Final Project 1

N Enos

2023-10-28

DTSA 5301 Final Project 1: NYPD Shooting Incident Data Report

This report analyzes the NYPD Shooting Incident Data. This data is supplied by the City of New York and provides information about shooting incidents from 2006 through the end of the previous calendar year. It is extracted and reviewed quarterly by the Office of Management Analysis and Planning before it is posted on the NYPD website. Each row is a shooting incident in NYC.

The purpose of this report is to identify trends in the timing of shootings and the demographics of the perpetrators and victims of shootings.

Link to data source: https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8

Data Import: Step 1

This is the markdown document for the DTSA 5301 Shootings Data Week 3 Project. The first section of code imports the shooting dataset.

```r
# Read in shooting data and display a summary
shootings <- 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.
```

```r
summary(shootings)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:27312       Length:27312      Length:27312
##  1st Qu.: 63860880   Class :character   Class1:hms        Class :character
##  Median : 90372218   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :120860536                      Mode  :numeric
##  3rd Qu.:188810230
##  Max.   :261190187
##
##  LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:27312       Min.   :  1.00   Min.   :0.0000    Length:27312
##  Class :character   1st Qu.: 44.00   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 68.00   Median :0.0000    Mode  :character
##                     Mean   : 65.64   Mean   :0.3269
##                     3rd Qu.: 81.00   3rd Qu.:0.0000
##                     Max.   :123.00   Max.   :2.0000
##                                      NA’s   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:27312       Mode :logical           Length:27312
##  Class :character   FALSE:22046             Class :character
##  Mode  :character   TRUE :5266              Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:27312       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000028   1st Qu.:182834   1st Qu.:40.67
##  Mode  :character   Median :1007731   Median :194487   Median :40.70
##                     Mean   :1009449   Mean   :208127   Mean   :40.74
##                     3rd Qu.:1016838   3rd Qu.:239518   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA’s   :10
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:27312
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA’s   :10
```

Data Cleanup: Step 2

This step cleans the data and displays the new summary. I have selected the OCCUR_DATE, PRECINCT, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VIC_AGE_GROUP, VIC_SEX, and VIC_RACE columns and have converted OCCUR_DATE into a date. I have also factored the PERP_AGE_GROUP, PERP_SEX, VIC_AGE_GROUP, VIC_SEX, and PRECINCT columns. Additionally, I filtered the age groups to include only the valid age ranges.

```r
# Clean up data by selecting only needed columns,
# converting dates, and factoring columns as appropriate.
shootings <- shootings %>%
  select(OCCUR_DATE,
         PRECINCT,
         PERP_AGE_GROUP,
         PERP_SEX,
         PERP_RACE,
         VIC_AGE_GROUP,
         VIC_SEX,
         VIC_RACE) %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE))
shootings$PERP_AGE_GROUP <- factor(shootings$PERP_AGE_GROUP)
shootings$PERP_SEX <- factor(shootings$PERP_SEX)
shootings$VIC_AGE_GROUP <- factor(shootings$VIC_AGE_GROUP)
shootings$VIC_SEX <- factor(shootings$VIC_SEX)
shootings$PRECINCT <- factor(shootings$PRECINCT)

# Set valid age range values
valid_age_ranges <- c('<18', '18-24', '25-44', '45-64', '65+', 'UNKNOWN')
shootings <- shootings %>% filter(VIC_AGE_GROUP %in% valid_age_ranges)
shootings <- shootings %>% filter(PERP_AGE_GROUP %in% valid_age_ranges)

# Display a summary of the cleaned data
summary(shootings)
```

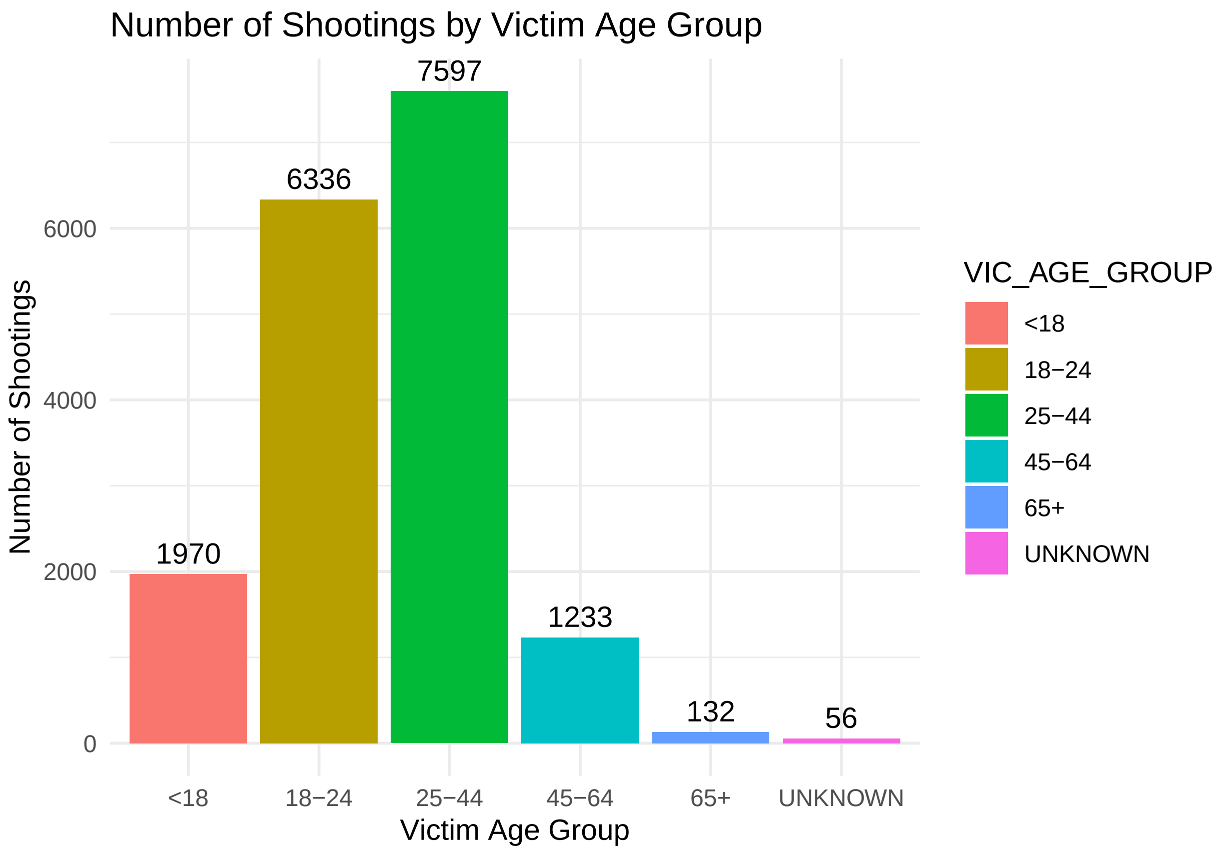
```
##    OCCUR_DATE            PRECINCT     PERP_AGE_GROUP   PERP_SEX
##  Min.   :2006-01-01   75     :  959   18-24  :6221   (null):    0
##  1st Qu.:2008-07-13   73     :  828   25-44  :5687   F     :  424
##  Median :2011-08-01   47     :  678   UNKNOWN:3148   M     :15435
##  Mean   :2013-01-08   44     :  663   <18    :1591   U     : 1465
##  3rd Qu.:2017-06-24   46     :  646   45-64  : 617
##  Max.   :2022-12-31   79     :  582   65+    :  60
##                       (Other):12968   (Other):   0
##   PERP_RACE          VIC_AGE_GROUP  VIC_SEX     VIC_RACE
##  Length:17324       <18    :1970   F: 1850   Length:17324
##  Class :character   1022   :   0   M:15466   Class :character
##  Mode  :character   18-24  :6336   U:    8   Mode  :character
##                     25-44  :7597
##                     45-64  :1233
##                     65+    : 132
##                     UNKNOWN:  56
```

Visualizations: Step 3

My first visualization displays the count of shootings by victim age group using a bar plot. I grouped the data by VIC_AGE_GROUP and summarized by counting the number of shootings.

```r
# Set valid age range values
valid_age_ranges <- c('<18', '18-24', '25-44', '45-64', '65+', 'UNKNOWN')

# Group the data to get the counts of shootings by victim age group
```

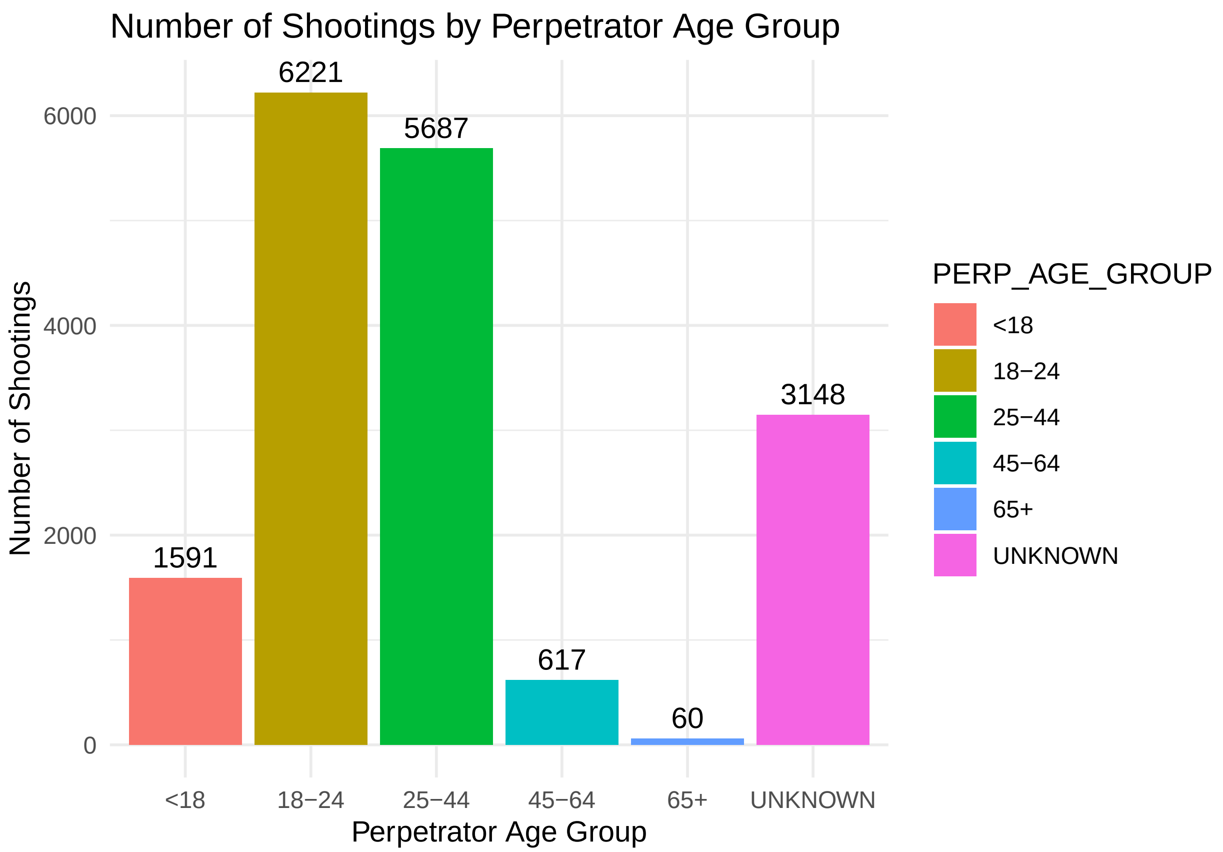
```
shootings_by_vic_age <- shootings %>%
  group_by(VIC_AGE_GROUP) %>%
  summarize(shooting_count = n()) %>%
  select(VIC_AGE_GROUP, shooting_count) %>%
  ungroup()

# Bar chart of shootings by victim age group
ggplot(shootings_by_vic_age, aes(x=VIC_AGE_GROUP,
                                 y=shooting_count,
                                 fill=VIC_AGE_GROUP)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label=shooting_count), vjust=-.5) +
  theme_minimal() +
  labs(title = "Number of Shootings by Victim Age Group",
       x = "Victim Age Group",
       y = "Number of Shootings")
```



My second visualization displays the count of shootings by perpetrator age group using a bar plot. I grouped the data by PERP_AGE_GROUP and summarized by counting the number of shootings.

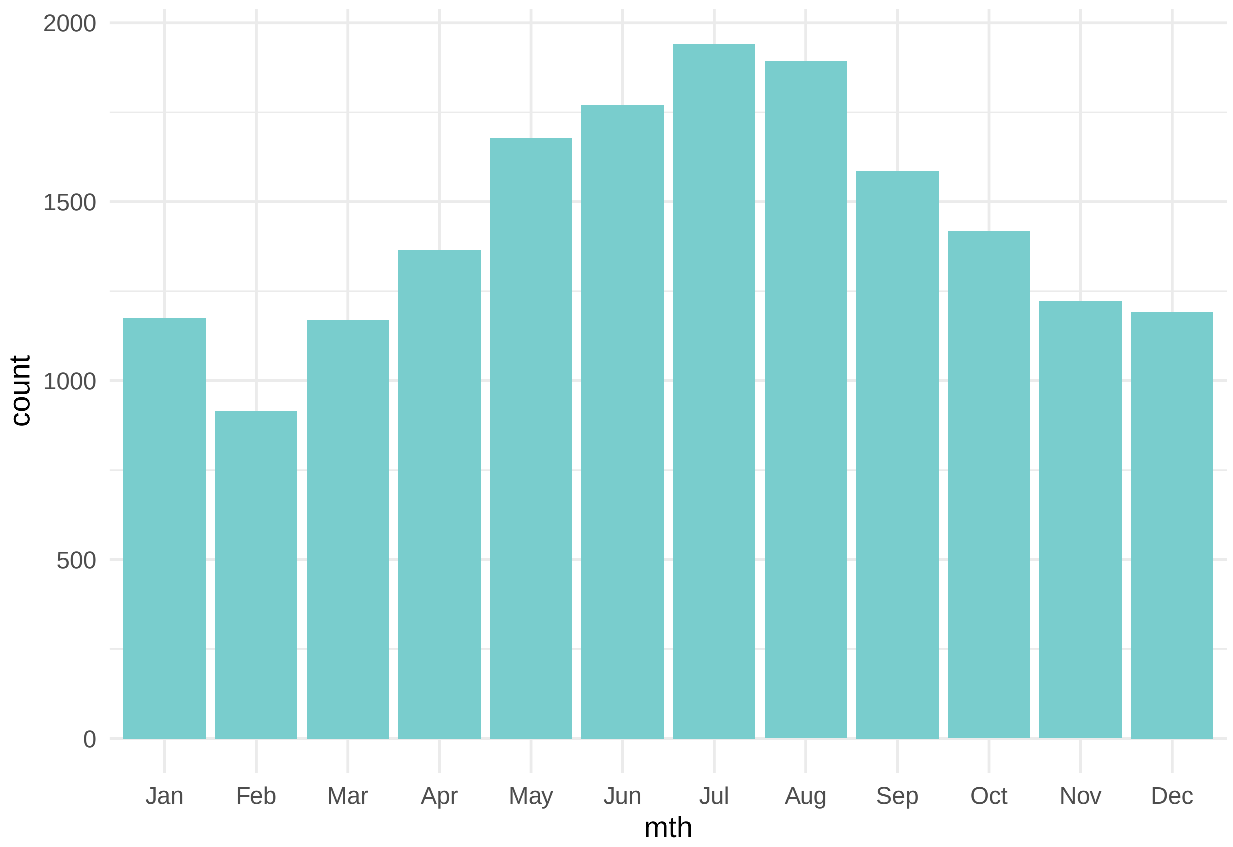
```
# Do the same for perp age group
shootings_by_perp_age <- shootings %>%
  group_by(PERP_AGE_GROUP) %>%
  summarize(shooting_count = n()) %>%
  select(PERP_AGE_GROUP, shooting_count) %>%
  ungroup()
```

```
# Bar chart of shootings by perp age group
ggplot(shootings_by_perp_age, aes(x=PERP_AGE_GROUP,
                                  y=shooting_count,
                                  fill=PERP_AGE_GROUP)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label=shooting_count), vjust=-.5) +
  theme_minimal() +
  labs(title = "Number of Shootings by Perpetrator Age Group",
       x = "Perpetrator Age Group",
       y = "Number of Shootings")
```



My second visualization displays the count of by month. I added a month column to the data set with the abbreviated name of the month, and used a bar chart to display the shooting count for each month.

```
# add month and year to shootings data
shootings$mth <- month(shootings$OCCUR_DATE,
                       label=TRUE, abbr=TRUE)
shootings$yr <- year(shootings$OCCUR_DATE)

# graph shootings by month
ggplot(shootings, aes(mth)) +
  geom_bar(fill='darkslategray3') +
  theme_minimal()
```



Data Analysis: Step 3

For my analysis, I determined the months with the least and most shootings in 2021 by filtering the data to 2021, grouping the counts by month, and then using the slice_min and slice_max functions.

I also determined the percentage of shooting perpetrators and victims by age group. I utilized the table functionality to break the age ranges into rows and get the percentage of shootings attributed to each range, for both the PERP_AGE_GROUP and VIC_AGE_GROUP columns.

```
# get month with the most and least numbers of shootings in 2021
shootings_2021 <- shootings %>% filter(yr == 2021)
shootings_2021_by_month <- shootings %>%
  group_by(mth) %>%
  summarize(shooting_count = n()) %>%
  select(mth, shooting_count) %>%
  ungroup()
print('The month with the least shootings in 2021:')
```

```
## [1] "The month with the least shootings in 2021:"
```

```
print(shootings_2021_by_month %>% slice_min(shooting_count))
```

```
## # A tibble: 1 x 2
##   mth   shooting_count
##   <ord>          <int>
## 1 Feb              914
```

```
print('The month with the most shootings in 2021:')
```

```
## [1] "The month with the most shootings in 2021:"
```

```
print(shootings_2021_by_month %>% slice_max(shooting_count))
```

```
## # A tibble: 1 x 2
##   mth   shooting_count
##   <ord>          <int>
## 1 Jul             1942
```

```
# get percentage of shooting perps by age group
perp_counts <- table(shootings$PERP_AGE_GROUP)
total_count <- sum(perp_counts)
percentage_by_perp_age <- (perp_counts / total_count) * 100
perc_by_perp_age <- data.frame(percentage_by_perp_age)
perc_by_perp_age <- perc_by_perp_age %>%
  rename(
    age_group = Var1,
    percentage = Freq
  )
perc_by_perp_age <- perc_by_perp_age %>% filter(percentage > 0)
print('The percentage of shooting perpetrators in each age group:')
```

```
## [1] "The percentage of shooting perpetrators in each age group:"
```

```
print(perc_by_perp_age)
```

```
##   age_group percentage
## 1       <18  9.1837913
## 2     18-24 35.9097206
## 3     25-44 32.8272916
## 4     45-64  3.5615331
## 5       65+  0.3463403
## 6   UNKNOWN 18.1713230
```

```
# get percentage of shooting victims by age group
vic_counts <- table(shootings$VIC_AGE_GROUP)
total_count <- sum(perp_counts)
percentage_by_vic_age <- (vic_counts / total_count) * 100
perc_by_vic_age <- data.frame(percentage_by_vic_age)
perc_by_vic_age <- perc_by_vic_age %>%
  rename(
    age_group = Var1,
    percentage = Freq
  )
perc_by_vic_age <- perc_by_vic_age %>% filter(percentage > 0)
print('The percentage of shooting victims in each age group:')
```

```
## [1] "The percentage of shooting victims in each age group:"
```

```
print(perc_by_vic_age)
```

```
##   age_group percentage
## 1       <18 11.3715077
## 2     18-24 36.5735396
## 3     25-44 43.8524590
## 4     45-64  7.1172939
## 5       65+  0.7619487
## 6   UNKNOWN  0.3232510
```

Some questions we might want to investigate based on this analysis:

1. Why is the age of so many shooting perpetrators unknown?
2. Why does February have so few shootings compared to other months in 2021?
3. Why does July have so many shootings compared to other months in 2021?
4. Are these monthly trends similar in other years?
5. The age ranges are pretty broad. What are the actual mean and median ages for shooting perpetrators and victims?

Bias Identification: Step 4

Shootings in NYC are a major problem. It's important to investigate when the majority of shootings occur, who typically commits them, and who the victims are. Some sources of bias may be that certain neighborhoods in New York are more heavily policed and surveilled. This may result in an under representation of other areas that are not included as much in the data. Additionally, the identification of perpetrators often relies on witnesses that are notoriously unreliable. It's important to recognize that this data might be incomplete or inaccurate. My recommendation would be to understand that there is a degree of uncertainty and inaccuracy of this data and it should not be considered a complete source of truth.