Car Crashes in New York City

Sawyer Cremer, Sophie Chikhladze, Adarsh Gadepalli, Rahul Prakash

Table of contents

Introduction

1.3 million people die each year as a result of road traffic crashes, and between 20 and 50 million people every year suffer non-fatal injuries resulting from car accidents. Other than the millions of lives lost, the U.S. Department of Transportation's most recent estimate of the annual economic cost of crashes is $340 billion. This is an issue affecting virtually every country in the world, including the United States. In New York City, New York, the New York Police Department is required to be fill out a report for every collision where someone is injured or killed, or where there is at least $1000 worth of damage. We decided to investigate this dataset and ask questions about which zip codes or boroughs have the most car crashes and why, which days of week and hours are more accident-prone, does weather have an effect on the number of accidents in a day, and more!

Primary Dataset

The Motor Vehicle Collisions dataset contains information about car crashes in NYC from the years 2016-2023. It contains information from all police reported motor vehicle collisions in NYC. We accessed the dataset from data.gov, provided by the City of New York. Each entry represents a car crash that ocurred in New York City. The dataset contains 28 columns that describe a variety of data, including time of accident, accident severity(deaths/injuries), pinpointed locations (neighborhoods/streets), causes for the accident, and vehicle type. Using this data, we hope to analyze and identify chronological and geographical patterns in New York City to pinpoint high-risk environments and circumstances. We are assuming that the data is continuous in terms of time, and that the specific time stamps and event details are accurate. Every plot in this document employs the primary data set, unless specified otherwise.

Plots

Number of Car Crashes over Time

First, we decided to plot a time series of the number of car crashes each day starting from 2012 up until 2023. Each point represents the total number of crashes that occurred on each day, plotted chronologically.

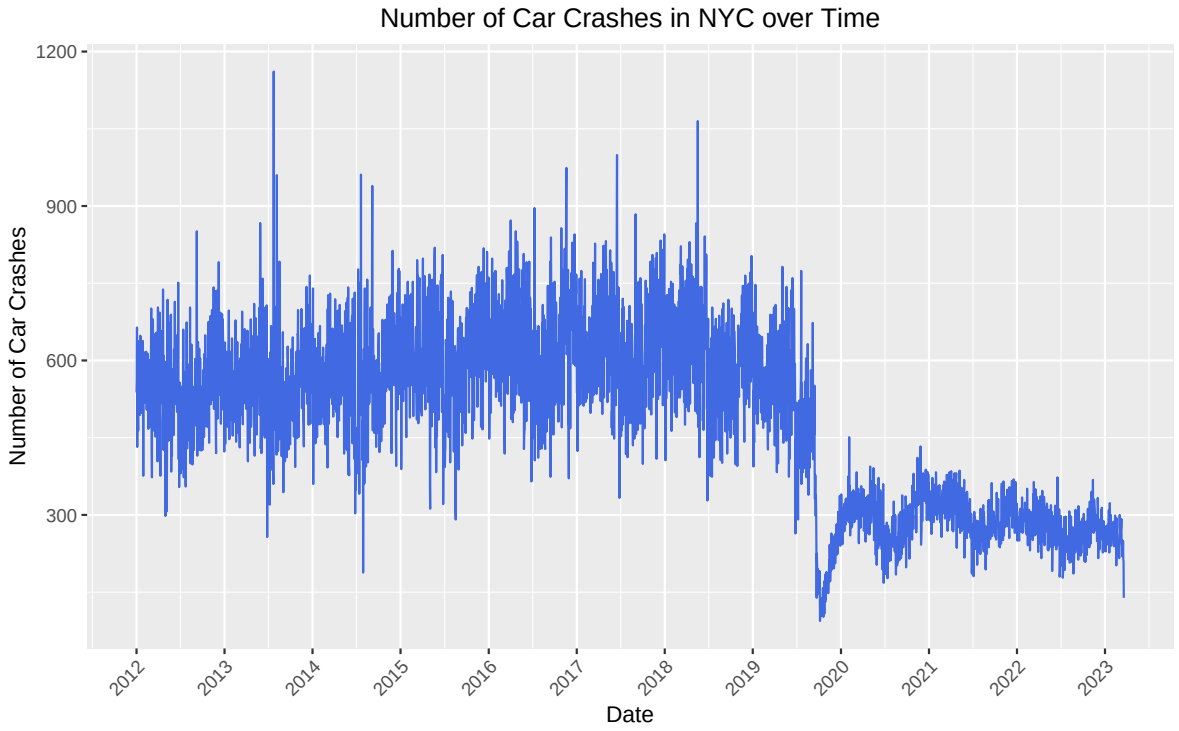


Figure 1: Number of Car Crashes over Time

Looking at the number of crashes over time, from 2012 to 2023, we can see that the number of crashes remained fairly consistent from 2012 to the beginning of 2020, with certain exceptions like big spikes or drops. Until 2020, there is no year with an exceptionally low or high number of crashes. However, we see a clear change in the number of crashes in the beginning of 2020, which can be interpreted as the effects of COVID-19 as that is roughly the date it started. Less people had to go to work, school became virtual, and therefore less people were driving, leading to a lowered number of car crashes per day. We could also further study the exact dates of this drop versus when the lockdown started.

Number of Car Crashes and Deaths per Borough

For this plot, we grouped the data by 5 boroughs of New York City: Bronx, Brooklyn, Manhattan, Queens, and Staten Island. First, we calculated the total number of car crashes in each borough, and plotted it in tens of thousands since there was a large amount. Then, we wanted to see if the number of casualties in each borough was proportional to the number of car crashes. Therefore, we calculated the total number of deaths from car accidents per borough and plotted that beside the number of car crashes.

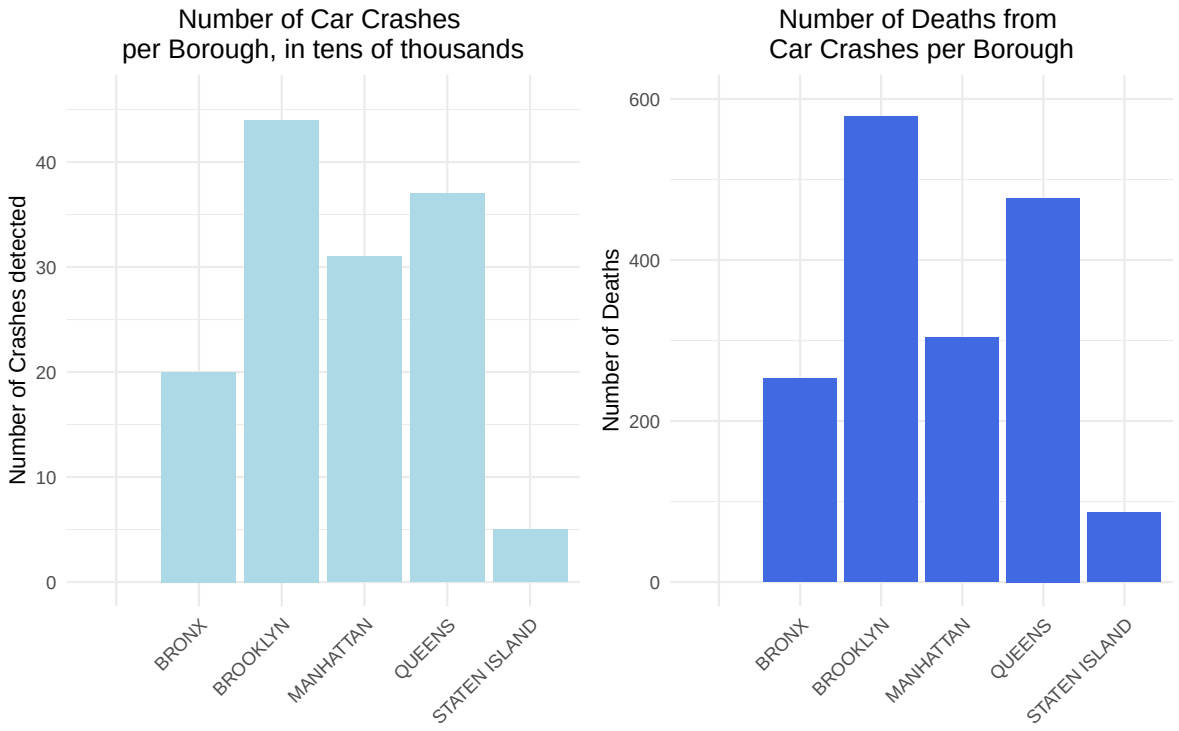


Figure 2: Number of Car Crashes and Deaths per Borough

From the left plot, we can see that Brooklyn has the highest amount of crashes, followed by Queens and Manhattan. Staten Island has the lowest number of crashes recorded, by a significant amount. Despite the fact that Staten Island has the highest number of cars per household, it is sensible that the rest of the boroughs have more car accidents, since there is more general traffic there. Furthermore, if we compare the number of crashes per borough to the number of casualties per borough, We can see that the proportions of deaths are similar to the number of accidents, except for Manhattan, which is slightly lower.

Average Accidents per Day per Season

Next, we asked ourselves - is an accident more or less likely to occur during a specific season? For this, we turned each date into a season and calculated the number of car crashes per day for each season. The, we plotted 4 boxplots to represent each season portraying the number of car crashes per day.

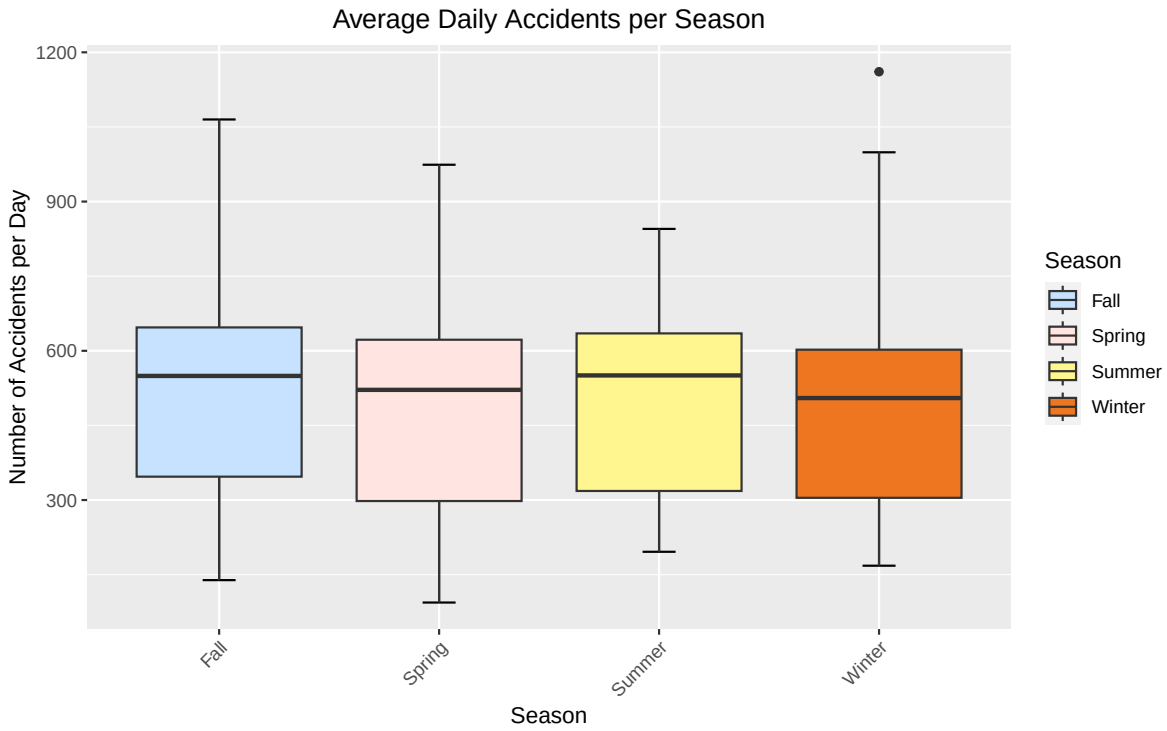


Figure 3: Average Accidents per Day per Season

As we can see, there is a relatively similar amount of car crashes across all the seasons, suggesting that the season does not significantly affect the likelihood of crashes occurring. One interesting point is that there is a much larger variance and spread of car crashes per day in the fall while there is a significantly smaller variance of car crashes per day in the summer. There is also an outlier in the Winter, where over 1200 crashes happened on one day. However, this is just an outlier since the median number of crashes in Winter is slightly lower than the median number of crashes in the Summer and Fall.

Regression over Total Number of Car Crashes vs Precipitation per Day in NYC

This is a linear regression over a plot of car crashes vs precipitation in New York City, New York. Displayed in the table below are the intercept and the coefficient for precipitation.

Table 1: Regression over Total Number of Car Crashes vs Precipitation per Day in NYC

	Estimate	Standard Error	t value	Pr(>|t|)
(Intercept)	294.693	1.539	191.424	0.0000***
Precipitation	0.045	0.292	0.156	0.8764

*Signif. codes: 0 <= '***' < 0.001 < '**' < 0.01 < '*' < 0.05*

Residual standard error: 42.4 on 850 degrees of freedom

Multiple R-squared: 2.847e-05, Adjusted R-squared: -0.001148

F-statistic: 0.0242 on 850 and 1 DF, p-value: 0.8764

The plot depicting the relationship, as well as further analysis, is on the next page.

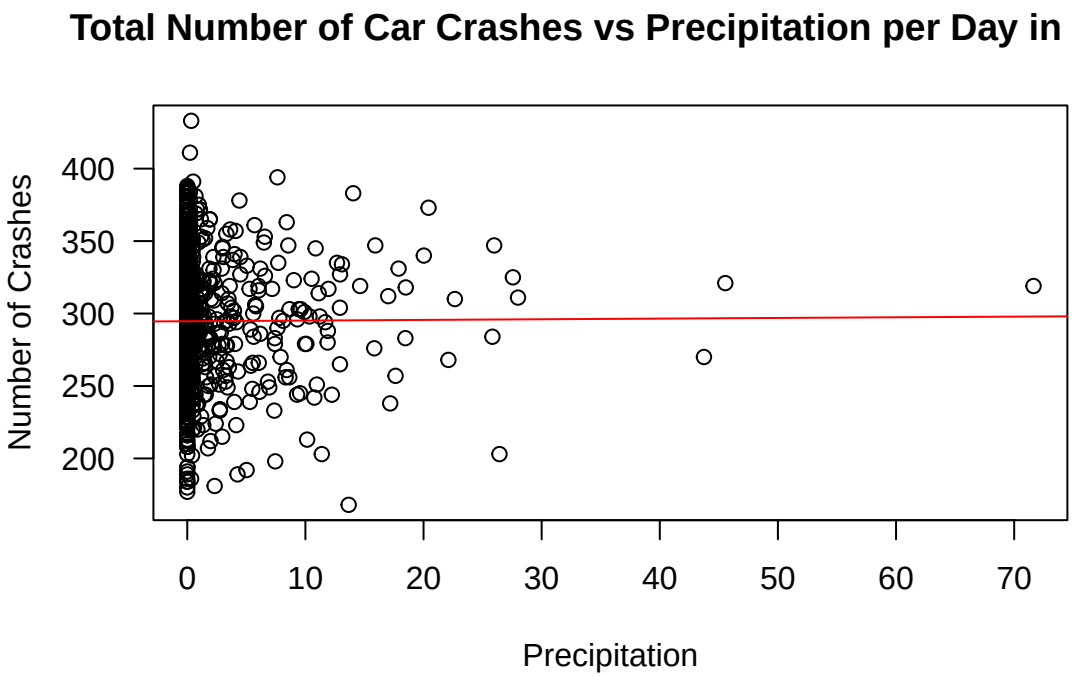


Figure 4: Total Number of Car Crashes vs Precipitation per Day in NYC

The summary of the regression indicates that as there is a one unit increase in precipitation, the average number of car crashes increases by 0.04545. The intercept shows that the estimated average number of car crashes on a day with no precipitation in NYC is 294.69. The precipitation coefficient is not statistically significant because of its high p-value of 0.876, and the t-value of 0.156 further reinforces the lack of statistical significance in the coefficient.

Number of Casualties By Weather Conditions and Number of Cars Involved

The following is a series of bar plots, plotting the number of casualties caused by vehicle crashes to weather conditions during the times of those accidents. Each plot represents the data from car crashes involving 3,4, and 5 cars.

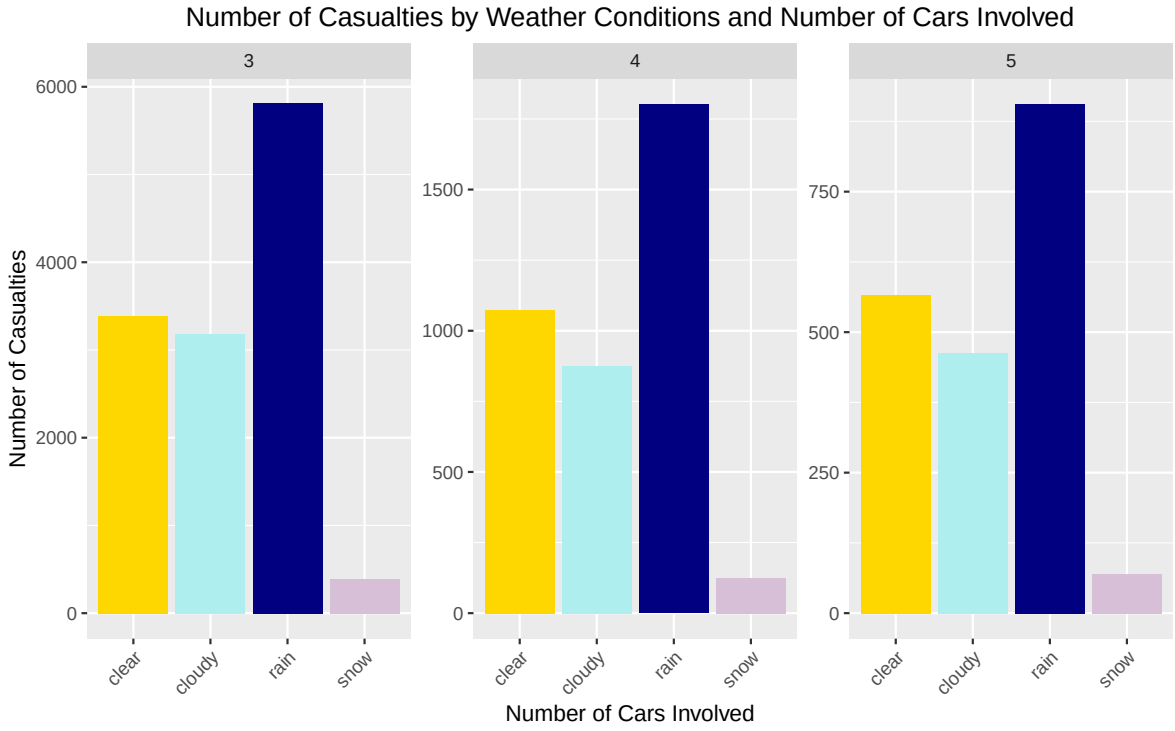


Figure 5: Number of Casualties By Weather Conditions and Number of Cars Involved

The plots display a large plurality of crashes occurring during rainy weather, and a noticeable minority of crashes occurring during snowy weather, which most likely is due to the fact that it doesn't snow for the majority of the year but may also be because many people aren't on the roads during snowy weather. Furthermore, each of the bar plots maintains a relatively consistent casualty distribution across the weather condition categories, with the exception of cloudy weather seemingly having a higher proportion of casualties for 3 car accidents. We also note the casualty count differences across the plots, as the total casualty count is much greater for 3 car accidents, then 4 car, then 5 car.

Number of Car Crashes Over/Under average per Day of Week

The plot below demonstrates a heat map of the number of crashes over/under average per day of week per borough. The numbers represent daily averages for each borough subtracted from the number of crashes per day. This is to demonstrate how the number of crashes varies per day of week. Green represents below average number of crashes, yellow represents an average number of crashes, and orange and red represent an above average number of crashes.

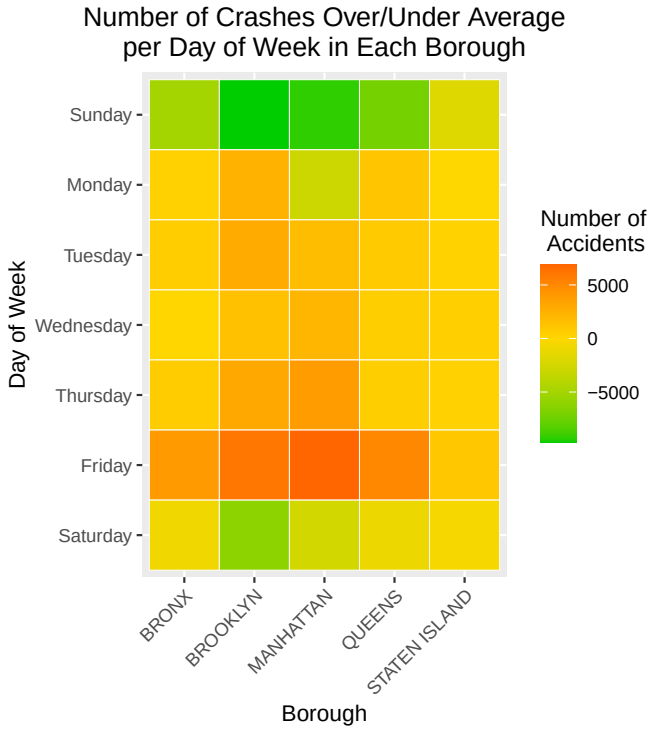


Figure 6: Number of Car Crashes Over/Under average per Day of Week

The heat map shows that in all boroughs, the highest number of crashes occurs on Fridays. The lowest in all boroughs occurs on Sundays. In general, a higher number of crashes happen on weekdays, which makes sense since a lot of New York City residents who work drive to and from work on weekdays. The plot also demonstrates variation in car crashes per borough. Staten Island has the least variation, but also has the least number of crashes as we saw in Figure 2. Manhattan has the highest variation, followed by Brooklyn.

Number of Car Crashes per ZIP code

The plot below demonstrates the number of car crashes per ZIP code in New York City, New York. The plot displays more and less accident-prone zones. As demonstrated in the legend, light yellow is the lower number of crashes, while dark red is the higher number of crashes. The ZIP codes that are gray had no data - indicating that either no car crashes happened in that zone, or none that were recorded. The x axis in the plot represents longitude and the y axis represents latitude.

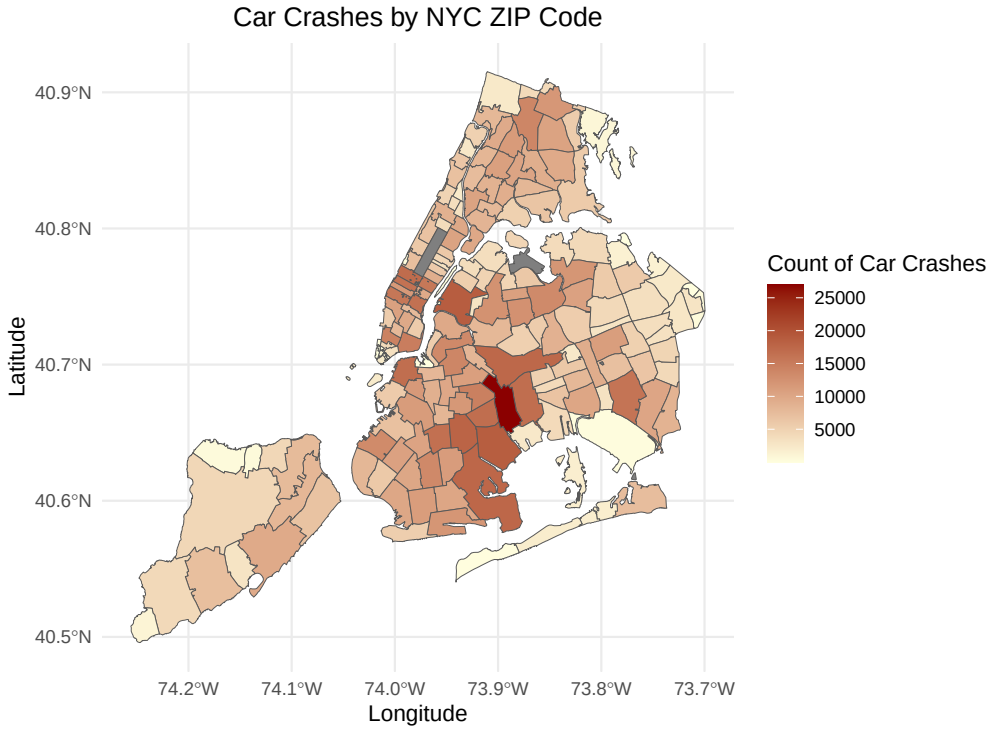


Figure 7: Number of Car Crashes per ZIP code

From the plot, it is visible that some of these darker regions include areas where major bridges lead into and out of Manhattan to Queens and Brooklyn, which might be a factor for why there is a high volume of crashes in these areas. The lighter shaded ZIP codes on the map signify regions with fewer crashes. Generally, there are less crashes in the borough of Staten Island based on this graphic, which might be due to a lower traffic and population. The gray block in Manhattan is central park, so it is logical that no crashes occurred in that ZIP code.

Number of Car Crashes by Hour of Day

The polar bar chart shows the number of car crashes in New York City by hour of the day. Each bar segment corresponds to an hour, labeled from 12:00AM to 11:00PM. A longer bar indicates a higher number of crashes occurred during that hour.

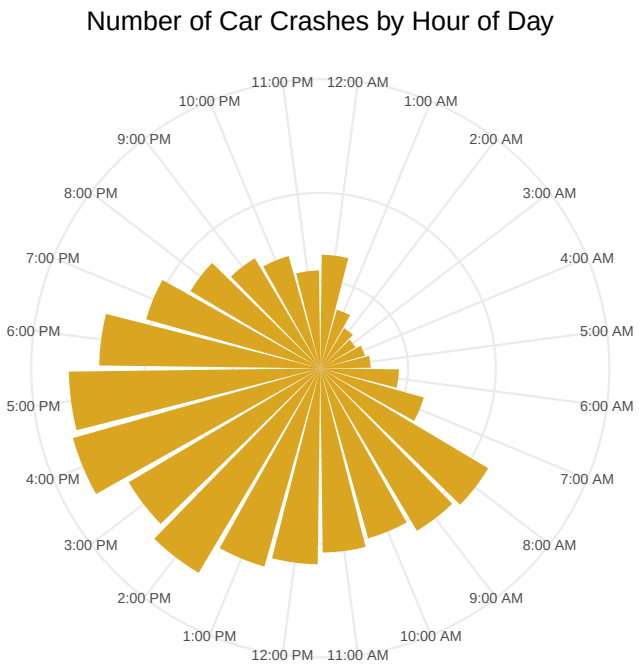


Figure 8: Number of Car Crashes by Hour of Day

There is a substantial increase in the amount of crashes during the late afternoon hours known as "rush hour", peaking around 4:00 PM. This is because there is increased traffic as people commute back from work. The number of crashes decreases significantly in the early morning hours, showing that there are fewer incidents when there are fewer vehicles on the road.

Number of Car Crashes by Type of Vehicle per Year

Below is a histogram of accidents by year, with each bar filled based on the top 3 types of vehicles involved in those accidents. These three categories are passenger vehicle, depicted in green, station wagon or sport utility vehicle, depicted in turquoise, and finally taxi, depicted in magenta. The x axis represents the year and y axis represents the number of accidents that occurred.

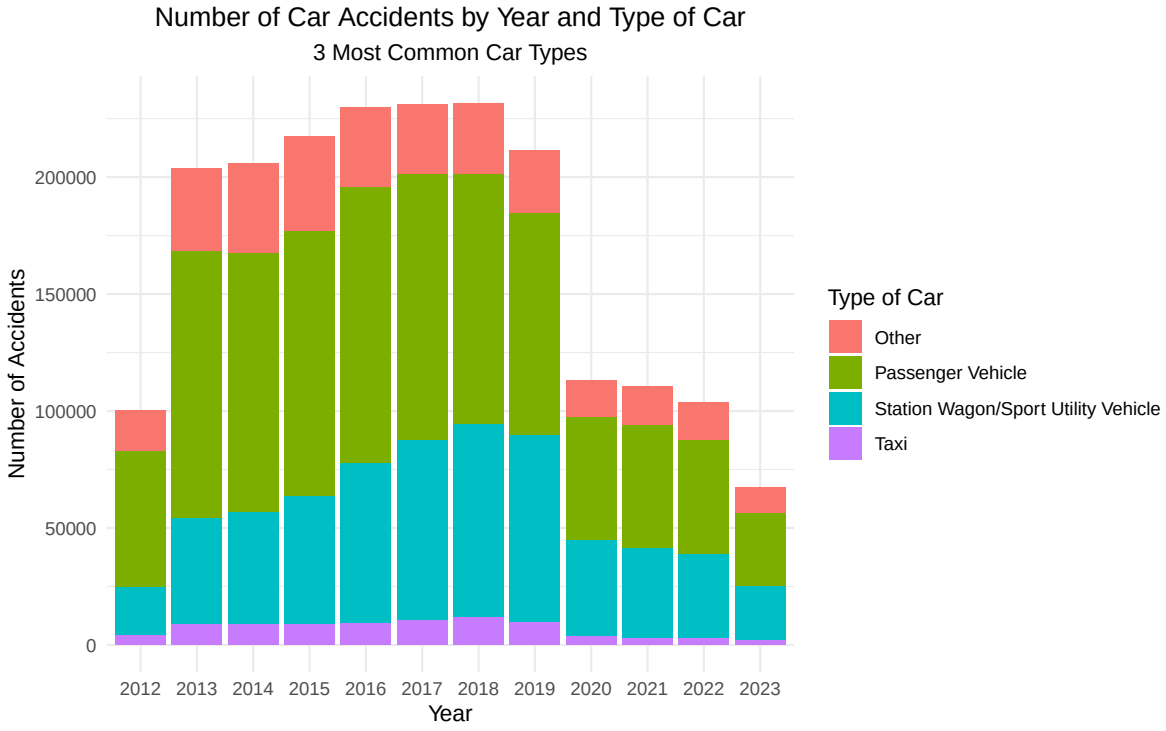


Figure 9: Number of Car Crashes by Type of Vehicle per Year

As we can see, there is a relatively large portion of each bar designated as 'Other', displaying the widespread variety of vehicles in accidents. Besides that, the next most common vehicle is a passenger vehicle. After that, the most prevalent was a station wagon or a sport utility vehicle. The last, much smaller category included in the plot is Taxi. This order remains the same throughout all years. Furthermore, we see a sharp decrease in the number of accidents in 2020, most likely due to the onset of COVID, but what's interesting to note is that accident numbers have not returned to what they once were post COVID. We can also see that the relative proportion of accidents with passenger vehicles, station wagons, and taxi's remain similar throughout the years.

Density of Casualties by Contributing Factor

This is a density plot visualizing total casualties for all accidents by contributing factor in our data set, where a casualty is either a person injured or a person killed. The plot visualizes the probability density function (pdf) of total casualties of car crashes in NYC. It shows the distribution of car crash casualties in NYC for 4 most common contributing factors. This plot gives us insight into how different contributing factors can influence the number of casualties.

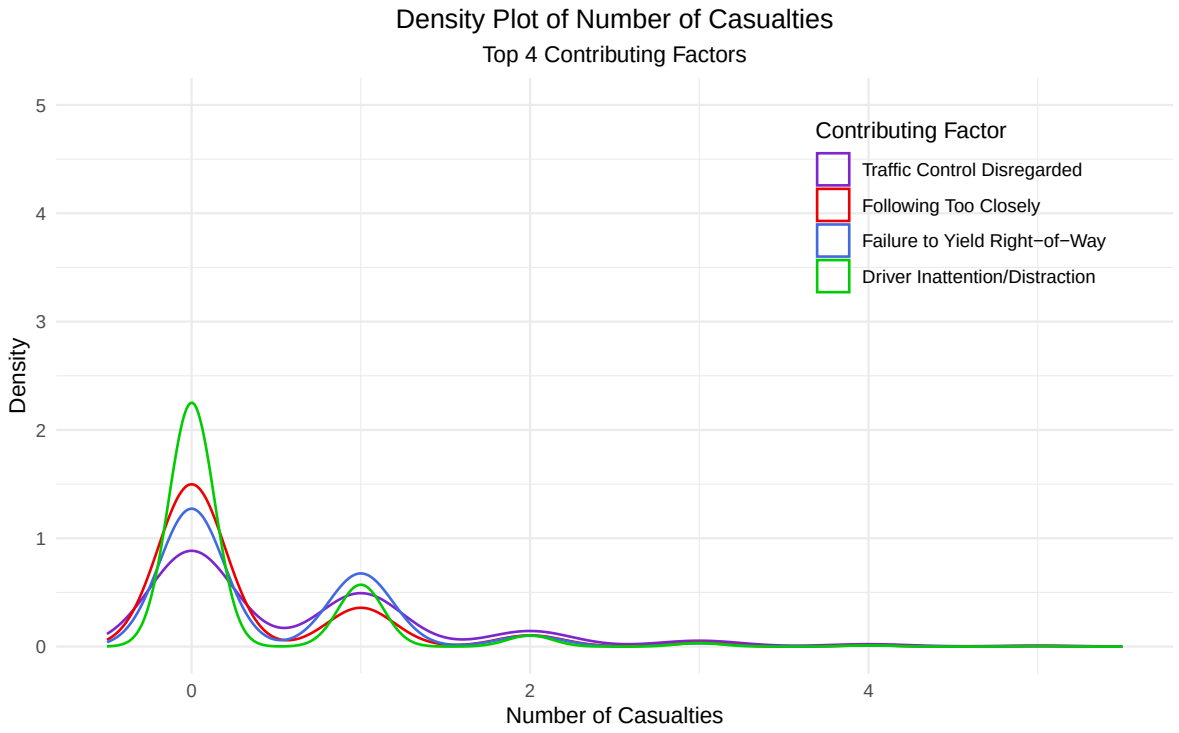


Figure 10: Number of Car Crashes by Type of Vehicle per Year

The plot shows that there is a large difference between low casualty (0 or 1) accidents caused by Failure to Yield Right of Way as compared to the rest of the contributing factors. However, for higher casualty incidents (greater than 2), this difference largely decreases and there is a relatively similar amount of high casualty accidents caused by Failure to Yield Right of Way as to caused by Driver Inattention and Distraction. This suggests that there is a much higher likelihood of a low casualty accident being caused by Failure to Yield Right of Way than a high casualty accident. Furthermore, accidents with 1 or more casualty are more likely to be caused by failure to yield, followed by driver inattention.

Number of Crashes by Weather Conditions

The Box-Dot plot below displays the number of crashes that occur for a given day for the various categories of weather conditions. The weather conditions are listed on the x axis, while the number of car accidents is on the y axis. Each blue dot represents a day - its x coordinate is the weather that day, and the y coordinate is the number of accidents that occurred that day. The boxes show the first and third quartiles, as well as the median number of car crashes for each weather. This plot, in addition to the primary data set, utilizes the secondary data set, Weather in New York City, New York from 2020-09-01 to 2022-12-31 [2].

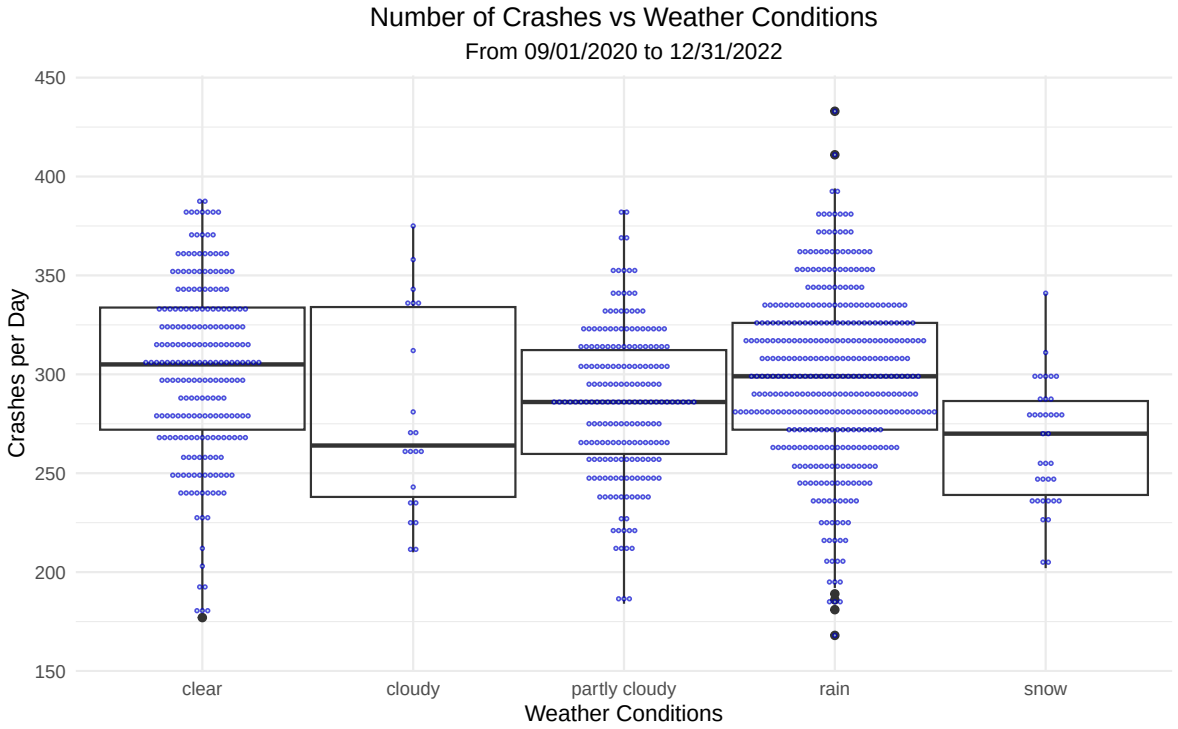
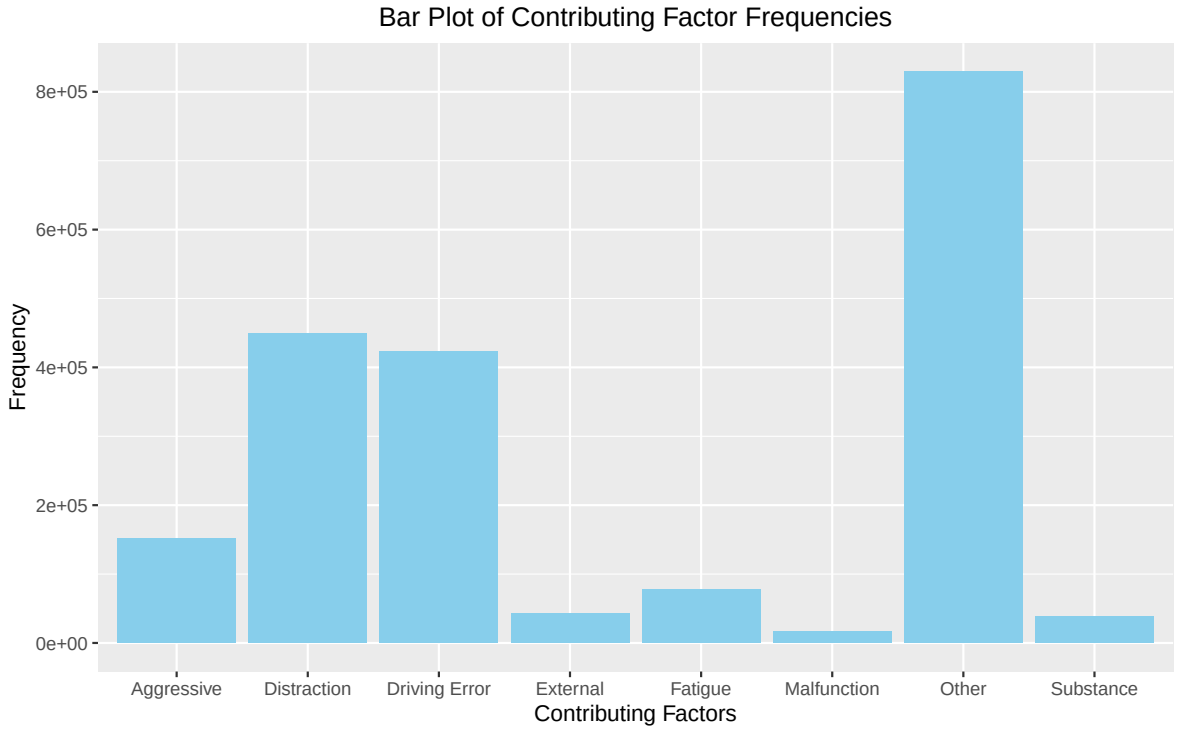


Figure 11: Number of Crashes by Weather Conditions

The plot shows that from 09/01/2020 to 12/31/2022, there was a majority of rainy days, followed by clear days, and partly cloudy days. This can be concluded by the number of blue dots on each section of the x axis. The median number of car accidents for a clear day is slightly higher than the number of car accidents on the rainy days. However, the rain box has two outliers, where the highest number of car accidents happened out of all days. What is quite surprising to note is that the median number of accidents on snowy days is lower than the median number of accidents on clear days, which may be the case because less people are out on the road during snowy weather.

Number of Car Crashes per Contributing Factors

For this plot, we decided to perform text mining on the Contributing Vehicle Factors. This column contained many different factors, and there was correlation between many of these factors, so in order to grasp a better picture of this variable, we decided to group them based on more broad overarching categories including: Aggressive, Distraction, Driving Error, External, Fatigue, Malfunction, Substance, and Other.



From the plot above, the "Other" category is the largest, which makes sense since the category also includes "Unspecified", and a lot of the contributing factors could not be determined. However, after that, "Distraction" is the biggest category, followed by "Driving Error". Other factors include "Aggressive", which includes behavior that is not an error or distraction but is what is commonly referred to as "road rage". Fatigue is also a big factor, followed by external. Close to this is also substance, which includes alcohol and drugs.

Resources

1. Primary Data Set: Motor Vehicle Collisions in New York City, New York
2. Secondary Data Set: Weather in New York City, New York from 2020-09-01 to 2022-12-31

3. Road Traffic Injuries - World Health Organization