NYPD-Project-Report

Saikat Sengupta

2023-05-02

Step 0: Import Library

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

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
```

Step 1: Load Data

```
df = 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 2: Tidy and Transform Data

```
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] 10
##
## $Longitude
## [1] 10
```

```
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:27309       Length:27309       Length:27309      BRONX        : 7935
##  Class :character   Class :character   Class1:hms        BROOKLYN     :10932
##  Mode  :character   Mode  :character   Class2:difftime   MANHATTAN    : 3572
##                                        Mode  :numeric    QUEENS       : 4094
##                                                          STATEN ISLAND:  776
##
##
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP     PERP_SEX             PERP_RACE
##  Mode :logical           (null) :  640   (null) :  640   BLACK         :11431
##  FALSE:22043             <18    : 1591   F      :  424   Unknown       :11146
##  TRUE :5266              18-24  : 6222   M      :15436   WHITE HISPANIC: 2339
##                          25-44  : 5687   Unknown:10809   BLACK HISPANIC: 1314
##                          45-64  :  617                   (null)        :  640
##                          65+    :   60                   WHITE         :  283
##                          Unknown:12492                   (Other)       :  156
##  VIC_AGE_GROUP      VIC_SEX                                VIC_RACE
##  <18    : 2839   F      : 2615   AMERICAN INDIAN/ALASKAN NATIVE:   10
##  1022   :    1   M      :24683   ASIAN / PACIFIC ISLANDER      :  404
##  18-24  :10085   Unknown:   11   BLACK                         :19438
##  25-44  :12279                  BLACK HISPANIC                : 2646
##  45-64  : 1863                  Unknown                       :   66
##  65+    :  181                  WHITE                         :  698
##  UNKNOWN:   61                  WHITE HISPANIC                : 4047
##     Latitude       Longitude
##  Min.   :40.51   Min.   :-74.25
##  1st Qu.:40.67   1st Qu.:-73.94
##  Median :40.70   Median :-73.92
```

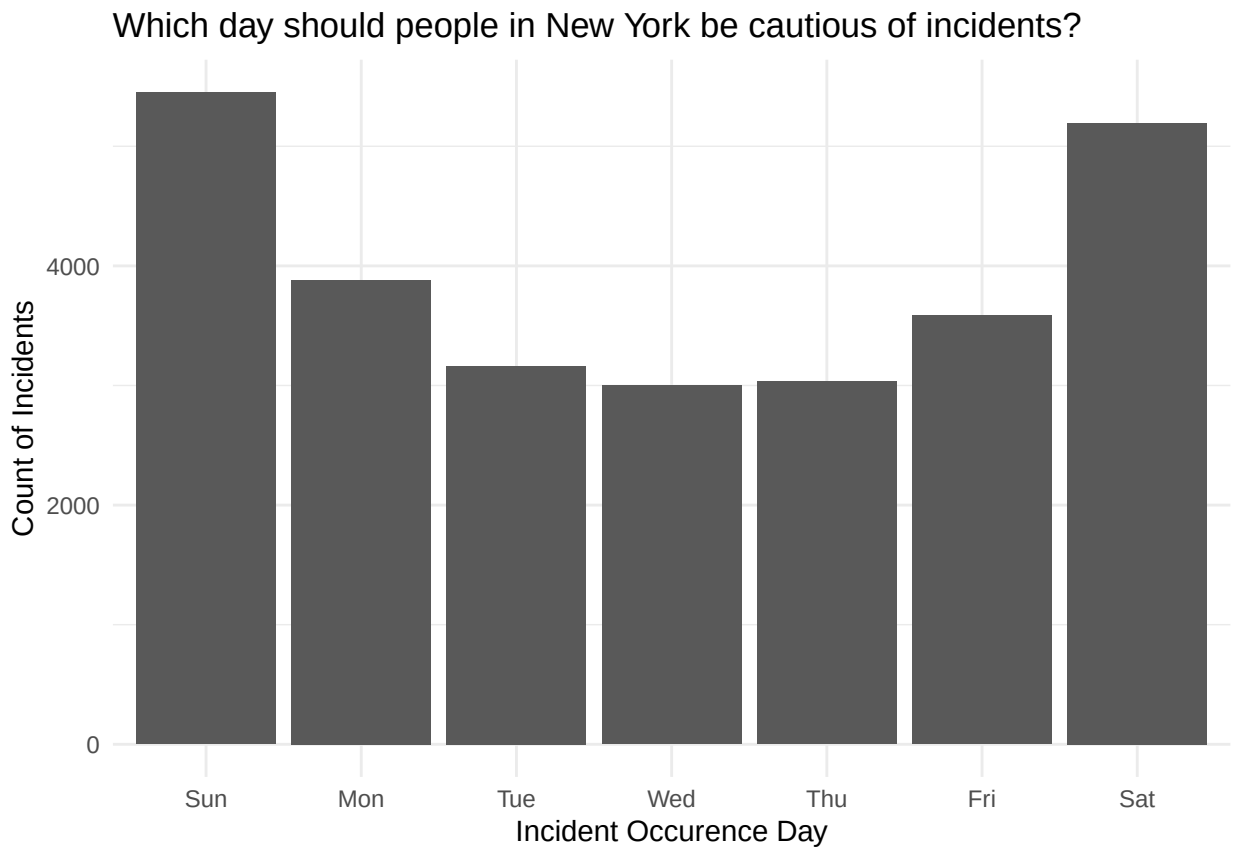
```
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##  NA's   :10      NA's   :10
```

Step 3: Visualizations and Analysis

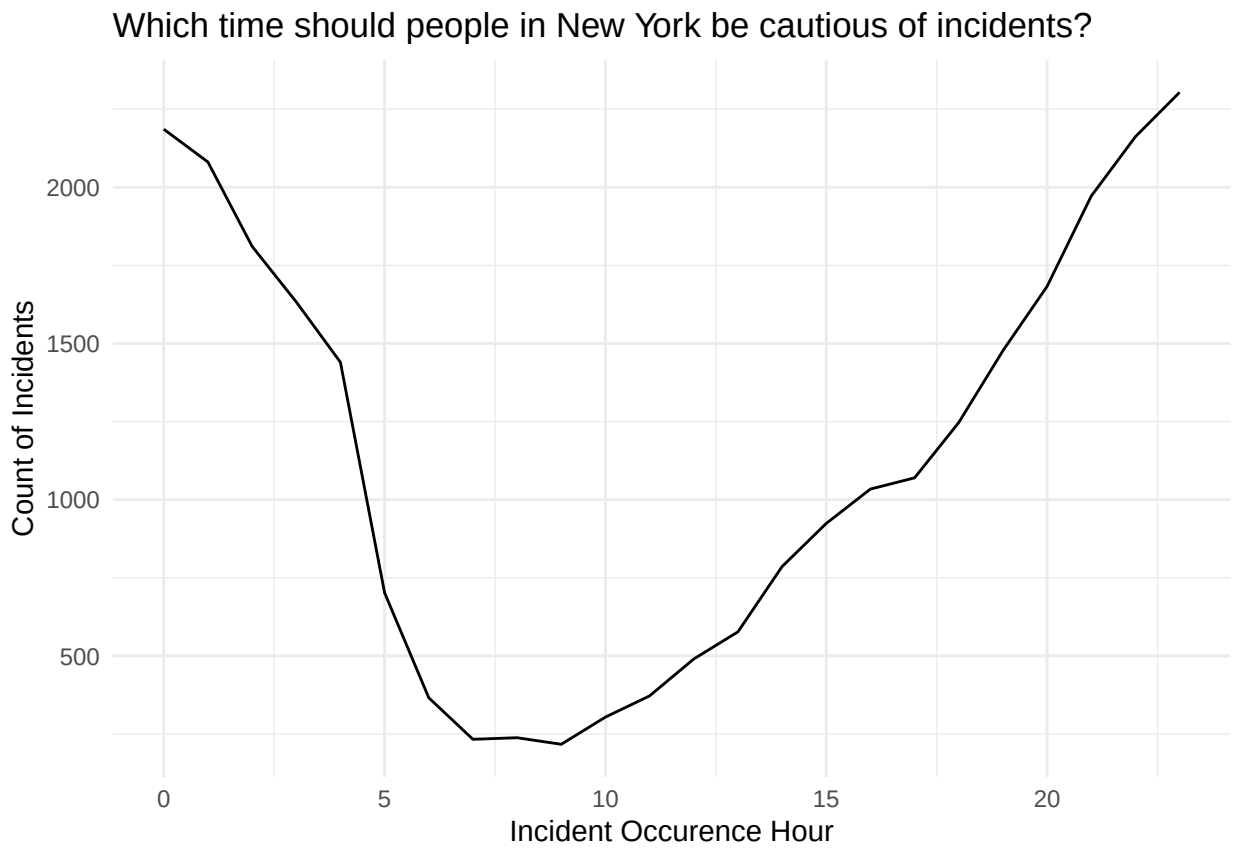
Which day and time should people in New York be cautious of falling into victims of crime? • Weekends in NYC have the most chances of incidents. Be cautious! • Incidents historically happen in the evening and night time. If there's nothing urgent, recommend people staying at home!

```
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
```

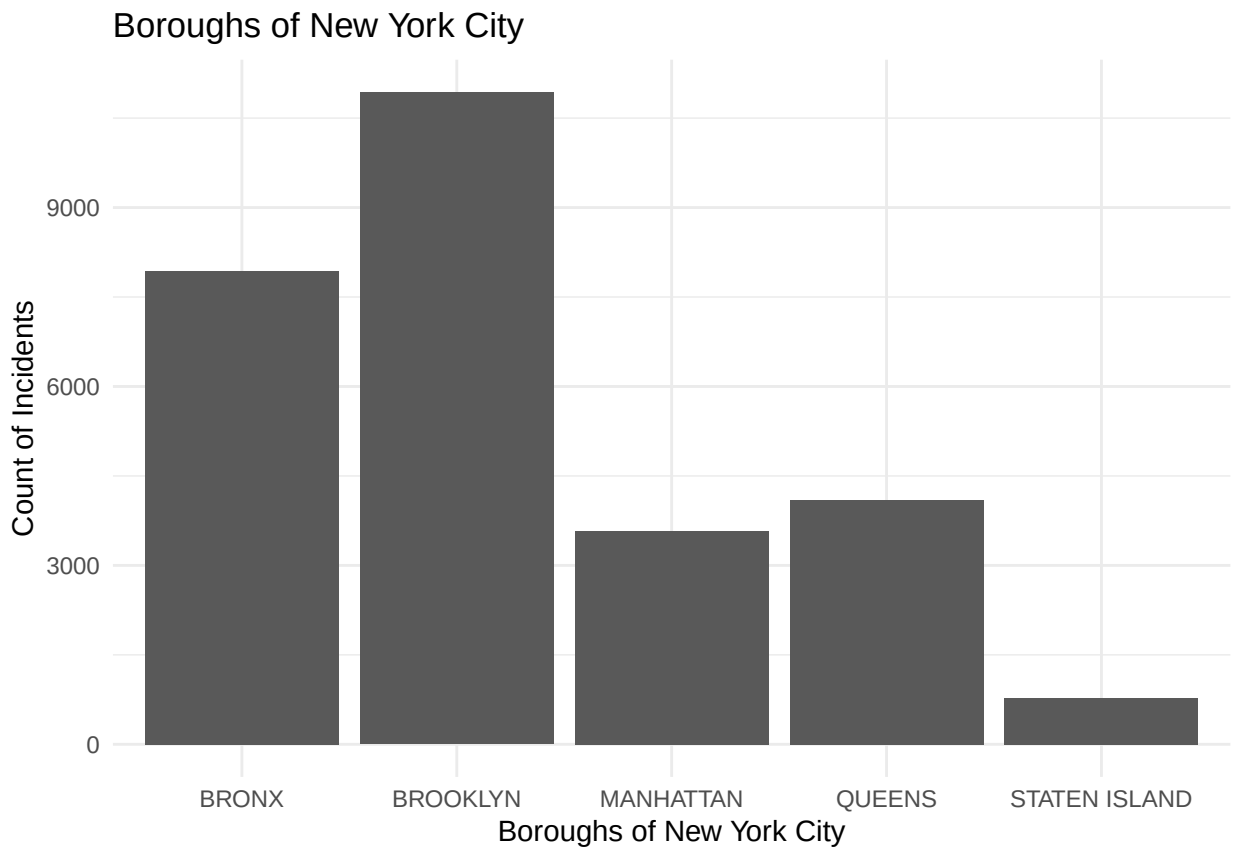



```
g <- ggplot(df_4, aes(x = OCCUR_HOUR, y = n)) +
  geom_line() +
  labs(title = "Which time should people in New York be cautious of incidents?",
       x = "Incident Occurence Hour",
       y = "Count of Incidents") +
  theme_minimal()
g
```



2. 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
```



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 1022 18-24 25-44 45-64  65+ UNKNOWN
##   (null)    57    0   181   340    57    5       0
##   <18      484    0   621   397    77   10       2
##   18-24    788    1  2758  2294   329   40      12
##   25-44    262    0  1516  3352   479   43      35
##   45-64     20    0    76   327   177   12       5
##   65+        0    0     1    25    23   11       0
##   Unknown 1228    0  4932  5544   721   60       7
```

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

```
##
##               F     M Unknown
##   (null)     72   568       0
##   F          72   351       1
##   M        1666 13764       6
##   Unknown   805 10000       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.

```
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)
##
## Coefficients: (2 not defined because of singularities)
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          49.825253  19.849788   2.510 0.012069
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE -8.915610  84.241402  -0.106 0.915714
## PERP_RACEASIAN / PACIFIC ISLANDER        1.027692   0.295457   3.478 0.000505
## PERP_RACEBLACK                           0.583282   0.236967   2.461 0.013838
## PERP_RACEBLACK HISPANIC                  0.500464   0.246258   2.032 0.042125
## PERP_RACEUnknown                         0.114060   0.114303   0.998 0.318340
## PERP_RACEWHITE                           1.192839   0.268215   4.447  8.7e-06
## PERP_RACEWHITE HISPANIC                  0.732434   0.241341   3.035 0.002406
## PERP_SEXF                               -2.459168   0.264949  -9.282  < 2e-16
## PERP_SEXM                               -2.615159   0.239331 -10.927  < 2e-16
## PERP_SEXUnknown                                NA         NA      NA       NA
## PERP_AGE_GROUP<18                        2.232264   0.170345  13.104  < 2e-16
## PERP_AGE_GROUP18-24                      2.413127   0.160286  15.055  < 2e-16
## PERP_AGE_GROUP25-44                      2.726829   0.160268  17.014  < 2e-16
## PERP_AGE_GROUP45-64                      3.091787   0.179314  17.242  < 2e-16
## PERP_AGE_GROUP65+                        3.243423   0.310185  10.456  < 2e-16
## PERP_AGE_GROUPUnknown                          NA         NA      NA       NA
## OCCUR_HOUR                              -0.002167   0.001916  -1.131 0.257959
## OCCUR_DAY.L                             -0.040648   0.038500  -1.056 0.291074
## OCCUR_DAY.Q                             -0.079104   0.041301  -1.915 0.055455
## OCCUR_DAY.C                             -0.058826   0.041569  -1.415 0.157029
## OCCUR_DAY^4                             -0.012408   0.042343  -0.293 0.769489
## OCCUR_DAY^5                              0.017122   0.044427   0.385 0.699941
## OCCUR_DAY^6                             -0.075924   0.045700  -1.661 0.096645
## Latitude                                -0.383301   0.183827  -2.085 0.037058
## Longitude                                0.485996   0.234079   2.076 0.037875
##
## (Intercept)                          *
## PERP_RACEAMERICAN INDIAN/ALASKAN NATIVE
## PERP_RACEASIAN / PACIFIC ISLANDER    ***
## PERP_RACEBLACK                       *
## PERP_RACEBLACK HISPANIC              *
## PERP_RACEUnknown
## PERP_RACEWHITE                       ***
## PERP_RACEWHITE HISPANIC              **
## PERP_SEXF                            ***
## PERP_SEXM                            ***
```

```
## PERP_SEXUnknown
## PERP_AGE_GROUP<18                        ***
## PERP_AGE_GROUP18-24                      ***
## PERP_AGE_GROUP25-44                      ***
## PERP_AGE_GROUP45-64                      ***
## PERP_AGE_GROUP65+                        ***
## PERP_AGE_GROUPUnknown
## OCCUR_HOUR
## OCCUR_DAY.L
## OCCUR_DAY.Q                              .
## OCCUR_DAY.C
## OCCUR_DAY^4
## OCCUR_DAY^5
## OCCUR_DAY^6                              .
## Latitude                                 *
## Longitude                                *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 26775  on 27298  degrees of freedom
## Residual deviance: 25831  on 27275  degrees of freedom
##   (10 observations deleted due to missingness)
## AIC: 25879
##
## Number of Fisher Scoring iterations: 9
```

Step 4: Bias

This topic has the potential to generate unconscious discrimination and bias in individuals. Based on my personal experience of living near New York City, I would assume that the Bronx has the highest number of incidents and that women are more likely to be targeted than men. However, it is essential to support these beliefs with data to make a well-informed decision. It is interesting to note that Brooklyn has the highest number of incidents, followed by the Bronx and Queens, and the number of murders follows a similar pattern. Moreover, there are significantly more incidents involving males than females. It is important to test and verify these assumptions using a data-driven approach instead of relying solely on personal experience, which could be biased and incorrect towards certain groups and populations. My findings align with CNN's report on the surge of hate crimes and shooting incidents in New York City, where shooting incidents increased by 73% in May 2021 compared to May 2020.