SML MPA 1 Inferences Document

1. In this MPA Adult Income Data is used to do Statistical Analysis
2. The Data has below columns

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 age 32561 non-null int64

1 workclass 32561 non-null object

2 fnlwgt 32561 non-null int64

3 education 32561 non-null object

4 education.num 32561 non-null int64

5 marital.status 32561 non-null object

6 occupation 32561 non-null object

7 relationship 32561 non-null object

8 race 32561 non-null object

9 sex 32561 non-null int64

10 capital.gain 32561 non-null int64

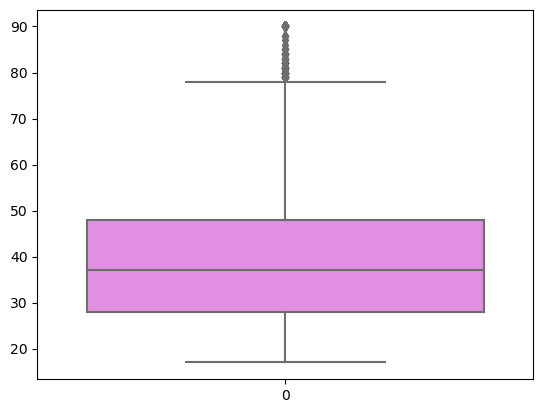
11 capital.loss 32561 non-null int64

12 hours.per.week 32561 non-null int64

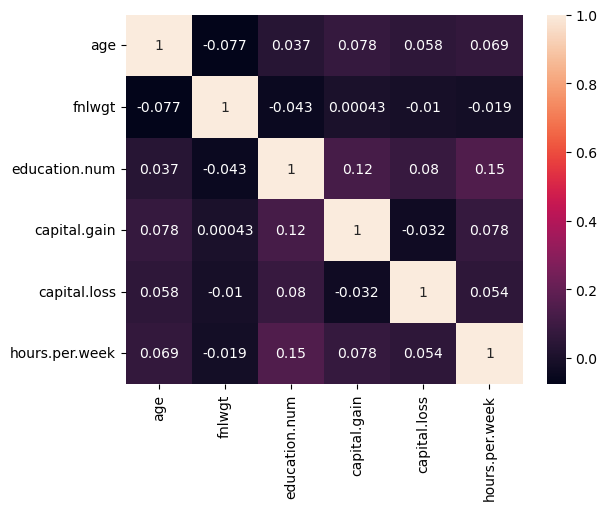
13 native.country 32561 non-null object

14 income 32561 non-null object

1. The Statistical Analysis needs to be done to understand Measure of central tendencies and variances of different Numerical variables of the data.
2. Usual Data Cleaning verifications are done.
3. Converted Numerical to and fro Categorical where necessary
4. **Mean of age** of workers is found to be around **37,39** by applying mean() function
5. **Most people have a qualification** of Higher Secondary Grade **(HS-grad)** using the mode of the data.
6. Median central tendency is studied using median functions applying group methods.
7. Variances are identified to be much more than normal (299.98 for working hours)
8. Interquartile range (**IQR**) is identified to be **5 for working hours**.
9. Using this IQR calculated above, **limits** to work for more number of hours and very less number of hours in the given week are identified as **52.5 and 32.5** respectively
10. From skew function we see that **education.num is negatively skewed**
11. Using **Kurtosis analysis** below observations are made for different variables (features)
    1. age is platykurtic
    2. fnlwgt, education.num, capital.gain, capital.loss, hours.per.week are leptokurtic
    3. There are no mesokurtic
12. With box plot extreme values in age are identified to be **around above 78**

****

1. Using correlation plot it is observed that most of the variable are not having much of correlation as they are not in the range of + or 0.6, - 0.8 to 1 but far too less.



16) The remaining questions are on probability where different probabilities are found using Poisson, binomial Distribution probability theories.