Detailed Report

# Introduction

As a team we analyzed the given data set regarding the Panel Survey of Income Dynamics. The data set consist of 4856 tuples with age, education, earnings, working hours, number of kids and marital status. The data set has some outliers like having 98 kids and education level is 99. Those are unrealistic data. So in our sample, we removed them and analyzed the remaining data. First we analyzed the distribution of each important parameters by visualizing them in charts. Then we analyzed the relationship among the parameters and make a hypothesis. We analyzed further to prove our hypothesis and concluded with the results.

Group Members

1.Nuwani Senanayake 199362N

2.Upekka Pathirana 199306A

3.Miyuru Sonnadara 199367J

4.Nilusija Nadarasamoorthy 199352J

# 2. Initial analysis on data set

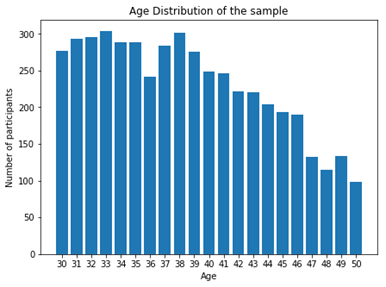
This is the initial analysis on the data set without any pre processing.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Min | Max | Avg | SD | Median |
| Age | 30 | 50 | 38.46 | 5.59 | 38 |
| Education | 0 | 99 | 16.38 | 18.45 | 12 |
| Earnings | 0 | 240000 | 14244.5 | 15983.8 | 11000 |
| Hours | 0 | 5160 | 1235.33 | 947.08 | 1517 |
| Kids | 0 | 99 | 4.48 | 14.89 | 2 |
| Hourly Wage | 0 | 364.24 | 9.81 | 16.41 | 7.3 |

**Note:** Hourly wage is calculated as (Earnings / No of hours worked).

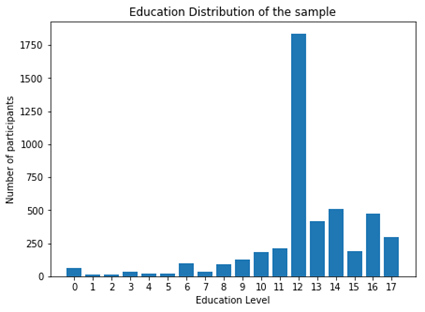
### 2.1 Age

The given data has the age group 30 to 50. The data set is biased. The figure below is the age distribution of the sample.



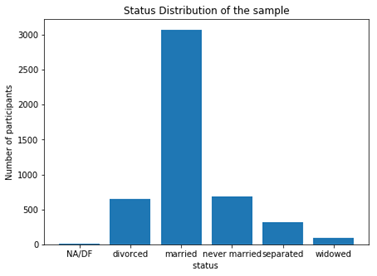
### 2.2 Education

The data set has some uncertain data like education level 98. The figure below is the education distribution of the sample. The uncertain data has been omitted for the graph. In this sample, 1804 people has 12 as their education level. 81.20% of people has 12 or above 12 as their education level.

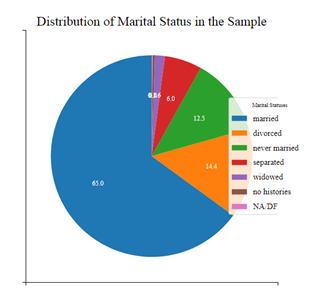


### 2.3 Marital Status

The figure below is the marital status distribution of the sample. 65% of this sample participants are married. And divorced and never married participants of this sample is nearly equal.

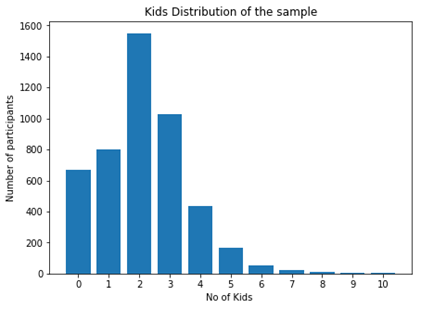


This pie chart show the marital status without preprocessing the data. In the data set, marital status shown as “no histories” has the number of child as 90+. So in the preprocessing step, we have removed those data. But those data may refer the orphanage childrens count. Those are like outliers. We need more information to analyze those data and better to analyze those data separately.



### 2.4 Number of Kids

The figures below show the marital kids distribution of the sample. The data set has some uncertain data like having 98 children. Those values were omitted for the graph. Most of this people in this sample have two children. The people are willing to have less number of children. May be because of the financial problem and the parents want to give the better education and facilities to their children.



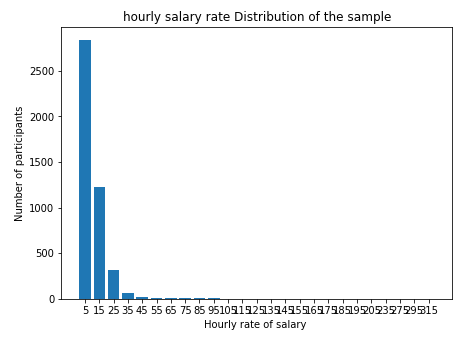
# 3. Secondary Analysis

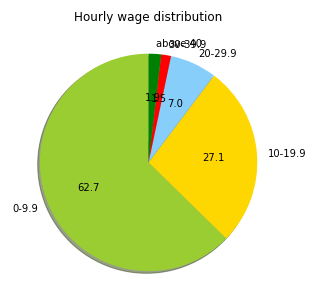
This is the analysis after removing unrealistic data.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Min | Max | Avg | SD | Median |
| Age | 30 | 50 | 38.46 | 5.59 | 38 |
| Education | 0 | 17 | 12.36 | 3.05 | 12 |
| Earnings | 0 | 240000 | 14244.5 | 15983.8 | 11000 |
| Hours | 0 | 5160 | 1235.33 | 947.08 | 1517 |
| Kids | 0 | 10 | 2.14 | 1.43 | 2 |
| Hourly Wage | 0 | 364.24 | 9.81 | 16.41 | 7.3 |

**Note:** Hourly wage is calculated as (Earnings / No of hours worked).

Hourly wage distribution of the sample



When we analyzed the hourly wage of the people in the sample, the hourly rate is too low. 62.7% of the people in the sample getting their hourly salary in the range of 0 to 9.99. Most of the people in this sample are suffering from poverty. The pie chart drawn for ['0-9.9', '10-19.9', '20-29.9','30-39.9','above 40' ] this interval is given below. Green indicate for above 40 + hourly rate wage receivers and red indicate the 30-39.9 range as the hourly wage receivers. 

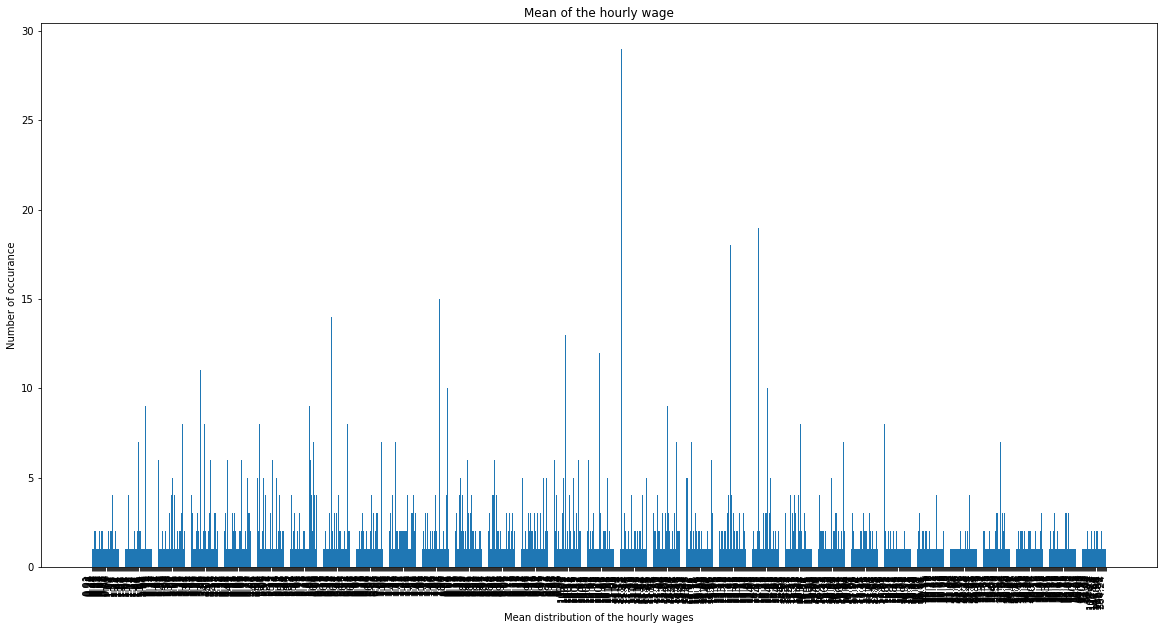
# 4. Hypothesis Test

H0 : The average hourly wage = 13.5

Ha : The average hourly wage < 13.5

First we calculated the hourly wage from the division of earnings by hours. For the people whose working hours were zero the hourly wage was also considered zero. The hourly rate distribution graph shows a right skewed distribution. The hourly rate is very low.

In order to calculate the true mean of the hourly wage of the population, the following procedure was followed. Take random 500 samples from the hourly wage and calculate their average. Repeat this step 2000 times. This makes the sample count to 2000. The random sample count for each iteration and the iteration count were tuned to obtain the best fitted normal distribution. The instances where the hourly wage equals to zero were removed to approximate properly to the normal distribution. The following graph shows the distribution.

Mean = 13.27

Standard deviation = 19.08

Sample count = 2000

With 20% confidence we can say that the true mean lies between [8.5, 18.04]

**Link to the Python Code**

**.**\source-code\MeanDistribution.ipynb

# 5. Executive Summary

From our analysis we can conclude that the education level and the working hours affect the earnings of a person. But working hard or the higher education level does not give a higher income.

* The people in the age range 35 – 40 tend to get the highest earnings.
* The people with education level 12 and above shows higher earning level compared to the people with lower education level.

Here we can conclude that the people are suffering from poverty. If there is any kind of help for self-employment schemes, better to include these people. It may help them to come out of the financial issues and have a better lifestyle. Encourage these people by doing some awareness programs and help them to get better job opportunities. These people have a better education level. Guiding them will help to overcome the poverty and increase their income as well.

Based on the analysis we also can state that the average hourly income of a person is around 13.5 dollars, which is very low.

# 6. Assumptions

* The people who are less than 30 years and greater than 50 years are not earning
* Kids are not earning and they don’t have educations as well.
* The ones having 98 or 99 kids may be orphanage conductor. So we consider them as outliers and not take into our calculation.

# 7. References

* https://psidonline.isr.umich.edu/
* <https://psidonline.isr.umich.edu/data/Documentation/UserGuide2015.pdf>
* <https://onlinewritingtraining.com.au/2014/04/bullet-points-or-prose-in-executive-summaries/>