Strathmore UNIVERSITY

School of Computing and Engineering Sciences

Master of Science in Energy Transitions

End of Semester Examination

MSET 8103: Data Science Concepts

Date:3rd May 2023 **Time 2 1/2 Hours**

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

Question ONE (20 Marks) (Compulsory)

a) Discuss the goals of regression analysis? [3 marks]

b) Suppose a dataset consisting energy consumption in a neighborhood of Nairobi were recorded as follows 567,1823, 517, 583, 317, 367, 250, 503, 317, 567, 583, 517, 650, 567, 450 and 350. Using r-software syntax give the code segments that can be used to;

i. Compute the mean energy consumption [2 marks]

ii. Generate a tabular presentation of the data set [2 marks]

iii. Identify the value that occurs most often [2 marks]

iv. Compute the standard deviation of the energy consumption [2 marks]

c) Suppose you are conducting a study that involves the collection of data and upon creating a scatter plot to visualize the dataset you realize a number of outliers are evident in the dataset. With reasons explain how you would handle the outlier values in the dataset. [3 marks]

d) The company "General Electric Inc." claims that a certain brand of its flashlight battery lasts on average 300 hours of flashlight use. You suspect that the population of batteries average fewer than 300 hours. You select a random sample of 49 batteries and obtain a sample mean of 290 and a sample standard deviation S=70.

i. Perform a one-tailed hypothesis test with the company's claim in the null hypothesis. Use a level of significance of .10. [3 marks]

ii. Calculate the probability value of the test statistic. Interpret the results. [3 marks]

Question TWO (15 Marks)

a) The following table shows the hours of sunshine, x, during nine days in October and the number of ice creams y sold by a beach shop in Mombasa.

x	4.3	6.9	0.0	10.4	5.2	1.8	8.0	9.2	2.1
y	224	208	123	419	230	184	362	351	196

i. Establish an equation of the regression line of y on x. [4 marks]

ii. Calculate the residuals for the days when the number of hours of sunshine was 8.0 and 6.9. [3 marks]

iii. One of the days the shop was closed early to allow the owner to attend a birthday party. Suggest, with reasons, which day this was. [2 marks]

b) Explain the importance of Data mining analysis and give examples of data mining techniques. [3 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) Momanyi and Kantai did a study on feelings of stress and life satisfaction due to low quality of electrical energy in their locations. Participants completed a measure on how stressed they were feeling on a scale of 1 to 30 and a measure of how satisfied they felt with their lives measured on a scale of 1 to 10. The table below indicates the participants' scores. Use this data to answer the questions that follow:

Participant #	**Stress score (x)**	**Life satisfaction(y)**
1	11	7
2	25	1
3	19	4
4	7	9
5	23	2
6	6	8
7	11	8
8	22	3
9	25	3
10	10	6
Sum Mean Sd	159 15.9 7.23	51 5.1 2.70

i. Calculate the correlation (r) between stress and life satisfaction. [6 marks]

ii. Write a brief interpretation of the correlation, including the strength, direction and an explanation of the effect. [2 marks]

iii. Can you conclude that being more stressed causes a lower level of life satisfaction? Why or why not? [2 mark]

b) 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]

c) The frequency of a certain feature's values can be presented visually by both box plots and histograms. What is the difference between a box plot and a histogram? [2 marks]

Question FOUR (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. [7 marks]

b) In hypothesis testing, we are often observing a sample and not an entire population. It is therefore possible that a conclusion may be wrong resulting into type I and type II errors.

i. What type of error do we directly control? [1 marks]

ii. What type of error is associated with decisions to retain the null? [1 marks]

iii. What type of error is associated with decisions to reject the null? [1 marks]

iv. State the two correct decisions that a researcher can make. [1 marks]

c) "Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed." -Arthur Samuel (1959). The most popular types of machine learning are supervised and unsupervised learning. What are the differences between supervised and unsupervised learning? [4 marks]