**Seattle Collision Severity**

**Introduction/Business Problem**

The problem at hand is how to predict the severity of a car accident given the dataset. Car accidents can happen in a multitude of ways and not all are alike. They can happen anywhere, but do they happen more frequently in certain places? Different vehicles have different accident rates and our dataset includes many variables outside of just these two. The end goal of this project is to help predict the severity based on all these variables, but also have it become interpretable by a large audience. Our audience would be neighborhoods where we see the common predictors are more likely than the norm. Value resides in how well you can identify the predictors, which can hopefully be used to prevent accidents or at a minimum reduce severity in the long run. I’m very curious about the dataset and trying multiple approaches to see what methods yield the most valuable insights.

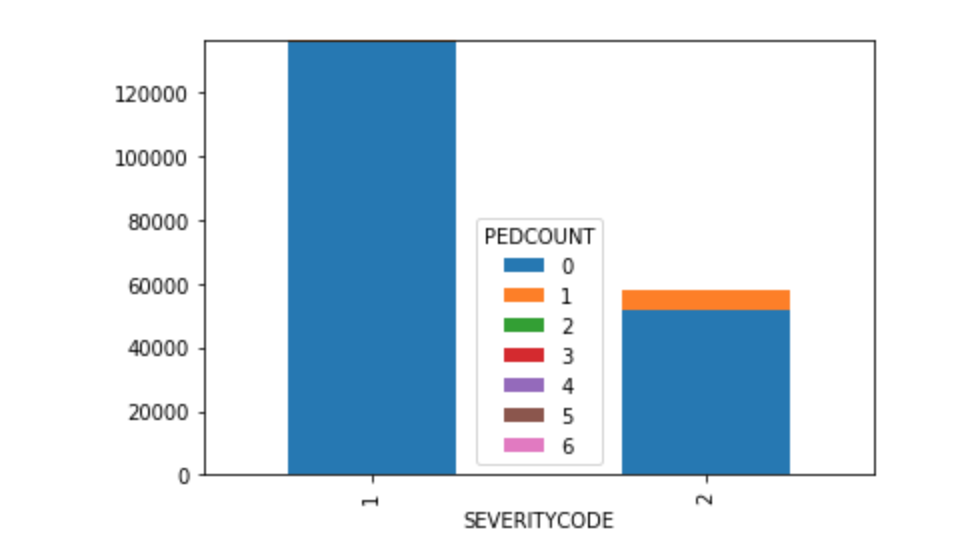
**Data**

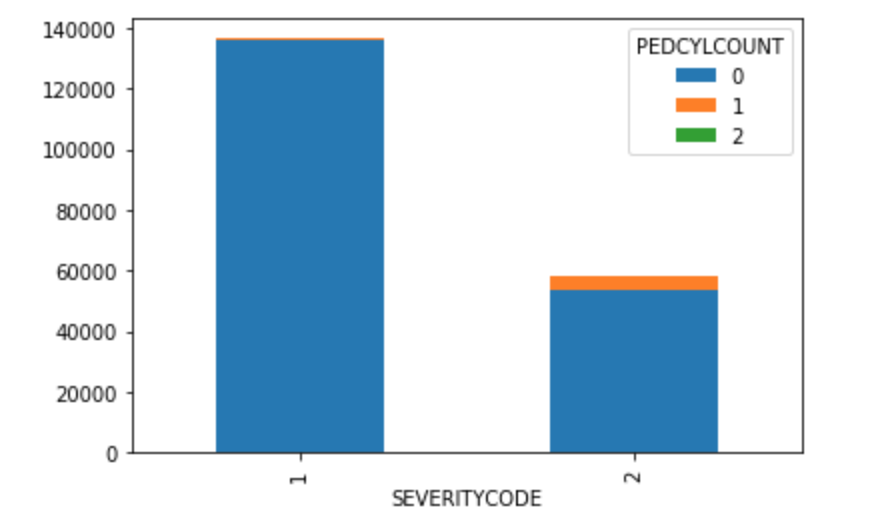
The data is structured well and consists of 194673 observations of 37 variables, numeric and categorical. The first column is the severity code which is either 1 or 2 which indicates that we will use some logistic regression to solve if an accident was or wasn’t severe rather than the specific severity on a continuous numeric scale. There is also locational data so we can see if there is a higher density of severe accidents in a specific location. Most of the variables that we will use as predictors are categorical which are better used to solve the problem because communities can take action if they fall into the vulnerable categories, rather than adjust a numerical measurement accordingly.

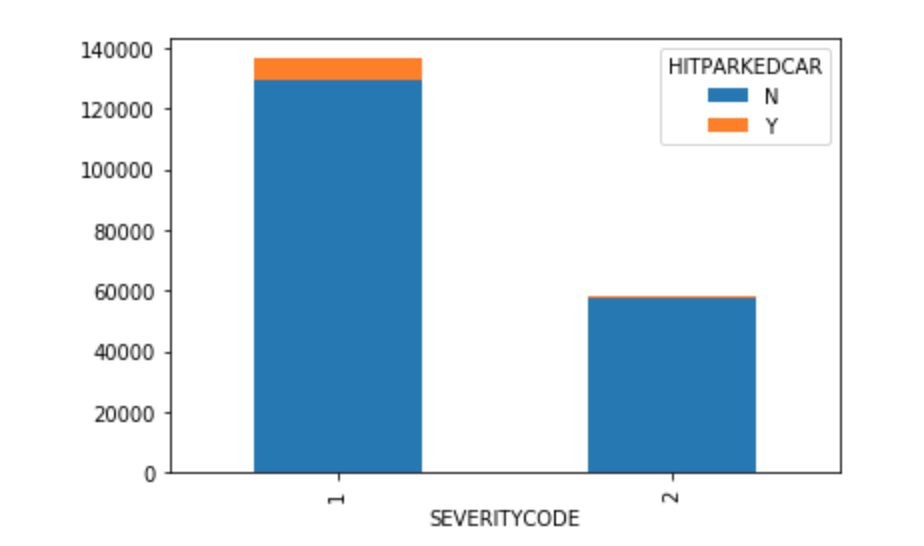
When exploring nulls and duplicates a few concerns were raised. For instance, we see that over 75% of observations in the speeding and inattention columns were null. If we had not observed this, when we get to 3e in our analysis we would’ve assumed all observations of crashes are from inattention and speeding, but we know thats not the case from cleaning the data. We leave all observations in. There were no duplicate rows so we didn’t remove any rows.

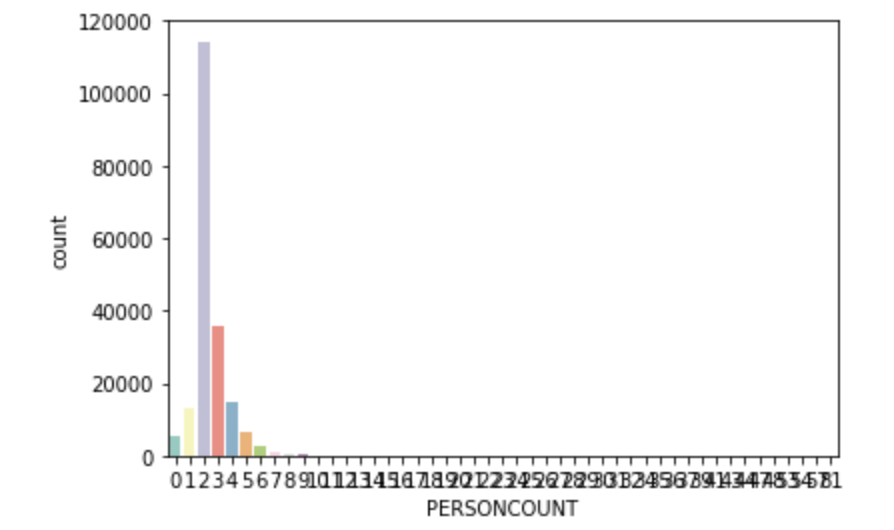
**Exploratory Analysis**

**Who**

To dive into the who of our dataset I looked at four key fields. These fields told us whether or not pedestrians, cyclists, or parked cars were involved and how many people were involved in the crash. These fields were all explored individually with the division of severity code.



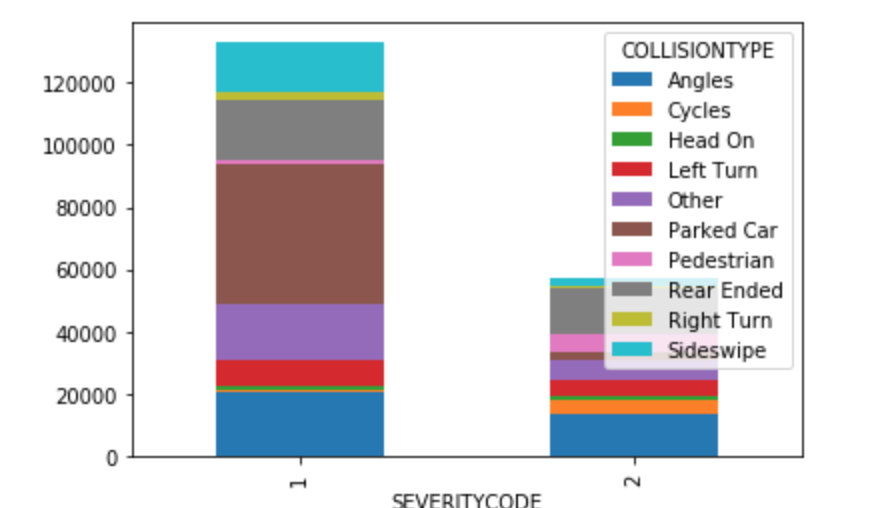




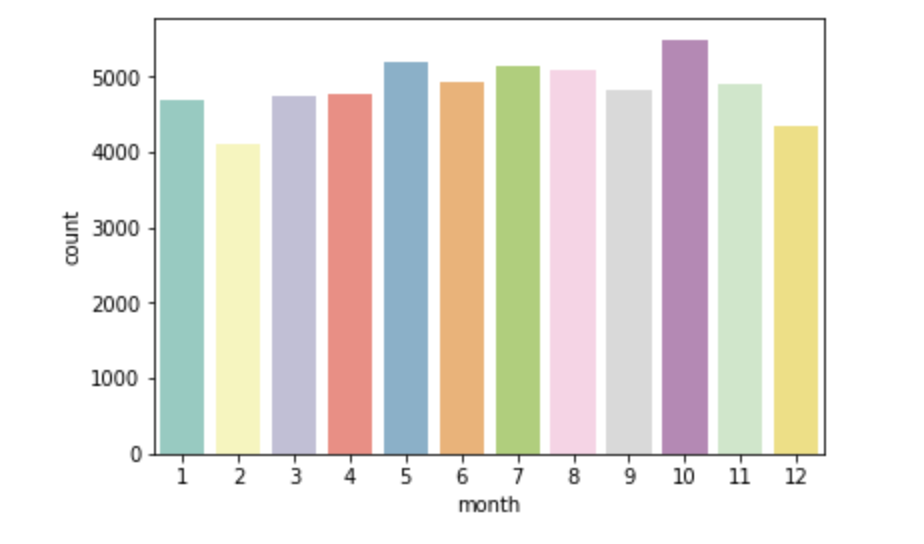
From the four graphs above, we can observe a few things regarding the dataset. Firstly, there is a much higher frequency of only 2 people in a collision than any other number. Meaning it’s mostly just 1 person in each car at the time of the collision. Also, we see that pedestrians usually aren’t involved, but when they are its more likely that it is severe rather than not. The same case is made for cyclists. The opposite occurs for parked cars which leads us to believe this could be parking lot collisions.

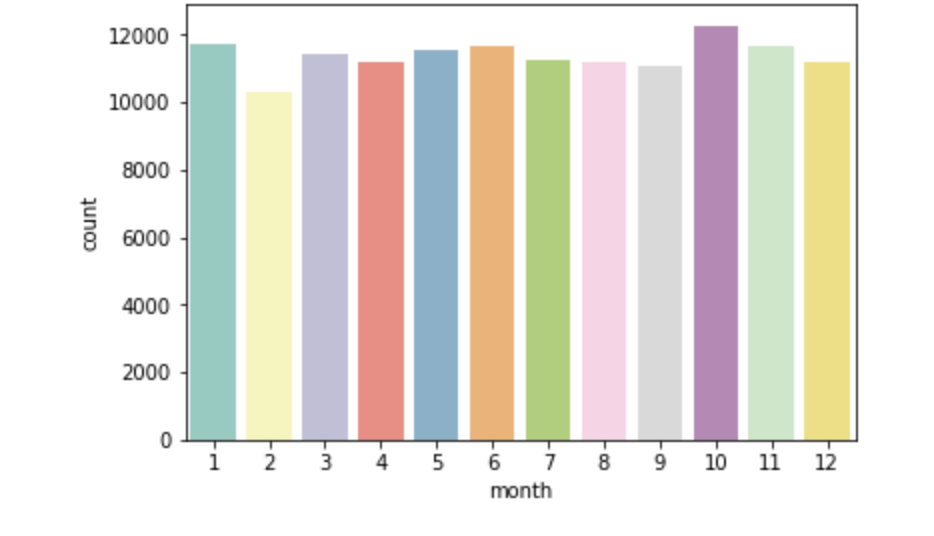
**What**

To explore the what regarding the dataset I looked at only one variable, the collision type. Separating it by severity code we can see a few things. From our who analysis the What supports that parked cars are not severe and make up a large majority of non severe crashes. Additionally supporting our who analysis we see way more cyclists in the severe column. Interestingly enough, left turns are more dangerous than right just observing proportions from this graph.



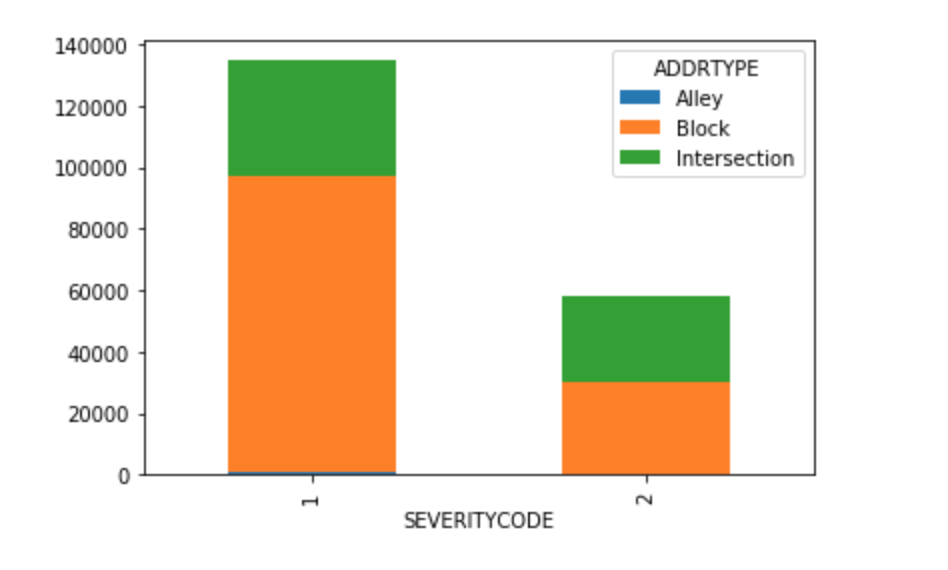
**When**

In exploring the dataset I found that the datetime column doesn’t contain many times so unfortunately I couldn’t explore the exact time of day that severe and non-severe accidents happened. I graphed month instead. The first graph below is the severe frequencies and the lower one is the non-severe. We don’t see a huge change between the two. It is curious however that most accidents occur in october for both categories. February with the least for each as well.



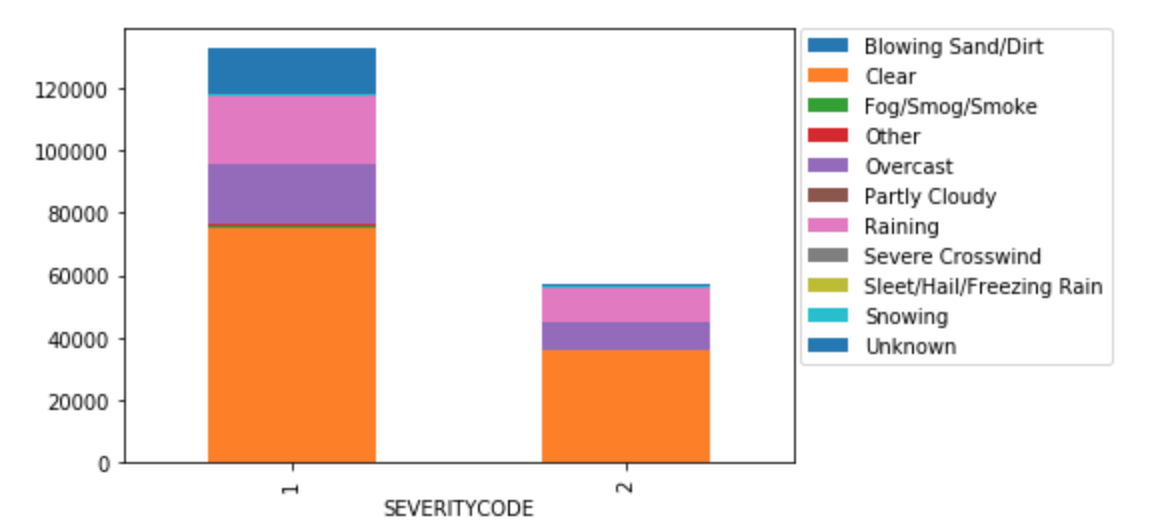
**Where**

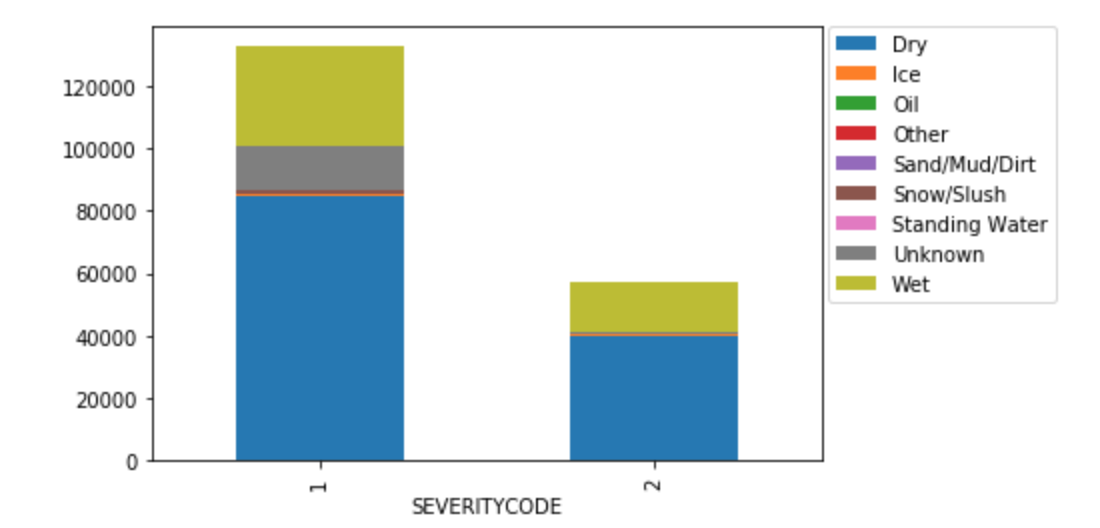
This part of the exploratory analysis was also just one variable as I will get into clustering in the modeling portion. This is just to see at a glance what type of place the collision occurred rather than where exactly in geospatial coordinates. Alleys seem to be very safe as we don’t see any visible blue. Further, intersections make up nearly 50% of severe crashes, where as they look to be around 25% of non-severe. This is probably due to the nature of the intersection with colliding directions of traffic and more traffic laws to obey to.

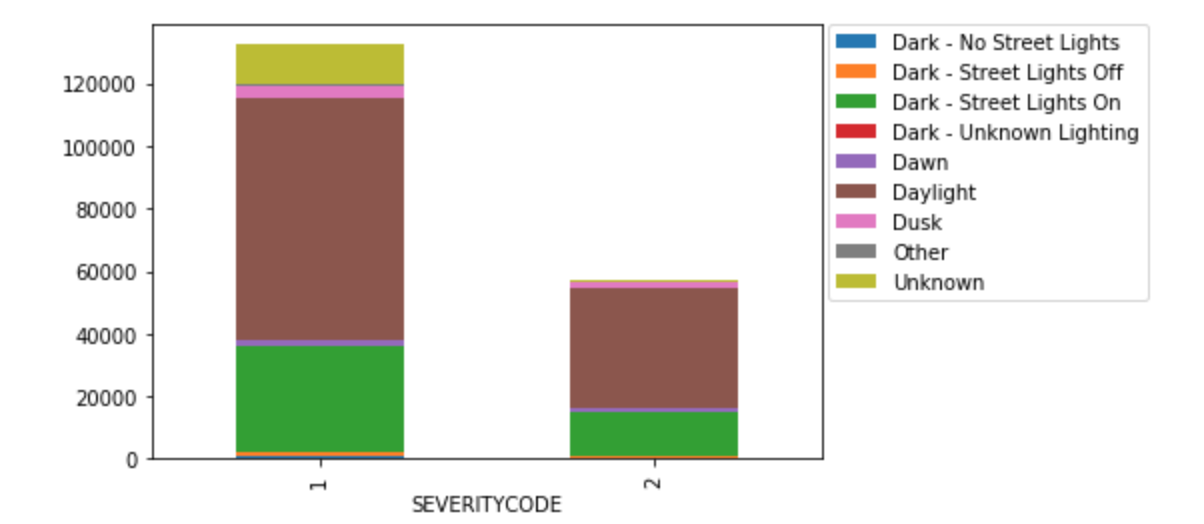
****

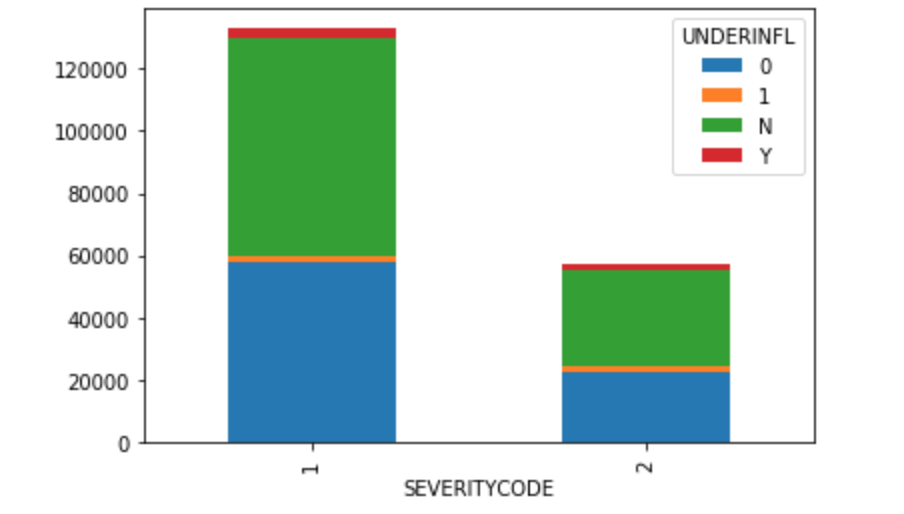
**Why**

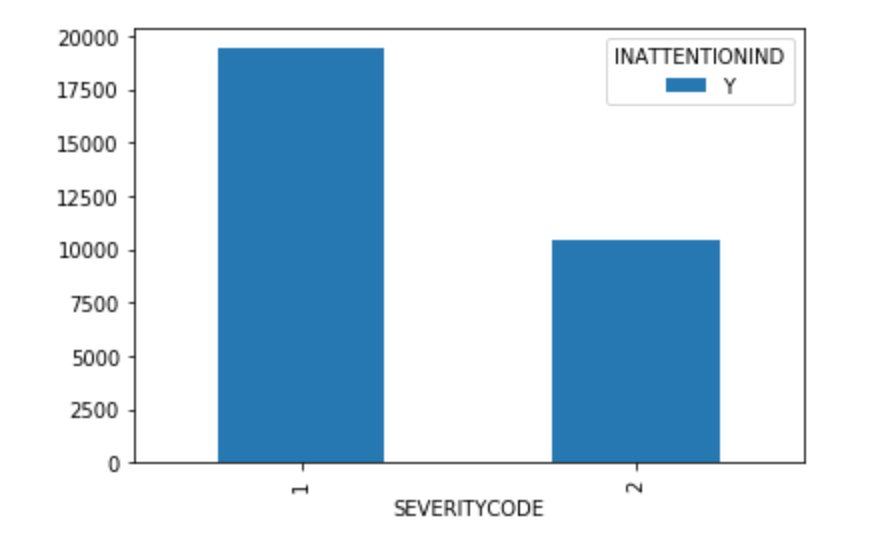
For the why I tried to explore more features than above because it is a more vague facet of the exploratory analysis. I looked at weather, road condition, lighting, inattention, speeding and under the influence. The graphs are below. From the weather graph we see that most collisions occur on clear days just because that’s most likely the most frequent sort of weather. Overcast and raining seem to hold the same proportion of accidents severe and non severe. For the road condition, we don’t see anything peculiar. Lighting conditions aren’t peculiar either and surprisingly neither is the under the influence chart. We see that speeding and inattention graphs are completely blue indicating that every collision was from inattention and speeding, but from our data cleaning we know that those columns are mostly filled with nulls so these two fields will be disregarded in our modeling.

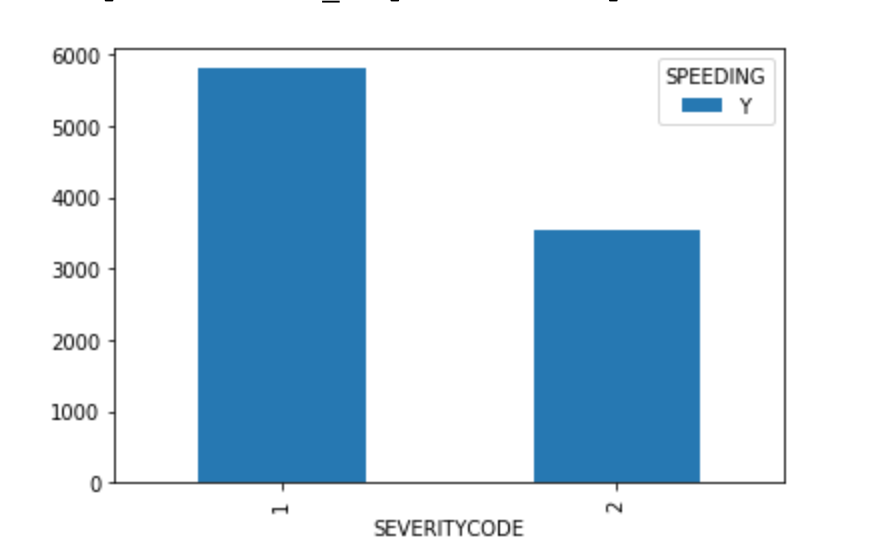






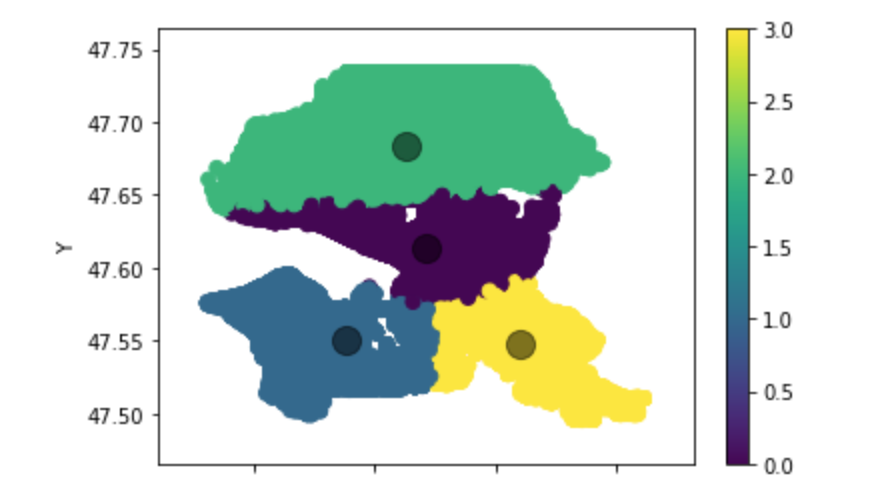


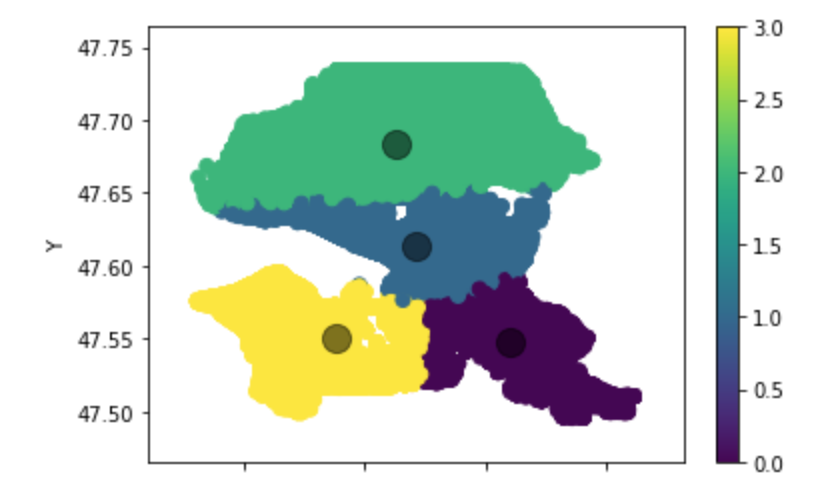
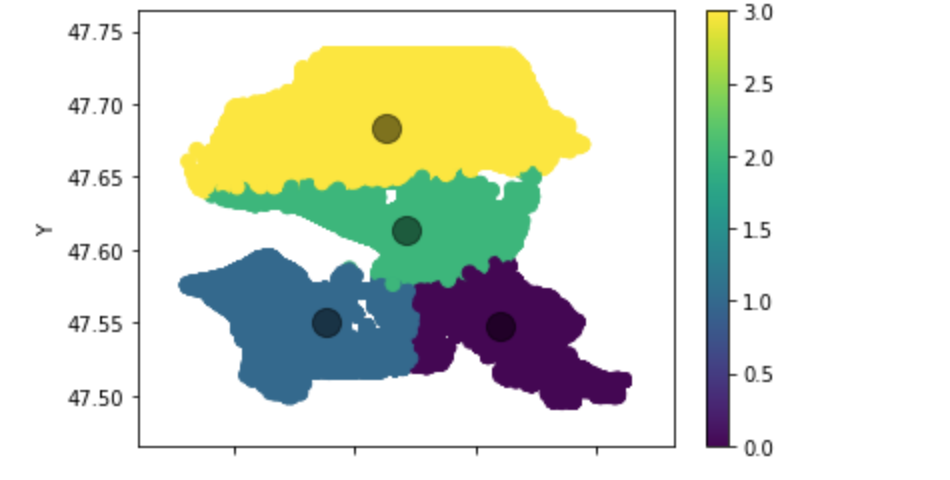




**Predictive Modeling**

**Clustering**

Clustering was the predictive modeling approach I took when tackling this problem. To do so I reinspected the data for cleansing. There were a few thousand rows without X or Y coordinates so I went and removed those before doing any clustering. From there I took three approaches with the data. Left it all as one, filtered for severe and filtered for non-severe. Using these three k means clustering graphs I was hoping to find some visual discrepancies between the three to give us some insight. I chose the same about of clusters for each plot because the elbow curve allowed for it and I wanted to stay consistent. In order below are the combined collisions, severe then non-severe graphs. Unfortunately, from the three graphs I can’t observe any extreme difference between the three. 



**Conclusions**

The conclusions from this project are very observational. We see clearly that intersections have proportionally more severe accidents than non-severe. Further, when any cyclists or pedestrians are involved it’s almost always severe. Another conclusion is that left turns are proportionally more severe than right turns which correlates with the intersection conclusion.

**Future Directions**

Next time I look at this dataset again I want to take a different approach than geo spatial clustering. I thought it was the correct approach because the prior lessons had shown me how effective it could be, but I think in this case logistic regression is needed. Considering our response was 1(non-severe) and 2(severe) we could treat them as a 0 and 1 respectively and find what factors play a significant role in deciding. I’d like to redo this project with logistic regression and then utilize more model fitting like goodness of fit, r^2 and least squares. I wish the dataset contained more severity codes like the data description sheet alluded to. If there were more time observations I’d be curious to see at what time of the day accidents occur. Next time I will use a correlation matrix to see if the variables correlate before I jump to conclusions about both.