NYPD Shooting Incident Data Science Report

Tom Plunkett

2025-03-03

Introduction

This data science project analyzes New York City Police Department Shooting Incident data from the period 2006 through 2023, aiming to understand how each borough's number of reported shooting incidents have changed over time on an annual basis.The analysis focuses on he Boroughs of New York City: Brooklyn, Queens, The Bronx, Manhattan, and Staten Island.

Data Ingestion and Data Cleansing

I use the tidyverse and lubridate libraries. I then import the csv file from New York City's open Data website at the URL below

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)

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

shootings <- read_csv(url)
```

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

I trim down to the data I will need for analysis to make the analysis perform faster. The columns I will need include the incident key, the date of the shooting incident, the borough, and the murder flag. The original data set had a lot of columns that I will not need. I then group the data by date and borough, and finally tally the number of shooting incidents by year.

```
by_boro <- shootings %>%
  mutate(INCIDENT_KEY = as.character(INCIDENT_KEY),
  OCCUR_DATE = mdy(OCCUR_DATE),
  BORO = factor(BORO),
  STATISTICAL_MURDER_FLAG = factor(STATISTICAL_MURDER_FLAG),)%>%
  select(c(INCIDENT_KEY,BORO,OCCUR_DATE,STATISTICAL_MURDER_FLAG))
```

Next I create three more tables which are the boroughs grouped by daily, monthly, and yearly incidents

```
by_boro_yearly <- by_boro %>%
  group_by(BORO, year = floor_date(OCCUR_DATE, 'year'))%>%
  tally(name = "INCIDENT_COUNT")
```

I then take a look at the total number of shooting incidents within each of the Boro’s in the next section.

Annual Shooting Incident Analysis

Grouping shooting incidents in the boroughs by total number of annual shootings.

The outputs of each model summary should provide us with clues about the shootings in the city and each bourough. The p-value of each model will determine if the passage of time is a statistically significant predictor of the number of shooting incidents in each borough. Additionally, we can look at the model summary for the entire set to gain insight into the predictions for any given borough in the city.

```
mod_total <- lm(INCIDENT_COUNT ~ year, data = by_boro_yearly)
summary(mod_total)
```

```
##
## Call:
## lm(formula = INCIDENT_COUNT ~ year, data = by_boro_yearly)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -323.62 -158.75  -64.41  170.81  539.99
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 627.65585  208.12899   3.016  0.00335 **
## year         -0.01909    0.01272  -1.501  0.13693
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 228.7 on 88 degrees of freedom
## Multiple R-squared:  0.02496,    Adjusted R-squared:  0.01388
## F-statistic: 2.253 on 1 and 88 DF,  p-value: 0.1369
```

```r
incidents_lm<-by_boro_yearly %>%
  ungroup()%>%
  mutate(pred = predict(mod_total))

mod_bk <- lm(INCIDENT_COUNT ~ year, data = by_boro_yearly%>%filter(BORO == "BROOKLYN"))
summary(mod_bk)
```

```
##
## Call:
## lm(formula = INCIDENT_COUNT ~ year, data = by_boro_yearly %>%
##     filter(BORO == "BROOKLYN"))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -216.890  -62.386   -8.781   63.188  312.768
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1634.60647  269.46423   6.066 1.64e-05 ***
## year          -0.06179    0.01647  -3.752  0.00174 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 132.4 on 16 degrees of freedom
## Multiple R-squared:  0.4681, Adjusted R-squared:  0.4348
## F-statistic: 14.08 on 1 and 16 DF,  p-value: 0.00174
```

```r
incidents_lm_bk<-by_boro_yearly %>%
  filter(BORO == "BROOKLYN") %>%
  mutate(pred = predict(mod_bk))

mod_queens <- lm(INCIDENT_COUNT ~ year, data = by_boro_yearly%>%filter(BORO == "QUEENS"))
summary(mod_queens)
```

```
##
## Call:
## lm(formula = INCIDENT_COUNT ~ year, data = by_boro_yearly %>%
##     filter(BORO == "QUEENS"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -81.846 -37.960  -5.352  38.506  90.858
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 440.683892 114.189077   3.859  0.00139 **
## year         -0.012515   0.006978  -1.793  0.09183 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.1 on 16 degrees of freedom
## Multiple R-squared:  0.1674, Adjusted R-squared:  0.1153
## F-statistic: 3.216 on 1 and 16 DF,  p-value: 0.09183
```

```r
incidents_lm_queens<-by_boro_yearly %>%
  filter(BORO == "QUEENS") %>%
  mutate(pred = predict(mod_queens))

mod_bronx <- lm(INCIDENT_COUNT ~ year, data = by_boro_yearly%>%filter(BORO == "BRONX"))
summary(mod_bronx)
```

```
##
## Call:
## lm(formula = INCIDENT_COUNT ~ year, data = by_boro_yearly %>%
##     filter(BORO == "BRONX"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -172.50  -90.57   23.54   53.23  272.99
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 720.83850  234.80267   3.070  0.00733 **
## year         -0.01572    0.01435  -1.096  0.28950
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 115.4 on 16 degrees of freedom
## Multiple R-squared:  0.06978,    Adjusted R-squared:  0.01164
## F-statistic:   1.2 on 1 and 16 DF,  p-value: 0.2895
```

```r
incidents_lm_bronx<-by_boro_yearly %>%
  filter(BORO == "BRONX") %>%
  mutate(pred = predict(mod_bronx))

mod_mnht <- lm(INCIDENT_COUNT ~ year, data = by_boro_yearly%>%filter(BORO == "MANHATTAN"))
summary(mod_mnht)
```

```
##
## Call:
## lm(formula = INCIDENT_COUNT ~ year, data = by_boro_yearly %>%
##     filter(BORO == "MANHATTAN"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -101.51  -55.09  -13.87   47.20  138.62
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 240.600184 142.866301   1.684    0.112
## year         -0.001944   0.008731  -0.223    0.827
##
## Residual standard error: 70.19 on 16 degrees of freedom
## Multiple R-squared:  0.00309,    Adjusted R-squared:  -0.05922
## F-statistic: 0.04959 on 1 and 16 DF,  p-value: 0.8266
```

```
incidents_lm_mnht<-by_boro_yearly %>%
  filter(BORO == "MANHATTAN") %>%
  mutate(pred = predict(mod_mnht))

mod_si <- lm(INCIDENT_COUNT ~ year, data = by_boro_yearly%>%filter(BORO == "STATEN ISLAND"))
summary(mod_si)
```

```
##
## Call:
## lm(formula = INCIDENT_COUNT ~ year, data = by_boro_yearly %>%
##     filter(BORO == "STATEN ISLAND"))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.568  -2.917   1.028   4.633  15.881
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 101.55019   18.00048   5.642 3.68e-05 ***
## year         -0.00349    0.00110  -3.172  0.00591 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 8.844 on 16 degrees of freedom
## Multiple R-squared:  0.3861, Adjusted R-squared:  0.3477
## F-statistic: 10.06 on 1 and 16 DF,  p-value: 0.005913
```

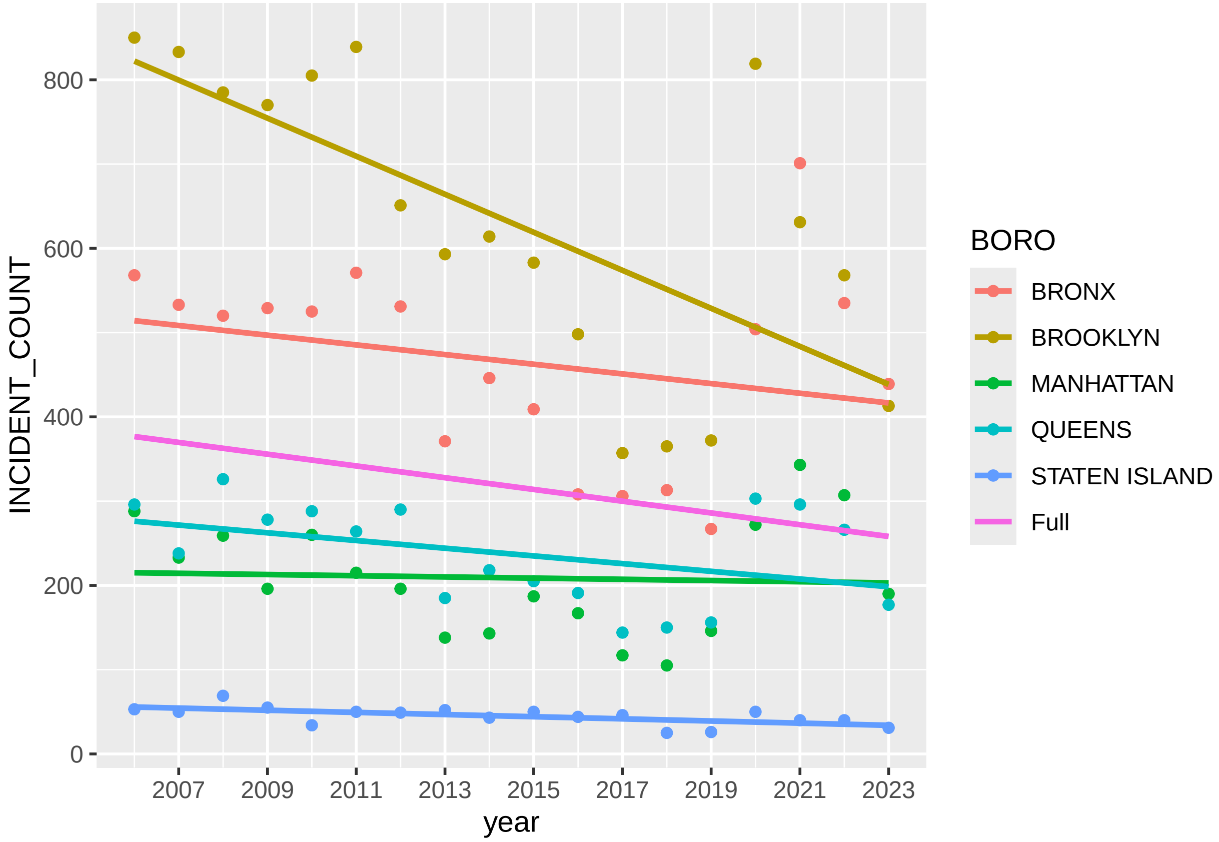
```
incidents_lm_si<-by_boro_yearly %>%
  filter(BORO == "STATEN ISLAND") %>%
  mutate(pred = predict(mod_si))


incidents_lm_total <- incidents_lm_bk %>%
  rbind(incidents_lm_bronx)%>%
  rbind(incidents_lm_queens)%>%
  rbind(incidents_lm_mnht)%>%
  rbind(incidents_lm_si)

incidents_lm <- incidents_lm %>%
  left_join(incidents_lm_total%>%select(lm_boro = "pred", BORO, year), by = c("BORO", "year"))
```

If the p-value is less than .05, we can say that there is a statistically significant linear relationship between the passage of time (increasing occurance date) and the total number of annual shootings in those boroughs. Looking at the p-values, we can infer that the progression of year is a good predictor of number of yearly shooting incidents in Brooklyn. The models for every other borough and any given borough show us that the progression of years might not be a statistically significant predictor of the number of yearly incidents.

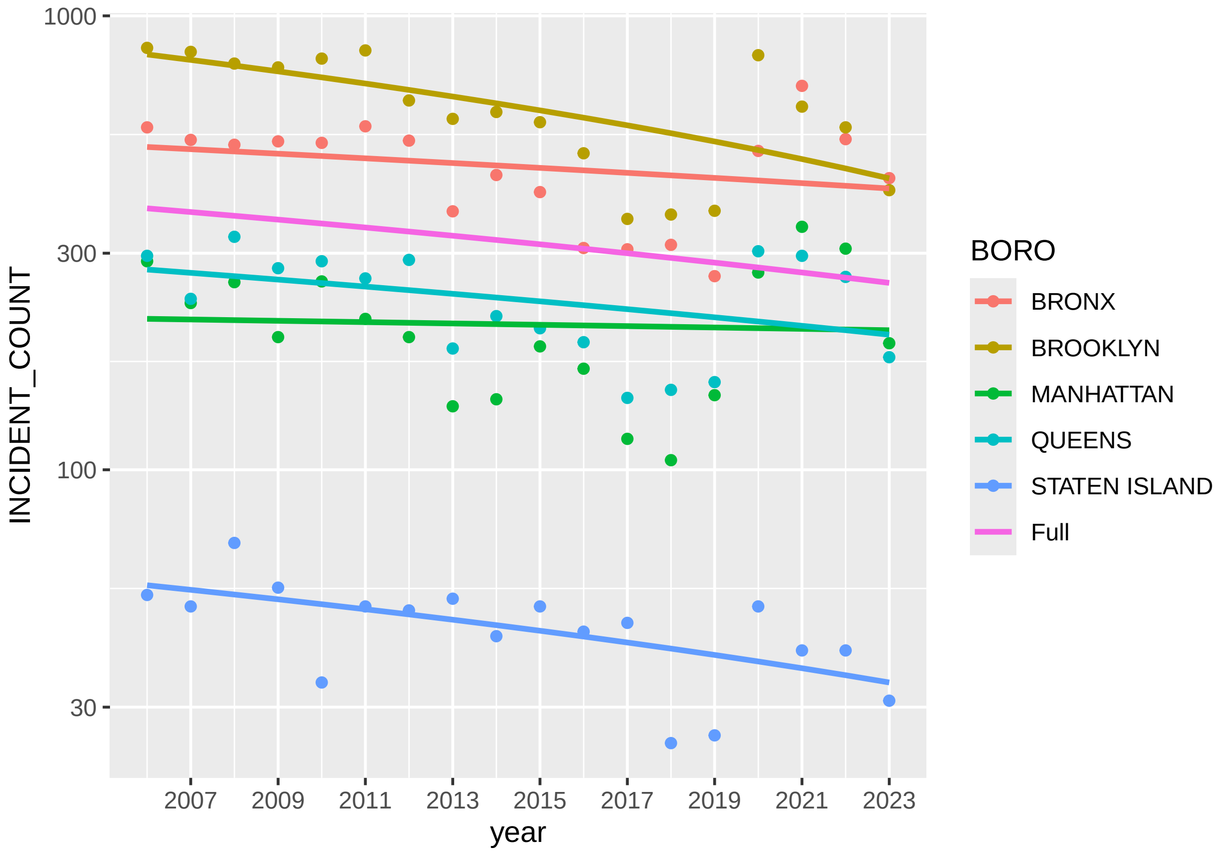
Plotting this results in the following

```
ggplot(incidents_lm, aes(x = year, y = INCIDENT_COUNT)) + geom_point(aes(color = BORO))+scale_x_date(da
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.
## This warning is displayed once every 8 hours.
## Call 'lifecycle::last_lifecycle_warnings()' to see where this warning was
## generated.
```



```
ggplot(incidents_lm, aes(x = year, y = INCIDENT_COUNT)) + geom_point(aes(color = BORO))+scale_x_date(da
```



We see a pretty prominent downward trend in the graphs for all of the boroughs except Manhattan. The most prominent is for Brooklyn (the prominence of which is what generated the p value discussion in the paragraph above).

The second plot transformed the y axis to a log scale. Compared to the initial plot, the log scale provides us a visual comparision of the percentage change of each borough as the years pass. It is much clearer through this plot that the rate of decline for Brooklyn, The Bronx, and Staten island are comparable while Queens and Manhattan show greater stagnation.

Bias

Bias can have impacted my data science project in several different ways.

First, the data set covers reported shooting incidents across one of the largest cities in the world. This data was not collected and collated by one person. Therefore, the different people assembling the data set might have followed different approaches, allowing bias to creep into the original data set.

Second, my data cleansing could have allowed bias to further modify the original data set.

Third, my choice of models and visualizations may also have allowed bias to enter the project.

Conclusion

This analysis reveals that reported shooting incidents in New York City are trending downwards on an annual basis in Brooklyn in a statistically significant manner in the period between 2006 and 2023. The other boroughs, except for Manhattan, also appear to be trending down, although perhaps not enough to be statistically significant.