NYPD Shooting Incident Data Report

2024-09-03

Load Packages

```
library(tidyverse)
library(ggplot2)
library(dplyr)
```

Import Data

The two data sets I will be using for my analysis are “NYPD Shooting Incident Data (Historic)” and “New York City Population by Borough, 1950 - 2040.” Both data sets are provided by the City of New York. Below I will import and load each to see what they contain.They will be called `nypd_main` and `nyc_boro_pop` respectively.

```
nypd_main <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
nyc_boro_pop <- read_csv("https://data.cityofnewyork.us/api/views/xywu-7bv9/rows.csv?accessType=DOWNLOA
```

```
nypd_main
```

```
## # 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>
```

```
nyc_boro_pop
```

```
## # A tibble: 6 x 22
##   ‘Age Group‘      Borough        ‘1950‘ ‘1950 - Boro share of NYC total‘ ‘1960‘
##   <chr>            <chr>           <dbl>                            <dbl>  <dbl>
## 1 Total Population NYC Total     7891957                           100    7.78e6
## 2 Total Population Bronx         1451277                            18.4  1.42e6
## 3 Total Population Brooklyn      2738175                            34.7  2.63e6
## 4 Total Population Manhattan     1960101                            24.8  1.70e6
## 5 Total Population Queens        1550849                            19.6  1.81e6
## 6 Total Population Staten Island  191555                             2.43 2.22e5
## # i 17 more variables: ‘1960 - Boro share of NYC total‘ <dbl>, ‘1970‘ <dbl>,
## #   ‘1970 - Boro share of NYC total‘ <dbl>, ‘1980‘ <dbl>,
## #   ‘1980 - Boro share of NYC total‘ <dbl>, ‘1990‘ <dbl>,
## #   ‘1990 - Boro share of NYC total‘ <dbl>, ‘2000‘ <dbl>,
## #   ‘2000 - Boro share of NYC total‘ <dbl>, ‘2010‘ <dbl>,
## #   ‘2010 - Boro share of NYC total‘ <dbl>, ‘2020‘ <dbl>,
## #   ‘2020 - Boro share of NYC total‘ <dbl>, ‘2030‘ <dbl>, ...
```

Clean/Transform NYPD Shooting Incident Data

The two variables I am interested in for this analysis are the borough and the year in which each shooting incident took place. Below I isolate those two variables by creating a new column `Year` referencing the year value from the `OCCUR_DATE` column. I then omit all other columns aside from `BORO` which I rename `Borough`. I name this data set `nypd_tidy`.

```
nypd_tidy <- nypd_main %>%
select(-c(INCIDENT_KEY, OCCUR_TIME, LOC_OF_OCCUR_DESC, STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP, PERP_SE
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))%>%
  mutate(Year = year(OCCUR_DATE)) %>%
  select(-OCCUR_DATE) %>%
  rename(Borough = BORO) %>%
  arrange(Borough,Year)

nypd_tidy
```

```
## # A tibble: 28,562 x 2
##    Borough  Year
##    <chr>   <dbl>
##  1 BRONX    2006
##  2 BRONX    2006
##  3 BRONX    2006
##  4 BRONX    2006
##  5 BRONX    2006
##  6 BRONX    2006
##  7 BRONX    2006
##  8 BRONX    2006
##  9 BRONX    2006
## 10 BRONX    2006
## # i 28,552 more rows
```

Clean/Transform Borough Population Data

Below I omit all columns other than those providing population data for the years 2000, 2010, and 2020 for each of the five boroughs. I will use these as population estimates for my analysis. I name this data set nyc_boro_pop_tidy.

```
nyc_boro_pop_tidy <- nyc_boro_pop[-c(1), ]%>%
  select(c(Borough, '2000', '2010', '2020'))

nyc_boro_pop_tidy
```

```
## # A tibble: 5 x 4
##   Borough        ‘2000‘  ‘2010‘  ‘2020‘
##   <chr>           <dbl>   <dbl>   <dbl>
## 1 Bronx         1332650 1385108 1446788
## 2 Brooklyn      2465326 2552911 2648452
## 3 Manhattan     1537195 1585873 1638281
## 4 Queens        2229379 2250002 2330295
## 5 Staten Island  443728  468730  487155
```

Combine the Two Data Sets

Below I create a new column `Population` in the `nypd_tidy` data set based on the population estimates above. I use the 2000 population estimate for years 2000-2009, the 2010 population estimate for years 2010-2019, and the 2020 population estimate for years 2020-2023. I call this data set `nypd_w_pop`.

```
nypd_w_pop <- nypd_tidy %>%
 mutate(Population = case_when(
    Borough == "BROOKLYN" & Year >= 2000 & Year <= 2009 ~ 2465326,
    Borough == "QUEENS" & Year >= 2000 & Year <= 2009 ~ 2229379,
    Borough == "BRONX" & Year >= 2000 & Year <= 2009 ~ 1332650,
    Borough == "MANHATTAN" & Year >= 2000 & Year <= 2009 ~ 1537195,
    Borough == "STATEN ISLAND" & Year >= 2000 & Year <= 2009 ~ 443728,
    Borough == "BROOKLYN" & Year >= 2010 & Year <= 2019 ~ 2552911,
    Borough == "QUEENS" & Year >= 2010 & Year <= 2019 ~ 2250002,
    Borough == "BRONX" & Year >= 2010 & Year <= 2019 ~ 1385108,
    Borough == "MANHATTAN" & Year >= 2010 & Year <= 2019 ~ 1585873,
    Borough == "STATEN ISLAND" & Year >= 2010 & Year <= 2019 ~ 468730,
    Borough == "BROOKLYN" & Year >= 2020 ~ 2648452,
    Borough == "QUEENS" & Year >= 2020 ~ 2330295,
    Borough == "BRONX" & Year >= 2020 ~ 1446788,
    Borough == "MANHATTAN" & Year >= 2020 ~ 1638281,
    Borough == "STATEN ISLAND" & Year >= 2020 ~ 487155,
    TRUE ~ NA_real_
  )) %>%
  group_by(Borough, Year)

nypd_w_pop
```

```
## # A tibble: 28,562 x 3
## # Groups:   Borough, Year [90]
##    Borough   Year Population
##    <chr>    <dbl>      <dbl>
##  1 BRONX     2006    1332650
##  2 BRONX     2006    1332650
##  3 BRONX     2006    1332650
##  4 BRONX    2006    1332650
##  5 BRONX    2006    1332650
##  6 BRONX    2006    1332650
##  7 BRONX    2006    1332650
##  8 BRONX    2006    1332650
##  9 BRONX    2006    1332650
## 10 BRONX    2006    1332650
## # i 28,552 more rows
```

Questions and Visualizations

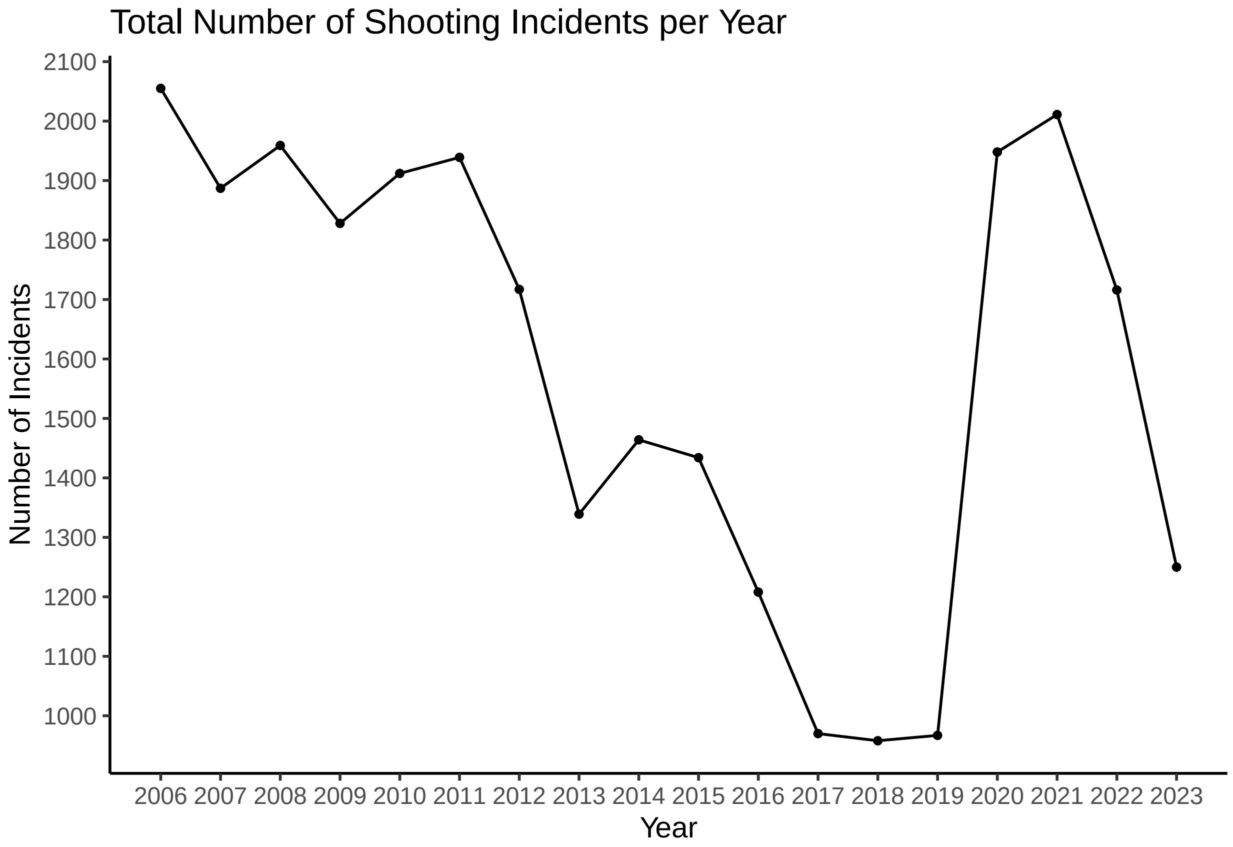
Question 1: What was the total number of shooting incidents each year?

Below I calculate the total number of shooting incidents in each year and create a line graph showing how that figure changed over time.

```
year_totals <- nypd_w_pop %>% group_by(Year) %>%
  summarize(Incidents = n())

year_totals
```

```
## # A tibble: 18 x 2
##     Year Incidents
##    <dbl>     <int>
##  1  2006      2055
##  2  2007      1887
##  3  2008      1959
##  4  2009      1828
##  5  2010      1912
##  6  2011      1939
##  7  2012      1717
##  8  2013      1339
##  9  2014      1464
## 10  2015      1434
## 11  2016      1208
## 12  2017       970
## 13  2018       958
## 14  2019       967
## 15  2020      1948
## 16  2021      2011
## 17  2022      1716
## 18  2023      1250
```

```
ggplot(year_totals, aes(x=Year, y=Incidents)) +
  geom_line(linewidth = .5, stat="identity") +
  geom_point(size = 1) +
  xlab("Year") + ylab("Number of Incidents") +
  scale_x_continuous(breaks = seq(2006, 2023, by = 1)) +
  scale_y_continuous(breaks = seq(0, 2100, by = 100)) +
  ggtitle("Total Number of Shooting Incidents per Year") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  theme_classic()
```



From the above graph we see a downward trend in the total number of shooting incidents for years 2006-2019 until a massive upward swing in 2020. This upward trend appears to peak in 2021 and begin to quickly decrease again through 2023.

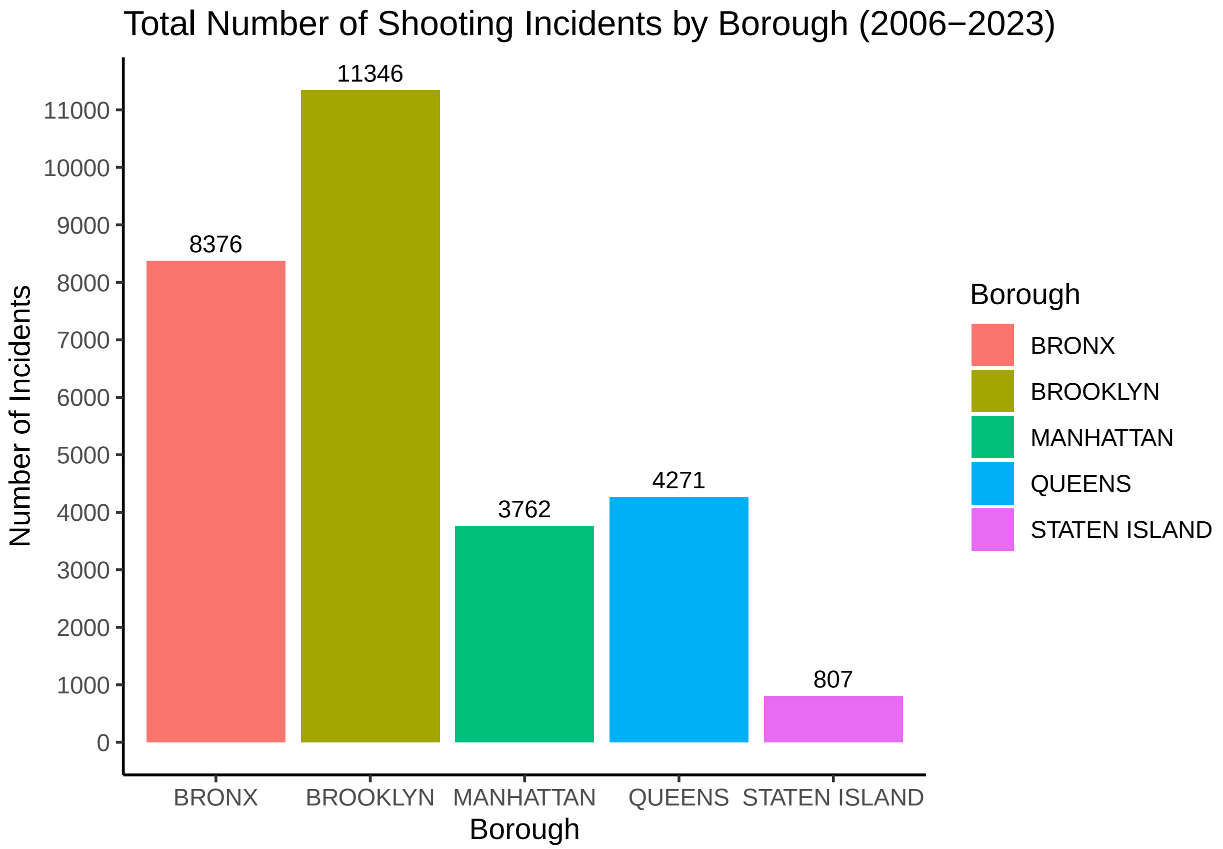
Question 2: What was the total number of shooting incidents in each of the five boroughs between 2006 and 2023?

Below I calculate the total number of shooting incidents in each of the five boroughs between 2006 and 2023 and visualize the figures using a bar chart.

```
boro_totals <- nypd_w_pop %>%
  group_by(Borough) %>%
  summarise(Incidents = n()) %>%
  arrange(desc(Incidents))

boro_totals
```

```
## # A tibble: 5 x 2
##   Borough       Incidents
##   <chr>             <int>
## 1 BROOKLYN          11346
## 2 BRONX              8376
## 3 QUEENS             4271
## 4 MANHATTAN          3762
## 5 STATEN ISLAND       807
```

```
ggplot(boro_totals, aes(x=Borough, y=Incidents, fill=Borough)) +
  geom_bar(stat="identity") +
  xlab("Borough") + ylab("Number of Incidents") +
  ggtitle("Total Number of Shooting Incidents by Borough (2006-2023)") +
  geom_text(aes(label = Incidents), vjust = -0.5, size = 3) +
  theme_classic() +
  scale_y_continuous(breaks = seq(0, 12000, by = 1000))
```

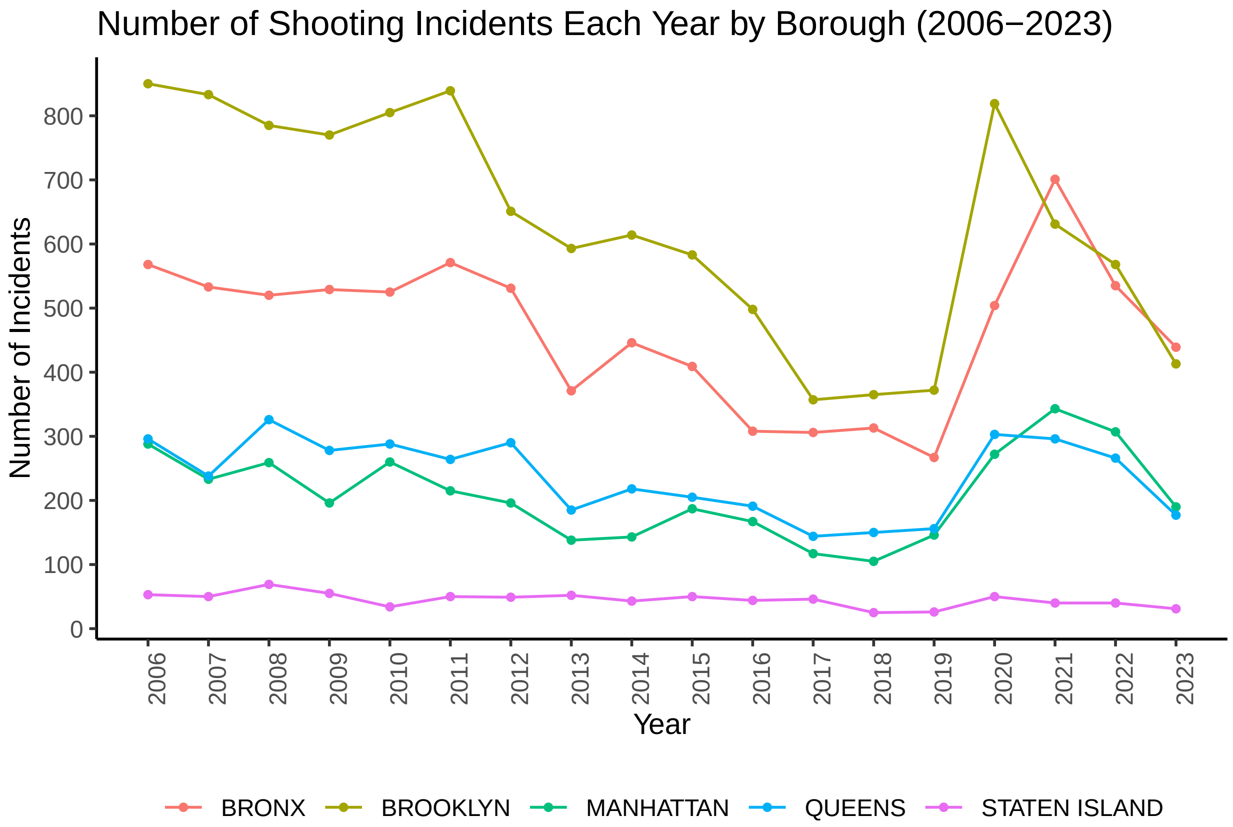


Brooklyn experienced the highest number of shooting incidents over the given time period (11,346). It is followed by the Bronx (8,376), Queens (4,271), Manhattan (3,762), and Staten Island (807).

Question 3: How did the number of shooting incidents in each borough compare between 2006 and 2023?

Below I calculate the number of shooting incidents per year in each borough and plot the data for each together in a line graph for comparison.

```
nypd_w_pop_2 <- nypd_w_pop %>%
  group_by(Borough, Year) %>%
  summarise(Count = n(), .groups = 'drop')

nypd_w_pop_2 %>%
  ggplot(aes(x = Year, y = Count, group = Borough, color = Borough)) +
    geom_line(linewidth = .5) +
    geom_point(size = 1) +
    labs(title = "Number of Shooting Incidents Each Year by Borough (2006-2023)",
         y = "Number of Incidents") +
    theme_classic() +
    theme(axis.text.x = element_text(angle = 90),
          legend.position = "bottom",
          legend.title = element_blank()) +
    scale_x_continuous(breaks = seq(2006, 2023, by = 1)) +
  scale_y_continuous(breaks = seq(0, 1000, by = 100))
```



The number of shooting incidents in each borough appears to trend downward until a significant upswing in 2020 as we also observed previously. Brooklyn generally experienced the highest number of shooting incidents each year with the exception of 2021 and 2023 in which it was overtaken by the Bronx. A similar patter is observed between Queens and Manhattan in which Queens trended higher until 2021 after which Manhattan overtook. Staten Island consistently experienced the lowest number of shooting incidents.

Question 4: How did the total number of shooting incidents in each borough between 2006 and 2023 compare when accounting for population size?

Below I create a new column `Per_100K` in which I calculate the total number of shooting incidents in each borough each year per 100,000 residents. I call this new data set `nypd_per_100k`. I then create a bar chart to visualize the total number of shooting incidents in each borough between 2006 and 2023 per 100,000 residents. I also create a line graph including said data for each of the five boroughs to show how the figure changed over time.

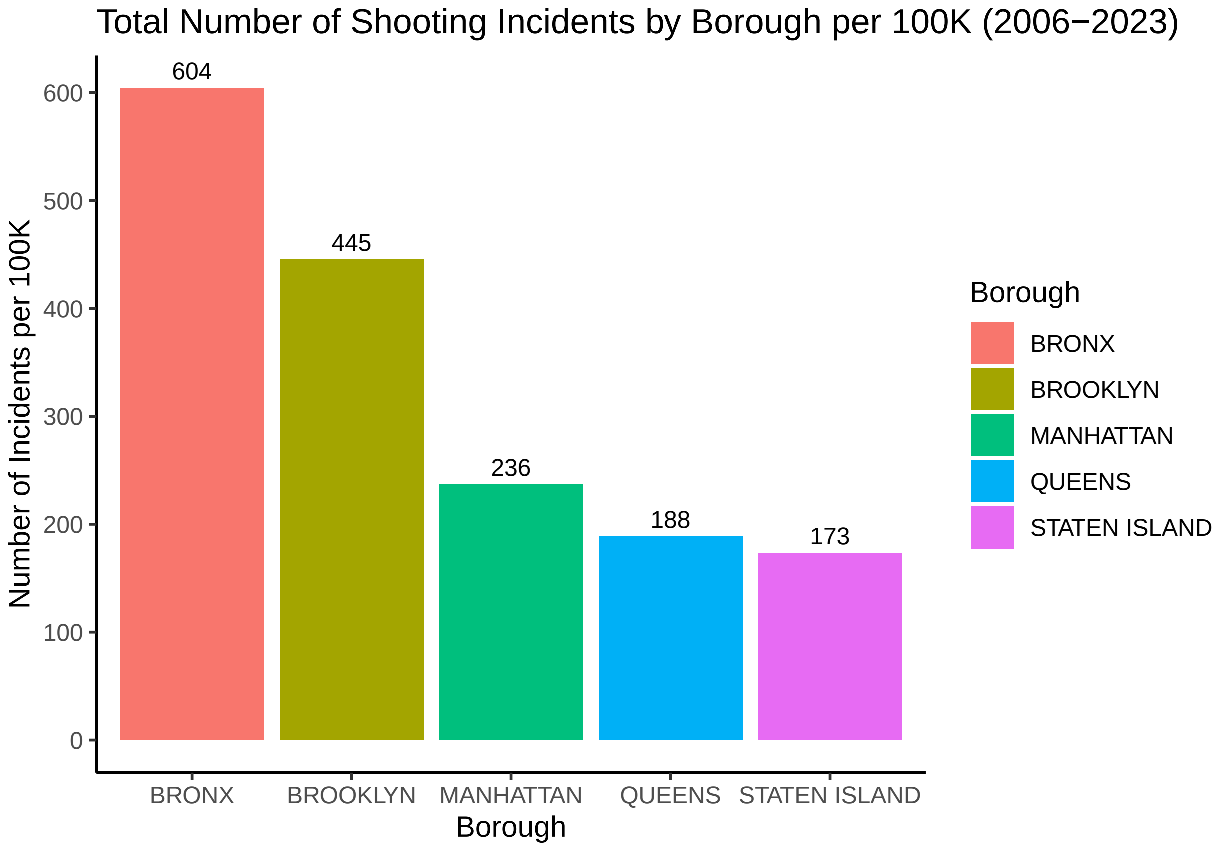
```
nypd_per_100k <- nypd_w_pop %>%
  select(Borough, Year, Population) %>%
  group_by(Borough, Year, Population) %>%
  summarise(Incidents = n()) %>%
  mutate(Per_100K= (Incidents/Population)*100000)
```

```
## `summarise()` has grouped output by 'Borough', 'Year'. You can override using
## the `.groups` argument.
```

```
nypd_per_100k
```

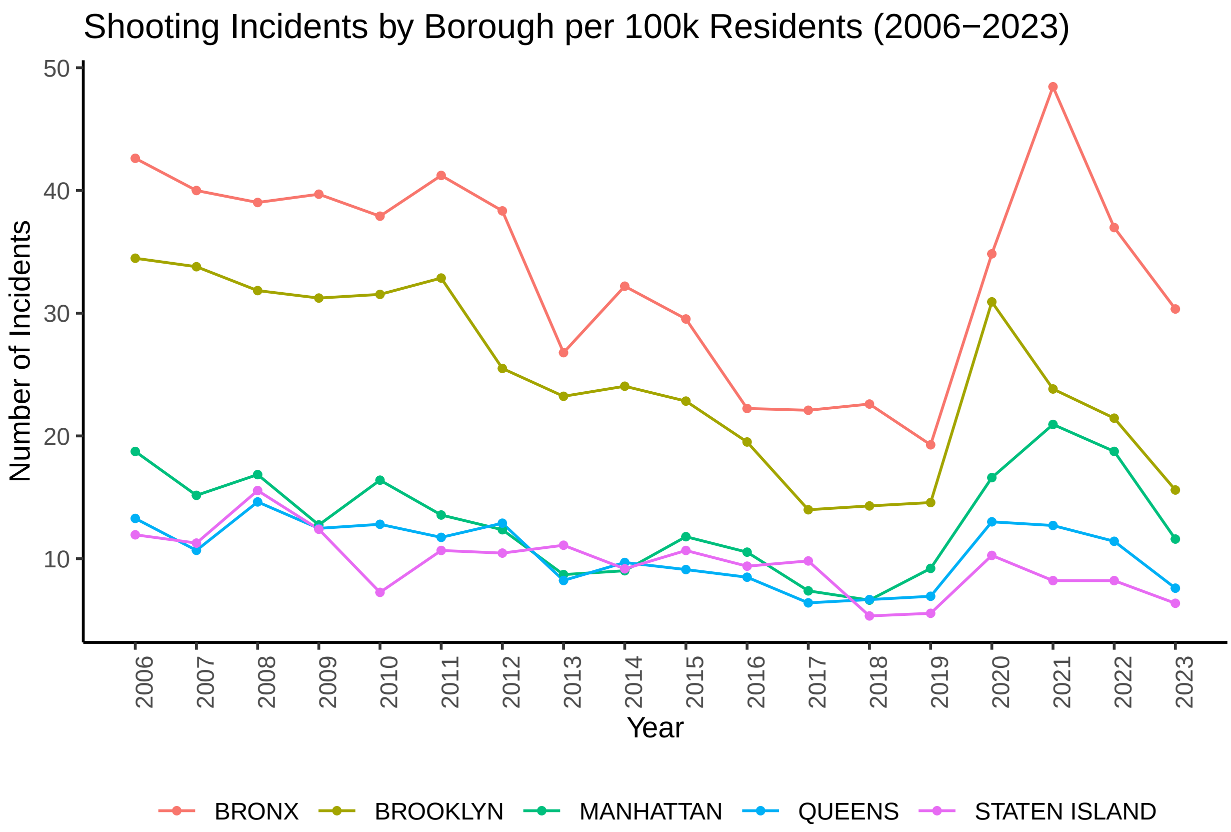
```
## # A tibble: 90 x 5
## # Groups:   Borough, Year [90]
##    Borough  Year Population Incidents Per_100K
##    <chr>   <dbl>      <dbl>     <int>    <dbl>
##  1 BRONX    2006    1332650       568     42.6
##  2 BRONX    2007    1332650       533     40.0
##  3 BRONX    2008    1332650       520     39.0
##  4 BRONX    2009    1332650       529     39.7
##  5 BRONX    2010    1385108       525     37.9
##  6 BRONX    2011    1385108       571     41.2
##  7 BRONX    2012    1385108       531     38.3
##  8 BRONX    2013    1385108       371     26.8
##  9 BRONX    2014    1385108       446     32.2
## 10 BRONX    2015    1385108       409     29.5
## # i 80 more rows
```

```
nypd_per_pop_sum <- nypd_per_100k %>%
  group_by(Borough) %>%
  summarize(Total_Per_100K = sum(Per_100K, na.rm = TRUE))

ggplot(nypd_per_100k, aes(x=factor(Borough, levels = unique(Borough)), y=Per_100K, fill=Borough)) +
  geom_bar(stat="identity") +
  geom_text(data=nypd_per_pop_sum, aes(x=Borough, y=Total_Per_100K, label=floor(Total_Per_100K)),
            vjust=-0.5, size=3) +
  xlab("Borough") + ylab("Number of Incidents per 100K") +
  ggtitle("Total Number of Shooting Incidents by Borough per 100K (2006-2023)") +
  theme_classic() +
  scale_y_continuous(breaks = seq(0, 700, by = 100))
```



Accounting for population size, the boroughs do not fall into the same order in terms of total number of shooting incidents for the period as they did with the gross number. The Bronx is now the clear leader (604) followed by Brooklyn (445), Manhattan (236), Queens (188), and Staten Island (173).

```
nypd_per_100k %>%
ggplot(aes(x = Year, y = Per_100K, group = Borough, color = Borough)) +
    geom_line(linewidth = .5) +
    geom_point(size = 1) +
    labs(title = "Shooting Incidents by Borough per 100k Residents (2006-2023)",
         x = "Year",
         y = "Number of Incidents") +
    theme_classic() +
    theme(axis.text.x = element_text(angle = 90),
          legend.position = "bottom",
          legend.title = element_blank()) +
    scale_x_continuous(breaks = seq(2006, 2023, by = 1))
```



### Modeling

Below I will run a linear regression to predict the number of shooting incidents per 100,000 residents for each borough for 2024-2028 based on the historic data.

```
nypd_per_100k$Borough <- as.factor(nypd_per_100k$Borough)

model <- lm(Per_100K ~ Year + Borough, data = nypd_per_100k)

summary(model)
```

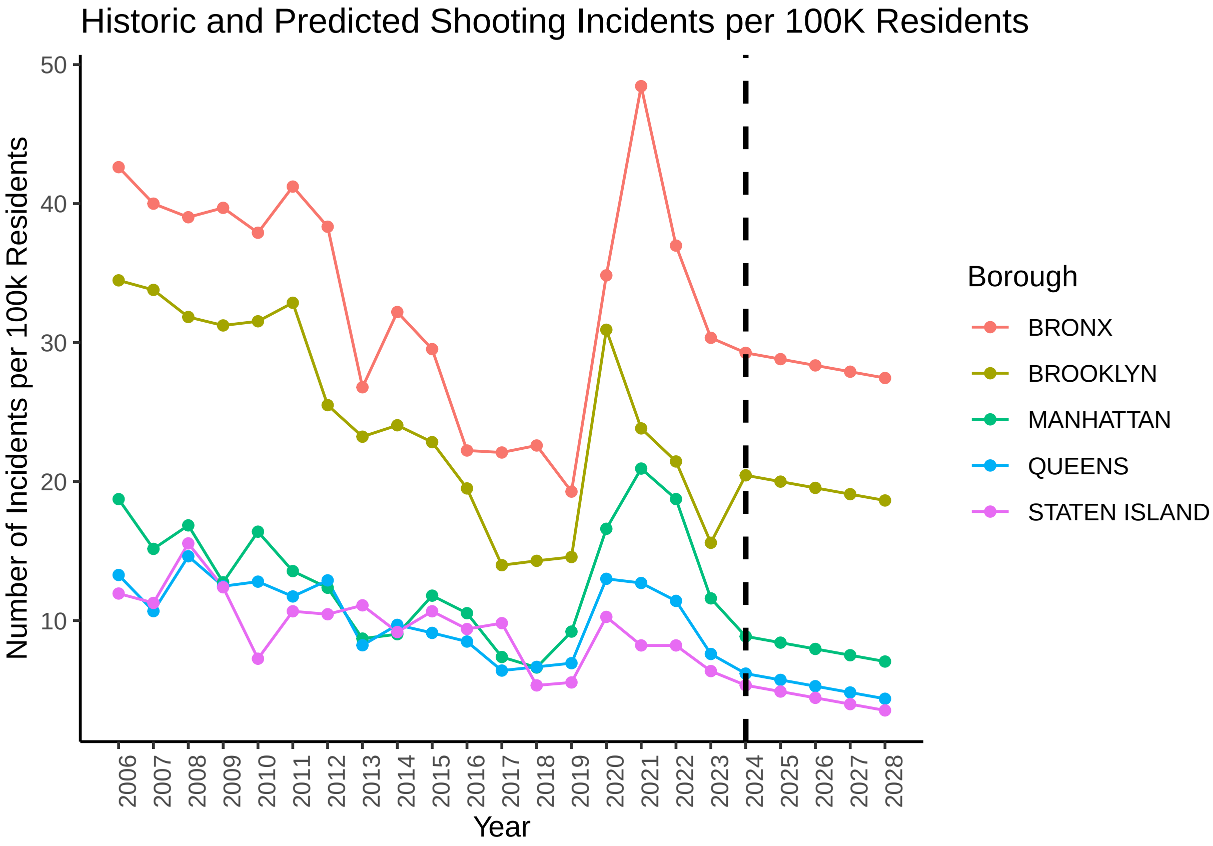
```
##
## Call:
## lm(formula = Per_100K ~ Year + Borough, data = nypd_per_100k)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.249  -2.166  -0.328   2.852  17.832
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             945.3713   204.6061   4.620 1.37e-05 ***
## Year                     -0.4526     0.1016  -4.456 2.55e-05 ***
## BoroughBROOKLYN          -8.8120     1.6663  -5.288 9.63e-07 ***
## BoroughMANHATTAN        -20.3998     1.6663 -12.243  < 2e-16 ***
## BoroughQUEENS           -23.0793     1.6663 -13.851  < 2e-16 ***
## BoroughSTATEN ISLAND -23.9180     1.6663 -14.354  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 4.999 on 84 degrees of freedom
## Multiple R-squared:  0.7993, Adjusted R-squared:  0.7874
## F-statistic: 66.92 on 5 and 84 DF,  p-value: < 2.2e-16
```

```
nypd_pred <- expand.grid(Year = 2024:2028, Borough = levels(nypd_per_100k$Borough))

nypd_pred$Predicted_Per_100K <- predict(model, newdata = nypd_pred)

nypd_pred
```

```
##    Year       Borough Predicted_Per_100K
## 1  2024         BRONX          29.262390
## 2  2025         BRONX          28.809767
## 3  2026         BRONX          28.357144
## 4  2027         BRONX          27.904521
## 5  2028         BRONX          27.451898
## 6  2024      BROOKLYN          20.450344
## 7  2025      BROOKLYN          19.997721
## 8  2026      BROOKLYN          19.545098
## 9  2027      BROOKLYN          19.092475
## 10 2028      BROOKLYN          18.639852
## 11 2024     MANHATTAN           8.862598
## 12 2025     MANHATTAN           8.409975
## 13 2026     MANHATTAN           7.957352
## 14 2027     MANHATTAN           7.504729
## 15 2028     MANHATTAN           7.052106
## 16 2024        QUEENS           6.183094
## 17 2025        QUEENS           5.730471
## 18 2026        QUEENS           5.277848
## 19 2027        QUEENS           4.825225
## 20 2028        QUEENS           4.372602
## 21 2024 STATEN ISLAND           5.344358
## 22 2025 STATEN ISLAND           4.891735
## 23 2026 STATEN ISLAND           4.439112
## 24 2027 STATEN ISLAND           3.986489
## 25 2028 STATEN ISLAND           3.533866
```

```
combined_data <- bind_rows(nypd_per_100k %>% mutate(Type = "Historical"),nypd_pred %>% mutate(Type = "P

ggplot(combined_data, aes(x = Year, y = ifelse(Type == "Historical", Per_100K, Predicted_Per_100K), col
  geom_line() +
  geom_point() +
  theme_classic() +
  theme(axis.text.x = element_text(angle = 90)) +
  geom_vline(xintercept = 2024, linetype = "dashed", color = "black", size = 1) +
  labs(title = "Historic and Predicted Shooting Incidents per 100K Residents",
       x = "Year",
       y = "Number of Incidents per 100k Residents") +
    scale_x_continuous(breaks = seq(2006, 2028, by = 1))
```

```
## 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.
```



Conclusion

The number of shooting incidents in New York City was in steady decline until a sharp upswing in 2020 which then peaked in 2021 (2,011 incidents) before again beginning to decline through 2023. Brooklyn experienced the highest overall number of shooting incidents (11,346), but accounting for population size, the Bronx experienced the greatest number (604) per 100,000 residents. Using a linear regression, I predict that the number of shooting incidents per 100k residents in each borough will continue to gradually decrease through 2028. The bias in this analysis should be fairly minimal. I chose to study these variables as there was no missing data and it did not include variables such as race, age, and gender with which bias is more likely. It is however possible there are unreported shootings that took place in each borough that are not included in this data. Shootings involving illegal activity or domestic disputes that were not reported to police could potentially have impacted the final analysis.