Project 10 – Energy Efficiency and Building Performance

Statistical Programming with R

Jaures Kuete Tchinda

February 2026

1 Introduction and Motivation

Energy consumption in the building sector represents a substantial share of total energy demand in industrialized economies and plays a central role in discussions on sustainability, climate change mitigation, and cost-efficient resource allocation. Residential and commercial buildings account for a large fraction of final energy use, primarily driven by heating and cooling requirements. Understanding how architectural and geometric design choices influence energy demand is therefore of considerable practical and economic importance.

From an analytical perspective, building energy performance is shaped by a combination of physical characteristics—such as compactness, surface exposure, height, and glazing properties—as well as orientation-related effects that influence solar gains and heat losses. While physical simulation models are commonly used in engineering applications, statistical analysis of simulated design data provides a complementary perspective. Such an approach allows researchers to explore systematic patterns, quantify uncertainty, and compare alternative modeling strategies without relying on strong structural assumptions.

In recent years, statistical learning methods have increasingly been applied to energy and sustainability research. Beyond classical regression-based approaches, supervised learning techniques enable the comparison of predictive models with different levels of flexibility, while unsupervised learning methods allow the identification of latent building typologies based solely on design characteristics. These tools are particularly well suited for exploratory analysis and prediction-focused tasks, where the primary objective is not causal inference but understanding patterns, dependencies, and predictive performance.

The present project provides an applied statistical analysis of simulated residential building data, with the aim of examining how building design characteristics are related to energy demand. The analysis follows a structured workflow that includes data preprocessing, exploratory data analysis, distributional and probabilistic reasoning, hypothesis testing, and both supervised and unsupervised learning. Initially, both heating and cooling loads are treated as outcomes of interest. Based on distributional properties and dependence patterns, subsequent inference and modeling focus on heating load as the primary response variable. Throughout the analysis, emphasis is placed on methodological transparency, reproducibility, and the comparison of alternative modeling approaches.

2 Description of the Data

2.1 Data Source and Structure

The empirical analysis is based on the Energy Efficiency dataset (ENB2012), which contains simulated energy performance data for residential buildings with varying architectural and geometric configurations. Each observation corresponds to a distinct building design generated through simulation, ensuring a purely cross-sectional data structure without temporal or spatial dependencies.

The final dataset consists of 768 observations, where the unit of observation is a single building. As the data are simulation-based, no sampling error in the conventional survey sense is present; however, substantial variability across building designs allows for meaningful statistical analysis and comparison of design alternatives.

2.2 Outcome Variables

The dataset includes two continuous response variables that measure energy demand:

- **Heating Load (Y1)**: the amount of energy required to maintain indoor thermal comfort during colder conditions
- **Cooling Load (Y2)**: the energy required to cool the building during warmer conditions.

Both variables are measured on a continuous scale and reflect distinct but related aspects of building energy performance. In the initial stages of the analysis, heating and cooling loads are treated symmetrically. Their joint distribution and dependence structure are examined in detail before a primary focus on heating load is adopted for subsequent statistical inference and modeling.

2.3 Design Variables

Energy demand is modeled as a function of several building design characteristics:

- **Relative Compactness (X1)**: a dimensionless measure capturing how compact the building shape is relative to a reference structure.
- **Surface Area (X2)**: total external surface area of the building envelope.
- **Wall Area (X3)**: area of walls.
- **Roof Area (X4)**: area of the roof
- **Overall Height (X5)**: total building height, reflecting the number of floors.
- **Orientation (X6)**: categorical variable indicating the building's orientation.
- **Glazing Area (X7)**: proportion of the building façade covered by windows.
- **Glazing Area Distribution (X8)**: categorical variable describing how glazing is distributed across façades.

These variables represent core architectural and geometric dimensions that directly influence heat transfer, solar exposure, and thermal efficiency. While several of the variables are highly correlated by construction (e.g., surface area, wall area, and roof area), each captures a distinct physical aspect of building design.

3 Data Preprocessing

Prior to formal statistical analysis, the dataset was subjected to a set of preprocessing steps aimed at improving clarity, interpretability, and numerical stability. All preprocessing decisions described in this section are implemented explicitly in the analysis code and are applied uniformly across subsequent analytical steps.

3.1 Variable Renaming and Data Types

To improve readability and reduce ambiguity, all variables were renamed using descriptive labels reflecting their physical interpretation. The original design variables were mapped to intuitive names such as *Relative Compactness*, *Surface Area*, *Wall Area*, *Roof Area*, and *Overall Height*, while the outcome variables were labeled *Heating Load* and *Cooling Load*.

Categorical design variables—**Orientation** and **Glazing Area Distribution**—were converted into factor variables with meaningful labels. This ensures correct handling in descriptive statistics, graphical analysis, and inferential procedures, and avoids imposing an artificial numerical ordering on nominal categories.

All remaining variables are treated as numeric.

3.2 Missing Values and Plausibility Checks

The dataset contains **no missing values**, eliminating the need for imputation or case-wise deletion. Summary statistics and range checks confirm that all variables lie within plausible bounds consistent with their physical interpretation.

Given the simulation-based nature of the data, measurement error in the conventional sense is absent; however, substantial variation across building designs remains, providing a rich basis for statistical analysis.

3.3 Multicollinearity Assessment and Variable Selection

As part of preprocessing, pairwise correlations among the numerical design variables were examined using a correlation matrix. This step serves to identify potential multicollinearity issues that may affect both statistical inference and predictive modeling.

The correlation analysis reveals a **very strong linear relationship between Roof Area and Overall Height**, with an absolute Pearson correlation coefficient exceeding $|\rho| > 0.95$. Such a high correlation indicates that the two variables convey largely redundant information about building geometry. From a statistical perspective, including both variables in regression-based models may lead to unstable coefficient estimates and reduced interpretability.

Formally, when two predictors X_1 and X_2 satisfy

$$|\text{corr}(X_1, X_2)| \approx 1,$$

the design matrix becomes nearly collinear, which can inflate variance estimates in ordinary least squares regression.

To address this issue, a variable selection decision is required. Between the two highly correlated variables, **Overall Height** is retained due to its clearer physical interpretation and direct relevance for architectural design decisions. **Roof Area**, by contrast, is largely determined by building height and footprint geometry and therefore adds limited independent information once height is included.

Consequently, **Roof Area is removed from subsequent analyses**. This decision is made at the preprocessing stage and is applied consistently across all later parts of the analysis.

4 Exploratory Data Analysis

At this stage of the analysis, both Heating Load and Cooling Load are treated symmetrically as outcomes of interest.

4.1 Distribution of Heating and Cooling Loads

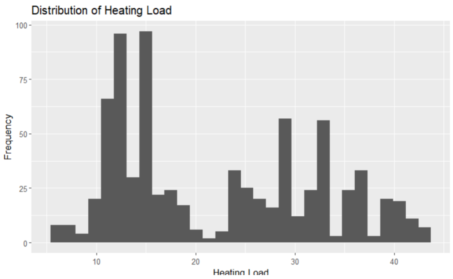


Figure 1: Distribution Heating Load

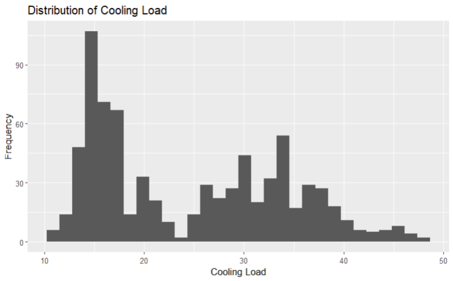


Figure 2: Distribution Cooling Load

Exploratory inspection of the empirical distributions of **Heating Load** and **Cooling Load** reveals that neither variable follows a simple unimodal Normal distribution. While both variables are continuous and exhibit smooth density shapes, kernel density estimates indicate the presence of **two distinct modes** in each distribution. This bimodal structure is visible in both histogram-based and density-based visualizations.

The observed bimodality suggests that the energy demand of buildings is not generated by a single homogeneous process. Instead, the data appear to reflect **two latent groups of building designs** with systematically different energy characteristics. These groups may correspond, for example, to compact versus non-compact structures, or to buildings with low versus high envelope exposure.

From a probabilistic perspective, this pattern can be interpreted as arising from a **mixture of two Normal distributions**. Let Y denote either Heating Load or Cooling Load. The marginal distribution of Y can be expressed as

$$f_Y(y) = \pi_1 \mathcal{N}\big(y \mid \mu_1, \sigma_1^2\big) + \pi_2 \mathcal{N}\big(y \mid \mu_2, \sigma_2^2\big), \quad \text{with } \pi_1 + \pi_2 = 1,$$

where π_k denotes the mixing weight of component k, and μ_k and σ_k^2 represent the component-specific means and variances. Although each component distribution is Gaussian, their weighted combination produces a bimodal marginal density.

4.2 Relationships Between Building Features and Energy Demand

Exploratory visualizations reveal several non-intuitive patterns. Higher Relative Compactness is associated with higher Heating and Cooling Loads. Surface Area shows a negative association with energy demand, while Wall Area exhibits a positive association.

Uniform Glazing Area Distribution is associated with substantially lower energy demand than non-uniform distributions.

4.3 Relationships Between Building Features and Energy Demand

To explore how building design characteristics relate to energy demand, scatterplots and boxplots are used to visualize associations between individual design variables and both Heating Load and Cooling Load. These visualizations reveal several systematic patterns.

4.3.1 Relative Compactness

Contrary to a naive physical expectation, exploratory scatterplots reveal that higher Relative Compactness is associated with higher Heating Load and higher Cooling Load. Buildings with lower compactness values tend to exhibit lower energy demand, while more compact designs cluster at higher levels of both heating and cooling energy.

The relationship appears approximately monotonic but displays substantial dispersion, indicating that compactness alone does not determine energy demand. This pattern suggests that Relative Compactness, as defined in the dataset, captures structural characteristics that are correlated with other design features—such as surface exposure, height, or glazing configuration—which jointly influence energy performance.

4.3.2 Surface Area and Wall Area

Exploratory scatterplots reveal distinct and opposing relationships between different surface-related design variables and energy demand.

Surface Area exhibits a negative association with both Heating Load and Cooling Load. Buildings with larger total surface area tend, on average, to display lower energy demand, while buildings with smaller surface area are more frequently associated with higher heating and cooling requirements. Although the relationship is not perfectly linear and substantial dispersion remains, the overall trend is clearly downward.

In contrast, Wall Area shows a positive association with energy demand. Buildings with larger wall areas tend to exhibit higher Heating Load and Cooling Load. This pattern is more pronounced for Heating Load and suggests that wall exposure plays a significant role in determining heat losses and overall energy requirements. The opposing signs of these relationships highlight that different geometric measures capture fundamentally different aspects of building design. While Surface Area reflects overall envelope size, Wall Area isolates vertical exposure, which appears to be more directly related to thermal losses in the simulated building configurations. These findings further illustrate that energy demand cannot be inferred from a single geometric measure and underscore the importance of multivariate analysis in subsequent modeling stages.

4.3.3 Overall Height

reveals distinct group-level differences. Visual comparisons indicate that taller buildings tend to display systematically different energy demand profiles compared to lower buildings. The separation is more pronounced for Heating Load than for Cooling Load, suggesting that vertical building structure may affect heat loss more strongly than cooling requirements.

4.3.4 Glazing Area

is positively associated with energy demand, particularly with Heating Load. Buildings with larger glazing proportions tend to exhibit higher heating requirements, consistent with increased heat transfer through window surfaces. Nevertheless, wide overlap across glazing levels indicates that glazing effects are moderated by other architectural characteristics such as compactness and surface exposure.

4.4 Categorical Design Variables and Energy Demand

Categorical building features introduce additional heterogeneity in energy demand.

4.4.1 Orientation

Boxplots by **Orientation** suggest modest differences in both Heating Load and Cooling Load across directional categories. While some orientations appear to be associated with slightly higher or lower average energy demand, within-group variability is substantial, and no single orientation dominates the overall distribution.

4.4.2 Glazing Area Distribution

Similarly, **Glazing Area Distribution** Boxplots of Heating Load and Cooling Load by Glazing Area Distribution reveal a clear and systematic pattern. Buildings with a uniform glazing area distribution exhibit substantially lower heating and cooling loads compared to all other distribution types.

In contrast, the remaining glazing distributions—where glazing is concentrated on a specific orientation—display very similar energy demand levels among themselves. Differences between these non-uniform categories are minor, and their interquartile ranges overlap strongly, indicating that orientation-specific glazing placement does not meaningfully differentiate energy demand within this group.

This pattern suggests that the uniform distribution of glazing is associated with more favorable energy performance, while concentrating glazing on a single façade tends to increase both heating and cooling requirements. At the same time, the similarity among the non-uniform categories implies that the presence of concentration itself, rather than its specific orientation, is the dominant factor affecting energy demand.

5 Distributional Analysis and Probability

Building on the exploratory insights from the previous section, this part of the analysis moves beyond descriptive visualization and focuses on distributional structure, variability, and probabilistic reasoning. In contrast to the exploratory analysis, the emphasis here is on formal characterization of uncertainty and dependence between outcomes, as well as on probability-based questions that are directly relevant for energy efficiency assessment.

Throughout this section, Heating Load and Cooling Load are analyzed jointly.

5.1 Joint Distribution of Heating and Cooling Loads

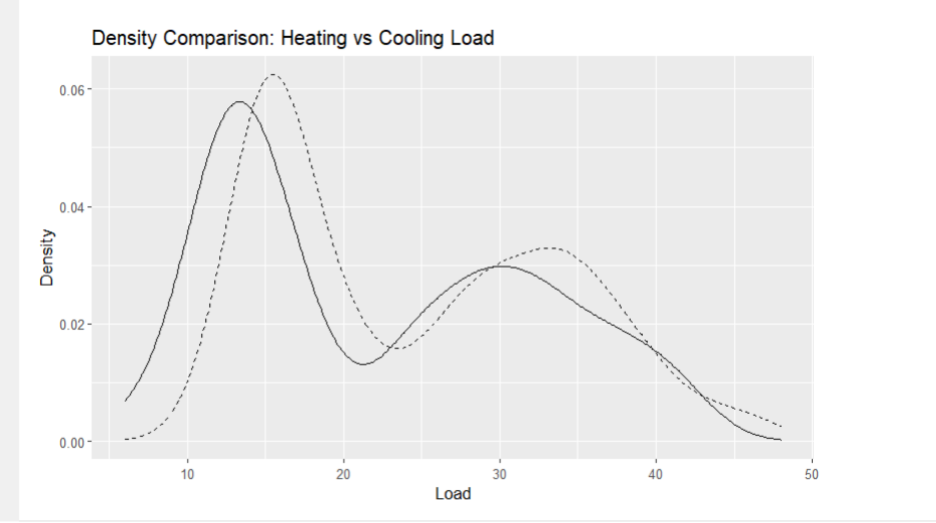


Figure 3: Overlaid kernel density estimates of Heating Load and Cooling Load.

Figure 3 presents overlaid density estimates of Heating Load and Cooling Load. While these curves represent marginal distributions, their comparison provides important insight into the **shared structure** of the two outcomes. In particular, both variables exhibit a **similar bimodal pattern**, with corresponding low-load and high-load regimes appearing at comparable ranges of energy demand. This alignment of modes suggests that Heating Load and Cooling Load are influenced by **common underlying building characteristics**, rather than representing independent aspects of energy performance.

From a joint-distribution perspective, the similarity in density shapes indicates that realizations of Heating Load and Cooling Load are not generated independently. Instead, the data are consistent with a situation in which buildings belong to latent energy-efficiency regimes that simultaneously affect both outcomes. Formally, this implies that the joint density

$$f_{Y_H,Y_C}(h,c)$$

cannot be approximated by the product of the marginal densities

$$f_{Y_H}(h)\, f_{Y_C}(c),$$

as such a factorization would require statistical independence.

The slight shifts in location and dispersion between the Heating Load and Cooling Load densities further suggest that, although the two variables share common structure, they respond differently to certain design features. For example, Cooling Load exhibits a heavier right tail, indicating that extreme energy demand may arise more frequently for cooling than for heating. This asymmetry implies that the joint distribution is not simply a scaled version of a single latent variable but reflects **both shared and outcome-specific variability**.

Taken together, the density comparison provides evidence of a **strong joint structure** between Heating Load and Cooling Load. Rather than treating the two outcomes as separate quantities, the joint distributional perspective highlights that they capture closely related dimensions of building energy efficiency. This observation motivates the formal dependence analysis conducted in the next subsection.

5.2 Variability and Dependence Between the Two Outcomes

To formally assess the dependence suggested by the distributional similarities discussed above, the relationship between Heating Load and Cooling Load is examined using a scatterplot and a quantitative dependence measure.

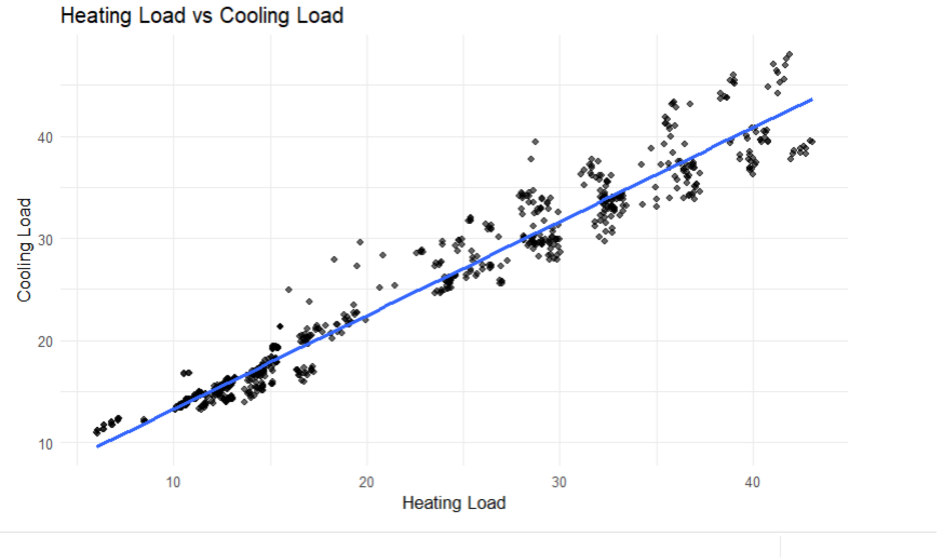


Figure 4: Scatterplot of Heating Load versus Cooling Load with a smooth fitted trend.

Figure 4 displays a scatterplot of Heating Load versus Cooling Load together with a smooth fitted trend. The plot shows that observations are concentrated along a narrow, upward-sloping band. As Heating Load increases, Cooling Load increases in a highly systematic manner across the entire range of values. This visual pattern indicates a strong monotonic association between the two outcomes. Although some dispersion around the fitted trend is present—particularly at higher load levels—the variability remains small relative to the overall co-movement. No evidence of independent regimes or divergent behavior between heating and cooling demand is observed.

To confirm this visual impression, dependence is quantified using the **Spearman rank correlation coefficient**, which is appropriate given the non-normal distributional features identified earlier. Spearman's rho is defined as

$$\rho_S = \mathrm{corr}\big(\mathrm{rank}(Y_H), \mathrm{rank}(Y_C)\big),$$

where Y_H denotes Heating Load and Y_C denotes Cooling Load.

The estimated correlation is

$$\rho_S = 0.9727.$$

This value indicates an **extremely strong positive dependence** and exceeds our threshold of **0.95** used in this study to identify near-redundant outcome variables. The result confirms that Heating Load and Cooling Load contain largely overlapping information.

Despite minor outcome-specific differences, the magnitude of the correlation implies that analyzing both variables as separate primary outcomes would introduce substantial redundancy. Therefore, based on the combined qualitative observation in Section 5.1 and the quantitative confirmation provided here, the subsequent analysis focuses on **Heating Load as a single representative outcome variable**.

5.3 Probability-Based Questions Relevant to Energy Efficiency Thresholds

Having established that **Heating Load** serves as a representative outcome variable, this subsection focuses on **concrete probability-based questions** that are directly evaluated in the analysis code. Rather than considering arbitrary thresholds, the probabilities are defined using **explicit and interpretable energy-demand ranges**.

Let Y_H denote the Heating Load.

5.3.1 Probability of High Heating Demand

The first quantity of interest is the probability that a building exhibits **high heating demand**, defined as a Heating Load exceeding 30 units:

$$P(Y_H > 30) = 0.2942.$$

Using the estimated Gaussian finite mixture model, this probability is given by

$$P(Y_H < 10) = 0.2942.$$

This result indicates that approximately **29.42%** of the simulated building designs fall into a high-energy-demand regime. From an energy-efficiency perspective, this proportion is substantial and suggests that nearly one third of the considered architectural configurations can be classified as energy-inefficient with respect to heating requirements. In practical terms, this probability quantifies the risk associated with selecting unfavorable design choices that lead to excessive heating demand.

5.3.2 Probability of Low Heating Demand

The second quantity of interest is the probability that a building exhibits **very low heating demand**, defined as a Heating Load below 10 units:

$$P(Y_H < 10).$$

Using the estimated Gaussian finite mixture model, this probability is given by

$$P(Y_H < 10) = 0.0522.$$

This result indicates that only about **5.2%** of the simulated building designs achieve very low heating energy demand. Consequently, highly energy-efficient buildings are relatively rare within the considered design space.

From an energy-efficiency perspective, this low probability suggests that achieving exceptionally low heating demand requires specific and potentially restrictive design choices. From a policy standpoint, this finding highlights the challenge of promoting best-practice energy-efficient designs, as they represent only a small fraction of all possible architectural configurations.

5.3.3 Probability of Moderate Heating Demand

To complement the analysis of extreme energy-demand regimes, the probability that Heating Load lies within an intermediate range between 10 and 25 units is also considered:

$$P(10 \leq Y_H \leq 25).$$

Based on the mixture-model estimates, this probability is

$$P(10 \leq Y_H \leq 25) = 0.5193.$$

This result shows that approximately **51.9%** of the building designs fall into a moderate heating-demand regime. This group constitutes the majority of the dataset and represents buildings with balanced energy performance, which are neither highly energy-efficient nor particularly energy-intensive.

From a practical perspective, this finding suggests that most architectural configurations yield intermediate heating requirements. Consequently, moderate energy performance appears to be the norm rather than the exception within the simulated design space, highlighting the importance of incremental design improvements to shift buildings from moderate to low energy-demand regimes.

5.3.4 Interpretation and Implications

Taken together, these three probabilities partition the distribution of Heating Load into **low**, **moderate**, and **high** energy-demand regimes. This probabilistic characterization goes beyond average behavior and explicitly accounts for the variability observed in the data.

Importantly, because Heating Load and Cooling Load were shown to be strongly dependent in Section 5.2, these probabilities also indirectly reflect cooling-related energy risk. As a result, focusing on Heating Load alone is sufficient for evaluating the likelihood of unfavorable energy outcomes in the present dataset.

6 Statistical Inference (Hypothesis Testing)

Based on prior exploratory findings and theoretical considerations, **two hypotheses** are formulated and tested. Each hypothesis is evaluated using a statistical test that is consistent with the scale of the variables, distributional properties, and methods introduced in the course.

6.1 Hypothesis 1: Effect of Building Height on Heating Load

The first hypothesis examines whether building height is associated with differences in average heating demand. Building height is represented by the categorical variable **Overall Height**, which distinguishes between low-rise and high-rise buildings.

6.1.1 Hypotheses

- **Null Hypothesis (H_0):** The mean Heating Load is the same for low-rise and high-rise buildings.
- **Alternative Hypothesis (H_1):** The mean Heating Load differs between low-rise and high-rise buildings.

Formally, letting μ_L and μ_H denote the mean Heating Load for low-rise and high-rise buildings, respectively, the hypotheses can be written as

$$H_0 : \mu_L = \mu_H, \qquad H_1 : \mu_L \neq \mu_H.$$

6.1.2 Choice of Statistical Test

To test this hypothesis, a **Welch two-sample t-test** is employed. This test compares the means of two independent groups without assuming equal variances. The choice of the Welch test is motivated by two considerations:

1. Heating Load is a continuous outcome variable.
2. There is no *a priori* reason to assume homogeneity of variances between low-rise and high-rise buildings.

Compared to the pooled t-test, the Welch test provides a more robust inference when group variances differ.

6.1.3 Results and Interpretation

The Welch two-sample t-test provides **very strong evidence** against the null hypothesis of equal mean Heating Load for low-rise and high-rise buildings. The test statistic is extremely large ($t = 53.86$), and the associated p-value is **smaller than** 2.2×10^{-16}. Consequently, the null hypothesis of equal means is rejected at any conventional significance level. The estimated mean Heating Load differs substantially between the two height groups. High-rise buildings exhibit an average Heating Load of **31.28**, whereas low-rise buildings show a considerably lower average value of **13.34**. The estimated difference in means is therefore large and economically meaningful. This difference is further supported by the 95% confidence interval for the mean difference, which ranges from **17.28 to 18.59**. Since this interval does not include zero, it confirms that the observed difference is not only statistically significant but also precisely estimated.

From a substantive perspective, these results indicate that **building height is strongly associated with heating energy demand**. Taller buildings systematically require substantially more heating energy than lower buildings in the simulated design space. This finding is consistent with increased heat loss and exposure effects associated with greater vertical building structures.

Overall, Hypothesis 1 is clearly supported by the data: building height has a significant and practically relevant effect on Heating Load.

6.2 Hypothesis 2: Association Between Surface Area and Heating Load (Results and Interpretation)

The second hypothesis examines whether the **Surface Area** of a building is associated with Heating Load. Surface Area is a continuous design variable capturing the total external envelope size of the building.

6.2.1 Hypotheses

- **Null Hypothesis (H_0):** Surface Area has no association with Heating Load.
- **Alternative Hypothesis (H_1):** Surface Area is associated with Heating Load.

Formally, in terms of monotonic association, the hypotheses are stated as

$$H_0 : \rho_S = 0, \qquad H_1 : \rho_S \neq 0,$$

where ρ_S denotes Spearman's rank correlation coefficient between Surface Area and Heating Load.

6.2.2 Choice of Statistical Test

Given the non-normal distributional properties identified earlier and the exploratory evidence of a monotonic relationship, the association is assessed using the **Spearman rank correlation test**. This test is robust to outliers, does not rely on parametric assumptions, and is fully consistent with the methods introduced in the course.

6.2.3 Results

The estimated Spearman correlation coefficient between Surface Area and Heating Load is

$$\rho_S = -0.622.$$

The associated p-value is **smaller than** 2.2×10^{-16}, providing overwhelming evidence against the null hypothesis of no association.

6.2.4 Interpretation

The results indicate a **strong and statistically significant negative monotonic relationship** between Surface Area and Heating Load. Buildings with larger surface area tend to exhibit **lower heating energy demand**, whereas buildings with smaller surface area are associated with higher Heating Load.

The magnitude of the correlation suggests that Surface Area explains a substantial portion of the variation in Heating Load, although it does not fully determine energy demand on its own. This finding is fully consistent with the exploratory visualizations presented earlier, where Surface Area displayed a clear downward trend with respect to Heating Load.

From a substantive perspective, this result highlights that Surface Area captures geometric characteristics that are inversely related to heating demand within the simulated design space.

Overall, Hypothesis 2 is strongly supported by the data: Surface Area exhibits a robust and meaningful association with Heating Load.

7 Supervised Learning

7.1 Definition of the Prediction Task

Based on the results of the distributional analysis and hypothesis testing, **Heating Load** is selected as the sole response variable for supervised learning. This choice is justified by the extremely strong dependence between Heating Load and Cooling Load (Spearman's $\rho = 0.9727 > 0.95$), which indicates that both outcomes convey largely redundant information.

The supervised learning task is therefore defined as:

Predict Heating Load using building design characteristics.

The predictor set consists exclusively of **design variables**, excluding the response variable to avoid information leakage. Specifically, the following predictors are used:

- Relative Compactness
- Surface Area
- Wall Area

- Overall Height
- Orientation
- Glazing Area
- Glazing Area Distribution

7.2 Training–Test Split and Evaluation Metrics

To evaluate predictive performance, the dataset is randomly split into a **training set (80%)** and a **test set (20%)**, using a fixed random seed to ensure reproducibility. All models are fitted on the training data and evaluated on the held-out test data.

Model performance is assessed using two standard regression metrics:

- **Root Mean Squared Error (RMSE)**

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- **Mean Absolute Error (MAE)**

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

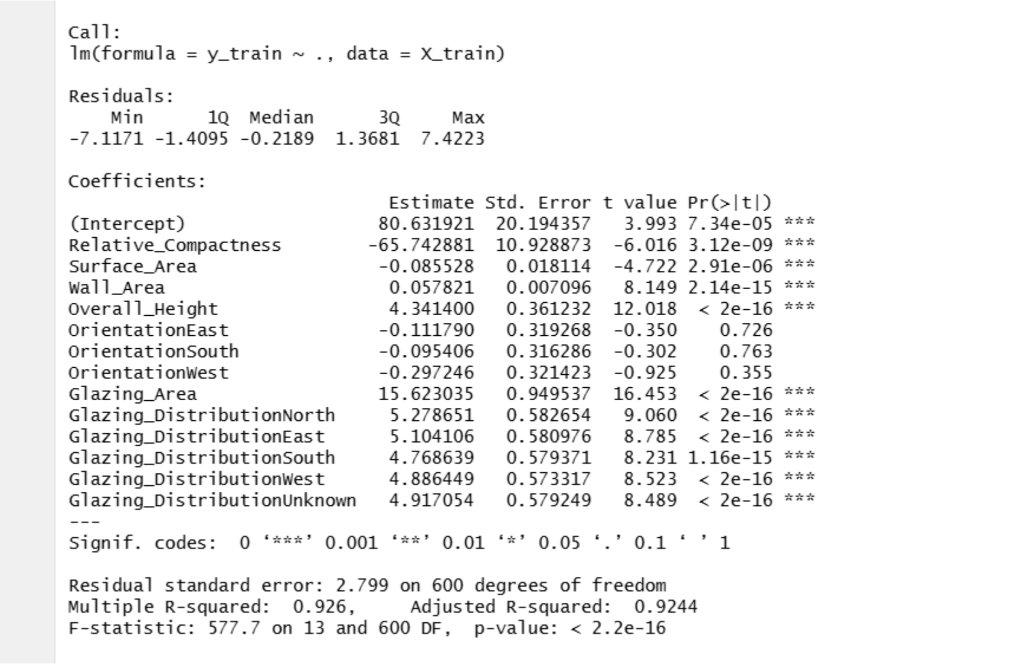
RMSE penalizes large prediction errors more strongly, while MAE provides a robust measure of average absolute deviation. Using both metrics allows for a balanced comparison of models.

7.3 Competing Supervised Learning Models

Four supervised learning models discussed in the course are fitted and compared.

7.3.1 Model 1: Linear Regression Model for Heating Load

We model the *Heating Load* as a function of the main building design characteristics using a multiple linear regression model estimated by ordinary least squares.



```
Call:
lm(formula = y_train ~ ., data = X_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.1171 -1.4095 -0.2189  1.3681  7.4223 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  80.631921  20.194357   3.993 7.34e-05 ***
Relative_Compactness        -65.742881  10.928873  -6.016 3.12e-09 ***
Surface_Area                 -0.085528   0.018114  -4.722 2.91e-06 ***
Wall_Area                     0.057821   0.007096   8.149 2.14e-15 ***
Overall_Height                4.341400   0.361232  12.018  < 2e-16 ***
OrientationEast              -0.111790   0.319268  -0.350    0.726    
OrientationSouth             -0.095406   0.316286  -0.302    0.763    
OrientationWest              -0.297246   0.321423  -0.925    0.355    
Glazing_Area                 15.623035   0.949537  16.453  < 2e-16 ***
Glazing_DistributionNorth     5.278651   0.582654   9.060  < 2e-16 ***
Glazing_DistributionEast      5.104106   0.580976   8.785  < 2e-16 ***
Glazing_DistributionSouth     4.768639   0.579371   8.231 1.16e-15 ***
Glazing_DistributionWest      4.886449   0.573317   8.523  < 2e-16 ***
Glazing_DistributionUnknown   4.917054   0.579249   8.489  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.799 on 600 degrees of freedom
Multiple R-squared:  0.926,	Adjusted R-squared:  0.9244 
F-statistic: 577.7 on 13 and 600 DF,  p-value: < 2.2e-16
```

Figure 5: Output linear model .

Mathematical formulation of the estimated linear model

Based on the output of the `lm()` function, the **prediction function for the Heating Load** (denoted $\widehat{Y}_H$) is given by:

$$\begin{aligned}
\widehat{Heating\_Load} = 80.632 \\
&- 65.743 \cdot Relative\_Compactness \\
&- 0.0855 \cdot Surface\_Area \\
&+ 0.0578 \cdot Wall\_Area \\
&+ 4.341 \cdot Overall\_Height \\
&- 0.112 \cdot Orientation_{East} \\
&- 0.095 \cdot Orientation_{South} \\
&- 0.297 \cdot Orientation_{West} \\
&+ 15.623 \cdot Glazing\_Area \\
&+ 5.279 \cdot Glazing\_Distribution_{North} \\
&+ 5.104 \cdot Glazing\_Distribution_{East} \\
&+ 4.769 \cdot Glazing\_Distribution_{South} \\
&+ 4.886 \cdot Glazing\_Distribution_{West} \\
&+ 4.917 \cdot Glazing\_Distribution_{Unknown}.
\end{aligned} \tag{1}$$

The **orientation** and **glazing distribution** variables are **indicator (dummy) variables**, interpreted **relative to a reference category** (the category that does not appear explicitly in the model).

Interpretation of the estimated coefficients

Intercept (80.63).

- Theoretical value of the **Heating Load** when all explanatory variables are equal to zero.
- **Not physically interpretable**, but necessary to fit the model.

Relative Compactness (65.74) *.**

- **Strong and negative effect**.
- An increase of 0.1 in relative compactness leads on average to a **decrease of approximately 6.6 units** in heating demand.
- More compact buildings are **significantly more energy efficient**.

Surface Area (0.0855) *.**

- An increase of one unit in surface area slightly reduces the heating demand.
- Statistically significant but **moderate effect**.

Wall Area (+0.0578) *.**

- A larger wall area increases thermal losses.
- **Positive and significant effect**, consistent with building physics.

Overall Height (+4.34) *.**

- An increase in building height substantially increases heating demand.
- **Highly significant effect** (p-values < 0.0001), indicating strong sensitivity to building verticality.

Orientation (East, South, West).

- **Non-significant coefficients** (p-values > 0.7).
- Building orientation has **no statistically detectable effect** on heating load in this model once other variables are controlled for.

Glazing Area (+15.62) *.**

- **Very strong and positive effect** (p-values < 0.0001).
- Increasing glazing area substantially increases heating demand.
- Large glazed surfaces are **energetically unfavorable** for heating.

Glazing Distribution (all significant).

- All glazing distribution categories increase the heating load relative to the reference category (Uniform Distribution).
- Effects range between **+4.7 and +5.3 units**.
- The spatial distribution of glazing significantly influences energy performance.

Overall model quality

Residual Standard Error = 2.799.

- Average prediction error of approximately ±2.8 **units** of Heating Load.
- Indicates **good predictive accuracy**.

F-statistic = 577.7, p-value $< 2.2 \times 10^{-16}$**.**

- Global model test:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{13} = 0$$

- The null hypothesis is **strongly rejected**.
- The model is **globally statistically significant**.

7.3.2 Model 2: Ridge Regression

To address potential multicollinearity among predictors, a Ridge regression model is estimated. Ridge regression applies an ℓ_2-penalty to shrink coefficients toward zero. The regularization parameter λ is selected via cross-validation on the training data.

7.3.3 Model 3: Lasso Regression

Lasso regression is also considered as an alternative regularized model. By applying an ℓ_1-penalty, Lasso allows for coefficient shrinkage and implicit variable selection. As with Ridge regression, the tuning parameter λ is chosen via cross-validation.

7.3.4 Model 4: Random Forest Regression

Finally, a Random Forest regression model is fitted to capture nonlinear relationships and interactions between predictors. The model is trained using 500 trees, with the number of variables considered at each split set to $\lfloor \sqrt{p} \rfloor$, following standard practice.

Unlike linear models, Random Forests make minimal assumptions about functional form and are well suited for complex, high-dimensional prediction tasks.

7.4 Model Comparison

All four models are evaluated on the same test set using RMSE and MAE. The results are summarized in Table 1.

Model	RMSE	MAE
Linear Regression	2.87	2.09
Ridge Regression	2.99	2.07
Lasso Regression	2.87	2.09
Random Forest	1.03	0.78

Table 1: Predictive performance of supervised learning models on the test set.

The linear and regularized regression models achieve similar predictive performance, with only minor differences in RMSE and MAE. Regularization does not substantially improve prediction accuracy relative to ordinary least squares in this setting. In contrast, the Random Forest model **outperforms all linear models by a large margin**, reducing RMSE by more than 60% and MAE by more than 50%. This indicates that nonlinear effects and interactions between building design variables play a major role in determining Heating Load.

7.5 Final Model Choice

Based on out-of-sample predictive performance, the **Random Forest regression model** is selected as the final supervised learning model. Its superior accuracy demonstrates that flexible, nonparametric approaches are better suited for predicting Heating Load than linear or regularized linear models.

At the same time, the comparison highlights an important trade-off: while linear models offer greater interpretability, they fail to capture complex relationships present in the data. Random Forests sacrifice some interpretability in exchange for substantially improved predictive performance.

8 Unsupervised Learning

The purpose of this section is to explore whether buildings can be grouped into **homogeneous clusters based solely on design characteristics**, without using any energy outcome variables. In contrast to supervised learning, unsupervised methods aim to uncover latent structure in the design space rather than optimize prediction accuracy.

Throughout this section, **Heating Load is explicitly excluded from the clustering procedure** and is only used **post hoc** for interpretation.

8.1 Variable Selection for Clustering

Clustering is performed using a subset of numerical **building design variables** that describe geometric and envelope-related characteristics:

- Relative Compactness
- Surface Area
- Wall Area

- Overall Height
- Glazing Area

Categorical variables (Orientation and Glazing Area Distribution) are excluded from clustering, as distance-based methods require numerical input and meaningful distance computations. Similarly, outcome variables such as Heating Load and Cooling Load are omitted to avoid information leakage and to ensure that clusters reflect **design similarity only**.

This variable selection is consistent with the objective of identifying **architectural building types** rather than energy-performance classes.

8.2 Preprocessing and Distance Measure

Before clustering, all selected variables are **standardized** to have zero mean and unit variance. This step is essential because the variables are measured on different scales and would otherwise contribute unevenly to distance calculations. Clustering is based on **Euclidean distance**, which is appropriate for continuous, standardized variables and is commonly used in both hierarchical clustering and K-means. Using the same distance metric across methods ensures comparability of results.

8.3 Hierarchical Clustering

Hierarchical clustering is first applied using **Ward's minimum variance method (Ward.D2)**. This method constructs clusters by minimizing within-cluster variance at each merging step and tends to produce compact and interpretable clusters.

The resulting dendrogram reveals a clear separation at a high linkage distance, suggesting a small number of well-separated clusters. Cutting the dendrogram at $k = 2$ yields two clusters of equal size (384 observations each), indicating a balanced partition of the building designs.

At this stage, hierarchical clustering is used primarily for **structural insight and robustness assessment**, rather than for final cluster selection.

8.4 K-Means Clustering and Choice of the Number of Clusters

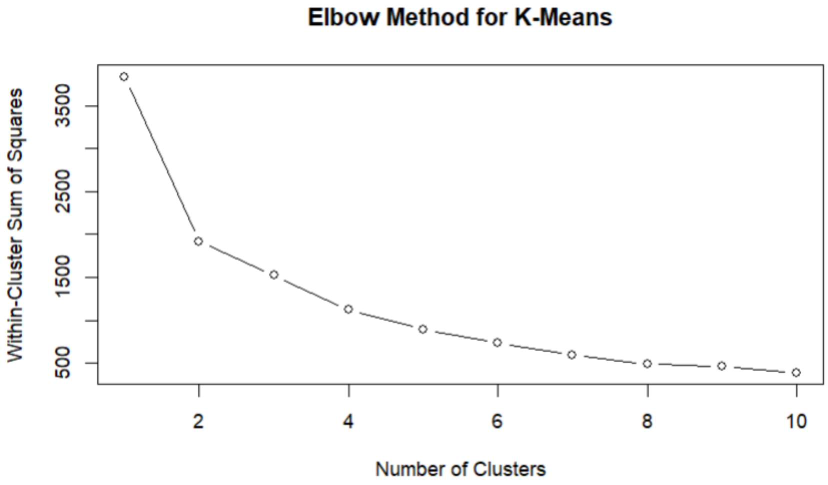


Figure 6: Elbow Method for K-Means.

To determine the appropriate number of clusters more formally, **K-means clustering** is applied in combination with the **Elbow Method**. The Elbow plot shows the within-cluster sum of squares (WSS) as a function of the number of clusters k.

The plot exhibits a pronounced decrease in WSS when moving from $k = 1$ to $k = 2$, followed by a much more gradual decline for larger values of k. This pattern indicates that $k = 2$ **represents the primary structural split in the data** and provides the best trade-off between cluster compactness and model simplicity.

Based on this criterion, K-means clustering is performed with $k = 2$. The resulting cluster sizes again consist of 384 observations per cluster, matching the hierarchical clustering solution.

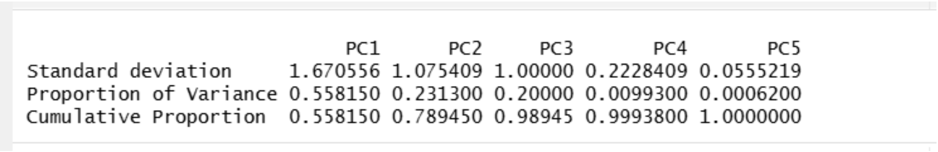
Importantly, the agreement between hierarchical clustering and K-means provides a **robust confirmation** that the two-cluster structure is not an artifact of a specific algorithm.

8.5 Comparison of Clustering Results

To assess consistency between the two unsupervised methods, cluster assignments from hierarchical clustering and K-means are cross-tabulated. The strong alignment between the two solutions indicates that both methods identify **essentially the same partition of the design space**.

This agreement strengthens confidence in the identified cluster structure and suggests that the underlying grouping reflects genuine patterns in building design characteristics.

8.6 PCA-Based Visualization of Clusters



```
                            PC1      PC2     PC3       PC4       PC5
Standard deviation     1.670556 1.075409 1.00000 0.2228409 0.0555219
Proportion of Variance 0.558150 0.231300 0.20000 0.0099300 0.0006200
Cumulative Proportion  0.558150 0.789450 0.98945 0.9993800 1.0000000
```

Figure 7: Principal Component.

To visualize the clusters in a low-dimensional space, **Principal Component Analysis (PCA)** is applied to the standardized design variables. PCA is used **exclusively for visualization purposes** and does not influence cluster formation.

The first two principal components explain approximately **79% of the total variance**, indicating that most of the variation in building design can be represented in a two-dimensional space. A scatterplot of the first two principal components, colored by cluster membership, shows a clear separation between the two clusters.

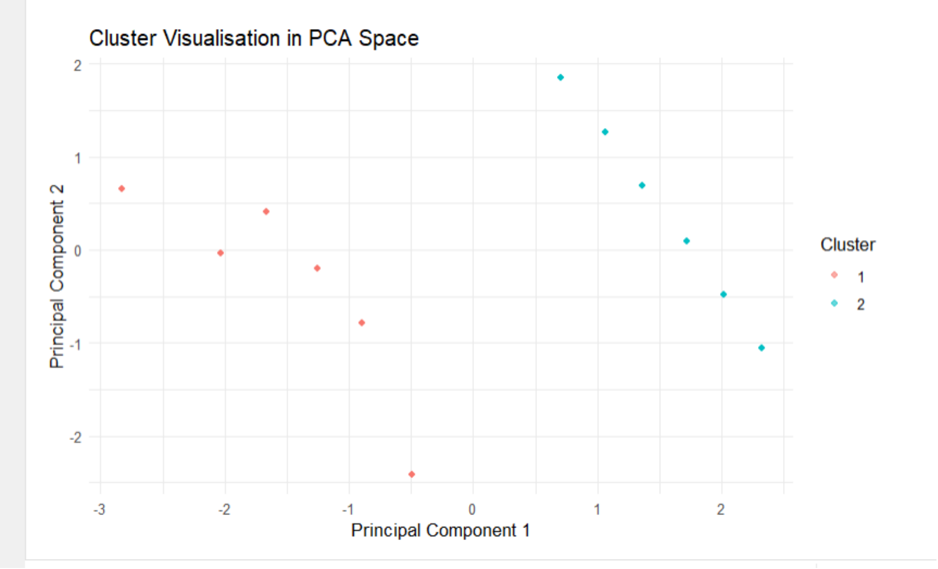


Figure 8: Visualisation of clusters

This visualization confirms that the clusters correspond to distinct regions in the design space rather than overlapping groups.

8.7 Interpretation and Implications

The unsupervised analysis identifies **two distinct building design clusters** that are stable across methods and clearly separable in the reduced PCA space. The consistency between hierarchical clustering and K-means supports the robustness of the two-cluster solution.

From an applied perspective, the results suggest that building designs naturally fall into two broad architectural types, which are also associated with different heating energy demands. This finding complements the supervised learning results by providing a **structural explanation** for heterogeneity in energy performance.

Overall, the unsupervised learning analysis demonstrates that meaningful latent structure exists in the design space and that this structure aligns with differences in energy efficiency, even when energy outcomes are not used during clustering.

9 Discussion and Limitations

9.1 Discussion

This study investigated the relationship between building design characteristics and energy demand using a structured statistical workflow that combined exploratory analysis, formal inference, supervised learning, and unsupervised learning. The analysis was based on simulated residential building data and focused primarily on **Heating Load** as the key outcome variable.

A central result of the analysis is the **very strong dependence between Heating Load and Cooling Load**. Both exploratory visualization and formal dependence measures demonstrated that the two outcomes are almost perfectly aligned, with a Spearman rank correlation of $\rho = 0.9727$. This finding justified the methodological decision to restrict subsequent inference and modeling to a single response variable. By focusing on Heating Load, redundancy was reduced without discarding meaningful information about overall energy performance.

The hypothesis testing results provide clear evidence that certain design characteristics are systematically associated with heating energy demand. In particular, building height exhibits a large and statistically significant effect: high-rise buildings require substantially more heating energy than low-rise buildings. In addition, surface-related characteristics show strong monotonic associations with Heating Load. Surface Area is negatively associated with heating demand, while Glazing Area exhibits a positive association. These findings highlight that different geometric measures capture distinct and sometimes counterintuitive aspects of building energy behavior, reinforcing the importance of empirical analysis over purely theoretical expectations.

From a predictive perspective, supervised learning results demonstrate that **model flexibility plays a crucial role**. Linear regression and regularized variants (Ridge and Lasso) perform similarly, suggesting that linear structure alone is insufficient to fully capture the complexity of the data. In contrast, the Random Forest model achieves substantially lower RMSE and MAE, indicating that nonlinear effects and interactions between design variables are central drivers of Heating Load. This result underscores the trade-off between interpretability and predictive accuracy: while linear models allow for coefficient-based interpretation, they fail to capture the full structure present in the data.

The unsupervised learning analysis complements these findings by revealing **latent structure in the design space**. Using only numerical design variables, both hierarchical clustering and K-means consistently identify two well-separated clusters. The agreement between clustering methods and the clear separation in PCA space indicate that the two-cluster solution is robust. Importantly, although Heating Load was excluded from clustering, the resulting clusters differ systematically in heating energy demand. This suggests that architectural design alone induces distinct energy-performance profiles, providing a structural explanation for the heterogeneity observed in both inference and prediction results.

Taken together, the results paint a coherent picture: building energy demand is shaped by complex, multivariate interactions between design features. Simple univariate explanations are insufficient, and both flexible predictive models and unsupervised structure discovery are essential for a comprehensive understanding.

9.2 Limitations

Despite the strengths of the analysis, several limitations should be acknowledged.

First, the dataset is **simulation-based rather than observational**. While this ensures internal consistency and eliminates measurement error, it also limits external validity. The results reflect the assumptions embedded in the simulation model and may not fully generalize to real-world buildings, where construction quality, occupant behavior, and climate variability play a significant role.

Second, the analysis is **cross-sectional**. All buildings are treated as independent observations, and no temporal dynamics are considered. Consequently, the study cannot address questions related to seasonal variation, temporal adaptation, or long-term energy performance.

Third, although strong associations are identified, the analysis does **not establish causality**. Hypothesis testing and correlation-based methods reveal statistical relationships but cannot disentangle causal mechanisms. For example, the observed associations between surface-related variables and Heating Load may reflect indirect effects mediated by other design characteristics.

Fourth, the Random Forest model, while highly accurate, suffers from **limited interpretability**. Although variable importance measures can provide some insight, the model does not yield simple parametric relationships. This limits its usefulness for policy-making contexts where transparent decision rules are often required.

Finally, the unsupervised clustering analysis relies on **distance-based methods** and a specific set of numerical design variables. Alternative distance measures, clustering algorithms, or feature representations could potentially reveal different latent structures. While robustness was partially addressed by combining hierarchical clustering and K-means, the clustering results should still be interpreted as exploratory rather than definitive.

9.3 Concluding Remarks

Within these limitations, the analysis demonstrates how a coherent statistical workflow can be applied to building energy data to move from exploration to inference, prediction, and structural interpretation. The combination of supervised and unsupervised methods provides complementary insights and highlights the value of integrating classical statistical reasoning with modern machine learning techniques.

10 Conclusion

This project applied a comprehensive statistical and machine learning framework to analyze building energy performance based on architectural design characteristics. Using simulated residential building data, the analysis systematically progressed from exploratory investigation to formal inference, predictive modeling, and unsupervised structure discovery.

A key finding is the near-redundancy between Heating Load and Cooling Load, which justified focusing subsequent analysis on Heating Load as a single representative outcome. Hypothesis testing revealed that building height and surface-related characteristics are strongly associated with heating energy demand, while supervised learning demonstrated that nonlinear models substantially outperform linear approaches in prediction accuracy.

Unsupervised learning further showed that building designs naturally cluster into two distinct architectural groups based solely on design variables, with these clusters exhibiting systematic differences in energy demand. Together, these results highlight the multivariate and nonlinear nature of building energy efficiency and emphasize the value of integrating classical statistical reasoning with modern machine learning methods.

Overall, the study illustrates how data-driven analysis can support the understanding of complex energy-performance patterns and inform the evaluation of architectural design choices in an energy-efficiency context.