Assignment Summary

1. Team & their responsibilities - TEAM_LSEG (Group Number 10)

Name	Index Number	Contribution
Prageeth Anjula	199304P	Data loading, data describing, outlier removal, sampling, hypothesis testing and documentation
Waruna Wickramasingha	199371P	Data visualization, data correlation measurements using pearson's correlation, hypothesis testing and documentation
Hesitha Wijayasinghe	199372U	Histogram viewing, Missing value replacing with mean, describing and documentation
Pumudinee Kumarasiri	199338X	Data visualization (Histograms, Normal Distributions), skewness removal and documentation

2. Git Repo Link
   https://github.com/PraAnj/DSGroupProject/commits/master
3. Hypothesis / Questions

   We claim that high earnings people are spending more time at work. We came to this hypothesis after visualizing histograms and measuring correlation between interested variables. Pearson's correlation value highlights high correlation between the 'hours' and 'earnings'. We prove this by taking 2 random samples hour data of equal size. First sample is taken from people who have high earnings than the median of earning values and second with earning values greater than the median earning values.
4. Assumptions
   1. Missing values are replaced with mean of the variable (Ex: Education column).
   2. Invalid data are dropped assuming they are mistakenly added by the dataset administrator. Following data are dropped,
      - i. Kids count > 20
      - ii. Education > 20 (There were education level beyond 90 as well)
      - iii. Earnings > 150000
      - iv. Hours of work == 0 (Assuming they are students hence has no contribution to our hypothesis testing
   3. Assume Hours data is normally distributed after removing people with 0 hours of work.
5. References

1. https://www.shanelynn.ie/using-pandas-dataframe-creating-editing-viewing-data-in-python/
2. https://machinelearningmastery.com/use-statistical-significance-tests-interpret-machine-learning-results/