NYPD Shooting Incident Data Report

T. Shreeve

12/12/2023

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

```
new_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.
```

```
head(df)
```

```
##
## 1 function (x, df1, df2, ncp, log = FALSE)
## 2 {
```

```
## 3     if (missing(ncp))
## 4         .Call(C_df, x, df1, df2, log)
## 5     else .Call(C_dnf, x, df1, df2, ncp, log)
## 6 }
```

Step 2: Tidy and Transform Data

We are going to focus on the location and time of these events so let's first eliminate the columns: **PRECINCT**, **JURISDICTION_CODE**, **LOCATION_DESC**, **X_COORD_CD**, **Y_COORD_CD**, and **Lon_Lat**.

```
df_2 = new_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
```

This shows us the missing values in the data set. We can see that there are a number of entries that do not have information about the perpetrator here. The data could be missing for a number of reasons such as the cases are being currently unsolved. In order to not data seemingly disappear, I will refer to any piece of data that is missing as a part of the group "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)
```

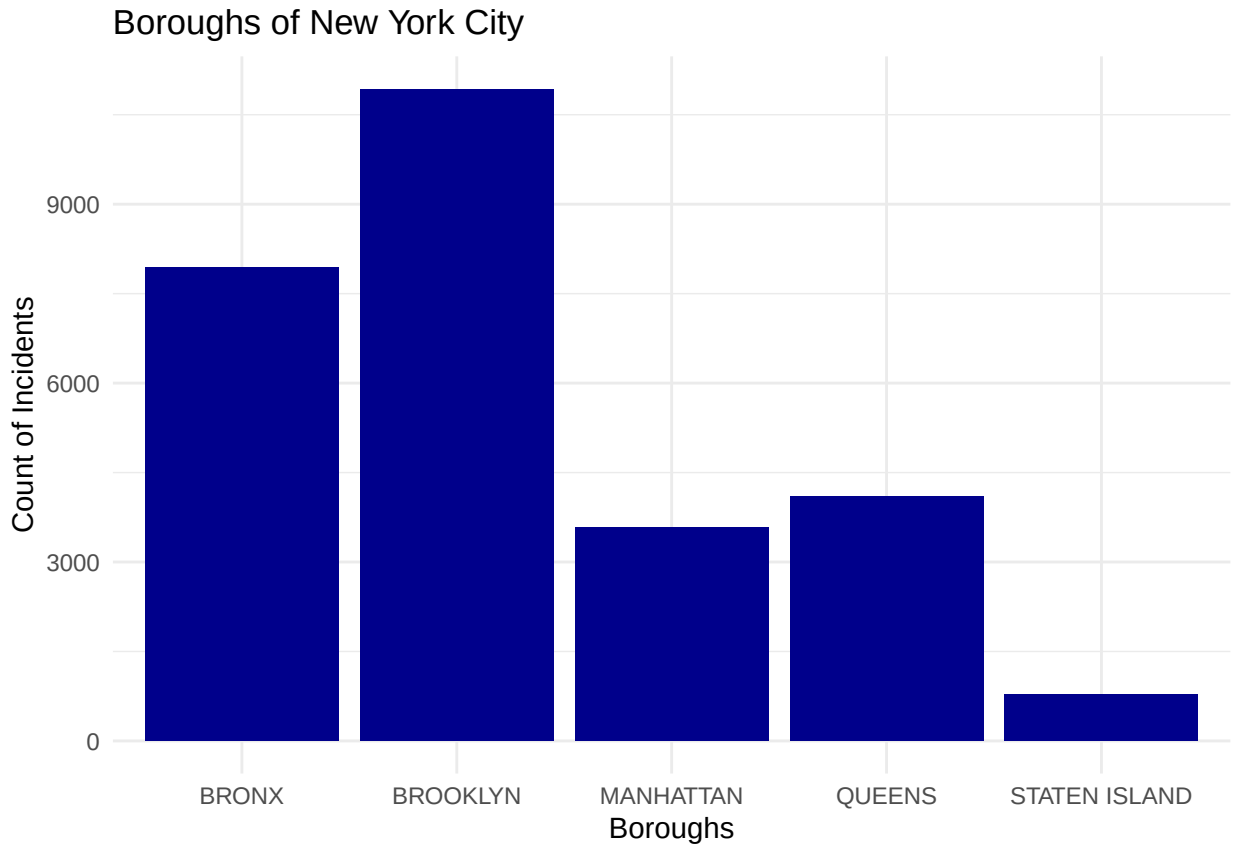
Step 3: Add Visualizations and Analysis

Research Question

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

Brooklyn has the highest number of incidents, followed by Bronx and Queens respectively.

```
g <- ggplot(df_2, aes(x = BORO)) +
  geom_bar(fill='darkblue') +
  labs(title = "Boroughs of New York City",
       x = "Boroughs",
       y = "Count of Incidents") +
  theme_minimal()
g
```



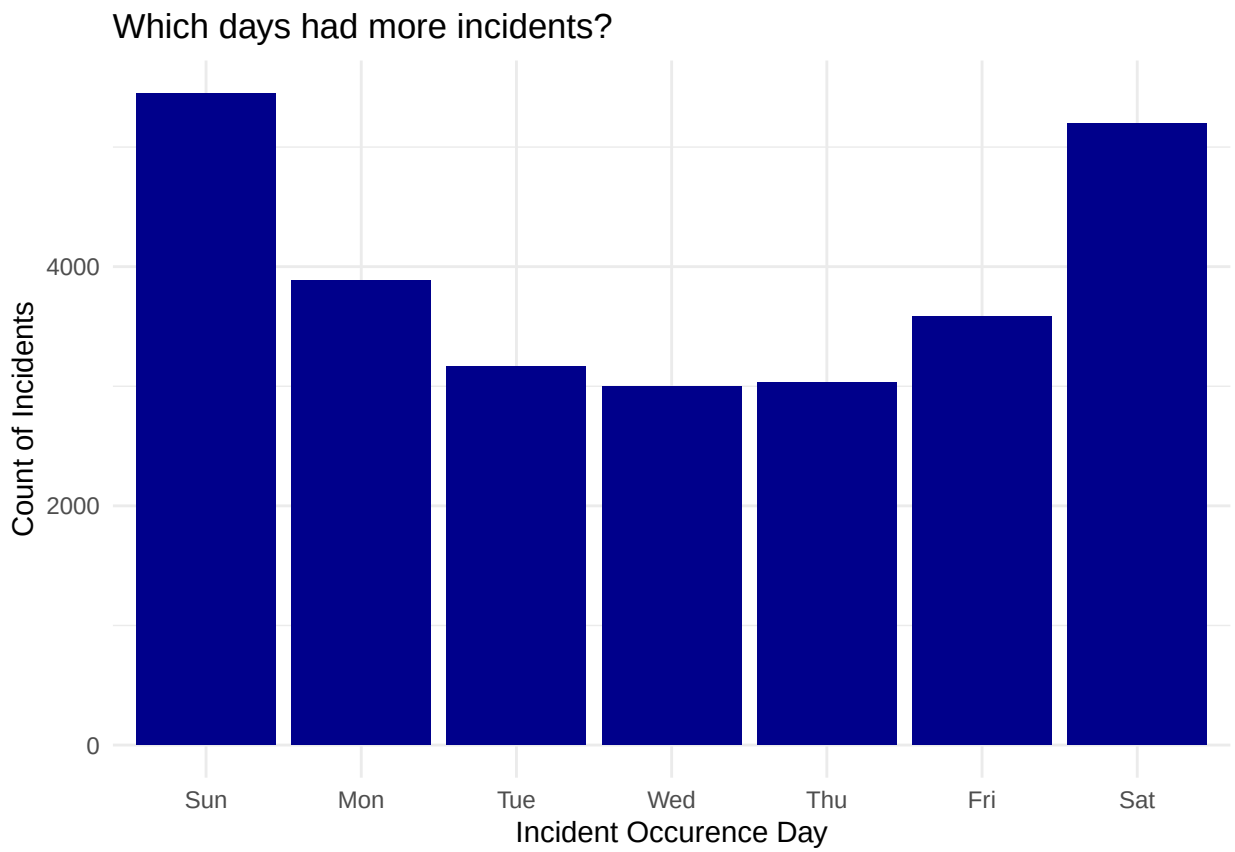

```
table(df_2$BORO, df_2$STATISTICAL_MURDER_FLAG)
```

```
##
##                FALSE TRUE
##   BRONX         6393 1542
##   BROOKLYN      8810 2122
##   MANHATTAN     2942  630
##   QUEENS        3284  810
##   STATEN ISLAND  614  162
```

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

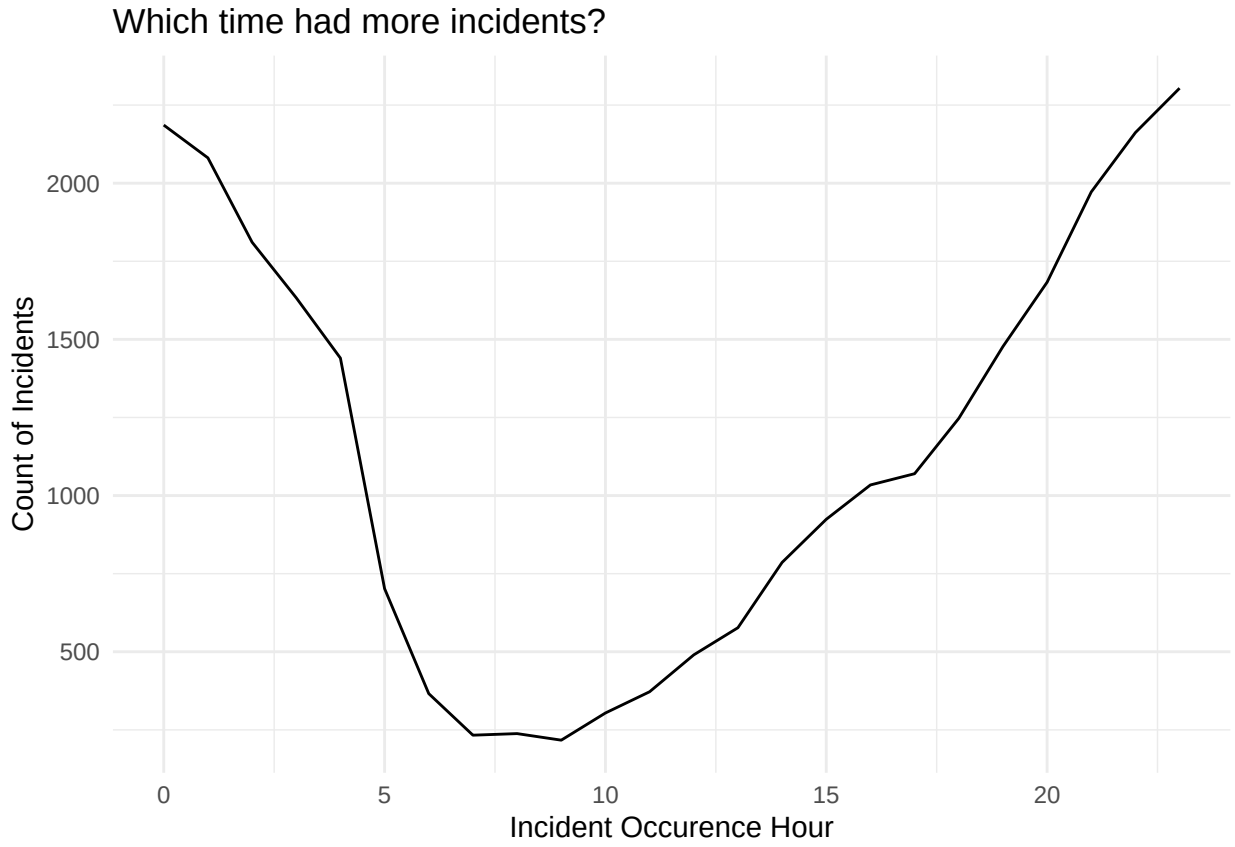
```
df_2$OCCUR_DAY = mdy(df_2$OCCUR_DATE)
df_2$OCCUR_DAY = wday(df_2$OCCUR_DAY, label = TRUE)
df_2$OCCUR_HOUR = hour(hms(as.character(df_2$OCCUR_TIME)))

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(fill='darkblue') +
  labs(title = "Which days had more 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 had more incidents?",
```

```
       x = "Incident Occurence Hour",
       y = "Count of Incidents") +
  theme_minimal()
g
```



It seems as though the most dangerous time and place to be would be Brooklyn on a Sunday around 11pm. Cross referencing this information could provide interesting information about what areas are the either safer or more dangerous during specific times.

Step 4: Identify Bias

Some of the boroughs of NYC have a significant wealth inequality which I thought could lead to those boroughs having a higher number of shootings. This is a bias and while it didn't exactly get in the way of my questions and data, it drove what questions I did ask. I attempted to fight this bias by looking more at the time and day that the shootings happen in order to see the data from another way. This instead lead to me wondering which areas were most likely to have a shooting base on the time and day. I had made the assumption that the three most prevelant areas would be Manhattan, the Bronx, and Brooklyn but only two of those are correct.