Roll No.: _

Amrita Vishwa Vidyapeetham

Amrita School of Computing, Coimbatore

B.Tech First Assessment Examinations – October 2022

Fifth Semester

Computer Science and Engineering

19CSE304 Foundations of Datascience

Duration: Two hours Maximum: 50 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

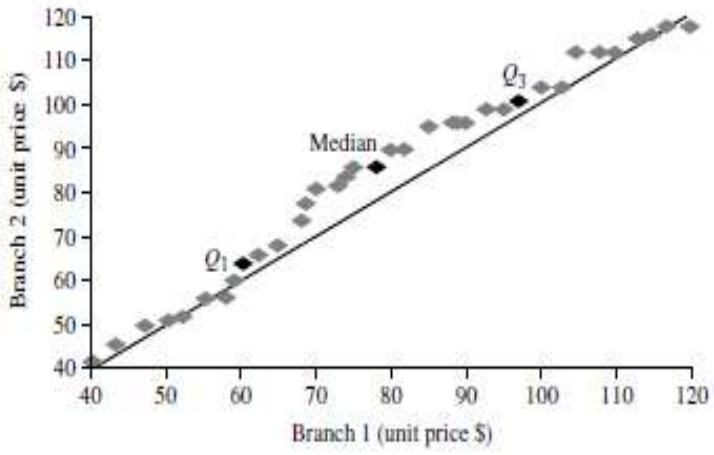
Answer all questions

Part A **(5 x 2 =10 Marks)**

1. Given the following data x1(0,2), x2(3,5), x3(6,5) and x4(0,5). What is the Manhattan distance between x1 and x3? [2][CO01][BTL2]

2. If covariance is zero between two sets of data, then can we assume the data to be independent? [2][CO01][BTL2]

3. Identify the plot given below and interpret the information on the unit sales price prevailing in the two branches. [2][CO03][BTL3]



4. Given the following data x1(4,2), x2(3,5), x3(6,5) and x4(0,5). Identify the maximum value of supremum distance amongst them. [2][CO01][BTL2]

5. Suppose the population figures (in millions) in the states of a country are:
[30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, and 110].
Find the prevailing standard deviation. [2][CO01][BTL3]

Part B **(8 x 5 = 40 Marks)**

6. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem. [5][CO02] [BTL2]

7. The mean of a normally distributed data set is **5** and the standard deviation is **1**.
(a) What is the corresponding z value of **7**? [2][CO01][BTL2].
(b) What percentage of this data distribution would you expect to be greater than **7**? [2][CO01][BTL2].
(c) Would this dataset qualify as a standard normal distribution? Justify. [1][CO01][BTL2].

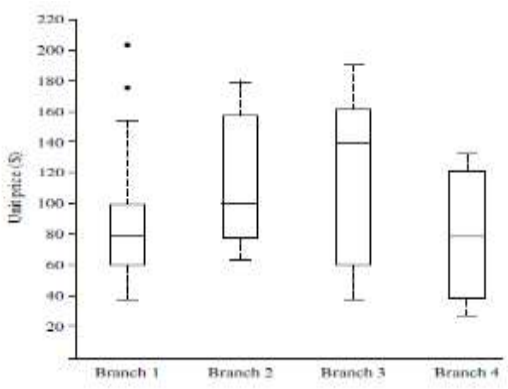
8. (a) The coefficient of kurtosis in respect of a dataset is calculated as **7**. Find the Excess Kurtosis present in the data. [1][CO03][BTL4]
(b) What inference do you make from this information regarding its distribution? [1][CO03][BTL4]
(c) Sketch the various types of kurtosis and name them**.** [3][CO03][BTL4]

9. (a) State the four fundamental classifications of scales of Measurement. Elucidate their properties, giving an example of each type. [2][CO01][BTL2]
(b) Distinguish them by their qualitative/ quantitative nature [1][CO01][BTL2]
(c) The central tendencies of data can be expressed by their Mean, Median and Mode. Map the best method of describing the central tendencies in respect of each of these. [2][CO01][BTL2]

10. (a) Find the similarity between two vectors 'x' and 'y', using Cosine Similarity.
x = { 3, 2, 0, 5 } and = { 1, 0, 0, 0 }. [4][CO01][BTL2].
(b) How would you describe their dissimilarity? [1][CO01][BTL2].

11. Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45,46, 52, 70.
(a) What is the mean, and mode of the data? [1][CO03][BTL3]
(b) Comment on the modality of the data. [1][CO03][BTL3]
(c) Give the 5 number summary of the data. [2][CO03][BTL3]
(d) Draw the box plot. [1][CO03][BTL3]

12. The box plot of the unit prices of items sold in four branches are shown below. Analyse the plot and identify the following: [5][CO03][BTL4]



(a) The 5 number summary in Branch2 and its corresponding IQR. Explain the variation in the lengths of the 2 whiskers in Branch2. [2][CO03][BTL4]
(b) Outliers, if any. [1][CO03][BTL4]
(c) Negatively skewed data [1][CO03][BTL4]
(d) Positively skewed data [1][CO03][BTL4]

13 (a) Find the Pearson Correlation Coefficients of the following data. [3][CO02][BTL3]

x	y
6	12
8	10
10	20

(b) Comment on the relationship between x and y parameters. [1][CO02][BTL3]
(c) Does correlation imply causation? [1][CO02][BTL3]

****

Course Outcome /Bloom's Taxonomy Level (BTL) Mark Distribution Table

CO	Marks	BTL	Marks
CO01	**23**	BTL 1	
CO02	**10**	BTL 2	**21**
CO03	**17**	BTL 3	**19**
CO04		BTL 4	**10**