Week 3 Final Assignment

Oanh

2023-08-21

NYPD Shooting Incident Data Report

Introduction:

To begin, we need to install these necessary packages:(tidyverse),(caret), (ggplot2), (knitr), (dplyr)

```
library(tidyverse)
library(caret)
library(ggplot2)
library(knitr)
library(dplyr)
```

Read the data from the link.

```
#Read the CSV file from URL
nypd_data<-read.csv("https://data.cityofnewyork.us/api/views/833y-fsy8/rows.csv?accessType=DOWNLOAD" )
str(nypd_data)
```

```
## 'data.frame':    27312 obs. of  21 variables:
##  $ INCIDENT_KEY           : int  228798151 137471050 147998800 146837977 58921844 219559682 85295722
##  $ OCCUR_DATE             : chr  "05/27/2021" "06/27/2014" "11/21/2015" "10/09/2015" ...
##  $ OCCUR_TIME             : chr  "21:30:00" "17:40:00" "03:56:00" "18:30:00" ...
##  $ BORO                   : chr  "QUEENS" "BRONX" "QUEENS" "BRONX" ...
##  $ LOC_OF_OCCUR_DESC      : chr  "" "" "" "" ...
##  $ PRECINCT               : int  105 40 108 44 47 81 114 81 105 101 ...
##  $ JURISDICTION_CODE      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ LOC_CLASSFCTN_DESC     : chr  "" "" "" "" ...
##  $ LOCATION_DESC          : chr  "" "" "" "" ...
##  $ STATISTICAL_MURDER_FLAG: chr  "false" "false" "true" "false" ...
##  $ PERP_AGE_GROUP         : chr  "" "" "" "" ...
##  $ PERP_SEX               : chr  "" "" "" "" ...
##  $ PERP_RACE              : chr  "" "" "" "" ...
##  $ VIC_AGE_GROUP          : chr  "18-24" "18-24" "25-44" "<18" ...
##  $ VIC_SEX                : chr  "M" "M" "M" "M" ...
##  $ VIC_RACE               : chr  "BLACK" "BLACK" "WHITE" "WHITE HISPANIC" ...
##  $ X_COORD_CD             : num  1058925 1005028 1007668 1006537 1024922 ...
##  $ Y_COORD_CD             : num  180924 234516 209837 244511 262189 ...
##  $ Latitude               : num  40.7 40.8 40.7 40.8 40.9 ...
##  $ Longitude              : num  -73.7 -73.9 -73.9 -73.9 -73.9 ...
##  $ Lon_Lat                : chr  "POINT (-73.73083868899994 40.662964620000025)" "POINT (-73.9249423
```

Display the first 10 rows of the dataset

```
head(nypd_data,5)
```

```
##   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
##   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
##   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
##   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
##                                      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)
```

Data Preparation and Cleaning

Missing Values

```
#Replace missing value with "N/A"
nypd_data<-nypd_data %>% mutate(across(everything(),~ifelse(is.na(.), "N/A", .)))
```

Making sure there is no missing values.

```
sum(is.na(nypd_data))
```

```
## [1] 0
```

Show the first 10 rows

```
head(nypd_data,10)
```

```
##    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
## 7      85295722 06/17/2012   22:47:00   QUEENS                       114
## 8      71662474 03/08/2010   19:41:00 BROOKLYN                        81
## 9      83002139 02/05/2012   05:45:00   QUEENS                       105
## 10     86437261 08/26/2012   01:10:00   QUEENS                       101
##    JURISDICTION_CODE LOC_CLASSFCTN_DESC           LOCATION_DESC
## 1                  0
## 2                  0
## 3                  0
## 4                  0
## 5                  0
## 6                  0
## 7                  0
## 8                  0
## 9                  0
## 10                 0                     MULTI DWELL - APT BUILD
##    STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
## 1                    false                                          18-24
## 2                    false                                          18-24
## 3                     true                                          25-44
## 4                    false                                            <18
## 5                     true          25-44        M     BLACK         45-64
## 6                     true                                          25-44
## 7                    false                                          25-44
## 8                     true                                          18-24
## 9                    false                                          25-44
## 10                   false          25-44        M     BLACK         25-44
##    VIC_SEX       VIC_RACE X_COORD_CD Y_COORD_CD         Latitude
## 1        M          BLACK    1058925   180924.0      40.66296462
## 2        M          BLACK    1005028   234516.0     40.810351863
## 3        M          WHITE    1007668   209836.5     40.742606633
## 4        M WHITE HISPANIC    1006537   244511.1     40.837782003
## 5        M          BLACK    1024922   262189.4 40.8862379180001
## 6        M          BLACK    1004234   186461.7 40.6784567180001
## 7        M          BLACK     998860   214885.0 40.7564823430001
## 8        M          BLACK    1002883   192219.7 40.6942640560001
## 9        M          BLACK    1054366   196628.4     40.706106731
## 10       M          BLACK    1053937   157130.4     40.597697198
##             Longitude                                       Lon_Lat
## 1  -73.7308386889999 POINT (-73.73083868899994 40.662964620000025)
## 2      -73.924942326  POINT (-73.92494232599995 40.81035186300006)
## 3      -73.915491742  POINT (-73.91549174199997 40.74260663300004)
## 4  -73.9194566149999  POINT (-73.91945661499994 40.83778200300003)
## 5      -73.852909509  POINT (-73.85290950899997 40.88623791800006)
## 6      -73.927952241 POINT (-73.92795224099996 40.678456718000064)
```

```
## 7      -73.947266494  POINT (-73.94726649399996 40.75648234300007)
## 8   -73.9328086369999 POINT (-73.93280863699994 40.694264056000065)
## 9      -73.747106539 POINT (-73.74710653899996 40.706106731000034)
## 10     -73.749064642  POINT (-73.74906464199995 40.59769719800005)
```

Remove any duplicates

```
nypd_data<-distinct(nypd_data)
nrow(nypd_data)
```

```
## [1] 27312
```

There are no duplicates

Incident vs Race analysis

Let's check unique values in VIC_RACE

```
unique(nypd_data$VIC_RACE)
```

```
## [1] "BLACK"                          "WHITE"
## [3] "WHITE HISPANIC"                 "BLACK HISPANIC"
## [5] "ASIAN / PACIFIC ISLANDER"       "UNKNOWN"
## [7] "AMERICAN INDIAN/ALASKAN NATIVE"
```

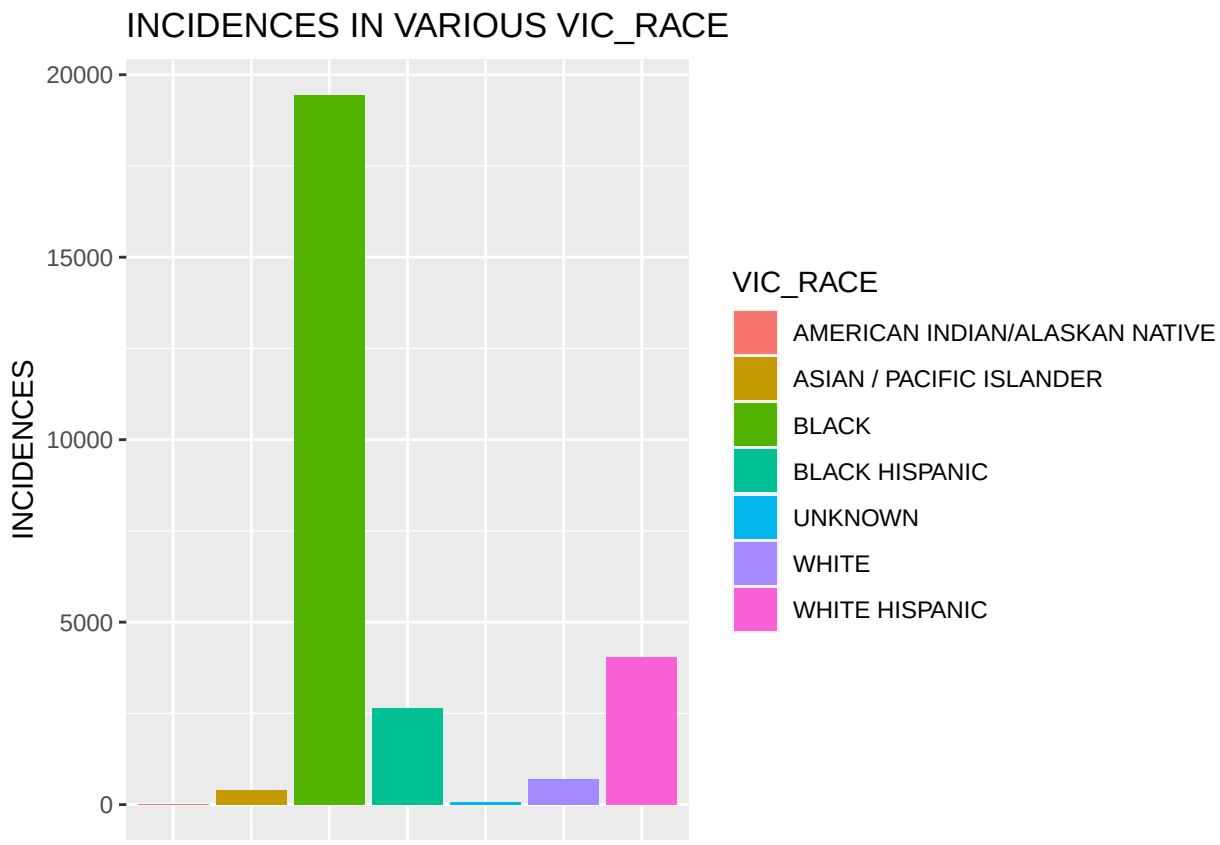
Sort the VIC_RACE in descending order to see which one has the most shootings.

```
nypd_data %>% group_by(VIC_RACE) %>% summarise(Total =n()) %>% arrange(desc(Total))
```

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

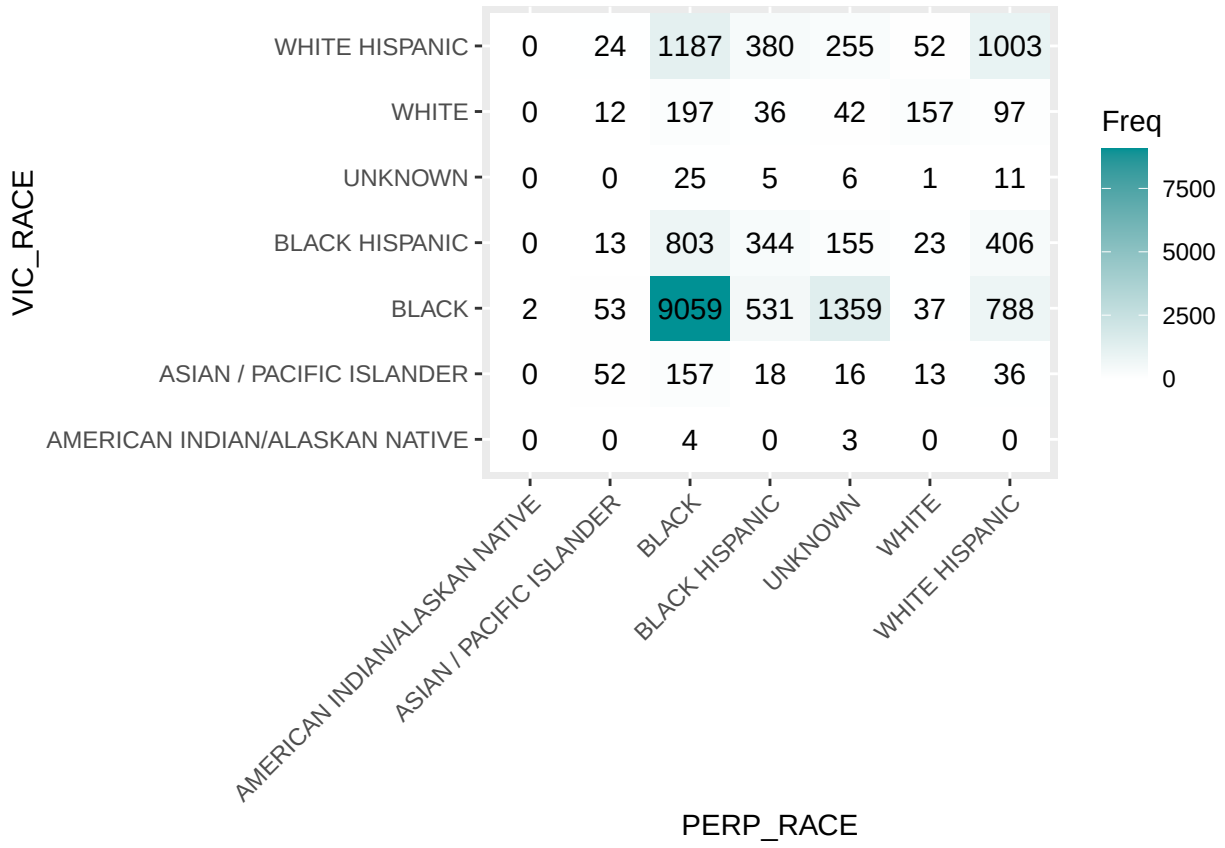
Make chart to see incidences.

```
#Group data by VIC_RACE and calculate the total number of incidents
VIC_RACE_shooting <- nypd_data %>% group_by(VIC_RACE) %>% summarise(incidents=n())
#Create a bar chart
ggplot(VIC_RACE_shooting, aes(x=VIC_RACE, y=incidents, fill=VIC_RACE)) + geom_bar(stat="identity") + xla
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank())
```



Make a confusion matrix between PERP_RACE and VIC_RACE.

```
filter_nypd = nypd_data[nypd_data$PERP_RACE != "" & nypd_data$PERP_RACE != "(null)", ]
cm <- confusionMatrix(factor(filter_nypd$PERP_RACE), factor(filter_nypd$VIC_RACE))

plt <- as.data.frame(cm$table)

ggplot(plt, aes(Prediction,Reference, fill= Freq)) +
        geom_tile() + geom_text(aes(label=Freq)) +
        scale_fill_gradient(low="white", high="#009194") +
        labs(x = "PERP_RACE",y = "VIC_RACE") +
  theme(axis.text.x = element_text(angle = 45, hjust=1))
```



Modeling

Here I made a new column for the population for each Race. I am combining `BLACK_HISPANIC` and `WHITE_HISPANIC` into a single value `HISPANIC` since the NYC demographic data only has population of `HISPANIC` in general.

```
#Filter and remove Unknown Race
nypd_data<-nypd_data[nypd_data$VIC_RACE!= "UNKNOWN", ]
#Make new column
nypd_data[nypd_data$VIC_RACE == "BLACK HISPANIC" | nypd_data$VIC_RACE == "WHITE HISPANIC", c("VIC_RACE"]
nypd_data <- nypd_data %>% mutate(Population = case_when(VIC_RACE =="BLACK" ~ 1947328, VIC_RACE=="WHITE
head(nypd_data,10)
```

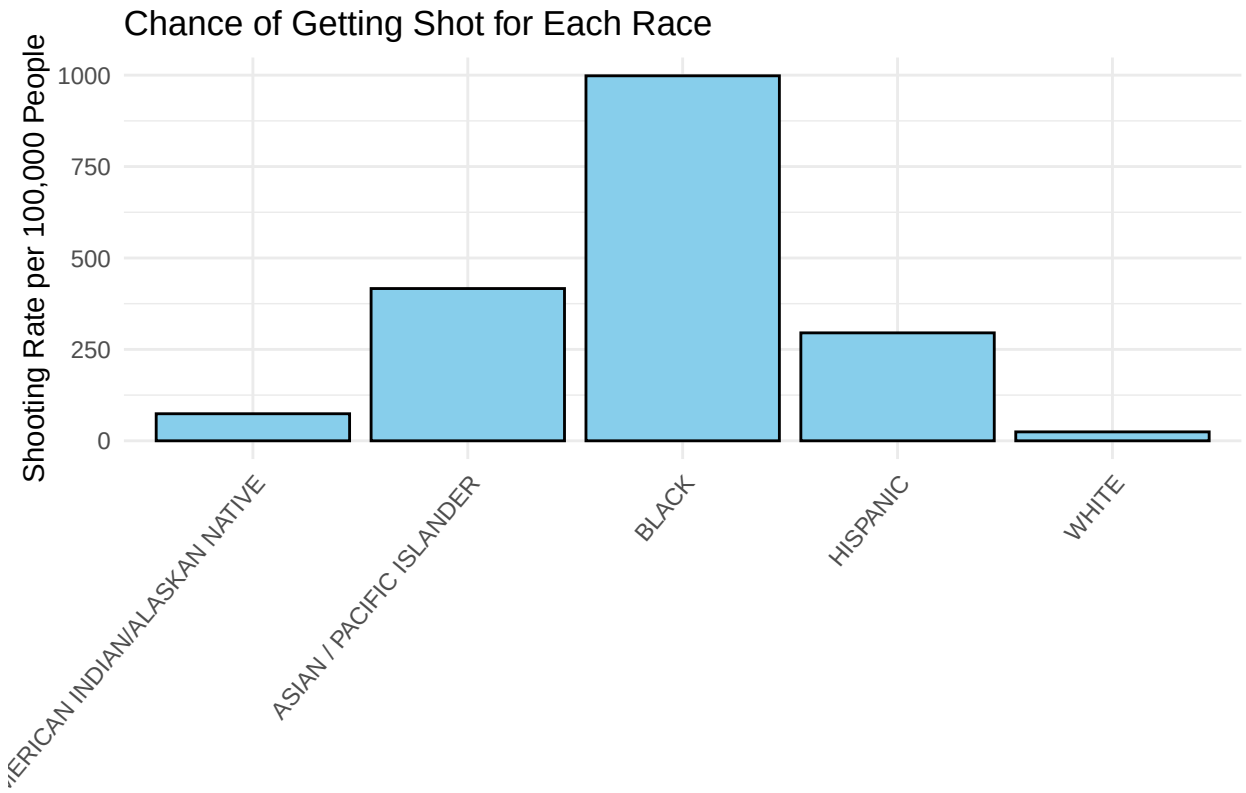
```
##    INCIDENT_KEY OCCUR_DATE OCCUR_TIME     BORO LOC_OF_OCCUR_DESC PRECINCT
## 1     228798151 05/27/2021   21:30:00   QUEENS                      105
## 2     137471050 06/27/2014   17:40:00    BRONX                       40
## 3     147998800 11/21/2015   03:56:00   QUEENS                      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
## 7      85295722 06/17/2012   22:47:00   QUEENS                      114
## 8      71662474 03/08/2010   19:41:00 BROOKLYN                       81
## 9      83002139 02/05/2012   05:45:00   QUEENS                      105
## 10     86437261 08/26/2012   01:10:00   QUEENS                      101
##    JURISDICTION_CODE LOC_CLASSFCTN_DESC          LOCATION_DESC
```

```
## 1                  0
## 2                  0
## 3                  0
## 4                  0
## 5                  0
## 6                  0
## 7                  0
## 8                  0
## 9                  0
## 10                 0                     MULTI DWELL - APT BUILD
##    STATISTICAL_MURDER_FLAG PERP_AGE_GROUP PERP_SEX PERP_RACE VIC_AGE_GROUP
## 1                    false                                          18-24
## 2                    false                                          18-24
## 3                     true                                          25-44
## 4                    false                                            <18
## 5                     true          25-44        M     BLACK         45-64
## 6                     true                                          25-44
## 7                    false                                          25-44
## 8                     true                                          18-24
## 9                    false                                          25-44
## 10                   false          25-44        M     BLACK         25-44
##    VIC_SEX VIC_RACE X_COORD_CD Y_COORD_CD         Latitude         Longitude
## 1        M    BLACK    1058925   180924.0      40.66296462 -73.7308386889999
## 2        M    BLACK    1005028   234516.0     40.810351863     -73.924942326
## 3        M    WHITE    1007668   209836.5     40.742606633     -73.915491742
## 4        M HISPANIC    1006537   244511.1     40.837782003 -73.9194566149999
## 5        M    BLACK    1024922   262189.4 40.8862379180001     -73.852909509
## 6        M    BLACK    1004234   186461.7 40.6784567180001     -73.927952241
## 7        M    BLACK     998860   214885.0 40.7564823430001     -73.947266494
## 8        M    BLACK    1002883   192219.7 40.6942640560001 -73.9328086369999
## 9        M    BLACK    1054366   196628.4     40.706106731     -73.747106539
## 10       M    BLACK    1053937   157130.4     40.597697198     -73.749064642
##                                       Lon_Lat Population
## 1  POINT (-73.73083868899994 40.662964620000025)    1947328
## 2   POINT (-73.92494232599995 40.81035186300006)    1947328
## 3   POINT (-73.91549174199997 40.74260663300004)    2854519
## 4   POINT (-73.91945661499994 40.83778200300003)    2267827
## 5   POINT (-73.85290950899997 40.88623791800006)    1947328
## 6  POINT (-73.92795224099996 40.678456718000064)    1947328
## 7   POINT (-73.94726649399996 40.75648234300007)    1947328
## 8  POINT (-73.93280863699994 40.694264056000065)    1947328
## 9  POINT (-73.74710653899996 40.706106731000034)    1947328
## 10  POINT (-73.74906464199995 40.59769719800005)    1947328
```

Calculate the shooting rate per 100,000 people and Plot the shooting rate for each RACE

```
nypd_shooting_rate<-nypd_data %>% group_by(VIC_RACE) %>% summarise(total_shooting=n(), population=unique
shooting_rate=total_shooting/(population/100000)) %>% arrange(desc(shooting_rate))
ggplot(nypd_shooting_rate, aes(x = VIC_RACE, y = shooting_rate)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  ggtitle("Chance of Getting Shot for Each Race") +
  xlab("VIC_RACE") +
```

```
  ylab("Shooting Rate per 100,000 People") +
  theme_minimal()+theme(axis.text.x = element_text(angle = 50, hjust=1))
```




```
nypd_shooting_rate %>%
  as_tibble() %>%
  select(VIC_RACE, shooting_rate) %>%
  mutate(shooting_rate = sprintf("%.2f", shooting_rate))
```

```
## # A tibble: 5 x 2
##   VIC_RACE                       shooting_rate
##   <chr>                          <chr>
## 1 BLACK                          998.24
## 2 ASIAN / PACIFIC ISLANDER       416.40
## 3 HISPANIC                       295.22
## 4 AMERICAN INDIAN/ALASKAN NATIVE 74.04
## 5 WHITE                          24.45
```

```
nypd_shooting_rate_per_person <- nypd_shooting_rate %>%
  mutate(shooting_rate_per_person = total_shooting / population) %>%
  select(VIC_RACE, shooting_rate_per_person) %>%
  mutate(shooting_rate_per_person = sprintf("%.6f", shooting_rate_per_person * 100)) %>%
  rename(`Borough` = VIC_RACE, `Shooting Rate per Person` = shooting_rate_per_person) %>%
  mutate(`Shooting Rate per Person` = paste0(`Shooting Rate per Person`, "%"))
print(nypd_shooting_rate_per_person)
```

```
## # A tibble: 5 x 2
##   Borough                       `Shooting Rate per Person`
##   <chr>                         <chr>
## 1 BLACK                         0.998240%
## 2 ASIAN / PACIFIC ISLANDER      0.416400%
## 3 HISPANIC                      0.295217%
## 4 AMERICAN INDIAN/ALASKAN NATIVE 0.074041%
## 5 WHITE                         0.024452%
```

Create the linear regression model and Print the summary of the model

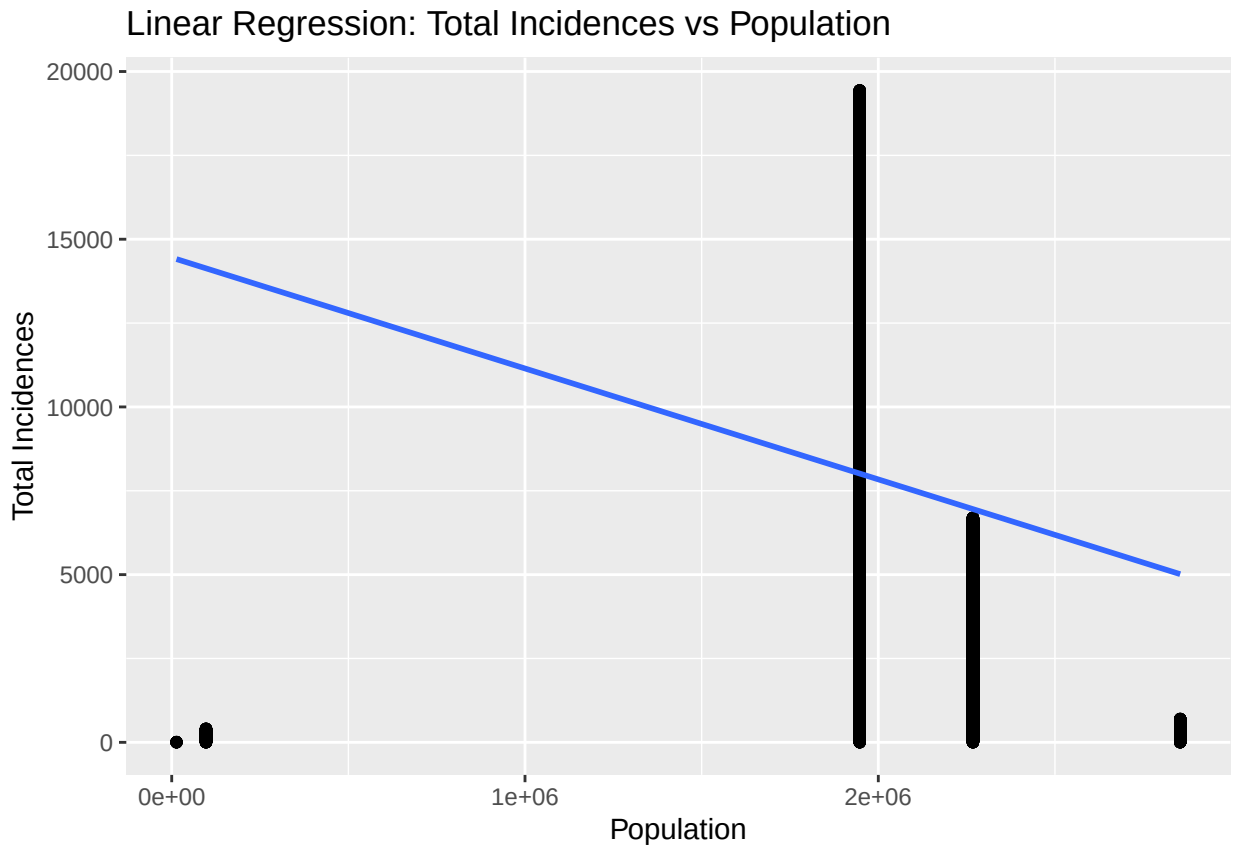
```
nypd_data <- nypd_data %>%
  mutate(Total = ifelse(!is.na(VIC_RACE), 1, 0)) %>%
  group_by(VIC_RACE) %>%
  mutate(Total = cumsum(Total))

lm_model <- lm(Total ~ Population, data = nypd_data)

summary(lm_model)
```

```
##
## Call:
## lm(formula = Total ~ Population, data = nypd_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14407  -4528  -1228   4614  11425
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.445e+04  2.303e+02   62.75   <2e-16 ***
## Population  -3.307e-03  1.127e-04  -29.35   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5672 on 27244 degrees of freedom
## Multiple R-squared:  0.03064,    Adjusted R-squared:  0.03061
## F-statistic: 861.2 on 1 and 27244 DF,  p-value: < 2.2e-16
```

Create a scatter plot with the regression line

```
ggplot(nypd_data, aes(x = Population, y = Total)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  xlab("Population") +
  ylab("Total Incidences") +
  ggtitle("Linear Regression: Total Incidences vs Population")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Conclusion

Bias

If we only see the INCIDENCES IN VARIOUS VIC_RACE chart, we can conclude that the Black people have the highest chances of getting shot compared to the others in New York City. However, if we also see the confusion matrix between PERP_RACE and VIC_RACE, we will know that the original cause is because back people shooting black people has the highest number cases. We can understand that because people in same race can involve in similar of activities and tends to live in the same neighborhood. Looking at only VIC_RACE charts will introduce us some bias in shooting rates of different races. However, we can mitigate that by exploring more fine-grained visualization such as confusion matrix between VIC_RACE vs PERP_RACE as I have done here.

Model prediction

Based on the linear regression results, we can conclude that there is a negative relationship between the number of shooting incidents and the population size in each Race. In other words, as the population size increases, the number of shooting incidents tends to decrease . The R-squared value of 0.03064 indicates that the model explains approximately 3% of the variability in the number of shooting incidents. However, it's important to note that correlation does not imply causation, and there may be other factors that contribute to the number of shooting incidents beyond just population size.