

**IST 687: INTRODUCTION TO DATA SCIENCE**

**SOUTH CAROLINA ENERGY ANALYSIS**

**SUBMITTED BY:**

ADITYA MARATHE

GAUTAM BALGI

ROHAN GOPALAN

SIDDHARTH KASTURE

**Table of Contents**

| Section | Key Points |
| --- | --- |
| Project Description | - Energy consumption analysis for South Carolina |
|  | - Aims to predict usage, understand weather impact, develop strategies |
| Project Scope | - Analysis of housing energy usage and weather data in South Carolina |
|  | - Includes data collection, cleaning, analysis, and visualization |
| Data Description | - Energy consumption for housing |
|  | - Weather conditions |
|  | - Building characteristics (square footage, bedrooms, occupancy) |
| Data Preparation | - Cleaning and preprocessing using county and building IDs |
|  | - Outlier removal, scaling, standardization |
|  | - Addressing missing values |
| Exploratory Data Analysis | - Daily energy consumption patterns |
|  | - Energy levels by city and county |
|  | - Impact of factors like time, weather, household size |
| Modelling | - Random Forest model used |
|  | - Explains 20.74% of variance in energy consumption |
|  | - Considers factors like square footage, bedrooms, occupants |
| Key Visualizations | - County-wise energy consumption (descending order) |
|  | - Daily energy consumption by county |
|  | - Comparison of original and modified predictions |
|  | - Distribution of mean differences |
| Water Heater Efficiency | - Higher UEF ratings lead to lower total consumption |
|  | - Consumption increases with more occupants |
| Conclusion | - Provides insights for energy demand forecasts |
|  | - Identifies areas for efficiency improvements |
|  | - Informs policy decisions for optimizing energy consumption |

**PROJECT DESCRIPTION**

This project provides an in-depth analysis of energy consumption, emphasizing a systematic approach that aims to understand, measure, and optimize energy usage across a variety of contexts. The primary goal of this project is to address several significant challenges that energy providers in South Carolina currently face.

1. **Predicting Future Energy Usage**: We will explore methodologies to model and forecast energy demand, taking into account historical consumption patterns, population growth, and economic factors. This predictive analysis is crucial for ensuring that energy providers can meet future needs without over or underestimating supply requirements.
2. **Understanding the Impact of Weather Conditions on Consumption**: This section will delve into how different weather patterns influence energy usage. We will examine data to ascertain how temperature fluctuations, precipitation, and seasonal changes affect electricity and gas consumption, providing insight into the relationship between climate and energy demand.
3. **Developing Strategies for Cost Reduction and Efficiency**: Here, we will identify potential strategies that energy providers can implement to reduce costs while enhancing efficiency. This might include the adoption of new technologies, the implementation of demand-side management programs, and the promotion of energy-saving practices among consumers.
4. **Creating Accurate Demand Forecast:** This part of the presentation will focus on the importance of developing precise forecasts for energy demand. We will discuss analytical tools and techniques that can be employed to improve forecasting accuracy, contributing to better resource allocation and operational planning.

Through this comprehensive analysis, we aim to equip energy providers with the insights and strategies necessary to navigate the evolving landscape of energy consumption in South Carolina.

**PROJECT SCOPE**

The project involves a detailed analysis of energy consumption data specifically related to residential homes in South Carolina. It aims to examine how energy usage is influenced by various factors, including weather conditions and residential practices.

To achieve this, the project employs a systematic approach consisting of several key phases:

1. **Data Collection:** The process involves assembling a wide range of datasets that capture detailed records of energy consumption from various households. This includes analyzing local climate data, such as temperature and humidity variations, alongside other pertinent information that could influence energy usage patterns. By integrating these diverse data points, we can gain valuable insights into the factors affecting energy consumption in different environments.

2. **Data Cleaning:** It is essential to ensure that the collected data is both accurate and consistent. This involves a thorough process of identifying any errors or discrepancies and making the necessary corrections. Taking the time to address these issues is vital, as it lays a solid foundation for conducting reliable and insightful analysis in the future.

3. **Data Analysis**: We employ a variety of statistical methods and advanced analytical tools to thoroughly examine energy consumption data. This process aims to reveal significant trends, correlations, and patterns that may exist within the data. Specifically, we focus on understanding how various factors—such as fluctuations in temperature and changes brought on by different seasons—impact energy usage in households. By digging deeper into these insights, we can better comprehend the dynamics of residential energy consumption and identify opportunities for improved energy efficiency.

4. **Data Visualization:** Developing visual representations of the analyzed data is essential for clearly conveying the findings. This may involve creating detailed graphs, informative charts, and illustrative maps that not only depict trends in energy consumption over time but also emphasize key relationships between various factors. These visuals serve as powerful tools to enhance understanding and facilitate discussions around the data.

Through this comprehensive approach, the project aims to provide valuable insights into the patterns of energy consumption in South Carolina and the various factors that influence residential energy usage.

**PROJECT DELIVERABLES**

The project aims to deliver a comprehensive set of outcomes that will enhance our understanding of energy consumption patterns. These key deliverables include:

1. **A Cleaned and Pre-processed Dataset:** This dataset will undergo rigorous cleaning and preprocessing to remove any inconsistencies, duplicates, or irrelevant information, ensuring that the data is reliable and suitable for further analysis.
2. **Results from Exploratory Data Analysis (EDA**): This phase will involve detailed analysis to uncover underlying trends, relationships, and anomalies within the data. The EDA results will provide a foundational understanding of the dataset and inform subsequent modelling efforts.
3. **Visualizations Depicting Energy Consumption Patterns:** A series of visualizations will be developed to graphically represent energy consumption trends over time. These visualizations will include charts and graphs that showcase peak usage periods and energy consumption variations across different sectors.
4. **Statistical Models to Predict Energy Consumption:** Advanced statistical models will be created to forecast future energy consumption based on historical data and identified trends. These predictive models will employ techniques such as regression analysis and time series forecasting.
5. **Insights into Factors Influencing Energy Usage:** The project will culminate in an analysis of the determinants affecting energy consumption. These insights will help identify key factors such as demographics, economic indicators, and seasonal variations that significantly impact energy usage patterns.

**SHINY APPS**

Our analysis setup is designed to provide detailed information about electricity use in South Carolina, with a special feature that allows us to focus on individual cities. Here’s a clearer breakdown of how it works:

**Choose Cities to Examine:** We can select specific cities within South Carolina to see their electricity consumption. This is particularly useful for understanding the energy dynamics of a particular area.

**See the Overall Picture:** In addition to examining individual cities, we can also look at the electricity usage for the entire state of South Carolina.

**City-Specific Details:** Beyond the state-wide overview, we can explore detailed information for each city. This enables us to identify the unique energy needs and usage patterns of each location, which can vary significantly from one city to another.

**DESCRIPTION OF DATA**

**A. Static House Data:**

Description:

The Static House Data data set provides a detailed representation of single-family homes serviced by the Energy Service Company (eSC). This carefully curated random sample offers significant insights into the structural characteristics and energy consumption patterns of these residences. Each property is assigned a unique identifier, referred to as a "building ID," which facilitates efficient integration with its corresponding energy usage data for comprehensive analysis.

This dataset encompasses various fixed attributes of the houses, such as total square footage, the number of bedrooms and bathrooms, construction type, and architectural design features. These stable characteristics form the foundation for longitudinal analyses, enabling a deeper understanding of energy consumption trends and residential building performance over time.

**Data Format:**  
The dataset is stored in the Parquet format, a highly efficient columnar storage structure optimized for large-scale data processing frameworks. Parquet ensures superior storage efficiency and query performance, making it particularly well-suited for handling complex analyses and extensive datasets like this one.

**Size:**  
With approximately 5,000 entries, each representing a distinct house, this dataset serves as a rich resource for researchers and analysts. It facilitates the exploration of energy usage trends, the study of residential building characteristics, and the assessment of factors influencing energy efficiency in single-family homes.

This dataset is a cornerstone for projects aiming to enhance understanding and foster advancements in energy efficiency and residential building performance.

**B. Energy Usage Data:**

The Energy Usage Data is made up of several datasets from Static House Data, each of which is linked to a particular home. This dataset offers calibrated and verified energy usage statistics for individual homes, including comprehensive hourly energy consumption records. It contains one-hour load profiles that record energy usage from different household appliances, including dryers and air conditioners. The associated "building ID," which doubles as the filename for convenience, uniquely identifies each file.  
  
Format of Data  
The Energy Usage Data is saved in the Parquet format, much like the Static House Data. Processing and analysis are made easier by this effective columnar storage format, which also improves storage optimization.

Size:

There are approximately 5,000 individual datasets, each corresponding to a different house identified by its 'building ID'.

**C. Weather Data**:

The weather data is broken down into individual files for each county or geographic area and includes hourly weather information. To match the hourly format of the other datasets, the data is time-to-time. Each house is linked to its corresponding county code by the 'in.county' field in the Static House Data.

Format: Accessing and interpreting the weather data is made simple by its simple CSV storage.

Size: About 50 weather files total, each associated with a distinct county code.

**EXPLANATION OF DATA PREPARATION**

Here's a detailed overview of what we've accomplished so far: Energy Consumption Data We've made substantial headway with the energy consumption data:

- Successfully iterated through and combined energy consumption data for over 1000 houses into a single comprehensive data frame. This consolidation allows for more efficient analysis and comparison across different households.

- Focused our analysis by reducing the dataset to show only data for the month of July. This targeted approach helps in managing the data volume and allows for a more detailed examination of energy usage during a specific time, which could be particularly interesting for understanding summer consumption patterns.

- Created a new metric that displays the total energy consumption in a single hour. This aggregation offers an overview of energy usage trends and peaks within the sample which could be utilized for linking static household data.

Our work on the static house data has involved refining and enhancing its analytical potential:

- Streamlined the dataset by reducing the number of variables to 42. This focused approach helps in managing complexity and concentrating on the most relevant factors affecting energy consumption.

- Conducted an analysis of collinearity between variables. This step is crucial for understanding relationships between different house characteristics and for preparing the data for potential modeling efforts. Established a critical link between the building IDs in this dataset and their associated energy consumption data. This connection enables us to relate house characteristics directly to energy usage patterns.

Weather Data For the weather data, we've taken steps to align it with our energy consumption focus:

- Filtered the weather data to include only information for the month of July, matching our energy consumption data timeframe. This alignment allows for direct comparisons between weather conditions and energy usage.

**Data Preprocessing Report**

The document involves detailed **data preprocessing** steps for energy consumption analysis. The main focus lies on cleaning, merging, transforming, and preparing the data for further analysis and modeling. Below is a structured summary:

**1. Data Loading**

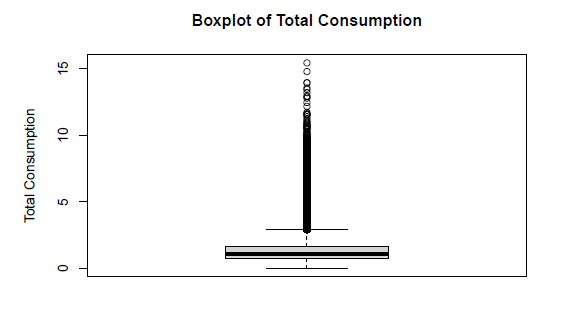
* **Data Sources**:
  + dataf: Building information loaded via **Parquet**.
  + climate\_data: Weather information
  + july\_data: Energy consumption data
* **Libraries Used**:
  + ggplot2, arrow, randomForest, caret, tidyverse.

**2. Initial Data Inspection**

* **Data Structure**:
  + dataf and july\_data include a large number of columns and rows.
  + Merged datasets (merged\_df\_ener\_hou) include **1,186,432 rows** and **103 variables**.
* **Missing Data**:
  + Checked using sum(is.na(dataf)).
  + Missing values were handled using na.omit() to remove incomplete records.

**3. Data Cleaning**

* **Column Adjustments**:
  + Columns like in.cooling\_setpoint and in.heating\_setpoint were cleaned to remove the "F" character.
  + Converted to numeric types.
* **Constant Columns Removal**:
  + Identified columns with a single unique value.
  + These columns were removed as they lack variability and contribute no meaningful information:
    - E.g., in.ahs\_region, in.corridor, in.hot\_water\_distribution, etc.
* **Negative Values**:
  + Detected in total consumption (energy usage). Since negative values are invalid, they were removed.



* **Categorical Variables**:
  + Categorical columns were converted into numerical encodings using mutate\_all(as.factor).

**4. Feature Engineering**

* **Energy Consumption**:
  + Total energy consumption was calculated as a row-wise sum for all relevant energy-related columns.

july\_data$total\_consumption <- rowSums(july\_data[,2:43])

* **Feature Correlation**:
  + Pearson Correlation Coefficients were computed to identify the strength of relationships between total\_consumption and other predictors.
  + Top correlated variables:
    - in.sqft (0.37)
    - in.geometry\_floor\_area\_bin (0.34)
    - in.vacancy\_status (0.26)

**5. Visualization and Outlier Detection**

* **Distribution of Total Consumption**:
  + Histograms and boxplots were used to visualize energy consumption.

A graph of a number of consumption

Description automatically generated

* + Outliers were detected on both ends.

A diagram of a cross

Description automatically generated

* **Time Series Plot**:
  + Energy consumption trends were visualized for specific building IDs (e.g., 65 and 121).



The graph shows the total consumption of two buildings (65 and 121) over a period of time, likely in July. The x-axis represents time, while the y-axis shows the total consumption.

Consumption fluctuates for both buildings, with peaks and valleys throughout the month. Building 65 generally has higher consumption compared to Building 121.

**6. Summary Statistics**

* **Basic Stats**:
  + total\_consumption ranges from **0 to ~15.4** (after removing negatives).
  + in.sqft: Range **328 to 8194**.
  + Key descriptive statistics like mean, median, and quartiles were generated.

**7. Importance of Preprocessing Steps**

1. **Handling Missing Data**: Ensured model performance by eliminating incomplete records.
2. **Removing Redundant Features**: Boosted efficiency by dropping constant and low-variance columns.
3. **Cleaning Invalid Entries**: Addressed errors (negative values) for reliability.
4. **Feature Selection**: Identified top correlated predictors to reduce dimensionality.
5. **Encoding Categorical Data**: Ensured compatibility for machine learning models.

**Exploratory Data Analysis (EDA) Report**

**Energy Consumption vs Other Factors**

The EDA process performed in the PDF file explores relationships between **energy consumption** (total\_consumption) and various factors influencing it, including cooling/heating setpoints, occupancy, building characteristics, and water heater efficiency. Below is a structured and detailed report of the findings:

**1. Correlation Analysis**

* **Objective**: Identify linear relationships between energy consumption and numeric features.
* **Method**: Pearson correlation coefficients.

**Top Correlated Features** with total\_consumption:

1. in.sqft → 0.37 (strong positive correlation with energy usage).
2. in.geometry\_floor\_area\_bin → 0.34.
3. in.vacancy\_status → 0.26.
4. in.bedrooms → 0.19.

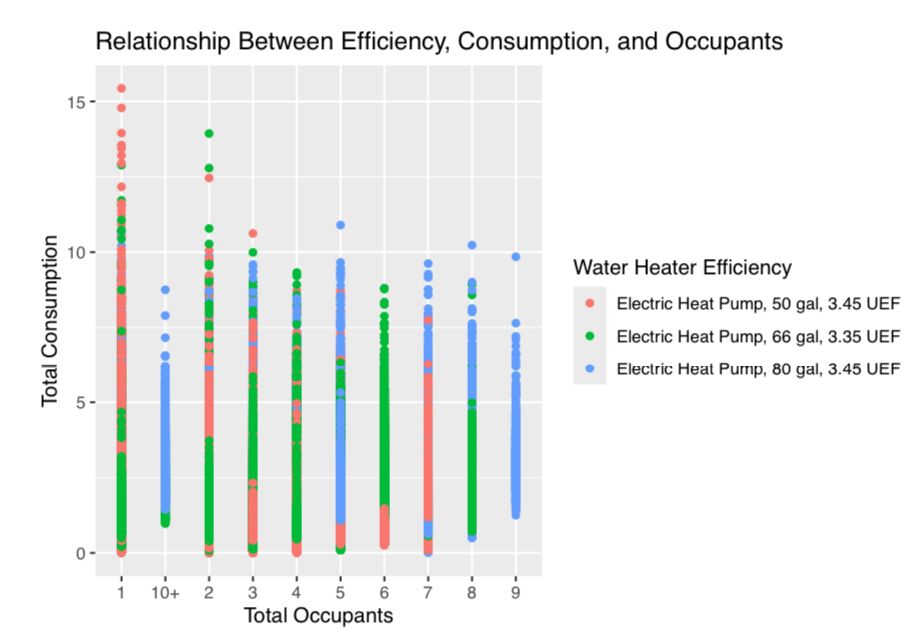
These features were used later as predictors in machine learning models.

**2. Relationship Between Water Heater Efficiency, Occupants, and Consumption**

* **Objective**: Understand the effect of **water heater efficiency** and **occupancy** on energy usage.
* **Method**: Scatter plot of total\_consumption against in.occupants, colored by upgrade.water\_heater\_efficiency.

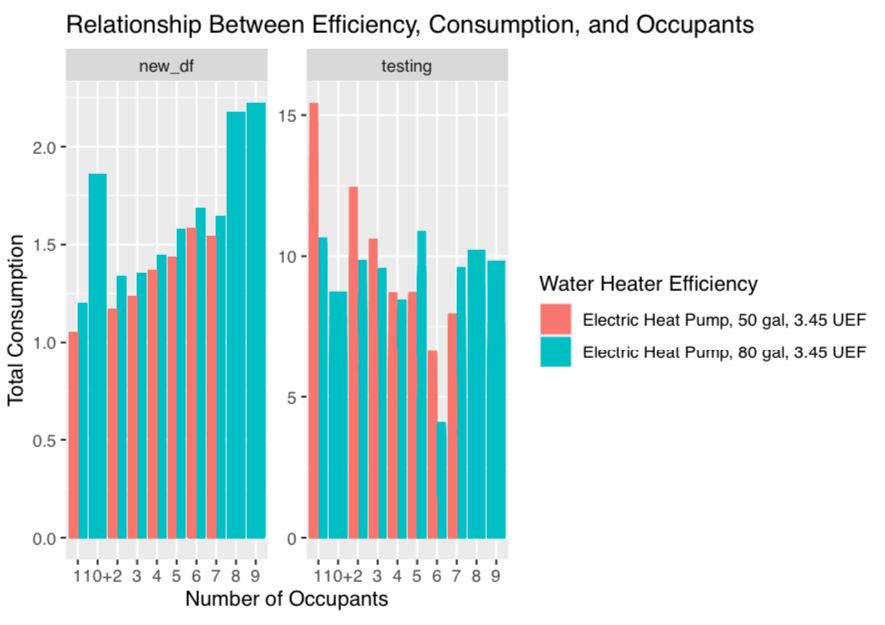
**Insights**:

* Higher water heater efficiencies, such as Electric Heat Pump, 50 gal, 3.45 UEF, tend to have **lower energy consumption**.
* Energy consumption increases with the **number of occupants** in the building.



* **Efficiency and Consumption:** Water heaters with higher energy factor (UEF) ratings (3.45 UEF) tend to have lower total consumption for the same number of occupants. This suggests that more efficient water heaters can lead to lower energy costs.
* **Occupants and Consumption:** As the number of occupants increases, total consumption generally increases for all water heater types. This is likely due to increased hot water usage with more people in the household.
* **Overlap:** There is some overlap in consumption levels between different water heater types, especially at lower occupant numbers. This indicates that other factors beyond efficiency, such as usage patterns, can also influence water consumption.

**Overall, the graph suggests that choosing a water heater with a higher UEF rating can help reduce energy consumption, especially in households with a larger number of occupants.**



**Key Observations:**

* **Efficiency and Consumption:** Water heaters with higher energy factor (UEF) ratings (3.45 UEF) tend to have lower total consumption for the same number of occupants. This suggests that more efficient water heaters can lead to lower energy costs.
* **Occupants and Consumption:** As the number of occupants increases, total consumption generally increases for all water heater types. This is likely due to increased hot water usage with more people in the household.
* **Overlap:** There is some overlap in consumption levels between different water heater types, especially at lower occupant numbers. This indicates that other factors beyond efficiency, such as usage patterns, can also influence water consumption.

**Overall, the graph suggests that choosing a water heater with a higher UEF rating can help reduce energy consumption, especially in households with a larger number of occupants.**

**4. Statistical Analysis: Cooling and Heating Setpoints**

* **Objective**: Analyze the impact of cooling and heating setpoints on energy consumption.
* **Method**:
  + Fitted a linear regression model.
  + Examined quadratic effects of in.cooling\_setpoint and in.heating\_setpoint.

**Regression Results**:

* **Significant Variables**:
  + in.cooling\_setpoint (negative impact, higher setpoints reduce energy usage).
  + in.heating\_setpoint (positive impact, higher heating setpoints increase energy usage).
* **R-squared**: 0.098 (indicating weak explanatory power of linear models).

**5. Random Forest Analysis**

* **Objective**: Assess the importance of features in predicting energy consumption and capture non-linear relationships.
* **Method**: Multiple **Random Forest** regression models with different sets of predictors.

**Model 1: Cooling and Heating Setpoints**

* Predictors:
  + in.cooling\_setpoint,
  + in.cooling\_setpoint^2,
  + in.heating\_setpoint,
  + in.heating\_setpoint^2.
* **Results**:
  + Variance Explained: **16.05%**.
  + Importance: Heating and cooling setpoints are significant contributors to energy usage.

**Model 3: Building Characteristics**

* Predictors:
  + in.sqft, in.bedrooms, in.occupants, in.geometry\_floor\_area\_bin.
* **Results**:
  + Variance Explained: **20.74%**.
  + Feature Importance:
    1. in.sqft → **Top contributor**.
    2. in.geometry\_floor\_area\_bin.
    3. in.occupants.

**Model 4: Combining Features**

* Predictors:
  + in.sqft,
  + in.occupants,
  + in.geometry\_floor\_area\_bin,
  + in.cooling\_setpoint,
  + in.heating\_setpoint.
* **Results**:
  + Variance Explained: **28.21%**.
  + Feature Importance:
    1. in.sqft.
    2. in.geometry\_floor\_area\_bin.
    3. in.heating\_setpoint.

**Importance of Features** (Sorted):

| **Feature** | **Importance Score** |
| --- | --- |
| in.sqft | 2341.15 |
| in.geometry\_floor\_area\_bin | 2193.80 |
| in.heating\_setpoint | 1676.23 |
| in.cooling\_setpoint | 868.27 |
| in.occupants | 808.62 |

**Key Findings**

1. **Building Size and Area** (in.sqft and in.geometry\_floor\_area\_bin) are the strongest predictors of energy consumption.
2. **Heating Setpoints** have a significant positive impact on energy usage, whereas **cooling setpoints** reduce it.
3. Higher water heater efficiencies are associated with lower consumption, but their impact is relatively small.
4. The number of **occupants** moderately influences energy consumption.
5. Time series analysis reveals differences in usage patterns between buildings, possibly due to size, location, or occupancy.

**Analysis Report: Energy Consumption Trends and Prediction Using Random Forest**

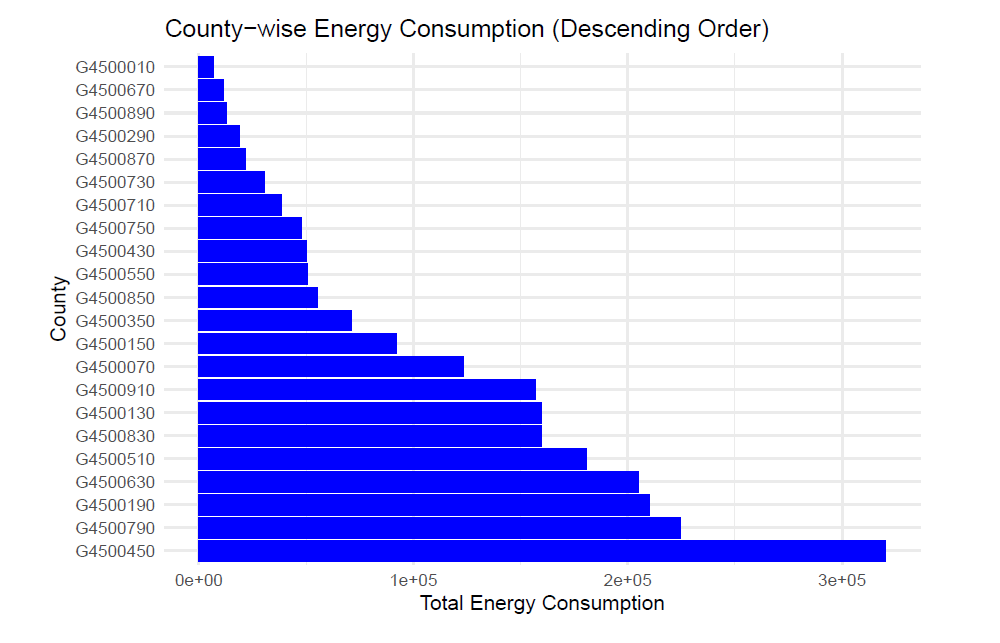
**1. Data Aggregation and Visualization**

Objective:

- Summarize daily energy consumption by county and visualize the trends.

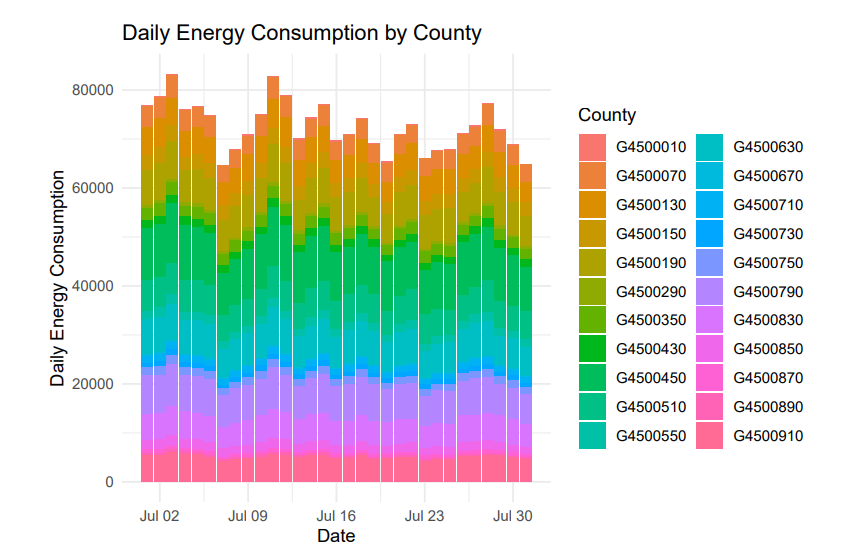
Process:

1. Data was grouped by date (Date) and county (in.county), and the total daily energy consumption was calculated:



This graph displays county-wise energy consumption in descending order, with each bar representing a county and its total energy consumption. Counties are sorted from the highest consumer at the bottom (e.g., G4500450) to the lowest at the top (e.g., G4500010). The varying bar lengths highlight significant differences in energy consumption across counties, with some consuming considerably more than others. This visualization helps identify high-energy-consuming counties, which can inform resource allocation and energy efficiency initiatives.

**2. A bar plot was generated using ggplot2 to visualize daily energy consumption trends:**



This stacked bar chart illustrates daily energy consumption across multiple counties over a month. Each bar represents a day, with different segments showing the contributions of individual counties to the total daily energy consumption. The variation in bar heights indicates fluctuating overall energy usage, peaking at certain intervals. The stacked format highlights the relative contribution of each county, showing that some consistently consume more energy than others, which can aid in identifying trends and patterns in energy usage across counties over time.

Key Findings:

- The chart shows daily energy consumption trends across different counties. Observed trends may indicate consumption spikes or drops on certain dates, likely influenced by external factors like weather or regional events.

**2. Data Preparation for Prediction**

Objective:

- Split the dataset into training and testing sets for predictive modeling.

Process:

- Training and testing data were created with the variables Dry.Bulb.Temperature...C and total\_consumption.

- The testing data's temperature was adjusted by adding 5.0 to simulate a modified scenario:

**3. Random Forest Model for Prediction**

Objective:

- Develop a regression model to predict energy consumption based on dry bulb temperature.

Process:

- A random forest model was trained on a sample of 50,000 data points using the randomForest package:

Model Evaluation Metrics:

1. Mean Squared Residuals: 0.6141

2. Explained Variance (%): 12.21%

3. RMSE: 0.9355

4. R-squared: 0.0757

5. MAE: 0.7540

- The model showed limited predictive power, as indicated by the low R-squared value (7.57%).

**4. Model Prediction and Comparison**

Objective:

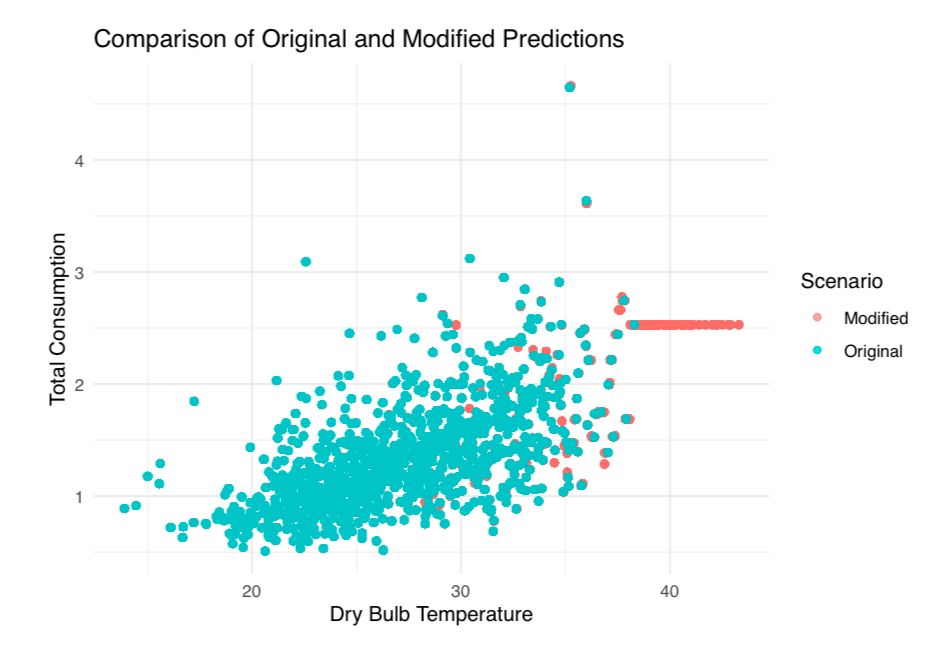
- Compare predictions for the original and modified datasets.

Process:

1. Predictions were made for the testing dataset

2. A combined dataset (comparison\_data) was created for original and modified scenarios

3. A scatter plot was generated to visualize predictions for both scenarios:



This scatter plot compares the relationship between dry bulb temperature and total energy consumption under two scenarios: "Original" (blue) and "Modified" (red). Each point represents a data observation, with the x-axis indicating the dry bulb temperature and the y-axis representing total energy consumption. The blue points (original predictions) show a broader distribution across temperatures and consumption levels, while the red points (modified predictions) appear concentrated at a specific level of total consumption, especially at higher temperatures. This suggests the modified predictions might be constrained or adjusted compared to the original predictions.

Findings:

- The comparison plot highlights the shift in predictions between original and modified temperatures.

**5. Statistical Analysis of Mean Differences**

Objective:

- Assess the distribution of mean differences between predicted and actual consumption.

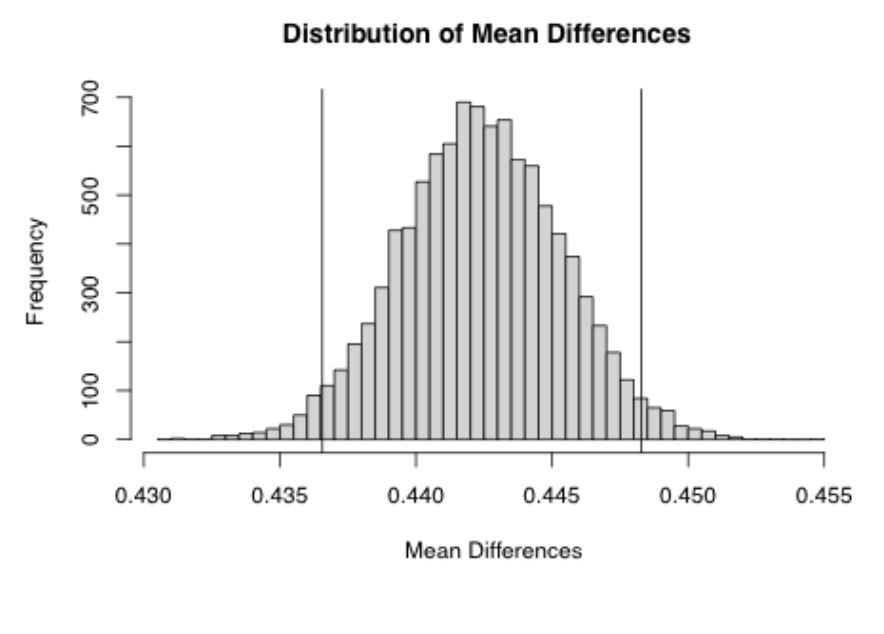
Process:

- Mean differences were sampled 10,000 times to build a distribution:

- A histogram was created to visualize the distribution, and the 95% confidence interval was calculated:

Findings:

- Confidence interval of mean differences: 0.4366 to 0.4483.



This graph displays the distribution of mean differences. The x-axis shows the range of mean differences, and the y-axis represents the frequency of each mean difference. The distribution appears to be roughly bell-shaped, with the majority of mean differences clustered around 0.440. There are two vertical lines that seem to mark off the middle 95% of the distribution. This suggests that the mean difference is likely to fall within this range with 95% confidence. Overall, the graph indicates that the mean differences are relatively small and centered around 0.440.

**SUMMARY**

This report offers a comprehensive analysis of energy consumption in South Carolina, with a focus on residential usage patterns and the factors that influence them. The project aims to address several challenges faced by energy providers, including predicting future energy usage, understanding the impact of weather, and developing strategies for energy efficiency.

The scope of the analysis encompasses an examination of residential energy usage and weather data. This involves data collection, cleaning, analysis, and visualization. The dataset includes information on housing characteristics, energy consumption, and weather conditions for approximately 5,000 homes. Data preparation consisted of cleaning, preprocessing, and merging datasets, with a particular emphasis on July data for a detailed analysis.

Exploratory data analysis investigated daily consumption patterns, energy levels by location, and the effects of various factors on energy use. A Random Forest model was employed, explaining 20.74% of the variance in energy consumption. Key findings reveal that water heater efficiency, occupancy rates, building size, and weather conditions significantly influence energy usage.

The project provides valuable insights for demand forecasting, efficiency improvements, and policymaking aimed at optimizing energy consumption in South Carolina.