RAVI K.RAJENDRAN
U30-05-9012

ADULTS

ABOUT DATASET

- Source : SGI Boston.
- Survey about the adults -age(17-100)
- Has about 48841 entries and 15 attributes
- Binary classification dataset

ABOUT DATASET

- Has two main classes (Earnings >50K and Earnings <=50K)
- Out of 48841 data
  - class(Earnings < 50K) 37154 76.07%
  - class(Earnings =>50K) 11687 23.93%
- Imbalanced Dataset
- Having missing values ‘?’

ABOUT DATASET

- 15 Attributes
- Continuous- Age, Final_weight, Education_Number, CapGain, CapLoss, Hours/week
- Categorical- Occupation(14), Relationship(6), Race(5), Sex(2), Native(41), Work_class(8), Education(12),Marital_status(7)
- Class - Earning(2)

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            1980            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
               15                242                983               1446               1490               2072
  Transport-moving                ?  Machine-op-inspct      Other-service              Sales       Adm-clerical
             2355               2809               3022               4923               5504               5610
   Exec-managerial      Craft-repair     Prof-specialty
             6086               6112               6172
```

IMPUTING THE MISSING VALUES

Here, NATIVE attribute has the third maximum and lower end is much low compared to missing values

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

        Holland-Netherlands                    Hungary                   Honduras                   Scotland
                          1                         19                         20                         21
                       Laos  Outlying-US(Guam-USVI-etc)                 Yugoslavia             Trinadad&Tobago
                         23                         23                         23                         27
                   Cambodia                       Hong                   Thailand                    Ireland
                         28                         30                         30                         37
                     France                    Ecuador                       Peru                     Greece
                         38                         45                         46                         49
                  Nicaragua                       Iran                     Taiwan                   Portugal
                         49                         59                         65                         67
                      Haiti                   Columbia                    Vietnam                     Poland
                         75                         85                         86                         87
                  Guatemala                      Japan         Dominican-Republic                      Italy
                         88                         92                        103                        105
                    Jamaica                      South                      China                    England
                        106                        115                        122                        127
                       Cuba                      India                El-Salvador                     Canada
                        138                        151                        155                        182
                Puerto-Rico                    Germany                Philippines                          ?
                        184                        206                        295                        857
                     Mexico              United-States
                        951                      43831
```

REMOVING THE MISSING VALUES

- Since imputing is not an option, better remove the rows of the data having the missing values.
- Final cleaned dataset has 45221 (7.4 % fall)
  - class(Earnings < 50K) 34013 75.21% (0.86% fall)
  - class(Earnings =>50K) 11208 24.79% (0.86% rise)

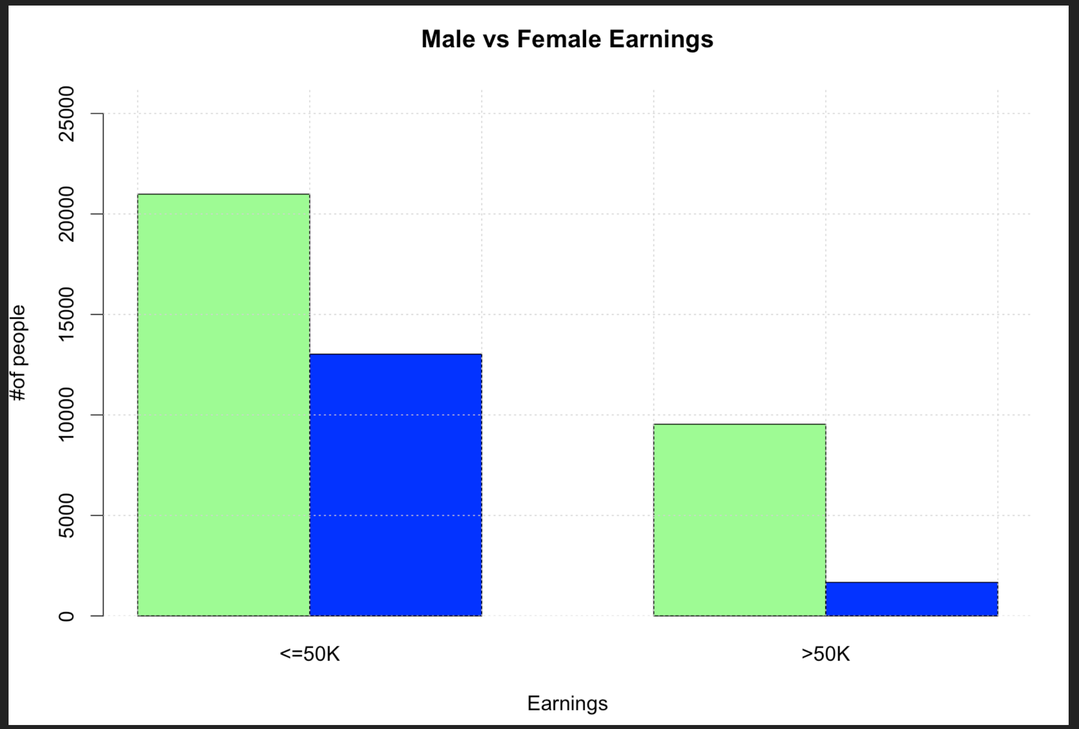
AGE VS EARNINGS

- Age ranges from 17 - 90

- Blue-> Low income
- Red-> High Income

GENDER VS EARNINGS

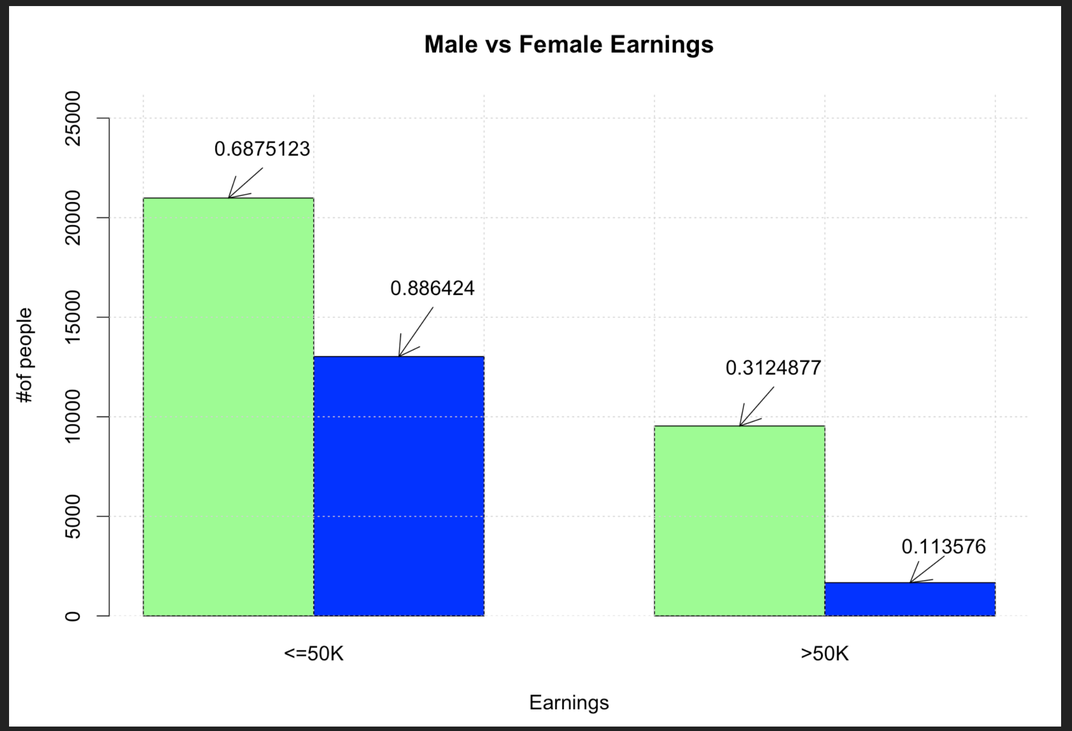
- Low income in both gender are relatively equal
  - 30526 males
  - 14695 females



- Blue-> Female
- Green-> Male

GENDER VS EARNINGS

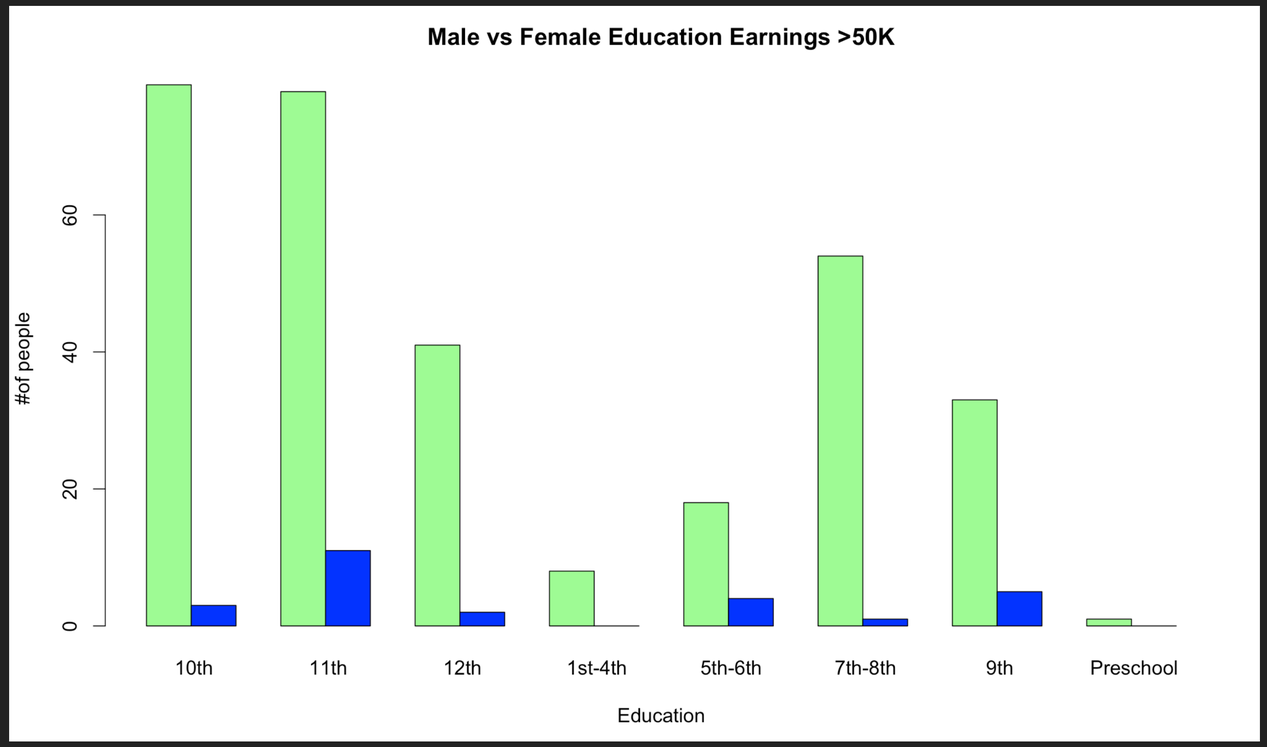
MALE		FEMALE	
30526		14695	
<=50K	>50K	<=50K	>50K
20987	9539	13026	1669



- Blue-> Female
- Green-> Male

EDUCATION VS EARNINGS -GENDER WISE

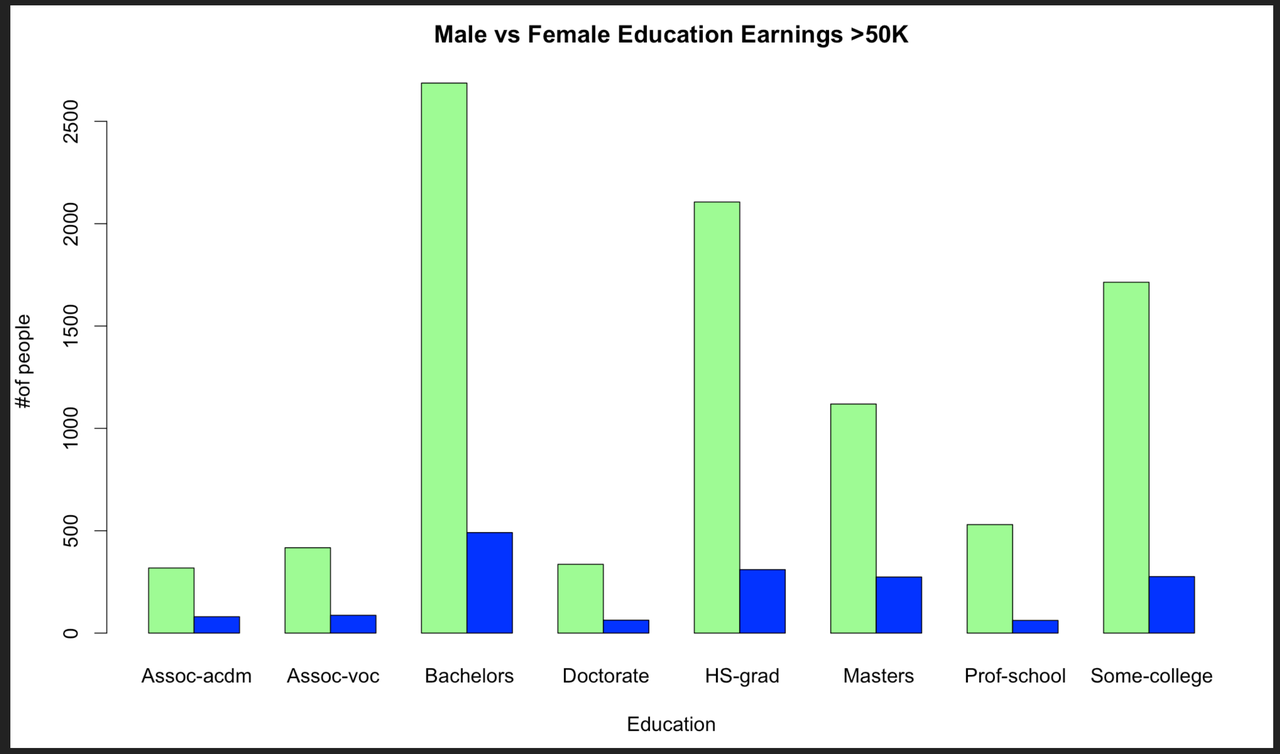
- This is for the persons earning more than 50K per year unto highschool



- Blue-> Female
- Green-> Male

EDUCATION VS EARNINGS -GENDER WISE

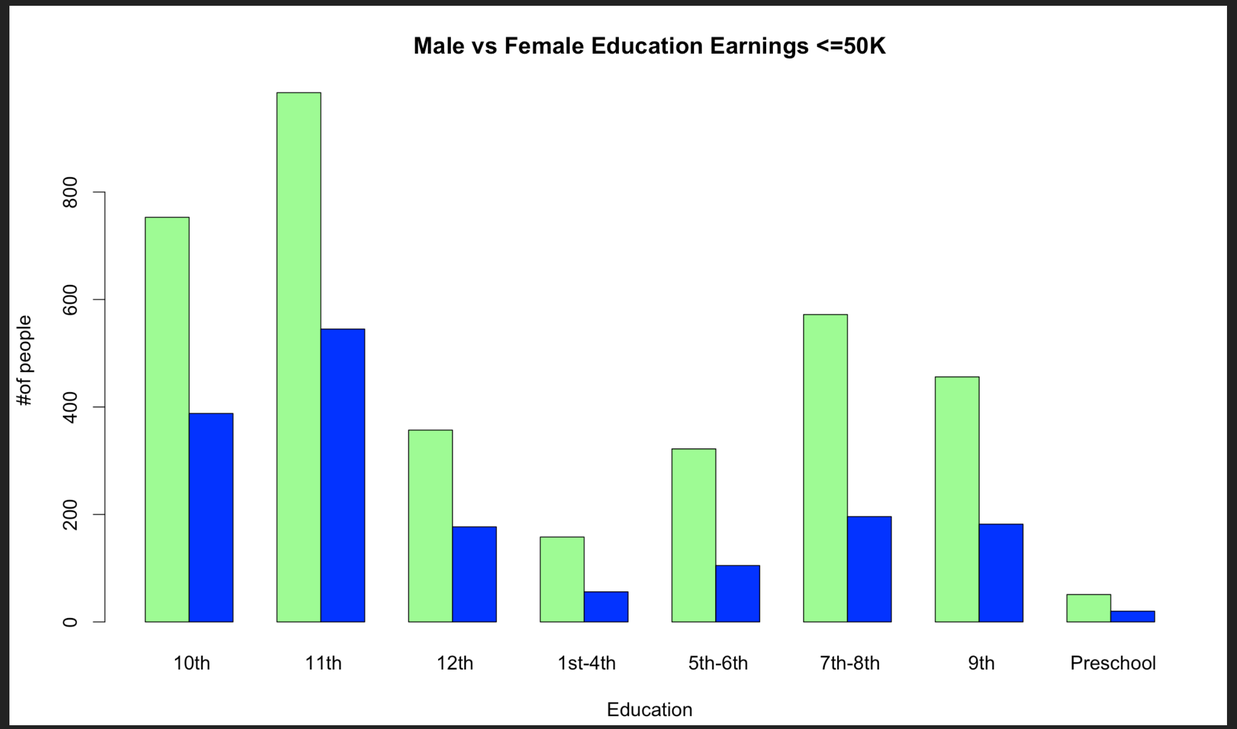
- This is for the persons earning more than 50K per year after highschool



- Blue-> Female
- Green-> Male

EDUCATION VS EARNINGS -GENDER WISE

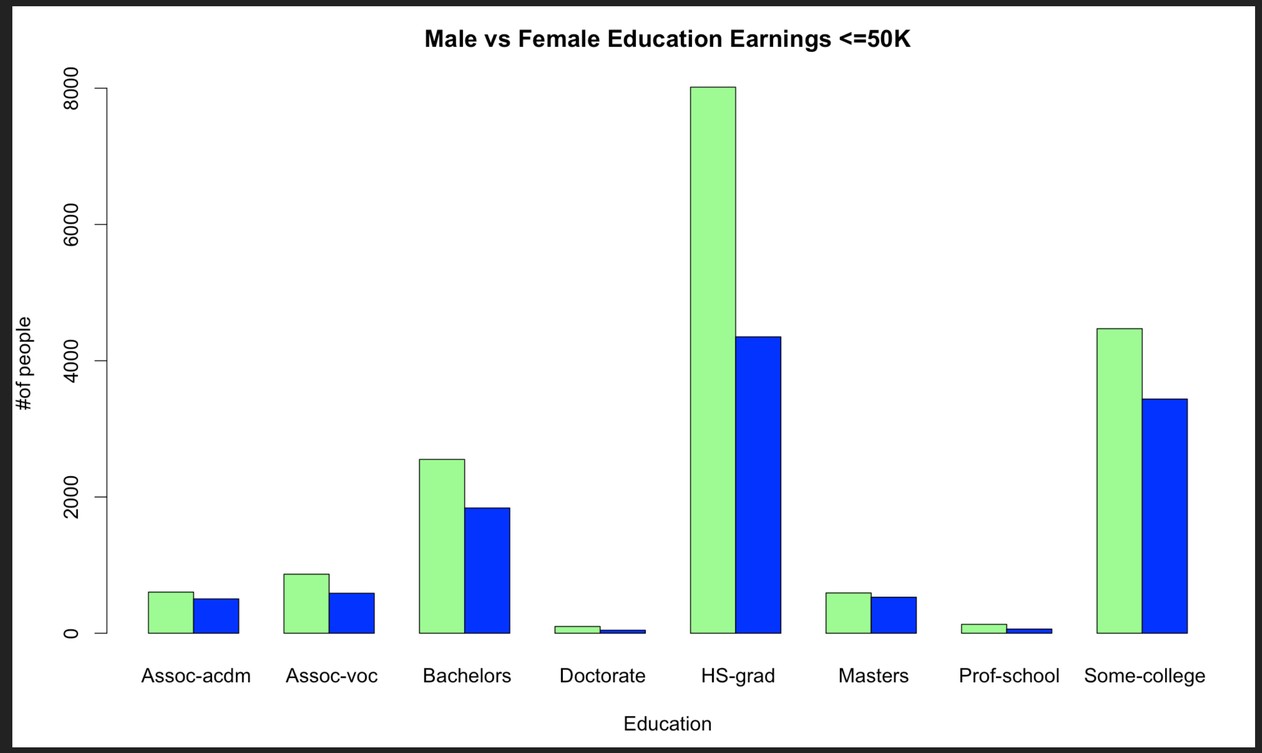
- This is for the persons earning less than 50K per year upto highschool



- Blue-> Female
- Green-> Male

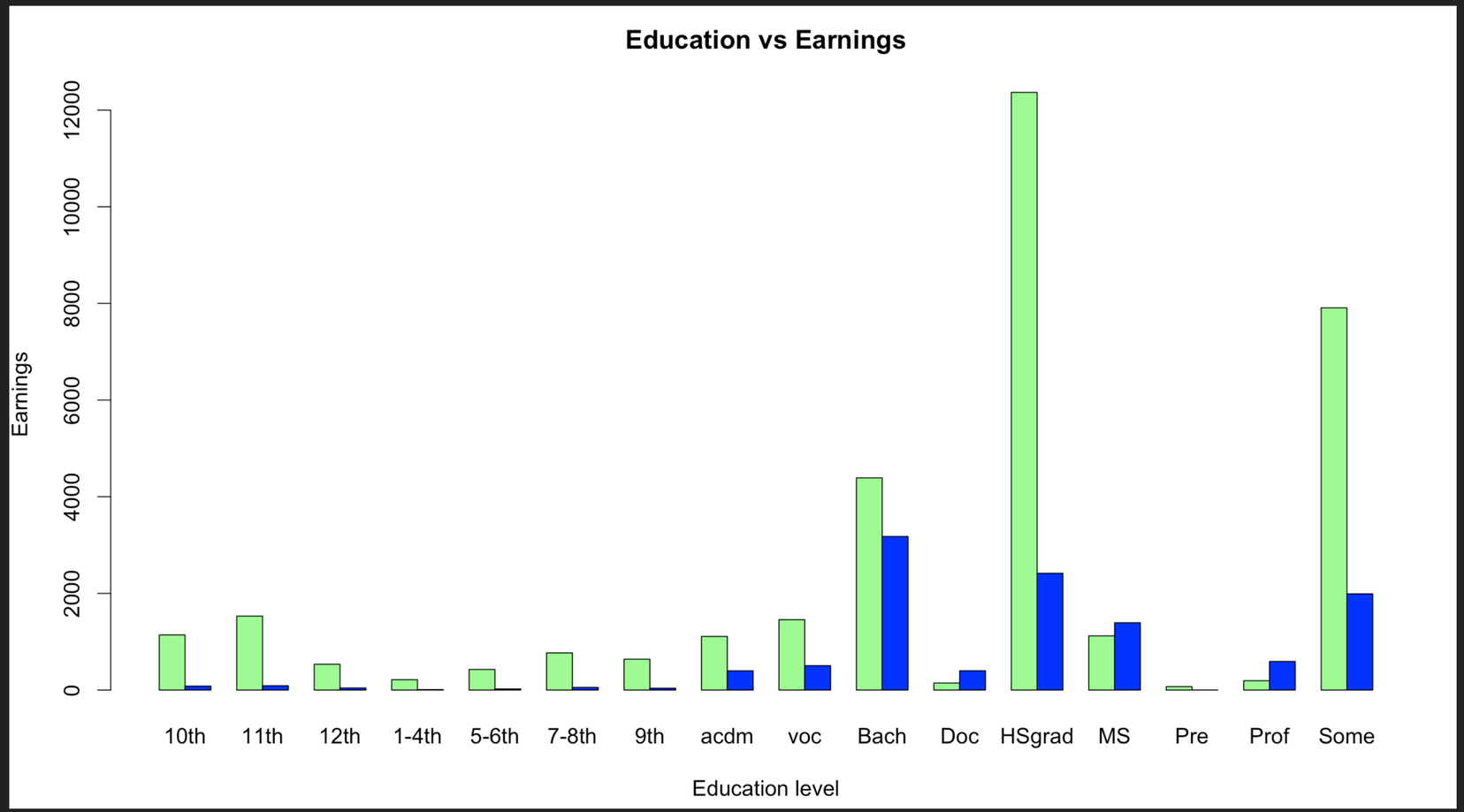
EDUCATION VS EARNINGS -GENDER WISE

- This is for the persons earning less than 50K per year after highschool



- Blue-> Female
- Green-> Male

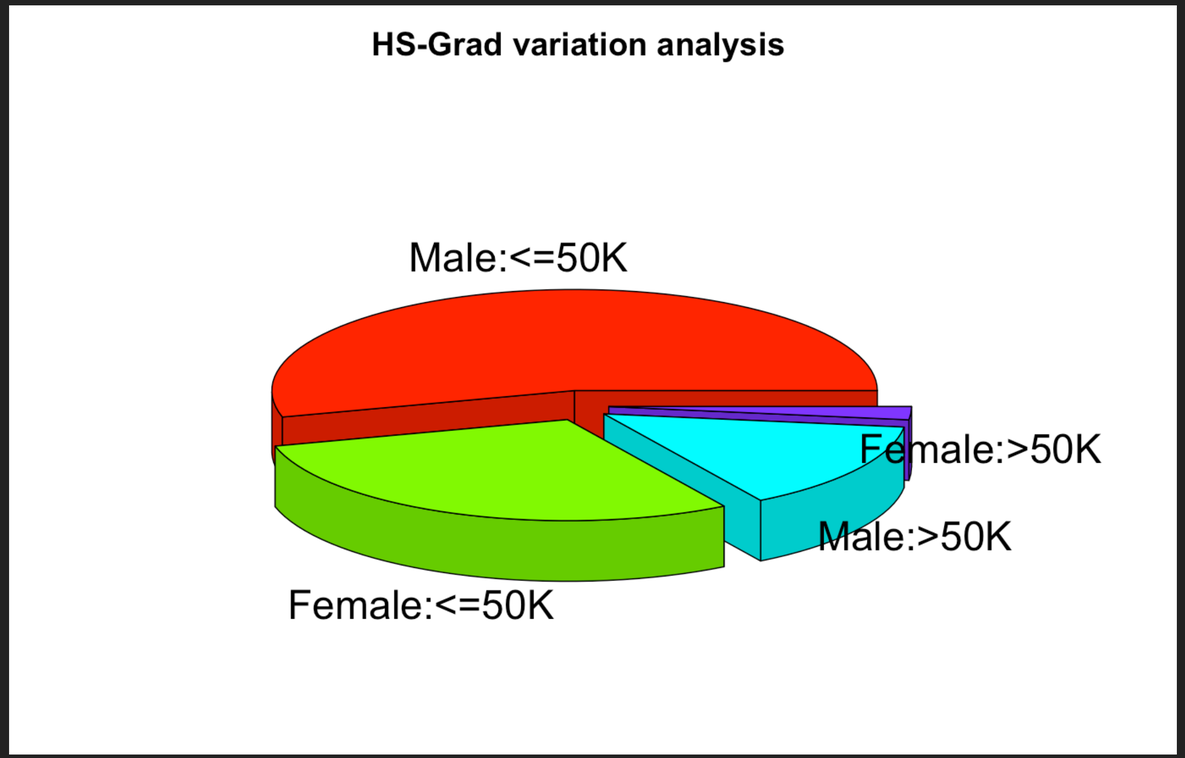
EDUCATION VS EARNINGS



- Blue-> Female
- Green-> Male

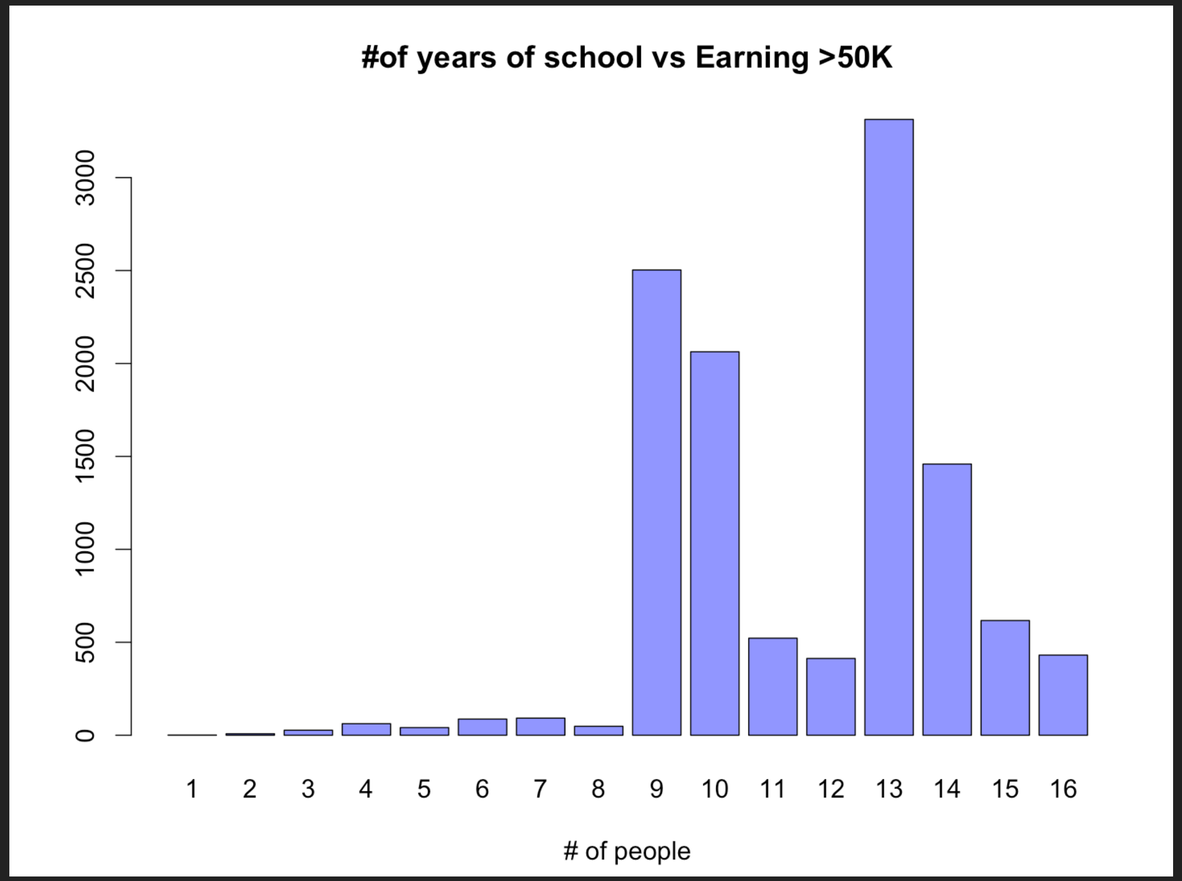
EDUCATION VS EARNINGS

- High school graduate variation analysis



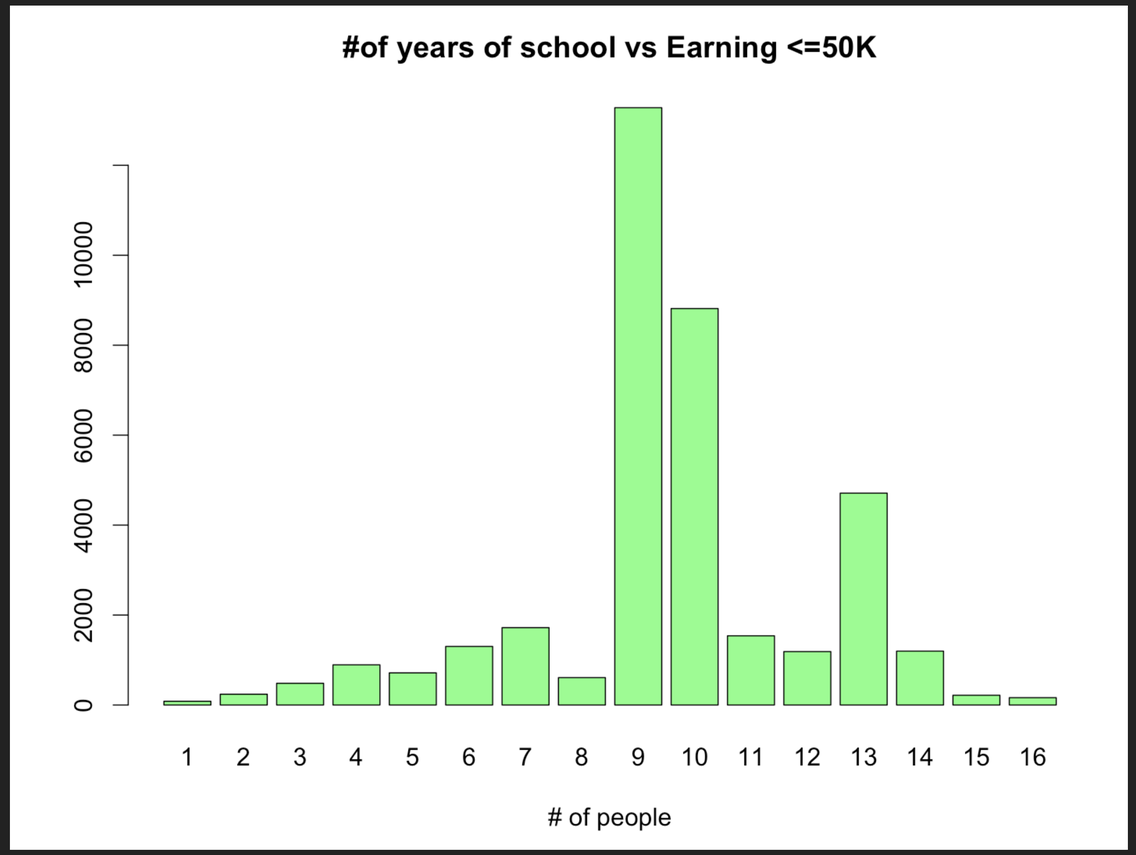
YEARS OF EDUCATION VS EARNINGS

- People earning more than 50K

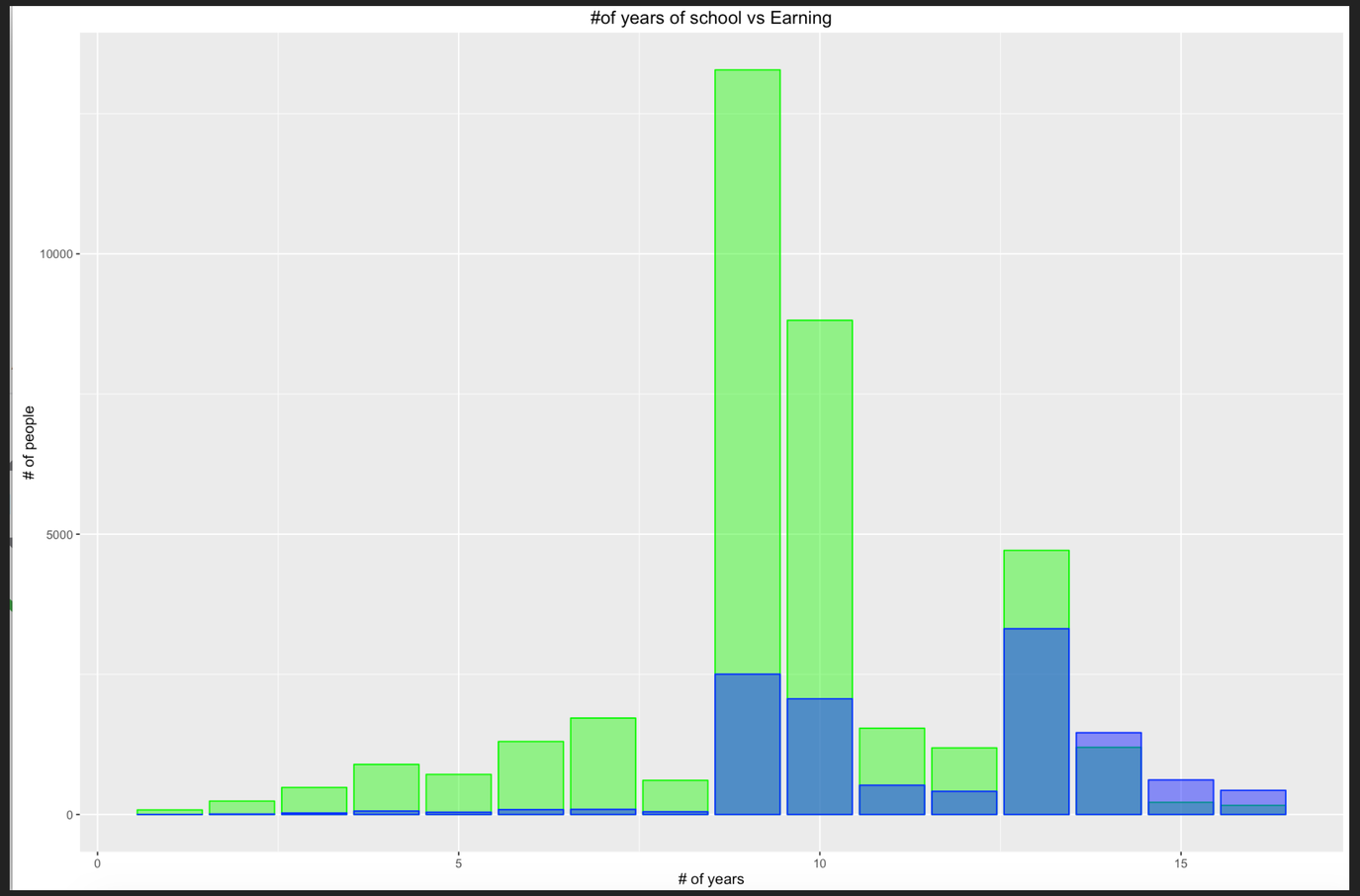


YEARS OF EDUCATION VS EARNINGS

- People earning less than 50K



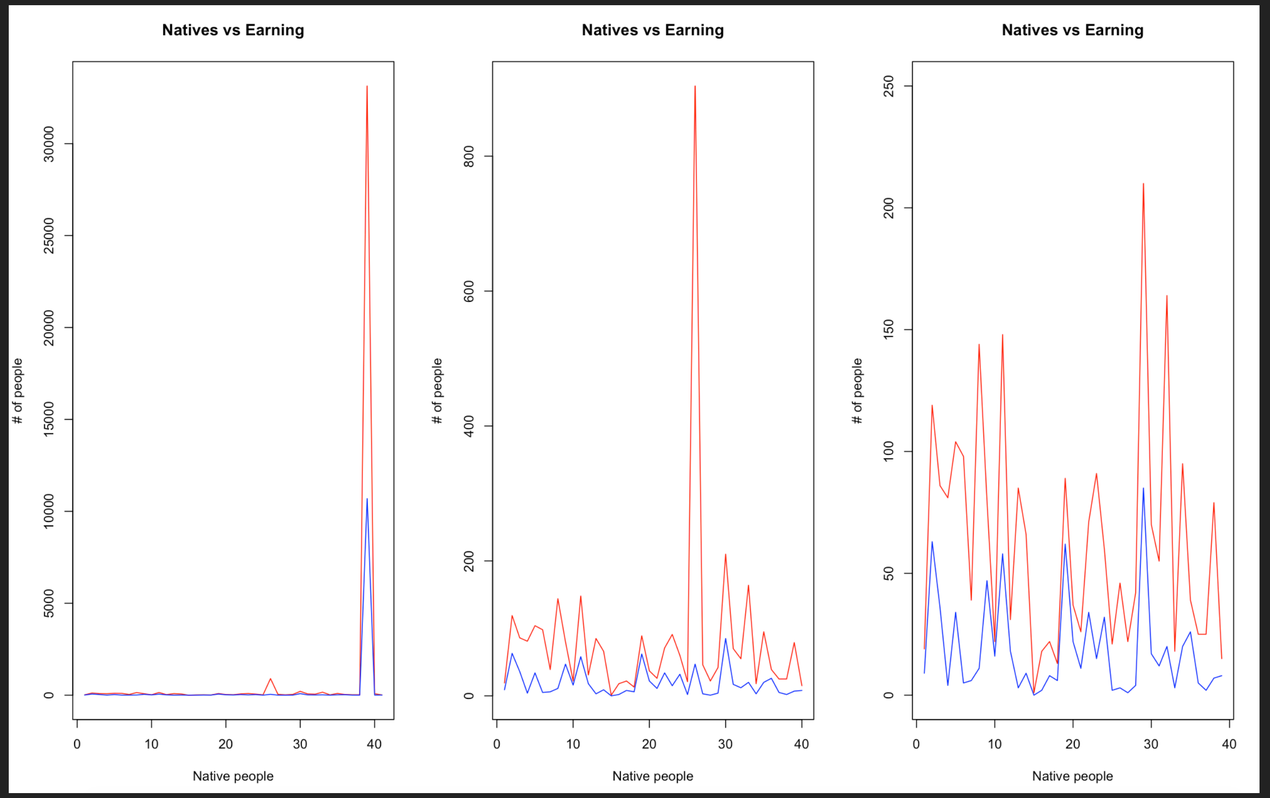
YEARS OF EDUCATION VS EARNINGS



- Blue-> >50K
- Green-> <=50K

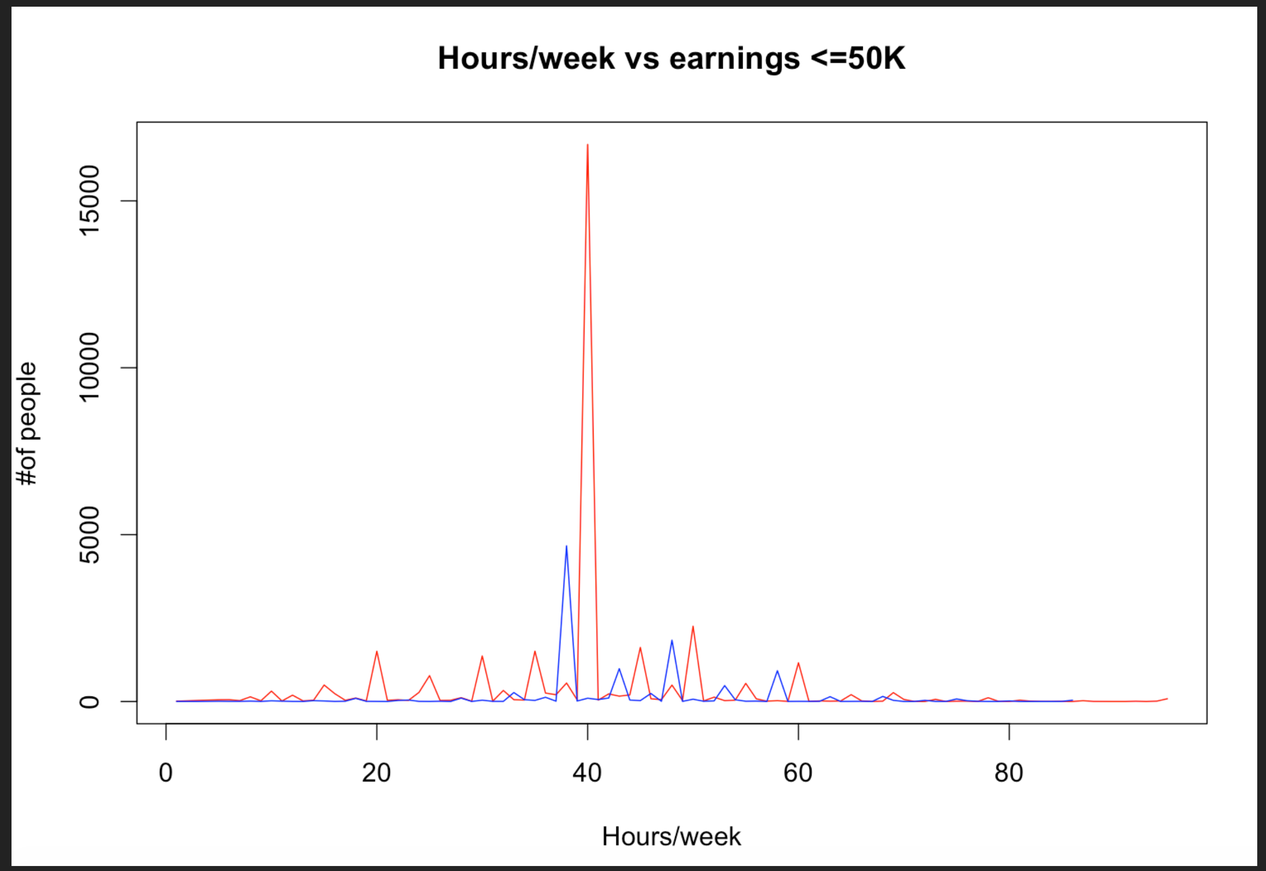
NATIVES VS EARNINGS

- Having an extreme maximum edge
- Removing the extreme and the local maxima



- Blue–> >50K
- Red–> <=50K

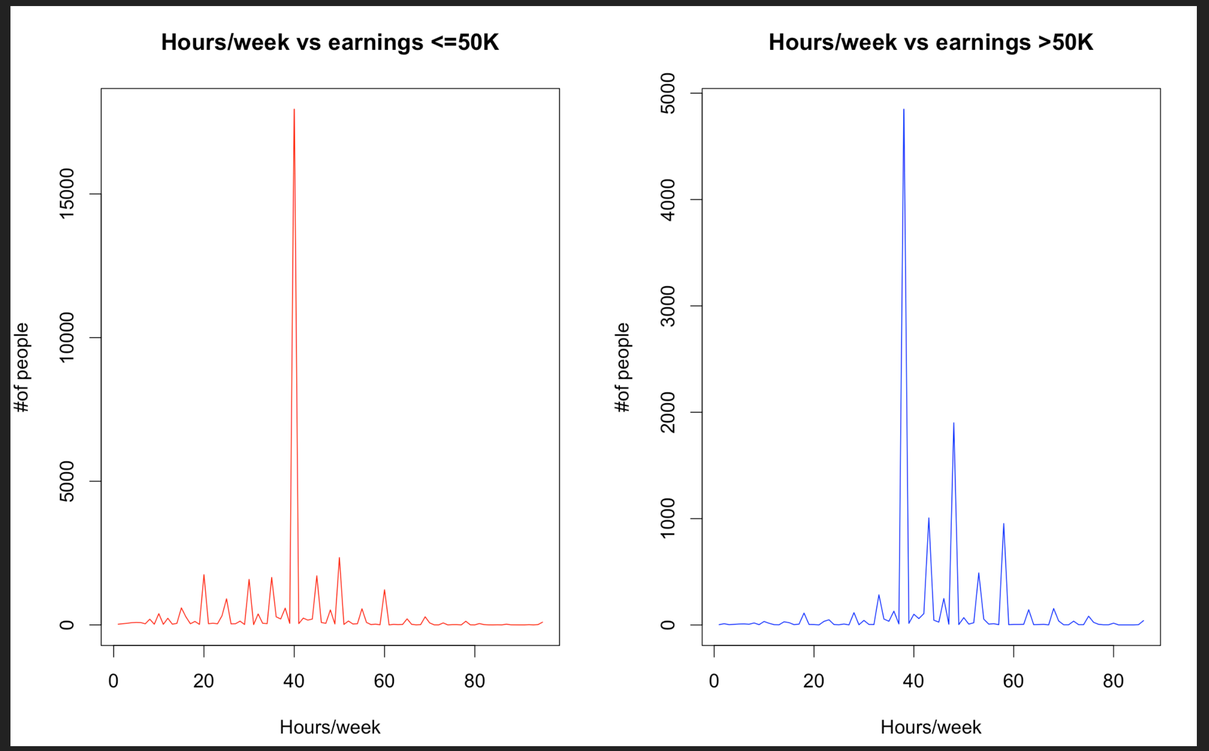
HOURS / WEEK VS EARNINGS



- Blue–> >50K
- Green–> <=50K

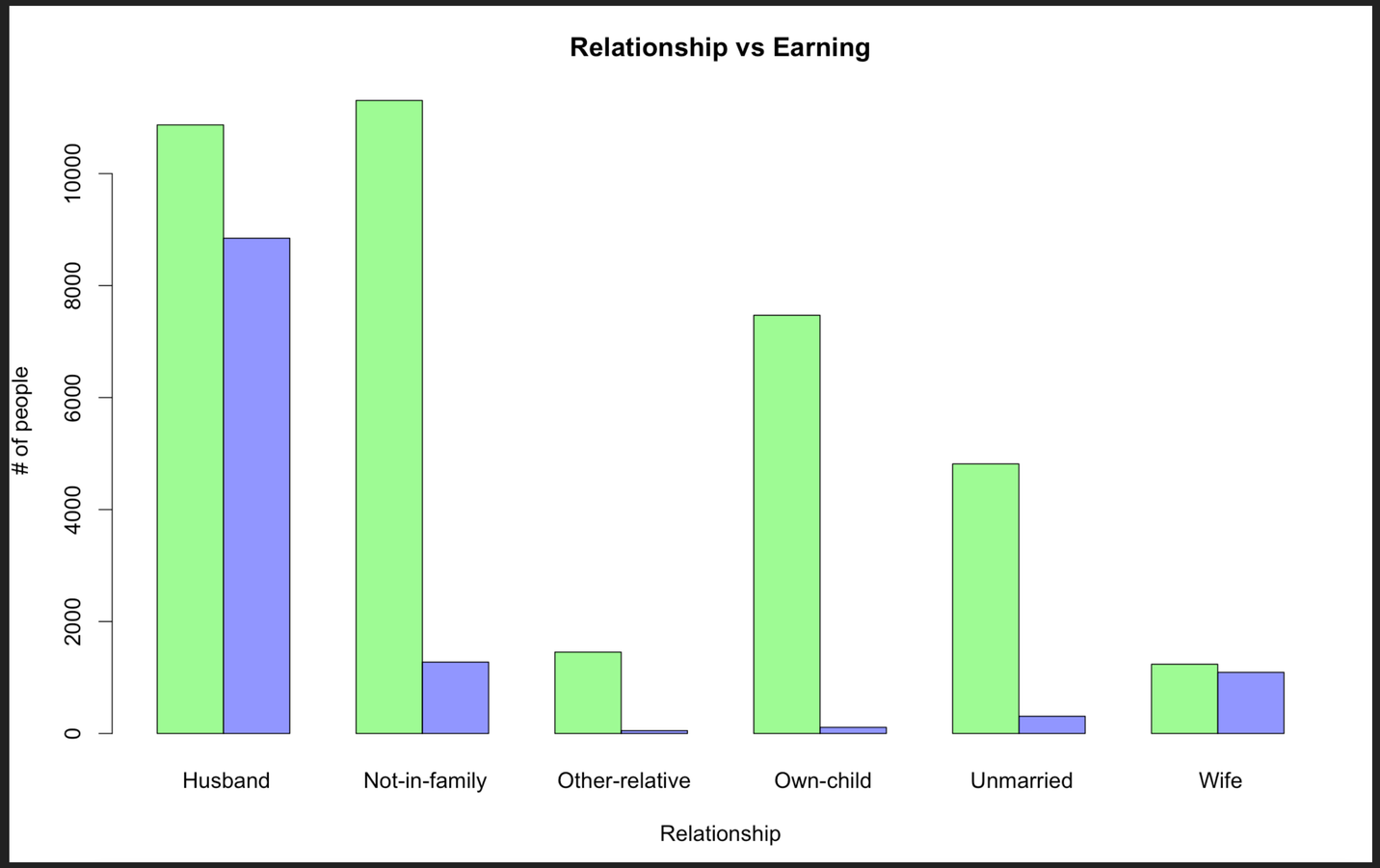
HOURS / WEEK VS EARNINGS

- # of peaks in the data shows majority



- Blue–> >50K
- Green–> <=50K

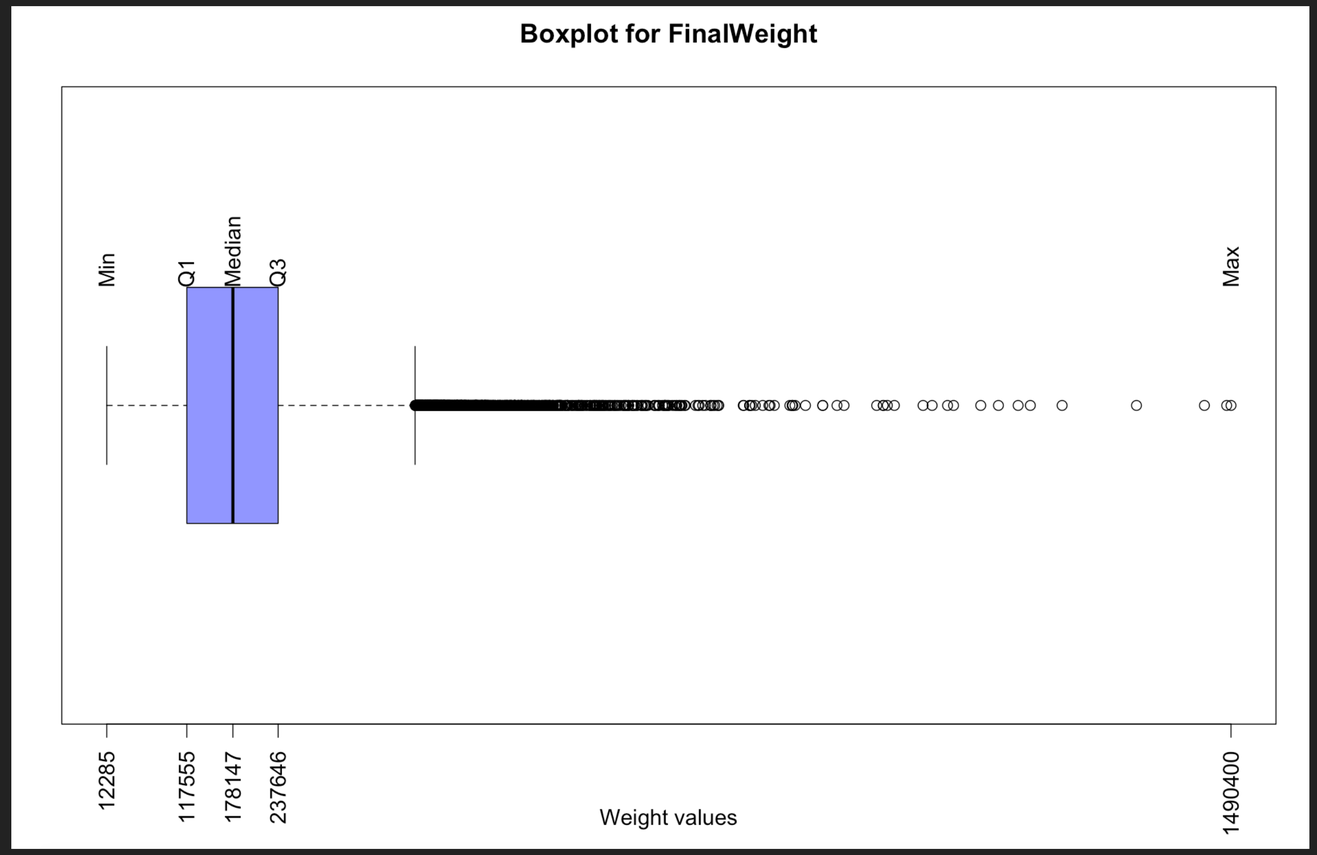
RELATIONSHIPS VS EARNIGNS



- Blue–> Female
- Green–> Male

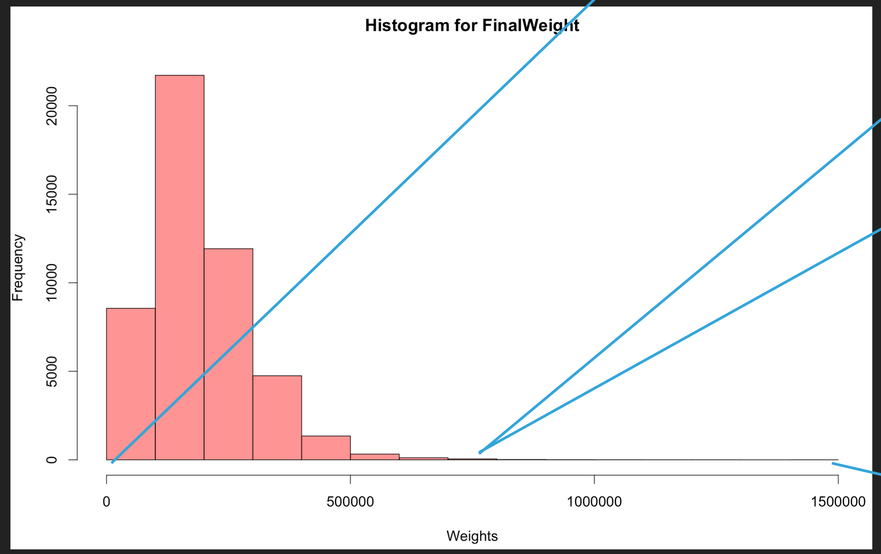
FINALWEIGHTS VS EARNINGS

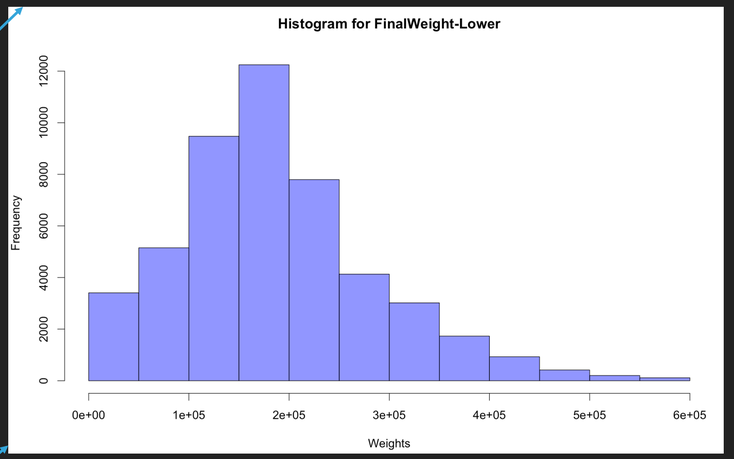
- Important attribute in the dataset

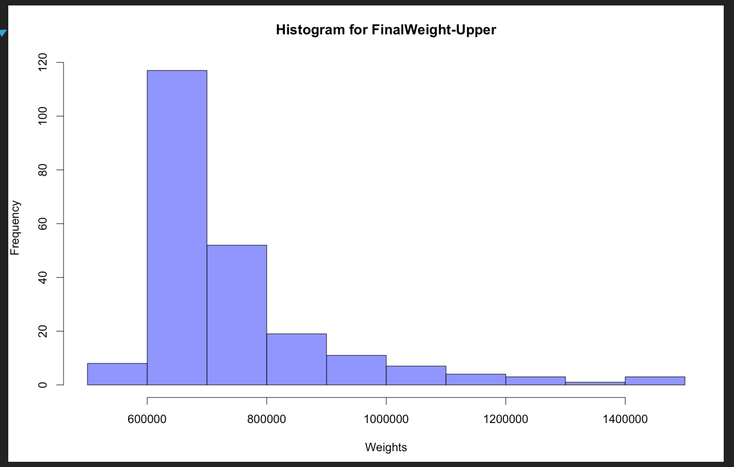


FINALWEIGHTS VS EARNINGS

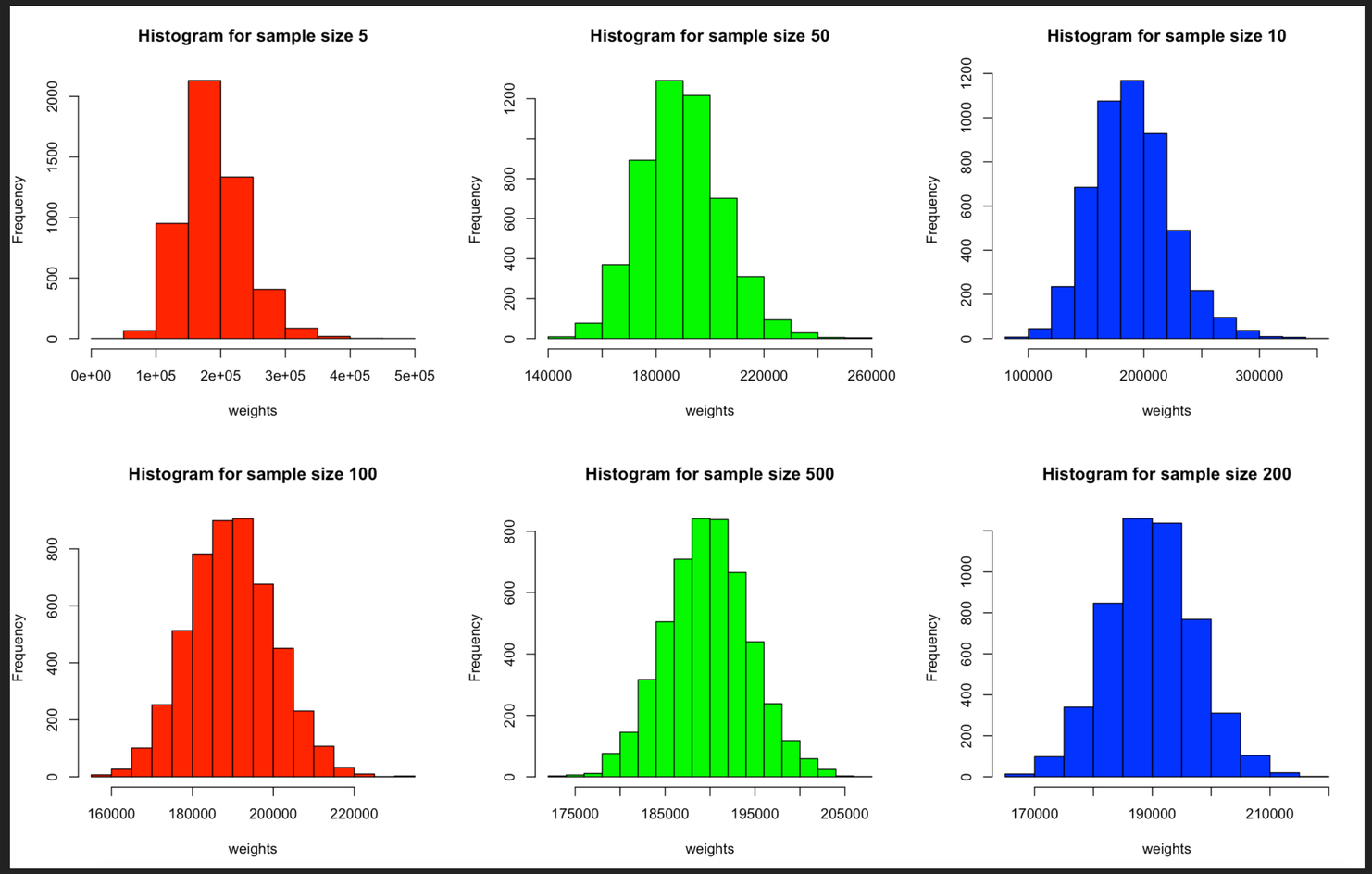
- Distribution of the values in this attribute is wide-spread with extreme boundaries



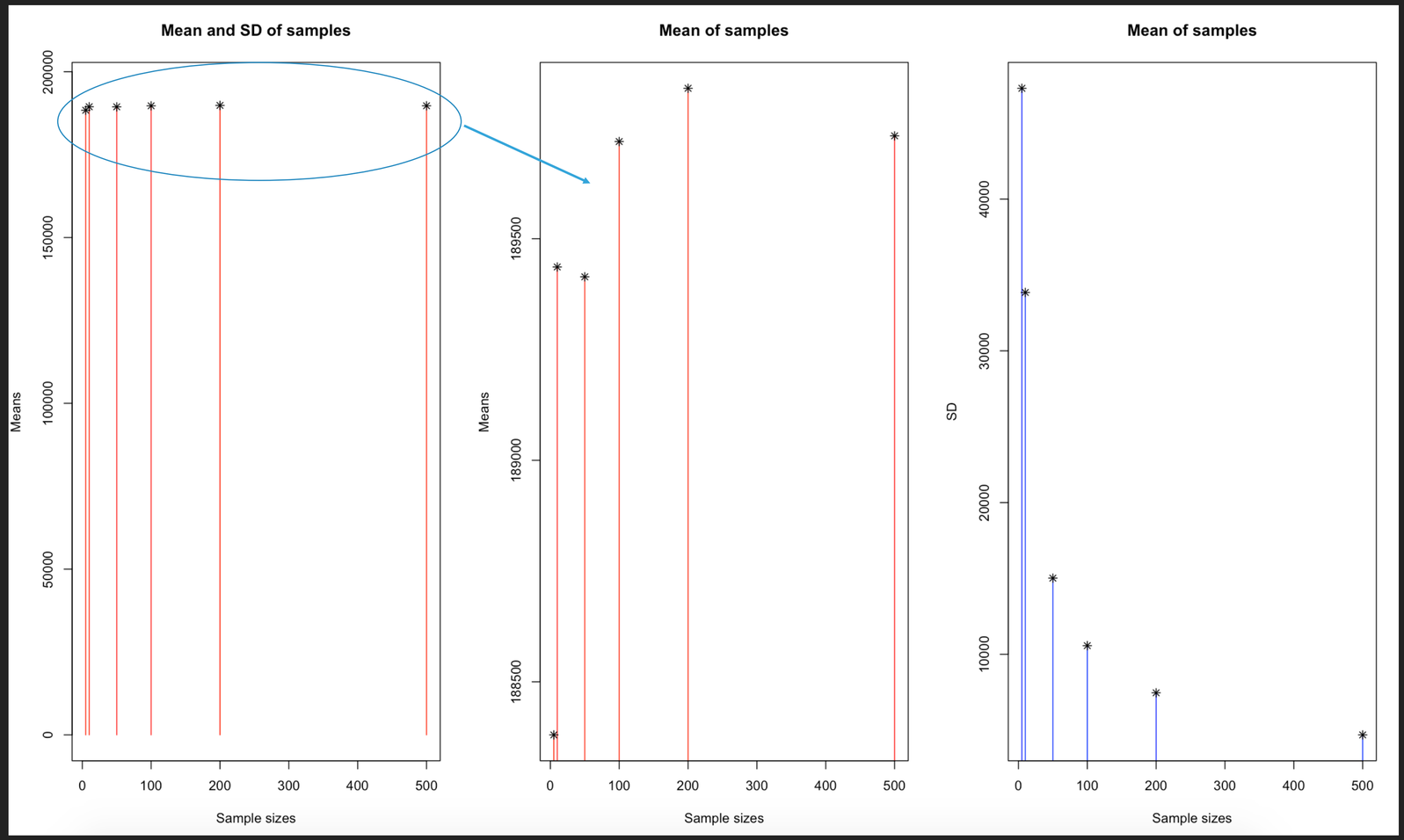




DISTRIBUTION OF MEANS FOR VARIOUS SAMPLES



MEAN AND STANDARD DEVIATION OF THE SAMPLES



SAMPLING RESULTS

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	TYPES	Cambodia	Canada	China	Columbia	England	France	Holand	Honduras	Hong	India	Italy	Japan	Mexico	Philippines	United State	Yugoslavia
2	SRS-WR	3	43	20	17	27	6	0	3	3	28	25	16	223	61	9068	6
3	SRS-WOR	5	37	24	15	25	7	0	5	9	31	23	10	212	69	9167	2
4	SYS-EQUAL	9	29	26	15	19	8	0	4	6	31	11	18	178	62	8249	6
5	SYS-UNEQUA	0	3	2	2	6	0	0	1	1	15	14	12	148	52	9534	4
6	STRATA	6	36	25	18	26	8	1	4	6	33	22	20	199	63	9130	5

CONFIDENCE INTERVALS — 80% AND 90 %

Population Mean falls inside the Confidence intervals

```
Sample.size :  5
80% confidence intervals = 124835.4  -  245737.8
80% confidence intervals = 114983.2  -  235885.6
80% confidence intervals = 82337.81  -  203240.2
```

```
Sample.size :  5
90% confidence intervals = 118086  -  273464.4
90% confidence intervals = 132735.6  -  288114
90% confidence intervals = 131756.8  -  287135.2
```

```
Sample.size :  10
80% confidence intervals = 178474.8  -  263965.6
80% confidence intervals = 114302.8  -  199793.6
80% confidence intervals = 142237.1  -  227727.9
```

```
Sample.size :  10
90% confidence intervals = 160926  -  270795.2
90% confidence intervals = 138060.6  -  247929.8
90% confidence intervals = 115288.2  -  225157.4
```

```
Sample.size :  50
80% confidence intervals = 176338.4  -  214571.1
80% confidence intervals = 180163.4  -  218396.1
80% confidence intervals = 187108.9  -  225341.6
```

```
Sample.size :  50
90% confidence intervals = 145807  -  194942
90% confidence intervals = 168332.4  -  217467.4
90% confidence intervals = 180660.3  -  229795.3
```

```
Sample.size :  100
80% confidence intervals = 187745.1  -  214779.7
80% confidence intervals = 191343.8  -  218378.3
80% confidence intervals = 174422.2  -  201456.8
```

```
Sample.size :  100
90% confidence intervals = 189594.1  -  224337.8
90% confidence intervals = 174309.2  -  209052.8
90% confidence intervals = 163396.6  -  198140.3
```

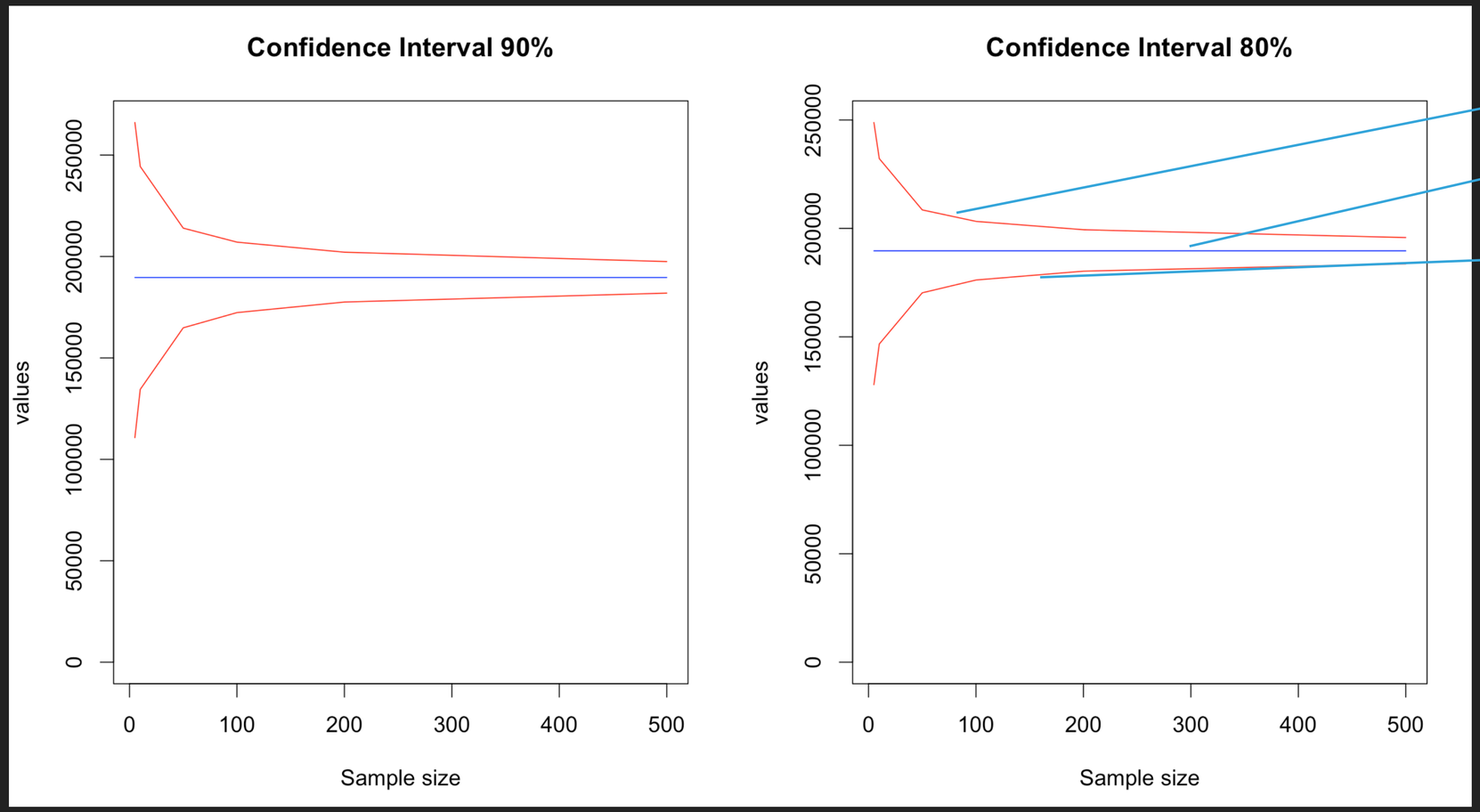
```
Sample.size :  200
80% confidence intervals = 191852.4  -  210968.7
80% confidence intervals = 185861.7  -  204978
80% confidence intervals = 180887.3  -  200003.6
```

```
Sample.size :  200
90% confidence intervals = 166835.6  -  191403.1
90% confidence intervals = 174147.9  -  198715.4
90% confidence intervals = 172323.1  -  196890.6
```

```
Sample.size :  500
80% confidence intervals = 176562.9  -  188653.2
80% confidence intervals = 183336.4  -  195426.6
80% confidence intervals = 179546.8  -  191637.1
```

```
Sample.size :  500
90% confidence intervals = 180658.8  -  196196.7
90% confidence intervals = 186675.8  -  202213.7
90% confidence intervals = 178751.4  -  194289.2
```

GRAPH FOR CI LIMITS



NOTES

- Replacing the missing values in dataset “?” -> hard
- COMPLETE.CASES doesn't work unless na is stated as NA in dataset.
- Sampling in imbalanced dataset
- Systematic unequal probabilities ->categorical data
- Strata in imbalanced datasets