Business Problem

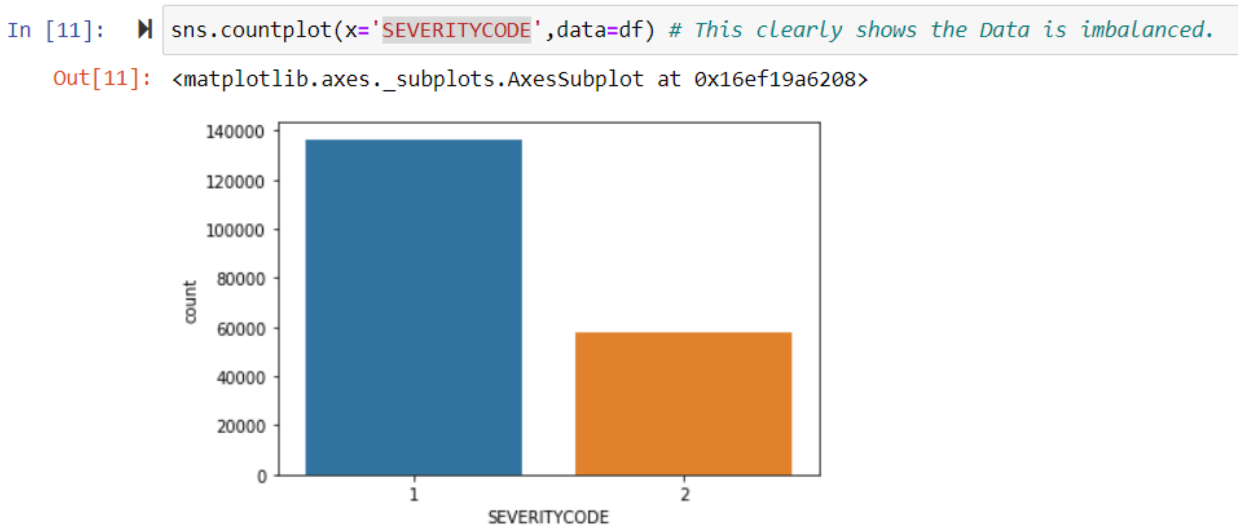
Every Year many people loose their valuable lives because of collisions happening on roads. We have data collected everywhere from police department , if we are able to analyze the data and create awareness among the people we can avoid many collisions.

Seattle one of the busiest cities in USA reports 1.3 million crashes every year, analyzing this crashes and taking measures will definitely help in reducing the crashes.

Approach Followed for Analysis

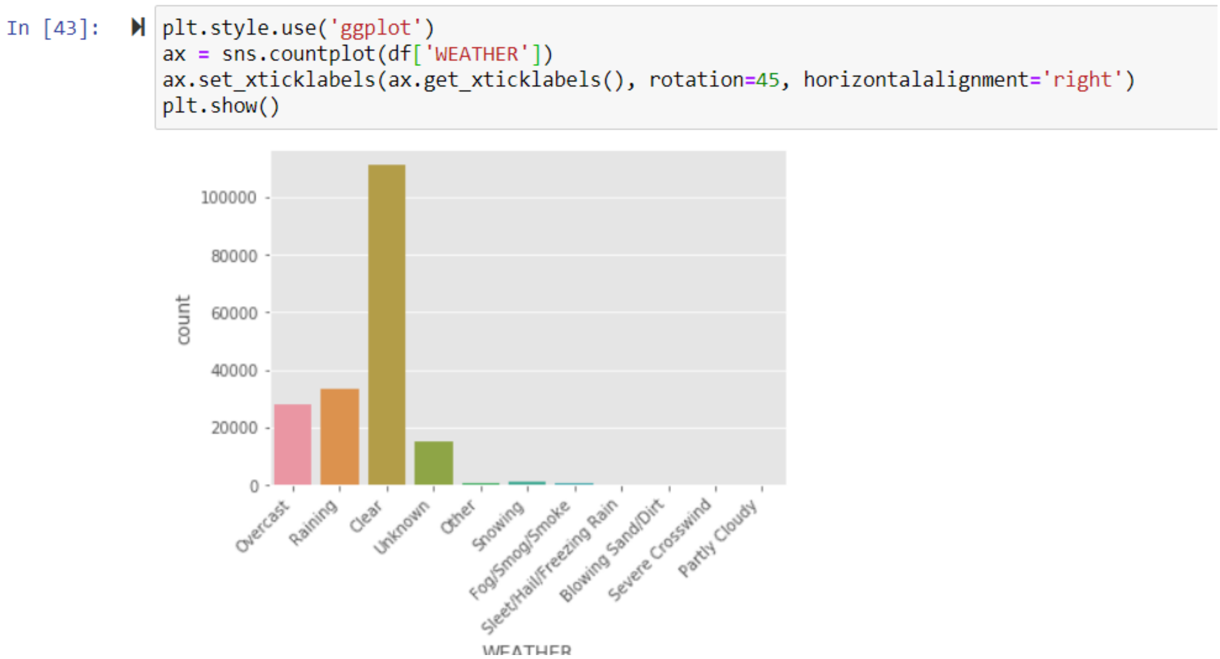
Here I am going to use the dataset given as part of course by CourseEra which is part of real data from Seattle collisions. The dataset consists of 38 columns with very less information about the features.

The Data is highly imbalanced Data , so we need to balance the data so the model dont get biased to the the Majority class.



```
In [11]:  sns.countplot(x='SEVERITYCODE',data=df) # This clearly shows the Data is imbalanced.
Out[11]:  <matplotlib.axes._subplots.AxesSubplot at 0x16ef19a6208>
```

Fig: This shows the Imbalance between the classes.



```
In [43]:  plt.style.use('ggplot')
          ax = sns.countplot(df['WEATHER'])
          ax.set_xticklabels(ax.get_xticklabels(), rotation=45, horizontalalignment='right')
          plt.show()
```

fig: This shows the weather conditions on various days, so weather has very less effect on the crashes.

Feature Engineering and Null value Handling

All categorical columns are encoded using the numerical values.

Missing values are dropped as we have huge dataset and cleaning missing values does not have any impact.

Train test split is done using Sklearn and Random forest classifier is applied.

since the Dataset is imbalanced accuracy is very less.

Further analysis

Driver inattentive is the major cause for many collisions we can create awareness sessions among the drives and reduce collisions.

Lot of feature engineering needs to be done with experts so model performance can be improved a lot.