NYPD Shooting Incidence Report

N. A. Sackey

06/10/2021

Introduction

This report is concerned with the victims of shooting incidents in New York City especially with regard to the age group, sex and race of the victims. It is mainly focused on determining whether there are any groups of people who are most often the victims of shooting incidents and who they might be.

The dataset used is a list of every incidence of shooting in New York City from 2006 to 2020 and records a variety of information regarding the incident such as the date, time, location, precinct as well as demographic information about both the perpetrator and the victim.

Tidying and Transfroming the Data

```r
# Downloads and reads in the dataset
data_url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_data <- read.csv(data_url)
```

Below is a summary of the data after it has been imported into Rstudio.

```r
summary(nypd_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:23568       Length:23568       Length:23568
##  1st Qu.: 55317014   Class :character   Class :character   Class :character
##  Median : 83365370   Mode  :character   Mode  :character   Mode  :character
##  Mean   :102218616
##  3rd Qu.:150772442
##  Max.   :222473262
##
##     PRECINCT      JURISDICTION_CODE LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Min.   :  1.00   Min.   :0.0000    Length:23568       Length:23568
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   Class :character
##  Median : 69.00   Median :0.0000    Mode  :character   Mode  :character
##  Mean   : 66.21   Mean   :0.3323
##  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:23568       Length:23568       Length:23568       Length:23568
##  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:23568       Length:23568       Length:23568       Length:23568
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     Latitude       Longitude        Lon_Lat
##  Min.   :40.51   Min.   :-74.25   Length:23568
##  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
##
```

From the summary, it is clear that there are many unneeded columns such as 'Latitude', 'Longitude', 'Lon_Lat' etc. hence there is a need to clean up the dataset by getting rid of these unnecessary columns.

```
# Keep only the columns needed
nypd_data %>% select(OCCUR_DATE:VIC_RACE) %>% select(-LOCATION_DESC) %>% select(-JURISDICTION_CODE) ->
```

After the unneeded columns have been removed, the next step is to change the data types of variables to the appropriate data type, namely, the factor and date types.

```
# Change the data types for the appropriate data type
nypd_data %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE)) -> nypd_data
nypd_data$BORO <- as.factor(nypd_data$BORO)
nypd_data$PRECINCT <- as.factor(nypd_data$PRECINCT)
nypd_data$STATISTICAL_MURDER_FLAG <- as.factor(nypd_data$STATISTICAL_MURDER_FLAG)
nypd_data$PERP_AGE_GROUP <- as.factor(nypd_data$PERP_AGE_GROUP)
nypd_data$PERP_SEX <- as.factor(nypd_data$PERP_SEX)
nypd_data$PERP_RACE <-as.factor(nypd_data$PERP_RACE)
nypd_data$VIC_AGE_GROUP <- as.factor(nypd_data$VIC_AGE_GROUP)
nypd_data$VIC_SEX <- as.factor(nypd_data$VIC_SEX)
nypd_data$VIC_RACE <- as.factor(nypd_data$VIC_RACE)
```

Next, the data frame is checked to ensure that there are no problems with the data after transforming the dataset such that the variables now have their appropriate data types.

```
summary(nypd_data)
```

```
##    OCCUR_DATE          OCCUR_TIME                     BORO         PRECINCT
##  Min.   :2006-01-01   Length:23568       BRONX        :6700   75     : 1367
##  1st Qu.:2008-12-30   Class :character   BROOKLYN     :9722   73     : 1282
##  Median :2012-02-26   Mode  :character   MANHATTAN    :2921   67     : 1102
##  Mean   :2012-10-03                      QUEENS       :3527   79     :  920
##  3rd Qu.:2016-02-28                      STATEN ISLAND: 698   44     :  842
##  Max.   :2020-12-31                                           47     :  815
##                                                               (Other):17240
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX          PERP_RACE
##  false:19080                    :8459    : 8425   BLACK         :9855
##  true : 4488             18-24  :5448   F:  334                 :8425
##                          25-44  :4613   M:13305   WHITE HISPANIC:1961
##                          UNKNOWN:3156   U: 1504   UNKNOWN       :1869
##                          <18    :1354             BLACK HISPANIC:1081
##                          45-64  : 481             WHITE         : 255
##                          (Other):  57             (Other)       : 122
##  VIC_AGE_GROUP   VIC_SEX                             VIC_RACE
##  <18    : 2525   F: 2195   AMERICAN INDIAN/ALASKAN NATIVE:    9
##  18-24  : 9000   M:21353   ASIAN / PACIFIC ISLANDER      :  320
##  25-44  :10287   U:   20   BLACK                         :16846
##  45-64  : 1536             BLACK HISPANIC                : 2244
##  65+    :  155             UNKNOWN                       :  102
##  UNKNOWN:   65             WHITE                         :  615
##                            WHITE HISPANIC                : 3432
```

From the summary above, some columns with missing data are noticed. These columns include 'PERP_AGE_GROUP', 'PERP_SEX' and 'PERP_RACE'. However, all these columns have a value that denotes unknown data therefore the missing data will be replaced with that value for those columns i.e. either "UNKNOWN" or "U".

```
# Replace the missing data values with 'UNKNOWN' or 'U'
nypd_data$PERP_AGE_GROUP[nypd_data$PERP_AGE_GROUP == ""] <- "UNKNOWN"
nypd_data$PERP_SEX[nypd_data$PERP_SEX == ""] <- "U"
nypd_data$PERP_RACE[nypd_data$PERP_RACE == ""] <- "UNKNOWN"
```

With this, the summary now shows no missing data for any of the rows and thus the analysis can proceed.

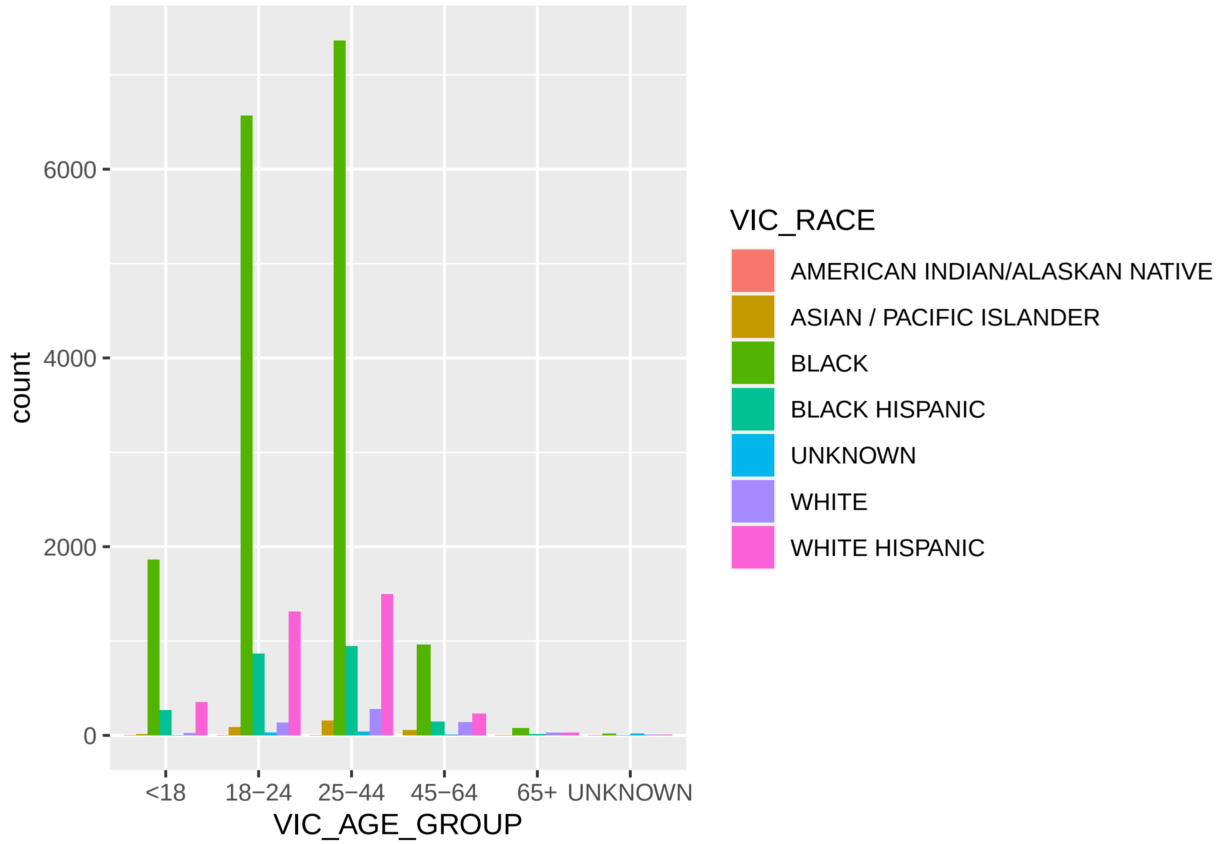
```
# Display a summary of the transformed data
summary(nypd_data)
```

```
##    OCCUR_DATE          OCCUR_TIME                    BORO         PRECINCT
##  Min.   :2006-01-01   Length:23568       BRONX        :6700   75     : 1367
##  1st Qu.:2008-12-30   Class :character   BROOKLYN     :9722   73     : 1282
##  Median :2012-02-26   Mode  :character   MANHATTAN    :2921   67     : 1102
##  Mean   :2012-10-03                      QUEENS       :3527   79     :  920
##  3rd Qu.:2016-02-28                      STATEN ISLAND: 698   44     :  842
##  Max.   :2020-12-31                                           47     :  815
##                                                               (Other):17240
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP  PERP_SEX
##  false:19080             UNKNOWN:11615    :    0
##  true : 4488             18-24  : 5448   F:  334
##                          25-44  : 4613   M:13305
##                          <18    : 1354   U: 9929
##                          45-64  :  481
##                          65+    :   54
##                          (Other):    3
##                        PERP_RACE      VIC_AGE_GROUP    VIC_SEX
##  UNKNOWN                 :10294   <18    : 2525   F: 2195  
##  BLACK                   : 9855   18-24  : 9000   M:21353  
##  WHITE HISPANIC          : 1961   25-44  :10287   U:   20  
##  BLACK HISPANIC          : 1081   45-64  : 1536            
##  WHITE                   :  255   65+    :  155            
##  ASIAN / PACIFIC ISLANDER:  120   UNKNOWN:   65            
##  (Other)                 :    2                            
##                             VIC_RACE    
##  AMERICAN INDIAN/ALASKAN NATIVE:    9  
##  ASIAN / PACIFIC ISLANDER      :  320  
##  BLACK                         :16846  
##  BLACK HISPANIC                : 2244  
##  UNKNOWN                       :  102  
##  WHITE                         :  615  
##  WHITE HISPANIC                : 3432  
```

Analysis

The focus of this analysis will be the victims of shooting incidents in New York. The first visualization would be a grouped bar chart showing the number of shooting incidents against age group and race of the victim.

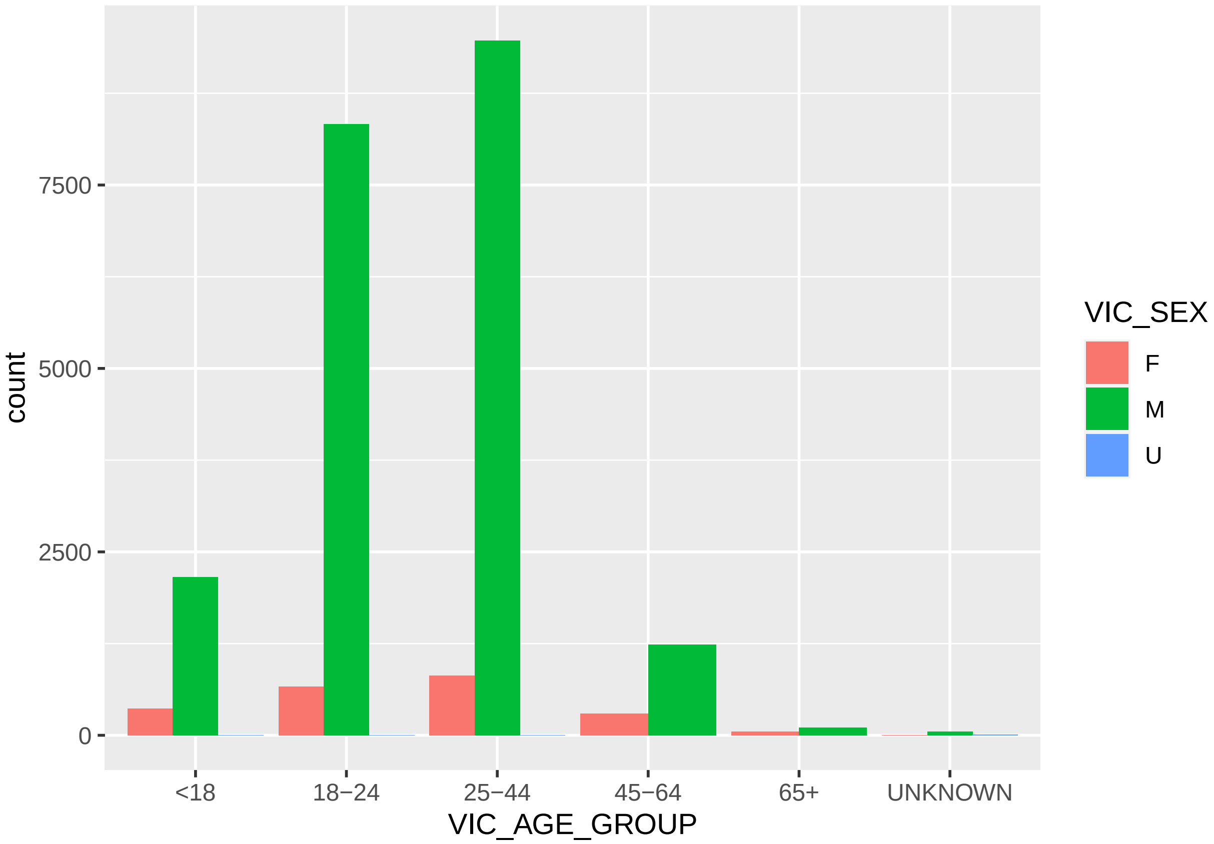
```
# Plot a bar chart of number of incidents vs age and race
ggplot(nypd_data, aes(x = VIC_AGE_GROUP, fill = VIC_RACE)) + geom_bar(position="dodge")
```



This bar chart shows that the most victims of shooting incidents are black victims aged 25-44. The next highest are white Hispanic victims while the least appears to be American Indian/Alaskan Natives.

The second visualization is another group bar chart but one showing the number of shooting incidents against the sex and age group of the victims.

```
# Plot a bar chart of number of incidents vs age and sex
ggplot(nypd_data, aes(x = VIC_AGE_GROUP, fill = VIC_SEX)) + geom_bar(position="dodge")
```



The bar chart above shows that males aged 25-44 are the most common victims of shooting incidents. This analysis does raise many questions particularly with regard to the unknown data and whether there may be additional variables and factors which could be considered. For example, population data or the number of people of a specific race or age group who reside within New York City or even the number of males as opposed to the female residents of the city.

Conclusion

The analysis carried out shows that black males aged 25-44 are most often the victims of shooting incidents in New York City. The main source of personal bias would be the choice of topic and data as the decision to choose to analyze the victims was motivated by personal curiosity and interest in which demographic was more affected by shootings in New York City. One way of attempting to mitigate bias would be the inclusion of the data observations with missing data when cleaning the data. By retaining the measurements which held missing values instead of discarding them, any aspect of exclusion bias would hopefully be mitigated.

Appendix

```
# Provide info about R Session
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.3 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0
##
## locale:
##  [1] LC_CTYPE=en_CA.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_CA.UTF-8        LC_COLLATE=en_CA.UTF-8
##  [5] LC_MONETARY=en_CA.UTF-8    LC_MESSAGES=en_CA.UTF-8
##  [7] LC_PAPER=en_CA.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_CA.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.8.0 forcats_0.5.1   stringr_1.4.0   dplyr_1.0.7
##  [5] purrr_0.3.4     readr_2.0.2     tidyr_1.1.4     tibble_3.1.5
##  [9] ggplot2_3.3.5   tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.1 xfun_0.26        haven_2.4.3      colorspace_2.0-2
##  [5] vctrs_0.3.8      generics_0.1.0   htmltools_0.5.2  yaml_2.2.1
##  [9] utf8_1.2.2       rlang_0.4.11     pillar_1.6.3     glue_1.4.2
## [13] withr_2.4.2      DBI_1.1.1        dbplyr_2.1.1     modelr_0.1.8
## [17] readxl_1.3.1     lifecycle_1.0.1  munsell_0.5.0    gtable_0.3.0
## [21] cellranger_1.1.0 rvest_1.0.1      evaluate_0.14    labeling_0.4.2
## [25] knitr_1.36       tzdb_0.1.2       fastmap_1.1.0    fansi_0.5.0
## [29] highr_0.9        broom_0.7.9      Rcpp_1.0.7       scales_1.1.1
## [33] backports_1.2.1  jsonlite_1.7.2   farver_2.1.0     fs_1.5.0
## [37] hms_1.1.1        digest_0.6.28    stringi_1.7.5    grid_4.1.1
## [41] cli_3.0.1        tools_4.1.1      magrittr_2.0.1   crayon_1.4.1
## [45] pkgconfig_2.0.3  ellipsis_0.3.2   xml2_1.3.2       reprex_2.0.1
## [49] rstudioapi_0.13  assertthat_0.2.1 rmarkdown_2.11   httr_1.4.2
## [53] R6_2.5.1         compiler_4.1.1
```