TABLE OF CONTENTS

Problem Statement

The increase in building energy consumption (EBP) has led to concerns of energy waste and its negative impact on our environment. There are laws and policies set in place seeking to regulate energy consumption by ensuring that buildings meet particular building codes (varies across location). Buildings play a massive role in energy consumption since it takes a lot of energy to power heat and cool air. Thus, engineers and architects have been working together to find ways to create/maintain energy efficient buildings.

There are different building features that play a role in a building's energy efficiency (e.g.: height of building). Energy efficiency of a building can be measured by observing heating load and cooling load. Since it is unrealistic to measure the heating and cooling load of every single building in a given area, engineers have utilized building simulations and machine learning to analyze energy efficiency of buildings across building features. Based on a sample energy efficiency dataset, I hope to create a machine learning model that can analyze building features across one aspect of energy efficiency (heating load).

Data Wrangling

The energy efficiency dataset was provided as an open source file by UC Irvine, and contains roughly 768 rows and nine columns. Our dataset did contain any missing values. The raw energy efficiency dataset contained column names that only contained X or y labels, along with its respective position (e.g.: X3 or y2). Thus, I renamed each of these columns based on the associated X or y variable (X1 as Relative Compactness or y1 as Heating Load). All X columns appeared to be relevant to our analysis. I dropped Cooling Load from analysis, since I was only interested in observing Heating Load.

Exploratory Data Analysis

After renaming and dropping irrelevant columns, I explored the dataset to get an idea of its composition. Relative Compactness, Surface Area, Wall Area, Roof Area, Overall Height, Glazing Area, Heating Load, and Cooling Load were represented using floats, while Orientation and Heating Load were represented using integer values. Although Overall Height contained continuous values, this column only contained two unique values (3.5 and 7, which is why we treated these values as categories). I wanted to find the average Heating Load level across the data frame, which averaged out to 22.31. I conducted univariate analysis to get a look at each variable in isolation. I utilized Seaborn to create subplots of histograms across eight different building features (Relative Compactness, Surface Area, Wall Area, Roof Area, Overall Height, Orientation, Glazing Area, and Glazing Area Distribution) (**Figure 1**), along with a histogram for Heating Load (**Figure 2**.). With the exception of Wall Area and Roof Area, the distribution across the X variables were relatively uniform in distribution. I also utilized a boxplot to check for outliers across all variables (**Figure 3-4.**). No outliers were observed within the dataset.

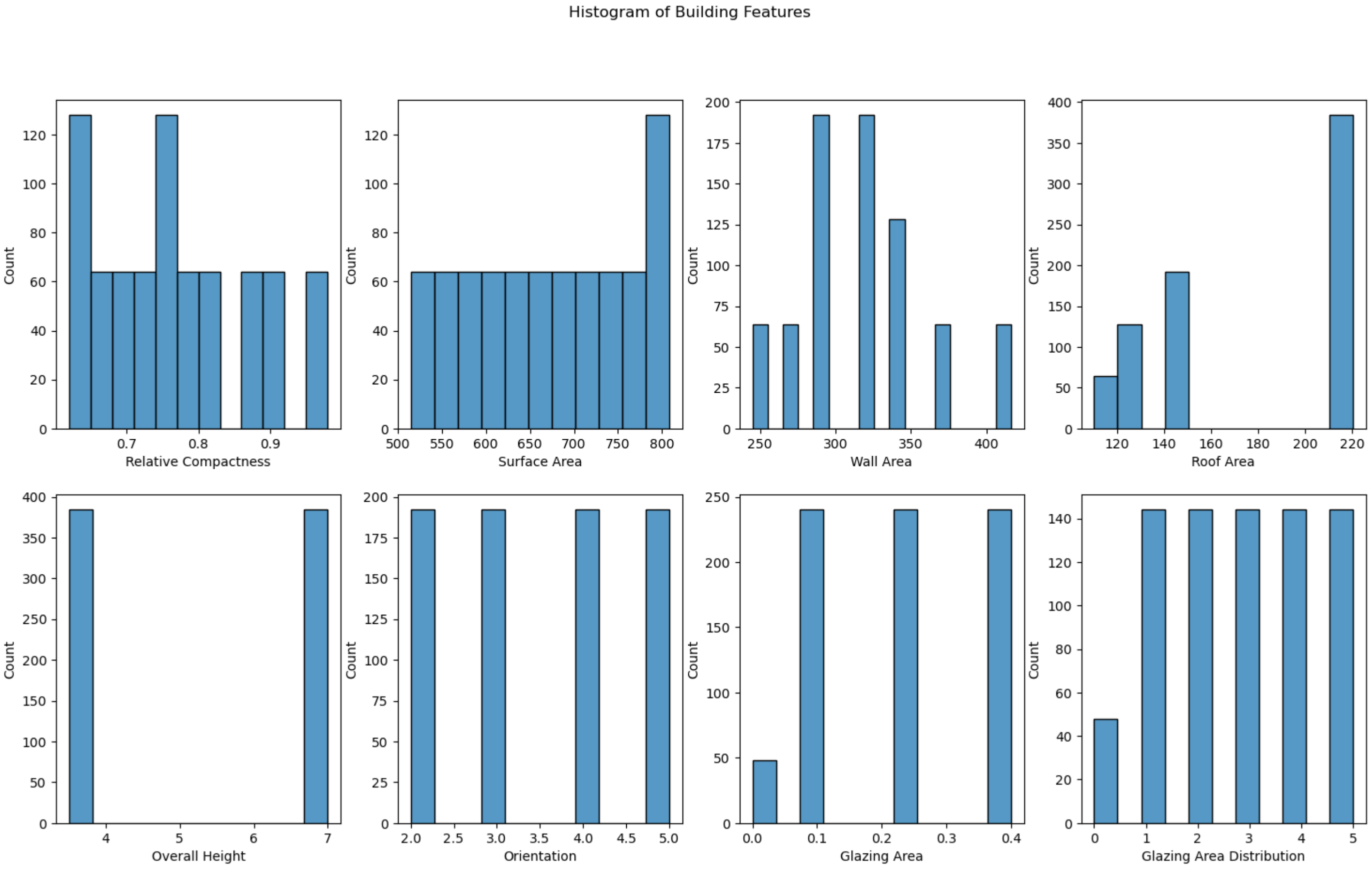


Figure 1. Histogram of Building Features

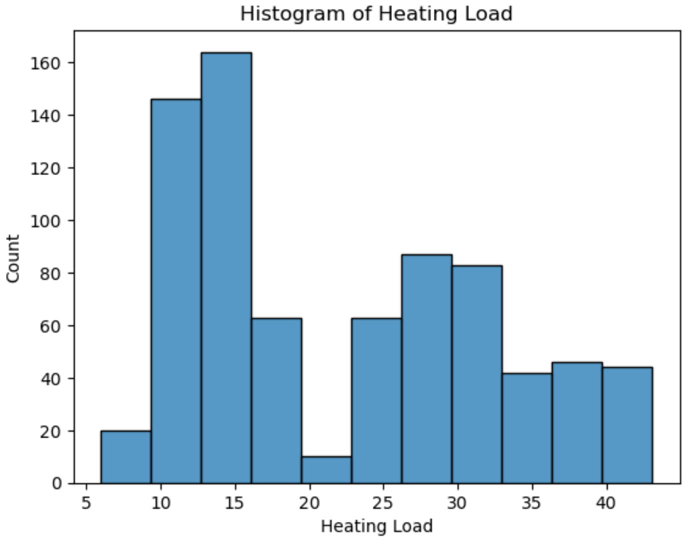


Figure 2. Histogram of Heating Load

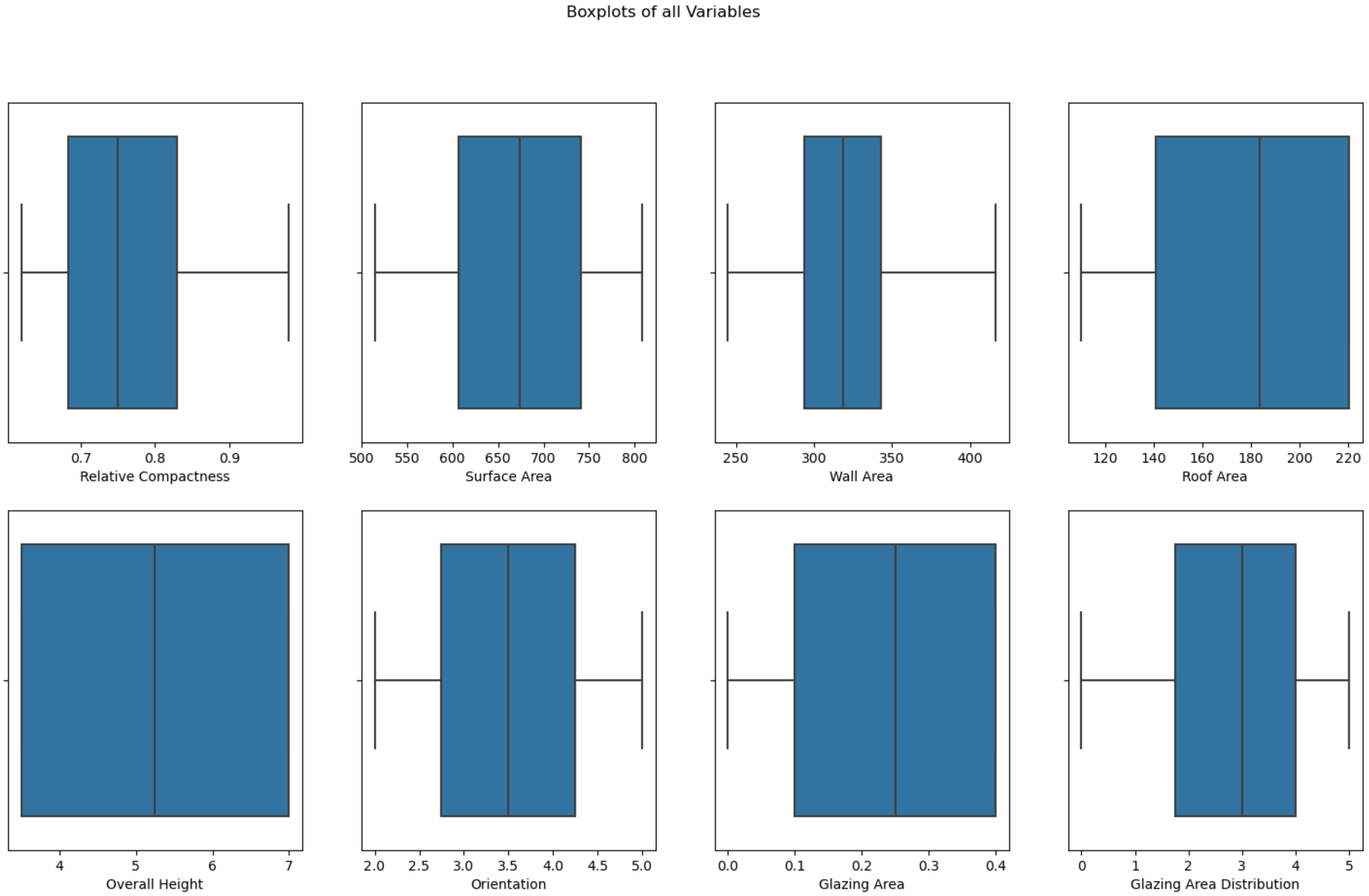


Figure 3. Boxplot of all building features

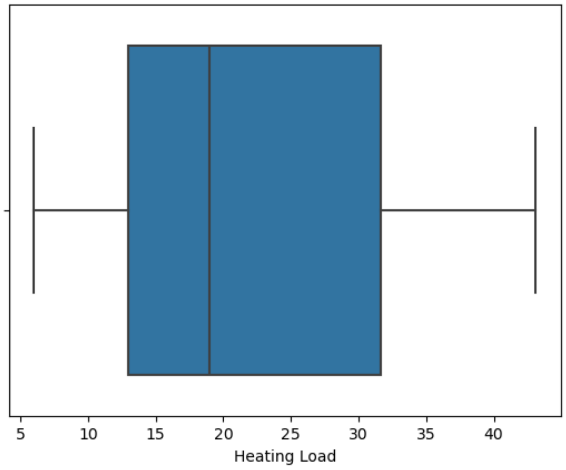
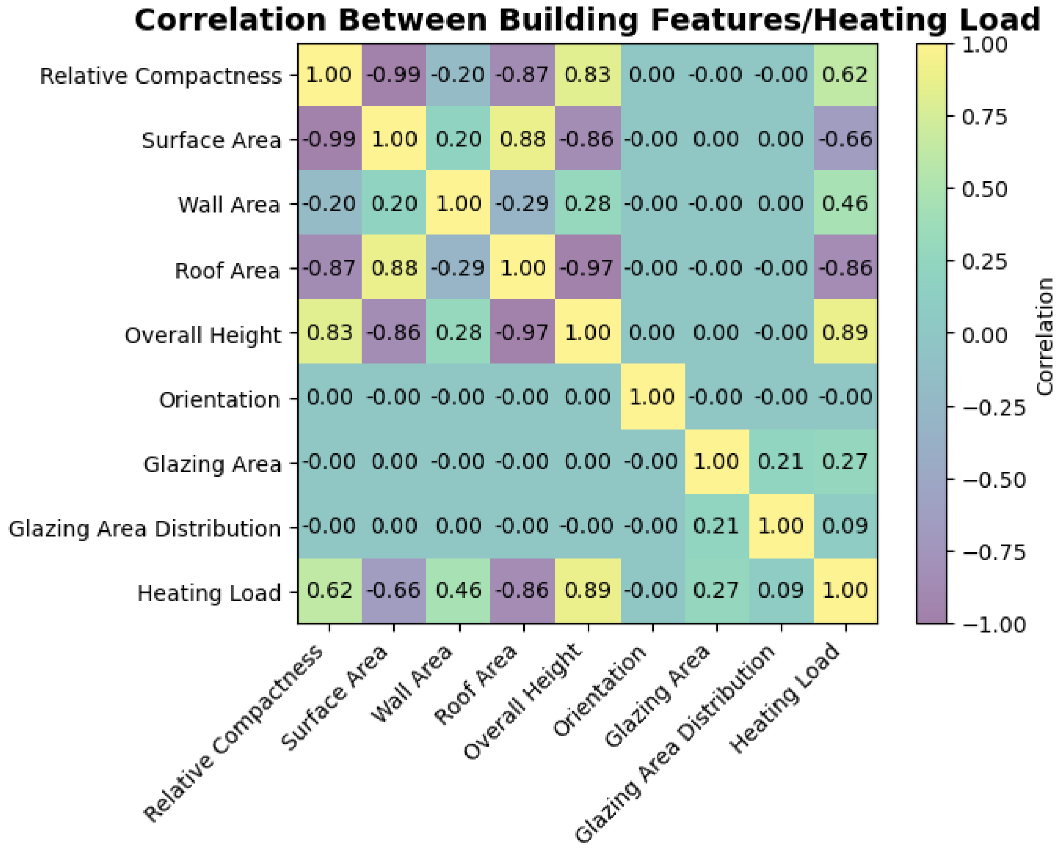


Figure 4. Boxplot of Heating Load

I also conducted bivariate analysis across building features and Heating Load to get a better understanding of the relationship between our X and y variables. Our pairgrid (see notebook) displayed all possible relationships between each of the variables (both X and y) in the dataframe. We were primarily interested in looking at the relationship between each X variable across the y variable (Heating Load). The scatterplot didn't provide us with clear information about the relationship of the X and y variables. The kernel density plot provided a better visual to analyze the relationship between building features and Heating Load, by displaying where points on a graph were densely populated. By looking at the kernel density plot, we were able to observe a potential positive correlation between Relative Compactness and Heating Load, Overall Height and Leading Load, as well as a negative correlation between Surface Area and Heating Load, and Roof Area and Heating Load. We explored the relationship between these variables further using a correlation matrix.

The correlation matrix (**Figure 5**.) displays a heatmap containing the correlation across multiple building features (X variables) and Heating Load (y variable). The numbers listed (-1 to 1) represent correlation where: 1 represents a very strong positive correlation (purple), 0 represents no correlation (bluegreen), and -1 represents a strong negative correlation (yellow). Although the graph displays the varying dependencies of building features, I was interested in the correlation of all building features across Heating Load. Orientation (correlation 0.0) possessed no correlation to Heating Load. Glazing Area Distribution (correlation 0.09) and Glazing Area (correlation 0.27) possessed almost no correlation to Heating Load. Wall Area (correlation 0.46) possessed some correlation to Heating Load. However, there were some build features worth exploring in further details. There was a moderate positive correlation between Relative Compactness (0.62) and Heating Load. There was a very strong positive correlation between Overall Height of building (0.89) and Heating Load. Additionally, the graph revealed a moderate negative correlation between Surface Area (correlation -0.66) and Heating Load. Lastly, the visual revealed a strong negative correlation between Roof Area (correlation -0.86) and Heating Load.



	Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazing Area Distribution	Heating Load
Relative Compactness	1.00	-0.99	-0.20	-0.87	0.83	0.00	-0.00	-0.00	0.62
Surface Area	-0.99	1.00	0.20	0.88	-0.86	-0.00	0.00	0.00	-0.66
Wall Area	-0.20	0.20	1.00	-0.29	0.28	-0.00	-0.00	0.00	0.46
Roof Area	-0.87	0.88	-0.29	1.00	-0.97	-0.00	-0.00	-0.00	-0.86
Overall Height	0.83	-0.86	0.28	-0.97	1.00	0.00	0.00	-0.00	0.89
Orientation	0.00	-0.00	-0.00	-0.00	0.00	1.00	-0.00	-0.00	-0.00
Glazing Area	-0.00	0.00	-0.00	-0.00	0.00	-0.00	1.00	0.21	0.27
Glazing Area Distribution	-0.00	0.00	0.00	-0.00	-0.00	-0.00	0.21	1.00	0.09
Heating Load	0.62	-0.66	0.46	-0.86	0.89	-0.00	0.27	0.09	1.00

Figure 5. Correlation Matrix of Building Features

Preprocessing/Feature Selection

After completing the data exploration process, I generated dummy variables for our discrete features, namely, Orientation and Glazing Area Distribution. Despite their integer nature, both Orientation and Glazing Area functioned like categorical variables. Thus, I employed the pandas.get_dummies() function to specifically choose Orientation and Glazing Area Distribution, transforming them into dummy variables. I also used dummy encoding on the Height variable, which only contained two unique values, 3.5 and 7. After dummy encoding our variables, the shape of our distribution changed, as we now had 17 columns (previous nine columns).

I utilized the MinMaxScaler which placed data between a range of 0-1. Our dataset was composed of variables with different units and extreme scale differences (e.g.: the scale of Surface Area vs scale of Relative Compactness). I split the data into training and test (using a 75/25) split. I performed fit only on the training data, mitigating the risk of data leakage and overfitting on the test set. After splitting and transforming data, I prepared the data for modeling through feature selection.

I needed to look for relevant and significant features to improve the model and reduce the complexity of the model. To analyze the correlation between our functionally categorical X variables, I discretized our y variable to be utilized for Chi-squared testing. I wanted to see which variables were strongly correlated with the target variable (Heating Load). Chi2 displayed both the F-score and p-value. A high F-score is indicative of the X variable being strongly correlated to the target variable. Using alpha of 0.05, I observed that the p-values of the Orientation and Glazing Distribution Area columns were not sufficient to reject the null hypothesis, which implied a weak relationship between these variables and Heating Load. However, the Height columns possessed extremely low p-values (1.593149e-55), which indicated a highly significant relationship with the target variable.

I took a look at all of the continuous variables. After looking at the correlation matrix for a second time, I observed that Glazing Area (correlation 0.27) possessed almost no correlation to Heating Load. Thus, I decided not to utilize this variable for the model, as this attribute wouldn't help predict linear regression model. Ultimately, I decided to include Relative Compactness, Overall Height, Surface Area, and Roof Area for our linear regression model (based on their associated strength to Heating Load).

Model/Results

I decided to utilize multiple linear regression to observe how Relative Compactness, Overall Height, Surface Area, and Roof Area could be used in the model to predict Heating Load. After identifying the X variables I wanted to include in the model, I conducted a train/test/split (using 75/25 split). I also fit our linear regression training model. Our model received approximately 0.86 R-squared score, which is a score that indicates the goodness of fit for our linear regression model. According to our R-squared score, about 86% of the variation in our output can be explained by our X values.

I also utilized the Mean Squared Error (MSE), which takes the squared differences from the actual data points and predicted values. A lower MSE suggests a model with better performance. An MSE of 0 would suggest that a model perfectly aligns with the actual values (which is unrealistic in real world scenarios). I obtained an MSE of 15.23, which suggests a moderate level of error relative to the scale of Heating Load (values ranging as low as 5 and as high as 43). It is important to take into consideration our model's R-squared value and MSE in determining our model's overall effectiveness.The graph above displays our predicted Heating Load values and actual Heating Load values.

Our graph (**Figure 6.**) shows how some of our predicted values overestimated our actual values (e.g.: Predicted Heating Load of 11 vs Actual Heating Load of 5). In other instances our model underestimated our actual values (e.g.: Predicted Heating Load of 30 vs Actual Heating Load of 42.) Our model is able to closely predict some of our actual values (e.g.: Predicted Heating Load of 15 and Actual Heating Load of 15). Looking at our model's scatterplot (of predicted vs actual values), we can see that our model can use additional refinement to better predict our target variable.

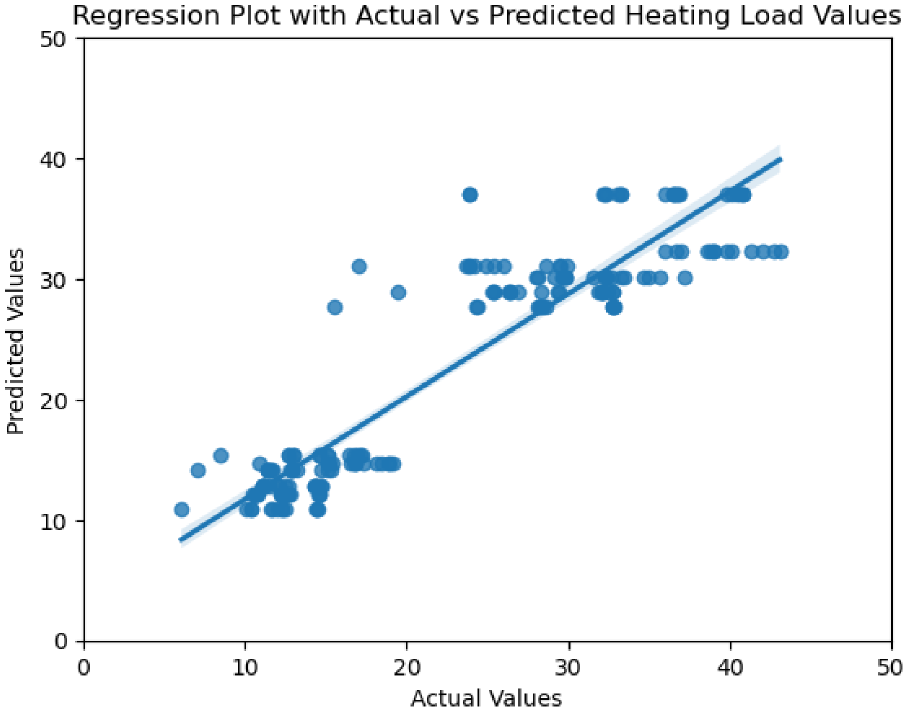


Figure 6. Regression Plot with actual vs. predicted values

Future Discussion

Additional factors such as a building's city, occupancy size, and usage (e.g., apartment versus studio) can be studied to further understand how to maximize energy efficiency. Exploring the relationship between building energy consumption and heating load, as well as local regulations, offers insights for architects, engineers, and policymakers striving to advance environmental and building sustainability. These findings denote the importance of a well-rounded approach to building analysis, as well as a call for collaborative efforts to create effective solutions to fight energy waste.