NYPD Shooting Incident Data Report

Ryan Masad

2023-09-30

NYPD Shooting Incident Data Report

The NYPD Shooting Incident Data provides insight into cases of shooting since 2006. The data set provides information on when the event took place, where it took place, whether it resulted in a death, and any information that is available about the perpetrator and the victim. The question that I want to dive into is about how shootings and murders correlate within various boroughs.

Loading the Data

Start by loading the data from the New York shooting report and examining the data for any immediate fixes.

```
url_file = "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
ny_shooting = read.csv(url_file)
ny_shooting %>% head(5)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME   BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    228798151 05/27/2021   21:30:00 QUEENS                       105
## 2    137471050 06/27/2014   17:40:00  BRONX                        40
## 3    147998800 11/21/2015   03:56:00 QUEENS                       108
## 4    146837977 10/09/2015   18:30:00  BRONX                        44
## 5     58921844 02/19/2009   22:58:00  BRONX                        47
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0                                                   false
## 2                 0                                                   false
## 3                 0                                                    true
## 4                 0                                                   false
## 5                 0                                                    true
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                                          18-24       M          BLACK
## 2                                          18-24       M          BLACK
## 3                                          25-44       M          WHITE
## 4                                            <18       M WHITE HISPANIC
## 5          25-44        M     BLACK         45-64       M          BLACK
##   X_COORD_CD Y_COORD_CD Latitude Longitude
## 1    1058925   180924.0 40.66296 -73.73084
## 2    1005028   234516.0 40.81035 -73.92494
## 3    1007668   209836.5 40.74261 -73.91549
## 4    1006537   244511.1 40.83778 -73.91946
## 5    1024922   262189.4 40.88624 -73.85291
```

```
##                                       Lon_Lat
## 1 POINT (-73.73083868899994 40.662964620000025)
## 2  POINT (-73.92494232599995 40.81035186300006)
## 3  POINT (-73.91549174199997 40.74260663300004)
## 4  POINT (-73.91945661499994 40.83778200300003)
## 5  POINT (-73.85290950899997 40.88623791800006)
```

```
summary(ny_shooting)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class :character   Class :character
##  Median : 90372218   Mode  :character   Mode  :character   Mode  :character
##  Mean   :120860536
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Length:27312            Length:27312
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  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:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000029   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude         Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
```

```
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

Cleaning the Data

After initial examination of the above data the following initial cleaning were identified.

1. Convert Date to a Date type variable
2. Create a column with Year, Month, and Monthly Date
    a. This is needed because there isn't a shooting every day. So tracking trends over time seems more useful.
3. Convert STATISTICAL_MURDER_FLAG into a binary variable so we can sum it up to see how many murders occurred
4. Clean up the data set and remove unneeded columns. INCIDENT_KEY, TIME, PRECINCT, and JURISDICTION CODE do not seem necessary

```
ny_shooting = ny_shooting %>%
  mutate(OCCUR_DATE = as.Date(OCCUR_DATE, format = "%m/%d/%Y")) %>%
  mutate(YEAR = year(OCCUR_DATE)) %>%
  mutate(MONTH = month(OCCUR_DATE)) %>%
  mutate(MONTHLY_DATE = floor_date(OCCUR_DATE, unit="month")) %>%
  mutate(STATISTICAL_MURDER_FLAG = ifelse(STATISTICAL_MURDER_FLAG == "true",1,0)) %>%
  select(OCCUR_DATE, MONTHLY_DATE, YEAR, MONTH, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC:VIC_RACE)

ny_shooting %>% head(5)
```

```
##   OCCUR_DATE MONTHLY_DATE YEAR MONTH   BORO LOC_OF_OCCUR_DESC
## 1 2021-05-27   2021-05-01 2021     5 QUEENS
## 2 2014-06-27   2014-06-01 2014     6  BRONX
## 3 2015-11-21   2015-11-01 2015    11 QUEENS
## 4 2015-10-09   2015-10-01 2015    10  BRONX
## 5 2009-02-19   2009-02-01 2009     2  BRONX
##   LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
## 1                                                        0
## 2                                                        0
## 3                                                        1
## 4                                                        0
## 5                                                        1          25-44
##   PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                           18-24       M          BLACK
## 2                           18-24       M          BLACK
## 3                           25-44       M          WHITE
## 4                             <18       M WHITE HISPANIC
## 5        M     BLACK         45-64       M          BLACK
```

Population Data Import

After Cleaning the data, Population seems like it could aid in analyzing the following data. So I found a population file off the same site. I chose to average population for 2010 and 2020 as there weren't records for every year and I think the average will give a decent idea of the population. Using just 2020 or 2010 would be valid as well.

```
boro_file = "https://data.cityofnewyork.us/resource/xywu-7bv9.csv"
boro_data = read.csv(boro_file)

boro_data = boro_data %>%
  mutate(borough = trimws(str_to_upper(borough))) %>%
  mutate(population = round((X_2020+X_2010)/2,0)) %>%
  select(borough,population)
boro_data
```

```
##         borough population
## 1     NYC TOTAL    8396798
## 2         BRONX    1415948
## 3      BROOKLYN    2600682
## 4     MANHATTAN    1612077
## 5        QUEENS    2290148
## 6 STATEN ISLAND     477942
```

Join the Data

Here I am joining in the Population data to the NY_shooting database.

```
ny_shooting = left_join(ny_shooting, boro_data, by = join_by("BORO" == "borough"))
ny_shooting %>% head(5)
```

```
##   OCCUR_DATE MONTHLY_DATE YEAR MONTH   BORO LOC_OF_OCCUR_DESC
## 1 2021-05-27   2021-05-01 2021     5 QUEENS
## 2 2014-06-27   2014-06-01 2014     6  BRONX
## 3 2015-11-21   2015-11-01 2015    11 QUEENS
## 4 2015-10-09   2015-10-01 2015    10  BRONX
## 5 2009-02-19   2009-02-01 2009     2  BRONX
##   LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
## 1                                                        0
## 2                                                        0
## 3                                                        1
## 4                                                        0
## 5                                                        1          25-44
##   PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE population
## 1                           18-24       M          BLACK    2290148
## 2                           18-24       M          BLACK    1415948
## 3                           25-44       M          WHITE    2290148
## 4                             <18       M WHITE HISPANIC    1415948
## 5        M     BLACK         45-64       M          BLACK    1415948
```

New York Shootings by Month

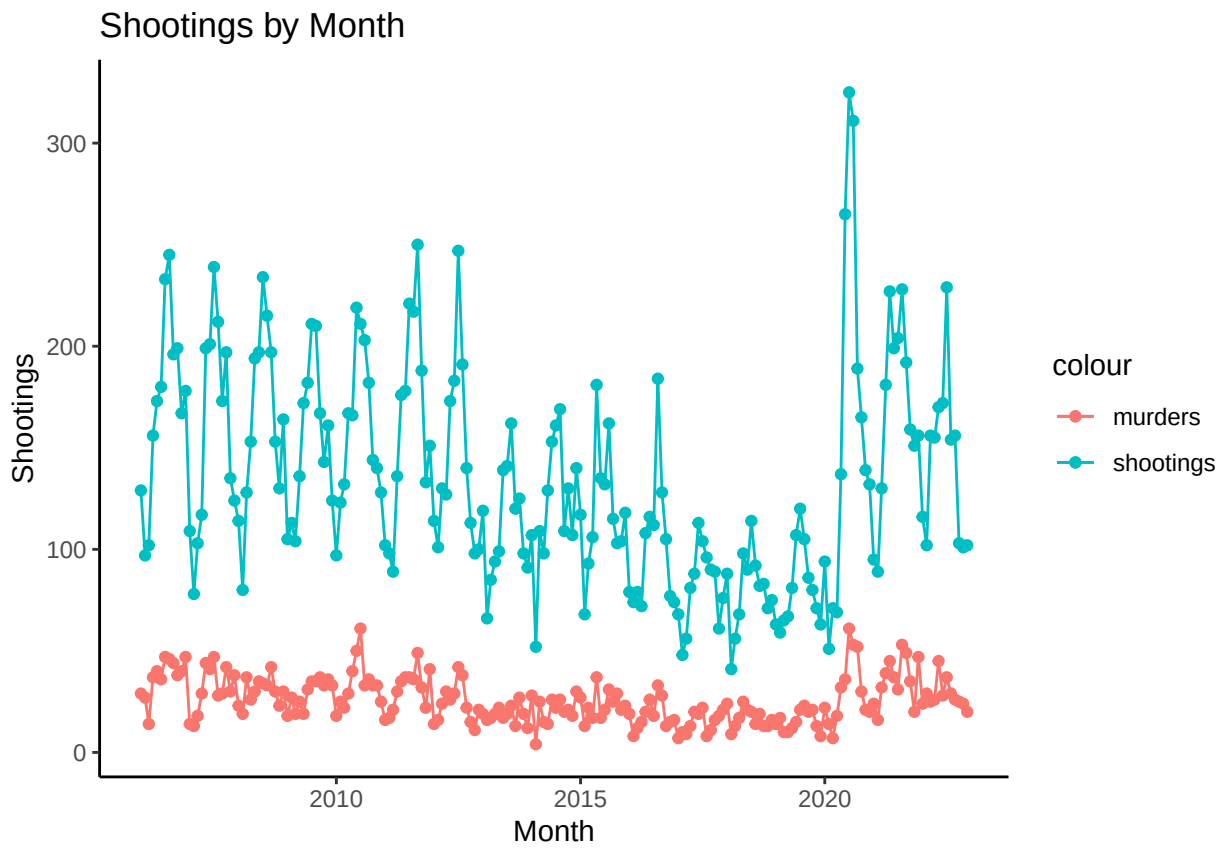
Starting off with a simple analysis, I wanted to see how shootings and murders trended for the entirety of New York. I am looking at the data by month.The first thing that stands out to me, is the seasonal nature to shootings. The data spikes in some seasons and then drops. It also looks like there was a trend toward less shooting up until 2020. Perhaps Covid had a severe effect on crime in New York.

```
shootings_by_month = ny_shooting %>%
  group_by(MONTHLY_DATE,MONTH) %>%
  summarise(shootings = length(OCCUR_DATE), murders = sum(STATISTICAL_MURDER_FLAG))
```

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

```
shootings_by_month
```

```
## # A tibble: 204 x 4
## # Groups:   MONTHLY_DATE [204]
##    MONTHLY_DATE MONTH shootings murders
##    <date>       <dbl>     <int>   <dbl>
##  1 2006-01-01       1       129      29
##  2 2006-02-01       2        97      27
##  3 2006-03-01       3       102      14
##  4 2006-04-01       4       156      37
##  5 2006-05-01       5       173      40
##  6 2006-06-01       6       180      36
##  7 2006-07-01       7       233      47
##  8 2006-08-01       8       245      46
##  9 2006-09-01       9       196      44
## 10 2006-10-01      10       199      38
## # i 194 more rows
```

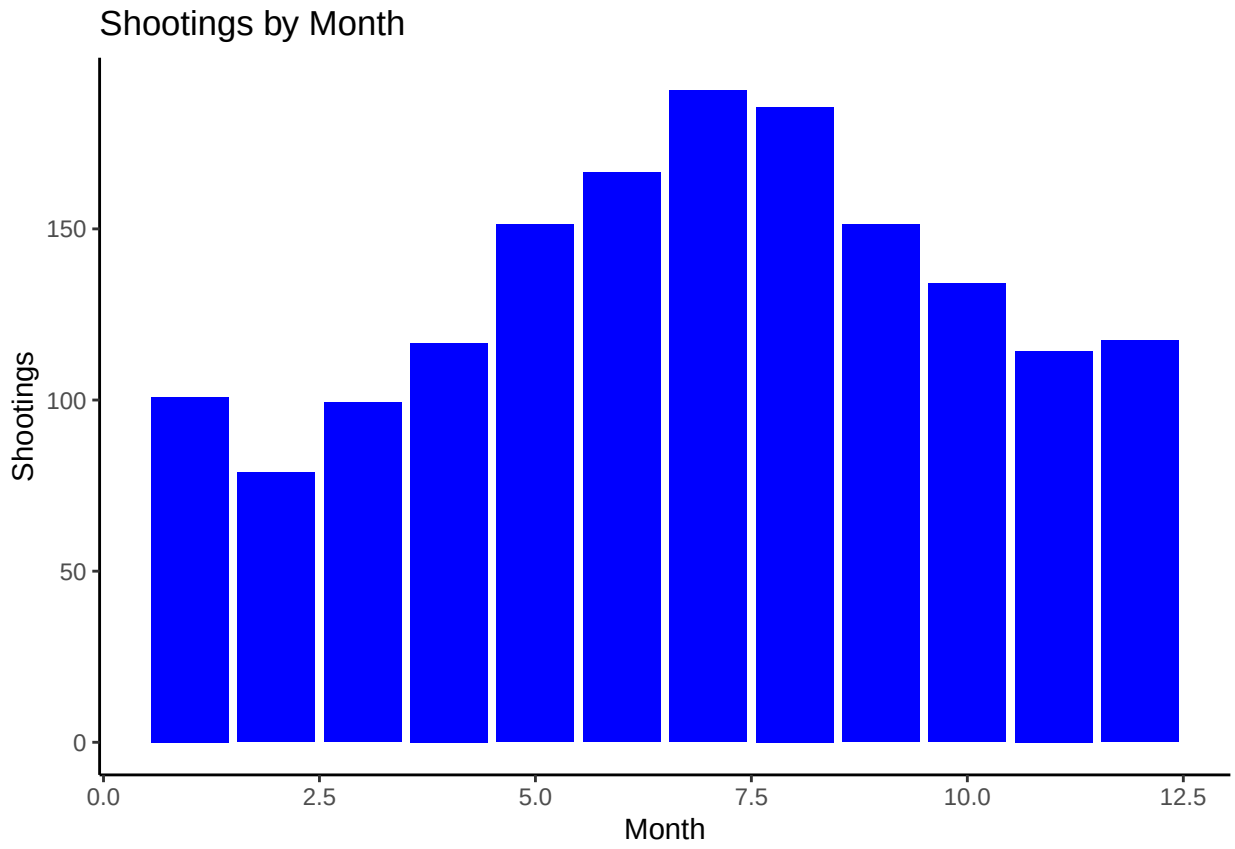


Diving Deeper into Shootings By Month

Now i want to look at the average # of shootings over the last 15 years by month. When diving into this, you can see a peak during the summer and it dropping during the winter. This makes sense as people tend to stay home when it snows likely resulting in lower crime rates.

```
shootings_by_month_no_year = shootings_by_month %>%
  group_by(MONTH) %>%
  summarise(shootings = mean(shootings), murders = mean(murders))
shootings_by_month_no_year
```

```
## # A tibble: 12 x 3
##    MONTH shootings murders
##    <dbl>     <dbl>   <dbl>
##  1     1     101.     20
##  2     2      78.8    16.5
##  3     3      99.3    19.2
##  4     4     117.     23.2
##  5     5     151.     29.8
##  6     6     166.     29.5
##  7     7     190.     33.8
##  8     8     186.     31.9
##  9     9     151.     30.7
## 10    10     134.     26.6
## 11    11     114.     22.2
## 12    12     117.     26.4
```



Shootings by Boro

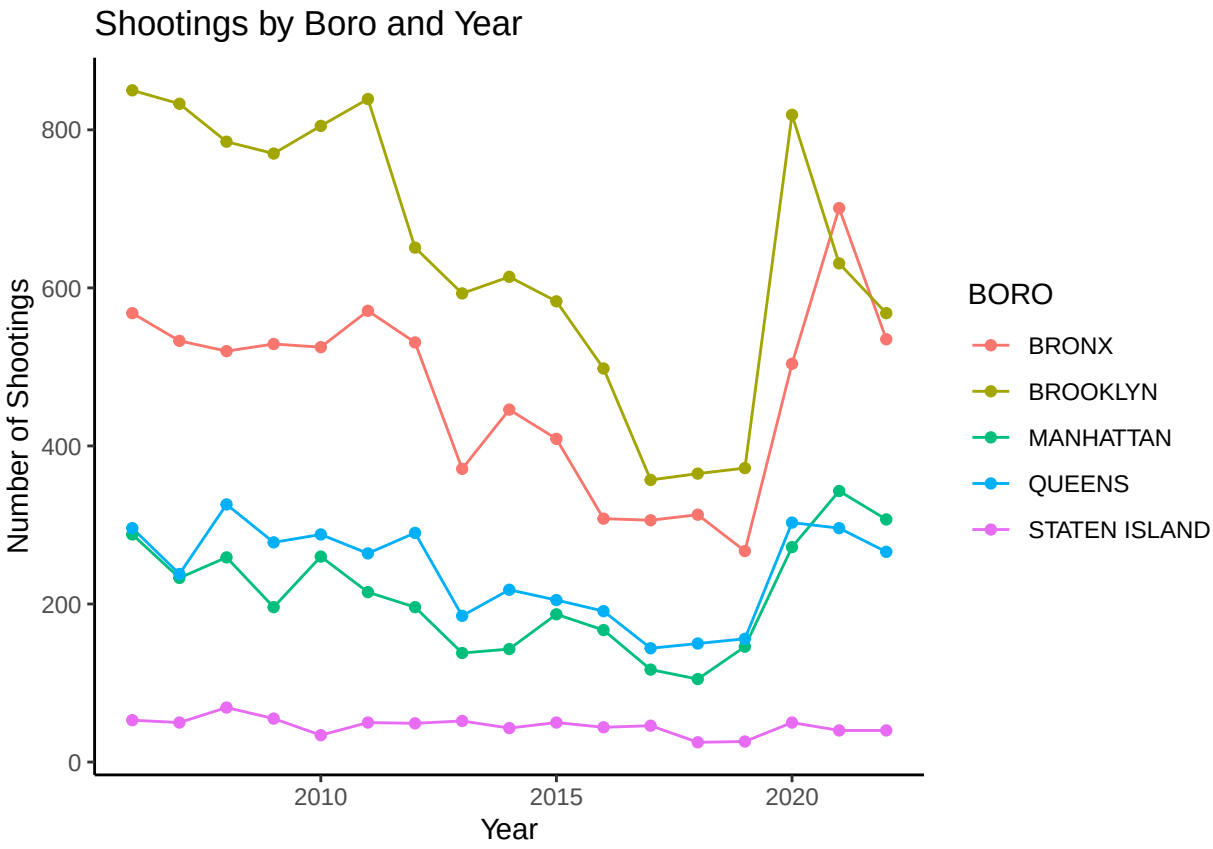
Next I want to look at Shootings by Year by BORO to focus more on the decline in shootings over the years followed by the spike. I wanted to see how boroughs see crime rates differently. I start by graphing shootings and shootings per million by boroughs. This shows that while Brooklyn has the most shootings, Bronx has the highest shootings per million people. The other 3 boroughs all have similar shootings per million. Murders show similar results.

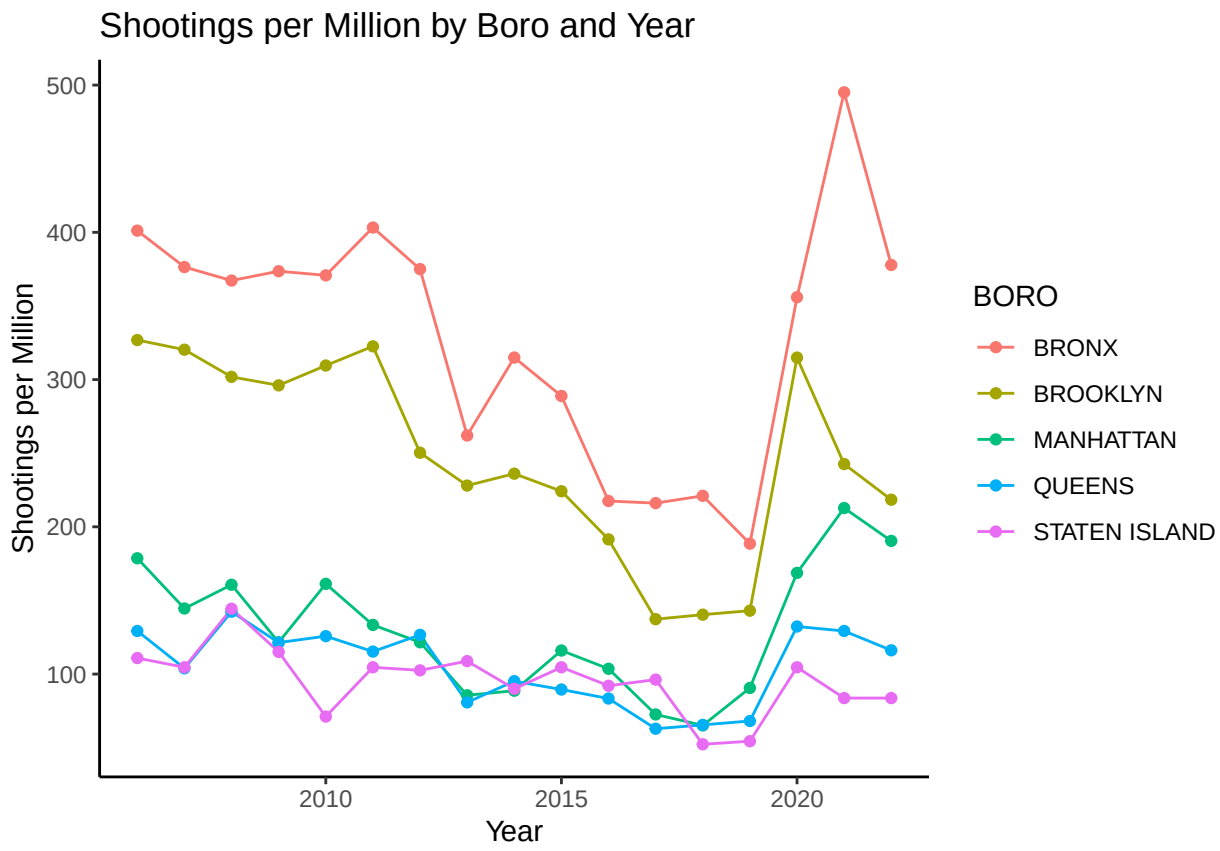
```
shootings_by_boro = ny_shooting %>%
  group_by(BORO, YEAR) %>%
  summarise(shootings = length(OCCUR_DATE), murders = sum(STATISTICAL_MURDER_FLAG), population = max(po
  mutate(shootings_per_pop = shootings/population*1000000) %>%
  mutate(murders_per_pop = murders/population*1000000)
```

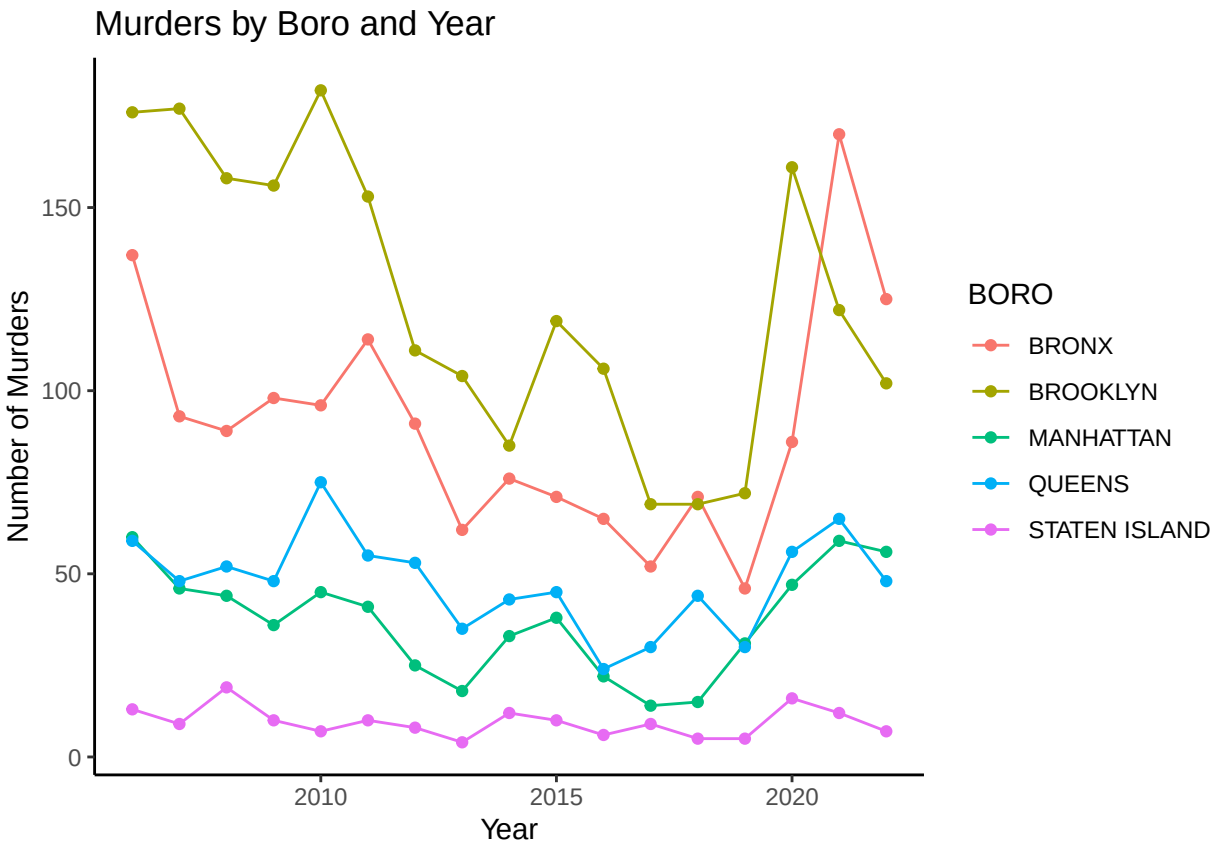
```
## `summarise()` has grouped output by 'BORO'. You can override using the
## `.groups` argument.
```

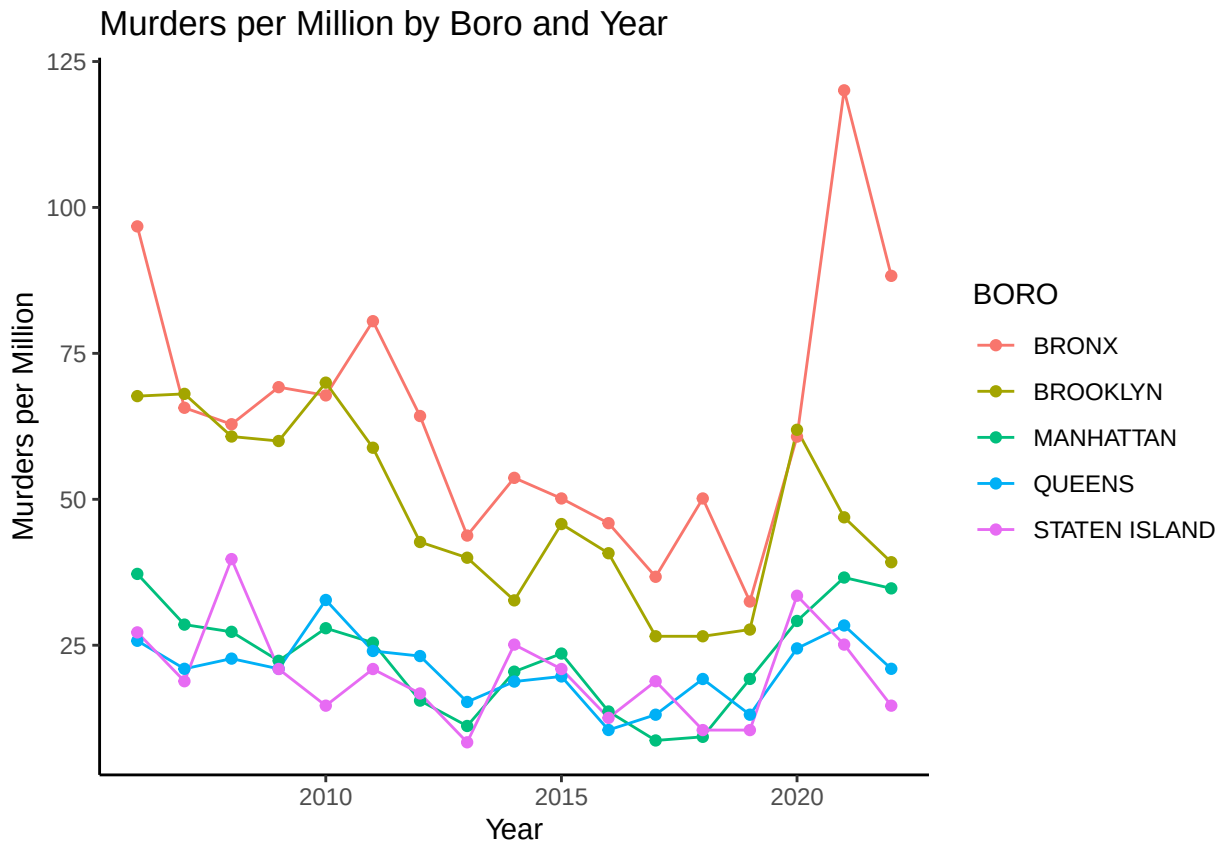
```
shootings_by_boro %>% head(5)
```

```
## # A tibble: 5 x 7
## # Groups:   BORO [1]
##   BORO   YEAR shootings murders population shootings_per_pop murders_per_pop
##   <chr> <dbl>     <int>   <dbl>      <dbl>             <dbl>           <dbl>
## 1 BRONX  2006       568     137    1415948              401.            96.8
## 2 BRONX  2007       533      93    1415948              376.            65.7
## 3 BRONX  2008       520      89    1415948              367.            62.9
## 4 BRONX  2009       529      98    1415948              374.            69.2
## 5 BRONX  2010       525      96    1415948              371.            67.8
```


Shootings by Boro and Year
Number of Shootings
800
600
400
200
0
2010
2015
2020
Year
BORO
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND


Shootings per Million by Boro and Year
Shootings per Million
500
400
300
200
100
2010
2015
2020
Year
BORO
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND


Murders by Boro and Year
Number of Murders
150
100
50
0
2010
2015
2020
Year
BORO
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND



Linear Model of Shootings against Murders

Now i want to verify that murders per million people is a function of Shootings per million people. We also found out that there is high correlation between shootings and murders with R^2 being ~ 97%.

```
model = lm(shootings_per_pop ~ murders_per_pop + factor(BORO) + factor(YEAR), data = shootings_by_boro)
summary(model)
```

```
##
## Call:
## lm(formula = shootings_per_pop ~ murders_per_pop + factor(BORO) +
##     factor(YEAR), data = shootings_by_boro)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.730 -13.170  -0.777  13.038  42.236
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 143.3173    23.4883   6.102 7.12e-08 ***
## murders_per_pop               2.9335     0.2624  11.181  < 2e-16 ***
## factor(BORO)BROOKLYN        -35.3206     8.4628  -4.174 9.37e-05 ***
## factor(BORO)MANHATTAN       -78.9078    13.0373  -6.052 8.64e-08 ***
## factor(BORO)QUEENS          -97.6873    13.5145  -7.228 7.99e-10 ***
## factor(BORO)STATEN ISLAND -104.7933    13.7061  -7.646 1.49e-10 ***
```

```
## factor(YEAR)2007            11.4340    13.8100   0.828   0.4108
## factor(YEAR)2008            18.1393    13.7040   1.324   0.1904
## factor(YEAR)2009            12.0969    13.9078   0.870   0.3877
## factor(YEAR)2010             2.7027    13.7062   0.197   0.8443
## factor(YEAR)2011            12.8121    13.7355   0.933   0.3545
## factor(YEAR)2012            19.9948    14.3721   1.391   0.1691
## factor(YEAR)2013             3.4890    15.2988   0.228   0.8203
## factor(YEAR)2014            -3.4099    14.5889  -0.234   0.8160
## factor(YEAR)2015            -9.2427    14.4131  -0.641   0.5237
## factor(YEAR)2016           -14.7283    15.1841  -0.970   0.3358
## factor(YEAR)2017           -23.8997    15.6741  -1.525   0.1323
## factor(YEAR)2018           -38.9611    15.3714  -2.535   0.0138 *
## factor(YEAR)2019           -31.4457    15.6982  -2.003   0.0495 *
## factor(YEAR)2020            12.2798    13.7354   0.894   0.3747
## factor(YEAR)2021             1.8907    13.5326   0.140   0.8893
## factor(YEAR)2022             1.2546    13.8559   0.091   0.9281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.4 on 63 degrees of freedom
## Multiple R-squared:  0.9696, Adjusted R-squared:  0.9595
## F-statistic: 95.81 on 21 and 63 DF,  p-value: < 2.2e-16
```

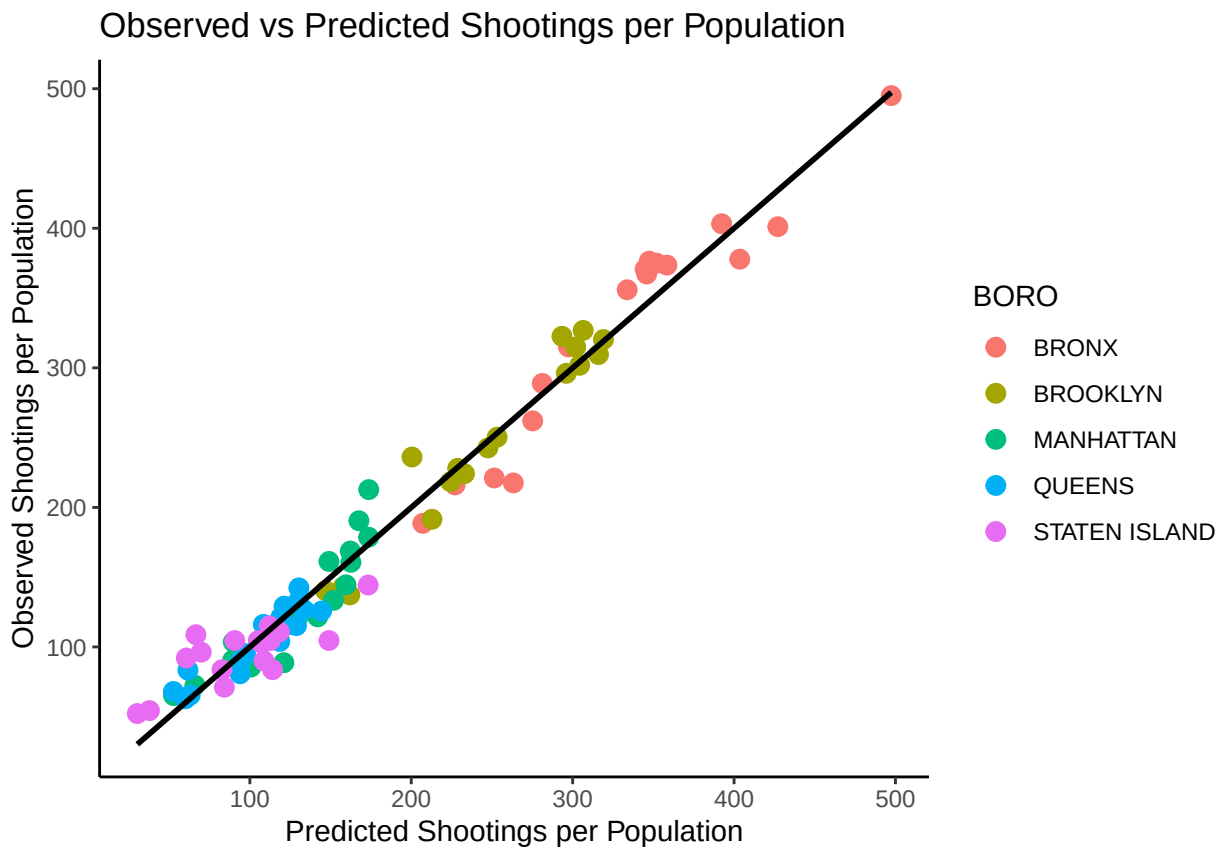
```
predictions = predict(model, newdata = shootings_by_boro)

shootings_by_boro$predicted_shootings_per_pop = predictions
shootings_by_boro %>% head(5)
```

```
## # A tibble: 5 x 8
## # Groups:   BORO [1]
##   BORO   YEAR shootings murders population shootings_per_pop murders_per_pop
##   <chr> <dbl>     <int>   <dbl>      <dbl>             <dbl>           <dbl>
## 1 BRONX  2006       568     137    1415948              401.            96.8
## 2 BRONX  2007       533      93    1415948              376.            65.7
## 3 BRONX  2008       520      89    1415948              367.            62.9
## 4 BRONX  2009       529      98    1415948              374.            69.2
## 5 BRONX  2010       525      96    1415948              371.            67.8
## # i 1 more variable: predicted_shootings_per_pop <dbl>
```

Finally let's plot this to show a strong correlation between the predictions and the actual values.

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



Continuation.

Now that we know murders and shootings share similar rates across boroughs, I want to dive deeper into the victims, by looking at Race and Sex involved in shootings by borough.

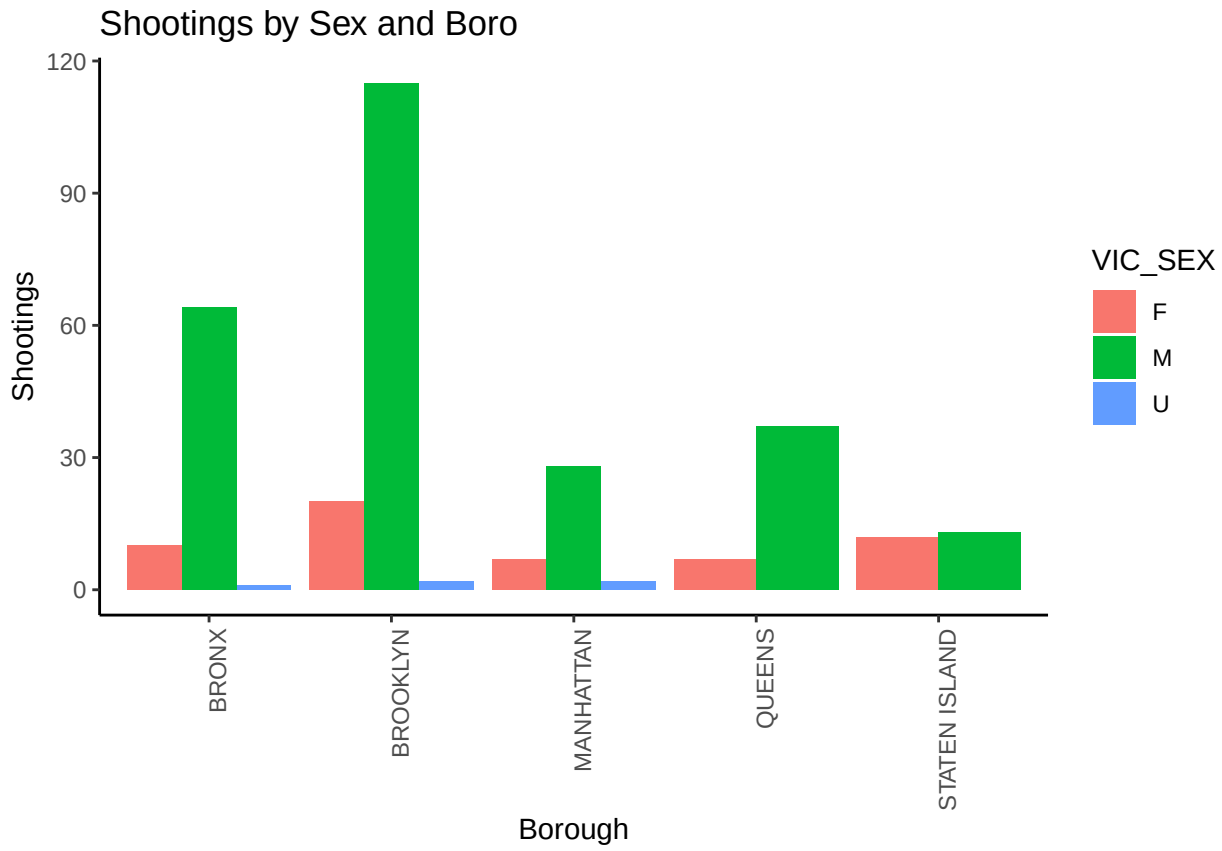
```
shootings_by_victim = ny_shooting %>%
  group_by(BORO,YEAR,MONTHLY_DATE,MONTH,VIC_SEX, VIC_RACE) %>%
  summarise(shootings = length(OCCUR_DATE), murders = sum(STATISTICAL_MURDER_FLAG))
```

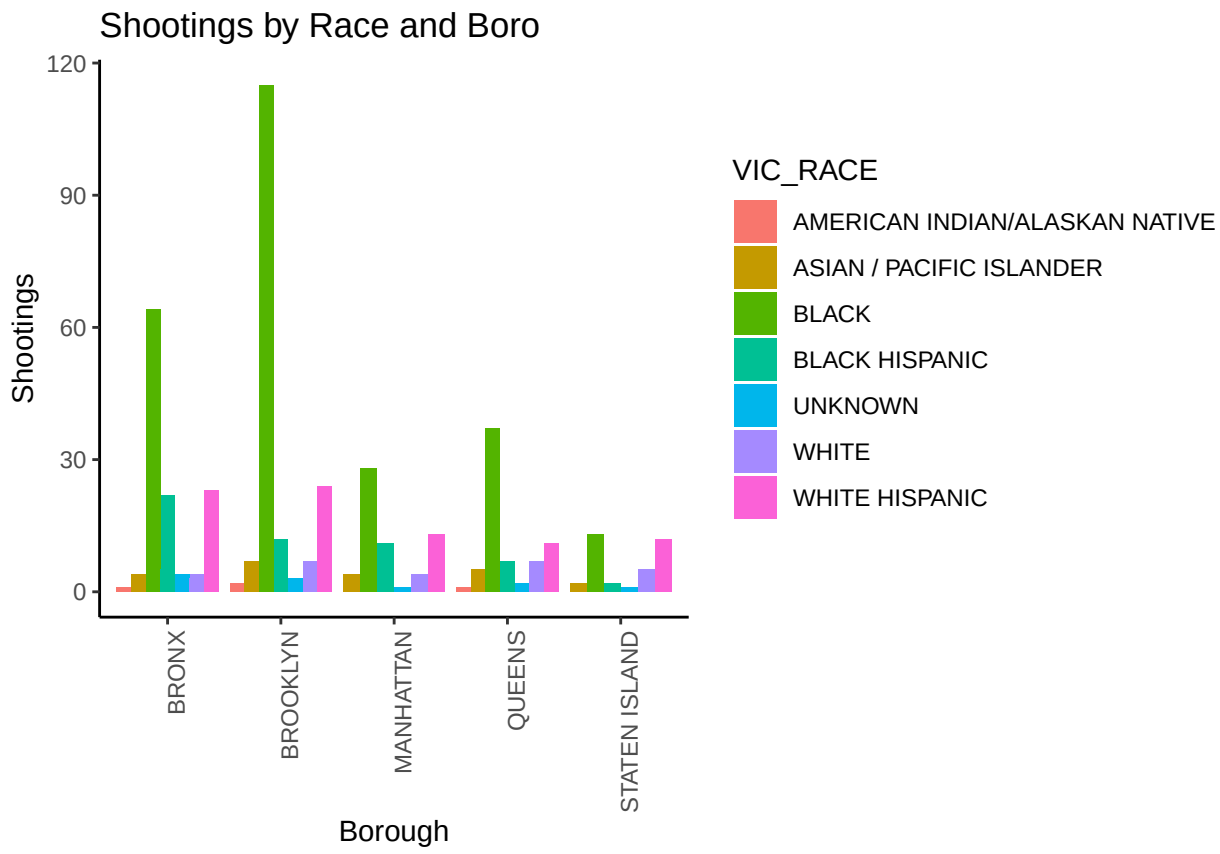
```
## `summarise()` has grouped output by 'BORO', 'YEAR', 'MONTHLY_DATE', 'MONTH',
## 'VIC_SEX'. You can override using the `.groups` argument.
```

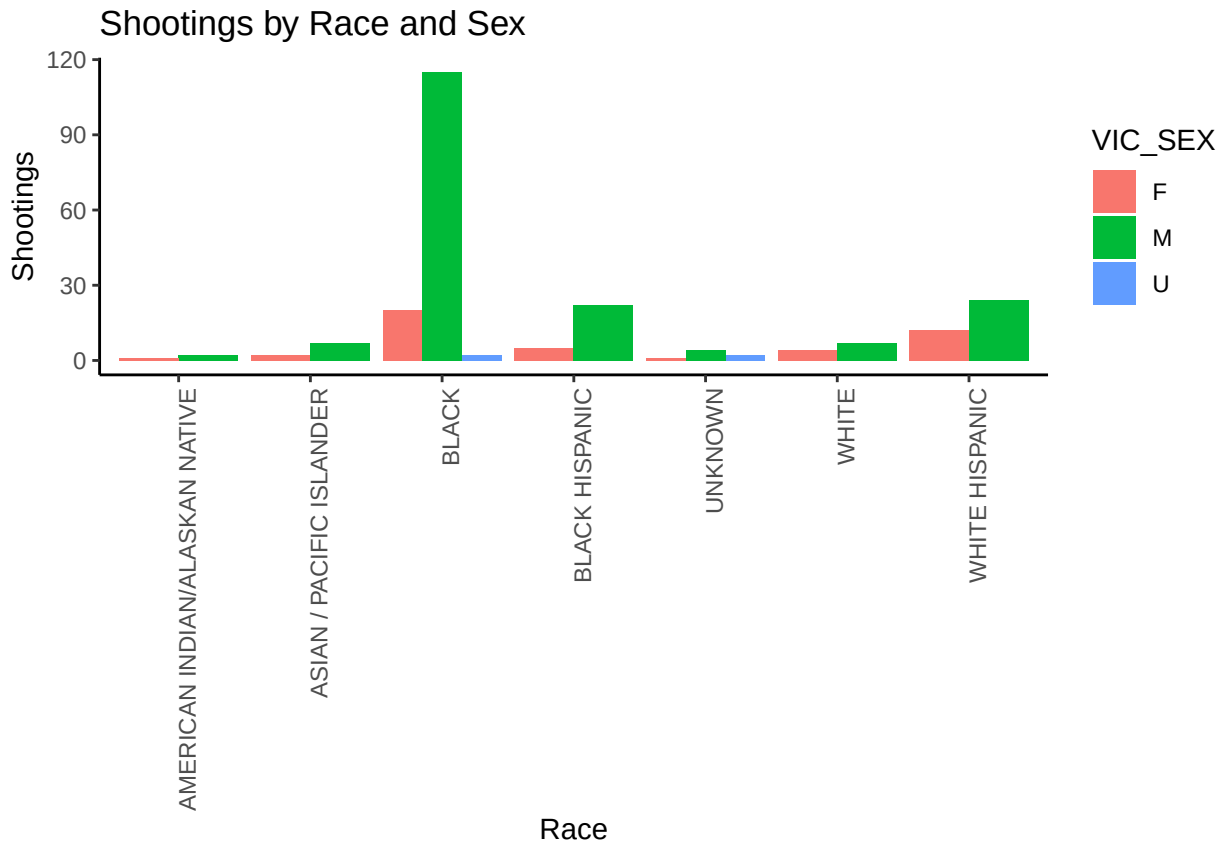
```
shootings_by_victim
```

```
## # A tibble: 4,226 x 8
## # Groups:   BORO, YEAR, MONTHLY_DATE, MONTH, VIC_SEX [1,736]
##    BORO   YEAR MONTHLY_DATE MONTH VIC_SEX VIC_RACE             shootings murders
##    <chr> <dbl> <date>       <dbl> <chr>   <chr>                    <int>   <dbl>
##  1 BRONX  2006 2006-01-01       1 F       BLACK                        1       0
##  2 BRONX  2006 2006-01-01       1 F       WHITE HISPANIC               1       1
##  3 BRONX  2006 2006-01-01       1 M       BLACK                       22       5
##  4 BRONX  2006 2006-01-01       1 M       BLACK HISPANIC               4       0
##  5 BRONX  2006 2006-01-01       1 M       WHITE                        1       1
##  6 BRONX  2006 2006-01-01       1 M       WHITE HISPANIC              11       2
```

```
##  7 BRONX  2006 2006-02-01       2 F       WHITE HISPANIC              2       1
##  8 BRONX  2006 2006-02-01       2 M       ASIAN / PACIFIC ISL~        2       0
##  9 BRONX  2006 2006-02-01       2 M       BLACK                       6       2
## 10 BRONX  2006 2006-02-01       2 M       BLACK HISPANIC              2       0
## # i 4,216 more rows
```




Shootings by Race and Boro
120
90
60
30
0
Shootings
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND
Borough
VIC_RACE
AMERICAN INDIAN/ALASKAN NATIVE
ASIAN / PACIFIC ISLANDER
BLACK
BLACK HISPANIC
UNKNOWN
WHITE
WHITE HISPANIC



Conclusion

We have seen heavy correlation between shooting and murders and strong seasonality with the data.While Male victims occur much more frequently, Female victims nearly match male victims in Staten Island, which posted the lowest number of shootings.Additionally, while Race also showed much higher shootings involving Black Victims, White Hispanics were also incredibly high in Bronx and Staten Island.

Some areas of bias, that affect this data set are:

1. The spike of violent crimes write around 2020. This may not be representative of a Normal New York City, and may lead to a bad fitting model. We could remove this data, but that would also likely remove the return to pre improvement that New York was achieving.
2. Races and Age also are an area of Bias. with race and age being included, you could attempt to manipulate the data into making certain races or age groups look worse.
    a. I chose to not analyze race and age for my analysis, but this could lead to false conclusions.
3. Boroughs to shooting ratio. Some boroughs have much higher shootings. This could make a boroughs look significantly worse or better even if it isn't. I chose to normalize the data by pulling in population to find the average shootings per million people.