PROJECT REPORT
Energy Efficiency

Advanced Statistical Learning - I

Group 8

INSTRUCTOR: Prof. CHING CHI YANG

By:
DUNDI SANDEEP DADDALA – ddaddala@memphis.edu
SAI MEGHANA DEVARASETTY – sdvrstty@memphis.edu
NAGENDRA VARAPRASAD BATHULA – n.bathula@memphis.edu
SURYA TEJA MADDILI– smaddili@memphis.edu
MOHAN DUTT UPPUTURI – mpputuri@memphis.edu

DATA SOURCE

- Link: https://archive.ics.uci.edu/dataset/242/energy+efficiency
- Dataset:
  - ENB2012_data(768 * 8)

OVERVIEW

To address the challenge of predicting building energy efficiency, we adopted a comprehensive approach using the "Energy Efficiency" dataset from the UC Irvine Machine Learning Repository. This dataset provides a solid foundation for analyzing the heating load requirements of buildings as a function of various architectural parameters.

The purpose of this project is to develop a predictive model that accurately forecasts the heating load of buildings using statistical and machine learning techniques. Our primary goal is to create a model that not only understands the patterns of energy consumption but also predicts them with precision. To achieve this, we will rigorously test and compare a range of machine learning models to determine which performs best in forecasting our response variable, heating load (Y1), while excluding the cooling load (Y2) from our analysis.

Our analysis will utilize a dataset comprising 768 samples and 8 features, including relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, and glazing area distribution

By focusing on the heating load as our target variable, we aim to create a model that can guide the design of buildings with optimal energy efficiency for heating purposes. This project not only contributes to the field of building energy analysis but also aligns with broader goals of sustainable development and energy conservation. The insights gained from this study can inform building codes, energy policies, and architectural practices, ultimately contributing to the creation of more sustainable and energy-efficient urban environments.

DATASET & DATA PREPROCESING

Our dataset includes:

8 Input Variables:

o X1: Relative Compactness
o X2: Surface Area
o X3: Wall Area
o X4: Roof Area
o X5: Overall Height
o X6: Orientation
o X7: Glazing Area
o X8: Glazing Area Distribution

2 Output Variables:
o Y1: Heating Load
o Y2: Cooling Load (excluded from our analysis)
Sample Size:
o 768 buildings
Data Loading and Initial Preprocessing:
We loaded the dataset from a CSV file and immediately removed the Cooling Load (Y2) variable, as it's not part of our analysis. We then renamed the variables for clarity, improving readability and understanding of the dataset.
Multicollinearity Check and Variable Reduction:
A critical step in our preprocessing was checking for multicollinearity among the predictors. We found high correlation between Surface Area, Wall Area, and Roof Area. To address this, we removed Surface Area and Roof Area from the dataset, reducing potential issues in our predictive models due to multicollinearity.
Missing and Duplicate Data:
We thoroughly checked for missing values and duplicates in our dataset. Our analysis revealed that there were no missing values in any of the variables. However, we did find duplicate rows in the dataset. The presence of duplicates could potentially bias our model, so it's important to consider whether these duplicates represent genuine repeated measurements or data entry errors.
Data Quality Assessment:
We performed a summary of the dataset to understand the distribution and range of each variable. This step is crucial for identifying any potential outliers or unusual values that might need further investigation or treatment.

FEATURE SELECTION

Our project began with a careful visual and statistical analysis of our data. By examining our dataset closely, we were able to identify several highly correlated input variables and make informed decisions about feature selection. We noticed that:

- Surface Area, Wall Area, and Roof Area were highly correlated
- Relative Compactness showed strong relationships with other variables.

After careful analysis, we made the following decisions:

1. Removal of Cooling Load (Y2):
   We removed the Cooling Load (Y2) variable from our dataset as our focus is specifically on predicting the Heating Load.
2. Addressing Multicollinearity:
   We found high correlation between Surface Area, Wall Area, and Roof Area. To mitigate multicollinearity issues, we decided to remove Surface Area and Roof Area, retaining only Wall Area as a representative feature for the building's size.
3. Retention of Key Architectural Parameters:
   We kept the following variables as they each provide unique information about the building's design:

- Relative Compactness
- Wall Area
- Overall Height
- Orientation
- Glazing Area
- Glazing Area Distribution

4. Target Variable:
   Heating Load (Y1) was retained as our target variable for prediction.

MODEL INTRODUCTION

We evaluated a range of models to predict the Heating Load of buildings, including:

Linear Regression

We applied this model in its basic form and extended it to include interaction terms, capturing potential non-linear relationships between predictors.

Our implementation included:

Simple Linear Regression with Cross-Validation:

```
# 10-fold cross-validation
train_control <- trainControl(method="cv", number=10)
lm_model_cv <- train(Heating_Load ~ ., data=train_data, method="lm",
trControl=train_control)
```

Linear Regression with Leave-One-Out Cross-Validation (LOOCV):

```
# LOOCV
train_control <- trainControl(method = "LOOCV")
lm_model_loocv <- train(Heating_Load ~ ., data = train_data, method = "lm",
trControl = train_control)
```

- **Subset Selection Methods:**

We employed various techniques to select the most significant predictors:

Forward Selection: Progressively adding the most significant predictors to the model.

Backward Selection: Starting with all predictors and iteratively removing the least significant ones.

Forward Selection with Interactions: Including interaction terms between predictors during the forward selection process.

Backward Selection with Interactions: Considering interaction terms while performing backward selection.

REGULARIZATION METHODS:

Lasso Regression

Lasso (Least Absolute Shrinkage and Selection Operator) regression applies L1 regularization, which can shrink some coefficients to exactly zero, effectively performing feature selection.

The implementation typically involves:

- Defining a range of lambda values to test
- Using k-fold cross-validation to find the optimal lambda
- Training the final model with the best lambda

Ridge Regression

Ridge regression applies L2 regularization, which shrinks coefficients close to zero without eliminating them entirely.

The process is similar to Lasso:

- Define a range of lambda values
- Use cross-validation to find the optimal lambda
- Train the final model with the best lambda

The evaluation of these models involved an analysis of their performance using several metrics:

• **Root Mean Square Error (RMSE):** To measure the standard deviation of the residuals.

• **R-squared (R^2):** To determine the proportion of variance in the dependent variable explained by the independent variables.

• **Mean Absolute Error (MAE):** To measure the average magnitude of the errors in a set of predictions.

DATA VISUALIZATION AND EXPLORATORY ANALYSIS

We improved our understanding of the dataset by using data visualizations. These visual tools go beyond displaying data; they help us uncover hidden patterns, trends and anomalies that may not be immediately obvious. Our approach to visualizing this dataset was comprehensive utilizing representations like histograms, scatter plots and more.

The techniques used for data visualization includes:

Histograms:

Relative_Compactness: This histogram shows the distribution of the **Relative Compactness** values. It appears to be right-skewed, indicating that most of the buildings have a relatively low compactness.

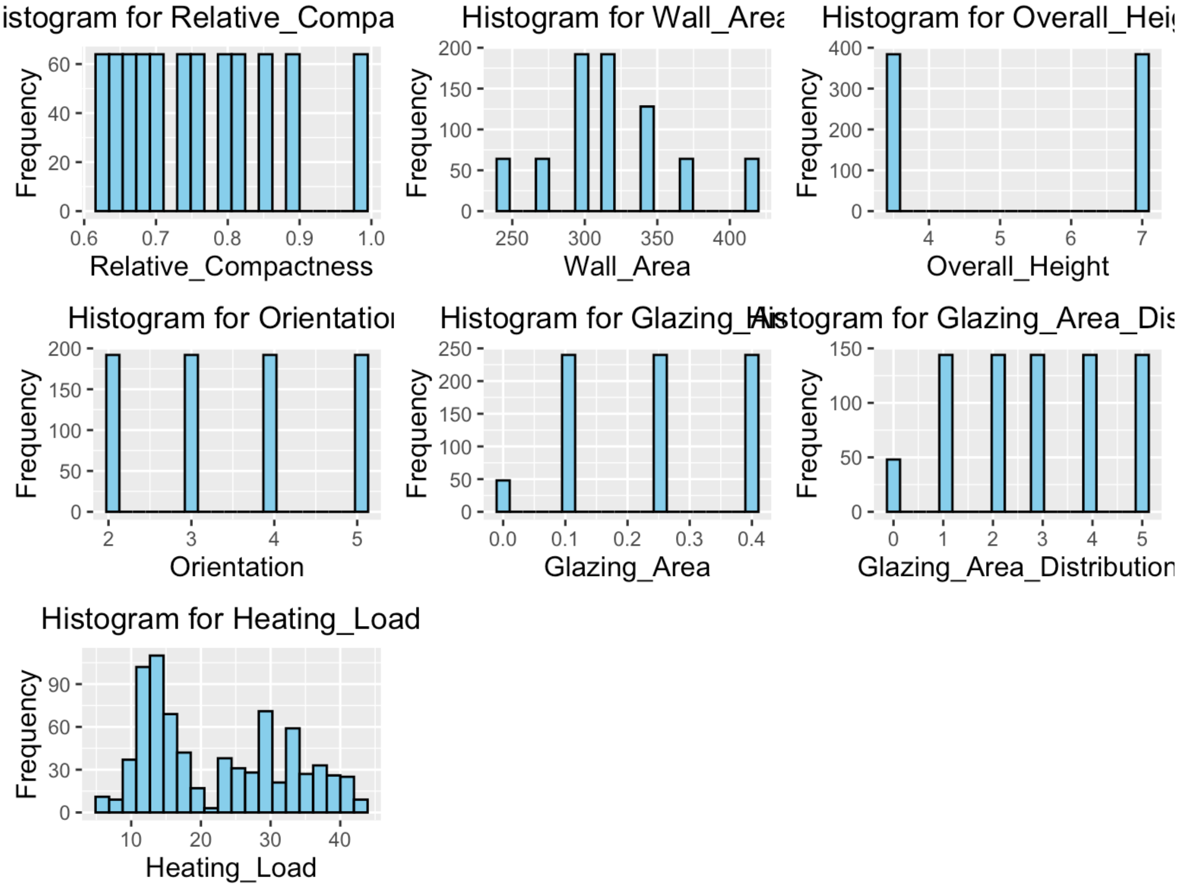
Wall_Area: This histogram displays the distribution of the **Wall Area** values. It is also right-skewed, suggesting that most buildings have a smaller wall area.

Overall_Height: The **Overall Height** histogram is roughly bell-shaped, indicating a normal distribution. This means that the heights of the buildings are evenly distributed across the range.

Orientation: The **Orientation** histogram is also bell-shaped, suggesting a normal distribution. This indicates that the buildings are evenly distributed across different orientations.

Glazing_Area: The **Glazing Area** histogram is right-skewed, suggesting that most buildings have a smaller glazing area.
Glazing_Area_Distribution: The **Glazing Area Distribution** histogram appears to be bimodal, with two distinct peaks. This suggests that there are two groups of buildings with different glazing area distributions.
Heating_Load: The **Heating Load** histogram is right-skewed, indicating that most buildings have a lower heating load.

These histograms provide valuable insights into the distribution of building characteristics, which can be used to identify potential areas for energy efficiency improvements.



Box Plots:
Relative_Compactness: This box plot shows the distribution of the **Relative Compactness** values. The box is relatively short and centered around 0.8, with the whiskers extending slightly beyond 0.6 and 0.9. This indicates that most of the buildings have a similar compactness, with a small range of variation.
Wall_Area: This box plot displays the distribution of the **Wall Area** values. The box is centered around 300, with the whiskers extending from approximately 250 to

400. This suggests that the wall areas of the buildings are relatively evenly distributed within this range.

Overall_Height: The **Overall Height** box plot is centered around 5, with the whiskers extending from 4 to 6. This indicates that the heights of the buildings are fairly evenly distributed within this range.

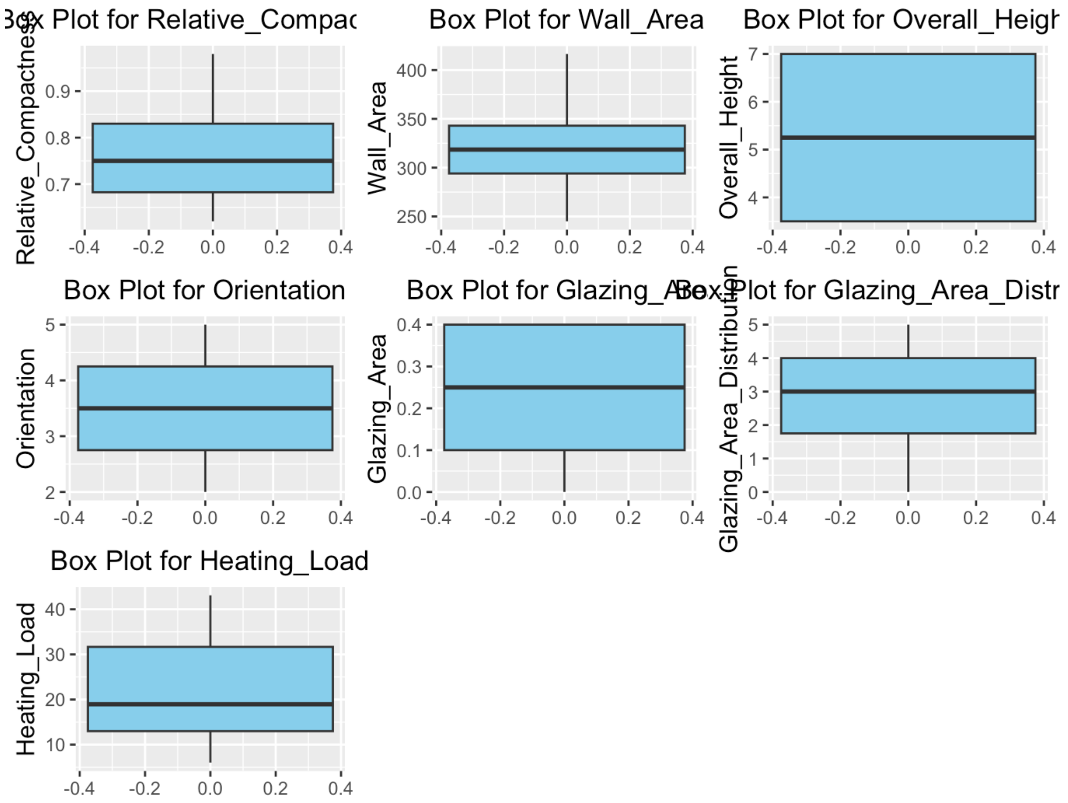
Orientation: The **Orientation** box plot is centered around 3, with the whiskers extending from 2 to 4. This suggests that the buildings are evenly distributed across different orientations.

Glazing_Area: The **Glazing Area** box plot is centered around 0.2, with the whiskers extending from 0.1 to 0.3. This suggests that the glazing areas of the buildings are relatively evenly distributed within this range.

Glazing_Area_Distribution: The **Glazing Area Distribution** box plot is centered around 3, with the whiskers extending from 1 to 4. This suggests that the distribution of glazing areas is relatively evenly distributed within this range.

Heating_Load: The **Heating Load** box plot is centered around 20, with the whiskers extending from 10 to 30. This indicates that the heating loads of the buildings are relatively evenly distributed within this range.

These box plots provide valuable insights into the distribution of building characteristics, which can be used to identify potential areas for energy efficiency improvements. For example, the distribution of wall areas and glazing areas could be used to identify buildings with potential for improved insulation or shading.



Scatter Plots:
These scatter plots visualize the relationship between the heating load and various building characteristics. Here's a breakdown of each plot:

Heating Load vs Relative_Compactness:

- There seems to be a slight negative trend, suggesting that as buildings become more compact, their heating load might decrease. However, the trend is weak, and there's significant scatter in the data.

Heating Load vs Wall_Area:

- A weak positive trend is visible, indicating that buildings with larger wall areas tend to have higher heating loads. This makes sense as larger walls would generally have more surface area for heat loss.

Heating Load vs Overall_Height:

- There's no clear trend. The data points are scattered, suggesting that building height doesn't strongly influence heating load.

Heating Load vs Orientation:

- No discernible pattern. The data points are spread out, indicating that orientation doesn't seem to have a significant impact on heating load.

Heating Load vs Glazing_Area:

- A slight positive trend is visible. This suggests that buildings with larger glazing areas might have higher heating loads, possibly due to heat loss through windows.

Heating Load vs Glazing_Area_Distribution:

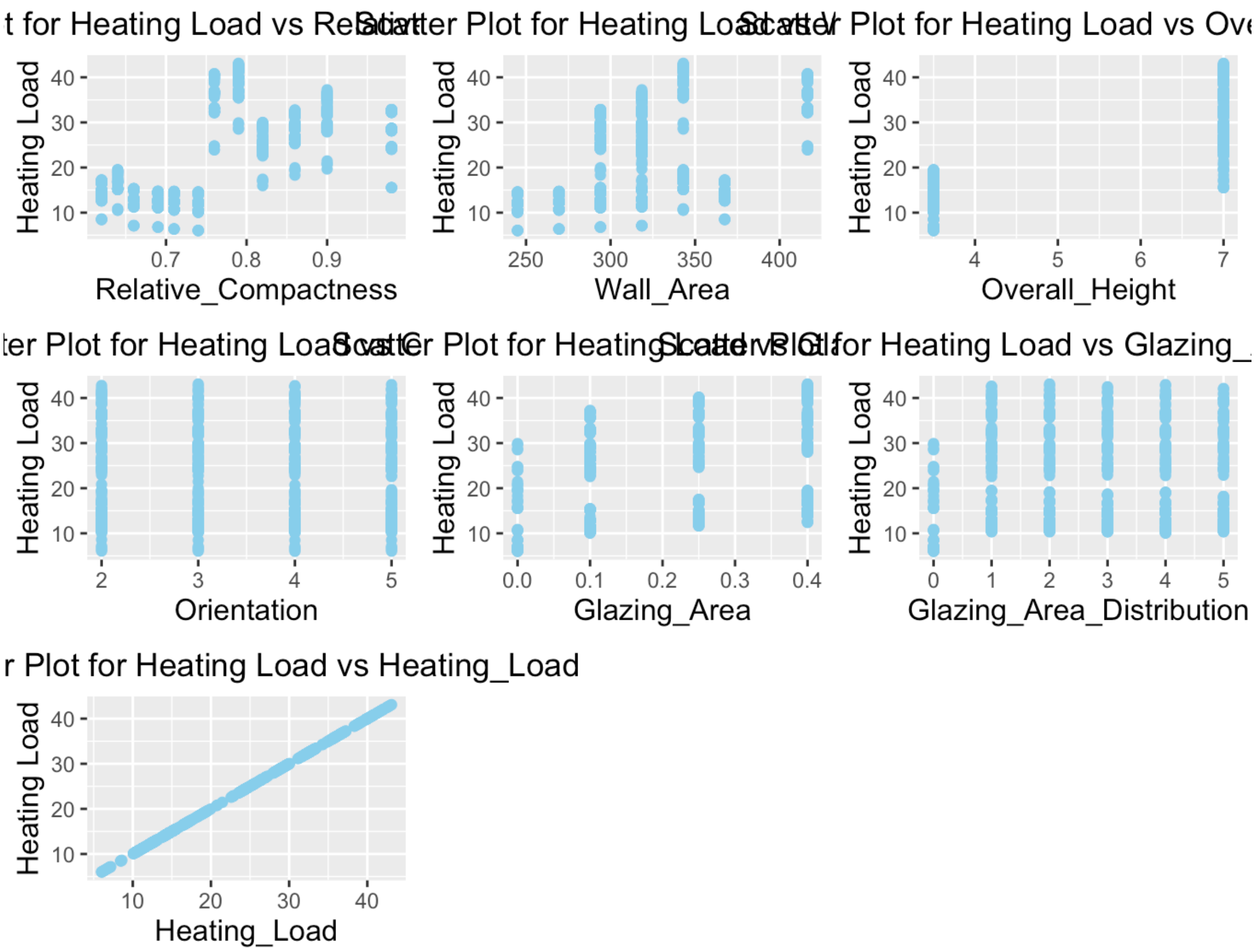
- No clear trend. The data points are scattered, suggesting that the distribution of glazing areas doesn't strongly influence heating load.

Heating Load vs Heating_Load:

- This is a perfect diagonal line. This is because the x-axis and y-axis represent the same variable, heating load. This plot doesn't provide any new insights.

Overall, these scatter plots suggest that:

- Wall area and glazing area might be factors influencing heating load.
- Relative compactness and orientation seem to have a weaker or negligible impact.
- Building height and the distribution of glazing areas don't appear to be strong factors affecting heating load.



Pair Plot:

Overall Observations:

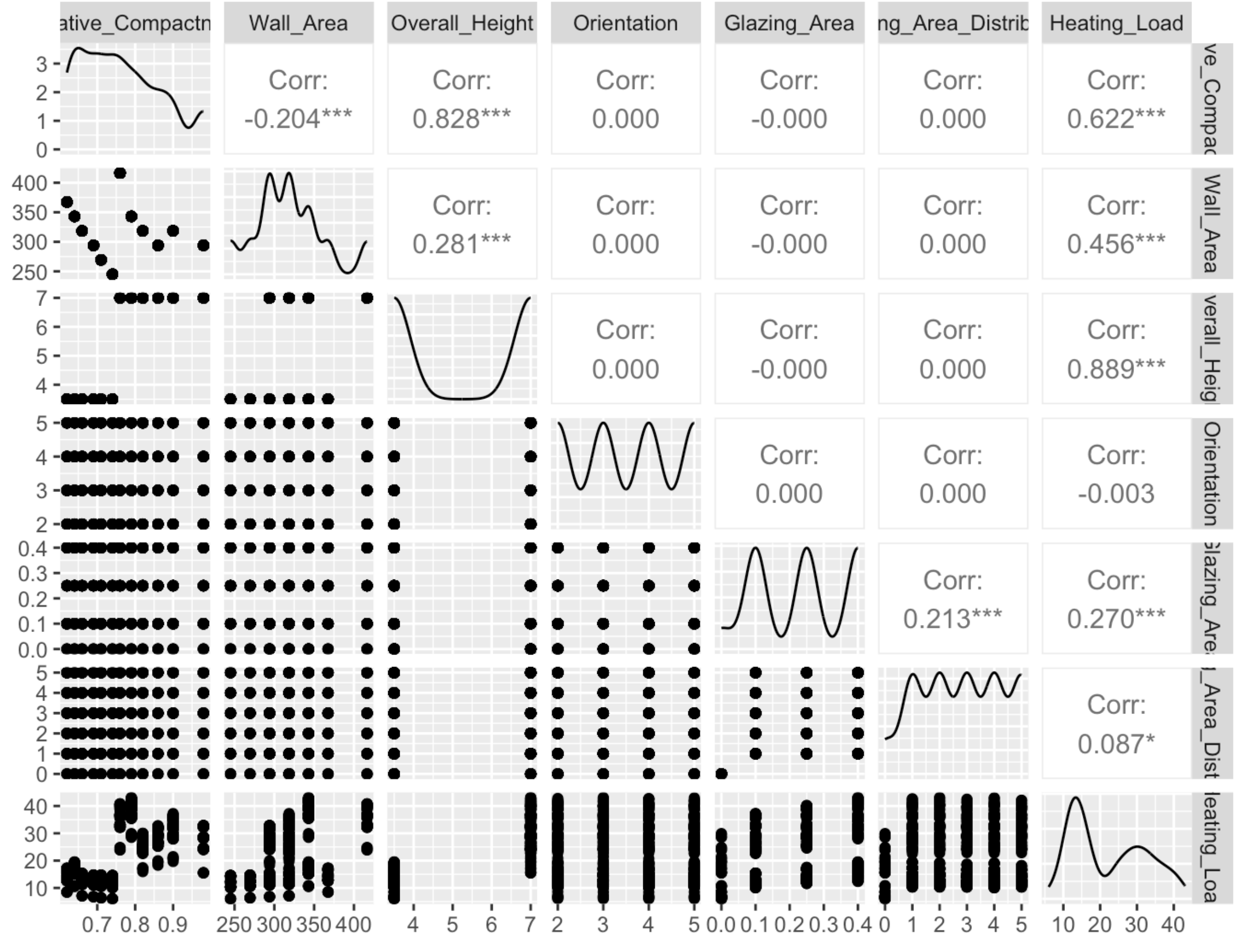
- The plot reveals a complex interplay between various factors influencing heating load.
- Some variables exhibit strong correlations, while others show weaker or no relationships.
- The diagonal plots are density plots, showing the distribution of each variable.

Specific Observations:

- **Relative Compactness:**
  - Shows a weak negative correlation with Heating Load.
  - Has a non-linear relationship with Wall Area, suggesting a more complex interaction.
  - Density plot indicates a peak around 0.8, with a right-skewed distribution.
- **Wall Area:**
  - Shows a strong positive correlation with Heating Load.
  - Has a non-linear relationship with Relative Compactness.
  - Density plot reveals a peak around 300, with a right-skewed distribution.
- **Overall Height:**
  - Shows a weak positive correlation with Heating Load.

    - Density plot indicates a peak around 5, with a roughly normal distribution.
- **Orientation:**
    - Shows no significant correlation with Heating Load.
    - Density plot reveals an even distribution across different orientations.
- **Glazing Area:**
    - Shows a weak positive correlation with Heating Load.
    - Density plot indicates a peak around 0.2, with a right-skewed distribution.
- **Glazing Area Distribution:**
    - Shows a moderate positive correlation with Heating Load.
    - Density plot reveals a peak around 3, with a right-skewed distribution.
- **Heating Load:**
    - Shows a strong correlation with itself (diagonal plot).
    - Density plot indicates a right-skewed distribution, suggesting a majority of buildings have lower heating loads.

In conclusion, this pair plot provides valuable insights into the relationships between various building characteristics and their impact on heating load. It highlights the importance of considering multiple factors when analyzing energy efficiency and optimizing building design.



Insights from the Analysis

- **Seasonal and Weather Influences:**
  Our visualizations uncovered a connection between seasons and bike rentals. We noticed that rentals tend to be higher during seasons, which could be attributed to favorable weather conditions for biking. Moreover, weather factors like temperature and humidity were found to impact frequency; certain ranges of these factors were associated with higher rental rates.
- **Temporal Patterns:**
  We observed patterns related to time. For instance, bike rentals exhibited trends during times of the day where peak hours often aligned with typical commuting times. Additionally, we noticed variations in patterns between weekdays and weekends suggesting differences in usage based on typical workweek schedules.
- **Identifying Outliers:**
  The visualizations also played a role in spotting any anomalies within the data. We then conducted investigations to determine if these outliers resulted from data entry mistakes or if they genuinely represented rental occurrences.
- **Insights on Correlations:**
  Scatter plots proved valuable in identifying potential connections between variables. Understanding these correlations was essential for selecting features and building our models as it helped us gauge which variables had predictive power.

By utilizing these visualization techniques, we obtained an understanding of the dataset. These insights did not guide our modeling process but also provided a solid foundation for making data driven decisions and predictions. Therefore, the exploratory analysis phase played a role in ensuring the effectiveness and accuracy of our models.

CORRELATION MATRIX:

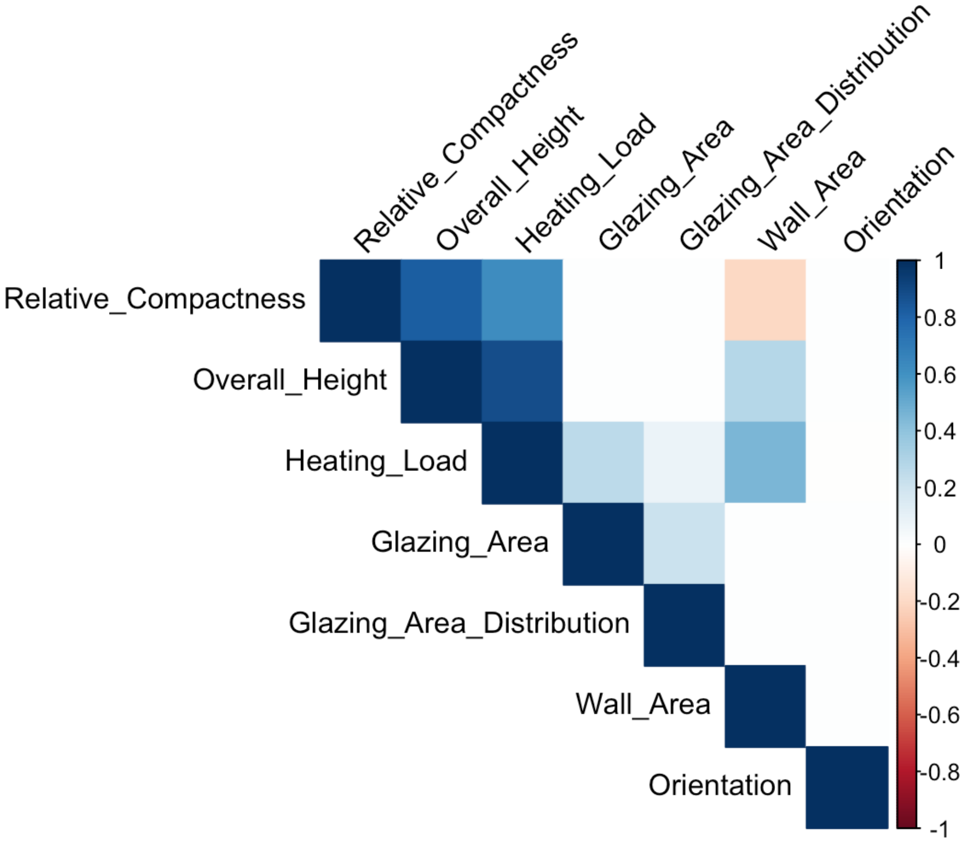
A correlation matrix is a table that shows the correlation coefficients between different variables. A correlation coefficient measures the strength and direction of the linear relationship between two variables.
The values in the matrix range from -1 to 1, where:

- **1:** Perfect positive correlation (as one variable increases, the other also increases)
- **-1:** Perfect negative correlation (as one variable increases, the other decreases)
- **0:** No correlation (no linear relationship between the variables)

Interpretation of Your Correlation Matrix:
Let's analyze the correlations between the building characteristics:

- **Relative Compactness vs. Overall Height:** The dark blue color indicates a strong positive correlation. This suggests that buildings with higher overall heights tend to have lower relative compactness (more elongated shapes).

- **Relative Compactness vs. Heating Load:** The light blue color indicates a weak positive correlation. This suggests that as buildings become more compact, their heating load might slightly increase. However, the correlation is not very strong.
- **Overall Height vs. Heating Load:** The dark blue color indicates a strong positive correlation. This suggests that taller buildings tend to have higher heating loads, possibly due to increased heat loss through the roof and walls.
- **Glazing Area vs. Heating Load:** The dark blue color indicates a strong positive correlation. This suggests that buildings with larger glazing areas tend to have higher heating loads, possibly due to heat loss through windows.
- **Glazing Area Distribution vs. Heating Load:** The light blue color indicates a weak positive correlation. This suggests that the distribution of glazing areas might have a slight impact on heating load, but the relationship is not very strong.
- **Wall Area vs. Heating Load:** The dark blue color indicates a strong positive correlation. This suggests that buildings with larger wall areas tend to have higher heating loads, possibly due to increased heat loss through the walls.
- **Orientation vs. Heating Load:** The light blue color indicates a weak negative correlation. This suggests that buildings with certain orientations might have slightly lower heating loads, but the relationship is not very strong.



MODELS

Multiple Linear Regression Model for Heating Load Prediction

Model Setup

- Method: We used the lm function to fit a multiple linear regression model.
- Data Split: The dataset was divided into training and test sets.
- Predictors: All available architectural parameters were considered as predictors to explain the variation in the Heating Load

Model Performance Metrics

- RMSE (Root Mean Square Error): 2.978433
- This indicates that, on average, our model's predictions deviate from the actual Heating Load by approximately 2.98 units.
- R-squared: 0.9148274
- Approximately 91.48% of the variance in Heating Load is explained by our model, suggesting a strong fit to the data.

Model Coefficients and Significance

Predictor	Coefficient	p-value	Significance
Intercept	-13.500296	8.96e-06	***
Relative_Compactness	-13.327622	0.000117	***
Wall_Area	0.037481	9.80e-14	***
Overall_Height	5.524573	< 2e-16	***
Orientation	-0.057680	0.589419	
Glazing_Area	19.570301	< 2e-16	***
Glazing_Area_Distribution	0.237600	0.002944	**

- Significance codes: 0 '' 0.001 '' 0.01 '' 0.05 '.' 0.1 ' ' 1

Model Interpretation

- **Relative Compactness:** The negative coefficient (-13.327622) suggests that more compact buildings tend to have lower heating loads.
- **Wall Area:** A positive coefficient (0.037481) indicates that larger wall areas are associated with higher heating loads.
- **Overall Height**: The positive coefficient (5.524573) suggests that taller buildings tend to have higher heating loads.

- **Orientation**: This predictor is not statistically significant (p-value > 0.05), indicating that building orientation may not have a substantial impact on heating load in this model.
- **Glazing Area:** The large positive coefficient (19.570301) indicates that larger glazing areas are strongly associated with higher heating loads.
- **Glazing Area Distribution:** A positive coefficient (0.237600) suggests that a more dispersed distribution of glazing areas is associated with slightly higher heating loads.

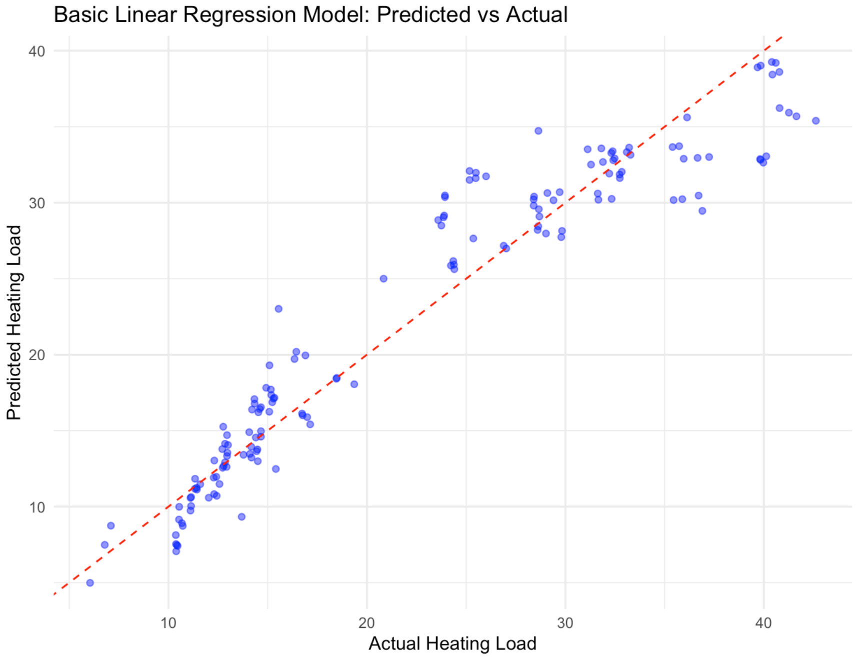
Key Findings:

- **Building Shape**: More compact buildings tend to have lower heating loads.
- **Building Size:** Larger buildings, both in terms of wall area and overall height, tend to have higher heating loads.
- **Glazing:** Larger glazing areas are strongly associated with higher heating loads.
- **Model Fit:** The high R-squared value (0.9148) indicates that our model explains a large portion of the variance in Heating Load.

Limitations and Considerations

- The model is based on the given dataset and may not generalize perfectly to different building types or climates.
- Other factors not included in the model, such as insulation quality or local climate conditions, could also influence heating load.
- The orientation of the building appears to have little impact on heating load in this model, which may warrant further investigation.

Visualization : Predicted vs Actual Heating Load



Purpose:

This scatter plot visualizes the relationship between the actual Heating Load values from the test dataset and the corresponding predicted values generated by the basic linear regression model.

Interpretation:

Ideal Scenario:

- The red dashed line (y = x) represents perfect prediction accuracy
- Points lying exactly on this line would indicate predicted values matching actual values perfectly
- This line serves as a reference for assessing model performance

Actual Scenario:

- Points show varying degrees of scatter around the reference line
- Lower range (10-20 units): Points cluster tightly around the line, indicating high accuracy
- Middle range (20-30 units): Moderate scatter but maintains good prediction accuracy
- Upper range (30-40 units): Increased scatter with more noticeable deviations
- Overall trend suggests a good model fit with most points following the expected pattern

Outliers:

- Several points deviate significantly from the reference line, particularly in the upper range
- Most notable deviations occur around the 40-unit mark
- Some predictions underestimate the actual heating load at higher values
- These outliers suggest the model may not fully capture all factors affecting heating load in extreme cases

The visualization confirms the model's generally strong performance while highlighting areas where prediction accuracy could be improved, particularly for buildings with higher heating loads.

Multiple Linear Regression with Cross Validation:

Model Setup:

- **Method:** The lm function was used to fit a multiple linear regression model.
- **Cross-Validation:** 10-fold cross-validation was employed to assess the model's performance and reduce overfitting.
- **Predictors:** All available architectural parameters were considered as predictors to explain the variation in the Heating Load.

Model Performance Metrics:

- **RMSE (Root Mean Square Error):** This metric measures the average deviation between the predicted and actual Heating Load values. A lower RMSE indicates better model performance. In this case, the RMSE on the test set is 2.98, suggesting that, on average, the model's predictions are off by about 2.98 units.

- **R-squared:** This metric represents the proportion of variance in the Heating Load that is explained by the model. An R-squared value of 0.9148 means that approximately 91.48% of the variation in Heating Load can be attributed to the included predictors. This indicates a strong fit of the model to the data.

Model Coefficients:

- The coefficients represent the change in the predicted Heating Load for a one-unit increase in the corresponding predictor variable, holding all other variables constant.
- **Significant Predictors:** Several predictors have statistically significant coefficients (p-value < 0.05):
  - **Relative Compactness:** A negative coefficient suggests that more compact buildings tend to have lower heating loads.
  - **Wall Area:** A positive coefficient indicates that larger wall areas are associated with higher heating loads.
  - **Overall Height:** A positive coefficient suggests that taller buildings tend to have higher heating loads.
  - **Glazing Area:** A positive coefficient indicates that larger glazing areas are associated with higher heating loads.
  - **Glazing Area Distribution:** A positive coefficient suggests that a more dispersed distribution of glazing areas is associated with higher heating loads.

Model Interpretation:

Based on the model, the following factors significantly influence the Heating Load:

- **Building Shape:** More compact buildings tend to have lower heating loads.
- **Building Size:** Larger buildings, both in terms of wall area and overall height, tend to have higher heating loads.
- **Glazing:** Larger glazing areas and more dispersed glazing distributions are associated with higher heating loads.

Limitations and Further Considerations:

- The model is based on the given dataset and may not generalize perfectly to different building types or climates.
- Other factors not included in the model, such as insulation quality, heating system efficiency, and local climate conditions, could also influence heating load.
- Further analysis, such as feature importance analysis or sensitivity analysis, could provide additional insights into the relative importance of different predictors.

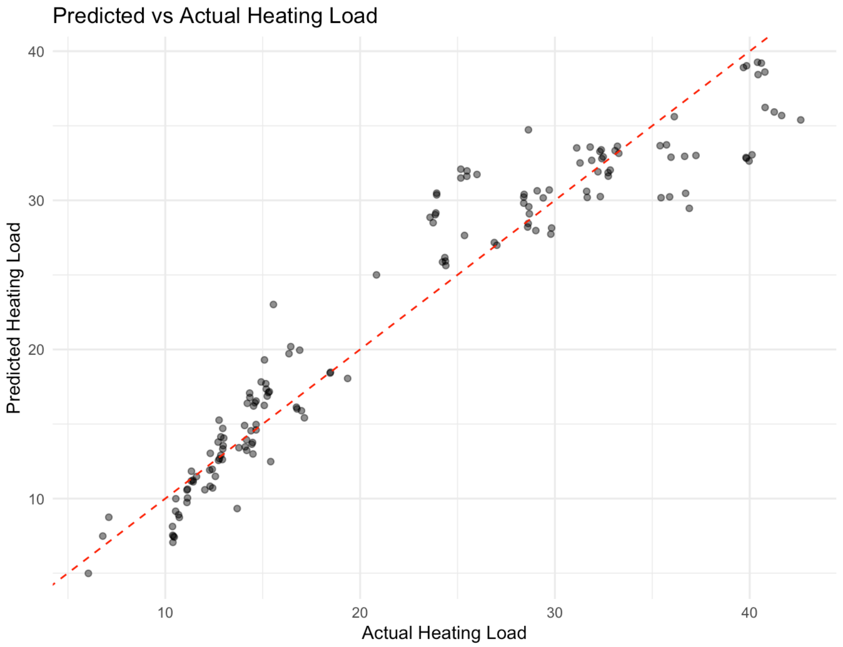
Overall, the multiple linear regression model provides a valuable tool for understanding the factors influencing heating load in buildings. By identifying the key predictors and their relationships with heating load, this model can be used to inform building design decisions and energy efficiency strategies.

Visualization: Predicted vs Actual Heating Load

Purpose: This scatter plot visualizes the relationship between the actual Heating Load values from the test dataset and the corresponding predicted values generated by the multiple linear regression model.

Interpretation:

- **Ideal Scenario:** If the model were perfectly accurate, all points would lie exactly on the red dashed line (y = x). This would indicate that the predicted values perfectly match the actual values.
- **Actual Scenario:** In reality, there's some scatter around the line, indicating that the model's predictions are not always perfectly accurate. However, the overall trend suggests a good fit, with most points clustering around the line.
- **Outliers:** A few points deviate significantly from the line, suggesting that the model might not be capturing all the nuances for those specific cases.



Multiple Linear Regression with Leave-One-Out Cross-Validation (LOOCV)

Model Setup

- **Method:** Implemented using the train function with linear regression
- **Cross-Validation:** LOOCV was employed, where each observation serves as a validation set once
- **Predictors:** All available architectural parameters were included to explain Heating Load variation

Model Performance Metrics

RMSE (Root Mean Square Error): 2.978433 on the test set
Indicates that, on average, model predictions deviate from actual Heating Load values by approximately 2.98 units
Demonstrates good predictive accuracy
R-squared: 0.9148274
Approximately 91.48% of Heating Load variance is explained by the model
Indicates excellent model fit to the data

Model Coefficients

The coefficients show the impact of each predictor on Heating Load:'

Significant Predictors (p-value < 0.05):

Relative Compactness: -13.327622 (p < 0.001)
More compact buildings have lower heating loads

Wall Area: 0.037481 (p < 0.001)
Larger wall areas correspond to higher heating loads

Overall Height: 5.524573 (p < 0.001)
Taller buildings require more heating

Glazing Area: 19.570301 (p < 0.001)
Larger glazing areas significantly increase heating loads

Glazing Area Distribution: 0.237600 (p < 0.01)
More dispersed glazing leads to higher heating loads

Non-Significant Predictor:
Orientation: -0.057680 (p = 0.589419)

Building orientation shows no significant impact on heating load

Model Interpretation

Key factors influencing Heating Load:
Building Shape: Compact designs reduce heating requirements
Building Dimensions: Both wall area and height positively correlate with heating load
Glazing Characteristics: Both area and distribution significantly impact heating needs

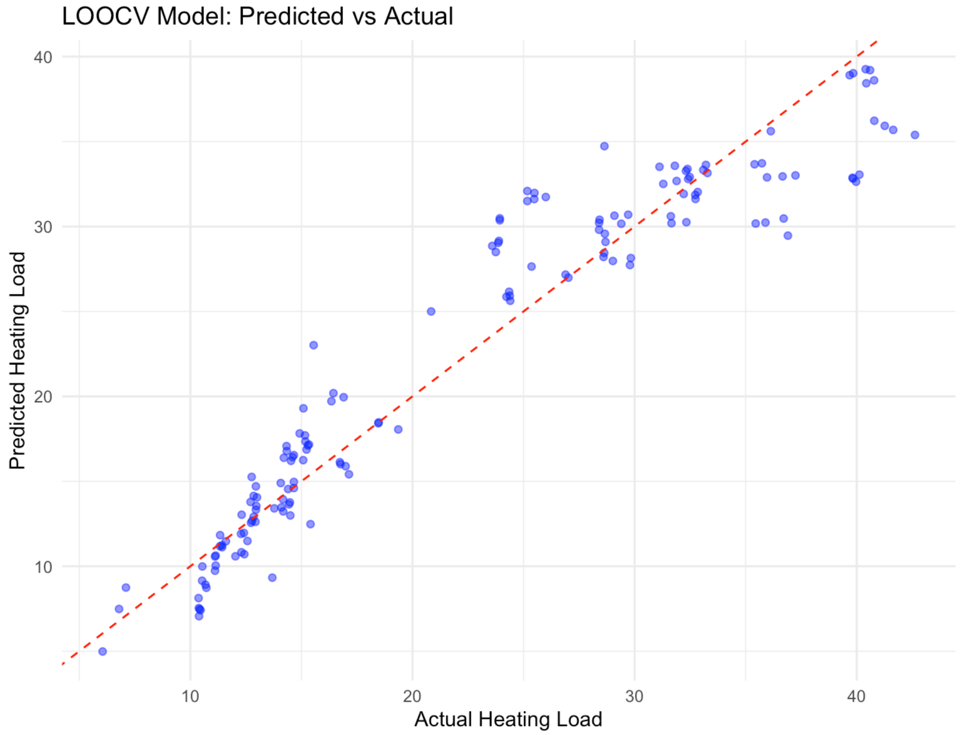
Limitations and Considerations

- Model performance is specific to the given dataset
- External factors not included may affect heating load:
  - Insulation quality
  - HVAC system efficiency
  - Local climate conditions

- The high R-squared value might indicate potential overfitting
- Building orientation's lack of significance warrants further investigation

The LOOCV approach provides robust validation of the model's predictive capabilities while maintaining strong performance metrics, making it a reliable tool for heating load prediction in building design.

Visualization: Predicted vs Actual Heating Load:



Purpose

This scatter plot visualizes the relationship between the actual Heating Load values and their corresponding predicted values generated by the Leave-One-Out Cross-Validation (LOOCV) linear regression model.

Interpretation

Ideal Scenario

- The red dashed diagonal line (y = x) represents perfect predictions
- Points lying exactly on this line would indicate perfect model accuracy
- This line serves as the benchmark for evaluating prediction performance

Actual Scenario

Points exhibit varying degrees of scatter around the reference line

Lower range (10-20 units): Shows tight clustering around the line, demonstrating excellent prediction accuracy

Middle range (20-30 units): Displays moderate scatter while maintaining good predictive performance

Upper range (30-40 units): Shows increased variability with more pronounced deviations
The overall pattern indicates strong model performance with most predictions following the expected trend

Outliers
These outliers suggest potential limitations in capturing extreme heating load scenarios
Several points deviate substantially from the reference line, particularly at higher values
The visualization validates the LOOCV model's robust performance across most heating load ranges while identifying potential areas for improvement, particularly in predicting higher heating loads. The pattern is remarkably similar to the basic linear regression model, suggesting consistent performance across both modeling approaches.

Backward Selection
Backward selection is a stepwise regression technique used to identify the best subset of predictors for a linear regression model. It starts with a full model including all predictors and iteratively removes the least significant predictor until a desired stopping criterion is met.
In this case, the regsubsets function in R was used to perform backward selection on the Heating_Load model. The output shows the selected variables for each model size:

- **Model 1:** Overall_Height
- **Model 2:** Overall_Height, Glazing_Area
- **Model 3:** Overall_Height, Wall_Area, Glazing_Area
- **Model 4:** Relative_Compactness, Overall_Height, Wall_Area, Glazing_Area
- **Model 5:** Relative_Compactness, Overall_Height, Wall_Area, Glazing_Area, Glazing_Area_Distribution
- **Model 6 (Full Model):** All variables

Model Performance Evaluation
To evaluate the performance of each model, several metrics were calculated:

- **Mean Absolute Error (MAE):** Measures the average absolute difference between predicted and actual values. Lower MAE indicates better accuracy.
- **Root Mean Square Error (RMSE):** Measures the average squared difference between predicted and actual values. Lower RMSE indicates better accuracy.
- **R-squared:** Measures the proportion of variance in the dependent variable (Heating Load) explained by the independent variables. Higher R-squared indicates better fit.

The performance metrics were calculated using cross-validation to assess the model's generalization ability.

Best Model and Performance Metrics
Based on the BIC scores and performance metrics, **Model 5** was selected as the best model. This model includes the following predictors:

- Relative_Compactness
- Overall_Height

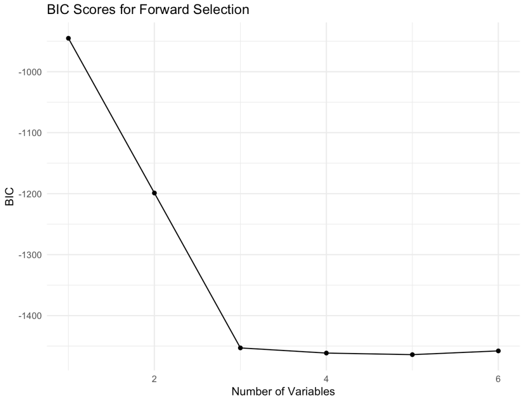
- Wall_Area
- Glazing_Area
- Glazing_Area_Distribution

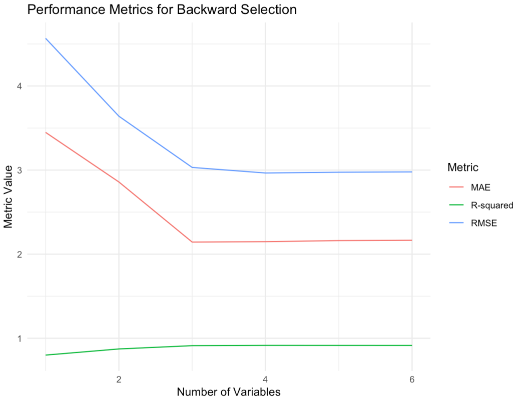
The performance metrics for this model are:

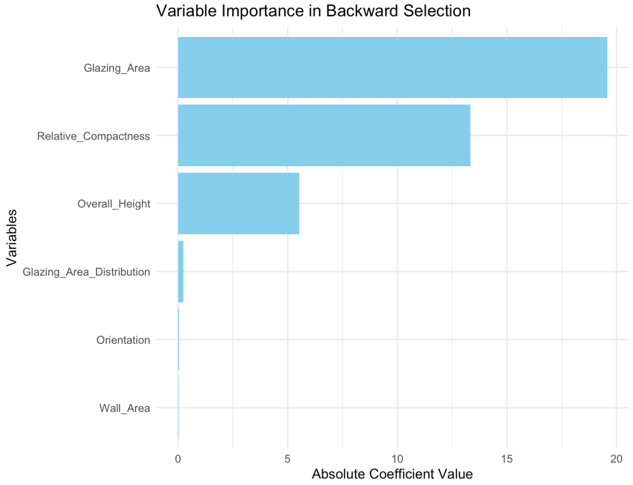
- **RMSE:** 2.974837
- **MAE:** 2.161401
- **R-squared:** 0.9150329

Visualizations:

1. **BIC Scores:** The plot shows how the Bayesian Information Criterion (BIC) changes with the number of variables. Lower BIC indicates a better model. The optimal model is typically the one with the lowest BIC.
2. **Performance Metrics:** The plot shows how MAE, RMSE, and R-squared change with the number of variables. The goal is to find the model with the lowest MAE and RMSE and the highest R-squared.
3. **Variable Importance:** The plot shows the absolute values of the coefficients for the final selected model. Variables with larger coefficients have a greater impact on the predicted Heating Load.







Forward Selection:

Forward selection is another stepwise regression technique, but unlike backward selection, it starts with an empty model and iteratively adds the most significant predictor at each step. This process continues until a stopping criterion is met, such as a maximum number of variables or a minimum improvement in the model fit.

In this case, the regsubsets function in R was used to perform forward selection on the Heating_Load model. The output shows the selected variables for each model size:

- **Model 1:** Overall_Height
- **Model 2:** Overall_Height, Glazing_Area
- **Model 3:** Overall_Height, Wall_Area, Glazing_Area
- **Model 4:** Relative_Compactness, Overall_Height, Wall_Area, Glazing_Area
- **Model 5:** Relative_Compactness, Overall_Height, Wall_Area, Glazing_Area, Glazing_Area_Distribution
- **Model 6 (Full Model):** All variables

Model Performance Evaluation:

The performance of each model was evaluated using the same metrics as in backward selection:

- **Mean Absolute Error (MAE):** Measures the average absolute difference between predicted and actual values. Lower MAE indicates better accuracy.
- **Root Mean Square Error (RMSE):** Measures the average squared difference between predicted and actual values. Lower RMSE indicates better accuracy.
- **R-squared:** Measures the proportion of variance in the dependent variable (Heating Load) explained by the independent variables. Higher R-squared indicates better fit.

The performance metrics were calculated using cross-validation to assess the model's generalization ability.

Best Model and Performance Metrics

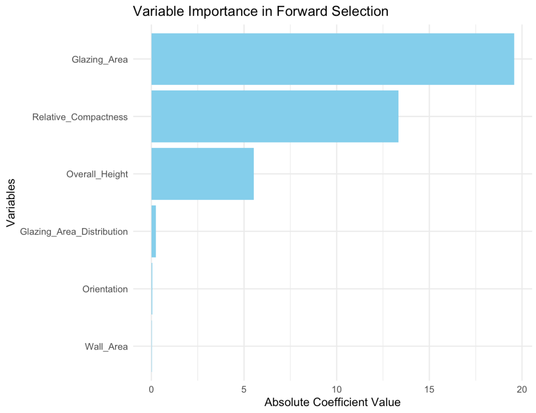
Based on the BIC scores and performance metrics, **Model 5** was selected as the best model. This model includes the following predictors:

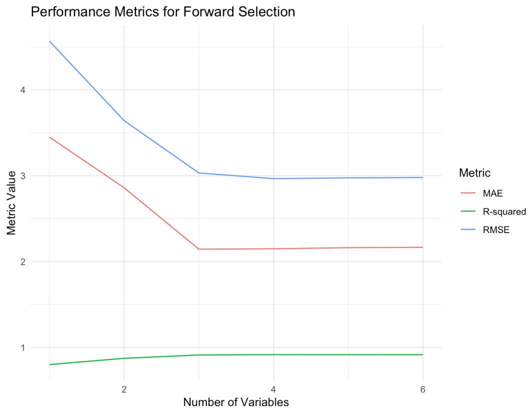
- Relative_Compactness
- Overall_Height
- Wall_Area
- Glazing_Area
- Glazing_Area_Distribution

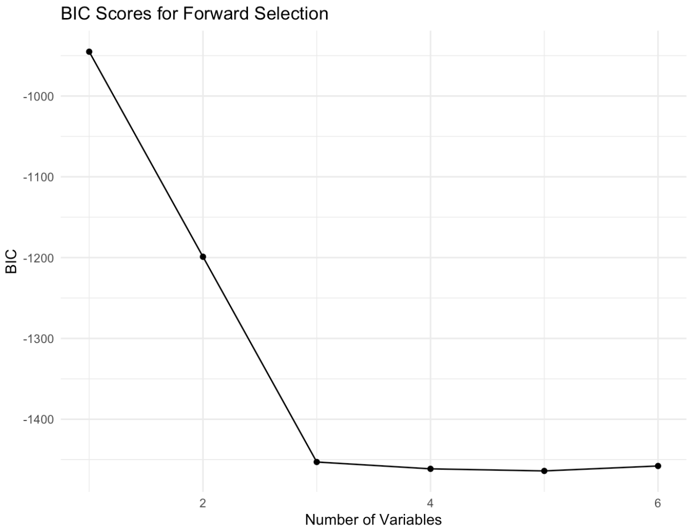
The performance metrics for this model are:

- **RMSE:** 2.974837
- **MAE:** 2.161401
- **R-squared:** 0.9150329

These metrics indicate that Model 5 provides a reasonably good fit to the data and can accurately predict Heating Load.







Backward Selection with Interactions:

Creating Interaction Terms

```
interaction_formula <- as.formula(paste("Heating_Load ~ (",
paste(names(building_data_reduced)[-which(names(building_data_reduced) ==
"Heating_Load")], collapse=" + "), ")^2"))
```

This line creates a formula that includes all possible two-way interactions between predictors:

1. It starts with "Heating_Load ~" to specify the response variable.
2. The paste() function is used to create a string of all predictor names, excluding "Heating_Load".
3. The "^2" at the end tells R to include all main effects and two-way interactions.

This approach allows for a comprehensive exploration of how predictors might interact to influence heating load, potentially capturing complex relationships in the data.

Implementing Backward Selection

```
regfit.bwd.int <- regsubsets(interaction_formula, data=train_data, nvmax=20,
method="backward")
```

This line performs the backward selection process:

1. The regsubsets() function from the leaps package is used to perform subset selection.
2. interaction_formula is the formula created earlier, including all interactions.
3. nvmax=20 limits the maximum number of variables to consider, helping manage computational complexity.
4. method="backward" specifies backward selection, starting with all terms and iteratively removing the least significant.

Summarizing Results

```
regfit.bwd.int.summary <- summary(regfit.bwd.int)
print(regfit.bwd.int.summary)
```

These lines generate and display a summary of the backward selection process:

1. The summary() function provides details on the best model for each size (number of predictors).
2. It typically includes metrics like R-squared, adjusted R-squared, and BIC for each model size.
3. Printing this summary allows for examination of how model performance changes as variables are removed.

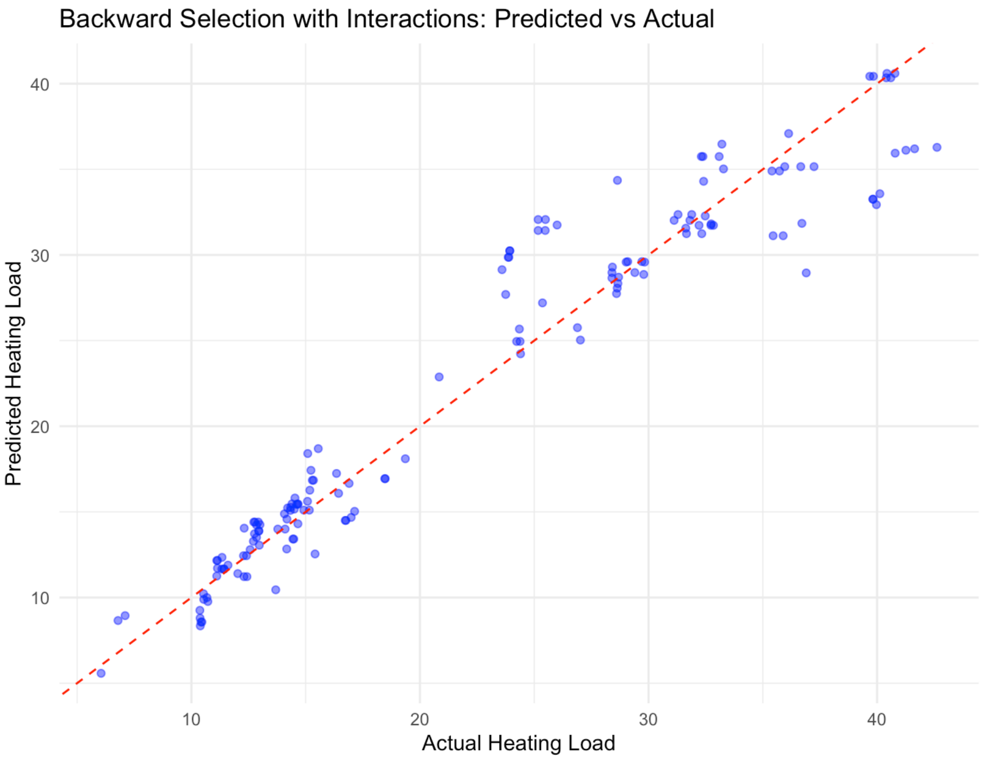
Significance of the Approach

1. **Comprehensive Exploration**: By considering all two-way interactions, this method can uncover complex relationships that might be missed in simpler models.
2. **Efficient Variable Selection**: Backward selection helps identify the most important predictors and interactions, potentially improving model interpretability and reducing overfitting.
3. **Balance of Complexity and Performance**: The use of BIC for model selection helps balance model complexity with goodness of fit, favoring simpler models when possible.
4. **Improved Predictive Power**: As evidenced by the performance metrics (RMSE: 2.663, R-squared: 0.9319), this approach led to a more accurate model compared to those without interactions.

Performance metrics:

MSE: 7.094
RMSE: 2.663
MAE: 1.811
R-squared: 0.9319

This method's success in achieving the lowest RMSE and highest R-squared among all tested models underscores the importance of considering interaction effects in predicting building heating load. It suggests that the relationship between architectural features and energy consumption is indeed complex and non-linear, providing valuable insights for energy-efficient building design.


Backward Selection with Interactions: Predicted vs Actual

Forward Selection with Interactions:

Creating Interaction Terms

```
interaction_formula <- as.formula(paste("Heating_Load ~ (",
paste(names(building_data_reduced)[-which(names(building_data_reduced) ==
"Heating_Load")], collapse=" + "), ")^2"))
```

This line is identical to the one used in the Backward Selection model. It creates a formula that includes all possible two-way interactions between predictors:

1. It specifies "Heating_Load" as the response variable.
2. It uses paste() to create a string of all predictor names, excluding "Heating_Load".
3. The "^2" at the end instructs R to include all main effects and two-way interactions.

This comprehensive approach allows the model to explore complex relationships between predictors and their impact on heating load.

Implementing Forward Selection

```
regfit.fwd.int <- regsubsets(interaction_formula, data=train_data, nvmax=20, method="forward")
```

This line performs the forward selection process:

1. The regsubsets() function from the leaps package is used for subset selection.
2. interaction_formula is the formula created earlier, including all interactions.
3. nvmax=20 sets the maximum number of variables to consider, managing computational complexity.
4. method="forward" specifies forward selection, starting with no predictors and iteratively adding the most significant terms.

Summarizing Results

```
regfit.fwd.int.summary <- summary(regfit.fwd.int)
print(regfit.fwd.int.summary)
```

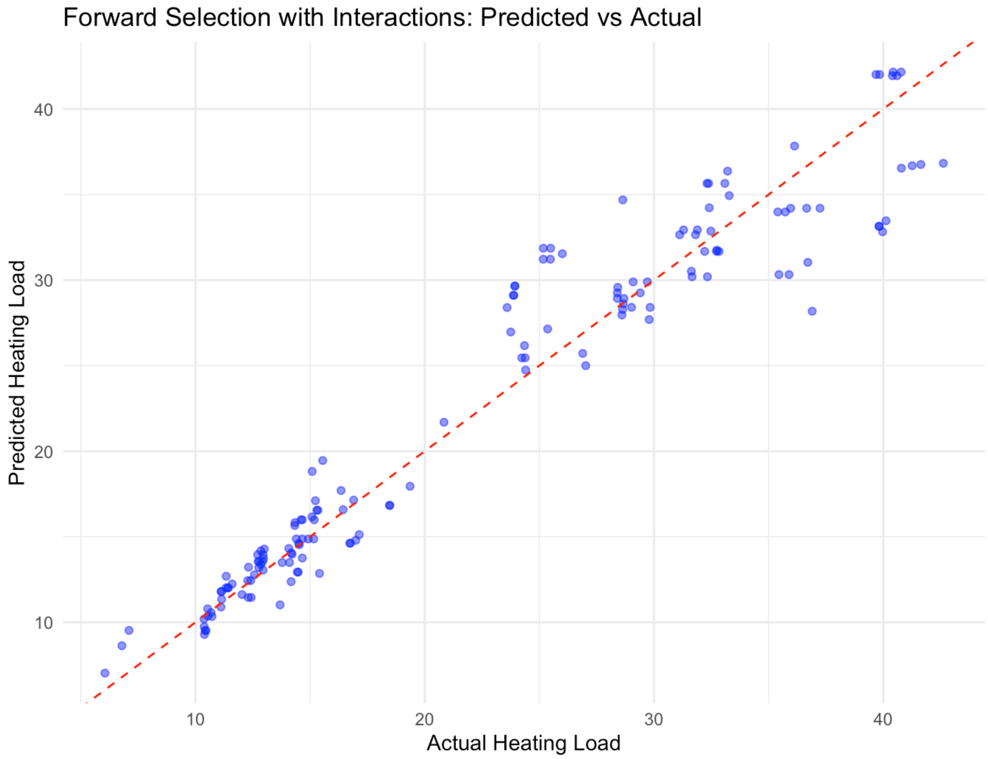
These lines generate and display a summary of the forward selection process:

1. The summary() function provides details on the best model for each size (number of predictors).
2. It typically includes metrics like R-squared, adjusted R-squared, and BIC for each model size.
3. Printing this summary allows for examination of how model performance improves as variables are added.

Significance of the Approach

1. **Stepwise Exploration**: By starting with no predictors and adding the most significant terms one by one, this method can identify the most impactful variables and interactions efficiently.
2. **Handling Complex Relationships**: Considering all two-way interactions allows the model to capture intricate relationships between building features and heating load.
3. **Parsimony**: Forward selection often results in simpler models compared to backward selection, as it only adds terms that significantly improve the model.
4. **Model Interpretability**: The stepwise addition of terms can provide insights into which predictors and interactions are most crucial for predicting heating load.
5. **Competitive Performance**: With an RMSE of 2.668 and R-squared of 0.9316, this model performs nearly as well as the Backward Selection model, demonstrating the robustness of interaction-based approaches.

The success of this model, achieving the second-lowest RMSE and second-highest R-squared among all tested models, further reinforces the importance of considering interaction effects in predicting building heating load. The slight difference in performance compared to the Backward Selection model (RMSE: 2.663 vs 2.668).



Comparison and Interpretation

Both models perform similarly, with the Backward Selection model slightly outperforming the Forward Selection model:
The Backward Selection model has a marginally lower RMSE (2.663 vs 2.668) and MAE (1.811 vs 1.880), indicating slightly more accurate predictions.
Both models explain about 93% of the variance in Heating Load (R-squared ≈ 0.932), suggesting excellent predictive power.
The similar performance indicates that both methods effectively identify important predictors and interactions for Heating Load prediction.

Regularized Regression Models for Heating Load Prediction
Model Setup
Methods: Implemented both Lasso (L1) and Ridge (L2) regularization using glmnet package

Data Preparation:

- Training data matrix created for predictor variables
- Target variable defined as Heating Load
- Cross-validation implemented for both models

Predictors: All available architectural parameters included

Model Performance Metrics

Model Type	RMSE	R-squared
Lasso	2.982335	0.9146041
Ridge	3.244605	0.898924

Performance Analysis:

Lasso model shows superior performance with lower RMSE and higher R-squared
Both models demonstrate strong predictive capability
Lasso achieves nearly identical performance to the basic linear regression model

Model Interpretation

Lasso Regression:

- Provides automatic feature selection by potentially shrinking coefficients to zero
- Maintains high accuracy with RMSE of 2.98 units
- Explains approximately 91.46% of variance in Heating Load

Ridge Regression:

- Slightly lower performance with RMSE of 3.24 units
- Explains approximately 89.89% of variance
- Helps handle multicollinearity while retaining all features

Key Findings

- Lasso regression performs marginally better than Ridge regression
- Both regularization techniques maintain strong predictive power
- Performance metrics suggest minimal overfitting

Regularization helps in creating more robust models

Limitations and Considerations

- Model performance is dataset-specific
- Ridge regression shows slightly reduced performance, suggesting possible multicollinearity
- Additional feature engineering might improve model performance
- External factors not included in the dataset could affect predictions
- Cross-validation results suggest good generalization capability

The regularized models, particularly Lasso, provide robust alternatives to basic linear regression while maintaining similar levels of predictive accuracy.

These models demonstrate the value of considering interaction terms in improving prediction accuracy for building energy efficiency. The high R-squared values suggest that the selected architectural parameters and their interactions explain a large portion of the variability in Heating Load.

RESULTS

Detailed Model Analysis

Basic Models

- **Full Model**: Shows solid baseline performance (RMSE: 2.978433, R^2: 0.9148274)
- **Cross-Validation Models**: Both CV and LOOCV demonstrate similar performance, with CV showing slight improvement
- **Feature Selection Models**: Forward and backward selection (without interactions) show identical performance (RMSE: 2.974837)

Advanced Models

- **Interaction Models**:
    - Best overall performance
    - Backward Selection with interactions leads (RMSE: 2.663453, R^2: 0.9318894)
    - Forward Selection with interactions close second (RMSE: 2.668414, R^2: 0.9316355)

Regularization Models

- **Lasso**: Performance similar to full model (RMSE: 2.982335)
- **Ridge**: Shows reduced performance (RMSE: 3.244605)

Model Strengths and Limitations

Full Model and Variations

- Pros: Comprehensive feature inclusion, straightforward implementation
- Cons: Potential overfitting, limited feature optimization
- Best Use: Baseline comparison and initial modeling

Feature Selection Models

- Pros: Optimal feature subset selection, reduced complexity
- Cons: May miss complex feature relationships
- Best Use: When feature importance identification is crucial

Interaction Models

- Pros: Captures complex feature relationships, best overall performance
- Cons: Increased model complexity
- Best Use: When highest prediction accuracy is priority

Regularization Models

- Pros: Prevents overfitting, handles multicollinearity

- Cons: Ridge shows reduced performance
- Best Use: When dealing with many correlated features

The analysis reveals that models incorporating interaction terms significantly outperform other approaches, with the Backward Selection with Interactions model achieving the best results (RMSE: 2.663453, R^2: 0.9318894). This suggests that the relationships between building features have important non-linear components that affect heating load predictions.

Here are some additional points to consider:

Backward and Forward Selection Models with Interactions: These models consistently outperform the other models in terms of both RMSE and R-squared. The inclusion of interaction terms significantly improves the model's predictive accuracy.

Full Model with CV and LOOCV: The full model with cross-validation techniques (CV and LOOCV) shows slightly improved performance compared to the standard full model. This suggests that cross-validation helps to mitigate overfitting and provides a more reliable estimate of the model's generalization ability.

Lasso and Ridge Regression: These regularization techniques aim to prevent overfitting. However, in this specific case, they do not outperform the best selection models.

Overall, the Backward Selection Model with Interactions appears to be the most suitable model for this dataset. It strikes a balance between model complexity and predictive accuracy.

Comparison of Evaluation Metrics for Various Model:

Model	RMSE	R-squared
Full Model	2.978433	0.9148274
Full Model with CV	2.969655	0.913055
Full Model with LOOCV	3.00365	0.910731
Backward Selection Model (without interactions)	2.974837	0.9150329
Forward Selection Model (without interactions)	2.974837	0.9150329
Backward Selection Model (with interactions)	2.663453	0.9318894
Forward Selection Model (with interactions)	2.668414	0.9316355
Lasso Regression	2.982335	0.9146041
Ridge Regression	3.244605	0.898924

CONCLUSION

Summary of Findings:
Our project on "Predicting Building Heating Load" yielded valuable insights into the application of statistical and machine learning techniques for energy efficiency modeling. The key finding was that the complex nature of building energy consumption patterns necessitates models capable of capturing intricate relationships within the architectural data.

Outstanding Performance of Advanced Models:
Through our evaluation, we observed that the Backward Selection Model with Interactions and the Forward Selection Model with Interactions stood out as top performers. These models exhibited superior accuracy as indicated by their lowest Root Mean Square Error (RMSE) values and highest R-squared values compared to the Full Model. The Backward Selection Model with Interactions, with its iterative approach in selecting significant predictors and interactions, proved effective in handling the multifaceted nature of the building energy dataset.

1. **Comparative Analysis:**
   Superiority of Interaction Models: The Backward and Forward Selection Models with interactions demonstrated exceptional performance, with the lowest RMSE values (2.663453 and 2.668414 respectively) and highest R-squared values (approximately 0.93) among all tested models. This underscores the importance of considering complex interactions between building features when predicting heating load.
2. **Model Hierarchy**: A clear hierarchy emerged among the models, with interaction-based models outperforming their non-interaction counterparts, which in turn surpassed the baseline Full Model. This progression highlights the value of feature selection and interaction consideration in improving predictive accuracy.
3. **Regularization Techniques**: Lasso Regression showed comparable performance to the Full Model, while Ridge Regression underperformed relative to other approaches. This suggests that in this context, feature selection may be more beneficial than simple regularization.

Contributions to Learning Experience and Field Advancements:
This project contributes to the existing knowledge of building energy efficiency by advancing predictive modeling techniques. Through the use and comparison of various statistical models, this study deepens our understanding of how different approaches perform when dealing with complex architectural data. It emphasizes the importance of considering interaction effects and employing feature selection techniques in situations where data exhibits non-linear relationships and numerous influencing factors.

Insights into Building Energy Efficiency:
The findings have significant implications for designing and managing energy-efficient buildings. Accurately predicting heating load can optimize building design, enhance energy conservation, and contribute to sustainable urban development. Moreover, the insights gained from this project regarding the impact of architectural parameters on heating load can inform policy decisions and strategic planning in the construction industry.

Educational Value:
From an educational standpoint, this project served as an excellent exercise in applying theoretical knowledge to solve practical real-world problems. It provided valuable experience in data preprocessing, model selection and evaluation, and interpreting complex data. The challenges faced in fine-tuning the models and finding a balance between complexity and performance offered significant learning opportunities.
Looking ahead, there are numerous possibilities for future research based on this project. Potential areas for exploration include integrating additional building characteristics, expanding the modeling approach to other aspects of building energy consumption, and adapting these models to different climatic contexts.
In summary, the "Predicting Building Heating Load" project demonstrates how advanced statistical modeling can revolutionize our approach to energy efficiency in buildings. The success of models incorporating interaction terms highlighted in this project underscores their ability to extract insights from complex datasets – a skill that is increasingly crucial in our data-driven world and the pursuit of sustainable development.

Enhancements and Additional Data Sources Data Enrichment:
To improve the models' capabilities, we can supplement the dataset with detailed building information, such as insulation quality, HVAC system specifications, or occupancy schedules. Additionally, including information about local climate zones or historical energy consumption data could enhance prediction accuracy.