Predictive Analysis of Energy Usage for eSC

Introduction:

The project focuses on analyzing energy usage patterns for eSC, with a primary goal of understanding the drivers of energy consumption and promoting energy-saving behaviors among customers. This analysis is crucial for eSC(Energy Smart Communities) to address concerns about global warming, prevent blackouts caused by excessive demand on the electrical grid, and align with environmental sustainability objectives.
To extract insights from the data, use of descriptive statistics, data visualization, and machine learning modeling technique is done.

Assumptions & Considerations:

- The dataset used for analysis is assumed to be representative of the population of residential properties in South Carolina and North Carolina
- Considered only July data as focus is on high energy consumption in hot summers especially for cooling purposes
- Considered total energy consumption required only for cooling purposes as a part of modelling.
- Targeted buildings only in hot-humid climate zones

Business Questions Addressed:

The primary goals of our project include:

- **Predicting energy usage for eSC based on weather data, building characteristics, and historical energy consumption:**

  The project aims to predict energy usage for eSC based on weather data, building characteristics, and historical energy consumption. This aligns with eSC's goal of understanding the key drivers of energy usage.

- **Identifying key factors influencing energy usage and their relative importance:**

  By identifying key factors influencing energy usage and providing actionable insights for energy management and resource planning, the project contributes to eSC's objective of encouraging customers to save energy.

- **Providing actionable insights for energy management and resource planning in eSC:**
  eSC's concern about potential blackouts due to excessive demand on the electrical grid is addressed by the project's focus on reducing energy usage, especially during periods of high demand such as 'extra hot'(hot-humid) summers. The project's emphasis on reducing energy usage aligns with eSC's environmental sustainability objectives by promoting energy-saving behaviors.

These business questions align with eSC's goal of understanding energy usage patterns, identifying opportunities for energy conservation, and promoting sustainable practices to address concerns about global warming and maintain grid reliability.

Data Acquisition, Cleansing, Transformation, Munging:

The dataset includes static house data, energy usage data, and weather data. It was appropriately transformed, cleaned, and munged to prepare it for analysis.

1. **Data Acquisition:**

   House Data Retrieval:
   Acquired comprehensive house information from Parquet files stored in an AWS S3 bucket using **read_parquet()** function to extract relevant details such as house ID, county ID, and other attributes essential for subsequent analysis.
   url: https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/static_house_info.parquet

   Energy Data Retrieval :
   Fetched energy data for each house from an AWS S3 bucket using the **arrow** package's **read_parquet()** function.
   url: https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/2023-houseData/102063.parquet

   Weather Data Retrieval :
   Weather data for each county was obtained from CSV files stored in an AWS S3 bucket using the **read_csv()** function from the **readr** package.
   url: https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/weather/2023-weather-data/G4500010.csv

   Data Dictionary/Metadata Retrieval :
   Metadata information for all datasets is retrieved from csv file stored in AWS S3 bucket using read_csv() function from readr package.
   url: https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/data_dictionary.csv

2. **Cleansing and Transformation:**

   - Removed rows with missing values in any column of the energy and weather data frames using the **complete.cases()** function.
   - Filtered out houses based on specific conditions, such as being in the "Hot-Humid" climate zone to get count 1639 out of 5710 in total.
   - Merged the house, energy, and weather data frames using the **merge()** function, combining them into a single data frame for analysis.
   - Processed and transformed the data to calculate total energy consumption for cooling purposes.

3. **Data Munging:**

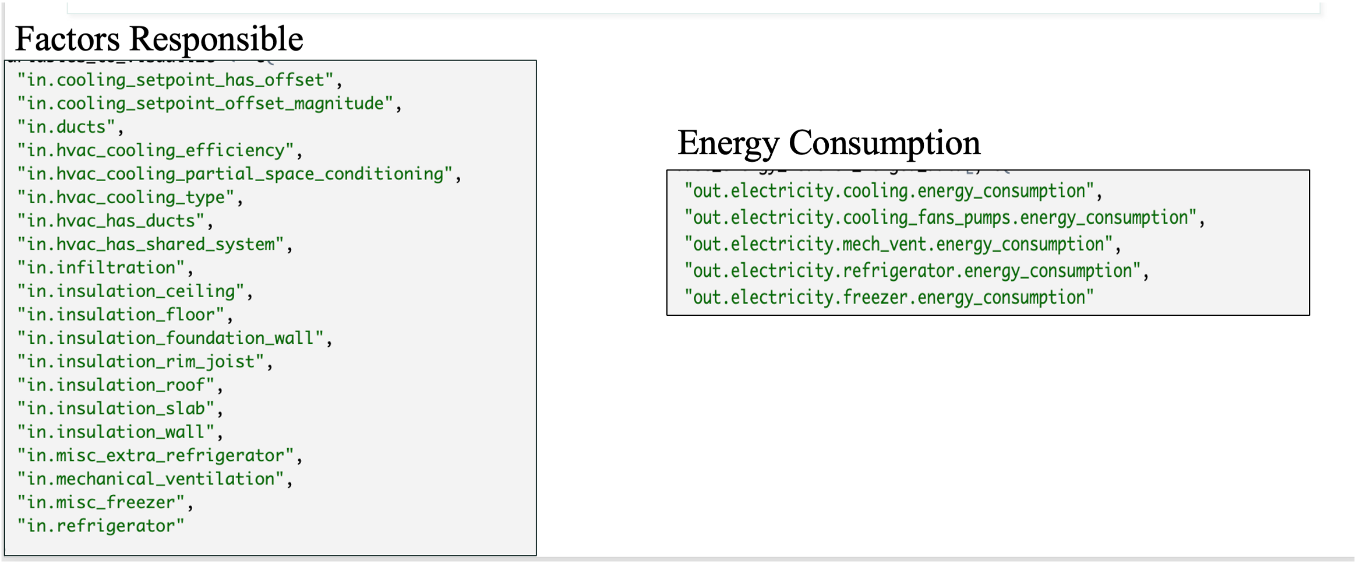
   - Leveraged functions like **parLapply()** and **mclapply()** to execute tasks concurrently across multiple cores, enhancing computational efficiency and reducing processing time. This approach ensured faster data transformation and preparation for subsequent analysis and modeling stages.
   - Handled exceptions while processing weather data to ensure robustness.

- Conducted exploratory data analysis (EDA) to identify patterns and relationships in the data using visualization techniques like histograms and bar plots.
- Prepared the data for modeling by selecting relevant columns, creating training, and testing datasets, and building a linear regression model to predict energy usage.

Descriptive Statistics & Visualization:

Descriptive statistics were utilized to provide context and a basic understanding of the data. Visualizations, including histograms, scatter plots, and bar charts, were utilized to illustrate patterns and relationships in the data. Factors responsible for total energy consumption for cooling purposes were identified.

Below are the factors responsible and electricity energy consumption for cooling purposes:



Factors Responsible

```
"in.cooling_setpoint_has_offset",
"in.cooling_setpoint_offset_magnitude",
"in.ducts",
"in.hvac_cooling_efficiency",
"in.hvac_cooling_partial_space_conditioning",
"in.hvac_cooling_type",
"in.hvac_has_ducts",
"in.hvac_has_shared_system",
"in.infiltration",
"in.insulation_ceiling",
"in.insulation_floor",
"in.insulation_foundation_wall",
"in.insulation_rim_joist",
"in.insulation_roof",
"in.insulation_slab",
"in.insulation_wall",
"in.misc_extra_refrigerator",
"in.mechanical_ventilation",
"in.misc_freezer",
"in.refrigerator"
```

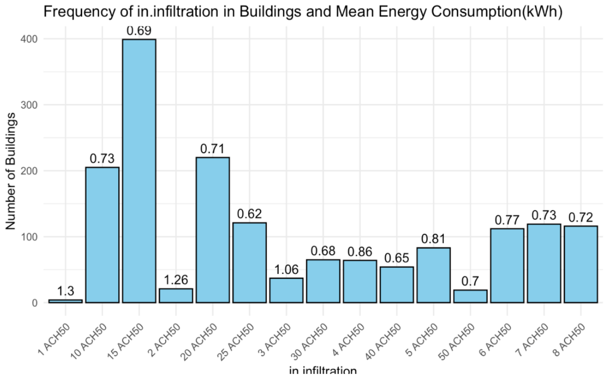
Energy Consumption

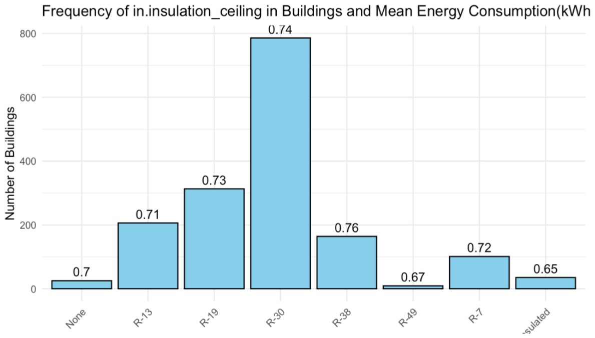
```
"out.electricity.cooling.energy_consumption",
"out.electricity.cooling_fans_pumps.energy_consumption",
"out.electricity.mech_vent.energy_consumption",
"out.electricity.refrigerator.energy_consumption",
"out.electricity.freezer.energy_consumption"
```

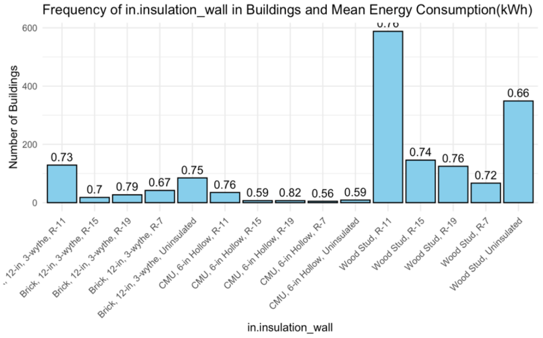
Total energy consumption was calculated adding these energy consumptions and taken it as dependent variable in modelling.

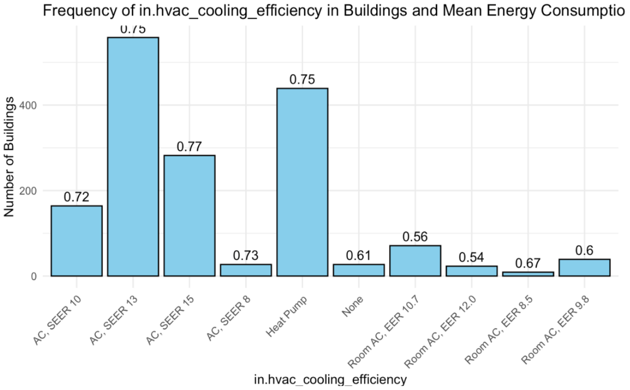
Identification of weather conditions as predictors/independent variables for this energy consumption was done. This is covered in detail in section- Use of Modeling Techniques & Visualizations.

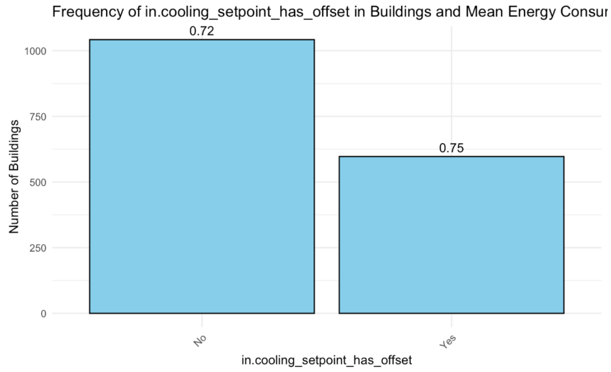
Below is the visual showing distributed count of buidings using each above factor and hourly average energy consumption by each buidling category. Using this visual, one can see where extra energy is getting utilized and factors repsonsible for that. Based on this, suggestions for alternatives/solutions to consume less energy are given. Some of these insights in section- Interpretation of Results/Actionable Insights are below:

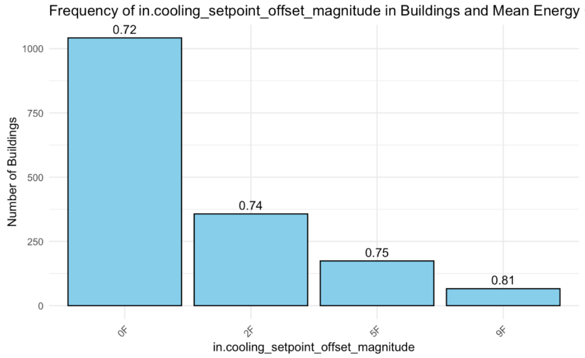

Frequency of in.infiltration in Buildings and Mean Energy Consumption(kWh)

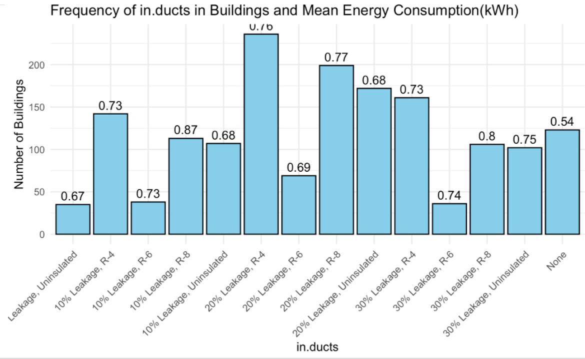

Frequency of in.insulation_ceiling in Buildings and Mean Energy Consumption(kWh

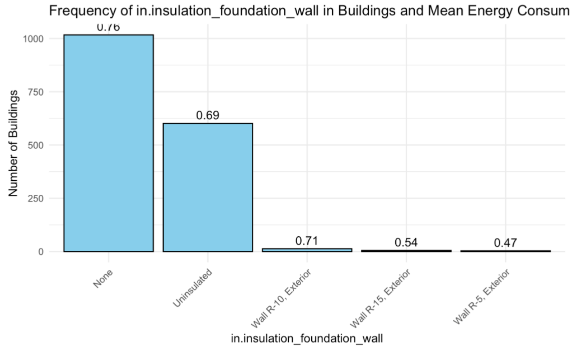

Frequency of in.insulation_wall in Buildings and Mean Energy Consumption(kWh)

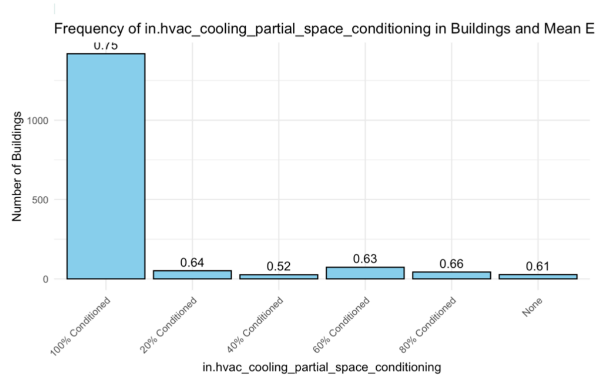

Frequency of in.hvac_cooling_efficiency in Buildings and Mean Energy Consumptio

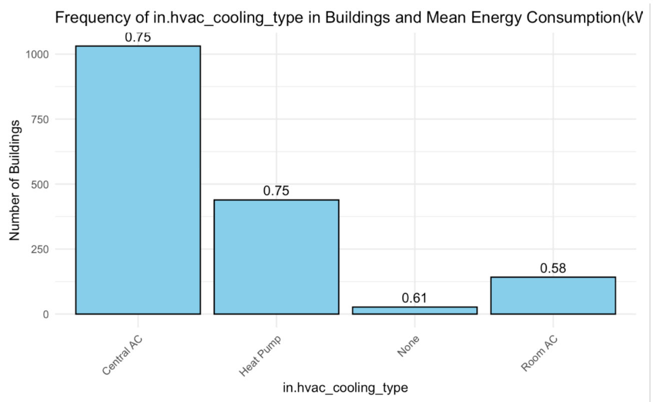

Frequency of in.cooling_setpoint_has_offset in Buildings and Mean Energy Consur

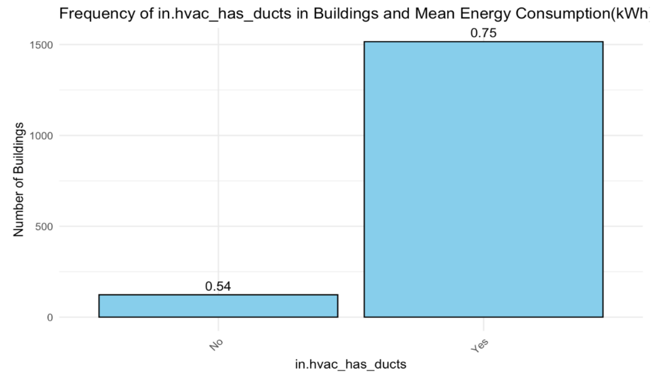

Frequency of in.cooling_setpoint_offset_magnitude in Buildings and Mean Energy

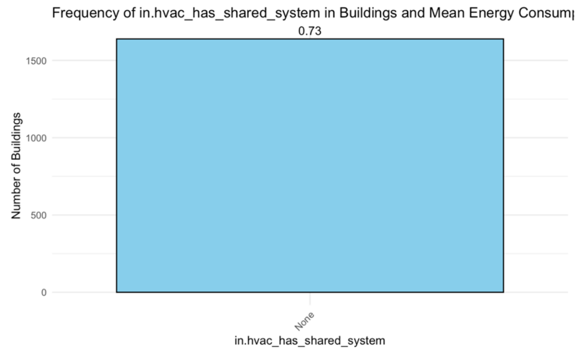

Frequency of in.ducts in Buildings and Mean Energy Consumption(kWh)

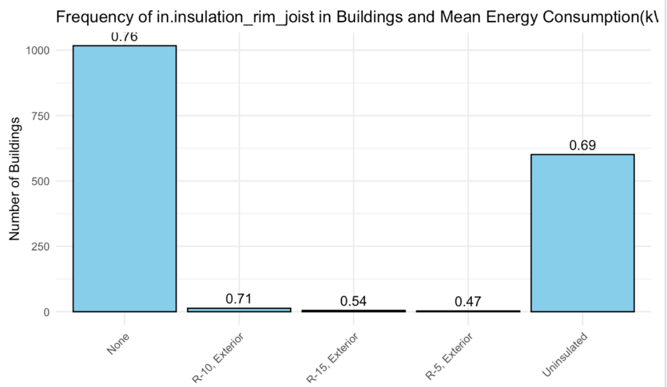

Frequency of in.insulation_foundation_wall in Buildings and Mean Energy Consum

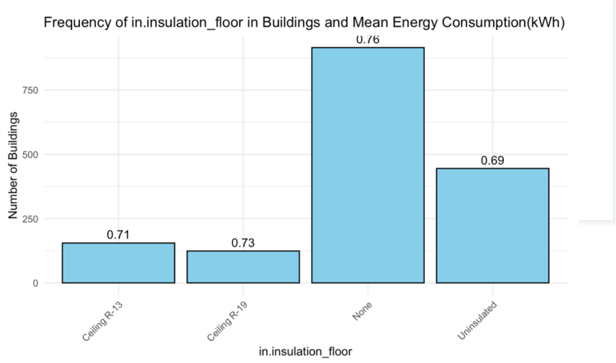

Frequency of in.hvac_cooling_partial_space_conditioning in Buildings and Mean E
Number of Buildings
1000
500
0
0.75
0.64
0.52
0.63
0.66
0.61
100% Conditioned
20% Conditioned
40% Conditioned
60% Conditioned
80% Conditioned
None
in.hvac_cooling_partial_space_conditioning

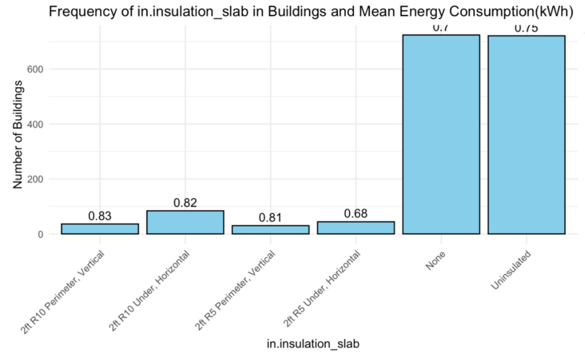

Frequency of in.hvac_cooling_type in Buildings and Mean Energy Consumption(kV
Number of Buildings
1000
750
500
250
0
0.75
0.75
0.61
0.58
Central AC
Heat Pump
None
Room AC
in.hvac_cooling_type

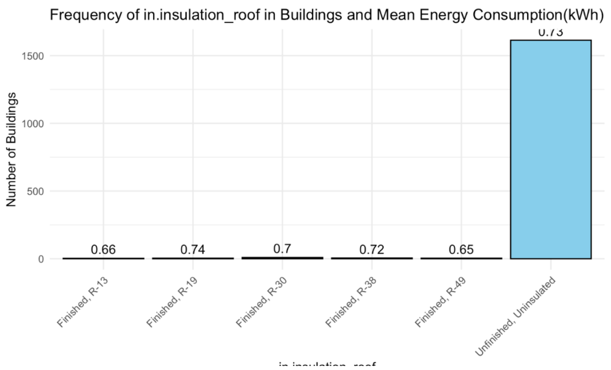

Frequency of in.hvac_has_ducts in Buildings and Mean Energy Consumption(kWh
Number of Buildings
1500
1000
500
0
0.54
0.75
No
Yes
in.hvac_has_ducts

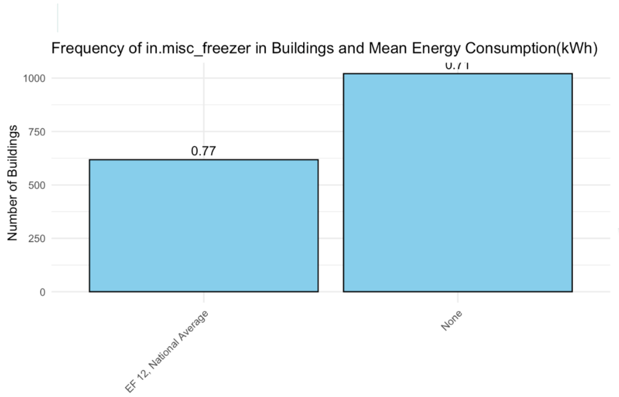

Frequency of in.hvac_has_shared_system in Buildings and Mean Energy Consum
Number of Buildings
1500
1000
500
0
0.73
None
in.hvac_has_shared_system

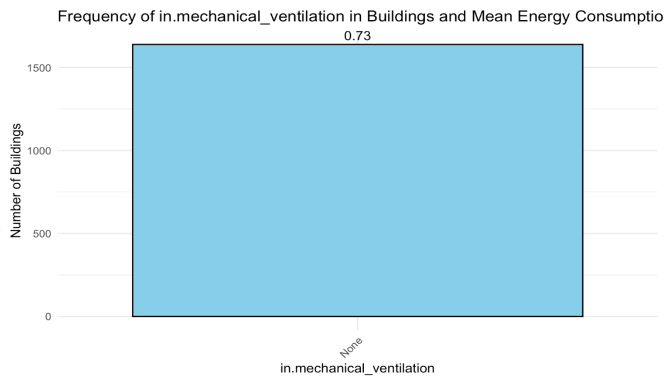

Frequency of in.insulation_rim_joist in Buildings and Mean Energy Consumption(k\
Number of Buildings
1000
750
500
250
0
0.76
0.71
0.54
0.47
0.69
None
R-10, Exterior
R-15, Exterior
R-5, Exterior
Uninsulated

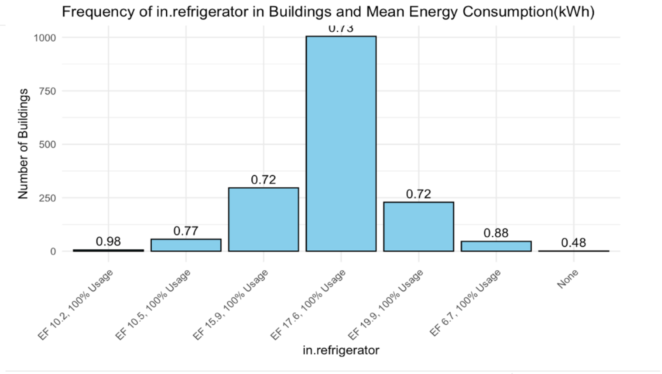

Frequency of in.insulation_floor in Buildings and Mean Energy Consumption(kWh)
Number of Buildings
750
500
250
0
0.71
0.73
0.76
0.69
Ceiling R-13
Ceiling R-19
None
Uninsulated
in.insulation_floor

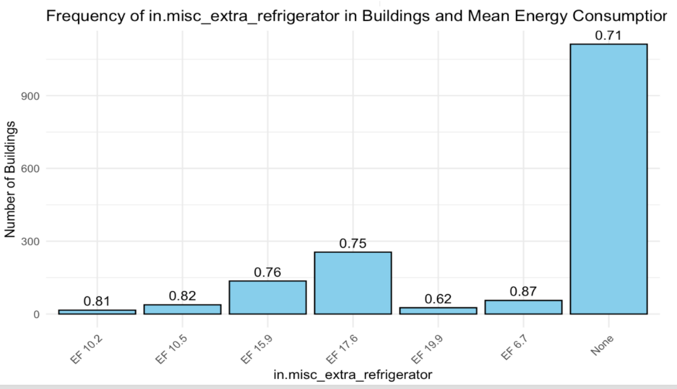

Frequency of in.insulation_slab in Buildings and Mean Energy Consumption(kWh)
Number of Buildings
600
400
200
0
0.83
0.82
0.81
0.68
0.7
0.75
2ft R10 Perimeter, Vertical
2ft R10 Under, Horizontal
2ft R5 Perimeter, Vertical
2ft R5 Under, Horizontal
None
Uninsulated
in.insulation_slab


Frequency of in.insulation_roof in Buildings and Mean Energy Consumption(kWh)
Number of Buildings
1500
1000
500
0
0.66
0.74
0.7
0.72
0.65
0.73
Finished, R-13
Finished, R-19
Finished, R-30
Finished, R-38
Finished, R-49
Unfinished, Uninsulated
in.insulation_roof









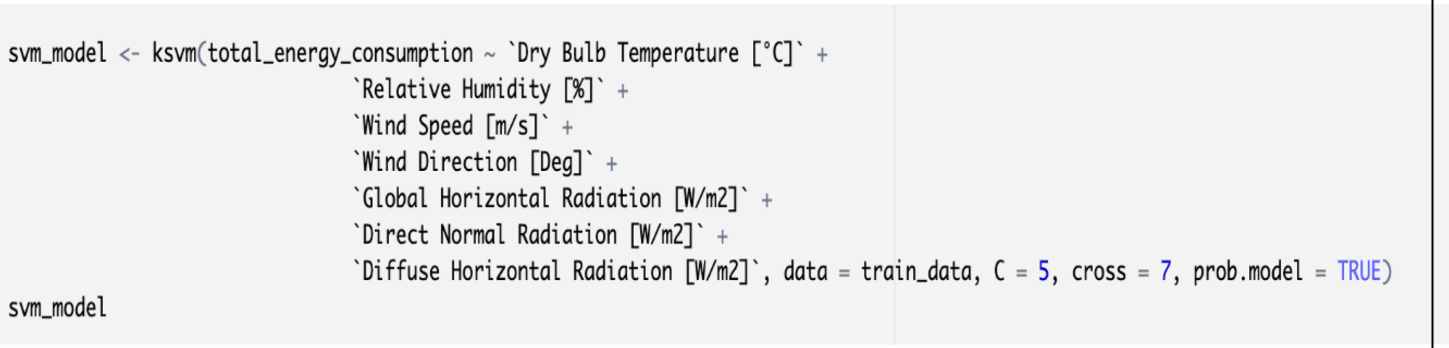
Use Of Modelling Techniques & Visualizations:

Different modeling techniques were explored, including supervised-k-fold cross-validation with SVM and linear regression.

Initially exploration of supervised-k-fold cross-validation was done with SVM trying different cross number of cross folds keeping 60% data for training the model due to its potential for capturing complex relationships.

However, extensive computational requirements led to long training times. Despite attempts to optimize parameters and fold settings, SVM remained impractical for our project's time constraints. As a result, I opted to pursue alternative modeling approach discarding SVM to ensure efficient model training and evaluation.

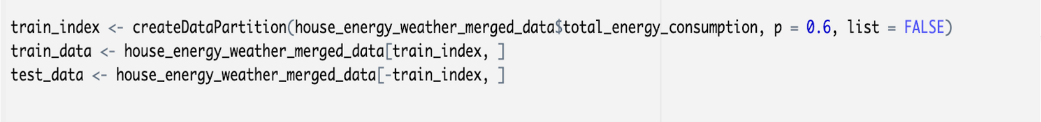
supervised-k-fold cross-validation with SVM:



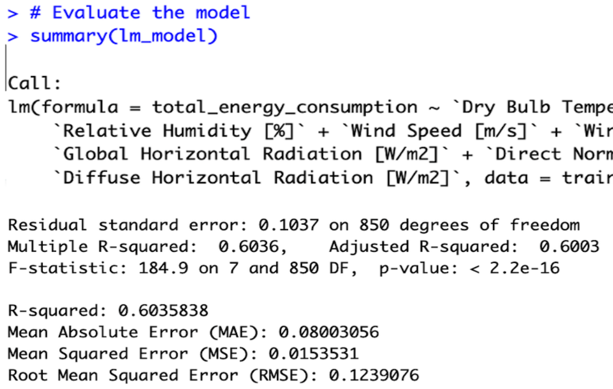
```
svm_model <- ksvm(total_energy_consumption ~ `Dry Bulb Temperature [°C]` +
                            `Relative Humidity [%]` +
                            `Wind Speed [m/s]` +
                            `Wind Direction [Deg]` +
                            `Global Horizontal Radiation [W/m2]` +
                            `Direct Normal Radiation [W/m2]` +
                            `Diffuse Horizontal Radiation [W/m2]`, data = train_data, C = 5, cross = 7, prob.model = TRUE)
svm_model
```

Model evaluation metrics such as R-squared, Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error were used to evaluate model based on which I decided to go with linear regression model with partition to train model where 60% of the data is trained reserving 40% of the data for validation or testing.

Linear Regression:



```
train_index <- createDataPartition(house_energy_weather_merged_data$total_energy_consumption, p = 0.6, list = FALSE)
train_data <- house_energy_weather_merged_data[train_index, ]
test_data <- house_energy_weather_merged_data[-train_index, ]
```



```
> # Evaluate the model
> summary(lm_model)

Call:
lm(formula = total_energy_consumption ~ `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]`, data = train_data)

Residual standard error: 0.1037 on 850 degrees of freedom
Multiple R-squared:  0.6036,    Adjusted R-squared:  0.6003
F-statistic: 184.9 on 7 and 850 DF,  p-value: < 2.2e-16

R-squared: 0.6035838
Mean Absolute Error (MAE): 0.08003056
Mean Squared Error (MSE): 0.0153531
Root Mean Squared Error (RMSE): 0.1239076
```

Approximately 60% of the variance in the dependent variable is explained by the independent variables

Validation details are described in detail in further Validation section of report.

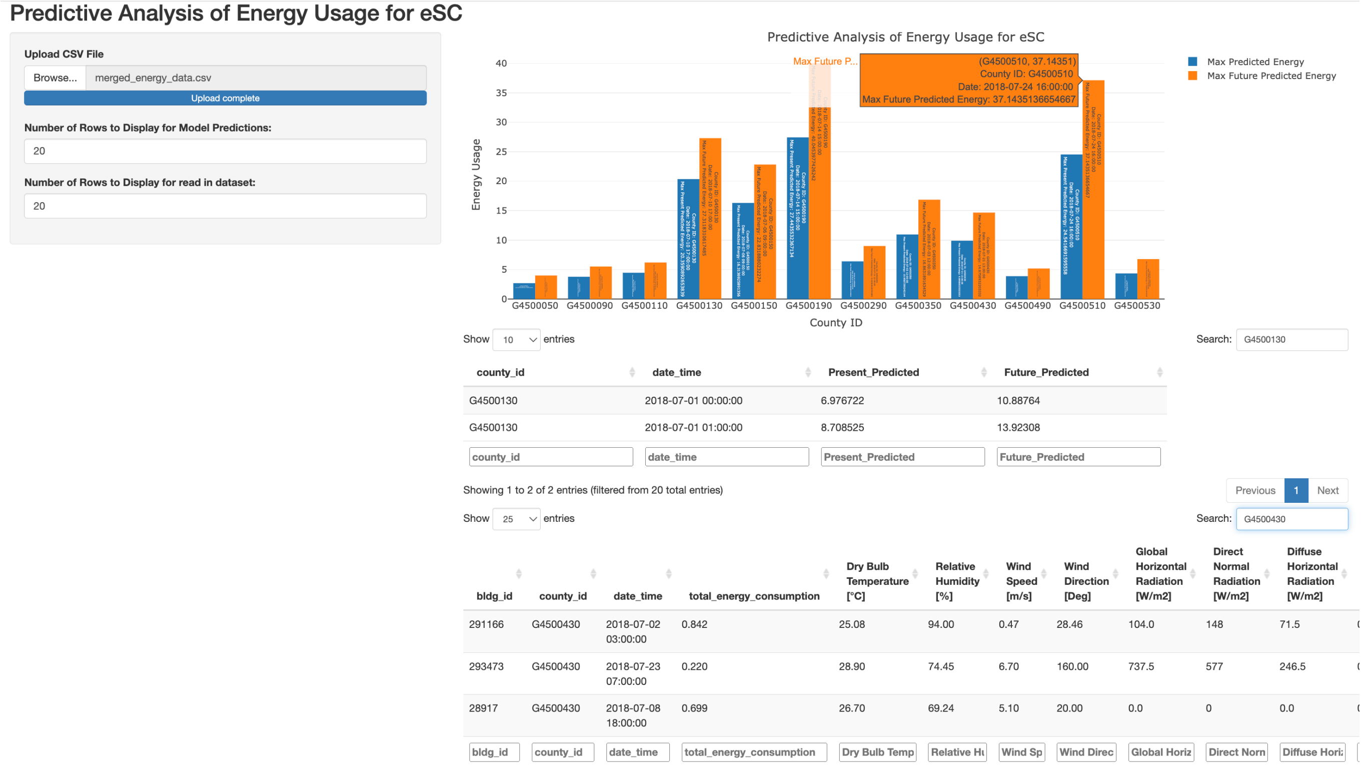
Visualizations of model predictions (present and future total energy usage) per county were generated to evaluate difference in energy consumptions (present and future) for same set of users when 5 degrees of atmospheric temperature is increased. It would be helpful to identify areas of improvement to reduce energy consumption.Below is the visual of our predictions using model on shiny App, where there are few rows of output of predictions along with few rows of read in merged data used.

Shiny App:

The Shiny app in question is an interactive and visually structured platform designed for the analysis of weather data. Utilizing the R programming language and various libraries such as tidyverse, ggplot2, and shiny dashboard, the application offers a polished user interface.

In essence, this Shiny app provides a sophisticated yet accessible tool for users to delve into and interpret weather data, serving as a valuable resource for meteorological analysis.

Here is the deployed application: https://mugdha-karodkar.shinyapps.io/IST687_Final_Project/



Predictive Analysis of Energy Usage for eSC

Upload CSV File

Browse... merged_energy_data.csv

Upload complete

Number of Rows to Display for Model Predictions: 20

Number of Rows to Display for read in dataset: 20

Show 10 entries

Search: G4500130

county_id	date_time	Present_Predicted	Future_Predicted
G4500130	2018-07-01 00:00:00	6.976722	10.88764
G4500130	2018-07-01 01:00:00	8.708525	13.92308
county_id	date_time	Present_Predicted	Future_Predicted

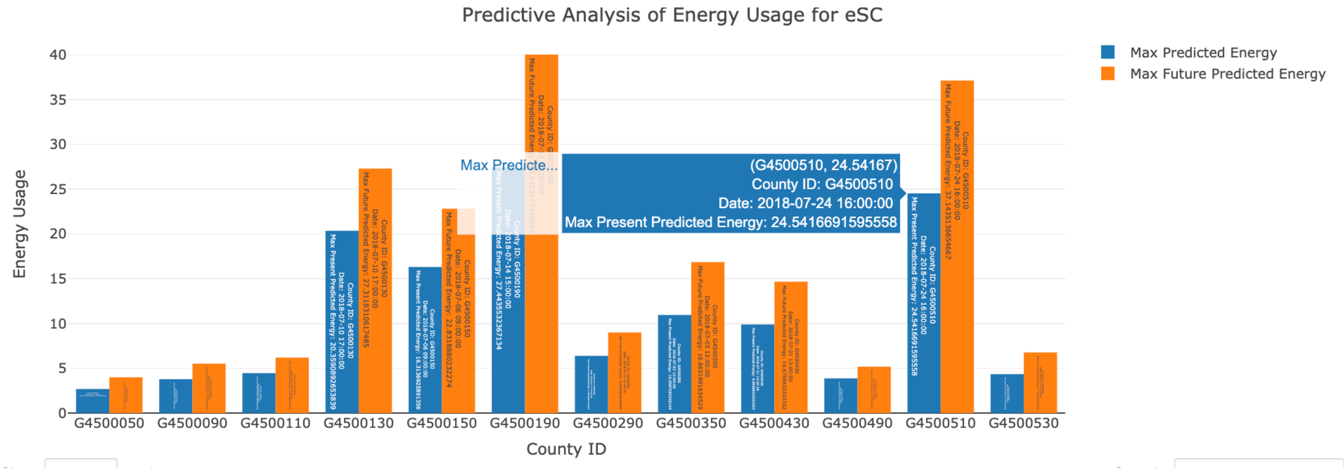
Showing 1 to 2 of 2 entries (filtered from 20 total entries)

Previous 1 Next

Show 25 entries

Search: G4500430

bldg_id	county_id	date_time	total_energy_consumption	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]
291166	G4500430	2018-07-02 03:00:00	0.842	25.08	94.00	0.47	28.46	104.0	148	71.5
293473	G4500430	2018-07-23 07:00:00	0.220	28.90	74.45	6.70	160.00	737.5	577	246.5
28917	G4500430	2018-07-08 18:00:00	0.699	26.70	69.24	5.10	20.00	0.0	0	0.0
bldg_id	county_id	date_time	total_energy_consumption	Dry Bulb Temp	Relative Hu	Wind Sp	Wind Direc	Global Horiz	Direct Norn	Diffuse Hori:



Interpretation Of Results/Actionable Insights:

The results of our analysis provide valuable insights into the factors driving energy usage in eSC. Key findings include the significant impact of weather conditions, building characteristics, and historical energy consumption on future energy usage.

Actionable insights derived from the results can inform energy management strategies, such as optimizing HVAC systems, improving insulation, and implementing energy-efficient appliances.

Interpretation of Results:

The results of the modeling techniques were interpreted to derive actionable insights. Strategies for peak energy demand reduction and energy usage optimization were proposed based on the analysis. Recommendations include upgrading insulation, minimizing air leakage, improving HVAC efficiency, and optimizing wall insulation. Also, to balance over energy consumption for cooling purposes, here are some insights to save energy in factors like washers, dryers, ceiling fans etc.

Below are the insights:

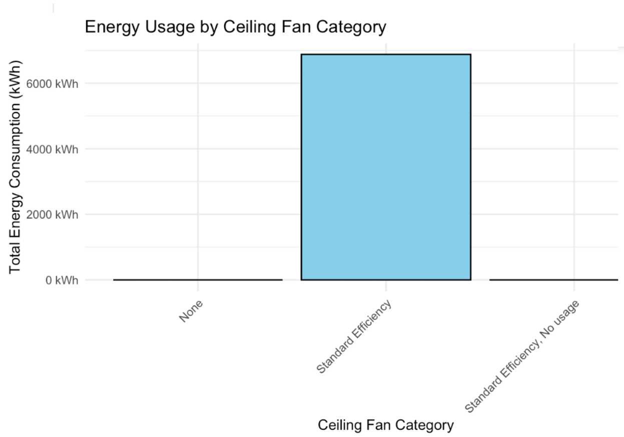
1. **Ceiling Insulation:** Homes with uninsulated ceilings require insulation to improve energy efficiency. Consider upgrading ceiling insulation to at least R-30 or higher, especially in top floors, to enhance thermal resistance and minimize heat transfer.

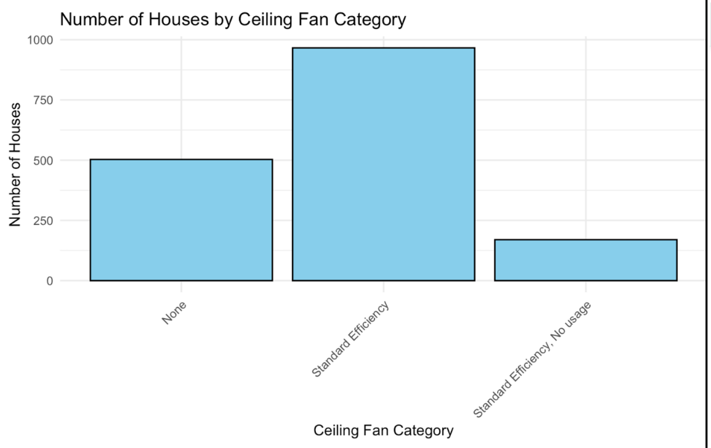
2. **Infiltration (Air Leakage Rates):** Minimize air leakage by sealing gaps and cracks in the building envelope. Aim for lower air change rates per hour (ACH50), such as "1 ACH50" or "2 ACH50," to reduce the infiltration of hot outdoor air into the conditioned space. Proper sealing helps maintain indoor comfort levels and prevents energy loss.

3. H**VAC Cooling Efficiency:** Enhance cooling efficiency by upgrading to a more efficient cooling system with a higher SEER rating (Seasonal Energy Efficiency Ratio). Options like "AC, SEER 13" or higher offer better energy performance. Additionally, consider heat pumps, which are highly efficient and can provide both cooling and heating functions.

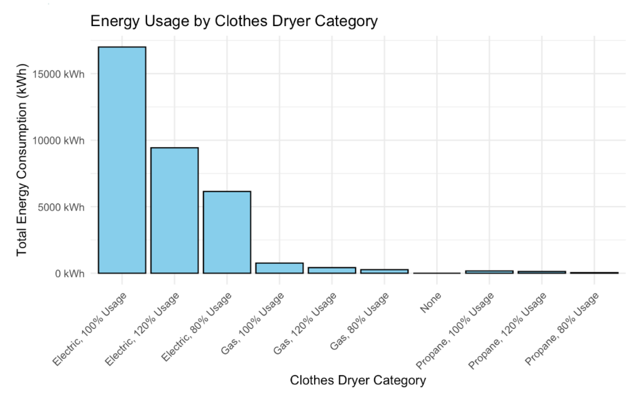
4. **Wall Insulation:** Optimize energy efficiency for the summer months by selecting appropriate wall insulation based on construction type and R-value. Choose options like "Wood Stud, R-15" or "CMU, 6-in Hollow, R-19" to minimize heat transfer through walls. Adequate insulation reduces cooling load, enhances comfort, and contributes to overall energy savings.

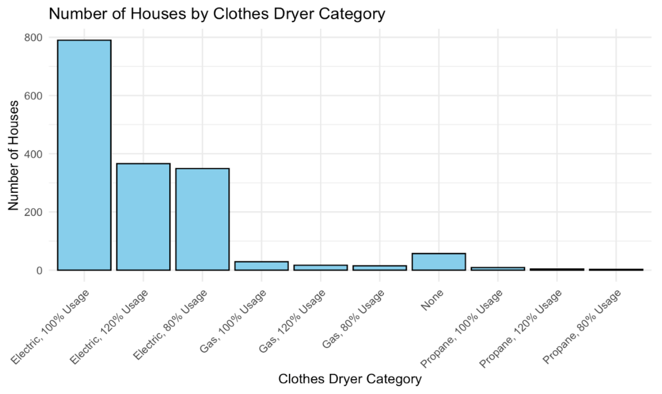
5. Encourage houses without **standard efficient ceiling fans** to install them for improved air circulation and reduced AC reliance. For houses with ceiling fans already in place, promote their usage to decrease AC energy consumption in July.



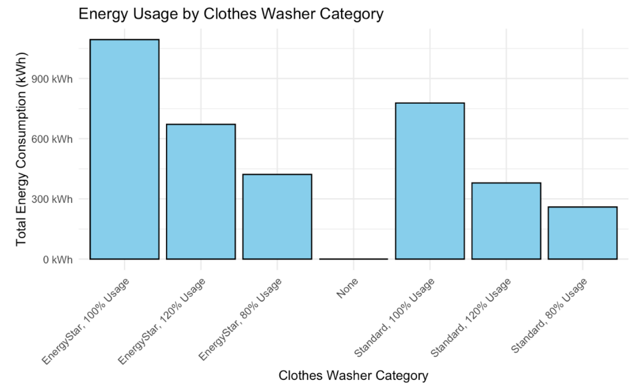


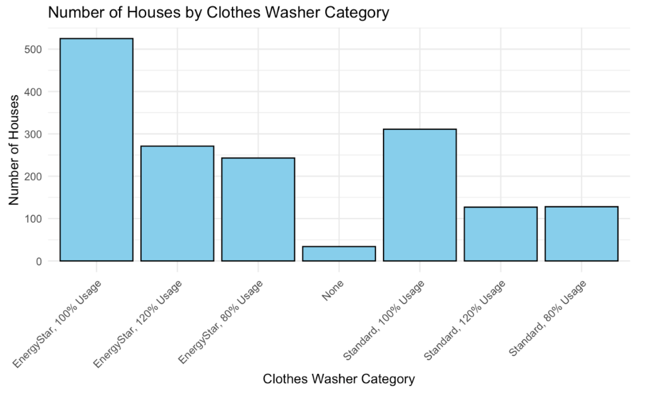
6. Rrecommend **heat pump dryers** as they are gaining popularity for their energy efficiency. They utilize a heat pump system to extract moisture from the air inside the dryer and recycle heat, making them an eco-friendly choice.





7. Recommend **Energy Star-rated Front load washers.**





Energy Star-rated washers are generally more energy-efficient compared to standard washers.

Here's why: Energy Star-rated washers must meet specific energy efficiency criteria set by the **Environmental Protection Agency (EPA)** in the United States. These criteria typically include requirements for water usage, energy consumption during washing and spinning cycles, and overall efficiency. Standard washers may not meet these stringent criteria and may consume more energy and water.

Front-load washers are generally more energy-efficient than top-load washers because they use less water and require less energy to operate. Energy Star-rated front-load washers are even more efficient as they meet strict energy efficiency criteria set by the Environmental Protection Agency (EPA).

Validation:

To ensure the accuracy and reliability of our linear regression predictive model, several validation techniques were employed. These techniques aimed to assess the performance of the linear regression model in predicting energy consumption based on various weather parameters.

Referring to the selected linear regression model details in section -Use of Modeling Techniques & Visualizations:

1. **Training-Test Split:**

   - Partitioned the dataset into training and test sets using a 60-40 split. The training set was used to train the linear regression model, while the test set was kept separate to evaluate the model's performance on unseen data.

2. **Model Evaluation Metrics**:
   Upon training the model on the training data, its performance was evaluated using several metrics:
   - **Residual Standard Error (RSE):**
     The RSE measures the average deviation of the observed values from the predicted values. A lower RSE indicates better model fit.

   - **R- Squared (R^2)**:
     The R-squared value represents the proportion of variance in the dependent variable (energy consumption) explained by the independent variables (weather parameters). An R^2 close to 1 indicates a better fit of the model to the data.

   - **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, **and Root Mean Squared Error (RMSE)**:
     These metrics quantify the average magnitude of errors between predicted and observed values. Lower values indicate better predictive accuracy.
   - **F-statistic and p-value**:
     The F-statistic tests the overall significance of the model, while the associated p-value assesses the probability of obtaining such results by chance. A p-value <

0.05/ p-value < 2.2e-16 suggests that the model's coefficients are statistically significant.

3. **Interpretation of Results**:

   The summary statistics of our linear regression model indicate promising performance. The adjusted R-squared value of 0.6003 suggests that approximately 60% of the variance in energy consumption can be explained by the weather parameters included in the model. Additionally, the p-value (< 2.2e-16) associated with the F-statistic confirms the statistical significance of the model.

Furthermore, the model's performance is supported by the following metrics:

- **Residual Standard Error (RSE)**:
  The RSE, with a value of 0.1037, indicates that the average deviation of the observed energy consumption values from the predicted values is relatively low, indicating a good fit of the model to the data.

- **Mean Absolute Error (MAE)**:
  With a value of 0.08003056, the MAE represents the average magnitude of errors between predicted and observed energy consumption values. A lower MAE signifies better predictive accuracy.

- **Mean Squared Error (MSE)** and **Root Mean Squared Error (RMSE)**:
  The MSE (0.0153531) and RMSE (0.1239076) quantify the average squared and square root of errors, respectively. Both metrics provide additional insights into the model's predictive performance, with lower values indicating better accuracy.