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

# Remove unnecessary columns
shooting_data <- select(shooting_data, -c(X_COORD_CD, Y_COORD_CD, Latitude, Longitude))

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

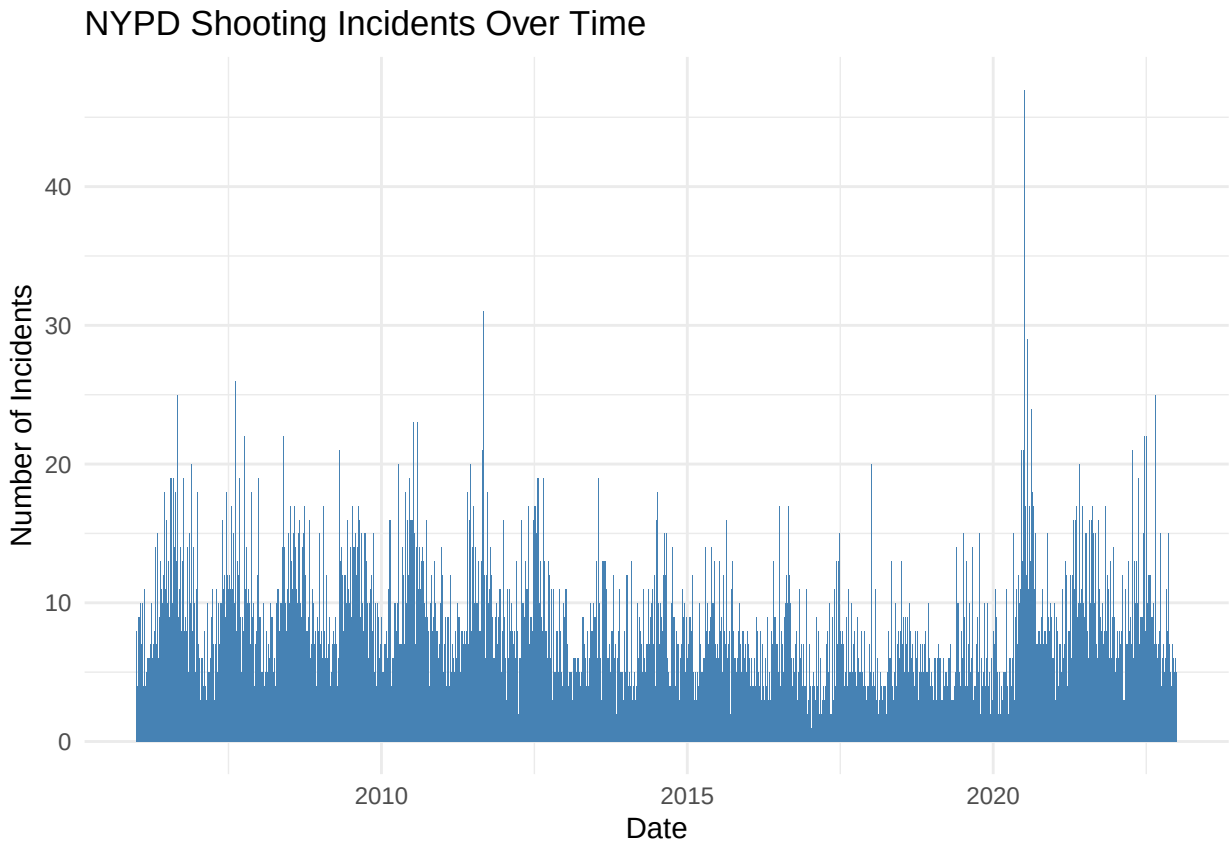
```
## 'data.frame':    27312 obs. of  17 variables:
##  $ INCIDENT_KEY           : int  228798151 137471050 147998800 146837977 58921844 219559682 85295722
##  $ OCCUR_DATE             : Date, format: "2021-05-27" "2014-06-27" ...
##  $ OCCUR_TIME             : chr  "21:30:00" "17:40:00" "03:56:00" "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
##  $ Lon_Lat                : chr  "POINT (-73.73083868899994 40.662964620000025)" "POINT (-73.9249423
```

Data Visualization and Analysis

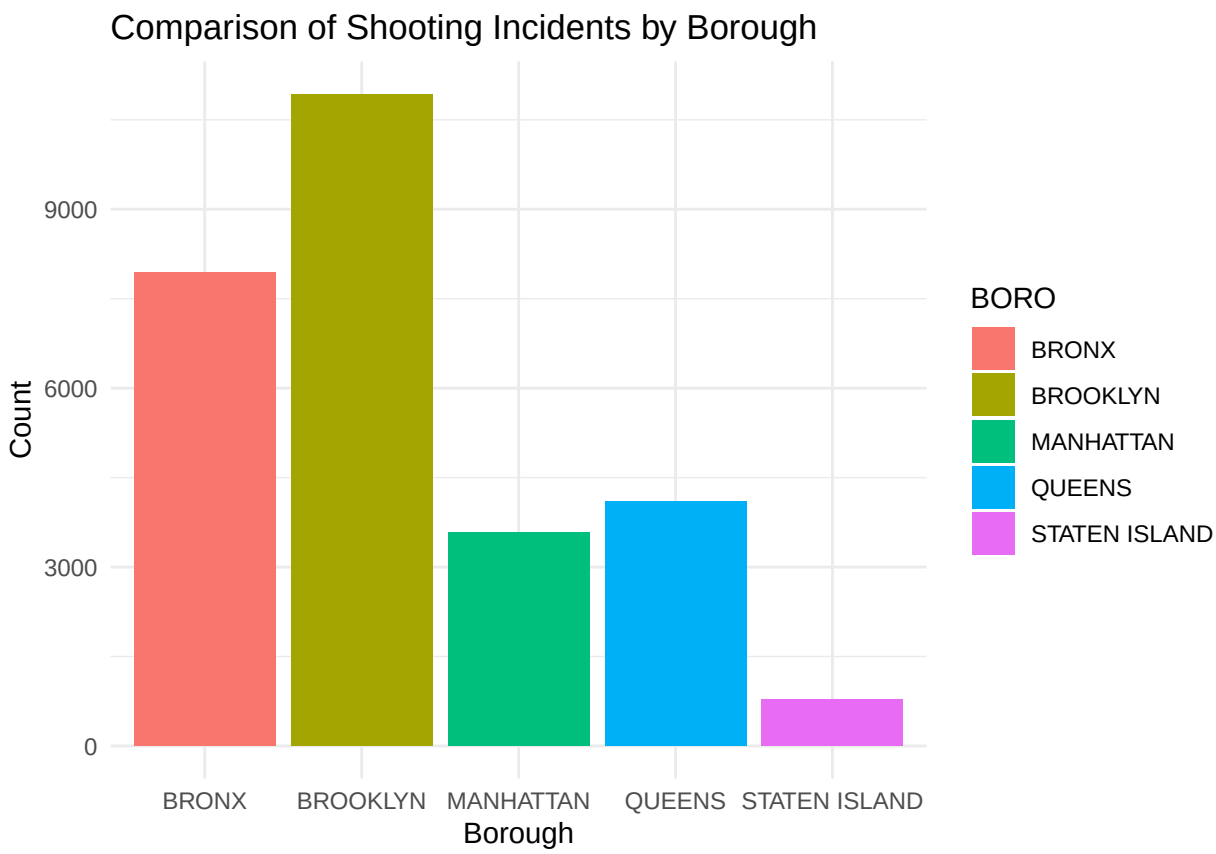
We will create two different visualizations to explore the dataset further and perform some basic analysis.

```
library(ggplot2)

ggplot(shooting_data, aes(x = OCCUR_DATE)) +
  geom_histogram(stat="count", binwidth = 10, fill="steelblue") +
  theme_minimal() +
  labs(title = "NYPD Shooting Incidents Over Time", x = "Date", y = "Number of Incidents")
```

```
## Warning in geom_histogram(stat = "count", binwidth = 10, fill = "steelblue"):
## Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```




```
ggplot(shooting_data, aes(x = BORO, fill = BORO)) +
  geom_bar() +
  theme_minimal() +
  labs(title = "Comparison of Shooting Incidents by Borough", x = "Borough", y = "Count")
```

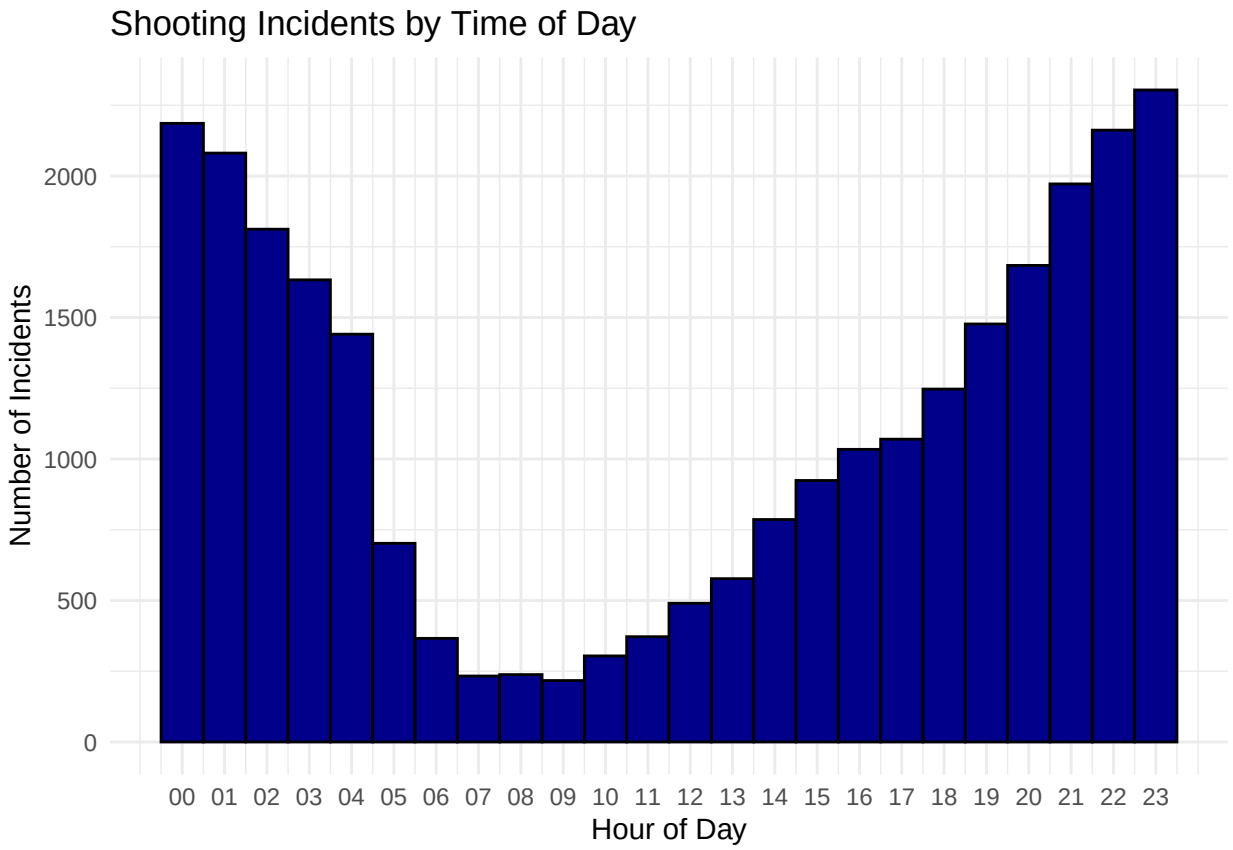



```r
library(hms)
```

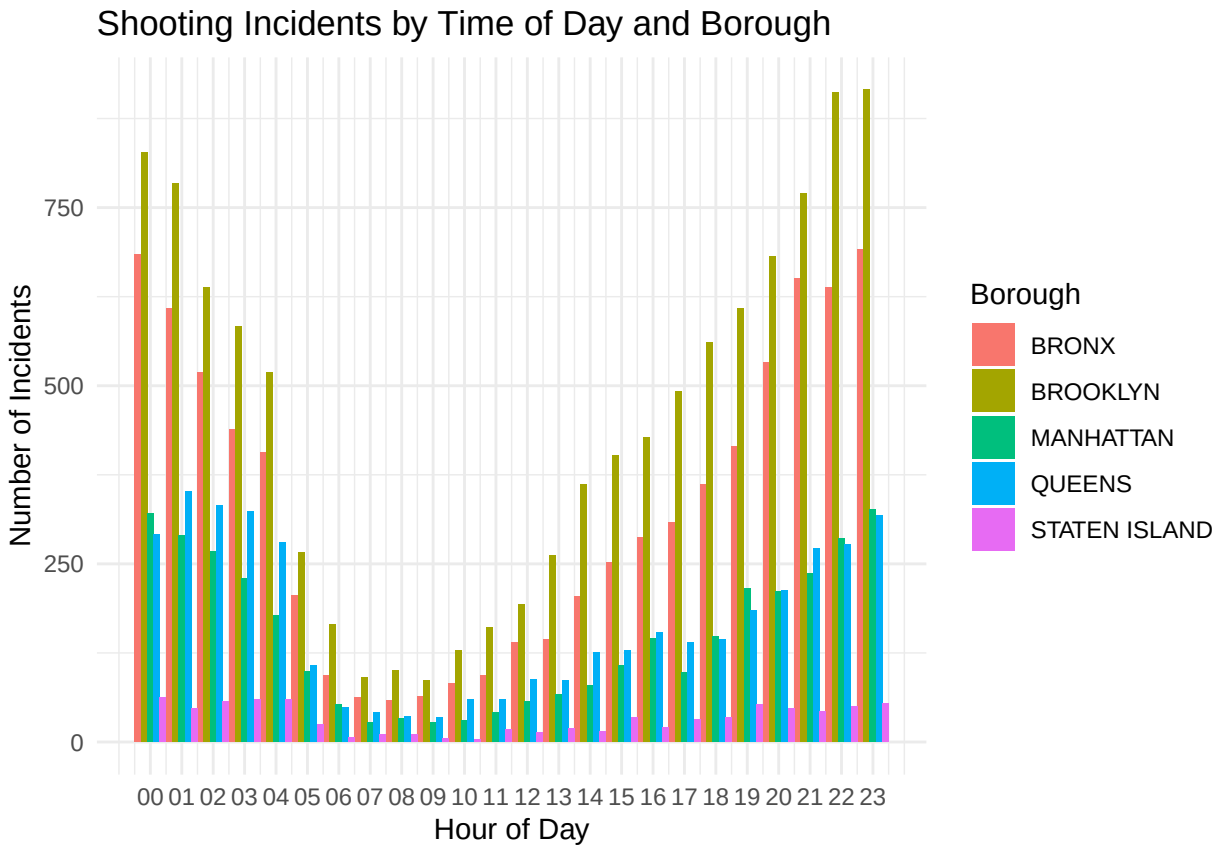
```
##
## Attaching package: 'hms'
```

```
## The following object is masked from 'package:lubridate':
##
##     hms
```

```r
library(ggplot2)

# Convert OCCUR_TIME to HMS format
shooting_data$OCCUR_TIME_HMS <- as_hms(shooting_data$OCCUR_TIME)

# Extract hour from OCCUR_TIME_HMS
shooting_data$HOUR <- hour(shooting_data$OCCUR_TIME_HMS)

# Plotting incidents by time of day
ggplot(shooting_data, aes(x = HOUR)) +
  geom_histogram(binwidth = 1, fill="darkblue", color="black") +
  scale_x_continuous(breaks = 0:23, labels = sprintf("%02d", 0:23)) +
  theme_minimal() +
  labs(title = "Shooting Incidents by Time of Day", x = "Hour of Day", y = "Number of Incidents")
```




```
ggplot(shooting_data, aes(x = HOUR, fill = BORO)) +
  geom_histogram(position = "dodge", binwidth = 1) +
  scale_x_continuous(breaks = 0:23, labels = sprintf("%02d", 0:23)) +
  theme_minimal() +
  labs(title = "Shooting Incidents by Time of Day and Borough", x = "Hour of Day", y = "Number of Incid
```



Modelling

To incorporate a model into our analysis of the NYPD shooting incident data, a simple approach could be to predict the number of shooting incidents based on time of day and borough.

Given the nature of the data, a Poisson regression model could be appropriate. Firstwe ensure your data is aggregated appropriately for the model.

```
# Aggregate data for modeling
incident_counts <- shooting_data %>%
  group_by(HOUR, BORO) %>%
  summarise(Incident_Count = n(), .groups = 'drop')

# View the aggregated data
head(incident_counts)
```

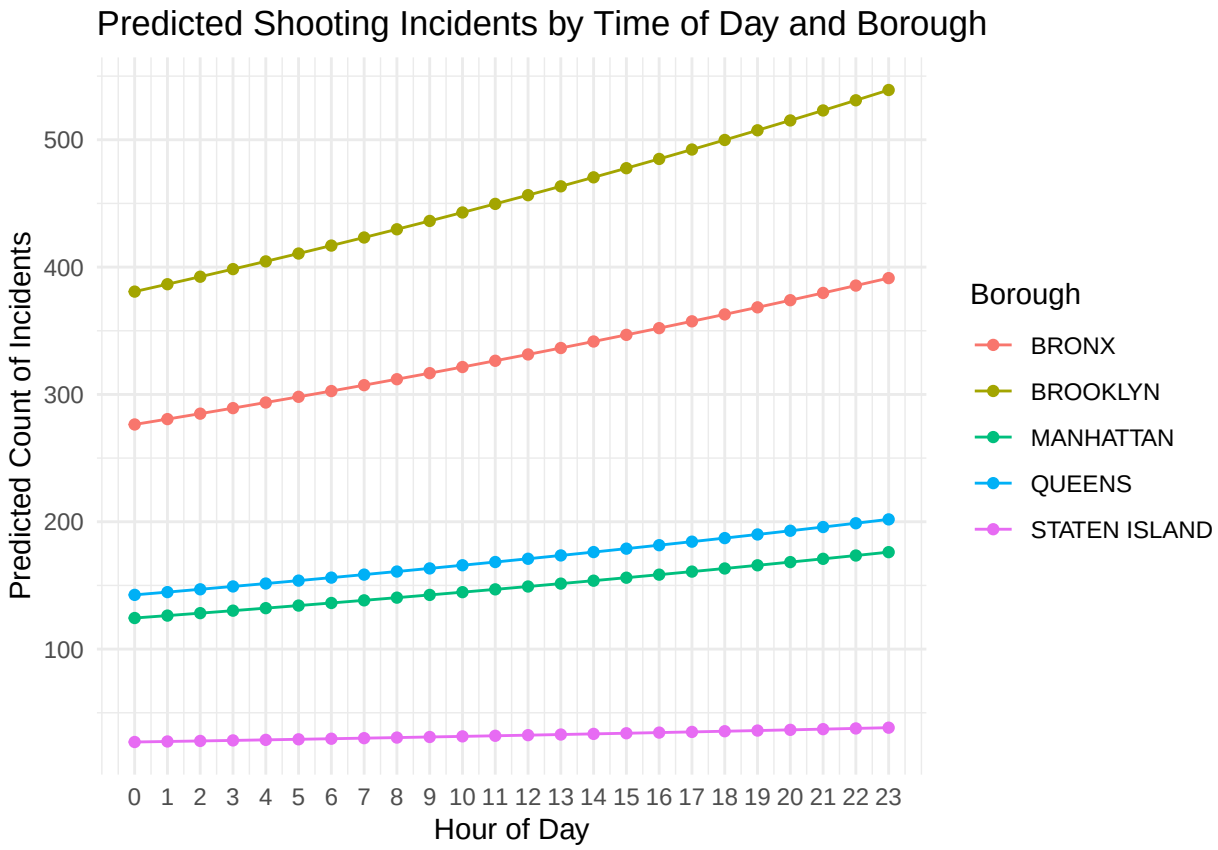
```
## # A tibble: 6 x 3
##    HOUR BORO          Incident_Count
##   <int> <fct>                  <int>
## 1     0 BRONX                    684
## 2     0 BROOKLYN                 827
## 3     0 MANHATTAN                321
## 4     0 QUEENS                   291
## 5     0 STATEN ISLAND             63
## 6     1 BRONX                    609
```

Next we fit a Poisson regression model using glm function

```r
model <- glm(Incident_Count ~ HOUR + BORO, data = incident_counts, family = poisson(link = "log"))

# Summary of the model
summary(model)
```

```
##
## Call:
## glm(formula = Incident_Count ~ HOUR + BORO, family = poisson(link = "log"),
##     data = incident_counts)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        5.621906   0.015521  362.21   <2e-16 ***
## HOUR               0.015118   0.000877   17.24   <2e-16 ***
## BOROBROOKLYN       0.320250   0.014746   21.72   <2e-16 ***
## BOROMANHATTAN     -0.798410   0.020148  -39.63   <2e-16 ***
## BOROQUEENS        -0.662013   0.019242  -34.41   <2e-16 ***
## BOROSTATEN ISLAND -2.325138   0.037612  -61.82   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 24161  on 119  degrees of freedom
## Residual deviance: 11184  on 114  degrees of freedom
## AIC: 11997
##
## Number of Fisher Scoring iterations: 5
```

```r
# Create a new data frame for predictions
hours <- 0:23
boroughs <- levels(incident_counts$BORO)
prediction_data <- expand.grid(HOUR = hours, BORO = boroughs)

# Generate predictions from the model
prediction_data$Incident_Count_Pred <- predict(model, newdata = prediction_data, type = "response")

# Plotting the predictions
ggplot(prediction_data, aes(x = HOUR, y = Incident_Count_Pred, color = BORO)) +
  geom_line() +
  geom_point() +
  theme_minimal() +
  labs(title = "Predicted Shooting Incidents by Time of Day and Borough",
       x = "Hour of Day",
       y = "Predicted Count of Incidents",
       color = "Borough") +
  scale_x_continuous(breaks = 0:23)
```



Additional Questions Raised

- Is there a relation between the time of year and the number of shootings?
- How do victim and perpetrator demographics influence shooting incidents?
- Is there a relationship between socio-economical characteristics of the borough and the number of shooting incidents

Conclusion

This report has provided an initial analysis of the NYPD shooting incident data. We have uncovered patterns that suggest the time of day and borough significantly affect the number of shooting incidents. Additionally, our predictive model offers a framework for anticipating the count of incidents based on these factors. However, the explanation of complexities of shooting incidents, might be influenced by other factors such as demographics and socio-economic conditions, warrant further investigation.

Consideration of Bias

Data Collection Bias: The potential for underreporting or misclassification 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 choice of a Poisson regression model is based on the nature of the data but may not capture all nuances. Future analyses could explore alternative modeling approaches to better fit the data and reduce potential model bias.