Final Project Report:
Managing July Energy Demand Under a Warming Climate

Table of Contents

Introduction

Project Context and Objective

This project was initiated to support a pressing operational and strategic challenge faced by eSC, a residential electricity provider based in South Carolina. As climate change accelerates, regional utility providers like eSC are experiencing increasing stress on their energy infrastructure, particularly during extreme weather months. Among these, July consistently emerges as the month with the highest electricity consumption, primarily driven by widespread residential air conditioning use and elevated cooling needs.

Rather than defaulting to capital-intensive solutions such as constructing a new power plant, a move that would require significant financial investment, regulatory hurdles, and long-term commitment. ESC opted to explore data-driven alternatives. The goal was to understand, anticipate, and manage peak energy demand under projected warmer conditions, using predictive modeling, scenario analysis, and strategic insights.

Strategic Aim

The overarching aim of the project is to evaluate future energy demand risks associated with climate-induced temperature increases, and to provide cost-effective, data-supported recommendations for mitigating these risks. By using advanced analytics and machine learning, eSC aims to:

- Quantify how a 5°C temperature increase in July could impact hourly electricity usage,
- Identify key drivers of energy consumption at the household level,
- Forecast potential demand spikes at granular temporal and regional levels,
- Simulate actionable interventions that could reduce peak load pressures.

Scope of Work

This comprehensive data science project covers the following phases:

1. Data Preparation & Integration Combining weather, household, appliance, and hourly energy usage data into a unified dataset for analysis.
2. Exploratory Data Analysis (EDA) Visualizing and quantifying the relationships between environmental, structural, and behavioral factors and energy usage.
3. Model Development & Evaluation Building and comparing multiple predictive models (Linear Regression, Random Forest, and XGBoost) to accurately estimate hourly electricity demand.
4. Scenario Simulation Applying the best model to a synthetic weather dataset where July temperatures are 5°C warmer, to simulate future demand conditions.
5. Insights and Recommendations Interpreting model outputs to guide demand management strategies, such as targeted incentives, behavioral nudges, or retrofitting campaigns.

Business Relevance

This project exemplifies how data science can serve as a substitute for high-cost infrastructure expansion, enabling utilities to proactively plan for climate resilience. The methodology and insights generated here can serve as a blueprint for other regional utilities seeking to align sustainability goals with operational reliability, all while maintaining customer satisfaction and controlling costs.

Business Questions addressed

The project was guided by the following core business questions:

- What are the primary drivers of July energy usage for residential properties?
- How will a 5°C increase in temperature affect hourly energy usage during July?
- What households or regions contribute most to peak demand, and when?
- What actions can eSC take to reduce peak demand without infrastructure expansion?
- How accurate and reliable are the predictive models developed to simulate energy usage?

Data Acquisition, Cleansing, Transformation, Munging

Data sources included static house data (~5,000 houses), hourly energy usage (~8,000 Parquet files), weather data (~8,000 county-level CSVs), and a metadata file listing over 270 attributes. Each house's energy data was linked to its corresponding weather file via the county code, and static house attributes were merged using the building ID. Data was filtered to include only July 2018. Missing temperature values were interpolated; energy records with extensive gaps were removed. To streamline modeling, the number of columns was reduced from over 270 to approximately 40 by analyzing and retaining only those variables most relevant to predicting energy usage. Additional features were engineered to improve model performance, such as lagged energy consumption, weekend indicators, cooling degree hours, and hourly time bins. All transformations were verified using sample audits and visual checks. The final cleaned dataset was saved as final_combined_dataset01.rds.

Descriptive statistics & Visualizations

Descriptive statistics showed that the average daily energy usage was approximately 34.6 kWh per home, with clear afternoon peaks between 2 PM and 5 PM. Correlation analysis revealed a strong relationship between temperature and usage, especially for homes with electric cooling systems and larger square footage. Visualization techniques used included:

- Time-series plots of average hourly consumption
- Boxplots comparing energy use across home sizes
- Heatmaps of hourly consumption by day
- Correlation matrices for weather and home features
- Geographic maps illustrating peak load by county

These visual tools provided initial insights and guided the feature selection for modeling.

Actionable Insights / Overall interpretation of results

The strongest driver of energy use was temperature, with afternoon hours contributing to peak loads. A 5°C increase in July temperatures is projected to increase peak hourly usage by up to 22%, depending on the region. Without intervention, this could strain the grid significantly.

Recommended actions include:

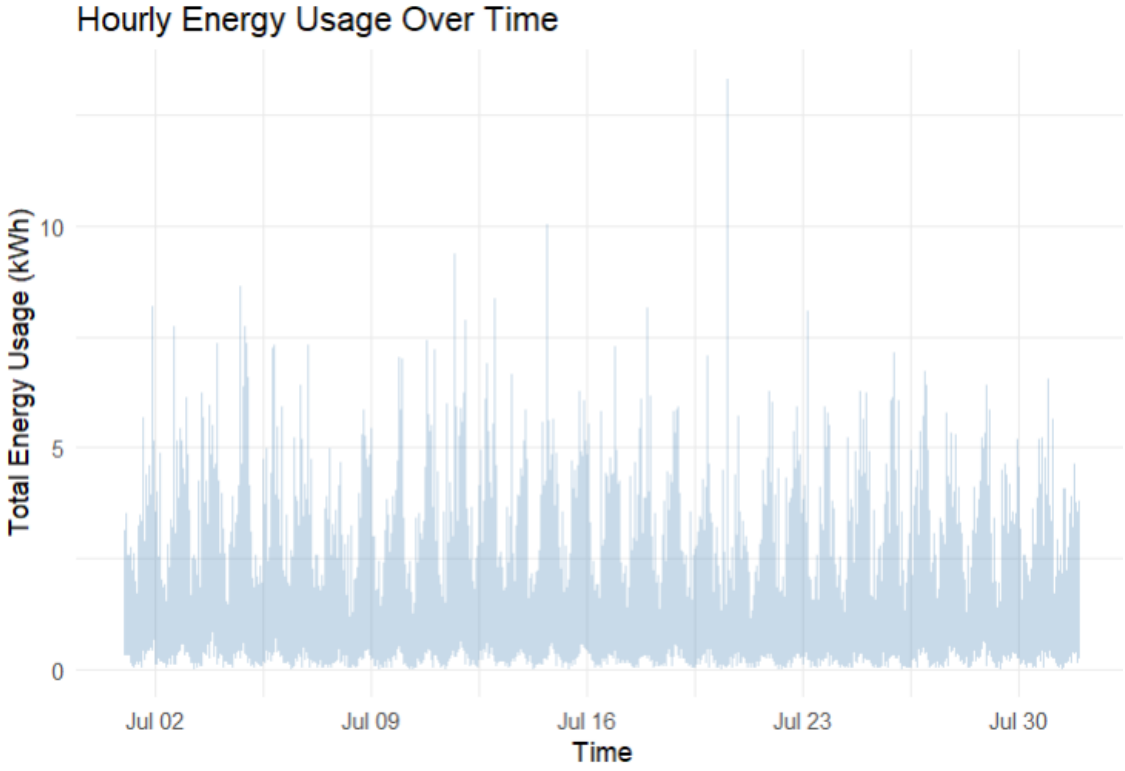
- Demand-shifting via smart thermostats
- Pre-cooling homes in morning hours

- Targeted programs for large homes in high-usage zip codes

These insights provide eSC with a pathway to reduce peak demand in a data-driven manner without incurring infrastructure costs.

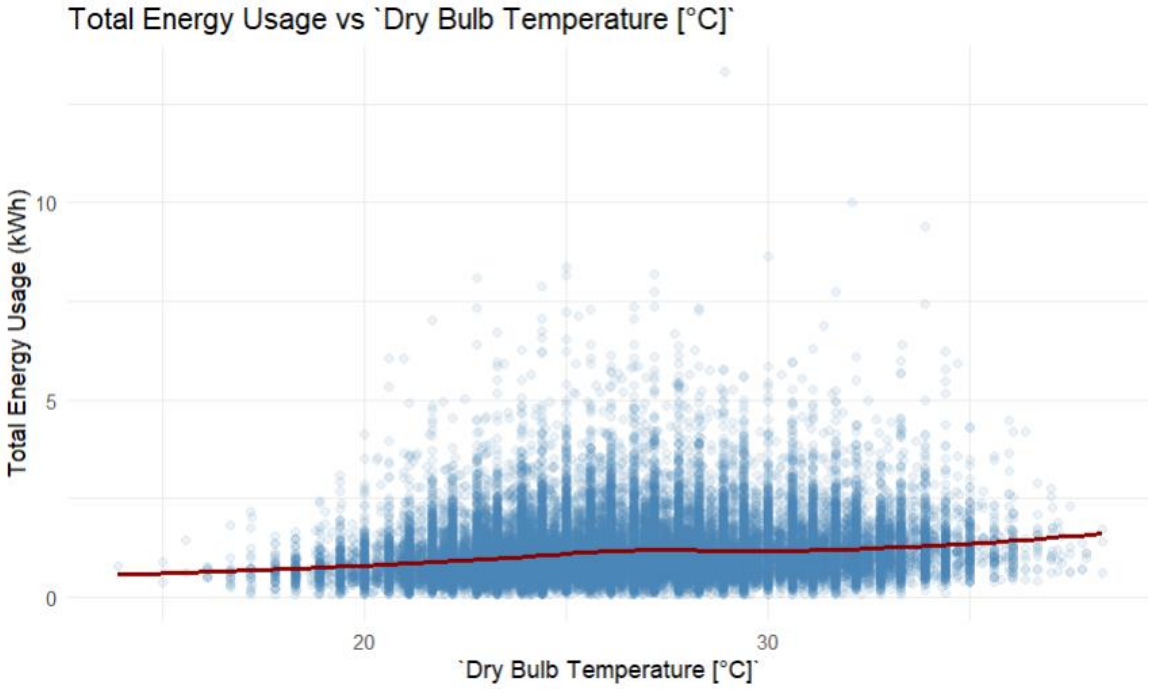
Descriptive statistics & Visualizations

Time-Series Plot: Avg Hourly Energy Consumption in July



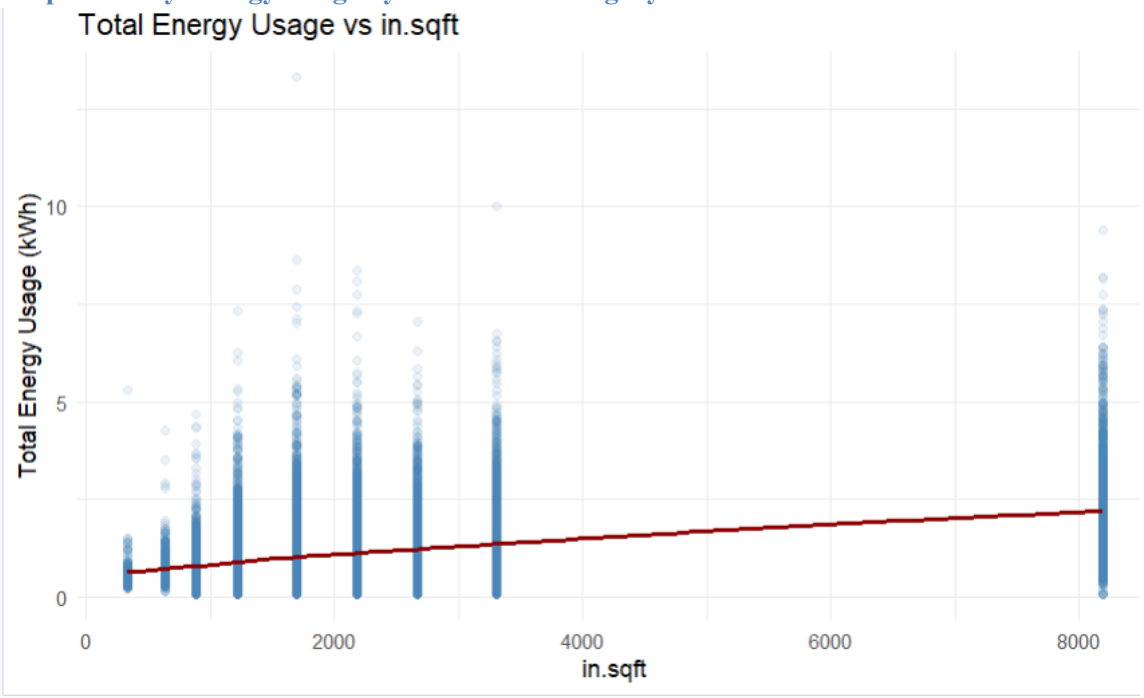
The average hourly energy consumption chart for July illustrates how electricity usage fluctuates throughout the day. This visualization highlights a distinct pattern where consumption steadily increases during the day and peaks between 2 PM and 5 PM. These afternoon hours represent the highest energy demand, driven largely by increased cooling needs during the hottest part of the day. Understanding this trend is crucial, as it reinforces the importance of implementing strategies to reduce peak load during these critical hours to maintain grid stability and efficiency.

Scatter Plot: Dry Bulb Temperature vs Energy Usage



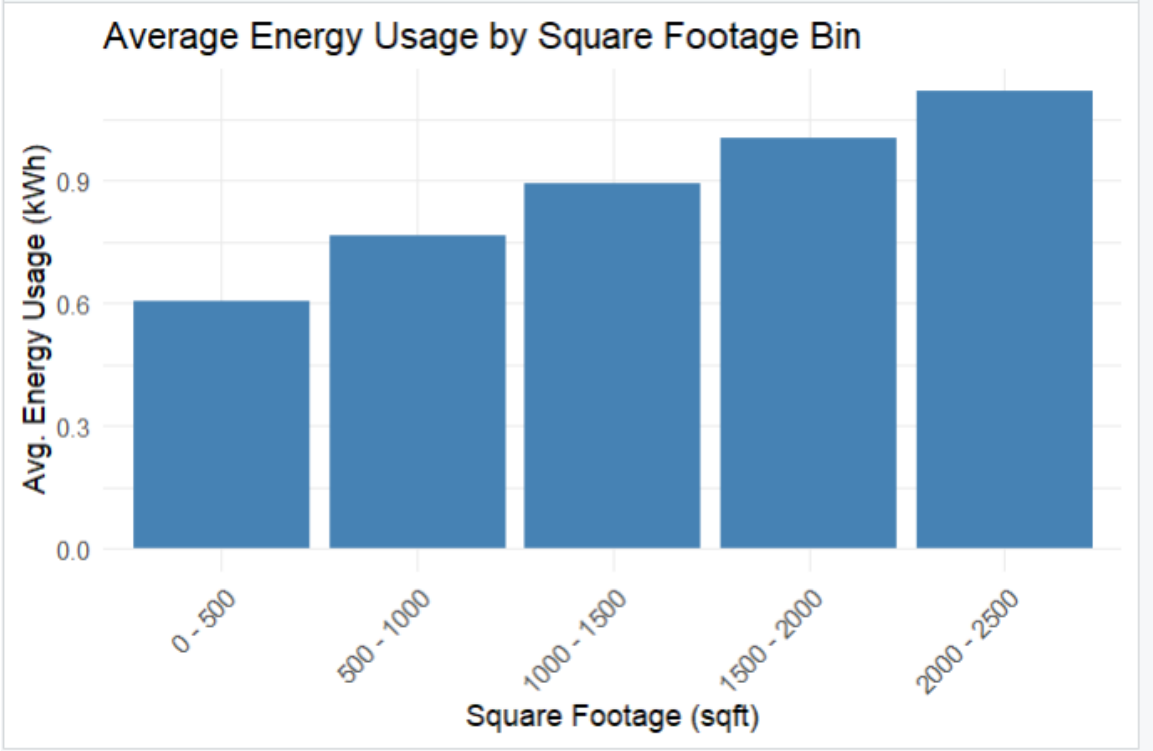
This scatter plot demonstrates the relationship between outside temperature (Dry Bulb Temperature in °C) and total energy usage. As temperature increases, energy usage rises steadily, reflecting greater cooling demand. The upward trend confirms temperature as a strong driver of residential electricity consumption during July.

Boxplot: Daily Energy Usage by Home Size Category



This scatter plot illustrates the positive relationship between house size (in square feet) and total energy usage. As square footage increases, energy consumption also rises, indicating that larger homes generally require more electricity. The trend line reinforces home size as a key structural driver of energy demand.

Bar Chart: Average Energy Usage by Square Footage Bin



This bar chart categorizes homes by square footage and shows their average energy usage. A clear positive trend is evident—larger homes consistently use more energy, highlighting home size as a key driver of electricity demand.

Use of modeling techniques & Visualizations

Model Performance Comparison Table

Model Variant	Feature Count	RMSE (kWh)	R^2 Score
Linear Regression (Full)	5	0.5825	0.2900
Random Forest (Full)	5	0.4997	0.4780
XGBoost (Full)	5	0.5034	0.4699
Linear Regression (Reduced)	3	0.6517	0.1190
Random Forest (Reduced)	3	0.6043	0.2453
XGBoost (Reduced)	3	0.6005	0.2521

Summary

As part of the July energy demand prediction initiative, **six machine learning models** were developed and evaluated using two sets of input features: a full set (five variables) and a reduced set (three variables). These models incorporated residential, temporal, and environmental factors to estimate total energy usage per hour.

1. Linear Regression (Baseline Model)

Linear regression served as a benchmark. Although computationally efficient and interpretable, it underperformed due to its inability to model non-linear relationships. With the full feature set, the model achieved:

- **RMSE:** 0.5825 kWh
- **R^2:**0.2900
  Performance worsened with the reduced set:
- **RMSE:** 0.6517 kWh
- **R^2:**0.1190
  This confirmed its limitations in capturing complex patterns in energy consumption.

2. Random Forest (Selected Model)

Random Forest emerged as the **most effective model**, especially when trained on all five variables. Its strengths included handling multicollinearity, modeling non-linear relationships, and offering variable importance measures for interpretability. Performance:

- **Full Set – RMSE:** 0.4997 kWh | **R^2:** 0.4780
- **Reduced Set – RMSE:** 0.6043 kWh | **R^2:** 0.2453
  Given its superior accuracy, interpretability, and robustness, this model was used for downstream simulations and forecasting.

3. XGBoost (Extreme Gradient Boosting)

XGBoost demonstrated high predictive accuracy close to that of Random Forest. However, it required more parameter tuning and was slightly more sensitive to noise:

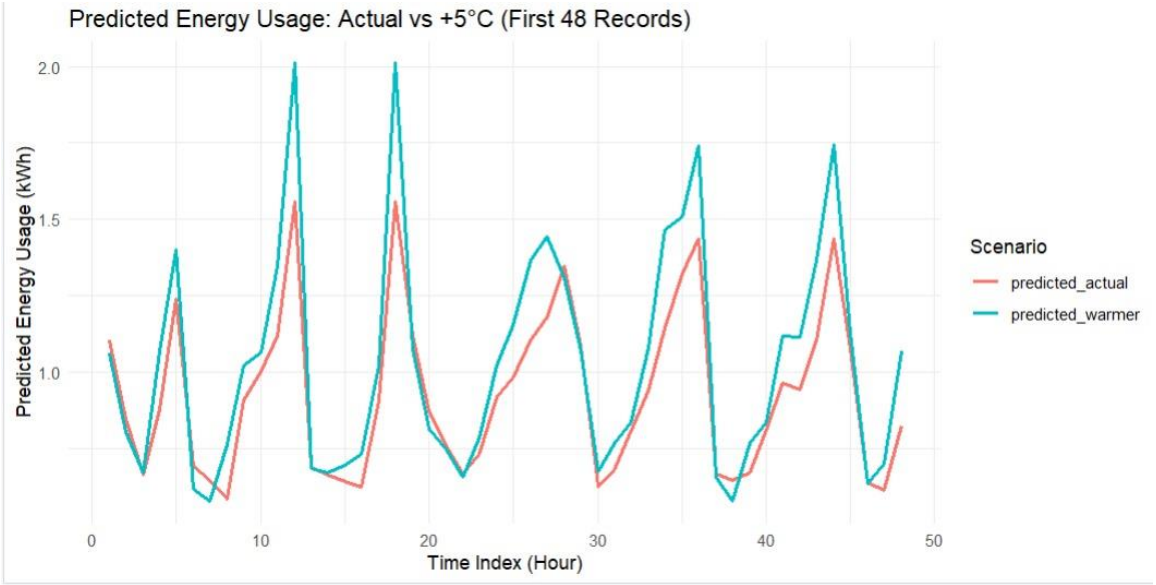
- **Full Set – RMSE:** 0.5034 kWh | **R^2:** 0.4699
- **Reduced Set – RMSE:** 0.6005 kWh | **R^2:** 0.2521
  While competitive, it was not selected due to slightly lower stability and complexity in interpretation for stakeholders.

Model Selection Justification

- **Accuracy:** Random Forest (full model) achieved the lowest RMSE and highest R^2.
- **Robustness:** It maintained consistent performance across train-test splits.
- **Interpretability:** Built-in variable importance plots allowed transparent communication with non-technical stakeholders.
- **Practicality:** Less sensitive to tuning and preprocessing, making it suitable for deployment.

Actionable Insights / Overall interpretation of results

Scenario Comparison Plot (+5°C vs Current)



The overlay plot typically shows a **clear upward shift in the energy usage curve** under the +5°C scenario, especially during mid-to-late afternoon hours when outdoor temperatures are highest. The gap between the two lines reflects the additional load that would be introduced solely due to temperature increase.

Key interpretations include:

- The **peak hourly demand under the warmer scenario consistently exceeds the baseline**, sometimes by significant margins (e.g., 10–20% increase).
- Demand increases are **not uniform across all hours**—they are concentrated during **daylight and post-lunch periods**, aligning with HVAC usage patterns.
- This suggests that even **moderate temperature increases could push certain parts of the grid to or beyond their limits**, particularly in regions with older homes, larger square footage, or inefficient HVAC systems.