Interpreting Causal Relationships between Contributing Factors and Motor Vehicle Accidents in the New York Metropolitan Area

Jian Ruan, Zach Yu, Carrie Hashimoto, Sophia Lyu (Undergraduate), Patrick Yee (Graduate)

March 1, 2023

1 Introduction.

The aim of this project is to analyze contributing factors for motor vehicle accidents and related injuries in the New York Metropolitan area. A better statistical understanding of these factors could help guide more intelligent decisions in the interest of public safety by preemptively identifying conditions under which accidents are more likely. These analyses would be of interest to officials in both the public (legislative, infrastructual) and private (insurance, engineering) sectors.

The primary dataset for this project is a list of motor vehicle accidents in New York City between 1 Januray 2013 and 31 December 2022 (MVCC). The data are collected by the New York Police Department and maintained by the City of New York [1] (1.97m rows $\times$ 29 cols) and contains date/time, vehicle type, location, and injury data for each raw observation. Since the MVCC set is complete population data (eg, contains every known car accident over the specified 10-year interval and is not a sample), the main focus of this study will be on regression and causality, not statistical inference.

2 Detailed Location Data.

To better understand the geographical distribution of the car crashes in NYC, each incident is first mapped as a blue dot (x = lattitude,y = longitude) based on the OpenStreetMap API. However, the distribution map alone is not sufficient to draw conclusion on the difference in occurances across different regions. Thus, a countour map is plotted to further demonstrate in which area a car crash is more likely to occur.

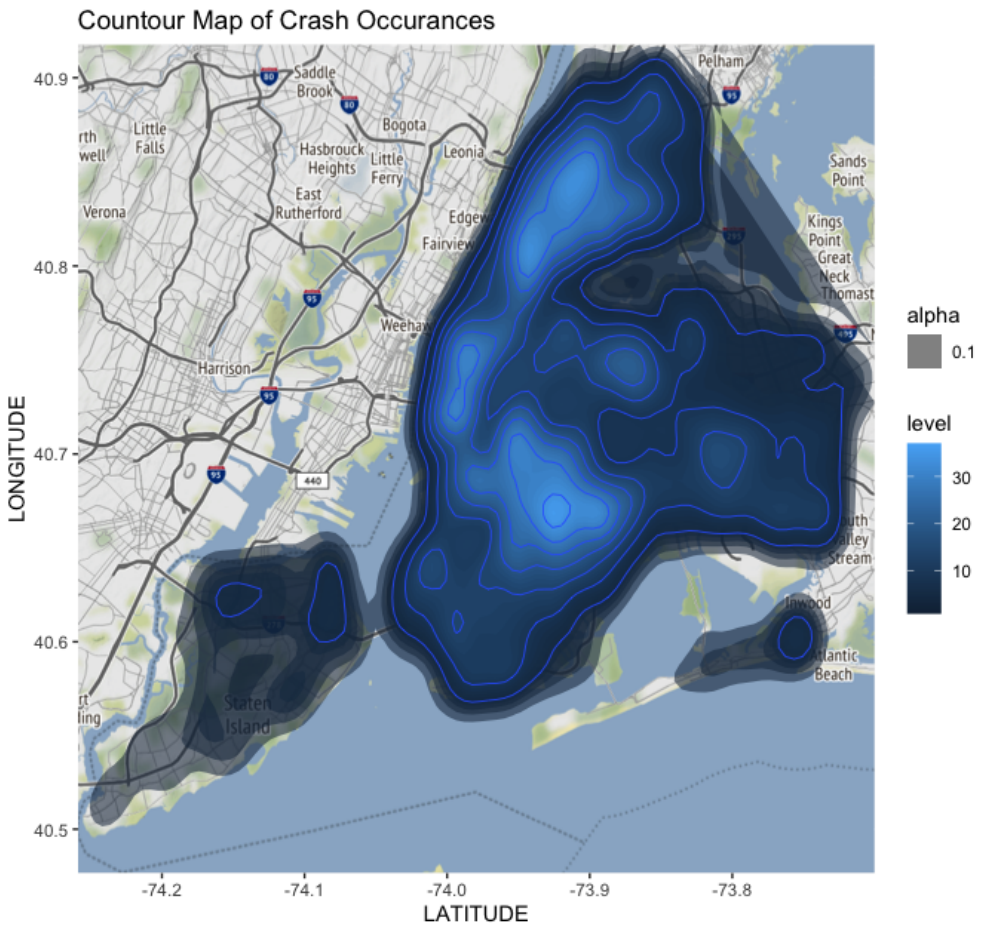

Figure 1: Heatmap of high-risk areas

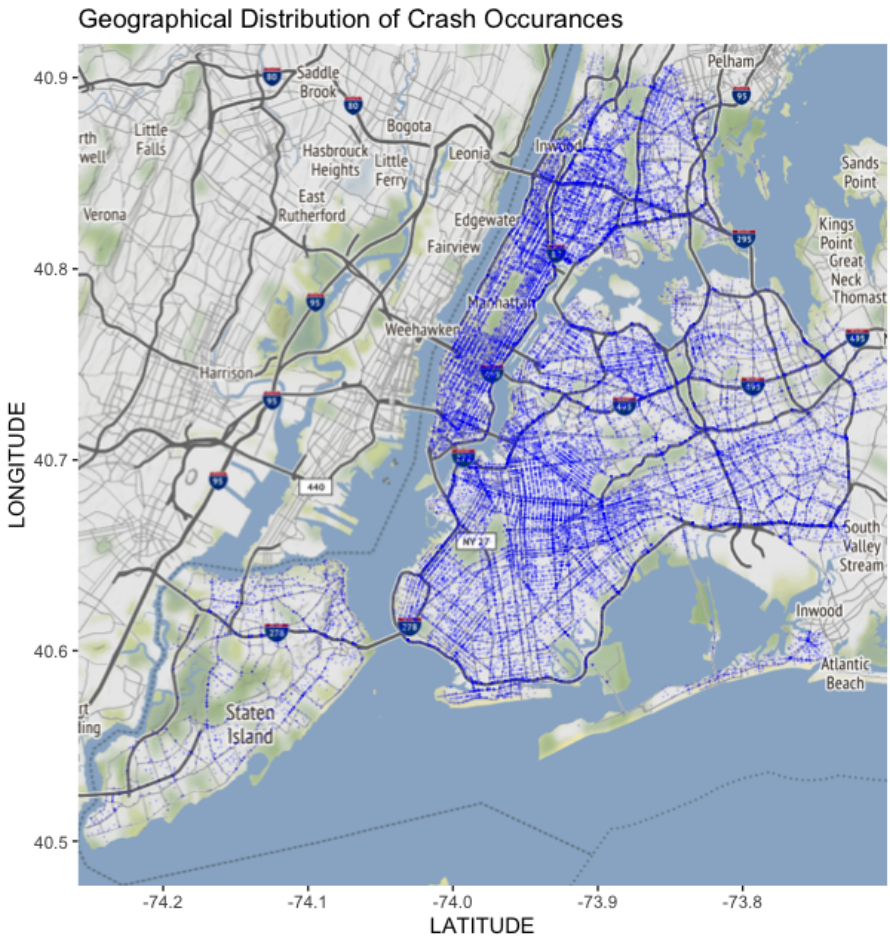

Figure 2: Exact locations of crashes

The lighter the blue color, the more frequent car crashes occur. Based on the borough map of NYC, Brooklyn and Manhattan are the top two areas for car crash incidents with a local maximum in the center of each borough. The result is reasonable based on city data where the density of population and GDP is the highest in Brooklyn and Manhattan.

3 Detailed Time Data

It was hypothesized that motor accidents are more likely to occur at busier times of day. 10,000 samples were drawn from the population data sorted into 15-minute blocks. An empirical distribution was then calculated and smoothed using native ggplot functionality.

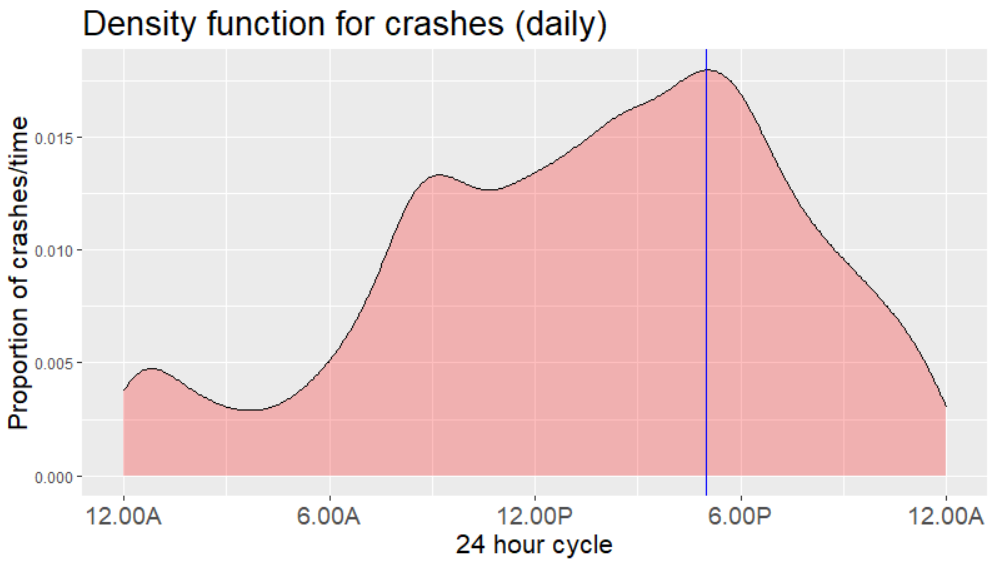

Figure 3: Density function for accident counts over a 24-hour period.

Crashes are most likely to happen at 5.00pm, where they are nearly 4 times more likely to occur than at night. This suggests that high traffic during rush hour might be a contributing factor. We can further explore crashes as a function of time on a larger scale by plotting the mean number of crashes/day/month, with a ribbon for the safest and least safe days for each entry.

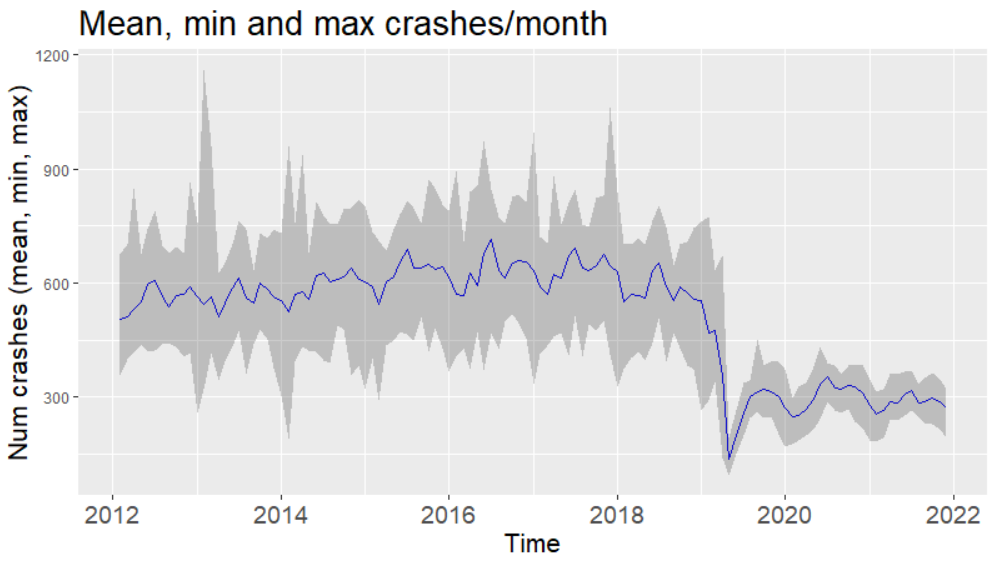

Figure 4: There is a sudden decrease in motor vehicle accidents coinciding with the start of the COVID-10 pandemic.

This plot corroborates our cumulative crash plot from a previous draft, in which it was hypothesized that the COVID-19 pandemic may have influences car crashes. In particular, a sudden decrease in accidents can be observed right at the onset of the pandemic.

4 Contributing factors.

It was taken into consideration what were the top contributing factors for New York crashes (excluding the unspecified causes). Many crash factors were accounted for so the top 5 contributing factors (excluding unspecified) were examined. Within the data of top 5 contributing crash factors, the proportion of vehicle type involved was implemented by color. Below is the bar graph of the percentage of crashes occurring with the top 5 crash factors and the type of vehicle taken into account.

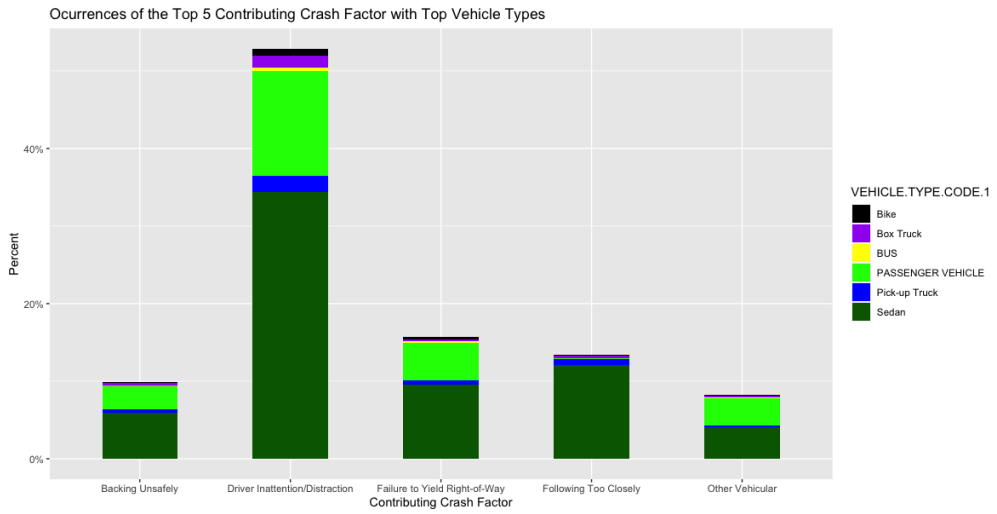

Figure 5: Barplot of contributing factors, accounting for vehicle type.

The bar graph shows that driver inattention/ distraction accounts significantly for the proportion of crashes in New York city. In addition, within each of the top 5 crash factors, sedan accounts for the most number of crashes, supporting the fact that sedan is involved in the most amount of crashes. We can further explore these high-risk contributing factors by visualizing raw relative counts. An array of pie charts show the relative frequency of each of the top ten contributing factors to car crashes by borough and number of fatalities. Feature engineering was performed to combine related contributing factors into new categories such as "Fatigued/drowsy/sleepy/unconscious," and entries with unknown borough were dropped.

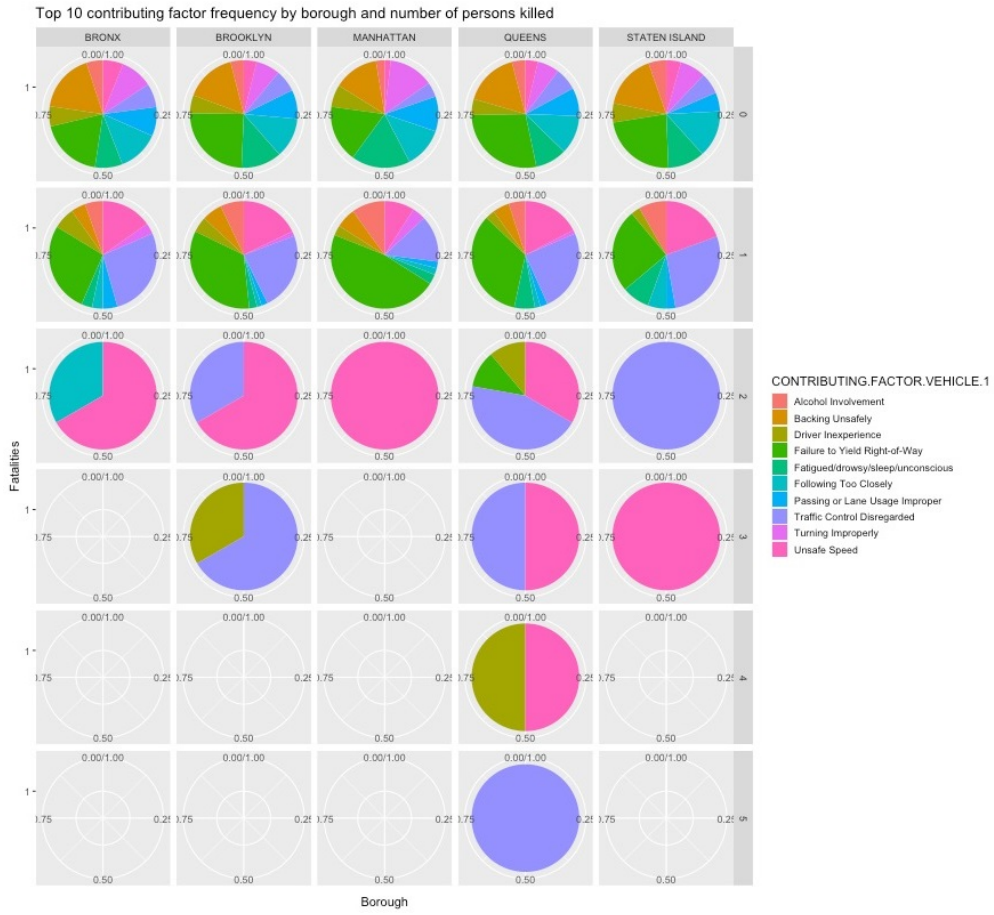

Figure 6: Pie charts of relative frequencies of contributing factors

The pie charts show most notably that traffic control disregard is a more common contributing factor to crashes where a person is killed than crashes where nobody is killed. Crashes with higher numbers of fatalities are less common, which may be why there are no crashes where over two people were killed for some given contributing factor and borough combinations. Traffic control disregard, unsafe speed, and driver inexperience accounted for the causes of all crashes with fatalities over three people, while backing unsafely was a relatively more common cause of accidents with no fatalities.

In previous analyses, unexpected interactions were discovered between number of injuries and seemingly unrelated factors such as borough and time of day. To identify additional "hidden" factors, a scatterplot measuring injuries against month and population density was created.

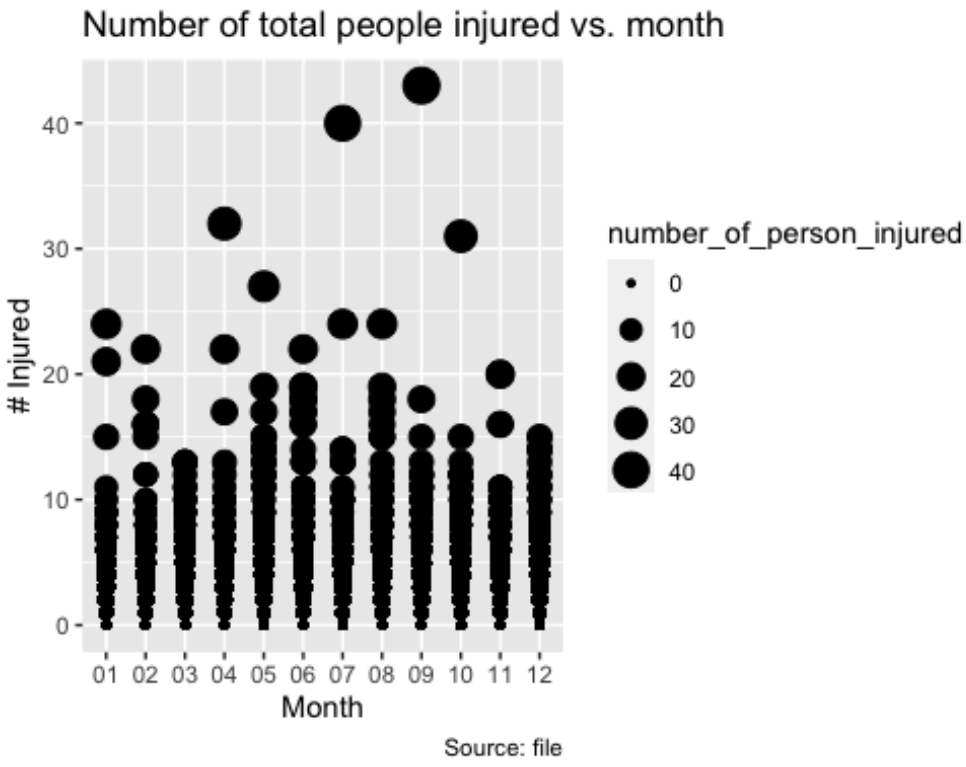

Figure 7: Number of people injured against month and contributing factor

Above, the x-axis was month and the y-axis represents the number of people injured in car accidents. Each plot corresponds to one accident, and the size of the plot is proportional to the population density. It was concluded that there is no significant correlation between the number of people injured and the month. There were some outliers in July and September, which shows, more people were injured in car accidents during summer compared to winter.

5 Next Steps.

In a previous draft, contributing factors and descriptive statistics were explored in relationship to car crashes in the New York City area. In this submission, these factors are analyzed in greater detail. In particular, advanced visualizations for location, time, and contributing factors are created and briefly discussed.

Our analyses reveal possible interactions between population density and risk of car crashes. There is also strong evidence that COVID-19 has an effect on accidents in this area. In future drafts, free text data will be analyzed to explore qualitative public attitude about COVID-19 and car accidents.