Predicting Crash Severity for Seattle Car Collisions

Applied Data Science Capstone

Dimitris Paschalidis - 13 OCT 2020

BACKGROUND:

The number of traffic collisions and their victims has been a rising trend globally due to increases in population and motorization. Traffic collisions disturb the traffic operations, break down the traffic flow, and cause severe urban problems worldwide. Major traffic accidents can sometimes lead to irreparable damages, injuries, and even fatalities. In order to take necessary actions to control this ever-growing problem, extensive research has been carried out into the prediction of traffic collisions in both developed and developing countries using various statistical techniques. Different factors involved in traffic collisions have a substantial effect on each other, thus making it difficult to individually consider any of the parameters when explaining the severity of traffic collisions.

Realizing traffic accidents as a preventable problem developed countries have implemented different policies and measures to reduce this problem. These include enforcement, education, training and engineering improvements. Any part of this report can be utilized by the government authorities for making necessary policy changes to avoid collisions or to minimize their severity.

OBJECTIVES OF THIS PROJECT:

The main objective of the research is to investigate the role of factors in collision severity using Seattle Department of Transportation data and predictive models. Specific objectives include:

1) Exploring the underlying variables such as human characteristics, vehicle characteristics, roadway characteristics, and environmental characteristics that impact collision severity.
2) Predicting collision severity using Decision Tree and Logistic Regression

DESCRIPTION OF THE DATASET:

Governments, states, provinces and municipalities collect and manage data for their internal operations. In the last decade, an open data movement has emerged that encourages governments to make the data they collect available to the public as "open data". Open data is defined as "structured data that is machine-readable, freely shared, used and built on without restrictions.

The data set used here is taken from the open data website of the Seattle City. It is published by the Seattle Department of Transportation. The dataset contains information about 194673 collisions, recorded between 2004-01-01 00:00:00 and 2020-05-20 00:00:00.

The data can be accessed through the following link: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

The metadata for the same can be accessed through: https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf

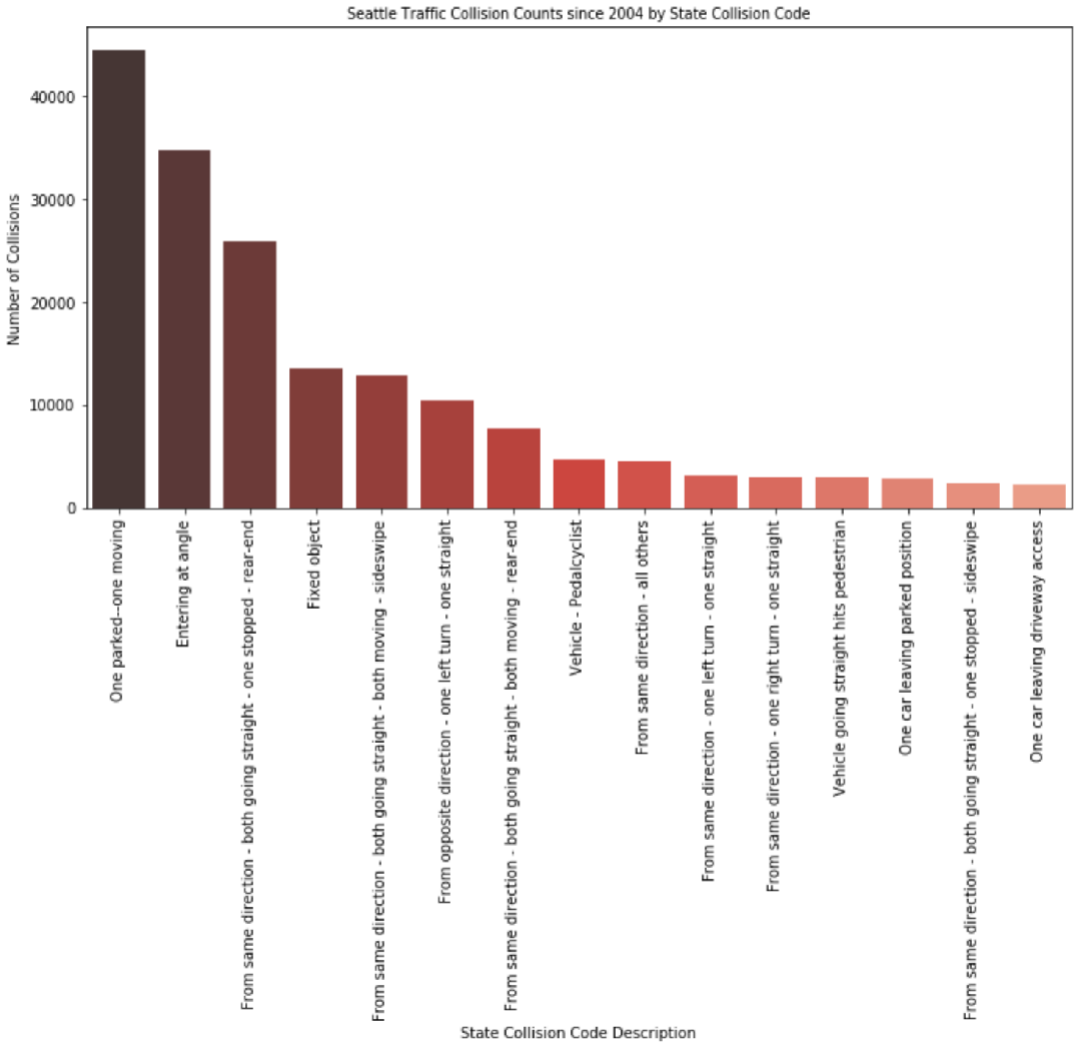
APPROACH:

First of all, the data set will be analyzed using data visualization tools and libraries in python to identify trends in collisions and parameters affecting the collisions. Then the data set will be modeled to predict collision severity. The data set mentions 2 levels of collision severity: 1- Property Damage Only Collision 2- Injury Collision The approach for modeling collision severity involves statistical modeling considering severity as a dependent variable while road conditions, speeding, driver attention, influence of drugs/ alcohol on driver, junction type where the collision occurred and a few environmental factors as the independent variables.

ASSUMPTIONS:

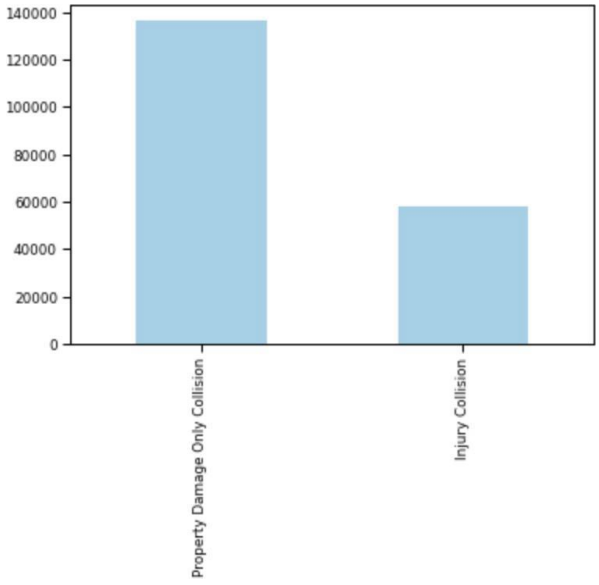
A few of the columns in the data set contained categorical values, 'Y': Yes and NaN. It is assumed that the NaN values correspond to 'N':No. It is also assumed that the data values - 'Other' and 'Unknown' correspond to Null as they tell us nothing about the features in the dataset.

Exploratory Data Analysis:
The Seattle Department of Transportation has categorized the types of collisions into 62 categories. Let us look at some of the most frequent types of collisions.



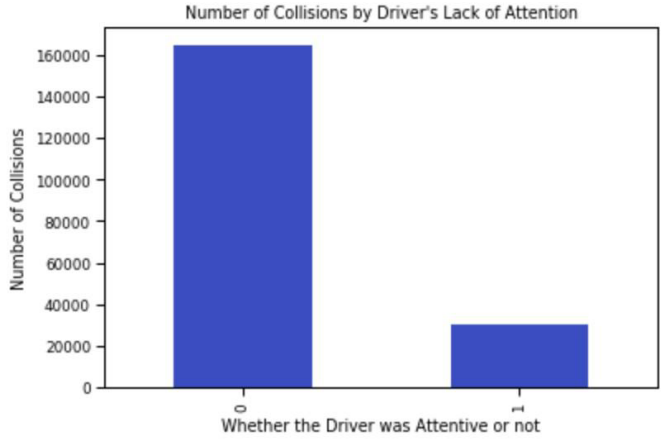
We see that the most frequent category is where one vehicle is parked and one is moving followed by another category which is where the vehicle is entering at an angle. Another category that must be considered is when both the vehicles are going in the same direction and one of the vehicles stops suddenly and the other vehicle hits the former's rear end.
Inferences: These categories suggest that the main causes of collision are driver's inattention or speeding or the driver being under influence of drugs/alcohol.

Now we will plot the collision severity to identify the number of collisions in the 2 severity categories.



Now let us plot the INATTENTIONIND variable to identify the number of collisions caused by driver's inattention.

```
INATTENTIONIND:
 0    164868
1     29805
Name: INATTENTIONIND, dtype: int64
```

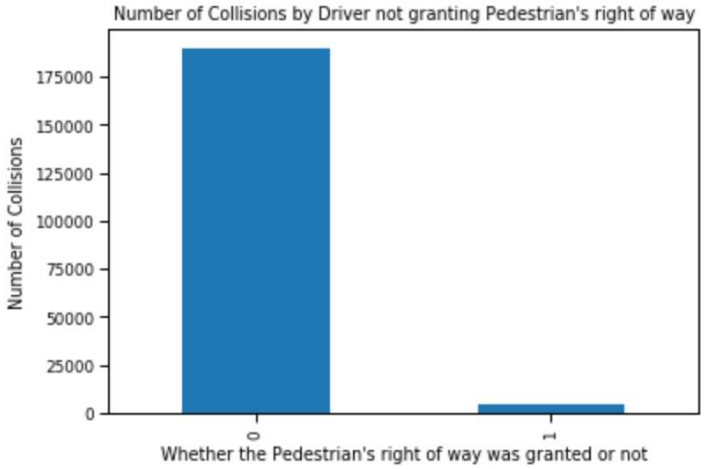


We now look at the Number of collisions caused by speeding of the vehicle:

```
SPEEDING:
 0    185340
1      9333
Name: SPEEDING, dtype: int64
```

It is seen that speeding is not the major factor leading to collisions, however it is still significant enough to be considered in decision making by the government authorities. Every state provides the pedestrian with some form of right of way to cross the street in a crosswalk. So, whether the pedestrian's right of way was granted by the driver or not becomes an important factor in identifying the cause of collisions.

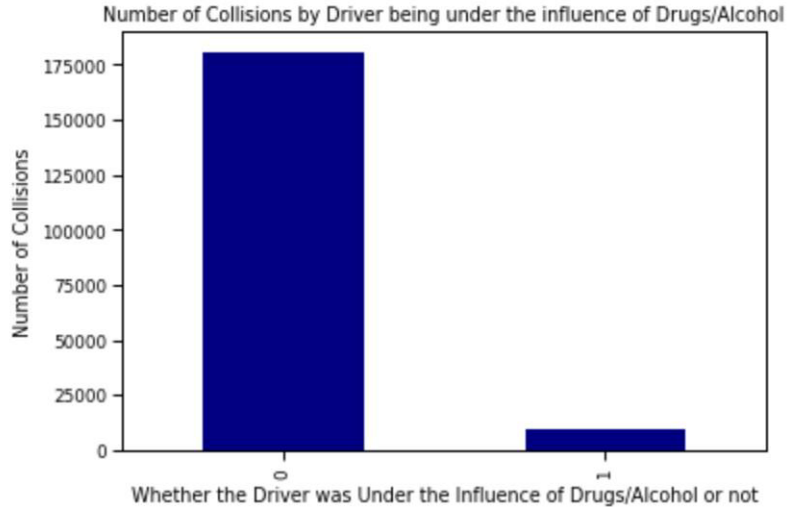
```
PEDROWNOTGRNT:
 0    190006
1      4667
Name: PEDROWNOTGRNT, dtype: int64
```



Hence, even though the state provides pedestrians right of way, the drivers tend to overlook which, as can be seen in the plot above, is causing collisions.

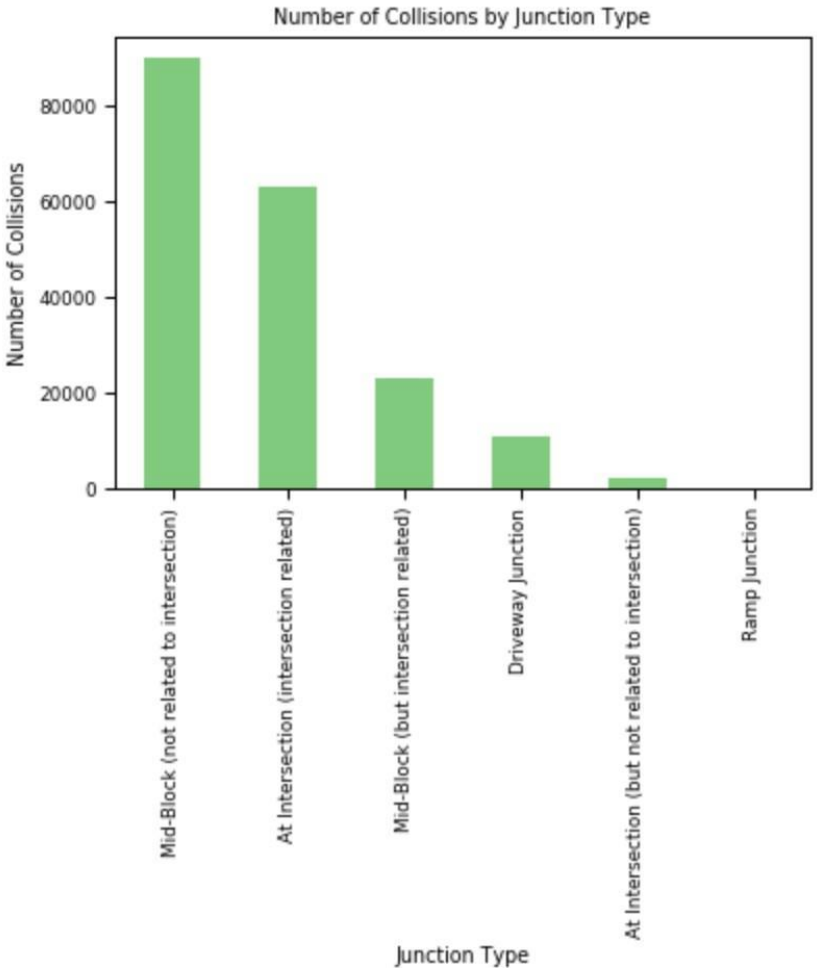
Drink and drive is a serious offence and has been one of the major factors causing collisions leading to property damage and fatalities. It is important to consider whether the driver was under the influence of drugs or alcohol when the collision took place.

```
UNDERINFL:
 0    180668
1      9121
Name: UNDERINFL, dtype: int64
```



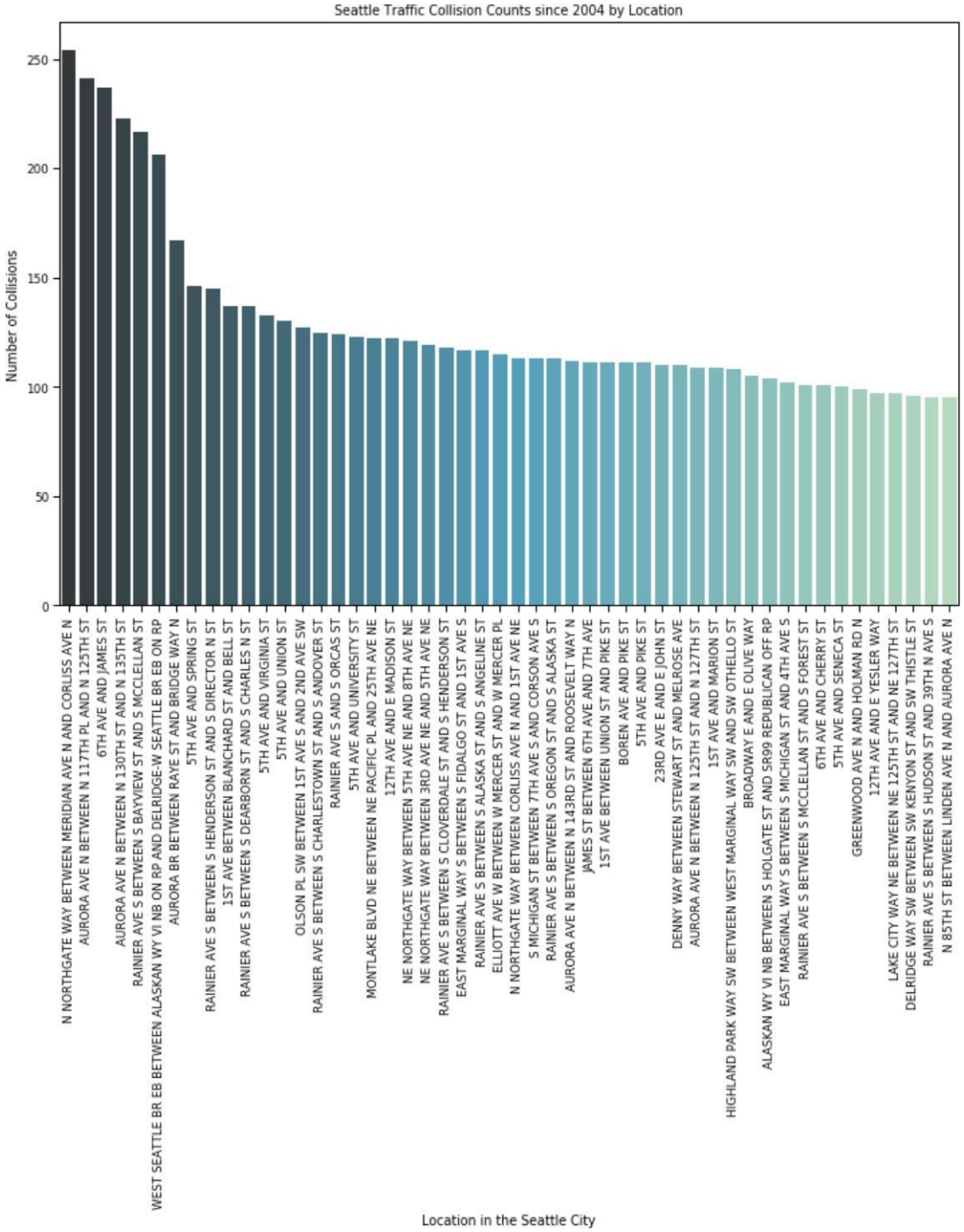
Apart from the location where the collision took place and the position of vehicles, pedestrians and pedal-cyclists, another parameter which is crucial is the junction type. The following plot gives us the number of collisions corresponding to various types of junctions.

```
JUNCTIONTYPE:
 Mid-Block (not related to intersection)              89800
At Intersection (intersection related)                62810
Mid-Block (but intersection related)                  22790
Driveway Junction                                     10671
At Intersection (but not related to intersection)      2098
Ramp Junction                                           166
Name: JUNCTIONTYPE, dtype: int64
```

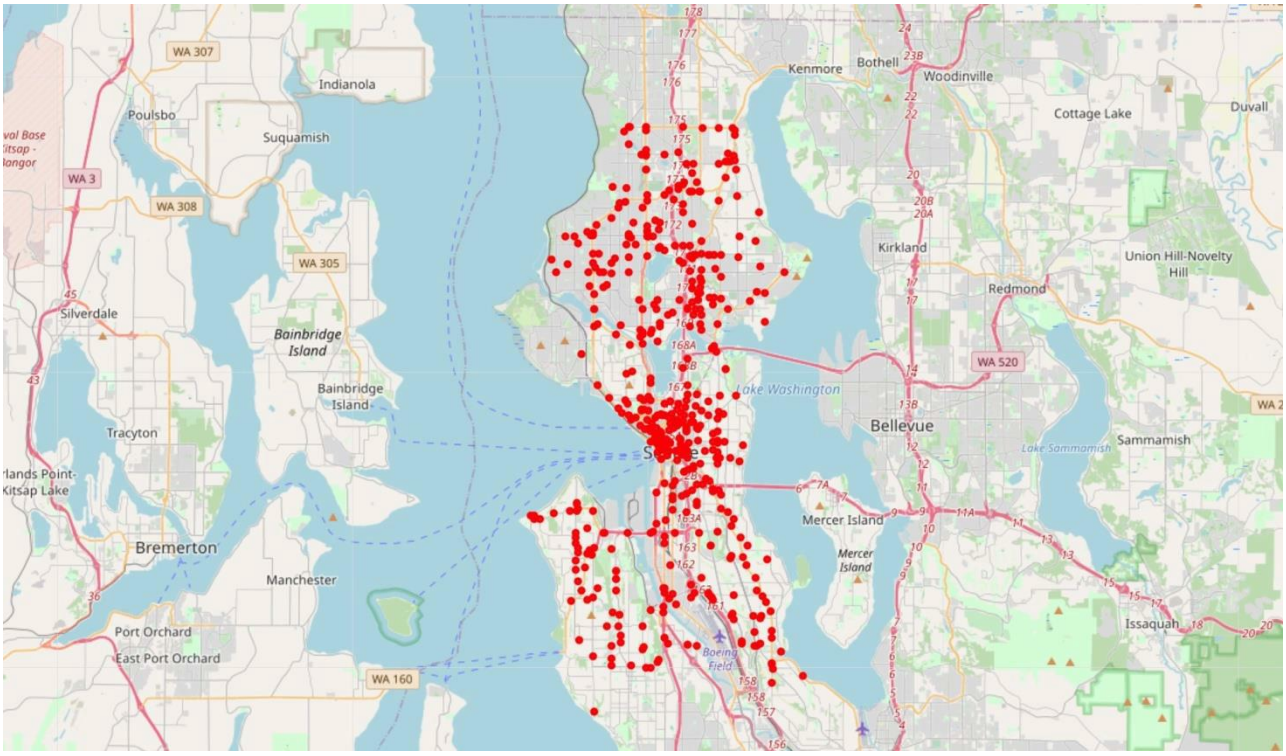


It can be seen that the number of collisions that take place at the intersection of roads or intersection related collisions are more than those not related to intersections. This information can be used to frame rules specifically for the intersections.

Now we will Dataframe using the State Collision Code value counts to identify the frequency of the type of collisions in the Seattle City.

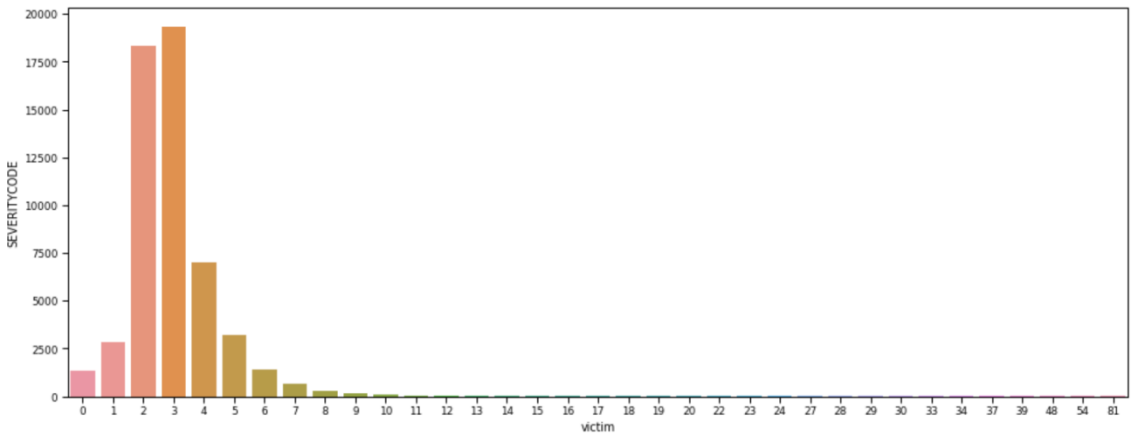


Let us also look these locations on the world map:

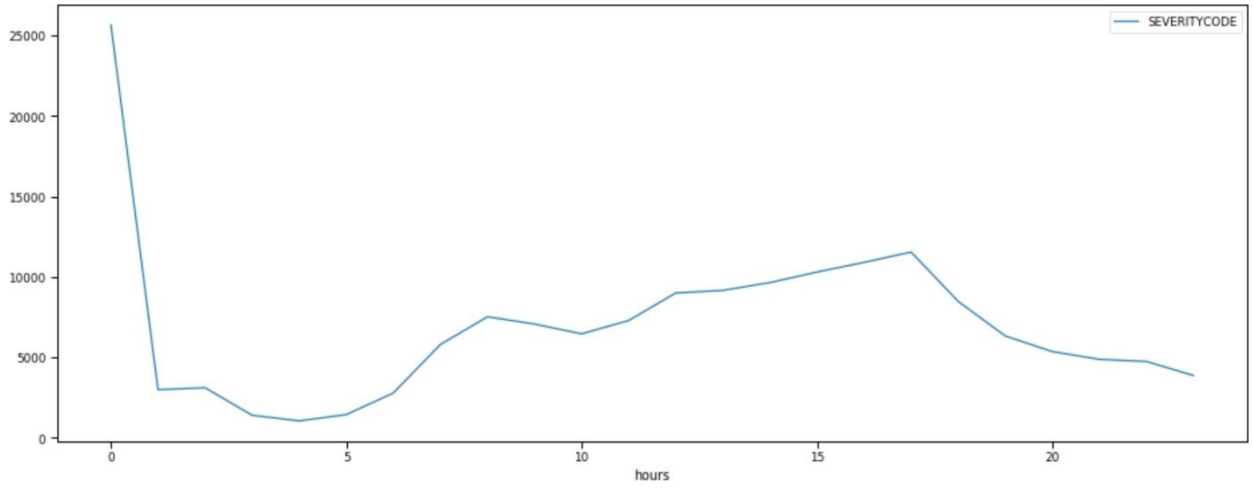


The bar plot of the collisions numbers by location and their markings on the world map suggest that the location that is the most prone to collisions is the Downtown Seattle.

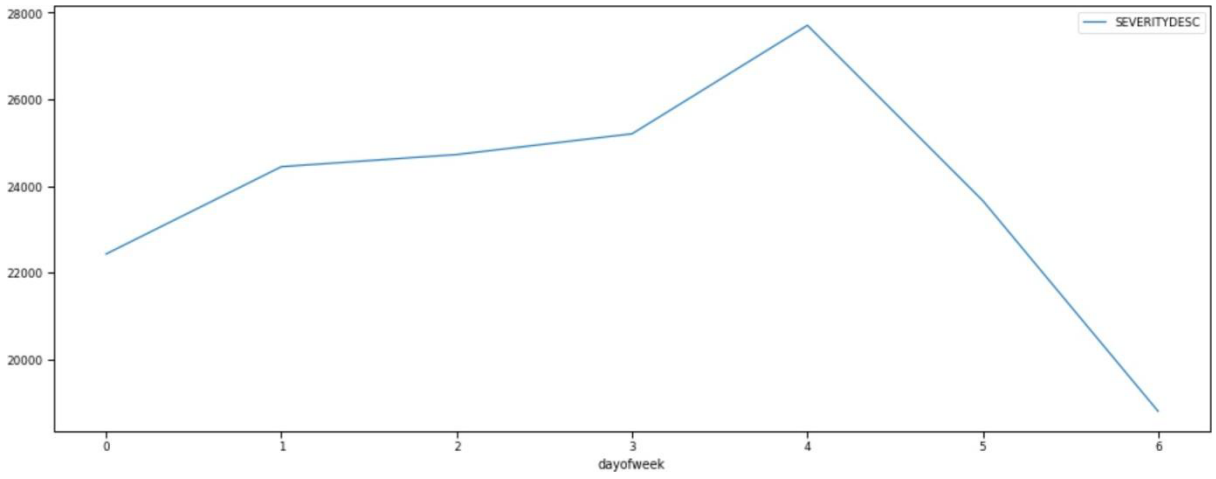
The data set contains collisions leading property damage and property damage with injuries. We will now take subset of our original dataset to identify the number of people injured per collision and the most frequent number of victims.

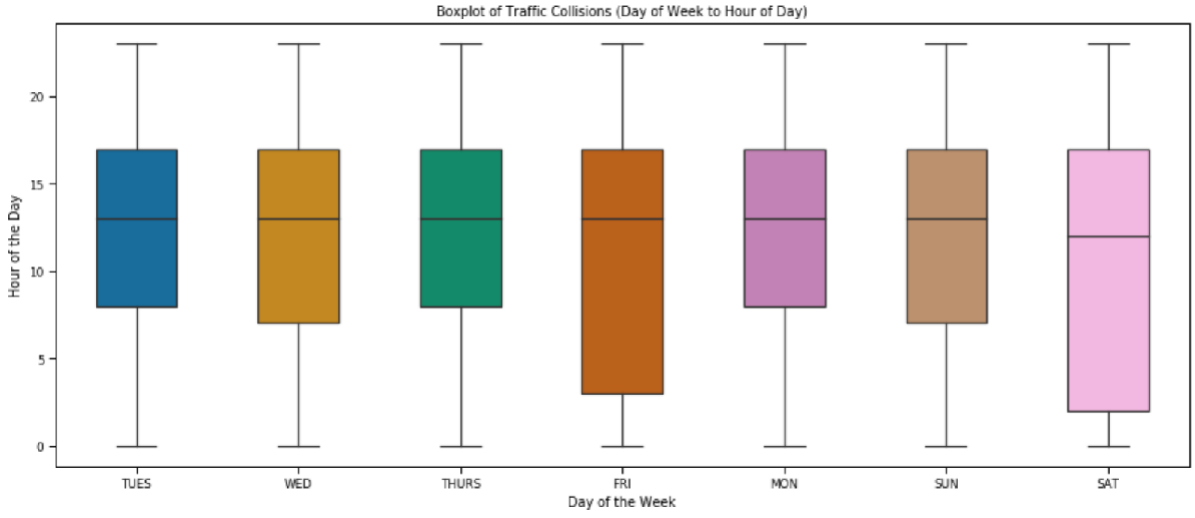


It can be seen that in general, 2-3 people are injured during severe collisions. Looking at the time periods at which the highest number of collisions occured:



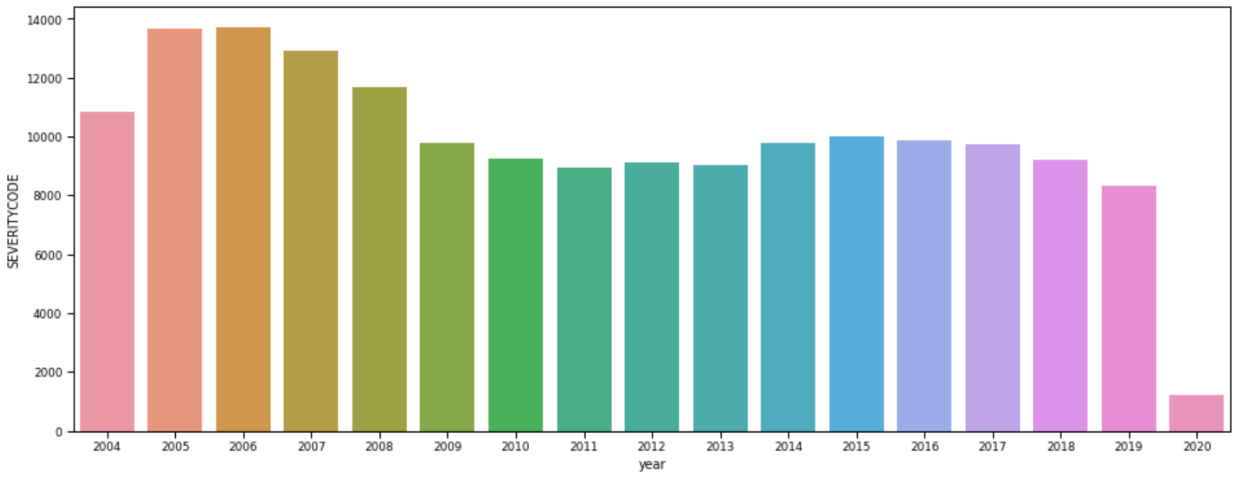
Surprisingly, most of the collisions occured around the midnight. Other time periods when the number of collisions is high are the office and lunch hours i.e., in the morning around 9am, during lunch hours around 1-2pm and in the evening around 6pm.



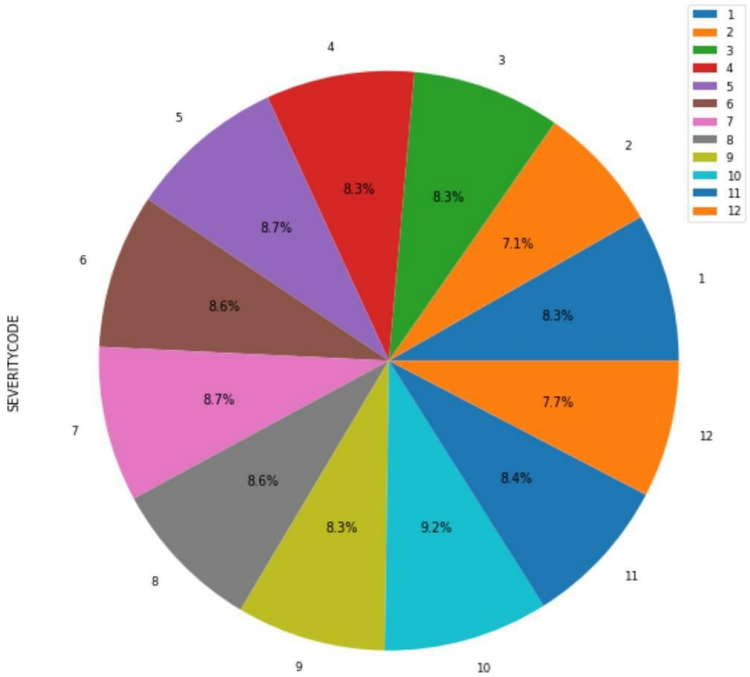


In the plot above, 0 : Monday 1 : Tuesday 2 : Wednesday 3 : Thursday 4 : Friday 5 : Saturday 6 : Sunday
Combining the day of week and the time when the number of collisions are high suggests that the highest number of collisions happen on Friday nights.

Year on Year Trend of Collisions (2004 - 05,2020):

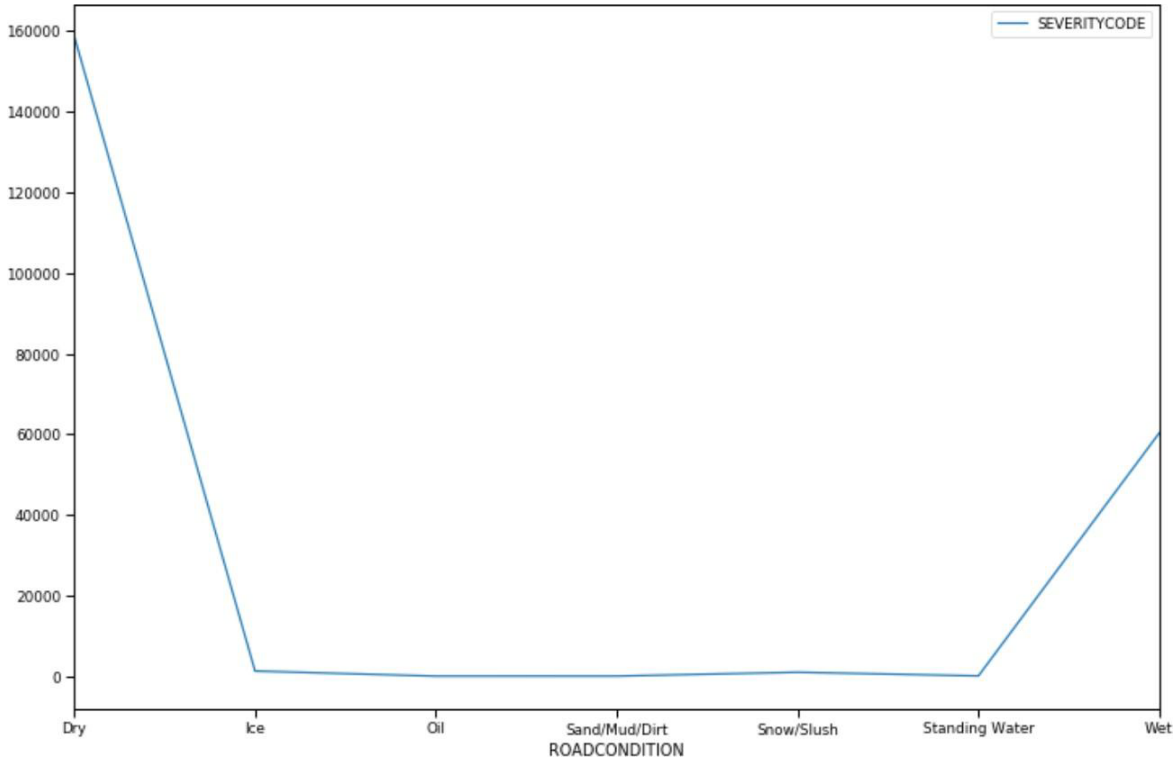


Monthly Trend of Collisions:

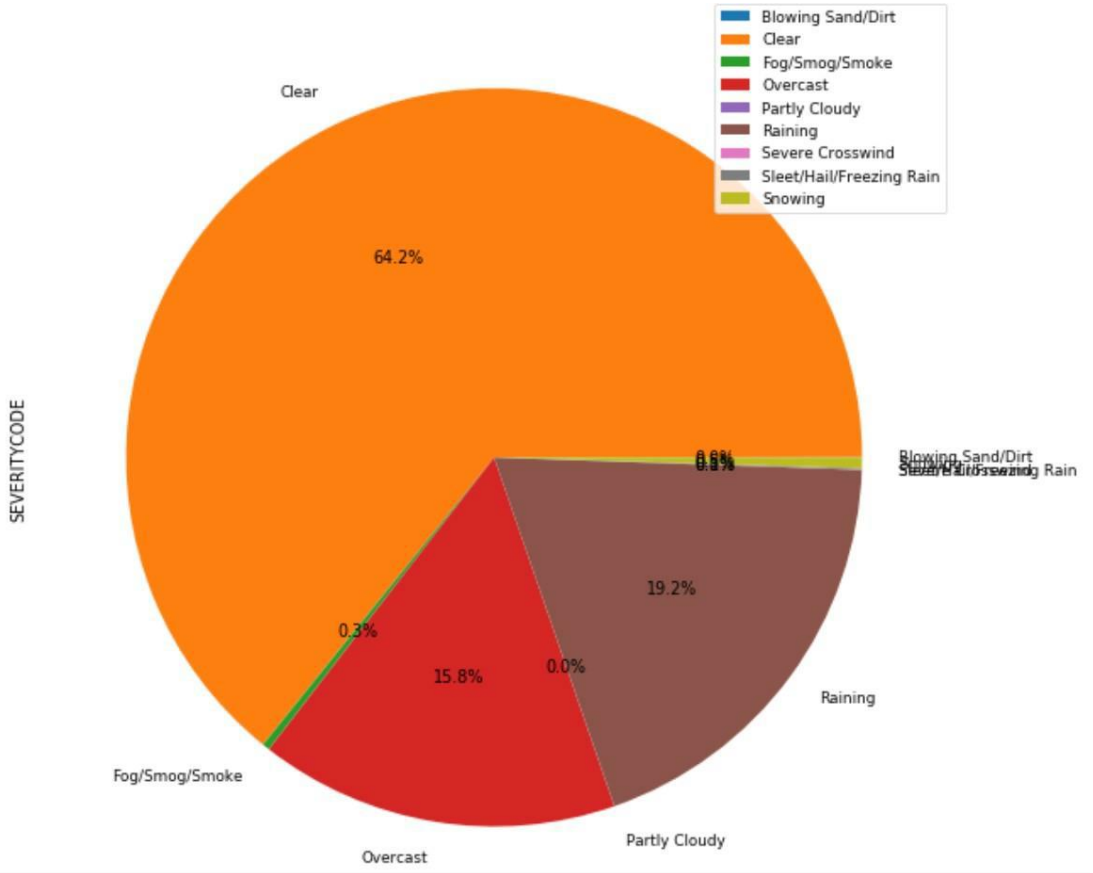


We observe no particular trend when the numbers of collisions are plotted against years or months.

The condition of the road is another parameter that must be considered for the collisions:



Percentage of Collisions by Weather:



As we can see above, most of the coefficient values are insignificant, hence it is safe to assume that the features corresponding to those coefficients do not contribute in predicting the severity of the collisions. Let us model our data using only the features with significant values of coefficients and check if the accuracy of the model is affected or not.

MODELLING THE DATASET:

As we can see, all of our features are categorical. Therefore, for modelling the dataset, we will use two of the most popular classification algorithms, the decision tree and the logistic regression. The feature selection was done by looking at the coefficients of the independent variables in the decision tree.
First of all, I modelled all the relevant features:
COLLISIONTYPE : A feature which tells whether the collision happened at an angle, sideswipe or with a parked car
UNDERINFLUENCE: This feature tells us whether the driver was under the influence of drugs/Alcohol
JUNCTIONTYPE: Tells us about the type of junction, whether the collision took place at the intersection, mid block, at a driveway junction or at a ramp junction
INATTENTION: Tells us whether the driver was attentive when the collision took place or not
WEATHER: The weather at the time of collision
ROADCONDITION: The condition of the road at the location of the collision
LIGHTCONDITION: The light condition at the time of the collision, whether there was daylight or it was dark
SPEEDING: If the driver was speeding
On checking the significance of the coefficients of these independent variables, it was seen that some of them could be considered irrelevant as their coefficients were too low.
It was seen that only two parameters has significant coefficients: COLLISIONTYPE, UNDERINFLUENCE.

RESULT:

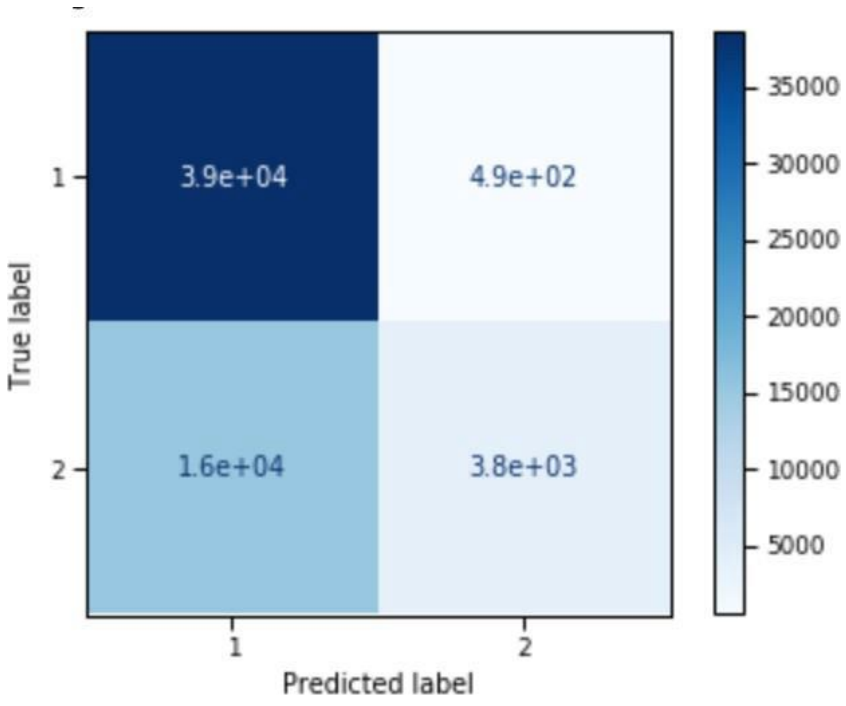
We see that both of our classification models can predict the severity of the collision up to accuracy of 73% given the independent features.

CONFUSION MATRIX OF THE MODELS:

```
[[38650   489]
 [15506  3796]]
              precision    recall  f1-score   support

           1       0.71      0.99      0.83     39139
           2       0.89      0.20      0.32     19302

    accuracy                           0.73     58441
   macro avg       0.80      0.59      0.58     58441
weighted avg       0.77      0.73      0.66     58441
```



DISCUSSION:

Hence, we see that the major parameters that contribute towards predicting severity of a collision are the collision type and the parameter identifying whether the driver was under influence of drugs/alcohol or not. Other than that, the Seattle Department of Transportation Code which classifies the collisions into various collision categories also

helps in predicting the collision severity. Actually, what can be seen is the parameters defining the details of collisions are mainly important for predicting the severity of collision. I believe that the parameters relating to road-condition, speeding, light condition, weather, junction type etc. can better be used for predicting whether a collision is likely to happen or not; whereas for predicting the collision severity, the type and number of vehicles, the angle at which the vehicle collided with another vehicle or person are the most pronounced parameters.

CONCLUSION:

Road traffic accident constitutes a serious problem and prediction of its magnitude using reliable approaches has become a necessity. An accident prediction model was developed using two classification algorithms through analysing the relationship between accidents and parameters affecting them for which data were available. In this project, I collected and cleaned traffic collision data, attempted to construct novel attributes, and tested a number of predictive models.