Introduction

For many travelers, having the freedom to explore the city to experience and enjoy various aspects is a great way to relax and enjoy weekends. Similarly, traveling via a vehicle, cycling, or walking are the most prominent ways for one to journey throughout the city for your everyday errands. Due to the large population that uses public roads, safety has become a top priority for the municipal government to ensure they are able to reduce the number of accidents that occur. Therefore, analyzing the various factors that could help predict accident severity can guide the board to implement changes in a timely manner that may reduce the number of fatalities & serious injuries. Consequently, this will ultimately lower the economic costs that would trickle into other fields within the economy.

Business Problem

The objective of this capstone project is to analyze the collision data set for Seattle, WA and determine the most pertinent factors including weather, road conditions, visibility, and various other factors that best predict accident severity. Using various analytical techniques and machine learning algorithms such as logistic regression. This project will be used to answer the business question: How can the city of Seattle, Washington best predict the severity of collisions that occur?

The intended audience for this project will be the Department of Transportation of Seattle, Washington. Due to the danger of vehicle collisions, providing solutions that may reduce the amount of accidents can significantly improve the quality of life of pedestrians & overall ensure public safety. Similarly, companies whose business model rely on customer transportation can benefit and alter their algorithms to avoid certain high risk areas.

Data

The data that will be used to conduct this analysis is the compiled collision dataset from 2004 to Present within Seattle, Washington. This data source includes 194,673 rows and 38 columns was taken from the Seattle Department of Transportation that is continuously updated weekly. This will include the severity of each accident, the type of vehicle involved if any, the location of where the collision took place, as well as weather conditions that may have had an adverse effects on the event. Specifically, weather conditions, light conditions, and types of junctions can collectively be used to determine the most dangerous parts of the road where severe accidents are most likely to occur. Most importantly, it contains a severity code that ranges from 0 (unknown) to 3 (fatality) that will be the main focus of this analysis. Being able to use the various features within the dataset to better predict this the level of severity of the collision can allow proper safety features to be placed in certain junctions where the probability is higher.

Methodology

I performed several analyses including:

- Created a visual heat map of all of the Nan values present within the database to see how much is missing and what pieces of the data is useful.
- Used bar graphs to visualize the difference between the relationships of various factors to the severity of the corresponding collisions.
- Gave each categorial value a numerical placeholder to facilitate the function of further analysis.
- Devised a regression analysis to predict the severity of future collisions based on the given data.

Discussion

Having this big of a dataset definitely helps to create a visual of various factors that normally involve collisions. With this dataset specifically, it was possible to predict the severity of each collision using a regression model that gave a score of 0.74. However, this set came with various abnormalities. The first one was the unfortunate large number of missing data that caused several columns that would have been useful to the analysis to be completely dropped as seen in figure 1.

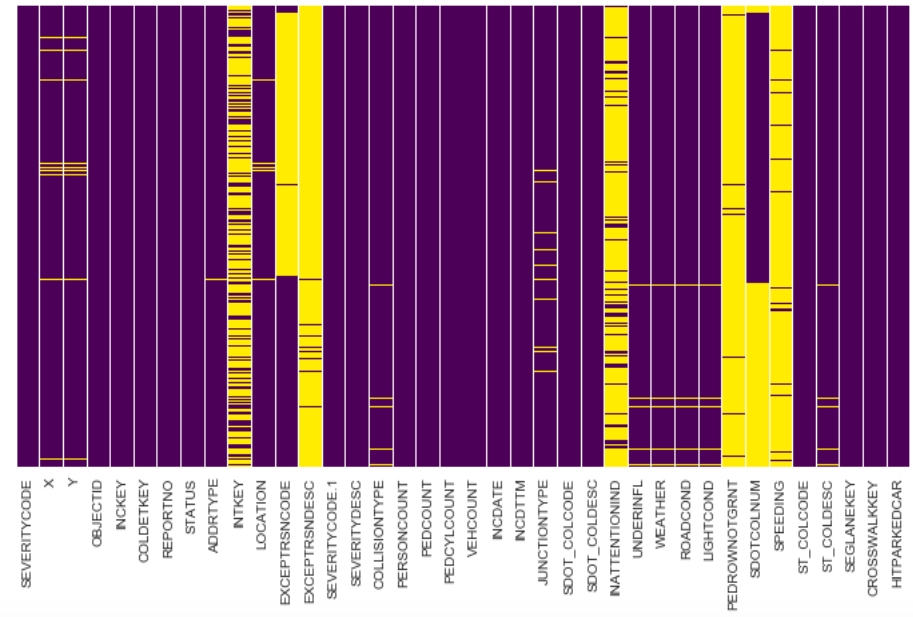

Figure 1 shows the empty or NaN values in yellow and filled values in purple for each column.

Secondly, the set was supposed to have collisions with a severity code the ranges from 0 to 3, each number signifying the outcome as follows: 0 – Unknown, 1 – Property Damage, 2 – Injury, 2b – Serious Injury, 3 – Fatality. In this case, the set only had severity codes for property damage and injury which fails to fully cover the range that it needs. Therefore, this limitation directly affects this analysis in the sense that it will ultimately only predict the severity for these two categories so a holistic overview is not ultimately possible.

Several factors were looked at to be able to determine the severity of collisions. One particular feature was the weather conditions during the time of the accident. I gathered the data and created a bar chart to visually see the differences between each within figure 2.

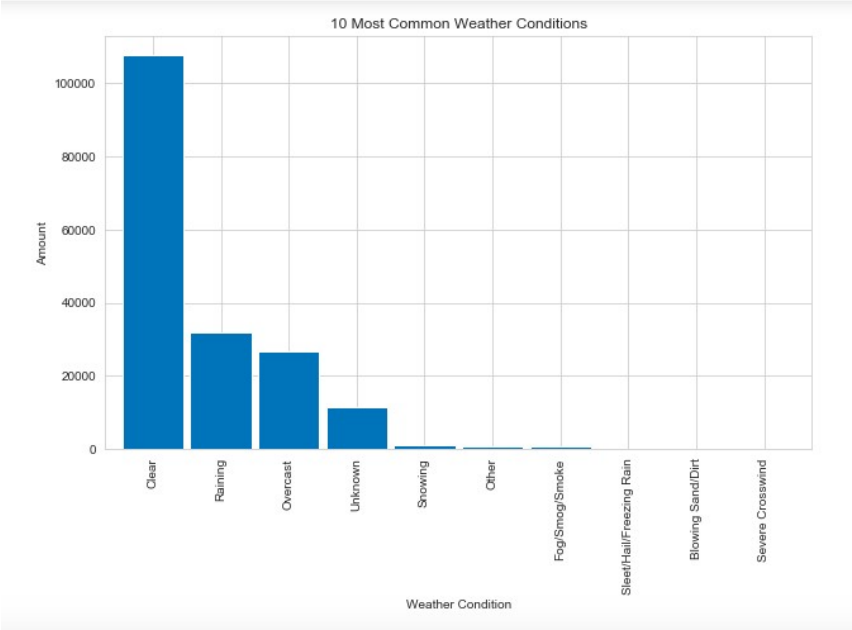

Figure 2 shows the relationship between the accidents and the weather present.

It was evident that clear weather was the most prominent weather condition within the majority of the accidents. Consequently, this rules out the initial hypothesis that weather has a major impact on the conditions that cause collisions but can potentially play a minor role. Equally important, was the analysis of the frequency of accidents on a weekday basis. The relationship for this was shown in figure 3.

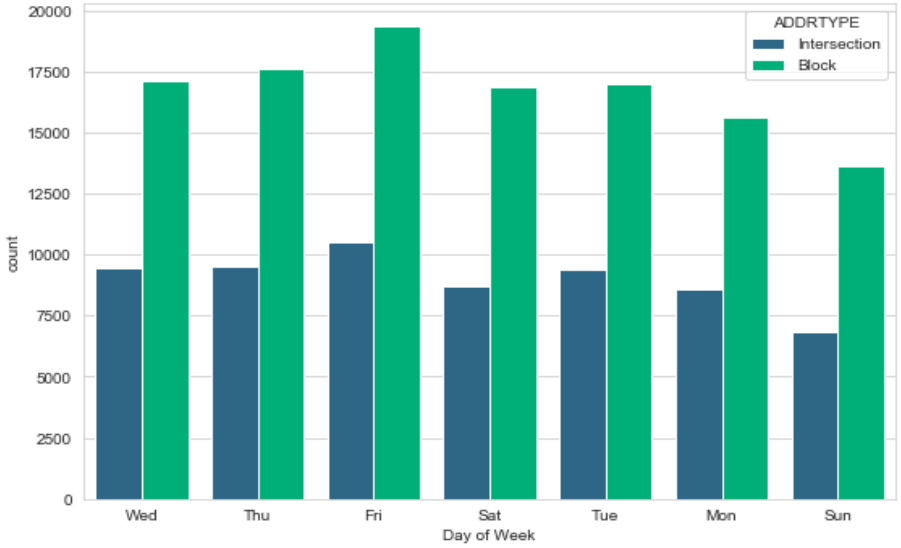

Figure 3 shows the frequency of accidents throughout the week with an emphasis on the type address type (Intersection & Block)

This graph clearly shows that common trend that was expected, higher collision rates on Fridays, the usual day the public decides to socialize in public settings thus increasing the population density in these areas. Similarly, the relationship between severity and collision type also reaffirms the previous graph. Within figure 4 the collision type and severity is seen.

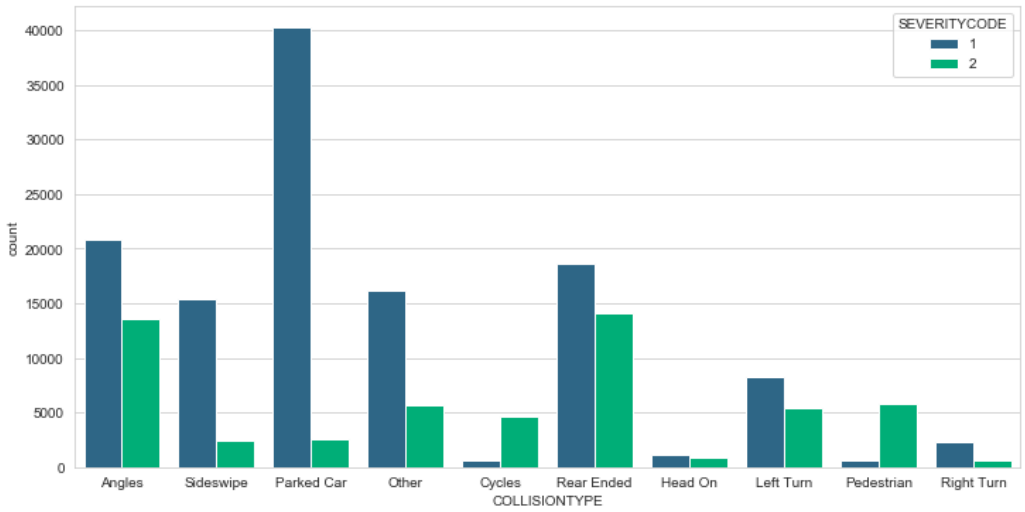

Figure 4 shows the number of collision types with an emphasis on the severity code of each.

The majority of the incidents appear to come from parked, angles, and rear ended accidents. All of which occur within a block setting as opposed to intersections. The one outlier being the Pedestrian column but the result is expected. Furthermore, the gaps for the rest of the features are large. This is primarily due to the difference between the severity present within the dataset. Mathematically, there was 2.33x more property damage than personal injury. Nonetheless, a logistic regression analysis was still able to be performed using the data that was left to decently predict the severity of future collisions.

Conclusion

There were several interesting relationships that came about during the analysis that can help shed light on the subject. Collisions that occur on the street as opposed to an intersection are far more prevalent. This could be due to complacency on the drivers side since only two opposing lanes are present. Similarly, nighttime collisions are still prevalent even in well lit areas coming in second behind clear days. Using a regression analysis worked perfectly for this dataset due to the binary characteristics of the severity codes present. The model was able to predict the results with a score of 0.74 and a weighted average of recall and precision of 0.74 & 0.75 respectively.