SDS 322E

Energy Efficiency

Fall 2022

Team: Anchita Raybarman, Shalini Das, Abdullah Hammamy, Julian Huang

Introduction and Objectives

Predicting accurate thermal load requirements during building construction is one of the most effective tools for optimizing energy usage and improving the economic value added for any property owner. In our project, we focused primarily on residential building types (ie. single family homes) and how to minimize HVAC costs and usage in this particular setting. Our project builds upon existing simulated data of 12 different building shape types with 8 common building parameters and their respective effects on heating and cooling load. Energy efficient buildings and homes can be achieved by lowering the heating and cooling load (AKA thermal loads) and maintaining building design parameters within an acceptable range.

For this project, our objectives were to: (1) find the impact of four specific variables – Building Surface Area [X2], Roof Area [X4], Overall Height [X5] and Glazing Area [X7] – on the heating and cooling loads. We chose to exclude the analysis of variables that have already been tested in prior literature. For example larger walls, regardless of surface or roof area, have been shown to have higher heating and cooling loads, resulting in lower energy efficiency (1). Lastly, we were motivated to (2) construct a model to predict and optimize energy efficiency of a residential building by determining the feature importance of building parameters to lower heating and cooling loads, respectively.

Data

This particular data set was donated to the UCI Machine Learning Repository by researcher, Athanasios Tsanas, and civil engineering industry professional, Angeliki Xifara, from the University of Oxford in 2012. Each object of the data set represents a distinct building type (modeled after one of 12 distinct residential building classes) that was generated in Autodesk Ecotect by simulating different values for each place of attributes [X1:X8]. Additionally, a value for heating load [Y1] and cooling load [Y2] were provided for each building.

In order to clean and prepare this dataset we needed to convert the original file from an .xlsx to .csv format. After being imported to R, however, the data set consisted of 1296 objects. To tidy this dataframe we chose to convert variables X1:X8 from a numerical to categorical data type. Once missing values were dropped in R, we were able to secure a dataframe with the original 768 building shapes simulated by Tsanas and Xifara's original study "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools." We proceeded with this relatively simple, but tidied version of the original data for the next portion(s) of the project.

Exploratory Analysis

For our exploratory analysis, we made visualizations of relationships between 4 independent variables and the heating and cooling loads using box plots in order to create 4

hypotheses. The independent variables that were used are Surface Area, Roof Area, Overall Height and Glazing Area of the buildings. Moreover, we will use the hypotheses to find the relationship of each of the independent variables with the energy efficiency.

Building's Surface Area relationship with heating and cooling loads

Figures 1 and 2 show that as the surface area of a building increases, the heating/cooling loads initially remain constant at around 30 BTU. However, after the surface area is higher than 661.5 lengths^2, we observe a significant decrease in heating/cooling loads. After the decrease, the heating/cooling loads remain constant at approximately 13 BTU-15 BTU. Even though there are outliers in the box plot, these outliers are within the range of heating/cooling loads of other box plots in **Figures 1 and 2.** Therefore, if the surface area of a building is larger than 661.5 lengths^2, the energy efficiency is high.

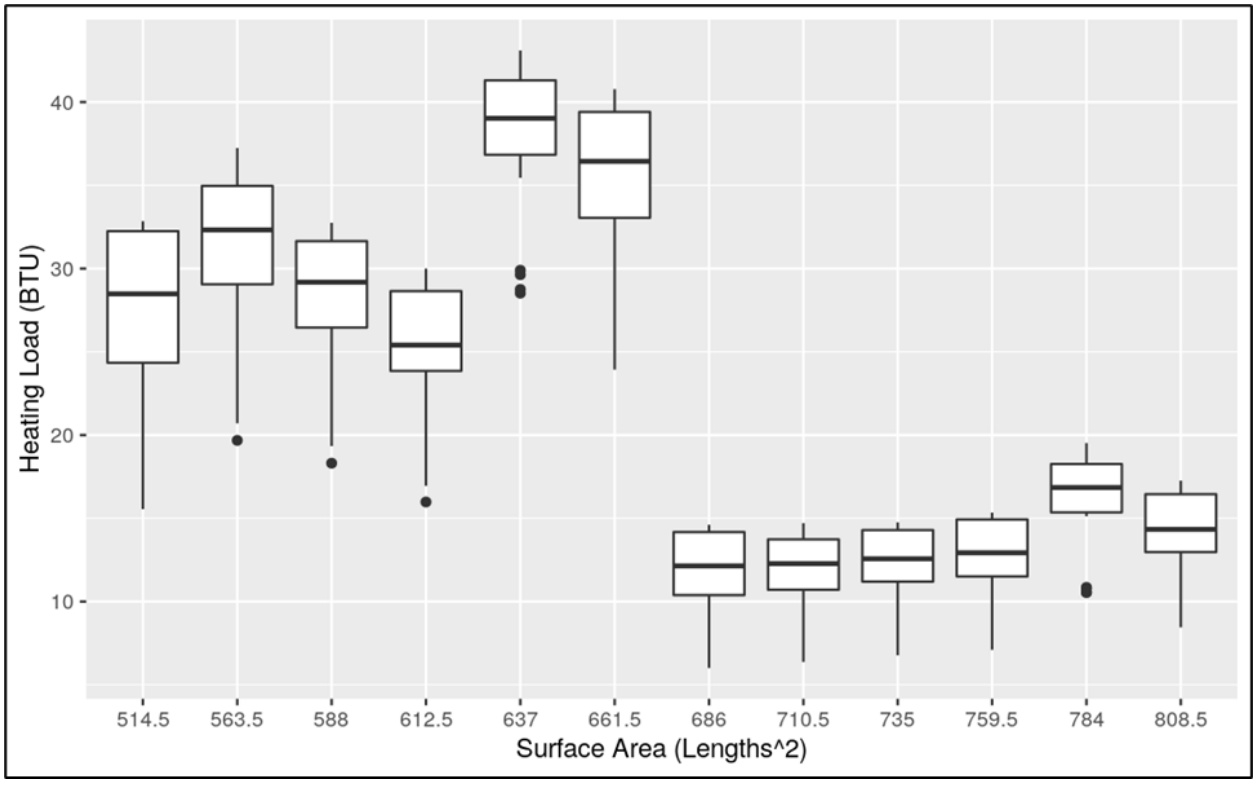

Figure 1. Relationship between Surface Area and Heating Load

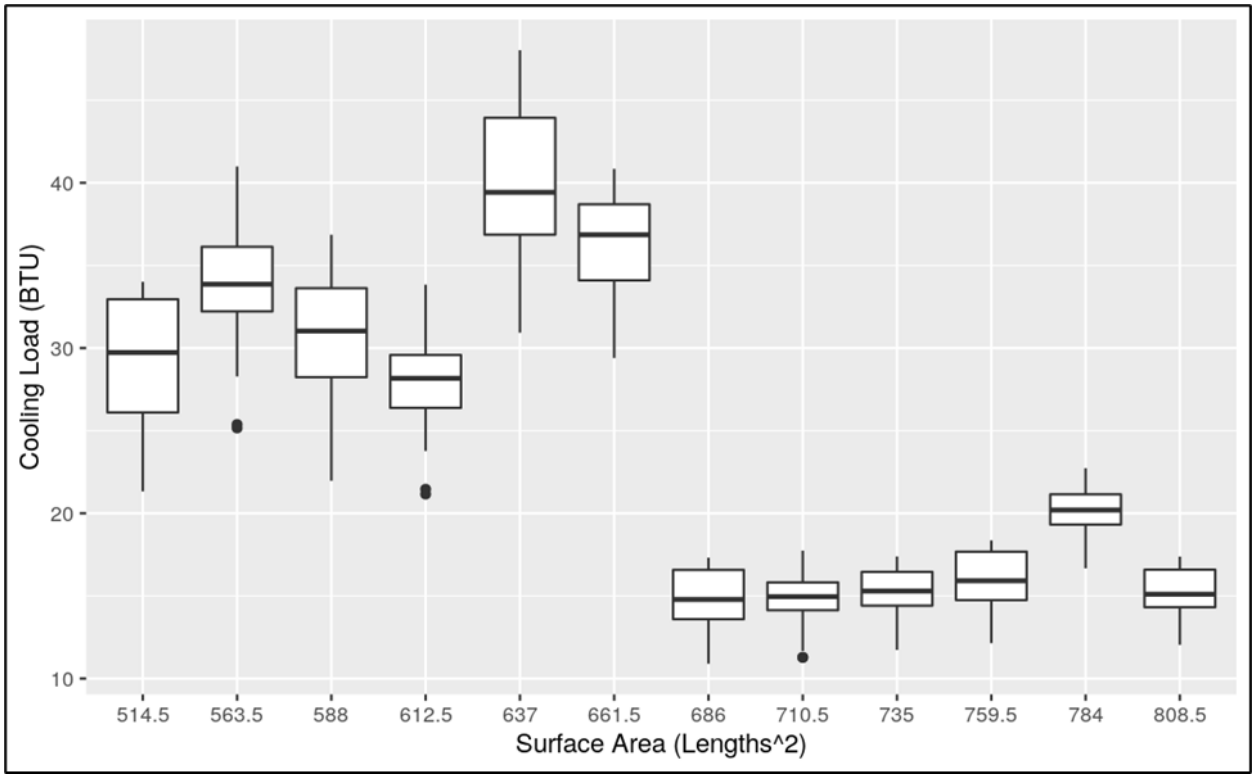

Figure 2. Relationship between Surface Area and Cooling Load

Building's Roof Area relationship with heating and cooling loads

Figures 3 and 4 shows that a similar pattern to that of **Figures 1 and 2. Figures 3 and 4** shows that as the roof area of a building increases, the heating/cooling loads initially remains constant at around 30 BTU. However, when the roof area is higher than 147 lengths^2, we observe a decrease in heating/cooling loads, where the heating/cooling loads becomes around 13 BTU-17 BTU. Overall, this indicates that buildings with roofs that have an area larger than 147 lengths^2 have a higher energy efficiency.

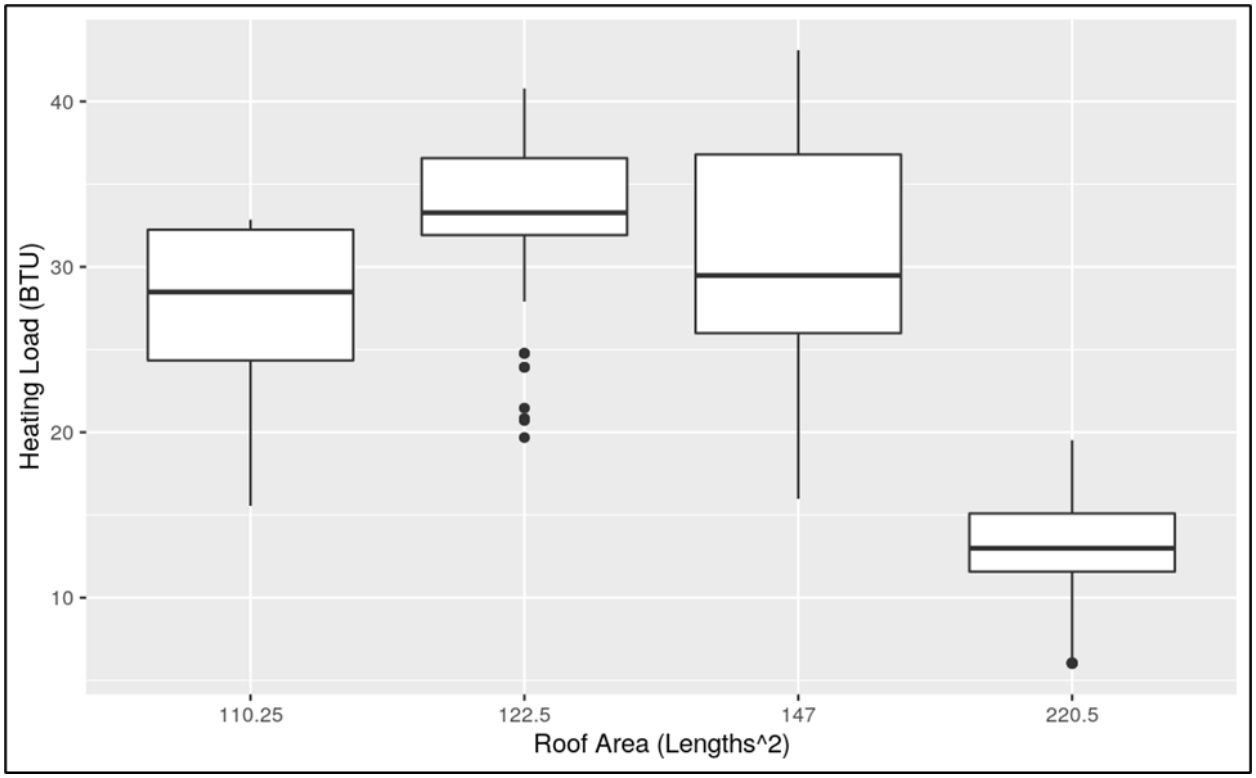

Figure 3. Relationship between Roof Area and Heating Load

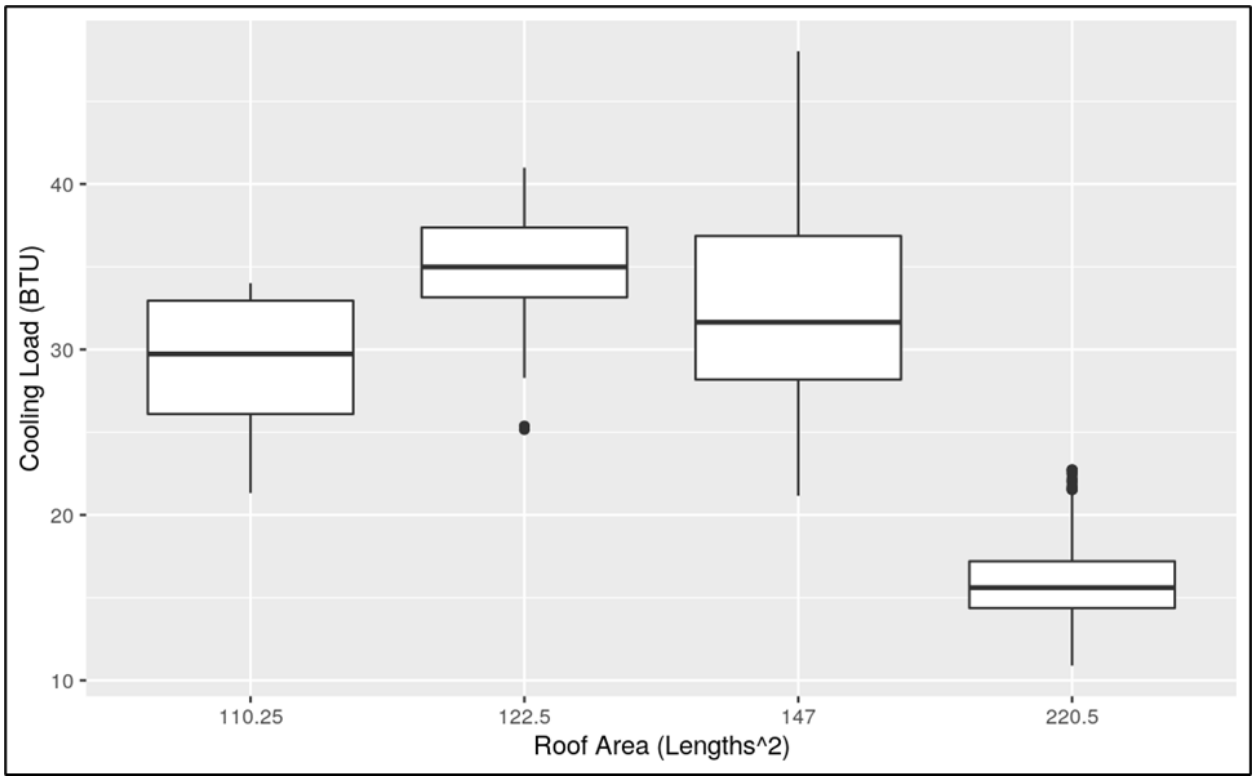

Figure 4. Relationship between Roof Area and Cooling Load

Building's Overall Height relationship with heating and cooling loads

Figures 5 and 6 shows that as the height of a building increases, we observe an increase in heating/cooling loads. This indicates that buildings with a relatively larger height have a lower energy efficiency.

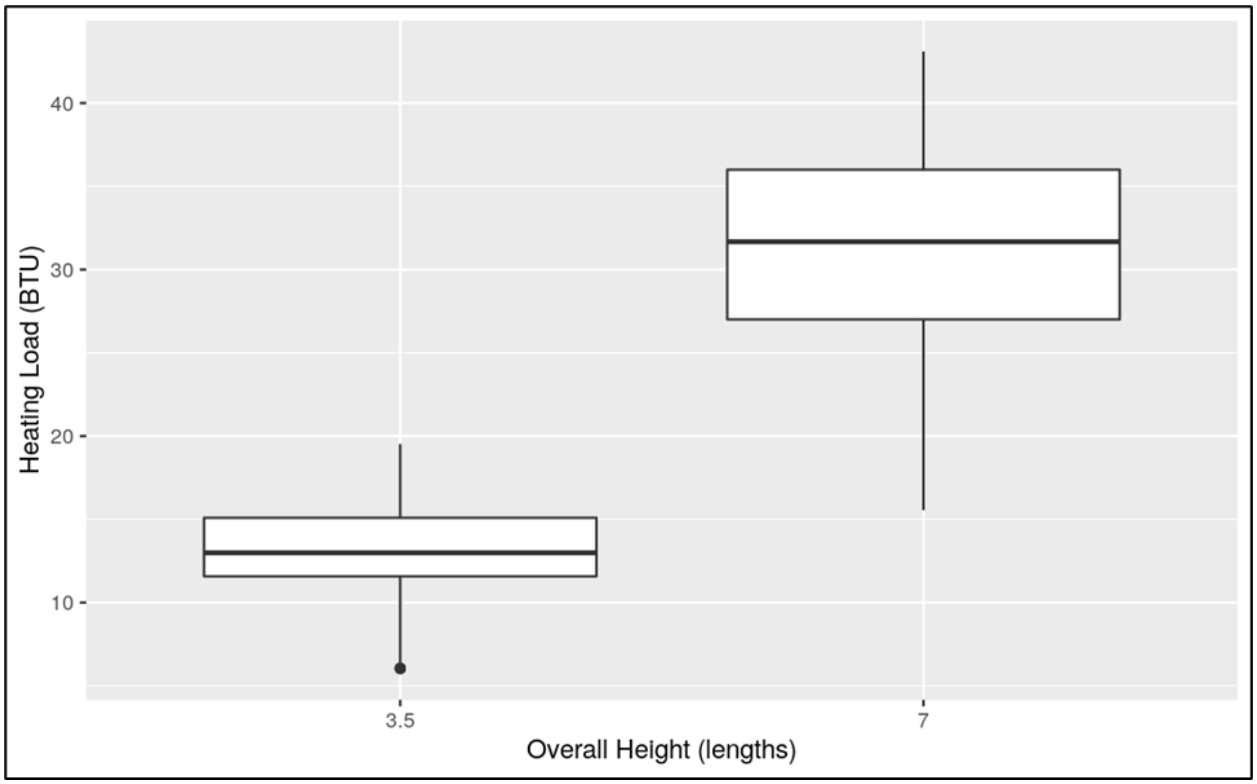

Figure 5. Relationship between Overall Height and Heating Load

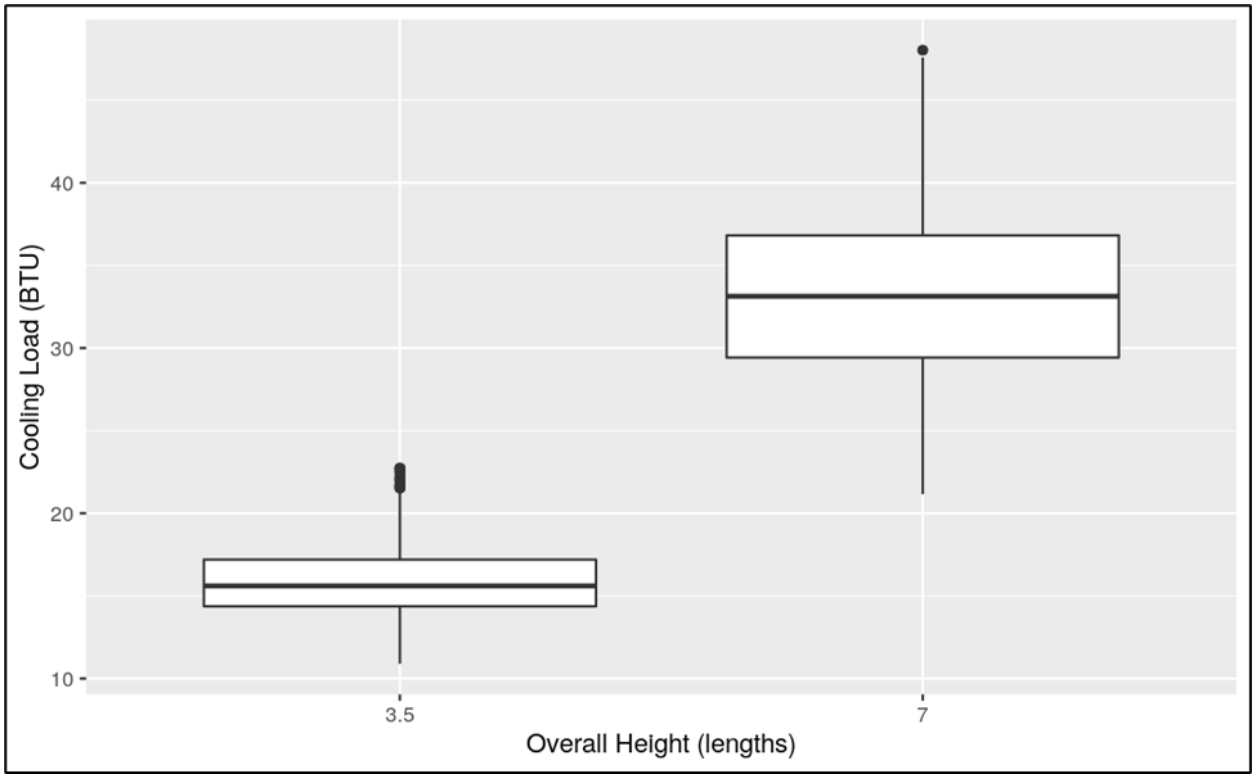

Figure 6. Relationship between Overall Height and Cooling Load

Building's Glazing Area relationship with heating and cooling loads

Similar to the overall height of a building, the higher the glazing area of the building is, the higher the heating/cooling loads are. **Figures 7 and 8** shows that there is a constant increase in heating/cooling loads. Research has shown that glazing area "is the most important predictor for both heating load and cooling load" [2]. Overall, a building with a relatively high glazing area leads to lower energy efficiency.

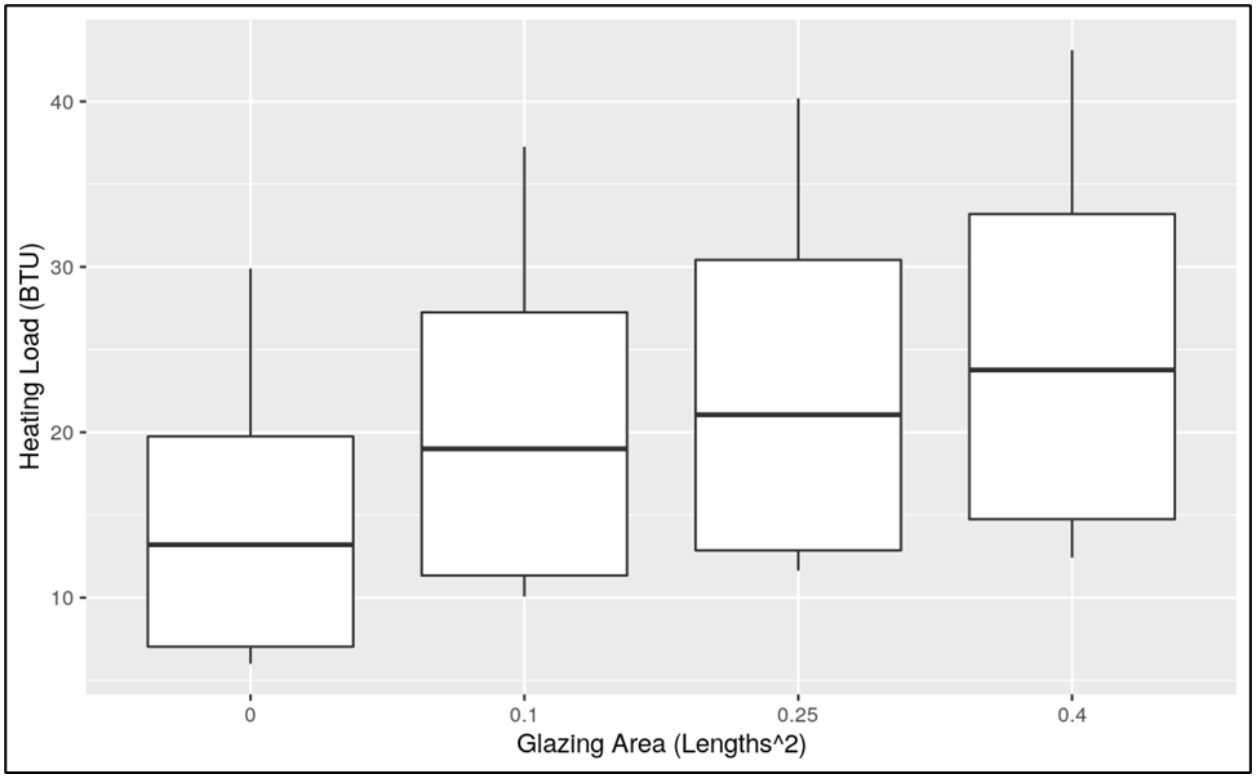

Figure 7. Relationship between Glazing Area and Heating Load

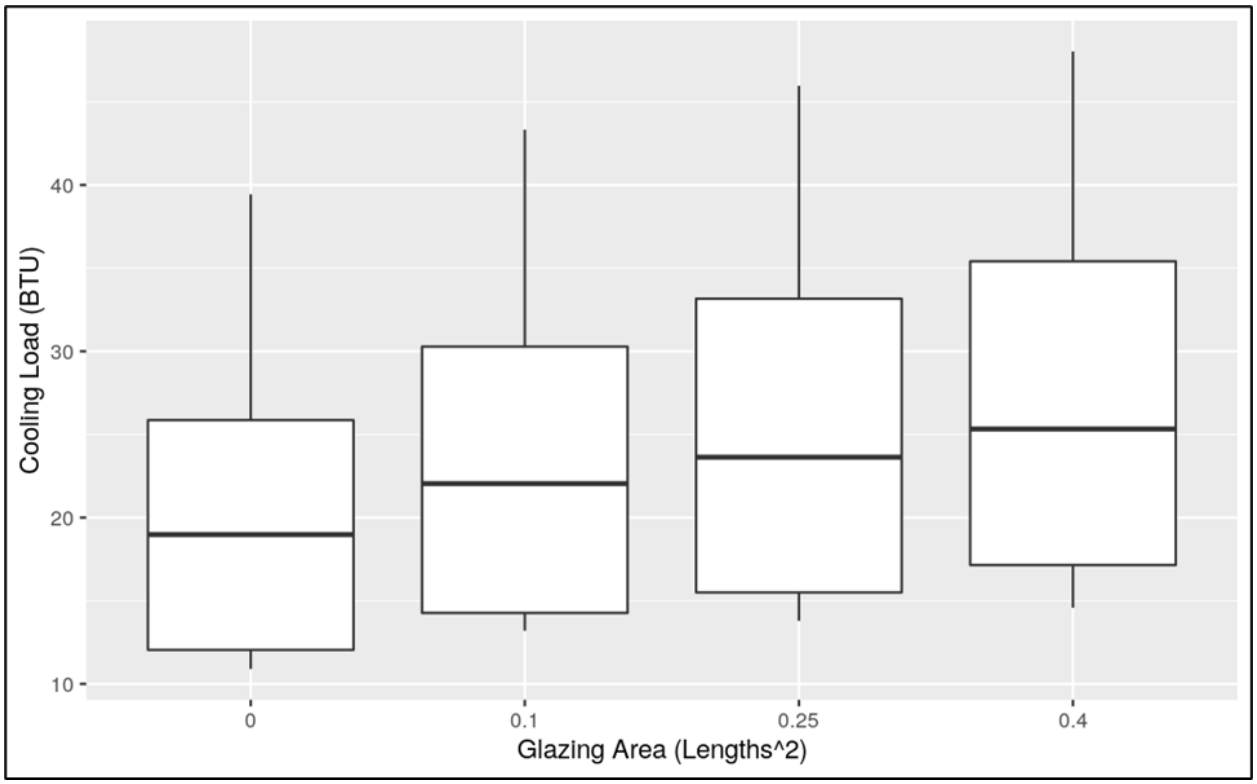

Figure 8. Relationship between Glazing Area and Cooling Load

In addition to the boxplots, we have also created a correlation matrix (**Figure 9**) for the Building's surface area, roof area, overall height and glazing area variables.

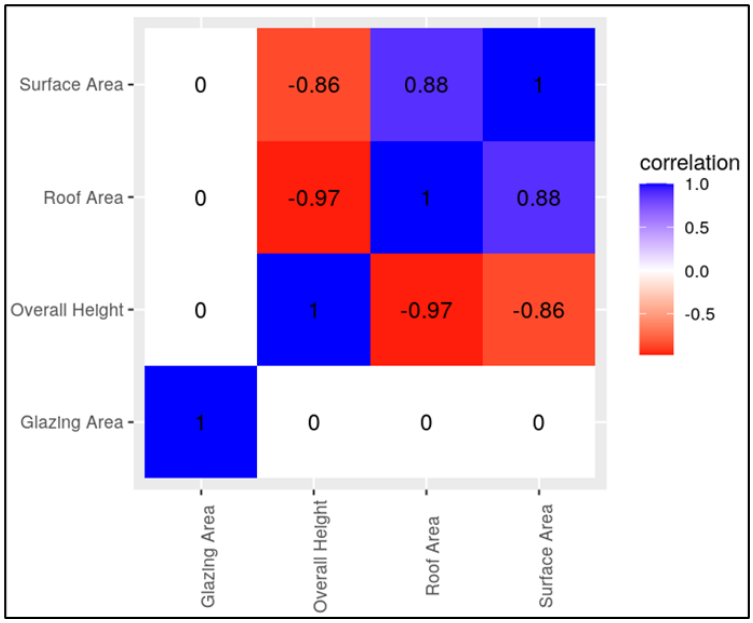

Figure 9. Correlation Matrix

In **Figure 9**, we see that the 'Surface Area' and 'Roof Area' variables are positively correlated variables and that might explain the reason the variables have similar relationships with heating and cooling loads. 'Overall Height' is negatively correlated with both 'Surface Area' and 'Roof Area' and this might be why the 'Overall Height' relationship with heating and cooling loads are the complete opposite to the 'Surface Area' and 'Roof Area' variables. Lastly, 'Glazing area' has no correlation with any of the three independent variables.

In summary, the 4 hypotheses that we made based on visualizing the data are as follows:

1. As the surface area of the building increases, the heating/cooling loads initially remain nearly constant at around 30 BTU but when the surface area is larger than 661.5 lengths^2, there is a sharp decline in the heating/cooling loads. The decline leads to a higher energy efficiency.

2. As the roof area of the building increases, the heating/cooling loads initially remain nearly constant at around 30 BTU but when the roof area is larger than 147 lengths^2 there is a decline in the heating/cooling loads; this results in an increase in energy efficiency.

3. As the overall height of the building increases, the heating/cooling loads increase which leads to a lower energy efficiency.

4. As the glazing area of the building increases, the heating/cooling loads increase constantly which leads to a lower energy efficiency.

Modeling

We approached setting up the predictive task for this project by attempting to determine the best model to reduce HVAC costs/usage for residents, owners and building designers/engineers. Assuming that being more energy efficient is closely tied to different parameters in building design over a range of continuous values – we opted for a regression task over classification in this project. A well performing regression model could be instrumental in helping determine the ideal parameters needed to optimize energy efficiency as accurately as possible.

For the baseline model we opted to use all eight original building parameter variables [X1:X8]. We used the scikit-learn library available in Python and used [Y1:Y2] as the target in our training data and visualized the results using Matplotlib as seen in **Figure 10.**

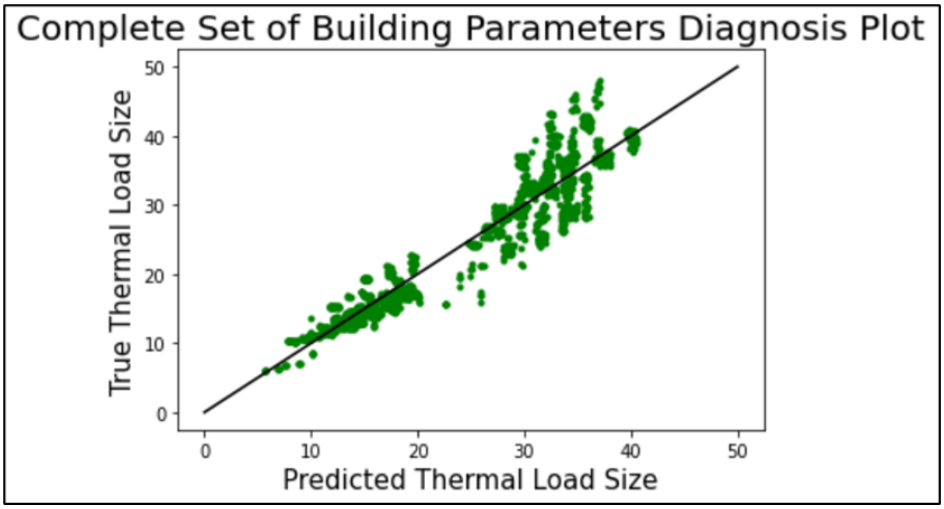

Figure 10. Linear Regression Plot Using Complete Set of Building Parameters

We also wanted to conduct a second linear regression using the building parameters we have isolated throughout this project for comparison. Visually speaking, as seen in **Figure 11**, there was much greater variance between the original data and the fitted linear regression. On further inspection, this second linear model using our subset, yielded a higher mean squared error in comparison to **Figure 10**. However, by constructing **Figure 11**, we revealed to ourselves our earlier selection bias when choosing glazing area as one of the parameters. Along with orientation and glazing distribution, glazing area was in the bottom-most tier of weakest correlation values in respect to heating and cooling load when we conducted our exploratory analysis. While our goal was to reduce noise by subsetting certain building parameters from the data set we had unknowingly lost important features, such as wall area and relative compactness.

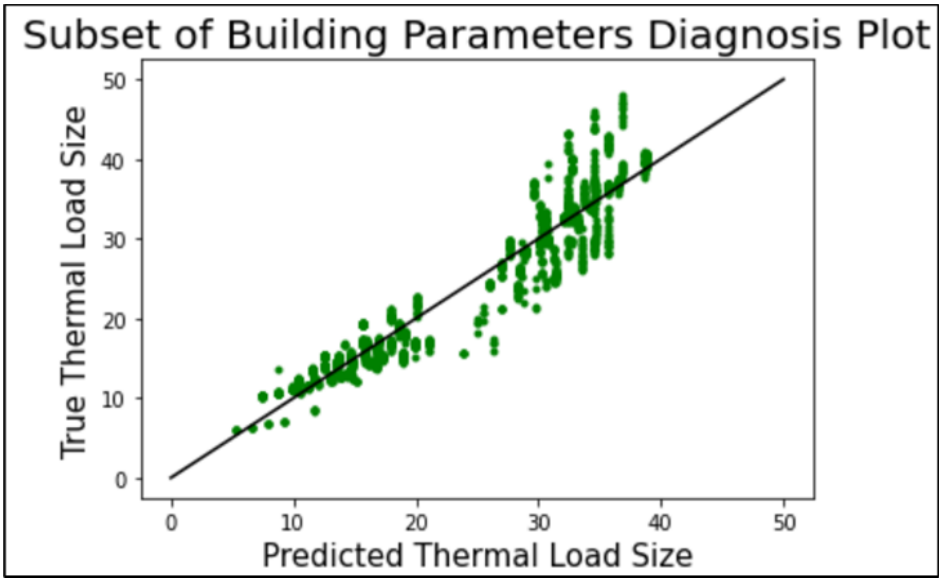

Figure 11. Linear Regression Plot Using Subset of Building Parameters

Our last model was a random forest plot using ensemble methods in machine learning from class. Using the complete set of building parameters, we set the number of trees to 1000 and random state to 42 to optimize the performance of the fitted model. As seen in **Figure 12**, the difference between predicted and true thermal load size narrowed significantly which was reflected when measuring performance metrics of each model.

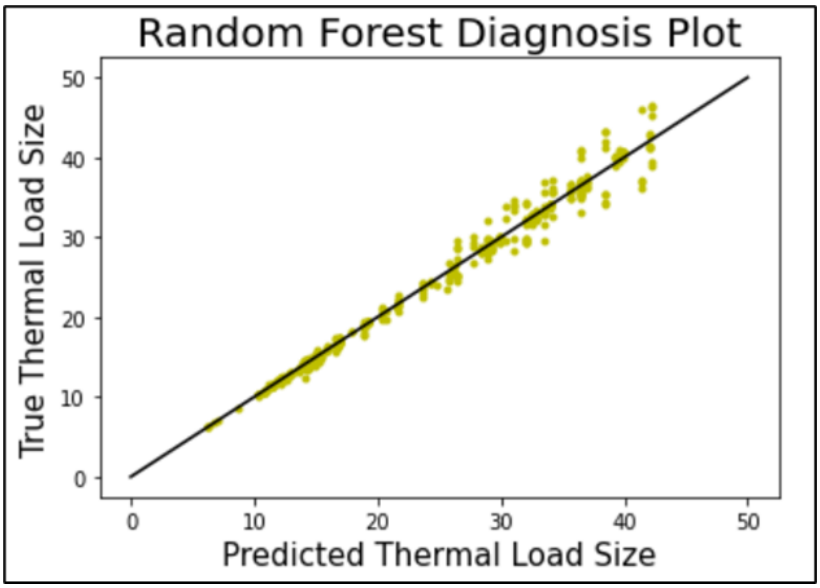

Figure 12. Random Forest Plot Using Complete Set of Building Parameters

Table 1: Summary Statistics

Metrics Observed	Complete Set of Building Parameters [X1: X8]	Subset of Building Parameters [X2, X4, X5, X7]	Random Forest Regressor [X1: X8]
MSE (Y1)	8.51	9.06	0.21
MSE (Y2)	10.14	10.69	2.95

Summarized above in **Table 1** are the cost functions we used to define and evaluate each regression with. Mean squared error offers more insight into the quality of each predictive model and is derived by measuring the Euclidean distance between true values and estimated values (for thermal loads, Y1 and Y2 in this case). MSE is a positive value in which the loss for a model decreases as it approaches 0.

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2.$$

In order to further predict the ideal parameters for minimizing thermal load, improving energy efficiency and minimizing HVAC costs, we would like to explore a quantile random forest regression and gradient boosting algorithm.

Discussion

In an attempt to create a more accurate model for our project, we chose 4 building parameters out of the complete set to create a regression around. The four building parameters that we selected—which were surface area, roof area, overall height, and glazing area—were chosen because we believed these 4 factors to be the most influential in predicting thermal load. Out of the four parameters selected, we found that overall height had the highest importance in determining thermal load at 0.47, then surface area at 0.31, and then roof area and glazing area at 0.15 and 0.07, respectively.

At first glance, with a quick visual inspection of **Figures 10 and 11**, there does not appear to be a significant change in the performance of the model between the selected building parameters and the baseline using the complete set of building parameters. However, looking at the respective summary statistics of the 2 regressions—as found in **Table 1**—we found that the accuracy decreased by 1.08% for heating load, and 0.88% for cooling load; as well as the MSE

being increased by 0.55 for heating load, and 0.55 for cooling load when we ran the model on the 4 selected building parameters compared to the baseline. While not a major change in the performance, this indicates that the model using the complete set of building parameters is the slightly better of the two.

In contrast, upon visual inspection and comparison of the Random Forest Regressor model in **Figure 12** compared to the previous **Figures 10 and 11**, it does appear that there is a significant difference in the performance of the model relative to the other two. When we compare the graph for the Random Forest Regressor model with the other two, the data points are in closer proximity to the regression line overall, leading us to believe that there is an improvement in performance with this model versus the others. Looking at the Random Forest Regressor model's summary statistics in **Table 1** confirms this, as we see a significant increase in accuracy to 98.45% and 96.26%—and a decrease in MSE to 0.21 and 2.95—for heating load and cooling load, respectively. This amounts to an additional increase in accuracy of 8.19% to heating load and 5.17% to cooling load—as well as a decrease in MSE of 8.3 for heating load and 7.19 for cooling load—when compared to the baseline model, which was established to be the more accurate of the 2 other models in our findings. This is a rather significant improvement in model performance, and, therefore, we concluded that the Random Forest Regressor model was the best model that we found for the purposes of accurately estimating the actual thermal load size.

Limitations

A major limitation with our project is the fact that the data set we are pulling from was simulated in Ecotect under very specific conditions in a very controlled environment, so it's possible that the model's accuracy or effectiveness is impacted when used on data from a different environment with different conditions.

These conditions include:

- The buildings being made solely of 18 elementary (3.5 x 3.5 x 3.5) cubes
- Material being the same/uniform for all 18 elements
    - "The materials used for each of the 18 elements are the same for all building forms. The selection was made by the newest and most common materials in the building construction industry and by the lowest U-value. Specifically, we used the following building characteristics (the associated U-values appear in parenthesis): walls (1.780), floors (0.860), roofs (0.500), windows (2.260)" (Tsanas, Xifara).
- Buildings are assumed to be "in Athens, Greece, residential with seven persons, and sedentary activity (70 W). The internal design conditions were set as follows: clothing: 0.6 clo, humidity: 60%, air speed: 0.30 m/s, lighting level: 300 Lux. The internal gains

were set to sensible (5) and latent (2 W/m^2), while the infiltration rate was set to 0.5 for air change rate with wind sensitivity 0.25 air changer per hour. For the thermal properties we used mixed mode with 95% efficiency, thermostat range 19–24 °C, with 15–20 h of operation on weekdays and 10–20 h on weekends" (Tsanas, Xifara).

The fact that the dataset we created the model from was under such specific conditions means that our model is only proven to be accurate under these conditions, and may perform differently with data simulated in other situations and with other circumstances. For example, a building with a different material, being composed of shapes other than elementary cubes, or in different humidity or activity could create a very different dataset which we have neither based our model around, nor tested its accuracy for. As such, it is difficult to tell how effectively our model is able to accurately predict thermal load when using datasets that do not follow these conditions. More extensive testing with other datasets would have to be done in order to determine if our model is still the best fit for those scenarios.

In addition to these conditions, another limitation is the fact that overall height was only separated into 2 distinct categories, when it would be better to simulate a wide range of building heights to obtain a more thorough understanding of the relationship between height and heating/cooling load. Looking at **Figures 1 & 2**, as well as **Figures 3 & 4**, we observe a relationship between the respective independent variables and the dependent variables wherein the heating and cooling loads initially remain approximately constant, but change rather abruptly upon hitting specific breakpoints as the independent variable continues to increase. This could also potentially be the case with the overall height building parameter's relationship with thermal load, but we are unable to support or disprove this possibility due to the imprecise nature of the 2 distinct values for height.

Conclusion

Our goal with this project was to determine the relationship between thermal load and the 4 chosen parameters of surface area, roof area, overall height, and glazing area. From our exploratory analysis, we created four hypotheses of the relationship of the selected independent variables to the thermal load based on our data visualizations (**Figures 1-8**) and findings. In addition, we determined the relative feature importance of the 4 parameters in the modeling section, and found that the overall height of the building was the most influential factor in determining thermal load, followed by the surface area, then the roof area, and finally the glazing area.

Using the simulated dataset from Ecotect, we then created 3 separate models, as found in **Figures 10-12**, which were designed to predict thermal load given data from the building

parameters provided. The goal of creating these models was to determine the accuracy of our predicted thermal loads using the parameters chosen versus a baseline model using the complete set of building parameters. Of the 3 models, we determined that the Random Forest Regression was the best performing, given its significantly improved degree of accuracy compared to the other models tested—as seen in **Table 1**.

In terms of potential directions or improvements to make in the future, several could be made to widen the scope and applicability of the model we created. As previously mentioned, one of the biggest issues in the general application of our model is that the dataset it was based on was simulated using a very specific set of conditions, which means that we cannot determine its effectiveness in predicting thermal load on data that is simulated or recorded in other scenarios. To improve this, data could be collected from simulations under different conditions using the same or similar software as this dataset was collected under—or even from real world building data—allowing us to confidently apply the model to a wider range of datasets. One more improvement that could be made to improve our understanding of the relationships of the variables within the project itself—without changing the overall scope—would be to change the data simulated for overall height to be a continuous range of values rather than the 2 discrete categories we had of 3.5 and 7.

Acknowledgements

- **Anchita Raybarman**: introduction - specified objectives, organized github repository, presentation
- **Shalini Das**: data cleaning, modeling, introduction -
- **Abdullah Hammamy:** Exploratory analysis - box plots, correlation matrix, hypotheses, coding, and writing this section in the report. Overall height limitation, presenting data to the class.
- **Julian Huang:** Discussion and conclusion
- **Justin Kim:** Attempted to contribute on the last two days of project work

Team Member	Percentage Contribution
Anchita Raybarman	100%
Shalini Das	100%
Abdullah Hammamy	100%
Julian Huang	100%
Justin Kim	10%

References

1. Joshua Kneifel, Beyond the code: Energy, carbon, and cost savings using conventional technologies, Energy and Buildings, Volume 43, Issue 4,Pages 951-959,ISSN 0378-7788 (2011). https://doi.org/10.1016/j.enbuild.2010.12.019
2. A. Tsanas, A. Xifara: 'Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools', Energy and Buildings, Vol. 49, pp. 560-567, (2012). https://doi.org/10.1016/j.enbuild.2012.03.003
3. Ling, J., Zhao, L., Xing, J. *et al.* Statistical analysis of residential building energy consumption in Tianjin. *Front. Energy* 8, 513–520 (2014). https://doi.org/10.1007/s11708-014-0327-5
4. Yan, S., Li, X. Comparison of space cooling/heating load under non-uniform indoor environment with convective heat gain/loss from envelope. *Build. Simul.* 14, 565–578 (2021). https://doi.org/10.1007/s12273-020-0708-0
5. Obrinsky, M., & Walter, C. (2016). Energy Efficiency in Multifamily Rental Homes: An Analysis of Residential Energy Consumption Data. *The Journal of Sustainable Real Estate*, *8*(1), 2–19. https://www.jstor.org/stable/24876479
6. Marius Zumwald, Benedikt Knüsel, David N. Bresch, Reto Knutti, Mapping urban temperature using crowd-sensing data and machine learning, *Urban Climate*, Volume 35, (2021). https://doi.org/10.1016/j.uclim.2020.100739