NYPD Data Analysis

Given - Data of every shooting incident in NYC since 2006. Each record includes information such as, location and time of occurence, suspect and victim information.

Step 1 - Package Installation

```
## Install the tidyverse package install.packages("tidyverse")

## Loading the tidyverse library
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
## Loading the lubridate library for date
library(lubridate)
```

Step 2 - Import Data

```
url_nypd_data <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

## Load data from url
nypd_data <- read_csv(url_nypd_data)
```

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## 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.
```

Step 3- Tidy and Transform the data

- Removing columns that are not required for the analysis.
- Handle null/unknown values in the columns.

```r
## Removing columns that are not required for the analysis
nypd_data <- nypd_data %>%
  select(-c("LOC_CLASSFCTN_DESC", "LOC_OF_OCCUR_DESC", "PRECINCT","LOCATION_DESC","JURISDICTION_CODE",X

summary(nypd_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:28562       Length:28562      Length:28562
##  1st Qu.: 65439914   Class :character   Class1:hms        Class :character
##  Median : 92711254   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :127405824                      Mode  :numeric
##  3rd Qu.:203131993
##  Max.   :279758069
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Mode :logical           Length:28562       Length:28562
##  FALSE:23036             Class :character   Class :character
##  TRUE :5526              Mode  :character   Mode  :character
##
##
##
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```r
## Check for null values in the columns
sapply(nypd_data, function(x) sum(is.na(x)))
```

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                       0                       0                       0
##                    BORO STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP
##                       0                       0                    9344
##                PERP_SEX               PERP_RACE           VIC_AGE_GROUP
##                    9310                    9310                       0
##                 VIC_SEX                VIC_RACE
##                       0                       0
```

```r
## Replace null values with Unknown
nypd_data <- nypd_data %>% replace_na(list(PERP_AGE_GROUP="UNKNOWN", PERP_SEX="UNKNOWN", PERP_RACE="UNKN

nypd_data$PERP_SEX <- recode(nypd_data$PERP_SEX, U="UNKNOWN")
nypd_data$VIC_SEX <- recode(nypd_data$VIC_SEX, U="UNKNOWN")

nypd_data$BORO <- as.factor(nypd_data$BORO)
nypd_data$PERP_AGE_GROUP <- as.factor(nypd_data$PERP_AGE_GROUP)
nypd_data$PERP_SEX <- as.factor(nypd_data$PERP_SEX)
nypd_data$PERP_RACE <- as.factor(nypd_data$PERP_RACE)
nypd_data$VIC_AGE_GROUP <- as.factor(nypd_data$VIC_AGE_GROUP)
nypd_data$VIC_SEX <- as.factor(nypd_data$VIC_SEX)
nypd_data$VIC_RACE <- as.factor(nypd_data$VIC_RACE)

summary(nypd_data)
```

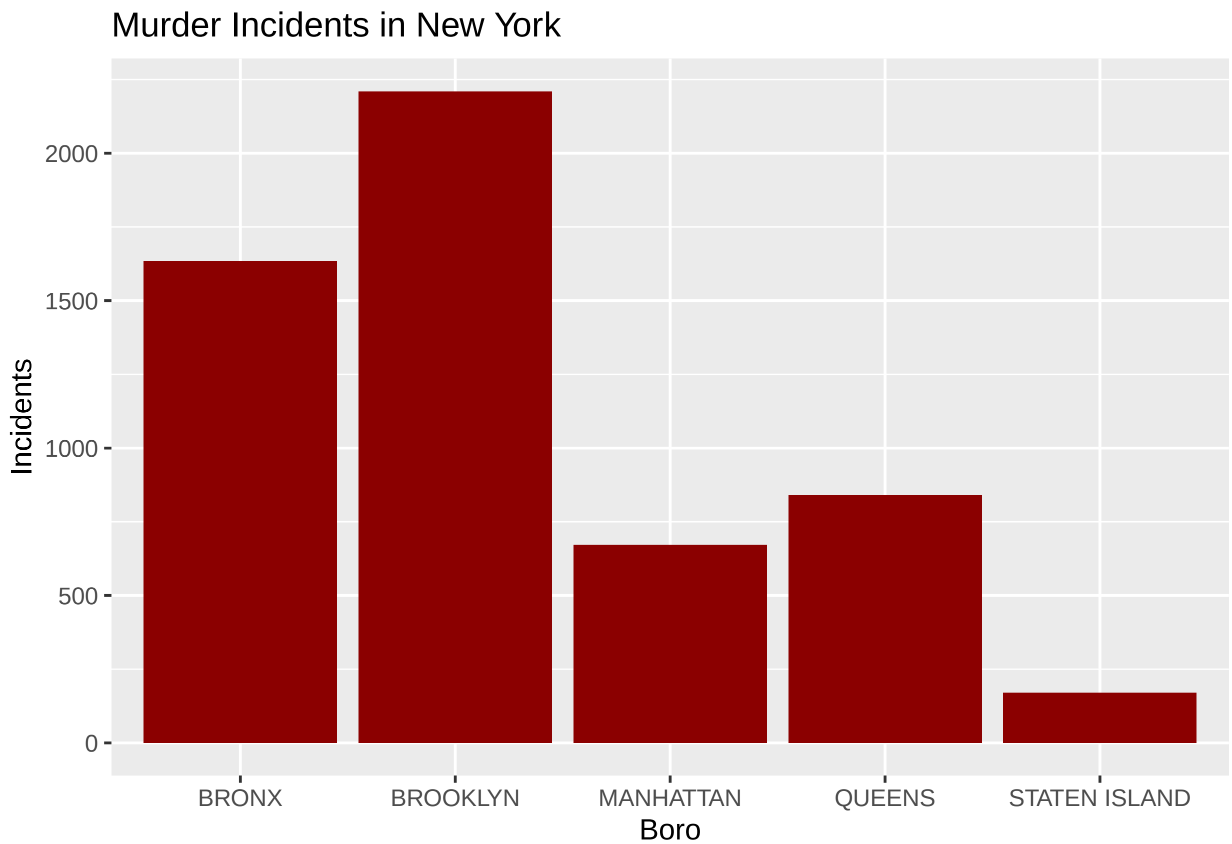
```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME                   BORO
##  Min.   :  9953245   Length:28562       Length:28562      BRONX        : 8376
##  1st Qu.: 65439914   Class :character   Class1:hms        BROOKLYN     :11346
##  Median : 92711254   Mode  :character   Class2:difftime   MANHATTAN    : 3762
##  Mean   :127405824                      Mode  :numeric    QUEENS       : 4271
##  3rd Qu.:203131993                                        STATEN ISLAND:  807
##  Max.   :279758069
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX              PERP_RACE
##  Mode :logical           UNKNOWN:12492   (null) : 1141   BLACK         :11903
##  FALSE:23036             18-24  : 6438   F      :  444   UNKNOWN       :11147
##  TRUE :5526              25-44  : 6041   M      :16168   WHITE HISPANIC: 2510
##                          <18    : 1682   UNKNOWN:10809   BLACK HISPANIC: 1392
##                          (null) : 1141                   (null)        : 1141
##                          45-64  :  699                   WHITE         :  298
##                          (Other):   69                   (Other)       :  171
##  VIC_AGE_GROUP      VIC_SEX                                VIC_RACE
##  <18    : 2954   F      : 2760   AMERICAN INDIAN/ALASKAN NATIVE:   11
##  1022   :    1   M      :25790   ASIAN / PACIFIC ISLANDER      :  440
##  18-24  :10384   UNKNOWN:   12   BLACK                         :20235
##  25-44  :12973                   BLACK HISPANIC                : 2795
##  45-64  : 1981                   UNKNOWN                       :   70
##  65+    :  205                   WHITE                         :  728
##  UNKNOWN:   64                   WHITE HISPANIC                : 4283
```

Step 4 - Visualize

1. Which Boro has the highest number of murders?

```r
murder_data <- nypd_data %>%
  filter(STATISTICAL_MURDER_FLAG == TRUE) %>%
  group_by(BORO) %>%
  mutate(count_per_boro = sum(STATISTICAL_MURDER_FLAG))

murder_data %>%
```

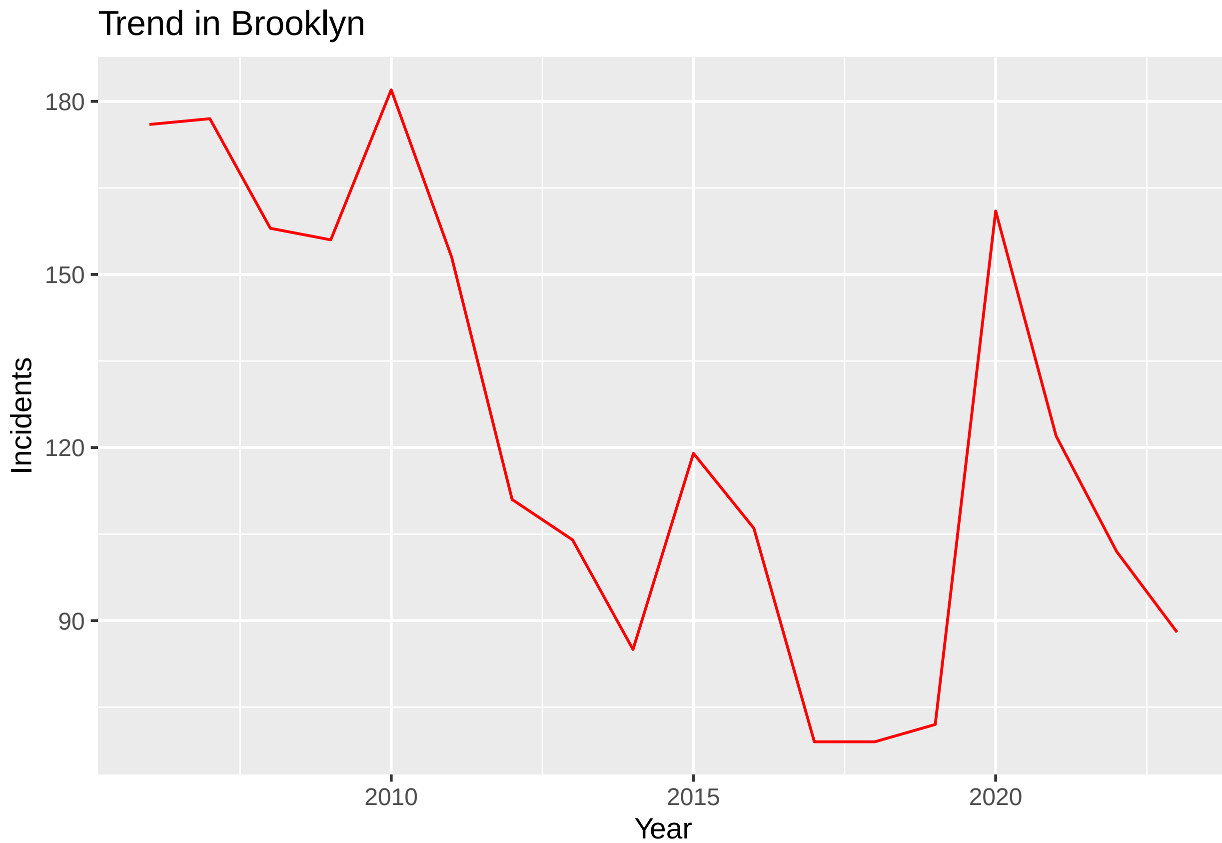
```r
  ggplot(aes(x = BORO)) +
  geom_bar(fill="darkred") +
  labs(title = "Murder Incidents in New York", x = "Boro", y = "Incidents")
```



Analysis - The bar graph gives us a view of the Boro that has the highest number of murders in NewYork city. Brooklyn has the highest followed by Bronx.

2. How has the crime rate changed across years in Brooklyn?

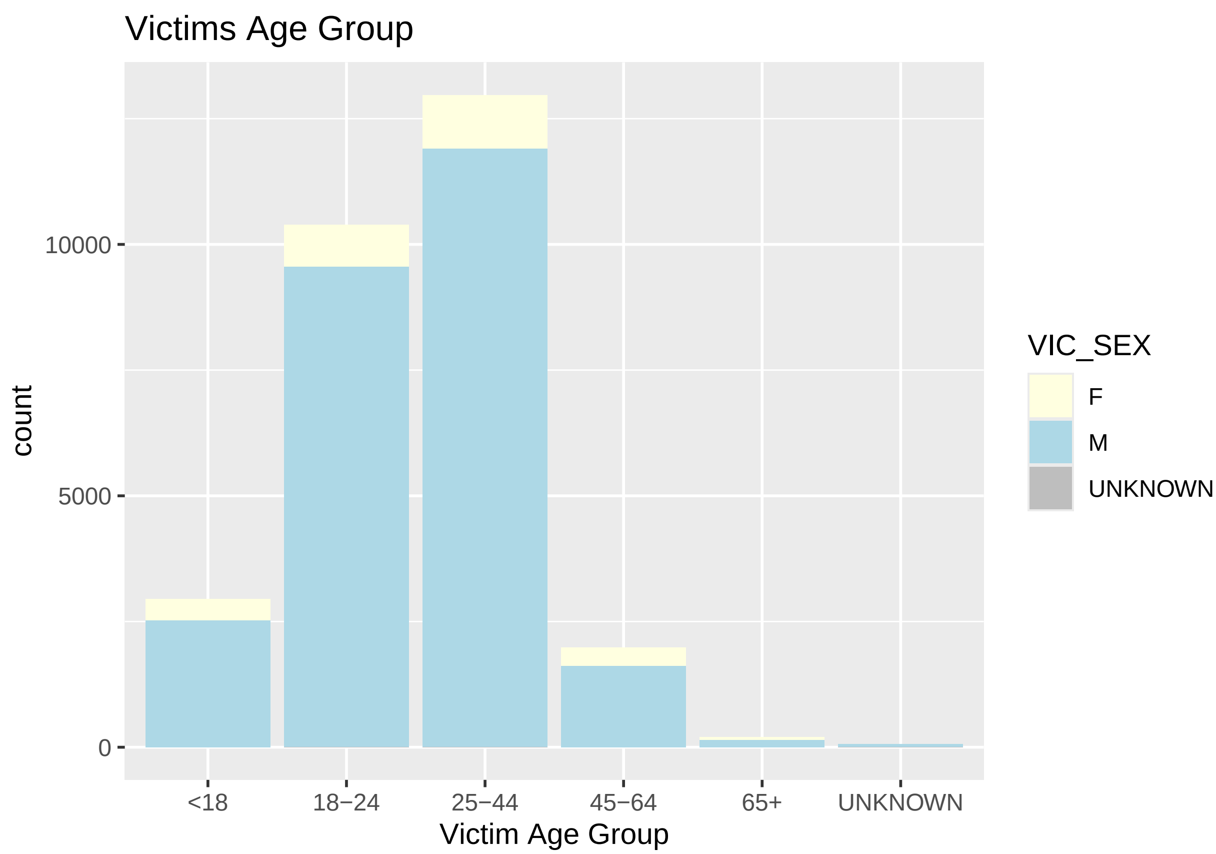
```r
Brooklyn_incidents <- nypd_data %>%
  filter(BORO=="BROOKLYN") %>%
  mutate(year = year(mdy(OCCUR_DATE))) %>%
  group_by(year) %>%
  mutate(cases_per_year = sum(STATISTICAL_MURDER_FLAG)) %>%
  select(year, cases_per_year)

Brooklyn_incidents %>%
  ggplot(aes(x= year)) +
  geom_line(aes(y=cases_per_year), color = "red") +
  labs(title = "Trend in Brooklyn ", x = "Year", y = "Incidents")
```



Analysis - The trends of crime show a dip in the years 2010 to 2017. But there is a sudden spike again in the year 2020. This might be due to other reasons such as - covid outbreak, political changes. To get an exact idea of the reason for this spike we have to investigate data related to other factors as well.

3. Which age group was the most affected by these crimes?

```
nypd_data %>%
  filter(VIC_AGE_GROUP!=1022) %>%
  ggplot(aes(x=VIC_AGE_GROUP, fill= VIC_SEX)) +
  labs(x = " Victim Age Group ",title="Victims Age Group")+
  geom_bar(position='stack')+
  scale_fill_manual(values=c('lightyellow','lightblue','grey'))
```



Analysis - It can be noted that 25-44 followed by 18-24 age groups have the largest number of victims. Also it can be seen that most of the victims in any age groups are men.

Step 5 - Model creation

```
model <- lm(STATISTICAL_MURDER_FLAG ~   PERP_RACE + VIC_RACE + VIC_SEX + VIC_AGE_GROUP, data= nypd_data
summary(model)
```

```
##
## Call:
## lm(formula = STATISTICAL_MURDER_FLAG ~ PERP_RACE + VIC_RACE +
##     VIC_SEX + VIC_AGE_GROUP, data = nypd_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4715 -0.2140 -0.1858 -0.1163  1.0089
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             -0.093708   0.119006  -0.787   0.4310
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -0.179391   0.277381  -0.647   0.5178
## PERP_RACEASIAN / PACIFIC ISLANDER        0.143122   0.032975   4.340 1.43e-05
## PERP_RACEBLACK                           0.068444   0.012167   5.625 1.87e-08
```

```
## PERP_RACEBLACK HISPANIC                  0.065919   0.015760   4.183 2.89e-05
## PERP_RACEUNKNOWN                         0.023127   0.012206   1.895   0.0581
## PERP_RACEWHITE                           0.202451   0.026896   7.527 5.33e-14
## PERP_RACEWHITE HISPANIC                  0.107127   0.014164   7.563 4.05e-14
## VIC_RACEASIAN / PACIFIC ISLANDER         0.198142   0.119724   1.655   0.0979
## VIC_RACEBLACK                            0.174106   0.118196   1.473   0.1408
## VIC_RACEBLACK HISPANIC                   0.141540   0.118406   1.195   0.2319
## VIC_RACEUNKNOWN                          0.062753   0.128394   0.489   0.6250
## VIC_RACEWHITE                            0.188567   0.119214   1.582   0.1137
## VIC_RACEWHITE HISPANIC                   0.176707   0.118330   1.493   0.1354
## VIC_SEXM                                -0.001019   0.007929  -0.129   0.8977
## VIC_SEXUNKNOWN                          -0.058683   0.119316  -0.492   0.6228
## VIC_AGE_GROUP1022                       -0.147822   0.391962  -0.377   0.7061
## VIC_AGE_GROUP18-24                       0.038029   0.008195   4.640 3.49e-06
## VIC_AGE_GROUP25-44                       0.088576   0.008017  11.048  < 2e-16
## VIC_AGE_GROUP45-64                       0.111449   0.011449   9.735  < 2e-16
## VIC_AGE_GROUP65+                         0.174219   0.028445   6.125 9.20e-10
## VIC_AGE_GROUPUNKNOWN                     0.117706   0.052315   2.250   0.0245
##
## (Intercept)
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE
## PERP_RACEASIAN / PACIFIC ISLANDER       ***
## PERP_RACEBLACK                          ***
## PERP_RACEBLACK HISPANIC                 ***
## PERP_RACEUNKNOWN                        .
## PERP_RACEWHITE                          ***
## PERP_RACEWHITE HISPANIC                 ***
## VIC_RACEASIAN / PACIFIC ISLANDER        .
## VIC_RACEBLACK
## VIC_RACEBLACK HISPANIC
## VIC_RACEUNKNOWN
## VIC_RACEWHITE
## VIC_RACEWHITE HISPANIC
## VIC_SEXM
## VIC_SEXUNKNOWN
## VIC_AGE_GROUP1022
## VIC_AGE_GROUP18-24                      ***
## VIC_AGE_GROUP25-44                      ***
## VIC_AGE_GROUP45-64                      ***
## VIC_AGE_GROUP65+                        ***
## VIC_AGE_GROUPUNKNOWN                    *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3919 on 28540 degrees of freedom
## Multiple R-squared:  0.01658,    Adjusted R-squared:  0.01586
## F-statistic: 22.92 on 21 and 28540 DF,  p-value: < 2.2e-16
```

Step 6 - Bias

From the above visualization and analysis done it can be noted that most of the victims are men. So there is a possibility that women have not reported their crimes. In order to get accurate results other factors should also be considered such as impact of covid, any political changes or change in the job market. This might be the possible bias in data. On a personal level, while analysing the data, the higher crime rate in a particular region has intrigued me to do a further analysis on that Boro. But this could have been looked at a different way and have analysed the crime rates between boros in New York to get more insights.

Conclusion

To conclude, Brooklyn had the most number of murders, followed by Bronx. The number of male victims is significantly larger than the female victims. And the age groups that are most affected fall between 18-44, which are most likely salaried persons. Therefore, it is suggested for people living in these areas, in this age group to be cautious.