NYPD Shooting Incident Data Report

2023-11-18

Dataset Description

List of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year.

This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. This data is manually extracted every quarter and reviewed by the Office of Management Analysis and Planning before being posted on the NYPD website. Each record represents a shooting incident in NYC and includes information about the event, the location and time of occurrence. In addition, information related to suspect and victim demographics is also included. This data can be used by the public to explore the nature of shooting/criminal activity.

(Please refer to https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic for additional information about this dataset.)

Step 0: Import Packages

```
library(tidyverse)
library(scales)
```

Step 1: Import the Data

- Copy the link address of the csv file and read in the data.

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

```
head(data)
```

```
## # 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>
```

Step 2: Tidy and Transform the Data

1. Remove the columns not needed

- The high-NA-ratio columns: **LOC_OF_OCCUR_DESC** (94%), **LOC_CLASSFCTN_DESC** (94%) and **LOCATION_DESC** (55%).
- The not important columns: **JURISDICTION_CODE**, **X_COORD_CD**, **Y_COORD_CD**, **Latitude**, **Longitude**, and **Lon_Lat**.

```
# Get the percentage of NA of each column
na_count = colSums(is.na(data))
data_count = nrow(data)
na_ratio = percent(na_count/data_count)
```

```
##            INCIDENT_KEY               OCCUR_DATE               OCCUR_TIME
##                "0.0000%"                "0.0000%"                "0.0000%"
##                     BORO       LOC_OF_OCCUR_DESC                 PRECINCT
##                "0.0000%"                "93.7170%"               "0.0000%"
##       JURISDICTION_CODE      LOC_CLASSFCTN_DESC           LOCATION_DESC
##                "0.0073%"                "93.7170%"               "54.8367%"
## STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP                 PERP_SEX
##                "0.0000%"                "34.2121%"               "34.0876%"
##                PERP_RACE           VIC_AGE_GROUP                  VIC_SEX
##               "34.0876%"                "0.0000%"               "0.0000%"
##                 VIC_RACE              X_COORD_CD              Y_COORD_CD
##                "0.0000%"                "0.0000%"               "0.0000%"
##                 Latitude                Longitude                   Lon_Lat
##                "0.0366%"                "0.0366%"               "0.0366%"
```

```
# Remove the high-NA-ratio and the not important columns
data_tidy = data %>%
  select(-c(LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION_DESC, JURISDICTION_CODE,
            X_COORD_CD, Y_COORD_CD, Latitude, Longitude, Lon_Lat))
```

2. Handle missing data for important columns

- Replace NA values with "UNKNOWN" in these columns: **PERP_AGE_GROUP** (34%), **PERP_SEX** (34%) and **PERP_RACE** (34%).

```
data_tidy = data_tidy %>%
  replace_na(list(PERP_AGE_GROUP = "UNKNOWN", PERP_SEX = "UNKNOWN", PERP_RACE = "UNKNOWN"))
```

3. Decide labels of factor

- Print the labels of factor. (Apply `table()` to each column.)
- Decide the labels of factor.

For column **PRECINCT**:

```r
# Print the labels of column PRECINCT
table(data_tidy$PRECINCT)

# Most of the labels are distinct
# Remove column PRECINCT
data_tidy = data_tidy %>%
  select(-PRECINCT)
```

For column **PERP_AGE_GROUP**:

```r
# Print the labels of column PERP_AGE_GROUP
table(data_tidy$PERP_AGE_GROUP)

# Remove typos and rename "(null)" to "UNKNOWN"
data_tidy = data_tidy %>% filter(PERP_AGE_GROUP != "1020" & PERP_AGE_GROUP != "224" &
                                 PERP_AGE_GROUP != "940")
data_tidy$PERP_AGE_GROUP = recode(data_tidy$PERP_AGE_GROUP, "(null)" = "UNKNOWN")
```

For column **PERP_SEX**:

```r
# Print the labels of column PERP_SEX
table(data_tidy$PERP_SEX)

# Rename "U" and "(null)" to "UNKNOWN"
data_tidy$PERP_SEX = recode(data_tidy$PERP_SEX, "U" = "UNKNOWN", "(null)" = "UNKNOWN")
```

For column **PERP_RACE**:

```r
# Print the labels of column PERP_RACE
table(data_tidy$PERP_RACE)

# Rename "(null)" to "UNKNOWN"
data_tidy$PERP_RACE = recode(data_tidy$PERP_RACE, "(null)" = "UNKNOWN")
```

For column **VIC_AGE_GROUP**:

```r
# Print the labels of column VIC_AGE_GROUP
table(data_tidy$VIC_AGE_GROUP)

# Remove the typo.
data_tidy = data_tidy %>% filter(VIC_AGE_GROUP != "1022")
```

For column **VIC_SEX**:

```r
# Print the labels of column VIC_SEX
table(data_tidy$VIC_SEX)

# Rename "U" to "UNKNOWN"
data_tidy$VIC_SEX = recode(data_tidy$VIC_SEX, "U" = "UNKNOWN")
```

For column **VIC_RACE**:

```
# Print the labels of column VIC_RACE
table(data_tidy$VIC_RACE)

# Don't need to change
```

4. Factoring the dataframe

- Apply `as.factor()` to each column.

```
# For column INCIDENT_KEY, apply `as.character()` instead of `as.factor()`
data_tidy$INCIDENT_KEY = as.character(data_tidy$INCIDENT_KEY)
data_tidy$BORO = as.factor(data_tidy$BORO)
data_tidy$PERP_AGE_GROUP = as.factor(data_tidy$PERP_AGE_GROUP)
data_tidy$PERP_SEX = as.factor(data_tidy$PERP_SEX)
data_tidy$PERP_RACE = as.factor(data_tidy$PERP_RACE)
data_tidy$VIC_AGE_GROUP = as.factor(data_tidy$VIC_AGE_GROUP)
data_tidy$VIC_SEX = as.factor(data_tidy$VIC_SEX)
data_tidy$VIC_RACE = as.factor(data_tidy$VIC_RACE)
```

```
summary(data_tidy)
```

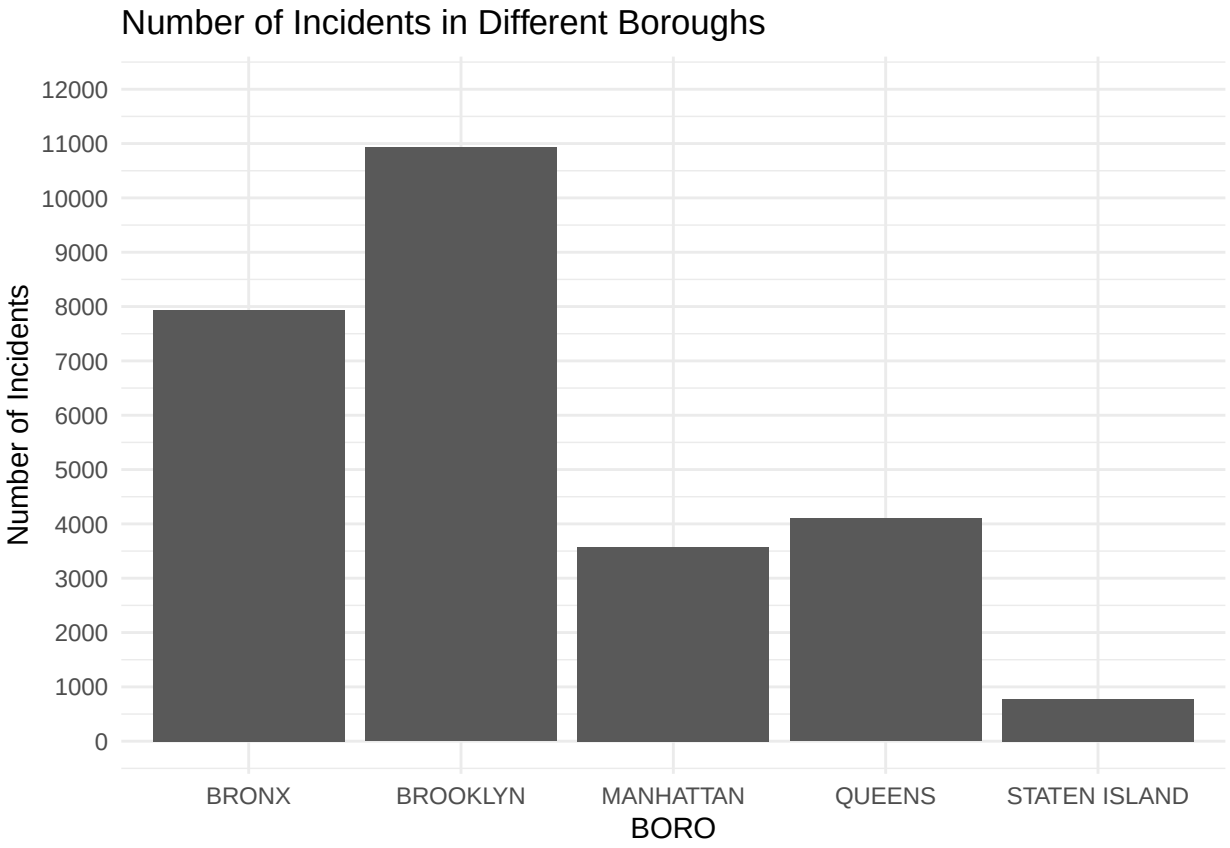
```
##  INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME                   BORO
##  Length:27308       Length:27308       Length:27308      BRONX        : 7935
##  Class :character   Class :character   Class1:hms        BROOKLYN     :10932
##  Mode  :character   Mode  :character   Class2:difftime   MANHATTAN    : 3571
##                                        Mode  :numeric    QUEENS       : 4094
##                                                          STATEN ISLAND:  776
##
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX
##  Mode :logical           <18    : 1591   F       :  424
##  FALSE:22042             18-24  : 6221   M       :15435
##  TRUE :5266              25-44  : 5687   UNKNOWN:11449
##                          45-64  :  617
##                          65+    :   60
##                          UNKNOWN:13132
##
##                            PERP_RACE     VIC_AGE_GROUP      VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    : 2839   F       : 2615
##  ASIAN / PACIFIC ISLANDER      :  154   18-24  :10085   M       :24682
##  BLACK                         :11430   25-44  :12279   UNKNOWN:   11
##  BLACK HISPANIC                : 1314   45-64  : 1863
##  UNKNOWN                       :11786   65+    :  181
##  WHITE                         :  283   UNKNOWN:   61
##  WHITE HISPANIC                : 2339
##                              VIC_RACE
##  AMERICAN INDIAN/ALASKAN NATIVE:   10
##  ASIAN / PACIFIC ISLANDER      :  404
##  BLACK                         :19437
##  BLACK HISPANIC                : 2646
##  UNKNOWN                       :   66
##  WHITE                         :  698
##  WHITE HISPANIC                : 4047
```

Step 3: Add Visualizations and Analysis

Question 1: How is the distribution of incidents across different boroughs?

- The borough with the highest number of incidents is **BROOKLYN**, followed by **BRONX** and **QUEENS**.

```
# Create a bar chart
ggplot_1 = ggplot(data_tidy, aes(x = BORO)) +
  geom_bar() +
  labs(title = "Number of Incidents in Different Boroughs", x = "BORO", y = "Number of Incidents") +
  theme_minimal() +
  scale_y_continuous(limits = c(0, 12000), breaks = seq(0, 12000, by = 1000))
ggplot_1
```



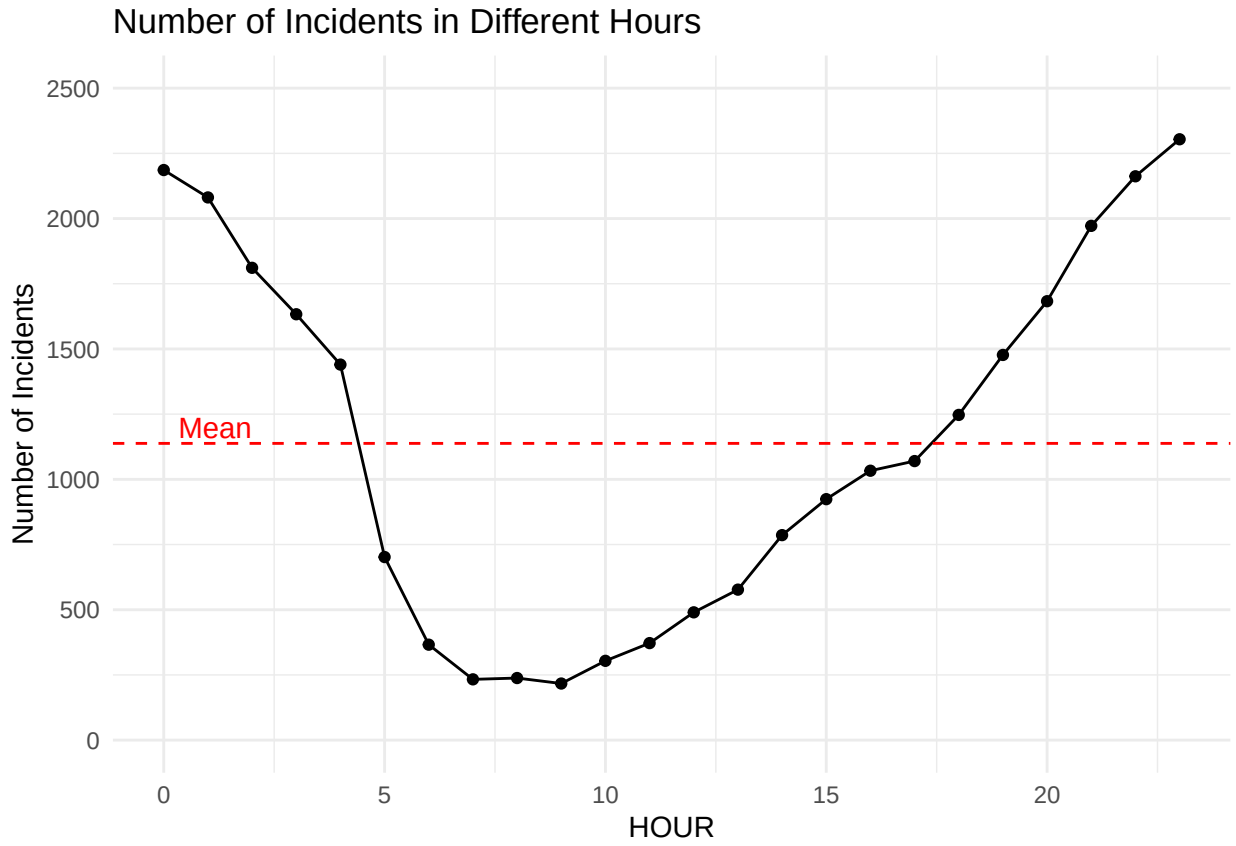
Question 2: How is the distribution of incidents across different hours?

- The number of incidents is below average after **5:00** and above average after **17:00**.
- The most incidents occurred between **23:00** and **00:00**.

```
# Get the Hour part from OCCUR_TIME
data_tidy$OCCUR_HOUR = hour(data_tidy$OCCUR_TIME)

# Create a line chart
```

```r
ggplot_2 = ggplot(data_tidy, aes(x = OCCUR_HOUR)) +
  geom_point(stat = "count") +
  geom_line(stat = "count") +
  labs(title = "Number of Incidents in Different Hours", x = "HOUR", y = "Number of Incidents") +
  theme_minimal() +
  scale_y_continuous(limits = c(0, 2500), breaks = seq(0, 3000, by = 500)) +
  geom_hline(yintercept = mean(table(data_tidy$OCCUR_HOUR)), color = "red", linetype = "dashed") +
  annotate("text", x = 0, y = 1200, label = "Mean", hjust = -0.2, color = "red")
ggplot_2
```



Question 3: How is the distribution of incidents involving different races?

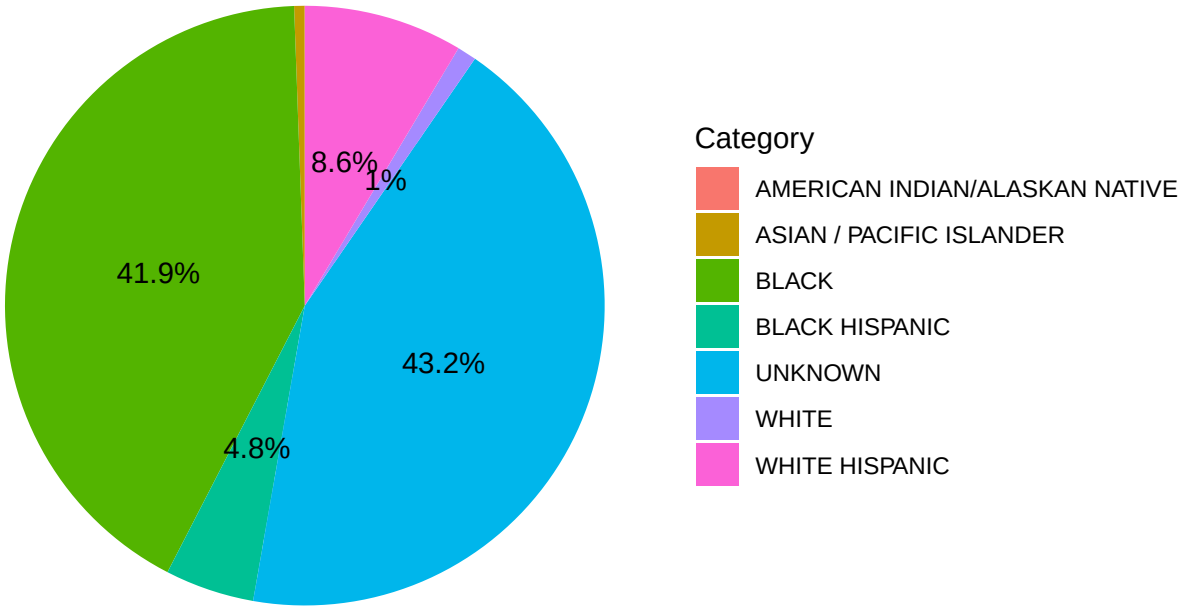
- PERP_RACE: excluding UNKNOWN, the largest proportion is **BLACK**, followed by **WHITE HISPANIC** and **BLACK HISPANIC**.
- VIC_RACE: same as PERP_RACE, the largest proportion is **BLACK**, followed by **WHITE HISPANIC** and **BLACK HISPANIC**.

```r
# Convert data table to data frame
data_PERP_RACE = as.data.frame(table(data_tidy$PERP_RACE))
colnames(data_PERP_RACE) = c("Category", "Count")

# Calculate percentage
data_PERP_RACE$Percentage = (data_PERP_RACE$Count / sum(data_PERP_RACE$Count)) * 100
```

```r
# Create a pie chart
ggplot_3 = ggplot(data_PERP_RACE, aes(x = 1, y = Percentage, fill = Category)) +
  geom_bar(stat = "identity") +
  coord_polar(theta = "y") +
  theme_void() +
  labs(title = "PERP_RACE") +
  geom_text(aes(label = ifelse(Percentage > 1, paste0(round(Percentage, 1), "%"), "")),
            position = position_stack(vjust = 0.5))
  ggplot_3
```

PERP_RACE

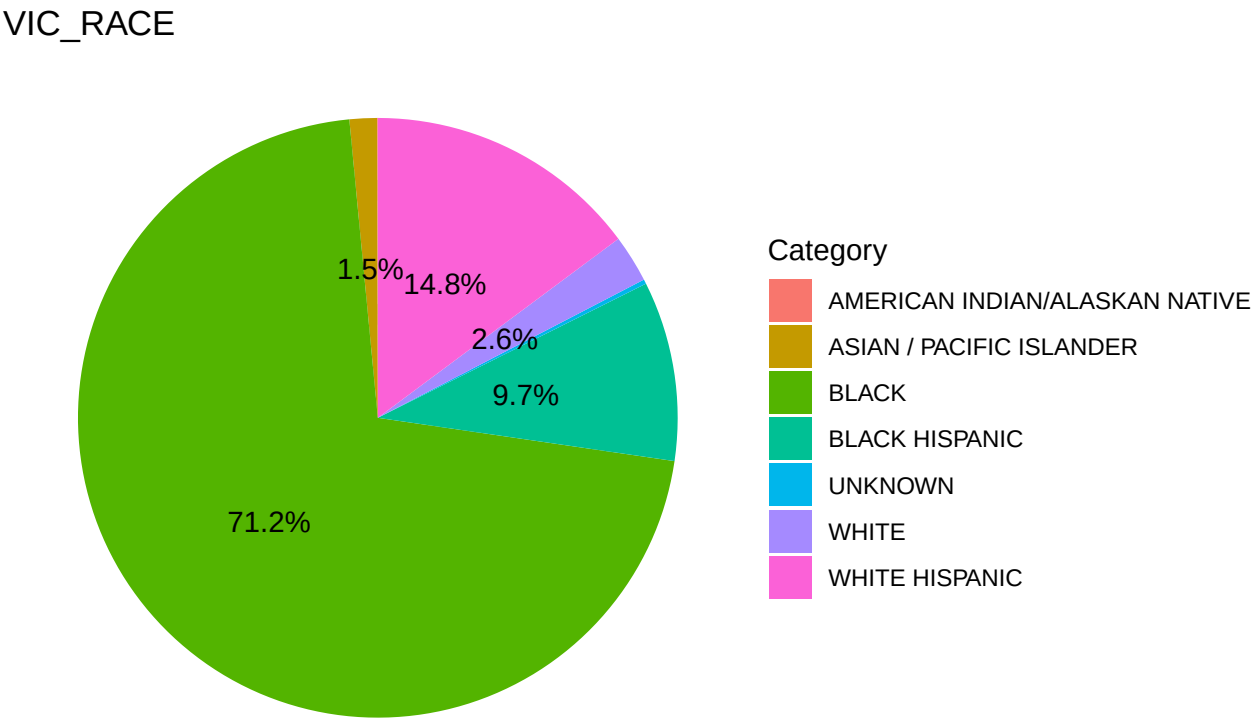



```r
# Convert data table to data frame
data_VIC_RACE = as.data.frame(table(data_tidy$VIC_RACE))
colnames(data_VIC_RACE) = c("Category", "Count")

# Calculate percentage
data_VIC_RACE$Percentage = (data_VIC_RACE$Count / sum(data_VIC_RACE$Count)) * 100

# Create a pie chart
ggplot_4 = ggplot(data_VIC_RACE, aes(x = 1, y = Percentage, fill = Category)) +
  geom_bar(stat = "identity") +
  coord_polar(theta = "y") +
  theme_void() +
  labs(title = "VIC_RACE") +
  geom_text(aes(label = ifelse(Percentage > 1, paste0(round(Percentage, 1), "%"), "")),
            position = position_stack(vjust = 0.5))
```

```
ggplot_4
```



Question 4: What is the correlation between STATISTICAL_MURDER_FLAG and other features?

- **Purpose**: Predict the probability of a murder case based on the following variables (OCCUR_DAY, OCCUR_HOUR, BORO, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, and VIC_RACE).
- **Methods**: Use a logistic regression model to evaluate feature importance.
  - Use `glm()` to build a logistic regression model.
  - Use `summary()` to view the coefficients of each feature.
- **Results**:
  - Statistical significance: **BOROMAHATTAN**, **PERP_AGE_GROUP** (25-44, 45-64, 65+, UNKNOWN), **PERP_SEXUNKNOWN**, and **VIC_AGE_GROUP** (18-24, 25-44, 45-64, 65+) have a statistically significant impact on predicting **STATISTICAL_MURDER_FLAG**.
  - Positive correlation: **PERP_AGE_GROUP** (25-44, 45-64, 65+), and **VIC_AGE_GROUP** (18-24, 25-44, 45-64, 65+).
  - Negative correlation: **BOROMANHATTAN**, **PERP_AGE_GROUPUNKNOWN**, and **PERP_SEXUNKNOWN**.

```r
#Convert date to day of week
data_tidy$OCCUR_DAY = wday(mdy(data_tidy$OCCUR_DATE), label = TRUE)

#Logistic regression model
model = glm(STATISTICAL_MURDER_FLAG ~ OCCUR_DAY + OCCUR_HOUR + BORO + PERP_AGE_GROUP +
            PERP_SEX + PERP_RACE + VIC_AGE_GROUP + VIC_SEX + VIC_RACE, data = data_tidy,
            family = "binomial")

summary(model)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ OCCUR_DAY + OCCUR_HOUR +
##     BORO + PERP_AGE_GROUP + PERP_SEX + PERP_RACE + VIC_AGE_GROUP +
##     VIC_SEX + VIC_RACE, family = "binomial", data = data_tidy)
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -23.940152 249.166868  -0.096  0.92346
## OCCUR_DAY.L                       -0.045999   0.038639  -1.190  0.23386
## OCCUR_DAY.Q                       -0.062112   0.041485  -1.497  0.13434
## OCCUR_DAY.C                       -0.059489   0.041714  -1.426  0.15383
## OCCUR_DAY^4                       -0.012525   0.042499  -0.295  0.76822
## OCCUR_DAY^5                        0.011236   0.044581   0.252  0.80101
## OCCUR_DAY^6                       -0.081435   0.045878  -1.775  0.07589 .
## OCCUR_HOUR                        -0.001721   0.001928  -0.893  0.37207
## BOROBROOKLYN                       0.030529   0.039637   0.770  0.44117
## BOROMANHATTAN                     -0.153222   0.053409  -2.869  0.00412 **
## BOROQUEENS                        -0.014103   0.050445  -0.280  0.77981
## BOROSTATEN ISLAND                 -0.099064   0.095914  -1.033  0.30168
## PERP_AGE_GROUP18-24                0.102130   0.073430   1.391  0.16427
## PERP_AGE_GROUP25-44                0.331319   0.074021   4.476 7.60e-06 ***
## PERP_AGE_GROUP45-64                0.620786   0.110223   5.632 1.78e-08 ***
## PERP_AGE_GROUP65+                  0.692069   0.280163   2.470  0.01350 *
## PERP_AGE_GROUPUNKNOWN             -2.337027   0.171152 -13.655  < 2e-16 ***
## PERP_SEXM                         -0.126735   0.114163  -1.110  0.26695
## PERP_SEXUNKNOWN                    2.477610   0.266053   9.312  < 2e-16 ***
## PERP_RACEASIAN / PACIFIC ISLANDER 12.003532 229.627337   0.052  0.95831
## PERP_RACEBLACK                    11.633739 229.627264   0.051  0.95959
## PERP_RACEBLACK HISPANIC           11.569950 229.627275   0.050  0.95981
## PERP_RACEUNKNOWN                  11.157062 229.627358   0.049  0.96125
## PERP_RACEWHITE                    12.167884 229.627305   0.053  0.95774
## PERP_RACEWHITE HISPANIC           11.775464 229.627269   0.051  0.95910
## VIC_AGE_GROUP18-24                 0.267503   0.063198   4.233 2.31e-05 ***
## VIC_AGE_GROUP25-44                 0.531597   0.062113   8.559  < 2e-16 ***
## VIC_AGE_GROUP45-64                 0.624194   0.080471   7.757 8.72e-15 ***
## VIC_AGE_GROUP65+                   0.884785   0.177171   4.994 5.92e-07 ***
## VIC_AGE_GROUPUNKNOWN               0.612678   0.320486   1.912  0.05591 .
## VIC_SEXM                           0.035787   0.053149   0.673  0.50073
## VIC_SEXUNKNOWN                    -0.382542   1.083860  -0.353  0.72413
## VIC_RACEASIAN / PACIFIC ISLANDER  10.781783  96.723592   0.111  0.91124
## VIC_RACEBLACK                     10.620272  96.723516   0.110  0.91257
## VIC_RACEBLACK HISPANIC            10.413822  96.723529   0.108  0.91426
```

```
## VIC_RACEUNKNOWN                     9.729084  96.724435   0.101  0.91988
## VIC_RACEWHITE                      10.691342  96.723565   0.111  0.91199
## VIC_RACEWHITE HISPANIC             10.683046  96.723523   0.110  0.91205
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26779  on 27307  degrees of freedom
## Residual deviance: 25667  on 27270  degrees of freedom
## AIC: 25743
##
## Number of Fisher Scoring iterations: 11
```

Step 4: Add Bias Identification

1. Personal bias

- **Analyst's subjective bias**: Before looking at the data, I might have thought that there were more female than male victims, but in fact the data shows that both the perpetrators and the victims are more male than female. Beyond that, during the step of tidying data, I treated longitude and latitude as unimportant data and removed them. Maybe they are important data but I didn't analyze them carefully. Ways to mitigate this bias include staying as objective as possible, avoiding personal interpretations of the data, and considering multiple explanations.

2. Analysis bias

- **Selectivity bias**: Selective selection or reporting of a specific subset of data to support a specific conclusion. Ways to mitigate this bias include openly explaining the selection of data subsets and providing complete analyses.
- **Statistical analysis bias**: Incorrect statistical methods can lead to bias. Ways to mitigate this bias include ensuring you use correct statistical methods and interpret statistical results appropriately.

Conclusion

Overall, factors such as BOROMAHATTAN, PERP_AGE_GROUP (25-44, 45-64, 65+, UNKNOWN), VIC_AGE_GROUP (18-24, 25-44, 45-64, 65+), and PERP_SEXUNKNOWN have statistically significant effects on predicting whether a incident is a murder case. This report analyzes NYPD Shooting Incident Data from 2006 to the recent past, and provides some data visualizations and brief analysis as a reference for future researchers.