Strathmore UNIVERSITY

School of Computing and Engineering Sciences

Master of Science in Information Technology

End of Semester Examination

MSSET 8103: Data Science Concepts

Date: 22nd April 2024 **Time: 18:00-20:30 Hours**

Instructions: Answer Question **ONE** and any other **TWO** Questions

Question ONE (20 Marks) (Compulsory)

a) How do classification algorithms differentiate and categorize data points based on their features and attributes, and what are the key steps involved in the classification process? [6 marks]

b) Explain the importance of Data mining analysis and give examples of data mining techniques. [3 marks]

c) Give three statistical measures for the characterization of data dispersion, and discuss how they can be computed efficiently in large databases. [6 marks]

d) A sample of 10 households were monitored for one year. The household income (in $1000s) and the amount of energy consumed (in 10^{10} joules) were determined. The results recorded were as follows:

Income	31	40	28	48	195	96	70	100	145	78
Energy	16	40	30	46	165	92	96	77	115	67

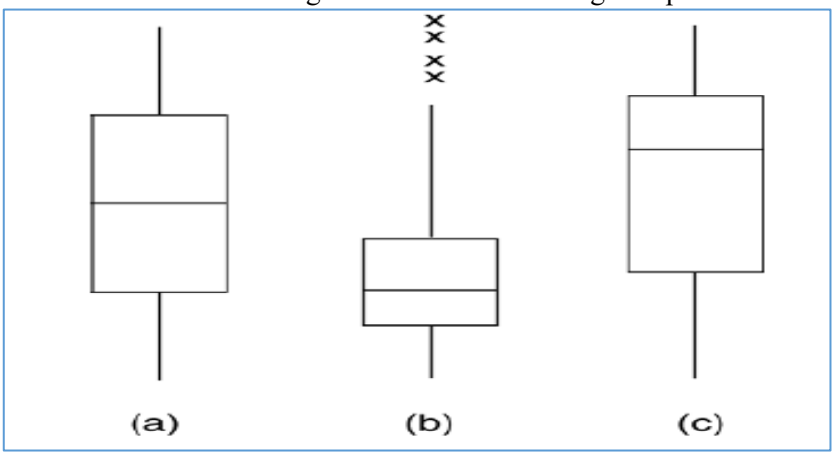
Determine the correlation coefficient between income and energy consumption. What does the results from the correlation coefficient indicate about income and energy consumption in the city? [5 marks]

Question TWO (15 Marks)

a) Suppose that the thickness of a part used in a semiconductor is its critical dimension and that measurements of the thickness of a random sample of 18 such parts have the variance s2 = 0.68, where the measurements are in thousandths of an inch. The process is considered to be under control if the variation of the thickness is given by a variance not greater than 0.36.

Assuming that the measurements constitute a random sample from a normal population, state the null hypothesis and test it against the alternative hypothesis at the $\alpha = .05$ significance level. [6 marks]

b) A data scientist conducting data visualization using box-plot obtained the figures below.



What inferences can be made from the three outcomes? [6 marks]

c) What is the difference between a variable and a random variable? Provide an example of a random variable. [3 marks]

Question THREE (15 Marks)

a) What are the primary techniques and methodologies used in data exploratory analysis, and how do they contribute to uncovering patterns, trends, and insights within a dataset? [6 marks]

b) Explain the k-means clustering algorithm and its steps. [3 marks]

c) Suppose a sample of 16 light trucks is randomly selected off the assembly line. The trucks are driven 1000 miles and the fuel mileage (MPG) of each truck is recorded. It is found that the mean MPG is 22 with a SD equal to 3. The previous model of the light truck got 20 MPG.

i. State the null hypothesis for the problem above. [2 marks]
ii. Conduct a test of the null hypothesis at p=0.05 and make appropriate conclusion. [4 marks]

Question FOUR (15 Marks)

a) Discuss the application of clustering in data mining in the energy sector. Further, with examples explain how the k-means algorithm can be done. [6 marks]

b) What is anomaly detection? Explain distanced based method for anomaly detection. [6 marks]

c) When conducting data analysis, different statistics are obtained that help in the interpretation of your results. One of these statistics is the p-value. Explain how you will interpret a p-value from your data analysis. [3 marks]