Analyzing Motor Collisions Data in New York City

Zijia Hu, Ken Yokokawa, Matthew Zhang

December 11th 2020

Abstract

This paper investigates different factors that surround motor vehicle collisions in New York City with the purpose of identifying risk factors so that future government policy and city planning can reduce such collisions. In our study, we analyzed datasets for vehicle collisions, traffic volume, and construction permits based on their location. We set the number of motor vehicle collisions in New York City as our dependent variable, and investigated its correlation with two independent variables: traffic volume and construction permits on roadways. We further tried to analyze the collision rate as an addition to collisions. However once we dove deeper into looking at specific intersections we were able to determine some other trends that could not easily be reflected in quantitative data.

Through quantitative analysis, we found a positive correlation between permits and collisions, and a negative correlation between traffic volume and collision rate. In our qualitative analysis, we found a trend that high collision intersections often have complicated layouts and merges that may be disorienting to drivers. These findings can help city planners design intersections that contribute to less collisions and benefit the residents of New York City, whether it be drivers, pedestrians, or cyclists.

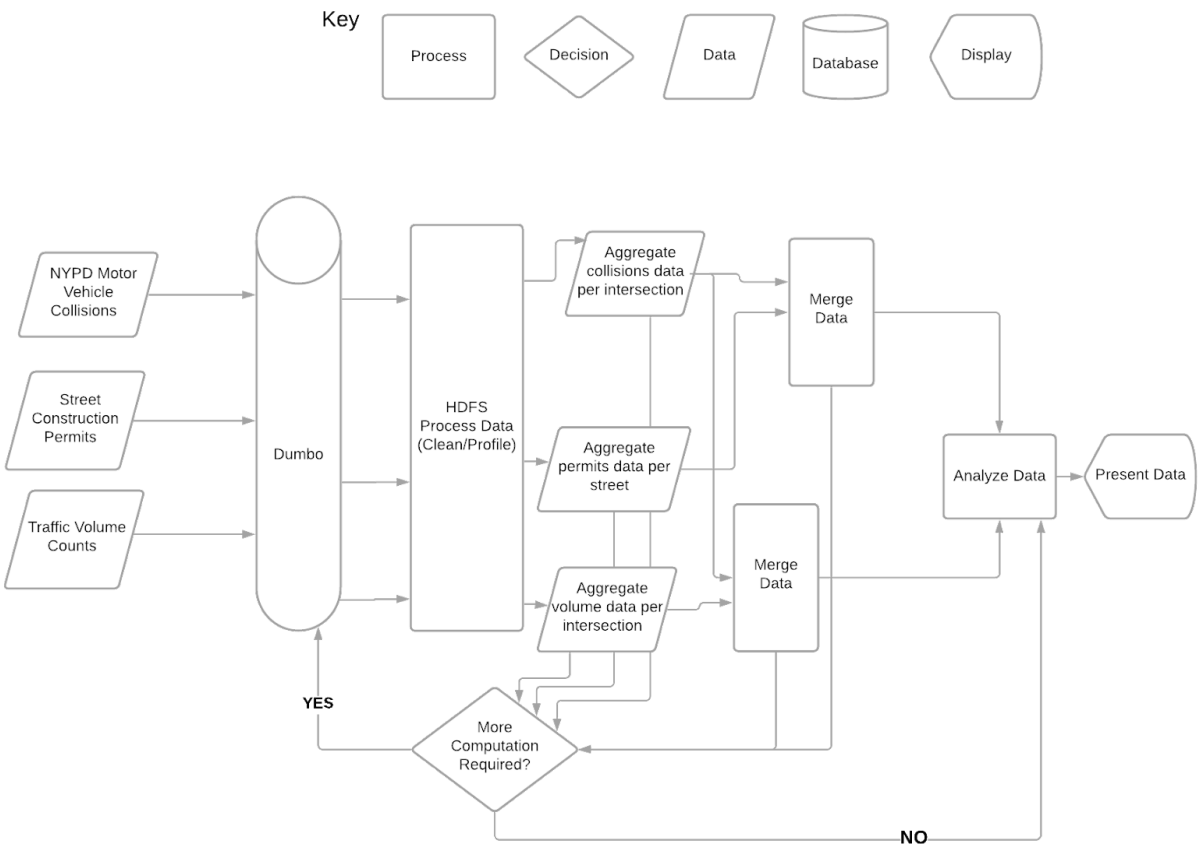
Introduction

Our team studied different factors that intuitively might cause higher motor vehicle collision rates in New York City. This topic was important to investigate because we knew in New York City people generally use four main modes of transportation: walking, cycling, riding in a car or riding on the subway. The first three modes often take place in close proximity to each other. Bikes and cars share the same street in many instances and perhaps even the same lanes, and many people walk on sidewalks next to busy streets. The streets themselves service a disproportionately large amount of people compared to smaller cities. As a result, motor vehicle collisions can have a larger than normal impact due to how transit is done. Reducing these collisions is generally important, but has an elevated urgency in the context of New York City for these reasons. Our study advances the body of knowledge on this subject by showing the exact significance of these common quantitative factors the city government has robust data on, and takes a deeper dive into the specifics of qualitative factors that the quantitative analysis guided us towards.

When analyzing the chosen datasets, our data was first downloaded from the City of New York website and put into New York University's Dumbo node cluster, specifically within the

Hadoop distributed file system. Hadoop MapReduce was used to clean and profile each dataset. Then the dataset with our independent variable was joined with both dependent variables individually using Hive. Finally, we used Hive to do introductory data exploration, and Spark to do linear regression.

Data Flow Diagram



Motivation

New York City is the most populous city in America and faces unique challenges with regards to transportation. There is a significant amount of motor vehicle accidents that happen relatively frequently which can have massive health and economic impacts. This paper is motivated from these concerns and attempts to reduce the frequency of accidents by analyzing potential factors that may correlate with these accidents. Our team is composed of three New York University students who have first hand experience with transit challenges in New York City, and we were drawn to this project as we wanted to make the city we lived in safer for all residents.

In New York, driving a motor vehicle is a means of transportation or even a means of livelihood for many people, so car accidents are among the top causes of national death. As we all know, traffic accidents often cause heavy loss of life and property. So, the reason for me to choose this topic for our analytic project is that I found it would be interesting to dive deeper into analyzing the factors that correlated to the vehicle accidents. Because analyzing the data to see what would be the factors that are correlated to the cause of vehicle accidents might give me a new perspective to view the cause of vehicle accidents or change my original thought on the reasons behind these accidents.

Related Work

There has been extensive research for motor vehicle safety, although we couldn't find research that analyzed New York City traffic specifically, and most research seemed to be on a much smaller sample size.

The paper "Effects of Traffic Volumes on Accidents: The Case of Romania's National Roads" by Rodica Dorina Cadari , Melania Rozalia Boitori, and Mara Dumitrescu analyze Roads' accident rates based on their traffic volume calculated by vehicles per day. The study focused on seven national roads in Romania and found using regression analysis that there was a positive correlation between a road's accident rate and its traffic volume. "Relationship Between Traffic Volume and Accident Frequency at Intersections" by Angus Eugene Retallack and Bertram Ostendorf draws takes a slightly different approach, where they looked at each intersection's accident rate based a traffic volume that changed throughout the observation period and related the accident rate to a time local traffic volume of the intersection. The study also found a positive correlation between a road's accident rate and its traffic volume. A big portion of our analysis was done with these two studies in mind, producing similar metrics from our data that would help us run a regression analysis that would produce a similar relationship. However, we found a negative correlation between a road's accident rate and its traffic volume that diverged from the findings of the two papers.

"Roundabout Safety Experience" by Philip Weber examined accident rates and severity to find ways that roundabouts could be designed and constructed to improve safety. We used this paper as a basis for our qualitative analysis that focused on looking at high-collision intersections and looking for possible ways in which the intersections could be improved to reduce safety.

Description of Datasets

The traffic volume counts dataset is provided by the New York City Department of Transportation. It contains traffic volume data per street segment from December 18th, 2018 to August 4th, 2020. The data schema is as follows: the first column is the Count ID which is an integer, the second row is the Segment ID which is an integer, the third column is the Roadway Name which is a string, the fourth and fifth columns are named "From" and "To" respectively

and are strings, the sixth column is the direction and it is a string, the seventh column is a date and it is a string, and the eighth to thirty first column are times, representing each hour of the day and are integers. Going into more detail, the ID column gives each row a specific ID, the segment ID gives each street segment a specific ID, the roadway name is the road the street segment is on, and the from and to columns give the streets that bound the road segment. For example, if a street segment is on road A, the from and to columns could be roads B and C which would mean that B and C intersect with road A, and the shortest distance between the intersections is considered the street segment evaluated. The direction column indicates the compass direction of the street segment, the date column represents the specific day that volume was being counted and the rest of the time columns give the volume counts for each hour segment of the day.

The traffic collisions dataset is provided by the New York City Department of Transportation and collected by the New York Police Department. It contains all available information on vehicle collisions reported to the police from July 2nd, 2012 to Present (the most recent recorded collision was on October 31st, 2020). The data schema is as follows: the first column and second columns are the crash date, and crash time respectively. Columns three through 10 describe the location of the crash, with column three containing Borough, column four containing zip code, and columns five, six, and seven containing latitude, longitude, and coordinates (the pair of columns five and six), respectively. Columns seven and eight contain the cross streets of the collision, although if the collision occurred too far from an intersection or off of the road, those columns are left blank and column 10 “off street name” may be filled in instead. The columns for location are not always filled completely, so many accidents do not have zip codes, coordinates, or street names associated with them. Columns 11 and 12 describe how many people were injured and killed in the collision, respectively, and columns 13 through 18 provide further detail by separating the people into pedestrians, cyclists, and motorists. Columns 19 through 23 contain “contributing factors” for the collision, which are descriptions such as “Failure to Yield Right-of-Way.” Column 23 contains the collision’s unique “Collision ID,” and columns 24 through 29 contain the types of vehicles involved in the collision, such as “sedan.”

The street construction permits dataset is provided by the New York City Department of Transportation. It contains all available information on over 150 different types of sidewalk and roadway construction permits from December 17th, 2018 to Present (the most recent recorded permit was on December 10th, 2020). The data schema is as follows: The first column is the description of the application type which is a string, the second column is the description of the permit series which is a String, and the third column is the secretion of the permit type which is a string. The fourth, fifth, sixth, eleventh, and twelfth columns contain a date and are strings. The seventh column is the borough name which is a string, the eighth column is the street name, the eighth and ninth columns are respectively the comments on permit purpose and permit location. Going into more details, in the original dataset, there are three columns called OnStreetName, FromStreetName, and ToStreetName. We split every row into two, one with the OnStreetName

and FromStreetName combined as one column called StreetName, separated by '/', and the other with the OnStreetName and ToStreetName combined as the StreetName.

Analytics Stages and Process

The data ingestion process involved downloading each dataset from the City of New York data repository website and then uploading it onto Hadoop distributed file system which is hosted on New York University's dumbo cluster.

For the volume dataset, it was cleaned by first removing columns, specifically the ID, Segment ID and Direction columns. Then the columns that held volume counts for each hour of the day were aggregated to one column that displays the day's volume count for each row. There were some technical difficulties in this, mainly the fact that counts over 999 have comma separated values and quotes around the string which represents the number. As a result the Java parse integer function could not take in the string and convert it to an integer. To solve this problem, we sanitized these strings by removing the commas and quotation marks. To profile the dataset, we ran a count on both the original and the cleaned dataset to make sure they both contained the same amount of words. For our analytics, we created an aggregate dataset where we split the data into intersections and averaged repeat recordings of each intersection. This was the dataset format we used to join with the dataset that contained the collision counts. Then we used Hive commands to get the top ten busiest intersections from the aggregate dataset, and for the joined dataset we used Spark's regression functionality to get a linear regression between the volume and collisions for each intersection.

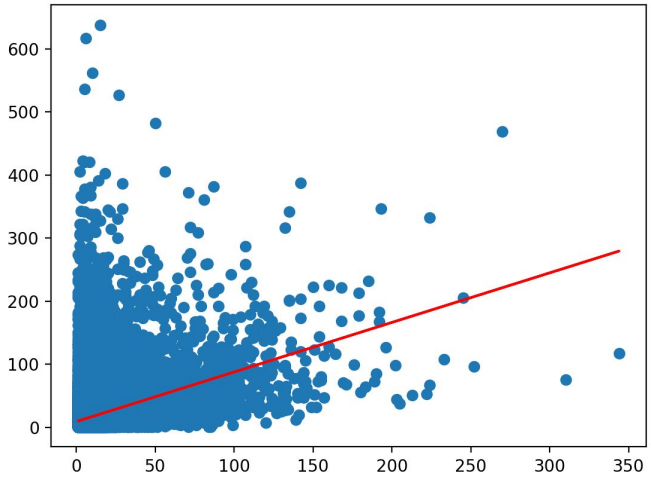
The collision dataset was cleaned by extracting rows for date, cross streets, borough, number of people injured, number of people killed, and time from the original dataset. Profiling the dataset, we counted the number of rows for the original and cleaned dataset to compare the amount of rows: the cleaned dataset had three fewer rows since those rows were not formatted correctly and the columns could not be properly extracted. For our analytics, again we aggregated this dataset similarly to the volume dataset where we aggregated the total number of collisions as well as injuries and deaths resulting from those collisions by the intersection they occured on. From the aggregated dataset, we used Hive to get the top ten intersections in terms of collisions for further qualitative analysis. We also used Hive to join the aggregate collision dataset with the aggregate volume and permit datasets for the regression analysis mentioned for the volume dataset.

The permit dataset was cleaned by extracting rows for ApplicationTypeShortDesc, PermitSeriesShortDesc, PermitTypeDesc, PermitIssueDate, IssuedWorkStartDate, IssuedWorkEndDate, BoroughName, OnStreetName, FromStreetName, ToStreetName, PermitPurposeComments, PermitLocationComments, CreatedOn, and ModifiedOn. To profile the dataset, we convert the format of every column that contains the date in the selected columns to match the format of date in the other two datasets. Also, we split every row into two, one with the OnStreetName and FromStreetName combined as one column called StreetName, separated

by '/', and the other with the OnStreetName and ToStreetName combined as the StreetName. To compare if the number of rows is accurate, we counted the number of rows for the dataset after row splitting and the original dataset. We multiplied the number we got from counting the number of rows in the original datasets by two to see if it matches the number of rows in the row-split dataset. For our analytics, we aggregated this dataset similarly to the other two datasets where we aggregated the total number of permits by the street name they occurred on. From the aggregated dataset, we used Hive to get the top ten intersections with the most permits issued for further qualitative analysis.

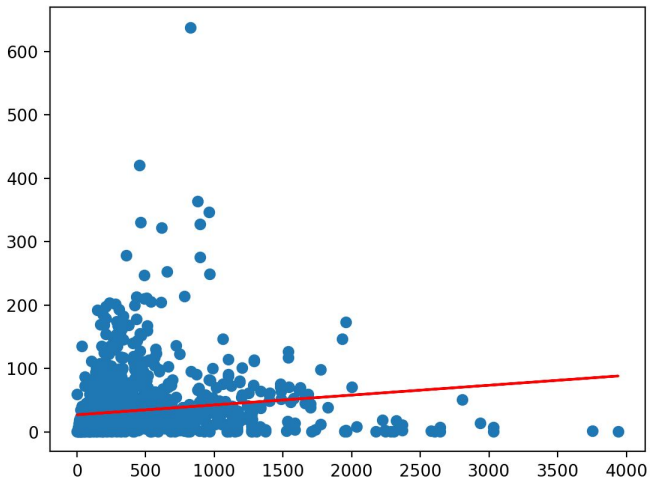
Graphs

Collisions vs. Permits



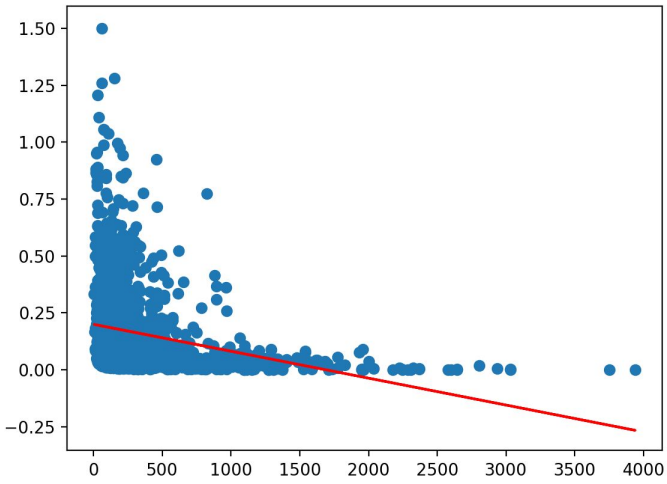
Using Spark to regress collisions on the number of permits for each intersection, we got an equation of "Collisions = 9.48 + 0.76*Permits". For a summary of the model, the Root Mean Squared Error was 23.79, the R Squared of the model was 0.14. From the graph and the regression equation, we can see a clear positive correlation between the number of permits that an intersection has and the number of collisions.

Collisions vs. Traffic



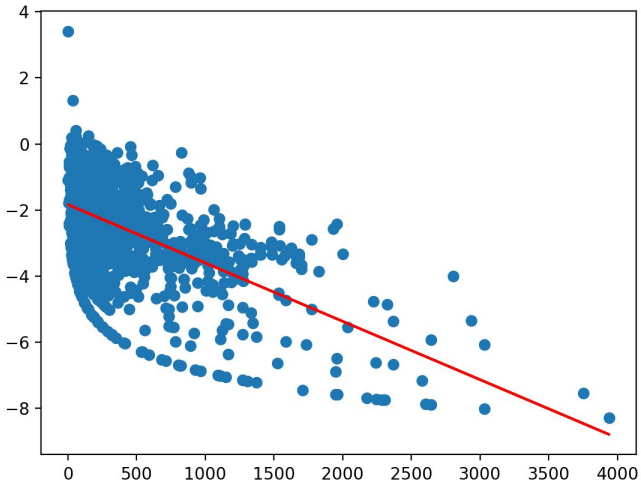
Using Spark to regress collisions on the traffic volume for the intersection, we got an equation of "Collisions = 27.30 + 0.0149*Volume". For a summary of the model, the Root Mean Squared Error was 42.01, the R Squared of the model was 0.02. From the graph and the regression equation, we can see a weak correlation between the traffic volume of an intersection and the number of collisions. We expected the correlation to be much stronger and the regression equation to have a higher coefficient, since the size effect would lead us to predict that a street with more traffic would have more collisions if the collision rate of the street was kept constant. To better understand how collisions relate to traffic volume, we focused our analysis on the collision rate of an intersection (the total number of collisions divided by the traffic volume) rather than the number of collisions.

Collision Rate vs. Traffic*



Using Spark to regress collisions on the collision rate on permits for each intersection, we found a statistically significant relationship between the two variables. However as we can see from the graph above, the relationship looks far from linear so we took the natural log of the collision rate to perform a linear regression that fit better.

ln(Collision Rate) vs. Traffic*



Using Spark to regress the natural log of the collision rate on the traffic volume for each intersection, we got an equation of "ln(Collision Rate) = -2.04 + -0.00114*Volume". For a summary of the model, the Root Mean Squared Error was 1.10, the R Squared of the model was 0.28. From the graph and the regression equation, we can see a notable negative correlation between the traffic volume of an intersection and the collision rate, with the correlation being strongest for streets with traffic volume of less than 1,500. The interpretation of the regression

model may seem counterintuitive (and contradicts the finding of the two papers discussed above): streets with higher traffic volumes are associated with lower collision rates - meaning that cars driving in busier streets are less likely to be involved in collisions.

Conclusion

Our research shows that there is a positive correlation between the number of permits and number of collisions, and there is a negative correlation between volume and collision rate. While the fit of each linear regression may not be very good, the correlation is still evident when looking at the graphs. This suggests that the New York City government should improve upon safety measures around construction sites and on lower volume roads that have high collision rates which may signify other issues with the road itself. This leads onto an important idea. When we tried looking at a more qualitative factor like how difficult a merge or intersection was to navigate, we were more successful in finding potential issues. Instead of looking directly at the currently available quantitative data which might not be able to capture factors that are harder to measure like the difficulty of a merging lane, there should be an effort to analyze these more qualitative factors. This may require quantifying them since the sets of data generated may be so massive that aggregation tools might be the only reasonable way to take into account all the data and reduce complexity. Future research may be able to use this qualitative approach to get more actionable insights, although there is a tradeoff with objectivity and ease of data collection and analysis.