Data Science Assignment 01

Full Report

Group 09 (DS_2019_007)

MSc in Computer Science 2019
Department of Computer Science and Engineering
University of Moratuwa

Executive Summary

After analysing the PSID data which is based on families and individuals in United States, we have found out below details.

Specially during the hypothesis testing, where we considered earnings and education level in which we found out, when education rate is becoming higher, the earnings are also increased.

Also we have observed 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

We used T-test as the method of hypothesis testing.

Summary

Group Members and Responsibilities

199329V, H.E.A.K.P. Hewawasam
Developing Analytical Insights, Other Analytics, Executive Summary, Detailed Report

199355V, W.L.D.W.P. Perera
Developing Analytical Insights, Visualizations, Hypothesis Testing, Detailed Report

199342E, C.G. Madage
Developing Analytical Insights, Hypothesis Testing, Executive Summary, Detailed Report

199361K, M.D.R.N. Senanayake
Developing Analytical Insights, Sampling, bias & relevant matters, Detailed Report

Github Repository

https://github.com/crishalm/dataScience-assignment1

Hypothesis Testing

People with high education level have high earnings than people with low education levels.

Ph : Average earnings of people with education level higher than education median
Pl : Average earnings of people with education level lower than education median
H0 : Ph = Pl
Ha : Ph > Pl

Assumptions

- Entries above education level 20 have minimum effect on overall statistics.
- Entries above number of kids 20 have minimum effect on overall statistics.
- Entries above earnings of USD 150,000 does not affect the overall statistics.
- Even the “persnum” is same, we have considered them as separate individuals

Detailed Analytical Report

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.

The provided data set only included 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.

When we analyse the data, the main area we considered is how earnings vary against education where we came up with the hypothesis "People with high education level have high earnings than people with low education levels".

In-order to evaluate the alternative hypothesis, we have provided relevant statistical details and graphs and finally we have figured out that when education level is higher, the earnings also increased.

Sampling

Since in the PSID itself mentioned that they have considered nearly 18,000 individuals for the research but the data provided only containing nearly 4900 related data, we can come to a conclusion that the data is already sampled
But for analytical purposes, we have sampled the given data set with 800 individual records in which the case we can generate more precise outcome.

Bias

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 biased.

Explanations

Workspace setup and importing data

We created jupyter notebook with Python 3 to do this assignment. Following packages were imported to the notebook.

- Pandas
- Numpy
- Matplotlib.pyplot
- math
- Scipy.stats

Basic Statistics

Using describe() that comes with scipy.stats, we were able to generate following values for each attribute in the dataset.

- count
- mean
- standard deviation
- min
- max
- 25%, 50% and 75%

	Seq No	intnum	persnum	age	educatn	earnings	hours	kids
count	4856.000000	4856.000000	4856.000000	4856.000000	4855.000000	4856.000000	4856.000000	4856.000000
mean	2428.500000	4598.101318	59.213550	38.462932	16.377137	14244.506178	1235.334843	4.481260
std	1401.950784	2761.971174	79.748556	5.595116	18.449502	15985.447449	947.175837	14.887856
min	1.000000	4.000000	1.000000	30.000000	0.000000	0.000000	0.000000	0.000000
25%	1214.750000	1905.000000	2.000000	34.000000	12.000000	85.000000	32.000000	1.000000
50%	2428.500000	5464.000000	4.000000	38.000000	12.000000	11000.000000	1517.000000	2.000000
75%	3642.250000	6655.000000	170.000000	43.000000	14.000000	22000.000000	2000.000000	3.000000
max	4856.000000	9306.000000	205.000000	50.000000	99.000000	240000.000000	5160.000000	99.000000

Understanding Data

Then we generated some visualizations to get an high level idea of the given data set. All these visualizations will be discussed in next section.

We wanted to remove outliers to get accurate analytical results. First, we removed the entries with 20 or more number of kids (outliers found near 100 from original sample for kids possibly due to incorrectly measured data). Secondly, we removed entries above earnings of USD 150,000 and finally we removed entries above education level 20. Our assumption for these tasks have already given in summary.

Analytical Tasks

Following steps were performed:

- Calculated the median value of “educatn”, the result was 12.0
- Categorized the total population into two variables (high_educated, low_educated) as shown in below (`median_education = 12`):

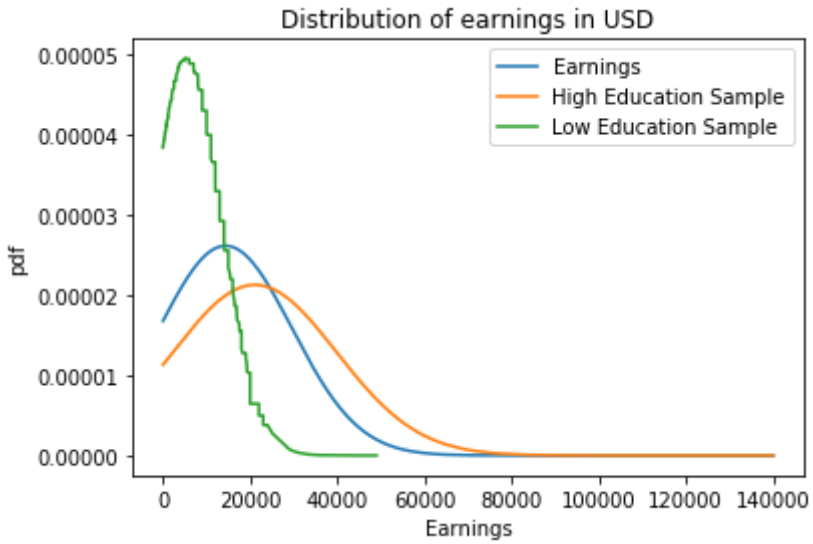
```
high_educated = orig[orig.educatn > median_education]
low_educated = orig[orig.educatn < median_education]
```

- We took two random two samples, for both categories, from the original population with the size of 800.
- Sorted the “earnings” values of both samples.

```
High_educated_earnings =
high_educated.sample(sample_size).earnings.sort_values()

Low_educated_earnings =
low_educated.sample(sample_size).earnings.sort_values()
```

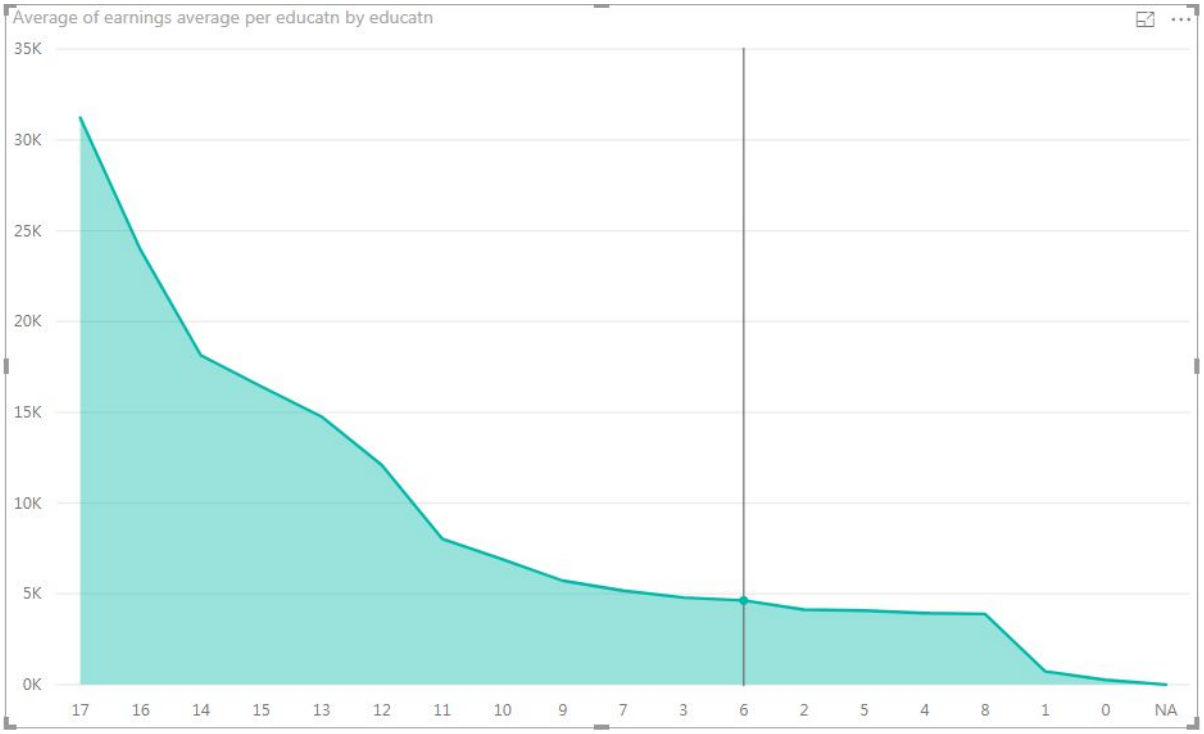
- Drew the graph of “Distribution of earnings in USD” for Original Population and two samples.



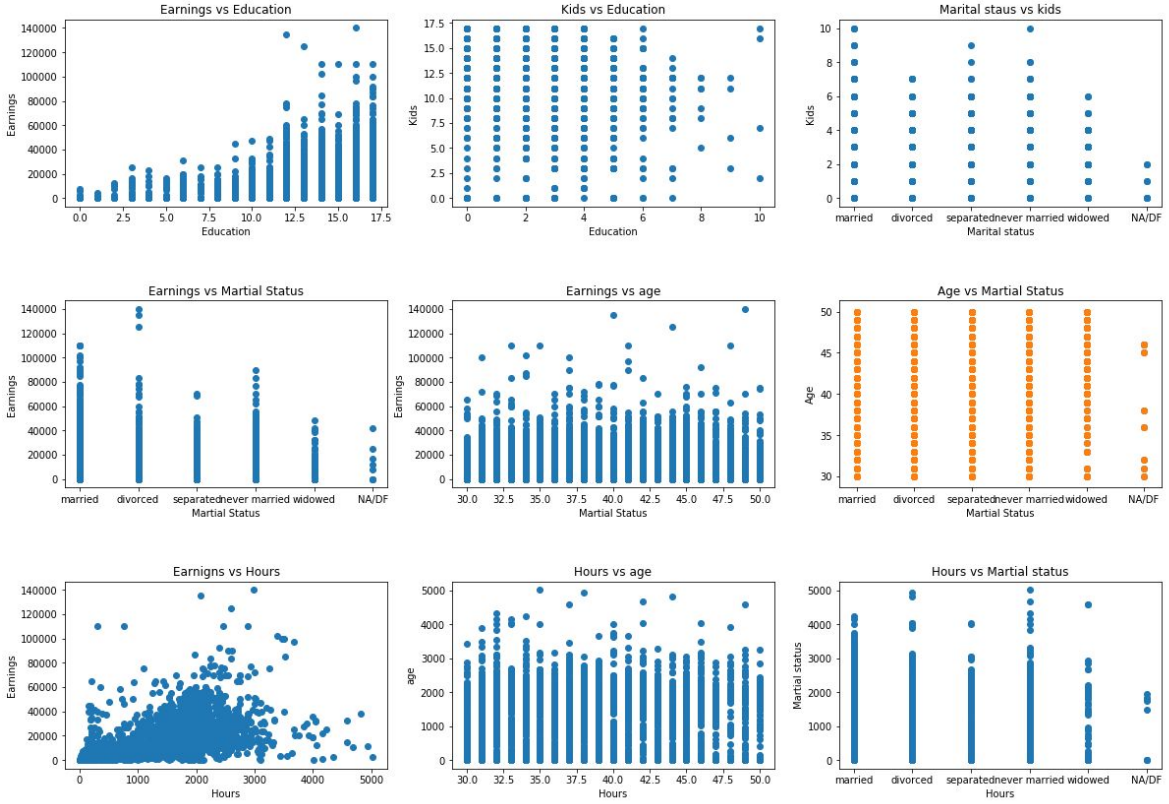
- Performed the hypothesis testing using the T-test method while calculating the p-value and statistics.

Visualizations

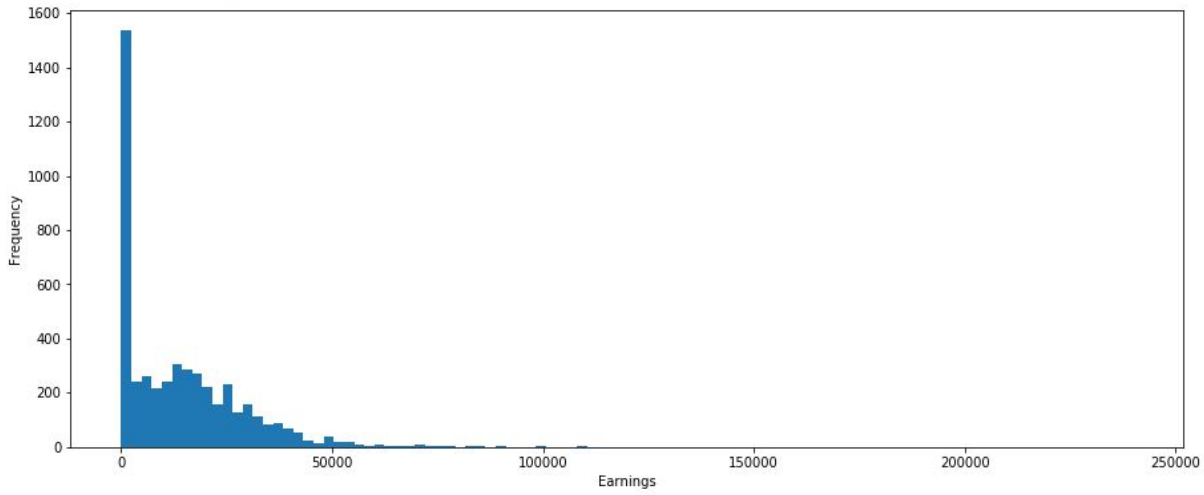
- Average earnings by education

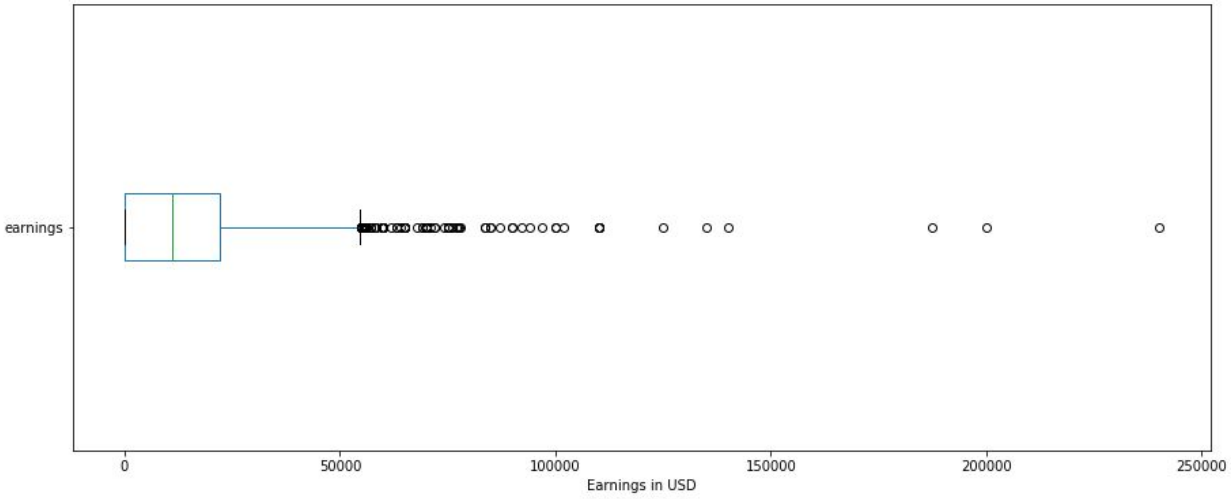


- Correlations

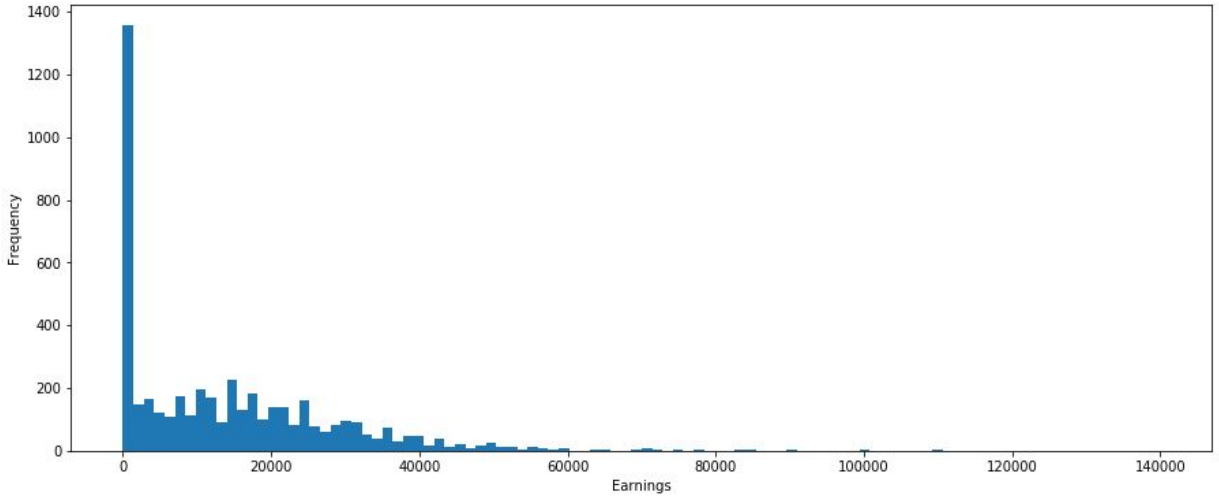


- Distribution of earnings before removing outliers:

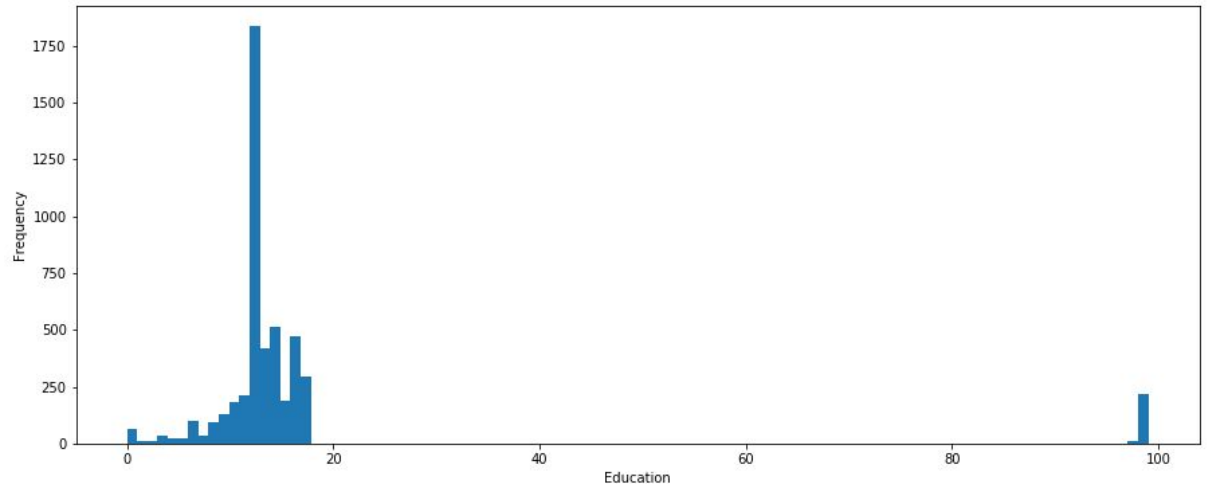


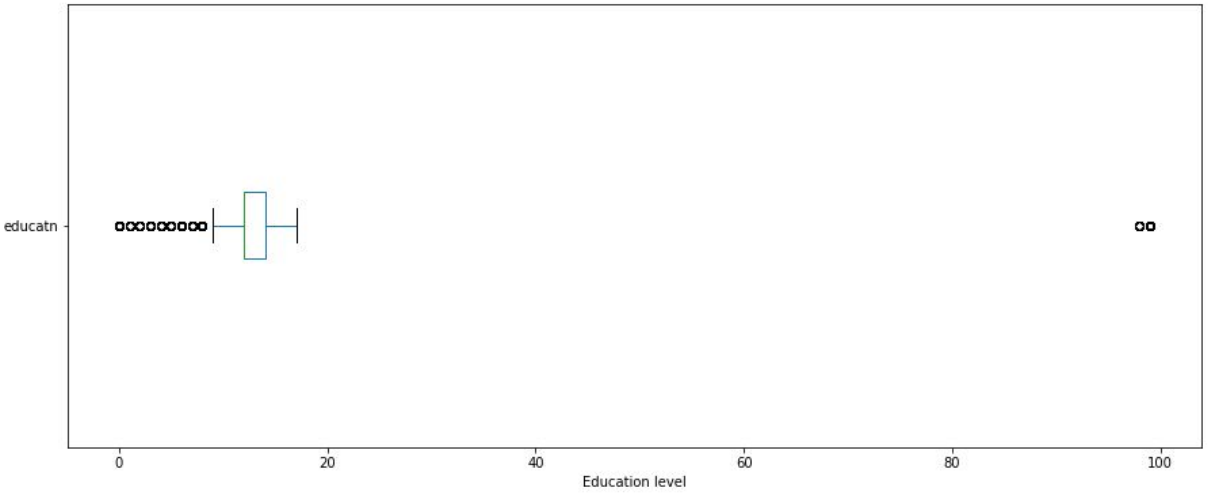


- Distribution of earnings after removing outliers (earnings > 150000):

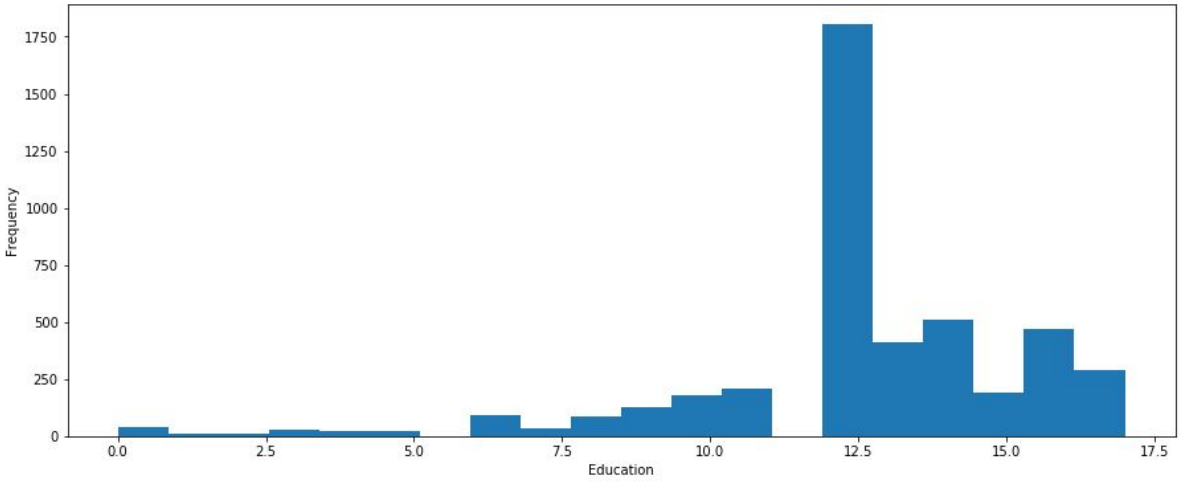


- Distribution of education level value (educatn) before removing outliers:



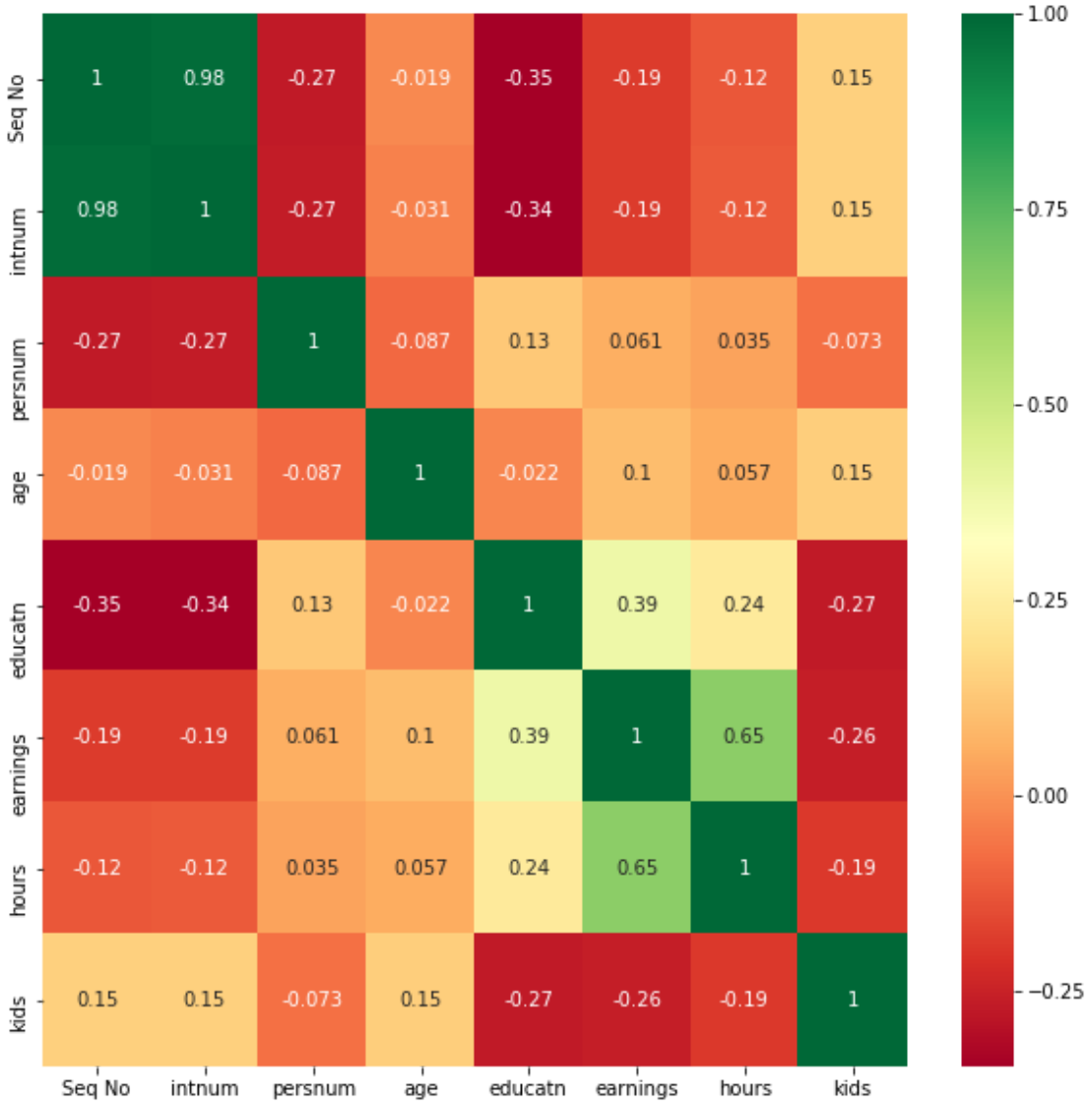


- Distribution of education level value (educatn) before removing outliers (“educatn” value > 20)



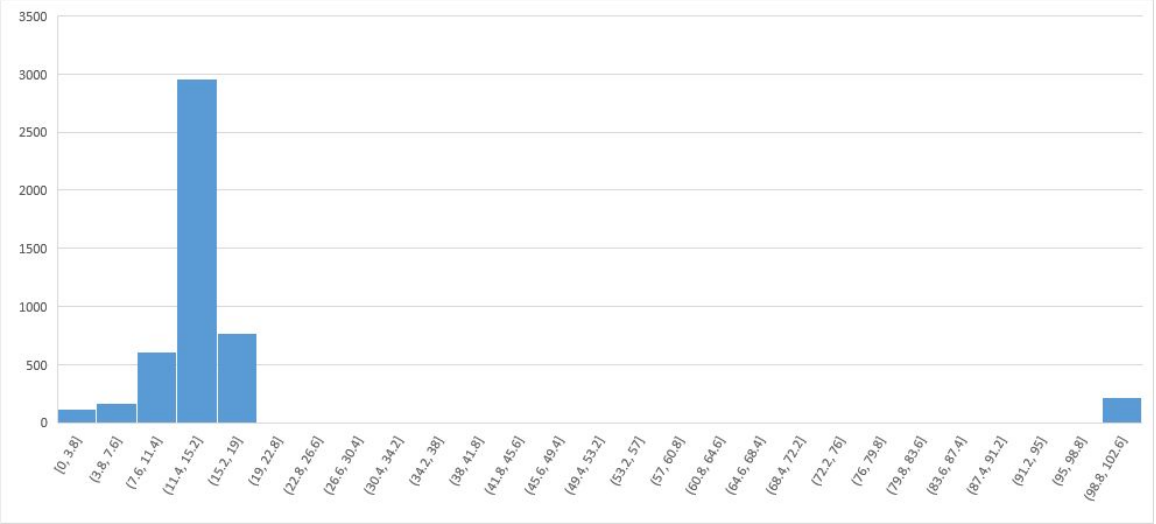
Other Useful Analytics

- Correlations values of each of the attributes of the given data set can be visualized in a heatmap as below:



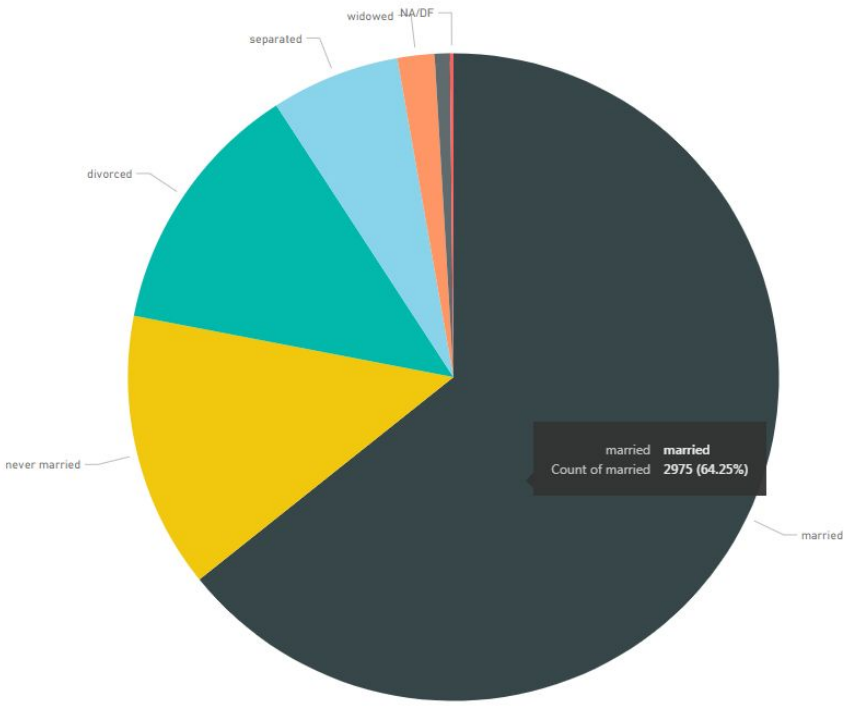
- There were multiple entries in the column of educatn (education) with the value of 98 and 99 seemed outliers as illustrated in below diagram;

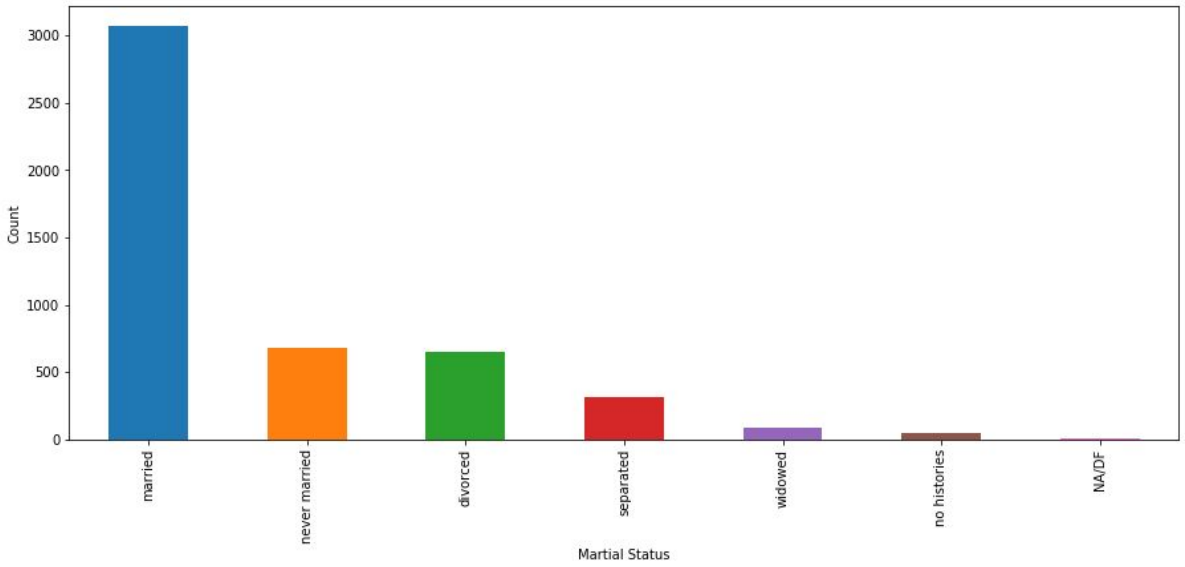
  X axis: educatn
  Y Axis: count



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

- It was noted that most of the people whose data had taken into this data set, were in the married status (64.25%).





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
[4] https://psidonline.isr.umich.edu/data/Documentation/UserGuide2015.pdf