NYC Motor Vehicle Collision: An Exploratory Analysis

December 19, 2018

Authors: Micaela Flores (mrf444), Trevor Mitchell (tim225), Jason Li (yl2813), Laureano Nisenbaum (lvn218)

Abstract: In this project, we focused on exploring data from the NYPD concerning motor vehicle accidents that occur in the city during the period November 1st, 2017 to October 31st 2018. After cleaning the data, we observed the reasons for which these collisions occurred and bucketed them into more general groups. Using these, we explored the number of collisions, the top three reasons for collisions per borough, and the number of fatalities that resulted from the accidents. Our results indicated that Manhattan has the highest number of accidents per borough, that *Bad Driving*, *Impaired Driving*, and *External Factors* are the main causes of accidents, and that impaired driving causes the greatest number of fatalities across all boroughs.

1. Introduction and Motivation

Nearly 1.3 million people die in road crashes each year and, on average, there are 3,287 deaths a day due to motor vehicle collisions in the United States. Moreover, car accidents are not only responsible for deaths, but material damage as well. Since New York City is the most populous city in the US, we want to know what are the top three zip codes in each borough with the most accidents and what causes them. This exploratory analysis can aid and advise New York City law enforcement to prioritize their officers and resources. It would help local governments diagnose the major issues in these high accident-prone zip codes and offer potential solutions for the underlying causes.

2. Methodology

The raw data was taken from the NYPD Motor Vehicle Collisions Data Set, where each row in the original data included the date and time of the accident, the borough and zip code in which it occurred, location, and street and cross street, etc. We extracted only the features we deemed relevant and necessary to carry out this exploratory analysis: date, time, borough, zip code, number of persons injured/killed, number of pedestrians injured/killed, number of motorcyclists injured/killed, and contributing factor (the reason of the collision) and put it into a pandas dataframe. After dropping rows that included NaNs, we proceeded to bucket the original reasons for the collisions into more general groups for easier analysis. For example, the original reasons of *Eating/Drinking* and *Texting* were binned as *Distracted Driving*, *Alcohol Involvement* and *Prescription Medication* were binned as *Impaired Driving*. This was done via the sklearn preprocessing LabelEncoder, which takes categorical variables and encodes them into integers. This was then used to map the original reasons and our generated groups. Furthermore, the heat maps used in the "Main Causes for Collisions" section below were created with the use of geopandas package by utilizing the NYPD Motor Collision Data set and the zip code geo-spatial data from NYC Open Data.

3. Results

We began our analysis by plotting the total number of collisions per borough (Fig. 1). From the resulting plot, we noticed that the raw number of collisions per borough followed the population distribution, i.e., more collisions occur in more populated areas.

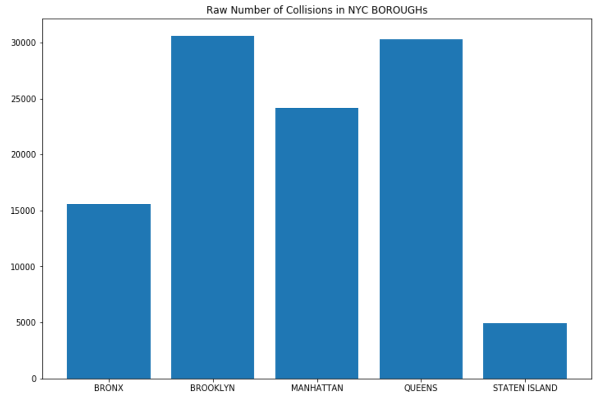

Figure 1. Raw number of Collisions in NYC Boroughs.

After this, we plotted the same data but per capita (Fig. 2). With this rescaling, it appeared Manhattan leads the ranking with the most number of accidents. This could be explained by the fact that even though Manhattan is not the most populated borough, it concentrates the greatest amount of touristic, economical, and employment activities. In other words, the number of residents in the Manhattan borough is probably not the best indicator to describe the amount of people that traffic this area everyday, leading to the greater possibility of motor vehicle collisions.

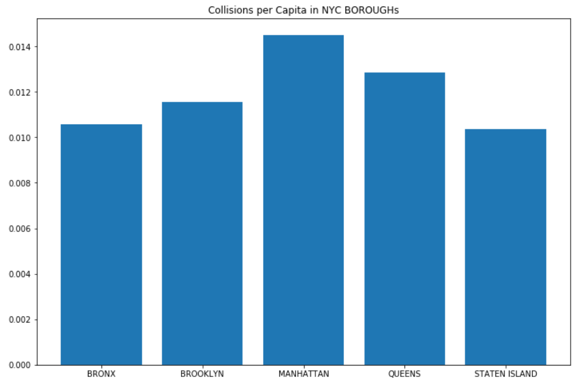

Figure 2. Number of Collisions per capita in NYC Boroughs.

Additionally, we also explored the effect of time on the amount of collisions. We picked month and day of the week as group criterion. The results are shown in Fig. 3 and Fig. 4 respectively.

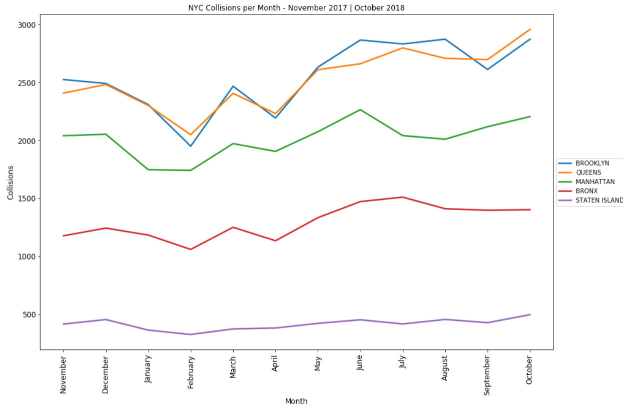

Figure 3. NYC Collisions by month. Nov. 2017 - Oct. 2018

From Fig. 3, we observe that the amount of collisions seem to peak during the summer months and are lower during the winter. This may signify that people are less prone to use their vehicles during the colder month of the year due to the poor driving conditions from snow or rain.

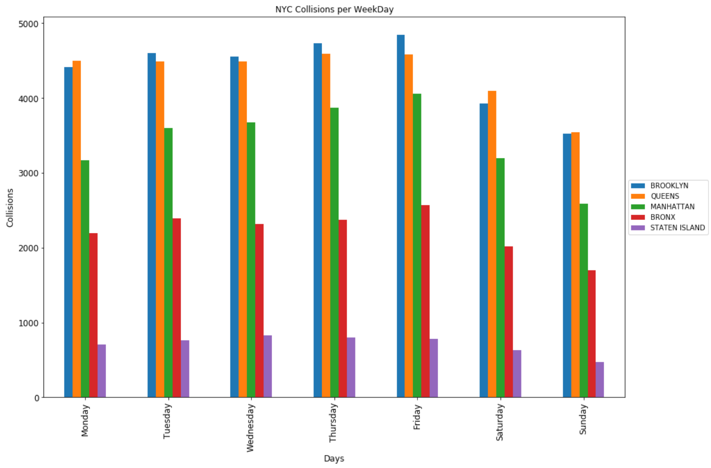

Figure 4. Number of Collisions by day of the week.

Fig. 4 tells us that the number of collisions tend to go down during the weekends. This results seem reasonable as one would expect more drivers on the street during workdays.

Main Causes for Collisions

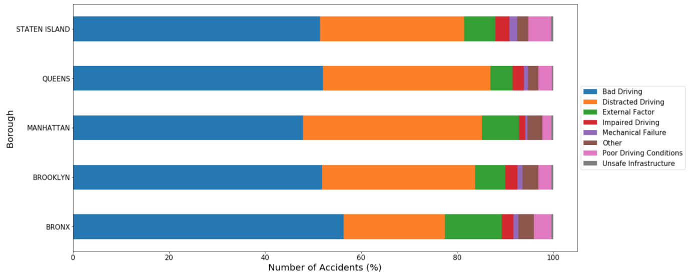

Figure 5. Main causes of collisions by borough.

Fig. 5 tells us that the top three factors are Bad Driving, Distracted Driving and External Factors across all boroughs.

In order to identify the top three causes of the collisions within each borough, we plotted a stacked histogram. We noticed that on at least

50% of the collisions from each borough is from *Bad Driving*; 25% is from *Distracted Driving*, and 15% from *External Factors* (outside car distraction, reaction to uninvolved vehicle, etc.).
.

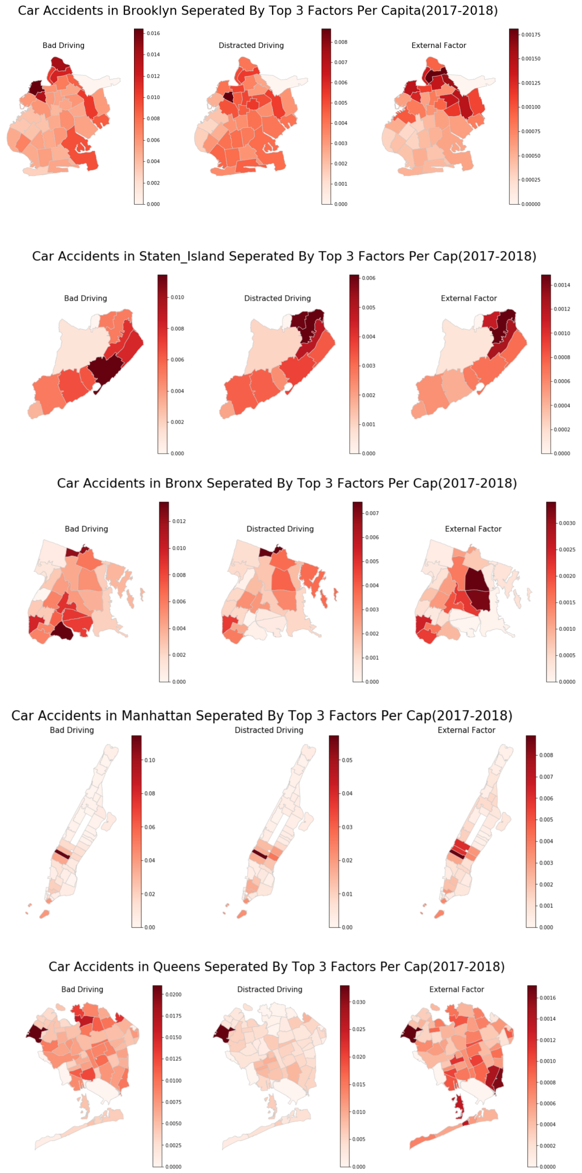

Figure 6. Heatmaps of collisions per Capita by ZIP Code.

Fig. 6 shows heatmaps of car accidents in each borough separated by the top three factors per capita.

After we identified the top three causes for each borough from the stacked bar plot, heatmaps were used to better visualize within each zip code the concentration of the number of collisions due to each factor per borough. For each borough, we tend to see that in the most concentrated (highest number of accidents per capita) area for each collision reason is around the outer areas of the boroughs. This make senses because lot of the traffic comes from people commuting to or from the city. We can see this extreme example in Manhattan, where the most concentrated area is the Lincoln Tunnel.

Deaths Per Borough

In order to observe trends in vehicle accident fatalities in NYC, we had to process the data related specifically to collision fatalities. There were four columns of particular interest. These were “Number of Persons Killed”, “Number of Pedestrians Killed”, “Number of Cyclist Killed”, and “Number of Motorist Killed”. In order to verify the integrity of the raw data set, we checked the number of records by filtering the data in two ways. One query was when the “Number of Persons Killed” is greater than zero. The second query was where the “Number of Pedestrians Killed” or “Number of Cyclist Killed” or the “Number of Motorist Killed” was greater than zero. This was done to verify that the number of persons killed equaled the sum of the number of pedestrians, cyclists, and motorists killed. There were 100 records brought back in each scenario.

Next, we simply filtered the data on the “Number of Persons Killed” that were greater than zero. This provided us with a fatalities dataset to work with moving forward. The dataset was then grouped by borough. Only the relevant columns “Number of Persons Killed” and “Reason” were used from the data frame. After that, we simply retrieved each group by borough and the resulting data frames were then grouped by reason. For each borough being grouped by reason, we retrieved the number of deaths and plotted the results.

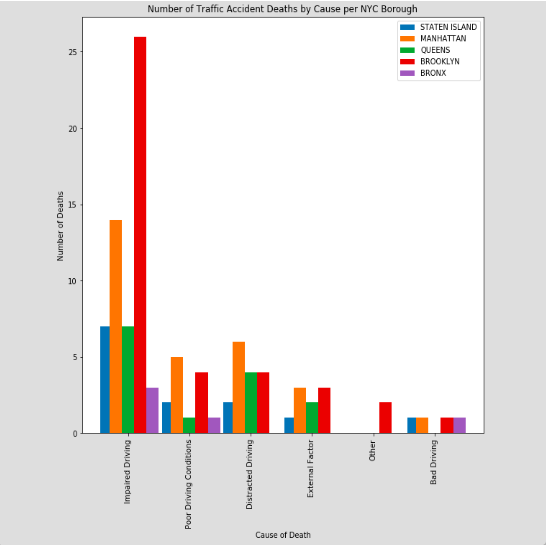

Figure 7. Fatalities Barplots by borough and main cause.

Fig. 7 shows that impaired driving is the biggest culprit of death across all 5 boroughs.

4. Discussion

In summary, Manhattan has the highest concentration of motor vehicle collisions per capita than the other boroughs. This could be due to the greatest amount of pedestrian and motor vehicle activity, which increases the likelihood of accidents.

In addition, the heatmap provides the NYPD with a comprehensive visualization of where the accidents are concentrated. This model can be further improved by mapping the longitude and the latitude of each accident in order to get a more precise area of where accidents occur.

Lastly, the highest amount of fatalities are due to impaired driving and the number of deaths in Brooklyn was significantly higher than the other boroughs. It would be interesting to do an analysis on whether this was because of a particularly bad car accident from November 1st, 2017 to October 31st 2018 or if this is a recurring trend year after year.

5. Conclusion

The methodology we used was satisfactory as it allowed us to identify those geographic areas where accidents are more likely to happen. Since most accidents occur near the edge of boroughs, the NYPD should focus on areas along the border of the boroughs where commuter highways and tunnels are located.

Possible future analysis could include factoring in the time of day. Additionally, a more granular time analysis could be performed where weekdays could be binned by hour or periods of 4 hours, and checked at what time each weekday the collisions peak. By doing so, we could better inform the NYPD when to deploy officers for traffic controls as a preventive measure. Moreover, other factors such as type of vehicle involved in the accident or weather factors could also be included in the analysis.

In the future, we recommend that the NYPD dedicate more resources and training to checking for impaired drivers in Brooklyn. Additional policies to prevent impaired driving should also be enacted.

References:

"NYC Open Data (2017-2018). NYPD Motor Vehicle Collisions Data Set."
https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95"

"Kaggle. Exploring Vehicle Collisions in New York.
https://www.kaggle.com/adhok93/exploratory-data-analysis"