5301 NYPD Shooting Incident Report

Anubhav Sharma

11/04/2022

NYPD Shooting Incident Data Report

NYPD Shooting Incident Data Set is a breakdown of shooting incidents 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.

Variables pertaining to the geo-location has been removed along with others which can be used to identify the incidents individually.

Breakdown of this report:

1. Part 1: Tidying up the NYPD Data Set for further Analysis
2. Part 2: Visualizations and Analysis
3. Part 3: Model
4. Part 4: Bias Sources
5. Part 5: Conclusion

Part 1: Tidying up the NYPD Data Set for further Analysis

This Part Contains all the steps carried out during the process aiming to convert the raw data into a data on which further analysis could be done.

The code chunks contain a brief description of the process itself and further efforts to explain important code chunks have been made in this Part for further understanding of the readers.

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.0     v forcats 0.5.1

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

```
library(ggplot2)
library(chron)
```

```
## Warning: package 'chron' was built under R version 4.1.3
```

```
library(ggthemes)
```

```
## Warning: package 'ggthemes' was built under R version 4.1.3
```

In the first step we will import and read the data for

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

Correcting the format of date

```
nypd_data <- nypd_data %>%
  mutate(OCCUR_DATE = chron(dates = OCCUR_DATE))
```

Here we will filter the data to create a data set we could use for further analysis.

```
drops <- c('X_COORD_CD', 'Y_COORD_CD', 'Latitude','Longitude', 'Lon_Lat',
           'OCCUR_TIME', 'INCIDENT_KEY', 'PRECINCT', 'JURISDICTION_CODE')
nypd_data <- nypd_data[, !(names(nypd_data) %in% drops)]
```

Note: Columns related to Geo-Locations and time of the incidents have been removed to ensure anonymity. Others are deemed as not crucial for this analysis.

```
summary(nypd_data)
```

```
##    OCCUR_DATE           BORO           LOCATION_DESC
##  Min.   :01/01/06   Length:23585       Length:23585
##  1st Qu.:12/31/08   Class :character   Class :character
##  Median :02/27/12   Mode  :character   Mode  :character
##  Mean   :10/05/12
##  3rd Qu.:03/02/16
##  Max.   :12/31/20
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Length:23585            Length:23585       Length:23585
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##
##
##
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:23585       Length:23585       Length:23585       Length:23585
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```
nypd_data <- nypd_data %>%
  filter(PERP_AGE_GROUP != "",PERP_AGE_GROUP != "UNKNOWN", PERP_AGE_GROUP !="940",
         PERP_AGE_GROUP !="1020", PERP_AGE_GROUP !="224"
         , LOCATION_DESC != "", LOCATION_DESC != "NONE", PERP_RACE != "UNKNOWN",
         VIC_RACE != "UNKNOWN", PERP_SEX != "U", VIC_AGE_GROUP != "UNKNOWN")
```

Note: In the above code chunk. Unknowns and blanks have been removed along with outliers in the columns where wrong data was recorded in the data set.

Creating a binary murder column so that we could examine murders resulting from the shootinf incidents in Part 2.

```
nypd_data <- nypd_data %>% rename(DATE = "OCCUR_DATE", LOCATION = "BORO",
                                  MURDER_FLAG = "STATISTICAL_MURDER_FLAG") %>%
  mutate(MURDER_FLAG = ifelse(MURDER_FLAG == "true", 1,0))
```

```
nypd_data<- nypd_data %>% mutate(DATE_POSIXct = as.POSIXct(DATE, format = "%Y/%m/%d")) %>%
  mutate(YEAR = format(DATE_POSIXct, format = "%Y"),
         MONTH = format(DATE_POSIXct, format = "%m"),
         DAY = format(DATE_POSIXct, format = "%d"))
```

Note: Above column is necessary to obtain 'YEAR' column used in the further analysis in Part 2 and Part 3.

This is an additional step done with the aim to create a better order of the column. It can be skipped.

```
column_order = c('DATE', 'YEAR', 'MONTH', 'DAY', 'LOCATION', 'LOCATION_DESC',
                 'MURDER_FLAG','PERP_AGE_GROUP', 'PERP_SEX', 'PERP_RACE', 'VIC_AGE_GROUP','VIC_SEX','VI
nypd_data <- nypd_data[,column_order]

nypd_data <- nypd_data %>% arrange(DATE)
```

```
summary(nypd_data)
```

```
##       DATE              YEAR              MONTH               DAY
##  Min.   :01/01/06   Length:5341        Length:5341        Length:5341
##  1st Qu.:09/19/08   Class :character   Class :character   Class :character
##  Median :07/05/11   Mode  :character   Mode  :character   Mode  :character
##  Mean   :03/18/12
##  3rd Qu.:04/08/15
##  Max.   :12/29/20
##    LOCATION         LOCATION_DESC       MURDER_FLAG     PERP_AGE_GROUP
##  Length:5341        Length:5341        Min.   :0.0000   Length:5341
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :0.0000   Mode  :character
##                                        Mean   :0.2691
##                                        3rd Qu.:1.0000
##                                        Max.   :1.0000
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:5341        Length:5341        Length:5341        Length:5341
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
##
##
##
##    VIC_RACE          DATE_POSIXct
##  Length:5341        Min.   :2006-01-01 05:30:00
##  Class :character   1st Qu.:2008-09-19 05:30:00
##  Mode  :character   Median :2011-07-05 05:30:00
##                     Mean   :2012-03-18 16:28:55
##                     3rd Qu.:2015-04-08 05:30:00
##                     Max.   :2020-12-29 05:30:00
```

Modified NYPD Data Set has 5341 observations and 14 variables. This Data Set will be used to create different Visualizations and linear model.

Part 2: Visualizations And Analysis

In Part 2, we have taken seven major question into considerations:

1. In which day and month most of the incidents happen? Can we identify a visible trend in the findings?
2. In which day and month most of the murders happen? Can we identify a visible trend in the findings?
3. At which type of locations most of incidents and deaths happen in the New York?
4. What is the distribution of Incidents leading to Murders across the different regions in New York? That is to find which region reported most murders and which region reported the least.
5. What is the trend of Shooting Incidents and Murder Incidents across the Years?
6. What is the proportion of different genders falling under the categories of Victims and Perpetrators across the Years?
7. What is the proportion of different races falling under the categories of Victims and Perpetrators across the Years?

This is a preliminary analysis finding to find which regions have highest number of incidents across the period per month. While Bronx took top spot for the month of January Brooklyn is the region with most number of incidents in rest of the 11 months.

```
nypd_data %>%
  group_by(MONTH) %>%
  count(LOCATION) %>%
  rename(Count = n) %>%
  top_n(1) %>%
  arrange(desc(Count))
```

```
## Selecting by Count
```

```
## # A tibble: 12 x 3
## # Groups:   MONTH [12]
##    MONTH LOCATION Count
##    <chr> <chr>    <int>
##  1 08    BROOKLYN   228
##  2 06    BROOKLYN   216
##  3 07    BROOKLYN   211
##  4 04    BROOKLYN   185
##  5 05    BROOKLYN   184
```

```
##  6 09    BROOKLYN   177
##  7 11    BROOKLYN   177
##  8 10    BROOKLYN   169
##  9 01    BRONX      165
## 10 03    BROOKLYN   149
## 11 12    BROOKLYN   137
## 12 02    BROOKLYN    93
```

```
nypd_data %>%
  group_by(DAY) %>%
  count(LOCATION) %>%
  rename(Count = n) %>%
  top_n(1) %>%
  arrange(desc(Count))
```

```
## Selecting by Count
```

```
## # A tibble: 32 x 3
## # Groups:   DAY [31]
##    DAY   LOCATION Count
##    <chr> <chr>    <int>
##  1 07    BROOKLYN   104
##  2 25    BROOKLYN    89
##  3 21    BROOKLYN    88
##  4 18    BROOKLYN    80
##  5 01    BROOKLYN    77
##  6 10    BROOKLYN    77
##  7 22    BROOKLYN    75
##  8 19    BROOKLYN    74
##  9 02    BROOKLYN    73
## 10 11    BROOKLYN    71
## # ... with 22 more rows
```

```
nypd_data %>%
  group_by(MONTH) %>%
  count(MONTH) %>%
  rename(incidents = n) %>%
  arrange(desc(incidents))
```

```
## # A tibble: 12 x 2
## # Groups:   MONTH [12]
##    MONTH incidents
##    <chr>     <int>
##  1 08          558
##  2 07          516
##  3 06          501
##  4 05          486
##  5 09          475
##  6 10          452
##  7 01          437
##  8 04          426
##  9 11          405
```

```
## 10 03          394
## 11 12          387
## 12 02          304
```

```
month_incidents <- nypd_data %>%
  group_by(MONTH) %>%
  count(MONTH) %>%
  rename(incidents = n) %>%
  arrange(desc(incidents))
```

```
month_incidents %>%
  ggplot(aes(x = MONTH, y = incidents, group = 1))+
  geom_point()+
  geom_line()+
  theme_stata()+
  xlab("Months")+
  ggtitle("Shootings across the months")
```

From the above table and graphs we can see that numbers of incidents increase from February in a linear fashion before they reach their peak in August and starts falling down again. There can be many reasons why there exists a seasonal cycle like this. We need data from other sources to find out the conclusive reasons why this trend is happening in New York.

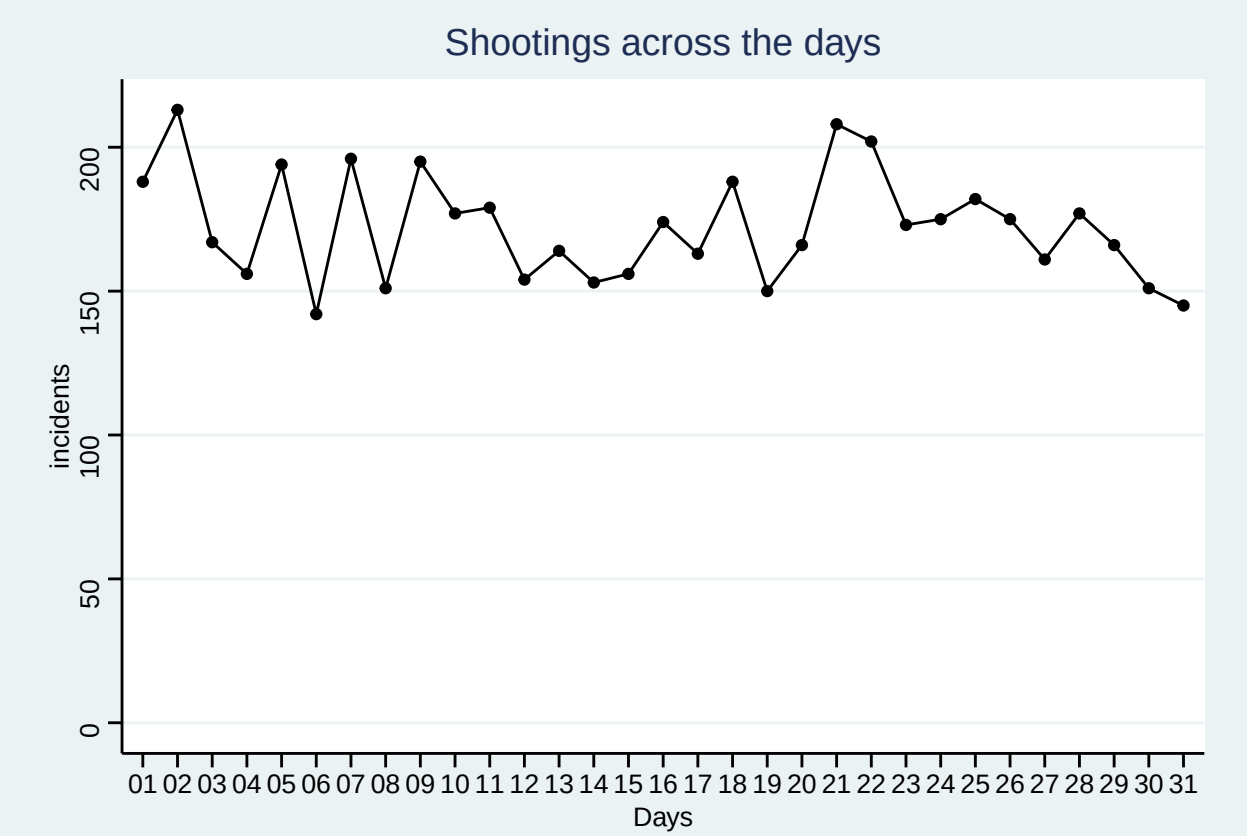
```
nypd_data %>%
  group_by(DAY) %>%
```

```
  count(DAY) %>%
  rename(incidents = n) %>%
  arrange(desc(incidents))
```

```
## # A tibble: 31 x 2
## # Groups:   DAY [31]
##    DAY   incidents
##    <chr>     <int>
##  1 02          213
##  2 21          208
##  3 22          202
##  4 07          196
##  5 09          195
##  6 05          194
##  7 01          188
##  8 18          188
##  9 25          182
## 10 11          179
## # ... with 21 more rows
```

```
day_incidents <- nypd_data %>%
  group_by(DAY) %>%
  count(DAY) %>%
  rename(incidents = n) %>%
  arrange(desc(incidents))
```

```
day_incidents %>%
  ggplot(aes(x = DAY, y = incidents, group = 1))+
  geom_point()+
  geom_line()+
  theme_stata()+
  xlab("Days")+
  ggtitle("Shootings across the days")+
  expand_limits(y=0)
```



From the above table and graph we can see that unlike months analysis there does not exist a clearly visible trend. Nonetheless, we can see that following the day when number of incidents go beyond 170 in general number of incidents dip. Due to the lack of the data we can't know why a few days have higher number of incidents than others and why there is a dip in number of incidents after it goes beyond 170 in general.

```
nypd_data %>%
  group_by(MONTH) %>%
  filter(MURDER_FLAG == 1) %>%
  count(MONTH) %>%
  rename(incidents = n) %>%
  arrange(desc(incidents))
```

```
## # A tibble: 12 x 2
## # Groups:   MONTH [12]
##    MONTH incidents
##    <chr>     <int>
##  1 09          143
##  2 08          141
##  3 07          137
##  4 05          133
##  5 10          125
##  6 06          123
##  7 01          119
##  8 04          119
##  9 12          110
## 10 11          103
```

```
## 11 03           95
## 12 02           89
```

```
month_deaths <- nypd_data %>%
    group_by(MONTH) %>%
  filter(MURDER_FLAG == 1) %>%
  count(MONTH) %>%
  rename(deaths = n) %>%
  arrange(desc(deaths))
```

```
month_deaths %>%
  ggplot(aes(x = MONTH, y = deaths, group = 1))+
  geom_point()+
  geom_line()+
  theme_stata()+
  xlab("Months")+
  ggtitle("Deaths due to shooting across the months")
```

The above table and the graph is quite similar to the the month_incidents data set with both increasing from the month of February and reaching the peak in September and August respectively. That being said month_deaths is reaching the peak in September when months_incidents starting to fell creating the month of Spetember only outlier where deaths are increasing even though shooting incidents are decreasing.

Also, like incidents, deaths follow a seasonal pattern and to analyse it further we would need data from other sources which could explain us why there exists a seasonal cycle for shooting incidents and deaths.q

The above relationship between incidents and deaths suggest existence of a linear relationship between them. We will look further into this in Part 3: Model.

Now we will find the days on with most number of deaths happened and a graph for that.

```
nypd_data %>%
  group_by(DAY) %>%
  filter(MURDER_FLAG == 1) %>%
  count(DAY) %>%
  rename(deaths = n) %>%
  arrange(desc(deaths))
```

```
## # A tibble: 31 x 2
## # Groups:   DAY [31]
##    DAY   deaths
##    <chr>  <int>
##  1 02        69
##  2 24        64
##  3 05        61
##  4 18        59
##  5 21        59
##  6 22        57
##  7 06        54
##  8 07        54
##  9 01        51
## 10 09        51
## # ... with 21 more rows
```

```
day_deaths <- nypd_data %>%
  group_by(DAY) %>%
  filter(MURDER_FLAG == 1) %>%
  count(DAY) %>%
  rename(deaths = n) %>%
  arrange(desc(deaths))
```

```
day_deaths %>%
  ggplot(aes(x = DAY, y = deaths, group = 1))+
  geom_point()+
  geom_line()+
  theme_stata()+
  xlab("Days")+
  ggtitle("Deaths due to shooting across the days")+
  expand_limits(y=0)
```

Deaths across the days mirrors the incidents across the days. While again we can see a decrease in deaths when the number goes beyond 45 previous day. And we will need more data to find why a few days record more deaths and incidents than the others.

Now we will find out which location_descriptions recorded most number of incidents and murders.

```
nypd_data %>%
  group_by(LOCATION_DESC) %>%
  count(LOCATION_DESC, sort = TRUE) %>%
  rename(incidents = n) %>%
  ungroup() %>%
  mutate(incident_percent = incidents/sum(incidents)*100) %>%
  select(LOCATION_DESC, incident_percent)
```

```
## # A tibble: 35 x 2
##    LOCATION_DESC             incident_percent
##    <chr>                                <dbl>
##  1 MULTI DWELL - PUBLIC HOUS           40.4
##  2 MULTI DWELL - APT BUILD             29.2
##  3 PVT HOUSE                            8.82
##  4 GROCERY/BODEGA                       6.03
##  5 BAR/NIGHT CLUB                       5.77
##  6 RESTAURANT/DINER                     1.93
##  7 COMMERCIAL BLDG                      1.69
##  8 BEAUTY/NAIL SALON                    1.03
##  9 FAST FOOD                            0.824
## 10 SOCIAL CLUB/POLICY LOCATI            0.618
```

```
## # ... with 25 more rows
```

```
nypd_data %>%
  group_by(LOCATION_DESC) %>%
  filter(MURDER_FLAG == 1) %>%
  count(LOCATION_DESC, sort = TRUE) %>%
  rename(deaths = n) %>%
  ungroup() %>%
  mutate(deaths_percent = deaths/sum(deaths)*100) %>%
  select(LOCATION_DESC, deaths_percent)
```

```
## # A tibble: 29 x 2
##    LOCATION_DESC             deaths_percent
##    <chr>                              <dbl>
##  1 MULTI DWELL - APT BUILD           33.8
##  2 MULTI DWELL - PUBLIC HOUS         33.8
##  3 PVT HOUSE                         11.8
##  4 BAR/NIGHT CLUB                     5.98
##  5 GROCERY/BODEGA                     4.94
##  6 COMMERCIAL BLDG                    1.67
##  7 HOTEL/MOTEL                        0.974
##  8 RESTAURANT/DINER                   0.974
##  9 BEAUTY/NAIL SALON                  0.905
## 10 LIQUOR STORE                       0.835
## # ... with 19 more rows
```

From the above table we can see that 69.5% of shooting incidents and 67.6% of the deaths resulting from them in the New York happened in multi-dwelling buildings Apartment Buildings and Public Housing units. Interestingly enough public places like bars, clubs, and stores only amount to 21.7% of the shooting incidents and 20.7% of the deaths meaning that around 80% of the deaths and incidents happen in the residential areas implying that in New York residential areas are more suspectible to such events.

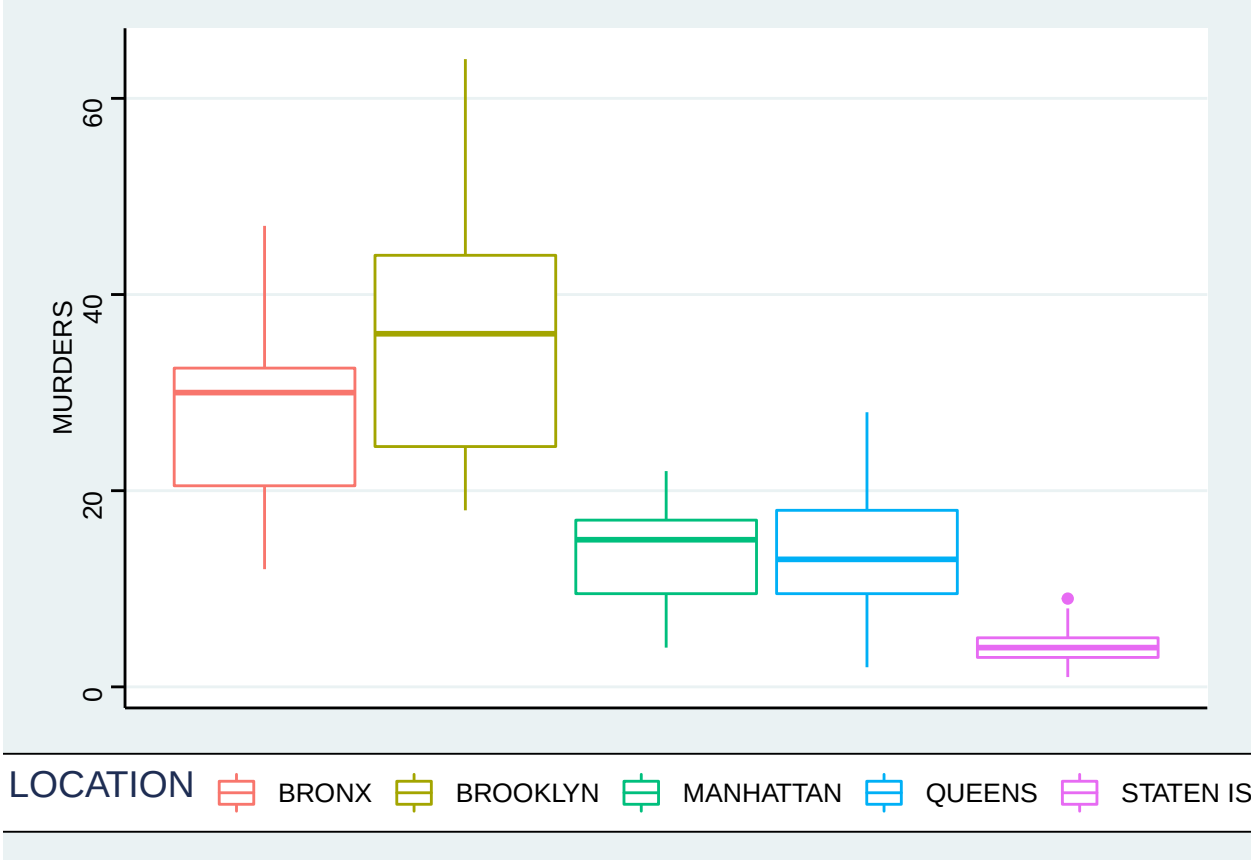
This can be happening because of various reasons like maybe 80% of the New York is residential area or some other reason we can not know about from the data we have. We need data from other sources to find out why 80% of these events are happening in the residential areas.

Now we will find out the distribution of the incidents leading to murders across the different regions/locations of the New York.

```
nypd_murders <- nypd_data %>%
  group_by(YEAR, LOCATION) %>% summarize(MURDERS = sum(MURDER_FLAG))
```

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

```
nypd_murders %>%
  ggplot(aes( y = MURDERS,  color = LOCATION))+
  geom_boxplot()+
  theme_stata()+
  theme(axis.ticks.x = element_blank(),
        axis.text.x = element_blank())
```



From this we can see that Brooklyn recorded highest record of murders during the period. In working below we can see that Brooklyn has a mean of 36.2 Murders reported in each year while Staten Island has a mean of only 4.2 Murders per year.

Brooklyn also have the highest standard deviation as we can see from the box plot, and from the working below.

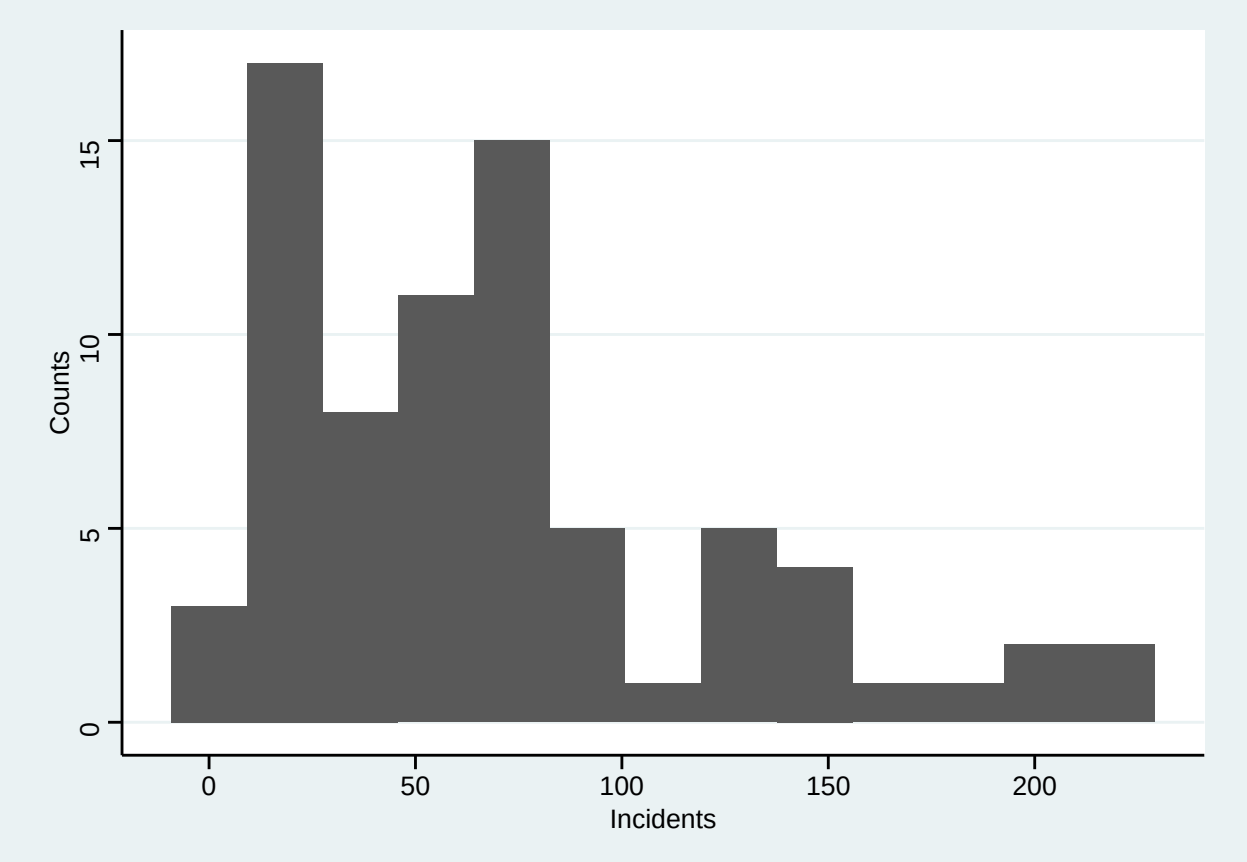
```
nypd_murders %>% filter(LOCATION == "BROOKLYN") %>%
  summarize(mean_years = mean(MURDERS)) %>%
  summarize(mean_brooklyn =  mean(mean_years),
  sd_brooklyn = sd(mean_years))
```

```
## # A tibble: 1 x 2
##   mean_brooklyn sd_brooklyn
##           <dbl>       <dbl>
## 1          35.9        15.3
```

```
nypd_murders %>% filter(LOCATION == "STATEN ISLAND") %>%
  summarize(mean_years = mean(MURDERS)) %>%
  summarize(mean_staten_island =  mean(mean_years),
            sd_staten_island = sd(mean_years))
```

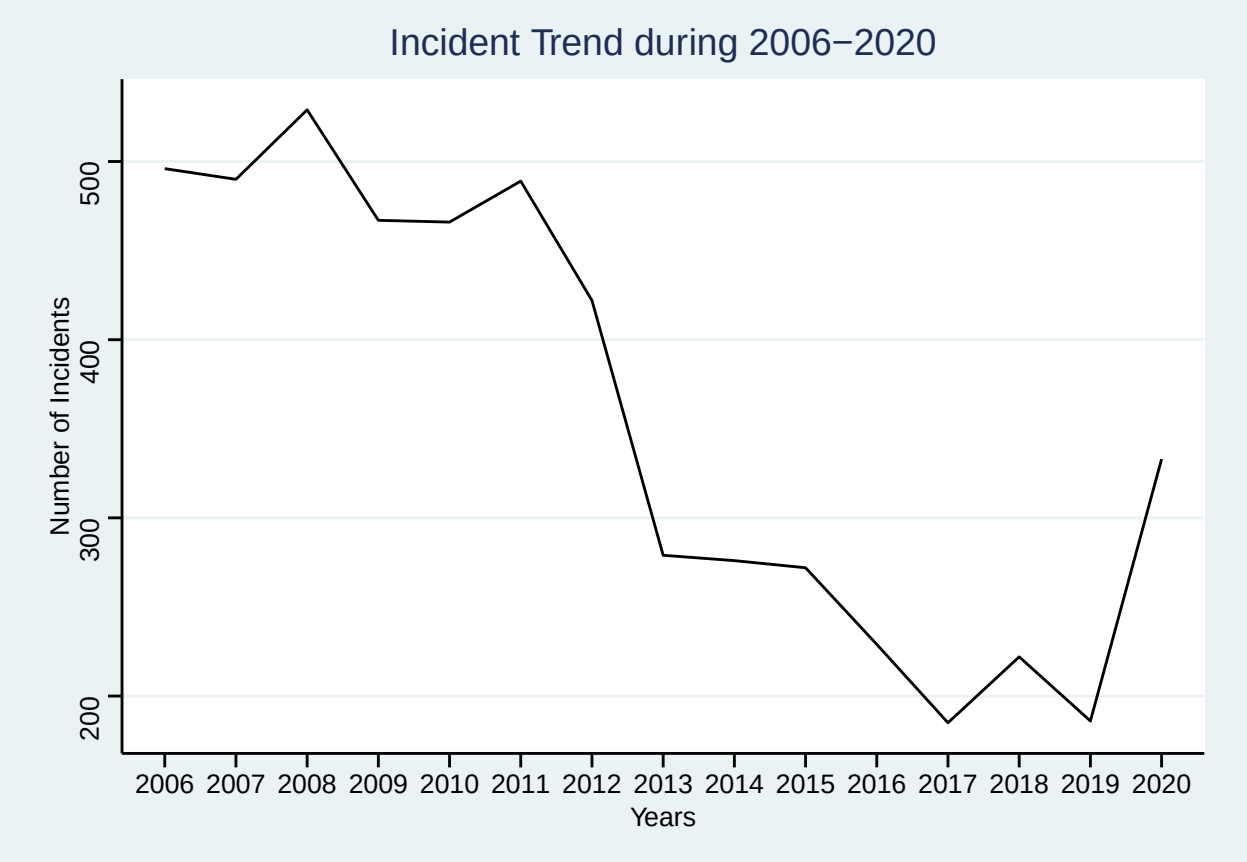
```
## # A tibble: 1 x 2
##   mean_staten_island sd_staten_island
##                <dbl>            <dbl>
## 1                4.2             2.21
```

```r
nypd_count <- nypd_data %>%
  group_by(YEAR, LOCATION) %>%
  count()

ggplot(nypd_count, aes(x = n))+
  geom_histogram(bins = 13)+
  theme_stata()+
  xlab("Incidents")+
  ylab("Counts")
```



The above histogram shows that the distributions of the Shooting Incidents is right skewed, this is an optimistic finding telling that most of the time each location record less than 100 incidents each year.

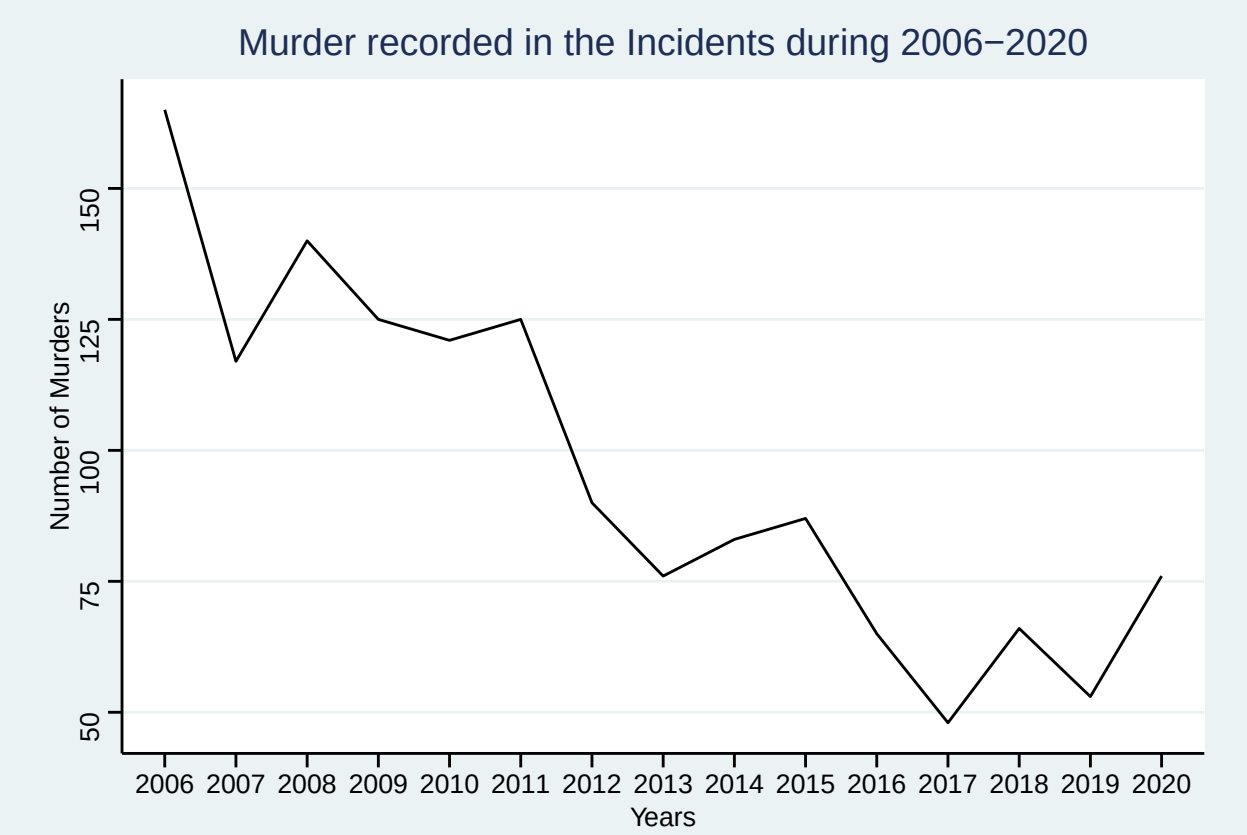
```r
nypd_count[,c('YEAR','n')] %>% summarize(n = sum(n)) %>%
  ggplot(aes(y = n, x = YEAR, group = 1)) +
  geom_line()+
  theme_stata()+
  ggtitle("Incident Trend during 2006-2020")+
  xlab("Years")+
  ylab("Number of Incidents")
```



By looking at this trend like we can see that while incident rates have been falling across the Years in the New York a sudden spike has been recorded in Year 2020.

While we don't have appropriate data to infer the reasons for this, based on the general understanding and similar analysis in the past, this spike can be attributed to the various factors like COVID Pandemic, Rising Inflation, Lack of Optimism among the Public in general, and several other Macro and Micro Level factors.

```
nypd_murders[,c('YEAR','MURDERS')] %>% summarize(n = sum(MURDERS)) %>%
  ggplot(aes(y = n, x = YEAR, group = 1)) +
  geom_line()+
  theme_stata()+
  ggtitle("Murder recorded in the Incidents during 2006-2020")+
  xlab("Years")+
  ylab("Number of Murders")
```



Just like incident trend, the Murders also spiked in 2020. We believe that this also happened due to the same reasons, concrete analysis can be made with data from other sources.

Besides this, from this we can see that there is a linear trend between the Murder Rates and Incident Rates. This relationship will be further analysed in Part 3: Model.

```
x1 = nypd_data %>%
  select(VIC_AGE_GROUP, VIC_SEX, VIC_RACE) %>% group_by(VIC_AGE_GROUP) %>%
  count() %>%
  rename(Count_VIC = n, Age_Group = VIC_AGE_GROUP)
y1 =nypd_data %>%
  select(PERP_AGE_GROUP, PERP_SEX, VIC_RACE) %>% group_by(PERP_AGE_GROUP) %>%
  count() %>% rename(Count_Perp = n)
x1
```

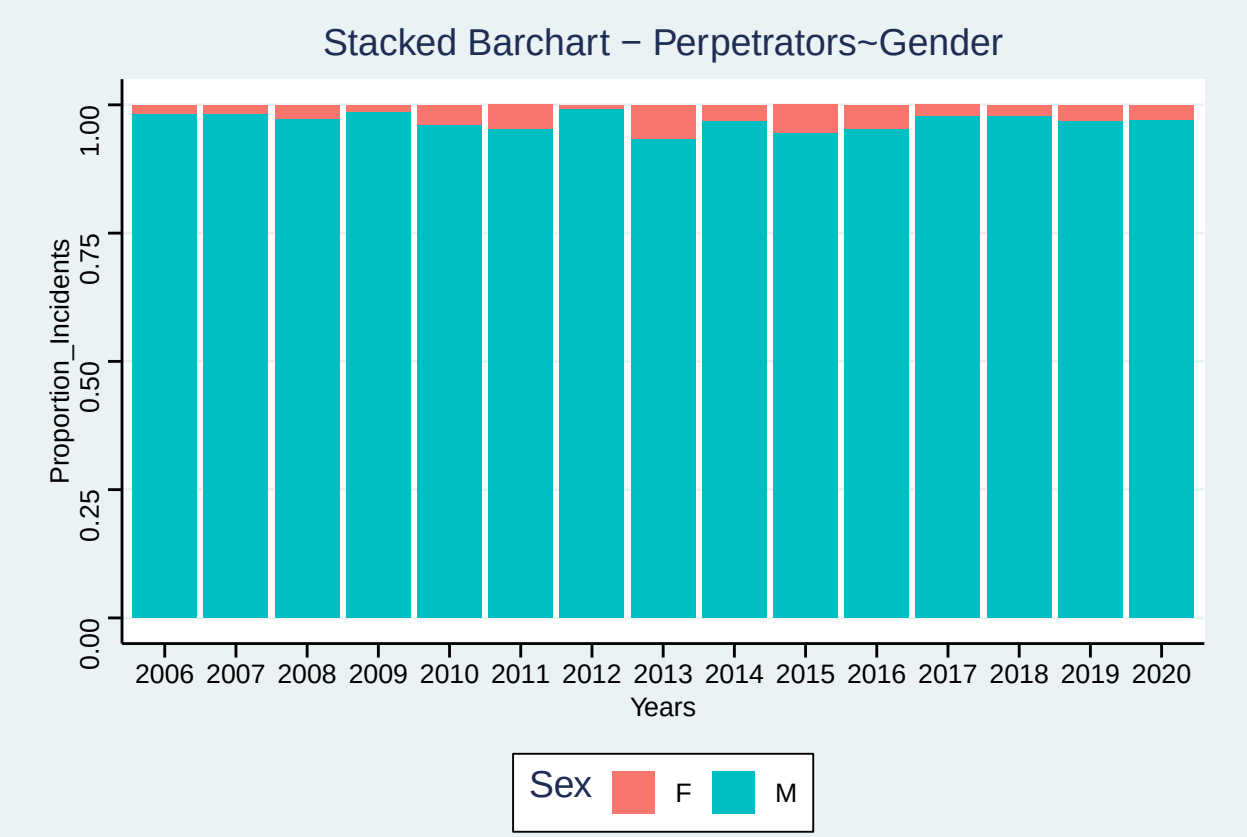
```
## # A tibble: 5 x 2
## # Groups:   Age_Group [5]
##   Age_Group Count_VIC
##   <chr>         <int>
## 1 <18             567
## 2 18-24          1910
## 3 25-44          2376
## 4 45-64           435
## 5 65+              53
```

```
table1 = cbind(x1, y1[,'Count_Perp'])
table1
```

```
## # A tibble: 5 x 3
## # Groups:   Age_Group [5]
##   Age_Group Count_VIC Count_Perp
##   <chr>         <int>      <int>
## 1 <18             567        547
## 2 18-24          1910       2385
## 3 25-44          2376       2144
## 4 45-64           435        229
## 5 65+              53         36
```

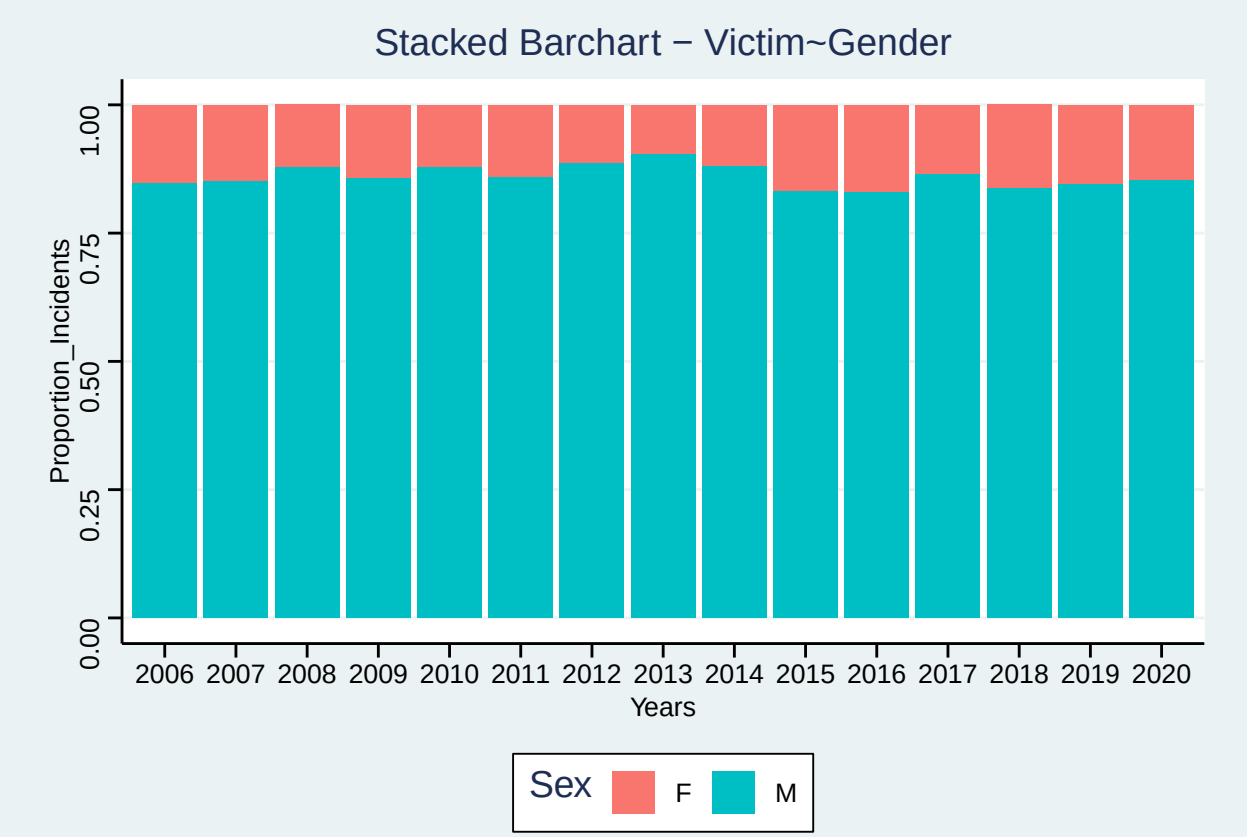
Table 1 is showing the number of Victims and Perpetrators across the Age Group. We can see that for the Age Group above 44 the number of Victim exceeds the number of Perpetrators with almost twice the numbers while the Age Group of 18-24 has highest numbers of Perpetrators.

```
year1 = c(rep("2006",2), rep("2007",2), rep("2008",2), rep("2009",2), rep("2010",2)
          ,rep("2011",2) , rep("2012",2), rep("2013",2), rep("2014",2), rep("2015",2)
          ,rep("2016",2) , rep("2017",2), rep("2018",2), rep("2019",2), rep("2020",2))

table2 <- nypd_data %>%
  select(YEAR, PERP_RACE, PERP_SEX) %>% group_by(YEAR, PERP_SEX) %>%
  count() %>%
  rename(Counts = n)

Sex <- rep(c("F","M"),15)

ggplot(table2, aes(fill=Sex, y=Counts, x=year1)) +
  geom_bar(position="fill", stat="identity")+
  theme_stata()+
  ggtitle('Stacked Barchart - Perpetrators~Gender')+
  xlab("Years")+
  ylab('Proportion_Incidents')
```



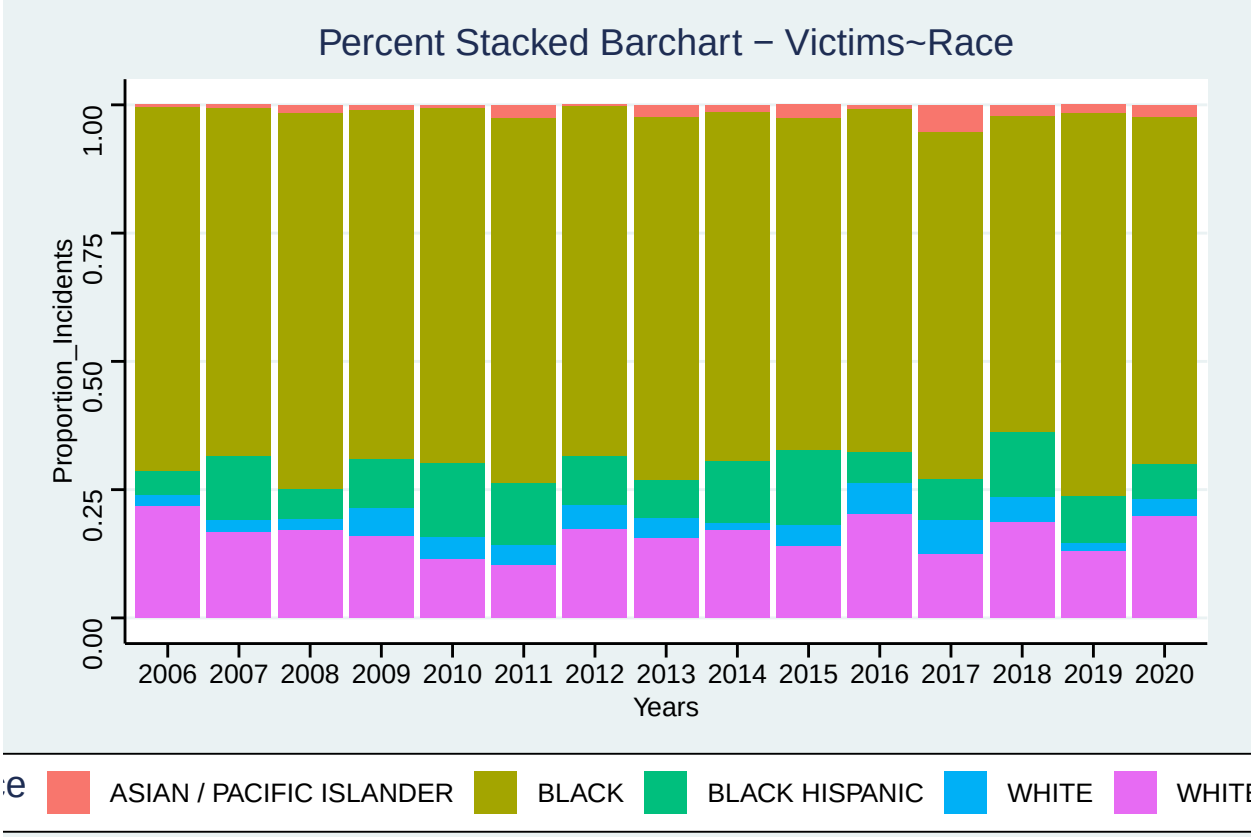
The above Bar Plot was made to find the proportion of Gender for the Perpetrator Group across the year. In all the years, more than 90% of the perpetrators were Men.

```
table3 <- nypd_data %>%
  select(YEAR, VIC_RACE, VIC_SEX) %>% group_by(YEAR, VIC_SEX) %>%
  count() %>%
  rename(Counts = n)

ggplot(table3, aes(fill=Sex, y=Counts, x=year1)) +
  geom_bar(position="fill", stat="identity")+
  theme_stata()+
  ggtitle('Stacked Barchart - Victim~Gender')+
  xlab("Years")+
  ylab('Proportion_Incidents')
```



From the above visual we can see that while again Men are the major targets of the incidents proportion of Women targeted far exceeds their proportion of being the perpetrators of the incidents.

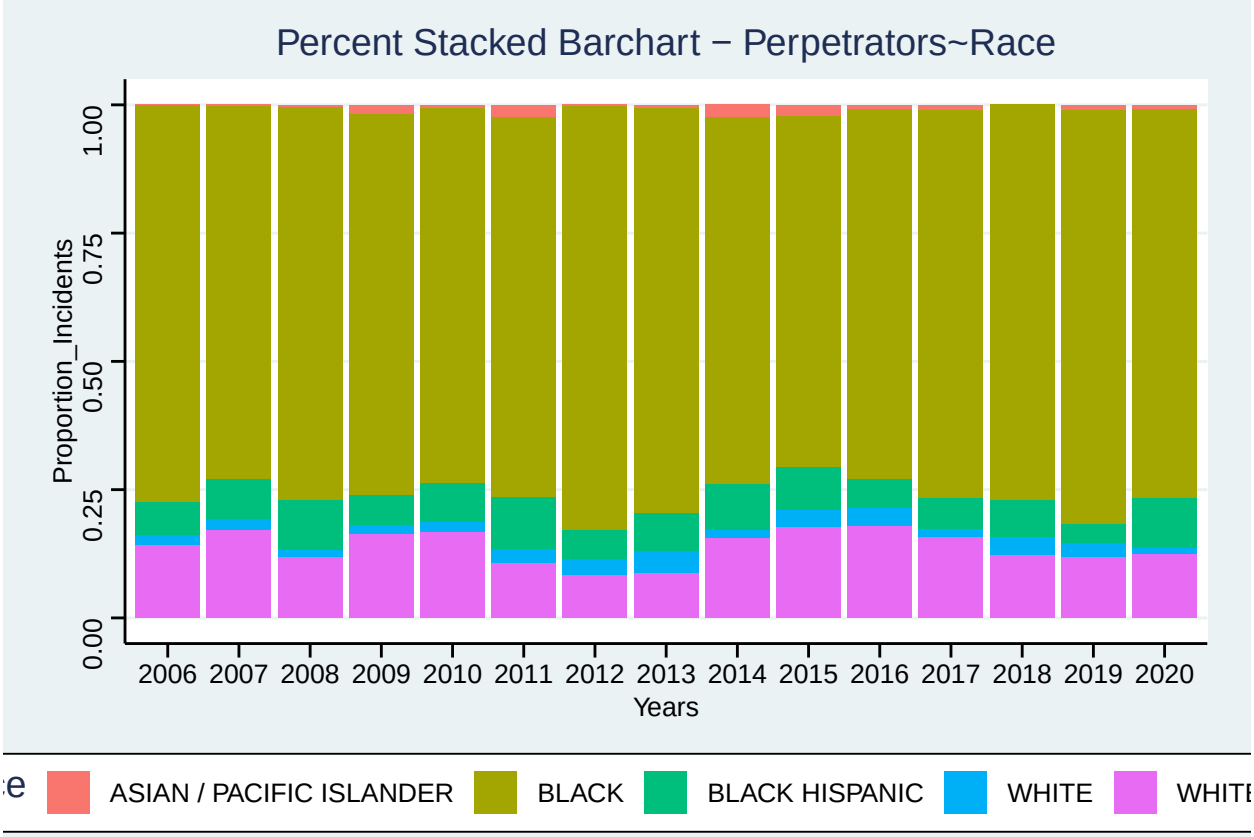
```
year2 = c(rep("2006",5), rep("2007",5), rep("2008",5), rep("2009",5), rep("2010",5)
          ,rep("2011",5) , rep("2012",5), rep("2013",5), rep("2014",5), rep("2015",5)
          ,rep("2016",5) , rep("2017",5), rep("2018",5), rep("2019",5), rep("2020",5))
Race = rep(c('ASIAN / PACIFIC ISLANDER', 'BLACK', 'BLACK HISPANIC', 'WHITE',
             'WHITE HISPANIC'),15)

table4 <- nypd_data %>%
  select(YEAR, VIC_RACE, VIC_SEX) %>% group_by(YEAR, VIC_RACE) %>%
  filter(VIC_RACE != 'AMERICAN INDIAN/ALASKAN NATIVE')%>%
  count() %>%
  rename(Counts = n)

ggplot(table4, aes(fill=Race, y=Counts, x=year2)) +
  geom_bar(position="fill", stat="identity")+
  theme_stata()+
  ggtitle('Percent Stacked Barchart - Victims~Race')+
  xlab("Years")+
  ylab('Proportion_Incidents')
```



Note: 'AMERICAN INDIAN/ALASKAN NATIVE' had 1 count in 2012 and 2018 only. This fact is not represented in the graph above in order to make the graph smoother to read for the readers.

The above bar plot shows that most of victims in the shooting incidents across the New York were Black while showing that across the years this proportion has not changed much with at least 65% of the victims belonging to this race every single year.

```
table5 <- nypd_data %>%
  select(YEAR, PERP_RACE, PERP_SEX) %>% group_by(YEAR, PERP_RACE) %>%
  filter(PERP_RACE != 'AMERICAN INDIAN/ALASKAN NATIVE')%>%
  count() %>%
  rename(Counts = n)
table5_2 <- data.frame(YEAR = "2018", PERP_RACE = 'ASIAN / PACIFIC ISLANDER', Counts = 0)

table5 = rbind(table5, table5_2)
table5 = table5 %>% arrange((PERP_RACE)) %>% arrange(YEAR)

ggplot(table5, aes(fill=Race, y=Counts, x=year2)) +
  geom_bar(position="fill", stat="identity")+
  theme_stata()+
  ggtitle('Percent Stacked Barchart - Perpetrators~Race')+
  xlab("Years")+
  ylab('Proportion_Incidents')
```



Note: 'AMERICAN INDIAN/ALASKAN NATIVE' had 1 count in year 2009. This fact is not represented in the graph above in order to make the graph smoother to read for the readers.

The above graph shows clearly that approximately more than 75% of the perpetrators of the shooting cases across the years belonged to Black race. Again, when we look at the data across the years this proportion has remained approximately same.

Further actions should be taken by the authorities of New York to correct this kind of imbalance in both of the proportions.

Part 3: Model

In this part a linear model will be made to predict the Number of Shooting Incidents across the years and locations on the basis of the number of Murders recorded.

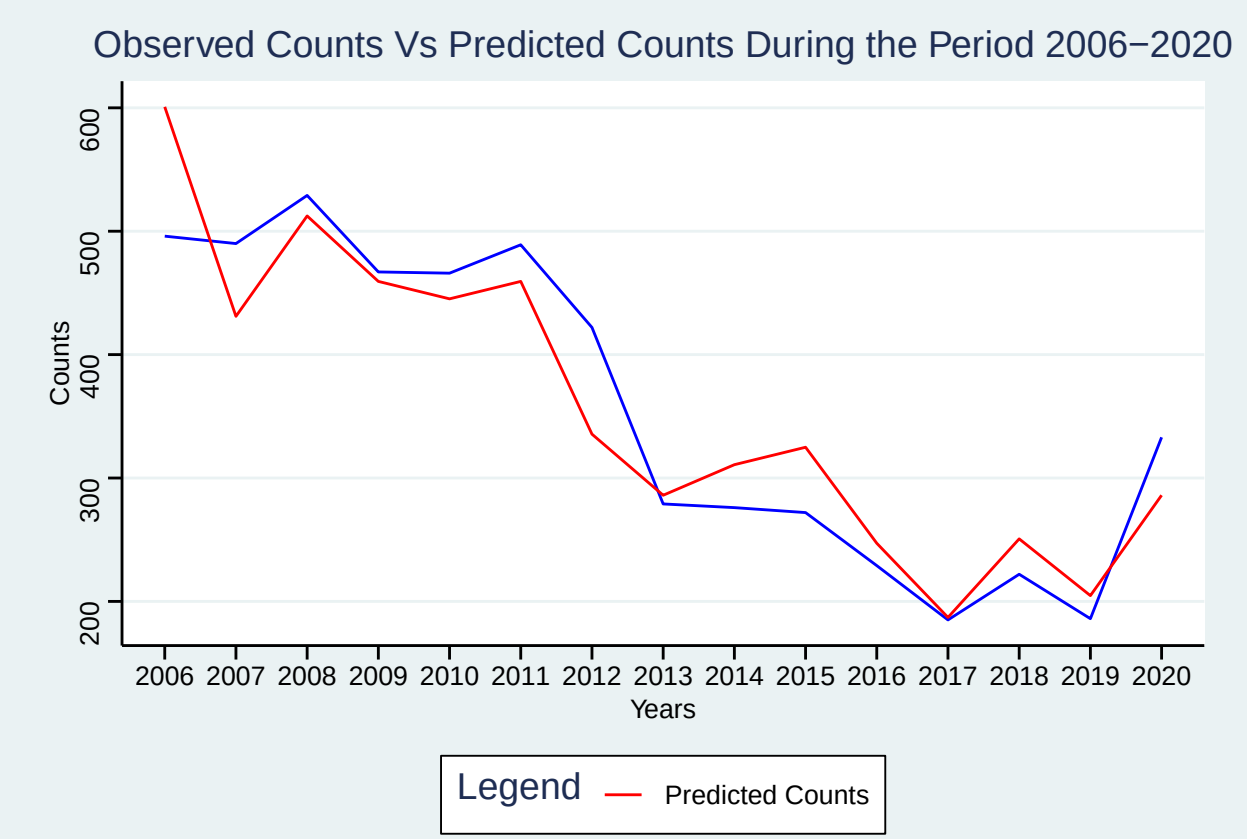
In Part 2 we already saw that the trend lines reflecting the existence of the presence of a linear relationship between the Number of Shooting Incidents and the Murders. We will now further explore this relationship with the help of a model.

```
#Note: Model to predict Murders can be obtained by reversing the position pf the elements in the below

nypd_total_counts <- cbind(nypd_count, nypd_murders[,"MURDERS"])
nypd_total_counts <- nypd_total_counts %>%
  rename(CountsofIncidents = 'n', Murders = "MURDERS")

model = lm(CountsofIncidents ~ Murders, data = nypd_total_counts)
nypd_total_counts <- cbind(nypd_total_counts, predCounts = predict(model))
```

```r
summary(model)
```

```
##
## Call:
## lm(formula = CountsofIncidents ~ Murders, data = nypd_total_counts)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.670  -8.343  -1.600   7.413  43.695
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.4538     3.0419   1.135     0.26
## Murders       3.5365     0.1269  27.878   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.84 on 73 degrees of freedom
## Multiple R-squared:  0.9141, Adjusted R-squared:  0.913
## F-statistic: 777.2 on 1 and 73 DF,  p-value: < 2.2e-16
```

```r
nypd_total_counts[,c('YEAR','CountsofIncidents', 'predCounts')] %>%
  group_by(YEAR) %>%
  summarize(CountsofIncidents = sum(CountsofIncidents),
            predCounts = sum(predCounts)) %>%
  ggplot(aes(group =1))+
  geom_line(aes(x = YEAR, y = CountsofIncidents), color = "blue")+
  geom_line(aes(x = YEAR, y = predCounts), color = 'red', show.legend = TRUE)+
  scale_colour_manual(name = 'Legend', values = c("Predicted Counts" = 'red'))+
  theme_stata()+
  ggtitle("Observed Counts Vs Predicted Counts During the Period 2006-2020")+
  xlab("Years")+
  ylab("Counts")
```



Predicted counts, derived on the Murder Counts, follows the observed counts quite properly. From the above graph we can see that Model, by using Murder Counts, was able to predict the motions/changes in the Shooting Incidents across the period; this confirm the there indeed exist a linear relationship between the Number of Shooting Incidents and Number of Murders and we can infer the other if we have one of them with quite a confidence.

Part 4: Bias Sources

Bias identification is an important part of any Data Science Project. Below are the Bias Sources identified throughout the whole reporting process:

1. A preconceived bias about Brooklyn known popular for Shooting Incidents on the basis of media reports. To mitigate it all the regions were considered equally in the process.
2. Any information which could have been used to identify individual case has been removed during the Part 1. This information could have been used to find when does most of the incidents occur in the different locations while being able to predict race of Victim and Perpetrator. Also, the GPS data could have been further used to find the real names and locations of all the parties recorded being involved in the incidents. All this Data was removed to ensure maximum anonymity in each case.
3. A preconceived bias that a particular race falls under the category of Perpetrators most of the time. This bias originated once again from media reports I obtained in my country. To mitigate it, all the races were considered equally in the concerned processes.
4. A bias in general that a huge number of shooting incidents happen in the US. This bias helped me to do more research than required in the assignment's grading scheme to fully satisfy my curiosity.
5. A few filters were introduced during the process to make more visually appealing and easy to understand graphs. To mitigate it, special notes have been introduced in the report.

Conclusion

NYPD Data Set is a complex data set with a lot of data one can use to unearth a vast amount of facts about the reality of Shooting Incidents in New York in general. From the analysis in the report, we can see summarize our findings as below:

1. Shooting incidents and deaths across the months for the period follow a seasonal cycle with both of them having lowest count in February and continues to a rise in a linear fashion until they reach the peaks in the month of August and September respectively before eventually falling down, again in a linear fashion.
2. Shooting incidents and deaths across the days for the period don't follow any specific cyclic pattern but whenever they goes beyond a threshold on a specific day they fall back within the threshold in the next day. More data is required to examine this erratic trend.
3. 78.4% of the total shooting incidents and 79.4% of the deaths resulting from such incidents have been recorded from residential areas containing multi-dwelling units like apartments and public housing units and private homes. This means that only 20% of such events happen in public places like stores and bars. We need more data to examine this conclusion because as it stands it means that staying outdoor is much safer than being indoor in the New York.
4. Incident Rate and Murder Rates has been falling in general but spiked in 2020. It can be a temporary spike due to global pandemic amid other factors, need more data besides this data set to gain a complete picture on this point.
5. Not every regions in New York has equal number of incidents and incidents leading to general. The number of incidents leading to murders vary quite a bit with Brooklyn having mean incidents of 36.2 Murders reported in each year while Staten Island has a mean of only 4.2. And since we have seen a linear relationship between Murder Rates and Incident Rates in Part 3 we can infer that same must be true for number of incidents.
6. The Number of Incident follows a right skew distribution which is a welcoming plus point indicating that most of the time each region reports less than 100 incidents in a year.
7. There is a linear relationship between the Number of Incidents and Numbers of Murders in the regions across the years.
8. There is a huge imbalance in the proportion of perpetrators and victims in New York with Blacks covering them at least 75% and 65% respectively. This shows that most of the individuals involved in the shooting incidents in New York are Black, and mostly Men, as we saw that more than 90% of all the perpetrators in the shooting incidents are men.

All the above conclusions are based on the process followed in the report and are subjected to the bias sources identified in Part 4. Hence, we must be cautious in accepting these conclusions as they could change if one could use data from different overarching sources.