NYPD Shooting Project

William Wilson

2023-06-28

The following packages and customizations are used in this report:

```
library(tidyverse)
library(lubridate)
library(forcats)
library(ggplot2)
library(ggthemes)
library(ggmap)
library(osmdata)
library(knitr)
library(modelr)


# A nice theme for our plots
theme_set(theme_economist())

#Size up our figures as needed
opts_template$set(map_fig = list(fig.height = 10, fig.width = 10))
```

Tidying and Transforming Data

The data for this project was taken from https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic. The file is in .csv format and contains a "List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year."

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

First, we convert the data to a tibble and remove unnecessary columns.

```
#Convert data to a tibble and use mutate to drop unwanted columns
shooting_data <- shooting_data %>% as_tibble() %>% mutate(
  INCIDENT_KEY = NULL,
  OCCUR_TIME = NULL,
  LOC_OF_OCCUR_DESC = NULL,
  JURISDICTION_CODE = NULL,
  LOC_CLASSFCTN_DESC = NULL,
  LOCATION_DESC = NULL,
  STATISTICAL_MURDER_FLAG = NULL,
  X_COORD_CD = NULL,
  Y_COORD_CD = NULL,
  Lon_Lat = NULL
  )
```

Second, we convert OCCUR_DATE to Date Object.

```
#convert to date object using lubridate mdy function
shooting_data <- shooting_data %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

Next, convert the following columns to use factors:

```
#Select the various columns that can be viewed as factors
shooting_data <- shooting_data %>% mutate(PERP_AGE_GROUP = as_factor(PERP_AGE_GROUP))
shooting_data <- shooting_data %>% mutate(PERP_SEX = as_factor(PERP_SEX))
shooting_data <- shooting_data %>% mutate(PERP_RACE = as_factor(PERP_RACE))
shooting_data <- shooting_data %>% mutate(VIC_AGE_GROUP = as_factor(VIC_AGE_GROUP))
shooting_data <- shooting_data %>% mutate(VIC_SEX = as_factor(VIC_SEX))
shooting_data <- shooting_data %>% mutate(VIC_RACE = as_factor(VIC_RACE))
shooting_data <- shooting_data %>% mutate(BORO = as_factor(BORO))
```

A summary of the data shows the following:

```
summary(shooting_data)
```

```
##    OCCUR_DATE                    BORO         PRECINCT      PERP_AGE_GROUP
##  Min.   :2006-01-01   QUEENS       : 4094   Min.   :  1.00          :9344
##  1st Qu.:2009-07-18   BRONX        : 7937   1st Qu.: 44.00   18-24  :6222
##  Median :2013-04-29   BROOKLYN     :10933   Median : 68.00   25-44  :5687
##  Mean   :2014-01-06   MANHATTAN    : 3572   Mean   : 65.64   UNKNOWN:3148
##  3rd Qu.:2018-10-15   STATEN ISLAND:  776   3rd Qu.: 81.00   <18    :1591
##  Max.   :2022-12-31                         Max.   :123.00   (null) : 640
##                                                              (Other): 680
##    PERP_SEX              PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##         : 9310   BLACK         :11432   18-24  :10086   M:24686
##  M     :15439                 : 9310   25-44  :12281   F: 2615
##  U     : 1499   WHITE HISPANIC: 2341   <18    : 2839   U:   11
##  F     :  424   UNKNOWN       : 1836   45-64  : 1863
##  (null):  640   BLACK HISPANIC: 1314   65+    :  181
##                 (null)        :  640   UNKNOWN:   61
##                 (Other)       :  439   1022   :    1
##                              VIC_RACE        Latitude       Longitude
##  BLACK                         :19439   Min.   :40.51   Min.   :-74.25
##  WHITE                         :  698   1st Qu.:40.67   1st Qu.:-73.94
##  WHITE HISPANIC                : 4049   Median :40.70   Median :-73.92
##  BLACK HISPANIC                : 2646   Mean   :40.74   Mean   :-73.91
##  ASIAN / PACIFIC ISLANDER      :  404   3rd Qu.:40.82   3rd Qu.:-73.88
##  UNKNOWN                       :   66   Max.   :40.91   Max.   :-73.70
##  AMERICAN INDIAN/ALASKAN NATIVE:   10   NA's   :10      NA's   :10
```

There is blank data in several columns. It can be dealt with by assigning both blank and null values to 'UNKNOWN,' as it is already a category in each column.

```
#Fill in blank and (null) cells
shooting_data$PERP_AGE_GROUP <- shooting_data$PERP_AGE_GROUP %>%
  fct_collapse(UNKNOWN = c("UNKNOWN","","(null)"))

shooting_data$PERP_SEX <- shooting_data$PERP_SEX %>%
  fct_collapse(U = c("U","","(null)"))

shooting_data$PERP_RACE <- shooting_data$PERP_RACE %>%
  fct_collapse(UNKNOWN = c("UNKNOWN","","(null)"))
```

There are also several errors in the Victim's Age Groups. Since we do not now the original intent of the data we shall categorize it as "UNKNOWN". This is dealt with below:

```
#All impossible age groups are changed to unknown
shooting_data$VIC_AGE_GROUP <- shooting_data$VIC_AGE_GROUP %>%
  fct_recode("UNKNOWN" = "1022")
shooting_data$PERP_AGE_GROUP <- shooting_data$PERP_AGE_GROUP %>%
  fct_recode("UNKNOWN" = "940")
shooting_data$PERP_AGE_GROUP <- shooting_data$PERP_AGE_GROUP %>%
  fct_recode("UNKNOWN" = "224")
shooting_data$PERP_AGE_GROUP <- shooting_data$PERP_AGE_GROUP %>%
  fct_recode("UNKNOWN" = "1020")
```

Here is our revised summary:

```
summary(shooting_data)
```

```
##    OCCUR_DATE                    BORO          PRECINCT      PERP_AGE_GROUP
##  Min.   :2006-01-01   QUEENS       : 4094   Min.   :  1.00   UNKNOWN:13135
##  1st Qu.:2009-07-18   BRONX        : 7937   1st Qu.: 44.00   25-44  : 5687
##  Median :2013-04-29   BROOKLYN     :10933   Median : 68.00   18-24  : 6222
##  Mean   :2014-01-06   MANHATTAN    : 3572   Mean   : 65.64   45-64  :  617
##  3rd Qu.:2018-10-15   STATEN ISLAND:  776   3rd Qu.: 81.00   <18    : 1591
##  Max.   :2022-12-31                         Max.   :123.00   65+    :   60
##
##  PERP_SEX                           PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  U:11449   UNKNOWN                       :11786   18-24  :10086   M:24686
##  M:15439   BLACK                         :11432   25-44  :12281   F: 2615
##  F:  424   BLACK HISPANIC                : 1314   <18    : 2839   U:   11
##            ASIAN / PACIFIC ISLANDER      :  154   45-64  : 1863
##            WHITE HISPANIC                : 2341   65+    :  181
##            WHITE                         :  283   UNKNOWN:   62
##            AMERICAN INDIAN/ALASKAN NATIVE:    2
##                           VIC_RACE        Latitude       Longitude
##  BLACK                         :19439   Min.   :40.51   Min.   :-74.25
##  WHITE                         :  698   1st Qu.:40.67   1st Qu.:-73.94
##  WHITE HISPANIC                : 4049   Median :40.70   Median :-73.92
##  BLACK HISPANIC                : 2646   Mean   :40.74   Mean   :-73.91
##  ASIAN / PACIFIC ISLANDER      :  404   3rd Qu.:40.82   3rd Qu.:-73.88
##  UNKNOWN                       :   66   Max.   :40.91   Max.   :-73.70
##  AMERICAN INDIAN/ALASKAN NATIVE:   10   NA’s   :10      NA’s   :10
```

A few NA's still exist in Longitude and Latitude, but we will ignore these rows when creating map data. Finally, we reorder the factors so that age groups are presented in the correct order:

```
#order so that it displays as youngest to oldest
shooting_data$PERP_AGE_GROUP <-
  factor(shooting_data$PERP_AGE_GROUP, c('<18','18-24','25-44','45-64','65+','UNKNOWN'))
shooting_data$VIC_AGE_GROUP <-
  factor(shooting_data$VIC_AGE_GROUP, c('<18','18-24','25-44','45-64','65+','UNKNOWN'))
```

Data Visualization and Analysis

Age Data

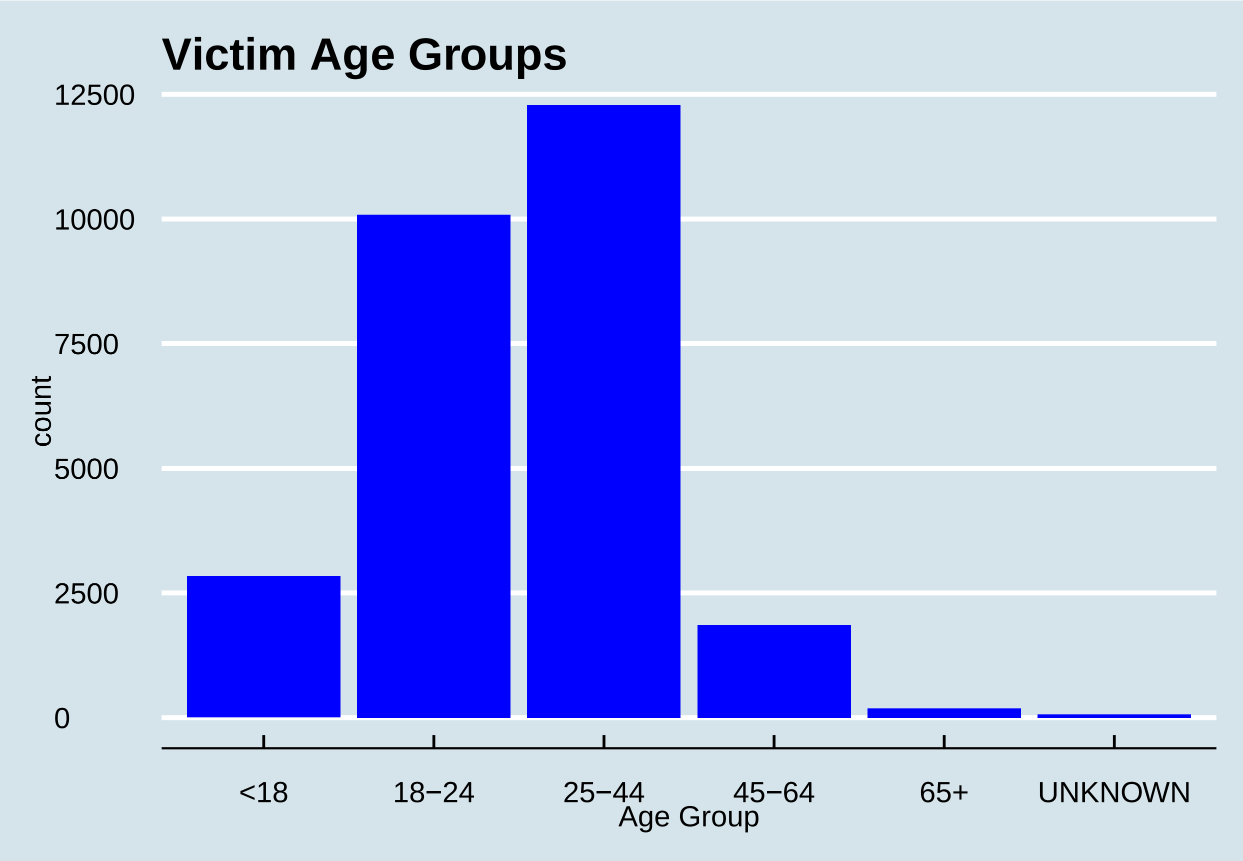
In this section we will analyze and visualize the age data.

First, let's look at the age information for victims:

```
#groups by age group factor and summarizes
shooting_data_by_age_V <- shooting_data %>% group_by(VIC_AGE_GROUP)
summarize(shooting_data_by_age_V, count=n())
```

```
## # A tibble: 6 x 2
##   VIC_AGE_GROUP count
##   <fct>         <int>
## 1 <18            2839
## 2 18-24         10086
## 3 25-44         12281
## 4 45-64          1863
## 5 65+             181
## 6 UNKNOWN          62
```

```
#make bar chart
ggplot(data=shooting_data,aes(x=VIC_AGE_GROUP)) +
  geom_bar(fill='blue') +
  labs(title='Victim Age Groups',x='Age Group')
```



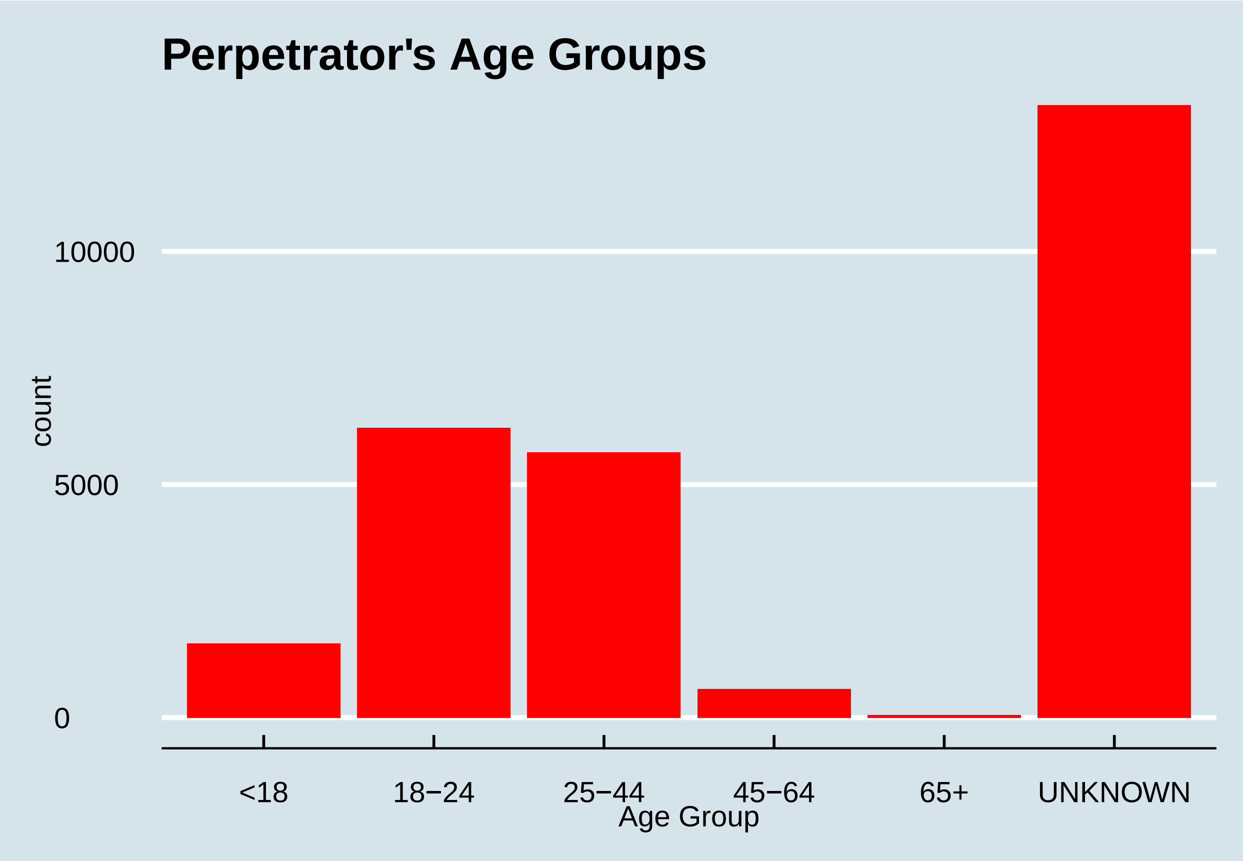
Next, let's look at the age information for perpetrators :

```
#groups by age group factor and summarizes
shooting_data_by_age_P <- shooting_data %>% group_by(PERP_AGE_GROUP)
summarize(shooting_data_by_age_P, count=n())
```

```
## # A tibble: 6 x 2
##   PERP_AGE_GROUP count
##   <fct>          <int>
## 1 <18             1591
## 2 18-24           6222
## 3 25-44           5687
## 4 45-64            617
## 5 65+               60
## 6 UNKNOWN        13135
```

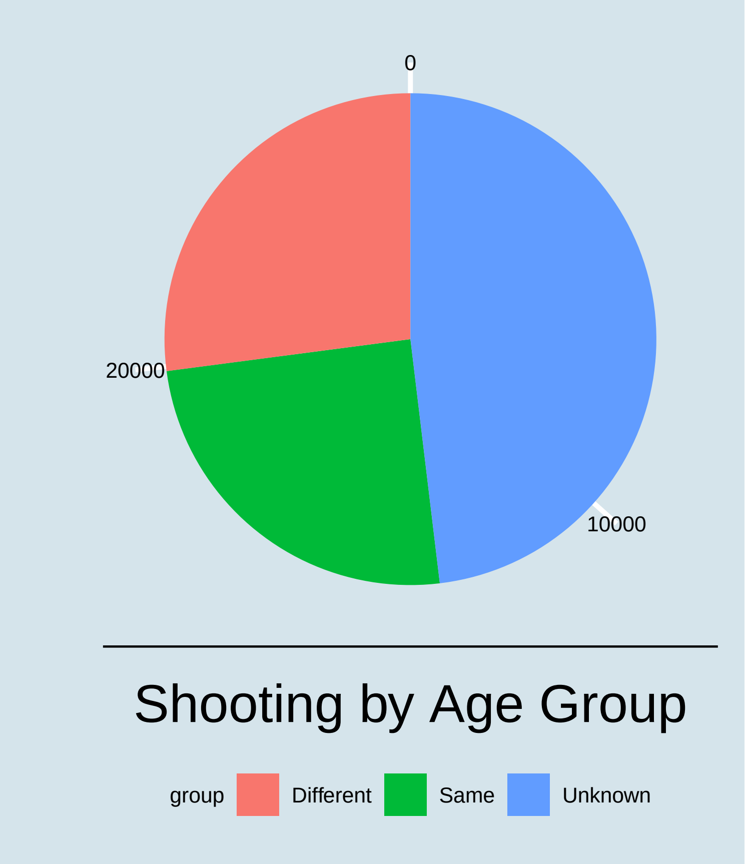
A chart of the perpetrator's age distribution follows:

```
#make bar chart
ggplot(data=shooting_data,aes(x=PERP_AGE_GROUP)) +
  geom_bar(fill='red') +
  labs(title="Perpetrator\'s Age Groups",x='Age Group')
```



Let's compare the difference of ages in a single incident. In other words, are most crimes being committed within the same age group?

```
#first filter out unknown as a comparison can not be made.
sd_known_ages <- filter(shooting_data, PERP_AGE_GROUP != 'UNKNOWN')

#Now calculate total of same age shootings and total known shootings
num_same_age <-  sum(sd_known_ages$PERP_AGE_GROUP == sd_known_ages$VIC_AGE_GROUP)

# pull converts tibble into vector
#so it can be included in the pie chart

num_total <-  pull(count(sd_known_ages))

num_diff <- num_total - num_same_age
num_unknown <- pull(count(shooting_data)) - num_total

percentage <-  round((num_same_age / num_total*100),digits=2)

#Create a Pie chart that shows results
df <- data.frame(value = c(num_same_age, num_diff,num_unknown),
group = c("Same","Different","Unknown"))
ggplot(df, aes(x = "", y = value, fill = group)) +
geom_col() +
coord_polar(theta = "y") +
  labs(y="Shooting by Age Group", x= "") + theme(
  legend.position = "bottom",
  legend.title = element_text(size=8),
  axis.title = element_text(size=20),
  axis.text = element_text(size=8),
    legend.text = element_text(size=8)

 )
```



```
#add information on chart
print(paste('Out of',toString(num_total),'known,'
            ,toString(num_same_age),
            'shootings were committed within the same age group which is',
            percentage,'%.'))
```

```
## [1] "Out of 14177 known, 6782 shootings were committed within the same age group which is 47.84 %."
```

Analysis of Age Material

The shooting data revealed many interesting trends regarding the age of the victims and perpetrators. The majority of shooting victim's where age 25-44 and the 18-24 group were second most common. Victim's age 65 and above were the least frequent in this data. Most of the known perpetrators were in the age group 18-25. The second most common was age 25-44. As with victim's age the least frequent known age was 65 and above. Additionally, it was interesting to see that shootings that involved different age groups (for example a 18-24 year old victim and a 25-44 perpetrator) made up more than half (52.16%) of all shootings.

These findings do raise additional questions. Would grouping the age data with other socio-economic factors help establish further trends regarding the similarities and differences between victims and perpetrators? Also, if the age groups were more specific, what trends would emerge?

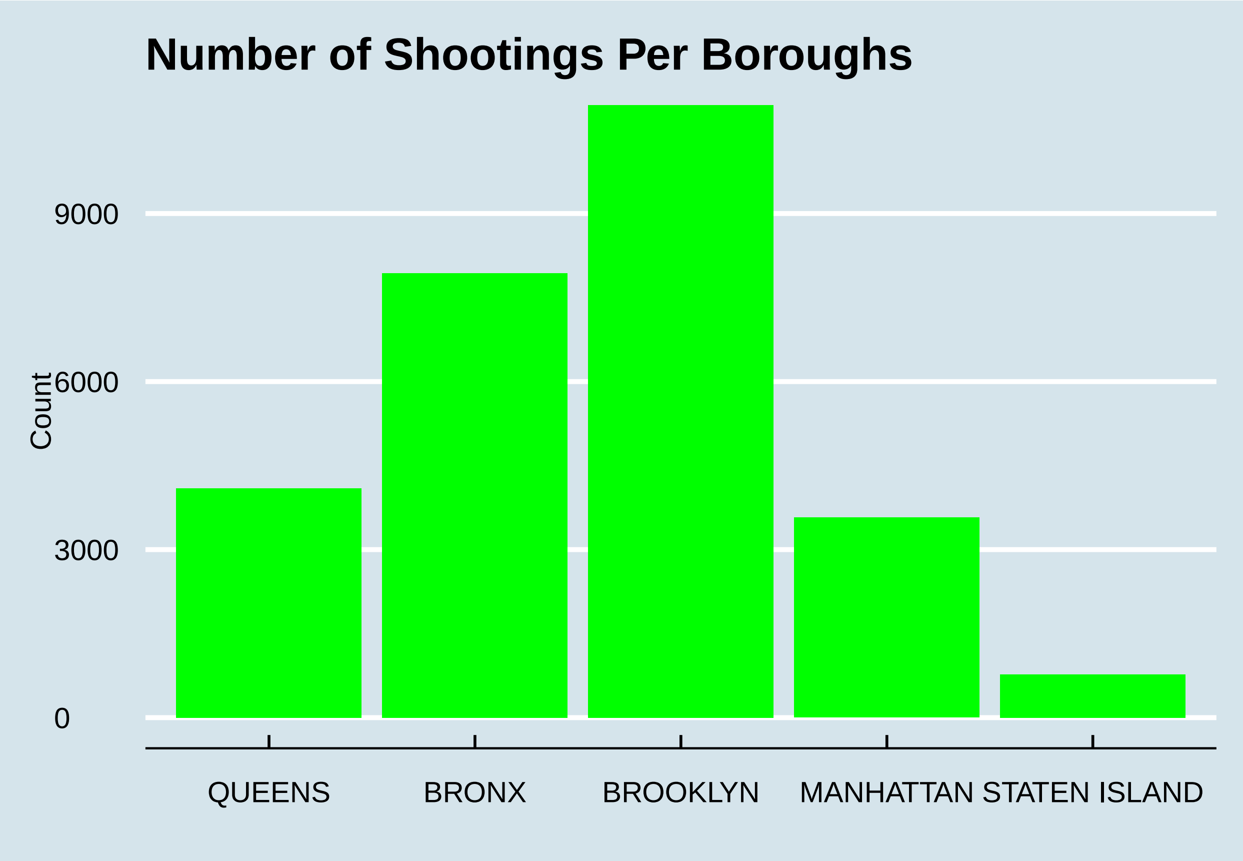
Geographic Analysis

Given the Borough classifications and the Longitude and Latitude information a geographical analysis is also possible.

```
#groups by Boro and summarizes
shooting_data_by_boro <- shooting_data %>% group_by(BORO)
summarize(shooting_data_by_boro, count=n())
```

```
## # A tibble: 5 x 2
##   BORO          count
##   <fct>         <int>
## 1 QUEENS         4094
## 2 BRONX          7937
## 3 BROOKLYN      10933
## 4 MANHATTAN      3572
## 5 STATEN ISLAND   776
```

```
#make bar chart
ggplot(data=shooting_data,aes(x=BORO)) +
  geom_bar(fill='green') +
  labs(title='Number of Shootings Per Boroughs',x='',y='Count')
```



Mapping

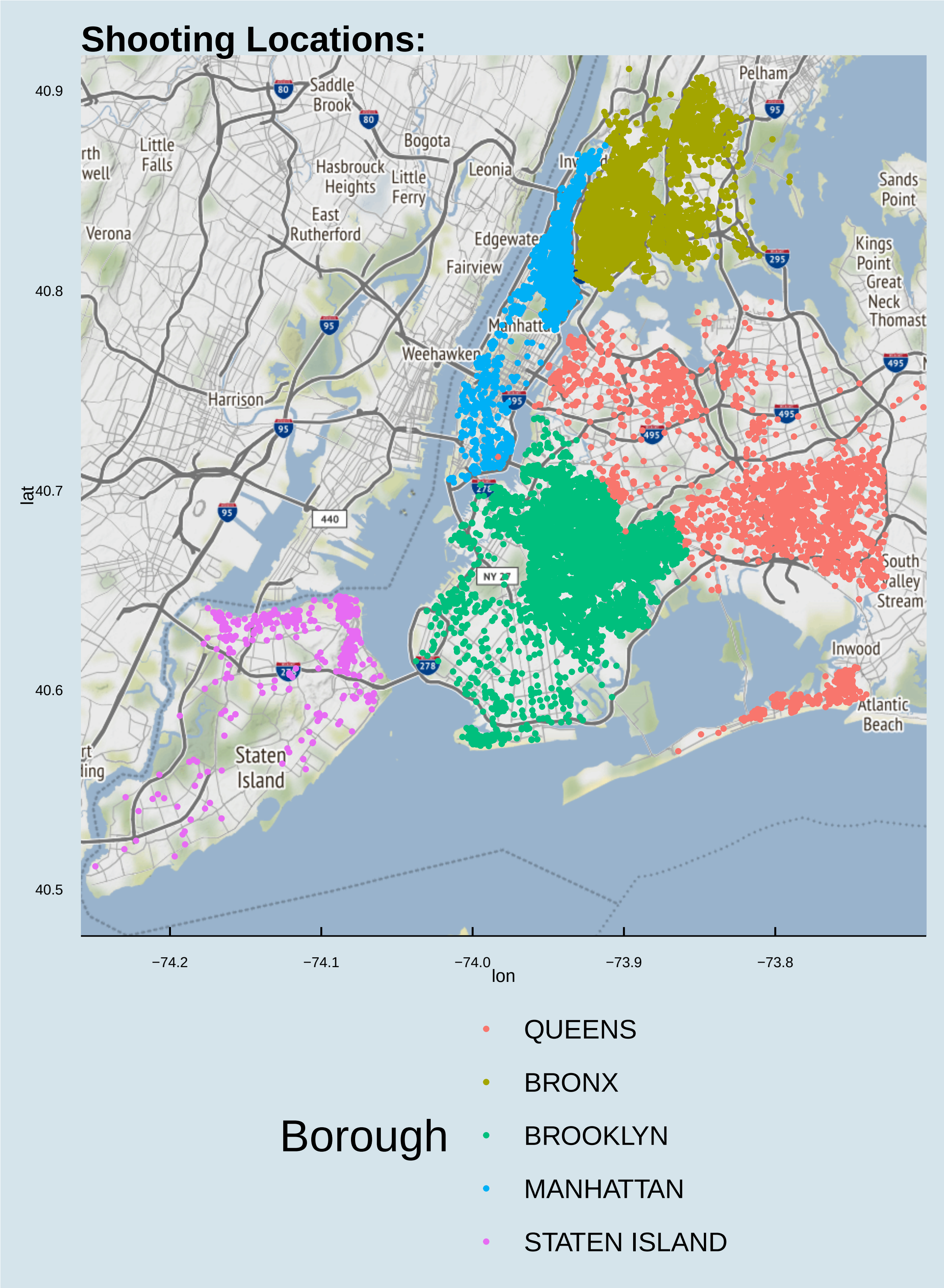
Using Open Source Maps, Stamen Tiles, and the ggmap library, the latitude and longitude variables can be plotted to show the geographical spacing of the shootings. In the Shooting Locations figure that follows each shooting is indicated by a dot and color coded according to borough.

```
#filter locations with no longitude or latitude
shooting_data_ll <- shooting_data %>% filter(!is.na(Longitude),!is.na(Latitude))

#Get map using Open Source Maps and Stamen Tiles
ny_map <- get_map( getbb('New York City, New York'), source="stamen")
```

```
## i Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.
```

```
ggmap(ny_map) +
  geom_point(data = shooting_data_ll,
             aes(x = Longitude, y = Latitude, color=BORO), size = 1)  +
  theme(legend.position='bottom',
            legend.text = element_text(size=15),
            legend.title = element_text(size=25),
              axis.title = element_text(size=10),
        axis.text = element_text(size=8),
plot.title = element_text(size=20)
        ) +
```

```r
  guides(color=guide_legend(nrow=5, byrow=TRUE)) +
  labs(title="Shooting Locations:",color="Borough")
```


Shooting Locations:
40.9
40.8
40.7
40.6
40.5
lat
−74.2
−74.1
−74.0
−73.9
−73.8
lon
Saddle Brook
Bogota
Little Falls
Hasbrouck Heights
Little Ferry
Leonia
Verona
East Rutherford
Edgewater
Fairview
Manhattan
Weehawken
Harrison
Pelham
Sands Point
Kings Point
Great Neck
Thomast
South Valley Stream
Inwood
Atlantic Beach
Staten Island
Borough
● QUEENS
● BRONX
● BROOKLYN
● MANHATTAN
● STATEN ISLAND

Geographical Analysis

The Geographical data included in the NYPD Shooting Data reveals several interesting trends. According to the 2010 census[1] the population of the boroughs can be seen as follows:

Bronx	Brooklyn	Manhattan	Queens	Staten Island
1,385,108	2,504,700	1,585,873	2,230,722	468,730
16.9%	30.6%	19.4%	27.3%	5.7%

Now let's add the shooting data by borough we computed earlier to see if any borough has a higher than expected shooting rate.

Bronx	Brooklyn	Manhattan	Queens	Staten Island
Population:				
1,385,108	2,504,700	1,585,873	2,230,722	468,730
16.9%	30.6%	19.4%	27.3%	5.7%
Shootings:				
7937	10933	3572	4094	776
29.06%	40.03%	13.08%	14.99%	2.84%

Here we can see that the Bronx has 29% of shootings but only 17% of the popultion making it the most shootings per capita of the five boroughs. Whereas Staten Island (and Queens a close second) has the lowest rate of shootings per capita with 5.7% of the population but only 2.84% of shootings. These densities can be clearly seen in the map above.

Date Model

The following model shows the yearly trend of shootings based on the month. A scatter plot is shown with each dot representing the total shootings for a month of a particular year, a linear model (shown in red) is computed to predict the total number of shootings given the month of the year.

```r
#first we create a new month column
sd_by_date <- shooting_data
sd_by_date$month <-  month(sd_by_date$OCCUR_DATE)
sd_by_date$year <-  year(sd_by_date$OCCUR_DATE)

#now we summarize based on month and year
sd_by_date <- sd_by_date %>% group_by(month, year)
months_count <- sd_by_date %>% summarise(count = n())
```

```
## ‘summarise()‘ has grouped output by ’month’. You can override using the
## ‘.groups‘ argument.
```

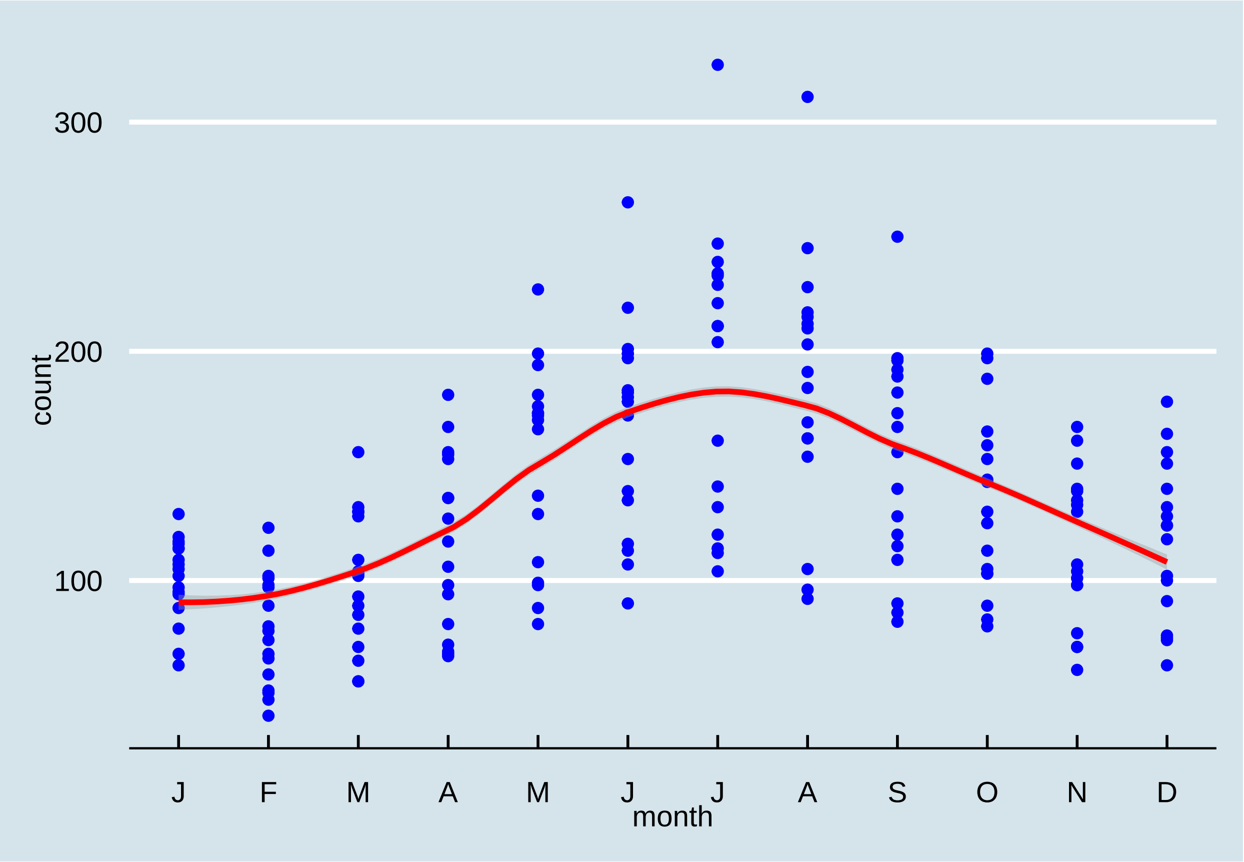
```r
#we create a model - includes polynomial to accomidate curved data
mod <- lm(count ~ month +  I(month^2) + I(month^3) + I(month^4), data = months_count)
#show a summary of the model's performance
summary(mod)
```

[1] Dept of City Planning, "New York City Population Projections by Age/Sex & Borough 2000-2030 . . . " NYC.gov. Accessed June 30, 2023. https://www.nyc.gov/assets/planning/download/pdf/data-maps/nyc-population/projections_report.pdf.

```
##
## Call:
## lm(formula = count ~ month + I(month^2) + I(month^3) + I(month^4),
##     data = months_count)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -85.453 -29.385   5.246  26.534 141.279
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  181.37998   25.75376   7.043 3.00e-11 ***
## month       -116.87784   25.04055  -4.668 5.59e-06 ***
## I(month^2)    41.99264    7.41350   5.664 5.11e-08 ***
## I(month^3)    -4.84117    0.84112  -5.756 3.23e-08 ***
## I(month^4)     0.17633    0.03216   5.483 1.26e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 40.68 on 199 degrees of freedom
## Multiple R-squared:  0.4107, Adjusted R-squared:  0.3989
## F-statistic: 34.67 on 4 and 199 DF,  p-value: < 2.2e-16
```

```
#let's add a new column to show prediction based on month
months_count <- months_count %>% ungroup() %>% mutate(pred = predict(mod))

#now we chart the data according to the new grouping
my_labels <- c('J','F','M','A','M','J','J','A','S','O','N','D')
ggplot(data=months_count) + geom_point(aes(x = month, y=count),color='blue') +
  geom_smooth(aes(x=month, y=pred),color='red') +
  scale_x_continuous(breaks=seq(1,12,1),labels=my_labels)
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Conclusion

In this report we have looked at the relationship between age and geographical location in the NYPD Shooting Dataset. The results showed that shooting victims are most likely to be 25-44 and perpetrators 18-24. Also, shootings were virtually just as likely to be in the same age group as not. Next, we pinpointed the Bronx as having the most shootings per capita and Staten Island as having the least. And finally our linear model predicts that July will have the most shootings.

Potential Bias

Regarding the question of bias in the data, there is possible bias in its collection as well as my analysis. The identification of race is never as straightforward as black vs white vs Hispanic, much of the population does not fall completely into one category. How this was collected and reported would require further investigtion. My own bias shows in choosing age and geography as determining factors. I expected that certain age groups would be more likely to be involved than others, and certain geographic regions as well. This proved to be the case according to my analysis. However, no age or geographical data was given preferential treatment which would have lead to these results.