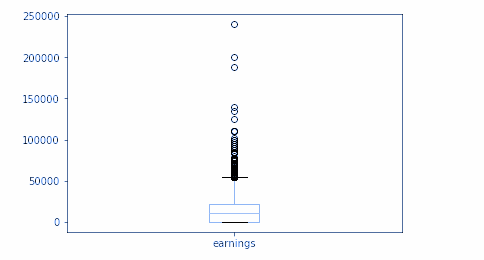
# Analytical Report

The PSID data set was collected for the Panel Study of Income Dynamics (PSID). This is a representative sample of over 18000 individuals from 5000 families in the United State. It contains data about employment, income, wealth, expenditures, health, marriage, childbearing, child development, philanthropy, education, and numerous other topics.

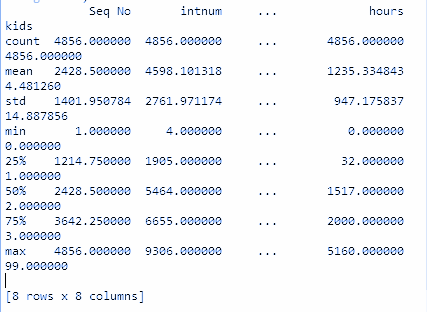
Since this is sample collected from the whole individuals in United State, there is a no need of sampling again.

Here the earnings is the key factor. First we try to find out are there any outliers in data set. After plotting a box plot, it can be seen that there are any outliers available. Outliers were omitted to do an accurate analysis.

Following is the output for boxplot.



After transformation of the data, next step is to do descriptive statistics for analysis of data set. Here we use “describe” method of panda class. Output is generated as follow.



According above output below table demonstrates the count, mean, standard deviation, min, max, Q1, median and Q3.

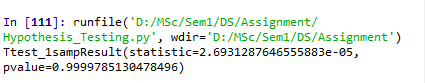
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Seq No** | **intnum** | **persnum** | **age** | **educatn** | **earnings** | **hours** | **kids** |
| **Mean** | **2428.5** | **4598.1** | **59.2** | **38.5** | **16.4** | **14244.5** | **1235.3** | **4.5** |
| **Median** | **2428.5** | **5464** | **4** | **38** | **12** | **11000** | **1517** | **2** |
| **SD** | **1402.0** | **2762.0** | **79.7** | **5.6** | **18.4** | **15985.4** | **947.2** | **14.9** |
| **Max** | **4856.0** | **9306.0** | **205.0** | **50.0** | **99.0** | **240000.0** | **5160.0** | **99.0** |
| **Min** | **1.0** | **4.0** | **1.0** | **30.0** | **0.0** | **0.0** | **0.0** | **0.0** |
| **Q1** | **1214.8** | **1905.0** | **2.0** | **34.0** | **12.0** | **85.0** | **32.0** | **1.0** |
| **Q3** | **3642.3** | **6655.0** | **170.0** | **43.0** | **14.0** | **22000.0** | **2000.0** | **3.0** |

**Hypothesis Testing**

stats library of scipy is used for statistical inference. Ttest\_1samp function consists with one sample t-test. So that, to test the mean of earnings against the null hypothesis with the mean of **14244.5.**

This returns two values. They are t : float or array (t-statistic) and prob : float or array (two-tailed p-value).

The output returned is as below.

From the above numerical output, it can be seen that the p-value = 0.9999 is greater than alpha=0.05, hence we can’t say that the average is not equal to **14244.5.**  

**Visualization**

There are many modules available for visualization in python. Here what we use is matplotlib library and searborn modules.

The below are output what we get as data visualization.

