NYPD Shooting Incidents

I. Dawud

2023-03-03

This document will analyze a list of shooting incidents that occurred in New York City starting in 2006 till 2021. Data is collected manually and reviewed by the Office of Management Analysis and Planning before it is posted to the NYPD website. The data being used today is from data.gov and is allowed for public access and use.

Importing Libraries

```
library(tidyverse)
library(lubridate)
library(chron)
```

Loading and Importing Data

Reading from a csv file from the cityofnewyork website and importing it into NYPDdata table.

```
NYPDFile = "NYPDdata.csv"

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

```
NYPDdata <- read.csv(url_in)
summary(NYPDdata)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME          BORO
## Min.   :  9953245   Length:25596       Length:25596       Length:25596
## 1st Qu.: 61593633   Class :character   Class :character   Class :character
## Median : 86437258   Mode  :character   Mode  :character   Mode  :character
## Mean   :112382648
## 3rd Qu.:166660833
## Max.   :238490103
##
##    PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
```

```
##   Min.   :  1.00   Min.   :0.0000   Length:25596      Length:25596
##   1st Qu.: 44.00   1st Qu.:0.0000   Class :character   Class :character
##   Median : 69.00   Median :0.0000   Mode  :character   Mode  :character
##   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
##
```

Clean Up Data

Remove unnecessary columns and define it as another dataframe JURISDICTION_CODE,X_COORD_CD,Y_COORD_CD
I also need to make sure to replace missing spaces with NA in order to make sure I do not have incorrect
analysis.

```r
# Remove unnecessary columns and define it as another dataframe JURISDICTION_CODE,X_COORD_CD,Y_COORD_CD

NYPDdata_cleaned <- subset(NYPDdata, select = -c(JURISDICTION_CODE,X_COORD_CD,Y_COORD_CD,Latitude,Longit

#There's a lot of missing data , replacing all the missing space with NA (Non applicable).

NYPDdata_cleaned <- replace(NYPDdata_cleaned, NYPDdata_cleaned=='1020', NA)
NYPDdata_cleaned <- replace(NYPDdata_cleaned, NYPDdata_cleaned=='224', NA)
NYPDdata_cleaned <- replace(NYPDdata_cleaned, NYPDdata_cleaned=='940', NA)


NYPDdata_cleaned <- replace(NYPDdata_cleaned, NYPDdata_cleaned=='', NA)
```

```r
NYPDdata_cleaned <- replace(NYPDdata_cleaned, NYPDdata_cleaned=='UNKNOWN', NA)
NYPDdata_cleaned <- replace(NYPDdata_cleaned, NYPDdata_cleaned=='U', NA)


summary(NYPDdata_cleaned)
```

```
##   OCCUR_DATE          OCCUR_TIME             BORO              PRECINCT
##  Length:25596       Length:25596       Length:25596       Min.   :  1.00
##  Class :character   Class :character   Class :character   1st Qu.: 44.00
##  Mode  :character   Mode  :character   Mode  :character   Median : 69.00
##                                                           Mean   : 65.87
##                                                           3rd Qu.: 81.00
##                                                           Max.   :123.00
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:25596       Length:25596            Length:25596
##  Class :character   Class :character        Class :character
##  Mode  :character   Mode  :character        Mode  :character
##
##
##
##    PERP_SEX           PERP_RACE          VIC_AGE_GROUP         VIC_SEX
##  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_RACE
##  Length:25596
##  Class :character
##  Mode  :character
##
##
##
```

Investigations

Investigation 1

What brououghs suffer from the most shootings since 2006?

```r
NYPD_by_boro <- NYPDdata_cleaned %>%
  group_by(BORO) %>%
  summarize(Total_Incidents=n(),
            .groups = 'drop')
NYPD_by_boro
```

```
## # A tibble: 5 x 2
```

```
##   BORO          Total_Incidents
##   <chr>                   <int>
## 1 BRONX                    7402
## 2 BROOKLYN                10365
## 3 MANHATTAN                3265
## 4 QUEENS                   3828
## 5 STATEN ISLAND             736
```

```
g <- ggplot(NYPDdata_cleaned, aes(x = BORO)) +
  geom_bar() +
  labs(title = "Boroughs of New York",
      x = "Boroughs",
      y = "Total Incidents") +
  theme_classic()
g
```

Boroughs of New York
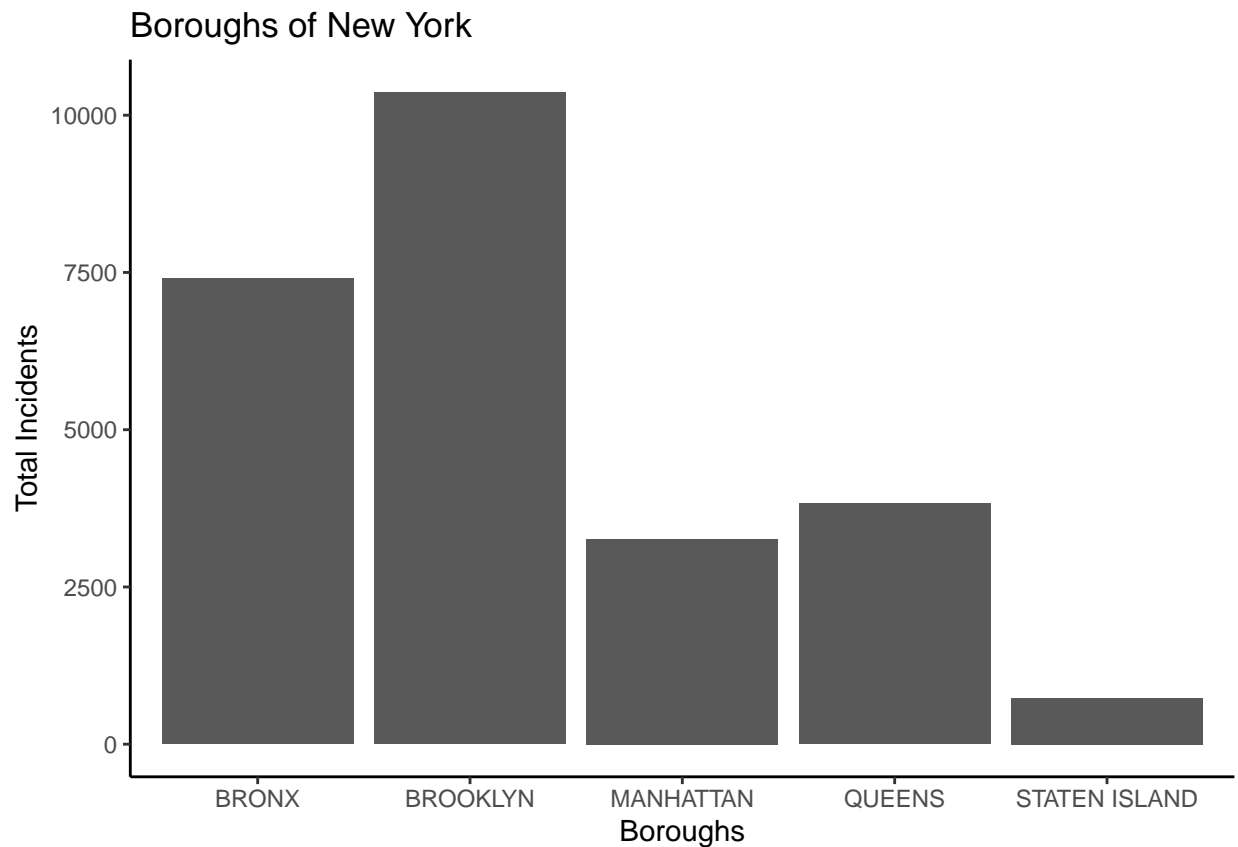
Conclusion

- We can see that Brooklyn is number one in the number of incidents that have occured. Staten Island shows to have the least reported incidents.

Investigation 2

4

Who are the victims being targeted in these incidents? Who are the perpetrators?

```r
NYPD_perp_race <- NYPDdata_cleaned %>%
  group_by(PERP_RACE) %>%
  summarize(Total=n(),
            .groups = 'drop')
NYPD_perp_race
```

```
## # A tibble: 7 x 2
##   PERP_RACE                   Total
##   <chr>                       <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE    2
## 2 ASIAN / PACIFIC ISLANDER        141
## 3 BLACK                         10668
## 4 BLACK HISPANIC                 1203
## 5 WHITE                           272
## 6 WHITE HISPANIC                 2164
## 7 <NA>                          11146
```

```r
NYPD_vic_race <- NYPDdata_cleaned %>%
  group_by(VIC_RACE) %>%
  summarize(Total=n(),
            .groups = 'drop')

NYPD_vic_race
```

```
## # A tibble: 7 x 2
##   VIC_RACE                    Total
##   <chr>                       <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE    9
## 2 ASIAN / PACIFIC ISLANDER        354
## 3 BLACK                         18281
## 4 BLACK HISPANIC                 2485
## 5 WHITE                           660
## 6 WHITE HISPANIC                 3742
## 7 <NA>                             65
```

```r
g <- ggplot(NYPDdata_cleaned, aes(y = VIC_RACE)) +
  geom_bar() +
  labs(title = "Race of Victims",
       y = "Race",
       x = "Total Incidents") +
  theme_classic()
g
```

Race of Victims
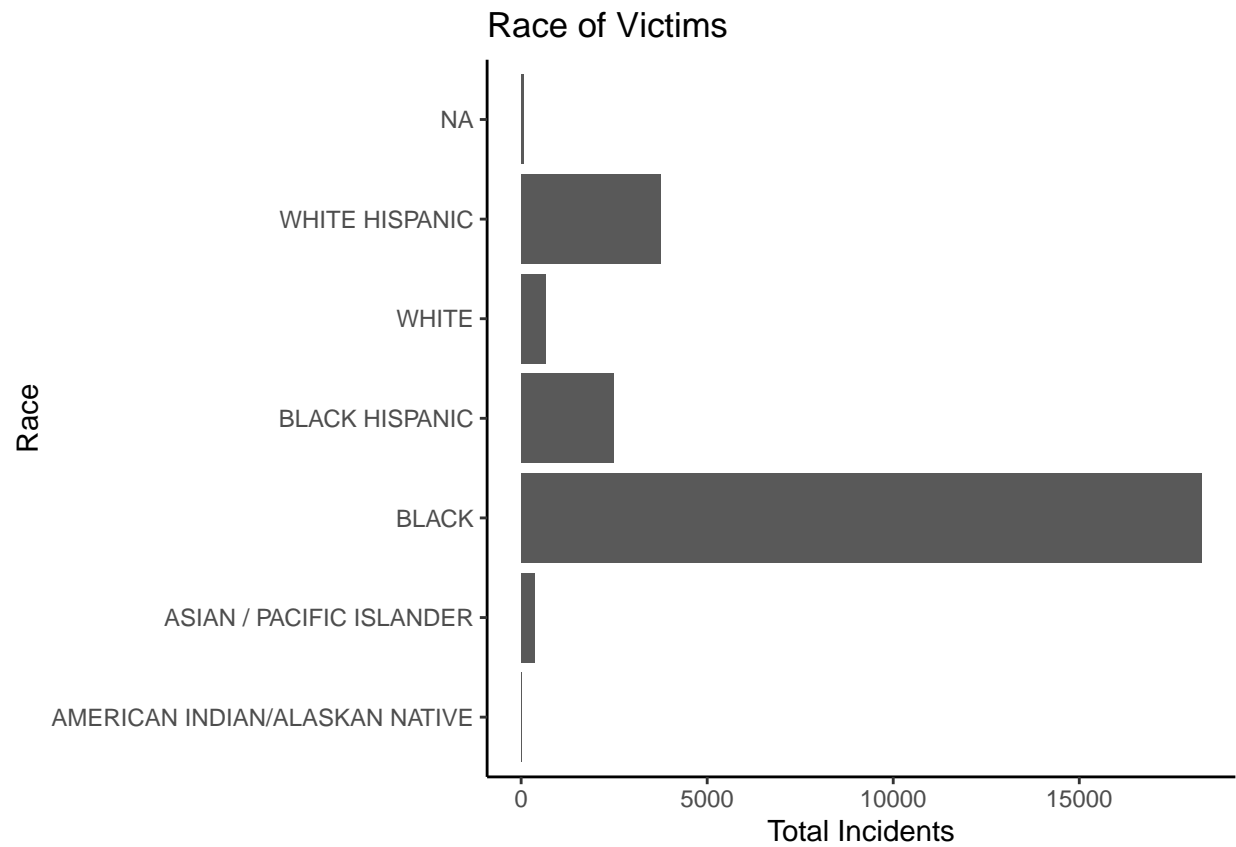

```
g <- ggplot(NYPDdata_cleaned, aes(y = PERP_RACE)) +
  geom_bar() +
  labs(title = "Race of Perpetrators",
       y = "Race",
       x = "Total Incidents") +
  theme_classic()
g
```

Race of Perpetrators

NA	
WHITE HISPANIC	
WHITE	
BLACK HISPANIC	
BLACK	
ASIAN / PACIFIC ISLANDER	
AMERICAN INDIAN/ALASKAN NATIVE	

Race

Total Incidents
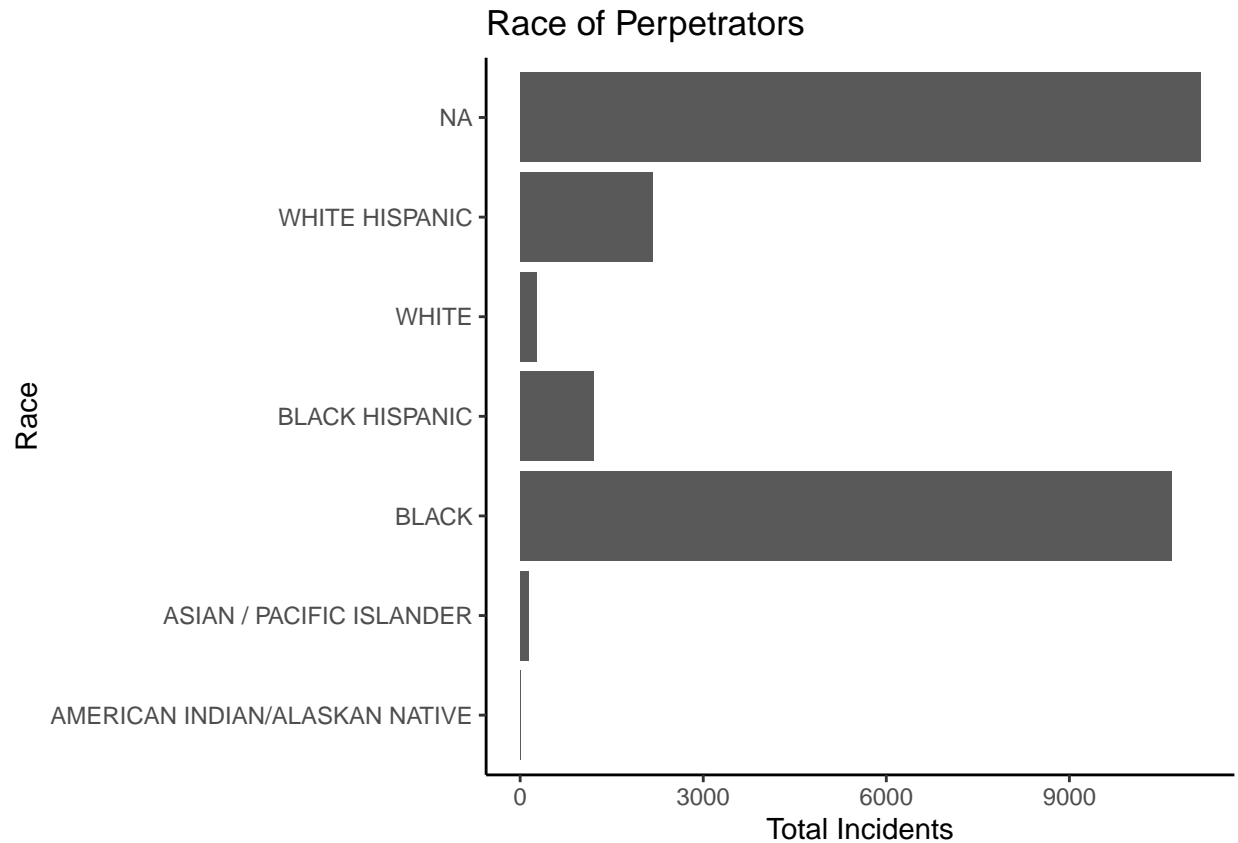
```
model <- glm(as.factor(STATISTICAL_MURDER_FLAG) ~ as.factor(PERP_RACE) + as.factor(VIC_RACE), data = NYI
summary(model)
```

```
##
## Call:
## glm(formula = as.factor(STATISTICAL_MURDER_FLAG) ~ as.factor(PERP_RACE) +
##     as.factor(VIC_RACE), family = binomial, data = NYPDdata_cleaned)
##
## Deviance Residuals:
##    Min      1Q   Median      3Q     Max
## -1.0673  -0.6799  -0.6799  -0.6327   1.8925
##
## Coefficients:
##                                           Estimate Std. Error z value
## (Intercept)                                 -23.79     281.24  -0.085
## as.factor(PERP_RACE)ASIAN / PACIFIC ISLANDER  11.67     229.63   0.051
## as.factor(PERP_RACE)BLACK                     11.22     229.63   0.049
## as.factor(PERP_RACE)BLACK HISPANIC            11.12     229.63   0.048
## as.factor(PERP_RACE)WHITE                     12.03     229.63   0.052
## as.factor(PERP_RACE)WHITE HISPANIC            11.38     229.63   0.050
## as.factor(VIC_RACE)ASIAN / PACIFIC ISLANDER   11.49     162.37   0.071
## as.factor(VIC_RACE)BLACK                      11.22     162.37   0.069
## as.factor(VIC_RACE)BLACK HISPANIC             11.06     162.37   0.068
## as.factor(VIC_RACE)WHITE                      11.39     162.37   0.070
## as.factor(VIC_RACE)WHITE HISPANIC             11.34     162.37   0.070
```

7

```
##                                    Pr(>|z|)
## (Intercept)                          0.933
## as.factor(PERP_RACE)ASIAN / PACIFIC ISLANDER   0.959
## as.factor(PERP_RACE)BLACK            0.961
## as.factor(PERP_RACE)BLACK HISPANIC   0.961
## as.factor(PERP_RACE)WHITE            0.958
## as.factor(PERP_RACE)WHITE HISPANIC   0.960
## as.factor(VIC_RACE)ASIAN / PACIFIC ISLANDER    0.944
## as.factor(VIC_RACE)BLACK             0.945
## as.factor(VIC_RACE)BLACK HISPANIC    0.946
## as.factor(VIC_RACE)WHITE             0.944
## as.factor(VIC_RACE)WHITE HISPANIC    0.944
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15041  on 14408  degrees of freedom
## Residual deviance: 14952  on 14398  degrees of freedom
##   (11187 observations deleted due to missingness)
## AIC: 14974
##
## Number of Fisher Scoring iterations: 11
```

Conclusion

- Looking at this data, we can see that the race of the victims tends to be Black, followed by White/Hispanic. The Race of the perpetrators reported is shown to be black with a good majority also being NA.

Investigation 3

What time of day are shootings occuring?

```
# Need to be able to first clean the data, find what day of the month an incident occurs, then convert
NYPDdata_cleaned$OCCUR_DAY <- mdy(NYPDdata_cleaned$OCCUR_DATE) #Day of the month
NYPDdata_cleaned$OCCUR_DAY <- wday(NYPDdata_cleaned$OCCUR_DAY, label = TRUE) #Day of the week

NYPDdata_cleaned$OCCUR_TIME <- hour(hms(as.character(NYPDdata_cleaned$OCCUR_TIME))) #Hour of a day

summary(NYPDdata_cleaned)
```

```
##   OCCUR_DATE          OCCUR_TIME        BORO              PRECINCT
## Length:25596       Min.   : 0.00   Length:25596       Min.   :  1.00
## Class :character   1st Qu.: 3.00   Class :character   1st Qu.: 44.00
## Mode  :character   Median :15.00   Mode  :character   Median : 69.00
##                    Mean   :12.19                      Mean   : 65.87
##                    3rd Qu.:20.00                      3rd Qu.: 81.00
##                    Max.   :23.00                      Max.   :123.00
##
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:25596       Length:25596            Length:25596
```

```
##   Class :character   Class :character        Class :character
##   Mode  :character   Mode  :character        Mode  :character
##
##
##
##
##      PERP_SEX           PERP_RACE          VIC_AGE_GROUP          VIC_SEX
##   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_RACE           OCCUR_DAY
##   Length:25596       Sun:5156
##   Class :character   Mon:3597
##   Mode  :character   Tue:2945
##                      Wed:2818
##                      Thu:2809
##                      Fri:3384
##                      Sat:4887
```

```r
NYPD_time_of_day <- NYPDdata_cleaned %>%
  group_by(OCCUR_TIME) %>%
  count()

NYPD_day_of_week <- NYPDdata_cleaned %>%
  group_by(OCCUR_DAY) %>%
  count()


g <- ggplot(NYPD_time_of_day, aes(x = OCCUR_TIME, y = n)) +
  geom_col() +
  labs(title = "Incidents During Different Times of the Day",
       x = "Time of Day",
       y = "Number of Incidents") +
  theme_classic()
g
```
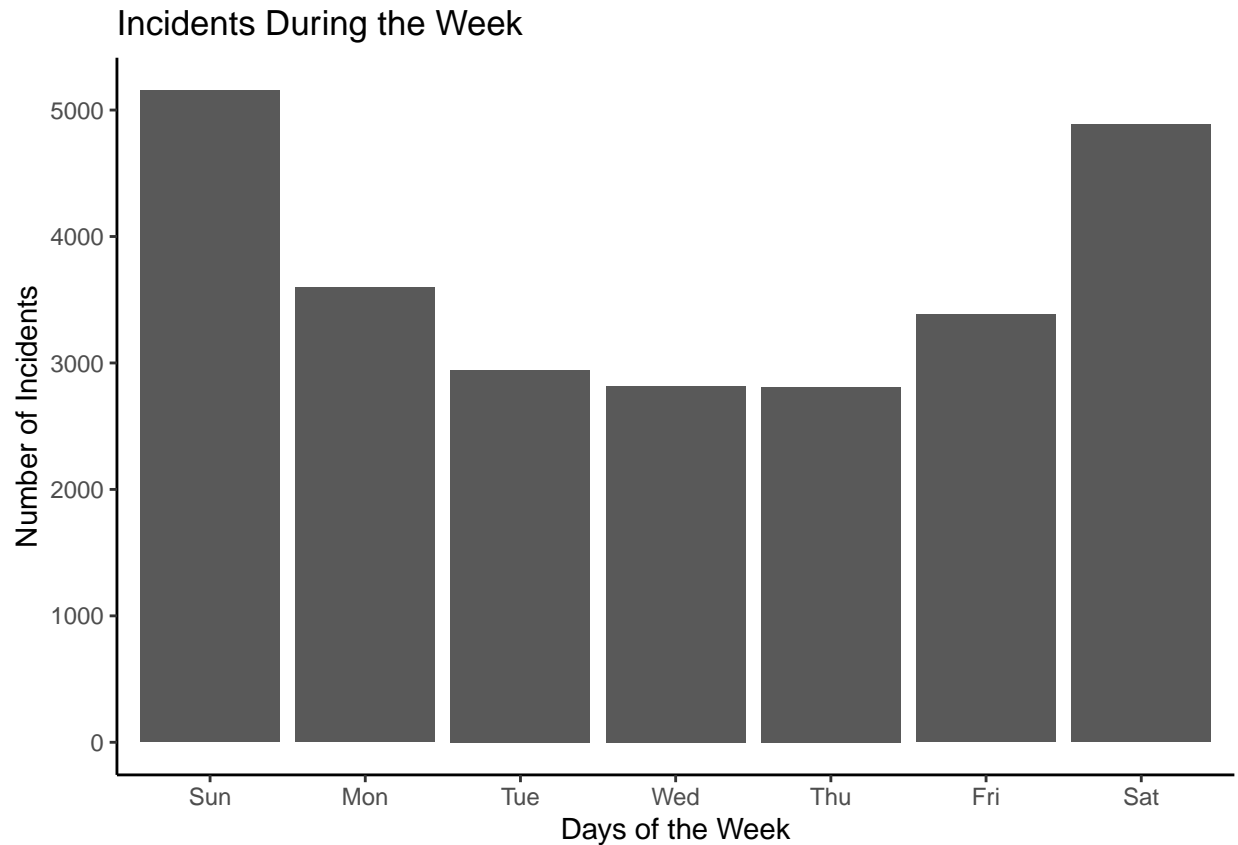
Incidents During Different Times of the Day
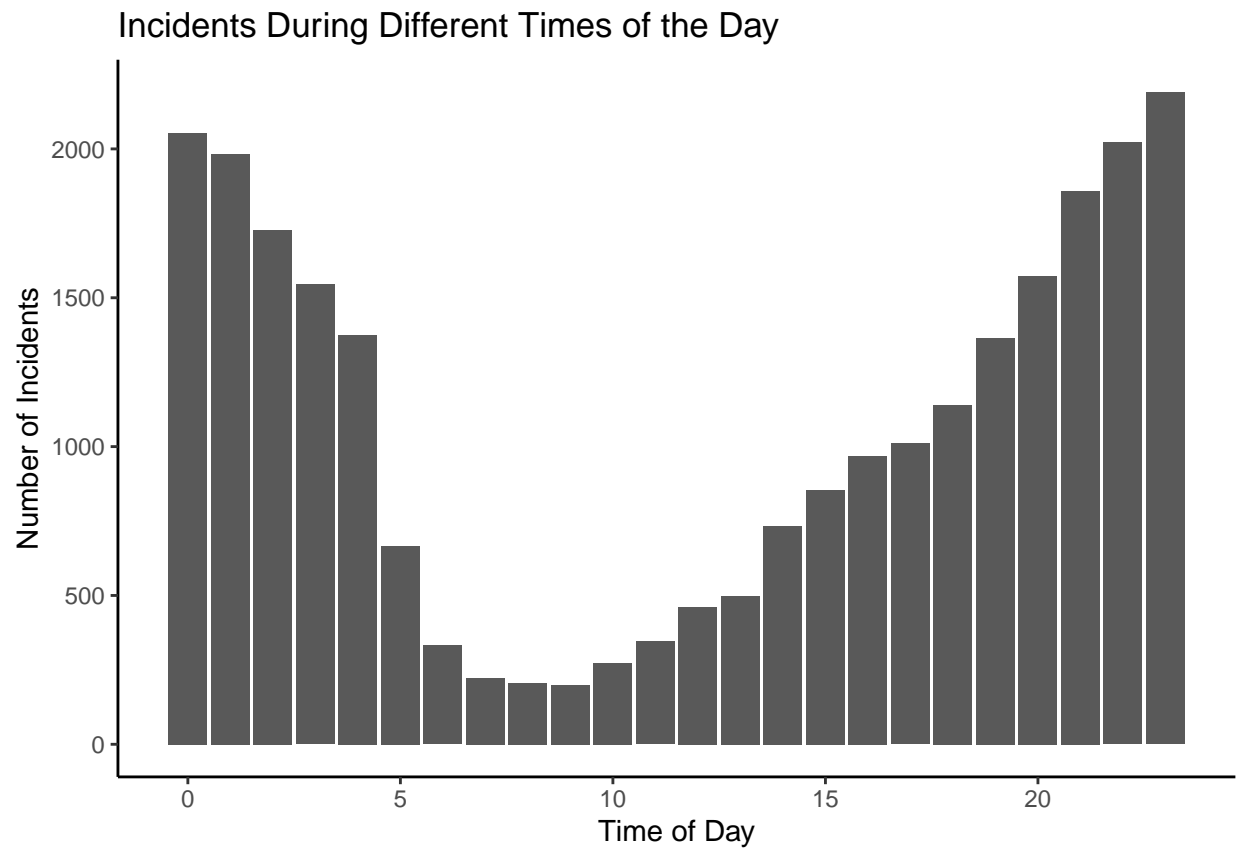

```
g <- ggplot(NYPD_day_of_week, aes(x = OCCUR_DAY, y = n)) +
  geom_col() +
  labs(title = "Incidents During the Week",
       x = "Days of the Week",
       y = "Number of Incidents") +
  theme_classic()

g
```

Incidents During the Week

Conclusion

- Based on the time of day, we can see that it is more likely for an incident to occur late at night, being especially high 6pm and 4am. We can also see that during the week, Saturday and Sunday are the prominent days for these incidents to occur. What is somewhat surprising is seeing that Monday is higher than Friday with the reportings.

Bias

Within this topic, implicit bias and discrimination can occur. I personally have never been to New York City or really know any of the borough's all that well except for what I see on tv. I do know that me being a minority, I have a certain perspective of the world based on public media, having seen what has occured to various minority groups throughout history especially when dealing with police. There are also other forms of bias in that the data set I have used has some missing areas that are considered Unknown or NA which can skew the data in various ways. Making sure to clean up the data and remove any lines that are missing data is vital to my analysis. Having the ability to take this data and reveal what truly happens in the world is far more valuable then opinions not being driven by facts.