**1. Introduction**

**1.1 Background**

The city council of Seattle has recently made a pledge to try and reduce the number of injury collisions in their city after a high-profile case involving a kid getting injured made national news.

Seattle has been ranked as one of the nation’s worst cities to drive in, with the cost of car ownership being extremely high, accidents being extremely likely, days of lots of precipitation, and traffic infrastructure being inadequate.

<https://wallethub.com/edu/best-worst-cities-to-drive-in/13964/#methodology>

https://komonews.com/news/local/study-seattle-ranks-as-one-of-nations-worst-cities-to-drive-in

They fear that with the Covid-19 pandemic, more people will be driving instead of taking public transportation, and want to try and mitigate the number of injury collisions.

However, many members hold different opinions on what the most effective solution for this is.

**1.2 Problem**

They've pulled up the SDOT accident collision data to see which factors (if any) can predict if an accident will be property damage or an injury collision (labeled as SEVERITYCODE within the metadata).

From there, they can see which factor has the greatest predictive power and implement the appropriate solution.

Here are some of the different competing factors (and solutions) that many of the city council members believe are the root cause:

1. Certain locations are dangerous, so if this has the highest correlation, then we need to find the intersections with the highest number of accidents to address this.

2. Time of day may play a large role, so if this has the highest correlation, then we may need to install more streetlights.

3. The type of address (Block, Intersection, Alley) and Collision (Rear-end, Left Turn, etc.) may play a role, which means reviewing how the city approaches designing those types of roadways.

4. Other miscellaneous factors, such as # of pedestrians or bikes, may play a role.

This project will examine these hypotheses one by one to see if these may be predictive factors in injury collisions and see if there are any others that may be useful in coming up with calls to actions.

**1.3 Interest**

In addition to figuring out things for Seattle citizens and the city council, this model may help to predict injury collisions in similar cities as Seattle (for example, Portland) where a similar climate or infrastructure currently exists.

**2. Data acquisition and cleaning**

**2.1 Data sources**

I used the provided dataset but supplemented it with additional data such as GeoJSON’s for neighborhood and Zip codes. There were a number of missing values within a number of the fields, which led to me having to adopt certain approaches talked about within previous courses. I included a dataset with specific boundaries of Seattle (Neighborhoods, Zip codes) to see if that might help in defining specific problem areas for injury collision within Seattle.

**2.2 Data cleaning**

To first start using this data, I first had to deal with a number of problems within the dataset.

First, several identifiers, such as INATTENTIONIND and UNDERINFL, could not be used due to the dataset missing too many variables. As a result, I removed those columns from the dataset as they would not be good predictors.

Secondly, some identifiers, such as OBJECTID and INCKEY, used unique ESRI or other unique/secondary identifiers, that could not be correlated without a paid subscription to ARCGIS or access to seattle.gov. I also eliminated a number of these columns for the same reason.

Lastly, some of the data values were organized in a way which made standardization nearly impossible. For example, LOCATION was not useful due to the non-standard way that it was categorized, and SDOT\_COLDESC was not able to be parsed in a meaningful way. Since a number of these columns were redundant, I ended up not using these predictors.

For the values remaining, I removed the missing rows of data as it was not reliable to use any of the other cleaning techniques learned (such as averaging out a column of data).

**2.3 Feature selection**

After data cleaning, there were 184,346 samples and 39 features in the data.

However, upon examining the meaning of each feature, there were a number of features that were considered redundant. For example, Location, which is a description of the block where the accident occurred, can also be found through the usage of X and Y co-ordinates. Likewise, SDOT\_COLDESC, a description of the incident, can also be replicated with SD\_COLCODE.

|  |  |  |
| --- | --- | --- |
| **Kept Variables** | **Dropped Variables** | **Reason for Dropping variables** |
| SEVERITYCODE | SEVERITYCODE.1, SEVERITYDESC | Similar values |
| X,Y, ST\_COLCODE,SDOT\_COLCODE | LOCATION | Similar values |
| LIGHTCOND, ROADCOND, WEATHER | INCDATE, INCDTTM, ST\_COLDESC, SDOT\_COLDESC | Similar values |
| ADDRTYPE, COLLISIONTYPE, PERSONCOUNT, PEDCOUNT,  PEDCYLCOUNT, VEHCOUNT | JUNCTIONTYPE, HITPARKEDCAR | Similar values |
|  | INATTENTIONIND, UNDERINFL, SPEEDING, PEDROWNOTGRNT, EXCEPTRSNCODE, EXCEPTRSNDESC, STATUS | Too many missing values |
|  | SEGLANEKEY, CROSSWALKKEY, INTKEY, SDOTCOLNUM, OBJECTID, INCKEY, COLDETKEY, REPORTNO | Unique keys which are not useful for predicting SEVERITYCODE |

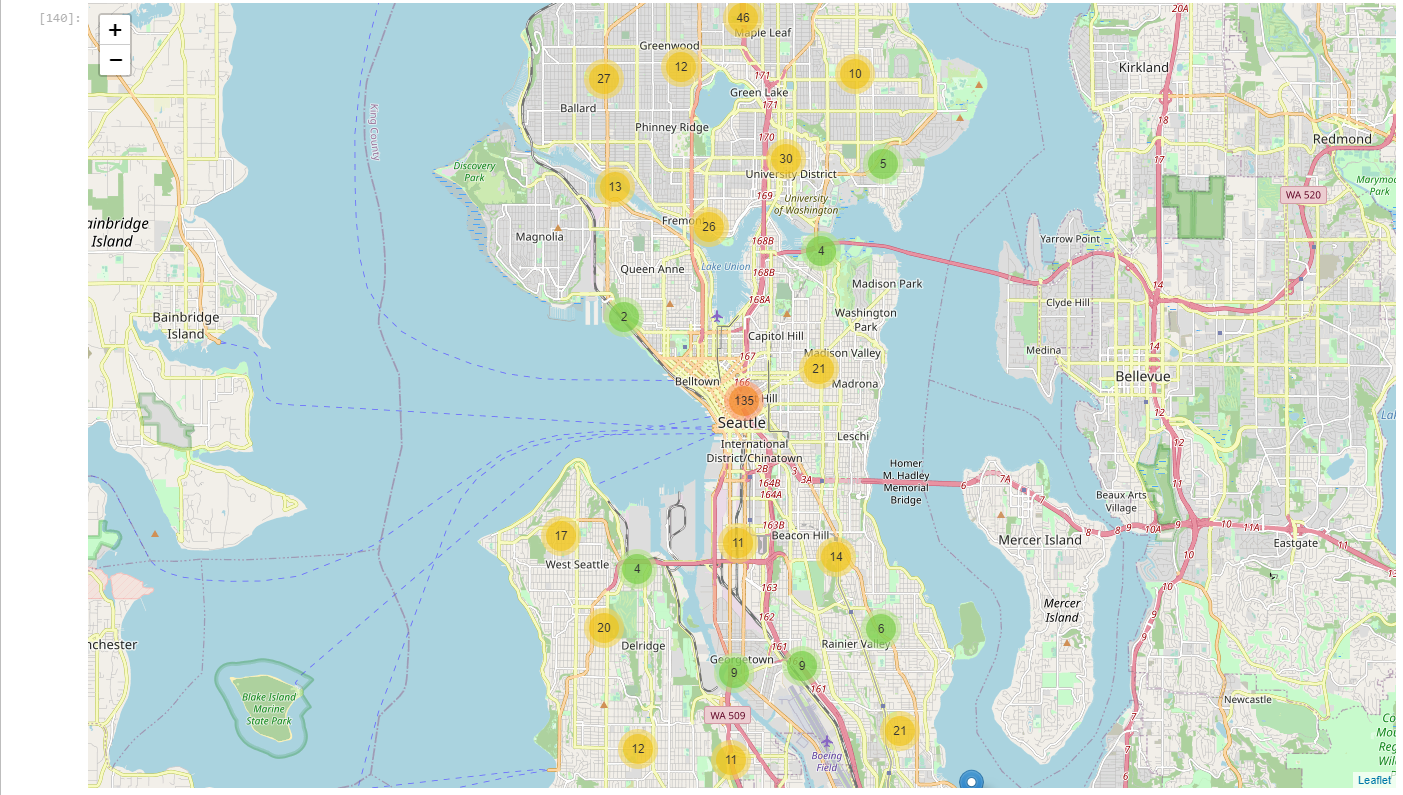
After filtering out redundant features, as well as those missing too many values and those not relevant to the problem at hand, I was left 13 features to use as predictors as well as the value that I was predicting, SEVERITYCODE.

**3. Exploratory Data Analysis**

**3.1 X-Y Co-ordinates**

One of the discussed hypotheses was that there may be high risk areas within the city that need to be addressed. For this to be true, there should have been certain areas with higher than normal accident collisions.

The X-Y coordinates were easier to work with when compared with LOCATION, so I cleaned up the X-Y data and exported a limited set of co-ordinates to export on to a Folium Map. From initial glance, it looked as though there were some promising results when the data was first returned.

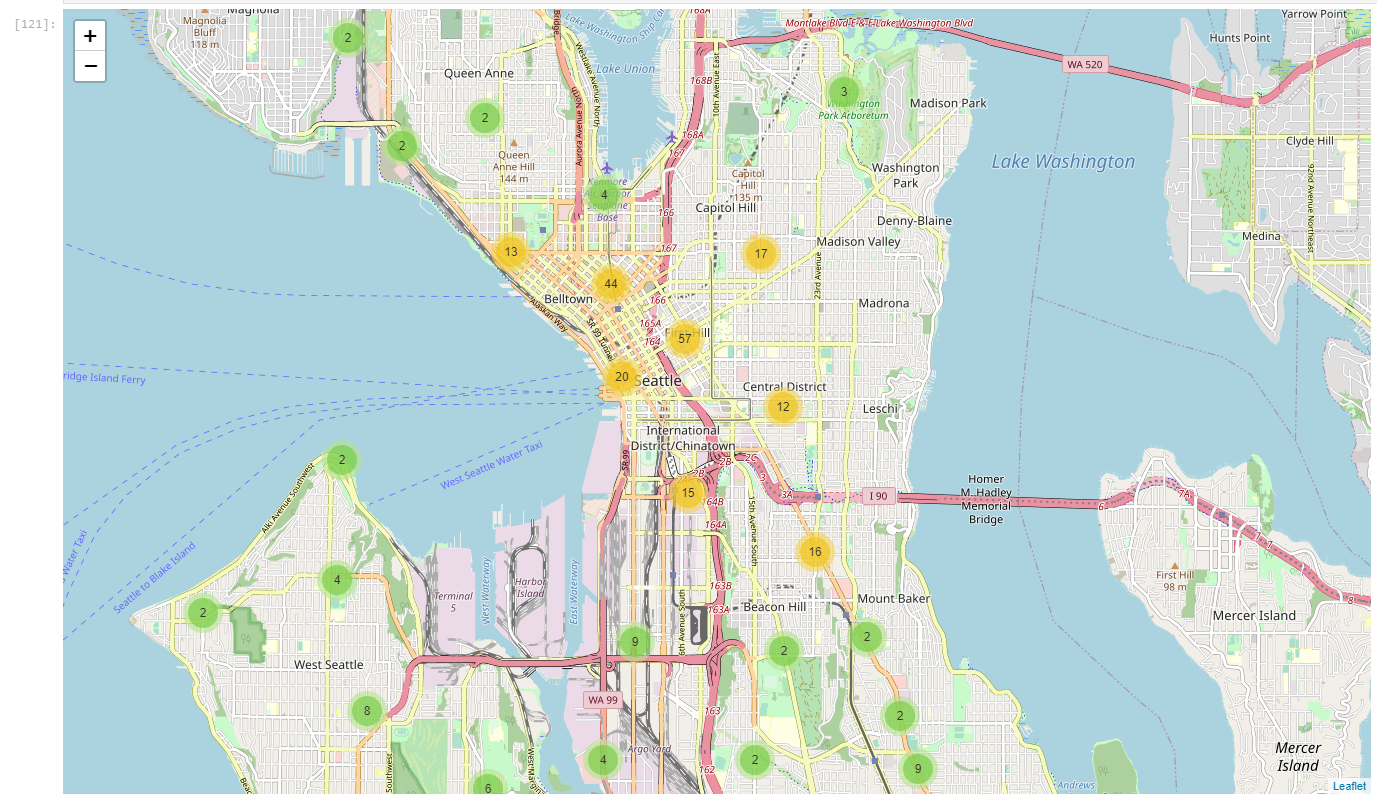


It seemed as though there were some clusters of data which seemed to suggest that certain areas, particularly Downtown Seattle, had higher than normal accident collisions. When another map was created, specifically sampling data from injury collisions, it seemed to strengthen this result.

Map

Description automatically generated

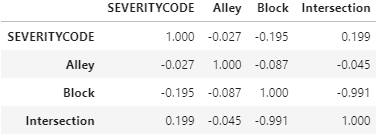
While it might be ideal to implement initiatives across all neighborhoods, it seems there is one specific area where implementing changes would be most beneficial. Another version of the map, which examines where more than one accident has occurred, seems to solidify this approach.



However, simply looking at the location of accidents is not a good predictor by itself. While there are certain locations that have higher numbers of severe accidents, that is simply because there are more accidents within that area. There also seem to be places where there are a high number of accidents, but few of them are severe. Therefore, we need another metric to examine this.

**3.2 Type of accident and street**

To figure out if any of the factors are good predictors for injury collision, I first looked at the values ADDRTYPE and COLLISIONTYPE to see if there was anything about the accidents that might stand out. To do that, I created several dummy variables for both of these variables and compared them to SEVERITYCODE to see if anything stood out.

Table

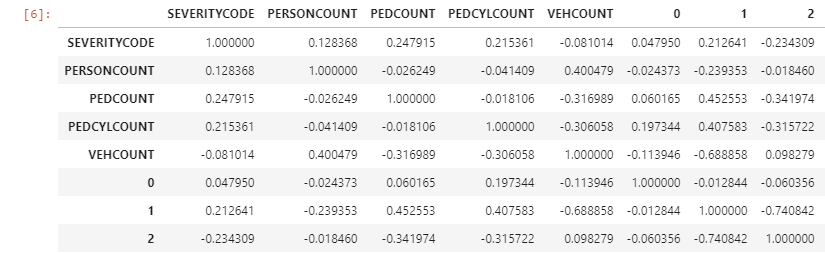
Description automatically generated

A couple of correlations stand out in this list. Intersections seem to have some correlation with Severity, as does two other categories: Cycles and Pedestrians. The hypothesis here is that the most significant factor affecting the severity of the accident is the number of person involved who are not in vehicles.

Therefore, it makes sense to look at another set of related variables.

**3.3 # of Person/Cycle/Vehicle Count.**

There is a set of variables that keeps track of the number of persons and transportation involved. Looking at this chart, we see that there are a number of slightly significant correlations here: both PEDCOUNT and PEDCYLCOUNT seem to be indicators of SEVERITYCODE, as well as when the VEHCOUNT is set to 1.

****

This seems to further back our hypothesis: injury collisions are more likely between a single vehicle and alternative method of transportation, such as a cyclist or pedestrian.

**3.4 Time of day**

Another one of the hypotheses suggested was that time of day might play a role in types of collisions.

Table

Description automatically generated

However, looking at the LIGHTCOND data, which showcases the time of day, does not suggest that this plays a role in the severity of accident collisions.

**3.5 Weather**

Another one of the hypotheses suggested was that weather might play a role in types of collisions.



However, looking at the ROADCOND and WEATHER data, which showcases the condition of the weather and road, does not suggest that this plays a role in the severity of accident collisions.

**4. Predictive Modeling**

There are two types of models, regression and classification, that can be used within predictive modeling. However given that SEVERITYCODE, is a discrete variable with two classes, property damage and injury collision, we will be using classification predictive modeling.

For that end, we’ve talked about several variables that might be good predictors: PEDCOUNT, PEDCYLCOUNT, (ADDRTYPE = Intersection), (VEHCOUNT = 1), and X & Y data. However, these variables should be further simplified to make the modeling easier.

Given that VEHCOUNT was a subset of the PEDCOUNT and PEDCYLCOUNT (since if there was 1 vehicle, there must either be 1 pedestrian or cyclist), I eliminated that as a variable. I also eliminated X&Y data, as it was simply a measure of where the initiative should be located.

Therefore, there are two variables that should be generated:

* Whether the accident involves a non-vehicle (i.e. pedestrian/cyclist) or not.
* Whether the accident occurs at an intersection or not.

And from there, seeing whether these variables can be useful in predicting one of two classes, SEVERITYCODE = 1 and SEVERITYCODE = 2.

For this part, I used Logistic regression and random forest models. Logistic regression returned the more accurate predictive model vs. Random Forest.

A picture containing table

Description automatically generatedA picture containing receipt, screenshot

Description automatically generated

**5. Discussion/Conclusions**

The predictive factors for accident severity make sense for a number of reasons regarding the city of Seattle.

First, Seattle has been classified as one of the worst cities to drive a car in. This may result in people using other modes of transportation, such as buses, bicycles, or simply walking, when their destinations are nearby.

As a result, places where pedestrians or cyclists are most common, such as the downtown Seattle area highlighted by the X & Y data, are where accidents between a single car and a pedestrian/cyclist are most likely to occur.

As for intersections, these are places where pedestrians/cyclists are most likely to cross (due to the presence of crosswalks), so it makes sense for accidents to occur there rather than the middle of the block.

Based on these factors, it suggests that the city council should start their initiatives to reduce injury collisions in Downtown Seattle intersections. Certain measures, such as speed limit trackers or speed bumps to reduce car speed as well as possibly protective barriers for crosswalks may have an effect on reducing the amount of injury collisions that occur.

**6. Future work**

One of the major hurdles that I had to work around was not being able to implement ArcGIS into the Jupyter notebook. As a result, I could not conduct much more than exploratory analysis on things such as Location and X/Y data. Through value\_counts, I found that there were several intersections where a lot of collisions occurred, but I could not narrow it down to a specific place.

The other thing I would probably do is try to find supplementary data which would improve my classification models. I was able to achieve 74% accuracy within the classification model, but it could be further improved by adding additional variables such as high/low speed or other crosswalk data.