NYDP Shooting Incident Data

Garey Salinas

2024-11-05

Contents

NY Shooting Incident Data

The NY Shooting Incident data set provides a comprehensive record of every shooting incident reported in New York City from 2006 through the end of the previous calendar year. The New York city data set is a csv file and can be downloaded from https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

Import Libraries

```
library(stringr)
library(readr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v purrr     1.0.1
## v forcats   1.0.0     v tibble    3.2.1
## v ggplot2   3.5.1     v tidyr     1.3.0
## v lubridate 1.9.2
## -- 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 error
```

```
library(knitr)
```

Load Data

I will start by reading in the data from the link provided above.

```
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
incidents <- read_csv(url, show_col_types = FALSE)
```

Data

View data set

```
incidents
```

```
## # A tibble: 28,562 x 21
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      LOC_OF_OCCUR_DESC PRECINCT
##           <dbl> <chr>      <time>     <chr>     <chr>                <dbl>
##  1    244608249 05/05/2022 00:10      MANHATTAN INSIDE                  14
##  2    247542571 07/04/2022 22:20      BRONX     OUTSIDE                 48
##  3     84967535 05/27/2012 19:35      QUEENS    <NA>                   103
##  4    202853370 09/24/2019 21:00      BRONX     <NA>                    42
##  5     27078636 02/25/2007 21:00      BROOKLYN  <NA>                    83
##  6    230311078 07/01/2021 23:07      MANHATTAN <NA>                    23
##  7    229224142 06/07/2021 19:55      QUEENS    <NA>                   113
##  8    231246224 07/22/2021 01:47      BROOKLYN  <NA>                    77
##  9    228559720 05/22/2021 18:39      BRONX     <NA>                    48
## 10    238210279 12/22/2021 23:17      BRONX     <NA>                    49
## # i 28,552 more rows
## # 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>
```

Summary of data set

```
summary(incidents)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:28562       Length:28562      Length:28562
##  1st Qu.: 65439914   Class :character   Class1:hms        Class :character
##  Median : 92711254   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :127405824                      Mode  :numeric
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Mode :logical           Length:28562
##  Class :character   FALSE:23036             Class :character
##  Mode  :character   TRUE :5526              Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
##  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:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

After looking at the data set, I want to tidy the data set by removing the INCIDENT_KEY, X_COORD_CD, Y_COORD_CD, PRECINCT, LOC_OF_OCCUR_DESC, JURISDICTION_CODE, LOC_CLASSFCTN_DESC, LOCATION_DESC, Latitude, Longitude, Lon_Lat.

```
incidents_clean <- incidents %>% select(-c(INCIDENT_KEY,
                                           X_COORD_CD,
                                           Y_COORD_CD,
                                           PRECINCT,
                                           LOC_OF_OCCUR_DESC,
                                           JURISDICTION_CODE,
                                           LOC_CLASSFCTN_DESC,
                                           LOCATION_DESC,
                                           Latitude, Longitude, Lon_Lat))
incidents_clean
```

```
## # A tibble: 28,562 x 10
##    OCCUR_DATE OCCUR_TIME BORO     STATISTICAL_MURDER_F~1 PERP_AGE_GROUP PERP_SEX
##    <chr>      <time>     <chr>    <lgl>                  <chr>          <chr>
##  1 05/05/2022 00:10      MANHATT~ TRUE                   25-44          M
##  2 07/04/2022 22:20      BRONX    TRUE                   (null)         (null)
##  3 05/27/2012 19:35      QUEENS   FALSE                  <NA>           <NA>
##  4 09/24/2019 21:00      BRONX    FALSE                  25-44          M
##  5 02/25/2007 21:00      BROOKLYN FALSE                  25-44          M
##  6 07/01/2021 23:07      MANHATT~ FALSE                  <NA>           <NA>
##  7 06/07/2021 19:55      QUEENS   TRUE                   <NA>           <NA>
##  8 07/22/2021 01:47      BROOKLYN FALSE                  <NA>           <NA>
##  9 05/22/2021 18:39      BRONX    FALSE                  <NA>           <NA>
## 10 12/22/2021 23:17      BRONX    TRUE                   25-44          M
## # i 28,552 more rows
## # i abbreviated name: 1: STATISTICAL_MURDER_FLAG
## # i 4 more variables: PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>
```

Count the number of missing values in each column

```
# Count NA values for each column in incidents_clean
na_counts <- sapply(incidents_clean, function(x) sum(is.na(x)))

# Convert the result to a DataFrame with specified column names
na_summary <- tibble(
  Columns = names(na_counts),
  NA_Count = na_counts
)

kable(na_summary)
```

Columns	NA_Count
OCCUR_DATE	0
OCCUR_TIME	0
BORO	0
STATISTICAL_MURDER_FLAG	0
PERP_AGE_GROUP	9344
PERP_SEX	9310
PERP_RACE	9310
VIC_AGE_GROUP	0
VIC_SEX	0
VIC_RACE	0

Converting `OCCUR_DATE` object into a date object

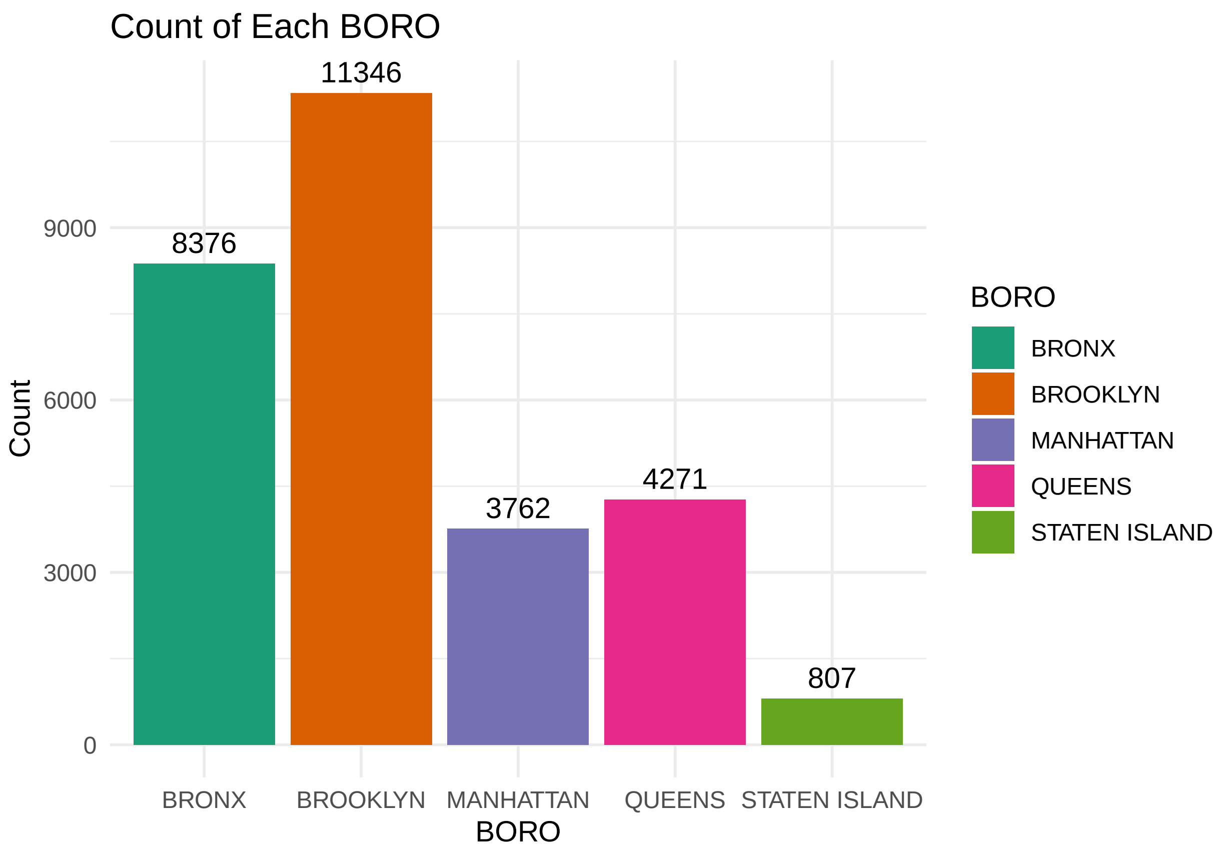
```
incidents_clean$OCCUR_DATE <- mdy(incidents_clean$OCCUR_DATE)
incidents_clean
```

```
## # A tibble: 28,562 x 10
##    OCCUR_DATE OCCUR_TIME BORO     STATISTICAL_MURDER_F~1 PERP_AGE_GROUP PERP_SEX
##    <date>     <time>     <chr>    <lgl>                  <chr>          <chr>
##  1 2022-05-05 00:10      MANHATT~ TRUE                   25-44          M
##  2 2022-07-04 22:20      BRONX    TRUE                   (null)         (null)
##  3 2012-05-27 19:35      QUEENS   FALSE                  <NA>           <NA>
##  4 2019-09-24 21:00      BRONX    FALSE                  25-44          M
##  5 2007-02-25 21:00      BROOKLYN FALSE                  25-44          M
##  6 2021-07-01 23:07      MANHATT~ FALSE                  <NA>           <NA>
##  7 2021-06-07 19:55      QUEENS   TRUE                   <NA>           <NA>
##  8 2021-07-22 01:47      BROOKLYN FALSE                  <NA>           <NA>
##  9 2021-05-22 18:39      BRONX    FALSE                  <NA>           <NA>
## 10 2021-12-22 23:17      BRONX    TRUE                   25-44          M
## # i 28,552 more rows
## # i abbreviated name: 1: STATISTICAL_MURDER_FLAG
## # i 4 more variables: PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>
```

Plots

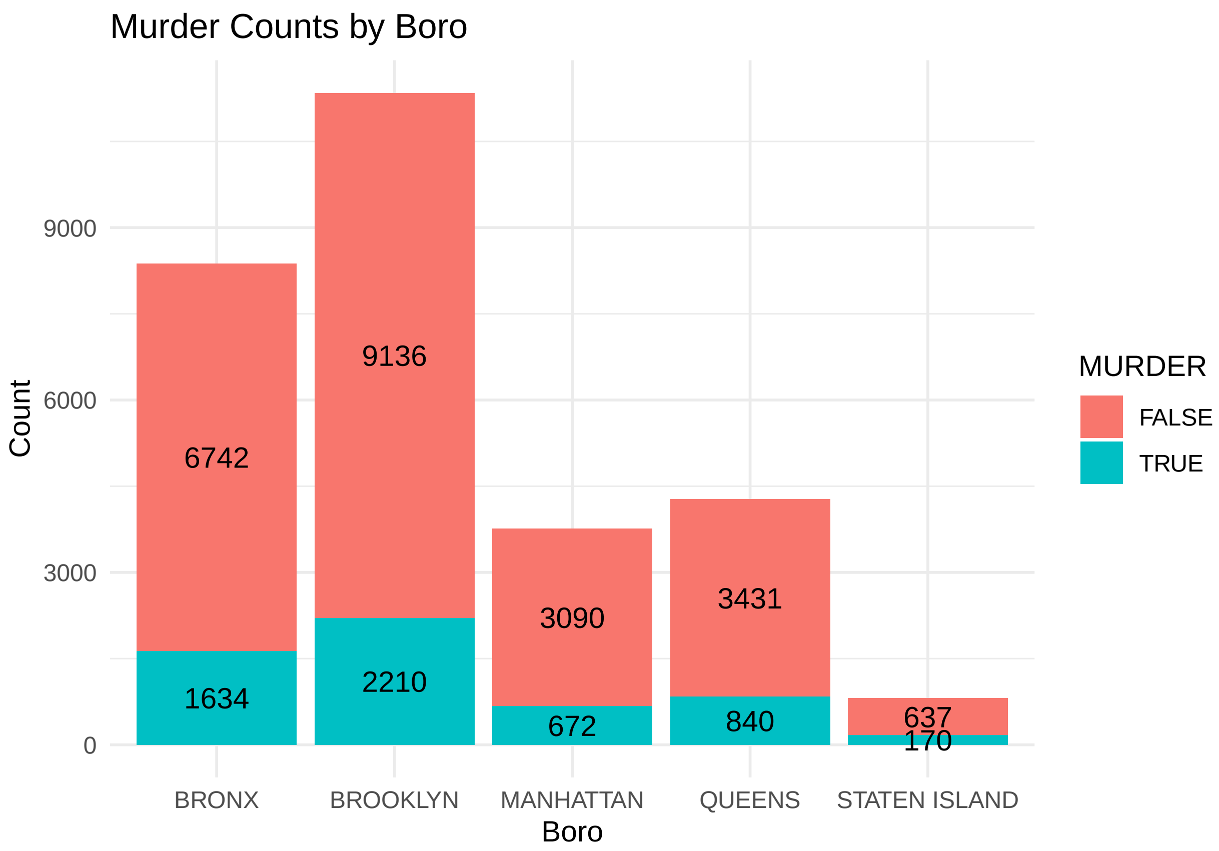
Bar chart of shooting incidents by BORO

```
ggplot(incidents_clean, aes(x = BORO, fill = BORO)) +
  geom_bar() +
  geom_text(stat = "count", aes(label = after_stat(count)), vjust = -0.5) +
  labs(title = "Count of Each BORO", x = "BORO", y = "Count") +
  theme_minimal() +
  scale_fill_brewer(palette = "Dark2")
```



Stacked bar chart of `STATISTICAL_MURDER_FLAG` (TRUE/FALSE) in each `BORO`

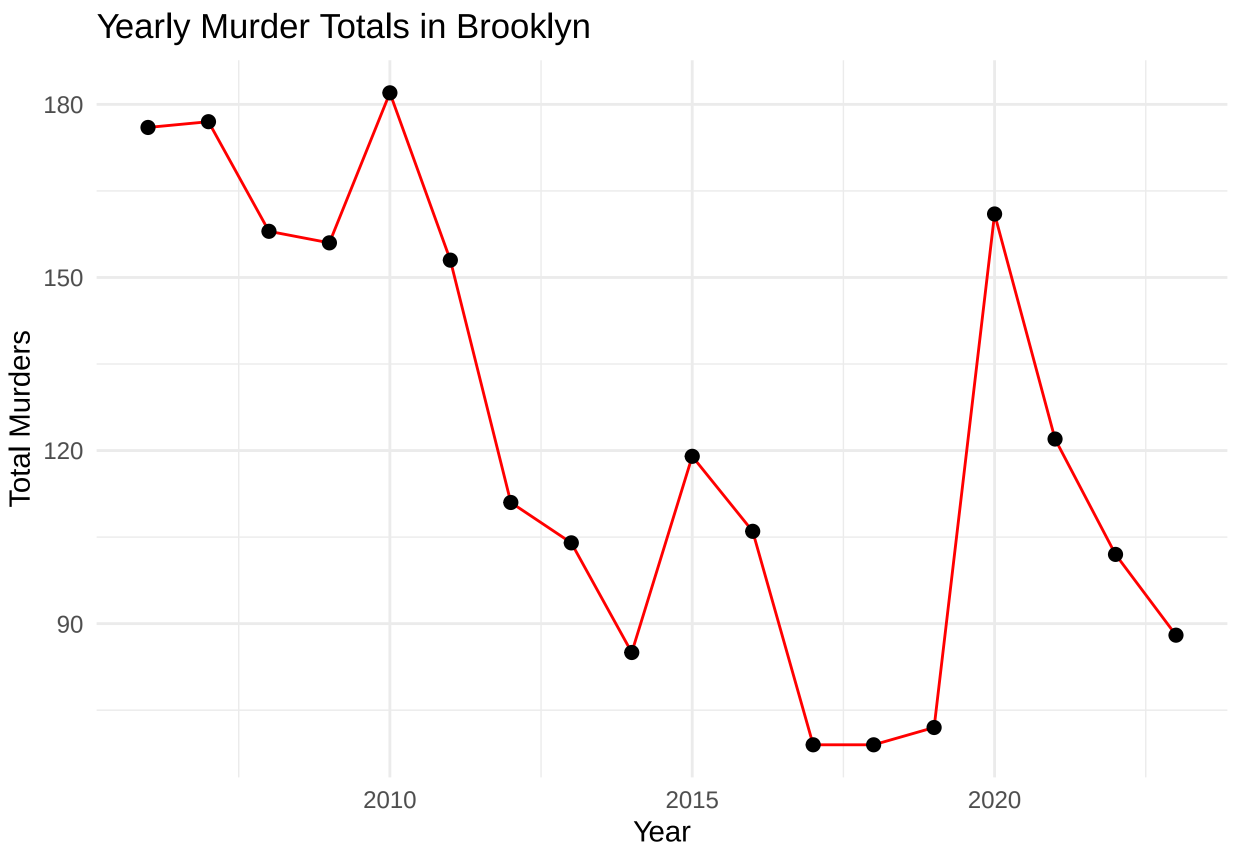
```r
ggplot(incidents_clean, aes(x = BORO, fill = STATISTICAL_MURDER_FLAG)) +
  geom_bar() +
  geom_text(stat = "count", aes(label = after_stat(count)),
            position = position_stack(vjust = 0.5)) +
  labs(title = "Murder Counts by Boro", x = "Boro", y = "Count", fill = "MURDER") +
  theme_minimal()
```



Showing a line plot of Brooklyn Murders.

```
# Filter for Brooklyn murders and aggregate by year
brooklyn_murders_yearly <- incidents_clean %>%
  filter(BORO == "BROOKLYN" & STATISTICAL_MURDER_FLAG == TRUE) %>%  # Filter for Brooklyn murder incide
  mutate(year = year(OCCUR_DATE)) %>%                                # Extract year from date
  group_by(year) %>%                                                 # Group by year
  summarize(total_incidents = n())                                   # Count murders per year

# Plot the line chart
ggplot(brooklyn_murders_yearly, aes(x = year, y = total_incidents)) +
  geom_line(color = "red") +                         # Line plot for yearly totals
  geom_point(color = "black", size = 2) +            # Add points at each year for clarity
  labs(title = "Yearly Murder Totals in Brooklyn",
       x = "Year",
       y = "Total Murders") +
  theme_minimal()
```



Analysis

- Brooklyn has the highest total incidents among all the boroughs.
- The Bronx and Brooklyn have the highest murder counts, with 1,634 and 2,210 murders.
- Manhattan and Queens have moderate murder counts, while Staten Island has the lowest murder count.
- The proportion of murders to non-murders varies between boroughs. For example, while Brooklyn has the highest number of murders, it also has a very high count of non-murders.
- Murder incidents appear to be highly variable across the years.
- The later years in the data set, especially 2020 onward, show relatively lower and more consistent incident counts.
- Incidents peaked from 2008 to 2010, with a decrease in incidents from 2011 to 2019, then spiking again around 2020.

Bias

- The analysis did not account for socioeconomic and demographic factors. Income, employment rates, and population density can influence incident levels.
- Some neighborhoods may experience higher police presence and higher reporting rates, which can skew the data toward these areas.

- Not all crimes could have been reported, especially in under-resourced communities.

Analysis Conclusion

The Project examined the murder trend in Brooklyn. My analysis identified several key findings. Brooklyn has the highest number of murders; it also has a very high count of non-murders. Incidents peaked from 2008 to 2010, with decreased incidents from 2011 to 2019, then spiking again around 2020. High murder rates could indicate a period of economic depression, such as the mortgage crisis and the COVID-19 pandemic. Other factors, such as employment and other socioeconomic factors, contribute to the number of Incidents. Periods of low or no incidents suggest effective policing or community engagement.

Question

- Can we provide a model that predicts the number of murder incidents in Brooklyn to help law enforcement target resources more effectively?

Model

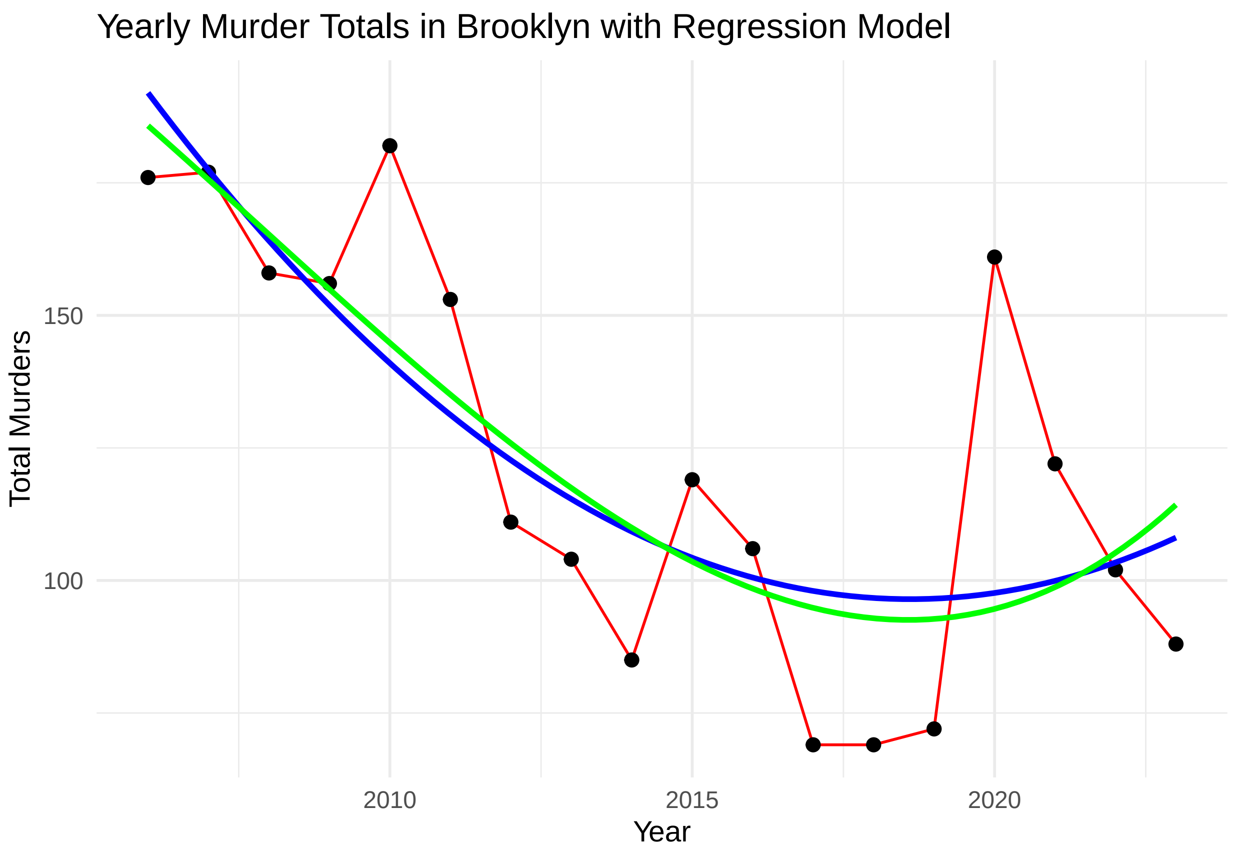
The Brooklyn yearly trend seem to be non-linear so we will model the yearly trends using a polynomial regression of degree 2 or degree 3 and check which one is a better fit.

```r
# Fit a polynomial regression model of degree 2
model_poly2 <- lm(total_incidents ~ poly(year, 2), data = brooklyn_murders_yearly)

# Fit a polynomial regression model of degree 3
model_poly3 <- lm(total_incidents ~ poly(year, 3), data = brooklyn_murders_yearly)

# Original line plot
p <- ggplot(brooklyn_murders_yearly, aes(x = year, y = total_incidents)) +
  geom_line(color = "red") +
  geom_point(color = "black", size = 2) +
  labs(title = "Yearly Murder Totals in Brooklyn with Regression Model",
       x = "Year",
       y = "Total Murders") +
  theme_minimal()

# Add the polynomial regression line (degree 2)
p + geom_smooth(method = "lm", formula = y ~ poly(x, 2), color = "blue", se = FALSE) +
  geom_smooth(method = "lm", formula = y ~ poly(x, 3), color = "green", se = FALSE)
```



Model Summary

```r
# Summary of the quadratic model
summary(model_poly2)
```

```
##
## Call:
## lm(formula = total_incidents ~ poly(year, 2), data = brooklyn_murders_yearly)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.018 -19.053  -3.734  12.401  63.361
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)      122.78       6.29   19.52 4.49e-12 ***
## poly(year, 2)1  -108.63      26.69   -4.07  0.00101 **
## poly(year, 2)2    61.11      26.69    2.29  0.03695 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.69 on 15 degrees of freedom
## Multiple R-squared:  0.5925, Adjusted R-squared:  0.5382
## F-statistic: 10.91 on 2 and 15 DF,  p-value: 0.001191
```

```r
# Summary of the cubic model
summary(model_poly3)
```

```
##
## Call:
## lm(formula = total_incidents ~ poly(year, 3), data = brooklyn_murders_yearly)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.275 -19.273  -5.238  13.472  66.368
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     122.778      6.452  19.030 2.11e-11 ***
## poly(year, 3)1 -108.626     27.373  -3.968   0.0014 **
## poly(year, 3)2   61.109     27.373   2.232   0.0424 *
## poly(year, 3)3   13.895     27.373   0.508   0.6196
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 27.37 on 14 degrees of freedom
## Multiple R-squared:  0.5999, Adjusted R-squared:  0.5141
## F-statistic: 6.996 on 3 and 14 DF,  p-value: 0.004157
```

Based on the similar R-squared, higher RSE, and non-significant cubic term, the quadratic model (degree 2) is a better choice. It provides a similar fit with fewer terms, making it more straightforward and interpretable.

Using AIC and BIC to help choose the best model

```r
AIC(model_poly2, model_poly3)
```

```
##             df      AIC
## model_poly2  4 174.0310
## model_poly3  5 175.7027
```

```r
BIC(model_poly2, model_poly3)
```

```
##             df      BIC
## model_poly2  4 177.5925
## model_poly3  5 180.1546
```

The quadratic model is better based on lower AIC and BIC values, simplicity, and interpretability.

Prediction

```r
# Create a data frame for the year 2025
murder_prediction <- data.frame(year = 2025)

predicted_value_2025 <- predict(model_poly2, newdata = murder_prediction)

print(predicted_value_2025)
```

```
##        1 
## 121.0497
```

Model predicts that 121 murders will occur in Brooklyn in the year 2025.

Session Information

```
sessionInfo()
```

```
## R version 4.1.0 (2021-05-18)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 26100)
## 
## Matrix products: default
## 
## locale:
## [1] LC_COLLATE=English_United States.1252 
## [2] LC_CTYPE=English_United States.1252   
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C                          
## [5] LC_TIME=English_United States.1252    
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] knitr_1.48      lubridate_1.9.2 forcats_1.0.0   dplyr_1.1.2    
##  [5] purrr_1.0.1     tidyr_1.3.0     tibble_3.2.1    ggplot2_3.5.1  
##  [9] tidyverse_2.0.0 readr_2.1.4     stringr_1.5.1  
## 
## loaded via a namespace (and not attached):
##  [1] highr_0.11         RColorBrewer_1.1-3 pillar_1.9.0       compiler_4.1.0    
##  [5] tools_4.1.0        bit_4.0.5          digest_0.6.31      lattice_0.20-44   
##  [9] nlme_3.1-152       timechange_0.2.0   evaluate_1.0.1     lifecycle_1.0.4   
## [13] gtable_0.3.6       mgcv_1.8-35        pkgconfig_2.0.3    rlang_1.1.4       
## [17] Matrix_1.3-3       cli_3.6.1          rstudioapi_0.17.1  curl_5.0.0        
## [21] parallel_4.1.0     yaml_2.3.7         xfun_0.48          fastmap_1.1.1     
## [25] withr_3.0.2        generics_0.1.3     vctrs_0.6.5        hms_1.1.3         
## [29] bit64_4.0.5        grid_4.1.0         tidyselect_1.2.1   glue_1.6.2        
## [33] R6_2.5.1           fansi_1.0.4        vroom_1.6.1        rmarkdown_2.28    
## [37] farver_2.1.1       tzdb_0.3.0         magrittr_2.0.3     splines_4.1.0     
## [41] scales_1.3.0       htmltools_0.5.8.1  colorspace_2.1-0   labeling_0.4.3    
## [45] utf8_1.2.3         stringi_1.7.12     munsell_0.5.1      crayon_1.5.3
```