Assignment 1: Object-Oriented Data Cleaning and Preprocessing

Deadline: Friday, May 30th 2025

Part 4: Short-Answer Questions (Upload to Blackboard)

1) Please provide the link to your public **GitHub Repository**.

2) Provide key summary statistics for the messy dataset and the cleaned dataset. Discuss any notable changes in the dataset after preprocessing.

3) How many rows/columns were removed due to missing data? Why?

4) How many features were removed due to redundancy?

5) How did the preprocessing steps affect the logistic regression model's performance? Be quantitative! On the original dataset, prediction accuracy is approximately 85%.

6) **Critical Thinking (BONUS)**
   a) Could preprocessing steps (e.g., imputation or redundancy removal) introduce bias? How can this be mitigated?

   b) What improvements or additional preprocessing steps would you recommend?