Predicting Accident Severity

# Introduction:

## Background:

The data used in this project is provided by Seattle Police Department. The goal of this project is to predict the severity of an accident based on the given collision data in the csv file. There are several groups who would be interested in these predictions; however, the main focus is on the police department and other first responders would like to know the severity of the accidents to maneuver the help that is needed for the victims as soon as possible. Further, the everyday commuters would also benefit from knowing this information, which can help them decide how to re-route their travel plans.

## Problem Statement:

In a metropolitan city like Seattle, there is always a high chance of severe accidents based on the heavy traffic during the rush hours. It would be extremely beneficial for the Police Department and other first responders to be able to predict the severity of a given accidents based on the preliminary information they obtain; this can help the first responders response faster and could help save lives.

## Interests:

A prediction of severity of accidents could help the first responders to respond to a given situation swiftly and with optimum tools. This could not only save lives but also help save time and resources, could prevent the first responders from overwhelming situations on a given day.

Major risk could be over-prediction and thus engaging a large number of first responders even when the accident is relatively less severe. Though this could be tackled by updating the data and remodeling the model on new/ updated model.

## Objective:

Objective of this project is to predict the severity of the given accident based on the preliminary information which is known to the first responders when they receive a 911 call. In order to develop a machine learning model, the CSV file that was downloaded will be used. It has several parameters/ attributes. The data will be carefully filtered through the elementary data analysis phase. The filtered data will be used to develop the model. This dataset to help create and train the model, and a subset will be used for testing this model.

# Data Acquisition and Cleaning:

## Data acquisition:

The data used in this project was shared by the IBM team in the week one of the Applied Data Science Capstone project. A link to this page with CSV file has been attached [here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv). The data set has 194673 rows of data, and 38 columns. There are two columns with ‘Severity code’ repetition, thus leaving 37 total attributes.

## Data Cleaning:

Out of the 38 columns, there were 19 columns which had missing values. Out of these 19 columns column for speeding was modified to show ‘N’ as in not speeding for the missing values. Thus, reducing number of missing value columns to 19. It was noticed that there were several columns which were added to provide the administrative information during report filling by the Seattle Police Department. These included columns such as ‘INTKEY’, 'EXCEPTRSNCODE', 'EXCEPTRSNDESC'. These columns do not provide information necessary for the analysis in this project. Hence these columns were removed. A list of removed columns will be provided in table 1 below in the next sub-section.

## Feature Selection:

As mentioned above there were missing values in 18 columns. There were two columns providing the severity code in the data frame, one of them was removed to reduce the redundancy.

The X, Y co-ordinates provided, was used to plot a map using folium to understand the distribution of the locations around the Seattle city limits.

Further, based on a exploratory data analysis, the columns of interest were selected and distributed into two groups Natural Factors and Vehicular Factors. Natural Factors included columns such as Weather condition, Light conditions and road conditions. The Vehicular Factors included the columns such as Collision type, Vehicle Count, Pedestrian Count, Pedestrian-Cycle count, Person Count, Speeding. The third category was location factors which included Junction and Crosswalk.

In the next sections, a preliminary data analysis is shown. This will be followed by methodology to develop a ML model using DecisionTree model and followed by training and testing results on the datasets.

# Elementary Data Exploration:

## Basic statistical analysis:

## Relationship between