NY Shooting Incident Analysis by Boro

Data Science Student

January 2, 2022

Introduction

I would like to investigate which boro in New York is the safest place to live. To do so, I will load NYPD shooting incident data for analysis. The NYPD Shooting Incident Data (Historic) dataset lists each shooting incident in New York City through end of the previous calendar year. More information about this dataset can be found on data.gov.

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

Dataset Description: This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity.

Dataset Summary Information

```
summary(data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596      Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms        Class :character
##  Median : 86437258   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA’s   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596       Length:25596      
##  Class :character   Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD    
##  Length:25596       Length:25596       Min.   : 914928   Min.   :125757  
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782  
##  Mode  :character   Mode  :character   Median :1007715   Median :194038  
##                                        Mean   :1009455   Mean   :207894  
##                                        3rd Qu.:1016838   3rd Qu.:239429  
##                                        Max.   :1066815   Max.   :271128  
##                                                                          
##     Latitude       Longitude        Lon_Lat         
##  Min.   :40.51   Min.   :-74.25   Length:25596      
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character  
##  Median :40.70   Median :-73.92   Mode  :character  
##  Mean   :40.74   Mean   :-73.91                     
##  3rd Qu.:40.82   3rd Qu.:-73.88                     
##  Max.   :40.91   Max.   :-73.70                     
##
```

Updating Data Types and Removing Extraneous Columns

To prepare the data, the date and time fields were combined and cast as a POSIX datetime object. Categorical data were assigned a factor data type. The incident key, which is unique for each incident was cast as a character type.

There were numerous fields with no values. These were re-labeled to match the Unknown category for the appropriate field.

As this analysis will not focus on longitude and latitude, these geographical fields were dropped from the data frame.

```r
data[['DATETIME']] <- paste(data$OCCUR_DATE, "-" ,data$OCCUR_TIME)
data[['DATETIME']] <- as.POSIXct(data[['DATETIME']],format = "%m/%d/%Y - %H:%M:%S")
data[['YEAR']] <- as.numeric(format(data$DATETIME, "%Y"))
data[['VIC_SEX']] <- as.factor(data$VIC_SEX)
data[['VIC_RACE']] <- as.factor(data$VIC_RACE)
data[['VIC_AGE_GROUP']] <- as.factor(data$VIC_AGE_GROUP)
data[['PERP_SEX']] <- as.factor(data$PERP_SEX)
data[['PERP_RACE']] <- as.factor(data$PERP_RACE)
data[['PERP_AGE_GROUP']] <- as.factor(data$PERP_AGE_GROUP)
data[['BORO']] <- as.factor(data$BORO)
data[['PRECINCT']] <- as.factor(data$PRECINCT)
data[['INCIDENT_KEY']] <- as.character(data$INCIDENT_KEY)
data[['JURISDICTION_CODE']] <- as.factor(data$JURISDICTION_CODE)

data = subset(data, select = -c(Lon_Lat,Longitude,Latitude,Y_COORD_CD,X_COORD_CD,OCCUR_TIME,OCCUR_DATE)

data['PERP_AGE_GROUP'][is.na(data['PERP_AGE_GROUP'])] <- "UNKNOWN"
data['PERP_SEX'][is.na(data['PERP_SEX'])] <- "U"
```

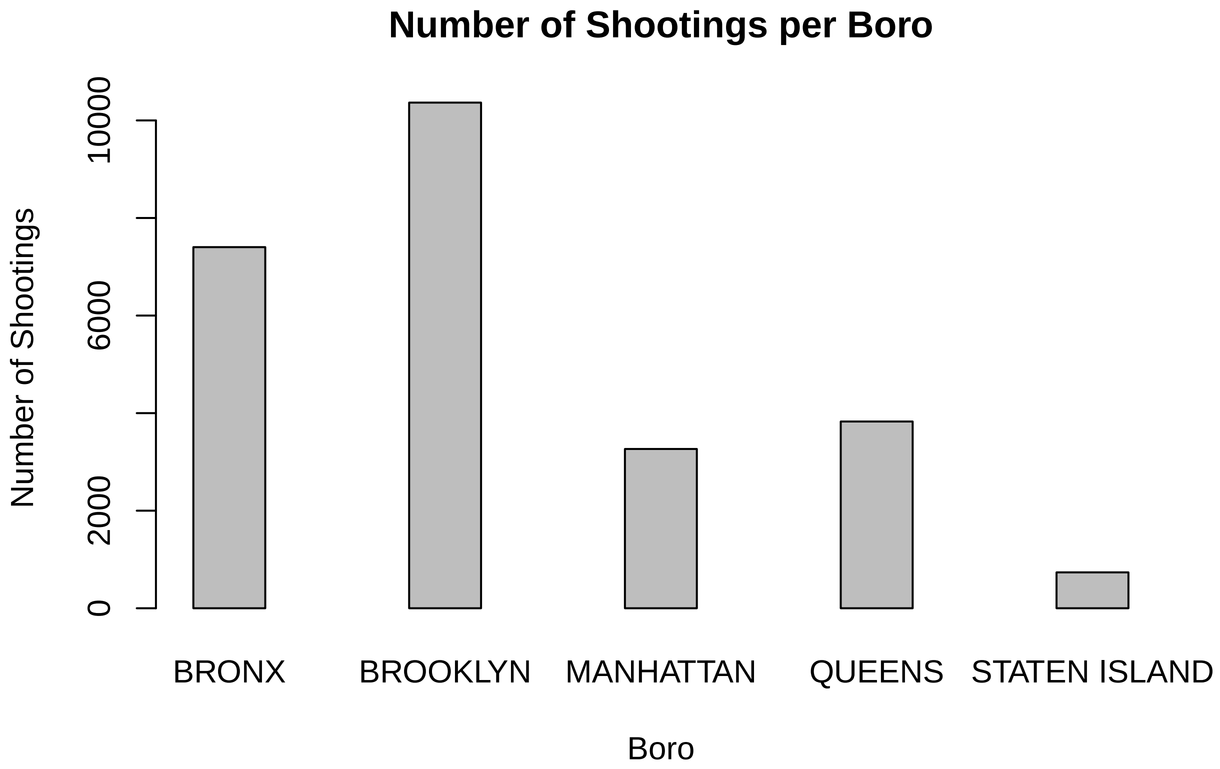
```
data['PERP_RACE'][is.na(data['PERP_RACE'])] <- "UNKNOWN"
data['VIC_AGE_GROUP'][is.na(data['VIC_AGE_GROUP'])] <- "UNKNOWN"
data['VIC_SEX'][is.na(data['VIC_SEX'])] <- "U"
data['VIC_RACE'][is.na(data['VIC_RACE'])] <- "UNKNOWN"

summary(data)
```

```
##  INCIDENT_KEY                  BORO          PRECINCT     JURISDICTION_CODE
##  Length:25596       BRONX        : 7402   75     : 1470   0   :21321
##  Class :character   BROOKLYN     :10365   73     : 1372   1   :   59
##  Mode  :character   MANHATTAN    : 3265   67     : 1160   2   : 4214
##                     QUEENS       : 3828   79     :  982   NA’s:    2
##                     STATEN ISLAND:  736   44     :  949
##                                           47     :  903
##                                           (Other):18760
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP   PERP_SEX
##  Length:25596       Mode  :logical          UNKNOWN:12492   F:  371
##  Class :character   FALSE:20668             18-24  : 5844   M:14416
##  Mode  :character   TRUE :4928              25-44  : 5202   U:10809
##                                             <18    : 1463
##                                             45-64  :  535
##                                             65+    :   57
##                                             (Other):    3
##                            PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    : 2681   F: 2403
##  ASIAN / PACIFIC ISLANDER      :  141   18-24  : 9604   M:23182
##  BLACK                         :10668   25-44  :11386   U:   11
##  BLACK HISPANIC                : 1203   45-64  : 1698
##  UNKNOWN                       :11146   65+    :  167
##  WHITE                         :  272   UNKNOWN:   60
##  WHITE HISPANIC                : 2164
##                             VIC_RACE        DATETIME
##  AMERICAN INDIAN/ALASKAN NATIVE:    9   Min.   :2006-01-01 02:00:00.00
##  ASIAN / PACIFIC ISLANDER      :  354   1st Qu.:2009-05-10 04:05:00.00
##  BLACK                         :18281   Median :2012-08-26 01:05:00.00
##  BLACK HISPANIC                : 2485   Mean   :2013-06-14 04:24:56.10
##  UNKNOWN                       :   65   3rd Qu.:2017-07-01 00:20:15.00
##  WHITE                         :  660   Max.   :2021-12-31 19:23:00.00
##  WHITE HISPANIC                : 3742
##       YEAR
##  Min.   :2006
##  1st Qu.:2009
##  Median :2012
##  Mean   :2013
##  3rd Qu.:2017
##  Max.   :2021
##
```

Analysis

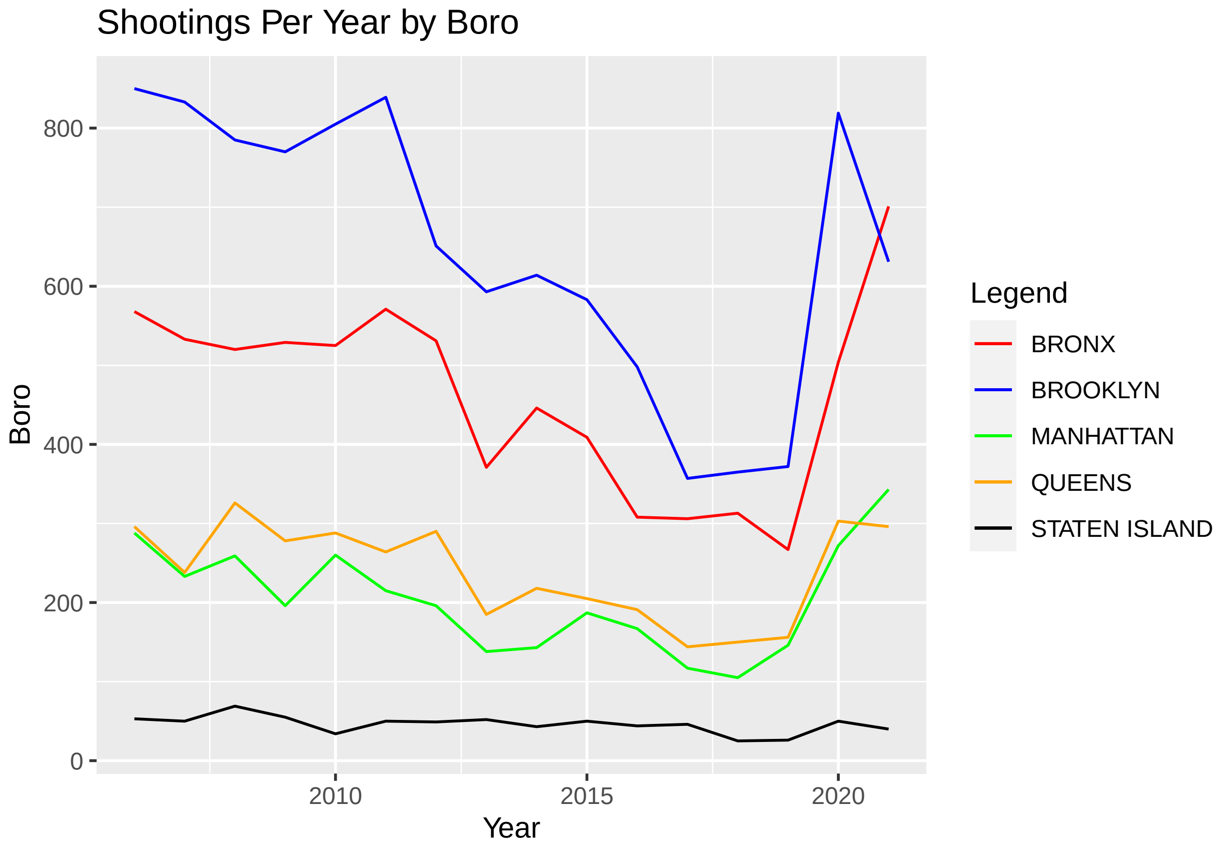
A first question to be evaluated is whether the number of shootings are different across different parts of New York. First a simple bar chart has been provided to exhibit the total number of shootings that have taken place in each New York boro since the beginning of the data set. This shows that the greatest number of shoots took place in Brooklyn, with over 10,000, while the Bronx trails slightly, with nearly 8,000 shootings over that time period.

```
barplot(table(data$BORO),xlab="Boro",ylab="Number of Shootings",main="Number of Shootings per Boro",spac
```



A second question is whether the number of shootings in each boro has flutated over time. Therefore, a line graph has been provided to plot the number of shootings per year in each boro. It shows through 2019 there were steep declines in the number of shooting throughout the city, with the greatest drop taking place in Brooklyn. However, from 2020 onwards, the numbers have risen strikingly. This corresponds with the years of the pandemic.

```
BORO_YEAR <- data %>% group_by(BORO,YEAR) %>%
  summarise(total_count=n(),.groups = 'drop') %>%
  as.data.frame()
BORO_YEAR <- BORO_YEAR %>% pivot_wider(names_from = BORO, values_from = total_count)
BORO_YEAR['STATEN'] <- BORO_YEAR['STATEN ISLAND']

colors <- c("BRONX" = "red", "BROOKLYN" = "blue", "MANHATTAN" = "green", "QUEENS" = "orange", "STATEN IS
ggplot(BORO_YEAR, aes(x=YEAR)) +
  geom_line(aes(y = BRONX, color = "BRONX")) +
  geom_line(aes(y = BROOKLYN, color="BROOKLYN")) +
  geom_line(aes(y = MANHATTAN, color="MANHATTAN")) +
  geom_line(aes(y = QUEENS, color="QUEENS")) +
  geom_line(aes(y = STATEN, color="STATEN ISLAND")) +
  labs(x = "Year", y = "Boro", color = "Legend", title = "Shootings Per Year by Boro") +
  scale_color_manual(values = colors)
```



Future Investigation

After creating these two plots, I would next want to investigate how I could compare the boros using the same scale. For instance, while the graphs are true in absolute terms, comparisons between the boros are misleading because they may have different populations. I would want to import population statistics to the model and create a shootings per 1,000 people measure. Then I could better compare the incidence of shootings.

Conclusion

The intial analysis suggests that Staten Island would be the safest part of the New York in which to live. The number of shootings trails behind the other boros. However, while providing some interesting first analysis of the data set, bias is present because of the differences in population between the boros. It will be necessary to create a better measure to compare the number of shooting incidents in each boro. Personal bias includes the fact that I am only looking at shootings as a measure of safety. I live in a city with gun violence, so I assume that gun violence is a proxy for crime in general. However, perhaps other forms of crime are more prevasive or destructive than gun violence in New York. To get a better understanding of whether Staten Island is the safest boro, it would necessary to expand the data model to include other crime categories.

Loaded Libraries

```
(.packages())
```

```
##  [1] "forcats"   "stringr"   "dplyr"     "purrr"     "readr"     "tidyr"    
##  [7] "tibble"    "ggplot2"   "tidyverse" "stats"     "graphics"  "grDevices"
## [13] "utils"     "datasets"  "methods"   "base"
```