NYP Shooting Incident - Data Analysis

2024-03-03

```
knitr::opts_chunk$set(echo = TRUE)
```

Introduction

This report aims to provide an analysis of the historical NYPD shooting incident data. Our goal is to understand and identify the possible trends, patterns, and any underlying issues within the data. We also want to see if it's possible to reliably forecast the number of shooting incident over time. We will also do our best to acknowledge potential biases in data collection and analysis, aiming for an objective analysis.

Data Import and Description

The NYPD shooting incident dataset is a historical compilation of shooting incidents reported by the New York Police Department. This section outlines the steps to import and initially describe the dataset.

```
# Load necessary libraries
library(readr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Import dataset
url_NYPD <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
shooting_data <- read.csv(url_NYPD)

# Display the structure and summary of the dataset using the new variable name
str(shooting_data)
```

```
## 'data.frame':    27312 obs. of  21 variables:
##  $ INCIDENT_KEY           : int  228798151 137471050 147998800 146837977 58921844 219559682 85295722
##  $ OCCUR_DATE             : chr  "05/27/2021" "06/27/2014" "11/21/2015" "10/09/2015" ...
##  $ OCCUR_TIME             : chr  "21:30:00" "17:40:00" "03:56:00" "18:30:00" ...
```

```
##  $ BORO                    : chr  "QUEENS" "BRONX" "QUEENS" "BRONX" ...
##  $ LOC_OF_OCCUR_DESC       : chr  "" "" "" "" ...
##  $ PRECINCT                : int  105 40 108 44 47 81 114 81 105 101 ...
##  $ JURISDICTION_CODE       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LOC_CLASSFCTN_DESC      : chr  "" "" "" "" ...
##  $ LOCATION_DESC           : chr  "" "" "" "" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "false" "true" "false" ...
##  $ PERP_AGE_GROUP          : chr  "" "" "" "" ...
##  $ PERP_SEX                : chr  "" "" "" "" ...
##  $ PERP_RACE               : chr  "" "" "" "" ...
##  $ VIC_AGE_GROUP           : chr  "18-24" "18-24" "25-44" "<18" ...
##  $ VIC_SEX                 : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE                : chr  "BLACK" "BLACK" "WHITE" "WHITE HISPANIC" ...
##  $ X_COORD_CD              : num  1058925 1005028 1007668 1006537 1024922 ...
##  $ Y_COORD_CD              : num  180924 234516 209837 244511 262189 ...
##  $ Latitude                : num  40.7 40.8 40.7 40.8 40.9 ...
##  $ Longitude               : num  -73.7 -73.9 -73.9 -73.9 -73.9 ...
##  $ Lon_Lat                 : chr  "POINT (-73.73083868899994 40.662964620000025)" "POINT (-73.9249423
```

```
summary(shooting_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class :character   Class :character
##  Median : 90372218   Mode  :character   Mode  :character   Mode  :character
##  Mean   :120860536
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Length:27312            Length:27312
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE          VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE            X_COORD_CD        Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
```

```
##  Class :character   1st Qu.:1000028   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

Data Cleaning

In this step we will convert appropriate variables to factor and date types and remove unnecessary columns.

```
# Convert date columns to Date type
shooting_data$OCCUR_DATE <- as.Date(shooting_data$OCCUR_DATE, format="%m/%d/%Y")

# Convert categorical variables to factors
categorical_vars <- c("BORO", "PRECINCT", "JURISDICTION_CODE",
                      "VIC_SEX", "VIC_RACE", "PERP_SEX", "PERP_RACE")

shooting_data[categorical_vars] <- lapply(shooting_data[categorical_vars],
                                          factor)

# Check the structure after cleaning
str(shooting_data)
```

```
## 'data.frame':    27312 obs. of  21 variables:
##  $ INCIDENT_KEY          : int  228798151 137471050 147998800 146837977 58921844 219559682 85295722
##  $ OCCUR_DATE            : Date, format: "2021-05-27" "2014-06-27" ...
##  $ OCCUR_TIME            : chr  "21:30:00" "17:40:00" "03:56:00" "18:30:00" ...
##  $ BORO                  : Factor w/ 5 levels "BRONX","BROOKLYN",..: 4 1 4 1 1 2 4 2 4 4 ...
##  $ LOC_OF_OCCUR_DESC     : chr  "" "" "" "" ...
##  $ PRECINCT              : Factor w/ 77 levels "1","5","6","7",..: 63 23 66 27 30 52 72 52 63 59 ..
##  $ JURISDICTION_CODE     : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LOC_CLASSFCTN_DESC    : chr  "" "" "" "" ...
##  $ LOCATION_DESC         : chr  "" "" "" "" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "false" "true" "false" ...
##  $ PERP_AGE_GROUP        : chr  "" "" "" "" ...
##  $ PERP_SEX              : Factor w/ 5 levels "","(null)","F",..: 1 1 1 1 4 1 1 1 1 4 ...
##  $ PERP_RACE             : Factor w/ 9 levels "","(null)","AMERICAN INDIAN/ALASKAN NATIVE",..: 1 1 1
##  $ VIC_AGE_GROUP         : chr  "18-24" "18-24" "25-44" "<18" ...
##  $ VIC_SEX               : Factor w/ 3 levels "F","M","U": 2 2 2 2 2 2 2 2 2 2 ...
##  $ VIC_RACE              : Factor w/ 7 levels "AMERICAN INDIAN/ALASKAN NATIVE",..: 3 3 6 7 3 3 3 3 3
##  $ X_COORD_CD            : num  1058925 1005028 1007668 1006537 1024922 ...
##  $ Y_COORD_CD            : num  180924 234516 209837 244511 262189 ...
##  $ Latitude              : num  40.7 40.8 40.7 40.8 40.9 ...
##  $ Longitude             : num  -73.7 -73.9 -73.9 -73.9 -73.9 ...
##  $ Lon_Lat               : chr  "POINT (-73.73083868899994 40.662964620000025)" "POINT (-73.9249423
```

Next, we are going to do some additional data cleaning (missing date, duplicates)

```
shooting_data <- shooting_data %>%
  filter(!is.na(OCCUR_DATE))

shooting_data <- shooting_data %>%
  distinct()

str(shooting_data)
```

```
## 'data.frame':    27312 obs. of  21 variables:
##  $ INCIDENT_KEY          : int  228798151 137471050 147998800 146837977 58921844 219559682 85295722
##  $ OCCUR_DATE            : Date, format: "2021-05-27" "2014-06-27" ...
##  $ OCCUR_TIME            : chr  "21:30:00" "17:40:00" "03:56:00" "18:30:00" ...
##  $ BORO                  : Factor w/ 5 levels "BRONX","BROOKLYN",..: 4 1 4 1 1 2 4 2 4 4 ...
##  $ LOC_OF_OCCUR_DESC     : chr  "" "" "" "" ...
##  $ PRECINCT              : Factor w/ 77 levels "1","5","6","7",..: 63 23 66 27 30 52 72 52 63 59 ..
##  $ JURISDICTION_CODE     : Factor w/ 3 levels "0","1","2": 1 1 1 1 1 1 1 1 1 1 ...
##  $ LOC_CLASSFCTN_DESC    : chr  "" "" "" "" ...
##  $ LOCATION_DESC         : chr  "" "" "" "" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "false" "true" "false" ...
##  $ PERP_AGE_GROUP        : chr  "" "" "" "" ...
##  $ PERP_SEX              : Factor w/ 5 levels "","(null)","F",..: 1 1 1 1 4 1 1 1 1 4 ...
##  $ PERP_RACE             : Factor w/ 9 levels "","(null)","AMERICAN INDIAN/ALASKAN NATIVE",..: 1 1 1
##  $ VIC_AGE_GROUP         : chr  "18-24" "18-24" "25-44" "<18" ...
##  $ VIC_SEX               : Factor w/ 3 levels "F","M","U": 2 2 2 2 2 2 2 2 2 2 ...
##  $ VIC_RACE              : Factor w/ 7 levels "AMERICAN INDIAN/ALASKAN NATIVE",..: 3 3 6 7 3 3 3 3 3
##  $ X_COORD_CD            : num  1058925 1005028 1007668 1006537 1024922 ...
##  $ Y_COORD_CD            : num  180924 234516 209837 244511 262189 ...
##  $ Latitude              : num  40.7 40.8 40.7 40.8 40.9 ...
##  $ Longitude             : num  -73.7 -73.9 -73.9 -73.9 -73.9 ...
##  $ Lon_Lat               : chr  "POINT (-73.73083868899994 40.662964620000025)" "POINT (-73.9249423
```

We notice that the number of observations didn't change, meaning that there was no duplicates or missing date record.

Lastly, as our research question is about the month of the year, we will create an additional column which contains the month of the incident based on the date

```
shooting_data <- shooting_data %>%
  mutate(month = format(OCCUR_DATE, "%m"))

shooting_data <- shooting_data %>%
  mutate(month = factor(month, levels = c("01", "02", "03", "04", "05", "06",
                                          "07", "08", "09", "10", "11", "12"),
                        labels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun",
                                   "Jul", "Aug", "Sep", "Oct", "Nov", "Dec")))
```

Data Visualization and Analysis

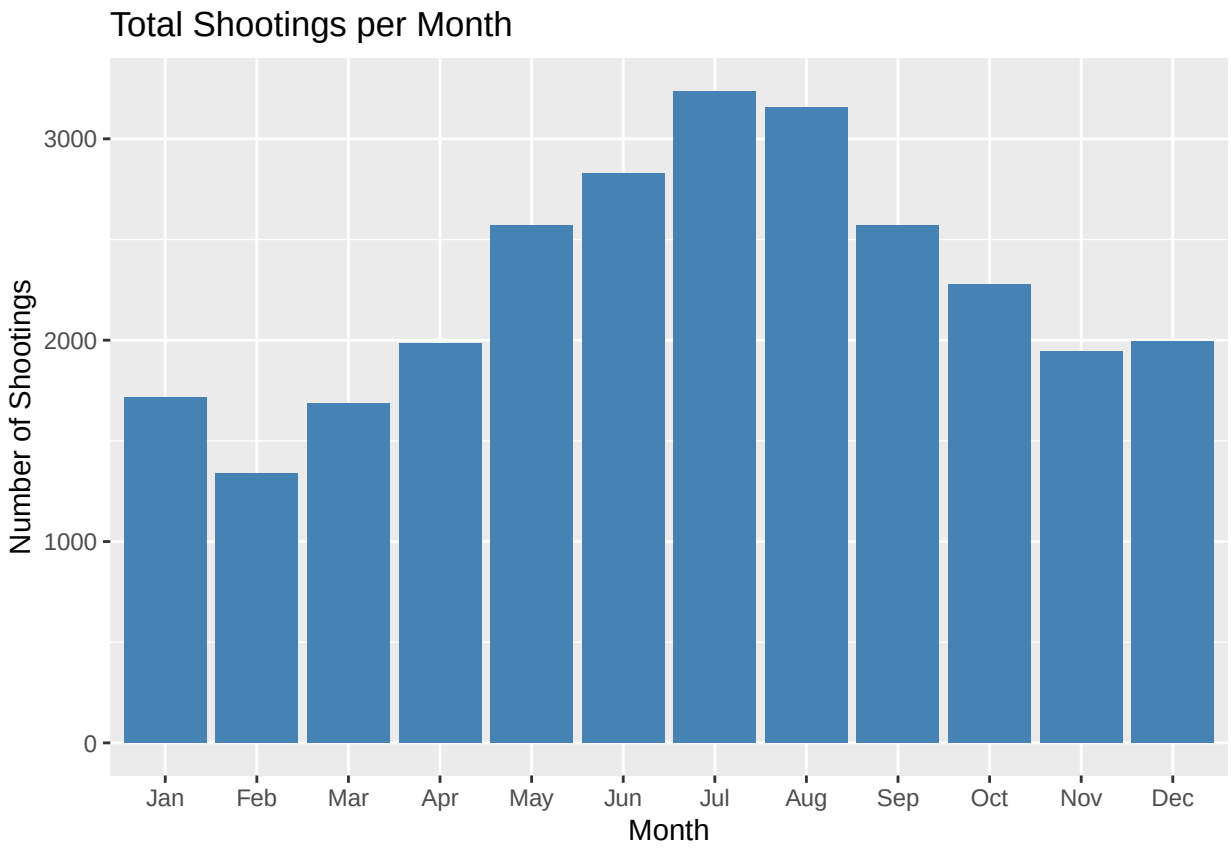
We will create a few different visualizations to explore the dataset further and perform some basic analysis.

```
summary(shooting_data)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:27312
##  1st Qu.: 63860880   1st Qu.:2009-07-18   Class :character
```

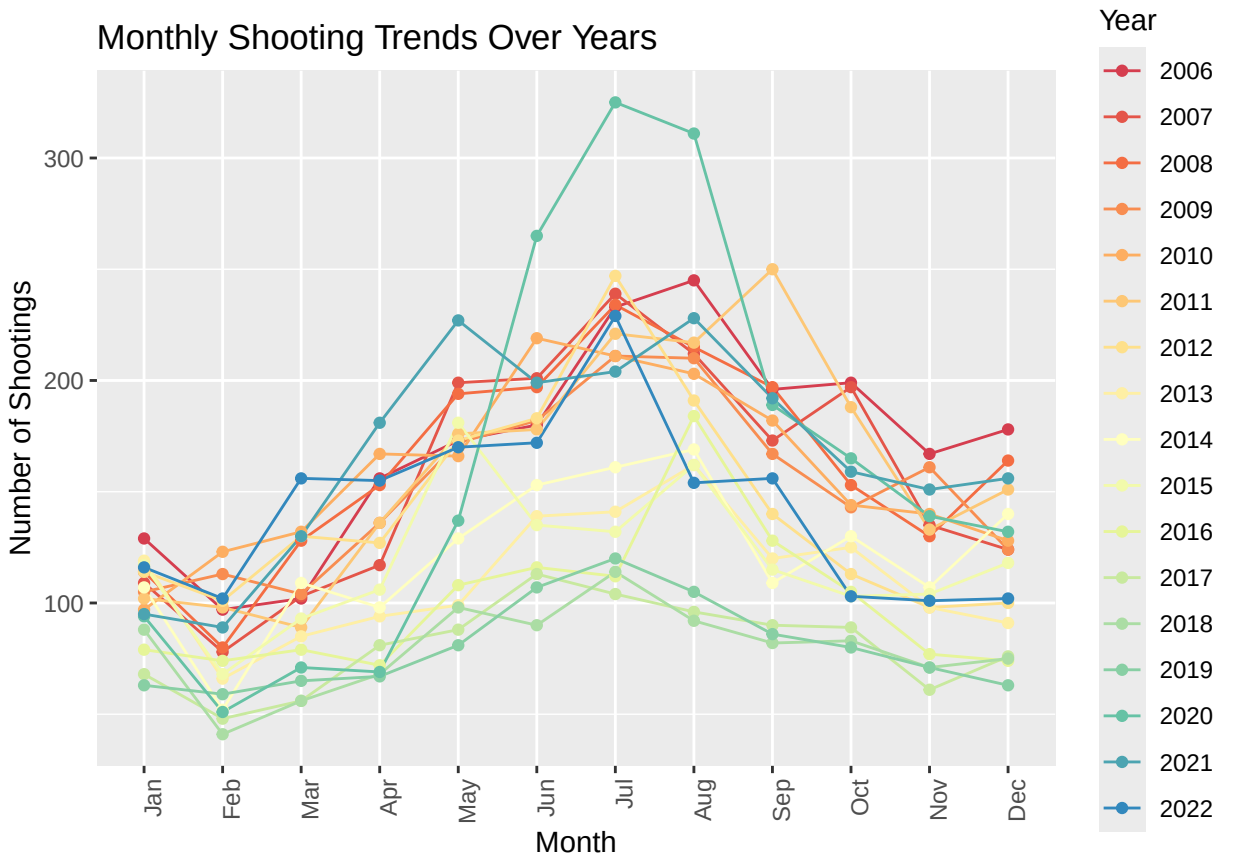
```
##  Median : 90372218   Median :2013-04-29   Mode  :character
##  Mean   :120860536   Mean   :2014-01-06
##  3rd Qu.:188810230   3rd Qu.:2018-10-15
##  Max.   :261190187   Max.   :2022-12-31
##
##            BORO       LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE
##  BRONX        : 7937   Length:27312       75     : 1557   0   :22809
##  BROOKLYN     :10933   Class :character   73     : 1452   1   :   74
##  MANHATTAN    : 3572   Mode  :character   67     : 1216   2   : 4427
##  QUEENS       : 4094                      44     : 1020   NA's:    2
##  STATEN ISLAND:  776                      79     : 1012
##                                           47     :  953
##                                           (Other):20102
##  LOC_CLASSFCTN_DESC LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  PERP_AGE_GROUP       PERP_SEX              PERP_RACE     VIC_AGE_GROUP
##  Length:27312            : 9310   BLACK         :11432   Length:27312
##  Class :character   (null):  640                 : 9310   Class :character
##  Mode  :character   F     :  424   WHITE HISPANIC: 2341   Mode  :character
##                     M     :15439   UNKNOWN       : 1836
##                     U     : 1499   BLACK HISPANIC: 1314
##                                     (null)        :  640
##                                     (Other)       :  439
##  VIC_SEX                              VIC_RACE       X_COORD_CD
##  F: 2615   AMERICAN INDIAN/ALASKAN NATIVE:   10   Min.   : 914928
##  M:24686   ASIAN / PACIFIC ISLANDER      :  404   1st Qu.:1000028
##  U:   11   BLACK                         :19439   Median :1007731
##            BLACK HISPANIC                : 2646   Mean   :1009449
##            UNKNOWN                       :   66   3rd Qu.:1016838
##            WHITE                         :  698   Max.   :1066815
##            WHITE HISPANIC                : 4049
##    Y_COORD_CD        Latitude       Longitude        Lon_Lat
##  Min.   :125757   Min.   :40.51   Min.   :-74.25   Length:27312
##  1st Qu.:182834   1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :194487   Median :40.70   Median :-73.92   Mode  :character
##  Mean   :208127   Mean   :40.74   Mean   :-73.91
##  3rd Qu.:239518   3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :271128   Max.   :40.91   Max.   :-73.70
##                   NA's   :10      NA's   :10
##      month
##  Jul    : 3238
##  Aug    : 3156
##  Jun    : 2829
##  Sep    : 2572
##  May    : 2571
##  Oct    : 2279
##  (Other):10667
```

```r
library(ggplot2)

ggplot(shooting_data, aes(x=month)) +
  geom_bar(fill="steelblue") +
  xlab("Month") + ylab("Number of Shootings") +
  ggtitle("Total Shootings per Month")
```




```r
library(RColorBrewer)

colors <- colorRampPalette(brewer.pal(9, "Spectral"))(17)

# Add 'year' and 'month' columns
shooting_data <- shooting_data %>%
  mutate(year = format(OCCUR_DATE, "%Y"),
         month = format(OCCUR_DATE, "%m"),
         month_year = paste(year, month, sep = "-")) %>%
  mutate(month = factor(month, levels = sprintf("%02d", 1:12),
                        labels = c("Jan", "Feb", "Mar", "Apr", "May", "Jun",
                                   "Jul", "Aug", "Sep", "Oct", "Nov", "Dec")))

# Aggregate shootings per month-year
monthly_counts <- shooting_data %>%
  group_by(month, year) %>%
  summarise(shootings = n(), .groups = 'drop')

ggplot(monthly_counts, aes(x = month, y = shootings, group = year, color = factor(year))) +
  geom_line() +
```

```
  geom_point() +
  scale_color_manual(values = colors) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(x = "Month", y = "Number of Shootings",
       title = "Monthly Shooting Trends Over Years", color = "Year")
```



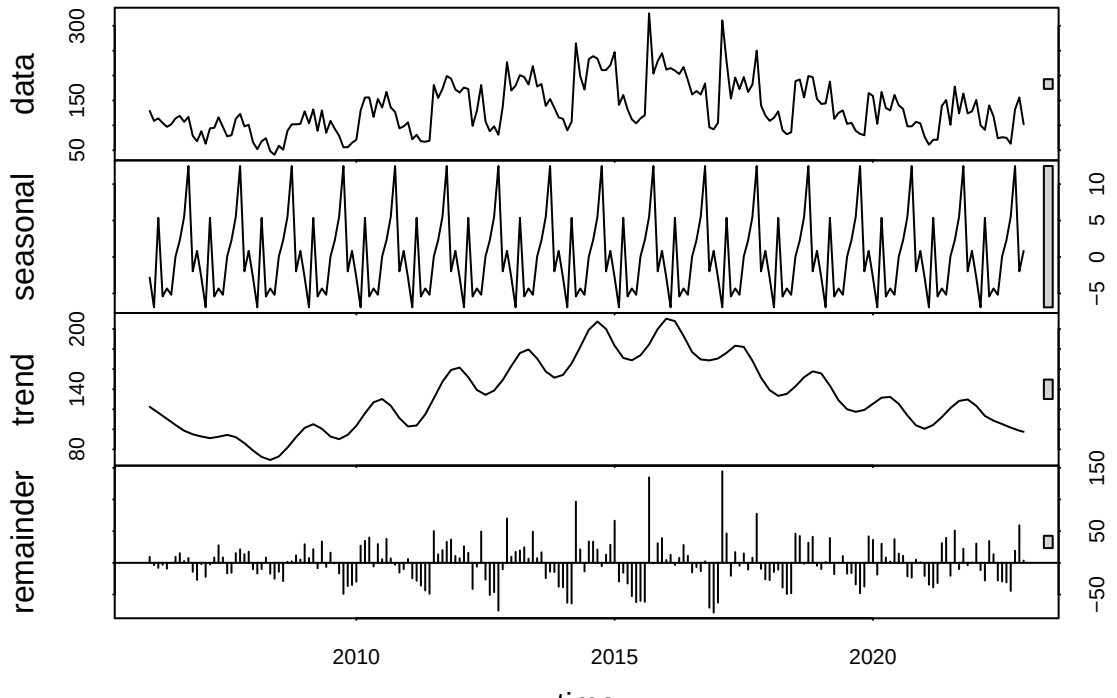
From those two visualizations, it seems that there might be some complex effects at play : - Most of the years have higher number of incident during summer - The trend number of incident seems to be decreasing over the years (less shootings in 2014-2020 than in 2006-2011) but recently there as been some increase (2020-2022)

In order to explore those effects, we need to use time-series analysis technique to explore further the seasonality and trends of the number of shooting incidents

```
library(forecast)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```
ts_data <- ts(monthly_counts$shootings, frequency=12, start=c(2006,1))
decomposed <- stl(ts_data, s.window="periodic")
plot(decomposed)
```



After performing STL analysis, we can notice that there is indeed a seasonality in the number of shooting incidents and a trend effects that was increasing from 2010 to 2015 and decreasing from 2015 to 2020.

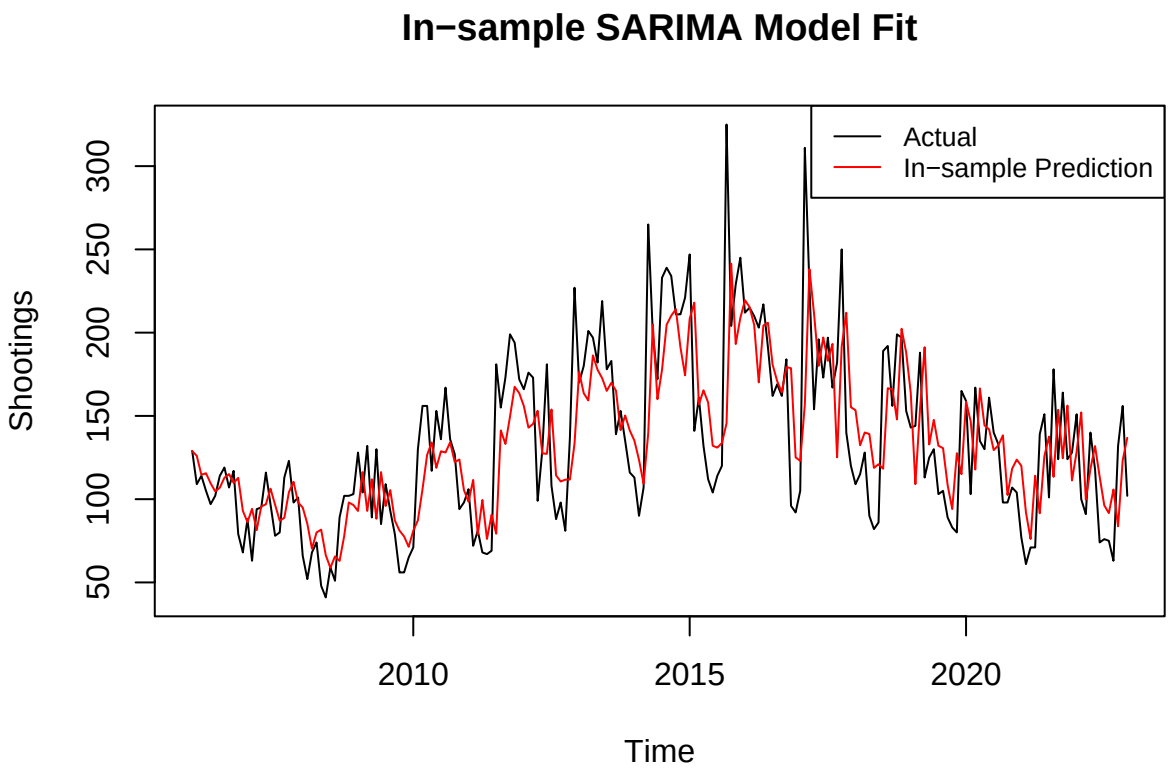
Though there some large remainder, especially around the year 2015 which suggest that this decompostion might not account perfectly for the variability in the number of shooting incidents.

We will now try to forecast the number shooting incidents using SARIMA model

```
sarima_model_seasonal <- auto.arima(ts_data, seasonal = TRUE,
                                    stepwise = FALSE, approximation = FALSE)

# Use the model to predict the in-sample values
in_sample_forecasts <- fitted(sarima_model_seasonal)

# Quantitative Evaluation: Compare the in-sample forecasts to the actual data
actuals <- ts_data
predictions <- in_sample_forecasts

mae <- mean(abs(actuals - predictions))
rmse <- sqrt(mean((actuals - predictions)^2))
mape <- mean(abs((actuals - predictions) / actuals), na.rm = TRUE) * 100

cat("In-sample MAE:", mae, "\nIn-sample RMSE:",
                rmse, "\nIn-sample MAPE:", mape, "%\n")
```

```
## In-sample MAE: 26.99603
## In-sample RMSE: 36.37683
## In-sample MAPE: 21.80282 %
```

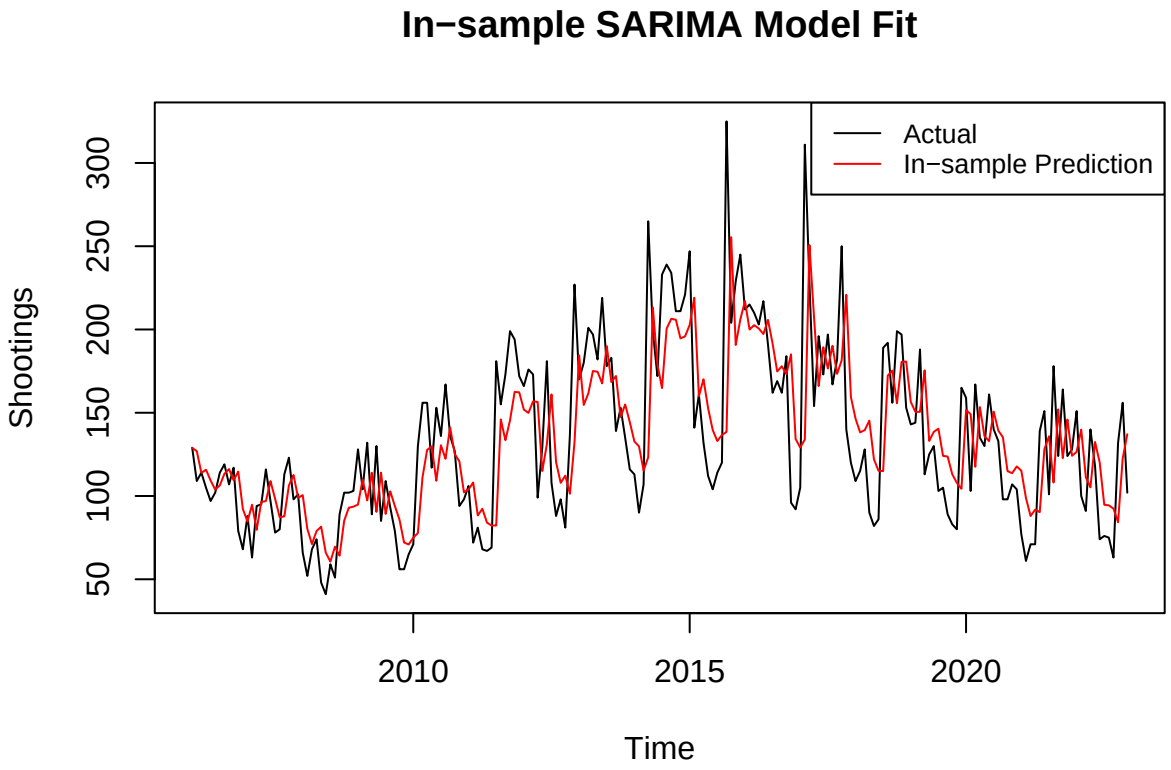
```
# Visual Evaluation: Compare the in-sample forecasts to the actual data
plot(ts_data, main="In-sample SARIMA Model Fit", ylab="Shootings")
lines(predictions, col = 'red')
legend("topright", legend=c("Actual", "In-sample Prediction"),
       col=c("black", "red"), lty=1, cex=0.8)
```




```
sarima_model_non_seasonal <- auto.arima(ts_data, seasonal = FALSE,
                                        stepwise = FALSE, approximation = FALSE)

# Use the model to predict the in-sample values
in_sample_forecasts <- fitted(sarima_model_non_seasonal)

# Quantitative Evaluation: Compare the in-sample forecasts to the actual data
actuals <- ts_data
predictions <- in_sample_forecasts

mae <- mean(abs(actuals - predictions))
rmse <- sqrt(mean((actuals - predictions)^2))
mape <- mean(abs((actuals - predictions) / actuals), na.rm = TRUE) * 100

cat("In-sample MAE:", mae, "\nIn-sample RMSE:", rmse, "\nIn-sample MAPE:",
    mape, "%\n")
```

```
## In-sample MAE: 27.74873
## In-sample RMSE: 37.68257
## In-sample MAPE: 22.1899 %
```

```
# Visual Evaluation: Compare the in-sample forecasts to the actual data
plot(ts_data, main="In-sample SARIMA Model Fit", ylab="Shootings")
lines(predictions, col = 'red')
legend("topright", legend=c("Actual", "In-sample Prediction"),
       col=c("black", "red"), lty=1, cex=0.8)
```



Additional Questions Raised

- Would incorporating geographic or demographic data provide more precise forecasts
- What accounts for the substantial errors indicated by the MAPE and RMSE? Are there underlying trends or cyclic patterns not captured by the current model, or do external shocks or events drive these discrepancies?
- Given the observed limitations of the SARIMA model, what alternative forecasting models or machine learning approaches might yield better predictive performance?

Conclusion

The combination of STL decomposition and SARIMA modeling has provided valuable insights into the seasonal patterns and long-term trends of shooting incidents. However, the performance metrics reveal a need for model refinement and possibly the exploration of alternative forecasting approaches to enhance predictive accuracy. The analysis underscores the complex nature of crime trends, suggesting that factors beyond historical patterns likely influence the incidence of shootings.

The substantial MAPE and RMSE values, in particular, highlight the challenges in forecasting such incidents accurately, emphasizing the potential role of unmodeled external influences or events. This analysis serves as a foundational step, illuminating the path for further research that could incorporate a broader array of data sources, consider external socio-economic factors, and apply more sophisticated modeling techniques.

In conclusion, while the current model offers a baseline understanding of shooting incident dynamics, the insights gained and questions raised from this analysis underscore the importance of continued research. Such efforts should aim not only to improve forecasting accuracy but also to enhance the practical application of these forecasts in policy-making, law enforcement resource allocation, and community safety initiatives.

Consideration of Bias

Data Collection Bias: The potential for under-reporting or mis-classification remains a concern. Efforts to cross-reference incident data with other crime reporting databases could mitigate some of these issues, ensuring a more comprehensive dataset.

Analysis Bias: While we have aimed for objective analysis methods, biases towards certain boroughs or demographic groups could influence interpretation. By expanding our analysis to include socio-economic and demographic factors, and by employing statistical controls where appropriate, we aim to provide a more balanced and nuanced understanding of the data.

Modeling Bias: The assumptions and limitations inherent in the STL and SARIMA methodologies may not fully account for the complex, multifaceted nature of crime dynamics. For instance, linear assumptions might oversimplify the relationships between different variables, neglecting non-linear interactions, threshold effects, or feedback loops present in the real world.