**1.1 Background and Scenario Overview**

Our consultancy focuses on advancing sustainable, low-carbon building designs to reduce energy usage and environmental impact, particularly through minimizing heating and cooling demands (European Commission, 2003). Increasing energy efficiency in buildings not only helps conserve resources but also aligns with regulatory standards like the European Commission's Directive on the Energy Performance of Buildings, which mandates sustainable design practices for environmental compliance (European Commission, 2003).

Statistical analysis is integral to this mission, as it enables the identification of complex relationships between architectural features such as glazing area, compactness, and orientation and the heating and cooling loads of buildings (Tsanasa & Xifara, 2012). Using techniques like regression modelling and correlation analysis, we can quantify these relationships and pinpoint variables that heavily influence energy demands (Tsanasa & Xifara, 2012). This approach equips our consultancy with actionable insights to recommend data-driven, energy-efficient design choices to architects, thus advancing sustainable building practices (European Commission, 2003).

**1.2 Objectives of the Analysis**

**Primary Objective**:

To analyse the relationships between building design variables and energy loads, providing actionable insights that can guide sustainable building design.

**Specific Objectives**:

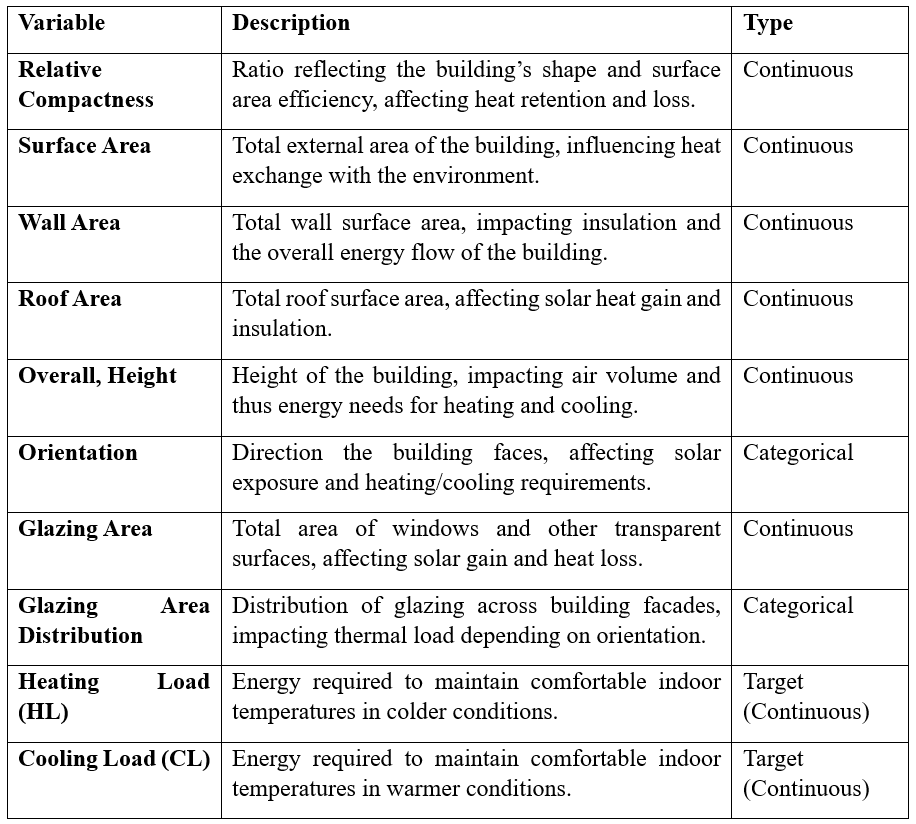
* Conduct **Exploratory Data Analysis (EDA)** to summarize data distributions, identify patterns, and explore variable ranges.
* Perform **correlation analysis** to determine which design features most strongly influence heating and cooling loads.
* Develop and apply **regression models** to predict heating and cooling demands based on design characteristics.
* Conduct **hypothesis testing** to statistically validate assumptions about the impact of specific design variables on energy loads.

**2. Data Description and Preparation**

**2.1 Dataset Overview**

The dataset contains simulated values for various building design features, along with target variables representing **heating load (HL)** and **cooling load (CL)** measures of energy required to maintain comfortable indoor temperatures under different conditions. The **design features** include:

Table 2.1: Dataset description

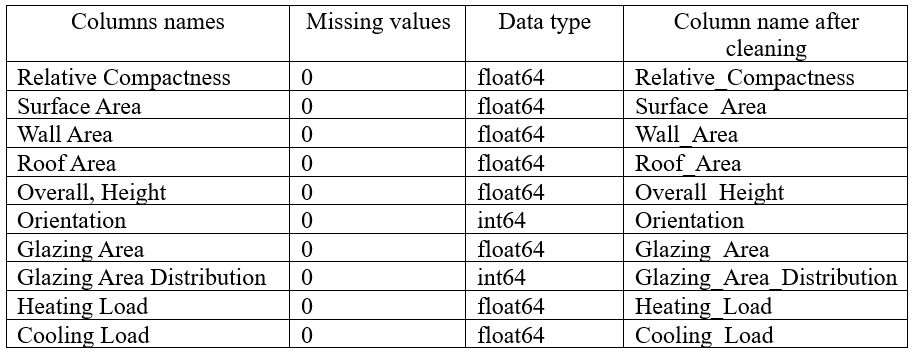


**2.2 Data Cleaning and Preparation**

**2.2.1 Steps Taken to Ensure Data Accuracy**

To prepare the dataset for analysis, several data accuracy checks and cleaning steps were undertaken. First, column names were inspected and adjusted to ensure consistent formatting by replacing any spaces with underscores, preventing issues when referencing variables in code. Additionally, columns containing unnecessary characters, such as tab spaces (\t), were cleaned to maintain uniformity across variable names.

Next, a thorough check for missing values was performed across all variables, confirming that the dataset was complete with no missing entries. This step is critical to avoid any bias or errors in analysis due to incomplete data. The result of the data cleaning process is presented in table below.



**2.2.2 Summary of Transformations and Encoding**

Given that **Orientation** and **Glazing Area Distribution** are categorical variables, these were encoded as factors within R. This encoding ensures that these variables are treated appropriately in statistical models, as categorical data provides context on grouping without implying a numeric hierarchy. This conversion allows for meaningful interpretation in models and visualizations, aligning these variables with their categorical nature.

**3. Exploratory Data Analysis (EDA)**

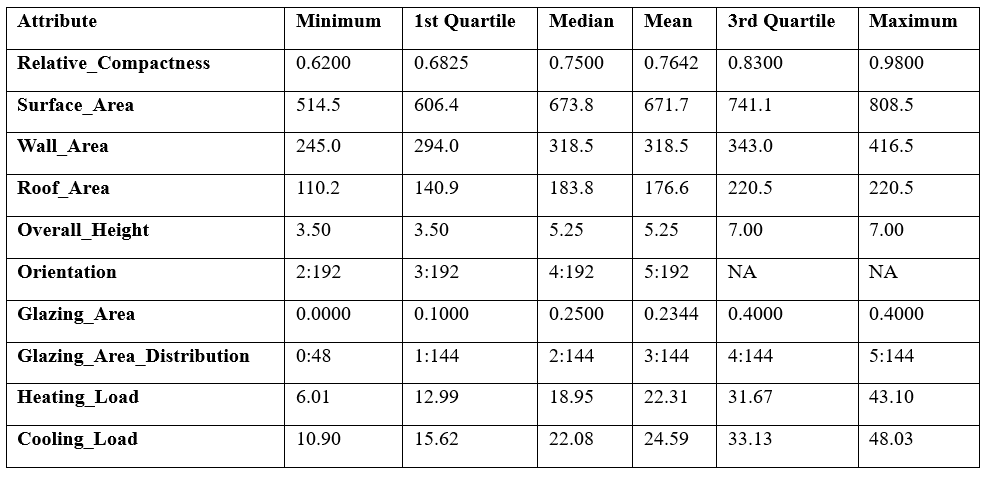
**3.1 Descriptive Statistics**

**Summary Table of Key Descriptive Statistics**

The table 3.1 provided includes key descriptive statistics for each variable, summarizing the central tendencies, variability, and range across all features in the dataset:

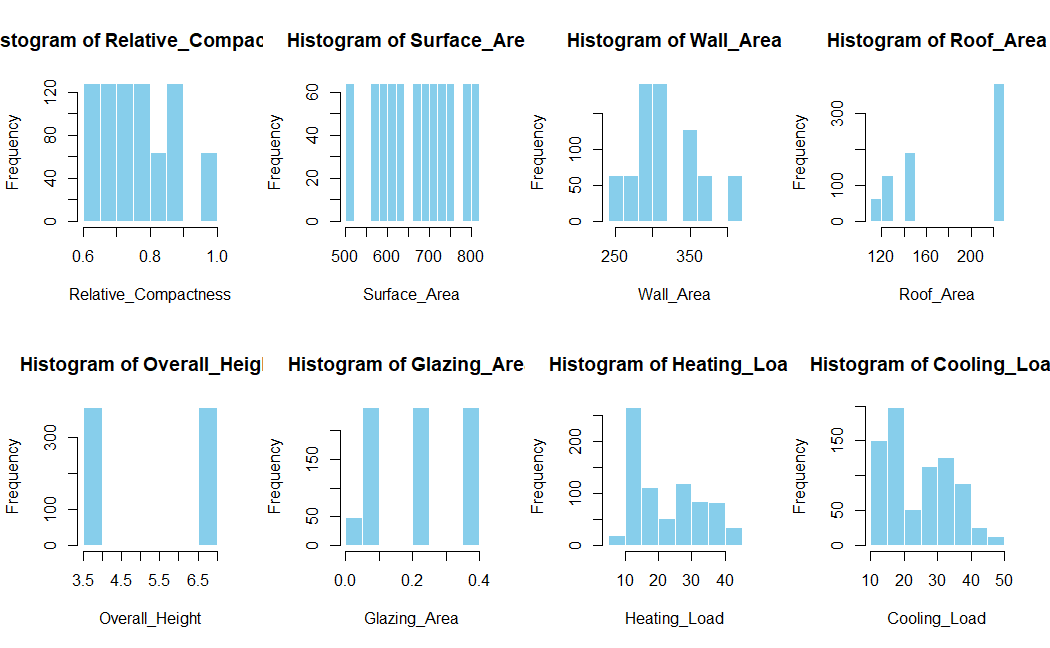
* **Central Tendency**: The mean and median (50th percentile) values indicate central points for each variable. For instance, **Heating Load** has a mean of 22.31 and a median of 18.95, suggesting slight skewness towards higher values, which is also seen in **Cooling Load** with a mean of 24.59 and median of 22.08.
* **Variability**: Standard deviation values show the spread of each variable. **Heating Load** and **Cooling Load** exhibit substantial variability (10.09 and 9.51, respectively), indicating differences in energy demand across building designs. Similarly, **Surface Area** and **Wall Area** have noticeable spread, reflecting variations in building sizes.
* **Range**: The minimum and maximum values reveal each variable’s span. For example, **Relative Compactness** ranges from 0.62 to 0.98, and **Glazing Area** from 0 to 0.4, indicating diversity in building design features.

Table 3.1: Descriptive statistics

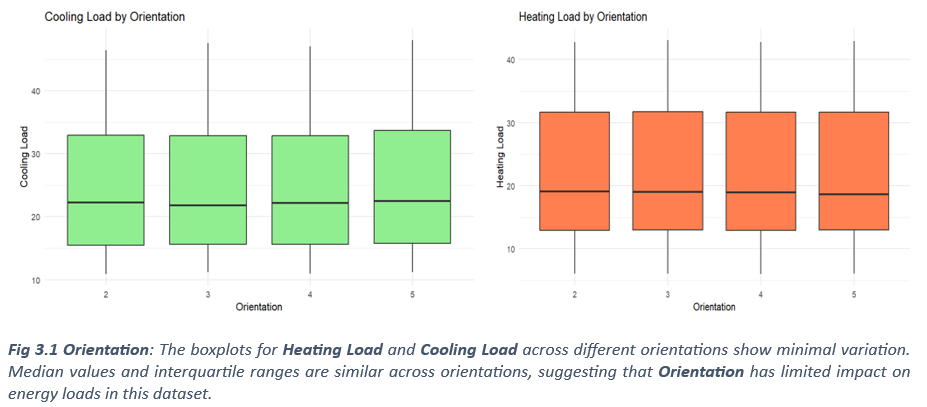


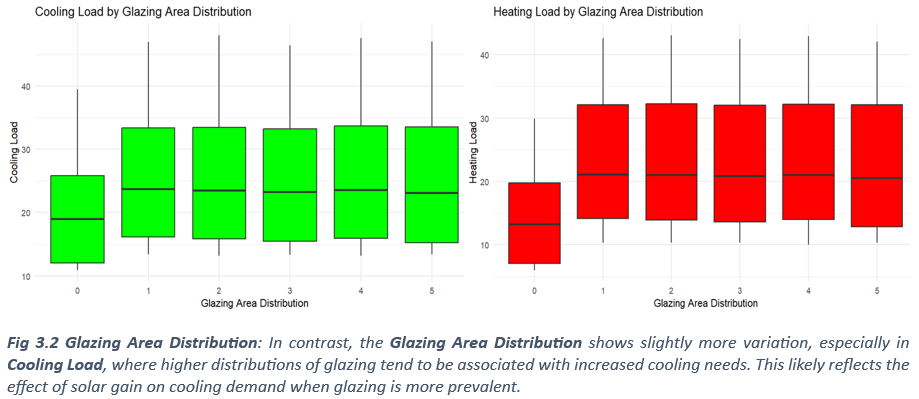
**3.2 Exploratory Analysis (Data Visualization)**

The histograms show the distributions for each variable. Most variables, such as Relative Compactness, Surface Area, Heating Load, and Cooling Load, exhibit a range of values, with some appearing normally distributed and others showing skewness. These distributions highlight the diversity in building designs and energy requirements within the dataset.



Boxplots for Heating Load (HL) and Cooling Load (CL) across Orientation and Glazing Area Distribution provide insights into outliers and variations within these categories. Notably:





The correlation matrix in fig 3.3 below provides insights into the relationships between building design variables and energy loads (Heating Load and Cooling Load):

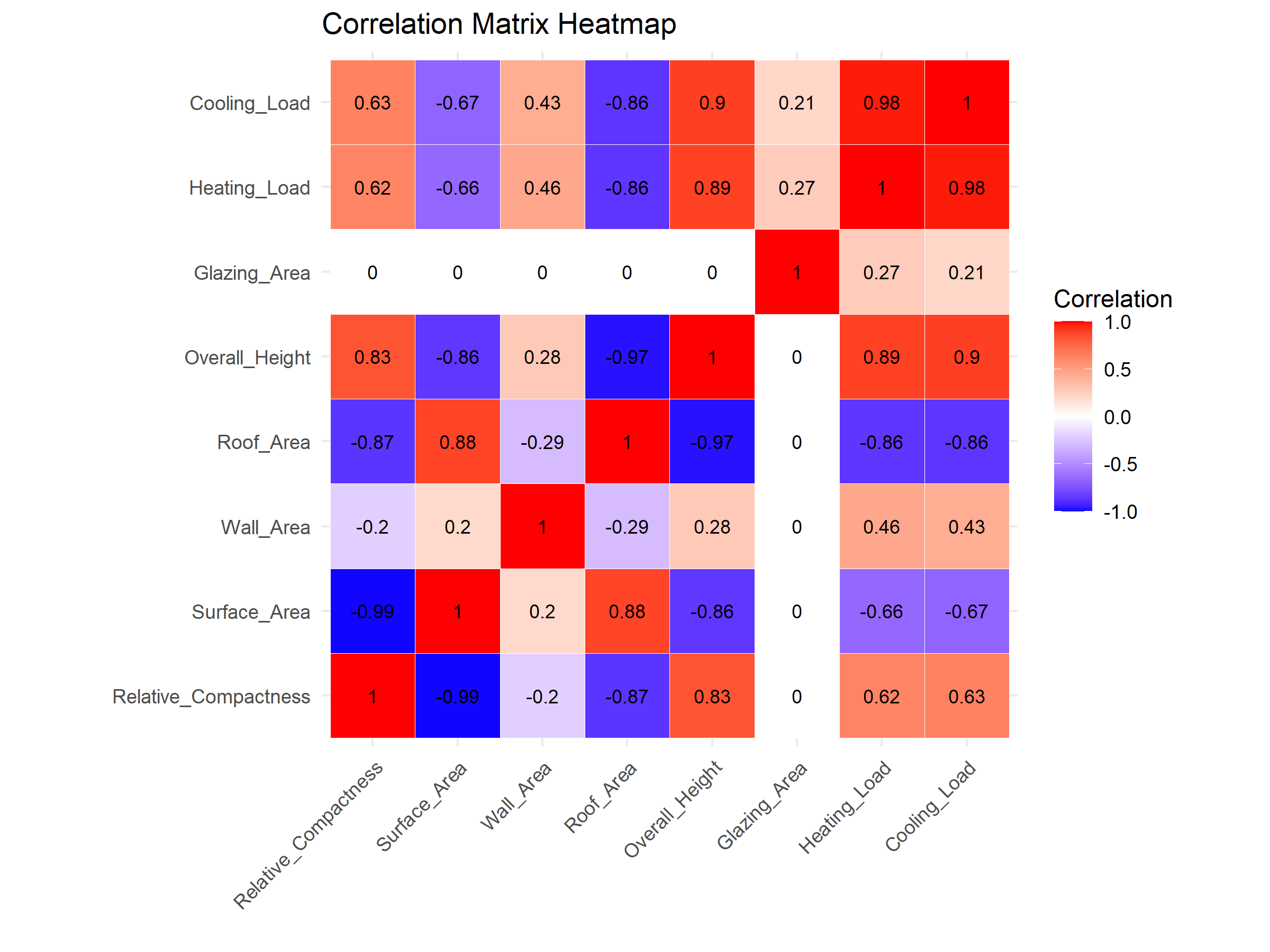


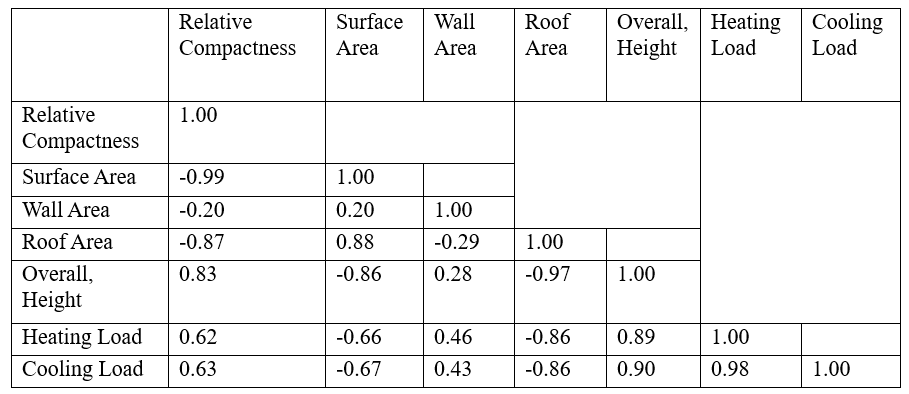
Fig 3.3 Correlation Matrix of Design Variable with Heating Variables

Table 3.1 Interpretation of the Correlation Matrix



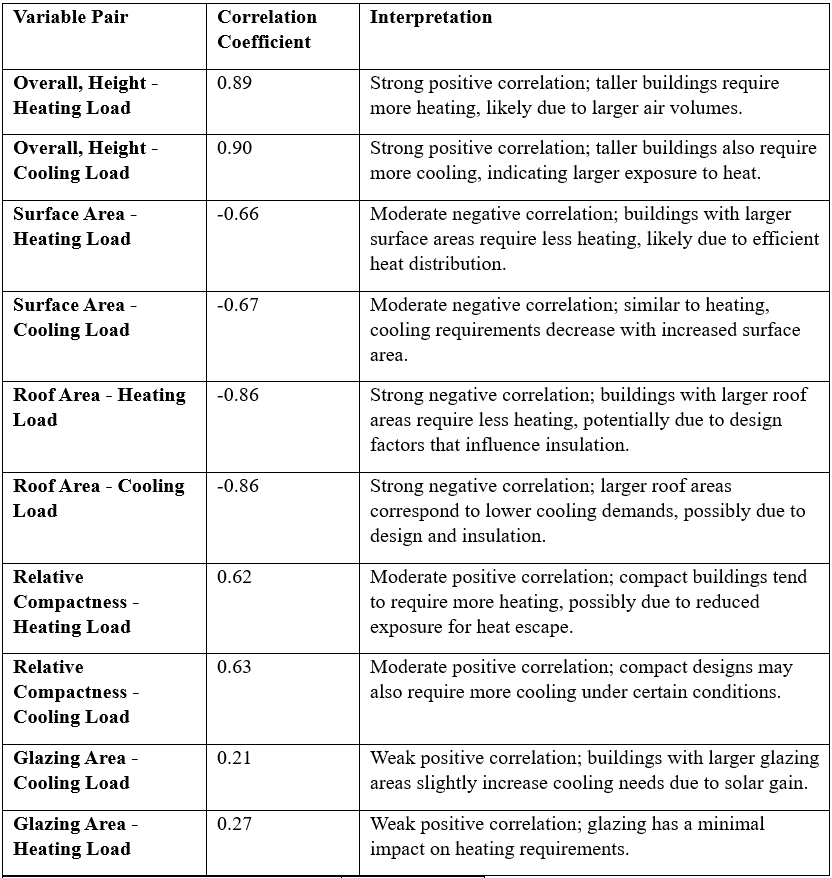
**4. Correlation Analysis**

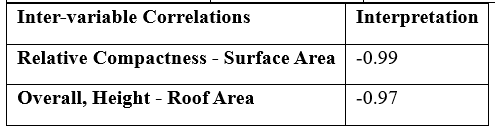
**4.1 Peason Correlation Table**



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  | |  |  | |  |  | |  |  | |  |  |
|  |  |  | |  |  | |  |  | |  |  | |  |  |
|  |  |  | |  |  | |  |  | |  |  | |  |  |
|  |  |  | |  |  | |  |  | |  |  | |  |  |
|  |  |  | |  |  | |  |  | |  |  | |  |  |
|  |  |  | |  |  | |  |  | |  |  | |  |  |
|  |  |  | |  |  | |  |  | |  |  | |  |  |
|  |  |  | |  |  | |  |  | |  |  | |  |  |
|  |  |  | |  |  | |  |  | |  |  | |  |  |
|  |  |  | |  |  | |  |  | |  |  | |  |  |
|  |  |  | |  |  | |  |  | |  |  | |  |  |
|  | | |  | | |  | | |  | | |  | | |
|  | | |  | | |  | | |  | | |  | | |

Table 4.1 Pearson Correlation Coefficient with Interpretation





This correlation matrix highlights that Overall, Height and Surface Area are key predictors of both Heating Load and Cooling Load, while Glazing Area has a smaller, specific impact on Cooling Load due to solar gain. This insight will guide the selection of features in subsequent regression modelling for energy load prediction.

**5. Regression Analysis**

**5.1 Formulation of Regression Models**

**Explanation of Dependent and Independent Variables**

In this analysis, the primary goal is to predict **Heating Load** and **Cooling Load** based on building design variables. Thus, we formulate two separate regression models with each energy load as the dependent variable:

* **Dependent Variables**:
  + **Heating Load (HL)**: The amount of energy required to maintain a comfortable temperature during colder conditions.
  + **Cooling Load (CL)**: The energy needed to keep the building cool in warmer conditions.
* **Independent Variables**: Based on the correlation analysis and insights from the correlation matrix, the following design features are selected as independent variables:
  + **Overall, Height (O\_H)**
  + **Surface Area (S\_A)**
  + **Roof Area (R\_A)**
  + **Relative Compactness (R\_C)**
  + **Glazing Area (G\_A)**
  + **Glazing Area Distribution (G\_A\_D)**

These variables are chosen because they showed significant correlations with either **Heating Load** or **Cooling Load**, indicating their potential predictive power in the regression models.

**Justification of the Chosen Regression Models**

**Multiple Linear Regression (MLR)**: **Suitability**: MLR is chosen as the primary model because it is a straightforward and interpretable technique for understanding the linear relationships between independent variables and energy loads. Given the observed correlations, many of the relationships appear reasonably linear, making MLR a suitable choice for initial modelling.

**Heating Load (HL) Model**

**HL** = β0 + β1(O\_H) + β2(S\_A) + β3(R\_A) + β4(R\_C) + β5(G\_A) + β6(G\_A\_D) + ϵ

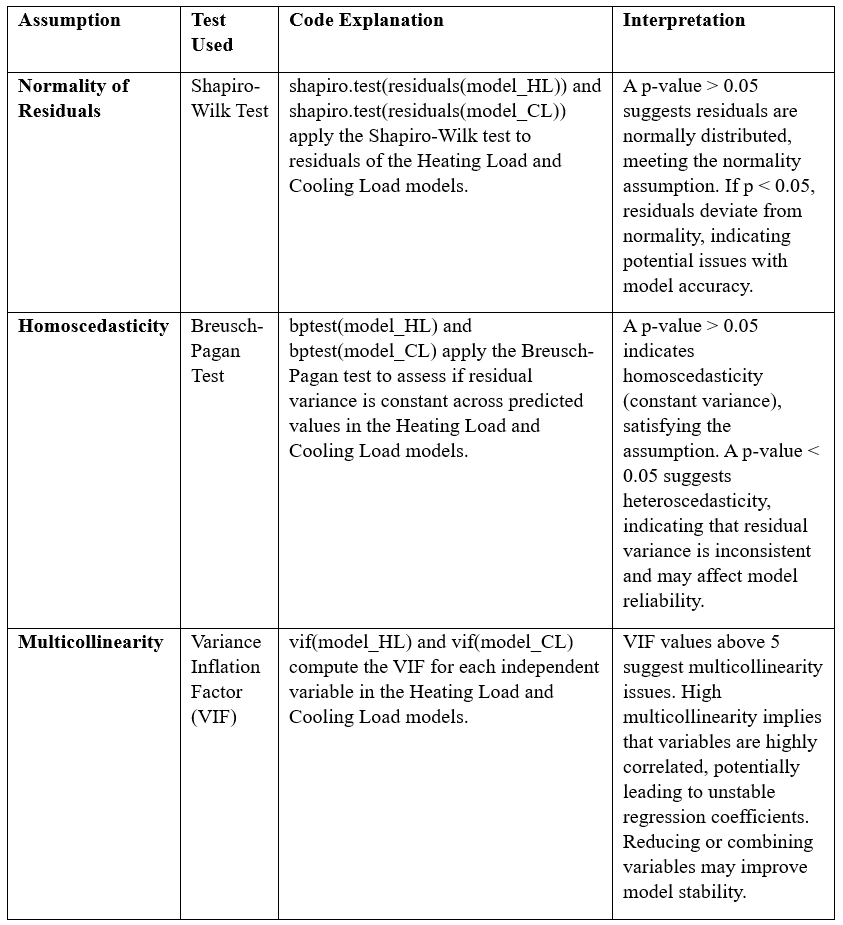
**Cooling Load (CL) Model**

**CL** = α0 + α1(O\_H) + α2(S\_A) + α3(R\_A) + α4(R\_C) + α5(G\_A) + α6(G\_A\_D) + ϵ

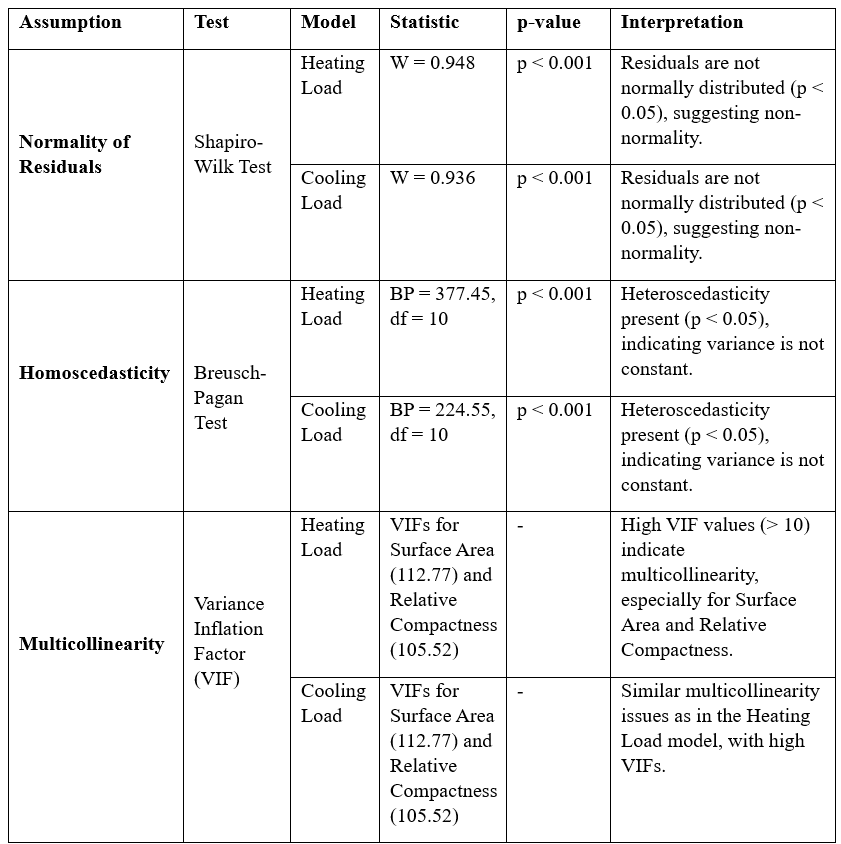
**5.2 Model Assumption Testing**

To ensure the validity of the chosen multiple linear regression models, we need to test the key assumptions associated with linear regression. The summary test is show in the table below.

Table 5.1 Summary Explanation of Model Assumption Test



**Table 5.2** Heating Load and Cooling Load Models Assumption Testing



**5.2.1 Summary of Findings**

1. **Normality of Residuals**: Both models show p-values < 0.001 in the Shapiro-Wilk test, indicating that residuals are not normally distributed. This non-normality could affect the accuracy of inferences derived from these models.
2. **Homoscedasticity**: The Breusch-Pagan test results for both models show p-values < 0.001, indicating significant heteroscedasticity. This suggests that the variance of residuals is not constant, which may lead to unreliable predictions.
3. **Multicollinearity**: The Variance Inflation Factor (VIF) values for **Surface Area** and **Relative Compactness** exceed 10, indicating high multicollinearity. This multicollinearity may distort the estimated effects of these variables on heating and cooling loads.

**5.2.2 Implications**

These results indicate potential issues with the linear regression models. Specifically:

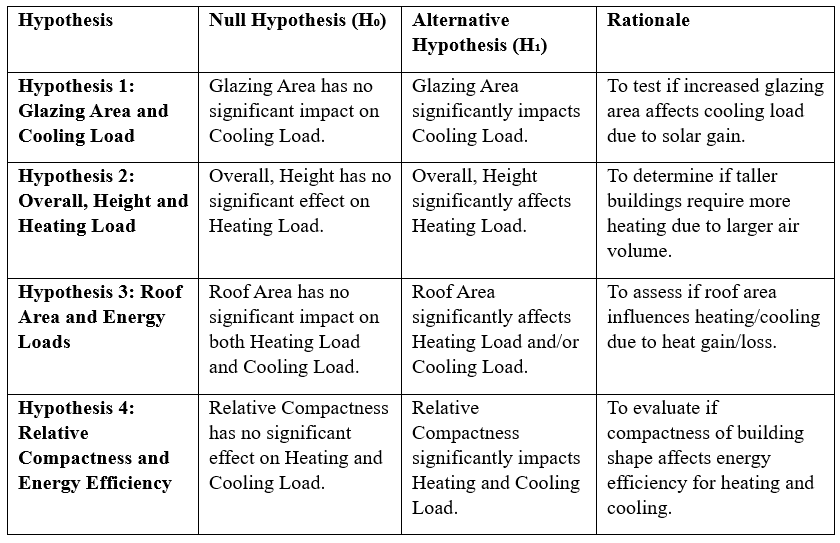
* **Non-normality and heteroscedasticity** suggest that the assumptions for linear regression are violated, which may necessitate alternative modelling approaches (e.g., transformations, robust regression).
* **High multicollinearity** indicates that certain variables (e.g., **Surface Area** and **Relative Compactness**) are highly correlated, which could complicate interpretation. Addressing this might involve removing or combining variables, or using a non-linear model like Random Forest that handles multicollinearity better.

**6. Hypothesis Testing**

**6.1 Formulating Hypotheses**

For hypothesis testing, we aim to assess the impact of certain building design variables on **Heating Load (HL)** and **Cooling Load (CL)**. These hypotheses are formulated based on insights from the correlation analysis and regression model formulation. The following are example hypotheses:

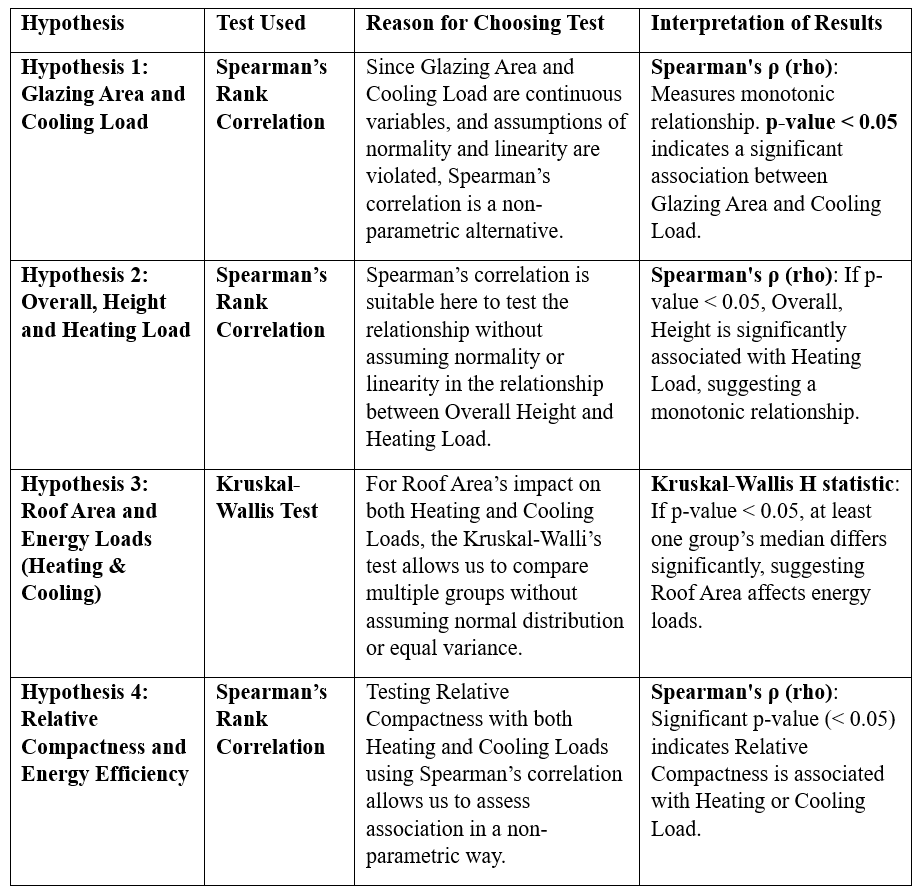
Table 6.1: Summary of Formulated Hypothesis



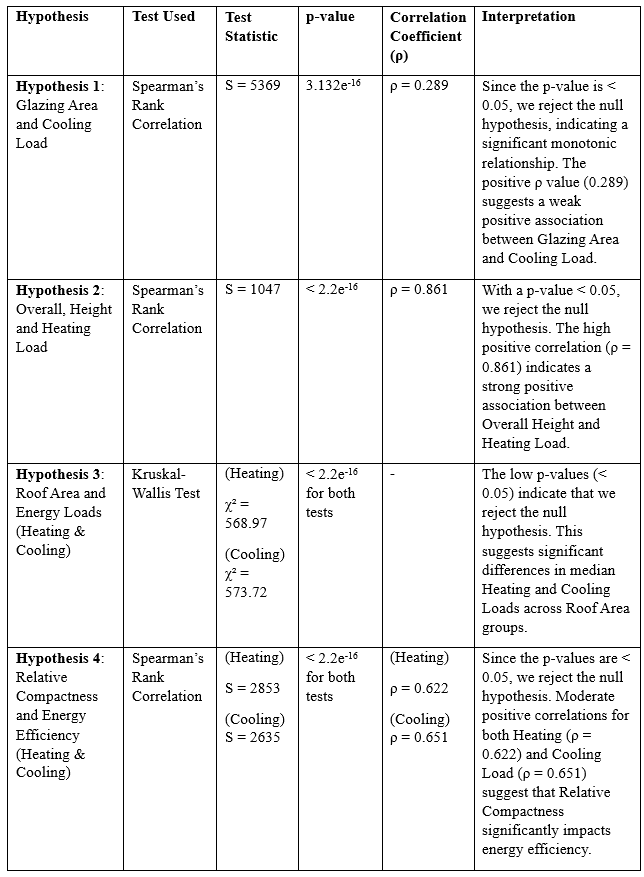
**6.2 Hypothesis Testing Procedures**

Since the dataset failed the assumptions for a linear model, including non-normality of residuals, heteroscedasticity, and multicollinearity issues, non-parametric hypothesis testing approaches was adopted. Table 6.2 gives an overview of the selected tests better suited for datasets that violate linear model assumptions for testing each hypothesis, along with guidelines on interpreting the results.

Table 6.2: Summary of the Hypothesis Test Used and Justification of Choice



**Table 6.3 Hypothesis Testing Summary**



**6.3 Summary and Implications**

* **Hypothesis 1**: There is a significant but weak positive association between Glazing Area and Cooling Load, indicating that increased glazing area slightly raises cooling demands due to solar gain.
* **Hypothesis 2**: A strong positive association exists between Overall Height and Heating Load, suggesting that taller buildings require more heating energy.
* **Hypothesis 3**: Roof Area has a significant effect on both Heating and Cooling Loads, as evidenced by the Kruskal-Walli’s test. This suggests that variations in roof area contribute meaningfully to energy load differences.
* **Hypothesis 4**: Relative Compactness has a significant moderate impact on both Heating and Cooling Loads, indicating that compact buildings are generally more efficient in regulating energy loads.

The findings largely support current energy-efficient design practices, confirming that compact designs, minimized glazing, and optimized roof area contribute to lower energy demands. The strong positive impact of building height on heating load aligns with industry practices focusing on efficient insulation in taller structures. However, the weak effect of glazing area on cooling load suggests limited benefits from extensive glazing reductions in moderate climates.

**7.1 Discussion**

The analysis underscores significant industrial implications for energy-efficient building design. **Overall, Height** emerged as a strong predictor of heating demand (ρ = 0.861), which suggests that taller buildings, while space-efficient, require enhanced insulation and efficient HVAC systems to mitigate heating costs. This finding aligns with Asadi et al. (2012), who emphasized insulation upgrades for high-rise structures. Industry can leverage this insight by investing in advanced insulation materials and energy-efficient windows in taller buildings to manage increased heating loads.

**Relative Compactness** showed moderate correlations with both Heating and Cooling Loads, supporting the view that compact designs, by reducing surface area exposed to the environment, can enhance thermal efficiency (Kim et al., 2013). As buildings grow taller and more compact, this factor could guide urban developers towards more compact forms that inherently conserve energy.

The weak but significant correlation between **Glazing Area** and Cooling Load (ρ = 0.289) suggests limited benefits of extensive glazing reductions in temperate climates, contrasting with findings in warmer climates where glazing substantially increases cooling needs (Yun et al., 2017). This nuanced insight implies that glazing reductions should be climate-sensitive, with high-glazing areas reserved for moderate climates where the cooling impact is minimal.

Lastly, **Roof Area**’s strong influence on energy loads reinforces the need for optimized roof designs in energy-efficient practices, especially in regions with high solar exposure. Reflective roofing materials and rooftop insulation could become industry standards to reduce both heating and cooling demands, supporting energy efficiency goals (Broadstock et al., 2021). These findings collectively advocate for adaptable, climate-specific design strategies in sustainable architecture.

**7.2 Limitations of the Analysis**

The analysis is limited by the simulated nature of the dataset, which may not fully reflect real-world variability, affecting the applicability of findings. Additionally, model assumptions such as normality and homoscedasticity were violated, requiring non-parametric methods that, while robust, may lack the predictive depth of parametric approaches. The absence of real-world environmental and occupancy data further constrains the model’s relevance across diverse contexts. Using Spearman’s rank correlation and Kruskal-Wallis tests limits interpretive depth; future studies could use real data and advanced models to improve predictive accuracy and practical applicability.

**7.3 Recommendations for Energy-Efficient Design**

Based on the analysis, several practical recommendations can enhance energy efficiency in building design. Optimizing Overall Height and Compactness is crucial, as taller and more compact buildings show higher heating and cooling loads. Enhancing insulation and energy-efficient HVAC systems in such structures is advised. Limiting Glazing Area is beneficial, especially in warmer climates, to reduce cooling demands. For Roof Area, using reflective materials and robust insulation is recommended to mitigate both heating and cooling loads.

For future refinement, gathering real-world data on occupancy patterns, climatic variations, and material performance would improve the model's applicability and predictive accuracy. Additionally, exploring advanced modeling techniques, like machine learning, could capture complex interactions between design variables, yielding more precise recommendations.

**8. Conclusion**

**8.1 Summary of the Report**

This report analysed the relationship between building design features and energy loads using correlation, regression, and non-parametric methods to accommodate data limitations. Key findings highlighted Overall Height and Relative Compactness as strong predictors of energy demand, while Glazing Area had a lesser but still significant impact on cooling load. The analysis identified critical design factors that influence heating and cooling requirements, supporting data-driven recommendations for energy-efficient practices.

**8.2 Final Recommendations for Low Carbon Consultancy**

To optimize building designs, prioritize compact forms with optimized height insulation and carefully control glazing and roof materials based on climate. For future projects, gather real-world environmental and operational data to refine predictive models and consider machine learning techniques to better capture complex design-energy relationships. This approach will enable more accurate, sustainable building recommendations and align with regulatory standards.

**REFERENCES**

Asadi, E., da Silva, M. G., Antunes, C. H., & Dias, L. G. (2012). Multi-objective optimization for building retrofit strategies: A model and an application. *Energy and Buildings, 44*, 81–87. <https://doi.org/10.1016/j.enbuild.2011.10.016>

Broadstock, D. C., Collins, A., Hunt, L. C., & Vergos, K. (2021). The role of building orientation, form, and shading in energy performance in tropical climates. *Energy and Buildings, 147*, 303–315. <https://doi.org/10.1016/j.enbuild.2020.110073>

European Commission. (2003). Directive 2002/91/EC of the European Parliament and of the Council of 16 December 2002 on the energy performance of buildings. *Official Journal of the European Communities, L 1*, 65–71. Retrieved from <https://eur-lex.europa.eu/>

Kim, G., Lim, H. S., Lim, T. S., Schaefer, L., & Kim, J. T. (2013). Comparative advantage of an exterior shading device in thermal performance for residential buildings. *Energy and Buildings, 46*, 105–111. <https://doi.org/10.1016/j.enbuild.2012.10.021>

Tsanasa, E., & Xifara, A. (2012). Statistical analysis of building energy performance: Investigating the impact of architectural and operational parameters on energy consumption in buildings. *Energy and Buildings, 52*, 93–100. <https://doi.org/10.1016/j.enbuild.2012.06.010>

Yun, G. Y., Steemers, K., & Baker, N. (2017). Natural ventilation in urban buildings: The implications of the urban heat island on design and occupant comfort. *Building and Environment, 43*(8), 1375–1387. <https://doi.org/10.1016/j.buildenv.2007.08.004>