NYDP Shooting Project

Maher Ali Rusho

2023-01-20

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

#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 NYPD Shooting Incident Data (Historic) - CKAN for additional information about this dataset. ## Step 0: Import Library

```
# install.packages("tidyverse")
library(tidyverse)
library(lubridate)
```

Step 1: Load Data

#`read_csv()` reads comma delimited files, read_csv2() reads semicolon separated files (common in countries where , is used as the decimal place), read_tsv() reads tab delimited files, and read_delim() reads in files with any delimiter.

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

```
## Rows: 25596 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## 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(df)
```

```
## # A tibble: 6 x 19
```

```
##   INCIDE~1 OCCUR~2 OCCUR~3 BORO  PRECI~4 JURIS~5 LOCAT~6 STATI~7 PERP_~8 PERP_~9
##      <dbl> <chr>   <time>  <chr>   <dbl>   <dbl> <chr>   <lgl>   <chr>   <chr>
## 1   2.36e8 11/11/~ 15:04   BROO~      79       0 <NA>    FALSE   <NA>    <NA>
## 2   2.31e8 07/16/~ 22:05   BROO~      72       0 <NA>    FALSE   45-64   M
## 3   2.31e8 07/11/~ 01:09   BROO~      79       0 <NA>    FALSE   <18     M
## 4   2.38e8 12/11/~ 13:42   BROO~      81       0 <NA>    FALSE   <NA>    <NA>
## 5   2.24e8 02/16/~ 20:00   QUEE~     113       0 <NA>    FALSE   <NA>    <NA>
## 6   2.28e8 05/15/~ 04:13   QUEE~     113       0 <NA>    TRUE    <NA>    <NA>
## # ... with 9 more variables: 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>, and abbreviated variable
## #   names 1: INCIDENT_KEY, 2: OCCUR_DATE, 3: OCCUR_TIME, 4: PRECINCT,
## #   5: JURISDICTION_CODE, 6: LOCATION_DESC, 7: STATISTICAL_MURDER_FLAG,
## #   8: PERP_AGE_GROUP, 9: PERP_SEX
```

Step 2: Tidy and Transform Data

Let's first eliminate the columns I do not need for this assignment, which are: **PRECINCT**,**JURISDICTION_CODE**,**LO**
X_COORD_CD, **Y_COORD_CD**, and **Lon_Lat**.

```
df_2 = df %>% select(INCIDENT_KEY,
                     OCCUR_DATE,
                     OCCUR_TIME,
                     BORO,
                     STATISTICAL_MURDER_FLAG,
                     PERP_AGE_GROUP,
                     PERP_SEX,
                     PERP_RACE,
                     VIC_AGE_GROUP,
                     VIC_SEX,
                     VIC_RACE,
                     Latitude,
                     Longitude)
# Return the column name along with the missing values
lapply(df_2, 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] 0
##
## $Longitude
## [1] 0
```

Understanding the reasons why data are missing is important for handling the remaining data correctly. There's a fair amount of unidentifiable data on perpetrators (age, race, or sex.) Those cases are possibly still active and ongoing investigation. In fear of missing meaningful information, I handle this group of missing data by calling them as another group of "Unknown".

Key observations on 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
df_2 = df_2 %>%
  replace_na(list(PERP_AGE_GROUP = "Unknown", PERP_SEX = "Unknown", PERP_RACE = "Unknown"))
# Remove extreme values in data
df_2 = subset(df_2, PERP_AGE_GROUP!="1020" & PERP_AGE_GROUP!="224" & PERP_AGE_GROUP!="940")
df_2$PERP_AGE_GROUP = recode(df_2$PERP_AGE_GROUP, UNKNOWN = "Unknown")
df_2$PERP_SEX = recode(df_2$PERP_SEX, U = "Unknown")
df_2$PERP_RACE = recode(df_2$PERP_RACE, UNKNOWN = "Unknown")
df_2$VIC_SEX   = recode(df_2$VIC_SEX, U = "Unknown")
df_2$VIC_RACE   = recode(df_2$VIC_RACE, UNKNOWN = "Unknown")
df_2$INCIDENT_KEY = as.character(df_2$INCIDENT_KEY)
df_2$BORO = as.factor(df_2$BORO)
df_2$PERP_AGE_GROUP = as.factor(df_2$PERP_AGE_GROUP)
df_2$PERP_SEX = as.factor(df_2$PERP_SEX)
df_2$PERP_RACE = as.factor(df_2$PERP_RACE)
df_2$VIC_AGE_GROUP = as.factor(df_2$VIC_AGE_GROUP)
df_2$VIC_SEX = as.factor(df_2$VIC_SEX)
```

```
df_2$VIC_RACE = as.factor(df_2$VIC_RACE)
# Return summary statistics
summary(df_2)
```

```
##  INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME                  BORO
##  Length:25593       Length:25593       Length:25593      BRONX         : 7400
##  Class :character   Class :character   Class1:hms        BROOKLYN      :10364
##  Mode  :character   Mode  :character   Class2:difftime   MANHATTAN     : 3265
##                                        Mode  :numeric    QUEENS        : 3828
##                                                          STATEN ISLAND:  736
##
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX
##  Mode :logical           <18    : 1463   F       :  371
##  FALSE:20665             18-24  : 5844   M       :14413
##  TRUE :4928              25-44  : 5202   Unknown:10809
##                          45-64  :  535
##                          65+    :   57
##                          Unknown:12492
##
##                            PERP_RACE     VIC_AGE_GROUP      VIC_SEX
##  AMERICAN INDIAN/ALASKAN NATIVE:    2   <18    : 2681   F       : 2403
##  ASIAN / PACIFIC ISLANDER      :  141   18-24  : 9603   M       :23179
##  BLACK                         :10667   25-44  :11384   Unknown:   11
##  BLACK HISPANIC                : 1203   45-64  : 1698
##  Unknown                       :11146   65+    :  167
##  WHITE                         :  272   UNKNOWN:   60
##  WHITE HISPANIC                : 2162
##                              VIC_RACE       Latitude       Longitude
##  AMERICAN INDIAN/ALASKAN NATIVE:    9   Min.   :40.51   Min.   :-74.25
##  ASIAN / PACIFIC ISLANDER      :  354   1st Qu.:40.67   1st Qu.:-73.94
##  BLACK                         :18280   Median :40.70   Median :-73.92
##  BLACK HISPANIC                : 2485   Mean   :40.74   Mean   :-73.91
##  Unknown                       :   65   3rd Qu.:40.82   3rd Qu.:-73.88
##  WHITE                         :  660   Max.   :40.91   Max.   :-73.70
##  WHITE HISPANIC                : 3740
```

Step 3: Add Visualizations and Analysis

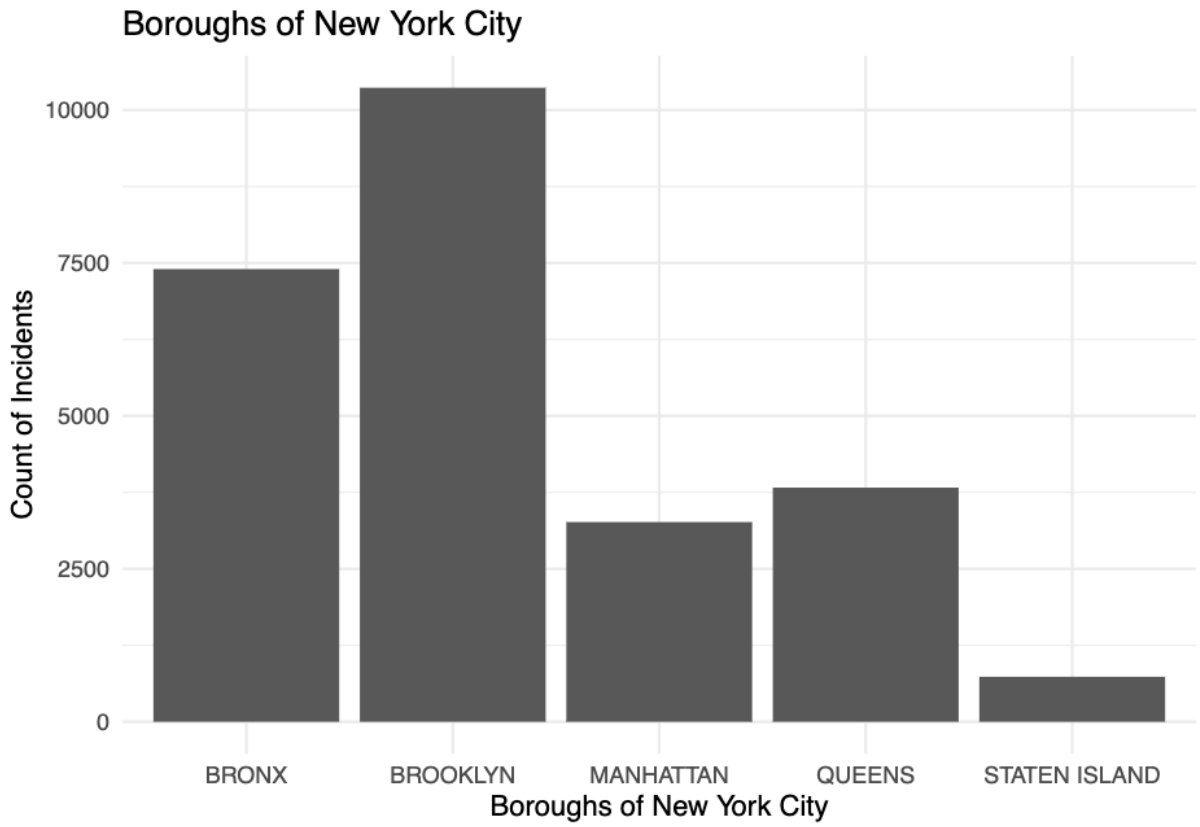
Research Question

1. Which part of New York has the most number of incidents? Of those incidents, how many are murder cases?

Brooklyn is the 1st in terms of the number of incidents, followed by Bronx and Queens respectively. Likewise, the number of murder cases follows the same pattern as that of incidents.

```
g <- ggplot(df_2, aes(x = BORO)) +
  geom_bar() +
  labs(title = "Boroughs of New York City",
       x = "Boroughs of New York City",
```

```
       y = "Count of Incidents") +
  theme_minimal()
g
```




```
table(df_2$BORO, df_2$STATISTICAL_MURDER_FLAG)
```

```
##
##                FALSE TRUE
##   BRONX         5983 1417
##   BROOKLYN      8344 2020
##   MANHATTAN     2691  574
##   QUEENS        3066  762
##   STATEN ISLAND  581  155
```

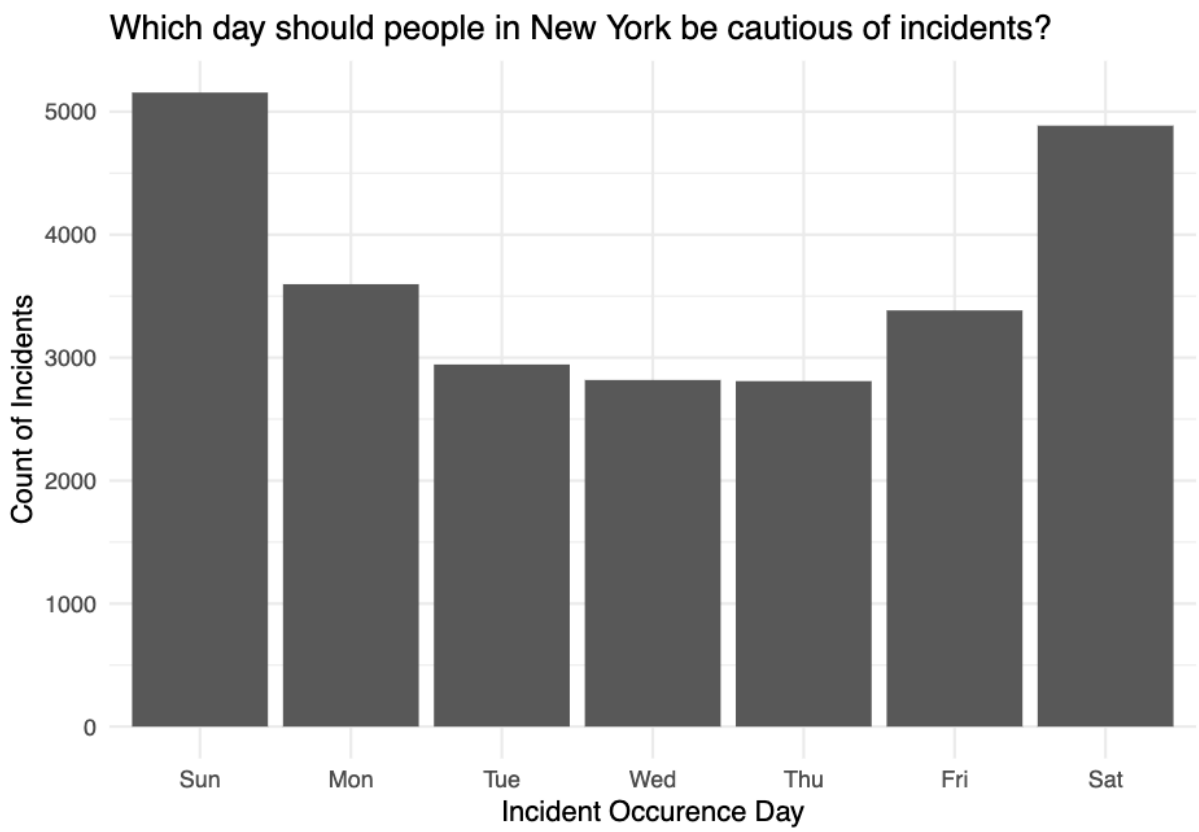
2. Which day and time should people in New York be cautious of falling into victims of crime?

- Weekends in NYC have the most chances of incidents. Be cautious!
- Incidents historically happen in the evening and night time. If there's nothing urgent, recommend people staying at home!

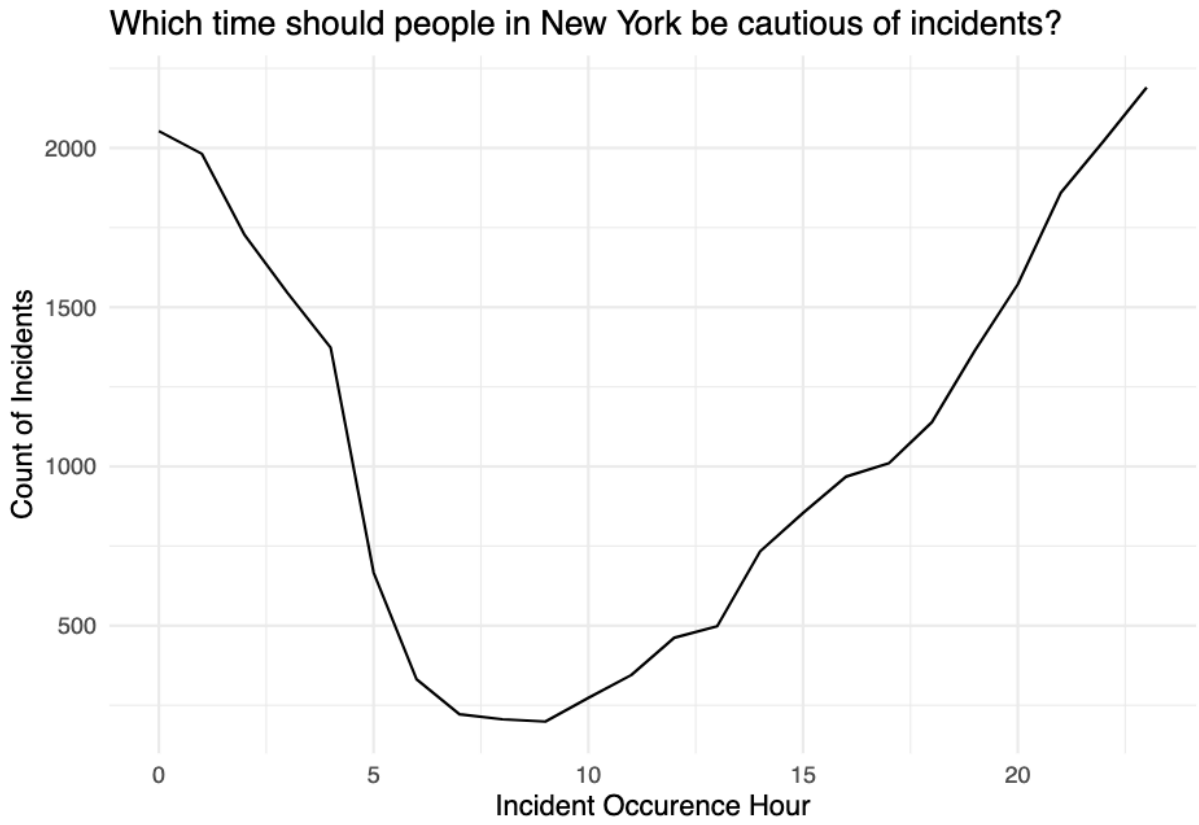
```
df_2$OCCUR_DAY = mdy(df_2$OCCUR_DATE)
df_2$OCCUR_DAY = wday(df_2$OCCUR_DAY, label = TRUE)
df_2$OCCUR_HOUR = hour(hms(as.character(df_2$OCCUR_TIME)))
```

```
df_3 = df_2 %>%
  group_by(OCCUR_DAY) %>%
  count()
df_4 = df_2 %>%
  group_by(OCCUR_HOUR) %>%
  count()
```

```
g <- ggplot(df_3, aes(x = OCCUR_DAY, y = n)) +
  geom_col() +
  labs(title = "Which day should people in New York be cautious of incidents?",
       x = "Incident Occurence Day",
       y = "Count of Incidents") +
  theme_minimal()
g
```




```
g <- ggplot(df_4, aes(x = OCCUR_HOUR, y = n)) +
  geom_line() +
  labs(title = "Which time should people in New York be cautious of incidents?",
       x = "Incident Occurence Hour",
       y = "Count of Incidents") +
  theme_minimal()
g
```



3. The Profile of Perpetrators and Victims

- There's a striking number of incidents in the age group of 25-44 and 18-24.
- Black and White Hispanic stood out in the number of incidents in Boroughs of New York City.
- There are significantly more incidents with Male than those of Female.

```
table(df_2$PERP_AGE_GROUP, df_2$VIC_AGE_GROUP)
```

```
##
##             <18 18-24 25-44 45-64  65+ UNKNOWN
##    <18      445   584   353    70    9       2
##    18-24    742  2607  2141   305   37      12
##    25-44    247  1417  3033   431   40      34
##    45-64     19    62   290   148   11       5
##    65+        0     1    23    23   10       0
##    Unknown 1228  4932  5544   721   60       7
```

```
table(df_2$PERP_SEX, df_2$VIC_SEX)
```

```
##
##                F     M Unknown
##    F          58   312       1
##    M        1540 12867       6
##    Unknown   805 10000       4
```

```
table(df_2$PERP_RACE, df_2$VIC_RACE)
```

```
##
##                                  AMERICAN INDIAN/ALASKAN NATIVE
##   AMERICAN INDIAN/ALASKAN NATIVE                              0
##   ASIAN / PACIFIC ISLANDER                                    0
##   BLACK                                                       4
##   BLACK HISPANIC                                              0
##   Unknown                                                     5
##   WHITE                                                       0
##   WHITE HISPANIC                                              0
##
##                                  ASIAN / PACIFIC ISLANDER BLACK BLACK HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE                        0     2              0
##   ASIAN / PACIFIC ISLANDER                             43    51             13
##   BLACK                                               135  8470            749
##   BLACK HISPANIC                                       17   481            320
##   Unknown                                             113  8523            999
##   WHITE                                                11    34             21
##   WHITE HISPANIC                                       35   719            383
##
##                                  Unknown WHITE WHITE HISPANIC
##   AMERICAN INDIAN/ALASKAN NATIVE       0     0              0
##   ASIAN / PACIFIC ISLANDER             0    11             23
##   BLACK                               24   183           1102
##   BLACK HISPANIC                       5    34            346
##   Unknown                             24   187           1295
##   WHITE                                1   156             49
##   WHITE HISPANIC                      11    89            925
```

4. Building logistic regression model to predict if the incident is likely a murder case or not?

Logistic regression is an instance of classification technique that you can use to predict a qualitative response. I will use logistic regression models to estimate the probability that a murder case belongs to a particular profile, location, or date & time.

The output shows the coefficients, their standard errors, the z-statistic (sometimes called a Wald z-statistic), and the associated p-values. **PERP_SEXUnknown**, **PERP_AGE_GROUP45-64**, **PERP_AGE_GROUP65+**, **PERP_AGE_GROUPUnknown**, and **PERP_AGE_GROUP25-44** are statistically significant, as are the **latitude** and **longitude**. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

- The person in the age group of 65+, versus a person whose age < 18, changes the log odds of murder by 1.03.

```
# Logistics Regression
glm.fit <- glm(STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX + PERP_AGE_GROUP + OCCUR_HOUR + OCCUR_DAY
summary(glm.fit)
```

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ PERP_RACE + PERP_SEX +
```

```
##     PERP_AGE_GROUP + OCCUR_HOUR + OCCUR_DAY + Latitude + Longitude,
##     family = binomial, data = df_2)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9208  -0.6805  -0.6070  -0.2244   2.9329
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     41.010992  86.674411   0.473 0.636098
## PERP_RACEASIAN / PACIFIC ISLANDER  9.970625  84.227789   0.118 0.905769
## PERP_RACEBLACK                   9.522565  84.227590   0.113 0.909985
## PERP_RACEBLACK HISPANIC          9.400786  84.227620   0.112 0.911131
## PERP_RACEUnknown                 9.048532  84.227842   0.107 0.914448
## PERP_RACEWHITE                  10.152437  84.227687   0.121 0.904059
## PERP_RACEWHITE HISPANIC          9.679434  84.227601   0.115 0.908509
## PERP_SEXM                       -0.190155   0.120994  -1.572 0.116043
## PERP_SEXUnknown                  2.429474   0.268105   9.062  < 2e-16 ***
## PERP_AGE_GROUP18-24              0.171250   0.075386   2.272 0.023108 *
## PERP_AGE_GROUP25-44              0.505662   0.075080   6.735 1.64e-11 ***
## PERP_AGE_GROUP45-64              0.837458   0.114547   7.311 2.65e-13 ***
## PERP_AGE_GROUP65+                1.008423   0.282806   3.566 0.000363 ***
## PERP_AGE_GROUPUnknown           -2.240328   0.171340 -13.075  < 2e-16 ***
## OCCUR_HOUR                      -0.002730   0.001978  -1.380 0.167503
## OCCUR_DAY.L                     -0.057112   0.039780  -1.436 0.151090
## OCCUR_DAY.Q                     -0.085781   0.042774  -2.005 0.044915 *
## OCCUR_DAY.C                     -0.056694   0.043019  -1.318 0.187541
## OCCUR_DAY^4                     -0.022589   0.043831  -0.515 0.606290
## OCCUR_DAY^5                     -0.009570   0.046033  -0.208 0.835305
## OCCUR_DAY^6                     -0.059742   0.047397  -1.260 0.207499
## Latitude                        -0.434742   0.190205  -2.286 0.022274 *
## Longitude                        0.461736   0.241068   1.915 0.055445 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 25076  on 25592  degrees of freedom
## Residual deviance: 24150  on 25570  degrees of freedom
## AIC: 24196
##
## Number of Fisher Scoring iterations: 9
```

###Question 7: How does the frequency distribution of the time of stop look like?

Step 4: Identify Bias

In this topic, it can spur discrimination and implicit bias unbeknownst among individuals. If I based my judgement on prior experience after living near New York City for a while, I would personally believe that Bronx must have had the most number of incidents. I might make an assumption that the incidents are more likely to occur with women than those of men. However, I must validate all the conviction with data, so I can make a better, well-informed decision. It's intriguing to find out that Brooklyn is the 1st in terms of the number of incidents, followed by Bronx and Queens respectively. Likewise, the number of murder

cases follows the same pattern as that of incidents. In addition, there are significantly more incidents with Male than those of Female. It's best to test and validate the assumption in a data-driven way rather than believing in your experience it all, which may be seriously wrong and biased towards a certain group and population. My finding is consistent with CNN's report on "Hate crimes, shooting incidents in New York City have surged since last year", especially that "shooting incidents in NYC increase by 73% for May 2021 vs. May 2020."