

CW: Analysis of E-Commerce Customer Purchases Using R

CM 1606: Computational Mathematics

2024/25 Semester 03 (Jan Intake)

Contents

1 Scenario	1
2 Assignment Instructions	1
3 Tasks	2
3.1 Exploratory Data Analysis	2
3.2 Probability Analysis	2
3.3 Distribution Fitting	3
3.4 Predictive Modeling	3
4 Marking Criteria	3
4.1 Marking Rubric	3
4.2 Final Grade Calculation Rules	4

1 Scenario

A growing e-commerce company is analyzing customer purchase behavior to optimize their marketing strategies and website design. The company has provided a dataset of 2000 customer transactions with the following characteristics:

- Customer demographics (region)
- Behavioral metrics (time spent on site)
- Transaction data (purchase amounts)
- Historical data (previous purchases)
- Promotion data (discount usage)

2 Assignment Instructions

- The coursework deadline is **3rd Aug 2025, 11:59 p.m. (IST)**
- Dataset attached separately.
- You can complete the assignment in a single R Markdown file using RStudio and submit it as a PDF document.
- Between R chunks, you can use LaTeX to write text or mathematical equations.
- Word Limit is 800 words (**Only for the interpretations**)
- **Plagiarism** (except for code) should be **less than 15%**.
- Submit Via Moodle at least 2 hours before deadline.

- No late submissions accepted without prior approval.
- If you have any questions, contact a module team member before the due date.
- There will be an individual viva after the submission.

3 Tasks

3.1 Exploratory Data Analysis

3.1.1 Required Analysis

- Generate summary statistics for all numerical variables
- Create the following visualizations:
 - Histogram of purchase_amount with density overlay
 - Boxplot of time_spent_on_site by region
 - Scatterplot of purchase_amount vs time_spent_on_site
- Identify and handle any missing values
- Detect and comment on outliers

3.1.2 Interpretation Guidelines

- Describe the distribution of key variables
- Note any unusual patterns or data quality issues

3.2 Probability Analysis

3.2.1 Required Calculations

- Calculate basic probabilities:
 - $P(\text{Purchase} > \$75)$
 - $P(\text{Used Discount} | \text{Purchase} > \$100)$
- Create contingency tables:
 - Region vs Used Discount
 - Previous Purchases vs Discount Usage
- Calculate conditional probabilities by region

3.2.2 Interpretation Guidelines

- Explain what probabilities reveal about customer behavior
- Compare probabilities across different segments
- Suggest how marketing could target specific groups
- Discuss limitations of the probability analysis

3.3 Distribution Fitting

3.3.1 Required Analysis

- a. Fit distributions to:
 - Poisson for number_of_previous_purchases
 - Normal for purchase_amount

3.3.2 Interpretation Guidelines

- Interpret the Fitted distributions

3.4 Predictive Modeling

Focus on the relationship between `purchase_amount` and `time_spent`. Then build a linear regression model:

- a. Comment on the type of association between `time_spent` (X -axis) and `purchase_amount` (Y -axis)
- b. Obtain the output and the fitted values, and show the regression line over the data
- c. Write down the fitted equation of the form $Y = \beta_0 + \beta_1 X$
- d. Predict the `purchase_amount` for a `time_spent` of 12 minutes

4 Marking Criteria

4.1 Marking Rubric

- All the sections of the rubric will be evaluated through individual viva assessments.

Criterion	A	B	C -	D	E	F / NS
Exploratory Data Analysis	Year and comprehensive EDA with meaningful insights, well-handled missing values and outliers	Mostly complete EDA, minor gaps in interpretation or missing value handling	Basic summaries and visuals, limited insight	Minimal EDA or unclear interpretations	Weak EDA, major issues with clarity or accuracy	No EDA presented
Probability Analysis	Accurate probability calculations with insightful interpretation and proper use of conditional probabilities	Mostly correct calculations with some interpretation gaps	Basic probability work with some interpretation	Incomplete or partially incorrect probability explanation	Incorrect or unclear work section	Not attempted or missing

CriterionA	B	C -	D	E	F / NS
Distribution Fitting	Appropriate choice of distributions, correct tests, strong visual comparisons and clear interpretation	Mostly correct distribution fitting and some interpretation issues	Attempted distribution fitting, but lacks depth or accuracy	Incomplete or poorly justified distribution fitting	Incorrect methods or visuals
Predictive Modeling	Accurate model building, diagnostics clearly presented, strong interpretations and business recommendations	Good model with minor issues in diagnostics or interpretation	Working model with basic diagnostics and conclusions	Weak model with missing checks or unclear explanations	Incorrect or incomplete model

4.2 Final Grade Calculation Rules

Final Grade	Condition
A	At least 2 A's , and no grade below C
B	At least 2 B's or higher , and no grade below C
C	At least 3 grades of C or higher , and no F or NS
D	At least 3 grades of D or higher , and no F or NS
E	At least 3 grades of E or higher
F / NS	Otherwise

End of the CW – Good Luck!

Prashan Rathnayaka, SoC, IIT-Colombo