NYPD Shooting Incident Data (Historic)

Chengxiao Yu

6/7/2023

Analysis of NYPD Shooting Incident Data (Historic)

First, data was loaded from "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD". This is a breakdown of every shooting incident that occurred in NYC going back to 2006 through the end of the previous calendar year. The goal of the analysis is to identify the frequency of shootings in 5 boroughs of NYC during each year and its breakdown percentages in each race.

The data were tidied by removing variables about the specific locations of the incidents, since they are not needed for my objective in this analysis. The 'OCCUR_DATE' was changed to 'date' format.

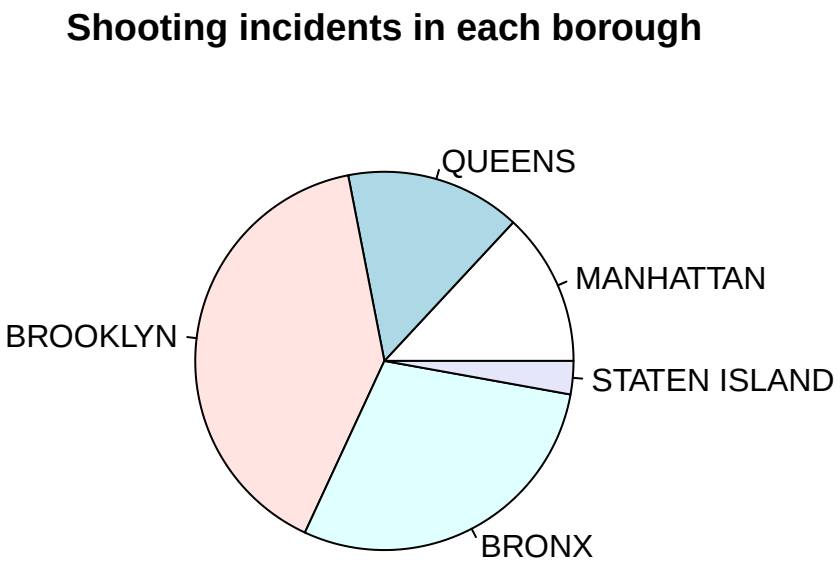
```
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
data <- read.csv(url)
data <- data %>% select(-c(X_COORD_CD,Y_COORD_CD,Lon_Lat,INCIDENT_KEY,Latitude, Longitude))
data$OCCUR_DATE <- as.Date(data$OCCUR_DATE, format = "%m/%d/%Y")
summary(data)
```

```
##    OCCUR_DATE          OCCUR_TIME            BORO           LOC_OF_OCCUR_DESC
##  Min.   :2006-01-01   Length:27312       Length:27312       Length:27312
##  1st Qu.:2009-07-18   Class :character   Class :character   Class :character
##  Median :2013-04-29   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2014-01-06
##  3rd Qu.:2018-10-15
##  Max.   :2022-12-31
##
##     PRECINCT      JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC
##  Min.   :  1.00   Min.   :0.0000    Length:27312       Length:27312
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   Class :character
##  Median : 68.00   Median :0.0000    Mode  :character   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
##  STATISTICAL_MURDER_FLAG PERP_AGE_GROUP       PERP_SEX
##  Length:27312            Length:27312       Length:27312
##  Class :character        Class :character   Class :character
##  Mode  :character        Mode  :character   Mode  :character
##
##
##
##
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE
```

```
##  Length:27312       Length:27312       Length:27312       Length:27312
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

Analysis of shooting incidents in five boroughs

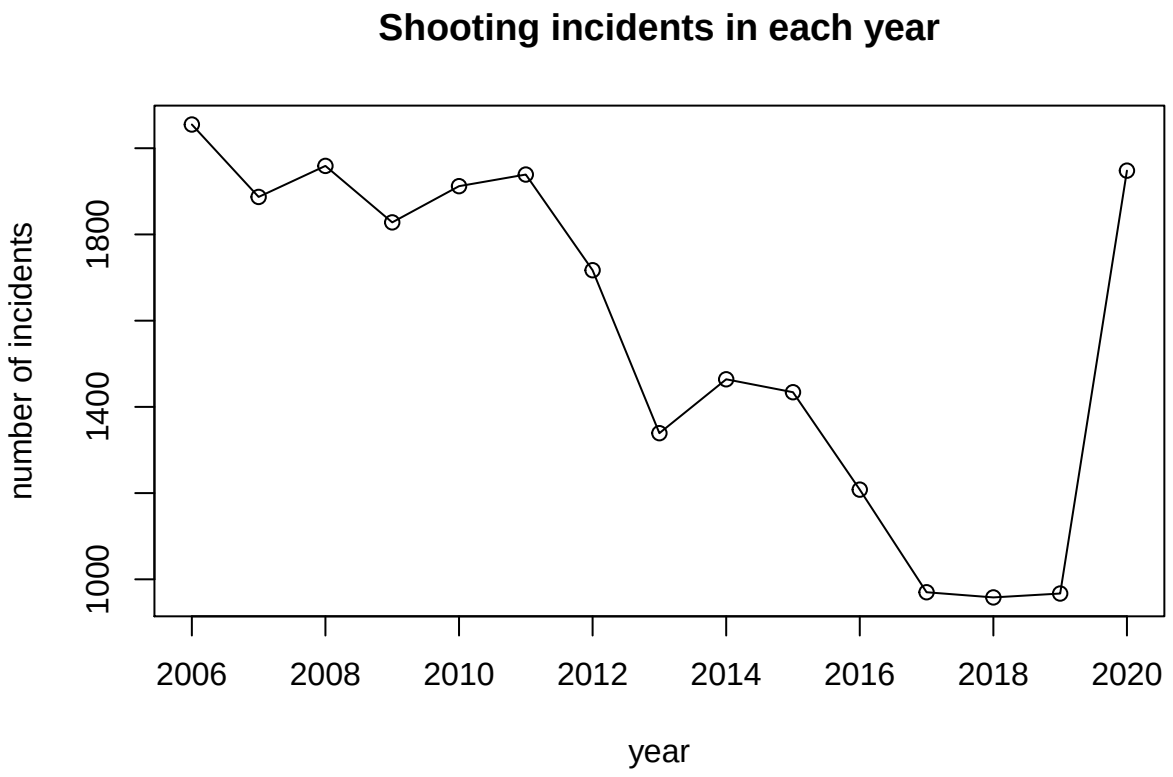
```
boro <- c('MANHATTAN','QUEENS','BROOKLYN','BRONX','STATEN ISLAND')
boro.p <- c()
for (p in boro){
  boro.p <- c(boro.p,mean(str_count(data$BORO,p)))
}
pie(boro.p, boro, main="Shooting incidents in each borough")
```



From, this pie plot, we can see that historically shootings happened in Brooklyn most frequently and Staten Island has the least number of incidents. ### Analysis of shooting incidents in each year

```
date <- data$OCCUR_DATE
year <- as.numeric(format(date,format="%Y"))
x <- seq(from=2006,to=2020,length=15)
y <- rep(NA,length(x))
for (i in 1:length(y)){
```

```
  y[i] <- sum(year==x[i])
}
plot(x,y,type='l',pch=19,main='Shooting incidents in each year',
     ylab = 'number of incidents',
     xlab = 'year')
points(x,y)
```



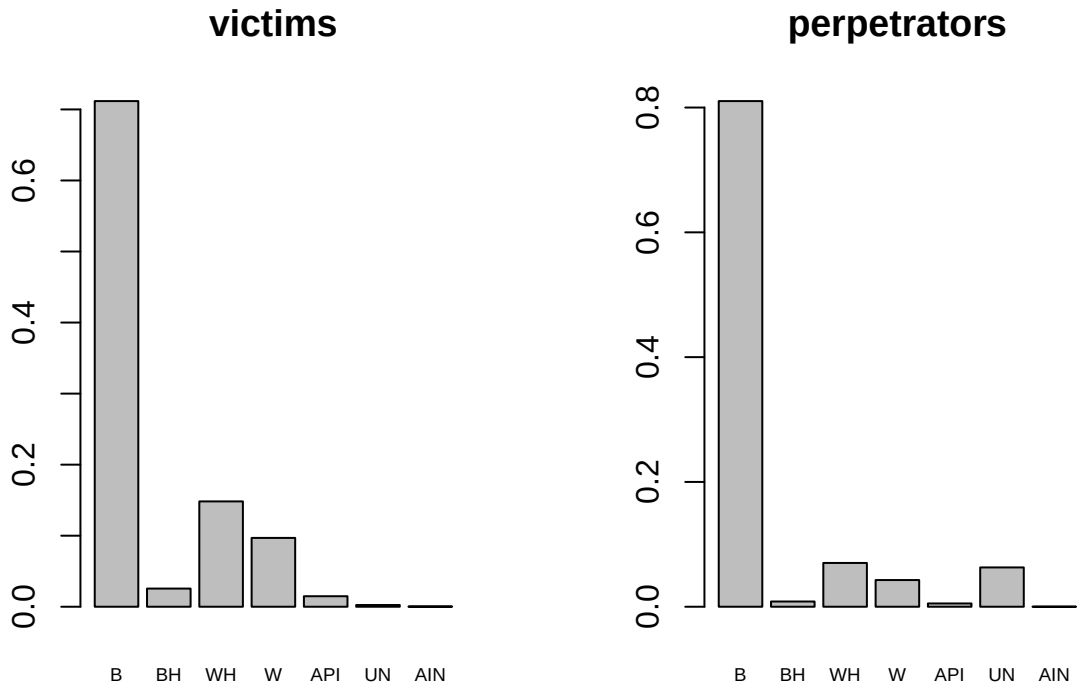
From this histogram, we can see that in recent 15 years, 2006 had the most incidents. The number of shootings dropped sharply in 2007 and decreased slowly over time until 2020 when there was a significant increase.

Victim and perpetrator by race

```
race <- data$VIC_RACE[!duplicated(data$VIC_RACE)]
vic_race <- c()
perp_race <- c()
for (r in race){
  vic_race <- c(vic_race,mean(data$VIC_RACE==r))
  perp_race <- c(perp_race,mean(data$PERP_RACE==r))
  perp_race <- perp_race/sum(perp_race)
}

par(mfrow=c(1,2))
lab <- c("B","BH","WH","W","API","UN","AIN")
```

```r
barplot(height=vic_race,names=lab,cex.names=0.6,main = "victims")
barplot(height=perp_race,names=lab,cex.names=0.6,main = "perpetrators")
```

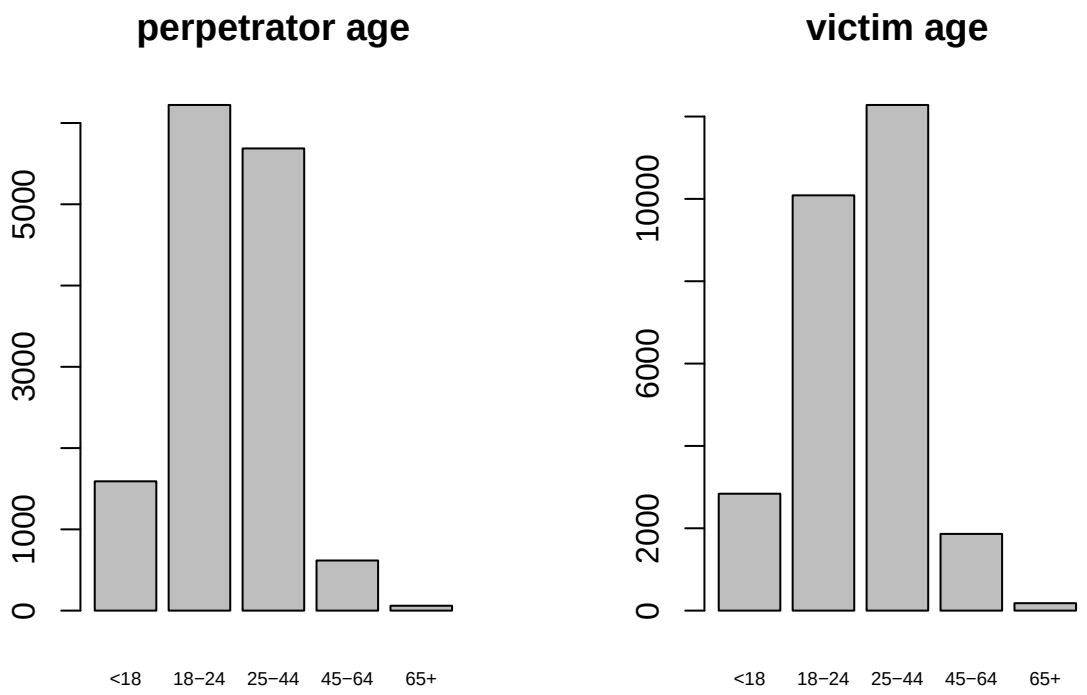


It's shown from the figure that around 70% of victims and 80% of perpetrators are black.

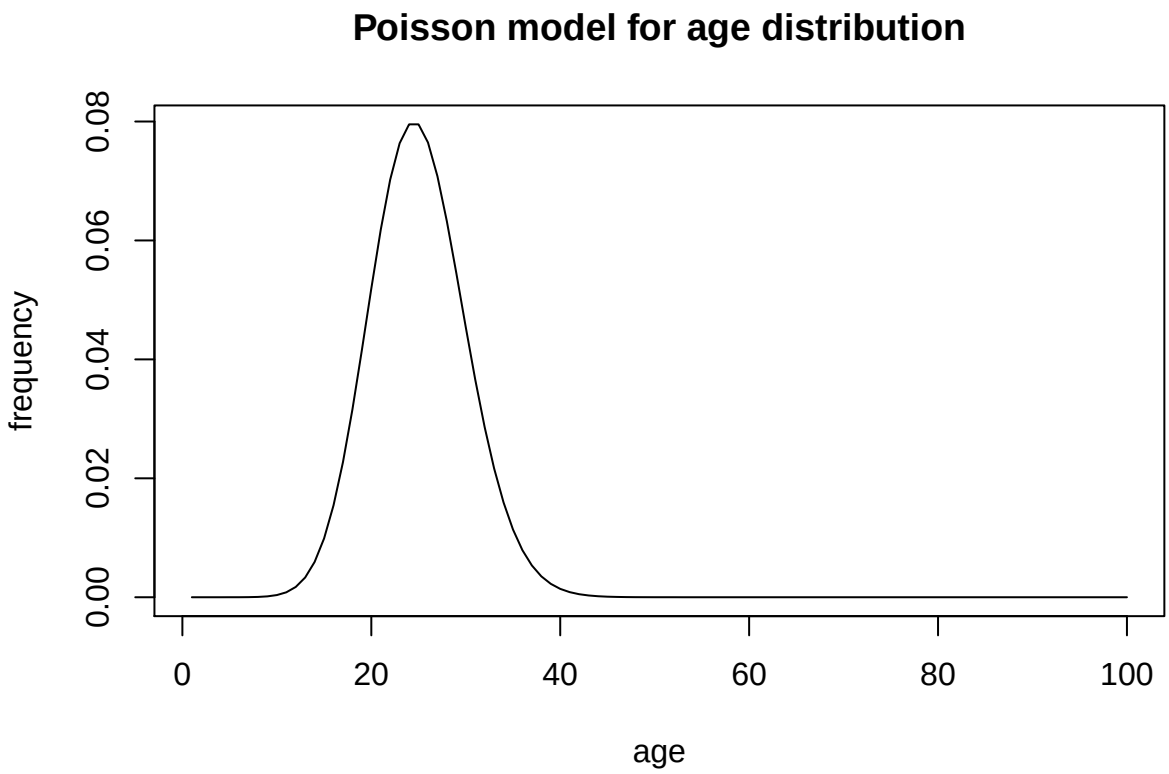
Victim and perpetrator by age

```r
age_group <- c("<18", "18-24", "25-44", "45-64", "65+", "UNKNOWN")
vic_age_count <- matrix(0,1,6)
for (i in data$VIC_AGE_GROUP){
  for (j in 1:length(age_group)){
    if(i==age_group[j]){
      vic_age_count[j] <- vic_age_count[j] + 1
    }
  }
}
perp_age_count <- matrix(0,1,6)
for (i in data$PERP_AGE_GROUP){
  for (j in 1:length(age_group)){
    if(i==age_group[j]){
      perp_age_count[j] <- perp_age_count[j] + 1
    }
  }
}
```

```
par(mfrow=c(1,2))
barplot(perp_age_count[1:5],names=age_group[1:5],cex.names = 0.6,main = "perpetrator age")
barplot(vic_age_count[1:5],names=age_group[1:5],cex.names = 0.6,main = "victim age")
```




```
par(mfrow=c(1,1))
plot(1:100,dpois(1:100,25),type='l',main='Poisson model for age distribution',xlab = 'age',ylab = 'freq
```



The most frequent age group for perpetrators is 18-24 and the most frequent age group for victims is 25-44. If given the accurate ages, they can be modeled as a Poisson distribution, Poisson(theta), where theta is the mean of the group.

Possible bias in the analysis

First, the data itself can be biased. Some variables in the data are "NA", which might decrease or increase the effect of a specific group. Second, the choice I made was biased. I chose these aspects to analyze because I have some prior belief that there should be some trends inside the data. For example, I expect that the number of shootings in 2020 would increase because of the COVID-19 and the election. More data analysis should be done to demonstrate the trend is indeed true.

Summary

Based on the historical data on shooting incidents of NYC, it's shown that Brooklyn has the most cases. The next step could be to search and analyze data on why this is the case. There could be some demographic or socioeconomic cause for this phenomenon. It's also noticed that black people are significantly overrepresented in both victims and perpetrators groups. We may ask the question that what is the underlying reason for this phenomenon? This could be what we focus on next.

```
sessionInfo()
```

```
## R version 4.0.5 (2021-03-31)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
```

```
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] forcats_0.5.1   stringr_1.4.0   dplyr_1.0.6     purrr_0.3.4
## [5] readr_1.4.0     tidyr_1.1.3     tibble_3.1.2    ggplot2_3.3.6
## [9] tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.1 xfun_0.39        haven_2.4.1      colorspace_2.0-1
##  [5] vctrs_0.3.8      generics_0.1.0   htmltools_0.5.3  yaml_2.2.1
##  [9] utf8_1.2.1       rlang_1.1.1      pillar_1.6.1     glue_1.4.2
## [13] withr_2.5.0      DBI_1.1.1        dbplyr_2.1.1     modelr_0.1.8
## [17] readxl_1.3.1     lifecycle_1.0.2  munsell_0.5.0    gtable_0.3.0
## [21] cellranger_1.1.0 rvest_1.0.0      evaluate_0.14    knitr_1.33
## [25] fastmap_1.1.0    fansi_0.5.0      highr_0.9        broom_1.0.1
## [29] Rcpp_1.0.9       backports_1.2.1  scales_1.1.1     jsonlite_1.7.2
## [33] fs_1.5.2         hms_1.1.0        digest_0.6.27    stringi_1.6.2
## [37] grid_4.0.5       cli_3.4.0        tools_4.0.5      magrittr_2.0.1
## [41] crayon_1.4.1     pkgconfig_2.0.3  ellipsis_0.3.2   xml2_1.3.2
## [45] reprex_2.0.0     lubridate_1.7.10 assertthat_0.2.1 rmarkdown_2.21
## [49] httr_1.4.2       rstudioapi_0.13  R6_2.5.0         compiler_4.0.5
```