Seattle: Understanding The street design, Driver Behavior and the Surrounding Environment Contribute to Crashes

1 Introduction / Business Problem

All around the world, roads are shared by many motorized vehicles that have made transportation faster and more comfortable while supporting many countries' economic and social development. However, these vehicles cause a global problem. Car accidents are responsible for 1.35 million deaths on roadways every year. Almost 3,700 people are killed globally in road traffic crashes, where more than half of those killed are pedestrians, motorcyclists, and cyclists.

While Seattle has seen a 30 percent decline in traffic fatalities over the last decade, traffic collisions are still a leading cause of death for Seattle residents age 5 to 24. Older adults are also disproportionately affected, so this trend could grow as the population ages. To supplement the findings of the City's Bicycle and Pedestrian Safety Analysis project and provide policy makers and engineers with actionable information for developing and implementing interventions, we need have a deeper study for Collision data from Seattle.

The objective of this project is to define the problem, to find the factors that can have a relevant weight in the quantity and seriousness of the accidents, so that any organism, company or enterprise interested in reducing these figures, can focus the resources in points where these conditions converge.

My study will focus in three part:

1. How street design impact collisions possibility
2. How the Surrounding Environment Contribute to Crashes
3. How the Driver Behavior contributes to Crashes
4. What is the most correlative factor?

2 Data

2.1 Data description

For an accurate prediction of the magnitude of damage caused by accidents, they require many reports on traffic accidents with accurate data to train prediction models. The data set provided for this work allows the analysis of a record of 200,000 accidents in the state of Seattle, from 2004 to the date it is issued, in which 37 attributes or variables are recorded and the codification of the type of accident is allowed, grouped according to 84 codes. The information can be extracted from it:

1. Weather
2. Road Condition
3. Light Condition
4. Speeding
5. The consequence of accidents: Fatalities, Injuries, etc

However, for this case study, not all the attributes are useful as the main objective is to predict an accident's probability and severity. Therefore, the Dataset needs deep understanding and analysis before choosing the right attributes to reach our goal.

For example, SDOTCOLNUM, X, Y, LOCATION, INCDTTM, INCDATE, REPORTNO, COLDETKEY, INCKEY and OBJECTID are features that give descriptive and detailed information about an accident, and are then not relevant to predict the severity of an accident in general.

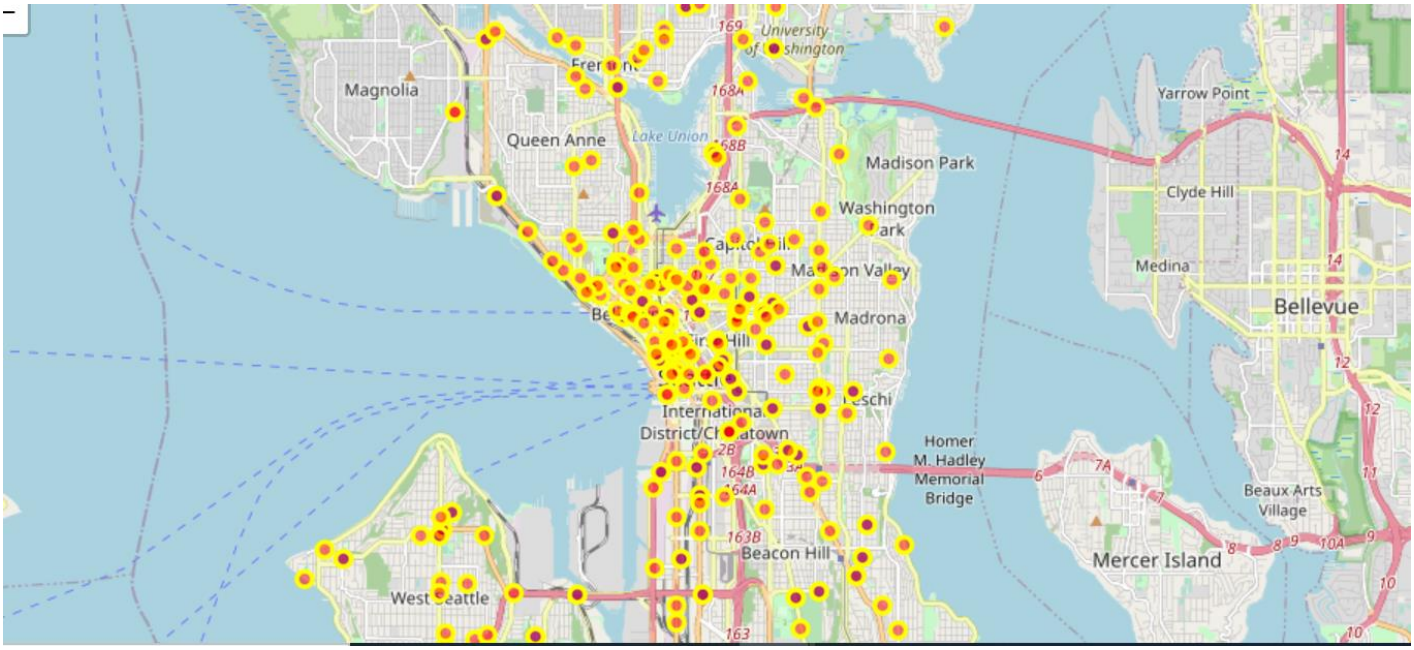
Moreover, EXCEPTRSNCODE, EXCEPTRSNDESC, PEDROWNOTGRNT, SPEEDING, INATTENTIONIND and INTKEY have a high number of missing data that would skew and bias our predictive model.

After selecting the appropriate features, the new Dataset is balanced and preprocessed before feeding it to a supervised machine learning model that will learn to predict in the future the probability of a car accident.

3 Methodology

We used Jupyter Notebook to do the data analysis. To generate the table and graph for the dataset, we imported Python libraries (Pandas, Numpy, Matplotlib, and Seaborn).

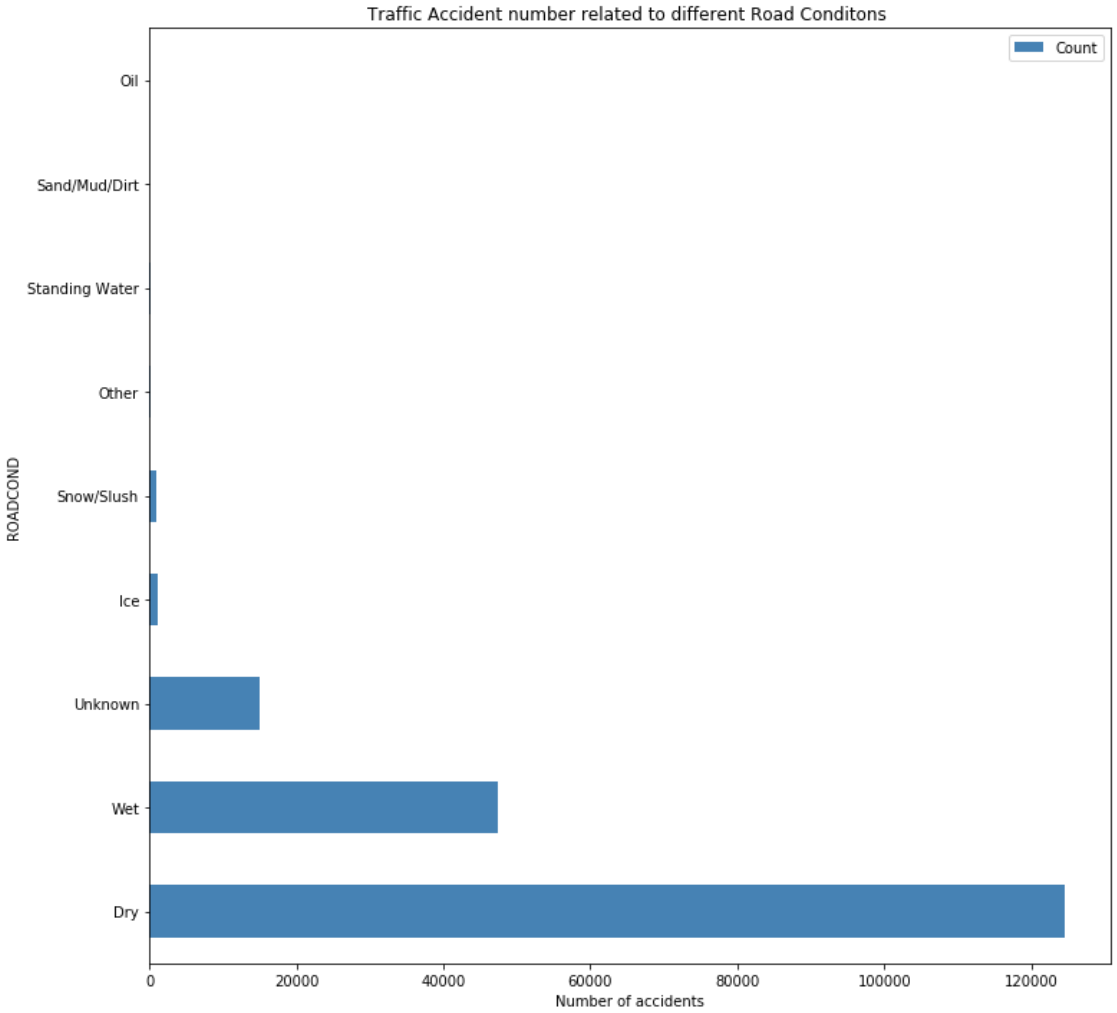
In order to clearly visualize the dataset with its labels proportion, I used folium library to obtain geographic details of Seattle and its accidents. I created a map of Seattle with markers that indicate the location of each accident and its severity. The markers in Orange represent the accidents of type 1 severity, where property damage occurred, and the markers in Red represent accidents of type 2 severity, in place of injury. However, to clearly visualize the map, I decided to only plot the 500 accidents in the map.



Next, we will explore the data and understand how a particular data column is distributed while most of our data columns are categorical and we need to know how surrounding parameters affect the total number of the accident.

3.1 Relationship Between Road Condition and Traffic Accident

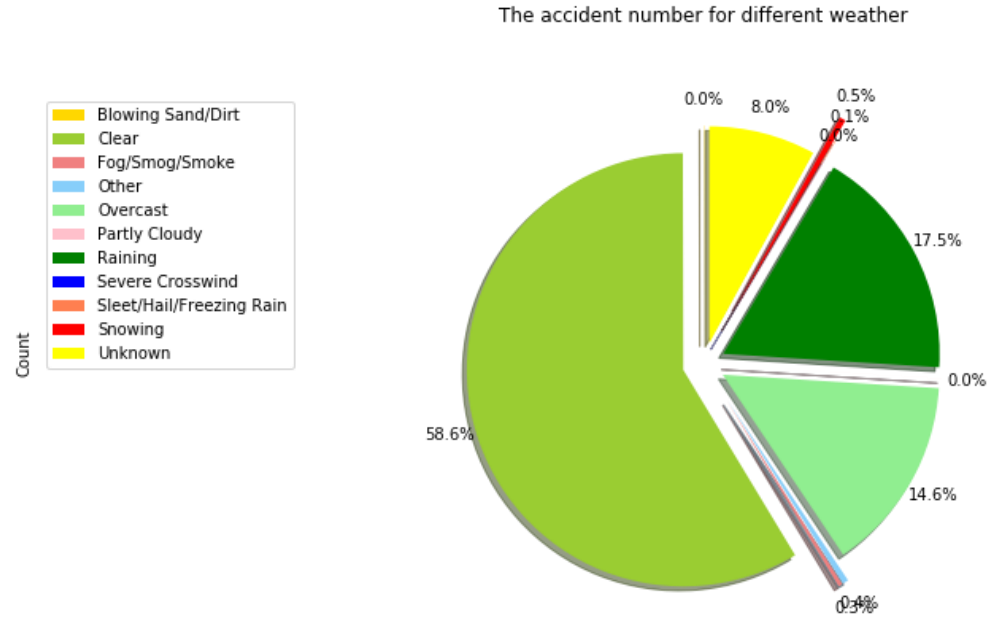
The Seattle's traffic data is imported through pd.read_csv. We noticed that it had 194,673 rows and 38 columns. Therefore, we narrowed it down to 3 columns ('Severity', 'Roadcond') and delete the missing values, which made the final dataset with 184,167 observations and 2 variables. In the study of relationship between road condition and traffic accident, I focused more on the graphical data and the value count for this categories. The horizontal bar graphic presented below.



Contrary to our usual thinking, the more dangerous and sliding the road is, the higher the number of accidents, but the truth is that dry roads are more dangerous than wet ones. What is even stranger is that car crashes are 80% more likely to happen on dry than on snowy and icy roads. One explanation is that drivers prefer to choose driving in a road with a good condition without snows or sands, so will be less car pass for roads with bad condition and less accidents possibility. Second explanation is that when drivers drive in road with bad conditions, they will have a better safety consciousness and pay attention to surrounding situation. In the other hand, it means that drivers relax vigilance in dry normal road with good condition.

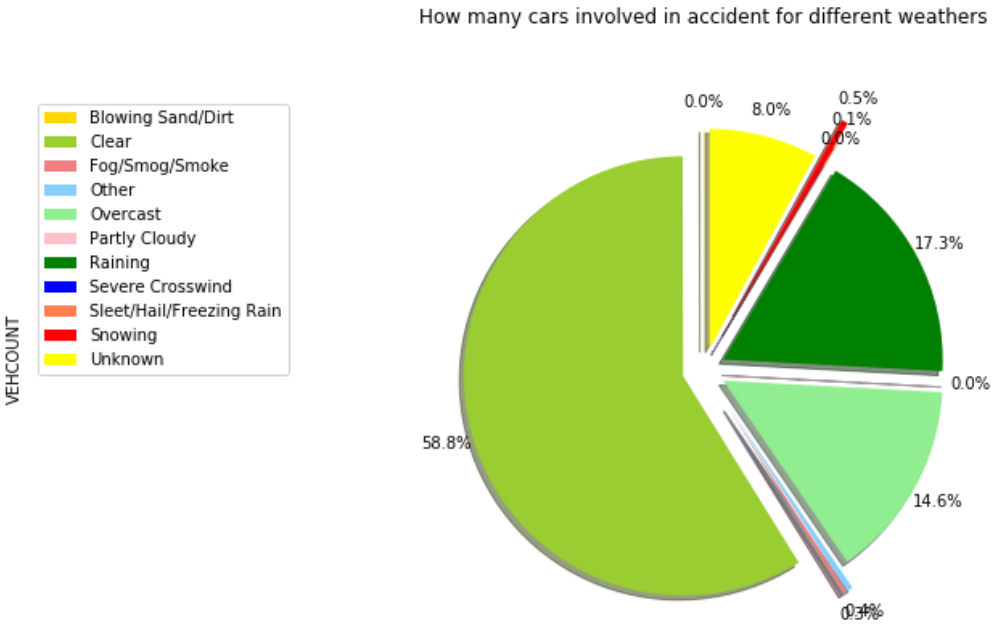
3.2 Relationship between weather condition and traffic accidents

After that, we checked about the weather. We calculated the total number of car accidents under different weather conditions. The result is show by a pie chart as below.



As we saw in the pie chart, about 60% traffic accidents happened in a clear weather condition, about 18% were for raining condition and 15% were for overcast condition. Contrary to our imagination, the most dangerous weather like smog, snowing, dirt and severe Crosswind totally contribute less than 1% traffic accidents.

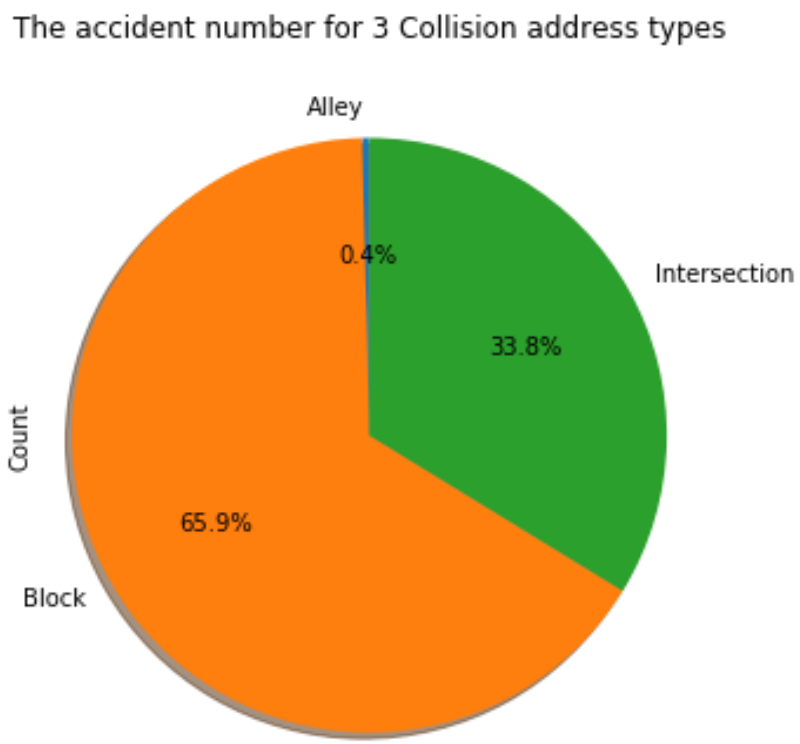
Another analysis realized for how many cars involved in accident for different weathers is showed in the pie chart below.



As we saw in the pie chart, it show nearly the same proportion distribution for different weather in involved car number as in the former pie chart. However, we can still see that there is a little increase for Clear Weather in involved car number compared with traffic accident number in the same situation. It can be explained that it will be less car pass in road when a bad weather condition, so it leads to less cars involved in accidents. When there is a good weather condition, it will be more cars drive in the road, more accident possibility.

3.3 Relationship between Collision address types and traffic accidents

It is also important to find out where most accidents take place. Upon analyzing the data, it turned out that intersections are the most common accident zones. This could be because drivers don't heed the stop sign, or maybe some intersections can use more stop signs, or maybe there need to be more pedestrian crossings. In any case, this should be an area to look into more in-depth.



Discussion

In the beginning of my analysis, I was trying to find out the severity and frequency of road accidents based on weather condition, road conditions, and other factors and other factors. Even though our data was a good size, there were a number of missing elements and we needed to clean the data in order to get a good result. We had to drop 'SPEED' because there were too many missing elements but I think that is an important factor that should be considered. From the analysis, it is clear that most accidents involve solo drivers, on wet roads, bad weather, at intersections, and are minor in nature. This could be helpful to the police department in understanding where to install more stop signs, or maybe adding cameras to intersections to compel people to slow down. We also live in a technologically friendly world so maybe we can develop some inbuilt technology in our cars that warn us when the road and weather conditions are bad, or the car is approaching a stop sign.

Conclusions

This project and analysis did not turn out as I had expected it to at all. As can be deduced in the Results section, the bad weather condition, the bad road condition in areas do not appear to have a big impact or correlation to accidents. The problem have to be focus in drivers' safety consciousness, which means that

drivers relax vigilance in dry normal road with good condition. Every driver need to pay attention to surrounding situation even though they are in good road and weather.

If I was working for the Seattle DOT or the city itself and was given this set of data to analyze, this would have been part of the full analysis done. Once I pulled this string and saw where it ended, I would have switched gears to look at the data at a much higher level in order to determine which conditions appear to have the most impact on accidents. I also propose to have more traffic sign in the road to remind our drivers to keep safety consciousness or some online trainee about this part.