CS 544
Final Project

Adults Census Income

Tzupin Kuo

U36-17-2777

Introduction – Adults Census Income

Topic:

An individual's annual income results from various factors. Intuitively, it is influenced by the individual's education level, age, gender, occupation, and etc.

Source From:

https://www.kaggle.com/uciml/adult-census-income

Acknowledgement:

This data was extracted from the 1994 Census bureau database by Ronny Kohavi and Barry Becker (**UCI machine learning repository)**

Introduction

15 Attributes

Continuous - age, fnlwgt(final weight), educational_num, capital.gain, capital.loss, hours.per.week

Categorical - occupation, relationship, race, gender, native.country, workclass, education, marital.status

Class - Income (1. >50K 2. <=50K)

Imputing the missing values

The missing values contribute much to this attribute **workclass**

```
> sort(table(adult$workclass))

    Never-worked      Without-pay      Federal-gov     Self-emp-inc        State-gov                ?        Local-gov
              10               21             1432             1695             1981             2799             3136
Self-emp-not-inc          Private
            3862            33906
```

Imputing the missing values makes the attribute **occupation** biased towards the upper end

```
> sort(table(adult$occupation))

    Armed-Forces   Priv-house-serv   Protective-serv     Tech-support   Farming-fishing Handlers-cleaners  Transport-moving                 ? Machine-op-inspct    Other-service
              15               242               983              1446              1490              2072              2355              2809              3022              4923
           Sales     Adm-clerical   Exec-managerial     Craft-repair    Prof-specialty
            5504              5611              6086              6112              6172
```

Imputing the missing values

native.country has the third maximum and lower end is much low compared to missing values

```
> sort(table(adult$native.country))

       Holand-Netherlands                   Hungary                  Honduras                  Scotland                      Laos Outlying-US(Guam-USVI-etc)
                        1                        19                        20                        21                        23                        23
               Yugoslavia           Trinadad&Tobago                  Cambodia                      Hong                  Thailand                   Ireland
                       23                        27                        28                        30                        30                        37
                   France                   Ecuador                      Peru                    Greece                 Nicaragua                      Iran
                       38                        45                        46                        49                        49                        59
                   Taiwan                  Portugal                     Haiti                  Columbia                   Vietnam                    Poland
                       65                        67                        75                        85                        86                        87
                Guatemala                     Japan    Dominican-Republic                     Italy                   Jamaica                     South
                       88                        92                       103                       105                       106                       115
                    China                   England                      Cuba                     India                 El-Salvador                    Canada
                      122                       127                       138                       151                       155                       182
              Puerto-Rico                   Germany               Philippines                         ?                    Mexico             United-States
                      184                       206                       295                       857                       951                     43832
>
```

Removing the missing values

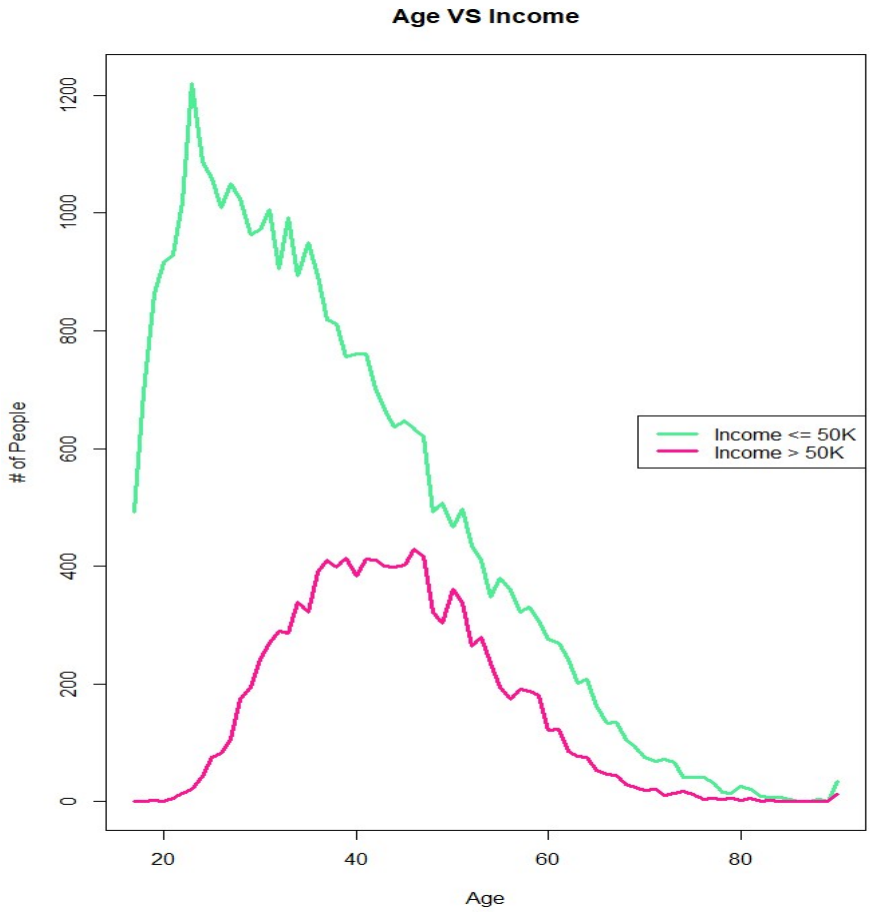
Because we found out that imputing missing values is not the best option, we need to remove the rows of data of those missing values.

Final cleaned dataset has 45,222

- class(income <= 50K) has 34,013 (75.21%) (original: 37,155)
- class(income >50K) has 11,208 (24.79%) (original: 11,687)

Analysis – Age vs Income

Age ranges from 17 - 90



Analysis – Gender vs Income

Male

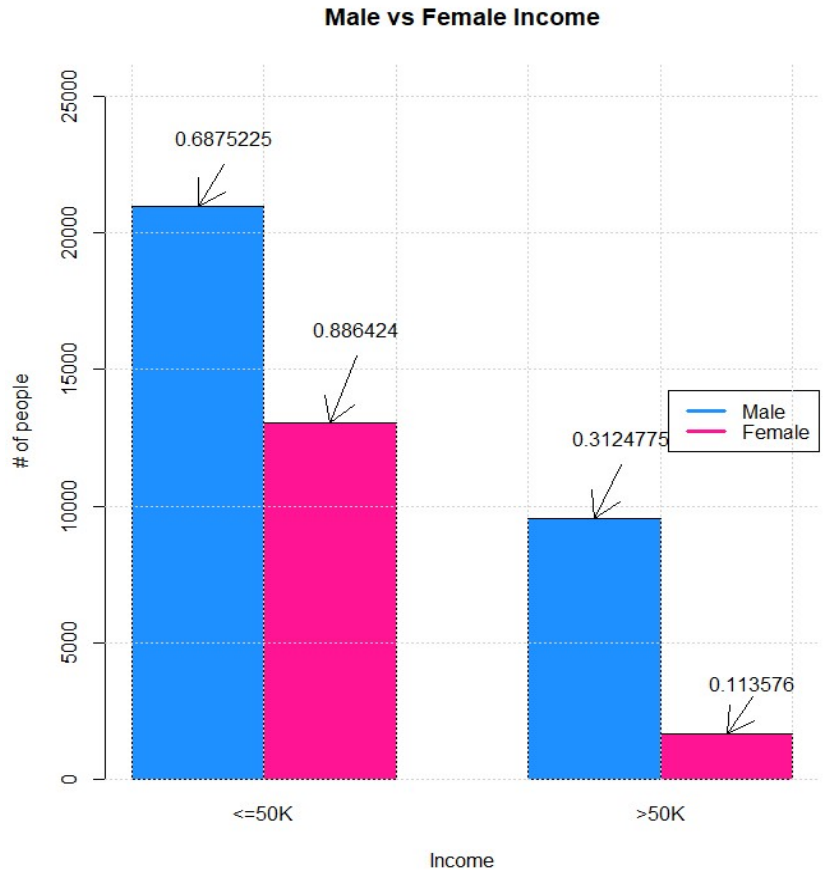
<= 50 K : 20,988

> 50 K : 9,539

Female

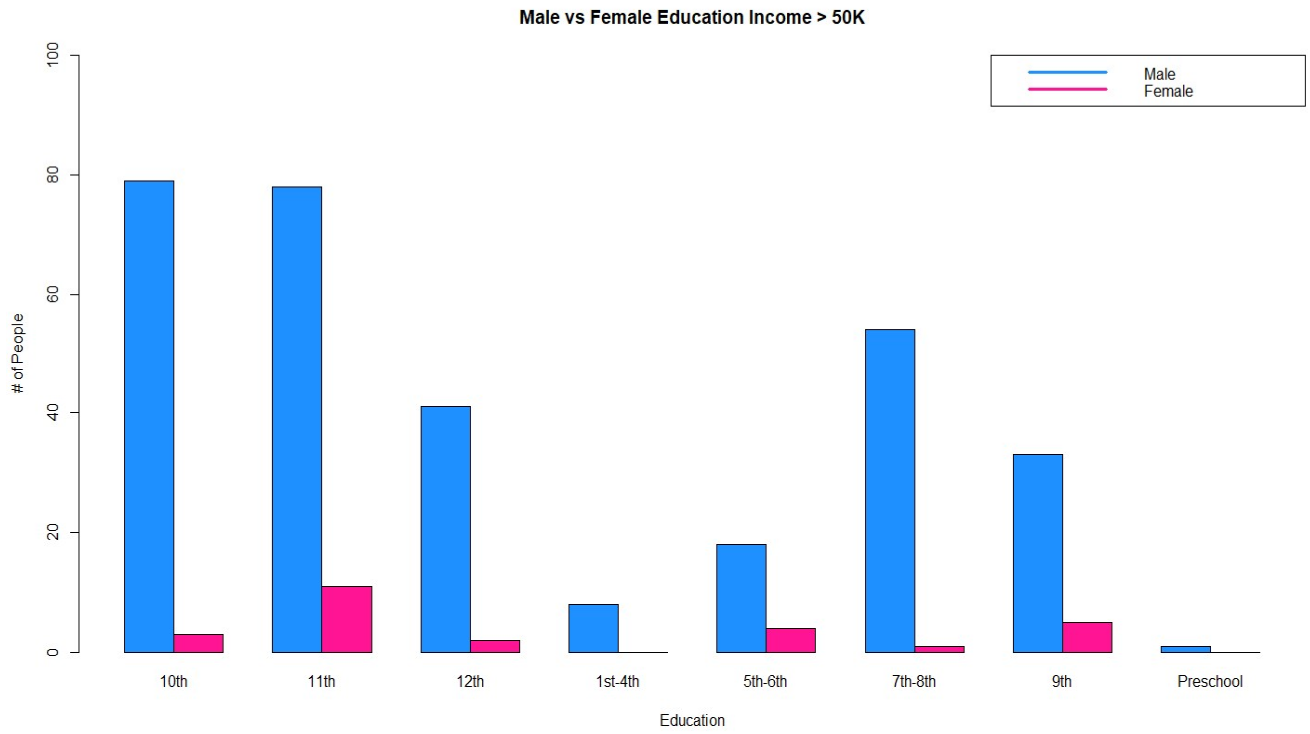
<= 50 K : 13,026

> 50 K : 1,669



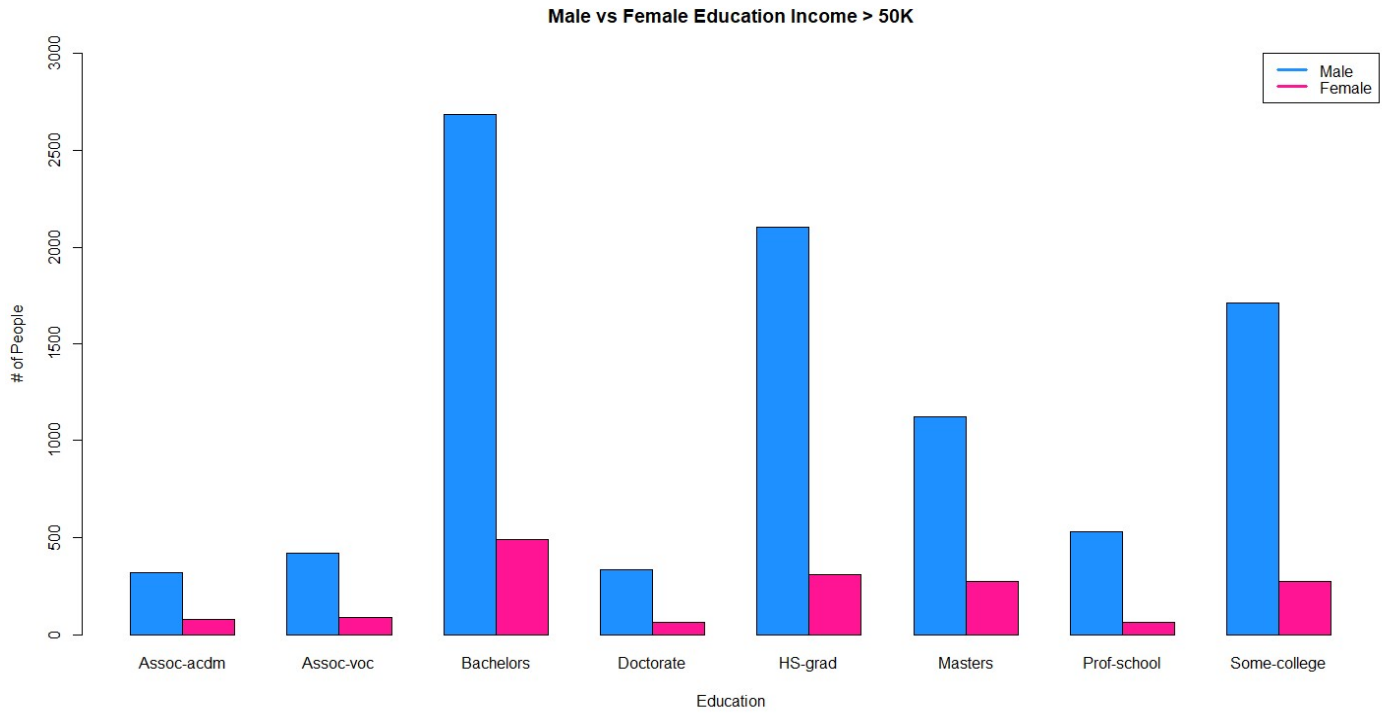
Analysis – Education vs Income

People who earned > 50K per year and received education up to high school



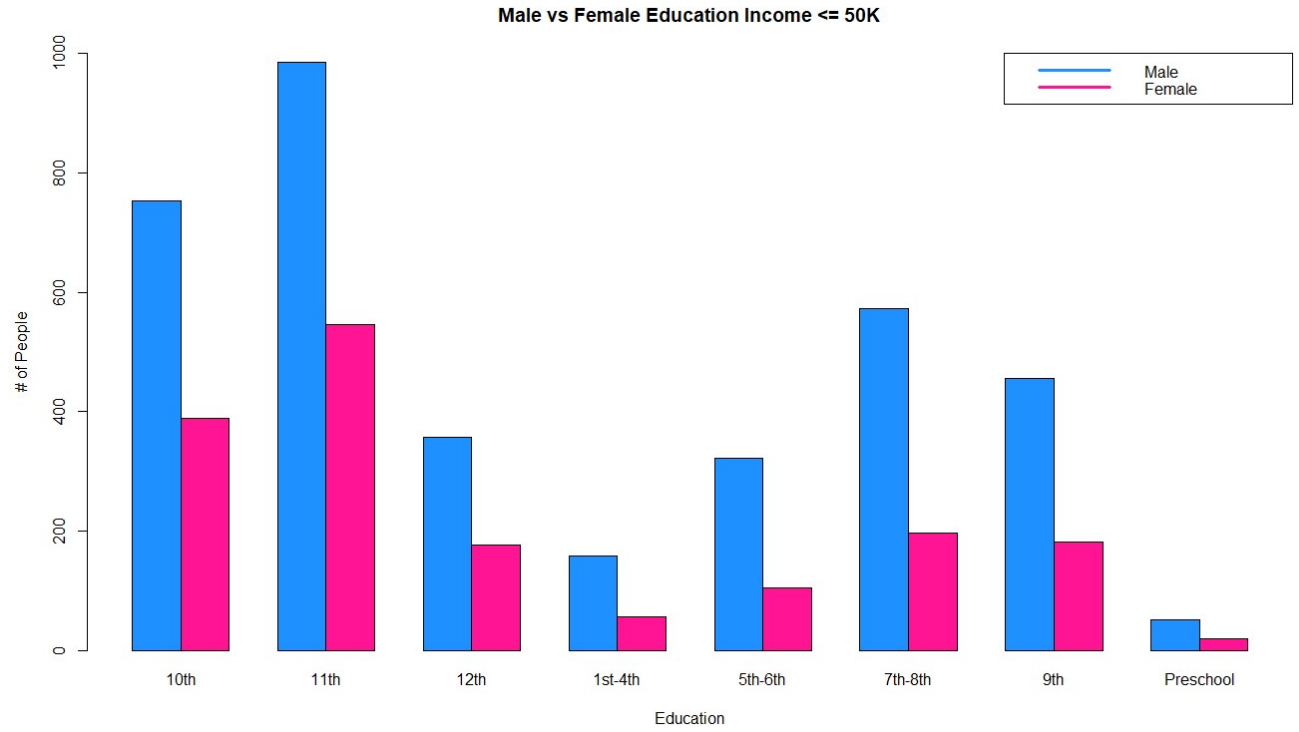
Analysis – Education vs Income

People who earned > 50K per year and received education after high school



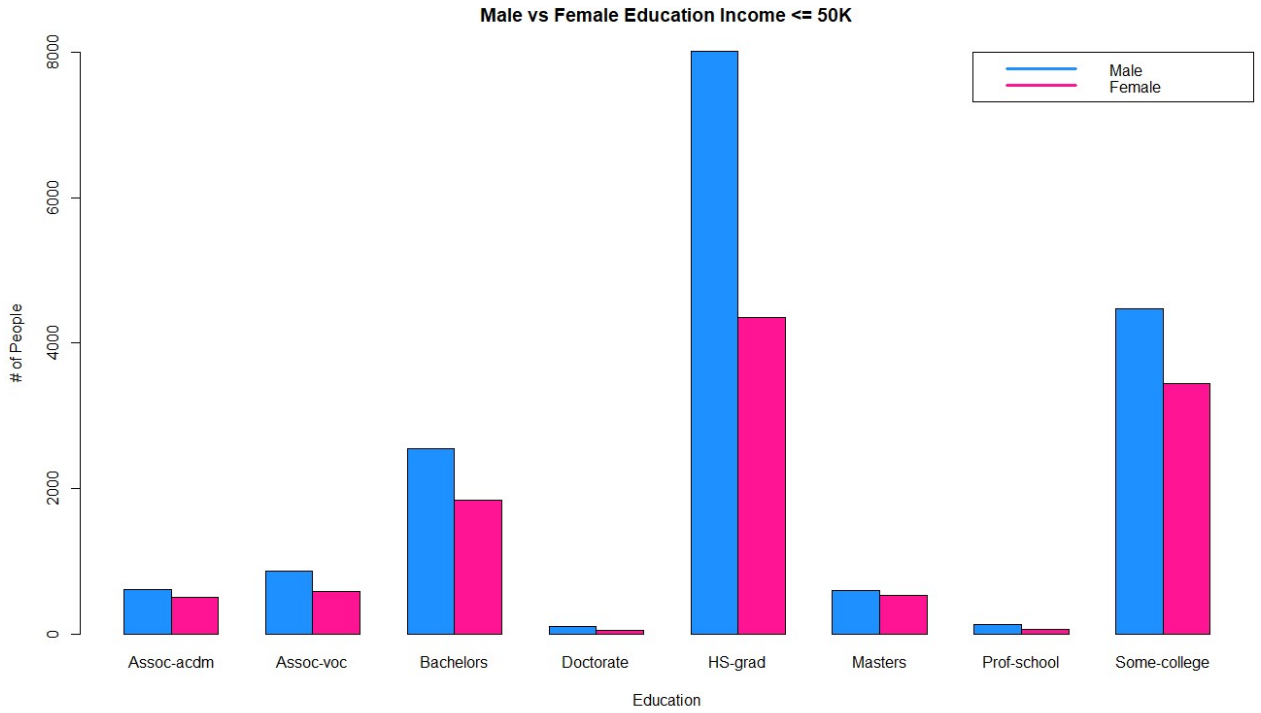
Analysis – Education vs Income

People who earned <= 50K per year and received education up to high school

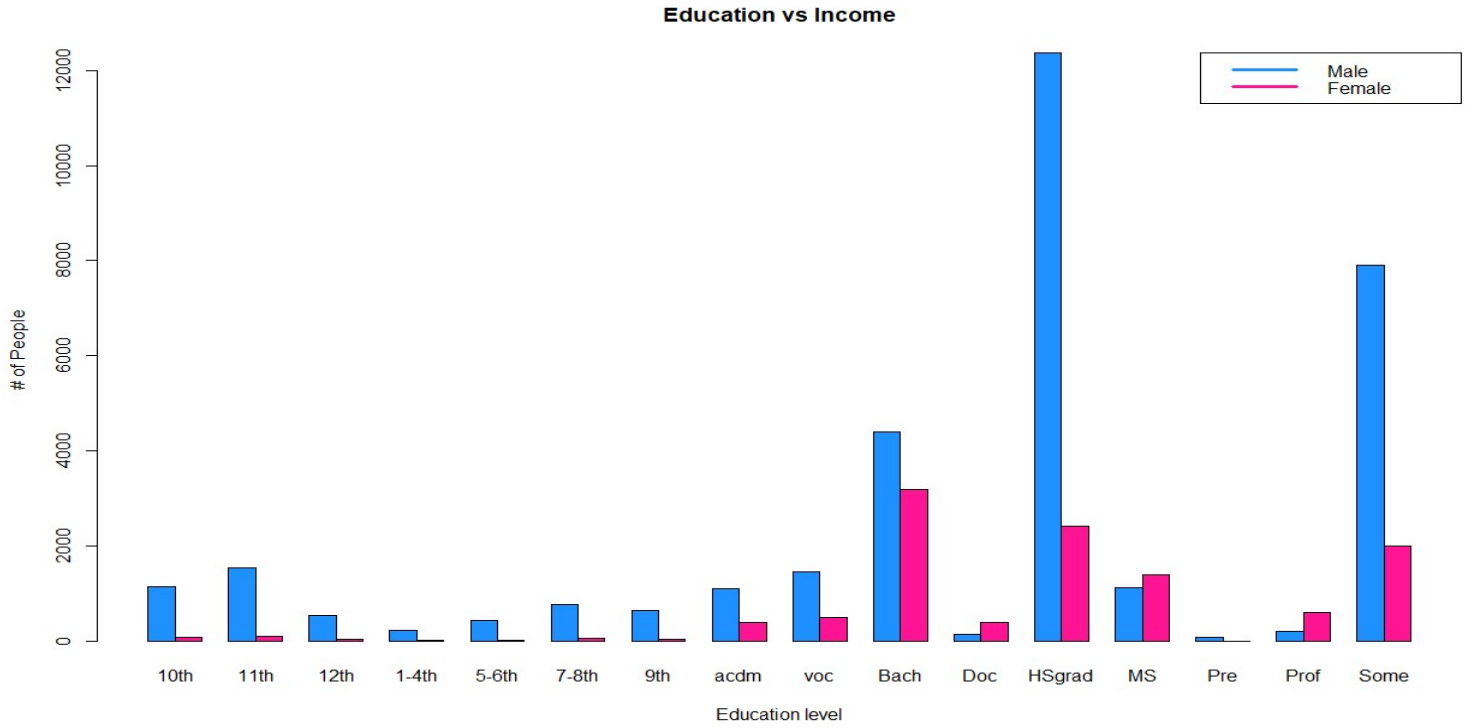


Analysis – Education vs Income

People who earned <= 50K per year and received education after high school

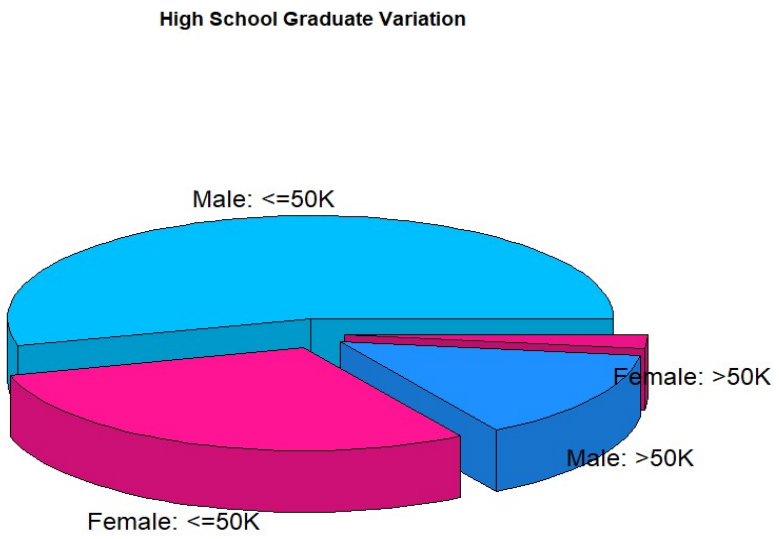


Analysis – Education vs Income



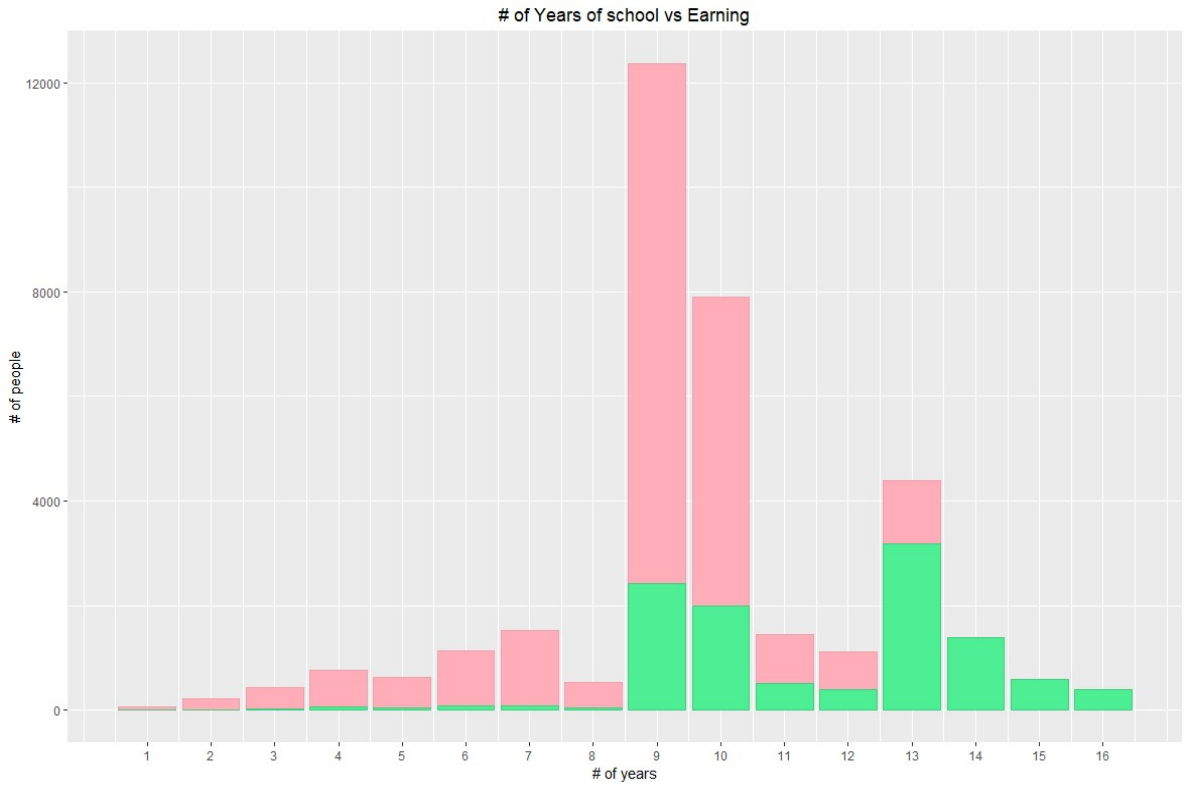
Analysis – Education vs Income

High School Graduate Variation Analysis

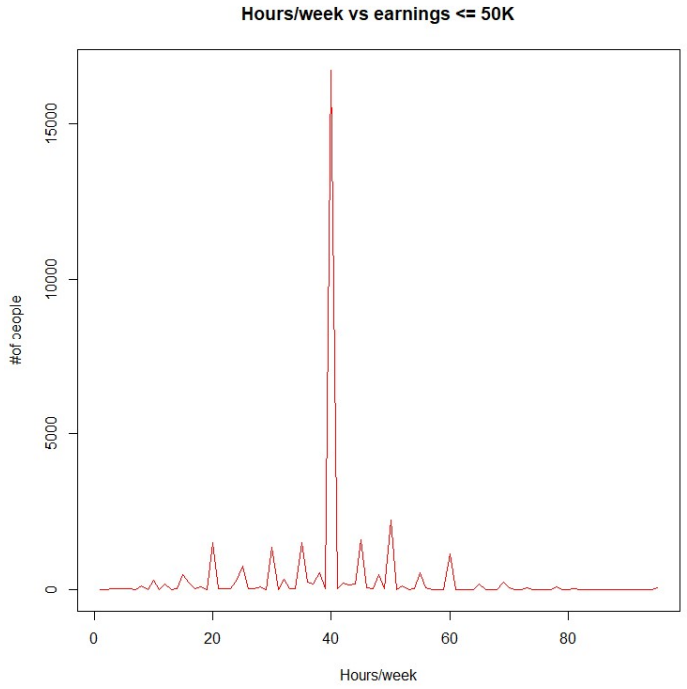


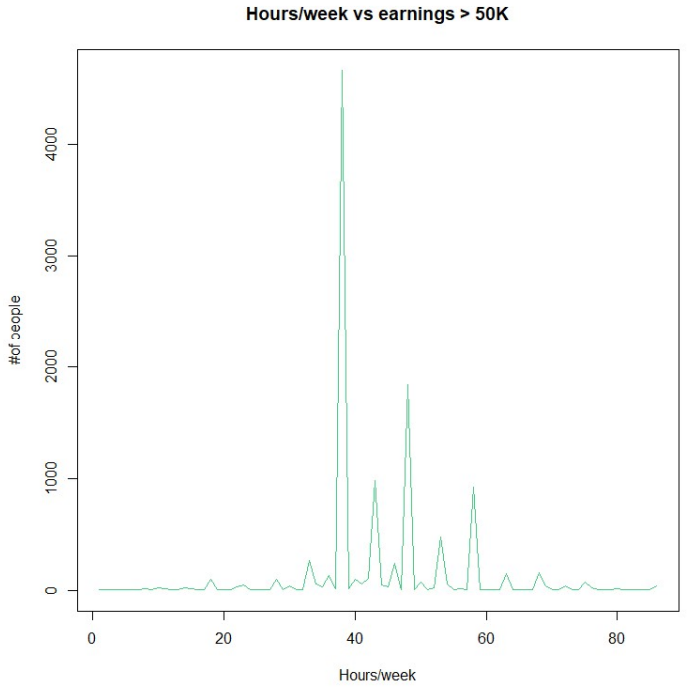
Analysis – Years of Education vs Income

Green : Income >50K
Red: Income <=50K



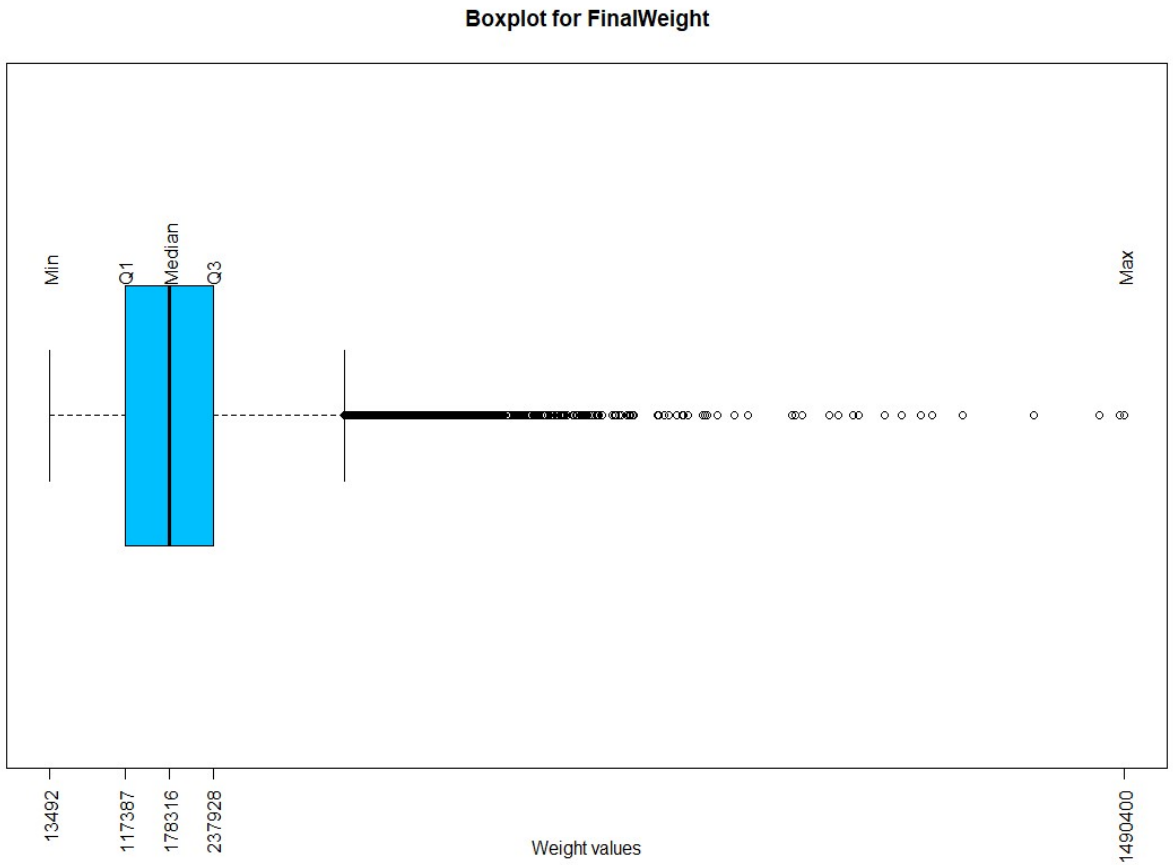
Analysis – Hours/Week vs Income





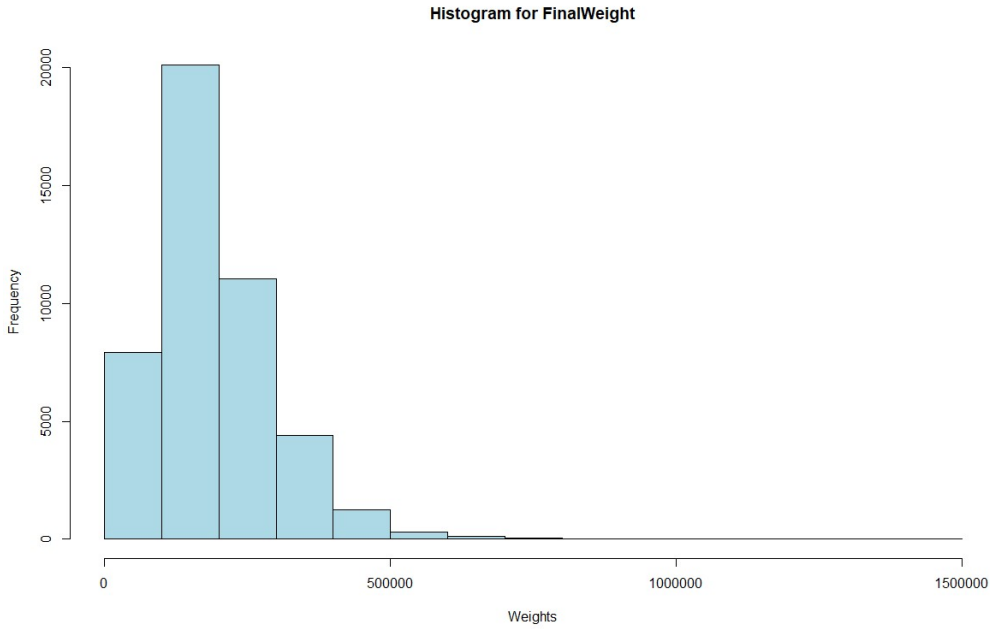
Analysis – Final Weights

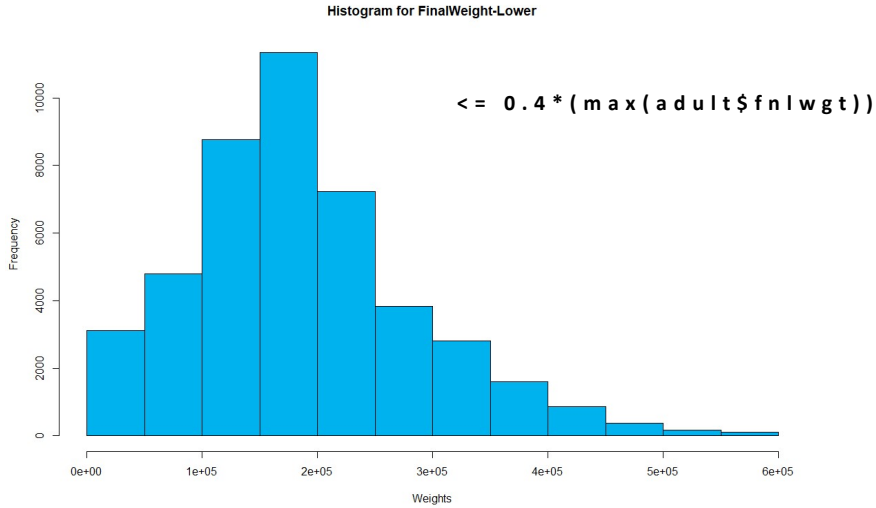
- Main predictor of the class
- Most values are in the lower end
- Need to split the set into upper/lower

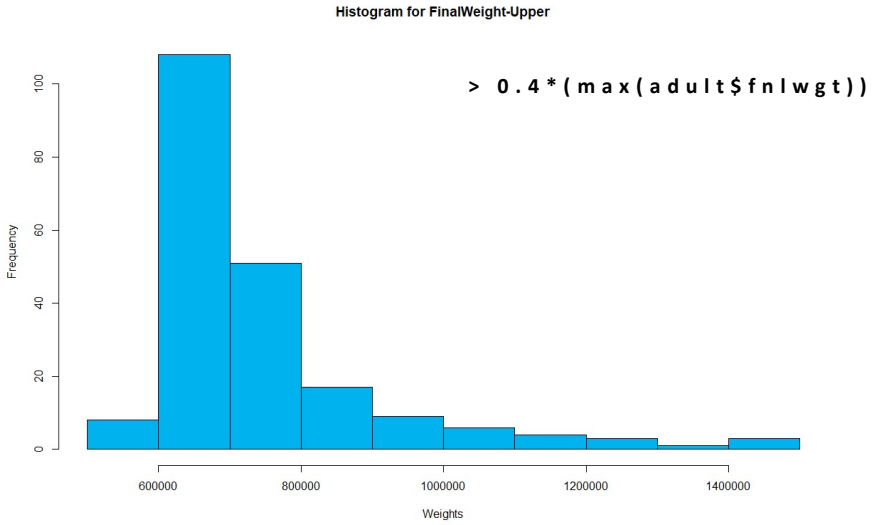


Analysis – Final Weights

- **Distribution of the values in this attribute is wide-spread**

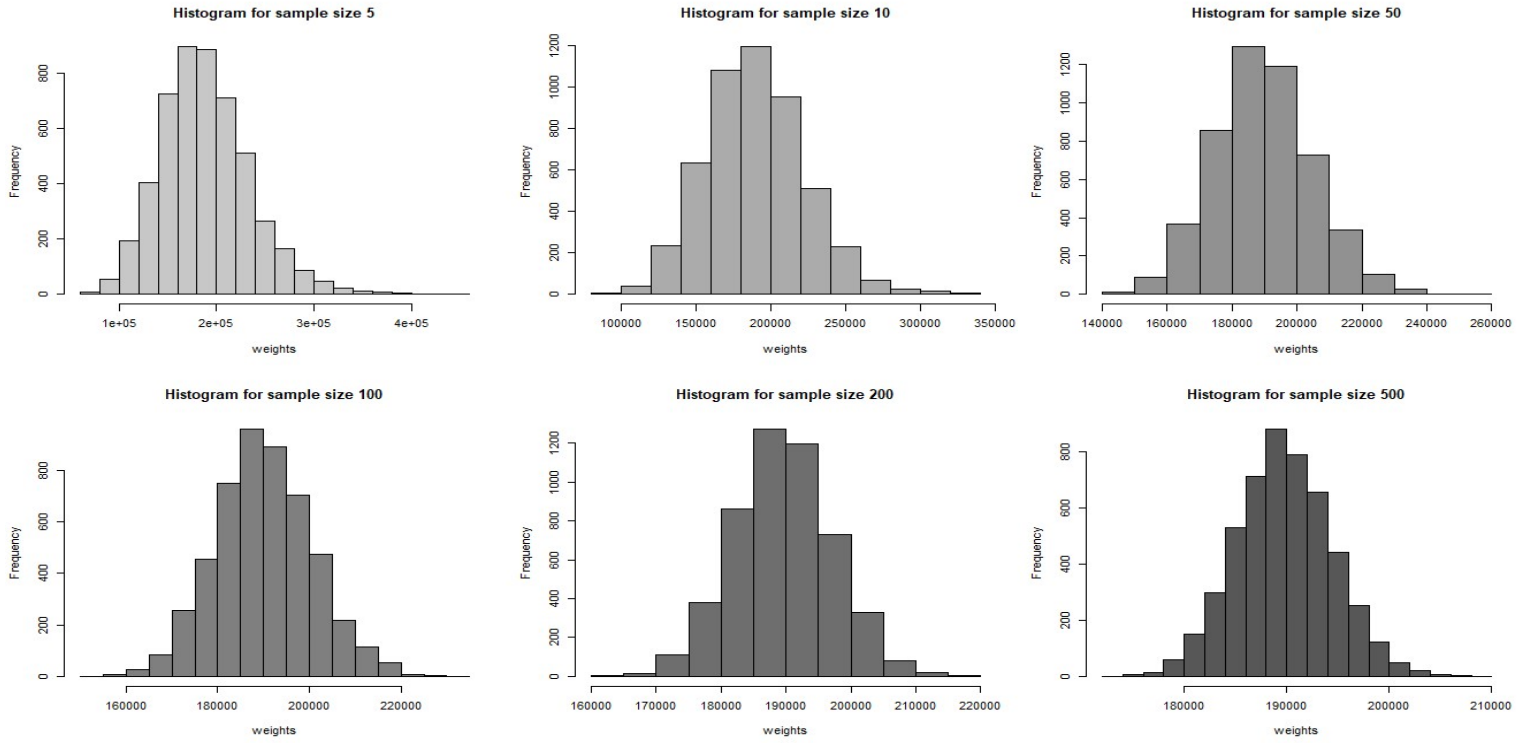






Central Limit Theorem

- **Distribution of Means for various sample size**



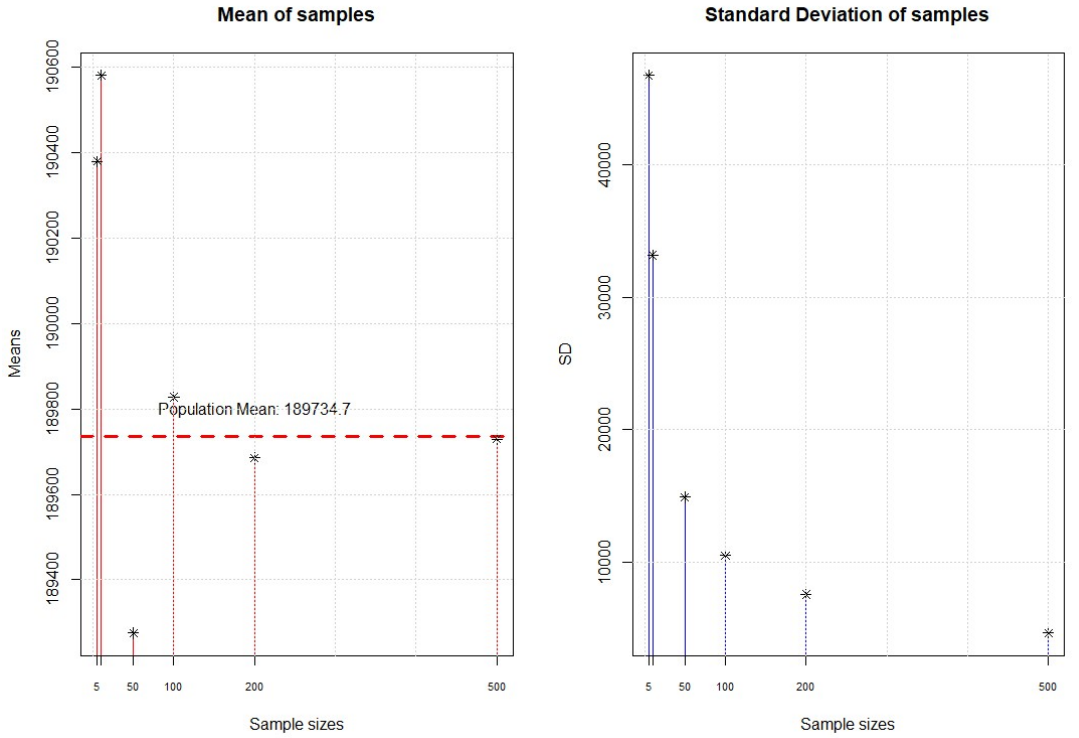
Central Limit Theorem

Mean and Standard Deviation of the samples

- **Population standard deviation: 105,639.2**

Sample size increases

- **The mean of samples are around the population mean**
- **The standard deviation of samples decreases**



Sampling

Sampling results

```
> unique(adult$native.country) # All conuntries taken by the population
 [1] "United-States"          "Peru"                   "Guatemala"              "Mexico"                 "Dominican-Republic"
 [6] "Ireland"                "Germany"                "Philippines"            "Thailand"               "Haiti"
[11] "El-Salvador"            "Puerto-Rico"            "Vietnam"                "South"                  "Columbia"
[16] "Japan"                  "India"                  "Cambodia"               "Poland"                 "Laos"
[21] "England"                "Cuba"                   "Taiwan"                 "Italy"                  "Canada"
[26] "Portugal"               "China"                  "Nicaragua"              "Honduras"               "Iran"
[31] "Scotland"               "Jamaica"                "Ecuador"                "Yugoslavia"             "Hungary"
[36] "Hong"                   "Greece"                 "Trinadad&Tobago"        "Outlying-US(Guam-USVI-etc)" "France"
[41] "Holand-Netherlands"
> length(unique(adult$native.country)) # the number of all countries, which is 41
[1] 41
>
```

```
> nrow(table(adult.srs.w.rep$native.country))
[1] 40
> nrow(table(adult.srs.wo.rep$native.country))
[1] 40
> nrow(table(adult.sys$native.country))
[1] 40
> nrow(table(adult.sys.unequal$native.country))
[1] 38
> nrow(table(adult.strata$native.country))
[1] 41
```

Stratified sampling effectively picked up the samples with all 41 countries included

Confidence Intervals

Confidence Intervals – 80% & 90%

Population Mean falls inside the Confidence Intervals

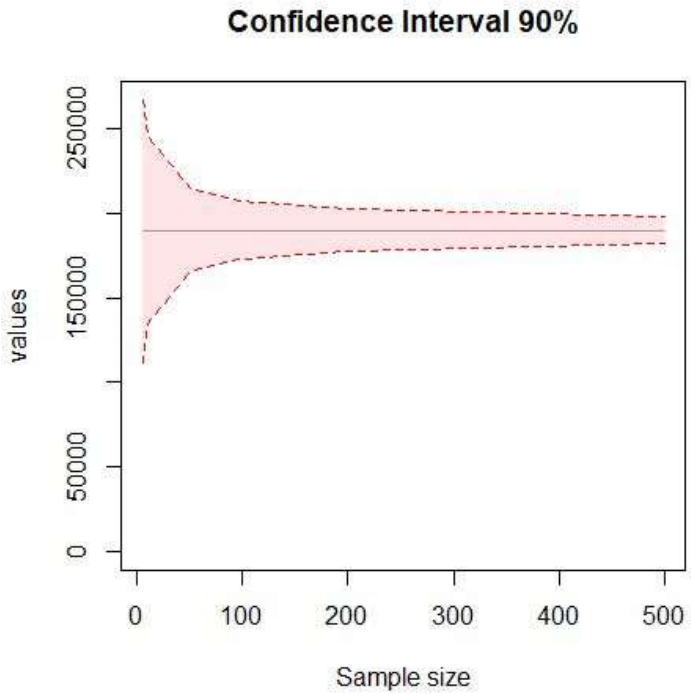
- **Population Mean: 189734.7**

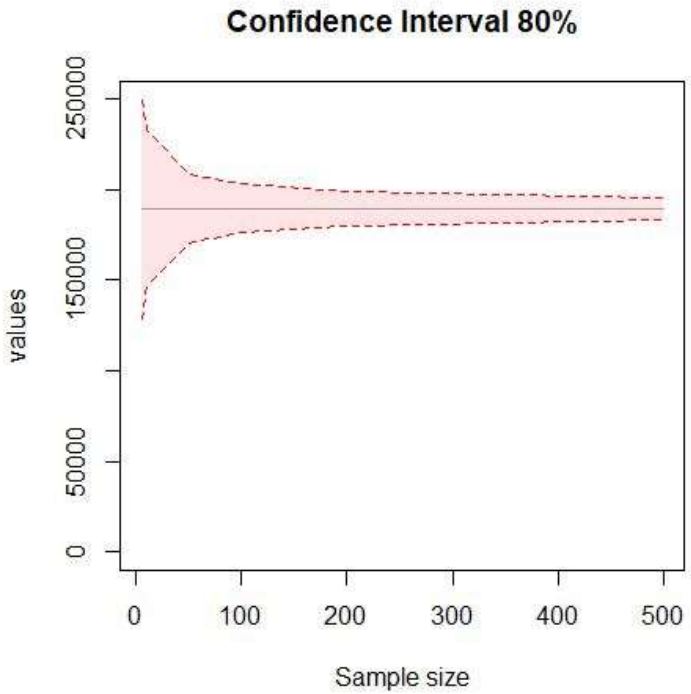
```
Sample.size :  5
80% confidence intervals = 93168.6  -  214111.4
80% confidence intervals = 193268.8  -  314211.6
80% confidence intervals = 275804.6  -  396747.4
Sample.size :  10
80% confidence intervals = 104354.3  -  189873.7
80% confidence intervals = 140879.1  -  226398.5
80% confidence intervals = 137640.6  -  223160
Sample.size :  50
80% confidence intervals = 168829.3  -  207074.8
80% confidence intervals = 178119.7  -  216365.2
80% confidence intervals = 191721.9  -  229967.4
Sample.size :  100
80% confidence intervals = 169269.3  -  196312.9
80% confidence intervals = 189430.8  -  216474.4
80% confidence intervals = 180042  -  207085.7
Sample.size :  200
80% confidence intervals = 182669.5  -  201792.2
80% confidence intervals = 183368.9  -  202491.6
80% confidence intervals = 174512.3  -  193635
Sample.size :  500
80% confidence intervals = 181294.1  -  193388.4
80% confidence intervals = 188696.8  -  200791.1
80% confidence intervals = 183644.2  -  195738.5
```

```
Sample.size :  5
90% confidence intervals = 174330  -  329760.4
90% confidence intervals = 143400.2  -  298830.6
90% confidence intervals = 75647.2  -  231077.6
Sample.size :  10
90% confidence intervals = 97273.75  -  207179.6
90% confidence intervals = 138113.6  -  248019.4
90% confidence intervals = 140349.2  -  250255
Sample.size :  50
90% confidence intervals = 156566.8  -  205718.2
90% confidence intervals = 168660.6  -  217812
90% confidence intervals = 155145.9  -  204297.3
Sample.size :  100
90% confidence intervals = 175087.4  -  209842.7
90% confidence intervals = 166754.7  -  201510
90% confidence intervals = 180906.4  -  215661.7
Sample.size :  200
90% confidence intervals = 190176.6  -  214752.3
90% confidence intervals = 181088.6  -  205664.3
90% confidence intervals = 185063.6  -  209639.3
Sample.size :  500
90% confidence intervals = 178699.3  -  194242.3
90% confidence intervals = 179941.4  -  195484.5
90% confidence intervals = 174830.2  -  190373.2
```

Confidence Intervals

Graph for CI Limits





Boston University

CS 544

Thank You

Foundations of Analytics

Final Project

Dataset: Adult Census Income

Student: Tzupin Kuo

StudentID: U36-17-2777