NYPD_Data_Historic

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.0     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'

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

Introduction

This rmd document describes the data set being imported and the reproducible steps applied to import it into R.

Dataset & questions to answer

The dataset consists of a list of all shooting incidents in New York City going back to 2006 until the end of 2019. Each record in the list represents a shooting incident and additional data on the event including suspect and victim demographics, location time and more. This data is manually reported and extracted every quarter and reviewed before being published on the NYPD website. We will analyze the data to identify any contributing factors based on patterns in these shooting incidents

Importing data

There are multiple methods to importing this dataset. If you are using R studio,download the files to your computer locally and then use the **Import Dataset" UI option in the** Environment** panel. You can also use the below command to import data from a local file on your computer (file path is relative):

```
# dataset <- read.csv("~/./NYPD_Shooting_Incident_Data__Historic_.csv")
```

You can also use the url to the file in the **Import Dataset" UI option or use the below command to read data from the url into your workspace:

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

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   INCIDENT_KEY = col_double(),
##   OCCUR_DATE = col_character(),
##   OCCUR_TIME = col_time(format = ""),
##   BORO = col_character(),
##   PRECINCT = col_double(),
##   JURISDICTION_CODE = col_double(),
##   LOCATION_DESC = col_character(),
##   STATISTICAL_MURDER_FLAG = col_logical(),
##   PERP_AGE_GROUP = col_character(),
##   PERP_SEX = col_character(),
##   PERP_RACE = col_character(),
##   VIC_AGE_GROUP = col_character(),
##   VIC_SEX = col_character(),
##   VIC_RACE = col_character(),
##   X_COORD_CD = col_number(),
##   Y_COORD_CD = col_number(),
##   Latitude = col_double(),
##   Longitude = col_double(),
##   Lon_Lat = col_character()
## )
```

The data directory also contains additional file formats of the dataset like JSON, RDF and XML if preference is to work with these.

Cleaning the dataset

To start, we will remove columns we will not need for the analysis, **(INCIDENT_KEY, JURISDICTION_CODE, STATISTICAL_MURDER_FLAG, X_COORD_CD, Y_COORD_CD, Lotitude, Longitude, Lon_Lat)**

```
nypd_data <- dataset %>%
  select(OCCUR_DATE, OCCUR_TIME, BORO, PRECINCT, LOCATION_DESC, PERP_AGE_GROUP, PERP_SEX, PERP_RACE, VI
```

Next, we will rename some of the columns in this data set

```
nypd_data <- nypd_data %>%
  rename(DATE = OCCUR_DATE, TIME = OCCUR_TIME, LOCATION = LOCATION_DESC)

summary(nypd_data)
```

```
##      DATE               TIME             BORO             PRECINCT
##  Length:23568       Length:23568      Length:23568       Min.   :  1.00
##  Class :character   Class1:hms        Class :character   1st Qu.: 44.00
##  Mode  :character   Class2:difftime   Mode  :character   Median : 69.00
##                     Mode  :numeric                       Mean   : 66.21
##                                                          3rd Qu.: 81.00
##                                                          Max.   :123.00
##    LOCATION         PERP_AGE_GROUP       PERP_SEX          PERP_RACE
##  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_AGE_GROUP        VIC_SEX            VIC_RACE
##  Length:23568       Length:23568       Length:23568
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
```

Transformation, Analysis & Visualization

We will start transforming data for analysis by creating slices of smaller datasets

```
data_occurrence <- nypd_data %>%select(DATE, TIME, BORO)
```

```
data_occurrence<- data_occurrence %>% arrange(DATE)
```

```
analyze_occurrence <- data_occurrence %>%
  group_by(DATE) %>%
  count(DATE)

summary(analyze_occurrence)
```

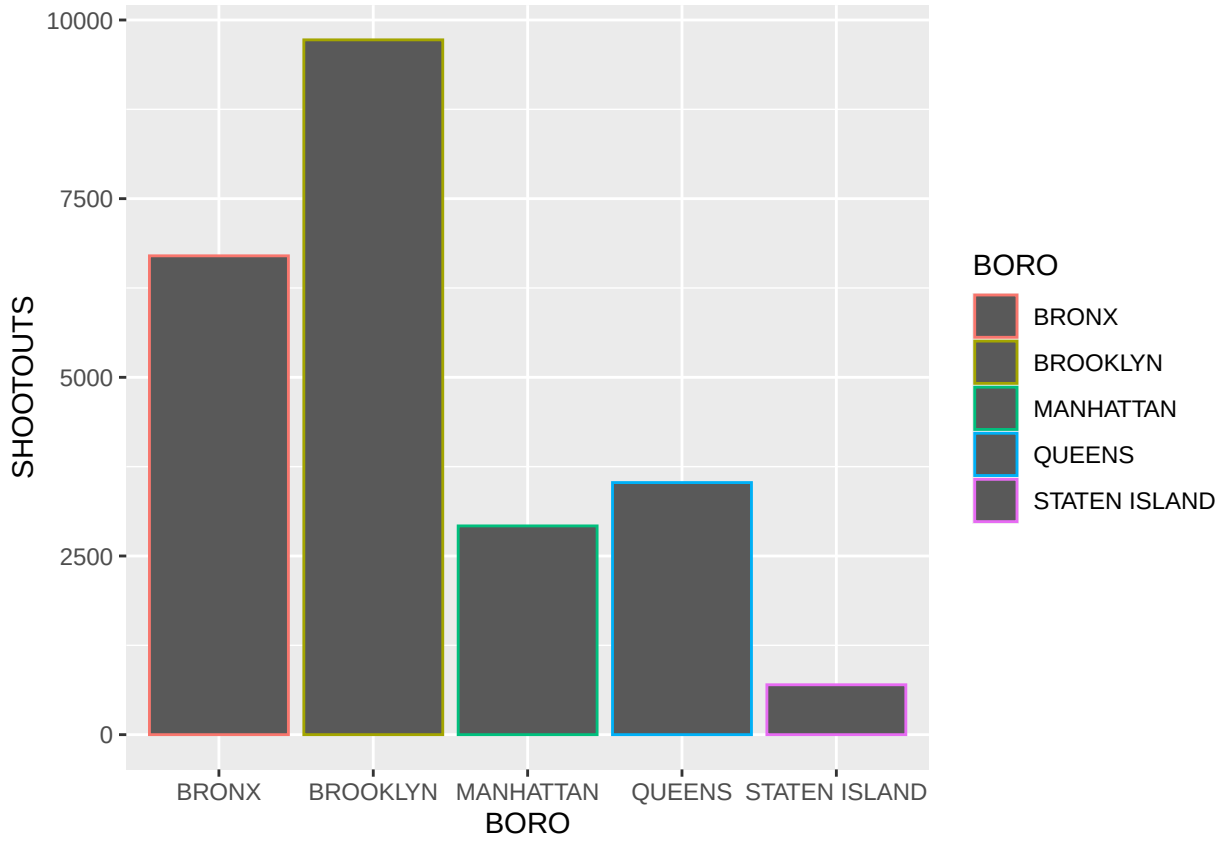
```
##      DATE                 n
##  Length:5054        Min.   : 1.000
##  Class :character   1st Qu.: 2.000
##  Mode  :character   Median : 4.000
##                     Mean   : 4.663
##                     3rd Qu.: 6.000
##                     Max.   :47.000
```

```
data_boro <- data_occurrence %>%
  select(DATE, BORO) %>%
  count(BORO) %>%
  rename(SHOOTOUTS = n)
```

```
data_boro <- data_boro %>%
  arrange(desc(SHOOTOUTS))
analyze_boro <- data_boro
summary(analyze_boro)
```

```
##      BORO             SHOOTOUTS
##  Length:5           Min.   : 698
##  Class :character   1st Qu.:2921
##  Mode  :character   Median :3527
##                     Mean   :4714
##                     3rd Qu.:6700
##                     Max.   :9722
```

```
ggplot(analyze_boro, aes(x=BORO, y = SHOOTOUTS, color = BORO)) + geom_col()
```



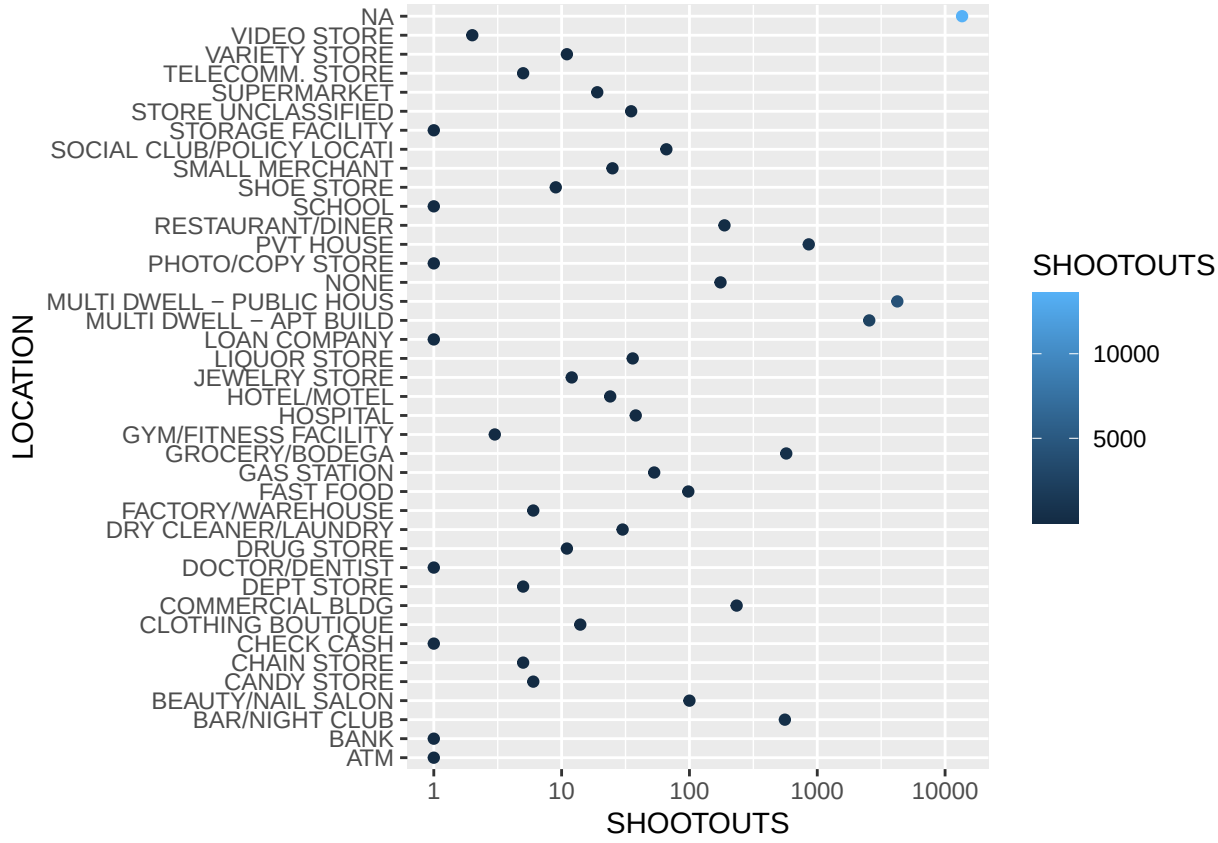

```
data_location <- nypd_data %>%
  count(LOCATION)
```

```
data_location <- data_location %>%
  rename(SHOOTOUTS = n) %>%
  arrange(desc(SHOOTOUTS))

analyze_location <- data_location

summary(analyze_location)
```

```
##    LOCATION           SHOOTOUTS
##  Length:40          Min.   :    1.0
##  Class :character   1st Qu.:    4.5
```

```
##  Mode  :character   Median :   16.5
##                     Mean   :  589.2
##                     3rd Qu.:   98.5
##                     Max.   :13581.0
```

```
ggplot(analyze_location, aes(x = SHOOTOUTS, y = LOCATION, color = SHOOTOUTS)) + geom_point() + scale_x_
```

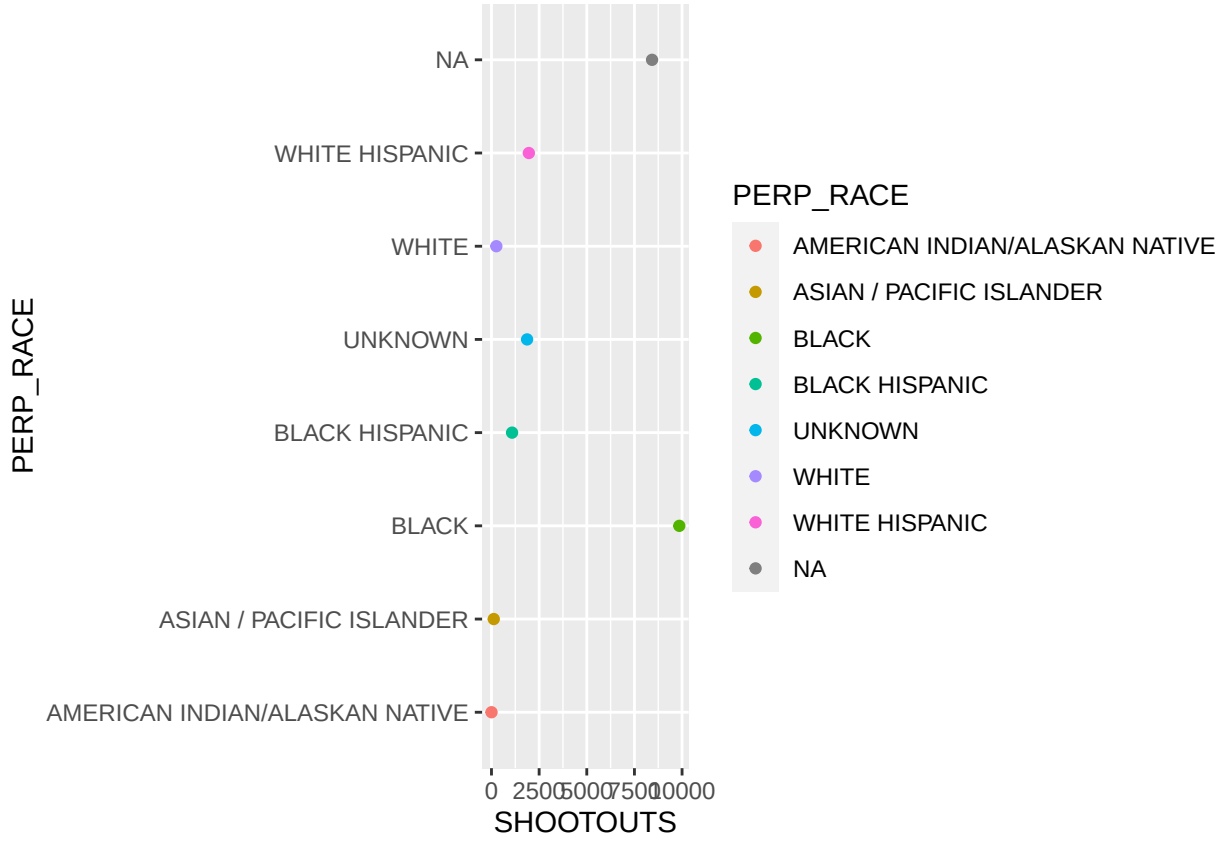



```
data_race <- nypd_data %>%
  select(DATE, PERP_RACE) %>%
  count(PERP_RACE)
```

```
data_race <- data_race %>%
  rename(SHOOTOUTS = n) %>%
  arrange(desc(SHOOTOUTS))

analyze_perp_race <- data_race
summary(analyze_perp_race)
```

```
##   PERP_RACE           SHOOTOUTS
##  Length:8           Min.   :   2.0
##  Class :character   1st Qu.: 221.2
##  Mode  :character   Median :1475.0
##                     Mean   :2946.0
##                     3rd Qu.:3577.0
##                     Max.   :9855.0
```

```
ggplot(analyze_perp_race, aes(x = SHOOTOUTS, y = PERP_RACE, color = PERP_RACE)) + geom_point()
```




```
data_perp_age <- nypd_data %>%
  select(DATE, PERP_AGE_GROUP) %>%
  count(PERP_AGE_GROUP)
```

```
data_perp_age <- data_perp_age %>%
  rename(SHOOTOUTS = n) %>%
  arrange(desc(SHOOTOUTS))

analyze_perp_age <- data_perp_age
summary(analyze_perp_age)
```

```
##  PERP_AGE_GROUP       SHOOTOUTS
##  Length:10          Min.   :   1.00
##  Class :character   1st Qu.:  14.25
##  Mode  :character   Median : 917.50
##                     Mean   :2356.80
##                     3rd Qu.:4248.75
##                     Max.   :8459.00
```

```
data_precinct <- nypd_data %>%
  select(DATE, PRECINCT) %>%
  count(PRECINCT) %>%
```

```
  rename(SHOOTOUTS = n) %>%
  arrange(desc(SHOOTOUTS))

analyze_precinct <- data_precinct
summary(data_precinct)
```

```
##     PRECINCT        SHOOTOUTS
##  Min.   :  1.00   Min.   :   1.0
##  1st Qu.: 32.00   1st Qu.:  61.0
##  Median : 66.00   Median : 166.0
##  Mean   : 63.32   Mean   : 306.1
##  3rd Qu.:100.00   3rd Qu.: 439.0
##  Max.   :123.00   Max.   :1367.0
```

Analysis Summary

There is **no co-relation** between the dates of the shootouts to the event itself. I do not see a pattern of higher or lower occurrence of events based on the day/date.

69.5% of the shootouts between 2006 and 2020 have happened in the boro's of **Brooklyn and Bronx**. If there was access to additional demographics data such as income and population, we could have analyzed this further. Why do most shootings happen in these boro's?

The location has not been captured for ~45% of events. The next highest number of shootouts have happened at **public housing and apartment buildings** when we do not account for events in which the location was not captured.

Analysis of the race of the perpetrator is inconclusive since ~43% of the events do not have a race captured.

Analysis of the age of the perpetrator is inconclusive as well since ~35% of the events do not have the agene pf the perpetrator captured. If we do not account for these events, ~23% of the perpetrators are in the 18-24 age group and ~20% of the perpetrators are in the 25-44 age group.

Analysis of precinct data is also inconclusive as it does not show a co-relation between the occurrence of shootings. Having precinct data mapped to the boro and income would have to be looked at to analyze this further.

Conclusion

At the level of granularity we have been able to get to with this dataset and analysis, the only insight we can confidently put out is that there is a strong relationship with the boro's of Brooklyn, Bronx and the incidents of shootings. The influencing factors are not clear with this data set and there are questions that can be answered with additional demographics data. The one obvious and possible source of bias is the race of perpetrators. It'll be easy to assume that these events are perpetuated by people of a specific race.

```
sessionInfo()
```

```
## R version 4.0.5 (2021-03-31)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.0/Resources/lib/libRblas.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.0/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_1.7.10 forcats_0.5.1    stringr_1.4.0    dplyr_1.0.5
##  [5] purrr_0.3.4      readr_1.4.0      tidyr_1.1.3      tibble_3.1.0
##  [9] ggplot2_3.3.3    tidyverse_1.3.1
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.0  xfun_0.22         haven_2.4.1       colorspace_2.0-0
##  [5] vctrs_0.3.7       generics_0.1.0    htmltools_0.5.1.1 yaml_2.2.1
##  [9] utf8_1.2.1        rlang_0.4.10      pillar_1.6.0      glue_1.4.2
## [13] withr_2.4.1       DBI_1.1.1         dbplyr_2.1.1      modelr_0.1.8
## [17] readxl_1.3.1      lifecycle_1.0.0   munsell_0.5.0     gtable_0.3.0
## [21] cellranger_1.1.0  rvest_1.0.0       evaluate_0.14     labeling_0.4.2
## [25] knitr_1.33        curl_4.3.1        fansi_0.4.2       highr_0.9
## [29] broom_0.7.6       Rcpp_1.0.6        scales_1.1.1      backports_1.2.1
## [33] jsonlite_1.7.2    farver_2.1.0      fs_1.5.0          hms_1.0.0
## [37] digest_0.6.27     stringi_1.5.3     grid_4.0.5        cli_2.4.0
## [41] tools_4.0.5       magrittr_2.0.1    crayon_1.4.1      pkgconfig_2.0.3
## [45] ellipsis_0.3.1    xml2_1.3.2        reprex_2.0.0      assertthat_0.2.1
## [49] rmarkdown_2.8     httr_1.4.2        rstudioapi_0.13   R6_2.5.0
## [53] compiler_4.0.5
```