An Analysis of NYC Motor Vehicle Collision Data

Patrick Hutecker
Harvey Mudd College
Claremont, CA 91711
phutecker@g.hmc.edu

Abstract

The NYC Motor Vehicle Collisions data set is a rich data set describing a variety of road accidents in New York City from 2012 to 2024. This analysis offers an insight into the shape of the data and its basic features. It also includes an introductory analysis of the data's features as well as the patterns which we begin to see in the data.

1 Data Set Background

In 1999, the New York City Police Department implemented the Traffic Accident Management System or TAMS. TAMS was the first standardized program to collect traffic accident data in NYC. Due to the success of other similar programs, NYCPD sought to collect - and eventually analyze - traffic collision data. The larger goal of the program was to improve traffic safety in NYC and to hopefully reduce the severity and frequency of accidents in the city. As the years went by - the data gathering techniques of the department were only improved and data was gathered at larger and more detailed scale.

The current data set is the "Motor Vehicle Collision crash table." A row of the table corresponds to a singular crash or accident. However, not all vehicle accidents in NYC are reported in the table. To qualify, an accident must have an injury, death, or at least 1000 dollars worth of damage.

Like the NYCPD - the goal of this paper's analysis is to gather patterns and information which could benefit traffic safety in NYC.

2 Initial Exploration

The data set was initially acquired as a `.csv`. To decrease the load time - I converted the file to a `.parquet`. However, there does not seem to be a significant improvement - as the load time was already trivial to begin with.

The data set houses **2,069,104** rows or collision entries which is well above our requirement of 500,000 data points. All fields in an entry fall into one of the 5 categories: date and time, location, cause of accident, injuries / death, and vehicle information. Bellow - I detail the initial exploratory analysis I did for each category.

2.1 Date and Time

For each collision, we are provided with the date and time. To explore the data - I first plotted the number of entries or collisions by the year they were recorded. The time range of collection spans from 2012 to 2024 and can be seen in Figure 1.

Although the year of the data is not of particular concern - it is beneficial to know that a majority of the data is from before 2020. There could be a variety of reasons for this. Perhaps there has already

Preprint. Under review.

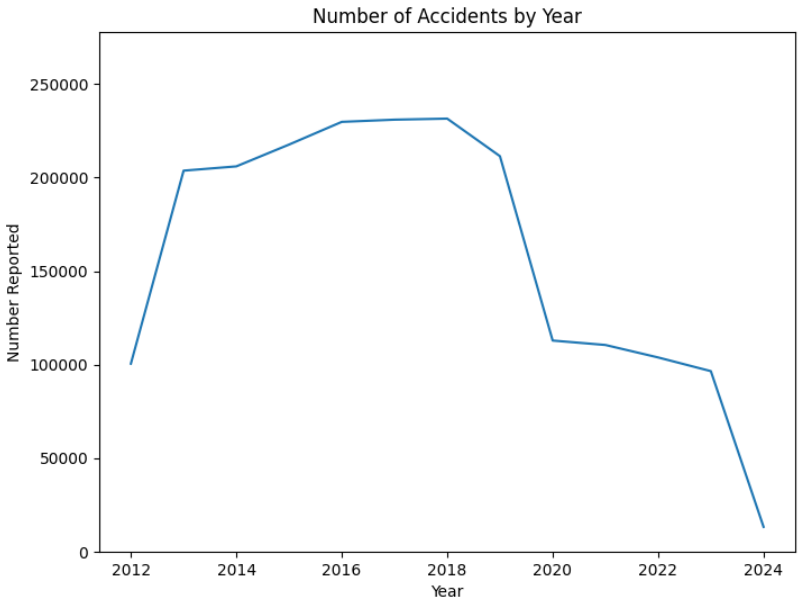

Figure 1: Number of accidents reported in the dataset by year.

been an improvement in traffic safety in NYC or less people are driving. COVID could have had an impact on these statistics.

I then looked into the number of collisions grouped by the day of the week. The graph for this data is in Figure 2.

There is a clear spike on Fridays and a drop off on the weekend. Friday marks the end of the workweek and is generally the day of the week where the most number of people drive. Furthermore, on weekends people generally do not need to commute to work. These results are generally in line with what is expected.

2.2 Location

To analyze the location, I plotted the number of accidents by boroughs. In NYC, boroughs are a good indicator of social class, race, population density, and culture. In Figure 3 we see the results.

The graph is not uniform. Staten Island has far less entries than the other boroughs. It could be worthwhile - in future analysis - to split the analysis by borough. Because NYC is large to begin with - different boroughs sometimes operate as cities of their own - with their own qualities and patterns.

2.3 Cause of Accident

The data set provides a cause of accident, but does not follow a uniform labelling system. Because of this - a lot of fields were lost and had to discarded in the analysis. A better parsing / comparison system would allow us to retain more data. Figure 4 showcases a bar graph of different causes of accident.

The most prominent field is "Driver Inattention/Distraction". Although this is a vague - it is very distinct from something like road rage - where a driver can be attributed to having some intent in causing the collision. Other interesting factors are "Backing Unsafely" and "Failure to Yield Right-of-Way". These can be interesting from an educational stand point - as when training and educating drivers one can emphasize these weak points.

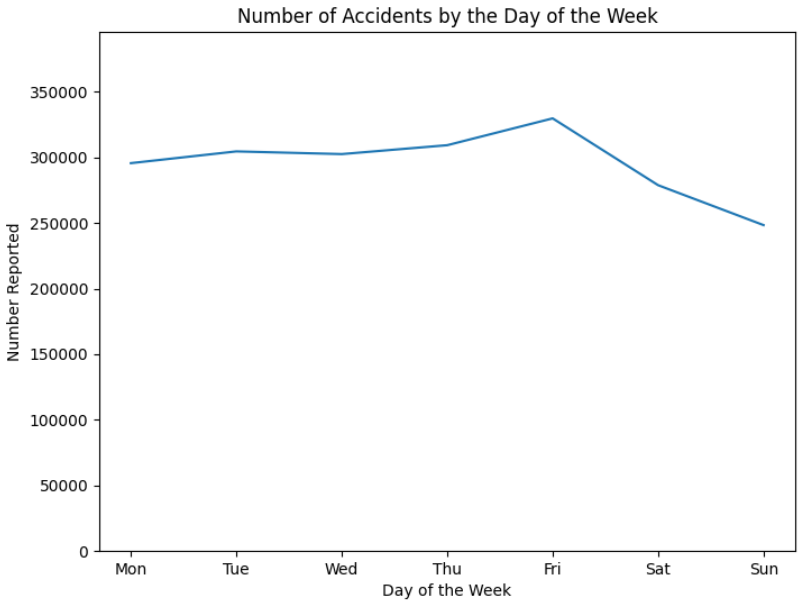

Figure 2: Number of accidents reported in the dataset by the day of the week.

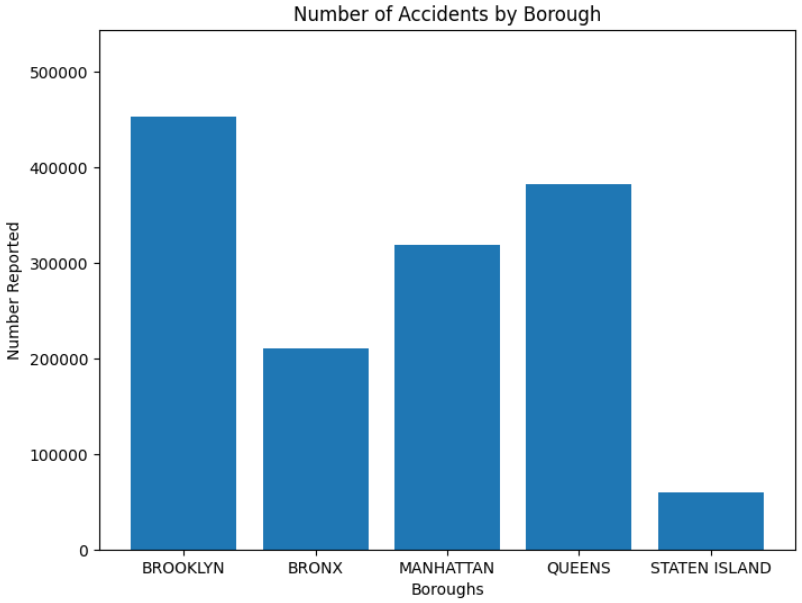

Figure 3: Number of accidents reported in the dataset by the borough.

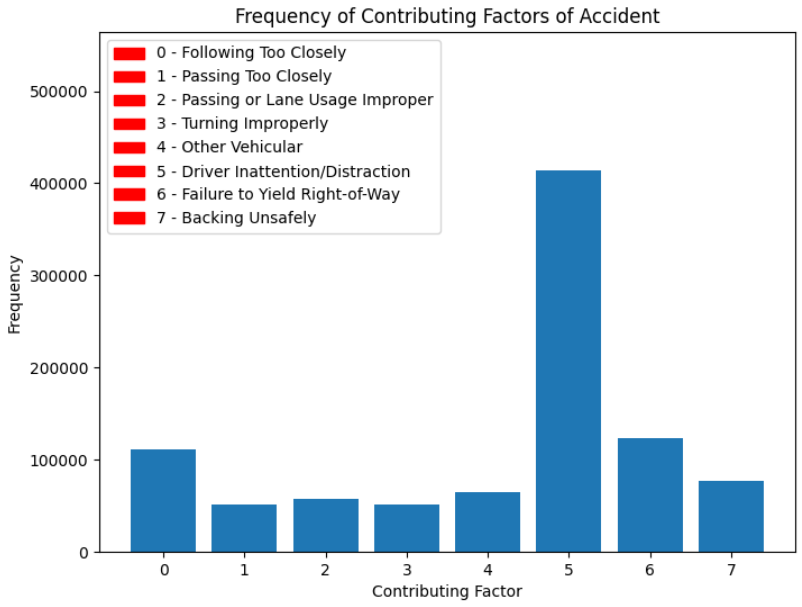

Figure 4: Number of accidents reported in the dataset by the cause of the accident.

Furthermore, the causes of accident are extremely useful for prediction models - as different behaviors - such as "Backing Unsafely" - can be much more likely to affect pedestrians - and can thus be useful for predicting the severity or nature of accidents.

2.4 Injuries and Deaths

Injuries and death are another key field in the data set. Both are measured with respect to persons, pedestrians, cyclists, and motorists. Therefore, the injury and death data can be seen as four dimensional vectors - where each dimension corresponds to persons, pedestrians, cyclists, and motorists respectively.

At first - I disregarded these vectors and simply considered how many people total were injured or died in an accident. Figure 5 showcases the frequencies of these two.

We see that it is quite uncommon for people to be killed in an accident and especially uncommon for multiple people to be killed. Injuries are far more common, but also has a steep drop off past 1 person injured.

This indicates that identifying what accidents result in deaths is similar to identifying outliers in a data set - but - because of the small sample size - noise might make such identification impossible.

Out of all the fields - I felt that these two were the most important - as understanding what kind of accidents cause injuries and deaths could help in the reduction of them.

To look into these fields more closely, I considered the four dimensional vector representation of injuries and deaths. I then applied a k-means algorithm on each respective field to see if there were any common flavors of injuries and deaths. For example, it may be the case that whenever a cyclist is injured so is a motorist.

The results of this analysis were fairly trivial. Out of 2 million recorded accidents, the major groups were the following: 1,600,000 no injuries / deaths, 200,635 Motorist injured, and 111,329 Pedestrian injured. The result then was a grouping of trivial basis vectors - not anything interesting like the aforementioned cyclist and motorist pairing.

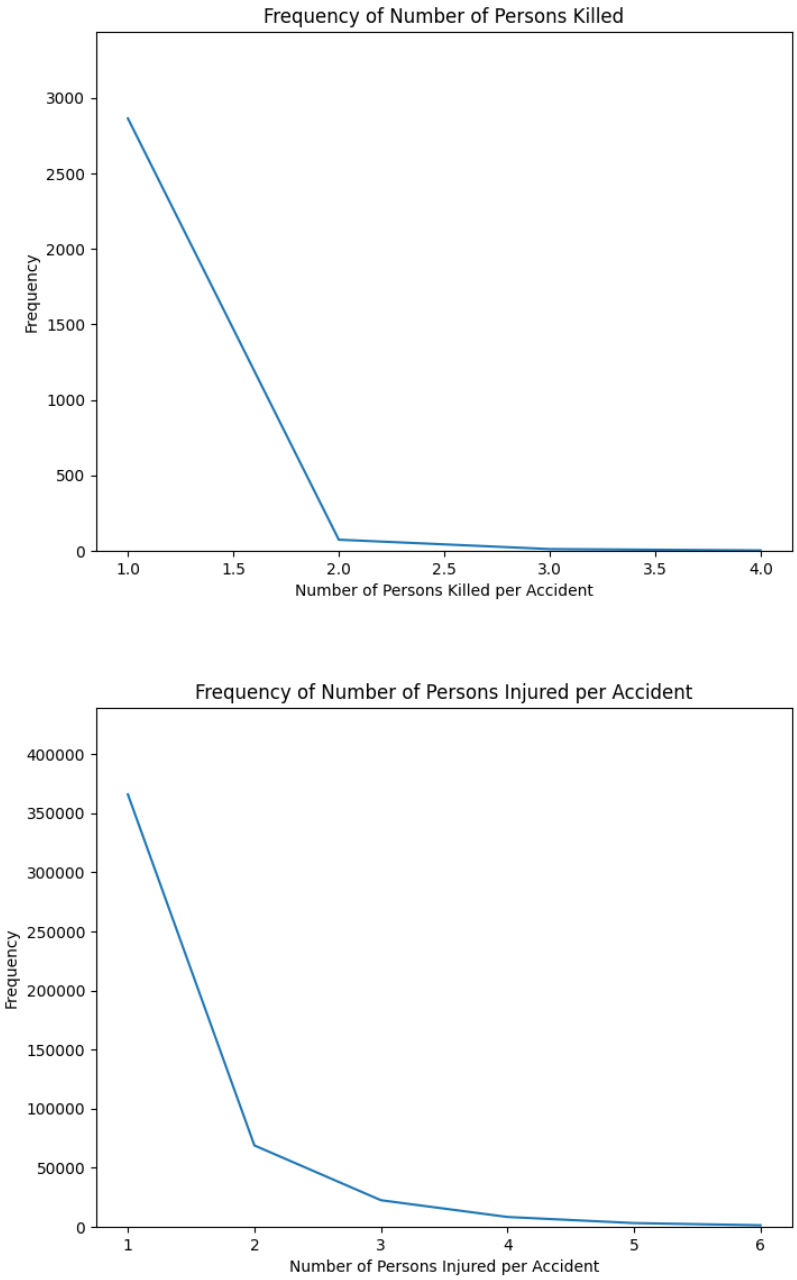

Figure 5: Number of accidents reported in the dataset by deaths and injuries.

3 Future Exploration

The initial exploratory analysis generally served as a survey of the data set. The analysis was very limited to each of the described categories and fairly straightforward.

In this section, I describe steps I could take to further the analysis of different categories - as well as how I might begin to consider categories in conjunction.

It also serves as a space to brainstorm - as there are many avenues for the analysis and many subsets of the data which could result in insights.

3.1 Date and Time

Although we looked at yearly and day of the week statistics - we could also see if there is a correlation between accidents and the time of the day. We can expect there to be a non-trivial correlation as throughout the day we have factors such as rush hour or decreased visibility during the night.

It might also be interesting to look at the Friday spike. Perhaps certain car models or accident types are more frequent on certain days of the week. Take weekends - where we expect more people to be on the streets enjoying their free time - rather than being at work.

It is not unreasonable to expect the time of the day or the day of the week to have an effect on accidents and collisions.

3.2 Location

The data set provides extensive locational data. Besides Boroughs, there are fields such as longitude and latitude and the street name where the collision occurred. These could be looked into more detail but would be difficult to work into prediction models - as the longitude and latitude requires high precision and as there are thousands of streets in NYC.

3.3 Cause of Accident

A much better analysis is needed for the Cause of Accidents. At the moment - the parser matches strings exactly. However, it might be worthwhile to import some language model to group causes of accident with the same meaning but different syntax's.

As mentioned - the cause of the accident is very important in determining the nature of the accident. Comparing the cause of accident with other fields is also of great importance - such as the Borough - day of the week - and even Vehicle Model.

3.4 Injuries and Deaths

Because the k-mean analysis did not yield anything interesting - it could be worthwhile to repeat the analysis but with more fields in consideration. For example, we could also consider the day of the week of an accident - meaning we would have a five dimensional vector - where the day of the week is the fifth dimension and the other four remain the same (persons, pedestrians, cyclists, and motorists).

This could tell us if deaths or injuries are more common on days like Fridays or weekends. Information like this could be useful in tuning the cities traffic safety by the day of the week.

It is also important to note that Injuries and Deaths can serve as a severity indicator of the accident. Accidents with more injuries or deaths are objectively more severe - and can thus be handled differently in analysis.

3.5 Vehicle Information

I was unable to include the vehicle information due to a lack of a good parser. Vehicle Information was much like Cause of Accident - in that the description of the vehicle varied even if the same vehicle was being described.

Therefore - it is of great importance that I implement a good parser - so that I can begin to group the Vehicle Information and include it in my analysis.

Vehicles can differ in size, speed, and breaks. Therefore - it is logical to expect the vehicle information to affect the severity of a collision. For example, a big truck has much more difficulty stopping than a sedan and thus could proportionally be in more accidents - as well as more severe ones.

4 Conclusion

The Motor Vehicle Collisions data set is very rich. There are a variety of provided fields each of which have their part and importance in describing an accident.

To continue my analysis, I wish to select a subset of the fields and categories and begin a closer comparative analysis between the subset. A covariance matrix might be useful in deciding this subset.

Furthermore, it might be worthwhile to work towards a predictive model - which could predict things like the severity of the accident (injuries and deaths) - as the model could then serve as an extrapolation tool or rather for measuring the danger of certain roads or boroughs which are not necessarily in the data set.

Regardless - for future analysis I wish to identity "hot spots" in the data. What roads are the most dangerous? Perhaps what times of the day? Or perhaps what Boroughs? Do previous accidents seem to affect later ones? Are certain vehicles more likely to be in an accident?

These questions and many more are of importance as the goal of this analysis is to clearly identify issues in NYC road safety. In doing so - we can hope to better the road safety.

References

[1] New York City Police Department (2024) *Motor vehicle collisions - Crashes - catalog*. https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes