NYC & Toronto Motor Vehicle Collision Analysis

Matthew Cihlar, Nomin Ganzorig, Samhita Vinay, Zi Wang

Table of contents

Introduction

The dataset that we selected contains entries of motor vehicle crashes in New York City from 2012 to 2023, and our supporting dataset was something similar but for Toronto in order to get an idea of similar data across different cities. Our motivation for selecting this data was determining what the main factors behind crashes in urban areas are and where officials tasked with preventing these accidents should funnel their resources. Our datasets are filled with precise geographical information about crash time, borough, zip code, latitude, and longitude in order to pinpoint the location of these crashes to determine where they occur most frequently. They also include the number of people injured and killed, number of pedestrians injured and killed, contributing factors of driver ability such as inattention and distracted driving, and vehicle type information to determine, after finding out where they occur, why they might have occurred. The Toronto dataset also had information about the visibility conditions at the time of the crash, which we believe should have an impact on whether or not a crash may occur.

Overall, this report seeks to explore the links between the frequency and lethality of these incidents and the various details we have outlined. We also know that COVID was a seriously disruptive event when it came to cars on the road, as fewer people were traveling, especially for work which in an urban environment is likely notable. We looked for evidence in our dataset for any abnormalities that might indicate that COVID or work from home was a significant factor in the number of crashes that occurred in these big cities. We also believe that factors such as the time of day or vehicle type involved may have a tangible impact on the result of the accidents, and much of our data analysis is built on probing the links between these phenomena. We hypothesized that crashes would be higher during the summer, as more people would be traveling, during the afternoon, as people return from work, and we also hypothesized that the main cause of these crashes was distracted driving. We made these hypotheses because we drive cars on a daily basis, so we are familiar with some conditions in which it is most unsafe to drive. We wanted to use these datasets as a point for confirming or clarifying our initial assumptions in order to make us and those we present to safer drivers.

Data Analysis & Visualization

We wanted to investigate the distribution of crashes per day for each borough. This plot demonstrates the number of daily crashes by boroughs. Each dot in the plot represents the number of crashes that took place

in the corresponding borough on a particular day. The darker the region is, the more observations fall into that level. We can see that Brooklyn have the most crashes per day on average while Staten Island has the least. This is likely due to the population density of each region, and also with the concentration of commuters and offices. This data falls in line with what one would expect from the more active and populous regions. One other noteworthy aspect is that while the average crashes per day for Bronx and Manhattan are similar, Bronx has a much smaller variance and range compared to Manhattan, which are on the same scale as Brooklyn and Queens. Population density and the number of cars on the road is likely extremely correlated with the number of crashes that would occur, and so this data should not be surprising and falls in line with what we had assumed would be the case. We also see that this is the case for Toronto, and both of our datasets seem to agree that population density is largely the most important factor when determining how many crashes might occur.

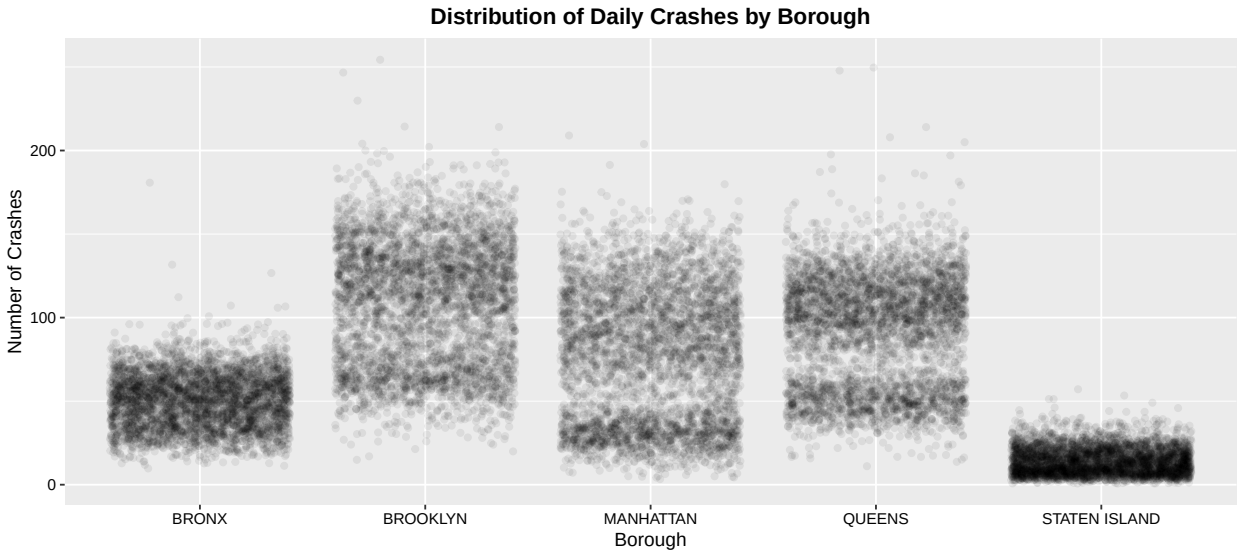

Figure 1. Crashes per day distributed for each of the 5 boroughs of NYC, with Bronx & Staten Island being the most concentrated and Manhattan, Brooklyn, & Queens being more sparse.

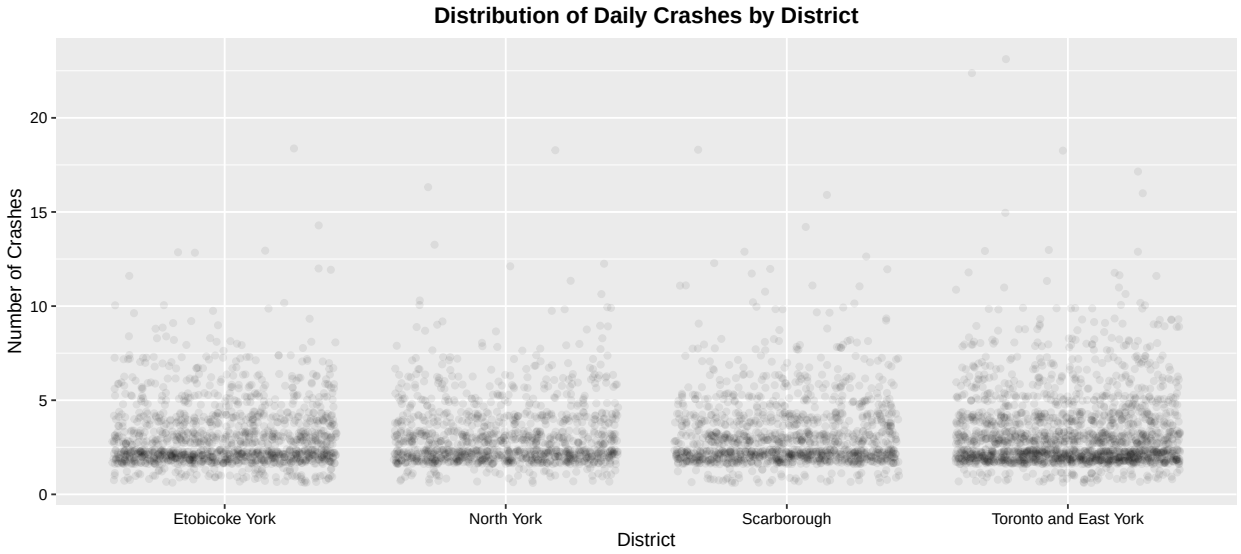

Figure 2. Crashes per day distributed for each of the 5 districts of Toronto, with Toronto & East York having the most crashes. Overall, crashes were more sparse than those in NYC.

We also wanted to investigate location at a more granular level, so we investigated zip code of the New York City crash data. More specifically, we looked at the change over time in crashes from 2012 to 2023 in different zip codes. Because there were too many zip codes, we captured the top 10 changes and noticed a decrease in the number of crashes for all the changes in the corresponding zip codes. This data will be helpful when thinking about mapping the crashes. From this data, we can better understand more granularly where these crashes occur and why they might appear in these places. Because these zip codes are the most densely populated, we also figured that COVID likely had a role in their decline, along with the increase in remote tech work and other work from home setups. The data suggests that the fewer cars on the road, the fewer crashes occur, which is of course what one would expect. Furthermore, it is important to note which zip codes had a more dramatic variation from before, during, and after the COVID period. Again, because we are looking at densely populated zip codes, the evidence suggests that people are simply moving around far less in these populated areas, and that they are not returning to the peak travelling that they were engaged in prior to COVID. This enables us to conclude that there is likely a correlation between the number of the crashes and how prolific work from home is, and by tracking work from home (which is likely common in these densely populated areas) we could track how many crashes are expected to occur.

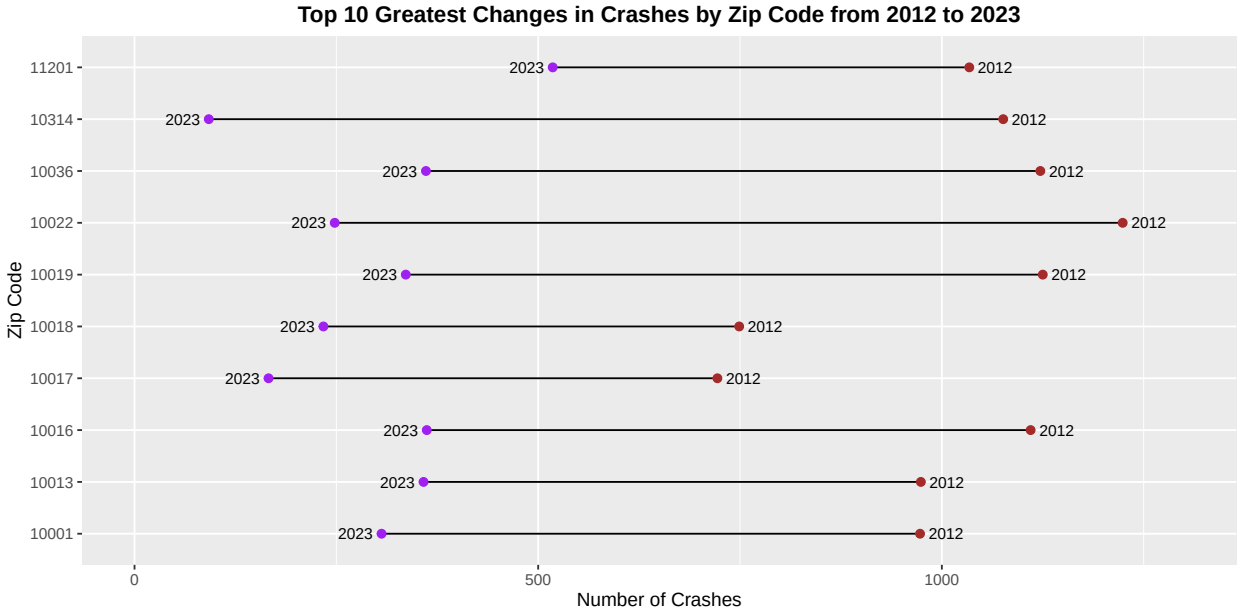

Figure 2. The top 10 changes in the number of crashes by zip code from 2012 to 2023 (negative change for all 10 zip codes) in NYC. The brown dots represent the number of crashes in 2012, while the purple dots represent the number of crashes in 2023.

Since the dataset contains the year in which crashes happened, so it is valuable to analyze the number of crashes per year as a trend line. The overall number of crashes in NYC seemed to peak during 2018, but has gone down since. As a result, the line of best fit is negatively sloping. Similar to the trend for each zip code, if we look at the overall trend within the city, we can see a similar pattern. We can then determine that the most likely cause of this effect is probably COVID and the increase in Work-From-Home jobs which obviously require less commute than traditional office jobs. Again, we see a similar pattern with the previous few plots. It is no surprise that COVID has had such an outsized impact on the number of crashes because it is simply the case fewer people are traveling, and therefore fewer people are getting into car crashes. We see something similar for Toronto. Notably, the downtrend continues, and has not returned to its previous levels. We do not suspect that COVID has suddenly caused people to become better drivers, but rather than there are still fewer cars on the road. Data about the number of cars on the road at any given time would have allowed us to draw conclusions about the skills of the drivers on the road before, during, and after COVID, which could have been particularly interesting.

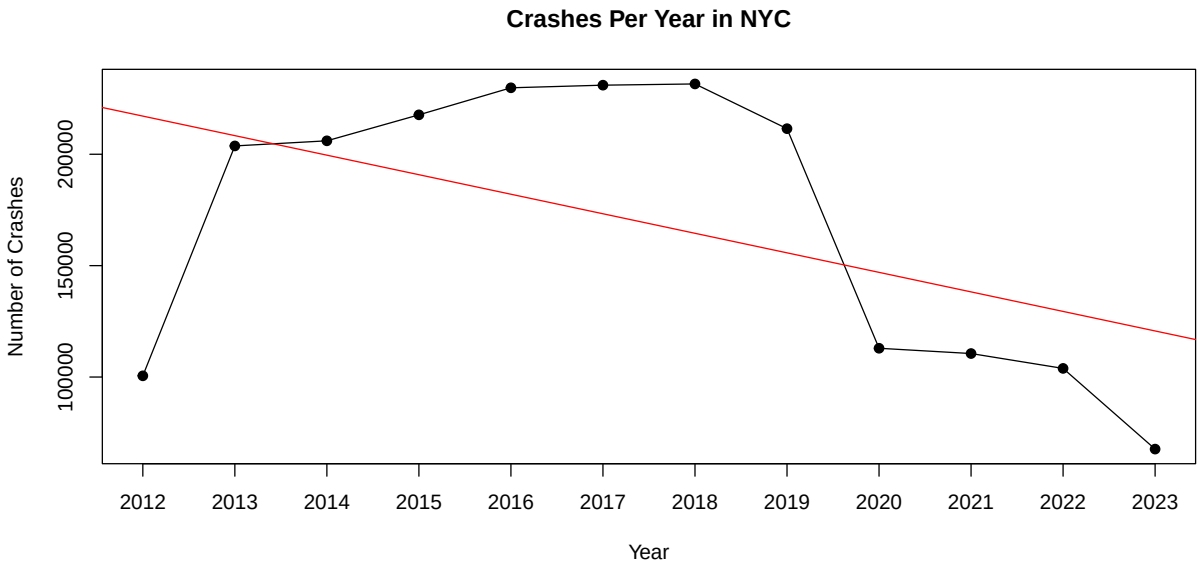

Figure 3. The number of crashes for every year in New York City from 2012-2023 visualized using a line along with the line of best fit (in red).

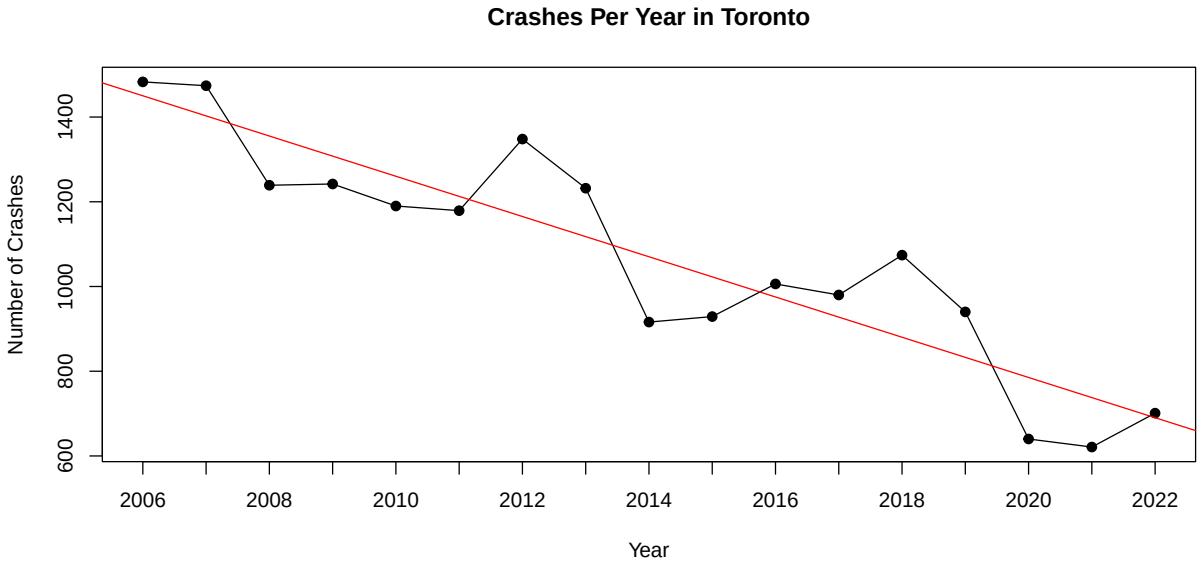

Figure 4. The number of crashes for every year in Toronto from 2006-2022 visualized using a line along with the line of best fit (in red).

We wanted to look more specifically at how the number of crashes trends through out the time horizon. To do so, we create a plot showing the number of crashes that take place for each month within the time span of the dataset. We can see that the number of crashes peaks from late 2016 to around late 2019, and then drastically declines at the start of 2020 potentially due to the advent of COVID-19, and then remains at a moderately low level. This is also the case for the Toronto dataset, though slightly less pronounced. We expect that the incomplete nature of the Toronto dataset had something to do with this, however, because we can still clearly extract this pattern from our incomplete dataset, we are more confident that this is a significantly important pattern and has a high degree of explanatory power.

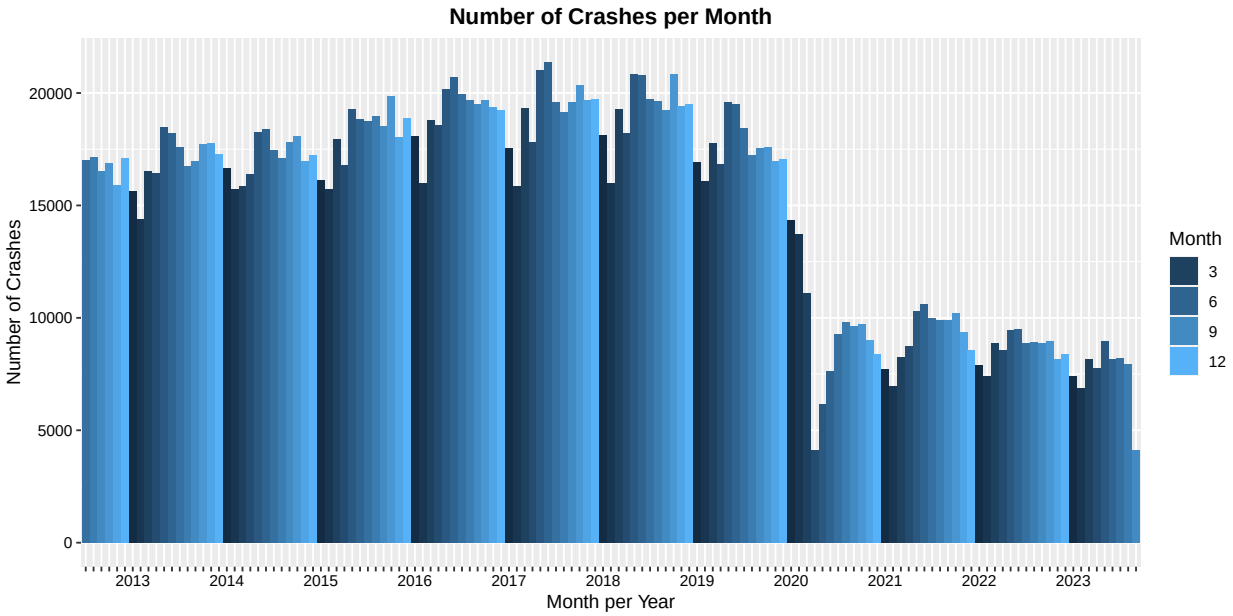

Figure 5. Bars with colors going from dark blue (January) to light blue (December), showing the distribution of crashes per month across time from 2012-2023 in NYC.

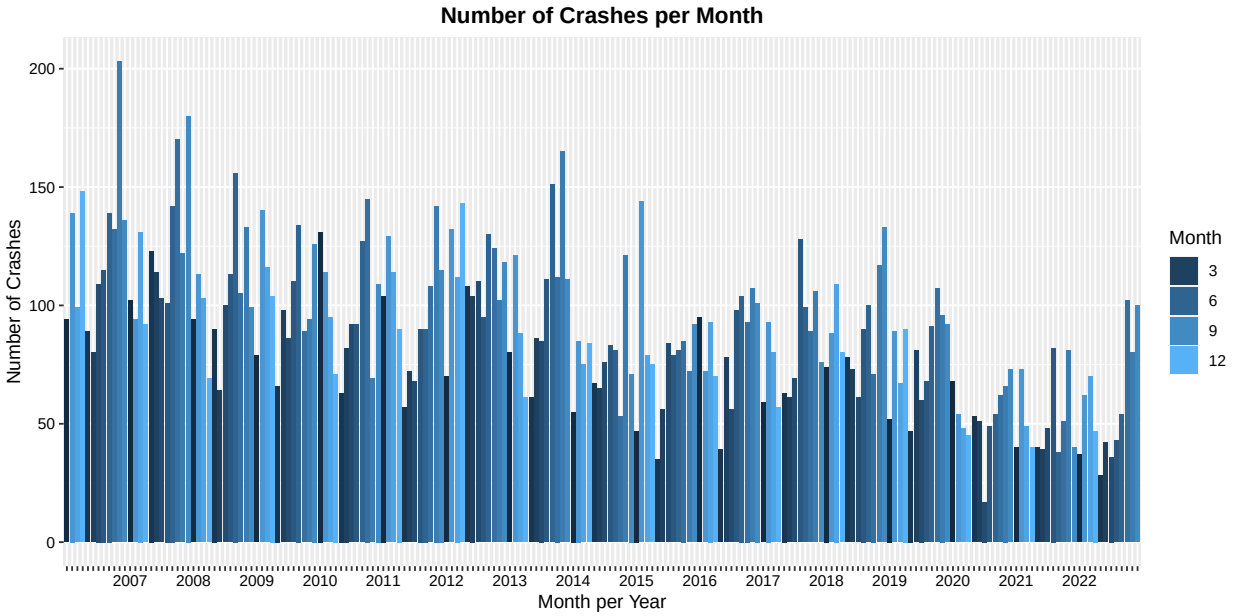

Figure 6. Bars with colors going from dark blue (January) to light blue (December), showing the distribution of crashes per month across time from 2006-2022 in Toronto.

We then wanted to combine the comparisons by borough and by time. Therefore, we made a stacked bar graph displaying these factors. We found that for all boroughs, crashes peak at around hour 15, while they are the lowest at around hour 4. This makes intuitive sense, since we would expect 3pm to be among the busiest times, as this is when people are getting off work and need to travel between their homes and the offices. Further, we would expect the early morning to have relatively few crashes, since relatively few people would be on the road during these times. This falls into line with much of our previous data. Moreover, Brooklyn and Queens had the highest number of crashes overall, while Staten Island had the lowest number of crashes, which again we saw earlier. This is more evidence for the work from home trend having an

outsized effect on our data. Because fewer people are traveling to and from work, these peaks are smoothed, and hence there is less of a "rush hour" risk and fewer crashes at this time. Because we are able to determine that this rush hour period greatly increases the number of crashes on the road, we would naturally expect the reduction of the rush hour effect from the reduction of people returning from work, and as such this data provides further evidence for our hypotheses.

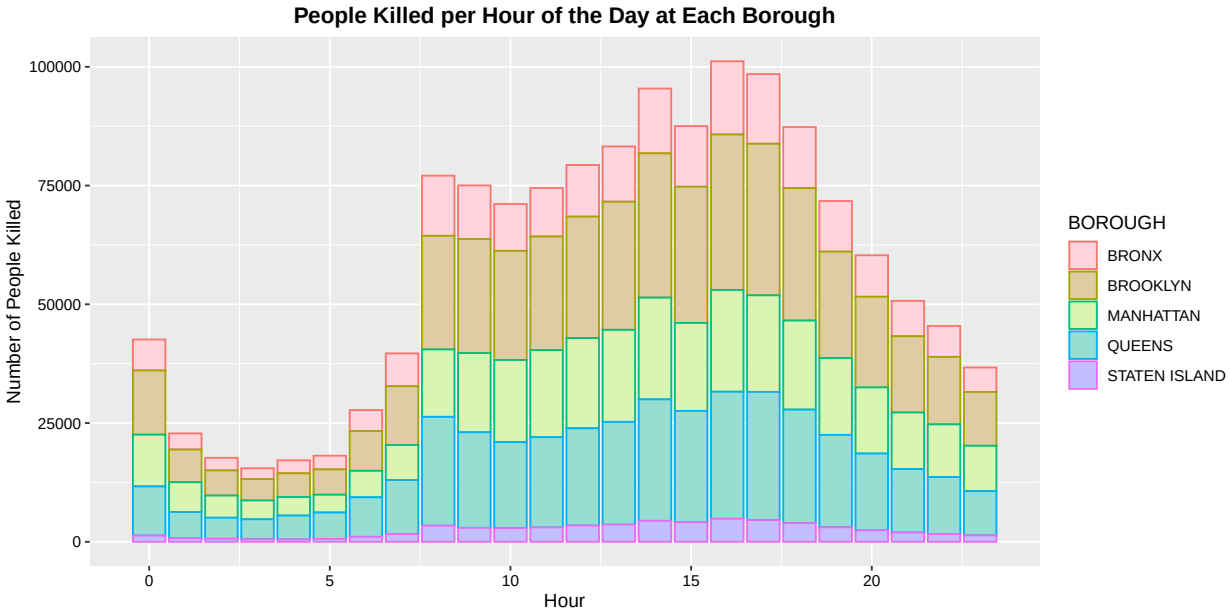

Figure 7. A stacked bar plot showing the number of crashes per hour of day for each borough of NYC. Pink represents the Bronx, brown represents Brooklyn, green represents Manhattan, blue represents Queens, and purple represents Staten Island.

Text mining is another technique that can be used in R to extract important information from data. In the New York City dataset, an interesting aspect that could be investigated through text mining was the type of road that the crash occurred on. Of course, crashes vary depending on factors such as road size, location, and busyness, and we wanted to investigate this by taking a look at whether the road was a street, avenue, ramp, etc. As shown in the plot below, avenues were where the most crashes took place, followed by street, boulevard, parkway, expressway, and road. This somewhat contradicts the notion that most crashes take place on large roads, since avenues are often mid-sized. Streets were not too far behind avenues, and this supports the conjecture that most crashes happened on mid-sized roads. Expressways and parkways, which are larger roads, had far fewer crashes than the aforementioned mid-sized roads. Perhaps this occurs because drivers do not take mid-sized roads as seriously as highways, causing them to forget certain safety precautions while driving. Much more analysis would have to be done when analyzing why certain roads are more prone to crashes than others, as this could have something to do with road quality, where in the city they are located, etc. that were not in this dataset.

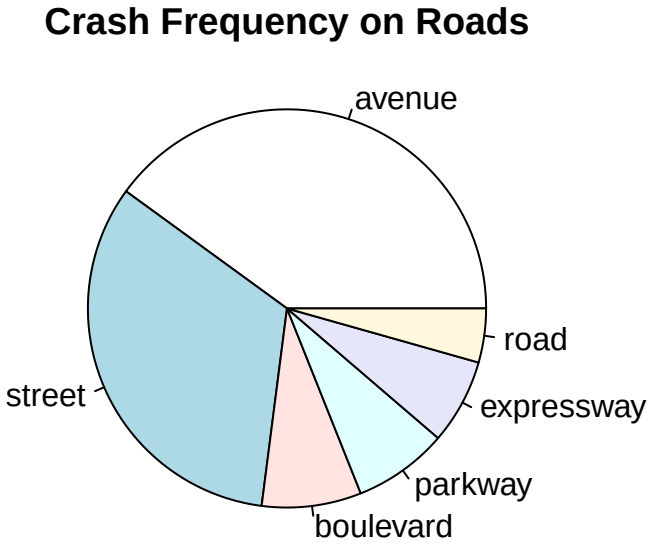

Figure 8. A pie chart displaying the relative frequencies of crashes occurring on different types of roads of NYC, obtained by text mining of the data.

We can also investigate the underlying causes of crashes individually, perhaps giving us insight into what specifically causes crashes on different roads. In this plot, we can view which vehicle types are involved in the most amount of crashes. Largely, this data falls in line with the proportion of vehicle types that are already on the road, as sedans are the most common vehicle, and so forth. We would need more data to check whether or not there is a link between a certain type of car and crashes, as from this we haven't found much evidence that one is skewed more than the other.

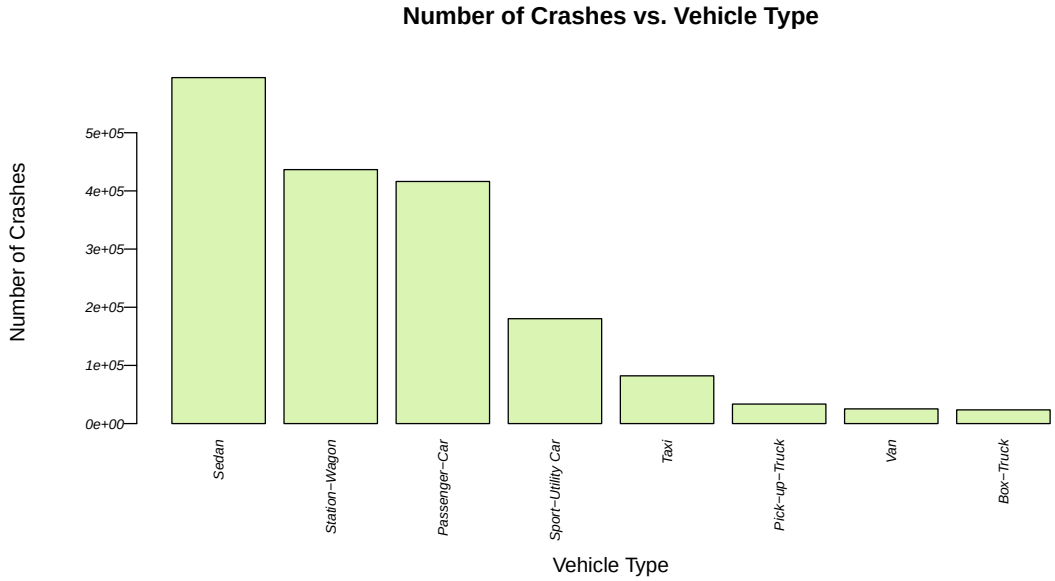

Figure 9. Barplot demonstrating the number of crashes per day for each vehicle type, sorted from highest number of crashes (left) to the lowest number of crashes (right).

There were also behaviors listed as being the cause of each crash. The majority of these come down to user error, so to speak, as driver inattention and other poor driving habits were far and away the most common cause of a crash. Notably, mechanical failure or any other external factor did not play a large role in these

crashes, which some may expect, but the cause of crashes almost universally comes down to human error in one way or another. Whether it be texting or any other distraction, crashes seem to increasingly be a result of poor driving over anything else. This could also explain the rush hour phenomenon. When people are stuck in traffic, it is not uncommon for them to take out their phones and check their emails or send some texts, and as we see from this data, this is incredibly harmful and an extremely important factor in whether or not a crash occurs. Many of these separate factors are correlated and help to explain each other.

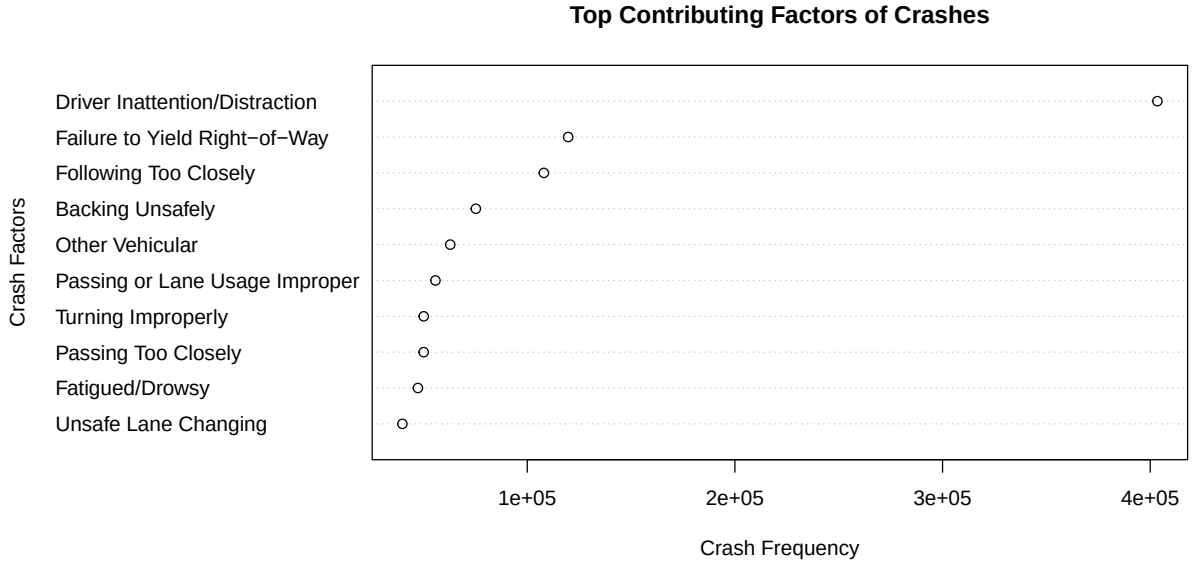

Figure 10. A dotplot displaying the crash frequency for each factor causing the crash, sorted from lowest frequency (bottom) to highest frequency (top).

From this graph, we can examine how the visibility conditions at the time of the crash affect the likelihood of a crash occurring. We would expect darker and foggier atmospheres to have an increased likelihood of crashes, and we see this reflected in the data. Rainier conditions also have a large cluster, which makes sense. However, we were surprised that snowier conditions did not have even more of a representation than they already do, given that Toronto is in Canada. Again, we wish we could have seen a more complete Toronto data set to further highlight these disparities, but however, we can still extract this theme from the dataset, as we would expect this general trends to scale with the dataset. This still falls in line with what we would expect from our hypotheses.

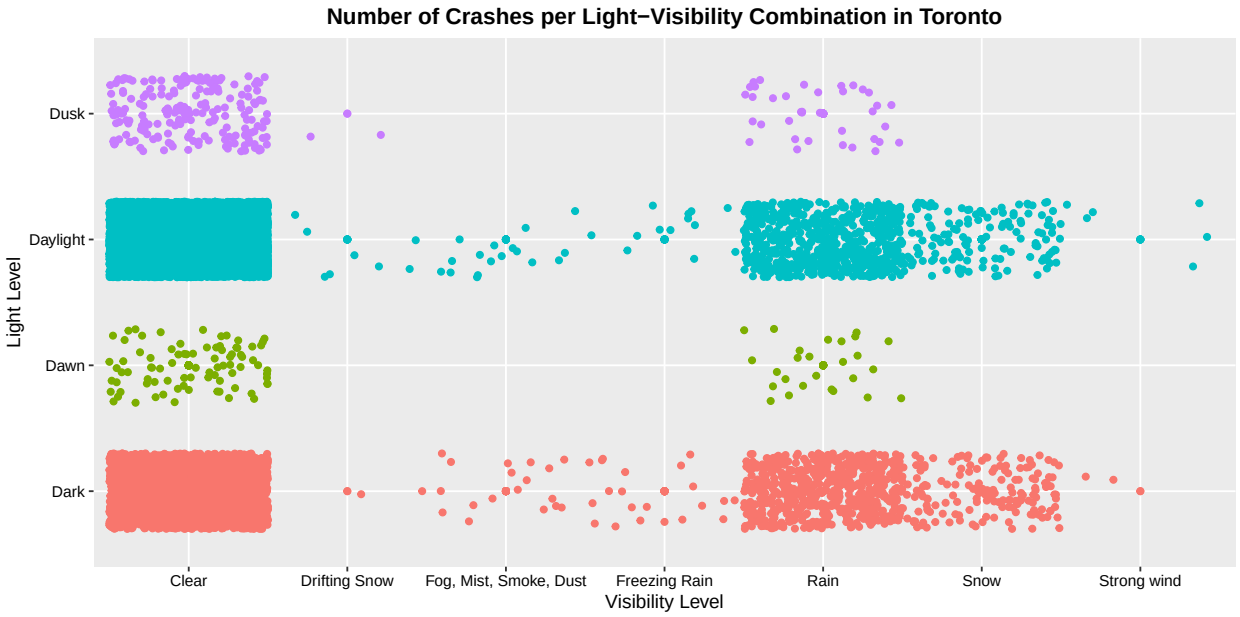

Figure 11. This plot shows the distribution of crashes in Toronto by light and visibility conditions. More dots centered around a particular condition pair demonstrates a greater density of crashes occurring in that pair of conditions.

Also, we can analyze the injuries per month that are caused by the top four contributing factors of crashes, respectively. Each of the factors corresponds to a different color in the plot. What we can derive from the plot is that the number of crashes caused by "Driver Inattention/Distraction" per month tends to vary largely by time, while the number of crashes caused by "Backing Unsafely" per month varies the least. The general trends of the number of crashes per month caused by each of the factors are all similar, which means that proportion of crashes caused by one factor with respect to any other factors is relatively stable throughout the time span. This provides good evidence that the proportion of causes of the underlying crashes are not random and the few at the top are indeed more harmful than the others.

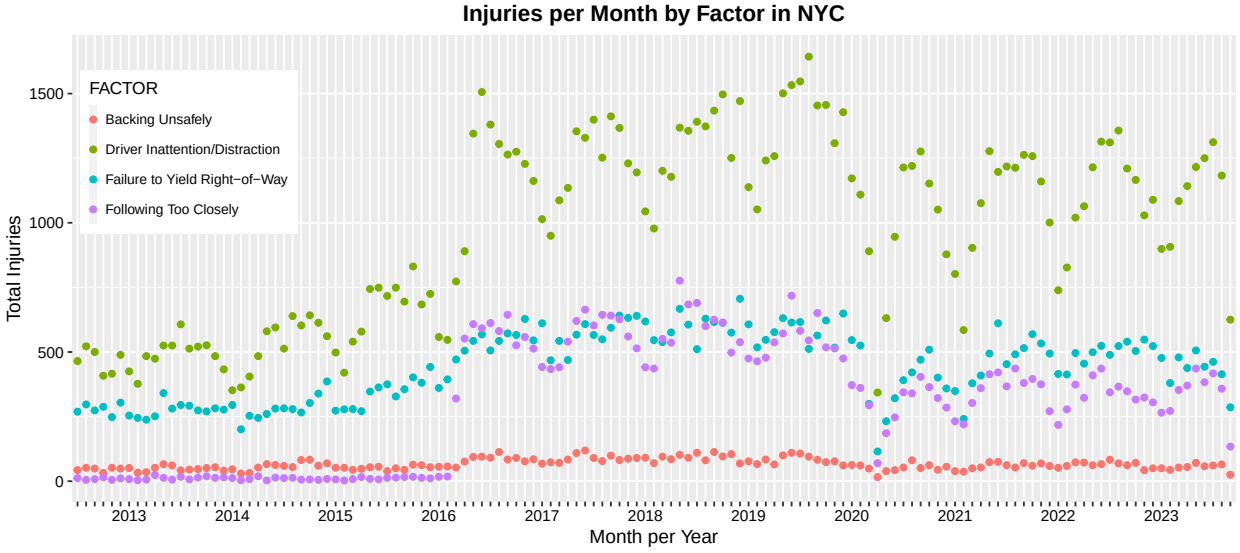

Figure 12. This plot shows the distribution of causes of injuries to the injuries themselves over time. Green dots represent driver inattention/distraction, teal dots represent failure to yield right-of-way, purple dots represent following too closely, and salmon dots represent backing unsafely.

Overall, in order to visualize the most impactful factors that our analysis produced, we generated killer plots for both the New York City and Toronto datasets. The plots showcase the top three contributors to each of five factors of the New York City and Toronto datasets. The plot resembles a traffic light, which makes sense given that our datasets are about vehicle crashes. Moreover, the red circles are generally bigger than the yellow and green circles, and this is because the factors in red circles are the highest contributors to crashes, followed by yellow, then green. The circles are different sizes according to their relative percentages, symbolizing the relative impact of each specific factor. The plot for New York City demonstrates that most crashes happen due to Driver Inattention, in Brooklyn, and with Sedans. Also, the plot for Toronto shows that the most crashes happen in automobiles/sedans, in the district of Toronto & East York, and due to not yielding. *Note: in the presentation, there were Shiny sliders and dropdowns that allowed the user to focus in on different factors and toggle between the New York City and Toronto plots respectively.*

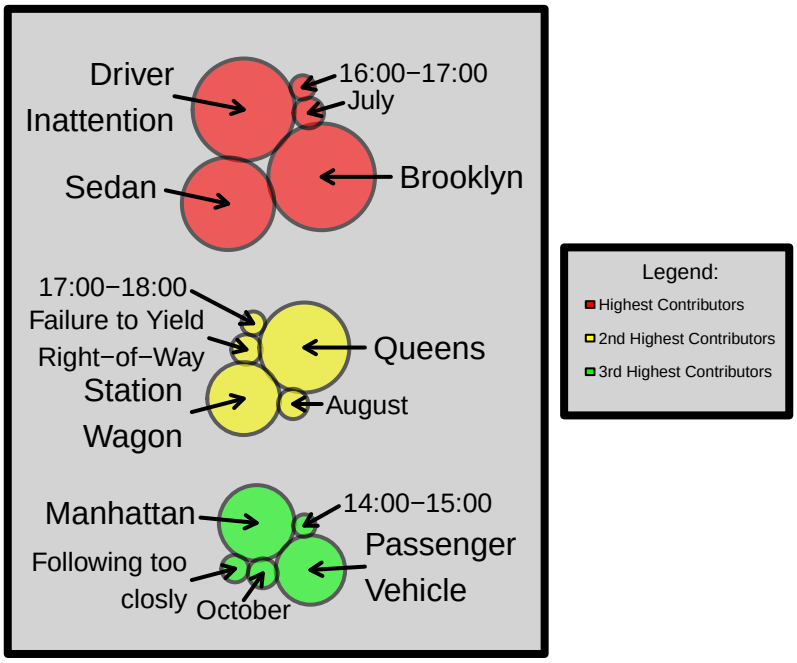

Figure 13. This plot shows the relative frequencies of the top three contributors of each of five factors for crashes in NYC. The red circles are the largest contributors, the yellow circles are the second largest contributors, and the green circles are the third largest contributors. The sizes of the circles are proportional to their relative impacts on crashes in NYC.

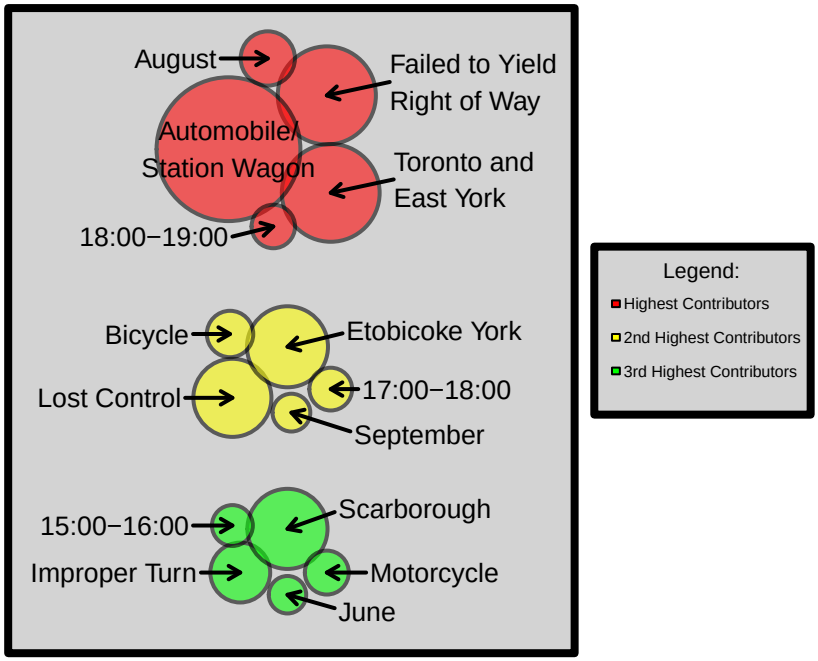

Figure 14. This plot shows the relative frequencies of the top three contributors of each of five factors for crashes in Toronto. The red circles are the largest contributors, the yellow circles are the second largest contributors, and the green circles are the third largest contributors. The sizes of the circles are proportional to their relative impacts on crashes in Toronto.

These plots all highlight various aspects of our data that give us a better understanding of the anatomy of crashes in New York City and Toronto. The time of day, borough, vehicle type, and more are all factors of a crash that come together in various ways to form our dataset. We have determined that the time of day certainly has an influence on the number of crashes that occur and also in how deadly they might be. Further, we have deduced that most crashes are a result of driver error, such as distracted driving or disobeying safety laws, and not the result of mechanical failure or other external factors.

Modeling

For our modeling, we chose to investigate the statistical link between the time of the day and the frequency of crashes. The model regresses the total number of crashes happening every quarter on a categorical variable that represents the time of a day. We categorize the 24 hours of a day into six time periods, namely 00:00 to 04:00 (late night), 04:00 to 08:00 (early morning), 08:00 to 12:00 (morning), 12:00 to 16:00 (afternoon), 16:00 to 20:00 (early evening), and 20:00 to 24:00 (evening). Our assumptions for the model include:

1) Linear relationship between time periods and crashes

2) Crashes in different time periods are independent

3) Variability of the number of crashes are similar across time

4) No linear relationship between explanatory variables

5) No correlation between errors and explanatory variables

6) Time periods specification is appropriate.

Based on the model assumptions, we want to test if there's statistical evidence of certain time periods having more (or less) crashes taking place than the others.

Table 1:

Regression Data for each Time Group

Groups	Estimate	Std. Error	t value	Pr(>|t|)
afternoon (12-16)	30,711.312	1,550.699	19.805	1.452e-34
early evening (16-20)	1,698.750	2,193.020	0.775	4.406e-01
early morning (4-8)	-20,619.000	2,193.020	-9.402	5.007e-15
evening (20-24)	-13,129.688	2,193.020	-5.987	4.301e-08
late night (0-4)	-21,356.812	2,193.020	-9.739	9.984e-16
morning (8-12)	-4,180.875	2,193.020	-1.906	5.978e-02

This table summarizes the results for the linear regression on crashes during each time of day. This shows the estimate, standard error, t-value, and the probability of being greater than the t-value for each time period: late night, early morning, morning, afternoon, early evening, and evening.

The regression model sets the period "afternoon" as the default category. Hence, all comparisons are between "afternoon" and the other periods. As the summary chart shows, there's statistically significant evidence that the number of crashes happening during "early morning", "evening", and "late night" are greatly lower than the "afternoon" period, as their corresponding p-values are infinitely close to zero. The difference between the numbers of crashes happening during "early evening" and during "afternoon" is quite insignificant, and the difference between "morning" and "afternoon" is more significant but is still rejected on a 5% significance level. Overall, the result shows that crashes happening from 20:00 to 08:00 the next day (when the traffic is typically less busy) are less than crashes happening from 08:00 to 20:00 (when the traffic is typically more busy).

Conclusion

As we have seen, there is strong statistical evidence for the time of day playing a large role in the frequency of crashes during the day. Further, we have seen that crashes are deadlier at different times of day, such as the very early morning, which we assume is the result of many different factors affecting driving at 3 AM. We have not found any evidence for the type of vehicle being a significant factor in the likelihood that the vehicle will be involved in a crash, but more statistical investigation against the population of vehicles in the city would be necessary to draw conclusions. Further, we have found that COVID and work from home policies have more than likely played a role in reducing the number of crashes that occur every year, as fewer cars are on the road nowadays as compared to years where the total number of crashes was much greater. More work will be needed to investigate whether or not a person on the road is more likely to be involved in a crash before or after COVID, as we do not have data on the total population of motorists between the two time frames. It would be useful to determine whether drivers relative skills were better or worse during COVID, and which types of people were on the roads during each time. This would have required several more datasets that we did not have access to, though it would have certainly added a number of interesting factors to consider. Perhaps we can conduct this type of analysis in the future. Overall, the trend of COVID having an extreme impact on the number of total crashes in NYC, and the fact that the reason that these crashes occur is almost entirely based on user error and driver failure should allow any officials to make educated decisions regarding where to send resources to mitigate these crashes. We suggest more driver education and heightened caution as more and more individuals return from working at home and the streets become more filled.

References

City of New York. (2023). *Motor Vehicle Collisions - Crashes* [Data set]. https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes

City of Toronto (2023). Motor Vehicle Collisions involving Killed or Seriously Injured Persons [Data set]. https://open.toronto.ca/dataset/motor-vehicle-collisions-involving-killed-or-seriously-injured-persons/

Fell, J. C., Freedman, M., Page, J. F., Bellis, E. S., Scheifflee, T. G., Hendricks, S. L., Steinberg, G. V., &; Lee, K. C. (1999). *Background*. The Relative Frequency of Unsafe Driving Acts in Serious Traffic Crashes. https://one.nhtsa.gov/people/injury/research/udashortrpt/background.html