eSC Energy

Predictions and Recommendations for Future Energy Use

Final Report

IST 687 - M003

Fall 2023

Date: 12/8/2023

Table of Contents:

1. Project Goal…… 3

2. Understanding & Merging the Data…… 3 3. Visualizations…….. 6 4. Modeling……17

a. Time Series……17

b. ANOVA for Variable Choice……20 c. Random Forest……..22 d. Neural Networks……24

e. Linear Regression……24 5. Final Predictions & Future Peak Energy Demand……25 6. Shiny Application……26 7. Recommendations to eSC……28 8. Work Log (Who Did What)……34

Project Goal

For this project our ultimate goal was to provide information to eSC regarding their customer's energy usage in order to understand next year's energy needs. We did this through four different steps. First, we wanted to understand the historical energy usage from the 2018 data we were provided and determine which various factors had the most impact on energy usage trends. Then, we considered what would happen next summer if the temperature was to increase by 5℃ in order to make energy usage predictions. Based on these predictions, we provided a number representing the ultimate grid capacity that eSC would need to ensure their energy grid allowed so that they could account for maximum future energy needs. Finally, we provided recommendations on how to reduce their customer's energy usage so as to not overwhelm their grid.

Understanding & Merging the Data

Now that we have the project goal, our next step was to figure out how to gather and best organize the data. The data provided was split into three different types: static house information on each house surveyed, energy data for each house split by day and hour, and finally weather data for each county also split by day and hour. Since eSC was solely interested in July data, the month with the highest energy usage, we filtered the energy and weather data to only include dates in July. Additionally, in order to understand the *total* energy use at each hour for each house, we added a column that summed each individually calculated energy metric. Therefore, each house now had total energy for each hour in July, and this is the column that we would use for all future predictions. Then, we faced the decision of how to merge all of the data together in a comprehensive and meaningful way.

To merge the data, we ran two different loops in R: one that cycled through bldg_id

(building id, unique for each house in the static house data) in order to merge each house with its equivalent energy data, and another that then cycled through in.county (county id) to merge each house with its equivalent county weather. Below is a screenshot of these two loops in R:

Energy Merge Loop

```r
#####energy#####
#initialize storage for combined energy df
combined_energy = NULL
#loop for energy data
for (i in static_house_info$bldg_id) {
  #create energy url for specific house id
  energy_url = paste0(
    "https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/2023-houseData/",i,".parquet")
  energy = read_parquet(energy_url)
  energy <- energy[energy$time >= as.POSIXct("2018-07-01 00:00:00") &
                     energy$time <= as.POSIXct("2018-07-31 23:00:00"), ] #only keep july dates
  energy$total_energy_usage = rowSums(energy[1:42])
  energy$bldg_id = i
  energy = na.omit(energy)
  if (is.null(combined_energy)) {
    combined_energy = energy
  } else {
    combined_energy = rbind(combined_energy, energy)
  }
}
```

Weather Merge Loop

```r
#####weather#####
#grab unique county codes
county_codes <- unique(house_energy_less_columns$in.county)
#initialize storage for combined_weather df
combined_weather = NULL
#loop for weather data
for (i in county_codes) {
  #create weather url for specific county
  weather_url <- paste0(
    "https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/weather/2023-weather-data/",i,".csv")
  weather = read_csv(weather_url)
  #only keep july dates
  weather = weather[weather$date_time >= as.POSIXct("2018-07-01 00:00:00", tz = "UTC") &
                      weather$date_time <= as.POSIXct("2018-07-31 23:00:00", tz = "UTC"), ]
  weather$in.county = i
  if (is.null(combined_weather)) {
    combined_weather = weather
  } else {
    combined_weather = rbind(combined_weather, weather)
  }
}
```

Within this initial merge, one thing we did not include was any information other than bldg_id and in.county from the static house data. Since static house data contained 171 total columns,

when we attempted to merge all of it together at once it was simply too large to be useful. Instead, we decided to keep our combined weather and energy data mostly separate from the static house data, and only merge necessary columns when we got to that step in the analysis process. This allowed us the flexibility to treat the two datasets as one without physically combining them and

overwhelming our storage. As an example, this is a condensed sample of what our dataset looked like at this stage, with the reminder that we would merge any necessary static house information in using bldg_id at various points in our analysis. There are 744 rows for each bldg_id, with each row holding information on energy and weather for that house for one hour of July.

Initial Merged Dataset Structure

	bldg_id	in.county	datetime	total_energy	*other energy columns*	Temp (°C)	*other weather columns
1	47	G4500010	07-01-18 00:00:00	0.80	…	25	…
2	47	G4500010	01:00:00	1.20	…	26	…
3	47	G4500010	02:00:00	0.43	…	28	…

… *The rest of July rows for house with bldg_id 47*

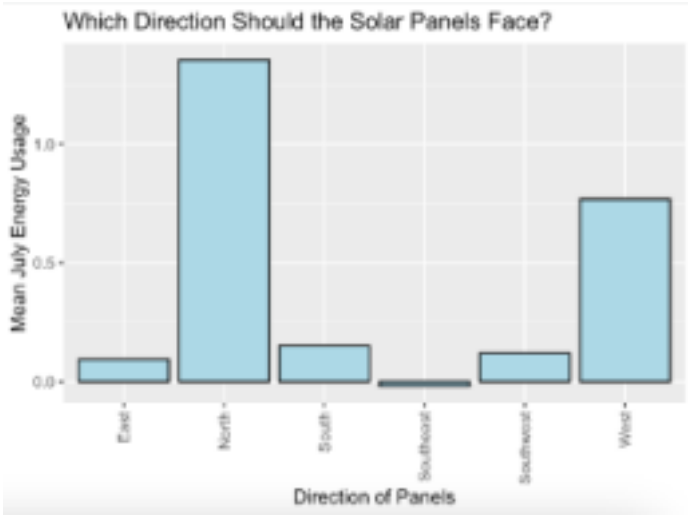
0.96 … 23 …

745 68 G4500045 07-01-18 00:00:00

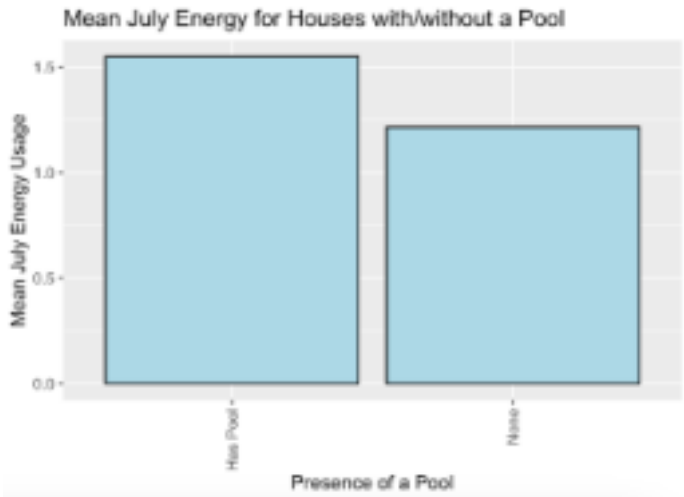
Other necessary steps included creating separate date and time columns from the combined datetime column, as well as making sure the datetime format was correct. Additionally, when it came to running models on the data, we did various aggregation techniques in order to make the

process smoother. This will be explained in more detail in the model section, but for example we aggregated by county for some models and grouped by day for others.

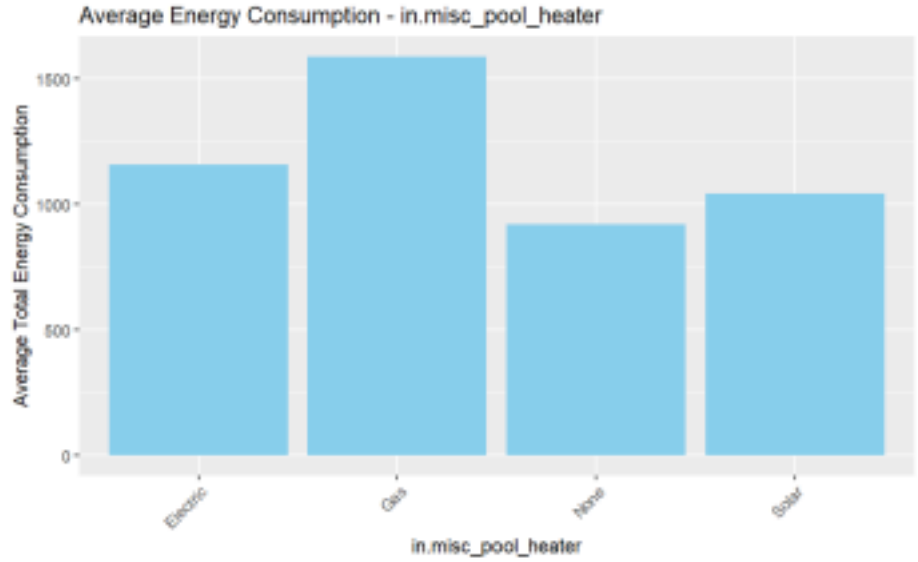
Visualizations



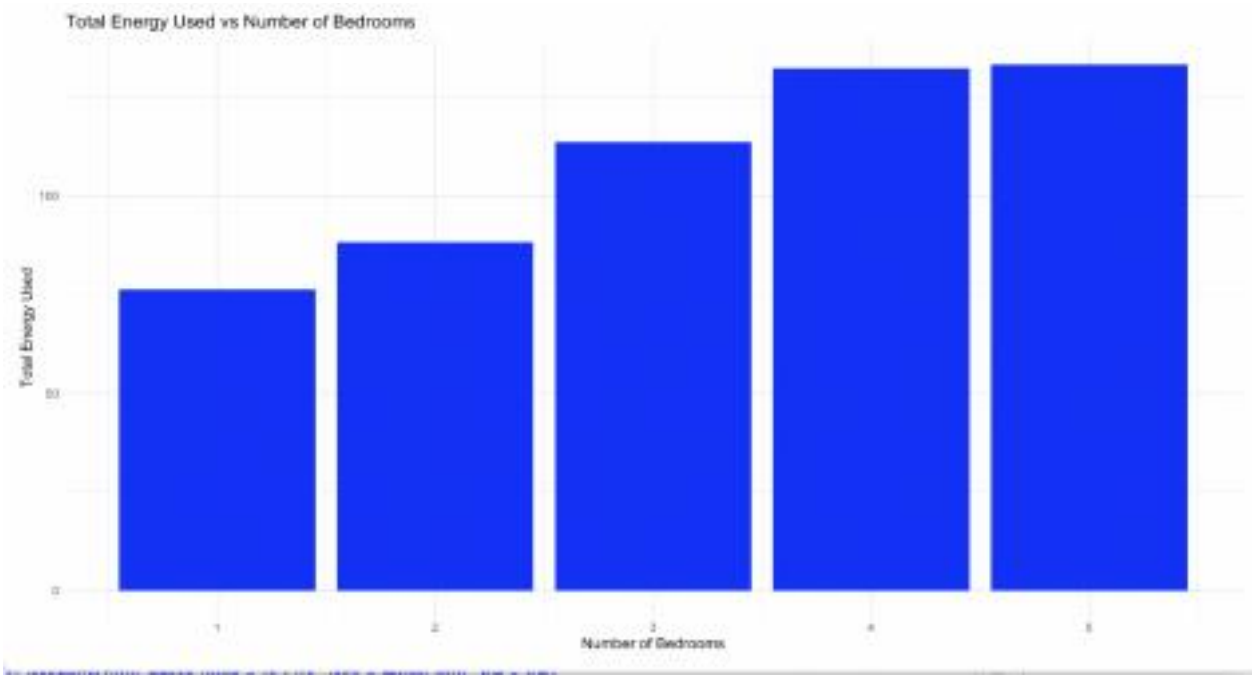
The graph indicates that solar panels facing Southeast are the most efficient, as shown by their lower average energy usage in July. This suggests that Southeast-facing panels require less energy to achieve the desired output, possibly due to optimal sun exposure in this region during the month. It highlights the importance of considering directional orientation when installing solar panels to maximize efficiency.



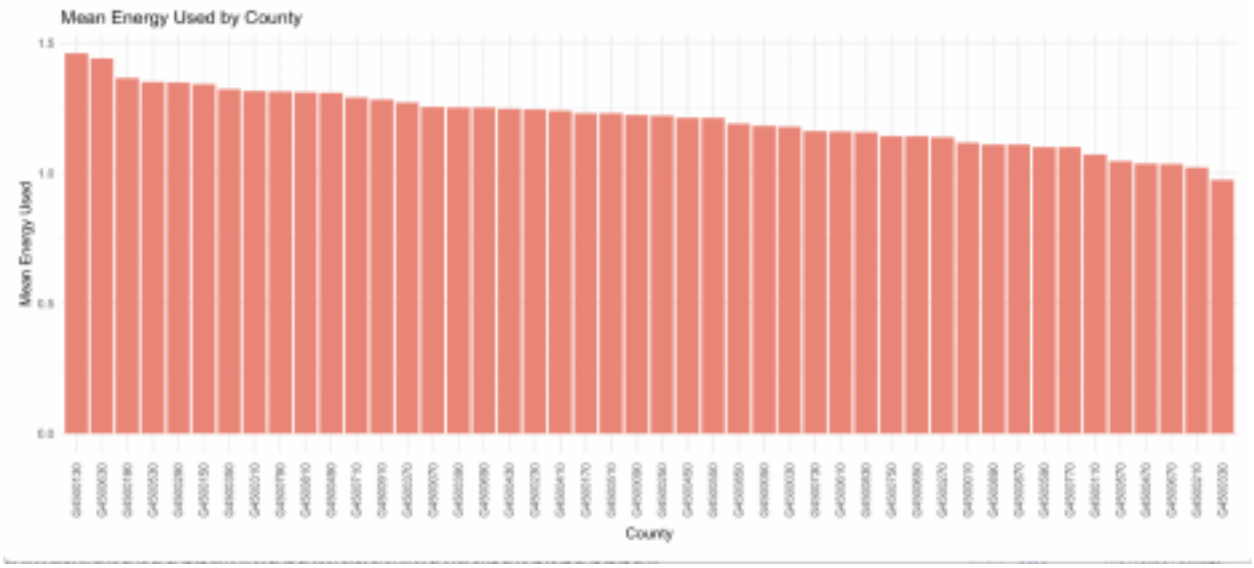
Houses with a pool tend to have higher average energy usage in July compared to houses without a pool. This could indicate that pools, likely due to maintenance and heating or the fuel type used (which we will analyze further), contribute to increased energy consumption during this summer months.



The graph suggests that homes with gas pool heaters have the highest average energy consumption, while those with solar heaters consume the least. Electric and hybrid heaters result in moderate energy usage. This implies that choosing solar pool heaters could be more energy-efficient compared to gas, electric, or hybrid heating options.



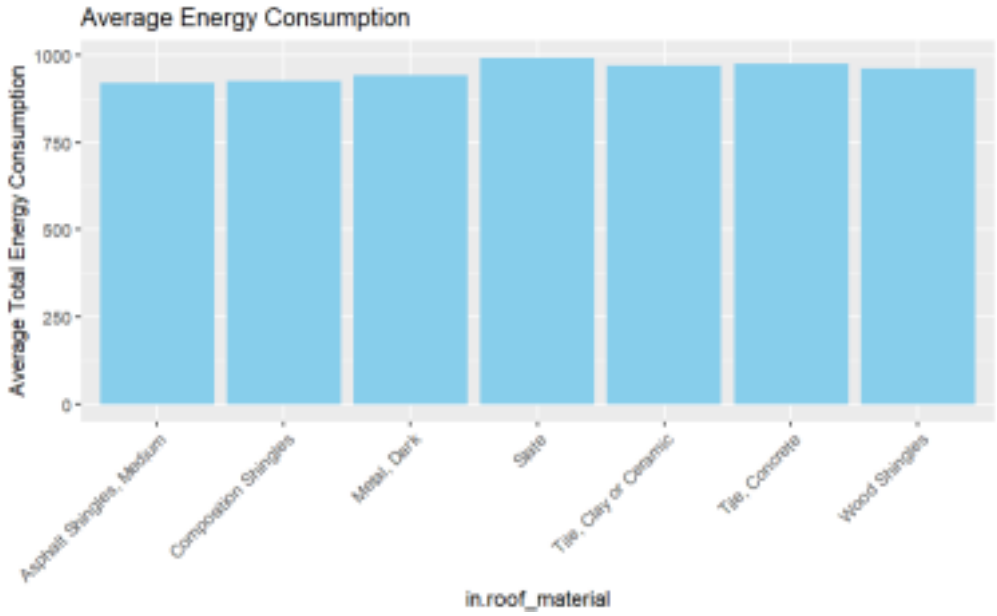
This graph shows that as the number of bedrooms increases, the total energy consumption also increases. This is because more bedrooms typically indicate a larger home size, which often requires more energy for heating, cooling, and lighting. Because of the upward trend, it suggests a positive correlation between house size (as indicated by the number of bedrooms) and total energy consumption.



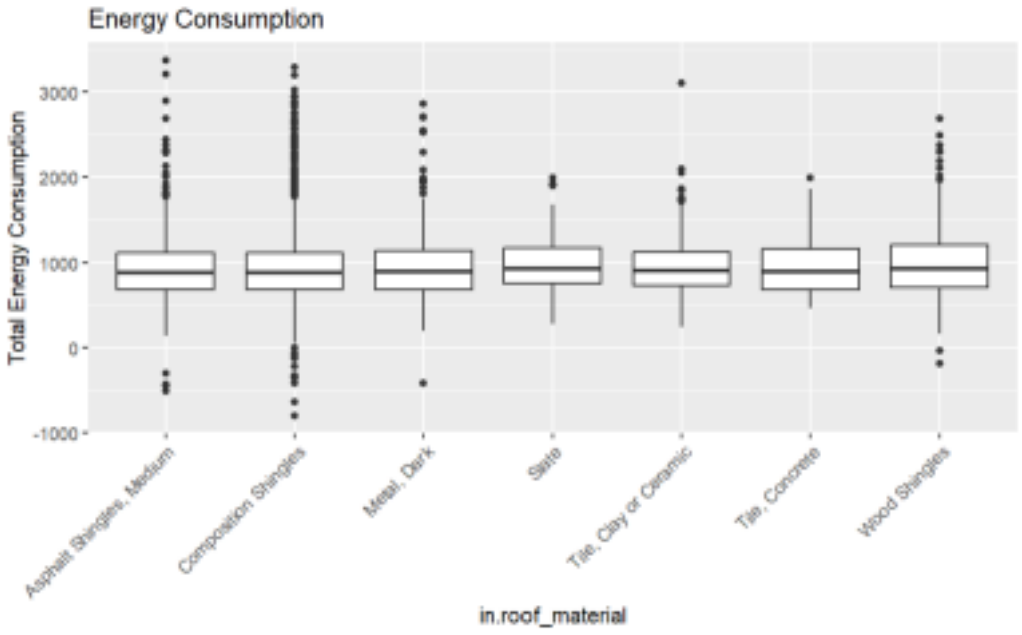
This graph shows the mean energy consumed per county. This highlights the average usage of energy of every single county and it can be derived that the county “G4500130” has the highest average energy usage.

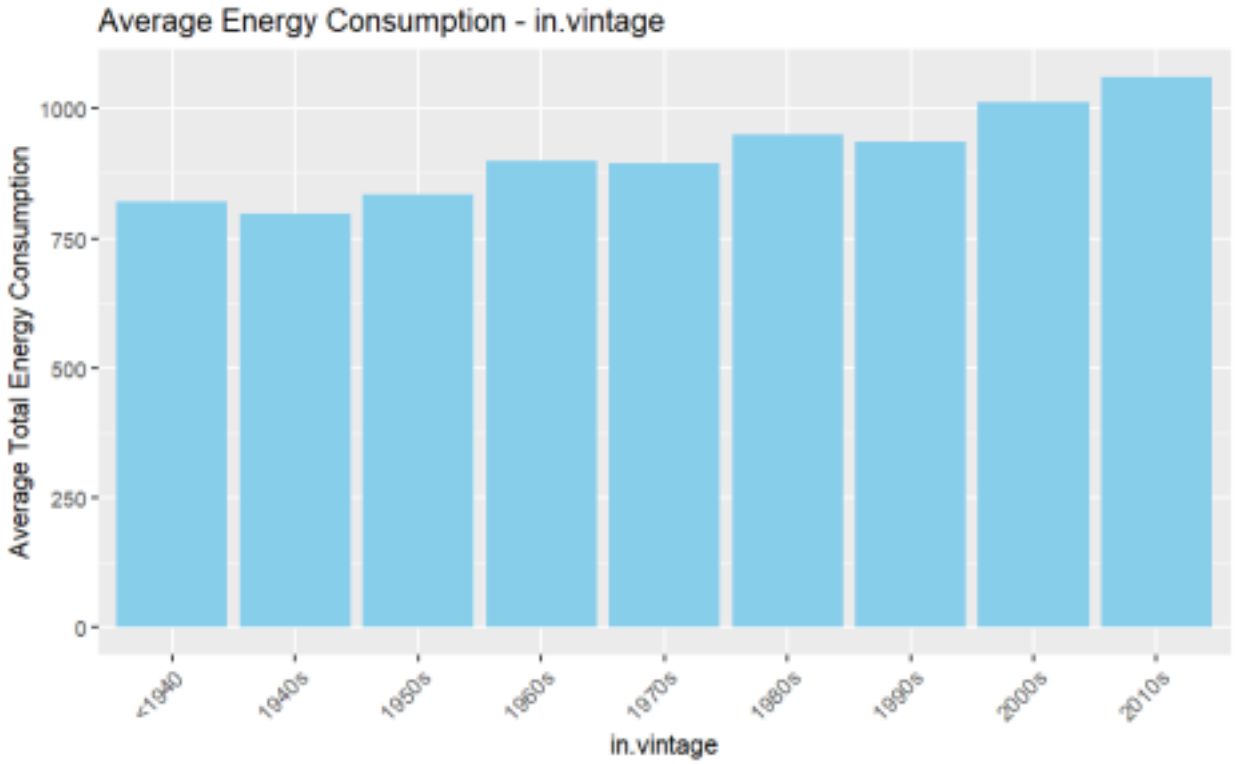
The above image shows us a geographical view of the state of South Carolina and a scatterplot

was plotted on top of it using geom_point(). The points indicate the cities and the size of the point indicates the amount of the Total Energy Usage. Since eSC services a small number of North Carolina areas as well, there are a few points not located within South Carolina

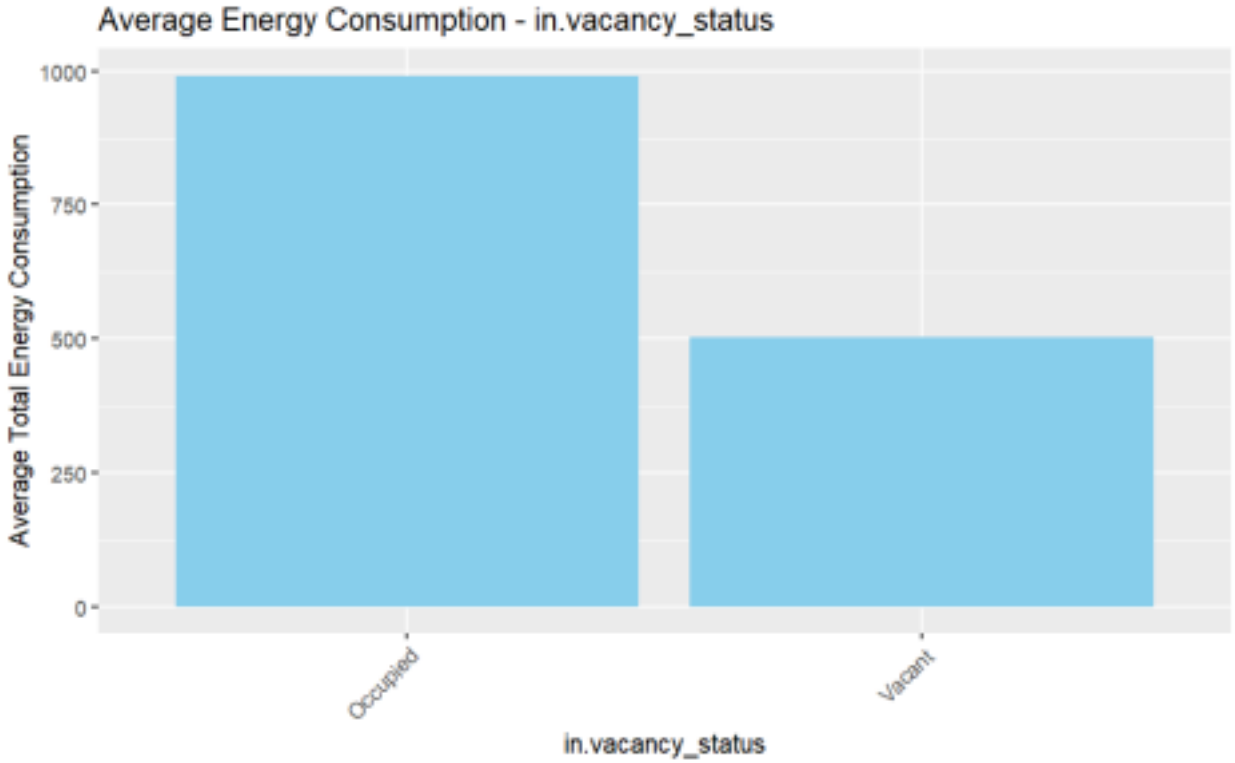


The above bar graph shows Average Total Energy Consumption vs Roof material of houses. It can be derived from this graph that the houses who have their roof material made of Asphalt Shingles are consuming the least energy on average. Also the roofs which are made with Slate or Tile and Concrete seem to have the most average energy consumption. Below is a boxplot of roof material showing a similar trend, with shingle types having a high number of high outliers.

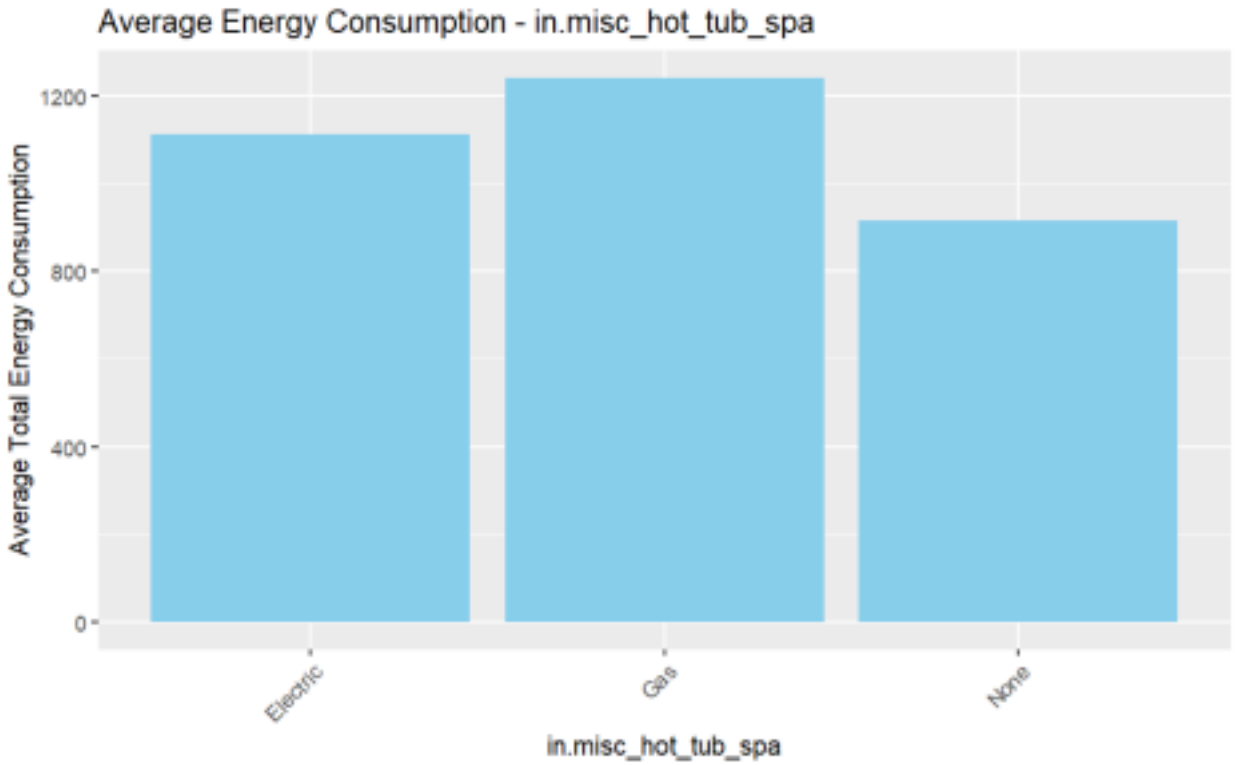




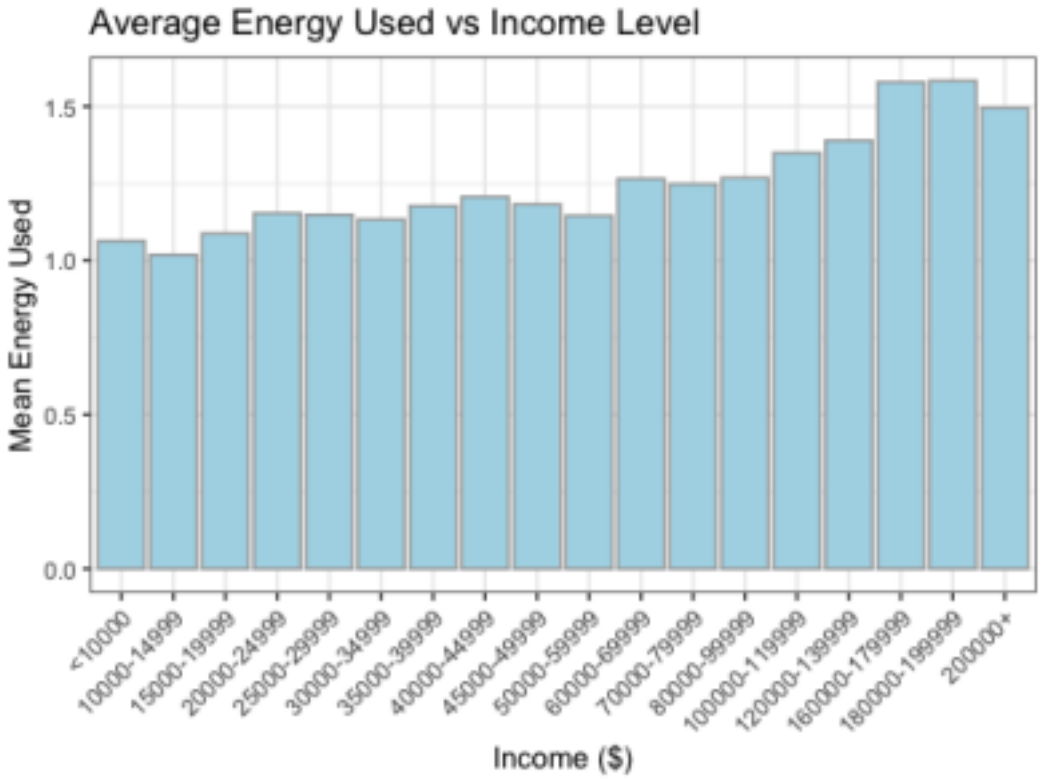
The above graph shows the average total energy consumption of the building according to the period it was constructed in. As there seems to be a pattern that as the years passed by, the average total energy usage increased. The buildings which were constructed in the 1940s seem to have the least total energy usage while the most recent buildings have the highest usage.



This graph shows us that the average total energy is consumed way more in the occupied residences than the vacant residences. The data was derived from 2019 public-use microdata samples [IPUMS] which is an indicator of dwelling units as primary residence or not. This makes a lot of sense since if a building is empty it will use less electricity and energy overall.



The above graph indicates the presence of a hot tub and which type of fuel is used for it. The hot tubs which are fueled by gas seem to consume a higher number of average total energy relative to the hot tubs with electric fuel.



This graph shows us a pattern that higher the income level, the more likely they are to consume higher mean energy. Even though there might be a bit of fluctuation in the pattern of this graph, we can see that the last three groups have the highest mean usage of energy while the first three groups have the lowest.

The bar graph above indicates the baseline cooling setpoints with no offset in a house and how much average energy is consumed by them. The houses with their baseline cooling setpoint at 60F seems to consume the most while the houses with their baseline cooling setpoint at 80F consumes energy the most. During summers it is common for houses to have a lower cooling setpoint thus resulting in a high energy usage. Typically to save money and energy, some houses will purposefully not set their thermostats too low.

Modeling

For the modeling process, we tested a variety of different models as well as different

methods of aggregating our data. Overall, we tested linear models, time series, neural networks, and random forest. Additionally, we tested grouping the data by county, by hour, and by day. To begin with, let's look at the time series model we tested.

Time Series

Initially, we believe a time series model would work the best with our data since it is structured in order of hour for a full month of July data. However, we ran into the issue where each house would need its own individual time series in order to correctly order the data. This is simply not feasible with over 5,000 houses, so our first step was to aggregate the data by county so that each county had its own time series for the month of July, with temperature and energy data at each hour. Since weather data was the same for all houses in a single county, we did not have to do any additional aggregation there. For energy, we used a sum of total energy for all houses in that county as our new county-wide total energy metric. Then we made sure that our data was in the correct order within each county and ran a loop that cycled through each county code in order to produce a time series using time and weather to predict total energy. Additionally, within the loop we used the model created to forecast next year's predicted energy using the increased temperatures. The basic loop code used is below.

Time Series Modeling Loop

Now that we have both the model and the forecast results from the time series output, we can create a plot for each county that shows historical energy use compared to future predicted energy use. The plot below shows the time series output for the county G4500010.

Time Series Plot of Historical vs Predicted Energy Use

The black line in the graph is the energy use for 2018, and the blue line is the predicted energy use for 2019, with the light gray as the confidence interval. As you can see the predictions for next year are higher on average than the 2018 energy usage, which makes sense due to the rise of temperature. However, the confidence intervals appear to be fairly wide, meaning that the model might not be highly accurate.

Although the predictions and plots generated by the time series model were relevant and interesting, we decided not to go with this as our final model for one main reason: we could not include any additional input variables other than datetime and temperature. Time series models are not meant to have a large number of external regressors, especially not as many as we would have liked to include from the static house data. Going along with this, because we had to aggregate by county, it would be impossible to include house-level data in this analysis anyways. If we wanted to include any information from static house info, we needed to keep the individual

houses represented in the model. Our next step was to determine which house variables are significant to predicting energy usage, and so we decided to implement some ANOVA testing.

ANOVA - Determining Significant Variables

During our data exploration and visualization phase, we identified a handful of columns that might be statistically significant predictors of total energy consumption. However we also wanted our column choices to be data driven so we decided to do the ANOVA F-test for the entire set of static house categorical variables. ANOVA F-test looks at how a numeric dependent variable changes across the different levels of a categorical independent variable. Using a for loop, we ran an ANOVA F-test on all the static house data columns. The columns with p-values below the alpha value are statistically significant. Below is an example of an anova result for a single variable, in this case roof material.

ANOVA Example Output

Here you can see that the variable is not significant since the p-value is not below the alpha level of 0.05.

Below are all the columns that our ANOVA F-test did determine as statistically significant. We included all of these columns, as well as mean daily temperature, in all of our future models.

Random Forest

After determining which variables should be included in our model, our next step was to group the data by day (sum up total energy at every hour for each house, as well as calculate the mean temperature at that house for that day) and test out a new model type: random forest. A random forest model can be great because it outputs the variable importance, meaning that once the model runs it provides you with information on which variables were most significant to the model making decisions. After aggregating our energy and temperature data by day and merging

it with the static house data to attach the chosen variables, we used an 80/20 training and testing split on the data and wrote the following code to run the random forest model. *Random Forest Model Code*

The last two lines of code use the model to predict on the test set, and from these predictions, we generated an RMSE of 3.69. This means that on average, we are only 3.69 kWh off of the true total energy used for a house on any given day. We also calculated the R^2 value manually using the following code.

R^2 Random Forest Value

The $R^2$for the random forest model was approximately 92.5%, meaning that we can explain 92.5% of the change in energy using temperature and the variables from static house data that we included in the model. This is a very good predictive value, so we wanted to next look at what the predictions for next year (with the 5℃ increase in temperature) would look like. We predicted next year's data and generated a plot of what these predictions look like compared to the historical data.

Random Forest Predictions Plot

After looking at this plot, we realized that something did not look right, and after some research we found that random forest models are generally not good at extrapolating. This means that since we put in temperature values for prediction that were much higher than what we trained the model on, the model did not actually hold a lot of predictive value, despite the high accuracy on the training/testing data. Therefore, we decided to move on and test other model

types in order to find something that had a similar accuracy to random forest during the training phase, but was better at predicting the projected data for next year.

Neural Networks

After recognizing that our Random Forest model wasn't picking up all the patterns and relationships in our data, we decided to go with a Neural Networks model. Neural Network models are good at recognizing complex patterns and ignoring the noise in data. For our Neural Networks model we had 2 hidden layers with 4 neurons each. We went with a 33% validation split and a batch size of 10. Our model had a Mean Squared Error of 3.648 after running 100

epochs. While our Neural Networks model did give us a good MSE, the model itself took way too long to run. So we decided to go with simpler models that could provide an accurate output without running into vector memory exhaustion issues.

Linear Regression

The final model that we chose for predicting our future energy usage was a Linear Model. We aggregated each house's total energy consumption by day, averaged hourly temperature to get the daily temperature and merged the total energy consumption and temperature with the shortlisted house static data columns. We made a 80-20 train-test split and ran the following code:

Our linear model gave us an adjusted R-squared value of 85%. This means that 85% of the variation in total energy usage can be explained by our independent variables. The Root Mean Squared Error for our model was 5.15. This means that on average, our predictions are only 5.15 kWh off from the true daily energy usage. Below is the output from the model. *Linear Regression Model Results*

Since our linear regression model was accurate and simple but effective, we decided to use this model for our future energy usage predictions. Therefore, we created a new column in the dataset with all of the mean temperatures increased by 5℃ and predicted on this data (keeping all static house columns the same) using the finalized linear regression model.

Final Predictions & Future Peak Energy Demand

Our energy usage predictions for the next year are plotted in the graph below using the red line, and the historic energy data is depicted by the blue line. Ultimately, we can see that the future predictions are much higher than the historical energy usage, which makes sense when considering the large increase in temperature that we accounted for.

Model Predictions vs Past Energy Usage

Overall, the ultimate grid capacity that we found to be necessary for eSC to support all energy needs next year is 241659.27 kWh daily. eSC should be able to support slightly higher than that however, in order to account for slight fluctuations. The household maximum that we expect is about 99 kWh daily.

Shiny Application

Below is a link to our Shiny webpage, as well as some screenshots of the application for explanation purposes.

Link to Shiny Application webpage: https://megkratzer.shinyapps.io/FinalProjectShiny/ 26

Screenshots of the Shiny App

This is the main page of our shiny app, here users can select either any single county or all the counties to plot a line graph.

After selecting any option, A graph will be plotted. As we can see, the blue line indicates the present energy usage (2018) and the red line indicates the predicted energy usage (2019). The text box below also gives us more details regarding highest peak energy usage for both present and future along with the exact date and the number in kWh.

You can also select any individual county to plot data for and find the daily maximum energy usage as well.

Recommendations to eSC

During our data exploration phase we discovered that one of the biggest factors that affected a house's energy consumption was the presence of solar panels. This bar graph depicts the average energy consumption in South Carolina for houses that have solar panels versus houses that don't.

As you can see the presence of solar panels drastically reduces energy usage. We recommend that eSC incentivise their customers to install solar panels in their homes to reduce energy consumption. One simple way of doing this is providing a rebate to customers that prove they have installed solar panels in their home. This is cost effective in the long run since it saves eSC money if their houses have solar panels so they do not have to pay for as much energy. eSC can use these "future" savings to provide the rebates now. Combining this with a marketing campaign that tells consumers about the cost saving and environment saving benefits of solar panels would increase the likelihood that a customer installs them. A partnership with a solar panel provider would also be ideal so that customers do not have to spend time searching for options on their own.

Another important discovery regarding solar panels that we made is that the direction the solar panel is facing has an impact on the energy it produces.

We recommend that eSC's customers orient their solar panels in the South East direction, in order to maximize energy production. This is likely due to the fact that they get more light in this direction. In order to accomplish this, we can give information to customers about which direction they should install their panels if they are considering them, as well as suggest turning already installed ones if possible.

The size of a house's solar panels also has an impact on the energy it produces. We recommend

solar panels of 11 kW DC. Like mentioned above, we can provide this information to customers with the rebate information, as well.

Additionally, early on during our EDA process, we noticed that whether a house has a pool or not made a big difference in its energy consumption. Since eSC can't ask its customers to cut down on pool use during the summer, instead we recommend that they switch to solar power to power their pool heater. This graph clearly shows how much energy usage decreases by making the shift to solar.

We recognize that the high cost of installing solar panels can prevent customers from making the switch to solar. However, as we can see from the bar graph below, eSC customers with the highest income are also the ones with the highest energy usage. eSC could recommend specifically that these customers invest in Solar energy now in order to save on energy bills in the future.

A secondary recommendation that we would give to eSC is to incentivise their customers to make the switch to LED lighting. The bar graph below depicts the average energy usage in South Carolina for each type of lighting. Evidently, LED lighting results in lower energy consumption.

We suggest that eSC encourage customers to switch to solar power and LED lighting by providing financial incentives. For instance, eSC could offer rebates and discounts for customers to make the switch. Governments often provide tax incentives to get citizens to adopt renewable

energy solutions. eSC could educate its customers about available tax credits for solar installations. eSC could also provide customers with education materials like brochures that explain the benefits of switching to solar power and LED lighting. For example, they could educate customers on the potential savings on energy bills in the long run.

Work Log (Who Did What):

Akshay:

- Explored datasets in order to derive variables for visualizations.
- Created visualizations for columns which were shortlisted as part of EDA.
- Worked on the UI of Shiny app.

Punami:

- Ran ANOVA F-test on house static data to shortlist columns
- Created visualizations with the shortlisted columns
- Neural Network and Linear Model

Vaibhav:

- Explored all the datasets in order to derive which variables to use.
- Worked on Linear model using the variables shortlisted from ANOVA F-Test.
- Created Shiny App.

Divya:

- Worked on a different approach on how to merge datasets using the purr package (without the for loop).

● Plotted some visualizations such as ggmaps and bar graphs as part of Exploratory Data Analysis.

● Worked on the final presentation.

Megan:

● Downloading, cleaning, and merging the data using for loops

● Time Series & Random Forest modeling

● Variety of graphs such as energy vs temperature and energy vs income level 35