[Year]

UK ROAD TRAFFIC ANALYSIS AND PREDICTIONS

MODULE NAME: BIG DATA & DATA MINING

SIMIDOLA LAWANI

UNIVERSITY OF HULL | ID: 202110433

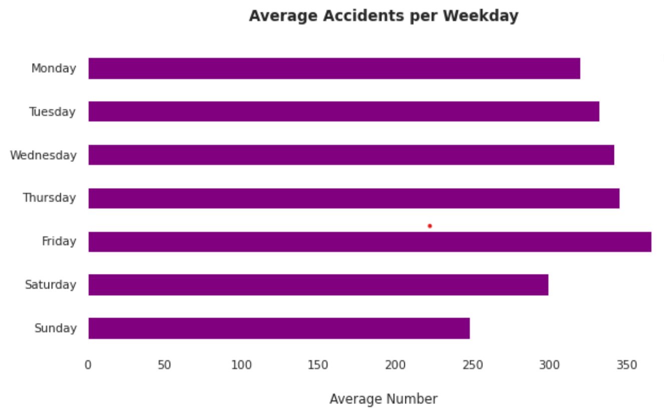
Contents

INTRODUCTION

The UK road safety 2019 of three datasets will be used in this project. The datasets are: **Road Safety Data - Accidents 2019, Road Safety Data - Casualties 2019, and Road Safety Data- Vehicles 2019.** The number of entries in the Road Safety Accident data was initially 117536, but during data cleaning, some null values were eliminated from the dataset to avoid data inconsistency and errors during data presentation and prediction modelling. To address certain issues, data analysis will be carried out in the first stage of the project by determining where, when, and why accidents occur in the UK, with the use of some visuals generated from the dataset. The focus of the second portion of this project will be on modelling utilising certain scikit-learn algorithms, with machine learning modelling being used to determine which of the algorithms performs better depending on the results each algorithm provides.

ANALYSIS AND VISUALIZATION

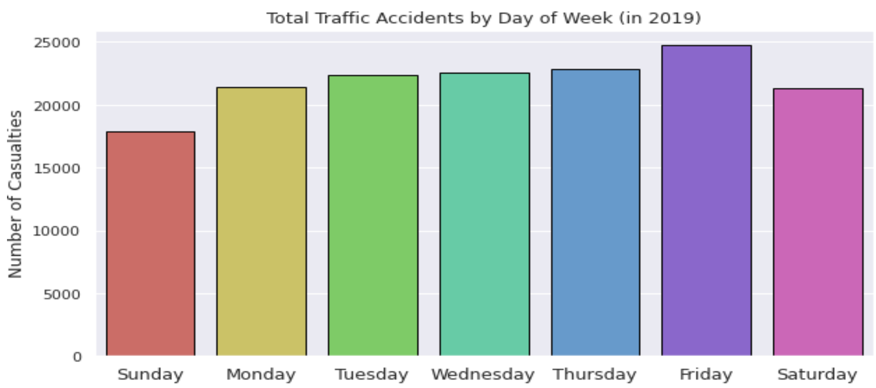
Hour of the day, and days of the week, on which accidents occur.



Weekday	Average_Accidents
Friday	366.134615
Monday	319.923077
Saturday	299.038462
Sunday	248.557692
Thursday	345.461538
Tuesday	331.943396
Wednesday	341.653846

Figure 1: Average Number of accidents per weekday

From the visualization generated in figure 1, it can be noticed that Friday has the highest average number of accidents per weekday



In fig 2, it is evident from the visualization that Friday has the highest number of casualties based on the accidents that occurred in 2019.

Figure 2: Number of Casualties per weekday

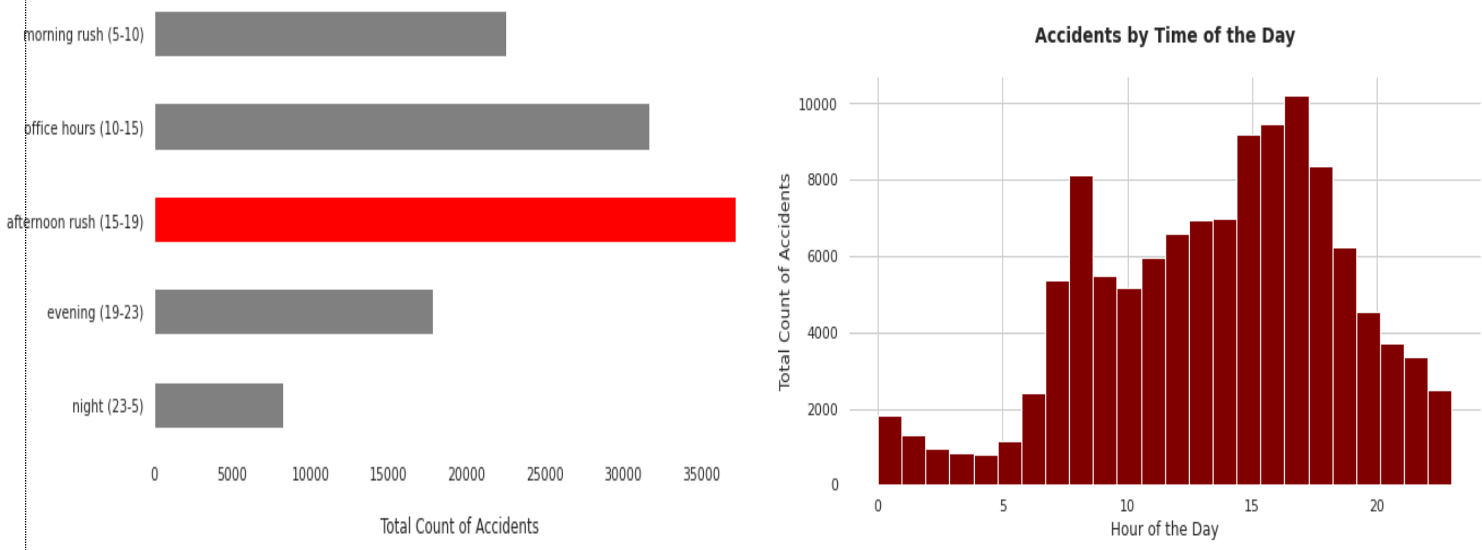


Figure 3: Accidents by hour.

Figure 3 displays the data visualization of the number of accidents according to the time of the day, afternoon rush period between the hour of 15:00pm and 19:00pm is the period with the most accidents and 17:00pm happens to be the peak period.

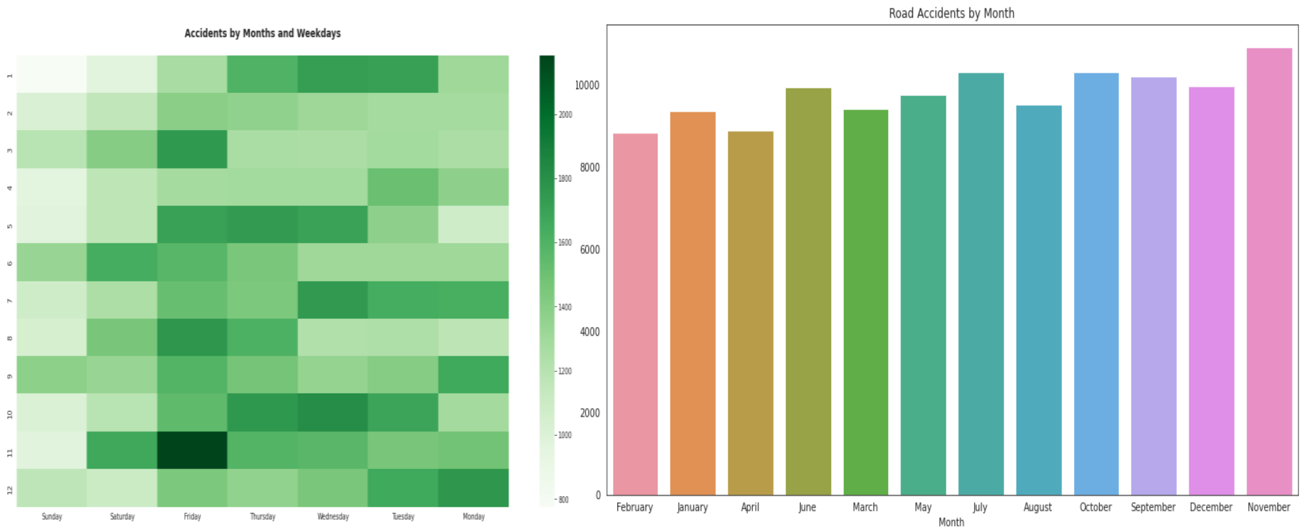


Figure 4: Road accidents by weekday and months

Judging by the graphical representation in figure 4, November appears to be the month with the highest accidents in the year 2019

Hours **of the day, and days of the week, on which accidents occur for motorbikes.**

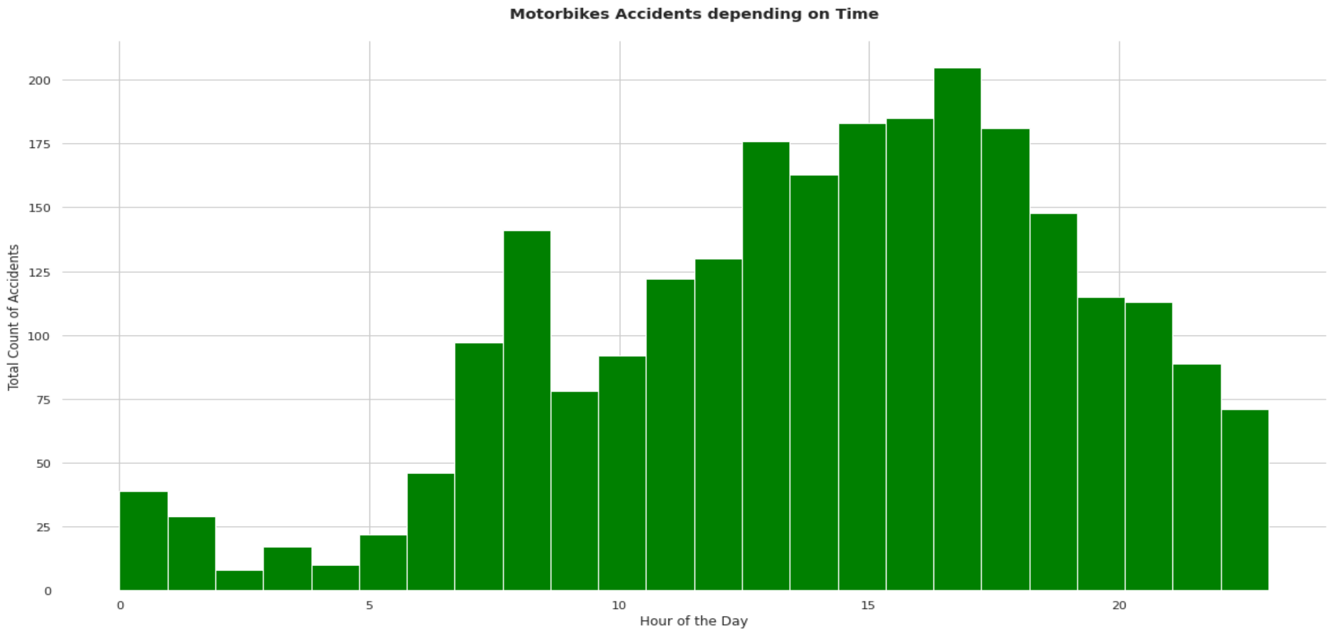


Figure 5:Motorbikes accidents according to time

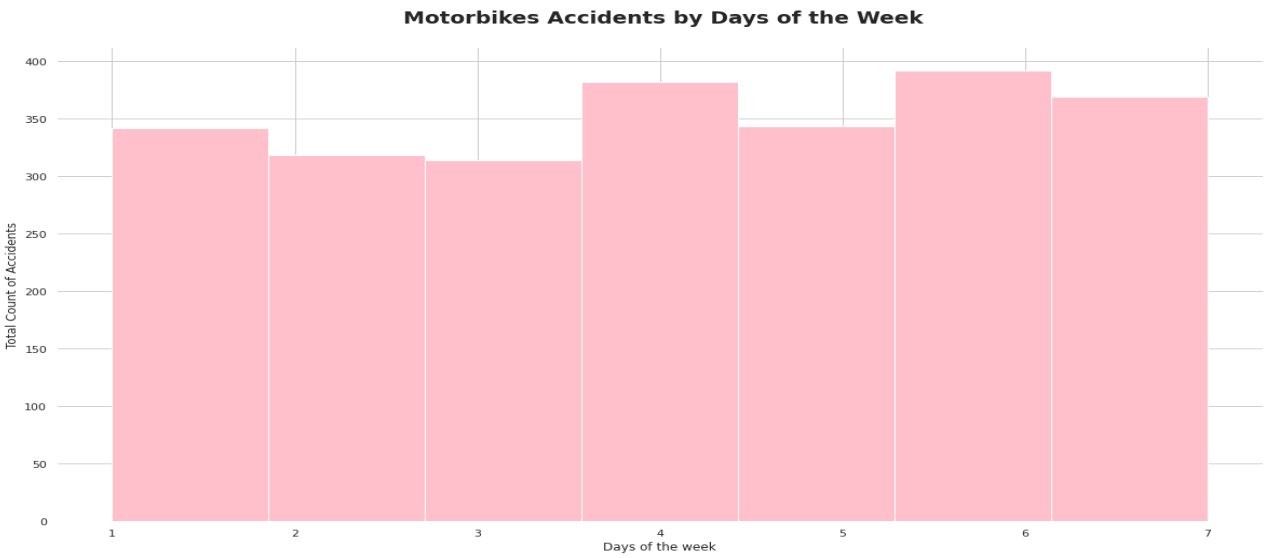


Figure 6: Motorbikes accidents according to days of the week

In figure 5, motorbike accidents occur mostly in the evenings during rush hour and on Fridays (figure 6).

Hours of the day, and days of the week, on which pedestrians are more likely to be involved in accidents.

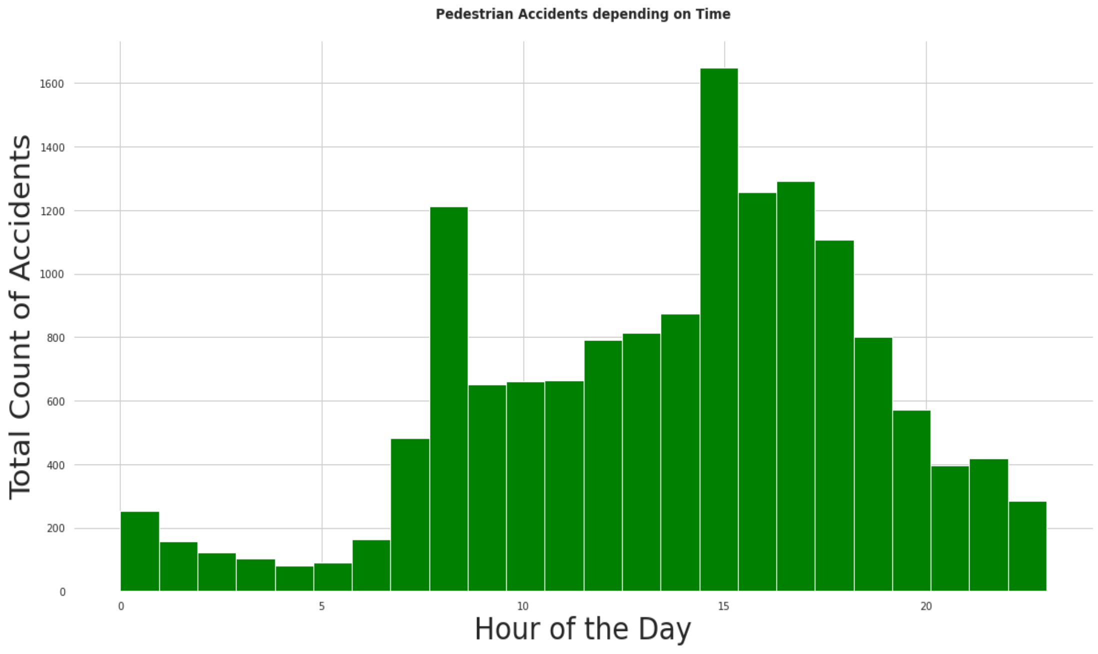


Figure 7: Pedestrian Accidents by Hour of the day

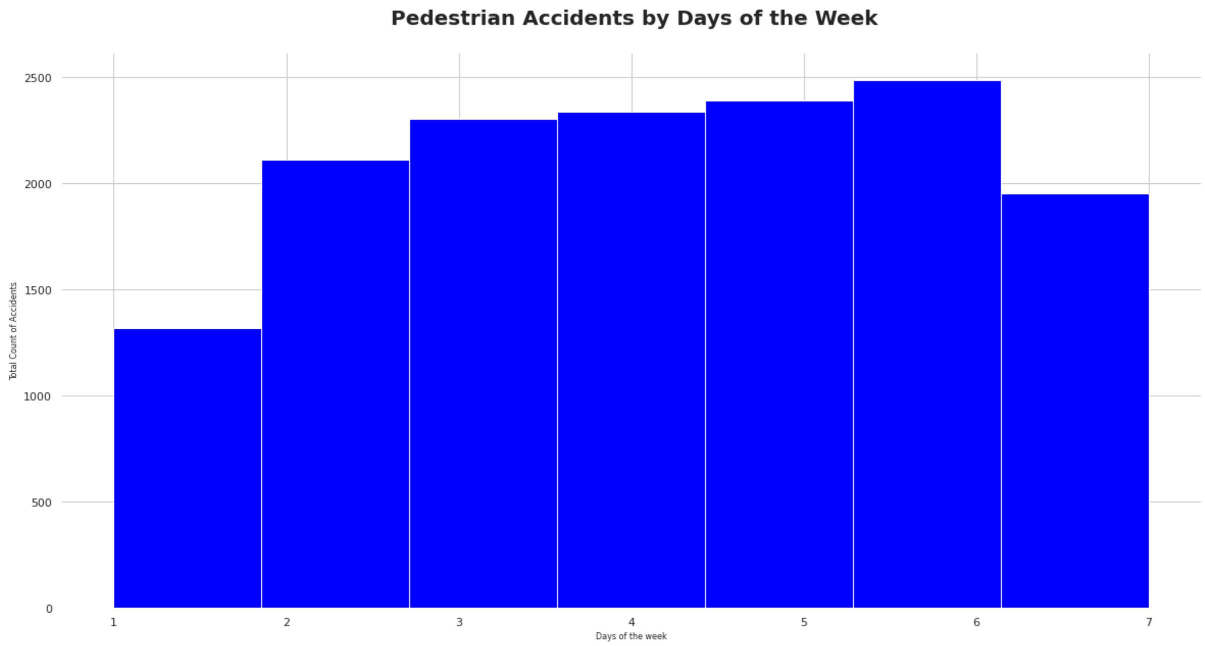


Figure 8: Pedestrian Accidents according to the days of the week

Figure 7, shows that the majority of pedestrian fatalities occurred around 15:00 p.m., while figure 8 shows that the majority of fatalities occurred on Fridays during the week. As a result, pedestrians in the UK are more likely to be involved in accidents during these times of the week in 2019.

Impact of daylight savings on road traffic accidents in the week/month after it starts and stops.

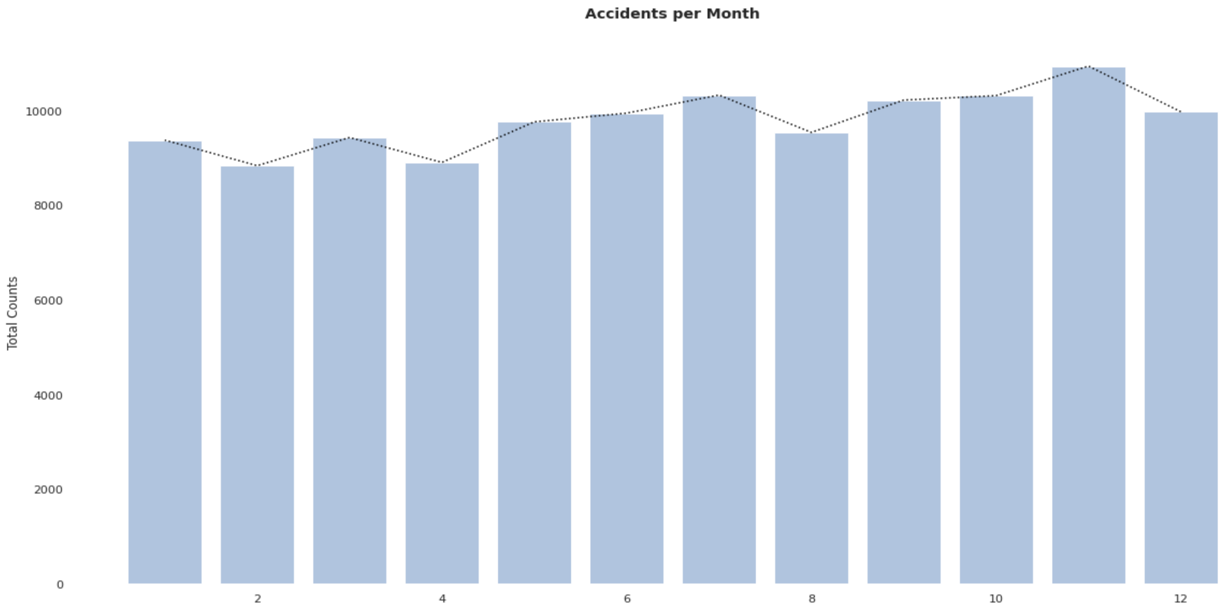


Figure 9: Effect of Daylight saving

In the year 2019, daylight saving time began on the 31st of March in the United Kingdom and finished on the 27th of October (TIME AND DATE, 2022). As seen in figure 9, there was a reduction in traffic accidents a few weeks after daylight saving time began in the month of April. However, there was an increase in road accidents in the weeks after the termination of daylight saving time..

Impact of sunrise and sunset times on road traffic accidents

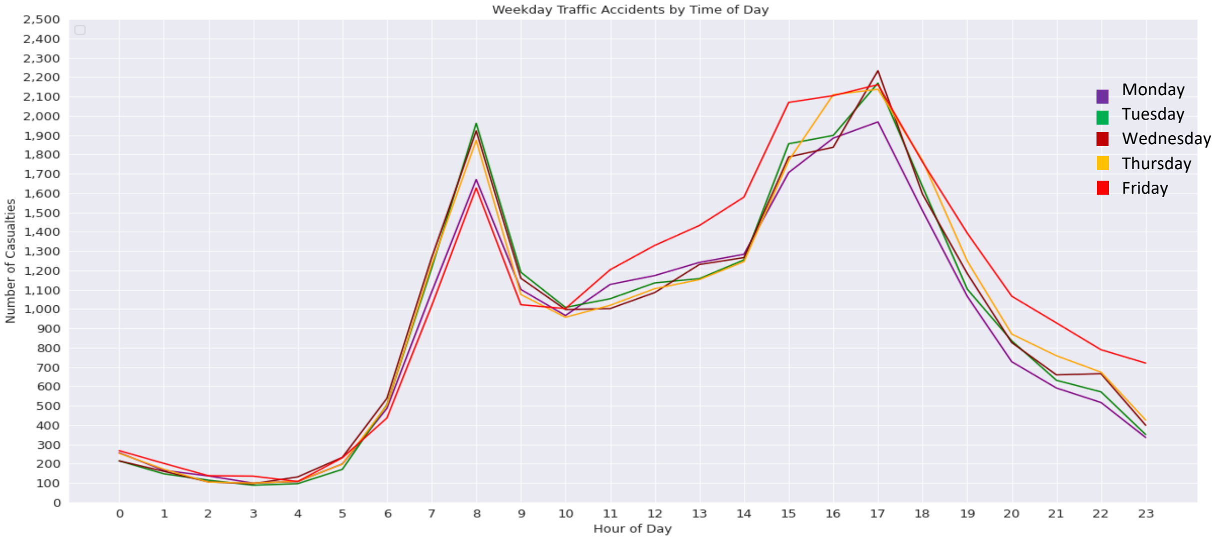


Figure 10: Weekday traffic accidents according time of day

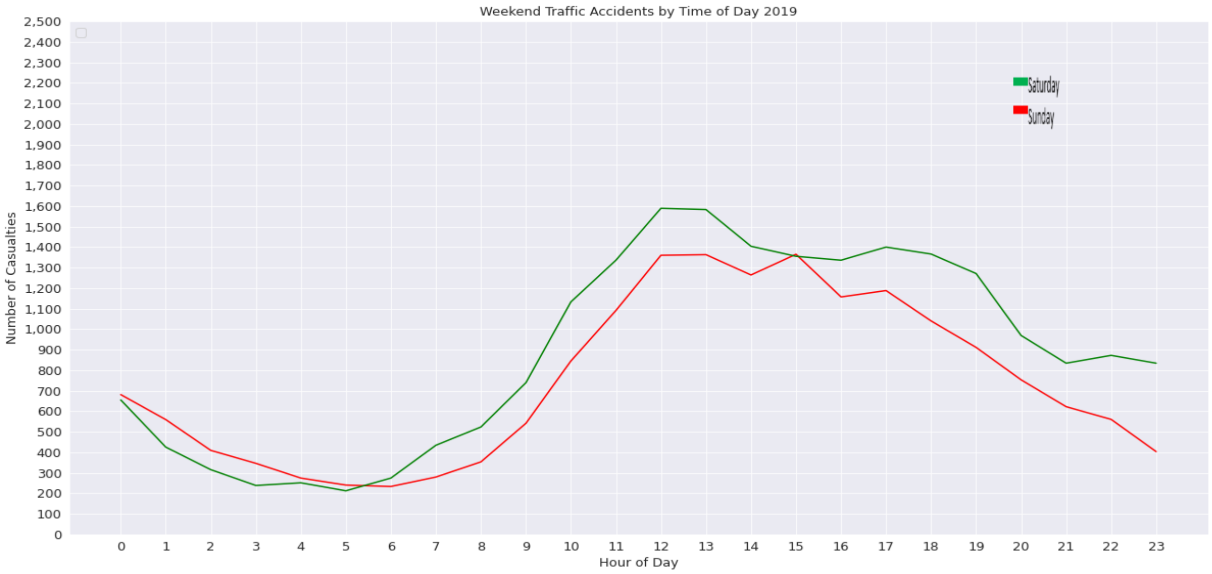


Figure 11: Weekend Traffic Accidents by Time of day

Figure 10 shows that traffic accidents were relatively low in the UK in 2019 during the late-night hours of 0:00 to 5:00am, steadily increased after 6am, and then declined again about 9am. However, it began to rise again about 14:00, peaked at 17:00, and then began to fall until around 23:00.

Weekend traffic accidents, like weekdays, were low in the late night hours but increased from 9 a.m., with a progressive increase in traffic up until 13 p.m., as seen in figure 11. Despite the fact that the decline was not dramatic, traffic began to decline until 23:00pm.

Vehicles (engine capacity, age of vehicle, etc.) that are more frequently involved in road traffic accidents.

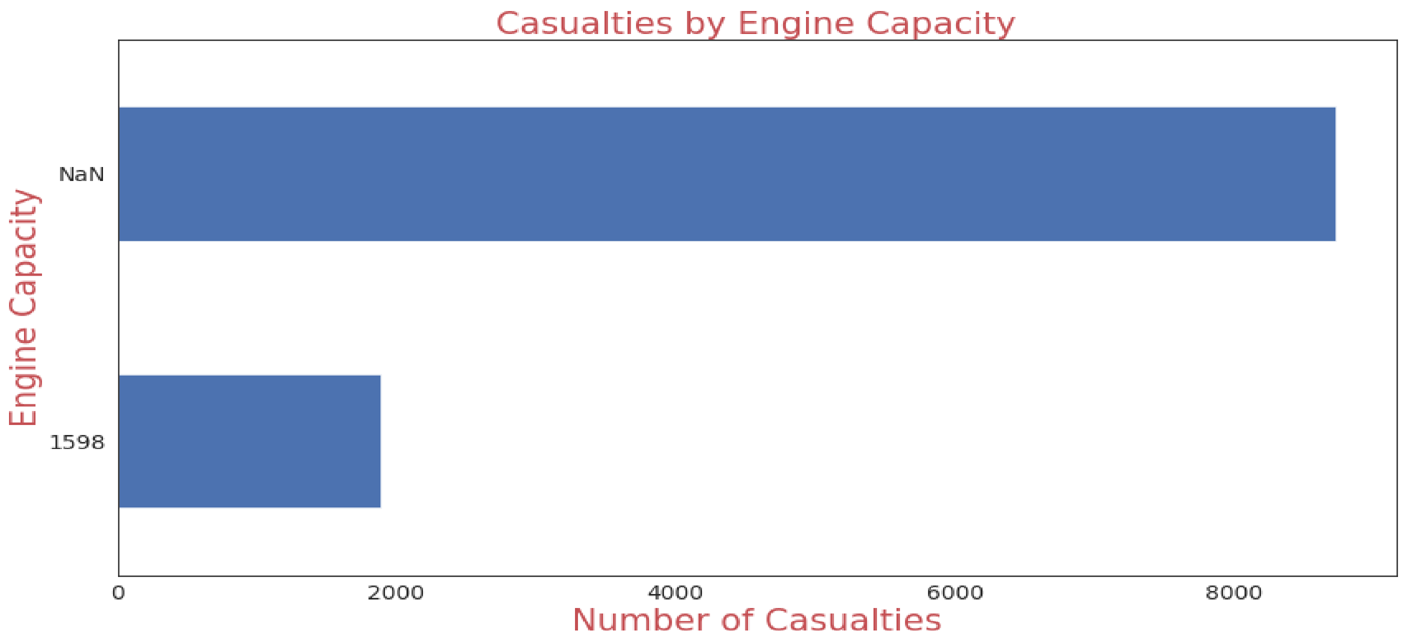


Figure 12. Engine Capacity count vs Number of Casualties

Because there are many NaN values in the dataset presented, there is no considerable information on what engine capacity has been involved more frequently in traffic incidents, as seen in figure 12.

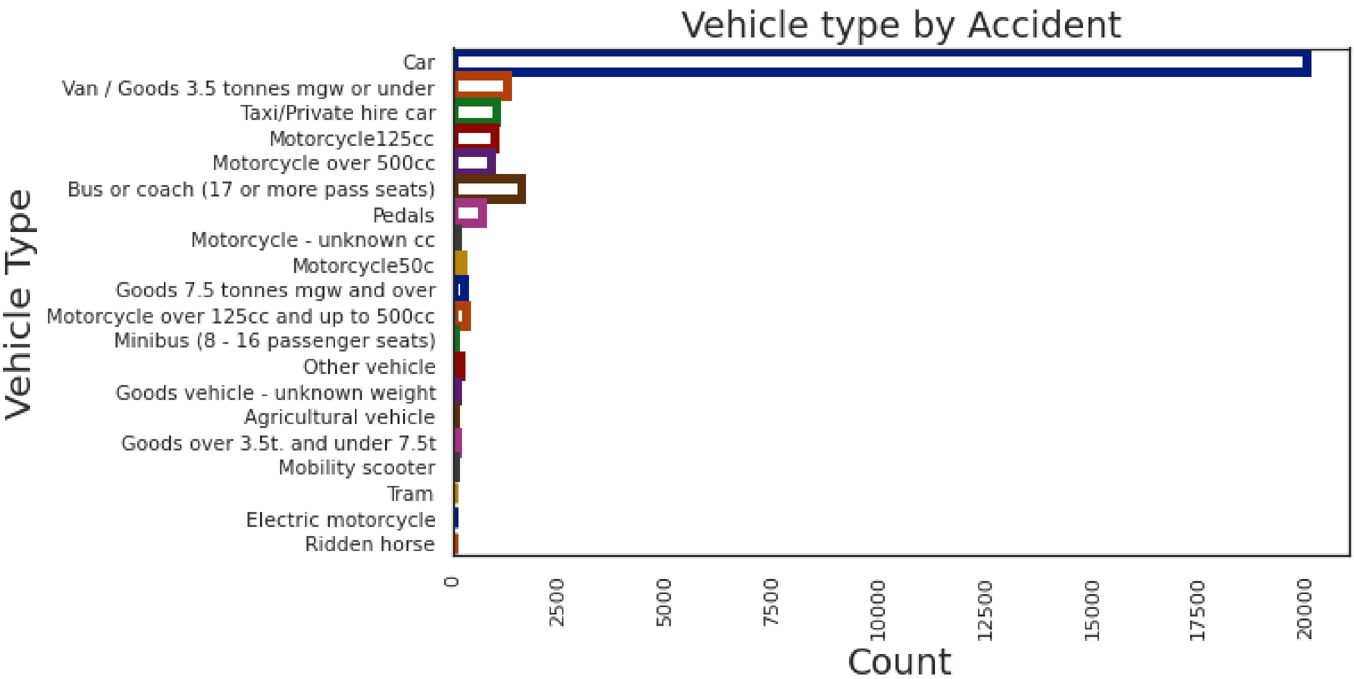


Figure 13: Traffic accidents according to Vehicle type

Cars appear to be the most engaged in traffic accidents by a large margin, followed by buses or coaches with 17 or more passenger seats, as seen in figure 13.

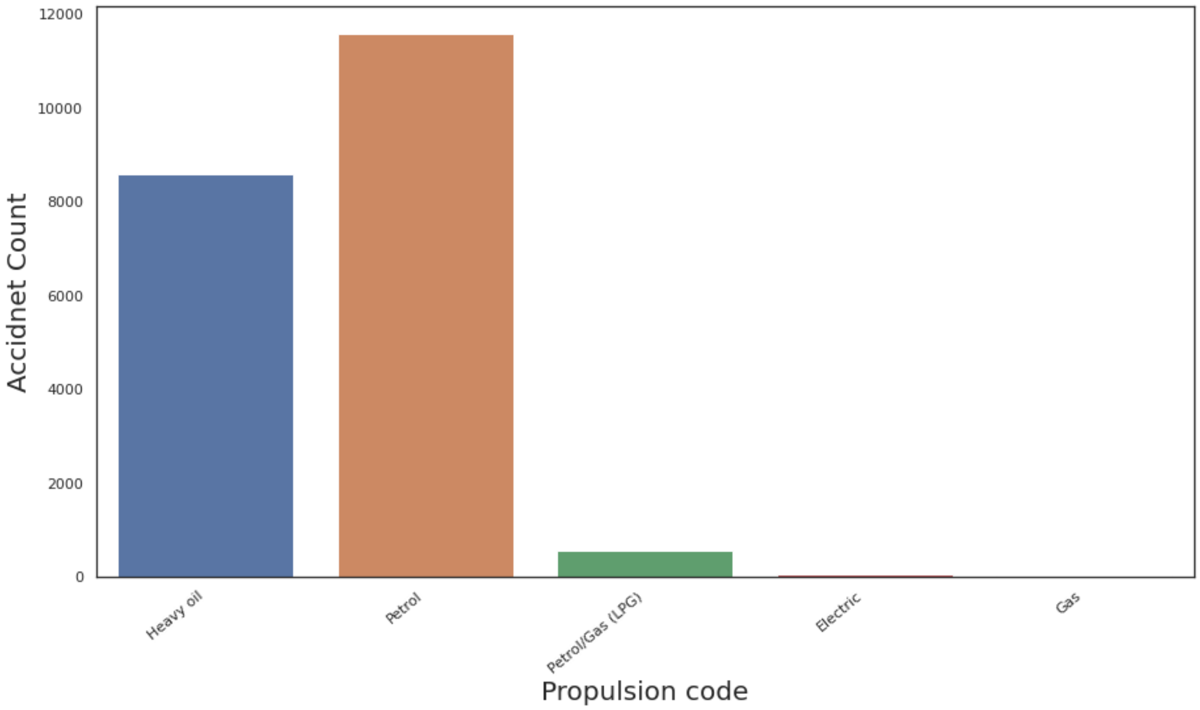


Figure 14: Propulsion of Vehicle

According to i figure 14, vehicles with petrol propulsion appear to be the most involved in accidents in 2019, while electric and gas vehicle propulsion appear to be the least involved in accidents.

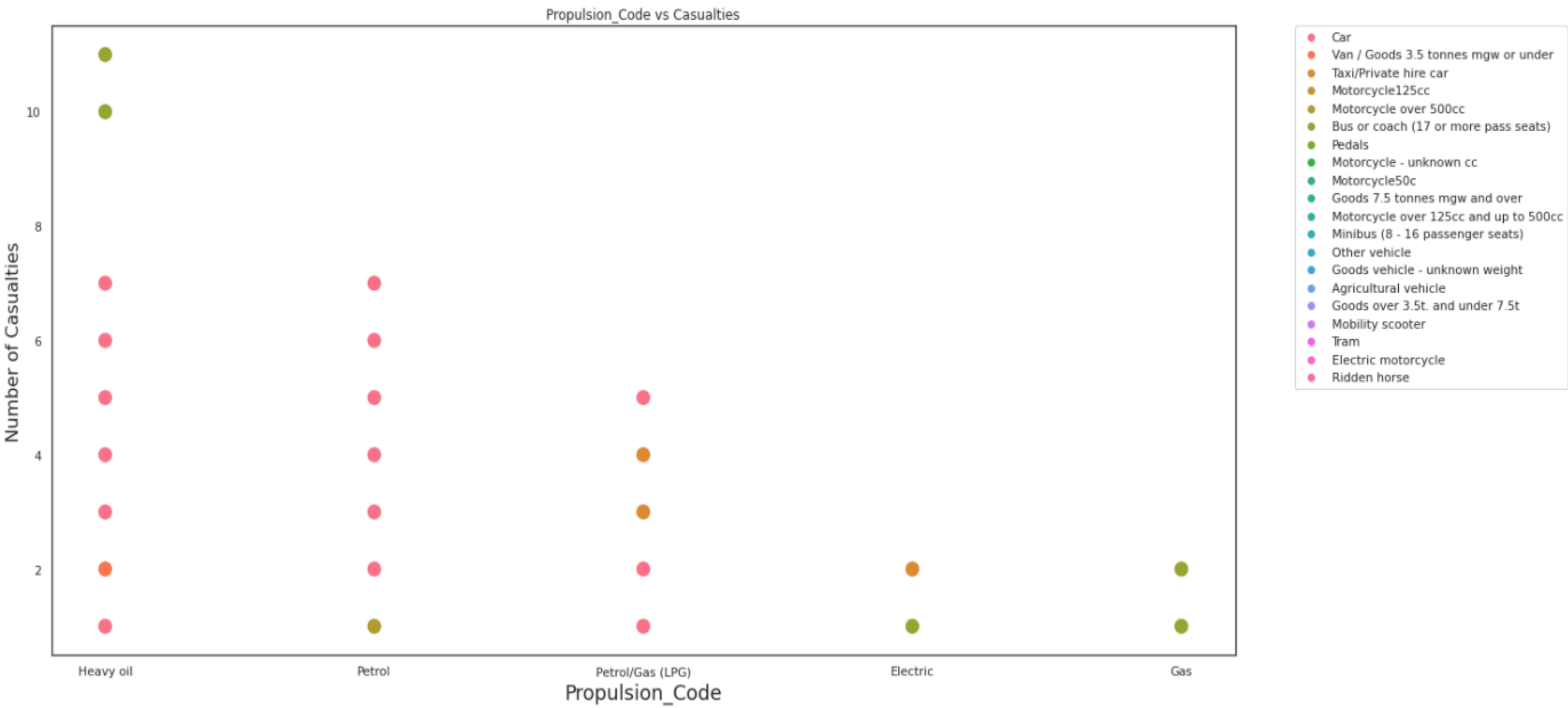


Figure 15: Propulsion Code Vs Casualties and Vehicle type

Based on the information in figure 15, vehicles having heavy oil seem to be more involved in accidents with the most casualties in 2019.

Conditions (weather, geographic location, situations) that generate more road traffic accidents.

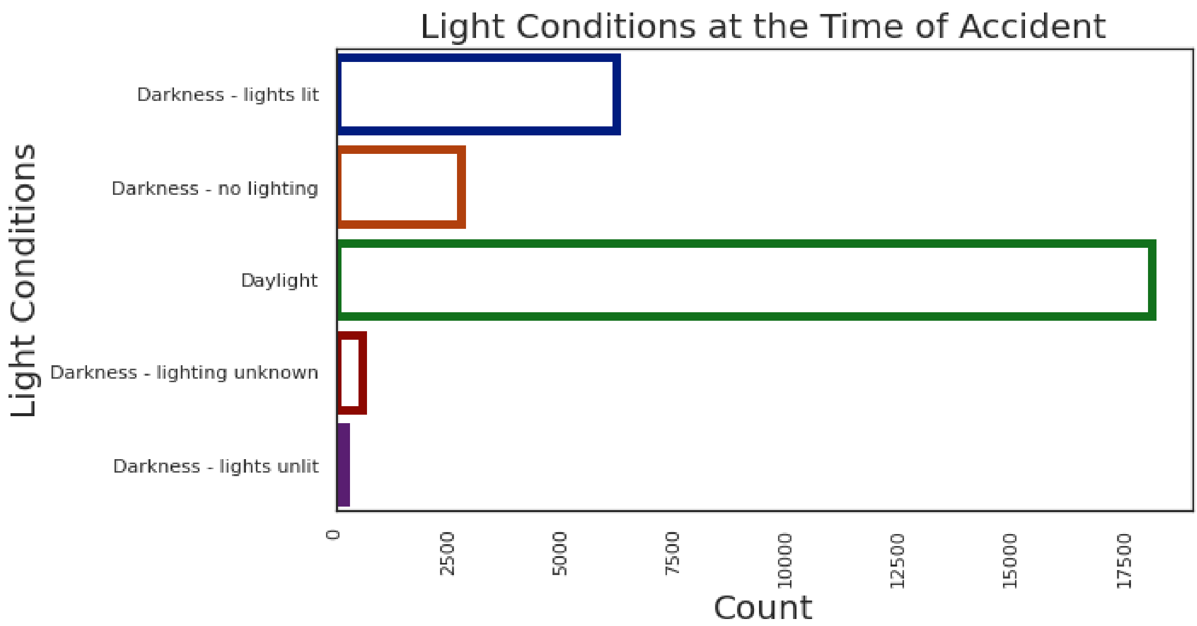


Figure 16: Light Conditions at time of Traffic Accidents

Figure 16 shows that accidents occur more frequently in daylight, with approximately 17500 counts; the second highest light condition is Darkness – lights lits, followed by Darkness – no lighting..

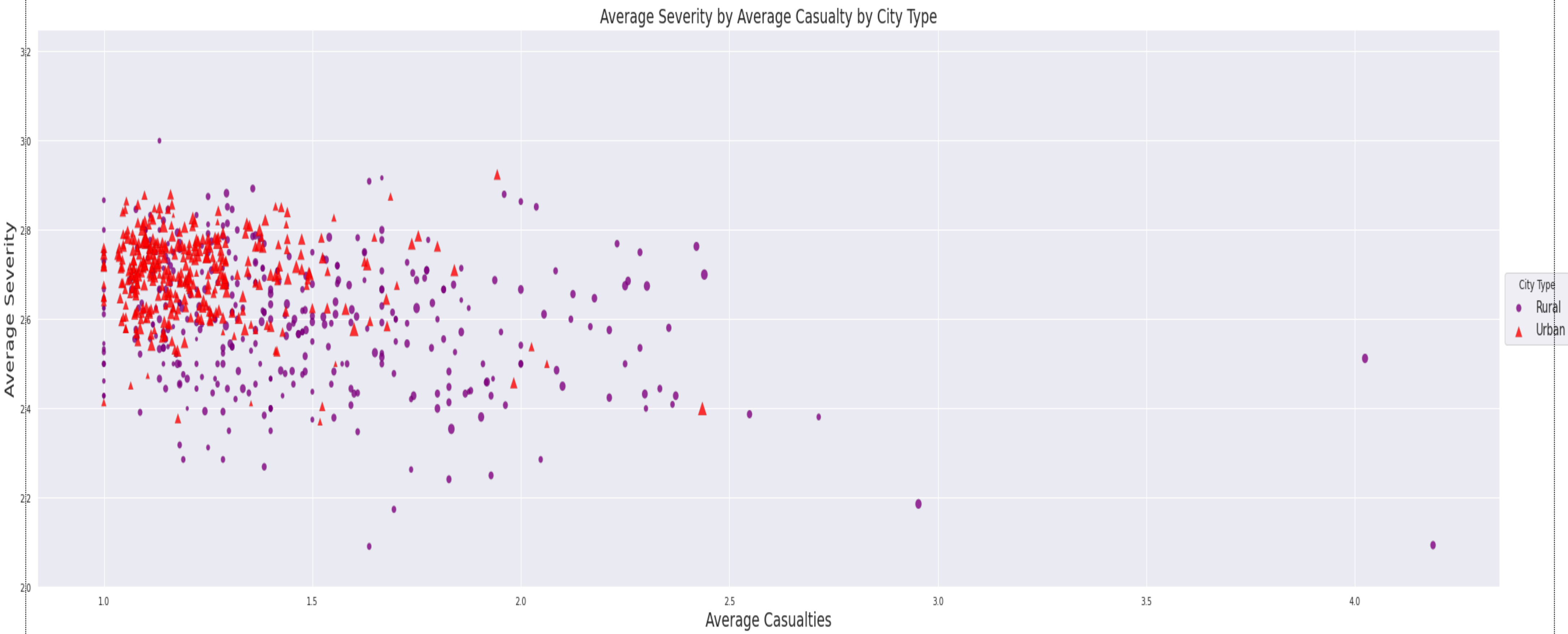


Figure 17:Average Severity by Average Casualty by City Type

In 2019, the severity of traffic accidents in the UK's metropolitan areas is between 2.5 and 3, however in the rural areas, there appear to be more casualties as a result of traffic accidents.

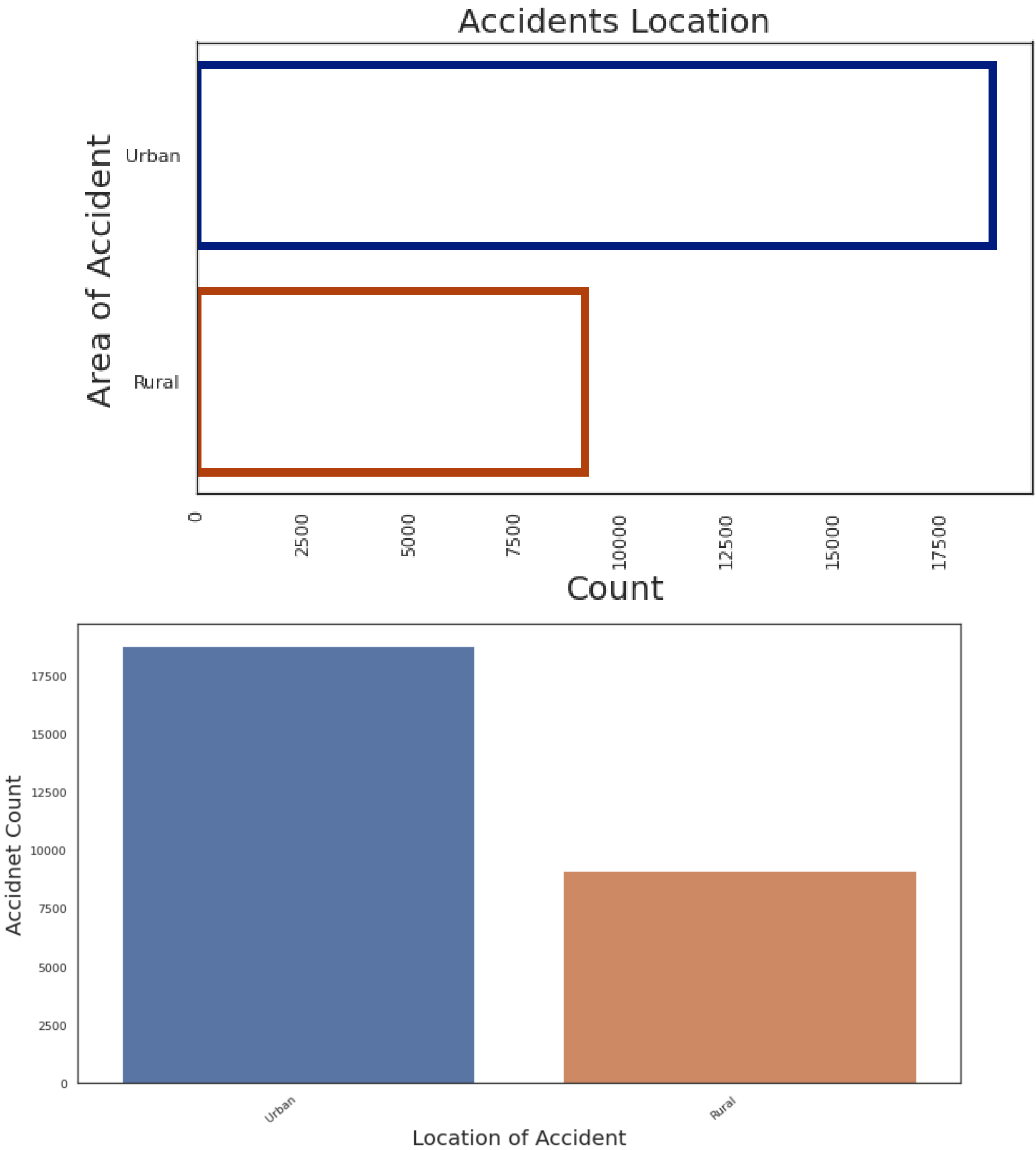


Figure 18: Location of Traffic Accidents

In the year 2019, the graphical depiction in figure 18 demonstrates that traffic accidents in the urban region double the size of traffic accidents in the rural area in the United Kingdom.

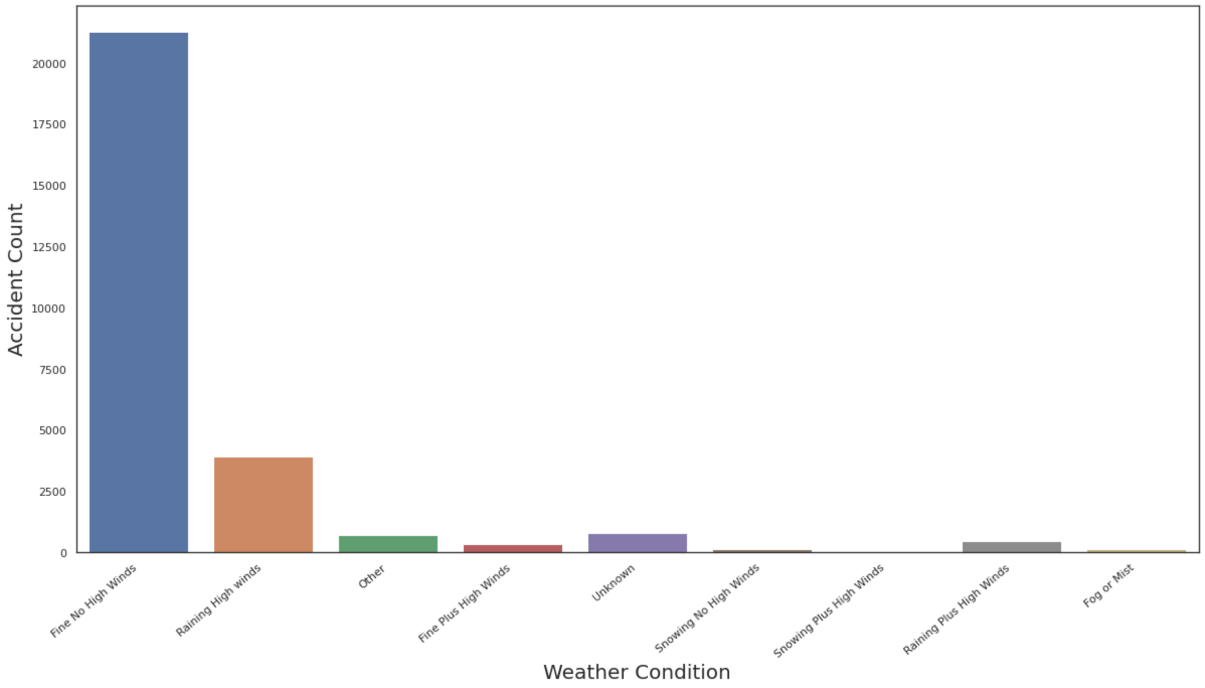


Figure 19: Weather Condition

Fine weather condition with no high winds was when the most traffic accidents occurred in the UK in 2019 followed by when it was raining with high winds with about 2,750 traffic accidents as seen in figure 19.

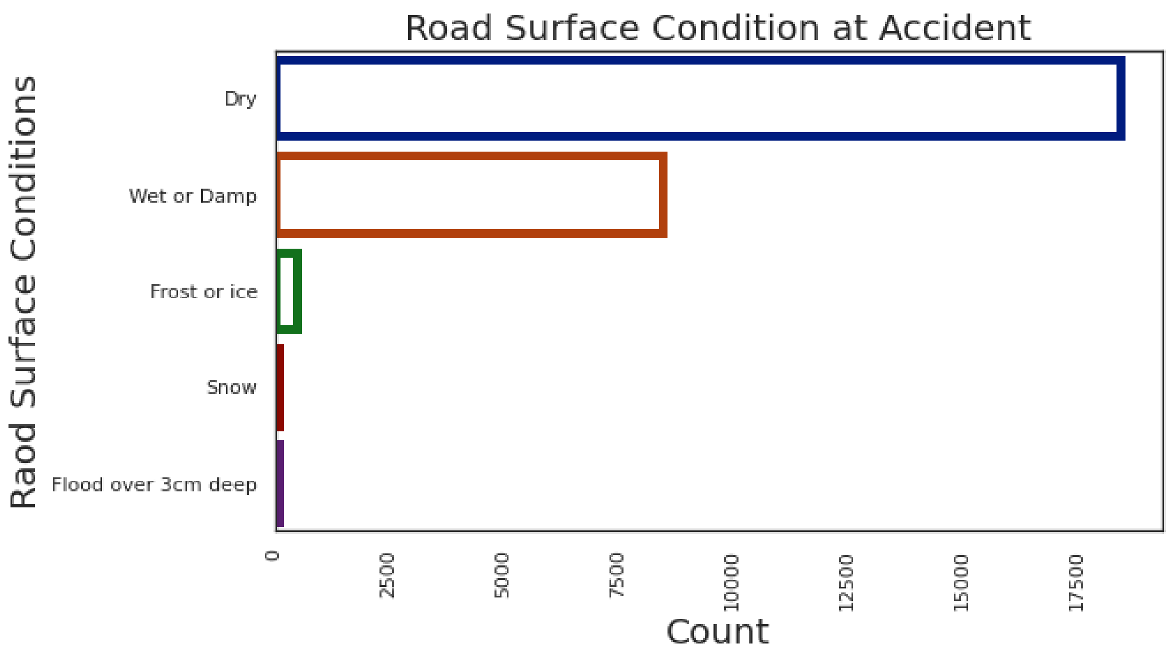


Figure 20: Road Surface conditions

The majority of traffic incidents in the UK in 2019 occurred on dry road surfaces, with wet or damp road surfaces coming in second. The flooded road surface condition, which is over 3cm deep, is the lowest, followed by the snow road surface condition, as shown in figure 20. In the year 2019, it snowed a few times in the United Kingdom (Office, 2019).

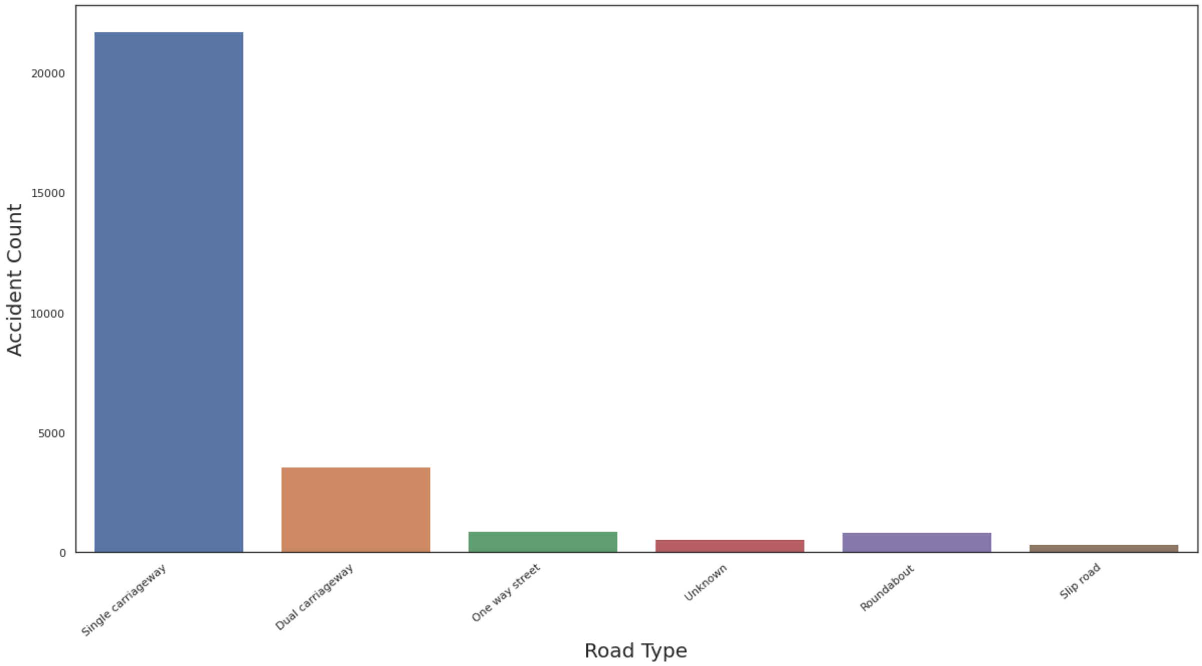


Figure 21: Road Type

Majority of the traffic accidents occurred in single carriageway road type with a huge margin compared to rest of the road types as seen in figure 21.

Driver related variables that affect the outcome

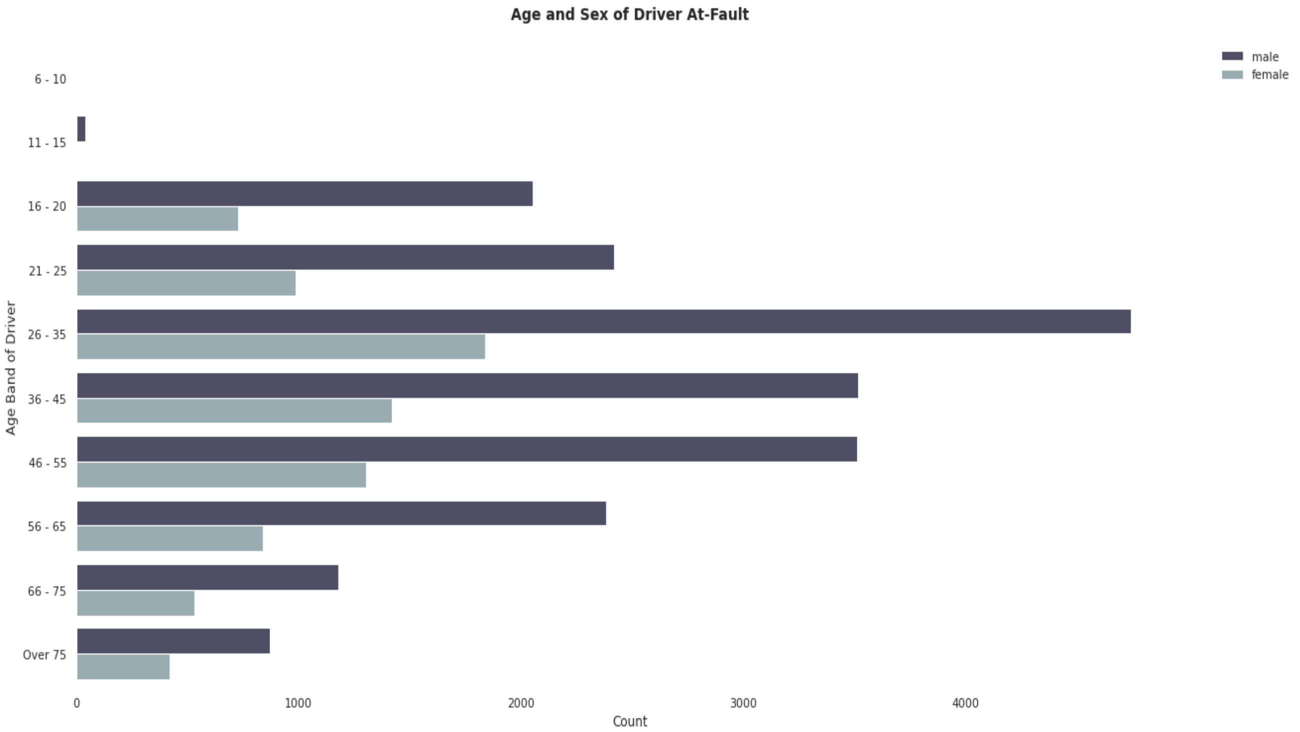


Figure 22: Age and Gender of Driver involved in traffic accidents

Male drivers, regardless of age, are more likely to be involved in traffic accidents than female drivers in the UK in 2019. Also, the age group of drivers engaged in the most traffic accidents were between the ages of 26 and 35, followed by the age groups of 36 and 45, 46 and 55, and finally 46 and 55. It's odd to see some male drivers between the ages of 11 and 15 get into accidents despite the fact that it's against the law to drive at that age. (Gov.UK, 2022)

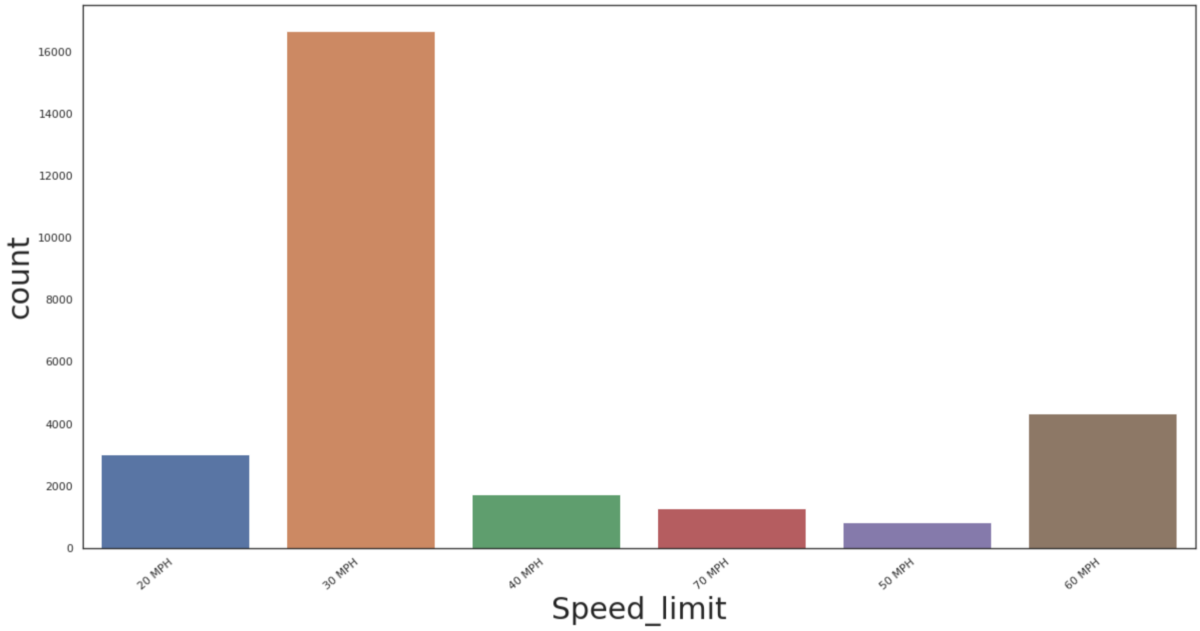


Figure 23: Speed Limit

In figure 23, vehicles moving with the speed limit of 30MPH got involved in the most traffic accidents while 50MPH had the least.

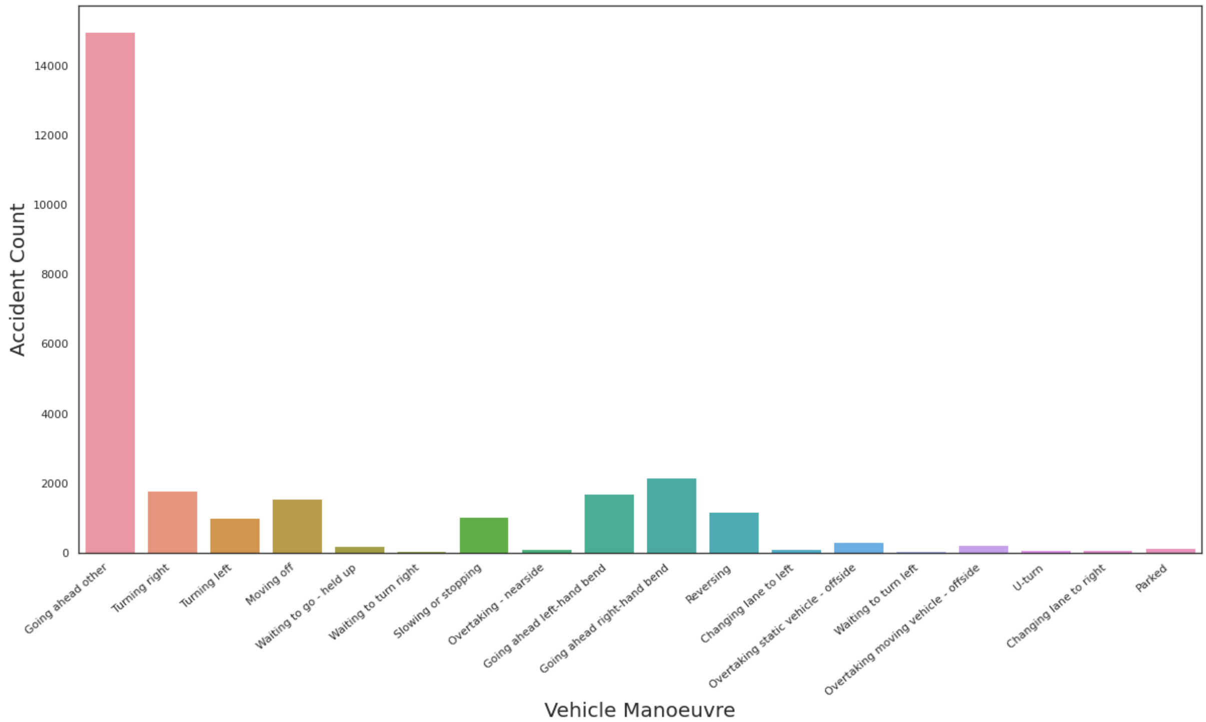


Figure 24: Vehicle Manoeuvre

It can be seen that vehicles going ahead of another got involved in the most accidents as compared to other manoeuvring conditions as seen in figure 24.

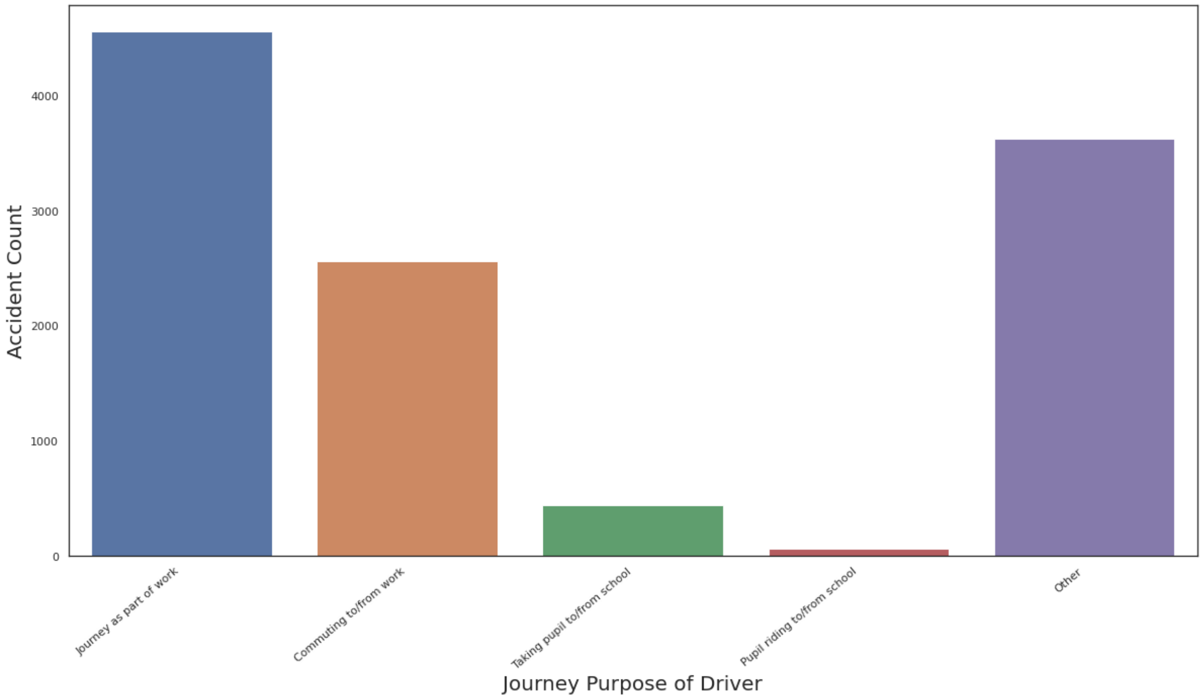


Figure 25: Journey Purpose of Driver

Work destination drivers appears to have the most traffic accidents followed closely by other unknown driver destinations as seen on figure 25.

MODELLING

Step 1. Pre-process data, deleted missing values and converted categorical variables to dummy variables, ended up with this Dataset: (463192, 45) for location prediction, (121801, 11)for injury severity and (463217, 14) for time prediction.

Step 2: Divide the dataset into a training and a test set in a 75:25 ratio. In this case, 75% of the records are used for model training and 25% are used for validation.

Step 3: Determined the highest performing KNN, Decision Tree, Logistic Regression, Random Forest, and Multinomial NB models upon stacking them and using RepeatedStratifiedFold of n-splits of 10 and 3 n-repeats.

Step 4. Using KFold cross-validation, I chose Random-forest to model further since it works well with imbalanced datasets and reduces overfitting (Ahmed, Hossain, Bhuiyan, & Ray, 2021). 100 n-estimators, random state, and n-jobs are among the hyperparameters. Synthetic minority oversampling technique (SMOTE) was also utilised. "SMOTE" is an oversampling technique in which synthetic samples are generated for the minority class. The SMOTE algorithm works by removing the problem of overfitting caused by oversampling (Satpathy, 2020).

INDEPENDENT FEATURES AND THEIR TARGET VARIABLES

1. Used these selected independent features for the target variable of **Casualty Severity** for modelling:

```
['Day_time',
 'Road_Type',
 'Speed_limit',
 'Urban_or_Rural_Area',
 'Number_of_Casualties',
 'Casualty_Type',
 'Did_Police_Officer_Attend_Scene_of_Accident',
 'Weather_Conditions',
 'Road_Surface_Conditions',
 'Light_Conditions']
```

2. Used these selected independent features for the target variable of **Rural_or_Urban_Area** for modelling:

```
['Day_time',
 'Pedestrian_Location',
 'Pedestrian_Movement',
 'Speed_limit',
 'Light_Conditions',
 'Did_Police_Officer_Attend_Scene_of_Accident',
 'Number_of_Casualties',
 'Weather_Conditions',
 'Road_Surface_Conditions',
 'Road_Type']
```

3. Used these selected independents features for the target variable of **Day_time** modelling:

```
['Urban_or_Rural_Area',
 'Age_Band_of_Driver',
 'Special_Conditions_at_Site',
 'Did_Police_Officer_Attend_Scene_of_Accident',
 'Longitude',
 'Latitude',
 'Pedestrian_Movement',
 'Speed_limit',
 'Weather_Conditions',
 'Number_of_Casualties',
 'Road_Surface_Conditions',
 'Road_Type',
 'Light_Conditions']
```

PERFORMANCE METRICS FORMULAR

Precision is the number of correct predictions the classifier was able to identify, recall refers to the of positives instances the classifier was able to pick and F1-score is the percentage of positives predictions that were correct. (Kohli, 2019)

$$Accuracy = \frac{Number\ of\ correct\ prediction}{Number\ of\ observation} \times 100 \tag{7}$$

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \tag{8}$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \tag{9}$$

$$F1 - Score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{10}$$

Figure 26: Performance Metrics Calculation formular

RESULTS

	RESULTS UPON STACKING THE DIFFERENT MODELS FOR THE PREDICTION EXPERIMENTS				
SITUATIONS TO PREDICT	DECISION TREE SCORES	KNN SCORES	LOGISTIC REGRESSION SCORES	RANDOM FOREST SCORES	MULTINOMIAL-NB
Casualty Severity	78%	78%	81%	80%	81%
Time of Traffic Accidents	84%	72%	40%	83%	x
Urban or Rural Area	88%	87%	86%	88%	82%

STACKING RESULTS

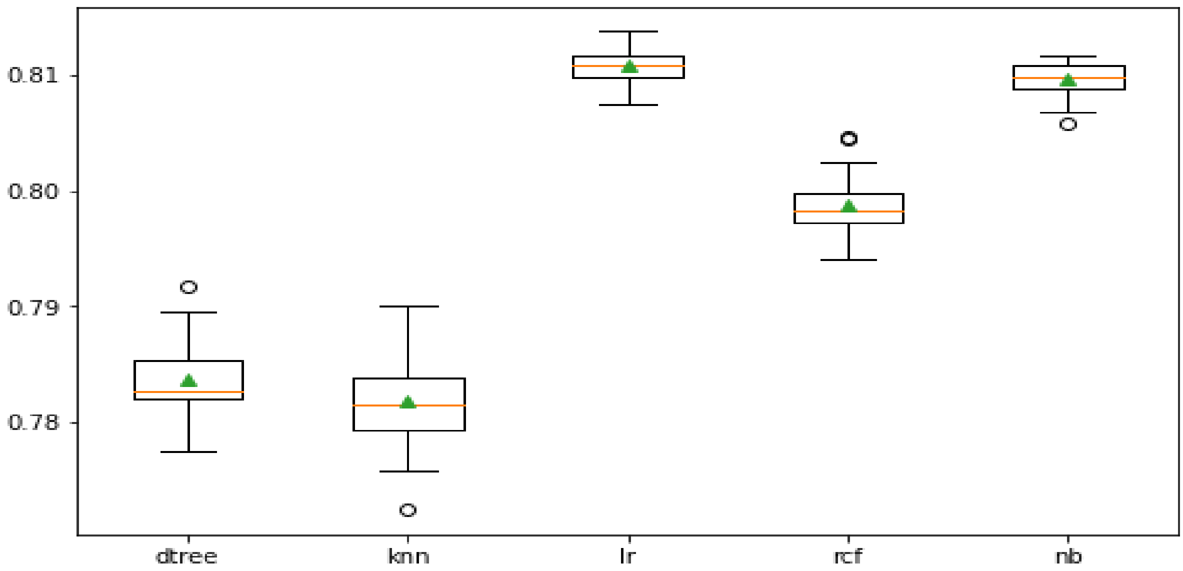


Figure 27: Boxplot of modelling results of Casualty Severity

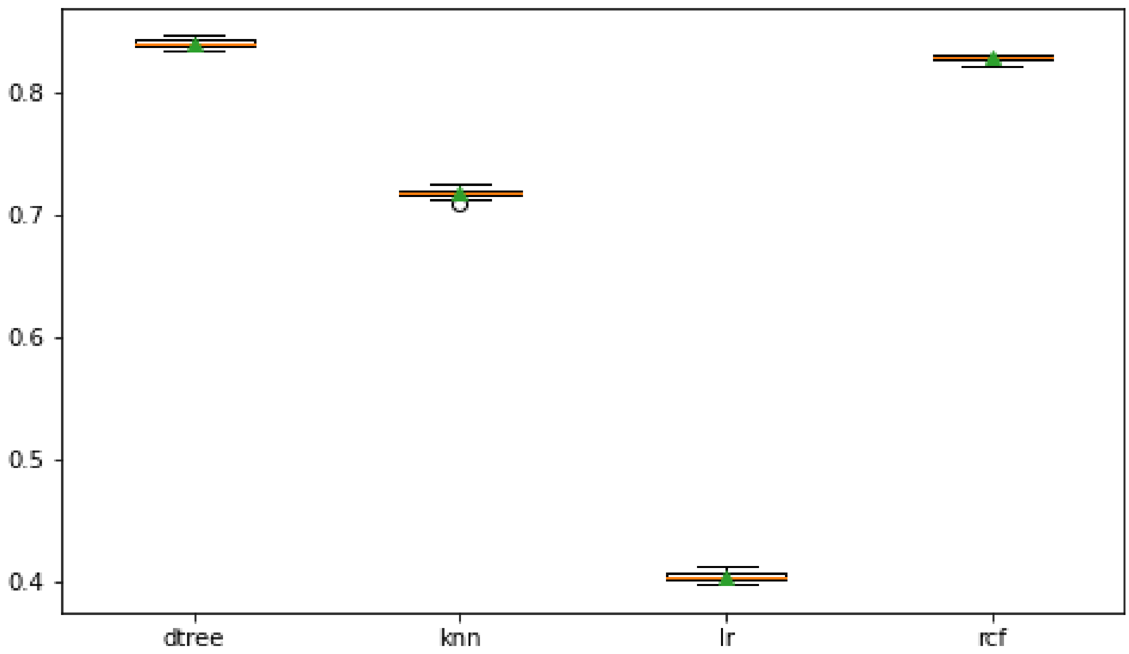


Figure 28: Boxplot of modelling results of time prediction

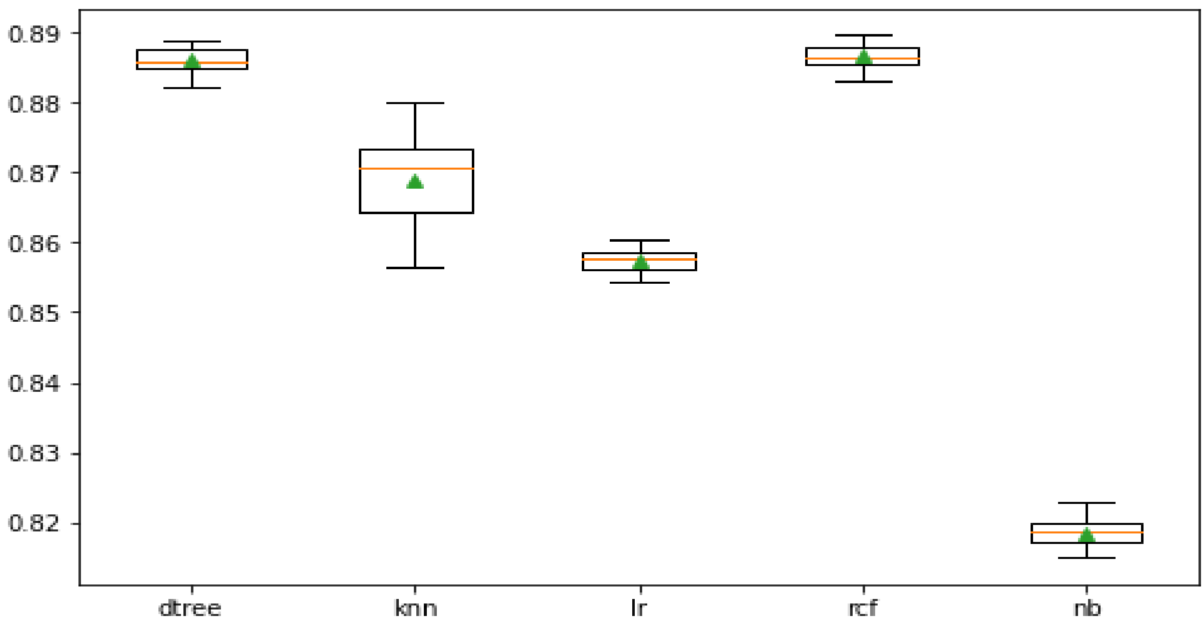


Figure 29: Boxplot of modelling result of location prediction

RF PREDICTION FOR CASUALTY SEVERITY

Classification Report Random Forest :

	precision	recall	f1-score	support
1	0.16	0.03	0.05	279
2	0.39	0.12	0.18	4076
3	0.82	0.96	0.89	18483
accuracy			0.80	22838
macro avg	0.46	0.37	0.37	22838
weighted avg	0.74	0.80	0.75	22838

Classification Report Random Forest - with Entropy and SMOTE Upsampling:

	precision	recall	f1-score	support
1	0.71	0.86	0.78	18678
2	0.62	0.57	0.59	18309
3	0.71	0.60	0.65	18408
accuracy			0.68	55395
macro avg	0.68	0.68	0.67	55395
weighted avg	0.68	0.68	0.67	55395

The accuracy results reduced from 80% to 68% after performing the modelling with SMOTE, however one can see a better improvement in the precision, recall and f1 score metrics of the five target variables after SMOTE was applied.

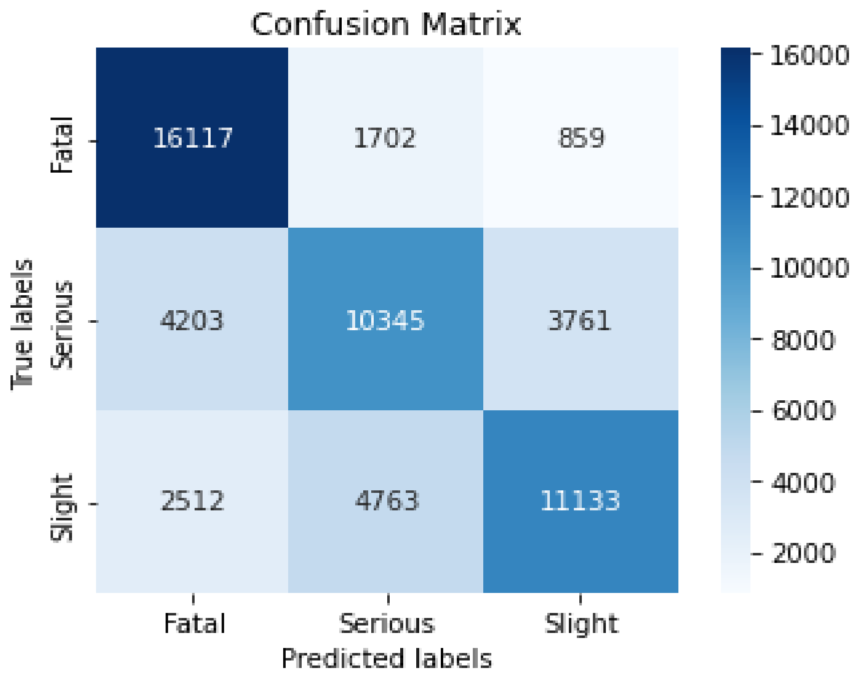


Figure 30: Confusion matrix for Casualty Severity Prediction upon upsampling the model.

Upon further modelling with random forest using smote, the model produced the results in fig 29 and 30. The classifier did not do well in classifying the training dataset into their rightful target variable, however it did better upon classifying by upsampling the dataset. The confusion matrix presents a better representation of the classification, about 16117 of our training set of 18,678 of the fatal accidents have been rightly predicted while, 10345 of the Serious accidents were rightly predicted and 11133 of the slight traffic accidents were rightly predicted.

PROBABILITY PREDICTIONS FOR CASUALTY SEVERITY

	Accident_Index	Fatal	Serious	Slight
0	2019010155191	0.010000	0.210738	0.779262
5	2019010155195	0.000000	0.076667	0.923333
12	2019010155198	0.000000	0.052843	0.947157
14	2019010155207	0.020292	0.904842	0.074866
17	2019010155216	0.513893	0.323277	0.162829
...	...	...	...	...
121785	201963DF01819	1.000000	0.000000	0.000000
121786	201963DF02019	0.406173	0.289812	0.304015
121793	201963DF02919	0.654249	0.235758	0.109993
121796	201963DF03019	0.927449	0.060384	0.012167
121798	201963DF03419	0.623032	0.318400	0.058568

30515 rows × 4 columns

Above is the probability prediction for Casualty severity merged with their Accident Index in the dataframe.

RF PREDICTION FOR TIME OF TRAFFIC ACCIDENT

	precision	recall	f1-score	support
1	0.82	0.78	0.80	15535
2	0.86	0.88	0.87	27535
3	0.84	0.85	0.84	29082
4	0.86	0.87	0.86	13638
5	0.90	0.84	0.87	6854
accuracy			0.85	92644
macro avg	0.85	0.84	0.85	92644
weighted avg	0.85	0.85	0.85	92644

Classification Report Random Forest - and SMOTE Upsampling:

	precision	recall	f1-score	support
1	0.85	0.88	0.86	36277
2	0.87	0.87	0.87	36092
3	0.87	0.83	0.85	36323
4	0.91	0.91	0.91	36203
5	0.94	0.95	0.95	36365
accuracy			0.89	181260
macro avg	0.89	0.89	0.89	181260
weighted avg	0.89	0.89	0.89	181260

The accuracy of the model improved by 4% upon upsampling with SMOTE and one can see a better distribution of the precsion, the recall and f1-score of the five target variables.

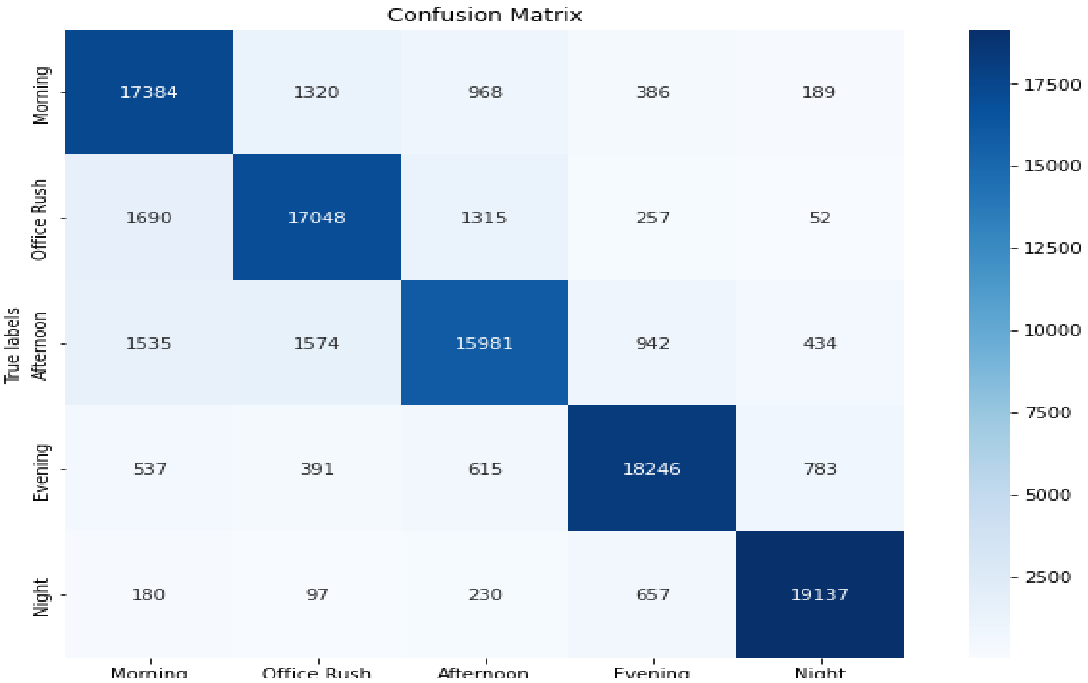


Figure 31: Confusion Matrix of Time of Traffic accident

After using SMOTE to upsample the training set, the modelling accuracy increased to 86% and from our confusion matrix in fig 31 , the randon forest classifer was able to identtify a large portion of the time of traffic accidents according to the time they occurred based on the data visualization of our confusion matrix.

PROBABILITY PREDICTIONS FOR TIME

	Accident_Index	Morning	Office Rush	Afternoon	Evening	Night
0	2019010155191	0.06	0.00	0.02	0.31	0.61
5	2019010155195	0.00	0.00	0.00	0.00	1.00
7	2019010155196	0.00	0.00	0.00	0.00	1.00
13	2019010155207	0.00	0.00	0.00	0.00	1.00
19	2019010155220	0.00	0.00	0.00	0.00	1.00
...	...	...	...	...	...	...
169609	201963DF02519	0.06	0.94	0.00	0.00	0.00
169611	201963DF02719	0.06	0.94	0.00	0.00	0.00
169613	201963DF02719	0.01	0.99	0.00	0.00	0.00
169617	201963DF02919	0.42	0.37	0.19	0.02	0.00
169622	201963DF03019	0.01	0.03	0.96	0.00	0.00

42463 rows × 6 columns

Here we have the probability predictions for time of traffic accidents prediction merged with the Accidents index in the data-frame.

RANDOM FOREST PREDICITION FOR LOCATION OF ACCIDENT

	precision	recall	f1-score	support
1	0.87	0.93	0.90	60729
2	0.91	0.84	0.88	55069
accuracy			0.89	115798
macro avg	0.89	0.88	0.89	115798
weighted avg	0.89	0.89	0.89	115798

Classification Report Random Forest - and SMOTE Upsampling:

	precision	recall	f1-score	support
1	0.86	0.92	0.89	60683
2	0.92	0.85	0.88	60911
accuracy			0.89	121594
macro avg	0.89	0.89	0.89	121594
weighted avg	0.89	0.89	0.89	121594

The difference in the performance metrics is not much here as it can be observed that only slight changes was made after upsampling the model.

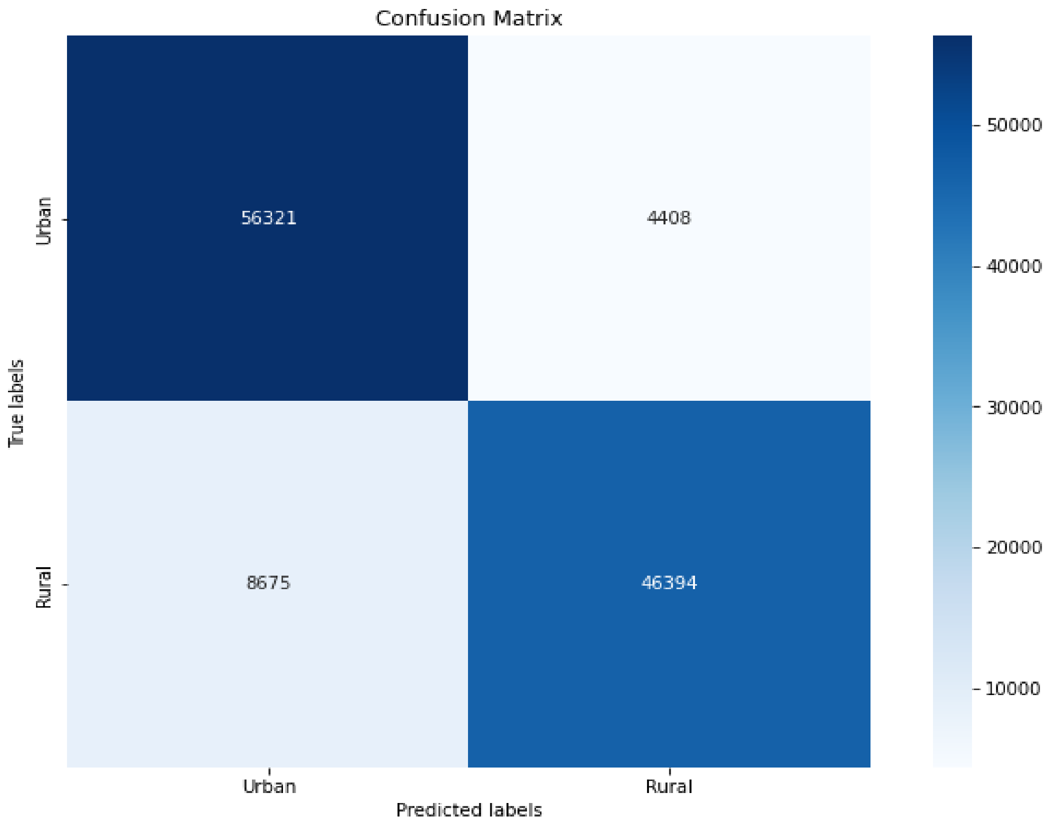


Figure 32: Confusion Matrix of Location Prediction before upsampling

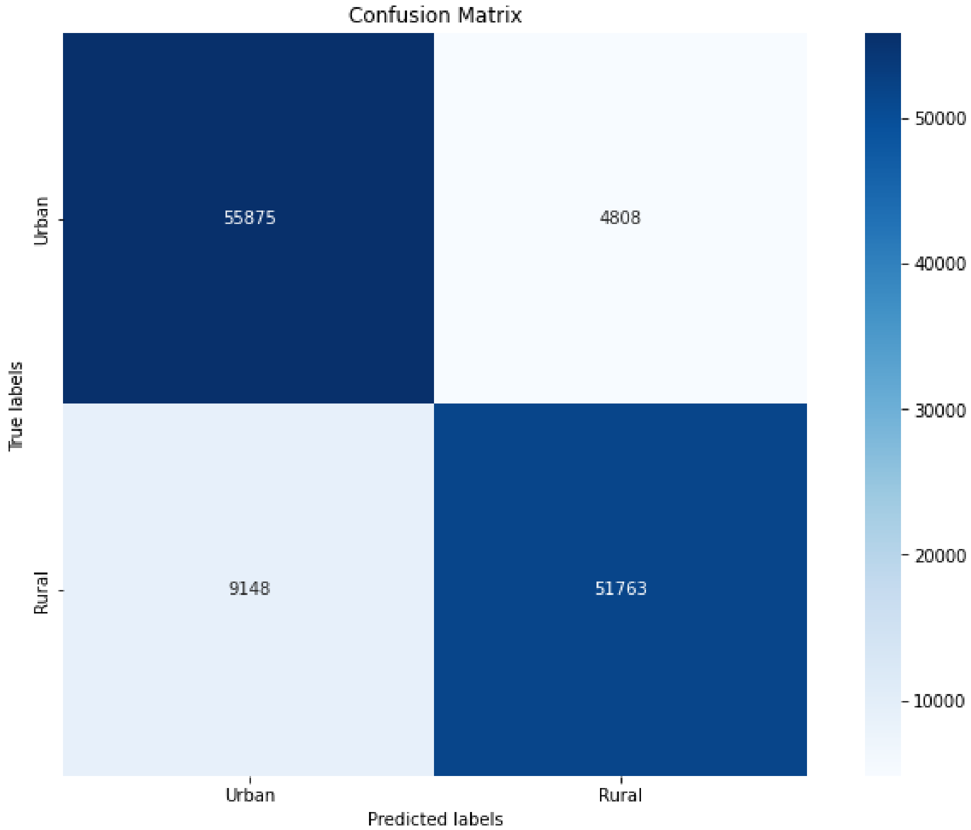


Figure 33: Confusion Matrix of Location Prediction after upsampling

For location prediction, the classifier did a good job in classifying before and after upsampling, however this time around it seems the classifier did better before upsampling the training set.

PROBABILITY PREDICTIONS FOR URBAN AND RURAL

	Accident_Index	Urban	Rural
5	2019010155194	0.851183	0.148817
13	2019010155195	0.925247	0.074753
14	2019010155195	0.925247	0.074753
29	2019010155196	1.000000	0.000000
32	2019010155198	0.582132	0.417868
...	...	...	...
463177	201963DF03019	0.016301	0.983699
463182	201963DF03019	0.016301	0.983699
463183	201963DF03019	0.016301	0.983699
463184	201963DF03019	0.016301	0.983699
463188	201963DF03419	0.016301	0.983699

115834 rows × 3 columns

What the data-frame looks like after merging the probability result with their predicted Accident Index for Rural or Urban location prediction.

LOGISTIC REGRESSION AND ANN MODELLING ON CASUALTY SEVERITY

	precision	recall	f1-score	support
1	0.58	0.71	0.64	18678
2	0.53	0.34	0.41	18309
3	0.57	0.64	0.60	18408
accuracy			0.57	55395
macro avg	0.56	0.56	0.55	55395
weighted avg	0.56	0.57	0.55	55395

With logistic regression the model produced an accuracy of 57% upon exploring the classifier further. From the classification report, some classes were better predicted than some. The recall score of the second class (Serious class) showed a low score.

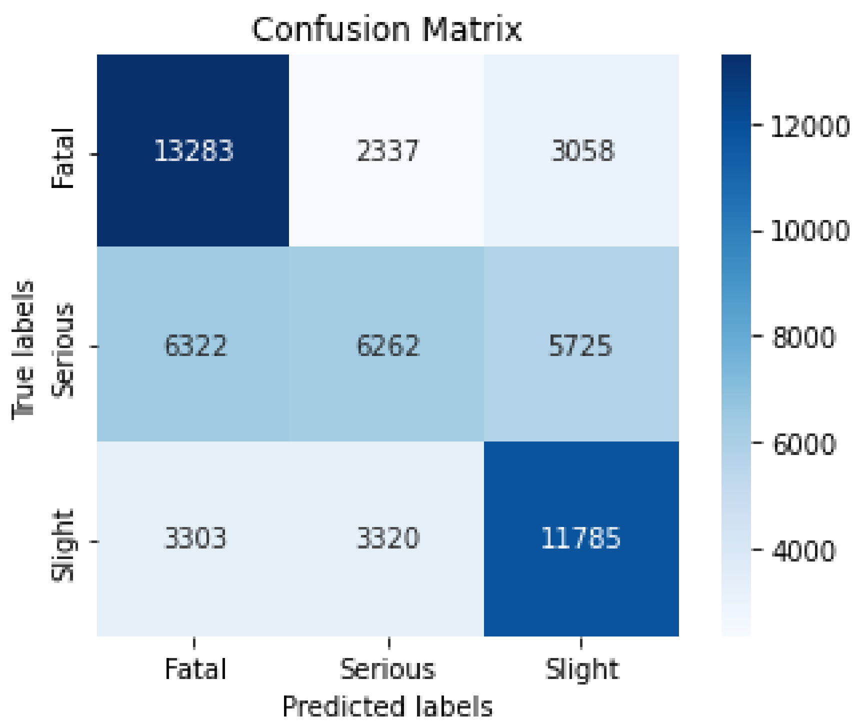


Figure 34: Logistic Regression Confusion Matrix for Casualty Severity

ANN

```
3426/3426 [==============================] - 6s 2ms/step - loss: -16546394.0000 - accuracy: 0.0132
Epoch 9/20
3426/3426 [==============================] - 6s 2ms/step - loss: -22977812.0000 - accuracy: 0.0132
Epoch 10/20
3426/3426 [==============================] - 6s 2ms/step - loss: -30840942.0000 - accuracy: 0.0132
Epoch 11/20
3426/3426 [==============================] - 6s 2ms/step - loss: -40272200.0000 - accuracy: 0.0132
Epoch 12/20
3426/3426 [==============================] - 6s 2ms/step - loss: -51397000.0000 - accuracy: 0.0132
Epoch 13/20
3426/3426 [==============================] - 6s 2ms/step - loss: -64356920.0000 - accuracy: 0.0132
Epoch 14/20
3426/3426 [==============================] - 7s 2ms/step - loss: -79294368.0000 - accuracy: 0.0132
Epoch 15/20
3426/3426 [==============================] - 6s 2ms/step - loss: -96321728.0000 - accuracy: 0.0132
Epoch 16/20
3426/3426 [==============================] - 6s 2ms/step - loss: -115564208.0000 - accuracy: 0.0132
Epoch 17/20
3426/3426 [==============================] - 6s 2ms/step - loss: -137197952.0000 - accuracy: 0.0132
Epoch 18/20
3426/3426 [==============================] - 6s 2ms/step - loss: -161315072.0000 - accuracy: 0.0132
Epoch 19/20
3426/3426 [==============================] - 7s 2ms/step - loss: -188065408.0000 - accuracy: 0.0132
Epoch 20/20
3426/3426 [==============================] - 7s 2ms/step - loss: -217577328.0000 - accuracy: 0.0132
<keras.callbacks.History at 0x7fcecaafa450>
```

The final accuracy score of 13% for ANN was low. Not a useful architecture for this dataset.

RECOMMENDATION

Traffic accidents seem to occur more on Fridays; therefore, it is imperative that the government put in more resources into making it safer for everyone commuting on Fridays. More government agencies, such as the police, should be on duty to ensure that no one is driving while inebriated and that cars are not exceeding the speed restrictions for the designated regions. Since traffic accidents are more common in urban areas, residents should be educated about the significance of driving safely. Drivers between the ages of 26 and 35 should be required to take easy driving exams once a month to refresh their knowledge of the rules of the road. When car drivers breach the laws of the road, they should be heavily sanctioned.

REFERENCES

Ahmed, S., Hossain, M., Bhuiyan, M. I., & Ray, S. K. (2021). A Comparative Study of Machine Learning Algorithms to Predict Road Accident Severity. *20th International Conference on Ubiquitous Computing and Communications (IUCC/CIT/DSCI/SmartCNS)*, 392-399.

Gov.Uk. (2022, Februray 26). *Driving lessons and learning to drive*. Retrieved from Gov.UK: https://www.gov.uk/driving-lessons-learning-to-drive#:~:text=You%20can%20apply%20for%20a,Personal%20Independence%20Payment%20(PIP).

Kohli, S. (2019, November 18). *Understanding a Classification Report For Your Machine Learning Model*. Retrieved from Medium: https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397

Office, M. (2019, December 23). *2019: A year in review*. Retrieved from Met Office: https://www.metoffice.gov.uk/about-us/press-office/news/weather-and-climate/2019/weather-overview-2019

Satpathy, S. (2020, October 6). *Overcoming Class Imbalance using SMOTE Techniques*. Retrieved from Analystics Vidhya: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/

TIME AND DATE. (2022, April 20). *Daylight Saving Time Around the World 2019*. Retrieved from Time and Date: https://www.timeanddate.com/time/dst/2019.html