Analysis of NYPD Shooting Incident Data

Giri Kunche

5/18/2021

Project Description

Gun violence continues to be one of the major issues in United States. Gun Violence has resulted in tens of thousands of deaths and injuries very year. I have done the analysis of shootings in New York City which is one of the largest cities in United States. The analysis is based on the data provided by NYC Open Data :https://data.cityofnewyork.us .

The data set contains breakdown of every shooting incident that occured in NYC going back to 2006 through the end of 2020. Each record represents a shooting event and includes information about the event like the location, date and time of occurrence, borough where the shooting incident occurred, precinct where the shooting incident occurred, juridiction code, location , preperator's age group,sex and race, victim's sex, age and race , longitude and latitude of the location.

New York City is divided into five borough's : Booklyn, Bronx, Manhattan, Staten Island and Queens. My analysis will explore year on year shooting trend. We will examine patterns of shootings by borough's.

Load Libraries

Load tidyverse and lubridate libraries.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

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

```
library(knitr)
```

Downloading Data

The data is downloaded from https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD. This data is a free public data published by New York City agencies.

```
##Get NYC Shooting Incident Data from https://catalog.data.gov/dataset
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
raw_data <-read_csv(url_in)
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   INCIDENT_KEY = col_double(),
##   OCCUR_DATE = col_character(),
##   OCCUR_TIME = col_time(format = ""),
##   BORO = col_character(),
##   PRECINCT = col_double(),
##   JURISDICTION_CODE = col_double(),
##   LOCATION_DESC = col_character(),
##   STATISTICAL_MURDER_FLAG = col_logical(),
##   PERP_AGE_GROUP = col_character(),
##   PERP_SEX = col_character(),
##   PERP_RACE = col_character(),
##   VIC_AGE_GROUP = col_character(),
##   VIC_SEX = col_character(),
##   VIC_RACE = col_character(),
##   X_COORD_CD = col_number(),
##   Y_COORD_CD = col_number(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

Data Analysis

There are total 19 columns and 23568 rows in the dataset. Data set consist of columns INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, BORO, PRECINCT, JURISDICTION_CODE, LOCATION_DESC, STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, VIC_RACE, X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat.

Remove INCIDENT_KEY, X_COORD_CD, Y_COORD_CD, Latitude, Longitude and Lon_Lat columns as they are not needed for data analysis.
OCCUR_DATA has date of shootings and is in character format. We will convert this column to date format.

Following is the output of NYC data removing above columns.

```
df <- raw_data %>% mutate(OCCUR_DATE=mdy(OCCUR_DATE))
nyc_data <-select(df,-INCIDENT_KEY,-X_COORD_CD,-Y_COORD_CD,-Latitude,-Longitude,-Lon_Lat)
kable(nyc_data[1:5,])
```

OCCUR_DATE	OCCUR_TIME	BORO	PRECINCT	JURISDICTION_CODE	LOCATION_DESC	STATISTICAL_MURDER_FLAG	PERP_AGE_GROUP	PERP_SEX	PERP_RACE	VIC_AGE_GROUP	VIC_SEX	VIC_RACE
2019-08-23	22:10:00	QUEENS	103	0	NA	FALSE	NA	NA	NA	25-44	M	BLACK
2019-11-27	15:54:00	BRONX	40	0	NA	FALSE	<18	M	BLACK	25-44	F	BLACK

OCCUR_DATE	OCCUR_TIME	BORO	PRECINCT	JURISDICTION_CODE	LOCATION_DESC	STATISTICAL_MURDER_FLAG	PERP_AGE_GROUP	PERP_SEX	PERP_RACE	VIC_AGE_GROUP	VIC_SEX	VIC_RACE
2019-02-02	19:40:00	MANHATTAN	23	0	NA	FALSE	18-24	M	WHITE HISPANIC	18-24	M	BLACK HISPANIC
2019-10-24	00:52:00	STATEN ISLAND	121	0	PVT HOUSE	TRUE	25-44	M	BLACK	25-44	F	BLACK
2019-08-22	18:03:00	BRONX	46	0	NA	FALSE	25-44	M	BLACK HISPANIC	18-24	M	BLACK

Summarizing Data

We have summarized data into five groups :
1. group_by_boro - Groups the data by Boro , Year and Number of Shootings.
2. nyc_data_year - Mutates OCCUR_DATE and created new column called Year. Also created column Murder, which is based on STATISTICAL_MURDER_FLAG.
3. shootings_by_year - Summarize shootings and murders by Year.
4. shootings_by_boro - Summarize shootings by Boro
5. murders_by_boro - Summarize murders by Boro

```
group_by_boro <- nyc_data %>%group_by(BORO,OCCUR_DATE) %>%summarise(Shootings=n())
```

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

```
nyc_data_year<-nyc_data%>%mutate(Year=year(OCCUR_DATE),Murder=(STATISTICAL_MURDER_FLAG == TRUE)*1)%>%gr
shootings_by_year<- nyc_data_year %>%group_by(Year) %>%summarise(Murder=sum(Murder),Shootings=n())
shootings_by_boro<-nyc_data_year %>% group_by(Year,BORO) %>% summarise(Shootings=n())
```

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

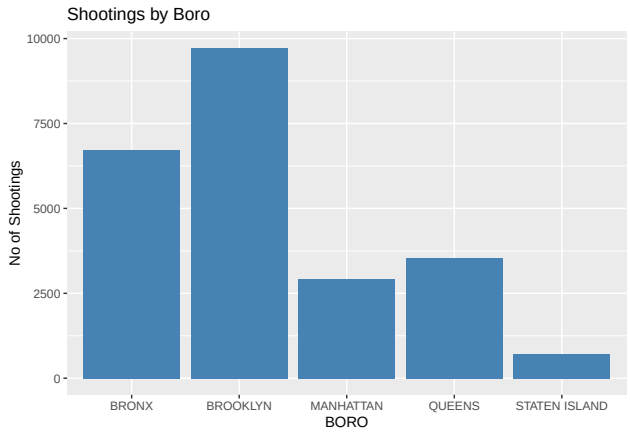
```
murders_by_boro<-nyc_data_year %>% group_by(Year,BORO) %>% summarise(Murder=sum(Murder))
```

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

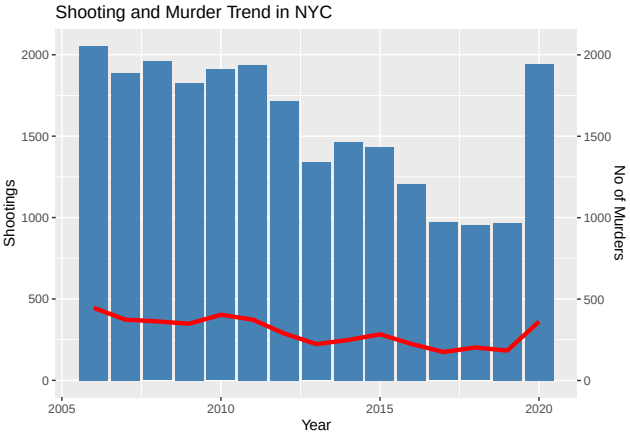
Visualization

Lets visualize the shooting and murder data by Boro's. Below bar plot shows the number of shootings in several boroughs. Brooklyn has the highest shootings recorded followed by Bronx. Shooting and Murder trend is shown in next chart. There is decreasing trend in both shootings and murders since 2006. However, there is a spike in both in 2020.

```
group_by_boro %>% ggplot(aes(x= BORO, y = Shootings))+geom_bar(stat="identity",fill="steelblue")+theme(
```

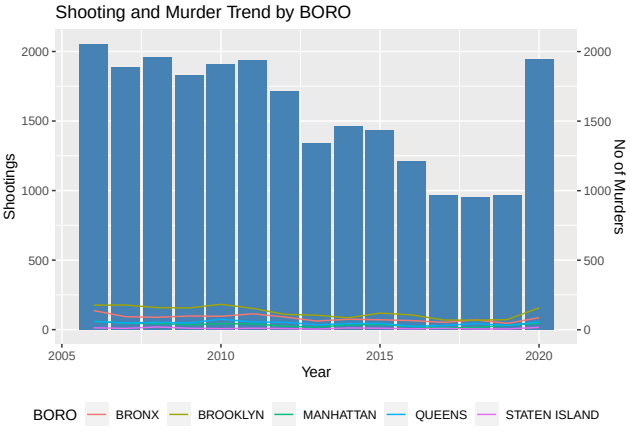



```
ggplot(shootings_by_year,aes(x=Year)) + geom_bar(stat="identity",aes(y=Shootings),fill="steelblue") + g
```

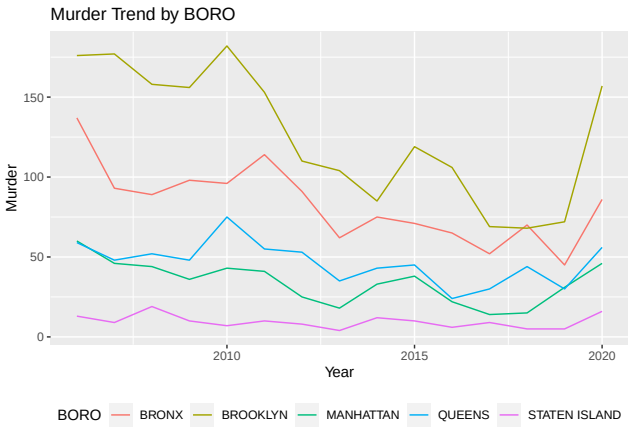


Further drill down of trend by Boro's shows Murder trend by Boro's as well as comparison of Number of Shootings and Murders since 2006.

```
ggplot(shootings_by_boro,aes(x=Year))+geom_bar(stat="identity",aes(y=Shootings),fill="steelblue") + geo
```

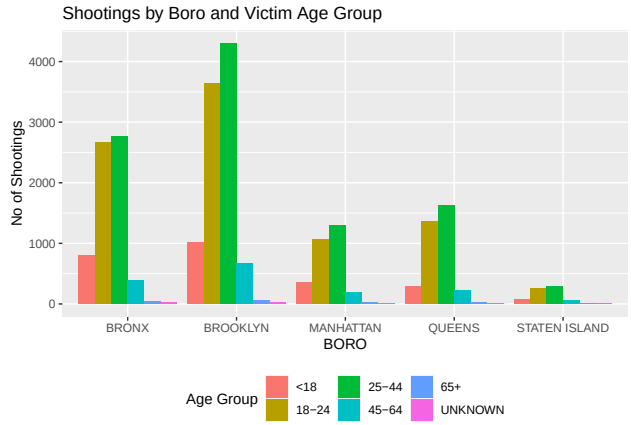



```
ggplot(murders_by_boro,aes(x=Year)) + geom_line(aes(y=Murder,group=BORO,color=BORO))+labs(title="Murder
```

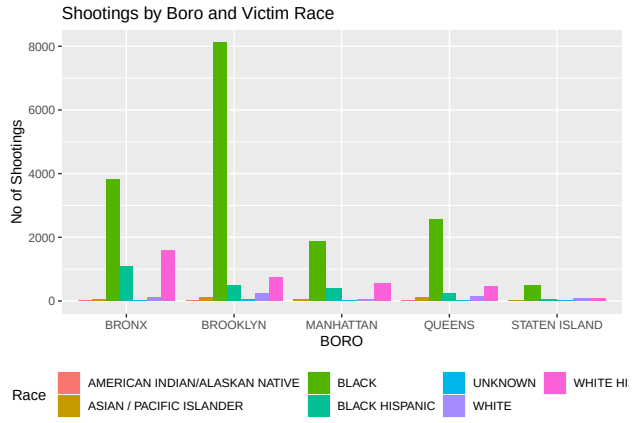


Below chart shows the number of shootings by Boro's and Victim Age Group. Age group 25-44 are in high proportion compared to other age groups. Most of the victims are males.

```
ggplot(nyc_data_year,aes(x=BORO,fill=VIC_AGE_GROUP))+geom_bar(position=position_dodge())+labs(title ="Sh
```




```
ggplot(nyc_data_year,aes(x=BORO,fill=VIC_RACE))+geom_bar(position=position_dodge())+labs(title ="Shootin
```



Modelling

Lets consider how the number of murders are related to number of shootings. To examine this relationship, we will first summarize shooting and murder data from nyc_data_year.
Below is the Shootings and Murders summarized by Year.

```
shooting_murder_analysis<-nyc_data_year%>%group_by(Year)%>%summarise(Murder=sum(Murder),Shootings=n())
kable(shooting_murder_analysis)
```

Year	Murder	Shootings
2006	445	2055
2007	373	1887
2008	362	1958
2009	348	1828
2010	403	1910
2011	373	1939
2012	287	1717
2013	223	1339
2014	248	1464
2015	283	1434
2016	223	1208
2017	174	969
2018	202	951
2019	183	967
2020	361	1942

Next step is to create a linear regressing model and find the relationship between Murders and Shootings. Following is the output of linear regression model.

```
mod<-lm(shooting_murder_analysis$Murder ~ shooting_murder_analysis$Shootings, data=shooting_murder_analy
summary(mod)
```

```
##
## Call:
## lm(formula = shooting_murder_analysis$Murder ~ shooting_murder_analysis$Shootings,
##     data = shooting_murder_analysis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.364 -16.069  -1.058  10.493  45.708
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     -25.86201   27.32092  -0.947    0.361
## shooting_murder_analysis$Shootings   0.20689    0.01688  12.260 1.62e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.52 on 13 degrees of freedom
## Multiple R-squared:  0.9204, Adjusted R-squared:  0.9143
## F-statistic: 150.3 on 1 and 13 DF,  p-value: 1.616e-08
```

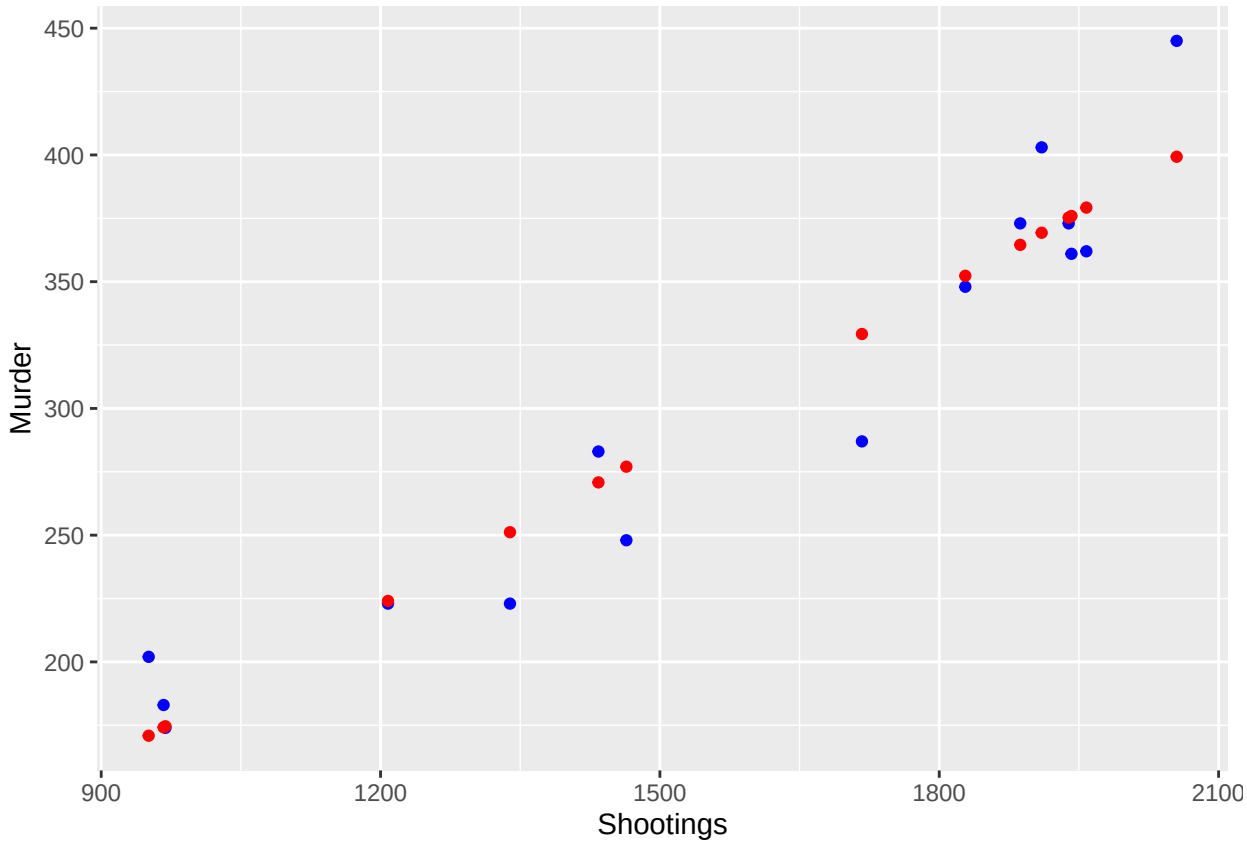
We will add the murder prediction (Murder_Predicted) from linear model to a new data set murder_pred.

```
murder_pred<-shooting_murder_analysis%>% mutate(Murder_Predicted=predict(mod))
kable(murder_pred)
```

Year	Murder	Shootings	Murder_Predicted
2006	445	2055	399.2923
2007	373	1887	364.5351
2008	362	1958	379.2242
2009	348	1828	352.3288
2010	403	1910	369.2936
2011	373	1939	375.2933
2012	287	1717	329.3642
2013	223	1339	251.1607
2014	248	1464	277.0216
2015	283	1434	270.8150
2016	223	1208	224.0584
2017	174	969	174.6122
2018	202	951	170.8882
2019	183	967	174.1984
2020	361	1942	375.9140

We will now plot between the actual murders vs predicted murders and check how will the model fits. The blue dots are the actual murders and the red dots are the predicted murders.
The model does well when there is cluster of data nearby. Overall the model prediction is closer to the actual values.

```
murder_pred %>% ggplot() + geom_point(aes(x=Shootings, y = Murder),color="blue") + geom_point(aes(x=Sho
```



Bias

Only Number of Shootings and Murders are considered to predict the murders. We have left out Perperator Race, Sex, Age as well as location of the shooting. These variables may have influence on the murder or shootings.