IST 687: Introduction to Data Science

Group 3

Final Project

Understanding and Reducing Summer Energy Usage for eSC Customers

Submitted by:

Vrushali Lad
Shilpa Pillai
Rijul Ugawekar
Meghana Inavilli
Abhijeet Baviskar

TABLE OF CONTENTS

Table of Contents

PROJECT DESCRIPTION

Executive Summary

This document outlines eSC's proactive approach to addressing the increasing challenges posed by global warming and its impact on summer energy demand. Facing the potential for blackouts due to insufficient grid capacity, eSC embarks on a project to understand key drivers of energy consumption and develop effective solutions to encourage customer conservation.

- **The eSC Challenge:**

  eSC, a prominent energy provider in South Carolina and parts of North Carolina, observes a concerning trend: rising summer temperatures fueled by global warming are leading to significant increases in energy demand. This surge, particularly driven by air conditioning needs, threatens to exceed the company's existing grid capacity.

- **Blackouts as a Potential Threat:**

  eSC anticipates that an "extra hot" summer could lead to unprecedented demand, exceeding the current infrastructure's capabilities. This scenario raises the specter of widespread blackouts, disrupting essential services and creating significant challenges for both eSC and its customers.

- **Alternative to Infrastructure Expansion:**

Rather than investing in costly infrastructure expansion through a new power plant, eSC seeks a more sustainable and cost-effective solution. This project focuses on understanding the factors influencing residential energy usage during peak periods, particularly in July.

- **Delving into Customer Behavior:** eSC aims to gain deeper insights into customer behavior and consumption patterns by analyzing July energy usage data across diverse customer segments. This analysis will consider various factors, including household size, income level, housing type, appliance efficiency, and others.

- **Informing Effective Interventions:** By identifying key drivers of energy consumption, eSC can develop targeted interventions to encourage customer conservation. This knowledge will be crucial for designing effective strategies that reduce overall demand.

- **Embracing Sustainability:** This project transcends the immediate goal of preventing blackouts. It represents a commitment to a more sustainable future. By empowering customers to become active partners in energy conservation, eSC contributes to environmental responsibility and a brighter future for all.

PROJECT SCOPE

Project's scope revolves around thoroughly examining various aspects related to energy usage in households served by the company (eSC) to address the imminent challenges of escalating energy demand due to global warming.

Static House Data Analysis: The project encompasses an in-depth examination of static house data. This includes attributes such as house size, construction materials, age, and other factors that remain constant over time. Analyzing this data aims to discern correlations between house characteristics and energy usage patterns.

Energy Usage Examination: The scope involves detailed scrutiny of energy usage data, collected at an hourly level for individual houses. These datasets provide calibrated information about various energy sources used in households (e.g., air conditioning, dryers). The analysis seeks to unveil trends, peak usage periods, and factors influencing energy consumption at a granular level.

Weather Data Integration: Hourly weather data for different geographic areas, captured using county codes, will be integrated. The analysis will involve assessing how weather fluctuations, particularly during the hottest month of July, impact energy usage patterns in residential properties.

Predictive Model Development: The project ambit extends to the development of robust predictive models focused on forecasting peak energy demand for July. These models will leverage the amalgamation of static house data, energy usage patterns, and weather information to create accurate predictions.

PROJECT DELIVERABLES

Data Collection and Preparation:

- Collect historical data on electricity consumption, weather patterns, and demographic factors.
- Ensure dataset integrity through thorough cleaning, addressing any missing or incorrect data to ensure reliability.

Energy Consumption Analysis:

- Employ linear regression models to discern influential factors impacting energy usage, especially during high-demand summer months.
- Conduct a comprehensive analysis to understand the significance of these factors on overall energy demand.

Predictive Modeling for Demand Forecasting:

- Implement advanced predictive models, like Support Vector Machines (SVM), to forecast electricity demand for the upcoming summer.
- Derive actionable insights to inform the energy company's strategic planning and response measures.

Shiny Apps:

The Shiny application developed aims to offer an intuitive interface for the client to interact with and gain valuable insights from the energy prediction models and future energy needs. It provides two key functionalities:

- **Model's Energy Prediction Insight:** The application allows the client to explore and comprehend the energy prediction models. Users can input various parameters or scenarios to visualize how these factors influence energy predictions. This interactive feature offers a clear understanding of the model's predictive capabilities and the impact of different variables on energy forecasts.

- **Future Energy Needs Exploration:** With a focus on understanding future energy requirements, the Shiny app provides an interface for users to analyze and explore the drivers behind these projected energy needs. Through interactive visualizations and data exploration tools, clients can delve into influential factors such as weather patterns, demographic shifts, or changes in energy consumption behaviors. This functionality empowers the client to anticipate and comprehend potential future energy demands more comprehensively.

DESCRIPTION OF DATA

Static House Data

Description: A file containing fundamental details about single-family houses serviced by eSC.

Contents: Information for each house, including unchanging attributes like size and a unique building ID used to access corresponding energy usage data.

Format: 'parquet' format optimized for efficient storage.

Location: [Static House Data Link] (https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/static_house_info.parquet)

Size: Around 5,000 houses represented in this dataset.

Energy Usage Data

Description: Individual datasets capturing hourly energy usage for each house.

Contents: Calibrated and validated energy consumption details with 1-hour load profiles, encompassing various sources (e.g., air conditioning systems, dryers) per house.

Format: Each file is in 'parquet' format for efficient storage.

Location: Stored in an AWS folder, with each file named after the house's 'building ID'.

Size: Approximately 5,000 individual datasets, each corresponding to a house.

Meta Data

Description: A comprehensive data description file explaining attribute fields used across different housing data files.

Contents: Human-readable descriptions of attributes found in both static house data and energy usage data.

Format: CSV file accessible via the provided link.

Weather Data

Description: Hourly weather information collected and categorized by geographic areas (counties).

Contents: Timeseries weather data stored based on county codes, providing detailed weather insights.

Format: Individual files for each geographic area in an accessible format.

DATA PREPARATION

Objective: The objective of this data processing and analysis pipeline is to extract, process, and analyze energy usage and weather data for a specific region during the month of July 2023. The analysis aims to understand patterns, trends, and relationships between energy consumption and weather conditions.

Libraries: The pipeline utilizes several R libraries for efficient data manipulation, including dplyr, readr, arrow, lubridate, and future.apply. These libraries provide essential functions for data processing, handling various data formats, managing dates, and enabling parallel processing.

Data Sources:

a. Static House Information:

- The static house information is sourced from a parquet file hosted on Amazon S3 (**static_house_info.parquet**).
- This dataset contains essential details about houses, including unique building IDs.

b. Energy Usage Data:

- The energy usage data is retrieved in chunks, with each chunk representing a subset of unique building IDs.
- Parallel processing is employed to expedite data extraction and processing.

- The processed data is then combined into a single dataframe (**all_data_combined**), which includes energy usage information for the specified month.

c. Weather Data:

- Weather data is obtained for each county in the dataset for the month of July 2023.
- The data is filtered to include only relevant information for analysis.
- Individual county weather dataframes are combined into a single dataframe (**weather_data**), facilitating a comprehensive analysis.

DATA CLEANING

1. Cleaning Static House Information

Step 1: Handling Missing Values and Empty Strings

- Identified columns with missing values or empty strings. Dropped columns with missing values or empty strings.

Step 2: Handling "None" and "Not Applicable" Values

- Identified and dropped columns with values like "None" or "Not Applicable".
- Removed columns deemed irrelevant for the analysis.
- Modified selected columns by removing the "F" suffix.
- The cleaned static house data was exported to a CSV file.

2. Cleaning Weather Data

Step 1: Outlier Handling

- We loaded the weather data from "WeatherData.csv" and conducted essential preprocessing, focusing on the 'time' column. A comprehensive structural overview and statistical analysis of the dataset were performed to understand its characteristics.
- Using the IQR method, outliers were identified and subsequently rectified by replacing them with 'NA' values. Lastly, the refined weather dataset, now free from outliers, was exported to a CSV file, ensuring data reliability for subsequent analyses.

3. Processing Energy Usage Data

- The energy usage data was loaded from the file "EnergyUsageData.csv."
- A comprehensive check of the dataset's structure and summary statistics was conducted for an initial understanding.
- The 'time' column was converted to POSIXct format to standardize the time representation.
- Data was filtered specifically for the month of July to focus the analysis.
- A new column named 'total_energy_usage' was added to calculate the total energy usage.
- Rows with negative total energy usage were removed to maintain data integrity.
- The dataset was aggregated by building and date, computing the total energy usage for each.
- Finally, the processed energy usage data for July, refined and aggregated, was exported to a CSV file. These steps ensured the dataset was suitably prepared for subsequent analyses or modeling tasks.

DATA MERGING

The merging process focused on integrating the 'energy_static_data' and the aggregated weather data ('weather_data_agg') using specific columns: 'in.county' and 'time'. This merging logic aimed to consolidate both datasets based on county-level information and temporal details extracted from the 'time' column.

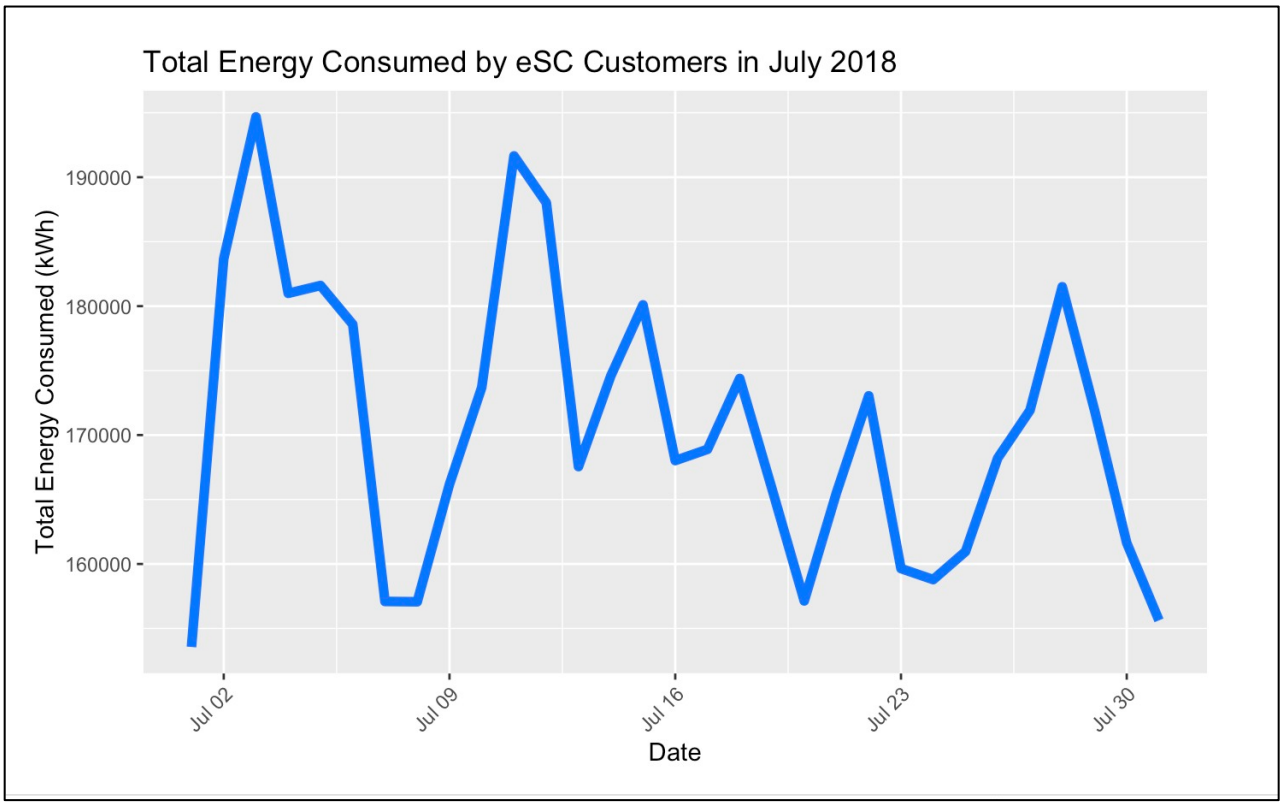
The merging occurred by aligning entries from both datasets using the 'in.county' column as a geographical identifier and the 'time' column as a temporal indicator. This integration facilitated the creation of the 'energy_static_weather_data' dataframe, combining energy-related information from 'energy_static_data' with aggregated weather details.

Ultimately, this merging strategy aimed to create a comprehensive dataset for subsequent analysis, exploring the interconnectedness between energy-related factors and weather attributes while ensuring data integrity by eliminating rows with "None" values.

EXPLORATORY DATA ANALYSIS

For EDA we made several exploratory plots employing various visualization techniques. Some of them are explained below.

1) Total Energy Consumed by eSc Customers in July 2018:



The line graph delineates eSC customers' total energy consumption trends throughout July 2018, exhibiting fluctuations over the month. Peaks on specific days indicate heightened energy usage during those periods. This visual representation offers crucial insights into consumption patterns, aiding eSC in managing and forecasting energy demand effectively.

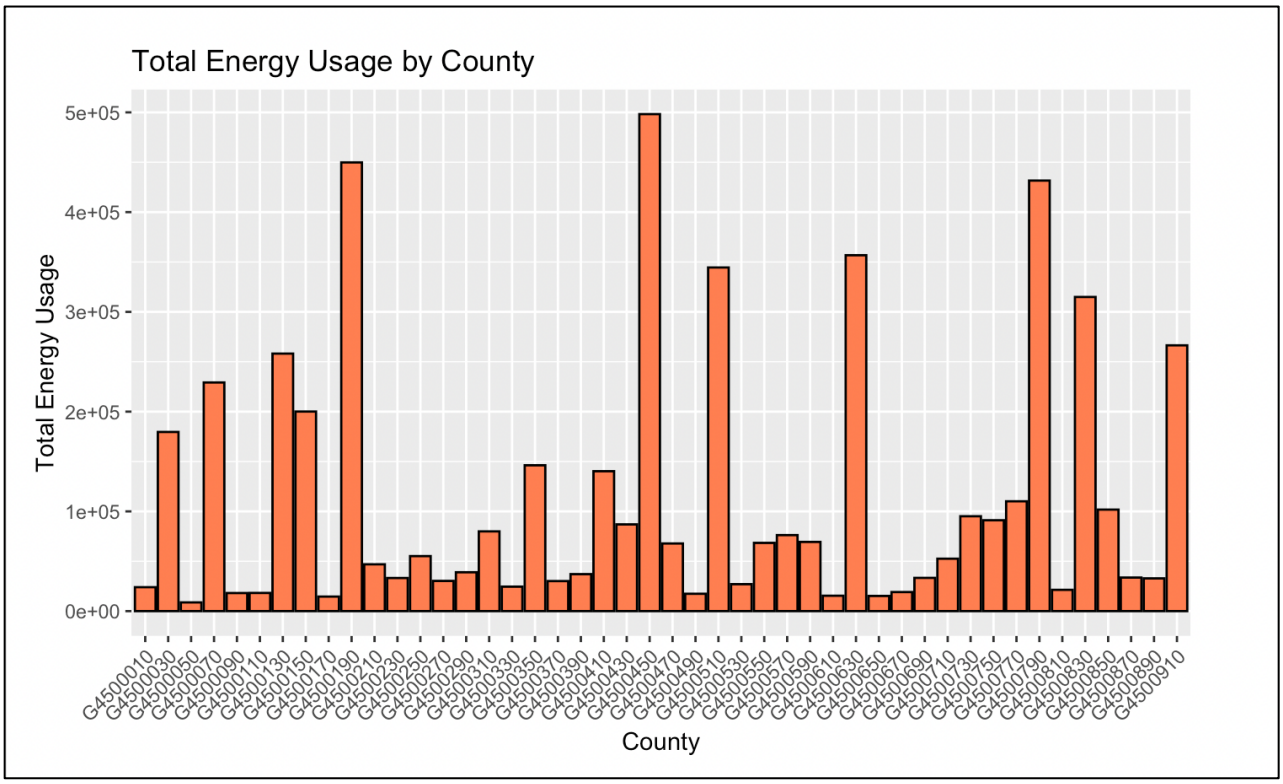
By comprehending these consumption patterns, eSC can proactively strategize and allocate resources to manage energy supply during peak periods. Understanding the days and times when energy consumption surges enables better preparation for potential strain on the grid. It allows for

anticipatory measures, such as load balancing or deploying additional resources, to meet heightened demand without disruptions.

Moreover, analyzing historical consumption data aids in forecasting future energy demands. Recognizing recurring patterns or anomalies on specific days facilitates more accurate predictions, enabling eSC to optimize energy distribution and resource allocation.

Ultimately, leveraging insights from this graph empowers eSC to make informed decisions regarding energy management strategies. By proactively addressing consumption peaks and leveraging forecasting capabilities, eSC can enhance operational efficiency, ensure grid stability, and provide reliable energy services to customers.

2) Total Energy Usage by County:



The bar graph provides a comparative view of total energy usage among different counties, highlighting substantial variations in consumption levels. The disparities observed across counties suggest potential factors such as differences in population density, industrial activities, or the effectiveness of existing efficiency measures.

Counties exhibiting notably higher energy usage levels compared to others warrant closer scrutiny and consideration for targeted interventions. This data serves as a valuable resource to identify specific areas ripe for energy efficiency improvements or further investigation into usage patterns.

For counties with elevated energy consumption, initiatives focusing on energy efficiency enhancements become pivotal. Implementing targeted efficiency programs tailored to the unique needs of these areas can lead to significant reductions in energy usage. These initiatives may involve promoting energy-efficient technologies, incentivizing conservation practices, or conducting energy audits to identify and rectify inefficiencies.

Moreover, this data allows for a more detailed examination of usage patterns within each county. It facilitates a deeper understanding of the factors contributing to higher consumption rates, enabling more informed decision-making regarding resource allocation and strategic planning.

By leveraging this data to pinpoint areas with higher energy consumption, authorities can effectively allocate resources and implement tailored strategies aimed at optimizing energy usage. This approach aligns with broader sustainability goals, promoting efficient energy utilization and reducing the environmental impact associated with excessive energy consumption.

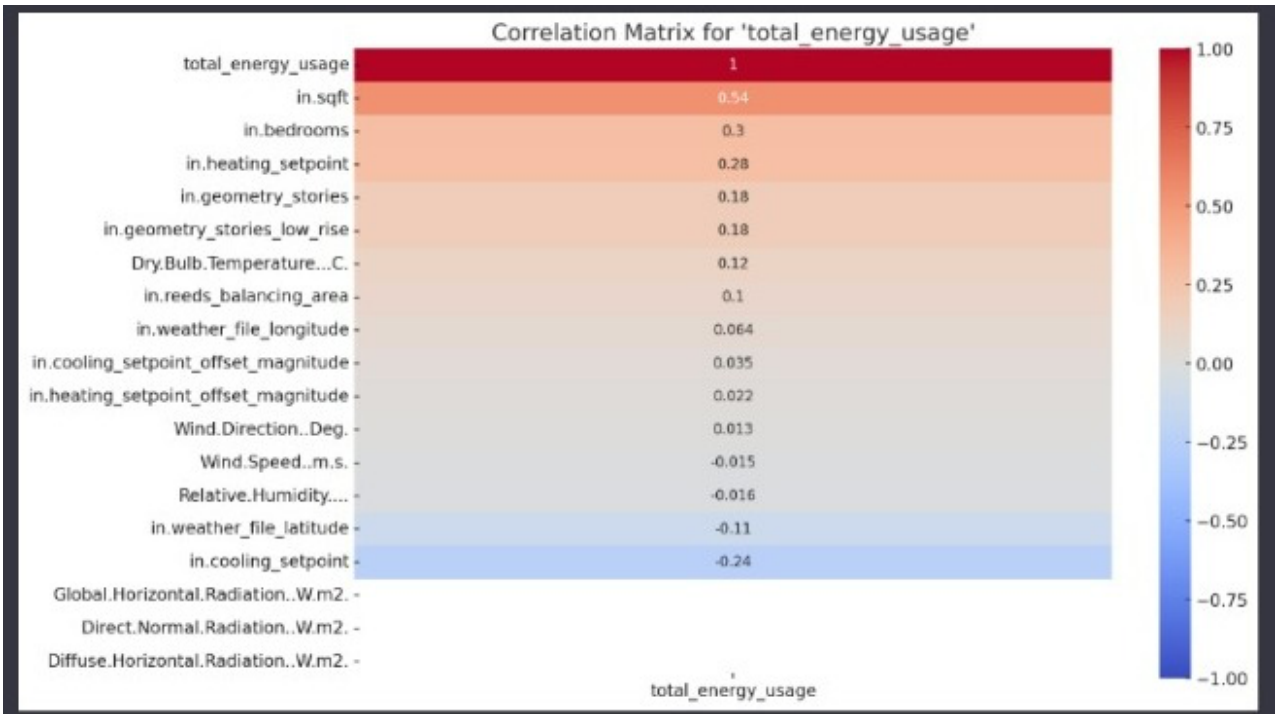
FEATURE ENGINEERING AND CORRELATION ANALYSIS

In our exploratory data analysis, we embarked on a comprehensive investigation of the **energy_static_weather_data** dataset, employing feature engineering techniques and correlation analysis to unveil insights into the relationships between various variables. Initially, we identified the numeric columns within the dataset using the **sapply** function, generating a logical vector denoting the nature of each column. This step was pivotal as subsequent analyses would focus exclusively on numeric variables.

Following the identification of numeric columns, we calculated a correlation matrix using the **cor** function, which allowed us to assess the pairwise correlations between variables. To enhance our understanding of these relationships, a visual representation of the correlation matrix was created using the **corrplot** library. This color-coded plot provided an intuitive depiction of the strength and direction of correlations, facilitating the identification of potential patterns within the dataset.

To narrow our focus, we honed in on the variable "total_energy_usage" and filtered the correlation matrix to identify variables highly correlated with it. Using a threshold of 0.5, we aimed to capture strong correlations, considering both positive and negative associations through the use of the **abs** function. The resulting set of highly correlated variables was then displayed, offering a concise list of potential influencers or indicators associated with energy usage.

In conclusion, this feature engineering and correlation analysis lay a robust foundation for further exploration and decision-making. The visualizations and identified variables provide valuable insights into the dataset, particularly in relation to energy usage. This analysis serves as a pivotal step towards a more in-depth understanding of the underlying patterns and dynamics within the data, offering actionable intelligence for subsequent stages of analysis and modeling.



MODELING

MODEL 1 – LINEAR REGRESSION

The aim is to create a linear regression model that forecasts 'total energy usage' by analyzing factors such as cooling setpoint, heating setpoint, number of bedrooms, square footage, time, and dry bulb temperature. This model intends to uncover how these factors impact the overall energy consumption.

Process Overview:

Data Preparation:

The dataset (data_subset) is assumed to be cleaned and curated, containing pertinent variables for analysis.

To ensure reproducibility, the set.seed(123) function is employed, ensuring consistent outcomes in random processes (like data splitting) across multiple runs of the code.

Data Splitting:

The dataset is divided into training and test sets to assess the model's performance.

train_indices are generated using the sample function, randomly picking 80% of data indices for training. This random selection ensures a representative subset.

Subsequently, train_data encompasses 80% of the dataset, while test_data holds the remaining 20%. This segregation enables training the model on a significant portion while retaining a separate set for evaluating its performance.

Model Construction:

A linear regression model (model) is built using the lm function in R.

The model predicts total_energy_usage based on various independent variables: in.cooling_setpoint, in.heating_setpoint, in.bedrooms, in.sqft, time, and Dry.Bulb.Temperature...C..

These variables are chosen for their expected influence on total energy consumption. The summary(model) function offers a comprehensive review of the model's performance, encompassing predictor significance, coefficients, R-squared value, and other crucial statistical metrics.

MODEL 2 – DECISION TREE MODEL

In the evaluation of 'total energy usage,' a comprehensive process was employed involving:

Data Preparation:

The initial step involved meticulous preparation of the dataset ('data_subset'), ensuring it contained essential variables necessary for analysis. Rigorous cleaning and curation guaranteed the inclusion of pertinent factors such as cooling setpoint, heating setpoint, bedrooms, square footage, time, and dry bulb temperature. Additionally, measures were taken to ensure data consistency and reliability, laying the groundwork for subsequent analysis.

Data Splitting:

The dataset was strategically divided into distinct subsets for training and evaluation purposes. Utilizing the sample function, 80% of the data was randomly allocated to the training set ('train_data'), while the remaining 20% constituted the test set ('test_data'). This segregation allowed the model to be trained on a significant portion of the data while retaining an independent set for assessing its performance, mitigating the risk of model overfitting and ensuring a robust evaluation.

Model Construction:

A linear regression model was developed using the lm function in R. This model aimed to predict 'total energy usage' based on a set of independent variables—cooling setpoint, heating setpoint, bedrooms, square footage, time, and dry bulb temperature—selected for their presumed influence on energy consumption patterns. Leveraging these variables, the model was designed to unveil insights into how each factor impacts overall energy usage, facilitating a deeper understanding of energy consumption dynamics.

In essence, meticulous data preparation, strategic data splitting for training and testing, and the construction of a robust linear regression model constituted the framework employed to comprehensively evaluate 'total energy usage' and elucidate the influencing factors within the dataset.

CHOICE OF MODEL

The choice between the linear regression and Decision Tree regression models was based on the nature of the data and the characteristics of the problem being addressed, aiming to optimize the prediction of 'total energy usage.'

The Linear Regression Model was selected due to its:

1. Interpretability: Its simplicity allows for a clear understanding of how each predictor influences 'total energy usage.' Coefficients provide direct insights into the magnitude and direction of these influences.

2. Assumption of Linearity: If there's an assumption that the relationships between predictors (cooling setpoint, heating setpoint, bedrooms, square footage, time, and dry bulb temperature) and 'total energy usage' are predominantly linear and additive, this model is a suitable choice.

On the other hand, the Decision Tree Regression Model was chosen for its:

1. Capability to Capture Nonlinearities: It excels in capturing nonlinear and complex relationships among variables. In cases where interactions or nonlinear relationships between predictors and 'total energy usage' are expected, this model can provide more accurate predictions.

2. Handling Complex Interactions: Decision Trees are adept at uncovering intricate patterns and interactions within the data, making them suitable for scenarios where relationships are not strictly linear.

The rationale behind selecting these models rested on their respective strengths in handling linear or nonlinear relationships between predictors and the target variable, 'total energy usage.' The decision was made to optimize predictive accuracy and facilitate a deeper understanding of how various factors impact overall energy consumption within the dataset.

MODEL 1 - LINEAR MODELING

The statistical analysis showcases a linear regression model predicting total energy usage based on several factors: cooling setpoint, heating setpoint, number of bedrooms, and dry bulb temperature across different time points or periods.

The model's accuracy, as indicated by the Multiple R-squared of 0.4943, suggests that approximately 49.43% of the variability in energy usage can be explained by the included predictors. This level of explanatory power signifies a moderate ability of the model to account for fluctuations in energy consumption based on the given factors.

Moreover, the significant F-statistic underscores that the model itself holds statistical significance. It indicates that this model is statistically superior to a model devoid of any predictors, implying that these variables collectively contribute to explaining changes in energy usage.

However, while statistically significant, the practical significance of the model's predictive power depends on the context and specific application. The model's 49.43% explanatory power might be considered moderate, highlighting that additional factors beyond those included in the model could also influence energy usage.

Therefore, while the model holds statistical significance and offers insights into the relationship between predictors and energy usage, its practical utility would benefit from a contextual understanding and considering other potential variables that might impact energy consumption for more robust predictions.

```
Call:
lm(formula = total_energy_usage ~ in.cooling_setpoint + in.heating_setpoint +
    in.bedrooms + in.sqft + time + Dry.Bulb.Temperature...C.,
    data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-44.489  -6.253  -1.212   5.107  85.234

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -1.385e+01  7.948e-01  -17.43   <2e-16 ***
in.cooling_setpoint       -8.108e-01  6.290e-03 -128.91   <2e-16 ***
in.heating_setpoint        6.007e-01  3.853e-03  155.92   <2e-16 ***
in.bedrooms                1.090e+00  3.285e-02   33.19   <2e-16 ***
in.sqft                    4.962e-03  2.038e-05  243.45   <2e-16 ***
time2018-07-02T04:00:00Z   4.585e+00  1.938e-01   23.65   <2e-16 ***
time2018-07-03T04:00:00Z   4.952e+00  1.951e-01   25.39   <2e-16 ***
time2018-07-04T04:00:00Z   4.548e+00  1.934e-01   23.52   <2e-16 ***
time2018-07-05T04:00:00Z   4.504e+00  1.937e-01   23.25   <2e-16 ***
time2018-07-06T04:00:00Z   4.831e+00  1.941e-01   24.89   <2e-16 ***
time2018-07-07T04:00:00Z   4.936e+00  2.021e-01   24.43   <2e-16 ***
time2018-07-08T04:00:00Z   4.546e+00  2.002e-01   22.71   <2e-16 ***
time2018-07-09T04:00:00Z   4.888e+00  1.963e-01   24.90   <2e-16 ***
time2018-07-10T04:00:00Z   4.380e+00  1.941e-01   22.57   <2e-16 ***
time2018-07-11T04:00:00Z   4.273e+00  1.957e-01   21.84   <2e-16 ***
time2018-07-12T04:00:00Z   4.817e+00  1.945e-01   24.76   <2e-16 ***
time2018-07-13T04:00:00Z   3.718e+00  1.942e-01   19.14   <2e-16 ***
time2018-07-14T04:00:00Z   4.377e+00  1.933e-01   22.64   <2e-16 ***
time2018-07-15T04:00:00Z   4.589e+00  1.932e-01   23.76   <2e-16 ***
time2018-07-16T04:00:00Z   3.907e+00  1.942e-01   20.12   <2e-16 ***
time2018-07-17T04:00:00Z   3.930e+00  1.939e-01   20.27   <2e-16 ***
time2018-07-18T04:00:00Z   3.770e+00  1.935e-01   19.48   <2e-16 ***
time2018-07-19T04:00:00Z   3.999e+00  1.950e-01   20.50   <2e-16 ***
time2018-07-20T04:00:00Z   3.560e+00  1.976e-01   18.02   <2e-16 ***
time2018-07-21T04:00:00Z   4.504e+00  1.964e-01   22.93   <2e-16 ***
time2018-07-22T04:00:00Z   5.618e+00  1.952e-01   28.78   <2e-16 ***
time2018-07-23T04:00:00Z   5.099e+00  2.004e-01   25.44   <2e-16 ***
time2018-07-24T04:00:00Z   3.925e+00  1.971e-01   19.91   <2e-16 ***
time2018-07-25T04:00:00Z   4.544e+00  1.981e-01   22.94   <2e-16 ***
time2018-07-26T04:00:00Z   4.138e+00  1.947e-01   21.25   <2e-16 ***
time2018-07-27T04:00:00Z   3.782e+00  1.939e-01   19.51   <2e-16 ***
time2018-07-28T04:00:00Z   4.885e+00  1.932e-01   25.28   <2e-16 ***
time2018-07-29T04:00:00Z   5.179e+00  1.950e-01   26.56   <2e-16 ***
time2018-07-30T04:00:00Z   3.939e+00  1.960e-01   20.10   <2e-16 ***
time2018-07-31T04:00:00Z   3.694e+00  1.988e-01   18.58   <2e-16 ***
Dry.Bulb.Temperature...C.  1.636e+00  2.055e-02   79.63   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.243 on 141572 degrees of freedom
Multiple R-squared:  0.4943,    Adjusted R-squared:  0.4942
F-statistic:  3954 on 35 and 141572 DF,  p-value: < 2.2e-16

[1] 85.46257
```

MODEL 2 - DECISION TREE MODEL

The analysis employed a decision tree model to predict total energy usage based on variables such as square footage, heating and cooling setpoints, number of bedrooms, time, and dry bulb temperature. Decision trees segment data into branches, establishing rules at each node to make predictions and improve predictive accuracy while controlling overfitting.

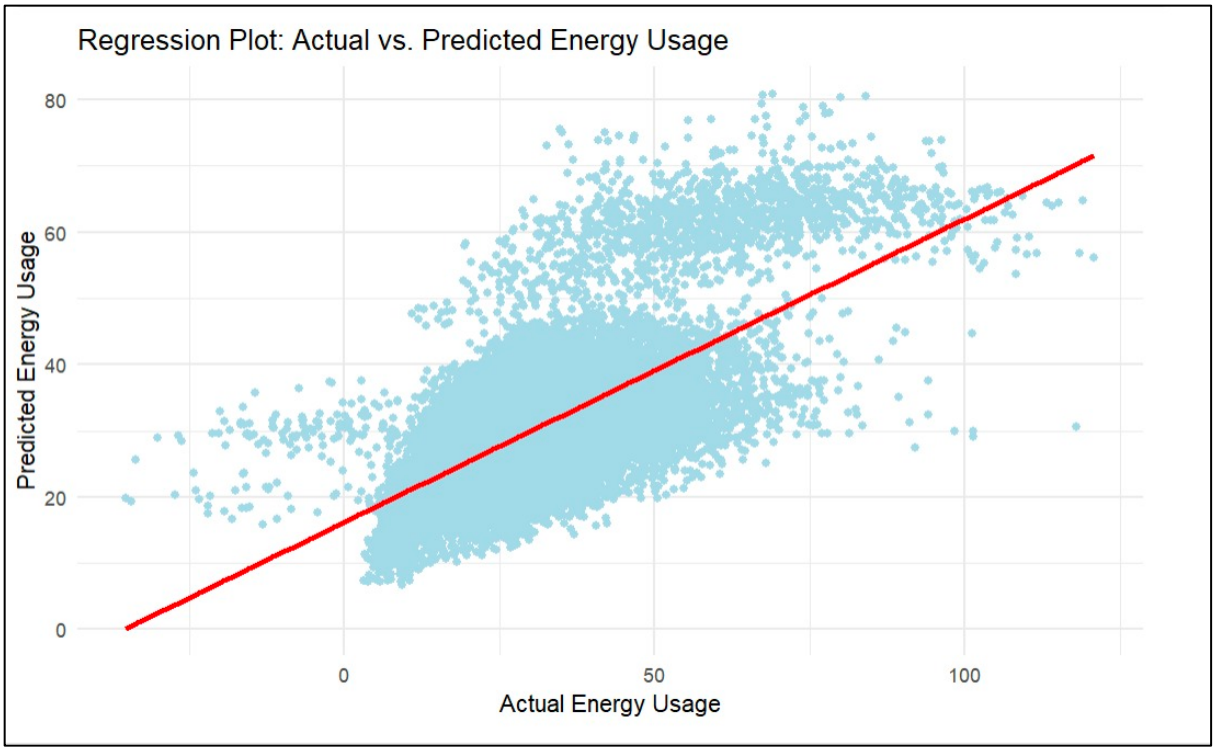
The model evaluation relied on mean squared error (MSE), indicating the average squared difference between predicted and actual values. The provided MSE of 130.28 reflects the model's predictive error, signifying the average discrepancy between predicted and observed energy usage.

Additionally, the reported R-squared value of 0.48 suggests that around 48% of the variability in energy usage can be explained by the included predictors. This demonstrates a moderate level of explanatory power in accounting for energy usage variance based on the selected variables.

However, a comprehensive assessment considers the context and practical implications of the findings. While MSE and R-squared provide insights into model performance, a deeper understanding of the dataset's characteristics and the specific application's requirements is essential for interpreting these metrics accurately.

```
      Dry.Bulb.Temperature...C. < 24.26555   to the left,  agree=0.519, adj=0.002, (0 split)
      in.bedrooms               < 4.5        to the right, agree=0.519, adj=0.002, (0 split)

Node number 20: 33486 observations
  mean=24.92464, MSE=57.13434

Node number 21: 37083 observations
  mean=29.8259, MSE=77.80777

Node number 22: 23739 observations
  mean=32.44219, MSE=107.4859

Node number 23: 25535 observations,    complexity param=0.01093825
  mean=38.28483, MSE=122.8807
  left son=46 (13374 obs) right son=47 (12161 obs)
  Primary splits:
      Dry.Bulb.Temperature...C. < 26.74708   to the left,  improve=0.08337502, (0 missing)
      in.sqft                   < 2982       to the left,  improve=0.07838883, (0 missing)
      in.cooling_setpoint       < 69         to the right, improve=0.02558068, (0 missing)
      in.bedrooms               < 4.5        to the left,  improve=0.02146086, (0 missing)
      in.heating_setpoint       < 73.5       to the right, improve=0.01990554, (0 missing)
  Surrogate splits:
      time                < 1531670000 to the right, agree=0.624, adj=0.211, (0 split)
      in.sqft             < 2982       to the left,  agree=0.537, adj=0.028, (0 split)
      in.cooling_setpoint < 66         to the right, agree=0.527, adj=0.007, (0 split)
      in.bedrooms         < 4.5        to the left,  agree=0.526, adj=0.004, (0 split)

Node number 46: 13374 observations
  mean=35.23263, MSE=96.58312

Node number 47: 12161 observations
  mean=41.64148, MSE=130.2891
```

FUTURE ENERGY DEMAND



The regression plot shows the actual energy usage against the predicted values generated by a model. In this visualization, the red line serves as a reference, depicting the scenario where the predicted values perfectly align with the actual observed values.

The dispersion of individual data points around this red line signifies the prediction errors made by the model. Points scattered farther from the red line indicate larger disparities between the predicted and actual values, highlighting instances where the model's estimations deviate from the observed reality.

Furthermore, the density or concentration of points close to the red line signifies the accuracy of the model. A denser cluster of points along the line indicates that the model's predictions closely match

the actual values, demonstrating higher accuracy and reliability in its estimations. Conversely, a sparser distribution implies more variability in prediction accuracy.

This visualization essentially provides a visual assessment of how well the model performs in predicting energy usage. A tighter alignment of points around the red line indicates a more accurate and precise model, while a scattered or dispersed pattern suggests potential areas for improvement in the model's predictive capabilities.

SHINY APPS

Our Shiny application stands as a cornerstone tool for clients vested in understanding and predicting their energy consumption patterns. Its seamless integration of diverse inputs—ranging from fundamental factors like square footage and the number of bedrooms to more dynamic variables such as prevailing weather conditions—allows for a comprehensive estimation of monthly energy usage. At its core, this application serves as a proactive resource, enabling clients to anticipate and manage their energy demands effectively.

One of the application's key features is the intuitive pie chart visualization, a pivotal component aiding clients in comprehending the relative impact of each input parameter on their overall energy consumption. This visualization acts as a compass, delineating the primary drivers behind their energy usage. By providing a clear breakdown of how different factors contribute to the total energy consumption, clients can discern and prioritize areas for potential optimization. This, in turn, empowers them with actionable insights, enabling informed decisions to curtail excessive energy consumption.

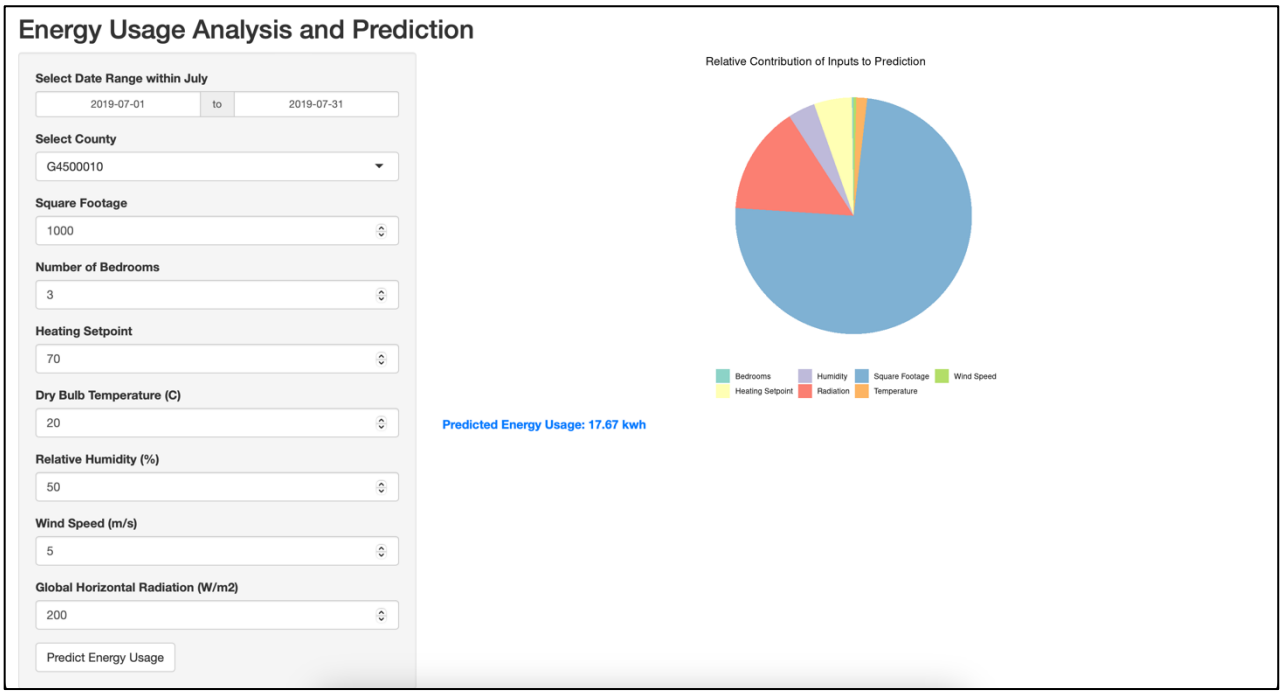
The predictive capability embedded within this application holds particular significance during July's peak demand period. As this month often marks heightened energy usage due to varying factors like increased cooling needs amidst warmer temperatures or amplified usage in residential areas during vacations, the ability to foresee and manage these peaks is crucial. Our tool equips clients with the foresight to anticipate and strategize for these surges, offering them a proactive approach to mitigate spikes in energy consumption. By enabling clients to prepare and adjust their

usage patterns during these critical periods, our application aids in averting potential strain on the energy grid, ultimately contributing to the collective effort to mitigate blackouts.

Moreover, the application's underlying purpose aligns seamlessly with eSC's objective to address peak energy demands without resorting to infrastructure expansion. By empowering clients with comprehensive insights into their energy usage and highlighting areas for optimization, our tool facilitates a proactive approach to manage and potentially reduce energy consumption during peak periods. This proactive stance is pivotal in supporting eSC's broader goal of maintaining a stable and reliable energy grid without the necessity for extensive infrastructure development.

Furthermore, the application's role in offering predictive insights not only aids clients in managing their energy consumption efficiently but also resonates with broader environmental conservation efforts. By enabling clients to optimize their energy usage based on predictive analysis rather than reactive measures, our tool fosters a culture of sustainability and conscientious energy consumption practices. This, in turn, contributes positively to environmental preservation endeavors by reducing unnecessary energy waste and promoting more efficient usage patterns.

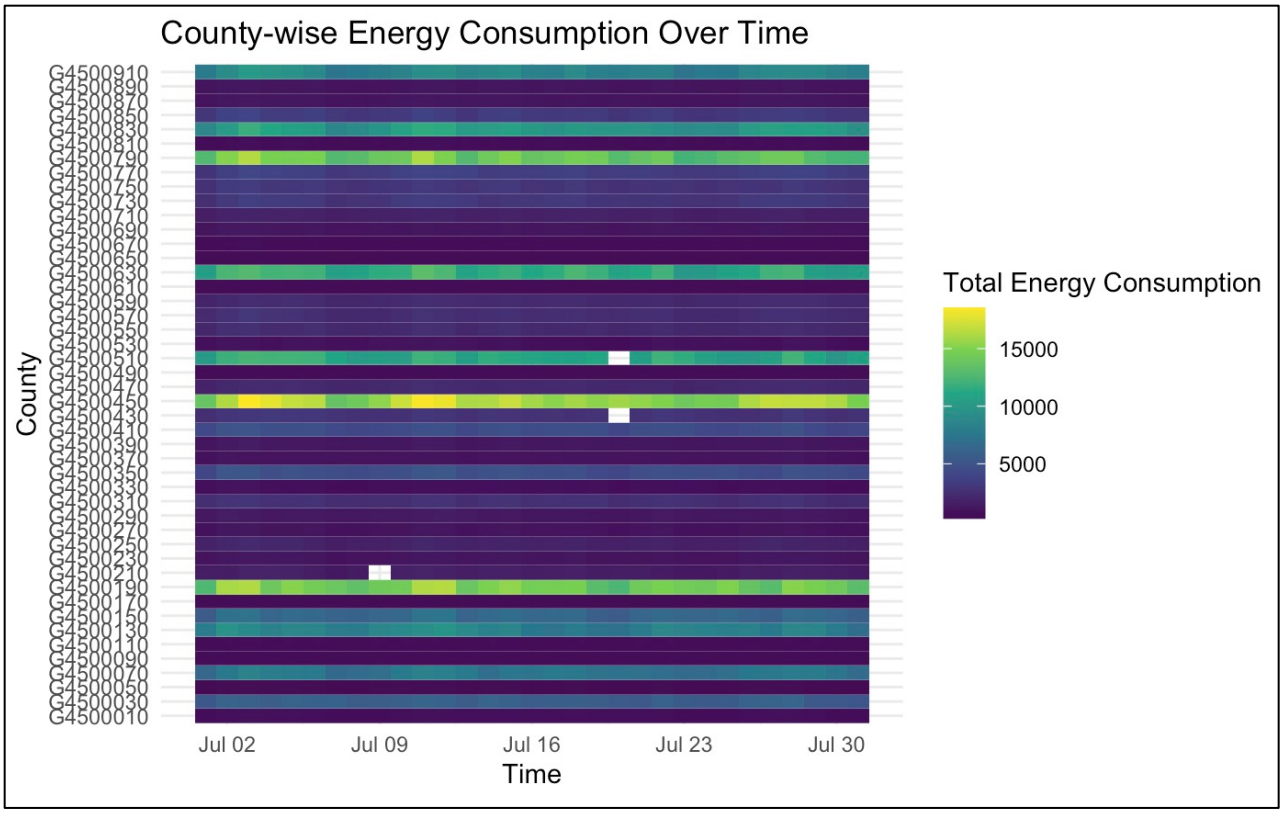
In essence, our Shiny application stands as an indispensable tool, fostering a proactive approach to energy management, empowering clients with predictive insights, and contributing to eSC's objective of sustainable energy grid management without substantial infrastructure expansion.



APP URL - https://rijulugawekar17.shinyapps.io/IDSFINAL/

APPROACH TO REDUCE PEAK ENERGY DEMAND

a. Mitigate peak energy demand



The visualization of county-wise energy consumption serves as a valuable resource for spotting trends and anomalies in usage across regions. It enables proactive strategies to mitigate peak energy demand:

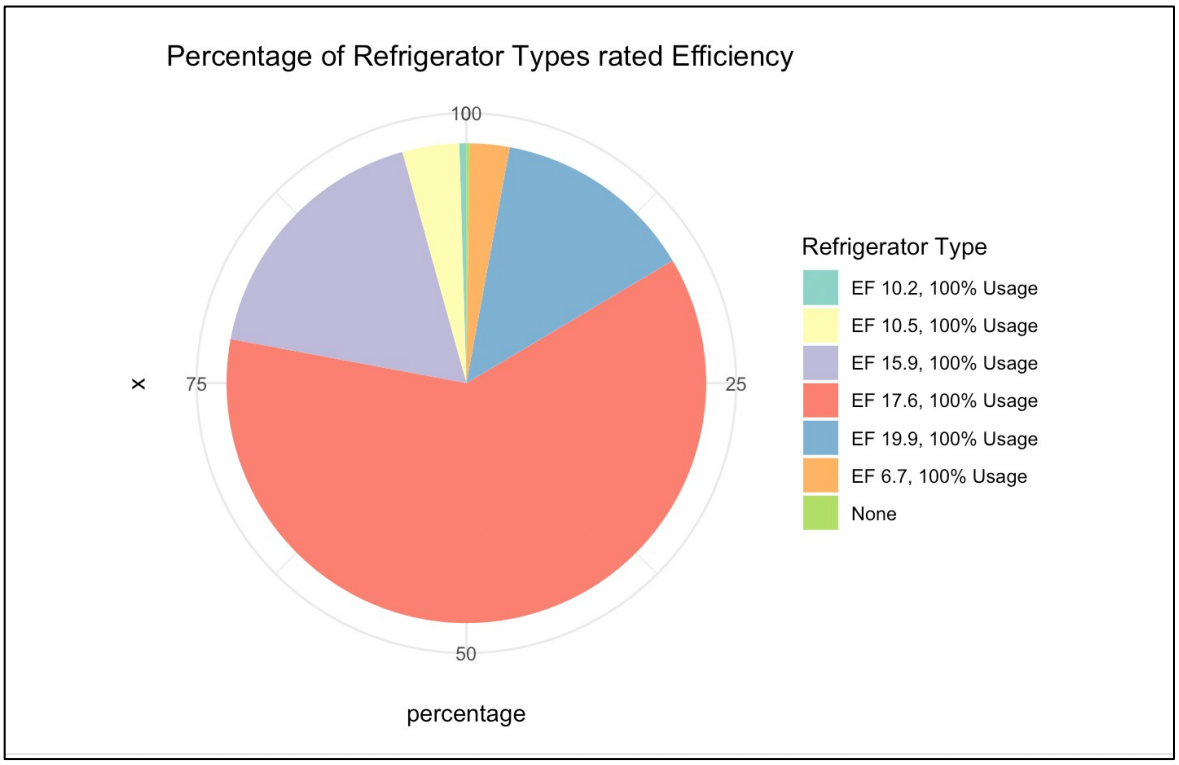
By identifying bright streaks indicating spikes in consumption, proactive measures can manage heightened demand. Targeted response programs and load-shifting strategies address high-usage

times and regions. Brighter shades prompt region-specific efficiency campaigns, educating on energy-saving practices. Continuous analysis refines predictive models for precise demand forecasts, aiding preparedness.

Strategic infrastructure upgrades in high-spike areas enhance grid capacity during peaks. Insights drive behavioral adjustments through awareness campaigns, encouraging energy-efficient habits. The data supports smart grid implementation, enabling real-time responses for optimized distribution.

Leveraging these insights fuels targeted strategies, including demand responses, efficiency drives, predictive models, infrastructure enhancements, behavioral shifts, and smart grid advancements. Collectively, these measures align with reducing and managing peak energy demand effectively.

b. Upgrade Refrigerator to EF 19.9+ models



The pie chart provides a snapshot of the distribution of refrigerator types categorized by their Energy Factor (EF) ratings within a sample population. The EF rating serves as a crucial metric in evaluating the energy efficiency of appliances, where higher EF values signify more efficient energy utilization.

The dominant segment within the chart represents refrigerators with an EF rating of 17.6, indicating their prevalent usage within the sample population. Contrastingly, a smaller portion of the sample is represented by refrigerators with higher EF ratings, notably EF 19.9 models.

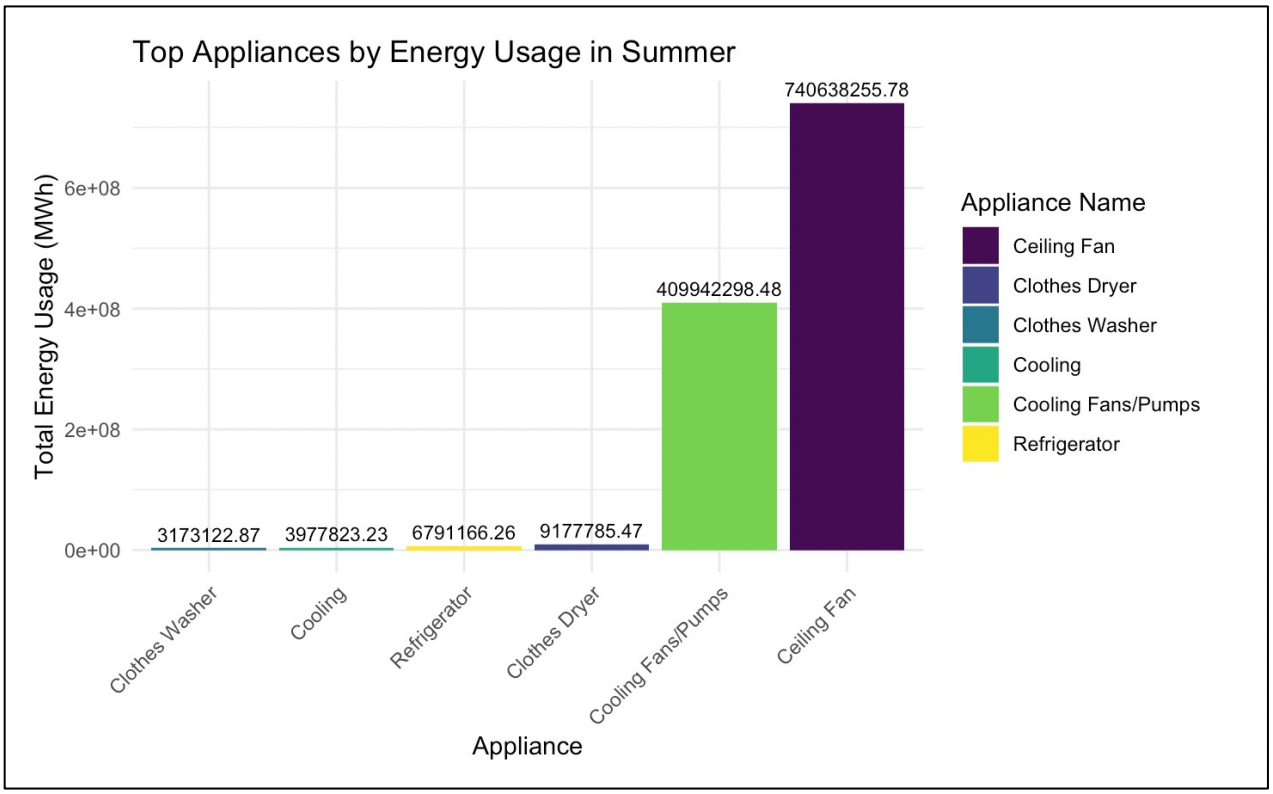
The key insight from this visualization lies in the potential for substantial energy conservation and cost efficiency by transitioning from EF 17.6 models to more efficient EF 19.9+ models. The

comparison between these two segments reveals that while EF 17.6 models are prevalent, EF 19.9+ models represent an opportunity for significant energy savings.

By upgrading a substantial portion of the existing EF 17.6 models to more energy-efficient EF 19.9+ models, there's a potential for substantial energy conservation. These higher EF-rated models utilize energy more effectively, resulting in reduced energy consumption and long-term cost benefits for users.

This transition aligns seamlessly with energy-saving initiatives and environmental goals by significantly reducing energy usage and its associated environmental impact. By advocating for and facilitating the adoption of more energy-efficient appliances, there's a tangible contribution to broader sustainability objectives, promoting a more environmentally conscious approach to energy consumption.

c. Use Energy star rated ceiling fans



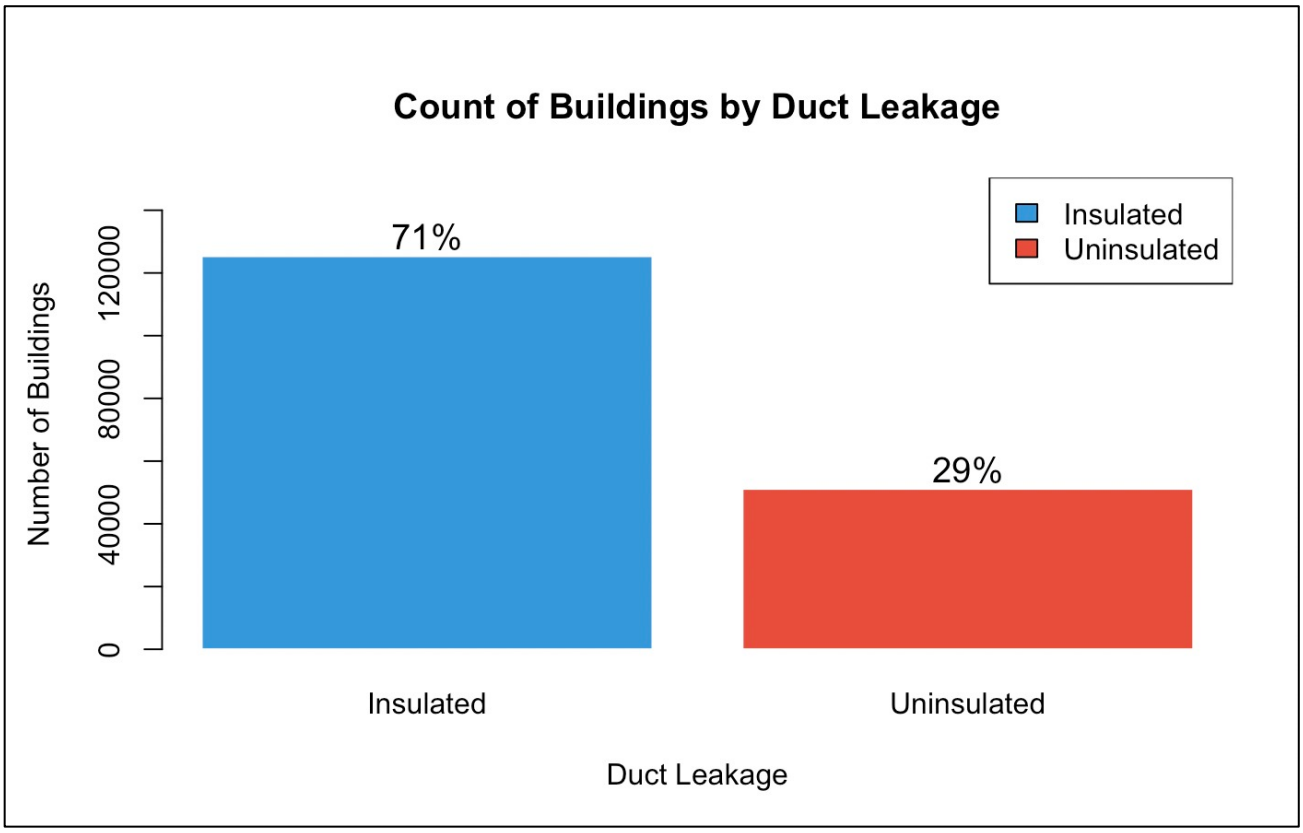
The bar graph offers a comparative view of energy usage by various appliances specifically during the summer season, revealing ceiling fans as the highest energy consumers among the appliances depicted. This insight prompts a closer examination of strategies to mitigate energy usage and promote efficiency.

The potential for significant energy savings lies in upgrading to Energy Star-rated ceiling fans. These fans are designed for superior energy efficiency, consuming notably less energy while delivering the same cooling benefits. Transitioning from conventional fans to Energy Star-rated models presents a tangible opportunity to reduce energy consumption significantly.

Moreover, complementing this transition with user education becomes pivotal. Empowering users with knowledge on optimal fan usage practices, such as turning them off when not in use or adjusting settings for efficiency, can further curtail energy waste. Simple behavioral adjustments can yield substantial energy savings over time.

By advocating for the adoption of Energy Star-rated ceiling fans and fostering user awareness on efficient fan usage habits, there's a dual approach to reducing energy consumption. This multifaceted strategy aligns with sustainability goals by not only introducing energy-efficient technologies but also encouraging conscious user behaviors that collectively contribute to reduced energy waste and environmental impact.

d. Seal duct leakage



The bar chart visually represents the comparison between buildings with insulated and uninsulated ductwork, highlighting a notable proportion with uninsulated systems. This observation underscores a substantial potential for energy savings and efficiency enhancements within these structures.

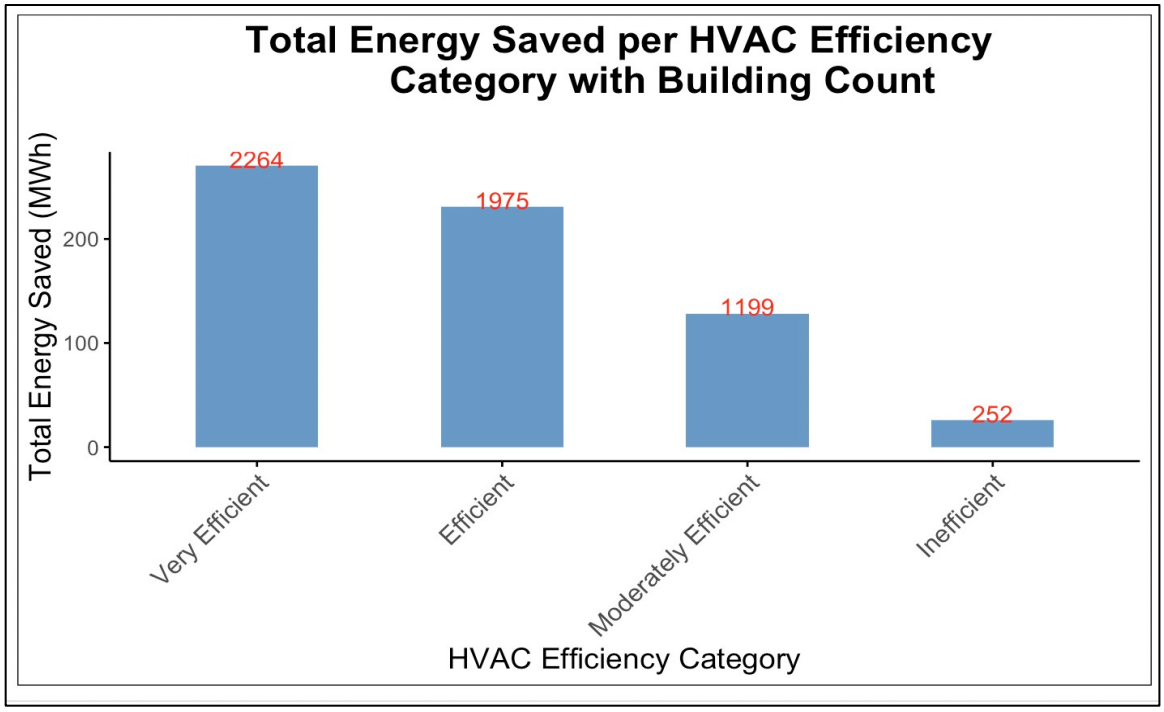
Insulating ductwork and sealing any leaks present significant opportunities for conserving energy. Studies indicate that these measures can lead to remarkable reductions in heating and cooling costs, with potential savings reaching up to 30%. This underscores the immense impact that such upgrades can have on energy efficiency within buildings.

The key lies in addressing the uninsulated ductwork present in a significant portion of buildings. By insulating these ducts and effectively sealing any leaks, the overall energy consumption for heating and cooling can be significantly reduced. This not only translates to considerable cost savings for building owners and occupants but also aligns with broader energy conservation and sustainability initiatives.

The opportunity for energy efficiency improvements in buildings with uninsulated ductwork is substantial. Introducing insulation and sealing measures not only enhances energy efficiency but also contributes to reducing the environmental footprint associated with excessive energy consumption. These actions align with broader sustainability goals, promoting responsible energy usage and fostering a more environmentally conscious approach to building operations.

e. Upgrade inefficient AC systems

hvac_efficiency_category	hvac_efficiency_values
Very Efficient	AC, SEER 15 | Heat Pump | Room AC, EER 12.0
Moderately Efficient	AC, SEER 10 | Room AC, EER 10.7 | Room AC, EER 9.8
Efficient	AC, SEER 13
Inefficient	Shared Cooling | Room AC, EER 8.5 | None | AC, SEER 8



The bar graph showcases the disparity in energy savings among various HVAC efficiency categories, notably highlighting that buildings equipped with 'Very Efficient' HVAC systems yielded the highest energy savings. This observation underscores a substantial potential for both energy conservation and improved comfort levels by addressing the majority of buildings currently utilizing inefficient AC systems, accounting for 68.52%.

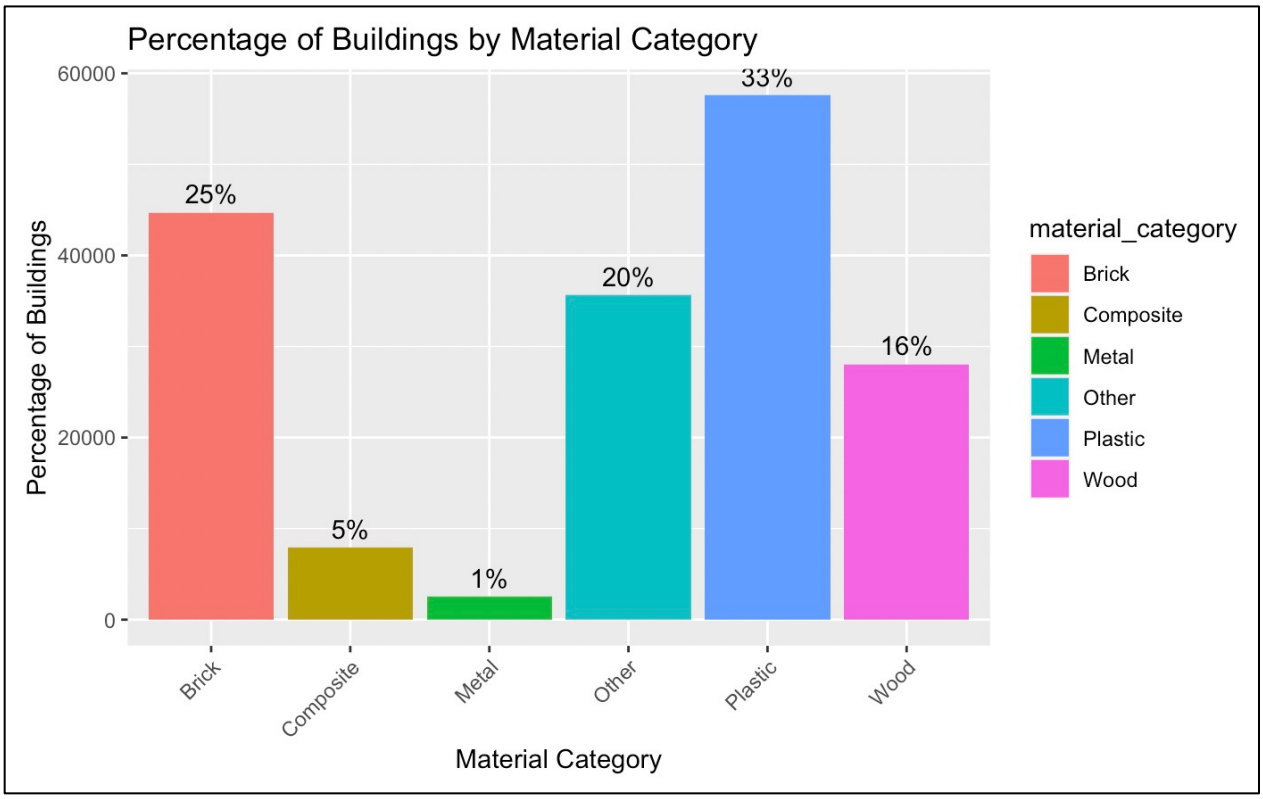
The data emphasizes a clear trend: buildings equipped with 'Very Efficient' HVAC systems demonstrate the highest levels of energy savings. This highlights the immense opportunity for considerable energy conservation and enhanced comfort by transitioning the significant percentage of buildings currently relying on inefficient AC systems to more efficient alternatives.

Upgrading inefficient AC systems to 'Very Efficient' models presents a tangible avenue for reducing energy consumption. Not only does this switch promise substantial energy savings, but it also translates into enhanced comfort for building occupants. More efficient systems offer improved temperature control, better air quality, and often quieter operation, contributing to a more comfortable indoor environment.

Addressing the inefficiency prevalent in the majority of buildings' AC systems aligns with broader energy conservation goals. Such transitions not only contribute significantly to reducing energy consumption and associated costs but also play a crucial role in minimizing the environmental impact stemming from excessive energy usage.

In summary, the data underscores the substantial potential for energy savings and enhanced comfort by upgrading inefficient AC systems to 'Very Efficient' models. This strategic shift not only promises significant energy conservation but also aligns with sustainability goals, fostering more comfortable and environmentally conscious building environments.

f. Implement targeted material for construction



The chart presents the distribution of building materials used in construction, notably highlighting significant percentages attributed to 'Plastic' and 'Other' categories. This distribution emphasizes the diverse range of materials employed in building construction, with these categories holding considerable shares.

For buildings constructed using materials like 'Plastic' and 'Other,' conducting regular energy audits emerges as a pivotal strategy. These audits serve as diagnostic tools to identify and pinpoint areas where energy efficiency improvements are most needed within structures constructed with these materials.

Energy audits enable a comprehensive assessment of a building's energy performance, shedding light on potential inefficiencies and areas for improvement. For buildings constructed using materials like 'Plastic' and 'Other,' these audits can specifically target areas where these materials might affect energy consumption or insulation properties.

By leveraging energy audits, building owners and managers gain valuable insights into optimizing energy usage and enhancing efficiency within structures constructed with these materials. These insights enable the implementation of targeted measures to improve insulation, reduce energy waste, and enhance overall energy performance.

Regular energy audits serve as a proactive approach to identify specific areas within buildings constructed using 'Plastic' or 'Other' materials that might benefit from energy efficiency improvements. This strategy aligns with broader sustainability goals by optimizing energy consumption and reducing the environmental impact associated with inefficient building practices.

OPTIMIZING PEAK ENERGY USAGE

1. Demand Response Programs: Implement demand response initiatives that encourage customers to adjust their energy usage during peak demand periods. Incentivize load-shifting strategies, encouraging consumers to reduce usage during high-demand hours.

2. Energy Efficiency Incentives: Introduce programs that incentivize customers to invest in energy-efficient appliances, smart thermostats, or home insulation. Offer rebates or discounts for adopting these energy-saving measures.

3. Customer Education Campaigns: Launch educational campaigns to raise awareness about energy-saving practices. Provide tips, workshops, or online resources on optimizing energy use during hot weather, such as using fans effectively or setting air conditioning temperatures wisely.

4. Time-of-Use Tariffs: Introduce flexible pricing plans that encourage consumers to consume energy during off-peak hours, thereby reducing pressure on the grid during peak times.

5. Smart Grid Implementation: Invest in smart grid technologies that allow for real-time monitoring and adaptive responses. Smart grids enable better management of energy distribution and can automatically adjust energy flow during peak demand.

6. Collaboration with Local Communities: Partner with local communities to promote sustainability initiatives. Support community-driven efforts for renewable energy adoption or local energy-saving campaigns.

7. Predictive Maintenance: Utilize predictive analytics to anticipate equipment failures or maintenance needs within the grid. This proactive approach can prevent downtime during critical periods.

By implementing these strategies, eSC can effectively manage and potentially reduce peak energy demand, thus mitigating the risk of blackouts during periods of high energy usage. These initiatives not only enhance grid reliability but also align with environmental conservation goals, promoting a more sustainable energy future.

CONCLUSION

The confluence of escalating temperatures and heightened energy demand casts a challenging trajectory for energy suppliers like eSC. Through our comprehensive analysis, we've unveiled the intricate relationship between climatic patterns and energy usage, offering eSC valuable insights to fortify its strategies in the face of escalating demand.

The predictive models we've employed, steeped in regression analysis and predictive analytics, serve as a beacon guiding eSC's approach towards mitigating the impending challenges posed by surging energy demands. These models elucidate the nonlinear effects of temperature on energy usage, forming the bedrock for anticipatory strategies during climatically intense periods, particularly the scorching months of July.

The crux of our findings lays bare the criticality of proactive measures. eSC's preemption of potential grid overload scenarios and blackouts hinges on harnessing the predictive capabilities ingrained in our data-driven approach. By deciphering historical patterns and correlations, eSC is primed to foresee and manage peaks in demand, averting disruptions in energy supply and ensuring uninterrupted service to customers.

Our recommendations, born from meticulous analysis, extend beyond mere short-term mitigation. They encompass a comprehensive suite of strategies encompassing demand response initiatives, energy-efficient practices, and tailored customer engagement. These initiatives collectively serve a dual purpose: immediate reduction of energy usage and the cultivation of a culture of responsible energy consumption among eSC's clientele.

The future-oriented stance of eSC, driven by our data-backed insights, propels the company towards sustainable grid management. By integrating smart grid technologies and fostering collaborative ties within communities, eSC strides not only to meet but to surpass contemporary challenges. This deliberate approach

not only reinforces grid reliability but also champions eSC's commitment to environmental stewardship and customer satisfaction.

In summation, our data-centric strategies pave the path for eSC to navigate the labyrinthine landscape of climate-induced energy demand fluctuations. This forward-thinking approach not only assures grid stability but also positions eSC as a vanguard in forging a sustainable, resilient energy ecosystem.

GROUP TASK DIVISION

Task	Group Member
Data Analysis, Data Preparation	Meghana Inavilli Abhijeet Baviskar
Data Cleaning, Data Merging	Vrushali Lad Abhijeet Baviskar
Data Modeling	Shilpa Pillai Vrushali Lad
Shiny Apps	Rijul Ugawekar Meghana Inavilli