Untitled

2022-07-04

NYPD ASSIGNMENT

Project Files and Importing Data

For the purpose of this assignment I am going to use data file located on data.gov website titled "NYPD Shooting Incident Data (Historic)" This document contains information about shooting incidents in the city of New York from year 2006 until the end of the year 2021.

Throughout this document I will be using tidyverse library

```
shootings <- read_csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD")
```

```
## 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

Straight away we notice that date is a character vector and not a date object, so we can mutate it using lubridate library.

```
library(lubridate)
shootings <- shootings %>% mutate(OCCUR_DATE = lubridate::mdy(OCCUR_DATE))
```

In order to simplify our data set we can remove information we will not use in our analysis, such as INCIDENT_KEY, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD, Latitude, Longitude and Lon_Lat.

```
shootings <- select(shootings, -c(INCIDENT_KEY, JURISDICTION_CODE, X_COORD_CD, Y_COORD_CD, Latitude, Lon
```

On further inspection we notice that there are many NA values in various columns, however in this case I have decided to leave them there, as they simply might suggest that not all information is available to NYPD for that particular shooting incident (e.g. perpetrator is not known, therefore gender, race or age group columns have NA value).

Additionally I decided to organize my data by the date.

```
shootings <- shootings %>% arrange(ymd(shootings$OCCUR_DATE))
```

Visualizing, Analyzing and Modeling Data

We can look at the total number of cases in different areas of NY using table() funcition.

```
table(shootings$BORO)
```

```
## 
##         BRONX      BROOKLYN     MANHATTAN        QUEENS STATEN ISLAND 
##          7402         10365          3265          3828           736
```

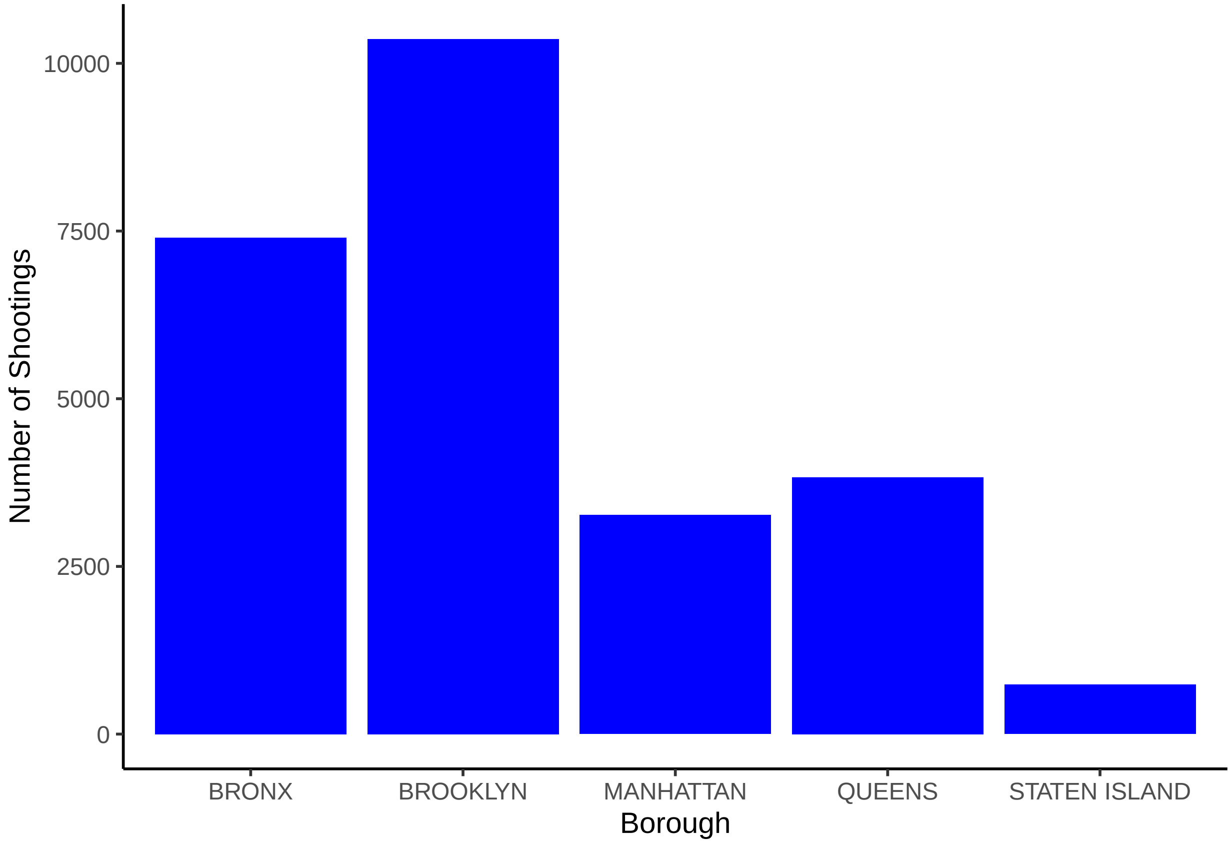
We can notice that NYPD uses 5 different areas in their statistics:

- Manhattan
- Brooklyn
- Bronx
- Queens
- Staten Island

First of all I would like to look at the information about number of incidents and their proportion as per main five boroughs of NYC.

```
borough <- shootings %>%
group_by(BORO) %>%
summarize(cases = n())

# Draw a bar chart using our data
borough %>%
ggplot(aes(x = BORO, y = cases)) +
geom_bar(stat = "identity", fill = "blue") +
theme_classic() +
xlab("Borough") +
ylab("Number of Shootings")
```



It is obvious from our table that most of the shootings happen in the borough of Brooklyn, followed by Bronx. It would require further study to find out what is the reason behind this (most obvious would be comparing amount of shootings with the population numbers, population density and many other factors).

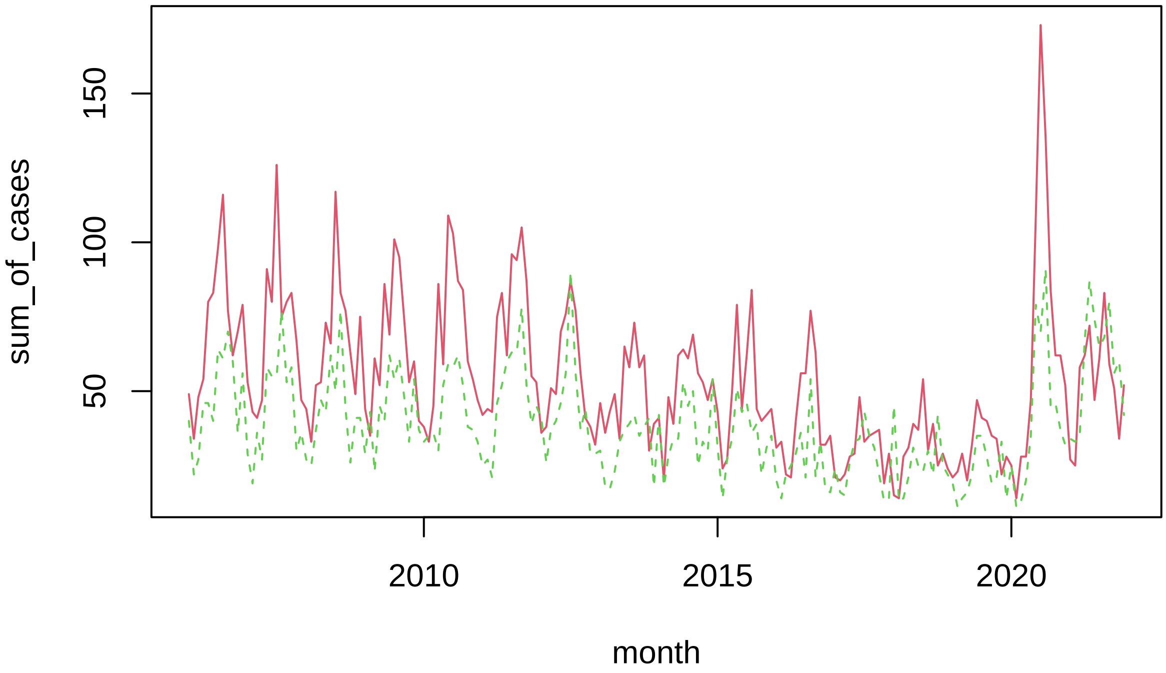
Shootings by months in Boroughs

For my own analysis I have decided to look at each of these areas and notice any patterns. Since we are dealing with a large time frame, I grouped shooting incidents by month.

```
shootings$new <- c(1)
Bronx <- subset(shootings, BORO == 'BRONX') %>%
group_by(month = lubridate::floor_date(OCCUR_DATE, 'month')) %>%
summarize(sum_of_cases = sum(new))
Brooklyn <- subset(shootings, BORO == 'BROOKLYN') %>%
group_by(month = lubridate::floor_date(OCCUR_DATE, 'month')) %>%
summarize(sum_of_cases = sum(new))
Manhattan <- subset(shootings, BORO == 'MANHATTAN') %>%
group_by(month = lubridate::floor_date(OCCUR_DATE, 'month')) %>%
summarize(sum_of_cases = sum(new))
Queens <- subset(shootings, BORO == 'QUEENS') %>%
group_by(month = lubridate::floor_date(OCCUR_DATE, 'month')) %>%
summarize(sum_of_cases = sum(new))
Staten_Island <- subset(shootings, BORO == 'STATEN ISLAND') %>%
group_by(month = lubridate::floor_date(OCCUR_DATE, 'month')) %>%
summarize(sum_of_cases = sum(new))
```

In order to obtain a clearer plot, I decided to have look at two areas with the highest amount of shootings *Brooklyn* and *Bronx*.

```
plot(Brooklyn, type = "l", col = 2)
lines(Bronx, type = "l", lty = 2, col = 3)
```

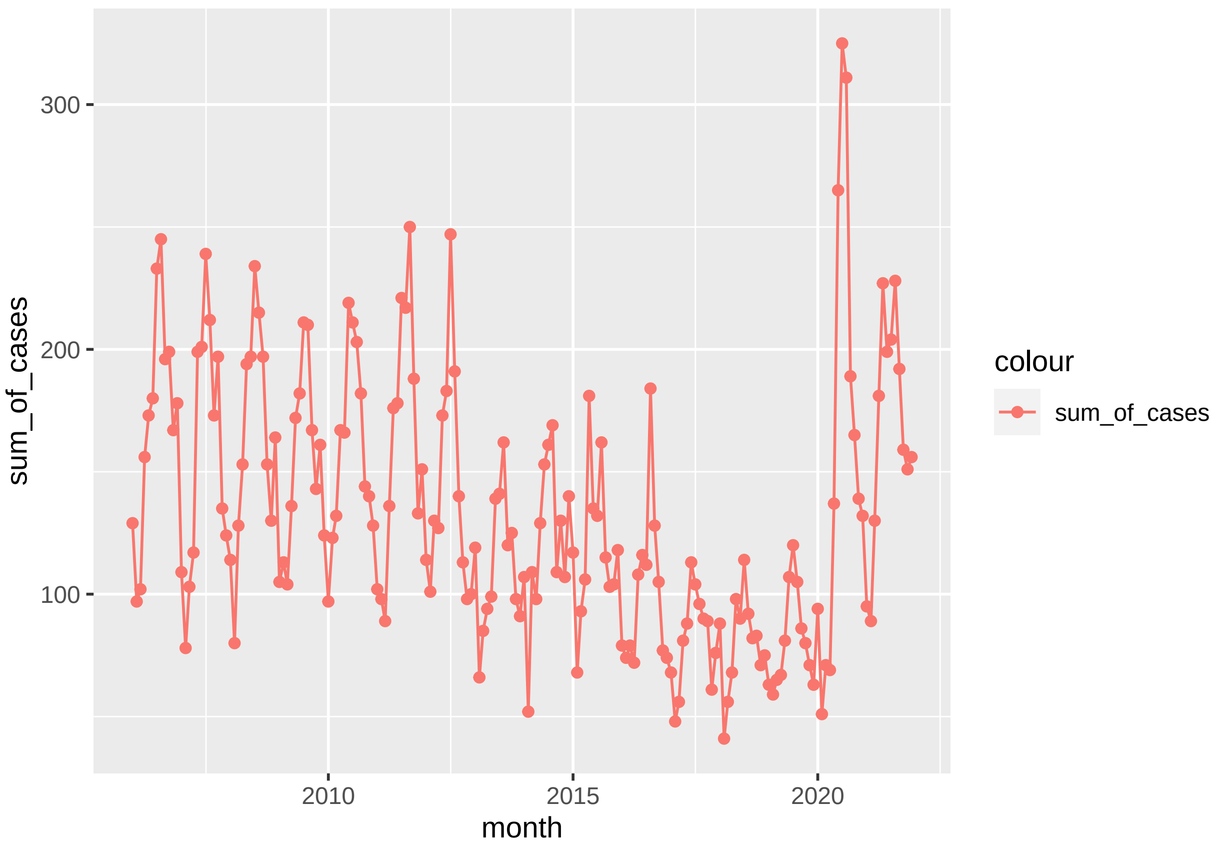


We can see that Brooklyn has by far the highest amount of shooting incidents, followed by Bronx. When we plot our monthly summaries for both of these areas onto the same graph, it is interesting to notice that both of these areas have peaks of incidents matching, i.e. incidents peaking during the same time and equally "incidents lows" occur more or less at the same time frame. Another interesting fact to notice is that there appears to be certain level of seasonality and pattern into highs and lows (e.g. there are 5 high peaks and 5 lows in each 5 year period). However that would require further study and analysis to establish what might be the underlying reason behind this observation.

Finally we can look at the visualization of total number of shooting incidents as per NYPD.

```
Total <- shootings %>% group_by(month = lubridate::floor_date(OCCUR_DATE, 'month')) %>%
summarize(sum_of_cases = sum(new))
```

```
Total %>%
ggplot(aes(x = month, y = sum_of_cases)) +
geom_line(aes(color = "sum_of_cases")) +
geom_point(aes(color = "sum_of_cases"))
```



Bias Identification

My personal bias for this particular assignment could be the fact, that I am not US citizen and I do not live in the USA, therefore I do not know much about situation in NYC or USA in general, except the information that I receive from media. In order to understand fully and be able to analyze these information one needs to have better information about sociological background of NYC, current policies and issues that city of New York might be facing. For this specific reason, it is important that I stick to this data presented here and I do not transfer my personal opinion or limited information into the analysis.

```
sessionInfo()
```

```
## R version 4.2.0 (2022-04-22)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur/Monterey 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2/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] lubridate_1.8.0 forcats_0.5.1   stringr_1.4.0   dplyr_1.0.9    
##  [5] purrr_0.3.4     readr_2.1.2     tidyr_1.2.0     tibble_3.1.7   
##  [9] ggplot2_3.3.6   tidyverse_1.3.1
## 
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.2 xfun_0.31        haven_2.5.0      colorspace_2.0-3
##  [5] vctrs_0.4.1      generics_0.1.2   htmltools_0.5.2  yaml_2.3.5      
##  [9] utf8_1.2.2       rlang_1.0.2      pillar_1.7.0     glue_1.6.2      
## [13] withr_2.5.0      DBI_1.1.2        bit64_4.0.5      dbplyr_2.2.0    
## [17] modelr_0.1.8     readxl_1.4.0     lifecycle_1.0.1  munsell_0.5.0   
## [21] gtable_0.3.0     cellranger_1.1.0 rvest_1.0.2      evaluate_0.15   
## [25] labeling_0.4.2   knitr_1.39       tzdb_0.3.0       fastmap_1.1.0   
## [29] curl_4.3.2       parallel_4.2.0   fansi_1.0.3      highr_0.9       
## [33] broom_0.8.0      backports_1.4.1  scales_1.2.0     vroom_1.5.7     
## [37] jsonlite_1.8.0   farver_2.1.0     bit_4.0.4        fs_1.5.2        
## [41] hms_1.1.1        digest_0.6.29    stringi_1.7.6    grid_4.2.0      
## [45] cli_3.3.0        tools_4.2.0      magrittr_2.0.3   crayon_1.5.1    
## [49] pkgconfig_2.0.3  ellipsis_0.3.2   xml2_1.3.3       reprex_2.0.1    
## [53] assertthat_0.2.1 rmarkdown_2.14   httr_1.4.3       rstudioapi_0.13 
## [57] R6_2.5.1         compiler_4.2.0
```