3818 R Project Example

*** Student Name ***

8/21/2020

I am looking through the American Consumer Survey for the year 2016 in Colorado.

```r
#load the data into R from your computer:
# similar for PC's, use the other backslash though to find where your file is located, right click on f
ACS_df <- read.csv("~/Desktop/3818/R Project/example_project/ACS_2016_CO.csv")
```

The dataset has 52333 observations of individuals in Colorado and for each individual there are 39 variables. These variables cover worker characteristics, familial characteristics, and income. This is a one-time survey, so the data is cross-sectional, i.e. we only see one observation per individual. We will consider total family income in $ and categorical variables for gender, marital status, and education level.

For the total family income, after removing missing observations that were classified as an income of $9,999,999, we have the following summary statistics

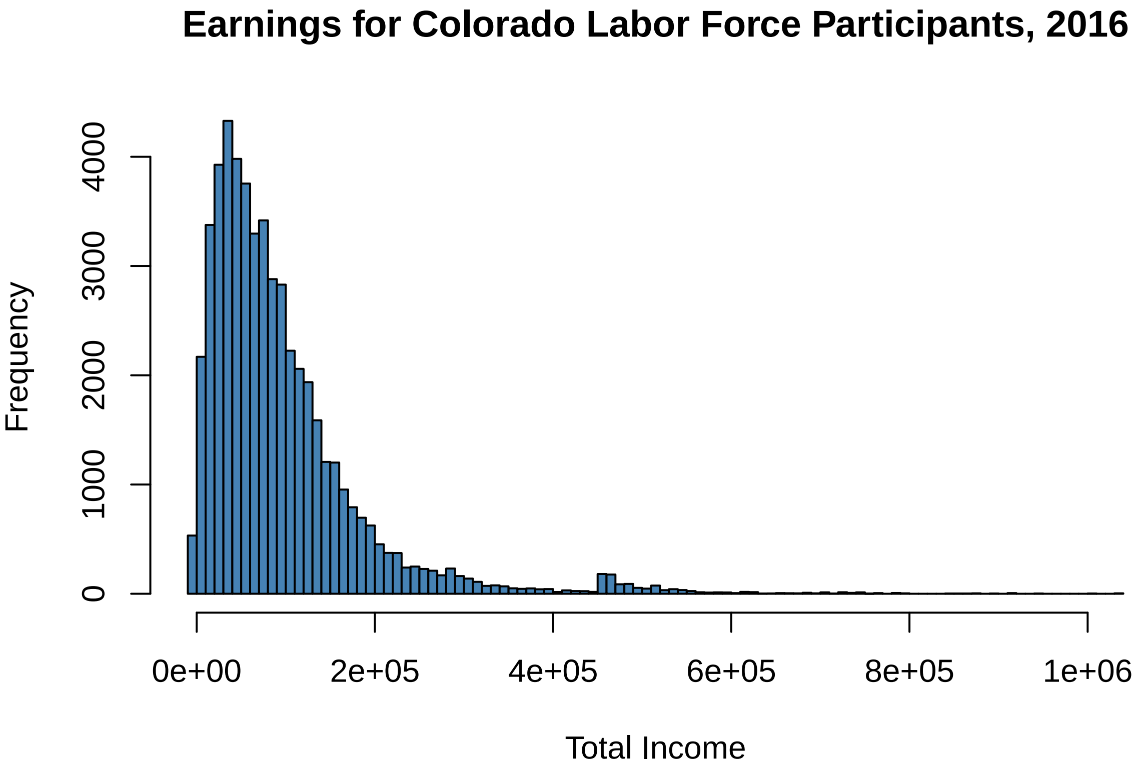
- Max: 1034000
- Min: -6700
- Mean: 9.6163531×10^4
- Standard Deviation: 9.34456×10^4

Among married families, the mean income raises to 1.1462141×10^5 which is larger than the total sample and the standard deviation lowers to 9.9175025×10^4. The maximum and minimum values don't change.

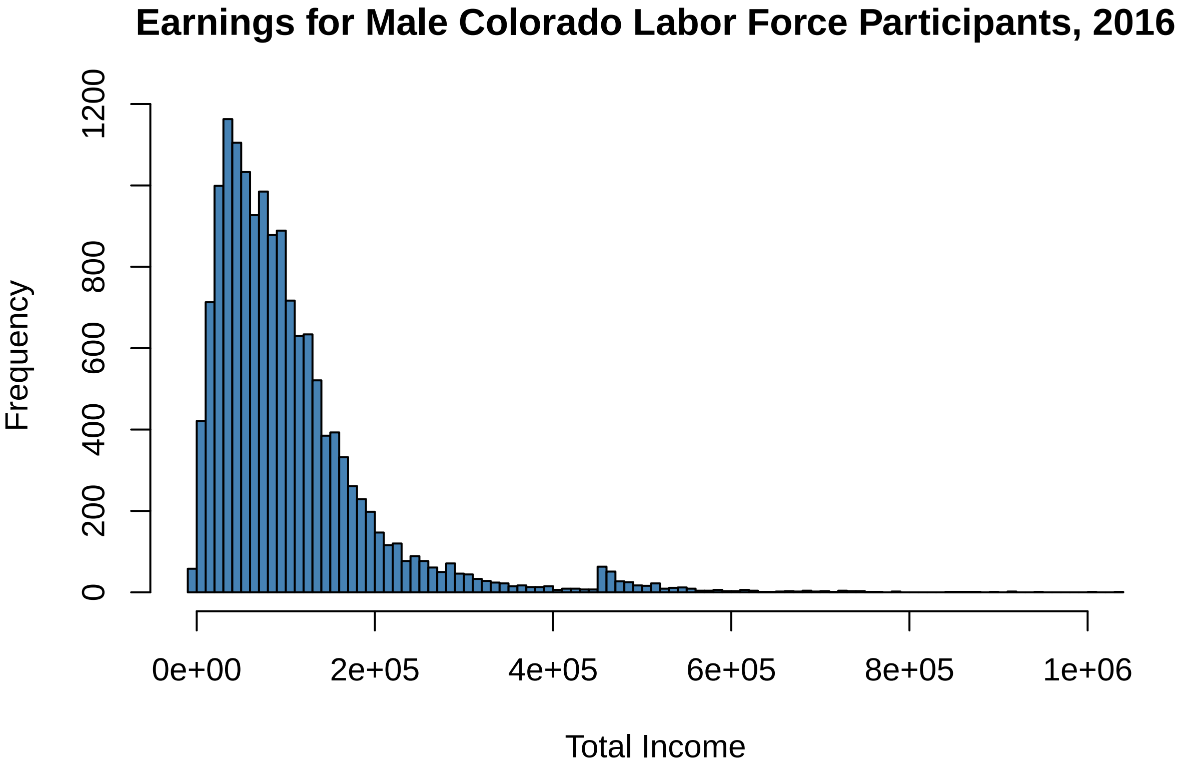
Now, I turn to look at only people who are active in the labor force and study how their income varies with different characteristics.

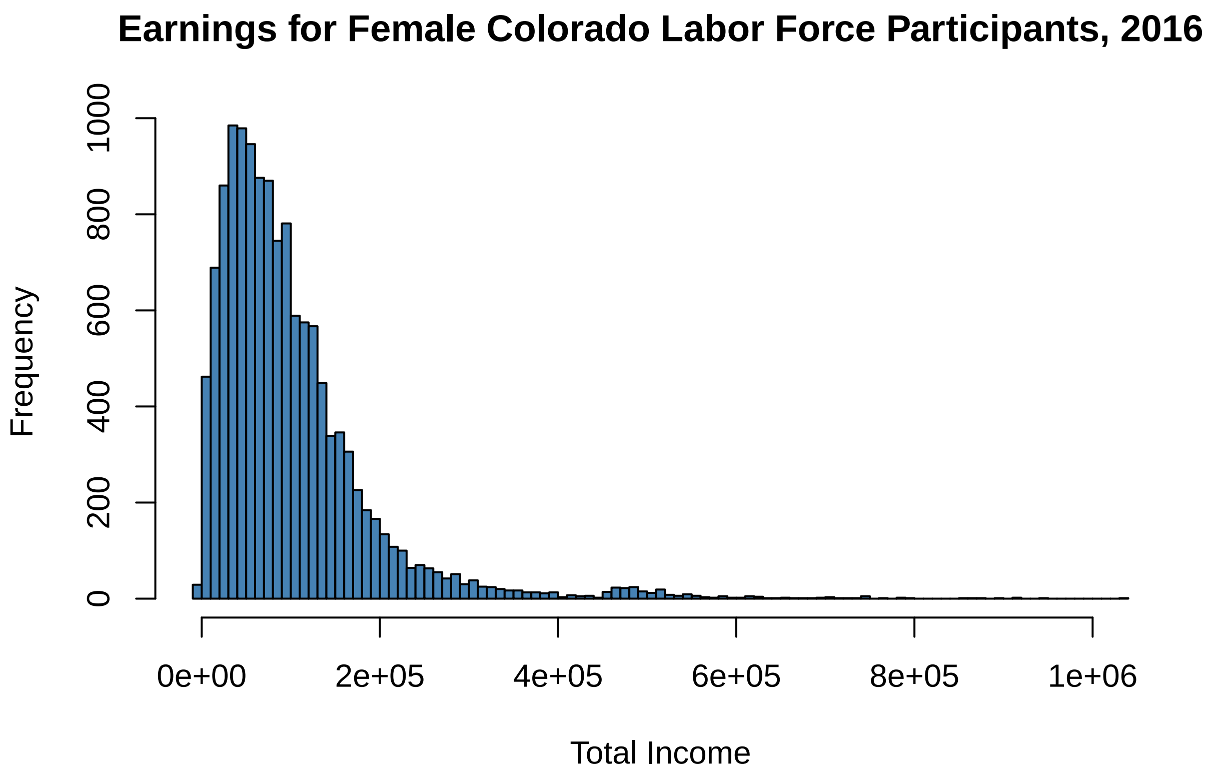
Distribution of Income within the Labor Force

Let's look at the distribution of income among the entire Labor Force sample first.



The first thing I'm considering is to see if the distribution of earnings differ between men and women. As you can see in the graph below, the distribution looks quite similar between females and male workers in Colorado.





The confidence intervals for mean income for men and women, respectively is:

```r
c(mean(ACS_df_lf_men$FTOTINC) - qnorm(.975) * sd(ACS_df_lf_men$FTOTINC)/sqrt(nrow(ACS_df_lf_men)),
    mean(ACS_df_lf_men$FTOTINC) + qnorm(.975) * sd(ACS_df_lf_men$FTOTINC)/sqrt(nrow(ACS_df_lf_men)))
```

```
## [1] 102926.5 105996.8
```

```r
c(mean(ACS_df_lf_women$FTOTINC) - qnorm(.975) * sd(ACS_df_lf_women$FTOTINC)/sqrt(nrow(ACS_df_lf_women))
    mean(ACS_df_lf_women$FTOTINC) + qnorm(.975) * sd(ACS_df_lf_women$FTOTINC)/sqrt(nrow(ACS_df_lf_women
```

```
## [1]  98497.43 101573.58
```

Notice that Women's mean income confidence interval is less than the Men's mean income confidence interval. I'm going to test if female income is lower than men's income on average in our data. From the two-sample t-test below, we can reject with 95% confidence that men and women don't earn the same income on average.

```r
# You'll see we have two different datasets representing these groups
## Remember, to conduct a hypothesis test, we'll need the mean, std deviation, and sample size of each
t.test(x= ACS_df_lf_men$FTOTINC, y= ACS_df_lf_women$FTOTINC,
       alternative= "two.sided",
       conf.level= 0.95
      )
```

```
##
##  Welch Two Sample t-test
##
## data:  ACS_df_lf_men$FTOTINC and ACS_df_lf_women$FTOTINC
## t = 3.992, df = 27909, p-value = 6.568e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2252.950 6599.333
## sample estimates:
```

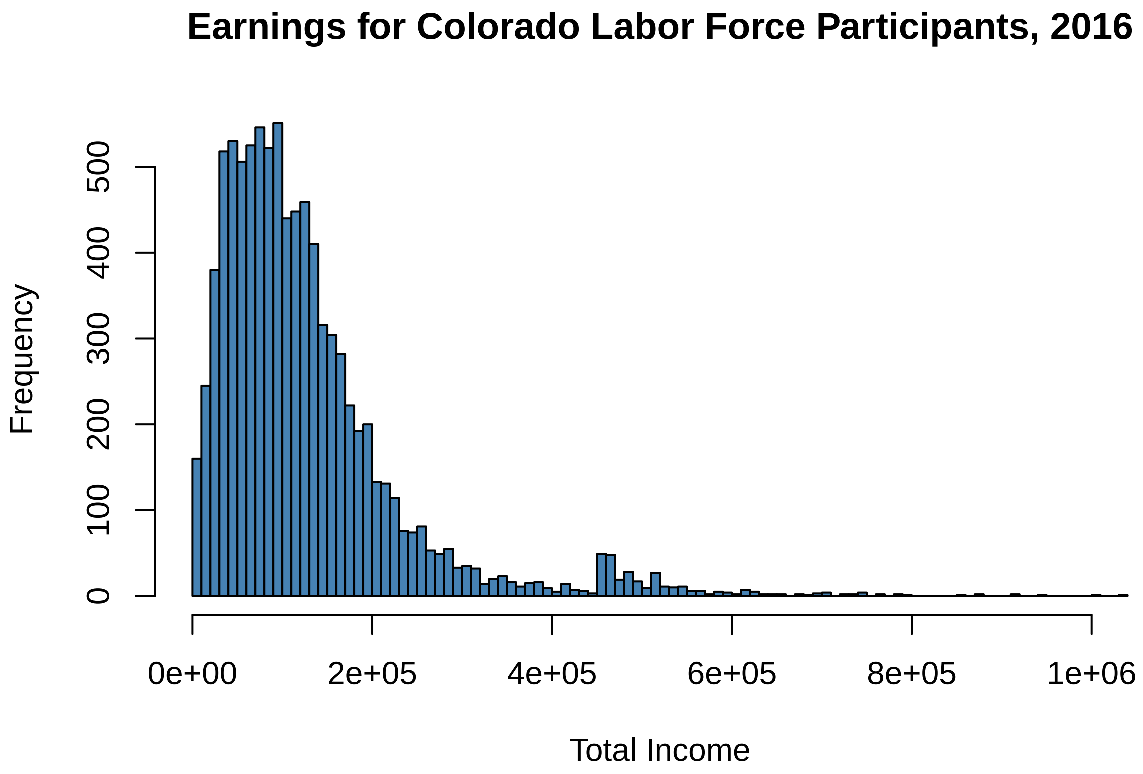
```
## mean of x mean of y
##  104461.6  100035.5
```

Scatter Plot

```
#relationship between age and earnings for college graduates
#let's subset to people of working age who have a degree and are employed
ACS_df_grads <- subset(ACS_df_lf, ACS_df_lf$AGE >= 18 & ACS_df_lf$AGE <= 55 & ACS_df_lf$EDUCD >=101 & AC
length(ACS_df_grads$FTOTINC)
```

```
## [1] 9084
```

```
#what does the distribution of earnings look like?
hist(ACS_df_grads$FTOTINC, main = "Earnings for Colorado Labor Force Participants, 2016"
     ,xlab = "Total Income", col = "steelblue", breaks = 100)
```

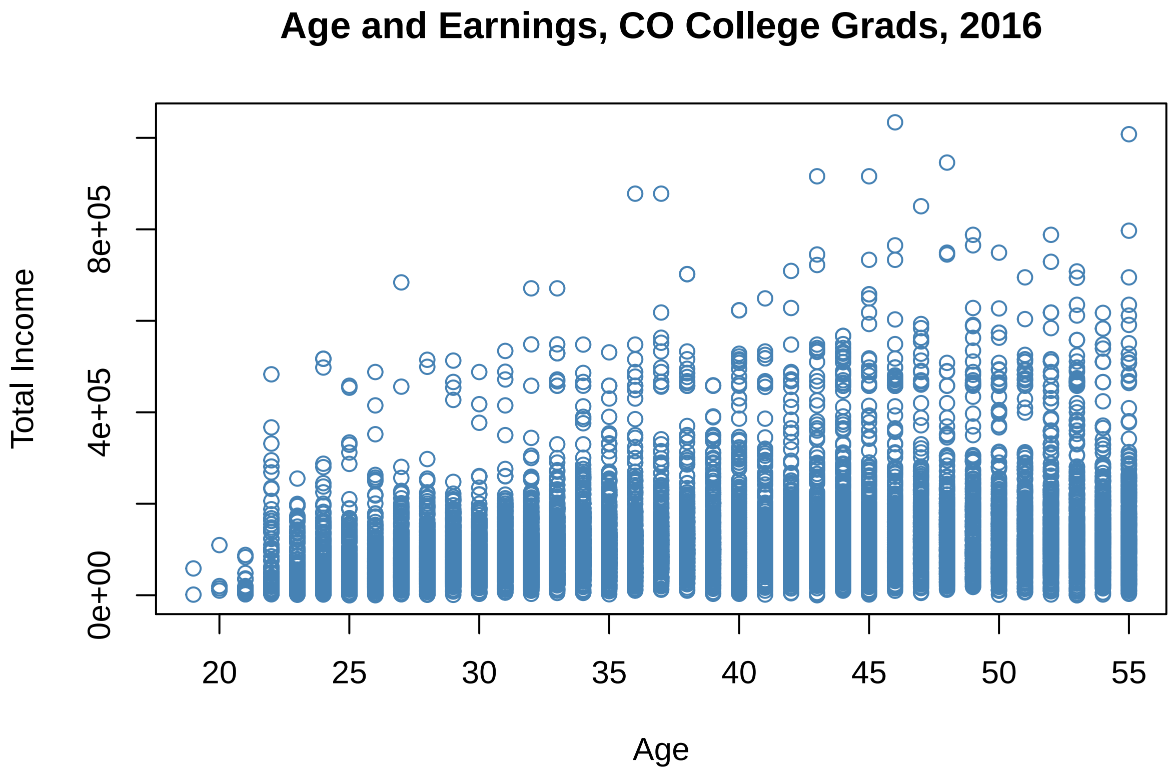


```
#let's drop outliers
#how do we make a scatterplot?
```

```
help("plot")
```

```
## Help on topic 'plot' was found in the following packages:
##
##   Package               Library
##   graphics              /Library/Frameworks/R.framework/Versions/4.0/Resources/library
##   base                  /Library/Frameworks/R.framework/Resources/library
##
##
## Using the first match ...
```

```r
plot(ACS_df_grads$AGE, ACS_df_grads$FTOTINC,
        main = "Age and Earnings, CO College Grads, 2016",
        xlab = "Age", ylab = "Total Income", col = "steelblue"
     )
```


Age and Earnings, CO College Grads, 2016
Total Income
0e+00
4e+05
8e+05
Age