Summary of the assignment

Team Contribution

- Thisun: Analysis & visualization
- Udeshika: Analysis, documentation, hypothesis testing and helps in visualization
- Kathusha: Analysis, documentation, sampling & testing
- Wasana: Analysis, documentation & testing

Git Repository Link

https://github.com/thisunhemakumara/In19-S1-CS5617-Data_Science_Assgnmnt_1

Hypothesis/Questions

H0 : mean of earnings of individuals in age range of 30-50 equals to 14245
Ha : mean of earnings of individuals in age range 30-50 not equal to 14245

The dataset considered was a sample of about 4857 taken from Panel Study of Income Dynamics(PSID) which was the longest running longitudinal survey in the world. We got a sample of 2000 from the existing sample dataset, considering dataset with 4857 dataset as population in order to carry-out hypothesis testing.

Assumptions

- The PSID dataset with 4857 tuples were considered as the population.
- The results done on the dataset which only reflects condition in USA applies to the whole population in the world.

References

[1] https://scipy-lectures.org/packages/statistics/index.html
[2] https://psidonline.isr.umich.edu/
[3] https://stackoverflow.com/questions/46711019/color-the-shaded-area-under-the-curve-distribution-plot-different-colors
[4] https://stackoverflow.com/questions/20011494/plot-normal-distribution-with-matplotlib/20026448
[5] http://www.datasciencemadesimple.com/standard-deviation-function-python-pandas-row-column/