Seattle Car/Bicycle Crash Analysis

Applied Data Science Capstone Project

Introduction

Car wrecks are a major problem, and there are several reasons for this. First off, safety is a concern. Car wrecks cost a lot of money too. Wrecks can also have effects on people that are not directly involved in the wreck itself. Traffic jams can waste time for a lot of people. If car wrecks can be predicted, they might be able to be prevented. Cyclist involved accidents are also a major concern; they are typically on the road with cars. There are large amounts of data involving car wrecks. Data Science can be used to predict when a wreck will occur. This information will be valuable to Departments of Transportation and all people that travel by road, not just drivers. A person who commutes by bus can appreciate the information. The models can help save their lives as well as help them to avoid traffic. Also, this information can be given to the Department of Transportation to help improve the actual roads and transportation system. If a specific location has a lot of vehicle/bicycle wrecks, the Department of Transportation might want to make changes in the area. They would be able to dive into the results from the models to determine the best way to prevent wrecks and injury.

Data Explanation

The collision data used is from Seattle, Washington. The data goes back to 2004 and goes to the present. There are many attributes in this data set: location, severity, collision type, number of cyclists/pedestrians, vehicles involved, injuries, fatalities, date, causes, weather, road conditions, and others. Using this data, models can be developed to help determine when a car wreck might occur. As the data is reviewed, it will determined which model will work best. Each incident in the data set, has the corresponding information filled out. Not all of the information is filled out every time, so the data will need to be processed to make sure the model turns out correctly. For example, one incident involving two vehicles occurred at an intersection. No pedestrians or cyclists were involved. The driver was not under the influence of drugs or alcohol. The road was wet, and the weather was listed as "Overcast". Given this sort of information for many car wrecks, a model can be developed to help predict a wreck given specific circumstances, weather, road conditions, time of day, etc.

Process Overview

- Import, review, and explore data
- Look for specific problem within data to explore in detail
- Will focus on cyclist involved crashes
- Map data
- Check the data for contributing factors
- Process data
- Develop machine learning models

Data Exploration

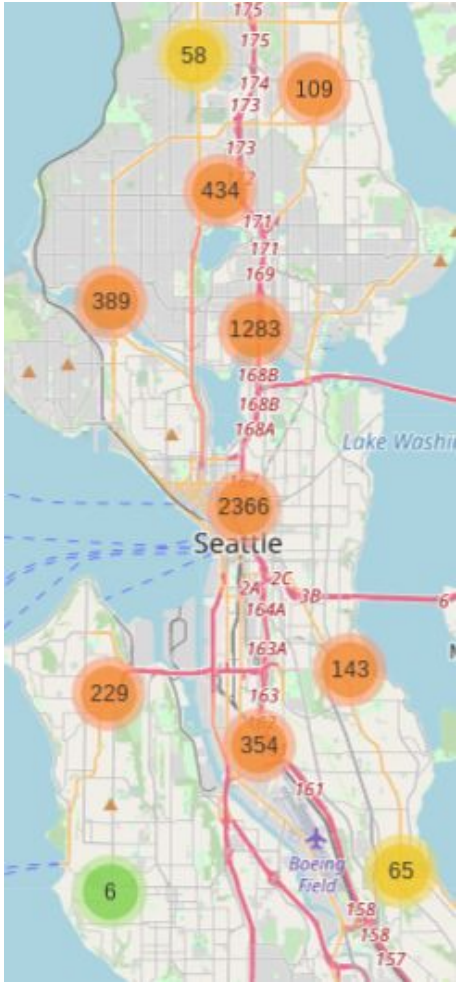
- Almost 200,000 incidents in data
- Look for specific problem within data to explore in detail
- There are no fatalities in the data set. This is surprising. The crashes with fatalities might have been removed for sensitivity purposes.
- Look into crash location
  - Block - 126,926
  - Intersection - 65,070
  - Alley - 751
- 5,484 incidents involve cyclists - will focus on these
- Map data

Data Processing

- Data types were mixed between integers, floats, and objects.
- Data types will be modified as necessary
- Many data points were lacking information
- Since there is so much data, it is best to simply remove these incidents rather than modify their contents.
- Standardized date format
- Sorted data by “Severity Code” among others
- The weather, road condition, and light condition information will need to be changed from words to numbers.

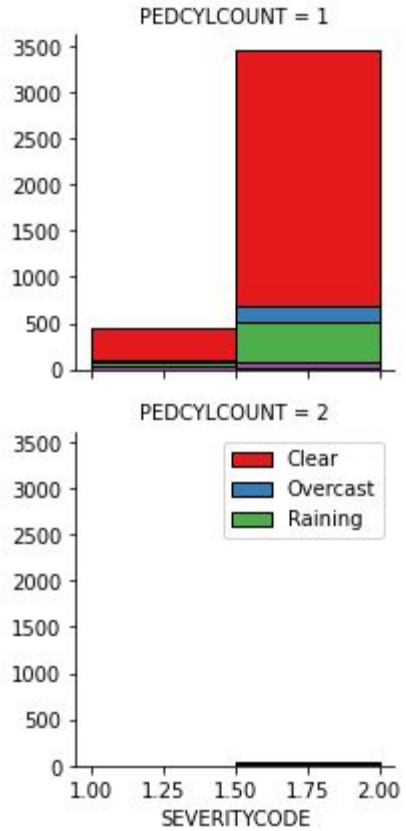
Map

- Interactive map on GitHub
- No surprise there are more cyclist involved accidents near the city center
- Accident numbers drop further from the city center



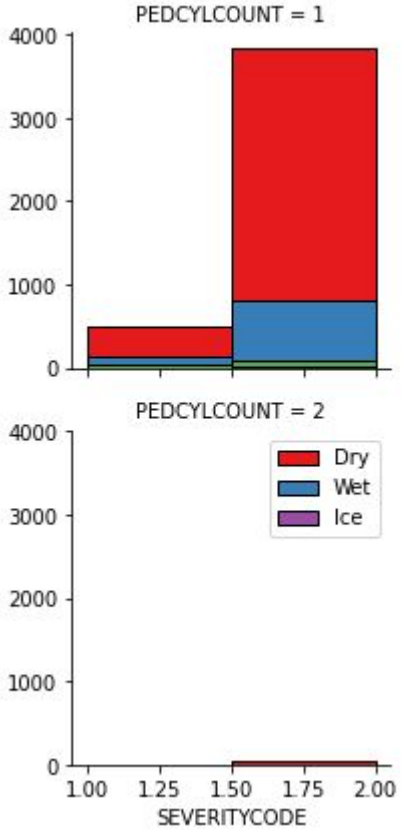
Weather

- The graphs to the right show the cyclist involved accidents broken out by the number of cyclists involved.
- Not many accidents involve multiple cyclists.
- The data is then divided into weather conditions.
- Weather does not seem to be a cause for accidents.
- By far, most incidents occur during clear weather.
- There could be fewer cyclists on the road during bad weather.



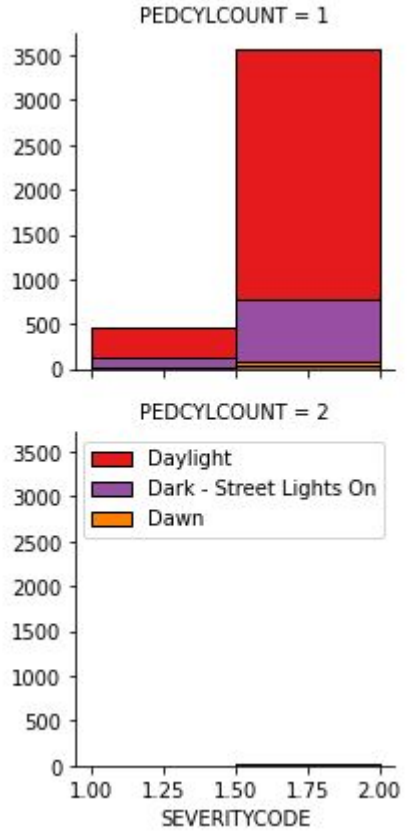
Road Conditions

- Road conditions do not appear to be a cause for cyclist related accidents.
- The majority of cyclist involved accidents occur with dry roads.
- Again, it is possible that not as many cyclists are on road during poor conditions.



Light Conditions

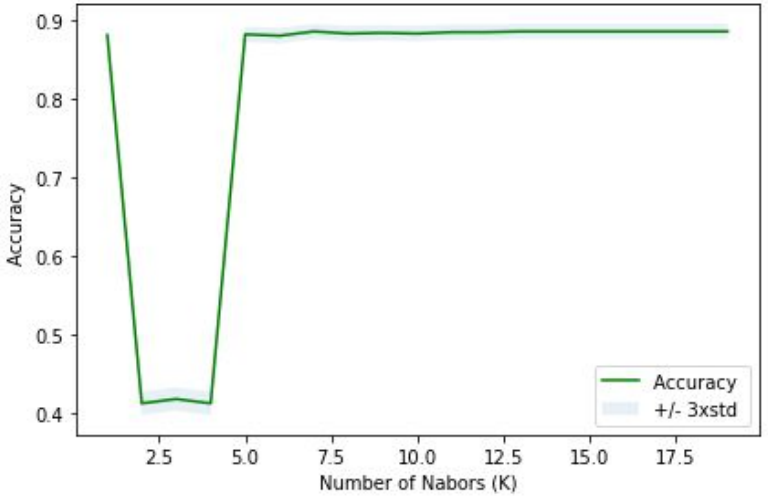
- Lighting does not appear to be a contributing factor for cyclist involved accidents.
- Most cyclist involved accidents occur during daylight hours.
- There are probably far fewer cyclists on the road when it is dark out.



Machine Learning Model

- K Nearest Neighbor was chosen because of a higher level of accuracy when compared to other models.
- K = 7
- Decision Tree, SVM, and Logistic Regression methods were tested but not used due to not being as good of a fit.

```
Train set Accuracy:  0.8739650413983441
Test set Accuracy:  0.8823529411764706
```




```
The best accuracy was with 0.8860294117647058 with k= 7
Avg F1-score: 0.8325
Jaccard score: 0.8860
```

Discussion - Observations

- Most of the cyclist involved accidents occur in the center of the city.
- This should not be too surprising because this is typically where more people commute by bicycle.
- There does not seem to be correlation between weather and cyclist involved accidents, so more research. This could be due to a sample bias. There are probably fewer cyclists on the road when the weather is bad and when it is dark.

Discussion - Recommendations

- Since cyclist involved accidents are not related to weather or lighting conditions, the Department of Transportation might want to spend additional resources on city wide bicycle awareness.
- This could be additional road signs, billboards, advertisements, etc.
- An education program for both cyclists and drivers could be beneficial.
- As the map suggests, these efforts might have the largest impact closer to the city center.

Conclusion

- Car/cyclist crashes are difficult for everyone involved.
- The map shows many more crashes near the city center.
- Weather and/or light are not contributing factors in most accidents.
- The K Nearest Neighbor model was used, k = 7.
- Increasing awareness for both drivers and cyclists would be beneficial.
- The Department of Transportation could use this data to help prevent car/cyclist crashes.
- Prevention would not only save people from injury, but it would also save time due to lowering traffic jams.