NYPD Data Report

1/29/2022

Introduction

In this R Markdown document, we are looking into NYPD shooting incidents. The data set that will be used is called NYPD Shooting Incident Data (Historic). This data set if found on https://catalog.data.gov/dataset. First, we will remove any of the columns that we do not plan on using in this report.

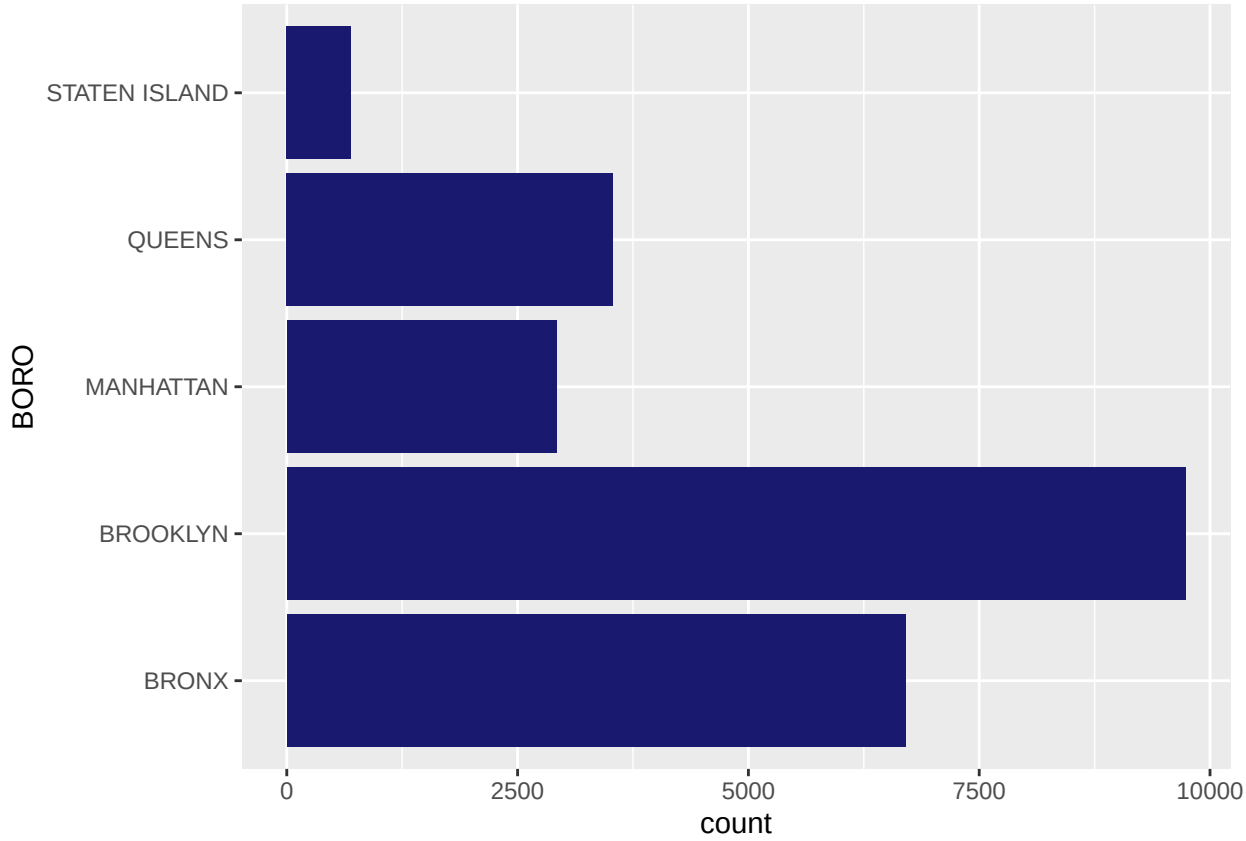
```
library("tidyverse")
library("ggpubr")
```

```
nypd_data <- read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
nypd_data <- nypd_data %>%
  select(INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, BORO, PRECINCT,STATISTICAL_MURDER_FLAG,
         VIC_AGE_GROUP, VIC_SEX, VIC_RACE)
```

Initial Impressions

Next, we are going to chart the number of shooting incidents by New York City borough. As you can see from the chart below, Brooklyn and the Bronx have the most number of shootings while Manhattan and Staten Island have the least. Since there is no data on total population present in the data set, we cannot calculate the shooting incidents per capita.

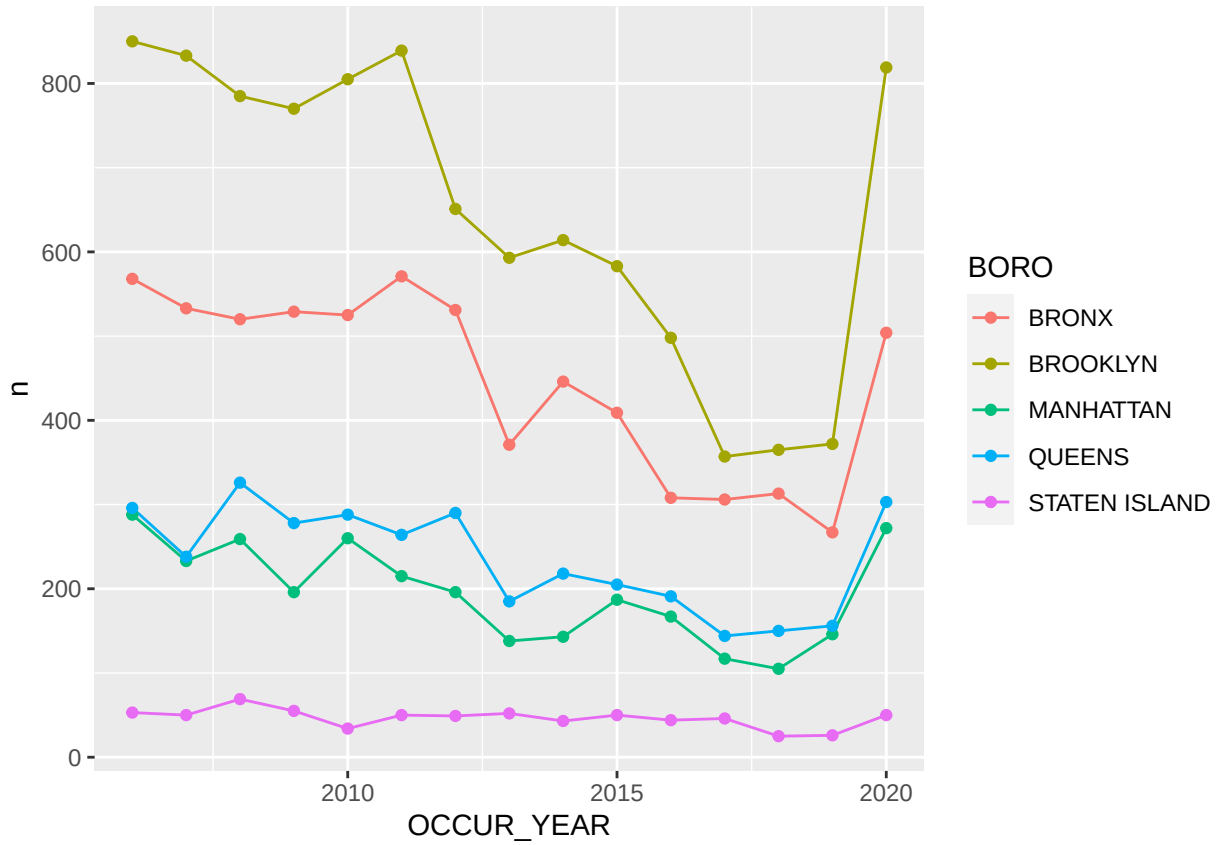
```
nypd_data %>%
  ggplot(aes(x = BORO))+
  geom_bar(fill = "midnightblue")+
  coord_flip()
```



The last chart gave us a total number of shootings across all boroughs over the whole time span from 2006 to 2020. Now let us look at the shootings over that time span by borough by year.

Shootings by Year

```
nypd_data %>%
  select(OCCUR_DATE, BORO) %>%
  mutate(OCCUR_YEAR = str_sub(OCCUR_DATE,-4)) %>%
  mutate(OCCUR_YEAR = as.numeric(OCCUR_YEAR)) %>%
  count(BORO, OCCUR_YEAR) %>%
  ggplot(aes(x = OCCUR_YEAR, y = n, group = BORO)) + geom_line(aes(color = BORO)) +
    geom_point(aes(color = BORO))
```



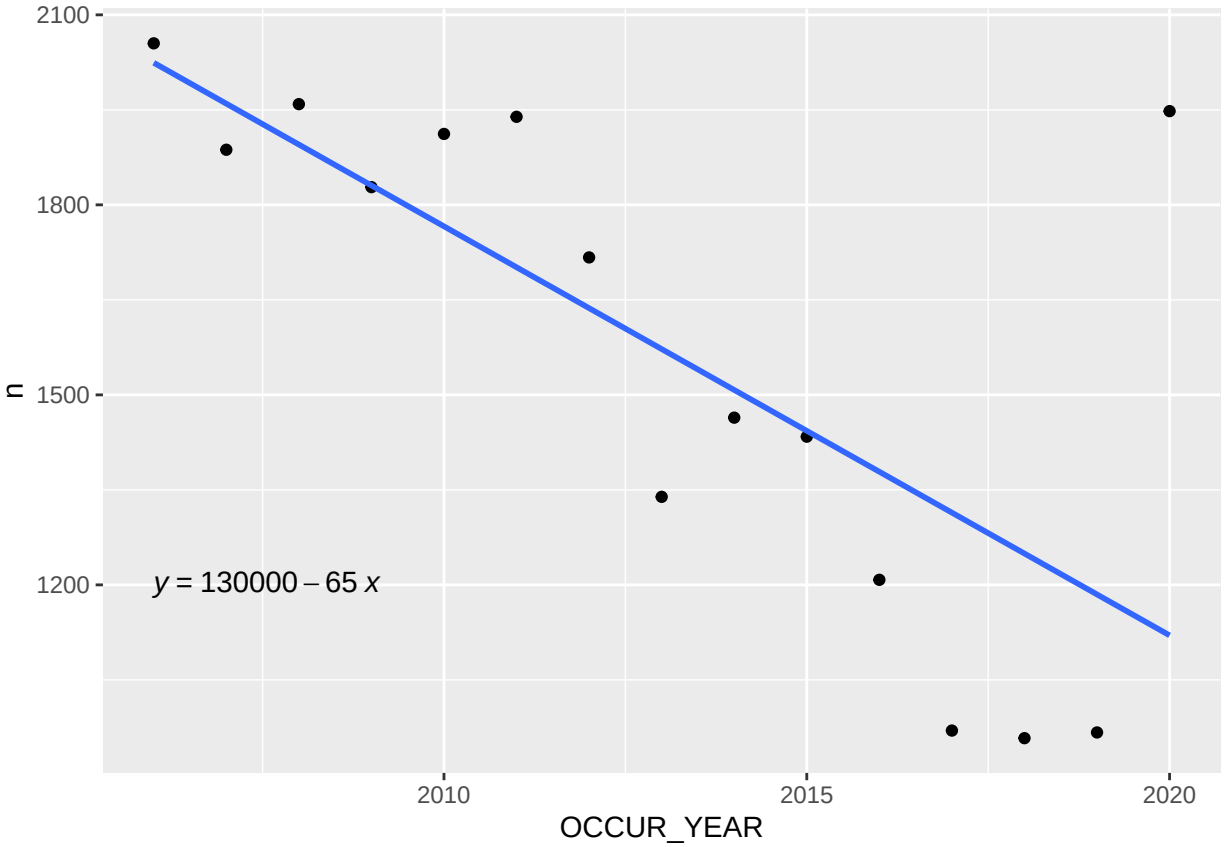
The data shows that from 2006 all the way up to 2018, there have mostly been a decrease in the total number of shootings in all New York City boroughs. The year with the highest number of shootings in each borough is roughly half of the lowest year outside of Staten Island. However, in 2020, the number of shootings have spiked for all boroughs up to 2010 numbers.

Modeling Future Shootings

Now let us look at the trend of total shootings in all of New York City. Since we do not have data in this data set on the population of each borough, we will combine the boroughs and shooting incidents together to get a general view of shooting trends in all of New York City.

```
nypd_data %>%
  select(OCCUR_DATE) %>%
  mutate(OCCUR_YEAR = str_sub(OCCUR_DATE,-4)) %>%
  mutate(OCCUR_YEAR = as.numeric(OCCUR_YEAR)) %>%
  count(OCCUR_YEAR) %>%
  ggplot(aes(x = OCCUR_YEAR, y = n)) + geom_point() + geom_smooth(method = lm, se = FALSE) +
    stat_regline_equation(label.x = 2006, label.y = 1200)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



In most of the years from 2006 to 2020, there has been a trend of decreasing shootings. The equation of the trend line y=130000-65x shows that each year there is a trend of 65 fewer shooting incidents a year in New York City. The 130000 would be the trend if we started in year 0 instead of year 2006. However, there was a spike in shooting incidents in 2020. Without more data in future years, we cannot conclude if that one year is an outlier or if there is a trend of more shootings in the future.

Possible bias in the data would be location bias and time lag bias. For location bias, police departments depending on borough may have different emphasis on reporting this data. For example, police in the Bronx and Brooklyn may see that reporting this data is a higher priority than the police in Staten Island. As for time lag data, the police might have reported data in one year when the original incident happened in a prior year. This may explain why there was a giant spike in 2020. When there were far fewer people on the street, the police might have had more time to report on numbers from a prior year.