Predicting the Likelihood of Shootings in New York City

Dan Beim, Shane Halpin, Nick Pelletier, Pranjal Singh

Syracuse University

IST 718: Big Data Analytics

Professor Chris Dunham

School of Information Studies

Introduction

New York City is a bucket list location for many people around the world. New York and the welcoming presence of the Statue of Liberty is often the first thing most people think of when they think of the United States. The "Big Apple" has many appealing attractions for tourists, such as numerous shopping destinations, a breadth of historical sites, the bright lights of Broadway and Times Square, the beauty of Central Park, and the world class sports franchises. With all of these features, it is no surprise that New York City continues to be one of the most visited cities in the world, welcoming 56.7 million visitors in 2022, according to New York City Tourism + Conventions. With this figure expected to grow to 63.3 million by the end of 2023, it is clear that the city will continue to host millions of more visitors each year.

When planning a trip to New York City, it is important to consider the safety aspects as well. New York City is one of the most dangerous places in America, with high rates of crime compared to most other major cities in the United States. One such crime that is more violent in nature, shootings, occur on a daily basis throughout the city's five boroughs. These shootings do not seem to be random, however. Certain locations within the city experience much higher rates of violence than others. In order to properly analyze which locations of the city might be safer to visit and stay in for tourists, visitors, or even people moving to the city full-time for work, a thorough analysis of these shooting occurrences must be performed. That is where our project comes in.

Our project objective is to use the data available to the public via the city of New York crime database to predict when a shooting will occur. We are setting out to properly analyze and gain a better understanding of the patterns of various crime complaints reported to the New York Police Department (NYPD) within the past year. Our goal is to properly identify high-risk areas, specific time frames where the most incidents occur and other factors influencing these complaints, assisting law enforcement in optimizing resources and public engagement. In addition, we look to make predictions about the number of specific complaints filed in locations over the course of a year, down to the type of crime, and time of day committed.

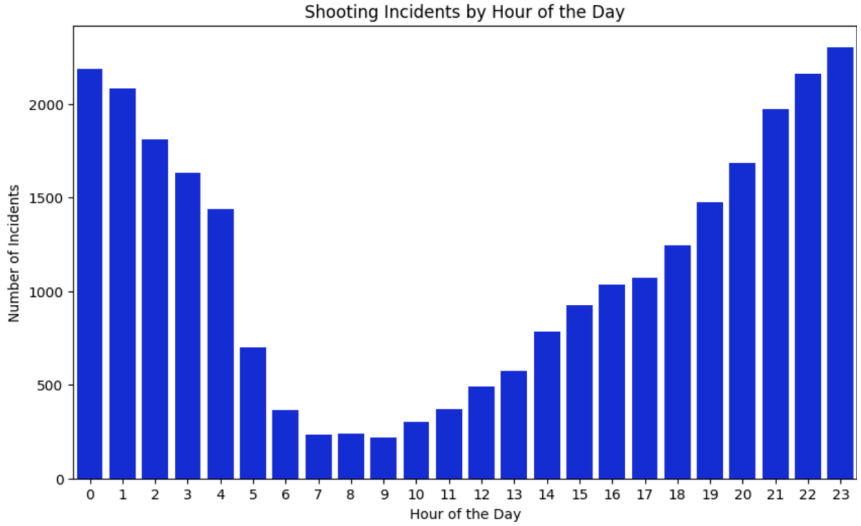
We will consider locations within our analysis as the main explanatory variable while leaving out socio-economic factors which could lead to biased predictions, leading to unequal distribution of implementations if the results were to be used as recommendations to the police department. Through our analysis, we hope to gain a better understanding of certain factors that play a role in shooting occurrences and build a model that takes these factors into account and makes accurate predictions about shooting likelihoods. In order to accomplish this goal, we will utilize a variety of data analysis tools learned in Big Data Analytics including visualization, regression trees, and classification analysis in order to help law enforcement better identify crime patterns as well as predictive analysis in order to help prevent future crimes.

Dataset

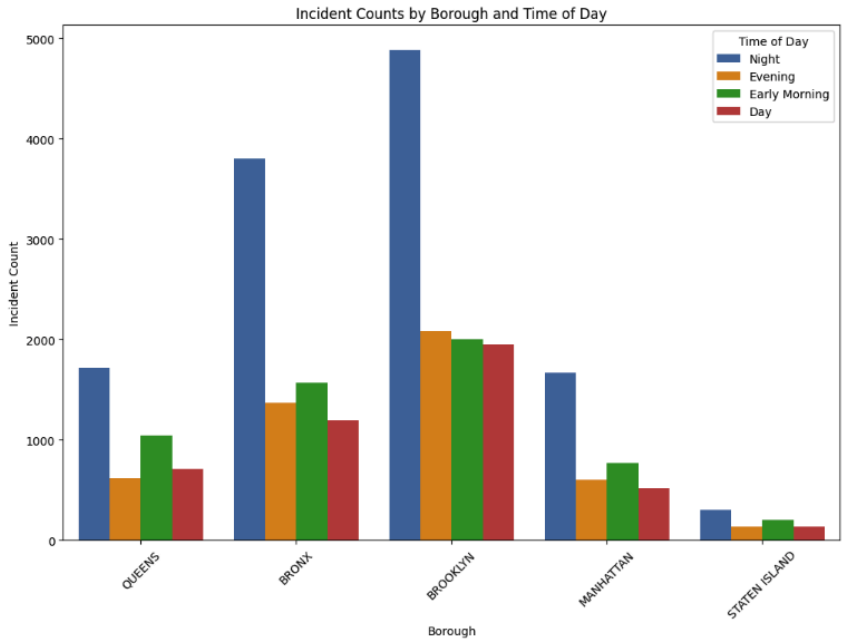
The dataset used in our analysis was made available by the city of New York through their online NYC Open Data portal. The Historic NYPD Shooting Incident Data is a list that encompasses every shooting incident reported by the NYPD from 2006 to 2022. The breakdown of each incident contains details regarding the specific complaint, including the location within the city, the time of day, demographics of perpetrator and victim, and other relevant attributes. There are 27,313 rows (or 27,313 total incidents reported) with 26 columns of descriptive indicators. Some sample predictors of interest include the Boro (Borough where the complaint was reported), Date and Time (Date and time of the complaint), Location (Specific location details of the reported complaint), and the precinct where the shooting occurred. Additionally, more specific indicators include information regarding the victim and the perpetrator, or the person committing the crime. These variables include their age group, sex, and race. However, given the serious nature of shootings, these fields are sometimes left blank. This is either because the police department has no witnesses of the crime, or witnesses are sometimes scared to speak up or intimidated by others to not cooperate with law enforcement during their investigation. Thus, it will be challenging to accurately predict shooting likelihoods, but perhaps easier in some scenarios where more specific and detailed information has been provided.

Data Exploration

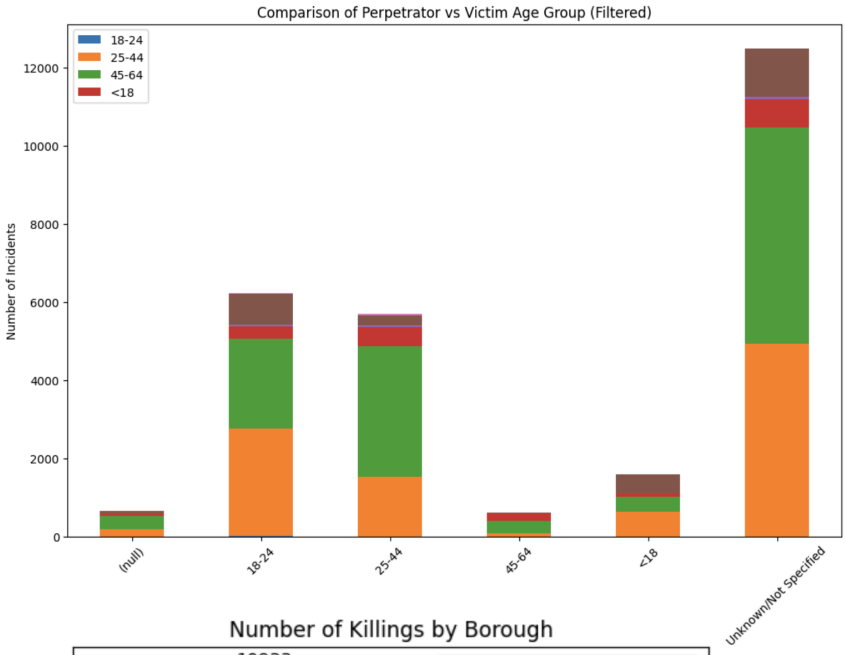
In order to properly identify which variables will assist us the most in making predictions, we first set out to generate visual plots to better represent the characteristics of the indicators in our dataset. Visualizations include heat maps to identify high-risk zones in NYC, time-series plots to see the trend of shooting incidents over time, and bar plots to determine patterns in perpetrator and victim demographics. In performing this analysis, we are properly examining the frequency of incidents across different boroughs, analyzing the time trends to determine if there are specific days or times with higher incidents, and investigating if certain locations (like residences or streets) have higher shooting occurrences. Lastly, we assess the relationship between perpetrator's demographics and the nature of the incident in order to identify the patterns in reported shooting incidents based on the individual backgrounds of both the victim and perpetrator.



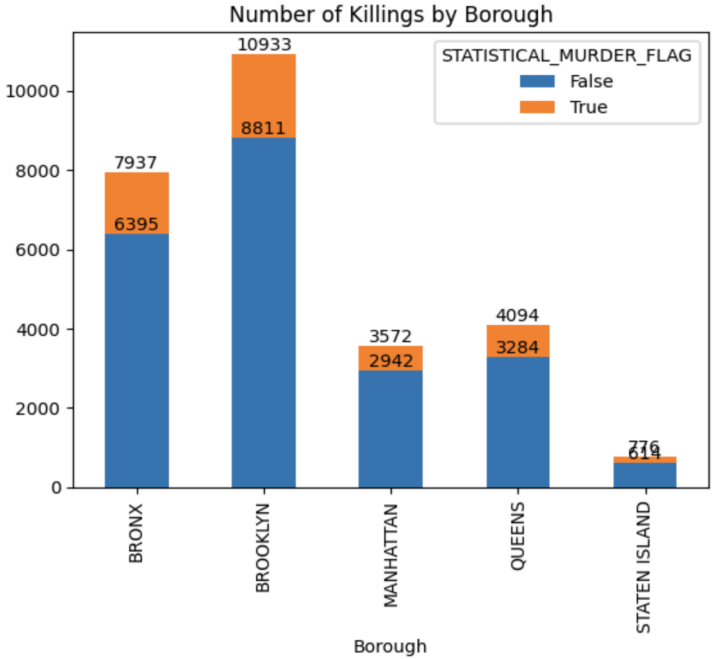
This bar chart displays the frequencies of shootings reported by the hour of the day. The majority of all shootings reported occur either later in the night or very early in the morning.



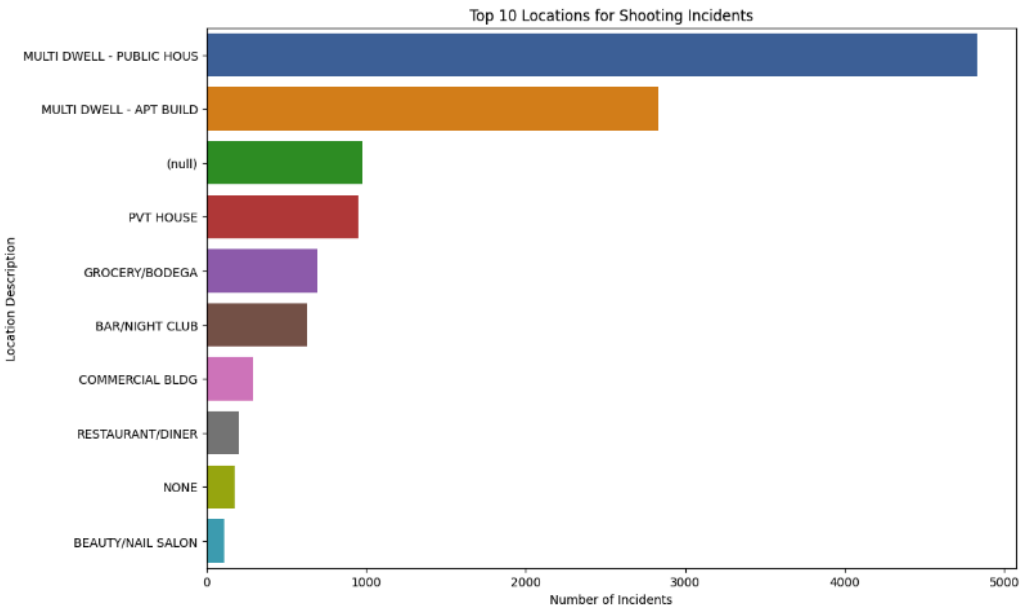
This visualization displays the total number of shootings in each borough, broken down by the time of day at which the report occurred. Given that there are 24 hours in a day, we decided to further break down those hours into four specific time groups, night, evening, day, and early morning. This plot shows that night is the most common time at which shootings are recorded across each borough, while also showing that Brooklyn is the borough with the most shooting incidents reported.



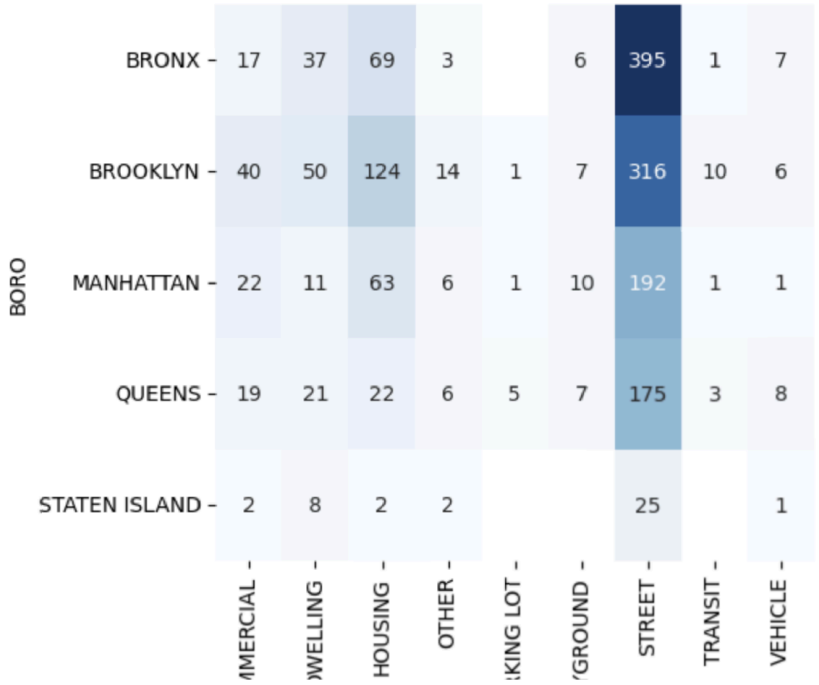
The next plot compares the age groups of the perpetrator and victim along with the number of instances shootings were reported for these age groups. We can see that the most common age groups for perpetrators (x-axis) are adults aged 18-24 and adults aged 25-44, who target older victims (making up the stacked bar plot) that are generally in the age group immediately after their own.



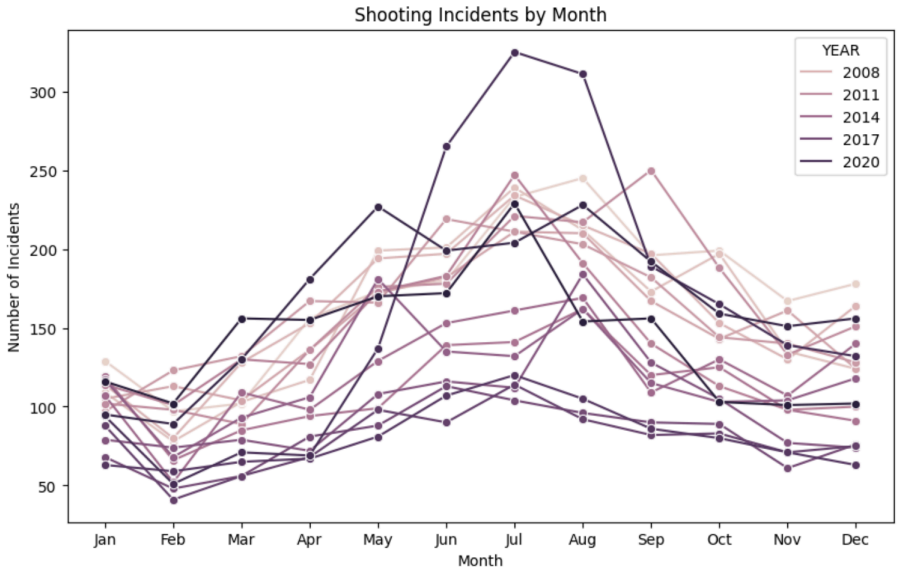
Lastly, this stacked bar plot displays the number of shootings by borough, along with an additional flag for if the shooting resulted in a murder. These types of shootings are most likely more premeditated in nature, as opposed to “random” shootings.



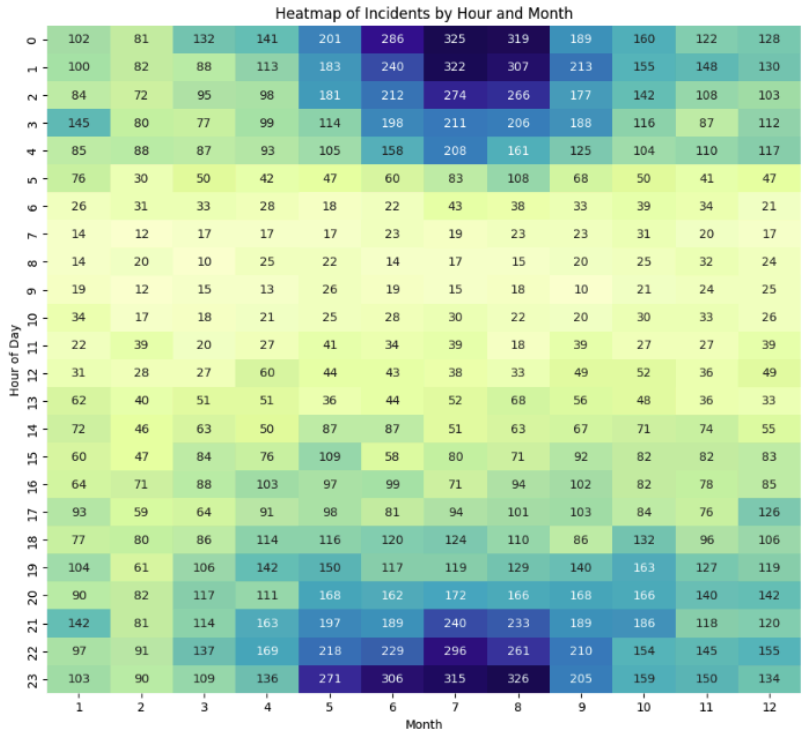
This horizontal bar chart displays the 10 most frequent locations at which shootings were reported. Shootings were most frequently reported in housing spaces, most commonly public or private housing, as well as in apartment buildings. This inherently makes sense as shootings in these locations are occurring out of the public eye, and have less likelihood of witnesses being able to describe the shooting, as well as a lesser chance of police catching the perpetrator.



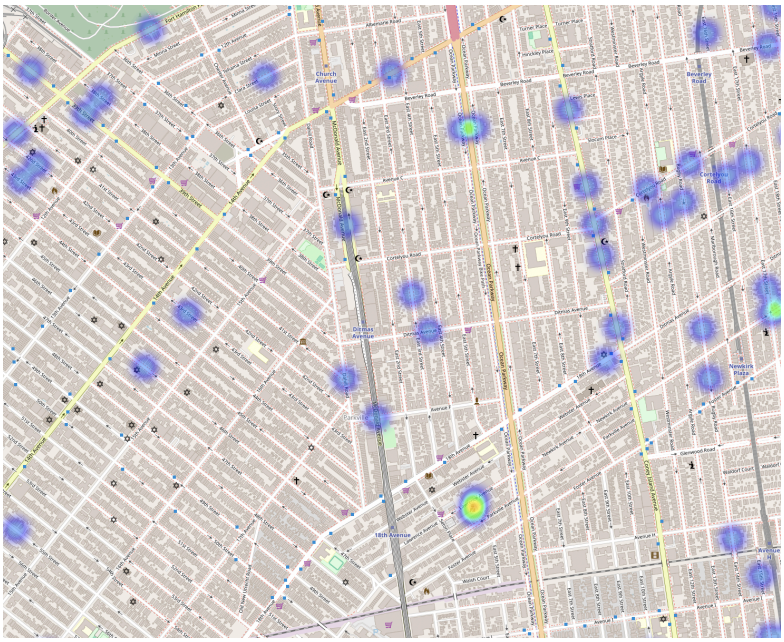
The heat map on the left side displays the most common locations at which a shooting was reported, broken down by the borough in which the shooting occurred.



This line chart displays the amount of shootings occurring in each year in the dataset by the month of the shooting. This is also an important variable to consider as it would assist with identifying certain shooting trends.



The last visualization made relating to data exploration is a heat map of all shooting incidents by the hour and month in which the shooting was reported. This is a high level heat map which would be chiefly important to the NYPD as they set out to staff locations and properly optimize shift rotations based on the observed shooting trends. As shown, most shootings occur in the summer months at the late hours of the night, and the early hours of the morning. Properly visualizing this information also allows us to recognize how important these variables will be for us as we model to make accurate predictions on shootings within the city. Additionally, we made an interactive heat map that displays areas with low and high densities of shootings reported throughout the entirety of New York City. Sample images are shown below representing specific areas that fit each criteria.



Area with Low Density of Shootings

Area with High Density of Shootings

Analysis

For our analysis we created four separate models to predict the time of day shootings would occur in a given borough of New York. Throughout all four models the preprocessing steps were the same. These steps included indexing the columns that we were using as the explanatory variables and then assembling those variables into a vector to be fed into the model as inputs. Because all of the explanatory columns were of type string in our original dataset, they had to be indexed, meaning that each string is given a unique integer value that the machine learning model can interpret. For example, one of our columns was 'boro' which contains a string value of the different boroughs in New York. Indexing converts a value such as 'Manhattan' to a value of 1 , 'Staten Island' to 2, and so on, for example. Another common practice in machine learning preprocessing is assembling all explanatory variables to be used in a certain model into a vector called the feature vector. After indexing was done, our explanatory variables were coalesced into this vector so that each machine learning algorithm could interpret them appropriately.

The first model we made was a random forest. This is a collection of n number of decision trees which each is only fed a portion of the explanatory variables. This is done so that the large number of trees each gets a small number of variables in different combinations to determine the impact each variable has on other variables when predicting the output label. This is a more powerful machine learning than single model algorithms because of the voting ability each tree has when determining what the prediction will be, allowing for more input in a single model, which most of the time results in a higher accuracy. Our random forest model had an accuracy of 45.7%. This may not seem overly impressive on its own, however, in a scenario such as ours with a four class variable, the baseline accuracy of random guessing would be 25%, meaning our first model performed 20% better than a random guess.

Our other three models performed generally just as good as the random forest but with slightly less accurate predictions. For our other models, we created a deep learning model with a neural network, a logistic regression model, and a second random forest model with different parameter settings to see if the performance would be improved. All these models had a 45% accuracy score, but was slightly less accurate than the original random forest.

Conclusion

The results of our model prove that it is difficult to correctly classify where and when shooting incidents will occur throughout NYC. While our models do in fact predict more accurately than the baseline of randomly guessing, none of our models returned an accuracy greater than 50%. Our model only used a few predictors such as precinct, locational predictors, season, and whether or not the shooting resulted in a murder to predict time of day the shooting

occurred. These predictors alone without using demographic information of both the perpetrator and the victim makes predicting time of shooting very difficult to accomplish. Despite being able to visually portray the trends in daily times where the most/fewest shootings occur, the results of our models allow us to conclude that the time of day shootings occur is mostly due to random variance and unexplained human elements. Ultimately, the data that comprise the details of each shooting, which we used as predictors, are not enough to explain where and when each shooting will occur.

Nonetheless, our best model returned an accuracy around 46%, which still predicts time of day the shooting will occur at a rate that is greater than the baseline of randomly guessing. While our conclusions allow us to decipher that the time of day shootings occur is mostly due to the nature of random chance, it is still interesting to note that we were able to explain some of this variance with the predictors in our model. These results along with the results of our visuals, specifically the heat maps, can help law enforcement better understand crime patterns. Additionally, making this information public will also help people avoid specific areas of New York City at varying times. This optimizes the safety of tourists and assists people who are looking to buy houses or apartments within the city's five boroughs.

In the future, we would like to continue exploring these data and patterns in shooting trends to explore a variety of potential insights. Specifically, we would like to gather more data that would provide additional information into each shooting, such as the motive of the shooting, the type of the shooting, and the location of the shot on the victim's body. Additionally, we could further test the relationship of certain variables and interact them with one another, as discussed in class. These could include the relationship between the amount of light in one specific area of a borough, as this could impact visibility and the likelihood of witnesses being able to report a crime and provide valuable feedback. The nature of this research paper is a rather grim one, with more complex questions than answers. If researched further, our level of curiosity will be a main component in identifying other predictors not yet mentioned in our modeling process, and help us develop more accurate predictions regarding shootings within New York City.

Ideally, the more information we have about each shooting that expands the overall dataset with which we are working, will likely boost our results. This information will also allow us to discern if there are any patterns/connections between the time of the shooting and these new potential factors that are now included in the data. Ultimately, there exists a plethora of information about each shooting that we do not currently have. By having police ask the right questions immediately following an incident, and gaining access to any additional data possible, we will be able to produce new insights to help law enforcement in NYC better predict the times and locations of shootings, and reduce crime. This will help to make New York City a safer place and a more enjoyable destination for tourists, families, and all people who wish to visit the city or call the Big Apple their home.