# ProjectRmd

Eric Yeung

2024-04-24

## Importing Data

We are going to download the NYPD Shooting Incident data from an URL. After that store the csv file into a data frame. Having a brief look on it. There are 21 variables and 28562 rows in the data set.

```
## Rows: 28562 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (12): OCCUR_DATE, BORO, LOC_OF_OCCUR_DESC, LOC_CLASSFCTN_DESC, LOCATION...
## 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.
```

```
##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:28562       Length:28562      Length:28562
##  1st Qu.: 65439914   Class :character   Class1:hms        Class :character
##  Median : 92711254   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :127405824                      Mode  :numeric
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Mode :logical           Length:28562
##  Class :character   FALSE:23036             Class :character
##  Mode  :character   TRUE :5526              Mode  :character
##
##
##
##
##    PERP_SEX          PERP_RACE         VIC_AGE_GROUP        VIC_SEX
```

```
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59

##   INCIDENT_KEY        OCCUR_DATE         OCCUR_TIME           BORO
##  Min.   :  9953245   Length:28562       Length:28562      Length:28562
##  1st Qu.: 65439914   Class :character   Class1:hms        Class :character
##  Median : 92711254   Mode  :character   Class2:difftime   Mode  :character
##  Mean   :127405824                      Mode  :numeric
##  3rd Qu.:203131993
##  Max.   :279758069
##
##  LOC_OF_OCCUR_DESC     PRECINCT     JURISDICTION_CODE LOC_CLASSFCTN_DESC
##  Length:28562       Min.   :  1.0   Min.   :0.0000    Length:28562
##  Class :character   1st Qu.: 44.0   1st Qu.:0.0000    Class :character
##  Mode  :character   Median : 67.0   Median :0.0000    Mode  :character
##                     Mean   : 65.5   Mean   :0.3219
##                     3rd Qu.: 81.0   3rd Qu.:0.0000
##                     Max.   :123.0   Max.   :2.0000
##                                     NA's   :2
##  LOCATION_DESC      STATISTICAL_MURDER_FLAG PERP_AGE_GROUP
##  Length:28562       Mode :logical           Length:28562
##  Class :character   FALSE:23036             Class :character
##  Mode  :character   TRUE :5526              Mode  :character
##
##
##
##
##    PERP_SEX           PERP_RACE          VIC_AGE_GROUP        VIC_SEX
##  Length:28562       Length:28562       Length:28562       Length:28562
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
```

```
##
##
##    VIC_RACE           X_COORD_CD        Y_COORD_CD        Latitude
##  Length:28562       Min.   : 914928   Min.   :125757   Min.   :40.51
##  Class :character   1st Qu.:1000068   1st Qu.:182912   1st Qu.:40.67
##  Mode  :character   Median :1007772   Median :194901   Median :40.70
##                     Mean   :1009424   Mean   :208380   Mean   :40.74
##                     3rd Qu.:1016807   3rd Qu.:239814   3rd Qu.:40.82
##                     Max.   :1066815   Max.   :271128   Max.   :40.91
##                                                        NA's   :59
##    Longitude        Lon_Lat
##  Min.   :-74.25   Length:28562
##  1st Qu.:-73.94   Class :character
##  Median :-73.92   Mode  :character
##  Mean   :-73.91
##  3rd Qu.:-73.88
##  Max.   :-73.70
##  NA's   :59
```

## Data Cleaning

Having overview of the data set, we found some variables contain null value and some redundant variables Those redundant variables are unrelated to our interest or repeated information.

First we remove those columns and store in a new data frame. We also find that the OCCUR_DATE is not in date format. Change it into data format

Looking at the data, there are a lot of missing data.Most are related to location.To deal with it, we will try to restore the LOC_CLASSFCTN_DESC if there is the same LOCATION_DESC. Otherwise,we will omit the records with missing data. As well as, transform the "(null)" into NULL value. For the PERP related data, the missing data will be recode into UNKNOWN or U

As a result, only 1238 records and 17 variables left.

```
##            INCIDENT_KEY              OCCUR_DATE              OCCUR_TIME
##                       0                       0                       0
##                    BORO       LOC_OF_OCCUR_DESC                PRECINCT
##                       0                   25596                       0
##       JURISDICTION_CODE      LOC_CLASSFCTN_DESC           LOCATION_DESC
##                       2                   25596                   14977
## STATISTICAL_MURDER_FLAG          PERP_AGE_GROUP                PERP_SEX
##                       0                    9344                    9310
##               PERP_RACE           VIC_AGE_GROUP                 VIC_SEX
##                    9310                       0                       0
##                VIC_RACE              X_COORD_CD              Y_COORD_CD
##                       0                       0                       0
##                Latitude               Longitude                 Lon_Lat
##                      59                      59                      59
```

```
##   INCIDENT_KEY         OCCUR_DATE                      OCCUR_TIME
##  Min.   :238531161   Min.   :2022-01-01 00:00:00.00   Length:1238
##  1st Qu.:245864734   1st Qu.:2022-05-06 00:00:00.00   Class1:hms
##  Median :252757824   Median :2022-10-04 00:00:00.00   Class2:difftime
##  Mean   :256436231   Mean   :2022-11-09 11:23:20.41   Mode  :numeric
```

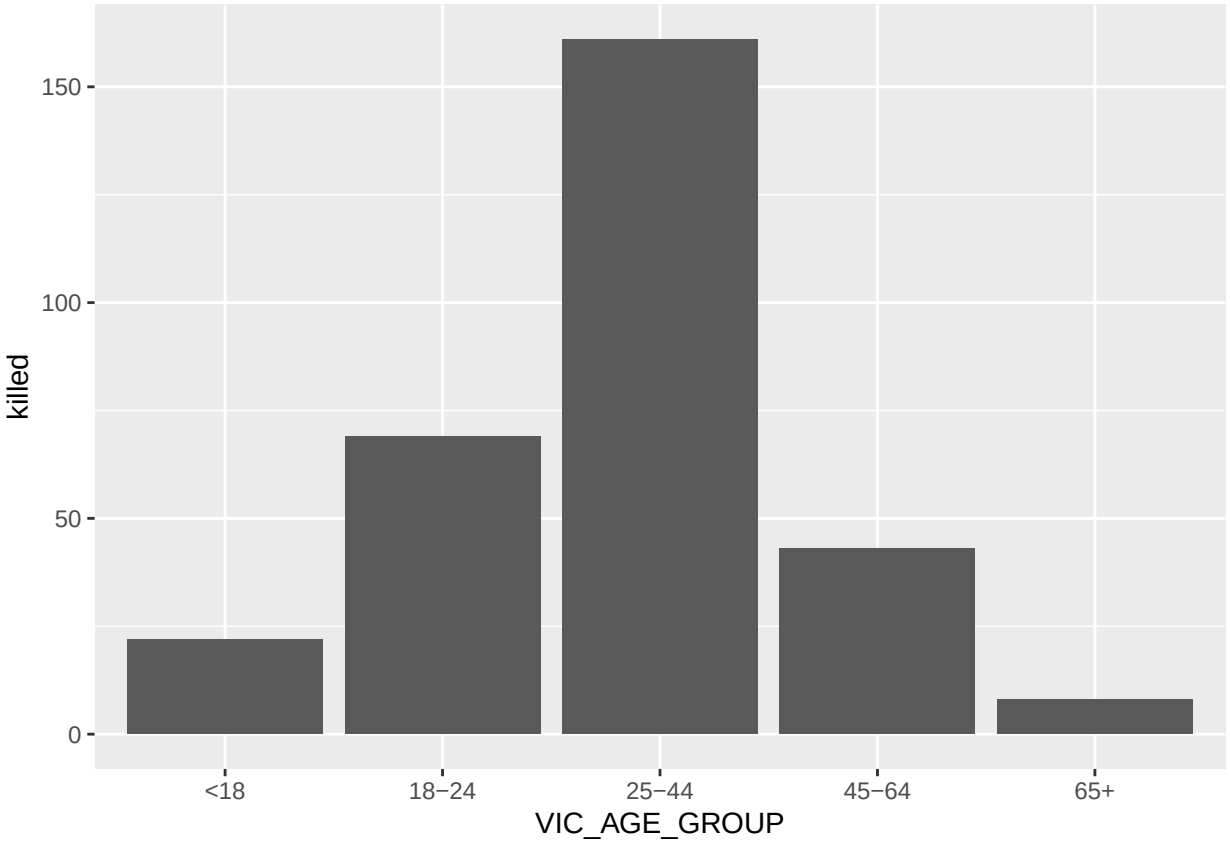
```
##  3rd Qu.:267917707   3rd Qu.:2023-05-07 00:00:00.00
##  Max.   :279683077   Max.   :2023-12-12 00:00:00.00
##                      NA's   :747
##      BORO           LOC_OF_OCCUR_DESC     PRECINCT      JURISDICTION_CODE
##  Length:1238        Length:1238        Min.   :  5.00   Min.   :0.0000
##  Class :character   Class :character   1st Qu.: 42.00   1st Qu.:0.0000
##  Mode  :character   Mode  :character   Median : 61.00   Median :0.0000
##                                        Mean   : 62.93   Mean   :0.5444
##                                        3rd Qu.: 79.00   3rd Qu.:2.0000
##                                        Max.   :123.00   Max.   :2.0000
##
##  LOC_CLASSFCTN_DESC LOCATION_DESC      STATISTICAL_MURDER_FLAG
##  Length:1238        Length:1238        Mode :logical
##  Class :character   Class :character   FALSE:935
##  Mode  :character   Mode  :character   TRUE :303
##
##
##
##
##  PERP_AGE_GROUP       PERP_SEX          PERP_RACE         VIC_AGE_GROUP
##  Length:1238        Length:1238        Length:1238        Length:1238
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    VIC_SEX            VIC_RACE          Lon_Lat
##  Length:1238        Length:1238        Length:1238
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

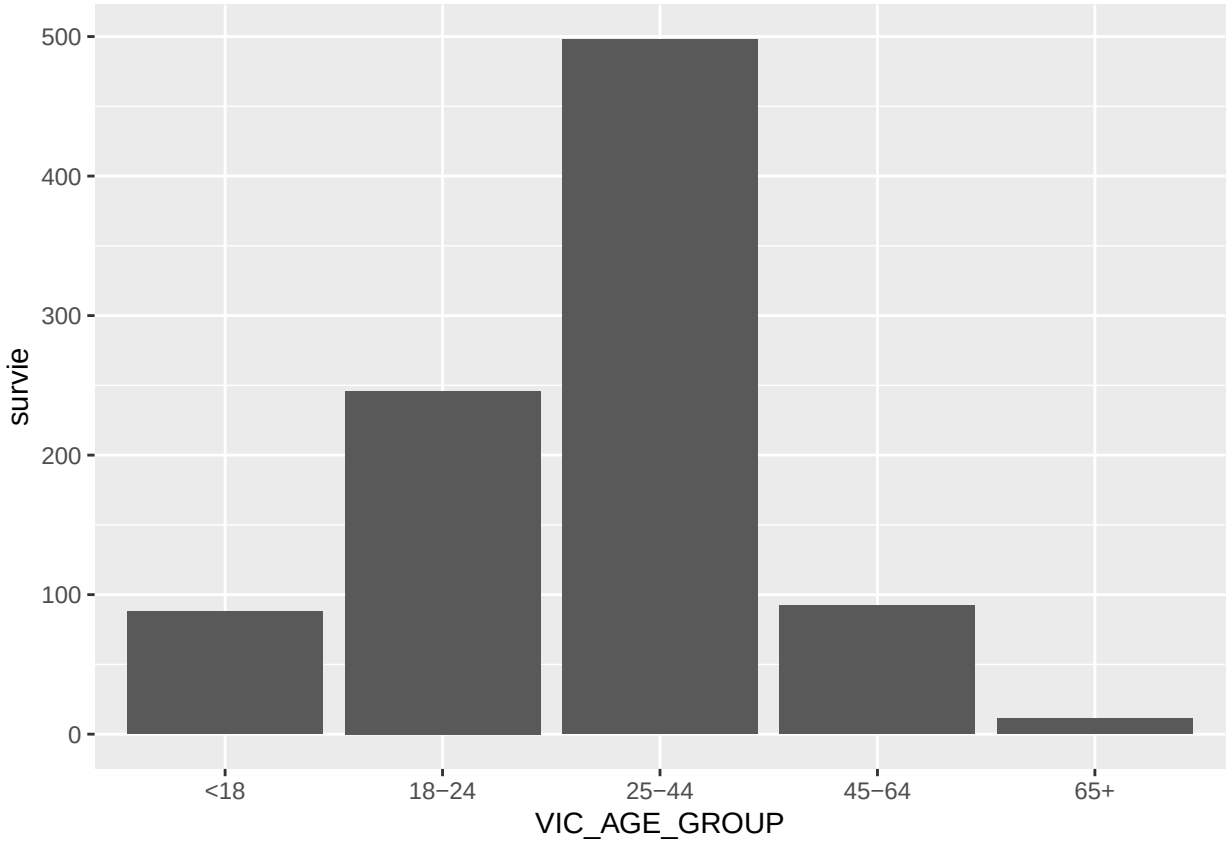
## Visualising Data

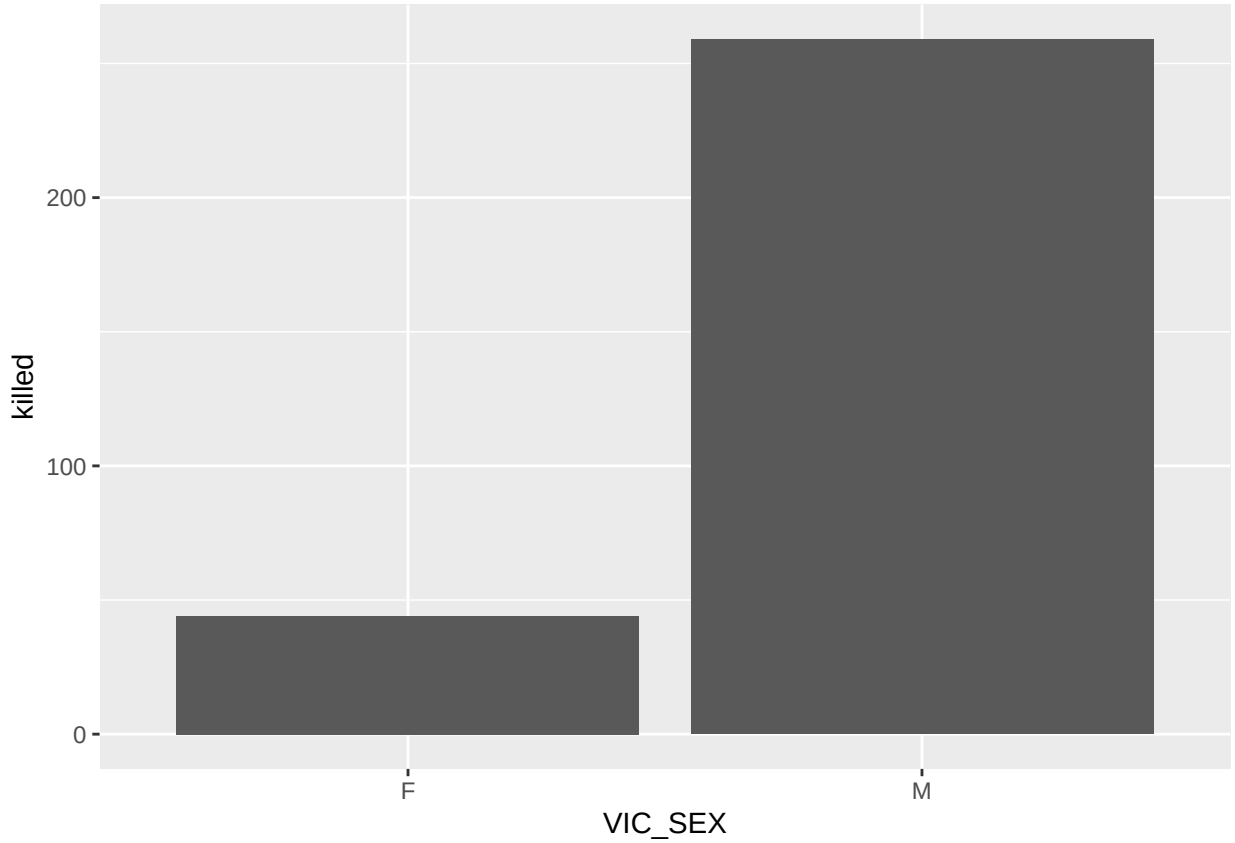
In this project, We would like to focus on the victims demographics and kill rate.

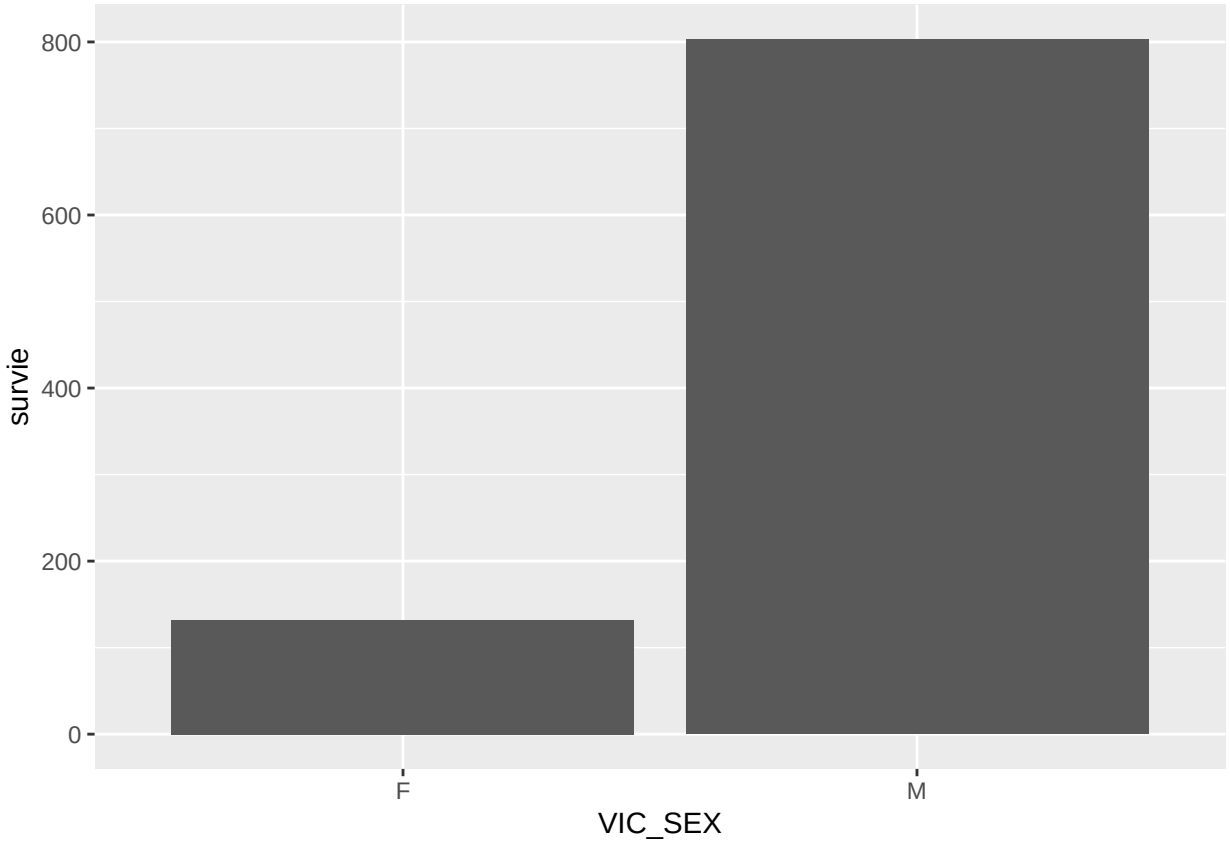
First group by the victims demographics and then plot bar chart for age group, gender and race. We found that most shooting victims were 25-44 adult, Male or Black.

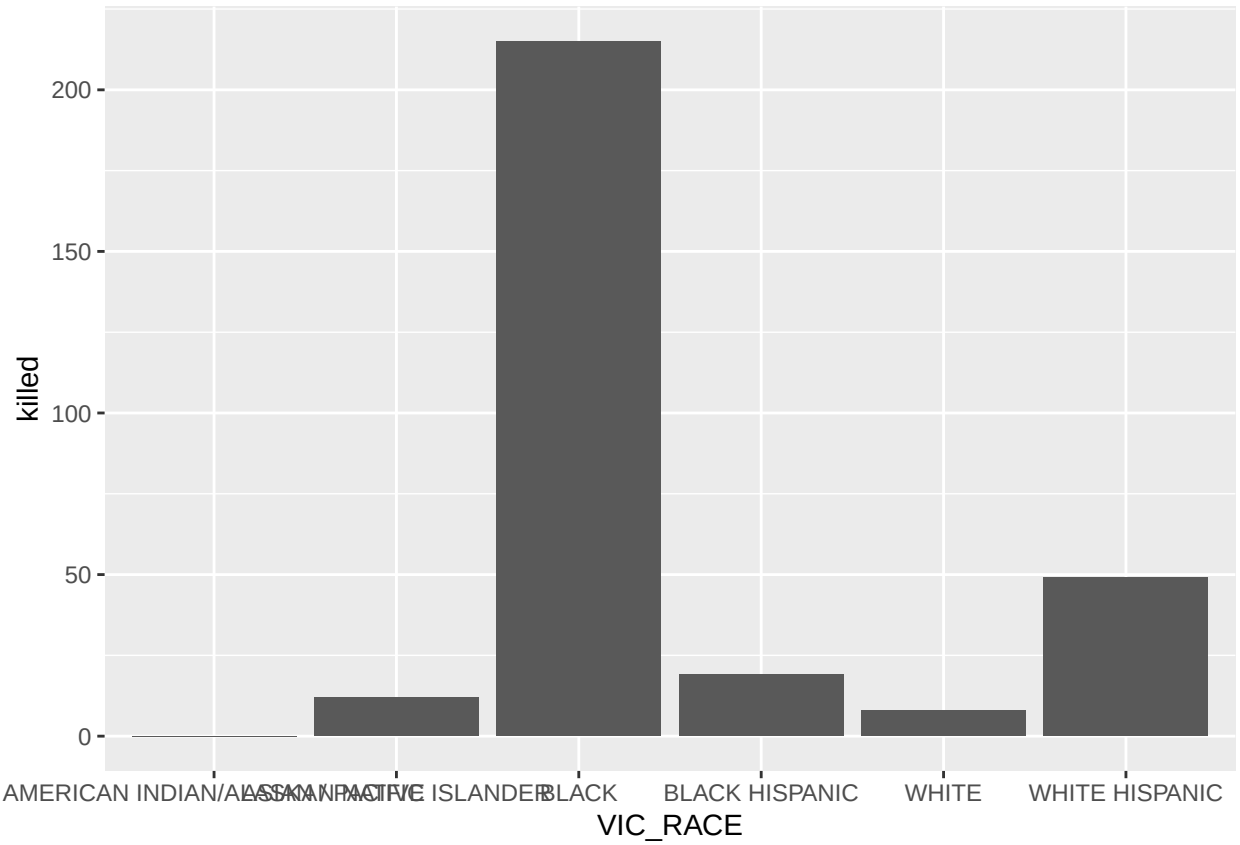
Briefly, we could not see main different of distribution between race, age or gender.

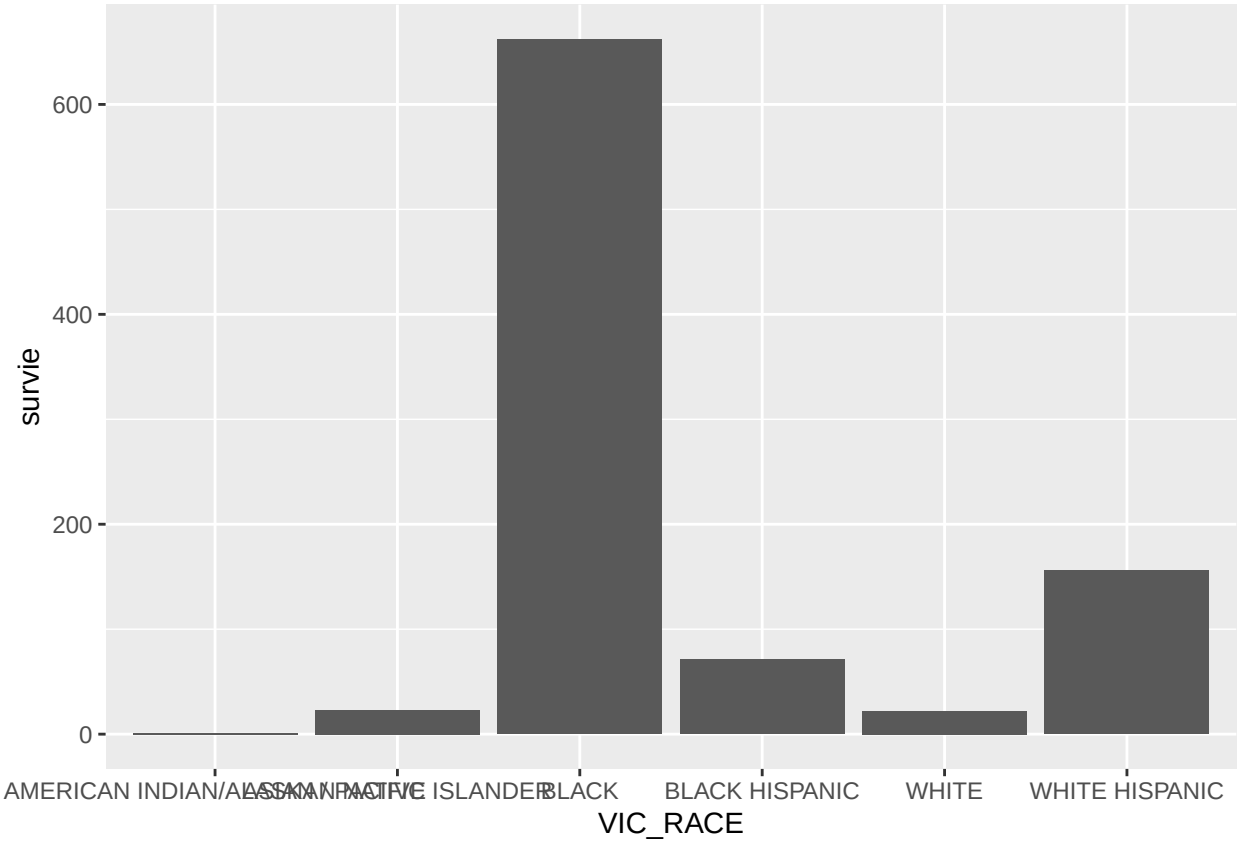

150
100
50
0
killed
<18
18−24
25−44
45−64
65+
VIC_AGE_GROUP


500
400
300
200
100
0
survie
<18
18−24
25−44
45−64
65+
VIC_AGE_GROUP


killed
200
100
0
F
M
VIC_SEX


800
600
400
200
0
survie
F
M
VIC_SEX


200
150
100
50
0
killed
AMERICAN INDIAN/ALASKAN NATIVE
ASIAN/PACIFIC ISLANDER
BLACK
BLACK HISPANIC
WHITE
WHITE HISPANIC
VIC_RACE




## Analysing Data

After constructing a cross table with killed and victims demographic, we test the association with chi-square test. The result show only the age group has weak significant association (p=0.05). the gender and race do not show significant association

```
##
##         AMERICAN INDIAN/ALASKAN NATIVE ASIAN / PACIFIC ISLANDER BLACK
##   FALSE                              1                       23   662
##   TRUE                               0                       12   215
##
##         BLACK HISPANIC WHITE WHITE HISPANIC
##   FALSE             71    22            156
##   TRUE              19     8             49
```

```
## Warning in chisq.test(clean_df$STATISTICAL_MURDER_FLAG, clean_df$VIC_RACE):
## Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  clean_df$STATISTICAL_MURDER_FLAG and clean_df$VIC_RACE
## X-squared = 2.8126, df = 5, p-value = 0.7289
```

```
##
##           F   M
##   FALSE 132 803
##   TRUE   44 259
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  clean_df$STATISTICAL_MURDER_FLAG and clean_df$VIC_SEX
## X-squared = 0.0064439, df = 1, p-value = 0.936
```

```
##
##         <18 18-24 25-44 45-64 65+
##   FALSE  88   246   498    92  11
##   TRUE   22    69   161    43   8
```

```
## Warning in chisq.test(clean_df$STATISTICAL_MURDER_FLAG,
## clean_df$VIC_AGE_GROUP): Chi-squared approximation may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  clean_df$STATISTICAL_MURDER_FLAG and clean_df$VIC_AGE_GROUP
## X-squared = 9.4874, df = 4, p-value = 0.05001
```

## Modelling Data

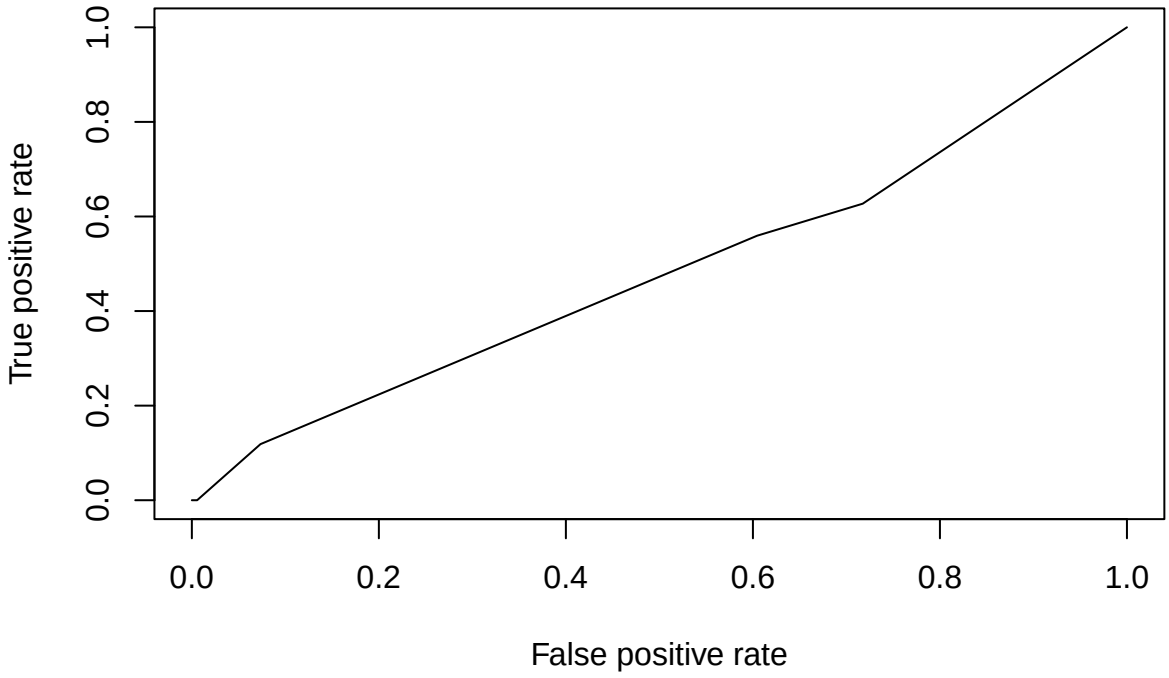
Even though the victim characteristics does not significantly affect the survival rate, we tried to model to fit the data. Using 80% data for training, and rest for testing.

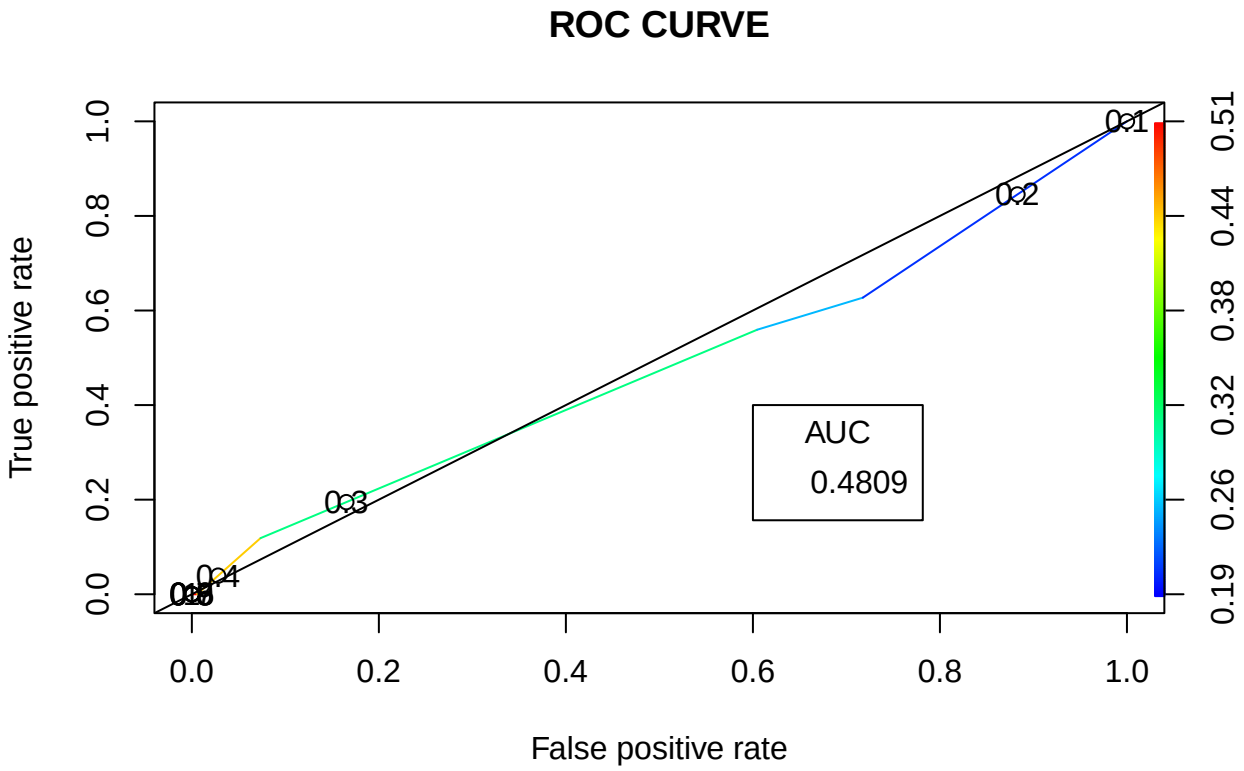
The model show that higher age significantly associate to the survival rate. AIC is high and the AUC is close to 0.5. That indicate that the model could not predict the survival rate. Its prediction is similar to randomly guess.

```
##
## Call:
## glm(formula = STATISTICAL_MURDER_FLAG ~ VIC_AGE_GROUP, family = "binomial",
##     data = train_reg)
##
## Coefficients:
##                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -1.32914    0.26507  -5.014 5.32e-07 ***
## VIC_AGE_GROUP18-24 -0.09883    0.31087  -0.318   0.7505
## VIC_AGE_GROUP25-44  0.23300    0.28310   0.823   0.4105
## VIC_AGE_GROUP45-64  0.53063    0.33247   1.596   0.1105
## VIC_AGE_GROUP65+    1.10599    0.54338   2.035   0.0418 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1112.4  on 1001  degrees of freedom
```

```
## Residual deviance: 1102.1  on  997  degrees of freedom
## AIC: 1112.1
##
## Number of Fisher Scoring iterations: 4

## [1] 0.4808963
```






### Bias Identification

To conclude, the victim demographic may not be associate to the survival rate in shooting incident. However, it might because of bias. First the data contain too many missing data. After cleaning data, there are not much valid data for analysis. Second, the data heavily contain specific group of victims. It also affect the power of the model and analysis. The serious bias is the race of perpetrator contain plenty of unknown. It might be because of racism, when the official would like to hide significant data of the perpetrator.