New York Shooting

2022-06-30

Required Libraries

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.1 ──
```

```
## ✓ ggplot2 3.3.6     ✓ purrr   0.3.4
## ✓ tibble  3.1.7     ✓ stringr 1.4.0
## ✓ tidyr   1.2.0     ✓ forcats 0.5.1
## ✓ readr   2.1.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

About the Data

The dataset was imported from the US catalog of shooting incidents in the city of New York from the link 'https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic/resource/c564b578-fd8a-4005-8365-34150d306cc4 (https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic/resource/c564b578-fd8a-4005-8365-34150d306cc4)'.It mainly shows the data collected for the past 2 decades regarding the shooting incidents in the city of new york like the victims age, gender etc.

```
ny_data = read_csv('https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD')
```

```
## Rows: 25596 Columns: 19
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## 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.
```

Tidying and Transforming

Getting an overview of the data that we are dealing with:

```
summary(ny_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:25596       Length:25596      Length:25596
##  1st Qu.: 61593633   Class :character   Class1:hms        Class :character
##  Median : 86437258   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :112382648                      Mode  :numeric
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:20668
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4928
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

Now, owing to the objective of our analysis, we will be removing certain columns that are not required which are:

1. Incident key
2. Precinct
3. Location
4. Perpetrator Age Group
5. Perpetrator Sex
6. Perpetrator Race

7. Victim Race
8. X coordinates
9. Y coordinates
10. Latitude
11. Longitude
12. Geo point
13. Jurisdiction Code

```
ny_data_mod = subset(ny_data, select = -c(INCIDENT_KEY ,OCCUR_TIME, PRECINCT, LOCATION_DESC, PER
P_AGE_GROUP,PERP_SEX , PERP_RACE,X_COORD_CD, Y_COORD_CD,Latitude,Longitude, Lon_Lat ) )
```

We will also remove the rows which contain any NA values.

```
ny_data_mod = ny_data_mod %>% drop_na()
```

Now, let us take a look at our cleaned data:

```
head(ny_data_mod)
```

```
## # A tibble: 6 × 7
##   OCCUR_DATE BORO      JURISDICTION_CODE STATISTICAL_MURD… VIC_AGE_GROUP VIC_SEX
##   <chr>      <chr>                 <dbl> <lgl>             <chr>         <chr>
## 1 08/27/2006 BRONX                     0 TRUE              25-44         F
## 2 03/11/2011 QUEENS                    0 FALSE             65+           M
## 3 04/14/2021 BRONX                     0 TRUE              18-24         M
## 4 12/10/2021 BRONX                     0 FALSE             25-44         M
## 5 02/22/2021 MANHATTAN                 0 FALSE             25-44         M
## 6 03/07/2021 BROOKLYN                  0 TRUE              25-44         M
## # … with 1 more variable: VIC_RACE <chr>
```

We can see that the occurence date column is in character which is not correct, so we need to change it to type date:

```
ny_data_mod$OCCUR_DATE = as.Date(ny_data_mod$OCCUR_DATE , format =  "%m/%d/%Y")
```

Visualization & Analysis

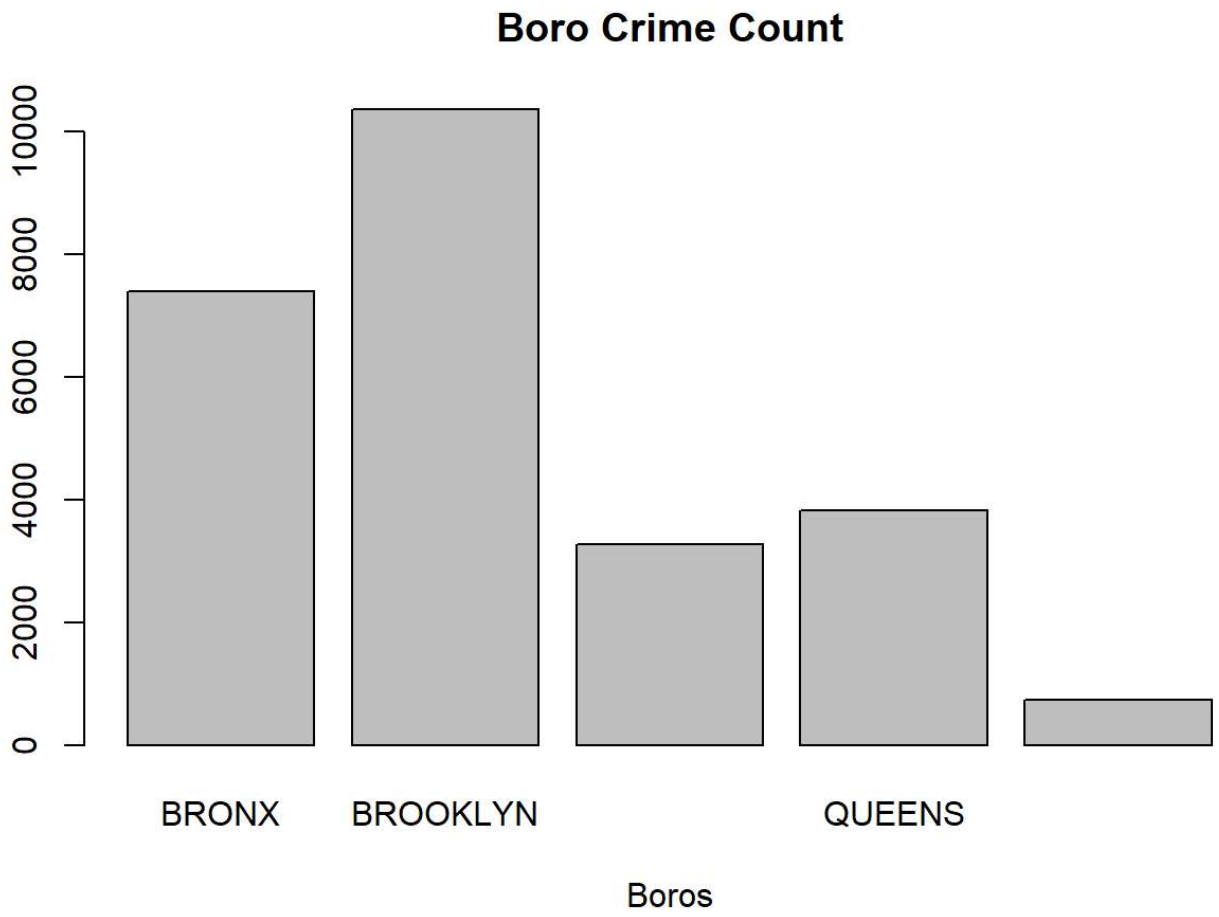
Our analysis will be based on answering the following question:

1. Which boro's has the highest crime rate?
2. Which Age group is the most exposed to crime?
3. Which Gender will make the person more liable to a criminal act?
4. Which Race will be more exposed?
5. Which age group is primarily exposed to murder?
6. Which month is the most crime savvy?

```
table(ny_data_mod$BORO)
```

```
##
##         BRONX      BROOKLYN     MANHATTAN        QUEENS STATEN ISLAND
##          7402         10365          3264          3827           736
```

```
barplot(table(ny_data_mod$BORO), main = 'Boro Crime Count' , xlab = 'Boros')
```

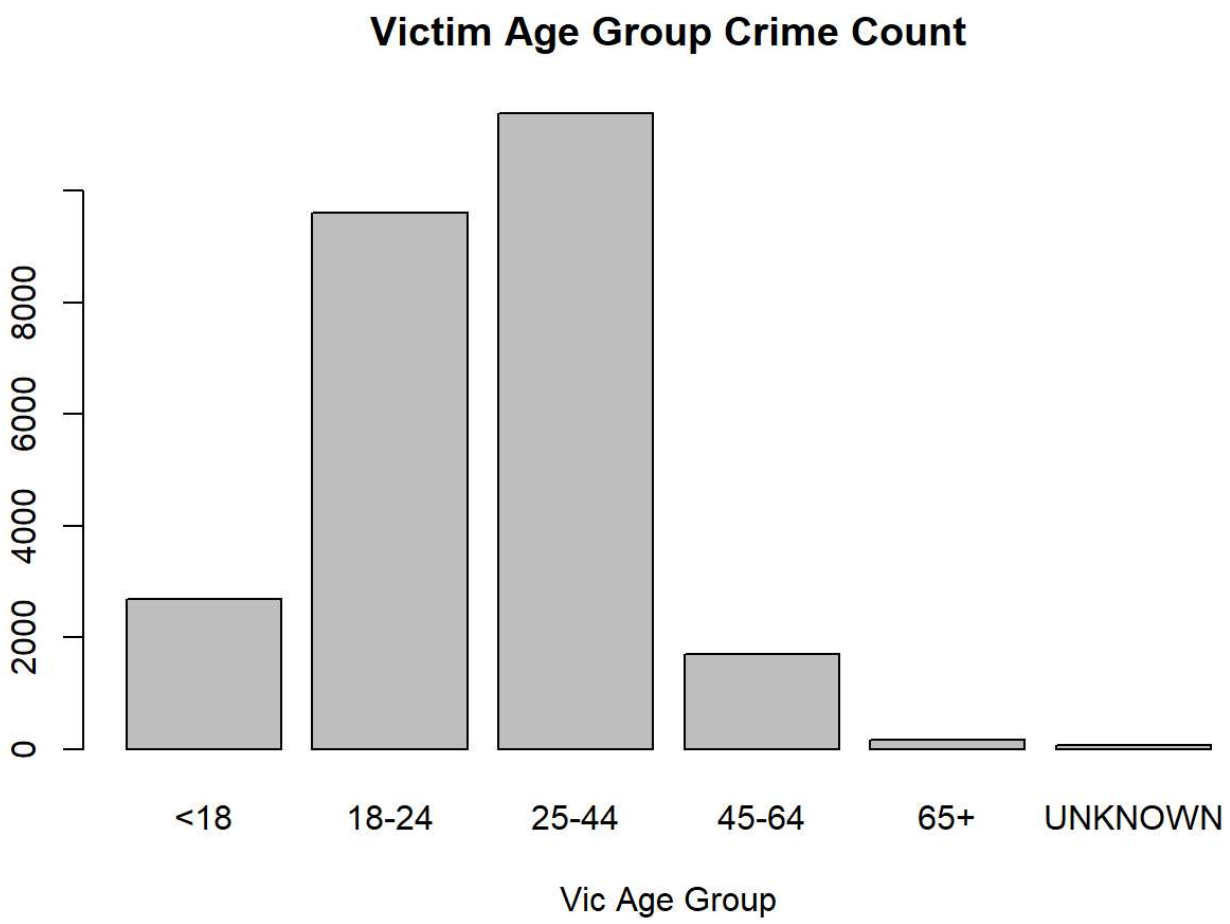


So, while looking at the above graph, we can say that 'Brooklyn' is the most dangerous area in terms of the total crimes committed over the years.

```
table(ny_data_mod$VIC_AGE_GROUP)
```

```
##
##     <18   18-24   25-44   45-64     65+ UNKNOWN
##    2681    9603   11385    1698     167      60
```

```
barplot(table(ny_data_mod$VIC_AGE_GROUP), main = 'Victim Age Group Crime Count', xlab = 'Vic Age
Group')
```

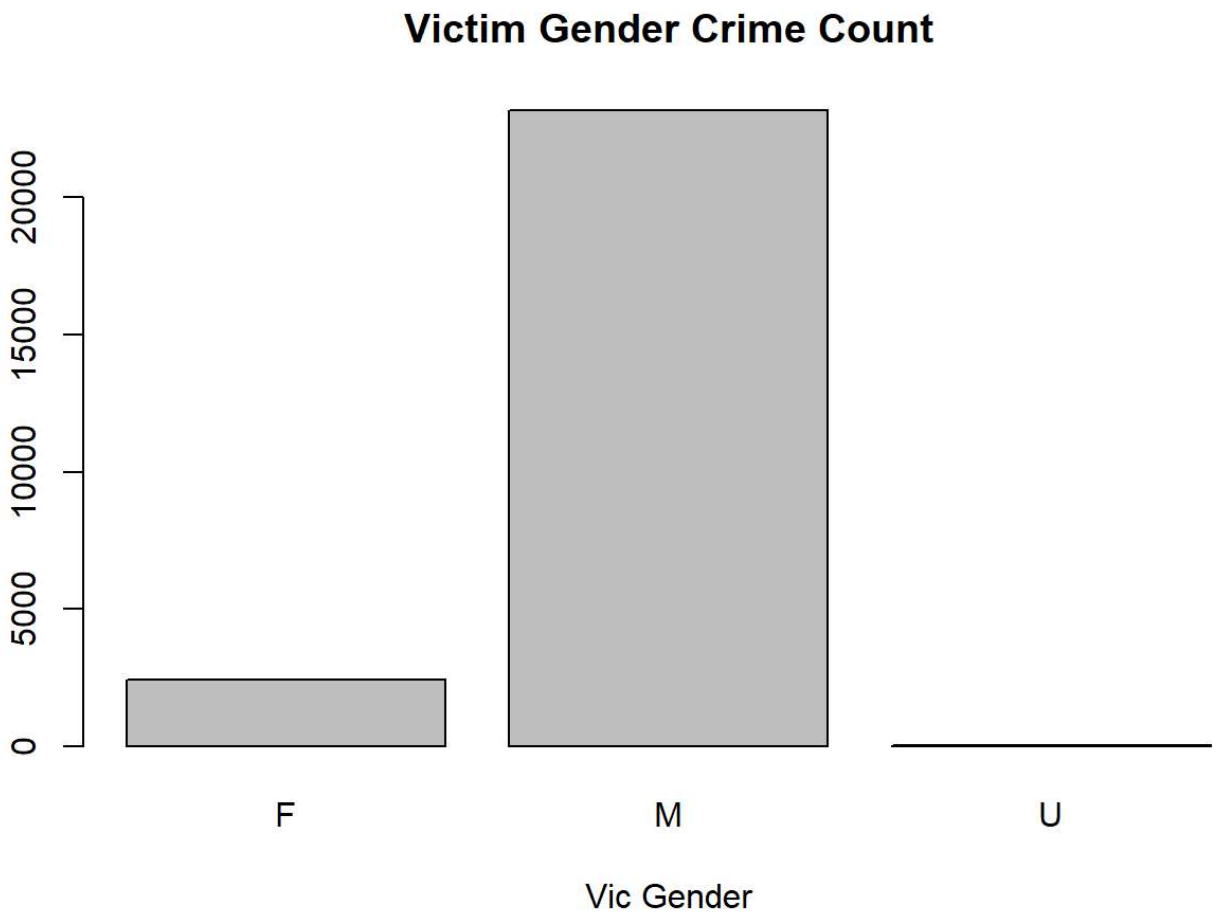


The most prone age group to report or be a victim to a crime is ‘25-44’ age group which is closely followed by the ‘18-24’ age group.

```
table(ny_data_mod$VIC_SEX)
```

```
##
##     F     M     U
##  2403 23180    11
```

```
barplot(table(ny_data_mod$VIC_SEX), main = 'Victim Gender Crime Count', xlab = 'Vic Gender')
```

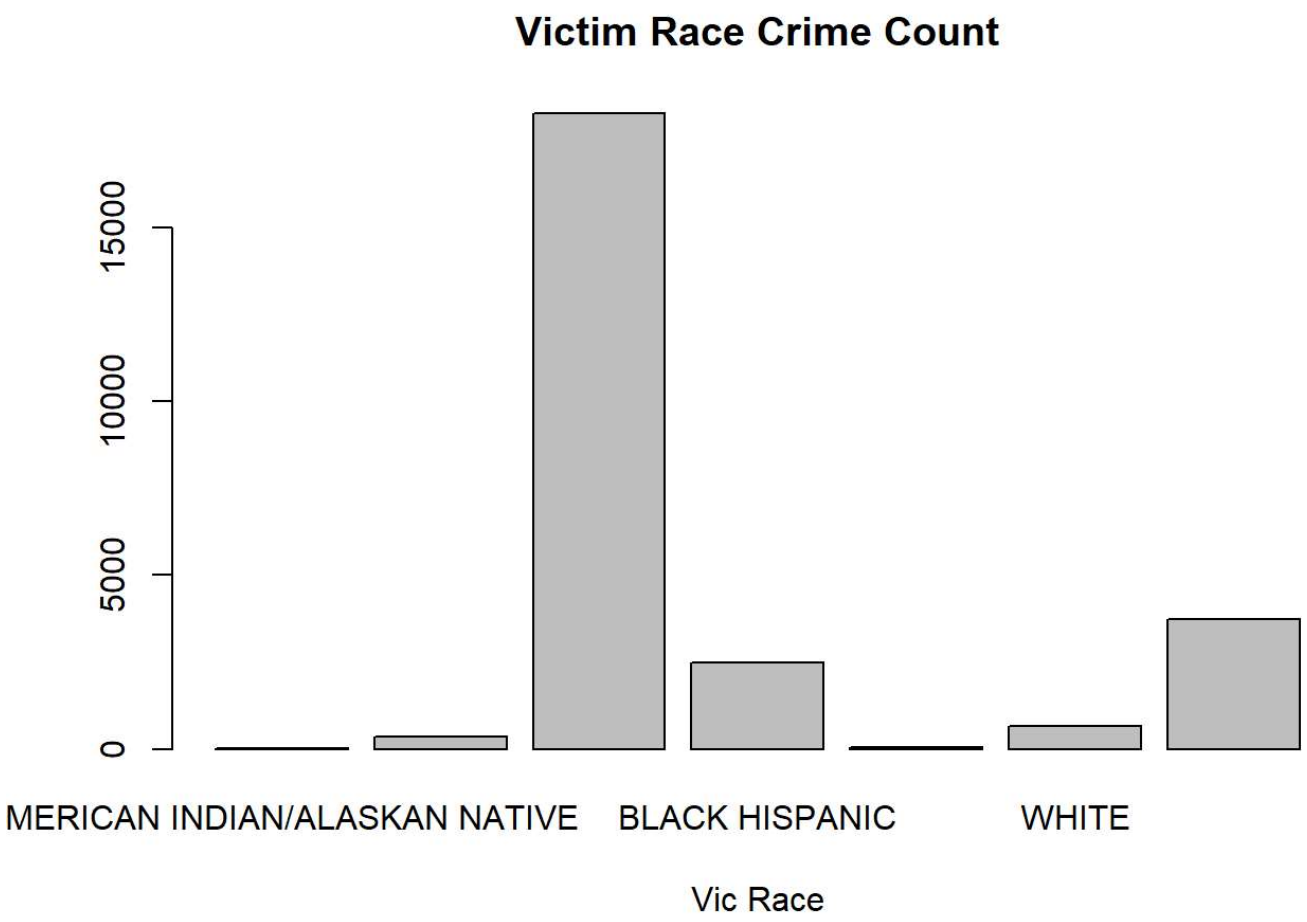


While comparing the victim's gender, males seem to be the most effected and the difference is a huge blow as well with more than 20000 reported male victims and almost 2500 female victims reported.

```
table(ny_data_mod$VIC_RACE)
```

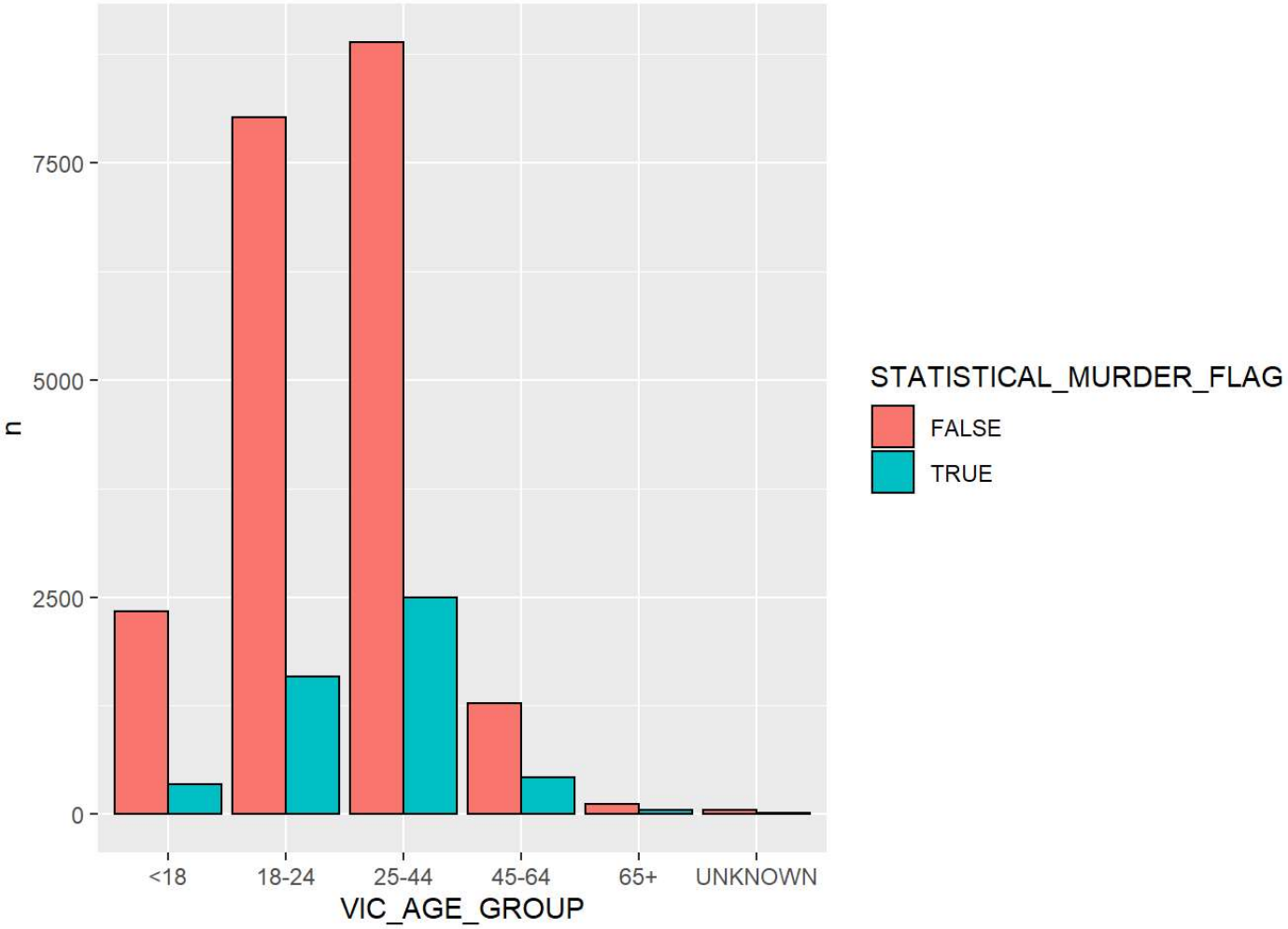
```
##
## AMERICAN INDIAN/ALASKAN NATIVE       ASIAN / PACIFIC ISLANDER
##                              9                            354
##                          BLACK                BLACK HISPANIC
##                          18280                           2485
##                        UNKNOWN                        WHITE
##                             65                            660
##                WHITE HISPANIC
##                           3741
```

```
barplot(table(ny_data_mod$VIC_RACE), main = 'Victim Race Crime Count', xlab = 'Vic Race')
```



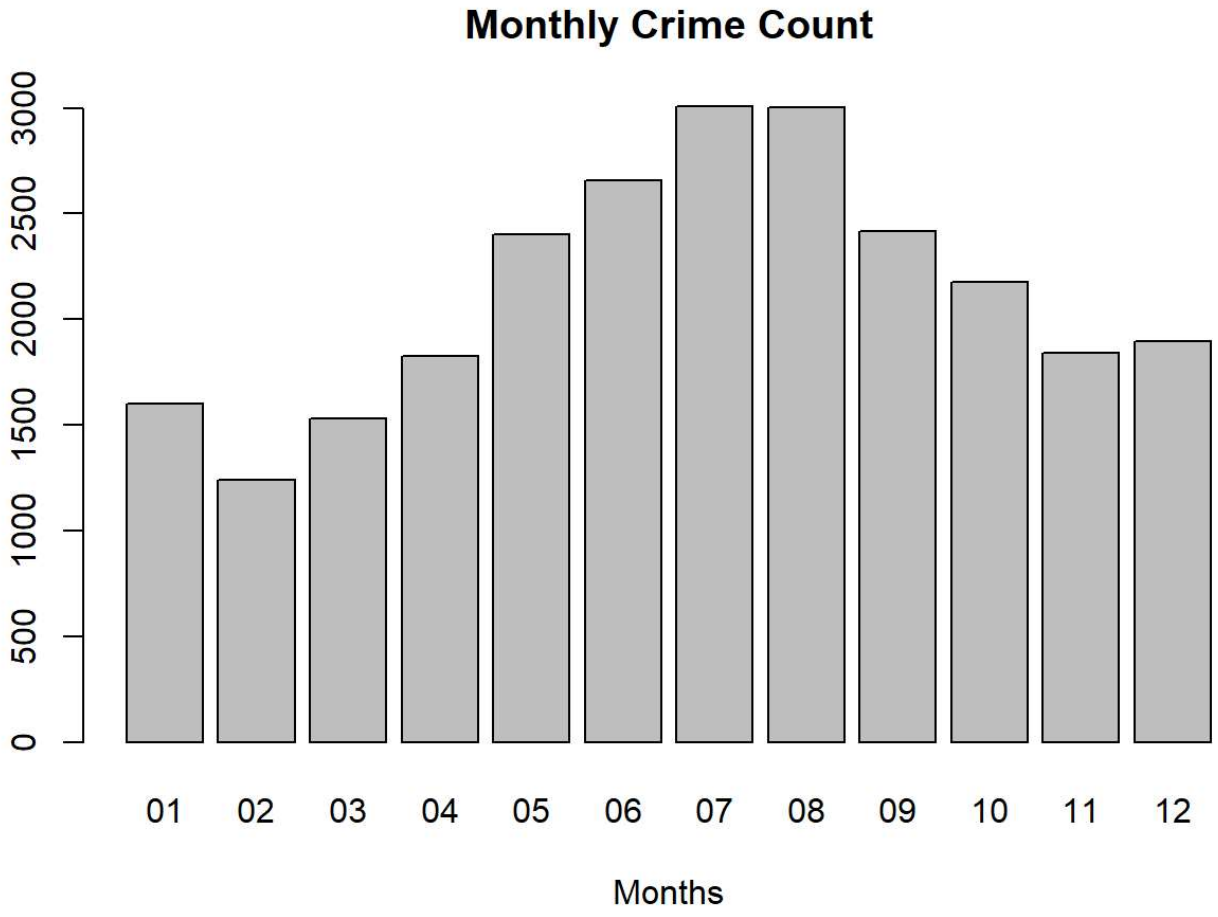
Black people are the most victims reported by the New York data over the years.

```
age_murder = ny_data_mod %>% count(VIC_AGE_GROUP, STATISTICAL_MURDER_FLAG)

ggplot(age_murder , aes(x = VIC_AGE_GROUP , y = n , fill = STATISTICAL_MURDER_FLAG)) + geom_bar
(stat='identity', color="black", position=position_dodge())
```



Victims who are in between 25-44 are most likely to be killed when shot at. This is synonymous with the overall observation that the victims who fall in this bracket are likely to face some sort of crime against them.

```
ny_data_mod$OCCUR_DATE = format(ny_data_mod$OCCUR_DATE, format="%m")

barplot(table(ny_data_mod$OCCUR_DATE), main = 'Monthly Crime Count', xlab = 'Months')
```



The months of July and August seem to be the most dangerous in terms of reported crimes.

Cumulative Analysis

We can sum up the takeaways from the analysis above as follows:

1. The age group between 25-44 seems to be the most likely victims of some sort of crime including murder. This observation might be skewed as the same age bracket is mainly composed of the prime working class so more exposure is also there.
2. The most dangerous area seems to be Brooklyn closely followed by Bronx.
3. The most deadly time of the year seems be the summers with the crime rate peaking in July and August so an additional police unit might be a good idea during these months especially.

Conclusion

A very basic analysis was carried out above which can be further improved by integrating the time of a particular crime and the precinct to effectively utilize and reinforce certain police units depending on the crime rate in the areas as well the time of the year.