Applied Data Science - Capstone

# **Introduction/Business Problem**

# The Outcome of this project is to predicting the severity of the accident, which is mainly targeted for Traffic departments. This will would also help to monitor the traffic due to accident since Traffic congestion is a big problem for all countries, especially in big cities. The higher severity level the accident is, the longer time the congestion lasted for. We analyse a wide variety of factors, including weather conditions, roadworks, traffic jams among others, an accurate prediction of the severity of the accidents can be performed. We use the collision data from Seattle to study and predict the severity level a road accident.

# 

The data made available to us by Coursera includes all types of collisions from the year 2004 to 2020-May. There were total 195K collision records with 37 variables. The data also contains many other useful information including the severity level, collision type, weather condition, road condition, the number of vehicles involved, the number of pedestrians involved, etc. variable SEVERITYCODE has 2 variables 1&2. 1 – low level, 2- high level.

* Data Count: 194673
* Attributes- 37
* Target Variable: SEVERITYCODE

## Nan variables.

* The columns which were not significant and had NAN have been dropped from dataset.
* Columns with 1 % of Nan compared to complete data set were dropped.

Encoding Data

* Used One Hot encoding for numerical transformation

# Identification Important Features:

Based on the dataset understanding and correlations have identified the below attributes

['OBJECTID','STATUS','PERSONCOUNT','PEDCOUNT', 'VEHCOUNT','INATTENTIONIND', 'PEDCYLCOUNT','WEATHER','UNDERINFL','ROADCOND','LIGHTCOND','COLLISIONTYPE','SPEEDING','SEVERITYCODE']

