Week 3 Assignment: Regression Analysis for Decision-Making

Course: Foundations of Data Analysis
Due Date: Sunday, February 15, 2026 | 11:59 PM (EAT)
Submission: ONE ZIP file per group named
`GroupX_Week3_Regression.zip`

Assignment Purpose

This assignment moves you from *running regression* to *using regression for real-world decisions*. You will build:

- An **OLS model** to inform credit allocation decisions
- A **Logistic regression model** to support clinical risk screening
- A **one-page executive brief** translating statistical findings into stakeholder action

The focus is not only statistical correctness — but **decision relevance in Kenyan contexts**.

Learning Outcomes

By the end of this assignment, you should be able to:

✓ Build and interpret an OLS regression model for a continuous outcome
✓ Implement logistic regression for binary classification
✓ Diagnose regression assumptions using residual analysis
✓ Interpret coefficients in business and clinical language
✓ Apply responsible modeling practices in healthcare and finance

Datasets & Variable Rules (STRICT)

You MUST use only the variables listed below.

Dataset	Outcome Variable	Allowed Predictors (Continuous ONLY)	Context
german_credit_data.csv	`Credit amount`	`Age`, `Duration`	Credit risk & microfinance decision-making
Heart.csv	`HD` (0 = No, 1 = Yes)	`Age`, `RestBP`, `Chol`, `MaxHR`, `Oldpeak`	Cardiovascular risk screening

Mandatory Constraint

You may ONLY use the continuous predictors listed above.

❌ Do NOT include categorical variables (e.g., Sex, ChestPain, Job, Housing).
❌ Do NOT engineer additional features.
❌ Do NOT remove outliers unless explicitly justified.

This restriction simulates real-world limited-data environments such as:

- Rapid digital loan approvals
- Rural clinical screening stations

Violation of this rule results in automatic deduction under **Statistical Rigor**.

PART A: OLS Regression - German Credit Data (40 Marks)

Notebook Name:
`GroupX_OLS_Analysis.ipynb`

1. Data Preparation (10 Marks)

- Load dataset
- Retain only: `Age`, `Duration`, `Credit amount`
- Check missing values
- Generate:
  - `df.describe()`
  - Correlation matrix
  - Scatter plots:
    - Age vs Credit amount
    - Duration vs Credit amount

2. Model Estimation & Diagnostics (20 Marks)

Required Outputs:

- Coefficients
- p-values
- R^2

Diagnostic Tests:

- Residuals vs Fitted Plot
- Q-Q Plot
- Breusch-Pagan test (heteroscedasticity)
- VIF (multicollinearity)

All plots must be exported and later combined in the required diagnostic image.

3. Business Interpretation (10 Marks)

For each question:

- Provide code
- Provide a one-sentence business interpretation

a) How does a **10-month increase in Duration** affect expected credit amount, holding Age constant?

b) Which predictor has stronger explanatory power?
Use standardized coefficients to justify.

PART B: Logistic Regression-Heart Disease (45 Marks)

Notebook Name:
`GroupX_Logistic_Analysis.ipynb`

1. Data Preparation (10 Marks)

- Retain only continuous predictors + `HD`
- Report class distribution (% with HD = 1)
- Plot KDE distributions by HD status for:
  - Age
  - RestBP
  - Chol
  - MaxHR
  - Oldpeak

Use `seaborn.kdeplot().`

2. Model Estimation & Evaluation (25 Marks)

Required Outputs:

- Coefficients
- Odds Ratios (exp(coef))
- p-values
- Model accuracy

Performance Evaluation:

- Confusion matrix (threshold = 0.5)
- ROC curve
- AUC value
- Optimal threshold using **Youden's J statistic**
- Sensitivity and specificity at optimal threshold

3. Clinical Interpretation (10 Marks)

Provide code + one-sentence explanation for each:

a) Which physiological variable has the strongest association with HD?
Interpret its odds ratio clinically.

b) Predict probability of HD for:

- Age = 55
- Oldpeak = 2.5
  (Use mean values for other predictors)

c) In a rural Kenyan clinic with limited ECG access, would you prioritize measuring **Oldpeak or MaxHR**?
Justify using both:

- Statistical results
- Practical feasibility

Align answers with:

- MOH Hypertension Guidelines
- UHC preventive screening priorities

PART C: Executive Summary (15 Marks)

File Name:
`GroupX_Week3_Summary.pdf`
(ONE page maximum)

Required Structure

Section	Requirement
Group Details	Group number + all 10 members
Key Finding 1	One-sentence OLS insight + Kenyan financial implication
Key Finding 2	One-sentence Logistic insight + Kenyan clinical implication
Model Limitation	One limitation of continuous-only modeling
Ethical Risk	One risk of deploying these models in Kenya

Final ZIP Structure

Your ZIP file MUST contain:

- `GroupX_OLS_Analysis.ipynb`
- `GroupX_Logistic_Analysis.ipynb`
- `GroupX_Week3_Summary.pdf`

Correct naming format:

```
Group3_Week3_Regression.zip
```

Grading Rubric (100 Marks)

Component	Marks
OLS Implementation	25
Logistic Implementation	30
Statistical Rigor	15
Executive Summary	15
Code Quality & Reproducibility	10
Formatting & Structure	5
TOTAL	**100**

Ethical & Professional Reminder

A statistically significant model can still:

- Discriminate unfairly (e.g., age bias in lending)
- Misclassify vulnerable patients
- Reinforce structural inequalities

You are not just building models.
You are influencing decisions.

Final Reminder

“A regression coefficient becomes powerful only when it changes a real-world decision.