Are Accidents Really Accidents?

Analyzing Motor Vehicle Accidents in NYC

DATASCI200 – Project 2
April 2024
Yanling Liu, Elizabeth (Lily) Maio, Diego Moss

Github repo link: https://github.com/thefivemayos/mids-datasci200-project2.git

Introduction

Last year approximately 40,990 people died in motor vehicle crashes in the US (NHTSA-1). Every year we see tens of thousands of people die in motor vehicle crashes and we call them tragic accidents. But when we say 'accident' colloquially we are implying something random and almost unavoidable. But it may be the case that accidents "are the result of larger systemic forces shaped by corporations and governments that intersect to create vulnerabilities in our environment; vulnerabilities that we don't all share equally (Singer, 2022)." This is the premise of the book 'There are no accidents' by Jesse Singer and the inspiration for this project.

Oftentimes when we hear about an accidental death we tend to think of the individual in that accident and potentially their risky behavior that contributed to the event. For example, a pedestrian struck and killed by a car might have been looking at their phone while crossing the street. When we blame the individual or assume an accident was random and unavoidable, we miss opportunities to ask, is this really an accident or can we change something about the situation to improve outcomes? As a society we can't necessarily control if that pedestrian stares at their phone when crossing the street. But we can control the speed limit, the visibility and distribution of designated pedestrian crosswalks, etc (Singer, 2022).

The goal of this research project was to evaluate the motor vehicle accidents dataset for NYC to determine if there are any patterns in the data to indicate social or built infrastructure gaps leading to accidents/deaths that may be preventable and or unevenly born by vulnerable groups.

We seek to understand the following questions:

1. Which borough has the highest occurrences of injuries and death from vehicle collisions? (including pedestrian, cyclist, persons overall)
2. What vehicle type and contribution factor is more common to cause deaths in a collision? What about collisions?
3. What's the correlation between characteristics analyzed above (i.e. borough, vehicle type, contribution factor to collision etc.) in the data and likelihood of death in a collision? What traits contribute to deaths in a collision if any?
4. Is there any correlation between median income of a zip code and number of deaths from collisions?

Source Data

- **NYC Motor Vehicle Incidents:** https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data

The main dataset for this analysis came from NYC Open Data. This data repository is managed by the Open Data Team under the NYC Office of Technology and Innovation (OTI). The mission of NYC Open Data is to make data available and encourage "the use of Open Data both within the government and throughout NYC" (NYCOD-1).

The dataset contains motor vehicle collision incidents recorded by the New York Police Department (NYPD) from 2012 through the present; over two-million records. The dataset is managed by NYC OpenData and automatically updated daily with new motor vehicle collision incidents and associated information (NYCOD-2).

Data Structure/Cleaning

Each record has 29 variables of information which include date and time of crash, location as zip, borough, lat and long, as well as number of injuries or deaths and the number of contributing vehicles for each crash incident.

We made the following minor edits to clean/format the data:

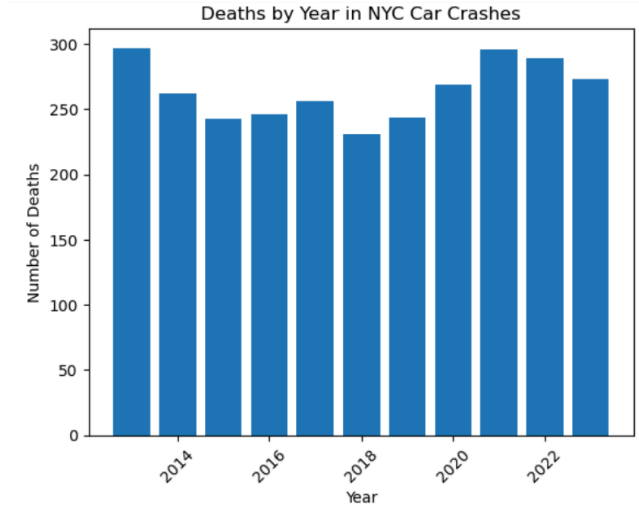
- Dataset Timeframe/Filter Out Partial Years
    - Full years from 2013-2023
- View Column Headings/Fix Formatting
    - Remove spaces, special characters, all lower
- Clean Zip Code format
    - Remove float

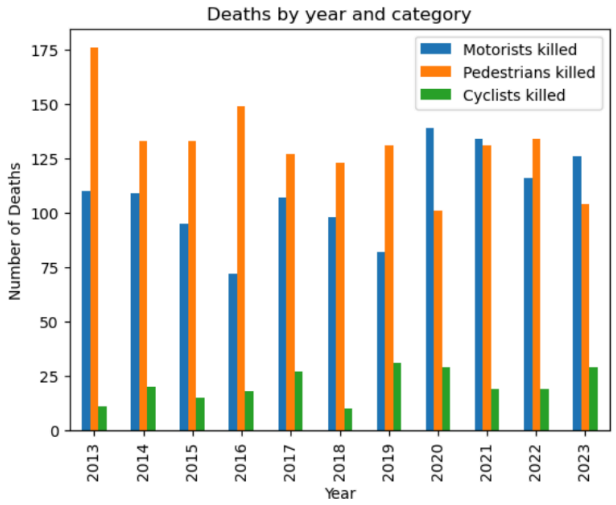
We performed a Non-null and Null Analysis on the variables in our dataset and found that there were a large number of records, ~ 600,00, with missing zip code and borough information, which is critical for our analysis. Much of this subset, however, did have latitude and longitude information and we were able to use Google query data to fill in the gaps, reducing our nulls from 600,000 to ~180,000.

The remaining 180,000 records do not have any latitude/longitude or zip information. This is about 9% of the dataset. There is a description about the location in the dataset, sometimes even cross streets, but there are over 8,000 unique descriptions, which makes this data difficult to clean/format in the timeframe of this project. Grouping certain keywords, we were able to see that over a quarter of these records involved a crash on a major highway, a bridge or a tunnel. Another quarter were accidents that took place 'off street'. And the remainder of the accidents did take place on the street, but their unique location identifiers were too numerous to process.

Initial Exploration

After cleaning the data we began by evaluating the death statistics and comparing the results to outside sources.





The histogram on the left shows total traffic deaths (motorists, pedestrians, cyclists) for each complete year in our dataset. Initially we were shocked by the relatively low total annual traffic deaths for a city of 8 million (0.003% death rate). However, data from the NYCDOH from 2000-2013 indicates total deaths around 300 people/year, the same rough order of magnitude as what we are seeing in the data (NYCDOH).

There appears to be a general downward trend in the total deaths/year. However, there is an increase around 2020-2021. This is consistent with the national trend reported by the National Highway Traffic Safety Administration (NHTSA-2). It was reported that 2021 saw a 16 year high in traffic fatalities(NHTSA). For NYC this was closer to a 10 year high in traffic fatalities.

The histogram on the right shows the deaths by category. You can see that the jump in deaths in 2020 came from motorists. Pedestrian deaths actually dropped dramatically, which may be the result of covid lockdowns. Motorist deaths may have gone up due to several factors such as decreased attention to our roadways/infrastructure as the country was dealing with the pandemic.

Pedestrian deaths have a declining trend with 2023 at the lowest annual rate in our dataset. However, motorist deaths are trending upwards, as are cyclist deaths. Following the pandemic, increased motorist deaths may be the result of more miles traveled on the road as people got out and traveled following lockdowns. Our findings are also consistent with data reported from other sources. According to Bloomberg, 2023 was an all time record low in pedestrian deaths for NYC in 114 years of record keeping. However, motorist and cyclist deaths are trending upwards (Surico, 2024). For cyclists part of the increased fatalities is due to the increase in ebikes on the road with 23 of the 30 cyclists killed in 2023 associated with ebikes (Bloomberg).

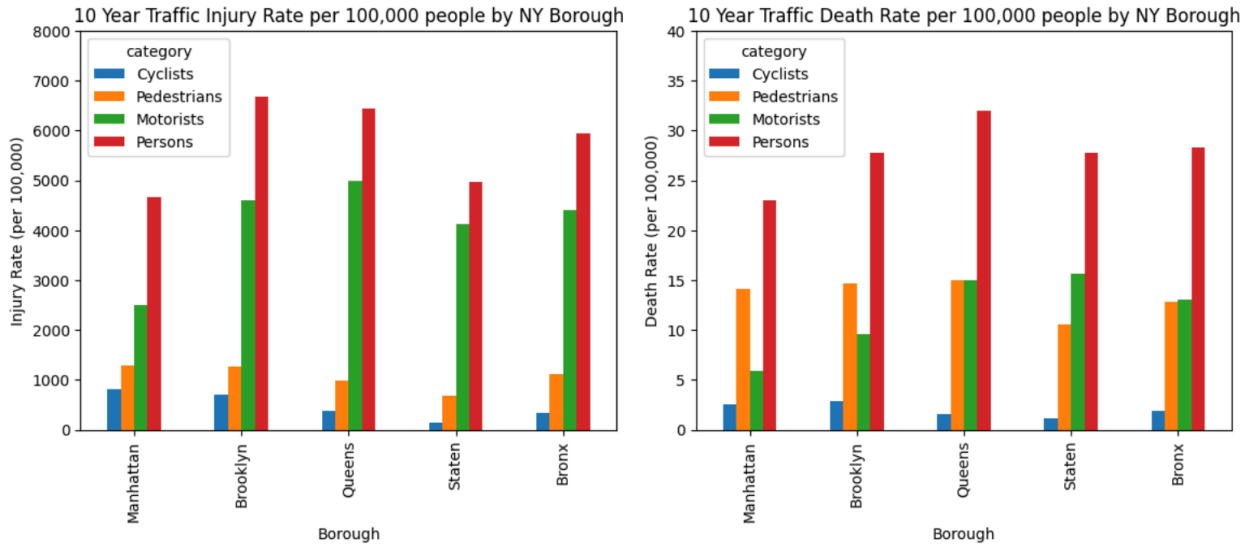
Overall our sanity check of the data revealed interesting trends consistent with what other trusted organizations have observed. Once we had a high level understanding of the dataset and traffic related deaths in NYC, we began to dive into our proposal questions.

Proposal Questions

Question 1: Which borough has the highest occurrences of injuries and death from vehicle collisions? (including pedestrian, cyclist, persons overall)

Based on median household income (Baruch data 2017), Manhattan is the wealthiest borough followed by Staten Island, Queens, Brooklyn and the Bronx. We wanted to understand if there were distinct trends in the data by borough that underscored a relationship between traffic accidents and a boroughs socioeconomic status. Would we see less collisions in weather Manhattan and more collisions in the Bronx due to less infrastructure investment, fewer NYPD resources, etc.?

Below is a table and bar plots representing the injury and death rate per 100,000 people for each borough of New York. The rate represents how many individuals, when controlling for borough population, have been involved over the span of the ten years of analysis. As a note, this is not the yearly rate, but the rate over the span of 10 years.



After aggregating and analyzing the data, the table shows mixed support for the idea that the poorer boroughs and richer boroughs experience differences in collision or death rates from traffic accidents. While Manhattan has substantially lower rates of total persons and motorists being injured or killed, it has about average rates (compared to the other boroughs) of injury and death for cyclists and pedestrians. This could be due to the frequency of pedestrians and cyclists in this borough compared to others, but that data was not gathered for this project.

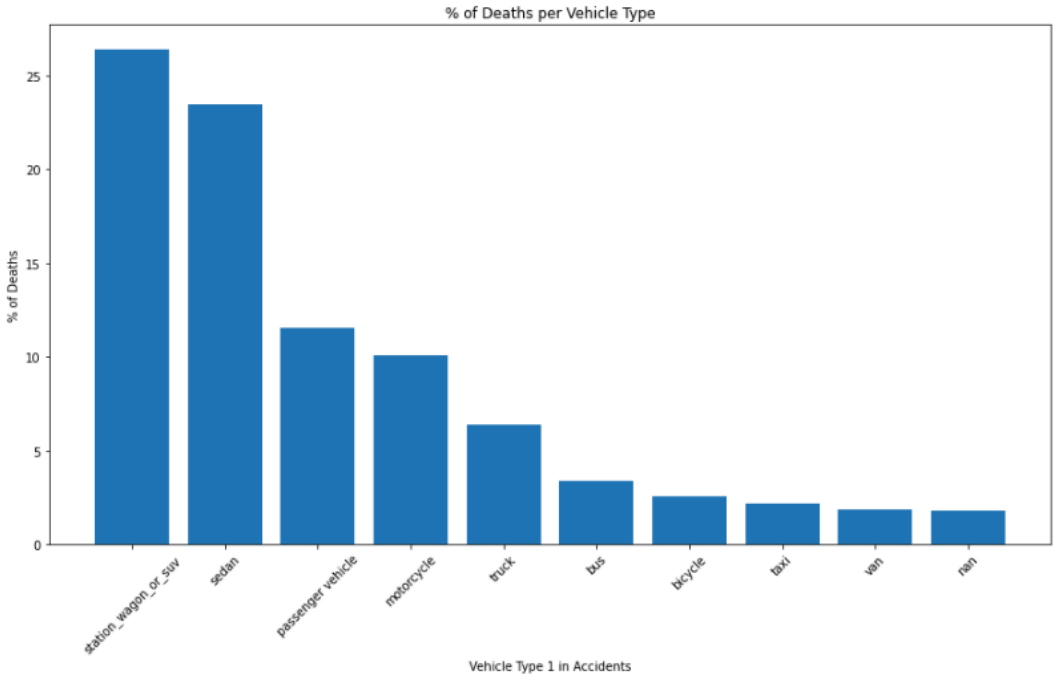
If there was a stronger association between socioeconomic class and rates of injury/death at the borough level, we would see rates increase in this order: Manhattan, Staten Island, Queens, Brooklyn, and Bronx. However, this is not a clear distinction in the data indicating that there are multiple contributing factors to traffic accidents across the boroughs.

Question 2: What vehicle type and contribution factor is more common to cause deaths in a collision? What about collisions?

To begin this part of the analysis, we took a closer look at vehicle_type_code_1 column, which contains values of the type of the main car in the car accident. However, we found that the values are not standardized and seem to be customized manually inputted values. Upon inspection, we cleaned up the column by implementing the following categorizations:

- sedan: values that contain “sedan”
- station_wagon_or_suv: values that contain “station wagon”, “sport utility”, “suv”
- truck: values that contain “truck”
- van: values that contain “van”
- motorcycle: values that contain “motorcycle”, “motorbike”
- bike: values that contain “bike”, “bicycle”

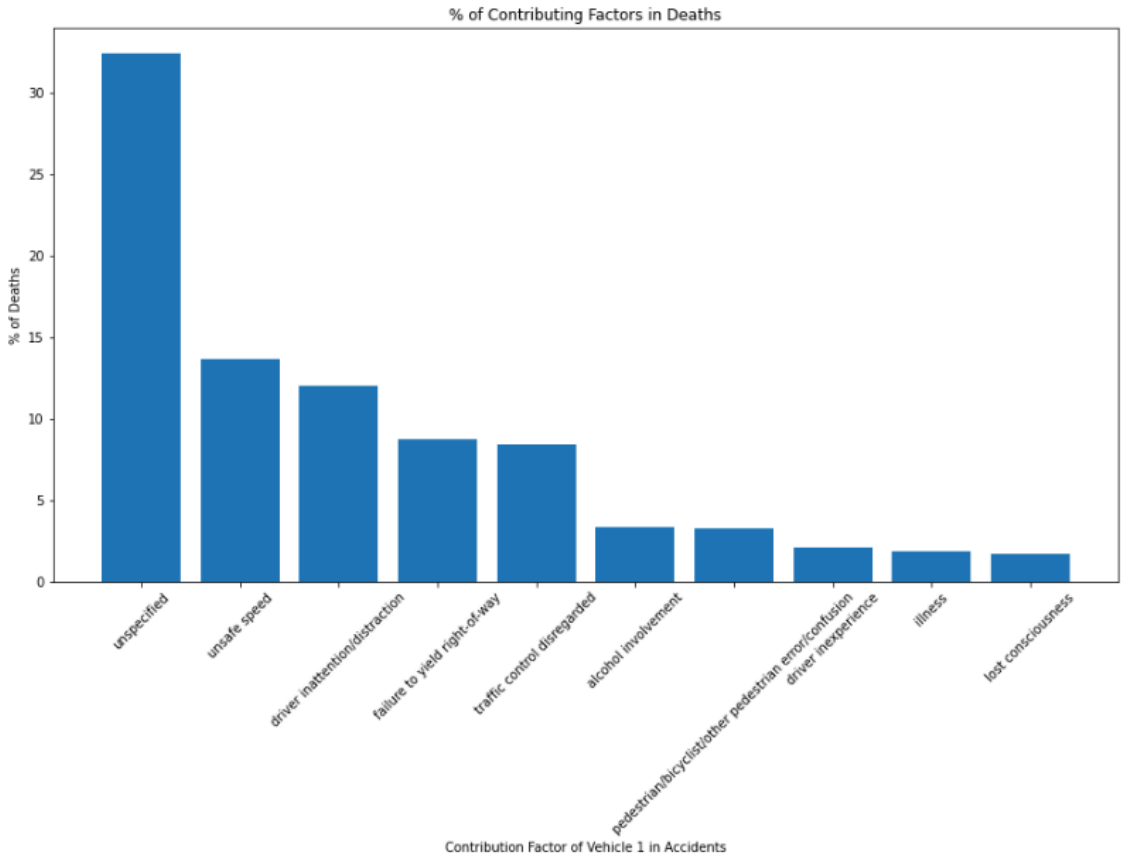
In the following graphs, we are showing how many percentages of deaths in the dataset are associated with each Vehicle Type. We can see that the top 5 vehicle types that contribute to death in a collision are station wagon or SUV, sedan, passenger vehicle, motorcycle and trucks, with station wagon or suv contributing to more than 1/4th of the total deaths.



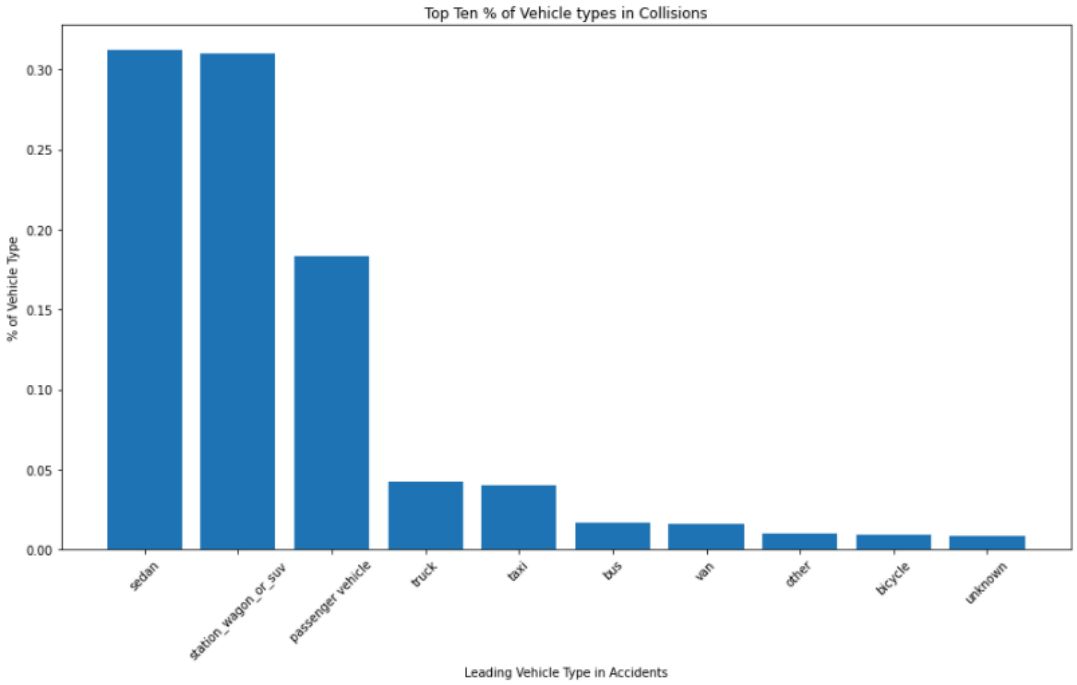
A study conducted by Urban Institute sought to evaluate the impact of increasing numbers of large vehicles like SUVs on the road and increased pedestrian deaths. While larger car sales were increasing 47% and 74% in 2009 and 2020, respectively, over this same time period pedestrian deaths increased from 4,000 to 6,800 nationally (juanlaw). It is speculated that while these larger vehicles offer more safety to their drivers, the increased size/mass and higher contact point mean decreased safety for pedestrians (juanlaw). This study further validates what we see in the dataset, that larger vehicles such as station wagons and SUVs contribute to more deaths on the road.

In terms of contributing factors, which means primary reason, of an accident that caused deaths, we found that almost 1/3rd of the dataset has unspecified reasons. Apart from them, unsafe speed contributes to ~14% of the deaths, ranked as the top contributing factor. Driver’s

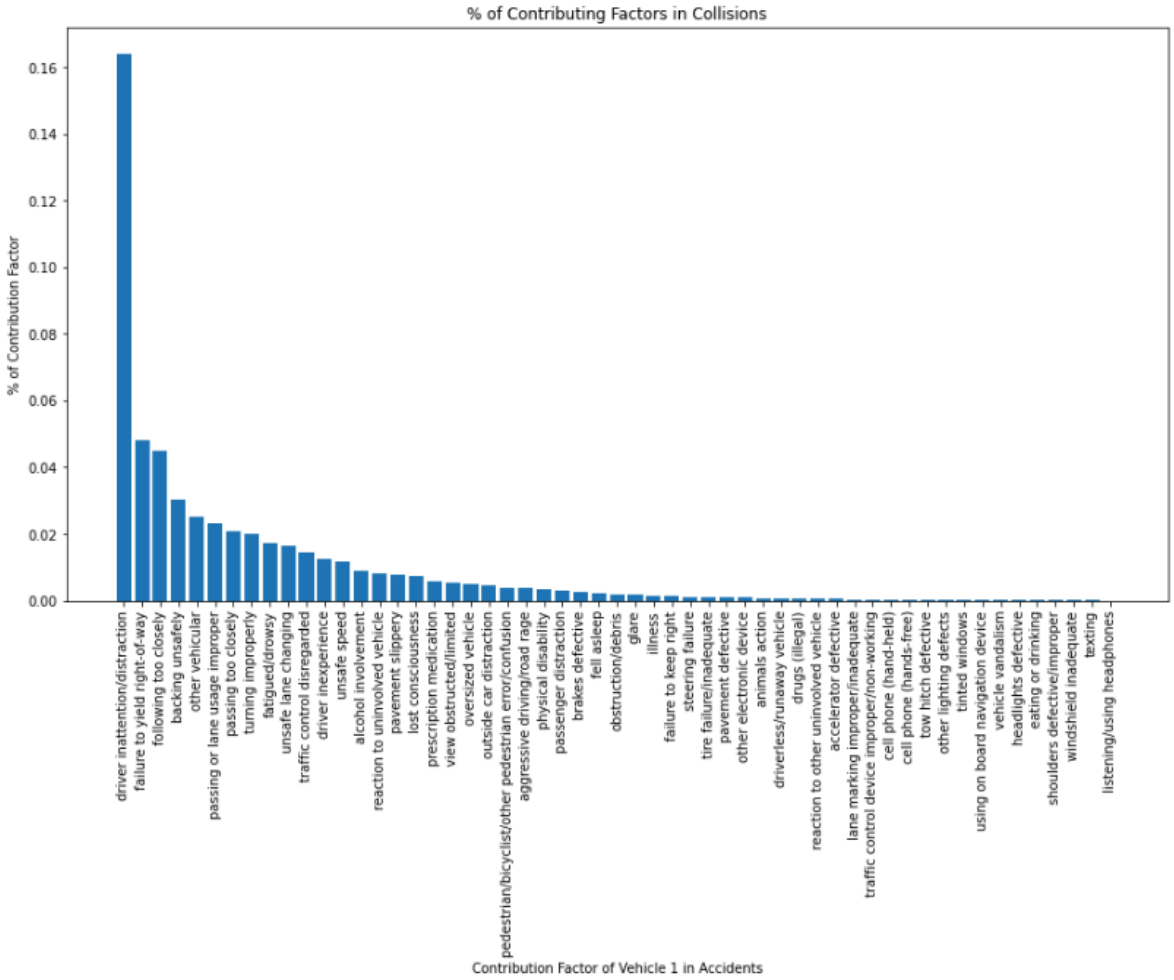
inattention and distraction, failure to yield right-of-way, disregarding traffic control, and alcohol consumption are among the top 5 contributing factors to death in an accident.



What about car collisions, regardless of if there were deaths? The results are not quite alike as those for accidents that caused deaths.
In the following graph, we are showing the top 10 vehicles that have the most collisions.Sedan surpassed station wagon or SUV as the vehicle type that has the most car accidents, though it is ranked after station wagon or SUV in the ranking for deaths. Passenger vehicles, trucks and taxis are also among the top 5 types that contribute to car accidents.



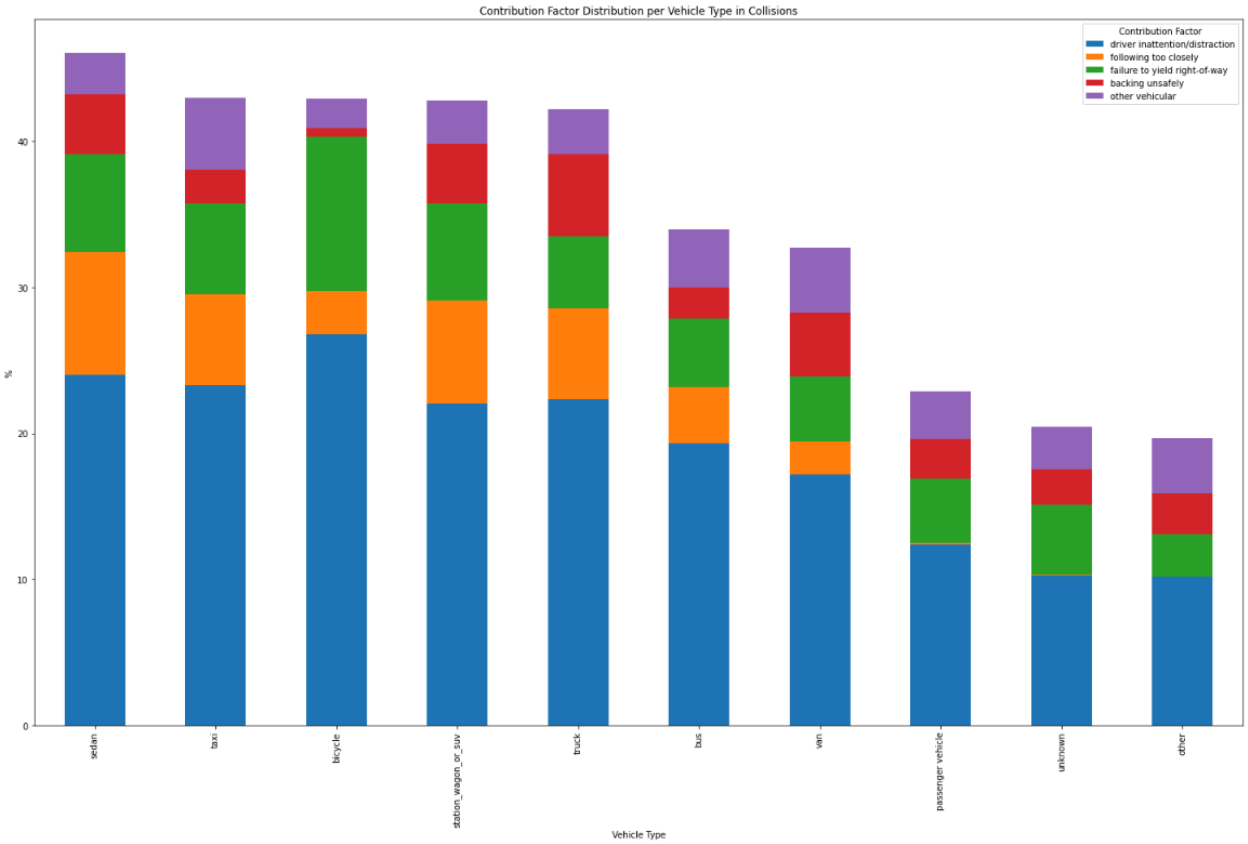
For contributing factors to collisions, we have an obvious winner - driver's inattention and distraction contributes to more than 16% of the accidents, while the second-ranked contributing factor(failure to yield right-of-way) only contributes to less than 5% of the total accidents. Following too closely, backing unsafely, passing too closely are the other top 5 contributions to car accidents.



Curious about how the top vehicle types and contributing factors relate to each other among the car accidents, we generated the following graph that shows for every top 10 car type that causes accidents, what is the distribution of contributing factors? Interestingly, the most common reason for bicycles to get in an accident is the rider's inattention and distraction.

Among all the car types, following too closely contributes to sedan accidents the most, while backing unsafely contributes to truck accidents the most. Last but not least, taxis have the most accidents due to the fault of others.

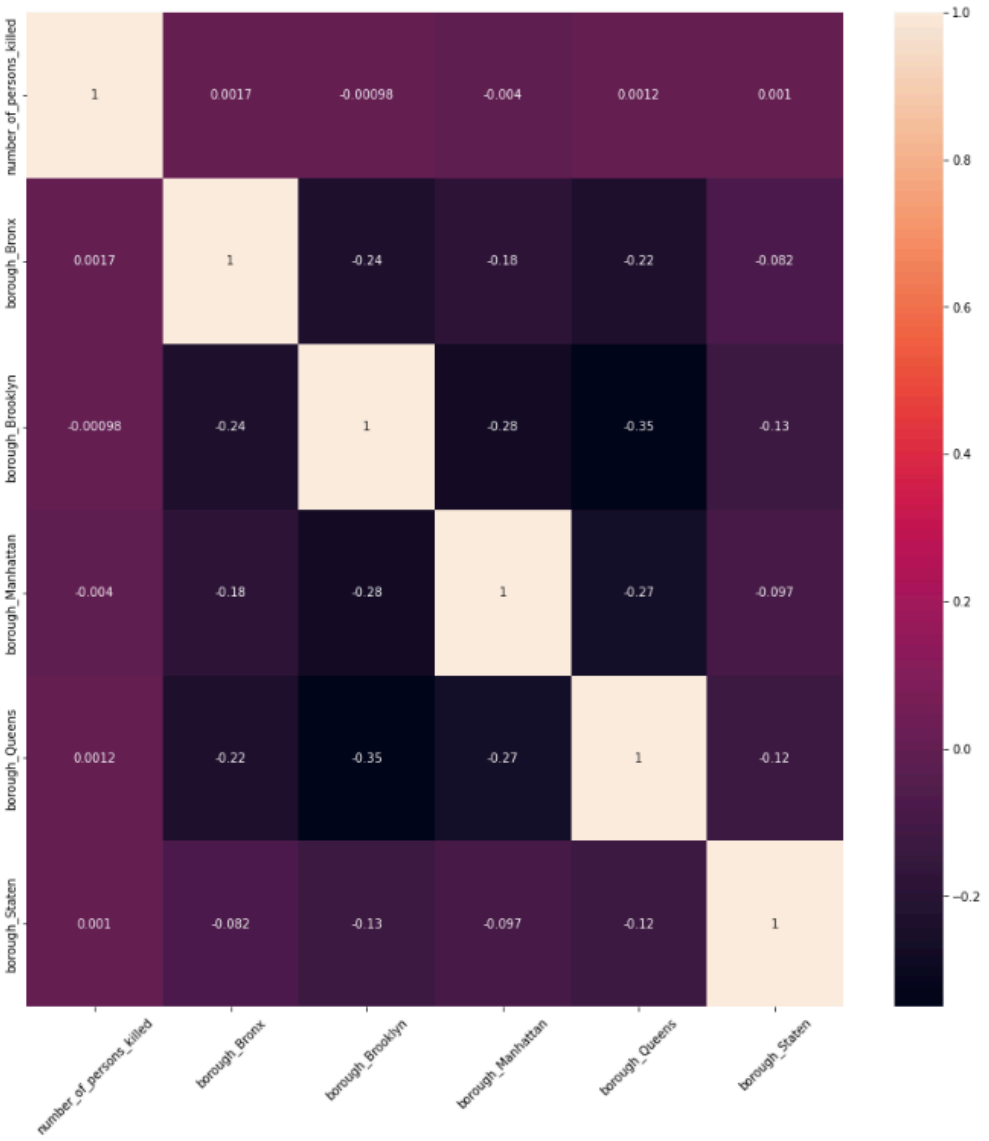
Our results reflect the fact that maybe more safety awareness training should be provided to bicycle riders so that they create less accidents due to them not paying attention on the road. Moreover, we may all know following too closely in sedans should be avoided, because sudden breaks can easily cause accidents, while it is proven in this dataset that following too closely causes the most accidents for sedans. For truck drivers, it's important to implement good practices when backing up. Since trucks are much bigger than other vehicles and have larger areas of blind spots, it's understandable why backing unsafely would cause more accidents for trucks.



Question 3: What's the correlation between characteristics analyzed in Question 1 & 2 (i.e. borough, vehicle type, contribution factor to collision) in the data and likelihood of death in a collision? What traits contribute to deaths in a collision if any?

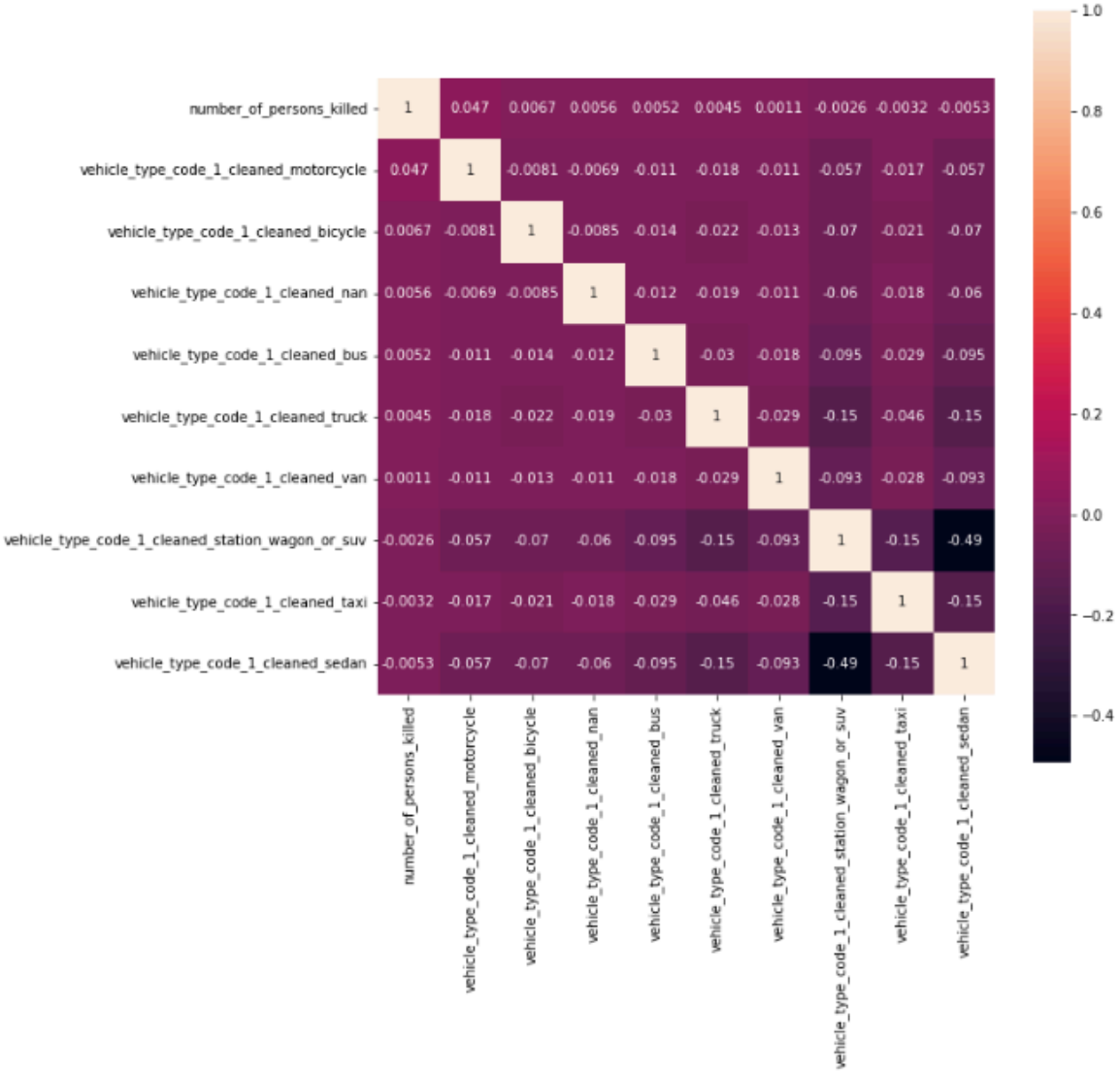
Based on the heatmap "Correlation between Boroughs and Car Accident Deaths", we can see that boroughs' correlation with deaths are all below 0.01. Since borough values are boolean values, (e.g. 1 if it's Manhattan, 0 if not), we would not expect high correlations. However, we can see from the heatmap that no individual borough was highly different from the average of the others' likelihood of death in traffic collisions.

Correlation between Boroughs and Car Accident Deaths

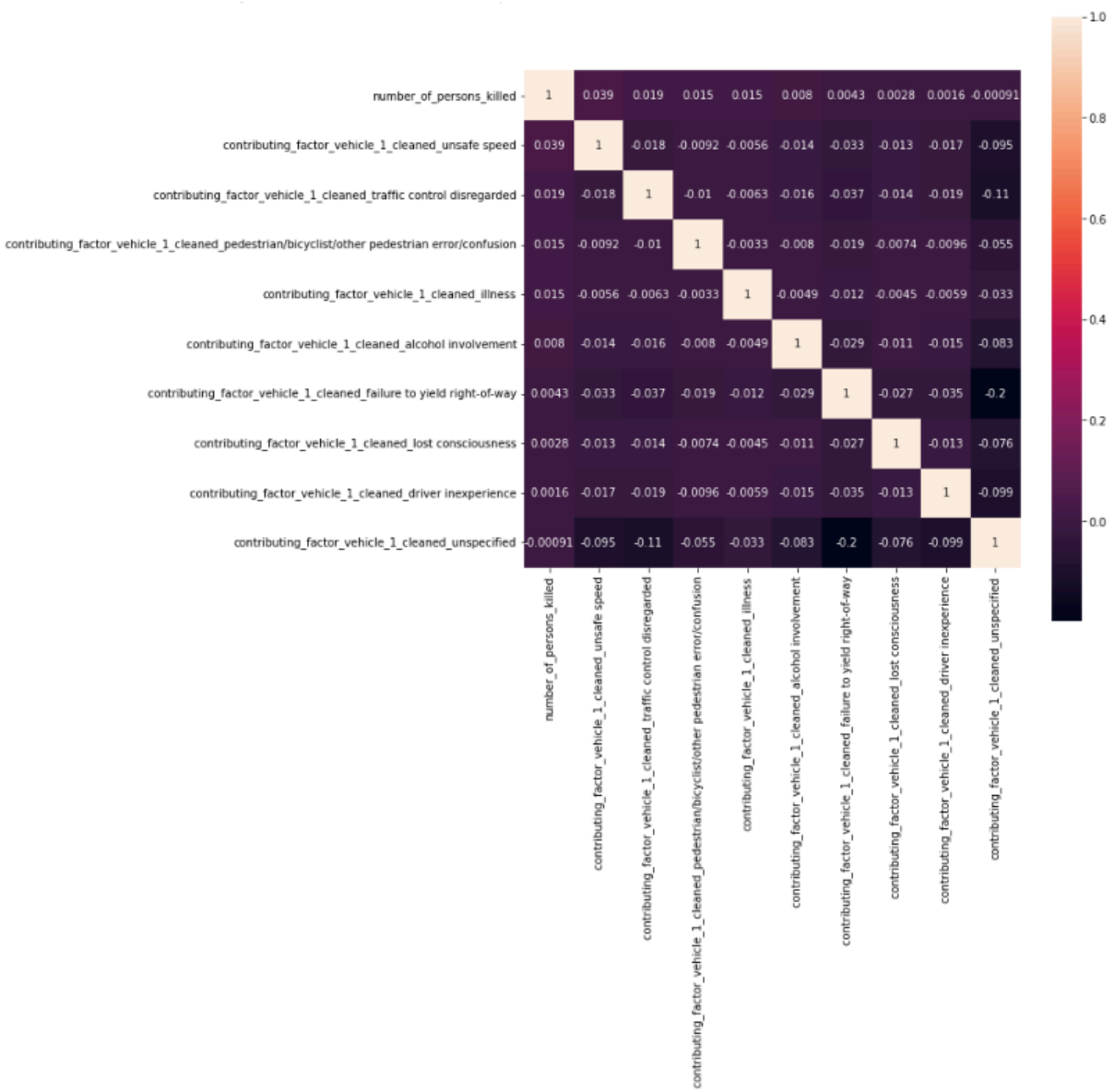


We calculated and graphed correlation between top 10 vehicles that contribute to deaths, and top 10 contribution factors for deaths. The results are similar to that of boroughs': no individual vehicle type stood out to have an obvious linear relationship with traffic collision deaths, and the same observation applies for contribution factors.

Correlation between Vehicle Type and Car Accident Deaths:



Correlation between Contributing Factor and Car Accident Deaths:



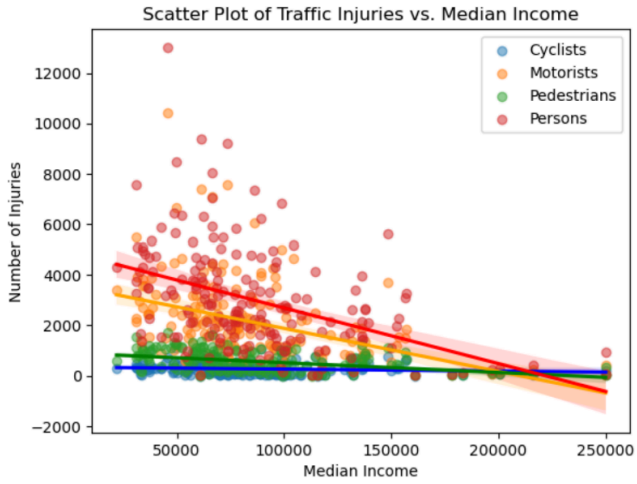
Question 4: Is there any correlation between median income of a zip code and number of deaths from collisions?

The relationship between income and traffic injuries/deaths is a primary focus on this project. It was expected that zip codes with higher median income would have lower numbers of injuries and deaths from traffic collisions than zip codes with lower median incomes.

To conduct this analysis, zip code median income data was scraped from simplemaps. The data comes from the 2022 American Community Survey run by the US Census Bureau. This data was imported as a Pandas dataframe and then merged with the main NYC Open Data motor vehicle accidents dataset on the grouped zip code.

With this new dataframe we were able to run a correlation analysis for the number of deaths and injuries by zip code median income. The results of this analysis by category (pedestrian, cyclist, motorist) are summarized in the table below and scatter plots below.

Correlations (Death Count and Zip Code Median Income):	
Total Deaths	-0.400
Pedestrian Deaths	-0.360
Cyclist Deaths	-0.150
Motorist Deaths	-0.318



Correlations (Injury Count and Zip Code Median Income):	
Total Injuries	-0.434
Pedestrian Injuries	-0.398
Cyclist Injuries	-0.129
Motorist Injuries	-0.431

When running descriptive correlations, we find small to medium sized negative correlations between motorist and total persons injured/killed and the median income of the zipcode, but we do not see a strong enough correlation when looking just at cyclist or pedestrian injuries or cyclists deaths.

When plotting the relationships using a scatter plot with lines of best fit, we find that the relationship between total person injured/killed and zip code median income is more similar to a negative curvilinear relationship. This means two things: 1) the relationship between total persons injured/killed and zip code median income is stronger than the correlations coefficient suggests, and 2) that the number of total persons injured/killed may be exponentially lower as the zip code median income increases. Preliminary results suggest this relationship is steeply negative until the median income is about 100k, then the relationship flattens slightly. Further

analysis is suggested to determine if this relationship is statistically better fit to a curvilinear rather than linear model.

We also see that there exists a negative relationship between pedestrian deaths and zip code median income, but we do not see a relationship with pedestrian injury. This may suggest that pedestrians are not more likely to be hit in lower income areas, but if they are hit, they are more likely to die.

Conclusion

Boroughs: Manhattan has less total persons and motorist death and injury, and Staten island had low injury rates all around. However, correlations show that the boroughs were essentially not very different from each other in injury/death. Further analysis is needed.

Vehicle Types: Station wagons or SUVs and Sedans are the most common types in collisions. However, this may be because of how common these types of cars are.

Contributing Factors: Driver Inattention/Distraction and speeding are most frequent causes, adding up to almost ⅛ of all reported causes in the past ten years. None of the common car types or contributing factors were found to be highly associated with injury or death.

Income and Injury/Death: There is a small to moderate negative correlation between zip code and most types of traffic accidents in NYC suggest additional factors beyond wealth are at play.

Returning to our guiding question - Are accidents really accidents? Inevitably, there are true accidents, but for many traffic related injuries/deaths oftentimes there is an element of choice that plays a role in the accident. We found in our research that speeding is a leading contributor to pedestrian deaths. The incidents themselves are accidents, but we made upstream choices about the speed limits on our roads and the distribution of speed bumps.

The leadership in NYC government has understood that choices play a role in what accidents we find acceptable and to what degree for 10 years. In the course of our research we came across the Vision Zero Program implemented in 2014 with the goal of eliminating traffic deaths through a combination of engineering, education, and enforcement" (Vision Zero). One Of the first steps Vision Zero took was to lower the city wide speed limit to 25 mph, suggesting that their research also indicated the high risk between speed limit and pedestrian deaths. Additionally Vision Zero developed and executed an automated speed camera enforcement program in school zones. Vehicles speeding in these designated zones were mailed a $50 dollar fine. This program has led to a 70% reduction in speeding in school zones and a14% decrease in injury (Vision Zero-2)."

Vision Zero is a fantastic example of a cultural shift away from accepting accidents to actively preventing them. With the right data available to inform your decisions and a shift in our thinking around the inevitability of accidents, we can make engineering, social and policy changes to save lives.

Appendix

Sub-findings:

- There appear to be fewer pedestrian deaths in Staten Island. It is speculated that this may be a function of fewer pedestrians in general in this borough given that it has the highest vehicle-to-resident ratio at 1.6 compared to 6.5 in Manhattan (Sillvie).
- There is low to moderate correlation between zip code median income and death/injury in NYC which suggests that while income may play a role in traffic accident risk, there are likely many other additional factors together that determine the overall risk of traffic accident in a given zip code.
- There is a cluster of high-death and high-injury zip codes all around Jamaica Bay in Brooklyn. Spending resources to investigate and mitigate in this cluster may be an efficient means to reduce traffic accidents.
- Pedestrian deaths appear to be clustered around major intersections, spread out across the city.

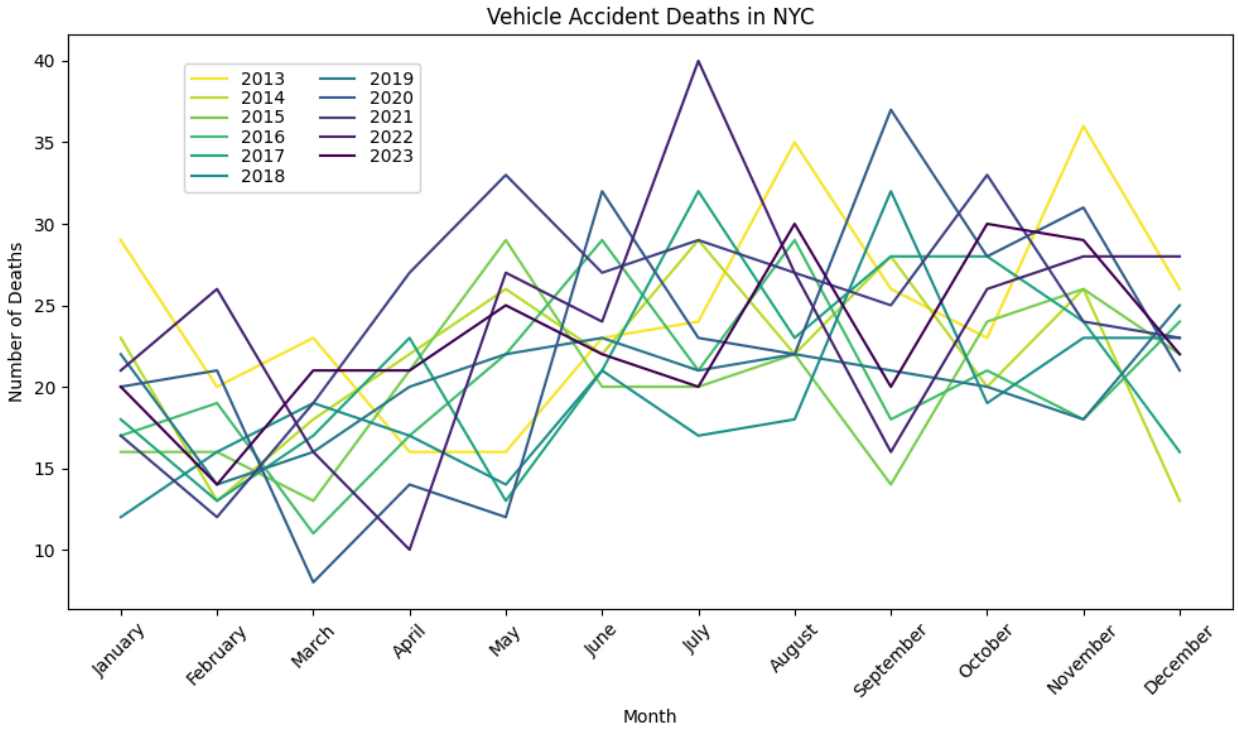
Future Questions for Exploration

- More research needs to be done to understand the reason for the cluster of high-injury and high-death zip codes around Jamaica Bay in Brooklyn. Both additional quantitative but also qualitative analysis through site visits is recommended.
- It is also recommended that an analysis of deaths and injuries by unit of area be conducted to determine if some of the high-injury and high-death zip codes are truly more deadly, or if they are just larger than many of the other zip codes.

Additional Data Evaluation/Findings

Seasonality

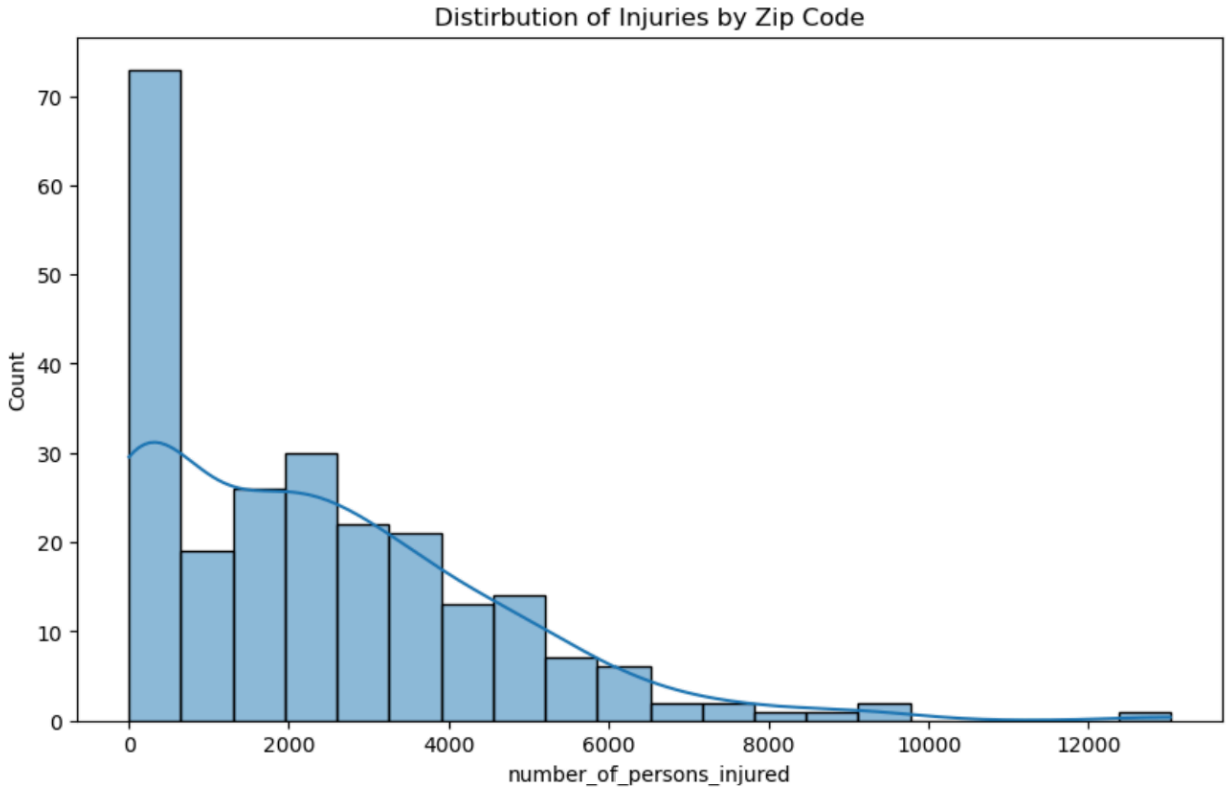
We also plotted monthly total traffic deaths for each year in our dataset to determine any seasonal trends. There is a distinct decrease in monthly deaths through the winter months. Then into the spring the monthly deaths begin to rise. Monthly deaths peak in the summer. This makes sense for NY which has distinct seasons, and fewer pedestrians and cars on the road in the winter months.



An interesting anecdote from researching NYC traffic incidents/fatalities - we came across an article from April 2022 that pointed out that the first quarter of 2022 was the deadliest Q1 of any year since 2014. “Traffic crashes killed 59 people in New York City during the first three months of 2022, a 44 percent increase over this point last year(Transport Alt).” When looking at the scatter plot above, you can see this sharp increase in monthly deaths for the first quarter of 2022. However, by the end of 2022 traffic deaths were down compared to 2021. And we wondered, what is the value in evaluating such a small snippet of data, just 3 months. where there is likely a high degree of variability. And this article demonstrated that almost anything with data can be turned into a news article.

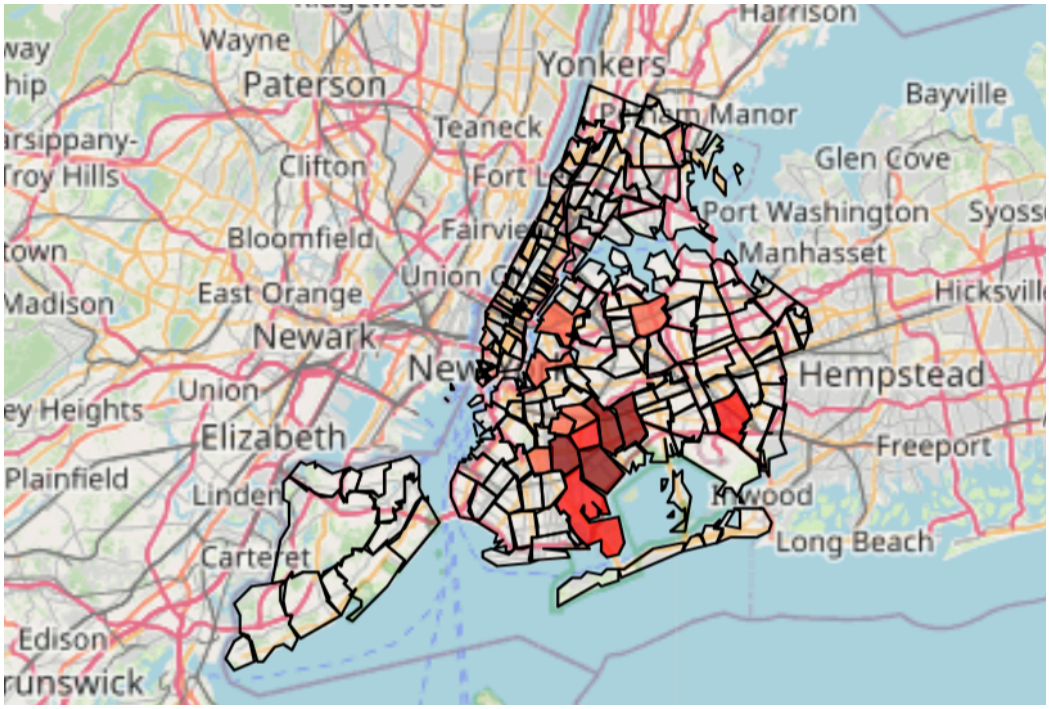
Spatial Distribution of Death and Injury

While evaluating the correlation between death and injury and zip code median household income, we noticed that there were some outlier zip codes where the rate of injury was dramatically higher than the majority of the other zip codes. The distribution plot below shows that the majority of zip codes had 5,000 injuries or less across our dataset period (2013-2023), with just a handful over 5,000 injuries.



We sorted the dataframe by injuries to find the outliers and plotted them on the map below.

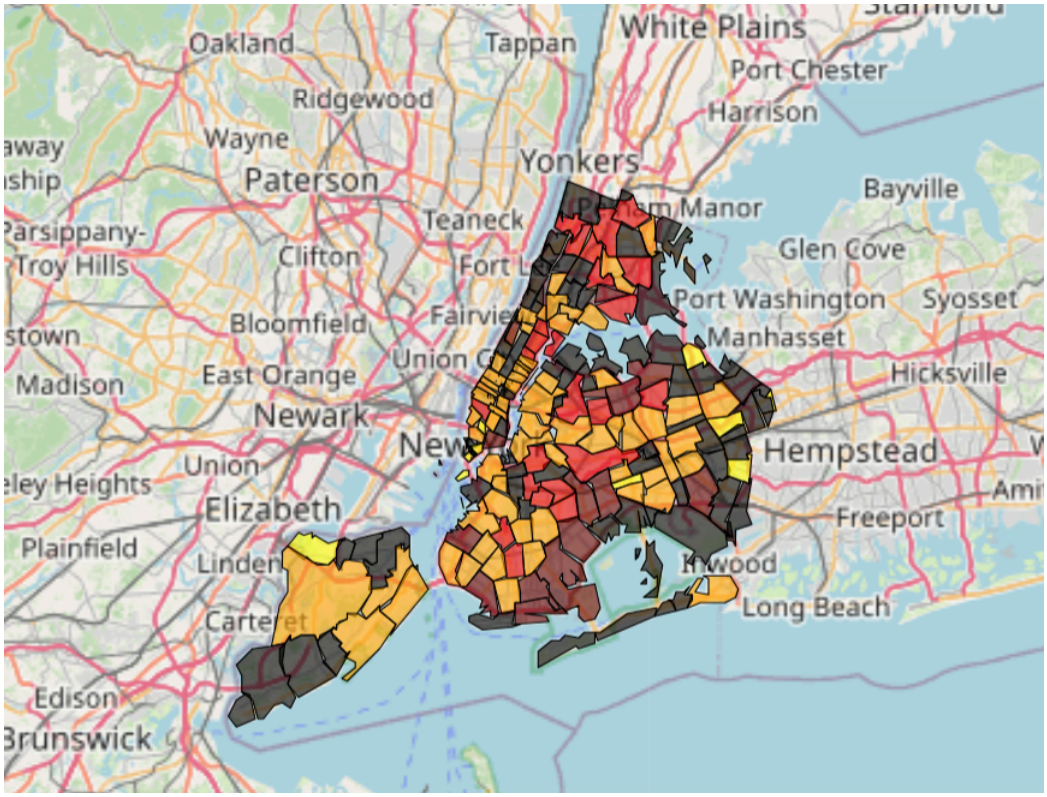
High-Injury Outlier Zip Codes



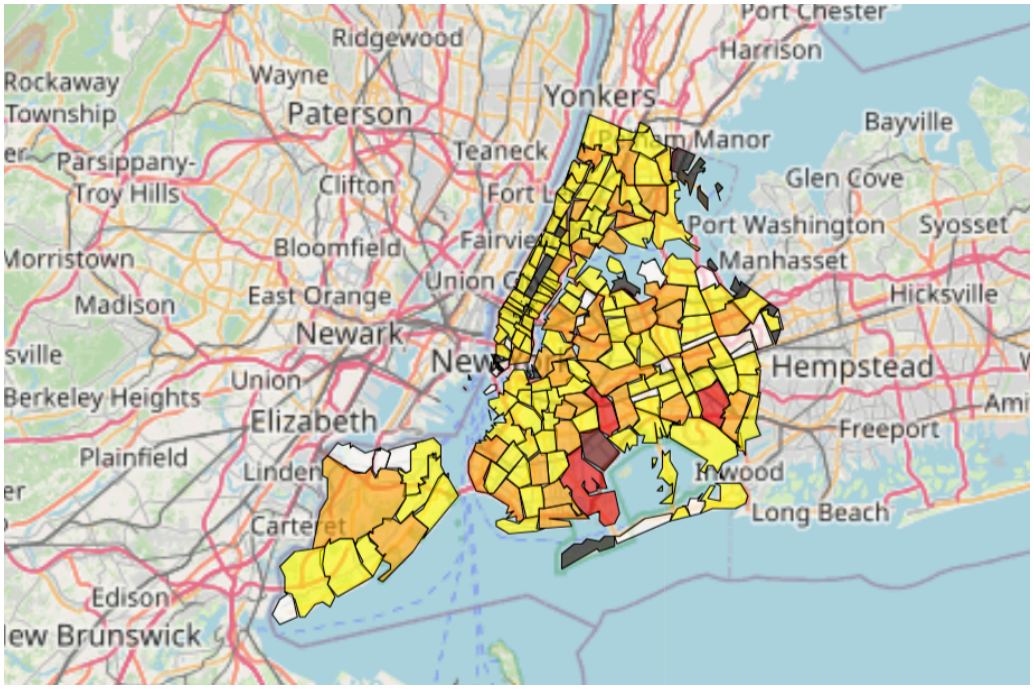
The location of the majority of these high-injury zip codes in Brooklyn is consistent with our earlier analysis indicating that Brooklyn had a high death and injury rate compared with the other boroughs. However, it was surprising to see several of these high-injury zip codes clustered together around Jamaica Bay. This may suggest something unique to this area, such as a major road system that crosses several of these high-injury zip codes may be contributing to higher injuries.

When looking at the number of pedestrian deaths by zip code and the number of overall deaths by zip code, shown in the two heat maps below, we see similar results to the high-injury distribution. There appears to be a cluster of zip codes in Brooklyn with high-injury and high-death rates.

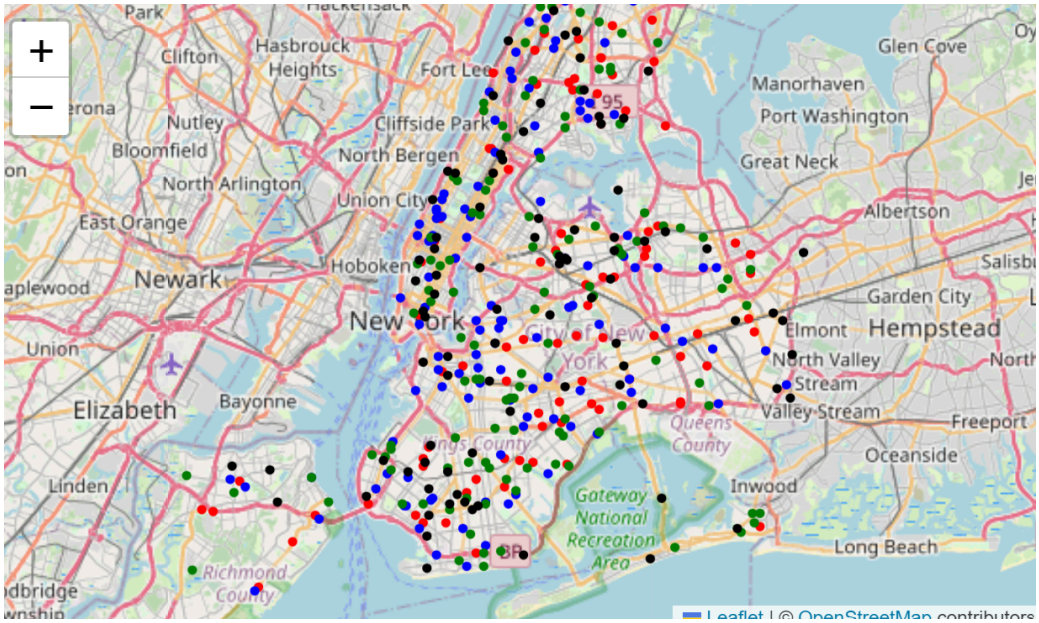
Pedestrian Deaths By Zip Code



Total Deaths by Zip Code



Lastly we sought to plot individual incidents of pedestrian death across the city. We quickly realized plotting all 10 years would result in a cluttered and unreadable map. Therefore the map below shows the location of a pedestrian death in NYC from 2017 - 2023. Some interesting observations from this map are the relatively low number of deaths in Staten Island. This we speculated earlier in the report may be the result of a higher person to car ratio in Staten Island compared to the rest of NYC. Additionally, this figure shows mini-clusters of pedestrian deaths around intersections and along major roads.



Resources

1. Surico, John "*New York City's Decade-Long Battle for Pedestrian Safety." Bloomberg, March 2024.* https://www.bloomberg.com/news/features/2024-03-11/new-york-city-s-decade-long-battle-for-pedestrian-safety?embedded-checkout=true
2. IIHS."Male and Female Fatality Facts." *Insurance Institute for Highway Safety*, Highway Loss Data Institute, accessed April 15, 2024, https://www.iihs.org/topics/fatality-statistics/detail/males-and-females.
3. Juan Law. "Which Pedestrians Are Most at Risk for Accident." *Juan Law*, no publication date, accessed April 15, 2024, https://www.juanlaw.com/car-accidents/which-pedestrians-are-most-at-risk-for-accident.
4. NHTSA-1. "2022 Traffic Deaths: 2023 Early Estimates." *National Highway Traffic Safety Administration*, U.S. Department of Transportation, accessed April 15, 2024, https://www.nhtsa.gov/press-releases/2022-traffic-deaths-2023-early-estimates.
5. NHTSA-2. "Early Estimate of 2021 Traffic Fatalities." *National Highway Traffic Safety Administration*, U.S. Department of Transportation, accessed April 15, 2024, https://www.nhtsa.gov/press-releases/early-estimate-2021-traffic-fatalities.
6. NYCDOH. "Data Brief No. 59: [Title of the Data Brief]." *New York City Department of Health*, [publication date if available], www.nyc.gov/assets/doh/downloads/pdf/epi/databrief59.pdf.
7. NYCgov. "Mayor Adams Launches Major Campaign to Tackle Traffic Violence: 'Speeding Ruins Lives, Slow Down'." *Office of the Mayor - New York City*, 2022, www.nyc.gov/office-of-the-mayor/news/265-22/mayor-adams-launches-major-campaign-tackle-traffic-violence-speeding-ruins-lives-slow-down-#/0.
8. NYCOD-1. "Overview." *NYC Open Data*, City of New York, accessed April 15, 2024, https://opendata.cityofnewyork.us/overview/.
9. NYCOD-2. "Motor Vehicle Collisions - Crashes: About the Data." *NYC Open Data*, City of New York, accessed April 15, 2024, https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95/about_data.
10. Williams, Juanita. "NYC Population Decline." *The New York Times*, 14 Mar. 2024, www.nytimes.com/2024/03/14/nyregion/nyc-population-decline.html.
11. Silvie. "Staten Island Has More Cars Per Capita Than Anywhere in NYC." *Staten Island Advance*, 2016, www.silive.com/news/2016/11/staten_island_has_more_cars_pe.html.
12. Simplemaps. "Median Household Income by ZIP Code in New York City." *SimpleMaps*, SimpleMaps.com, accessed April 15, 2024, https://simplemaps.com/city/new-york/zips/income-household-median.
13. Transport Alt. "New Data Shows 44 Percent Increase in Traffic Fatalities During First Three Months of 2022: Deadliest Start to Any Year Since Vision Zero Began in 2014." *Transportation Alternatives*, Transportation Alternatives, accessed April 15, 2024, https://transalt.org/press-releases/new-data-shows-44-percent-increase-in-traffic-fatalities-during-first-three-months-of-2022-deadliest-start-to-any-year-since-vision-zero-began-in-2014.
14. Singer, Jessie. "There Are No Accidents." *Vox*, Vox Media, 5 Apr. 2022, www.vox.com/23016529/there-are-no-accidents-jessie-singer.
15. Vision Zero. "Vision Zero." *NYC.gov*, City of New York, accessed April 15, 2024, https://www.nyc.gov/content/visionzero/pages.