IBM Data Science Capstone Project

Introduction

The problem at hand is to identify how certain conditions such as the weather, location and lighting affect the risks of driving. In particular, how likely are they to cause a severe collision.

This information can be used for navigation system providers who would direct the flow of traffic to less risky areas under certain conditions. In turn, a navigation company with such a tool could offer its services to insurance companies who would be willing to provide the data to their clients in order to reduce claims and increase profitability. Navigation companies could also sell the tools to public transportation regulation branches in whose interest it is to increase road safety and reduce traffic incidents.

The problem at hand considers understanding certain factors that affect the severity of collisions as well as the distribution of collisions over geographic areas. Such insight will be helpful in preventing future accidents or minimizing the damage from such accidents.

Data

Given data is provided by SDOT Traffic Management Division and it is updated on a weekly basis. The data contains 194,673 observations of road incidents in Seattle with various features such as the severity of the incident, number of vehicles and people involved, conditions during the incident, time of the incident, etc.

The data will be use to conduct an exploratory analysis and assess the effect of key factors on the severity of the collision as well as the distribution of the collision.

Methodology

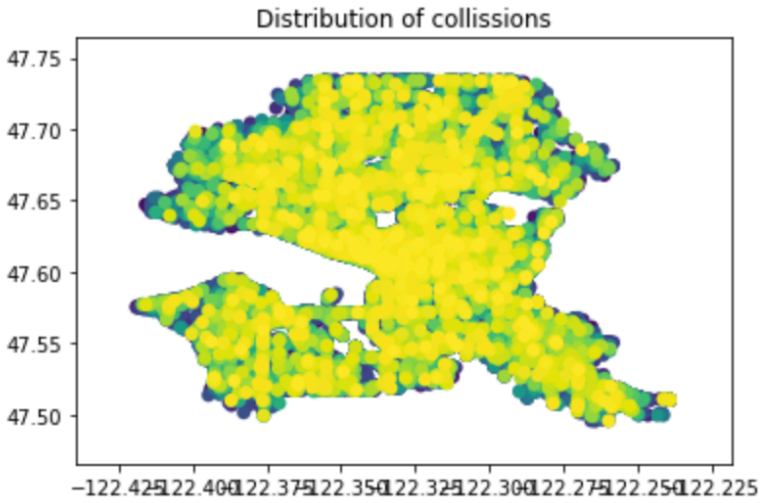
During the exploratory phase, I visualized the distribution of the colllissions over the geographic area. I used a heatplot map to plot thee coordinates of the collisions and thus see which areas where more prone to have collisions happen.

I also created new variables from the date_time variable provided in the raw data. These variables where DAY and HOUR. The represented day of the week and hour of the day respectively. After creating the variables, I plotted histograms to see the distribution of the collision over hour of the day and day of the week.

Another method I used for exploratory analysis was to simply count the values of various variables. This allowed me to explore how the collisions were distributed over various weather conditions, road conditions and light conditions. However, since I had no data on the distribution of weather events in the area, I used another method to see the effect of these conditions. In particular, I grouped the data based on these conditions separately. Then I created a separate column in the dataframe to see what was the distribution in the severity of the collision for each group separately. Thus, for example, if in a certain weather condition, the share of more severe accidents was larger than in another, it could be surmised that the particular weather condition was associated with more severe collisions.

Results

From the heatmap, it is visible that the distribution of the collisions is quite even across the city with slightly more concentration in center of the city than in the outskirts.

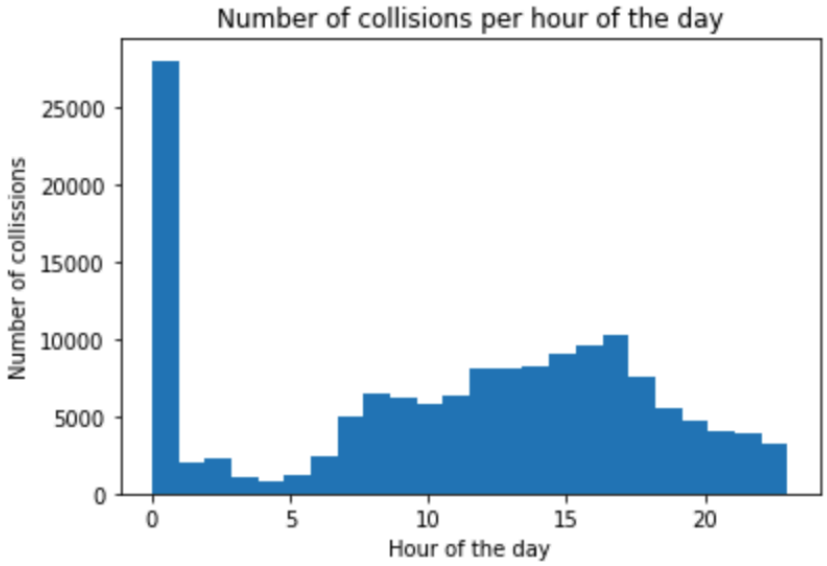


From the data, it is evident that there are certain locations that are more prone to collisions than others. In particular, “BATTERY ST TUNNEL NB BETWEEN ALASKAN WY VI NB AND AURORA AVE N” had 276 collisions, most out of all recorded locations. It was closely followed by “BATTERY ST TUNNEL SB BETWEEN AURORA AVE N AND ALASKAN WY VI SB” with 271 collisions and “N NORTHGATE WAY BETWEEN MERIDIAN AVE N AND CORLISS AVE N” with 265.

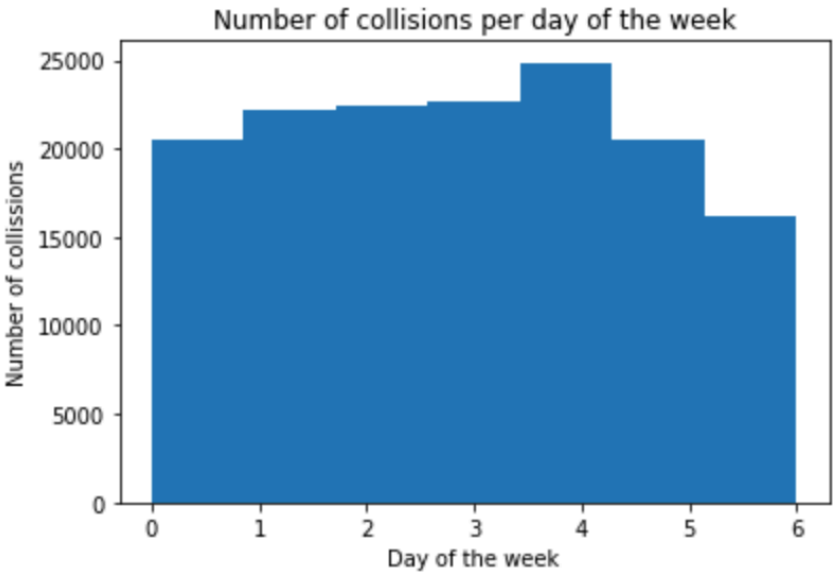
Vast majority of the collisions occurred in blocks. Half of that number was recorded at intersections while only a small percentage occurred in alleys.

Most collisions (47,987) where recorded on parked cars. There were 34,674 collisions at angles and 34,090 accidents where the car was rear-ended.

There was a peak in the number of collisions from 3 to 5 pm (peak at 0 hour is due to null values being converted to 0s).



For the days of the week, Friday was associated with most collisions while the number during the weekends were lower. The number of collisions increased during the weekdays but by a slight margin only.



When it comes to weather conditions, the share of severe collisions was higher during rain, fog and partly cloudy conditions. During snow, the collisions were list likely to be severe.

		SEVERITYCODE	SEVERITYDISTR
WEATHER	SEVERITYCODE		
Blowing Sand/Dirt	1	41	73.214286
	2	15	26.785714
Clear	1	75295	67.750934
	2	35840	32.249066
Fog/Smog/Smoke	1	382	67.135325
	2	187	32.864675
Other	1	716	86.057692
	2	116	13.942308
Overcast	1	18969	68.445551
	2	8745	31.554449
Partly Cloudy	2	3	60.000000
	1	2	40.000000
Raining	1	21969	66.281490
	2	11176	33.718510
Severe Crosswind	1	18	72.000000
	2	7	28.000000
Sleet/Hail/Freezing Rain	1	85	75.221239
	2	28	24.778761
Snowing	1	736	81.146637
	2	171	18.853363
Unknown	1	14275	94.592804
	2	816	5.407196

For the Light Conditions, more severe collisions were observed during dawn, daylight, and dusk.

		SEVERITYCODE	SEVERITYDISTR
LIGHTCOND	SEVERITYCODE		
Dark - No Street Lights	1	1203	78.269356
	2	334	21.730644
Dark - Street Lights Off	1	883	73.644704
	2	316	26.355296
Dark - Street Lights On	1	34032	70.158946
	2	14475	29.841054
Dark - Unknown Lighting	1	7	63.636364
	2	4	36.363636
Dawn	1	1678	67.066347
	2	824	32.933653
Daylight	1	77593	66.811610
	2	38544	33.188390
Dusk	1	3958	67.062013
	2	1944	32.937987
Other	1	183	77.872340
	2	52	22.127660
Unknown	1	12868	95.509538
	2	605	4.490462

It was also more likely for a severe collision to occur in dry and oily conditions.

		SEVERITYCODE	SEVERITYDISTR
ROADCOND	SEVERITYCODE		
Dry	1	84446	67.822665
	2	40064	32.177335
Ice	1	936	77.419355
	2	273	22.580645
Oil	1	40	62.500000
	2	24	37.500000
Other	1	89	67.424242
	2	43	32.575758
Sand/Mud/Dirt	1	52	69.333333
	2	23	30.666667
Snow/Slush	1	837	83.366534
	2	167	16.633466
Standing Water	1	85	73.913043
	2	30	26.086957
Unknown	1	14329	95.032498
	2	749	4.967502
Wet	1	31719	66.813414
	2	15755	33.186586

Discussion

There are several locations in the city that are characterized by an abnormally high number of collisions. It could be suggested to the relevant authorities to take measures accordingly and install road signs or equipment at those locations that would mitigate the risks of collisions.

As for the effect of weather and road conditions. The results were quite counterintuitive as icy roads, standing water,no street lights, and severe crosswind were all more likely to reduce the severity of the collision than normal conditions.

Distribution of the collisions over days of the week and hours of the day were as expected. The number of collisions was higher during the peak hours when there are more cars on the road. The number was also higher during the weekdays and especially Friday. On the weekends, when more people stay home, the number of collisions declined significantly.

Conclusion

While there is some insight to be drawn from the data, the effect of weather, light and road conditions showed counterintuitive results which could be caused by the absence of certain key factors in the data or the bias in the data. It was still possible to analyze the distribution of the collisions over geographic areas and time, however the results where as intuition would dictate. More collisions occurred in more congested areas and at times when there are more cars on the roads.