# NYPD Shooting Data Analysis

2022-10-05

## NYPD Shooting Incident Data (Historic)

This is an analysis of shooting data from the NYPD covering from 2006 to present day. Event information as well as data related to suspect and victim demographics are included. Data can be found here: https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD

## Question

I'd like to see if there are any trends in when the incidents happen or when murders happen. Another question I have about the data is if there is a relationship between the number of incidents and number of murders. I will tidy and transform the data before I start my analysis and visualization.

## Imports

```
library(tidyverse)
library(tinytex)
library(lubridate)
url.data = 'https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD'
nypd.data <- read_csv(url.data)
```

## Tidying and Transformation

For this analysis we are looking at NYPD shooting incident data. The next chunk will clean up and transform the data so that we can look closer at the trend of incidents per month and classify an incident as a murder using a binary column.

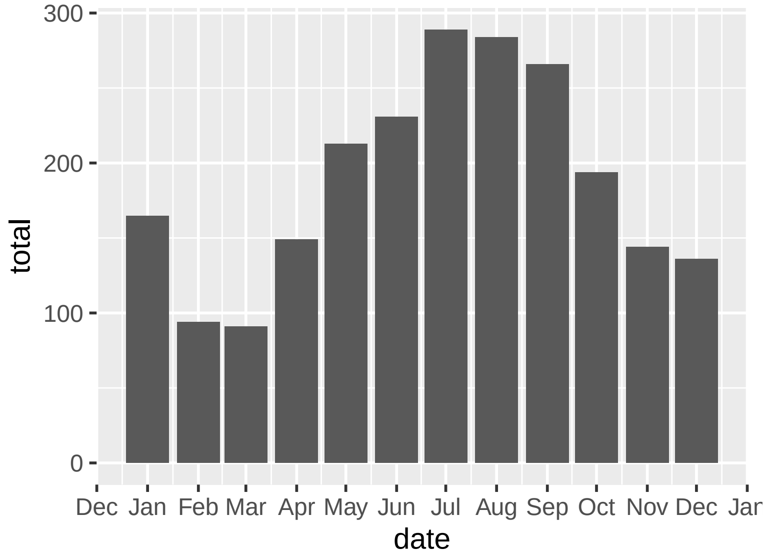
```
# Select important data and add binary column
nypd.grouped <- nypd.data %>% select(c(OCCUR_DATE,BORO,STATISTICAL_MURDER_FLAG,VIC_AGE_GROUP,VIC_SEX)) %
nypd.grouped <- nypd.grouped %>% mutate(murder_binary=as.numeric(STATISTICAL_MURDER_FLAG))

# Calculate total number of incidents per date
nypd.date <- nypd.grouped %>% group_by(OCCUR_DATE) %>% summarize(total = n(), total_murders = sum(murder

# Add numeric month column
nypd.date$month <- nypd.date$OCCUR_DATE %>% month()

# Normalizing the date by giving the same year and day of month
nypd.date <- nypd.date %>% mutate(date = ymd(paste("2000",month,"01",sep='-')))

nypd.grouped
```

```
## # A tibble: 25,596 x 6
##    OCCUR_DATE BORO      STATISTICAL_MURDER_FLAG VIC_AGE_GROUP VIC_SEX murder_b~1
##    <date>     <chr>     <lgl>                   <chr>         <chr>        <dbl>
##  1 2021-11-11 BROOKLYN  FALSE                   18-24         M                0
##  2 2021-07-16 BROOKLYN  FALSE                   25-44         M                0
##  3 2021-07-11 BROOKLYN  FALSE                   25-44         M                0
##  4 2021-12-11 BROOKLYN  FALSE                   25-44         M                0
##  5 2021-02-16 QUEENS    FALSE                   25-44         M                0
##  6 2021-05-15 QUEENS    TRUE                    25-44         M                1
##  7 2021-04-14 BRONX     TRUE                    18-24         M                1
##  8 2021-12-10 BRONX     FALSE                   25-44         M                0
##  9 2021-02-22 MANHATTAN FALSE                   25-44         M                0
## 10 2021-03-07 BROOKLYN  TRUE                    25-44         M                1
## # ... with 25,586 more rows, and abbreviated variable name 1: murder_binary
```

## Analysis & Visualization

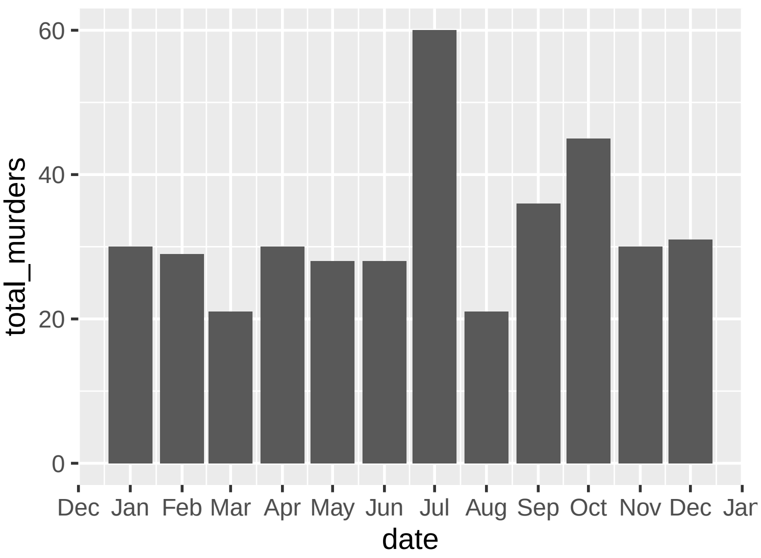
Let's see how the incidents break down by month. The histogram contains the total incidents per month for each year in the data set. There seems to be a trend in the summer months and January, possinly pointing to a correlation between high tourist times.

```
# Create histogram
nypd.date %>% ggplot(aes(x=date ,y=total)) + geom_bar(stat = 'sum') + scale_x_date(date_breaks = '1 mon
```



What about murder's per month? July seems to be an outlier with the most murders compared to the other months.

```
# Create histogram
nypd.date %>% ggplot(aes(x=date ,y=total_murders)) + geom_bar(stat = 'sum') + scale_x_date(date_breaks =
```



## Model

Is the number of incidents able to predict the number of murders? The data does not show a relation ship between the number of incidents per day and the number of murders.

```
mod <- lm(total ~ total_murders, data = nypd.date)
nypd.date <- nypd.date %>% mutate(pred = predict(mod))
summary(mod)
```

```
##
## Call:
## lm(formula = total ~ total_murders, data = nypd.date)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.8392 -2.1175 -0.6619  1.1103 24.3886
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.11754    0.04663   66.86   <2e-16 ***
## total_murders  1.77216    0.02986   59.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.785 on 5407 degrees of freedom
## Multiple R-squared:  0.3944, Adjusted R-squared:  0.3943
## F-statistic:  3522 on 1 and 5407 DF,  p-value: < 2.2e-16
```

## Conclusion

The data shows a trend of more shooting incidents in the summer and January, this could be connected to the number of tourists the are in the city as these could be heightened times. July has the highest number of murders while the other months are pretty average with each other, information on why the murder happened might help determine why this is the case. There does not appear to be a relation between the number of incidents and the number of murders based on my dataset.

## Bias

A potential source for bias is that in my background I am not used to data that is not mostly measurements and numeric data. There are a lot of different variables I did not explore. I make an assumption about when tourists are high in the city due to my view of NYC and summer vacation and New Year's Eve in Times Square. There are many other things that could be looked into that might explain the trends better.

```
sessionInfo()
```

```
## R version 4.1.3 (2022-03-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19042)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] lubridate_1.8.0 tinytex_0.41    forcats_0.5.2   stringr_1.4.1
##  [5] dplyr_1.0.10    purrr_0.3.4     readr_2.1.2     tidyr_1.2.1
##  [9] tibble_3.1.8    ggplot2_3.3.6   tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] assertthat_0.2.1    digest_0.6.29       utf8_1.2.2
##  [4] R6_2.5.1            cellranger_1.1.0    backports_1.4.1
##  [7] reprex_2.0.2        evaluate_0.16       highr_0.9
## [10] httr_1.4.4          pillar_1.8.1        rlang_1.0.5
## [13] googlesheets4_1.0.1 curl_4.3.2          readxl_1.4.1
## [16] rstudioapi_0.14     rmarkdown_2.16      labeling_0.4.2
## [19] googledrive_2.0.0   bit_4.0.4           munsell_0.5.0
## [22] broom_1.0.1         compiler_4.1.3      modelr_0.1.9
## [25] xfun_0.33           pkgconfig_2.0.3     htmltools_0.5.3
## [28] tidyselect_1.1.2    fansi_1.0.3         crayon_1.5.1
## [31] tzdb_0.3.0          dbplyr_2.2.1        withr_2.5.0
## [34] grid_4.1.3          jsonlite_1.8.0      gtable_0.3.1
## [37] lifecycle_1.0.2     DBI_1.1.3           magrittr_2.0.3
## [40] scales_1.2.1        cli_3.4.0           stringi_1.7.6
## [43] vroom_1.5.7         farver_2.1.1        fs_1.5.2
## [46] xml2_1.3.3          ellipsis_0.3.2      generics_0.1.3
## [49] vctrs_0.4.1         tools_4.1.3         bit64_4.0.5
## [52] glue_1.6.2          hms_1.1.2           parallel_4.1.3
## [55] fastmap_1.1.0       yaml_2.3.5          colorspace_2.0-3
## [58] gargle_1.2.1        rvest_1.0.3         knitr_1.40
## [61] haven_2.5.1
```