Analytical insight on sampling, bias & relevant matters

Sampling: The data set is a nationally representative sample of over 18,000 individuals living in 5,000 families in the United States. A representative sample is one that accurately represents, reflects, or “is like” the population. So in this case we can say that the sample is minimally biased.

We have used python and pandas library to explore the sample of 4856 records.

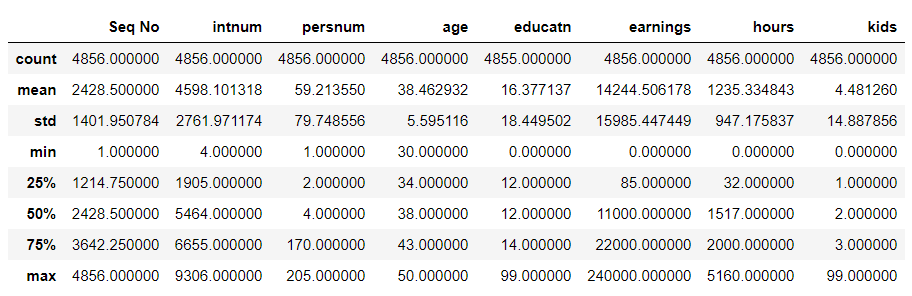


Image 1 : Summary of data set

From the summary of the data, it's clear that data set has many outliers. As an example the max value for number of kids Is 99 which seems impossible. With that outlier value, we got 4.48 as the mean value of number of kids in the population. Same issue is there with education level as 99 is the maximum of education level. As the next step we removed the outliers from the data set.

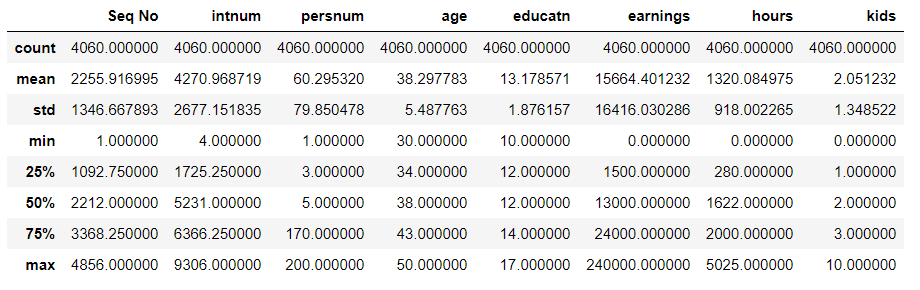


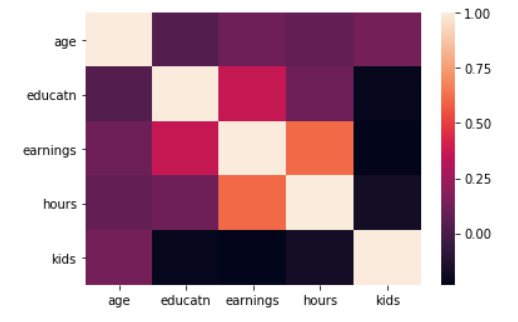
Image 2: Summary of Data set after removing outliers

Analytical insights on Hypothesis, Confident Intervals & Sample statistics

**Hypothesis 1**

* Education has n positive impact on a persons earnings. (Education and Earnings have a correlation)

To identify this possible hypothesis insights were taken from heatmap of columns



Apart from working hours and earnings, the other highest correlation is between education and earnings

If correlation is c (education and earnings)  
  
H0 : c = 0  
Ha : c > 0

from above correlation table, c between education and earnings is 0.364995 (c > 0)  
  
therefore, statistics support for alternative hypothesis with sufficient statistical significance.

but, using other means to confirm it (p value) we selected a significance level (99%) (α = 0.01). Calculations returned following results,

**correlation= 0.37 , p value= 0.0**

P value is less than α, (0 < 0.01), Therefore, null hypothesis is rejected.  
  
**Conclusion**: There is sufficient evidence to conclude there is a significant linear relationship between Education and Earnings.

**Hypothesis 2**

**Married people earn more than others**

Marital status was given as a categorical data. so it is converted to numerical value and added to data set. The numerical value is represented as IsMarried where 1 is married and 0 otherwise.

A heat map diagram is created to identify the possible correlation between marital status vs income.

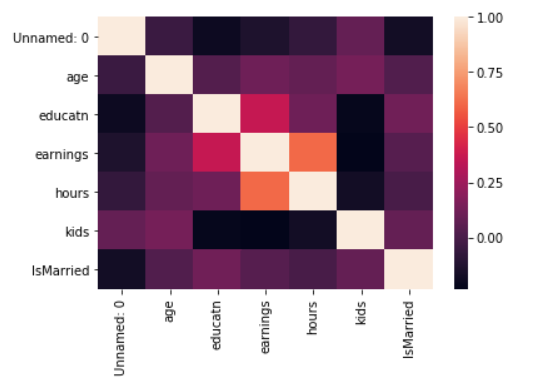


Fig: Heat map of Data with Marital status

We can clearly see from heatmap that there is a small relationship between earnings and marital status.

So Null and Alternative hypothesis has been created as follows.

**M1** : mean of earning of married people

**M2** : mean of earning of unmarried people

**Null Hypothesis H0 : M1 = M2**

**Alternative hypothesis Ha : M1 > M2**

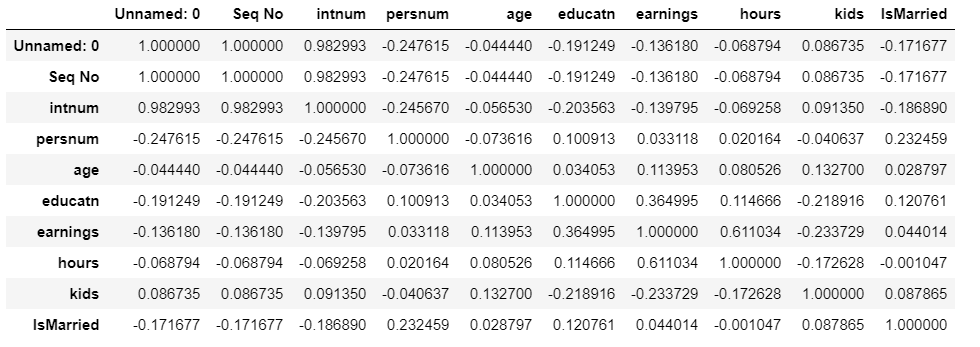
Then detailed descriptions generated for data rows which is IsMarried = 1 and IsMarried = 0. From that we got the Mean values for earnings for IsMarried = 1 and 0.

M1 = 16186.235007

M2 = 14664.219828

So it's clear that Null hypothesis in not valid and Alternative hypothesis is valid. But the difference between those are minimal.

Also the correlation is minimal as seen in Figure below



A scatter plot is created to visualize is.

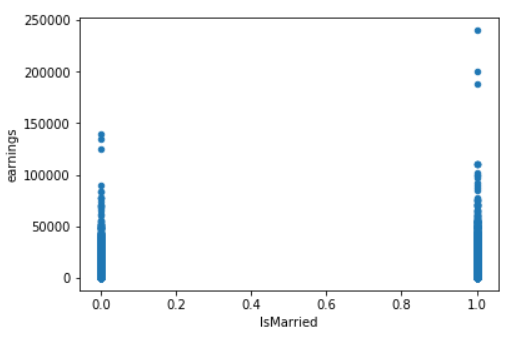


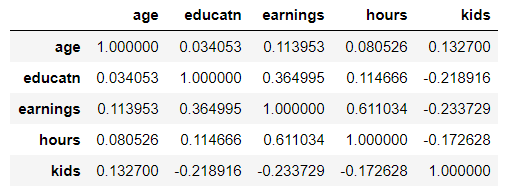
Figure: scatter plot between earnings and IsMarried

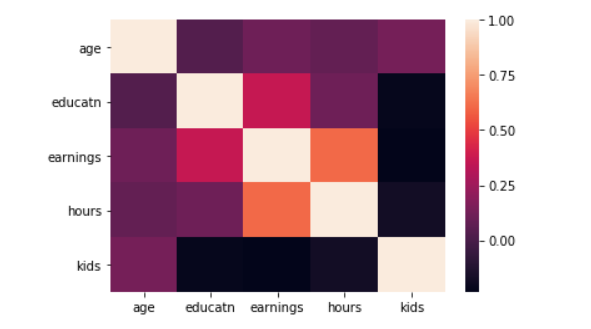
In conclusion we can say that Married people earn more than unmarried people. But the difference is minimal.

**Hypothesis 3**

**People who work longer hours earn more (Correlation between hours and earnings)**

Correlation values between all variables and their heatmap was generated to identify the possible correlated variables.





The above table and heatmap clearly shows that there is a strong correlation between hours worked and earnings, in fact that is the strongest correlation between any two different individual variables in the case.

Let the correlation between hours and earnings be c  
H0 : c = 0  
Ha : c > 0

from above correlation table, c between education and earnings is 0.611034 (c > 0)

therefore, statistics support for alternative hypothesis with sufficient statistical significance.

but, using other means to confirm it (p value) we selected a significance level (99%) (α = 0.01). Calculations returned following results,

**correlation= 0.44 , p value= 0.0**

P value is less than α, (0 < 0.01), Therefore, null hypothesis is rejected.  
  
**Conclusion**: This is evidential to conclude that there is a strong linear relationship between the hours worked worked and earnings. In other words, we would say “People who work longer hours usually earn more”