NYPD Shooting Incident Analysis

2022-07-28

NYPD Shooting Incident Data

Source: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic

Importing Data

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

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

```
url <- 'https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD'
main_data <- read_csv(url[1])
```

```
## Rows: 25596 Columns: 19
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (10): OCCUR_DATE, BORO, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_R...
## 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.
```

Tidying and Transforming Data

The NYPD Shooting Incident data file has a large amount of missing information pertaining to the perpetrator as well as information not needed for this analysis. This analysis will focus on the number of shootings and the date they occurred, therefore, all but the following information will be removed: OCCUR_DATE and BORO. Additionally, OCCUR_DATE will be changed to 'Date' and 'Date' will be changed into date(d-m-y) format.

```
main_data <- main_data %>%
  select(c(OCCUR_DATE, BORO)) %>%
  rename(Date = 'OCCUR_DATE') %>%
  mutate(Date = mdy(Date))
```

```
main_data <- main_data %>%
  count(Date, BORO) %>%
  rename(Shootings_Per_Day = 'n') %>%
  pivot_wider(names_from = BORO, values_from = Shootings_Per_Day) %>%
  dplyr::mutate(BRONX = replace_na(BRONX, 0)) %>%
  dplyr::mutate(BROOKLYN = replace_na(BROOKLYN, 0)) %>%
  dplyr::mutate(MANHATTAN = replace_na(MANHATTAN, 0)) %>%
  dplyr::mutate(QUEENS = replace_na(QUEENS, 0)) %>%
  rename(STATEN_ISLAND = 'STATEN ISLAND') %>%
  dplyr::mutate(STATEN_ISLAND = replace_na(STATEN_ISLAND, 0))
```

```
data2 <- main_data
```

Transforming data to have total number of shootings per month for each borough

```
data2 <- data2 %>%
  mutate(month=format(Date, "%m"), year = format(Date, "%Y")) %>%
  group_by(year, month) %>%
  summarise(total_month_Bronx = sum(BRONX), total_month_Brooklyn = sum(BROOKLYN), total_month_Manhattan
  add_column(day=1) %>%
  mutate(date = make_date(year, month, day))
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

Transforming data to create a new data set which will have total shootings per month for all of New York City

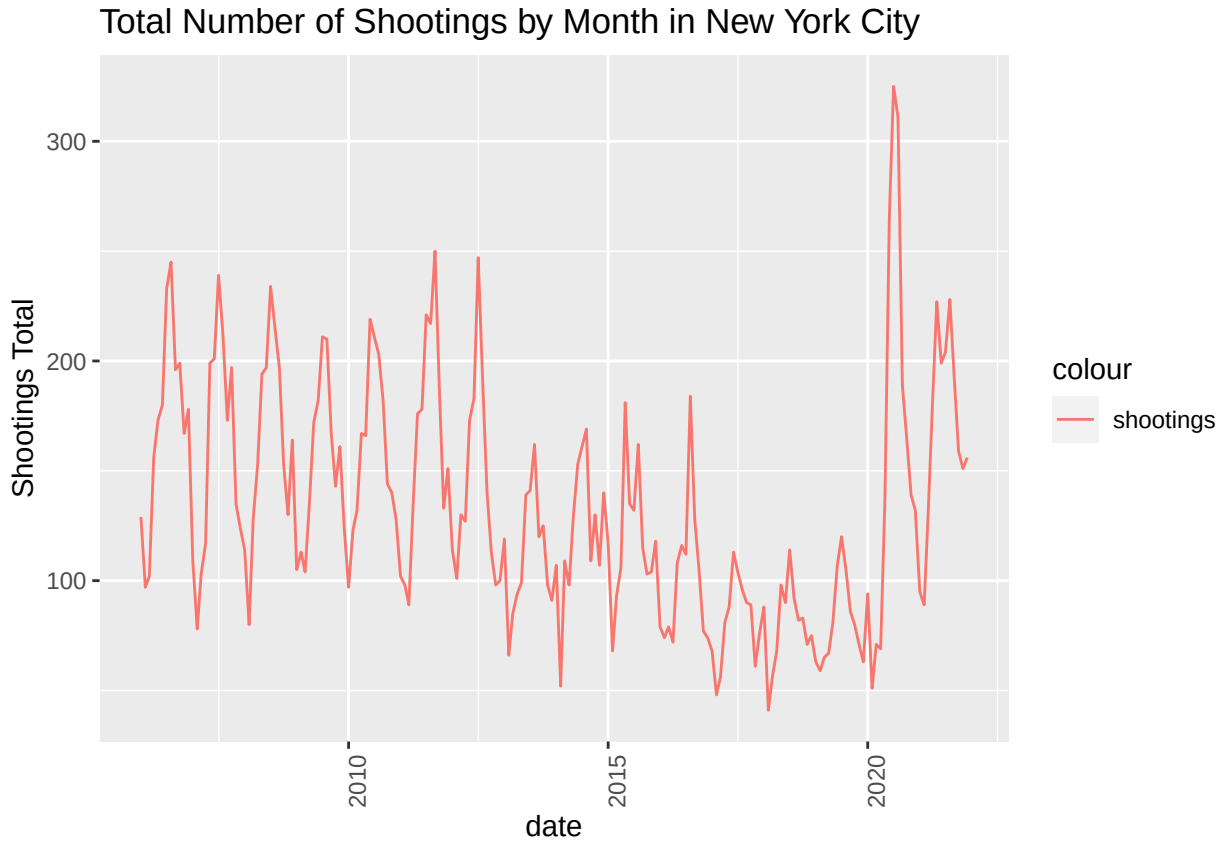
```
data3 <- data2 %>%
  group_by(year, month) %>%
  summarise(shootings = sum(total_month_Bronx, total_month_Brooklyn, total_month_Manhattan, total_month
  add_column(day=1) %>%
  mutate(date = make_date(year,month,day))
```

```
## `summarise()` has grouped output by 'year'. You can override using the
## `.groups` argument.
```

Visualizations

This first visualization shows the total number of shootings each month from 2006 to 2021

```
data3 %>%
  ggplot(aes(x=date, y = shootings)) +
  geom_line(aes(color = "shootings")) +
  theme(legend.position = NULL,
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Total Number of Shootings by Month in New York City", y = 'Shootings Total')
```

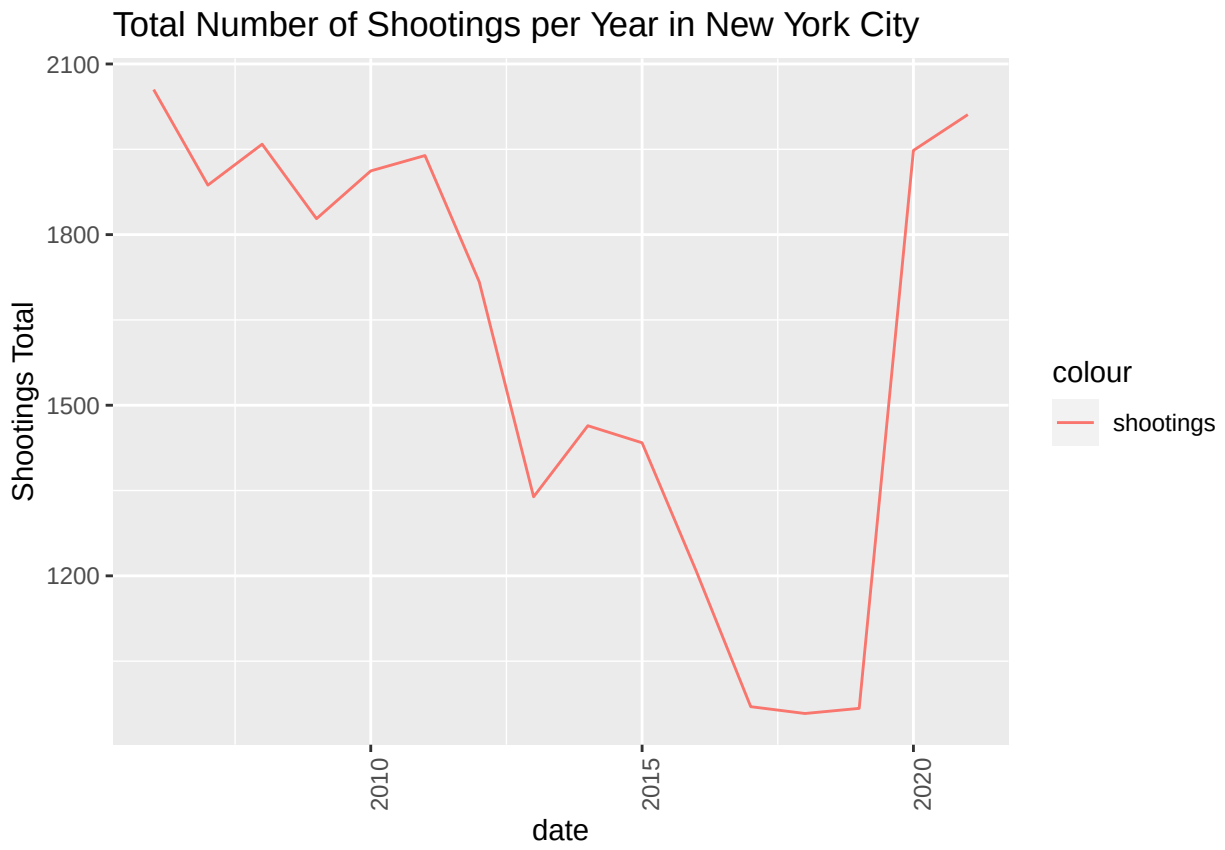



```
shootings_per_year <- data3 %>%
  group_by(year) %>%
  summarise(shootings = sum(shootings)) %>%
  add_column(day=1) %>%
  add_column(month=1) %>%
  mutate(date = make_date(year,month,day))
```

This visualization shows the total number of shootings each year from 2006 to 2021

```
shootings_per_year %>%
  ggplot(aes(x=date, y = shootings)) +
  geom_line(aes(color = "shootings")) +
  theme(legend.position = NULL,
```

```
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Total Number of Shootings per Year in New York City", y = 'Shootings Total')
```

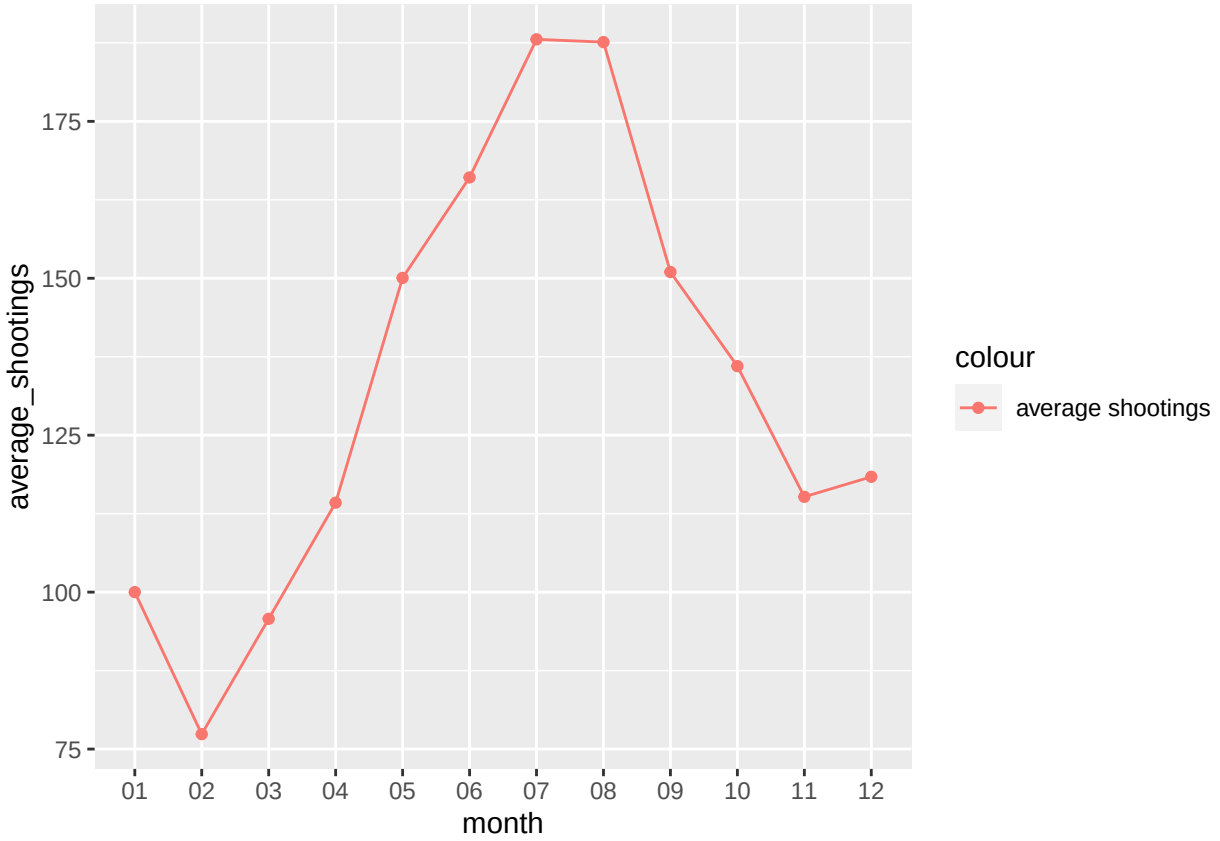


Analysis

From the first graph, it was apparent that there were more shootings in certain parts of the years than others. An analysis showing how many shootings occur each month on average will show which times of the year have the most and the least shootings

```
average_by_month <- data3 %>%
  group_by(month) %>%
  summarise(across(shootings, sum)) %>%
  mutate(shootings/16) %>%
  rename(average_shootings = 'shootings/16') %>%
  rename(total_shootings = 'shootings')
```

```
average_by_month %>%
  ggplot(aes(x=month, y = average_shootings)) +
  geom_line(aes(color = "average shootings", group=1)) +
  geom_point(aes(color = "average shootings"))
```




```
  theme(legend.position = NULL,
        axis.text.x = element_text(angle = 90)) +
  labs(title = "Average Number of Shootings in New York City by Month", y = 'Average Shootings')
```

```
## List of 4
##  $ axis.text.x    :List of 11
##   ..$ family       : NULL
##   ..$ face         : NULL
##   ..$ colour       : NULL
##   ..$ size         : NULL
##   ..$ hjust        : NULL
##   ..$ vjust        : NULL
##   ..$ angle        : num 90
##   ..$ lineheight   : NULL
##   ..$ margin       : NULL
##   ..$ debug        : NULL
##   ..$ inherit.blank: logi FALSE
##   ..- attr(*, "class")= chr [1:2] "element_text" "element"
##  $ legend.position: NULL
##  $ y              : chr "Average Shootings"
##  $ title          : chr "Average Number of Shootings in New York City by Month"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

This graph shows that on average, shootings in New York city tend to start off on the lower end at the

beginning of the year and except for February, they rise in frequency until they peak in July/August, and then decrease for the majority of the year afterwards.

Another analysis below will rank the days for each borough that have the highest number of shootings.

```
main_data %>%
  slice_max(MANHATTAN, n=3, with_ties = FALSE) %>%
  select(Date, MANHATTAN)
```

```
## # A tibble: 3 x 2
##   Date       MANHATTAN
##   <date>         <int>
## 1 2020-07-05        16
## 2 2014-10-02        13
## 3 2008-07-29        11
```

```
main_data %>%
  slice_max(BROOKLYN, n=3, with_ties = FALSE) %>%
  select(Date, BROOKLYN)
```

```
## # A tibble: 3 x 2
##   Date       BROOKLYN
##   <date>        <int>
## 1 2011-09-05       19
## 2 2020-07-26       19
## 3 2006-09-04       16
```

```
main_data %>%
  slice_max(BRONX, n=3, with_ties = FALSE) %>%
  select(Date, BRONX)
```

```
## # A tibble: 3 x 2
##   Date       BRONX
##   <date>     <int>
## 1 2018-01-06    19
## 2 2007-08-11    14
## 3 2012-08-26    14
```

```
main_data %>%
  slice_max(QUEENS, n=3, with_ties = FALSE) %>%
  select(Date, QUEENS)
```

```
## # A tibble: 3 x 2
##   Date       QUEENS
##   <date>      <int>
## 1 2008-05-18     14
## 2 2009-10-02     10
## 3 2010-02-22      9
```

```
main_data %>%
  slice_max(STATEN_ISLAND, n=3, with_ties = FALSE) %>%
  select(Date, STATEN_ISLAND)
```

```
## # A tibble: 3 x 2
##   Date       STATEN_ISLAND
##   <date>             <int>
## 1 2006-10-09            13
## 2 2008-05-28             8
## 3 2009-07-25             6
```

When viewing the three days with the highest number of shootings for each borough, the days generally fall int the warmer months of the year but not always. More analysis would be needed to know what other factors could cause a specific day or time of the year to have more shootings than others.

Conclusion and Bias

In conclusion, there is a measurable correlation between the time of year and number of shootings in New York City. Of course, correlation does not always indicate causation. It would be beneficial to see data from more years to see if this trend holds true based on older historical data. Also, more analysis is needed to determine the true cause in the uptick of shootings over the warmer months. A possible explanation is that people are simply outside for longer periods of time during the warmer months due to more tolerable temperatures, but it is possible that other factors are at play. There is likely more tourism and more people in the city overall in the summer months for the same reason.

It is important to consider my own biases in this analysis. First of all, I did not include any information on gender, age, or ethnicity. Though these factors could be used in a separate analysis, I did not want to include them because they had missing information and these three biases are difficult to fully ignore. I certainly have my own implicit biases for these factors but in choosing to not use them and purely look at numbers of shootings and dates, I have mitigated my bias.

I also have my own bias in regards to guns and gun violence. I believe that gun violence is a touchy subject for most and to mitigate this bias, I again have chosen to stick to using numerical data only.

```
sessionInfo
```

```
## function (package = NULL)
## {
##     z <- list()
##     z$R.version <- R.Version()
##     z$platform <- z$R.version$platform
##     if (nzchar(.Platform$r_arch))
##         z$platform <- paste(z$platform, .Platform$r_arch, sep = "/")
##     z$platform <- paste0(z$platform, " (", 8 * .Machine$sizeof.pointer,
##         "-bit)")
##     z$locale <- Sys.getlocale()
##     z$running <- osVersion
##     z$RNGkind <- RNGkind()
##     if (is.null(package)) {
##         package <- grep("^package:", search(), value = TRUE)
##         keep <- vapply(package, function(x) x == "package:base" ||
##             !is.null(attr(as.environment(x), "path")), NA)
##         package <- .rmpkg(package[keep])
##     }
##     pkgDesc <- lapply(package, packageDescription, encoding = NA)
##     if (length(package) == 0)
##         stop("no valid packages were specified")
```

```
##     basePkgs <- sapply(pkgDesc, function(x) !is.null(x$Priority) &&
##         x$Priority == "base")
##     z$basePkgs <- package[basePkgs]
##     if (any(!basePkgs)) {
##         z$otherPkgs <- pkgDesc[!basePkgs]
##         names(z$otherPkgs) <- package[!basePkgs]
##     }
##     loadedOnly <- loadedNamespaces()
##     loadedOnly <- loadedOnly[!(loadedOnly %in% package)]
##     if (length(loadedOnly)) {
##         names(loadedOnly) <- loadedOnly
##         pkgDesc <- c(pkgDesc, lapply(loadedOnly, packageDescription))
##         z$loadedOnly <- pkgDesc[loadedOnly]
##     }
##     z$matprod <- as.character(options("matprod"))
##     es <- extSoftVersion()
##     z$BLAS <- as.character(es["BLAS"])
##     z$LAPACK <- La_library()
##     l10n <- l10n_info()
##     if (!is.null(l10n["system.codepage"]))
##         z$system.codepage <- as.character(l10n["system.codepage"])
##     if (!is.null(l10n["codepage"]))
##         z$codepage <- as.character(l10n["codepage"])
##     class(z) <- "sessionInfo"
##     z
## }
## <bytecode: 0x000002276db165d8>
## <environment: namespace:utils>
```