NYC Crash Report

Authors

Kyle Rushing

Hayden Mann

Emails

kwr11@students.uwf.edu

hem21@students.uwf.edu

CAP 4770

11/25/2025

1. Abstract:

Over the years NYC has experienced a consistent increase in vehicular crashes. These crashes are caused by either human error or roadway design. This paper will go into detail on potential problem areas and characterization of these crashes. The dataset we are using is a cleaned version of the NYC “Motor Vehicle Collisions - Crashes”. The main features we are using are: borough, zip code, geographic coordinates, street, injury type, and etc. We used exploratory data analysis and spatial clustering algorithms to identify potential hotspots and reasons behind these crashes. We also identify other reasons for crashes such as: following too closely, driver inattention, and aggressive driving. A trend we see is the recurrent collisions near major bridges and expressways during certain timeframes within NYC.

2. Introduction:

New York City is the largest city on the east coast of the USA. It has a population of 8.5 million people. Every day, millions of people drive on the same roads, same streets, and obey the same traffic laws. Yet, with reckless driving, faulty vehicles, or malicious actions, incidents can occur, ranging from minor fender-benders, to serious accidents causing injuries, untold amounts of damages, and even fatalities. These horrible accidents can cause a city to become dangerous and weary drivers and pedestrians can increase the stress amongst a cities’ population.

However, wherever there are points of data, patterns can be formed. Time, date, location, severity, and contributing factors are all variables that can point towards certain trends and patterns in a traffic accident. These data points can be used to explore patterns, analyze trends, and predict where these crashes will take place. The data will be explored and a model will be created to predict the probability of a crash given the time and location within New York City.

The remaining portions of the paper will be structured as follows, section 3, Problem Statement, will go over the issue we are addressing in this paper. Section 4, Data Description, will describe the database in detail, including how we cleaned it to use in our data exploration and modeling. Section 5, Methodology, will go over our data exploration methods as well as our data modeling, and Section 6, Empirical Results, will go into further detail on our data showcasing what we discovered with graphs and other statistical results. Finally, Section 7 and 8, Conclusion and References respectively, we will give our final thoughts on the analysis as a whole, and credit any external sources used within the paper.

3. Problem Statement

New York City PD collects extensive data on vehicular crashes. They collect the time, how the crashes occurred, zip, vehicle types, amount of people injured, and etc. A ScienceDirect study has provided us with several methods to locate accident prone areas, so that transport agencies can help improve precise locations. Despite the results from ScienceDirect, additional research and analysis is needed to pinpoint NYC's most problematic streets, boroughs, intersections, and time of crashes.

4. Data Description:

The dataset used for this project uses columns that are directly used or useful for our problem statement.

A. Cleaned Dataset:

- Borough, zip code
- Latitude, longitude, location
- On_street_name, cross_street_name
- Location_string

- Vehicle_type_code_1-5
- Number_of_persons_injured, number_of_persons_killed

The cleaning process that we used was implemented to remove unnecessary and unused data. Sections such as vehicle types had 5 columns that were not needed; date of the accident was expanded on to provide a more clear and clean version. Columns 'Month', 'Hour', and 'day_of_week' were added to expand the data rather than just date.

B. Defining "Problem Areas"

The purpose of this paper is to provide a broader view of vehicular crashes in NYC. The data that we clean is used to create Spatial scatter plots, Temporal bar graphs, and a prediction model. This allows us to view the overall issue or "problem spots" in NYC.

5. Methodology:

A. Spatial Scatter Plot

The Spatial Scatter Plot is created on seeing a map sized view of all the crashes in NYC. When creating the spatial scatter plot key features used are:

1. **Coordinate Matrix:**

   Extracting the longitude and latitude from the dataset to store them into a NumPy array or pandas frame for a scatter plot.

2. **Intersection labels:**

   Using the in_street_name, cross_street_name, and location_string allows the data to be mapped by street name that is easier for us to understand and not a coordinate matrix.

B. Temporal Bar Graph

The Temporal graph is oriented on determining what time the crashes occur rather than problem areas. When creating the Temporal Bar Graph key features used are:

1. **Temporal Grouping:**

   When creating the bar graph time of the day is extracted from the dataset to form a correlation between other crashes.

2. **Crash Count:**

   The entire dataset is vehicular crashes. The crash count is based on the amount of entries in the dataset.

C. Crash Frequency Bar Graph

The Crash Frequency Bar Graph is oriented on determining what street has the most vehicular crashes.

1. **Crash Count:**

   Similar to how it’s explained in the Temporal Bar Graph, it's the total number of dataset entries.

2. **Categorical Grouping:**

   Groups based on the records of the location_string, provides an easier method to visualize problematic streets.

D. Injury Prediction Model

The prediction model is a random forest, binary tree classification model that predicts if you will be injured based on several factors, including date, time, location (borough, specifically), contributing factors, and if any other individuals were injured or killed in the accident.

E. Crashes By Borough

The Crashes By Borough graph is used to compare the crash data between the different boroughs in NYC. This helps identify which borough has a higher crash rate.

1. **Categorical Grouping:**

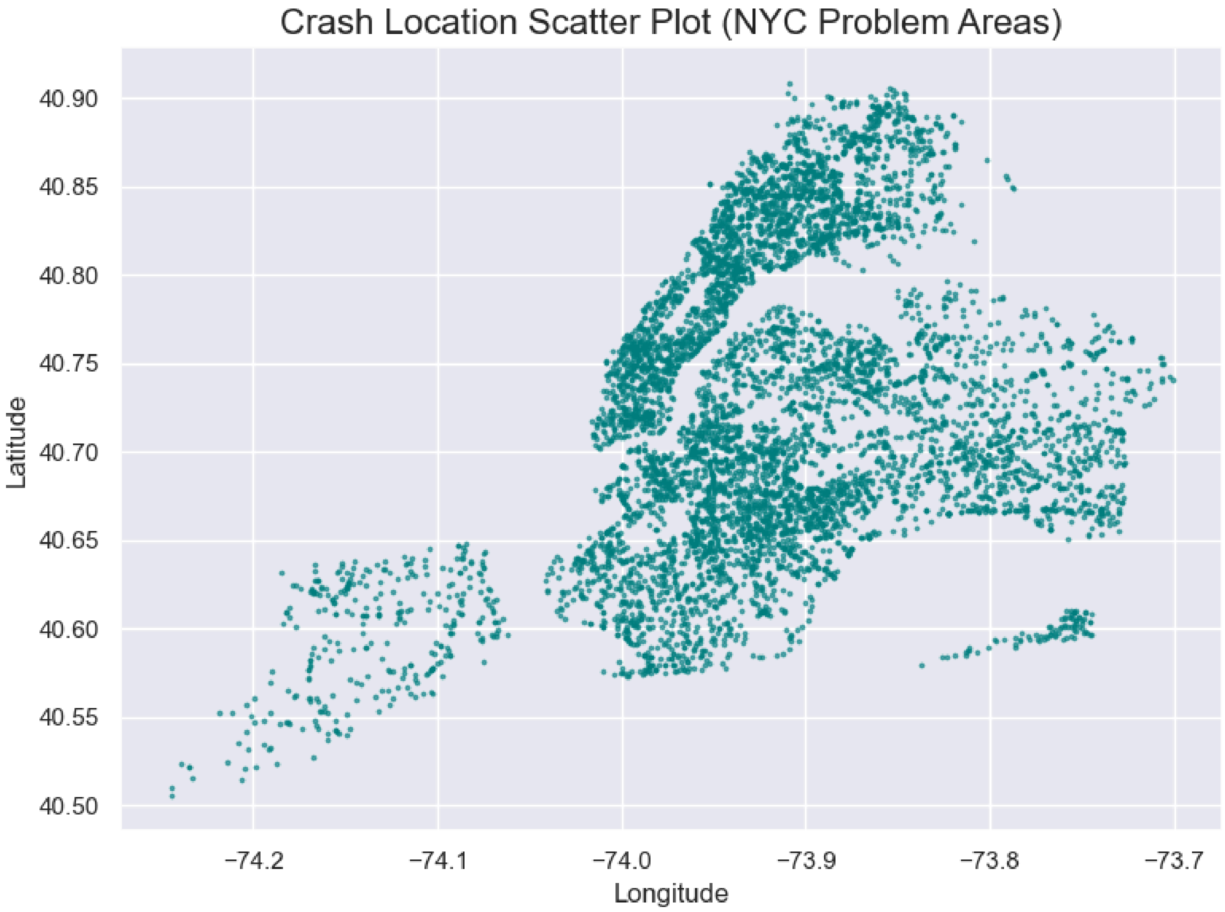
   This graph is grouped by the borough column to identify which borough the crash occurred in.

2. **Crash Count:**

   Counts the total number of crashes recorded in each borough.

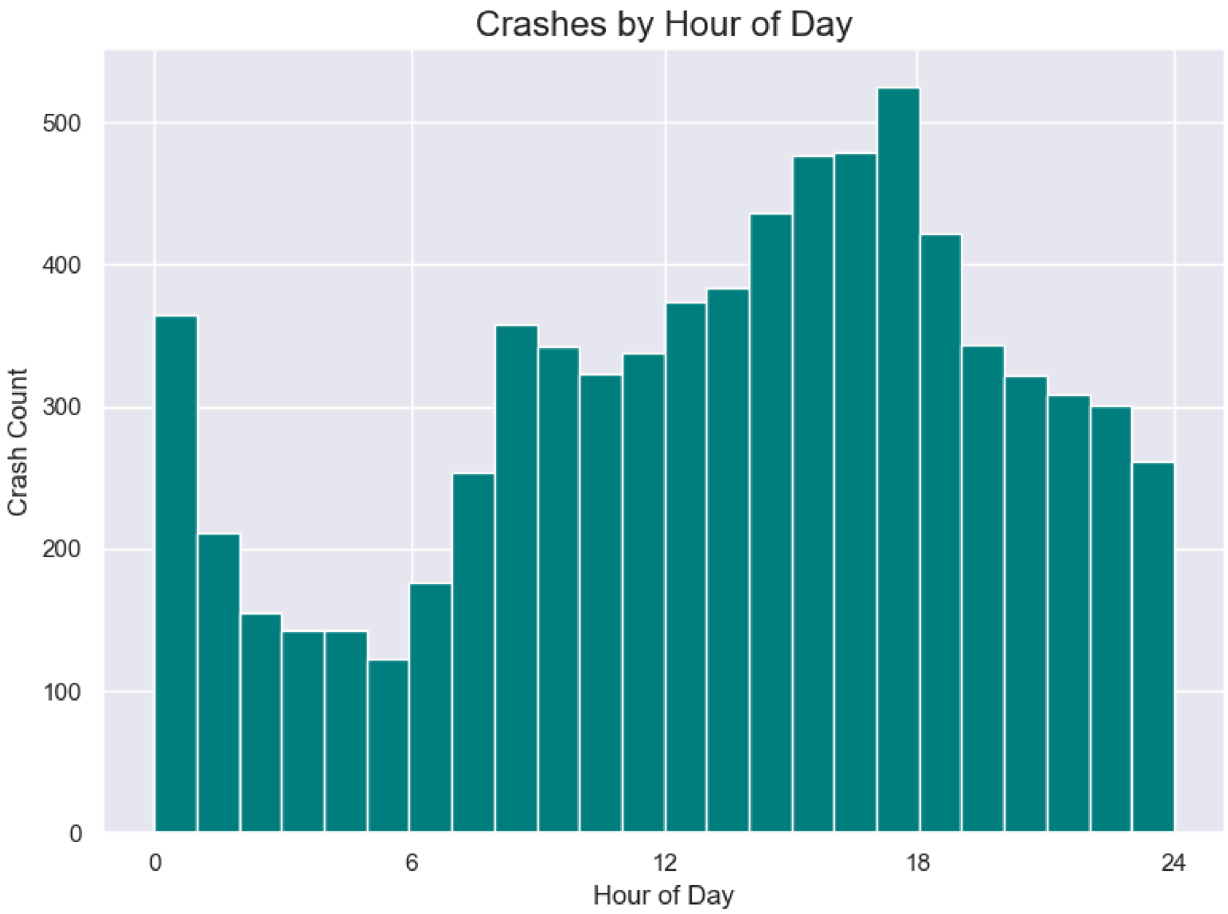
6. Empirical Results:

A. Spatial Scatter Plot



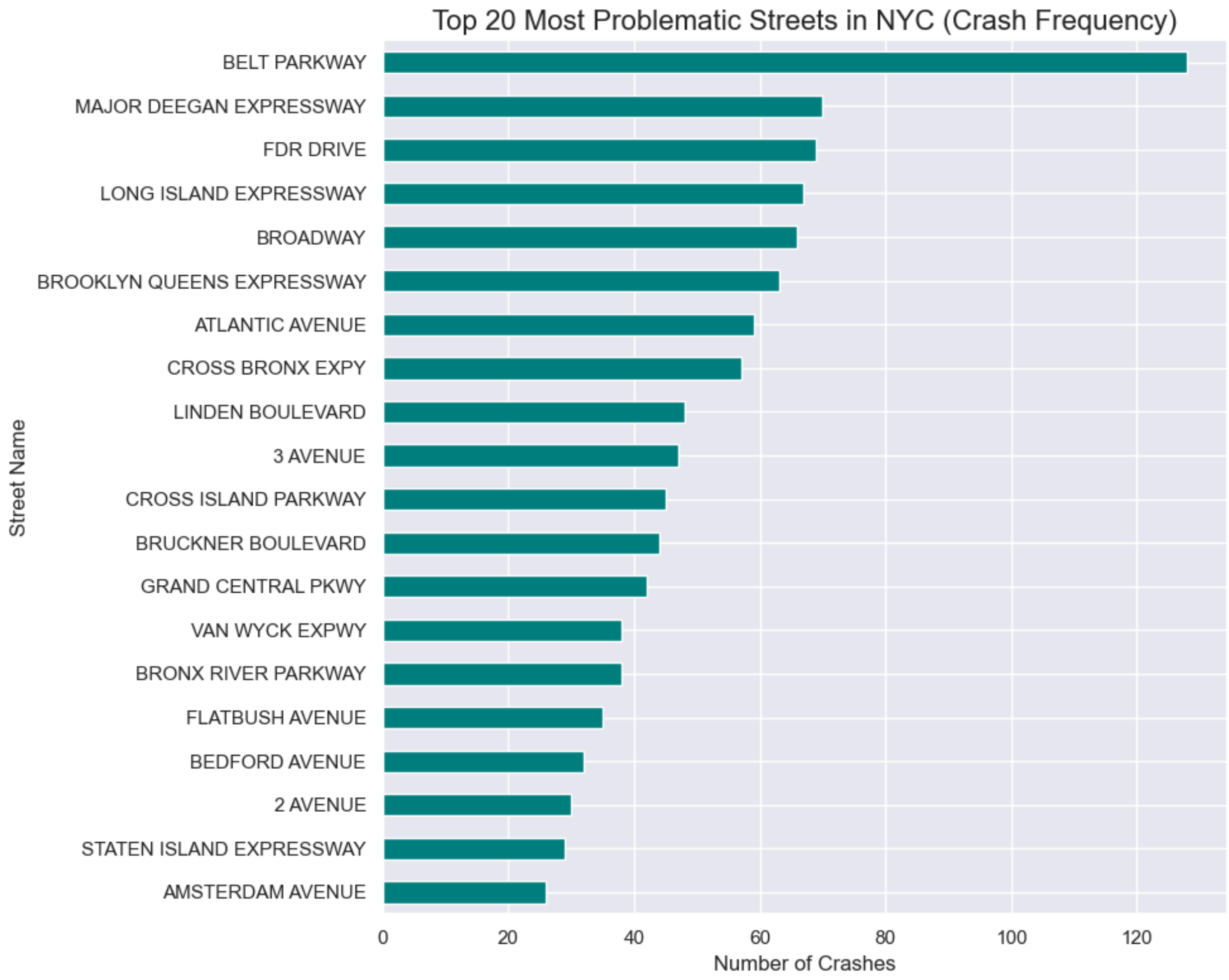
The Spatial Scatter Plot uses longitude and latitude to reveal several geographical problem areas across NYC. When observing the scatter plot each borough is almost mapped out, this shows dense concentrations of crashes across NYC. Every dense cluster of data is shown around or near traffic dense areas. The results show that crashes aren't evenly distributed across the city, but rather around predictable clusters of dense traffic regions.

B. Temporal Bar Graph



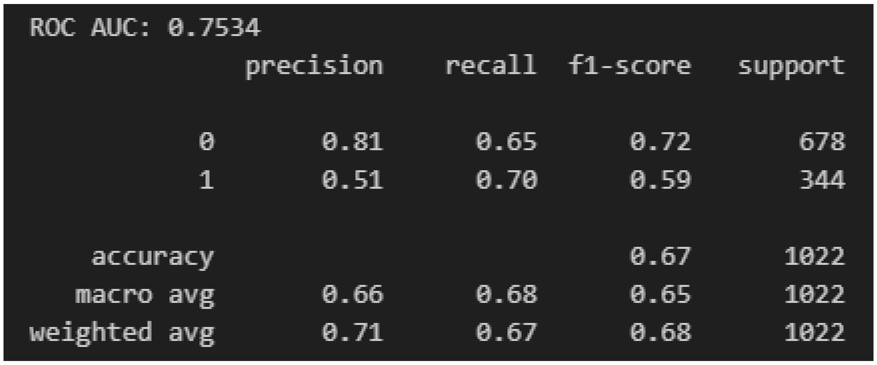
The temporal bar graph highlights a clear pattern in the timing of the vehicular accidents. When observing the graph the crash count spikes between 3 PM and 8 PM, suggesting that rush hour congestion, and uneven traffic flow contributes to the collision risk. The time of 12 AM and 1 AM suggests that drunk or tired drivers cause a majority of the accidents. Morning time between 7 AM and 10 AM show a moderate amount of crashes due to commuting patterns of the citizens of NYC. The overall graph displays the result that the peak crash count periods align with NYC commuting patterns.

C. Crash Frequency Bar Graph



The crash frequency bar graph identifies individual streets that have increased vehicular accidents. When processing the dataset the highest number of crashes occur on streets associated with expressways, multi-lane corridors, and bridge ramps. Streets such as major deegan expressway, fdr drive, and long island expressway hold the 2nd through 4th highest crash counts. The belt parkway has the highest crash count due to it being a culmination of a series of roads that stretch around the boroughs of Brooklyn. The results confirm that NYC vehicular crashes are heavily dependent on location and road length/complexity.

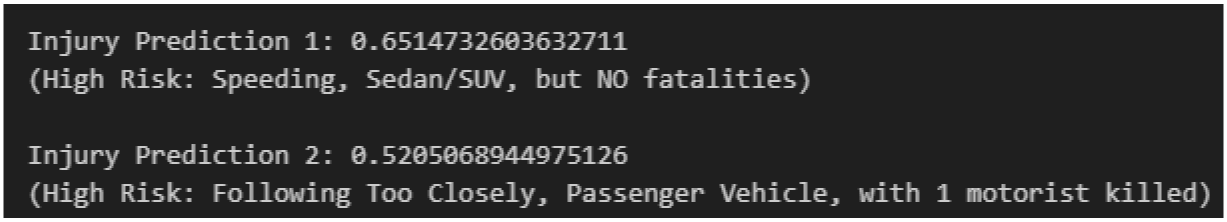
D. Injury Prediction Model



```
ROC AUC: 0.7534
              precision    recall  f1-score   support

           0       0.81      0.65      0.72       678
           1       0.51      0.70      0.59       344

    accuracy                           0.67      1022
   macro avg       0.66      0.68      0.65      1022
weighted avg       0.71      0.67      0.68      1022
```

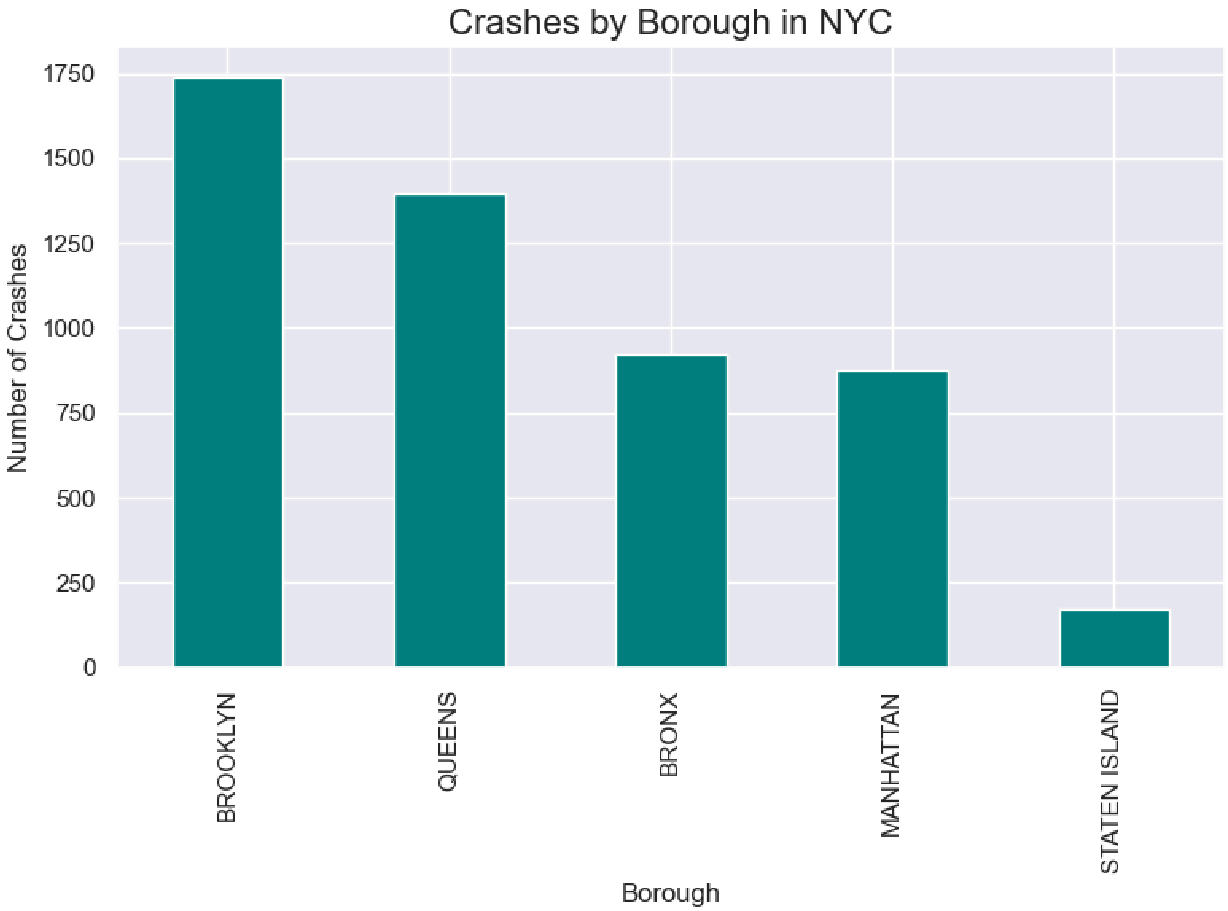


```
Injury Prediction 1: 0.6514732603632711
(High Risk: Speeding, Sedan/SUV, but NO fatalities)

Injury Prediction 2: 0.5205068944975126
(High Risk: Following Too Closely, Passenger Vehicle, with 1 motorist killed)
```

Our model used a random tree classification, with a series of binary trees. This means that our primary results are based on two classes, no injury (0), and injury (1). The model boasts a high precision score and moderate recall scores when predicting no injuries, but boats middling scores when predicting an injury. While the numbers are admittedly not as high as we’d like, this does fall in line with common risk prediction models, as it is better to present a false positive.

E. Crashes By Borough



The crashes by borough shows a hierarchy in crashes within NYC. Out of roughly 7500 data entries Brooklyn is recorded to be the borough with the highest amount of crashes. The 2nd highest record of crashes is Queens, which is driven by major expressways and bridge approaches. Manhattan had a moderate amount of crashes compared to Brooklyn because of its smaller geographical size with most of the accidents occurring in midtown. The Bronx was similar to Queens with most of its crashes being caused because of expressways and bridge approaches. Staten Island has the lowest number of crashes due to its size and low amount of traffic, dense roads and fewer expressways. This graph's results show that the increase of crash count is tied to roadway density and traffic.

7. Conclusion:

The analysis of NYC motor vehicle collisions displays that crash risk is concentrated in identifiable spatial, temporal, and roadway-based problems. Through the several graphs that have been displayed it is clear that these hotspots occur from roadway geometrics, merging conditions, and congested roads. If steps were to be taken to help alleviate these problems NYC should implement roadways safety such as redesigning problematic intersections(adding a four way stop), public announcements on safe driving behavior, law enforcement stationed near problematic areas, and speed bumps to slow down drivers. These actions can help reduce the amount of crashes and improve safety in these locations.

8. References:

1. City of New York, 2025. *Motor Vehicle Collisions – Crashes.* NYC OpenData / Data.gov. https://catalog.data.gov/dataset/motor-vehicle-collisions-crashes.
2. *GIS-Based Traffic Accident Hotspot Prediction Using Machine Learning and GIS.* ScienceDirect, 2025. https://www.sciencedirect.com/science/article/pii/S136984782500169X.

Kyle Rushing’s Statement

During the duration of the semester Hayden Mann and I both worked on our project “Vehicular Crashes and Problematic Streets Within NYC”. Hayden and I split the project evenly to help distribute the work between us. I helped work on the “Crashes by Hour of Day” graph and “Crashes by Borough” graph. I also created most of the PowerPoint slides and most of the paper.

When looking back at what I accomplished from this project, this assignment helped me gain experience working with real world data. I learned how to clean and process data to reduce redundancy and null values. At one point Hayden and I even came across an issue where we couldn’t import the cleaned version of our csv into MySQL. It forced us to clean the data within the file of our code. This issue helped me realize multiple effective ways to clean our data.

Working on this project gave me the experience of effectively visualizing data to reveal patterns. These patterns in our project correlate to peak crash hours, crashes by boroughs, crash predictions, and even our scatter plot of the crashes. Creating these graphs helped me to support our conclusions about our problem statement of finding problematic streets in NYC.

This project has overall improved my understanding of what is needed to be accomplished to correctly process, clean, and visualize data. It helped me clearly understand the process of problem solving and communication with group members.

Hayden Mann’s Statement

I worked with Kyle Rushing to complete this project for this class. We had considered getting a third member, but thought it was best to stick to ourselves, and focus on what we were good at. I completed most of the code work, including setting up the database itself with the MySQL server, and the injury prediction model, however Kyle also assisted with creating several of the graphs for data exploration. He also worked on the majority of the paper and powerpoint presentation.

For this project we played to our strengths, as I was better at coding, he was better at writing, and of course, we helped each other along the way and we both contributed to all pieces of the project. I enjoyed my time working with Kyle, and working on this project. This project served as a great way to apply a lot of what we learned in this class in a real way, and the freedom of choosing our own database and problem statement allowed us to get invested in what we wanted to do. It taught me a lot, and I certainly think I can improve in a lot of areas, but I do feel proud about what we did.

Working with a database and doing data analysis in this manner was all new to me, but I’m glad I was able to develop these skills for identifying data, cleaning it, exploring it, and creating a model for it.