

Statistical Modeling and Inferencing

Assignment-1

TOTAL MARKS: 20**DUE DATE: 3rd December 2025, 11:59 PM****Assignment Weightage: 20% of Final Grade**

Submission Format

Submission: Submit a Jupyter Notebook (or Google Colab link with "Anyone with the link can view" access enabled) containing your complete analysis. Use markdown cells to document your findings, interpretations, and insights throughout the notebook.

File Naming: YourRollNumber_SMI_Assignment1.ipynb

Dataset Selection

Choose **one** of the following datasets for your analysis:

Dataset 1: California Housing Dataset

- **URL:** [housing.csv](#) (~20,600 observations)

Dataset 2: Wholesale Customers Dataset

- **URL:** [Wholesale customers data.csv](#) (~440 observations)

Assignment Components

Part 1: Data Exploration and Preparation

(6 marks)

Load your dataset and present descriptive statistics with appropriate visualizations exploring distributions and relationships. Conduct a correlation analysis to understand how variables relate to each other and your target variable. Identify any data quality issues such as missing values and outliers, documenting your decisions on how to handle them. Encode categorical variables if present and describe any feature transformations, providing justification for your preprocessing choices.

Part 2: Model Development and Validation

(10 marks)

Apply appropriate analytical techniques based on your chosen dataset and validate your approach systematically.

For Dataset 1: Investigate relationships through regression modeling. Examine correlations among predictors and consider dimensionality reduction approaches if needed. Develop multiple models with different predictor combinations or transformations. Compare models using appropriate statistical metrics. Address multicollinearity concerns and identify influential features. Select your best model with clear justification.

For Dataset 2: Perform clustering analysis to identify natural customer groupings. Determine the optimal number of clusters. Visualize patterns using dimensionality reduction methods where

appropriate. Evaluate the quality of your chosen clustering solution. Document your analytical decisions and present key outputs clearly.

Part 3: Interpretation and Insights

(4 marks)

Interpret your key findings clearly and provide actionable insights. What patterns emerged from your analysis? Discuss the practical implications of your results, acknowledge limitations of your analysis, and provide recommendations based on your findings. Consider what additional data or analysis might strengthen your conclusions.

Document Organization

Your submission should follow this structure:

1. **Title Page:** Name, roll number, chosen dataset, and submission date
2. **Google Colab Link:** At the top (or note about Jupyter Notebook submission)
3. **Summary:** Brief overview of your approach and key findings (2-3 paragraphs)
4. **Part 1:** Data exploration and preparation with visualizations and documentation
5. **Part 2:** Model development, comparison, validation, and diagnostics
6. **Part 3:** Interpretation, insights, limitations, and recommendations

Timeline

Assignment Released: 23rd November 2025

Submission Deadline: 3rd December 2025, 11:59 PM

Questions? Post on the course discussion forum.