Exploratory Analysis of Gun Violence In New York

2022-07-22

Libraries Used

```
knitr::opts_chunk$set(echo = TRUE)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.7     ✓ dplyr   1.0.9
## ✓ tidyr   1.2.0     ✓ stringr 1.4.0
## ✓ readr   2.1.2     ✓ forcats 0.5.1
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Problem

For this report, New York City shooting incidents which include information from the date, time, boroughs, victim identifiers and perpetrator identifiers will be analyzed. The analysis is interested in describing the where, when and who is committing this crime in an effort to provide clarity with the intent for actionable measures to be taken to reduce the violence happening.

Data Description

The data being analyzed was collected and provided by the city of New York and includes data from 2006 through 2021.

Import Data

The data is initially imported allowing it to be analyzed

```
url_in<-"https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
urls<-str_c(url_in)
NY_gun<-read_csv(urls[1])
```

```
## Rows: 25596 Columns: 19
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## dbl   (7): INCIDENT_KEY, PRECINCT, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD...
## lgl   (1): STATISTICAL_MURDER_FLAG
## time  (1): OCCUR_TIME
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
NY_gun1<-na.omit(NY_gun)
NY_gun1
```

```
## # A tibble: 7,243 × 19
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO      PRECINCT JURISDICTION_CODE
##           <dbl> <chr>      <time>     <chr>        <dbl>              <dbl>
##  1    227950661 05/09/2021 02:50      BRONX           41                  2
##  2    225438895 03/10/2021 07:30      MANHATTAN       28                  0
##  3    232193235 08/13/2021 01:00      QUEENS         109                  0
##  4    225168410 03/04/2021 13:17      BRONX           40                  2
##  5    231965278 08/07/2021 23:57      BRONX           42                  2
##  6    225779049 03/17/2021 22:09      BRONX           52                  0
##  7    223553538 01/26/2021 09:03      MANHATTAN       23                  2
##  8    227496844 04/28/2021 09:10      BRONX           46                  0
##  9    236254076 11/13/2021 16:59      BRONX           44                  0
## 10    226450222 04/02/2021 13:25      BROOKLYN        61                  2
## # … with 7,233 more rows, and 13 more variables: 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>
```

Tidy Data

For Tidying and cleaning the data we chose to remove the values that belonged to variables PRECINCT,JURISDICTION_CODE,X_COORD_CD,Y_COORD_CD,Latitude,Longitude,Lon_Lat,INCIDENT_KEY, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE.

Since this is an exploratory analysis, the goal is to determine a pattern quickly that can then have new questions raised about it and evaluated further. The plan is to investigate which boroughs have the highest density of gun crime and then determine where and who in those areas is creating that crime.

For now, the analysis is not interested in whether the gun violence resulted in a murder or not. We believe solving the underlying issue of gun violence as a whole will bring down the murder count.

```
# Removing the variables stated above
NY_gun2<-select(NY_gun1,-c(PRECINCT,JURISDICTION_CODE,X_COORD_CD,Y_COORD_CD,Latitude,Longitude,L
on_Lat,INCIDENT_KEY, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE))

# Changing the variables to either date type or factor type
NY_gun2 <- NY_gun2 %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE), BORO = as.factor(BORO),LOCATION_DESC =
           as.factor(LOCATION_DESC), STATISTICAL_MURDER_FLAG = as.factor(STATISTICAL_MURDER_FLA
G),
            PERP_AGE_GROUP = as.factor(PERP_AGE_GROUP), PERP_SEX = as.factor(PERP_SEX))

# Providing a summary of the modified data
summary(NY_gun2)
```

```
##    OCCUR_DATE          OCCUR_TIME                   BORO
##  Min.   :2006-01-01   Length:7243       BRONX        :2019
##  1st Qu.:2008-03-02   Class1:hms        BROOKLYN     :2840
##  Median :2010-06-25   Class2:difftime   MANHATTAN    :1062
##  Mean   :2011-10-16   Mode  :numeric    QUEENS       :1055
##  3rd Qu.:2014-11-07                     STATEN ISLAND: 267
##  Max.   :2021-12-31
##
##                    LOCATION_DESC  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  MULTI DWELL - PUBLIC HOUS:2808   FALSE:5595              18-24  :2588
##  MULTI DWELL - APT BUILD  :2019   TRUE :1648              25-44  :2388
##  PVT HOUSE                : 601                           UNKNOWN:1368
##  BAR/NIGHT CLUB           : 452                           <18    : 600
##  GROCERY/BODEGA           : 448                           45-64  : 257
##  COMMERCIAL BLDG          : 165                           65+    :  39
##  (Other)                  : 750                           (Other):   3
##  PERP_SEX
##  F: 194
##  M:6458
##  U: 591
##
##
##
##
```

```
NY_gun2
```

```
## # A tibble: 7,243 × 7
##    OCCUR_DATE OCCUR_TIME BORO      LOCATION_DESC STATISTICAL_MUR… PERP_AGE_GROUP
##    <date>     <time>     <fct>     <fct>         <fct>            <fct>
##  1 2021-05-09 02:50      BRONX     MULTI DWELL … TRUE             25-44
##  2 2021-03-10 07:30      MANHATTAN MULTI DWELL … FALSE            25-44
##  3 2021-08-13 01:00      QUEENS    BAR/NIGHT CL… TRUE             18-24
##  4 2021-03-04 13:17      BRONX     MULTI DWELL … FALSE            25-44
##  5 2021-08-07 23:57      BRONX     MULTI DWELL … TRUE             <18
##  6 2021-03-17 22:09      BRONX     MULTI DWELL … TRUE             18-24
##  7 2021-01-26 09:03      MANHATTAN MULTI DWELL … TRUE             18-24
##  8 2021-04-28 09:10      BRONX     MULTI DWELL … FALSE            25-44
##  9 2021-11-13 16:59      BRONX     BAR/NIGHT CL… FALSE            18-24
## 10 2021-04-02 13:25      BROOKLYN  MULTI DWELL … FALSE            18-24
## # … with 7,233 more rows, and 1 more variable: PERP_SEX <fct>
```

Discussion for missing data

Given this is exploratory and the data set is significantly large with more then 25,000 rows and rows that contain missing data is significantly small, 19; rows that are missing data in any of their respective columns have been omitted.

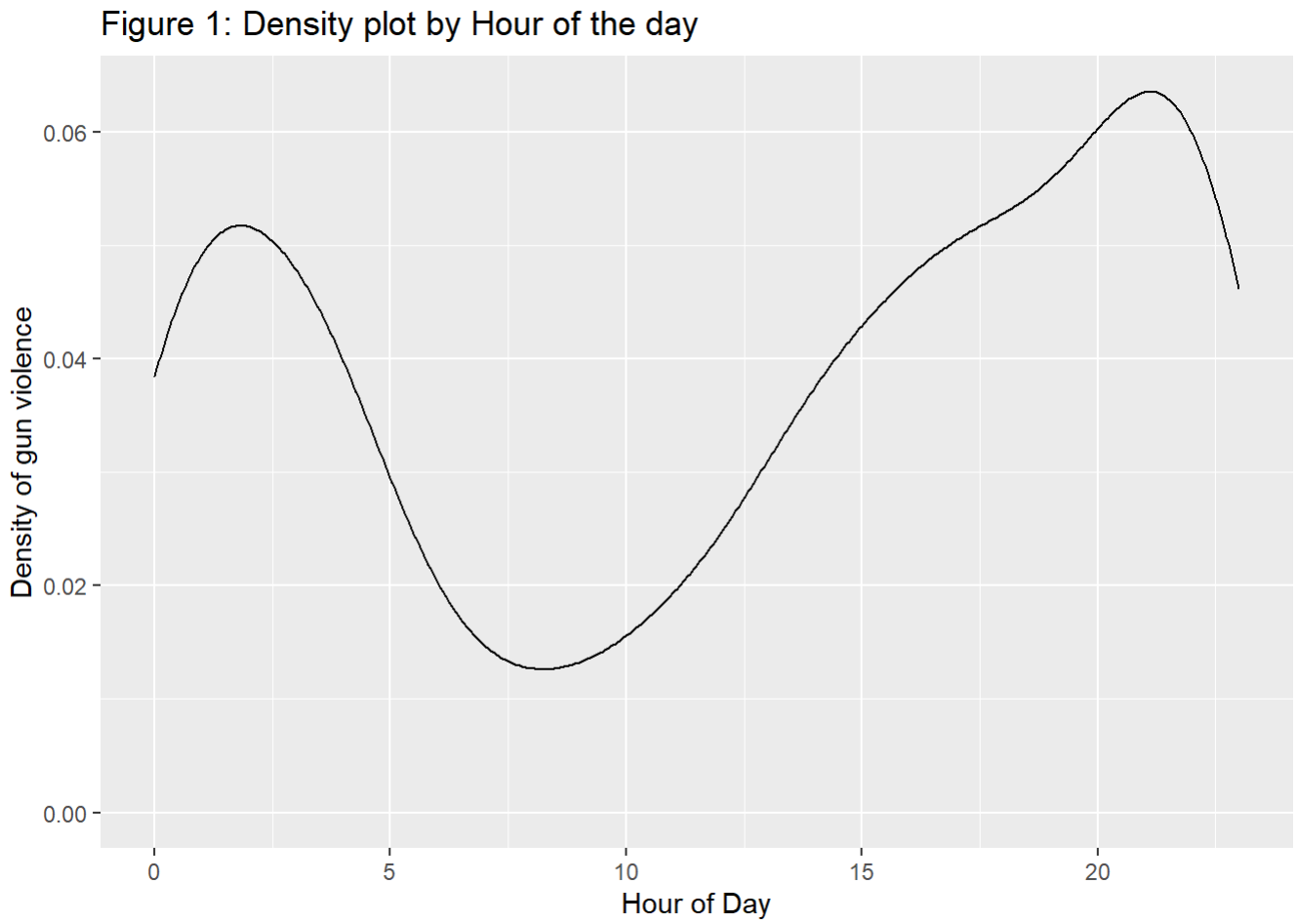
Analysis

After reviewing the summary of the filtered data, it is noted that Brooklyn has the highest amount of gun crime between all the boroughs. As a result, Brooklyn will be the area where the analysis will be focused.

```
NY_gun_BROOKLYN <- filter(NY_gun2, BORO == 'BROOKLYN')
summary(NY_gun_BROOKLYN)
```

```
##    OCCUR_DATE          OCCUR_TIME                   BORO
##  Min.   :2006-01-02   Length:2840       BRONX        :   0
##  1st Qu.:2007-11-18   Class1:hms        BROOKLYN     :2840
##  Median :2009-12-22   Class2:difftime   MANHATTAN    :   0
##  Mean   :2011-05-17   Mode  :numeric    QUEENS       :   0
##  3rd Qu.:2013-09-23                     STATEN ISLAND:   0
##  Max.   :2021-12-18
##
##                    LOCATION_DESC  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  MULTI DWELL - PUBLIC HOUS:1207   FALSE:2233              18-24  :988
##  MULTI DWELL - APT BUILD  : 739   TRUE : 607              25-44  :911
##  PVT HOUSE                : 210                           UNKNOWN:620
##  GROCERY/BODEGA           : 179                           <18    :229
##  BAR/NIGHT CLUB           : 144                           45-64  : 76
##  COMMERCIAL BLDG          :  63                           65+    : 15
##  (Other)                  : 298                           (Other):  1
##  PERP_SEX
##  F:  57
##  M:2506
##  U: 277
##
##
##
##
```

```
#Adding additional data
NY_gun_BROOKLYN$YEAR <- year(NY_gun_BROOKLYN$OCCUR_DATE)
NY_gun_BROOKLYN$HOUR <- hour(NY_gun_BROOKLYN$OCCUR_TIME)
NY_gun_BROOKLYN <- NY_gun_BROOKLYN %>%
  mutate(YEAR = as.factor(YEAR))

#Plotting occurrence time of gun violence to see times of peak gun violence
BROOKLYN_PLOT_TIMES <- ggplot(NY_gun_BROOKLYN, aes(HOUR))+
  geom_density(kernel="gaussian")+
  labs(x = "Hour of Day", y = "Density of gun violence", title = "Figure 1: Density plot by Hour
of the day")
BROOKLYN_PLOT_TIMES
```



Reviewing Figure 1, over all the data in Brooklyn, there is a peak in the early morning and later evening for gun violence. Is this constant over all the years of data taken?

```
# Creating a violin plot
BROOKLYN_PLOT_TIMEvYEAR <- ggplot(NY_gun_BROOKLYN, aes(x = YEAR, y = HOUR))+
  geom_violin(aes(color = YEAR, fill = YEAR))+
  labs(x = "Year", y = "Hour", title = "Figure 2: Violin plot of Year vs Hour")
BROOKLYN_PLOT_TIMEvYEAR
```

Figure 2: Violin plot of Year vs Hour

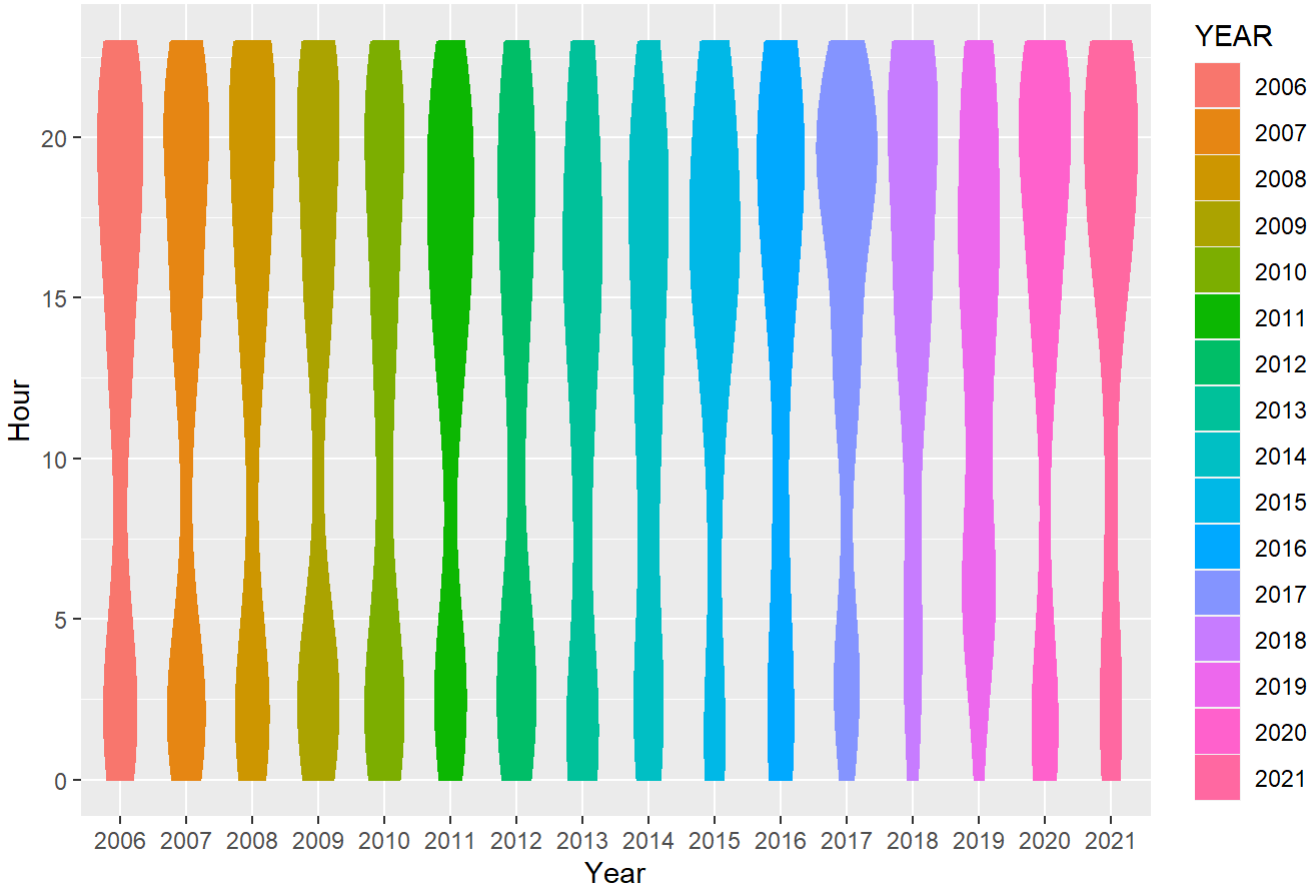
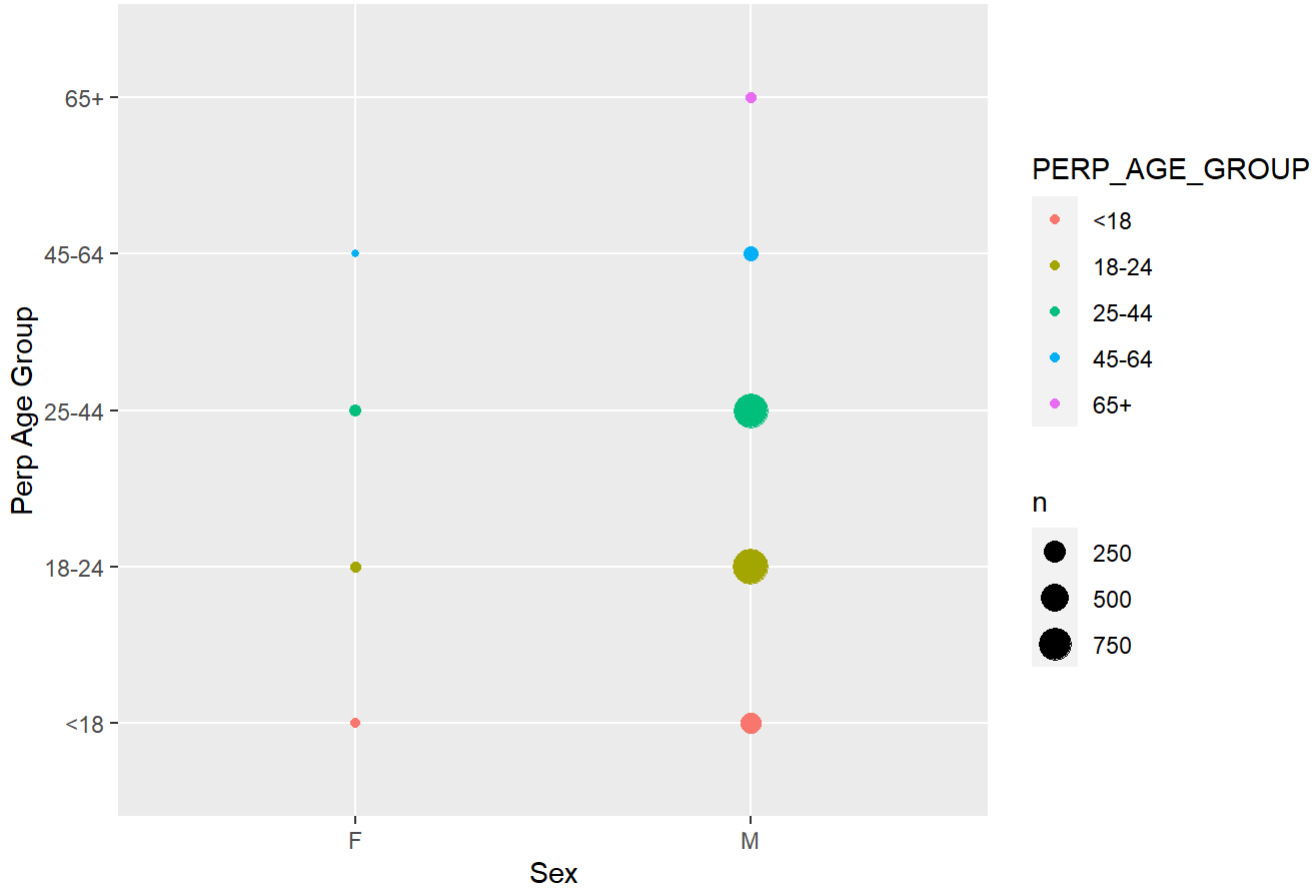


Figure 2, above, suggests the density plot over all years is accurate. We can see, for the most part, gun violence in the Brooklyn borough has been reported in the early morning hours and later evening hours. It is worth noting that the year 2019 has a more even distribution then the other years which could be attributed to different factors. Further analysis, with possibly more data from other sources needed, of that year would need to be done to identify why that occurred.

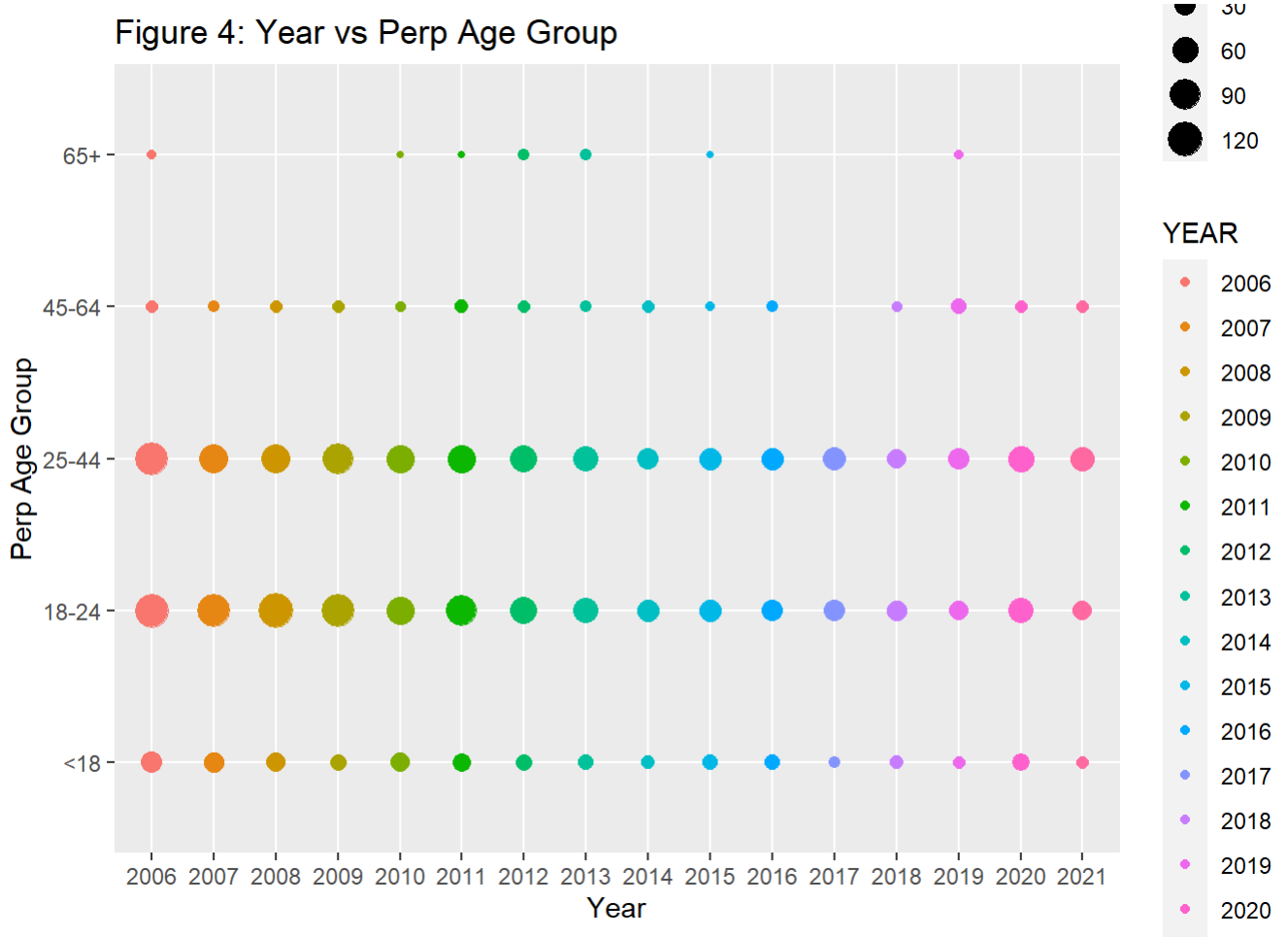
Next, the location of where the typical gun violence occurs is analysed again reviewing it over all years and by year

```
# Removing data that does not make sense in the data set, potentially a typo or other error of s
ome kind during entry
NY_gun_BROOKLYN_filter <- filter(NY_gun_BROOKLYN, PERP_AGE_GROUP != '940', PERP_AGE_GROUP != '22
4', PERP_AGE_GROUP != '1020', PERP_AGE_GROUP
                                 !='UNKNOWN',PERP_SEX != 'U')
# Creating plot to display sex vs perpetrator age
BROOKLYN_PLOT_SEXvAGE <- ggplot(NY_gun_BROOKLYN_filter, aes(x = PERP_SEX, y = PERP_AGE_GROUP))+
  geom_count(aes(color = PERP_AGE_GROUP))+
  labs(x = "Sex", y = "Perp Age Group", title = "Figure 3: Sex vs Perp Age Group")
BROOKLYN_PLOT_SEXvAGE
```

Figure 3: Sex vs Perp Age Group



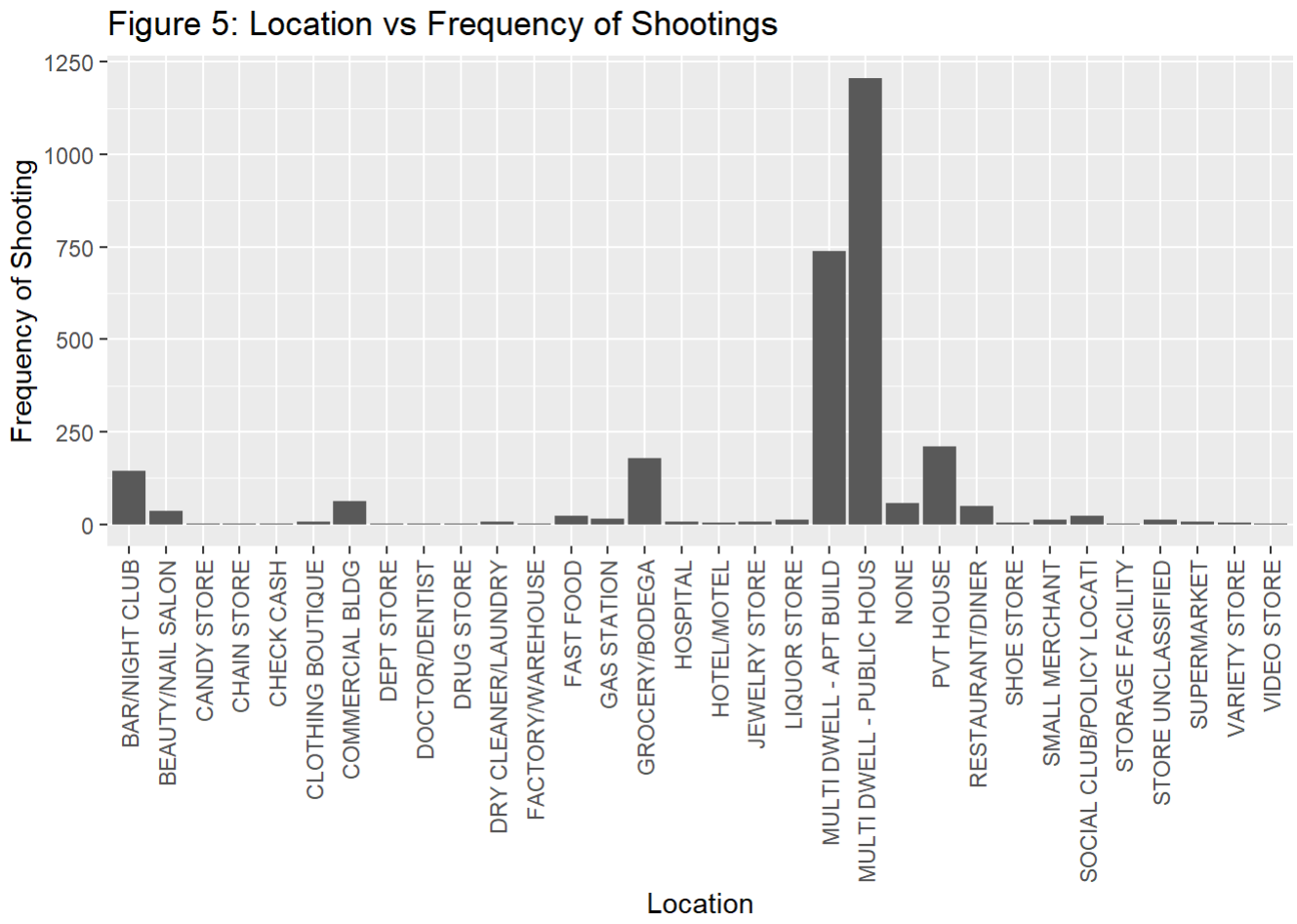

```
# Creating plot to display Perpetrator age vs year
BROOKLYN_PLOT_AGEvYEAR <- ggplot(NY_gun_BROOKLYN_filter, aes(x = YEAR, y = PERP_AGE_GROUP))+
  geom_count(aes(color = YEAR, fill = YEAR))+
  labs(x = "Year", y = "Perp Age Group", title = "Figure 4: Year vs Perp Age Group")
BROOKLYN_PLOT_AGEvYEAR
```



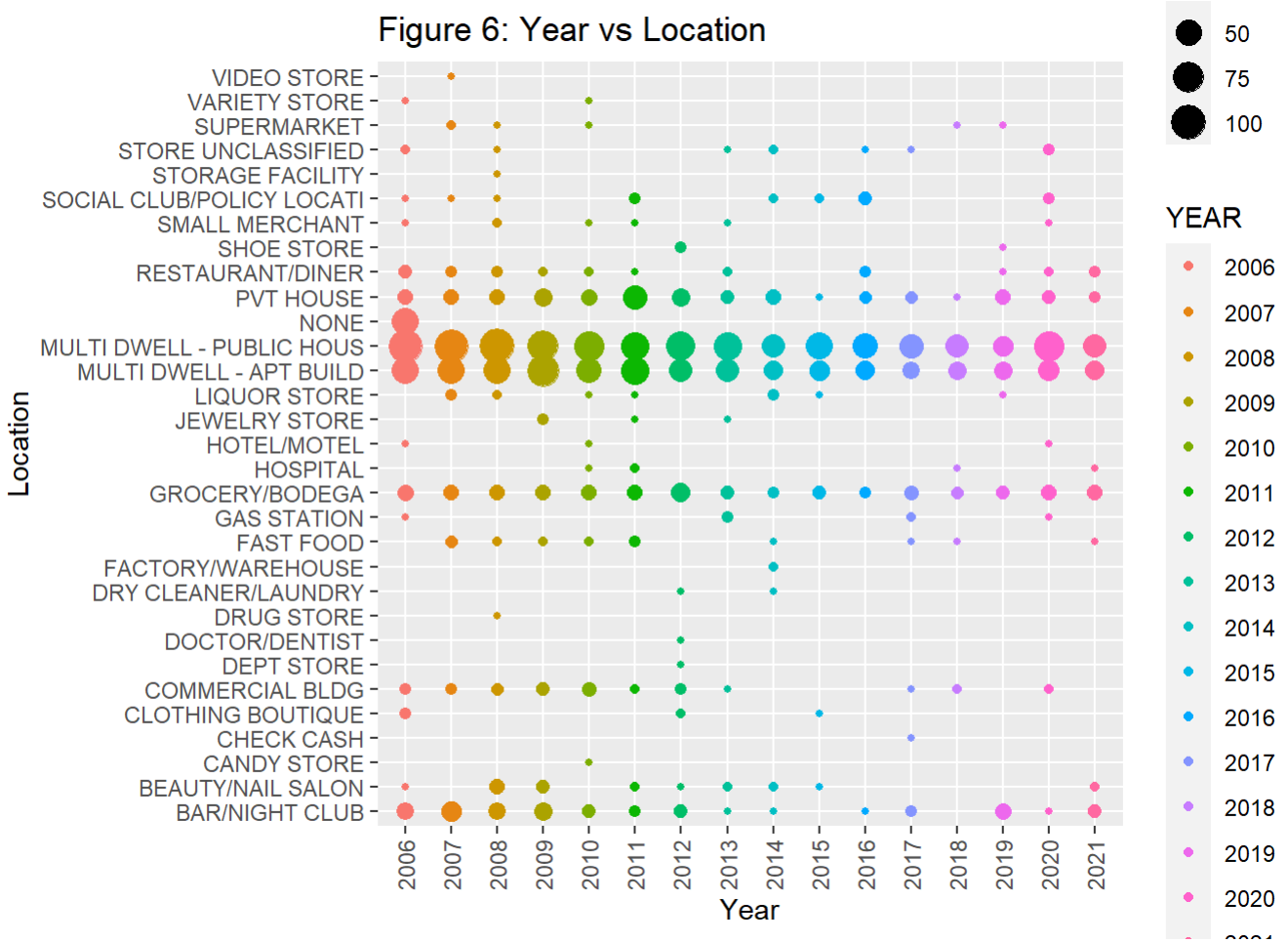
Reviewing Figures 3 and 4, it is noticed mostly males between the ages of 18-44 are the perpetrators of the violence with that age range staying consistent over all years in the data set.

Next, we want to know where the crime is typically occurring. The data set provides a location description variable so we will use that to answer this question. We will again analyze whether this has stayed constant year over year.

```
# Create plot to show frequency of gun violence at each specified location
BROOKLYN_PLOT_LOCFREQ <- ggplot(NY_gun_BROOKLYN, aes(LOCATION_DESC))+
  geom_bar()+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  labs(x = "Location", y = "Frequency of Shooting", title = "Figure 5: Location vs Frequency of
 Shootings")
BROOKLYN_PLOT_LOCFREQ
```




```
# Create plot to show year over year analysis
BROOKLYN_PLOT_LOCvYEAR <- ggplot(NY_gun_BROOKLYN_filter, aes(x = YEAR, y = LOCATION_DESC))+
  geom_count(aes(color = YEAR, fill = YEAR))+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  labs(x = "Year", y = "Location", title = "Figure 6: Year vs Location")
BROOKLYN_PLOT_LOCvYEAR
```



Reviewing Figure 5 we notice a concentration of gun crime happening around apartment buildings and public housing. This again stays constant when looked at year over year (Figure 6) .

Model

The final step in the analysis was to create a model, it makes sense to evaluate whether there is a relation between quantities of shootings (incidents) and how many murders occurred by year. Being able to analysis this should support a decision to increase efforts to reduce gun violence as a whole as opposed to just targeting violent offenders.

```
#filtering for murders
NY_gun_BROOKLYN_MURDER <- NY_gun_BROOKLYN %>%
  filter(STATISTICAL_MURDER_FLAG == "TRUE")%>%
  group_by(YEAR)
NY_gun_BROOKLYN_MURDER
```

```
## # A tibble: 607 × 9
## # Groups:   YEAR [16]
##    OCCUR_DATE OCCUR_TIME BORO     LOCATION_DESC  STATISTICAL_MUR… PERP_AGE_GROUP
##    <date>     <time>     <fct>    <fct>          <fct>            <fct>
##  1 2021-04-05 23:15      BROOKLYN MULTI DWELL -… TRUE             45-64
##  2 2021-02-07 01:23      BROOKLYN MULTI DWELL -… TRUE             25-44
##  3 2021-12-14 19:29      BROOKLYN GROCERY/BODEGA TRUE             25-44
##  4 2021-09-10 17:35      BROOKLYN MULTI DWELL -… TRUE             25-44
##  5 2021-10-01 16:53      BROOKLYN MULTI DWELL -… TRUE             18-24
##  6 2021-04-25 17:32      BROOKLYN MULTI DWELL -… TRUE             25-44
##  7 2021-04-25 17:32      BROOKLYN MULTI DWELL -… TRUE             25-44
##  8 2021-12-17 22:13      BROOKLYN GROCERY/BODEGA TRUE             25-44
##  9 2021-12-18 00:15      BROOKLYN PVT HOUSE      TRUE             25-44
## 10 2021-10-13 17:55      BROOKLYN PVT HOUSE      TRUE             25-44
## # … with 597 more rows, and 3 more variables: PERP_SEX <fct>, YEAR <fct>,
## #   HOUR <int>
```

```
#collecting murders by year
table(NY_gun_BROOKLYN_MURDER$YEAR)
```

```
##
## 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
##   89   55   56   65   36   47   29   36   18   25   26   24   21   20   38   22
```

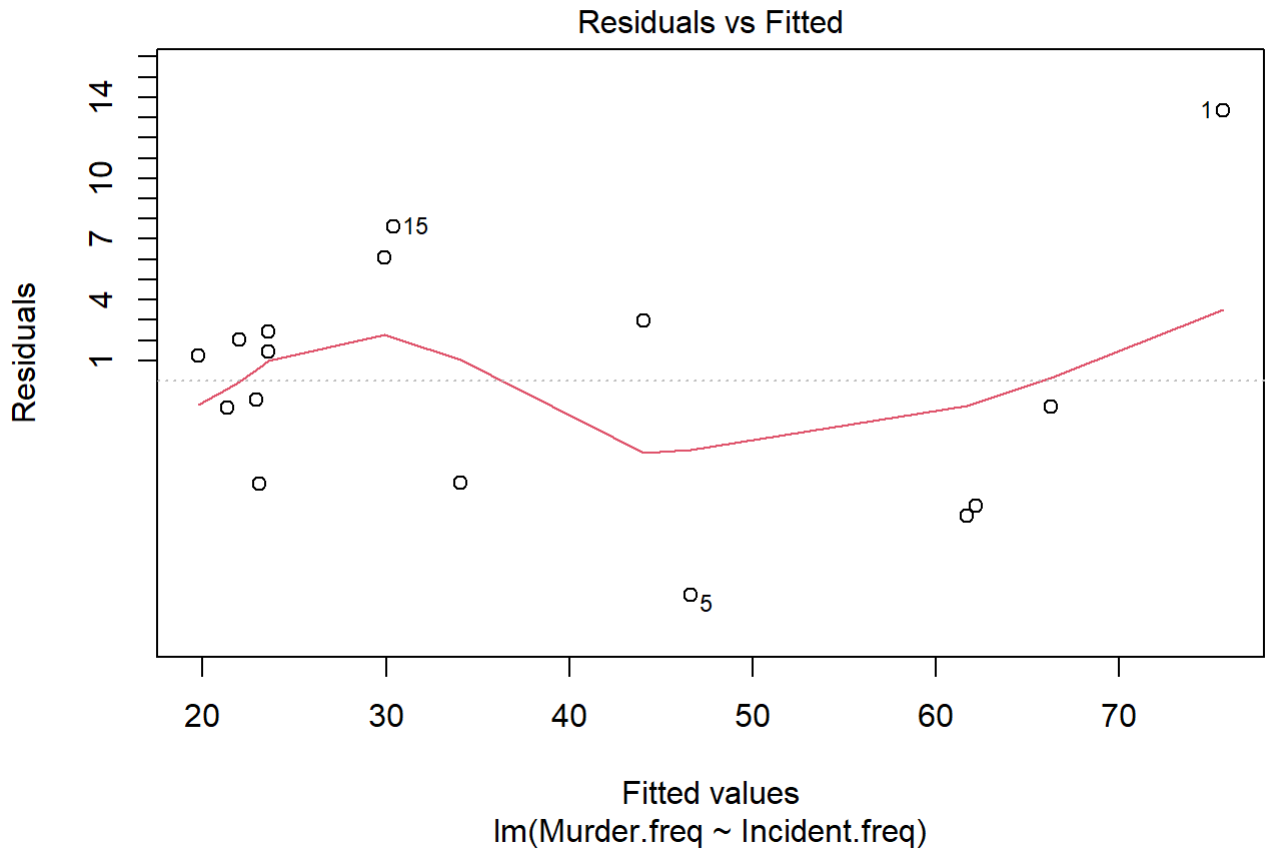
```
Murder.freq<-table(NY_gun_BROOKLYN_MURDER$YEAR)

#collecting all incidents by year
table(NY_gun_BROOKLYN$YEAR)
```

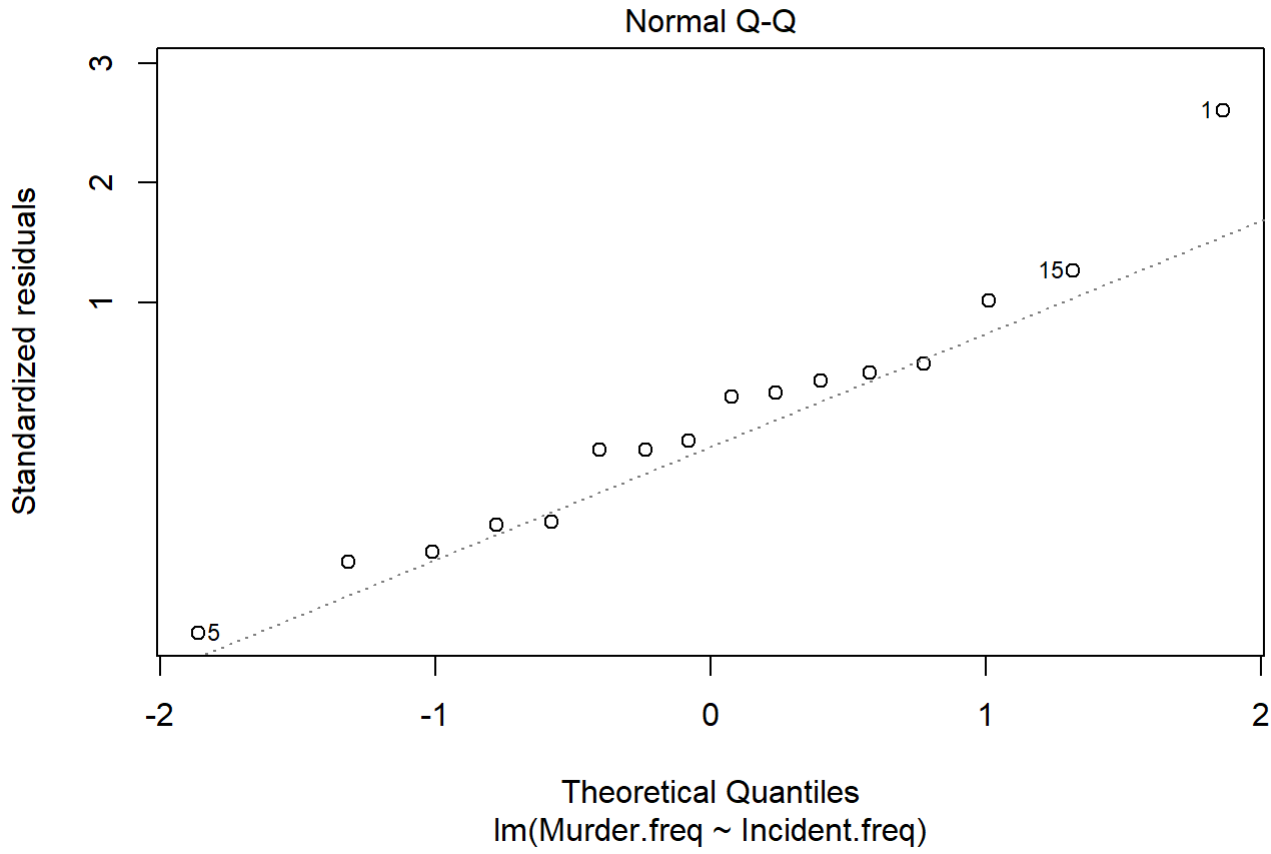
```
##
## 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021
##  415  327  330  356  232  216  153  127   84   87   87   77   63   73  130   83
```

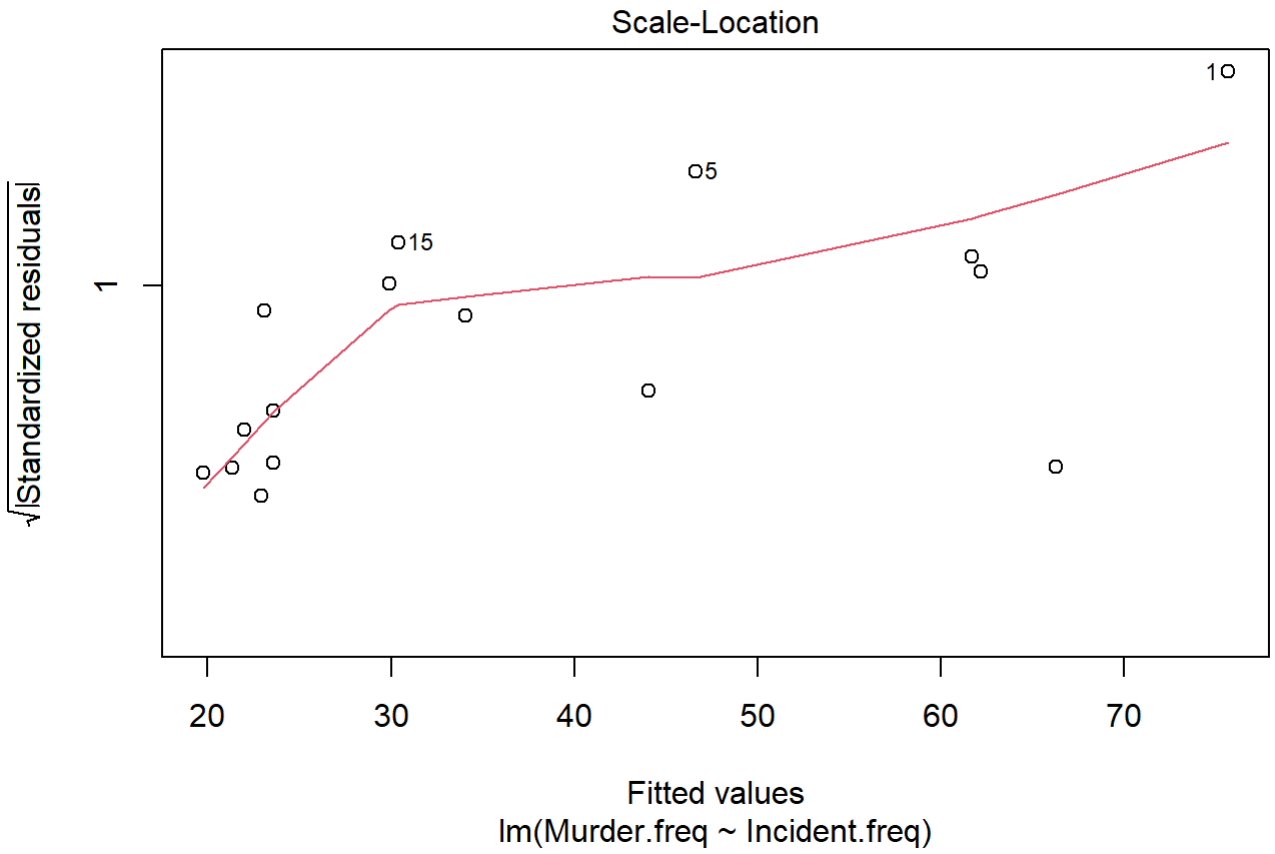
```
Incident.freq<-table(NY_gun_BROOKLYN$YEAR)

#creating a linear model
mod <- lm(formula = Murder.freq~Incident.freq, data = NY_gun_BROOKLYN)
summary(mod)
```

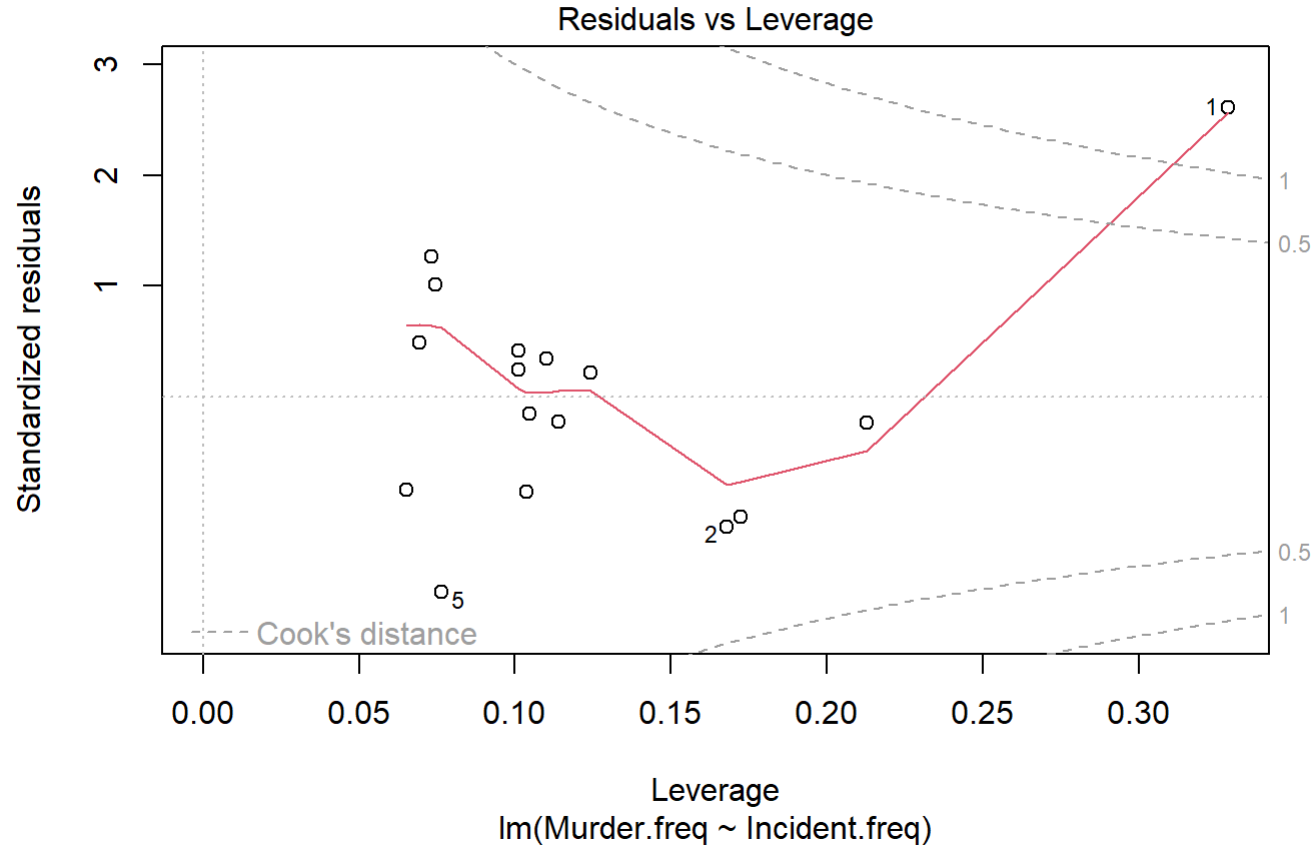
```
##
## Call:
## lm(formula = Murder.freq ~ Incident.freq, data = NY_gun_BROOKLYN)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.593  -5.057   0.159   2.564  13.343
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.74722    2.86890   3.398  0.00433 **
## Incident.freq  0.15882    0.01356  11.711 1.28e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.243 on 14 degrees of freedom
## Multiple R-squared:  0.9074, Adjusted R-squared:  0.9008
## F-statistic: 137.1 on 1 and 14 DF,  p-value: 1.278e-08
```

```
plot(mod)
```

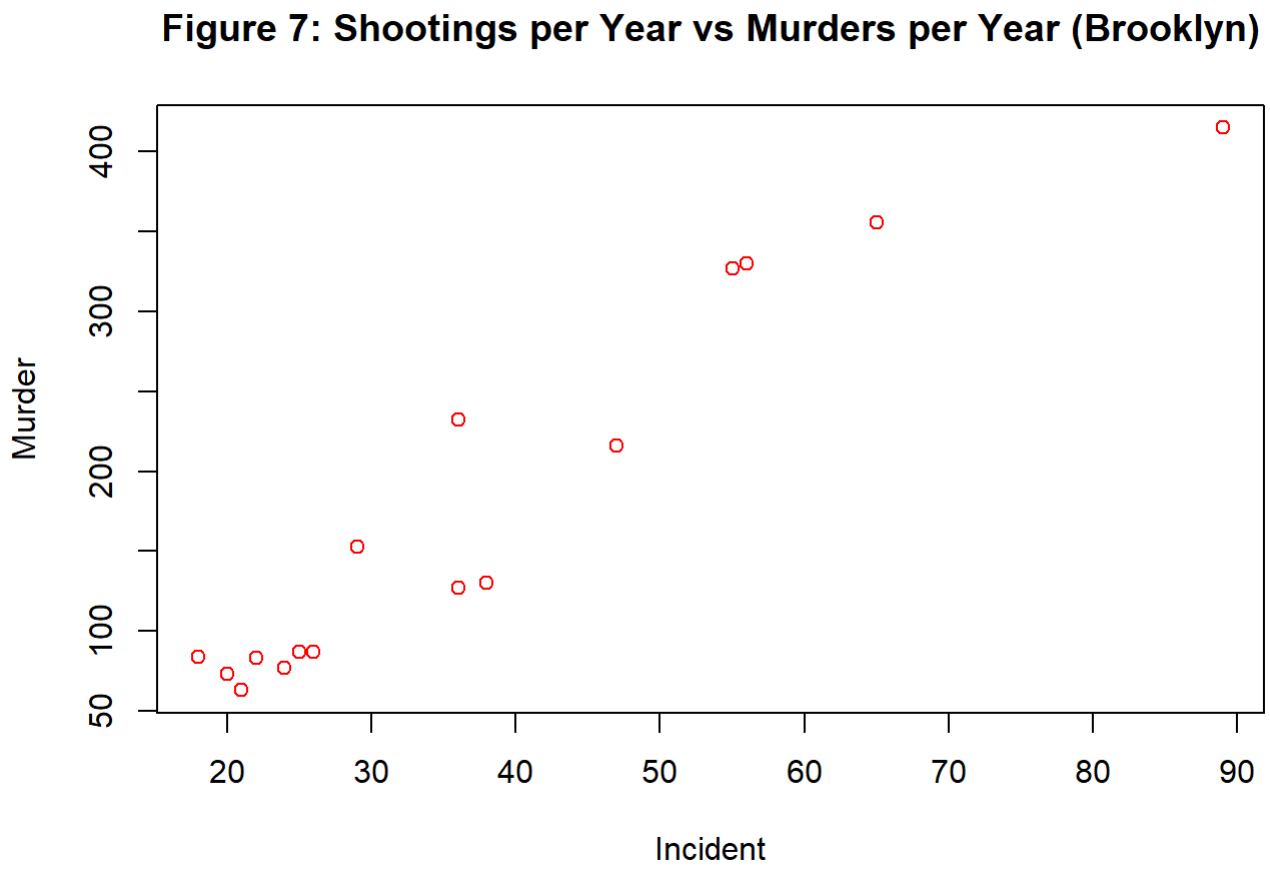

Residuals vs Fitted
Residuals
1
4
7
10
14
15
1
5
20
30
40
50
60
70
Fitted values
lm(Murder.freq ~ Incident.freq)


Normal Q-Q
Standardized residuals
1
2
3
1
15
5
-2
-1
0
1
2
Theoretical Quantiles
lm(Murder.freq ~ Incident.freq)


Scale-Location
√|Standardized residuals|
1
10
5
15
20
30
40
50
60
70
Fitted values
lm(Murder.freq ~ Incident.freq)


Residuals vs Leverage
Standardized residuals
3
2
1
10
1
0.5
0.5
1
2
5
Cook's distance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(Murder.freq ~ Incident.freq)

```
#storing murders by year and incidents by year in its own data frame
linear_data<-data.frame(Murder.freq,Incident.freq)

#create scatterplot of raw data and add line of best fit
plot(linear_data$Freq, linear_data$Freq.1, col='red', main='Figure 7: Shootings per Year vs Murd
ers per Year (Brooklyn)', xlab='Incident', ylab='Murder')
abline(lm(Murder.freq~Incident.freq))
```



Reviewing the model created, the p value is below .005 supporting that the model is statistically significant. When the Residuals vs Fitted plot is looked at, we can conclude that a linear regression model is appropriate for the dataset since the redline does not deviate from the horizontal dotted line substantially. Figure 7 helps visualize the model that was created and supports that there is a linear relationship between incident rate and murder rate, as the amount of shootings goes up so does the amount of murders in Brooklyn.

Bias

When reviewing the data set, we wanted to form a question that specifically did not require race to be evaluated in order for it to be answered. As it relates to gun violence, we did not believe to add significant value to an analysis because it has been shown that race does not necessarily correlate to violence and to include it in an analysis would only serve to introduce complexity. To state it a different way, there are concentrations of certain races, for various other factors, in these areas and as a result there will be a higher amount of those races committing these crimes. To mitigate this, race was removed from the data subset that would be analyzed.

Discussion and Summary

The analysis has successfully identified the highest concentration of gun crime in NYC and provided visualizations of when shootings occur by hour, where the occur most often and who, by sex and age, is committing the crimes. Additionally, the analysis provides a linear model to support that by reducing all gun violence, there would be a decrease in violent crime (murders).

The analysis should be used to develop actionable items that can be taken to reduce gun violence. These actions might include, but not be limited to, increasing police presence during hours where gun violence is more frequent and increasing funding to inner city programs that are focusing efforts around youth and their maturity from teenage years through young adult hood.

Future Analysis

Further steps should be taken to conduct the same analysis on the other boroughs in NYC. While it is expected the similar conclusions would be drawn, it should not be stated as fact without the analysis being completed. Additionally, this analysis should be completed on an annual basis as mitigation to the violence are rolled out to evaluate effectiveness of the efforts. Finally, the analysis should be continually updated and expanded as more data is included to ensure the right steps are being taken to curb the present issue of gun violence.