**Predicting the Severity of Car Crashes**

Lucas Kramer

September 19, 2020

**1. Introduction**

**1.1 Background**

We wish to understand the severity of a collision based on a variety of different aspects. We will focus on data that includes not only severity of collision and injuries, but also data that we feel could and would contribute to a collision. This information could be used in a variety of different aspects of the business world including but not limiting to: Insurance, Healthcare, Emergency Preparedness and Personnel, and Motor Vehicle sales and repair.

* + 1. **How the problem ties to the various business models**

**Insurance**: The severity of collision would indicate costs associated with such a collision. By being able to predict the severity of a collision based on some aspects of the surrounding environment, an insurance company could make better rates: increasing rates for drivers in poor conditions and decreasing rates for drivers under good conditions. You would only be able to use static variables such as location, climate, car type, etc. for this prediction as daily changing variables (except maybe averages) could not be included in a six month contract.

**Healthcare**: The severity of collisions would be very beneficial to healthcare. Knowing the conditions at which severe injury occurs would allow a hospital or emergency room to better plan both the need for supplies and the need to schedule personel at the appropriate times. This would allow the hospital to run more efficiently and smoothly.

**Emergency Preparedness**: The severity of collisions would allow cities, counties, and states to better plan their emergency personel including police presence, fire department size, and number of ambulances available at certain times. It would allow for long term planning making sure they have enough basic supplies and vehicles when static conditions, such as time of year, indicate that more severe accidents will occur as well as adjusting short term resources such as how many individuals are scheduled to be on call when easily changed variables, such as predicted rainfall, change.

**Motor Vehicle Sales**: If static conditions indicate there will be more severe collisions in a certain area, this will allow for better planning of supply as more vehicles will be needed to replace the vehicles that are totalled in the collision. It will also allow for the better planning of what types of vehicles are needed. If static conditions indicate that there will be more severe collisions, after a time the general public will want vehicles that are better at protecting the passengers in said vehicle.

**Motor Vehicle Repair**: If static conditions indicate there will be less severe collisions then businesses that repair vehicles can reasonably assume that the vehicles can be fixed or at least salvageable. This may indicate the need to hire more mechanics or expand garage space so that the business has the ability to handle a larger work load.

**2. Data Acquisition and Cleaning**

**2.1 Data Sources**

This data consists of crash data for the greater Seattle area. It includes data from accidents resulting in property damage to minor injury. The data consisting of intense injury or death is not included in this data. I would assume this is because of privacy and/or criminal investigations. The data is found [here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Data-Collisions.csv) and the metadata is found [here](https://s3.us.cloud-object-storage.appdomain.cloud/cf-courses-data/CognitiveClass/DP0701EN/version-2/Metadata.pdf).

**2.2 Data Cleaning**

All of the data came in one file, so no cleaning was required for multiple entries. However, there were a lot of columns that consisted of data names, unique identifier numbers, and duplicate data. All duplicate data was removed along with any unique identifiers. The unique identifiers were also removed, not just because they would not have any way to group and predict with them, but also because the numbers assigned might be dependent on the actual result. These occur after the fact and would not be able to predict the result. We also dropped many columns if they had over 30% NaN values as we felt that any column with that many NaN values will skew the data toward things that we don’t actually know.

As for issues within the dataset, there were several variables where there were many different options, such as weather, and some of these options had very few datapoints. Specifically, in weather, the variable values ‘Partly Cloudy’, ‘Blowing Sand/Dirt’, and ‘Severe Crosswind’ had only a few dozen entries, and hence were removed.

There were also minor revisions where some variables were recorded in multiple ways, such as ‘UNDERINFL’ which describes people under the influence of alcohol or drugs. The data was adjusted so that the variable values were consistent in such cases.

**2.3 Feature Selection**

When all adjustments were made there were 187548 data points with 26 features that were applicable to the various situations. These included the variable 'SEVERITYCODE' is the variable we wish to predict. Some of the variables are location variables such as ‘X’ and ‘Y’ are latitude and longitude. These values were intended to be used to find dangerous locations, but this idea was set aside due to time constraints. There were two variables that were time variables that we used to create a few graphs to compare different times of day. The rest of the data contains information about the conditions of the collision such as number of vehicles, lighting conditions, pedestrians present, and weather. These were used to a K-nearest neighbor model and a Decision Tree model. The complete list of used variables are listed below:

'SEVERITYCODE', 'X', 'Y', 'STATUS', 'ADDRTYPE', 'LOCATION', 'COLLISIONTYPE', 'PERSONCOUNT',

'PEDCOUNT', 'PEDCYLCOUNT', 'VEHCOUNT', 'INCDATE', 'INCDTTM', 'JUNCTIONTYPE',

'SDOT\_COLCODE', 'INATTENTIONIND', 'UNDERINFL', 'WEATHER', 'ROADCOND', 'LIGHTCOND',

'PEDROWNOTGRNT', 'SPEEDING', 'ST\_COLCODE', 'SEGLANEKEY', 'CROSSWALKKEY',

'HITPARKEDCAR'

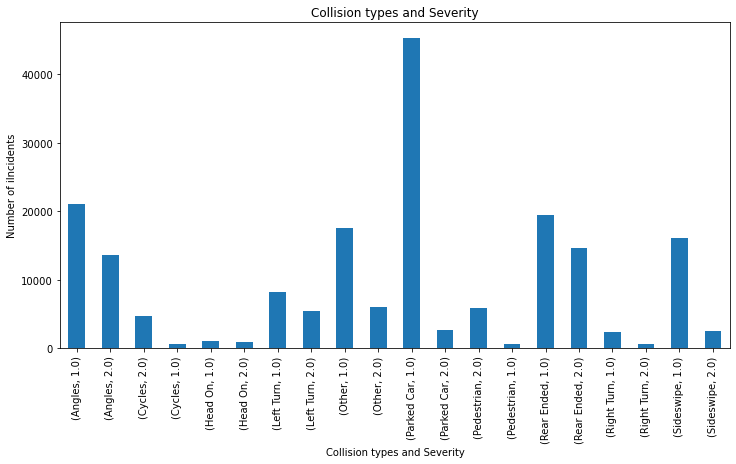
**3. Exploratory Data Analysis**

**3.1 Target Variable**

The variable of interest is SEVERITYCODE which indicates the severity of the car accident. The values that occur for this are 1 which means that there was only property damage or 2 which means that there were minor to moderate injury. There was no data on collisions with no damage or heavy bodily injury or death. It is suspected that this data is not publicly available due to criminal investigations and/or liability. The values and their descriptions were interchanged as necessary for analysis.

**3.2 Collision Type and Severity**

One of the initial graphs produced was a bar graph comparing collision types and severity.

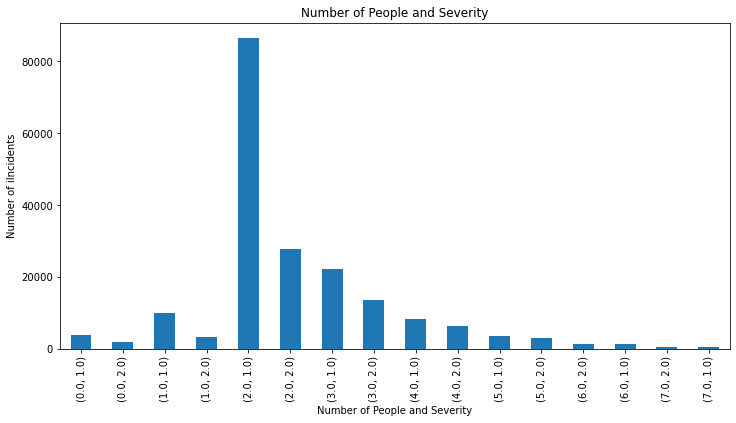


For space considerations we used the severity numbers instead of the names.

Based upon the ratio of heights, there are several variables that could indicate a difference in severity, such as Other, Parked Car, Pedestrian, and Sideswipe. You can see this more closely by looking at the two columns marked ‘Parked Car’ and the two columns marked ‘Rear Ended’. The ‘Rear Ended’ columns are much more equal implying that that type of collision implies more severe accidents.

**3.3 Severity and Number of Person involved**

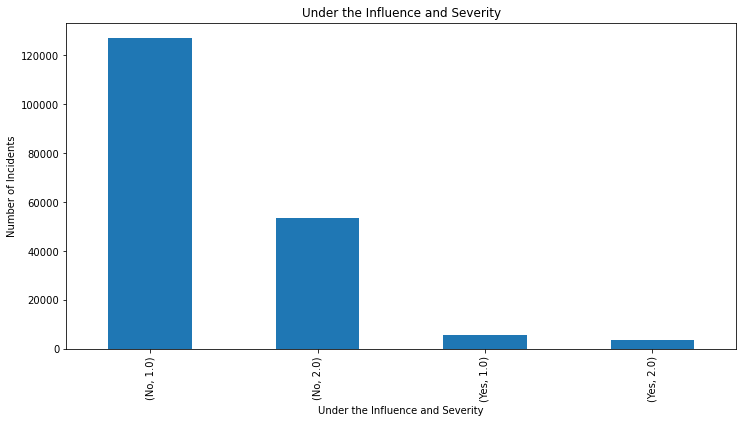
We speculated that if we increased the number of people involved in the accident, either by having more in a vehicle or by having multiple vehicles. That this would affect severity.



Here we can see the same sort of relationship between columns with the same number of people. When two people were involved there seems to be roughly 3 to 1 ratio of Property Damage to Personal Injury where with three involved there seems to be a 2 to 1 ratio of Property Damage to Personal Injury.

**3.4 Under the Influence and Severity**

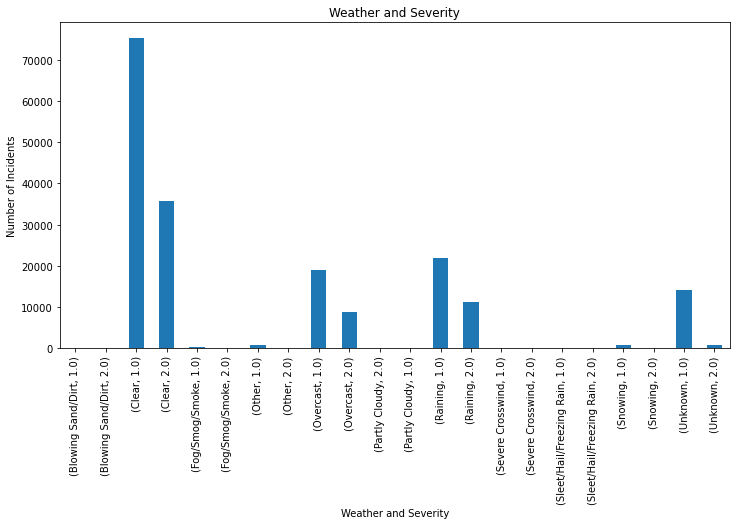
It is widely believed that drivers under the influence of alcohol or drugs increases the severity of accidents. While the number of collisions with those under the influence was small compared to those that were not, we can see there seems to be a relationship between the variables.



It should note however, that the most severe collisions are not included in this data so we should be careful about drawing strong conclusions.

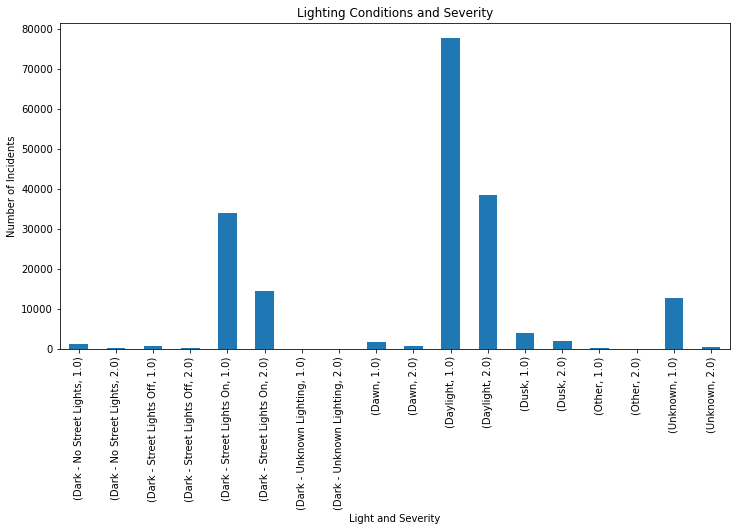
**3.5 Road Conditions**

Weather is thought to be a major contributor to damage and injury during a collision. It can reduce visibility and it harder to stop in time to avoid a collision. We found the following graphs.



From the graph, it is not easy to see if there is a relationship. We can see however that most of the collisions do not happen in some of the most extreme conditions. It might be that there are a lot less people out on the roads at this time and hence there are less collisions. It is impossible to say from the data. This does indicate that we cannot include some of these datapoints in the Machine Learning portion of this report

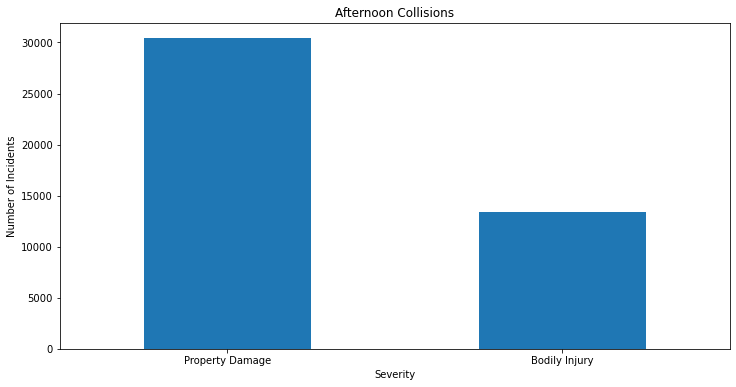
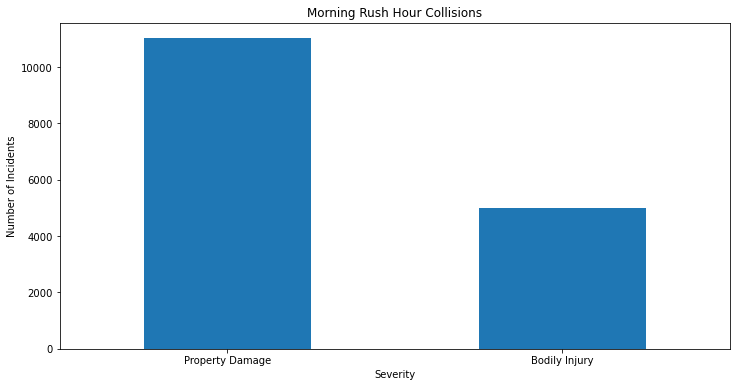
Lighting conditions are also believed to play a role in driver response time and hence collisions.

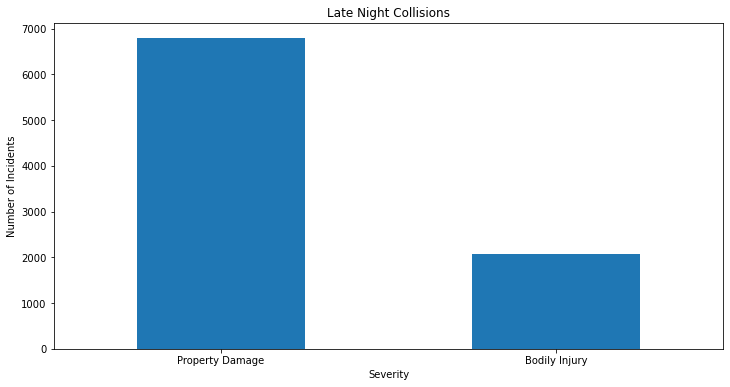
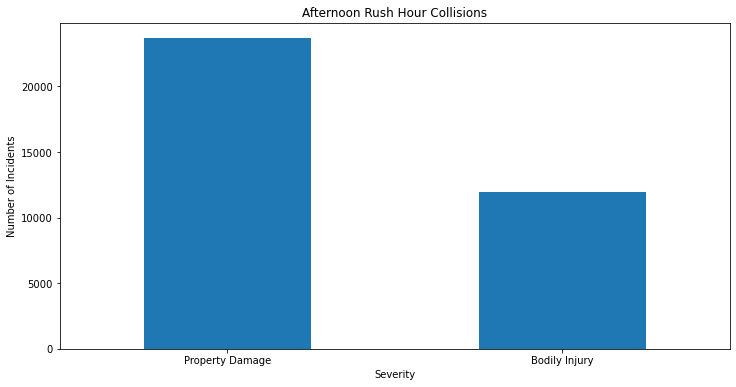


Again, it is less obvious that this plays a role in the same way we are uncertain about Weather. We also have a small number of certain variables which we will take under consideration in the Machine Learning component of the report.

**3.6 Time of Day**

While lighting conditions may or may not play a role. We wanted to consider certain times of day within those lighting conditions specifically. We looked at the afternoon hours, both rush hour times, as well as overnight, specifically when the bars and restaurants must close.



As you can see, there seems to be little consistency in the times of days causing a change in ratio of the types of collisions. It should be noted however, that the Late-Night Collisions ratio appears to be smaller. It is unknown at this time what the reason for this would be.

**3.7 – Other Comments**

There were several other comparisons done, such as Pedestrian Count, Bicyclists, as compared to Severity, but the number of such incidents was very low compared to other types of indicates and are not included.

**4. Predictive Modeling**

There are two types of models, regression and classification, that can be used to make predictions about the severity of the accidents. By far, most of our data is categorical in nature. Those of the variables with values, such as number of pedestrians, only have a couple values that are different than the vast majority of the data. This informs us that regression will not be a good indicator of these situations regardless of the accuracy of the model. We simply don’t have diverse enough numerical data to support such an analysis. Hence we will focus strictly on classification models.

**4.1 Classification Models**

We performed two classification models, a K-Nearest Neighbors model and a Decision Tree model. Both of these models required some slight manipulation of the data out of categorical values into numerical values. This was done with One Hot Encoding techniques. The data was then split into training and testing data in the ratio of 80/20 to be able to test the accuracy of the models we will be creating.

**4.1.1 K-Nearest Neighbors**

We investigated many values for the closest neighbors to consider. Due to calculation constraints we restricted the number of nearest neighbors to consider to be 15. Anything more than this was outside of the ability of our infrastructure to calculate. If we studied this data with a cloud computing software, we would drastically increase this value to see if we considered many more neighbors, if it would be beneficial.

We found the most accurate indicator for the calculation was when we considered 8 neighbors at a time. The Jaccard score for this model was 7.2 and the F1 score was 7.1 indicating that this model will currently be right roughly 70% of the time. If we consider the 8 closest types of accidents with those conditions, 70% of the time, we will predict the correct severity of the collision.

**4.1.2 Decision Tree**

In this model, we try to partition our data via the variable’s values. We tried depths of trees ranging from 3 to 10. This restriction was due to the limits of our computing power. If we had access to cloud computing, we would extend the depth of the tree out too see if there is a better indicator. The best indicator was at a depth of 6 with a Jaccard score of .75 and F1 score of .71. We then created a visualization of this graph. Due to its size, it is found in Appendix A.

According the findings here the variables that indicated the most identifying variables to indicate severity are hitting a parked car, the number of people involved in the collision, drivers under the influence, and levels of daylight.

**5. Conclusion**

In this study, we analyzed the relationship between a wide variety of variables and the severity of collisions that occurred in Seattle, WA from the year 2004 until May 2020. I identified that hitting a parked car, the number of people involved in the collision, drivers under the influence, and levels of daylight were the variables that indicated the severity of the collision the most. These models will be very useful to a variety of industries. Insurance companies can use them to help predict collisions and hence payouts the company would make. Cities and governments can design information campaigns to help reduce the number and severity of accidents based on my summary. They may also use this model to change fines, penalties, and possible charges in their civic code to try and influence the behavior of drivers to decrease accidents and severity.

**6. Future Directions**

We were able to predict severity in 70% of accidents using the models created. I think, if given different types of variables, we may be able to increase the reliability of these models. For example, if we included collisions where there were more severe injuries, we may be able to predict those as well. This would be especially useful, since if we could prevent or reduce these types of collisions, we could potentially save lives. We also do not have information on past behavior of the individuals in these collisions and information about previous speeding tickets or collisions could help us understand and predict the severity more accurately.

Also, we were unable to use the location predictors for our models. We believe that if these are incorporated, they could indicate what areas of the city or even individual intersections are more indicative of collisions and collision severity. This could be used to improve infrastructure. I would highly recommend including this in future analysis.

**--- Appendix A – Decision Tree Visual**

Due to its size and size and shape, I have shrunk down the graphic. If you cannot resize the graphic to the desired size, I can send you the pgn file directly. Please email me at lucas.kramer@briarcliff.edu.

