



Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SLOT: E2+TE2

Session: FN2



VIT

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2026-2027

REG.NO.:

SLOT: E2+TE2

Session: FN2

Client	Training Hours per Week, X	Muscle-Mass Gain (kg), Y
1	2	1
2	3	2
3	4	2
4	5	3
5	6	4
6	7	5
7	8	5
8	9	6

- Calculate the mean training time, $\bar{X}$, and the mean muscle-mass gain, $\bar{Y}$.
(2 marks)
- Compute the Pearson correlation coefficient r , between weekly strength-training hours and muscle-mass gain.
(5 marks)
- Draw a scatter plot with weekly strength-training hours (X) on the horizontal axis and muscle-mass gain (Y) on the vertical axis. Based on the scatter plot and the value of r obtained in part (b), interpret the direction and strength of the linear relationship between weekly strength-training hours and muscle-mass gain.
(3 marks)

3. A retail analytics team is conducting exploratory data analysis on a dataset containing several thousand products. The dataset includes four numerical variables and one categorical variable, as shown below.

Variable	Type	Description
Price	Numerical	Product price
Monthly Sales Volume	Numerical	Number of units sold per month
Customer Rating	Numerical	Average customer rating
Delivery Time	Numerical	Delivery duration
Product Category	Categorical	Category to which the product belongs

Available visualization toolkit: histogram, box plot, bar chart, scatter plot, correlation heatmap, pair plot, parallel coordinates, and Andrews curves.
For each requested analysis, identify the most suitable visualization technique from the toolkit. Briefly justify why the selected technique is more appropriate than one plausible alternative. Where a task explicitly requires more than one complementary visualization, identify and justify each one.

- The team wants to: (i) examine the shape of the distribution of Price, including features such as skewness or multiple peaks; (ii) separately identify potentially extreme Price values using the standard interquartile-range (IQR) convention; and (iii) compare the total Monthly Sales Volume across product categories. Select the most suitable visualization for each objective and justify the choices.
(4 marks)
- The team wants to examine whether Price and Customer Rating are associated and to visually identify individual products that depart substantially from the overall relationship. Select the most suitable visualization and justify why it is more informative for this task than one reasonable alternative.
(3 marks)
- The team wants to: (i) obtain an at-a-glance summary of the strength and direction of the pairwise linear associations among Price, Monthly Sales



SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2026-2027

	Volume, Customer Rating, and Delivery Time using Pearson correlation coefficients, in order to examine redundancy and potential multicollinearity when evaluating candidate predictors for a subsequent model; and (ii) after standardizing these four numerical variables to comparable scales, inspect product-level multivariate profiles across four explicit variable axes and look for products with similar profile shapes. Select the most suitable visualization for each objective and justify the choices. (3 marks)																			
4.	A data science consulting firm has received three time-series datasets from clients in different industries. A junior analyst must apply a common time-series analysis workflow: time indexing → visualization and smoothing → trend analysis → seasonality analysis → decomposition and interpretation and explain how the same analytical techniques can reveal different business insights across domains. Dataset Details: <table border="1"> <thead> <tr> <th>Domain</th> <th>DataFrame</th> <th>Index</th> <th>Variables and Coverage</th> </tr> </thead> <tbody> <tr> <td>Finance</td> <td>stock_df</td> <td>Date</td> <td>Close-daily closing stock price, 2015–2023.</td> </tr> <tr> <td>Healthcare</td> <td>df_health</td> <td>Date</td> <td>Patients-daily patient-visit count, 2019–2023.</td> </tr> <tr> <td>Retail</td> <td>df_retail</td> <td>Date</td> <td>Sales-daily revenue; IsHoliday-Boolean indicator for selected major shopping-holiday dates, such as Black Friday and Christmas, for one calendar year.</td> </tr> </tbody> </table> Assumptions: All three datasets have already been loaded into pandas DataFrames, sorted chronologically, and indexed by Date. For the healthcare dataset, assume that the observations form a regular daily time series with no unresolved missing dates. Using appropriate pandas and time-series analysis functions in Python, answer the following questions. a) For stock_df['Close'], calculate the 50-day moving average and 200-day moving average using an appropriate rolling-window function. Explain the usual interpretation of: • a golden cross, when the 50-day moving average crosses above the 200-day moving average; and • a death cross, when the 50-day moving average crosses below the 200-day moving average. Briefly state the business or investment interpretation of these signals, while recognizing that they do not guarantee future price movements. (4 marks) b) For df_health['Patients'], apply an appropriate time-series decomposition method to obtain the trend, seasonal, and residual components. Assume an annual seasonal period of 365 days. Explain what a rising long-term trend together with a pronounced winter seasonal peak would suggest for hospital capacity and resource planning. (3 marks) c) For df_retail, compare the mean sales on holiday and non-holiday dates and calculate the percentage holiday lift using: $\text{Holiday Lift (\%)} = \frac{S_H - S_N}{S_N} \times 100$ where $\hat{S}_H$ = mean sales on dates for which IsHoliday = True,	Domain	DataFrame	Index	Variables and Coverage	Finance	stock_df	Date	Close-daily closing stock price, 2015–2023.	Healthcare	df_health	Date	Patients-daily patient-visit count, 2019–2023.	Retail	df_retail	Date	Sales-daily revenue; IsHoliday-Boolean indicator for selected major shopping-holiday dates, such as Black Friday and Christmas, for one calendar year.	10	4	3
Domain	DataFrame	Index	Variables and Coverage																	
Finance	stock_df	Date	Close-daily closing stock price, 2015–2023.																	
Healthcare	df_health	Date	Patients-daily patient-visit count, 2019–2023.																	
Retail	df_retail	Date	Sales-daily revenue; IsHoliday-Boolean indicator for selected major shopping-holiday dates, such as Black Friday and Christmas, for one calendar year.																	



VIT

Vellore Institute of Technology
(Deemed to be University under Section 3 of UGC Act, 1956)

REG.NO.:

SLOT: E2+TE2

SCHOOL OF COMPUTER SCIENCE AND ENGINEERING
CONTINUOUS ASSESSMENT TEST - II
FALL SEMESTER 2026-2027

Session: FN2

	and $\bar{S}_N$ = mean sales on dates for which IsHoliday = False. Use an appropriate pandas grouping or aggregation operation. Interpret this value as a descriptive association rather than a causal effect, and briefly state the retail decisions that it could support. (3 marks)																																				
5.	An e-commerce company wants to investigate the relationship between the number of website visits it receives each week and the corresponding number of orders placed. The analysis will help the marketing team quantify the relationship between website traffic and order volume, use a fitted regression model for prediction, and understand the limitations of predictions made outside the observed data range. Data collected over ten weeks are given below: <table border="1"> <thead> <tr> <th>Week</th> <th>Website Visits, X (thousands)</th> <th>Orders Placed, Y (hundreds)</th> </tr> </thead> <tbody> <tr><td>1</td><td>4</td><td>3</td></tr> <tr><td>2</td><td>6</td><td>5</td></tr> <tr><td>3</td><td>7</td><td>4</td></tr> <tr><td>4</td><td>9</td><td>7</td></tr> <tr><td>5</td><td>11</td><td>8</td></tr> <tr><td>6</td><td>13</td><td>10</td></tr> <tr><td>7</td><td>16</td><td>11</td></tr> <tr><td>8</td><td>18</td><td>13</td></tr> <tr><td>9</td><td>21</td><td>14</td></tr> <tr><td>10</td><td>24</td><td>17</td></tr> </tbody> </table> The Head of Marketing wants to use these data not only to construct a predictive model but also to understand the practical interpretation of the fitted relationship and the risks associated with making predictions beyond the observed data range. a) Fit the simple linear regression model: $\hat{Y} = a + bX$ where X represents weekly website visits in thousands and Y represents weekly orders placed in hundreds. Using $n = 10$, compute the required quantities $\sum X$, $\sum Y$, $\sum X^2$, and $\sum XY$. Then determine the slope b and intercept a, showing the major calculation steps. Using the fitted regression equation, estimate the weekly number of orders when website visits are 12 thousand. (5 marks) b) Interpret the slope of the fitted regression model in the context of the business problem. Also interpret the intercept, commenting on whether it has a meaningful practical interpretation given the observed range of X. Then use the fitted model to estimate the weekly number of orders when website visits reach 50 thousand. Discuss the reliability of this prediction with reference to the observed range of X, which is 4 to 24 thousand visits, and explain the limitation associated with extrapolation. (3 marks) c) For the week in which the company received 18 thousand website visits, use the fitted regression equation to calculate the predicted number of orders, $\hat{Y}$. Then compute the residual, in hundreds of orders, using $e = Y - \hat{Y}$ where Y is the actual number of orders observed for that week. State whether the fitted model underpredicts or overpredicts the observed order volume, and briefly explain what the sign of the residual indicates. (2 marks)	Week	Website Visits, X (thousands)	Orders Placed, Y (hundreds)	1	4	3	2	6	5	3	7	4	4	9	7	5	11	8	6	13	10	7	16	11	8	18	13	9	21	14	10	24	17	10	4	4
Week	Website Visits, X (thousands)	Orders Placed, Y (hundreds)																																			
1	4	3																																			
2	6	5																																			
3	7	4																																			
4	9	7																																			
5	11	8																																			
6	13	10																																			
7	16	11																																			
8	18	13																																			
9	21	14																																			
10	24	17																																			
