Introduction to Data Science IST659

Understanding Key Drivers of Residential Energy Consumption in Peak Month

Group 1: Sophia Jaskoski, Marjan Abedini, Aaryan Wani, Sindy Siddarth Reddy Kolli

---

Introduction

eSC is a small energy company which provides electricity to residential properties in South Carolina and part of North Carolina. The firm wants to understand how they can encourage customers to save energy to avoid having to build another energy production facility amidst increasing energy consumption.

Problem Definition

The client needs to accurately assess the current demand for energy by different sources in their customers' homes. Electricity demand in known to increase during times of extreme weather (summer and winter), and therefore it is important to know what the maximum energy need is during those months to ensure that the demand does not outpace supply and cause power outages in the serviced areas. The client has determined that July is the month for which they expect energy consumption to peak and possibly outstrip their production capabilities. Using historical consumption, house attributes, and weather data, we will predict energy consumption for these homes for a hypothetical July which see all hourly temperatures increase by 5 degrees Celsius. Once we understand the change in magnitude of energy consumption in an extreme environment, we will pinpoint efficient and realistic ways in which eSC can incentivize its customers to reduce energy consumption and decrease the strain on the power facilities, even as temperatures increase in the future.

Available Data

eSC has provided us with data about the homes they service and the respective energy consumption of each home. Each home is identifiable by a unique building id and corresponds to a

county code. Each county has corresponding weather information. Weather and energy data are provided for every hour of every day for every month in 2018. The housing attribute data is static.

Housing Data Snapshot:

5710 observations (homes) with 171 columns (attributes)

bldg_id	upgrade	weight	applicability	in.sqft	in.ahs_region	in.ashrae_iecc_climate_zone_2004	in.ashrae_iecc_climate_zone_2004_2_a_split	in.bathroom_spot_vent_hour	in.bedrooms	in.building_america_climate_zone
65	10	242.131	TRUE	885	Non-CBSA South Atlantic	3A	3A	Hour23	3	Mixed-Humid
121	10	242.131	TRUE	1220	Non-CBSA South Atlantic	3A	3A	Hour20	2	Mixed-Humid
500	10	242.131	TRUE	1220	Non-CBSA South Atlantic	3A	3A	Hour11	3	Mixed-Humid
504	10	242.131	TRUE	1690	Non-CBSA South Atlantic	3A	3A	Hour13	3	Mixed-Humid
581	10	242.131	TRUE	1690	Non-CBSA South Atlantic	3A	3A	Hour22	3	Mixed-Humid
590	10	242.131	TRUE	2176	Non-CBSA South Atlantic	3A	3A	Hour5	2	Hot-Humid

The entirety of the data can be found here:

https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/static_house_info.parquet

Energy Data Snapshot:

8759 observations (energy usage for every hour of every day for every month in 2018) of 43 variables (sources which consume energy)

out.electricity.ceiling_fan.energy_consumption	out.electricity.clothes_dryer.energy_consumption	out.electricity.clothes_washer.energy_consumption	out.electricity.cooling_fans_pumps.energy_consumption
0.006	0	0	0.000
0.008	0	0	0.000
0.008	0	0	0.000
0.008	0	0	0.000
0.008	0	0	0.000
0.008	0	0	0.000

An example of energy data for building 102063 can be found here:

https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/2023-houseData/102063.parquet

Weather Data Snapshot:

8760 observations of 8 variables, including time, temperature in degrees Celsius, relative humidity, wind speed, wind direction, global horizontal radiation, direct normal radiation, and diffuse horizontal radiation.

date_time	Dry Bulb Temperature [Â°C]	Relative Humidity [%]	Wind Speed [m/s]	Wind Direction [Deg]	Global Horizontal Radiation [W/m2]	Direct Normal Radiation [W/m2]	Diffuse Horizontal Radiation [W/m2]
2018-01-01 01:00:00	-3.30	61.71	3.60	30.00	0.0	0.0	0.0
2018-01-01 02:00:00	-3.30	61.71	4.10	30.00	0.0	0.0	0.0
2018-01-01 03:00:00	-3.30	50.77	5.10	20.00	0.0	0.0	0.0
2018-01-01 04:00:00	-3.30	43.56	3.60	40.00	0.0	0.0	0.0
2018-01-01 05:00:00	-3.30	35.62	4.10	40.00	0.0	0.0	0.0
2018-01-01 06:00:00	-3.30	35.62	4.60	40.00	0.0	0.0	0.0

An example of the weather data for county G4500010 can be found here:

https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/weather/2023-weather-data/G4500010.csv

Exploratory Data Analysis

Given the volume of data provided, it was necessary to be tactical about which energy sources, home attributes and weather information to consider before any analysis was performed.

Below, we looked at cross-sections of the data to visualize the distribution of variables, understand the structure of the data, and summarize energy usage and weather changes across time.

Data Exploration and Preparation

Static House Data

The **static house data** contained many columns which contained information we deemed to be useless in terms of differentiating between the homes in this list. These columns only included one value, which was the same for every building in our data.

For example, the columns below are redundant given that we already know eSC services homes in South Carolina and parts of North Carolina. The region of these homes is the same for all, so it would not be a viable factor to consider when determining differentiators of energy usage.

in.census_division	in.census_division_recs	in.census_region
South Atlantic	South Atlantic	South
South Atlantic	South Atlantic	South
South Atlantic	South Atlantic	South
South Atlantic	South Atlantic	South
South Atlantic	South Atlantic	South
South Atlantic	South Atlantic	South
South Atlantic	South Atlantic	South

In total, we ended up with 93 columns from the original 171 that we found to contain distinctive information.

The code below was used to remove the following columns from the static house data and create a new data frame.

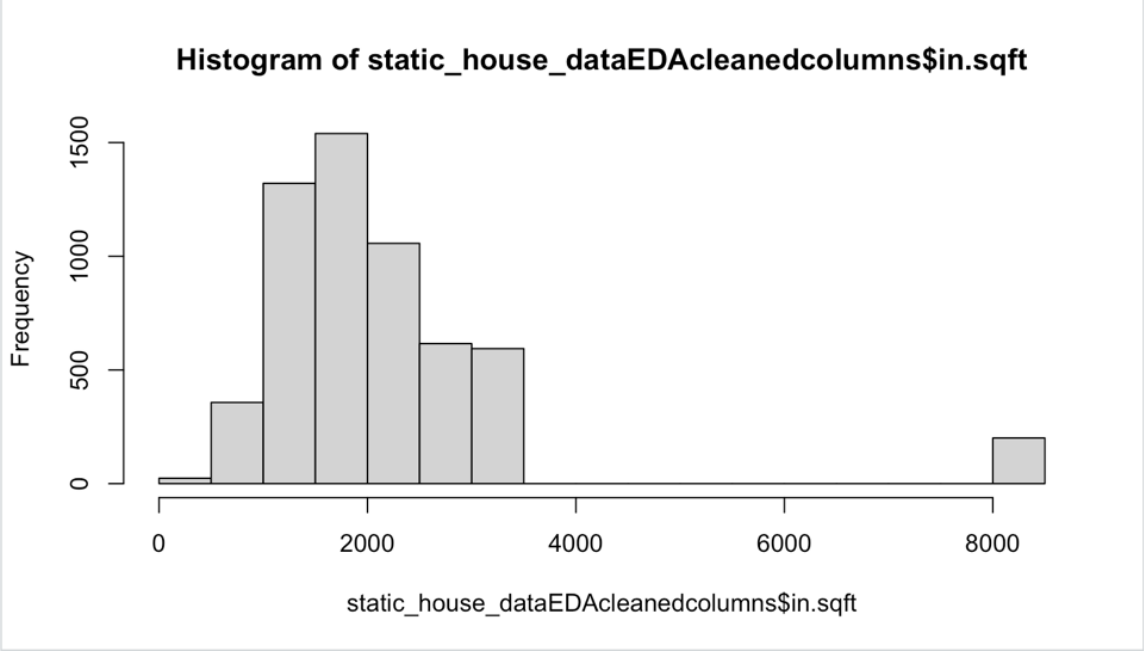
```
static_house_dataEDAcleanedcolumns<-static_house_dataEDA[,-c(2,3,4,6,7,8,12,14,15,16, 26, 29, 31,32,
                                                34,35,36,37,38,39,40,41,42,43,44,45,47,
                                                49,50,51,52,53,54,55,56,63,72,73,79,84,
                                                85,86,87,88,89,90,91,103,104,106,107,108
                                                ,109,120,121,124,126,131,136,137,138,139,
                                                140,141,142,143,144,146,153,159,161,162,
                                                163,165,166,167,168,169)]
```

The reduced data file consisted of integer, numeric, and (mostly) character data types.

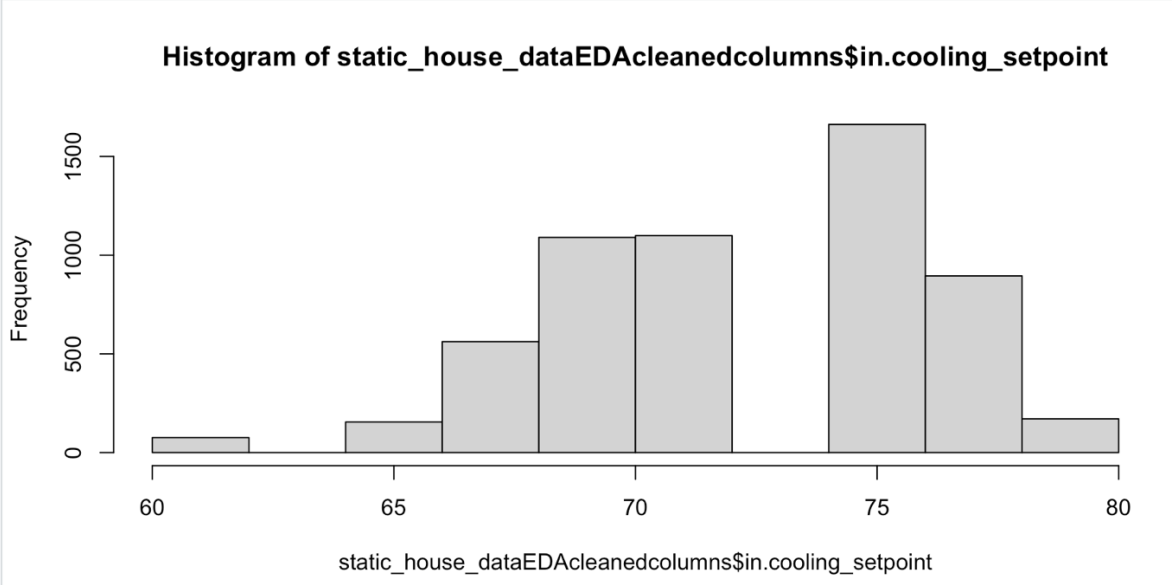
```
'data.frame':   5710 obs. of  93 variables:
 $ bldg_id                                : int  65 121 500 504 581 590 670 736 862 952 ...
 $ in.sqft                                : int  885 1220 1220 1690 1690 2176 885 2663 885 2663 ...
 $ in.bathroom_spot_vent_hour             : chr  "Hour23" "Hour20" "Hour11" "Hour13" ...
 $ in.bedrooms                            : int  3 2 3 3 3 2 2 4 2 3 ...
 $ in.building_america_climate_zone       : chr  "Mixed-Humid" "Mixed-Humid" "Mixed-Humid" "Mixed-Humid" ...
 $ in.ceiling_fan                         : chr  "Standard Efficiency" "None" "Standard Efficiency" "Standard Efficiency" ...
 $ in.city                                : chr  "SC, Rock Hill" "Not in a census Place" "Not in a census Place" "In another c
nsus Place" ...
 $ in.clothes_dryer                       : chr  "Gas, 100% Usage" "Electric, 100% Usage" "Electric, 80% Usage" "Electric, 80%
Usage" ...
 $ in.clothes_washer                      : chr  "Standard, 100% Usage" "EnergyStar, 100% Usage" "Standard, 80% Usage" "Energy
tar, 80% Usage" ...
 $ in.clothes_washer_presence             : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ in.cooking_range                       : chr  "Electric, 100% Usage" "Electric, 100% Usage" "Gas, 80% Usage" "Electric, 80%
Usage" ...
 $ in.cooling_setpoint                    : chr  "72F" "76F" "70F" "70F" ...
 $ in.cooling_setpoint_has_offset         : chr  "No" "No" "No" "Yes" ...
 $ in.cooling_setpoint_offset_magnitude   : chr  "0F" "0F" "0F" "2F" ...
 $ in.cooling_setpoint_offset_period      : chr  "None" "None" "None" "Night Setup +3h" ...
 $ in.county                              : chr  "G4500910" "G4500730" "G4500710" "G4500790" ...
 $ in.county_and_puma                     : chr  "G4500910, G45000502" "G4500730, G45000101" "G4500710, G45000400" "G4500790,
```

Variable Exploration:

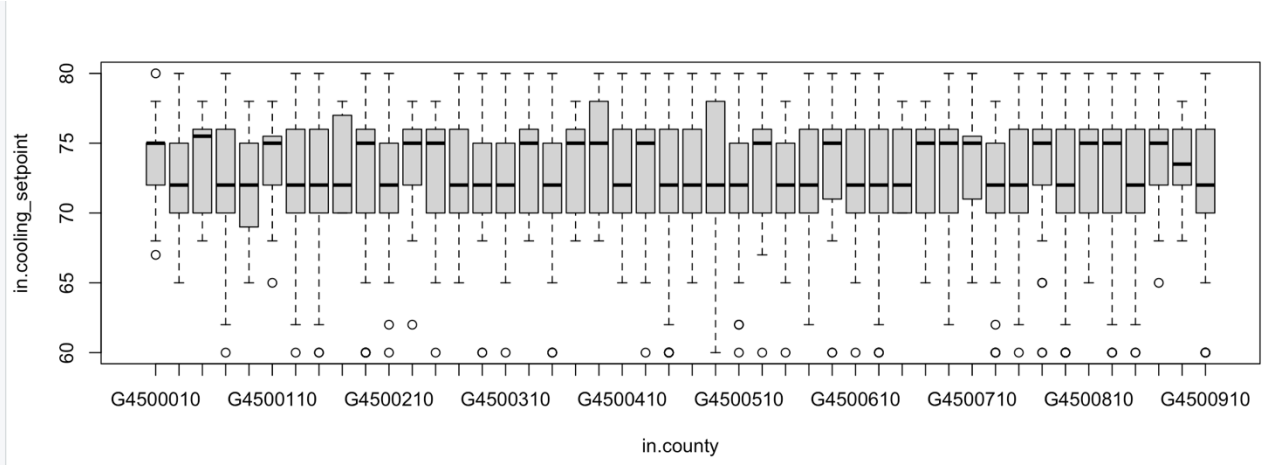
The **square footage variable** was skewed right and had outliers at large values of square feet:



The **cooling setpoint** is left-skewed, with the highest frequency of setpoint at 75 degrees Fahrenheit.



The boxplot comparing cooling setpoint by county shows overlap among all the counties and an average cooling setpoint of about 73 degrees F.



Energy Data

Inspecting one example file of the **energy data**, we found variables corresponding to the static attributes of the house data, including ceiling fan usage, cooling energy, interior lighting energy usage, and pool pump usage to name a few. Electrical units were measure in kWh and KWh/sqft.

The structure of the file revealed that we are dealing with numeric and POSIXct (date_time) data types.

```
Classes ‘tbl_df’, ‘tbl’ and 'data.frame':       8759 obs. of  43 variables:
 $ out.electricity.ceiling_fan.energy_consumption       : num  0.006 0.008 0.008 0.008 0
 $ out.electricity.clothes_dryer.energy_consumption     : num  0 0 0 0 0 0 0 0 0 0 ...
 $ out.electricity.clothes_washer.energy_consumption    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ out.electricity.cooling_fans_pumps.energy_consumption: num  0 0 0 0 0 0 0 0 0 0 ...
 $ out.electricity.cooling.energy_consumption           : num  0 0 0 0 0 0 0 0 0 0 ...
 $ out.electricity.dishwasher.energy_consumption        : num  0 0 0 0 0 0 0 0 0 0 ...
 $ out.electricity.freezer.energy_consumption           : num  0.021 0.028 0.028 0.028 0
 $ out.electricity.heating_fans_pumps.energy_consumption: num  0.145 0.16 0.162 0.164 0.
 $ out.electricity.heating_hp_bkup.energy_consumption   : num  0.194 0 0 0 0 0 0.502 0
 $ out.electricity.heating.energy_consumption           : num  1.78 2.07 2.1 2.11 2.1 ..
 $ out.electricity.hot_tub_heater.energy_consumption    : num  0 0 0 0 0 0 0 0 0 0 ...
 $ out.electricity.hot_tub_pump.energy_consumption      : num  0 0 0 0 0 0 0 0 0 0 ...
 $ out.electricity.hot_water.energy_consumption         : num  0.003 0.004 0.004 0.004 0
 $ out.electricity.lighting_exterior.energy_consumption : num  0.018 0.021 0.02 0.02 0.0
 $ out.electricity.lighting_garage.energy_consumption   : num  0 0 0 0 0 0 0 0 0 0 ...
```

A summary of the file shows that certain columns have all zero values, suggesting that no energy was consumed and/or the house does not have that appliance.

```
out.electricity.hot_tub_heater.energy_consumption out.electricity.hot_tub_pump.energy_consumption
Min.   :0                                         Min.   :0
1st Qu.:0                                         1st Qu.:0
Median :0                                         Median :0
Mean   :0                                         Mean   :0
3rd Qu.:0                                         3rd Qu.:0
Max.   :0                                         Max.   :0

out.electricity.hot_water.energy_consumption out.electricity.lighting_exterior.energy_consumption
Min.   :0.0030                               Min.   :0.00800
1st Qu.:0.0040                               1st Qu.:0.01200
Median :0.0040                               Median :0.01600
Mean   :0.0976                               Mean   :0.01833
3rd Qu.:0.1900                               3rd Qu.:0.02200
Max.   :0.4990                               Max.   :0.04000

out.electricity.lighting_garage.energy_consumption out.electricity.lighting_interior.energy_consumption
Min.   :0                                          Min.   :0.0300
1st Qu.:0                                          1st Qu.:0.0400
Median :0                                          Median :0.0790
Mean   :0                                          Mean   :0.1751
3rd Qu.:0                                          3rd Qu.:0.2640
Max.   :0                                          Max.   :0.8020
```

In the below example, we see that this building does not have a dishwasher, which gives context as to why dishwasher energy consumption in the corresponding energy file is zero throughout.

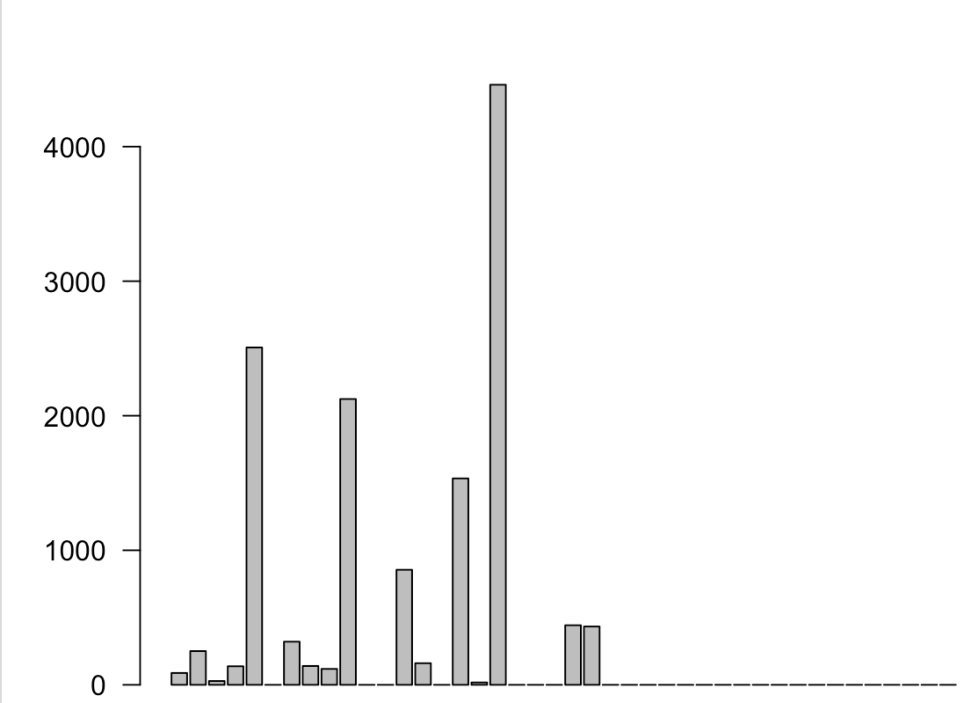
Building 102063:

in.county_and_puma	in.dehumidifier	in.dishwasher	in.door_area	in.doors	in.ducts	in.eaves	in.electric_vehicle
G4500690, G45000700	None	None	20 ft^2	Fiberglass	20% Leakage, R-4	2 ft	None

Energy output for building 102063:

out.electricity.dishwasher.energy_consumption
0
0
0
0
0
0
0
0

If we sum the energy sources, we see that for this particular building, the biggest contributors to energy consumption were **out.electricity.plug_loads.energy_consumption** (4460.016 kWh), **out.electricity.cooling.energy_consumption** (2507.141 kWh), and **out.electricity.heating.energy_consumption** (2124.227kWh), respectively. This conceptually makes sense, as heating and cooling are known to be some of the largest contributors to overall energy consumption.

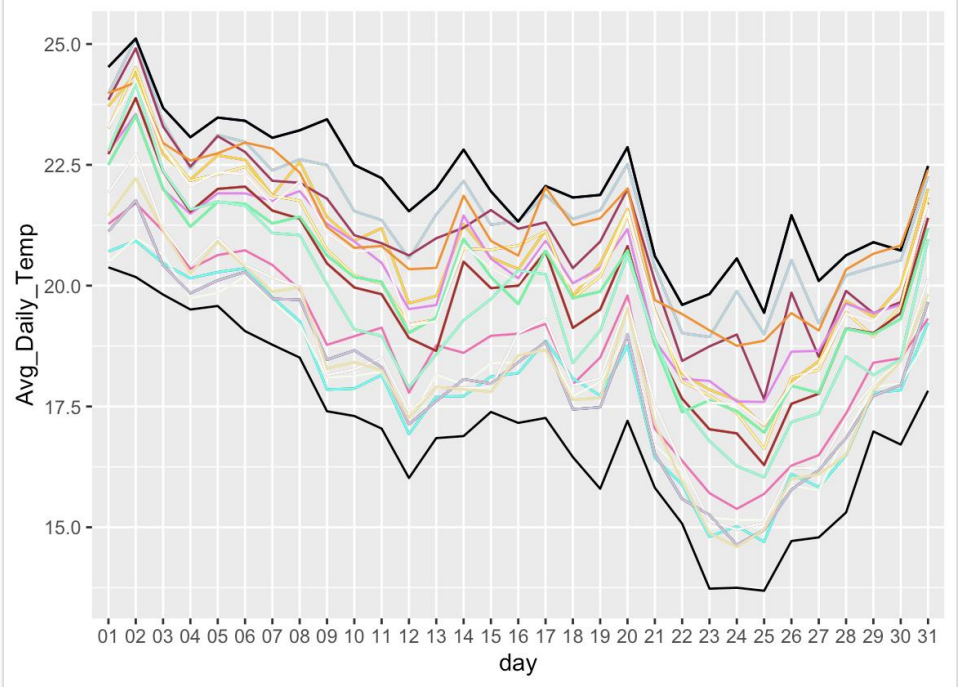


Also of note, we inspected the file and found one row with all NA values, which we made sure to be wary of in other energy files we explored.

Weather Data

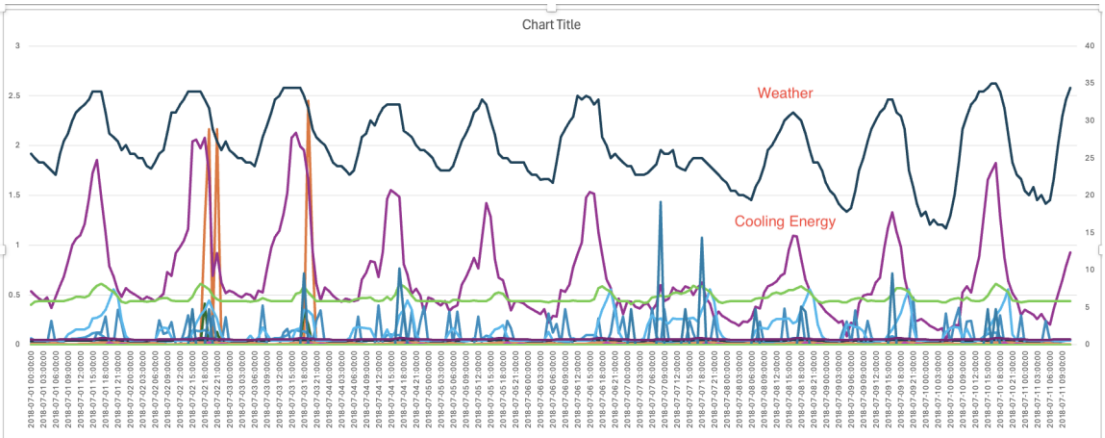
There are 46 counties represented by the houses in the static house data. We are primarily interested in the weather data for each county during the month of July, when conditions are expected to be the most extreme and have the most impact on energy usage.

A plot of average daily July temperatures in Celsius for each county is below:

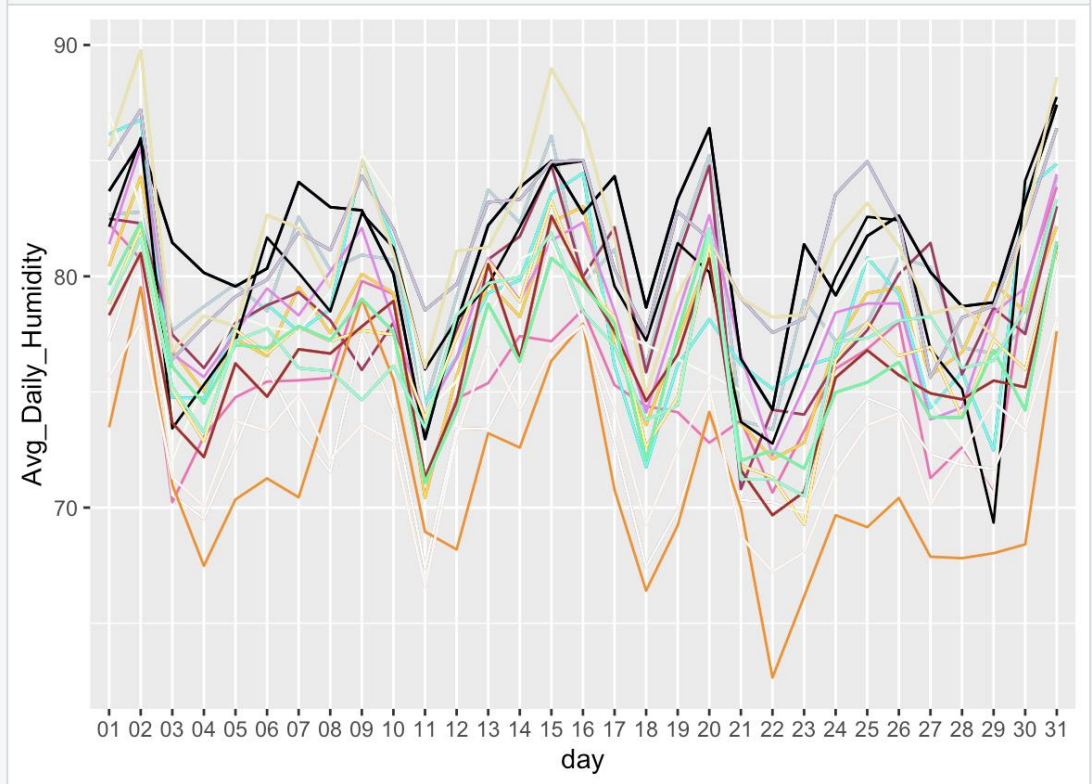


The lines in black represent on average, the hottest and coldest counties during the month of July. The plot shows that that while different temperatures were observed, the trends in temperatures were largely the same for all counties studied.

If we plot the temperatures for the 'hottest' county against the energy data for a random building in that county, we revealed an apparent correlation between the weather trend and the cooling energy usage, see in the graph below. Less apparent, the green line representing plug loads also appeared to increase and decrease with the temperature. The spikes of orange show the energy used by ovens. This graph was informative because at least two of the variables which appeared to have a relationship with temperature were also the variables we found to be the largest consumers of energy in the example energy files, cooling energy and plug loads. Consequently, we were interested in exploring this relationship further in our analysis.



An additional metric we wanted to plot was humidity, to understand if it shared a similar relationship to temperature. If we change the y-axis to measure relative humidity, we see much more overlap between the counties and see that the two black lines, representing the hottest and coldest counties, overlap significantly and do not represent the most extreme counties in terms of humidity levels.



Hypotheses and Variable Selection

Once we were familiar with the data, we were able to curate a set of business questions we would pursue as the scope of this project to help us in our overarching goal of determining key drivers of energy usage which can be addressed to lower consumption among consumers. These questions were designed with July in mind, which is why we primarily focused on systems and attributes we believed would be most affected by hot weather and comprise a significant proportion of a home's energy consumption.

Business Questions:

1. **Which sources of energy consume the most power and is this consistent across all homes?**
    a. In order to provide impactful recommendations to reduce energy consumption, we want to understand which sources consume the most power relative to total energy consumption. A small reduction in the energy consumption from these sources would likely have a large cumulative impact on reducing overall energy usage.
2. **Do consumers living in the hottest county use more total energy than those consumers living the "coldest" county in our data?**
    a. This business question will help us understand the magnitude of impact on energy usage (if it exists) between homes in hotter environments. This is useful to understand since we are concerned about rising temperatures due to global warming and ultimately want to predict how energy usage changes in an environment which is 5 degrees warmer.
3. **Does the type of HVAC cooling system in a home (Heat Pump vs Central AC) have an impact on cooling energy usage?**
    a. Heat pumps are touted a as more environmentally friendly alternative to Central AC, yet in our data set only about 21% of homes have one. If we can determine that Heat Pumps use less every than Central AC, it may be worthwhile to incentivize eSC customers to make the switch.
4. **Does the square footage of a house correspond to an increase in use of interior lighting energy and plug load energy consumption?**
    a. While air conditioning usage in the month of July is an obvious target for potential energy reduction, some less obvious major consumers of energy usage are outlets and lighting. We want to determine if a home's size has a meaningful effect on the amount of interior lighting and outlet energy usage. If so, convincing consumers to turn off lights or unplug electronics might be a more realistic and popular way to decrease consumer energy consumption.
5. **Do the top contributors to energy usage present a meaningful relationship with temperature?**
    a. Our ultimate goal in this project is to be able to predict energy usage in a climate which has temperatures increase by 5 degrees Celsius. It is important to understand whether the variables contributing most to energy usage have a relationship with the weather, separate from the static attributes of a house. We will consider the 'average' house by taking a sample of homes with the median value of square footage and compare the average energy usage of these homes at various temperatures.

Modeling

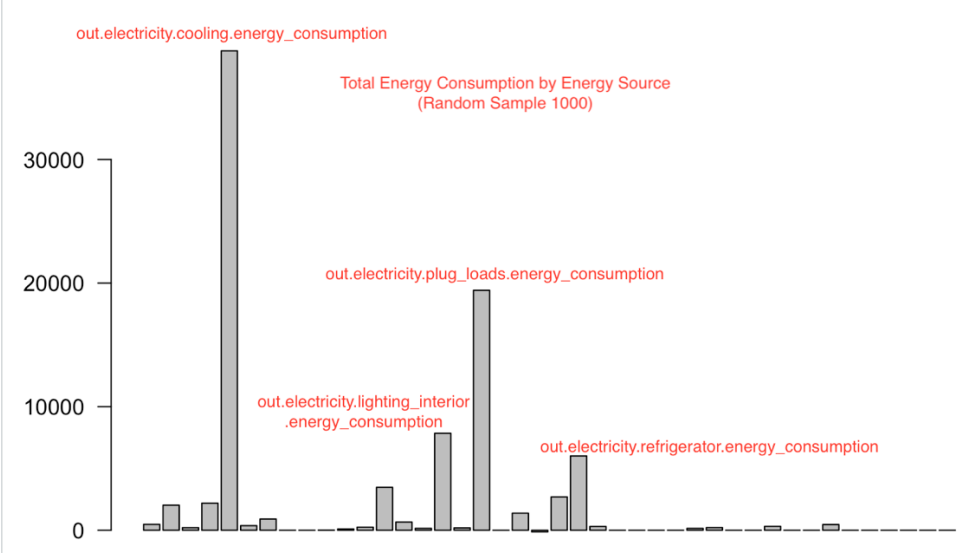
Business Question #1: **Which sources of energy consume the most power and is this consistent across homes?**

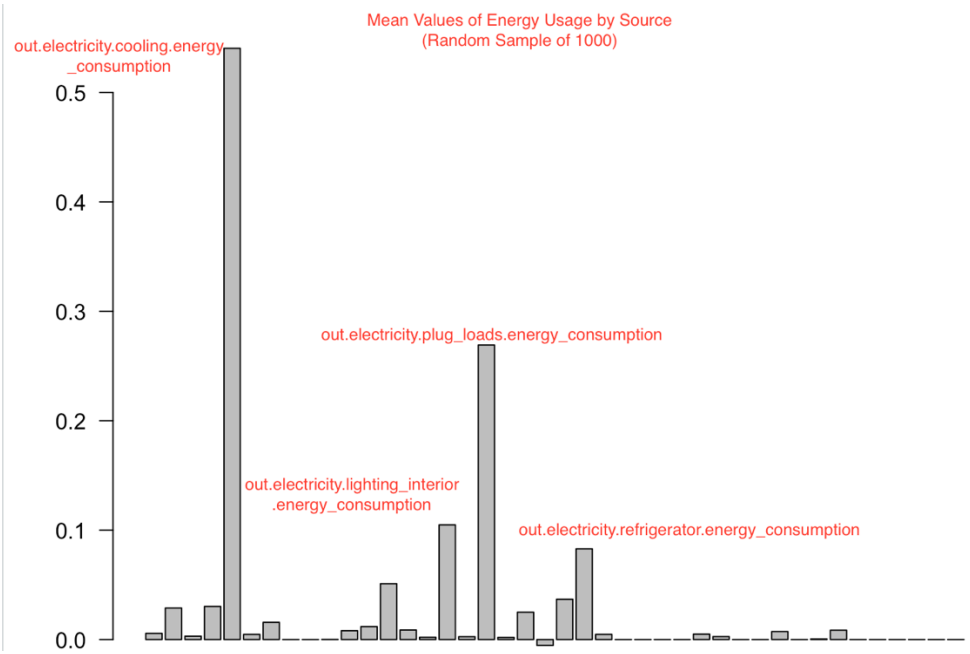
Approach:

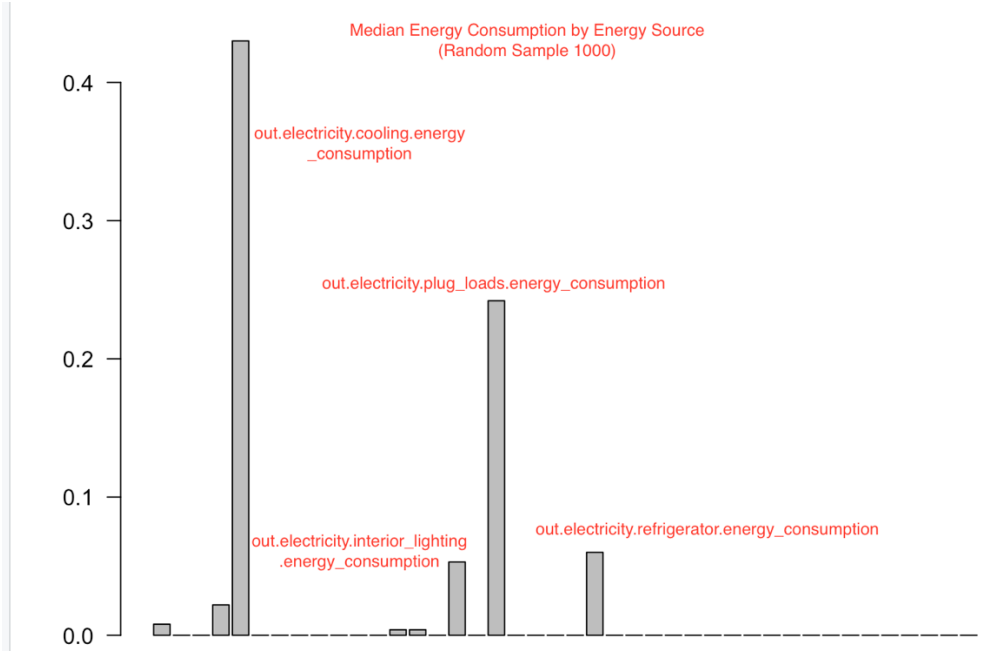
- Take a random sample of 1000 buildings from the static house data
- Import and bind the energy files corresponding to the random sample of buildings
- Use the colSums function to sum the energy usage over the course of July for each energy source
- Plot the energy totals to see which sources use the most energy in July for this sample of 1000 houses
- Repeat process for mean and median values of energy consumption

Result:

- We found that the top energy users across the board for the month of July were:
    - **out.electricity.cooling.energy_consumption**
    - **out.electricity.plug_loads.energy_consumption**
    - **out.electricity.lighting_interior.energy_consumption**
    - **out.electricity.refrigerator.energy_consumption**
- Taking the total, mean, and median energy usage for this sample gave us the same results:







Based on the results of this EDA, we are interested in diving deeper on which variables drive energy usage of these sources.

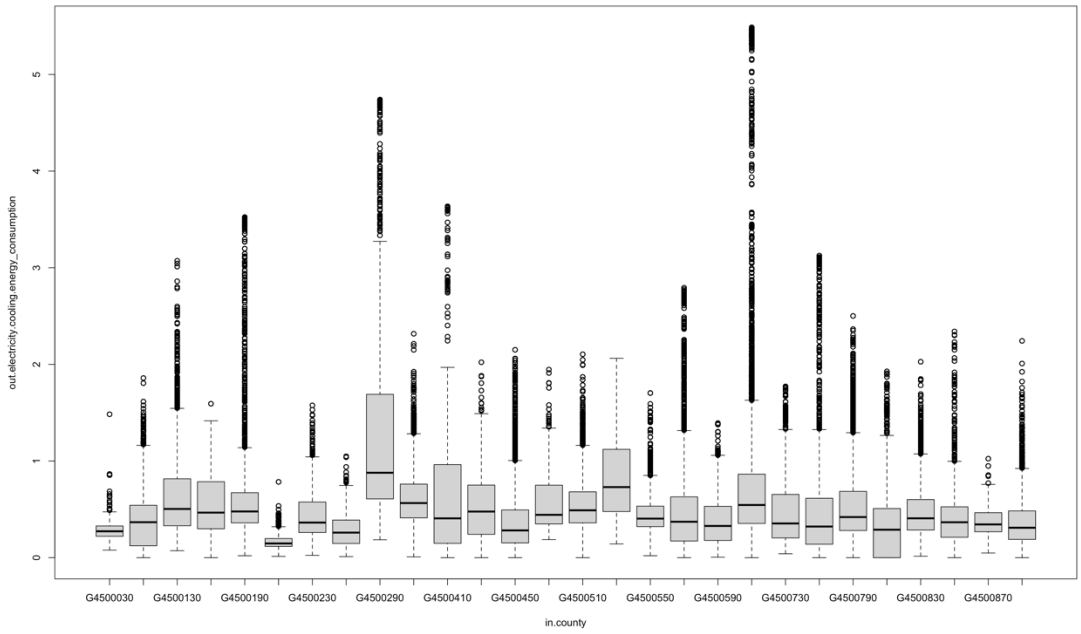
Business Question #2: **Do consumers living in the hottest county use more cooling energy than those consumers living the "coldest" county in our data?**

Solution Approach:

- Create a function to import the weather data for all 46 counties, adding a column with the county id for distinction.
- Filter the weather files for July and group by county to find the counties which have the highest and lowest average temperatures for July.
- Subset the static house data set to only include the homes in these counties.
- Create a function to import the energy files associated with those homes and create a column for building id
- Merge the static house and energy data on building id, then merge with the weather files on county id
- Visualize cooling energy consumption by county
- Create a linear model with predicts cooling energy based on county

Initial visualization suggested that there was a difference in the cooling energy used by certain counties, seen below. However, filtering to compare the “hottest” and “coldest” counties suggested that those counties in “extreme” temperature conditions actually used a comparable amount of cooling energy.

Cooling Energy by counties:

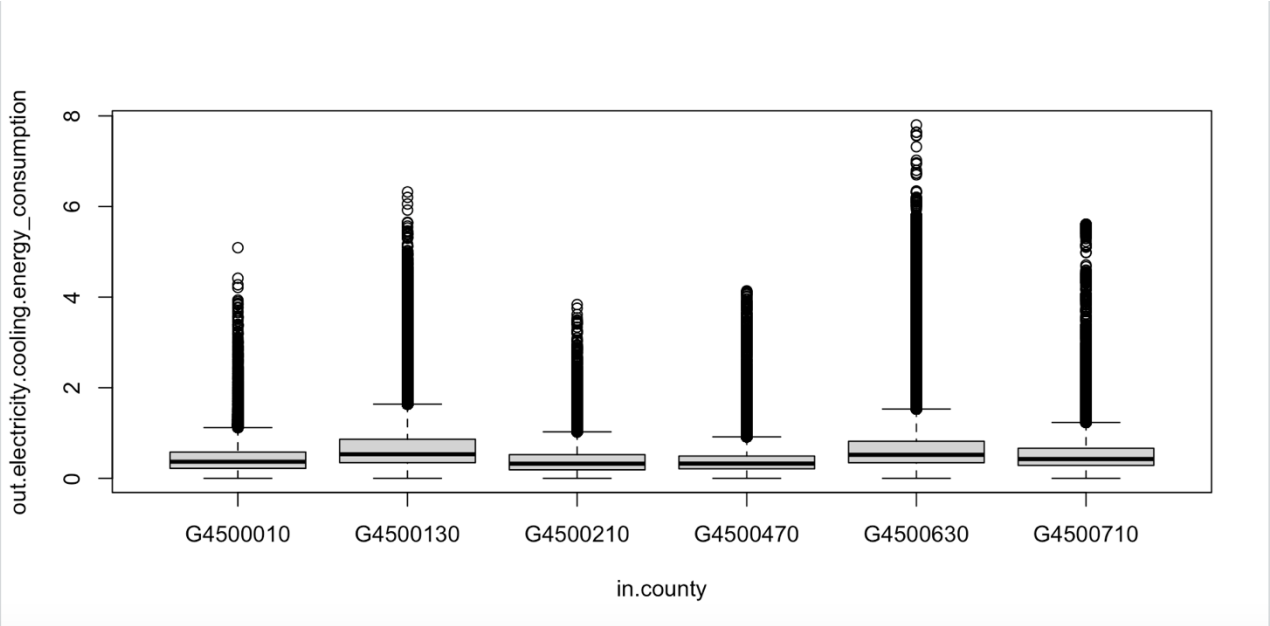


Hottest and Coldest Counties:

County_ID	Avg_Monthly_Temp
G4500630	27.98121
G4500710	27.98121
G4500130	27.96605

County_ID	Avg_Monthly_Temp
G4500210	24.60147
G4500010	25.04963
G4500470	25.04963

Cooling Energy Usage of Hottest and Coldest Counties:



An initial boxplot suggests that even though these counties reported the most extreme differences in July temperatures, homes within these counties used similar amounts of cooling energy. We would expect to see homes in the “hot” counties use more cooling energy than homes in the “cold” counties but did not see a substantial difference which would suggest that average monthly temperature alone accounts for these differences.

	in.county	Monthly_Cooling_Usage
1	G4500010	9567.535
2	G4500210	18352.544
3	G4500710	22021.046
4	G4500470	24902.049
5	G4500130	117596.295
6	G4500630	159988.663

```
Call:
lm(formula = out.electricity.cooling.energy_consumption ~ in.county,
    data = HotandColdALL)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6853 -0.2954 -0.1254  0.1417  7.1336

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.457580   0.003441  132.97   <2e-16 ***
in.countyG4500130  0.227721   0.003645   62.48   <2e-16 ***
in.countyG4500210 -0.047027   0.004169  -11.28   <2e-16 ***
in.countyG4500470 -0.060589   0.003974  -15.25   <2e-16 ***
in.countyG4500630  0.208781   0.003588   58.19   <2e-16 ***
in.countyG4500710  0.097736   0.004253   22.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4976 on 579678 degrees of freedom
Multiple R-squared:  0.04626,   Adjusted R-squared:  0.04625
F-statistic:  5623 on 5 and 579678 DF,  p-value: < 2.2e-16
```

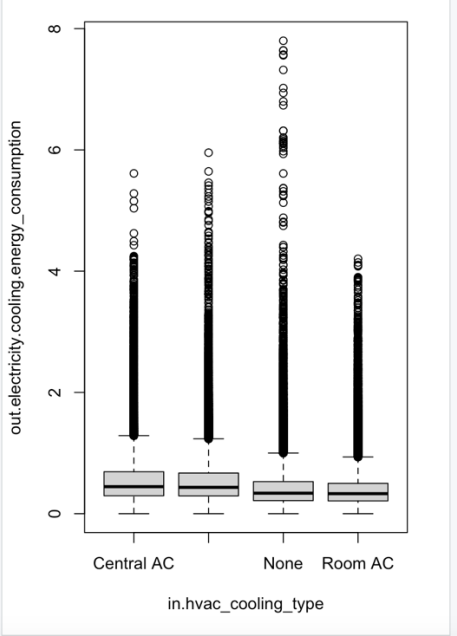
The model showed that a county by itself was not enough to account for the variation in cooling energy seen in our sample of “hot” and “cold” weather houses. We followed up this model with an additional model adding house attributes which might affect a home’s ability to keep cooling energy in/hot weather out and settled on the following variables to add to our model: Cooling setpoint, time of day, type of windows, infiltration, and duct leakage. This improved our original model to an R-Squared value of 0.1054, which told us that we were still missing the main attribute which determined cooling energy consumption.

```
Call:
lm(formula = out.electricity.cooling.energy_consumption ~ in.ducts +
    in.cooling_setpoint_has_offset + in.cooling_setpoint_offset_magnitude +
    in.cooling_setpoint_offset_period + time + in.infiltration +
    in.windows + HotandColdALL$in.building_america_climate_zone,
    data = HotandColdALL)

Residuals:
    Min      1Q  Median      3Q     Max
-1.5783 -0.2770 -0.1075  0.1503  6.7546

Residual standard error: 0.4819 on 579592 degrees of freedom
Multiple R-squared:  0.1055,    Adjusted R-squared:  0.1054
F-statistic: 751.4 on 91 and 579592 DF,  p-value: < 2.2e-16
```

Business Question #3: **Does the type of HVAC cooling system in a home (Heat Pump vs Central AC) have an impact on cooling energy usage?**

Solution Approach:

- Understand the types and prevalence of HVAC systems which exist in our static data: Central AC, Heat Pump, Room AC, and No AC
- Select an equal random sample of each type of HVAC system and merge the samples into one sample of 500 houses.
- Import the corresponding energy files and filter for July
- Create preliminary boxplots to visualize potential relationship:
    - There could be a slight difference between Central AC and Room AC, but most of the distributions seem to overlap. There appear to be many outliers and there is still cooling energy being used in homes with no AC system. This suggests that the type of HVAC system might not be related to the amount of cooling energy consumed.



A linear model determines that even though the type of HVAC is significant, the R-squared value is very low.

```
Call:
lm(formula = out.electricity.cooling.energy_consumption ~ in.hvac_cooling_type,
    data = HVACHouseEnergy)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5482 -0.2272 -0.0923  0.1148  7.3667

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   0.548211   0.001259 435.284  < 2e-16 ***
in.hvac_cooling_typeHeat Pump -0.010029   0.001781  -5.631 1.79e-08 ***
in.hvac_cooling_typeNone      -0.114959   0.001781 -64.544  < 2e-16 ***
in.hvac_cooling_typeRoom AC   -0.144034   0.001781 -80.868  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3781 on 360496 degrees of freedom
Multiple R-squared:  0.02717,   Adjusted R-squared:  0.02716
F-statistic:  3356 on 3 and 360496 DF,  p-value: < 2.2e-16
```

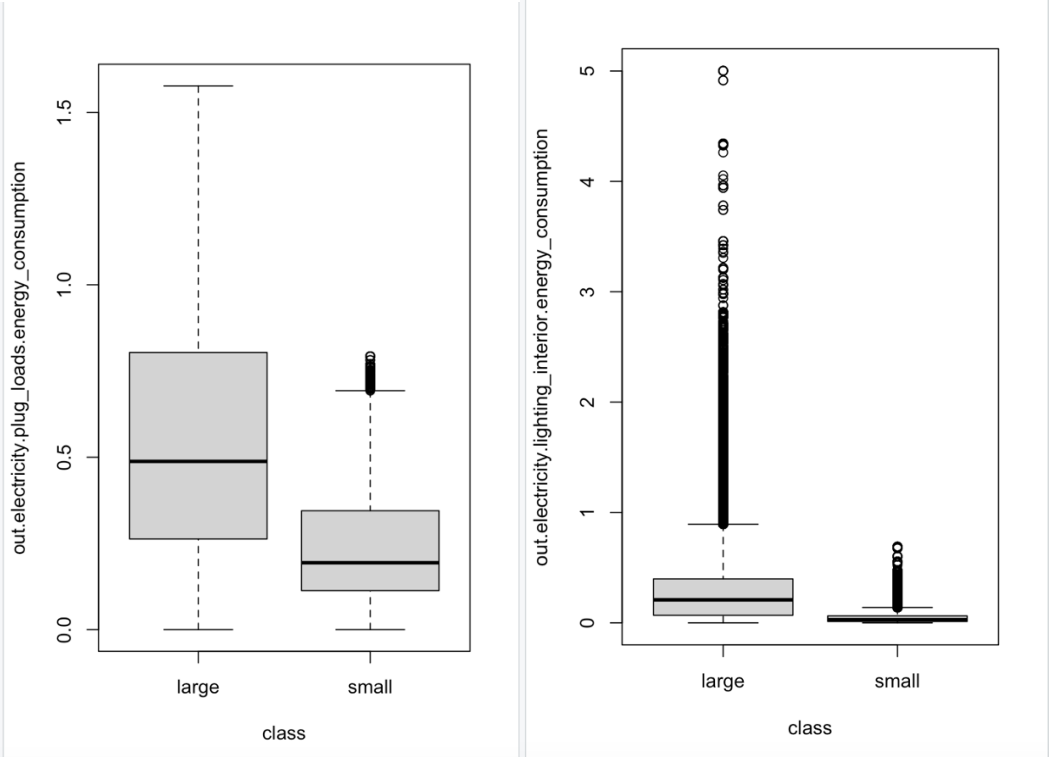
Adding similarly related variables (HVAC Cooling Efficiency, HVAC Cooling Partial Space Efficiency, and HVAC ducts) only improves our R-squared to 0.044, suggesting that a home’s HVAC system is not one of the main drivers of cooling energy usage.

```
Call:
lm(formula = out.electricity.cooling.energy_consumption ~ in.hvac_cooling_efficiency +
    in.hvac_cooling_partial_space_conditioning + in.hvac_has_ducts +
    in.hvac_cooling_type, data = HVACHouseEnergy)

Residuals:
    Min      1Q  Median      3Q     Max
-0.6187 -0.2242 -0.0883  0.1168  7.2982

Coefficients: (4 not defined because of singularities)
                                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                                           0.492995   0.003836 128.518   <2e-16 ***
in.hvac_cooling_efficiencyAC, SEER 13                -0.094644   0.003723 -25.421   <2e-16 ***
in.hvac_cooling_efficiencyAC, SEER 15                -0.058765   0.003978 -14.774   <2e-16 ***
in.hvac_cooling_efficiencyAC, SEER 8                  0.006769   0.008708   0.777    0.437
in.hvac_cooling_efficiencyHeat Pump                  -0.073746   0.003528 -20.902   <2e-16 ***
in.hvac_cooling_efficiencyNone                       -0.110170   0.003704 -29.743   <2e-16 ***
in.hvac_cooling_efficiencyRoom AC, EER 10.7           -0.223395   0.005993 -37.275   <2e-16 ***
in.hvac_cooling_efficiencyRoom AC, EER 12.0           -0.124553   0.006494 -19.180   <2e-16 ***
in.hvac_cooling_efficiencyRoom AC, EER 8.5            -0.225004   0.007985 -28.179   <2e-16 ***
in.hvac_cooling_efficiencyRoom AC, EER 9.8            -0.187096   0.005963 -31.376   <2e-16 ***
in.hvac_cooling_partial_space_conditioning20% Conditioned  0.130536   0.005191  25.145   <2e-16 ***
in.hvac_cooling_partial_space_conditioning40% Conditioned  0.065672   0.005412  12.135   <2e-16 ***
in.hvac_cooling_partial_space_conditioning60% Conditioned  0.071235   0.004820  14.779   <2e-16 ***
in.hvac_cooling_partial_space_conditioning80% Conditioned  0.040857   0.004579   8.923   <2e-16 ***
in.hvac_cooling_partial_space_conditioningNone                NA         NA      NA       NA
in.hvac_has_ductsYes                                  0.118933   0.001953  60.892   <2e-16 ***
in.hvac_cooling_typeHeat Pump                               NA         NA      NA       NA
in.hvac_cooling_typeNone                                    NA         NA      NA       NA
in.hvac_cooling_typeRoom AC                                 NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3748 on 360485 degrees of freedom
Multiple R-squared:  0.04402,   Adjusted R-squared:  0.04399
F-statistic:  1186 on 14 and 360485 DF,  p-value: < 2.2e-16
```

Business Question #4: **Does the square footage of a house correspond to an increase in use of interior lighting energy and plug load energy consumption?**

Solution Approach:

- Filter the static house data for the largest 100 homes and the smallest 100 homes
- Add a column to classify such homes as “large” or “small”
- Import the corresponding energy files and filter for July
- Combine static house data and energy files on building id
- Create preliminary boxplots to view potential relationships:
    - We can see that there does appear to be increased electricity usage for both sources in larger homes



- Create a linear model to measure the significance of the apparent relationships:
    - Our models show that for both energy sources, class is significant. However, the overall R-squared values of our models tell us that we are only capturing 25.17% and 18.79% of the variability in energy consumptions for plug loads and interior lighting by classifying houses into large and small based on square footage. We can expand upon our original business question by considering the ways in which other features of homes may further influence/compound the influence of square footage on energy usage.

```
> lmOutALL<-lm(formula=out.electricity.plug_loads.energy_consumption ~ class, data=ALL)
> summary(lmOutALL)

Call:
lm(formula = out.electricity.plug_loads.energy_consumption ~
    class, data = ALL)

Residuals:
     Min       1Q   Median       3Q      Max
-0.52879 -0.13188 -0.03988  0.13812  1.04821

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.528791   0.000947   558.4   <2e-16 ***
classsmall  -0.294912   0.001339  -220.2   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2543 on 144198 degrees of freedom
Multiple R-squared:  0.2517,    Adjusted R-squared:  0.2517
F-statistic: 4.849e+04 on 1 and 144198 DF,  p-value: < 2.2e-16
```

```
> lmOutALLlights<-lm(formula=ALL$out.electricity.lighting_interior.energy_consumption ~ class, data=ALL)
> summary(lmOutALLlights)

Call:
lm(formula = ALL$out.electricity.lighting_interior.energy_consumption ~
    class, data = ALL)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3375 -0.1295 -0.0318  0.0227  4.6645

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.337514   0.001106   305.1   <2e-16 ***
classsmall  -0.285762   0.001565  -182.7   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2971 on 144198 degrees of freedom
Multiple R-squared:  0.1879,    Adjusted R-squared:  0.1879
F-statistic: 3.336e+04 on 1 and 144198 DF,  p-value: < 2.2e-16
```

- Additional Variable Consideration:
    - **In.occupants**: Number of occupants living in a building
    - **Time**: Time of Day
    - **In.usage_level**: Usage of appliances relative to the national level
- Rerunning the models with these additional variables in mind greatly improved our model accuracy. Both usage level and the number of occupants showed to be significant predictors in addition to the size of the house. Time was not a significant predictor for plug loads. Our new model accounted for 75.18% of the variation in plug load energy usage.

```
Call:
lm(formula = out.electricity.plug_loads.energy_consumption ~
    class + in.usage_level + in.occupants + time, data = ALL)

Residuals:
     Min       1Q   Median       3Q      Max
-0.51128 -0.06587  0.03428  0.08460  0.58476

Coefficients:
                      Estimate Std. Error  t value Pr(>|t|)
(Intercept)         -6.751e-01  7.883e-01   -0.856  0.39182
classsmall          -2.729e-01  8.763e-04 -311.453  < 2e-16 ***
in.usage_levelLow   -5.034e-01  1.066e-03 -472.109  < 2e-16 ***
in.usage_levelMedium -3.695e-01  9.330e-04 -395.989  < 2e-16 ***
in.occupants10+      3.507e-01  5.560e-03   63.082  < 2e-16 ***
in.occupants2        2.646e-03  1.016e-03    2.604  0.00922 **
in.occupants3        8.086e-02  1.353e-03   59.787  < 2e-16 ***
in.occupants4        5.583e-02  1.372e-03   40.703  < 2e-16 ***
in.occupants5        8.210e-02  1.955e-03   41.987  < 2e-16 ***
in.occupants6        2.451e-01  2.922e-03   83.876  < 2e-16 ***
in.occupants7        3.503e-01  4.002e-03   87.532  < 2e-16 ***
in.occupants9        2.829e-01  5.560e-03   50.890  < 2e-16 ***
time                 9.502e-10  5.147e-10    1.846  0.06487 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1464 on 144187 degrees of freedom
Multiple R-squared:  0.7518,    Adjusted R-squared:  0.7518
F-statistic: 3.64e+04 on 12 and 144187 DF,  p-value: < 2.2e-16
```

- Our model for interior lighting improved as well, though we did not see the same improvement as we saw in plug loads. Here our R-squared only increased to 0.2145.

```
Call:
lm(formula = ALL$out.electricity.lighting_interior.energy_consumption ~
    class + time + in.occupants + in.usage_level, data = ALL)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5259 -0.1250 -0.0362  0.0372  4.6822

Coefficients:
                       Estimate Std. Error  t value Pr(>|t|)
(Intercept)          -8.678e+00  1.573e+00   -5.518 3.44e-08 ***
classsmall           -3.049e-01  1.748e-03 -174.418  < 2e-16 ***
time                  5.916e-09  1.027e-09    5.762 8.34e-09 ***
in.occupants10+      -1.479e-01  1.109e-02  -13.337  < 2e-16 ***
in.occupants2        -5.179e-03  2.028e-03   -2.554   0.0106 *
in.occupants3         1.614e-02  2.698e-03    5.983 2.20e-09 ***
in.occupants4        -8.379e-02  2.736e-03  -30.620  < 2e-16 ***
in.occupants5        -4.092e-02  3.901e-03  -10.491  < 2e-16 ***
in.occupants6        -3.734e-04  5.830e-03   -0.064   0.9489
in.occupants7        -1.268e-01  7.985e-03  -15.887  < 2e-16 ***
in.occupants9         3.837e-01  1.109e-02   34.595  < 2e-16 ***
in.usage_levelLow     1.930e-02  2.127e-03    9.075  < 2e-16 ***
in.usage_levelMedium -6.057e-02  1.861e-03  -32.543  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2922 on 144187 degrees of freedom
Multiple R-squared:  0.2145,    Adjusted R-squared:  0.2145
F-statistic:  3282 on 12 and 144187 DF,  p-value: < 2.2e-16
```

- Though it was not an initial part of our business question, since this combination of variables seemed to be good predictors for our key drivers, we made an additional model for cooling energy, which resulted in an R-squared value of 0.3714, which was highest R-squared for any of the cooling models we tried.

```
Residual standard error: 0.3473 on 720941 degrees of freedom
Multiple R-squared:  0.3714,    Adjusted R-squared:  0.3714
F-statistic:  7344 on 58 and 720941 DF,  p-value: < 2.2e-16
```

Business Question #5: **Do the top contributors to energy usage present a meaningful relationship with temperature?**

Solution Approach:

- Filter out the average houses (median sqft 1690) in the static house dataset
- Filtered out for the attributes that drive the energy consumption the most by taking the predictions done earlier. Also filtered out for the counties having houses which had high energy consumption.
- Took a subset of 250 buildings and merged in the energy files for all the buildings in the dataset.
- Merged the weather files of the locations of the buildings that were chosen.

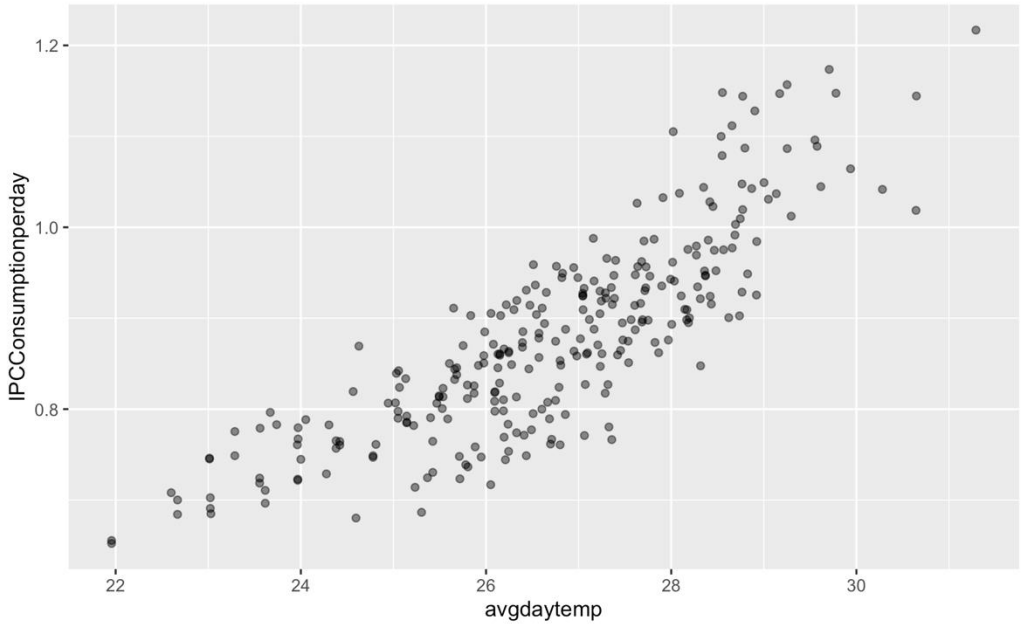
	in.county	day	avgdaytemp	avghumidity
1	G4500070	01	26.53261	78.47565
2	G4500070	02	27.04696	76.94043
3	G4500070	03	28.08870	73.09609
4	G4500070	04	27.39957	73.79652
5	G4500070	05	27.38087	74.75391
6	G4500070	06	26.43478	76.85826
7	G4500070	07	23.28652	84.25261
8	G4500070	08	23.96217	57.61913

We summed up the temperature and humidity of the counties on daily basis.

	in.county	day	TotalConsumptionperday	IPCConsumptionperday
1	G4500070	01	1.277614	0.9364350
2	G4500070	02	1.255098	0.9249033
3	G4500070	03	1.363493	1.0373517
4	G4500070	04	1.280783	0.9635850
5	G4500070	05	1.245868	0.9470883
6	G4500070	06	1.249543	0.9308283
7	G4500070	07	1.036375	0.7488033

In the energy file, we summed up the data by total energy consumption of each attribute and the energy consumption of the top 3 key drivers on daily basis.

Merged in the data of weather and energy consumption to check if the top contributors of energy usage have any relation with the temperature.



We can see that temperature and the top contributors of energy usage have a strong positive linear relationship. As it is seen, the higher the temperature, the higher the energy usage.

We determined that temperature is a significant predictor of energy usage with p-value of <2e-16.

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.4584784  0.0769434  -5.959 7.74e-09 ***
avgdaytemp   0.0526941  0.0022243  23.690  < 2e-16 ***
avghumidity -0.0009167  0.0004086  -2.244   0.0257 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05806 on 276 degrees of freedom
Multiple R-squared:  0.7193,    Adjusted R-squared:  0.7172
F-statistic: 353.6 on 2 and 276 DF,  p-value: < 2.2e-16
```

From the linear prediction model, we can say that whenever there is a 1-degree rise in temperature, the energy usage increases by 0.052 kWh for our key drivers in an average home.

Model Selection

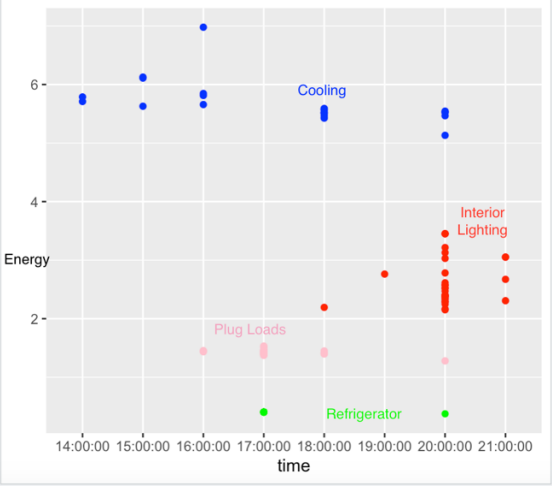
Our best model considered square footage, the number of occupants, the usage level, and the time of day to predict for plug load energy consumption, which we found in answering Business Question 4 regarding home size. By taking a subset of the largest and smallest homes in the dataset and classifying them into ‘large’ and ‘small’, we were able to determine that size was a significant predictor. However, it did not account for a large proportion of the variability in energy usage, so we considered other variables in our model, which is how we came to create a model for square footage, the number of occupants, the usage level, and the time of day. To make a useful prediction model for eSC, it was important to create a model which had both a high R-squared value but was also generalized enough to be applicable for prediction purposes. In other words, it was important to find the “common denominators” amongst all the static house variables to find which variables in combination contributed the most to energy consumption.

We felt it was important to include similar models for cooling energy and interior lighting energy because even though the R-squared values were not as high, we felt they still captured a relatively significant amount of variability and that these three energy sources were important to consider in conjunction with one another given the large proportion of energy consumption they make up.

Since time was a significant predictor in our models, we related the time of day to the weather data by averaging the hourly temperature across all counties in our data set. As expected, the hottest time of the day was in the early afternoon.

The peak daily temperatures were aligned with the times we observed the maximum amount of energy usage for our key drivers of consumption, specifically for cooling energy.

	time	Avg_Hourly_Temp
1	14:00:00	30.70004
2	15:00:00	30.61471
3	13:00:00	30.56206
4	16:00:00	30.26304
5	12:00:00	30.09978
6	17:00:00	29.84405



We related the time attribute with the average hourly temperature in our model so that we could manipulate the weather data to see how energy consumption would change at a particular hour of the day when average hourly temperature increases by 5 degrees.

Model Prediction

We created a data frame with example values for our predictors which we could test against the models to predict for different geographic regions and attributes and compare to the actual values of energy consumption.

Our plug loads model predicted usage for an **885sqft house with 3 occupants, medium usage level, in county G4500910 at midnight** to be 0.238 kWh. The actual value of energy consumption was 0.226 kWh, which provided evidence that our model was predicting within a reasonable range of the actual values.

Ex:

```
#Plug Loads:
lmPlugLoads<-lm(formula=out.electricity.plug_loads.energy_consumption ~ in.sqft + in.occupants + in.usage_level + Avg_Hourly_Temp + in.county, data=BaseForModel)
summary(lmPlugLoads)

predPlugLoadsDF <- data.frame(in.county = "G4500910", in.sqft= 885, in.occupants= "3", in.usage_level= "Medium",  Avg_Hourly_Temp= HottestHour$Plus5[HottestHour$time == "00:00:00"])
predict(lmPlugLoads, predPlugLoadsDF)
```

Similar tests were done for predicting cooling and lighting usage levels for the same home.

Actual Cooling: 0.316kwh

Predicted Cooling: 0.181kWh

Actual Lighting: 0.031kWh

Predicted Lighting: 0.048kWh

We expected these models to not perform as well, given that they had R-squared values of 0.3714 and 0.2145, respectively.

We then coded each of the three model and prediction data frames into our Shiny App, allowing the user to input values for each predictor.

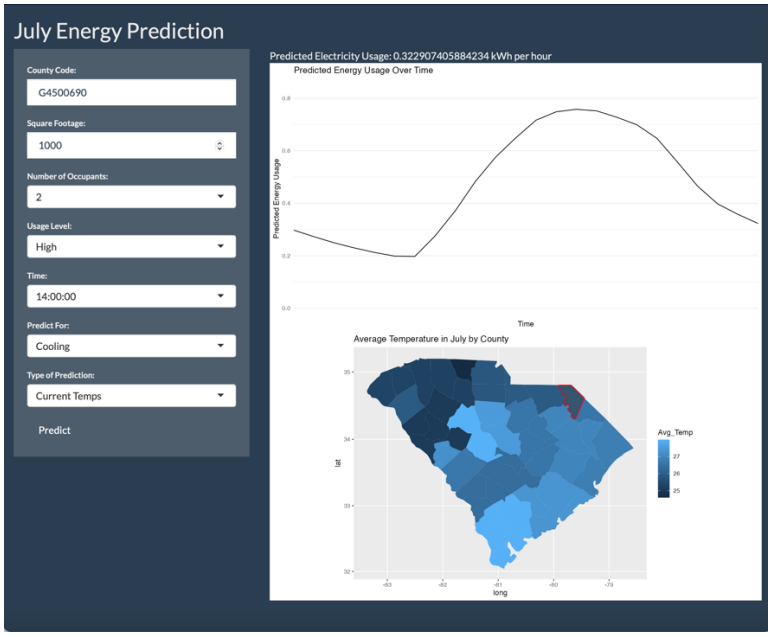
Prediction Results

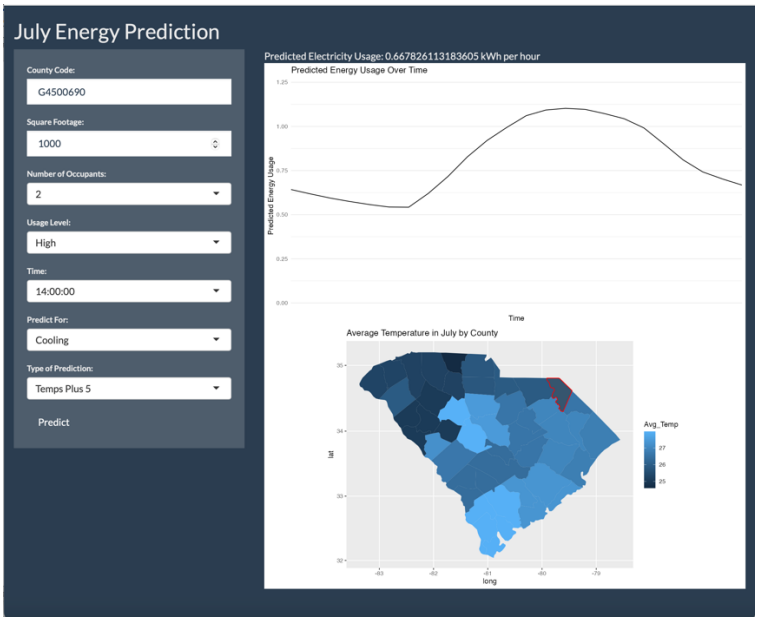
Our Shiny App allows the user to input values for specific attributes in their home and select the type of energy usage they wish to predict for and the time. In addition, the user can see their predicted future usage if the temperatures were to increase 5 degrees next July. In the example case below, cooling energy for this specific house is predicted to increase from 0.323 kWh to 0.668 kWh at the specified hour of 2pm if peak July temperatures increase by 5 degrees. In this case, cooling energy usage for this home at this hour has almost doubled. As the largest consumer of electricity, an increase in cooling energy usage among all houses would likely overwhelm the power grid if energy consumption were not reduced.

Interior lighting doubles in this example, from 0.058 kWh to 0.103 kWh, while plug loads seem least impacted by the increase in temperature, as evidenced by the plot showing a much flatter change in energy usage over time.

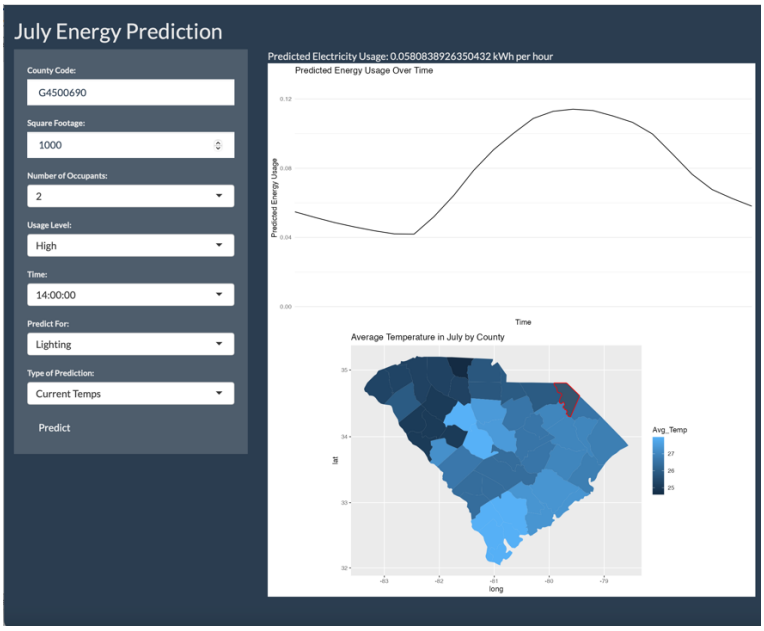
Keeping in mind the impact of global warming and increasing temperatures, cooling energy is predicted to be the most impacted by a warmer July. If all homes increase cooling usage by a similar amount, there is reason to believe that rising temperatures would overwhelm the electric grid.

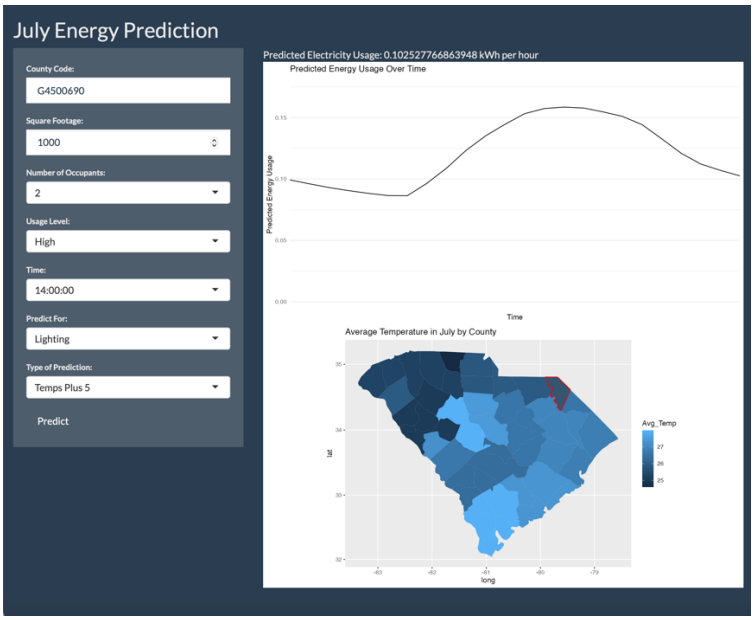
Example Cooling Energy Consumption (Current vs Future Temperatures):


July Energy Prediction
County Code:
G4500690
Square Footage:
1000
Number of Occupants:
2
Usage Level:
High
Time:
14:00:00
Predict For:
Cooling
Type of Prediction:
Current Temps
Predict
Predicted Electricity Usage: 0.322907405884234 kWh per hour
Predicted Energy Usage Over Time
Predicted Energy Usage
Time
Average Temperature in July by County
lat
long
Avg_Temp

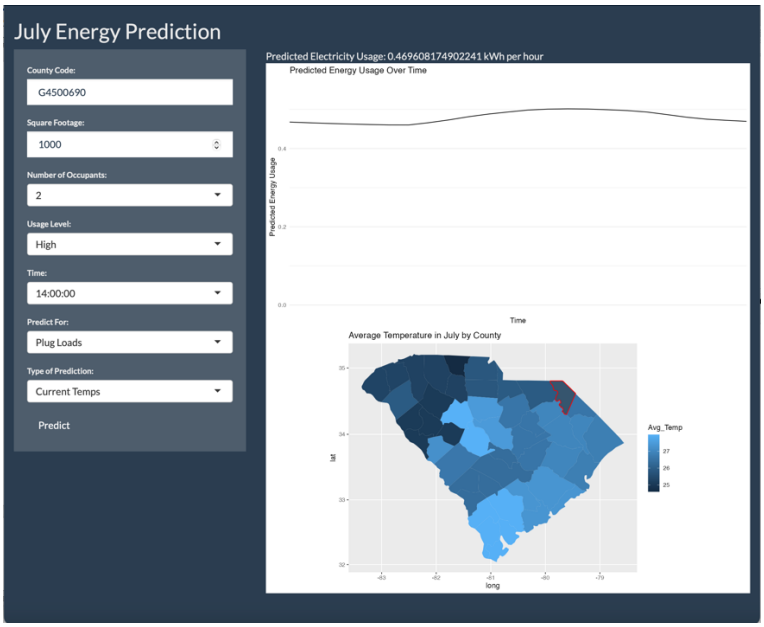

July Energy Prediction
County Code:
G4500690
Square Footage:
1000
Number of Occupants:
2
Usage Level:
High
Time:
14:00:00
Predict For:
Cooling
Type of Prediction:
Temps Plus 5
Predict
Predicted Electricity Usage: 0.667826113183605 kWh per hour
Predicted Energy Usage Over Time
Predicted Energy Usage
Time
Average Temperature in July by County
lat
long
Avg_Temp

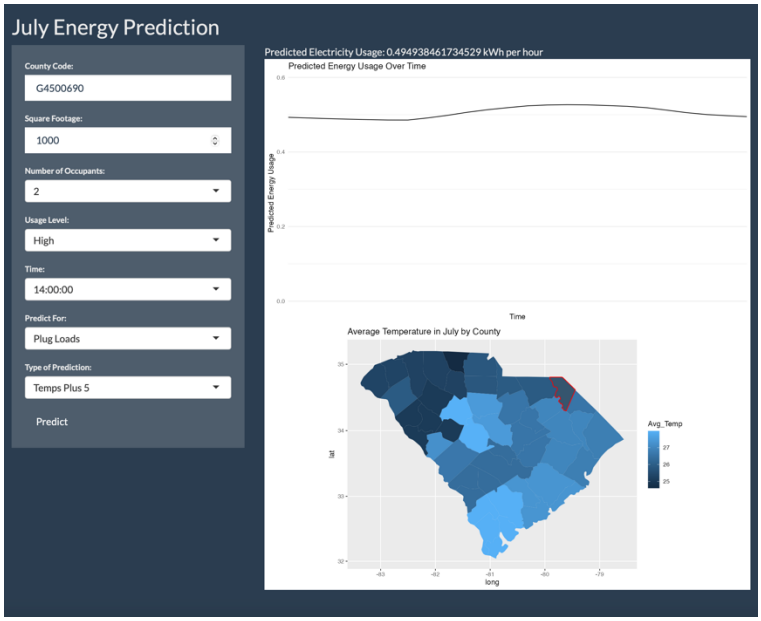
Example Interior Lighting Consumption (Current vs Future Temps):





Example Plug Loads Consumption (Current vs Future Temps):



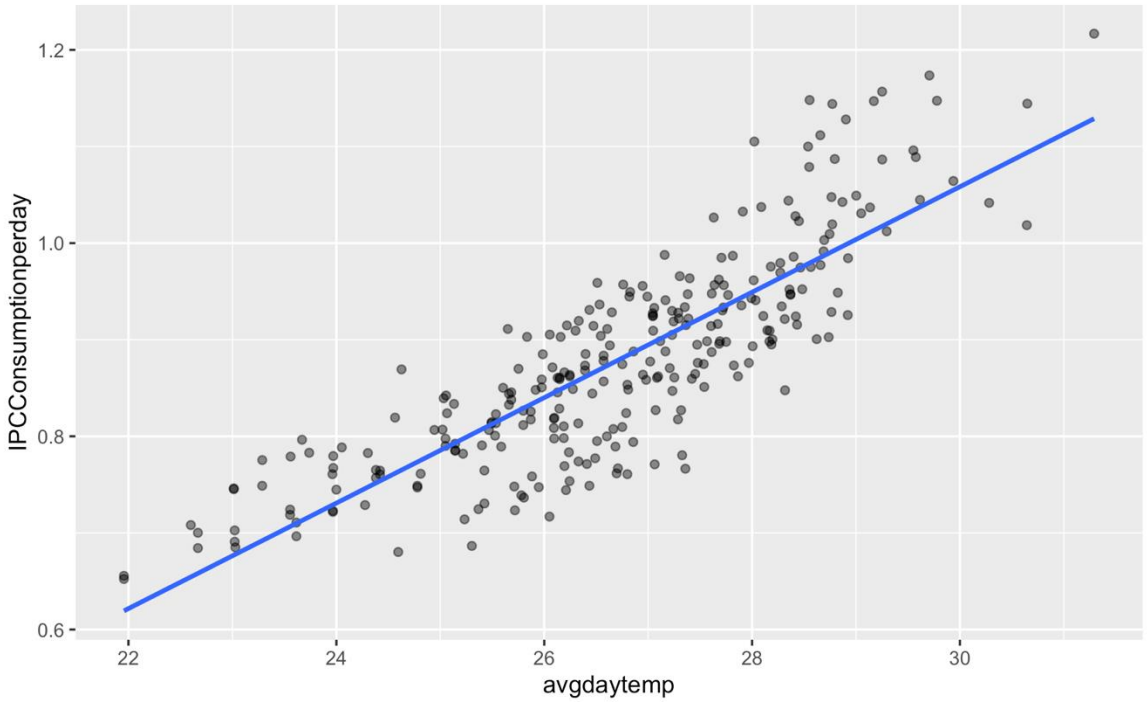


Insights and Recommendations

We set out to determine the leading sources of energy consumption among eSC's customers to provide a recommendation for how to incentivize customers to use less energy (specifically during the peak month of July). Our team was able to determine that the top sources of energy consumption across all homes were cooling energy, interior lighting and plug loads. We

hypothesized which home attributes might contribute to an increase in the consumption of these energy sources and how rising July temperatures would further compound on this increase in energy. We tested models which considered the effect of a home's HVAC system, local weather, number of occupants, square footage, relative energy consumption compared to the national average, the time of day, and the air leakage rates (via ducts, windows, etc), both independently and together. Using a random subset of 1000 homes, we were able to model 75.18% of the variability in plug load energy consumption by considering the size, number of occupants, usage levels, and time of day. This was our best model, followed by a similar model to predict for cooling energy and interior lighting energy using the same predictors. We felt it was important to model for these outcome variables given the large proportion of total energy consumption each source accounts for.

Evaluation of Future Peak Energy Demand



In order to understand the big picture of how rising temperature affect peak energy demand, we considered the combined consumption of interior lighting, plug loads, and cooling energy (called IPC energy) per day for an average sized house.

A strong, positive linear relationship is observed between energy consumption of these three sources and temperature.

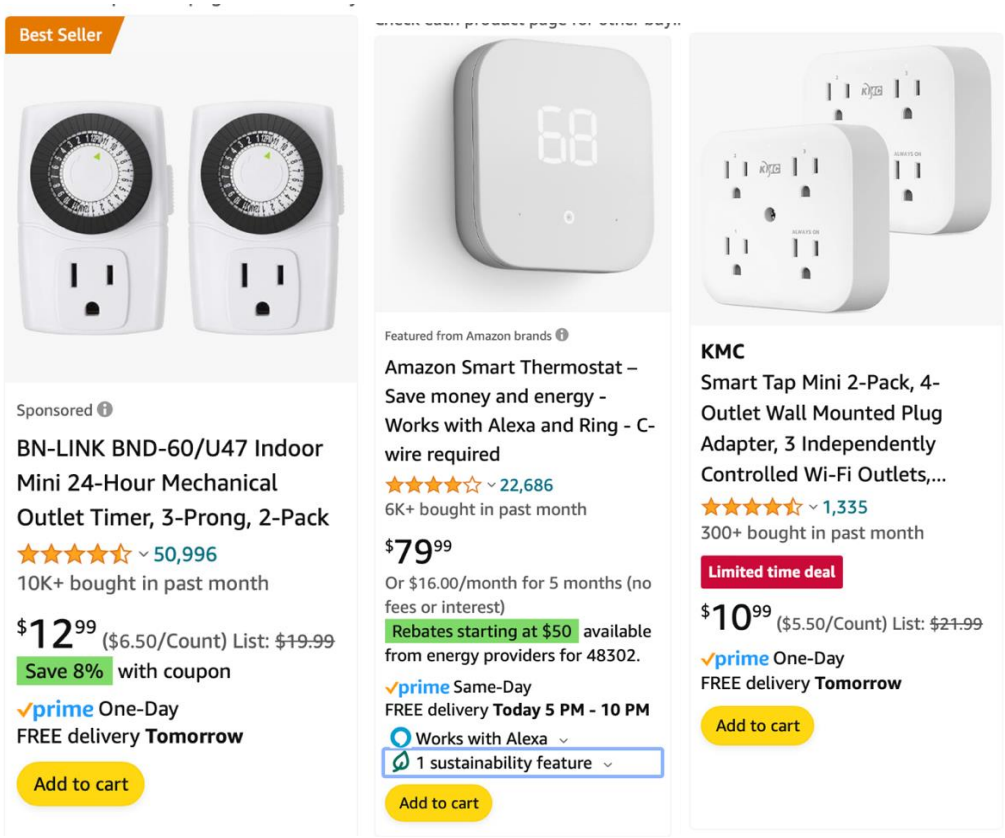
Holding all static house attributes constant and considering the median home size of 1690sqft, a model for IPC energy consumption based solely on average daily temperature reveals that for every 1 degree increase in temperature there is a 0.052kWh rise in energy consumption of cooling, interior lighting and plug loads per hour.

In the context of a hypothetical 5 degree increase in July temperatures, the average house is expected to increase energy usage by 0.26kWh per hour. If a similar trend was exhibited in all the 5710 homes, energy usage of the key driver variables would increase by 1,484.6 kWh per hour.

Recommendation

Based on the results of our investigation, we would recommend that eSC promote the reduction of energy consumption by supplying their customers with smart plugs, light timers, and smart thermostats. The number of smart plugs and light timers would be related to the number of occupants in the home, while each home might need one thermostat per floor of their home. Specifically, we would recommend that the “High” energy users be the primary targets of this rollout. We found that usage level compared to the national average was a significant predictor of cooling, interior lighting, and plug load energy consumption. Similarly, time was a significant predictor in determining cooling and lighting usage, which fits with the fact that temperatures are cooler at night, and we need light to see at night. Having automatic timers to shut off lights and outlets, as well as adjusting the temperature when consumers are out of the house during the day (when most cooling energy is used) or asleep (when many unused electronics would still be consuming energy) would likely make a significant difference in energy consumption during peak months.

The changes are cost effective too. Below are some examples of outlet adaptors and smart systems on the market which eSC could incentivize customers to purchase or provide to customers.



We can design an experiment to give smart plugs or thermostats to designated group of homes (High Usage homes for example) and monitor their energy usage for the key drivers over a fixed period (potentially Summer or Winter months when we expect temperatures to be most extreme). Homes matching the targeted attributes would be divided into test and control sets. The test group would be provided with the smart plugs, while the control group would have energy monitored under normal conditions. Using the prediction model we made, we can compare the actual energy usage to the predicted energy usage (under normal conditions with no smart devices). We would be able to measure both the accuracy of our prediction model and determine whether there is a significant difference in the energy consumption of like homes with and without smart devices. Another point of investigation that was not part of this analysis was the cost of electricity usage. Valuable insight would be gained if we could understand how cost effective providing smart devices/incentivizing customers to buy smart devices would be in the long run. From a consumer perspective, buying smart devices would help them save energy and money on their electric bill. The advantage for eSC is that decreased energy usage would eliminate the need to build another costly facility.