Analyzing car accidents in the city of Seattle in the past decade (2010-2020)

Diego Kaleb Samano Rodríguez

September, 2020

Culiacán, México

1.1 Introduction

1.2 Background

Seattle is a big city located in the Northwestern part of the United States, the Seattle Metropolitan Area has a population of 3,263,497 habitants. The city has gained the reputation of 'Rain City' because the of the constant precipitation it has all year long. It is important to notice that since 2013, Seattle has been in the top five of cities with the most annual growth rate, the increase of business, economical activity and population mean that the city has also seen and increase of vehicle activity. The climatological characteristics and its population size makes the city a very interesting study object for car accidents.

1.3 Problem

According to data from the ‘Centers for Desease Control and Prevention’ from the US government (cdc.gov), car accidents ate the leading cause of death in the country killing around 100 people each year and threatening the lives around 2.5 million drivers and passengers with injuries. Financially talking, the expense that car accidents generated was around 75 billion dollars.

1.4 Interest

The Seattle Metropolitan area is the 15th largest in the country and the city of Seattle the 24th largest, knowing the distribution of crashes, their tendencies and their causes is crucial for the creation of better road systems, insurance plans, traffic policies etc.

2.1 Data acquisition and cleaning

2.2 Data Source

The data used for the project was the one provided as CSV file from the IBM Course. It contains around 194673 registries spanning from 2004 all the way to the year 2020.

2.2 Data cleaning and feature selection

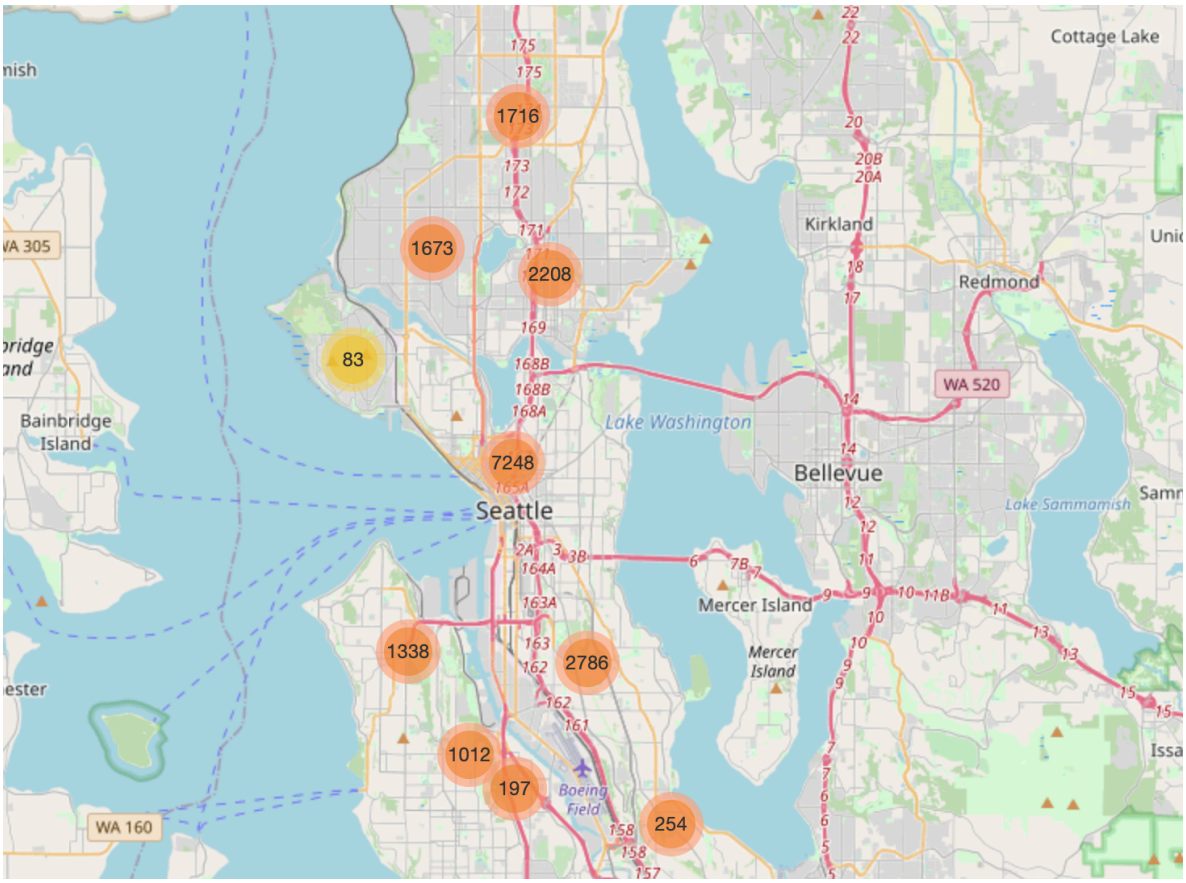
After downloading the CSV file from the link provided by the course a data exploration was made, the link containing the dataset metadata was visited in order to have a better understanding of the type of data each of the 38 different columns of the dataset had and its meaning. After doing so, the columns that were considered not as vital for the final goal were dropped. The dataset left had dimensions of 194673 x 25 instead of the original 194673 x 38. The columns were also renamed in order to have a better understanding of the information they contain.

3.1 Exploratory graphical data analysis

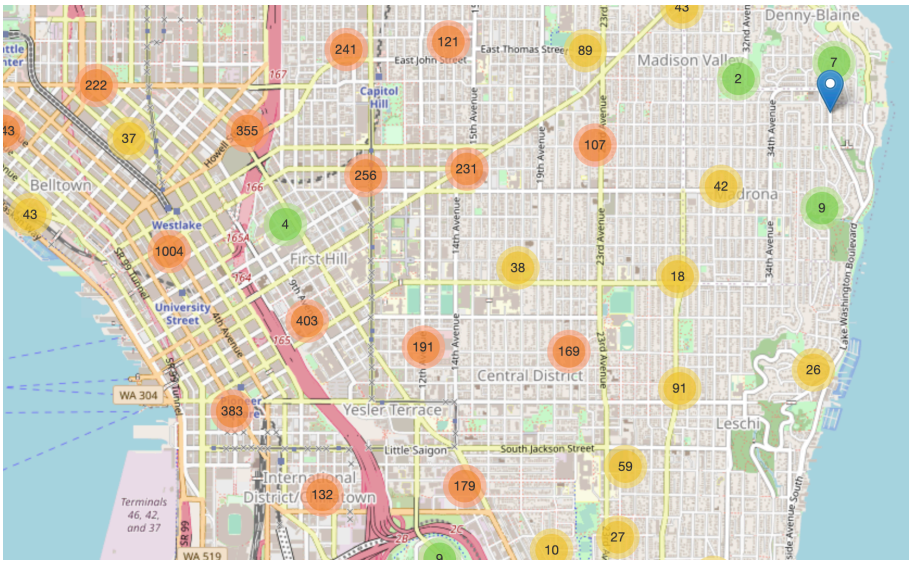
In order to understand better data it is recommended before building any kind model to do a graphical analysis and visualize the information to see what kind of challenge you are dealing with and see how you can tackle it.

3.2 Accident geographical distribution

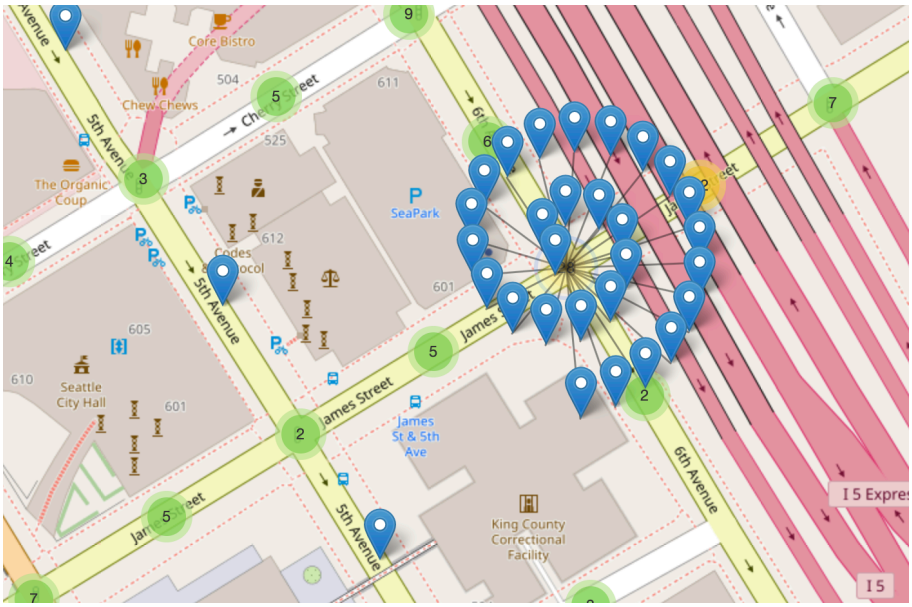
First and foremost, having a visualization of how the accidents were distributed throughout the city was interesting. The library used, Folium, allows to zoom in and out the map to see the location of every incident, accidents shown in the image are clustered together because it is easier to visualize.



The image below shows and example of a zoomed in portion of the map showing the city's downtown area, as we can see instead of having a very big cluster of incidents now we have smaller clusters.

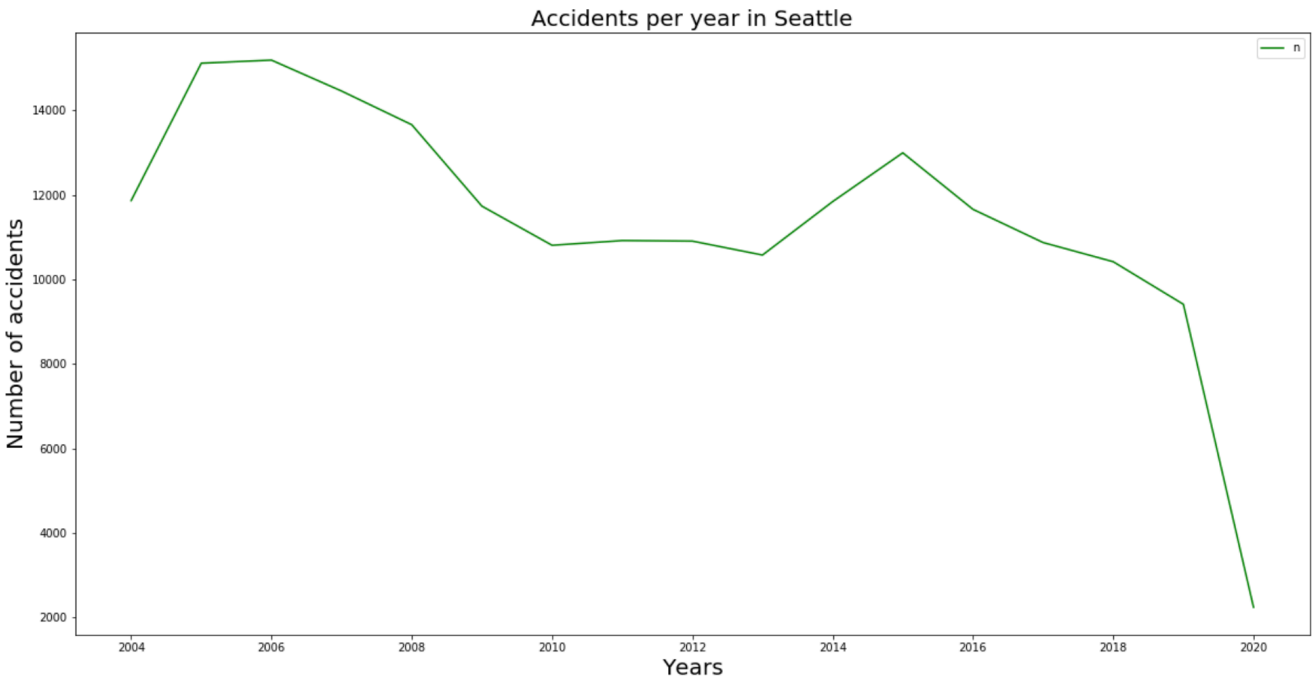


As an example, the image below shows markers of the exact locations of accidents that occurred in the freeway.



3.3 Accident frequency variability per year (2004-2020)

The data set spans 16 years, from 2004 to 2020. It has 194673 accidents records, it is interesting how the amount of accidents per year has either increased or decreased.



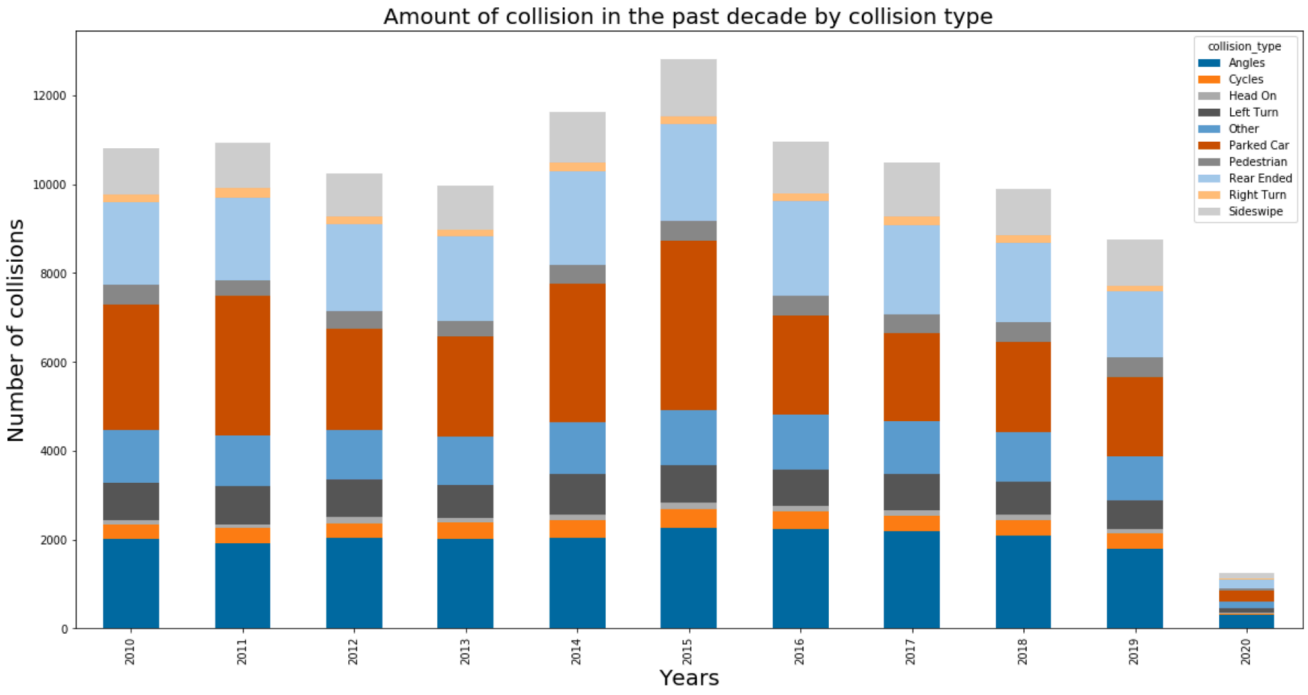
As we can see in the figure above the amount of accidents has seen a decline since the year 2005 were the see had a significant spike in comparison to 2004. If the graph line is followed we see another spike that occurred in the year 2015, after that year the amount of accidents declined little by little to finally plumb down in 2020. Many of the variations can be attributed to bigger magnitude events such as the coronavirus pandemic of this year which caused a serious decreased in traffic all around the world due to the quarantine protocols and home office measures taken by each country. The decline seen in the years 2006-2007 may be attributed, in part, to the opening of the South Lake Union Streetcar opening in 2007, a tram line that connects the South Lake Union neighborhood to Downtown Seattle, two of the most transited and active areas in the city.

3.4 Collision type distribution in the past decade

The dataset contains a column that indicates the type of collision of every accident registry. Each collision can be of 10 different types determined in the dataset, this are: angles, cycles, head on, left turn, parked car, pedestrian, rear ended, right turn, sideswipe and other (if the collision does not fit in the other nine values).

collision_type	Angles	Cycles	Head On	Left Turn	Other	Parked Car	Pedestrian	Rear Ended	Right Turn	Sideswipe
year										
2010	2004	333	112	841	1181	2820	436	1879	166	1034
2011	1924	332	89	870	1118	3165	340	1865	216	1000
2012	2033	339	128	844	1133	2272	389	1950	176	981
2013	2015	369	114	730	1095	2251	346	1914	153	990
2014	2053	395	124	894	1175	3120	418	2098	215	1140
2015	2254	437	135	857	1242	3811	442	2178	179	1285
2016	2247	380	144	794	1243	2229	453	2129	175	1157
2017	2196	345	129	807	1180	1981	432	2007	205	1206
2018	2093	337	132	753	1109	2028	446	1787	178	1037
2019	1782	357	96	660	972	1801	440	1490	121	1034
2020	304	33	19	92	164	240	58	196	19	136

The stacked bar graph below shows the distribution of every collision type in relation to the total amount of accidents of the given year.



As we can see from the graph above, in the past decade the year that had the most amount of car accidents was 2015, and the least (not including 2020) was 2019. The collision type distribution remains fairly the same every year. What is worthy of notice is that an increase or decrease of the ‘Parked Car’ collision type (red) is the one that the changes the most every year, followed by ‘Rear Ended and ‘Sideswipe’.

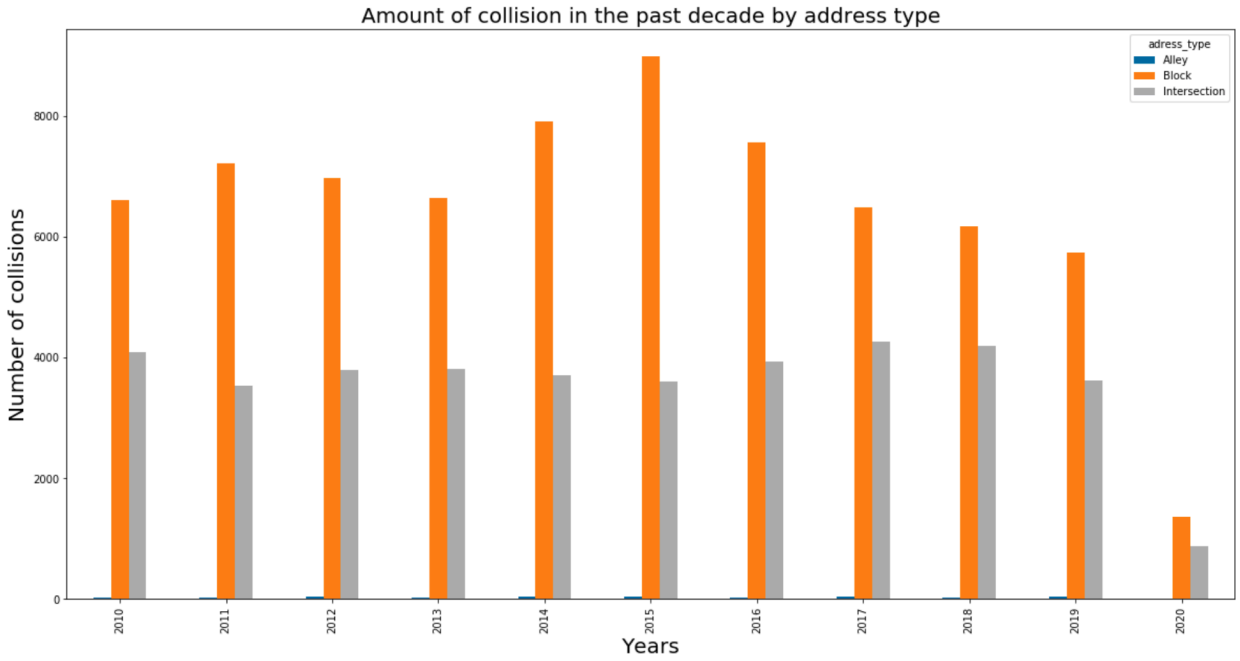
3.5 Accidents by address type in the past decade

adress_type year	Alley	Block	Intersection
2010	30	6606	4080
2011	31	7220	3535
2012	34	6969	3792
2013	22	6640	3807
2014	42	7917	3701
2015	45	8990	3599
2016	27	7566	3928
2017	42	6479	4256
2018	32	6166	4191
2019	39	5732	3618
2020	4	1357	881

The dataset contains a column called ‘address_type’, which contains the type of place (streetwise) where the accident occurred, the possible values are ‘Alley’, ‘Block’ and ‘Intersection’.

In the table at the left we can see the amount of accidents of every address type for every year in the past decade. The majority of the accidents occurred in address type ‘Block’ which means it occurred in a normal street.

In the graph below we can see in a graphical way the values contained in the table.



3.6 graphical analysis conclusion

Contrary to what was expected previous to the graphical analysis, there is not much of change in the distribution of the attribute values throughout the years, there is though a constant decrease in the amount of total accidents in the city of Seattle, which is a good thing.

4.1 Machine learning model

The machine learning model used for this project was a logistic regression model. Logistic Regression is a statistical model that in its simplest form uses a logistic function to model a binary dependent variable. In this case the variable to model was 'SEVERITYCODE' which con only be of values 1(only material damage in the accident) and 2 (injuries or deaths present).

In the below table we can see the results obtained for the model, as mentioned earlier, the dependent variable was 'SEVERITYCODE', the independent variables where related to the amount of persons and cyclists involved, road condition, address type, collision type and speeding. In total there were 18 independent variables.

	Precission	Recall	F1-Score	Support
1	0.75	0.96	0.84	15606
2	0.77	0.28	0.41	6925
Accuracy			0.75	225331
Macro avg.	0.76	0.62	0.62	22531
Weighted avg.	0.76	0.75	0.71	22531

The Jaccard similarity score was also calculated, obtaining a value of 0.7522

4.2 Results

The Logistic Regression model used has a very balanced precision for values 1 and 2, with values of 0.75 and 0.77 respectively, this means that in both cases 70+% of the the items in the model are relevant.

The recall values where 0.96 for value 1 and 0.28 for value, this means how many relevant items were selected. Finally we obtained and F1-Score of 0.75, this means that our model works fairly well, obtaining a value of 1 means that the model has a perfect performance.

5 Project conclusions

After obtaining the results of the machine learning model some advice can be given to the most interested entities regarding the situation, in this case, car accidents in the city of Seattle. This entities are the Seattle Transportation Department and possibly car insurance companies. Taking into account all of the independent variables in the model, both stakeholders can now determine depending on the status of all those variables the chances of a car accident resulting in property damage only or injured individuals. The Seattle Transportation Department can use this data to improve road safety guidelines and recommendations, as well as improving road infrastructure for particular vehicles, public transport, cyclists and pedestrians. Car insurance companies can develop new business strategies and protocols regarding cities because this model approach can be replicated in other places where similar car accident data is available.