Predicting Crash Severity for Seattle Car Collisions

Predicting Crash Severity for Seattle Car Collisions

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.

THE DATASET AND ITS AQUISITION:

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:

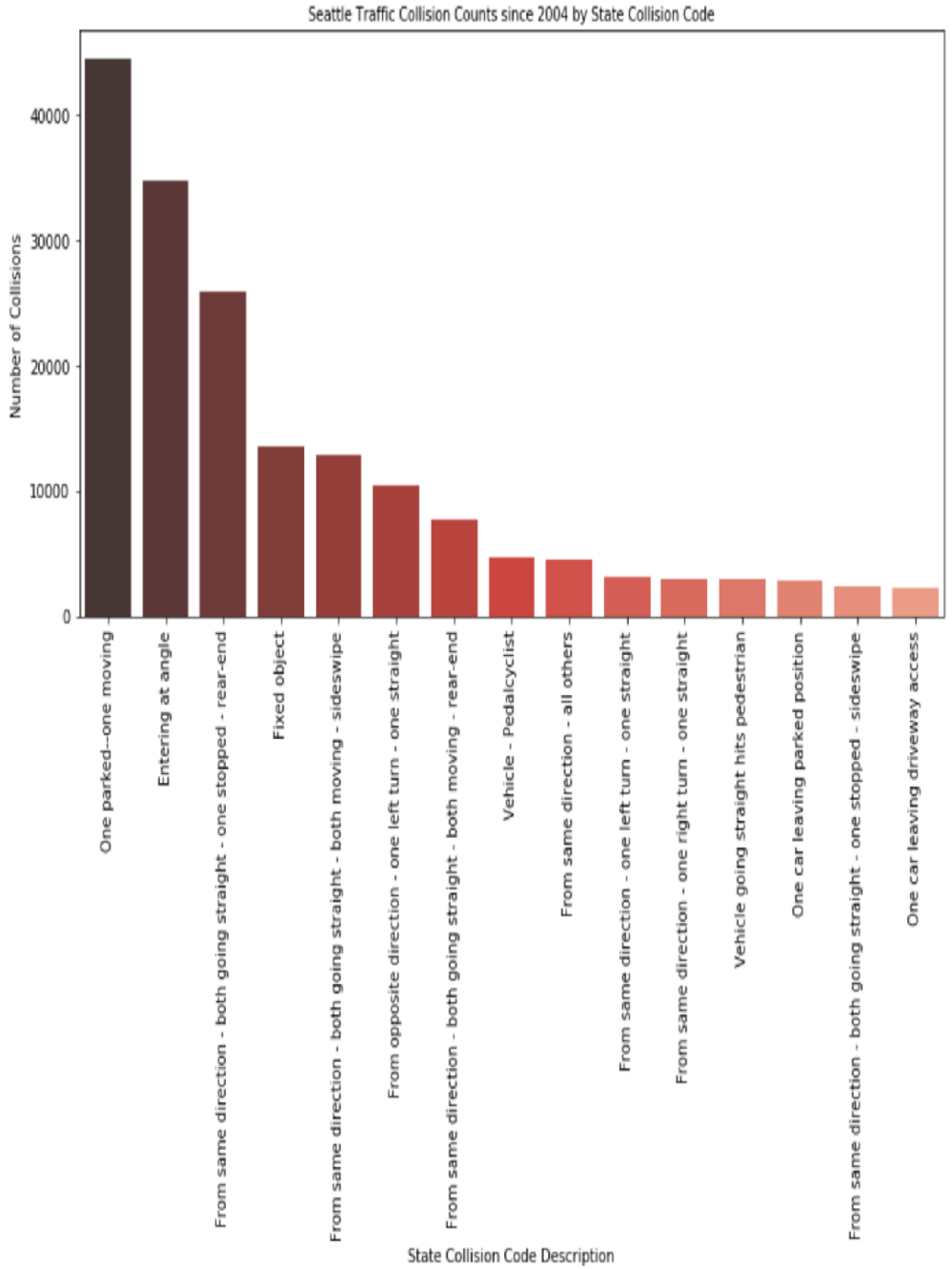
First of all, the data set will be analyzed using data visualization tools and libraries in python to identify trends in collisions and parameters affecting the collisions. Then the data set will be modelled to predict collision severity. The data set mentions 2 levels of collision severity: 1- Property Damage Only Collision 2- Injury Collision The approach for modelling collision severity involves statistical modelling 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

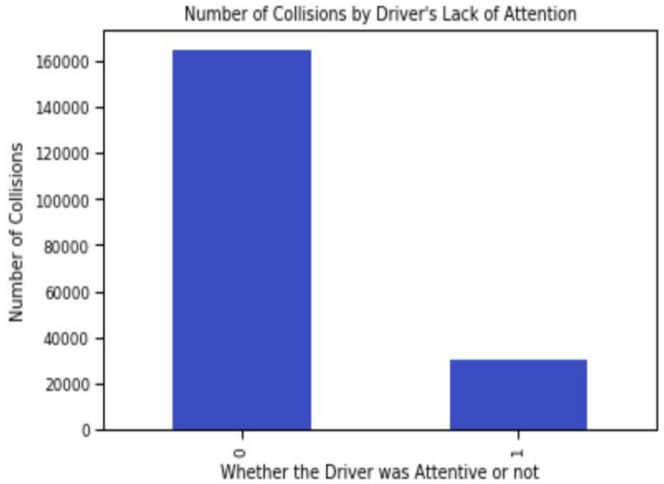
-'Other' and 'Unknown' correspond to Null as they tell us nothing about the features in the dataset.

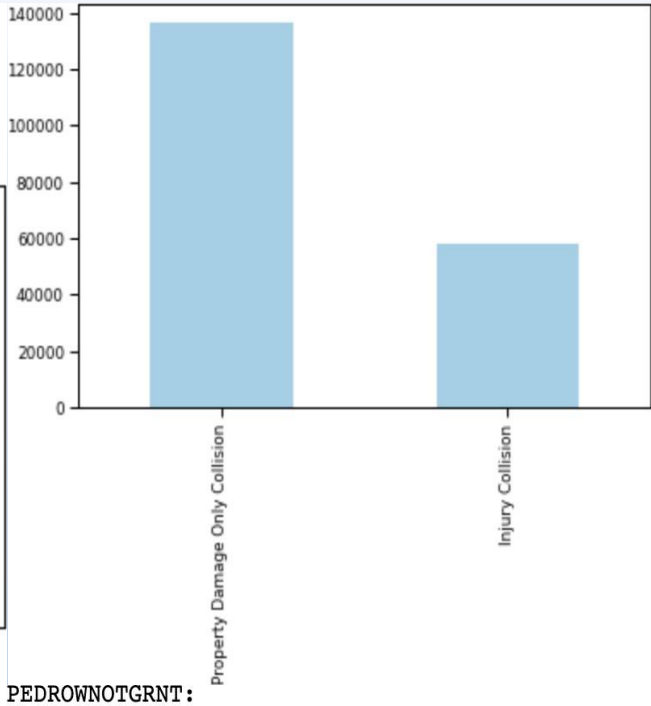
Exploratory Data Analysis



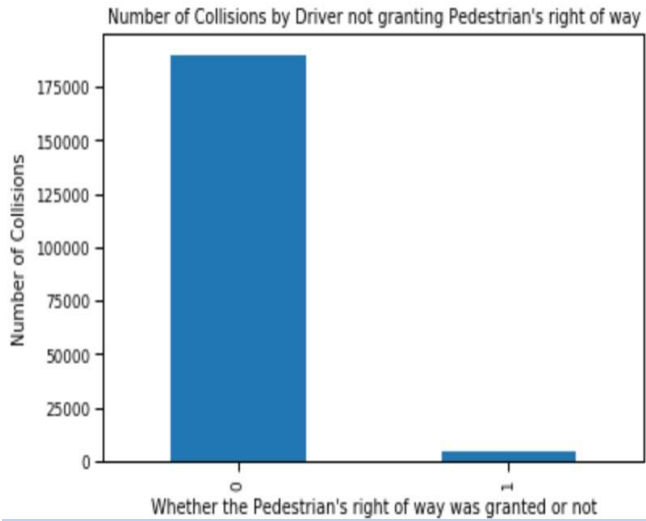

```
SPEEDING:
 0    185340
1      9333
Name: SPEEDING, dtype: int64
```

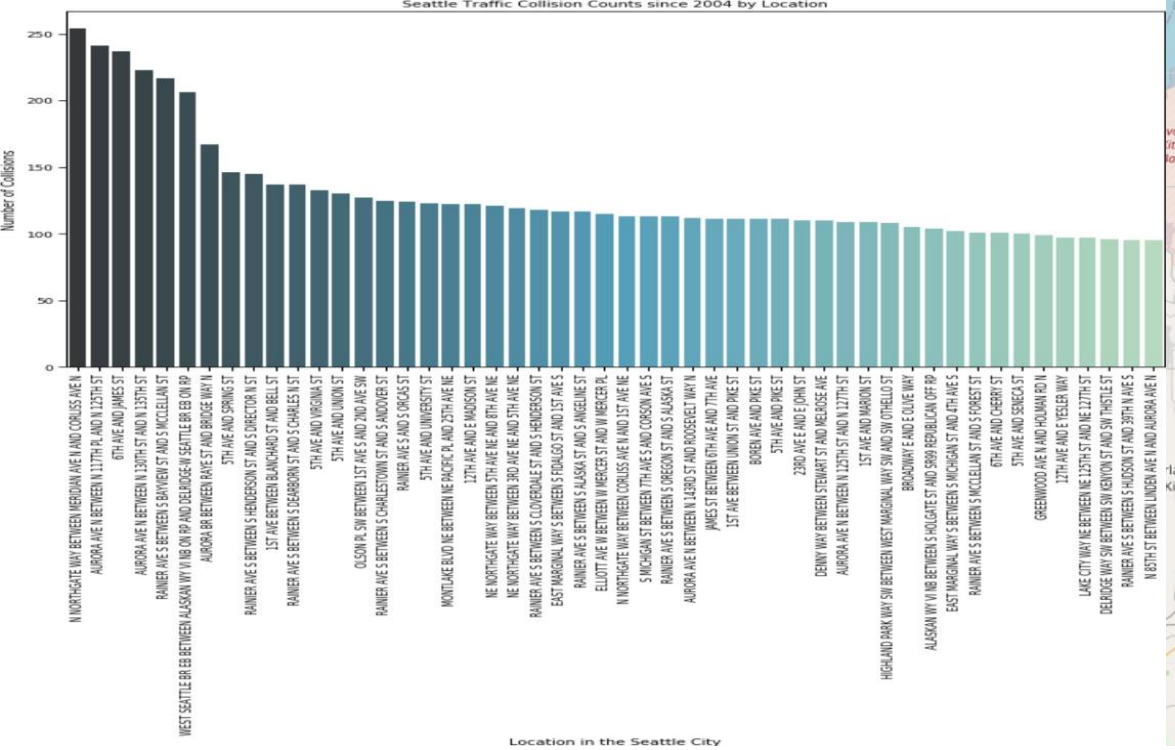

```
INATTENTIONIND:
 0    164868
1     29805
Name: INATTENTIONIND, dtype: int64
```

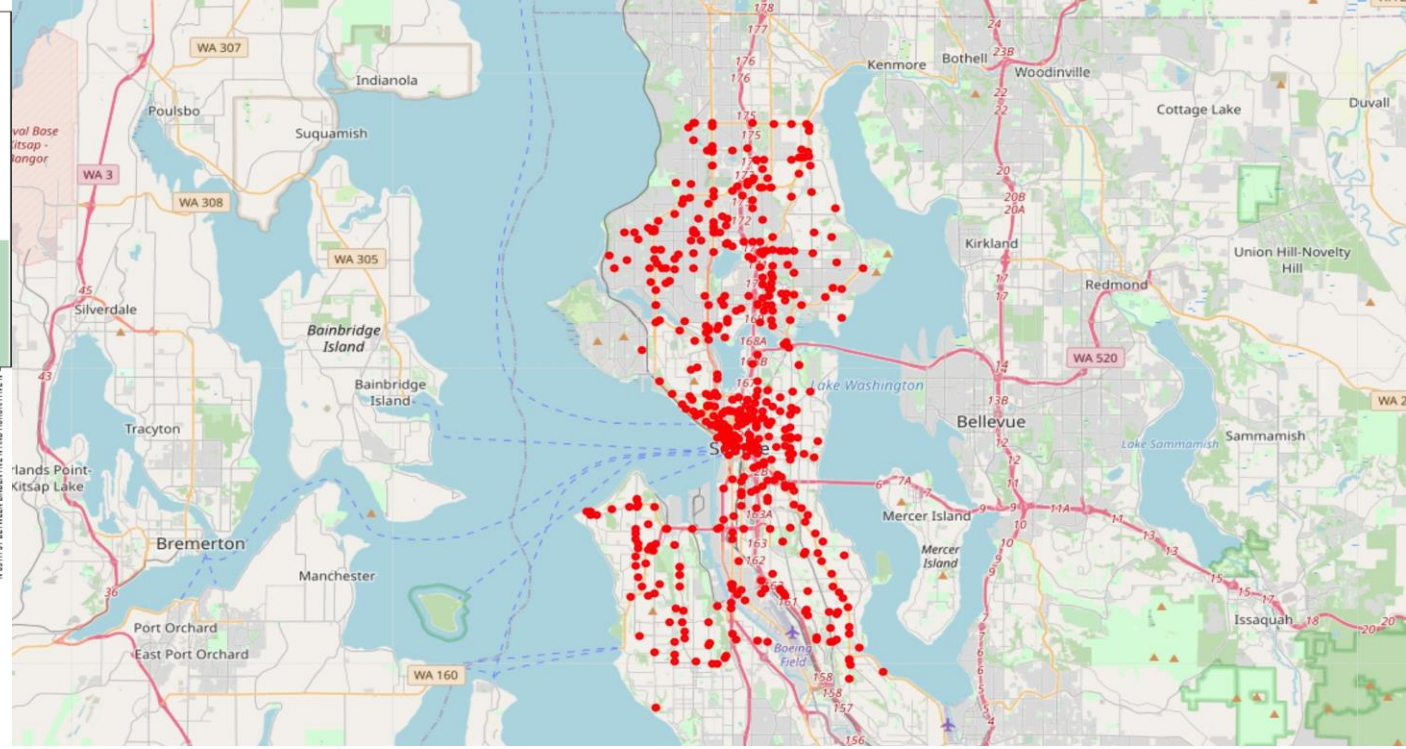


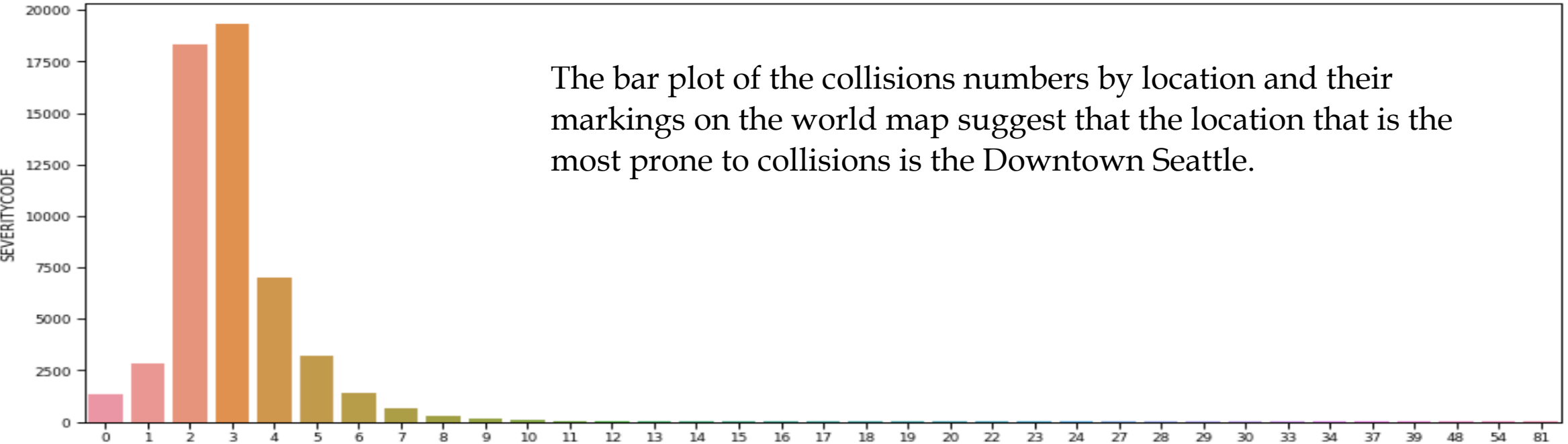



```
PEDROWNOTGRNT:
 0    190006
1      4667
Name: PEDROWNOTGRNT, dtype: int64
```

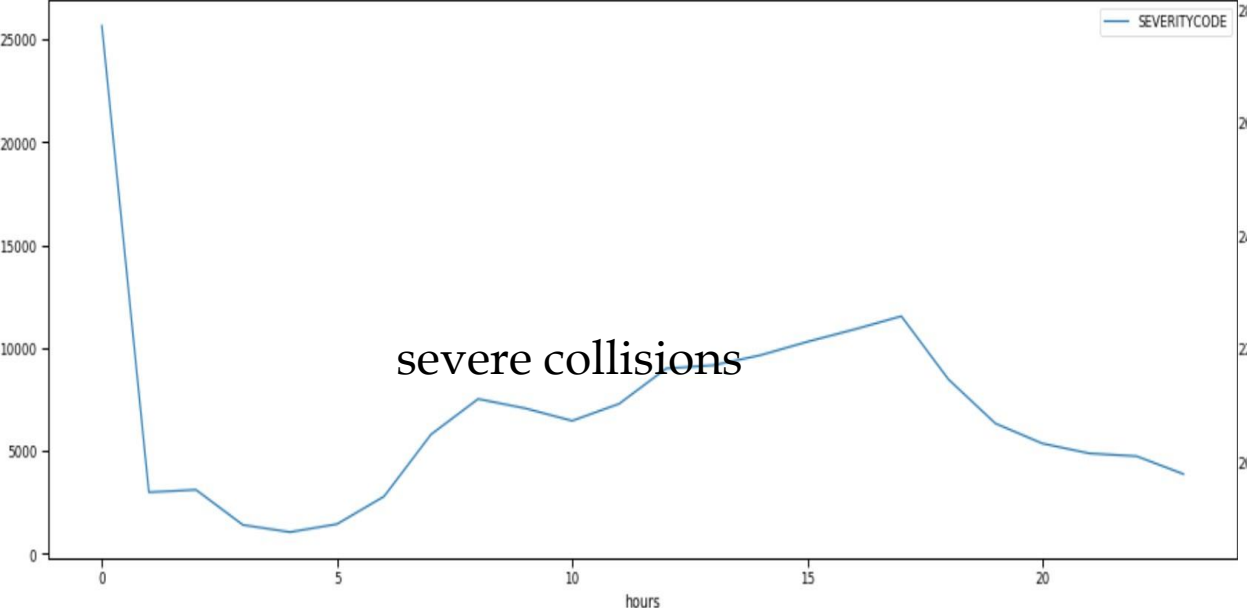




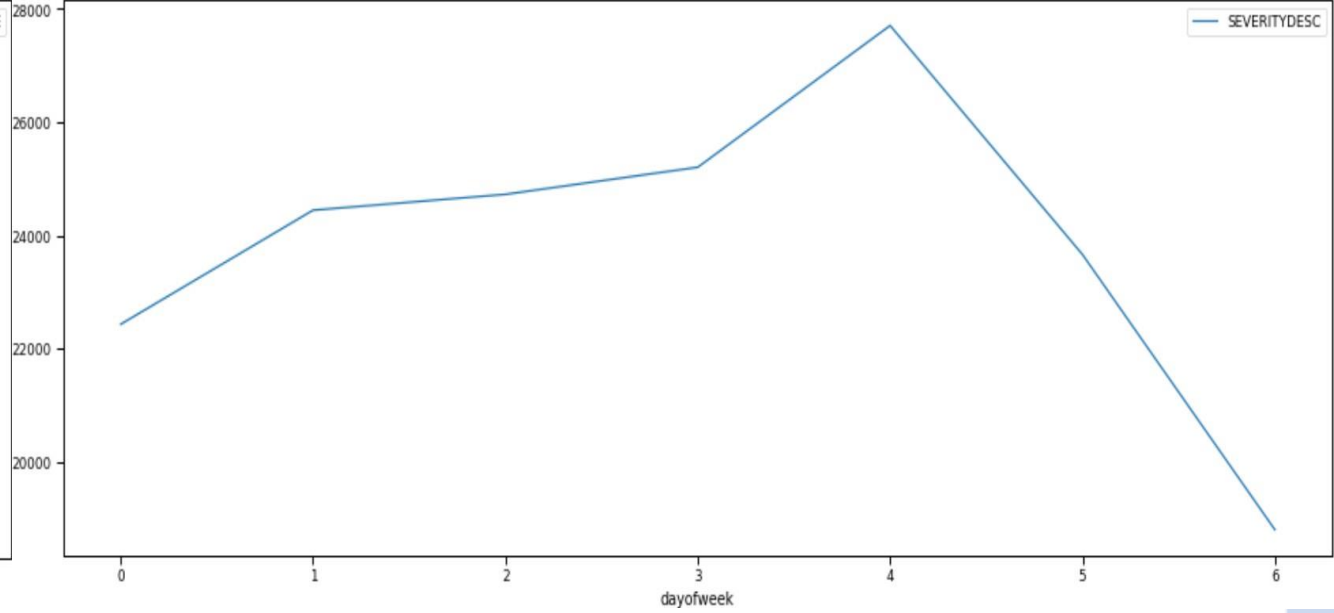


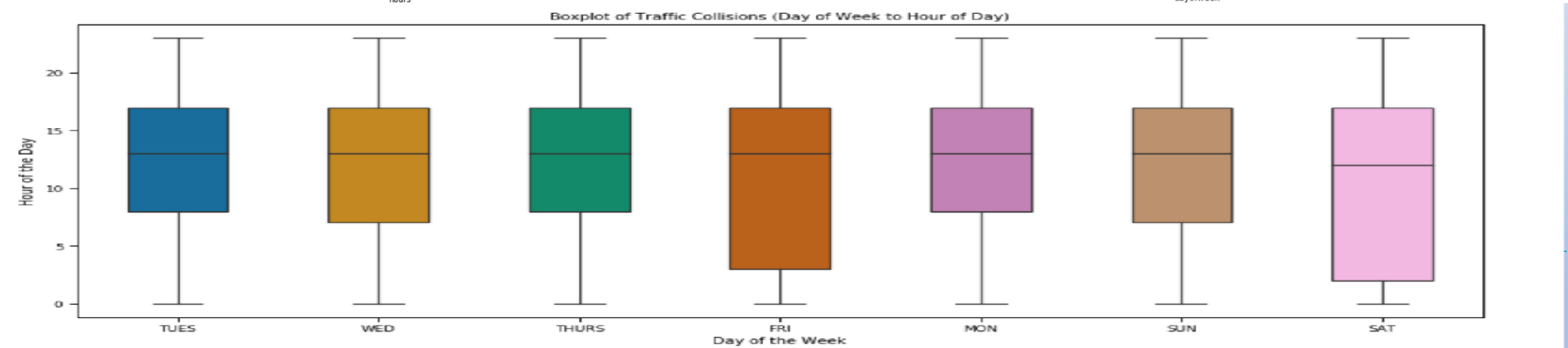


The bar plot of the collisions numbers by location and their markings on the world map suggest that the location that is the most prone to collisions is the Downtown Seattle.



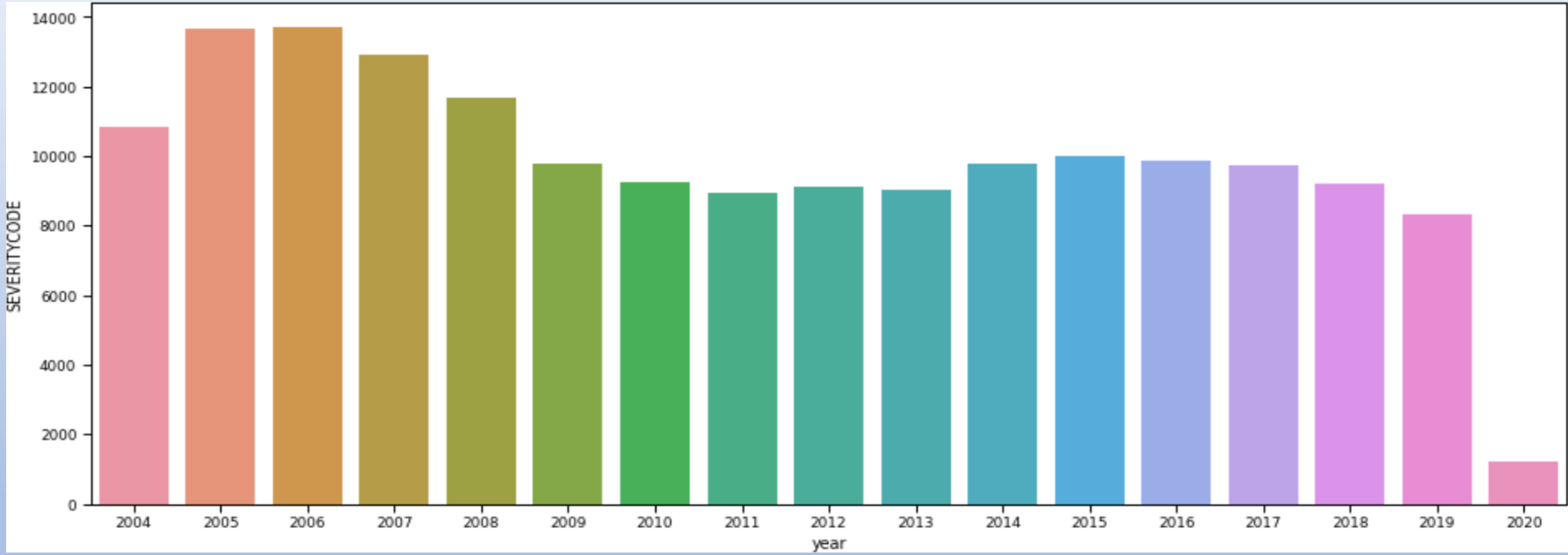
severe collisions



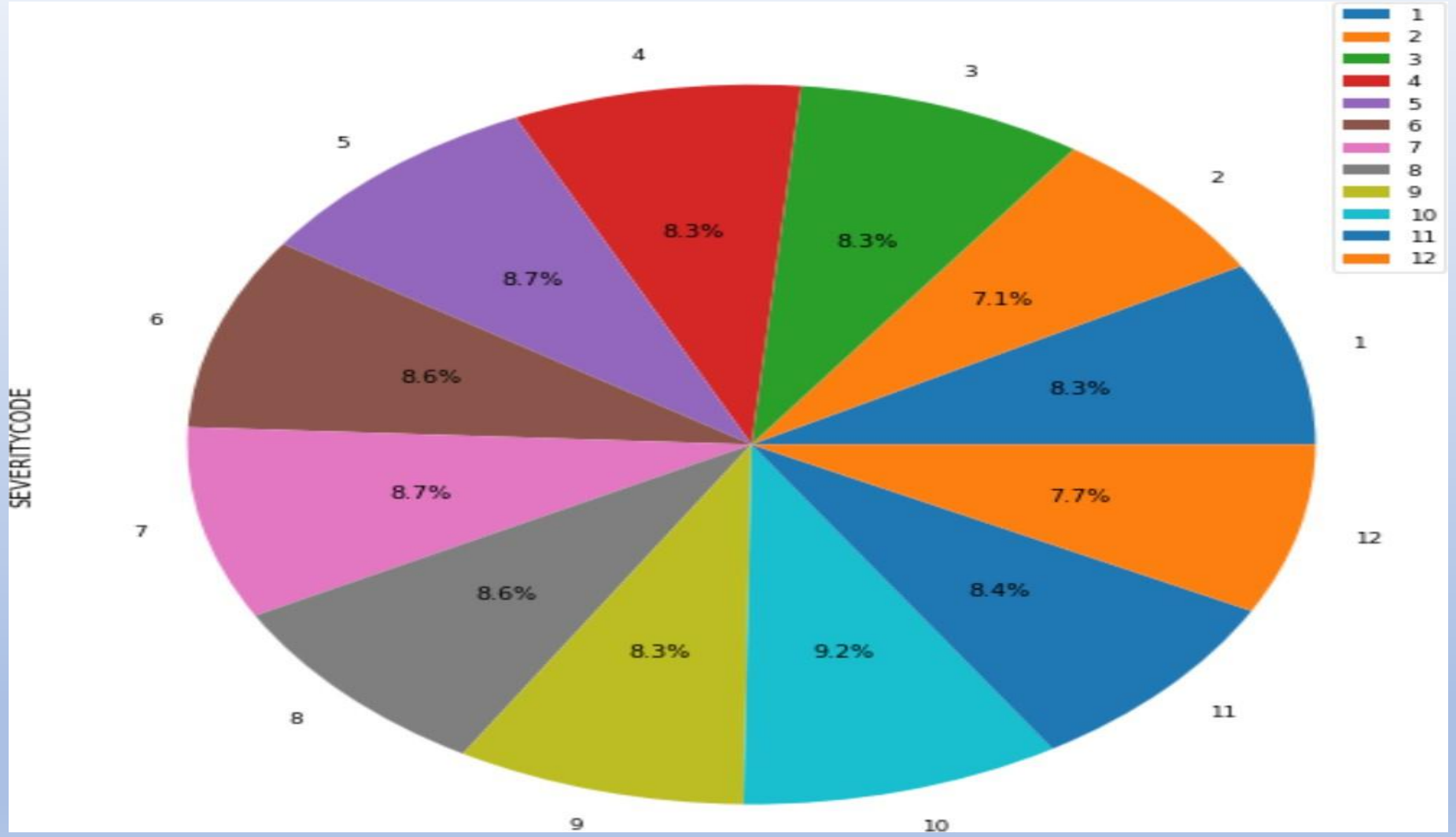


In the plot above, 0 : Monday 1 : Tuesday 2 : Wednesday 3 : Thursday 4 : Friday 5 : Saturday 6 : Sunday
Combining the day of week and the time when the number of collisions are high suggests that the highest number of collisions happen on Friday nights.

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

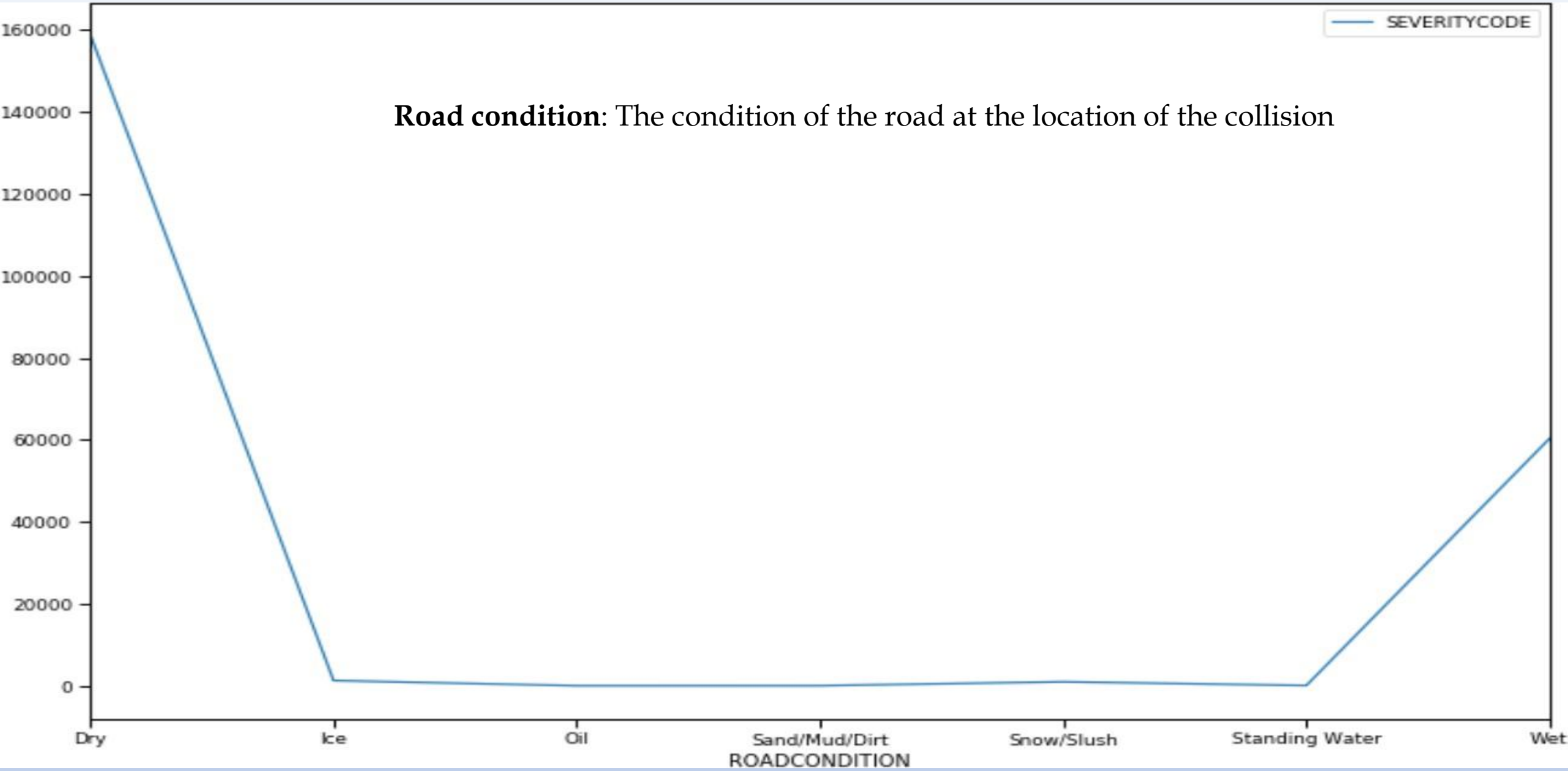


Monthly Trend of Collisions:

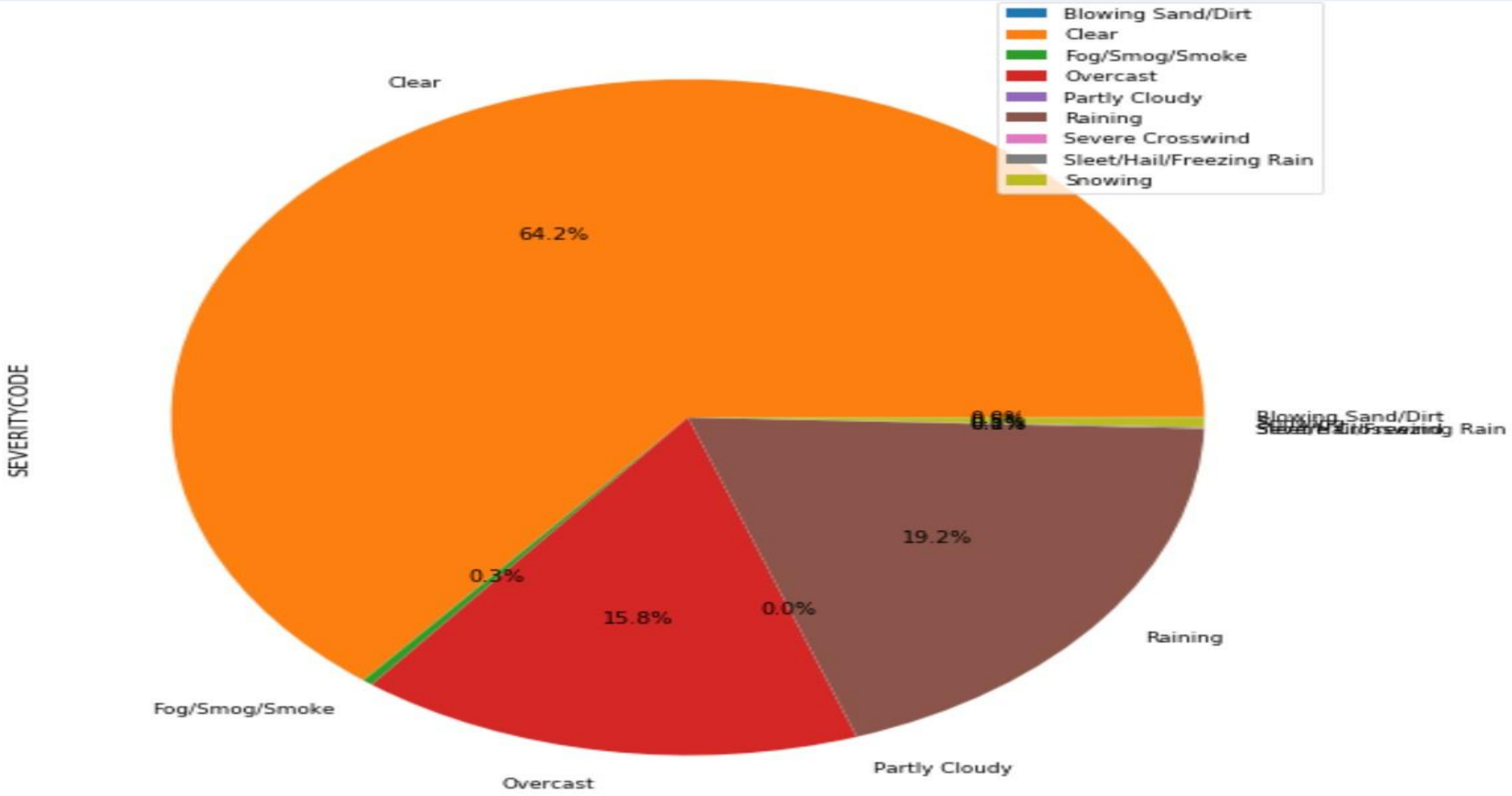


Road Condition Parameter

Road condition: The condition of the road at the location of the collision



Percentage of Collisions by Weather



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.

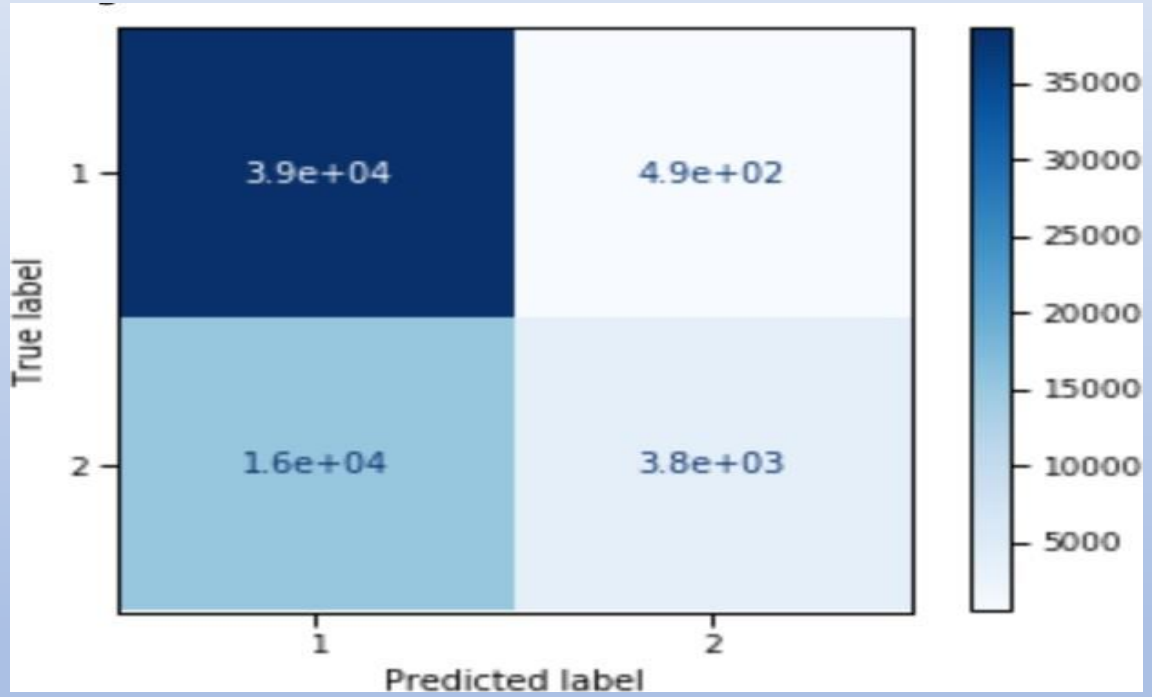
RESULT

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

```
[[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
```



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 analyzing 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