(c) Load a dataset from a CSV file using Pandas. Perform basic data cleaning by handling any missing values and renaming columns where necessary. Display the final Data Frame. (6)

7. (a) Using Matplotlib, create a histogram for any numerical variable in a dataset of your choice. Label the axes, add a title, and describe what the histogram reveals about the data distribution. (6)

(b) Differentiate between "*Series*" and "*DataFrame*" in Pandas. Can we use a dictionary datatype to create a DataFrame? Explain. (9)

[This question paper contains 8 printed pages.]

Your Roll No.....

Sr. No. of Question Paper : 3508 **I**

Unique Paper Code : 2514001001

Name of the Paper : Data Engineering and Analytics (GE)

Name of the Course : **B.Sc. (H) Electronics**

Semester : I

Duration : 3 Hours Maximum Marks : 90

Instructions for Candidates

1. Write your Roll No. on the top immediately on receipt of this question paper.
2. There are **seven** questions in all, out of which you have to attempt any **five** questions.
3. **All** questions carry equal marks.
4. **First** Question is Compulsory.
5. Use of a non-programable scientific calculator is allowed.

1. (a) Describe the purpose of scipy.stats and provide an example of a statistical function it offers. (3)

(b) Describe the meaning of the term 'p-hacking'. (3)

(c) Define mean, median, and mode, and explain their differences. (3)

(d) Draw the block diagram of ETL pipeline architecture. (3)

(e) What do you understand by the term Data Wrangling? (3)

(f) Explain the purpose of np.linspace() and provide a scenario where it would be useful. (3)

2. (a) In the population, the average IQ is 115 with a standard deviation of 15. A team of scientists wants to test a new medication to see if it has either a positive or negative effect on intelligence, or no effect at all. A sample of 30 participants who have taken the medication has a mean of 140. Did the medication affect intelligence? (6)

Consumer_ID	Name
C001	Ritesh
C002	Sunil
C003	Amit
Data Set-I	

Order_ID	Consumer_ID	Amount
201	C001	3500
202	C002	1500
203	C004	2000
Data Set-II		

(c) Load the following retail data into a DataFrame, then use the groupby() function to group data by Store and calculate the total Revenue for each store. Display the grouped DataFrame. (6)

Store	Product	Revenue
Store1	Shoes	1500
Store1	Bags	1000
Store2	Shoes	2000
Store2	Bags	1200

6. (a) Using NumPy, create a 3x3 matrix with random integers between 1 and 10. Calculate and display the mean, standard deviation, and variance for each row and column. (6)

(b) Describe the five V's of Big Data (Volume, Velocity, Variety, Veracity, and Value). Explain how these traits impact data storage and processing. (6)

P.T.O.

(c) Generate a line plot to visualize a time series of daily temperature data. Assume you have a DataFrame df with a Date and Temperature columns. (6)

(i) Convert the Date column to a DateTime format.

(ii) Plot the Temperature data against Date, setting appropriate labels for the axes and a title "Daily Temperature Over Time".

5. (a) Given the following sales data, use Pandas to calculate the mean, median, and standard deviation for each numerical column and interpret the results. (6)

Product_ID	Sales	Quantity	Price
PR001	1500	10	150
PR002	2000	8	250
PR003	3000	15	200
PR004	4000	5	800

(b) Given the two datasets below, perform an inner join on Customer_ID and display the merged DataFrame. (6)

(b) A company has two departments, Sales and Marketing, with each department hiring male and female employees. The company wants to analyze the success rate (number of successful hires) in each department by gender. (9)

(i) Calculate the success rate of hires for each gender (male and female) in each department separately.

(ii) Calculate the overall success rate of hires for males and females across both departments combined.

(iii) Explain whether Simpson's Paradox is observed in this case, and if so, describe how it affects the interpretation of the data.

(c) Ram is designing a game that involves flipping 3 fair coins and counting how many land showing "heads". Here is the sample space showing the possible outcomes when flipping 3 coins. Make a theoretical probability distribution table of the number of heads? (3)

0 heads	1 head	2 heads	3 heads
TTT	HTT	THH	HHH
	THT	HTH	
	TTH	HHT	

3. (a) A small business tracks the monthly sales (in thousands of dollars) for the past year. The data is as follows : (6)

Sales = [32,35,40,45,50,55,60,65,70,72,78,85]

(i) Determine the five-number summary (minimum, first quartile Q1, median Q2, third quartile Q3, and maximum) for the monthly sales data.

(ii) Construct a box-and-whisker plot based on the five-number summary to visualize the sales distribution over the year.

(b) What does a correlation coefficient tell you? What are the limitations of using correlation to infer causation? (6)

(c) A company has monthly sales data for different regions stored in a CSV file-sales.csv with columns: Region, Month, Sales. Write a Python program to : (6)

(i) Read the CSV file and load the data into a DataFrame.

(ii) Aggregate the total sales for each region across all months.

(iii) Identify the region with the highest total sales and display the result.

4. (a) Given a DataFrame df with columns Department, Employee, and Salary, write Python code to find the average Salary for each Department using the split-apply-combine approach. Also, display the results in descending order of average salary. (6)

(b) What are the key types of data repositories in data engineering, and how do they differ in terms of data storage, accessibility, performance, and use cases within a modern data pipeline? (6)