Amrita Vishwa Vidyapeetham

Amrita School of Computing, Coimbatore

**Lab Evaluation 1**

Fifth Semester

Computer Science and Engineering

19CSE304 Foundations of Data Science

Duration: 60mins

Maximum: 32 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 |

For the given data set (test1.csv) perform the following.

1. Import data and save it in your disk space (1)
2. Describe the data(1).
3. Display the first 3 rows(1).
4. Display the last 4 rows(1)
5. Locate the first row corresponding to the last 3 digits of your Roll No(2)
6. Rename column <=50K as "Income", and display all the columns(1)
7. Check if income of the highest row ending with the last 3 digits of your roll number is above 50K(1)
8. 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'] (2)
9. How many rows of data do you have? (1)
10. Display the first 5 columns of the first row(2)
11. Count the number of items per country(2)
12. Split people according to their gender into two groups: men and women. How many males are there?(2)

Contd..

-2-

1. Focus on high-income professionals (>=50K) separated by sex(2).
2. Print the percentage rate of males with high income >=50K(1)
3. Find the skew and kurtosis in terms of age and comment(1)
4. Analyse the proportion of high income male professionals in the database(2)
5. Evaluate the characteristics of the population distribution with mean, deviation, histograms, outliers, etc. (a) Average age of men (b) Average age of women (c) Average `age of high-income men (2)
6. Compute the mean and the variance of "hours per week" for men(1).
7. Find the median age of men and women(1)
8. Find the median age of men and women with high income(1)
9. 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).
10. Draw a BoxPlot of "age" vs "income"(2).

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