NYPD Shooting Incident Data Report

M. Jovanovski

2023-03-17

Libraries

The most important library for analyzing and visualizing data is tidyverse. This library consists of many libraries that can be used for data analysis and data visualization. I will use dlpyr and ggplot2 functions to perform most of the data analysis and data visualization tasks. I will also use library lubridate to convert date variable into date data type.

```
library(lubridate)

##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr   1.1.0     ✓ readr   2.1.4
## ✓ forcats 1.0.0     ✓ stringr 1.5.0
## ✓ ggplot2 3.4.1     ✓ tibble  3.1.8
## ✓ purrr   1.0.1     ✓ tidyr   1.3.0

## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force
all conflicts to become errors
```

Data Analysis Objective

This data analysis tries to answer the questions how gender, age group and hour of the day relate to being victim of shooting incidents occurred in New York.

Import Dataset

The dataset for this project was retrieved from https://catalog.data.gov/dataset. The observations represent the shooting incidents in different areas of New York over multiple years from 2006 to 2021. The link of the dataset is given below:

(https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic).I read the dataset from web using built-in function read.csv().

```
#reading data from csv file

NYPD_Shooting_Incident_Data__Historic <-
read.csv("https://data.cityofnewyork.us/api/views/833y-
fsy8/rows.csv?accessType=DOWNLOAD")
View(NYPD_Shooting_Incident_Data__Historic)

nypd <- data.frame(NYPD_Shooting_Incident_Data__Historic)
summary(nypd)

##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME            BORO
##  Min.   :  9953245   Length:25596       Length:25596       Length:25596
##  1st Qu.: 61593633   Class :character   Class :character   Class
:character
##  Median : 86437258   Mode  :character   Mode  :character   Mode
:character
##  Mean   :112382648
##  3rd Qu.:166660833
##  Max.   :238490103
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC
STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:25596       Length:25596
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   Class :character
##  Median : 69.00   Median :0.0000    Mode  :character   Mode  :character
##  Mean   : 65.87   Mean   :0.3316
##  3rd Qu.: 81.00   3rd Qu.:0.0000
##  Max.   :123.00   Max.   :2.0000
##                   NA's   :2
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE           X_COORD_CD        Y_COORD_CD
##  Length:25596       Length:25596       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.:1000011   1st Qu.:182782
##  Mode  :character   Mode  :character   Median :1007715   Median :194038
##                                        Mean   :1009455   Mean   :207894
##                                        3rd Qu.:1016838   3rd Qu.:239429
##                                        Max.   :1066815   Max.   :271128
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:25596
```

```
##  1st Qu.:40.67   1st Qu.:-73.94   Class :character
##  Median :40.70   Median :-73.92   Mode  :character
##  Mean   :40.74   Mean   :-73.91
##  3rd Qu.:40.82   3rd Qu.:-73.88
##  Max.   :40.91   Max.   :-73.70
##
```

Tidy and Transform

In this section, few variables of interest from the set of available variables are selected . These variables are gender, age, group and hour. The variable hour is extracted from the variable OCCUR_TIME. Because the selected variable are categorical, they are converted to factor data type.

```
#converting OCCUR_DATE to date data type
library(lubridate)
library('hms')

##
## Attaching package: 'hms'

## The following object is masked from 'package:lubridate':
##
##     hms

data <- nypd %>% select(OCCUR_DATE,OCCUR_TIME, VIC_AGE_GROUP,VIC_SEX) %>%
na.omit()
data$OCCUR_DATE <- mdy(data$OCCUR_DATE)
data$year <- year(data$OCCUR_DATE)
data$month <-month(data$OCCUR_DATE)
data$day <-day(data$OCCUR_DATE)
data$hour<-hour(parse_time(data$OCCUR_TIME))



data$year <- factor(data$year)
data$month <- factor(data$month)
data$day <- factor(data$day)
data$hour <- factor(data$hour)

data$gender <- factor(data$VIC_SEX)
data$age <- factor(data$VIC_AGE_GROUP)
```

The variables of interest are, OCCUR_TIME, VIC_SEX and VIC_AGE_GROUP. so I will select only these columns from data.

```
#selecting variables of interest
subData <- data%>%
  select(OCCUR_TIME, VIC_SEX, VIC_AGE_GROUP)
#checking null values in selected data
colSums(is.na(subData))
```

```
##    OCCUR_TIME        VIC_SEX VIC_AGE_GROUP
##             0              0             0
```

From above output, it is pretty evident that there is no null values in the dataset which means the dataset is already cleaned.

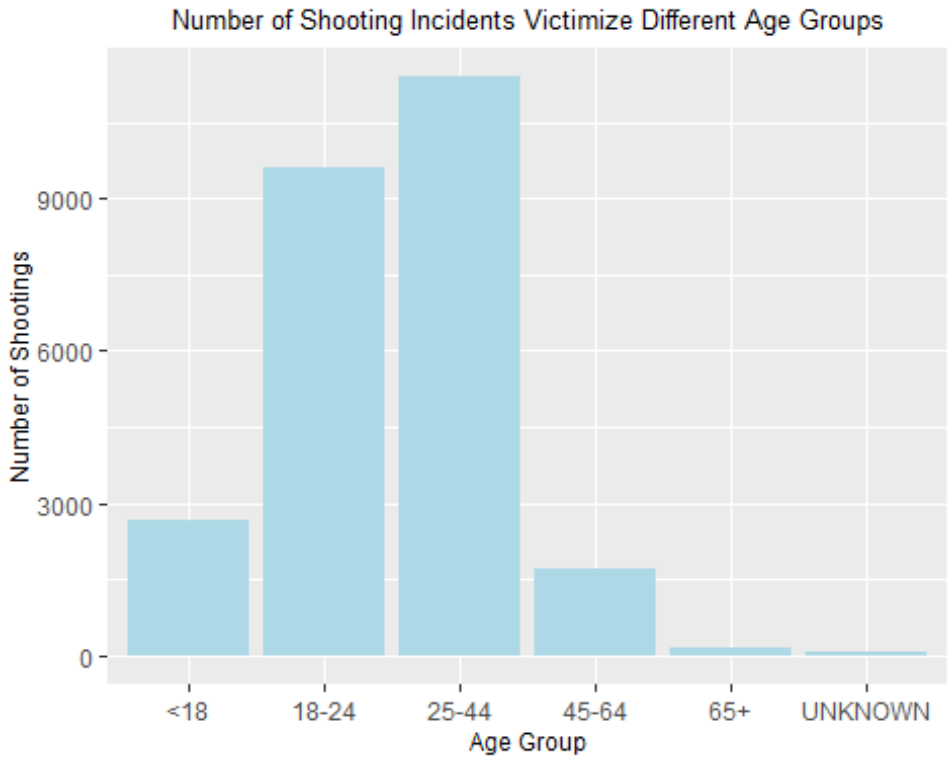
```
#summary of data
summary(subData)

##   OCCUR_TIME          VIC_SEX          VIC_AGE_GROUP
##  Length:25596       Length:25596       Length:25596
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character

glimpse(subData)

## Rows: 25,596
## Columns: 3
## $ OCCUR_TIME    <chr> "15:04:00", "22:05:00", "01:09:00", "13:42:00",
"20:00:0…
## $ VIC_SEX       <chr> "M", "M", "M", "M", "M", "M", "M", "M", "M", "M",
"M", "…
## $ VIC_AGE_GROUP <chr> "18-24", "25-44", "25-44", "25-44", "25-44", "25-
44", "1…

View(subData)
```

Above output shows the summary of three variables of interest.
1) There are 5 unique observations for victim age group. 2681 victims are less than 18 years, 9604 victims are between 18-24 years, 11386 victims are between 25-44 years, 1698 victims are between 1698 years and 167 victims are older than 65 years. 2) There are 2403 female victims and 23,182 male victims

Visualizations

In this section three visualizations are created. The first visualization depicts the distribution of being victim of shooting for different age groups. The second visualization shows the number of shooting that victimize males versus females. The last visualization shows the number of shootings for different hours in a day.

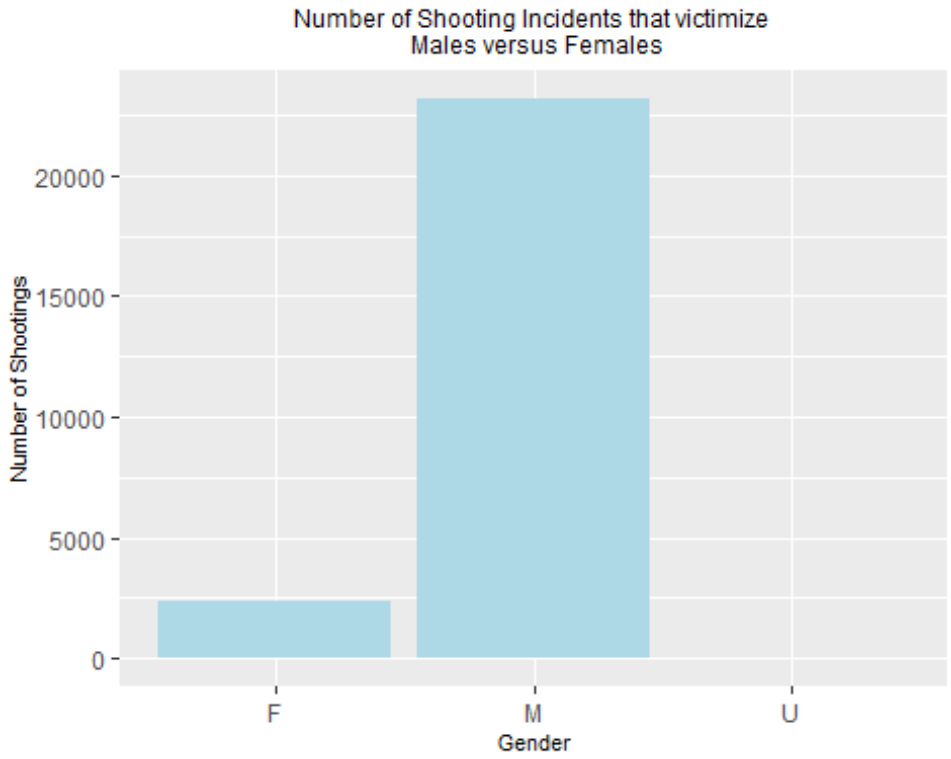
```
#plotting distribution of incidents by the victims'Age Group
ggplot(data=data, aes(x=age))+ geom_bar(fill='lightblue')+
labs(title='Number of Shooting Incidents Victimize Different Age Groups',
x='Age Group', y='Number of Shootings') +
theme(plot.title = element_text(hjust = 0.5,size=10)
, axis.title =element_text(hjust = 0.5,size=9) )
```



The above plot represents the distribution of number of shootings by the victim's Age group. From above plot, it can be seen that most number of shooting incidents occurred in the Age group between 25 and 44 years. Least number of shootings occurred in the Adults of between 45 to 64 years old.

In the second chart you can see how is the number of shooting distributions related to Gender.

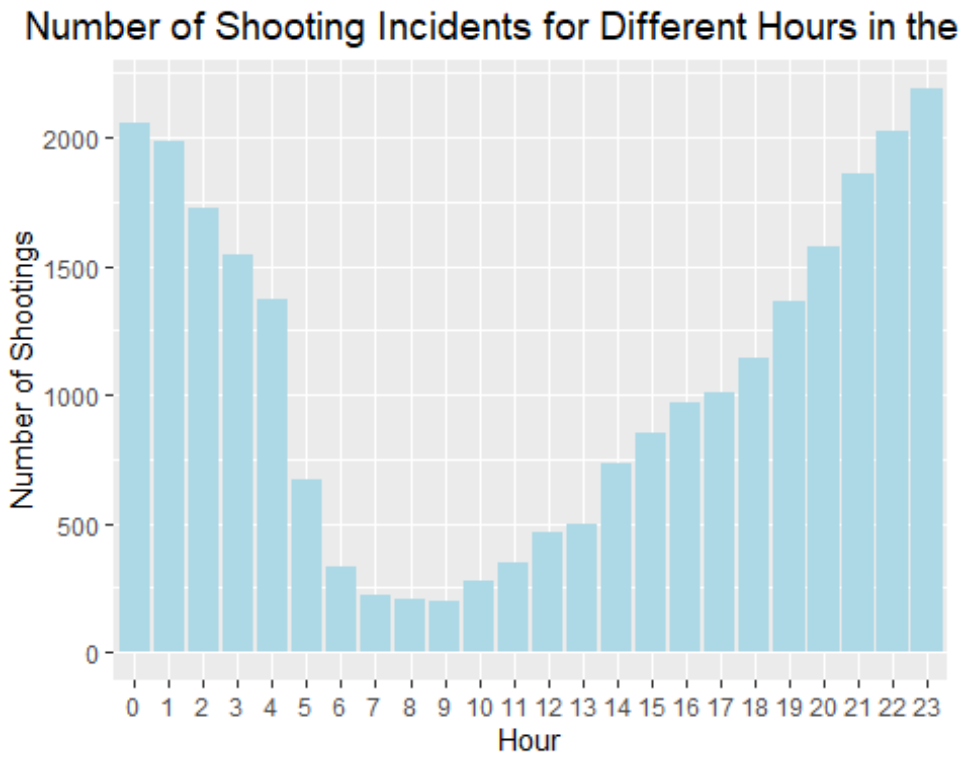
```
##plotting distribution of incidents by Gender
ggplot(data=data, aes(x=gender))+geom_bar(fill='lightblue')+
labs(title='Number of Shooting Incidents that victimize \n Males versus
Females' ,x='Gender'
,y='Number of Shootings') +
theme(plot.title = element_text(hjust = 0.5,size=9)
, axis.title =element_text(hjust = 0.5,size=8))
```



The above plot represents the distribution of shooting incidents by Gender. It shows that vast majority of victims in shooting incidents are Males.

In the third plot you can see how the hour of the day correlates with the number of shootings. The plot clearly shows that during the daylight number of shootings drastically decreases. In majority of the days, the peak hour when we see the maximum number of shooting is in the hour between 11pm and 12am.

```
ggplot(data=data, aes(x=hour))+geom_bar(fill='lightblue')+
labs(title='Number of Shooting Incidents for Different Hours in the Day'
,x='Hour'
,y='Number of Shootings') +
theme(plot.title = element_text(hjust = 0.5,size=14)
, axis.title =element_text(hjust = 0.5,size=11) )
```



Since both variables are categorical, I will use Chi-Squared test of independence for checking is there any association between perpetrators and gender of victims. The null and alternative hypotheses for Chi-Squared test of dependence are given below: H0: There is no association between between perpetrators and gender of victims. Ha: There is a significant association between between perpetrators and gender of victims. The significance level alpha = 0.05.

```
#implementing chi square test
sex <- nypd %>% select(VIC_SEX,PERP_SEX)
chisq.test(table(sex$VIC_SEX,sex$PERP_SEX),simulate.p.value = T)

##
##  Pearson's Chi-squared test with simulated p-value (based on 2000
##  replicates)
##
## data:  table(sex$VIC_SEX, sex$PERP_SEX)
## X-squared = 98.289, df = NA, p-value = 0.0004998

#Pearson's Chi-squared test with simulated p-value (based on 2000 ##
replicates)
```

Conclusion

Above output shows that the p-value is less than significance level alpha = 0.05, so I reject the null hypotheses and conclude that there is a significant association between perpetrators and gender of victims.

Modeling

In this section, a multiple linear regression model is fitted on the sliced data and the result is interpreted

```
data_agg <- data %>% group_by(gender,age,hour ) %>% summarise(count = n())

## `summarise()` has grouped output by 'gender', 'age'. You can override
using the
## `.groups` argument.

lm(count~gender+age+hour,data = data_agg)

##
## Call:
## lm(formula = count ~ gender + age + hour, data = data_agg)
##
## Coefficients:
## (Intercept)      genderM      genderU     age18-24     age25-44     age45-
64
##      51.704      162.751      -76.401      137.642      174.869      -
20.479
##      age65+   ageUNKNOWN        hour1        hour2        hour3
hour4
##     -57.742      -73.774        2.642      -21.909      -31.165      -
47.935
##       hour5        hour6        hour7        hour8        hour9
hour10
##     -93.991     -161.409     -157.851     -142.846     -160.407     -
151.550
##      hour11       hour12       hour13       hour14       hour15
hour16
##    -145.005     -134.369     -131.096     -109.732      -68.736      -
79.165
##      hour17       hour18       hour19       hour20       hour21
hour22
##     -84.550      -72.823      -39.799      -20.071       -7.369       -
2.583
##      hour23
##       8.314
```

According to the result, variable for Males (genderM), 23:00PM-24:00PM (hour23) , age group between 25-44 (age25-44) have highest positive coefficients and this imply these variables are positively correlated with the number of shooting incidents

Conclusion and Potential Source of Bias

In this project, the data for historic shooting incidence in New York was studied. It turned out that age group between 25 to 44 was more exposed to those shooting incidents compared to other groups. In addition, the shooting between 11:00 PM to 12:00 PM is high probable in comparison to other hours in a day. Moreover, Males are much more expected

to be victim of shooting than females. One potential source of bias could be the distribution of different age groups in the place that data was recorded. If one age group has the largest fraction of the population in a certain area, then it is rational for that group to have highest number of shooting victims in the data and this does not necessarily mean that shooters, in general, are more inclined to victimize this age group comparing to other age groups. Also, this data is provided by NYPD and not independent sources which might be another potential source of bias. This of course is not confirmed. But in general, it is a good practice data to be collected by an independent third party.