<u>IST 687 M009: Introduction to Data Science</u>

Project on

“*Energy Consumption Forecast and Analysis*”

Abstract

E.SC, the leading energy provider in South Carolina, is dedicated to delivering reliable and sustainable electricity to its customers. With the escalation of summer temperatures due to climate change, E.SC anticipates an increase in demand for electricity, particularly during peak heatwaves. This surge poses a threat to the power grid's stability, potentially resulting in disruptive blackouts.

To address this challenge, E.SC is actively initiating a comprehensive study to understand the primary determinants of residential electricity demand. The project seeks to uncover influential factors driving energy consumption, analyze its correlation with temperature changes in July, identify significant determinants, provide actionable recommendations for energy conservation, and demonstrate the efficacy of energy reduction strategies through predictive modeling.

1. Introduction

The project's primary objectives encompass uncovering the key factors influencing energy consumption, investigating its relationship with temperature variations in July, pinpointing significant determinants, offering practical recommendations for energy conservation, and demonstrating the effectiveness of energy reduction strategies using predictive modeling.

Our initial steps involve exploratory data analysis (EDA) and data cleaning, followed by the development of predictive models. Within the dataset under analysis lies a wealth of information about individual residences, encompassing unique building IDs and specific attributes associated with each property. Furthermore, the dataset provides intricate hour-by-hour details of energy consumption patterns in these residences.

Moreover, the dataset meticulously integrates hour-by-hour weather data, systematically organized according to geographic regions. This weather information serves as a critical external factor significantly shaping energy usage patterns across diverse regions.

The object of this project is to answer the following questions :

1. Identify Factors Affecting Energy Consumption.
2. Analyse Energy Consumption Variation with Temperature Increase in July.
3. Identify Significant Factors Impacting Energy Consumption.
4. Recommend for Decreasing Energy Consumption.
5. Demonstrate Energy Consumption Reduction via a Model.

2. Exploratory Data Analysis

To understand the most significant factors that are driving the consumption of energy we start cleaning the data. We began with the first dataset “static_housing” dataset by pulling it from a parquet file.

```{r}
library (arrow)
library(tidyverse)
library(writexl)
static_housing <- read_parquet("https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/static_house_info.parquet")
str(static_housing)
write_xlsx(static_housing, "static_housing.xlsx") #writng to excel for easier access (time consuming to pull repititively)
```

The static housing dataset comprises 5710 rows and 171 columns, with around 8 numeric columns and the rest of 163 character types. Each row is unique based on its building id.

Key variables such as building ID, size of house labeled as in. sqft, and characteristics of the house infrastructure like wall-to-window ratio, cooling and heating system setting, and a lot more could contribute to energy consumption.

```
 $ in.heating_fuel                               : chr  "Natural Gas" "Natural Gas" "Natural Gas" "Natural Gas" ...
 $ in.heating_setpoint                           : chr  "70F" "65F" "70F" "68F" ...
 $ in.heating_setpoint_has_offset                : chr  "No" "Yes" "No" "Yes" ...
 $ in.heating_setpoint_offset_magnitude          : chr  "0F" "3F" "0F" "3F" ...
 $ in.heating_setpoint_offset_period             : chr  "None" "Night -4h" "None" "Night -3h" ...
 $ in.holiday_lighting                           : chr  "No Exterior Use" "No Exterior Use" "No Exterior Use" "No Exterior Use" ...
 $ in.hot_water_distribution                     : chr  "Uninsulated" "Uninsulated" "Uninsulated" "Uninsulated" ...
 $ in.hot_water_fixtures                         : chr  "100% Usage" "100% Usage" "50% Usage" "50% Usage" ...
 $ in.hvac_cooling_efficiency                    : chr  "AC, SEER 15" "AC, SEER 13" "AC, SEER 13" "None" ...
 $ in.hvac_cooling_partial_space_conditioning    : chr  "100% Conditioned" "100% Conditioned" "100% Conditioned" "None" ...
 $ in.hvac_cooling_type                          : chr  "Central AC" "Central AC" "Central AC" "None" ...
 $ in.hvac_has_ducts                             : chr  "Yes" "Yes" "Yes" "No" ...
 $ in.hvac_has_shared_system                     : chr  "None" "None" "None" "None" ...
 $ in.hvac_has_zonal_electric_heating            : chr  "No" "No" "No" "No" ...
```

For example, it consists of a variable called “in.pv_system_size” which signifies how big the solar panels and other variables describe the insulation of the house.

```
 in.pv_system_size
 Length:5710
 Class :character
 Mode  :character
```

```
 in.insulation_ceiling in.insulation_floor
 Length:5710           Length:5710
 Class :character      Class :character
 Mode  :character      Mode  :character
```

On doing some descriptive analysis on the numerical columns we saw that the number of bedrooms and the size of the house (in square feet) both had a roughly normal distribution as expected.

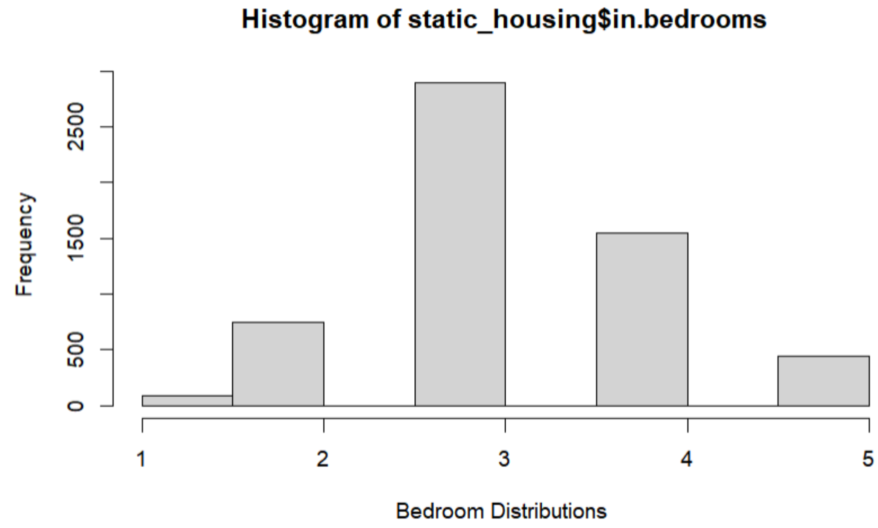

Fig. 2.1

Fig. 2.2

The other two variables were discrete values. Regional Energy Deployment type had two types 95 and 96. The number of stories a building had also had only three values from 1 up to 3 which is right skewed.

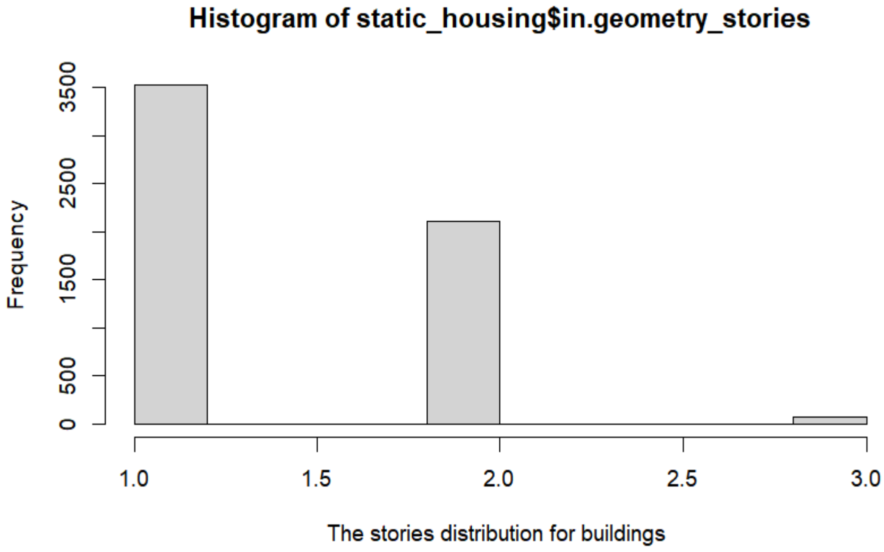

Fig. 2.3

To understand the correlation between numerical variables we created a correlation matrix. The analysis revealed a robust correlation between weather and reed, indicating a strong association between these variables. Consequently, we have opted to retain the variable at this stage of the analysis.

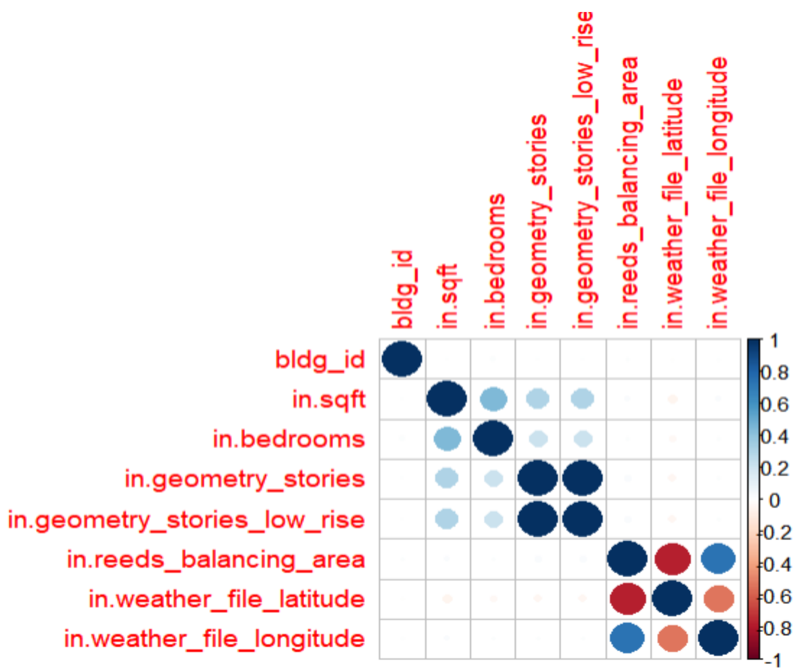

Fig. 2.4

We cleaned the data by checking for any NA columns and also got rid of all the rows that consisted of just 1 value as that would imply no variability.

This brought down the columns from 171 to 93 in number. Lastly, we checked for the percentage of blanks in columns. We saw lower than 5% of blanks and decided to leave it as it is.

	upgrade.water_heater_efficiency	upgrade.clothes_dryer	upgrade.cooking_range
Percentage of Blanks	0.910683	3.677758	1.17338
Percentage of Values	99.089317	96.322242	98.82662

We also plotted the density of buildings by county and saw an uneven distribution, which led to the conclusion that we cannot aggregate any of these factors for energy consumption by county. We see that Greenville has the highest density, followed by Colleton, Georgetown, Horry, and so on.

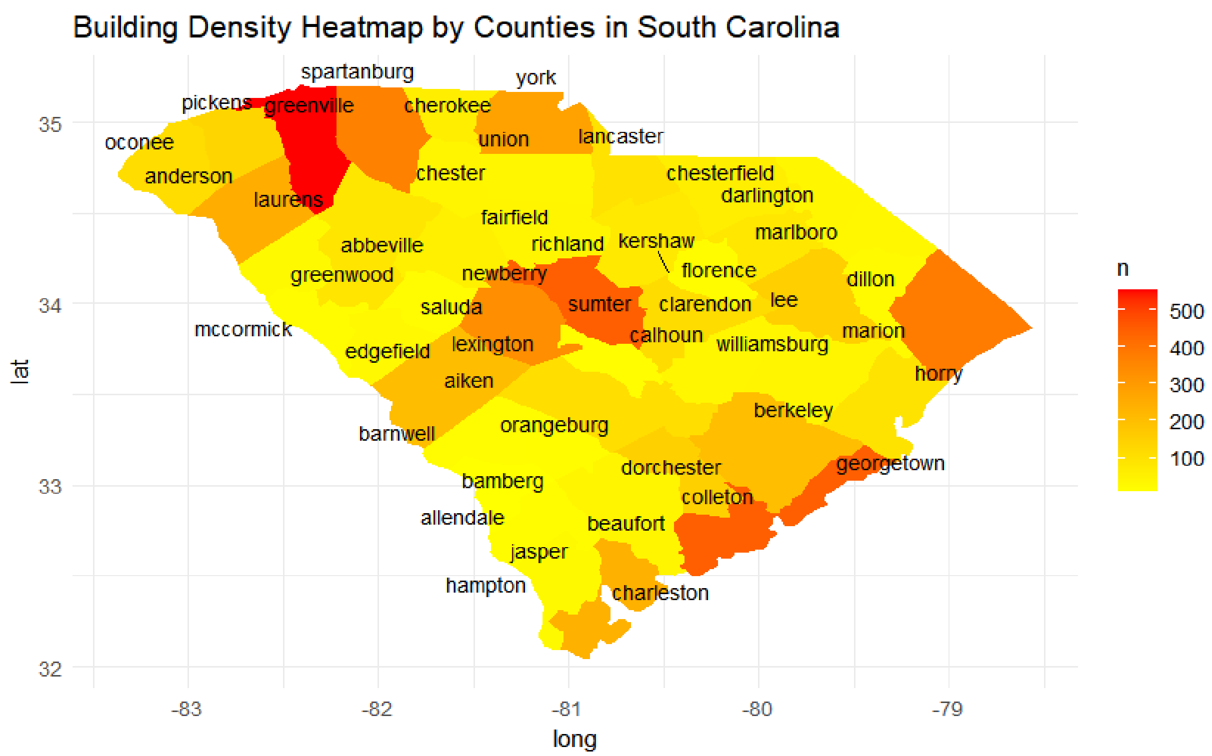

Fig. 2.5

Moving on to the energy consumption data. For each of the buildings, there was a data record for each hour for each month of 2018. We decided to pick July of 2018 for our analysis.

Our objective was to figure out an hourly increase in consumption of energy, so for each building house, we aggregated the energy by hour for all of July. This left us with around 130k rows to work with across 137 variables.

```
Rows: 137,040
Columns: 136
$ bldg_id                              <dbl> 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, …
$ in.sqft                              <dbl> 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 8…
$ in.bathroom_spot_vent_hour           <chr> "Hour23", "Hour23", "Hour23", "Hour23", "Hour23", "Hour23", "Hour23", "Hour2…
$ in.bedrooms                          <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 2…
$ in.building_america_climate_zone     <chr> "Mixed-Humid", "Mixed-Humid", "Mixed-Humid", "Mixed-Humid", "Mixed-Humid", "…
$ in.ceiling_fan                       <chr> "Standard Efficiency", "Standard Efficiency", "Standard Efficiency", "Standa…
$ in.city                              <chr> "SC, Rock Hill", "SC, Rock Hill", "SC, Rock Hill", "SC, Rock Hill", "SC, Roc…
$ in.clothes_dryer                     <chr> "Gas, 100% Usage", "Gas, 100% Usage", "Gas, 100% Usage", "Gas, 100% Usage", …
$ in.clothes_washer                    <chr> "Standard, 100% Usage", "Standard, 100% Usage", "Standard, 100% Usage", "Sta…
$ in.clothes_washer_presence           <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes",…
$ in.cooking_range                     <chr> "Electric, 100% Usage", "Electric, 100% Usage", "Electric, 100% Usage", "Ele…
$ in.cooling_setpoint                  <chr> "72F", "72F", "72F", "72F", "72F", "72F", "72F", "72F", "72F", "72F", "72F",…
```

We also noticed that output provided for a variety of utilities utilized in July. Since we need the total energy consumed per hour per building we aggregated the energy columns.

````
merged_house_Static_energy_sum_out<-merged_house_Static_energy
#aggregating all the energy coloumns and summing to Final_enery_KWH
merged_house_Static_energy_sum_out$Final_Energy_KWH<- merged_house_Static_energy_sum_out %>%select(starts_with("out")) %>% rowSums(na.rm = TRUE)#

# removing out coloumns
merged_house_Static_energy_sum_out<- merged_house_Static_energy_sum_out[, -out_cols]
glimpse(merged_house_Static_energy_sum_out)

```

Rows: 137,040
Columns: 95
$ bldg_id                              <dbl> 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65, 65,…
$ in.sqft                              <dbl> 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 885, 88…
$ in.bathroom_spot_vent_hour           <chr> "Hour23", "Hour23", "Hour23", "Hour23", "Hour23", "Hour23", "Hour23", "Hour23", "Hour23…
$ in.bedrooms                          <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 2, 2, 2, 2, …
$ in.building_america_climate_zone     <chr> "Mixed-Humid", "Mixed-Humid", "Mixed-Humid", "Mixed-Humid", "Mixed-Humid", "Mixed-Humid…
$ in.ceiling_fan                       <chr> "Standard Efficiency", "Standard Efficiency", "Standard Efficiency", "Standard Efficien…
$ in.city                              <chr> "SC, Rock Hill", "SC, Rock Hill", "SC, Rock Hill", "SC, Rock Hill", "SC, Rock Hill", "S…
$ in.clothes_dryer                     <chr> "Gas, 100% Usage", "Gas, 100% Usage", "Gas, 100% Usage", "Gas, 100% Usage", "Gas, 100%…
$ in.clothes_washer                    <chr> "Standard, 100% Usage", "Standard, 100% Usage", "Standard, 100% Usage", "Standard, 100%…
$ in.clothes_washer_presence           <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Ye…
````

Firstly, we looked into the total energy consumption by different variables. We saw that the energy consumption per hour in July was taking negative values for some homes. This was because of “in.pv” variables. They represented the solar power panels that generated a surplus of energy for various building ids.

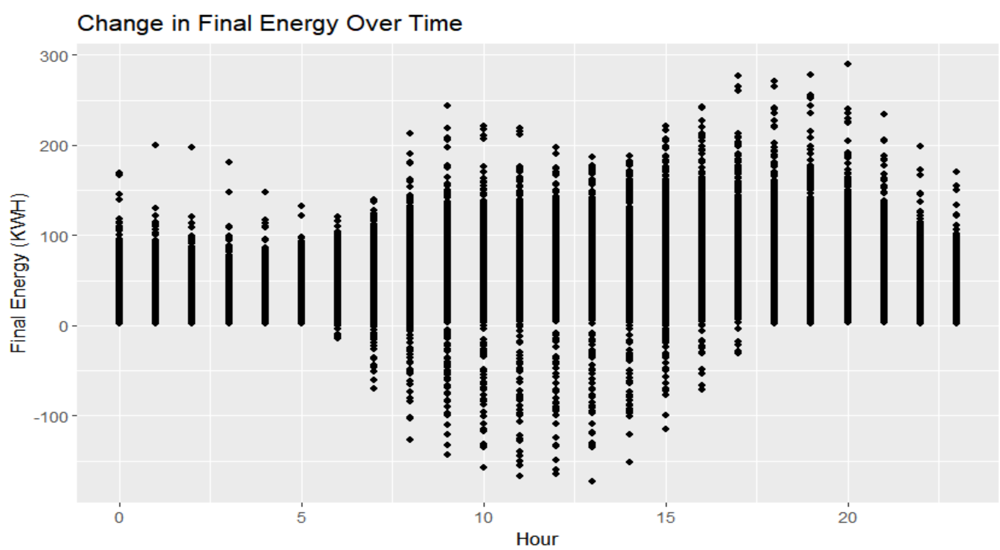

Fig. 2.6
````

Some of the other variables that had a strong positive correlation were the number of bedrooms and the size of a building. Below are the graphs showing a fitted linear regression line to showcase the same.

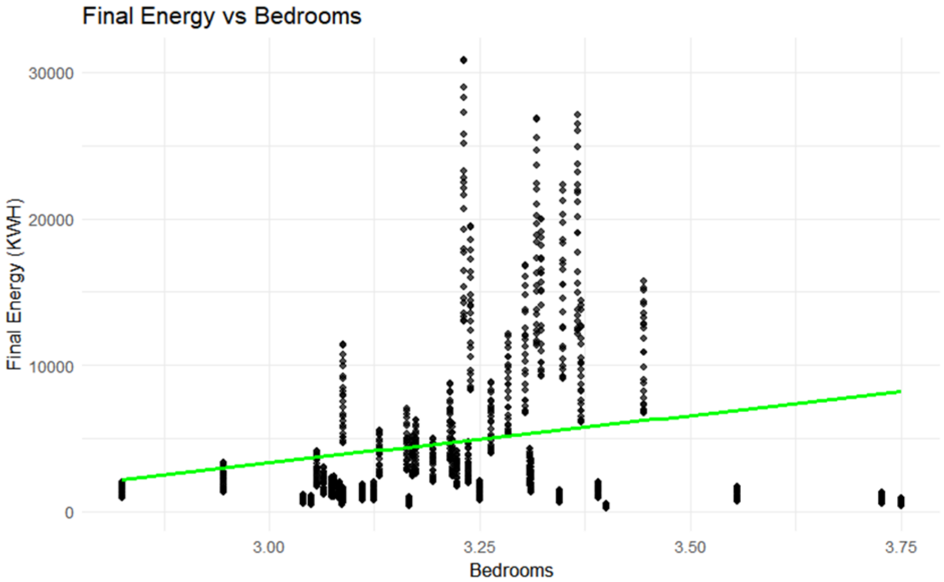

Fig. 2.7

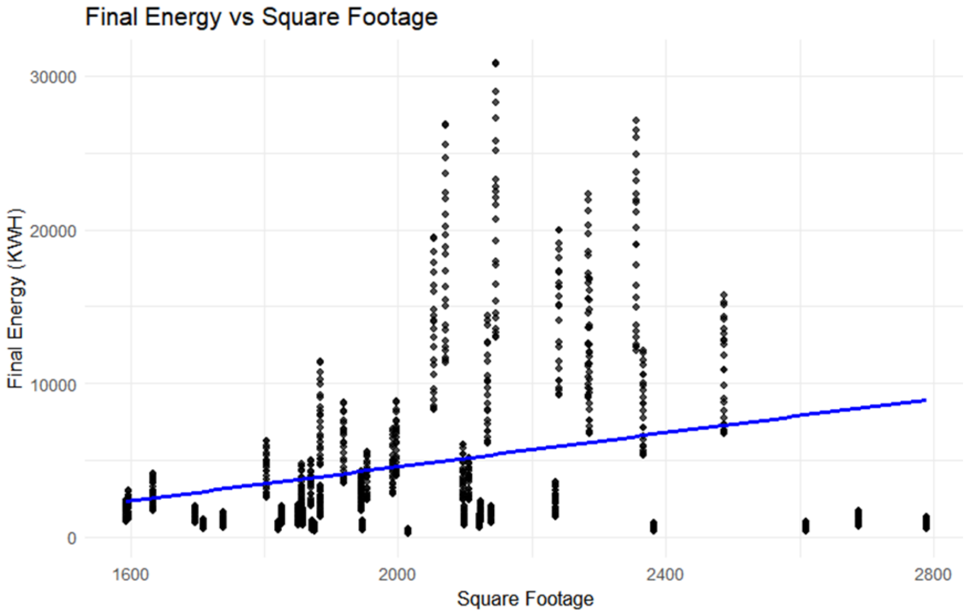

Fig. 2.8

There was also a similarity in pattern for energy consumption (by hour) for all of the counties but of different magnitude.

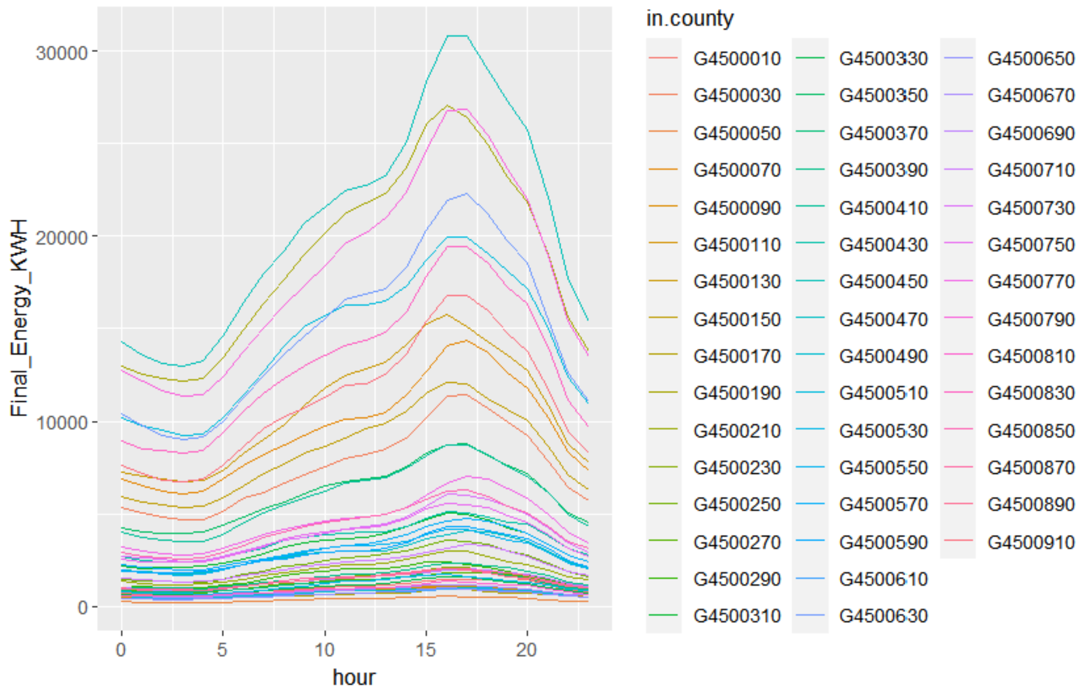

Fig. 2.9

We also explored the categorical variables to select features to put in the model. Based on humid or dry weather how we saw the mean consumption change :

Category of Weather <fctr>	Frequency <int>	Mean_Value <dbl>
Hot-Humid	39336	41.10031
Mixed-Humid	97704	37.87466

We also tested it for some of the variables that in theory would not contribute a lot to energy consumption. Like ceiling fans, but we found it to be contradictory.

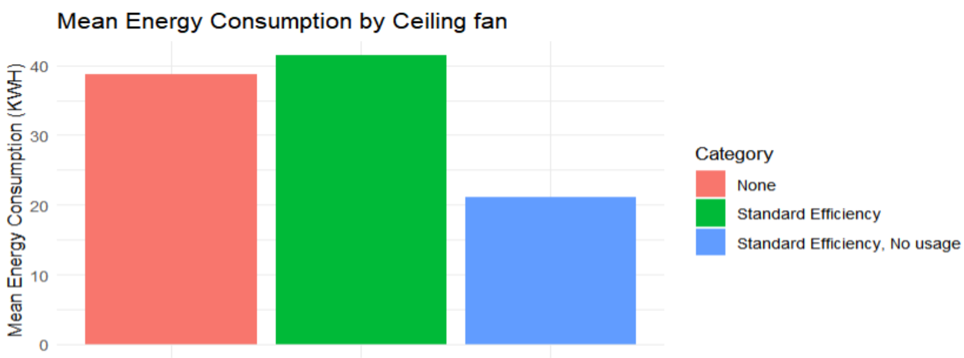

Fig. 2.10

Finally, weather data consisted of 1104 rows and 9 columns for every day of July, which were all numeric. For the sake of efficiency weather data was also averaged out for July to showcase daily averages of Temperature, Humidity, Wind, and a variety of radiations.

```
tibble [1,104 x 9] (S3: tbl_df/tbl/data.frame)
 $ in.county                        : chr [1:1104] "G4500010" "G4500010" "G4500010" "G4500010" ...
 $ hour                             : num [1:1104] 0 1 2 3 4 5 6 7 8 9 ...
 $ Dry Bulb Temperature [°C]        : num [1:1104] 22.4 22.1 21.8 21.6 21.5 ...
 $ Relative Humidity [%]            : num [1:1104] 95.2 95.7 96.6 96.9 96.9 ...
 $ Wind Speed [m/s]                 : num [1:1104] 1.089 0.932 0.978 0.729 0.956 ...
 $ Wind Direction [Deg]             : num [1:1104] 125.6 104.2 127.4 86 83.5 ...
 $ Global Horizontal Radiation [W/m2] : num [1:1104] 0 0 0 0 0 ...
 $ Direct Normal Radiation [W/m2]   : num [1:1104] 0 0 0 0 0 ...
 $ Diffuse Horizontal Radiation [W/m2]: num [1:1104] 0 0 0 0 0 ...
```

This data was merged with the Energy data for each house to generate the final file utilized for modeling. The final dataset had 130k rows and over 102 variables.

```
Rows: 137,040
Columns: 102
$ in.county                              <chr> "G4500010", "G4500010", "G4500010", "G4500010", "G4500010", "G4500010", "G4500010", "G4…
$ hour                                   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ `Dry Bulb Temperature [°C]`            <dbl> 22.35581, 22.35581, 22.35581, 22.35581, 22.35581, 22.35581, 22.35581, 22.35581, 22.3558…
$ `Relative Humidity [%]`                <dbl> 95.18613, 95.18613, 95.18613, 95.18613, 95.18613, 95.18613, 95.18613, 95.18613, 95.1861…
$ `Wind Speed [m/s]`                     <dbl> 1.0893548, 1.0893548, 1.0893548, 1.0893548, 1.0893548, 1.0893548, 1.0893548, 1.0893548,…
$ `Wind Direction [Deg]`                 <dbl> 125.5919, 125.5919, 125.5919, 125.5919, 125.5919, 125.5919, 125.5919, 125.5919, 125.591…
$ `Global Horizontal Radiation [W/m2]`   <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ `Direct Normal Radiation [W/m2]`       <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ `Diffuse Horizontal Radiation [W/m2]`  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ bldg_id                                <dbl> 410602, 465218, 473719, 29915, 102598, 402928, 450383, 350625, 107360, 376705, 32902, 2…
$ in.sqft                                <dbl> 1220, 2176, 3301, 2663, 1690, 2176, 1220, 2176, 1690, 1220, 1690, 2176, 1220, 3301, 266…
$ in.bathroom_spot_vent_hour             <chr> "Hour20", "Hour11", "Hour4", "Hour19", "Hour0", "Hour1", "Hour0", "Hour11", "Hour21", "…
$ in.bedrooms                            <dbl> 4, 4, 5, 3, 3, 4, 3, 4, 3, 2, 4, 3, 3, 5, 3, 3, 3, 3, 3, 3, 3, 3, 4, 1, 5, 4, 3, 2, 4, …
$ in.building_america_climate_zone       <chr> "Mixed-Humid", "Mixed-Humid", "Mixed-Humid", "Mixed-Humid", "Mixed-Humid", "Mixed-Humid…
$ in.ceiling_fan                         <chr> "Standard Efficiency", "Standard Efficiency", "Standard Efficiency", "Standard Efficien…
$ in.city                                <chr> "In another census Place", "Not in a census Place", "Not in a census Place", "Not in a …
```

Lastly, we referred to the metadata mainly while considering features for the model and also being conscious of the units of variables while aggregating and interpreting the final merged data.

3. Feature Engineering

In analyzing energy consumption patterns, feature engineering plays a pivotal role in preparing and enhancing the dataset for predictive modeling and analysis. It also allows relevant features from the dataset that might have a significant impact on energy consumption patterns.

Feature engineering also helps in selecting or creating the most relevant features, reducing dimensionality, and ensuring that the model does not overfit due to a feature that doesn't impact the model or incorrectly represents the energy consumption.

We looked at each of the 102 variables carefully and decided based on three major factors, if we wanted to keep that variable at all. First, we visualized the mean energy consumption using bar graphs for each categorical variable to discern potential variations. Subsequently, we delved into academic literature to investigate potential correlations between these variables and energy consumption. Finally, we subjected all the variables to a linear regression model to ascertain their statistical significance to energy consumption.

For example, we saw a significant increase in average energy consumption for hot water fixtures in the house. This was unexpected as during the summer there isn't a need for the hot water fixtures to run at a 200% percent capacity let alone 100%.

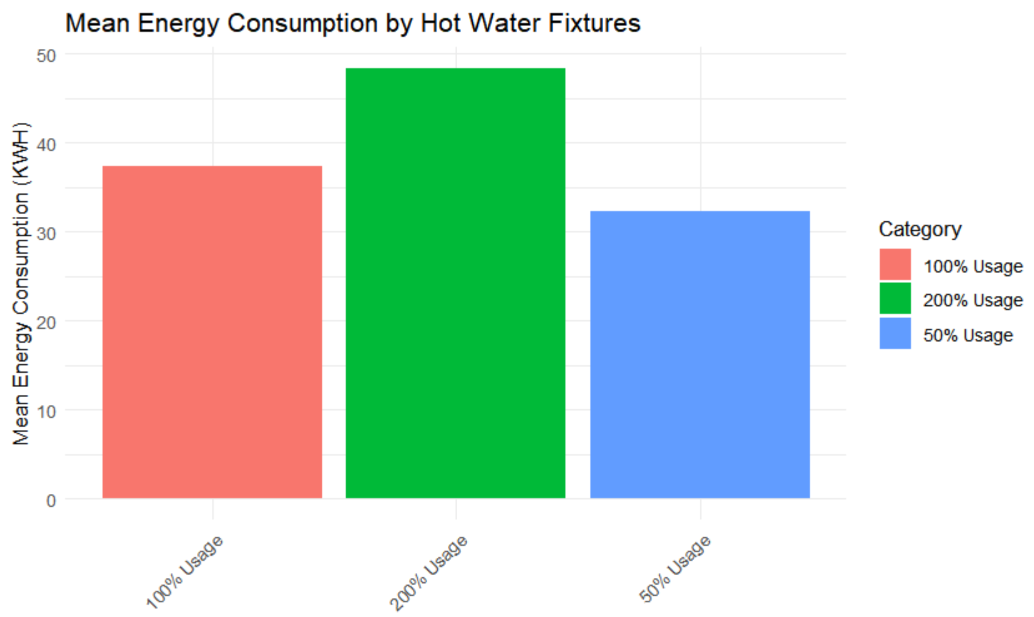

Fig. 3.0

We also saw a very low correlation for income (of about 0.04) and decided that it was not one of the major factors for energy consumption.

Now, one of the more interesting observations was for “in.infiltration” where we expected the energy consumption to increase with an increase in infiltration (which is essentially a leakage in air insulation from outside to inside and vice versa).

However, our analysis revealed inverse relationships between the two variables. This finding prompted a deeper investigation, leading us to the conclusion that higher the “ACH” (air change per hour) lowers the frequency at which the inside air is replaced with the outside.

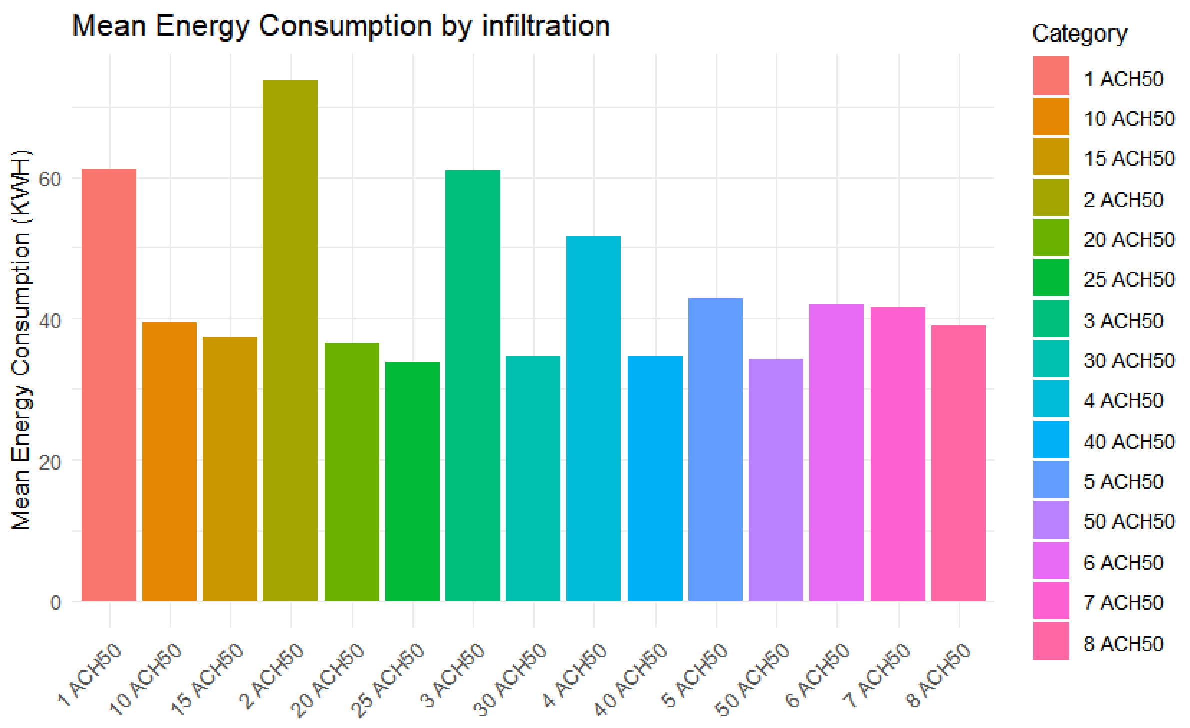

Fig. 3.1

With the same iterative process, we were able to choose around 40 variables. Finally, we explored the weather aspects of the dataset and found some strong linear relationships as shown below :

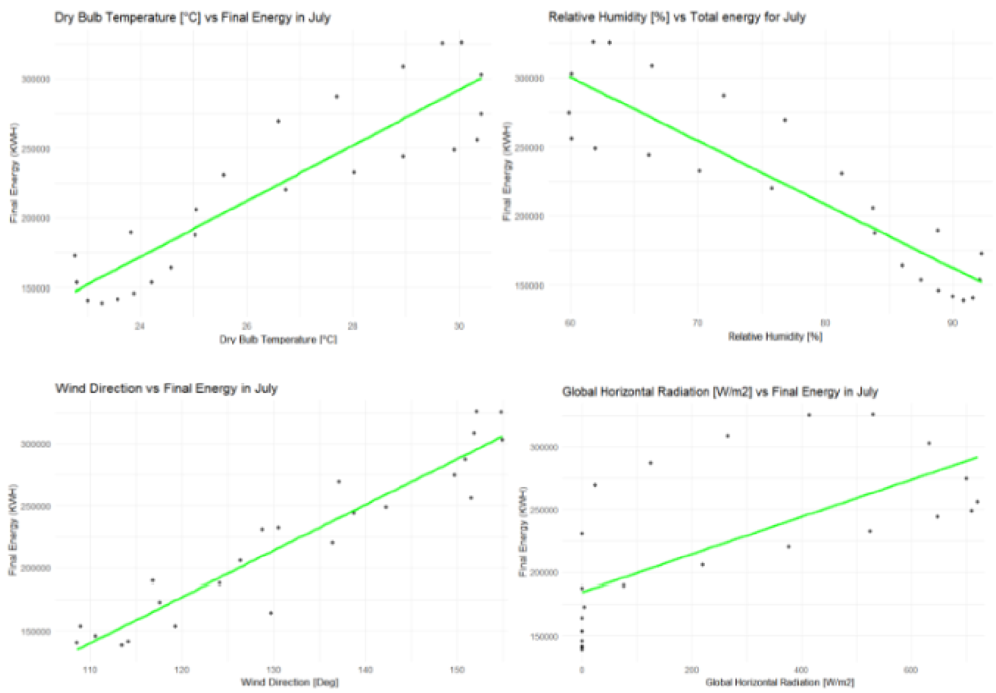

Fig. 3.2

Overall, all the weather-related variables had some sort of strong correlation with energy consumption. We also saw an inverse relation between energy and wind which was quite interesting.

Upon constructing a correlation matrix, we observed a robust relationship among the variables. Consequently, we opted to retain all of them for our modeling phase.

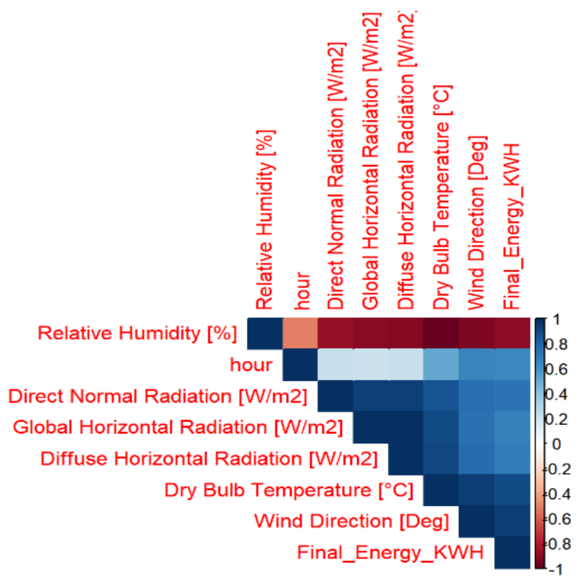

Fig. 3.3

4. Modelling

The modeling phase in our project involves testing out a variety of variables together, training, and evaluation. This was an iterative process and we mainly used linear regression, gradient boost, and XGBoost to do so. Our findings of each of the models are listed below.

Linear Regression: establishes a linear relationship between the independent variables and the dependent variable. It aims to minimize the difference between the observed and predicted values by fitting a straight line to the data.

We started experimenting with the linear regression model to find statistically significant variables for energy consumption. We tried up to 20 different models and over time with added variables, the adjusted r-squared decreased with an increase in multiple r-squared. This meant that the model was being penalized for using too many variables.

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.02 on 136898 degrees of freedom
Multiple R-squared:  0.6603,    Adjusted R-squared:  0.6599
F-statistic:  1887 on 141 and 136898 DF,  p-value: < 2.2e-16
```

It also peaked in terms of the adjusted r-squared at 69%. With this result, we concluded that linear regression was not a good fit for our data and moved to testing our data with Gradient Boost.

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.59 on 136848 degrees of freedom
Multiple R-squared:  0.6823,    Adjusted R-squared:  0.6818
F-statistic:  1539 on 191 and 136848 DF,  p-value: < 2.2e-16
```

Gradient Boosting: is an ensemble learning method that builds multiple decision trees sequentially. It minimizes errors by iteratively training new models to correct the mistakes of previous models. The model produces a strong predictive model by combining multiple weak models (decision trees), focusing on the residuals (errors) of the previous models.

Gradient Boosting was an ideal choice for predicting Final Energy Consumption due to its ability to capture complex non-linear relationships in the data. It's an ensemble learning technique that combines the strengths of multiple models, making it robust against outliers and providing high predictive accuracy.

Additionally, Gradient Boosting handles both numerical and categorical variables effortlessly, reducing the need for extensive data preprocessing. Its built-in regularisation techniques prevent overfitting, and feature importance scores help identify key predictors of energy consumption.

Overall, gradient-boosting flexibility and predictive power make it a strong candidate for accurate energy consumption predictions.

```r
# Define the parameters for the Gradient Boosting model
gbm_params <- list(
  distribution = "gaussian",
  n.trees = nrounds,
  interaction.depth = 8,
  shrinkage = 0.1,
  bag.fraction = 0.5
)

# Train the Gradient Boosting model
gbm_model <- gbm(
  formula = Final_Energy_KWH ~ .,
  data = train_data,
  distribution = gbm_params$distribution,
  n.trees = gbm_params$n.trees,
  interaction.depth = gbm_params$interaction.depth,
  shrinkage = gbm_params$shrinkage,
  bag.fraction = gbm_params$bag.fraction,
  verbose = FALSE
)
```

We tested the goodness of our model by Root mean squared error and R-squared going forward. RMSE is a measure of the average deviation of predicted values from the actual observed values. It represents the square root of the average of the squared differences between predicted and actual values. R-Squared talks about how much of the variability in the data the model accounts for.

In this case, the RMSE value of approximately 6.93 suggests that, on average, the predictions of the model are around 6.93 units away from the actual values. R- Squared had significantly improved for the model making it a better fit for the data than linear regression. This was a good result but we wanted to see if RMSE could be lowered further so we tested out Extreme Gradient boost onto our dataset.

```
> rmse_gbm <- sqrt(mean((predictions_gbm - test_data$Final_Energy_KWH)^2))
> print(paste("RMSE (GBM):", rmse_gbm))
[1] "RMSE (GBM): 6.93080183569941"
>
> # Compute R-squared
> SST_gbm <- sum((test_data$Final_Energy_KWH - mean(test_data$Final_Energy_KWH))^2)
> SSR_gbm <- sum((predictions_gbm - test_data$Final_Energy_KWH)^2)
> r_squared_gbm <- 1 - SSR_gbm/SST_gbm
> print(paste("R-squared (GBM):", r_squared_gbm))
[1] "R-squared (GBM): 0.902063067390577"
```

```
> importance_summary <- summary(gbm_model)
> importance_summary
                                                                   var      rel.inf
in.sqft                                                        in.sqft 2.041027e+01
`Dry Bulb Temperature [°C]`                `Dry Bulb Temperature [°C]` 1.213516e+01
in.vacancy_status                                    in.vacancy_status 9.165926e+00
hour                                                              hour 6.951033e+00
`Direct Normal Radiation [W/m2]`      `Direct Normal Radiation [W/m2]` 5.886754e+00
in.hot_water_fixtures                            in.hot_water_fixtures 5.767530e+00
in.county                                                    in.county 4.553274e+00
in.occupants                                              in.occupants 4.306346e+00
in.pv_system_size                                    in.pv_system_size 4.290164e+00
in.cooling_setpoint                                in.cooling_setpoint 4.202246e+00
in.misc_pool_pump                                    in.misc_pool_pump 3.036000e+00
`Relative Humidity [%]`                        `Relative Humidity [%]` 2.747264e+00
in.ducts                                                      in.ducts 2.698718e+00
in.lighting                                                in.lighting 1.895096e+00
in.infiltration                                        in.infiltration 1.876884e+00
in.cooling_setpoint_offset_magnitude  in.cooling_setpoint_offset_magnitude 1.189164e+00
in.insulation_wall                                  in.insulation_wall 1.160132e+00
in.clothes_washer                                    in.clothes_washer 8.862936e-01
`Diffuse Horizontal Radiation [W/m2]` `Diffuse Horizontal Radiation [W/m2]` 8.586879e-01
`Global Horizontal Radiation [W/m2]`   `Global Horizontal Radiation [W/m2]` 8.143707e-01
`Wind Speed [m/s]`                                  `Wind Speed [m/s]` 7.862158e-01
```

<u>XGBoost (Extreme Gradient Boosting)</u>: is an optimized and scalable implementation of gradient boosting. It enhances the gradient boosting method by employing techniques like regularisation, parallel computing, and tree pruning to improve accuracy and speed.

XGBoost is selected as our model primarily because it addresses the issue of overfitting by incorporating built-in regularisation techniques. This is crucial given the large number of variables and factors in our model. XGBoost is also effective at handling non-linear relationships between predictor and target variables, which is important for our analysis.

Additionally, it demonstrates robustness against outliers, a valuable feature considering the presence of energy houses of air bases, airports, and similar data points which are some outliers considering high energy consumption.

Moreover, XGBoost is more focused on making predictions rather than drawing inferences, aligning well with the project's prediction-oriented objectives. Furthermore, XGBoost provides feature importance scores, allowing us to pinpoint the variables that exert the most significant influence on the prediction, aiding in identifying key drivers of Final Energy Consumption.

```
cols_4<-c('hour',
            'in.county',
            'Dry Bulb Temperature [°C]','Relative Humidity [%]','Wind Speed [m/s]',
            'Wind Direction [Deg]','Direct Normal Radiation [W/m2]','Diffuse Horizontal Radiation [W/m2]',
            'Global Horizontal Radiation [W/m2]', 'in.sqft',
            'in.bedrooms',
            'in.building_america_climate_zone',
            'in.ceiling_fan',
            'in.clothes_dryer',
            'in.clothes_washer',
            'in.cooling_setpoint',
            'in.cooling_setpoint_has_offset',
            'in.cooling_setpoint_offset_magnitude',
            'in.dishwasher',
            'in.ducts',
            'in.geometry_foundation_type',
            'in.geometry_wall_type',
            'in.geometry_stories',
            'in.has_pv',
            'in.heating_fuel',
            'in.hot_water_fixtures',
            'in.hvac_cooling_partial_space_conditioning',
            'in.hvac_cooling_type',
            'in.hvac_heating_type',
            'in.hvac_heating_type_and_fuel',
            'in.infiltration',
            'in.insulation_ceiling',
            'in.insulation_wall',
            'in.lighting',
            'in.misc_extra_refrigerator',
            'in.misc_freezer',
            'in.misc_pool_pump',
            'in.occupants',
```

```
params <- list(
  objective = "reg:squarederror",
  eta = 0.1,
  max_depth = 8,
  subsample = 0.5,
  colsample_bytree = 0.5
)

nrounds <- 3000  # Number of boosting rounds. Adjust based on your dataset and needs
```

In this scenario, the RMSE value stands at approximately 6.31, which marks the lowest among the two models we evaluated. Although the difference isn't substantial, it could significantly impact the accuracy of energy consumption estimates in predictive analysis.

The R-squared value, approximately 0.919 (or 91.9%), denotes that roughly 91.9% of the variability in the dependent variable is explained by the independent variables in the model. Similar to that of the previous model but higher.

```
[1] "RMSE: 6.40537655545942"
[1] "R-squared: 0.916349437434933"
```

This analysis reveals the key variables crucial for predicting energy consumption. Factors such as building size, presence of ceiling fans, infiltration rates, solar panel size, and similar aspects emerge as major contributors—elements within our control that significantly influence consumption. Additionally, a majority of the weather-related factors also stand out as primary influencers of energy usage.

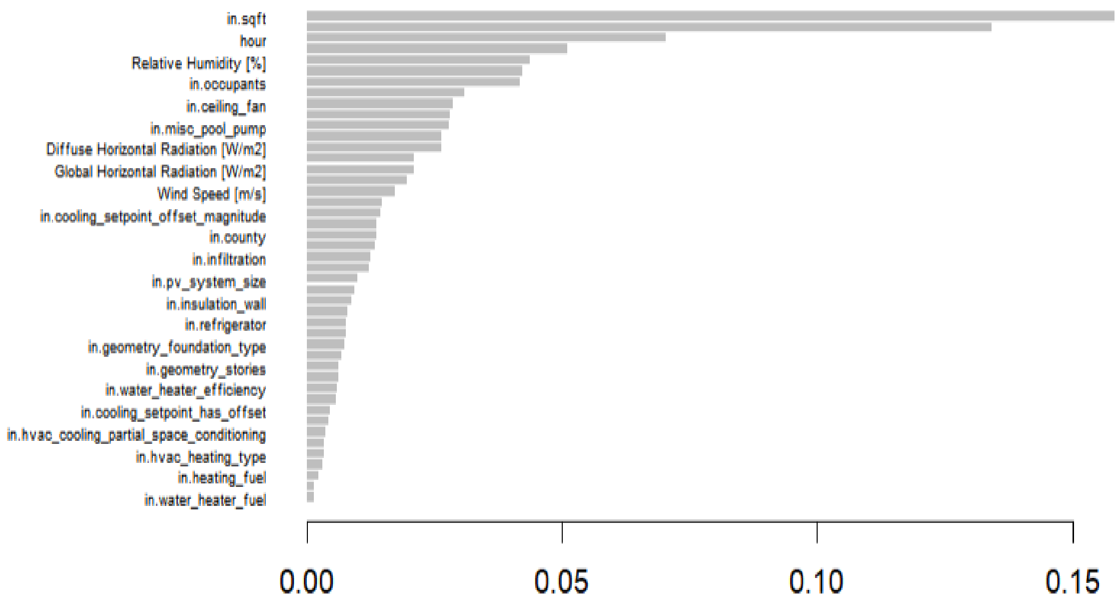

Fig. 4.1

With all of the above in mind, we decided to choose XG Boost as our final model, since it worked the best with our dataset and had the lowest root mean squared error as well as the highest accuracy.

5. Energy Prediction at Warmer Temperatures

With the final model, we predicted what the energy consumption would look like for all the buildings across the counties in South Carolina if there was a 5 ℃ increase in temperatures. Greenville has one of the highest energy consumptions followed by Colleton and so on.

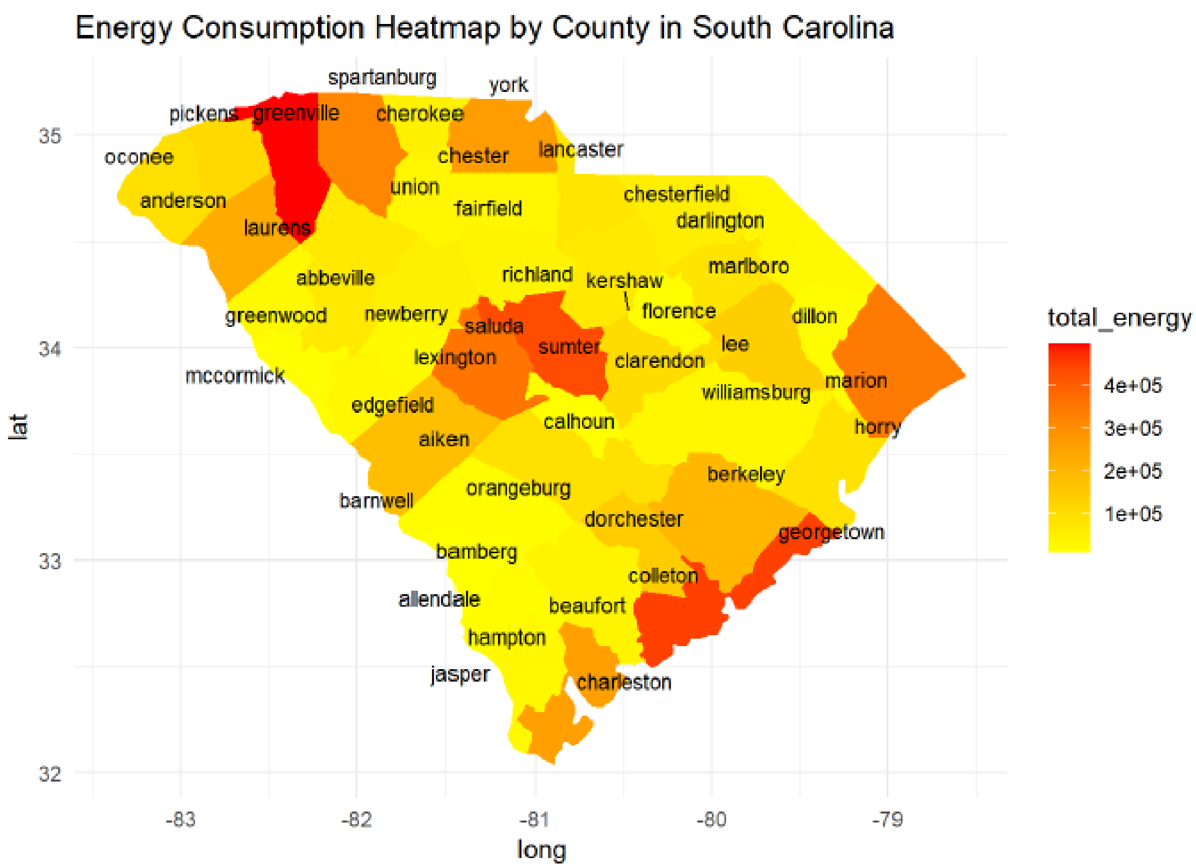

Fig. 4.2

This map is in line with the energy consumption we saw before the temperature increase which confirms no drastic change in energy consumption, just a general increase across the counties.

We also looked at county-wise percent increase in energy usage. Again, Greenville has one of the largest increases in terms of magnitude. However, Horry has had the largest percentage increase of 33% across all the counties as shown below, making it the county with the largest impact on temperature change.

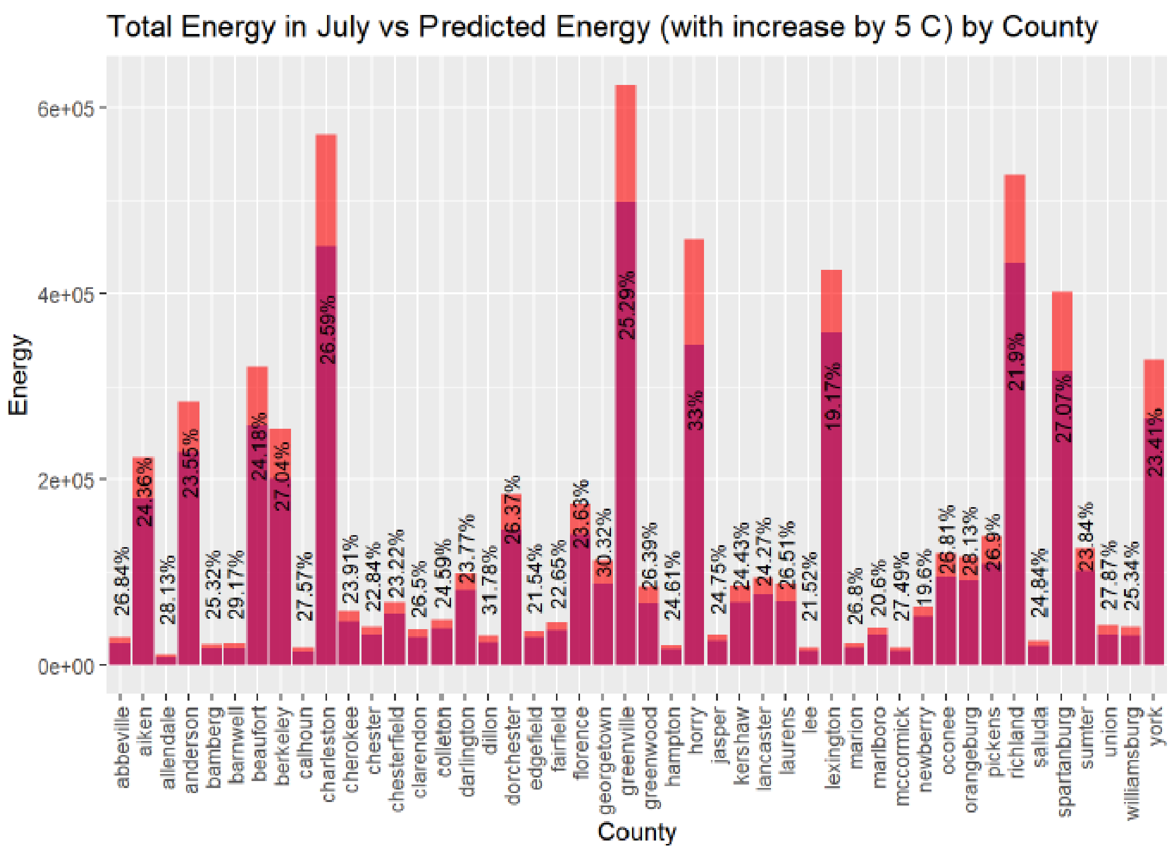

Fig. 4.3

This graph illustrates the hourly temperature increase throughout July. We see the energy consumption peak at around 4 pm and then it starts to come down from there.

Even with an increase in temperature, the pattern has not changed, just the magnitude of consumption has increased. Even with the temperature rise, the observed pattern in energy consumption remains consistent; only the magnitude of consumption has shown an increase.

We could implement load-shedding strategies to strategically schedule intentional blackouts during periods of anticipated high energy demand. By identifying consistent patterns, blackout schedules can be optimized to minimize disruption while efficiently managing energy usage during peak periods. Communicate with consumers about peak energy usage periods, encouraging them to adjust their usage behavior during these times to help alleviate strain on the grid. Further, Coordinating maintenance activities for power infrastructure during periods of relatively lower consumption minimizes disruptions for consumers while allowing necessary maintenance work on the grids.

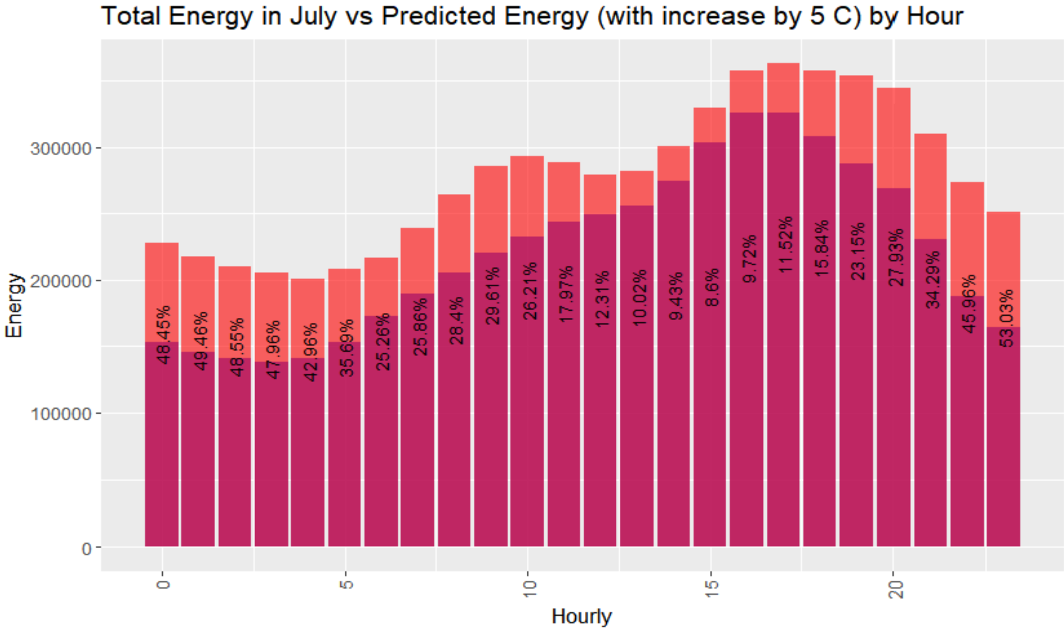

Fig. 4.4

In conclusion, the model revealed a notable 30% surge in energy consumption, marking a concerning escalation for a mere 5 degree Celsius increase. The sharp increase underscores the substantial influence that relatively minor temperature variations can wield on energy demand, signifying a critical aspect to consider in managing and planning for sustainable energy usage.

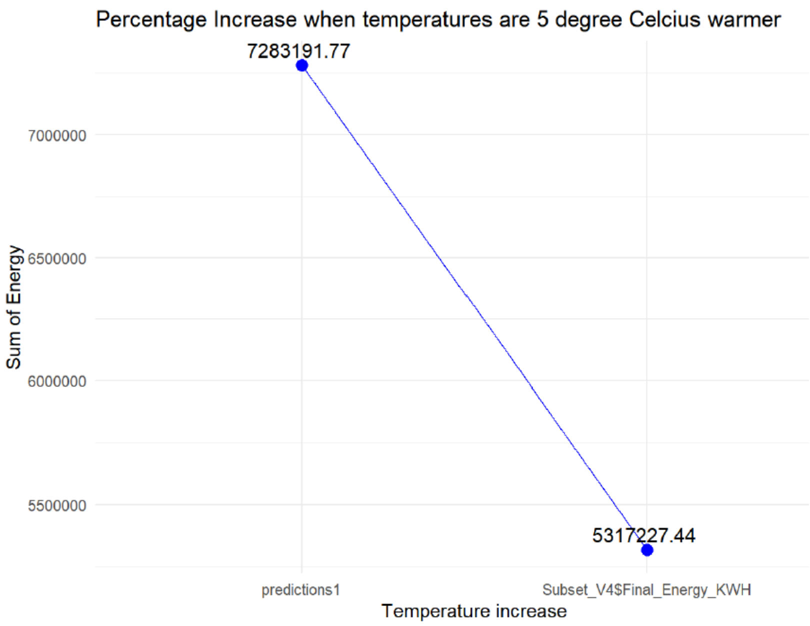

Fig. 4.5

6. Actionable Energy Efficiency Strategies

With the recent surge in energy consumption, particularly highlighted by a substantial 30% increase linked to just a 5-degree Celsius rise in temperature, the imperative for implementing actionable energy efficiency strategies becomes increasingly evident. The observed spike in energy demand not only underscores the sensitivity of energy consumption to minor temperature variations but also raises critical concerns about sustainability, resource management, and the ecological footprint of heightened energy usage.

In light of these developments, this section aims to explore and propose actionable energy efficiency strategies. These strategies, designed to curtail energy demand, mitigate environmental impact, and foster sustainability, are crucial for meeting the escalating energy needs while simultaneously striving for responsible resource utilization.

Firstly, we tried to change some of the more obvious variables that impact energy consumption. When we turned the ceiling fan to optimal efficiency, the hot water fixtures, and ACH levels to the most optimized energy usage we saw a minute decrease in energy consumption.

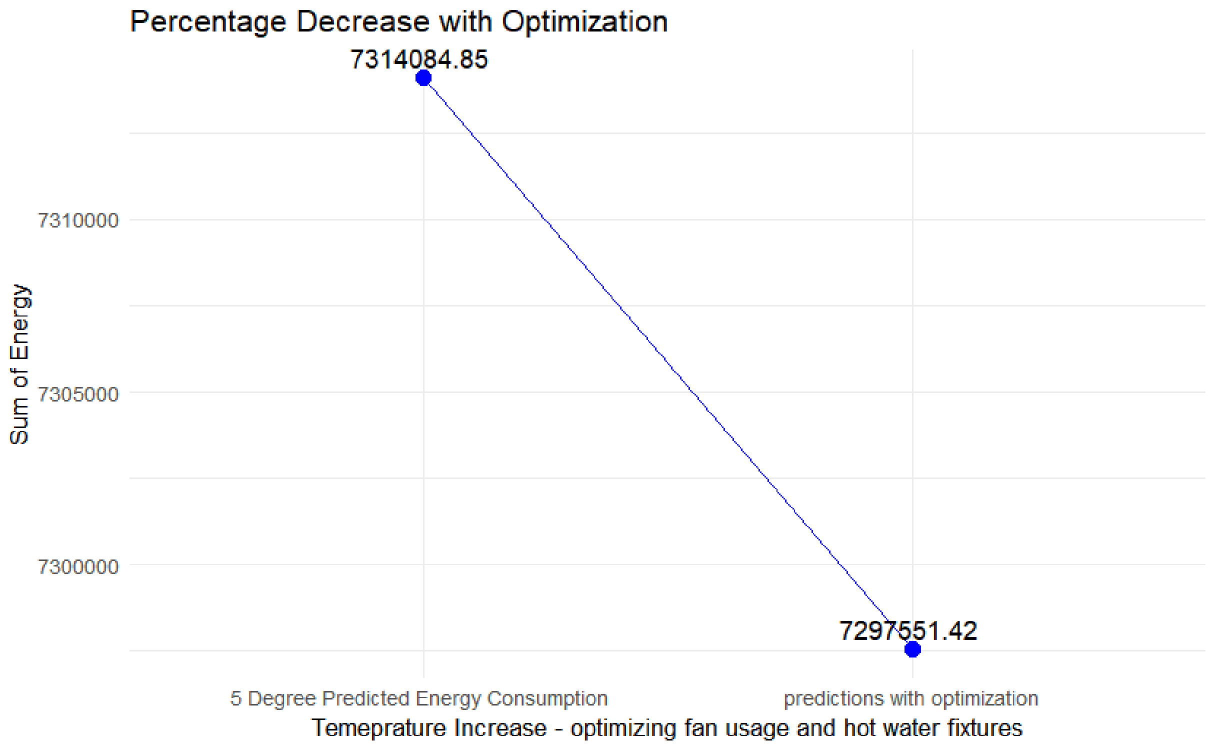

Fig. 4.6

Despite observing a 1% decrease in energy consumption despite adjustments in three significant factors, delving deeper into the dynamics of these changes revealed insightful findings.

Implementing such a strategy would be more cost effective, but its feasibility remains a concern. Every state mandates a minimum Air Changes per Hour (ACH) required for buildings to maintain, typically around 4 ACH. To achieve a mere 1% decrease in energy consumption, we would need to set it as high as 15 ACH. However, this significantly exceeds the legally required threshold, rendering this approach cost-effective but illogical in practice.

To make it easier to reduce energy we looked at the size of solar panels installed. Instead of increasing the sizes of solar panels in buildings with already existing solar panels. We recommended the buildings with no solar panels install the smallest one of the "1KwDC" Solar Panel Systems.

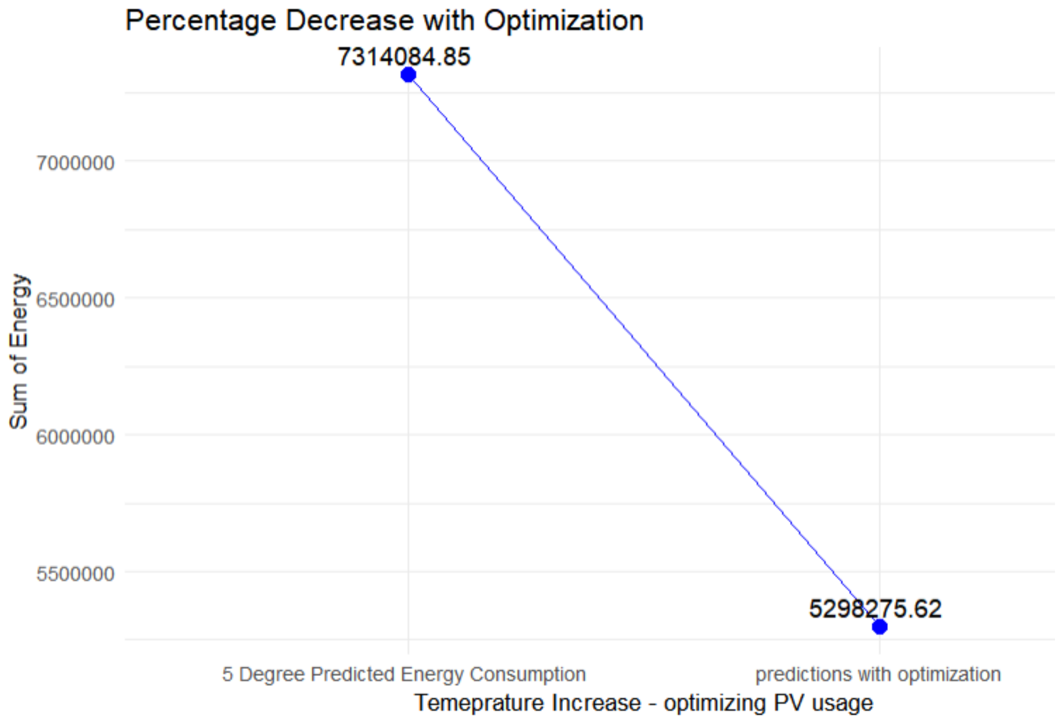

Fig. 4.7

7. Conclusion

Based on our findings, the recommendation to install '1KwDC' solar panels on buildings without existing solar panels yielded the most significant impact, resulting in a notable 28% reduction in energy consumption.

This decrease underscores the cost-effectiveness of this approach compared to alternatives such as altering architectural insulation within houses or enhancing the efficiency of appliances. House insulation enhancements necessitate comprehensive structural changes and material upgrades, potentially involving higher initial investment costs. Similarly, improving appliance efficiency often involves replacing existing appliances with higher-rated energy-efficient models, which might also incur substantial upfront expenses.

The cost-effectiveness of installing '1KwDC' solar panels resides in its ability to harness renewable energy directly from the sun, thereby reducing dependency on traditional energy sources.

Moreover, this installation incurs a one-time cost that, in the long run, offers sustainable energy generation, diminishing reliance on conventional grid-based electricity. The scalability and modularity of solar panel installations render them adaptable to various building sizes and energy requirements, making them a versatile and cost-efficient solution. In contrast, altering house insulation or upgrading appliance efficiency, while viable strategies, might involve more intricate and expensive modifications.

Here’s a link to Shiny App: https://keerthikrishnaaiyappan.shinyapps.io/ShinyApp/