**Introduction & Data**

As a consultant for eSC, we do a lot of work in regards to aiding them with many of their projects and work. Specifically we do a lot of work when it comes to predicting future energy usage. In particular, we currently want to predict energy consumption within the month of July for the houses amongst eSC’s coverage. With July typically being the hottest month of the year, it coincides with high energy usage amongst many people. Given this, we want to be as efficient with our energy usage as possible. However, more importantly we want to give eSC insight on how to potentially get their clients to save more energy by whatever means necessary. This would put less of a strain on their power grid, avoiding potential blackouts, and have potential benefits for the environment. At our disposal we have a multitude of housing metrics for the many houses in their coverage, the location of these houses across South Carolina, weather data from July 2023 within these regions, along with corresponding energy consumption statistics of these houses during this same time measured down to hour. Using these metrics we plan to give eSC a better sense of what to expect in future Julys in regards to the amount of energy they will have to supply.

**Business Questions/Objective**

How does the year a home was made affect its energy consumption during the month of July, and what cost-effective solutions can be implemented to improve

energy efficiency for homes of different ages? How does this differ for old homes (<1970s), middle aged homes (1970-2000), and newer homes (>2000s)

**Data Acquisition & Cleansing**

For the data we used three main sources. A file containing the list of all houses in the dataset (eSC’s coverage), along with house attributes such as size, materials, etc. A set of files containing the energy usage data, which was collected hour-by-hour, for the houses, with each house getting its own individual file. Lastly, there were file(s) that contained hour-by-hour weather information (one file for each geographic area).

Once the files were obtained, we split the house data into three different files based on the year the house was made. One for houses made before the 1970s, homes made between 1970-2000, and houses made after the 2000s. The files were then filtered for the preferred housing attributes. Every house was assigned a value in relation to this, called era.

The data was then filtered to include only one of the three possible era values. Next, a list of building IDs was created, with every house receiving a unique identifier. The energy data for the selected building, weather data for the county, and filtered static data for the house are then pulled and merged. The building ID is added to the energy data, which is subsequently merged with the weather data using date and time as a key. This merged dataset is then combined

with the house data using the building ID. The first building ID is removed from the list, and the process is repeated for the

remaining IDs using a for loop. Once all iterations are completed, the datasets are combined using rbind(). Finally, the data is filtered to only have the entries corresponding to July dates. Alternatively, there was also a dataset containing all of the entries throughout the year, with no filtering done for the month of July.

After the data has been split and formatted, the variables can then be selected. Our first step was to look through the metadata and see what the variables within the data represented. Then using this info, we cross checked with prior knowledge and additional information to see which variables would have the most significance. Then,the out./energy variables were examined to see which one would be selected as the dependent variable. There was no total energy consumption variable, so out.electricity.cooling.energy\_consumption was selected. Given that there is a focus on summer months and the effects on global warming, it made the most sense to select this as the outcome variable. The variables were then converted to type numeric. A correlation matrix was then created to see which variables had the highest correlation with our selected outcome. The top 28 variables were selected. These were said variables:

"Dry Bulb Temperature [°C]"

"Relative Humidity [%]"

"Wind Speed [m/s]"

"Wind Direction [Deg]"

"Global Horizontal Radiation [W/m2]"

"Direct Normal Radiation [W/m2]"

"Diffuse Horizontal Radiation [W/m2]"

"in.sqft"

"in.occupants"

"in.roof\_material"

"in.ducts"

"in.refrigerator"

"in.orientation"

"in.geometry\_stories"

"in.usage\_level"

"in.bedrooms"

"in.clothes\_dryer"

"in.clothes\_washer"

"in.cooking\_range"

"in.dishwasher"

"in.geometry\_building\_type\_height"

"in.geometry\_wall\_type"

"in.heating\_fuel"

"in.heating\_setpoint"

"in.hvac\_cooling\_efficiency"

"in.income"

"in.infiltration"

“hours”

**Modeling Methods & Visuals**

We started our modeling process by using four different models for each set:

Decision Trees (ctree()), Gradient Boosting (gbm()), Logistic Regression (glm()), Linear Regression (lm()). To prepare for the model, the hour data was extracted from the date and time variable. Then each data set is partitioned into a training and testing set at 0.7/0.3 splits. When making these models, we used Mean Absolute Error(MAE), Root Mean Square Root Error(RSME) and R Squared(R2) value to evaluate the performance. MAE is the average of the sum of all of the absolute values of the difference between the true value and the predicted value, with smaller values being more

Optimal. RMSE is the square root of the average of the squared differences of the

predicted and actual value, with lower values being favored. R-Squared is a

measurement that represents the correlation between the predictors and the response

variable, with higher values being the most optimal. We ran these models on the July and the Full Year dataset. Unfortunately, many of the devices used did not contain enough memory to fully run all the models on the Full Year dataset. This led to some models not being able to give adequate predictions, and some not able to run at all. Here are results for our corresponding models:

***Old(Full Year):***

*Decision Tree*

MAE: 0.0666971

RSME: 0.1340939

R2: 0.8271629

*Logistic Regression*

MAE: 0.1911806

RSME: 0.2912063

R2: 0.3985515

*Linear Regression*

MAE: 0.1911806

RSME: 0.2912063

R2: 0.3985515

*Gradient Boosting*

MAE: 0.1894238

RSME: 0.295824

R2: 0.3979404

***Old (July):***

*Decision Tree*

MAE: 0.1232393

RSME: 0.2045168

R2: 0.7037737

*Logistic Regression*

MAE: 0.1911806

RSME: 0.2912063

R2: 0.3985515

*Linear Regression*

MAE: 0.1911806

RSME: 0.2912063

R2: 0.3985515

*Gradient Boosting*

MAE: 0.1894722

RSME: 0.2958494

R2: 0.3981865

***Middle(Full Year):***

*Decision Tree*

MAE: 0.07897223

RSME: 0.1581843

R2: 0.817754

*Logistic Regression*

MAE: 0.1822417

RSME: 0.2835746

R2: 0.4139467

*Linear Regression*

MAE: 0.1822417

RSME: 0.2835746

R2: 0.4139467

*Gradient Boosting*

MAE: 0.1501468

RSME: 0.264628

R2: 0.499092

***Middle(July):***

*Decision Tree*

MAE: 0.1465791

RSME: 0.2376875

R2: 0.6890425

*Logistic Regression*

MAE:0.2190419

RSME: 0.3333306

R2: 0.3872922

*Linear Regression*

MAE: 0.2190419

RSME: 0.3333306

R2: 0.3872922

*Gradient Boosting*

MAE: 0.2145662

RSME: 0.3367807

R2: 0.3876322

***New(Full Year):***

*Decision Tree*

MAE: .08577422

RSME: .1746164

R2: .8110475

*Logistic Regression*

MAE: .1989929

RSME: .3091445

R2: .4073457

*Linear Regression*

MAE: .1990069

RSME: .3091445

R2: .4073457

*Gradient Boosting*

MAE: .1649724

RSME: .2882479

R2: .4942091

***New(July):***

*Decision Tree*

MAE: 0.1663364

RSME: 0.2768326

R2: 0.6612165

*Logistic Regression*

MAE: 0.2419463

RSME: 0.3747597

R2: 0.3781606

*Linear Regression*

MAE: 0.2419655

RSME: 0.3747598

R2: 0.3781606

*Gradient Boosting*

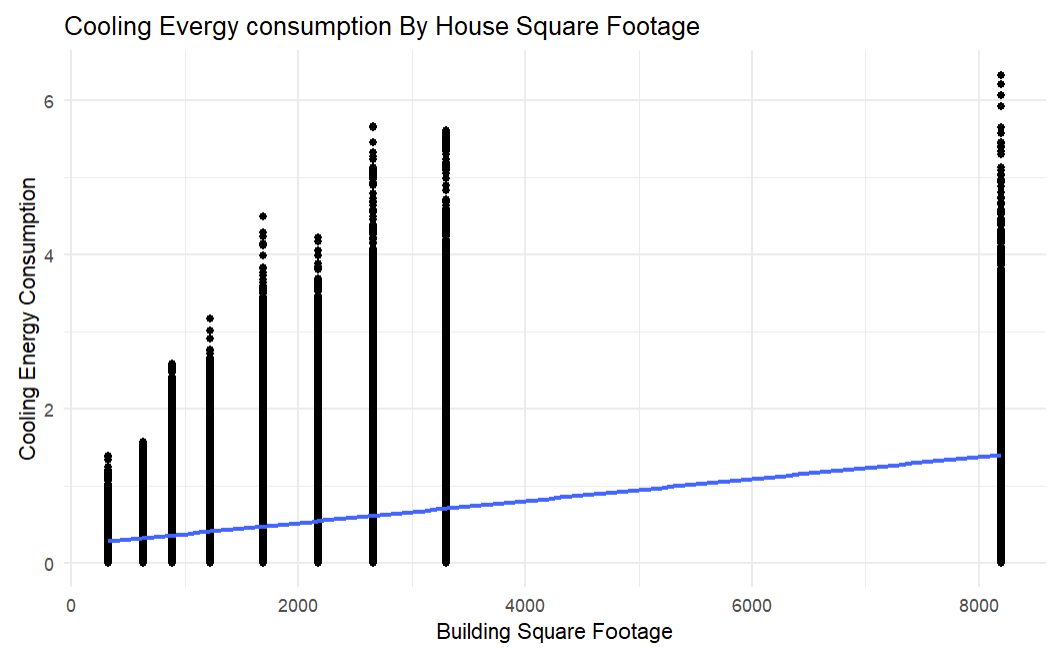
MAE: 0.2349473

RSME: 0.3767427

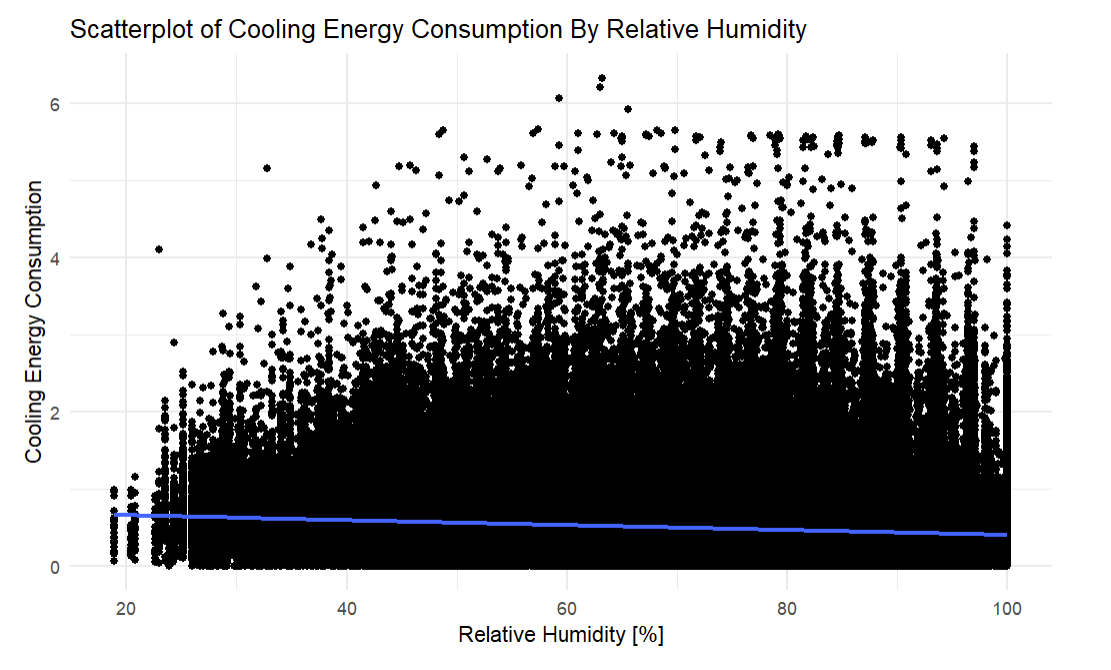
R2: 0.3844302

**Results & Visuals**

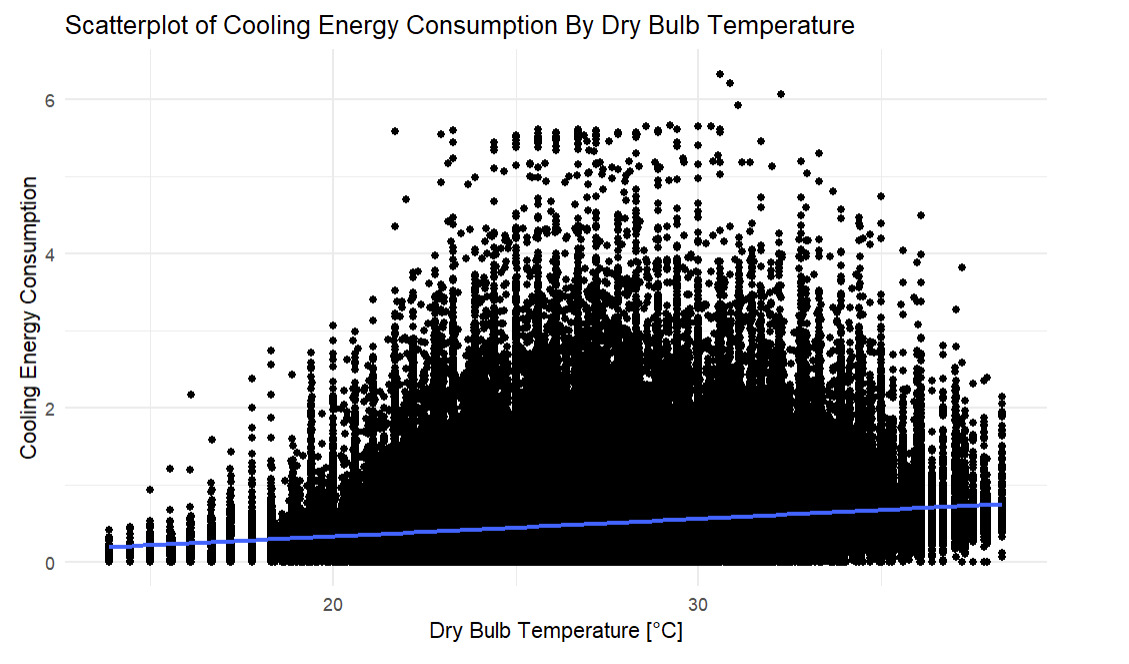
***Old:***

**

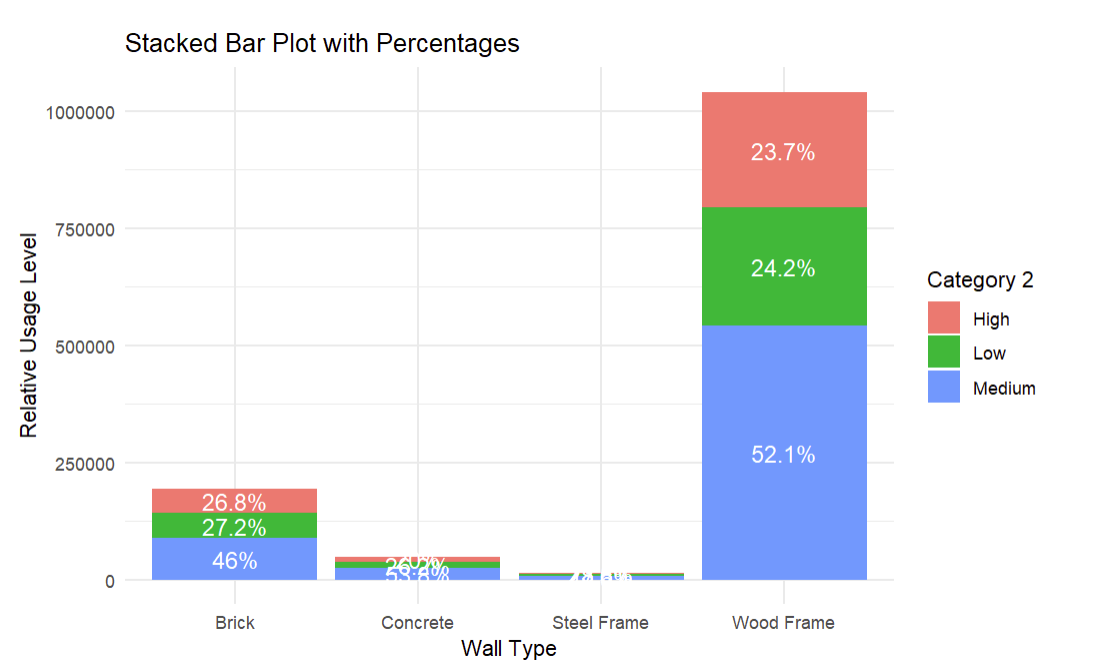
Based on the scatterplot between Cooling Energy Consumption and Building Square Footage, we can see that when a house is larger in size, there tends to be a larger amount of cooling energy consumed.



Based on the scatterplot showing Cooling Energy Consumption vs. Relative humidity, we can see that there tends to be a negative trend between the two variables.



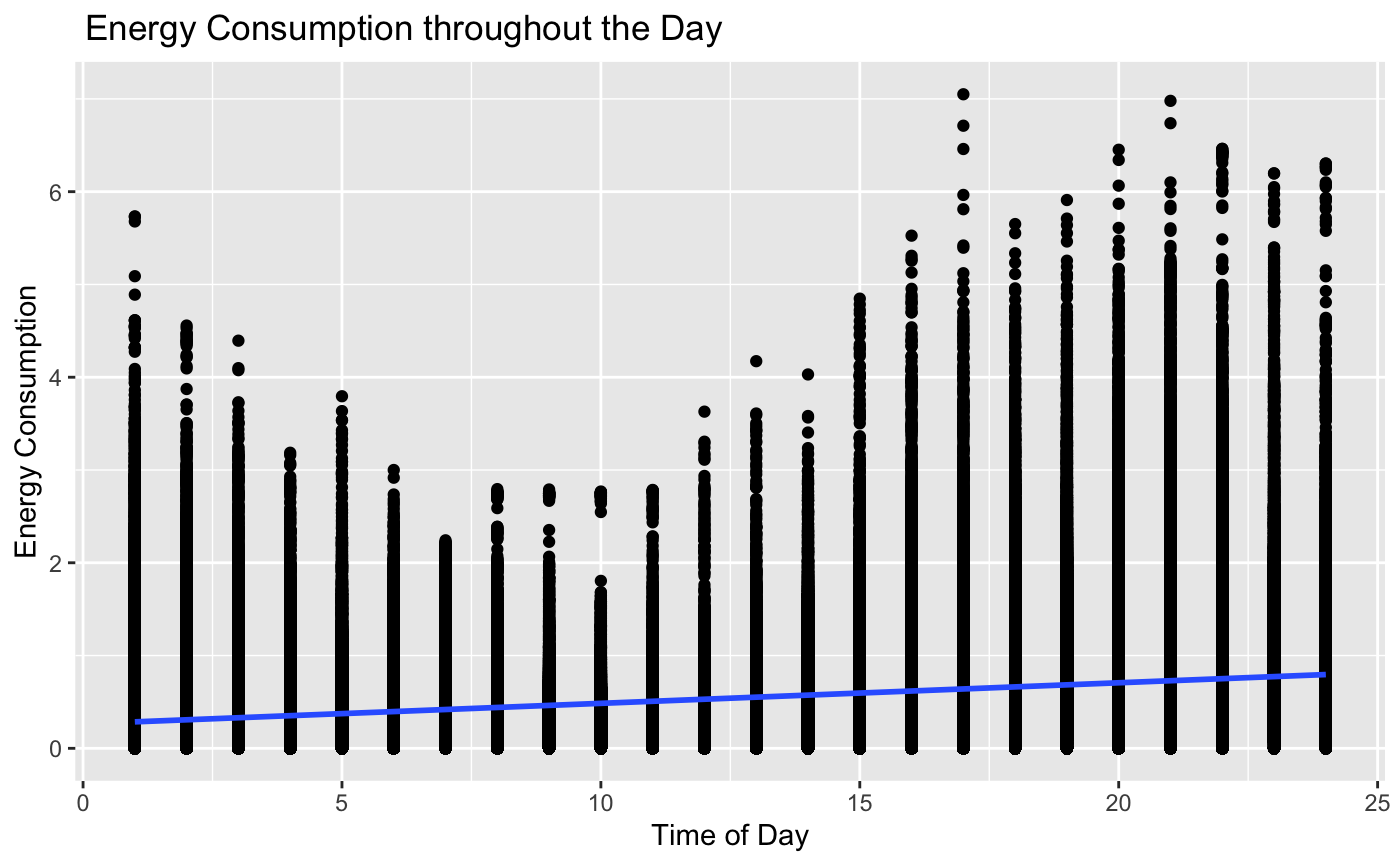
The scatterplot showing Cooling Energy Consumption vs. Dry Bulb Temperature in Degrees Celsius, shows us that there tends to be a positive trend between the two variables.



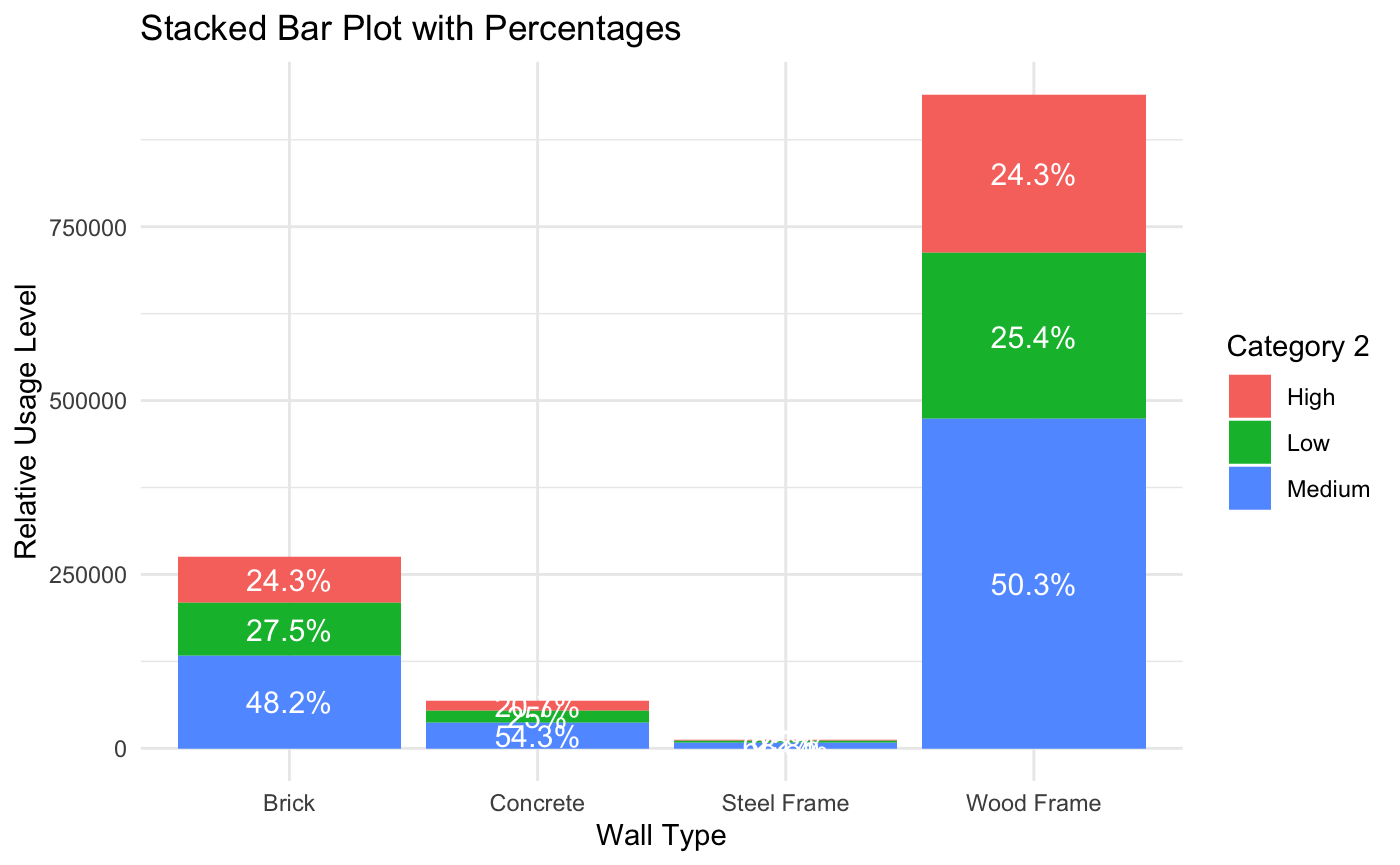
The main takeaway from this plot should be that wood frame wall types seem to be the most common out of all of the wall types. We can also see that the usage levels seem to be relatively consistent within each wall type.

***Middle:***

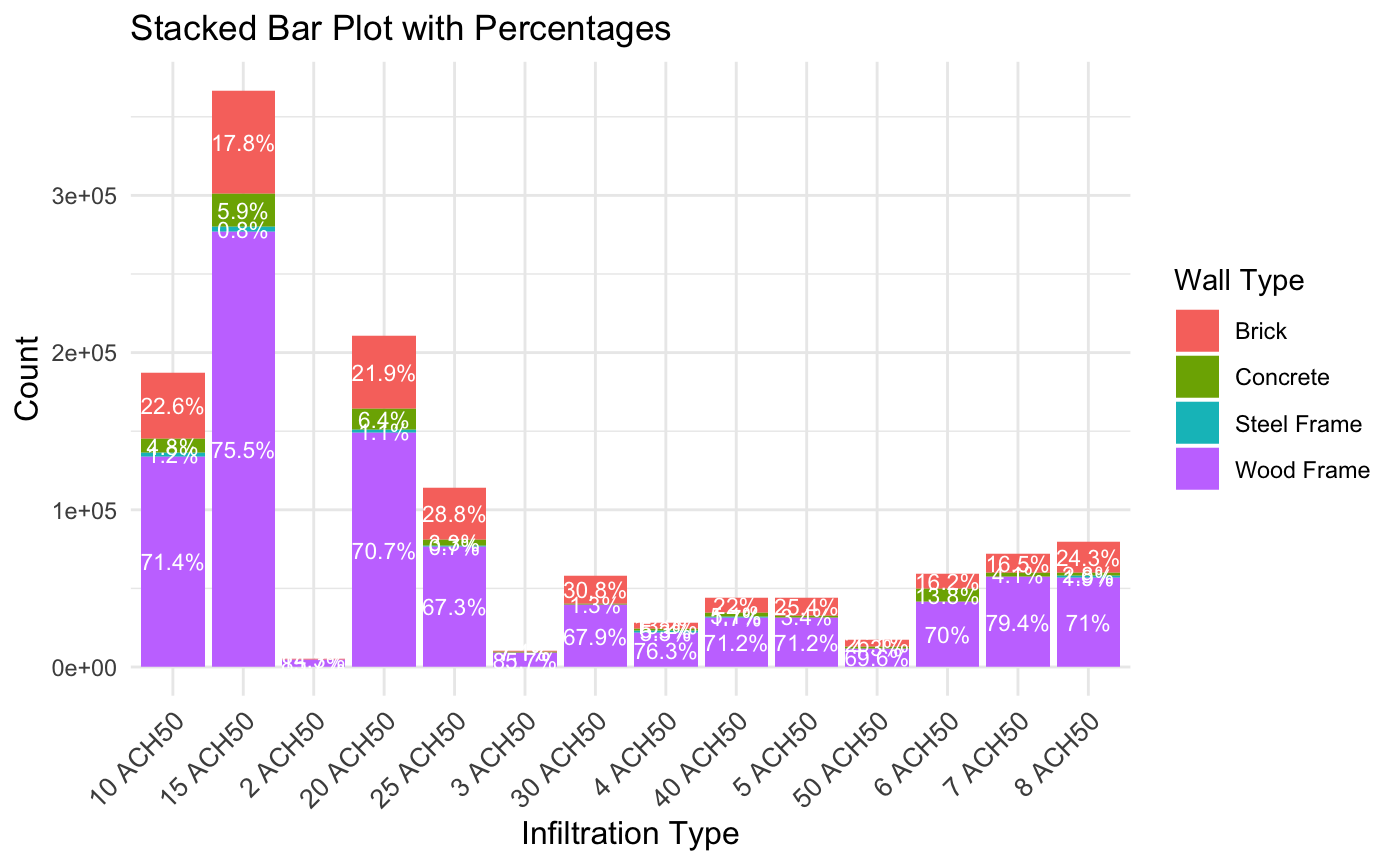
Percent Increase in Energy Consumption -

******

*The results of the graph above indicate that Energy Consumption and the Time of the Day are positively correlated. Consumption rises throughout the day after 10am, and begins to drop shortly after midnight(12am).*

******

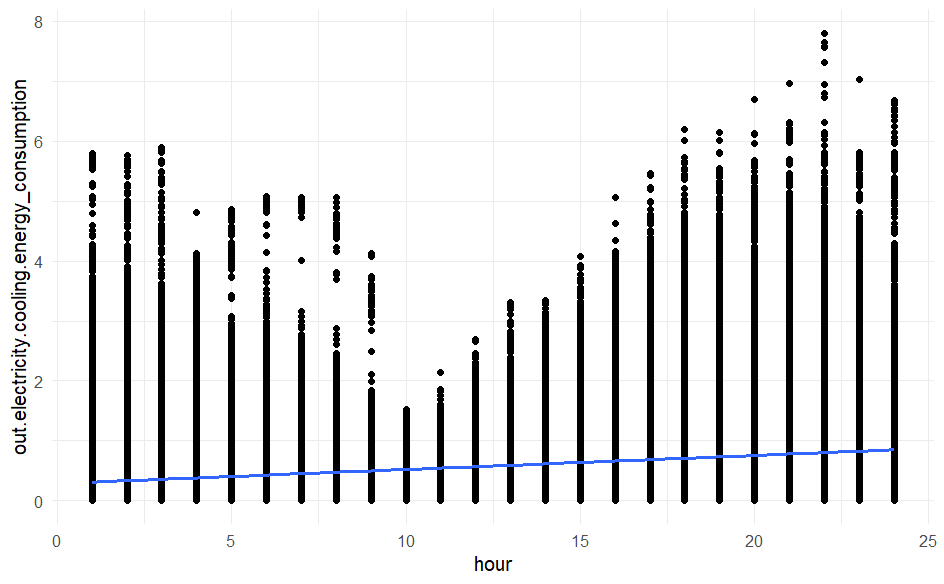
*This graph shows the relationship between Relative Usage level and Wall type. For the Middle data, relative usage across the 4 wall types does not vary much and wooden frames are the most common.*

******

*The plot indicates that across all infiltration types, wooden frames dominate. While brick and Concrete frames are more common in higher infiltration types.*

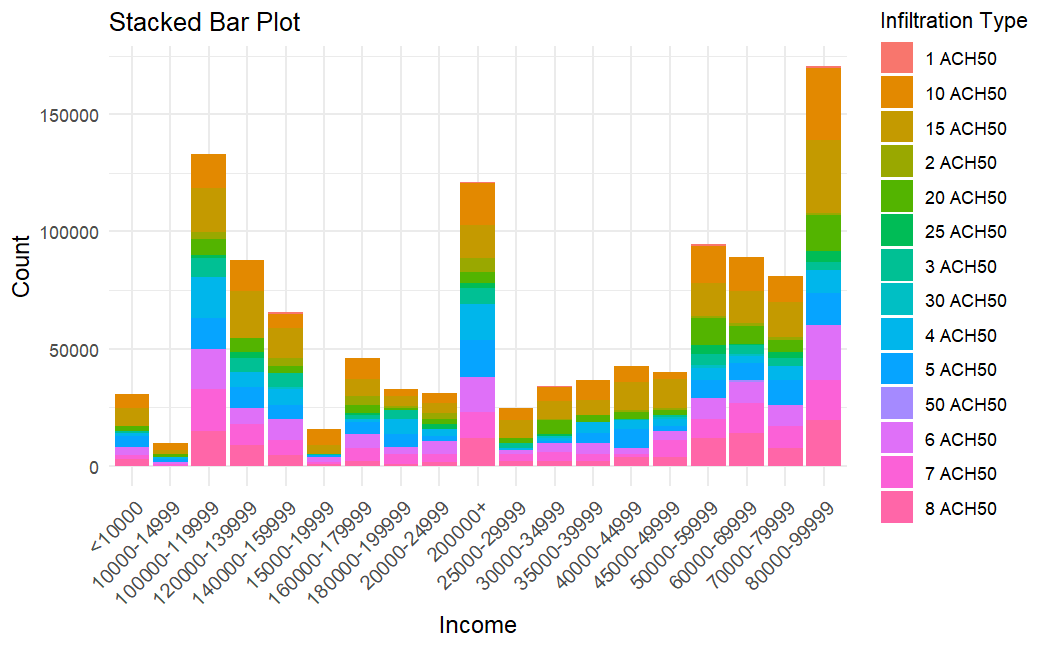
***New:***

Percent Increase in Energy Consumption - 1.958993% (July Data)



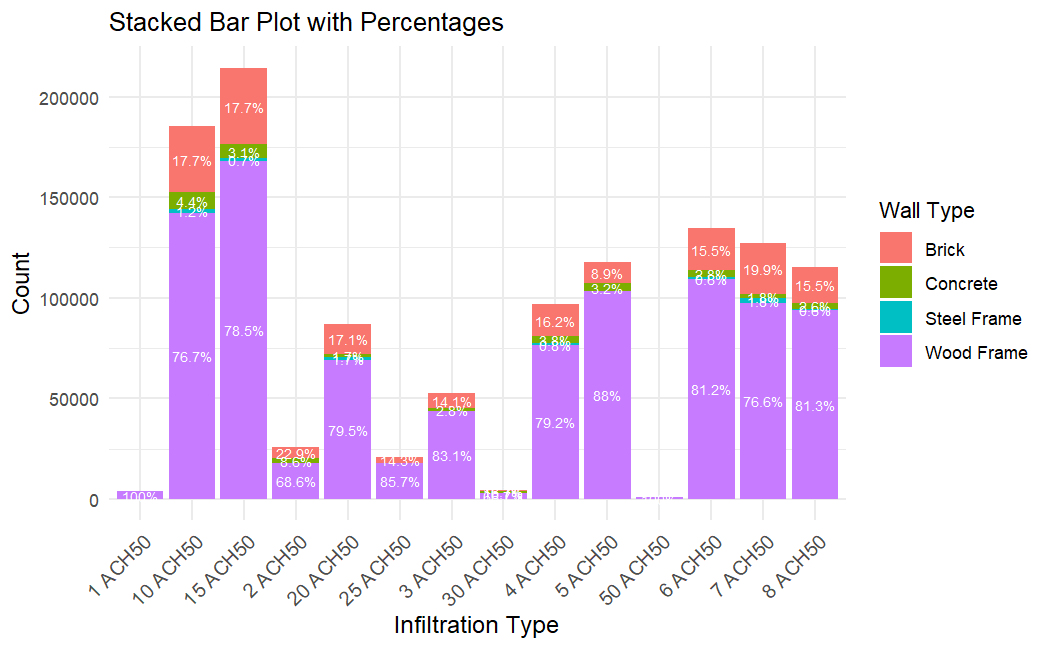
*Time of Day vs. Energy Consumption*

There is a positive correlation between the time of day and energy consumption, with supposed peaks in the beginning and the end of the day.



*Income vs. Infiltration Type*

Income seems to have an impact on the infiltration level of a home, and other features such as hvac cooling efficiency.



*Infiltration Type vs. Wall Type*

Wood Frame percentage seems to trend upwards as infiltration level worsen. However this could just be due to Wooden Frame being the most common wall type of the four.

**Analysis & Solution**

Our model's feature importance analysis identified hour of usage, HVAC cooling efficiency, usage level, income level, and square footage as the most significant predictors of energy consumption. While some features, such as the square footage of pre-existing houses, are immutable, practical interventions can be implemented to improve energy efficiency. The idea of constructing new houses with advanced, cost-effective materials is not a feasible solution. Instead, it would be advisable to emphasize the importance of maintaining an efficient HVAC system and encouraging its use only when necessary. Specifically, the creation of incentive programs that provide discounts and rebates, can motivate homeowners to upgrade to energy-efficient HVAC systems. However, it is understandable that with this plan may come unwanted costs to the and financial burdens to the company. Although subsidizing these costs may strain resources in the short term, this approach is more cost-effective than the potential expense of building or repairing power grids. Additionally, implementing surge pricing during peak energy usage

hours could further reduce demand. By having higher rates during periods of high consumption, consumers are incentivized to adjust their energy usage, and spread out

their energy usage among off-peak hours. This simple change can lead to a flattening of the demand curve, reducing strain on the power grid during critical times. This may cause potential problems, as lower income homes may find it difficult to cope with these fluid prices, but the addition of the incentives program should also alleviate much of the potential burden on these households. In theory if they are given discounts to upgrade their HVAC systems, they will be able to spend energy more efficiently, thus being able to do the same as before with less of the cost, all while receiving the upgrade at a cheaper price. With these strategic changes and the findings from our model, it is expected that peak energy demand would significantly decrease, ensuring more sustainable energy consumption within the state.

**References**

* *Mean absolute error*. Mean Absolute Error - an overview | ScienceDirect Topics. (n.d.). https://www.sciencedirect.com/topics/engineering/mean-absolute-error
* Frost, J. (2023, May 28). *Root mean square error (RMSE)*. Statistics By Jim. https://statisticsbyjim.com/regression/root-mean-square-error-rmse/
* *What are R2 and RMSE?*. Low-Cost Air Quality Monitoring & Measurement. (n.d.). <https://click.clarity.io/knowledge/r2-rmse>
* *What is root mean square error? calculation & importance*. Deepchecks. (2024, May 27). <https://www.deepchecks.com/glossary/root-mean-square-error/>
* Satellite Beach, FL. (n.d.). https://satellitebeach.gov/residents/sustainable\_satellite/living\_\_\_working\_sustainably/electricity\_at\_home.php