NYPD Shooting Incident Data Report

A.G.

2024-10-07

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(lubridate)
library(ggrepel)
```

```
## Warning: package 'ggrepel' was built under R version 4.3.3
```

```
### Get the NYPD data from the website
url_in_nypd <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"

nypd <- read_csv(url_in_nypd)
```

```
## Rows: 28562 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.
```

The data is imported from the appropriate site and read into the R markdown.

```
nypd <- subset(nypd, select = -c(INCIDENT_KEY, OCCUR_TIME,
          LOC_OF_OCCUR_DESC:STATISTICAL_MURDER_FLAG, X_COORD_CD:Lon_Lat))

nypd <- nypd %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))
summary(nypd)
```

```
##    OCCUR_DATE             BORO           PERP_AGE_GROUP       PERP_SEX
##  Min.   :2006-01-01   Length:28562       Length:28562       Length:28562
##  1st Qu.:2009-09-04   Class :character   Class :character   Class :character
##  Median :2013-09-20   Mode  :character   Mode  :character   Mode  :character
##  Mean   :2014-06-07
##  3rd Qu.:2019-09-29
##  Max.   :2023-12-29
##   PERP_RACE         VIC_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

The data set is then tidied. I got rid of columns that are not important to my analysis and changed certain variables into appropriate data types.

```
perp_freq <- table(nypd$PERP_AGE_GROUP)
perp_freq <- as.data.frame(perp_freq)
rows_to_remove <- c(1,3,4,6,10)
perp_freq <- perp_freq[-rows_to_remove, ]
colnames(perp_freq) <- c("Age_Group", "Incidence")
perp_freq$Percentage <- round(perp_freq$Incidence/sum(perp_freq$Incidence)*100, 1)
perp_freq
```

```
##    Age_Group Incidence Percentage
## 2        <18      1682        9.3
## 5      18-24      6438       35.6
## 7      25-44      6041       33.4
## 8      45-64       699        3.9
## 9        65+        65        0.4
## 11   UNKNOWN      3148       17.4
```

```
vic_freq <- table(nypd$VIC_AGE_GROUP)
vic_freq <- as.data.frame(vic_freq)
vic_freq <- vic_freq[-2, ]
colnames(vic_freq) <- c("Age_Group", "Incidence")
vic_freq$Percentage <- round(vic_freq$Incidence/sum(vic_freq$Incidence)*100, 1)
vic_freq
```

```
##   Age_Group Incidence Percentage
## 1       <18      2954       10.3
## 3     18-24     10384       36.4
## 4     25-44     12973       45.4
## 5     45-64      1981        6.9
## 6       65+       205        0.7
## 7   UNKNOWN        64        0.2
```

```
nypd_2020 <- nypd %>%
  filter(format(OCCUR_DATE, "%Y") == "2020")
ny_borough <- table(nypd_2020$BORO)
```

```r
ny_borough <- as.data.frame(ny_borough)
colnames(ny_borough) <- c("NYC Borough", "Incidence")
ny_borough$Percent_of_Incidence <- round(ny_borough$Incidence/sum(ny_borough$Incidence)*100, 1)
ny_borough
```

```
##     NYC Borough Incidence Percent_of_Incidence
## 1         BRONX       504                 25.9
## 2      BROOKLYN       819                 42.0
## 3     MANHATTAN       272                 14.0
## 4        QUEENS       303                 15.6
## 5 STATEN ISLAND        50                  2.6
```

```r
nyc_borough_pop <- data.frame(
    Borough = c("BRONX", "BROOKLYN", "MANHATTAN", "QUEENS", "STATEN ISLAND"),
    Population = c(1472654, 2736074, 1694251, 2405464, 495747),
    stringsAsFactors = FALSE
)

nyc_borough_pop$Percent_of_nyc <- round((nyc_borough_pop$Population/sum(
  nyc_borough_pop$Population))*100,1)
nyc_borough_pop
```

```
##         Borough Population Percent_of_nyc
## 1         BRONX    1472654           16.7
## 2      BROOKLYN    2736074           31.1
## 3     MANHATTAN    1694251           19.2
## 4        QUEENS    2405464           27.3
## 5 STATEN ISLAND     495747            5.6
```

```r
ny_borough$Percent_of_nyc <- nyc_borough_pop$Percent_of_nyc
ny_borough
```

```
##     NYC Borough Incidence Percent_of_Incidence Percent_of_nyc
## 1         BRONX       504                 25.9           16.7
## 2      BROOKLYN       819                 42.0           31.1
## 3     MANHATTAN       272                 14.0           19.2
## 4        QUEENS       303                 15.6           27.3
## 5 STATEN ISLAND        50                  2.6            5.6
```

```r
correlation <- cor(ny_borough$Percent_of_Incidence,ny_borough$Percent_of_nyc)

ny_regression <- lm(Percent_of_nyc ~ Percent_of_Incidence, data = ny_borough)
summary(ny_regression)
```

```
## 
## Call:
## lm(formula = Percent_of_nyc ~ Percent_of_Incidence, data = ny_borough)
## 
## Residuals:
##       1       2       3       4       5 
## -6.2838 -0.1083  2.2953  9.5779 -5.4811
```

```
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             9.7529     6.0395   1.615    0.205
## Percent_of_Incidence    0.5108     0.2516   2.030    0.135
##
## Residual standard error: 7.451 on 3 degrees of freedom
## Multiple R-squared:  0.5788, Adjusted R-squared:  0.4384
## F-statistic: 4.122 on 1 and 3 DF,  p-value: 0.1353
```

In this analysis, I looked at the frequencies of perpetrators by age group as well as the frequencies of victims by age group. There was some missing data that was removed, as well as few data points that didn't make sense in this context. Lastly, I looked at the incidences by the NYC boroughs. A further analysis/next step would be to then compare the relative population of each borough to their percentages of shootings, to see whether any borough proportionally has more shootings. Looking at just the number of shootings in each borough by themselves can be misleading, without comparing their populations.

I pulled that data from the US Census Bureau and added a column that showed what percent of the city each borough made up with their populations. That way, one could compare the incidences to its population proportion by borough.

Populations of each boroughs came from the 2020 census data:

Bronx - "https://data.census.gov/profile/Bronx_borough,_Bronx_County,_New_York?g=060XX00US3600508510"

Brooklyn - "https://data.census.gov/profile/Brooklyn_borough,_Kings_County,_New_York?g=060XX00US3604710022"

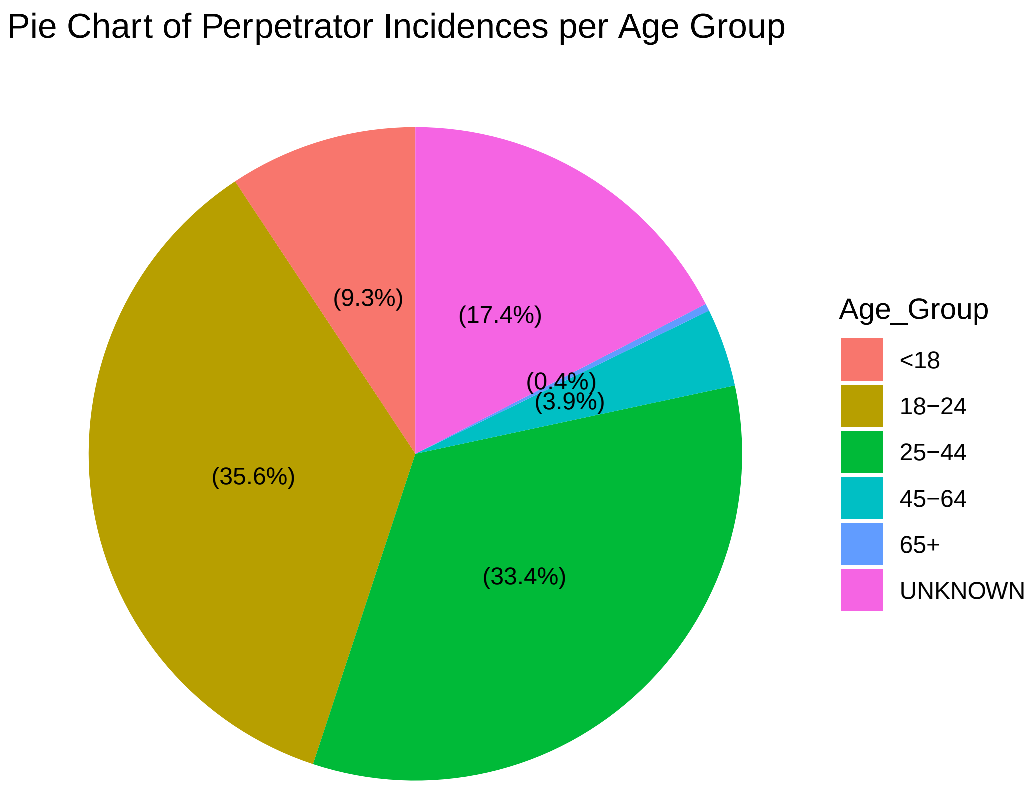
Manhattan - "https://data.census.gov/profile/Manhattan_borough,_New_York_County,_New_York?g=060XX00US3606144919"

Queens - "https://data.census.gov/profile/Queens_borough,_Queens_County,_New_York?g=060XX00US3608160323"

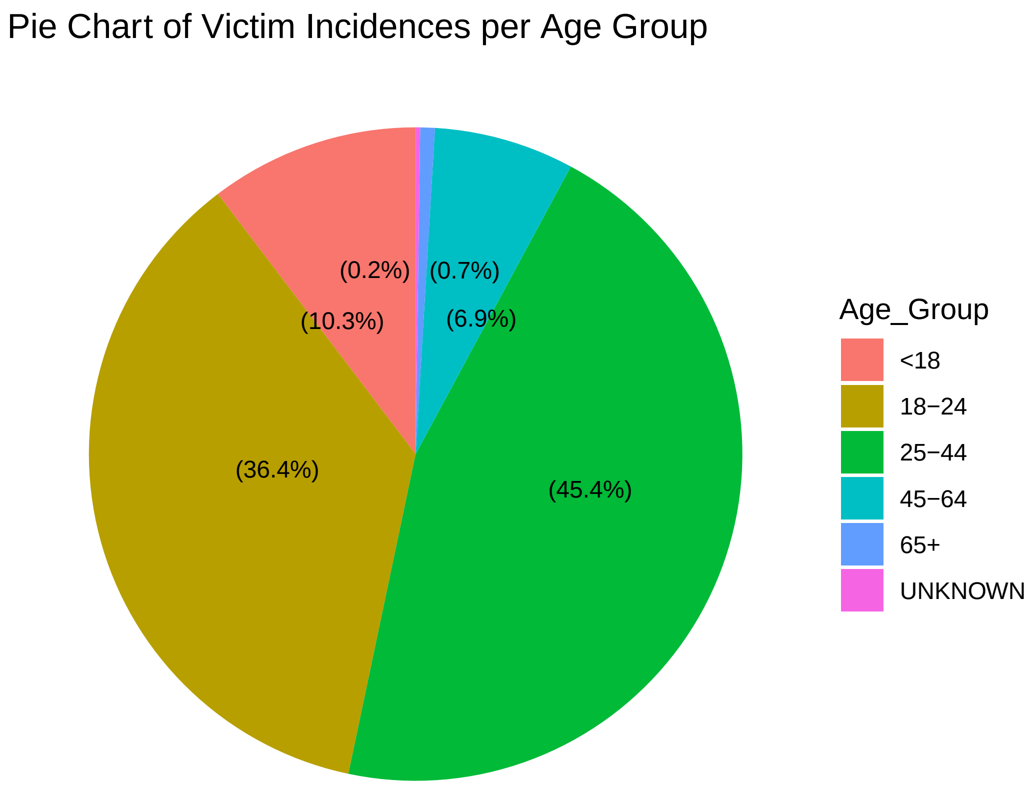
Staten Island - "https://data.census.gov/profile/Staten_Island_borough,_Richmond_County,_New_York?g=060XX00US3608570915"

Since the data is from the 2020 census, the data points have been filtered to only include incidents from 2020 to have a fair comparison.

```r
perp_freq$Label <- paste("(", perp_freq$Percentage, "%)", sep = "")
ggplot(perp_freq, aes(x = "", y = Incidence, fill = Age_Group)) +
    geom_bar(stat = "identity", width = 1) +
    coord_polar("y", start = 0) +
    theme_void() + geom_text(aes(label = Label),
                             position = position_stack(vjust = 0.5), size = 3) +
    labs(title = "Pie Chart of Perpetrator Incidences per Age Group")
```

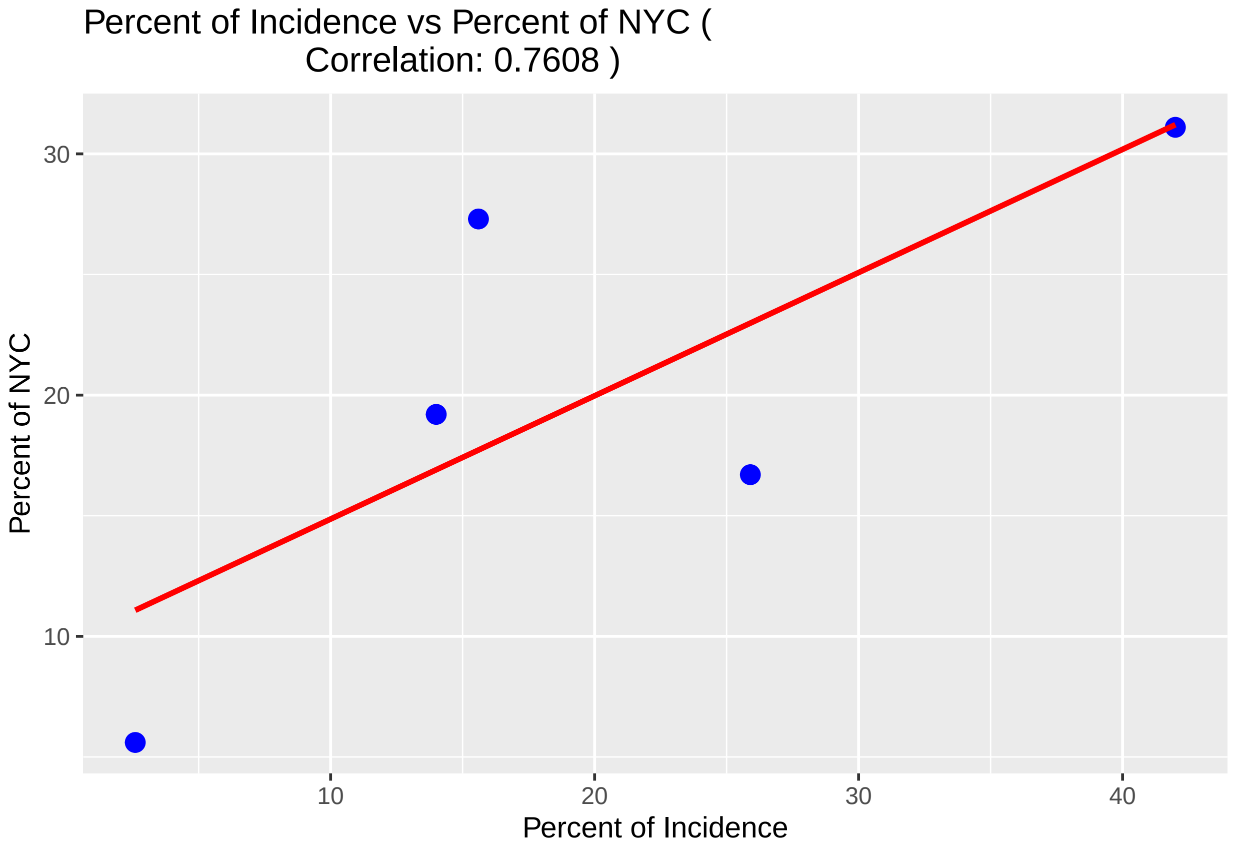


```
vic_freq$Label <- paste("(", vic_freq$Percentage, "%)", sep = "")
ggplot(vic_freq, aes(x = "", y = Incidence, fill = Age_Group)) +
    geom_bar(stat = "identity", width = 1) +
    coord_polar("y", start = 0) +
    theme_void() + geom_text_repel(aes(label = Label),
    position = position_stack(vjust = 0.5), size = 3) +
    labs(title = "Pie Chart of Victim Incidences per Age Group")
```



```
ggplot(ny_borough, aes(x = Percent_of_Incidence, y = Percent_of_nyc)) +
    geom_point(color = "blue", size = 3) +
    geom_smooth(method = "lm", color = "red", se = FALSE) +
    labs(title = paste("Percent of Incidence vs Percent of NYC (
                       Correlation:", round(correlation, 4), ")"),
         x = "Percent of Incidence",
         y = "Percent of NYC")
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



Here I've made two pie charts of what was analyzed above: the percentages of perpetrators and victims of the shootings by their age range.

There's also a line of best fit looking at what percent of NYC each borough makes up vs percent of incidences of the shootings. As can be seen, the correlation is there but not too strong: 0.7608 and r-squared value of 0.58. Given this, we cannot say that the differences in rate of incidents across boroughs is only accounted for by their populations. There are likely many other factors influencing the shooting incidence rates, like socio-economic factors: income, education, access to housing, etc.

This also raises questions like what does the percentages of each age group as the perpetrators and victims compare to the percentages of each age group in the actual population? This is a path to further explore.

Conclusion

To conclude, we can say that the differences in shooting incidences across boroughs can be mainly explained by the differences in their population density, but not totally. This is an area to explore, to see if socio-economic factors like income and access to housing can explain the rest of the differences. 18-24 year-olds and 25-44 year-olds make up a majority of the perpetrators of these shootings (35.6% & 33.4%) as well as the victims of the shootings as well (36.4% & 45.4%).

Looking at the perpetrator data and pie chart, there is a very big percent of perpetrators that are unknown and were unidentified (17.4%). This data could skew the results any which way. This gives us a sense of bias in that all we know comes from those that were caught, and those that weren't (and perhaps results in them committing more of the shootings in our data set) are not represented in the perpetrators data.

Another aspect of this data that is questionable comes from the fact that this data spans many years (17 years to be exact). Reporting practices and policies may have changed, and thus might reflect differently across the years. This bias might also include law enforcement practice biases. We have no knowledge of the differences in policing across the neighborhoods and boroughs, which would skew the amount of incidents reported and perpetrators caught across these locations. These practices might have changed across the years, leading to increases of decreases in policing in certain locations, subsequently resulting in increases/decreases of incidents reported due to these practices.

I tried not to let my personal biases affect my analysis. There may have been personal bias about which boroughs have higher incidents, which I tried to mitigate by bringing in data about each boroughs population. The best way to mitigate personal bias is to let the data tell you its patterns without looking for certain patterns from the start.