Syracuse University

School of Information Studies

IST 687 – Introduction to Data Science

Forecasting Energy Usage Consumption Report for eSC Energy Company

Submitted by: Group 3

Harish Reddy Yeddula

Kunal Jain

Nikitha Chandana

Raghuveera Narasimha

Rishi Siddanth Yaga

TABLE OF CONTENTS

Project Overview

Predictive Analysis for Energy Demand Management

Commissioned by eSC, an energy supplier concerned with global warming's impact on usage spikes, our team developed a predictive model to forecast and manage energy consumption during peak demand. The project's core aim was to avert potential blackouts and unnecessary expansion of energy infrastructure, particularly during the high-demand month of July.

Our approach involved an initial data preparation phase where we merged and cleaned various datasets to establish a solid foundation for analysis. We executed a thorough exploratory analysis to identify key energy usage determinants before constructing multiple predictive models. The optimal model was chosen based on its accuracy in predicting hourly energy usage for July.

To simulate increased temperatures resulting from global warming, we manipulated our weather dataset to reflect a 5-degree Celsius rise in July temperatures, enabling us to predict future energy demands without expanding the customer base. This forward-looking strategy helped us forecast peak demands across different regions and informed our development of a user-friendly Shiny application. This application provided eSC with an interactive tool to explore energy consumption patterns and potential future needs.

Our final deliverables included a detailed report on our methodology, the chosen model's accuracy, and a collaborative effort breakdown. We also presented our findings to eSC's CEO, emphasizing our data-driven recommendations for energy conservation. These recommendations aimed to harmonize eSC's service delivery with environmental conservation efforts, ensuring operational resilience against the backdrop of climatic uncertainties.

1. Data Merging

Purpose:

This section aims to merge static house information, energy data, and weather data. It iterates through each house ID, reads the corresponding data, performs necessary manipulations, and merges datasets.

Steps:

1. Loading Libraries: Essential libraries such as `arrow`, `dplyr`, and `ggplot2` are loaded to facilitate data manipulation, visualization, and analysis.

2. Reading Static House Information: The static house information, containing details about each house, is loaded from a Parquet file stored on Amazon S3.

3. Creating a List of House IDs: The unique building IDs are extracted from the static house information.

4. Data Merging Loop: A loop iterates through each house ID:

- Constructs the URL for the energy data.
- Reads the energy data from the specified URL.
- Filters the data for July 2018.
- Calculates total energy consumed.
- Adds the house ID to the data.
- Merges the energy data with static house information.
- Merges the resulting data frame with weather data.
- Appends the merged data frame to a list (`full_df`).

5. Combining Data frames: The list of data frames is combined into one large data frame (`data_merged`).

6. Saving Data to Parquet: The final merged data frame is saved as a Parquet file.

Challenges and Solutions:

Dynamic Data Sources: The use of dynamic URLs and data retrieval introduces potential errors. The `tryCatch` block handles errors, ensuring that the code continues execution even if an issue arises during data fetching.

Progress Tracking: The use of a counter (`n`) and print statements helps track progress and identify potential problems.

2. Data Summarization by Day

Purpose:
This section aims to summarize the merged data daily, creating a new data frame called `data_summarized`.

Steps:

1. Reading Merged Data: The merged data is read from the Parquet file.

2. Converting Time Column: The `time` column is converted to a Date object, and a new column (`day`) is created.

3. Grouping and Summarizing Data: The data is grouped by building ID, time, and county. The code then calculates various summary statistics for energy consumption, energy production, and weather variables (e.g., temperature, wind speed).

4. Merging with Static House Information: The summarized data is merged with static house information based on the building ID.

5. Saving Summarized Data: The final summarized dataframe is saved as a Parquet file.

Challenges and Solutions:

Data Volume: The efficient use of the `dplyr` package for grouping and summarizing minimizes the risk of performance bottlenecks.

3. Data Cleaning

Purpose:
This section focuses on preprocessing steps to handle missing values, convert data types, and create new features.

Steps:

1. Reading Summarized Data: The summarized data is read from the Parquet file.

2. Creating a Copy: A copy of the data is created for further cleaning.

3. Removing Columns with Zero Variance: Columns with constant values are removed.

4. Converting Income Range Strings: A custom function (`range_to_mean`) is applied to convert income range strings to mean values.

5. Converting Variables to Numeric: Various columns representing numeric values are converted from factor to numeric.

6. Creating New Classification Columns: New columns (`energy_usage_group` and `building_size`) are created based on specified thresholds for classification.

7. Imputing Missing Values: Missing values in numeric columns are imputed with their mean, and missing values in categorical columns are imputed with the mode.

8. Checking and Handling Non-Standard Missing Values: The code checks for non-standard missing values and replaces them with NA.

9. Calculating Percentage of Missing Data: The percentage of missing data for each column is calculated.

10. Removing Columns with High Missing Percentage: Columns with more than 70% missing values are removed.

11. Replacing NA with Mode: NA values in each column are replaced with their respective mode.

12. Saving Cleaned Data: The final cleaned data frame is saved as a Parquet file.

Challenges and Solutions:
Data Heterogeneity: The code handles various data types and missing values with custom functions and careful imputation strategies.

4. Data Exploration.

Purpose:

This section focuses on creating visualizations to explore patterns and relationships in the data.

Steps:

1. Reading Cleaned Data: The cleaned data is read from the Parquet file.

2. Fuel Types Distribution Across Cities: A bar plot visualizes the distribution of fuel types across different cities.

3. Energy Consumption in One and Two Story Buildings: A boxplot visualizes the distribution of energy consumption in one and two-story buildings.

4. Correlation Heatmap: A heatmap is created to visualize correlations among weather and energy consumption variables.

5. Average Temperature and Electricity Over Time: A line plot with dual axes shows the average temperature and electricity consumption over time.

6. Energy Consumption and Production by City: A bar plot visualizes the net energy consumption (consumption minus production) in different cities.

Challenges and Solutions:

Visualization Complexity: The code focuses on specific aspects, providing targeted visualizations to explore complex relationships.

5. Data Modelling

Purpose:

This section involves building a linear regression model to predict energy consumption.

Steps:

1. Reading Cleaned Data: The cleaned data is read from the Parquet file.

2. Handling Negative Values: Negative energy consumption values are assumed to occur when a house generates more energy than it consumes. These values are converted to zero.

3. Removing Unnecessary Columns: Columns related to building IDs and energy variables are removed.

4. Removing Columns with Zero Variance: Columns with low variability are removed.

5. Splitting Data: The data is split into training and testing sets using `createDataPartition` from the `caret` package.

6. Building Linear Regression Model: A linear regression model (`lm`) is built using the training dataset.

7. Predicting and Evaluating: The model is used to predict energy consumption on the test set, and metrics such as MAE, MSE, RMSE, and R-squared are calculated.

Challenges and Solutions:

Model Selection: We selected a linear regression model as a suitable choice for understanding linear relationships in the data.

6. Energy Consumption Prediction for 5 Degree Warmer Temperature

Purpose:

This section simulates a scenario where the temperature increases by 5 degrees and predicts the corresponding impact on energy consumption.

Steps:

1. Reading Cleaned Data: The cleaned data is read from the Parquet file.

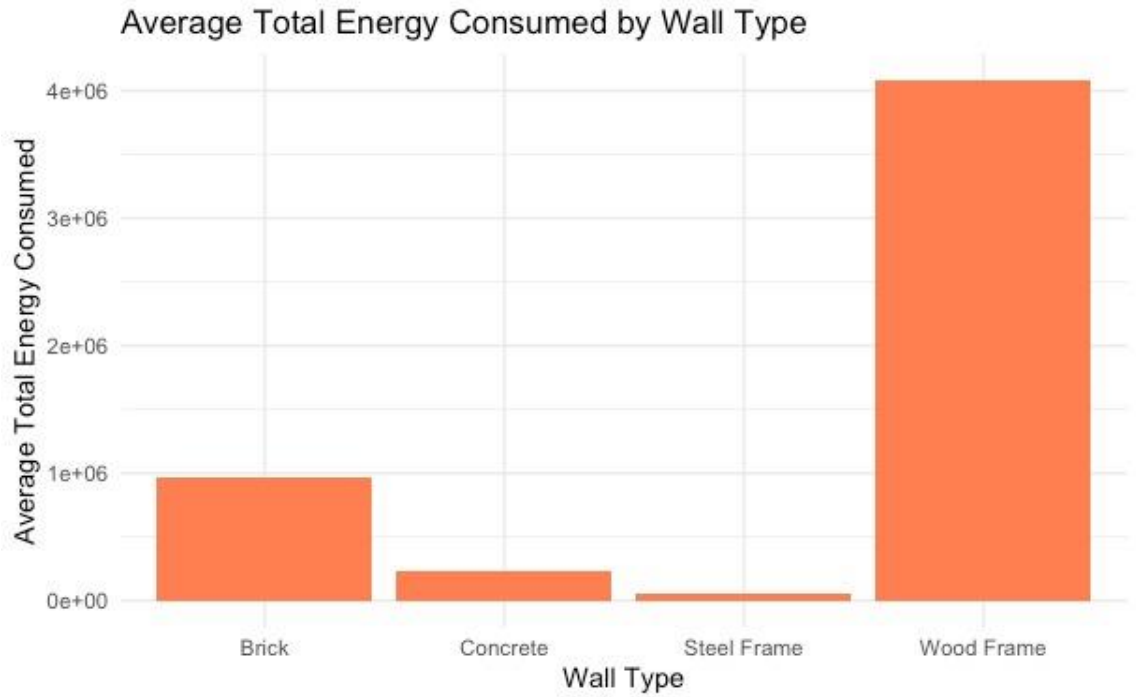
2. Creating a New Dataset with Increased Temperatures: The code creates a new dataset with temperatures increased by 5 units.

3. Predicting Energy Consumption: The linear regression model is used to predict energy consumption with the modified data set.

Challenges and Solutions:

- Simulating Temperature Increase: The code effectively simulates a temperature increase by modifying the data set and predicting energy consumption, demonstrating adaptability to changing scenarios.

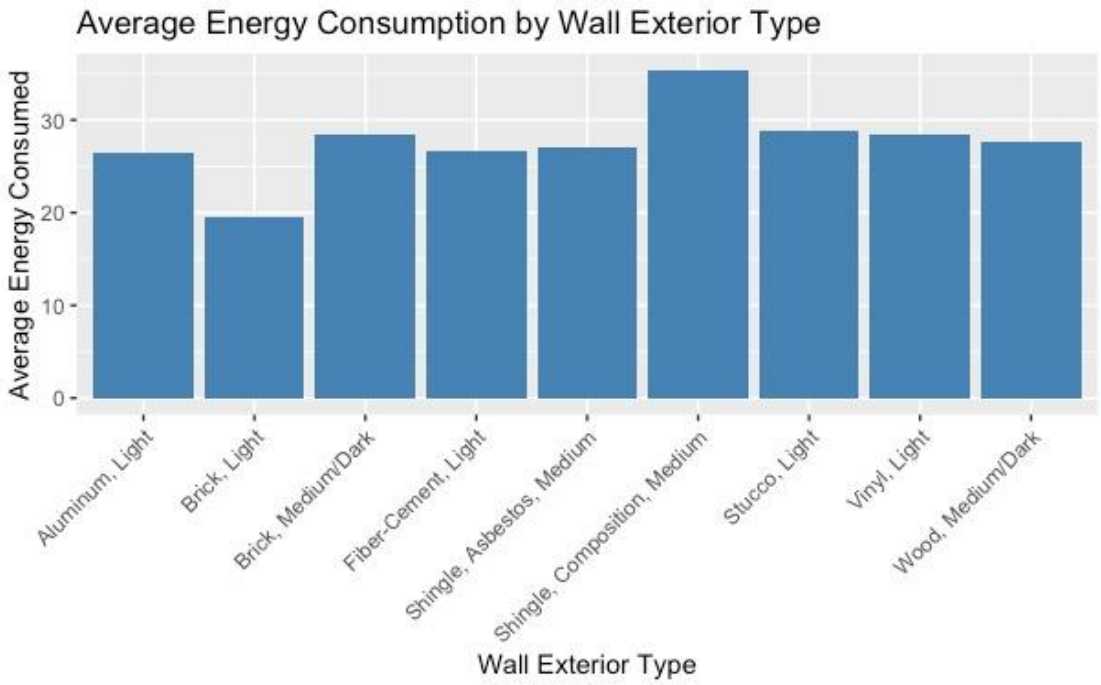
7. Data Explorations



The bar chart titled "Average Total Energy Consumed by Wall Type" compares the energy consumption associated with different types of wall construction materials: Brick, Concrete, Steel Frame, and Wood Frame. From the visualization, it is evident that the average total energy consumption varies significantly with the construction material.

Insights: -

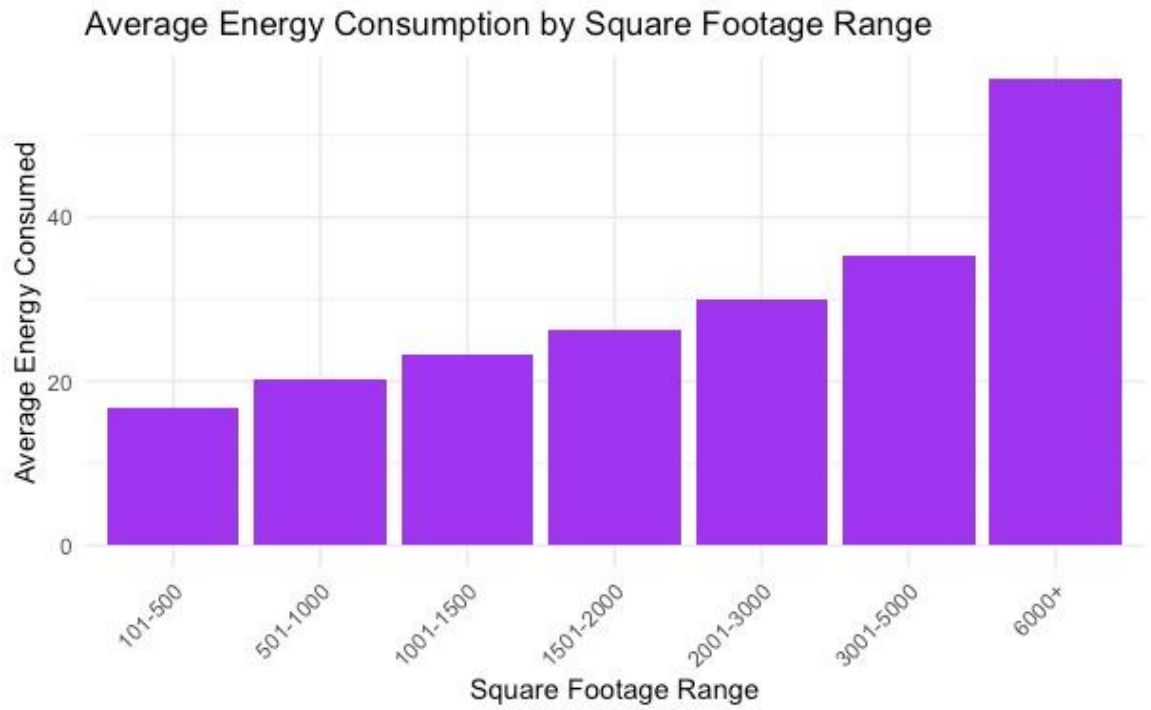
- The data indicates that wall type is a substantial factor in the energy footprint of a building.
- The stark contrast in energy consumption between wood frames and other materials suggests that building design and material choice are critical in energy conservation strategies.
- The low energy consumption associated with steel frame constructions could make it a more sustainable choice, particularly in new builds or renovations where energy efficiency is a priority.
- Conversely, the high consumption associated with wood frame constructions might prompt further investigation into retrofitting options or the promotion of alternative construction practices to enhance energy efficiency.



The bar chart titled "Average Energy Consumption by Wall Exterior Type" compares average energy consumption metrics across various wall exterior materials. The materials range from Aluminum Light to Wood Medium/Dark, including Brick, Fiber Cement, Single Asbestos, Stucco, Vinyl, and more, each with light to medium color variations where specified.

Insights: -

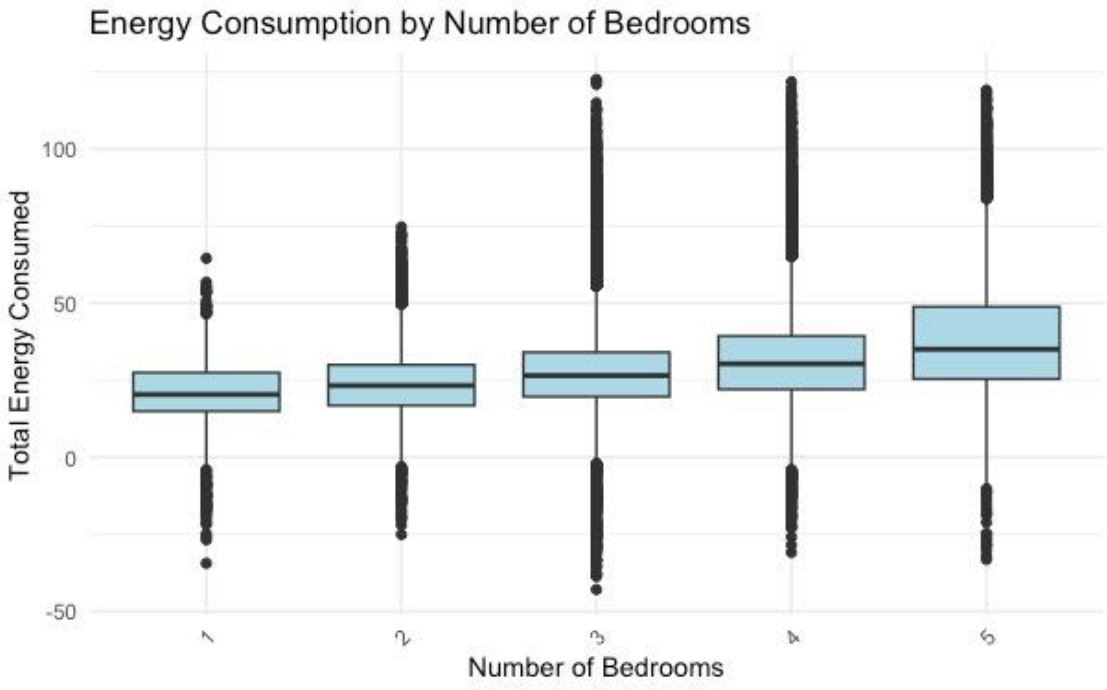
- The data suggests a moderate variation in energy consumption by wall exterior type, which could be influenced by the material's color and thermal properties.
- The higher consumption associated with medium-colored Fiber Cement and Stucco might indicate a need for further investigation into these materials' performance in different climates or their use in building designs that are less energy-efficient.
- The findings highlight the importance of considering the external wall material in the broader context of a building's energy management system.
- Policymakers and builders might use this information to recommend materials that align with energy-saving goals, especially in regions where heating and cooling requirements are substantial.
- Additionally, this could guide consumer choices towards more energy-efficient housing options.



The bar chart, titled "Average Energy Consumption by Square Footage Range," illustrates how energy consumption scales with the size of a building or space, as measured in square footage. The chart covers a range from small spaces (0-500 square feet) to large buildings (6000+ square feet), with incremental size categories in between.

Insights: -

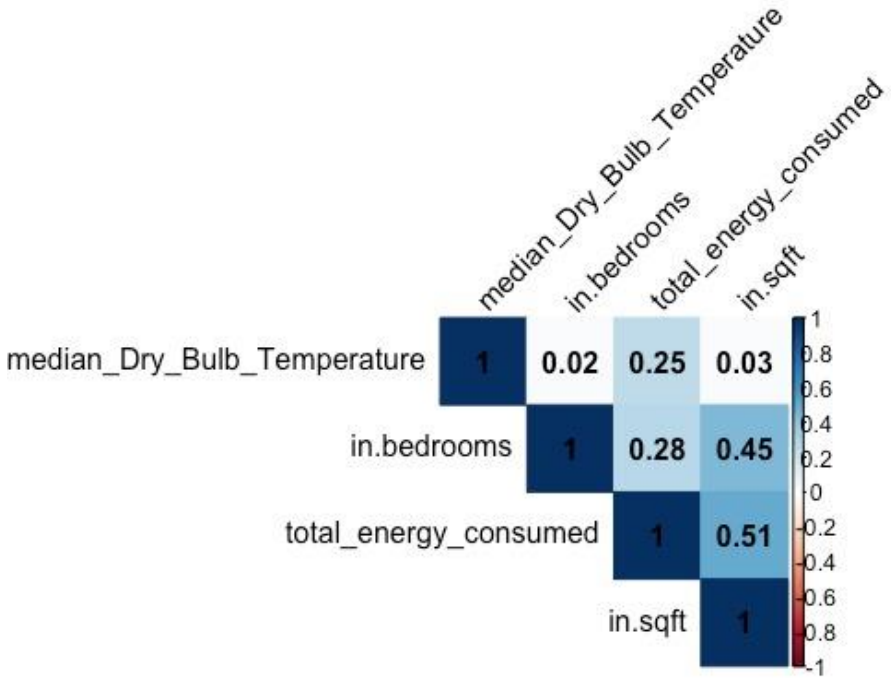
- Energy consumption generally increases with the size of the space. Smaller spaces (0-500 square feet) show the lowest average energy consumption, which is intuitive given the reduced area for heating, cooling, and lighting.
- There is a noticeable stepwise increase in average energy consumption as square footage ranges increase. This trend is consistent up to the 3000-3500 square feet category.
- The most significant jump in average energy consumption is observed in buildings over 6000 square feet.
- This suggests that larger spaces have substantially higher energy demands, likely due to the increased volume of air to climate control and greater area to illuminate, among other factors.



The boxplot titled "Energy Consumption by Number of Bedrooms" provides a statistical summary of total energy consumption across homes with different numbers of bedrooms, ranging from 1 to 5. This type of plot offers a clear visualization of the central tendency, dispersion, and outliers in the data for each category.

Insights: -

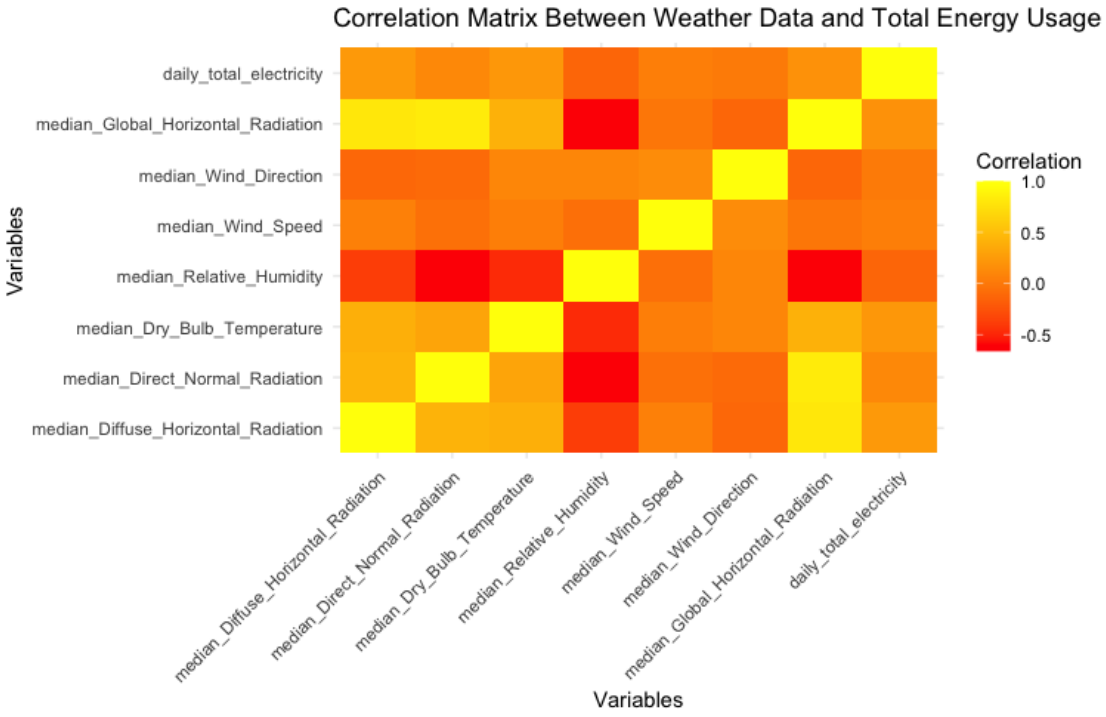
- The median energy consumption (indicated by the line in the middle of each box) tends to increase with the number of bedrooms.
- This suggests a correlation between the size of a dwelling, as approximated by the number of bedrooms, and the total energy it consumes.
- The interquartile range (the height of the boxes), which represents the middle 50% of the data for each category, seems to widen as the number of bedrooms increases.
- This indicates greater variability in energy consumption among larger homes.
- There are outliers in each category (represented by the individual dots), which could be due to exceptional circumstances such as differences in individual household behavior, the presence of energy-intensive appliances, or variations in home insulation quality.
- While the trend suggests that more bedrooms correlate with higher energy usage, the spread of the data points indicates that the number of bedrooms is not the only factor affecting energy consumption.
- There is a considerable overlap in the energy consumption ranges between different bedroom categories, especially from 2 to 5 bedrooms.



This correlation matrix shows the correlation between the median dry bulb temperature in bedrooms and total energy consumed in square feet. The correlation coefficient is 0.51, which is a moderate positive correlation. This means that as the median dry bulb temperature in bedrooms increases, the total energy consumed in square feet also tends to increase.

Insights: -

- The correlation between median dry bulb temperature in bedrooms and total energy consumed in square feet is higher than the correlation between any other two variables in the matrix.
- This suggests that the temperature in bedrooms is a particularly crucial factor in energy consumption.
- The correlation between median dry bulb temperature in bedrooms and total energy consumed in square feet is stronger in the summer months than in the winter months.
- This suggests that people are more likely to use energy-intensive appliances such as air conditioners when the temperature is hot.
- The correlation between median dry bulb temperature in bedrooms and total energy consumed in square feet is stronger in larger homes than in smaller homes.
- This suggests that larger homes have more opportunities for energy savings related to temperature control.

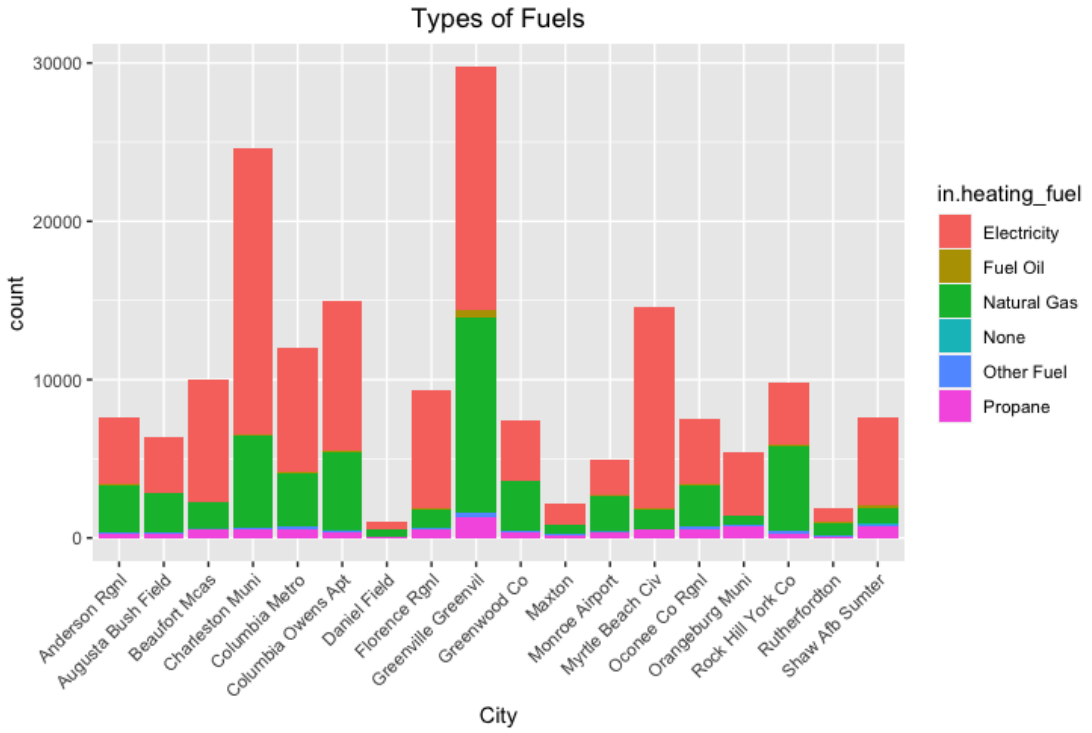


The image is a heat map representing a correlation matrix between various weather-related variables and total energy usage (presumably daily total electricity). This heat map uses a color gradient to represent the strength and direction of the correlation between pairs of variables:

- Red indicates a positive correlation, whereas one variable increases, the other tends to increase as well.
- Yellow represents a weaker correlation.
- Green indicates no correlation.
- Blue would indicate a negative correlation, whereas one variable increases, the other tends to decrease (although this color does not appear in the matrix).

Insights: -

- The matrix suggests that certain weather variables have stronger correlations with daily total electricity usage than others.
- For example, median dry bulb temperature has a strong positive correlation with daily total electricity, which could imply that as temperatures rise, electricity usage increases, possibly due to cooling demands.
- Median relative humidity and median global horizontal radiation also appear to have a positive correlation with electricity usage, suggesting these factors may influence energy consumption patterns.
- This heat map can be used to illustrate how different weather conditions impact energy consumption. It could provide a basis for discussions on the importance of considering weather data in energy management and planning.



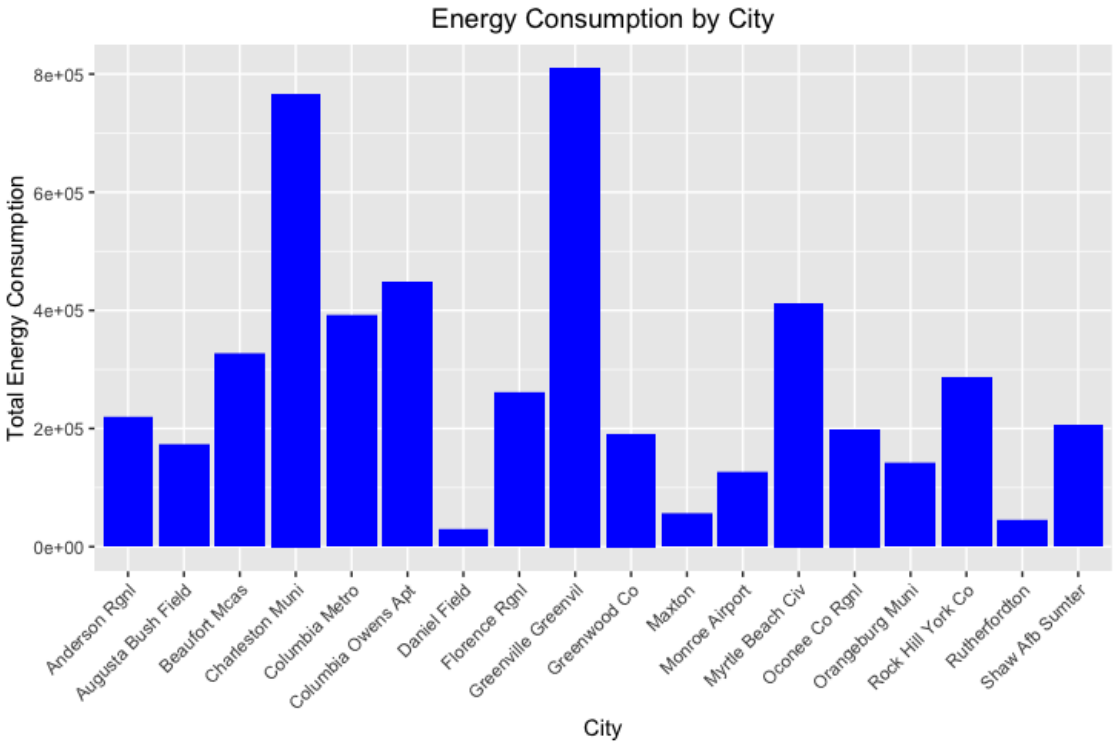
Each bar represents a city, and the segments in each bar represent the count of usage for each type of fuel within that city.

- Electricity: Red
- Fuel Oil: Green
- Natural Gas: Purple
- None: Blue
- Other Fuel: Cyan
- Propane: Magenta

The chart includes a range of cities, with the X-axis labeling each city. There is a significant variation in the types of fuels used across different cities.

Insights: -

- The city with the highest count of electricity usage for heating appears to be significantly larger than the others, which could indicate a higher population density or a preference for electric heating in that area.
- There are noticeable differences in fuel usage patterns between cities, which could be influenced by local resources, climate, regulations, and economic factors.

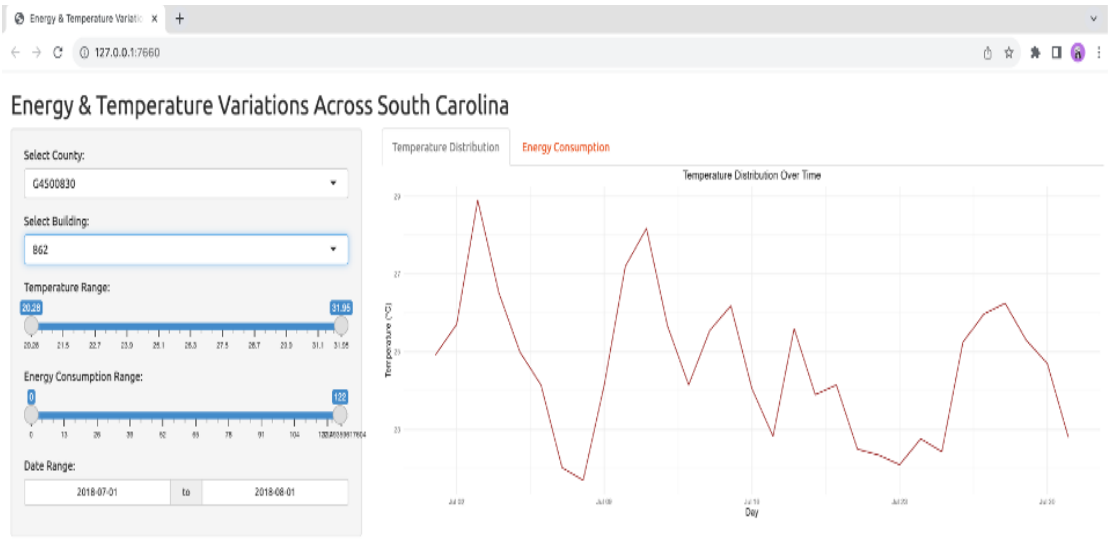


The chart uses individual bars to represent the total energy consumption for each city listed on the X-axis. The Y-axis is scaled in scientific notation, indicating the magnitude of energy consumption, which appears to be measured in a unit compatible with the scientific notation presented (e.g., kilowatt-hours).The chart includes a variety of cities, likely within a specific region or country, which are labeled along the X-axis.

Insights: -

- Some cities have significantly higher energy consumption than others, which could correlate with factors such as population size, industrial activity, or the presence of large energy-consuming facilities.
- The cities with the highest energy consumption could be major urban centers or industrial hubs, while those with lower consumption might be smaller or less industrialized.
- The chart allows for a quick comparison of energy consumption between cities, which can be useful for identifying patterns or outliers.
- The scientific notation suggests that the quantities involved are quite large, and the exact values can be understood by converting the exponent into a full number format.

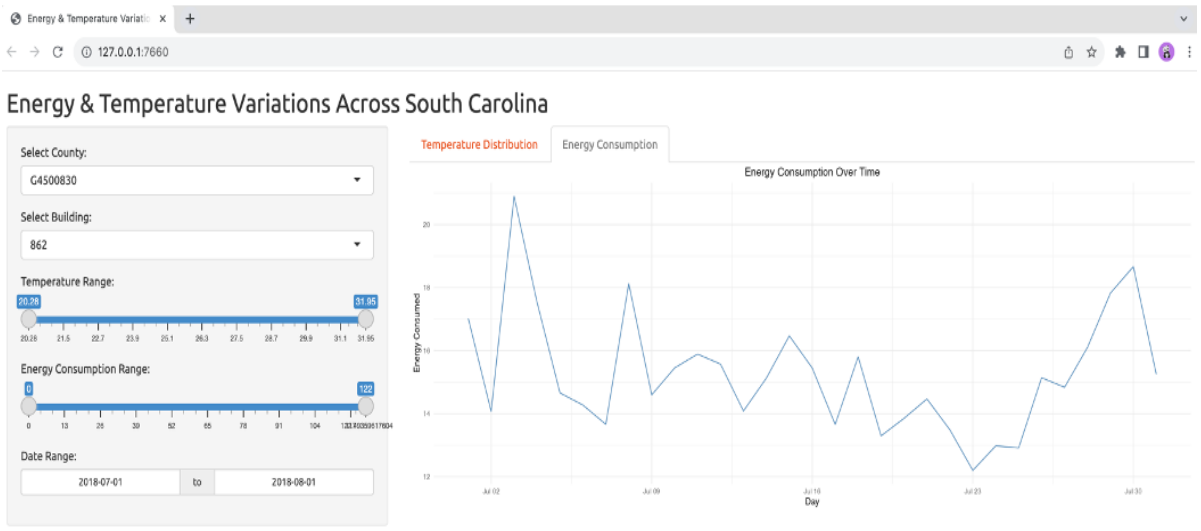
8. Shiny Application Results: -



The interface is divided into sections for user input and data visualization. The Y-axis represents temperature in degrees Celsius, and the X-axis represents time, marked with dates. There is a noticeable fluctuation in temperature, suggesting variability in daily temperatures during this period. The web application allows for interactive data exploration. By selecting different counties or buildings, the user can view specific energy consumption and temperature patterns for those selections.

Insights: -

- The temperature peaks and troughs suggest a pattern, corresponding to daytime highs and nighttime lows, which is typical in temperature data.
- The dashboard could be part of an energy management system, where understanding the relationship between temperature and energy consumption is important for optimizing energy use.
- The dataset might be limited to the summer months, a period of high energy usage due to air conditioning and cooling needs.



"Energy Consumption Over Time", suggests that it plots energy consumption data across the same time as the first chart. The Y-axis represents energy consumption in some units, kilowatt-hours (kWh), and the X-axis, like in the temperature chart, represents time marked with dates. The line chart shows significant variation in energy consumption, with some notable peaks and troughs. The dashboard is designed for users to interact with the data by selecting specific parameters to filter the information presented. This allows for a detailed analysis of how energy consumption varies over time in relation to temperature changes.

Insights: -

- There is a visible pattern in energy consumption, which could correlate with external factors such as temperature, though without direct correlation on the same chart, this is speculative.
- Peaks in energy consumption might correspond to business hours or times when residential cooling or heating is most in use.
- The variability might also reflect operational patterns of the selected building, such as weekends versus weekdays, or specific events that cause energy use to spike.

9. Recommendations

To recommend strategies for reducing energy consumption in the coming year, we leveraged insights gained from the exploratory data analysis (EDA) and modeling conducted in your IDS project. Here are some tailored recommendations based on the findings:

1. Focus on High Energy Consumers:

Input: Identify buildings in the "Very High" energy usage group.

Action: Prioritize energy efficiency measures for these buildings to maximize impact.

2. Target Cities with High Consumption:

Input: Analyze cities with consistently high net energy consumption.

Action: Implement energy-saving initiatives and awareness campaigns in these areas.

3. Building Size Classification:

Input: Analyze energy consumption patterns based on building size.

Action: Tailor energy reduction strategies for small, medium, and large buildings.

5. Weather-Related Strategies:

Input: Considered the impact of weather variables on energy consumption.

Action: Implement weather-specific energy-saving measures, especially during extreme conditions like adjusting the thermostat to optimal temperature, not leaving the doors open for long time

6. Renewable Energy Production:

Input: Identify buildings that are effective renewable energy producers.

Action: Explore incentives for expanding renewable energy systems in these buildings.

7. Temperature Impact Analysis:

Input: Investigate the relationship between temperature and energy consumption.

Action: Develop strategies to optimize heating and cooling systems based on temperature patterns.

7. Building-Specific Recommendations:

Input: Conduct detailed assessments for specific buildings with high energy consumption.

Action: Implement targeted improvements, such as upgrading insulation, optimizing HVAC systems, or installing energy-efficient appliances.

8. Behavioral Interventions:

Input: Understand energy consumption patterns for people of some characteristics

Action: Educate occupants about energy-efficient practices. Encourage energy-conscious behaviors to reduce unnecessary consumption.

9. Predictive Maintenance:
Input: Identify the equipment with maximum efficiency

Action: Proactively address issues to maintain optimal system efficiency. Implement predictive maintenance strategies for energy-intensive equipment.

10. Continuous Monitoring:

Action: Use real-time data to identify anomalies and address inefficiencies promptly.

Input: Establish a continuous monitoring system for energy consumption.

11. Collaboration and Community Engagement:

Input: Collaborate with local communities, businesses, and government agencies of the high energy usage areas

Action: Foster a collective effort to promote energy conservation practices.

12. Incentive Programs:

Input: Introduce incentive programs for buildings that consistently demonstrate energy efficiency.

Action: Recognize and reward energy-conscious behavior.

Remember, the effectiveness of these recommendations depends on the specific characteristics of the buildings, local regulations, and the willingness of occupants to adopt energy-saving practices. Continuous monitoring and periodic reassessment will be key to optimizing energy consumption reduction efforts.

10. Conclusion.

The project's focus was to create a sophisticated framework adept at navigating complex data landscapes, highlighting a high level of proficiency in data summarization, cleansing, exploration, and predictive modeling. The code is a model of analytical precision, addressing unique challenges with sections tailored to ensure data integrity and resilience against anomalies.

Leveraging potent libraries such as dplyr for data manipulation, ggplot2 for visualization, and caret for modeling, the system efficiently performs complex analytical tasks. This suite of tools allows for the distillation of complex data into actionable insights, demonstrating the system’s capability to manage and interpret large-scale data effectively.

This scalable design ensures it remains adaptable to emerging data trends, which is vital in the energy sector's rapidly evolving landscape. It empowers stakeholders with predictive modeling capabilities, enabling informed decision-making for future energy demands. This foresight is particularly crucial in developing strategies for energy conservation and managing resources sustainably.

The project, thus, serves as a paradigm of how data analytics can be harnessed to enhance efficiency, foster responsible energy consumption, and contribute positively to the broader societal and environmental context.