Car accident severity in Seattle (2004-2020)

Cesar Garcia Jara

Applied Data Science Capstone

November 1, 2020

1. Introduction

1.1. Problem and Background

According to the National Safety Council (nsc.org) during 2019 and 2018 the US experience a decline in the number of roadway deaths. It is estimated in 2019 approximately 38,000 people lost their lives due to car crashes. Even though this decline is encouraging the total of people seriously injured in 2019 were 4.4 million[1]. This numbers shows why traffic accidents are an important public safety challenge. Further more, in a global perspective, recent projections have estimated that as income and vehicle ownership levels rise in the developing world, “the global number of road traffic deaths would increase by approximately two-thirds by the year 2020”[2]. This explains why information is needed in order to guide local and national governments in what are the driving factors and how to prevent car accidents and fatalities.

Our goal is to find if the currently available data can help determine the severity of the car accidents in the City of Seattle. We believe that using data and the proper analysis we can create prediction models, using data analytics.

1.2. Interest

No one city is exactly the same as any other, but only research and quality analytics can help understand what data is needed, how to build de model and what results to expect of that model. Once a solid model is developed we may be able to apply it to different cities and countries, not only US.

2. Data acquisition and cleaning

2.1. Data sources

The data used is published by the City of Seattle. The first one is called “Data_Collision”, a text file provided by the Coursera web site. This is an extract of the data provided by the City of Seattle with a full record of all car accidents in the city, with more than 30 different attributes and expanding from 2004 to the first months of 2020. The data is created using ArcGIS, program used by the city. [3]

The second data source is “City_Clerk_Neighborhoods.geojson”, a geographical file with the boundaries of all the neighborhoods in the city.[4]

2.2. Data cleaning

The “Data_Collision” file is provided by SPD and recorded by Traffic Records. The data is provided as a text file. The application provides geographical data, latitude and longitude of each record, but also 35 other fields with pertinent data for our research. The data set has 195K records in total.

In particular our model will work on predicting SEVERITYCODE, as dependent variable, based on the behavior of other independent variables like collision type, date of incident, number of persons injured or dead, if the driver was under the influence, weather, if the vehicle was speeding among others.

After reviewing and removing records with missing information the final dataset includes 187K records. The records removed had missing geographical location.

2.3. Feature selection

Upon examining the meaning of each feature, it was clear that there was some features that are nor relevant to the model. Features like SDOT_COLCODE, SDOT_COLDESC, ST_COLCODE, ST_COLDESC which are descriptions of other features, HITPARKEDCAR , CROSSWALKKEY, SEGLANEKEY were also discarded.

After discarding redundant features, I inspected the correlation of independent variables, and found several pairs that were highly correlated, SEVERITYCODE is correlated to PERSONCOUNT and VEHCOUNT.

The below listed fields are considered to proceed with our analysis:

Table 1 Feature Selection

Kept Features	Dropped Features	Reasons for Dropping Features
SEVERITYCODE	PERSONCOUNT: The total number of people involved in the collision. VEHCOUNT: The number of vehicles involved in the collision.	Features correlated to SEVERITYCODE
OBJECTID X (longitud), Y (latitude) INCDTTM and INCDTTM: The date and time of the collision.		
ADDRTYPE: Collision address type:Alley, Block or Intersection.		
COLLISIONTYPE: Collision type.		
JUNCTIONTYPE:Category of junction at which collision took place.		
	ROADCOND: The condition of the road during the collision. 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) UNDERINFL: Driver under the influence of drugs or alcohol.	Features that show little incidence. Good weather, road and light conditions do not contribute to a car accident. Similarly speed and under-influence has marginal presence in the total number of records.

3. Exploratory data analysis

3.1. Target variable

Our dependent variable is SEVERITYCODE, severity of the collision, which has the values of 3 for a fatality, 2b for a serious injury, 2 for an injury, 1 for a property damage or 0 for unknown. When reviewing the data we noticed only two values are recorded out of the 4 available.

When reviewing the geographical location of each accident we noticed areas that show higher density, as shown in Figure 1. Areas like Belltown and the Business District, show are very accident prone. Also roads like Reiner Ave S and 15th AveW to 15Th Ave NW.

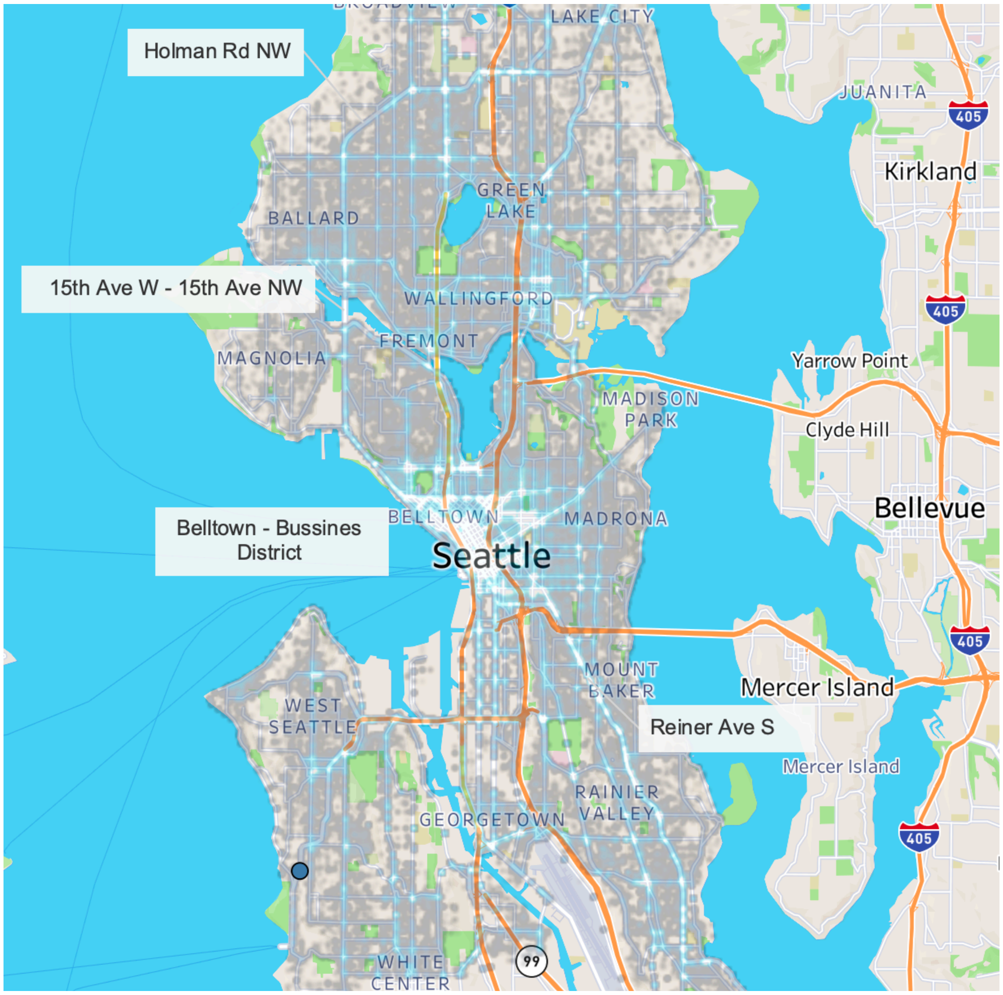


Figure 1: Density map of accidents in the City of Seattle.

When trying to find which neighborhood of the city with the most incidents we used the geographical data provided by the City of Seattle. In figure 2 it is noted the Industrial District show the highest incidence, even more than the Central Business District. This influenced by the largest area in the Industrial District. The third area with the most accidents is the University District.

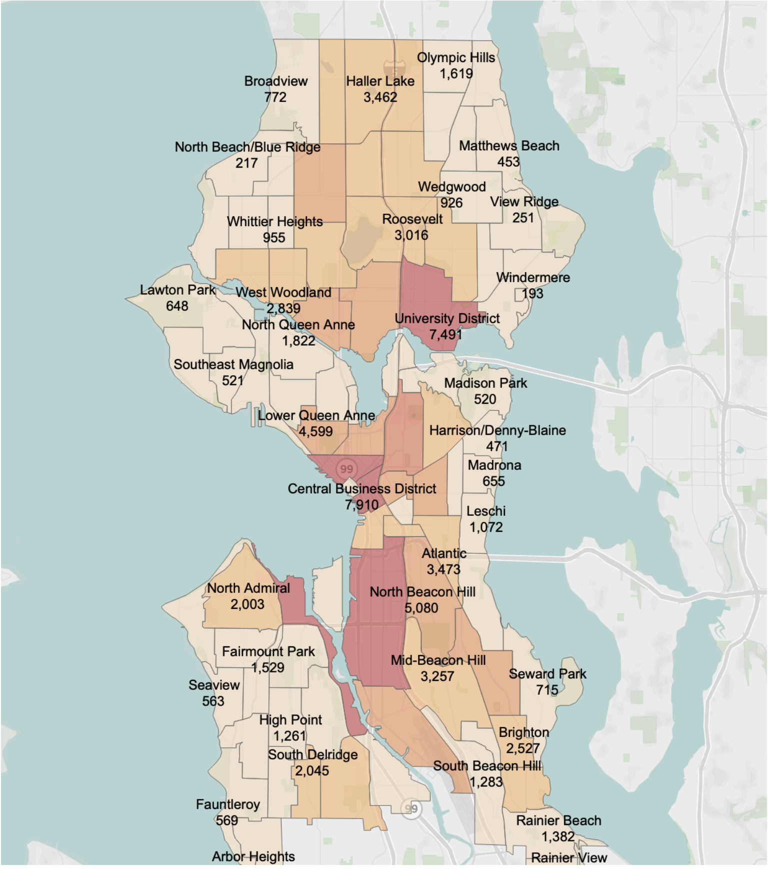


Figure 2: Seattle Neighborhoods with the most car accidents.

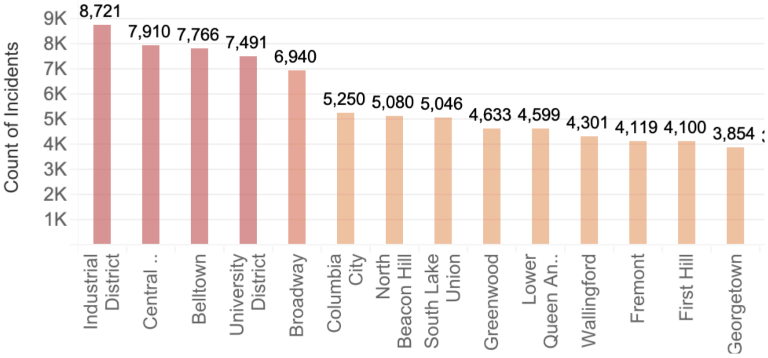


Figure 3: Number of car accidents by the Top 14 city neighborhood.

3.2. Relationship between Severity Code and Collision Type

There are several types of collisions and they affect differently the severity of the accidents. The most of the accidents, 25%, involve a parked car that causes property damage, but not personal injury; as shown in Figure 4. In order to make the model more efficient some of the attributes will be grouped: Cycles, Head On, Left Turn, Pedestrian and Right Turn will show as part of Other.

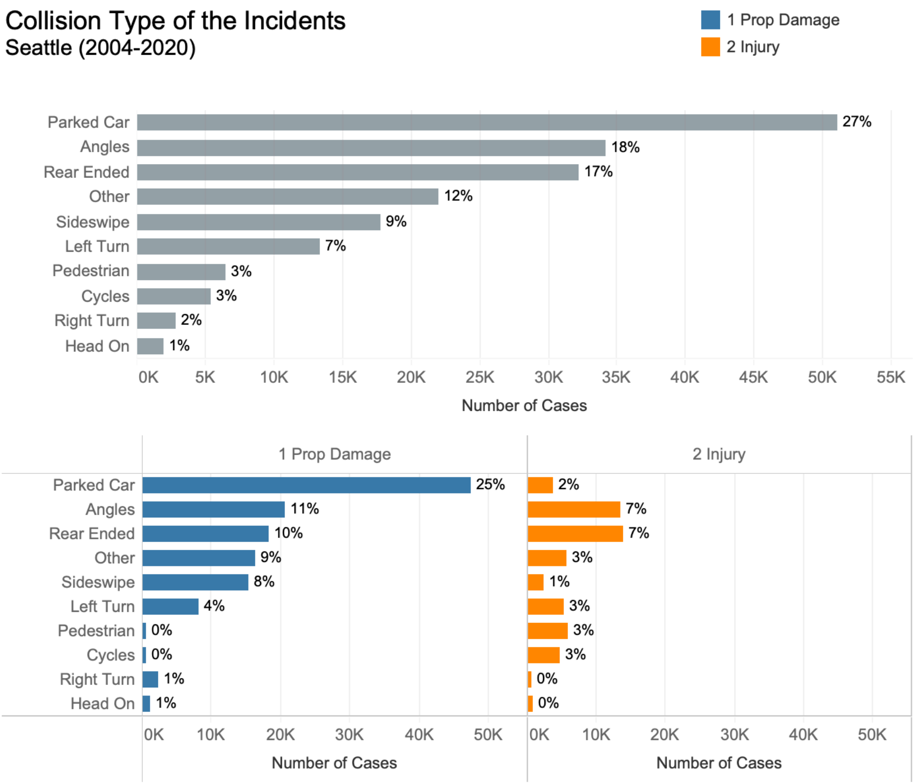


Figure 4: Relationship between Severity Code and Collision Type.

3.3.Relationship between Severity Code and Address Type

The feature used to identify if the accident occurs at an intersection or in middle of the block is the Address Type. Accidents at an intersection are 66% of the total, but they do not reflect equally in the severity of the accident. Half of the accidents in record occurred at mid block

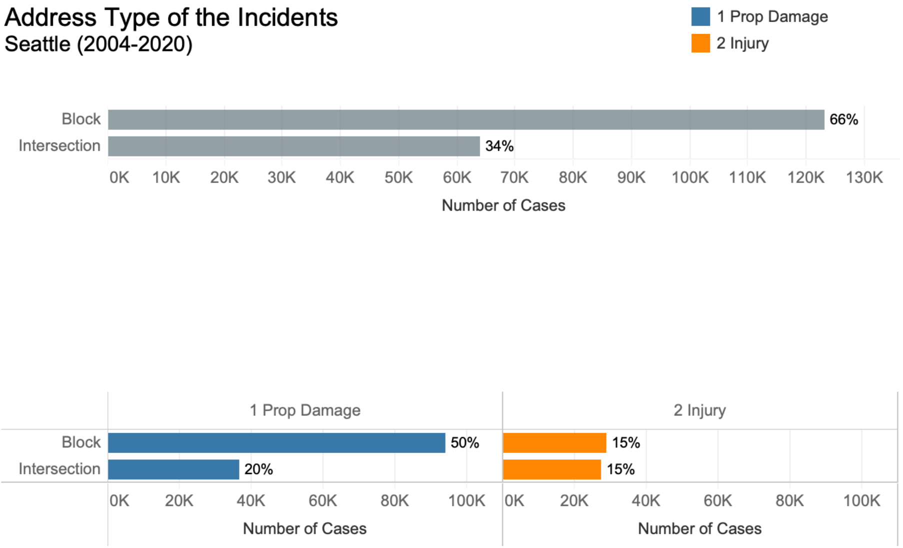


Figure 5: Relationship between Severity Code and Address Type.

and cause only property damage and that is three time the incidence of the accidents with personal injury; as indicated in Figure 5.

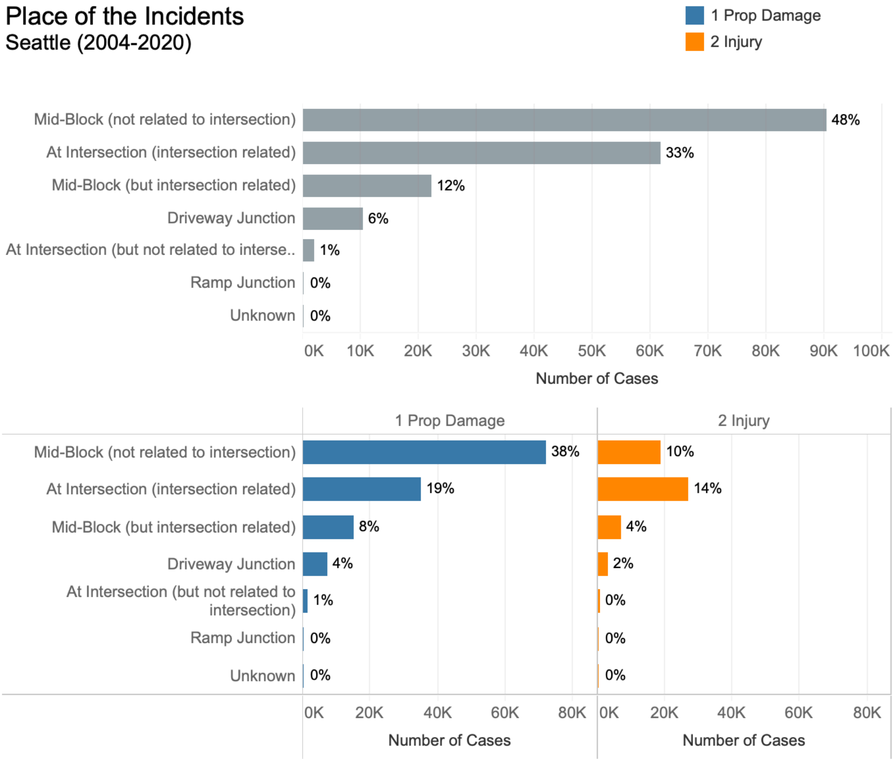


Figure 6: Relationship between Severity Code and Collision Place

3.4. Relationship between Severity Code and Collision Place

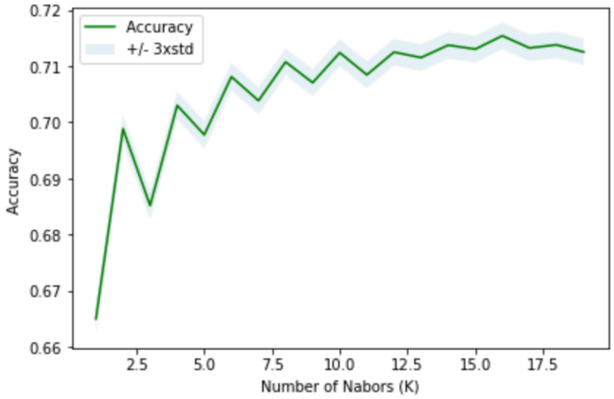
Junction place, or place of the incident, provides detail on the specifics of the location iof the junction of the collision. Most of the accident are categorized as mid-block but not related to an intersection, 48%. This accidents are mostly related to property damage, 38% of all accidents. See Figure 6.

In our final dataset we have dropped three features: At Intersection (but not related to intersection), Unknown and Ramp Junction; in order to make the model more efficient and considering their little present in the total data.

4. Predictive Modeling

We will focus on identify problematic areas and within those create clusters of locations, (using **K-mean** -ns clustering) and Logistic regression with the features that help the stakeholder identify problems and spark conversations for possible solutions.

4.1. Classification models



```
The best accuracy was with 0.7153797401499947 with k = 16
```

Figure 7: The value of K.

For the KNN model we used a K equal to 16 which provided the highest median, 0.7153. The training set provided a 75.59% accuracy and the test set reached 71.54%.

The second model to try is the Linear Regression. Once the model is trained and tested the accuracy on the test set reached 72.34%. A brief recap of the performance of the models is included in table 2.

Table 2. Performance of classification models.

Algorithm	Jaccard	F1-SCORE	Log Loss
KNN	71.54%	[0.8134132 0.40028934]	
LogisticRegression	72.26%	[0.81954538 0.40095819]	0.526

5. Conclusions

We were able to identify the main neighborhoods that have the highest number of incidents, Belltown, the Business District and the Industrial District; as well as the 2 roads that accumulate the highest density of accidents; Reiner Ave S and 15th AveW to 15Th Ave NW.

Some information may be missing or recorded incomplete. SEVERITY CODE only recorded 2 type of collision out of the 4 available, not considering the "unknown" value.

The three more common kind of collision types are: parked cars, a car hit by the angle or the car being rear ended.

Collision Type, Address Type, Junction Type, Latitude and Longitude are good predictor of the severity of a car accident.

Both the KNN and Logistic Regression algorithms present similar accuracy with the later being slightly better and also faster to perform.

6. Future directions

This research may need to expand the analysis to determine if, for example, the collision of parked cars has any relation with the location of nearby universities, in the University District, or retail location, in Belltown or the Central Business District.

“Incident Timestamp”, time stamp attribute, did not have the time value, only de date. This affected 16% of the data. Stakeholder may need to review internal process in order to improve the data collection and perform a more extensive analysis.

There is a huge opportunity to expand this research with more detail of the geography and transit loads of the city.

[1] The National Safety Council. Motor Vehicle Deaths Estimated to Have Dropped 2% in 2019. https://www.nsc.org/road-safety/safety-topics/fatality-estimates

[2] Kavi Bhalla et All. A Risk-Based Method for Modeling Traffic Fatalities. Risk Analysis, Vol. 27, No. 1, 2007

[3] https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv

[4] https://data.seattle.gov/dataset/Municipal-Boundaries/54dn-ah5p\