Sai_Omkar_PS7

Load packages and data

```
suppressMessages(library(dplyr))
library(ggplot2)
library(moments)
library(haven)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.4      v purrr   0.3.4
## v tidyr   1.1.3      v stringr 1.4.0
## v readr   2.0.1      v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(boot)
```

```
data <- read.csv("ppha312x2021.csv") #working directory is same as data directory
# set seed to ensure same results
set.seed(110821)
```

```
data <- data %>% filter(inctot >= 0)
```

```
summary(data$age)
```

```
##    Length     Class      Mode
##     10140 character character
```

```
#desc(data$age)
```

```
#max(data$age)
```

```
summary(data)
```

```
##       year        statefip           met2013             perwt
##  Min.   :2019   Length:10140      Length:10140       Min.   :   2.0
##  1st Qu.:2019   Class :character   Class :character   1st Qu.:  57.0
##  Median :2019   Mode  :character   Mode  :character   Median :  85.0
```

```
## Mean    :2019                                     Mean    : 115.6
## 3rd Qu.:2019                                       3rd Qu.: 133.0
## Max.    :2019                                      Max.    :1977.0
##     sex               age               race              hispan
## Length:10140      Length:10140      Length:10140      Length:10140
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##     bpl               educd             empstat           uhrswork
## Length:10140      Length:10140      Length:10140      Length:10140
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##     inctot            incwage
## Min.   :      0   Min.   :      0
## 1st Qu.:  11775   1st Qu.:      0
## Median :  35000   Median :  25000
## Mean   :1744935   Mean   :209709
## 3rd Qu.:  94000   3rd Qu.:  88250
## Max.   :9999999   Max.   :999999
```

```r
#New Column : African American
data %>% count(race)
```

```
##                            race    n
## 1 American Indian or Alaska Native    9
## 2      Black/African American/Negro 3899
## 3                           Chinese    9
## 4                          Japanese    3
## 5  Other Asian or Pacific Islander   61
## 6                  Other race, nec   93
## 7          Three or more major races   15
## 8                  Two major races  136
## 9                             White 5915
```

```r
data$isAfricanAmerican <- ifelse(data$race == 'Black/African American/Negro', 1, 0)
data <- data %>% mutate(isAfricanAmerican = as.factor(isAfricanAmerican))
summary(data$isAfricanAmerican)
```

```
##    0    1
## 6241 3899
```

```r
head(data$isAfricanAmerican)
```

```
## [1] 1 0 0 0 1 1
## Levels: 0 1
```

```r
data$isFemale <- ifelse(data$sex == 'Female', 1, 0)
data <- data %>% mutate(isFemale = as.factor(isFemale))
summary(data$isFemale)
```

```
##    0    1
## 4823 5317
```

```r
summary(data)
```

```
##       year         statefip            met2013              perwt
##  Min.   :2019   Length:10140       Length:10140       Min.   :    2.0
##  1st Qu.:2019   Class :character   Class :character   1st Qu.:   57.0
##  Median :2019   Mode  :character   Mode  :character   Median :   85.0
##  Mean   :2019                                         Mean   :  115.6
##  3rd Qu.:2019                                         3rd Qu.:  133.0
##  Max.   :2019                                         Max.   : 1977.0
##      sex                age                race               hispan
##  Length:10140       Length:10140       Length:10140       Length:10140
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      bpl                educd              empstat             uhrswork
##  Length:10140       Length:10140       Length:10140       Length:10140
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      inctot             incwage        isAfricanAmerican isFemale
##  Min.   :      0    Min.   :     0    0:6241            0:4823
##  1st Qu.:  11775    1st Qu.:     0    1:3899            1:5317
##  Median :  35000    Median : 25000
##  Mean   :1744935    Mean   :209709
##  3rd Qu.:  94000    3rd Qu.: 88250
##  Max.   :9999999    Max.   :999999
```

```r
data <- data %>% mutate(isEmployed = ifelse(data$empstat == 'Employed', 1, 0))
data <- data %>% mutate(isEmployed = as.factor(isEmployed))
summary(data$isEmployed)
```

```
##    0    1
## 5478 4662
```

```r
summary(data)
```

```
##       year         statefip            met2013              perwt
##  Min.   :2019   Length:10140       Length:10140       Min.   :    2.0
##  1st Qu.:2019   Class :character   Class :character   1st Qu.:   57.0
##  Median :2019   Mode  :character   Mode  :character   Median :   85.0
```

```
##   Mean   :2019                                        Mean    : 115.6
##   3rd Qu.:2019                                        3rd Qu.: 133.0
##   Max.   :2019                                        Max.   :1977.0
##       sex                age                race               hispan
##   Length:10140       Length:10140       Length:10140       Length:10140
##   Class :character   Class :character   Class :character   Class :character
##   Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       bpl                educd              empstat            uhrswork
##   Length:10140       Length:10140       Length:10140       Length:10140
##   Class :character   Class :character   Class :character   Class :character
##   Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##       inctot             incwage         isAfricanAmerican isFemale isEmployed
##   Min.   :      0   Min.   :      0     0:6241            0:4823   0:5478
##   1st Qu.:  11775   1st Qu.:      0     1:3899            1:5317   1:4662
##   Median :  35000   Median :  25000
##   Mean   :1744935   Mean   :209709
##   3rd Qu.:  94000   3rd Qu.:  88250
##   Max.   :9999999   Max.   :999999
```

```r
data <- data %>% mutate(age = as.numeric(age)) #formatting age to be numeric using as.numeric
```

```
## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
```

```r
question1_data <- data %>% filter(data$isFemale == '1' & data$isAfricanAmerican == '1' & data$isEmployed

summary(question1_data)
```

```
##      incwage            age
##   Min.   :     0   Min.   :16.00
##   1st Qu.: 15000   1st Qu.:33.00
##   Median : 29000   Median :44.00
##   Mean   : 33869   Mean   :44.21
##   3rd Qu.: 46000   3rd Qu.:56.00
##   Max.   :476000   Max.   :86.00
```

```r
corr <- function(x, indices){
  data <- x[indices, ]
  corr <- cor(data[,1], data[,2])
  return( corr )
}

coefficient <- boot(question1_data[],
                    corr,
                    10000)
```

```
coefficient
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = question1_data[], statistic = corr, R = 10000)
##
##
## Bootstrap Statistics :
##      original       bias    std. error
## t1* 0.1497379 0.0008852872  0.02755512
```

#The standard error of the correlation coefficient between labor wages, worker ages is ~ 0.0272992

####Question 2

#Filter white, men, employed

```
question2_data <- data %>% filter(data$race == "White" & data$isEmployed == '1' & data$isFemale == '0')

coefficient <- boot(question2_data[],
                    corr,
                    10000)


coefficient
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = question2_data[], statistic = corr, R = 10000)
##
##
## Bootstrap Statistics :
##      original       bias    std. error
## t1* 0.1748645 0.0004523001  0.02119825
```

#The standard error of the correlation coefficient between labor wages, worker ages is ~ 0.02086219