NYPD Shootings Social Economic Impact Analysis

6/12/2021

The Data

This is an analysis of shootings in the City of New York based on the "NYPD Shooting Incident Data (Historic)" sourced from
https://catalog.data.gov/dataset
with its details here
https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8.

There are 23,568 data points in 19 columns with a date range from 01/01/2006 to 12/31/2020. Note that data presented is an aggregation of this date range and not adjusted to intervals such as day, week or month.

Missing data for the original data as below.

description

missing

actions

Location of the shooting incident

13,581

ignore - ie house, building, bar

Perpetrator's age within a category

8,459

ignore - assume not seen

Perpetrator's sex description

8,425

ignore - assume not seen

Perpetrator's race description

8,425

ignore - assume not seen

For this analysis the above missing fields will not be used. In the case they are required for subsequent in depth analysis, it may be possible to derive location characteristics, was the shooting at a home, store or workplace, from the longitude and latitude data. Perpetrator characteristics are considered impossible to derive and assumed to be missing due to lack of witnesses.

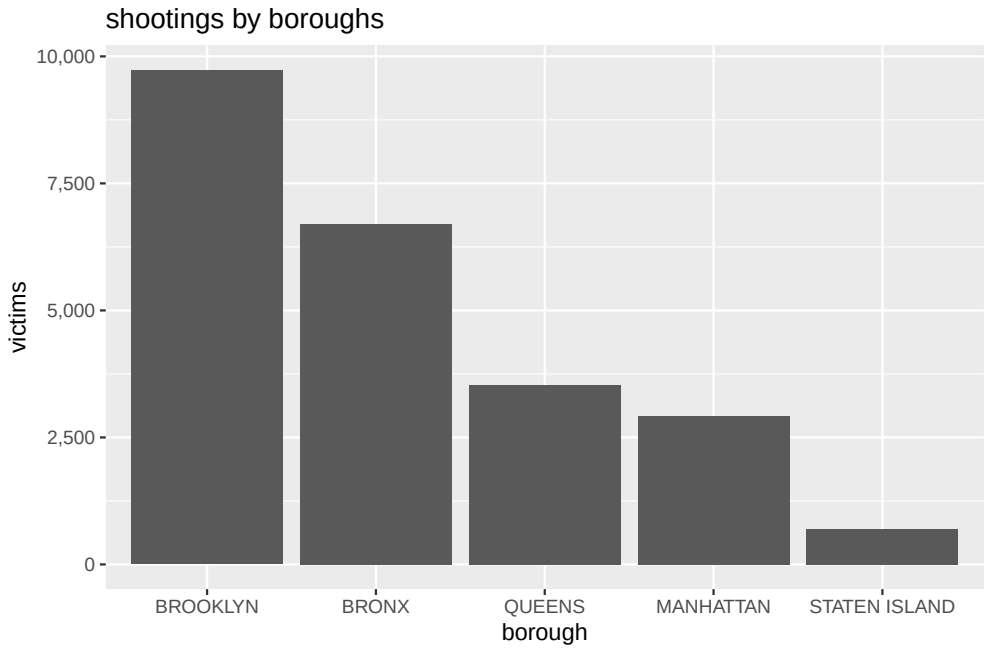
What I am Interested in

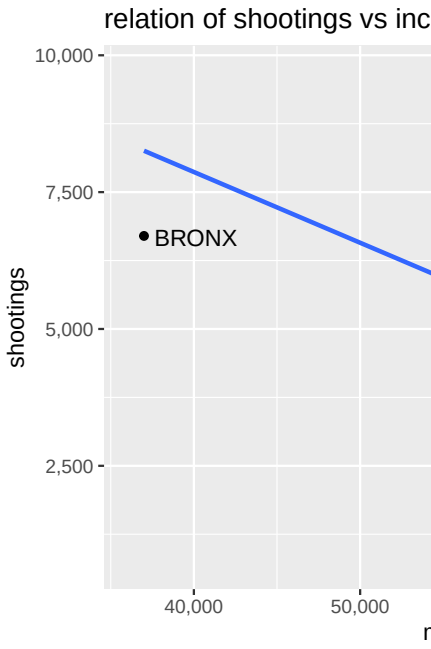
For this analysis I will focus on whether social economic factors can be correlated with shootings. Typically it is said that "better off neighborhoods" have lower crime and can be attributed to fewer criminals in the neighborhood and or higher security. A key characteristic of a "better off neighborhood" can be income which we will look at.

The Evidence

In the "shootings by boroughs" chart we can see higher levels of shootings in some boroughs vs others however based on the https://catalog.data.gov/dataset we can't see why with respect to income. I add data sourced from Quora and a regression model to understand income and shooting correlation as shown in "relation of shootings vs income" chart. Assumptions can be made on median income; lower income areas may attract more shootings as show with the regression line. As mentioned earlier, this can be due to fewer criminals in a high security area.

https://www.quora.com/What-is-the-richest-borough-in-NYC





What Next

This simple analysis helps us conclude that there are social economic factors such as income has some sort of impact on shootings. If we want to find long term sustainable solutions to crime I believe this finding tells us to further investigate improving social economic factors in order to reduce crime. There is data, not presented here, which can lead us to conclusions for tactical fixes such as more police in certain areas and at certain times or tell us "don't stay out at night". These tactical fixes may just be bandage.

Further analysis based on mathematical processes with larger and broader datasets are required to prove the validatity of this report and determine action plans. Some areas to investigate to better understand social economic impact on shootings are

- education and occupation have an impact on income but how much direct impact
- do familiy structures and how they spend time together have an impact.
- are there correlations between victim and perpetrator to crimes such as robbery or gang wars

---

In Hindsight

I use the data in this analysis to offset and at the same time support my bias of New York City including those on correlation of New York City areas and safety. I have lived and worked in Manhattan I have formed strong opinions over the years which have given me hints on where to start the analysis, what to do to offset my biases to better understand as well as educate and convince others on topics which interest me such as social economic issues.

I assume that since this data is from a government source, it is reasonably accurate and based on standard data collection techniques. When viewing the raw data I felt that the breadth of data available from this source is biased in that it seems to indicate race is an issue however is that so or rather is root cause of shootings due to a social economic factor which can be changed

I will, for future analysis, when adding other similar and complementary sources will

- Assess whether there are discrepancies between data used for this analysis and other sources. Discrepancies can indicate data quality issues or bias due to data collection techniques.
- Potentially leverage other sources fill in the missing data gaps such as "location", ie bar, home, apartment, which can give me more insight whether venue has an impact

---

Summary of original data set

```
##   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       Mode :logical
##  1st Qu.: 44.00   1st Qu.:0.0000    Class :character   FALSE:19080
##  Median : 69.00   Median :0.0000    Mode  :character   TRUE :4488
##  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       Min.   : 914928   Min.   :125757
##  Class :character   Class :character   1st Qu.: 999900   1st Qu.:182565
##  Mode  :character   Mode  :character   Median :1007645   Median :193482
##                                        Mean   :1009363   Mean   :207312
##                                        3rd Qu.:1016807   3rd Qu.:239163
##                                        Max.   :1066815   Max.   :271128
##
##     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
##
```

This report was generated in the following environment

```
## R version 4.1.1 (2021-08-10)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Big Sur 11.4
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/lib/libRlapack.dylib
##
## locale:
## [1] ja_JP.UTF-8/ja_JP.UTF-8/ja_JP.UTF-8/C/ja_JP.UTF-8/ja_JP.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] ggrepel_0.9.1     formattable_0.2.1 lubridate_1.7.10  forcats_0.5.1
##  [5] stringr_1.4.0     dplyr_1.0.7       purrr_0.3.4       readr_2.0.1
##  [9] tidyr_1.1.3       tibble_3.1.3      ggplot2_3.3.5     tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.7        lattice_0.20-44   assertthat_0.2.1  digest_0.6.27
##  [5] utf8_1.2.2        R6_2.5.0          cellranger_1.1.0  backports_1.2.1
##  [9] reprex_2.0.1      evaluate_0.14     httr_1.4.2        pillar_1.6.2
## [13] rlang_0.4.11      curl_4.3.2        readxl_1.3.1      rstudioapi_0.13
## [17] Matrix_1.3-4      rmarkdown_2.10    splines_4.1.1     labeling_0.4.2
```

```
## [21] htmlwidgets_1.5.3 bit_4.0.4         munsell_0.5.0     broom_0.7.9
## [25] compiler_4.1.1    modelr_0.1.8      xfun_0.25         pkgconfig_2.0.3
## [29] mgcv_1.8-36       htmltools_0.5.1.1 tidyselect_1.1.1  fansi_0.5.0
## [33] crayon_1.4.1      tzdb_0.1.2        dbplyr_2.1.1      withr_2.4.2
## [37] grid_4.1.1        nlme_3.1-152      jsonlite_1.7.2    gtable_0.3.0
## [41] lifecycle_1.0.0   DBI_1.1.1         magrittr_2.0.1    scales_1.1.1
## [45] cli_3.0.1         stringi_1.7.3     vroom_1.5.4       farver_2.1.0
## [49] fs_1.5.0          xml2_1.3.2        ellipsis_0.3.2    generics_0.1.0
## [53] vctrs_0.3.8       tools_4.1.1       bit64_4.0.5       glue_1.4.2
## [57] hms_1.1.0         parallel_4.1.1    yaml_2.2.1        colorspace_2.0-2
## [61] rvest_1.0.1       knitr_1.33        haven_2.4.3
```

Show the code for reproducibility

```
library(tidyverse)
library(lubridate)
library(ggplot2)
library(formattable)
library(ggrepel)

#
# pull and read NYPD Shooting Incident Data (Historic) data from data.gov
# tally up missing data
#
# data description
# https://opendatanetwork.herokuapp.com/dataset/data.cityofnewyork.us/833y-fsy8
#
url_in <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
nypd_shootings <- read_csv(url_in)

#
# show date range of data embedded in a sentence
There are `r format(nrow(nypd_shootings), big.mark = ",")` data points in `r format(ncol(nypd_shootings)

#
# prep and show the missing data counts and actions in a table
description = c(
    'Location of the shooting incident',
    'Perpetrator's age within a category',
    'Perpetrator's sex description',
    'Perpetrator's race description')
missing = accounting(c(
    sum(is.na(nypd_shootings$LOCATION_DESC)),
    sum(is.na(nypd_shootings$PERP_AGE_GROUP)),
    sum(is.na(nypd_shootings$PERP_SEX)),
    sum(is.na(nypd_shootings$PERP_RACE))), format = "d")
actions = c(
    'ignore - ie house, building, bar',
    'ignore - assume not seen',
    'ignore - assume not seen',
    'ignore - assume not seen')
missing_data <- data.frame(description, missing, actions)
```

```
`r formattable(missing_data, align = c("l","r","l"))`

#
# extract and convert data to be reused for the following graphs
nypd_shootings_analyze <- nypd_shootings %>%
  mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
  mutate(OCCUR_TIME = hms(OCCUR_TIME))

#
# prep data and graph for "shootings by borough" sorted by number of shootings
nypd_shootings_borough <- nypd_shootings_analyze %>%
  count(BORO) %>%
  arrange(desc(n))
nypd_shootings_borough_plot = ggplot(data=nypd_shootings_borough, aes(x = reorder(BORO, -n), y = n)) + g

#
# prep additional data source, add regression model and graph to correlate borough, median income and s
BORO = c('MANHATTAN', 'STATEN ISLAND', 'QUEENS', 'BROOKLYN', 'BRONX')
income = c(85000, 79000, 64000, 57000, 37000)
median_income_by_borough <- data.frame(BORO, income)
income_correlation <- merge(x=median_income_by_borough, y=as.data.frame(nypd_shootings_borough), by="BO
nypd_shootings_income_shooting_plot = ggplot(data=income_correlation, aes(x = income, y = n)) + geom_po

#
# output the graphs prepared above
plot(nypd_shootings_borough_plot)
plot(nypd_shootings_income_shooting_plot)

#
# show summary of original data set and session which details the environment this report was generated
summary(nypd_shootings)
sessionInfo()
```