NYPD Shooting Incident Report

Step 1: Import Data

The following block could make sure anyone who runs the code can reproduce the same analysis. This report uses NYPD Shooting Incident Data (Historic) from https://catalog.data.gov/dataset.

```
url = "https://data.cityofnewyork.us/api/views/833y-
fsy8/rows.csv?accessType=DOWNLOAD"
rawdata = read.csv(url)
#install.packages(tidyverse)
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ──────────────────────────────────────────
tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

Step 2: Tidy and Transform Data

To start with, I got rid of the columns that I do not think I will need for further analysis. I believe that the incident keys and exact locations like coordinates, latitude or longitude will not be needed in this report. Thus, I removed “INCIDENT_KEY”, “X_COORD_CD”, “Y_COORD_CD”, “Latitude”, “Longitude” and “Lon_Lat” from the raw dataset. Then, I transformed character cells to date for column “OCCUR_DATE” and I transformed character cells to time for column “OCCUR_TIME” as well. For column “STATISTICAL_MURDER_FLAG”, I apply integers 0 and 1 to character cells of “false” and “true’. For column”VIC_SEX”, I apply integers 0 and 1 to character cells of “female” and “male” respectively. In the end, I decided to let the remaining columns as factors in order to do further analysis.

```
summary(rawdata)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:27312       Length:27312       Length:27312
##  1st Qu.: 63860880   Class :character   Class :character   Class
:character
##  Median : 90372218   Mode  :character   Mode  :character   Mode
:character
##  Mean   :120860536
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Length:27312            Length:27312
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD         Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000029   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :10
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :10
```

```
data = rawdata[,2:16]

library(lubridate)
data$OCCUR_DATE = mdy(data$OCCUR_DATE)
library(chron)
```

```
##
## Attaching package: 'chron'
```

```
## The following objects are masked from 'package:lubridate':
##
##     days, hours, minutes, seconds, years
```

```
data$OCCUR_TIME = hms(data$OCCUR_TIME)

data$STATISTICAL_MURDER_FLAG[data$STATISTICAL_MURDER_FLAG == "true"] <- 1
data$STATISTICAL_MURDER_FLAG[data$STATISTICAL_MURDER_FLAG == "false"] <- 0
data$VIC_SEX[data$VIC_SEX == "M"] <- 1
data$VIC_SEX[data$VIC_SEX == "W"] <- 0

summary(data)
```

```
##    OCCUR_DATE           OCCUR_TIME                            BORO
##  Min.   :2006-01-01   Min.   :0S                          Length:27312
##  1st Qu.:2009-07-18   1st Qu.:3H 27M 0S                   Class :character
##  Median :2013-04-29   Median :15H 11M 0S                  Mode  :character
##  Mean   :2014-01-06   Mean   :12H 41M 31.7091388400731S
##  3rd Qu.:2018-10-15   3rd Qu.:20H 45M 0S
##  Max.   :2022-12-31   Max.   :23H 59M 0S
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Length:27312            Length:27312
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

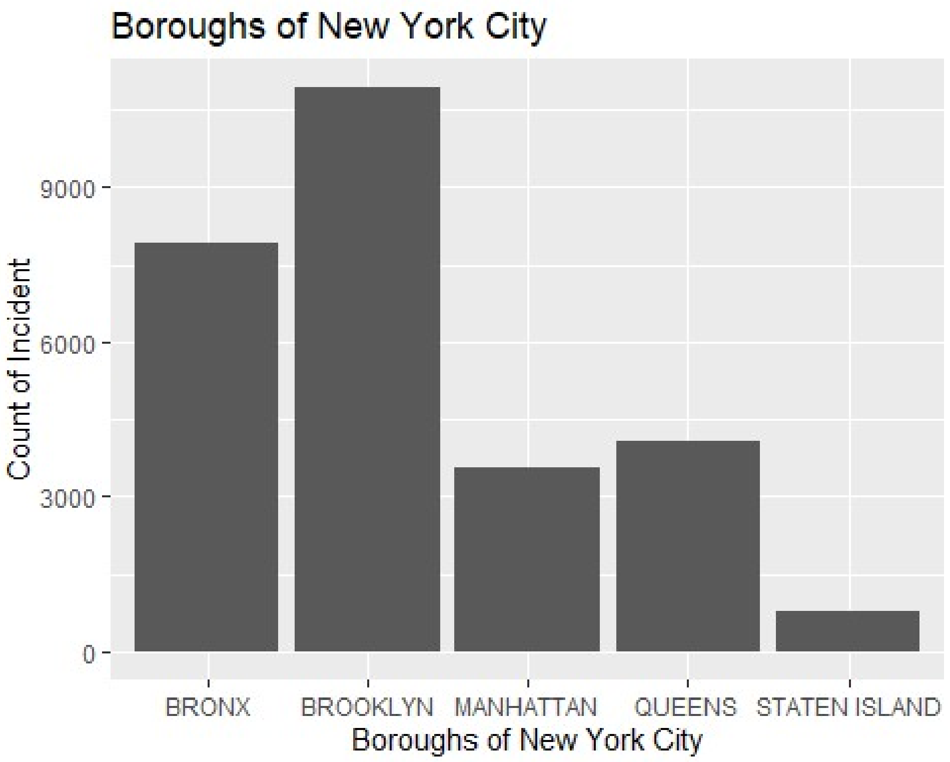
```
##
##
##
##
##    VIC_RACE
##  Length:27312
##  Class :character
##  Mode  :character
##
##
##
##
```

Step 3: Visualizations and Analysis

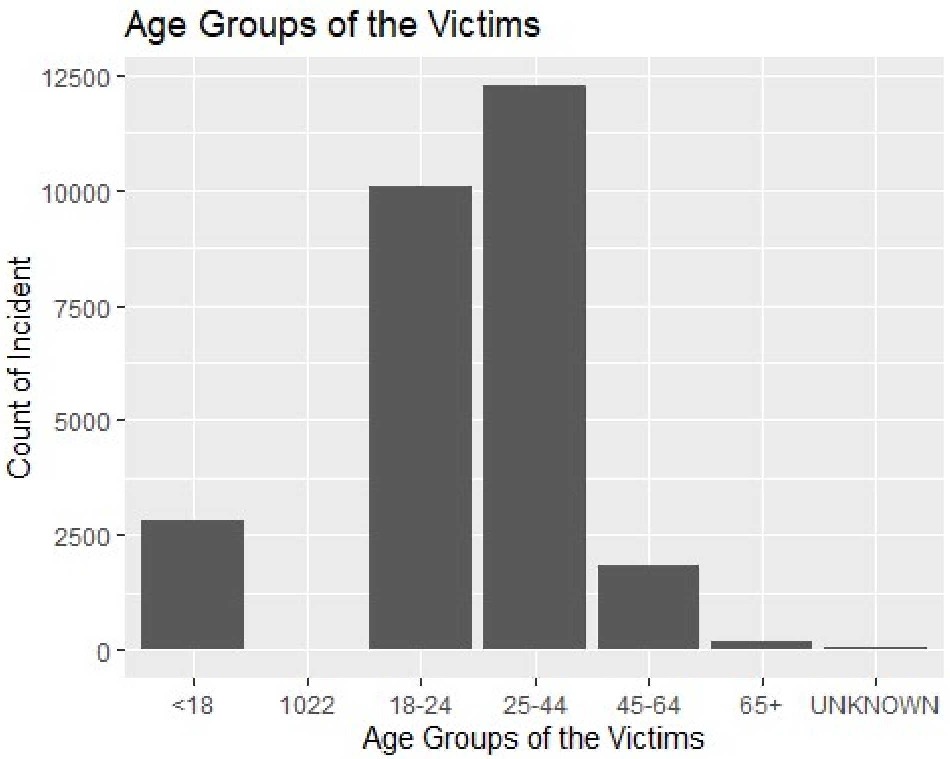
In this step, I generated a histogram plot of the incidents happened in New York City to investigate if the shooting incidents are related to Boroughs. According to the first chart below, we can conclude that Brooklyn has the most counts of incidents and Bronx has the second. Then, to detect the difference of the shooting incidents among the age groups, I generated a second histogram. This second histogram illustrated that there are two age groups in New York City that are more likely to get shot. The first one is individuals aged 25-44 and the second one is individuals aged 18-24.

```
library(ggplot2)

#visualization 1
ggplot(data,aes(x=BORO))+geom_bar()+labs(title="Boroughs of New York City",
x="Boroughs of New York City", y="Count of Incident")
```



```
# visualization 2
ggplot(data,aes(x=VIC_AGE_GROUP))+geom_bar()+labs(title="Age Groups of the
Victims", x="Age Groups of the Victims", y="Count of Incident")
```



According to the preliminary analysis of the data, I assume there exists a relationship between the statistical murder flag and the other factors like occur time, victim sex, or victim age. As the analysis below indicates, I concluded that victims aged 65+ are more likely to be involved in the statistical murder shooting incidents.

```
# model
lm(data$STATISTICAL_MURDER_FLAG~data$VIC_AGE_GROUP+data$VIC_SEX+data$VIC_AGE_
GROUP+data$OCCUR_TIME)
```

```
##
## Call:
## lm(formula = data$STATISTICAL_MURDER_FLAG ~ data$VIC_AGE_GROUP +
##     data$VIC_SEX + data$VIC_AGE_GROUP + data$OCCUR_TIME)
##
## Coefficients:
##               (Intercept)     data$VIC_AGE_GROUP1022
##                  0.128996                  -0.128996
##   data$VIC_AGE_GROUP18-24    data$VIC_AGE_GROUP25-44
##                  0.036890                   0.088489
##   data$VIC_AGE_GROUP45-64      data$VIC_AGE_GROUP65+
##                  0.118871                   0.177369
## data$VIC_AGE_GROUPUNKNOWN              data$VIC_SEXF
##                  0.126394                   0.009287
##             data$VIC_SEXU            data$OCCUR_TIME
##                 -0.125042                         NA
```

Step 4: Bias Identification

This report only investigated limited relationships in the data frame that interest or are relatively obvious to me. But there may be other important topics that I omitted. This could cause the original bias. Besides, there is some data missing in the given data set, this could be caused by various reasons and could also be another source of bias. Furthermore, there could be potential extreme points in the data set affecting the results as well. And the major bias concern towards this report might be the analysis of the age groups of the victims. I think how the data divided the age group could lead to bias as well.