncontrolled junctions due to the absence of proper traffic management mechanisms like traffic signals and bumps, these should be considered by the government.

REFERENCES

Carey, R. N., & Sarma, K. M. (2017). Impact of daylight-saving time on road traffic collision risk: a systematic review. Available online: https://doi.org/10.1136/bmjopen-2016-014319 [Accessed 2/04/2022]

David B. Zwiefelhofer (2022) FindLatitudeAndLongitude. Available online: https://www.findlatitudeandlongitude.com/l/51.5+-0.13/2912833/ [Accessed 2/04/2022]

Gan, J., Li, L., Zhang, D., Yi, Z. & Xiang, Q. (2020) An alternative method for traffic accident severity prediction: Using deep forests algorithm. *Journal of Advanced Transportation,* 2020 1257627.

Laura Lewis (2020) Predicting traffic accidents-CNN. Available Online : https://github.com/L-Lewis/Predicting-traffic-accidents-CNN [Accessed 2/04/2022]

Micheal P. Notter (2022) Advanced exploratory data analysis. Available online: https://miykael.github.io/blog/2022/advanced_eda/ [Accessed 2/04/2022]

Plainis, S., Murray, I. J. & Pallikaris, I. G. (2006) Road traffic casualties: Understanding the night-time death toll. *Injury Prevention,* 12 (2), 125-138.

World Health Organization (2018), *Global status report on road safety*, World Health Organization, Geneva, Switzerland, 2018. Available online: https://www.who.int/publications/i/item/9789241565684 [Accessed 2/04/2022]

Timeanddate (2022) Clock changes in London, England, United Kingdom 2019. Available online: https://www.timeanddate.com/time/change/uk/london?year=2019 [Accessed 2/04/2022]