MAKERERE UNIVERSITY
COLLEGE OF COMPUTING AND INFORMATION SCIENCES

(YEAR II) RECESS TERM 2019
FINAL 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

CONCEPT NOTE FOR ROAD ACCIDENTS PROBLEM

INTRODUCTION.

This part of the document shows the proposed area of interest for our recess data science project.

Our proposed area of interest is road traffic accidents. The major inspiration for this project is the increased death rate due to road traffic accidents. Some objectives of this project is to analyse data in order to come up with insights to help us answer various questions concerning accidents e.g the likelihood of death when an accident occurs, the likelihood of accidents for different age groups e.t.c. we also hope to come up with policy recommendations to help reduce road traffic accidents.

BACKGROUND

We got our data from https://www.kaggle.com/silicon99/dft-accident-data

The dataset was collected by the UK police forces on every vehicle collision for the period of 2005-2015. The data is usually collected when an accident happens by filling a Stats19 Report Form from the link

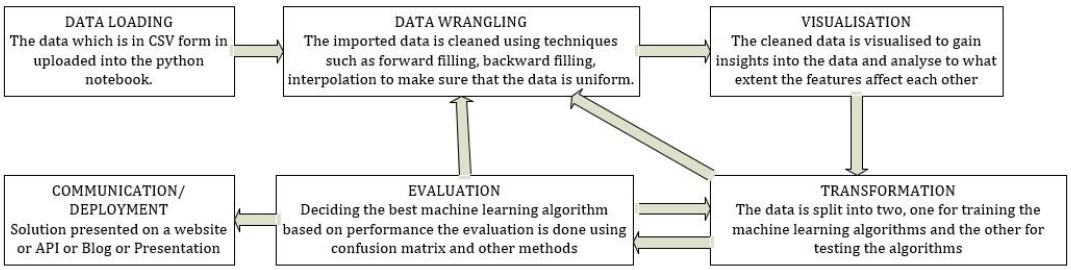
http://docs.adrn.ac.uk/888043/mrdoc/pdf/888043_stats19-road-accident-injury-statistics-report-form.pdf

The dataset is multivariate containing of three CSV files i.e accidents, casualities and vehicles. Accidents is the primary dataset which has 32 columns. It has the unique Accident_index which references the casualities and vehicles tables. The casualities table has 15 columns and the vehicles table has 22 columns. Both the vehicles and casualities table contain the Accident_index that references to the primary dataset(Accidents).

This dataset will help us to analyse the trend of accidents in the uk and also apply various data visualisation tools to come up with insights.

DATA PIPELINE

This illustration of the pipeline was creatively generated by the group and it is explained in the steps below.



A data pipeline enables the data scientist to transform data from one representation to another through a series of steps. It involvesthe following steps:

1. Data Loading

At this step, the data is loaded into the python notebook. The three csv files i .e accidents,casualities and vehicles are all loaded and transformed into dataframes using the pandas python library. The three datasets are then merged into a single dataframe for further analysis. This is easier because all the three datasets have a common Unique id Accident_Index.

2. Data Wrangling

Data Wrangling involves is the process by which data is cleaned so that it becomes easy to analyse and visualise it. The data wrangling process is as follows:

Checking missing values The dataset is checked to find out if there are any missing values and int this case we have two types of missing values which is -1 and ‘Nan’. various methods such as Backward fill, Foward fill, Interpolation, Dropping all missing values, Filling in the mean or median e.t.c can be used to remove the missing values.

Label encoding

This is the process where the categorical features of the data are transformed into numeric values usually 0’s and 1’s. for this case the dataset features are all numerical already so there is no need to label encode.

Feature scaling

This is a step of data pre-processing which is applied to independent variables or features of data. It helps to normalise the data within a particular range. It also helps in speeding up the calculations in an algorithm. The data set will be feature scaled in

order to bring the outliers closer to other values in the dataset. Two main libraries are used i .e RobustScaler and StandardScaler.

3. Visualisation

Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns. Data visualisation will help us to determine the spread of the dataset through measures like range,quartiles and the interquartile range,variance and standard deviation, determine the correlation between features. Various visualisation tools such as heat maps, histograms, scatter plot, line plots, box plots, regression analysis and they will be applied to the following key features of the dataset: Accident severity, Number of vehicles involved in the accident, day of the week the accident happened, time of the day when the accident happened, Road type on which the accident happened, speed limit, light and weather conditions, vehicle type, vehicle manouvre, age of the driver,sex of the driver, point of impact, age band, age of the vehicle, casuality class, casuality severity, casuality type, age of the casuality e.t.c. Visualisation tools like hiatograms, maps, scatter plot, line plots e.t.c will be used to analyse the spread of data which will in turn help us come up with better policy recommendations.

4. Transformation

At this step of the pipeline, machine learning algorithms are applied onto the data and they will help us to predict our final outcome which is the possibility of death when an accident happens. Machine learning algorithms include: logistic regression, linear regression, Support Vector Machine, Decision trees, K- Nearest Neighbours e.t.c

5. Evaluation

Theaccuracy of the above algorithms is then evaluated based on the scores in order to find out the best algorithm. This evaluation can be carried out using the precison, Recall, confusion matrix and cross validation performance measures.

6. Deployment/Communication

After the data has been analysed and visualised, the solution has to be presented or communicated. This can be done with the help of either of the following tools: A blog, a website, an API, a dashboard or a presentation.

SYSTEM REQUIREMENTS SPECIFICATIONS

PURPOSE

This part of the report describes what the system will do and how it will be expected to do it . This document will include a use case diagram which shows how the system interacts with the outside world.

Intended audience and Reading suggestions

The intended audience for this project includes, the data scientists(students working on the project), Road authorities and road users(drivers, riders, pedestrians, passengers).

The users of the Data Science Pipeline include the following:

- Data scientists

These are the developers of the system. They extract the relevant data from the data collected by the police and clean it, visualize it, model the data and interpret it.

- Data clerks in the Police.

These enter the collected data into the system. From they get the data from the Report forms in which information is recorded when an accident occurs.

- System Administrators

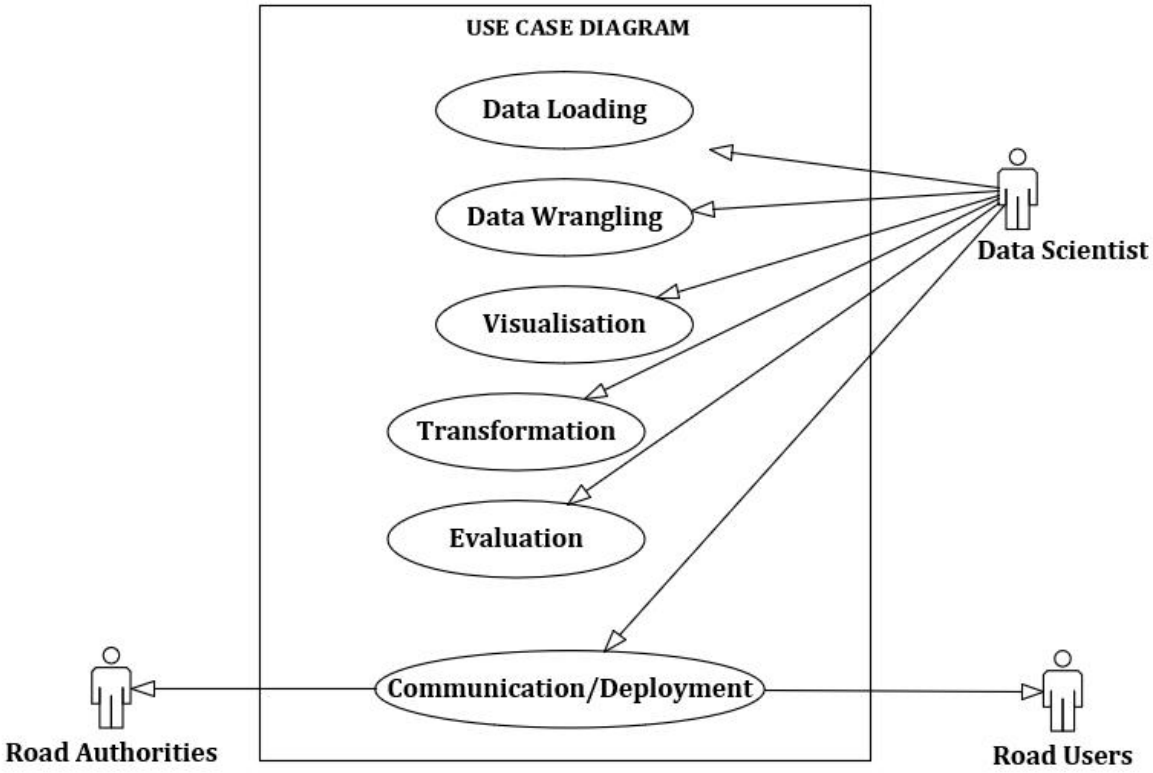
These have access rights to the system and carry out analysis of the data.

- System testers

These test the functionalities of the system, how correct the data output is and whether all needs of the users are fully satisfied.

Use Case Diagram

A use case diagram helps to describe the system's behaviour and how a system interacts with outside world. The use case diagram consists of actors (who interacts with the system) and use cases(this shows what the actors do with the system). A use case basically helps in understanding of the system form the end user's point of view.



Source: Generated by the group using Microsoft Visio

Roles of the Actors in the Use case Diagram

Use Case	Actors	Purpose	Justification
Data Loading	Data Scientist	Extract data from the files into the development environment	This is done so a to have data in an environment where they can be processed
Data Wrangling	Data Scientist	To attain data that is free from missing values, duplication, irregularities and in their right formats	The data has to be cleaned so as to have accurate and consistent operation of algorithms
Visualisation	Data Scientist	To have an overview on the behaviour of data	This enables one to understand the behaviour of data.
Transformation	Data Scientist	To prepare data such that they can be fed to the machine learning algorithms to produce the best results	This is done so as to have more accurate learning for the machine learning algorithms.
Evaluation	Data Scientist	To know which machine learning algorithm performs best.	This is done to know which machine learning algorithm performs better.
Communication/Deployment	Data Scientist Road	To communicate to the users of the system and the roads of the	The communication of the results aids the actors in better decision

	Authorities Road Users	outcome of the analysis	making when using or planning for the roads.

Use Case Detailed Description

Use case name:	Data Loading
Priority:	High
Actors:	Data scientist
Description:	The data scientist loads the data from the csv files having all the data into the development environment using python libraries.
Precondition:	The data has to be in csv files, preloaded by data clerks from the stats19-road-accident-injury-statistics-report-form
Conclusion:	This stage concludes when the data from the csv files are loaded successfully into the development environment and are assigned to variables that hold the data ready to be manipulated.
Assumptions:	The raw data from the forms has already been entered into csv files by the data clerks

Use case name:	Data Wrangling
Priority:	High
Actors:	Data scientist
Description:	The csv files may have data with irregularities such as missing values, incorrect formats and others, so in this stage the data scientist makes sure that data is cleaned using data cleaning techniques, so as the data is ready for analysis
Precondition:	The data must be loaded in the development environment
Conclusion:	This stage is concluded when the data no longer has irregularities and duplications

Use case name:	Visualisation
Priority:	High
Actors:	Data Scientist
Description:	The data scientist at this stage manipulates the data to create visualisations in order to understand the behaviour of data
Precondition:	The data must be clean
Conclusion:	This stage is concluded when the data scientist comes up with visualisations of different dimensions of the data.

Use case name:	Transformation
Priority:	High
Actors:	Data Scientist
Description:	at this stage, the data is split into two categories where one will be used to train the machine learning algorithms and the other set to test the accuracy of the algorithm.
Precondition:	The data must be clean and split into training data and test data
Conclusion:	The stage is concluded when the machine learning algorithms are successfully trained and tested

Use case name:	Evaluation
Priority:	High
Actors:	Data Scientist
Description:	The data scientist compares the performances of the machine learning algorithms used, and determines which one performs best.
Precondition:	The machine learning algorithms must be trained and tested.
Conclusion:	This stage concludes with the determination of the best machine learning algorithm.

Use case name:	Communication/Deployment
Priority:	High
Actors:	Data Scientist, Road Authorities, Road Users
Description:	The Data Scientist displays or communicates the findings or the analysis of the dataset to the Road authorities and road users, who use this information to influence their decision making for the use of the roads to reduce the death rate caused by road accidents. The data scientist displays the visualisations he generated to aid the other actors in understanding the results.
Conclusion:	This stage is concluded with display of information through visualisations to the other actors who interface with the system.

SOFTWARE DESIGN SPECIFICATIONS.

Introduction

The system design specifications part of the document describes the details of the pipeline all the way from datasources to the visualisation. It gives a detailed explanation of each of the components of the pipeline and how each of the activities involved shall be performed.

It also explains the libraries that support each of the activities in the pipeline and why those libraries are used. It furthermore explains the viualisation tools that we shall use for our accidents dataset and why those visualisation tools are used.

Representation of the pipeline from datasources to visualisation

Source: this diagram was creatively generated by the group using microsoft visio

Components of the pipeline

Data loading

The first step is data collection, in this case we collected our dataset from www.kaggle.com, which was in csv format.
After, the data is loaded into a dataframe from the csv file in the python notebook.
The is done using the method **"read_csv"** which is a data structure supported by the pandas library.Since we have three datasets we are going to combine them into one dataset. This component of the pipeline is supported by the python pandas library. The pandas library provides tools for writing and reading the data.

Data wrangling

The process of data wrangling(formerly known as data cleaning) involves activities like label encoding, checking and eliminating missing values as well as feature scaling.

Missing values

The first step is to check whether missing values exist in the dataset and this is done by running the line of code "**dataframeName.isnull().values.any()"** if it evaluates to true, it implies that there are missing values. The pandas library provides various methods to deal with missing values namely: foward fill, backward fill, filling with zero, mean or median and interpolation. We shall use the interpolation method to remove missing values in our data is its close to accurate.

Label encoding

This refers to converting the labels of textual columns into numeric form so as to make it machine-readable. Our dataset doesn't require this part of the pipeline since its already encoded.

Feature scaling

It basically helps to normalise the data within a particular range(removing outliers). Sometimes, it also helps in speeding up the calculations in an algorithm.
It is done by methods provided by the done by **scikit-learn.** This package contains a module sklearn.preprocessing that helps us to carry out machine learning. Feature scaling is done because many machine learning algorithms perform better or converge faster when features are on a relatively similar scale and/or close to normally distributed.
From the above package we can either use the standard scaler or robust scaler for feature scaling.

Robustscaler is used when we want to reduce the effect of the outliers. Minmax scaler doesnt reduce the effect of the outliers.
RobustScaler transforms the feature vector by subtracting the median and then dividing by the interquartile range (75% value — 25% value).
StandardScaler standardizes a feature by subtracting the mean and then scaling to unit variance. Unit variance means dividing all the values by the standard deviation. StandardScaler does not meet the strict definition of *scale* I introduced earlier.

Why scikit-learn for feature scaling

Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like Numpy and Scipy.

N.B: Both data loading and data wrangling are supported by the pandas and numpy libraries

Why pandas

-Pandas provides an organised form of data representation. This helps to analyze and understand data better. It can present data in a way that is suitable for data analysis via its Series and DataFrame data structures.

-The package contains multiple methods for convenient data filtering.

-Pandas has a variety of utilities to perform Input/Output operations in a seamless manner. It can read data from a variety of formats such as CSV, TSV, MS Excel, etc.

Why numpy(numerical python) library

Numpy (numerical python) is a library that helps to deal with scientific computations. It provides a high-performance multidimensional array and basic tools to compute with and manipulate these arrays.

Data Visualisation

After the data cleaning/wrangling exercise is done, the data is visualised in order to come up with plots, pictorial representations, that will help us to come up with various insights and conclusions. This exercise is done by importing the matplotliblibrary and seaborn library and then using the various methods to plot. These libraries can be intergrated by importing them before starting the actual visualisation.

The data can be visualised to come up with visualisations like scatter plots, Histogram, maps, multiple linear regression, logistic regression e.t.c

Why matplotlib library

It has a module named pyplot which makes things easy for plotting by providing feature to control line styles, font properties, formatting axes etc. It supports a very wide variety of graphs and plots namely - histogram, bar charts, scatter plots e.t.c which charts happen to be part of our visualisation plan.

Why seaborn library

Seaborn is a Python library created for enhanced data visualization.It is built on top of matplotlib and closely integrated with pandas data structures.

-seaborn provides an API that is dataset-oriented which helps in examining relationships between multiple variables

- Seaborn makes it easier to carry out the data visualisation exercise because it provides automatic estimation and plotting of linear regression models for different kinds of dependent variables

-seaborn library also provides tools for choosing color palettes that faithfully reveal patterns in your data

Reasons for the above mentioned visualisation tools

Scatter plot

-It shows the relationship between two variables

-It shows us the non-regression pattern

-It shows the range of data i.e. Maximum or minimum value can be determined

-Observation and reading is straight forward

Histogram

-This Is an accurate representation of the distribution of numerical data. It relates one variable.

-It helps to visualize the distribution of data

-It identifies different data, the frequency of the data occurring in the dataset and categories which are difficult to interpret in tabular form.

Maps

In our dataset we shall draw a map to show the distribution of accidents.

Maps add value to the reader and helps them visualize and understand the data geographically

Multiple linear regression

- Since we have more than one independent variable in our dataset multiple linear regression is the best type of linear regression to use. It establishes a relationship between dependent variables and the independent variables using the best fit straight line ($y = mx + c$).
- It has the ability to determine the relative influence of the predictor variables to the criterion value for example we could find that the vehicle type and the Age of the driver have a strong correlation to the rate of accidents while the sex of the driver has no correlation
- It also has the ability to identify outliers that is to say the variable that comes out to be out of range compared to others is considered as an outlier.

Logistic regression

Here we use data to find the probability of event=survived accident(less fatalities when an accident happens) and event=died from accident (many fatalities when an accident happens). And it is used when the dependent variable is binary.

Since of dependent variable is accident severity and is not binary i.e it includes slight, serious and fatal, we shall combine slight and serious to form non-fatal so as to use this model.

And in order to avoid over fitting and under fitting in this case, we shall use a stepwise method to estimate the logical regression. Logistic regression is efficient and straight foward in nature, it is also easy to implement.

K-Nearest neighbors

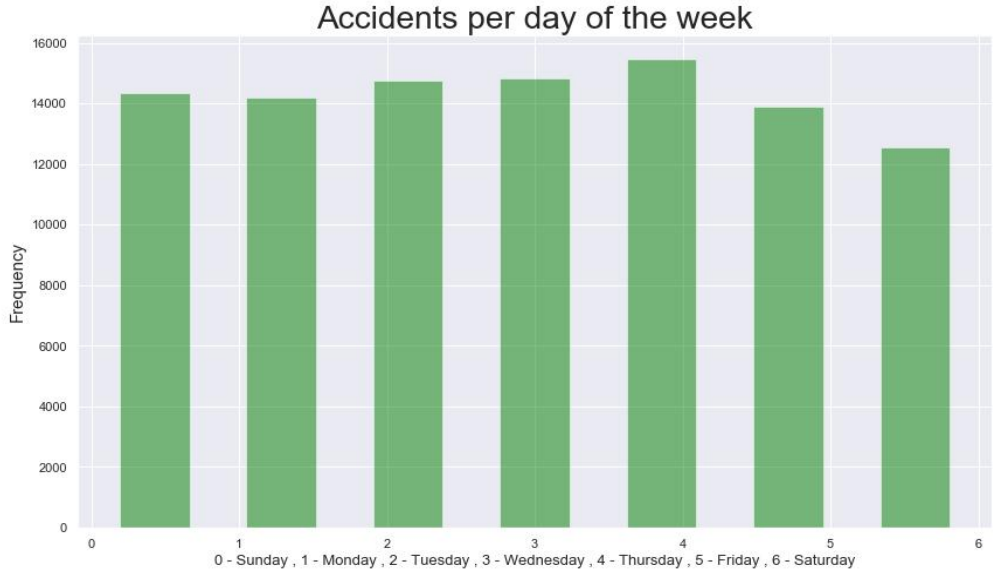
It is very simple to understand and equally implement.

It is a non-parametric algorithm which means there non assumptions to be met to implement K-NN.

K-NN is a memory-based approach,the classifier immediately adapts as we collect new training data.

IMPLEMENTATION REPORT

HISTOGRAM



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:

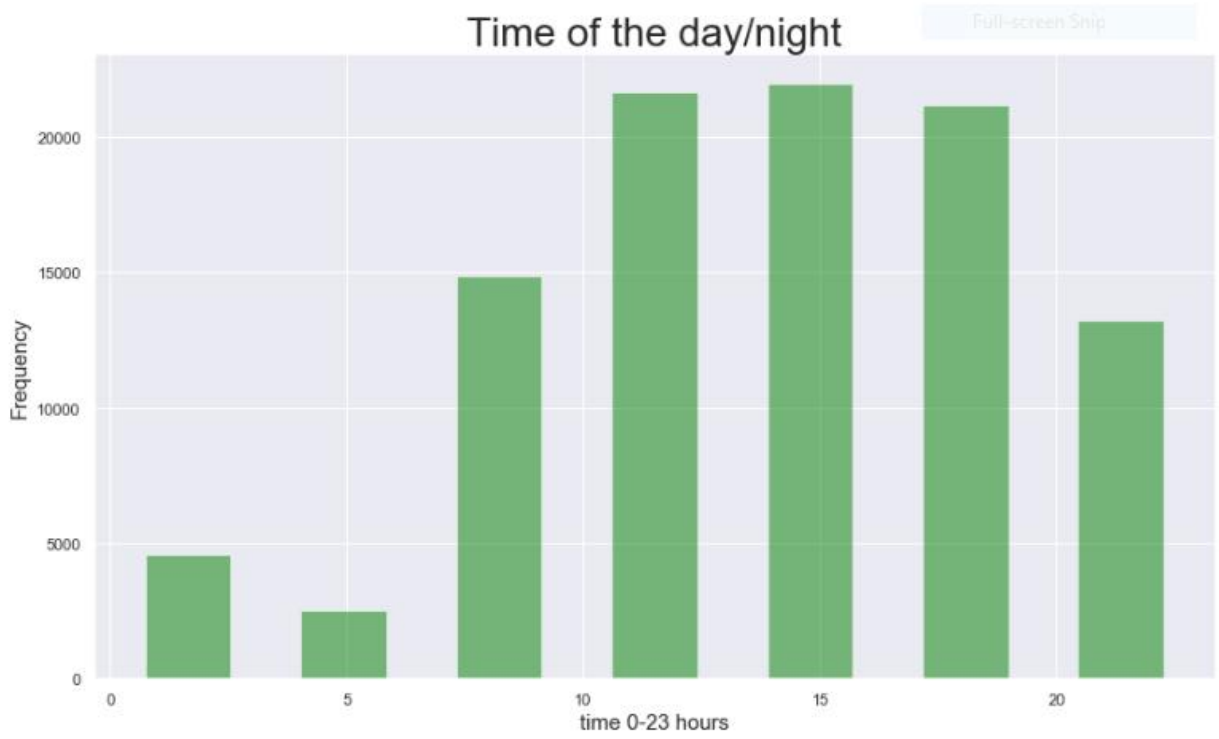
- Since Thursday is a week day, we can conclude that accidents may be high on this day since people are going to and from work which increases the amount of traffic hence increased chances of traffic accidents.
- And Saturday has the least number of accidents and this may be because it's a weekend and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

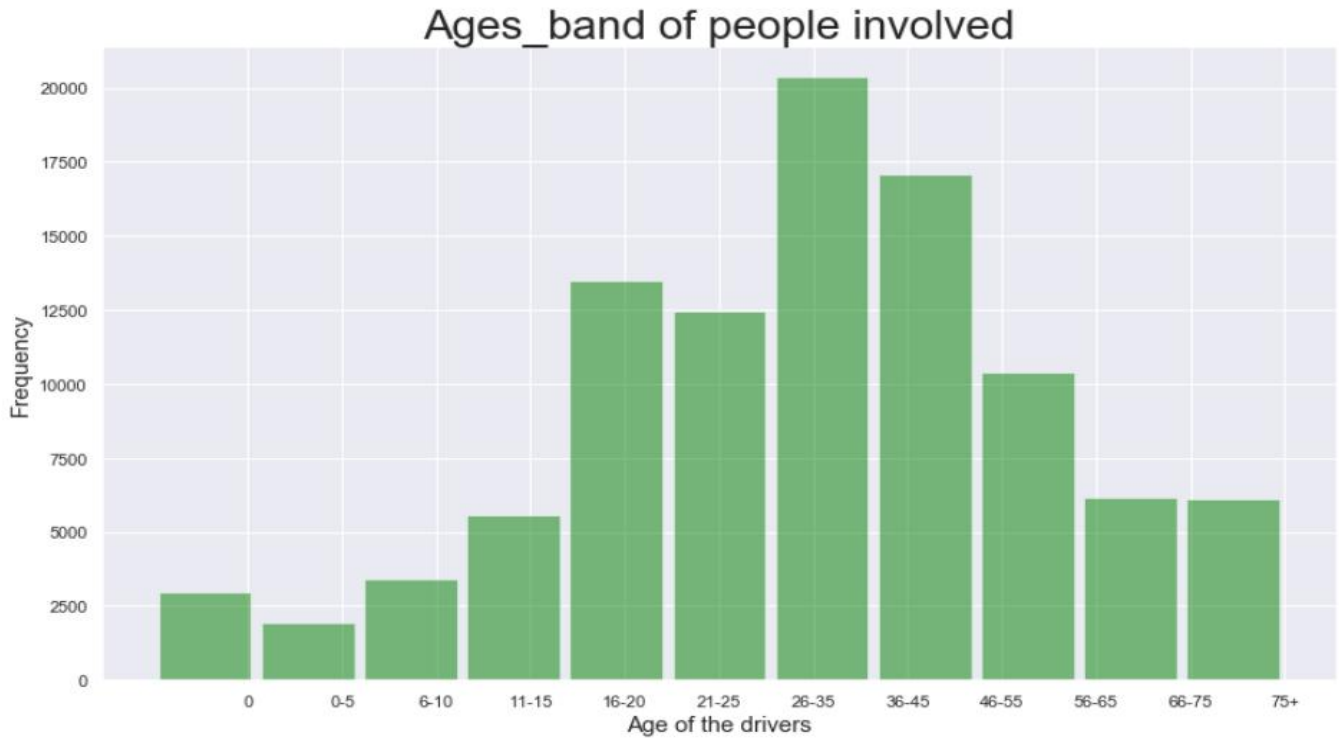
Frequency of accidents and time of the day

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.





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

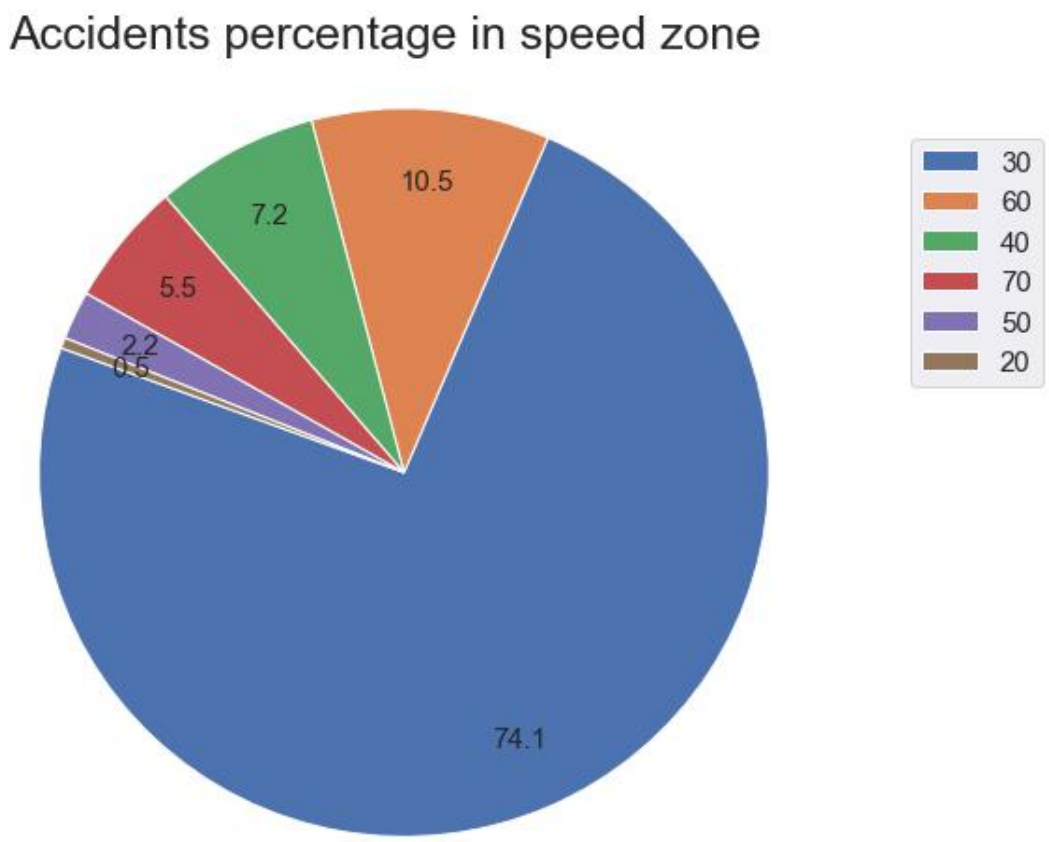
We assume that most of these young drivers are involved in accidents because they take many risks such as not wearing a seat belt, cutting in and out of traffic, speeding to impress passengers and performing other dangerous manvoeures.

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

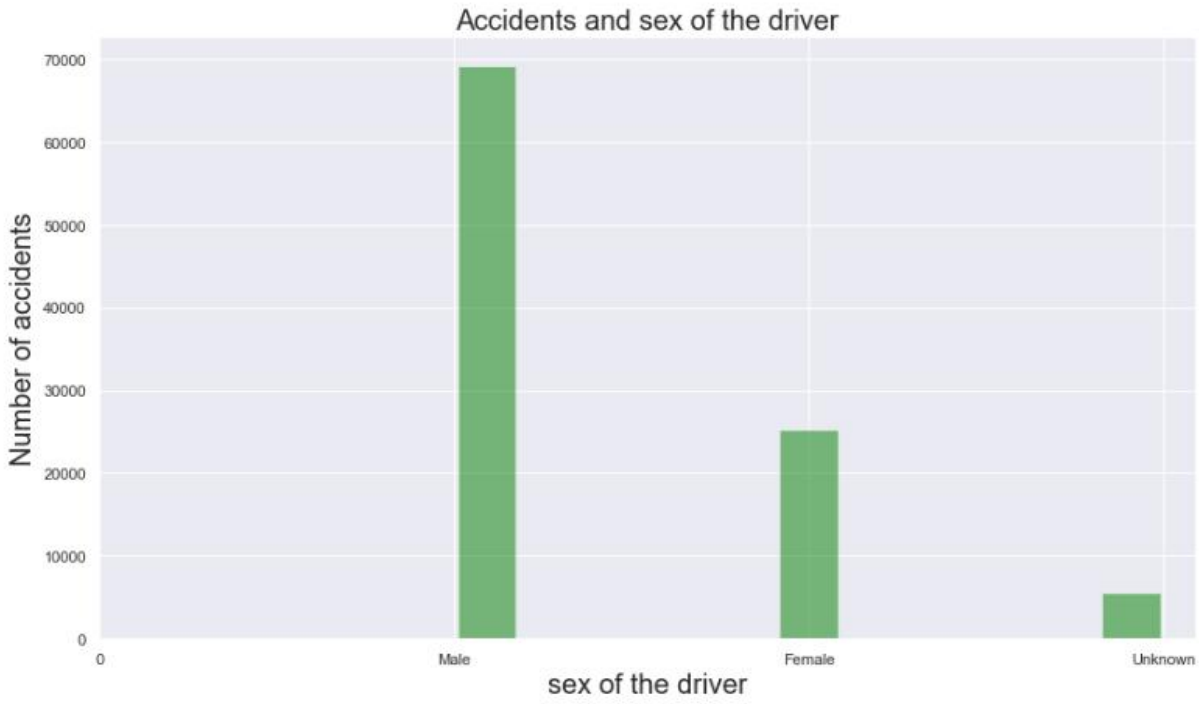
According to DRIVING-TEST.com, the speed limit of 30 is standard for built-up Urban roads in UK. We conclude that most accidents occur in Urban areas compared to Rural areas.

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



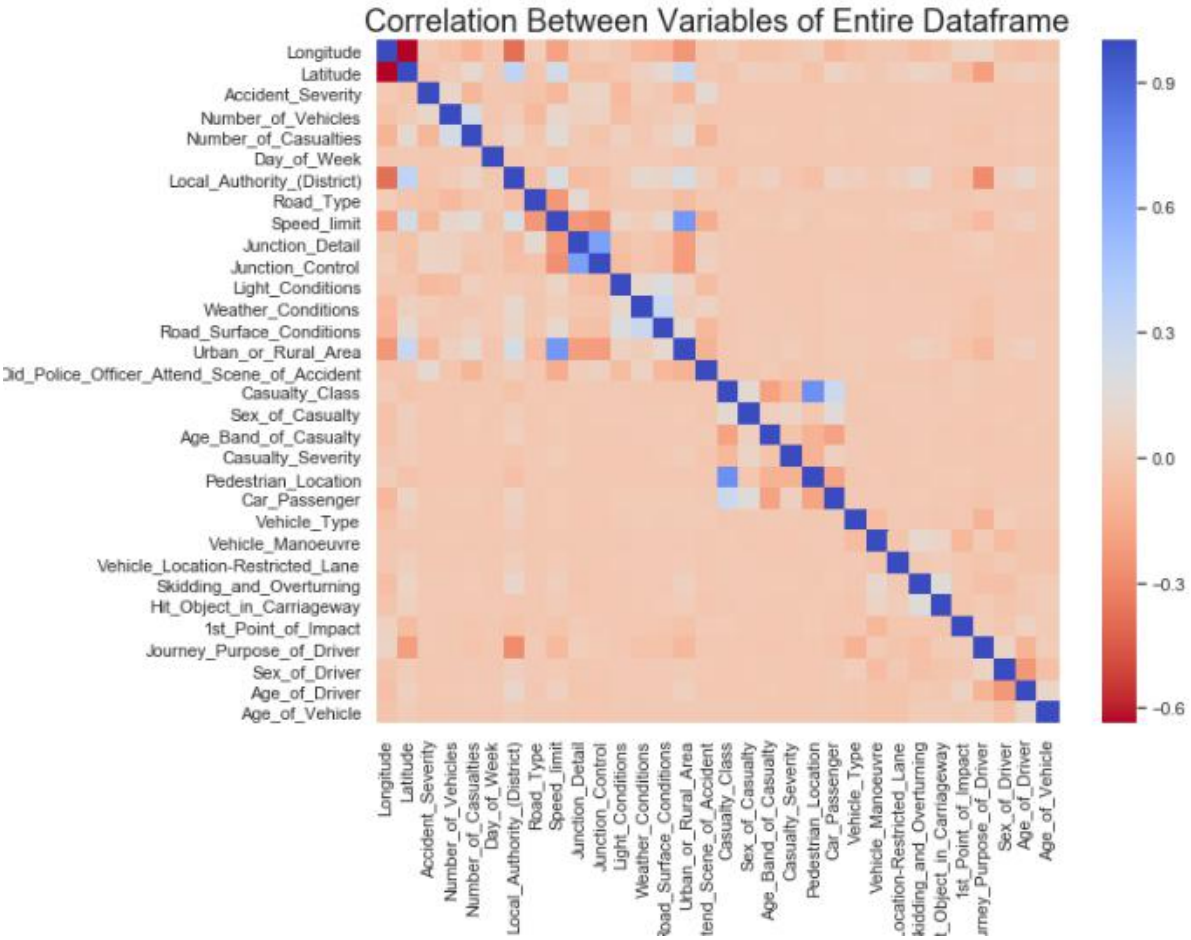
According to the histogram above, Males cause the most accidents compared to females and the Unknown has the least number of accidents. We can therefore conclude that females are more careful than men while on the road.

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



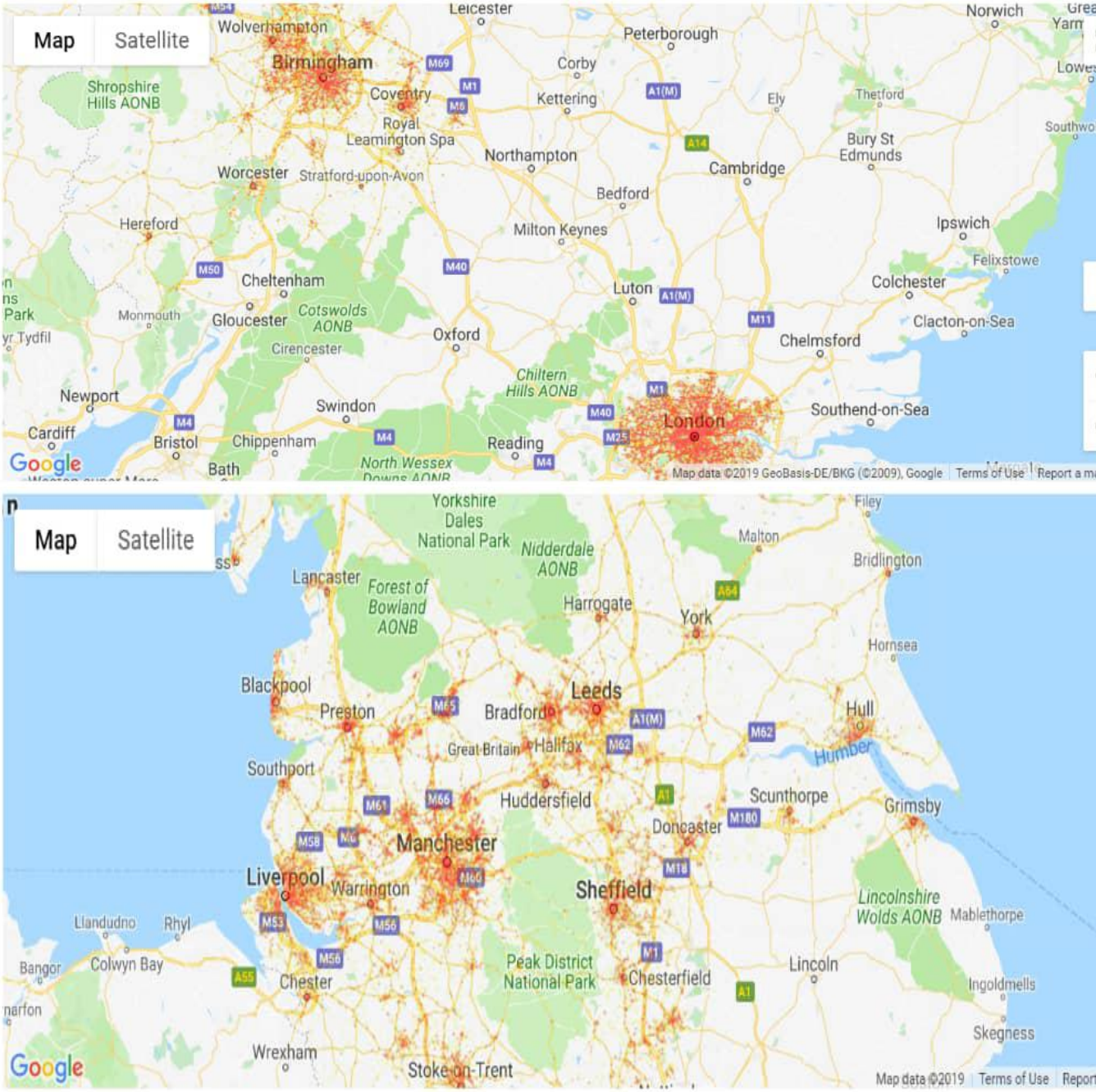
Explanation

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.kaggle.com/silicon99/dft-accident-data

https://dev.to/marsja/essential-python-libraries-for-data-science-machine-learning-and-statistics-5175

https://towardsdatascience.com/scale-standardize-or-normalize-with-scikit-learn-6ccc7d176a02

http://pandas.pydata.org

http://scikit-learn.org/_

https://seaborn.pydata.org/introduction.html

https://www.theguardian.com/science/2004/may/13/thisweekssciencequestions1