## 2. Data

### i. Data Source:

UK police forces collect data on every vehicle collision in the ukon a form called Stats19. Data from this form ends up at the DfT and is published at here and the data csv files found in kaggle datasets found here. This Database collected between (2005 to 2015).

### ii. Data Cleaning:

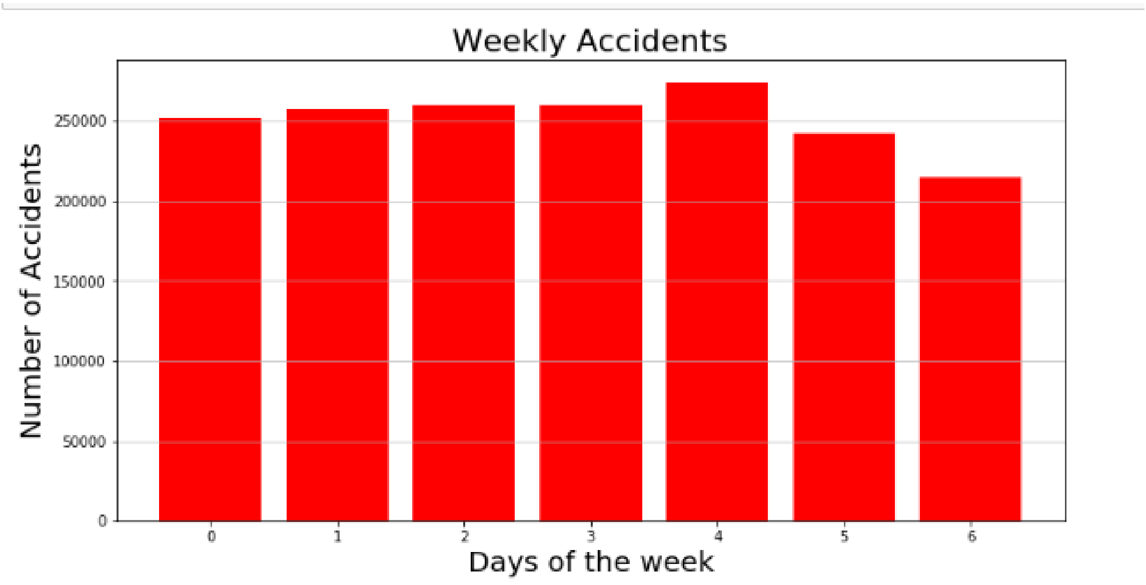
This data consists of three files that are accidents, causalities, and vehicles. However, we have one more file which is general information about the traffic count for the year 2000 to 2015. We can use general traffic information data for the machine learning part. There are two types of missing values '-1' and 'Nan'. So, We will solve missing data problem by investigating each column with total missing values. We will not be imputing any mean or median value since the dataset is big enough to perform analysis.

### iii. Data Visualization:

I will try here to describe the data that I have using visual figures to walk you through understanding the data and giving better view on what we are dealing with.

The first thing we can do is to find out about accidents time to get intuition and some driver's age who are involved in the accident.

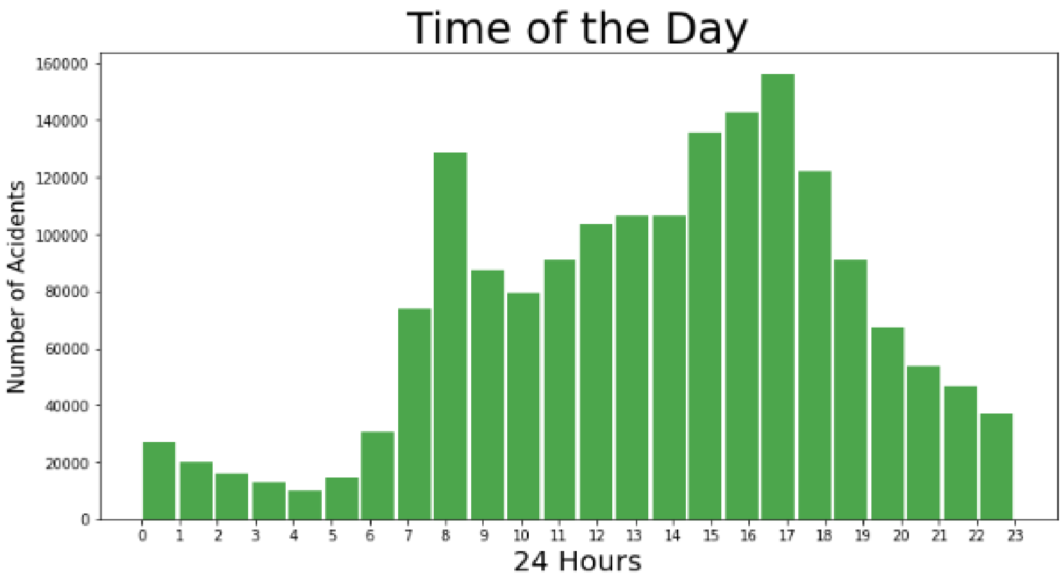
1- We can find out the number of accidents on the days of a week.



0-> Monday , 1-> Tuesday , 2-> Wednesday , 3-> Thursday , 4-> Friday , 5-> Saturday , 6->Sunday

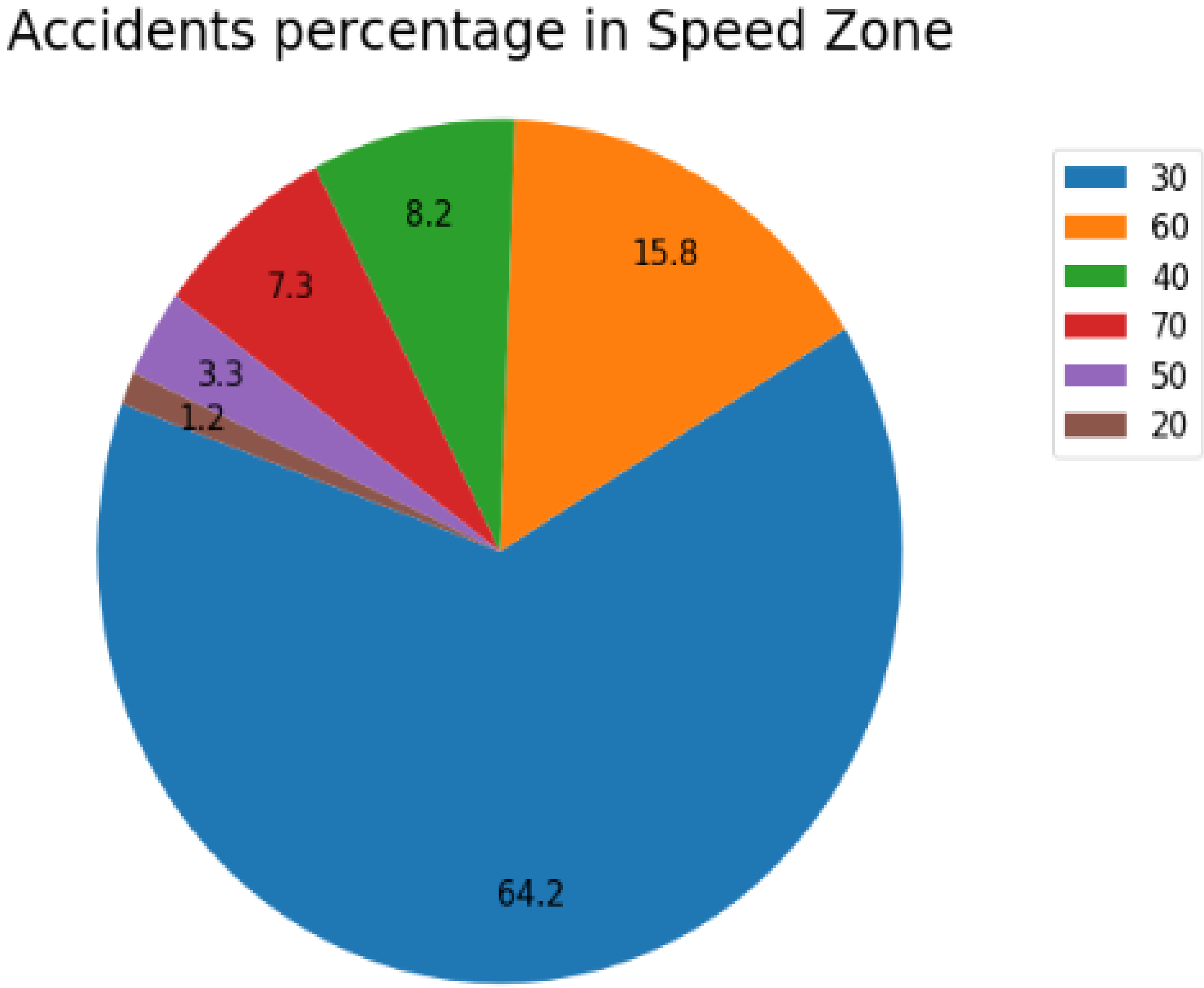
We can see that the most day accidents happened in UK is on Friday. Also, we have to keep in mind that accidents numbers could be depending on traffic amount on particular day.

<u>2- We can find out about the accidents number using hours of the day.</u>



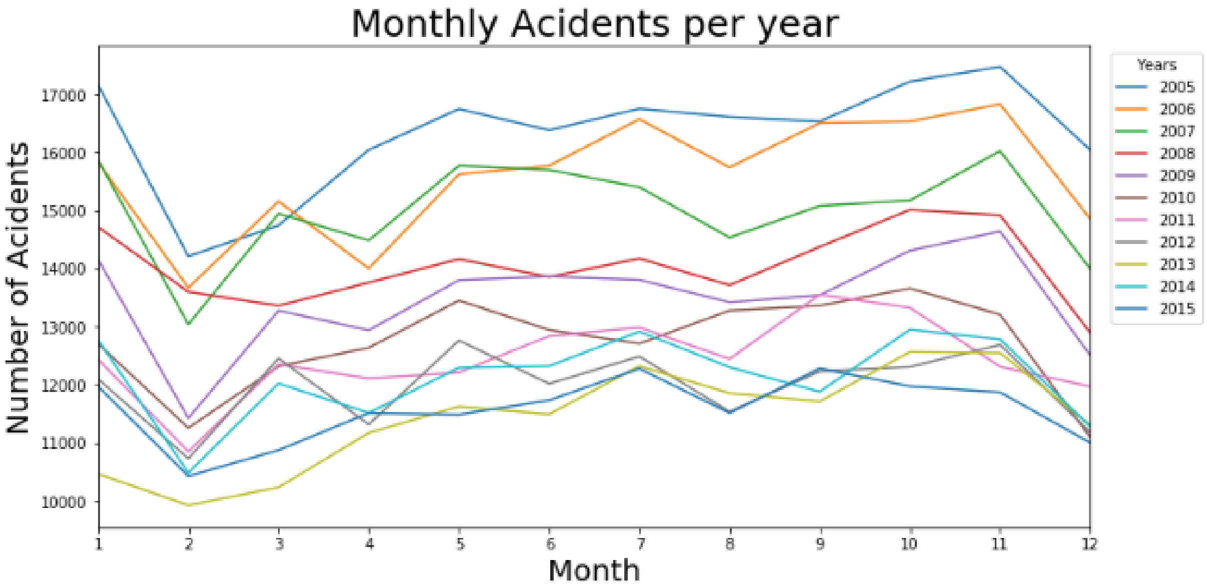
We can see that the most of accidents happened around after noon. We can assume that this time of the day has the most traffic moving such as people leaving from work.

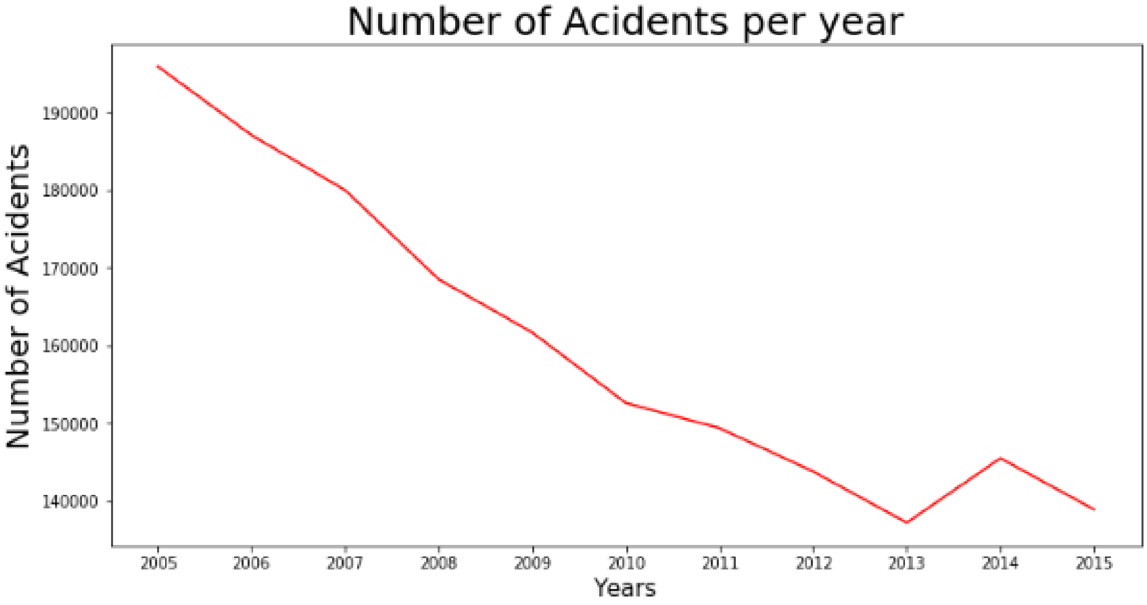
3- We will study the speed limits of the roads



Most of the accidents occurred on the road where the speed limit is 30. I was expecting more accidents on highway or major roadways. Some of the accidents could be cause of stop sign, changing lanes or turning into parking lot etc.

4- Is the number of accident increasing by the time or decreasing.



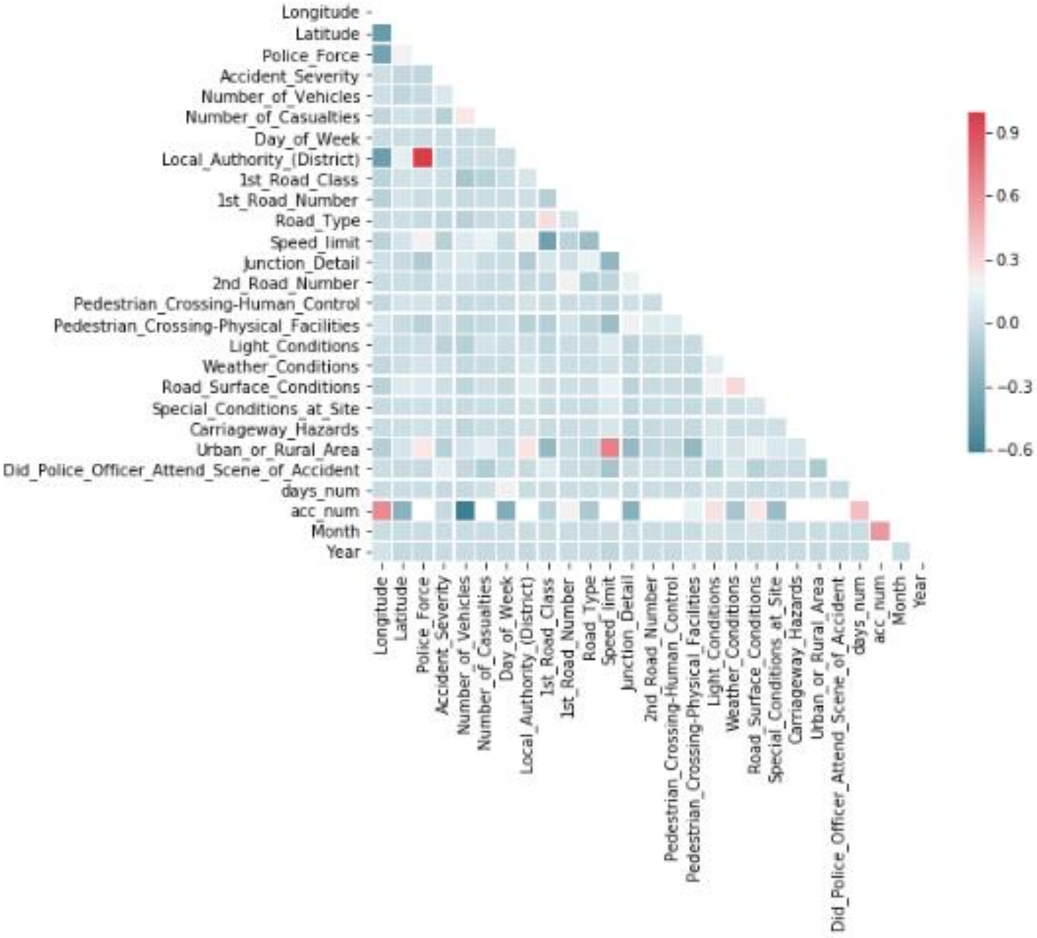


We can see that most of the accidents happened between (October – November) throughout the years.it may refer to the time of the year where Workers and Students finish their holiday and return to their work.

Also, The Second Line graph illustrates how the number of accidents decreases dramatically during the year (2005 – 2015).it may because the government set new traffic rules that brought more safety to the roads and applied careful driving on drivers.

*After all, we are trying to understand the data and how each feature effect on it. We will now try to find any relation between the features to use it to find a solution to predict accidents severity using machine learning models.*

**Correlation between all features:**



As shown in the figure that represents a heat map for the correlation between the features of the database. There are not many correlations that could help us during the process. We can see that there is a strong positive correlation between Local_Authority_ (District) and Police_force but it will not help us in our study. There is another strong positive correlation between Urban_or_Rural_Areas and Speed_limits we can use it.