coursera

Applied Data Science Capstone: Car Accident Severity Prediction

Car Accident Severity Prediction: Background

According to a report by CDC in 2019, road traffic crashes are a leading cause of death in the United States for people aged 1–54 and the leading cause of non-natural death for healthy U.S. citizens residing or traveling abroad. WHO in a 2018 report profiled the following facts concerning global road traffic injuries and deaths;

- Each year, 1.35 million people are killed on roadways around the world.
- Every day, almost 3,700 people are killed globally in road traffic crashes involving cars, buses, motorcycles, bicycles, trucks, or pedestrians. More than half of those killed are pedestrians, motorcyclists, and cyclists.
- Road traffic injuries are estimated to be the 8th leading cause of death globally and the leading cause of death for children and young people 5–29 years of age.

Car crashes are a public health concern both globally and in the United States but, these injuries and deaths are preventable and the impact and consequences of these accidents can be minimised. Conventional techniques have been used in the past to predict the severity of clashes, though these had a number of drawbacks in producing quality and accurate inferences. This project aims to predict the severity of accidents and how the impact can be minimised based on a number of factors and for the purposes of this project we will use data relating to Seattle city.

The solutions seeks to provide aid to the Seattle's Department of Transportation (SDOT) in its transportation infrastructure planning, building and maintenance to ensure that it tailor makes its road networks in a manner that addresses the rise in accidents.

The solution will also help the Department of Health in planning for relevant equipment and resources required based on the predicted accident severities. This will help them in acquiring resources that appropriately address the problems and injuries on a particular accident scene.

Road users in general, that is, pedestrians and motorists will be advised accordingly if the information is publicly availed to Seattle citizens and passer-by's to take precautionary measures to reduce severity of accidents.

Data Acquisition and Cleaning

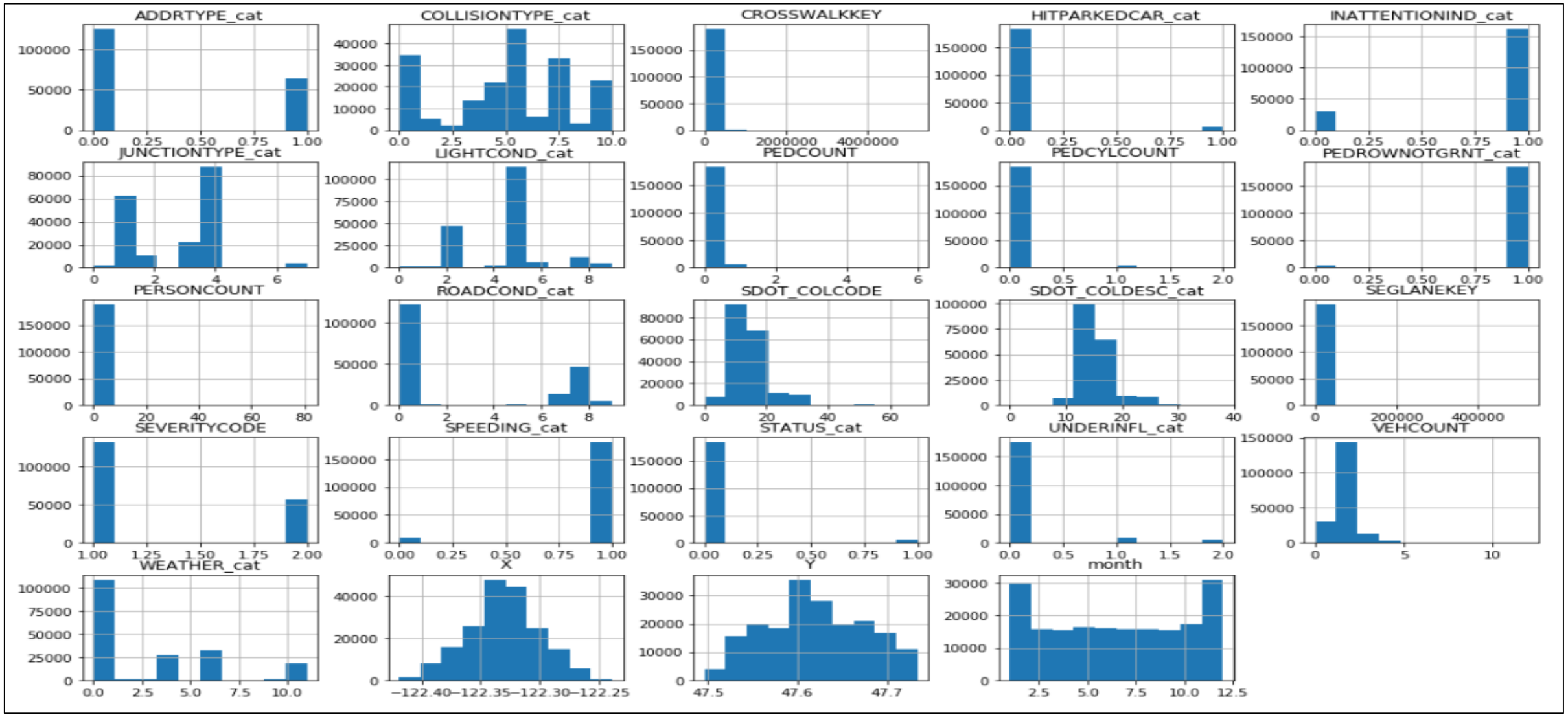
For the purposes of this project, the already existing Seattle City accident csv dataset will be used to predict the severity of an accident as such no further data gathering procedures or scraping processes were performed in an attempt to collect the data.

In general the whole dataset, shows the severity of an accident which we will in this particular case use as our label to predict using supervised learning. The model should be a more generalised state to allow it to be implemented in other similar instances avoiding over/under fitting problems. (SDOT) in its transportation infrastructure planning, building and maintenance to ensure that it tailor makes its road networks in a manner that addresses the rise in accidents.

The dataset contains ;

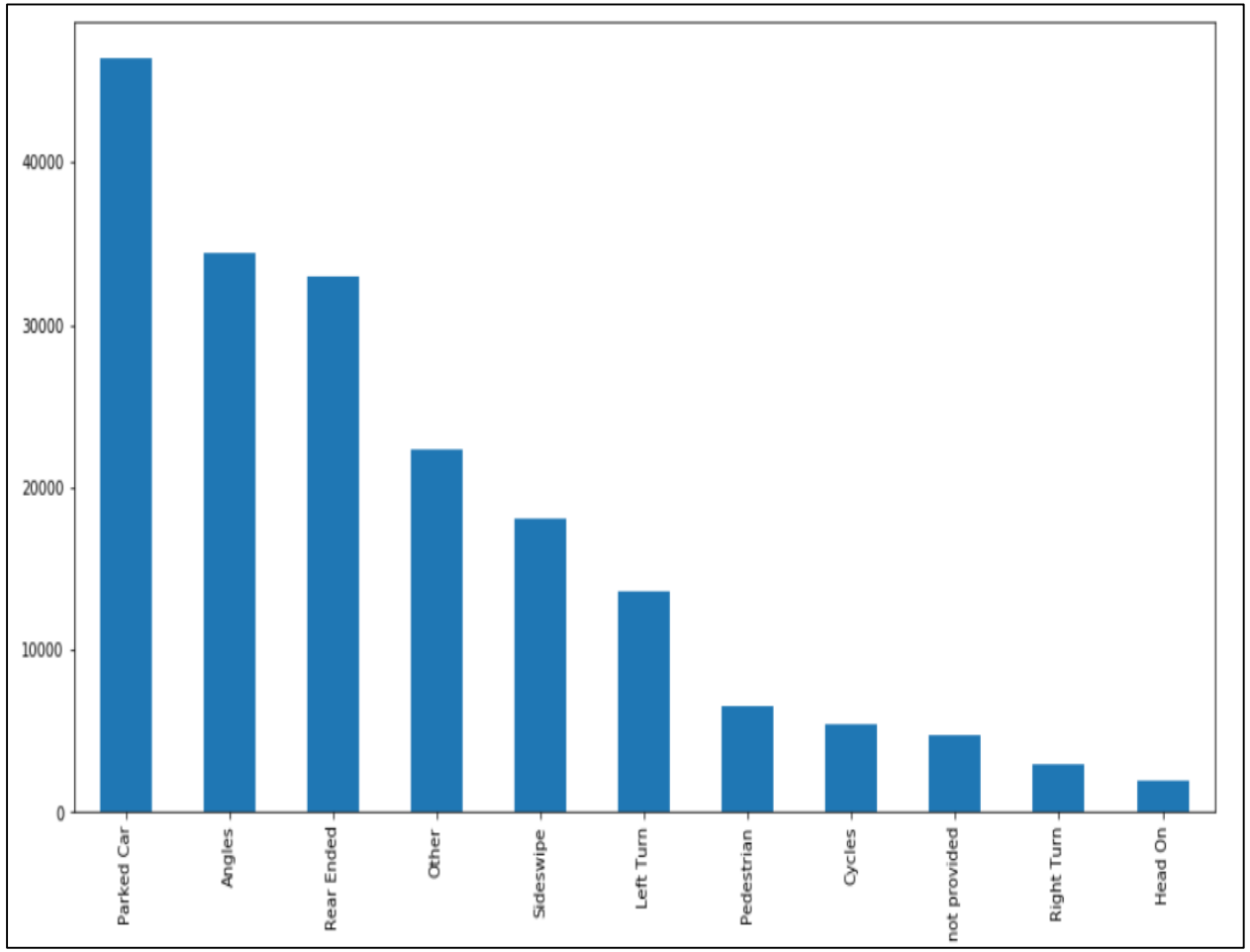
- 194,673 unique instances or rows and 38 features
- Cleaned data contains 21 features
- Duplicate and uses features were removed from the dataset
- Label Encoding was performed on categorical variables
- Null values for location data were dropped

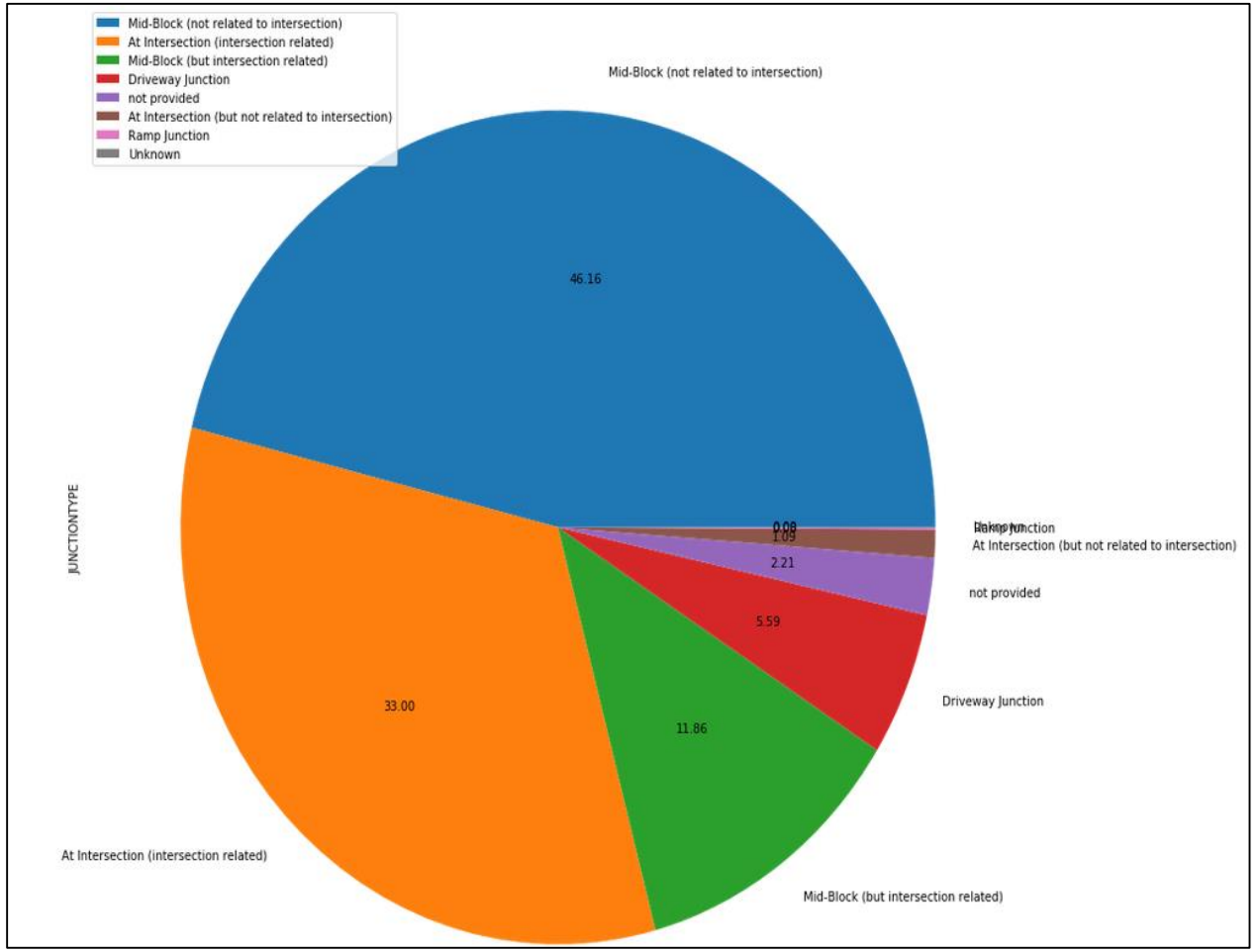
Exploratory Data Analysis



- Using the pandas histogram function to understand the distribution of data for the fields in the dataset

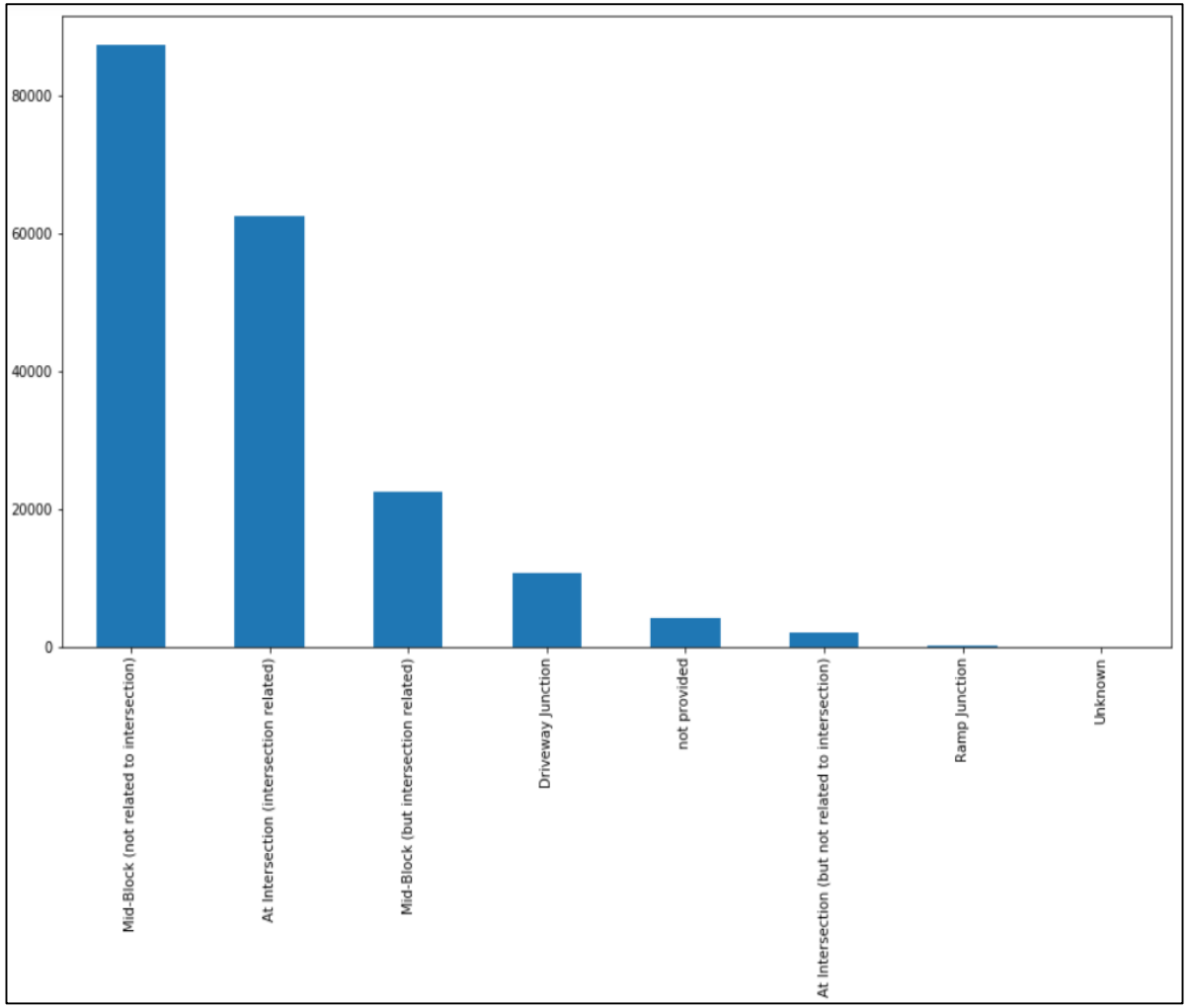
Exploratory Data Analysis

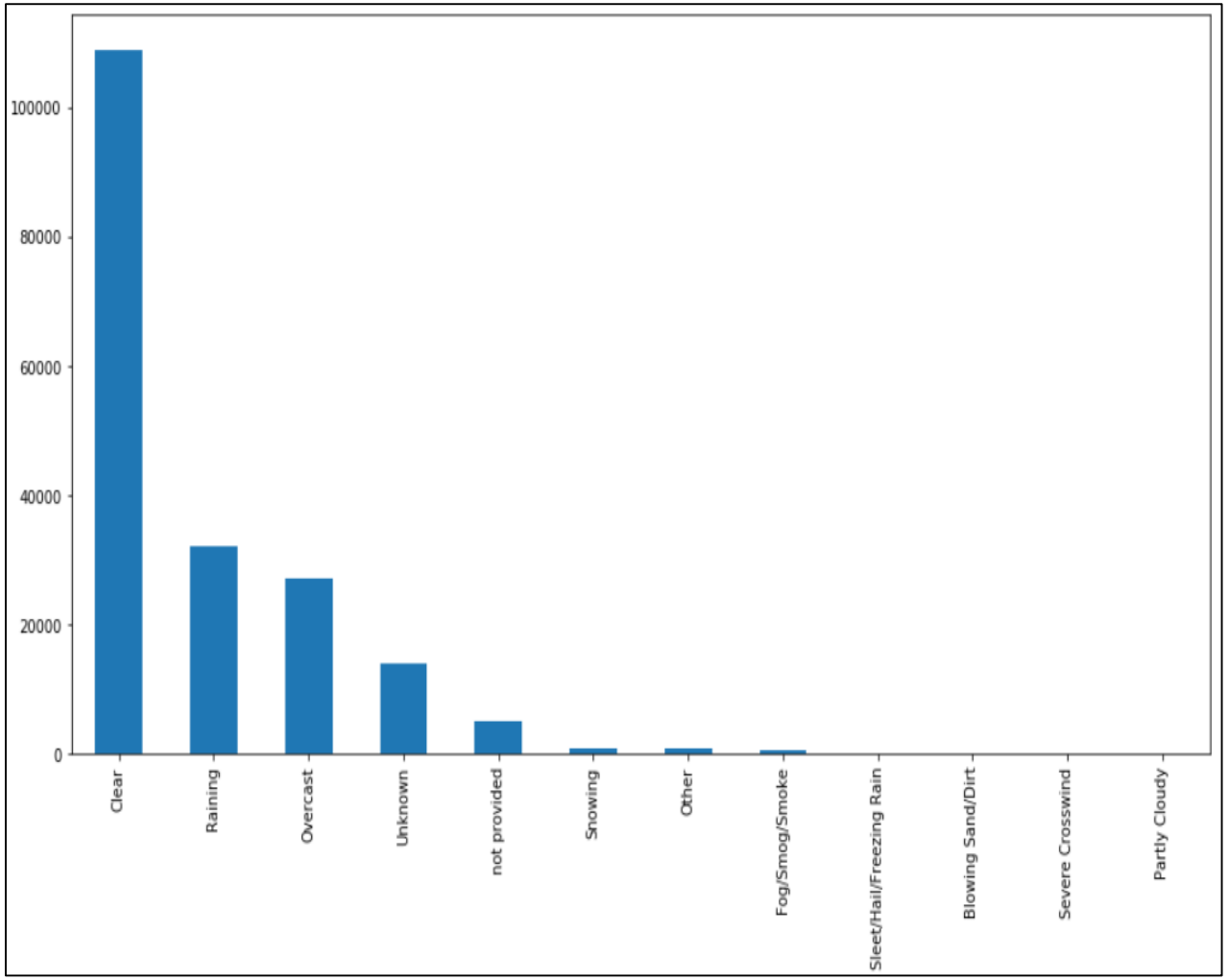




- The most common collision type was that with Parked Vehicles converting to about 25% of the total accidents.
- Mid-Block junctions not related to intersections resulted in 46% of the total accidents as compared to other junction types.

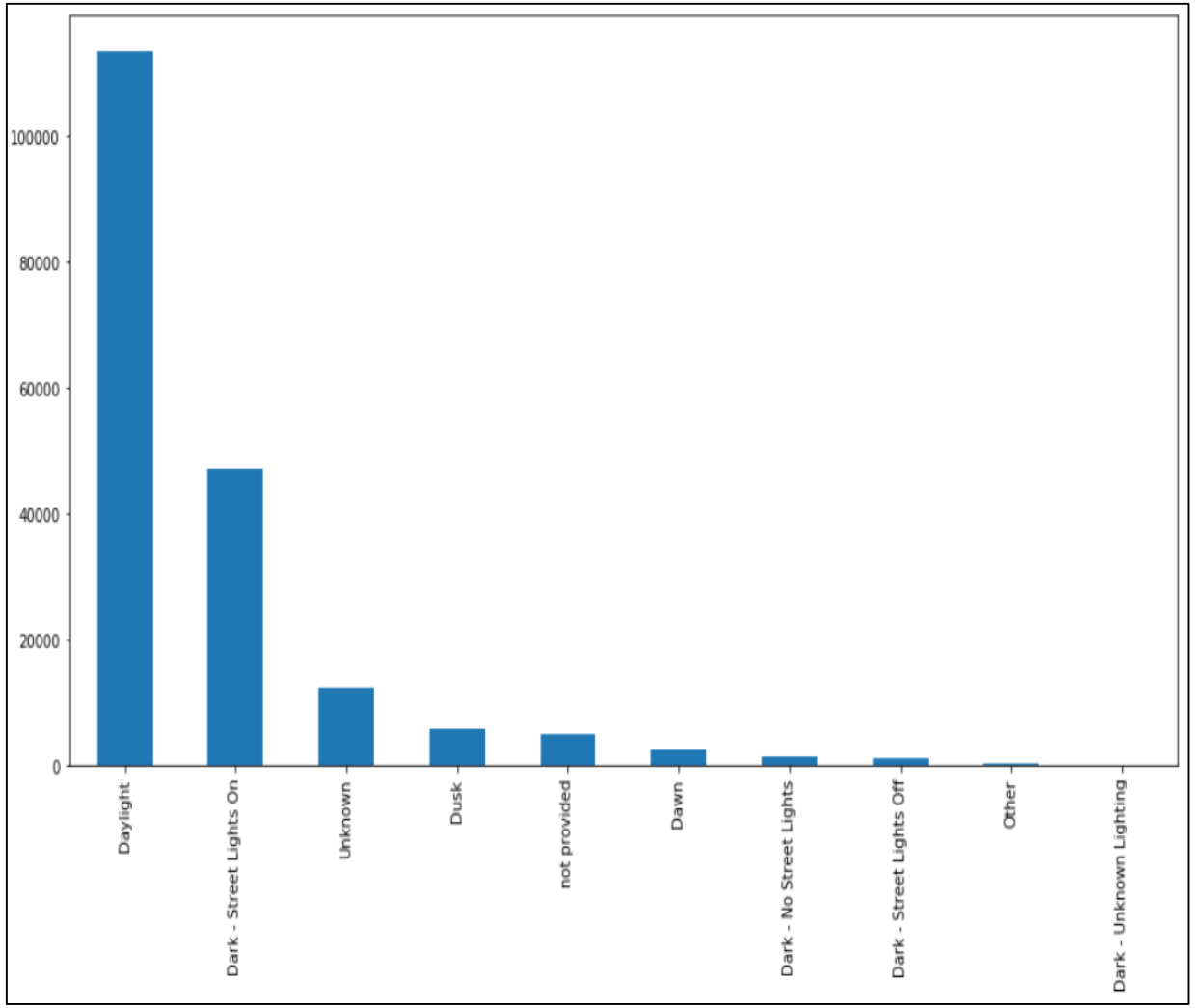
Exploratory Data Analysis

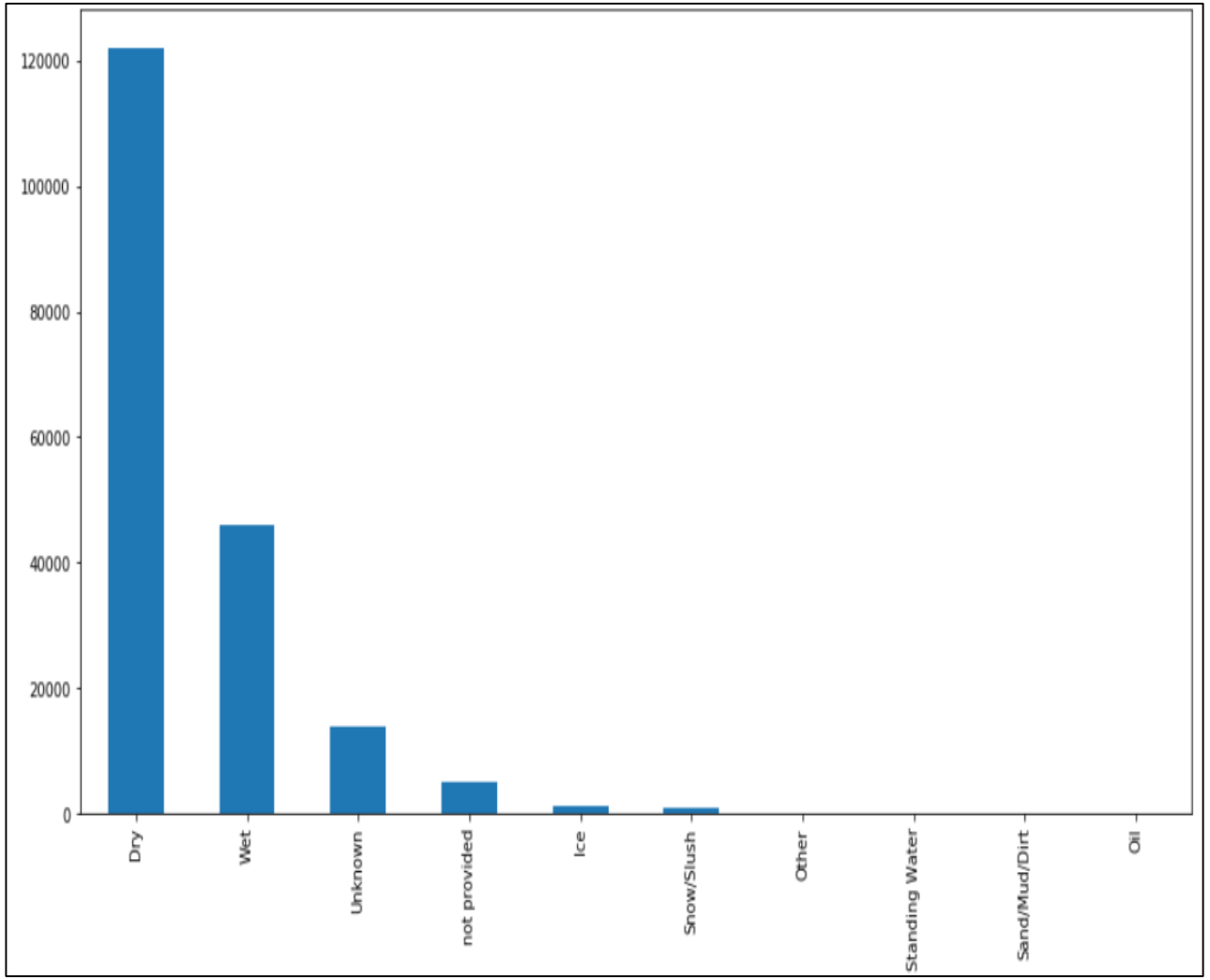




- It was observed that most accidents happened during times of clear weather which converts to about 58% of the total accidents.
- Mid-Block junctions not related to intersections resulted in 46% of the total accidents as compared to other junction types.

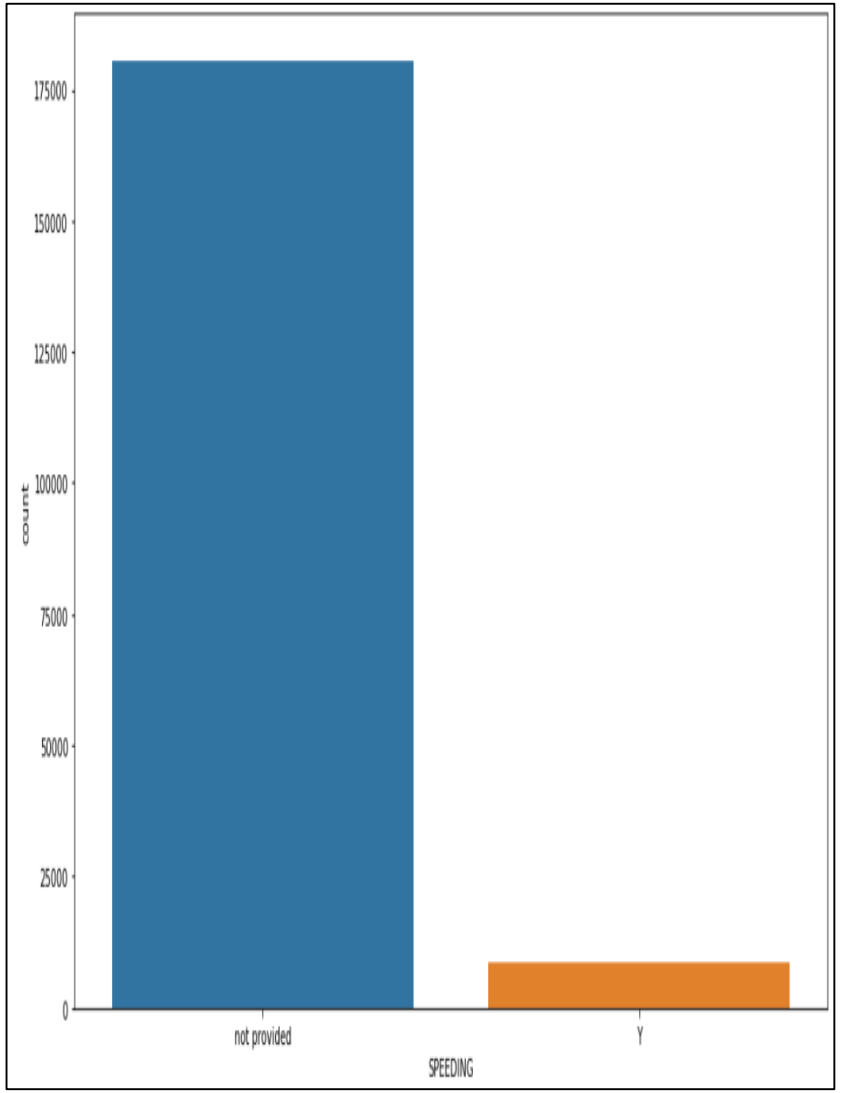
Exploratory Data Analysis

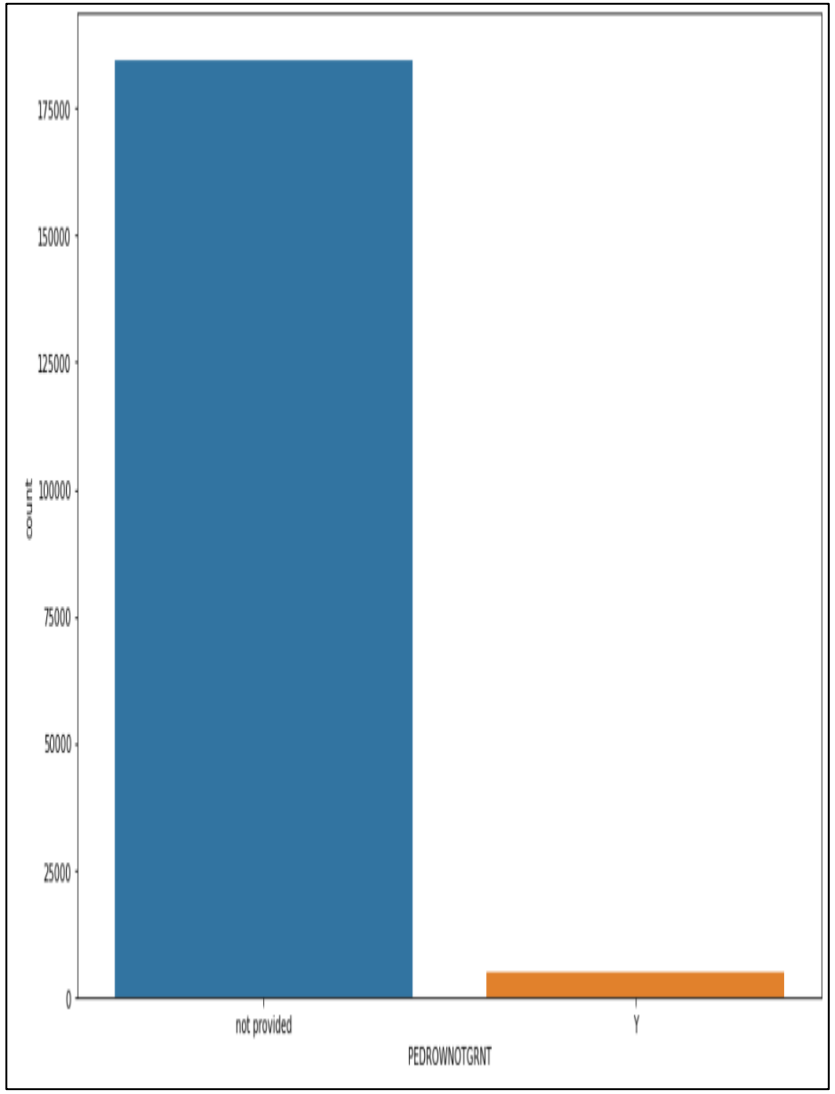


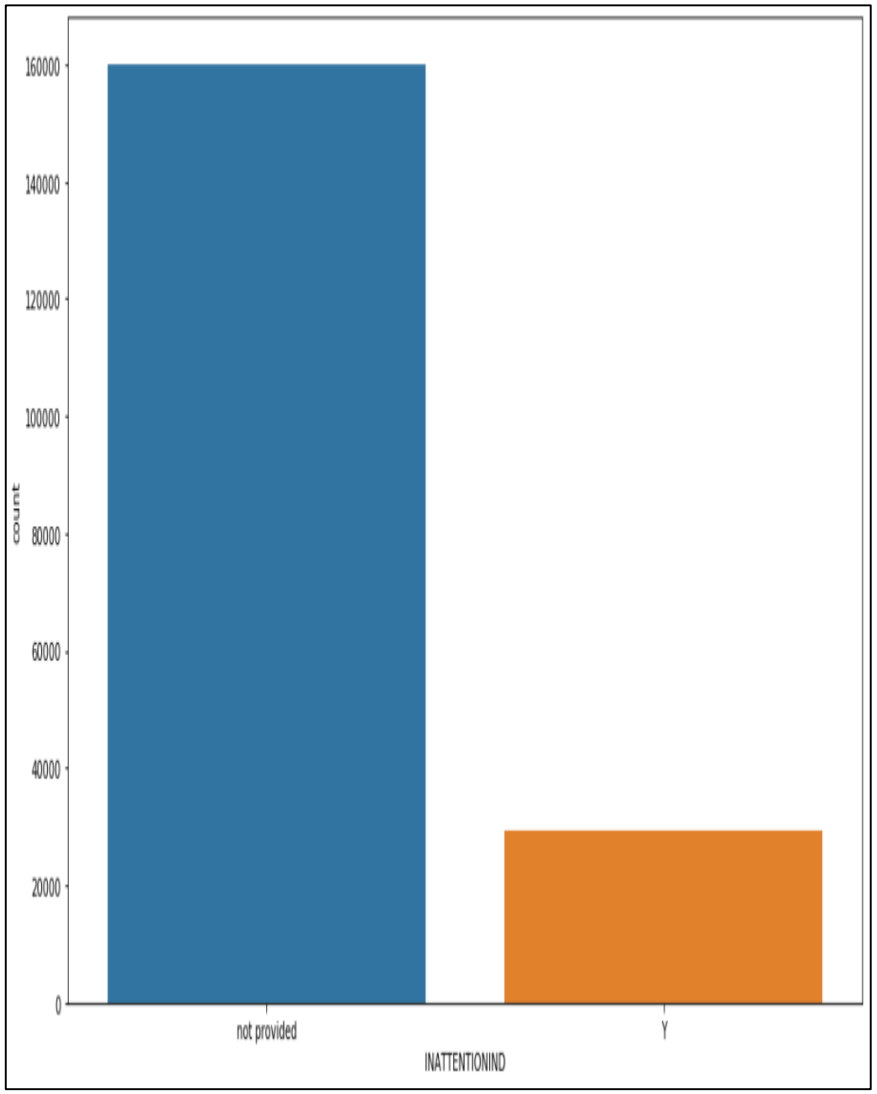


- Most accidents are taking place during daylight and when the road condition is dry. This is actually opposing the general notion that poor lighting and oily or wet roads cause accidents.

Exploratory Data Analysis

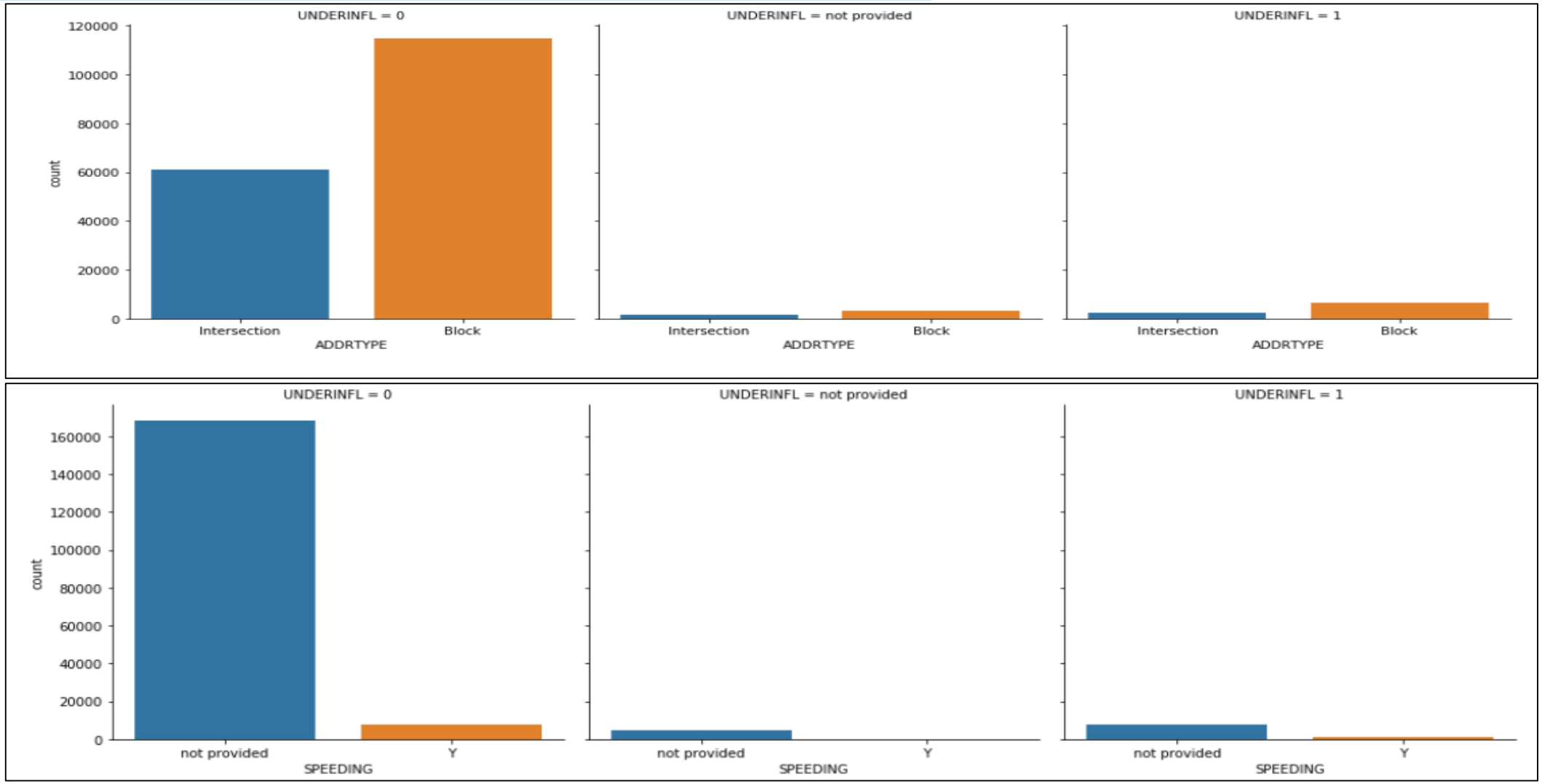






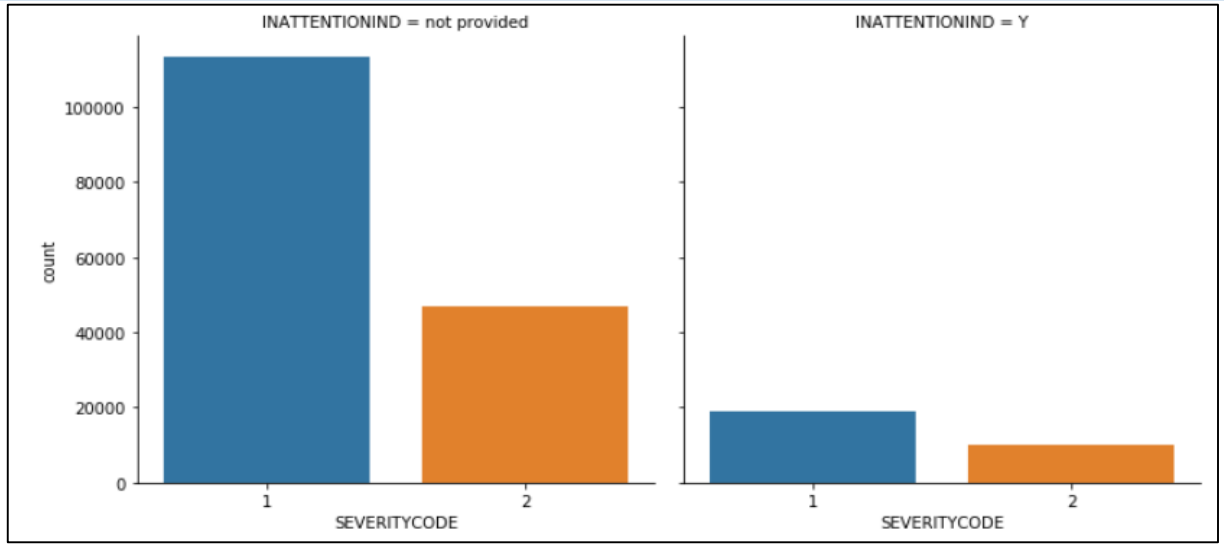
- It was observed that Speeding, not granting pedestrians right of way and inattention are not directly linked to most accidents happening in Seattle.

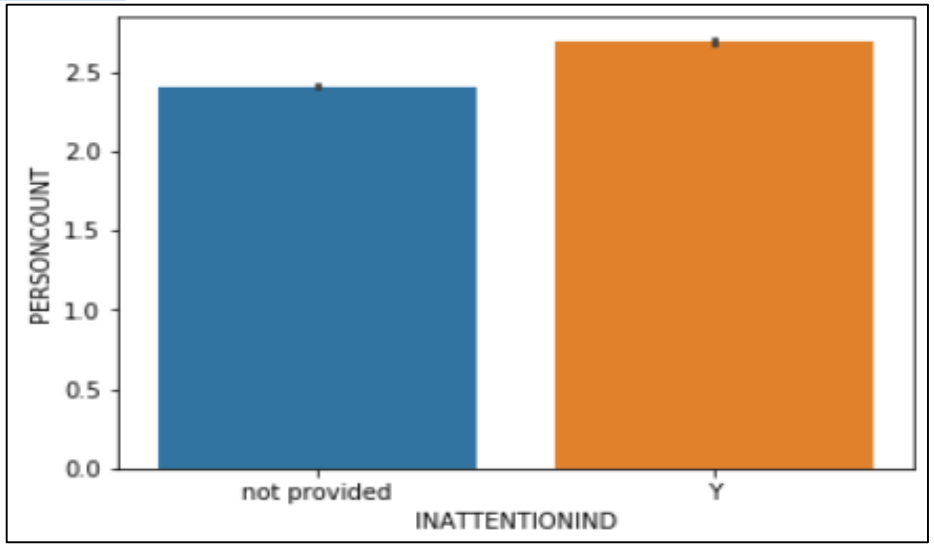
Exploratory Data Analysis

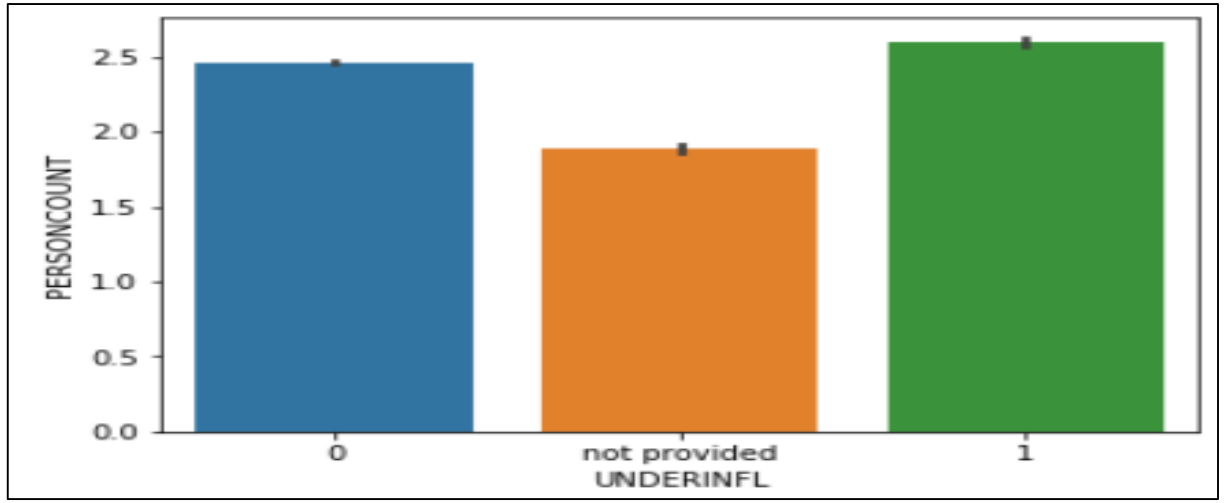


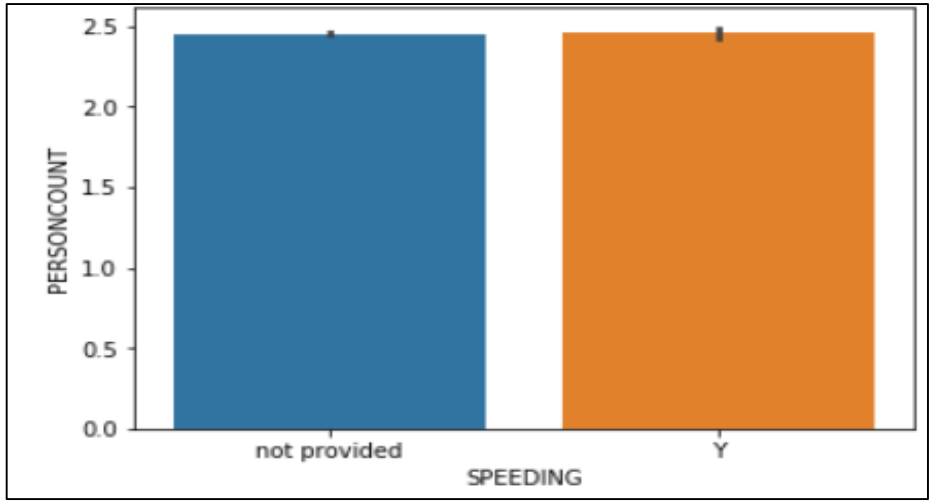
- From the data provided those who are not under the influence of alcohol or drugs tend to be involved in block accidents
- From the data provided most of the non speeding drivers were not under the influence of drugs or alcohol

Exploratory Data Analysis



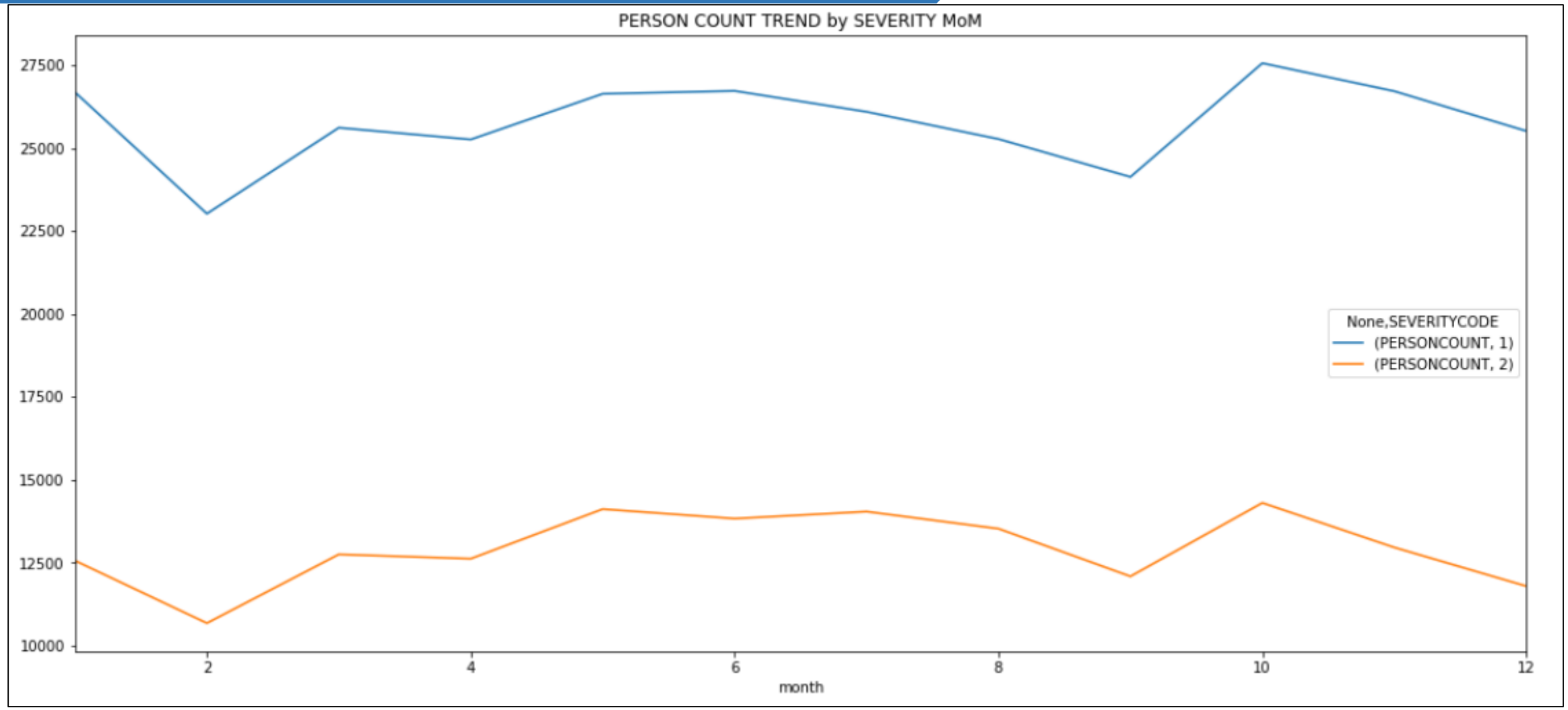






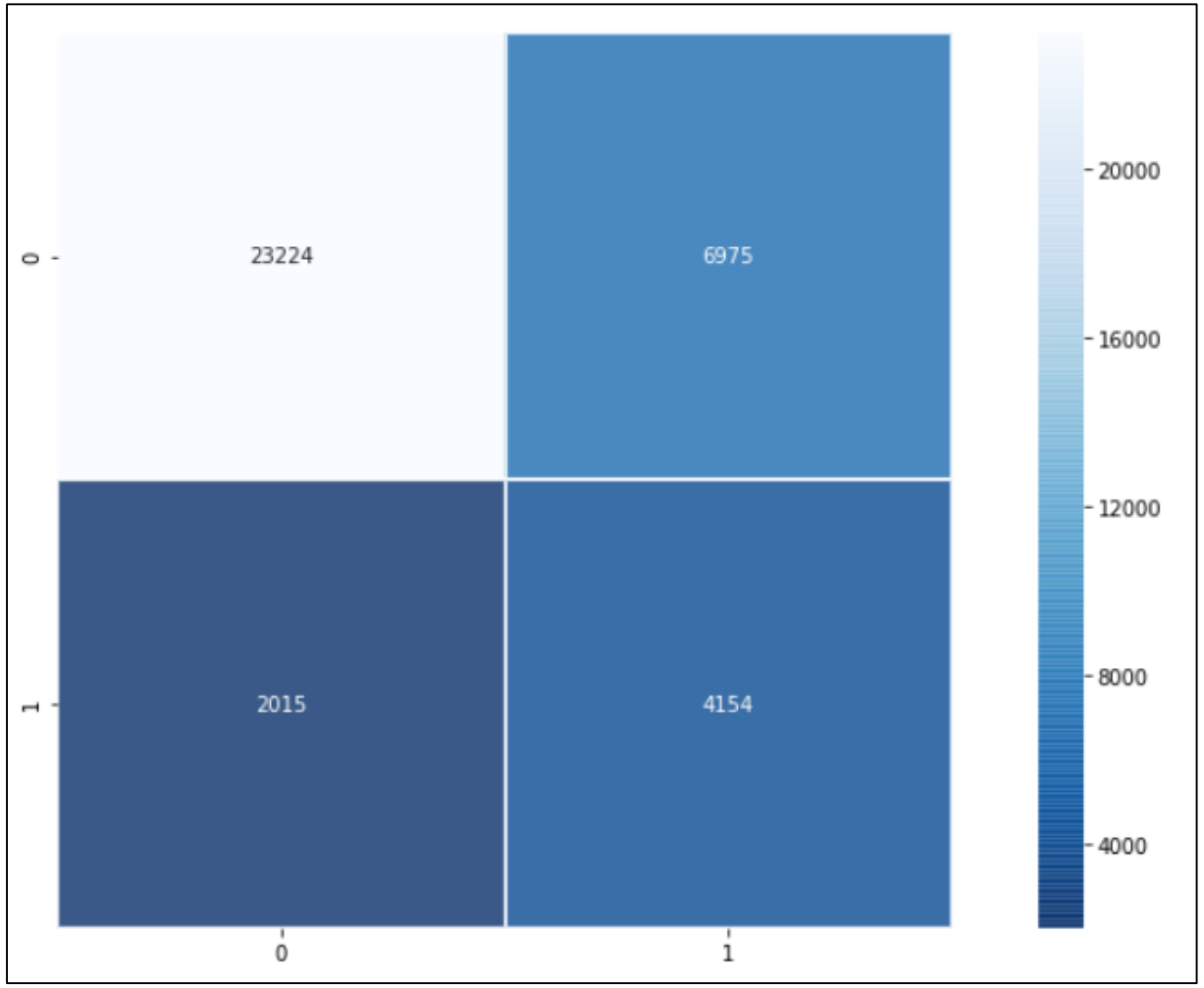
- It was observed that Inattention, driving under the influence of alcohol or drugs and speeding results in a number of people involved in an accident. However, these are not directly linked to the number of accidents that happen. It was observed that most injury collisions happen when motorists are well attentive as compared to when they are not paying attention. On the other hand, during times of inattention injury collisions do happen more as well.

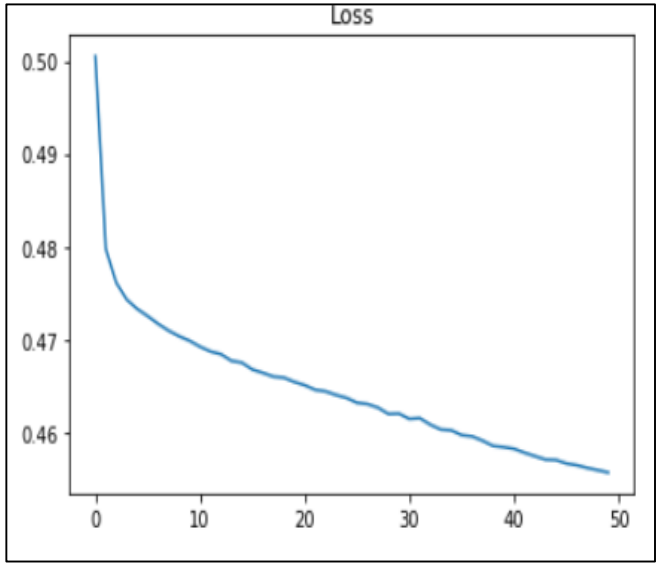
Exploratory Data Analysis

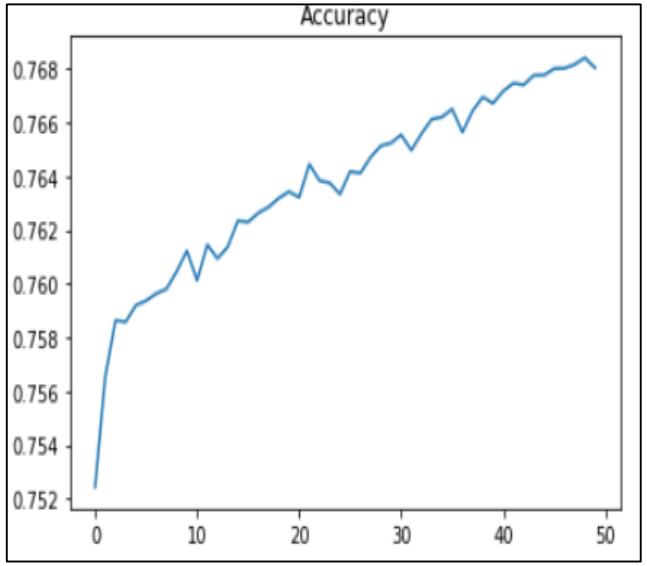


- It can be observed that people involved in collisions in both injury and property damage collisions do spike on the 10th month of the year. Also observed was that property damage accidents have a larger probability of involving more people than injury collisions.

Results and Model Evaluation







	precision	recall	f1-score	support
0	0.77	0.92	0.84	25239
1	0.67	0.37	0.48	11129
micro avg	0.75	0.75	0.75	36368
macro avg	0.72	0.65	0.66	36368
weighted avg	0.74	0.75	0.73	36368

- The classification algorithm retained an accuracy of 75% having a total of 27,378 true predictions out of a total number of 36,368.

Conclusion

The purpose of the project was to predict the severity of an accident based on the accident data for Seattle City. By applying supervised learning in the form of a multi-classification algorithm, we were able to able an accuracy of 75% which might indicate that the algorithm to a larger extent will be able to predict and help various stakeholders in planning against these accidents. The end result we all wish for will be a decline in the rate of these accidents. Exploratory Data Analysis apart from the model itself was done to give a clear picture to the authorities and various stakeholders on the certain conditions that are mostly resulting in accidents. Accuracy of the model has room for improvement.

There might be need to include more variables for model performance such as;

- Vehicle Type (Model and Make)
- Speed at the time of accident
- Gender of the driver