

# **Group 3: Crash Analysis NYC- Draft 3**

Kevin Lei, Micah Kepe, Zachary Kepe, Giulia Costantini

2024-02-27

## **Table of Contents**

|   |    |
|---|----|
| Introduction . . . . .  | 2  |
| Data Description . . . . .  | 2  |
| Data Cleaning and Pre-Processing . . . . .                                | 2  |
| Past Trends and Insights . . . . .  | 3  |
| Frequency of Car Crashes by NYC Borough (2012-2024) . . . . .             | 3  |
| Density of Car Crashes in NYC . . . . .                                   | 4  |
| Histogram of Car Crashes by Time of Day . . . . .                         | 5  |
| Line Plot of Top 3 Contributing Factors by Time of Day . . . . .          | 6  |
| Temporal Density Map by Both Day of Week and Hour of Day . . . . .        | 7  |
| Pie Chart of Contributing Factors to Car Crashes . . . . .                | 8  |
| Stack Area Plot of Contributing Factors and Crash Count by Year . . . . . | 9  |
| Jitter Plot of Fatalities and Injuries by Contributing Factor . . . . .   | 9  |
| Key Findings . . . . .  | 10 |
| Predictive Modeling . . . . .   | 11 |
| Decision Tree . . . . .   | 11 |
| Statistical Modeling with Generalized Linear Model . . . . .              | 12 |
| Conclusion . . . . .  | 14 |

## Introduction

In the bustling metropolis of New York City, the dynamics of motor vehicle collisions present a complex puzzle of public health, urban planning, and behavioral science.

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 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. The data set of crash incidents contains over 2 million records alone, 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. In addition to the crash dataset, NYC OpenData maintains two other datasets related to the vehicle(s) and people involved in each crash, respectively. When combined, these datasets provide a comprehensive view of each collision incident and the individuals involved. 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

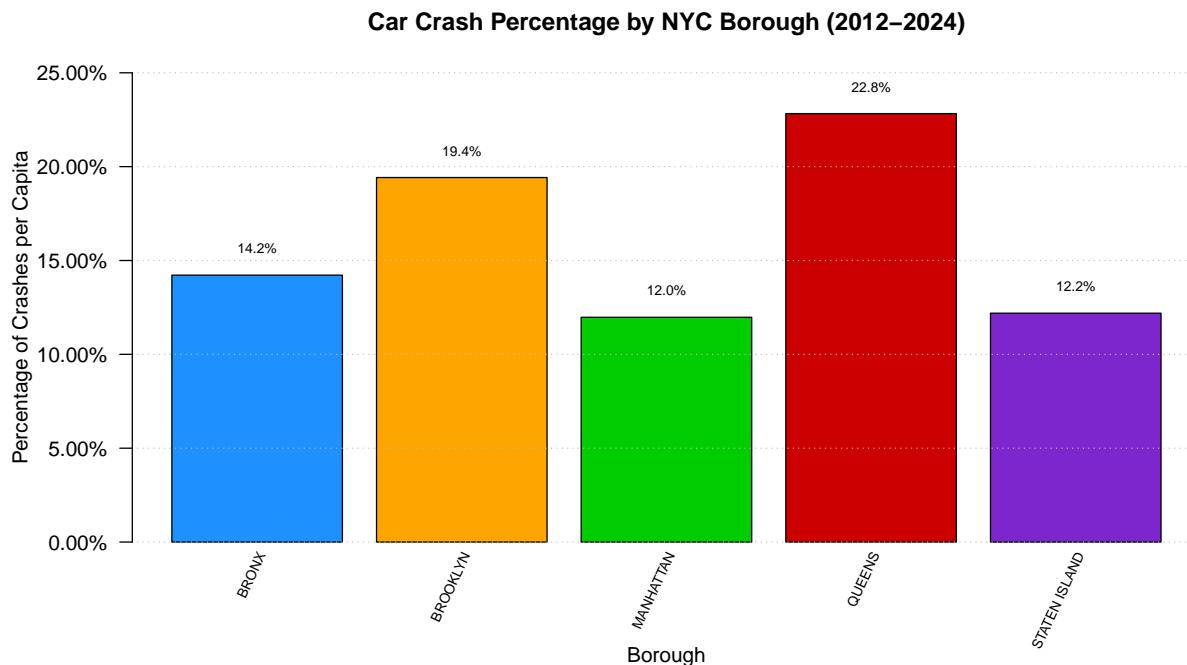
For the data cleaning process, we removed rows with missing or invalid values in crucial columns, such as BOROUGH, LATITUDE, LONGITUDE, CRASH.DATE, and CRASH.TIME. We also converted the CRASH.DATE column to a Date format and the CRASH.TIME column to a more standard format. We replaced empty strings with “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 primary 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

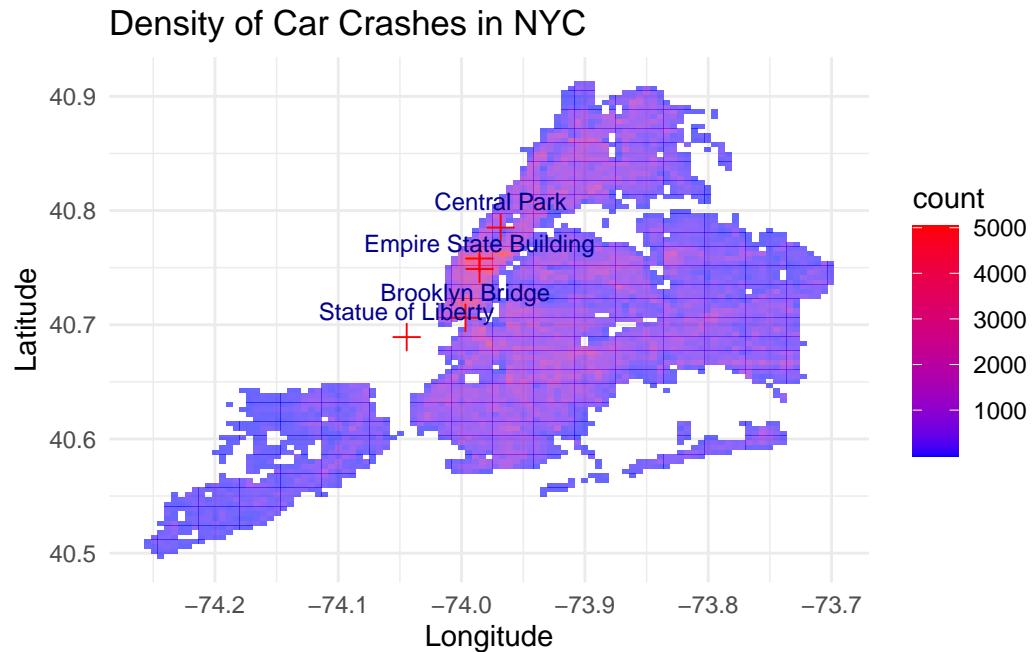
After cleaning the data, we performed an exploratory data analysis to identify trends and insights in the data and glean a better understanding of the factors that contribute to car crashes. Understanding the past trends of the data is crucial for informing our intuition leading into our predictive modeling and the recommendations we plan to make. Our tools were graphs and charts, each telling a different story about the data.

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

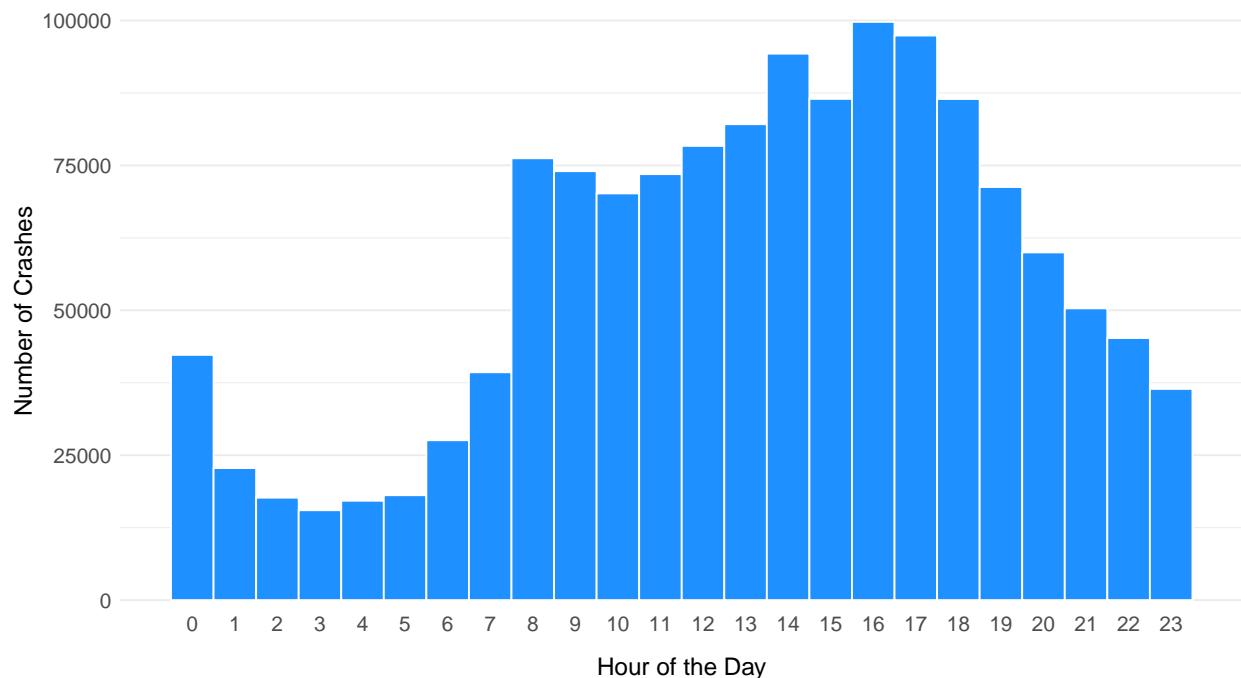


We plotted the frequency of car crashes per borough in NYC. The goal of this graph was to gain a better understanding of which boroughs were more likely to result in car crashes.

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.

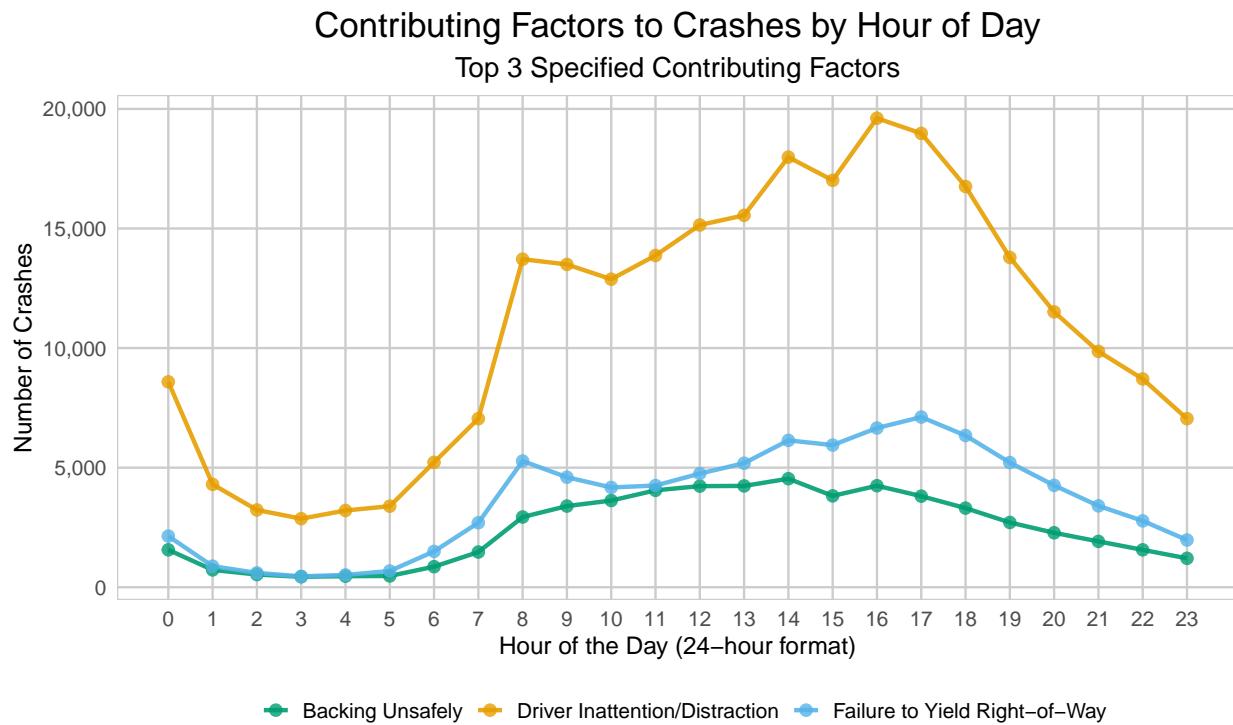
**Density of Car Crashes in NYC**

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****Car Crashes Frequency by Time of Day**

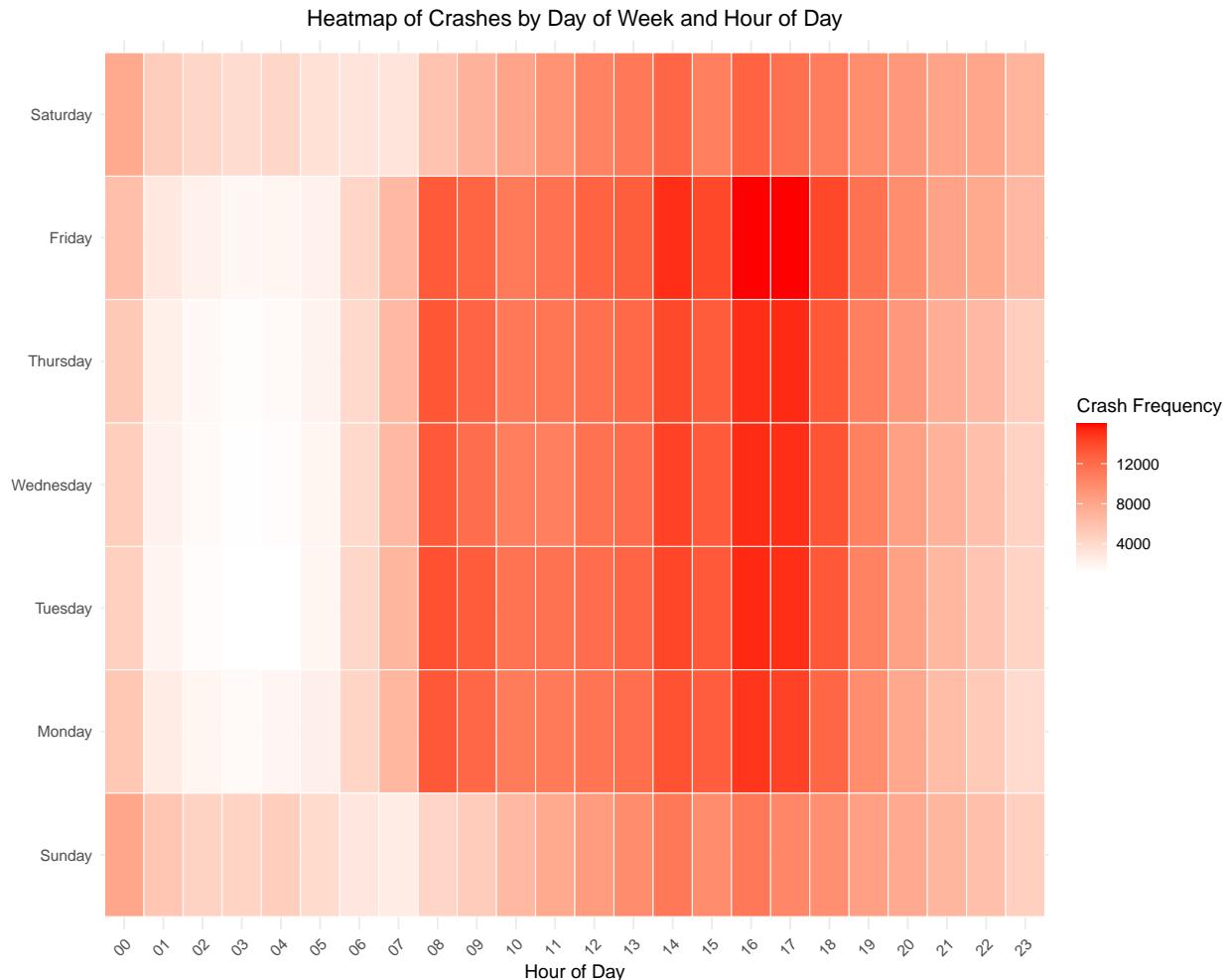
The histogram of car crashes by time of day shows that the most car crashes occur around the hours of 15:00 and 18:00. This is likely due to the fact that these are the hours when people are getting off work and are driving home. The least amount of car crashes occur around the hours of 3:00 and 4:00, 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 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

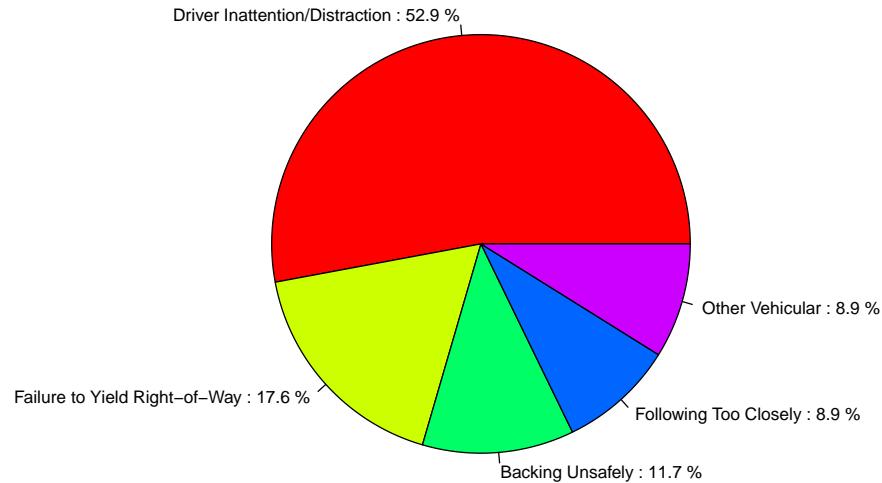


The scatter plot of the top 3 contributing factors by time of day shows the times that result in the most crashes in accordance to their contributing factor. What can be seen is that the hours to and from work result in the most crashes, such as 8 am and 5 pm. All contributing factors show that the crash 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



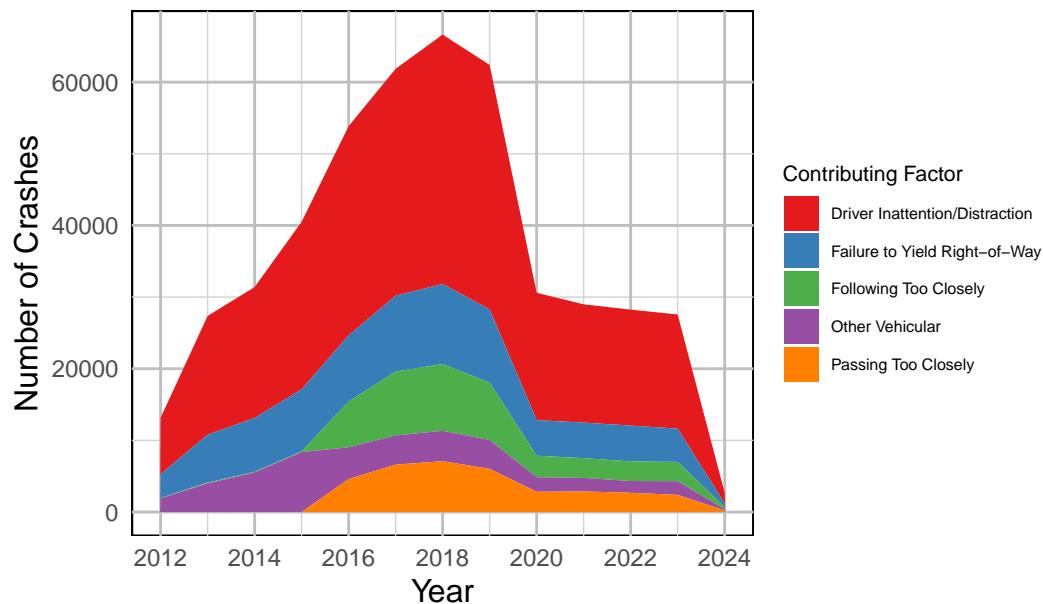
The temporal density map shows the frequency of car crashes by day of the week and hour of the day. There is an evident “safe zone” on weekdays between the hours of 1:00AM and 5:00AM. This is expected as these are not usually times where there are lots of people active. Interesting, “danger zones” are evident on weekdays between the hours of 8:00AM and 10:00AM and 4:00PM and 6:00PM. One 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****Top 5 Specified 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.

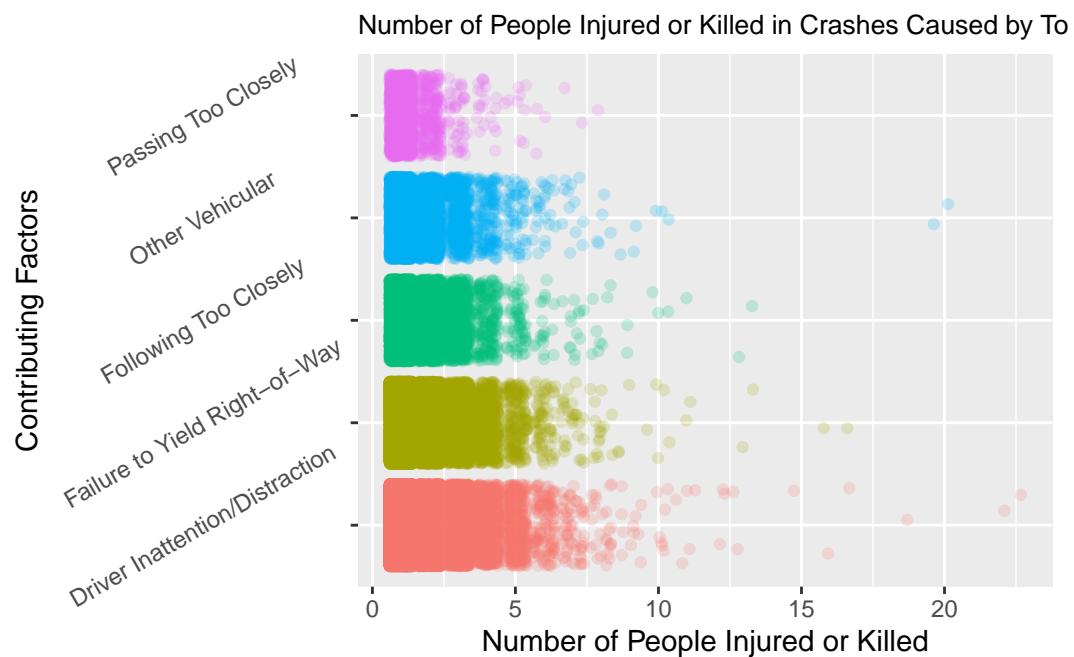
### Stack Area Plot of Contributing Factors and Crash Count by Year

#### Trends of Top 5 Contributing Factors to Crashes Over Time



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.

### **Key Findings**

**Borough-wise Crash Frequency:** We found that some boroughs experience more crashes than others, with Brooklyn leading the chart. This could relate to population density, traffic flow, or other urban factors.

**Crash Density and Landmarks:** A map showing where crashes happen most frequently revealed that central areas with heavy traffic and significant landmarks are hotspots for incidents.

**Time of Day Analysis:** Crashes peak during evening rush hours, suggesting that the end-of-day commute might be particularly hazardous.

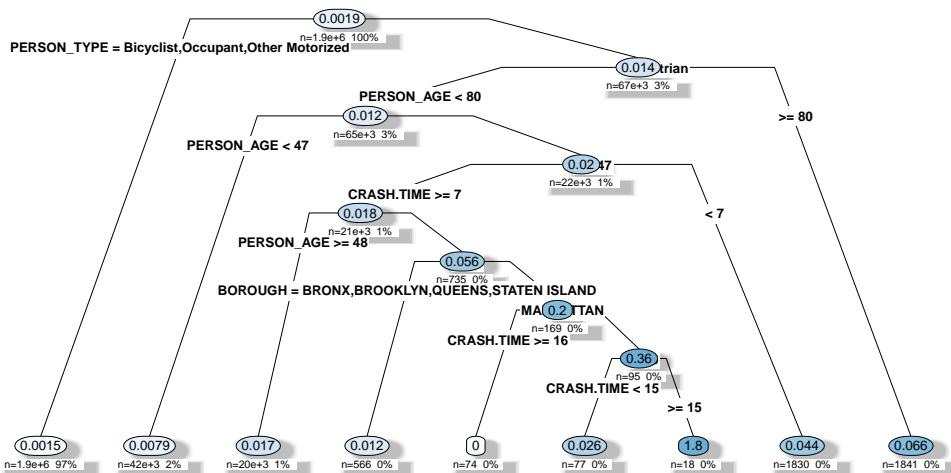
**Contributing Factors:** Distracted driving emerged as a leading cause of accidents, highlighting the need for increased awareness and possibly stricter regulations on driver attention.

## 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



Before interpreting the decision tree, it is important to understand the theoretical underpinnings of the model. Decision trees are a type of supervised learning algorithm that is mostly used in classification problems. It works for both categorical and continuous input and output variables. The goal of using a decision tree is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. The decision tree model is a flowchart-like structure in which each internal node represents a test on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. The paths from the root to the leaf represent classification rules. Each branch represents a choice or condition leading to different outcomes.

In this case, 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, and the complexity of the model is not necessary for the task at hand. However, the model is a good starting point for predicting the number of people killed in car crashes.

### Statistical Modeling with Generalized Linear Model

Call:

```
glm(formula = NUMBER.OF.PERSONS.INJURED + NUMBER.OF.PERSONS.KILLED ~
    CRASH.TIME + BOROUGH + DRIVER_SEX + PERSON_AGE + CRASH.DATE,
    family = poisson, data = subsampled_data)
```

Coefficients:

|                      | Estimate | Std. Error | z value | Pr(> z )     |
|----------------------|----------|------------|---------|--------------|
| (Intercept)          | -0.90359 | 0.08081    | -11.182 | < 2e-16 ***  |
| CRASH.TIMEEvening    | 0.21774  | 0.03646    | 5.973   | 2.33e-09 *** |
| CRASH.TIMELate Night | 0.13263  | 0.05125    | 2.588   | 0.00965 **   |
| CRASH.TIMEMorning    | -0.08898 | 0.04040    | -2.202  | 0.02764 *    |
| BOROUGHBROOKLYN      | -0.03479 | 0.04342    | -0.801  | 0.42301      |
| BOROUGHMANHATTAN     | -0.50420 | 0.05498    | -9.171  | < 2e-16 ***  |
| BOROUGHQUEENS        | -0.10024 | 0.04515    | -2.220  | 0.02640 *    |
| BOROUGHSTATEN ISLAND | 0.40392  | 0.06678    | 6.049   | 1.46e-09 *** |
| DRIVER_SEXF          | 0.29368  | 0.05501    | 5.339   | 9.35e-08 *** |
| DRIVER_SEXM          | 0.32319  | 0.04965    | 6.510   | 7.53e-11 *** |
| DRIVER_SEXU          | 0.44889  | 0.29255    | 1.534   | 0.12493      |
| PERSON_AGE25-34      | -0.04782 | 0.05454    | -0.877  | 0.38063      |
| PERSON_AGE35-44      | -0.10588 | 0.05714    | -1.853  | 0.06389 .    |
| PERSON_AGE45-54      | -0.09734 | 0.05896    | -1.651  | 0.09872 .    |
| PERSON_AGE55-64      | -0.16729 | 0.06394    | -2.616  | 0.00889 **   |
| PERSON_AGE65+        | -0.21497 | 0.07247    | -2.966  | 0.00302 **   |
| PERSON_AGEUnder 18   | -0.26254 | 0.05977    | -4.393  | 1.12e-05 *** |
| CRASH.DATEMonday     | -0.11529 | 0.05587    | -2.063  | 0.03907 *    |
| CRASH.DATESaturday   | 0.05179  | 0.05311    | 0.975   | 0.32956      |
| CRASH.DATESunday     | 0.10476  | 0.05421    | 1.933   | 0.05329 .    |
| CRASH.DATEThursday   | 0.12289  | 0.05217    | 2.356   | 0.01849 *    |
| CRASH.DATETuesday    | -0.09449 | 0.05559    | -1.700  | 0.08920 .    |
| CRASH.DATEWednesday  | -0.06625 | 0.05489    | -1.207  | 0.22742      |
| <hr/>                |          |            |         |              |

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

```
Null deviance: 14215 on 9998 degrees of freedom
Residual deviance: 13845 on 9976 degrees of freedom
(1 observation deleted due to missingness)
AIC: 20073
```

Number of Fisher Scoring iterations: 6

Generalized Linear Models (GLMs) are an extension of traditional linear regression models that allow for response variables to have error distributions other than the normal distribution. GLMs are used to model relationships between a scalar-dependent variable and one or more independent variables. The model consists of three components: (1) the random component, which specifies the distribution of the response variable; (2) the systematic component, which is the linear combination of the predictors; and (3) the link function, which connects the mean of the distribution of the response variable to the linear predictor. In this analysis, we utilize a Poisson distribution, appropriate for count data such as the number of people injured or killed in car crashes. The Poisson distribution models data where events occur independently within a fixed interval and assumes that the number of events is proportional to the length of the interval.

The generalized linear model was trained to predict the number of people injured or killed in car crashes based on the following features: CRASH.TIME, BOROUGH, DRIVER\_SEX, PERSON\_AGE, and CRASH.DATE. The model was trained on a subsample of 10,000 rows of the data in order to speed up the model fitting process. The fitted GLM (whose results are shown above) provides a summary of the model's coefficients, standard errors, z-values, and p-values. The coefficients represent the estimated effect of each predictor variable on the response variable, while the p-values indicate the statistical significance of these effects. The results indicate that the time of day, borough, driver's sex, the person's age group, and the day of the week significantly influence the likelihood of injuries or fatalities. For instance, evening times and late nights have higher associated counts of injuries or fatalities compared to other times of the day, highlighting riskier periods for road users. Manhattan, compared to other boroughs, shows a significantly lower count, suggesting geographical differences in traffic safety or reporting practices. Additionally, the model reveals the older age groups (55-64, 65+) have a lower likelihood of being involved in injurious or fatal crashes compared to younger age groups. The day of the week also plays a role, with Thursday showing a higher likelihood of crashes compared to other days, suggesting possible variations in traffic patterns or behaviors.

It is important to note, however, that while GLM provides valuable insights, it is based on associations and does not imply causation. The model's assumptions, including the independence of events and the suitability of the Poisson distribution for this data, should be carefully considered when interpreting the results and planning interventions.

## Conclusion

Through our exploration, we've unearthed patterns and insights that shed light on the complex nature of car crashes in New York City. Our findings point toward actionable insights, such as the need for targeted interventions during high-risk times of the day, the importance of considering geographical differences in traffic safety, and the potential for tailored interventions for different age groups. While our study offers detailed analysis, it is only the beginning of a broader effort to understand and address the multifaceted issue of car crashes. Future work could involve more advanced predictive modeling, such as ensemble methods or neural networks, to improve the accuracy of our predictions. Additionally, further research could explore the impact of external factors, such as weather conditions, road infrastructure, and vehicle safety features, on the likelihood of car crashes. By continuing to investigate and address the complex issue of car crashes, we can work toward a safer and more sustainable future for all road users.

As it stands, some key findings from our analysis can be summarized as follows into these practical insights:

- Most crashes occur during morning and evening rush hours, suggesting the need for targeted interventions during these high-risk times.
- The borough of Manhattan shows a significantly lower count of injuries or fatalities compared to other boroughs, suggesting geographical differences in traffic safety or reporting practices
- The leading cause of car crashes is driver inattention or distraction, highlighting the need for increased awareness and possibly stricter regulations on driver attention.

The next steps in this project could involve more advanced predictive modeling, such as neural networks, to improve the accuracy of our predictions. A good possible type of neural network to use would be a recurrent neural network (RNN) which is good for time series data. Additionally, ShapeFiles, which are a type of geographic information systems (GIS) file, could be used to map out the locations of car crashes and identify hotspots with a high degree of accuracy.

So far, this project has revealed much but also shown us how much more there is to learn.