Title: Motor Vehicle Collisions - Crashes in New York City

Name: Anushka Chandrashekar

DUIs and traffic signs exist to reduce the rate of drivers driving while under the influence of alcohol or other drugs. The state of New York only permits hands-free cell phone use while driving (cite). However, multiple other factors significantly contribute to motor vehicle collisions in NYC, as we will explore. My goals in this project are to inspire greater awareness of these contributing factors, predict the number of people injured based on how many minutes into the day it had been on the day of the crash, and predict how important specific features are in their contribution to injuries. These predictions have implications for the prevention of motor vehicle collision injuries in NYC.

The City of New York's Open Data website contains Motor Vehicle Collisions data tables. These tables have information from all police-reported motor vehicle collisions in NYC.

We can find this at https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95. The original dataset has 250,000 rows and 34 columns. We choose to limit the dataset to 20,000 rows. Each column describes a specific facet of the nature of the crash. 5 columns list the contributing factor (i.e., A driver is distracted) of the crash for each vehicle involved in the incident. There is a maximum number of 5 vehicles in the crash. We create a column for the number of contributing factors for each crash. We use this column in our features that we train to predict the importance of features in their contribution to injuries. Next, we create a column that displays a 1 if there were injuries and 0 if there were no injuries. We use this column to predict whether anyone was injured in the collision. We create another column that displays a 1 if there were deaths and 0 if there were no deaths.

We start by exploring the 10 most common contributing factors to collisions. We drop an outlier of 80 from the column that lists the contributing factors of vehicle 1. There are two plots in Figure 1. In the first plot, each bar represents the number of crashes that were caused by a particular contributing factor. We see that the 2nd most common contributing factor, the driver being distracted or inattentive while driving, causes about 7,000 crashes. The most minor, common contributing factor, drivers disregarding traffic rules, causes only about 1,000 crashes. In the second plot, each bar represents the top ten factors most likely to cause an injury. It shows that using headphones while driving has about a 90% chance of causing an injury.

Next, we want to predict the number of people injured after a certain number of minutes of the day has passed. Figure 2 is a scatterplot that contains 6 clusters. Each cluster represents the number of people injured a certain number of minutes into the day. We can see that collisions occur in clusters by the time of day due to the frequency of accidents during rush hour and other periods of the day. For example, we can see that the largest cluster occurs between 0-220 minutes of the day. The column containing the time of the crash is used to calculate the number of minutes. Next, we generate a data frame with 9 features that we will use to make our predictions. We fill missing values of columns of the data frame with 0, indicating that an empty value was left out because it was 0. We create our model by first using a StandardScaler to scale the data, so all attributes are of equal weight. Then, we cluster numbers 1 to 14 by twos and determine the number of clusters to use. We find that 7-9 clusters are reasonable. Then we use KMeans to get each inertia, which is the sum of squares to the cluster center.

Lastly, we predict the importance of 10 features in their contribution to injuries. Figure 3 is a bar plot where each bar represents the importance of a feature on a scale of 0-0.35. We examine the contributing factors for vehicle 1, the borough in NYC (a borough is a smaller city within a large metropolis), and the number of contributing factors. We fill missing values with 0 and conduct a train-test split on our data. Then, we use a RandomForestClassifier to predict which feature has the highest importance. Our model has a score of 71.6 %. The number of contributing factors has the highest importance (0.33), which shows that the most significant indicator of an injury is the number of contributing factors for vehicle 1.

To conclude, our analysis shows that the most common contributing factors to motor vehicle collisions vary slightly with the most likely factors to cause an injury. Also, injuries are more likely to be very common early into the day, and the risk for injuries depends mainly on the number of contributing factors in the collision. *Special thanks to my friend Christopher for helping me with some of the difficult parts of this project.*

Figures:

Figure 1: Top 10 Contributing Factors in Collissions

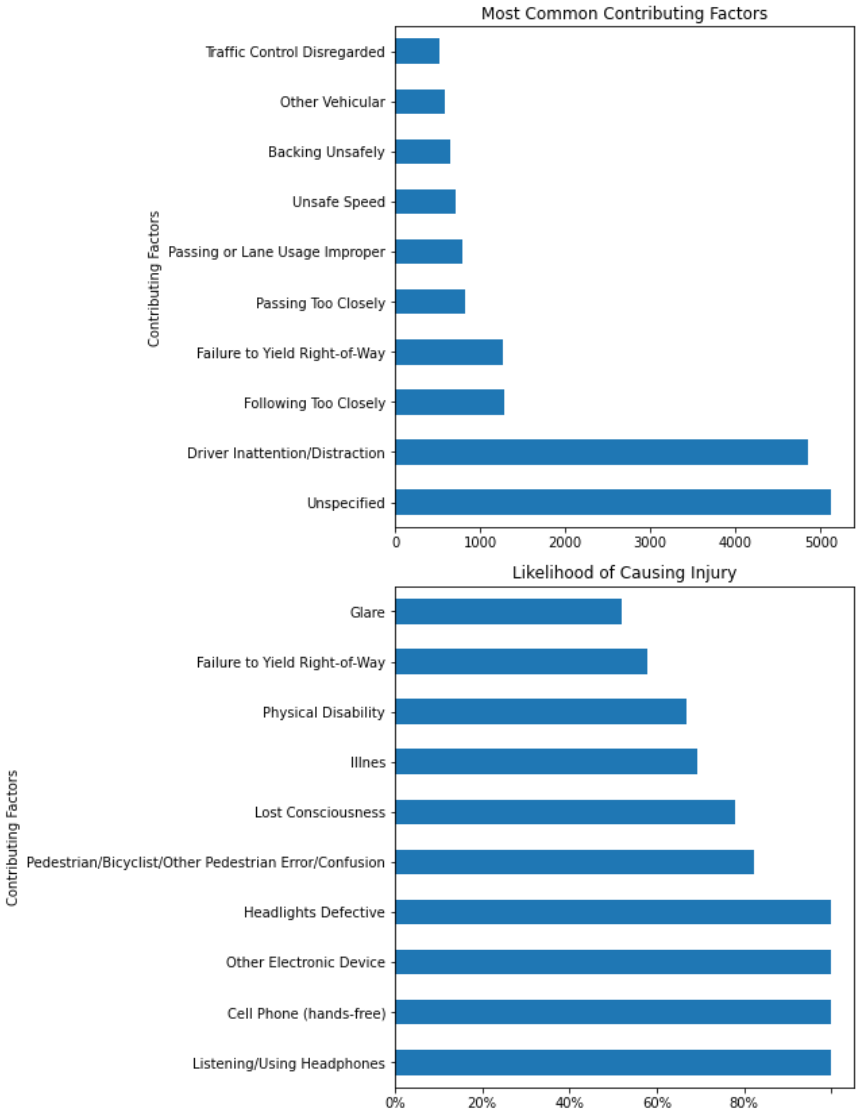

Figure 2: Number of people injured vs. Time passed

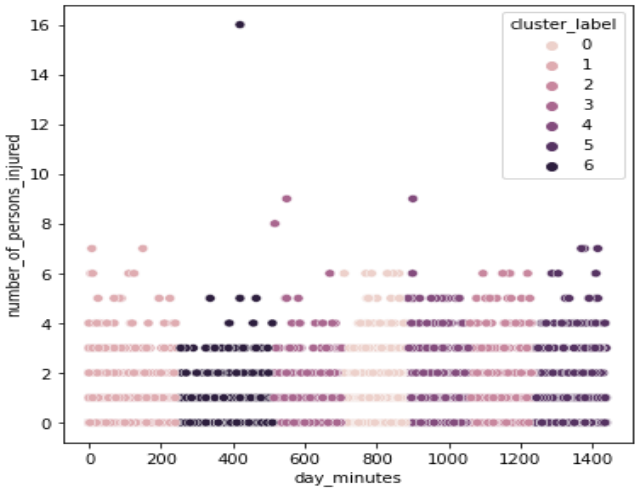

Figure 3: Predicting Injuries - Feature Importance