# NYPD Shooting Incident Analysis

Project Team 1: Akhona Msibi, Cheting Meng, Ryan Sequeira, Zhongyan Zheng

# Table of Contents

[**NYPD Shooting Incident Analysis 1**](#_1ebzc9one9qn)

[**Table of Contents 1**](#_wc3uyrz42d3b)

[**Introduction 2**](#_7lzpzp99r0fl)

[**Dataset Description 2**](#_azlgtdpgq49t)

[**Motivation 3**](#_v2ethy6gyjp3)

[**Literature Review 3**](#_vcqh38c7e52c)

[**Exploratory Data Analysis (EDA) 4**](#_ws31rxj9hqlq)

[**Statistical Analysis 4**](#_83vtumsqlfp2)

[Trend analysis 4](#_75fuz4oqu0vy)

[ARMA-GARCH 5](#_77aasz50gode)

[Multivariate Analysis 10](#_bwudlmvdstjt)

[Granger Causality 10](#_8mekpo5psp4d)

[VAR 11](#_smdubyic1e9)

[Model Summary 11](#_15cmq1k7f89k)

[Goodness of fit: Failed 12](#_1lkzkgsa5ar0)

[Normality 12](#_5u0zmwczlmpt)

[Heteroskedasticity 12](#_l34yua3hydri)

[Serial Correlation 12](#_cl3nmefaeo5s)

[Prediction Results 12](#_u1d2zyuu9qj9)

[ARMAX 13](#_6hl55e5gr7zs)

[Model Summary 13](#_2yzi5dmcxb5w)

[Goodness of fit: Satisfied 13](#_r904arxo919g)

[Normality 13](#_rq4kovptvmk8)

[Heteroskedasticity 14](#_y2lmp928dyh5)

[Serial Correlation 14](#_shqzzeh2cw28)

[Prediction Results 14](#_vfbg0gw51igd)

[**Subject Matter Statements 14**](#_ky28p159ya00)

[**Discussion and Future Research 15**](#_prvcgckr3r9r)

[**References 16**](#_s0dlrww1rtl1)

# Introduction

In recent times, police shooting incidents have become topical in American society due to high-profile confrontations involving officers across the country. This has been a motivating factor in seeking greater understanding of the nature and characterization of these incidents and their occurrences over time. New York City has been identified as a useful case study due to its diverse population and historical data availability, which allows for analysis over time and across unique demographic factors. The overarching aim of this study is to understand the temporal and spatial patterns, demographic trends, and underlying factors associated with shooting incidents in New York City. It seeks to identify key vulnerabilities, potential causes, and correlations, which could offer actionable insights to policy makers, law enforcers, and communities at large.

# Dataset Description

The chosen dataset is the New York Police Department (NYPD) Shooting Incident Data (Historic) dataset. This dataset was downloaded from a file last updated on September 2nd, 2023 and contains a breakdown of every police shooting incident recorded in New York City between January 1, 2006 and December 31, 2022.

This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website - we can therefore assume high quality in the dataset. Each record represents a shooting incident in NYC and includes information about the event, the responsible precinct, the location and time of occurrence. In addition, information related to victim demographics is also included.

The dataset is in tabular form with 27312 rows and 21 columns. Furthermore, the time component is detailed to minute level, and hence suitable for detailed time series analysis. We plan to aggregate the data to the day-level or to specific hours if necessary. Additionally, we will segment the data based on demographic information of the victims to identify how shooting incidents change by victim profile in addition to time. Our chosen segmentations include sex, age group, race, and borough (regions of New York City).

| **Demographic** | **Categories** |
| --- | --- |
| Age | <18, 18-24, 25-44, 45-65, 65+ |
| Race | American Indian, Asian/Pacific Islander, Black, Black Hispanic, White, White Hispanic |
| Sex | Male, Female |
| Borough | Bronx, Brooklyn, Manhattan, Queens, Staten Island |

**Table 1: Demographics of Dataset**

# Motivation

Firstly, an examination of historical occurrences of police shooting incidents is carried out, assessing these counts with daily frequency. This is performed to gain insights on incident trends, as well as potential seasonality in when these take place. Additionally, the demographic characteristics of police shooting incidents are assessed. This involves examining categorical factors of victims such as age (at time of incident), race, and gender. The objective is to determine if certain demographic groups are disproportionately affected by these incidents. Moreover, we're keen on tracking the discernible trends in these frequencies over time.

Furthermore, a key facet of the investigation centers around the spatial and temporal patterns of shootings. It starts by analyzing the distribution of these incidents across various time periods and locations, designated by borough. This is supported by a time-series analysis, assessing whether certain boroughs exhibit consistently greater incident frequencies, and changing trends over time. This information is key for police officer distribution and future planning by management.

Finally, the researcher conducts a multivariate analysis of all dimensions to identify any correlations or lead/lag relationships between the factors. This assists in proposing informed strategies or interventions based on combined insights from all categories to curtail the incidence of shootings in the city and adequately prepare for future occurrences.

# Literature Review

Prior to this investigation, research on this subject matter was performed to guide intuition on the expected results from the study. Firstly, existing literature on situational predictors of fatal police shootings highlights how non-White suspects and officers are a significant predictor of deadly force incidents (White, 2002). This suggests that non-White racial groups will exhibit higher shooting frequencies within our data.

Additionally, research conducted by the National Policing Institute on officer-involved shooting situations identified certain demographic factors that may increase the likelihood of use of force by the police (Grieco & Robbins, 2019). This report detailed how police are more likely to use force against young male suspects as well as those from racial and ethnic minority groups, originating from a prior study by Terrill & Mastrofski on this subject matter (Terrill & Mastrofski, 2002). This assessment provides insight on the expected results across three of the demographic factors. It is expected that higher shooting frequencies will occur for younger victims, male victims, and non-White victims in comparison to their respective alternatives.

Finally, temporal trends may not be present in the data, as described in existing literature on racial inequality in shooting incidents (Lett et al., 2020). This describes how the rate of shootings of victims of color had remained constant in the period between 2015 and 2020. This may extend to other demographic groups, where differences in frequency are present between them but no long-term trends present themself in the analysis. The analysis is conducted to assess if these expectations hold true.

# Exploratory Data Analysis (EDA)

In analyzing the daily time series for each of our demographic groups, we found that our data was too sparse for our time period (many days with 0 incidents) and too cluttered due to several years of daily data. An example of such a plot is shown below. To counteract this, we aggregated our data to the month level for each of our time series.

# 

# 

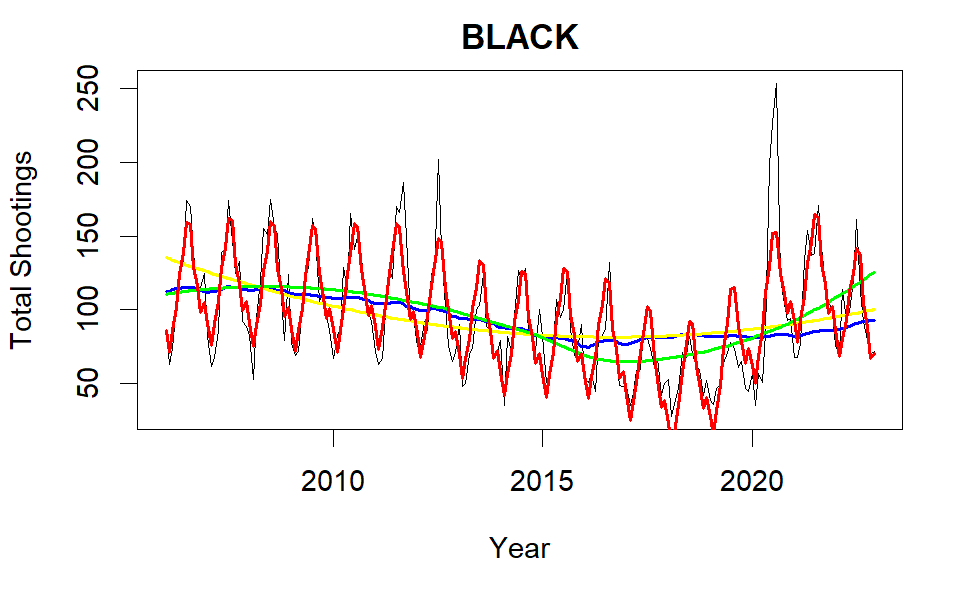
**Figure 1: Daily Time Series for Age < 18 (Left) and Monthly Time Series for Age < 18 (Right)**

From the time series plots and all of our ACF plots (see attached code), we see strong seasonality in the pattern of shootings with peaks in summer months. We find that there is a trend of decreased total shootings in the years up to 2020 and an increase directly after. The data does not seem to be stationary.

# Statistical Analysis

## Trend analysis

We attempted to fit multiple models to the time series for each demographic. We found consistently that the splines model fit best with moving average, quadratic, and polynomial models performing roughly equal to each other.



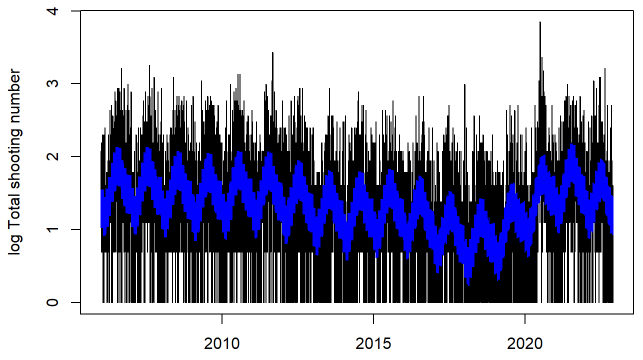
**Figure 2: Time series (black victims) - splines (red), polynomial (green), quadratic (yellow), moving average (blue) models**

Though the splines model fit best, we found that it still had a low R-squared coefficient (0.483 for time series above). Significant factors include trend, and multiple months - confirming our previous observations.

## ARMA-GARCH

**The first way** we plan to do this is to firstly do a logarithm of the daily data for Total shooting data, and secondly, using the gam model, go about fitting a non-parametric trend using splines together with ANOVA day-of-the-week and day-of-the-month seasonality. Training data is the daily shooting data from 01-01-2006 to 11-30-2022, the testing data is from 12-01-2022 to 12-31-2022. We then do ARIMA-GARCH modeling of the residuals of this model.

logtotal.train ~ s(pts) + weekday + month



**Figure 3: Fitted gam model and Daily shooting time series**

This gam model is good fit enough that only weekday7 and month3 is greater than the significant level of 0.05.

Firstly, using ARIMA selection on the gam model’s residual. Selecting the best model based on AIC with both p and q up to 5, d up to 1:

| **p** | **d** | **q** | **AIC** |
| --- | --- | --- | --- |
| 2 | 0 | 2 | 11400.67 |
| 3 | 0 | 1 | 11399.46 |
| 1 | 0 | 3 | 11399.44 |
| 1 | 0 | 2 | 11397.44 |
| 2 | 0 | 1 | 11397.44 |

**Table 2: ARIMA selection to gam model’s residuals**

We choose the order of (2,0,1), then do the box test to check the serial correlation and heteroskedasticity

|  | **X-squared** | **df** | **p-value** |
| --- | --- | --- | --- |
| **Serial correlation** | 1.1714 | 1 | 0.2791 |
| **Heteroskedasticity** | 7.9873 | 1 | 0.004711 |

**Table 3: Box test of ARIMA model (2,0,1) residuals**

The tests for serial correlation show that at a significance level of p = 0.05, there is no serial correlation in the residuals. Though there appears to have significant heteroskedasticity, according to the results of the tests for the ARCH effect.

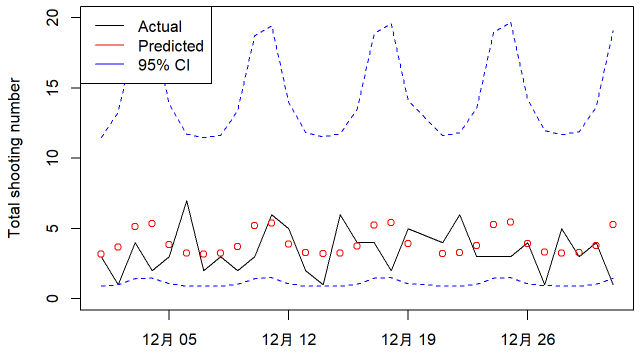
So we use the Heteroskedasticity Modeling to check the order of the ARMA-GARCH model, starting with the ARMA=(2,1). Here is the three steps of the Heteroskedasticity Modeling:

| *1. Select GARCH order given ARMA=(2,1)* | | | *2. arma update, given GARCH=(0,0)* | | | *3. GARCH update, given ARMA = (1,2)* | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **m** | **n** | **BIC** | **p** | **q** | **BIC** | **m** | **n** | **BIC** |
| 0 | 0 | 2.000657 | 1 | 0 | 1.999787 | 1 | 1 | 2.000622 |
| 1 | 0 | 2.000656 | 2 | 2 | 1.999146 | 2 | 0 | 2.000582 |
| 2 | 0 | 2.000618 | 3 | 1 | 1.999133 | 0 | 2 | 2.000455 |
| 1 | 0 | 1.999147 | 1 | 3 | 1.999111 | 1 | 0 | 1.999112 |
| 0 | 0 | 1.999147 | 2 | 1 | 1.997637 | 0 | 1 | 1.999110 |
| 0 | 0 | 1.997637 | 1 | 2 | 1.997602 | 0 | 0 | 1.997602 |

**Table 4: The best ARMA-GARCH selection for gam model residuals**

Still the garch(0,0). We will use the garch(0,1) instead of (0,0), use arma(1,2) instead of (2,1). Build the arma-garch (1,2)x(0,1) model.

Because we are doing an arma-garch model fitting to the residuals of the gam model, and there are two different confidence intervals in the final prediction, which cannot be simply summed up, we need to synthesize the prediction variance of the two models. Since we did not involve learning for the change content, we used the confidence intervals of the ARMA-GARCH model as an approximation of what might be considered sufficiently accurate for practical applications.Regarding the predicted values, we use the gam model predictions and the arma-garch model predictions added together.

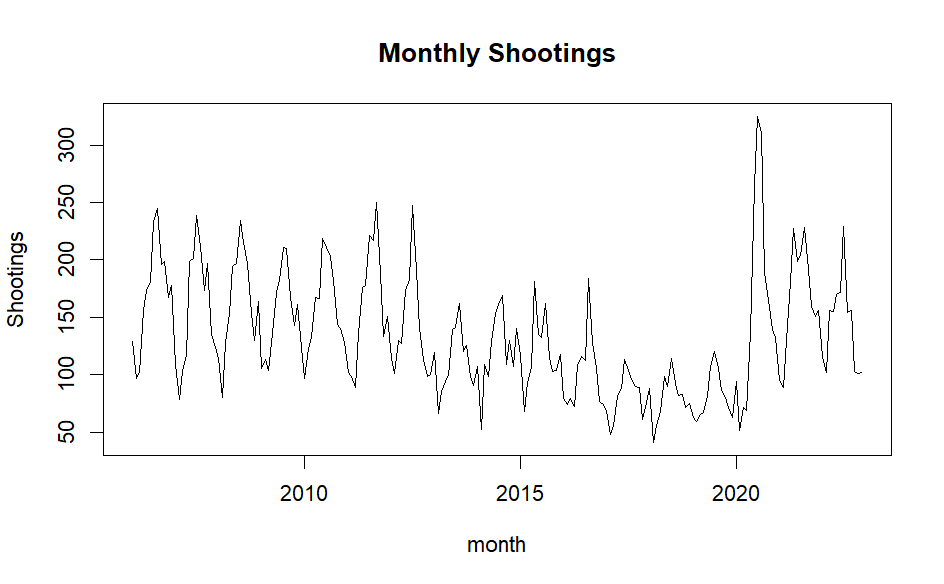


**Figure 4: Prediction of the next 30 points along with the test data**

This plot shows that the predictions approximate the actual daily shooting data not closely. So we try to use another approach to explore a better model.

**Second method**

Since the total shooting data itself is daily data, we try to recognize this series in a more clear way by aggregating daily data into monthly data for easy calculations. We use data from January 2006 through December 2021 as our training data, and a full year of data from 2022 as our test data.



**Figure 5: Monthly shooting data time series**

Firstly, using ARIMA selection on the log Monthly Total shooting data. Selecting the best model based on AIC with both p and q up to 5, d up to 2:

| **p** | **d** | **q** | **AIC** |
| --- | --- | --- | --- |
| 5 | 1 | 5 | -64.45662 |
| 5 | 0 | 5 | -64.75619 |
| 2 | 1 | 3 | -65.75298 |
| 4 | 1 | 5 | -65.77985 |
| 3 | 1 | 3 | -66.16617 |

**Table 5: ARIMA selection to Monthly shooting data**

We choose the ARIMA model (3,1,3), all coefficients are significant at level 0.15 except ar2.

Then we do the Box test to check the serial correlation and arch effect of this model:

|  | **X-squared** | **df** | **p-value** |
| --- | --- | --- | --- |
| **Serial correlation** | 1.691 | 1 | 0.1935 |
| **Heteroskedasticity** | 35.719 | 1 | 2.279e-09 |

**Table 6: Box test to ARIMA model (3,1,3) residual**

The tests for serial correlation show that at a significance level of p = 0.05, there is no serial correlation in the residuals. Though there appears to have significant heteroskedasticity, according to the results of the tests for the ARCH effect.

Starting with the ARIMA orders of (3,1,3) to model the Log Monthly total shooting training data and using a max possible order of (5,2,5)x(3,3), use minimum BIC to determine the best ARIMA-GARCH order on the training data.

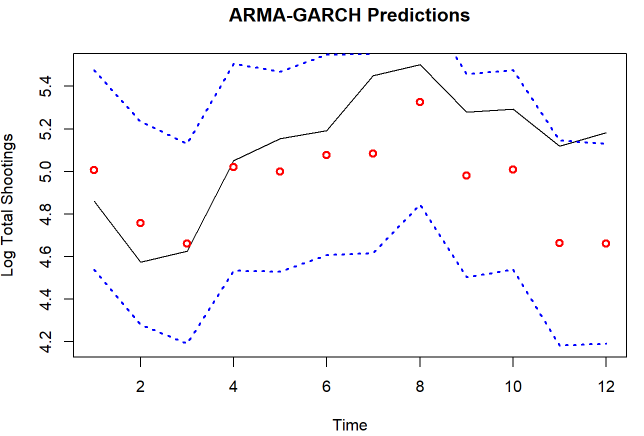
Then update the ARIMA model by GARCH model (0,0) based on BIC, we change the ARIMA model into the order of (5,1,5). After that, update the GARCH model:

| **m** | **n** | **BIC** | **p** | **d** | **q** | **BIC** | **m** | **n** | **BIC** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 3 | 0.1880831 | 5 | 1 | 0 | 0.1199186 | 1 | 1 | 0.1735944 |
| 1 | 1 | 0.1607284 | 5 | 1 | 1 | 0.1199186 | 0 | 2 | 0.1735522 |
| 0 | 2 | 0.1607139 | 5 | 1 | 2 | 0.1199186 | 2 | 0 | 0.1733151 |
| 1 | 0 | 0.1340242 | 5 | 1 | 3 | 0.1199186 | 0 | 1 | 0.1462116 |
| 0 | 1 | 0.1333456 | 5 | 1 | 4 | 0.1199186 | 1 | 0 | 0.1460127 |
| 0 | 0 | 0.1070525 | 5 | 1 | 5 | 0.1199186 | 0 | 0 | 0.1199186 |

**Table 7: The best ARMA-GARCH selection for monthly shooting time series**

So that the optimal orders for modeling the Log monthly shooting training data are (5,1,5)x(1,0). The p-values for the Ljung-Box test for the residuals are small but for the squared residuals are large, indicating that there is serial correlation but no heteroskedasticity present in the model residuals of this ARIMA-GARCH model.

In the end, we use the 1-time rolling prediction to predict the log monthly total data in 2022:

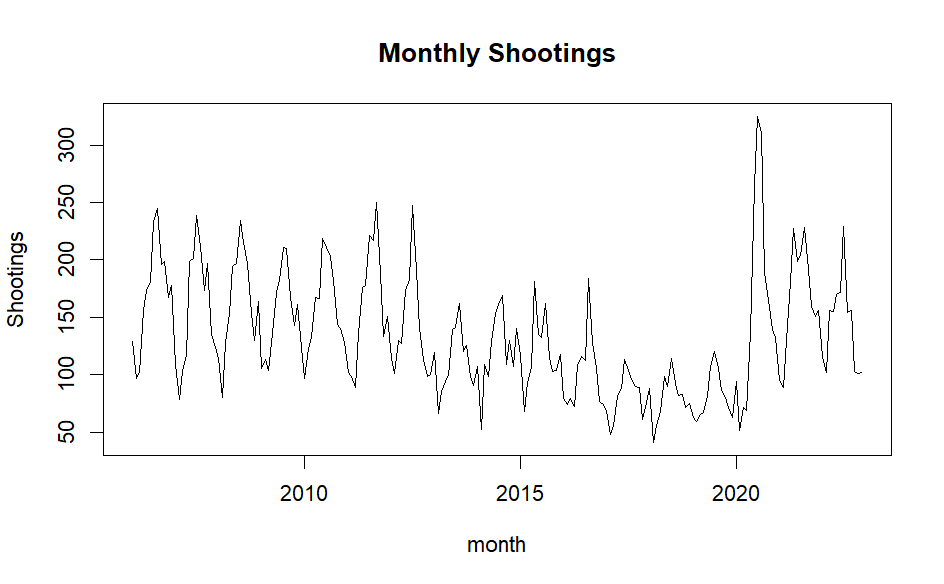


**Figure 6: 1-step-ahead rolling prediction of the next 12 points along with the test data**

This plot shows that the predictions approximate the actual Log total monthly shooting data very closely.

## Multivariate Analysis

The monthly shooting data shows a declining trend and strong seasonality, with a disruptive spike at the beginning of 2020 as a result of Covid. Thirteen other time series are curated to help predict the monthly shooting total in NYC, they are economic indicators from the Census Bureau and weather data from NOAA. Public available economic datasets are mostly by month, hence the entire multivariate analysis will be conducted on the monthly-level.



**Figure 7: monthly shooting total time series**

### Granger Causality

All thirteen time series are investigated on their granger causality with the monthly shootings using the grangertest(order=3) function in r, the order is set to 3 due to the clear seasonality pattern observed in the monthly shooting time series. Three time series granger causes monthly shooting incidents, they are labor partition rate, unemployment rate and average temperature.

Labor partition rate is the percentage of the total population in the labor force regardless of employment, and the unemployment rate is the percentage of the labor force without a job. These two indicators can be easily inferred as a person with a stable job is much less likely to commit a crime.

However, the causal relationship between weather and crime is unexpected and needs further deduction. These three significant factors will be deployed in the multivariate time series analysis in the following sections

| **Time Series** | **Granger Test P-value** | **Time Series** | **Granger Test P-value** |
| --- | --- | --- | --- |
| Condo Price | 0.311 | High Education Rate | 0.601 |
| House Price | 0.932 | Home Ownership Rate | 0.393 |
| Income Inequality Index | 0.3 | Hourly Wage | 0.337 |
| Labor Partition Rate | 0.034 | GDP in Tristate Area | 0.421 |
| Median Income | 0.826 | Unemployment Rate | 0.00017 |
| Residential Population | 0.0634 | Avg. Temperature | 2.417e-8 |
| State Min. Wage | 0.898 |  |  |

**Table 8: p-value table of multivariate granger test to monthly shooting total**

### VAR

#### Model Summary

The VAR model aims to give accurate predictions of the monthly shooting total city wide to generate prior expectation to help the policing department set up predictive plannings. The last 2 years (24 data points) are segregated to use as a validation set, others are used in training . The aforementioned granger-cause time series together with monthly shooting total are used as multivariate time series. We use unrestricted var and selected order p=5 based on both AIC and HQ.

The equation for Shooting:

Shootings = Shootings.l1 + labor.l1 + unr.l1 + weather..l1 + Shootings.l2 + labor.l2 + unr.l2 + weather.l2 + Shootings.l3 + labor.l3 + unr.l3 + weather.l3 + Shootings.l4 + labor.l4 + unr.l4 + weather.l4 + Shootings.l5 + labor.l5 + unr.l5 + weather.l5 + const + trend

Significant factors are:

| **Shooting.l1** | **Shooting.l2** | **const** | **trend** |
| --- | --- | --- | --- |
| 4.53e-06 | 0.0184 | 0.0217 | 0.0123 |

**Table 9: table of significant VAR coefficients**

Only the shooting time series itself, constant and trend are significant coefficients, this shows that the VAR(p) model might be too restrictive using endogenous, and cannot capture cross-dependencies.

#### Goodness of fit: Failed

##### Normality

JB test & Skewness only test & Kurtosis only test (multivariate): p-value < 2.2e-16.

Reject null hypothesis and therefore the model does not follow normality assumptions.

##### Heteroskedasticity

Arch test: p-value=0.411

Cannot reject null hypothesis and therefore cannot reject that the residual does not have heteroskedasticity.

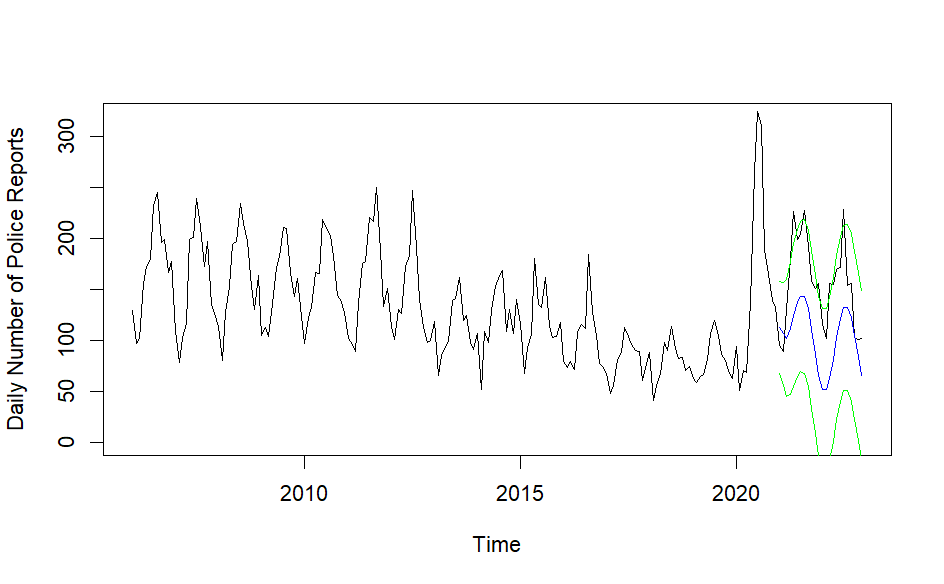
##### Serial Correlation

Portmanteau Test: p-value = 0.6293

Cannot reject null hypothesis and therefore cannot reject that the model does not have serial correlation.

#### Prediction Results

The VAR model achieves mse=3754.017, is noticeably lower than the observed data and can barely contain it within 95% confidence interval. Therefore, the prediction is not a good fit and requires exogenous multivariate models.



**Figure 8: VAR model prediction of the next 24 months along with the test data**

### ARMAX

#### Model Summary

The ARMAX model aims to give accurate predictions of the monthly shooting total city wide to generate prior expectation to help the policing department set up predictive plannings. The last 2 years (24 data points) are segregated to use as a validation set, others are used in training . We selected ARMAX up to orders p,q = 0~4 and r,s = 0~2 with period set to 12. without differencing and using the three aforementioned granger-cause time series as exogenous factors. The order selected by AIC is (4,0,4)x(1,0,0).

Only the intercept and sar1 parameters are significant. While this suggests exogenous factors are not significant, removing them and using an ARIMA model alone would result in substantial drop in performance. Considering the good fit of the final ARMAX model, and their granger causality, the exogenous factors should be considered influential to the model.

| **ar1** | **ar2** | **ar3** | **ar4** | **ma1** | **ma2** | **ma3** | **ma4** | **sar1** | **intercept** | **xreg1** | **xreg2** | **xreg3** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 8.39E-01 | 1.53E-01 | 9.63E-01 | 5.63E-01 | 2.58E-01 | 9.62E-01 | 9.88E-01 | 7.69E-01 | 3.54E-10 | 5.70E-16 | 7.92E-01 | 7.54E-01 | 7.13E-01 |

**Table 10:**

#### Goodness of fit: Satisfied

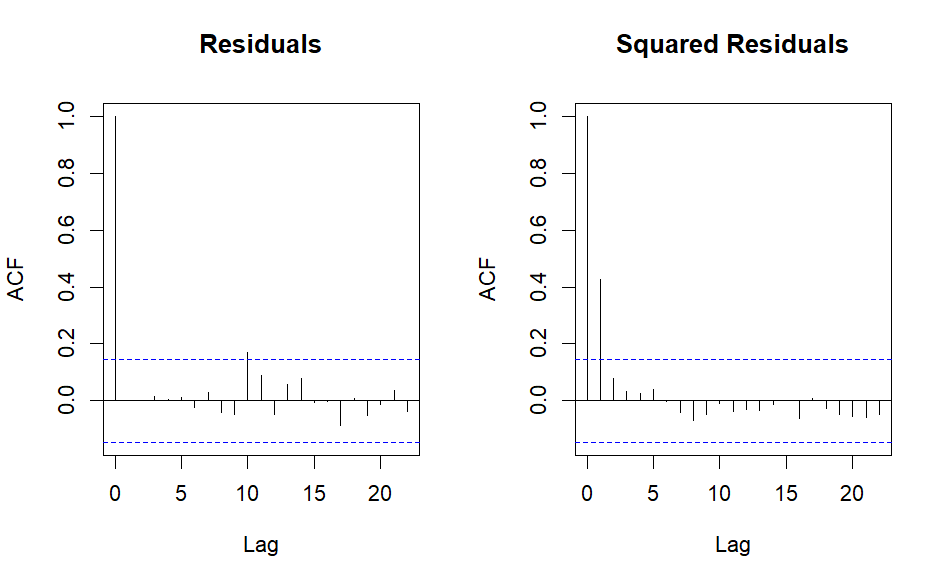
##### Normality

Shapiro-Wilk normality test on ARMAX residuals: p-value = 0.0067.

Cannot reject null hypothesis and therefore cannot reject that the residual follows normality assumption.

##### Heteroskedasticity

The residual and squared residual both show signs of stationarity, indicating that there is no heteroskedasticity left in the time series.



**Figure 9-1: ARMAX residual ACF plot**

**Figure 9-2: ARMAX squared residual ACF plot**

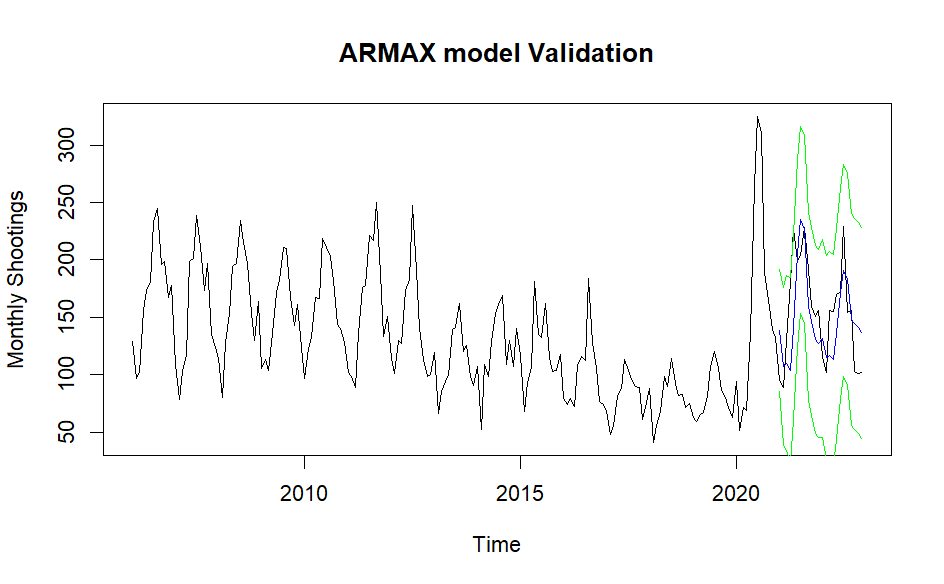
##### Serial Correlation

Box-Ljung test on ARMAX residuals (lag=(p+q+1),df=(p+q)): p-value = 0.2866

Cannot reject null hypothesis and therefore cannot reject that the residual have on serial correlation.

#### Prediction Results

The ARMAX model achieves mse=1294.655, stays close to the observed data and contains it within 95% confidence interval.



**Figure 10: ARMAX model prediction of the next 24 months along with the test data**

# Subject Matter Statements

It is important to note that the data provided are simply records of incidents with very basic and hopefully objective information recorded. The dataset does not provide insights into and motivation or bias of the individuals involved. Any conclusions that we make as a result of this study represent further points to investigate and experiment as our data does not provide a comprehensive context of the incidents reported.

# Discussion and Future Research

Areas of future research include investigating the effect of police presence on shooting incidents and experimenting with the addition/subtraction of police units. Additionally, situational factors like police callout reasoning and officer demographics can be studied alongside this investigation for further corollary research. We aim to use the trends and patterns discovered to further describe the evolving atmosphere of violence in New York City. Further applications of our work can explore the effects of the COVID-19 pandemic on shooting incidents as we noticed an increase around the beginning of this time. We also hope to incorporate additional datasets into our multivariate causal analysis.

# References

1. Grieco, J. & Robbins, T.G. (2019).Officer-involved shooting situations, responses, and data: An analysis of information from major city police agencies. *National Policing Institute and Arnold Ventures*. Retrieved https://www.policinginstitute.org/publication/officer-involved-shooting-situations-responses-and-data-an-analysis-of-information-from-major-city-police-agencies-2/
2. Lett, E., Asabor, E. N., Corbin, T., & Boatright, D. (2020). Racial inequity in fatal US police shootings, 2015–2020. *Journal of Epidemiology and Community Health, 75*(4), 394–397. https://doi.org/10.1136/jech-2020-215097
3. Terrill, W., & Mastrofski, S. D. (2002). Situational and officer-based determinants of police coercion. *Justice Quarterly, 19*(2), 215–248. https://doi.org/10.1080/07418820200095221
4. White, M.D. (2002). Identifying situational predictors of police shootings using multivariate analysis. *Policing: An International Journal, 25*(4), 726-751. https://doi.org/10.1108/13639510210450659