Predicting major cause for car accidents in Seattle

Rafael Aguirre

03 of September 2020

1. Introduction

1.1 Background

Seattle being one of the major cities with most precipitations per year in the US have a considerable amount of car accidents per year. With this said many people can conclude that this precipitation is the main factor for car accidents in the city, but this may not be 100% true. This city has been doing an effort to reduce the amount of car accidents per year and in some way, they've accomplished some reduction in this numbers.

1.2 Problem

Use the data for car accidents in Seattle from 2006-2020 to predict the main factors of this car collisions and, although the city can't stop people from going out on their cars, advise the people on which conditions it's less recommended to go out.

1.3 Interest

Obviously, the people and the city, car accidents cost a great amount of money to the people whose cars need to be repaired and to the city whose public property sometimes gets considerable damage due to this accident.

2. Data acquisition and cleaning

I wanted to analyze data that once reaching a conclusion it can make a greater impact. Searching through Kaggle a couldn't find a data set with all the right characteristics for this project so I decide to analyze the Seattle car accident data set given by the course.

For a start the data set had 194673 rows and 38 columns. As our purpose was to predict the major factor for car accidents, we decided to eliminate those columns that didn't met that purpose. Columns such as 'PEDROWGRNT', 'INTKEY'

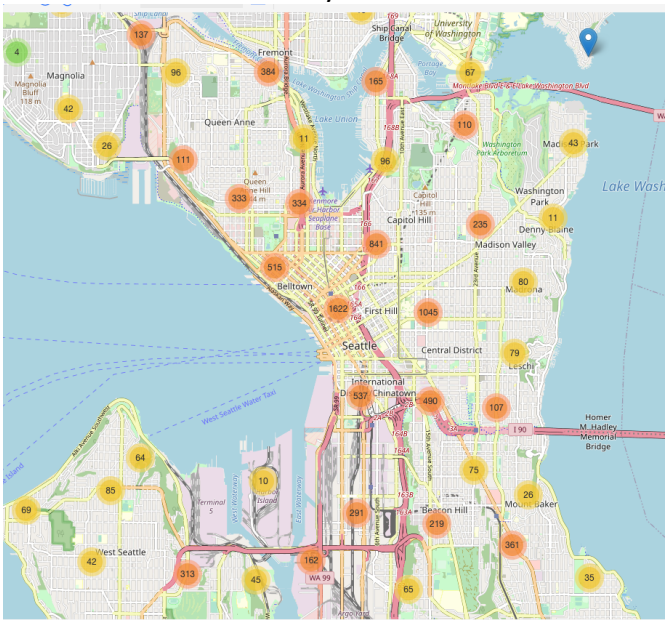
and ‘SPEEDING’ the majority of its values are Nan values that we can’t process. Also columns such as ‘SEGLANEKEY’, ‘ CROSSWALKEY’ and ‘HITPARKEDCAR’ had a continue value for almost all of its rows and this tells us that this isn’t a major factor for the outcome of the crash.

Eliminated Column
INATTENTIONIND
INCKEY
COLDETKEY
REPORTNO
INTKEY
SEGLANEKEY
CROSSWALKKEY
SDOTCOLNUM
PEDROWNOTGRNT
INCDTTM
PEDCYLCOUNT
SEVERITYCODE.1
EXCEPTRSNDESC
EXCEPTRSNCODE
SPEEDING

Once those columns were off the data set, we continued to change the remaining columns names for more easy ones to use. As the remaining columns had the necessary properties for me to analyze them this was all the cleansing I needed to do.

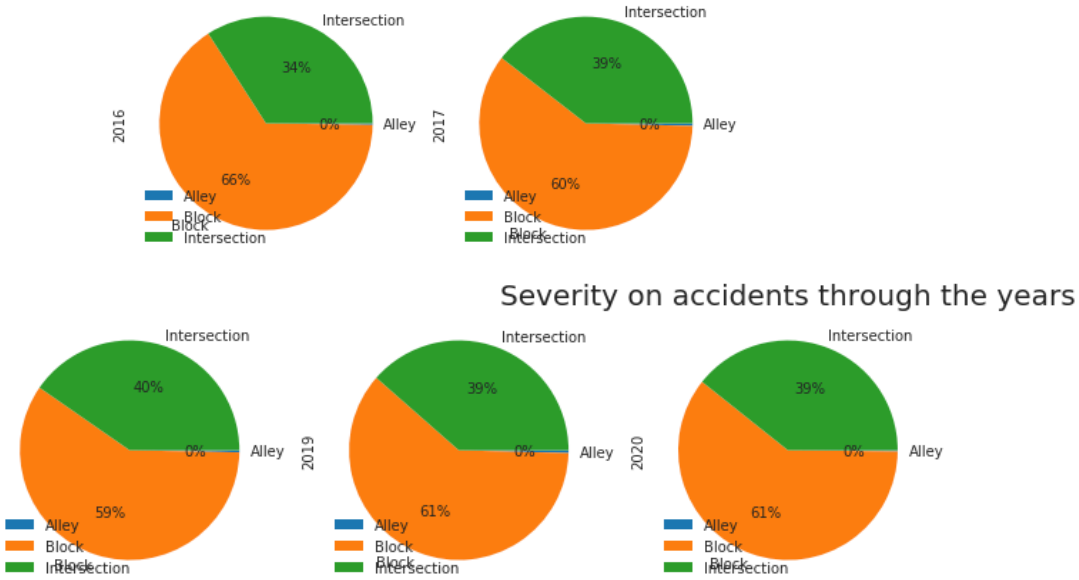
3. Data Analysis

We started visualization the zone of the city with the most car accidents:



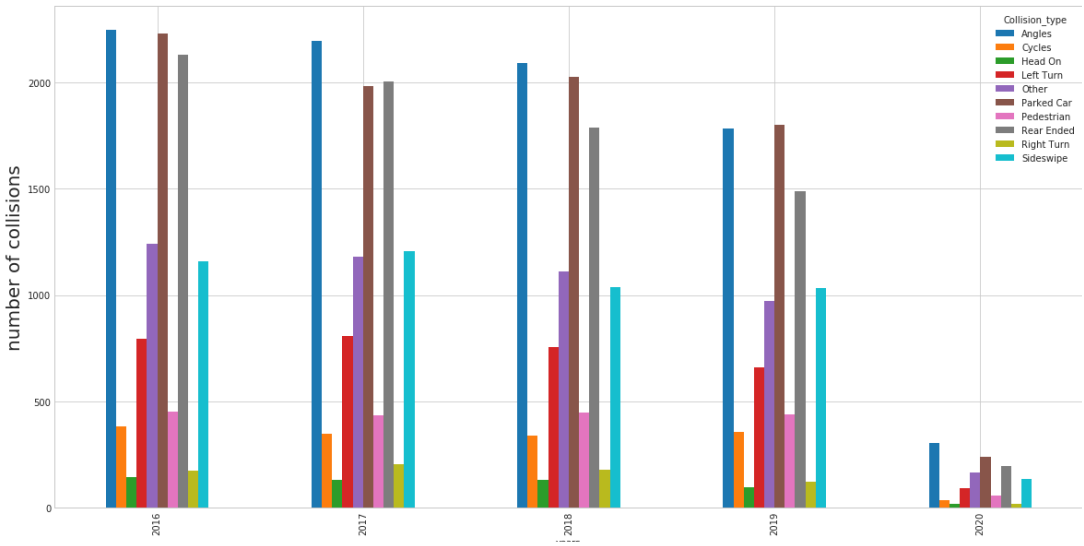
We can see the central area of Seattle is one of the zones with more car accidents in the city. This covers the zones of Chinatown, First Hill, Central district, Belltown, International District and beacon Hill.

Once we analyzed the data, we see that the majority of this zones is covered by block areas. Also, the column 'Adress_type' tells us that the address types where accidents occur the most are block areas:



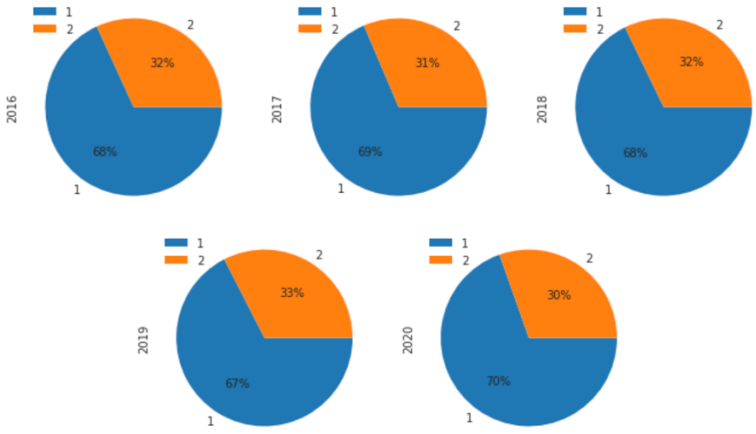
Adress_type	
Alley	751
Block	126926
Intersection	65070

Once we concluded that the zones with more accidents are the block areas, I analyze with the data set which type of this accident cause more severity for the people:



Angles, Parked car and rear ended are the top 3 most common types of accidents in Seattle. All three are common for areas with high traffic which is related with the block areas mentioned above. Also, in block areas people tend to parked their car next to the side walk which would also explained why this type of collision is one of the top three.

Finally, we must see what the severity per year of this accidents is to know what type of solution we must give. I chose to create some pie chart showing the percentage of which type of severity (material damage or injury) is most concurrent in this accident:



Machine Learning Model:

I decide to make a logistic regression model in which we use the variables: 'Person_count', 'Ice','Oil','Other','Sand/Mud/Dirt','Snow/Slush','Standing Water','Unknown','Wet','Block','Intersection','Cycles', 'Head On','Left Turn', 'Other','Parked Car','Pedestrian','Rear Ended','Right Turn','Sideswipe' and for our ´Y´we chose ´SEVERITYCODE and this were the results.

	precision	recall	f1-score	support
1	0.76	0.96	0.85	27425
2	0.75	0.26	0.39	11510
micro avg	0.76	0.76	0.76	38935
macro avg	0.75	0.61	0.62	38935
weighted avg	0.75	0.76	0.71	38935

The precision refers to how many of the selected items from the model are relevant and this is calculated by dividing true positives by true positive and false positive. The recall tells how many relevant items were selected and it is calculated by dividing true positives by true positive and false negative. Finally, the f1-score is a measure of accuracy of the model, which is the mean of the model's precision and recall.

4. Conclusions

The model has decent precision values for the 1 and 2 values of 76 and 75 respectively giving us 70%+ relevant values in the model. As for the recall we had 96% and 26% respectively of relevant values selected and finally we had an accuracy score of 0.75 which tells us that this model performs pretty well for this value.

Once reviewing all the data the recommendations I may give to the city officials is that they know with this how many of these accidents have occurred in a place where road or light conditions and which type of crash happened in this places and lunch public projects to repair this errors to minimize the results of this factors. As for the car owners they could use this data to take extra precautions certain roads where the light condition, road condition and weather perform badly in order to avoid a severe accident.