National University of Computer and Emerging Sciences, Lahore Campus

Course Name:	Data Science	Course Code:	CS-4048
Degree Program:	BS(Computer Science)	Semester:	Spring 2023
Exam Duration:	60 Minutes	Total Marks:	36
Paper Date:	8-April-2023	Weight	15%
Sections:	All	No of Page(s):	3
Exam Type:	Midterm-II		

Student : Name:_______________________________ Roll No.________________ Section:_______

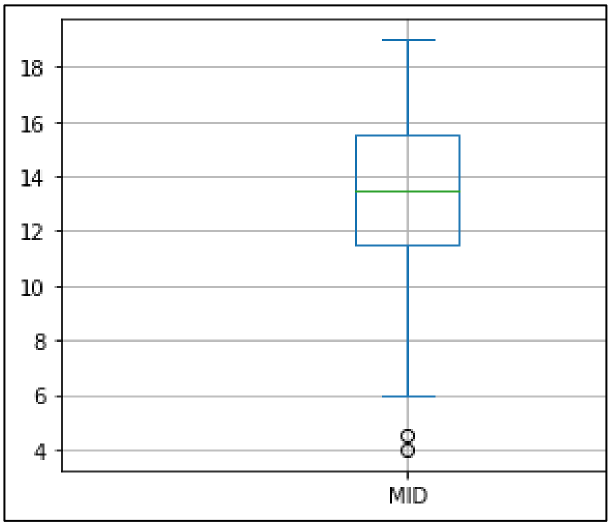
Instruction/Notes: There are 3 questions. Attempt all question in sequence.

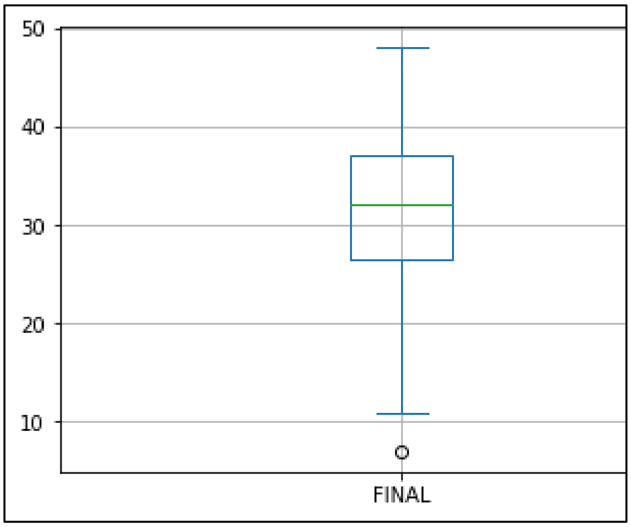
Problem#1 (CLO-3) **[2+2+8 =12 Marks]**

a) How can visualizations be used to identify outliers, anomalies and correlations within data sets?

b) Suppose you're given a dataset that tracks the daily sales of three different products (Product A, Product B, and Product C) over the course of a year. Which type of graph or chart would be best to visualize this data? Explain your reasoning.

c) Fill the given table based on the box plots shown below. Q1, Q2 and Q3 indicates the quartiles. Distribution type could be normal, bimodal, multimodal, left skewed, right skewed, etc.





	Min	Max	Q1	Q2	Q3	Median	No. of Outliers	Distribution Type
MID								
FINAL								

Problem#2 (CLO-3) **[2+2+4+4=12 Marks]**

a) Differentiate covariance and correlation.

b) Explain the term "Curse of Dimensionality".

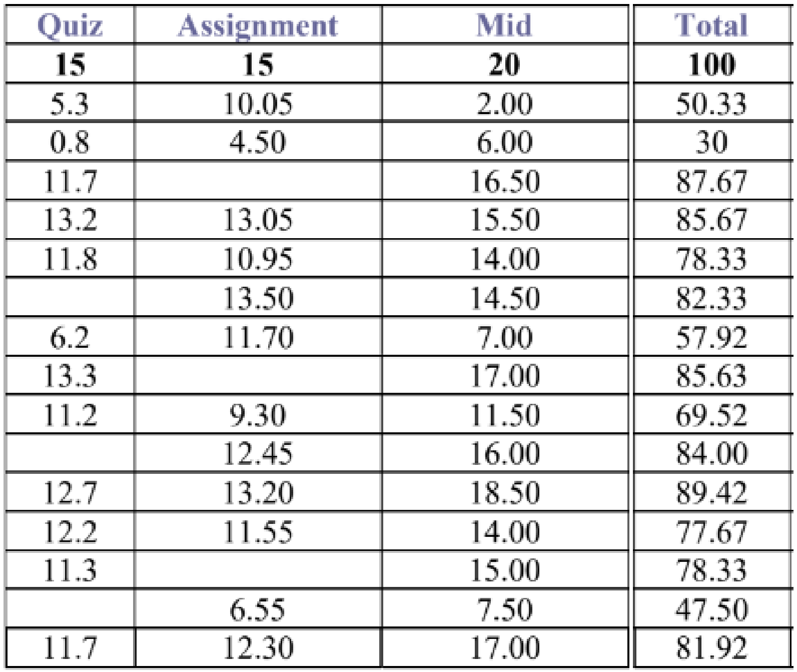
OR

b) Given the eigenvalues λ1 = 1.284028 and λ2 = 0.04908323, along with their corresponding eigenvectors v1 = [0.6778736, 0.7351785] and v2 = [-0.7351785, 0.6778736], Calculate the percentage of total variance (information) carried by each component?

c) Fill the missing values in the given dataset using 'ffill' method and then normalize Mid and Total columns of given dataset. Skip first row containing total marks for each grading component.

d) A subset of given dataset containing first 5 rows and first three column is selected for applying PCA. Eigen vectors and Eigen values are given below. Transform the subset and reduce it to have two features.

Eigen values: [66.095, 1.66, 17.011]
Eigen Vector: [[-0.63, -0.73, -0.26]
[-0.18, 0.46, -0.87]
[-0.76, 0.50, 0.42]]



Quiz	Assignment	Mid	Total
15	**15**	**20**	**100**
5.3	10.05	2.00	50.33
0.8	4.50	6.00	30
11.7		16.50	87.67
13.2	13.05	15.50	85.67
11.8	10.95	14.00	78.33
	13.50	14.50	82.33
6.2	11.70	7.00	57.92
13.3		17.00	85.63
11.2	9.30	11.50	69.52
	12.45	16.00	84.00
12.7	13.20	18.50	89.42
12.2	11.55	14.00	77.67
11.3		15.00	78.33
	6.55	7.50	47.50
11.7	12.30	17.00	81.92

Only For Section BSCS-8A

Problem#3 (CLO-3) **[4x3 = 12 Marks]**

Consider the data set given in problem 2. A linear regression model is trained using first 10 records and the values of slope are 3.57210049, 0.15607937, 0.86148416 respectively. Intercept is 24.94 and R2 score is 0.957.

a) Interpret the meaning of the values of slope, intercept, and R2 score.
b) Fill the missing values using 'ffill' method. Write regression line equation. Predict Total marks for last 5 records.
c) Calculate and interpret MAE and RMSE.

Only For Section BSCS-8B

Problem # 3 **[9+3= 12 Marks]**

Part a:

Given the following data on the weight (in pounds) and height (in inches) of 5 individuals:

Person	*Height (in)*	*Weight (lbs)*
1	68	150
2	70	160
3	62	120
4	75	190
5	66	135

Calculate the coefficient of determination for the relationship between weight and height. What does this value tell you about the strength and direction of the relationship?

Correlation formula: **r = (nΣXY - ΣXΣY) / sqrt[(nΣX^2 - (ΣX)^2)(nΣY^2 - (ΣY)^2)]**

Part b:

Suppose you have a dataset with one missing value in the predictor variable X and a strong positive correlation between X and the response variable Y. You decide to use linear regression to impute the missing value. The available data points are:

X	**Y**
2	5
4	9
5	11
6	13
8	17

What is the missing value in X if the predicted value of Y is 15?

Given: Θ_0= 0.7273 and Θ_1= 2.0404