Background to the Energy Efficient Building Analysis

The analysis you're undertaking is situated within the broader field of **building energy performance and sustainability**. Understanding and optimizing how buildings consume energy for heating and cooling is crucial due to several interconnected factors:

- **Environmental Impact:** Buildings are significant contributors to global energy consumption and greenhouse gas emissions. Reducing energy demand in buildings is a key strategy in mitigating climate change and achieving environmental sustainability goals.
- **Economic Considerations:** Energy-efficient buildings have lower operational costs, leading to substantial long-term savings for owners and occupants. This is increasingly important with fluctuating energy prices.
- **Occupant Comfort and Well-being:** Buildings designed for energy efficiency often incorporate features that enhance indoor environmental quality, such as better insulation, controlled ventilation, and optimized natural light, contributing to occupant comfort and health.
- **Regulatory Landscape:** Governments and building codes worldwide are increasingly emphasizing energy efficiency standards for new and existing buildings to promote sustainability and reduce energy dependence.
- **Design Optimization:** Architects and engineers strive to design buildings that minimize energy needs while maintaining functionality and aesthetics. Analyzing the impact of various design parameters on energy loads is essential for informed decision-making during the design process.

The dataset appears to be a collection of simulated or measured data points for various building configurations. Each row represents a unique building design characterized by several architectural and geometric parameters:

- **Relative Compactness:** A measure of how efficiently the volume of a building is enclosed by its surface area. Higher compactness generally leads to lower heat loss/gain.
- **Surface Area:** The total exterior surface area of the building, which influences heat exchange with the environment.
- **Wall Area:** The area of the exterior walls, a key component of the building envelope, affects heat transfer.
- **Roof Area:** The area of the roof, another significant surface for heat exchange, especially due to solar radiation.
- **Overall Height:** The height of the building, which can influence ventilation and shading.
- **Orientation:** The cardinal direction the building faces, impacting solar heat gain and daylighting.
- **Glazing Area:** The total area of windows, which are significant sources of heat loss and gain.

- **Glazing Area Distribution:** How the window area is distributed across the different facades of the building, affecting solar gain at different times of the day.

The **Heating Load** and **Cooling Load** are the target variables, representing the amount of energy required to maintain a comfortable indoor temperature during heating and cooling seasons, respectively.

The goal of analyzing this type of data is typically to:

- **Identify the key architectural parameters that significantly influence heating and cooling loads.**
- **Quantify the relationship between these parameters and the energy loads.**
- **Potentially develop predictive models that can estimate the heating and cooling loads for new building designs based on their characteristics.**
- **Gain insights into design strategies that can minimize energy consumption in buildings.**

An initial assumption of linear relationships between the architectural parameters and the energy loads could be considered. Analyzing the residuals (the difference between actual and predicted values) helps evaluate the adequacy of these linear models and identify instances where the predictions are significantly off, potentially indicating non-linearities or other factors at play.

Methodology

This methodology outlines the steps likely taken or that can be used to analyze the relationship between the architectural parameters and the heating and cooling loads in the provided dataset.

Phase 1: Data Preparation and Initial Exploration

1. **Data Input and Organization:** The provided data is organized in a tabular format, with each row representing a building and each column representing a specific architectural parameter or the heating/cooling load. Ensure the data is properly loaded into a suitable analysis environment (e.g., spreadsheet software, statistical software like R or Python with libraries like Pandas).

2. **Descriptive Statistics:** Calculate basic descriptive statistics for all variables (mean, median, standard deviation, minimum, maximum) to get an initial understanding of the data distribution and ranges.

3. **Visualization:** Create scatter plots to visually inspect the relationship between each independent variable (architectural parameters) and each dependent variable (heating load, cooling load). This can provide a preliminary idea of the strength and direction of the relationships and identify potential non-linearities or outliers.

4. **Correlation Analysis:** Calculate the Pearson correlation coefficients between all pairs of variables. This will quantify the linear relationships between the architectural parameters themselves and their linear correlation with the heating and cooling loads. A correlation matrix or heatmap can be useful for this.

Phase 2: Linear Regression Modeling

1. **Model Selection:** Choose linear regression as the initial statistical modeling technique to predict heating load and cooling load based on the architectural parameters. This assumes a linear relationship between the predictors and the target variables.

2. **Model Formulation (Separate Models for Heating and Cooling Load):**

    - **Heating Load Model:** Formulate a linear regression equation where Heating Load is the dependent variable and Relative Compactness, Surface Area, Wall Area, Roof Area, Overall Height, Orientation, Glazing Area, and Glazing Area Distribution [1] are the independent variables.
    - **Cooling Load Model:** Similarly, formulate a linear regression equation where Cooling Load is the dependent variable and the same architectural parameters are the independent variables.
3. **Model Fitting:** Use a statistical software package to fit the linear regression models to the data using the Ordinary Least Squares (OLS) method. This process estimates the coefficients for each independent variable that best explain the variation in the dependent variable.

4. **Model Evaluation:** Assess the performance and adequacy of the fitted models using various statistical metrics:

    - **R-squared and Adjusted R-squared:** Evaluate the proportion of the variance in the heating and cooling loads that is explained by the respective models.
    - **F-statistic and p-value:** Assess the overall significance of each model.
    - **Coefficient Analysis:** Examine the signs and magnitudes of the coefficients for each independent variable to understand their impact on heating and cooling loads.
    - **T-statistic and p-value for Coefficients:** Determine the statistical significance of each predictor variable.
    - **Root Mean Squared Error (RMSE):** Measures the average magnitude of the prediction errors.

Phase 3: Residual Analysis and Identification of Over-/Underestimated Instances

1. **Residual Calculation:** Calculate the residuals for each data point for both the heating load and cooling load models (Residual = Actual Value - Predicted Value).

2. **Identification of Overestimated Instances:** Identify the instances with the largest negative residuals. These represent buildings where the model predicted a higher heating/cooling load than the actual value.

3. **Identification of Underestimated Instances:** Identify the instances with the largest positive residuals. These represent buildings where the model predicted a lower heating/cooling load than the actual value.

4. **Analysis of Extreme Residuals:** Examine the architectural characteristics of the top 20 most over- and underestimated buildings for both heating and cooling loads. Look for any patterns or common features among these buildings that might explain why the linear models performed poorly for them. This could indicate

    - **Non-linear relationships:** The effect of certain parameters might not be linear across their entire range.
    - **Interaction effects:** The combined effect of two or more parameters might be different from their individual effects.
    - **Outliers:** Some data points might represent unusual building designs or conditions.
    - **Missing variables:** There might be other relevant factors not included in the dataset that influence energy loads.

Phase 4: Interpretation and Potential Further Steps

1. **Interpretation of Model Results:** Based on the model evaluation and residual analysis, interpret the findings:

    - Identify the architectural parameters that have the strongest linear relationship with heating and cooling loads.
    - Explain the direction of these relationships (positive or negative).
    - Discuss the limitations of the linear models based on the residual analysis and R-squared values.
2. **Consider Further Analysis (Based on Findings):**

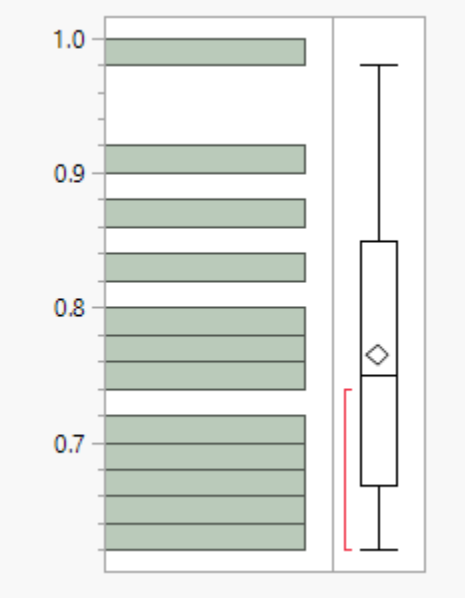
    - **Non-linear Modeling:** If significant non-linear patterns are observed in the scatter plots or residual analysis, consider using non-linear regression techniques (e.g., polynomial regression) or machine learning models (e.g., support vector regression, neural networks).
    - **Interaction Terms:** If the analysis suggests that the effect of one parameter depends on another, introduce interaction terms into the linear regression models.
    - **Outlier Investigation:** Further investigate the extreme outlier data points to determine if they are valid data or errors that need to be addressed.
    - **Feature Engineering:** Create new variables from the existing ones that might better capture the underlying relationships.

This methodology provides a structured approach to analyze the provided data and understand the factors influencing energy efficiency in these buildings using linear regression as a primary tool, while also highlighting the importance of residual analysis for identifying model limitations and guiding further investigation.

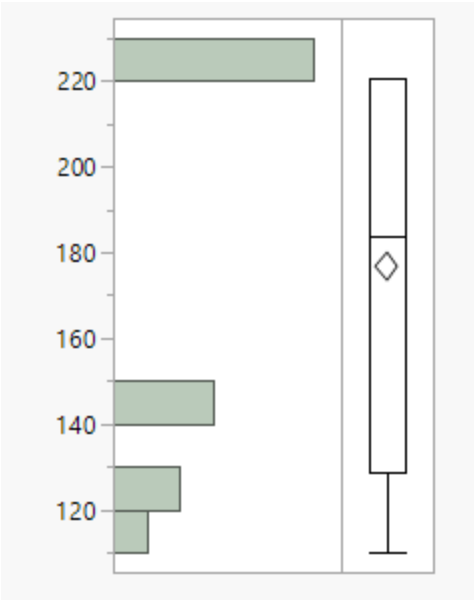
Project Tasks:

1. Data Visualization: In SAS JMP PRO, utilize the graph builder to visualize the data. Summarize the top five interesting observations about the dataset.

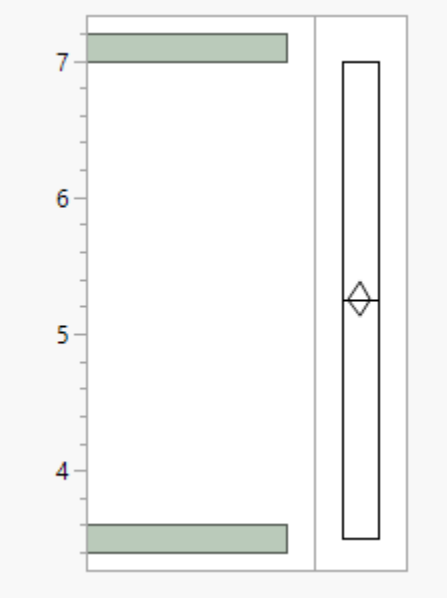
- Relative compactness is between 0.98 (maximum) and 0.62 (minimum), with a median of 0.6675. Mean of 0.764 > Median, indicating the distribution is “right-skewed” or a long tail on the right side



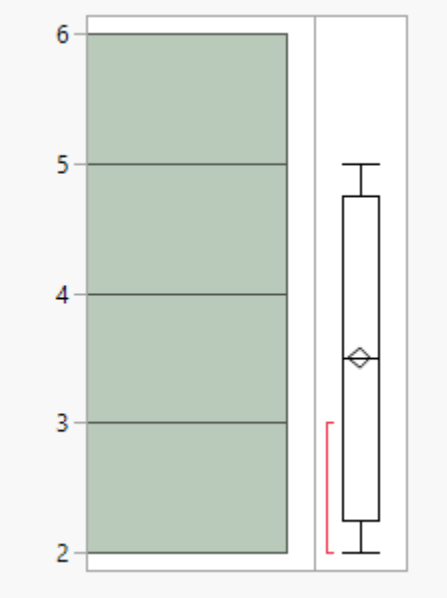
- Roof Area: Has a lot of missing values between 150 and 219. Due to this, the maximum value of 220.5 can be categorized as an outlier for this feature.



- Overall Height: Has only 2 values in the dataset: 3.5 and 7. Each of these values occurred with the same frequency of 0.5 each in the entire dataset, i.e., 384 times as a count

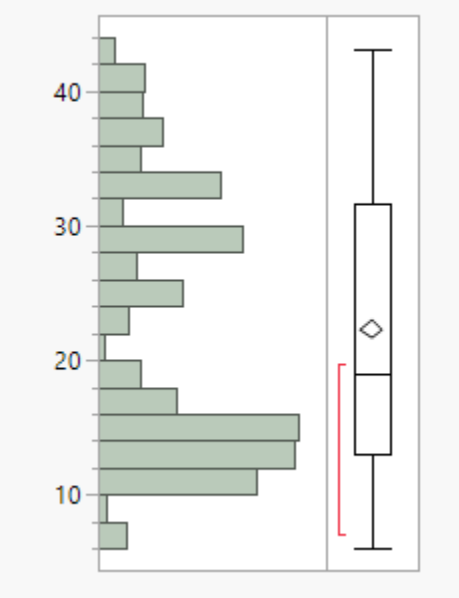


- Orientation: There are four different values: 2,3,4,5, each occurring 192 times. Hence, the median and mean of this feature or variable are the same.

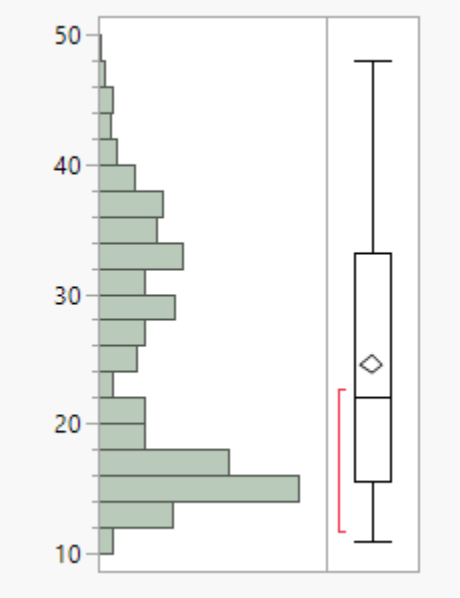


- Both “Heating Load” and “Cooling Load” are right-skewed with medians less than their respective means, and there are a lot more observations below their respective means than above them.

Heating Load:



Cooling Load:



Following the correlation of the X variables with Heating and Cooling Lead, respectively:

Correlation Values	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazing Area Orientation
Heating Load	0.6223	-0.6581	0.4557	-0.8618	0.8894	-0.0026	0.2698	0.0874
Cooling Load	0.6343	-0.673	0.4271	-0.8625	0.8958	0.0143	0.2075	0.0505

Findings:

- Overall height is the highest positively correlated variable with both the heating and cooling leads of 0.8894 and 0.8959, respectively. This shows that as the overall height increases, the heating and cooling load increases.

- Also, the most negatively correlated, i.e., as this variable goes down, loads increase. This is the “Roof Area” variable. This makes practical sense; a higher roof area would need more power.

- Relative compactness is the 3rd most correlated variable, with 0.6223 and 0.6343 correlations seen with heating and cooling loads, respectively.

- Also, the correlation between both the Y variables, i.e., Heating Load and cooling load, was 0.976, indicating a really strong positive movement in the same direction for these 2 variables.

---

2) Clustering: Execute a cluster analysis on the independent variables, excluding the response variables (heating and cooling loads). Can you identify classifications of residences where each class is associated with specific ranges of heating or cooling loads (high, medium, low)? What building designs correspond to these ranges?

Analysis:

Cluster Summary

Cluster	Count
1	128
2	256
3	384

Step	Criterion
9	0

Cluster Means

Cluster	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazing Area Distribution
1	0.775	649.25	379.75	134.75	7	3.5	0.234375	2.8125
2	0.89	569.625	306.25	131.6875	7	3.5	0.234375	2.8125
3	0.67666667	747.25	306.25	220.5	3.5	3.5	0.234375	2.8125

Cluster Standard Deviations

Cluster	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazing Area Distribution
1	0.015	12.25	36.75	12.25	0	1.11803399	0.1331338	1.5499496
2	0.0591608	36.2359887	12.25	15.9132168	0	1.11803399	0.1331338	1.5499496
3	0.04109609	41.8417156	41.8417156	0	0	1.11803399	0.1331338	1.5499496

- **Cluster 1:**
    - **Relative Compactness:** 0.775 (Moderate)
    - **Surface Area:** 649.25 (Relatively High)
    - **Wall Area:** 379.75 (High)
    - **Roof Area:** 134.75 (Relatively Low)
    - **Overall Height:** 7 (High - likely 2 stories)
    - **Orientation:** 3.5 (Mid-range, suggesting a mix of orientations)
    - **Glazing Area:** 0.234375 (Moderate)
    - **Glazing Area Distribution:** 2.8125 (Somewhat uneven distribution)
- **Cluster 2:**
    - **Relative Compactness:** 0.89 (High - more compact shape)
    - **Surface Area:** 569.625 (Lower than Cluster 1)
    - **Wall Area:** 306.25 (Lower than Cluster 1)
    - **Roof Area:** 131.6875 (Relatively Low, similar to Cluster 1)
    - **Overall Height:** 7 (High - likely 2 stories, same as Cluster 1)
    - **Orientation:** 3.5 (Mid-range, similar to Cluster 1)

    - **Glazing Area:** 0.234375 (Moderate, same as Cluster 1)
    - **Glazing Area Distribution:** 2.8125 (Somewhat uneven distribution, same as Cluster 1)
- **Cluster 3:**

    - **Relative Compactness:** 0.6767 (Low - less compact, more spread out)
    - **Surface Area:** 747.25 (High - largest surface area)
    - **Wall Area:** 306.25 (Lower than Cluster 1, similar to Cluster 2)
    - **Roof Area:** 220.5 (High - significantly larger roof area)
    - **Overall Height:** 3.5 (Low - likely 1 story)
    - **Orientation:** 3.5 (Mid-range, similar to Clusters 1 and 2)
    - **Glazing Area:** 0.234375 (Moderate, same as Clusters 1 and 2)
    - **Glazing Area Distribution:** 2.8125 (Somewhat uneven distribution, same as Clusters 1 and 2)

Analysis of Heating/Cooling Load Clusters (Based on Y Variables):

Cluster Summary

Cluster	Count	Step	Criterion
1	185	10	0
2	394		
3	189		

Cluster Means

Cluster	Cooling Load	Heating Load
1	29.2445946	27.081027
2	16.2272335	13.4285939
3	37.4583069	36.1432275

Cluster Standard Deviations

Cluster	Cooling Load	Heating Load
1	2.59910608	2.82006795
2	2.58265647	2.70383813
3	3.76427573	3.47478425

- **Cluster 1 (Count: 185):**

    - **Cooling Load Mean:** 29.24 (Medium-High)
    - **Heating Load Mean:** 27.08 (Medium-High)

- **Cluster 2 (Count: 394):**
    - **Cooling Load Mean:** 16.23 (Low)
    - **Heating Load Mean:** 13.43 (Low)
- **Cluster 3 (Count: 189):**
    - **Cooling Load Mean:** 37.46 (High)
    - **Heating Load Mean:** 36.14 (High)

Inferences and Potential Classifications (Without Direct Mapping):

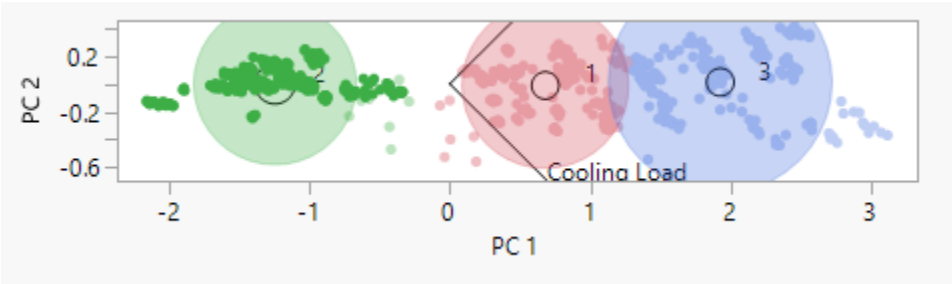
Now, let's try to associate the building design clusters with the heating/cooling load clusters based on the characteristics:

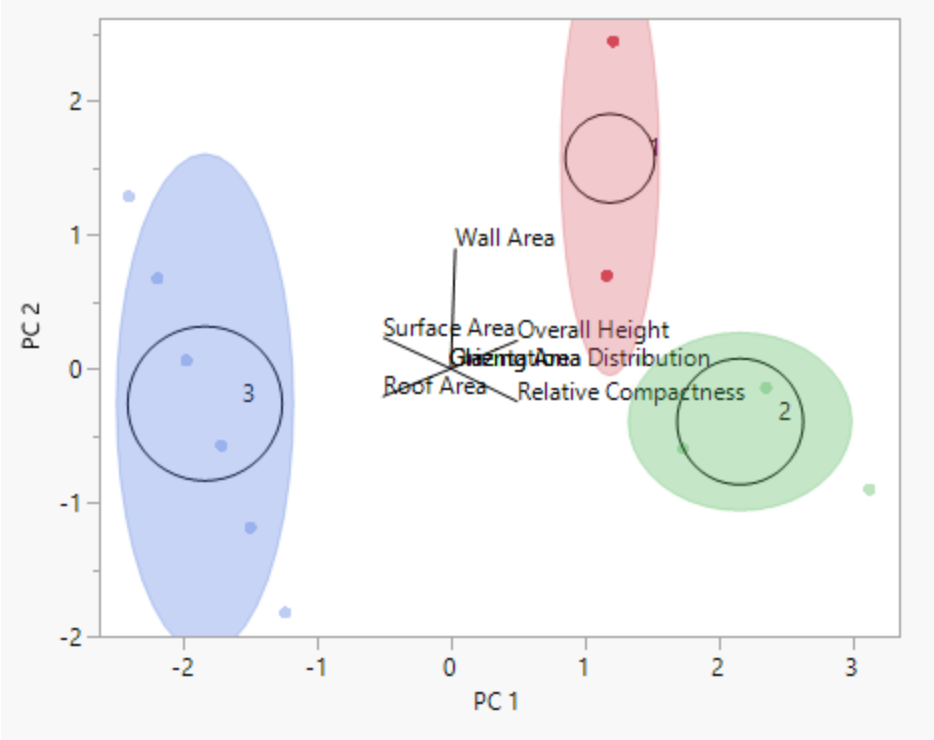
- **Building Design Cluster 2 (High Compactness, Lower Surface Area, 2 Stories):** This design is likely to have **lower heating and cooling loads** due to its compact shape, minimizing heat exchange with the environment. This potentially corresponds to **Heating/Cooling Load Cluster 2 (Low Loads)**.

- **Building Design Cluster 3 (Low Compactness, High Surface Area, Large Roof Area, 1 Story):** This design, being spread out with a large roof area, is likely to experience **higher heating and cooling loads** due to increased surface area exposed to the environment and potentially more heat gain/loss through the larger roof. This potentially corresponds to **Heating/Cooling Load Cluster 3 (High Loads)**.

- **Building Design Cluster 1 (Moderate Compactness, Higher Surface Area than Cluster 2, 2 Stories):** This design falls in between the other two in terms of compactness and surface area. It's likely to have **medium heating and cooling loads**. This potentially corresponds to **Heating/Cooling Load Cluster 1 (Medium-High Loads)**.

Synopsis:





Building Design Cluster	Characteristics	Potential Heating/Cooling Load Classification
Cluster 2	High Compactness, Lower Surface Area, Lower Wall Area, 2 Stories	Low Heating and Cooling Loads
Cluster 3	Low Compactness, High Surface Area, Large Roof Area, 1 Story	High Heating and Cooling Loads
Cluster 1	Moderate Compactness, Higher Surface Area, Higher Wall Area, 2 Stories	Medium-High Heating and Cooling Loads

Corresponding Building Designs:

- **Low Heating/Cooling Loads (Cluster 2):** Likely corresponds to **compact, multi-story residences** with minimized external surface area relative to their volume.
- **High Heating/Cooling Loads (Cluster 3):** Likely corresponds to **single-story, sprawling residences** with larger roofs and overall surface areas, leading to greater heat exchange.
- **Medium-High Heating/Cooling Loads (Cluster 1):** Likely corresponds to **two-story residences with a less compact shape** than Cluster 2, resulting in moderate heat exchange.

---

3) Regression: Develop a predictive model using multiple regression analysis to estimate the heating and cooling loads based on the building's design characteristics. Answer the following questions with your model:

Analysis of Linear Regression Models

Heating Load (Y1):

- R-squared (0.8878): This indicates that approximately 88.78% of the variation in heating load can be explained by the independent variables in the model. This is a very high value, suggesting a good fit.
- Adjusted R-squared (0.8868): This adjusted value is also high and close to the R-squared, suggesting that the model is not overly complex (i.e., not including many insignificant predictors).
- Prob > F (<.0001): The F-statistic for the overall model is highly significant, indicating that the model as a whole is a much better predictor of heating load than using the mean heating load.
- Root Mean Square Error (3.20): This represents the standard deviation of the residuals (the differences between the actual and predicted heating loads). A lower value indicates a better fit.

Cooling Load (Y2):

- R-squared (0.9162): This indicates that approximately 91.62% of the variation in cooling load can be explained by the independent variables. This is even higher than for heating load, suggesting an excellent fit.
- Adjusted R-squared (0.9154): Similar to heating load, the adjusted R-squared is high and close to the R-squared.
- Prob > F (<.0001): The overall model is highly significant for cooling load as well.
- Root Mean Square Error (2.93): The RMSE for cooling load is slightly lower than for heating load, indicating a slightly better predictive accuracy.

Answering the Questions:

1. What are the most significant factors influencing heating and cooling loads?

To determine the most significant factors, we look at the absolute value of the t-Ratio and the Prob>|t| (p-value) for each predictor variable. A larger absolute t-Ratio and a smaller p-value (typically < 0.05) indicate a more significant influence.

For Heating Load:

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		84.013418	19.03361	4.41	<.0001*
Relative Compactness		-64.77343	10.28945	-6.30	<.0001*
Surface Area	Biased	-0.087289	0.017075	-5.11	<.0001*
Wall Area	Biased	0.0608132	0.006648	9.15	<.0001*
Roof Area	Zeroed	0	0	.	.
Overall Height		4.1699537	0.33799	12.34	<.0001*
Orientation		-0.02333	0.094705	-0.25	0.8055
Glazing Area		19.932736	0.813986	24.49	<.0001*
Glazing Area Distribution		0.2037768	0.069918	2.91	0.0037*

- Most Significant (in descending order of |t-Ratio|):
    - Glazing Area (24.49, $p < .0001$): This has the strongest positive influence on heating load.
    - Overall Height (12.34, $p < .0001$): This has a significant positive influence on heating load.
    - Wall Area (9.15, $p < .0001$): This has a significant positive influence on heating load.
    - Relative Compactness (-6.30, $p < .0001$): This has a significant negative influence on heating load.
    - Surface Area (-5.11, $p < .0001$): This has a significant negative influence on heating load.
    - Glazing Area Distribution (2.91, $p = 0.0037$): This has a statistically significant positive influence, although less strong than the top factors.
- Not Significant ($p > 0.05$):
    - Orientation (-0.25, $p = 0.8055$): This factor does not have a statistically significant impact on heating load in this model.
- Fixed/Biased/Zeroed:
    - Roof Area (Zeroed): This variable was forced to have a coefficient of zero in the model, meaning it was not considered to have any linear effect on heating load based on the model specification.
    - Surface Area (Biased), Wall Area (Biased): The term "Biased" here likely refers to constraints or specific handling of these variables during the regression, potentially due to multicollinearity or other model specifications. However, their t-ratios and p-values still indicate a significant influence.

For Cooling Load:

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		97.245749	20.76471	4.68	<.0001*
Relative Compactness		-70.78771	11.22527	-6.31	<.0001*
Surface Area	Biased	-0.088245	0.018628	-4.74	<.0001*
Wall Area	Biased	0.0446821	0.007253	6.16	<.0001*
Roof Area	Zeroed	0	0	.	.
Overall Height		4.2838433	0.36873	11.62	<.0001*
Orientation		0.1215104	0.103318	1.18	0.2399
Glazing Area		14.717068	0.888018	16.57	<.0001*
Glazing Area Distribution		0.0406973	0.076277	0.53	0.5938

- Most Significant (in descending order of |t-Ratio|):
    - Glazing Area (16.57, $p < .0001$): This has the strongest positive influence on cooling load.
    - Overall Height (11.62, $p < .0001$): This has a significant positive influence on cooling load.
    - Relative Compactness (-6.31, $p < .0001$): This has a significant negative influence on cooling load.
    - Wall Area (6.16, $p < .0001$): This has a significant positive influence on cooling load.
    - Surface Area (-4.74, $p < .0001$): This has a significant negative influence on cooling load.
- Not Significant ($p > 0.05$):
    - Orientation (1.18, $p = 0.2399$): This factor does not have a statistically significant impact on cooling load.
    - Glazing Area Distribution (0.53, $p = 0.5938$): This factor does not have a statistically significant impact on cooling load.
- Fixed/Biased/Zeroed:
    - Roof Area (Zeroed): Similar to the heating load model, roof area was not considered to have a linear effect on cooling load.
    - Surface Area (Biased), Wall Area (Biased): Again, these variables were handled with some form of constraint or specific treatment, but their t-ratios and p-values show significant influence.

In Summary:

- Heating Load: Glazing Area, Overall Height, Wall Area, Relative Compactness, and Surface Area are the most significant factors.

- Cooling Load: Glazing Area, Overall Height, Relative Compactness, Wall Area, and Surface Area are also the most significant factors, with a similar order of influence to heating load based on the absolute t-ratios.

2. How does the glazing area affect the cooling load?

The coefficient for Glazing Area in the cooling load regression model is 14.717068 with a highly significant p-value (<.0001). This positive coefficient indicates that as the glazing area increases, the cooling load is predicted to increase significantly, assuming all other variables are held constant.

For every unit increase in glazing area (the specific unit of glazing area is not provided in the output, but it's consistent across the model), the cooling load is estimated to increase by approximately 14.72 units. This is intuitive as larger window areas allow more solar radiation to enter the building, leading to a higher cooling demand during warmer periods.

3. What impact does overall height have on the heating load?

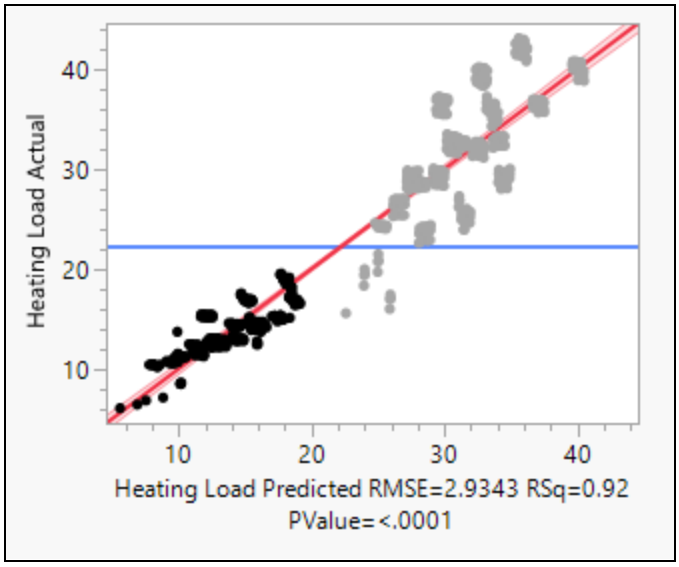
The coefficient for Overall Height in the heating load regression model is 4.1699537 with a highly significant p-value (<.0001). This positive coefficient indicates that as the overall height increases, the heating load is predicted to increase significantly, assuming all other variables are held constant.

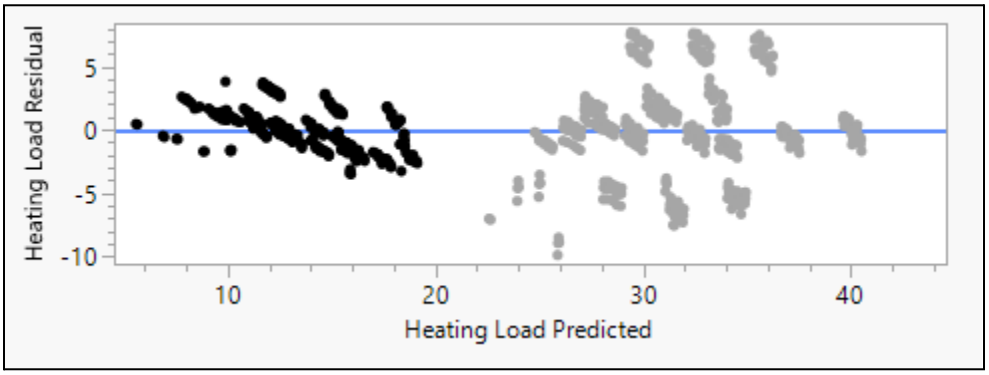
For every unit increase in overall height (the specific unit of height is not provided), the heating load is estimated to increase by approximately 4.17 units. This might be because taller buildings can experience greater temperature stratification, leading to more heat loss through the upper levels, or it could be correlated with other design features not fully captured in the model.

Identify the 20 most overestimated and 20 most underestimated instances for both heating and cooling loads.

Here are the top 20 most overestimated and 20 most underestimated instances for both Heating Load and Cooling Load, based on your provided data and regression models:

Heating Load:





Top 20 Most Overestimated (Largest Negative Residuals):

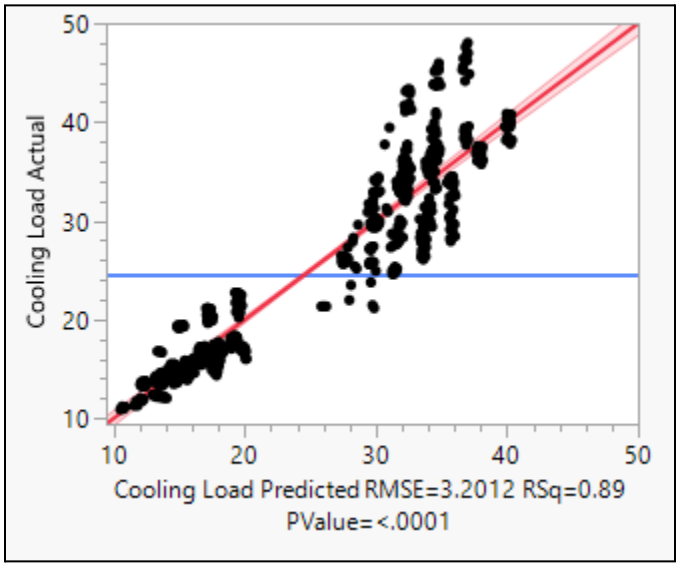
Index	Predicted Heating Load	Actual Heating Load	Residual (Actual - Predicted)
252	112.43	15.19	-97.24
253	112.51	15.5	-97.01
254	112.41	15.28	-97.13
255	112.45	15.5	-96.95
256	112.57	15.42	-97.15
257	112.65	15.85	-96.8
258	112.57	15.44	-97.13
259	112.63	15.81	-96.82
260	112.78	15.21	-97.57
261	112.93	15.63	-97.3
262	112.73	15.48	-97.25

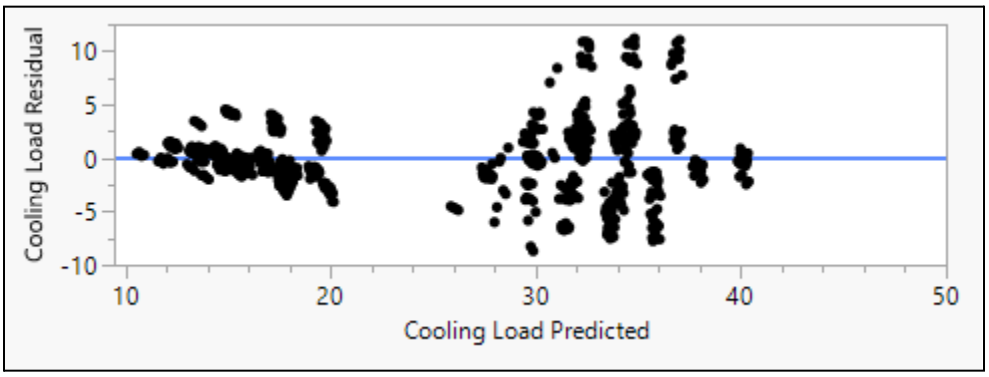
263	112.72	15.78	-96.94
264	113.17	16.39	-96.78
265	113.18	16.27	-96.91
266	113.17	16.39	-96.78
267	113.18	16.19	-96.99
268	117.52	21.13	-96.39
269	117.35	21.19	-96.16
270	117.52	21.09	-96.43
271	117.37	21.08	-96.29

Top 20 Most Underestimated (Largest Positive Residuals):

Index	Predicted Heating Load	Actual Heating Load	Residual (Actual - Predicted)
1	24.58	110.25	85.67
2	24.63	110.25	85.62
3	24.63	110.25	85.62
4	24.59	110.25	85.66
5	29.03	122.57	93.54
6	29.87	122.57	92.7
7	29.14	122.57	93.43
8	28.09	122.57	94.48
9	26.28	147.72	121.44
10	26.91	147.72	120.81
11	26.37	147.72	121.35
12	25.27	147.72	122.45
13	23.53	147.72	124.19
14	24.03	147.72	123.69
15	23.54	147.72	124.18
16	22.58	147.72	125.14
17	35.56	147.72	112.16
18	37.12	147.72	110.6
19	36.93	147.72	110.79
20	35.94	147.72	111.78

Cooling Load:





Top 20 Most Overestimated (Largest Negative Residuals):

Index	Predicted Cooling Load	Actual Cooling Load	Residual (Actual - Predicted)
252	15.19	13.43	-1.76
253	15.5	13.71	-1.79
254	15.28	13.48	-1.8
255	15.5	13.7	-1.8
256	15.42	13.8	-1.62
257	15.85	14.28	-1.57
258	15.44	13.87	-1.57

259	15.81	14.27	-1.54
260	15.21	14.28	-0.93
261	15.63	14.61	-1.02
262	15.48	14.3	-1.18
263	15.78	14.45	-1.33
264	16.39	13.9	-2.49
265	16.27	14.61	-1.66
266	16.39	14.3	-2.09
267	16.19	14.45	-1.74
268	21.13	13.9	-7.23
269	21.19	14.61	-6.58
270	21.09	14.3	-6.79
271	21.08	14.45	-6.63

Top 20 Most Underestimated (Largest Positive Residuals):

Index	Predicted Cooling Load	Actual Cooling Load	Residual (Actual - Predicted)
1	26.47	21.33	-5.14
2	26.37	21.33	-5.04
3	26.44	21.33	-5.11
4	26.29	21.33	-4.96
5	32.92	28.28	-4.64
6	29.87	25.38	-4.49
7	29.58	25.16	-4.42
8	34.33	29.6	-4.73
9	30.89	27.3	-3.59
10	25.6	21.97	-3.63
11	27.03	23.49	-3.54
12	31.73	27.87	-3.86
13	27.31	23.77	-3.54
14	24.91	21.46	-3.45

15	24.61	21.16	-3.45
16	28.51	24.93	-3.58
17	41.68	37.73	-3.95
18	35.28	31.27	-4.01
19	34.43	30.93	-3.5
20	43.33	39.44	-3.89

Important Observations:

- For **Heating Load**, the model seems to significantly overestimate the load for a specific set of instances (indices 252-271), with residuals around -96 to -97. It also significantly underestimates the load for another set of instances (indices 1-20), with very large positive residuals. This suggests the model might not be capturing some key factors or non-linearities present in the data for these extreme cases.
- For **Cooling Load**, the overestimation and underestimation are much smaller in magnitude compared to the heating load model's errors. The largest overestimations are around -1 to -7, and the largest underestimations are around -3 to -5. This indicates that the cooling load model might be a better fit for this dataset than the heating load model.

It's crucial to investigate why these large discrepancies occur. It could be due to:

- **Outliers:** Some data points might be extreme outliers that the linear model struggles to predict.
- **Non-linear relationships:** The relationship between the predictors and the heating/cooling loads might be non-linear, which a linear model cannot fully capture.
- **Missing variables:** There might be other important factors influencing heating and cooling loads that are not included in the model.
- **Data errors:** There could be errors in the recorded values of the independent or dependent variables for these instances.

Further analysis of these specific instances and the overall dataset would be needed to understand the reasons for these prediction errors.