Final project

Liza Lunardi Lemos

December 15, 2017

Dataset

This work is the final project of Literate Programming and Statistics. The dataset the is going to be analyzed is US Homicides, which it has homicides reports from 1980 to 2014. This dataset includes the age, race, sex, ethnicity of victims and perpetrators, in addition to the relationship between the victim and perpetrator and weapon used.

Download the data

The dataset is available in: https://www.kaggle.com/jyzaguirre/us-homicide-reports/downloads/database.csv

The dataset must be in the same directory as .Rmd file.

```r
library(readr)
#URL <- "https://www.kaggle.com/jyzaguirre/us-homicide-reports/downloads/database.csv"
df <- read_delim("database.csv", delim=",")
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   Year = col_integer(),
##   Incident = col_integer(),
##   `Victim Age` = col_integer(),
##   `Perpetrator Age` = col_integer(),
##   `Victim Count` = col_integer(),
##   `Perpetrator Count` = col_integer()
## )

## See spec(...) for full column specifications.

## Warning in rbind(names(probs), probs_f): number of columns of result is not
## a multiple of vector length (arg 2)

## Warning: 1 parsing failure.
## row # A tibble: 1 x 5 col      row             col   expected actual           file expected    <int
```

```r
df
```

```
## # A tibble: 638,454 x 24
##    `Record ID` `Agency Code` `Agency Name`    `Agency Type`      City
##          <chr>         <chr>         <chr>            <chr>     <chr>
##  1      000001       AK00101     Anchorage Municipal Police Anchorage
##  2      000002       AK00101     Anchorage Municipal Police Anchorage
##  3      000003       AK00101     Anchorage Municipal Police Anchorage
##  4      000004       AK00101     Anchorage Municipal Police Anchorage
##  5      000005       AK00101     Anchorage Municipal Police Anchorage
##  6      000006       AK00101     Anchorage Municipal Police Anchorage
##  7      000007       AK00101     Anchorage Municipal Police Anchorage
##  8      000008       AK00101     Anchorage Municipal Police Anchorage
##  9      000009       AK00101     Anchorage Municipal Police Anchorage
## 10      000010       AK00101     Anchorage Municipal Police Anchorage
## # ... with 638,444 more rows, and 19 more variables: State <chr>,
## #   Year <int>, Month <chr>, Incident <int>, `Crime Type` <chr>, `Crime
## #   Solved` <chr>, `Victim Sex` <chr>, `Victim Age` <int>, `Victim
## #   Race` <chr>, `Victim Ethnicity` <chr>, `Perpetrator Sex` <chr>,
## #   `Perpetrator Age` <int>, `Perpetrator Race` <chr>, `Perpetrator
## #   Ethnicity` <chr>, Relationship <chr>, Weapon <chr>, `Victim
## #   Count` <int>, `Perpetrator Count` <int>, `Record Source` <chr>
```

Load the necessary packages:

```
library(dplyr);
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr);
library(ggplot2);
library(gridExtra);
```

```
##
## Attaching package: 'gridExtra'

## The following object is masked from 'package:dplyr':
##
##     combine
```

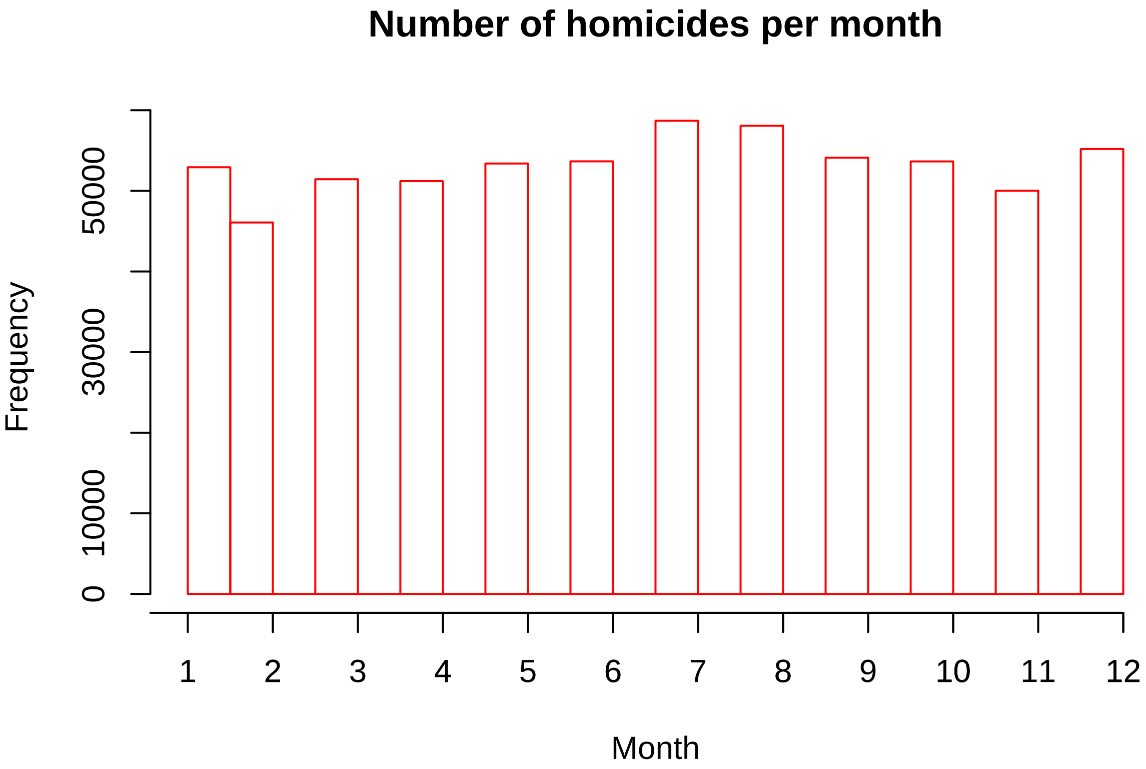
The possible questions are:

1) If the month influence the homicides
2) What is the sex and race of the victim that is most frequent? or What is the sex and race of the perpetrator that is most frequent?

1) If the month influence the homicides

First, we have to change the column 'Month' that is a string to integer. Thus, it is possible to plot a histogram. From the histogram, we can notice that there is no relation between the month and the quantity of homicides.

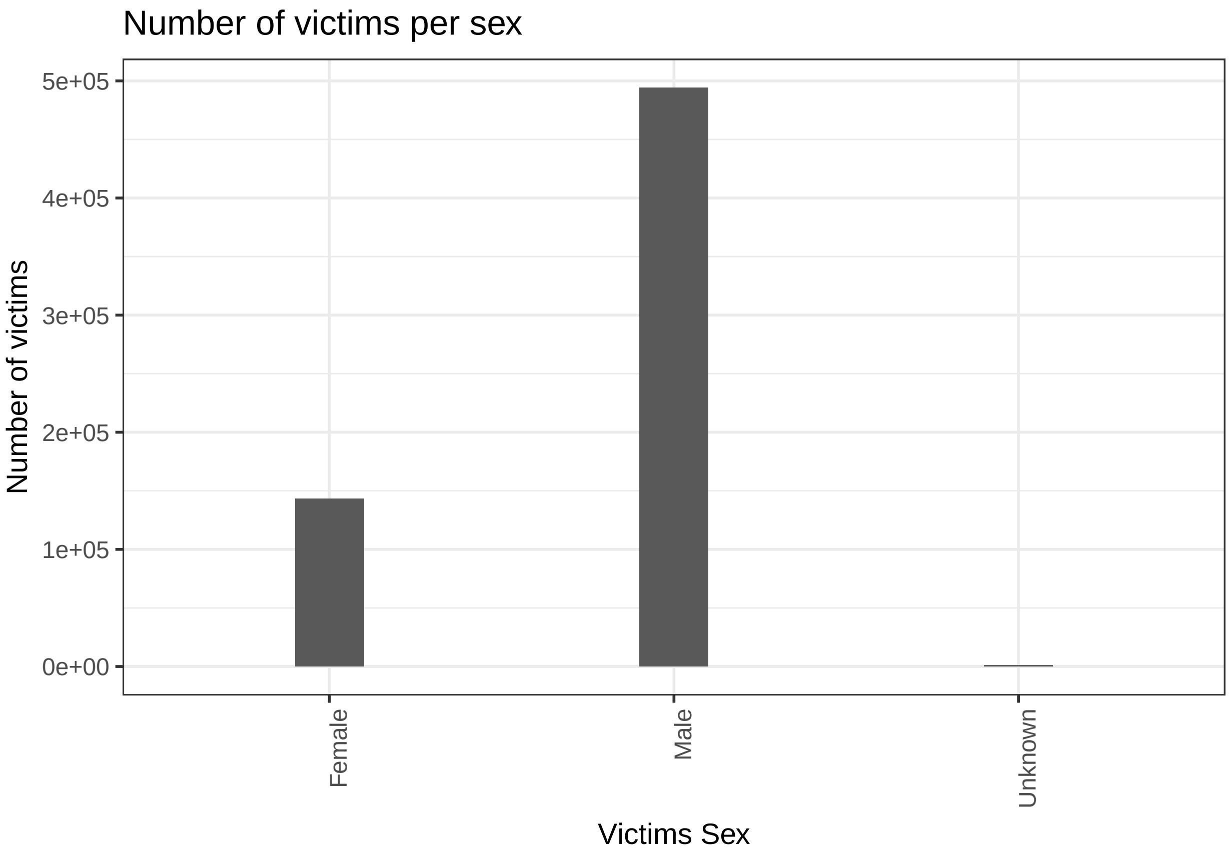
```
a = df$Month
new_col = df %>%
 mutate(nr_month = if_else(a == 'January', 1,
               if_else(a == 'February', 2,
               if_else(a == 'March', 3,
               if_else(a == 'April', 4,
               if_else(a == 'May', 5,
               if_else(a == 'June', 6,
               if_else(a == 'July', 7,
               if_else(a == 'August', 8,
               if_else(a == 'September', 9,
               if_else(a == 'October', 10,
               if_else(a == 'November', 11,
               if_else(a == 'December', 12,
                       NA_real_)))))))))))))

# change the column nr_month to integer
df = transform(new_col, nr_month = as.integer(new_col$nr_month))


hist(df$nr_month, xlab = 'Month', main = 'Number of homicides per month',  border="red")
axis(side=1, at=seq(0,12, 1))
```



2.1) What is the sex and race of the victim that is most frequent?

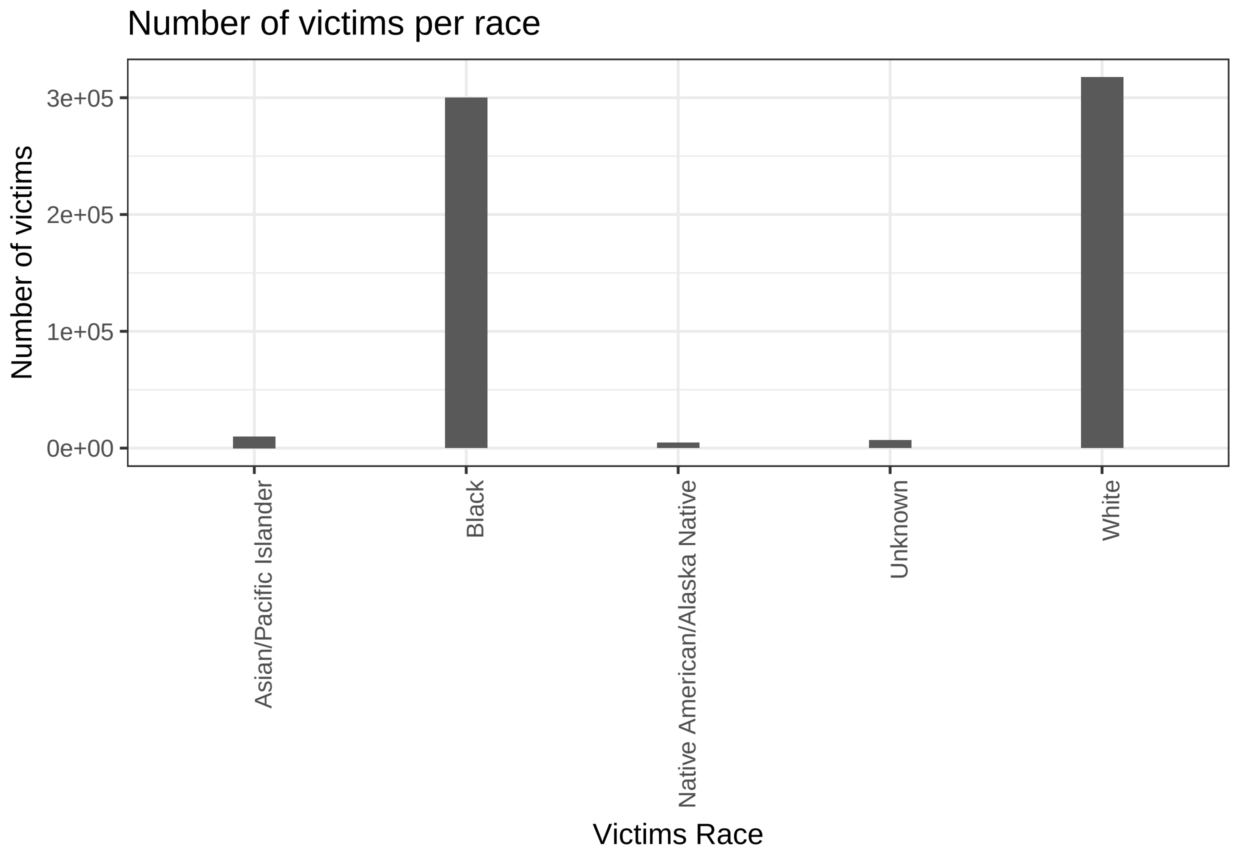
First, we plot the number of victims per sex. We can see that most of victims are man.

```
df %>% group_by(Victim.Sex) %>% summarise(number_victins_per_sex = n()) %>% ggplot(aes(x=Victim.Sex, y=
```



Second, we plot the number of victims per race. We can notice that the races ‘white’ and ‘black’ suffer more homicides.

```
df %>% group_by(Victim.Race) %>% summarise(number_victins_per_race = n()) %>% ggplot(aes(x=Victim.Race,
```



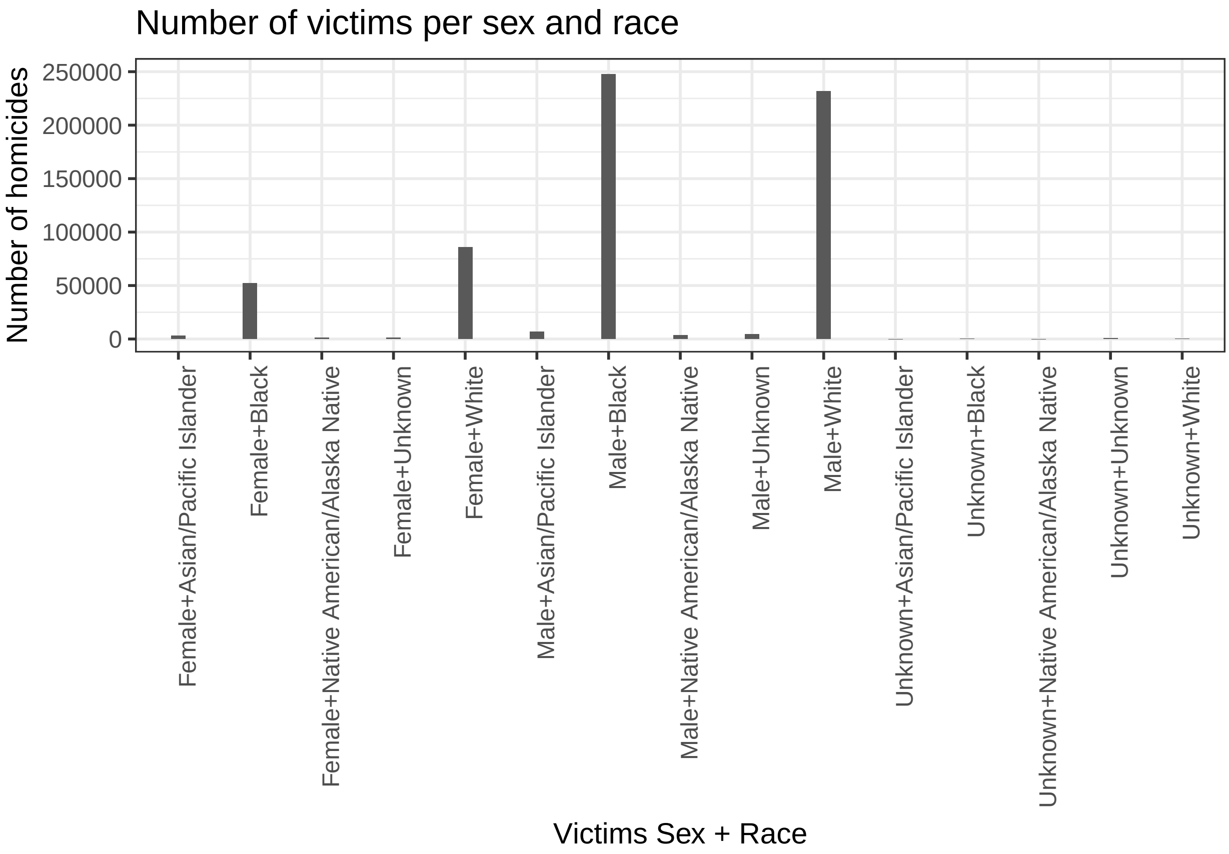
After, we group by the data set by columns ‘Victim.Sex’ and ‘Victim.Race’. Thus, we count how many victims it had for each sex and race.

Since it is difficult to compare too many numbers, we made a bar plot to show which variable has more victims, the variable in question is a group by of sex and race.

```
df %>% group_by(Victim.Sex, Victim.Race) %>% summarise(number_victins_per_sex_race = n())
```

```
## # A tibble: 15 x 3
## # Groups:   Victim.Sex [?]
##    Victim.Sex                   Victim.Race number_victins_per_sex_race
##         <chr>                         <chr>                       <int>
##  1     Female        Asian/Pacific Islander                        2953
##  2     Female                         Black                       52083
##  3     Female Native American/Alaska Native                        1218
##  4     Female                       Unknown                        1352
##  5     Female                         White                       85739
##  6       Male        Asian/Pacific Islander                        6935
##  7       Male                         Black                      247775
##  8       Male Native American/Alaska Native                        3348
##  9       Male                       Unknown                        4439
## 10       Male                         White                      231628
## 11    Unknown        Asian/Pacific Islander                           2
## 12    Unknown                         Black                          41
## 13    Unknown Native American/Alaska Native                           1
## 14    Unknown                       Unknown                         885
## 15    Unknown                         White                          55
```

```
df %>% group_by(Victim.Sex, Victim.Race) %>% summarize(number_victins_per_sex_race = n()) %>% mutate(vi
```

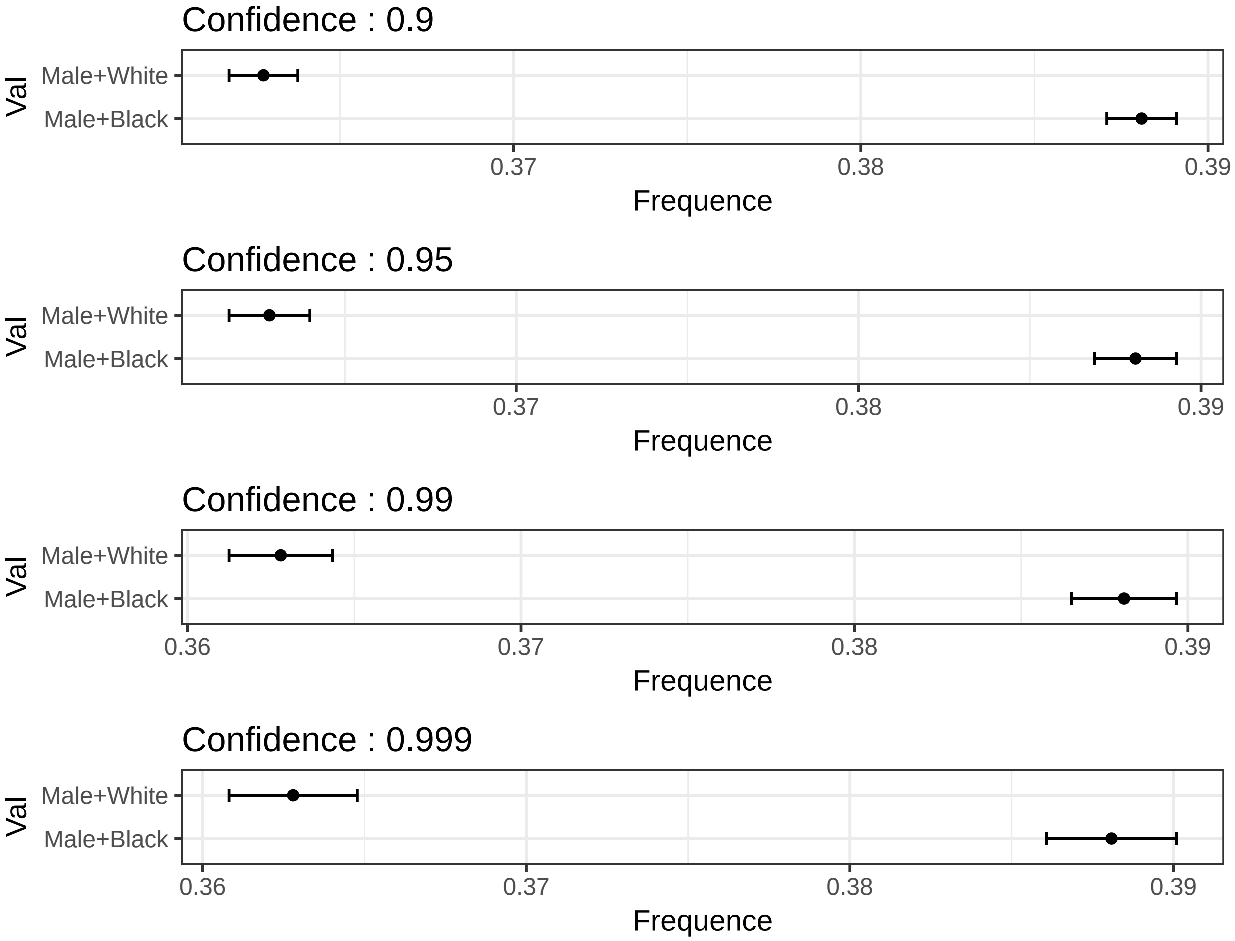


As the 'Male + White' and 'Male + Black' bars are very similar and difficult to identify the difference between them, we do a statistical test with different levels of confidence. Thus, we can identify which is the sex and race of the victim who most suffers homicides.

The conclusion is that even with a high level of confidence 'Male+Black' are the highest victims of homicides according to this dataset.

```
 Calculate_error <-function(Confidence = 0.95 )
 {

  Phi_alpha= qnorm(1-(1-Confidence)/2) ;

  sample_size = nrow(df)


df %>%
    group_by(Victim.Sex, Victim.Race) %>%
    summarize(number_victins_per_sex_race = n()) %>%
    mutate(vic_sex_race = paste(Victim.Sex, Victim.Race, sep = '+')) %>%
    mutate(Freq=number_victins_per_sex_race/sample_size) %>%
    mutate(Estimated_std_deviation=sqrt(Freq*(1-Freq)),
         Erreur=Phi_alpha*Estimated_std_deviation/sqrt(sample_size)) %>%
    filter((Victim.Sex == 'Male')&((Victim.Race == 'White') | Victim.Race == 'Black')) %>%
    ggplot(aes(x=Freq,xmin=Freq-Erreur,xmax=Freq+Erreur,y=(vic_sex_race))) +
    geom_point()+
```

```r
    geom_errorbarh(height=.3)+
    xlab("Frequence")+ylab ("Val") +
    labs(title = paste("Confidence :",Confidence))+
    theme_bw()
 }

list(0.9,0.95,0.99,0.999) %>%
lapply(function(Param_Confidence) {
  Calculate_error(Confidence = Param_Confidence)
}) %>%
grid.arrange(ncol = 1,grobs=.);
```



2.2) What is the sex and race of the perpetrator that is most frequent?

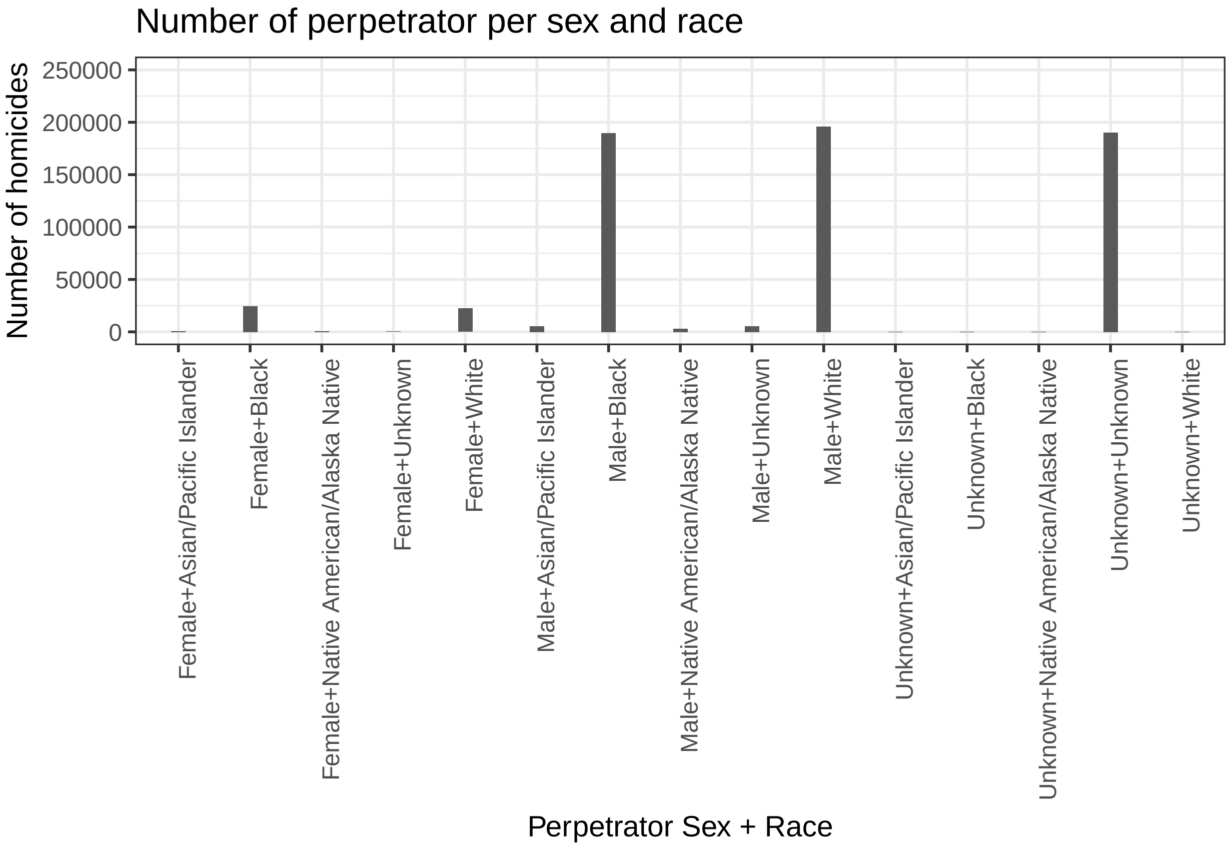
We do the same analysis of victims to perpetrator. From the plot, we can see that most of perpetrator that we have information are male, white or black. However, the bar of perpetrator that are male (white or black) is almost the same of unknown sex and race, what is possible, that we do not know who is the perpetrator.

The conclusion is that even with a high level of confidence 'Male+White' is the highest perpetrator of homicides according to this dataset. An important point here, the value 'Unknown' for sex and race is high, pretty close to 'Male+Black', their error bars approach each other.

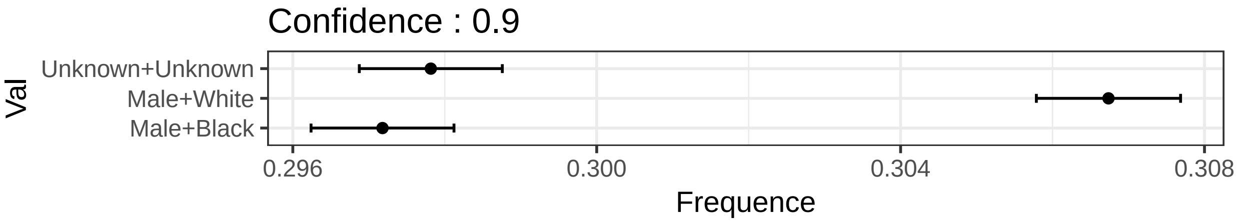
```
df %>% group_by(Perpetrator.Sex, Perpetrator.Race) %>% summarise(number_perp_per_sex_race = n())
```

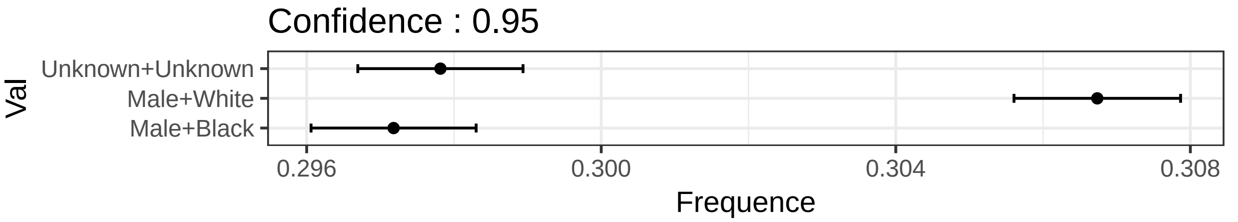
```
## # A tibble: 15 x 3
## # Groups:   Perpetrator.Sex [?]
##    Perpetrator.Sex              Perpetrator.Race number_perp_per_sex_race
##              <chr>                         <chr>                    <int>
##  1          Female        Asian/Pacific Islander                      577
##  2          Female                         Black                    24648
##  3          Female Native American/Alaska Native                      578
##  4          Female                       Unknown                      403
##  5          Female                         White                    22342
##  6            Male        Asian/Pacific Islander                     5449
##  7            Male                         Black                   189736
##  8            Male Native American/Alaska Native                     3017
##  9            Male                       Unknown                     5502
## 10            Male                         White                   195837
## 11         Unknown        Asian/Pacific Islander                       20
## 12         Unknown                         Black                      132
## 13         Unknown Native American/Alaska Native                        7
## 14         Unknown                       Unknown                   190142
## 15         Unknown                         White                       64
```

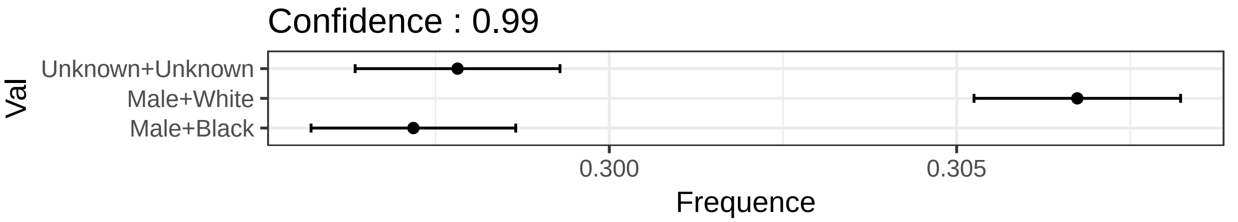
```
df %>% group_by(Perpetrator.Sex, Perpetrator.Race) %>% summarise(number_perp_per_sex_race = n()) %>% mu
```

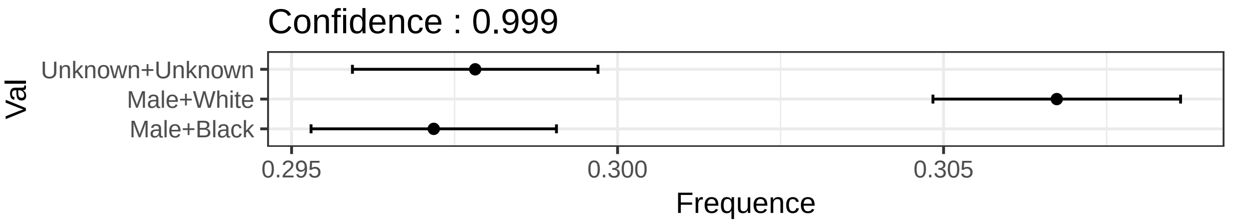


```
Calculate_error <-function(Confidence = 0.95 )
{
  Phi_alpha= qnorm(1-(1-Confidence)/2) ;

  sample_size = nrow(df)


df %>%
    group_by(Perpetrator.Sex, Perpetrator.Race) %>%
    summarize(number_victins_per_sex_race = n()) %>%
    mutate(vic_sex_race = paste(Perpetrator.Sex, Perpetrator.Race, sep = '+')) %>%
    mutate(Freq=number_victins_per_sex_race/sample_size) %>%
    mutate(Estimated_std_deviation=sqrt(Freq*(1-Freq)),
         Erreur=Phi_alpha*Estimated_std_deviation/sqrt(sample_size)) %>%
    filter((Perpetrator.Sex == 'Male')&((Perpetrator.Race == 'White') | Perpetrator.Race == 'Black') |
    ggplot(aes(x=Freq,xmin=Freq-Erreur,xmax=Freq+Erreur,y=(vic_sex_race))) +
    geom_point()+
    geom_errorbarh(height=.3)+
    xlab("Frequence")+ylab ("Val") +
    labs(title = paste("Confidence :",Confidence))+
    theme_bw()
 }

list(0.9,0.95,0.99,0.999) %>%
lapply(function(Param_Confidence) {
  Calculate_error(Confidence = Param_Confidence)
}) %>%
grid.arrange(ncol = 1,grobs=.);
```









Conclusion

The mojority of victims are male and black. On the other hand, the mojoriry of perpetrator are male and white. However, we do not have much information about the perpetrator, because the quantity of homicides that the perpetrator is unknown is high.