UK Traffic Data Analysis

1. Introduction

When it comes to data analysis, it is described as the process of cleansing, manipulating, and modeling data in order to identify information that may be used for corporate decision-making. The objective of data analysis is to extract valuable information from data and to make decisions based on the information derived from the data analyses (Bazerman, M.H. & Moore, D., 2013). So, her in this project we are going to analyze a dataset so that according to the driven information we could take some decisions. The data that we are going to analyze is related to Road Traffic Accidents in United Kingdom.

There are multiple files of the dataset. They contain detailed information on the circumstances of personal injury road accidents that occurred in the United Kingdom from 2005 onwards, as well as information about the types of automobiles involved and the number of individuals who were injured as a result of the accident. There are no other accidents included in the data since they are all reported to the police and subsequently documented using the STATS19 accident reporting form. There are certain exceptions to this rule, including incidents involving just property damage and no human casualties, as well as accidents happening on private roads or in parking lots.

Although it is well known that a considerable number of non-fatal injury accidents do not result in police involvement, it is not known how many fatal accidents do not result in police involvement. The fatality data refer to individuals who died within 30 days of the accident or within a short period of time after the event occurred. In general, this is the international definition, which was defined by the Vienna Convention in 1968 and is the most often used.

Additionally, in addition to including information on the event's date, time, and place, the accident file provides a summary of all registered autos and pedestrians involved in road accidents, along with the total number of casualties, broken down by severity level. The "Accident Index" feature in the casualty and vehicle files may be used to link information included in the files to the proper accident or incident report. The values for longitude and latitude are based on the World Geodetic System (WGS) 1984 coordinate system.

Information on the process of data collection, including the form (STATS19) that was used to collect the statistics, the instructions for completing the form (STATS20), and the definitions that were used in these data can be found on the Department of Transport's website, which can be found at the following address:

http://www.dft.gov.uk/statistics/series/road-accidents-and-safety/

So, we analyze these datasets and will come to some conclusions about which safety measures should we take to avoid a traffic accident.

2. Data Cleaning & Analysis

2.1. Cleaning

There were three data files, and they were each loaded independently before being combined into a single data frame. And then, prior to conducting data analysis, I performed data preprocessing. Data It is necessary to employ preprocessing while scanning a record set, table, or database in order to identify and rectify any faulty or erroneous records (or removed). Cleaning data refers to the process of finding and replacing dirty or coarse data with clean or correct data in order to enhance the overall quality of the data and eliminate mistakes and omissions. It is also referred to as data transformation. For this dataset, I initially checked for null values in the dataset, and there were just 4 percent missing values in one column of the dataset, which was a good result. As a result, I eliminated the characteristics in which the NAN was present. Once I had that information, I looked into the data types of features and discovered that there are some features whose data type is incorrect, such as the data type of the date feature being object, which is incorrect. I changed the data type of their data to the proper one.

2.2. Analysis

After I started performing data analysis. As far as data analysis is concerned, it may be summarized as the process of cleansing, changing, and modeling data in order to discover information that can be used to make choices for the benefit of an organization. Aims of data analysis include the extraction of useful information from data and the formulation of judgments based on the knowledge obtained from the data analysis process.

To provide an example of data analysis at its most fundamental level, when we make decisions in our everyday lives, we explore the ramifications of those decisions by considering what has happened in the past or what will happen if we choose that specific option (Košcielniak, H. & Puto, A., 2015). What we are doing is nothing more than conducting an inquiry into our history or future, followed by the development of conclusions based on our findings. In order to do this, it is vital to first gather memories from our past as well as ambitions for the future. As a result, at this point, everything is based on statistical analysis. When it comes to achieving corporate objectives, data analysis is a term that is used to describe the same job that an analyst does: collecting and analyzing data.

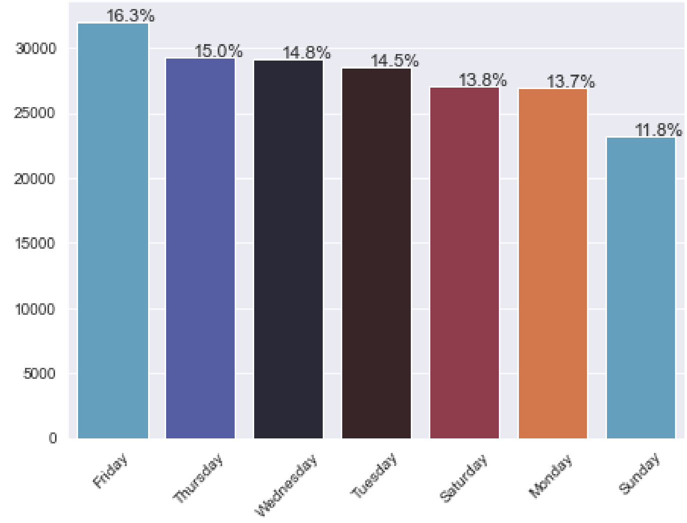
A little amount of analysis might sometimes be all that is needed to assist you in growing your business or even moving forward in your life. If your firm is not expanding, you must first acknowledge and accept responsibility for your faults before developing a new strategy to prevent repeating the same mistakes. It is necessary to prepare for the future, even if your company is expanding, in order to ensure that it continues to develop. Nothing more than a comprehensive analysis of your business data and activities will be asked of you. Data Analysis approaches are listed below in no particular order and are the most often employed in the field of data analysis.

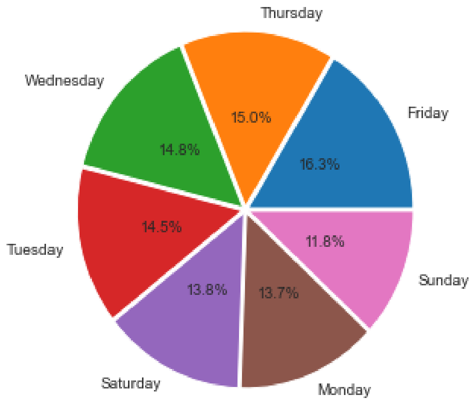
- Exploratory Data Analysis
- Descriptive Analysis
- Text Analysis
- Predictive Analysis

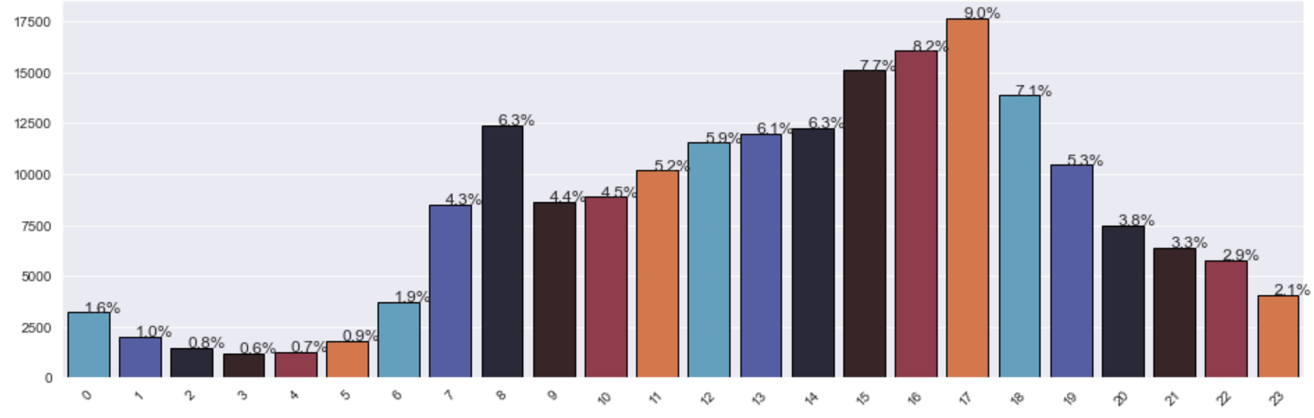
So, in this project I took Exploratory Data Analysis and Predictive Analysis into considerations and did following analysis.

a) Are there significant hours of the day, and days of the week, on which accidents occur?

For this I extracted hour from the time feature and made following two graphs.



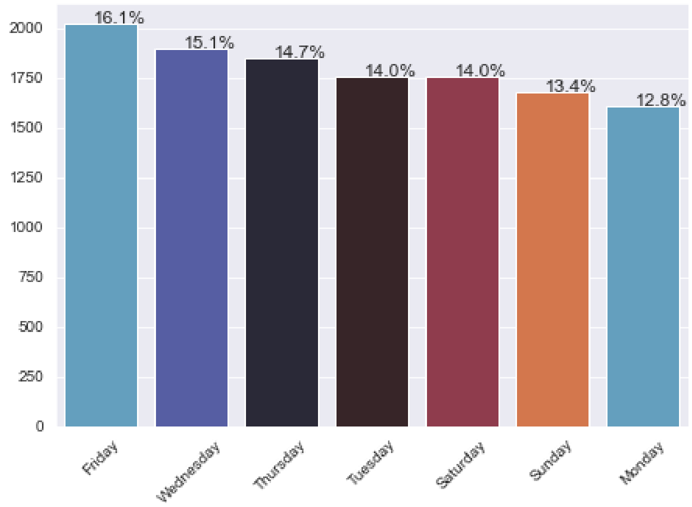


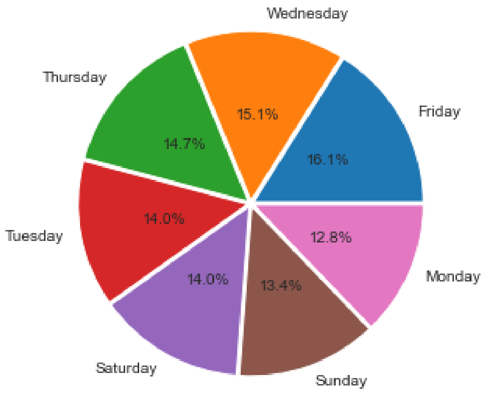


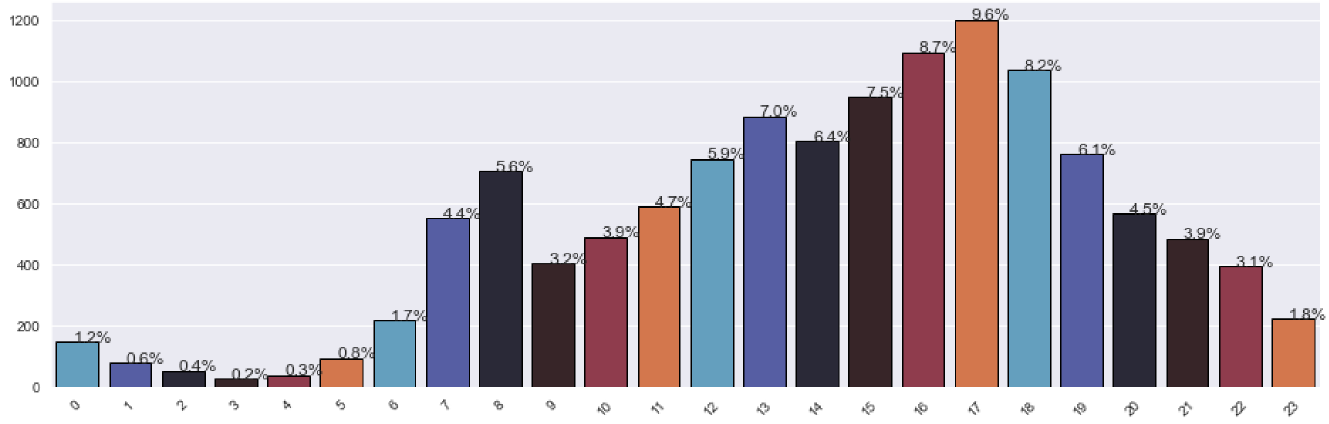
From the first graph we can see that the accidents are more on Friday than on any other day.
From the second graph we can see that the accidents are more in hour 8 and in hour 16 and 17.

b) For motorbikes, are there significant hours of the day, and days of the week, on which accidents occur?

For this problem first of all I took the Vehicle Type feature and then filtered out the data only for motor bikes and for that I took help from Variable Lookup file. On the filtered data I made following plots.





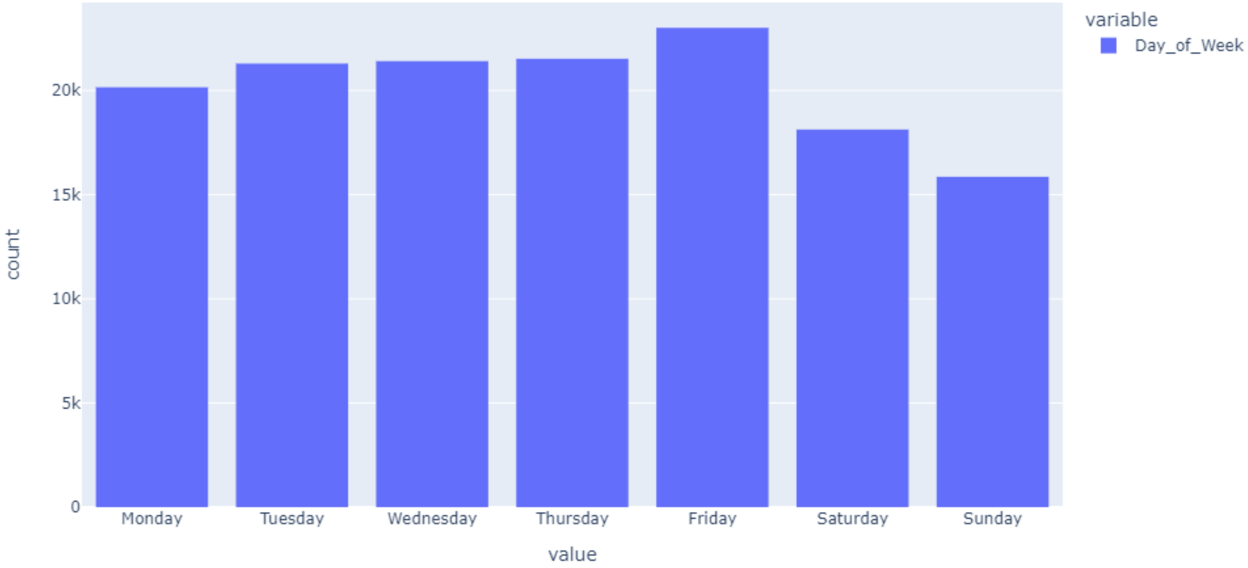


From these two graphs I observed that there are no specific days and hours for the bikers on which accident occurs.

c) **For pedestrians involved in accidents, are there significant hours of the day, and days of the week, on which they are more likely to be involved?**

There were two type of pedestrians one was Human and second was some other objects like they can be animal, or some other living or non-living objects. So, I filtered out the data for both of them separately and analyzed them. I came to know that there was no such impact in the time and days of week that were likely to involved. The graphs spikes were same as they were in the original graphs.

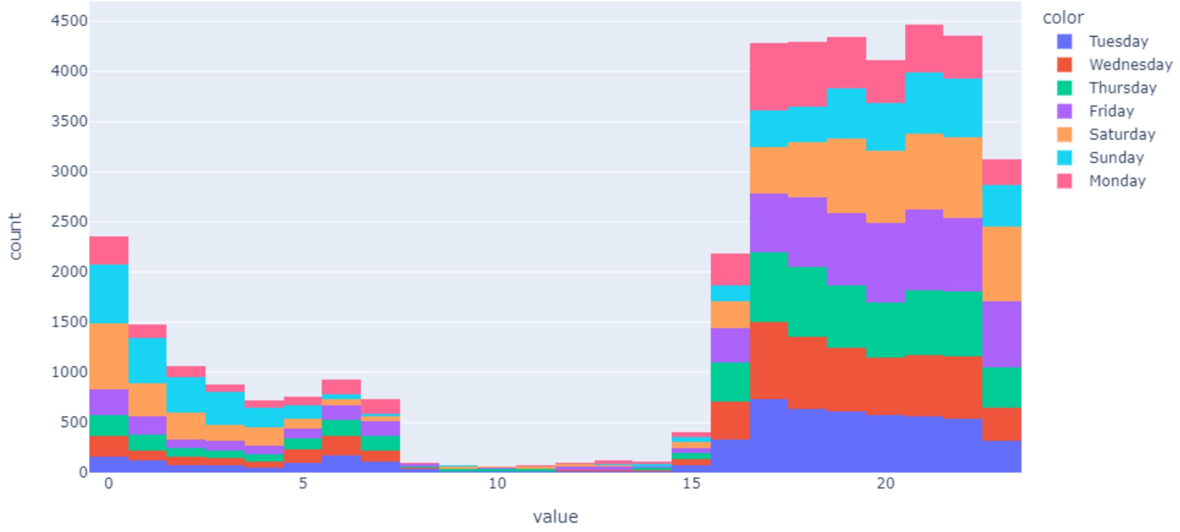
d) **What impact, if any, does daylight savings have on road traffic accidents in the week after it starts and stops?**



From this graph we can see that there is such impact on the number of accidents. Again, accidents are on peak on Friday.

e) What impact, if any, does sunrise and sunset times have on road traffic accidents?

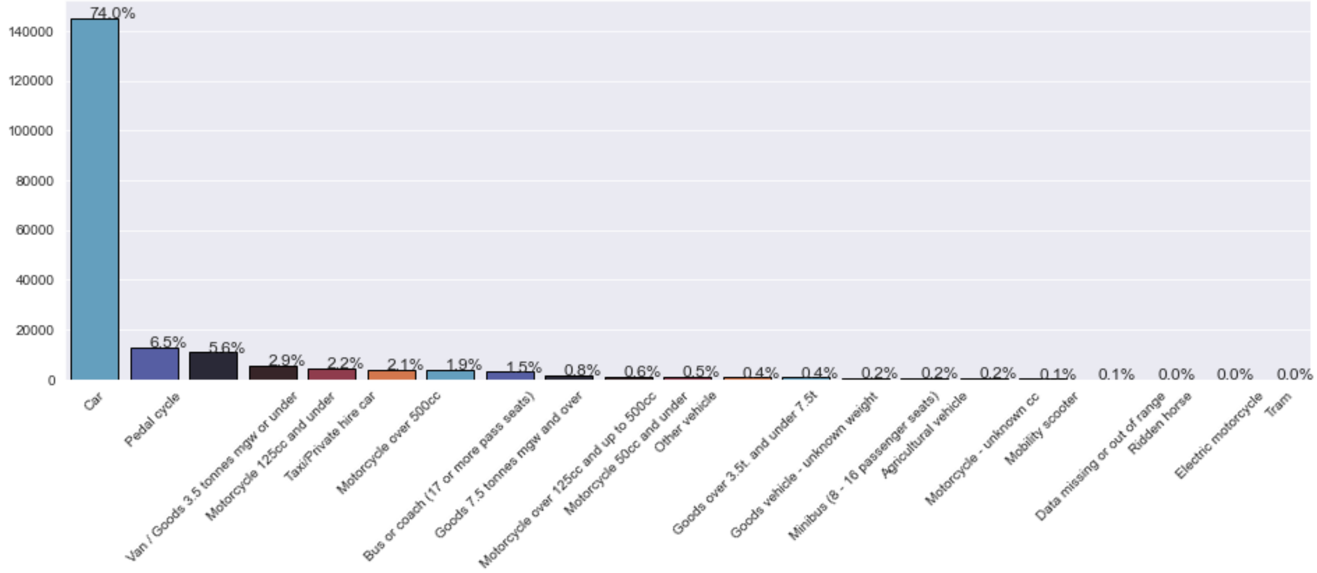
First of all, I filtered out the data and then made following plots.

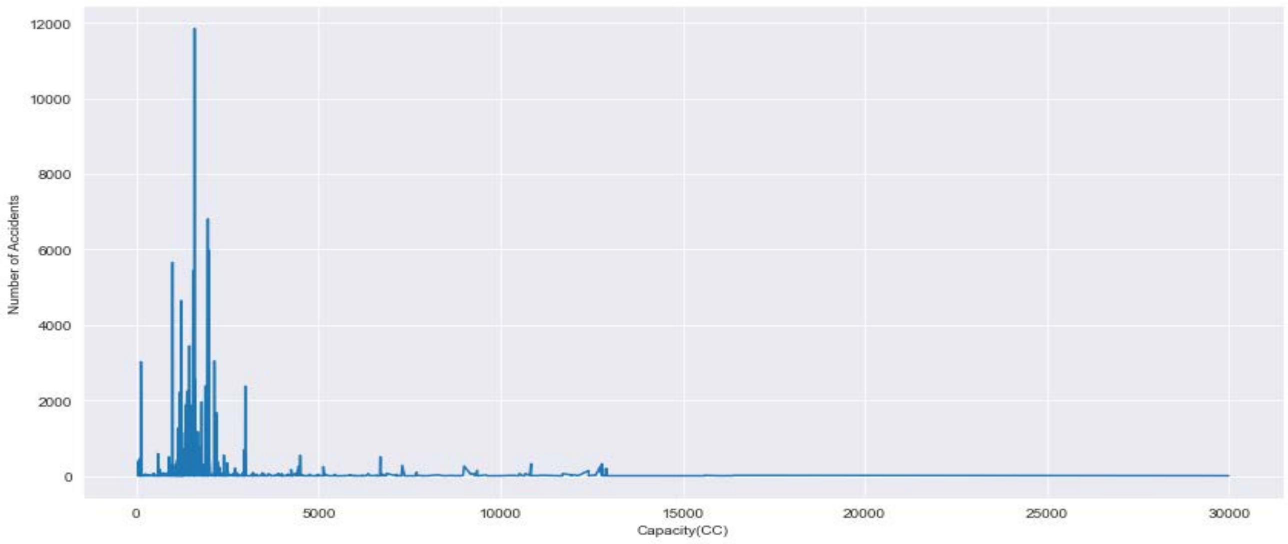


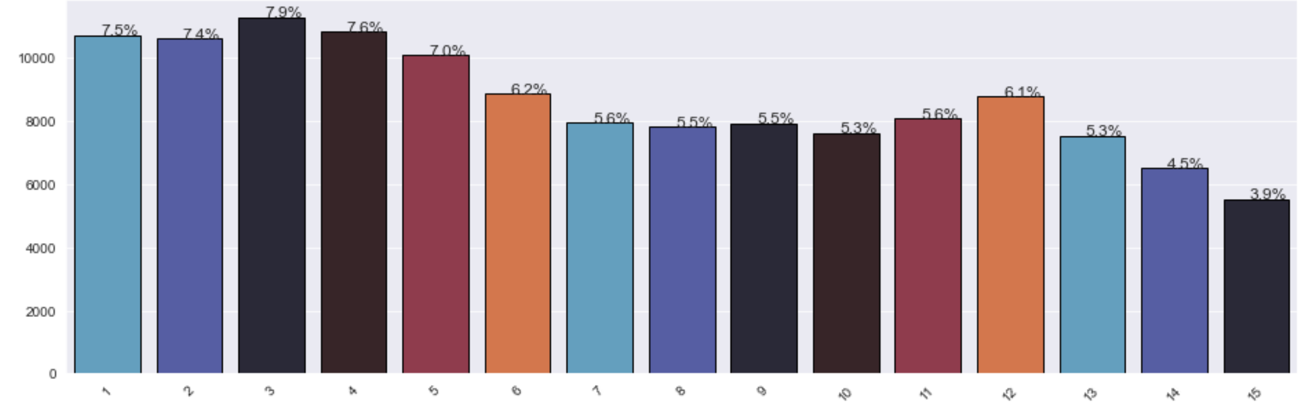
From this graph we can see that Yes, the Sun rise have an impact on the ratio of accident specially at the end of day means in the evening. There are a smaller number of accidents when the daylight is present like in hours from 7 to 15. But except this interval there are a greater number of accidents.

f) Are there particular types of vehicles (engine capacity, age of vehicle, etc.) that are more frequently involved in road traffic accidents?

I made analysis for the Vehicle type, capacity and age of vehicle. For these three graphs are given below.



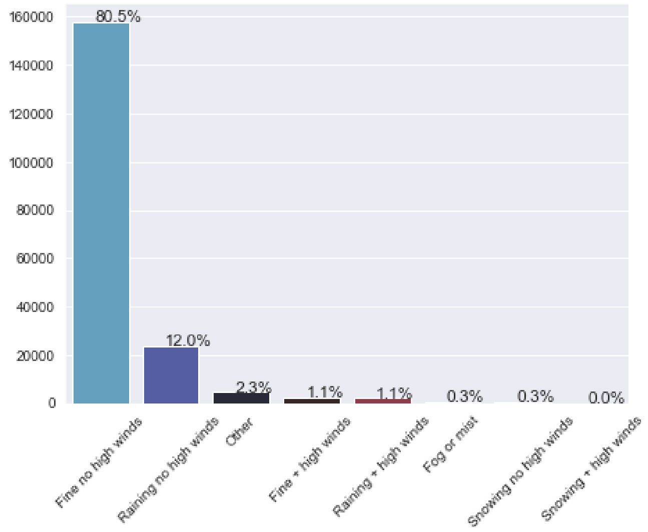


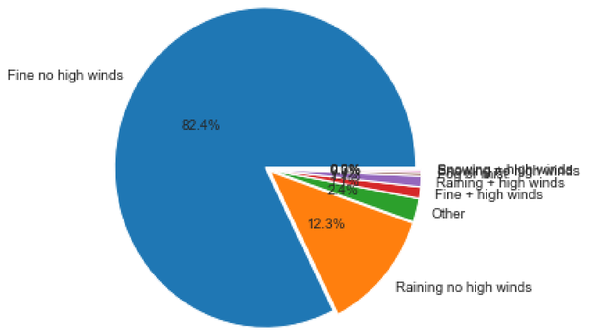


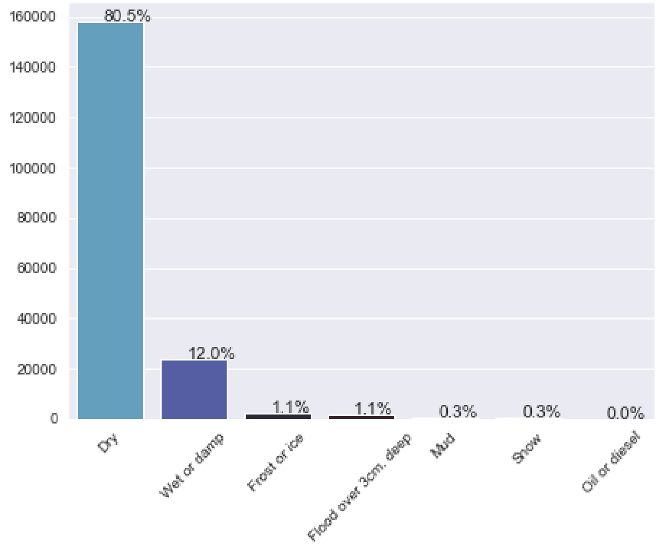
The first graph shows that cars and pedal bicycles are the most frequently engaged in road accidents. We can observe from the second graph that the majority of the accidents occur in cars with engine capacities ranging between one thousand and three thousand horsepower. Vehicles with a lot of power have a lower rate of accidents than others. The third graph shows that vehicles that are 3 and 4 years old are more likely to be involved in an accident.

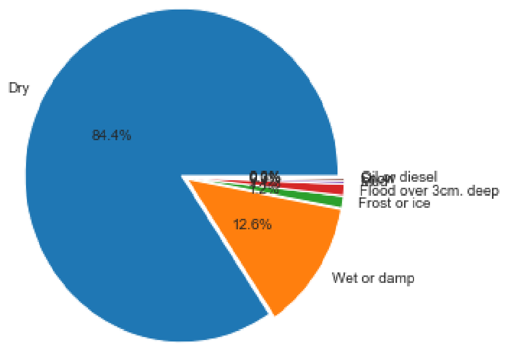
g) Are there particular conditions (weather, geographic location, situations) that generate more road traffic accidents?

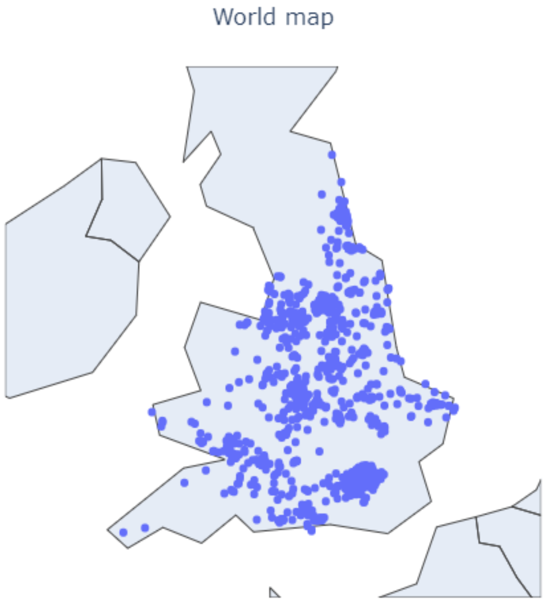
In this part I checked the number of accidents in different conditions like weather and on geographical locations. There are the plots that I made

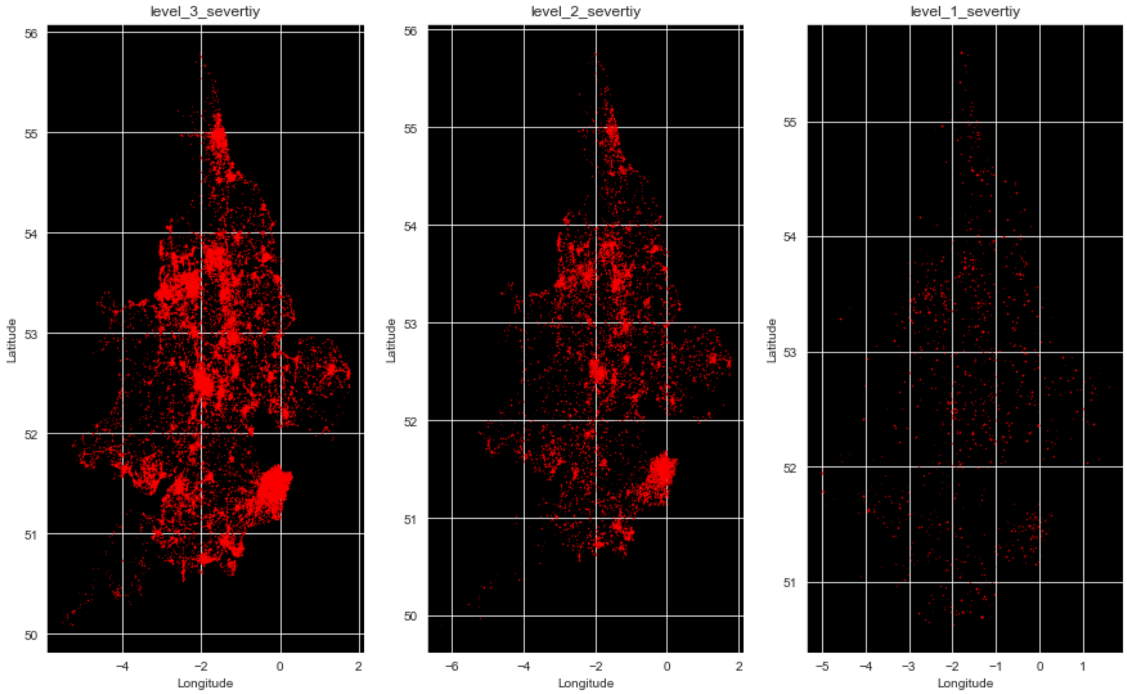












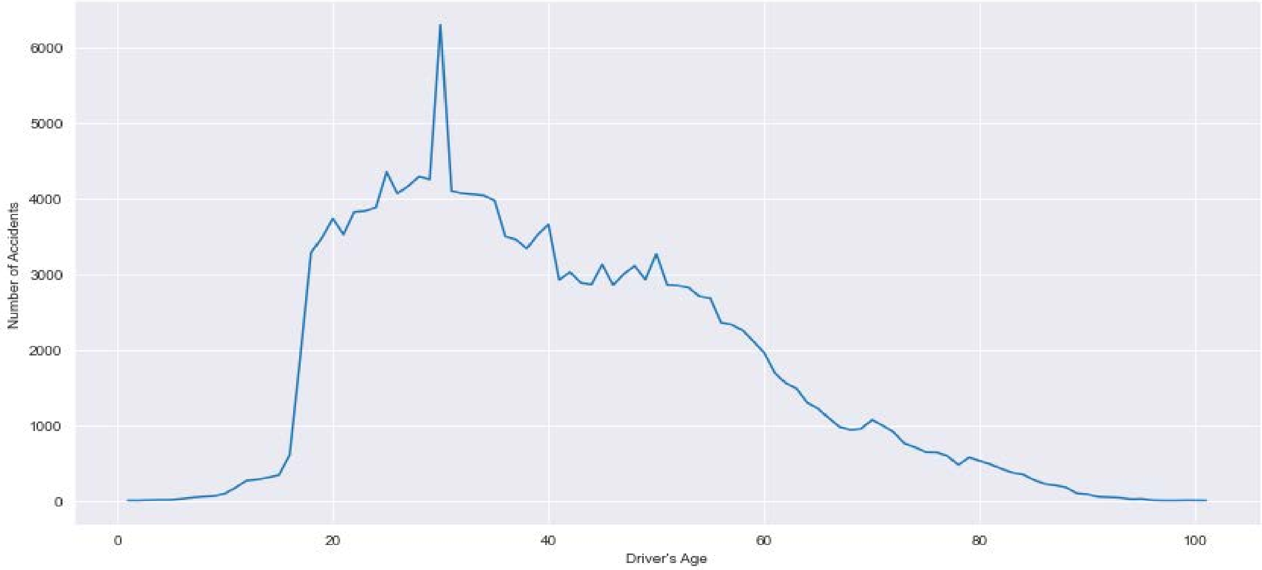
After Fine weather the accidents are high for Raining Environment.

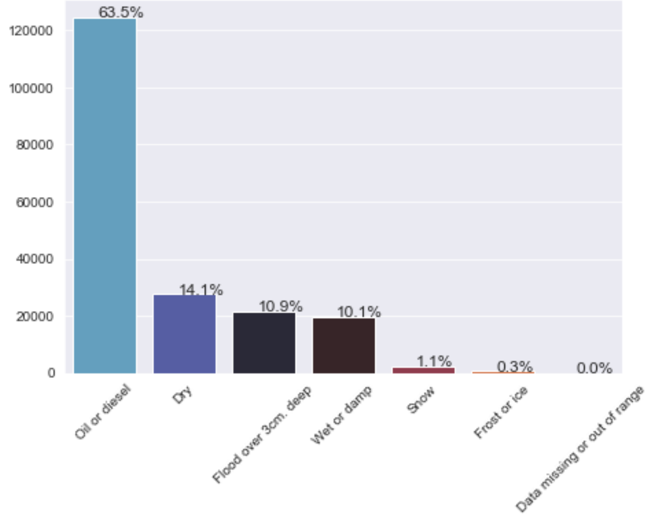
Wet and Damp Road has a greater number of accidents after Dry Road.

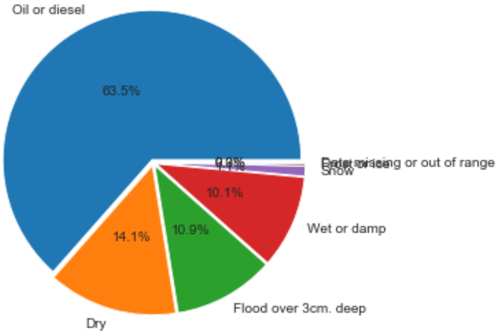
There are some specific areas where accidents are higher like at Latitude=51.4

h) How does driver related variables affect the outcome (e.g., age of the driver, and the purpose of the journey)?

Driver influences a lot in road accidents (G.S. Larue, C. Wullem, 2019).First of all, I filtered the data for only the records in which the driver detail is given. Then I did following analysis.







Drivers with age from 30 to 35 become the cause of more accidents. Along with that most of the accidents happened when the purpose of the driver was Oli or Diesel.

i) **Can we make predictions about when and where accidents will occur, and the severity of the injuries sustained from the data supplied to improve road safety? How well do our models compare to government models?**

In this part we basically have to built some machine learning models that will predict when and where the accident will occur and what will be the severity of that accident. And for prediction I made three models (WHEN, WHERE, SEVERITY). The machine learning algorithm that I selected is Decision Tree but before building the model I applied PCA (Principal Component Analysis) that is used for feature reduction. When extracting information from a high-dimensional space, Principal Component Analysis (PCA) is a linear dimensionality reduction approach that may be used to reduce the dimensionality of the space.

PCA is utilized in exploratory data analysis as well as in the development of prediction models. It is often used for dimensionality reduction, in which each data point is projected onto just the first few main components in order to get lower-dimensional data while retaining as much of the data's variance as feasible, according to the literature.

So, I built three models on reduced data that can be used for predictions. Model was decision tree. Deterministic trees (DTs) are a type of non-parametric supervised learning approach that may be used for classification and regression tasks. To do this, a model that predicts the value of a target variable must be built by learning basic decision rules that may be inferred from the data's characteristics. A tree can be thought of as an approximation to a piecewise constant.

Our model predict that accidents is highly to occur at the time range 17 (5pm ... rush hour as we said) and the zone with latitude: 51.508057 and longitude: -0.153842 (London zone), whereas with severity 3 (slight)

3. Recommendations

Keep in mind to solve big problems like this we should go more deeply in each question and in each point one by one to see what is the cause! And analyze the exact reasons to solve them properly. For example, here we reach a point that from 5pm till 8pm the number of accidents increases and at this time it is the rush hour (people are going back from work school etc...) but we should go further and analyze and detect that even if there is a rush hour (where the probability of accident occur is high) what is the exact reason leading the people to accidents?

Here are some main recommendations:

- Traffic wardens should be in action at the office time as we have seen there are a greater number of accidents at these times.
- There should be some more restrict rules and regulations for Pedestrians.
- There should be separate routes for Pedal cycles because after car they are the one who are the cause of accidents.
- We should focus deeply on Drivers with age 30 to 35, why they are mostly involved in accidents? Driving in a high speed? Without Focusing on road due to music? Etc...
- Some strict actions should be taken where number of accidents are high as we seen such places in the geo graphs, (London)... stricter rules should be applied there.

References

Bazerman, M.H. & Moore, D., 2013. Judgement in Managerial decision-making. Volume 8th.

G.S. Larue, C. Wullem, 2019. A new method for evaluating driver behavior and interventions for passive railway level crossings with pneumatic tubes. p. 150–166.

Koŝcielniak, H. & Puto, A., 2015. BIG DATA in decision-making Processes of Enterprises.