IST 687

Introduction to Data Science

Final Project Report

Group 2

Submitted by

Smita Deulkar
Harika Gangu
Indraneel Karandikar
Sukesh Meda
Sejal Sardal

Under the guidance of

Professor Erik Anderson

Spring 2024

Table of Contents

Project Overview

Goal

Build a machine learning model to address concerns about increased energy demand during hot summers by understanding energy usage drivers. Develop strategies to encourage energy conservation and use data-driven modeling to predict and manage peak energy demand.

Description

This project entails collaboration with eSC, an energy company servicing residential properties in South Carolina and parts of North Carolina. With mounting concerns over the impact of rising temperatures on energy demand, particularly during peak periods like July, the aim is to devise data-driven solutions to manage energy consumption efficiently, mitigate blackout risks, and promote sustainability. Tasks include analysing energy usage data, constructing predictive models, devising conservation strategies, and developing interactive tools to aid the company in making informed decisions regarding energy demand management.

By addressing these challenges, eSC seeks to cultivate awareness about energy usage and savings among residents, ultimately reducing overall energy consumption and avoiding the need for constructing new power plants.

Business Requirements

Energy Demand Management: Develop strategies to effectively manage energy demand during peak periods, especially in July, to prevent potential blackouts and ensure uninterrupted energy supply to customers.

Sustainability Promotion: Promote sustainability by reducing overall energy consumption and mitigating the need for constructing new power plants, thereby minimizing environmental impact.

Informed Decision-Making: Predict the future energy usage and provide actionable insights and tools to assist the energy company in making informed decisions about energy demand management and resource allocation.

Customer Engagement: Raise awareness among customers about alternative energy sources and saving practices to encourage active participation in conservation efforts, fostering a culture of responsible energy consumption.

General Overview

Data Provided

Static House Data:

Contains static information about residential properties served by the energy company eSC in South Carolina. It contains data on about 5,000 single-family homes that use eSC energy. It has house attributes such as the building ID, house size, and other static details. The file is saved in the ’parquet’ format, which is optimal for storage.

Energy Usage Data:

Provides hourly energy consumption statistics for each residence. It contains hourly energy usage data, including consumption from various sources such as air conditioning systems, dryers, and other appliances. Each residence has one dataset file that contains validated energy usage that is traced on hourly basis. Describes each house's energy usage per hour from numerous sources (e.g., air conditioning system, dryer).

Meta Data:

A description file that has information on the fields used in the various housing data files. A simple, human-readable CSV file with attribute descriptions.

Weather Data:

Contains hourly weather data for a specific geographic area or county within South Carolina. Time-series weather data is gathered and stored based on a county code. Each house's county code is identified by the 'in.county' column in the Static House Data. The file is stored in a CSV format which is easy to use.

Tasks and Deliverables

Data Collection and Integration: Collect and integrate diverse datasets, including energy usage data, weather data, and house information into a single merged dataset.

Exploratory Analysis: Gain insights on the merged data. Clean the data and make it effective and usable.

Model Development and Evaluation: Develop predictive models using machine learning and statistical techniques. Evaluate the accuracy and performance of these models using appropriate metrics and validation techniques to forecast energy usage based on different strategies.

Demand Analysis: Assume high temperatures, daily usage and seasonal variations in temperature to predict the future demand.

Demand reduction strategy: Propose strategies to enable reduction in demand based on the analysis made. This should help eSC in reducing their overhead costs.

Tool Development(Shiny): Develop an interative application to present to the CEO of eSC.

Other Technical Concerns

Scalability and Performance: Design the solution to be scalable and performant, capable of handling large volumes of data from thousands of residential properties while ensuring real-time or near-real-time processing capabilities.

Security and Privacy: Secure sensitive data, comply with data privacy laws, preserve consumer information, and uphold confidence by putting strong security measures in place.

Maintenance and Support: Establish procedures for ongoing maintenance, monitoring, and support of the implemented solution to ensure its reliability and effectiveness over time.

Documentation and Knowledge Transfer: Create comprehensive documentation, including data dictionaries, model documentation, and user guides, to facilitate knowledge transfer and support ongoing maintenance and future enhancements.

Detailed Overview

Data Collection and Integration

1) Cleaning Static House Information:

- The script extracts data from a Parquet file containing details about residential properties.
- It selectively removes specified columns based on predefined criteria.
- The resulting dataset is stored as "static house info."

2) Calculating Total Energy for Each Building:

- An empty dataframe named "result df daywise" is initialized to aggregate total energy usage for individual buildings.
- The script iterates through each building, retrieves its energy consumption data from the Parquet file, and calculates the total energy used in July.
- The calculated totals are appended to "result df daywise."

3) Processing July Weather Data:

- The script fetches weather data for various counties during the month of July.
- Median values for different weather parameters are computed.
- The corresponding columns in "static house info" are updated with these median values.

4) Creating Final Output Dataframe:

- A final output dataframe, "merge static house info df4," is constructed by selecting specific columns from "static house info".

5) Merging Dataframes:

- The script aims to merge multiple data frames:
  I. static house info (original dataset)
  II. result df daywise (energy dataset daywise)
  III. weather final data(weather dataset including date and county)
  IV. static house info df1 (merging of static house info and result df datewise)
  V. merge static house info df (merging of static house info df1 and weather final including date and county)
  VI. merge static house info df4 (dataset after cleaning the original one)
- A left join is performed using the merge function and later using the left join function from the dplyr package.
- Columns not necessary for further analysis are removed from the merged dataframe.

Exploratory Analysis

6) Understand the Data:

- Examine the datasets using functions like str(), head(), summary(), and dim().
- Identify the types of variables (numerical, categorical, date/time, etc.) and search for any data that is missing.

7) Descriptive Statistics:

- Use summary() to compute basic statistics for numerical variables.

- Use histograms, box plots, or density plots to examine the distribution of numerical variables.

8) Categorical data:

- To analyse the distribution of categorical data, use frequency tables, bar charts, or pie charts.
- Search for uncommon or unusual categories.

9) Correlation Analysis:

- To examine correlations between numerical data, utilize correlation matrices or scatter plots.
- Boxplots are a great visual aid for correlation analysis.

10) Data Cleaning:

- To deal with missing values, apply imputation or elimination.
- Look for anomalies and outliers, and then decide whether to keep or delete them based on domain expertise.

Model Development and Evaluation

11) Data Splitting:

- The dataset is split into training (80%) and testing (20%) sets using the createDataPartition function.

12) Managing Categorical Variables:

- The unique values in the character columns are located.
- Include only those rows in the test data that have values that match the unique values in the training data.

13) Finding Constant Variables:

- Locating single level variables(constant variables) in both training and testing data and eliminating them.

14) Linear Regression Model:

- The training set of data is fitted to a linear regression model using the lm function.

15) Evaluation of the Model:

- Predictions are based on the test results, and the median, minimum, and maximum of the target variable are calculated and reported.
- The Mean Absolute Percentage Error, or MAPE, is a used for evaluating the accuracy of a model.

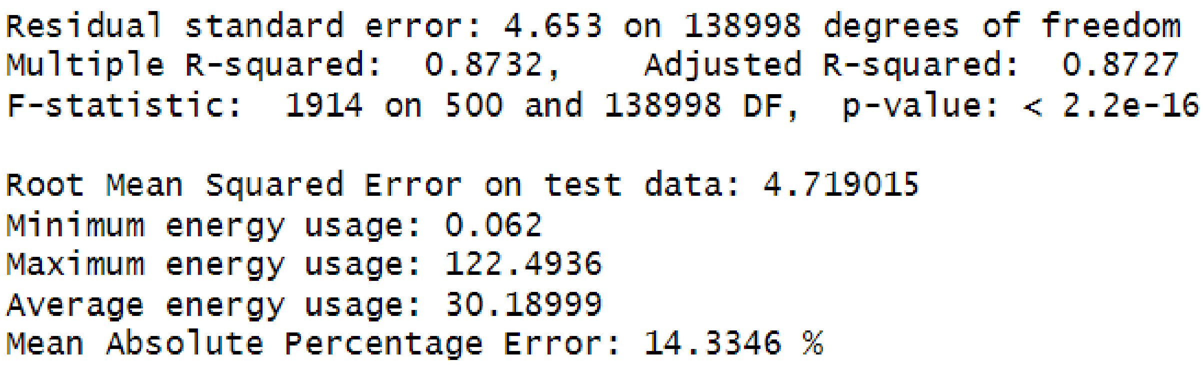
16) Multiple R-squared (R^2):

- This metric indicates the extent to which the predictors account for the variance in the response variable (total energy).
- In this case, the model accounts for about 87.27% of the variation.
- The number of predictors in the model is taken into consideration by the Adjusted R-squared, a shortened version of R-squared.

17) P-value:

- The model's p-value for the F-statistic is very near to zero (higher than 2.2e-16), suggesting that it may be statistically significant.
- This disproves the null hypothesis, which states that all coefficients are 0.

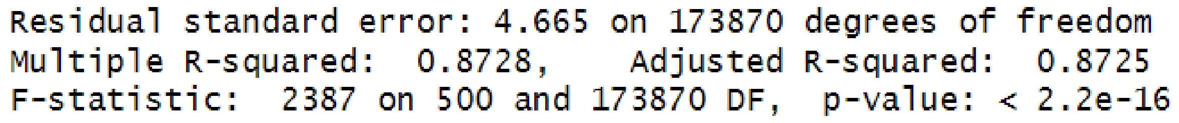
Fig: Per hour energy usage prediction



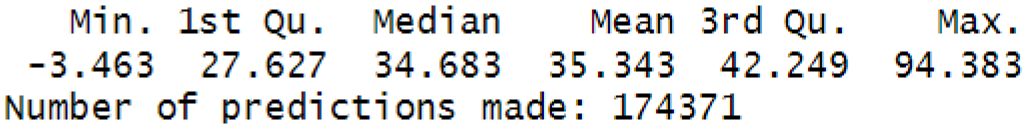
```
Residual standard error: 4.653 on 138998 degrees of freedom
Multiple R-squared:  0.8732,    Adjusted R-squared:  0.8727
F-statistic:  1914 on 500 and 138998 DF,  p-value: < 2.2e-16

Root Mean Squared Error on test data: 4.719015
Minimum energy usage: 0.062
Maximum energy usage: 122.4936
Average energy usage: 30.18999
Mean Absolute Percentage Error: 14.3346 %
```

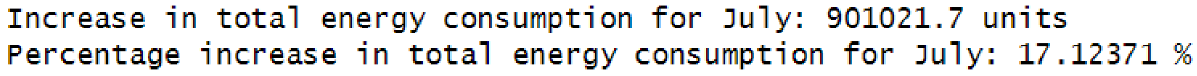
Fig: Peak future energy demand



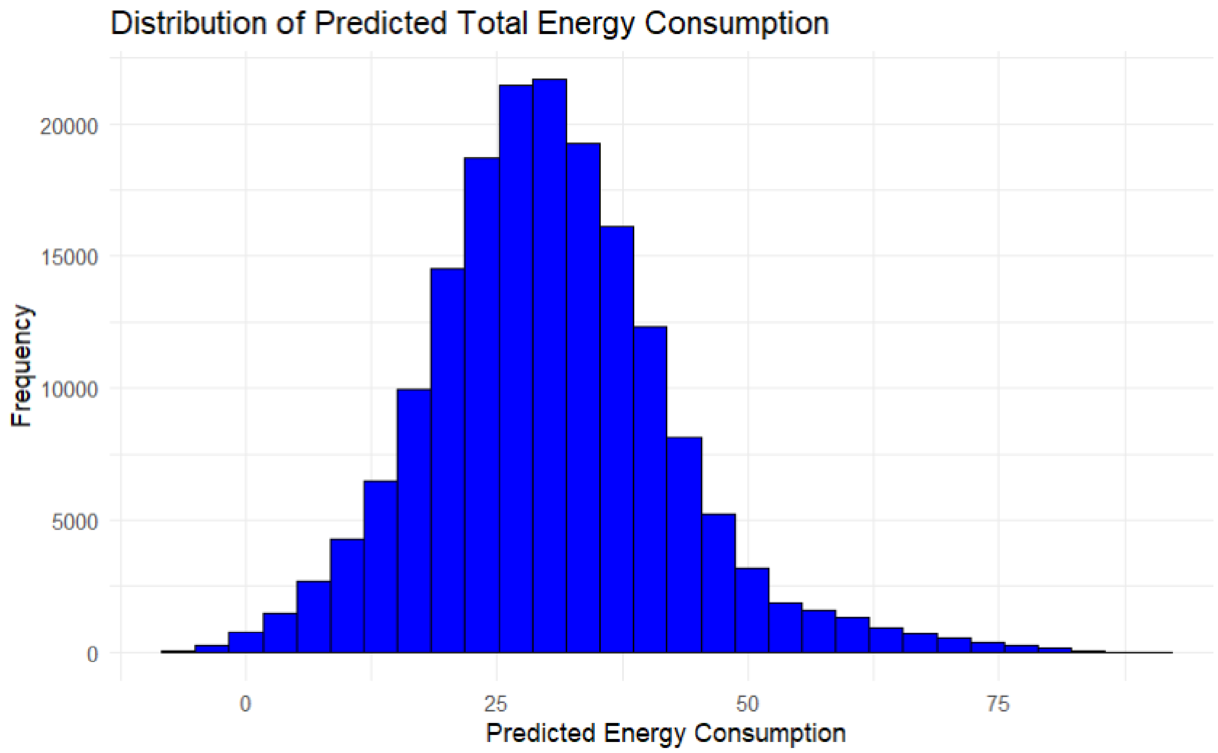
```
Residual standard error: 4.665 on 173870 degrees of freedom
Multiple R-squared:  0.8728,    Adjusted R-squared:  0.8725
F-statistic:  2387 on 500 and 173870 DF,  p-value: < 2.2e-16
```



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 -3.463  27.627  34.683  35.343  42.249  94.383 
Number of predictions made: 174371
```



```
Increase in total energy consumption for July: 901021.7 units
Percentage increase in total energy consumption for July: 17.12371 %
```

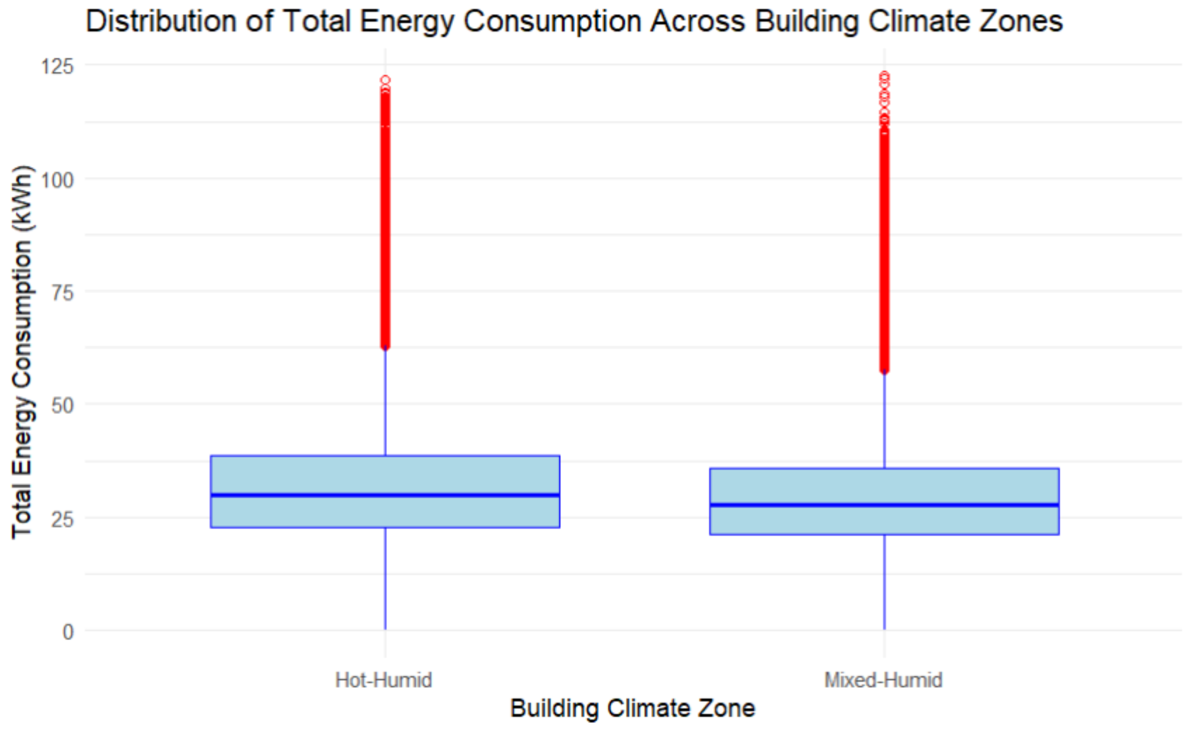


The above histogram shows the distribution of predicted total energy consumption values. The x-axis represents the predicted energy consumption values, while the y-axis represents the frequency or count of occurrences for each energy consumption value.

- The distribution follows an approximately normal or bell-shaped curve.
- The peak of the distribution lies around the value of 40-45 units, suggesting that this range represents the most common or typical predicted energy consumption.
- The distribution is slightly skewed to the right, which could potentially be attributed to factors such as variations in weather conditions, housing characteristics, or other variables that influence energy consumption.
- It allows for assessing the central tendency (mean or median) and the spread or variability of the predicted values, which can be useful for understanding the general patterns and identifying potential outliers or deviations from the expected energy consumption levels.

This type of analysis can be valuable for energy demand management strategies, as it helps to identify the typical range of energy consumption and potential areas where targeted interventions or conservation efforts may be most effective.

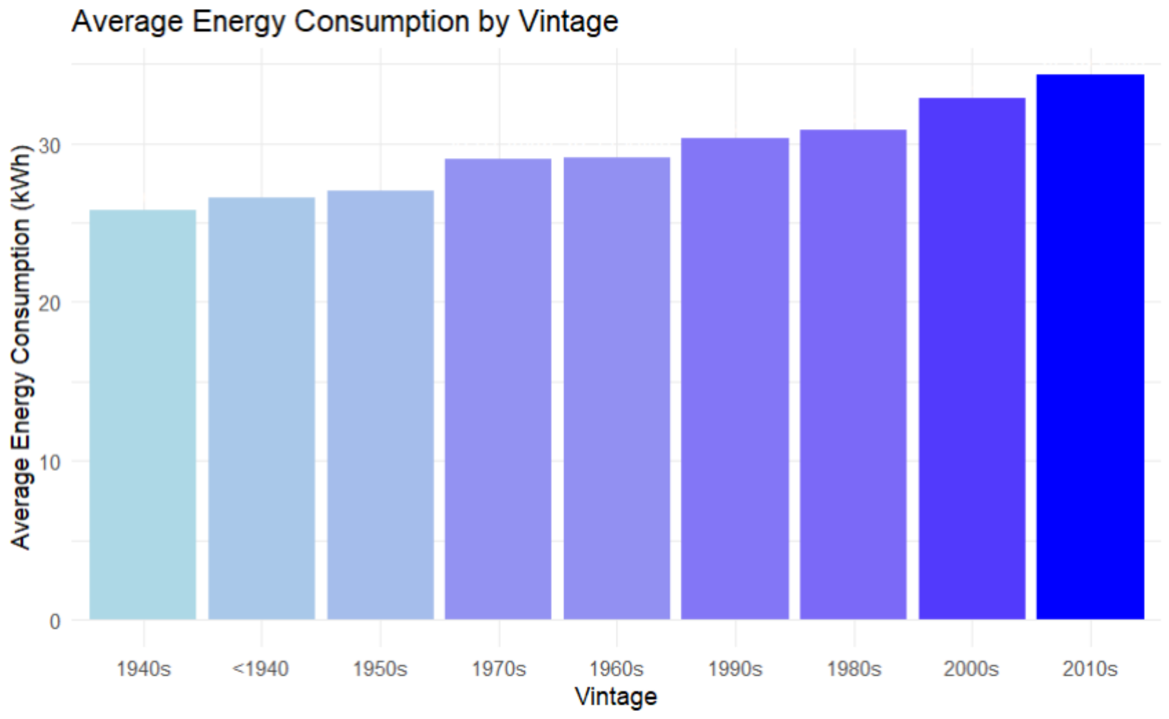
Demand Analysis



The above box plot displays the distribution of total energy consumption across different building climate zones. The x-axis represents the building climate zones ("Hot-Humid" and "Mixed-Humid.") The y-axis shows the total energy consumption in kilowatt-hours (kWh).

- **Energy consumption variation:** The box plot shows the variation in energy consumption within each climate zone. It represents the interquartile range (IQR), which covers 50% of the data points, while the whiskers extend to the minimum and maximum values within a certain range.
- **Comparison of median energy consumption:** The median energy consumption, represented by the horizontal line within each box, is higher for the "Mixed-Humid" climate zone (around 25 kWh) compared to the "Hot-Humid" climate zone (around 25 kWh).
- **Outliers:** The visualization reveals the presence of outliers, represented by individual data points (red dots) above the whiskers. These outliers indicate buildings with exceptionally high energy consumption compared to the rest of the data in their respective climate zones.
- **Spread of energy consumption:** The length of the boxes and whiskers provides insight into the spread of energy consumption. The "Mixed-Humid" climate zone appears to have a larger spread, with a wider box and longer whiskers, indicating greater variability in energy consumption among buildings within this zone.
- **Climate zone influence:** The difference in energy consumption patterns between the "Hot-Humid" and "Mixed-Humid" climate zones suggests that climate characteristics may play a role in influencing energy consumption.

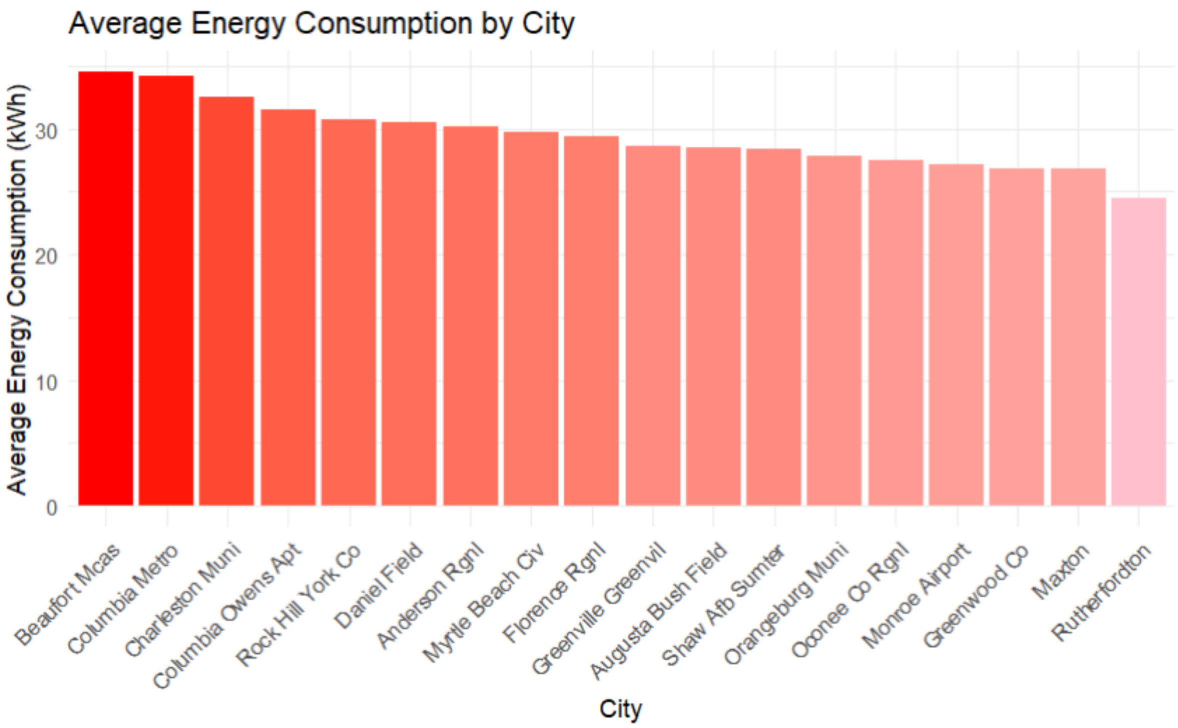
Overall, this visualization allows for a comparative analysis of energy consumption patterns across different building climate zones. It highlights the potential impact of climate conditions on energy usage and identifies outliers that may warrant further investigation or energy-efficiency measures.



The above bar chart depicts the average energy consumption across different vintages (time periods) of buildings. The x-axis shows the vintage categories, ranging from buildings constructed in the 1940s to those built in the 2010s, including a category for buildings older than 1940. The y-axis represents the average energy consumption in kilowatt-hours (kWh).

- There is a clear upward trend in average energy consumption as the vintage of buildings becomes more recent. Buildings constructed in the 2010s have the highest average energy consumption, while those from the 1940s have the lowest.
- The difference in average energy consumption between the oldest (1940s) and newest (2010s) vintages is substantial, indicating that older buildings tend to be more energy-efficient or have lower energy demands.
- There is a noticeable increase in average energy consumption starting from buildings constructed in the 1980s, suggesting potential changes in building practices, materials, or energy-consuming appliances and systems over time.
- The gradual increase in average energy consumption across vintages could be attributed to factors such as larger building sizes, increased use of energy-intensive appliances and systems, or changes in insulation and construction techniques affecting energy efficiency.
- The visualization highlights the importance of considering the vintage or age of buildings when analyzing and addressing energy consumption patterns, as newer buildings may require different energy-saving strategies compared to older ones.

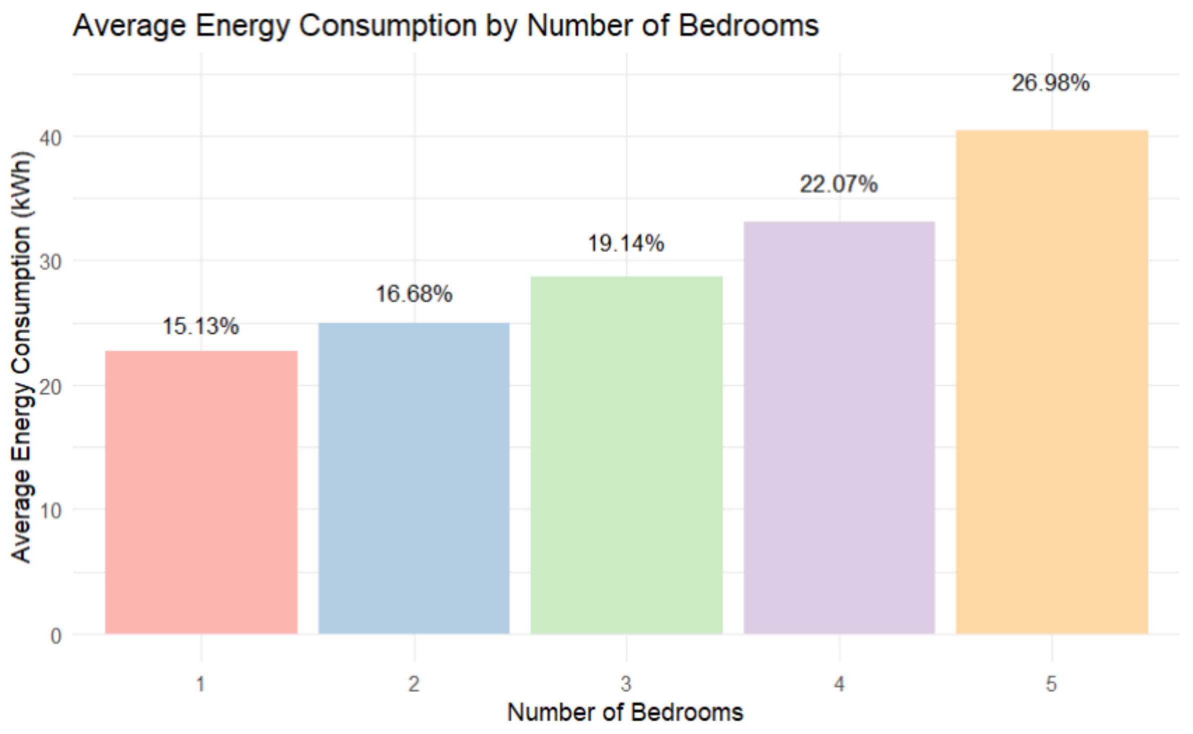
Overall, this bar chart effectively illustrates the relationship between the vintage of buildings and their average energy consumption, providing valuable insights for energy conservation efforts and building energy management strategies.



The bar chart displays the average energy consumption across different cities in the region. The y-axis represents the average energy consumption, while the x-axis lists the city names.

- There is a significant variation in average energy consumption among the cities shown.
- Batesburg-Leesville and Columbia have the highest average energy consumption, far exceeding the other cities.
- Cities like Ridgeland, Ruffin, Newberry, North Augusta, and Rock Hill have relatively lower average energy consumption compared to the others.
- Most of the cities cluster around a similar, moderate level of average energy consumption, except for the two outliers with much higher values.
- The disparities in average energy consumption could be influenced by factors such as population density, climate conditions, prevalence of energy-intensive industries, building characteristics, and energy efficiency practices within each city.
- Identifying the underlying reasons behind the high or low energy consumption patterns in specific cities could provide valuable insights for developing targeted energy conservation strategies or policies.

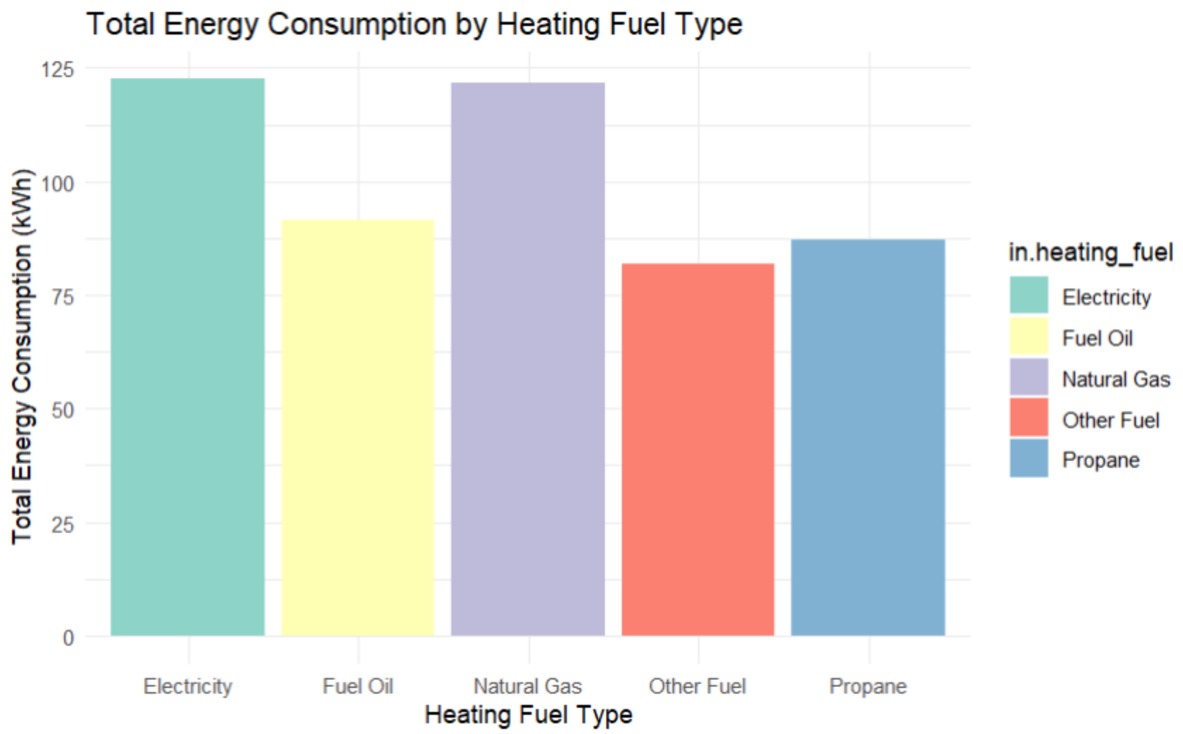
Overall, the visualization effectively highlights the significant variations in average energy consumption across different cities, emphasizing the importance of considering location-specific factors when addressing energy consumption issues and implementing energy management initiatives.



The bar chart illustrates the relationship between the number of bedrooms in a residential property and the average energy consumption. The x-axis represents the number of bedrooms, ranging from 1 to 5, while the y-axis shows the average energy consumption in kilowatt-hours (kWh).

- **Positive correlation:** The average energy consumption and the number of bedrooms are positively correlated. With the increase in the number of bedrooms, the average energy usage rises.
- **Substantial increase in energy consumption**: The increase in average energy consumption is substantial as the number of bedrooms increases. For instance, properties with 5 bedrooms have an average energy consumption of around 26.98 kWh, which is significantly higher than properties with only 1 bedroom (15.13 kWh).
- **Larger households:** Properties with more bedrooms typically accommodate larger households or families. The higher energy consumption can be attributed to increased occupancy and the associated energy demands for heating, cooling, lighting, and the operation of various appliances and electronics.
- **Energy efficiency considerations:** The visualization highlights the importance of considering the number of bedrooms or household size when evaluating energy efficiency and developing energy conservation strategies. Larger households may require different approaches or targeted interventions to reduce energy consumption effectively.

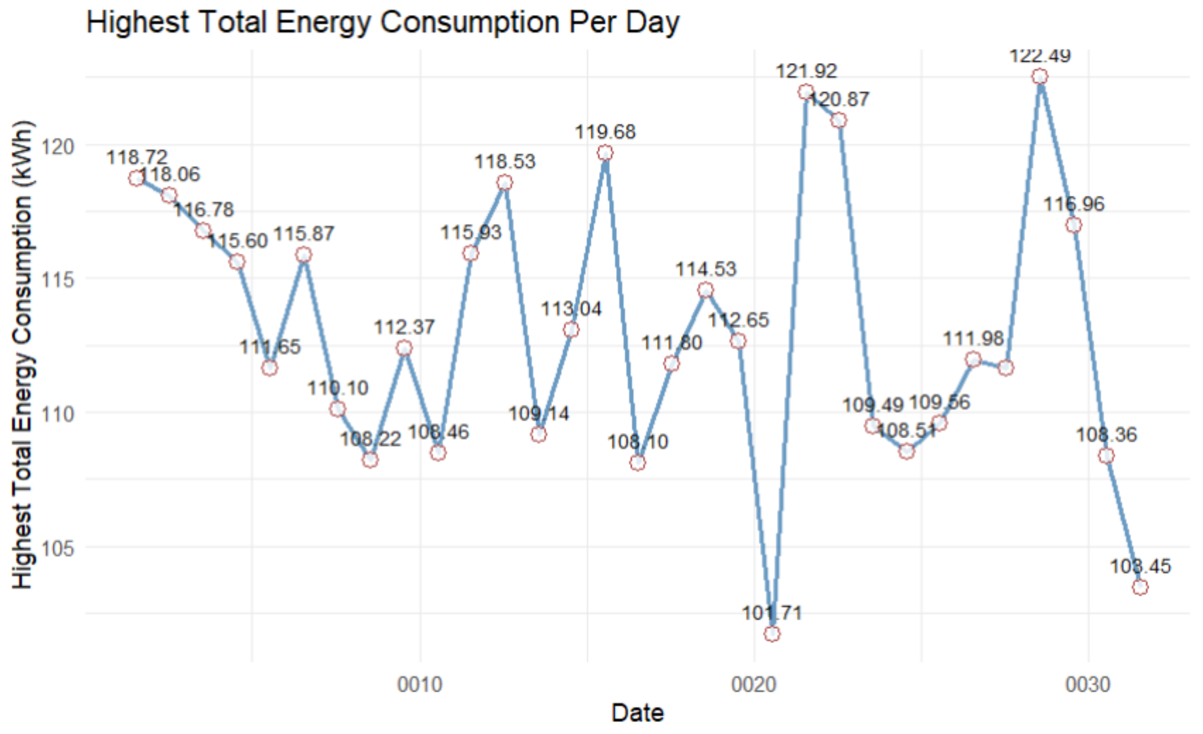
Overall, this bar chart effectively communicates the relationship between the number of bedrooms and average energy consumption, providing valuable insights for energy management and conservation efforts in residential settings.



The bar chart presents the total energy consumption across different heating fuel types used in residential properties. The x-axis lists the various heating fuel types, including Electricity, Fuel Oil, Natural Gas, Other Fuel, and Propane. The y-axis displays the total energy consumption in kilowatt-hours (kWh).

- **Dominant fuel type:** Electricity appears to be the most widely used heating fuel type.
- **Potential energy efficiency**: The variation in energy consumption across fuel types suggests differences in energy efficiency or heating requirements. Certain fuel types, such as Electricity or Natural Gas, may be more energy-intensive or commonly used for larger properties or specific heating needs.
- **Alternative fuel sources:** The chart includes alternative fuel sources like Propane and Other Fuel, indicating the presence of diverse heating options beyond the more conventional choices like Electricity, Fuel Oil, and Natural Gas. Other fuel (Solar energy, wind energy) can be used as an alternative to reduce the consumption of electricity.
- **Potential for targeted strategies:** By identifying the dominant heating fuel types and their associated energy consumption levels, targeted strategies can be developed to address energy efficiency and conservation efforts specific to each fuel type or residential segment.
- **Fuel mix and energy mix:** The visualization provides insights into the fuel mix and energy mix used for heating purposes in the residential sector, which can inform energy planning, resource allocation, and policymaking related to residential energy consumption.

Overall, this bar chart offers a clear comparison of total energy consumption across different heating fuel types, highlighting potential areas for energy efficiency improvements, alternative fuel exploration, and targeted energy conservation strategies based on the prevalent heating fuel types in residential properties.

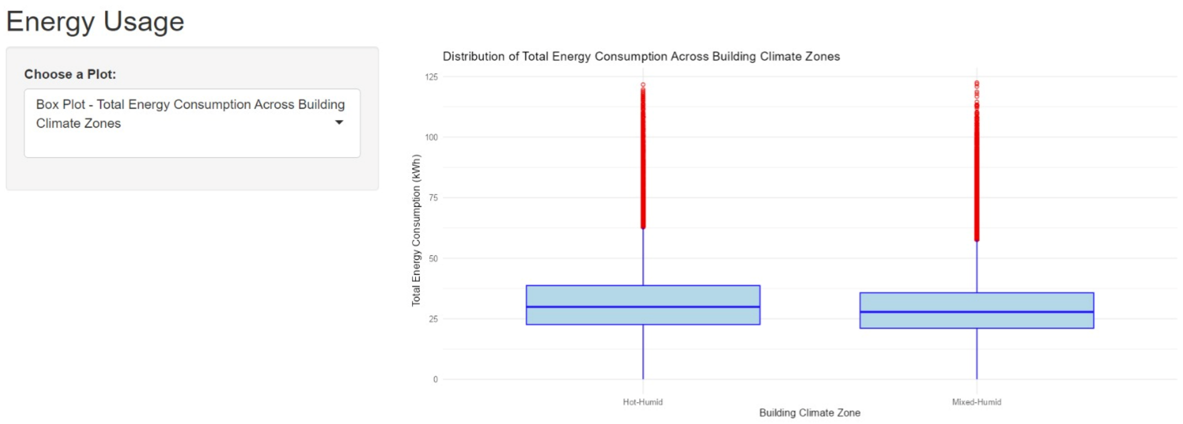


The visualization shows the highest total energy consumption per day during a period that appears to be in July. The x-axis represents the date, and the y-axis represents the total energy consumption in kWh.

- It depicts the significant fluctuations in energy consumption across different days. The peak energy consumption is observed on July 22nd, reaching 122.49 kWh, followed closely by July 21st at 121.92 kWh. These two days(weekends) likely experienced the hottest temperatures or highest demand for cooling, leading to increased energy usage.
- The chart also highlights other days(weekdays) with relatively high energy consumption. These days likely had similarly high temperatures or cooling demands, resulting in elevated energy consumption levels.
- On the other hand, there are days with relatively lower energy consumption, such as July 10th (110.10 kWh), July 11th (110.46 kWh), and July 12th (109.14 kWh). These days might have experienced milder temperatures or lower cooling requirements, leading to reduced energy usage.
- The visualization suggests a strong correlation between temperature or cooling demand and energy consumption, as the highest energy consumption levels are clustered around specific dates, potentially indicating periods of extreme heat or increased cooling needs.
- The difference in energy consumption on weekends and weekdays can also be attributed to most people being at home on weekends and using all the appliances.

This information can provide valuable insights for energy demand management strategies, such as implementing energy-saving measures or promoting conservation efforts during peak demand periods. Additionally, it can aid in forecasting future energy demands and allocating resources more effectively to meet the energy needs of residential properties in the region.

Tool Development(Shiny)

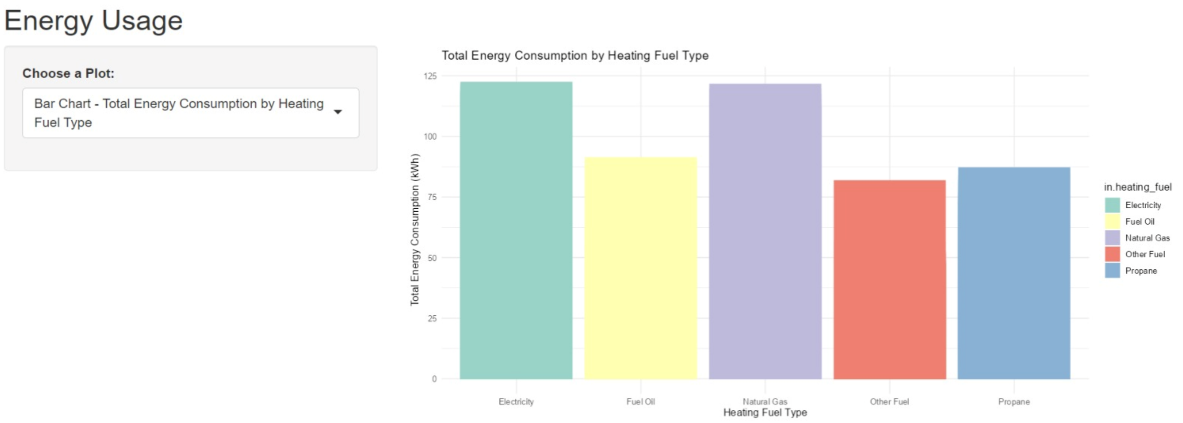


Energy Usage

Choose a Plot:

Box Plot - Total Energy Consumption Across Building Climate Zones

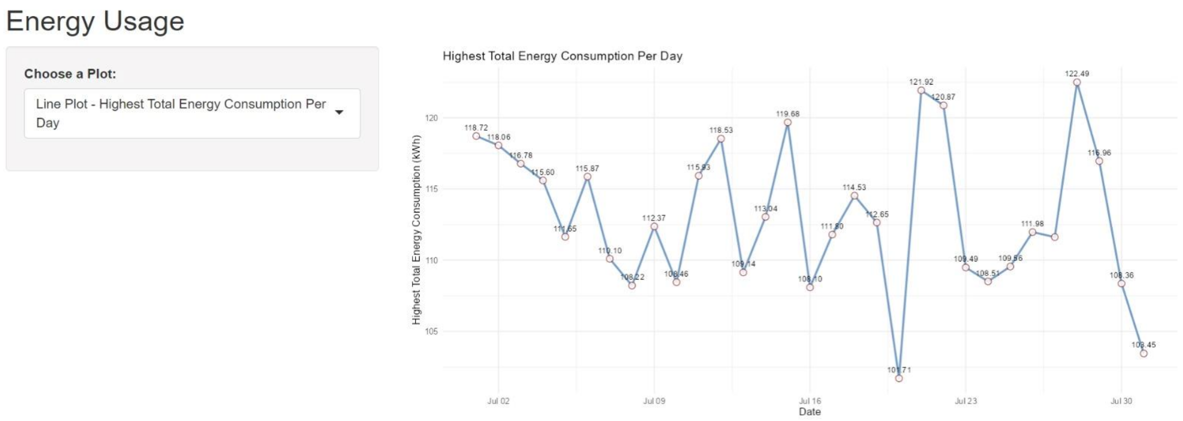
Distribution of Total Energy Consumption Across Building Climate Zones
Total Energy Consumption (kWh)
125
100
75
50
25
0
Hot-Humid
Mixed-Humid
Building Climate Zone



Energy Usage

Choose a Plot:

Bar Chart - Total Energy Consumption by Heating Fuel Type



Energy Usage

Choose a Plot:

Line Plot - Highest Total Energy Consumption Per Day

Shiny App URL: https://smitadeulkar.shinyapps.io/shinyapp/

Demand Reduction Strategy

1. Incentive Programs for Energy-Saving Appliances:
   - Promote the adoption of energy-efficient appliances through partnerships, subsidies, and targeted marketing efforts.

2. Incorporating Solar Power:
   - Evaluate the feasibility of incorporating solar power systems.
   - Assess the building's solar energy output potential.
   - Install solar panels on roofs to generate sustainable energy.

3. Integration of IoT Devices for Enhanced Energy Management:
   - Use IoT technologies to interconnect utility providers and buildings.
   - Improve energy distribution efficiency through real-time data exchange. Examples include smart meters and demand response systems.

4. Peak Hour Pricing:
   - Implement variable pricing to encourage shifting energy use to off-peak times.

5. Community Energy Awareness Campaigns:
   - Organize programs to promote energy-efficient practices.
   - Encourage residents to adopt energy-saving habits.

6. Maintenance on a regular basis:
   - Create a maintenance schedule for all systems and equipment.
   - Ensure that the HVAC, lighting, and other systems are operating at maximum efficiency.
   - Regular inspections and filter replacements are two examples.

7. Evaluate and test:
   - Pilot initiatives on a modest scale should be implemented to test novel energy solutions.
   - Before implementing each method widely, assess its effectiveness.

8. Exemplification:
   - Before broad deployment, test energy-efficient solutions in specific locations.

Challenges

Data Availability:
Ensuring access to accurate and comprehensive datasets, including static house information, energy usage data, and weather data, may pose a challenge. Data collection from various sources, especially through platforms like AWS, could require careful coordination and management.

Data Quality:
For appropriate analysis and modeling, it is imperative to ensure the quality and dependability of the data. The validity of the findings may be impacted by problems with data collection such as missing numbers, outliers, or inconsistent patterns.

Consistency in Column Names and Types:
Ensuring consistency in column names and data types across different datasets is crucial for successful data integration and analysis.

Inconsistencies could lead to errors during data merging and processing.

Data Matching: Verifying that the identifiers used for matching data across different datasets (e.g., building IDs) are consistent and accurate can be challenging. Mismatches or discrepancies may result in inaccurate analysis and modeling outcomes.

Memory Considerations: Processing and merging large datasets, especially with hourly energy usage data for thousands of residential properties, may require substantial computational resources and memory. Efficient memory management techniques may be necessary to handle such large volumes of data effectively.

Model Selection: Selecting the most suitable machine learning models for predicting energy usage accurately can be challenging. It requires careful evaluation of different algorithms and techniques to determine the best-performing model for the specific context of the project.

Scalability: As the project involves analysing data from many residential properties, scalability becomes a concern. Ensuring that the data processing and modeling pipelines can scale efficiently to handle increasing volumes of data is essential.

Interpretability of Models: To obtain insights into the factors driving energy use, it is crucial to make sure predictive models are interpretable and explicable when they are being built. The underlying causes of energy consumption are difficult to comprehend despite complex models' capacity to produce accurate predictions while being difficult to interpret.

Environmental Factors: External factors such as changes in weather patterns, unexpected events like natural disasters, or shifts in consumer behaviour can impact energy usage and demand. Anticipating and adapting to these environmental factors adds complexity to the project's planning and execution.

Contributions

Smita: Collect and merge data, Data Cleaning, Data Visualization
Harika: Data Cleaning, Data Visualization, Presentation, Report
Indraneel: Data Cleaning, Data Visualization, Data Modeling
Sukesh: Shiny App, Presentation
Sejal: Data Modeling, Data Visualization

Conclusion

In conclusion, this project endeavors to tackle the pressing issue of increased energy demand during hot summers, particularly in South Carolina and parts of North Carolina. By leveraging machine learning and data-driven strategies, the goal is to develop solutions that not only manage energy consumption effectively but also promote sustainability and prevent potential blackouts. Through collaborative efforts with eSC, the energy company, the project aims to raise awareness about energy usage, encourage conservation, and reduce overall energy consumption. By addressing these challenges, the project seeks to contribute to a more sustainable future while meeting the energy needs of residential properties in the region.