Prelim Report STAT373

Riley McIvor

11/4/2025

Abstract

GITHUB REPOSITORY: https://github.com/Derboshr/Assignment-Repository
Mass data collection presents many opportunities to independent researchers, who utilise transformative techniques in order to divulge findings relating to a pertinent questions they may present, regarding the data. Within this study, a large data set collected on New York City vehicle collision statistics is examined in order to explore the utility of these data sets in producing faithful results. While the data is largely consistent in producing results found in academic studies, these results are founded on contextual knowledge, and the information previously obtained in other studies.

Contents

Introduction

In this report, data pertaining to vehicle collisions in New York city is to be dissected, transformed and observed in order to gain insight on the dynamics of the 5 boroughs of the city, as well as assessing state of data aggregation throughout this set.

To begin, the concepts surrounding the data will be introduced, as well as studies that may need to be considered in relation to this report. After that, the research questions will be discussed in full, and what answering them may achieve.

Following the descriptions of key data used in the report, the analysis that drew from this data is to be discussed, and used to present notions towards the research questions.

Finally, conclusions that may come from this exploratory analysis are to discussed, as well as limitations and other insights that may need to be considered.

Background

The topic of interest is that of New York vehicle collisions. In exploratory analysis, it is important to understand key terminology and definitions that may not be entirely intuitive. This data set refers to 'vehicle collisions', which NYC-Open-Data (2025) refers to as "collisions where someone is injured or killed, or where there is at least $1000 worth of damage". This means this data set is not completely representative of all collisions, as small crashes in which there was little damage, or the dispute was handled without police intervention, i.e. cash payouts, are not included within the data.

Additionally, vehicles are defined as any transportation method involved in the collision, meaning that vehicles such as bikes, tractors, E-scooters, snow plows and other unusual entries are included, as well as the usual vehicles.

Finally, the data set refers to boroughs, which is an administrative district of a larger encompassing area. In New York specifically, the five boroughs are Manhattan, Brooklyn, Queens, The Bronx and Staten Island.

A research paper from Haddon and McCarroll (1963) presents results from a thorough and carefully implemented study. The study was conducted on fatal crashes in New York, with a sample size of 43, and concluded that the most common differences between drivers involved in the crashes and those who were not were alcohol concentration, proximity to their home, and those involved in criminal activity.

This paper had extreme attention to detail in its data set, notably attributable to the small sample size. This meant that each crash would be able to be catagorised by variables such as driver alcohol content, medical history, age, gender (although there were no female drivers in this study), and socioeconomic status.

Another paper reporting on research conducted between 2013-2023 suggests that the decline of taxis, as well as the registration of newer vehicles has led to a decline in traffic accidents. Mittal and Lim (2024) also observed the dynamics of accidents in relation to Covid-19, presenting the notable drop in all traffic accidents, especially those of high severity during spikes of coronavirus cases.

It is through first observing these cases that we may assess the practicality of the main data set; considering if we may achieve the same conclusions these studies have come to.

Research Question

The aim of this report is to produce practical, intuitive and stable results through transformation and tidying of the aforementioned data set. Successful results will mean the following questions are able to be answered reliably:

- How have the dynamics of vehicle collisions changed over time?
- What are the common attributes of a serious collision?
- Which borough is the most "dangerous" in relation to driving?

In answering these questions, it would be enlightening to make comparisons with existing results, and thus creating an answer for the final question; does the existence of this data set (and other sets similar in implementation) allow for useful analysis?

The scope of this data set is unimaginably wide, and presenting the findings of this study, alongside the findings of other sources, allows for a deeper understanding of useful/impractical data.

The importance of answering the final question comes from the notion of efficient and effective data sampling. Understanding if effective data-wrangling may produce rich results even under scrutinous conditions is key to exploratory data analysis.

Rationale

The need for this study comes from the interest in vehicle collision dynamics throughout varying environments. It is widely known that crash dynamics vary greatly from urban to rural areas, in the sense that rural areas tend to have less crashes, while having proportionally more fatal crashes, and urban areas having many more minor crashes. These are the dynamics of two extremely different environments, with clear cause and effects. The interest in the opportunity this data analysis provides arises from the diversity of New York's environment. Each borough is clearly distinct from each other, and inferences about their crash dynamics may be constructed by geographical location, population density etc. The data analysis aims to verify these inferences, and create an understanding of crash dynamics that for most may be only surface level.

Data Descriptions

The data set is a free, public use set found from the US Government's open data base. The data frame has over 2 million observations, ranging from July 2012 to March 2025. Each observation has 29 variables, including dates, locations, casualties, vehicle codes for up to 5 vehicles and contributing factors for up to 5 vehicles.

This data set, due to its exhaustive coverage, has many issues and limitations, as well as inconsistencies. For example, Vision Zero, an initiative for traffic safety was started in 2014, and had a fairly slow roll-out. This initiative emphasized succinct data collection. Before the took full effect, vehicle type codes were observed using the standard DMV codes, which slowly transitioned into a more detailed collection of data. The prime example of this is the "passenger vehicle" class in this data set, which was in later observations divided further into classes such as sedans, coupes and station wagons.

The data is further limited by the data collection, which is largely inconsistent. Many observations have blank details in multiple variables, with the only consistently filled variable being date. Due to the data being compiled largely from 3rd party sources, it is understandable that the data collection is inconsistent.

In comparison to the sample used in the study by Haddon and McCarroll (1963), this set contains much more inconsistencies, and a lot less variables that would aid in an extremely thorough analysis of the microcosms of the vehicle collisions.

For the analysis, the prime variables that will be observed are borough, vehicle type code, contributing factor, date and injuries. The analysis of these variables will allow for a deep understanding of the severity, location, and details of the collisions, without complicating the data.

- Vehicle type code would allow to identify top contributers to crashes, and observe the most 'dangerous' vehicle.
- Contributing factor allows for an understanding of the dynamics of the crash, and may indicate which vehicles are involved in crashes for recurring reasons.
- Data is useful for plotting relationships over time, and adds structure to the data set.
- Injuries (as well as fatalities) indicate the severity of a crash.

Exploratory Data Analysis

In order to observe the data in a more concise way, sub-frames must be created, each with separate purposes. The data has been separated into 4 main frames, each with varying amounts of tidying:

Motor_crash_better:

The code is first filtered to remove entries with blank vehicle observations:

```
Motor_crash_better <- filter(Motor_Vehicle_Collisions_Crashes,
    VEHICLE.TYPE.CODE.1 != "" & VEHICLE.TYPE.CODE.2 != "" | VEHICLE.TYPE.CODE.1 !=
        "" & VEHICLE.TYPE.CODE.2 != "" & VEHICLE.TYPE.CODE.3 !=
        "" | VEHICLE.TYPE.CODE.1 != "" & VEHICLE.TYPE.CODE.2 !=
        "" & VEHICLE.TYPE.CODE.3 != "" & VEHICLE.TYPE.CODE.4 !=
        "" | VEHICLE.TYPE.CODE.1 != "" & VEHICLE.TYPE.CODE.2 !=
        "" & VEHICLE.TYPE.CODE.4 != "" & VEHICLE.TYPE.CODE.3 !=
        "" & VEHICLE.TYPE.CODE.5 != "")
```

Afterwards, key variables are selected for analysis, and a new variable, 'NO.VEHICLES' which displays the amount of vehicles involved in an observation is created. The frame is further filtered to remove cases with null borough entries, and finally the 'CRASH.DATE' variable is altered so that R will recognise the entries as dates.

```
Motor_crash_better <- select(Motor_crash_better, CRASH.DATE,
    BOROUGH, NUMBER.OF.PERSONS.INJURED, NUMBER.OF.PERSONS.KILLED,
    VEHICLE.TYPE.CODE.1, VEHICLE.TYPE.CODE.2, VEHICLE.TYPE.CODE.3,
    VEHICLE.TYPE.CODE.4, VEHICLE.TYPE.CODE.5, COLLISION_ID)
Motor_crash_better <- Motor_crash_better %>%
    mutate(NO.VEHICLES = ifelse(VEHICLE.TYPE.CODE.5 != "", 5,
        ifelse(VEHICLE.TYPE.CODE.4 != "", 4, ifelse(VEHICLE.TYPE.CODE.3 !=
            "", 3, ifelse(VEHICLE.TYPE.CODE.2 != "", 2, 1)))))
Motor_crash_better <- filter(Motor_crash_better, BOROUGH != "")
Motor_crash_better$CRASH.DATE <- as.Date(Motor_crash_better$CRASH.DATE,
    "%m/%d/%Y")
```

Using Motor_crash_better, we are able to create our next frame with relative ease.

Motor_crash_better5

Motor_crash_better5 is a set created simply through filtering the previous frame, to only include observations with crashes involving 5 cars (the maximum in this set):

```
Motor_crash_better5 <- filter(Motor_crash_better, NO.VEHICLES ==
    5)
```

Observing this set may allow us to understand the dynamics of the most severe crashes only, and for a comparison with the dynamics of the previous set.

Motor_crash_vehicle_slim

This frame requires a considerable amount of set-up. The aim of this frame is to only include the top 10 most common vehicles, since there are countless duplicate or inadmissible entries. This is most likely due to the non-standardised collection of data before 2014.

The first step is to combine variable names for similar names. This has only been done for the first 20 variables in this frame, for simplicity.

```
Motor_crash_better$VEHICLE.TYPE.CODE.1[Motor_crash_better$VEHICLE.TYPE.CODE.1 ==
    "TAXI"] = "Taxi"
Motor_crash_better$VEHICLE.TYPE.CODE.2[Motor_crash_better$VEHICLE.TYPE.CODE.2 ==
    "TAXI"] = "Taxi"
Motor_crash_better$VEHICLE.TYPE.CODE.3[Motor_crash_better$VEHICLE.TYPE.CODE.3 ==
    "TAXI"] = "Taxi"
Motor_crash_better$VEHICLE.TYPE.CODE.4[Motor_crash_better$VEHICLE.TYPE.CODE.4 ==
    "TAXI"] = "Taxi"
Motor_crash_better$VEHICLE.TYPE.CODE.5[Motor_crash_better$VEHICLE.TYPE.CODE.5 ==
    "TAXI"] = "Taxi"
```

Then we create a decreasing table of vehicle types and subsequent tibble of the table.

```
vehicle2_sort <- sort(table(Motor_crash_better$VEHICLE.TYPE.CODE.1),
    decreasing = TRUE)
vehicle_names <- names(vehicle2_sort)

vehiclesorted <- tibble(vehicle_names, vehicle2_sort)
```

From this, we are able to see the top 10 vehicle entries, and use filter to create the new frame.

Motor_factor_10

This frame aims to make use of the Factor variable, which is extremely inconvenient, unless a significant amount of tidying is implemented. The set up for Motor_crash_fvehicle_slim is identical to that of Motor_crash_better, with the addition of selecting the factor variables, and filtering for null entries;

```
Motor_crash_fvehicle_slim <- filter(Motor_crash_fvehicle_slim,
    CONTRIBUTING.FACTOR.VEHICLE.1 != "Unspecified" & CONTRIBUTING.FACTOR.VEHICLE.2 !=
        "Unspecified" | CONTRIBUTING.FACTOR.VEHICLE.1 != "Unspecified" &
        CONTRIBUTING.FACTOR.VEHICLE.2 != "Unspecified" & CONTRIBUTING.FACTOR.VEHICLE.3 !=
        "Unspecified" | CONTRIBUTING.FACTOR.VEHICLE.1 != "Unspecified" &
        CONTRIBUTING.FACTOR.VEHICLE.2 != "Unspecified" & CONTRIBUTING.FACTOR.VEHICLE.3 !=
        "Unspecified" & CONTRIBUTING.FACTOR.VEHICLE.4 != "Unspecified" |
        CONTRIBUTING.FACTOR.VEHICLE.1 != "Unspecified" & CONTRIBUTING.FACTOR.VEHICLE.2 !=
            "Unspecified" & CONTRIBUTING.FACTOR.VEHICLE.3 !=
            "Unspecified" & CONTRIBUTING.FACTOR.VEHICLE.4 !=
            "Unspecified" & CONTRIBUTING.FACTOR.VEHICLE.5 !=
            "Unspecified")
```

After that, pivot longer is used, so that we can view each vehicle as a singular case, rather than viewing them as vehicle1, vehicle2 etc., and a new variable 'Factor' is created, to achieve the same effect for the 5 factor variables.

```
Motor_factor_slim_long <- pivot_longer(Motor_crash_fvehicle_slim,
    cols = 5:9, names_to = "Contributing_Vehicle", values_to = "Vehicle")

Motor_factor_slim_long <- Motor_factor_slim_long %>%
    mutate(Factor = ifelse(Contributing_Vehicle == "VEHICLE.TYPE.CODE.1",
        CONTRIBUTING.FACTOR.VEHICLE.1, ifelse(Contributing_Vehicle ==
            "VEHICLE.TYPE.CODE.2", CONTRIBUTING.FACTOR.VEHICLE.2,
```

```
            ifelse(Contributing_Vehicle == "VEHICLE.TYPE.CODE.3",
                CONTRIBUTING.FACTOR.VEHICLE.3, ifelse(Contributing_Vehicle ==
                  "VEHICLE.TYPE.CODE.4", CONTRIBUTING.FACTOR.VEHICLE.4,
                  CONTRIBUTING.FACTOR.VEHICLE.5)))))
```

After that, a tibble is made of the top 10 factor variables is made, similar to in the previous frame.

```
factor2_sort <- sort(table(Motor_factor_slim_long$Factor), decreasing = TRUE)
factor_names <- names(factor2_sort)

factorsorted <- tibble(factor_names, factor2_sort)
# making slice able to be used

factorsorted10 <- slice_tail(factorsorted, n = 61)
# getting rid of cases where ' ' is the top factor, due to
# crashes with < 5 vehicles
factorsorted10 <- slice_head(factorsorted10, n = 10)
```

Dissimilar to the previous frame, due to the fact that all vehicle and factor variables have been combined, this means that the top entry for either variable is now ” “.

```
slice_head(factorsorted, n=10)
```

```
## # A tibble: 10 x 2
##    factor_names                     factor2_sort
##    <chr>                            <table[1d]>
##  1 ""                               443504
##  2 "Driver Inattention/Distraction" 108308
##  3 "Other Vehicular"                 42544
##  4 "Failure to Yield Right-of-Way"   20220
##  5 "Passing or Lane Usage Improper"  13839
##  6 "Following Too Closely"           12136
##  7 "Unspecified"                     11064
##  8 "Backing Unsafely"                10051
##  9 "Passing Too Closely"              9988
## 10 "Traffic Control Disregarded"      8758
```

Making use of the fact that the slice function can be used on tibbles, we can create a frame displaying the top 10 factor variables. After using filter on Motor_factor_slim_long for the top 10 factors, the frame Motor_factor_10 is created.

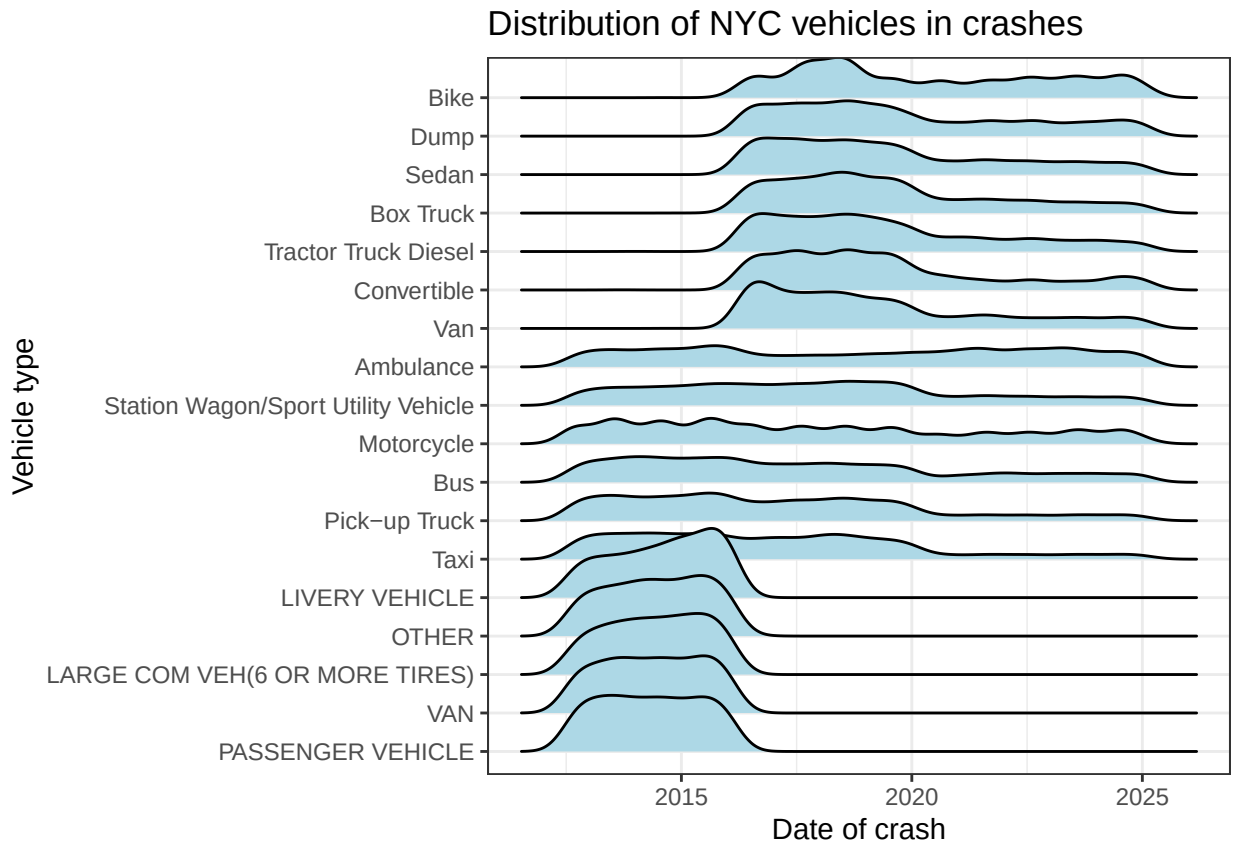
Analysis

Using these frames, we can now aim to answer the 3 research questions, the first one being ‘How have the dynamics of vehicle collisions changed over time?’

To answer this, we will use the following ridge plot, which displays density of the top 18 vehicle types over the period of the data set.

```
Motor_crash_vehicle_slim %>% mutate(VEHICLE.TYPE.CODE.1 = fct_reorder(VEHICLE.TYPE.CODE.1,CRASH.DATE,mea
  ggplot() +
  geom_density_ridges(aes(x=CRASH.DATE,y=VEHICLE.TYPE.CODE.1),fill='lightblue')+theme_bw()+
  labs(x = "Date of crash", y = "Vehicle type", title = "Distribution of NYC vehicles in crashes")
```

```
## Picking joint bandwidth of 118
```



As is displayed in the plot, we may come to the same conclusion that was stated by Mittal and Lim (2024), in the fact that following the first COVID-19 outbreak in 2020, there seems to have been a sharp decline in most vehicle types appearing in recorded collisions. We see that only 2 vehicle types do not follow this trend, Ambulances, which may have even had an increased road appearance during the pandemic, and bicycles, which did not see an increase or decrease.

```
summary(Motor_crash_vehicle_slim$CRASH.DATE)
```

```
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2012-07-01" "2014-08-21" "2016-10-11" "2017-04-20" "2019-05-23" "2025-03-18"
```

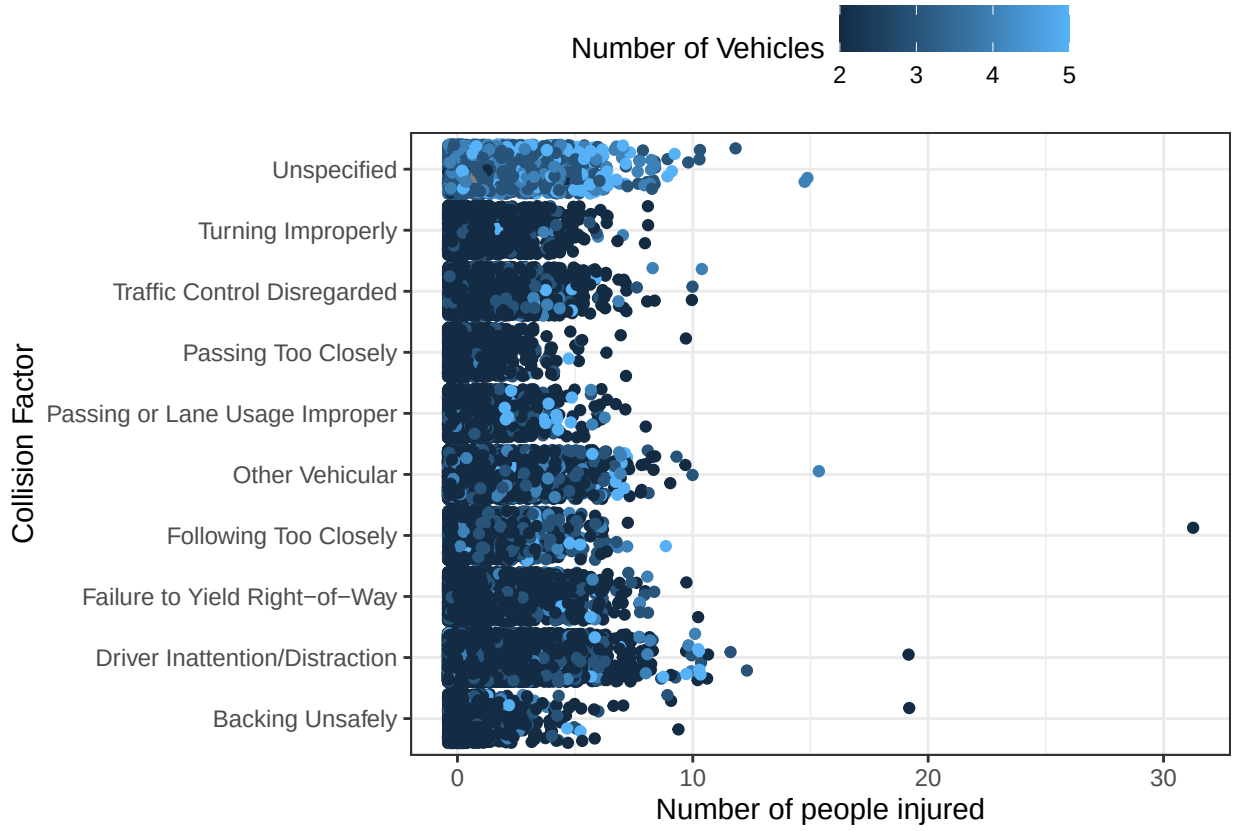
This data is also supported by a summary table of the CRASH.DATE variable, displaying that 75% of these notable collisions occurred before 2020, despite 75% of the distance between 2012-2025 being mid to late 2021.

Another characteristic to note relating to the ridgeplot is the bottom 5 entries. Taking note of the sharp decrease in these classes at ~2016, as well as the sharp increase in some classes around 2016 displays a substantial flaw in this data set. The cause is the change in reporting, relating to Vision Zero, which was implemented starting 2014. The recording of vehicle type codes was changed to be standardised, and since only the top classes were combined in the data tidying, this inconsistency remains.

Since the raw amount of unique data entries in this data set is 478 (including typos, capitalisation differences etc), the recommendation to make this data set reliable and practical is to simply disregard data before 2016, as extensive manual tidying is required to standardise vehicle type codes.

Addressing the second question, 'What are the common attributes of a serious collision?', we turn to the Motor_factor_10 data frame.

```
ggplot(Motor_factor_10)+
  geom_jitter(aes(x=NUMBER.OF.PERSONS.INJURED, y=Factor, col = NO.VEHICLES), position = 'jitter')+
  theme_bw()+
  labs(x="Number of people injured", y="Collision Factor", col = "Number of Vehicles")+
  theme(legend.position = "top")
```



The general trend that this plot shows is that as the number of people increases, so does the number of vehicles involved, with a few notable exceptions. The factor 'Unspecified' is almost entirely comprised of crashes involving above 3 cars. The cause of this is most likely that as the collision becomes more chaotic, it becomes increasingly difficult to ascertain the root cause. Additionally, despite having only 2 vehicles involved in the top 2 observations, both have injury counts over 18. To explain this, we extract a summary table of the frame, only including observations involving bus vehicles:

```
Bus_factor<-filter(Motor_factor_10, Vehicle=="Bus")
table(Bus_factor$Vehicle, Bus_factor$NUMBER.OF.PERSONS.INJURED)
```

```
##
##            0    1    2    3    4    5    6    7    8    9   15   19   31
##   Bus 3473  494  119   52   27   14    6    5    4    1    1    1    1
```

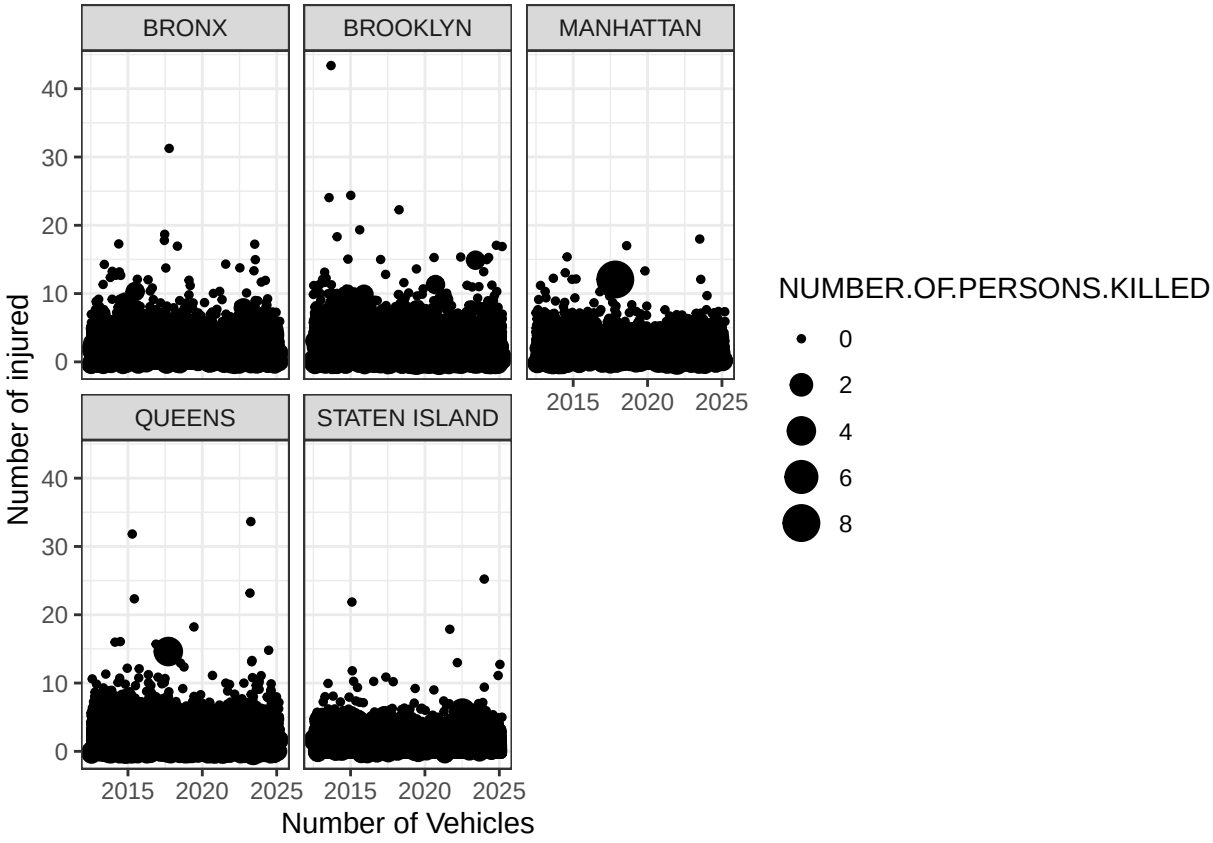
Therefore, we can conclude that the reason for the egregiously high injury count, despite the low vehicle count is due to the involvement of buses.

Observing the factors themselves, the results produced are intuitive in a contextual sense. Factors that have low injury counts are those that one would expect to happen in a slow or high traffic situation (backing unsafely, passing too closely, lane usage improper), while factors that may occur in higher speed situations (driver distraction, traffic control disregarded) have much higher injury counts.

Regarding the final question, 'Which borough is the most "dangerous" in regards to driving?', we may plot the number of people injured against time, whilst separating boroughs, and displaying the number of people killed.

```
ggplot(Motor_crash_better)+
  geom_jitter(aes(x=CRASH.DATE, y=NUMBER.OF.PERSONS.INJURED, size = NUMBER.OF.PERSONS.KILLED))+
  facet_wrap(vars(BOROUGH))+
  theme_bw()+
  labs(x="Number of Vehicles", y="Number of injured")
```

```
## Warning: Removed 4 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Whilst this plot may display the general density of crashes, it does not give a very helpful indication of the severity of crashes. Filtering this to remove cases where there were zero fatalities gives the following:

```
ggplot(filter(Motor_crash_better, NUMBER.OF.PERSONS.KILLED>0))+
  geom_jitter(aes(x=CRASH.DATE, y=NUMBER.OF.PERSONS.INJURED, size = NUMBER.OF.PERSONS.KILLED))+
  facet_wrap(vars(BOROUGH))+
  theme_bw()+
  labs(x="Date", y="Number of injured")
```

From this plot, we are able to see that Staten Island is most likely the least dangerous in terms of road fatalities, whilst Brooklyn has not only a higher abundance of crashes, the crashes usually have much higher fatalities than other boroughs.

Observing other studies, particularly one conducted between 2013 and 2023 finds similar results. Mittal and Lim (2024) concludes that Brooklyn accounts for most of the crashes due to its "heavy traffic congestion, complex street network, frequent construction", while Staten Island represents a much lower percentage of accidents, "due to its lower population density and reduced traffic volume".

In this regard, we are able to ascertain a majority of this information from the data, with reliable results that reflect academic studies.

Conclusion

While the prospect of using such a large data set may seem enticing, due to the larger pool from which statistics can be gathered, there are many detriments in the formation of this data. The collection of extremely large data sets can prove to be more surface level than dedicated studies, or the method of collection may present inconsistencies in the data, as a result of the large coverage. In addition, the transformation of such data for usage in analysis may present further challenges.

Such data as discussed above presents its use in supporting dedicated studies, or for a broader analysis of the related interests. While the research questions were able to be answered sufficiently, it is only through the comparison to existing data that those findings are vindicated.

Bibliography

Haddon, William, and James R. McCarroll. 1963. "A CONTROLLED STUDY OF FATAL AUTOMO BILE ACCIDENTS IN NEW YORK CITY*." *Journal of Chronic Diseases* 15 (8): 811–26.

Mittal, Vikram, and Elliot Lim. 2024. ""Patterns and Analysis of Traffic Accidents in New York City Between 2013 and 2023." *Urban Science* 8 (4).

NYC-Open-Data. 2025. "Motor Vehicle Collisions - Crashes." https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes.

posit. 2024. "Rmarkdown :: Cheatsheet." https://rstudio.github.io/cheatsheets/html/rmarkdown.html?_gl=1*1yhug4d*_ga*NTQ2ODc3NjE2LjE3NDU4ODA3Nzg.*_ga_2C0WZ1JHG0*czE3NDcyNzE1MT%0A%20%20kkbzUkZzEkdDE3NDcyNzE1MjMkajAkbDAkaDA.

Xie, Yihui and Dervieux, Christophe and Riederer, Emily. 2025. "R Markdown Cookbook." https://bookdown.org/yihui/rmarkdown-cookbook/source-script.html.