NYPD_Document

MDSD_SB

2023-07-12

NEW YORK SHOOTING INCIDENT DATA REPORT

In this assignment we took New York Police Department Shooting Indident data from the year 2006-2022 for data analysis.

PROJECT STEP 1: How to import Dataset in a reproducible manner

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

```
## Rows: 27312 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.
```

```
summary(dataset)
```

```
##   INCIDENT_KEY        OCCUR_DATE          OCCUR_TIME           BORO
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms        Class :character
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :120860536                      Mode  :numeric
##  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       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              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
```

From the summary, we can see there are 2 missing values in the Jurisdiction_code column and 10 missing values in longitude and latitude columns of the dataset.

PROJECT STEP 2: Tidying and Transforming the data

TIDYING

Code to find out the row numbers of the missing data

```
which(is.na(dataset$JURISDICTION_CODE))
```

```
## [1]  3031 19981
```

```
which(is.na(dataset$Latitude ))
```

```
##  [1]  1407 25598 25599 25833 25939 26274 26742 26815 26876 27206
```

```
which(is.na(dataset$Longitude))
```

```
##  [1]  1407 25598 25599 25833 25939 26274 26742 26815 26876 27206
```

Since we have the row numbers with missing data,we impute the missing values by substituting each of them with an estimate.

```
dataset[3031,'JURISDICTION_CODE']=0.3269
dataset[19981,'JURISDICTION_CODE']=0.3269
dataset[1407,  'Latitude'] = 40.74
dataset[25598, 'Latitude'] = 40.74
dataset[25599, 'Latitude'] = 40.74
dataset[25833, 'Latitude'] = 40.74
dataset[25939, 'Latitude'] = 40.74
dataset[26274, 'Latitude'] = 40.74
dataset[26742, 'Latitude'] = 40.74
dataset[26815, 'Latitude'] = 40.74
dataset[26876 ,'Latitude'] = 40.74
dataset[27206, 'Latitude'] = 40.74
dataset[1407,  'Longitude'] = -73.91
dataset[25598, 'Longitude'] = -73.91
dataset[25599, 'Longitude'] = -73.91
dataset[25833, 'Longitude'] = -73.91
dataset[25939, 'Longitude'] = -73.91
dataset[26274, 'Longitude'] = -73.91
dataset[26742 ,'Longitude'] = -73.91
dataset[26815 ,'Longitude'] = -73.91
dataset[26876 ,'Longitude'] = -73.91
dataset[27206, 'Longitude'] = -73.91
```

We can now see no missing values in the dataset.

```
summary(dataset)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO          
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312      
##  1st Qu.: 63860880   Class :character   Class1:hms        Class :character  
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character  
##  Mean   :120860536                      Mode  :numeric                      
##  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                      
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical          Length:27312
##  Class :character   FALSE:22046            Class :character
##  Mode  :character   TRUE :5266             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
##    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
```

```
head(dataset)
```

```
## # A tibble: 6 x 21
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     LOC_OF_OCCUR_DESC PRECINCT
##          <dbl> <chr>      <time>     <chr>    <chr>                <dbl>
## 1    228798151 05/27/2021 21:30      QUEENS   <NA>                   105
## 2    137471050 06/27/2014 17:40      BRONX    <NA>                    40
## 3    147998800 11/21/2015 03:56      QUEENS   <NA>                   108
## 4    146837977 10/09/2015 18:30      BRONX    <NA>                    44
## 5     58921844 02/19/2009 22:58      BRONX    <NA>                    47
## 6    219559682 10/21/2020 21:36      BROOKLYN <NA>                    81
## # i 15 more variables: JURISDICTION_CODE <dbl>, LOC_CLASSFCTN_DESC <chr>,
## #   LOCATION_DESC <chr>, STATISTICAL_MURDER_FLAG <lgl>, PERP_AGE_GROUP <chr>,
## #   PERP_SEX <chr>, PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>,
## #   VIC_RACE <chr>, X_COORD_CD <dbl>, Y_COORD_CD <dbl>, Latitude <dbl>,
## #   Longitude <dbl>, Lon_Lat <chr>
```

TRANSFORMING

Removing unwanted and repeated columns

Most of the attributes or columns have missing entries which dont contribute much for data exploration, so they were removed.

```
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
```

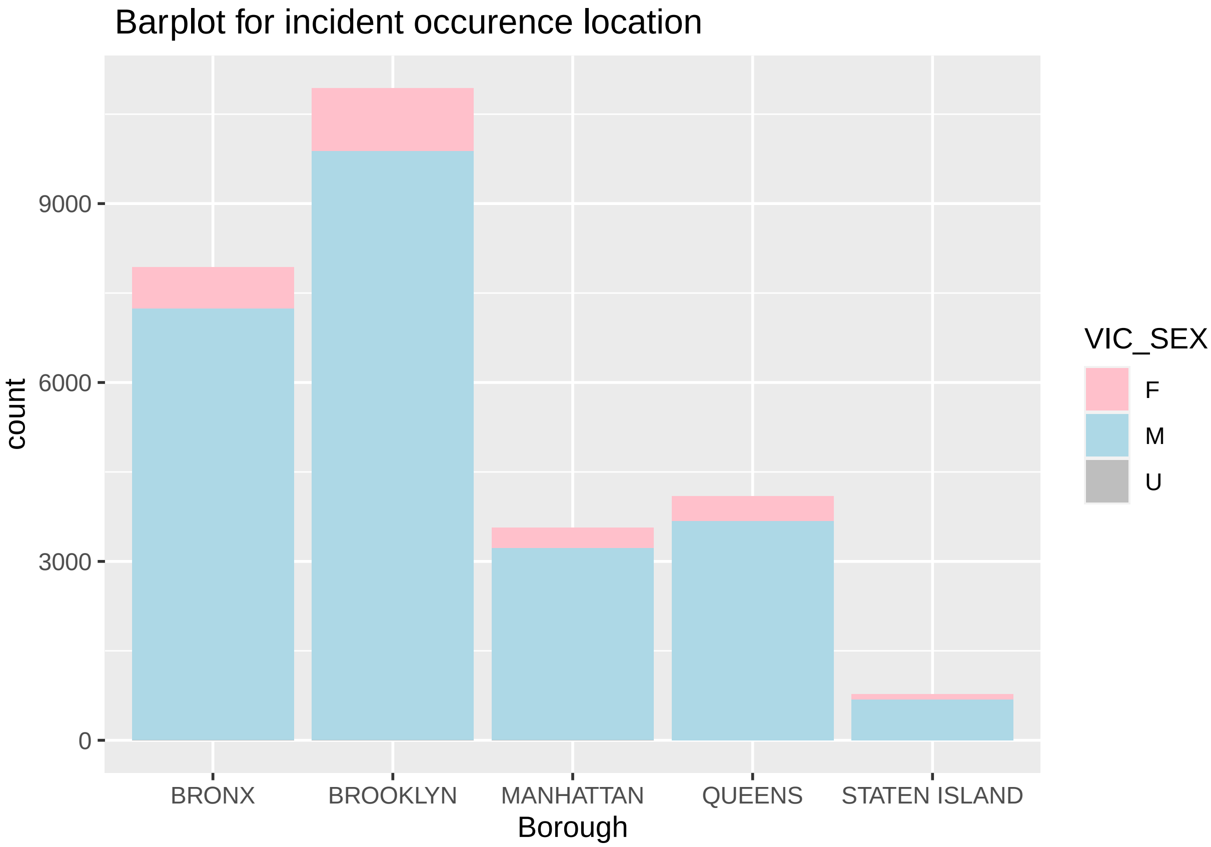
```
dataset2 <- dataset
dataset2 <- select(dataset, -c(PRECINCT,
                               LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC,STATISTICAL_MURDER_FLAG,
                               PERP_AGE_GROUP, PERP_SEX, PERP_RACE,
                               X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))

#dim(dataset) # 27312    21
#dim(dataset2) # 27312     9
head(dataset2)
```

```
## # A tibble: 6 x 9
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO     JURISDICTION_CODE LOCATION_DESC
##          <dbl> <chr>      <time>     <chr>                <dbl> <chr>
## 1    228798151 05/27/2021 21:30      QUEENS                   0 <NA>
## 2    137471050 06/27/2014 17:40      BRONX                    0 <NA>
## 3    147998800 11/21/2015 03:56      QUEENS                   0 <NA>
## 4    146837977 10/09/2015 18:30      BRONX                    0 <NA>
## 5     58921844 02/19/2009 22:58      BRONX                    0 <NA>
## 6    219559682 10/21/2020 21:36      BROOKLYN                 0 <NA>
## # i 3 more variables: VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>
```

Project Step 3: Visualizations and Analysis

1.Plot a barplot for incident location(BOROUGH)

```
library(ggplot2)
ggplot(dataset2,aes(x=BORO, fill= VIC_SEX )) +
  labs(x = " Borough ",title=" Barplot for incident occurence location ")+
  geom_bar(position='stack')+
  scale_fill_manual(values=c('pink','lightblue','grey'))
```

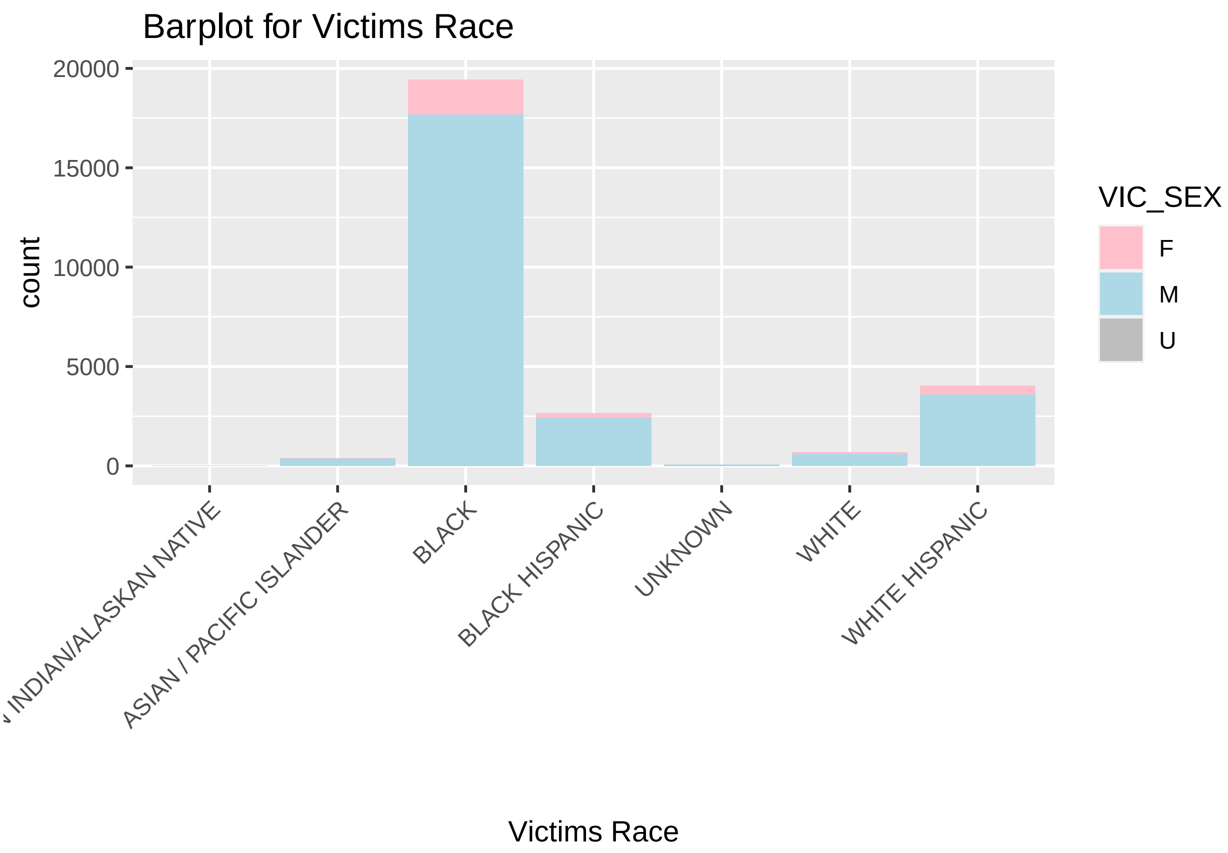


Analysis

From the above plot, most of the incidents took place in Brooklyn and Bronx compared to other 3 places. Also the ratio of males victims is higher than female victims.

2.Plot a barplot for victims race

```
library(ggplot2)
ggplot(dataset2, aes(x=VIC_RACE, fill= VIC_SEX)) +
  labs(x = " Victims Race ",title=" Barplot for Victims Race ")+
  geom_bar(position='stack')+
  scale_fill_manual(values=c('pink','lightblue','grey'))+theme(axis.text.x = element_text(angle = 45, v
```

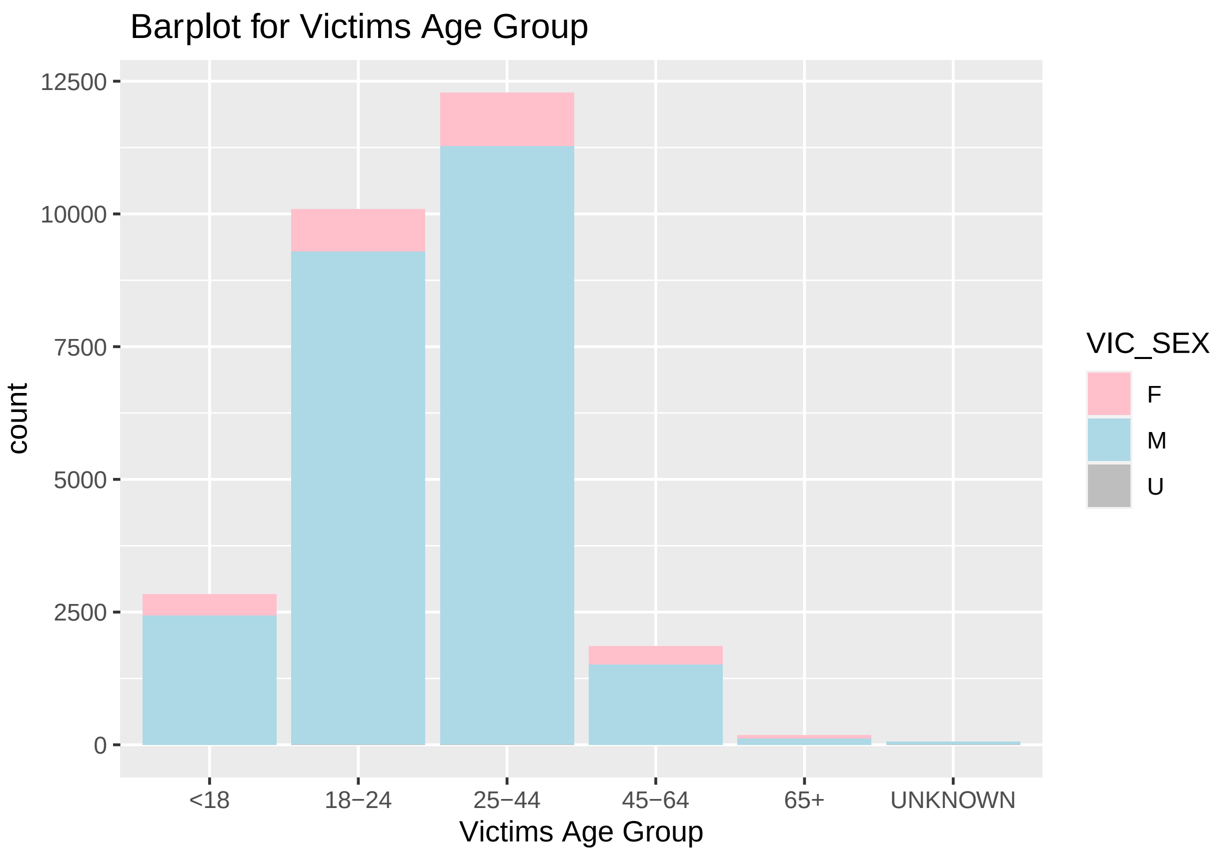


Analysis

From the above plot, we can say that black people are the highest victims, followed by white hispanic and black hispanic. Racial disparity existence is evident from the plot.

3.Plot a barplot for victims age group

```
ggplot(dataset2[dataset2$VIC_AGE_GROUP!=1022,],aes(x=VIC_AGE_GROUP, fill= VIC_SEX)) +
  labs(x = " Victims Age Group ",title=" Barplot for Victims Age Group ")+
  geom_bar(position='stack')+
  scale_fill_manual(values=c('pink','lightblue','grey'))
```

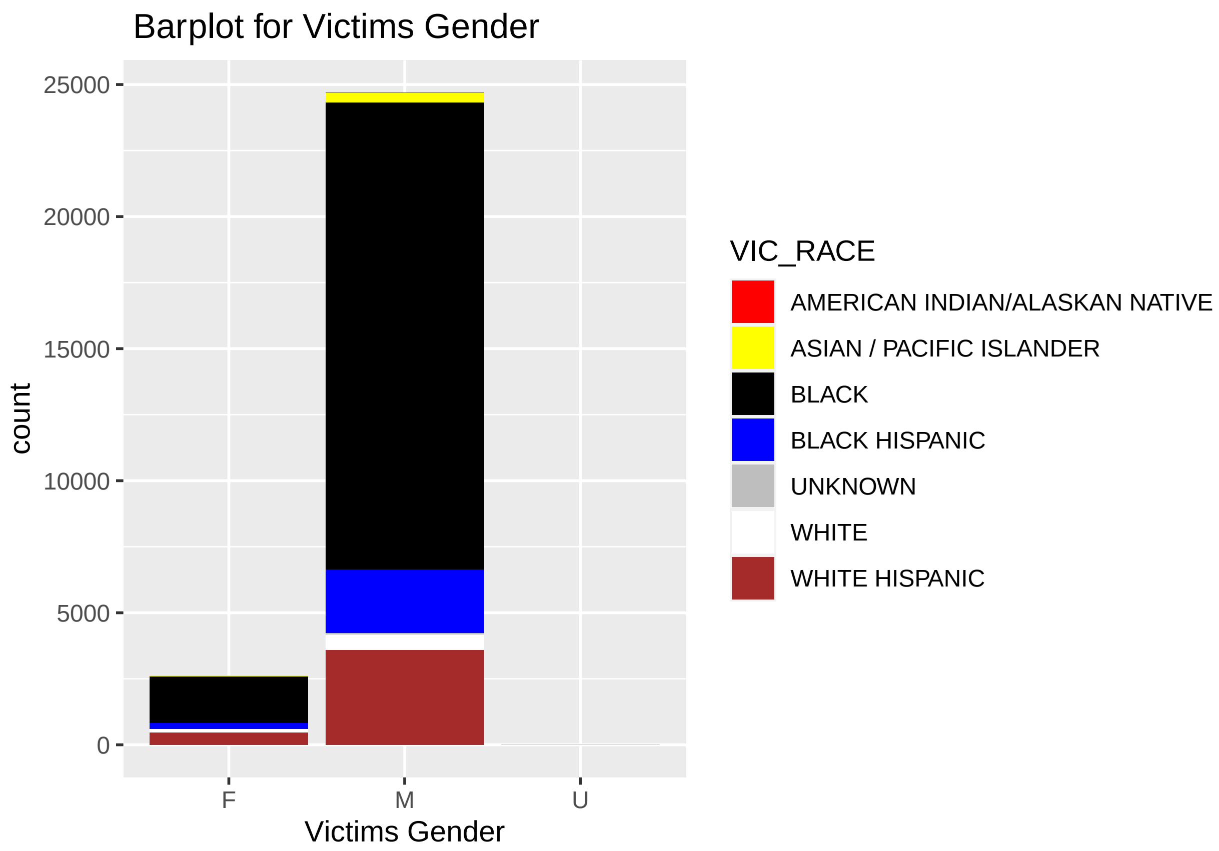


Analysis

From the above plot, most of the victims are in the age group of 18-45, due to the fact that they are the most active and independent age group to stay out and engage in various activities.Also most of the victims are males.

4. Plot a barplot for victims gender

```
ggplot(dataset2,aes(x=VIC_SEX, fill= VIC_RACE)) +
  labs(x = " Victims Gender ",title=" Barplot for Victims Gender ")+
  geom_bar(position='stack')+
  scale_fill_manual(values=c('red','yellow','black','blue','grey','white','brown'))
```



Analysis

From the above plot,it is clear that males are the most targetted victims and among them are black race males.Even among the females, even though they are less than males the ratio of balck females is high suggesting them as the targetted race.

How to extract year from the DATE

```
library(lubridate)
```

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

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

```
dataset2$Year <- format(as.POSIXct(dataset$OCCUR_DATE, format = "%m/%d/%Y "), format="%Y")
cases_by_boro <- dataset2 %>% group_by(BORO, Year) %>% summarize (Cases = n())
```

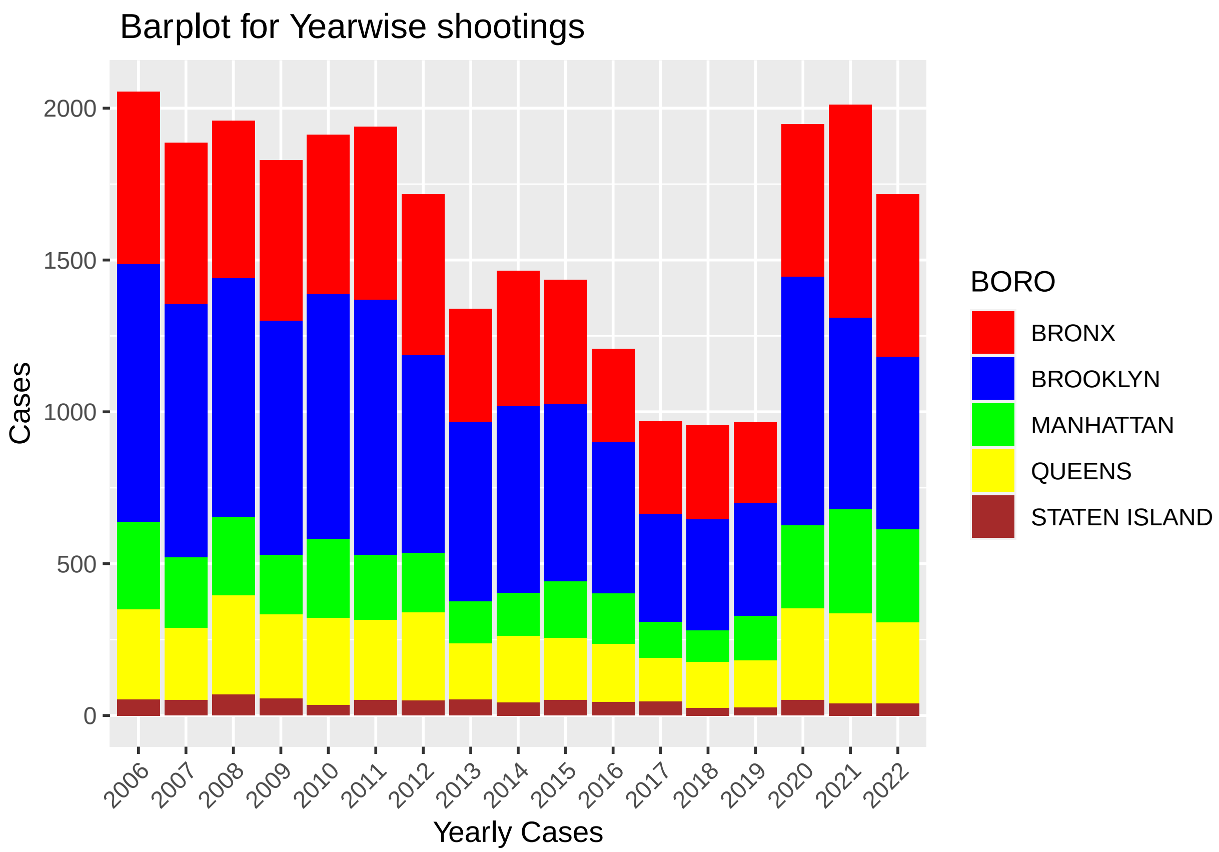
```
## `summarise()` has grouped output by 'BORO'. You can override using the
## `.groups` argument.
```

```
cases_by_boro
```

```
## # A tibble: 85 x 3
## # Groups:   BORO [5]
##    BORO  Year  Cases
##    <chr> <chr> <int>
##  1 BRONX 2006    568
##  2 BRONX 2007    533
##  3 BRONX 2008    520
##  4 BRONX 2009    529
##  5 BRONX 2010    525
##  6 BRONX 2011    571
##  7 BRONX 2012    531
##  8 BRONX 2013    371
##  9 BRONX 2014    446
## 10 BRONX 2015    409
## # i 75 more rows
```

5. Plot a barplot for yearly cases

```
ggplot(cases_by_boro, aes(x= Year, y = Cases, fill = BORO))+
  labs(x = " Yearly Cases ",title=" Barplot for Yearwise shootings ")+
  geom_bar(stat = "identity")+
  scale_fill_manual(values=c('red','blue','green','yellow','brown'))+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))
```



Analysis

From the above plot,we can see that the number of cases declined between 2017-2019, and again increased during covid pandemic.

Modelling

```
glm.fit <- glm(STATISTICAL_MURDER_FLAG ~   PERP_RACE + VIC_RACE + VIC_SEX + VIC_AGE_GROUP , family= bin
#glm.fit <- glm(STATISTICAL_MURDER_FLAG ~ . , family= binomial, data= dataset )
summary(glm.fit)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ PERP_RACE + VIC_RACE +
##     VIC_SEX + VIC_AGE_GROUP, family = binomial, data = dataset)
##
## Coefficients:
##                                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)                              -12.78608  111.33724  -0.115  0.90857
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -10.93734  229.56939  -0.048  0.96200
## PERP_RACEASIAN / PACIFIC ISLANDER         0.89290    0.21249   4.202 2.64e-05
## PERP_RACEBLACK                            0.44595    0.11399   3.912 9.14e-05
## PERP_RACEBLACK HISPANIC                   0.41293    0.13225   3.122  0.00179
## PERP_RACEUNKNOWN                         -0.84249    0.14637  -5.756 8.62e-09
## PERP_RACEWHITE                            1.17987    0.17662   6.680 2.38e-11
## PERP_RACEWHITE HISPANIC                   0.63473    0.12287   5.166 2.39e-07
## VIC_RACEASIAN / PACIFIC ISLANDER         10.86255  111.33725   0.098  0.92228
## VIC_RACEBLACK                            10.71464  111.33717   0.096  0.92333
## VIC_RACEBLACK HISPANIC                   10.51037  111.33719   0.094  0.92479
## VIC_RACEUNKNOWN                          10.09813  111.33816   0.091  0.92773
## VIC_RACEWHITE                            10.79986  111.33722   0.097  0.92273
## VIC_RACEWHITE HISPANIC                   10.78845  111.33718   0.097  0.92281
## VIC_SEXM                                 -0.13260    0.05958  -2.226  0.02604
## VIC_SEXU                                 -0.25433    1.13013  -0.225  0.82195
## VIC_AGE_GROUP1022                       -10.80797  324.74370  -0.033  0.97345
## VIC_AGE_GROUP18-24                        0.30806    0.07259   4.244 2.20e-05
## VIC_AGE_GROUP25-44                        0.53152    0.07045   7.545 4.53e-14
## VIC_AGE_GROUP45-64                        0.61828    0.09241   6.691 2.22e-11
## VIC_AGE_GROUP65+                          0.85048    0.20015   4.249 2.14e-05
## VIC_AGE_GROUPUNKNOWN                      0.53914    0.35156   1.534  0.12514
##
## (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_SEXU
## 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
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 17951  on 18001  degrees of freedom
## Residual deviance: 17465  on 17980  degrees of freedom
##   (9310 observations deleted due to missingness)
## AIC: 17509
##
## Number of Fisher Scoring iterations: 11
```

Project Step 4: Conclusions

From the data we have,it can be concluded that the black males within the age group of 18-45 are mojority of the victims of shooting in the areas of New York. Most of the incidents took place at Brooklyn and Bronx. It is unclear whether the victims are visitors or residents of Newyork. To have a more clear understanding about the magnitude of gun violence, the given data which has lots of missing entries should be filled. Appropriate measures such as increased patrol, awareness of gun violence should be taken to reduce the number of race related shootings.