National University of Computer and Emerging Sciences, Lahore Campus

	Course Name:	Data Science	Course Code:	CS4048
	Degree Program:	BS(Computer Science)	Semester:	Spring 2022
	Exam Duration:	60 Minutes	Total Marks:	40
	Paper Date:	24-Mar-2022	Weight	10%
	Section:	ALL	Page(s):	4
	Exam Type:	Midterm-I		

Student : Name:_____ **Roll No.**_____ **Section:**_____

Instruction/Notes: Attempt all questions. Programmable calculators are not allowed.

Q1. [10 marks]

1. Which statement is IN-CORRECT:
   a. In the interval scale, distances between each value on the scale are equal
   b. In the ordinal scale, distances between each value on the scale are not equal
   c. The ratio scale is the most informative measurement scale
   d. We can calculate mean on the nominal scale
   e. We can identify outliers in the interval scale

2. Which statement is IN-CORRECT:
   a. Jersey numbers for a football team is a nominal scale
   b. Military rank is an ordinal scale
   c. Shoe size is an ratio scale
   d. Year of birth is an interval scale
   e. T-shirt size (small, medium, large) is a ordinal scale

3. Which of the following is NOT-NECESSARY for valid data collection and analysis:
   a. Identify sub-groups during data collection
   b. Precisely define target group for data collection
   c. Ensure proportional participation by all sub-groups
   d. Always perform internet surveys for maximum participation
   e. Identify stigmatized respondents for possible omissions

4. What is IN-CORRECT about missing data:
   a. Missing data at random (MAR) is easier to handle than Missing completely at random (MCAR)
   b. MNAR can only be identified by considering external factors in addition to the collected data
   c. MAR can be identified using only the collected data
   d. Simpler methods like mean substitution or regression can be used with MCAR
   e. More advanced methods should be used for MAR/MNAR

5. Which statement is IN-CORRECT:
   a. Data scientists spend most of their time applying machine learning algorithms.
   b. Data science requires good business understanding
   c. A data scientist is an excellent communicator who is able to work with multi-disciplinary teams
   d. A data scientist should be well versed in statistical techniques
   e. Data science requires research aptitude

Q2. [20 marks]

Gradient descent - Linear regression - single variable

Hypothesis:	$h_\theta(x) = \theta_0 + \theta_1 x$
Cost function:	$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$
Gradient descent algorithm:	$\text{repeat until convergence } \{$ $\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$ $\} \qquad (\text{for } j = 0 \text{ and } j = 1)$
Partial derivatives	$\frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1) = \sum_{i=1}^{m} (h_\theta(x_i) - y_i)$ $\frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1) = \sum_{i=1}^{m} ((h_\theta(x_i) - y_i) x_i)$
Parameter updates will be simultaneous	$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$ $\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$ $\theta_0 := \text{temp0}$ $\theta_1 := \text{temp1}$

Consider following values of X (predictor variable) and Y (target variable)

X	Y
3	6
6	13

Also consider the following values

m = 2 (number of data points)
$\alpha = 0.001$ (learning rate)
$\theta 0 = 0$
$\theta 1 = 1$

Calculate the following (show working):

1. Calculate the value of Cost Function

$J(\theta_0, \theta_1) =$ 14.5

2. Perform one iteration of gradient descent algorithm and calculate new values of θ_0, θ_1

θ0 = 0.005 (0.01 without m)

θ1 = 1.0255(1.051 without m)

3. Using new values of θ_0, θ_1 calculate updated value of the cost function

$J(\theta_0, \theta_1) =$ 13.83 (13.18 without m)

Have your model learned by reducing error? _____

Q3. [10 marks]

Draw Box-and-Whisker Plot for the following data. Also, identify outliers if there are any.

25,35,42,47,48,49,50,50,52,53,55,56,58,75,81

You are also required to calculate the following values:

Median = 50

Q1 = 47

Q3 = 56

IQR (Q3 - Q1) = 9

Outliers = 25, 81, 75

Minimum: 25

Maximum: 81