**Project Overview, Introduction To Data Science**

**Project Description**

ESC is a company that provides electricity to residential properties in South Carolina and a small part of North Carolina. The company is concerned about global warming, specifically the impact of global warming on the demand for their electricity. Specifically, they are worried that next summer will put too much demand on their electrical grid and on their ability to supply electricity to their customers when they want to cool their homes. If this happens there will be blackouts, which eSC wants to avoid.

Rather than build out the capability to deliver more energy to their clients, they want to understand the key drivers of energy usage, and how they could encourage their customers to save energy. In short, their goal is to reduce energy usage if next summer is ‘extra hot’, so that they can meet demand. eSC is focused on July energy usage as the company thinks that July is typically the highest energy usage month.

**Data Preparation**

The static house data provided by the client is available in a parquet file format and contains information for 5,710 houses. The energy data is also provided in parquet format, with each house having its own energy usage data. The weather data, on the other hand, is provided in CSV format, with data available for each county in the static house dataset.

To load the data, we used the tidyverse package for handling CSV files and the arrow library for working with parquet files. We implemented for loops to retrieve the relevant datasets for energy and weather based on the building ID and county information from the static house dataset.

Then we dropped attributes based on two conditions; variables with similar value throughout the records such as none. Then we filter out the variables based on their NA counts.

**Understanding current energy usage patterns**

The dataset provided comprises three main sources of data: static house information, energy usage data per building per hour, and weather conditions per county. The static house dataset contains 5700 rows, each representing a unique building. The energy usage dataset records 8700 observations per building per hour, totaling a significant number of records. Additionally, the weather dataset includes weather conditions for 8900 observations per county.

To merge these datasets effectively, considering the large volume of data, we opted cloud services for data loading and merging. The merging process involved several steps. Firstly, we merged the static house data with the energy usage data based on the unique building ID, resulting in a dataset with a common 'time' column representing hourly intervals. Subsequently, we merged this combined dataset with the weather data based on the 'time' column and the 'county' id. To make our analysis more manageable, we looked at the counties that use the most energy, and at the counties that had a different climate from each other. This makes it so that we can see how energy use changes when temperatures are different, going from cold to warmer. This helps us understand the energy needs from people when it's really hot outside, and it makes it easier to predict the energy usage based on weather and temperature.

**Static House Data:**

* Includes 5,710 houses and 171 house attributes.
* Contains a wide range of information about houses like
* Geographic information: county, latitude and longitude coordinates and climate zone.
* Features: time period of construction, number of bedrooms, number of building stories, size of garage, area of the house's living space, dwelling unit foundation, exterior finish material, roof material type: The material used for the roof of the house.
* Resident: number of occupants, their income.
* Systems and Appliances related to energy : information on the presence and usage of various energy consuming appliances, heating and cooling systems, insulation levels, lighting and renewable energy systems.

**Energy Usage Data:**

* Includes 50,013,890 records of hourly energy usage data for each house taken between 2018-01-01 and 2018-12-31.
* Consists of 44 attributes with building id, time and 42 of the attributes describing;
* energy consumption from energy sources (electricity, fuel oil, natural gas, and propane) under different scenarios and time periods.
* energy consumption for various appliances and systems in relation to the size of the dwelling unit.

**Weather Data:**

* Includes 402,960 hourly weather data taken at different time points between 2018-01-01 and 2018-12-31 for each county. It has 9 attributes describing the weather in terms of temperature, humidity, wind and radiation.

We decided to start off with 5 business questions that we wanted to explore further. We thought this would help us in understanding our data more and getting to know the factors that had a big influence on our energy data. Our business questions were:

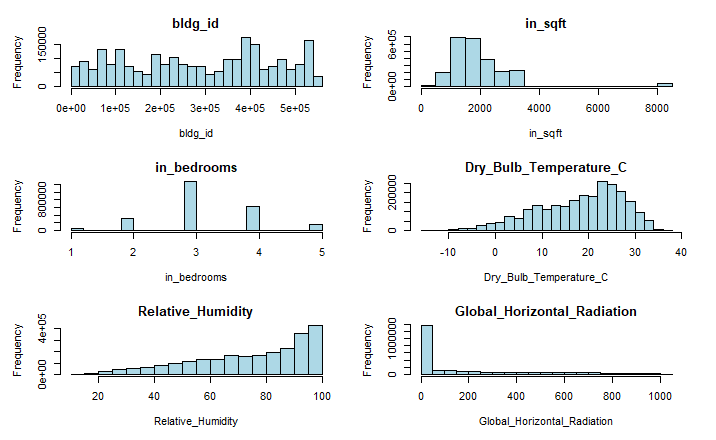
* Influence of the insulation in a house for energy usage.
* Influence of type of clothing washer/dryer on energy usage.
* Influence of weather humidity based on energy usage.
* Influence of ceiling fans based on energy usage.
* Hour in the week that has the most energy activity based on mixed humid or hot humid.
* Most energy usage based on highest temperatures

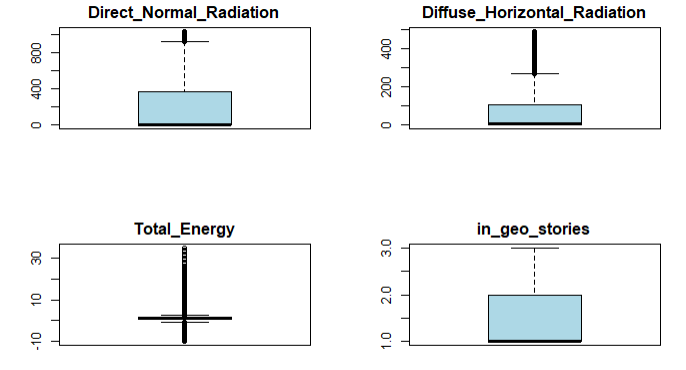
These were the 5 business questions we wanted to start off with. We were aware that they were not set in stone and that if anything had to change or on the way we found something different out it was okay for us to change it.

In our project we will talk about a column that is called total energy consumptions or sum of rows. For this column we aggregated all the energy columns from the energy dataset. These columns encompass various energy consumption sources, ranging from heating and cooling to lighting and appliances. By summing these values, we derived the total energy consumption metric, which serves as the target variable for our predictive model.

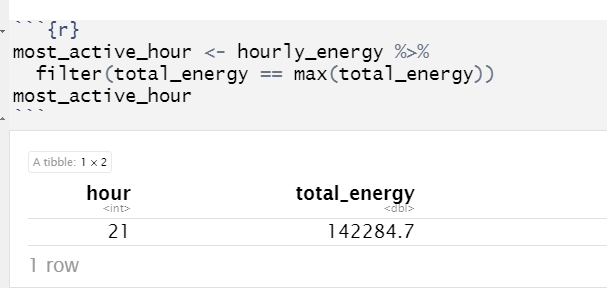
After successfully loading the data, we conducted exploratory data analysis, which provided insights on how to further proceed with modeling. This existed out of looking at the structure of the data, the summary and any futures of the dataset that caught our attention. Some columns of the energy contained negative values, or non unique values and after taking closer looks we decided what would be best to do with these values ( remove, take the average etc.)

After cleaning the data we decided to take a closer look at the numerical columns. We saw that some columns exhibit bias, and there were outliers present. However, these outliers cannot be removed as they are part of the original data, and none of them appear to be exceptional when compared to the overall dataset. There might be instances when energy usage was slightly higher what resulted in these outliers

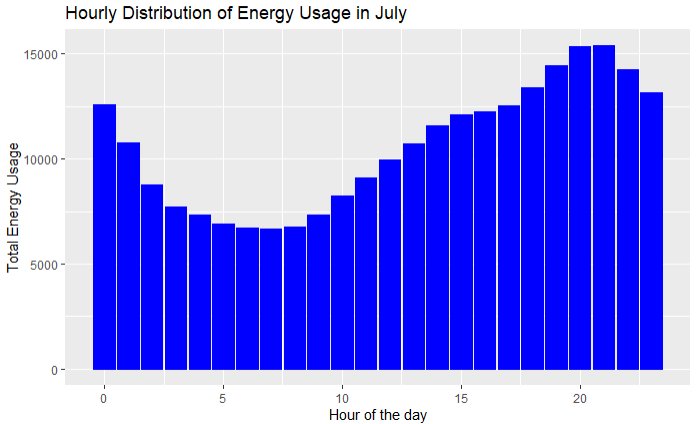




Furthermore, we filtered energy based on the hour and identified the most active hour, which is the 21st hour of the day (9:00PM), exhibiting a high total energy consumption.



We decided to plot the distribution of energy usage in the month of July based on the hour of the day. The bar graph below illustrates that the later hours of the day have the highest usage, with hours 20 and 21 exhibiting the highest levels.



More exploring of our data:

Graph that shows the total energy usage in the month of July:

A graph showing a graph showing a number of times

Description automatically generated with medium confidence

Graph that shows the temperature in the month of July:

A graph showing the time of a wave

Description automatically generated

Graph that shows the humidity in the month of July:

A graph showing the time of the month

Description automatically generated

A screenshot of a computer program

Description automatically generated

The energy consumption for the first 15 subsetted counties

A graph of energy consumption

Description automatically generated

The energy consumption for the next 15 subsetted counties

A graph of energy consumption

Description automatically generated

Energy Consumption for the last 16 counties

A graph of energy consumption

Description automatically generated

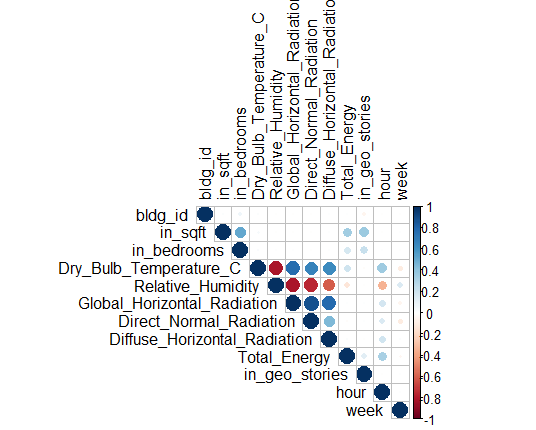
The top 4 highest energy consuming countries derived from the above subsetted graphs.

A graph of a number of people

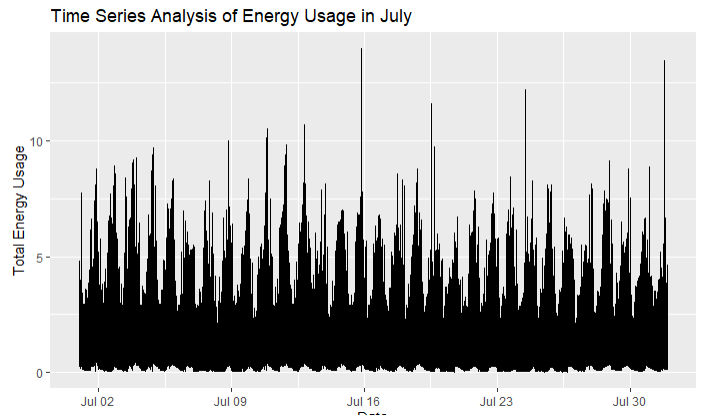
Description automatically generated

The hourly energy consumption in the month of July

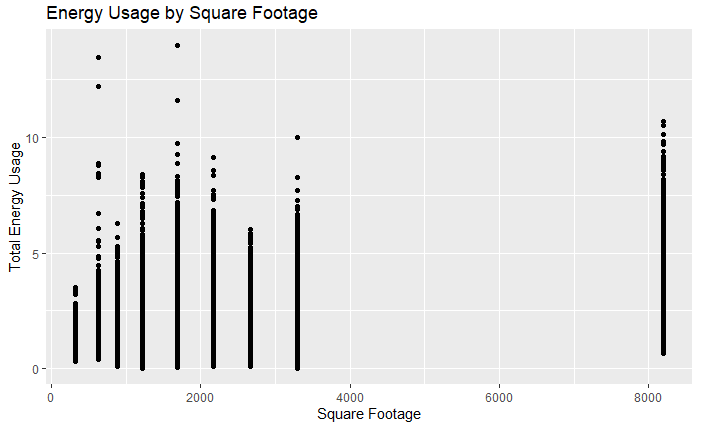
We also conducted correlation analysis to examine the relationship between various columns and our Total Energy column. The correlation plot below reveals that the columns Dry\_Bulb\_Temperature\_C, Relative\_Humidity, Direct\_Normal\_Radiation, and Diffuse\_Horizontal\_Radiation exhibit both positive and negative correlations with Total Energy.



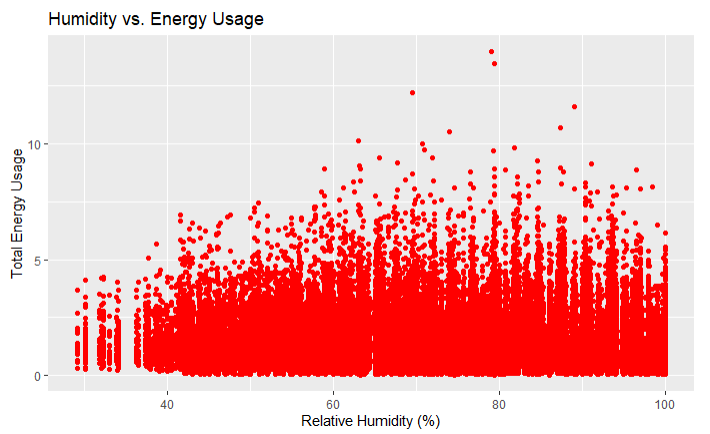
We then decided to perform a time series analysis to see energy consumption in july.



Next, we wanted to check whether energy usage is related to the size of the house as this might be of big influence on the total energy consumption. The analysis indicates that there somewhat a correlation between square footage and total energy usage as shown below.

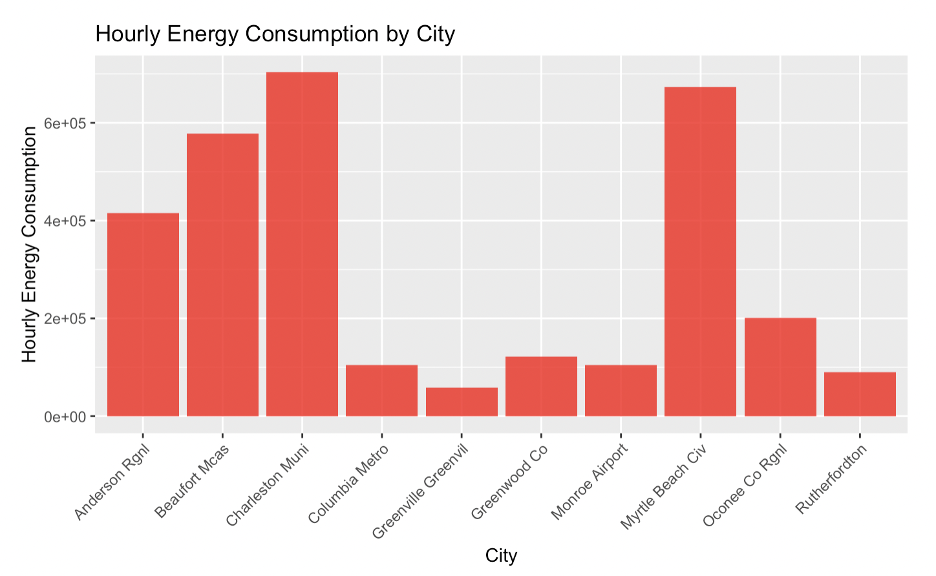


We also wanted to investigate whether humidity has any effect on total energy consumption. The analysis suggests that humidity does have some effect on total energy usage. Especially when humidity gets above 40%

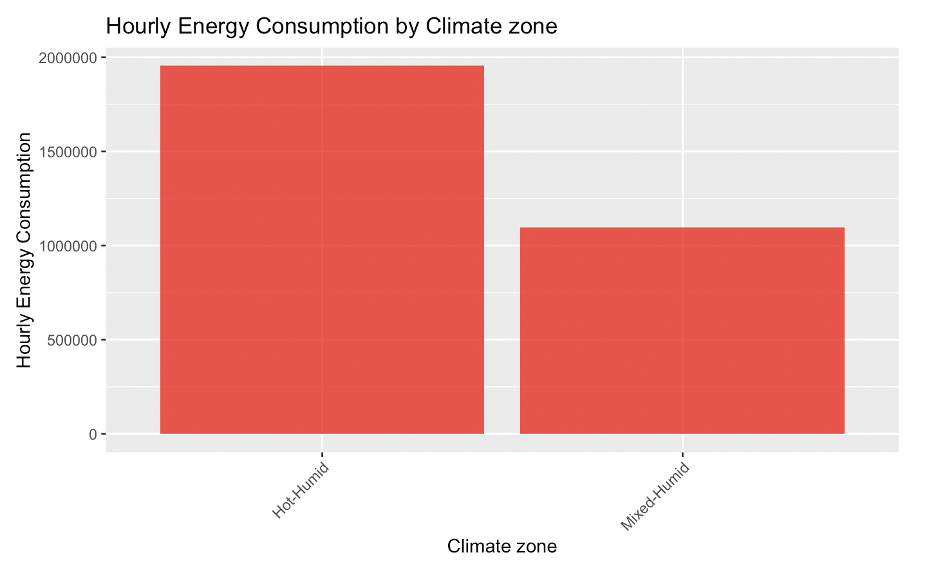


We decided to look some more into the City and counties and their energy behaviors

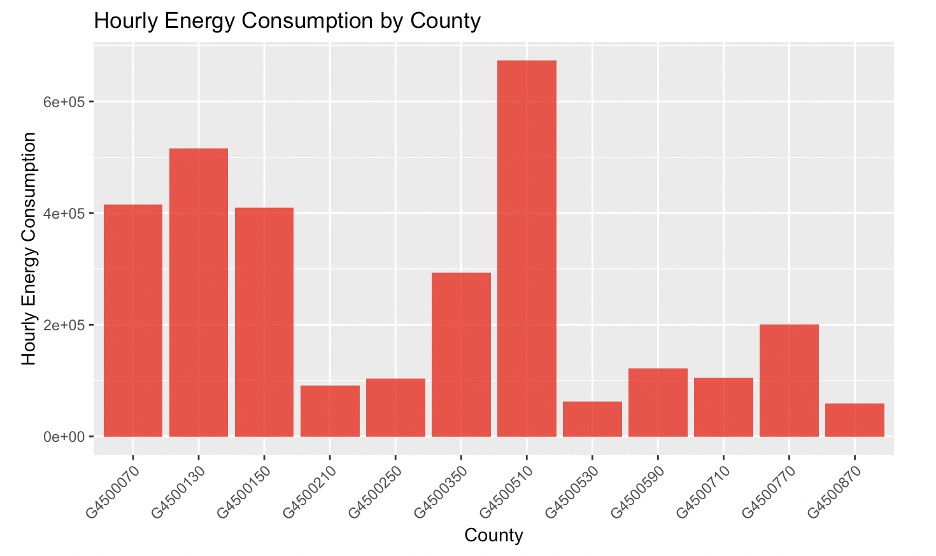
We used a sample of 25 counties from mix-humid and hot-humid. We picked mixed-humid and hot-humid as most of our business questions were based on these climates and that it will be helpful to see how climate influences the energy usage as well and if there is a big difference based on cities.



From the above graph we can see that the cities Charleston Muni and Myrtle beach Civ have the highest hourly energy consumption.



From the above graph we can see that the counties in the Hot-Humid climate zone have the highest hourly energy consumption as compared to the counties in the Mixed-Humid climate zone.



From the above graph we can see that the county G4500510 has the highest hourly energy consumption for the month of July.

After looking at all the outside factors we wanted to look more at factors within the house. For this we looked at county G4500730 which is a county in a mixed humid climate. This was a county with 110 houses.

After looking at the data and the columns we decided that we were deleting some columns that had the same value in here. This value was 0 and represented no energy usage all over the year of 2018 for all the houses.

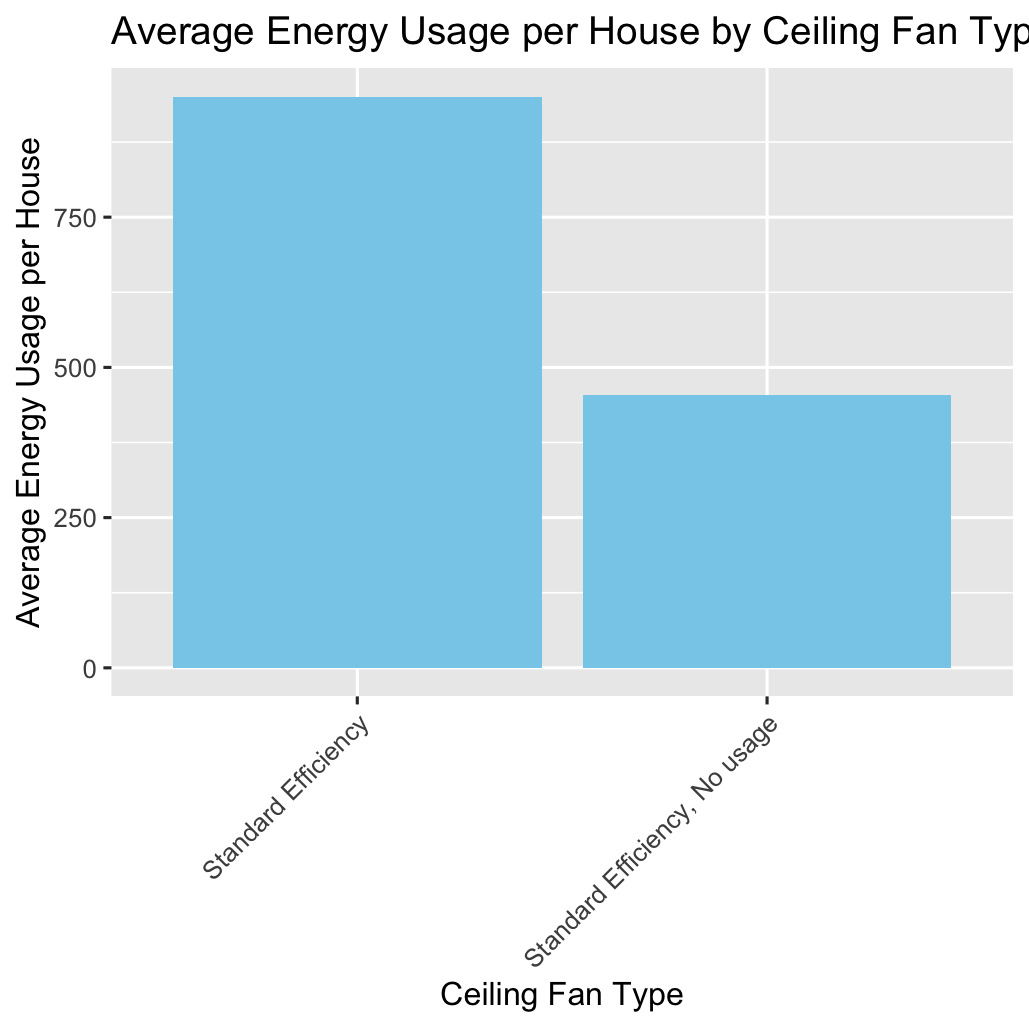
As we did with our other dataset as well, We decided to add an extra column that would sum up all the data except for the time column. We called this extra column sums\_of\_rows. We filtered the data for this dataset based on the month July as July is typically the highest energy usage month.

As above we took a look at the influence of seasonality, square foot and humidity. We wanted to get a better understanding of the energy usage of applications that are used in the houses. We decided to take a better look at the influence of the type of clothes washer, type of clothes dryer and type of ceiling fan.

We wanted to take these factors as we saw that these might be the big energy drivers within the houses in our dataset. To get a more visual view of the data we decided to make some plots to see in what way the different applications can be energy drivers.

We did this by grouping the different types of clothes washers, clothes dryers and ceiling fans to show what the influence is on the average energy usage by houses in County\_ G4500730.

Below is the barchart for average energy usage per house by ceiling fan type.

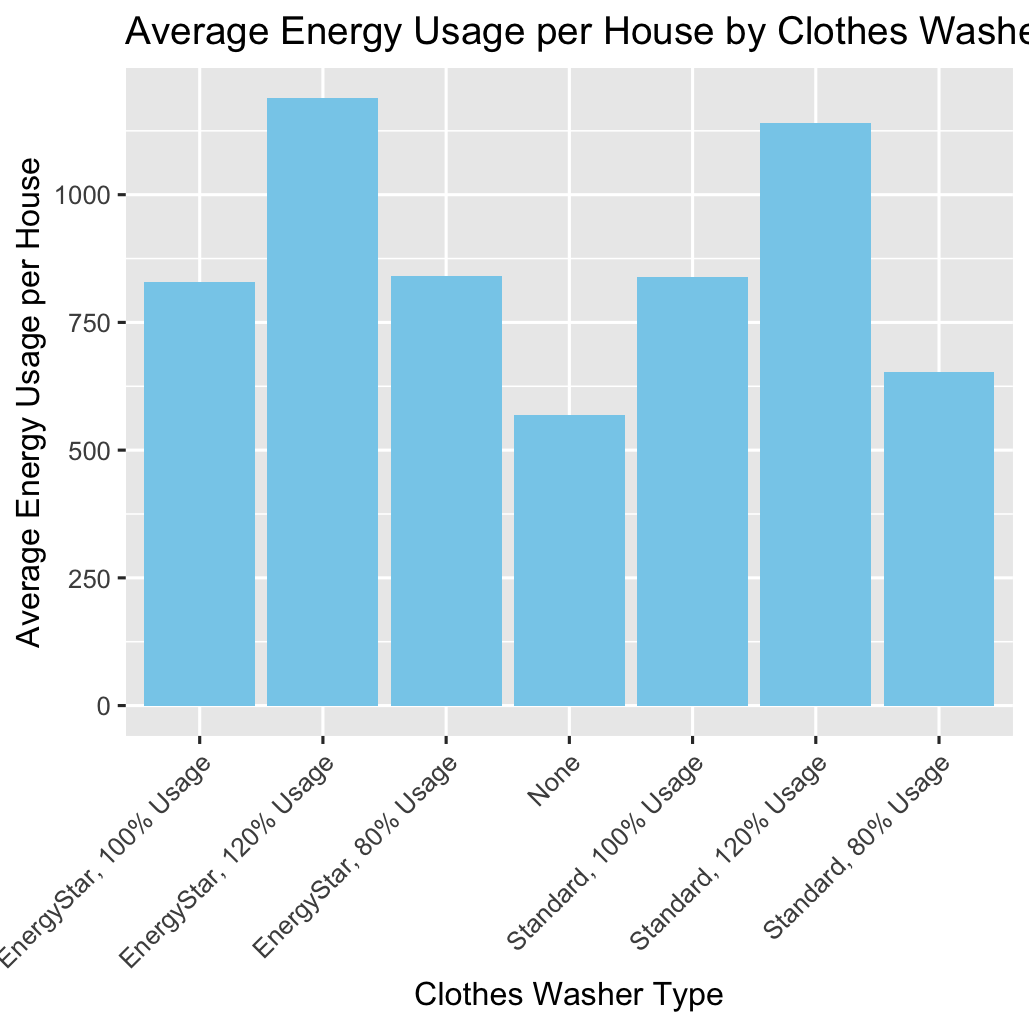


First taking the type of ceiling fan. We grouped the ceiling fans with standard efficiency together and the houses that had the standard efficiency and had no usage/didn't have a ceiling fan together. These were the results from our barplot. We can say that ceiling fans have an influence on the overall usage in the houses. So this is a variable that can have a more than average effect on the total energy consumption

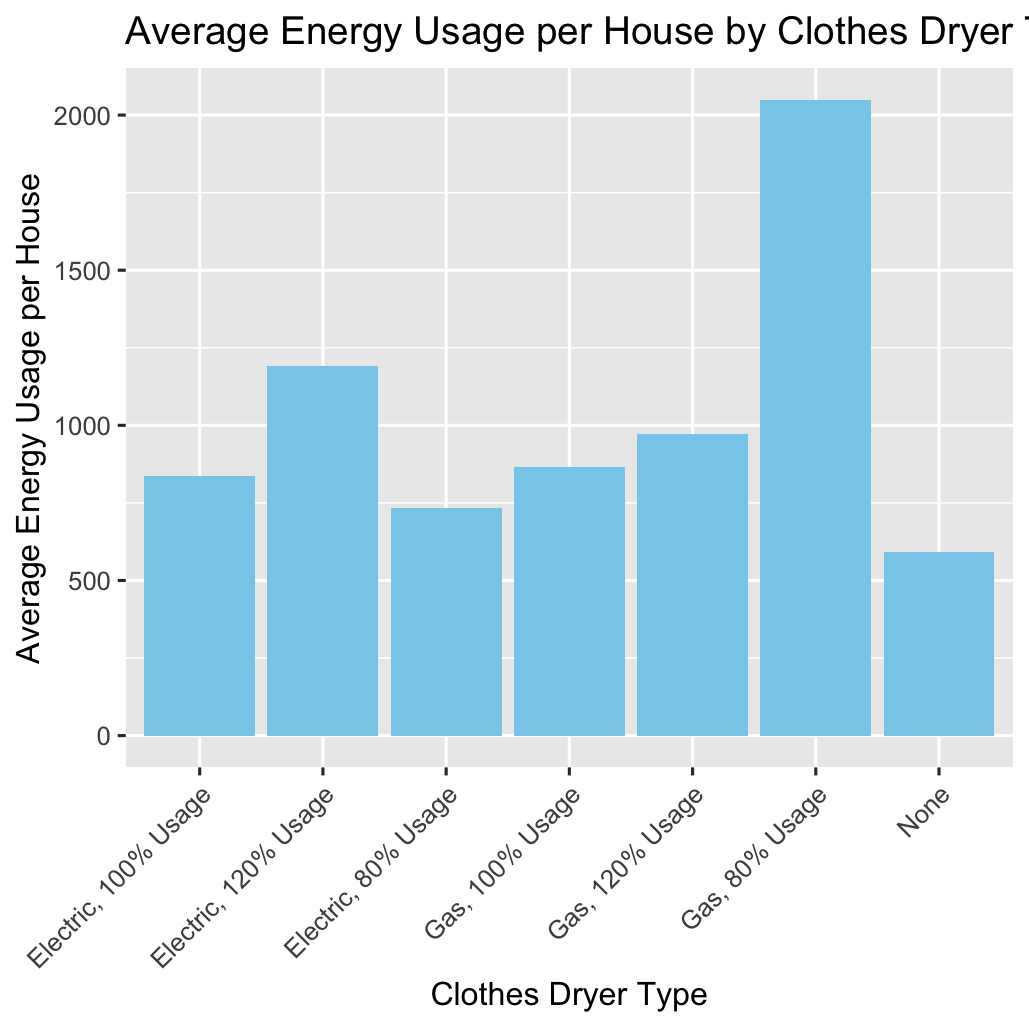
A graph showing the amount of energy consumption

Description automatically generated

after we looked at the clothes washers. This barplot was a little less representative. There are still some differences but they are more equal based on the type of usage. You see that the 120% usage are the 2 highest numbers so these might have an above average effect on the total energy usage.



Lastly we looked at the energy usage per house based on clothes dryers. You see that gas 80% usage is the highest one, but looking at the gas 120% usage is lower so this bar plot gives mixed results.



Overall this gave us some insights and let us look more into the type of ceiling fan related to the total energy usage in the houses.

We used str(), summary(), bar plots, line plots and scatter plots to get to know the data better and see what influences different factors outside and inside the houses would give us. After getting to know our data and getting more valuable insights we decided to start with the modeling. We tried different approaches to find out the best way to predict the energy usage and make sure we picked the best model for that.

First we made the correlation matrix. This correlation matrix shows what coefficients have a big effect on our response variable (total\_consumption) which stands for total energy usage in the houses.

In the correlation table below, the key drivers of energy usage based on energy consumption are highlighted for the ones that are greater than 0.3.

|  |  |
| --- | --- |
| **Cor(July\_consumption)** | **total\_consumption** |
| out.electricity.ceiling\_fan.energy\_consumption | 0.221107848 |
| out.electricity.clothes\_dryer.energy\_consumption | 0.40694943 |
| out.electricity.clothes\_washer.energy\_consumption | 0.204986703 |
| out.electricity.cooling\_fans\_pumps.energy\_consumption | 0.744517072 |
| out.electricity.cooling.energy\_consumption | 0.76993103 |
| out.electricity.dishwasher.energy\_consumption | 0.139228735 |
| out.electricity.freezer.energy\_consumption | 0.057435547 |
| out.electricity.heating\_fans\_pumps.energy\_consumption | -0.006910707 |
| out.electricity.heating.energy\_consumption | -0.007987264 |
| out.electricity.hot\_tub\_heater.energy\_consumption | 0.138751218 |
| out.electricity.hot\_tub\_pump.energy\_consumption | 0.152027386 |
| out.electricity.hot\_water.energy\_consumption | 0.266716261 |
| out.electricity.lighting\_exterior.energy\_consumption | 0.387483724 |
| out.electricity.lighting\_garage.energy\_consumption | 0.244223518 |
| out.electricity.lighting\_interior.energy\_consumption | 0.52795176 |
| out.electricity.mech\_vent.energy\_consumption | 0.06661666 |
| out.electricity.plug\_loads.energy\_consumption | 0.53428475 |
| out.electricity.pool\_heater.energy\_consumption | 0.088107126 |
| out.electricity.pool\_pump.energy\_consumption | 0.189573925 |
| out.electricity.pv.energy\_consumption | 0.09397985 |
| out.electricity.range\_oven.energy\_consumption | 0.337091386 |
| out.electricity.refrigerator.energy\_consumption | 0.127609235 |
| out.electricity.well\_pump.energy\_consumption | 0.150134567 |
| out.natural\_gas.fireplace.energy\_consumption | 0.044533751 |
| out.natural\_gas.grill.energy\_consumption | 0.059307666 |
| out.natural\_gas.hot\_tub\_heater.energy\_consumption | 0.063435251 |
| out.natural\_gas.lighting.energy\_consumption | 0.031857255 |
| out.natural\_gas.pool\_heater.energy\_consumption | 0.056240127 |
| total\_consumption | 1 |

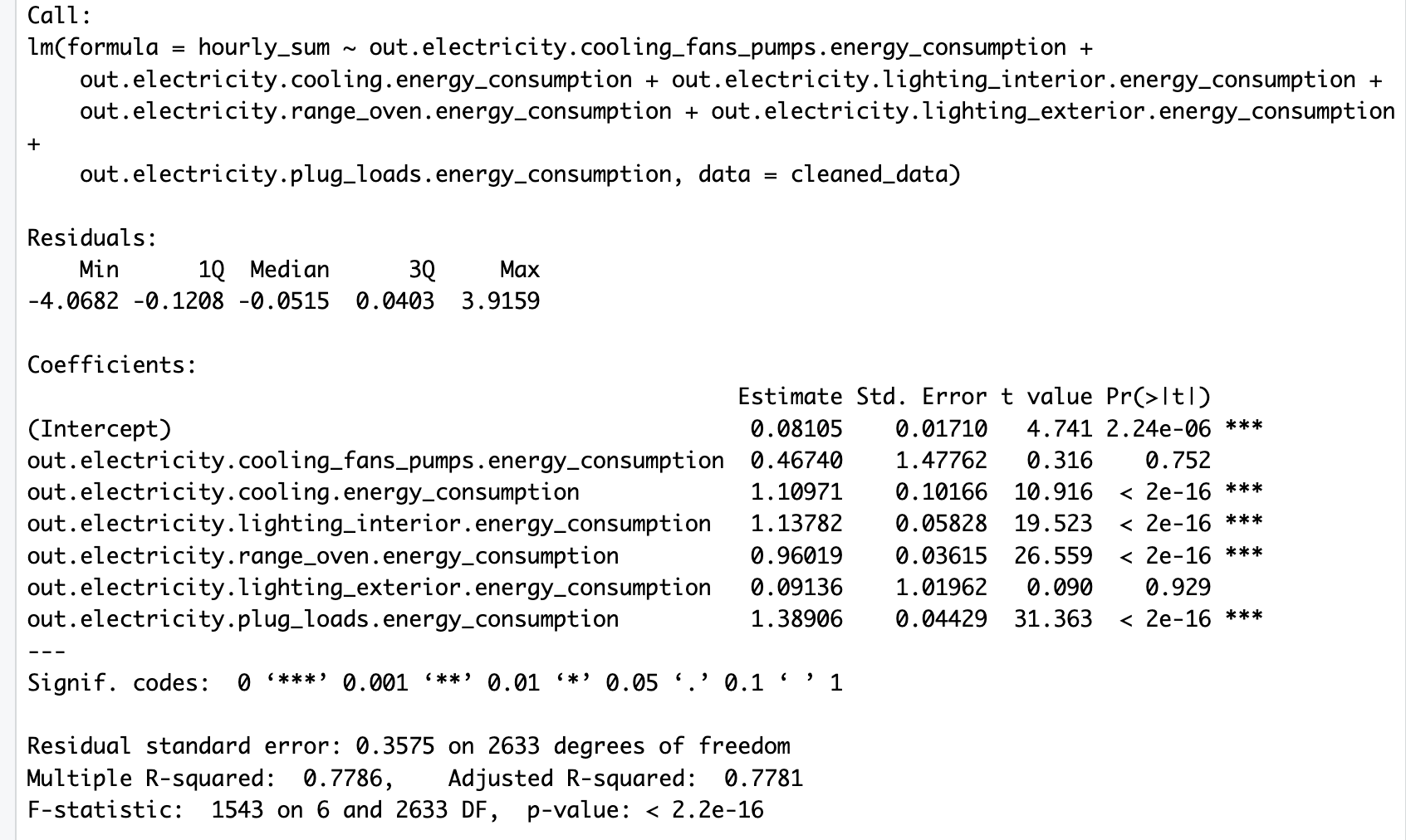
First we made a lm model with the 3 variables that had the biggest correlation. These were:

out.electricity.cooling\_fans\_pumps.energy\_consumption + out.electricity.cooling.energy\_consumption + out.electricity.lighting\_interior.energy\_consumption

Our R-sqaured value for this lm model was:

"R-squared: 0.626260230363969"

The value was not high enough in our opinion so we decided to run a new model that had the high correlation values in there but also the moderate correlation values in there:



This model gave us a R-squared of 0.7781 which was better than the model we used previously. You see that there are 2 columns that are not significant, these are out.electricity.cooling\_fans\_pumps.energy\_consumption and out.electricity.lighting\_exterior.energy\_consumption. With p values of 0.0687 and 0.547. We ran other models to see what would give us the highest p value but we decided to use the lm\_model\_updated because after trying all other variants this was the model that gave us the highest values

R-squared: 0.7786257

Adjusted R-squared: 0.7781212

Root Mean Squared Error (RMSE): 0.3569874

Mean Absolute Error (MAE): 0.163867

Mean Absolute Percentage Error (MAPE): 12.03538 %

R-squared (R²): This value, 0.7786257, indicates the proportion of the variance in the dependent variable (the variable you're trying to predict) that is predictable from the independent variables (the variables used for prediction) in the model. An R-squared of 0.7786 suggests that approximately 77.86% of the variance in the dependent variable is predictable from the independent variables included in the model. It's a measure of how well the model fits the data.

Adjusted R-squared: This is a modified version of R-squared that adjusts for the number of predictors in the model. It penalizes the addition of unnecessary predictors. An adjusted R-squared of 0.7781212 suggests that the model still explains approximately 77.81% of the variance in the dependent variable after adjusting for the number of predictors.

Root Mean Squared Error (RMSE): RMSE is a measure of the differences between values predicted by the model and the observed values. It represents the square root of the mean of the squared differences between predicted and observed values. In your case, the RMSE is 0.3569874, which means, on average, the model's predictions are off by approximately 0.357 units.

Mean Absolute Error (MAE): MAE is another measure of the errors between predicted and observed values. It represents the average of the absolute differences between predicted and observed values. In your case, the MAE is 0.163867, indicating that, on average, the model's predictions are off by approximately 0.164 units.

Mean Absolute Percentage Error (MAPE): MAPE measures the percentage difference between predicted and observed values relative to the observed values. In your case, the MAPE is 12.03538%, meaning, on average, the model's predictions are off by approximately 12.04% relative to the observed values.

These are some values that show the fit of our model. After doing these calculations we remembered that this is time series data. So a linear regression model might not be the best fit. We decided to go along and make a different model.

After using the lm regression model we decided to make a SVM model. We used the same independent variables to predict the dependent variable: energy consumption. We used this to create a SVM model and a KSVM model as well.

We decided to go with a SVM and KSVM model. We did this because overall SVM and KSVM models are better in complex relationships. As our project involves analyzing complex relationships between various factors such as energy consumption, weather conditions, housing characteristics, and appliance usage. SVM and KSVM models are more flexible in capturing non-linear relationships compared to linear regression models, making them better suited for our project. They also tend to perform better in predictive accuracy when dealing with complex and high-dimensional datasets. Given the diverse range of factors influencing energy consumption in our project, a model that can capture these nuances effectively would be advantageous.

**This is the code for the model**

# Load required library

library(e1071)

# Set seed for reproducibility

set.seed(123)

# Split the data into training and testing sets (80% train, 20% test)

train\_index <- sample(1:nrow(cleaned\_data), 0.8 \* nrow(cleaned\_data))

train\_data <- cleaned\_data[train\_index, ]

test\_data <- cleaned\_data[-train\_index, ]

# Define the formula for the regression model

formula <- as.formula(paste("hourly\_sum ~", paste(variables\_of\_interest, collapse = " + ")))

# Train the SVM regression model on the training set

svm\_model <- svm(formula, data = train\_data, type = "eps-regression", kernel = "radial")

#train the ksvm regression model on the training set

ksvm\_model <- ksvm(formula, data = train\_data, type = "eps-svr", kernel = "rbfdot")

#predictions on the testing set

predictions <- predict(ksvm\_model, newdata = test\_data)

# Make predictions on the testing set

predictions <- predict(svm\_model, newdata = test\_data)

Comparing the performance metrics of the SVM and KSVM models:

SVM Model:

* Root Mean Squared Error (RMSE): 0.350
* Mean Absolute Error (MAE): 0.129
* R-squared (R²) coefficient: 0.806

KSVM Model:

* Root Mean Squared Error (RMSE): 0.411
* Mean Absolute Error (MAE): 0.157
* R-squared (R²) coefficient: 0.734

Based on these metrics:

RMSE: The SVM model has a lower RMSE (0.350) compared to the KSVM model (0.411). A lower RMSE indicates better accuracy in prediction, suggesting that the SVM model performs better in terms of predicting the dependent variable.

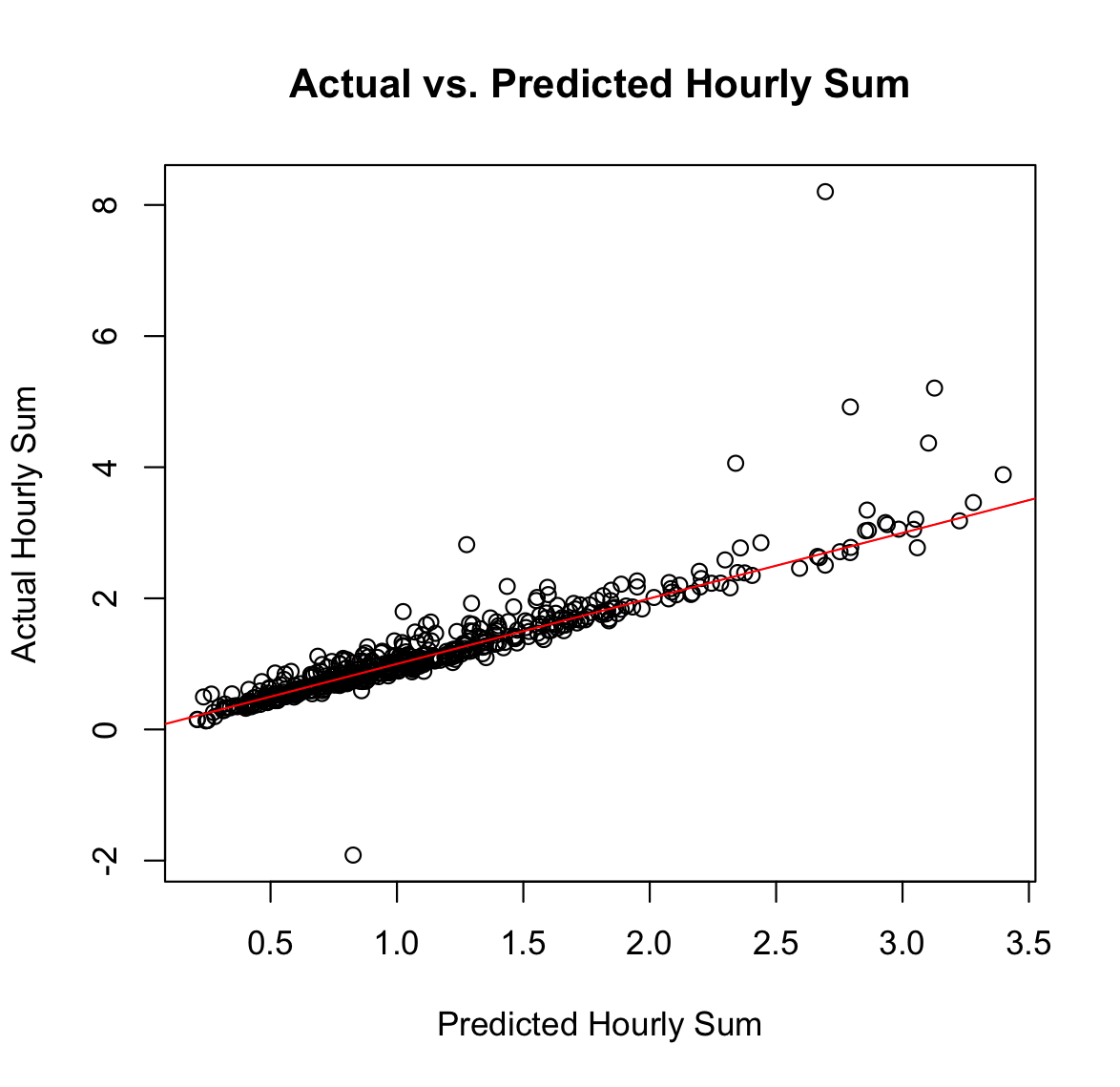
MAE: Similarly, the SVM model has a lower MAE (0.129) compared to the KSVM model (0.157). A lower MAE indicates better accuracy in prediction on average.

R-squared (R²) coefficient: The SVM model also has a higher R-squared value (0.806) compared to the KSVM model (0.734). A higher R-squared value indicates that a larger proportion of the variance in the dependent variable is explained by the independent variables in the model.

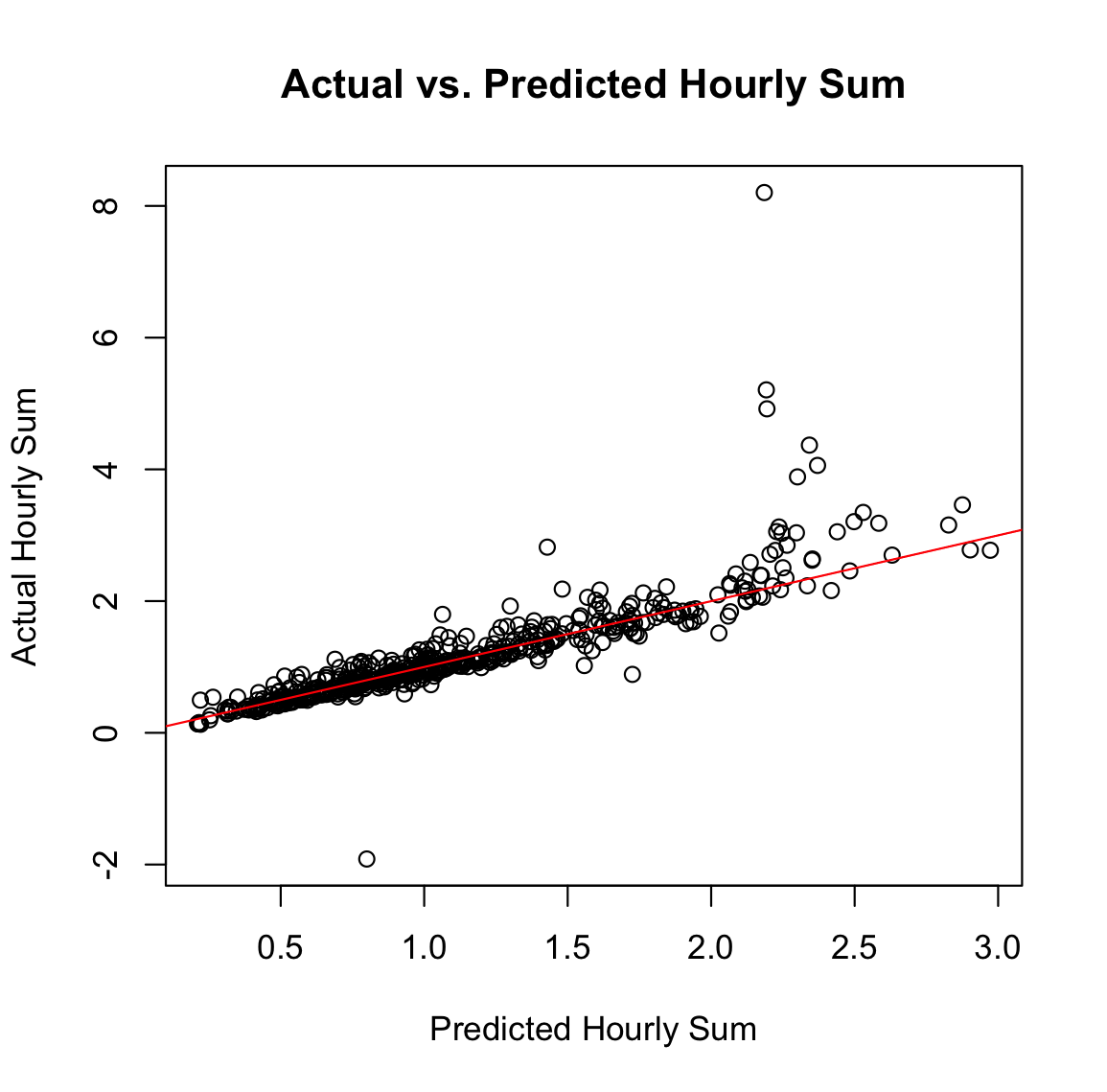
Based on these results, it seems that the SVM model outperforms the KSVM model in terms of prediction accuracy and explaining the variance in the dependent variable. We think accuracy and interpretability are important considerations what makes us believe that the SVM model is preferred over the KSVM mode

We decided to run our models with a test and training set, to see how good they are in predicting our dependent variable ( total energy consumption). We visualized our results into a plot:

**#based on our svm model**



**#based on our ksvm model**



After running these SVM and KSVM models we were happy with our r-squared value. We decided to run the model on different parts of our dataset as well to see in what way it would work on that and to make sure we were not overfitting or underfitting our model.

So we ran our KSVM model and SVM model based on a sample of 1000 houses and their total energy consumption of the 4 counties with the highest energy usage. We did this by creating a training and testing dataset based on our sample. The counties that we used were:

high\_consumption\_county<- c("G4500790", "G4500450", "G4500190", "G4500630")

A graph of a number of people

Description automatically generated

And these were our results for our models:

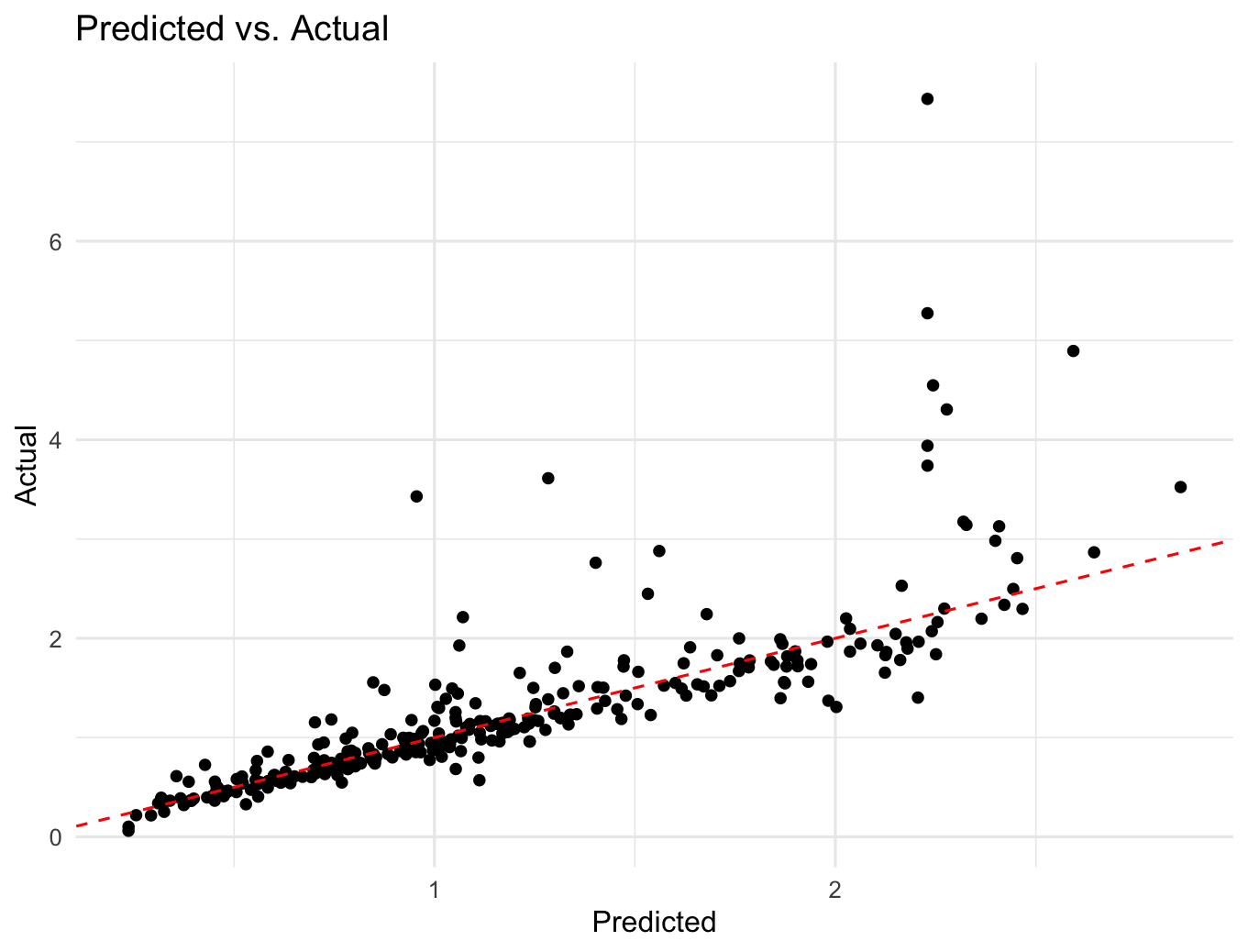


Based on our KSVM model

Mean Absolute Error (MAE): 0.234187

Root Mean Squared Error (RMSE): 0.5396112

R-squared (Coefficient of Determination): 0.6340126

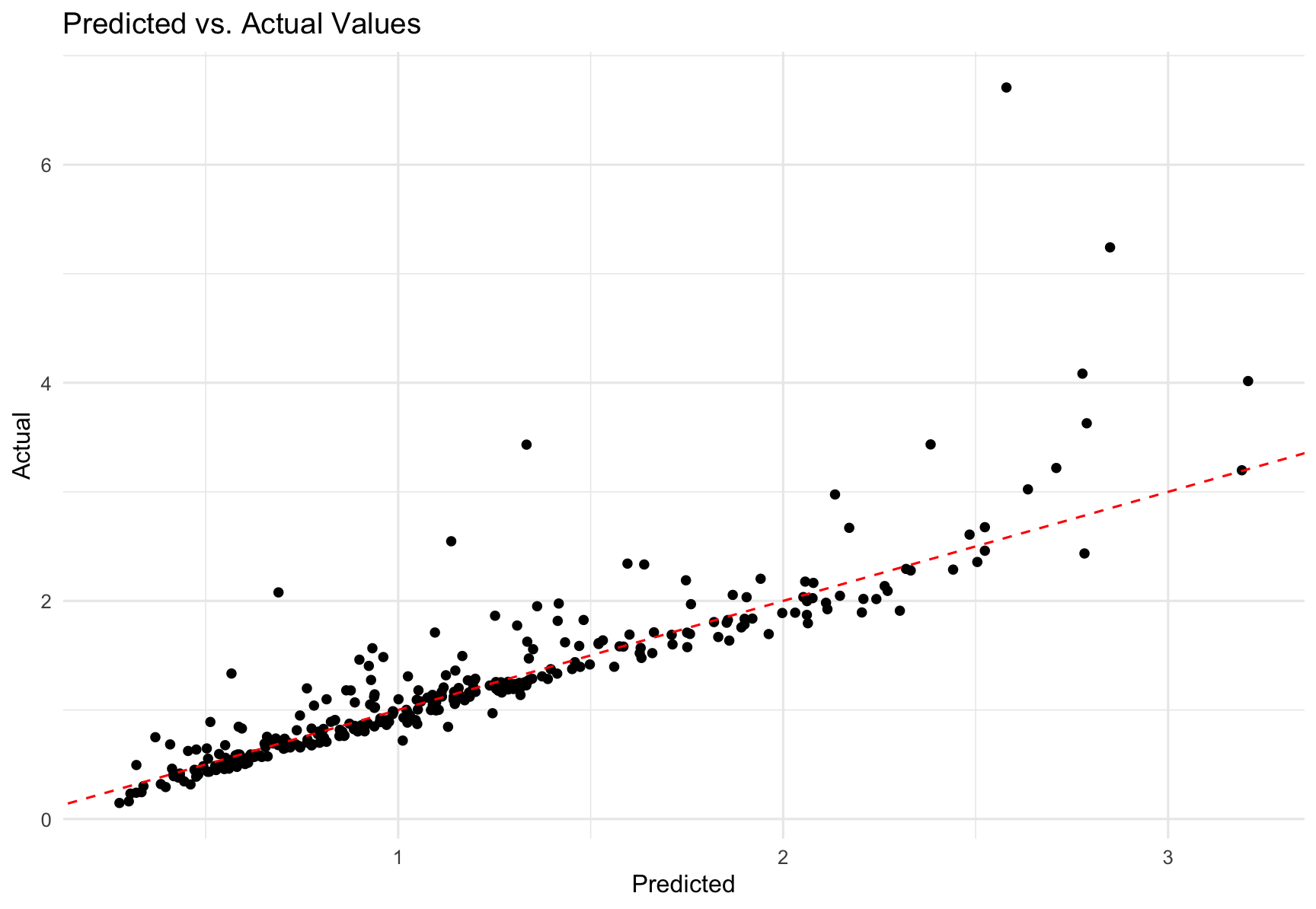


Based on our SVM model

"Mean Absolute Error (MAE): 0.175172808030973"

"Root Mean Squared Error (RMSE): 0.392067037877588"

"R-squared (R²) coefficient: 0.768781504770023"



We can say that our KSVM model and SVM model based on these variables can predict moderately well what the energy consumption will be for the different counties.

Variables of interest:

"out.electricity.cooling\_fans\_pumps.energy\_consumption", "out.electricity.cooling.energy\_consumption",

"out.electricity.lighting\_interior.energy\_consumption",

"out.electricity.range\_oven.energy\_consumption",

"out.electricity.lighting\_exterior.energy\_consumption",

"out.electricity.plug\_loads.energy\_consumption",

What we got out of this is that these are the main drivers of energy consumption within the dataset. We can use this model to evaluate peak future energy demand.

First looking at what these 6 columns mean

* Out.electricity.cooling\_fans\_pumps.energy\_consumption: Electric energy consumed by cooling fans and pumps
* Out.electricity.cooling.energy\_consumption: Electric energy consumed by electric cooling systems
* Out.electricity.lighting\_interior.energy\_consumption: Electric energy consumed by interior lighting
* out.electricity.range\_oven.energy\_consumption:Electric energy consumed by cooking range and oven
* out.electricity.lighting\_exterior.energy\_consumption:Electric energy consumed by exterior lighting
* Out.electricity.plug\_loads.energy\_consumption: Electric energy consumed by plug loads

To evaluate peak future energy demand, we think it is important for ESC to look into counties that have a hot-humid climate and have an above average consumption for these 6 attributes. As we saw above as well, the 21 hour in the day is the hour where most energy is being used. This is 9pm at night.

We grabbed the data of the 4 counties with the highest energy usage, these are:

high\_consumption\_county<- c("G4500790", "G4500450", "G4500190", "G4500630")

Looking at the first column, Out.electricity.cooling\_fans\_pumps.energy\_consumption there are 4 different variables in our static housing data set that we think might influence this.. They are "in.cooling\_setpoint",

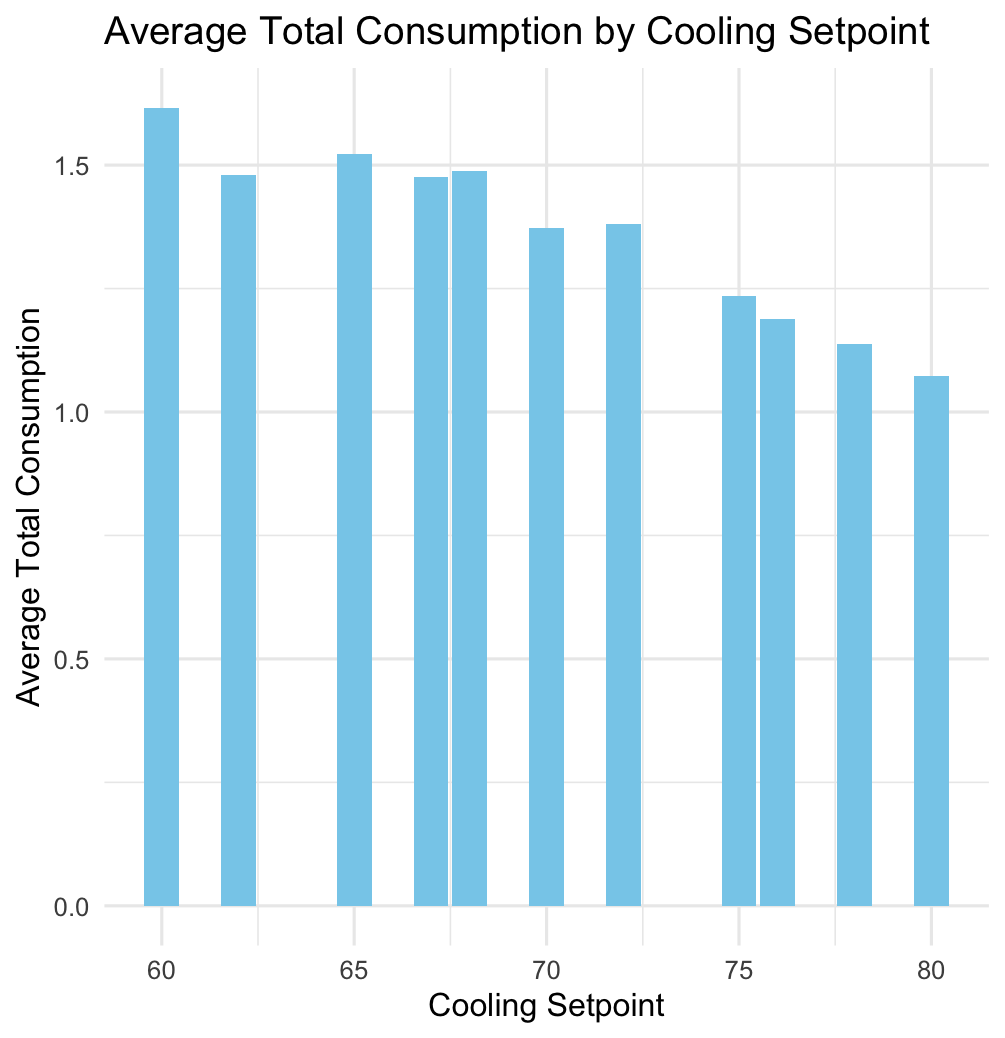
"in.cooling\_setpoint\_has\_offset",

"in.cooling\_setpoint\_offset\_magnitude",

"in.cooling\_setpoint\_offset\_period

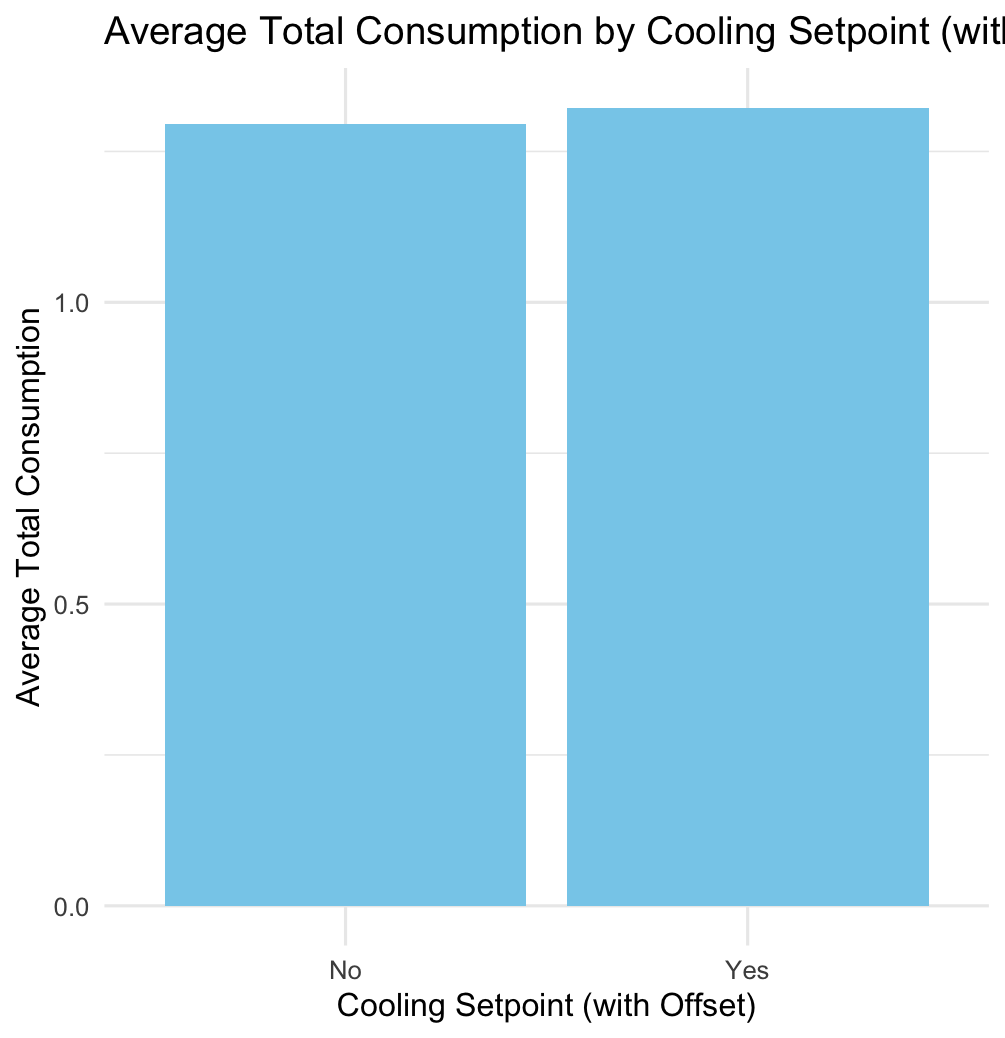
We made a barcharts to see what had the most influence on Out.electricity.cooling\_fans\_pumps.energy\_consumption.

First looking at in.cooling\_setpoint:



You see that how lower the cooling setpoint is, the more energy is being used.

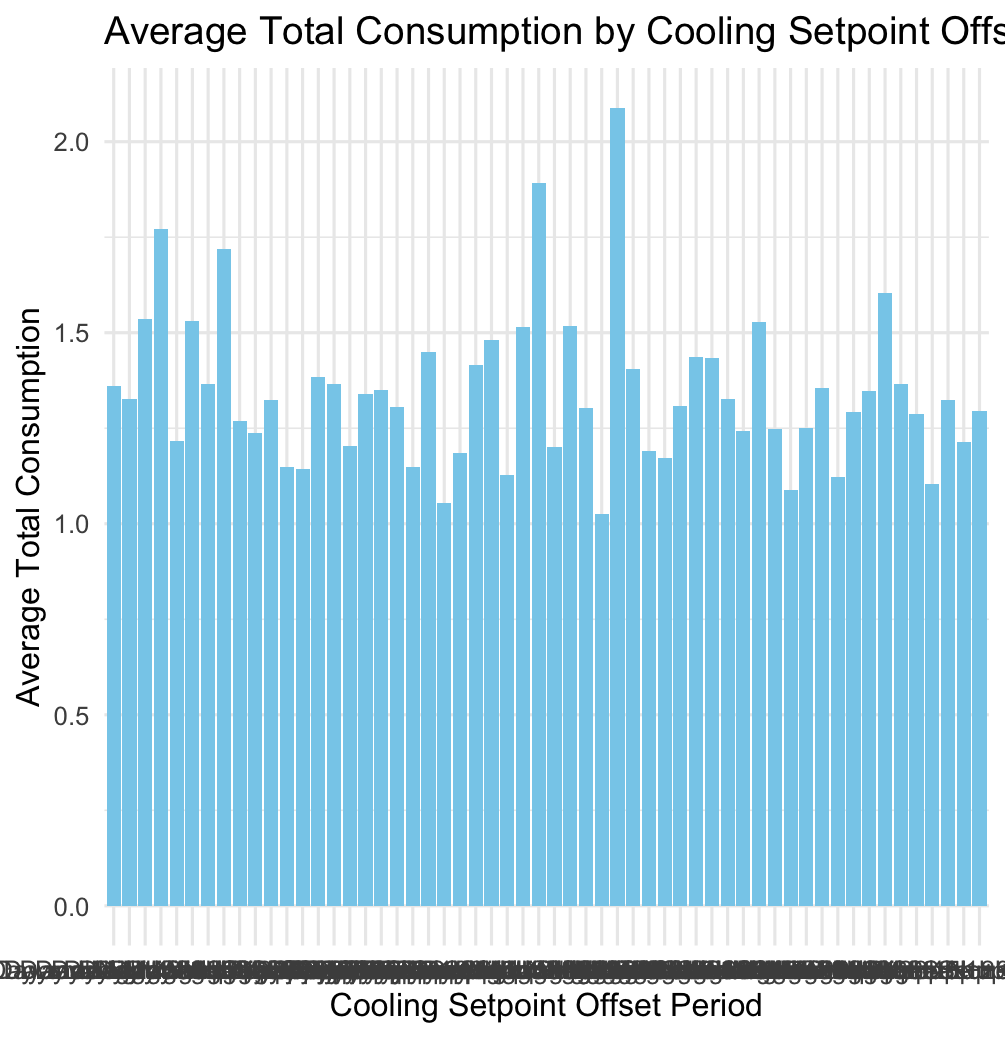
Looking at in.cooling\_setpoint\_has\_offset, you see that there is almost no difference between the answer yes and no.



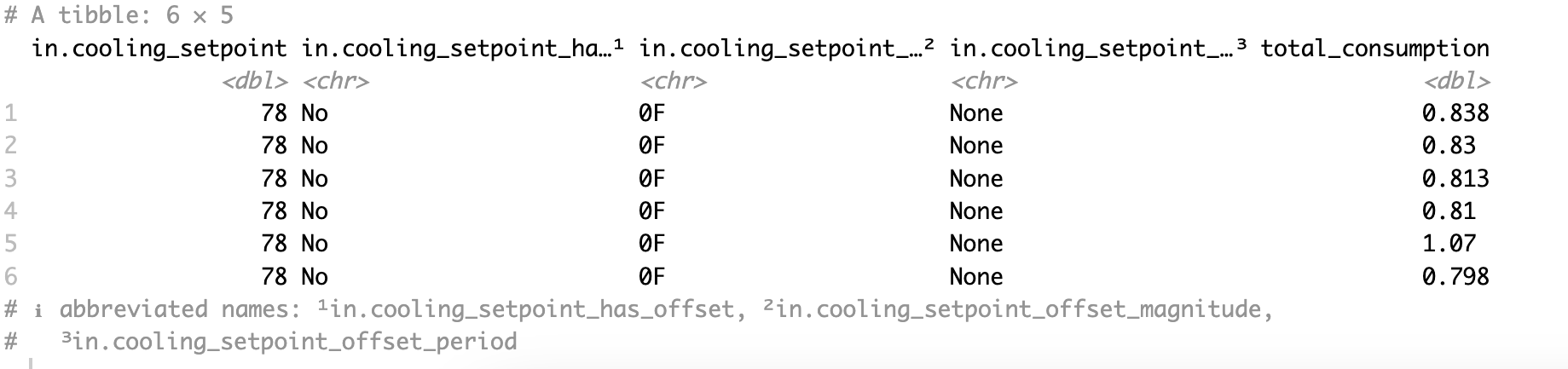
Looking at in.cooling\_setpoint\_offset\_magnitude, you see that 5F is a little higher than the other ones, but not not a big difference.



For in.cooling\_setpoint\_offset\_period you see that the data is very different across the whole barchart.



We decided to show the head to see the highest values to give a better presentation of what the main drivers are of the cooling setpoint offset period:



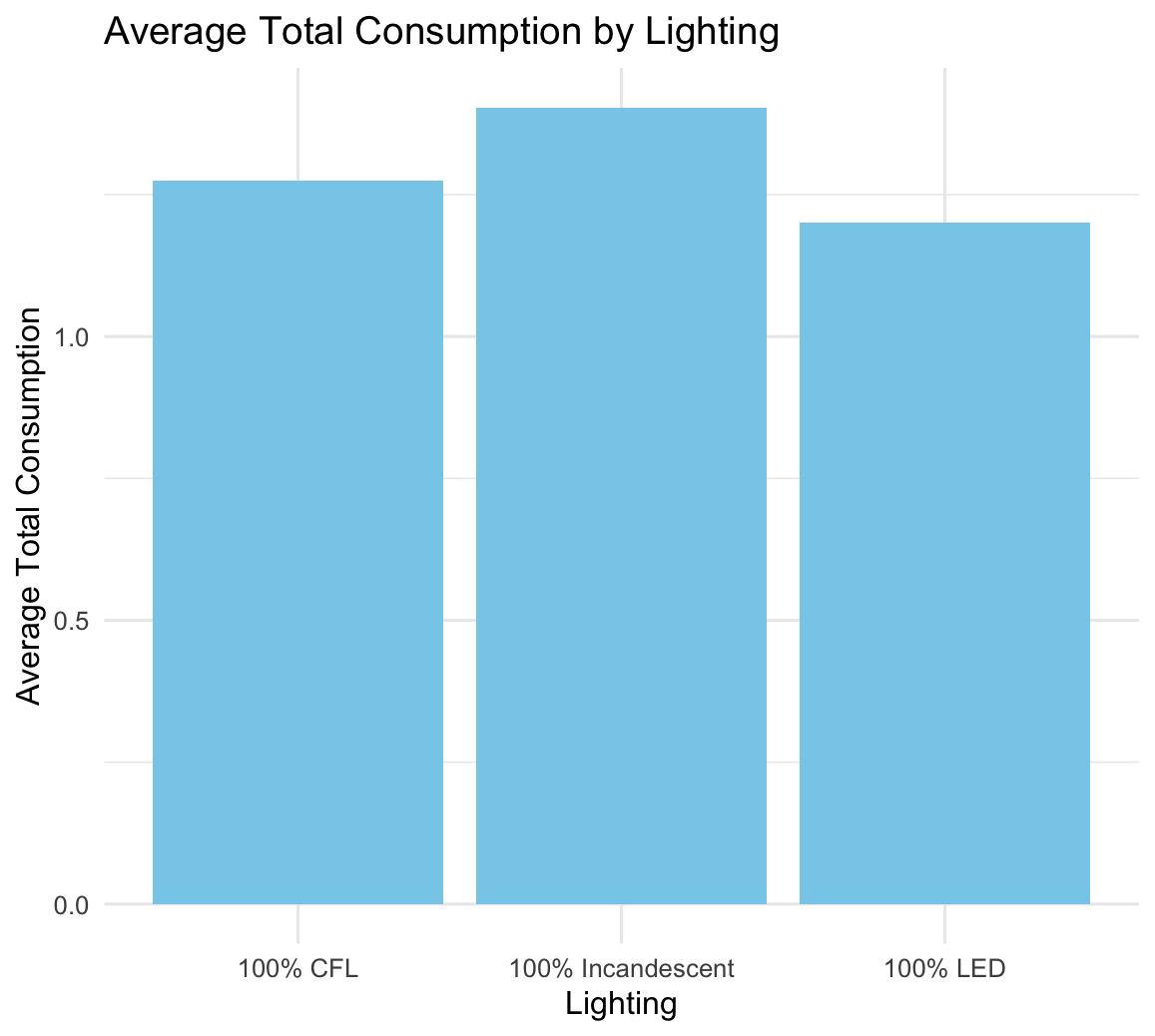
We also looked at energy consumption based on cooling efficiency, plot is listed below:

A graph of energy consumption

Description automatically generated

So for our first column that was out.electricity.cooling\_fans\_pumps.energy\_consumption we see that in.cooling\_setpoint had an influence on total energy consumption. We saw that the colder the in.cooling\_setpoint was set, the higher the average energy usage was.

Then we looked at the column for lighting. This was in.lighting:



We see that 100% let has the least energy usage based on the energy consumption.

After we decided to take a closer look at the plug\_loads column

A graph showing a number of plugs

Description automatically generated

Energy consumption by Plug Load graph shows that houses with 100%plug load diversity have the highest energy consumption while houses with 50% plug load have the lowest energy consumption.

We wanted to try a different approach and see the influence of the heating and cooling factors on total energy consumption in residential buildings. We created two datasets: one for heating and another for cooling.

First, we looked at houses in counties that had a hot and humid climate. We chose houses with three bedrooms to see if that would be a significant factor. Then, we blended information about these houses with data on their energy usage and the weather conditions they face.

Our goal was to understand how variables like temperature and electricity usage for heating are linked to the overall energy consumption of these houses. Through correlation analysis, we saw that there were some connections. Armed with this insight, we experimented with two predictive models: linear regression and decision trees.

Similarly, for the cooling dataset, the same preprocessing steps were followed. Strong correlations were found between total energy consumption and electricity usage for cooling-related purposes. Again, both linear regression and decision tree models were built. After doing all the analysis we got an r-squared of 1, what told us that the model was overfitted on the data. We decided to not continue with this business question as the results were not what we hoped.

Even Though this process didn’t go exactly as planned, it still gave us some valuable insights and made us decide to take it from a different approach. So far all our models mostly took only energy usage in consideration. But we wanted to create a model that could help out ESC more. We decided to create a model that looked not only at the energy data but took other factors in consideration as well.

As above we saw that in\_sqft and humidity had some correlation with the total energy usage. We wanted these in our model to predict based on more than just the energy usage. First we decided to pick the variable in\_hvac\_cooling\_type. This variable shows the presence and type of cooling system based on central AC, Heat pump, none or room ac. We saw in our barchart above that these have a big influence on the total energy consumption. And as cooling energy were two out of the 6 big drivers of our models, we decided that this would be a good predictor to add.

We decided we wanted to add Dry\_Bulb\_Temperature\_C in there as well as we wanted to be able to predict if the temperature would get 4 degrees warmer, and this is the variable that shows what the influence of temperature is on the total energy consumption. We added Relative\_Humidity as well, in our analysis above you can see that this has an influence on the energy consumption as well and makes us able to differentiate in what climate we are predicting.

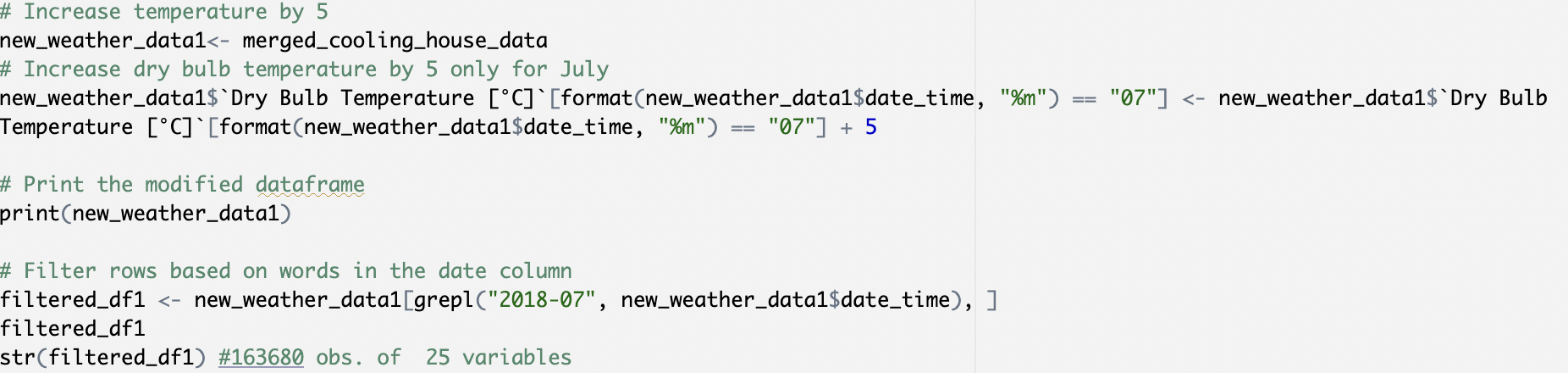
We added the total\_energy column in our prediction as well as this is our dependent variable that will be predicted by these in\_hvac\_cooling\_type, in\_sqft, in\_bedrooms, Dry\_Bulb\_Temperature\_C, Relative\_Humidity, Global\_Horizontal\_Radiation and Direct\_Normal\_Radiation. We decided to add a column for county in there as well to show different geographic locations. We picked these variables based on performing a correlation analysisexploratory data analysis. This made us select the best features that would give us the best model to predict energy usage based on location and weather temperature.

In total we developed 4 models:

|  |  |  |
| --- | --- | --- |
| Model Type | Features Considered | Accuracy |
| Linear Regression | All | 0.5934 |
| Decision Tree | All | 0.6397 |
| Linear Regression | Selected from analysis | 0.6829 |
| **Decision Tree** | **Selected from analysis** | **0.7038** |

We decided to use our decision tree and linear regression in our shiny app.

We also created a dataset with weather of 5 degrees warmer. We did this so we are able to predict when the temperature in july is warmer than it is currently. This is our code for that dataset:



**2 potential approaches to reduce peak energy demand based on lighting and cooling energy:**

First, we wanted to understand peak energy demand periods and we were focusinging on different energy drivers such as lighting and cooling system usage patterns. We identified opportunities for reducing energy usage during these peak periods.

To model the impact, we have established baselines for both lighting-related and cooling-related energy consumption by analyzing data during high-demand times. This will serve as the foundation for the development of energy-saving behavior encouragement programs and smart thermostat programs.

The energy-saving behavior encouragement program incentivized consumers to adopt energy-saving practices, such as turning off lights when not in use, through rebates or discounts on energy-efficient lighting products like LED bulbs as we saw that these are correlated with less energy usage. Simultaneously, the smart thermostat program enabled consumers to remotely control and schedule cooling settings based on their usage patterns and preferences. Smart thermostats optimize energy usage by adjusting temperatures according to occupancy and outdoor weather conditions, and can change the cooling setpoint to a higher level when occupants aren’t home to help in reducing energy usage .

During the pilot programs, we will monitor energy consumption data to assess their effectiveness. By collecting data on energy usage before and after program implementation, we analyzed the impact of both programs on reducing lighting-related and cooling-related energy consumption.

Statistical analysis will be used to quantify the reduction in energy consumption achieved through the programs, calculating metrics such as the percentage decrease in energy usage during peak demand periods. Additionally, we will conduct cost-benefit analyses to evaluate the economic feasibility of scaling up both programs. This involved comparing the costs of implementing the programs with the savings that will be achieved through reduced energy consumption.

Based on the results of the pilot programs, we provided recommendations on the scalability and potential impact of implementing both energy-saving behavior encouragement campaigns and smart thermostat programs across all geographic regions.

**Conclusion**

In conclusion, our project focused on understanding and predicting energy usage patterns in residential properties served by ESC in South Carolina and parts of North Carolina. Through thorough data analysis and modeling, we aimed to identify key drivers of energy consumption and propose strategies to mitigate peak energy demand.

We began by exploring various aspects of the data, including static house information, energy usage data, and weather conditions. This initial exploration led us to formulate business questions and hypotheses regarding factors influencing energy usage, such as insulation, appliance types, weather conditions, and time of day.

Our analysis revealed several important findings. We found that certain appliances and systems, such as cooling systems and interior lighting, were significant drivers of energy consumption. Additionally, we observed patterns in energy usage across different geographic regions and climate zones, with hotter and more humid areas exhibiting higher energy demand.

Using regression modeling techniques, we developed predictive models to estimate energy consumption based on relevant variables. We compared the performance of linear regression and support vector machine (SVM) models, ultimately finding that the SVM model outperformed the others in terms of prediction accuracy.

Furthermore, we conducted a pilot program to encourage energy-saving behaviors among consumers, focusing on reducing energy consumption related to lighting. By analyzing data before and after the program implementation, we quantified the impact of the program on energy usage reduction.

Overall, our project provides valuable insights into the factors influencing residential energy consumption and offers practical recommendations for ESC to manage peak energy demand effectively. By implementing targeted strategies to promote energy efficiency and behavior change, ESC can improve grid reliability, reduce environmental impact, and enhance customer satisfaction.