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       # represents Age
 1   workclass      32561 non-null  object      # working sector
 2   fnlwgt         32561 non-null  int64       #
 3   education      32561 non-null  object      # qualification degree
 4   education.num  32561 non-null  int64       # qualification converted to numerical
 5   marital.status 32561 non-null  object      # self explainatory
 6   occupation     32561 non-null  object      # Posting details
 7   relationship   32561 non-null  object      # pertaining to family
 8   race           32561 non-null  object      # breed
 9   sex            32561 non-null  int64       # gender
 10  capital.gain   32561 non-null  int64       # self explainatory
 11  capital.loss   32561 non-null  int64       # self explainatory
 12  hours.per.week 32561 non-null  int64       # self explainatory
 13  native.country 32561 non-null  object      # self explainatory
 14  income         32561 non-null  object      # self explainatory
```

3) The Statistical Analysis needs to be done to understand Measure of central tendencies and variances of different Numerical variables of the data.
4) Usual Data Wrangling was done.
5) Converted Numerical to and for Categorical for sex column

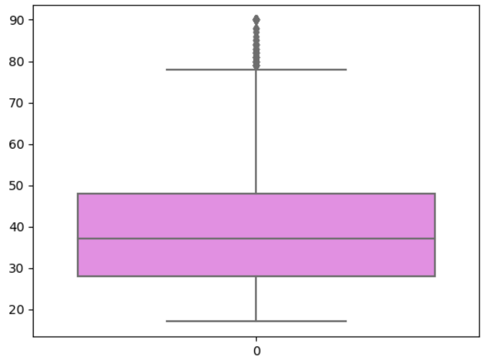
6) **Mean of age** of workers is found to be around **37,39** by applying mean() function

```
Female    39.0
Male      37.0
```

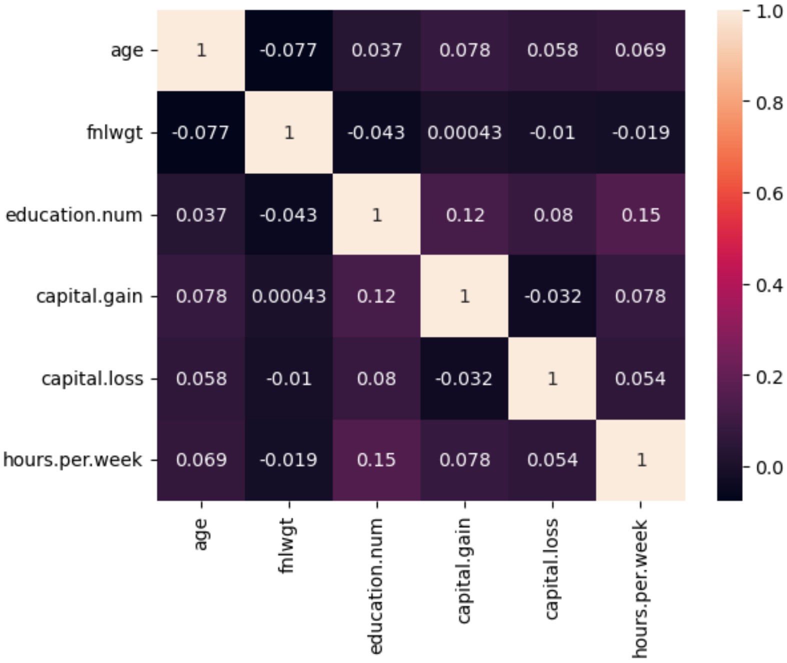
7) **Most people have a qualification** of Higher Secondary Grade **(HS-grad)** using the mode of the data.

8) Median central tendency is studied using median functions applying group methods.
   More are less all age groups work in all sectors and pay scale.
   Also, it is observed that people around 58 age are taking up non-profit jobs.

9) Variances are identified to be much more than normal
   It can be observed that Farming-Fishing has maximum variance.
   Followed by Priv-House-serv
10) Interquartile range (**IQR**) is identified to be **5 for hours per week**.
11) Using this IQR calculated above, **limits** to work for +& -1.5*IQR to Q1 and Q3 number of hours and very less number of hours in the given week are identified as **52.5 and 32.5** respectively
12) From skew function we see that **education.num is negatively skewed**
13) 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
14) With box plot extreme values in age are identified to be **around above 78**



15) 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. Detailed steps are provided in ipynb file as comments.