UTM
UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING
UTM Johor Bahru

Semester II 2024/2025

Subject	:	PROBABILITY & STATISTICAL DATA ANALYSIS (SECI 1143/SCST 1223)
Task	:	ASSIGNMENT 4 - Chapter 7 (60%) & Chapter 8 (40%)
Due Date	:	**JUNE 2025 (before 5 pm)** **1 week after release (please refer to the section's lecturer)**

INSTRUCTION:

1. This is a **GROUP** assignment. Please clearly write the group members **NAME & MATRIC NUMBER** in the front page of the submission.
2. This assignment contributes to 5% of overall course marks.
3. Only **HANDWRITTEN** submission is accepted:
   a. Submissions using any reporting or statistical tools (e.g.: MS Word, MS Excel, etc.,) will be **REJECTED**.
   b. Make sure the submission is neatly written. Any submission with handwriting that is unreadable, will be **REJECTED**.
   c. For answer that need to draw graphs, using graph paper is optional. You can use plain paper.
   d. Round your answers to **THREE** decimal places.
   e. Please scan/snapshot your work and save as a PDF file.
4. Submission via eLearning – only **ONE** group member needs to submit on behalf of the group.

PART 1 CHAPTER 7: CORRELATION AND REGRESSION (60%)

QUESTION 1 (10 MARKS)

A bakery production manager wants to investigate the linear relationship between the number of pastries produced and the production cost. To pursue his/her objective, the manager recorded the data on the number of pastries produced per day and the production cost per day (in thousands of Malaysian Ringgit) for 10 consecutive days as depicted in **Table 1**.

Table 1: Daily pastries produced and production cost for 10 consecutive days

Days	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Number of pastries, x	35	50	45	60	70	55	40	65	75	80
Production cost, y	48	65	60	72	83	62	50	75	90	95

a) Calculate the correlation coefficient, *r*. (8 marks)

b) Based on the correlation coefficient, *r* obtained, make a conclusion on the linear relationship between the number of pastries produced and the production cost.

(2 marks)

$$r = \frac{\sum xy - \frac{(\sum x \sum y)}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

QUESTION 2 (25 MARKS)

A social media analytics company is investigating the relationships between various factors to understand how they influence the engagement on posts. The company has collected data on the following variables (**Table 2**) for 6 different social media posts:

- **Number of Likes**: The number of likes received by the post.
- **Number of Comments**: The number of comments received by the post.
- **Number of Shares**: The number of times the post was shared.
- **Post Length**: The length of the post in characters.
- **Engagement Score**: A score (from 1 to 100) representing the overall engagement on the post.
- **Sentiment Score**: An ordinal variable representing the sentiment of the post (1: Very Negative, 2: Negative, 3: Neutral, 4: Positive, 5: Very Positive).

Table 2: Factors influencing engagement on posts

Post ID	Likes	Comments	Shares	Post Length	Engagement Score	Sentiment Score
1	150	20	30	200	85	4
2	100	10	20	100	70	3
3	200	25	40	250	90	5
4	80	8	15	150	60	2
5	170	22	35	220	88	5
6	120	15	25	180	75	3

a) Compute the correlation coefficient for Engagement and Sentiment Score variables. Interpret the result. (9 marks)

b) Compute the correlation coefficient for Likes and Share. Interpret the result. (9 marks)

c) Test the null hypothesis that there is no correlation between 'Engagement Score' and 'Sentiment Score' against the alternative hypothesis that there is a significant correlation. Use a significance level of 0.05 and provide the test statistics and p-values. (7 marks)

QUESTION 3 (25 MARKS)

A software development company wants to predict the time required to complete new software projects based on the estimated lines of code (LOC). The company has collected data on the development time (in weeks) and LOC for previous projects. Using this data presented in **Table 3,** answer the following questions in 2 decimal places:

Table 3: Length of code (LOC) and development time

Project ID	LOC (K)	Development Time (week)
1	150	40
2	100	35
3	200	50
4	80	30
5	170	45
6	120	38
7	160	42
8	90	28
9	250	55
10	130	37

a) Plot a scatter plot of the data with LOC on the x-axis and Development Time on the y-axis. (3 marks)

b) Calculate the correlation coefficient between LOC and Development Time. (8 marks)

c) Fit a simple linear regression model using LOC as the independent variable and Development Time as the dependent variable. Provide the regression equation and interpret the coefficients. (5 marks)

d) Use the regression model to predict the development time for a new project with an estimated 180K LOC. (2 marks)

e) Find value of SSR, SST and R-Squared. Interpret value of R-Squared. (6 marks)

PART 2 CHAPTER 8: ANOVA (40%)

QUESTION 4 (20 MARKS)

A local agricultural researcher is conducting a study to determine whether different types of fertilizers affect plant height after 30 days. Three fertilizers (A, B, and C) are applied to separate groups of plants. Each group contains 5 plants, and all other growing conditions (light, water, soil) are kept constant. The **Table 4** below shows the height (in cm) of plants after 30 days.

Table 4: Plant Heights after 30 Days (cm)

Fertilizer A	Fertilizer B	Fertilizer C
262	235	223
246	271	223
266	255	233
288	230	201
244	250	204

Conduct the ANOVA test for the above data by;

a) Define the hypothesis statement. (2m)

b) Calculate mean and variance. (3m)

c) Calculate the test statistics. (10m)

d) Calculate numerator and denominator degree of freedom. Use α = 0.05 (2m)

e) State the critical value. (1m)

f) Test the claim and state the conclusion. (2m)

QUESTION 5 (20 MARKS)

A car manufacturer wants to test the effectiveness of four different brands of brake tires. The stopping distances (in meters) under identical conditions were measured using 5 test runs for each tire brand:

Table 4: Plant Heights after 30 Days (cm)

Tire Brand	Stopping Distance (meters)
L	35.5, 34.8, 36.1, 35.2, 34.9
M	38.2, 37.7, 38.5, 37.9, 38.0
N	33.9, 34.2, 33.7, 34.0, 33.5
O	36.8, 37.0, 36.5, 36.9, 37.1

At the 0.05 significance level, conduct the ANOVA test whether the mean stopping distances are the same for all four tire brands.

a) Define the hypothesis statement. (2m)

b) Calculate mean and variance. (3m)

c) Calculate the test statistics. (10m)

d) Calculate numerator and denominator degree of freedom. Use α = 0.05 (2m)

e) State the critical value. (1m)

f) Test the claim and state the conclusion. (2m)