NYPD Historical Shooting Analysis

Johnny Capra

2024-10-04

Introduction

In this analysis, I used the historical NYPD shootings dataset from Data.gov to examine murder trends across New York City's five boroughs. The goal is to determine which boroughs have experienced the highest number of murders and to assess whether murder rates have been increasing or decreasing over time.

Preparing the data

To prepare the data for analysis I first loaded the data. I then cleaned the data by selecting the proper columns and formatting them as you will see in the code chunks below.

```
url_nypd <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

nypd_df <- read_csv(url_nypd, show_col_types = FALSE)
```

```
# Clean DF drop na select by col for boroughs and murder values
boro_murder_df <- nypd_df %>%
  select(STATISTICAL_MURDER_FLAG, BORO) %>%
  drop_na(STATISTICAL_MURDER_FLAG, BORO)

# Filter count only true murders group by borough
boro_murders <- boro_murder_df %>%
  filter(STATISTICAL_MURDER_FLAG == 1)%>%
  group_by(BORO) %>%
  summarise(total_murders = n())
```

```
# Clean DF drop na and changing date to proper date, selecting date, borough and murder columns
murder_historic_df <- nypd_df %>%
  mutate(occur_date = mdy(OCCUR_DATE)) %>%
  select(occur_date, BORO, STATISTICAL_MURDER_FLAG) %>%
  drop_na(occur_date, BORO, STATISTICAL_MURDER_FLAG)

# Filter murders to true and count, change date to be by year and summarize.
# Dropped groups to keep from messing up future aggregation.
murder_historic_df_clean <- murder_historic_df %>%
  filter(STATISTICAL_MURDER_FLAG == 1) %>%
  mutate(year = year(occur_date)) %>%
  group_by(BORO, year) %>%
  summarise(total_murders = n(), .groups = 'drop')
```

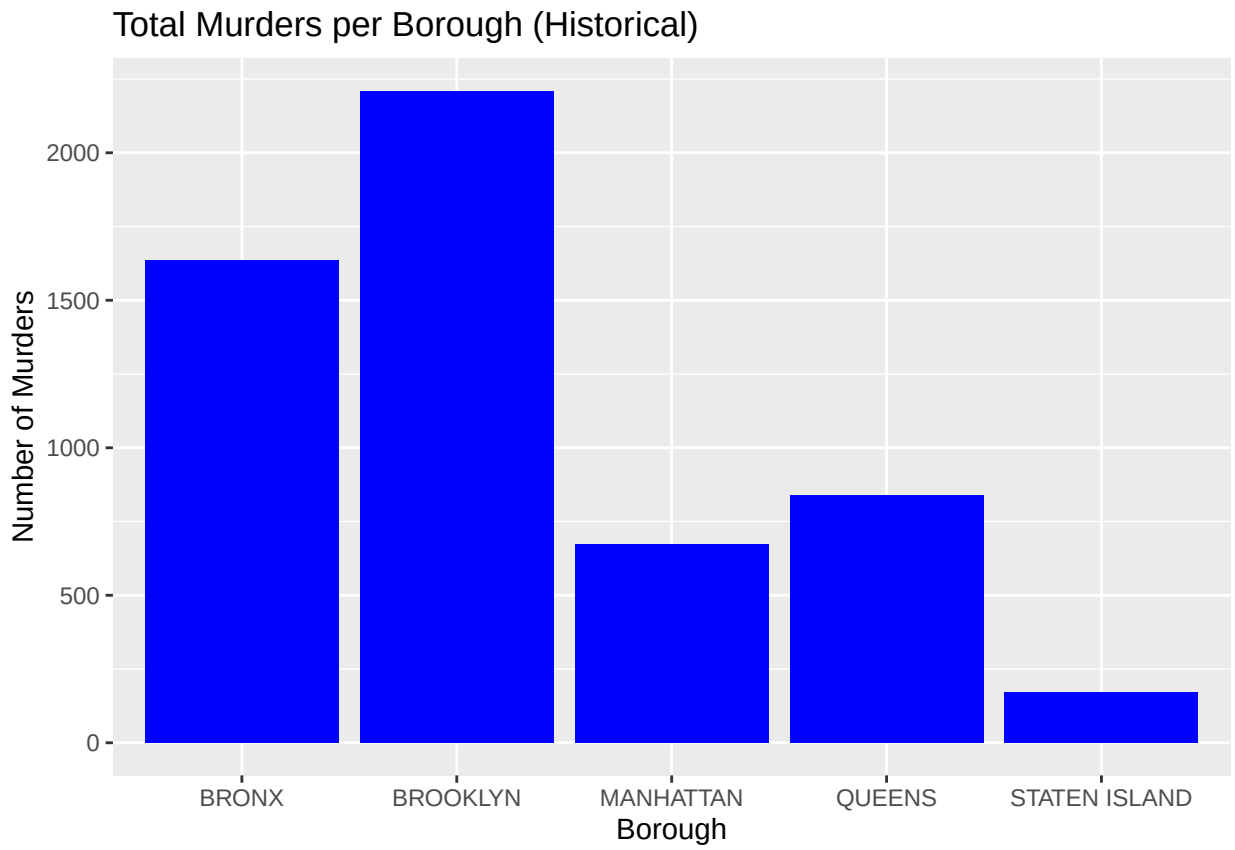
Analysis

Plot Total Murders by Borough

This Bar graph visualizes total historic murders in each borough.

```
#Bar graph of total murders in each borough
murder_plot <- ggplot(boro_murders, aes(x = BORO, y = total_murders)) +
  geom_bar(stat = 'identity', fill = 'blue' ) +
  labs(title = "Total Murders per Borough (Historical)",
       x = "Borough",
       y = "Number of Murders")

murder_plot
```



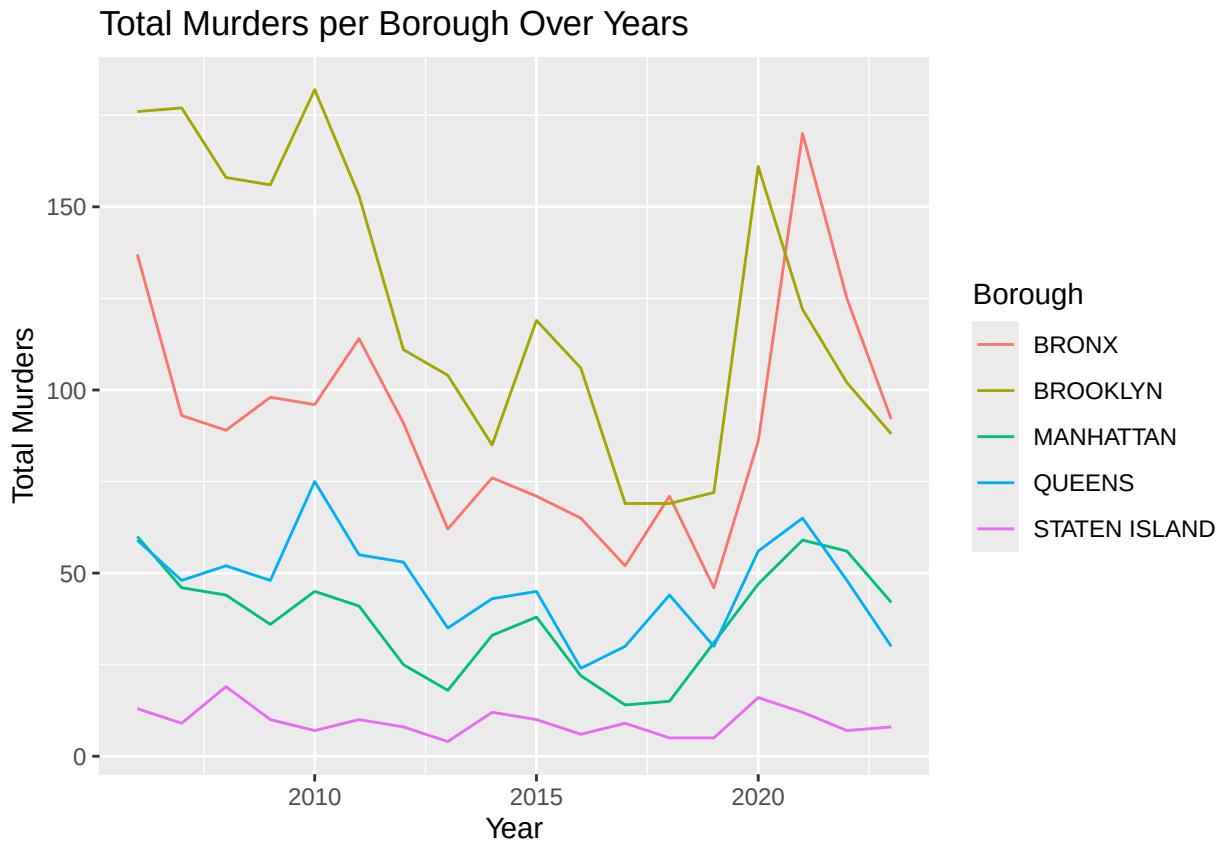
In the plot above we can see that Brooklyn has the highest amount of murders over the years, almost 2500, with the Bronx being the runner up with over 1500. Queens and Manhattan are pretty close in number being off by just a couple hundred of each other and Staten Island the lowest with under 250 total murders. This plot shows that historically speaking Brooklyn has had the most murders out of all the burrows.

Plot Total yearly murders in each borough

This line graph visualizes total murders per year in each borough.

```r
# Plot totals murders of each borough per year over time
murder_over_time_plot <- ggplot(murder_historic_df_clean, aes(x = year, y = total_murders, color = BORO
  geom_line() +
  labs(title = "Total Murders per Borough Over Years",
       x = "Year",
       y = "Total Murders",
       color = "Borough")

murder_over_time_plot
```



The line graph above shows the total historic murders for each year of the data set. We can clearly see that Brooklyn and the Bronx have had the most murder cases over time. Its also interesting to notice the spike around 2020 in each of the boroughs, could this possible have a relationship with covid? What could have caused the spike in 2010? It would be interesting to find out. Overall it looks like historically murders each year are dropping in number for each of the boroughs.

Linear Regression Model

I fit a linear regression model to determine if murders were rising or falling between each borough over the years.

```r
# Clean DF so that borough is factor and year is number value
murder_historic_df_clean <- murder_historic_df_clean %>%
  mutate(BORO = as.factor(BORO), year = as.numeric(year))
```

```r
# Create Model
murder_model <- lm(total_murders ~ year + BORO, data = murder_historic_df_clean)


# Predictions
murder_historic_df_clean <- murder_historic_df_clean %>%
  mutate(predicted_murders = predict(murder_model))
```

Explaining the Model

I used the year and borough of this model as predictors to try and explain the variation in number of murders. The summary shows these results below:

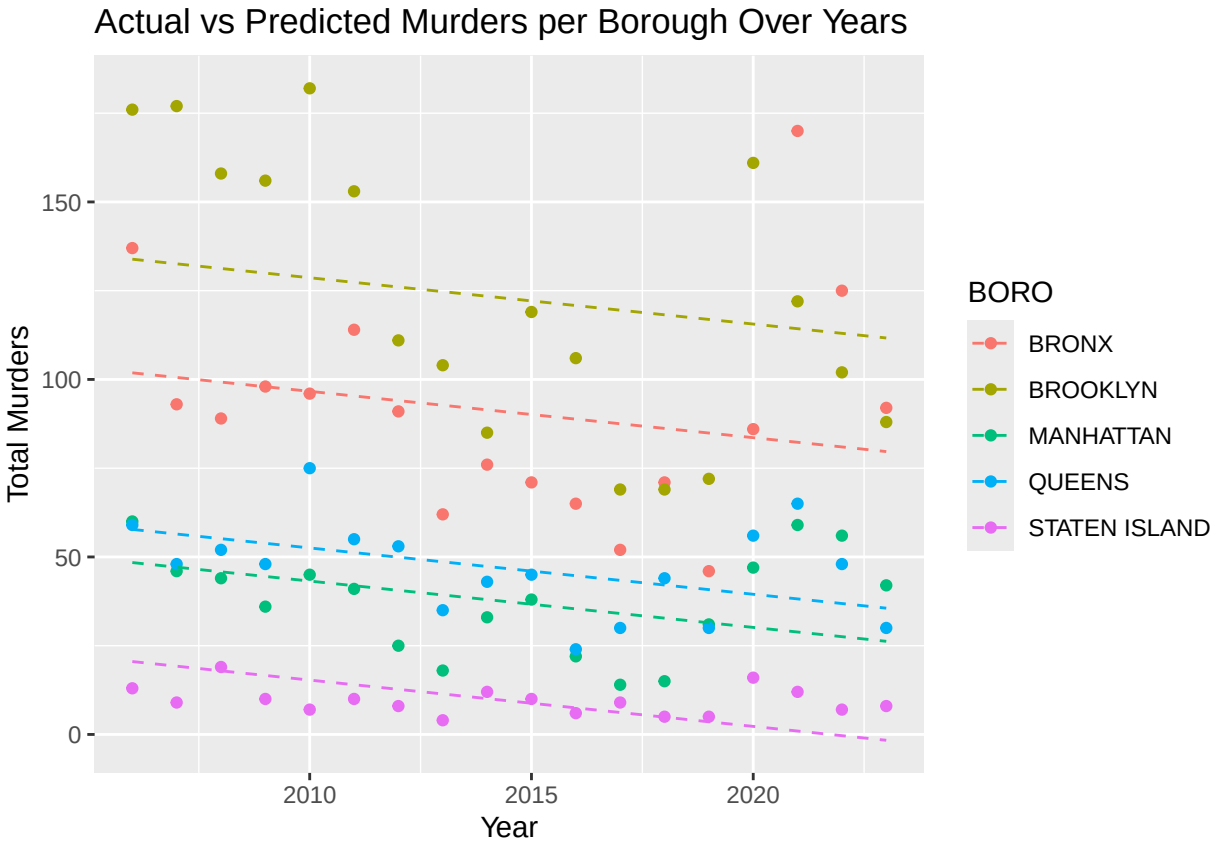
1. Year: There is a negative coefficient for year at -1.3049, indicating, on average, a decrease of 1.3 murders per year.
2. Borough Coefficients: The Bronx was used as a baseline and compares the coefficients for the different boroughs murder counts.
3. Brooklyn: +32 murders compared to the baseline.
4. Manhattan: -53.44 murders compared to the baseline.
5. Queens: -44.11 murders compared to the baseline.
6. Staten Island: -81.33 murders compared to the baseline.
7. R-squared: The r-squared value for this model is 77% meaning year and borough are statistically significant for predicting murder rates

Plot linear regression model

This plot shows the linear regression of each borough of actual murders vs predicted.

```r
# Plot actual vs predicted murders
model_plot <- ggplot(murder_historic_df_clean, aes(x = year, y = total_murders, color = BORO)) +
  geom_point() +
  geom_line(aes(y = predicted_murders), linetype = "dashed") +
  labs(title = "Actual vs Predicted Murders per Borough Over Years",
       x = "Year", y = "Total Murders")

model_plot
```



Conclusion

Historically, we can see distinguishable differences in murders across the boroughs of New York City. Demographics, culture, economic conditions and population density may play a vital role in the differences in violence through out each borough.

The results of the linear regression show a statistically significant decline in approximately 1.3 murders through each borrow over the years compared to the Bronx. The decline is most noticeable in Staten Island and Manhattan but still decreasing in the other boroughs. It makes me wonder if this could be due to gentrification or better law enforcement. The decline offers hope for a less violent future in New York City.

Areas for improvement in this analyses would be to focus on a larger scope of violent crime. This analysis is primarily focused on reported murders not capturing whole picture of what could be causing them. Another limitation is that the linear regression does not account for education levels, available police, or poverty rates. I would like to incorporate these findings in future analysis.

Identifying Bias

There are more than a few sources of bias in this analysis. First, demographic, the analysis doesn't account for income, education or employment rates. Secondly, geographical, this analysis does not take into consideration of population density and the model may be treating all boroughs as comparable units. Thirdly, selection bias, under reporting or incorrect classification of crimes may skew results. Lastly, time bias, the natural flow of the world may not be as linear as the model shows. We see a steady decline in murders that may no be true in a fast paced and changing society.