Shooting Incidents NYC

—-

11/25/2021

Load Data

For this report we use data from the Data Repository of the US government. We load one csv file.

```
## Get current Data
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
```

With this url we now read in the data.

```
NYPD <- read_csv(url)
```

We use glimps and given that we immediately observe many NAs we check for missing values using sapply. As a next step we convert OCCUR_DATE to date.

```
glimpse(NYPD)
```

```
## Rows: 23,585
## Columns: 19
## $ INCIDENT_KEY            <dbl> 24050482, 77673979, 203350417, 80584527, 90843~
## $ OCCUR_DATE              <chr> "08/27/2006", "03/11/2011", "10/06/2019", "09/~
## $ OCCUR_TIME              <time> 05:35:00, 12:03:00, 01:09:00, 03:35:00, 21:16~
## $ BORO                    <chr> "BRONX", "QUEENS", "BROOKLYN", "BRONX", "QUEEN~
## $ PRECINCT                <dbl> 52, 106, 77, 40, 100, 67, 77, 81, 101, 106, 71~
## $ JURISDICTION_CODE       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ LOCATION_DESC           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ STATISTICAL_MURDER_FLAG <lgl> TRUE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE~
## $ PERP_AGE_GROUP          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ PERP_SEX                <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ PERP_RACE               <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA~
## $ VIC_AGE_GROUP           <chr> "25-44", "65+", "18-24", "<18", "18-24", "<18"~
## $ VIC_SEX                 <chr> "F", "M", "F", "M", "M", "M", "M", "M", "M", "~
## $ VIC_RACE                <chr> "BLACK HISPANIC", "WHITE", "BLACK", "BLACK", "~
## $ X_COORD_CD              <dbl> 1017542, 1027543, 995325, 1007453, 1041267, 10~
## $ Y_COORD_CD              <dbl> 255918.9, 186095.0, 185155.0, 233952.0, 157133~
## $ Latitude                <dbl> 40.86906, 40.67737, 40.67489, 40.80880, 40.597~
## $ Longitude               <dbl> -73.87963, -73.84392, -73.96008, -73.91618, -7~
## $ Lon_Lat                 <chr> "POINT (-73.87963173099996 40.86905819000003)"~
```

```r
NYPD %>%
  summarise(count = sum(is.na(NYPD)))
```

```
## # A tibble: 1 x 1
##   count
##   <int>
## 1 38400
```

```r
  sapply(NYPD, function(x) sum(is.na(x)))
```

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                       0                       0                       0
##                    BORO                PRECINCT       JURISDICTION_CODE
##                       0                       0                       2
##           LOCATION_DESC STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP
##                   13581                       0                    8295
##                PERP_SEX               PERP_RACE           VIC_AGE_GROUP
##                    8261                    8261                       0
##                 VIC_SEX                VIC_RACE              X_COORD_CD
##                       0                       0                       0
##              Y_COORD_CD                Latitude               Longitude
##                       0                       0                       0
##                 Lon_Lat
##                       0
```

```r
#Convert date chr to date
NYPD <- NYPD %>% mutate(OCCUR_DATE = mdy(OCCUR_DATE))
```

Data quality is overall good. The fact that perpetrator data is missing seems reasonable, it would be interesting by what criteria the data is provided (can it be based on testimonies, or are only identified (arrested) perpetrators considered?)

Geospatial Display

We want to better understand where those shootings occured and therefore display incidents as dots on a ggplot map (entire history available)

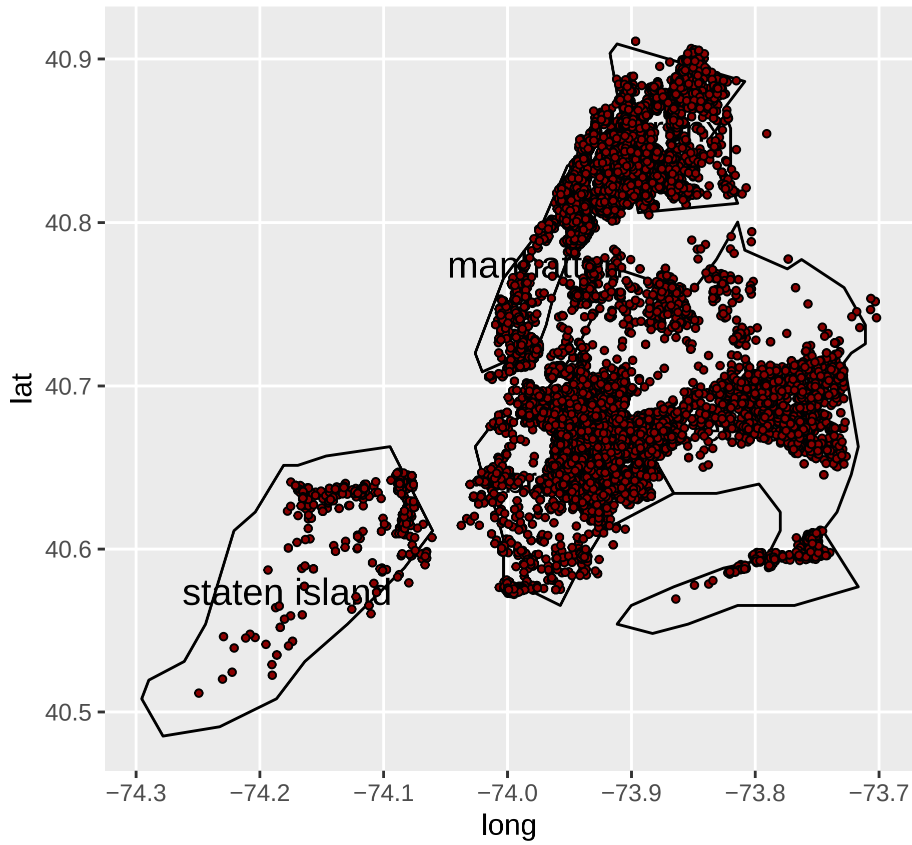
```r
### get all counties of the State of New York and filter for the 5 boroughs of New York City
counties <- map_data("county","New York")
counties <- as_tibble(counties)
nyc <- c("bronx", "kings","new york", "queens", "richmond")
counties <- counties %>%
  filter(subregion %in% nyc)

### rename the counties to borough names
 counties <- counties %>%
   mutate(subregion = replace(subregion, subregion == "kings", "brooklyn")) %>%
   mutate(subregion = replace(subregion, subregion == "richmond", "staten island")) %>%
   mutate(subregion = replace(subregion, subregion == "new york", "manhattan"))

# create centered names for the map
```

```
cnames <- aggregate(cbind(long, lat) ~ subregion, data=counties,FUN=function(x)mean(range(x)))

#plot the map of NYC
ggplot(counties, aes(long, lat)) +
    geom_polygon(aes(group=group), colour='black', fill=NA) +
    geom_text(data=cnames, aes(long, lat, label = subregion), size=5) +
    coord_map() +
  geom_point(data = NYPD, aes(x = Longitude, y = Latitude), size = 1,
        shape = 21, fill = "darkred")
```

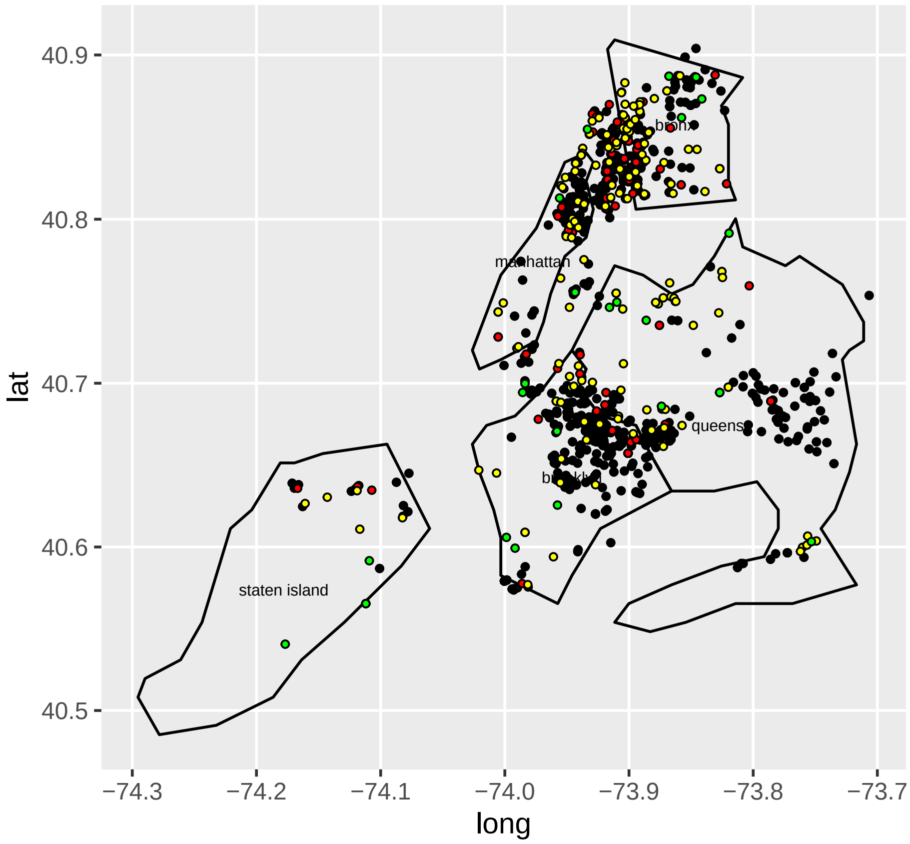


There are simply too many data points and given that we want to visualize other attributes provided in the report as well, we display data just for one year.

```
#plot the map of NYC and shootings for selected year

NYPD_selected <- NYPD %>% filter(OCCUR_DATE > "2018-12-31" & OCCUR_DATE < "2020-01-01")
NYPD_selected_black <- NYPD_selected %>% filter(VIC_RACE == "BLACK")
NYPD_selected_black_hispanic <- NYPD_selected %>% filter(VIC_RACE == "BLACK HISPANIC")
NYPD_selected_white <- NYPD_selected %>% filter(VIC_RACE == "WHITE")
NYPD_selected_white_hispanic <- NYPD_selected %>% filter(VIC_RACE == "WHITE HISPANIC")
  ggplot(counties, aes(long, lat)) +
    geom_polygon(aes(group=group), colour='black', fill=NA) +
    geom_text(data=cnames, aes(long, lat, label = subregion), size=2) +
    coord_map() +
  geom_point(data = NYPD_selected_black, aes(x = Longitude, y = Latitude), size = 1,
        shape = 21, fill = "black") +
```

```
  geom_point(data = NYPD_selected_black_hispanic, aes(x = Longitude, y = Latitude), size = 1,
        shape = 21, fill = "red") + geom_point(data = NYPD_selected_white_hispanic, aes(x = Longitude, y
        shape = 21, fill = "yellow") + geom_point(data = NYPD_selected_white, aes(x = Longitude, y = Lat
        shape = 21, fill = "green")
```



We marked the data by race of the victims: * black dots stand for blacks * red for black hispanics * green for whites and * yellow for white hispanics.

Overall we see incidents spread widely accross all NYC, with staten island being less affected. Obviously black and white hispanic together with black hispanic persons represent the vast majority of victims.

Time Series

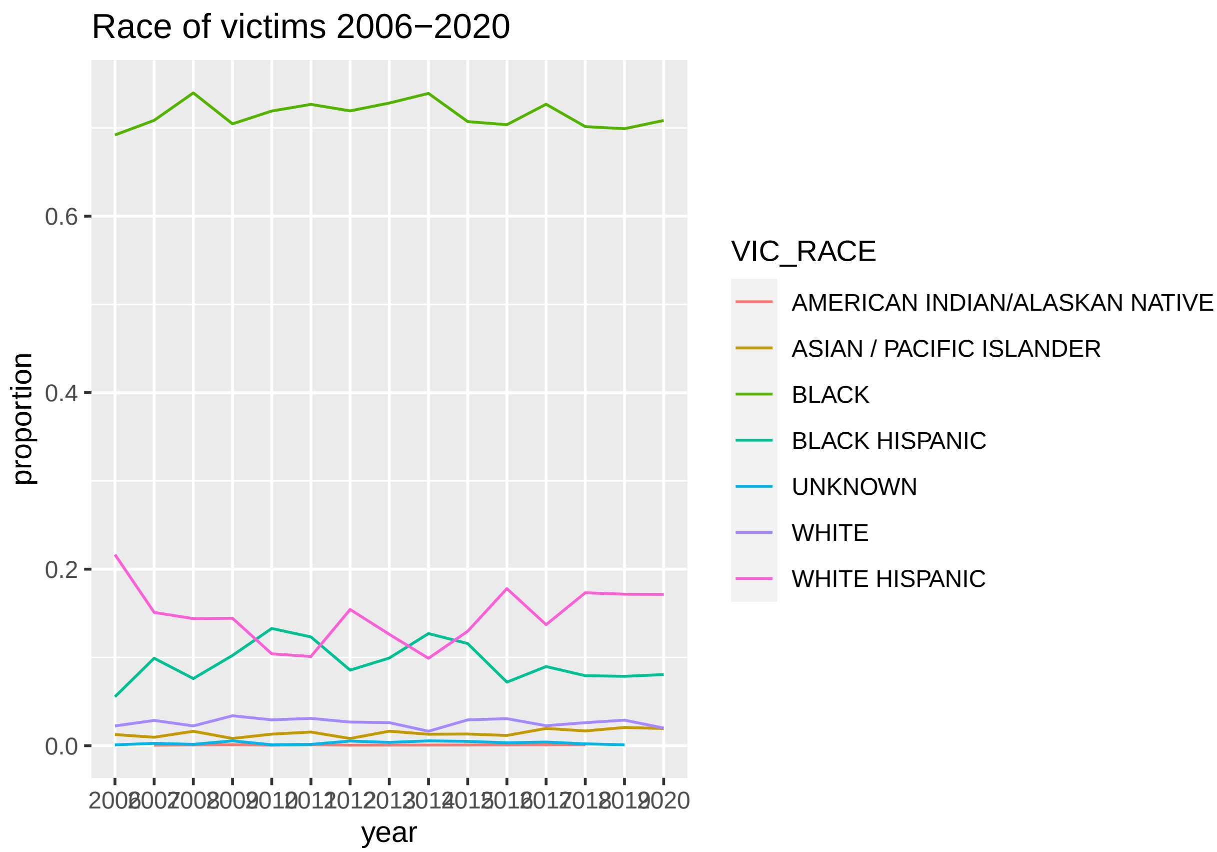
Now we are going to group the data by race of victim and plot the timeseries.

```
#select INCIDENT_KEY, OCCUR_DATE, BORO, PERP RACE and VIC RACE
NYPD_race <- NYPD %>% select ("INCIDENT_KEY","OCCUR_DATE", "VIC_RACE") %>%
  mutate(OCCUR_DATE = format(OCCUR_DATE, "%Y")) %>%
  group_by(OCCUR_DATE, VIC_RACE) %>%
  summarise(n = n()) %>%
  mutate(prop = n / sum(n))
```

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

```
#Plot the timeseries
ggplot(NYPD_race,aes(x=OCCUR_DATE, y=prop, group=VIC_RACE, color=VIC_RACE)) +
    geom_line() +
  ggtitle("Race of victims 2006-2020") +

    ylab("proportion") + xlab("year")
```



The high percentage of black victims is striking. Given that according to the demographic data provided by data.io the biggest group by race is white (non-hispanic) with 32% contrasted by just 22% black - demographic data obviously does not explain the distribution of victims by race in the shooting incident report.

Model the missing perpetrator race using a multinomial regression with categorical predictors

We saw that more than 8000 records did not show the perpetrators race. But we observed in addition to the high percentage of black victims also a very high share of black perpetrators. In general same race incidents where perpetrator has the same race as the victim prevails for all races (except for BLACK HISPANICS).

Please see a summary below:

```
#PERP RACE and VIC RACE
NYPD_race_total <- NYPD %>% select ("INCIDENT_KEY","OCCUR_DATE", "PERP_RACE", "VIC_RACE") %>%
  mutate(OCCUR_DATE = format(OCCUR_DATE, "%Y")) %>%
  group_by(PERP_RACE, VIC_RACE) %>%
```

```
  summarise(n = n()) %>%
  mutate(prop = n / sum(n))
NYPD_race_total %>% print(n = Inf)
```

```
## # A tibble: 45 x 4
## # Groups:   PERP_RACE [8]
##    PERP_RACE                      VIC_RACE                           n     prop
##    <chr>                          <chr>                          <int>    <dbl>
##  1 AMERICAN INDIAN/ALASKAN NATIVE BLACK                              2 1
##  2 ASIAN / PACIFIC ISLANDER       ASIAN / PACIFIC ISLANDER          39 0.320
##  3 ASIAN / PACIFIC ISLANDER       BLACK                             39 0.320
##  4 ASIAN / PACIFIC ISLANDER       BLACK HISPANIC                    12 0.0984
##  5 ASIAN / PACIFIC ISLANDER       WHITE                             11 0.0902
##  6 ASIAN / PACIFIC ISLANDER       WHITE HISPANIC                    21 0.172
##  7 BLACK                          AMERICAN INDIAN/ALASKAN NATIVE     4 0.000399
##  8 BLACK                          ASIAN / PACIFIC ISLANDER         126 0.0126
##  9 BLACK                          BLACK                           7975 0.796
## 10 BLACK                          BLACK HISPANIC                   687 0.0685
## 11 BLACK                          UNKNOWN                           24 0.00239
## 12 BLACK                          WHITE                            165 0.0165
## 13 BLACK                          WHITE HISPANIC                  1044 0.104
## 14 BLACK HISPANIC                 ASIAN / PACIFIC ISLANDER          17 0.0155
## 15 BLACK HISPANIC                 BLACK                            448 0.409
## 16 BLACK HISPANIC                 BLACK HISPANIC                   279 0.255
## 17 BLACK HISPANIC                 UNKNOWN                            5 0.00456
## 18 BLACK HISPANIC                 WHITE                             33 0.0301
## 19 BLACK HISPANIC                 WHITE HISPANIC                   314 0.286
## 20 UNKNOWN                        AMERICAN INDIAN/ALASKAN NATIVE     3 0.00163
## 21 UNKNOWN                        ASIAN / PACIFIC ISLANDER          16 0.00871
## 22 UNKNOWN                        BLACK                           1359 0.740
## 23 UNKNOWN                        BLACK HISPANIC                   155 0.0844
## 24 UNKNOWN                        UNKNOWN                            6 0.00327
## 25 UNKNOWN                        WHITE                             42 0.0229
## 26 UNKNOWN                        WHITE HISPANIC                   255 0.139
## 27 WHITE                          ASIAN / PACIFIC ISLANDER          11 0.0431
## 28 WHITE                          BLACK                             29 0.114
## 29 WHITE                          BLACK HISPANIC                    18 0.0706
## 30 WHITE                          UNKNOWN                            1 0.00392
## 31 WHITE                          WHITE                            151 0.592
## 32 WHITE                          WHITE HISPANIC                    45 0.176
## 33 WHITE HISPANIC                 ASIAN / PACIFIC ISLANDER          32 0.0161
## 34 WHITE HISPANIC                 BLACK                            648 0.326
## 35 WHITE HISPANIC                 BLACK HISPANIC                   352 0.177
## 36 WHITE HISPANIC                 UNKNOWN                           11 0.00553
## 37 WHITE HISPANIC                 WHITE                             84 0.0423
## 38 WHITE HISPANIC                 WHITE HISPANIC                   861 0.433
## 39 <NA>                           AMERICAN INDIAN/ALASKAN NATIVE     2 0.000242
## 40 <NA>                           ASIAN / PACIFIC ISLANDER          86 0.0104
## 41 <NA>                           BLACK                           6369 0.771
## 42 <NA>                           BLACK HISPANIC                   742 0.0898
## 43 <NA>                           UNKNOWN                           18 0.00218
## 44 <NA>                           WHITE                            134 0.0162
## 45 <NA>                           WHITE HISPANIC                   910 0.110
```

Using a model will help to understand if based on the data available shooting incidents involving people of same race are highly likely. We load again the full data set and will use borough, race of victim and sex of victim to predict the race of the perpetrator.

```r
url <- "https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD"
data.raw <- read.csv(url, stringsAsFactors = TRUE)
keeps <- c("PERP_RACE", "BORO", "VIC_RACE", "VIC_SEX")
data.raw <- data.raw[keeps]
train <- filter(data.raw, data.raw$PERP_RACE == "BLACK" | data.raw$PERP_RACE == "WHITE" | data.raw$PERP_
#Train the model
train <- droplevels(train)

model <- nnet::multinom(PERP_RACE ~., data = train)
```

```
## # weights:  70 (52 variable)
## initial  value 21704.879687
## iter  10 value 10807.183776
## iter  20 value 10039.505805
## iter  30 value 9816.774761
## iter  40 value 9625.664352
## iter  50 value 9267.086744
## final  value 9265.093761
## converged
```

```r
print(summary(model))
```

```
## Call:
## nnet::multinom(formula = PERP_RACE ~ ., data = train)
##
## Coefficients:
##                 (Intercept) BOROBROOKLYN BOROMANHATTAN BOROQUEENS
## BLACK             7.9118458    0.3154263     0.6825386 -0.7797258
## BLACK HISPANIC    0.7429062   -0.7372130     0.5306910 -1.7319666
## WHITE             0.8104980    0.2570805     0.8222006 -0.2706723
## WHITE HISPANIC    0.4861163   -0.7757124     0.4265348 -1.1577733
##                 BOROSTATEN ISLAND VIC_RACEASIAN / PACIFIC ISLANDER VIC_RACEBLACK
## BLACK                   0.1125366                       -6.9851658     -3.006480
## BLACK HISPANIC         -1.4147992                       -1.4247524      1.638532
## WHITE                   1.0388629                       -2.3111381     -1.463978
## WHITE HISPANIC         -0.7224033                       -0.5594103      2.308543
##                 VIC_RACEBLACK HISPANIC VIC_RACEUNKNOWN VIC_RACEWHITE
## BLACK                       -4.3183021       0.9767291    -5.5543411
## BLACK HISPANIC               2.0625521       6.9720511     0.4669812
## WHITE                       -0.7790796       5.1452190     1.4020066
## WHITE HISPANIC               2.6377337       8.0279764     1.6475863
##                 VIC_RACEWHITE HISPANIC  VIC_SEXM  VIC_SEXU
## BLACK                       -4.3912275 0.4729121  7.997041
## BLACK HISPANIC               1.7410284 0.7394862 -2.452513
## WHITE                       -0.3938387 0.2420567 -1.108506
## WHITE HISPANIC               3.0660291 0.6148265 -3.265150
##
## Std. Errors:
##                 (Intercept) BOROBROOKLYN BOROMANHATTAN BOROQUEENS
```

```
## BLACK              8.418764    0.2676437     0.4099650  0.2544361
## BLACK HISPANIC    15.097696    0.2777220     0.4160402  0.2740809
## WHITE             15.073865    0.3312708     0.4753548  0.3276562
## WHITE HISPANIC    16.421603    0.2732941     0.4131233  0.2602054
##                BOROSTATEN ISLAND VIC_RACEASIAN / PACIFIC ISLANDER VIC_RACEBLACK
## BLACK                  0.5471478                         8.417561      8.416907
## BLACK HISPANIC         0.5997592                        15.098218     15.096310
## WHITE                  0.5931525                        15.075142     15.073314
## WHITE HISPANIC         0.5604364                        16.421355     16.420406
##                VIC_RACEBLACK HISPANIC VIC_RACEUNKNOWN VIC_RACEWHITE
## BLACK                        8.419487        20.36536      8.420241
## BLACK HISPANIC              15.097692        23.59174     15.098820
## WHITE                       15.075245        23.59432     15.074026
## WHITE HISPANIC              16.421659        24.33334     16.422161
##                VIC_RACEWHITE HISPANIC  VIC_SEXM VIC_SEXU
## BLACK                        8.417805 0.2590881 84.46288
## BLACK HISPANIC              15.096796 0.2786668 26.66671
## WHITE                       15.073305 0.3131165 32.10285
## WHITE HISPANIC              16.420814 0.2673411 23.54248
##
## Residual Deviance: 18530.19
## AIC: 18634.19
```

```
# Make predictions
predicted.classes <- model %>% predict(train)
head(predicted.classes)
```

```
## [1] BLACK BLACK BLACK BLACK BLACK BLACK
## 5 Levels: ASIAN / PACIFIC ISLANDER BLACK BLACK HISPANIC ... WHITE HISPANIC
```

```
# Model accuracy
mean(predicted.classes == train$PERP_RACE)
```

```
## [1] 0.739211
```

```
data.raw$predicted.PERP_RACE<-model %>% predict(data.raw)
head(data.raw)
```

```
##   PERP_RACE     BORO       VIC_RACE VIC_SEX predicted.PERP_RACE
## 1              BRONX BLACK HISPANIC       F               BLACK
## 2             QUEENS          WHITE       M               WHITE
## 3           BROOKLYN          BLACK       F               BLACK
## 4              BRONX          BLACK       M               BLACK
## 5             QUEENS          BLACK       M               BLACK
## 6           BROOKLYN          BLACK       M               BLACK
```

Please see below the grouped data using the predicted perpetrator race:

```
data.raw.tbl <- tibble(data.raw)
#PERP RACE PREDICTED and VIC RACE
NYPD_race_total_predicted <- data.raw.tbl %>% select ("predicted.PERP_RACE","PERP_RACE", "VIC_RACE") %>%
```

```
  group_by(predicted.PERP_RACE,VIC_RACE) %>%
  summarise(n = n()) %>%
  mutate(prop = n / sum(n))
NYPD_race_total_predicted %>% print(n = Inf)
```

```
## # A tibble: 9 x 4
## # Groups:   predicted.PERP_RACE [3]
##   predicted.PERP_RACE VIC_RACE                           n     prop
##   <fct>               <fct>                          <int>    <dbl>
## 1 BLACK               AMERICAN INDIAN/ALASKAN NATIVE     9 0.000413
## 2 BLACK               ASIAN / PACIFIC ISLANDER         327 0.0150
## 3 BLACK               BLACK                          16869 0.775
## 4 BLACK               BLACK HISPANIC                  2245 0.103
## 5 BLACK               UNKNOWN                           65 0.00299
## 6 BLACK               WHITE                            403 0.0185
## 7 BLACK               WHITE HISPANIC                  1848 0.0849
## 8 WHITE               WHITE                            217 1
## 9 WHITE HISPANIC      WHITE HISPANIC                  1602 1
```

Model accuracy on training data was just 74%, therefore results need to be interpreted cautiously. Nevertheless we see that the percentage of incidents with a black perpetrator and a black victims stays almost the same as for the original data. For white perpetrators and hispanic perpetrators the model predicts a 100% probabiltiy that the victim is of the same race.

Bias

Being a non-US citizen my bias is driven by news coverage of gun-violence and crime by various international media. In the recent month the focus in the news was on police violence against black people. Regarding the analysis, I personally believe that the statistics of the NYPD, especially when it comes to race of victims are credible. There might be a bias when it comes to perpetrators. Nevertheless the observed high percentage of black persons being either victim or perpetrator is known in the US as black-on-black violence phenomenon. The latter one gets supported by the model (multinomial regression with categorical predictors) performed aboved. The shooting incidents are obviously not related to the demographic composition of NYC by race.

```
sessionInfo()
```

```
## R version 4.1.1 (2021-08-10)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19042)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_Switzerland.1252  LC_CTYPE=English_Switzerland.1252
## [3] LC_MONETARY=English_Switzerland.1252 LC_NUMERIC=C
## [5] LC_TIME=English_Switzerland.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
```

```
##  [1] mapproj_1.2.7   maps_3.4.0      lubridate_1.8.0 forcats_0.5.1
##  [5] dplyr_1.0.7     purrr_0.3.4     readr_2.1.0     tidyr_1.1.4
##  [9] tibble_3.1.6    ggplot2_3.3.5   tidyverse_1.3.1 stringr_1.4.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.7       assertthat_0.2.1 digest_0.6.28    utf8_1.2.2
##  [5] R6_2.5.1         cellranger_1.1.0 backports_1.3.0  reprex_2.0.1
##  [9] evaluate_0.14    highr_0.9        httr_1.4.2       pillar_1.6.4
## [13] rlang_0.4.12     curl_4.3.2       readxl_1.3.1     rstudioapi_0.13
## [17] rmarkdown_2.11   labeling_0.4.2   bit_4.0.4        munsell_0.5.0
## [21] broom_0.7.10     compiler_4.1.1   modelr_0.1.8     xfun_0.27
## [25] pkgconfig_2.0.3  htmltools_0.5.2  nnet_7.3-16      tidyselect_1.1.1
## [29] fansi_0.5.0      crayon_1.4.2     tzdb_0.2.0       dbplyr_2.1.1
## [33] withr_2.4.2      grid_4.1.1       jsonlite_1.7.2   gtable_0.3.0
## [37] lifecycle_1.0.1  DBI_1.1.1        magrittr_2.0.1   scales_1.1.1
## [41] cli_3.1.0        stringi_1.7.5    vroom_1.5.6      farver_2.1.0
## [45] fs_1.5.0         xml2_1.3.2       ellipsis_0.3.2   generics_0.1.1
## [49] vctrs_0.3.8      tools_4.1.1      bit64_4.0.5      glue_1.4.2
## [53] hms_1.1.1        parallel_4.1.1   fastmap_1.1.0    yaml_2.2.1
## [57] colorspace_2.0-2 rvest_1.0.2      knitr_1.36       haven_2.4.3
```