Project 10: Energy Efficiency and Building Performance

Statistical Programming with R

1. Project Overview

The objective of this project is to analyze the energy performance of buildings using simulated design characteristics. Students will investigate how geometric and architectural features of buildings affect energy demand, focusing on heating and cooling loads.

The project emphasizes:

- data cleaning and preprocessing,
- exploratory data analysis (EDA),
- multivariate response analysis,
- distributional analysis and probability,
- hypothesis testing,
- supervised and unsupervised learning,
- interpretation of results in an energy and sustainability context.

Important: The dataset is purely *cross-sectional*. No time series, spatial, or simulation-based modeling is required or expected.

2. Dataset Description

The dataset contains simulated energy performance data for residential buildings with different design configurations. Each row corresponds to a building shape generated through simulation.

- Unit of observation: building
- Data structure: cross-sectional
- Number of observations: 768

- Outcomes of interest: heating load and cooling load

The simulations vary building geometry, glazing characteristics, orientation, and height in order to study their impact on energy consumption.

3. Variable Description

The dataset includes the following variables:

- **Design characteristics:**
  - `X1`: relative compactness
  - `X2`: surface area
  - `X3`: wall area
  - `X4`: roof area
  - `X5`: overall height
  - `X6`: building orientation
  - `X7`: glazing area
  - `X8`: glazing area distribution
- **Response variables:**
  - `Y1`: heating load
  - `Y2`: cooling load

Students must discuss the physical and economic interpretation of these variables and justify any transformations or recoding decisions.

4. General Project Guidelines

- The project must be completed **individually**.
- A written report of approximately **20–30 pages** is required.
- All analyses must be conducted using `R`.
- Complete and well-documented `R` code must be submitted.
- The raw dataset must **not** be modified manually.
- All preprocessing steps must be implemented programmatically.

Reproducibility requirements:

- Use `set.seed()` whenever randomness is involved.
- Clearly comment all code.
- Ensure that the analysis runs from start to finish without errors.

5. Suggested Analysis Tasks

The tasks below are intended as guidance. Students are expected to justify all methodological choices.

5.1 Data Cleaning and Preprocessing

- Verify variable types and ranges.
- Examine correlations among design variables.
- Consider appropriate scaling of numerical predictors.

5.2 Exploratory Data Analysis

- Explore the distribution of heating and cooling loads.
- Visualize relationships between building features and energy demand.
- Examine trade-offs between heating and cooling requirements.

5.3 Distributional Analysis and Probability

- Study the joint distribution of heating and cooling loads.
- Discuss variability and dependence between the two outcomes.
- Formulate probability-based questions relevant to energy efficiency thresholds.

5.4 Hypothesis Testing

- Formulate hypotheses about the effect of design characteristics on energy demand.
- Choose and justify appropriate statistical tests.
- Interpret results in terms of energy efficiency and building design.

5.5 Supervised Learning

- Define a suitable prediction task using one or both response variables.
- Identify appropriate models discussed in the course.
- Fit and compare **multiple competing models**.
- Interpret results and justify the final model choice.

5.6 Unsupervised Learning

- Apply unsupervised methods to group buildings with similar energy profiles.
- Justify variable selection and preprocessing.
- Interpret clusters in terms of building design and energy performance.

Optional extension: Students may discretize the energy load variables to define energy-efficiency categories, provided the discretization is clearly justified.

6. Report Structure

A suggested report structure is:

1. Introduction and motivation
2. Description of the data
3. Data preprocessing
4. Exploratory analysis
5. Statistical inference
6. Supervised and unsupervised modeling
7. Discussion and limitations
8. Conclusion

7. Presentation

Each student will give a **12–15 minute presentation** summarizing:

- research questions and motivation,
- key empirical findings,
- methodological choices and model comparisons,
- interpretation of results in an energy efficiency context.

This will be followed by questions.