Week3 Assignment

2024-04-19

NYPD Shootings 2006 - 2022 Analysis

Data Description

This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity. Please refer to the attached data footnotes for additional information about this data set.

Goals

We will analyze various aspects of the shooting data and try to extract meaningful insights from it. Some of the insights we are seeking relate to the total shooting incident count per borough and total murder incident count per borough. We would also like to model the shootings to be able to predict future shootings in each borough.

Data Import

Let us begin with loading the downloaded csv data into a tibble

```
library(tidyverse)
nypd_data <- read_csv("NYPD_Shooting_Incident_Data__Historic_.csv")
```

Let's take a peek at the data imported

```
head(nypd_data)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
## 1    228798151 05/27/2021 21:30      QUEENS   <NA>                   105
## 2    137471050 06/27/2014 17:40      BRONX    <NA>                    40
## 3    147998800 11/21/2015 03:56      QUEENS   <NA>                   108
## 4    146837977 10/09/2015 18:30      BRONX    <NA>                    44
## 5     58921844 02/19/2009 22:58      BRONX    <NA>                    47
## 6    219559682 10/21/2020 21:36      BROOKLYN <NA>                    81
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

Data wrangling and cleanup

It looks like the data is mostly tidy with each observation as a row, and each variable as a column. We will convert the date column to a date object and remove the columns which are not needed for this analysis such as latitude and longitude data

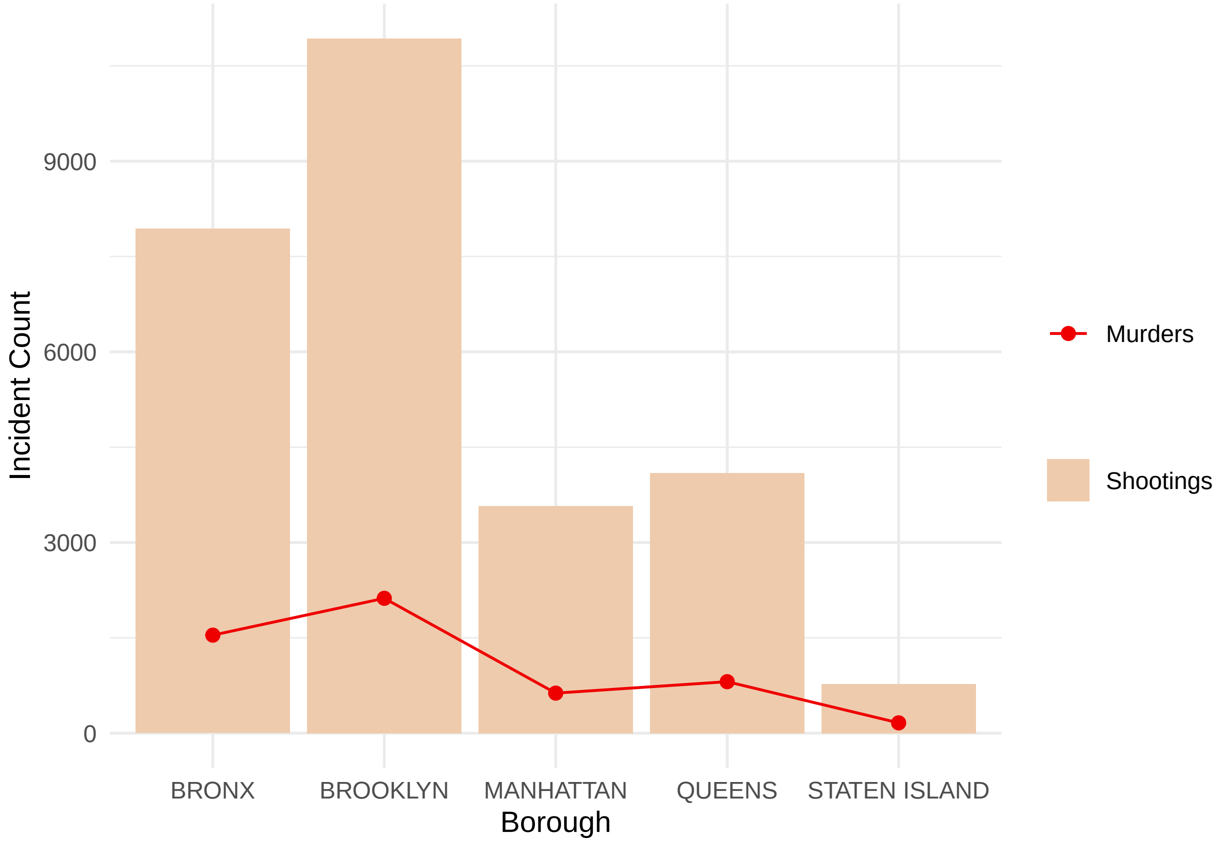
```
nypd_data <- nypd_data %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  select(-X_COORD_CD,-Y_COORD_CD,-Latitude,-Longitude,-Lon_Lat,-JURISDICTION_CODE,-PRECINCT)
head(nypd_data)
```

```
## # A tibble: 6 x 14
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO   LOC_OF_OCCUR_DESC LOC_CLASSFCTN_DESC
##          <dbl> <date>     <time>     <chr>  <chr>             <chr>
## 1    228798151 2021-05-27 21:30      QUEENS <NA>              <NA>
## 2    137471050 2014-06-27 17:40      BRONX  <NA>              <NA>
## 3    147998800 2015-11-21 03:56      QUEENS <NA>              <NA>
## 4    146837977 2015-10-09 18:30      BRONX  <NA>              <NA>
## 5     58921844 2009-02-19 22:58      BRONX  <NA>              <NA>
## 6    219559682 2020-10-21 21:36      BROOK~ <NA>              <NA>
## # i 8 more variables: LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>,
## #   PERP_AGE_GROUP <chr>, PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>,
## #   VIC_SEX <chr>, VIC_RACE <chr>
```

Analysis

Now that the data has been cleaned up, let us analyze shooting and murders based on Borough locations using a bar chart

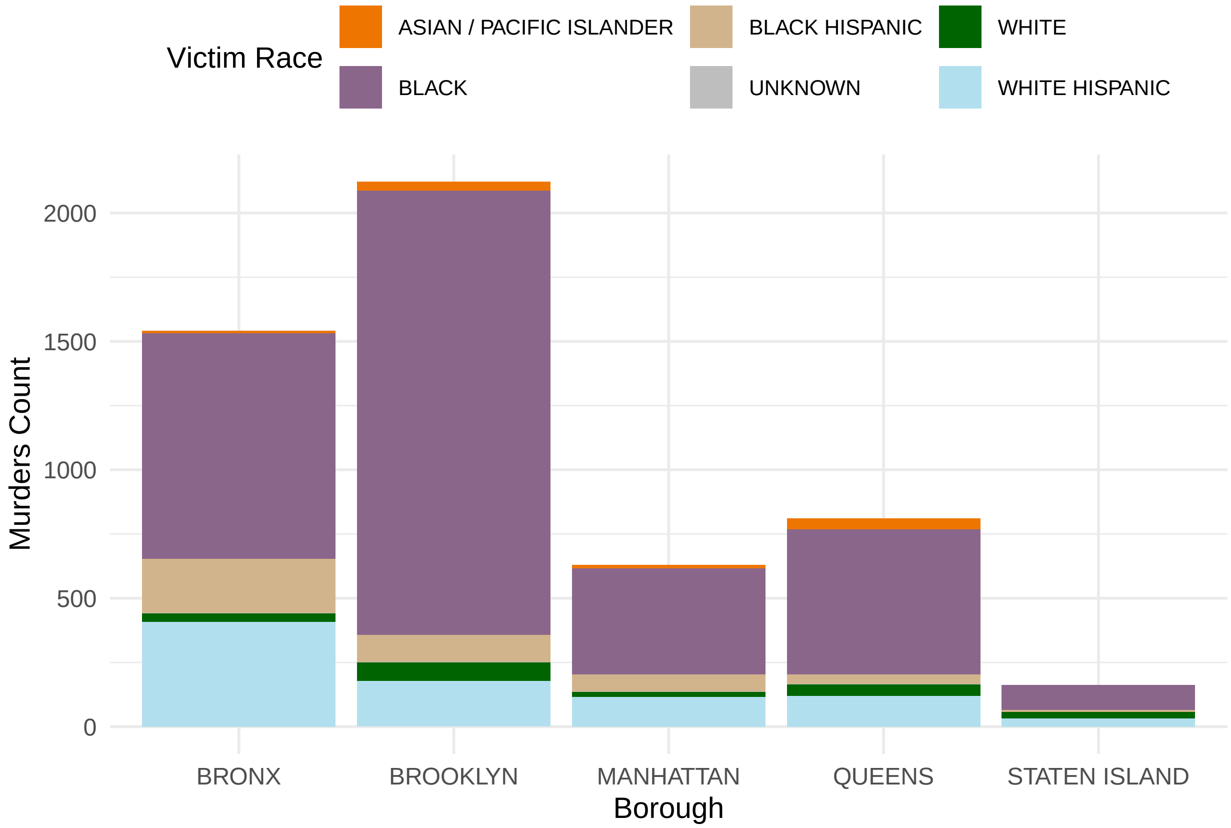
```
nypd_boro <- nypd_data %>%
  group_by(BORO) %>%
  summarize(Shootings = n(), Murders = sum(STATISTICAL_MURDER_FLAG == TRUE))
colors=c("Murders" = "red2", "Shootings" = "peachpuff2")
ggplot(nypd_boro) +
  geom_bar(aes(x = BORO, y = Shootings, fill = "Shootings"), stat = "identity") +
  geom_line(aes(x = BORO, y = Murders, group = 1, color = "Murders")) +
  geom_point(aes(x = BORO, y = Murders, color = "Murders"), size = 2) +
  labs(x = "Borough", y = "Incident Count") +
  scale_colour_manual("", values = colors) +
  scale_fill_manual("",values = colors) +
  theme_minimal()
```



We see that Brooklyn and Bronx have the highest shooting and murder incidents. Also, the murder incidents are fairly proportional to the shooting counts in the corresponding boroughs.

Let us investigate the murders based on victim racial group

```
group_colors <- c("AMERICAN INDIAN/ALASKAN NATIVE" = "darkorchid3",
"ASIAN / PACIFIC ISLANDER" = "darkorange2",
"BLACK" = "plum4",
"BLACK HISPANIC" = "tan",
"UNKNOWN" = "grey",
"WHITE" = "darkgreen",
"WHITE HISPANIC" = "lightblue2")
nypd_filtered <- nypd_data %>%
  filter(STATISTICAL_MURDER_FLAG == TRUE)
ggplot(nypd_filtered) +
  geom_bar(aes(x = BORO, fill = VIC_RACE)) +
  labs(x = "Borough", y = "Murders Count", fill = "Victim Race") +
  scale_fill_manual(values = group_colors) +
  theme_minimal() +
  theme(legend.text = element_text(size = 8),
        legend.position = "top")
```



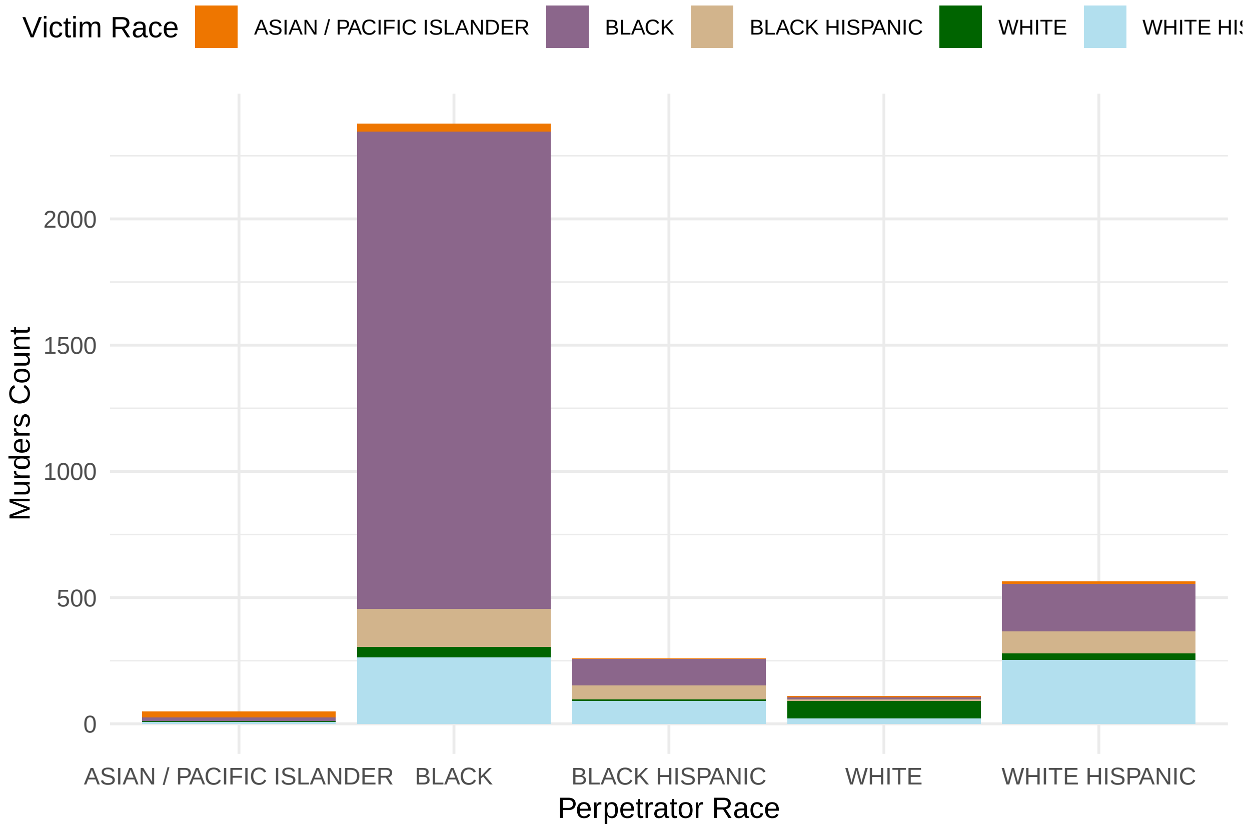
Black victims seem to be the highest murder victims in all boroughs, followed by White Hispanics as a distant second.

Let us analyze the racial group of the perpetrators vs victims in these murders.

We will remove the rows with unknown/null race information for perpetrators and victims.

```r
nypd_perp_filtered <- nypd_filtered %>%
  filter(PERP_RACE != "(null)" & !is.na(PERP_RACE) & PERP_RACE != "UNKNOWN") %>%
  filter(VIC_RACE != "(null)" & !is.na(VIC_RACE) & VIC_RACE != "UNKNOWN")

ggplot(nypd_perp_filtered) +
  geom_bar(aes(x = PERP_RACE, fill = VIC_RACE)) +
  labs(x = "Perpetrator Race", y = "Murders Count", fill = "Victim Race") +
  scale_fill_manual(values = group_colors) +
  theme_minimal() +
  theme(legend.text = element_text(size = 8),
        legend.position = "top")
```



This plot reveals some interesting insights that murders are more likely between the same race than murders between dissimilar races. It also shows that black-on-black murders are disproportionately high compared to all other combinations of perpetrator-victim racial groups.

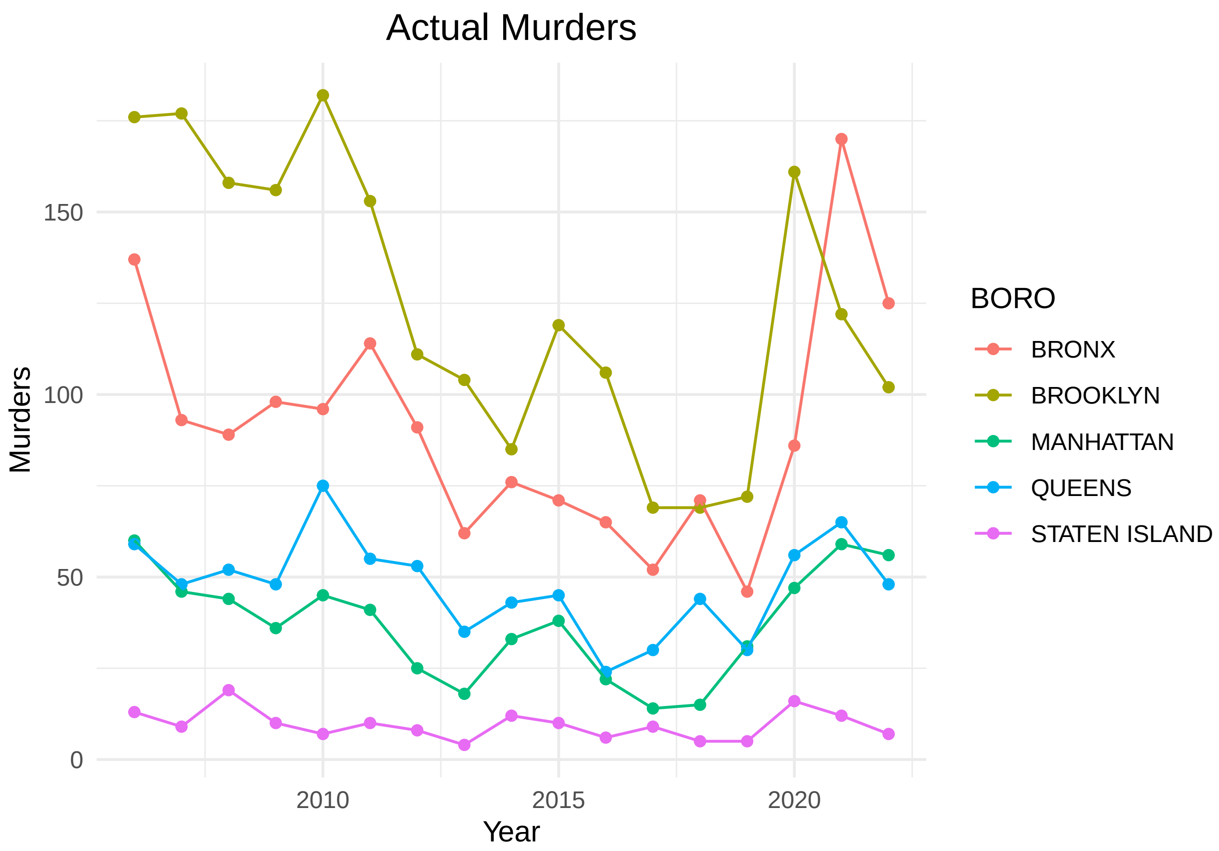
Model Creation

Now we would like to build a model which can predict Murders incidents for each borough based on this historical data. This can be useful to take preemptive measures to forecast and control future trends in violent crimes.

Let us setup a tibble 'nypd_model' to use for building the model and plot it

```
nypd_model <- nypd_data %>%
  filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
  mutate(Year = year(OCCUR_DATE)) %>%
  group_by(Year, BORO) %>%
  summarize(Murders = n())

ggplot(nypd_model) +
  geom_point(aes(x = Year, y = Murders, color = BORO)) +
  geom_line(aes(x = Year, y = Murders, color = BORO)) +
  labs(title = "Actual Murders") +
  theme_minimal() +
  theme(plot.title = element_text(size = 14,hjust = 0.5))
```



Now we can build out the build using this data and predict the murders for the same data

```
model <- lm(Murders ~ Year + BORO, data = nypd_model)
Murders_pred = predict(model, newdata = nypd_model)
results <- cbind(nypd_model, Murders_model = Murders_pred)
head(results)
```

```
## # A tibble: 6 x 4
## # Groups:   Year [2]
##    Year BORO          Murders Murders_model
##   <dbl> <chr>           <int>         <dbl>
## 1  2006 BRONX             137         101.
## 2  2006 BROOKLYN          176         136.
## 3  2006 MANHATTAN          60          47.8
## 4  2006 QUEENS             59          58.4
## 5  2006 STATEN ISLAND      13          20.3
## 6  2007 BRONX              93         100.
```

Model Results

There are negative values in the output of the model 'Murders_model' which is expected. As our dependent variable cannot be negative, let us set negative values to 0. Next, we will plot the actual murders as points and the predicted murders as lines as seen in the graph below.

```r
results <- results %>%
  mutate(Murders_model = ifelse(Murders_model < 0, 0, Murders_model))
ggplot(results) +
  geom_point(aes(x = Year, y = Murders, color = BORO)) +
  geom_line(aes(x = Year, y = Murders_model, color = BORO)) +
  labs(title = "Actual Murders vs Predicted Murders") +
  theme_minimal() +
  theme(plot.title = element_text(size = 14,hjust = 0.5))
```

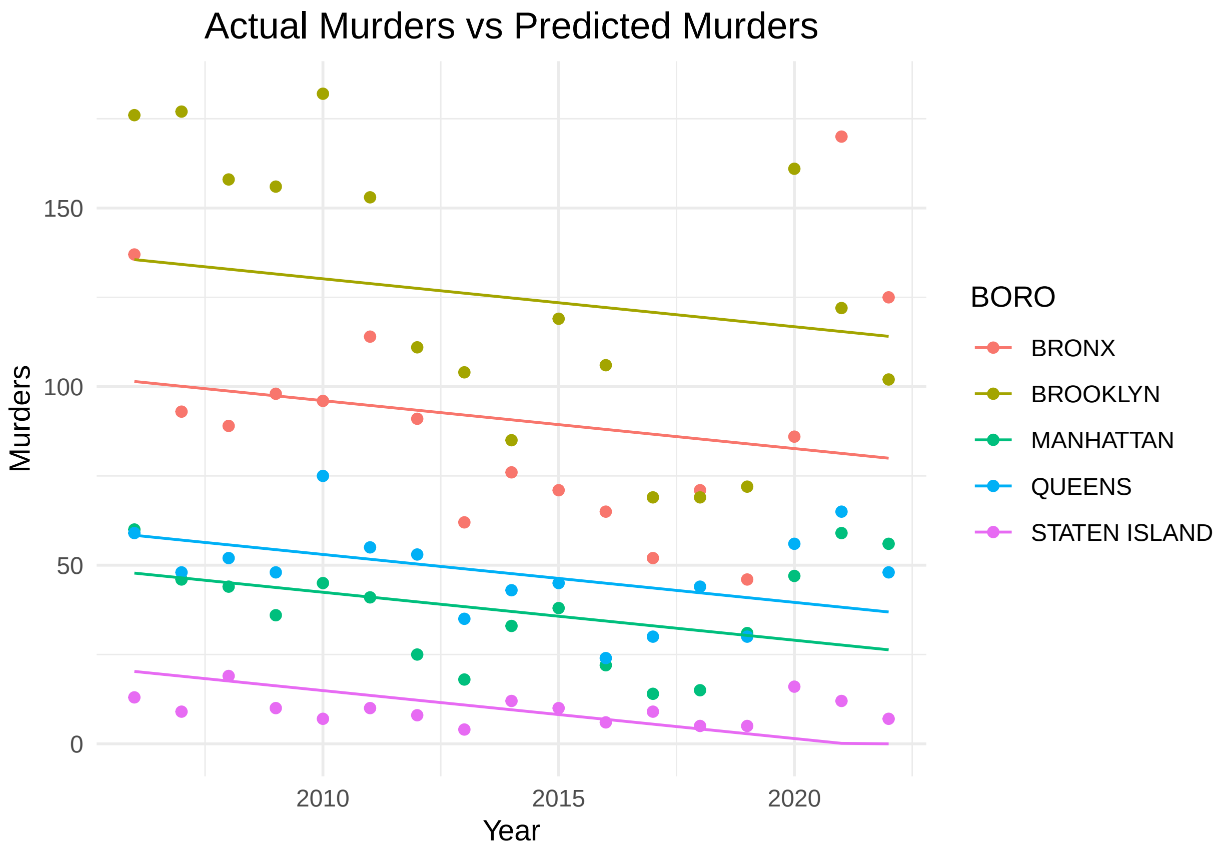


Figure 1: Actual Murders vs Predicted Murders

Conclusion

The model does a decent job at predicting the murders until the year 2020 following which there has been a resurgence of shootings/murders. This could be due to the extraordinary events which occurred during 2020 and after. Overall, the murders were trending downwards until 2020. The model accuracy can be improved by using a non-linear regression or including other independent variables which are not a part of the dataset used in this analysis.

Potential Bias

Sampling Bias:

Sampling Bias occurs when certain groups or individuals in the population are more likely to be included or excluded in the sample. In this analysis, it could arise if only certain shooting incidents are reported among the total shootings

Measurement Bias:

There is always a possibiity of Measurement Bias due to inaccuracies or inconsistencies in the measurement process, leading to mis-classification or misrepresentation of data.

Personal Bias:

Crime incidents are often attributed to racial and demographic factors in social media, few news outlets and other sources. This could lead to some bias in the analysis if not properly addressed by the analyst.