

---

# ENERGY CONSUMPTION ANALYSIS AND PREDICTION ON HOUSING DATA

---

Submission to Erik Anderson  
Group 4

# Table of Contents

<b>INTRODUCTION</b>	3
Background:	3
Context:	3
Scope:	3
<b>BUSINESS QUESTIONS</b>	4
<b>DATA DESCRIPTION</b>	4
Static House Data:	4
Energy Usage Data:	4
Meta Data:	5
Weather Data:	5
Dependent Variable:	5
Energy Consumption:	5
Independent Variables:	6
Weather Variables:	6
Building Characteristics:	6
Temporal Variables:	6
Geographic Variables:	7
Appliance Usage:	7
<b>DATA MERGING, ACQUISITION, CLEANING, TRANSFORMATION, MUNGING</b>	7
<b>DESCRIPTIVE STATISTICS</b>	12
<b>DATA MODELLING</b>	21
Modelling Approach:	21
Classification Models:	22
Linear Regression Model:	22
Light Gradient Boosting Model	22
Support Vector Regression Model:	23
Interpretation of models:	24
<b>VISUALIZATION</b>	25
<b>SHINY APP</b>	26
Title Panel:	26
Input Controls:	26
Predict Button:	27
Predicted Value Display:	27
<b>CONCLUSION</b>	28
<b>CONTRIBUTIONS OF TEAM</b>	29

## List of figures

Figure 1: summary of energy consumption before data cleaning	5
Figure 2: data merging	8
Figure 3: reading the data	9
Figure 4: structure of the data	10
Figure 5: data filtration	11
Figure 6: data transformation	11
Figure 7: summary of energy consumption after data cleaning	11
Figure 8: bar plot between presence of rooftop PV and energy consumption	12
Figure 9: bar plot between climate zone and energy consumption	13
Figure 10: energy consumption patterns over 24 hours	14
Figure 11: scatter plot between energy consumption and dry bulb temperature	14
Figure 12: boxplot of energy consumption by heating set points	15
Figure 13: boxplot of energy consumption by heating set points	15
Figure 14: bar plot of energy consumption in the cities	16
Figure 15: boxplot of energy consumption by types of wall insulation	16
Figure 16: boxplot of energy consumption by types of cooking range	17
Figure 17: boxplot of energy consumption by types of lighting	18
Figure 18: sample average temperature map of each county in south carolina and some areas in north carolina at 2 pm on July 13, 2018	19
Figure 19: simulated average temperature map of each county in South Carolina and some areas in North Carolina if increases by 5 degree celsius	20
Figure 20: simulated average temperature map of counties that fall in hot-humid zone	20
Figure 21: linear regression model output	22
Figure 22: light gradient boosting model output	23
Figure 23: support vector regression model output	24
Figure 24: energy consumption across various energy sources	25
Figure 25: user interface of energy consumption in the shiny app	27

## INTRODUCTION

### **Background:**

The upcoming summer is anticipated to bring substantial temperature elevations, particularly in July, raising concerns about the strain on the power grid. Given eSC's role as a primary energy provider, there is a critical need to proactively address the potential challenges associated with increased energy demand. Our team has been engaged to analyze historical energy consumption data, weather patterns, and other relevant factors to develop robust forecasting models. By doing so, we aim to assist eSC in managing energy demand effectively, reducing operational costs, and advancing sustainability goals.

### **Context:**

Global warming poses a significant challenge, leading to rising temperatures worldwide, particularly during the summer season. As temperatures continue to increase, energy providers face mounting concerns regarding grid stability and escalating energy demand. In this context, eSC, as a prominent electricity supplier for residential properties in the region, must address the challenges posed by heightened energy consumption effectively. Failure to do so could result in severe consequences such as widespread blackouts and disruptions in service provision.

### **Scope:**

Our analysis focuses on understanding energy consumption patterns in South Carolina and select areas of North Carolina, particularly during the summer months. We aim to examine data related to residential properties served by eSC, including information on household characteristics, energy usage, and weather conditions. By leveraging this data, we seek to develop predictive models to forecast future energy demand accurately. Our analysis will provide actionable insights to help eSC anticipate and manage potential energy shortages, thereby enhancing grid stability and ensuring reliable service delivery to customers.

## BUSINESS QUESTIONS

The main goals we aim to address in this project are as follows:

- What are the main factors contributing to energy consumption in residential properties served by eSC?
- How do our efforts to reduce energy consumption align with our sustainability goals and commitments?
- How does energy consumption vary across different counties or locations within the service area of eSC?
- What solutions can we propose to eSC to support their energy efficiency projects and initiatives?

## DATA DESCRIPTION

The dataset provided consists of four main components:

### **Static House Data:**

This dataset contains basic information about single-family houses served by eSC. It includes attributes such as building ID, house size, geographic location, and various characteristics of the house that remain constant over time. The file is stored in Parquet format and contains approximately 5,000 entries.

### **Energy Usage Data:**

Energy consumption data for each house is collected hour-by-hour and stored in individual files. Each file represents the energy usage profile of a specific house, with the building ID serving as the filename. The dataset consists of calibrated and validated energy usage, detailing consumption from various sources such as air conditioning systems and dryers.

## Meta Data:

A data dictionary file provides descriptions of the fields used across the different housing data files. This human-readable file aids in understanding the attributes present in both the static house data and the energy usage data.

## Weather Data:

Hourly weather information is available for each geographic area, with one file per county. The weather data is stored based on county codes, allowing for analysis of weather patterns that may influence energy consumption.

## Dependent Variable:

### Energy Consumption:

This is the variable we are trying to understand and predict. It represents the total amount of energy used by households within a specific time period, typically measured in kilowatt-hours (kWh) units.

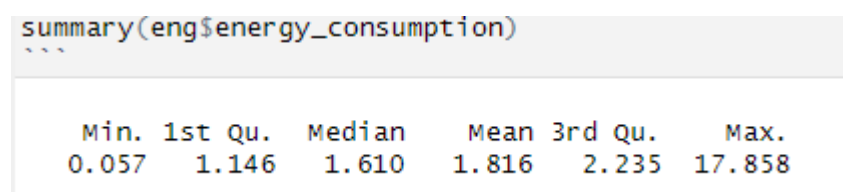


Figure 1: summary of energy consumption before data cleaning

According to the figure, the energy consumption summary shows a minimum of 0.057, 1st quartile at 1.146, median at 1.610, mean at 1.816, 3rd quartile at 2.235, and a maximum of 17.858. These statistics describe the range and central tendencies of energy consumption in the dataset, with 1.816 being the average energy consumption, and 1.610 as the median, indicating that half of the data points have energy consumption lower than this value. The maximum value of 17.858 highlights the extreme energy consumption in some instances.

## **Independent Variables:**

### **Weather Variables:**

#### **Temperature:**

Both dry bulb temperature and relative humidity can influence energy usage, particularly for heating and cooling systems.

#### **Solar Radiation:**

The amount of sunlight received can affect energy demand for lighting and cooling.

#### **Wind Speed:**

Higher wind speeds may impact heating systems and ventilation.

### **Building Characteristics:**

#### **House Size:**

The square footage of a house can affect its heating and cooling requirements.

#### **Building Type:**

Whether a house is single-family, multi-family, or commercial can influence energy usage patterns.

#### **Insulation Level:**

The quality of insulation can impact the efficiency of heating and cooling systems.

#### **Appliance Efficiency:**

The energy efficiency ratings of appliances such as HVAC systems, refrigerators, and water heaters can affect overall energy consumption.

### **Temporal Variables:**

#### **Time of Day:**

Energy usage may vary throughout the day due to factors like occupancy patterns and appliance usage.

**Day of Week:**

Weekdays and weekends may exhibit different energy consumption patterns.

**Season:**

Energy demand typically fluctuates with seasonal changes in weather and daylight hours.

**Geographic Variables:****Location:**

The geographic location of a house can influence weather patterns, building codes, and energy prices.

**County or Region:**

Energy consumption may vary between different counties or regions due to factors like climate, population density, and economic activity.

**Appliance Usage:****Appliance Types:**

The presence and usage patterns of energy-intensive appliances such as air conditioners, heaters, refrigerators, and washing machines can significantly impact energy consumption.

**DATA MERGING, ACQUISITION, CLEANING, TRANSFORMATION, MUNGING**

This section focuses on how we gather, combine, tidy up, change, and organize our data to make it useful and understandable for analysis.



```

#creating the links for each house energy data
house$details <- paste("https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/2023-houseData/",
                      house$building_id, ".parquet", sep = "")
#creating the links for each county for which has weather data
house$weather_data <- paste("https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/weather/2023-weather-data/",
                            house$in_county, ".csv", sep = "")
#view(house)
house <- house[house$in_building_america_climate_zone == "Hot-Humid",]
house <- house[house$in_has_pv == "No",]
#selecting data of the july month and sampling 50 rows and merging it with the
#weather data for 1st building
energy_data <- read_parquet(house$details[1])
energy_data <- energy_data[order(energy_data$time),]
energy_data$building_id <- house$bldg_id[1]
energy_data <- energy_data[energy_data$time >= as.POSIXct("2018-07-01") &
                          energy_data$time <= as.POSIXct("2018-07-31 23:00:00"),]
energy_data <- energy_data[as.numeric(format(energy_data$time, "%H")) >= 14 &
                          as.numeric(format(energy_data$time, "%H")) <= 18,]
energy_data <- na.omit(energy_data)
weather_data <- read_csv(house$weather_data[1])
energy_data <- merge(x=energy_data, y=weather_data, by.x = "time", by.y = "date_time")
#view(energy_data)
#Creating a loop to add 5 hours a day for july month from each house to energy data
for(i in 2:nrow(house)){
  energy_data1 <- read_parquet(house$details[i])
  energy_data1 <- energy_data1[order(energy_data1$time),]
  energy_data1$building_id <- house$bldg_id[i]
  energy_data1 <- energy_data1[energy_data1$time >= as.POSIXct("2018-07-01") &
                              energy_data1$time <= as.POSIXct("2018-07-31 23:00:00"),]
  energy_data1 <- energy_data1[as.numeric(format(energy_data1$time, "%H")) >= 14 &
                              as.numeric(format(energy_data1$time, "%H")) <= 18,]
  energy_data1 <- na.omit(energy_data1)
  weather_data1 <- read_csv(house$weather_data[i])
  energy_data1 <- merge(x=energy_data1, y=weather_data1, by.x = "time", by.y = "date_time")
  energy_data <- rbind(energy_data, energy_data1)
}
view(energy_data)
view(df)
df$hour <- as.numeric(format(energy_data$time, "%H"))
#write.csv(df, "combined_house_data.csv", row.names = FALSE)

```

Figure 2: data merging

- The code imports house information, filters houses in the "Hot-Humid" climate zone without photovoltaic systems, and merges energy consumption data with corresponding weather data for each house. It iterates through each house, selecting and merging July 2018 energy and weather data, appending the results into a single dataframe. The merged dataframe, energy\_data, contains energy consumption and weather information for all selected houses and time periods. However, there's a typo in the attempt to view the dataframes, and an erroneous attempt to add a column. Additionally, there's commented-out code to write the merged data to a CSV file named combined\_house\_data.csv.
- House and energy data were provided in Parquet format, a CSV file type optimized for storage. We utilized the read\_parquet() function from the Arrow package to import these files. Subsequently, we filtered houses based on their location in a "Hot-Humid" environment and their absence of solar power usage. Our analysis focused on the time

```
# Importing Data
house <- read_parquet("https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/static_house_info.parquet")
combined_data <- read_csv("C:/Users/tharu/Downloads/combined_house_data (1).csv")
energy_data <- combined_data
str(distinct(combined_data))
# Merge energy_data with house data
energy_data <- merge(x = energy_data, y = house, by.x = "building_id", by.y = "bldg_id")
View(energy_data)

# Calculate total energy consumption
energy_consumption <- rowSums(energy_data[, 3:44])

# Create data frame with energy information
eng <- data.frame(
  energy_data[, 1:2],
  energy_consumption,
  energy_data[, 45:221]
)

# Display the structure of eng
str(eng)
```

```
str(eng)
"data.frame":    252030 obs. of  180 variables:
 $ building_id   : num  390 390 390 390 390 390 390 390 390 390 ...
 $ time         : POSIXct, format: "2018-07-01 14:00:00" "2018-07-01 15:00:00" "2018-07-01 16:00:00" ...
 $ energy_consumption : num  1.49 1.85 2.21 2.31 2.04 ...
 $ dry_bulb_temperature...c: num  31.4 32.8 32.2 31.9 30.6 ...
 $ Relative_Humidity... : num  67.8 59.6 61.6 60.5 67.5 ...
 $ wind_speed...m.s.: num  3.1 4.6 5.1 5.65 4.6 2.55 3.1 3.8 3.85 4.1 ...
 $ wind_direction_deg : num  115 140 190 170 160 100 150 140 135 140 ...
 $ Global.Horizontal.Radiation..W.m2: num  858 872 701 560 367 ...
 $ Direct.Normal.Radiation..W.m2: num  664 840 668 708 606 ...
 $ Diffuse.Horizontal.Radiation..W.m2: num  218 120 180 118 95 ...
 $ hour          : num  14 15 16 17 18 14 15 16 17 18 ...
 $ upgrade      : fnt  10 10 10 10 10 10 10 10 10 10 ...
 $ weight       : num  242 242 242 242 242 ...
 $ applicability: logi TRUE TRUE TRUE TRUE TRUE TRUE ...
 $ fn_sift      : num  2176 2176 2176 2176 2176 2176 2176 2176 2176 ...
 $ fn_ahs_region: chr  "Non-CBSA South Atlantic" "Non-CBSA south Atlantic" "Non-CBSA south Atlantic" ...
 $ fn_cbsa      : chr  "3A" "3A" "3A" "3A" ...
 $ fn_ashrae_lecc_climate_zone_2004 : chr  "3A" "3A" "3A" "3A" ...
 $ fn_bathroom_spot_vent_hour : chr  "Hour5" "Hour5" "Hour5" "Hour5" ...
 $ fn_bedrooms  : fnt  2 2 2 2 2 2 2 2 2 ...
 $ fn_building_america_climate_zone : chr  "Hot-Humid" "Hot-Humid" "Hot-Humid" "Hot-Humid" ...
 $ fn_ccc_climate_zone : chr  "None" "none" "None" "None" ...
 $ fn_ceiling_fan_efficiency : chr  "Standard Efficiency" "Standard Efficiency" "Standard Efficiency" ...
 $ fn_census_division : chr  "South Atlantic" "South Atlantic" "South Atlantic" "South Atlantic" ...
 $ fn_census_division_rec : chr  "South Atlantic" "South Atlantic" "South Atlantic" "South Atlantic" ...
 $ fn_census_region : chr  "South" "south" "South" "South" ...
 $ fn_city      : chr  "SC, Goose Creek" "SC, Goose Creek" "SC, Goose Creek" "SC, Goose Creek" ...
 $ fn_clothes_dryer : chr  "Electric, 100% usage" "Electric, 100% usage" "Electric, 100% usage" ...
 $ fn_clothes_washer : chr  "EnergyStar, 100% usage" "Energystar, 100% usage" "energystar..." ...
 $ fn_cooking_range : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ fn_cooling_setpoint : chr  "Electric, 100% usage" "Electric, 100% usage" "Electric, 100% usage" ...
 $ fn_cooling_setpoint_has_offset : chr  "72F" "72F" "72F" "72F" ...
 $ fn_cooling_setpoint_offset_magnitude : chr  "Yes" "Yes" "Yes" "Yes" ...
 $ fn_cooling_setpoint_offset_period : chr  "Sf" "Sf" "Sf" "Sf" ...
 $ fn_county    : chr  "Night Setback" "Night Setback" "Night Setback" "Night Setback" ...
 $ fn_county_and_zip : chr  "Not Applicable" "not applicable" "Not Applicable" "not applicable" ...
 $ fn_dehumidifier : chr  "G4500150" "G4500150" "G4500150" "G4500150" ...
 $ fn_dishwasher : chr  "G4500150, G45001204" "G4500150, G45001204" "G4500150, G45001204" ...
 $ fn_door_area : chr  "None" "None" "None" "None" ...
 $ fn_ducts     : chr  "290 rated kwh, 100% usage" "290 rated kwh, 100% usage" "290 rated kwh, 100% usage" ...
 $ fn_door_sweeps : chr  "20 R=2" "20 R=2" "20 R=2" "20 R=2" ...
 $ fn_doors     : chr  "Fiberglass" "fiberglass" "Fiberglass" "Fiberglass" ...
 $ fn_ducts     : chr  "30N Leakage, R-6" "30N Leakage, R-6" "30N Leakage, R-6" "30N Leakage, R-6"
```

```

$ in.eaves : chr "2 ft" "2 ft" "2 ft" "2 ft" ...
$ in.electric_vehicle : chr "none" "none" "none" "none" ...
$ in.emissions_electricity_folders : chr "data/cambium/LRMR_MidCase_15_2025start,data/cambium/LRMR_
025start,data/cambium/LRMR_95Decarby" | __truncated__ "data/cambium/LRMR_MidCase_15_2025start,data/cambium/LRMR_LowRe
rt,data/cambium/LRMR_95Decarby" | __truncated__ "data/cambium/LRMR_MidCase_15_2025start,data/cambium/LRMR_LowReCost_15_202
a/cambium/LRMR_95Decarby" | __truncated__ "data/cambium/LRMR_MidCase_15_2025start,data/cambium/LRMR_LowReCost_15_202
ium/LRMR_95Decarby" | __truncated__ ...
$ in.emissions_electricity_units : chr "kg/Mwh,kg/Mwh,kg/Mwh,kg/Mwh" "kg/Mwh,kg/Mwh,kg/Mwh,kg/Mwh"
h,kg/Mwh,kg/Mwh" "kg/Mwh,kg/Mwh,kg/Mwh,kg/Mwh" ...
$ in.emissions_electricity_values_or_filepaths : chr "/11b/resources/data/cambium/LRMR_MidCase_15_2025start/SRVCC.cs
ources/data/cambium/LRMR_LowReCost_15" | __truncated__ "/11b/resources/data/cambium/LRMR_MidCase_15_2025start/SRVCC.cs
s/data/cambium/LRMR_LowReCost_15" | __truncated__ "/11b/resources/data/cambium/LRMR_MidCase_15_2025start/SRVCC.csv,/11b/
a/cambium/LRMR_LowReCost_15" | __truncated__ "/11b/resources/data/cambium/LRMR_MidCase_15_2025start/SRVCC.csv,/11b/resi
ium/LRMR_LowReCost_15" | __truncated__ ...
$ in.emissions_fossil_fuel_units : chr "lb/Mbtu,lb/Mbtu,lb/Mbtu,lb/Mbtu" "lb/Mbtu,lb/Mbtu,lb/Mbtu,lb/Mbtu"
tu,lb/Mbtu,lb/Mbtu,lb/Mbtu" "lb/Mbtu,lb/Mbtu,lb/Mbtu,lb/Mbtu" ...
$ in.emissions_fuel_oil_values : chr "195.9,195.9,195.9,195.9" "195.9,195.9,195.9,195.9" "195.9"
5.9" "195.9,195.9,195.9,195.9" ...
$ in.emissions_natural_gas_values : chr "147.3,147.3,147.3,147.3" "147.3,147.3,147.3,147.3" "147.3"
7.3" "147.3,147.3,147.3,147.3" ...
$ in.emissions_propane_values : chr "177.8,177.8,177.8,177.8" "177.8,177.8,177.8,177.8" "177.8"
7.8" "177.8,177.8,177.8,177.8" ...
$ in.emissions_scenario_names : chr "LRMR_MidCase_15_2025start,LRMR_LowReCost_15_2025start,LR
035_15_2025start,LRMR_LowReCost_15_2025start,LRMR_MidCase_15_2025start,LRMR_LowReCost_15_2025start,LRMR_95Decarby
rt,LRMR_LowReCost_15_2025start" "LRMR_MidCase_15_2025start,LRMR_LowReCost_15_2025start,LRMR_95Decarby2035_15_2025st
Cost_15_2025start" "LRMR_MidCase_15_2025start,LRMR_LowReCost_15_2025start,LRMR_95Decarby2035_15_2025start,LRMR_Low
art" ...
$ in.emissions_types : chr "CO2e,CO2e,CO2e,CO2e" "CO2e,CO2e,CO2e,CO2e" "CO2e,CO2e,CO2e"
2e,CO2e,CO2e" ...
$ in.emissions_wood_values : chr "200.0,200.0,200.0,200.0" "200.0,200.0,200.0,200.0" "200.0"
0.0" "200.0,200.0,200.0,200.0" ...
$ in.federal_poverty_level : chr "400%" "400%" "400%" "400%" ...
$ in.generation_and_emissions_assessment_region : chr "SRVCC" "SRVCC" "SRVCC" "SRVCC" ...
$ in.geometry_attic_type : chr "Vented Attic" "Vented Attic" "Vented Attic" "Vented Attic"
$ in.geometry_building_horizontal_location_sf : chr "none" "none" "none" "none" ...
$ in.geometry_building_level_sf : chr "none" "none" "none" "none" ...
$ in.geometry_building_number_units_sf : chr "none" "none" "none" "none" ...
$ in.geometry_building_number_units_sfa : chr "none" "none" "none" "none" ...
$ in.geometry_building_type_acs : chr "Single-Family Detached" "Single-Family Detached" "Single-F
"Single-Family Detached" ...
$ in.geometry_building_type_height : chr "Single-Family Detached" "Single-Family Detached" "Single-F
"Single-Family Detached" ...
$ in.geometry_building_type_recs : chr "Single-Family Detached" "Single-Family Detached" "Single-F
"Single-Family Detached" ...
$ in.geometry_floor_area : chr "2000-2499" "2000-2499" "2000-2499" "2000-2499" ...
$ in.geometry_floor_area_bin : chr "1500-2499" "1500-2499" "1500-2499" "1500-2499" ...
$ in.geometry_foundation_type : chr "Vented Crawlspace" "Vented Crawlspace" "Vented Crawlspace"
pace" ...
$ in.geometry_garage : chr "2 car" "2 car" "2 car" "2 car" ...
$ in.geometry_stories : chr "2 2 2 2 2 2 2 2 2 2" ...
$ in.geometry_stories_low_rise : chr "2 2 2 2 2 2 2 2 2 2" ...
$ in.geometry_story_bin : chr "<8" "<8" "<8" "<8" ...
$ in.geometry_wall_exterior_finish : chr "Stucco, Light" "Stucco, Light" "Stucco, Light" "Stucco, L
$ in.geometry_wall_type : chr "Wood Frame" "Wood Frame" "Wood Frame" "Wood Frame" ...

$ in.geometry_wall_type : chr "Wood Frame" "Wood Frame" "Wood Frame" "Wood Frame" ...
$ in.has_pv : chr "no" "no" "no" "no" ...
$ in.heating_fuel : chr "Natural Gas" "Natural Gas" "Natural Gas" "Natural Gas" ...
$ in.heating_setpoint : chr "72F" "72F" "72F" "72F" ...
$ in.heating_setpoint_has_offset : chr "no" "no" "no" "no" ...
$ in.heating_setpoint_offset_magnitude : chr "0F" "0F" "0F" "0F" ...
$ in.heating_setpoint_offset_period : chr "none" "none" "none" "none" ...
$ in.holiday_lighting : chr "No exterior use" "No exterior use" "No exterior use" "No exterior
...
$ in.hot_water_distribution : chr "Uninsulated" "Uninsulated" "Uninsulated" "Uninsulated" ...
$ in.hot_water_fixtures : chr "100% Usage" "100% Usage" "100% Usage" "100% Usage" ...
$ in.hvac_cooling_efficiency : chr "AC, SEER 15" "AC, SEER 15" "AC, SEER 15" "AC, SEER 15" ...
$ in.hvac_cooling_partial_space_conditioning : chr "100% conditioned" "100% conditioned" "100% conditioned" "100% con
d" ...
$ in.hvac_cooling_type : chr "Central AC" "Central AC" "Central AC" "Central AC" ...
$ in.hvac_has_ducts : chr "Yes" "Yes" "Yes" "Yes" ...
$ in.hvac_has_shared_system : chr "none" "none" "none" "none" ...
$ in.hvac_has_zonal_electric_heating : chr "no" "no" "no" "no" ...
$ in.hvac_heating_efficiency : chr "Fuel Furnace, 80% AFUE" "Fuel Furnace, 80% AFUE" "Fuel Furnace, 80% AFUE" "Fuel Furnace, 80% AFUE" ...
$ in.hvac_heating_type : chr "Ducted Heating" "Ducted Heating" "Ducted Heating" "Ducted Heating" ...
$ in.hvac_heating_type_and_fuel : chr "Natural Gas Fuel Furnace" "Natural Gas Fuel Furnace" "Natural Gas Fuel Furnace" "Natural Gas
rtnace" "Natural Gas Fuel Furnace" ...
$ in.hvac_secondary_heating_efficiency : chr "none" "none" "none" "none" ...
$ in.hvac_secondary_heating_type_and_fuel : chr "none" "none" "none" "none" ...
$ in.hvac_shared_efficiencies : chr "none" "none" "none" "none" ...
$ in.hvac_system_is_faulted : chr "no" "no" "no" "no" ...
$ in.hvac_system_single_speed_ac_kirflow : chr "none" "none" "none" "none" ...
$ in.hvac_system_single_speed_ac_charge : chr "none" "none" "none" "none" ...
[!list output truncated]

```

Figure 4: structure of the data

- With approximately 5000 houses in our dataset, each with its own unique set of field names, we devised a loop to iteratively process the data. This loop facilitated the merging of house data, energy data, and weather data within a consistent time frame.

```

67 # Check and filter columns with more than one level
68 x <- length(levels(eng[, 1])) != 1
69 for (i in 2:ncol(eng)) {
70   if (is.factor(eng[, i])) {
71     x <- append(x, length(levels(eng[, i])) != 1)
72   } else {
73     x <- append(x, length(unique(eng[, i])) != 1)
74   }
75 }
76 cleandf <- eng[, x]
77 clean <- eng[,x]

```

Figure 5: data filtration

- Before proceeding with model application, we ensured data integrity by converting variables containing characters or alphanumeric combinations into factor variables. Factors with only one level were excluded to maintain dataset coherence.

```

80
81 #Apply asinh transformation to numeric columns
82 for (i in 1:ncol(cleandf)) {
83   if (is.numeric(cleandf[, i])) {
84     cleandf[, i] <- asinh(cleandf[, i])
85   }
86 }
87

```

Figure 6: data transformation

- To optimize model performance, we transformed numeric variables using the sinh function. This transformation aimed to enhance the accuracy and adjusted R-squared of our models, thus improving the overall quality of our analysis.

```

# view summary of energy_consumption column in cleandf
summary(cleandf$energy_consumption)

```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.05697	0.98094	1.25427	1.27561	1.54405	3.57638

Figure 7: summary of energy consumption after data cleaning

- The summary statistics for energy consumption after cleaning the data show the minimum, 1st quartile, median, mean, 3rd quartile, and maximum values of energy

consumption in the cleaned dataset. The minimum energy consumption is 0.05697, which is the lowest energy consumption recorded after cleaning the data. The 1st quartile is 0.98094, meaning that 25% of the data points have energy consumption equal to or below this value in the cleaned dataset. The median is 1.25427, indicating that half of the data points have energy consumption equal to or below this value, and the other half have energy consumption equal to or above this value in the cleaned dataset. The mean is 1.27561, which represents the average energy consumption across all data points in the cleaned dataset. The 3rd quartile is 1.54405, which means that 75% of the data points have energy consumption equal to or below this value in the cleaned dataset. The maximum energy consumption observed in the cleaned dataset is 3.57638, representing the highest energy consumption recorded after cleaning the data.

## DESCRIPTIVE STATISTICS

Initially, we commenced by integrating the weather data with the energy consumption records for houses.

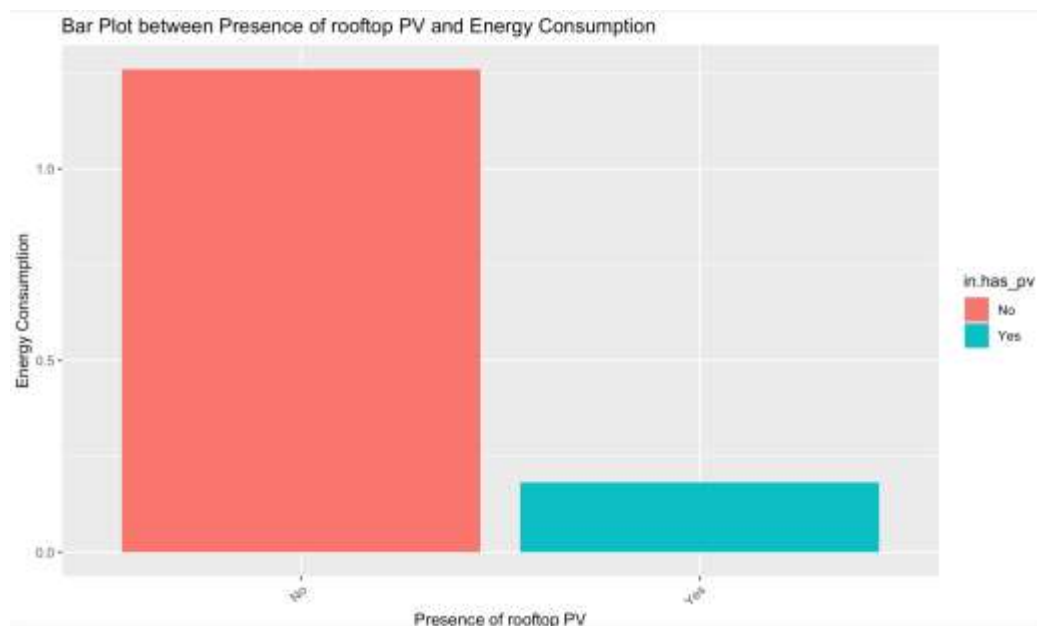


Figure 8: bar plot between presence of rooftop PV and energy consumption

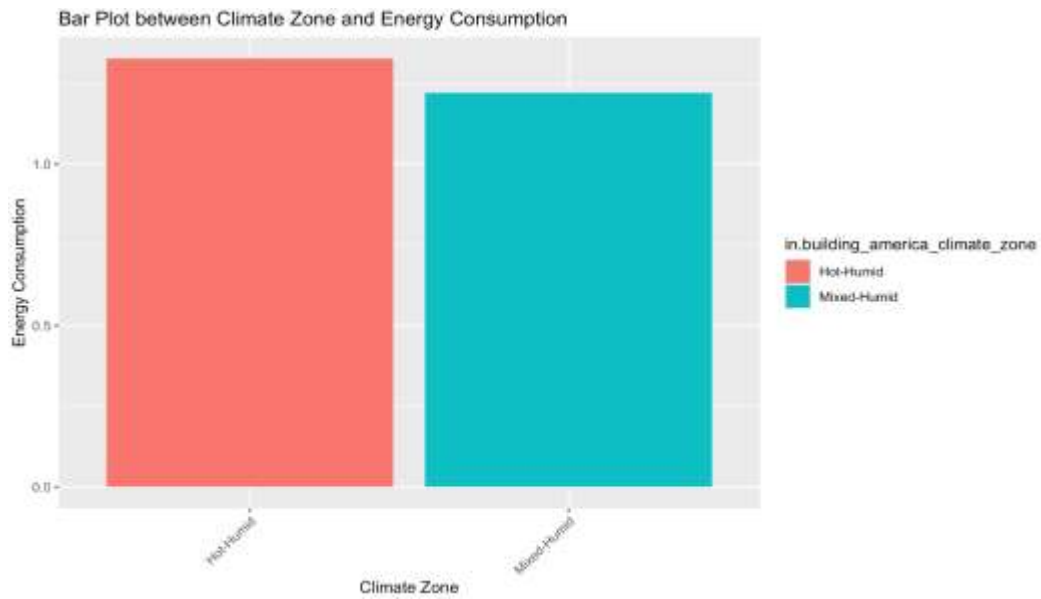


Figure 9: bar plot between climate zone and energy consumption

Upon thorough examination, a notable trend emerged in the figure 7 and figure 8, regions positioned within hot-humid climate zones demonstrated a propensity for heightened energy consumption relative to those in mixed-humid climate zones. Furthermore, a significant disparity was observed between houses equipped with rooftop photovoltaic (PV) systems and those without. Specifically, houses with rooftop PV installations exhibited lower energy consumption rates. With our overarching objective centred around reducing energy usage in areas characterized by elevated consumption levels, our focus narrowed to houses situated in hot-humid conditions lacking rooftop PV systems.



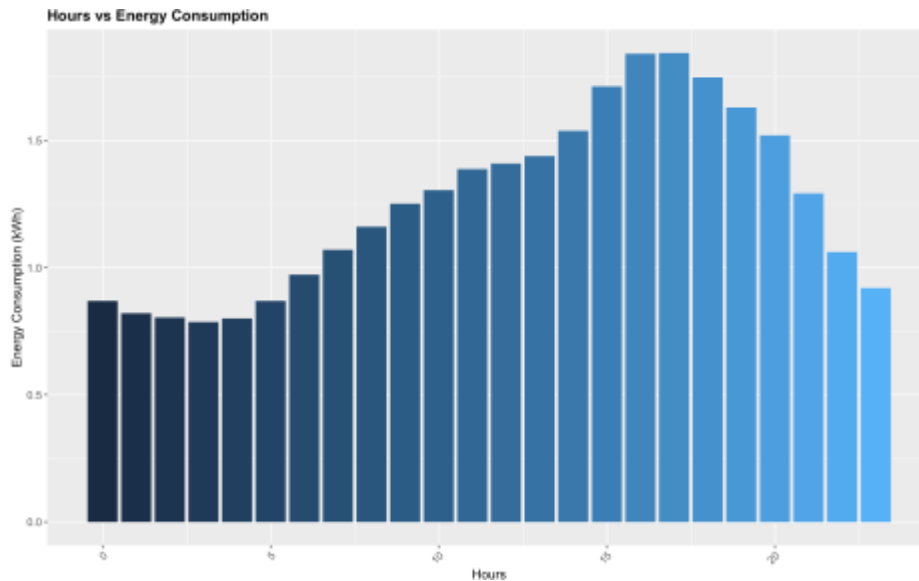


Figure 10: energy consumption patterns over 24 hours

In figure 9, our analysis revealed a notable surge in energy consumption between the hours of 2 pm and 7 pm. Therefore, we opted to exclusively focus on this time frame when constructing our prediction model

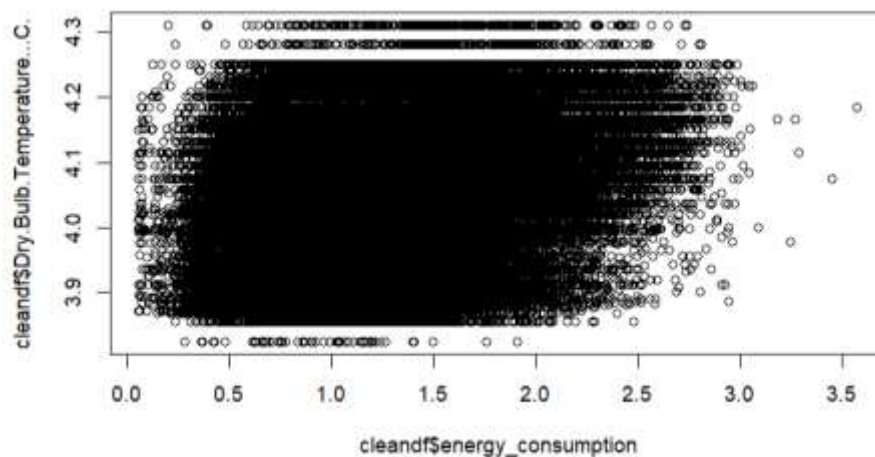


Figure 11: scatter plot between energy consumption and dry bulb temperature

In figure 10, the graph illustrates the relationship between clean and dry bulb temperature (°C) and clean dry energy consumption. The data points are densely populated, suggesting a potential correlation or pattern between these two variables. The scatter plot graph provides insights into how changes in temperature might impact energy consumption, which could be crucial for energy efficiency studies.

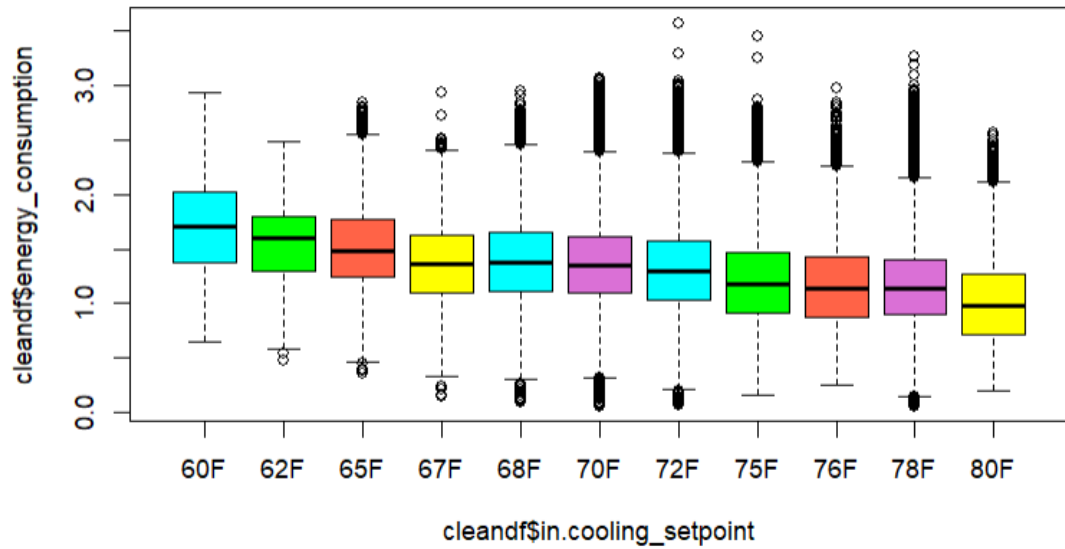


Figure 12: boxplot of energy consumption by heating set points

Since the time of study is in the summer when energy demand is high to cool houses, we expect energy consumption to be high at low cooling set points and low at high cooling set points, as shown in Figure 11.

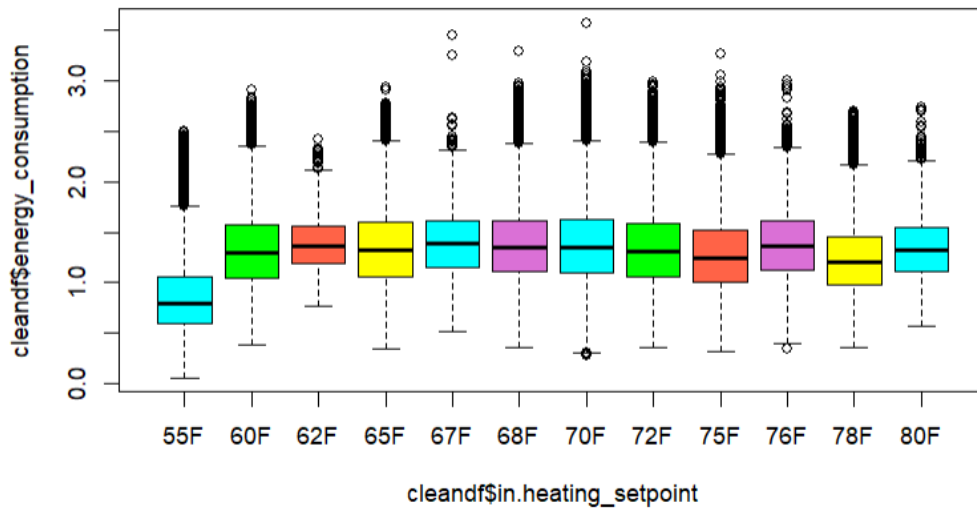


Figure 13: boxplot of energy consumption by heating set points

The boxplot of figure 12 shows energy consumption by different heating set points. Many houses set heating set point varies between 55 °F and 80 °F, and among which energy consumption does not varies significantly especially between 60 °F and 80 °F. However, they are significantly higher compared to the heating set point at 55 °F.



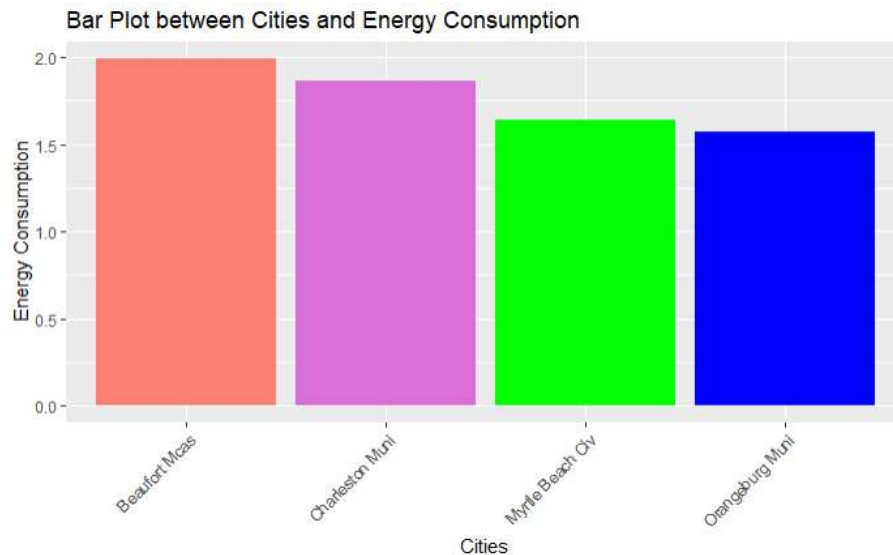


Figure 14: bar plot of energy consumption in the cities

In figure 13, the graph represents energy consumption in four cities: Beaverton, Clackamas, White Bear City, and Orangeburg. White Bear City has the highest energy usage, while Clackamas consumes the least.

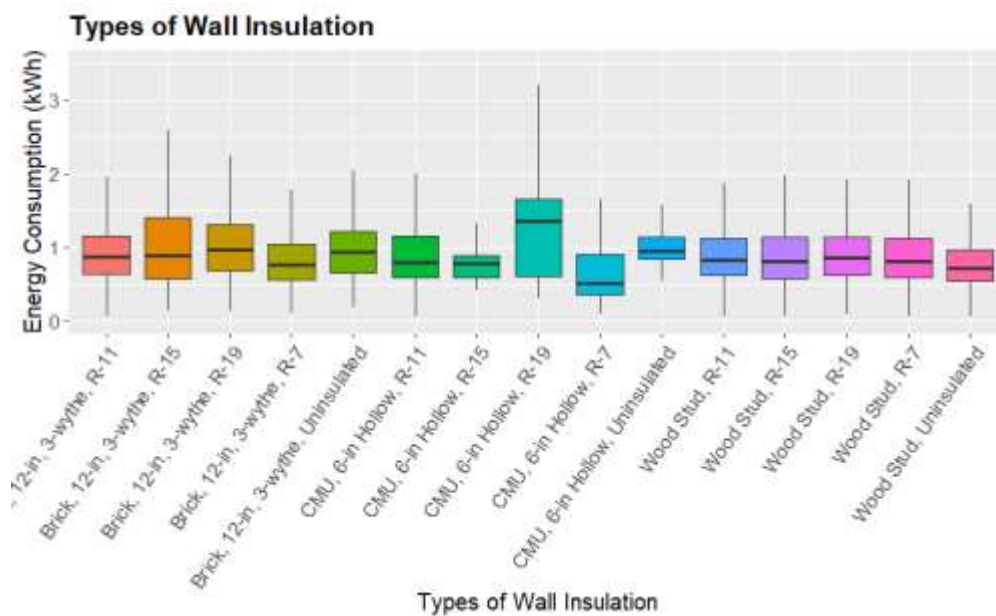


Figure 15: boxplot of energy consumption by types of wall insulation

In figure 14, the box plot likely compares different types of wall insulation. Notably, CMU (Concrete Masonry Unit) and 6-inch Hollow insulation stand out as energy-efficient options. Houses with CMU and 6-inch Hollow insulation consume less energy compared to other insulation types. This finding underscores the importance of selecting the right insulation material for optimal energy conservation. In summary, informed insulation choices can significantly impact both our comfort and environmental footprint.

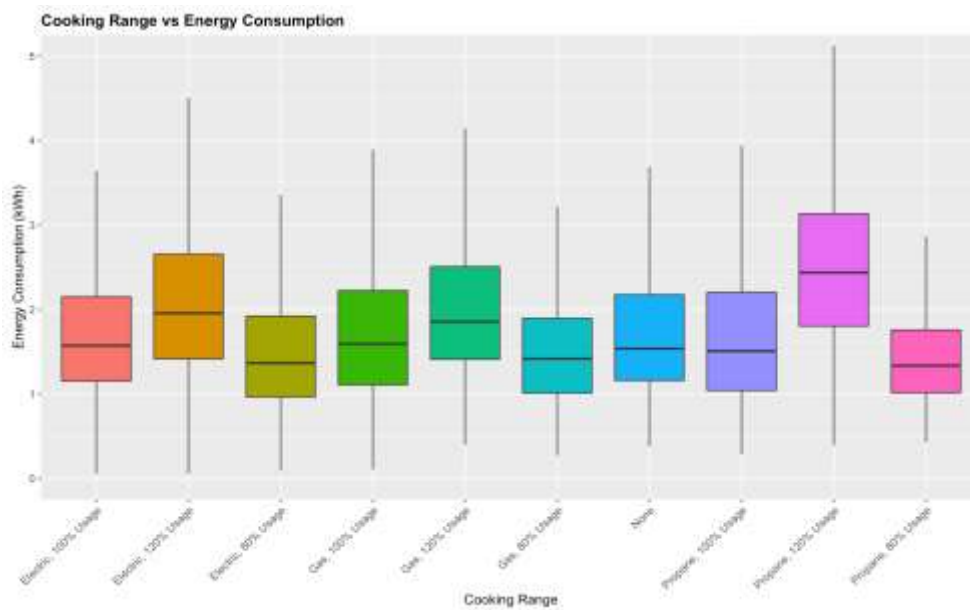


Figure 16: boxplot of energy consumption by types of cooking range

In figure 15, the box plot likely compares different cooking ranges (propane, electric, and gas) in terms of energy consumption. Each range's efficiency and impact on energy use are under scrutiny. The text emphasizes that using propane at 80% efficiency results in the least energy consumption. This finding suggests that households can significantly reduce energy usage by opting for propane-based cooking. Shifting from electric or gas cooking to propane-based cooking is a practical step toward energy savings. Propane's higher efficiency translates to lower energy bills and reduced environmental impact.

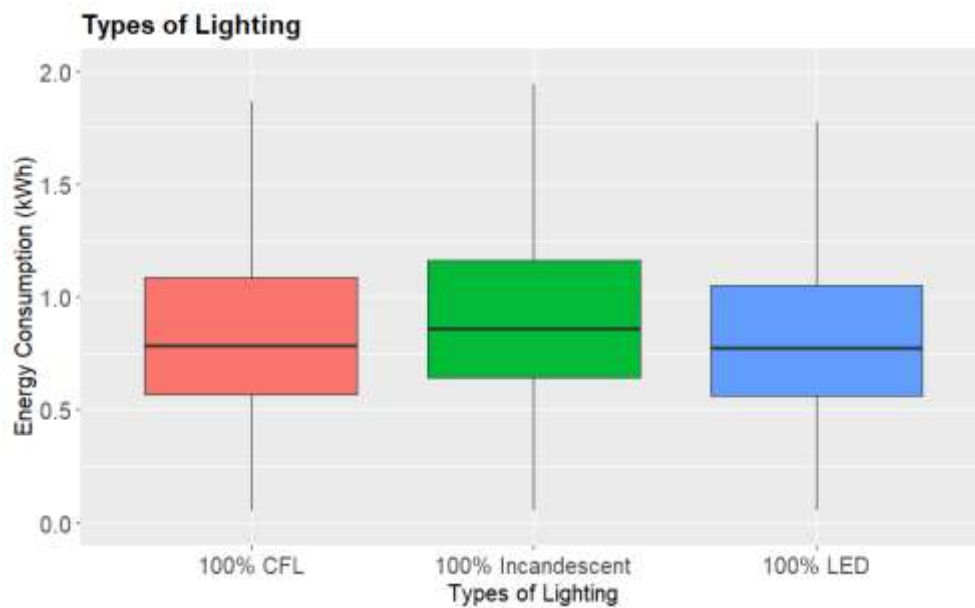
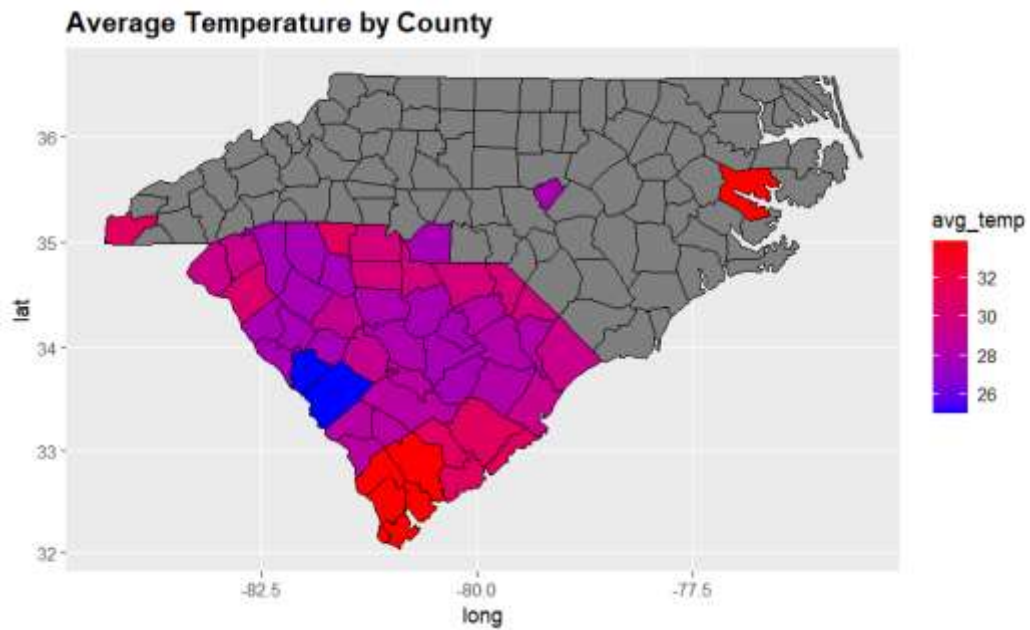


Figure 17: boxplot of energy consumption by types of lighting

Both CFLs and LEDs consume significantly less energy compared to traditional incandescent bulbs. Incandescent bulbs are less energy-efficient. As we can see in figure 16, the graph shows that they consume slightly more energy (approximately 0.3 kWh) compared to CFLs and LEDs. The shift from incandescent bulbs to CFLs or LEDs is a practical step toward energy conservation. By replacing outdated bulbs, households can reduce energy usage and contribute to environmental sustainability.



*Figure 18: sample average temperature map of each county in south carolina and some areas in north carolina at 2 pm on July 13, 2018*

On July 13, 2018, at 2 pm, a temperature map provided insights into the climate across South Carolina and parts of North Carolina. Charleston, being coastal, experienced milder temperatures, while Columbia had higher highs and lows. Greenville, further inland, maintained a moderate climate. The color-coded bands represented temperature ranges, with cooler blues and warmer reds. Overall, this snapshot offered insights into regional temperature distribution and sky conditions as shown in the figure 17.

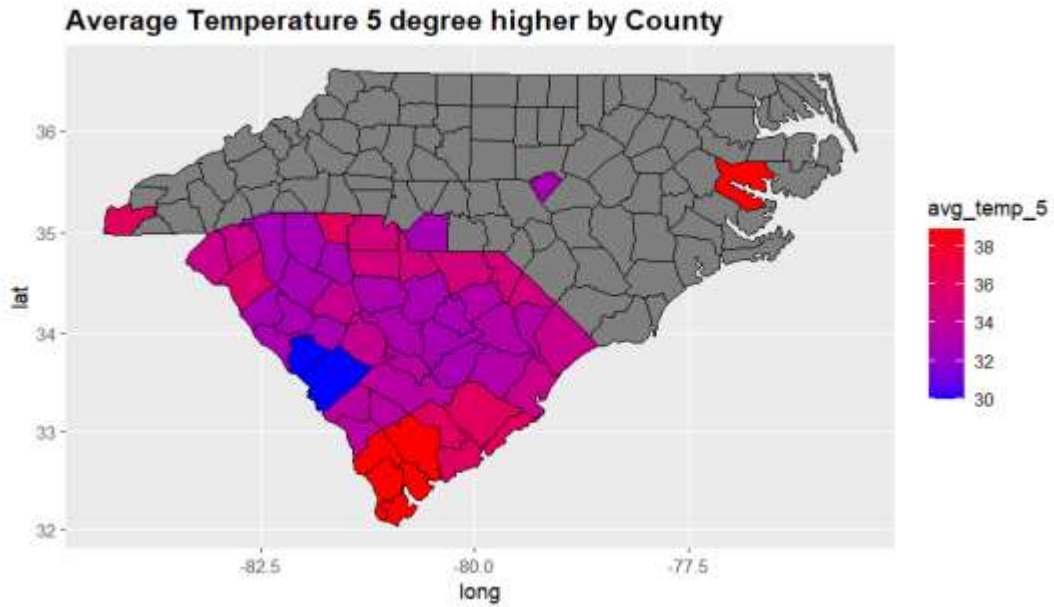


Figure 19: simulated average temperature map of each county in South Carolina and some areas in North Carolina if increases by 5 degree celsius

As per figure 18, ti appears to depict the impact of a 5-degree temperature increase across different counties. However, despite this rise in temperature, no significant change is observed.

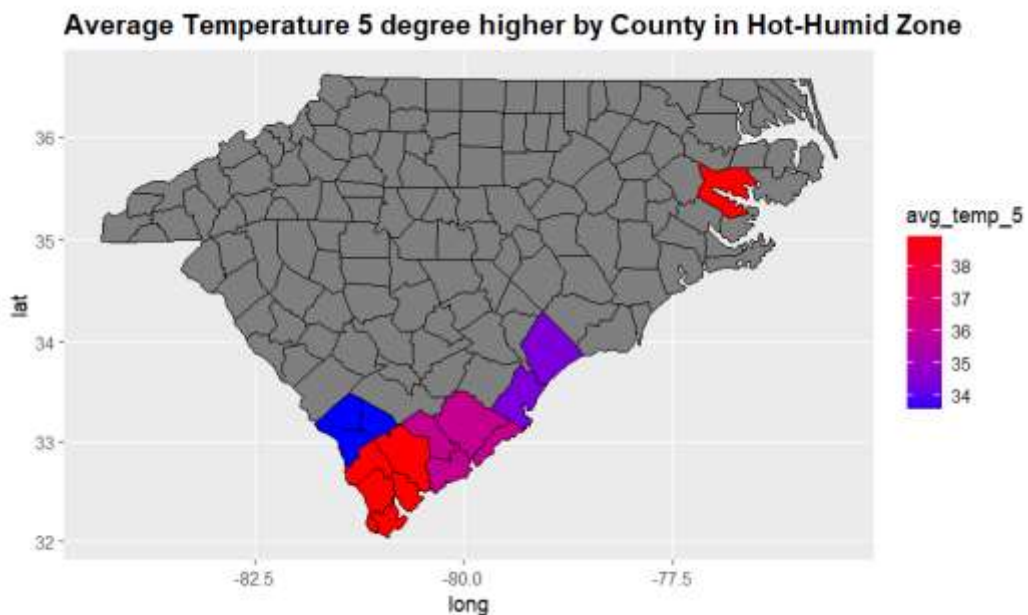


Figure 20: simulated average temperature map of counties that fall in hot-humid zone

This graph in figure 19 illustrates a simulated average temperature map for counties within the Hot-Humid zone.

## DATA MODELLING

### Modelling Approach:

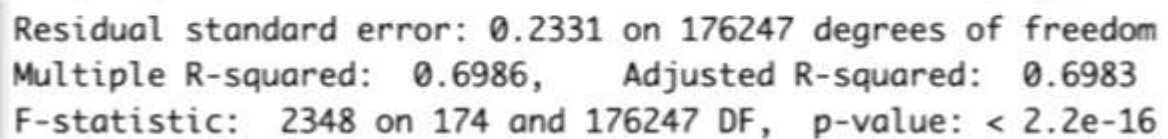
In the project, a combination of regression models was employed to predict energy consumption in households. The choice of models includes linear regression, support vector regression, and boosting methods such as light gradient boosting. This approach was taken to ensure a robust and accurate prediction model, as each of these methods has its strengths and weaknesses.

Before building the predictive models, the project team focused on extracting the most significant features in the dataset. These features were selected based on their high significance and commonality in almost all households. The most significant features used to train the model are in.heating\_setpoint, Dry.Bulb.Temperature...C., in.lighting, in.misc\_pool, in.misc\_hot\_tub\_spa, in.occupants, in.cooling\_setpoint, in.cooling\_setpoint\_offset\_magnitude, in.misc\_gas\_fireplace, in.window\_areas, in.income, in.misc\_freezer, Global.Horizontal.Radiation..W.m2., in.misc\_pool\_heater, in.sqft, in.cooking\_range, Direct.Normal.Radiation..W.m2., Wind.Speed..m.s., in.cooling\_setpoint\_offset\_period, in.geometry\_foundation\_type, in.misc\_gas\_grill, in.misc\_well\_pump, in.ducts, in.insulation\_wall, in.hot\_water\_fixtures, in.misc\_extra\_refrigerator, energy\_consumption. These 26 attributes were used to train the predictive model for energy consumption in households, with a combination of linear regression, support vector regression, and boosting methods.

## Classification Models:

### Linear Regression Model:

The Linear Regression model is a statistical technique used to model the relationship between a dependent variable (the one we want to predict) and one or more independent variables (predictors). It assumes a linear relationship between these variables. The model estimates the coefficients of the linear equation to make predictions. In the context of energy consumption, it helps predict energy usage based on relevant features or factors.



```
Residual standard error: 0.2331 on 176247 degrees of freedom
Multiple R-squared: 0.6986,    Adjusted R-squared: 0.6983
F-statistic: 2348 on 174 and 176247 DF,  p-value: < 2.2e-16
```

Figure 21: linear regression model output

The Linear Regression model, with an impressive R-squared value of 0.698, accurately predicts energy consumption. Its statistical significance, as indicated by a low p-value, underscores its reliability. This model holds great potential for strategic decision-making. The R-squared value of 0.698 suggests that 69.8% of the variability in energy consumption can be explained by the model. The small p-value ( $< 0.05$ ) indicates that the model's coefficients are statistically significant. With a robust R-squared and significant p-value, the model is reliable.

### Light Gradient Boosting Model

The LGB model is a powerful gradient boosting framework designed for efficient and distributed training. It utilizes tree-based learning algorithms to enhance prediction accuracy. Key features include histogram-based computation and leaf-wise growth, optimizing both speed and performance. Particularly useful for handling large datasets and achieving high accuracy.

```
[95]: test's rmse:0.332205
[96]: test's rmse:0.332143
[97]: test's rmse:0.332088
[98]: test's rmse:0.332054
[99]: test's rmse:0.331995
[100]: test's rmse:0.331943
Did not meet early stopping, best iteration is: [100]: test's rmse:0.331943
Warning: NAs introduced by coercion [1] "RMSE on test data: 0.358831184446306"
```

Figure 22: light gradient boosting model output

The LGB model yielded a respectable RMSE of 0.33. However, it did not surpass linear regression in performance. Despite this, the LGB model remains a valuable tool for various applications.

### Support Vector Regression Model:

Support Vector Regression (SVR) is a type of machine learning algorithm used for regression analysis. Unlike classification tasks, where the outcome is discrete (e.g., class labels), regression aims to predict a continuous target variable. SVR finds a function that approximates the relationship between input variables and the target variable while minimizing prediction error. It is an extension of Support Vector Machines (SVM), which are widely used for classification problems.



```
Call:
svm(formula = energy_consumption ~ ., data = train_data, type = "eps-regression",
     kernel = "radial")

Parameters:
  SVM-Type:  eps-regression
SVM-Kernel:  radial
    cost:    1
   gamma:   0.005617978
  epsilon:   0.1

Number of Support Vectors: 135786

[1] "RMSE: 0.2116875401452"
[1] "R_Squared: 0.750972509159475"
```

Figure 23: support vector regression model output

We executed the Support Vector Regression (SVR) model with the goal of achieving an R-squared value surpassing that of the Linear model. Impressively, the SVR yielded an R-squared value of 0.75 and an exceptionally low Root Mean Squared Error (RMSE) of 0.21. These metrics validate the SVR model's precision in predicting energy consumption.

### Interpretation of models:

Through our analysis of predictive models for energy consumption, we uncovered valuable insights. Linear Regression emerged as a robust performer, boasting an R-squared of 0.699, signifying its reliability in forecasting energy usage. LightGBM, while promising, exhibited a slightly higher RMSE compared to Linear Regression, suggesting room for improvement. However, the Support Vector Regression (SVR) model stole the spotlight with an exceptional R-squared of 0.75 and minimal RMSE, showcasing its unparalleled accuracy in predicting energy consumption patterns. This underscores SVR's superiority among the models evaluated. Its precision and low error rates make SVR the preferred choice for forecasting energy usage reliably. This comprehensive analysis equips us with actionable insights to guide strategic decision-making and support energy efficiency initiatives effectively.

## VISUALIZATION

To obtain a thorough grasp of energy consumption trends, we explore the information gleaned from our prediction models.

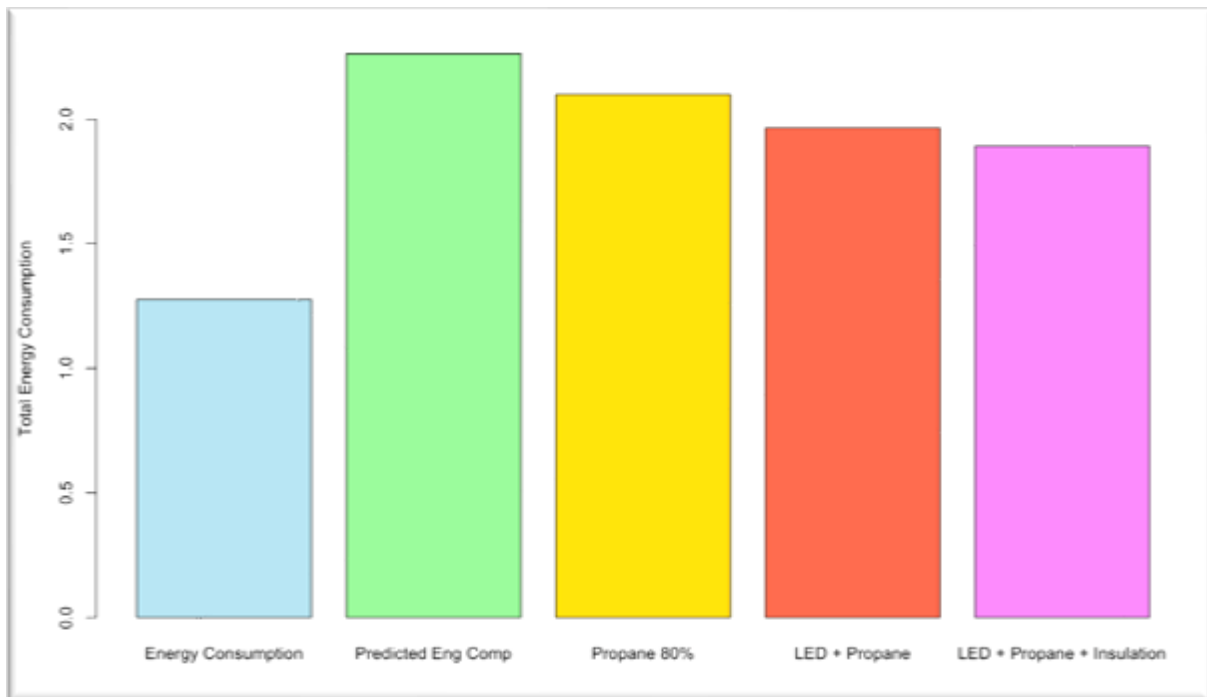


Figure 24: energy consumption across various energy sources

The bar plot in the figure 23 provides a visual representation of energy consumption across various energy source and insulation combinations. Each bar signifies a unique scenario, color-coded for clarity:

Light blue bars denote baseline energy consumption.

Light green bars represent predicted consumption based on the model.

Gold bars indicate energy use with 80% propane for cooking.

Tomato (red) bars signify the combination of LED lighting and propane.

Violet bars illustrate the most efficient blend of LED lighting, propane, and specific insulation.

Notably, the lowest energy usage occurs with LED bulbs, 80% propane cooking, and specific insulation. Encouraging adoption of this efficient combination could markedly reduce energy consumption, informing strategic energy management decisions for households and buildings.

## SHINY APP

### Title Panel:

At the top of the application, the title panel prominently showcases the title "SVR Model Prediction for Energy Consumption."

### Input Controls:

The input controls are neatly organized into three columns. Within each column, users can find dropdown menus, text inputs, and numeric inputs. These controls allow users to specify various parameters related to energy consumption, such as heating setpoint, dry bulb temperature, lighting type, occupants, cooling setpoint, window areas, square footage, cooking range, and more. They cover a wide range of factors influencing energy usage, including weather conditions, household characteristics, appliance usage, and geographical location.

### Prediction of Energy Consumption Using SVM Model

Heating Setpoint 72F	Window Areas F12 B12 L12 R12	Cooling Setpoint Offset Period Day and Night Setup
Dry Bulb Temperature (C) 40	Income 100000-119999	Geometry Foundation Type Ambient
Lighting 100% CFL	Misc Freezer EF 12, National Average	Misc Gas Grill Gas Grill
Misc Pool Has Pool	Global Horizontal Radiation (W/m2) 0	Misc Well Pump None
Hot Tub/Spa Electric	Misc Pool Heater Electric	Ducts 0% Leakage, Uninsulated
Occupants 1	Square Footage 1220	Wall Insulation CMU, 6-in Hollow, R-7
Cooling Setpoint 60F	Cooking Range Electric, 100% Usage	Hot Water Fixtures 100% Usage
Cooling Setpoint Offset Magnitude 0F	Direct Normal Radiation (W/m2) 1	Extra Refrigerator EF 10.2
Misc Gas Fireplace Gas Fireplace	Wind Speed (m/s) 1	

**Predict Energy Consumption**

**Your Predicted value is**

1.238372

Figure 25: user interface of energy consumption in the shiny app

#### Predict Button:

Located at the bottom of the input controls section, a predict button labelled "Predict Energy Consumption" is available. By clicking this button, users can trigger the prediction of energy consumption based on the specified input parameters.

#### Predicted Value Display:

Below the predict button as shown in the figure 24, there's a dedicated section for displaying the predicted energy consumption value. Whenever users click the predict button, this section dynamically updates to show the predicted energy consumption value. This feature enables users to swiftly view the predicted energy consumption output based on their selected input parameters.

## CONCLUSION

Addressing the business questions posed:

- **Main Factors Contributing to Energy Consumption:** The analysis revealed several significant factors influencing energy consumption in residential properties served by eSC. Notably, weather variables such as temperature, solar radiation, and wind speed exerted considerable influence on energy usage. Building characteristics, including house size, insulation level, and appliance efficiency, also played pivotal roles in determining energy consumption patterns.
- **Alignment of Energy Reduction Efforts with Sustainability Goals:** Our efforts to reduce energy consumption are closely aligned with sustainability objectives. By identifying key drivers of energy usage and developing predictive models, we aim to assist eSC in managing energy demand effectively, reducing operational costs, and promoting environmental sustainability. Encouraging the adoption of energy-efficient practices and technologies can further align our efforts with sustainability goals and commitments.
- **Variation in Energy Consumption Across Different Counties or Locations:** Our analysis revealed variations in energy consumption across different counties and locations within the service area of eSC. Factors such as climate, population density, and economic activity influenced energy usage patterns, with regions in hot-humid climate zones exhibiting higher energy consumption levels. Understanding these variations is crucial for tailoring energy management strategies to specific geographic areas and optimizing resource allocation.
- **Solutions to Support Energy Efficiency Projects:** To support eSC's energy efficiency projects and initiatives, we propose several solutions based on our analysis. These include promoting the adoption of energy-efficient appliances, incentivizing the use of renewable energy sources such as rooftop solar panels, implementing targeted

energy conservation programs, and providing personalized energy consumption insights to customers. By leveraging data-driven insights and innovative solutions, eSC can enhance its energy efficiency efforts and achieve sustainable outcomes.

In summary, our analysis provides valuable insights and recommendations to help eSC address energy consumption challenges effectively, advance sustainability goals, and ensure reliable service delivery to customers. By implementing proactive measures and adopting a holistic approach to energy management, eSC can navigate the evolving energy landscape with confidence and resilience.

## **CONTRIBUTIONS OF TEAM**

Sudhanshu Kumar Pawar – Data Modelling and Power point presentation

Teera Yong – Visualizations

Tharuni Tekula – Visualizations and Documentation

Vinay Kumar Chandra – Data merging, Data cleaning, Data modelling, Shiny app