NYPD Shooting Report

It's a secret :)

18/06/2021

INTRODUCTION

In this report, we will look at the NYPD shooting report. This report contains a breakdown of every shooting incident that occurred in NYC, from 2006 to 2020. In it, we find in which precinct the crime took place, the time and date, as well as if the person died or not. The data is available at the following URL: https://catalog.data.gov/dataset/nypd-shooting-incident-data-historic Analyzing this Data will allow us to get a better understanding of where the crimes generally take place, who are the victims and for what reasons.

Library

We will be using lubridate and tidyverse for our libraries to read and convert the dates as a date object.

Getting and Reading the data

The following chunks of code allows us to get access to the data and store it in the variable NYPD_data

```
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD "
# This line of code gives us access to the report
```

```
NYPD_data <- read_csv(url)
summary(NYPD_data)
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:23568       Length:23568       Length:23568
##  1st Qu.: 55317014   Class :character   Class1:hms         Class :character
##  Median : 83365370   Mode  :character   Class2:difftime    Mode  :character
##  Mean   :102218616                      Mode  :numeric
##  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
##
```

As we can see, this report contains 19 columns which describe who committed the crime, where, on whom, at what time, etc... However, some of these columns are not needed for our analysis, so let's get rid of them. We will also add a column, the population by Borough, as it will be useful if we want to calculate the average by borough

Tidying our data

```
BORO_Url <- "https://data.cityofnewyork.us/api/views/h2bk-zmw6/rows.csv?accessType=DOWNLOAD"
BORO_pop <- read_csv(BORO_Url)
BORO_pop <- BORO_pop %>% rename(BORO = Borough)
BORO_pop$BORO = toupper(BORO_pop$BORO)
NYPD_tidy <- NYPD_data %>% select(-c(X_COORD_CD:Lon_Lat)) %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE)) %>%
             rename(Date = OCCUR_DATE, Time = OCCUR_TIME)
NYPD_tidy <- NYPD_tidy %>% full_join(BORO_pop)
NYPD_tidy$STATISTICAL_MURDER_FLAG = as.numeric(NYPD_tidy$STATISTICAL_MURDER_FLAG )
NYPD_tidy
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 23,568 x 15
```

```
##    INCIDENT_KEY Date       Time  BORO  PRECINCT JURISDICTION_CO~ LOCATION_DESC
##           <dbl> <date>     <tim> <chr>    <dbl>            <dbl> <chr>
##  1    201575314 2019-08-23 22:10 QUEE~      103                0 <NA>
##  2    205748546 2019-11-27 15:54 BRONX       40                0 <NA>
##  3    193118596 2019-02-02 19:40 MANH~       23                0 <NA>
##  4    204192600 2019-10-24 00:52 STAT~      121                0 PVT HOUSE
##  5    201483468 2019-08-22 18:03 BRONX       46                0 <NA>
##  6    198255460 2019-06-07 17:50 BROO~       73                0 <NA>
##  7    194570529 2019-03-11 16:30 BROO~       81                0 <NA>
##  8    203211777 2019-10-03 01:45 BROO~       67                0 MULTI DWELL ~
##  9    193694863 2019-02-17 03:00 QUEE~      114                2 MULTI DWELL ~
## 10    199582060 2019-07-10 02:56 BROO~       69                0 <NA>
## # ... with 23,558 more rows, and 8 more variables:
## #   STATISTICAL_MURDER_FLAG <dbl>, PERP_AGE_GROUP <chr>, PERP_SEX <chr>,
## #   PERP_RACE <chr>, VIC_AGE_GROUP <chr>, VIC_SEX <chr>, VIC_RACE <chr>,
## #   Population <dbl>
```

Graph analysis

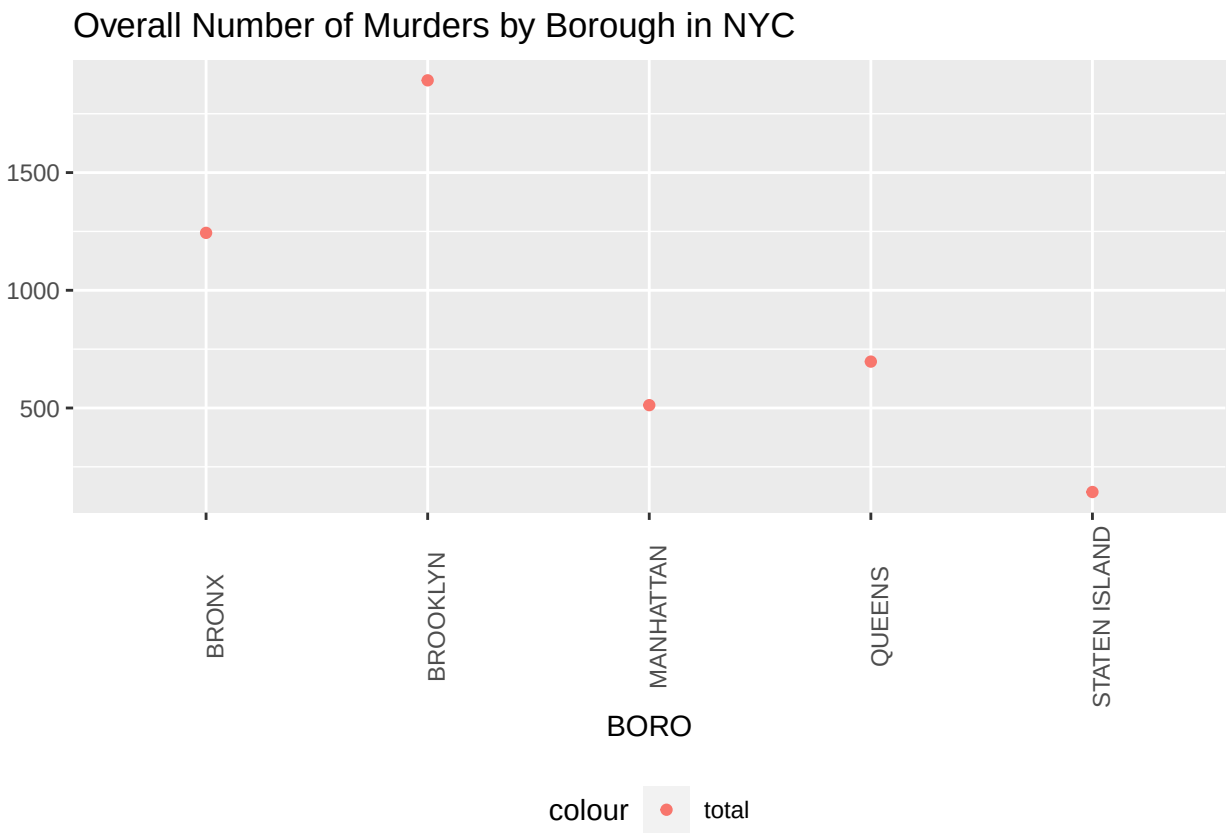
Here we have two graphs. The first one represents the number of murder by gun by Borough in NYC. The second graph allows us to visualize the number of victims of gun violence by race in NYC

```
Crime_Boro <- NYPD_tidy %>% group_by(BORO, PRECINCT, Date, Population) %>%
  summarize(deaths = sum(STATISTICAL_MURDER_FLAG)) %>%
  select(Date, BORO, PRECINCT, Population, deaths) %>%
  ungroup() %>% group_by(BORO) %>% summarize(total = sum(deaths)) %>% ungroup()
Crime_Boro
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?
```

```
## # A tibble: 5 x 2
##   BORO          total
##   <chr>         <dbl>
## 1 BRONX          1244
## 2 BROOKLYN       1892
## 3 MANHATTAN       512
## 4 QUEENS          697
## 5 STATEN ISLAND   143
```

```
Crime_Boro %>% ggplot(aes(x = BORO, y = total)) + geom_point(aes(color = "total")) +
  theme(legend.position = "bottom", axis.text.x = element_text(angle = 90)) +
  labs(title = "Overall Number of Murders by Borough in NYC", y = NULL)
```



Analysis of the graph

As we can see, some Boroughs are overly represented in the statistics. Staten Island has the fewest murder by guns in NYC, whereas Brookly has the highest. To have a better understanding and analysis of this fact, we can ask ourselves many questions. For instance, it would be important here to see if there is a correlation with the gun violence, and the average household income. We know that there tends to be more crime in poor areas, so we could see if this is true here. We could also see if these areas tend to under report gun crimes. Perhaps there is more gun crimes in Staten Island, but the inhabitants do not want to report it. Finally, perhaps these areas have less cops, so the inhabitants of the dangerous boroughs have no other choice to defend themselves.

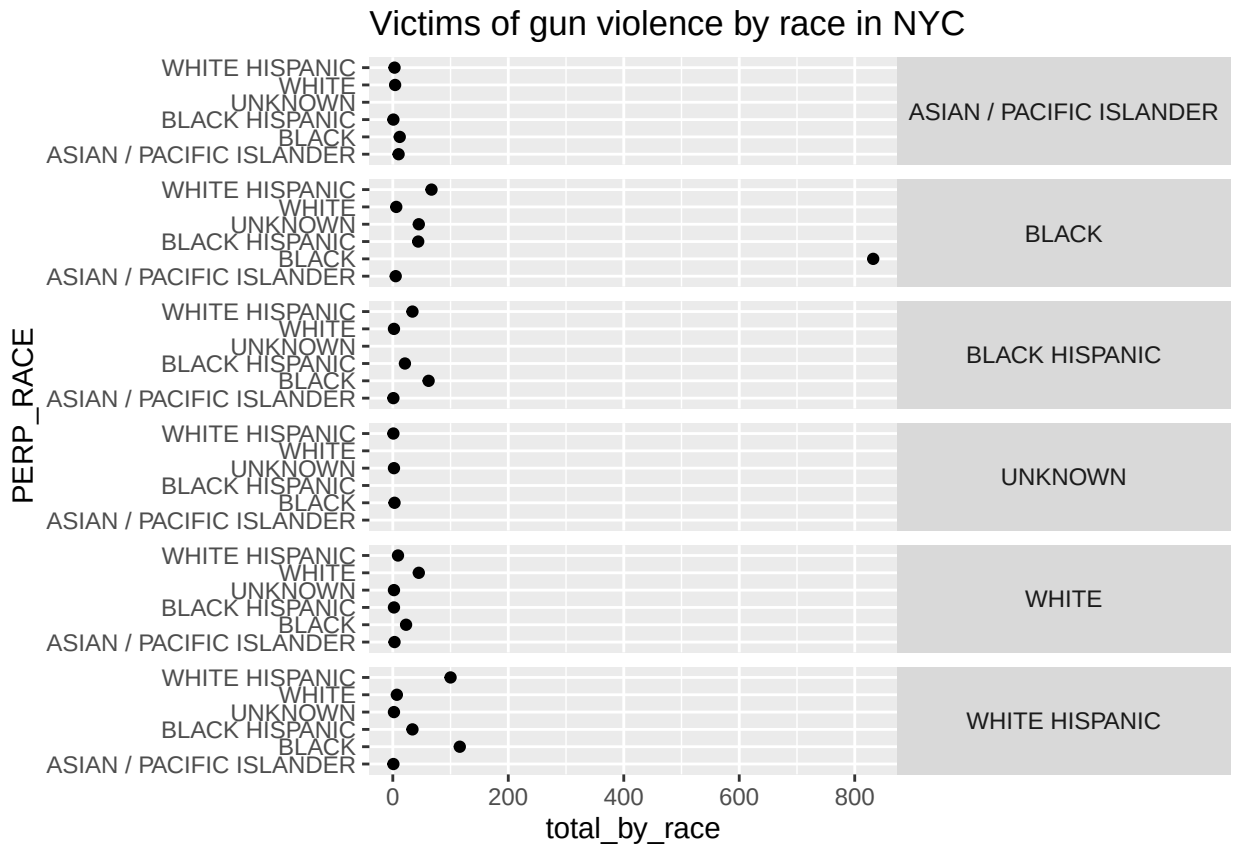
```
Crime_Race <- NYPD_tidy %>% filter(STATISTICAL_MURDER_FLAG >=1) %>% na.omit(PERP_RACE)%>% group_by(PERP_
  summarize(deaths = sum(STATISTICAL_MURDER_FLAG)) %>% select(PERP_RACE, VIC_RACE, Date, deaths) %>% ung
  group_by(PERP_RACE, VIC_RACE) %>% summarize(total_by_race = sum(deaths))
Crime_Race
```

```
## Warning: `...` is not empty.
##
## We detected these problematic arguments:
## * `needs_dots`
##
## These dots only exist to allow future extensions and should be empty.
## Did you misspecify an argument?

## # A tibble: 31 x 3
```

```
## # Groups:   PERP_RACE [6]
##    PERP_RACE                VIC_RACE                 total_by_race
##    <chr>                    <chr>                            <dbl>
##  1 ASIAN / PACIFIC ISLANDER ASIAN / PACIFIC ISLANDER            10
##  2 ASIAN / PACIFIC ISLANDER BLACK                                5
##  3 ASIAN / PACIFIC ISLANDER BLACK HISPANIC                       1
##  4 ASIAN / PACIFIC ISLANDER WHITE                                3
##  5 ASIAN / PACIFIC ISLANDER WHITE HISPANIC                       1
##  6 BLACK                    ASIAN / PACIFIC ISLANDER            12
##  7 BLACK                    BLACK                              832
##  8 BLACK                    BLACK HISPANIC                      62
##  9 BLACK                    UNKNOWN                              3
## 10 BLACK                    WHITE                               23
## # ... with 21 more rows
```

```
Crime_Race %>% ggplot(aes(total_by_race, PERP_RACE)) + geom_point() + facet_grid(rows = vars(Crime_Race$
  theme(strip.text.y =  element_text(angle = 0)) + labs(title = "Victims of gun violence by race in NYC
```



Analysis of this graph

This graph reads: " around 100 White Hispanic have killed Black Hispanic with a gun in NYC" (column 2) Once more, we can ask ourselves many questions. Is it possible that some murders have not been reported ? We could do a cross check by seeing who tends to live in which Boroughs. As we've seen earlier, some are more prone to gun violence than the rest; so perhaps economic inequality plays a role here. Finally, one should not generalize entire race or ethnicity by seeing this graph.Perhaps the source of the data is biased.

Linear model

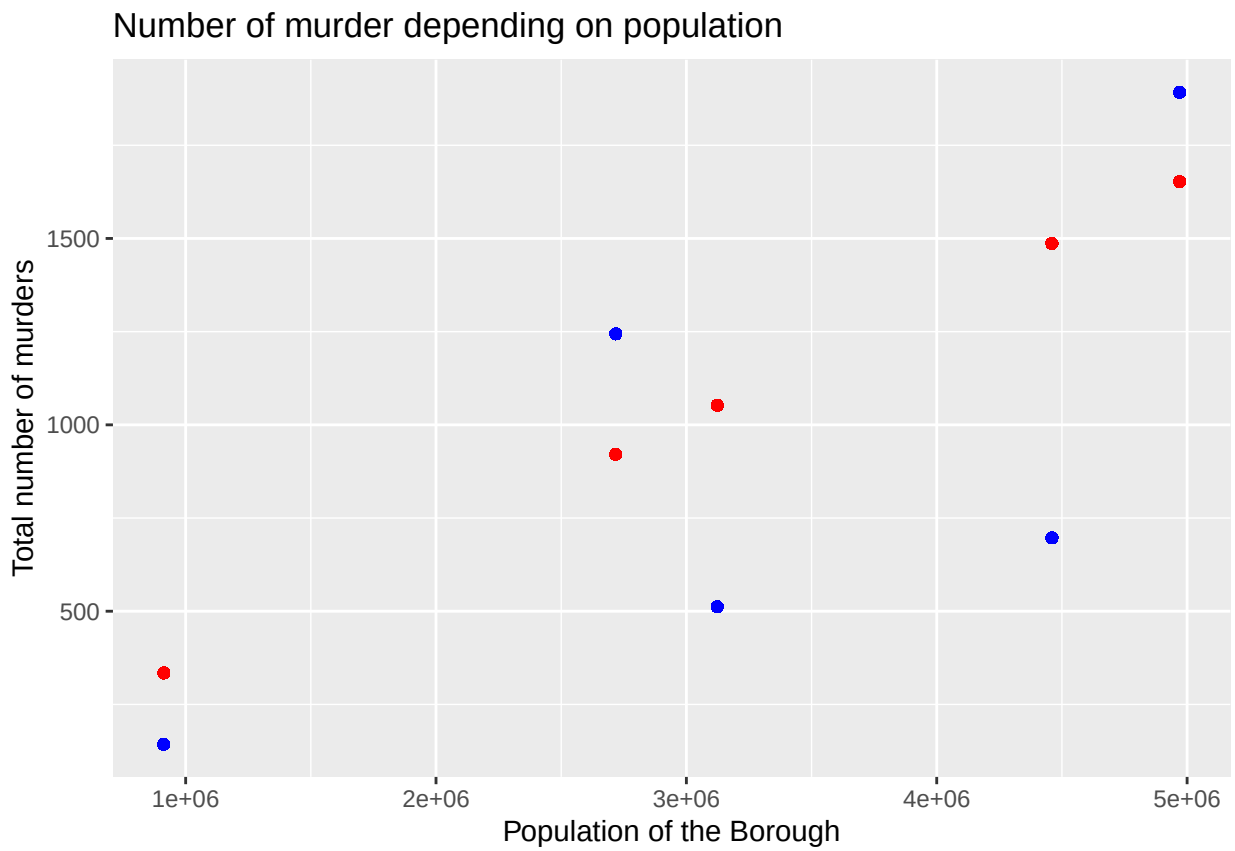
```
New_df <- NYPD_tidy %>% full_join(Crime_Boro)
```

```
## Joining, by = "BORO"
```

```
mod <- lm(total ~ Population, data = New_df)
summary(mod)
```

```
##
## Call:
## lm(formula = total ~ Population, data = New_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -789.6 -540.4  239.8  323.2  323.2
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.815e+01  1.013e+01   3.766 0.000166 ***
## Population  3.248e-04  2.494e-06 130.237  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 429 on 23566 degrees of freedom
## Multiple R-squared:  0.4185, Adjusted R-squared:  0.4185
## F-statistic: 1.696e+04 on 1 and 23566 DF,  p-value: < 2.2e-16
```

```
NYPD_w_pred <- New_df %>% mutate(pred = predict(mod))
NYPD_w_pred %>% ggplot() + geom_point(aes(x = Population, y = total ), color = "blue")+
  geom_point(aes(x = Population, y = pred), color ="red") + labs(title = "Number of murder depending on
  ylab("Total number of murders") + xlab("Population of the Borough")
```



Analysis of this model

I used a linear regression to see if the number of murder goes up with the population. The model is in red in this graph. As we can see with this graph, the correlation isn't as straightforward as one might think. We can infer from this graph that we must use other parameters in our model (like the average income or if the area is well connected with the police force and so on).

Conclusion

In conclusion, we could do a more thorough analysis with other variables and checking the correlation. However, it is really important that this type of graph should not be used to stigmatize entire population. Here, Racism and prejudice represent the two biggest biases, as they would use such a graphic to reinforce their preconceptions. As for one of my personal bias, as a French this type of categorizing data seem really strange to me, as any kind of race statistics are banned in France. Also, since we don't have guns, this type of problem seem foreign to me.

```
sessionInfo()
```

```
## R version 4.0.0 (2020-04-24)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19041)
##
## Matrix products: default
```

```
##
## locale:
## [1] LC_COLLATE=French_France.1252  LC_CTYPE=French_France.1252
## [3] LC_MONETARY=French_France.1252 LC_NUMERIC=C
## [5] LC_TIME=French_France.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] forcats_0.5.0   stringr_1.4.0   dplyr_1.0.2     purrr_0.3.4
##  [5] readr_1.3.1     tidyr_1.1.0     tibble_3.0.1    ggplot2_3.3.0
##  [9] tidyverse_1.3.0 lubridate_1.7.8
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.1.0  xfun_0.19         haven_2.3.0       colorspace_1.4-1
##  [5] vctrs_0.3.5       generics_0.1.0    htmltools_0.5.1.1 yaml_2.2.1
##  [9] utf8_1.1.4        rlang_0.4.10      pillar_1.4.7      glue_1.4.1
## [13] withr_2.3.0       DBI_1.1.0         dbplyr_2.0.0      modelr_0.1.8
## [17] readxl_1.3.1      lifecycle_0.2.0   munsell_0.5.0     gtable_0.3.0
## [21] cellranger_1.1.0  rvest_0.3.6       evaluate_0.14     labeling_0.4.2
## [25] knitr_1.30        curl_4.3          fansi_0.4.1       broom_0.7.2
## [29] Rcpp_1.0.4.6      scales_1.1.1      backports_1.1.6   jsonlite_1.7.2
## [33] farver_2.0.3      fs_1.4.1          hms_0.5.3         digest_0.6.25
## [37] stringi_1.4.6     grid_4.0.0        cli_2.2.0         tools_4.0.0
## [41] magrittr_2.0.1    crayon_1.3.4      pkgconfig_2.0.3   ellipsis_0.3.1
## [45] xml2_1.3.2        reprex_0.3.0      assertthat_0.2.1  rmarkdown_2.5
## [49] httr_1.4.2        rstudioapi_0.13   R6_2.5.0          compiler_4.0.0
```