## 1. Introduction

The purpose of this project is to perform a comprehensive regression analysis to assess the heating and cooling load requirements of buildings as a function of various building parameters. This project aims to understand the relationships between predictors and the target variables, build reliable regression models, and evaluate model performance using appropriate metrics.

### 1.1 Objectives and Research Questions

Objectives:  
- Understand how building parameters influence heating and cooling loads.  
- Develop and evaluate regression models to accurately predict heating and cooling loads.  
- Provide actionable insights for energy-efficient building design.

Research Questions:  
- What are the most significant predictors of heating and cooling loads?  
- How do different regression models (e.g., linear, polynomial, regularized) perform?  
- Can the results inform strategies for optimizing building energy efficiency?

## 2. Dataset Description

The dataset used in this project is the UCI Energy Efficiency Dataset, sourced from the UCI Machine Learning Repository. It contains 768 observations and 10 variables, including 8 predictors and 2 target variables.

Source: https://archive.ics.uci.edu/ml/datasets/Energy+efficiency

Variables:  
- X1: Relative Compactness (continuous)  
- X2: Surface Area (continuous)  
- X3: Wall Area (continuous)  
- X4: Roof Area (continuous)  
- X5: Overall Height (continuous)  
- X6: Orientation (categorical)  
- X7: Glazing Area (continuous)  
- X8: Glazing Area Distribution (categorical)  
- Y1: Heating Load (target, continuous)  
- Y2: Cooling Load (target, continuous)

## 3. Data Pre-Processing

Data pre-processing involved handling missing values, removing duplicates, transforming features, and scaling predictors for model compatibility. Key steps include:  
- Imputation for missing values (if any).  
- Removal of duplicate records.  
- Feature scaling using standardization.  
- Encoding categorical variables using one-hot encoding.

## 4. Exploratory Data Analysis (EDA)

### 4.1 Univariate Analysis

Summary statistics and visualizations such as histograms and boxplots were generated to understand distributions and identify outliers.

### 4.2 Bivariate Analysis

Scatter plots and correlation matrices were used to examine relationships between predictors and targets. Pair plots provided insights into interactions between variables.

### 4.3 Multivariate Analysis

Multivariate relationships were explored using heatmaps and dimensionality reduction techniques like Principal Component Analysis (PCA).

## 5. Regression Analysis

Several regression models were implemented, including:  
- Simple Linear Regression  
- Multiple Linear Regression  
- Polynomial Regression  
- Regularized Regression (LASSO, Ridge, and Elastic Net)  
- Quantile Regression

Each model was evaluated using metrics such as RMSE, MAE, and R². Quantile Regression provided insights into conditional quantiles of the target variables.

## 6. Model Evaluation and Comparison

The models were evaluated using RMSE, MAE, and R². Performance comparison revealed:  
- Polynomial Regression and Ridge Regression achieved the best RMSE and R².  
- LASSO effectively performed feature selection.  
- Quantile Regression offered robust insights into varying predictor effects across quantiles.

## 7. References

1. UCI Machine Learning Repository. (2012). Energy Efficiency Data Set. Retrieved from https://archive.ics.uci.edu/ml/datasets/Energy+efficiency

2. STAT 650 Course Notes by Dr. Yoonsung Jung. Retrieved from https://canvas.tamu.edu/

3. Pandas Documentation. Retrieved from https://pandas.pydata.org/docs/

4. Seaborn Documentation. Retrieved from https://seaborn.pydata.org/

5. Matplotlib Documentation. Retrieved from https://matplotlib.org/

6. Scikit-learn Documentation. Retrieved from https://scikit-learn.org/stable/

7. Statsmodels Documentation. Retrieved from https://www.statsmodels.org/stable/