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:	15
Paper Date:	8-April-2023	Weight	15%
Sections:	All	No of Page(s):	2
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**

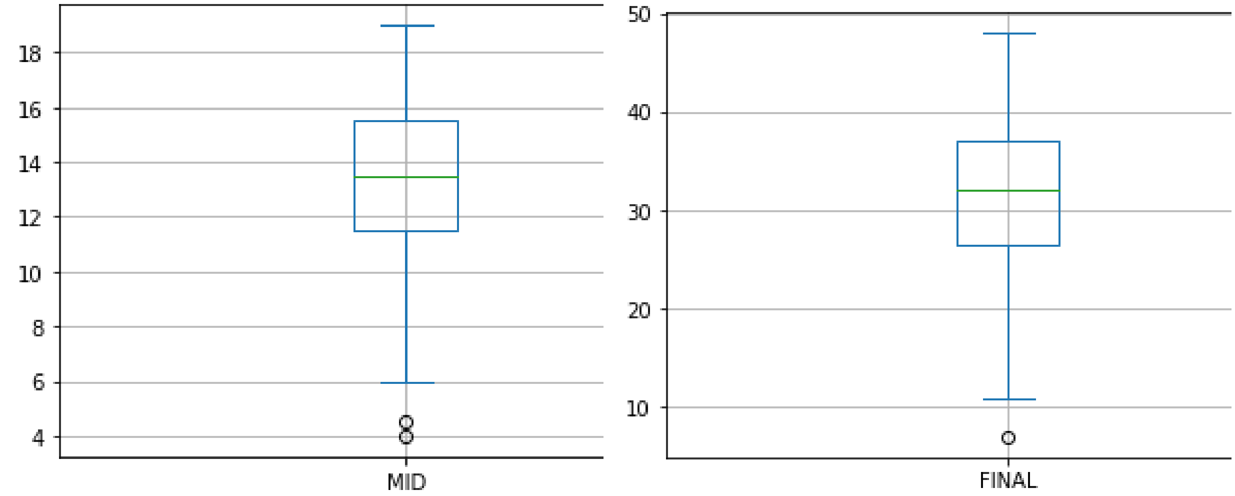
a) How can visualizations be used to identify outliers, anomalies and correlations within data sets?

Box and whisker plots are used to detect outliers and anomalies in the data. scatter plots can be used for the outlier analysis as well as correlation analysis.

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 .

Line chart. A line chart with a comparison of three products can represent the number of sales increased/decrease over a course of a year.

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, skewed etc.



	Min	Max	Q1	Q2	Q3	Median	No. of Outliers	Distribution Type
MID	6	19	11.5	13.5	15.5	13.5	2	left skewed
FINAL	11	48	26.5	32	37	32	1	left skewed

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

a) Differentiate covariance and correlation.
Covariance represents the direction of the relationship between two variables while correlation indicates the direction and strength of the relationship.

b) Explain the term "Curse of Dimensionality".
The curse of dimensionality basically refers to the difficulties a machine learning algorithm faces when working with data in higher dimensions. Too many features run the risk of massively overfitting our model and overburdening the computing power, time, and complexity.

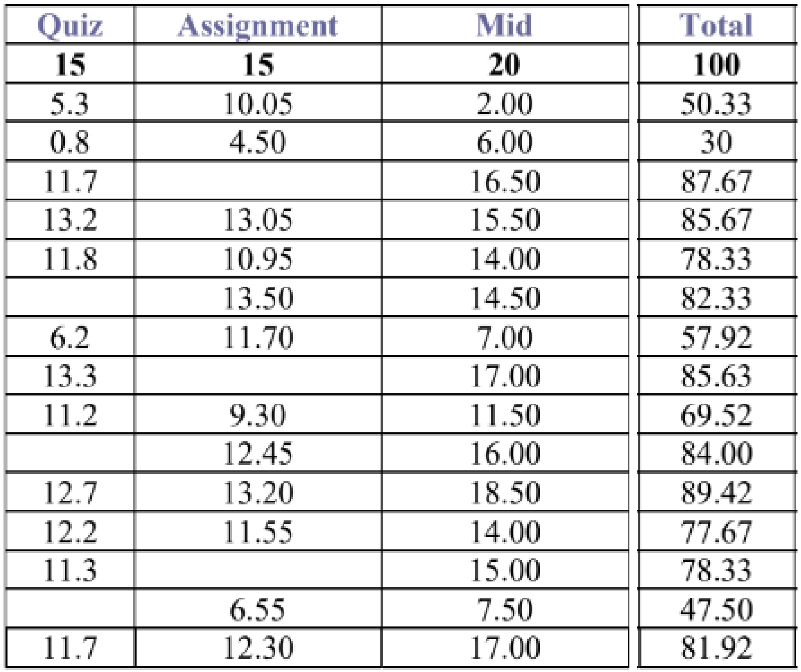
c) Fill the missing values using 'ffill' method and then normalize Mid and Total columns of given dataset.

```
[MID,        FINAL]
[0.         , 0.34214069],
[0.24242424, 0.        ],
[0.87878788, 0.97054864],
[0.81818182, 0.97054864],
[0.72727273, 0.8133625 ],
[0.75757576, 0.88067991],
[0.3030303 , 0.46987546],
[0.90909091, 0.93621676],
[0.57575758, 0.66509593],
[0.84848485, 0.90878492],
[1.         , 1.        ],
[0.72727273, 0.80225513],
[0.78787879, 0.8133625 ],
[0.33333333, 0.29451363],
[0.90909091, 0.87377987]
```

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]]
```

```
Reduced:        [[ -6.61586638,  -9.2859981 ],
                [ -5.83960932,  -1.60730559],
                [-20.64596773,  -0.05731228],
                [-22.33633098,  -8.30263978],
                [-19.95035994,  -6.73927113]]
```



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

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

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.

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

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

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)]

Solution:

Using the data from the table above:

ΣX = 68 + 70 + 62 + 75 + 66 = 341

ΣY = 150 + 160 + 120 + 190 + 135 = 755

ΣXY = (68 x 150) + (70 x 160) + (62 x 120) + (75 x 190) + (66 x 135) = 68300 (52000)

ΣX^2 = (68^2) + (70^2) + (62^2) + (75^2) + (66^2) = (23349)

ΣY^2 = (150^2) + (160^2) + (120^2) + (190^2) + (135^2) = (116825)

Plugging these values into the formula, we get:

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

r = (5 x 68300 - 341 x 755) / sqrt[(5 x 22045 - 341^2)(5 x 122500 - 755^2)]

r = 0.902

So the correlation coefficient between height and weight is 0.902, which indicates a strong positive relationship between the two variables.

To calculate the coefficient of determination, we square the correlation coefficient:

r^2 = (0.902)^2 = 0.8136

Therefore, the coefficient of determination is 0.8136, or 81.36%. This means that 81.36% of the variation in weight can be explained by the variation in height, and the remaining 18.64% is due to other factors not included in this relationship.

Question # 2:

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

Solution:

Using the given data, we can calculate the correlation coefficient between X and Y:

r = (nΣXY - ΣXΣY) / sqrt[(nΣX^2 - (ΣX)^2)(nΣY^2 - (ΣY)^2)] r = (5 x 342 - 25 x 55) / sqrt[(5 x 129 - 25)(5 x 598 - 55^2)] r = 0.9979

This indicates a strong positive correlation between X and Y.

Next, we can fit a linear regression model to the available data points to predict the value of Y for the missing value of X. We get:

Y = 2.0404X + 0.7273

Substituting the given value of Y = 15, we get:

15 = 2.0404X + 0.7273

Solving for X, we get:

X = (15 - 0.7273) / 2.0404

X = 7.074 (6.99)

Therefore, the missing value of X is approximately 7.074.

Since this value is closer to 6.5 than to 5.5, the answer is D) 6.5.

Question # 3:

Suppose for these two eigen values λ1 = 1.284028 and λ2 = 0.04908323 we have following two eigenvectors v1 = [0.6778736, 0.7351785] and v2 = [-0.7351785, 0.6778736]. compute the percentage of variance (information) accounted for by each component.