NYPD Shooting Data Analysis

This report tries to analyze and answer questions related to rate of crime from NYPD incident reports since 2005.

This report tries to analuze the data based on racial profiles of the victims and the perpetrators.

IMPORING THE LIBRARIES:

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

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

IMPORTING THE DATA:

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

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   INCIDENT_KEY = col_double(),
##   OCCUR_DATE = col_character(),
##   OCCUR_TIME = col_time(format = ""),
##   BORO = col_character(),
##   PRECINCT = col_double(),
##   JURISDICTION_CODE = col_double(),
##   LOCATION_DESC = col_character(),
```

```
##   STATISTICAL_MURDER_FLAG = col_logical(),
##   PERP_AGE_GROUP = col_character(),
##   PERP_SEX = col_character(),
##   PERP_RACE = col_character(),
##   VIC_AGE_GROUP = col_character(),
##   VIC_SEX = col_character(),
##   VIC_RACE = col_character(),
##   X_COORD_CD = col_number(),
##   Y_COORD_CD = col_number(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

CLEANING THE DATA:

To clean the data, first we drop the columns that are not required for our analysis. In this case, we are dropping the columns X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat.

```
shooting_data <- shooting_data %>% select(-c(X_COORD_CD,Y_COORD_CD,Latitude,Longitude,Lon_Lat))
summary(shooting_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:23568       Length:23568       Length:23568
##  1st Qu.: 55317014   Class :character   Class1:hms         Class :character
##  Median : 83365370   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :102218616                      Mode  :numeric
##  3rd Qu.:150772442
##  Max.   :222473262
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:23568       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:19080
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4488
##  Mean   : 66.21   Mean   :0.3323
##  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:23568       Length:23568       Length:23568       Length:23568
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE
##  Length:23568       Length:23568
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

Here, we are changing the datatype of OCCUR_DATE from chr to date. We do this by using the lubridate library.

```
shooting_data <- shooting_data %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

Checking for NULL values and omit them.

```
colSums(is.na(shooting_data))
```

```
##            INCIDENT_KEY               OCCUR_DATE               OCCUR_TIME
##                       0                        0                        0
##                    BORO                  PRECINCT        JURISDICTION_CODE
##                       0                        0                        2
##           LOCATION_DESC STATISTICAL_MURDER_FLAG           PERP_AGE_GROUP
##                   13581                        0                     8459
##                PERP_SEX                PERP_RACE            VIC_AGE_GROUP
##                    8425                     8425                        0
##                 VIC_SEX                 VIC_RACE
##                       0                        0
```

```
na.omit(shooting_data)
```

```
## # A tibble: 6,843 x 14
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME BORO          PRECINCT JURISDICTION_CODE
##           <dbl> <date>     <time>     <chr>            <dbl>              <dbl>
##  1    204192600 2019-10-24 00:52      STATEN ISLAND      121                  0
##  2    193694863 2019-02-17 03:00      QUEENS             114                  2
##  3    201436772 2019-08-21 23:34      STATEN ISLAND      120                  0
##  4    201852654 2019-08-31 07:42      BRONX               45                  0
##  5    193939359 2019-02-24 23:20      BRONX               44                  2
##  6    199247701 2019-07-03 00:04      QUEENS             114                  2
##  7    199134406 2019-06-29 05:48      BROOKLYN            69                  0
##  8    204971625 2019-11-10 14:03      BROOKLYN            63                  0
##  9    200365034 2019-07-28 14:35      MANHATTAN           30                  2
## 10    199422329 2019-07-07 10:50      BROOKLYN            60                  0
## # ... with 6,833 more rows, and 8 more variables: 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>
```

ANALYSIS:

We try to analyze the number of shooting incidents that occur during daytime and at night. By analyzing this data, we find that the number of incidents that happened during daytime is considerably higher than the number of incidents that happened at night.

```
cases_at_daytime <- shooting_data %>% filter(hour(OCCUR_TIME) <= 22 | hour(OCCUR_TIME) >= 6)
cases_at_night <- shooting_data %>% filter(hour(OCCUR_TIME) <= 6 | hour(OCCUR_TIME) >= 22)
nrow(cases_at_daytime)
```

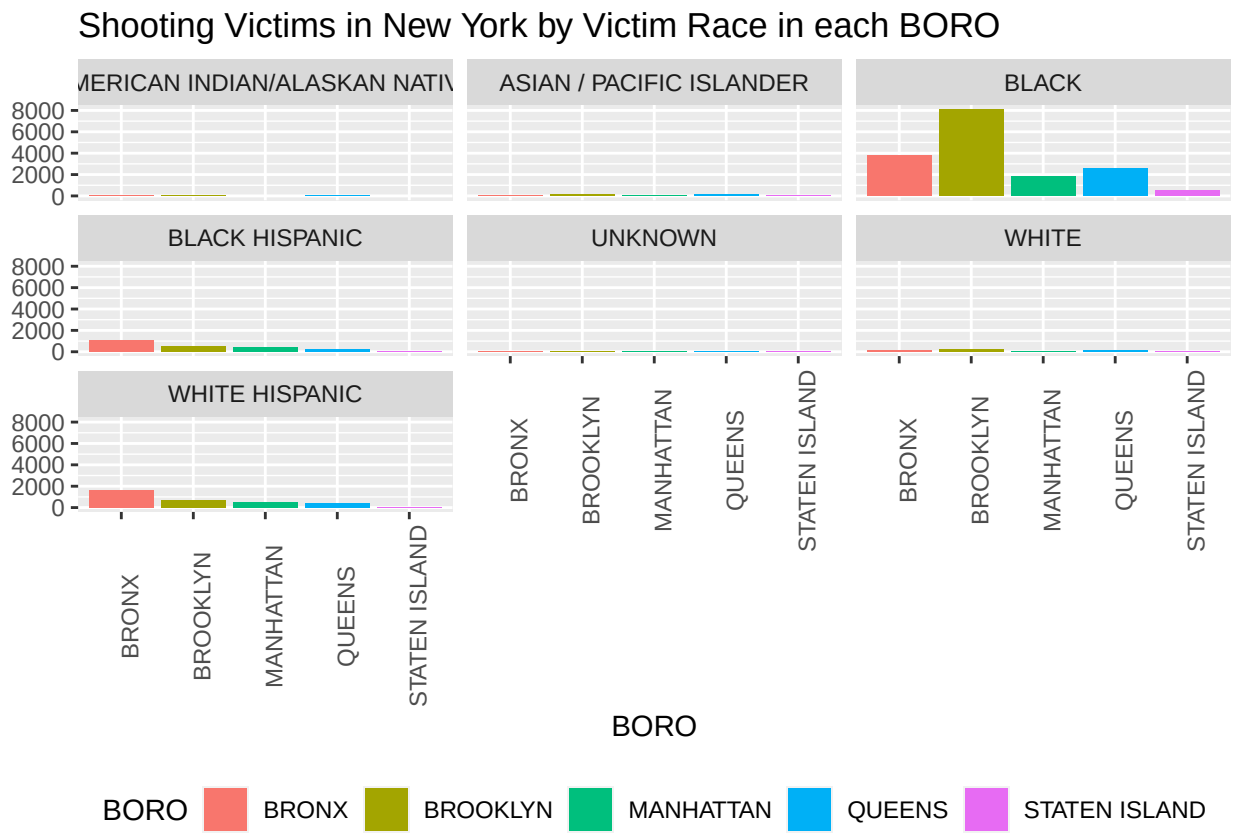
```
## [1] 23568
```

```
nrow(cases_at_night)
```

```
## [1] 12930
```

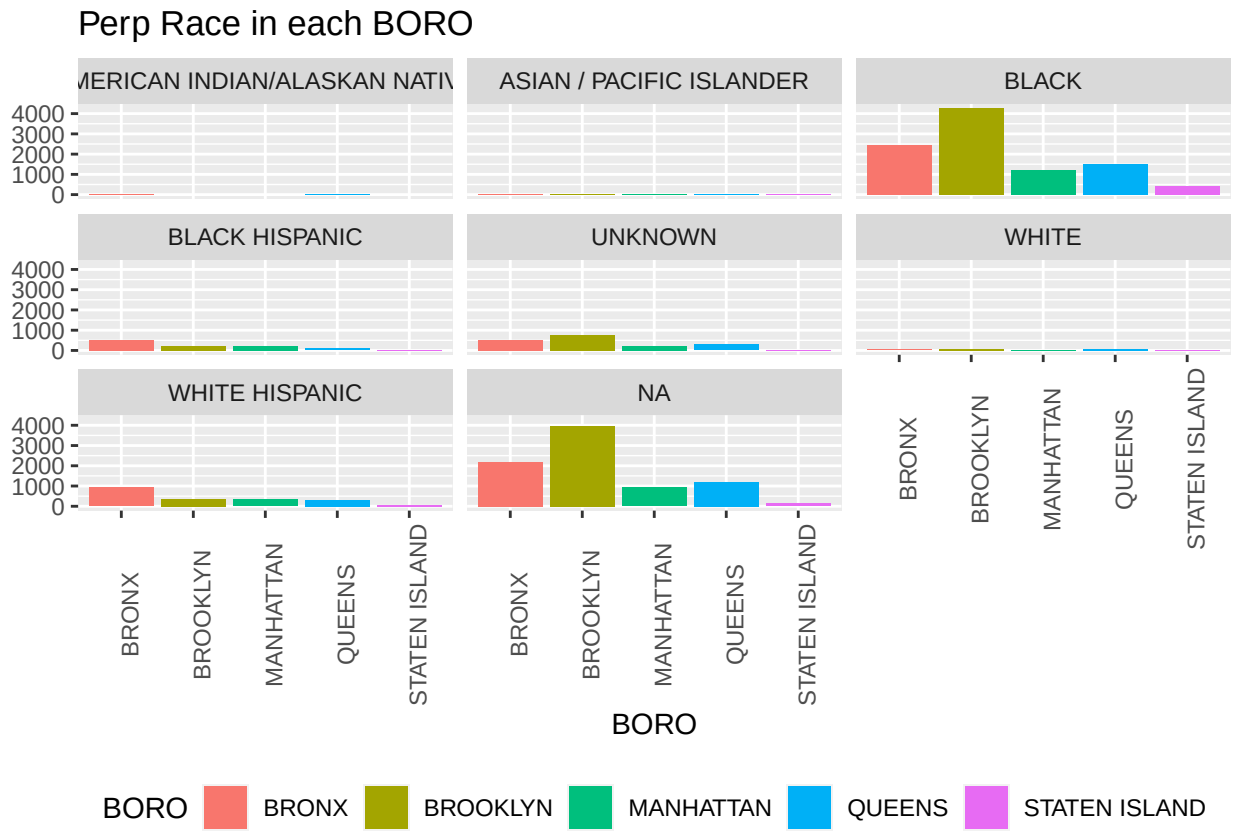
Visualizing the victim data by race in each BORO. By faceting the number of shooting incidents by victim race, we can see a clear breakdown of victims by race in each BORO.

```
ggplot(shooting_data) +
 geom_bar(aes(x = BORO, fill = BORO)) +
 facet_wrap(~VIC_RACE) +
 theme(legend.position = "bottom",
    axis.text.x = element_text(angle = 90)) +
 labs(title = "Shooting Victims in New York by Victim Race in each BORO", y = NULL)
```



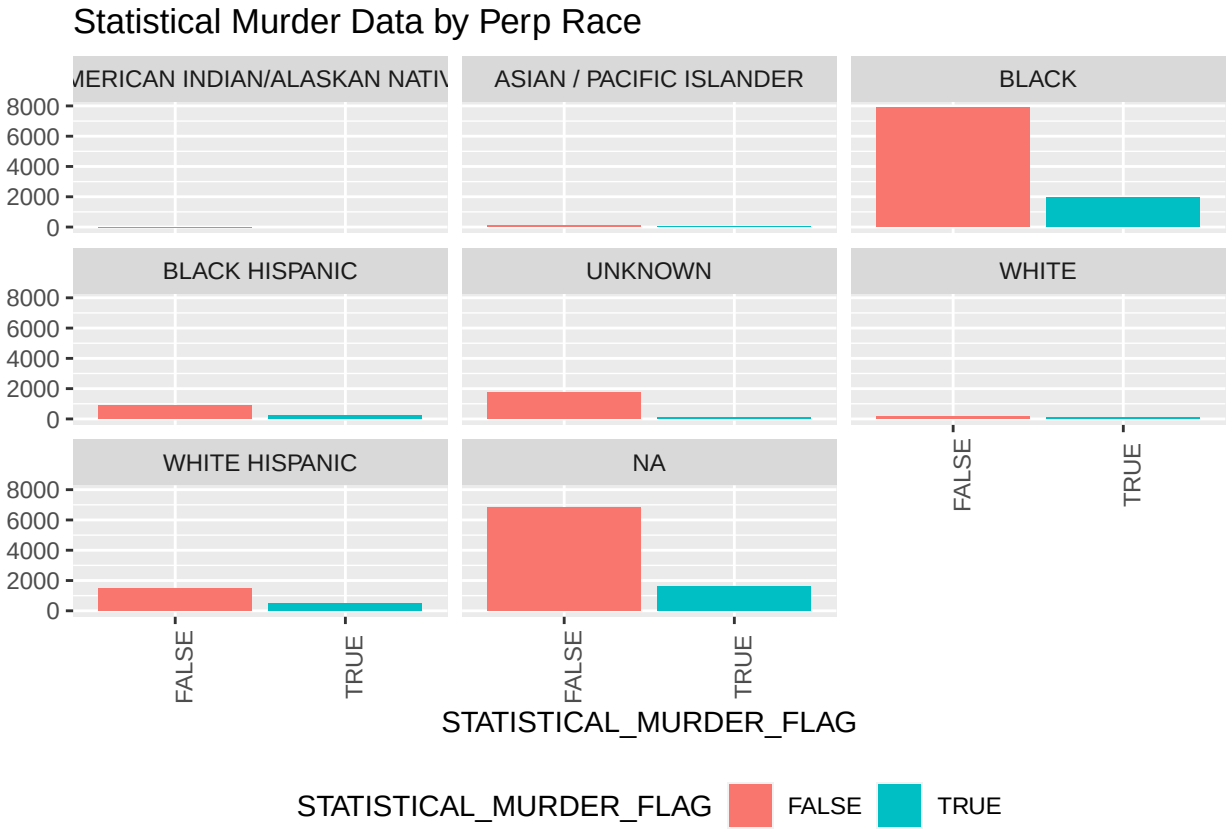
Similarly we visualize the Perp data according to race in each BORO.

```
ggplot(shooting_data) +
 geom_bar(aes(x = BORO, fill = BORO)) +
 facet_wrap(~PERP_RACE) +
 theme(legend.position = "bottom",
    axis.text.x = element_text(angle = 90)) +
 labs(title = "Perp Race in each BORO", y = NULL)
```



Here we analyze the statistical murder data by race.

```
ggplot(shooting_data) +
     geom_bar(aes(x = STATISTICAL_MURDER_FLAG, fill = STATISTICAL_MURDER_FLAG)) +
     facet_wrap(~PERP_RACE) +
     theme(legend.position = "bottom",
           axis.text.x = element_text(angle = 90)) +
     labs(title = "Statistical Murder Data by Perp Race", y = NULL)
```



CONCLUSION:

By analyzing the data, we find that the while the racial profile of majority of the perpetrators are black, the majority of victims are also black. This analysis has not considered the age group of the perpetrators. Not considering the age group if the perpetrators could be possible BIAS in this analysis, since it could possibly give more insights about statistical murder data.