MAKERERE UNIVERSITY

COLLEGE OF COMPUTING AND INFORMATION SCIENCES

(YEAR II) RECESS TERM 2019

IMPLEMENTATION REPORT

FOR PYTHON PROJECT

PROJECT MEMBERS

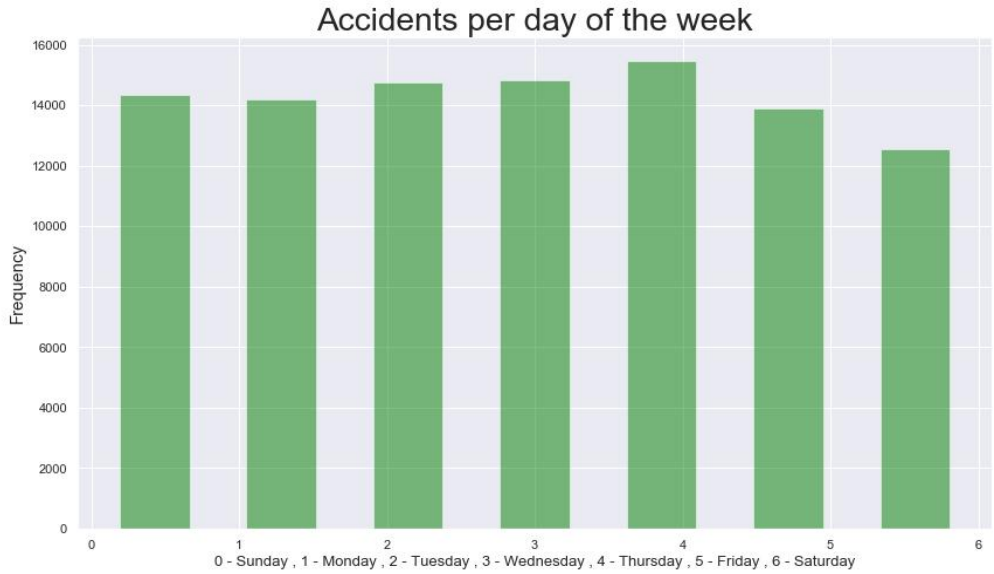
NAME	REGISTRATION NUMBER	STUDENT No
Nagaba Angel	17/U/726	217000189
Okello Marvin Kevin Ochira	17/U/9569/PS	217017015
Karungi Lydia	17/U/4676/PS	217002012
Wepukhulu Bruno	17/U/10891/PS	217012574

PROJECT COORDINATOR: MR. KAMULEGEYA GRACE

INTRODUCTION

After analysing the dataset we came up with visualisations which we used to come up with insights that are explained below.

HISTOGRAM OF ACCIDENTS AND WEEK DAYS



Explanation

Above is a histogram that shows the frequency of accidents for different Daysof the week, from here we can see that Thursday has the highest number of accidents from 2005 to 2015.

Insights

Keeping in mind that the accident numbers could be depending on traffic amount on a particular day:

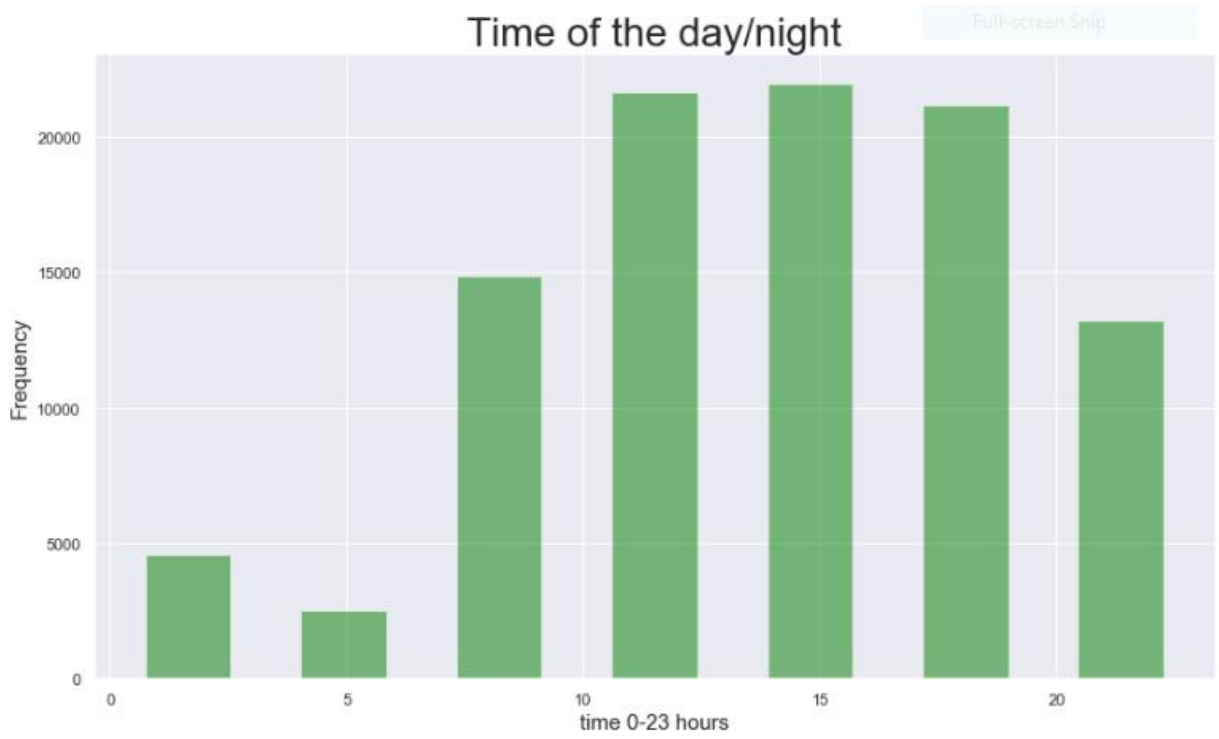
- Since Thursday is a week day, we can conclude that accidents may be high on this day since people are going to and from work which increases the amount of traffic hence increased chances of traffic accidents.
- And Saturday has the least number of accidents and this may be because it's a weekend and few people go to work and most people prefer to spend the weekends at their homes which reduces on the traffic.

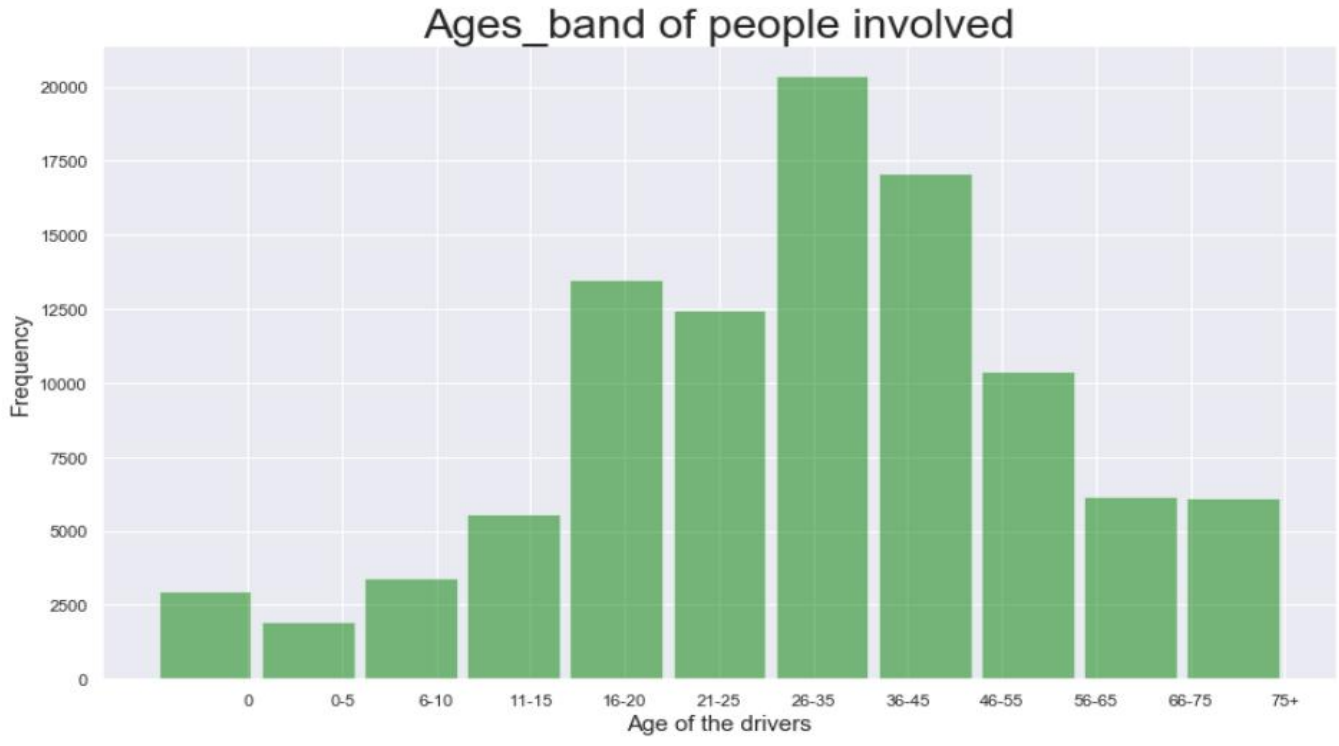
Frequency of accidents and time of the day

According to the histogrambelow,we found out that most of the accidents happened around after noon.And few accidents happen in nmorning hours.

Insight

We can assume that this time of the day has the most traffic since most people leave work in the afternoon.


Time of the day/night
Frequency
20000
15000
10000
5000
0
0
5
10
15
20
time 0-23 hours


Ages_band of people involved
Frequency
20000
17500
15000
12500
10000
7500
5000
2500
0
0
0-5
6-10
11-15
16-20
21-25
26-35
36-45
46-55
56-65
66-75
75+
Age of the drivers

Most of the casualties involved in accidents are aged around 25 to 35. We assume that drivers with the age 25 to 35 are more in number compared to drivers with other ages.

Insights/conclusions

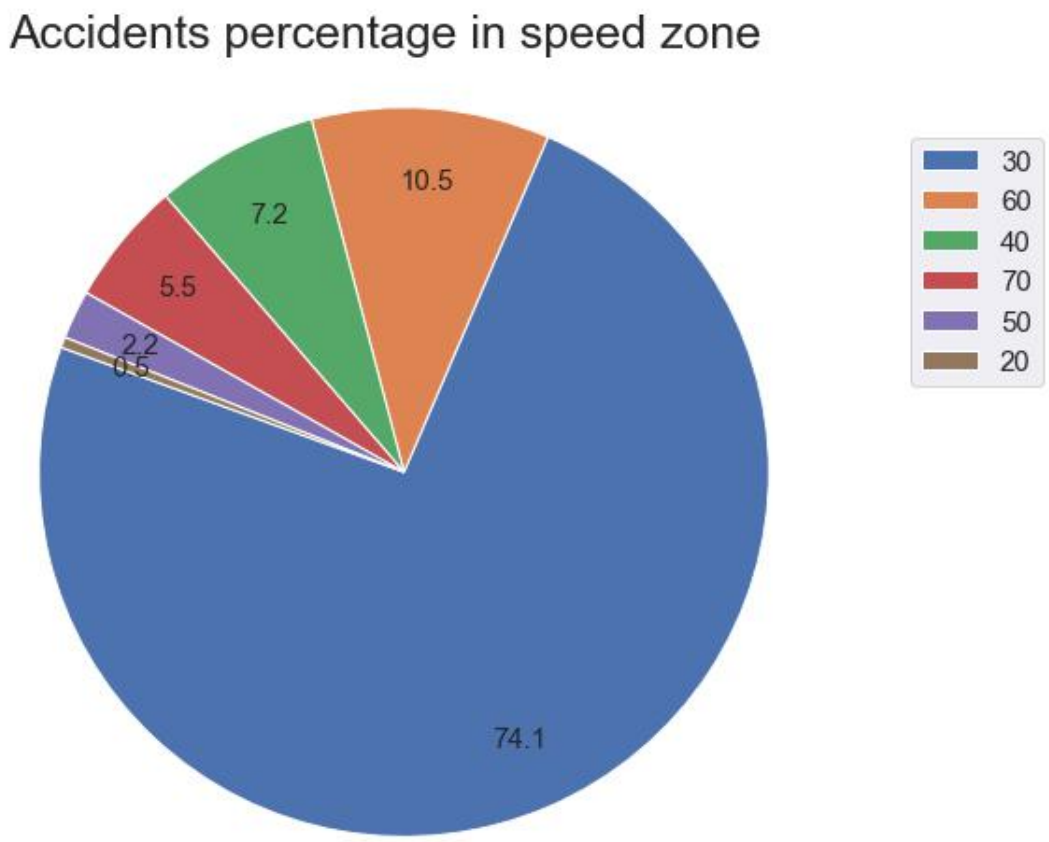
We assume that most of these young drivers are involved in accidents because they take many risks such as not wearing a seat belt, cutting in and out of traffic, speeding to impress passengers and performing other dangerous manvoeures.

Most of the young drivers drive under the influence of alcohol and drugs. Drink driving reduces co-ordination, slows down reaction, and impairs judgment to speed, distance and risk.

Recommendations

- We recommend that most of the sensitization of road usage should be frequently directed to the youth so as to reduce accidents.
- The government should set strict laws and regulations concerning such risks and ensure that they are implemented

PIE CHART



According to the above piechart, most accidents occurred on the roads where the speed limit is 30 compared to other speed limits. But we can also see accidents still occurred on roads with other speed limits though not as much the speed limit of 30.

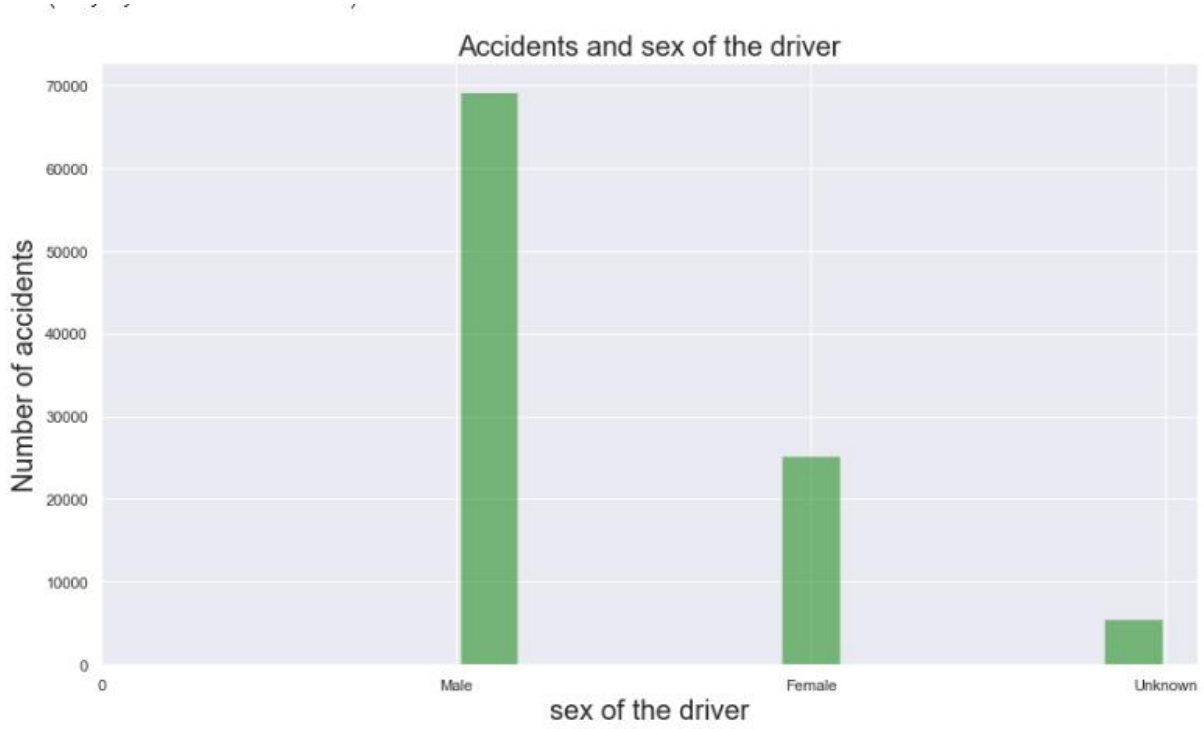
Insights

According to DRIVING-TEST.com, the speed limit of 30 is standard for built-up Urban roads in UK. We conclude that most accidents occur in Urban areas compared to Rural areas.

Recommendation from the piechart

- There needs to be a balance between higher speeds and lower speeds depending on the traffic
- We also advise drivers to drive at lower speed limit because the less time one spends thinking about cops, speeding tickets and slamming on the breaks, the more time one can actually spend driving well.

DISTRIBUTION OF ACCIDENTS ACCORDING TO DRIVER'S SEX



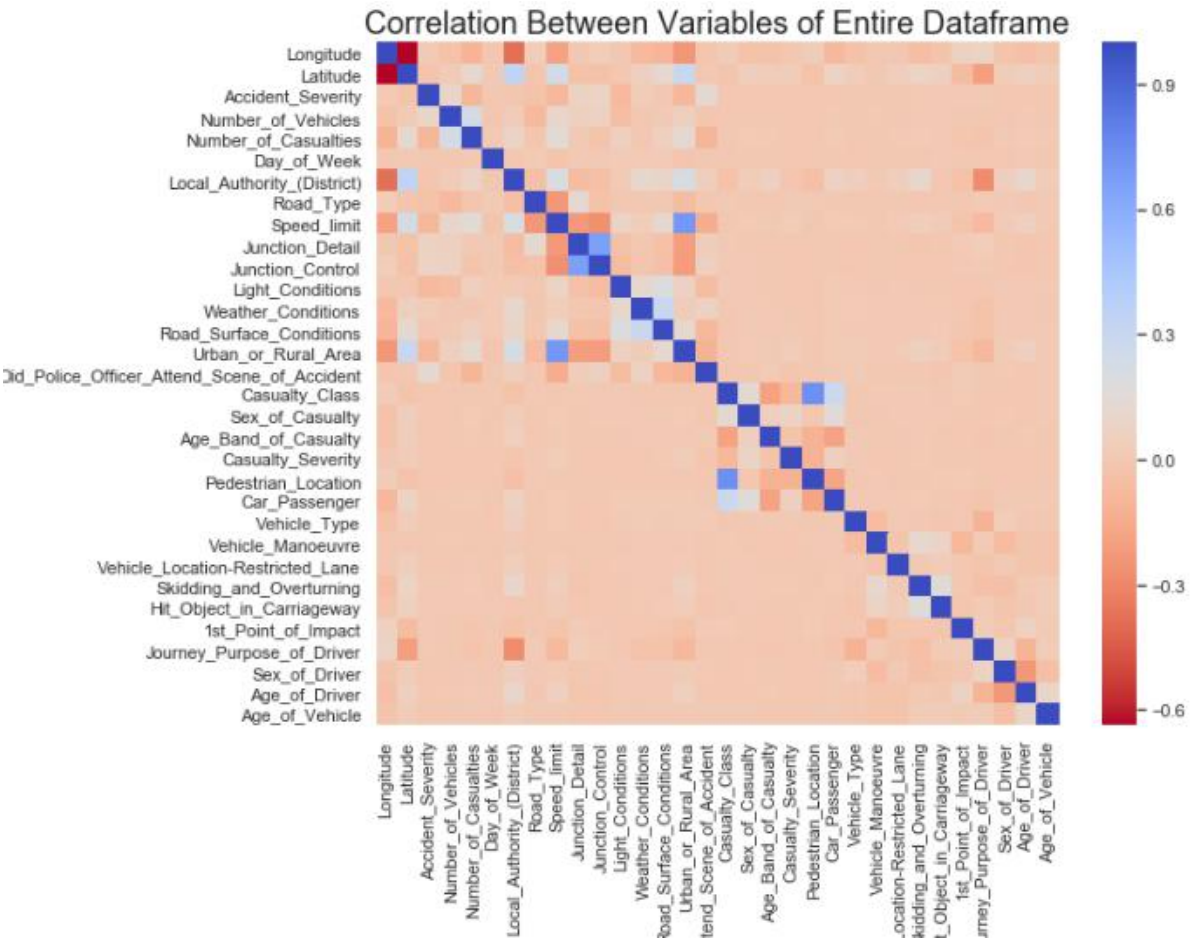
According to the histogram above, Males cause the most accidents compared to females and the Unknown has the least number of accidents. We can therefore conclude that females are more careful than men while on the road.

Insights

According to Steve Stradling an expert in transport psychology at Napier University in Edinburgh, Men are known to take more risks and that applies when there behind the wheel as much as anywhere else.

In conclusion, we think men may cause the most accidents because of some reasons like over speeding and over taking in sharp corners since there more of risk takers compared to women in the real world.

CORRELATION ANALYSIS



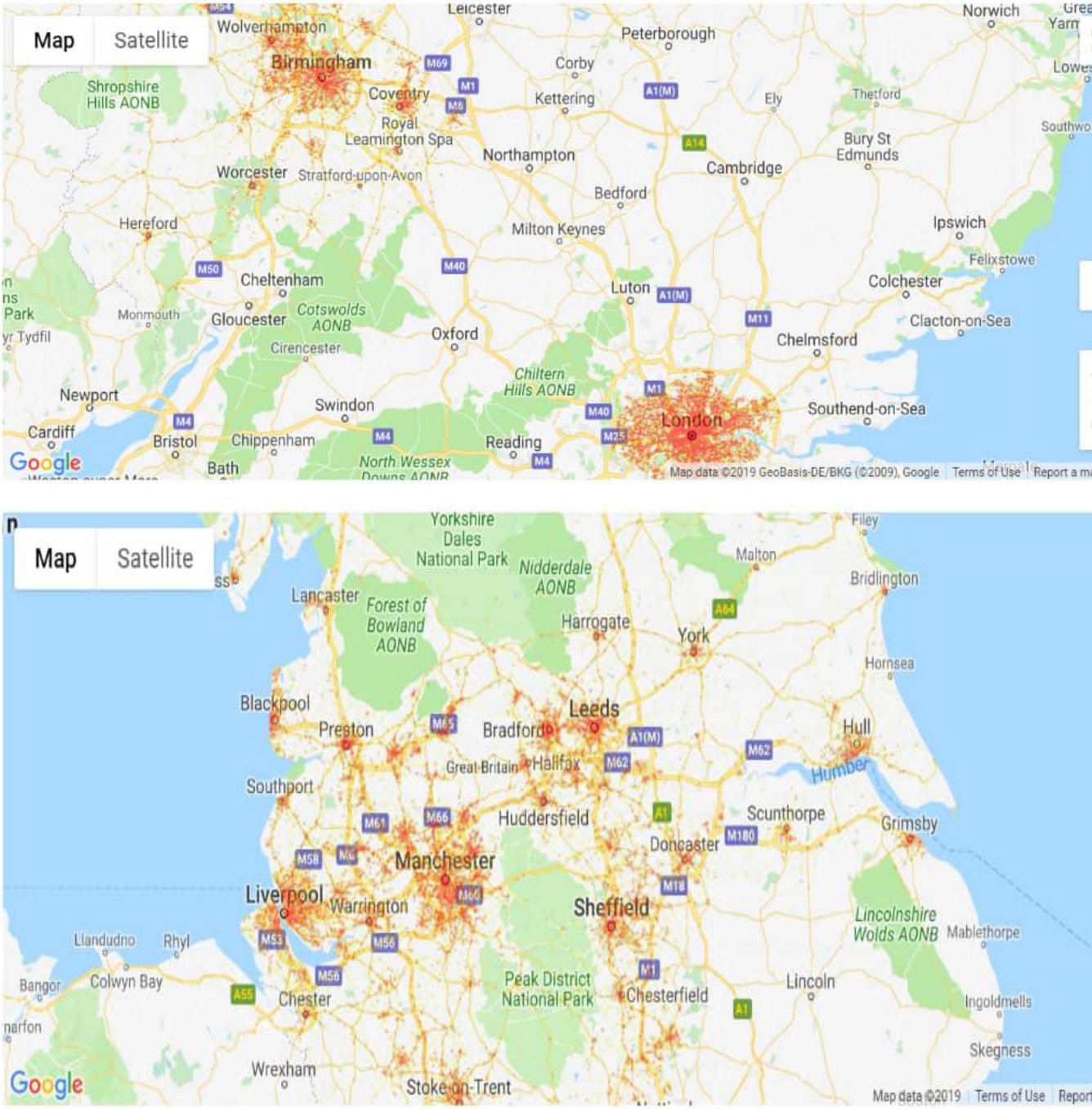
Explanation

The above heat-map shows the co-relation between variables in the dataset .And since the dataset is in numeric values,we could easily findout correlation between values.

Insights

As we see there is no much strong correlations between variables.There is only one strong correlation between speed limit and Urban or Rural area.

GOOGLE MAPS



Explanation

Above is part of the map of United Kingdom(UK), we choose to use the google maps because of the following reasons;

- They provide a panoramic view of the buildings and areas surrounding a particular street.
- Googles maps provide additional informational that may be useful for example current traffic load, Road work and road closures.

Insights

Most of the accidents occurred within cities instead of highways. As seen from the map the concetrated red areas show places with more accidents for example the cities of london, manchester, leeds and many others. This could be because traffic is more congested in cities than highways.

MACHINE LEARNING

We decided to use logistic Regression, K-Nearest Neighbors algorithm and Linear regression for our analysis. We divided our dataset into train data and test data considering our independent variable to be Accident_severity and other variables to be dependent variables.

1. Logistic Regression

The data was fed to a logistic regression, linear regression as well as K-NN to evaluate their performance and the following output was seen.

```
Classifiers:  LogisticRegression Has a training score of 86.0 % accuracy score
Classifiers:  KNeighborsClassifier Has a training score of 85.0 % accuracy score
Classifiers:  LinearRegression Has a training score of 5.0 % accuracy score
```

From the above, it can be seen that Logistic Regression has the best performance of the three machine learning algorithms.

On printing the classification report of the best algorithm in this case Logistic regression, the output is as follows:

	precision	recall	f1-score	support
0	0.99	1.00	0.99	29618
1	0.00	0.00	0.00	373
micro avg	0.99	0.99	0.99	29991
macro avg	0.49	0.50	0.50	29991
weighted avg	0.98	0.99	0.98	29991

Basing on the precision results which is 0.99 for none fatal(0) accidents,it means when our model predicts that accidents do not actually lead to death, it is correct 99% of the times

The recall results(1.00) implies that 100% of accidents are correctly predicted by the model.

Since f1 score(0.99) is the mean of precision and recall and it is high therefore both precision and recall of the classifier indicate good results.

Confusion matrix

A confusion matrix is used for finding the accuracy a classification model.And for this case the output is in 4 categories(True positives, false positives, true negatives, false positives) that is to say a 2*2 matrix as shown below

```
confusion_matrix(y_test,predictions)
```

```
array([[29612,     6],
       [  373,     0]], dtype=int64)
```

Accuracy

Since accuracy in this case refers to the number of correct predictions made by the predicting model over the rest of the predictions

When the accuracy is tested the output which is shown below, 0.987.. shows that our model and predictions are 100% accurate

```
accuracy_score(y_test,predictions)
```

```
0.9873628755293254
```

References

https://www.theguardian.com/science/2004/may/13/thisweeksciencequestions1