HEATING LOAD PROJECT

SID: 530062808
Submission date: 25 September 2024

I. Executive Summary

Efficient energy management is crucial for optimizing building operations, particularly in light of increasing energy consumption. In Australia, energy use rose by 2.0% in 2022–23, reaching 5,882 petajoules, highlighting the urgent need for effective energy solutions (Department of Climate Change, Energy, The Environment and Water, 2023). This report aims to develop a predictive model for Heating Load, which quantifies the daily energy required to maintain comfortable indoor temperatures in buildings.

We analyze a dataset that incorporates key factors, including building characteristics, environmental conditions, and occupancy, as outlined in Table 1. Our approach involves employing various model and variable selection techniques, such as Linear Regression, Forward and Backward Selection, K-Nearest Neighbors, Lasso, and Ridge Regression. We compare these models using metrics such as Root Mean Squared Error, Mean Squared Error, Mean Absolute Error, and Adjusted R-squared.

Ultimately, our goal is to mitigate excessive energy consumption, optimize heating systems, and promote sustainable energy practices in building management. This model will enhance energy efficiency consulting efforts, facilitating more strategic and cost-effective operations of heating systems.

Variable	Description
HeatingLoad	Total daily heating energy required (in kWh)
BuildingAge	Age of the building (in years)
BuildingHeight	Height of the building (in meters)
Insulation	Insulation quality (1 = Good, 0 = Poor
AverageTemperature	Average daily temperature (in °C)
SunlightExxposure	Solar energy received per unit area (in W/m^2)
WindSpeed	Wind speed at the building's location (in m/s)
OccupancyRate	Proportion of the building that is occupied (percentage)

Table 1. Description of Variables

II. Exploratory Data Analysis (EDA)

1. Data Understanding and Cleaning

The dataset contains 10,000 observations representing 10,000 buildings and includes 8 variables: one response variable, Heating Load, and seven predictors - Building Age, Building Height, Insulation, Average Temperature, Sunlight Exposure, Wind Speed, and Occupancy Rate - used to predict Heating Load. Among the variables, 7 are numerical, while Insulation is categorical, with 0 for low-quality insulation and 1 for good-quality insulation.

An initial assessment of the dataset using data_train.isnull().sum() reveals no missing values, eliminating the need for data cleaning in this regard (Appendix A). Furthermore, boxplots in Figure 1 and Appendix B illustrate no substantial outliers in the response variable or predictors, confirming the dataset's suitability for further analysis.

To mitigate order bias during model development, the dataset was shuffled using random_state=50 prior to analysis. This ensures the model learns general patterns in the data rather than memorizing any specific order of observations, contributing to more reliable and robust predictive performance (Dutta, 2024). Additionally, rather than splitting the data into training and validation sets, cross-validation will be consistently used throughout the analysis to evaluate model performance and preserve data integrity (Chanaka Prasanna, 2024).

2. Response Variable Y

The exploratory analysis of Heating Load in the dataset of 10,000 observations reveals that the average daily heating energy requirement is 260.07 kWh, with a range from 173.68 kWh to 793.92 kWh. Key statistics include a lower quartile (Q1) of 218.09 kWh, a median of 236.8 kWh, and an upper quartile (Q3) of 270.35 kWh, indicating that 50% of observations fall between 218.09 kWh and 270.35 kWh. The variance is 5,663.95, and the standard deviation is 74.59, reflecting considerable variability in the data.

The distribution analysis shows a right-skewed pattern, with histograms and boxplots indicating that most buildings require heating energy below 250 kWh. This skewness suggests the presence of a relatively small number of buildings that require exceptionally high energy daily, which extends the distribution's tail to the right.

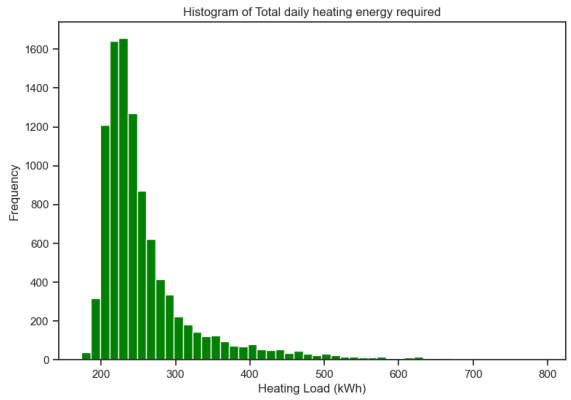


Figure 1. Histogram of HeatingLoad

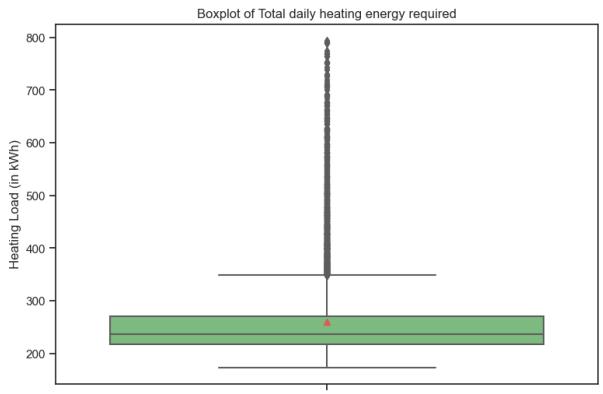


Figure 2. Boxplot of HeatingLoad

3. Categorical Predictor - Insulation

Since there is only one categorical variable—Insulation Quality—in the dataset, a brief evaluation will be conducted to examine any significant insights. The dataset contains 10,000 observations, with 5,960 buildings featuring high-quality insulation and 4,040 with low-quality insulation. An evaluation of Heating Load through the box plot in Appendix C, reveals no significant difference in mean heating load between these groups. This finding is further supported by the histogram of Heating Load based on Insulation Quality (see Appendix D).

4. Relationship Between Predictors Versus Response Variable

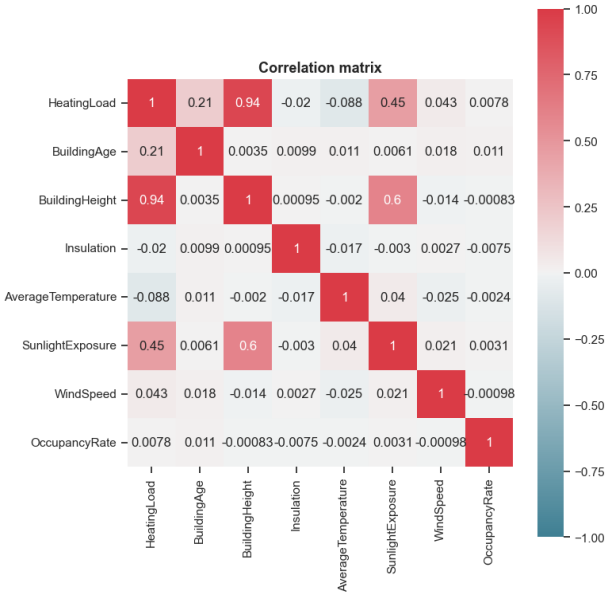


Figure 3. Heatmap of All Variables

The heatmap analysis reveals that the predictors most strongly correlated with Heating Load are Building Height (0.94), Sunlight Exposure (0.45), and Building Age (0.21), while the remaining predictors show low correlation. Most variables exhibit a positive correlation with Heating Load, except for Insulation and Average Temperature, which have a weak negative correlation.

Additionally, Building Height and Sunlight Exposure are highly correlated (0.6), raising potential collinearity concerns when included in the same regression model. Therefore, it is advisable to further examine this relationship using a more precise method, such as the Variance Inflation Factor (VIF).

Figure 5 and the pair plot in Appendix E illustrates a positive but not entirely linear relationship between Building Height and Heating Load. As Building Height increases, Heating Load rises at a decreasing rate, creating a convex relationship. To better capture this relationship, we incorporate the square of Building Height into our analysis.

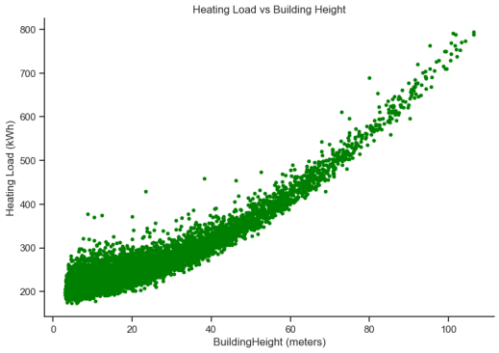


Figure 4. BuildingHeight vs HeatingLoad

5. Predictors in Test Dataset

The test dataset contains 4,944 observations. Without the response variable, we performed a brief exploratory analysis on the predictor distributions. The boxplots show that the distribution shapes of numerical variables closely match those in the training dataset (in Appendices B and F). Additionally, no significant outliers or errors were detected in the boxplots that could adversely affect the performance of our final model on the test set.

III. Modeling Selection and Evaluation

To ensure consistency in the model selection process, a 10-fold cross-validation procedure will be employed. Two key metrics, Mean Squared Error (MSE) and Mean Absolute Error (MAE), are used to evaluate and compare the performance of the models, providing reliable estimates of prediction error.

1. Linear Regression

1.1. Model 1: Basic Multiple Linear Regression

The first model is a standard Multiple Linear Regression (MLR) that includes all available predictors: Building Age, Building Height, Insulation, Average Temperature, Sunlight Exposure, Wind Speed, and Occupancy Rate. This model serves as a baseline for evaluating overall performance. Cross-Validation results in an average MSE of 287.64 and average MAE of 12.11.

Given the relatively high correlation (0.6) between BuildingHeight and SunlightExposure, we assessed potential collinearity using the Variance Inflation Factor (VIF). With a VIF of 1.558 (well below the threshold of 10), there is no evidence of multicollinearity affecting the model (CFI , 2022).

1.2. Model 2: MLR with Building Height Square

Building on the exploratory analysis, BuildingHeight exhibits a nonlinear relationship with HeatingLoad, suggesting a quadratic model may provide a better fit. A new predictor, BuildingHeight_sq (the square of BuildingHeight), is introduced while retaining the original BuildingHeight variable to respect the hierarchy principle and ensure interpretability (Penn State, n.d.).

Although BuildingHeight and BuildingHeight_sq are highly correlated (0.945), the VIF of 9.359 remains below the standard threshold of 10, indicating that collinearity is manageable.

Therefore, the predictors of this model include BuildingAge, BuildingHeight, Insulation, AverageTemperature, SunlightExposure, WindSpeed, OccupancyRate, and the additional predictor BuildingHeight_sq. The cross-validation MSE and MAE are 5.9642 and 1.9085 respectively, which means the addition of this quadratic term substantially improves the model's predictive accuracy.

1.3. Model 3: MLR with BuildingHeight_sq and Interaction terms

a) Trial and Error process

Since Model 2 yielded the best performance so far, additional interaction terms were explored to enhance predictive accuracy. The interaction terms were selected using a trial-and-

error process, focusing on combinations of two predictors at a time. The interaction between BuildingHeight and SunlightExposure (Height_Sunlight) resulted in the largest reduction in both MSE and MAE. A second interaction term, Wind_Height (between BuildingHeight and WindSpeed), was also introduced after further trials.

After attempting to include a third interaction term, no additional improvement was observed, leading to the decision to retain only two interaction terms. It's important to note that the model retains BuildingHeight, SunlightExposure, and WindSpeed, adhering to the Hierarchy Principle (Penn State, n.d.).

b) Building Model

The model we have so far - Model 3 - includes BuildingAge, BuildingHeight, BuildingHeight_sq, Insulation, AverageTemperature, SunlightExposure, WindSpeed, OccupancyRate, Height_Sunlight, and Wind_Height. It yields an MSE of 4.3656 and an MAE of 1.6704 in cross-validation.

To address potential multicollinearity, we evaluated the interaction terms by calculating the Variance Inflation Factor (VIF) for each against its original variables. The highest recorded VIF was 4.6772, which is comfortably below the acceptable threshold of 10, confirming that multicollinearity is not a concern in this model.

2. Forward and Backward Selection - Model 4

Although Model 3 demonstrated strong performance, its complexity raised concerns about the inclusion of potentially redundant predictors. To simplify the model while maintaining its predictive power, we applied Forward Selection to identify and retain only the most important variables. The result from Forward Selection is as follow:

Predictors Addition	Adj R-squared Before	Adj R-squared After	Change in Adj R-squared
BuildingHeight_sq	0	0.933211	0.933211
BuildingAge	0.933211	0.977575	0.044364
Average Temperature	0.977575	0.985778	0.008203
BuildingHeight	0.985778	0.99061	0.004832
SunlightExposure	0.99061	0.994994	0.004384
WindSpeed	0.994994	0.998106	0.003112
Insulation	0.998106	0.998908	0.000802
Height_Sunlight	0.998908	0.999194	0.000286
OccupancyRate	0.999194	0.999214	2E-05
Wind_Height	0.999214	0.999217	3E-06

Table 2. Result from Forward Selection of Model 3

The Forward Selection process resulted in a model that retained all variables from Model 3, yielding an adjusted R^2 of approximately 99.92%. However, to create a more parsimonious model and mitigate the risk of overfitting, we manually assessed the impact of each variable.

The addition of OccupancyRate led to only a minimal increase in adjusted R^2 (2E-05), far less significant than previous predictors. As a result, we terminated the addition process at the regressor Height_Sunlight.

The revised model, Model 4, now includes the following variables:

BuildingHeight_sq, BuildingAge, AverageTemperature, BuildingHeight, SunlightExposure, WindSpeed, Insulation, and Height_Sunlight. Cross-validation was conducted to evaluate the new model's performance, resulting in an MSE of 4.4915 and an MAE of 1.6941. While this model contains fewer variables, its predictive metrics are slightly worse than those of Model 3.

We also applied Backward Selection for comparison, but it produced the same result as Forward Selection, retaining all 10 variables from Model 3. This confirms that reducing variables in Model 4 did not significantly enhance performance and highlights the trade-off between simplicity and predictive accuracy.

3. K-Nearest Neighbors (kNN) - Model 5

We selected K-Nearest Neighbors (kNN) as a nonparametric approach to build a predictive model due to its robustness and flexibility in handling various relationships between predictors and the response variable. Unlike parametric models, kNN does not assume a specific functional form, making it a versatile choice for this task (James et al., 2013).

3.1. Trial and Error Process

Since different sets of predictors lead to different optimal numbers of neighbors (k), we employed a trial and error process to determine the best set of predictors. An important consideration with kNN is its computational intensity. Given the 10,000 samples in our training dataset, calculating the distances between each data point is resource-heavy. To avoid excessive computation time during model tuning, cells related to the trial and error process in the Jupyter Notebook were locked to optimize efficiency.

Initially, we tested single predictors to see which yielded the lowest Root Mean Squared Error (RMSE). Through cross-validation, we found that Building Height produced the lowest RMSE of 18.8, compared to other individual predictors (see Table 5). Therefore, Building Height was selected as the first predictor, and we proceeded to add a second one.

Among combinations of two predictors, Building Height and Building Age delivered the best performance, maintaining a lower RMSE compared to other two-predictor combinations. It is important to note that any two-predictor combinations excluding Building Height resulted in much higher RMSE values (see Table 5).

This process was repeated for additional predictors. We continued until a list of four predictors - Building Height, Building Age, Average Temperature, and Sunlight Exposure - was identified as optimal, achieving an RMSE of 9.83, as adding a fifth predictor resulted in an increase in RMSE.

The table below summarizes the key outcomes from the trial and error process, highlighting the optimal number of predictors and their associated RMSE values.

List of predictors	Chosen K	RMSE
kNN with 1 predictor		
[*Building Age*]	90	73.26
[*Building Height*]	38	18.81
[*BuildingHeight_sq*]	39	19.23
[*Insulation*]	33	74.81
[*Average Temperature*]	90	74.67
[*Sunlight Exposure*]	99	64.53
[*Wind Speed*]	100	74.9
[*Occupancy Rate*]	100	75.03
kNN with 2 predictors		
[Building Height, *Building Age*]	7	11.26
[Building Height, *Insulation*]	26	18.95
[Building Height, *Average Temperatur*e]	20	18.05
[Building Height, *Sunlight Exposur*e]	25	18.28
[Building Height, *Wind Speed*]	25	18.6
[Building Height, *Occupancy Rate*]	19	19.23
[*Insulation, Occupancy Rate*]*	92	75.01
kNN with 3 predictors		
[Building Height, Building Age, *Insulation*]	7	11.45
[Building Height, Building Age, *Average Temperature*]	8	9.9
[Building Height, Building Age, *Sunlight Exposure*]	5	10.37
[Building Height, Building Age, *Wind Speed*]	7	11.12
[Building Height, Building Age, *Occupancy Rate*]	11	12.3
kNN with 4 predictors		
[Building Height, Building Age, Average Temperature, *Insulation*]	7	10.48

[Building Height, Building Age, Average Temperature, *Sunlight Exposure*]	**5**	**9.83**
[Building Height, Building Age, Average Temperature, *Wind Speed*]	7	10.44
[Building Height, Building Age, Average Temperature, *Occupancy Rate*]	7	11.75
kNN with 5 predictors		
[Building Height, Building Age, Average Temperature, Sunlight Exposure, *Insulation*]	6	10.96
[Building Height, Building Age, Average Temperature, Sunlight Exposure, *Wind Speed*]	4	10.81
[Building Height, Building Age, Average Temperature, Sunlight Exposure, *Occupancy Rate*]	6	12.27

Table 3. Summary of Trial and Error Process in kNN

3.2. Model 5: Final kNN with 4 predictors

Through the trial and error process, the kNN model using a set of four predictors - Building Height, Building Age, Average Temperature, and Sunlight Exposure - produced the lowest RMSE of 9.83. However, despite this being the most effective kNN model, it yields significantly large CV_MSE of 96.582 and CV_MAE of 6.4015, making it less optimal overall for predicting Heating Load.

4. Lasso Regression

Given that linear models have provided the best performance so far, Lasso Regression was implemented to refine these models by introducing regularization. Lasso has the ability to shrink coefficients and potentially force some to zero, thereby reducing model complexity and improving interpretability (James et al., 2013).

We applied Lasso to the predictor sets from Model 2, Model 3, and Model 4. First, the predictor sets were standardized to ensure comparability across variables. Next, we conducted cross-validation to identify the optimal λ for each model, with λ values ranging from $10^{-2} \cdot 0.5$ to $10^{10} \cdot 0.5$. Once the optimal λ was found, we applied it to the respective models and performed cross-validation to compute the MSE and MAE. The results from 3 models are as table below:

Lasso Regression Model	Optimal λ	MSE	MAE
Model 6 (build on Model 2)	0.005	5.9646	1.9081
Model 7 (build on Model 3)	0.005	4.3662	1.6702
Model 8 (build on Model 4)	0.005	4.4920	1.6939

Table 4. Summary of Three Models in Lasso Regression

Interestingly, despite Lasso's potential for feature selection, none of the coefficients were reduced to zero, as observed in the coefficient tables (Appendices G, H and I). Model 7, which

utilizes the same set of predictors as Model 3 - the best-performing model thus far - demonstrated better performance compared to the other two models. However, the MSE and MAE values remained nearly identical to those from the corresponding MLR models. This indicates that Lasso Regression did not perform any effective variable selection in this case, nor did it improve the performance of the models.

5. Ridge Regression

Similar to Lasso Regression, Ridge Regression is conducted in the same process: standardize data, find optimal λ via cross validation for each model, calculate CV_MSE and CV_MAE for each model. The result are in the table below:

Ridge Regression Model	Optimal λ	MSE	MAE
Model 9 (build on Model 2)	0.005	5.9642	1.9085
Model 10 (build on Model 3)	0.1424	4.3656	1.6703
Model 11 (build on Model 4)	0.1424	4.4915	1.6941

Table 5. Summary of Three Models in Ridge Regression

Although Ridge Regression does not eliminate predictors from the model, this method effectively addresses multicollinearity by regularizing the coefficients (James et al., 2013). Model 10 performs better than the other two.

6. Model Comparison and Finalization

6.1. Model Comparison

Model	CV MSE	CV MAE
Model 1 (Basic MLR)	287.6441	12.1057
Model 2 (MLR with BuildingHeight_sq)	5.9642	1.9085
Model 3 (Model 2 with Interaction terms)	**4.3656**	**1.6704**
Model 4 (Forward (& Manual) Selection)	4.4915	1.6941
Model 5 (kNN with 4 predictors)	96.582	6.4015
Model 6 (Lasso based on Model 2)	5.9646	1.9081
Model 7 (Lasso based on Model 3)	4.3662	1.6702
Model 8 (Lasso based on Model 4)	4.4920	1.6939
Model 9 (Ridge based on Model 2)	5.9642	1.9085
Model 10 (Ridge based on Model 3)	4.3656	1.6703
Model 11 (Ridge based on Model 4)	4.4915	1.6941

Table 6. Comparison of Models

According to Table 6, the top-performing models are Model 3 and Model 10, with no significant differences in cross-validated MSE and MAE. However, Ridge Regression shows

limited effectiveness in this dataset, evidenced by a small optimal lambda that suggests minimal regularization. Furthermore, while Ridge aims to reduce multicollinearity by shrinking the coefficients of highly correlated predictors close to zero, the coefficients for BuildingHeight and BuildingHeight_sq remain substantial at 15.65 and 22.23 (see Appendices J, K and L). Thus, Ridge Regression does not significantly enhance the predictive model. Consequently, Model 3, with its robust performance and more interpretable coefficients, will be chosen as the final model.

6.2. Optimal Model Description

For the optimal model, the coefficients of the predictors on the training dataset are as follows. The model captures a non-linear effect through the squared term of Building Height (BuildingHeight_sq):

$$\widehat{HeatingLoad} = 201.4 + 1.26BuildingAge + 1.43BuildingHeight + 0.043BuildingHeight\_sq - 4.33Insulation - 1.56AverageTemperature - 0.017SunlightExposure + 1.6WindSpeed + 1.5OccupancyRate - 0.0005Height\_Sunlight + 0.003Wind\_Height.$$

The final model demonstrates an MSE of 4.3592 on the training set, closely aligning with the cross-validated MSE, and an R-squared of 99.9%. While such a high R-squared may raise concerns about potential overfitting—since it is derived solely from the training data—this model is still selected due to its lowest cross-validated MSE and MAE. Cross-validation method partially addresses the overfitting concern, supporting the model's robustness and generalizability.

IV. Performance on Test Data

In the final step, we let our optimal model perform on the test dataset.

References

CFI . (2022, December 5). *Variance Inflation Factor (VIF)*. Corporate Finance Institute. https://corporatefinanceinstitute.com/resources/data-science/variance-inflation-factor-vif/

Chanaka Prasanna. (2024, July 30). *Cross Validation Explained — Leave One Out, K Fold, Stratified, and Time Series Cross Validation Techniques*. Medium; Medium. https://medium.com/@chanakapinfo/cross-validation-explained-leave-one-out-k-fold-stratified-and-time-series-cross-validation-0b59a16f2223

Department of Climate Change, Energy, The Environment and Water. (2023, November 30). *Energy consumption | energy.gov.au*. Energy.gov.au. https://www.energy.gov.au/energy-data/australian-energy-statistics/energy-consumption

Dutta, S. (2024, July 16). *What is Shuffling the Data? A Guide for Students - Sanjay Dutta - Medium*. Medium; Medium. https://medium.com/@sanjay_dutta/what-is-shuffling-the-data-a-guide-for-students-0f874572baf6

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning : with applications in R*. Springer.

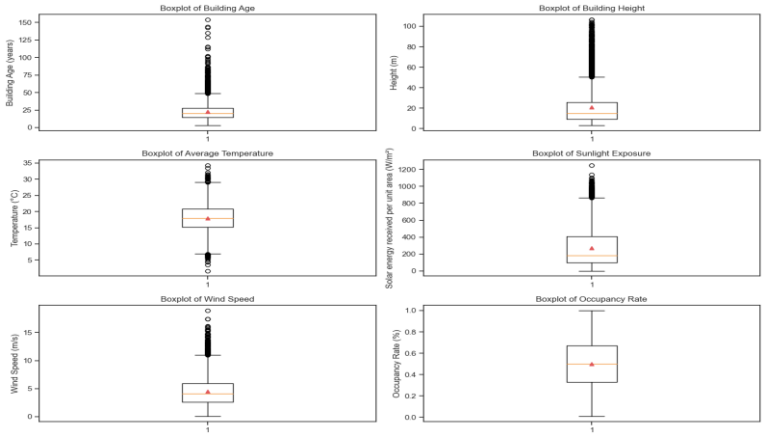
Penn State. (n.d.). *9.6 - Interactions Between Quantitative Predictors | STAT 501*. Online.stat.psu.edu. https://online.stat.psu.edu/stat501/lesson/9/9.6

Appendix

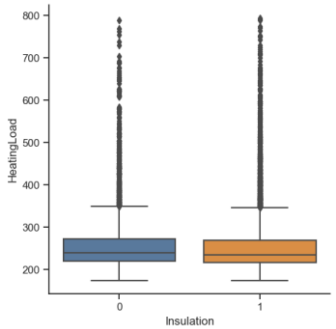
Appendix A - Null data counts over the variables

HeatingLoad	0
BuildingAge	0
BuildingHeight	0
Insulation	0
AverageTemperature	0
SunlightExposure	0
WindSpeed	0
OccupancyRate	0

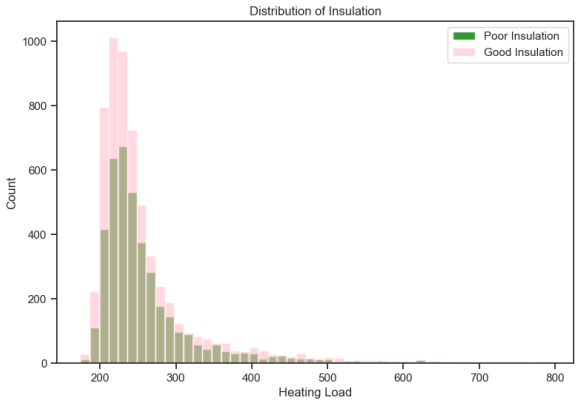
Appendix B - Boxplots of Numerical Variable In Training Set



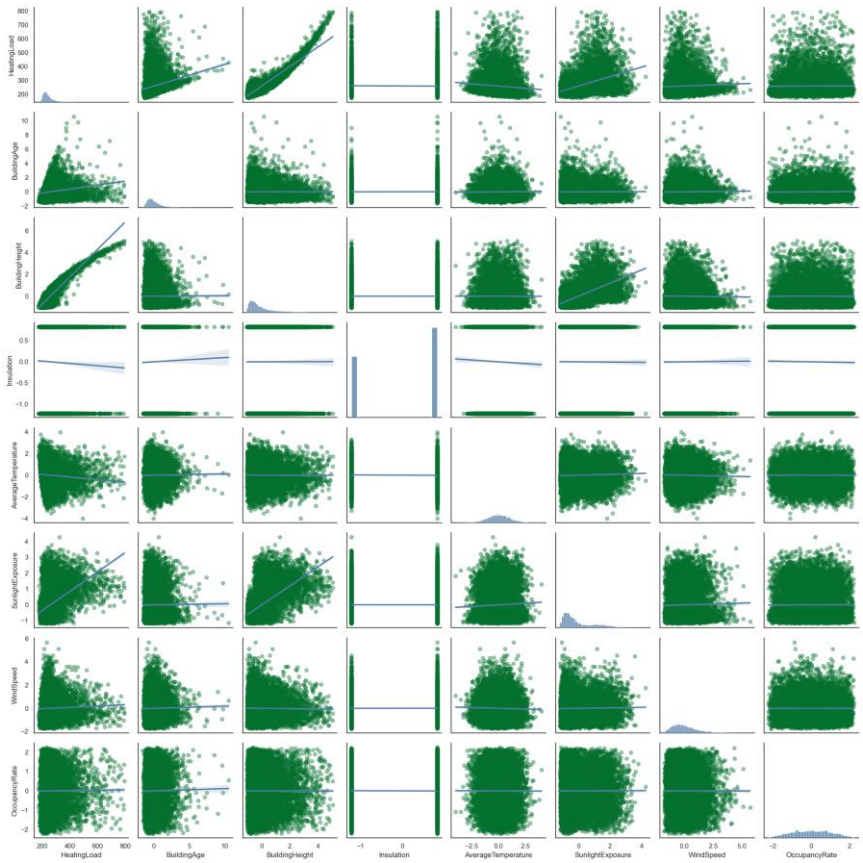
Appendix C - Boxplot of Insulation



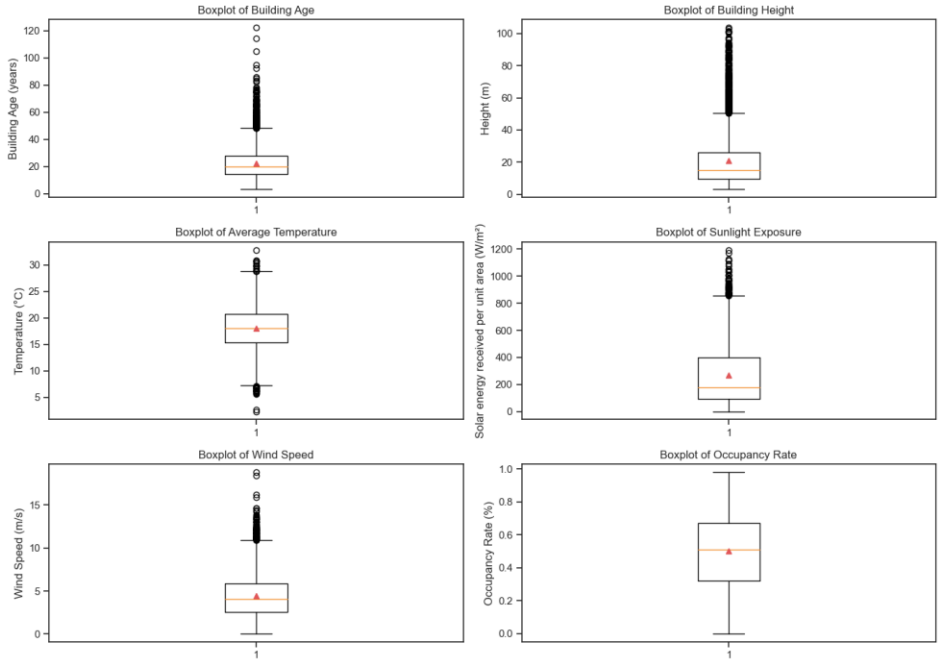
Appendix D - Distribution of HeatingLoad given the Insulation Quality



Appendix E - Pairwise plot



Appendix F - Boxplots of Numerical Variable In Test Set



Appendix G - Coefficients of Model 6 (Lasso)

BuildingHeight	BuildingAge	BuildingHeight_sq	AverageTemperature	SunlightExposure	OccupancyRate	Insulation	WindSpeed
23.674772	15.617178	53.089909	-6.358307	-6.58895	0.349751	-2.104268	4.166124

Appendix H - Coefficients of Model 7 (Lasso)

BuildingAge	BuildingHeight	BuildingHeight_sq	Insulation	AverageTemperature	SunlightExposure	WindSpeed	OccupancyRate	Height_Sunlight	Wind_Height
15.642043	22.287784	55.652711	-2.121114	-6.363808	-6.043025	4.211565	0.327564	-2.025056	0.120171

Appendix I - Coefficients of Model 8 (Lasso)

BuildingAge	BuildingHeight	BuildingHeight_sq	AverageTemperature	SunlightExposure	WindSpeed	Insulation	Height_Sunlight
15.644245	22.298675	55.622525	-6.365776	-6.03438	4.205203	-2.125409	-2.013447

Appendix J - Coefficients of Model 9 (Ridge)

BuildingHeight	BuildingAge	BuildingHeight_sq	AverageTemperature	SunlightExposure	OccupancyRate	Insulation	WindSpeed
23.641552	15.622397	53.127953	-6.363272	-6.59293	0.354588	-2.109589	4.171215

Appendix K - Coefficients of Model 10 (Ridge)

BuildingAge	BuildingHeight	BuildingHeight_sq	Insulation	AverageTemperature	SunlightExposure	WindSpeed	OccupancyRate	Height_Sunlight	Wind_Height
15.64483	22.243522	55.726889	-2.126504	-6.368758	-6.03895	4.21746	0.332056	-2.05092	0.12809

Appendix L - Coefficient of Model 11 (Ridge)

BuildingAge	BuildingHeight	BuildingHeight_sq	AverageTemperature	SunlightExposure	WindSpeed	Insulation	Height_Sunlight
15.64964	22.242537	55.696354	-6.37.0834	-6.032767	4.21069	-2.13096	-2.038384