NYPD Shooting Incidents

3/25/2022

```
library(tidyverse)
library(lubridate)
library(hms)
library(ggplot2)
library(patchwork) #to display to plots together
library(readxl)
```

Import Data Set from URL

Read the NYPD Shooting Incident Data (Historic) from https://catalog.data.gov/dataset and store it in **nypd** List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

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.

```
nypd <- read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv")
```

Tidy and Transform your data

First we deal with missing or missnamed data. We will set all the missing or unknown data to Na or Unknown. As well as the data that makes no sense like perp age group 1020 or 940.

We will start with the Location Description, as is the first one that has empty rows transforming the empty ones to ‘NONE’ that is one of the options.

```
nypd[nypd$LOCATION_DESC == "",]$LOCATION_DESC <- "NONE"
```

For PERP_AGE_GROUP the available options are: [“<18”, “18-24”, “25-44”, “45-64”, “65+”, “UNKNOWN”] options that aren’t an age group or ““,”UNKNOWN” I will consolidate change to “UNKONWN”. For PERP_SEX will change empty ones for “U” as in unknown same for PER_RACE empty ones will say “UNKNOWN

```
nypd[!(nypd$PERP_AGE_GROUP %in% c("<18", "18-24", "25-44", "45-64", "65+", "UNKNOWN")),]$PERP_AGE_GROUP
nypd[nypd$PERP_SEX == "" ,]$PERP_SEX <- "U"
nypd[nypd$PERP_RACE == "" ,]$PERP_RACE <- "UNKNOWN"
```

The rest of the columns don’t seem to have any missing values.

Second we tranform data type and get rid of columns we dont need Most of the data is stored as character but is more useful as categorical as it comes from categories like boroughs, race, age group, and if it was flagged as murder (the victim died) we change it to logical (TRUE, FALSE). We also get rid of the Lon_Lat column as we already have the long and lat separately. The location and longitude and latitude could be needed in case we want to find location hot-spots, locate the shootings in a map or find the distance between shootings so I will keep such data.

```
nyp <- nypd %>%
    mutate(OCCUR_DATE = mdy(OCCUR_DATE), OCCUR_TIME = as_hms(OCCUR_TIME), BORO = as.factor(BORO), PRECI
    select(-c(STATISTICAL_MURDER_FLAG,Lon_Lat))
```

Summary of the data

Check for missing data.

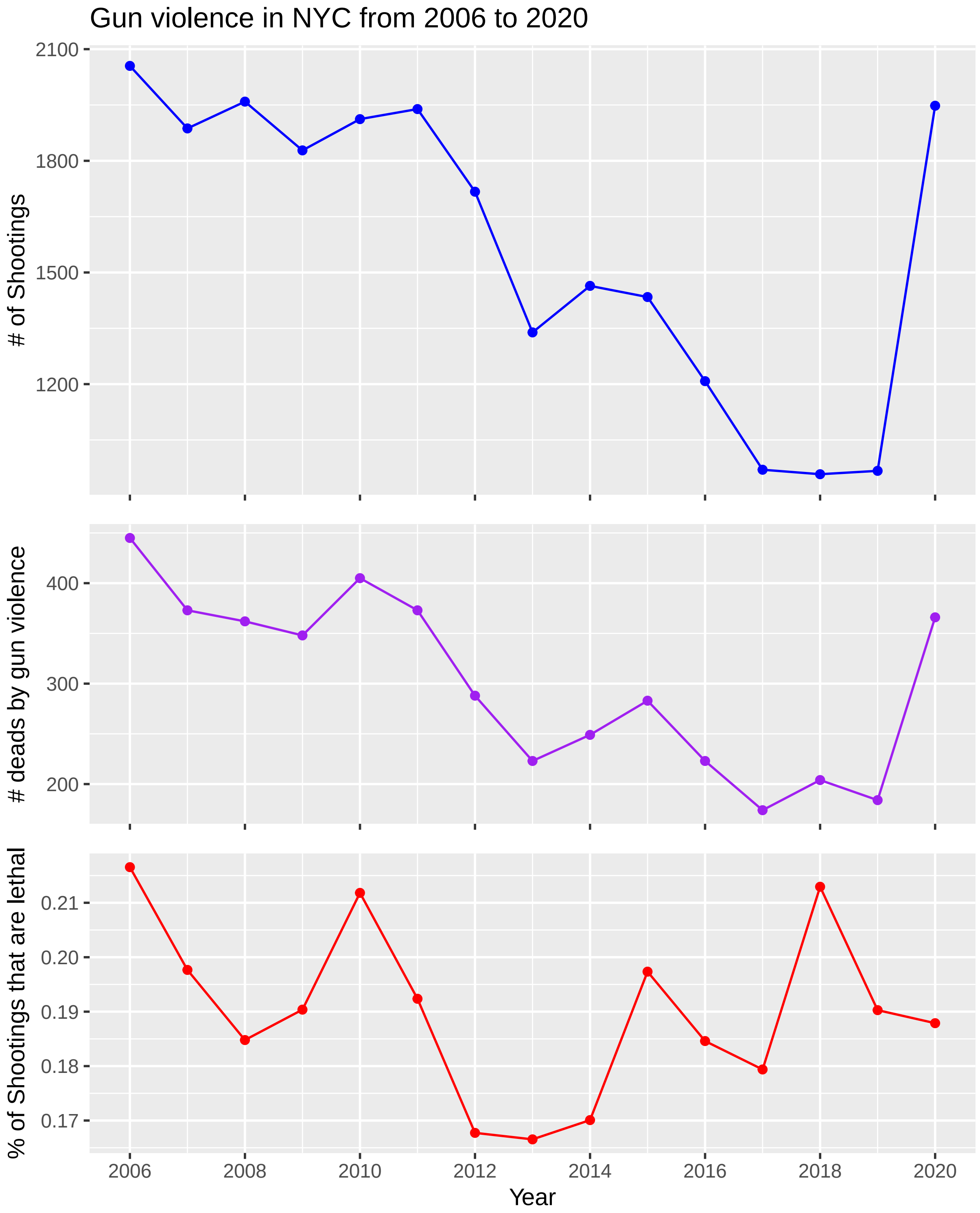
```
summary(nyp)
```

```
##   INCIDENT_KEY         OCCUR_DATE          OCCUR_TIME
##  Min.   :  9953245   Min.   :2006-01-01   Length:23585
##  1st Qu.: 55322804   1st Qu.:2008-12-31   Class1:hms
##  Median : 83435362   Median :2012-02-27   Class2:difftime
##  Mean   :102280741   Mean   :2012-10-05   Mode  :numeric
##  3rd Qu.:150911774   3rd Qu.:2016-03-02
##  Max.   :230611229   Max.   :2020-12-31
##
##             BORO         PRECINCT     JURISDICTION_CODE
##  BRONX        :6701   75     : 1375   0   :19629
##  BROOKLYN     :9734   73     : 1284   1   :   54
##  MANHATTAN    :2922   67     : 1101   2   : 3900
##  QUEENS       :3532   79     :  921   NA's:    2
##  STATEN ISLAND: 696   44     :  841
##                       47     :  818
##                       (Other):17245
##                     LOCATION_DESC   PERP_AGE_GROUP  PERP_SEX
##  NONE                     :13756   <18    : 1368   F:  335
##  MULTI DWELL - PUBLIC HOUS: 4240   18-24  : 5508   M:13490
##  MULTI DWELL - APT BUILD  : 2553   25-44  : 4714   U: 9760
##  PVT HOUSE                :  857   45-64  :  495
##  GROCERY/BODEGA           :  574   65+    :   54
##  BAR/NIGHT CLUB           :  562   UNKNOWN:11446
##  (Other)                  : 1043
##                            PERP_RACE     VIC_AGE_GROUP   VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    : 2525   F: 2204
##  ASIAN / PACIFIC ISLANDER      :  122   18-24  : 9003   M:21370
##  BLACK                         :10025   25-44  :10303   U:   11
##  BLACK HISPANIC                : 1096   45-64  : 1541
##  UNKNOWN                       :10097   65+    :  154
##  WHITE                         :  255   UNKNOWN:   59
##  WHITE HISPANIC                : 1988
##                             VIC_RACE       X_COORD_CD        Y_COORD_CD
##  AMERICAN INDIAN/ALASKAN NATIVE:    9   Min.   : 914928   Min.   :125757
##  ASIAN / PACIFIC ISLANDER      :  327   1st Qu.: 999925   1st Qu.:182539
##  BLACK                         :16869   Median :1007654   Median :193470
##  BLACK HISPANIC                : 2245   Mean   :1009379   Mean   :207300
##  UNKNOWN                       :   65   3rd Qu.:1016782   3rd Qu.:239163
##  WHITE                         :  620   Max.   :1066815   Max.   :271128
##  WHITE HISPANIC                : 3450
##     Latitude       Longitude      MURDER_FLAG
##  Min.   :40.51   Min.   :-74.25   Mode :logical
##  1st Qu.:40.67   1st Qu.:-73.94   FALSE:19085
##  Median :40.70   Median :-73.92   TRUE :4500
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

Visualizations.

First lets plot the total number of shootings and murders by gun violence per year in NYC and how it changed over the last 15 years. Including how lethal this shootings were.

```
nyp$Yr <- year(nyp$OCCUR_DATE)
yrl <- nyp %>%
  group_by(Yr) %>%
  summarize(shootings = n(), murders_gun = sum(MURDER_FLAG))
yrl$deadly_per = yrl$murders_gun/yrl$shootings

p1 = ggplot(yrl, aes(x=Yr,y = shootings)) +
  geom_point(color = 'blue' ) +
  geom_line(color = 'blue' ) +
  labs(x = NULL,y = "# of Shootings", title = "Gun violence in NYC from 2006 to 2020") +
  scale_x_continuous(breaks = seq(2006,2021,2), labels = NULL)

p2 = ggplot(yrl, aes(x=Yr,y=murders_gun)) +
  geom_point(color = 'purple' ) +
  geom_line(color = 'purple' ) +
  labs (x = NULL, y = "# deads by gun violence") +
  scale_y_continuous(breaks = c(200,300, 400)) +
  scale_x_continuous(breaks = seq(2006,2021,2),labels = NULL)

p3 = ggplot(yrl, aes(x=Yr,y = deadly_per)) +
  geom_point(color = 'red' ) +
  geom_line(color = 'red' ) +
  labs(x = "Year",y = "% of Shootings that are lethal") +
  scale_y_continuous(breaks = seq(0.15,0.25,0.01)) +
  scale_x_continuous(breaks = seq(2006,2021,2))

layout <- c(area(1,1, 3, 1),area(4,1,5, 1), area(6,1,7,1))
p1 + p2  + p3 +plot_layout(design = layout)
```

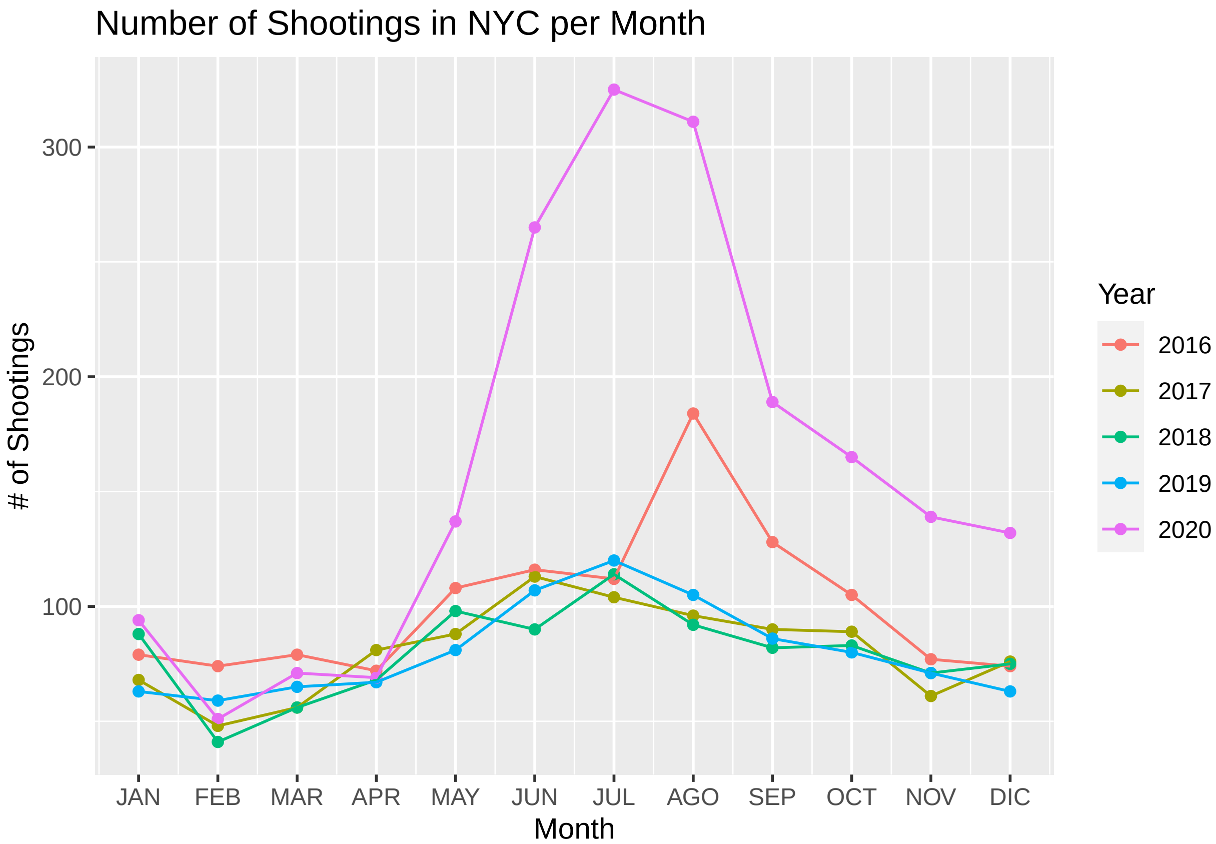


shootings-1.pdf

We can see that gun violence dropped slowly from 2006 to 2011 and then had a big drop until 2017, was stable until 2019 and then skyrocket back to 2011 levels. In 2020 the number of shootings doubled the ones of 2019. But at least the lethality did not increase.

I wanna see when this trend started so will analyze the year 2020 by month and compare it to previous years.

```
ny1920 <- nyp %>%
  filter(Yr > 2015) %>%
  mutate("Month"= month(OCCUR_DATE), "Year" = as.factor(Yr)) %>%
  group_by(Year,Month) %>%
  summarize(shootings = n(), murders_gun = sum(MURDER_FLAG))


f1920 <- ggplot(ny1920) +
  geom_point(aes(x=Month, y=shootings, colour=Year))
f1920 +
  geom_line(aes(x=Month, y=shootings, colour=Year)) +
  scale_x_continuous(breaks = seq(1,12),
                     labels =c('JAN',"FEB", "MAR","APR","MAY","JUN",
                               "JUL","AGO","SEP","OCT","NOV","DIC")) +
  labs(x = "Month",y = "# of Shootings",
       title = "Number of Shootings in NYC per Month")
```



comparison-1.pdf

We can see a yearly trend where shootings increase during the warmer months of the year. However the shootings of 2020 are almost 3 times more than on previous summers. Besides the 2020 peak that started in May, there is also a peak in August 2016.

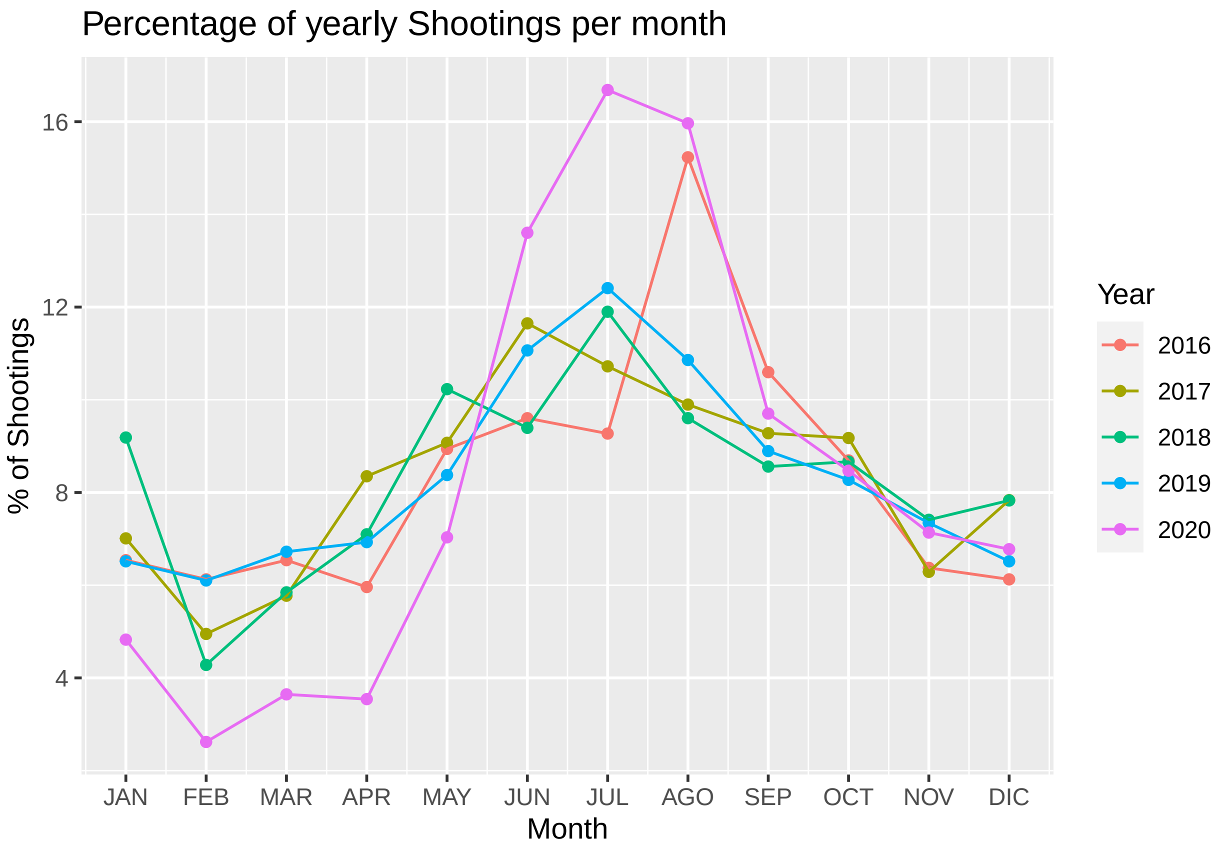
Whats the percantge of crimes per month in each year.

```
yrl_b <- yrl %>%
  rename(total_shootings = shootings, total_murders = murders_gun)
```

```
seasonal_y <- nyp %>%
  mutate("Month"= month(OCCUR_DATE)) %>%
  group_by(Yr,Month) %>%
  summarize(shootings = n(), murders_gun = sum(MURDER_FLAG))%>%
  ungroup()
```

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

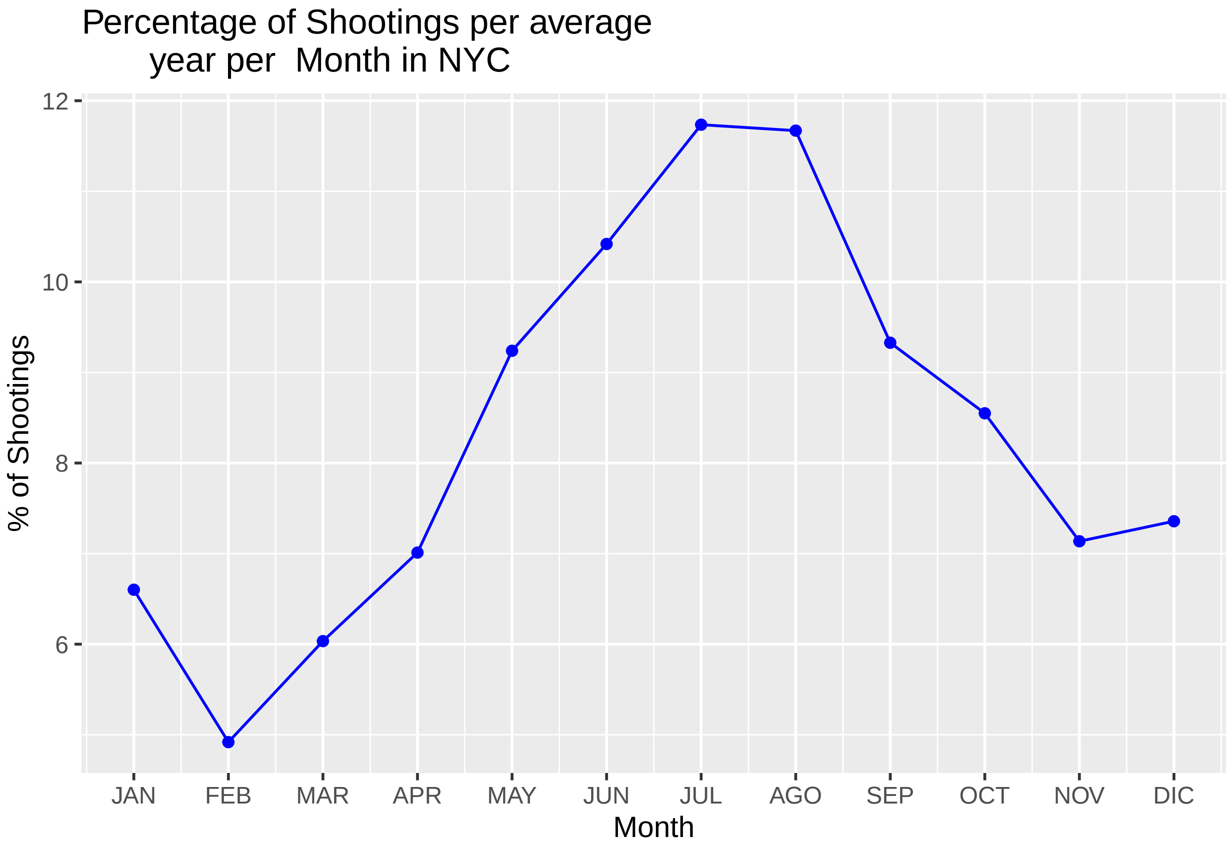
```
seasonal_y <- left_join(seasonal_y,yrl_b, by = "Yr")%>%
  mutate(per_shootings = 100*shootings/total_shootings,
         per_murder = 100*murders_gun/total_murders)


p1620 <- ggplot(seasonal_y[seasonal_y$Yr>2015,],aes(x=Month,y = per_shootings, colour = as.factor(Yr)))
  geom_point()+
  geom_line()+
  scale_x_continuous(breaks = seq(1,12), labels =c('JAN',"FEB", "MAR","APR","MAY","JUN","JUL","AGO","SE
  labs(x = "Month",y = "% of Shootings", title = "Percentage of yearly Shootings per month" , col = "Yea

p1620
```



montly nyc-1.pdf

If we average over all the years

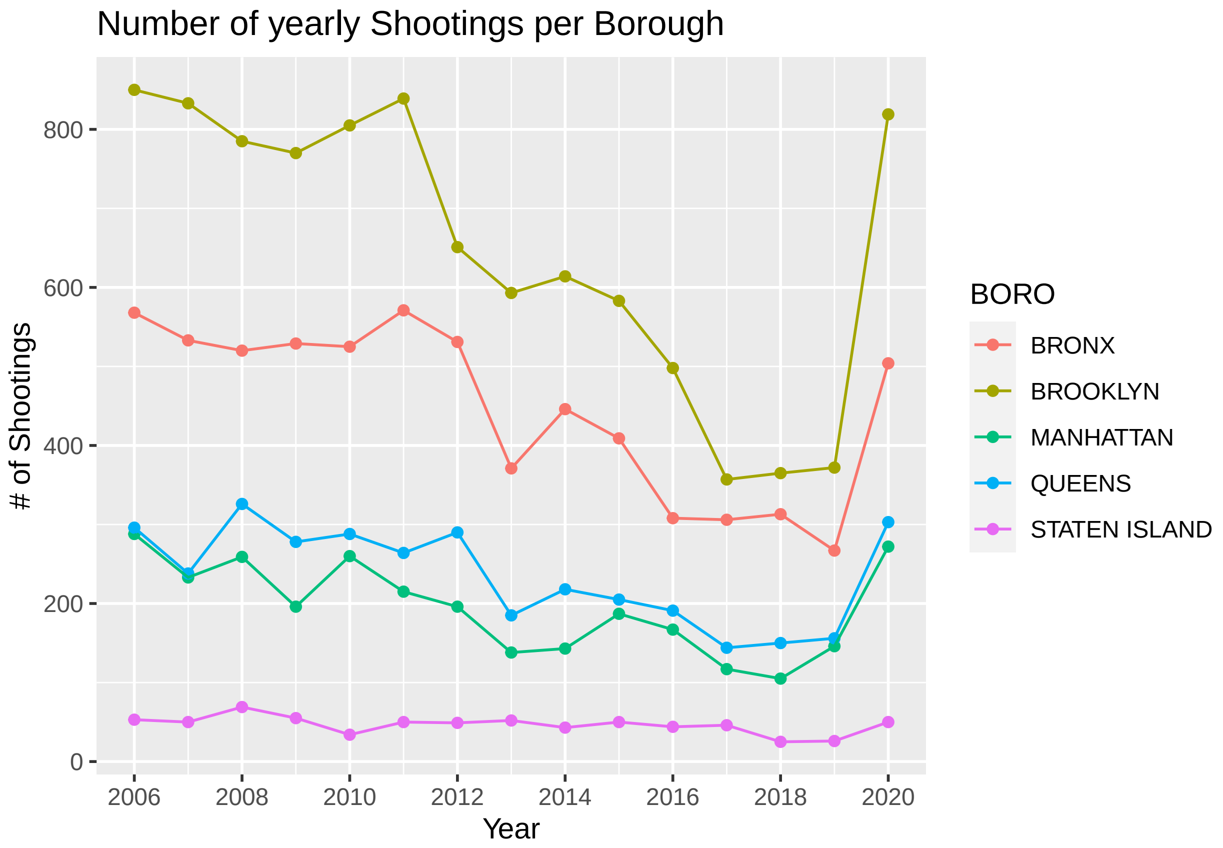
```
avg_year_nyc <- seasonal_y %>%
  group_by(Month)%>%
  summarise(shootings= sum(per_shootings)/n(),
            murders = sum(per_murder)/n())%>%
  ungroup()



pnyc_av <- ggplot(avg_year_nyc , aes(x = Month)) +
  geom_point(aes(y=shootings), col = 'blue') +
  geom_line(aes( y=shootings), col = 'blue') + scale_x_continuous(breaks = seq(1,12), labels =c('JAN',"
  labs(x = "Month",y = "% of Shootings",
       title = "Percentage of Shootings per average
       year per  Month in NYC")

pnyc_av
```



year-1.pdf

Let's check shootings by year by borough, to see if such increase was in all NYC or concentrated in some boroughs. First we need a column with the year and then we can group them and find how many shootings happened per year.

```
nyp_by_b <- nyp %>%
  group_by(BORO,Yr) %>%
  summarize(shootings = n(), murders = sum(MURDER_FLAG))
shb <- ggplot(nyp_by_b) +
  geom_point(aes(x=Yr, y=shootings, colour=BORO))
shb +
  geom_line(aes(x=Yr, y=shootings, colour=BORO)) +
  scale_x_continuous(breaks = seq(2006,2021,2)) +
  labs(x = "Year",y = "# of Shootings", title = "Number of yearly Shootings per Borough")
```



The shootings in every borough increased but it seems like the stepper increase was in Brooklyn that already had the highest gun violence rates in the city.

This raised the question of if the shootings per population are also the highest in Brooklyn or because Brooklyn is the borough with the highest population in the city it has the highest number of shootings.

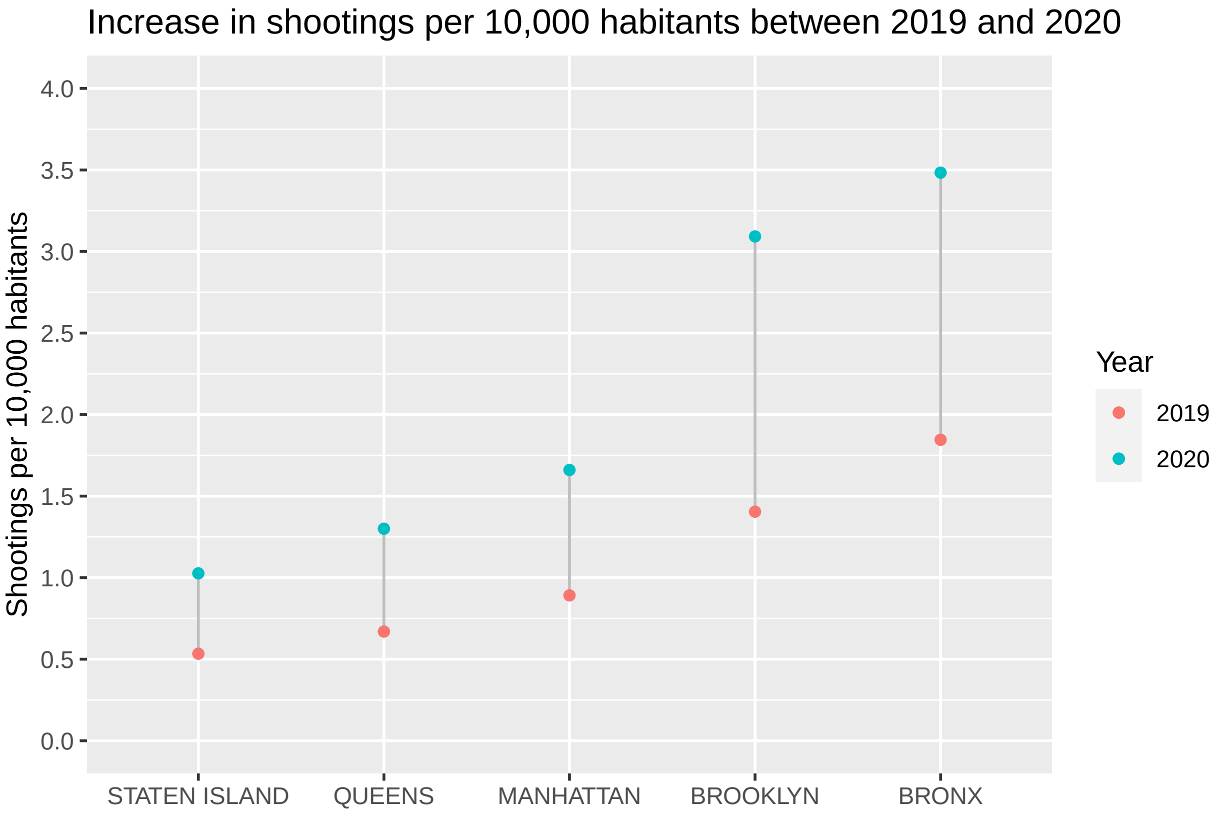
So to give a fair comparison between boroughs I will find the number of shootings per 10,000 habitants. However let's keep in mind that a perpetrator that lives out of a specific borough can shot a gun in the borough.

Because the increase was between 2019 and 2020, we will look at how much the murder rate per borough and borough population increased.

The data of The population by borough comes from https://data.cityofnewyork.us and I will use the data for 2020.

```
#Get NYC population by borough
pop <- read_csv("https://data.cityofnewyork.us/api/views/xywu-7bv9/rows.csv")

#Keep only 2020
population <- pop %>%
  select(Borough,`2020`) %>%
  rename('population'='2020', 'BORO'='Borough') %>%
  filter(BORO!= "NYC Total")%>%
  mutate(BORO = toupper(BORO))
```

```
#From the shootings data keep only 2019-2020 group it by Year and borough and rearange data to have a r
ny1920b <- nyp %>%
  filter(Yr > 2018) %>%
  mutate("Year" = as.factor(Yr)) %>%
  group_by(Year, BORO) %>%
  summarize(shootings = n()) %>%
  pivot_wider(names_from = Year,
              values_from = shootings,names_prefix = 'Y')

#Add population data to shootings data frame
ny1920b <-left_join(ny1920b,population,by = 'BORO')
#Get the number of shootings per 10,000 habitants and reorder data frame
nyper <- ny1920b %>%
  mutate('S_2019_per_10000'=10000*Y2019/population,'S_2020_per_10000'=10000*Y2020/population ) %>%
  rowwise()%>%
  mutate(mymean = mean(c(S_2019_per_10000,S_2020_per_10000))) %>%
  arrange(mymean) %>%
  mutate(Borough = factor(BORO, BORO), Increase =S_2020_per_10000/S_2019_per_10000 )

#Plot data using lolipop
ggplot(nyper)+
  geom_segment(aes(x = Borough, xend = Borough, y =S_2019_per_10000, yend=S_2020_per_10000), color = 'g
  geom_point(aes(x = Borough,  y =S_2019_per_10000, color = factor(2019)))+
  geom_point(aes(x = Borough,  y =S_2020_per_10000, color = factor(2020)))+
  ylim(0,4)+
  scale_y_continuous(breaks= seq(0,4,0.5), limits=c(0,4))+
  labs(x= "",y="Shootings per 10,000 habitants",title="Increase in shootings per 10,000 habitants betwe
```



data-1.pdf

We can see that even Brooklyn is the Borough with the most total number of shootings, the Bronx has more shootings per 10,000 people both in 2019 and 2020.

The model for this data would be that for each Borough the number of shooting of 2020 were 2 times the number of shootings of 2019 in that same borough.

Lets quickly check if gun violence increases during the summer in all boroughs. We will normalize the gun violence per year by the total number of shootings per year in all NYC.

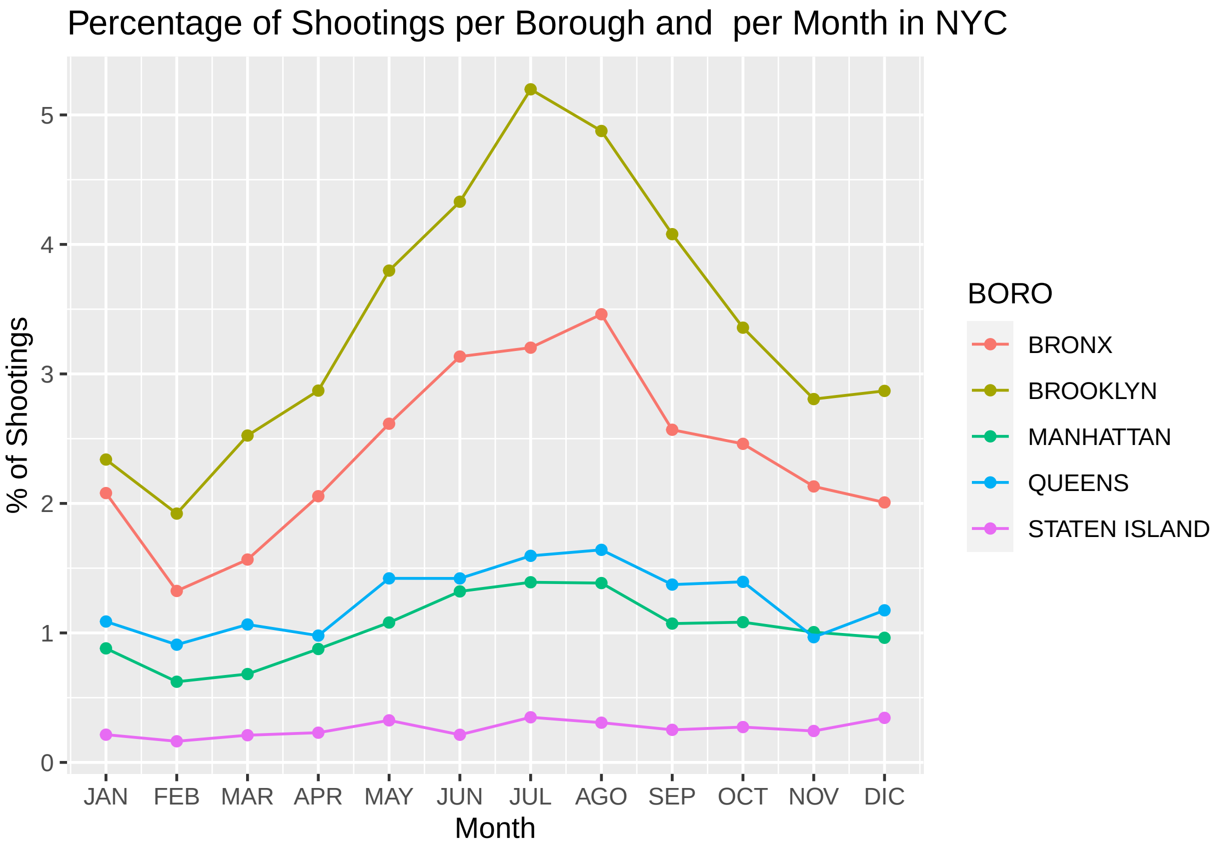
```
yrl_b <- yrl %>%
  rename(total_shootings = shootings, total_murders = murders_gun)


seasonal_b <- nyp %>%
  mutate("Month"= month(OCCUR_DATE)) %>%
  group_by(Yr,Month, BORO) %>%
  summarize(shootings = n(), murders_gun = sum(MURDER_FLAG))%>%
  ungroup()
```

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

```
seasonal_b <- left_join(seasonal_b,yrl_b, by = "Yr")%>%
  mutate(per_shootings = 100*shootings/total_shootings,
         per_murder = 100*murders_gun/total_murders)

avg_year <- seasonal_b %>%
  group_by(BORO,Month)%>%
  summarise(shootings= sum(per_shootings)/n(),
            murders = sum(per_murder)/n())%>%
  ungroup()
```

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

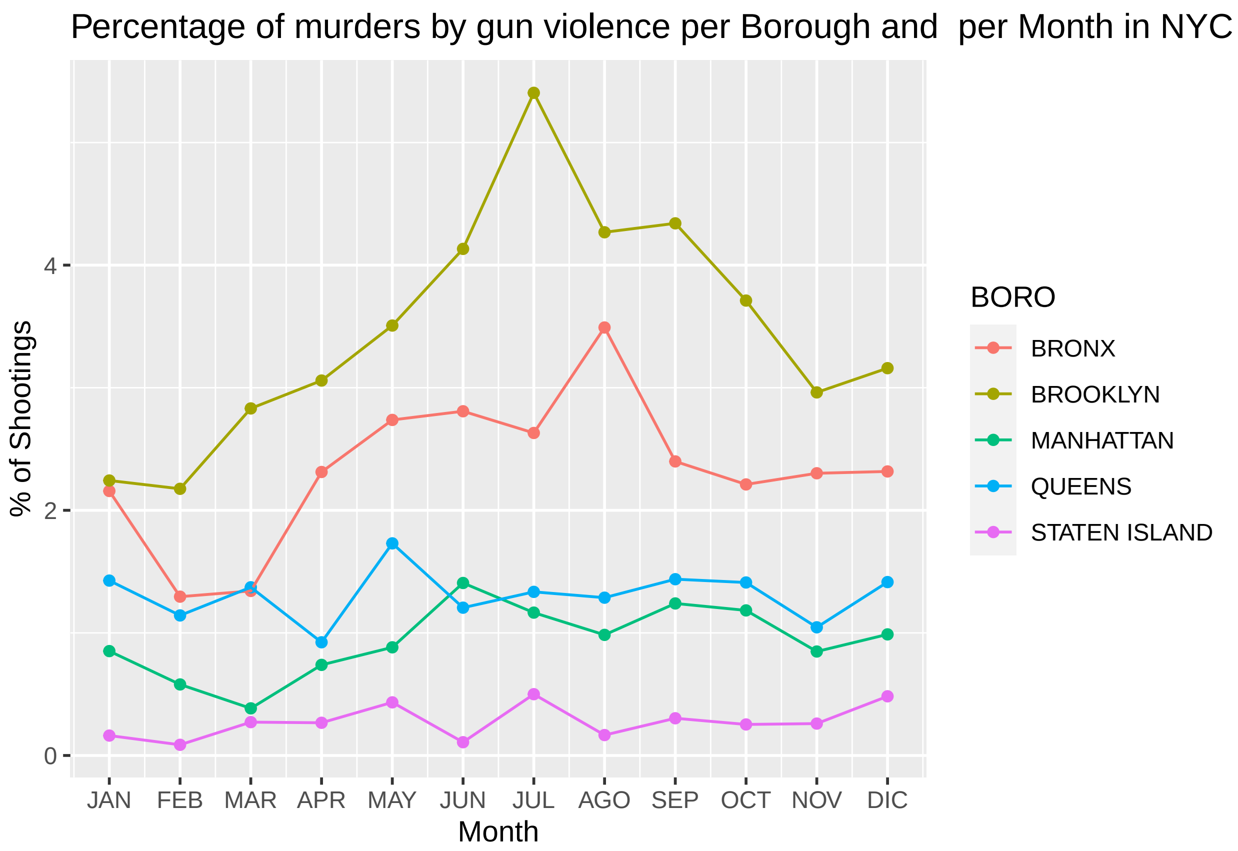
```
ggplot(avg_year, aes(x = Month)) +
  geom_point(aes(y=shootings, colour=BORO)) +
  geom_line(aes( y=shootings, colour=BORO)) + scale_x_continuous(breaks = seq(1,12), labels =c('JAN',"FE
```



boroughs-1.pdf
From the graph we can see that from all the shootings in NYC since 2006 5% happened in July in Brooklyn. And that in all the boroughs but Staten Island there is more shootings in the summer.

Now lets look at the same graph but for murders.

```
ggplot(avg_year, aes(x = Month)) +
  geom_point(aes(y=murders, colour=BORO)) +
  geom_line(aes( y=murders, colour=BORO)) + scale_x_continuous(breaks = seq(1,12), labels =c('JAN',"FEB
```



month-1.pdf

In Brooklyn July is the deadliest month while in the Bronx is August. In the other 3 Boroughs the difference is less clear.

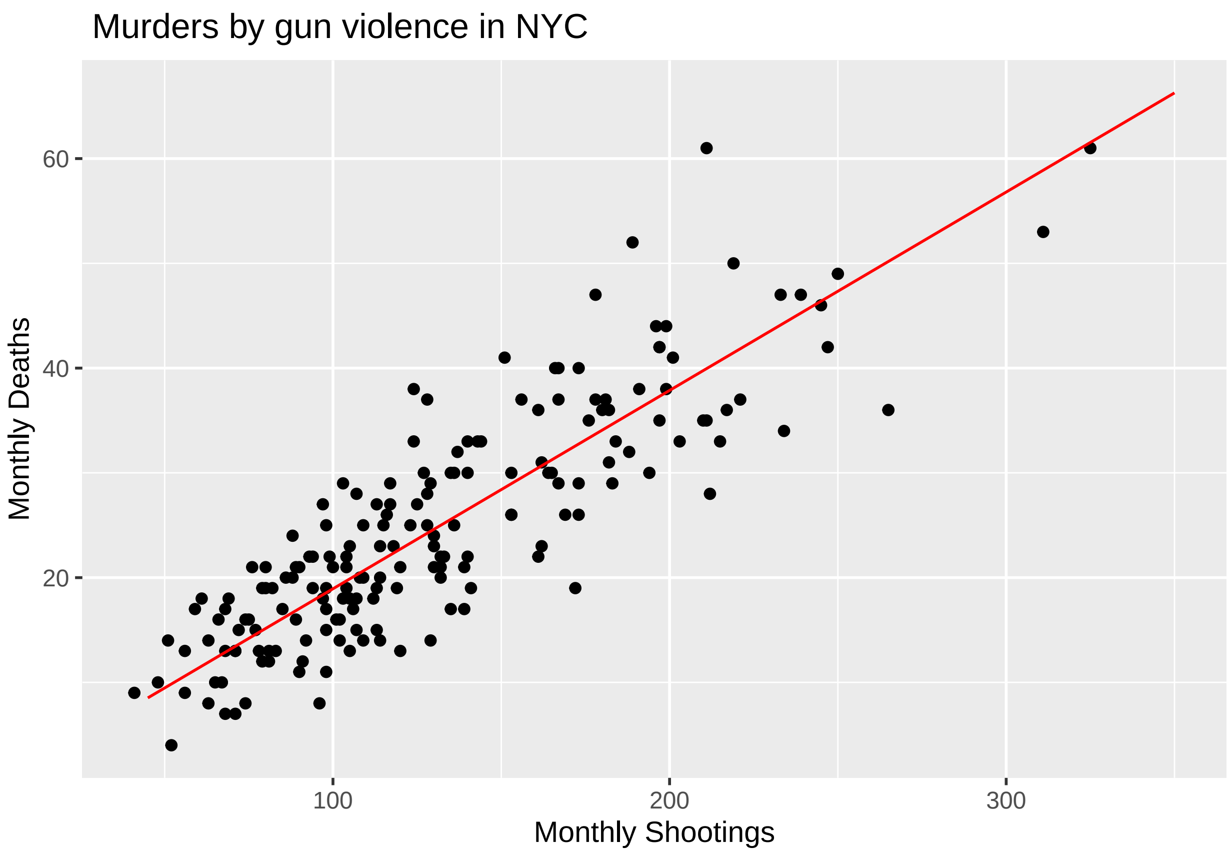
A simple model of murders by gun violence in NYC.

Lets just look to se if there is a linear relation citywide between monthly gun violence and gun murders.

```r
nyp$month_ <- month(nyp$OCCUR_DATE)
monthly <- nyp %>%
  group_by(Yr,month_) %>%
  summarize(shootings = n(), murders_gun = sum(MURDER_FLAG))%>%
  mutate(deadly_per = murders_gun/shootings,
         month_year = my(paste(month_,Yr)))%>%
  ungroup()
```

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

```r
mod = lm(murders_gun~0+shootings,data = monthly)
xmod = seq(45,350)
mod_f = data.frame(xm = xmod,ym = xmod*mod$coefficients)

ggplot(monthly, aes(x = shootings,y = murders_gun))+
  geom_point()+
  geom_line(color='red',data = mod_f, aes(x=xm, y=ym)) +labs(title= " Murders by gun violence in NYC ",
```



mortality-1.pdf

```
summary(mod)
```

```
##
## Call:
## lm(formula = murders_gun ~ 0 + shootings, data = monthly)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.1768  -3.5662   0.2223   4.0623  21.0479
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## shootings  0.18935    0.00301   62.91   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.704 on 179 degrees of freedom
## Multiple R-squared:  0.9567, Adjusted R-squared:  0.9565
## F-statistic:  3958 on 1 and 179 DF,  p-value: < 2.2e-16
```

Conclusions and bias Identification

The number of shootings registered in NYC had an steep increase in 2020 and reached the levels of 10 years ago and double than the values from 2019.

In the last five years there are more shootings during the summer than in any other season. Of the 5 Boroughs Brooklyn has the highest number of shootings per year, but the Bronx has the highest per ca-pita. Staten

Island has the least amount of total shootings and shootings per capita.

I avoided trying to explain why the number of shootings increased either in 2020 or in August of 2016 as I would be probably biased. I also didn't show any information of the gender or race or the perps as I saw in the summary that a lot of it said unknown and from the ones it did have information I don't how such information was collected and if it was biased when collected. My bias might be that I did not look into this information to try to avoid being biased.