NYPD Shooting Incident Data Report

5/5/2023

```
{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE)
```

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

```
{r library, message=FALSE, warning=FALSE} #
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.

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

Step 2: Tidy and Transform Data

Let’s first eliminate the columns I do not need for this assignment, which are: **PRECINCT**,**JURISDICTION_CODE**,**LOCATION_DESC**, **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)))
```

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)
```

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)
```

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)

table(df_2$PERP_SEX, df_2$VIC_SEX)

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

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 + Latitude + Longitude, data =
df_2, family = binomial)
summary(glm.fit)
```

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.”

Additional Resources

- NYPD Shooting Incident Data (Historic) - CKAN
- NYC, Chicago see another wave of weekend gun violence
- Hate crimes, shooting incidents in New York City have surged since last year, NYPD data show - CNN