Capstone project – predict car accident severity

Introduction

For the final capstone project in the IBM certificate course, we want to analyze the accident "severity" in terms of human fatality, traffic delay, property damage, or any other type of accident bad impact. The data was collected by Seattle SPOT Traffic Management Division and provided by Coursera via a link. This dataset is updated weekly and is from 2004 to present. It contains information such as severity code, address type, location, collision type, weather, road condition, speeding, among others.

The target audiences of this study are those people who really care about the traffic records, especially in the transportation department. Also, we want to figure out the reason for collisions and help to reduce accidents in the future.

Data

There are 194,673 observations and 38 variables in this data set. Since we would like to identify the factors that cause the accident and the level of severity, we will use SEVERITYCODE as our dependent variable Y, and try different combinations of independent variables X to get the result. Since the observations are quite large, we may need to filter out the missing value and delete the unrelated columns first. Then we can select the factor which may have more impact on the accidents, such as address type, weather, road condition, and light condition.

The target Data to be predicted under (SEVERITYCODE 1-prop damage 2-injury) label.

Other important variables include:

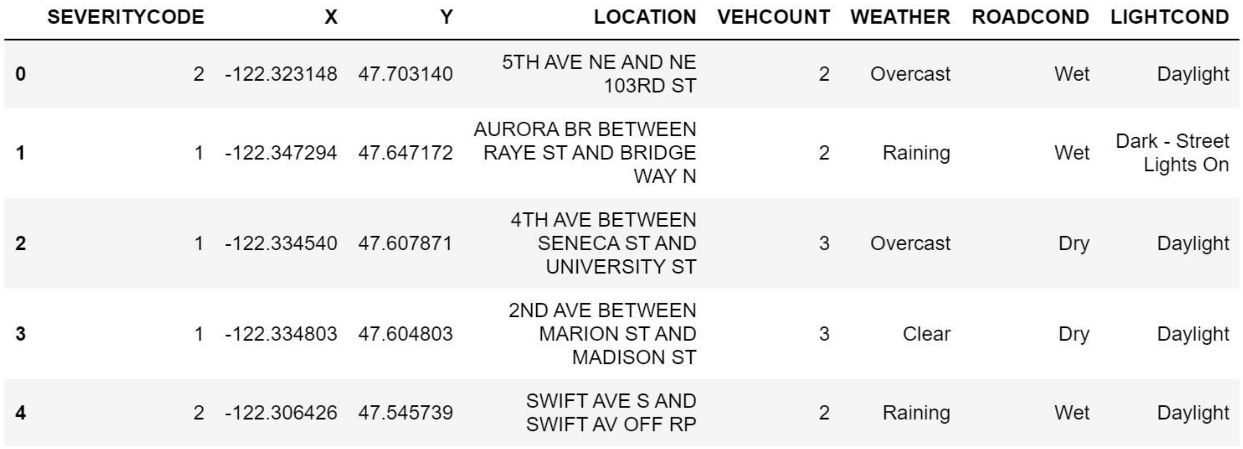
- ADDRTYPE: Collision address type: Alley, Block, Intersection
- LOCATION: Description of the general location of the collision
- PERSONCOUNT: The total number of people involved in the collision helps identify severity involved
- PEDCOUNT: The number of pedestrians involved in the collision helps identify severity involved
- PEDCYLCOUNT: The number of bicycles involved in the collision helps identify severity involved
- VEHCOUNT: The number of vehicles involved in the collision identify severity involved
- JUNCTIONTYPE: Category of junction at which collision took place helps identify where most collisions occur
- WEATHER: A description of the weather conditions during the time of the collision
- ROADCOND: The condition of the road during the collision
- LIGHTCOND: The light conditions during the collision
- SPEEDING: Whether or not speeding was a factor in the collision (Y/N)
- SEGLANEKEY: A key for the lane segment in which the collision occurred

- CROSSWALKKEY: A key for the crosswalk at which the collision occurred
- HITPARKEDCAR: Whether or not the collision involved hitting a parked car

Methodology

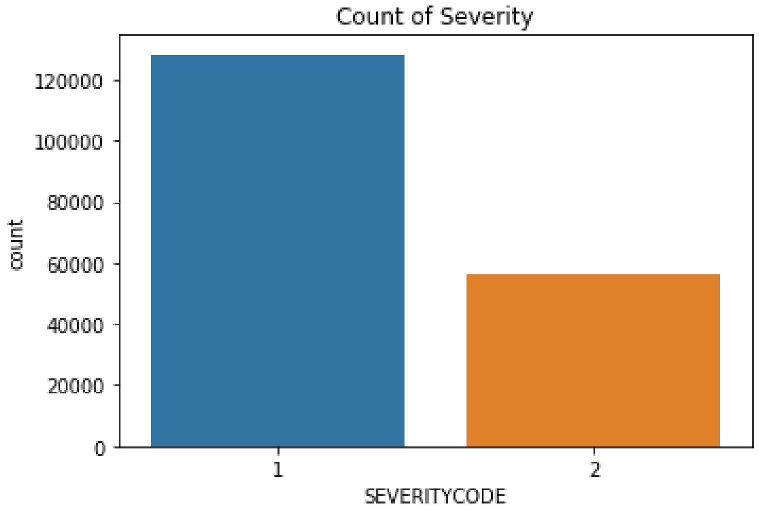
We used Jupyter Notebook to do the data analysis. To generate the table and graph for the dataset, we imported Python libraries (Pandas, Numpy, Matplotlib, and Seaborn).

First we imported the data through pd.read_csv. We noticed that it had 194,673 rows and 38 columns. Therefore, we narrowed it down to 8 columns ('Severity', 'X', 'Y', 'Location', 'Vehcount', 'Weather', 'Roadcond', 'Lighdcond') and delete the missing values, which made the final dataset with 184,167 observations and 8 variables.



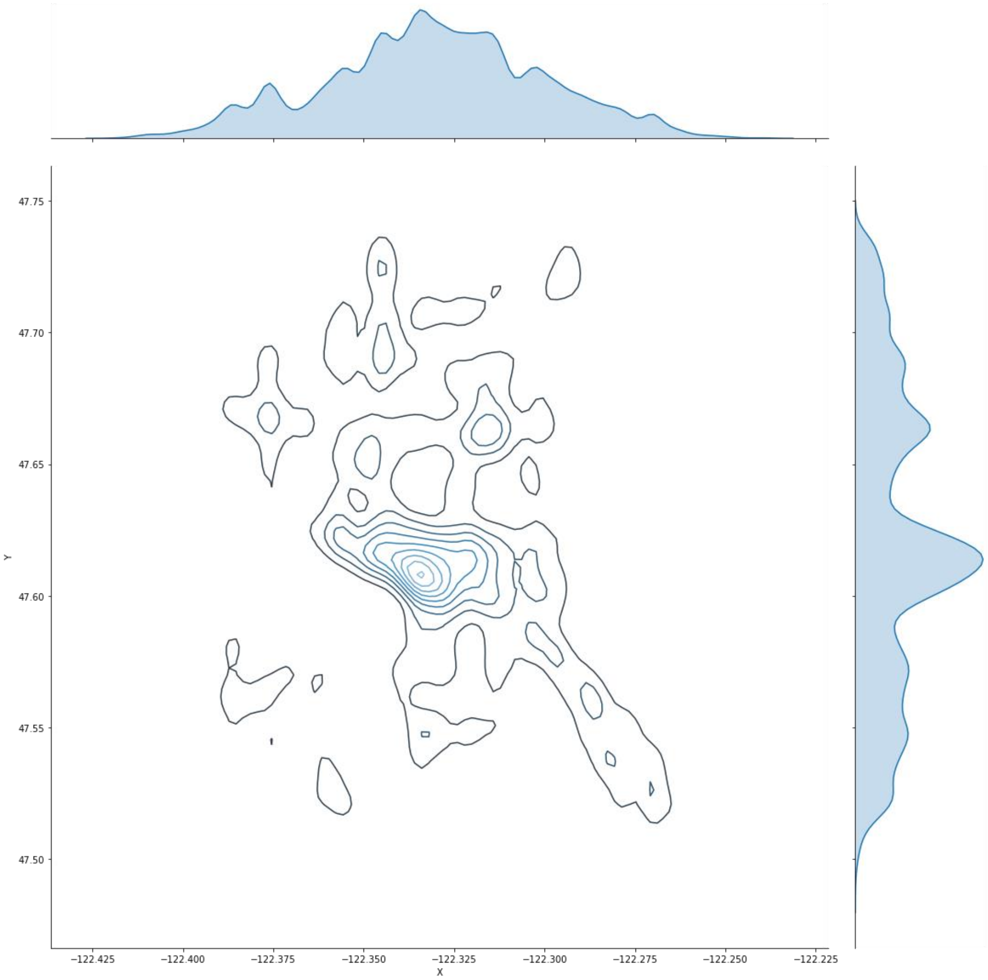
	SEVERITYCODE	X	Y	LOCATION	VEHCOUNT	WEATHER	ROADCOND	LIGHTCOND
0	2	-122.323148	47.703140	5TH AVE NE AND NE 103RD ST	2	Overcast	Wet	Daylight
1	1	-122.347294	47.647172	AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N	2	Raining	Wet	Dark - Street Lights On
2	1	-122.334540	47.607871	4TH AVE BETWEEN SENECA ST AND UNIVERSITY ST	3	Overcast	Dry	Daylight
3	1	-122.334803	47.604803	2ND AVE BETWEEN MARION ST AND MADISON ST	3	Clear	Dry	Daylight
4	2	-122.306426	47.545739	SWIFT AVE S AND SWIFT AV OFF RP	2	Raining	Wet	Daylight

Since most of the variable were categorical, it was hard to make the regression model. So, in this study, we focused more on the graphical data and the value count for different categories. There were around 135,000 (2/3) level 1 accidents and 60,000 (1/3) level 2 accidents.

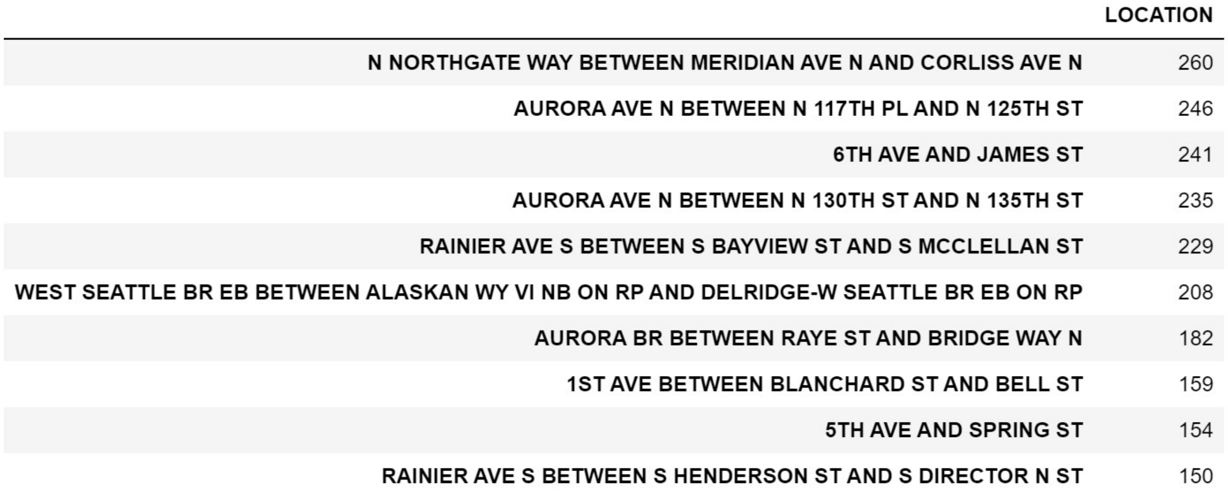


Results and Discussion

We generated the graphical information based on Seaborn library. The result showed that some locations did have more car accidents than the other places.

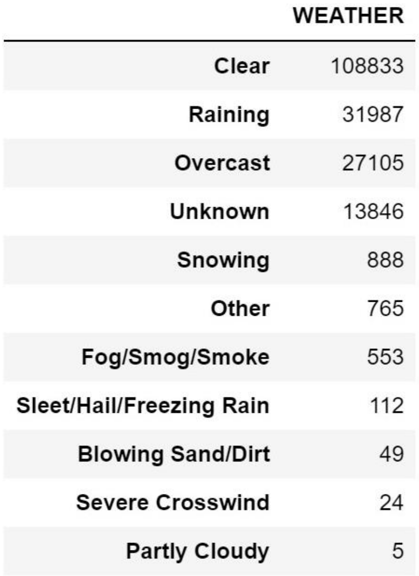


These places are listed as follow:

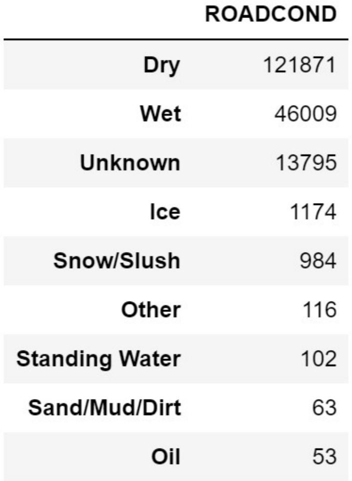


	LOCATION
N NORTHGATE WAY BETWEEN MERIDIAN AVE N AND CORLISS AVE N	260
AURORA AVE N BETWEEN N 117TH PL AND N 125TH ST	246
6TH AVE AND JAMES ST	241
AURORA AVE N BETWEEN N 130TH ST AND N 135TH ST	235
RAINIER AVE S BETWEEN S BAYVIEW ST AND S MCCLELLAN ST	229
WEST SEATTLE BR EB BETWEEN ALASKAN WY VI NB ON RP AND DELRIDGE-W SEATTLE BR EB ON RP	208
AURORA BR BETWEEN RAYE ST AND BRIDGE WAY N	182
1ST AVE BETWEEN BLANCHARD ST AND BELL ST	159
5TH AVE AND SPRING ST	154
RAINIER AVE S BETWEEN S HENDERSON ST AND S DIRECTOR N ST	150

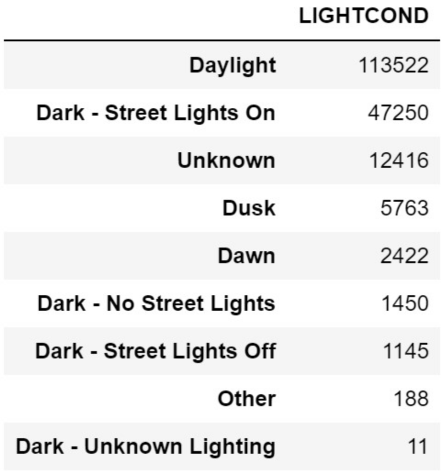
After that, we checked about the weather, road, and light condition. We calculated the total number of car accidents under different situations. There was no significant evidence showed that they might be the reason for the accidents.



	WEATHER
Clear	108833
Raining	31987
Overcast	27105
Unknown	13846
Snowing	888
Other	765
Fog/Smog/Smoke	553
Sleet/Hail/Freezing Rain	112
Blowing Sand/Dirt	49
Severe Crosswind	24
Partly Cloudy	5

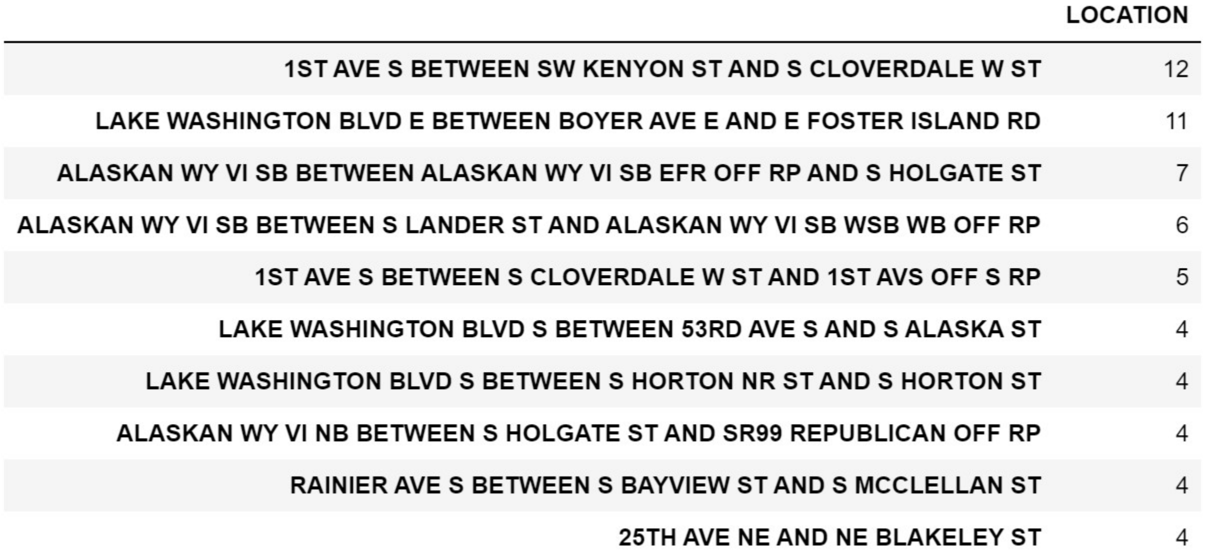


	ROADCOND
Dry	121871
Wet	46009
Unknown	13795
Ice	1174
Snow/Slush	984
Other	116
Standing Water	102
Sand/Mud/Dirt	63
Oil	53



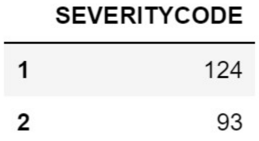
	LIGHTCOND
Daylight	113522
Dark - Street Lights On	47250
Unknown	12416
Dusk	5763
Dawn	2422
Dark - No Street Lights	1450
Dark - Street Lights Off	1145
Other	188
Dark - Unknown Lighting	11

Therefore, we created different subset under different conditions. Again, these conditions did not have any clear relationship with the severity code. However, we figured out that under dark light condition, different places will cause the car accidents, which might be helpful for the government to deal with it.

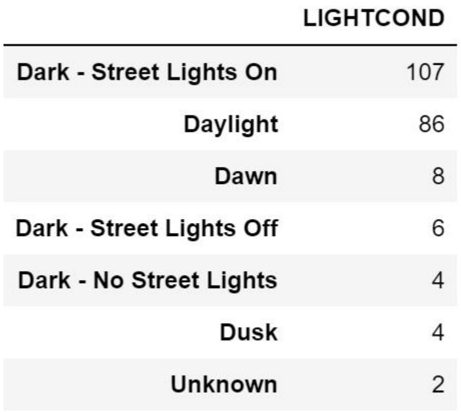


	LOCATION
1ST AVE S BETWEEN SW KENYON ST AND S CLOVERDALE W ST	12
LAKE WASHINGTON BLVD E BETWEEN BOYER AVE E AND E FOSTER ISLAND RD	11
ALASKAN WY VI SB BETWEEN ALASKAN WY VI SB EFR OFF RP AND S HOLGATE ST	7
ALASKAN WY VI SB BETWEEN S LANDER ST AND ALASKAN WY VI SB WSB WB OFF RP	6
1ST AVE S BETWEEN S CLOVERDALE W ST AND 1ST AVS OFF S RP	5
LAKE WASHINGTON BLVD S BETWEEN 53RD AVE S AND S ALASKA ST	4
LAKE WASHINGTON BLVD S BETWEEN S HORTON NR ST AND S HORTON ST	4
ALASKAN WY VI NB BETWEEN S HOLGATE ST AND SR99 REPUBLICAN OFF RP	4
RAINIER AVE S BETWEEN S BAYVIEW ST AND S MCCLELLAN ST	4
25TH AVE NE AND NE BLAKELEY ST	4

Finally, we put the number of vehicles into considerations. This time, we noticed that nearly 50% of accidents with 6 or more vehicles may cause series injuries. In addition, more than half of the accidents took place at dark conditions. This information was quite useful for improving the traffic situations.



	SEVERITYCODE
1	124
2	93



	LIGHTCOND
Dark - Street Lights On	107
Daylight	86
Dawn	8
Dark - Street Lights Off	6
Dark - No Street Lights	4
Dusk	4
Unknown	2

Conclusion

This project and analysis are quite helpful for the Seattle transportation department. Before I did the analysis, I thought that maybe weather, road, and light condition may cause more accidents, the results showed that it was not correct. However, we do figure out that the accidents are highly related to some specific locations. Thus, the traffic management division could try to improve the safety instructions or some other factors that could reduce the accidents.

Furthermore, there are some places which has more accidents during the dark time. For those places, adding lights might be a good solution to reduce the collisions. Also, when more cars involved in the accident, it seems that the level of severity will increase. They may need to be responded immediately to save more life.