NYPD - Shooting project

NYPD Shooting analysis and prediction

This project is made for the course DTSA-5301 Data Science in the field.

Libraries to use

First, let's import the libraries I will use for this project.

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(stringr)
library(lubridate)
library(ggplot2)
```

Loading data

In the first step, let's load the data from the data source.

```
url = 'https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD'

nypd_dataframe = read.csv(url)
head(nypd_dataframe)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    228798151 05/27/2021   21:30:00   QUEENS                        105
## 2    137471050 06/27/2014   17:40:00    BRONX                         40
## 3    147998800 11/21/2015   03:56:00   QUEENS                        108
## 4    146837977 10/09/2015   18:30:00    BRONX                         44
## 5     58921844 02/19/2009   22:58:00    BRONX                         47
## 6    219559682 10/21/2020   21:36:00 BROOKLYN                         81
```

```
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0                                                   false
## 2                 0                                                   false
## 3                 0                                                    true
## 4                 0                                                   false
## 5                 0                                                    true
## 6                 0                                                    true
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                                          18-24       M          BLACK
## 2                                          18-24       M          BLACK
## 3                                          25-44       M          WHITE
## 4                                            <18       M WHITE HISPANIC
## 5          25-44        M     BLACK         45-64       M          BLACK
## 6                                          25-44       M          BLACK
##   X_COORD_CD Y_COORD_CD Latitude Longitude
## 1    1058925   180924.0 40.66296 -73.73084
## 2    1005028   234516.0 40.81035 -73.92494
## 3    1007668   209836.5 40.74261 -73.91549
## 4    1006537   244511.1 40.83778 -73.91946
## 5    1024922   262189.4 40.88624 -73.85291
## 6    1004234   186461.7 40.67846 -73.92795
##                                    Lon_Lat
## 1 POINT (-73.73083868899994 40.662964620000025)
## 2  POINT (-73.92494232599995 40.81035186300006)
## 3  POINT (-73.91549174199997 40.74260663300004)
## 4  POINT (-73.91945661499994 40.83778200300003)
## 5  POINT (-73.85290950899997 40.88623791800006)
## 6 POINT (-73.92795224099996 40.678456718000064)
```

First lets look if there are null variables.

```
apply(is.na(nypd_dataframe), 2, sum)
```

```
##            INCIDENT_KEY             OCCUR_DATE             OCCUR_TIME
##                       0                      0                      0
##                    BORO      LOC_OF_OCCUR_DESC               PRECINCT
##                       0                      0                      0
##       JURISDICTION_CODE     LOC_CLASSFCTN_DESC          LOCATION_DESC
##                       2                      0                      0
## STATISTICAL_MURDER_FLAG         PERP_AGE_GROUP               PERP_SEX
##                       0                      0                      0
##               PERP_RACE          VIC_AGE_GROUP               VIC_SEX
##                       0                      0                      0
##               VIC_RACE             X_COORD_CD             Y_COORD_CD
##                       0                      0                      0
##               Latitude              Longitude               Lon_Lat
##                      10                     10                      0
```

Look! Latitude and Longitude and Jurisdiction code has nulls. Let's replace it.

```
nypd_dataframe <- nypd_dataframe %>%
  mutate(Longitude = if_else(is.na(Longitude), mean(Longitude), Longitude),
         Latitude = if_else(is.na(Latitude), mean(Latitude), Latitude),
         JURISDICTION_CODE = if_else(is.na(JURISDICTION_CODE), mean(JURISDICTION_CODE), JURISDICTION_CO
```

Visualizing the data

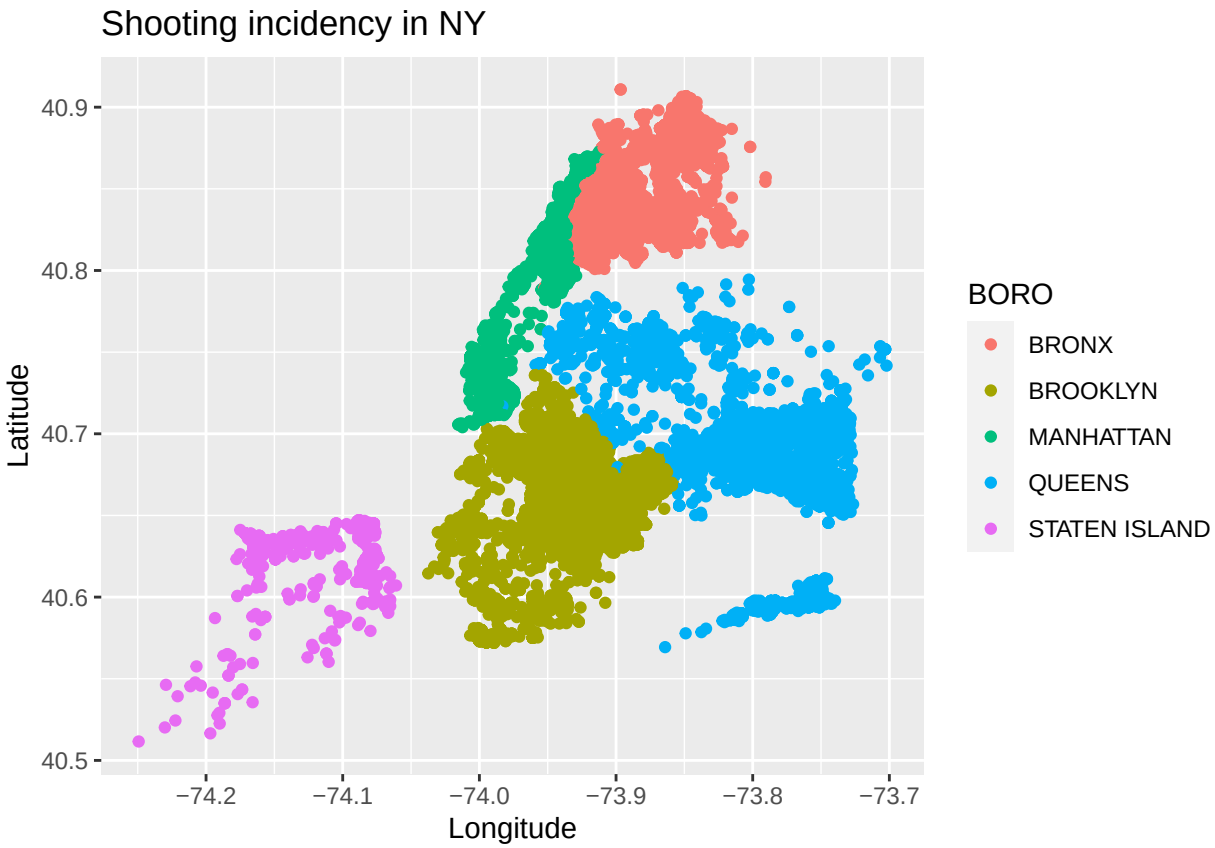
For this project I aim to make a predictor of how much I am in danger giving my geo position, race, sex, and age. So... Let's explore the data.

Geo located data

First, let's see which district of NY has more shootings.

```
ggplot(nypd_dataframe, aes(x=Longitude, y=Latitude, color=BORO)) +
  geom_point() +
  ggtitle("Shooting incidency in NY")
```

```
## Warning: Removed 10 rows containing missing values (`geom_point()`).
```



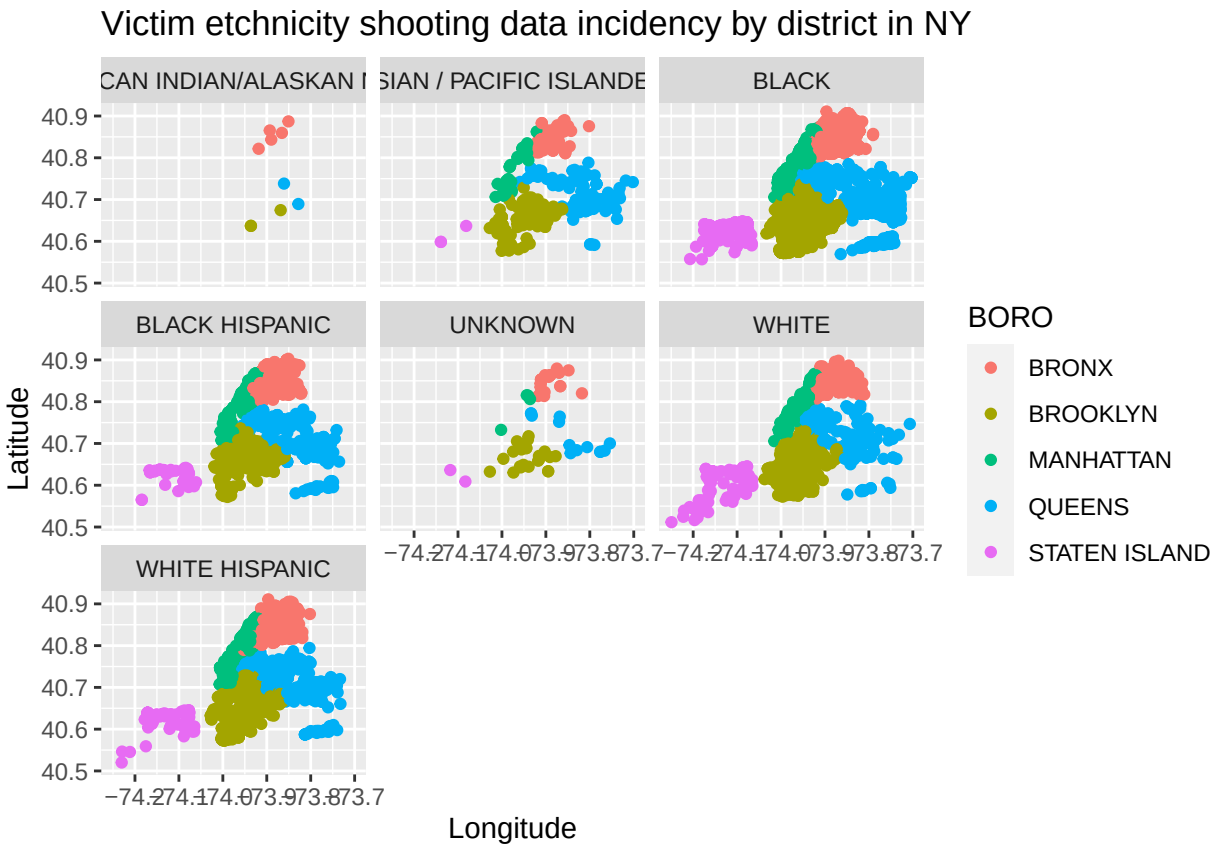
By ethnicity

What about my race in certain zones?

```
ggplot(nypd_dataframe, aes(x=Longitude, y=Latitude, color=BORO)) +
  geom_point() +
  facet_wrap(~VIC_RACE) +
  ggtitle("Victim etchnicity shooting data incidency by district in NY")
```

Victim data

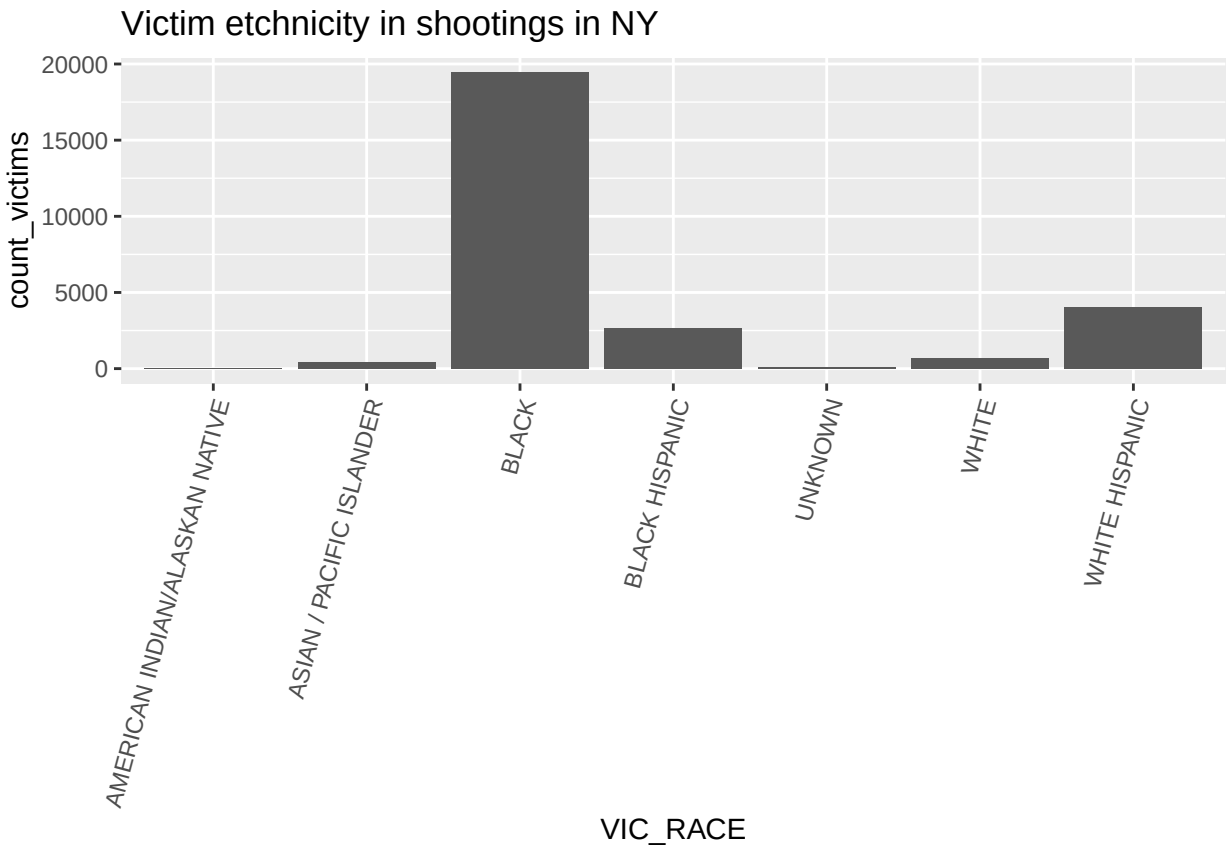
```
## Warning: Removed 10 rows containing missing values (‘geom_point()‘).
```




```
count_victims_by_race = nypd_dataframe %>%
  group_by(VIC_RACE) %>%
  summarize(count_victims = n()) %>%
  arrange(count_victims)
count_victims_by_race
```

```
## # A tibble: 7 x 2
##   VIC_RACE                         count_victims
##   <chr>                                    <int>
## 1 AMERICAN INDIAN/ALASKAN NATIVE              10
## 2 UNKNOWN                                     66
## 3 ASIAN / PACIFIC ISLANDER                   404
## 4 WHITE                                      698
## 5 BLACK HISPANIC                            2646
## 6 WHITE HISPANIC                            4049
## 7 BLACK                                    19439
```

```
ggplot(count_victims_by_race, aes(x = VIC_RACE, y = count_victims)) +
  geom_bar(stat='identity') +
  theme(axis.text.x = element_text(angle = 75, vjust = 1, hjust=1)) +
  ggtitle("Victim etchnicity in shootings in NY")
```

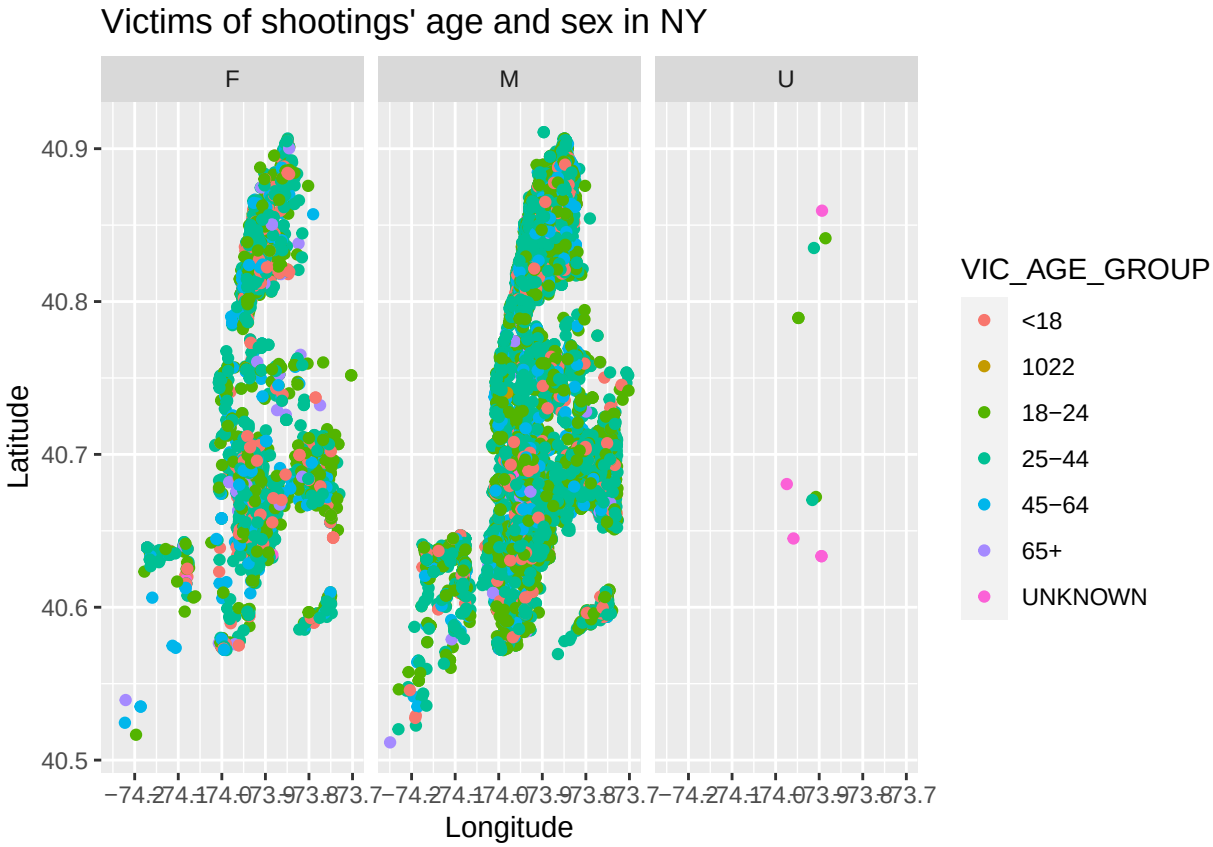


This previous plot shows which races recieved more shootings in NY. To make a ratio based prediction it is possible to support this barplot with the total number of people in NY by each race. But this is isolated to this project.

Age of victims and sex The next scatter plot shows that Male people around 18 to 44 years are more in danger to get shooted in NY.

```
ggplot(nypd_dataframe, aes(x=Longitude, y=Latitude, color=VIC_AGE_GROUP)) +
  geom_point() +
  facet_wrap(~VIC_SEX) +
  ggtitle("Victims of shootings' age and sex in NY")
```

```
## Warning: Removed 10 rows containing missing values (‘geom_point()‘).
```

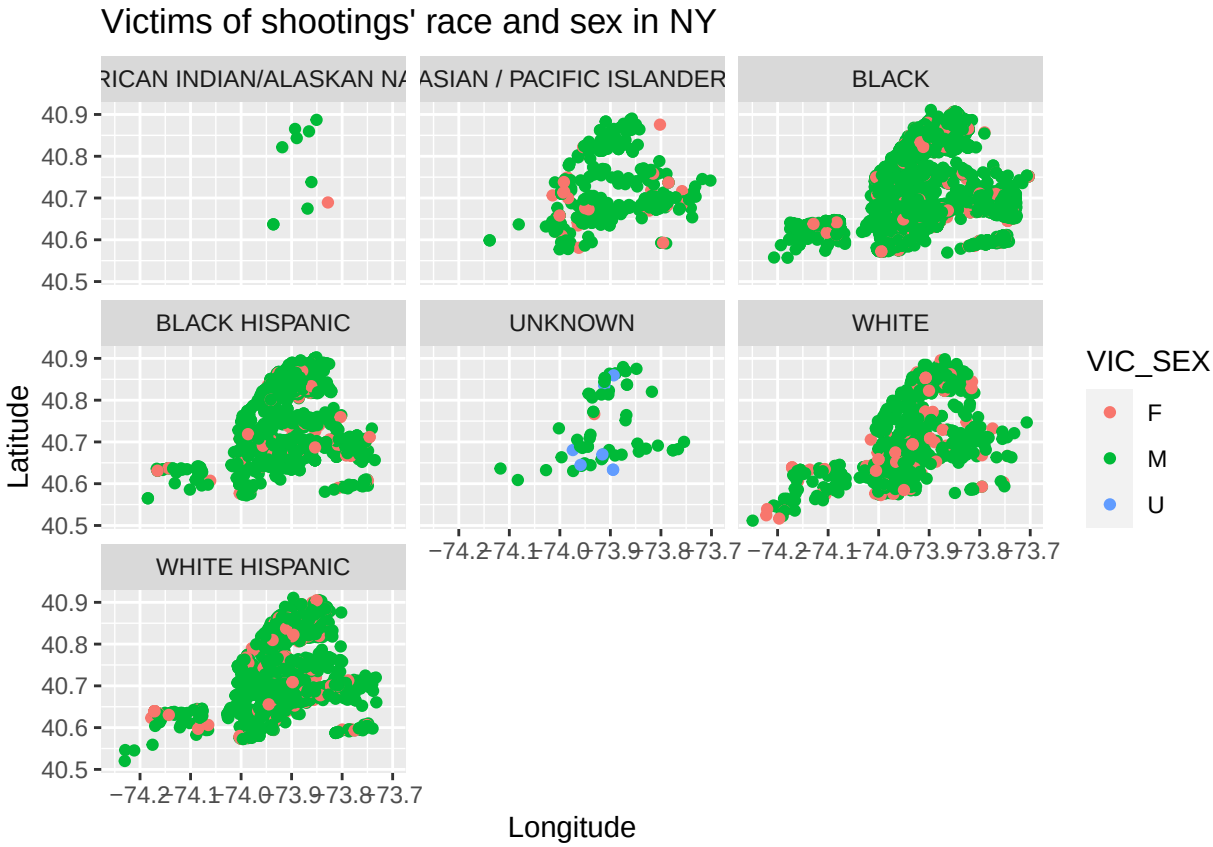


Victims' Sex and Race

Again looking at the victims ethnicity an sex, in the next plot you can see there are four ethnicities which are more victims of shootings (Hispanic Black, Hispanic White, Black, and White).

```
ggplot(nypd_dataframe, aes(x=Longitude, y=Latitude, color=VIC_SEX)) +
  geom_point() +
  facet_wrap(~VIC_RACE) +
  ggtitle("Victims of shootings' race and sex in NY")
```

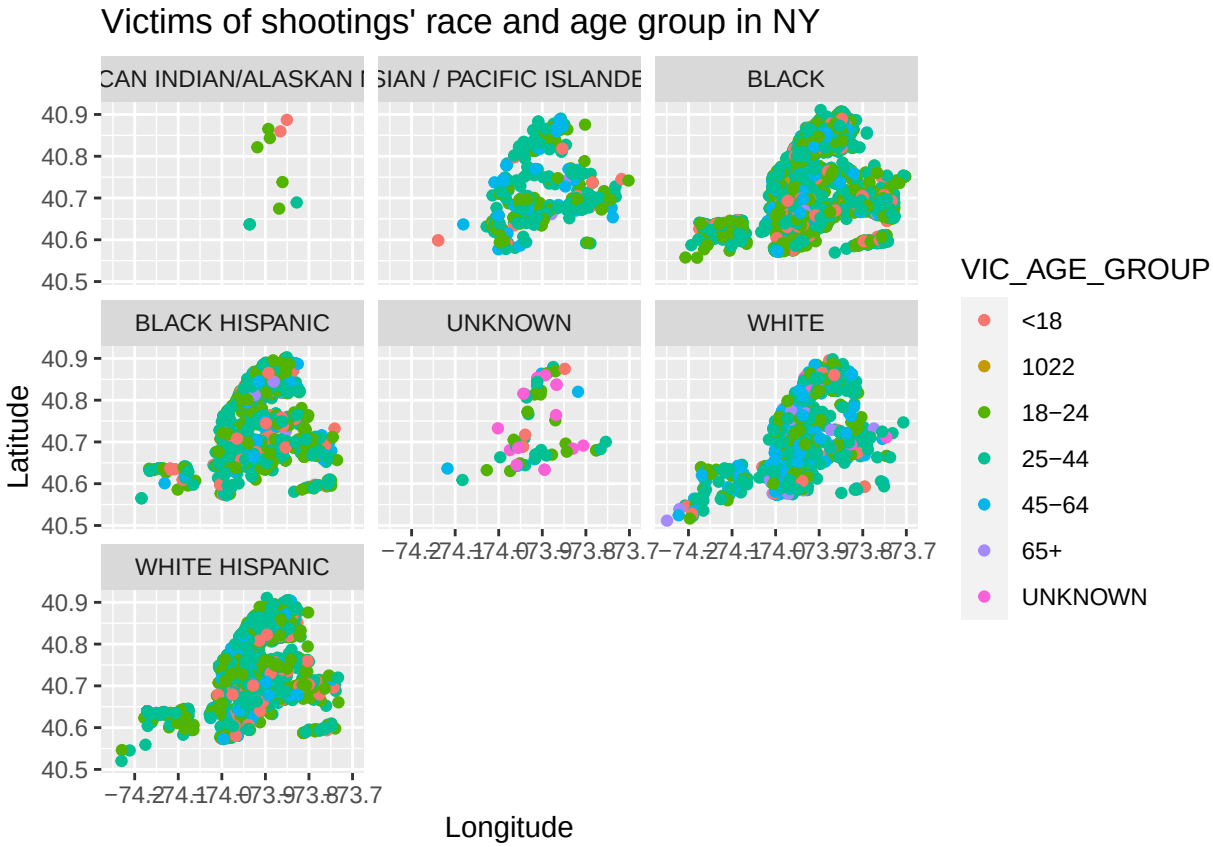
```
## Warning: Removed 10 rows containing missing values (‘geom_point()‘).
```



Victims Age and Race

```
ggplot(nypd_dataframe, aes(x=Longitude, y=Latitude, color=VIC_AGE_GROUP)) +
  geom_point() +
  facet_wrap(~VIC_RACE) +
  ggtitle("Victims of shootings' race and age group in NY")
```

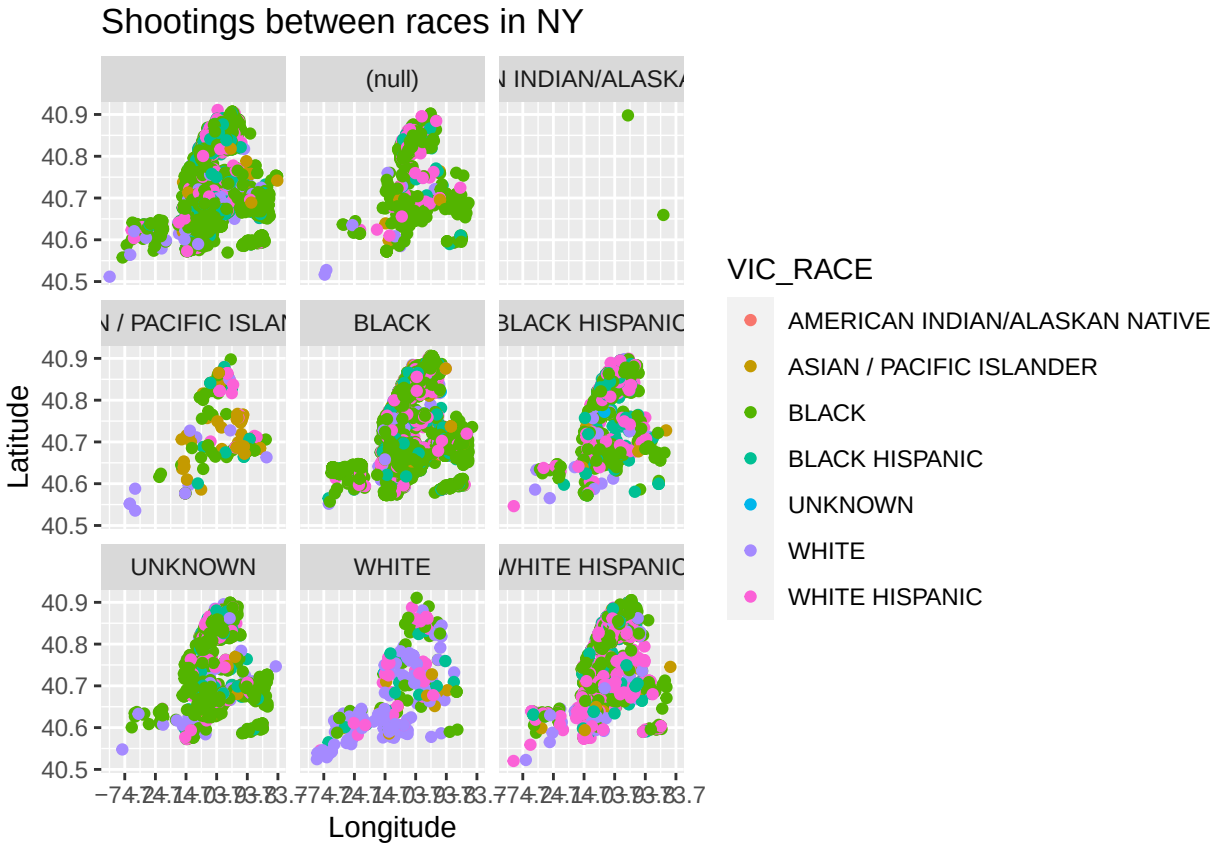
```
## Warning: Removed 10 rows containing missing values (‘geom_point()‘).
```



Shooter's data vs Victim's race The next plot could be interpreted by "which race is proned to get shoot by the races near by". This is unfair, but there are several assumptions you can assume looking it, but for ethic reasons we will not make any conclusion about it.

```
ggplot(nypd_dataframe, aes(x=Longitude, y=Latitude, color=VIC_RACE)) +
  geom_point() +
  facet_wrap(~PERP_RACE) +
  ggtitle("Shootings between races in NY")
```

```
## Warning: Removed 10 rows containing missing values (`geom_point()`).
```

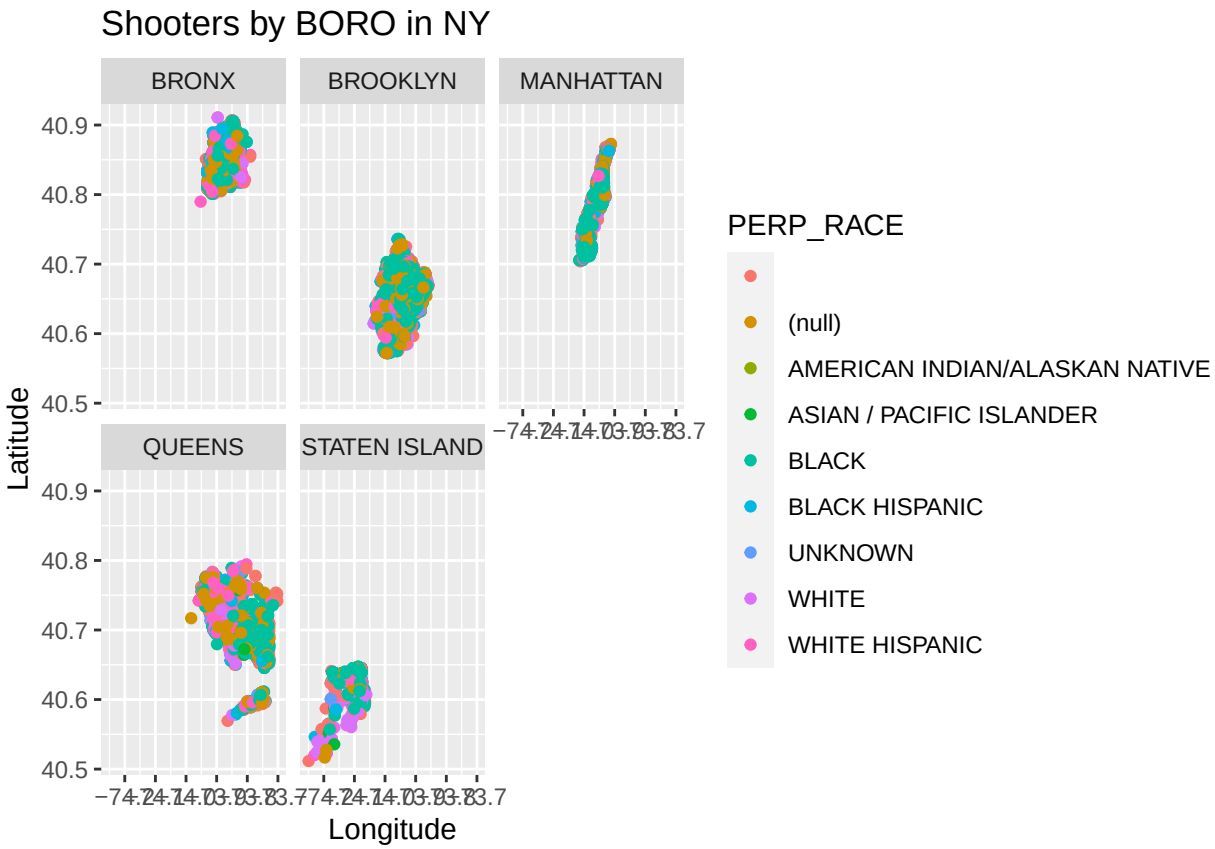


Shooters by BORO

The next plot shows which shooters are most common in each zone of NY.

```
ggplot(nypd_dataframe, aes(x=Longitude, y=Latitude, color=PERP_RACE)) +
  geom_point() +
  facet_wrap(~BORO) +
  ggtitle("Shooters by BORO in NY")
```

```
## Warning: Removed 10 rows containing missing values (`geom_point()`).
```



Discusion

There are several assumptions of the early visuallization. First, there are several important fields, such as BORO, Latitude, Longitude, etc, but we only consider for this study case BORO and Latitude and Longitude

Ethic Disclaimer

Some insights of this data may make the people think some races are more violent than others. This is FALSE. The true interpretation is made by more than early visualization made with less data than others. There are several other sources of data such as money invested in some areas that are important to development a "District" which is dangerous or with the absurd idea "some race" is dangerous than other, which is not. In this case we use the available data with the end to predict which district is dangerous than other or the possibility you are in danger in a certain area of NY. I will not use etchtnicy data because I consider this as a bias, there are several other data which makes the possibility of get shoot and as we can see in the model development.

The danger button predictor

We use a logistic regression for the predictor, and for educational purposes our target is if we are near or we are part of a shooting, we will survive to it. Specifically, we are talking to the column STATISTICAL_MURDER_FLAG.

First let's replace char variables to factors.

```
nypd_dataframe <- nypd_dataframe %>%
  mutate(VIC_SEX <- as.factor(VIC_SEX),
         VIC_RACE <- as.factor(VIC_RACE),
         VIC_AGE_GROUP <- as.factor(VIC_AGE_GROUP),
         BORO <- as.factor(BORO)
         )
head(nypd_dataframe)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    228798151 05/27/2021   21:30:00   QUEENS                       105
## 2    137471050 06/27/2014   17:40:00    BRONX                        40
## 3    147998800 11/21/2015   03:56:00   QUEENS                       108
## 4    146837977 10/09/2015   18:30:00    BRONX                        44
## 5     58921844 02/19/2009   22:58:00    BRONX                        47
## 6    219559682 10/21/2020   21:36:00 BROOKLYN                        81
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0                                                   false
## 2                 0                                                   false
## 3                 0                                                    true
## 4                 0                                                   false
## 5                 0                                                    true
## 6                 0                                                    true
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                                          18-24       M          BLACK
## 2                                          18-24       M          BLACK
## 3                                          25-44       M          WHITE
## 4                                            <18       M WHITE HISPANIC
## 5          25-44        M     BLACK         45-64       M          BLACK
## 6                                          25-44       M          BLACK
##   X_COORD_CD Y_COORD_CD Latitude Longitude
## 1    1058925   180924.0 40.66296 -73.73084
## 2    1005028   234516.0 40.81035 -73.92494
## 3    1007668   209836.5 40.74261 -73.91549
## 4    1006537   244511.1 40.83778 -73.91946
## 5    1024922   262189.4 40.88624 -73.85291
## 6    1004234   186461.7 40.67846 -73.92795
##                                     Lon_Lat VIC_SEX <- as.factor(VIC_SEX)
## 1 POINT (-73.73083868899994 40.662964620000025)                          M
## 2  POINT (-73.92494232599995 40.81035186300006)                          M
## 3  POINT (-73.91549174199997 40.74260663300004)                          M
## 4  POINT (-73.91945661499994 40.83778200300003)                          M
## 5  POINT (-73.85290950899997 40.88623791800006)                          M
## 6 POINT (-73.92795224099996 40.678456718000064)                          M
##   VIC_RACE <- as.factor(VIC_RACE) VIC_AGE_GROUP <- as.factor(VIC_AGE_GROUP)
## 1                           BLACK                                     18-24
## 2                           BLACK                                     18-24
## 3                           WHITE                                     25-44
## 4                  WHITE HISPANIC                                       <18
## 5                           BLACK                                     45-64
## 6                           BLACK                                     25-44
##   BORO <- as.factor(BORO)
## 1                  QUEENS
## 2                   BRONX
```

```
## 3                  QUEENS
## 4                   BRONX
## 5                   BRONX
## 6                BROOKLYN
```

Next, let's cast the target to 1 or 0.

```
labeler_murder <- function(val){
  if(val == 'true'){
    return(TRUE)
  }
  return(FALSE)
}
```

```
nypd_dataframe <- nypd_dataframe %>%
  mutate(STATISTICAL_MURDER_FLAG = sapply(STATISTICAL_MURDER_FLAG, labeler_murder))
head(nypd_dataframe)
```

```
##   INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1    228798151 05/27/2021   21:30:00   QUEENS                       105
## 2    137471050 06/27/2014   17:40:00    BRONX                        40
## 3    147998800 11/21/2015   03:56:00   QUEENS                       108
## 4    146837977 10/09/2015   18:30:00    BRONX                        44
## 5     58921844 02/19/2009   22:58:00    BRONX                        47
## 6    219559682 10/21/2020   21:36:00 BROOKLYN                        81
##   JURISDICTION_CODE LOC_CLASSFCTN_DESC LOCATION_DESC STATISTICAL_MURDER_FLAG
## 1                 0                                                  FALSE
## 2                 0                                                  FALSE
## 3                 0                                                   TRUE
## 4                 0                                                  FALSE
## 5                 0                                                   TRUE
## 6                 0                                                   TRUE
##   PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP VIC_SEX       VIC_RACE
## 1                                          18-24       M          BLACK
## 2                                          18-24       M          BLACK
## 3                                          25-44       M          WHITE
## 4                                            <18       M WHITE HISPANIC
## 5          25-44        M     BLACK         45-64       M          BLACK
## 6                                          25-44       M          BLACK
##   X_COORD_CD Y_COORD_CD Latitude Longitude
## 1    1058925   180924.0 40.66296 -73.73084
## 2    1005028   234516.0 40.81035 -73.92494
## 3    1007668   209836.5 40.74261 -73.91549
## 4    1006537   244511.1 40.83778 -73.91946
## 5    1024922   262189.4 40.88624 -73.85291
## 6    1004234   186461.7 40.67846 -73.92795
##                                      Lon_Lat VIC_SEX <- as.factor(VIC_SEX)
## 1 POINT (-73.73083868899994 40.662964620000025)                           M
## 2  POINT (-73.92494232599995 40.81035186300006)                           M
## 3  POINT (-73.91549174199997 40.74260663300004)                           M
## 4  POINT (-73.91945661499994 40.83778200300003)                           M
## 5  POINT (-73.85290950899997 40.88623791800006)                           M
## 6 POINT (-73.92795224099996 40.678456718000064)                           M
```

```
##   VIC_RACE <- as.factor(VIC_RACE) VIC_AGE_GROUP <- as.factor(VIC_AGE_GROUP)
## 1                            BLACK                                    18-24
## 2                            BLACK                                    18-24
## 3                            WHITE                                    25-44
## 4                  WHITE HISPANIC                                      <18
## 5                            BLACK                                    45-64
## 6                            BLACK                                    25-44
##   BORO <- as.factor(BORO)
## 1                   QUEENS
## 2                    BRONX
## 3                   QUEENS
## 4                    BRONX
## 5                    BRONX
## 6                 BROOKLYN
```

```
model <- glm(STATISTICAL_MURDER_FLAG ~ BORO + Latitude + Longitude + VIC_SEX + VIC_RACE + VIC_AGE_GROUP
model
```

```
##
## Call:  glm(formula = STATISTICAL_MURDER_FLAG ~ BORO + Latitude + Longitude +
##     VIC_SEX + VIC_RACE + VIC_AGE_GROUP, family = binomial, data = nypd_dataframe)
##
## Coefficients:
##                       (Intercept)                       BOROBROOKLYN
##                          23.86753                           -0.03512
##                     BOROMANHATTAN                         BOROQUEENS
##                          -0.11923                           -0.08284
##                BOROSTATEN ISLAND                           Latitude
##                           0.07951                           -0.17988
##                         Longitude                           VIC_SEXM
##                           0.39751                           -0.04936
##                          VIC_SEXU  VIC_RACEASIAN / PACIFIC ISLANDER
##                          -0.57958                           11.30357
##                     VIC_RACEBLACK            VIC_RACEBLACK HISPANIC
##                          11.01698                           10.84398
##                   VIC_RACEUNKNOWN                       VIC_RACEWHITE
##                          10.27060                           11.35995
##           VIC_RACEWHITE HISPANIC                  VIC_AGE_GROUP1022
##                          11.14450                          -10.54519
##               VIC_AGE_GROUP18-24                VIC_AGE_GROUP25-44
##                           0.28600                            0.61537
##               VIC_AGE_GROUP45-64                  VIC_AGE_GROUP65+
##                           0.76184                            1.01985
##             VIC_AGE_GROUPUNKNOWN
##                           0.87357
##
## Degrees of Freedom: 27301 Total (i.e. Null);  27281 Residual
##   (10 observations deleted due to missingness)
## Null Deviance:       26780
## Residual Deviance: 26490     AIC: 26530
```

Finally let's evaluate supposing if I am walking around New York.

```
df_test <- data.frame(BORO = c("STATEN ISLAND", "BROOKLYN", "QUEENS", "MANHATTAN"),
                 Longitude = c(-74.15, -74, -73.8, -74),
                 Latitude = c(40.6, 40.6, 40.7, 40.8),
                 VIC_SEX = c("M", "M", "M", "M"),
                 VIC_RACE = c("WHITE HISPANIC", "WHITE HISPANIC", "WHITE HISPANIC", "WHITE HISPANIC"),
                 VIC_AGE_GROUP = c("25-44", "25-44", "25-44", "25-44"))
df_test <- df_test %>%
  mutate(BORO = as.factor(BORO),
         VIC_SEX = as.factor(VIC_SEX),
         VIC_RACE = as.factor(VIC_RACE),
         VIC_AGE_GROUP = as.factor(VIC_AGE_GROUP)
         )
```

```
probabilities <- model %>% predict(df_test, type = "response")
predicted_my_walk_to_NY <- ifelse(probabilities > 0.5, "High Probability to get shoot", "Low Probabilit
predicted_my_walk_to_NY
```

```
##                              1                              2
## "Low Probability to get shoot" "Low Probability to get shoot"
##                              3                              4
## "Low Probability to get shoot" "Low Probability to get shoot"
```

Hey look! In my travel to New York I didn't die by shooting.