Prepared by : **Tanmay Prasad**
For : **IBM DataScience Capstone Project**
Submitted On : **Oct 6th 2020**

2. Data Preprocessing

2.1 Data Sources

Data provided with this project request is used for this analysis. A more raw data can from Seattle GeoData where most of the data and attributes can be studied. Even though raw dataset has records from 2004 and has data for all severitycodes and studied for possible utilization for this project, course provided dataset with just 2 severity codes was utilized due to very unbalanced nature of the data. Models were evaluated for utilizing unbalanced data but could not use it because of heavily unbalanced data that requires further analysis and utilization of advanced techniques. The imported data has 194673 observations of various attributes such as severity code, severity code description, Address type, Junction type, collision type, Road Condition, Light Condition, Weather Condition, Speeding, Driving Under Influence indicator, Longitude, Latitude, Injury count, Fatality count, SDOT code, description, Data and Time of Accident etc., The attribute Types and descriptions can be found at Seattle DOT

2.2 Data Cleaning

Data downloaded from data source with nulls and Nan values were replaced with Unknown/Other values. Data formats were converted to other formats as required. A few columns such as Severity Code, Description & SDOT Code, Description & ST Code and Description were combined to new columns for analysis purposes. X and Y co-ordinates missing Latitude and Longitude were substituted with Seattle's Latitude and Longitude. Upon cleaning, it can be observed that the observations are available from April 1st 2004 to May 20th 2020.

2.3 Exploratory Data Analysis

As the source data has many attributes, all the attributes were studied for possible selection as feature for this project. Individual Attribute was checked for percentage valid data that might be a possible reason for the accident. Such attributes were further analyzed with respect to the target attribute of severity. The following shows the results of this analysis:

Total Data Observations : 194673

Attribute	Selected(Y)/Not Selected (X)	Reason
OBJECTID	X	Doesn’t provide relevant information for Prediction
SHAPE	X	Doesn’t provide relevant information for Prediction
INCKEY	X	Doesn’t provide relevant information for Prediction