Group 3: Crash Analysis NYC- Draft 3

Kevin Lei, Micah Kepe, Zachary Kepe, Giulia Costantini

2024-02-27

Table of Contents

Introduction . . . 1
Data Description . . . 1
Data Cleaning and Pre-Processing . . . 2
Past Trends and Insights . . . 2
Frequency of Car Crashes by NYC Borough (2012-2024) . . . 2
Density of Car Crashes in NYC . . . 3
Histogram of Car Crashes by Time of Day . . . 4
Line Plot of Top 3 Contributing Factors by Time of Day . . . 4
Temporal Density Map by Both Day of Week and Hour of Day . . . 6
Pie Chart of Contributing Factors to Car Crashes . . . 7
Stack Area Plot of Contributing Factors and Crash Count by Year . . . 8
Jitter Plot of Fatalities and Injuries by Contributing Factor . . . 8
Predictive Modeling . . . 9
Decision Tree . . . 9
Multiple Linear Regression . . . 9
Conclusion . . . 9

Introduction

The NYC OpenData dataset contains details of motor vehicle collision occurrences reported by the NYPD from 2012 to 2024. Collisions are only reported if a person was injured or killed, or if there was more than $1000 in damage.

Car crashes in the United States represent a significant public health concern, with over 42,900 deaths involving over 61,000 vehicles in 2020 according to the Insurance Institute for Highway Safety. The economic impact of car crashes is also substantial, with an estimated cost of $340 billion in 2019 via the National Highway Traffic Safety Administration. Not only does preventing car crashes save lives, but it also saves money and resources.

Our project seeks to determine the most important factors that contribute to potentially fatal accidents. By analyzing the data, we hope to identify to provide recommendations for reducing the number of accidents.

Data Description

The data set contains over 2 million records, each detailing a motor vehicle collision. The data includes the date, time, location (latitude, longitude, borough), and the number of individuals affected (injured or killed) among drivers, pedestrians, and cyclists. The data set also includes information about the vehicles involved, such as the type of vehicle and the contributing factors to the crash. Our analysis will focus on

understanding patterns and trends within these incidents, and identifying the most important factors that contribute to potentially fatal accidents.

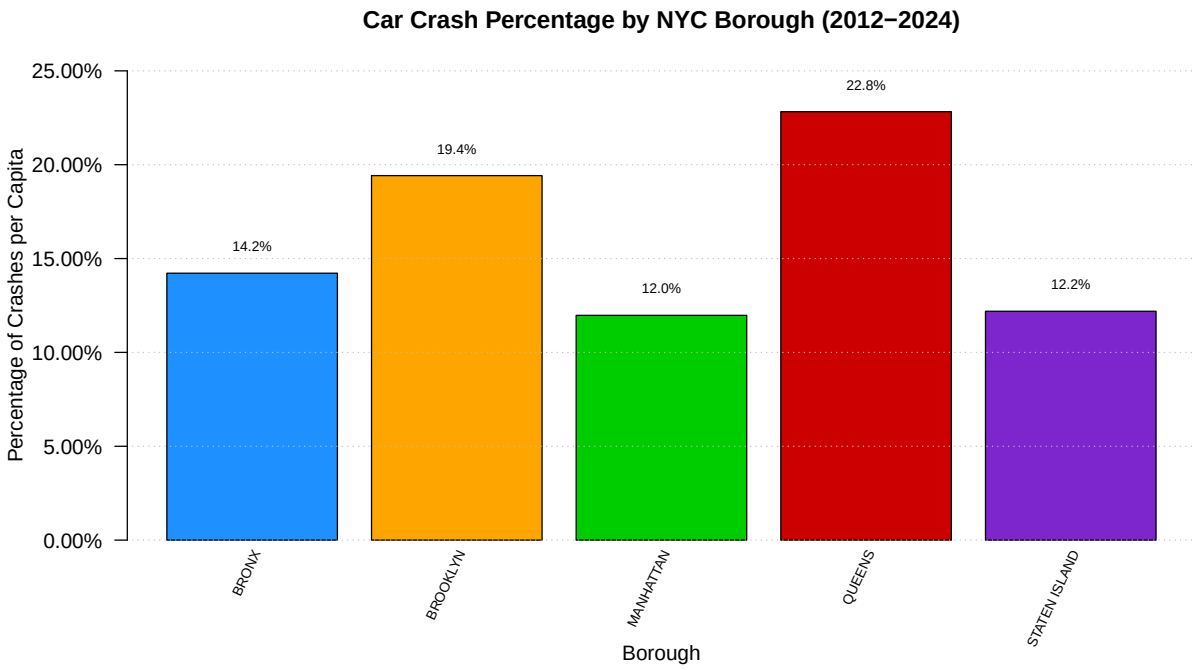
Data Cleaning and Pre-Processing

For the data cleaning process, we removed rows with missing or invalid values in crucial columns, such as BOROUGH, LATITUDE, LONGITUDE, CRASH.DATE, and CRASH.TIME. We also converted the CRASH.DATE column to a Date format and the CRASH.TIME column to a more standard format. We replaced empty strings with "UNKNOWN" for street names and dropped the LOCATION column as it was redundant. Finally, we removed any remaining rows with NA values in any column.

As a result of the data cleaning process, the data set went from having 2,065,192 rows and 29 columns to having 1,382,320 rows and 28 columns. While this is a significant reduction in the number of rows, the data set is still large enough to perform meaningful analysis.

Past Trends and Insights

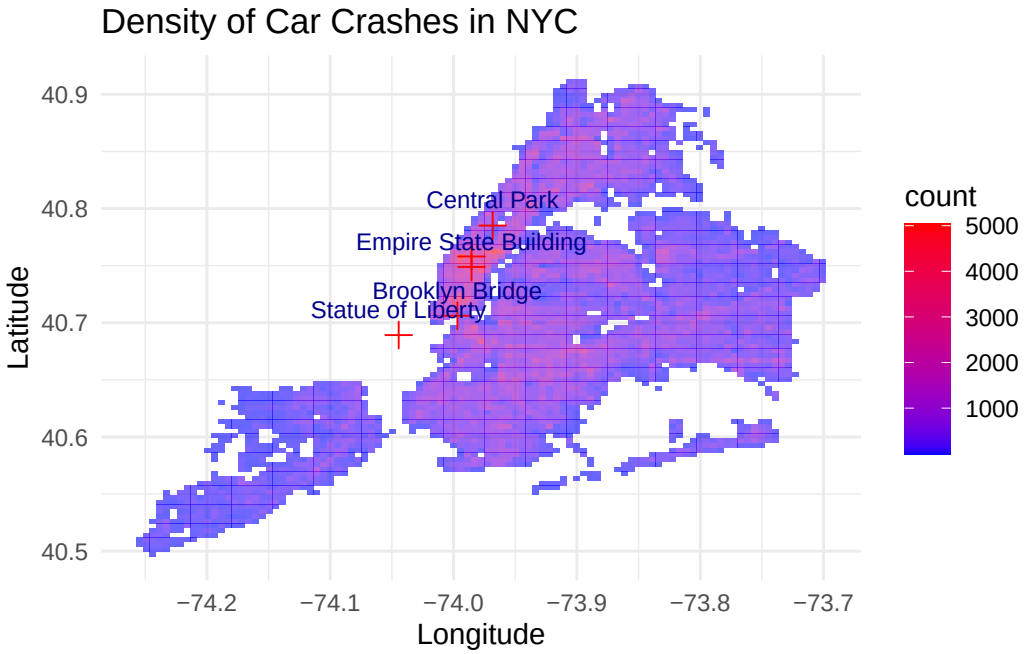
Frequency of Car Crashes by NYC Borough (2012-2024)



We plotted the frequency of car crashes per borough in NYC. The goal of this graph was to gain a better understanding of which boroughs were more likely to result in car crashes.

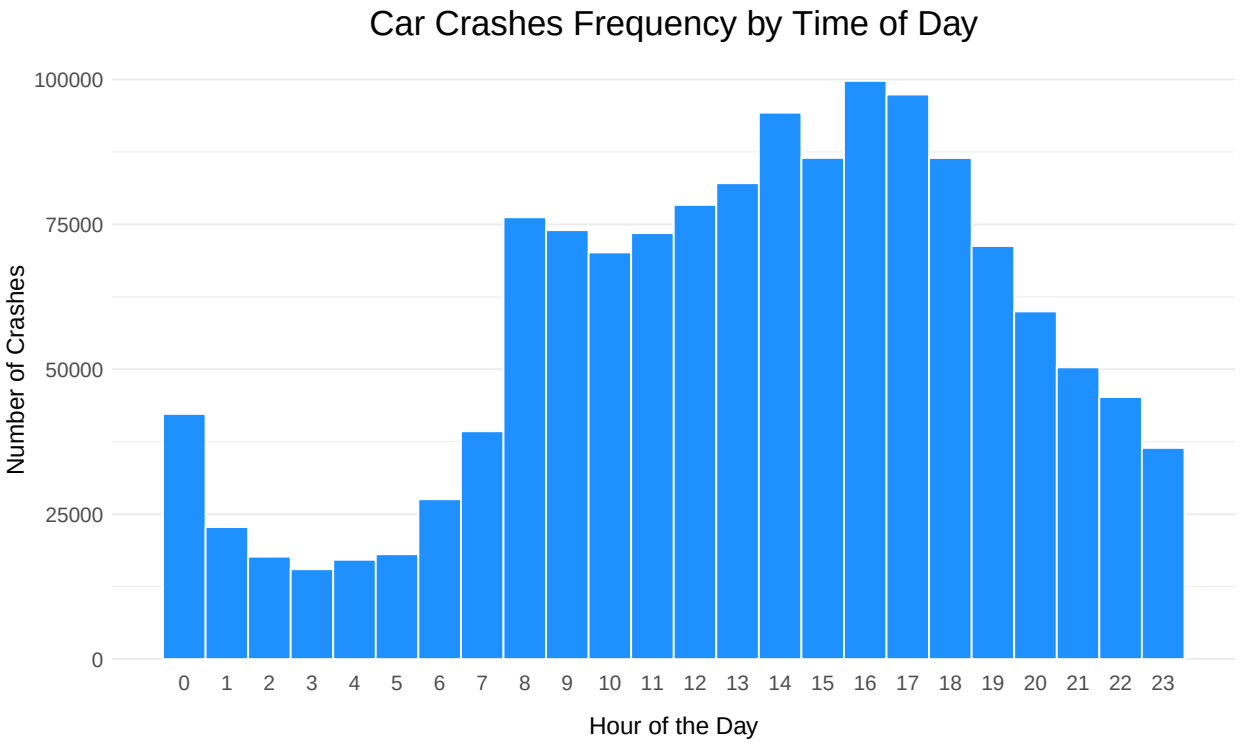
The borough with the least car crashes is Staten Island at around 75,000 car crashes, while the borough with the most car crashes is Brooklyn with around 450,000 car crashes. This graph gives us a better understanding of the likelihood that a car crash will occur in a certain borough. However, a consideration that is not specified is the population of each borough and the amount of traffic through them.

Density of Car Crashes in NYC



The heat map of car crashes in NYC shows that the highest density of car crashes occurs in the center of NYC. This is likely due to the fact that the center of NYC has the most traffic and the most people. The lowest density of car crashes occurs in the outskirts of NYC. This aligns with intuition, as the centers of cities are typically more crowded and have more traffic. This graph gives us a better understanding of where car crashes are most likely to occur in NYC.

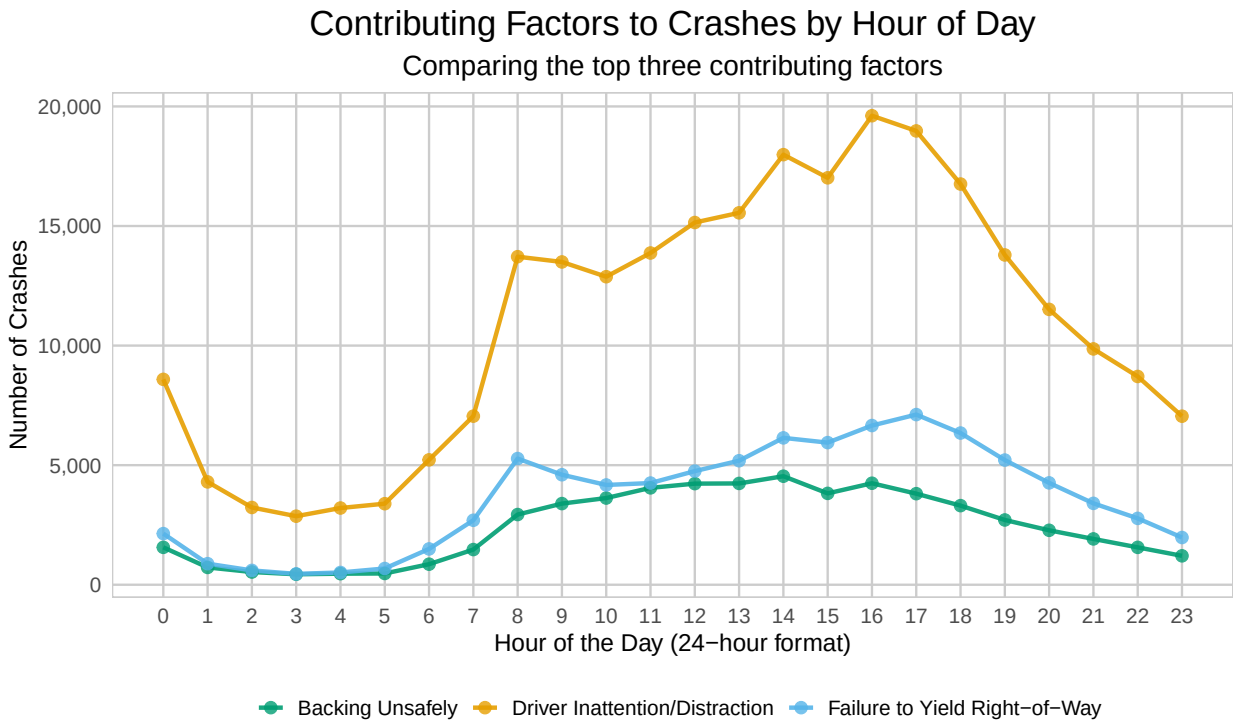
Histogram of Car Crashes by Time of Day



The histogram of car crashes by time of day shows that the most car crashes occur around the hours of 15:00 and 18:00. This is likely due to the fact that these are the hours when people are getting off work and are driving home. The least amount of car crashes occur around the hours of 3:00 and 4:00, which is likely due to the fact that these are the hours when people are sleeping and there is less traffic on the road. This graph gives us a better understanding of when car crashes are most likely to occur. An important consideration that is that traffic patterns may vary seasonally and by day of the week, which could affect the number of car crashes.

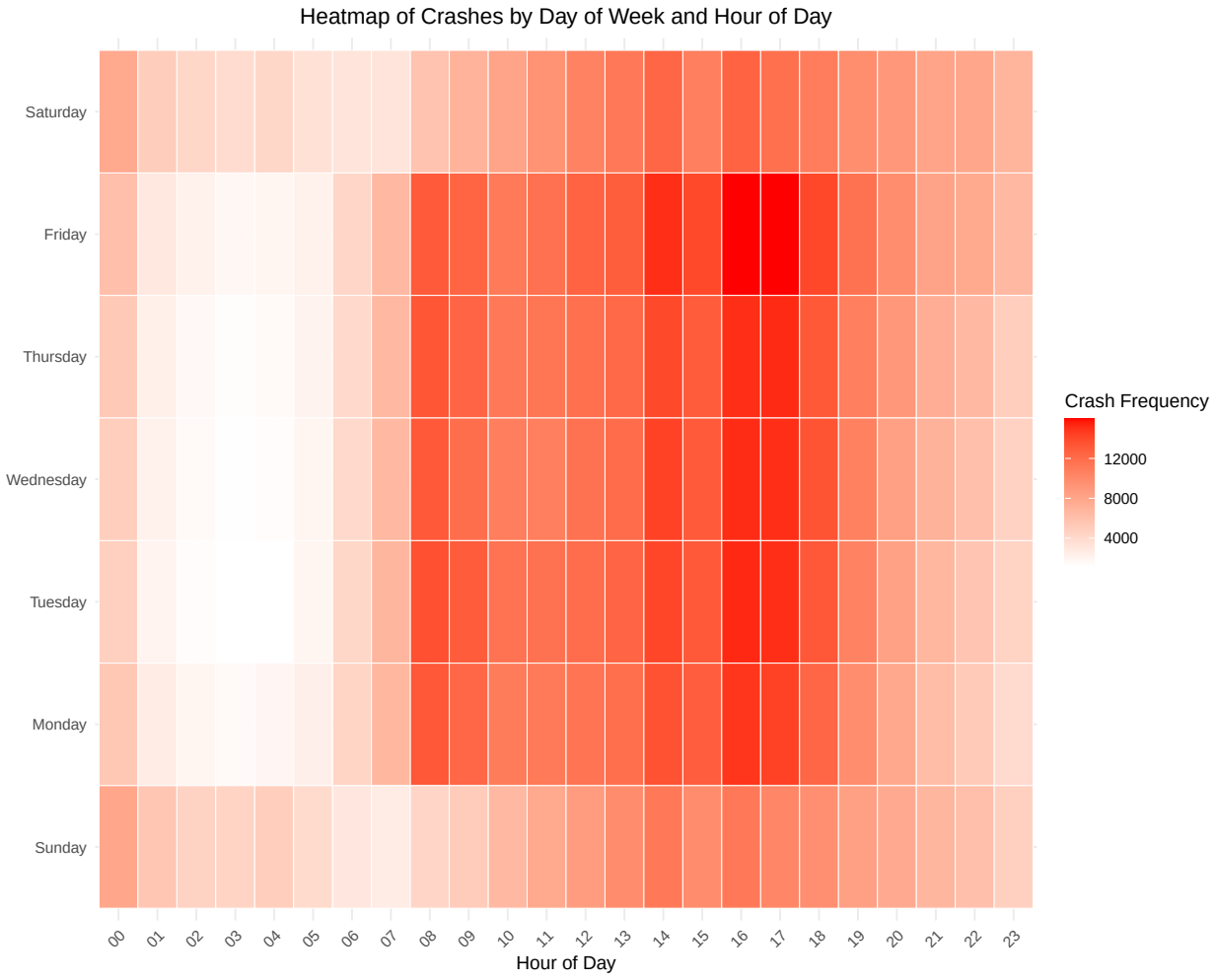
Line Plot of Top 3 Contributing Factors by Time of Day

```
Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use ‘linewidth‘ instead.
```



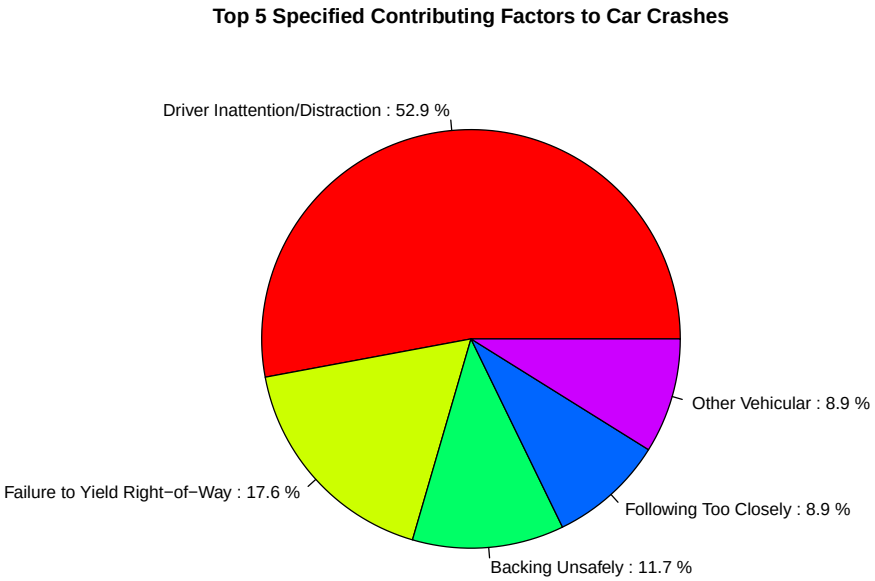
The scatter plot of the top 3 contributing factors by time of day shows the times that result in the most crashes in accordance to their contributing factor. What can be seen is that the hours to and from work result in the most crashes, such as 8 am and 5 pm. All contributing factors show that the crash amount increases between the hours of 8 am and 5 pm and decreases outside of those hours. This graph gives us a better understanding of the relationship between time of day, number of crashes, and contributing factors.

Temporal Density Map by Both Day of Week and Hour of Day



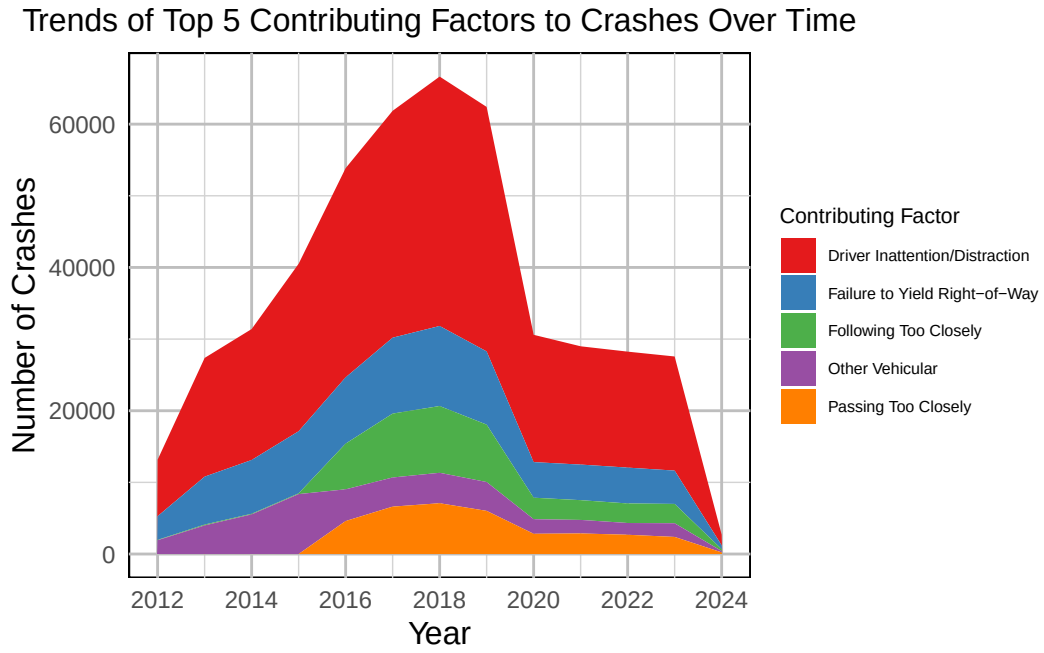
The temporal density map shows the frequency of car crashes by day of the week and hour of the day. There is an evident "safe zone" on weekdays between the hours of 1:00AM and 5:00AM. This is expected as these are not usually times where there are lots of people active. Interesting, "danger zones" are evident on weekdays between the hours of 8:00AM and 10:00AM and 4:00PM and 6:00PM. On possible explanation for this is that these are the hours when people are driving to and from work. This graph dispels the myth that most car crashes occur at night and shows that most car crashes actually occur during the day. This insight is crucial for eliminating preconceived notions about car crashes and understanding the true patterns of car crashes.

Pie Chart of Contributing Factors to Car Crashes



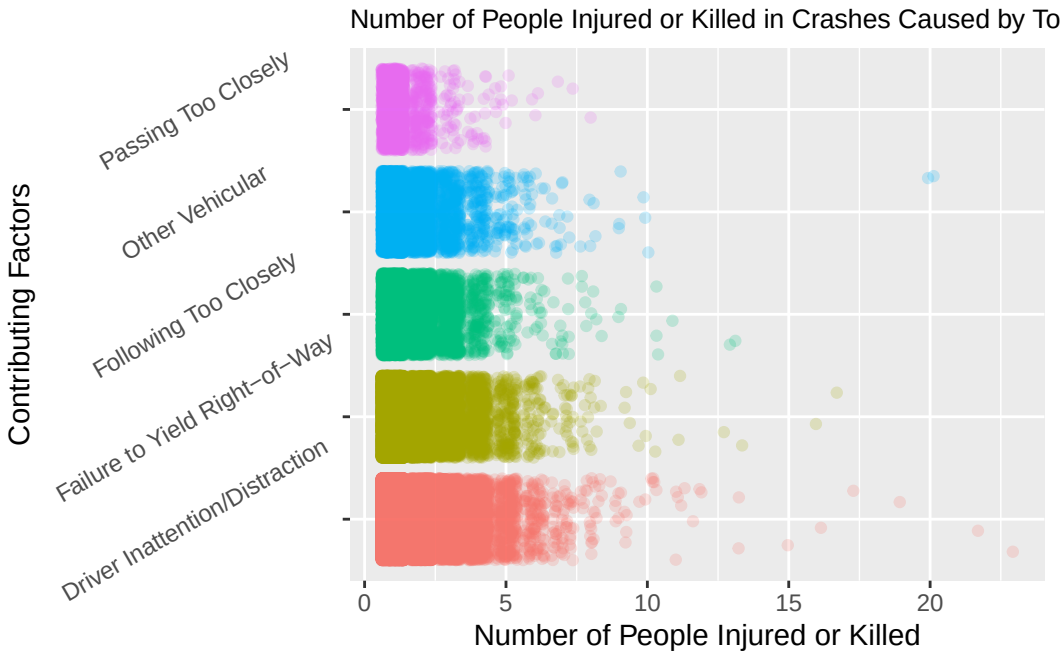
Disregarding the unspecified contributing factors, the top 5 contributing factors to car crashes are: Driver Inattention/Distraction, Failure to Yield Right-of-Way, Backing Unsafely, Following Too Closely, and Other Vehicular. The pie chart shows that the top specified contributing factor to car crashes is Driver Inattention/Distraction, which accounts for 52.9% of the observed car crashes. However, an important consideration is that a large amount of the contributing factors are unspecified, which could affect the accuracy of the data.

Stack Area Plot of Contributing Factors and Crash Count by Year



The stacked area plot of the top 5 contributing factors to car crashes by year shows how the top contributing factors changed over the years. What can be seen is that the number of crashes peaked in 2018 at around 67500 crashes and that driver inattention/distraction grew rapidly from 2012 to 2018. Then, all contributing factors started to decrease 2020 to 2024 most likely due to covid and stay at home orders. This graph gives us a better understanding of the relationship between the year, number of crashes, and contributing factors.

Jitter Plot of Fatalities and Injuries by Contributing Factor



The jitter plot shows the number of people injured or killed in crashes caused by the top 5 factors. The plot shows that the most common contributing factors to crashes are also correlated with more violent crashes. Driver inattention or distraction caused the most violent crashes. This plot gives more insight into how different contributing factors of crashes relate to injuries and deaths.

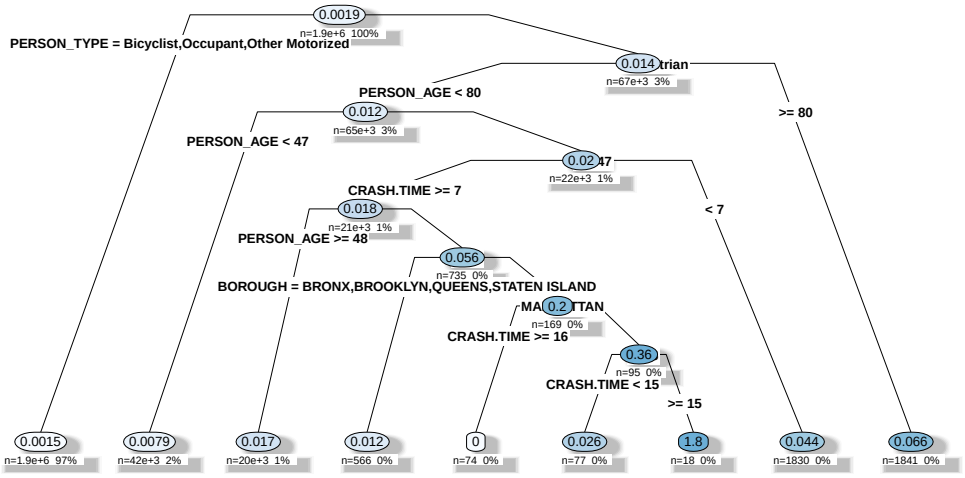
Predictive Modeling

Decision Tree

```
[1] "Mean Absolute Error (MAE): 0.00374560069164159"
```

```
[1] "Root Mean Squared Error (RMSE): 0.052475580656655"
```

Decision Tree Model for Predicting Fatalities in Car Crashes



The decision tree model was trained to predict the number of people killed in car crashes based on the following features: BOROUGH, PERSON_TYPE, CRASH.TIME, and PERSON_AGE. The model was trained on 70% of the data and tested on the remaining 30%. The model achieved a Mean Absolute Error (MAE) of 0.00374560069164159 and a Root Mean Squared Error (RMSE) of 0.052475580656655. The decision tree model is visualized above. The model is quite complex and difficult to interpret, but it is able to predict the number of people killed in car crashes with a high degree of accuracy. Most likely the model is overfitting the data and would need to be adjusted to be more generalizable. However, we will most likely switch to a different model for our final analysis.

Multiple Linear Regression

Conclusion

In our further research since our previous report draft, we have gathered more insights into our data. Of note is the insight that most accidents actually occur within weekdays and during the day, not at night. We have also gathered insights into the contributing factors of car crashes and how they relate to injuries and fatalities. With the addition of NYC OpenData's secondary dataset of people involved in car crashes, we

hope to uncover even more factors, trends, discoveries, etc. that will better guide our potential suggestions for the most meaningful and impactful changes to NYC's traffic safety.