NYPD Shooting Incident Data Report

Erick Maglalang

2023-01-28

```
library(lubridate)
```

```
## Loading required package: timechange
```

```
##
## Attaching package: 'lubridate'
```

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

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x lubridate::as.difftime() masks base::as.difftime()
## x lubridate::date()        masks base::date()
## x dplyr::filter()          masks stats::filter()
## x lubridate::intersect()   masks base::intersect()
## x dplyr::lag()             masks stats::lag()
## x lubridate::setdiff()     masks base::setdiff()
## x lubridate::union()       masks base::union()
```

```
#reading data from csv file
data <- read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
head(data)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO PRECINCT JURISDICTION_CODE
## 1    236168668 11/11/2021   15:04:00 BROOKLYN       79                 0
## 2    231008085 07/16/2021   22:05:00 BROOKLYN       72                 0
## 3    230717903 07/11/2021   01:09:00 BROOKLYN       79                 0
## 4    237712309 12/11/2021   13:42:00 BROOKLYN       81                 0
## 5    224465521 02/16/2021   20:00:00   QUEENS      113                 0
## 6    228252164 05/15/2021   04:13:00   QUEENS      113                 0
##   LOCATION_DESC STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX
## 1                                 false
## 2                                 false          45-64        M
## 3                                 false            <18        M
## 4                                 false
## 5                                 false
## 6                                  true
##                  PERP_RACE VIC_AGE_GROUP VIC_SEX                 VIC_RACE
## 1                                  18-24       M                    BLACK
## 2 ASIAN / PACIFIC ISLANDER         25-44       M ASIAN / PACIFIC ISLANDER
## 3                    BLACK         25-44       M                    BLACK
## 4                                  25-44       M                    BLACK
## 5                                  25-44       M                    BLACK
## 6                                  25-44       M                    BLACK
##   X_COORD_CD Y_COORD_CD Latitude Longitude
## 1     996313     187499 40.68132 -73.95651
## 2     981845     171118 40.63636 -74.00867
## 3     996546     187436 40.68114 -73.95567
## 4    1001139     192775 40.69579 -73.93910
## 5    1050710     184826 40.67374 -73.76041
## 6    1051329     196646 40.70618 -73.75806
##                                          Lon_Lat
## 1 POINT (-73.95650899099996 40.68131820000008)
## 2 POINT (-74.00866668999998 40.63636384100005)
## 3 POINT (-73.95566903799994 40.68114495900005)
## 4        POINT (-73.939095905 40.69579171600003)
## 5 POINT (-73.76041066999993 40.67374017600008)
## 6 POINT (-73.75806147399999 40.70617856900003)
```

Tidy and Transform

Converted the OCCUR_DATE variable to Date (month, year and day) for further analysis.

```r
#converting OCCUR_DATE to date data type
data <- data%>%
  mutate(OCCUR_DATE = as.Date(OCCUR_DATE, "%m/%d/%y"))
```

Converted all categorical variables into factor data type.

```r
#filtering out observations fro Unknown victim age group
data <- data%>%
  filter(VIC_AGE_GROUP != 'UNKNOWN')
#converting categorical variables to character data type
data <- data%>%
  mutate(BORO = as.factor(BORO),
         PERP_AGE_GROUP = as.factor(PERP_AGE_GROUP),
         PERP_SEX = as.factor(PERP_SEX),
         PERP_RACE = as.factor(PERP_RACE),
         PERP_AGE_GROUP = as.factor(PERP_AGE_GROUP),
         VIC_AGE_GROUP = as.factor(VIC_AGE_GROUP),
         VIC_SEX = as.factor(VIC_SEX),
         VIC_RACE = as.factor(VIC_RACE)
         )
```

Selected OCCUR_DATE, BORO and VIC_AGE_GROUP.

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

```
##    OCCUR_DATE          BORO VIC_AGE_GROUP
##             0             0             0
```

The above output shows that there is no null values in the dataset which means the dataset is already cleaned.

```
#summary of data
summary(subData)
```

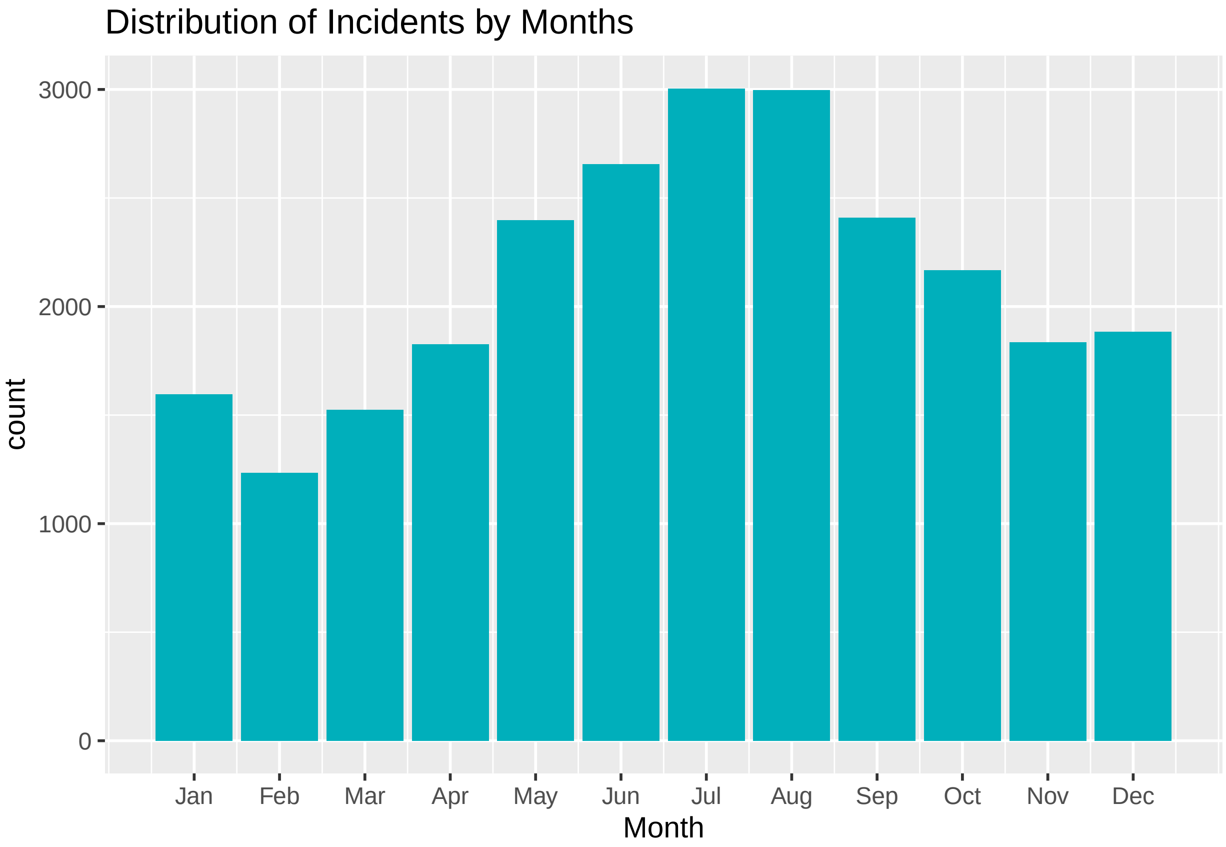
```
##    OCCUR_DATE                    BORO       VIC_AGE_GROUP
##  Min.   :2020-01-01   BRONX        : 7385   <18  : 2681
##  1st Qu.:2020-05-04   BROOKLYN     :10339   18-24: 9604
##  Median :2020-07-15   MANHATTAN    : 3260   25-44:11386
##  Mean   :2020-07-12   QUEENS       : 3817   45-64: 1698
##  3rd Qu.:2020-09-24   STATEN ISLAND:  735   65+  :  167
##  Max.   :2020-12-31
```

Above output shows the summary of three variables of interest. The date of incidents ranges from January 2020 to December 2020. For Boro there are 5 Boroughs in the state of New York. There are 7385 observations for Boro Bronx, 10339 observations for Boro Brooklyn, 3260 observations for Manhattan, and 3817 observations fro Queen and also 735 observations for Staten Island. 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 vitims are between 1698 years and 167 victims are older than 65 years.

Visualizations and Analysis

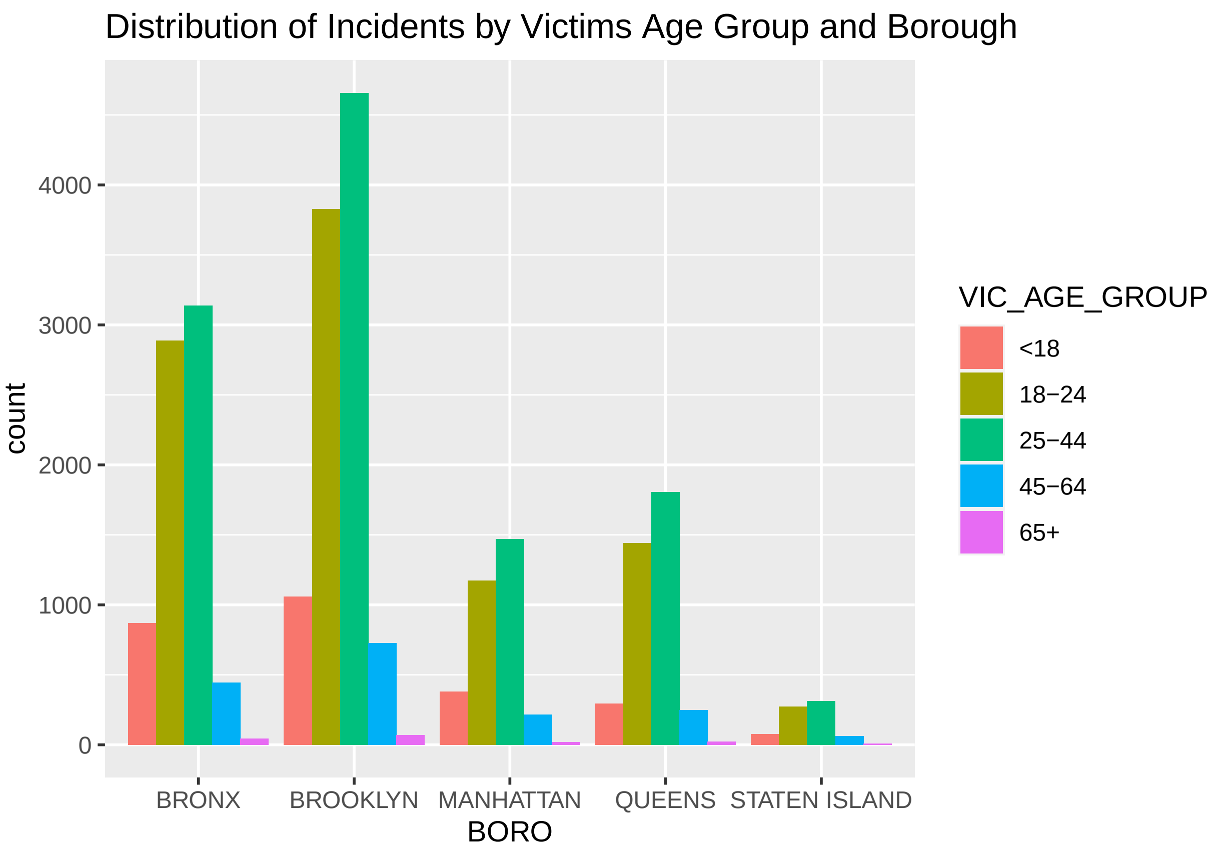
For visualizations and analysis, I will explore how the number of incidents varied over 12 months or a year and I also check whether there is an association between Borough and Victims' Age group, i.e is there any boroughs in which most incidents belong to particular age group of victims or not.

```
#creating new variable month
subData$Month <- month(subData$OCCUR_DATE)
#plotting number of incidents by month
ggplot(subData, aes(x = Month)) +
  geom_bar(fill = "#00AFBB") +
  scale_x_continuous(breaks = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12),
  labels = c("Jan", "Feb", "Mar","Apr","May","Jun", "Jul", "Aug", "Sep", "Oct", "Nov", "Dec")) +
  labs(title = "Distribution of Incidents by Months")
```



The plot represents the distribution of shooting incidents by month. It can be seen that most number of shooting incidents occurred in July followed by August, June and then September. It also shows that the least number of incidents occurred in February, followed by March and January. Next I want to analyze whether there is an association between Borough and Victims age group or not. The grouped bar chart for Distribution of Incidents by Borough and Victims is shown below:

```r
#plotting distribution of incidents by Victims' Age Group and Borough
ggplot(subData, aes(x =BORO, fill = VIC_AGE_GROUP)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of Incidents by Victims Age Group and Borough")
```



The plot represents the distribution of shooting incidents by Borough and Victims' Age Group. It shows that in almost all Boroughs, the most number of victims in shooting incidents are between the age of 25-44 followed by victims with age group of 18-24. The least number of victims in shooting incidents are older than 65 years followed by 45-64. The results show that distribution may be the same. However further analysis is needed before making any conclusions. Since both variables are categorical, I will use Chi-Squared test of independence for checking if there is any association between Borough and Victim Age group. The null and alternative hypotheses for Chi-Squared test of dependence are given below: H0: There is no association between Borough And Victim Age Group. Ha: There is a significant relation between Borough and Victim Age group. The significance level alpha = 0.05.

```
#implementing chi square test
chi <- chisq.test(subData$BORO, subData$VIC_AGE_GROUP)
chi$observed
```

```
##                  subData$VIC_AGE_GROUP
## subData$BORO      <18 18-24 25-44 45-64  65+
##   BRONX           869  2888  3139   445   44
##   BROOKLYN       1060  3826  4658   727   68
##   MANHATTAN       381  1173  1469   216   21
##   QUEENS          296  1443  1806   247   25
##   STATEN ISLAND    75   274   314    63    9
```

```
chi$expected
```

```
##                  subData$VIC_AGE_GROUP
## subData$BORO           <18     18-24    25-44     45-64       65+
##   BRONX          775.34402 2777.4726 3292.826 491.06086 48.296327
##   BROOKLYN      1085.48163 3888.4616 4609.957 687.48520 67.614857
##   MANHATTAN      342.26425 1226.0746 1453.570 216.77162 21.319706
##   QUEENS         400.74315 1435.5603 1701.925 253.80898 24.962367
##   STATEN ISLAND   77.16694  276.4309  327.722  48.87336  4.806743
```

```
chi
```

```
## 
##  Pearson’s Chi-squared test
## 
## data:  subData$BORO and subData$VIC_AGE_GROUP
## X-squared = 81.189, df = 16, p-value = 1.015e-10
```

The p-value is less than significance level alpha = 0.05, therefore I can reject the null hypotheses and conclude that there is a significant relation between Borough and Victims Age group. The observed and expected values for each category are also shown above. It can be concluded that there is a significant difference between the observed values as compared to the expected values. Therefore, significant correlation between Borough and Victims age group exists which means that Victim age groups do vary by Boroughs.

Bias Identification

Since this data is provided by NYPD which has been accused of racial bias and unfair treatment towards minority. There may be some biases as the data is not collected by independent sources. But anything about bias is not 100% confirmed and can not be validated, as I don’t know about the inner activities of New York Police Department. I analyzed and made conclusions based on whatever data we have and tried to avoid making any false conclusion about any type of bias in this dataset.