**Lab Evaluation 1**

Fifth Semester

Computer Science and Engineering

19CSE304 Foundations of Data Science

Duration: 60mins

Maximum: 20 Marks

**Course Outcomes (COs):**

|  |  |
| --- | --- |
| **CO** | **Course Outcomes** |
| CO01 | Understand the statistical foundations of data science. |
| CO02 | Apply pre-processing techniques over raw data so as to enable further analysis. |
| CO03 | Conduct exploratory data analysis and create insightful visualizations to identify patterns. |
| CO04 | Identify machine learning algorithms for prediction/classification and to derive insights |
| CO05 | Analyse the degree of certainty of predictions using statistical test and models |

A. Consider the dataset on Education and Income of Californian Adults. The Objective is to analyse the data for the year 2011, for adults (age >18).

https://github.com/flaviovdf/evcomp2018/raw/master/data/educ\_inc.csv

1.Group the table by Educational Attainment and sum the Population Count in each category(2).

2. Analyse the percentage distribution of educational attainment among adult Californians(2).

3. Using pivot , get a contingency table (a table of counts) of adult Californians cross-classified by Educational Attainment and Personal Income(2).

B. Consider the dataset on the salaries of NBA players.

'https://github.com/flaviovdf/evcomp2018/raw/master/data/nba\_salaries.csv'

4.List those NBA players who earned more than 10m USD(2).

5.How much money did each team pay for its players' salaries(2)?

C. Consider the dataset test1.csv

1. Locate the first tuple corresponding to the last 3 digits of your Roll No, and check if the income is above 50K(1)

7.Rename the columns as follows:[‘Srl No’, 'age', 'type\_employer', 'fnlwgt','education', 'education\_num','marital','occupation','relationship','race','sex','capital\_gain', capital\_loss', 'hr\_per\_week', 'country', 'income'] (1)

8.Analyse the proportion of high income professionals (>=50K) in the database (1).

9.Find the skew and kurtosis in terms of age(1).

10.Find the average age, and standard deviation for female population. (1)

11.Compute the mean and the variance of "hours per week" earned by men(1).

12.Draw a histogram for age of men. Histogram should be step-filled with bin of size 20. Find the skew of this distribution, and comment if it is positive or negative(2).

13. Draw a Box Plot of "age" vs "income"(2).

=====================================================================