# NYC_INCIDENTS_REPORT

STUDENT

2024-03-01

**NYC Incident Report**

This report is the data analysis of every shooting incident that occurred in New York City, NEW York, USA in 2006 through the end of the previous calendar year. Each record represents an incident of a crime in NYC and includes information about the Incident number, the location and time of occurrence. It also includes if this is a statistical murder flag. Information on the perpetrator and victim are included such as age, gender, and race.This data is public and can be found with the following link: https://catalog.data.gov/dataset?q=NYPD+Shooting+Incident+Data+%28Historic%29&sort=views_recent+desc&ext_location=&ext_bbox=&ext_prev_extent=

## STEP 1: Import Important Libraries

```
# install.packages("tidyverse")
library(tidyverse)
library(dplyr)
library(tinytex)
library(modelr)
```

## Step 2: Load Data

- `read_csv()` reads comma delimited files

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

```
## 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.
```

```
head(nyc)
```

```
## # 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>
```

```
view(nyc)
nyc_2= read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
```

```
## 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.
```

### Step 3: Tidy and Transform Data

I first eliminated the columns I do wish to use for this analysis, which are: **PRECINCT**,**JURISDICTION__CODE**,**LOCA**
**X__COORD__CD**, **Y__COORD__CD**, and **Lon__Lat**.

```
nyc =nyc %>% select(INCIDENT_KEY, OCCUR_DATE, OCCUR_TIME, BORO, STATISTICAL_MURDER_FLAG, PERP_AGE_GROUP


# After this, I need to see what is missing
lapply(nyc,function(x) sum(is.na(x)))
```

```
## $INCIDENT_KEY
## [1] 0
##
## $OCCUR_DATE
## [1] 0
##
## $OCCUR_TIME
## [1] 0
##
## $BORO
## [1] 0
##
```

```
## $STATISTICAL_MURDER_FLAG
## [1] 0
##
## $PERP_AGE_GROUP
## [1] 9344
##
## $PERP_SEX
## [1] 9310
##
## $PERP_RACE
## [1] 9310
##
## $VIC_AGE_GROUP
## [1] 0
##
## $VIC_SEX
## [1] 0
##
## $VIC_RACE
## [1] 0
##
## $Latitude
## [1] 10
##
## $Longitude
## [1] 10
```

It is essential to know why some of the data is missing. It seems at the time of the data collection, some information was no reported or it was not known as the victim may have been unsure if the perpetrator was a male or female, their age, or their race. Also, if the data set was a collection of solved cases, it is likely the missing data points are due to the investigation not being done yet. Therefore, out of respect of the investigation and those involved, missing points will be parked as "unknowns"

Key data type conversion are:

- **INCIDENT_KEY** should be treated as a string.
- **BORO** should be treated as a factor.
- **PERP_AGE_GROUP** should be treated as a factor.
- **PERP_SEX** should be treated as a factor.
- **PERP_RACE** should be treated as a factor.
- **VIC_AGE_GROUP** should be treated as a factor.
- **VIC_SEX** should be treated as a factor.
- **VIC_RACE** should be treated as a factor

```
# Tidy and transform data
nyc_2 = nyc_2 %>% replace_na(list(PERP_AGE_GROUP = "Unknown", PERP_SEX = "Unknown", PERP_RACE= "Unknown

#Clean up data
nyc_2$PERP_AGE_GROUP = recode(nyc_2$PERP_AGE_GROUP, UNKNOWN="Unknown")
nyc_2$PERP_SEX=recode(nyc_2$PERP_SEX, U= "Unknown")
nyc_2$PERP_RACE=recode(nyc_2$PERP_RACE, UNKNOWN="Unknown")
nyc_2$VIC_SEX=recode(nyc_2$VIC_SEX, U="Unknown")
nyc_2$VIC_RACE=recode(nyc_2$VIC_RACE, UNKNOWN="Unknown")
nyc_2$INCIDENT_KEY= as.character(nyc_2$INCIDENT_KEY)
```

```r
nyc_2$BORO=as.factor(nyc_2$BORO)
nyc_2$PERP_AGE_GROUP=as.factor(nyc_2$PERP_AGE_GROUP)
nyc_2$PERP_SEX= as.factor(nyc_2$PERP_SEX)
nyc_2$PERP_RACE=as.factor(nyc_2$PERP_RACE)
nyc_2$VIC_AGE_GROUP=as.factor(nyc_2$VIC_AGE_GROUP)
nyc_2$VIC_SEX=as.factor(nyc_2$VIC_SEX)
nyc_2$VIC_RACE= as.factor(nyc_2$VIC_RACE)

#Symmary of changes
 summary(nyc_2)
```

```
##  INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME                   BORO
##  Length:27312       Length:27312       Length:27312      BRONX        : 7937
##  Class :character   Class :character   Class1:hms        BROOKLYN     :10933
##  Mode  :character   Mode  :character   Class2:difftime   MANHATTAN    : 3572
##                                        Mode  :numeric    QUEENS       : 4094
##                                                          STATEN ISLAND:  776
##
##
##  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     PERP_SEX
##  Length:27312       Mode :logical           Unknown:12492   (null) :  640
##  Class :character   FALSE:22046             18-24  : 6222   F      :  424
##  Mode  :character   TRUE :5266              25-44  : 5687   M      :15439
##                                             <18    : 1591   Unknown:10809
##                                             (null) :  640
##                                             45-64  :  617
##                                             (Other):   63
##           PERP_RACE     VIC_AGE_GROUP      VIC_SEX
##  BLACK         :11432   <18    : 2839   F      : 2615
##  Unknown       :11146   1022   :    1   M      :24686
##  WHITE HISPANIC: 2341   18-24  :10086   Unknown:   11
##  BLACK HISPANIC: 1314   25-44  :12281
##  (null)        :  640   45-64  : 1863
##  WHITE         :  283   65+    :  181
##  (Other)       :  156   UNKNOWN:   61
##                             VIC_RACE       X_COORD_CD        Y_COORD_CD
##  AMERICAN INDIAN/ALASKAN NATIVE:   10   Min.   : 914928   Min.   :125757
##  ASIAN / PACIFIC ISLANDER      :  404   1st Qu.:1000029   1st Qu.:182834
##  BLACK                         :19439   Median :1007731   Median :194487
##  BLACK HISPANIC                : 2646   Mean   :1009449   Mean   :208127
##  Unknown                       :   66   3rd Qu.:1016838   3rd Qu.:239518
##  WHITE                         :  698   Max.   :1066815   Max.   :271128
##  WHITE HISPANIC                : 4049
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:27312
```

```
##  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
##  NA's   :10      NA's   :10
```

```
# The below tidying occurred during visual analysis

# Remove extreme values in data
nyc_2 = nyc_2[!(nyc_2$PERP_AGE_GROUP=="1020" | nyc_2$PERP_AGE_GROUP=="224" | nyc_2$PERP_AGE_GROUP=="940
```

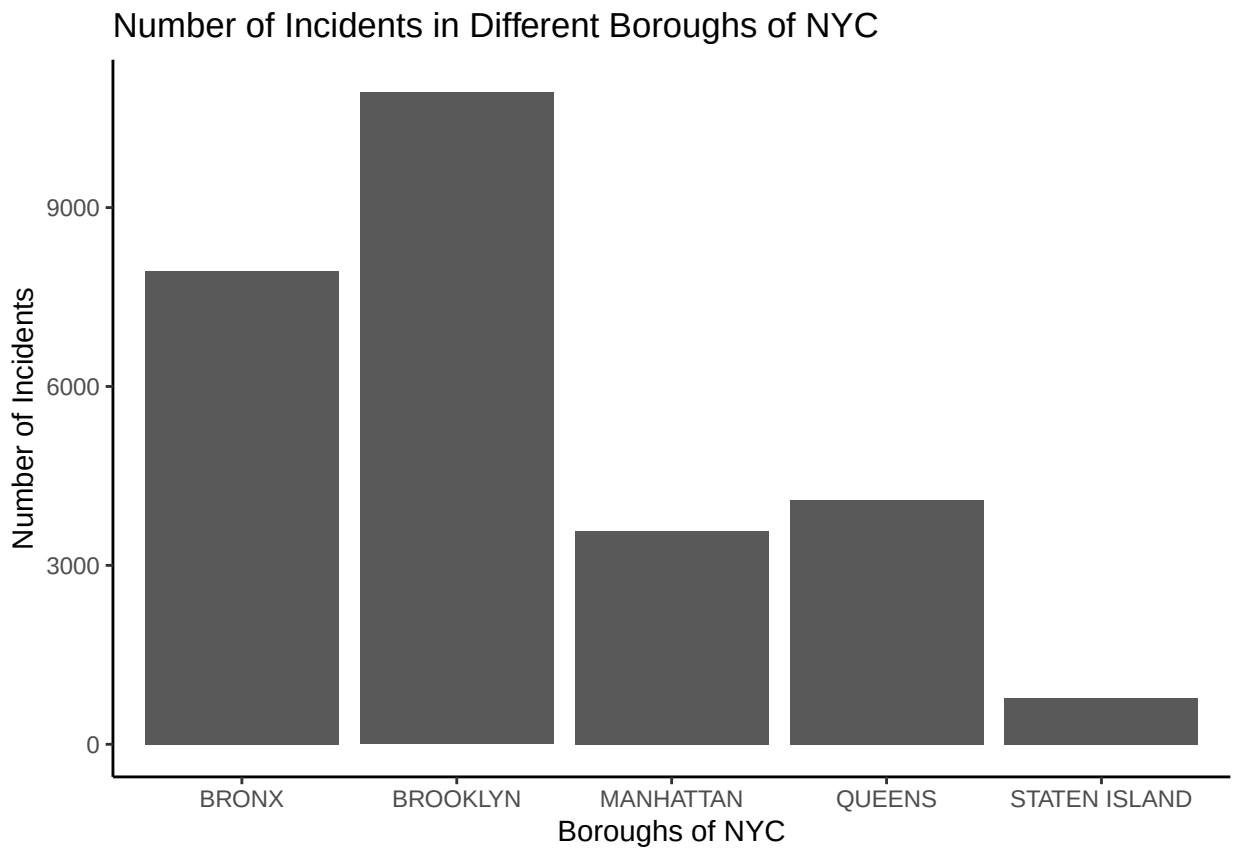
## Step 4: Visualizations and Analysis

1. Which park of New York is the most dangerous?

Based on the bar graph, Brooklyn is the most dangerous borough, followed by the Bronx, Queens, Manhattan, and Staten Island. Upon this analysis, I thought about if Brooklyn was really the most dangerous neighborhood. Therefore, outside analysis will be done by taking data on the size and populations of these boroughs and doing an analysis on number of incidents based on square mile and based on number of people. Based on data completed outside of R that will be part of the attachments, it was found that the Bronx may be the most dangerous borough based on the number of incidents per square mile and by person. Future analysis will need to be taken in this area.

```
g <- ggplot(nyc_2, aes(x=BORO))+ geom_bar()+ labs(title ="Number of Incidents in Different Boroughs of

g
```



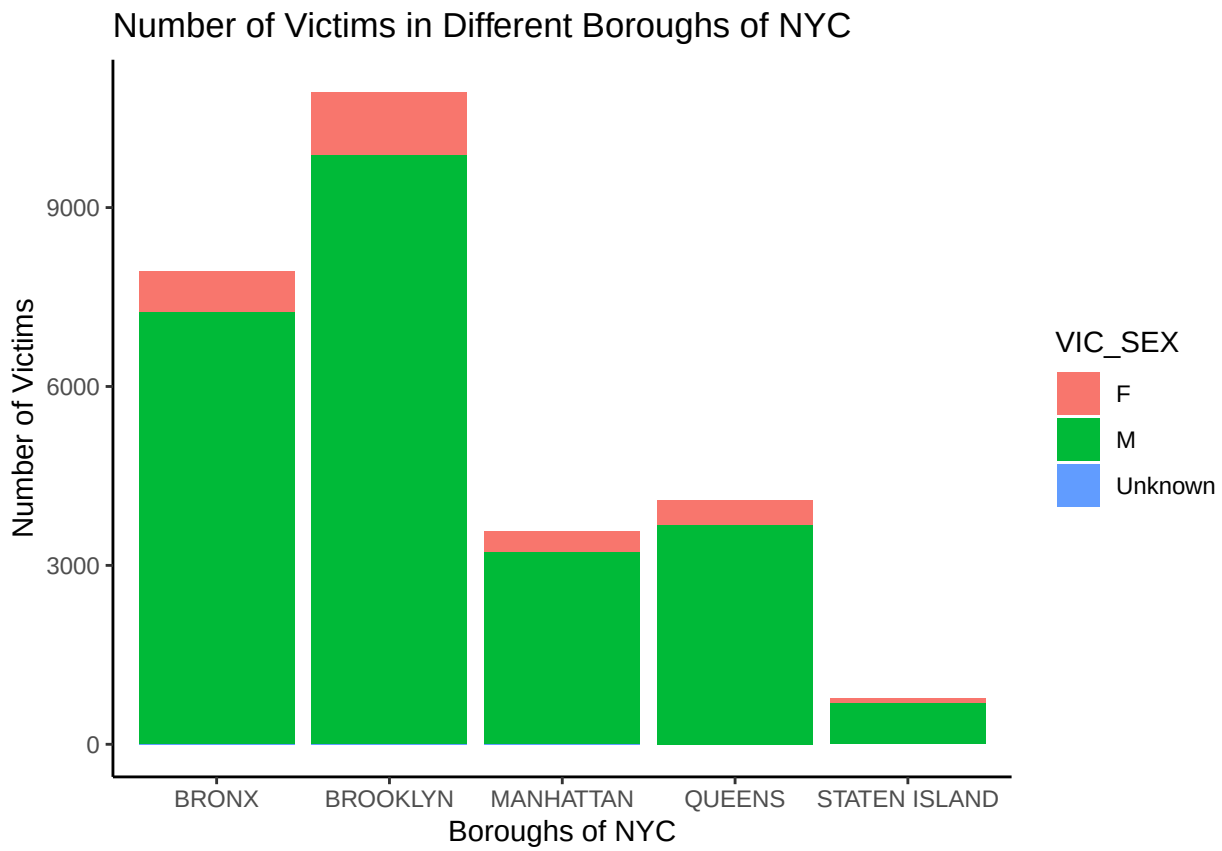

```
table(nyc_2$BORO,nyc_2$STATISTICAL_MURDER_FLAG)
```

```
##
##                  FALSE TRUE
##   BRONX           6393 1542
##   BROOKLYN        8810 2122
##   MANHATTAN       2942  630
##   QUEENS          3284  810
##   STATEN ISLAND    614  162
```

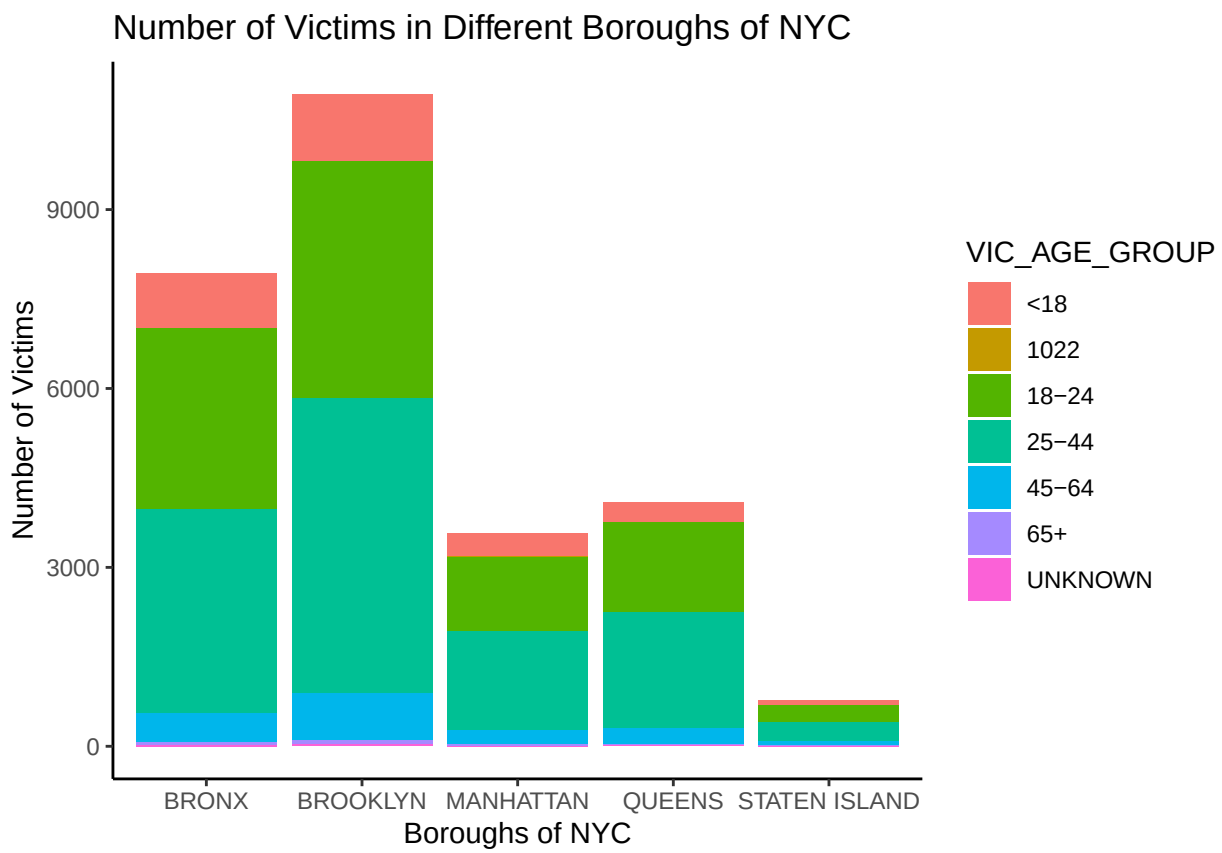
2. Is there a correlation between perpetrator and victim?

I am curious if there was any correlation between location and victim number based on the victim demographics. I therefore analyzed the Borough location based on the victim's race, gender, and age. There are some interesting correlations between these boroughs and victim profiles.

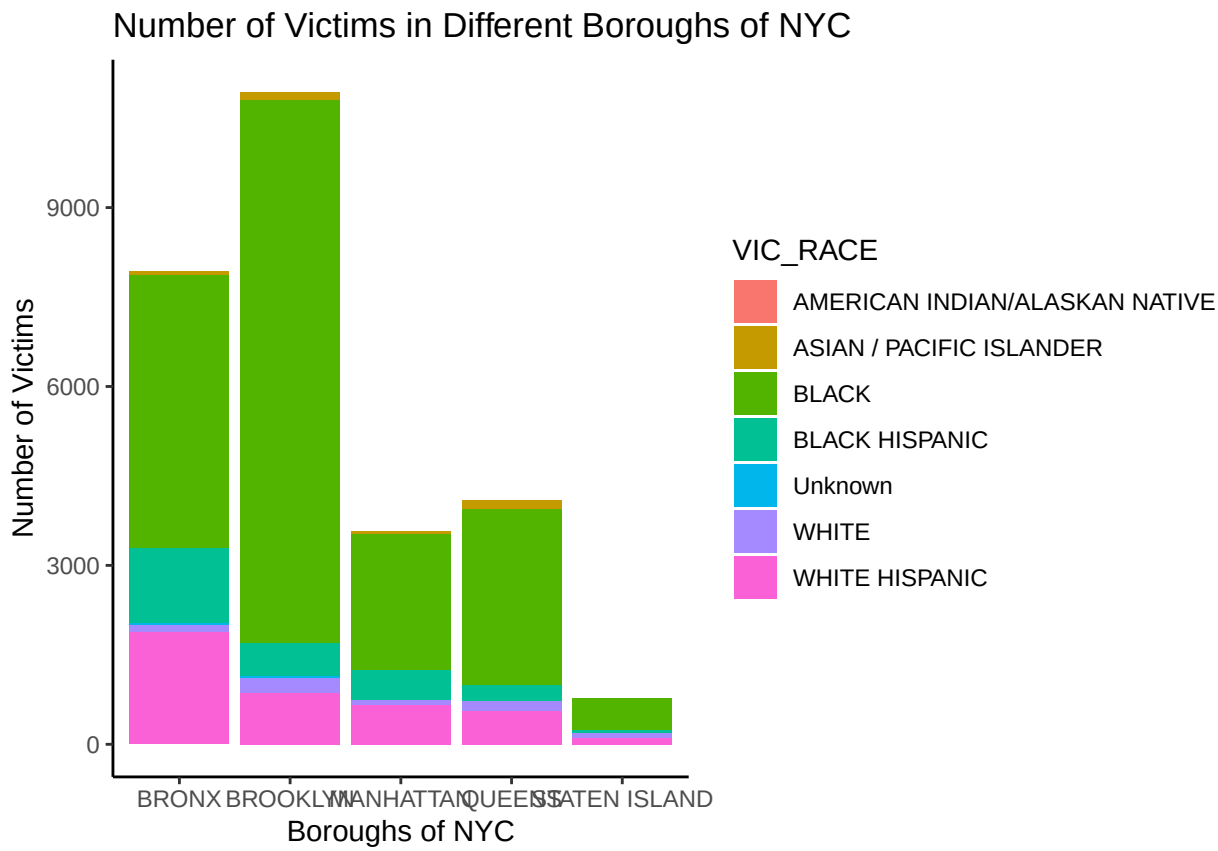
```
#Boroughs vs Victim Sex
 g <- ggplot(nyc_2, aes(x=BORO, fill=VIC_SEX))+ geom_bar()+ labs(title ="Number of Victims in Different
g
```



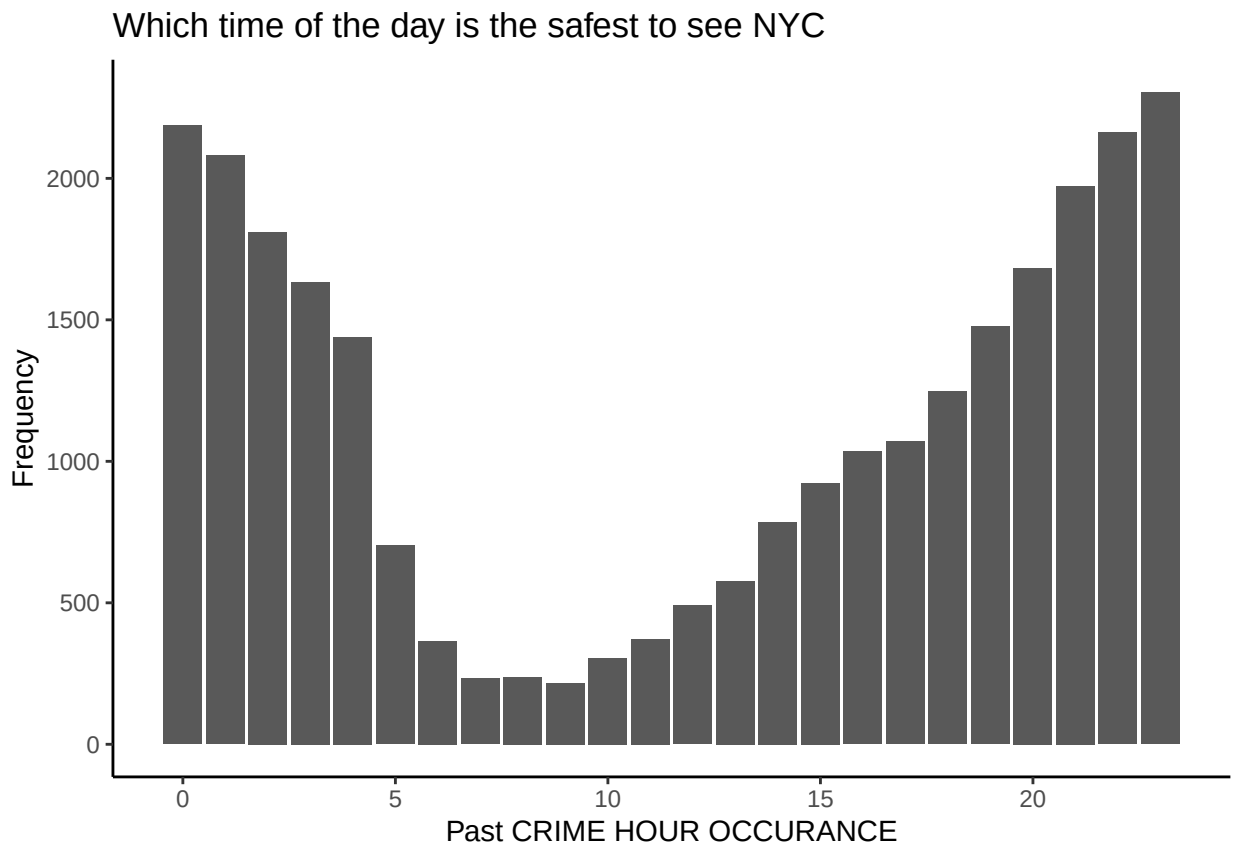

```
#Boroughs vs Victim Age
g <- ggplot(nyc_2, aes(x=BORO, fill=VIC_AGE_GROUP))+ geom_bar()+ labs(title ="Number of Victims in Diff
g
```




```
#Boroughs vs Victim RACE
g <- ggplot(nyc_2, aes(x=BORO, fill=VIC_RACE))+ geom_bar()+ labs(title ="Number of Victims in Different
g
```




```
nyc_2$OCCUR_HOUR=hour(hms(as.character(nyc_2$OCCUR_TIME)))
nyc_4 = nyc_2 %>%
  group_by(OCCUR_HOUR) %>%
  count()
 g <-ggplot(nyc_4, aes(x=OCCUR_HOUR,y=n))+geom_col()+labs(title = "Which time of the day is the safest
g
```




## Step 5: Identify Bias

Bias that occurred during this process was in selecting the topics to investigate. I became interested in investigating victim profiles because as a mother and an avid traveler, I want to make sure when I visit NYC, I keep safety as my top priority. It is important to keep bias out of data analysis as to not influence the data. It is important to also communicate these points during a presentation so this can be considered in data analysis. Also, I avoided using the perpetrator data due to bias. With all the cop crime stories in the news and the crimes and data showing racism and bias taught in the police academy, I avoided analyzing this data. Since there is a chance some of these crimes could have the perpetrators be falsely accused or committed, its important to further investigate the outcomes of the crimes in future data analysis.