

TRAFFIC ACCIDENTS SEVERITY PREDICTION

**Submitted in a partial fulfillment for the award of**

**Post Graduate Diploma in Big Data Analytics from C-DAC ACTS(Pune)**

Guided by:

Mr. Hari

Presented by:

Mr. Ankur Saini PRN: 190840125003

Mr. P N V Sai Ram Teja PRN: 190840125027

Mr. Vijay Pratap Pandey PRN: 190840125027

**Centre of Development of Advanced Computing(C-DAC), Pune**

****

**CERTIFICATE**

**TO WHOMSOEVER IT MAY CONCERN**

This is to certify that

Mr. Ankur Saini

Mr. P N V Sai Ram Teja

Mr. Vijay Pratap Pandey

Have Successfully completed their project on

TRAFFIC ACCIDENTS SEVERITY

PREDICTION

Under the guidance of Mr. Hari.

Project Guide: Project Supervisor

HOD ACTS

**ACKNOWLEDGEMENT**

This project “Traffic Accident Severity Prediction” was a great learning experience for us and we are submitting this work to Advanced Computing Training School (CDAC ACTS).

We all are very glad to mention the name of Mr.*Hari* for his valuable guidance to work on this project. His guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to Ms. Richa P.R. (Manager (ACTS training Centre), C-DAC), for her guidance and support whenever necessary while doing this course Post Graduate Diploma in *BIG DATA ANALYTICS(DBDA)*through C-DAC ACTS, Pune.

Our most heartfelt thanks go to *Ms. Shilpi Shalini* (Course Coordinator, PG-*DBDA*) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC ACTS, Pune.

**From:**

Mr. Ankur Saini (190840125003)

Mr. P N V Sai Ram Teja (190840125027)

Mr. Vijay Pratap Pandey (190840125053)

**TABLE OF CONTENTS**

**1.Abstract**

**2. Project Motivation**

**3.Introduction and Overview of Project**

**4.Data understanding and tools needed**

4.1 Type of data

4.2 Software needed to process data

**5.Data loading and pre-processing**

5.1Load datasets

5.2Concatenate all datasets into one

**6.Modelling**

**7.Evaluation**

**8.Business understanding and Analysis Question**

**9.System Requirements**

**10. Future Works**

**11. Conclusion**

**12. Bibliography**

|  |  |
| --- | --- |
| |  | | --- | | **1. Abstract** | |

Safety and accident issues are considered one of most important problem of the world. Accident severity analysis is done to identifying factors of accidents which are important for estimating accident cost, determining safety strategy and increasing road safety. Purpose of this study is to identify the factors which are more significant than other factors. The study collected (Accident Information dataset) from the UK data service over the year 2005 to 2015. Two separate analysis have been done for roadways and for junctions of Great Britain.

Depending on characteristics of data and outcome variables, different statistical models have been used in previous researches. To account the ordinal nature of response variables of dataset, ordered probit model has been used in this study. Nine explanatory variables are selected from a wide range of variables to use in analysis. These are: type of vehicle, gender of driver, age of driver, time of day, day of week, season of year, speed limit, light condition and weather condition. First, a preliminary analysis has been conducted by frequency distribution.

Finally, a multivariate analysis has been done by developing models for roadways and for junctions. The two road details, roadways and junctions, provide the similar result from frequency distribution analysis and model output. Eight variables provide significant factors for severe accidents. One variable shows no difference in probability on severity. Detected significant factors are: motor cycle, male driver, elderly driver age above 60 years, midnight and early morning, non-built-up areas with speed limit >30 mph, darkness without lighting, fine weather. Finally, recommendations have been provided to account the significant factors which will help to improve road safety. Keywords: Injury severity, significant factors, STATS19 data, probability model, traffic accident, casualty, slight injury, fatal and serious injury, road safety

**2. Project Motivation:**

The aim of this project is to use data science methodology and machine learning to gain an understanding on the problem at hand, and develop insights and prevention mechanisms for Traffic Accidents and Road Safety.

This project will use U.K Road Safety Data from (2005–2017). The dataset is published by Department of Transports, under Open Government Licence. The data consists of detailed road safety data about the circumstances of personal injury road accidents, the types of vehicles involved and the consequential casualties.

First off, we want to study and understand the nature of car accidents, and how it has changed throughout the years? how road safety has developed over time? where do accidents happen? what are the main causes of car crashes? whether we can predict the severity of accidents and prevent them before they happen?

This project will follow CRISP-DM methodology which provides a structed process to approach data science problems, it's constructed of 6 steps:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modelling
5. Evaluation

**3. Introductionand Overview of Project:**

Roads are very essential in our everyday lives for driving, riding, walking, travelling for obtaininggoods and services. Unfortunately, accidents onroads involved with killed and serious injury (KSI)bring unmeasurable sufferings to human and social

life. Worldwide more than one million people are killedin road accidents every year. As per World HealthOrganization (WHO), the estimated road crash death is 1.24 million and 2.4 million injured due to traffic crashes each year where the age group 15-29years are more severe and males are more killed orinjured.

91% of the world fatalities in road crashesoccur in low-income and middle-income countries, though they avail only half of the world’s vehicles. The projected annual road traffic fatalities will be 1.9 million by 2020, if no prevention would betaken. (WHO 2015)As per World Health Organization (WHO), roadaccident is going to be the 3rdleading cause of“Disability Adjusted Life Years” (DALY) by 2020,

which was at 9thposition in 1990.

From 2014WHO nominated world health day as a ‘RoadSafety Day’. (WHO 2004 cited in Road Safety as aHealth Problem n.d.). Fatalities due to road

accidents are getting higher than malaria, HIV/AIDS and tuberculosis as an epidemic, providing influence on global public health and contribute to the causes of poverty. As per UN report in 2009, traffic fatalities will increase from

1.3 million to more than 1.9 million over coming decade.

The Global Road Safety Commission suggested some actions to be taken to halt the rising in road injury by achieving year byyear reductions. The world can prevent 5 milliondeaths and 50 million serious injuries by the end of next decade.through implementing road safety measures at global, regional and national levels. (Commission for Global Road Safety 2009)Since 1990, number and severity of road accidentsin Great Britain has decreased. (House ofCommons 2013).

As per Daft (Department forTransport) report, 1,775 road deaths are reported in

2014 which is increased by 4%compared with2013. Likewise, serious injury increased by 5%against of 2.4% increase of vehicle traffic level onroads.

**4. Data understanding and tools needed**

In this project we use several python libraries and packages necessary for this project, we will be using:

1.Pandas: for loading csv data set as panda’s data frame for easy analysis

2. NumPy for basic analysis

3.Sklearn: for using various machine leaning algorithm

4.Plotly: for plotting

5.Sypder and Jupyter notebook: for running the machine learning code

6.Tableau software: for data visualization purpose

**5. Data loading and pre-processing:**

Data pre-processing is a process of cleaning the raw data i.e. the data is collected in the real world and is converted to a clean data set. In other words, whenever the data is gathered from different sources it is collected in a raw format and this data isn’t feasible for the analysis.  
Therefore, certain steps are executed to convert the data into a small clean data set, this part of the process is called as data pre-processing.

State of data in our project consist

1. Missing data: Missing data can be found when it is not continuously created or due to technical issues in the application.

2. Noisy data: This type of data is also called outliners; this can occur due to human errors (human manually gathering the data) or some technical problem of the device at the time of collection of data.

3. Inconsistent data: This type of data might be collected due to human errors (mistakes with the name or values) or duplication of data.

Which further categorised into Three Types of Data

1. Numeric e.g. speed limit

2. Categorical e.g. road type

3. Nominal e.g. Day\_of\_Week

|  |  |
| --- | --- |
| |  | | --- | |  | |

**How we did data pre-processing**

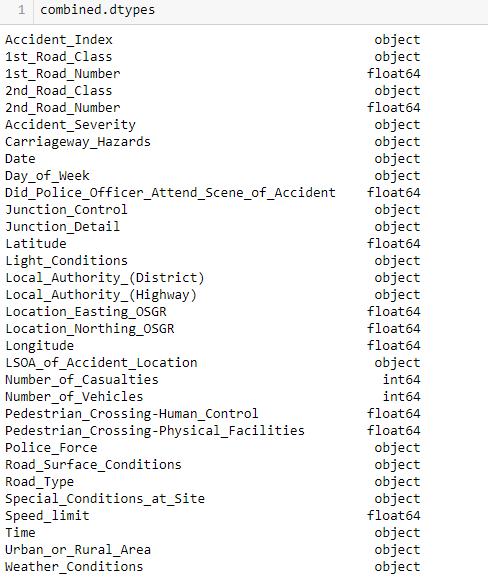
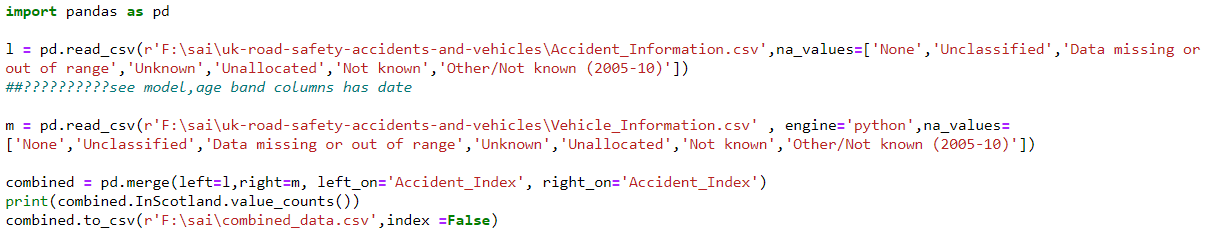
These are some of the basic pre — processing techniques that we used to convert raw data.

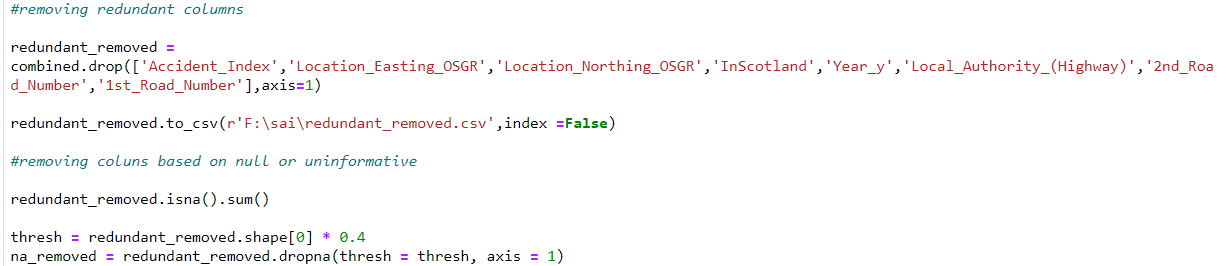
1. **Conversion of data**: As we know that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.

2**. Ignoring the missing values**: Whenever we encounter missing data in the data set then we can remove the row or column of data depending on our need. This method is known to be efficient but it shouldn’t be performed if there are a lot of missing values in the dataset.

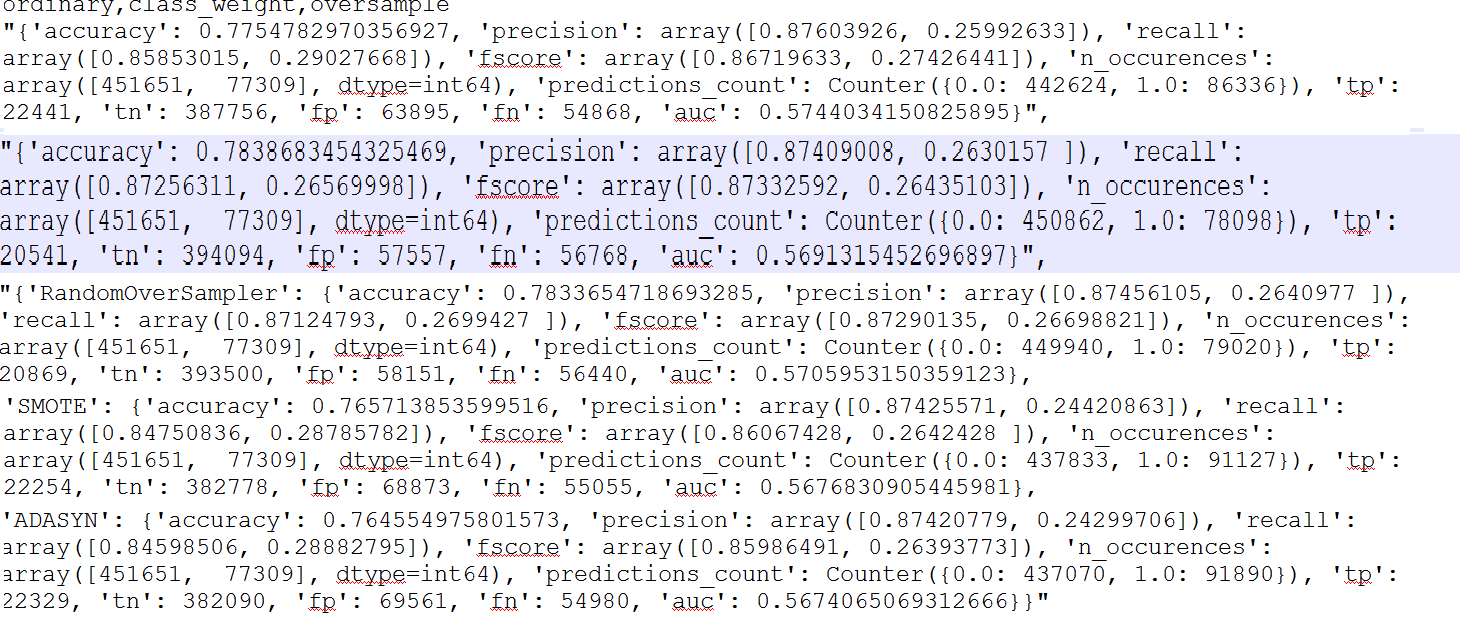
3. **Filling the missing values**: Whenever we encounter missing data in the data set then we can fill the missing data manually, most commonly the mean, median or highest frequency value is used.

4**. Machine learning**: If we have some missing data then we can predict what data shall be present at the empty position by using the existing data.

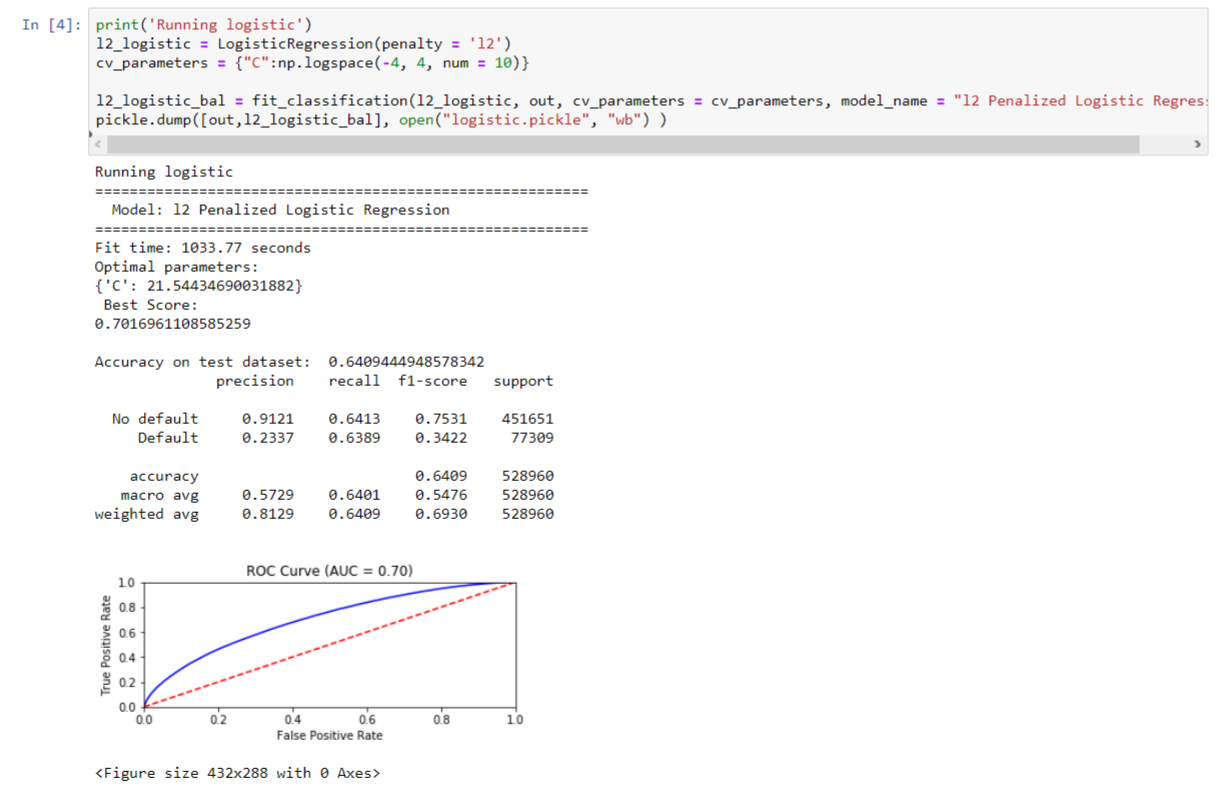
****

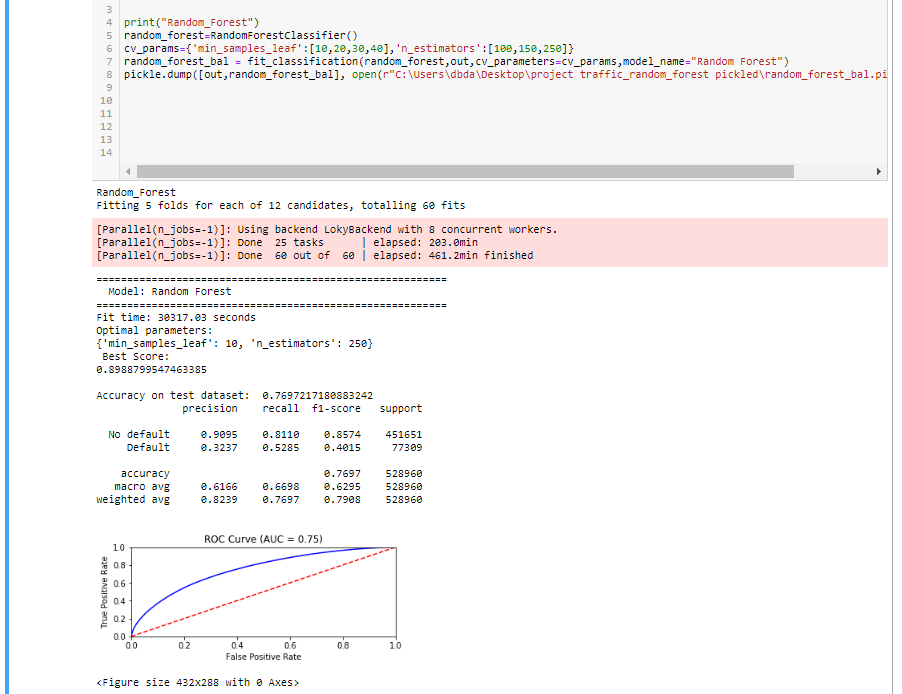
****

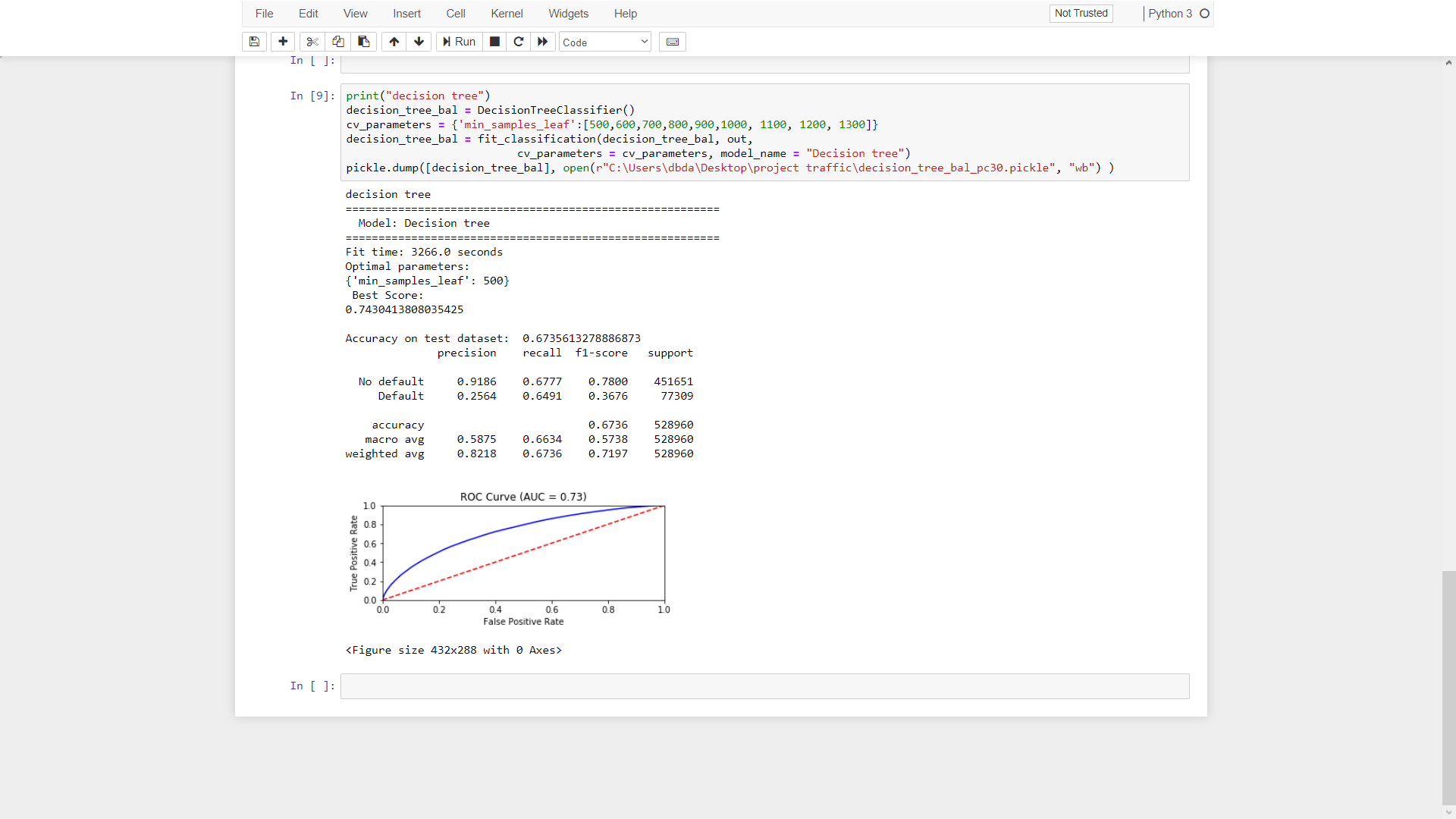
****

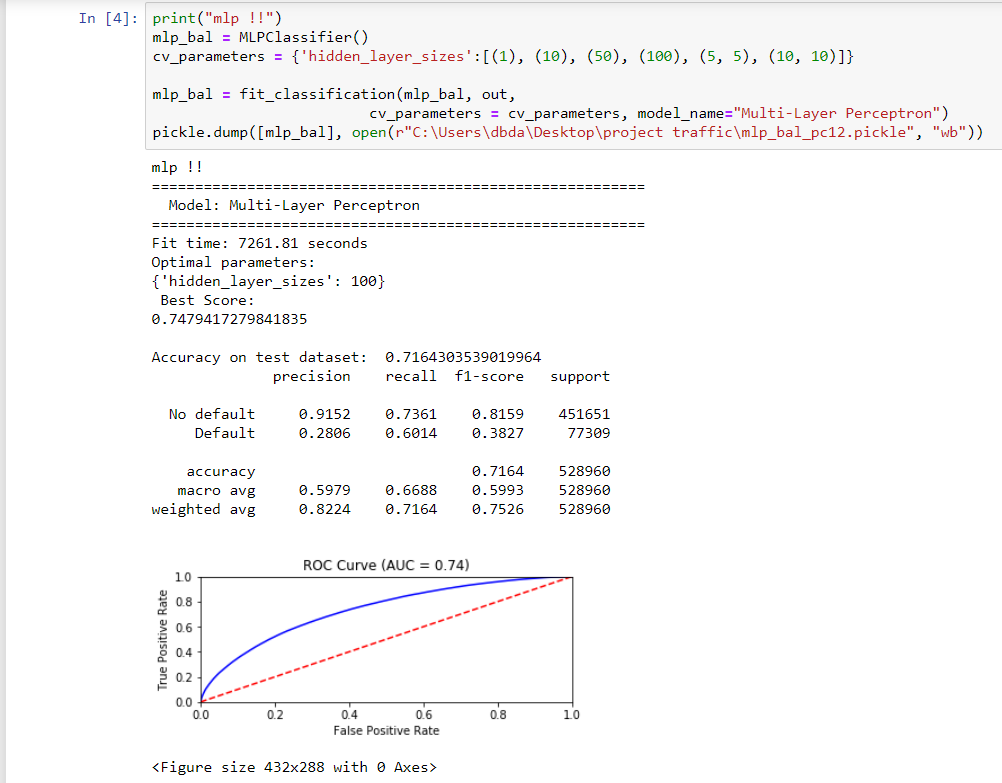
****

**6. Modelling**

****

****

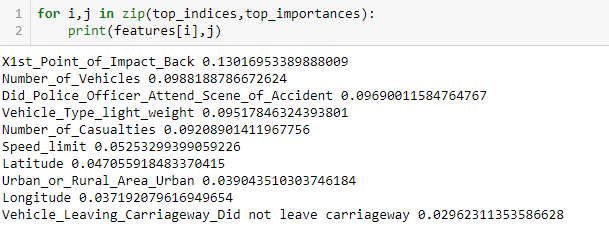
****

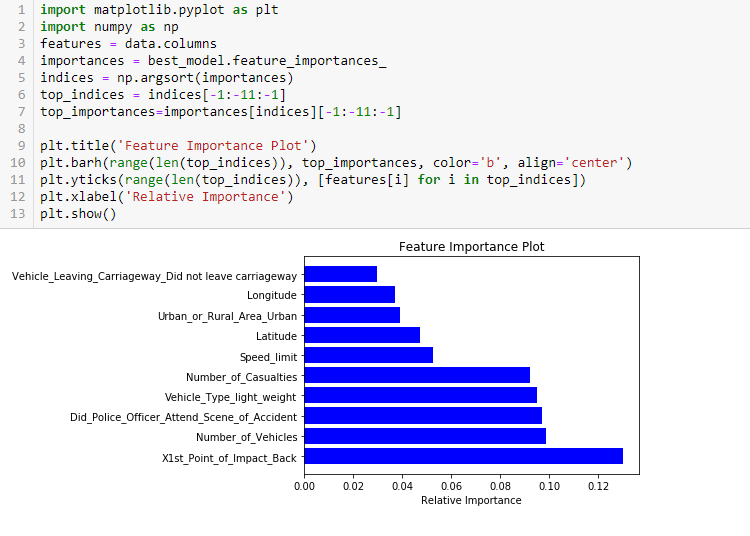
****

**7. Evaluation**

|  |  |  |
| --- | --- | --- |
| **MODEL NAME** | **ACCURACY(Kfold) %** | **ROC** |
| **Penalized logistic regression** | **70.16** | **0.7** |
| **Decision Tree** | **74.3** | **0.72** |
| **MLP** | **74.79** | **0.74** |
| **Random Forest** | **89.88** | **0.75** |

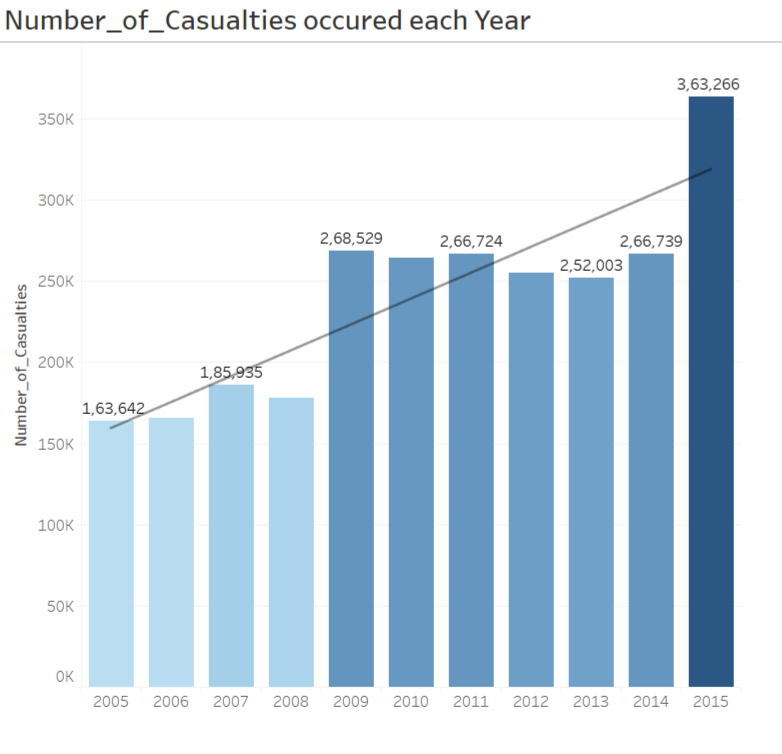


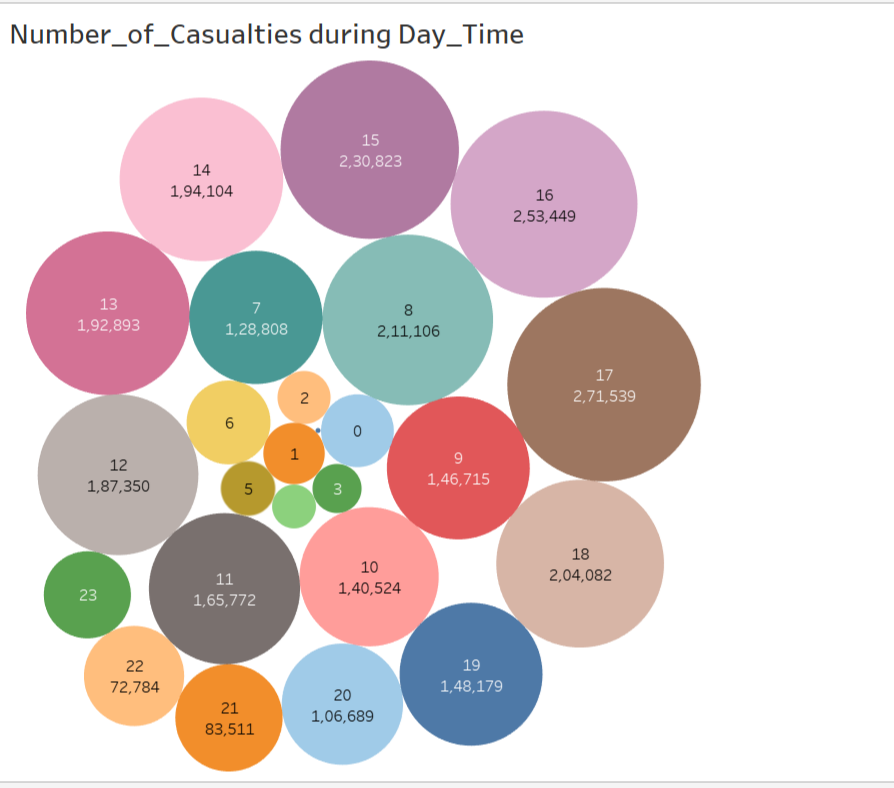




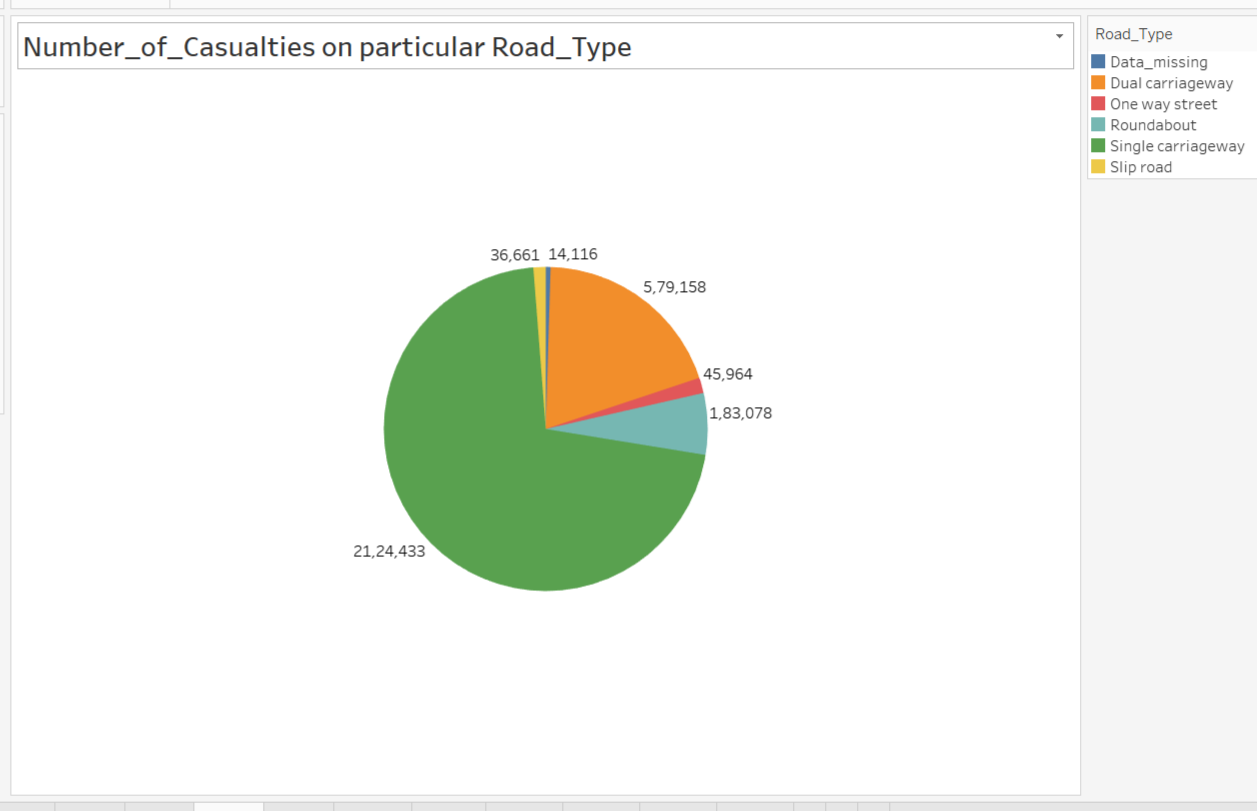
**8. Business understanding and Analysis Question**

To better guide us through the analysis, we formulated the problem into the following set of questions, so we can explore it at greater depth:

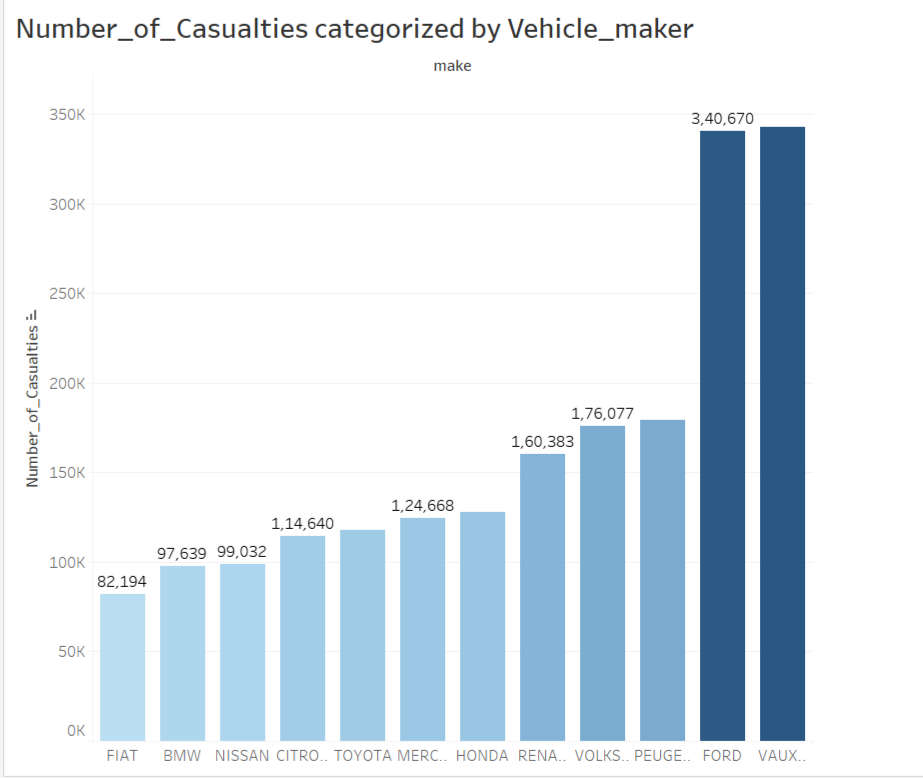
1. What is the severity of accidents over the last few years?
2. When do accidents usually happen?



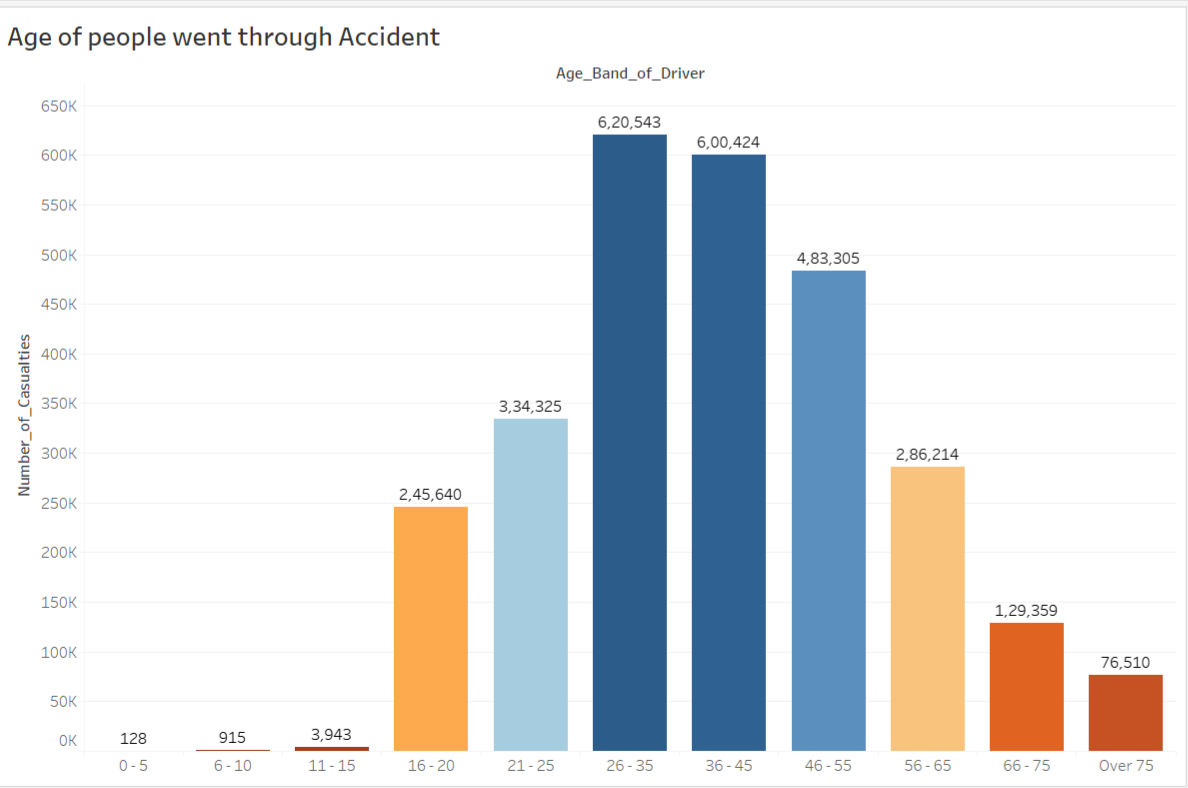
3.Number of causalities on a particular road type?



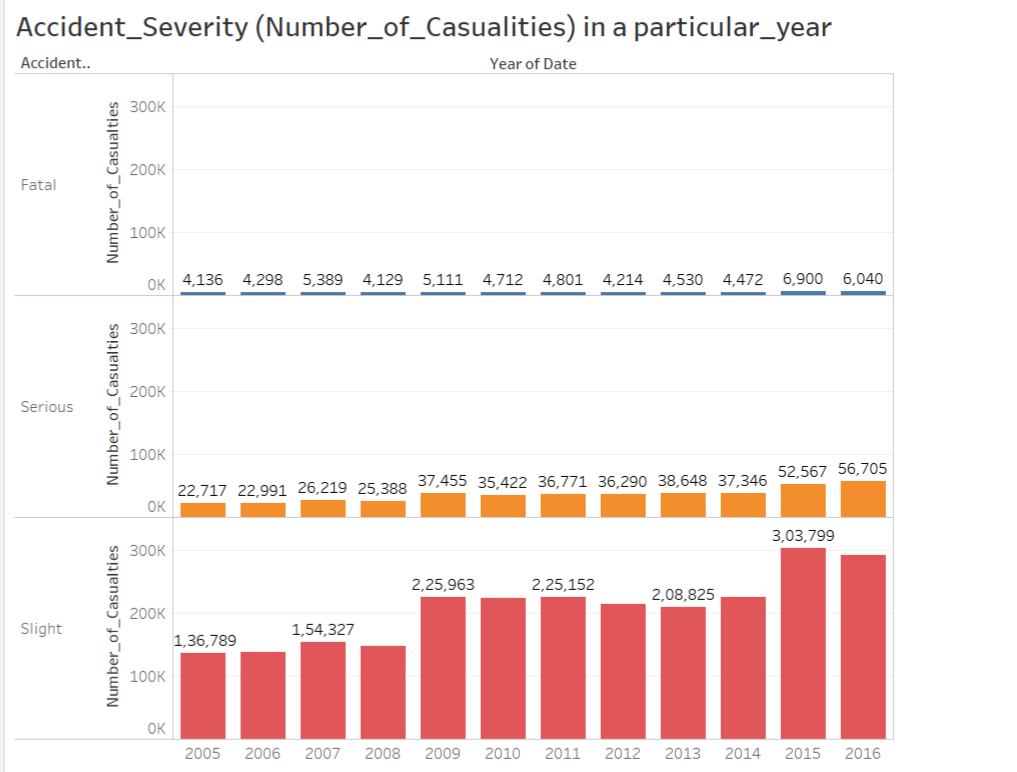
4.Number of causalities categorized according to the Vehicle Maker?



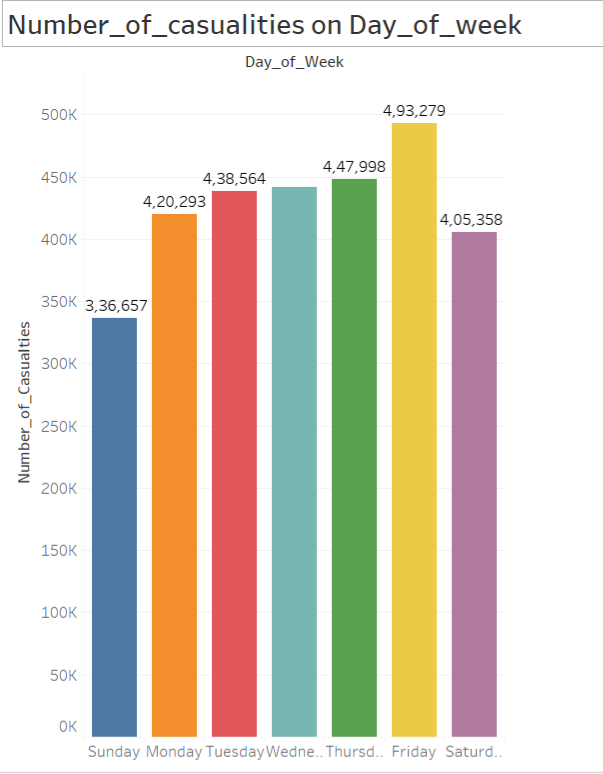
5. What is the age distribution of drivers involved in the accidents?



6.What are the no of causalities and accident severity in a particular year?



1. Number of causalities on a given day in on week?



**9. System Requirements**

16Gb –Ram

1TB ----Hard drive

8 core—processor

**10. Future Works**

In this analysis we explored the problem from different prospective, yet leaving so much to uncover. We can summarize our findings as follows:

1. From the day-hour heat maps, we can suggest to increasing response time during rush-hours, or construct roads to divert traffic from congested area.
2. Investigate high density points in map relocation analysis to discover construction needs.
3. From model feature importance we can recognize dangerous junctions which cause specific maneuver.
4. We explored the conditions of Vehicle and Casualties such as (Type of vehicle, Age of vehicles, Pedestrian movements and location when they got in accidents, etc..), these insights are useful for policy makers to understand how accidents happen and provide solutions to limit the cause.

Finally, I really appreciate the UK government efforts to provide this open data in very organized and well-documented format. I would like to explore similar dataset from my hometown!

Without making changes, road traffic injuries are predicted to become the fifth leading cause of death by 2030*.*

11**. Conclusion**

Some patterns were found in predicting the accidents and hence the according measures have to be taken by the authorities in rural areas and point of impact is the major factor in impacting the severity of the accident**.**

**12. Bibliography**

[**https://www.gov.uk/government/publications/road-accidents-and-safety-statistics-guidance**](https://www.gov.uk/government/publications/road-accidents-and-safety-statistics-guidance)

[**https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\_data/file/743853/reported-road-casualties-gb-notes-definitions.pdf**](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/743853/reported-road-casualties-gb-notes-definitions.pdf)

[**https://datamillnorth.org/dataset/road-traffic-accidents**](https://datamillnorth.org/dataset/road-traffic-accidents)