# Capstone Project Report
# Topics: Road Accident Severity

## 1. Introduction

It was summers and we decided to take a road trip to a nearby ski resort. We had our check in at 11:00 AM. Since we were students, it was an expensive resort for us, and we wanted to make the most of our time. We didn't check the weather forecast, and we assumed that our navigation would get us to the place in 4 hours. However, the weather was severe, and we had to drive with our usual speed to reach on time. That is when our car lost the grip on the road and was flung across the shoulder into a tree. Luckily, we survived the crash, however, the car was totaled. If there was a way which could analyze the weather conditions and the area from where we had to drive and informed us about the possibility of a severe accident, we would've left way early an driven at a much slower speed than the regular one. This would've saved us from the unfortunate crash and a hefty repair cost.

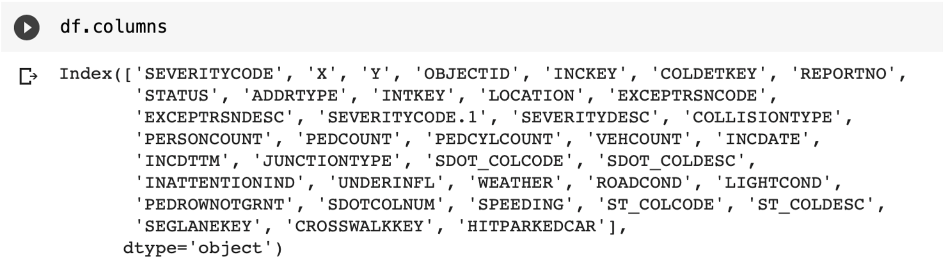
### 1.1 Business Problem

Unfortunate incidents often occur on the road. The unpredictability of these accidents makes them so dangerous. It is usually the impatient nature of the human which leads to an accident. But wouldn't it be a blessing if the number of accidents that occur on a daily basis be reduced? That is the problem which will be addressed in this project. As we know, accidents can happen anytime and anywhere. However, there are many factors which might influence the severity of the accident. For instance, these include the weather conditions, time of the day, the speed of the car and the area the car is being driven. These factors greatly contribute whether the accident will be severe or not. Having the information about the above-mentioned factors can be used to predict if the accident happens can be severe or not. The indication that there is a possibility of a severe accident if it happens might warn the car drivers to drive more carefully and hence prevent accidents.

This would greatly reduce the loss of life and also damage to the property. This would also help routing software to give a warning which would alert the drivers and their insurance companies which would help them in saving cost.

## 2. Data

The data used in the project is a collision data set provided by the Seattle Department of Transport, Traffic Management Division. The data set includes all the types of collisions occurred in Seattle from 2004 to present. The following screenshot below gives us what columns exit in the data.



```
df.columns
```

```
Index(['SEVERITYCODE', 'X', 'Y', 'OBJECTID', 'INCKEY', 'COLDETKEY', 'REPORTNO',
       'STATUS', 'ADDRTYPE', 'INTKEY', 'LOCATION', 'EXCEPTRSNCODE',
       'EXCEPTRSNDESC', 'SEVERITYCODE.1', 'SEVERITYDESC', 'COLLISIONTYPE',
       'PERSONCOUNT', 'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE',
       'INCDTTM', 'JUNCTIONTYPE', 'SDOT_COLCODE', 'SDOT_COLDESC',
       'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND',
       'PEDROWNOTGRNT', 'SDOTCOLNUM', 'SPEEDING', 'ST_COLCODE', 'ST_COLDESC',
       'SEGLANEKEY', 'CROSSWALKKEY', 'HITPARKEDCAR'],
      dtype='object')
```

## 2.2 Data Cleaning

Since the data consists of many unnecessary columns and missing values, the data was cleaned using different data cleaning methods. Since the prediction was mainly based on Person Count, Vehicle Count, Attention ID, under influence verification, Weather, Road conditions, Lighting conditions; the other trivial columns were removed.

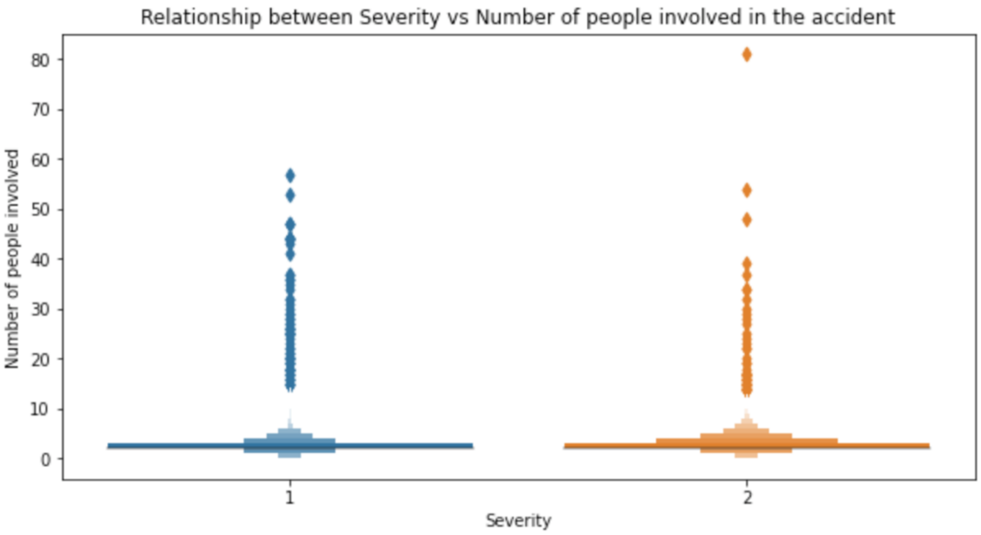
The columns in the data consisted of null values which were hindering in increasing the accuracy of the prediction model. Therefore, the missing values in the columns were replaced with the values of highest frequency. This prevented reduction in the training data set and provided higher efficiency.

Using the One-Hot-Encoding method, columns containing different categorical labels were split into more columns. Columns which contained the values initially were given a value 1 and the rest were given a value of 0.

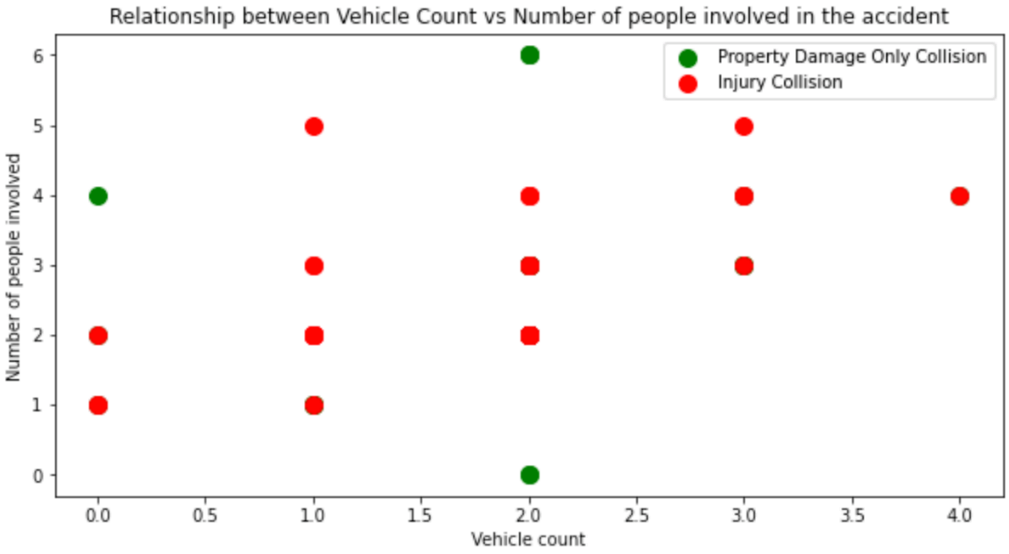
Many columns contained different values involving a number and a character. The values were standardized by replacing the characters ‘Y’ and ‘N’ with 1 and 0 respectively. This would ensure that the classification model would have integer values as input.

## 2.3 Exploratory Analysis

The data in the columns was used for exploratory analysis. Firstly, a boxplot of severity vs the number of people involved in the accident was plotted. It was observed that majority of the accidents involved around 1 to 2 people. Through this we could conclude that if there are higher people traveling, the possibility of severe collision is less since more people mean more traffic which leads to less possible speeding.
The box plot below shows this analysis.



Another important observation was found out by using the scatter plot by plotting the vehicle count and the number of people involved in the accident. Seeing this plot, we could conclude that conclude that when the number of vehicles involved in the accident were high, there was more injury collision than the collisions which only damaged the property.
The scatter plot below shows this analysis.



2.4 After the exploratory analysis, the data attributes required for the prediction are stored in variable X and the target variable is stored in variable y. After the separation, the attribute data is normalized to ensure equal contribution of each attribute.

This data will then be used to create a train/test data set. We use the train and test split in the data to prevent overfitting and ensure that there is some out-of-sample data for testing the model.