National University of Computer and Emerging Sciences

Lahore Campus

Data Science (CS4048)

Sessional-I Exam

Date: Sept 22th 2025

Course Instructor(s)

Ms. Maimoona Akram

Total Time (Hrs): 1

Total Marks: 45

Total Questions: 3

Roll No _____ Section _____ Student Signature _____

Do not write below this line

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

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

Q1: The following dataset (complete) is used to understand how employee performance and compensation are influenced by various factors such as their age, gender, department, educational background, years of experience, and office environment. Assume salary, age, and temperature are normally distributed. **[15 marks]**

Employee ID	Department	Gender	Education Level	Performance Rating	Salary (PKR)	Age	Experience Years	Average Office Temperature (°C)	Payment Method
EMP001	Engineering	Male	Masters	Excellent	78000	28	5	22.1	Direct Deposit
EMP002	Marketing	Female	Bachelors	Good	62000		3	23.4	Check
EMP003		Female	PhD	Outstanding	145000	35	12		Direct Deposit
EMP004	Sales	Male	High School	Fair	25000	22	1	21.8	
EMP005	HR		Bachelors	Good	68000	31	7	24.2	Direct Deposit
EMP006	Engineering	Male	Masters	Excellent	82000	45		22.7	Check
EMP007	Finance	Female	Masters		75000	29	6	23.8	Direct Deposit
EMP008	Marketing	Male	Bachelors	Good	65000	33	8	22.5	Check
EMP009	Sales	Female	Associates	Fair	55000	26	2	23.9	Direct Deposit
EMP010	Engineering	Male	PhD	Outstanding	95000	38	15	21.5	Check

1. Categorize each variable in above data according to following four categories. Ordinal, nominal, ratio, interval (2 marks)
2. Test the assumption of normality for **age**, and **temperature attributes** using appropriate statistical tests (2 marks)
3. Identify missing values in the dataset and state how you would fill these values (Write any assumptions you have made). (2 marks)
4. After filling all missing values, apply IQR method on **Salary** to identify outliers (show all working to obtain full marks). (5 marks)
5. Explain how you would handle these outliers and provide reasoning for your approach. (2 marks)

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

Q2 (a): You need to scrape product details including **name, price,** and **rating** from an e-commerce website that loads products dynamically as you scroll down (infinite scrolling). Some premium products require logging in to see their prices. The website uses JavaScript to load content after the initial page load. **[5 marks]**

- Briefly explain how you would approach scraping this data efficiently, covering:
- Handling the infinite scrolling and JavaScript-rendered content.
- Managing login authentication to access premium product data.
- Any strategies you would use to avoid being blocked or detected.

Q2 (b): Answer the following questions. **[5+5 = 10 marks]**

Student Data

School ID	Name	Type	Location
SCH01	Lincoln High	Public	Downtown
SCH02	Roosevelt Academy	Private	Suburbs
SCH03	Washington Prep	Charter	Midtown

School Data

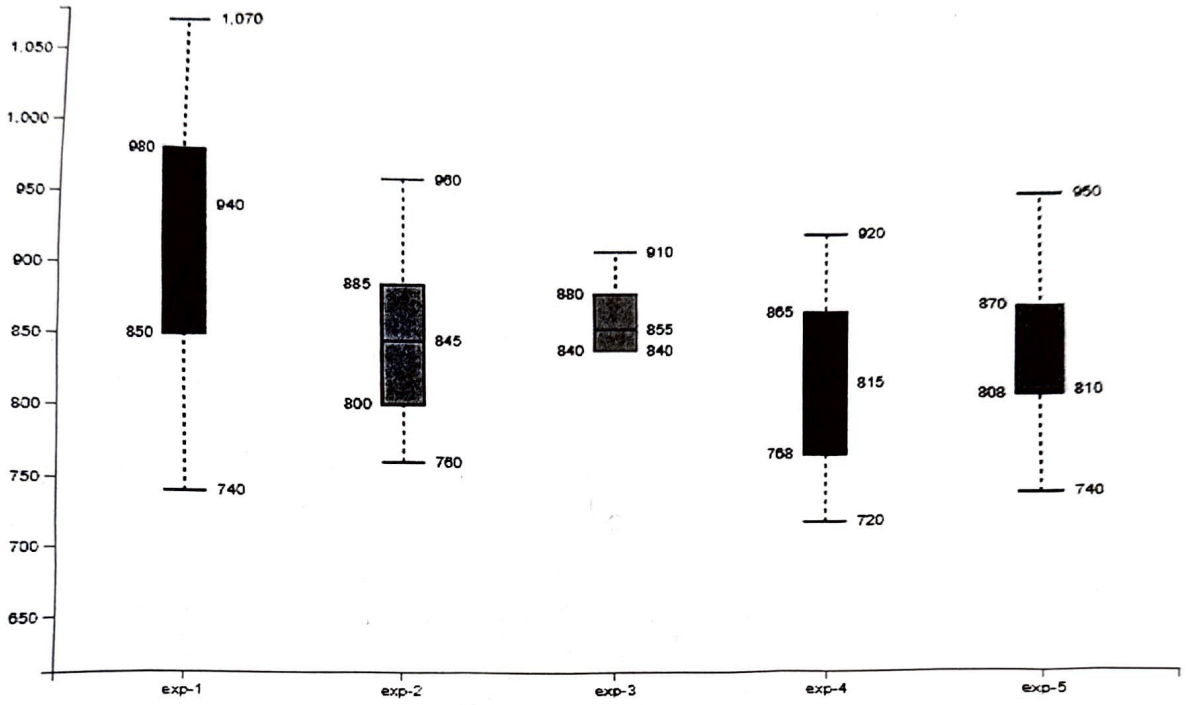
Student ID	Name	Ethnicity	School ID	Parent Education
S001	Alice Johnson	White	SCH01	Bachelors
S002	Sofia Rodriguez	Hispanic		High School
S003	James Wilson	Black	SCH01	PhD
S004	Emma Thompson	White	SCH02	
S005	Aisha Patel	Asian		Masters
S006	Michael Davis	Black	SCH03	High School

1. You are a district analyst who needs to create a comprehensive student report that includes both student information and their school details (school name, type, and location). Notice that some students (S002, S005) haven't been assigned to schools yet, but they must still appear in your final report for enrollment planning.
2. The school administration wants to understand the diversity of their student body across different schools. Using the dataset provided, which includes Student IDs, Names, Ethnicities, School IDs, and Parent Education levels, please create a clear report showing how many students from each ethnicity are enrolled in each school.
   - Your report should be in the form of a matrix or table where:
   - Each row represents a different school (by School ID).
   - Each column shows a different ethnicity.
   - Cells contain the count of students of that ethnicity at that school.

Explain how you would generate the reports for the two scenarios, given above, using data manipulation tools or functions (merge, join or pivot table) and briefly show an example of the report format based on the dataset.

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

Q3: Answer the following questions on the box plot chart shown below: **[Marks: 15]**



- Which experiment(s) show evidence of **skewness** in their data distribution? Explain your reasoning with reference to the position of the median and the whiskers.
- Compare the medians of all experiments. Which experiment has the **highest median value**, and what might this imply about the central tendency?
- For experiments exp-1 through exp-5, analyze the **consistency of their data** based on the size of their **interquartile ranges (IQR)**.
- Based on the box plots, which experiment would you say demonstrates the most **reliable or stable** results? Support your answer with evidence from the graph.
- Which experiment demonstrates the **greatest overall range**?

Note: Summarize your answers into 2-3 sentences.