Group 3: Crash Analysis NYC- Draft 4

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
- Pie Chart of Contributing Factors to Car Crashes . . . 5
- Histogram of Car Crashes by Time of Day . . . 6
- Line Plot of Top 3 Contributing Factors by Time of Day . . . 7
- Temporal Density Maps . . . 8
- Stack Area Plot of Contributing Factors and Crash Count by Year . . . 11
- Jitter Plot of Fatalities and Injuries by Contributing Factor . . . 12
- Bubble Plot Between Contributing Factor and Vehicle Type . . . 13
- Key Findings From the Exploratory Data Analysis . . . 14

Predictive Modeling . . . 15
- Decision Tree . . . 15
- Statistical Modeling with Generalized Linear Model . . . 16

Conclusion . . . 18

Introduction

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.

In the bustling metropolis of New York City, the largest city in the United States, motor vehicle crashes occur frequently and result in substantial economic costs. They cause injuries, deaths, financial damages, and create an contribute to an atmosphere of uncertainty and fear. Reducing the number of crashes, especially those that are fatal, is key to improving the safety of the city.

In this project, we will examine data on motor vehicle crashes in New York City to uncover underlying trends and discern the most important factors that contribute to potentially fatal accidents. By analyzing the data, we seek to provide stakeholders with actionable insights and recommendations to reduce the number of collisions and deaths.

Data Description

The NYC OpenData dataset contains details of motor vehicle collision occurrences reported by the NYPD since 2012 (the dataset is updated with new data frequently). 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

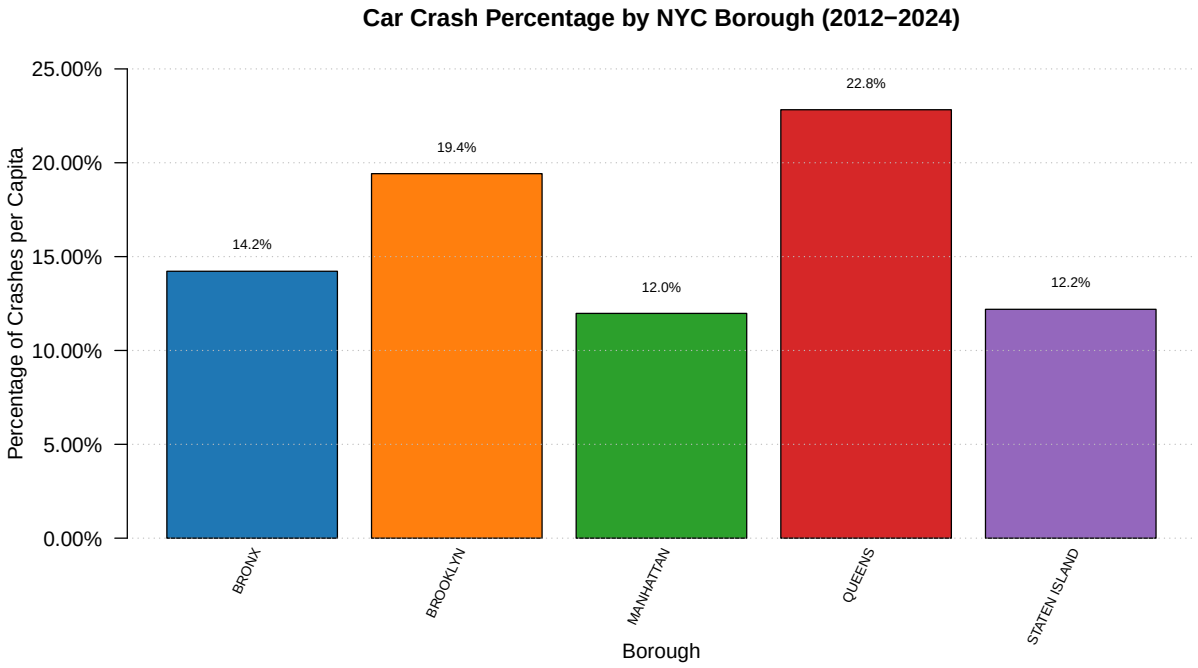
For the data cleaning process, we removed rows with missing or invalid values in crucial columns, such as BOROUGH, LATITUDE, LONGITUDE, CRASH.DATE, and CRASH.TIME. We also converted the CRASH.DATE column to a Date format and the CRASH.TIME column to a more standard format. We replaced empty strings with “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 allows us to gain intuition before we do predictive modeling and make recommendations. The graphs and charts below tell a story of the historic trends in NYC car crashes.

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

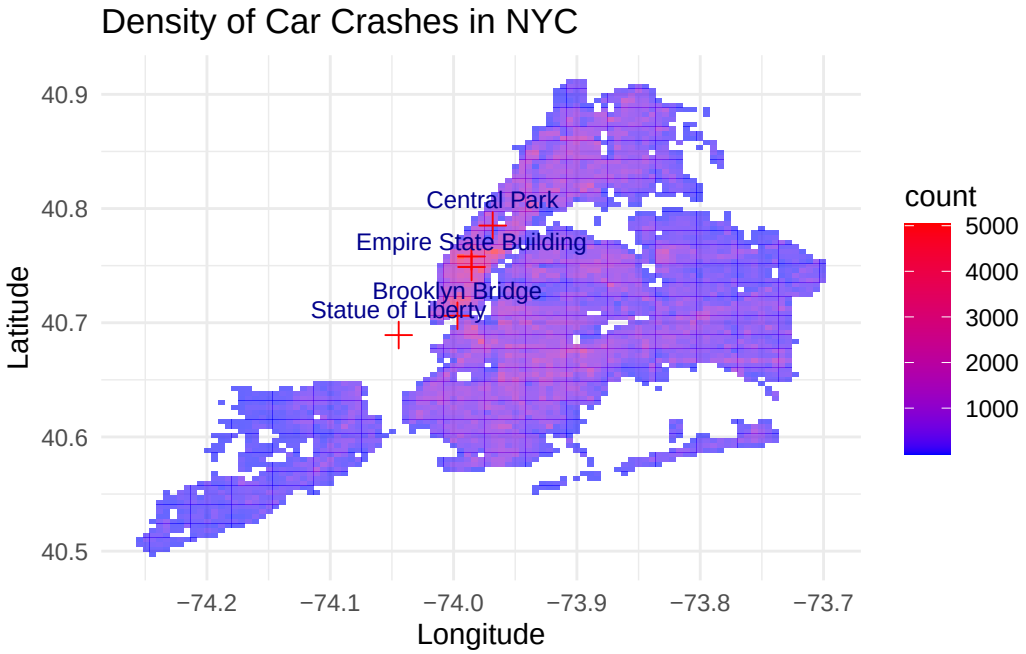


We plotted the frequency of car crashes per borough in NYC. The goal of this graph was to gain a better understanding of which boroughs were more likely to result in car crashes.

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 amount of traffic through the boroughs.

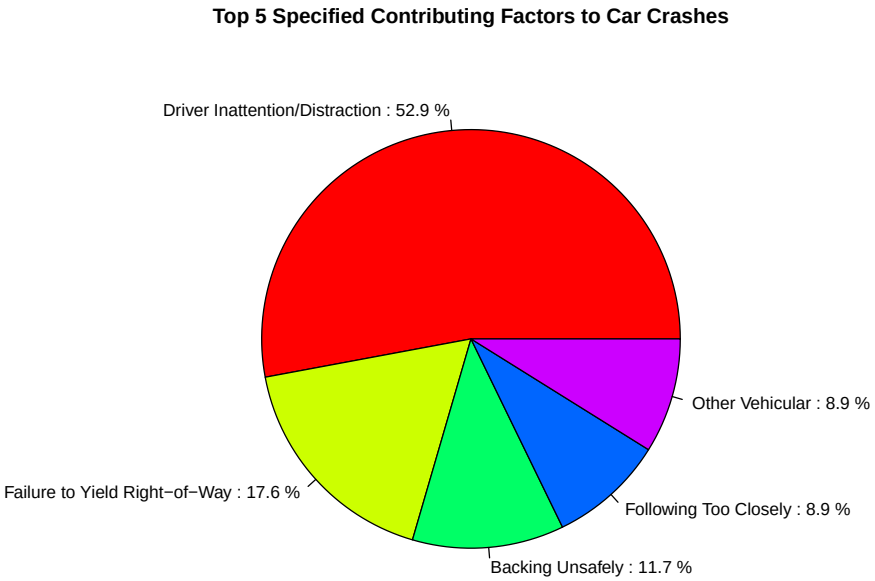
An important caveat for this graphic is that the population counts used to determine the percentages are from the most recent census, and we did not factor in population change across borough from 2012-2024. This affects the accuracy of the derived percentages, but the general trends are still useful for understanding the likelihood of car crashes in each borough.

Density of Car Crashes in NYC



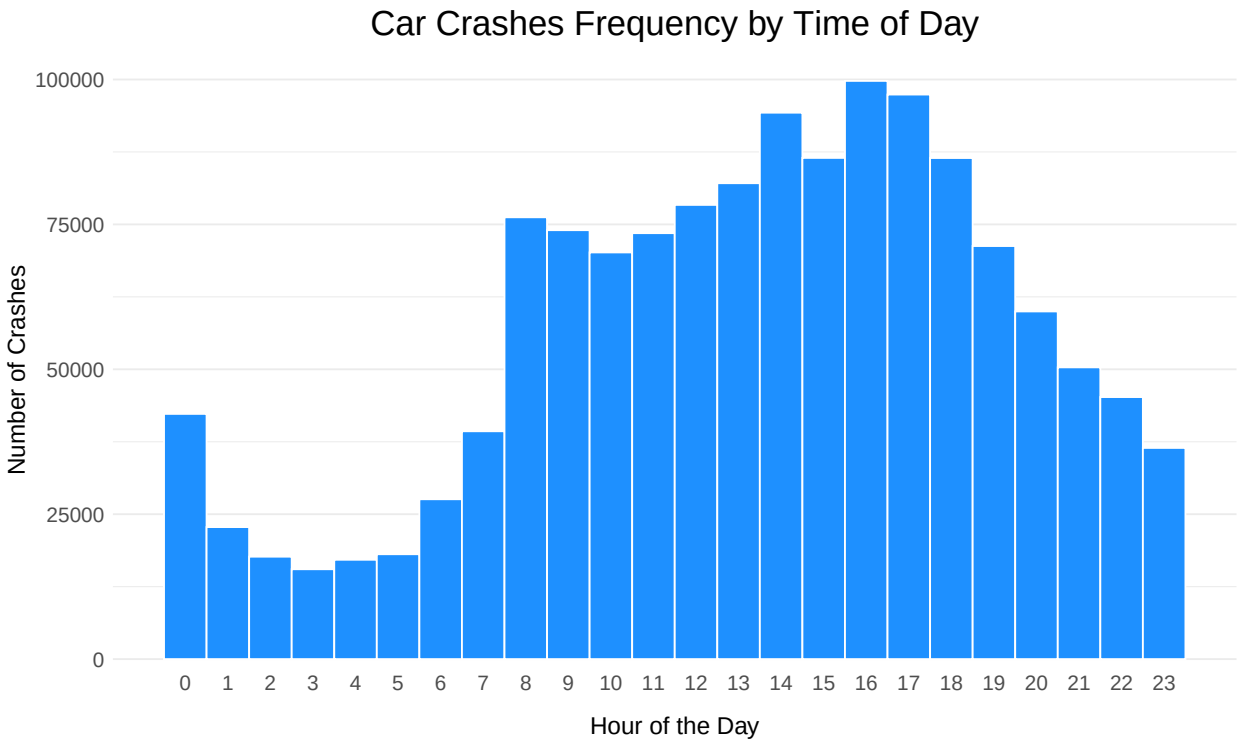
The heat map of car crashes in NYC shows that the highest density of car crashes occurs in the center of NYC. This is likely due to the fact that the center of NYC has the most traffic and the most people. The lowest density of car crashes occurs in the outskirts of NYC. This aligns with our intuition, 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.

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. 531,539 crashes in the cleaned data set have unspecified contributing factors, which is 38% of the all crashes in the data. There was around twice as many crashes with unspecified contributing factors than the number of crashes with driver inattention as a contributing factor.

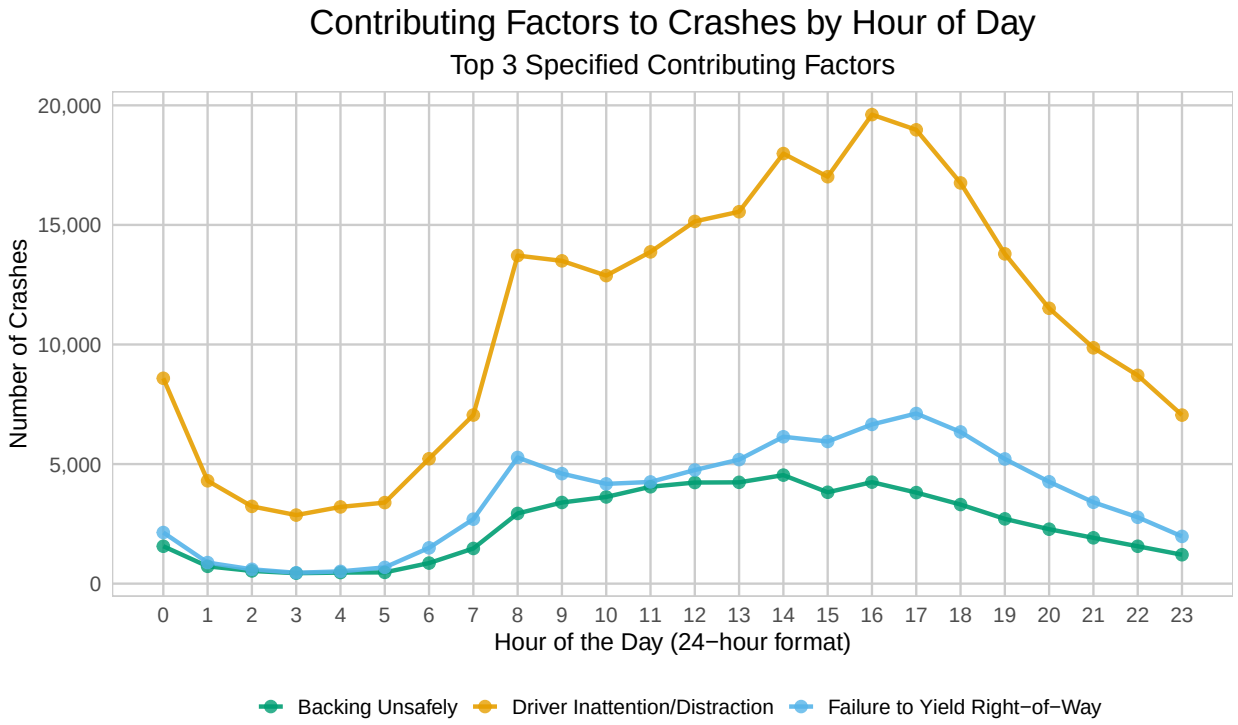
Histogram of Car Crashes by Time of Day



The histogram of car crashes by time of day shows that the most car crashes occur around the hours of 3:00PM and 6:00PM. 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 3:00AM and 4:00AM, 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.

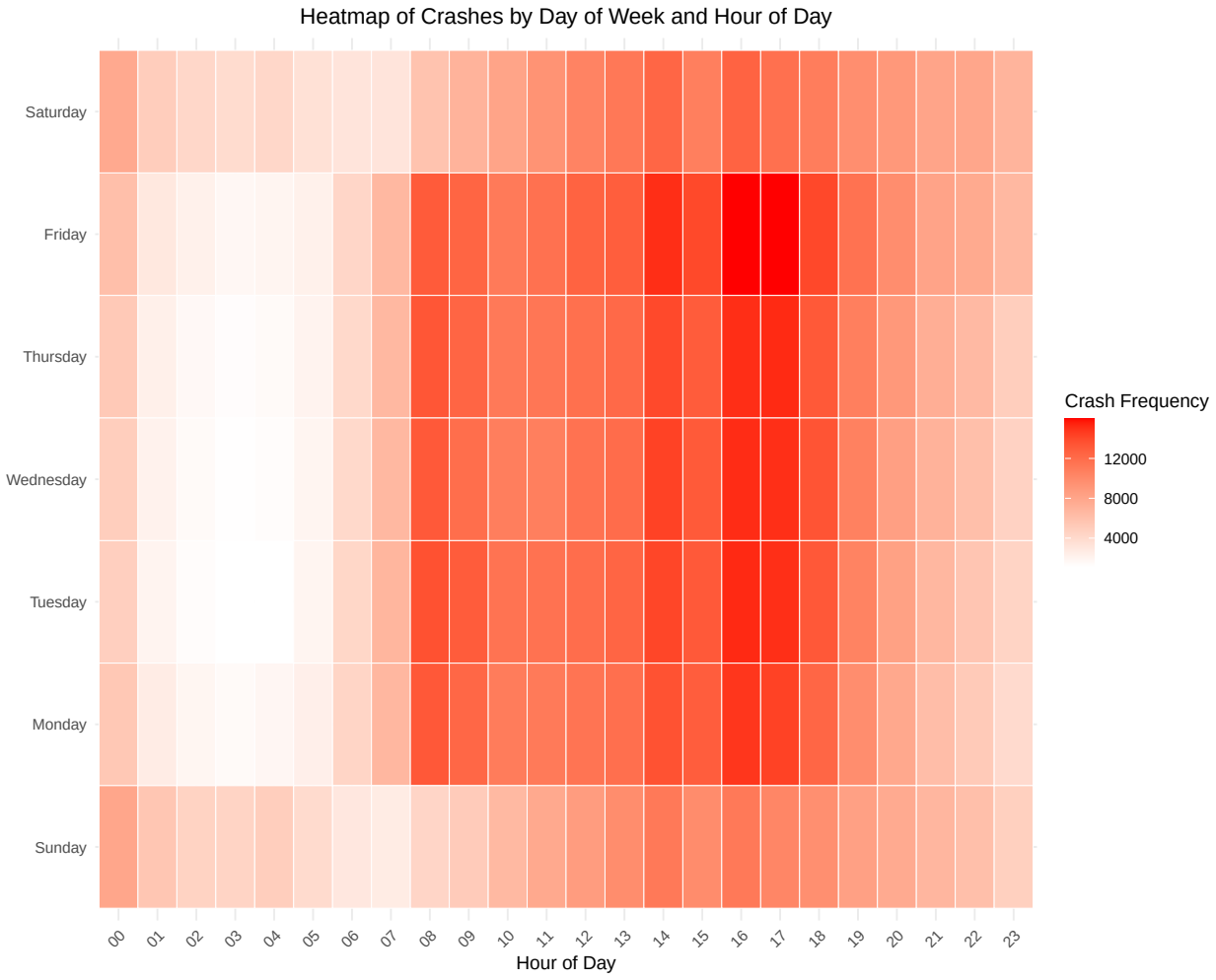
Line Plot of Top 3 Contributing Factors by Time of Day

From the previous histogram of car crashes by time of day, we observe that hour of the day does seem to influence the risk of car crashes. We now pose the question: Throughout the day, what factors contribute the most to car crashes?



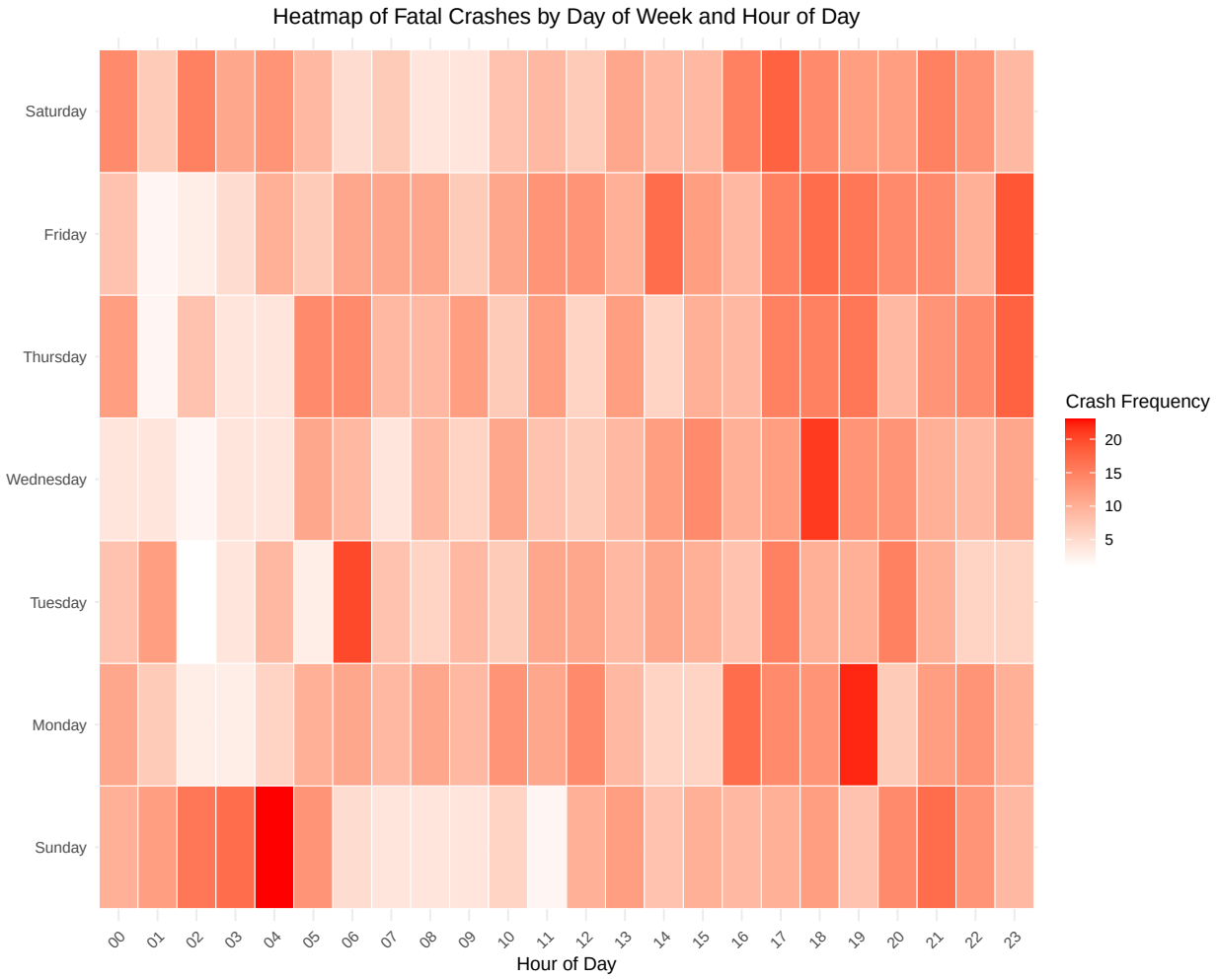
The scatter plot of the top 3 contributing factors by time of day show how the top factors contribute to the number of car crashes throughout the day. What can be seen is that distracted driving is consistently the biggest factor contributing to car crashes. During hours with the most crashes, such as 8:00AM and 5:00PM, all 3 factors reach peaks. All contributing factors show that the crash amount increases between the hours of 8:00AM and 5:00PM 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 Maps

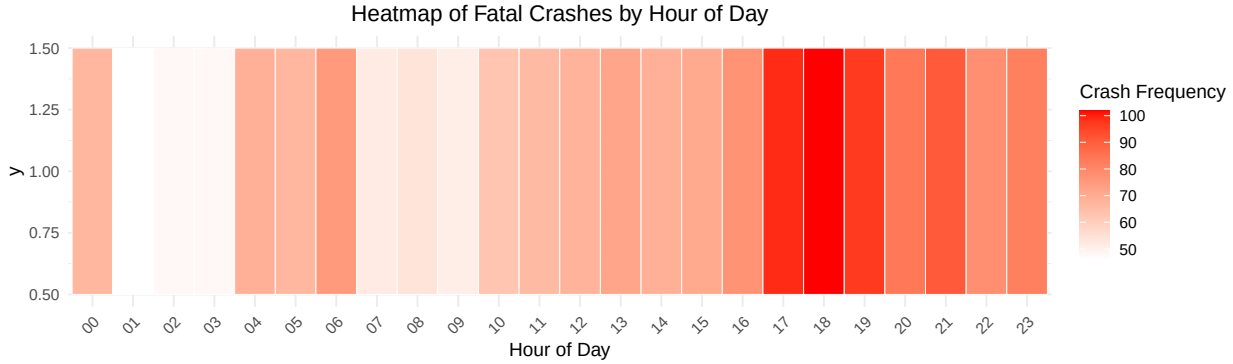


The temporal density map shows the frequency of car crashes by day of the week and hour of the day. There is an evident "safe zone" on weekdays between the hours of 1:00AM and 5:00AM. This is expected as these are not usually times where there are lots of people active. Interesting, "danger zones" are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

The most damaging car crashes are of course those that result in loss of life. Since one of the primary goals of our analysis is to examine potential trends that result in fatal car crashes, we pose the following question: When are fatal car crashes more likely to occur?



This temporal density map shows the frequency of fatal car crashes by day of the week and hour of the day. Unlike the previous heat map, there is no noticeable trend. This heat map suggests that time during the week has no impact on instances of fatal car crashes. However, we can narrow the scope further to examine if time of day in general influences the frequency of fatal car crashes. The following plot does just that.



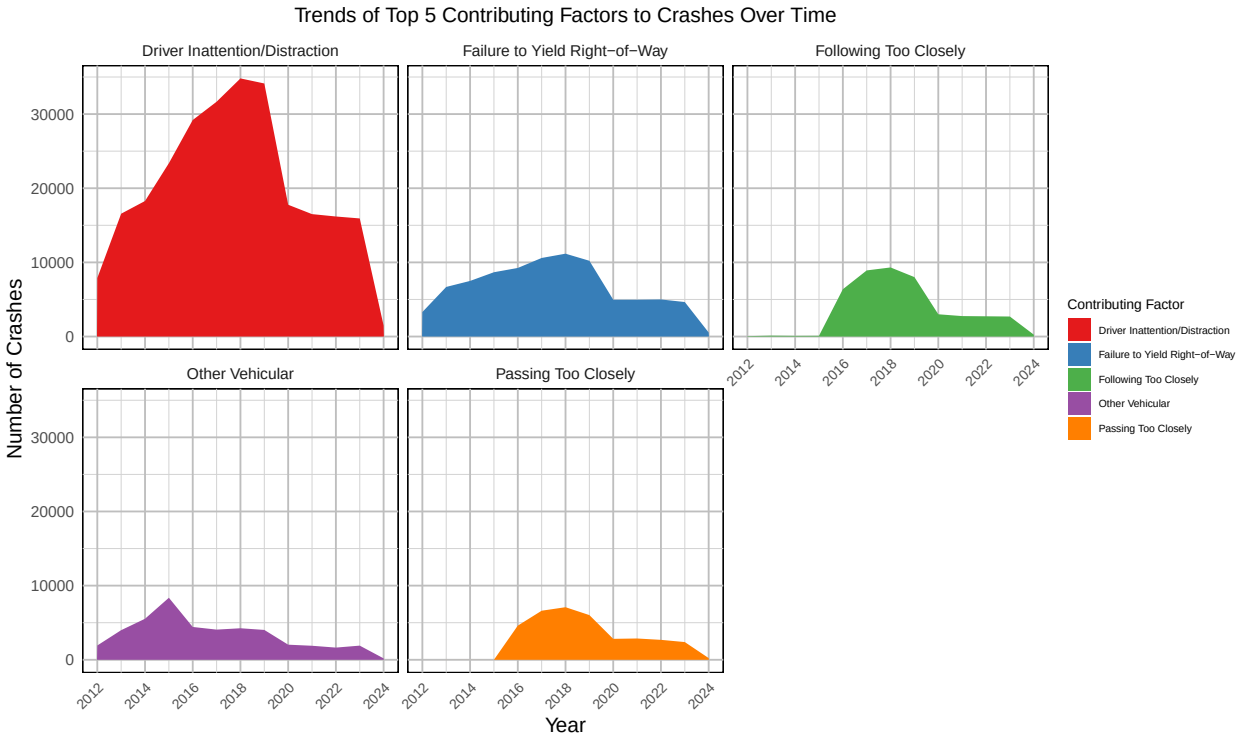
This temporal density map shows the frequency of fatal car crashes by only the hour of the day. This heat map suggests that time of the day does have a significant impact on instances of fatal car crashes. In particular, fatal car crashes seem to be more likely to occur in the evening or at night. This may be due to reduced visibility at later hours. The number of fatal car crashes peaks drastically between the hours of 5:00PM and 7:00PM. This is likely due to the compounding factors of reduced visibility at night and

increased traffic with people leaving work. In the early morning hours between 1:00AM and 9:00AM, there are less occurrences of fatal crashes. However, there does appear to be a significant number of fatal crashes that occur between the hours of 4:00AM and 6:00AM. This is in contrast to the "danger zone" observed between the hours of 8:00AM and 10:00AM in the first temporal density map. That "danger zone" was explained increased traffic as people commuted from work. Increased traffic can also partially explain the peak in fatal car crashes between 4:00AM and 6:00AM. However, reduced visibility in early morning hours likely compounded the risks.

The three temporal density plots above show that car crashes are more likely to occur during hours of high traffic when people are commuting to and from work. However, fatal car crashes are more likely caused by reduced visibility during darker hours, although amount of traffic is still a partially contributing factor.

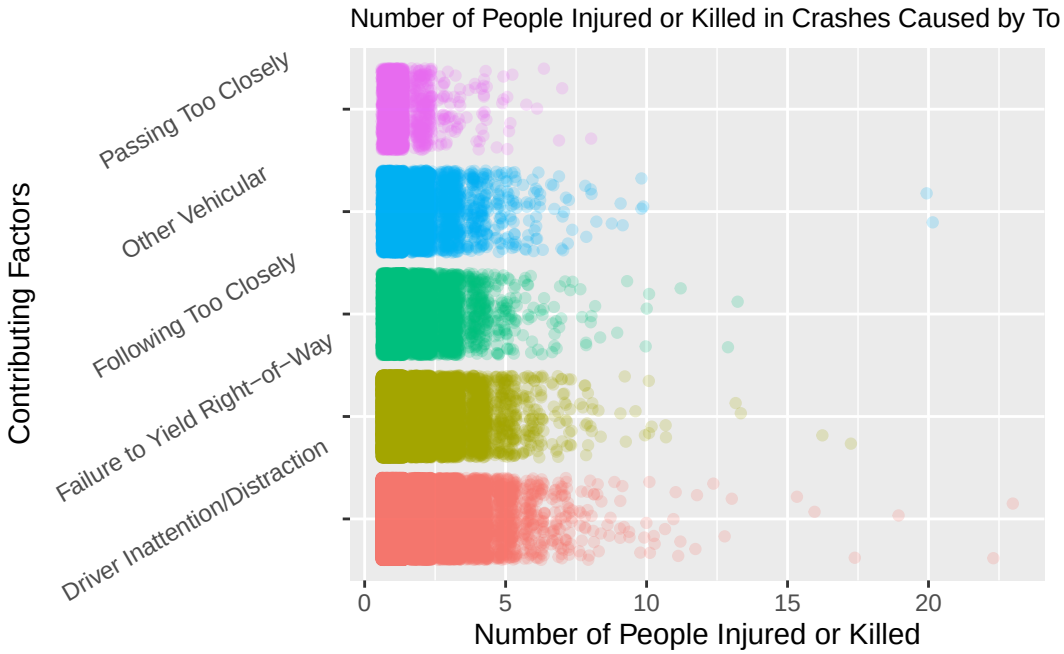
Stack Area Plot of Contributing Factors and Crash Count by Year

Now that we have an idea of which contributing factors to crashes are most common, we pose the following question: How have the frequency of contributing factors to crashes changed over time?



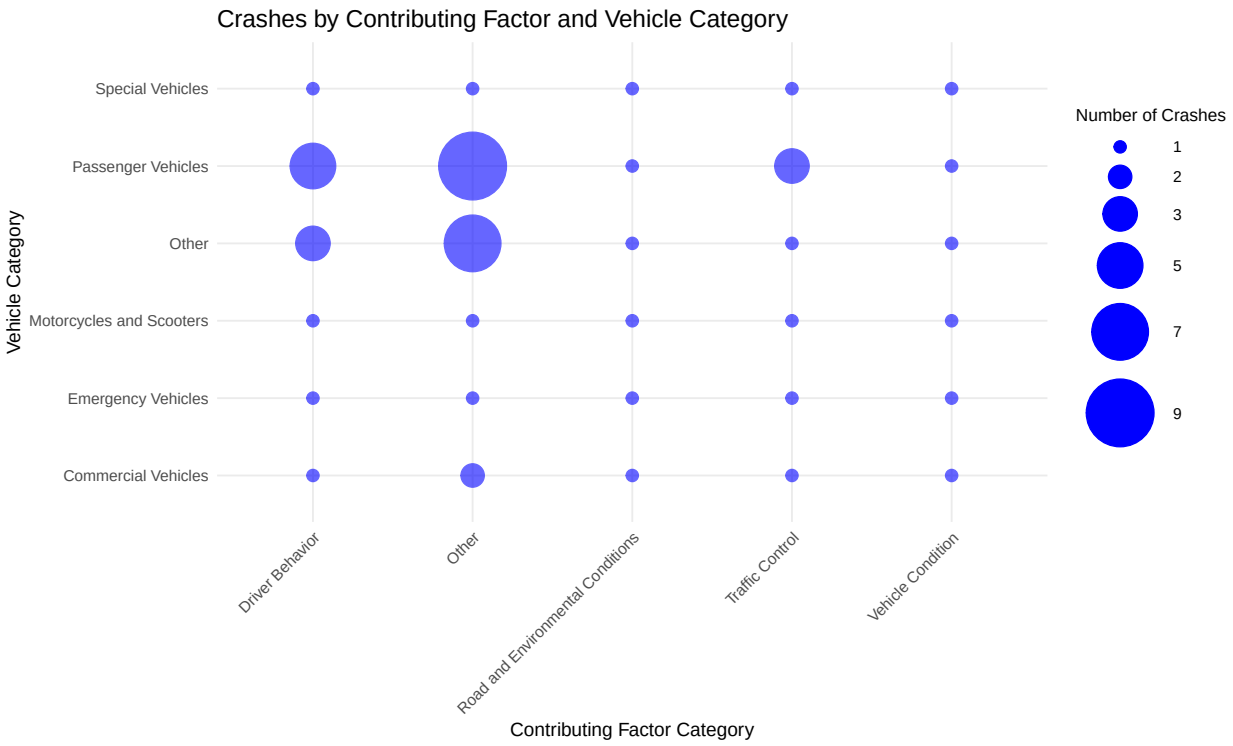
The stacked area plot of the top 5 contributing factors to car crashes by year shows how the top contributing factors changed over the years. What can be seen is that the number of crashes peaked in 2018 at around 67,500 crashes and that driver inattention/distraction grew rapidly from 2012 to 2018. 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.

Bubble Plot Between Contributing Factor and Vehicle Type



The bubble plot shows the relationship between contributing factors and vehicle types in car crashes. The size of the bubble represents the number of crashes. The plot shows that the most common contributing factors are related to driver behavior. The most common vehicle types involved in crashes are passenger vehicles. This plot gives us a better understanding of how contributing factors and vehicle types are related in car crashes.

Key Findings From the Exploratory Data Analysis

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. Fatal crashes are prominent during the evening and the night, when darkness causes reduced visibility.

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

Temporal Trends: The frequency of crashes caused by the top 5 contributing factors has been decreasing since 2020, which could be due to the COVID-19 pandemic.

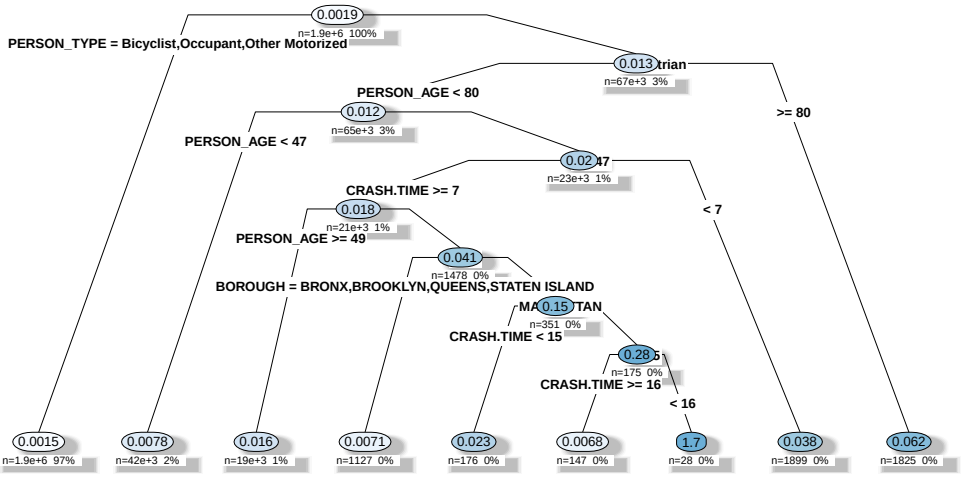
Predictive Modeling

Decision Tree

```
Mean Absolute Error (MAE): 0.003833112

Root Mean Squared Error (RMSE): 0.05435656
```

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.

Understanding the hyper parameters of the decision tree and the rationale behind is also important. The minimum number of observations that must exist in a node in order for a split to be attempted is 10. The complexity parameter (cp) is set to 0.001, which is the cost complexity parameter used for pruning the tree. The model is trained on 70% of the data and tested on the remaining 30%. The model uses ANOVA (Analysis of Variance) as the method for splitting the data.

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 0.003833112 and a Root Mean Squared Error (RMSE) of 0.05435656. 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

	NUMBER.OF.PERSONS.INJURED + NUMBER.OF.PERSONS.KILLED		
Predictors	Incidence Rate Ratios	CI	p
(Intercept)	0.51	0.43 – 0.59	**<0.001**
CRASH TIME [Evening]	1.24	1.15 – 1.33	**<0.001**
CRASH TIME [Late Night]	1.10	0.99 – 1.22	0.063
CRASH TIME [Morning]	0.91	0.84 – 0.99	**0.023**
BOROUGH [BROOKLYN]	0.97	0.89 – 1.05	0.451
BOROUGH [MANHATTAN]	0.61	0.54 – 0.67	**<0.001**
BOROUGH [QUEENS]	0.90	0.83 – 0.99	**0.027**
BOROUGH [STATEN ISLAND]	1.50	1.32 – 1.71	**<0.001**
DRIVER SEX [F]	1.05	0.95 – 1.16	0.381
DRIVER SEX [M]	1.13	1.04 – 1.24	**0.006**
DRIVER SEX [U]	0.28	0.05 – 0.87	0.073
PERSON_AGE25-34	0.92	0.83 – 1.03	0.135
PERSON_AGE35-44	0.78	0.70 – 0.88	**<0.001**
PERSON_AGE45-54	0.86	0.77 – 0.96	**0.008**
PERSON_AGE55-64	0.79	0.70 – 0.90	**<0.001**
PERSON AGE [65+]	0.76	0.66 – 0.87	**<0.001**
PERSON AGE [Under 18]	0.87	0.78 – 0.97	**0.015**
CRASH DATE [Monday]	0.89	0.80 – 0.99	**0.038**
CRASH DATE [Saturday]	1.06	0.95 – 1.17	0.306
CRASH DATE [Sunday]	1.12	1.00 – 1.24	**0.044**
CRASH DATE [Thursday]	1.13	1.02 – 1.26	**0.017**
CRASH DATE [Tuesday]	0.91	0.82 – 1.02	0.095
CRASH DATE [Wednesday]	0.94	0.84 – 1.05	0.251
Observations	9999		
R^2 Nagelkerke	0.043		

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.

- Fatal crashes are prominent during the evening and the night, when darkness causes reduced visibility.

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.