National University of Computer and Emerging Sciences

Lahore Campus

Data Science (CS4048)

Date: Sept 24th 2024

Course Instructor(s)

Ms. Eesha tur Razia Babar

Ms. Maimoona Akram

Mr. Muhammad Saif ul Islam

Sessional-I Exam

Total Time (Hrs): 1

Total Marks: 45

Total Questions: 3

Do not write below this line

Attempt all the questions. Summarize your answers into 2-3 sentences.

_____ _____ _____

Roll No Section Student Signature

CLO #1: Extract, clean, and transform data for analysis

Q1: The following data set is used to understand how student performance is influenced by various factors such as their year of birth, gender, ethnicity, parental level of education, weight etc. Assume weights, reading score and writing score are normally distributed. (15 marks)

ID	Year of Birth	Gender	Ethnicity	Parental Level of Education	Weight (kg)	Reading Score (out of 150)	Writing Score (out of 150)
1	2001	M	Asian	Middle School	50	2	-80
2	2004	F	African	High School	75	20	2
3	2003	F	American	Masters	76	57	8
4	2003	M	Hispanic	Middle School		99	9
5	2001	M	Irish			54	82
6	2003		African	Master's	56	97	33
7	2002	M		Master's	91		34
8	NA	F	Irish	High School	56	21	
9	2001	M	African	Master's	49	29	61
10	2002	F	Asian	Master's	64	85	63
11	2001	M	African	Bachelors	92	29	134

1. Categorize each variable in above data student according to following four categories. Ordinal, nominal, ratio, interval

   Id = nominal

   Date of birth = interval

   Gender = nominal

   Ethnicity = nominal

   Education = ordinal

   Weight = ratio

Marks = ratio

2. Name the specific feature encoding method which will be used for each categorical feature in the above data set. (2 marks)

   cat.codes for education
   one hot encoding for ethnicity and gender

   Identify missing values in the dataset and state how you would fill these values (Write any assumptions you have made). (2 marks)

   For categorical variables we will use mode.
   For numerical since it is mentioned that these variables are normally distributed we will use mean values.

3. After filling all missing values, apply IQR method on writing score to identify outliers (show all working to obtain full marks). (5 marks)

   Q1 = 8
   Q2 = 34
   Q3 = 63
   IQR = 55
   LB = -74.5
   UB = 145.5
   Outlier = -80

4. Explain how you would handle these outliers and provide reasoning for your approach. (2 marks)

Obvious approach can be to drop this outlier as it's impossible to have -80 marks. Given that our data set is very small the better approach will be to use some imputation method to impute value for this outlier.

5. What is the difference between forward fill and forward interpolation method for filling missing values? Explain using missing values in Weight feature as an example. (2 marks)

   forward fill involves replacing missing data points with the most recent known value forexample for forward fill method the missing value in the weight column will be replaced by 76. However Pandas interpolate() method is used to fill NaN values in the dataframe using various interpolation techniques to fill the missing values rather than hard-coding the value.

CLO #1: Extract, clean, and transform data for analysis

Q2: Answer the following questions. (5x3 = 15 marks)

Employee

	Name	Position	DeptID
0	Alice	Manager	101
1	Bob	Engineer	102
2	Charlie	Analyst	101

Department

	DeptID	Department	Location
0	101	HR	New York
1	102	Engineering	San Francisco
2	103	Finance	Boston

1. You have employee data from three different departments in separate DataFrames, and all DataFrames have the same structure (Name, Position, DeptID). How would you combine these DataFrames into a single DataFrame to analyze all employees? Which method would you choose: concat, merge, or join? Provide reason and demonstrate you implementation.

Since the structure of the files are same across department, we can use concat function to aggregate all files row-wise.
E.g. pd.concat([df1,df2,df3], axis=0)

2. You want to enrich the employee DataFrame by adding the department name and location from the Department table. Both tables share the DeptID as a common column. Suppose some employees in your Employee table do not have a DeptID assigned. You still want to include all employees in the final table How can you combine the two tables to add this information to the employee data? Which method would you choose: concat, merge, or join? Provide reason and demonstrate you implementation.

We can use merge function with specifying left join to include all the records from the employee table as we have acommon key in between.
pd.merge(employee_df, department_df, on='DeptID', how='left')

3. Compare and contrast melt and pivot_table function. Also provide their use case.

melt() unpivots a DataFrame from wide to long format, converting columns into rows, while pivot_table() reshapes data from long to wide format, aggregating values based on specified indexes.

melt() is useful for transforming datasets for visualization, while pivot_table() is ideal for creating summary tables by aggregating data.

4. What is the advantage of .agg function over groupby?

In agg function we can compare multiple aggregate function like mean, count, max, min, etc together.

5. How would you display the number of employees in each department?
   df_employee.DeptID.value_counts()

***CLO #2: Apply* tools for performing exploratory data analysis and visualization.**

Q3 (a): The following histogram summarizes the ages of the 25 employees who work in the IT Department of Elon's Company. (5 marks)

1. What is the age range with the highest number of employees?

   30-35

2. How many employees are older than 50 years?

   2

3. Based on the histogram, does the number of employees increase or decrease as the age of employees goes from 20 to 50 years?

   *As the age increases from 20 to 50 years, the number of employees **initially increases** (from 25 to 35 years), peaks around the 30-35 range, and then **decreases** as it goes towards the 50-year mark. Therefore, it generally **increases and then decreases**.*

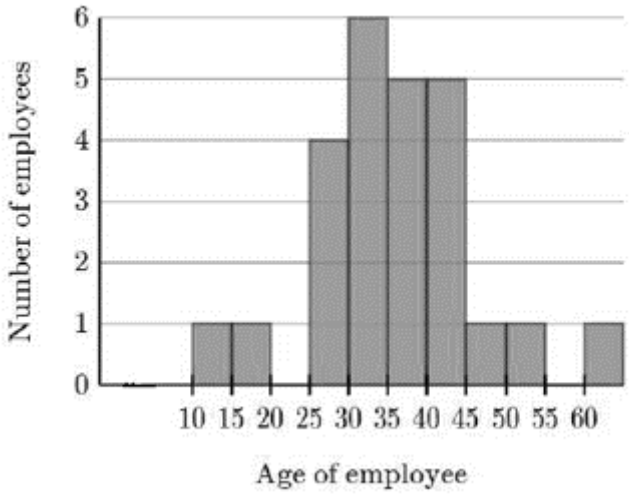
4. What percentage of the employees fall within the age range of 25 to 45 years?

   Calculate the percentage using the number of employees in the 25-45 range out of total employees.

   *(4+6+5+5)/25*100 = 80%*

5. If the average age of all employees who work for the company is 42, the median age of employees is 38, and the mode is 33. Is there any skewness in this data? If yes which one? Justify your answer.

   The data is right-skewed if mean > median > mode. The skewness is right or positive based on the provided average (42), median (38), and mode (33).



Q3 (b): Answer the following questions for the given box plot graph. (4 marks)

1. Which method shows skewness and high variability? Also, mention the skewness type.

   Method 3 *shows both* ***skewness and high variability****.*

   - *The long upper whisker and the distance between the median and the top of the box suggest* ***negative skewness (left-skewed)****.*
   - *The large range between the minimum and maximum values indicates high variability*

2. Which method shows the highest median?

   Method 4 *has the highest median score, as indicated by the position of the thick line (median) in the box, which is higher than the medians of the other methods.*

3. Which is the most effective teaching method among the four compared? Explain with reasoning.

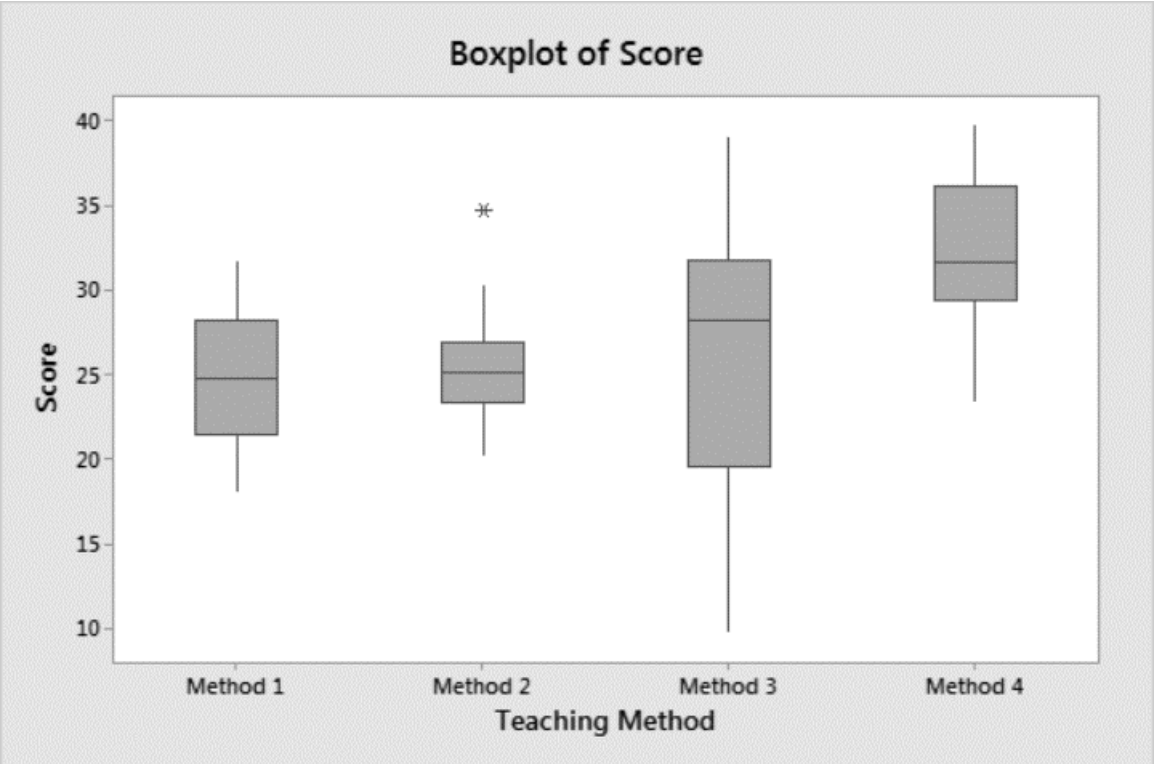
   Method 4 *appears to be the most effective teaching method because:*

   It has the ***highest median score****, meaning most students performed better using this method compared to the others.*

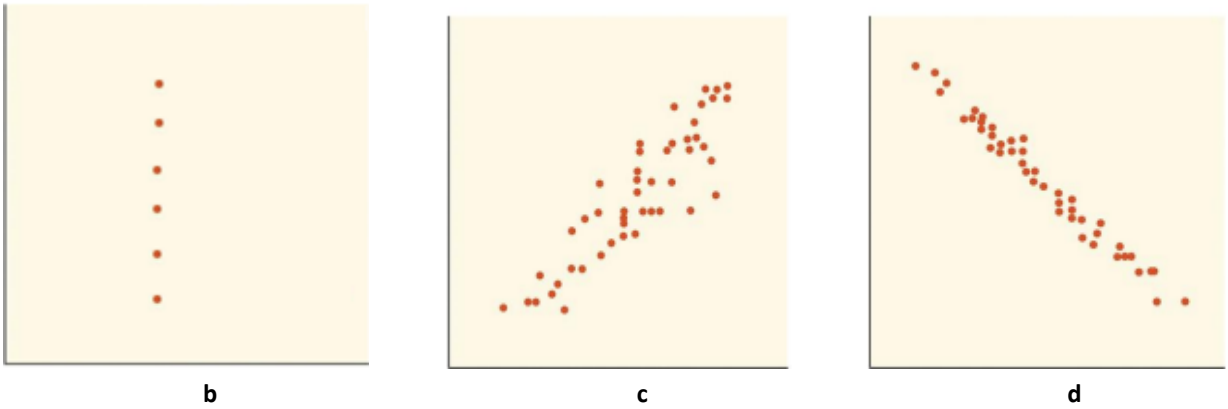
   The variability (spread of the scores) is moderate, suggesting relatively consistent results.

4. What does the size of the box indicate?

   The size of the box represents the ***interquartile range (IQR)****, which measures the variability within the middle 50% of the data. A larger box indicates a greater spread in the middle 50% of scores, and a smaller box indicates that most students' scores were closer together.*



Q3 (c): For each of the scatter graphs below, state whether or not there is a **correlation** and, if so, state the **strength** and **type of correlation**. (6 marks)



b **c** **d**

i. No correlation
ii. strong/moderate positive correlation (both are correct at this point)
iii. Strong negative correlation