Executive Summary 2

Analytical insight on sampling, bias & relevant matters 4

Assumptions 4

Hypothesis Testing 4

Other Useful Analytics 5

References 6

Executive Summary

In this assignment, we were supposed to analyse data set that was available publically. The data set is related to families who live in United States.

During the hypothesis testing, we took average earnings and the value given for education as parameters.

Anyway we have observer following key factors during the analysis:

- hours and earnings have a strong positive correlation inferring higher hours spent allow higher earnings
- education and earnings have a significant positive correlation inferring higher education levels allow higher earnings
- education and kids have significant negative correlation inferring residents with relatively higher education levels tend to have less children
- earnings and kids have a significant negative correlation inferring families with relatively high income tend to raise less children

Group Members:
- **199329V, H.E.A.K.P. Hewawasam**
- **199355V, W.L.D.W.P. Perera**
- **199342E, C.G. Madage**
- **199361K, M.D.R.N. Senanayake**

Analytical insight on sampling, bias & relevant matters

According to the Panel Study of Income Dynamics (PSID), web site PSID is the longest running longitudinal household survey in the world and this study began in 1968 with a nationally representative sample of over 18,000 individuals living in 5,000 families in the United States.

Since this is worldwide research their target population should be all the individuals in the world.
But they only collect data from families who live in the United States.
Therefore at the very beginning, we can conclude that this data set is geographically bias.

In the assignment, we get a data set of 4857 individuals. But according to their web site, they have collected more than 18000 individuals. Therefore data set given in the assignment was already sampled.

Assumptions

- 98 and 99 values in education level considered as outliers. So we remove them from our calculations.

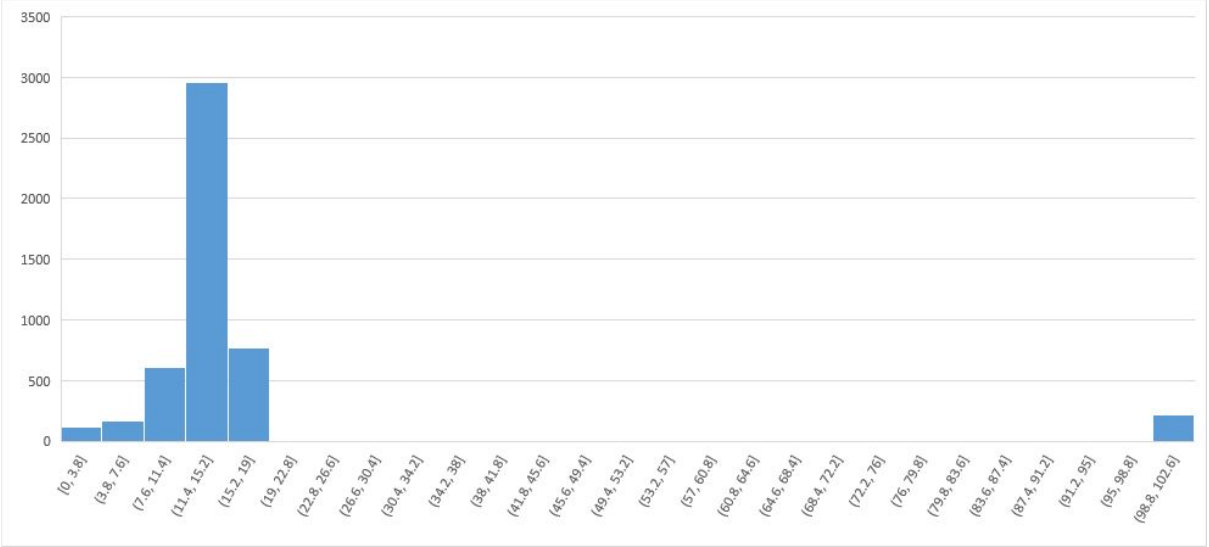
Hypothesis Testing

1. Is the average earnings per family is greater than 14000 if the education is greater than 12.
    a. Parameters interested
        i. Education
        ii. Earnings
    b. m: Education when average earnings per family is greater than 14,000
    c. H0: m = 14000
    d. Ha: m < 14000
2. If there are kids(greater than 0), the divorce rate is low
    a. Parameters interested
        i. Number of kids greater than 0
        ii. Marital status
    b. M: if there are kids most of them are married.

Other Useful Analytics

There were multiple entries in the column of educatn (education) with the value of 98 and 99 seemed outliers as illustrated in below diagram;

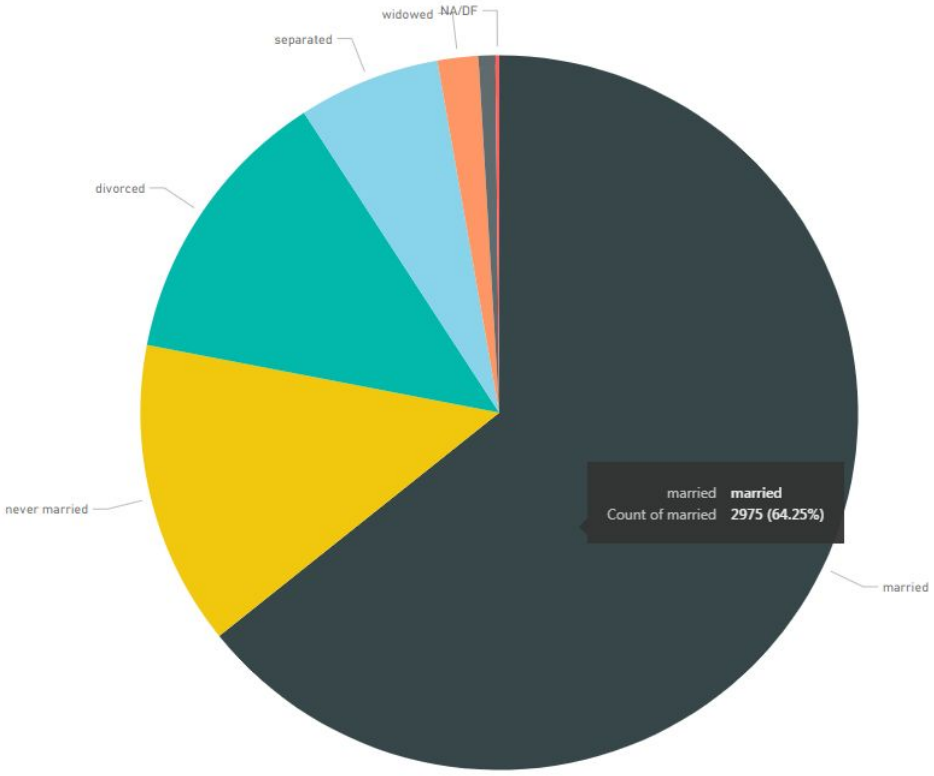
X axis: educatn
Y Axis: count



By analysing the data set we can see that most of the earnings from the people around age 35-45:

Further analysing, it was discovered that the count of age is also going with above graph. Hence, we could assume that the distribution is legitimate.

Our dataset have been organized as follows (based on their marital status):



References

[1] https://psidonline.isr.umich.edu/Guide/2015DataHighlights.pdf
[2] https://psidonline.isr.umich.edu/
[3] https://psidonline.isr.umich.edu/data/Documentation/UserGuide2013.pdf