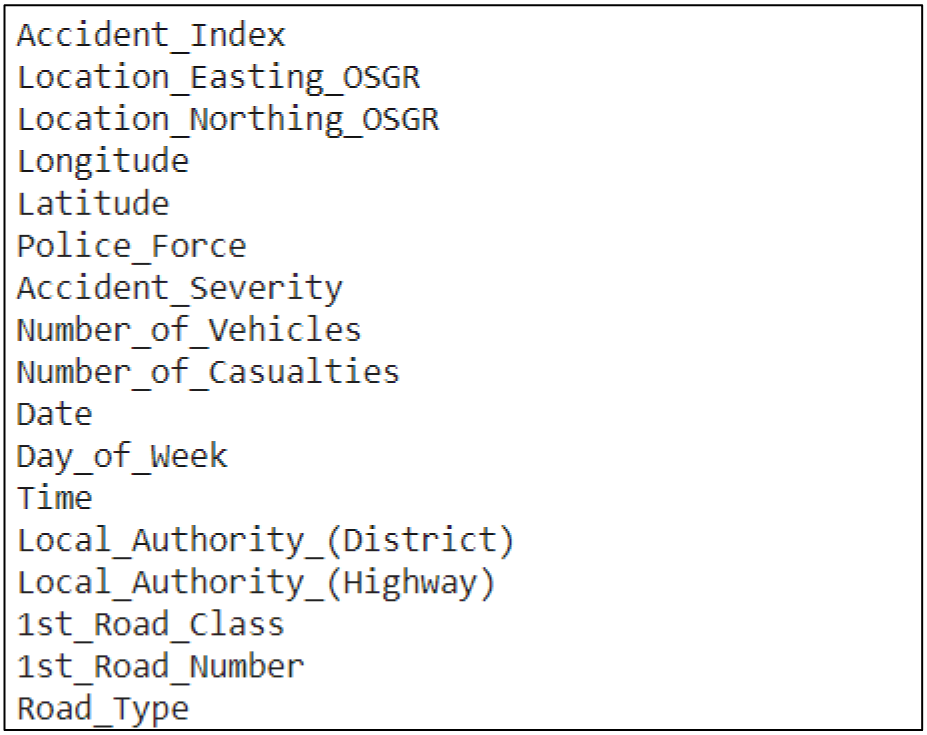
Data Description

The Open dataset that I will be using for this exercise is from UK Department of Transport from Kaggle. I have chosen this because it is an amassed traffic data for a total of six years with many factoring columns which is always good. The other weighing factor was also because the same dataset has been used by Data scientists before for different activities, as it has all necessary properties enabling us to model using it.

The shape of the dataset at the start was 1504150 rows and 33 columns, that is big enough for our needs. As we will be working on predictive analysis of the severity of the accidents, we need to know if the dataset is balanced between all categories. It turns out its not, but we can apply some methodologies to overcome this setback[2].

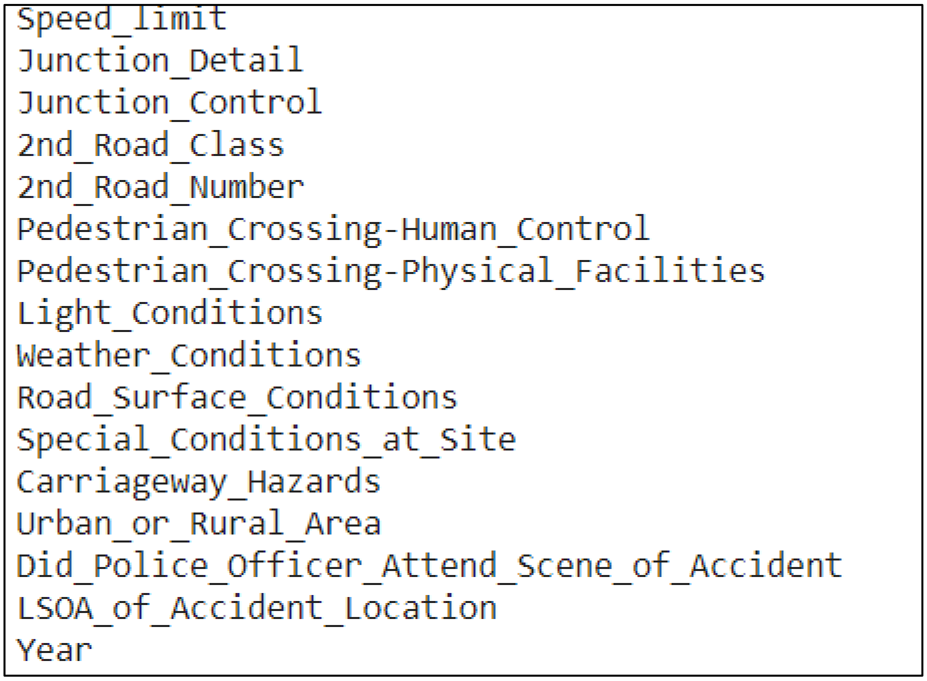
There are 33 columns to start with, but we would not be using all of those as they might not be significant enough. We will be deciding that on the basis on some statistical and distributive aspects of the data which will be discussed in detail in later sections of the report.

Also, it is observed that the data has some Null/Unknown values which will be treated or removed based on what is more fitting for us. The following is the list of columns we have initially in our dataset,



```
Accident_Index
Location_Easting_OSGR
Location_Northing_OSGR
Longitude
Latitude
Police_Force
Accident_Severity
Number_of_Vehicles
Number_of_Casualties
Date
Day_of_Week
Time
Local_Authority_(District)
Local_Authority_(Highway)
1st_Road_Class
1st_Road_Number
Road_Type
```

[2] To be discussed in the methodology section



```
Speed_limit
Junction_Detail
Junction_Control
2nd_Road_Class
2nd_Road_Number
Pedestrian_Crossing-Human_Control
Pedestrian_Crossing-Physical_Facilities
Light_Conditions
Weather_Conditions
Road_Surface_Conditions
Special_Conditions_at_Site
Carriageway_Hazards
Urban_or_Rural_Area
Did_Police_Officer_Attend_Scene_of_Accident
LSOA_of_Accident_Location
Year
```

Based on the columns and the overall size of the dataset, it looks great like a great dictionary for our use-case needs right now. The naming convention of the columns are all self-explanatory which makes it easier to work with it and not having to refer to the definition file of the dataset.