Prelim Report STAT373

Riley McIvor

11/4/2025

Abstract

github repository: https://github.com/Derboshr/Assignment-Repository

Contents

Introduction

Background

The topic of interest is that of New York vehicle collisions. When studying this data set, 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 controlled 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 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.

Research Question

In this report, analysis will be performed in order to investigate the dynamics of New York Boroughs regarding motor vehicle collisions. This will be broken down into several questions.

- Which New York Borough is the most “dangerous”?
- How have the dynamics of vehicle collisions changed over time?
- What are the most common attributes of a serious crash?

Additionally, we will examine if this data set, and the transformations on the data, are useful and/or practical for answering these questions.

In answering these questions, it would be enlightening to make comparisons with existing results, and thus creating an answer for the final question. 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 provide suitable analysis, some exhaustive data wrangling must be performed. The first step in analysis will be filtering the data heavily:

1. Creating a new data frame, which has been filtered to remove observations that have blank entries in vehicle type codes and contributing factor*.
2. Selecting 14* key variables, to remove unnecessary cluttering in the data frame.
3. Mutating to create a new variable, number of vehicles, which displays how many vehicles are involved in a collision
4. Further filtering on the frame to remove observations with blank spaces in Borough entries, as well as fixing the date variable so that it is recognised by R as a date.

Further, the data frame contains thousands of unique observations for vehicle type code, so making a frame that only includes the top 20 and top 5 vehicle most frequent vehicle types would aid in analysis. To do this courtesy of Schork (2021), a sorted table of the vehicle types must be made, which is then made into a tibble, as the slice function can not be done on tables. Once the tibble is sliced for the top 50 entries, unique identical entries (Taxi and TAXI for example) are combined, and new frames including the combined top 20 and top 5 vehicles are made.

Another data frame is to be created in which the data is pivoted longer with respect to the 5 vehicle type codes, so that each crash is represented in 2-5 observations. This will allow for each observation to have an individual contributing factor*.

From this, plots can be made with ggplot. The first planned plot is a point plot showing the relationship between number of vehicles and number of people injured, faceted by Borough, and coloured by vehicle code.

The second plot is a ridgeplot, showing crash density throughout time, by vehicle type.

These two plots in particular will be vital in answering the research questions, along with relevant summary statistics when required.

Moreover, plots can be made to show the relationships of contributing factor with other key variables*.

[*There may be a separate frame in which contributing factor is included, as well as an original where it has been deselected. Many observations do not have entries in contributing factor, and if the filtered version has too little observations this idea will be abandoned. The current version has selected 9 variables, not including contributing factor, but analysis on this could prove to be interesting]

Conclusion

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.

Schork, Joachim. 2021. "Extract Most Common Values from Vector in r (Example)." https://statisticsglobe.com/extract-most-common-values-from-vector-in-r.