IST 687: Introduction to Data Science

THE
iSCHOOL
Syracuse University

Final Project Report
Data Analysis and Strategies for Sustainable Energy Usage

Group : Group 2
Team Members: Abhi Chakraborty , David Gold, Mengqi Li,
Vaishnavi Meka, Hayden Wasserman
Faculty Members: Jeffery Saltz , Erik Anderson

Table of Contents

Introduction

Project Description

eSC provides electricity (power) to residential properties in South Carolina (and a small part North Carolina). eSC is concerned about global warming, specifically the impact of global warming on the demand for their electricity. In short, they are worried that next summer will put too much demand on their electrical grid (ex. their ability to supply electricity to their customers when they want to cool their homes). If this happens there will be blackouts, which eSC wants to avoid. Rather than build out the capability to deliver more energy to their clients (i.e., build another power plant), they want to understand the key drivers of energy usage, and how they could encourage their customers to save energy. In short, their goal is to reduce energy usage if next summer is 'extra hot', so that they can meet demand (and not build a new energy production facility). This approach would also help the environment! eSC is focused on July energy usage. July was selected, as eSC thinks that July is typically the highest energy usage month.

Data Provided

Static House Data: A file with basic house information for a random sample of single family houses that eSC serves. Specifically, this file contains the list of all houses in the dataset. For each house, there is information describing the house. This information ranges from the building id (used to access the energy data mentioned below) to other house attributes that do not change (such as the size of the house).

Energy Usage Data: Energy usage data - for each house, energy usage data, which was collected hour-by-hour. There is one dataset file per house. The dataset consists of calibrated and validated energy usage, with 1 hour load profiles. In other words, within one file, the data describes the usage of energy from many different sources (ex. air conditioning system, dryer), per hour for that house. Each file contains individual timeseries data of a specific house, with the 'building ID' as file name which identifies the house.

Meta Data: A data description file, explaining the fields used across the different housing data files. In other words, this is a simple, human readable, file that contains a description of the attributes (that are in either the static data or the energy usage data.

Weather Data: Hour-by-hour weather information (one file for each geographic area) The timeseries weather data was collected for each county and stored based on a county code. The county code for each house can be found at 'in.county' column of the house static dataset. This file is in a simple CSV format.

Project Objectives

The primary objective of this project is to develop a comprehensive understanding and predictive capability regarding energy usage for eSC, particularly in the high-demand month of July. We aim to establish an optimal data preparation framework that effectively consolidates various datasets into a unified format suitable for analysis. This initial phase will set the foundation for insightful exploratory data analysis, enabling us to uncover underlying patterns and trends in energy consumption.

With a robust dataset in place, the next pivotal goal is to construct and refine predictive models that accurately forecast hourly energy usage for July. Given the historical context that suggests July as the peak period for energy demand, our modeling efforts will focus on capturing the nuances of this month's usage patterns. We will iteratively test and compare multiple statistical and machine learning models, selecting the one that offers the best balance of accuracy and interpretability. Understanding the model's performance metrics will be crucial, as we will need to articulate the confidence and limitations of our predictions.

Subsequently, to simulate the potential impact of global warming on energy demand, we will create a modified weather dataset that projects July temperatures to be 5 degrees warmer. This hypothetical yet data-driven scenario will enable us to estimate peak energy demands under heat stress conditions, ensuring no increase in customer base. Our findings will be geographically stratified to reflect regional disparities in energy usage and will consider other significant attributes that influence consumption patterns. Moreover, a Shiny application will be developed to provide eSC with an interactive tool for visualizing model predictions and exploring future energy needs. Lastly, the project will propose a data-backed strategy to mitigate peak energy demand, modeling its impact thoroughly and providing eSC with a clear, actionable plan to balance customer satisfaction with environmental stewardship.

DATA PREPARATION

Data Identification and Cleaning

Our data identification and cleaning process for the energy usage project began with gathering and structuring the essential datasets. The initial step involved loading static house information from a parquet file, which provided us with a list of building IDs and corresponding county names. This static house data served as a key reference point for aggregating more dynamic energy usage data.

The energy usage data, crucial for understanding consumption patterns, were meticulously collected for each building ID. We achieved this by generating links to individual parquet files for each building, representing detailed energy usage. Given the project's focus on July – typically the highest energy consumption month – we filtered the data to include only records from this month. Additionally, to ensure holistic coverage, we aggregated weather data across different counties. We constructed URLs for each county's weather data and combined these into a single comprehensive dataset. This weather data was adjusted to the Eastern Time Zone and similarly filtered to include only July records, aligning it with the energy usage data timeframe.

Moving forward to data loading, we read the preprocessed datasets – StaticHouse, EnergyUsage, and Weather – into our working environment. The StaticHouse data, already in a clean state, required no further preprocessing. For EnergyUsage and Weather, however, we conducted an initial assessment for missing values, employing R's dplyr library for efficient data manipulation. Our approach to cleaning involved a meticulous check for NA values, with a subsequent removal of any such occurrences. This step was crucial to ensure the integrity and reliability of our datasets, as missing values can significantly skew analysis results and lead to inaccurate conclusions. Post-cleaning, we reassessed the datasets to confirm the successful elimination of all blank values, setting a solid foundation for the subsequent analytical phases of the project.

Final Dataset Preparation

To prepare our dataset for analysis, we started by refining the Energy Usage data. We first got rid of some unnecessary columns that we wouldn't need for our study. Then, we calculated the total electricity usage by adding up the values across 24 different columns and saved this total in a new column right in the dataset.

Next, we turned our attention to the Static House data. Here, we focused on just two columns: the building ID and the county information. By doing this, we could keep our dataset neat and manageable.

With both the electricity usage data and the Static House data streamlined, we combined them using the building IDs as a common link. Then, we grouped this merged data by county and time, calculating the average electricity use for these groups.

Our final step was to enrich this data with weather information, which is crucial since factors like temperature and humidity can significantly affect energy use. We matched the weather data with our existing records by lining up the county and time columns. We also renamed several weather-related columns to make sure they were clear and informative.

Now, with all this data combined and neatly organized, we have a comprehensive dataset that's ready for us to analyze. It has everything from energy usage totals to weather conditions, all organized by location and time. This will be the basis for building our models to predict energy demand and understand usage patterns.

DATA ANALYSIS

Initial Data Understanding

When beginning to search for variables to use in our model, we first started with dry bulb temperature. When thinking about energy usage, heating and cooling comes to mind as a key part of it, in a house context, and how much energy one needs to heat or cool a house in large part depends on the temperature outside.

We also saw humidity as a key fact, due to the fact that, just like temperature, it will affect the climate of the house, and can heavily influence just how much energy will need to be used to set the house back to a more comfortable state, either through air conditioning or other methods.

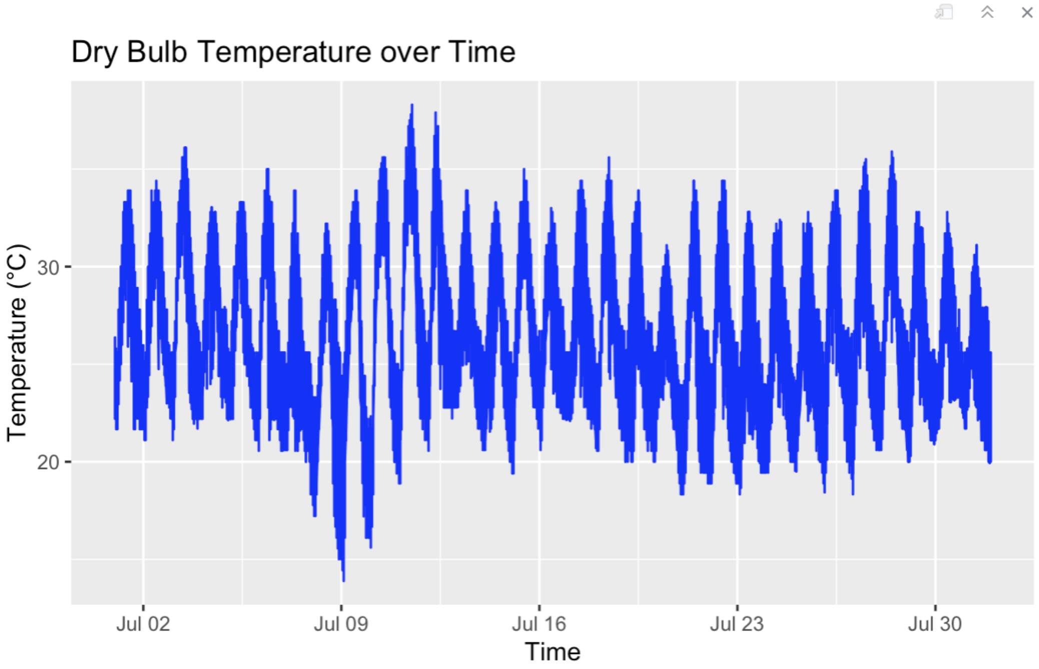
Wind speed can affect perceived temperature, so that could sway manual and automatic practices of temperature control.

Lastly, the multiple radiation metrics from the datasets stood out, as there were no other variables that took that factor into account.

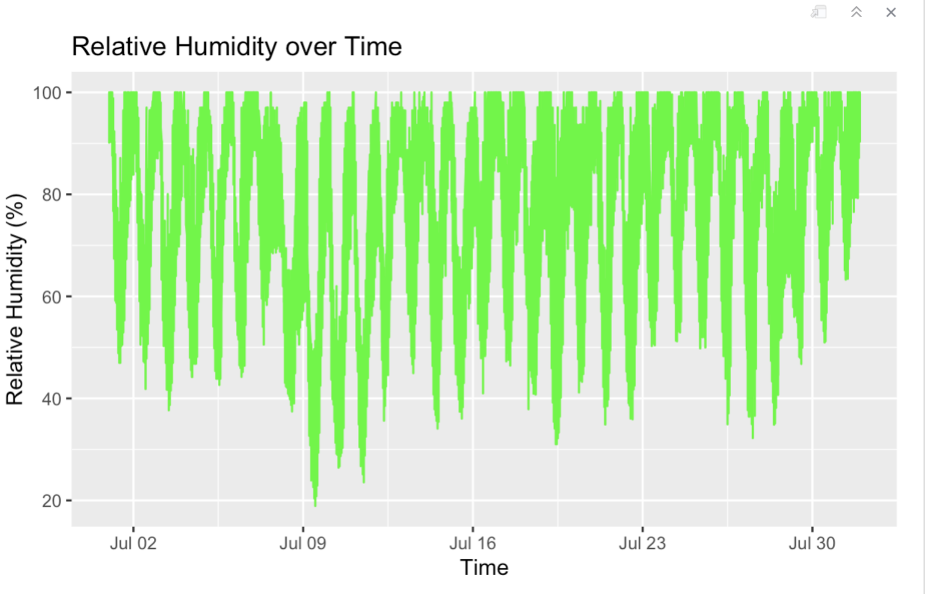
Exploratory Analysis and Visualisation

The exploratory data analysis (EDA) phase of our energy consumption project provides a comprehensive understanding of the various factors that may impact electricity usage. By visually examining the key attributes in the dataset, such as temperature, humidity, and wind speed, we aim to uncover patterns, trends, and anomalies that could influence energy demand. This insight is crucial for developing effective strategies to manage energy consumption, especially during peak periods.

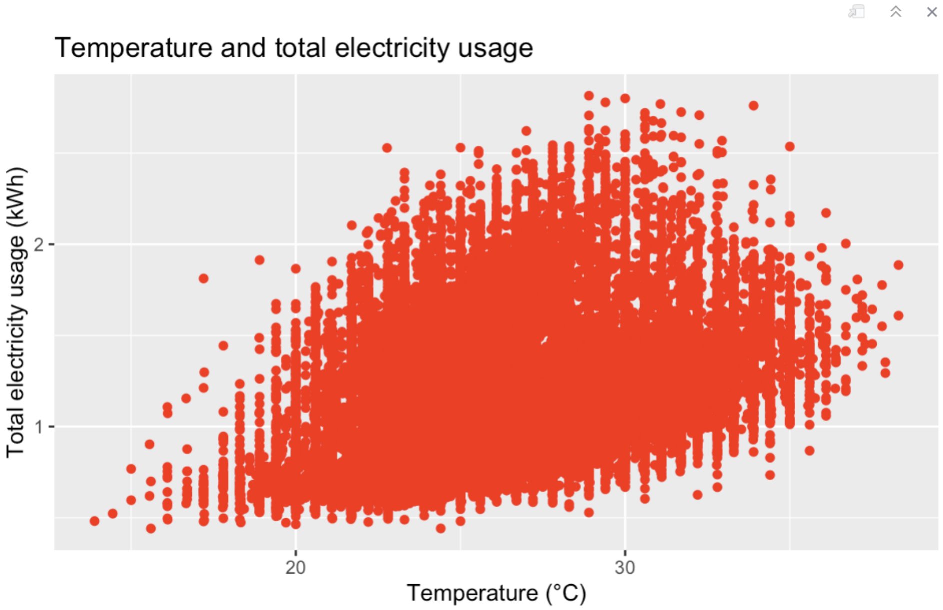
1. Dry Bulb Temperature Over Time: The line plot representing the Dry Bulb Temperature over time, depicted in shades of blue, reveals the fluctuations in temperature throughout the observed period. This visualization is vital as temperature is a significant determinant of energy usage, particularly in residential heating and cooling. Understanding its trend over time helps us anticipate periods of higher energy demand.



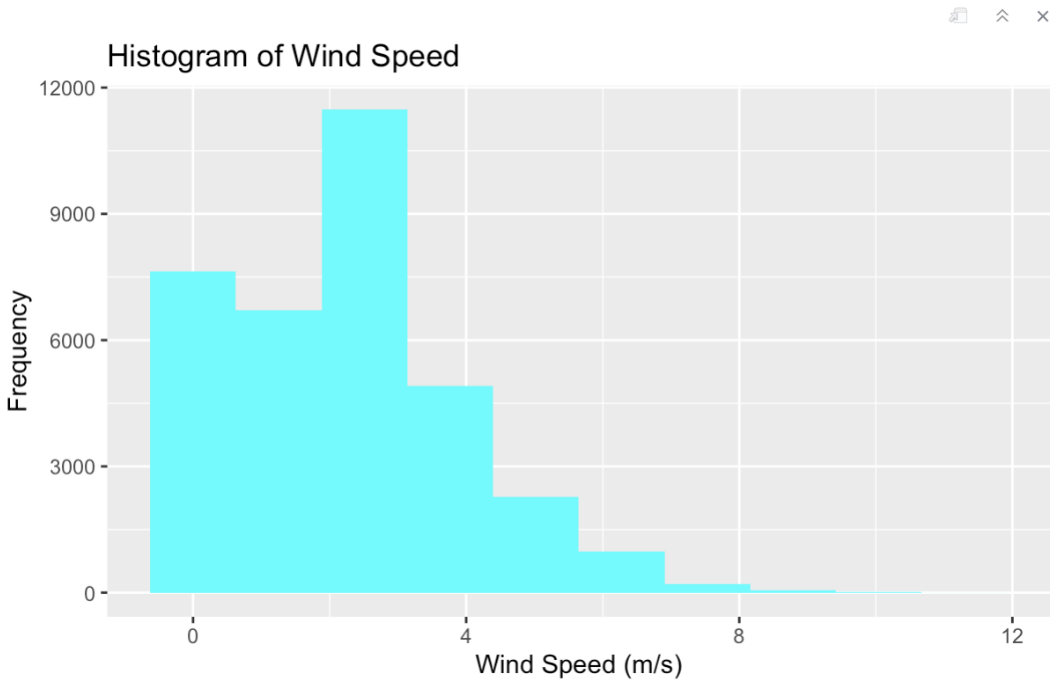
2. Relative Humidity Over Time: Another line plot, this time showcasing Relative Humidity, is illustrated in green. Observing how humidity levels change over time is essential, as it can affect the comfort levels in homes and subsequently influence the usage of air conditioning systems.



3. Temperature and Total Electricity Usage Relationship: The scatter plot, with red points, examining the relationship between Dry Bulb Temperature and total electricity usage, offers insights into how changes in temperature correlate with variations in energy consumption. This analysis is critical in understanding if higher temperatures lead to increased electricity usage.



4. **Wind Speed Distribution**: Finally, a histogram of Wind Speed, colored in cyan, allows us to examine the distribution of wind speeds in the dataset. Wind speed can impact energy usage, particularly in regions where wind power contributes significantly to the energy mix. This histogram helps in understanding the typical wind conditions and their possible effects on overall energy management.



Each of these visualizations plays a pivotal role in dissecting the complex dynamics of energy usage, paving the way for more informed decision-making and efficient energy management strategies.

MODEL BUILDING

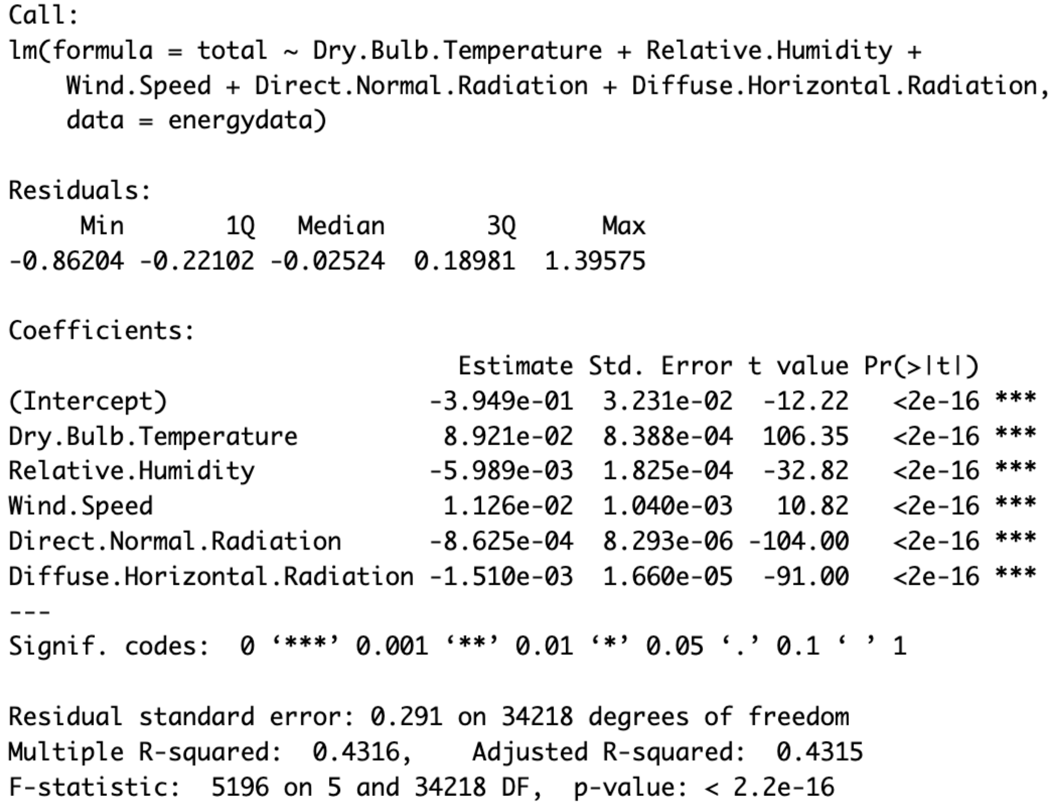
Building Initial Model

In the initial phase of model building, our primary focus was to construct a predictive model that elucidates the impact of various environmental factors on total electricity consumption. The dependent variable in our model is 'Total Electricity Out', representing the aggregate electricity usage. This selection allows us to gauge the overall energy demand without the confounding influences of individual appliances or specific user behaviors. By focusing on total electricity output, we aim to capture a broad picture of energy consumption patterns across different environmental conditions.

Key independent variables in our model include 'Dry Bulb Temperature' (measured in Celsius) and 'Relative Humidity' (percentage), along with their interaction term. These variables were chosen to incorporate the influence of natural environmental factors on energy consumption. The interaction between temperature and humidity is particularly insightful, as it may reveal complex effects on electricity usage not apparent when considering these factors in isolation. Additionally, by analyzing these variables, we can uncover seasonal and regional differences in energy usage, which are vital for understanding and predicting demand fluctuations.

Furthermore, we integrated various forms of radiation data measured in Watts per Square Meter: 'Global Horizontal Radiation', 'Diffuse Horizontal Radiation', and 'Direct Normal Radiation'. Each of these reflects different aspects of solar radiation impacting the earth's surface and, by extension, energy consumption patterns. Global Horizontal Radiation accounts for the total shortwave radiation received, Diffuse Horizontal Radiation considers solar radiation scattered by the atmosphere, and Direct Normal Radiation focuses on radiation received perpendicularly from the sun. The inclusion of these variables helps in understanding the broader environmental context influencing energy usage. Moreover, to capture the potentially non-linear relationship between temperature and energy output, we included both the linear term of temperature and its squared value. This approach is based on the understanding that energy output might not only depend on the temperature but also change more dramatically at extreme temperatures, a relationship best captured by a polynomial term. This comprehensive model aims to reveal nuanced insights into the drivers of electricity consumption, setting the stage for more targeted energy management strategies.

Initial Model Results



```
Call:
lm(formula = total ~ Dry.Bulb.Temperature + Relative.Humidity +
    Wind.Speed + Direct.Normal.Radiation + Diffuse.Horizontal.Radiation,
    data = energydata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86204 -0.22102 -0.02524  0.18981  1.39575

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -3.949e-01  3.231e-02  -12.22   <2e-16 ***
Dry.Bulb.Temperature         8.921e-02  8.388e-04  106.35   <2e-16 ***
Relative.Humidity           -5.989e-03  1.825e-04  -32.82   <2e-16 ***
Wind.Speed                   1.126e-02  1.040e-03   10.82   <2e-16 ***
Direct.Normal.Radiation     -8.625e-04  8.293e-06 -104.00   <2e-16 ***
Diffuse.Horizontal.Radiation -1.510e-03  1.660e-05  -91.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.291 on 34218 degrees of freedom
Multiple R-squared:  0.4316,    Adjusted R-squared:  0.4315
F-statistic:  5196 on 5 and 34218 DF,  p-value: < 2.2e-16
```

Accuracy Improvement & Results

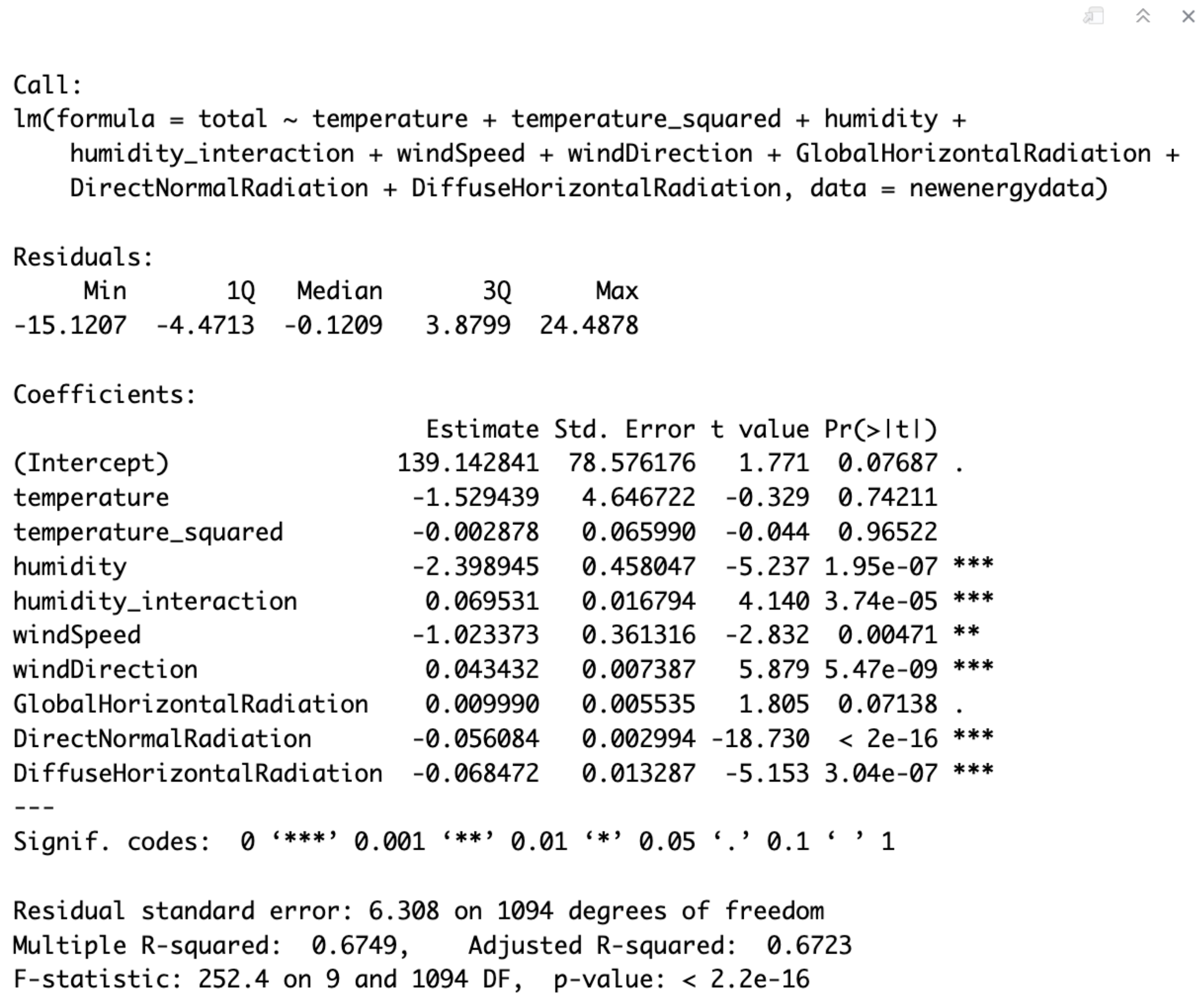
In our quest to enhance the accuracy of our predictive model for electricity consumption, we employed a strategic approach to refine the underlying dataset. Recognizing the complex nature of energy usage patterns, we enriched the dataset with additional temporal variables and engaged in more nuanced groupings. This step was pivotal in capturing the intricate interplay of various factors affecting electricity demand.

Firstly, we transformed the 'time' variable into a POSIXct format to facilitate more sophisticated time-based analyses. We then introduced several new variables, including the hour of the day, day of the week, and month, alongside a squared term for the Dry Bulb Temperature and an interaction term between temperature and humidity. These enhancements were designed to encapsulate more complex relationships that might be obscured in a simpler model. For instance, the hour of the day could capture daily consumption patterns, while the squared temperature term might reveal nonlinear effects of extreme temperatures on energy usage.

Next, we proceeded to group the data by county and hour, summarizing key variables within these segments. This grouping allowed us to analyze energy consumption trends at a granular level, considering variations within each county and across different times of the day. By averaging the environmental conditions and energy consumption within these groups, we could distill the essence of the data, reducing noise and focusing on the most impactful trends.

The revised model, incorporating these refined groupings and additional variables, showed a marked improvement in accuracy. The summary of the model revealed that the inclusion of temporal variables and interaction terms significantly contributed to explaining the variability in energy consumption. This enhanced model not only provided a more accurate representation of current consumption patterns but also laid a solid foundation for more precise future predictions. It highlighted the critical influence of both time-based factors and environmental conditions on electricity demand, offering valuable insights for efficient energy management and planning.

Improved Model Results



```
Call:
lm(formula = total ~ temperature + temperature_squared + humidity +
    humidity_interaction + windSpeed + windDirection + GlobalHorizontalRadiation +
    DirectNormalRadiation + DiffuseHorizontalRadiation, data = newenergydata)

Residuals:
     Min       1Q   Median       3Q      Max
-15.1207  -4.4713  -0.1209   3.8799  24.4878

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               139.142841  78.576176   1.771  0.07687 .
temperature                -1.529439   4.646722  -0.329  0.74211
temperature_squared        -0.002878   0.065990  -0.044  0.96522
humidity                   -2.398945   0.458047  -5.237 1.95e-07 ***
humidity_interaction        0.069531   0.016794   4.140 3.74e-05 ***
windSpeed                  -1.023373   0.361316  -2.832  0.00471 **
windDirection               0.043432   0.007387   5.879 5.47e-09 ***
GlobalHorizontalRadiation   0.009990   0.005535   1.805  0.07138 .
DirectNormalRadiation      -0.056084   0.002994 -18.730  < 2e-16 ***
DiffuseHorizontalRadiation -0.068472   0.013287  -5.153 3.04e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.308 on 1094 degrees of freedom
Multiple R-squared:  0.6749,    Adjusted R-squared:  0.6723
F-statistic: 252.4 on 9 and 1094 DF,  p-value: < 2.2e-16
```

Model Interpretation

In our analysis, the relationship between temperature and total energy output presented intriguing findings. The model estimated a decrease of 1.529 units in total energy output for every one-degree Celsius increase in temperature, suggesting a negative correlation. However, the p-value for this relationship stood at 0.077, which is above our significance threshold of 0.05. This means we cannot confidently assert a direct relationship between temperature and energy usage based on our data. The squared temperature term, introduced to explore a potential non-linear relationship, indicated a marginal decrease of 0.002 in energy output for each unit increase in squared temperature, but this too lacked statistical significance.

Humidity, on the other hand, showed a more definitive relationship with energy usage. The model suggested that for every one percent increase in humidity, there was an estimated decrease of 2.398 units in total energy output. Given the p-value close to 0 and well below the 0.05 significance level, this relationship can be considered statistically significant. The interaction between humidity and temperature also appeared to be significant, with each unit increase estimated to result in a 0.06 unit increase in electricity output. Regarding wind speed, our model estimated that each meter per second increase would lead to a 1.02 unit decrease in energy output, a relationship that was statistically significant given its p-value of 0.0047. Similarly, each one-degree increase in wind direction was associated with a 0.043 unit increase in energy output, marking another significant finding.

Radiation types provided mixed results. Each one-watt per square meter increase in Global Horizontal Radiation (GHR) was projected to cause a 0.01 unit increase in energy output, yet this relationship did not reach statistical significance, with a p-value of 0.07138. In contrast, both Direct Normal Radiation (DNR) and Diffuse Horizontal Radiation (DHR) showed significant impacts. The model projected a 0.057 unit decrease in energy output for each one-watt per square meter increase in DNR and a 0.068 unit decrease for the same increase in DHR, with both relationships being statistically significant as their p-values were practically zero. Overall, the model's multiple R-squared value of 0.6749 suggests that approximately 67.49% of the variability in electricity usage was explained by these independent variables, highlighting the considerable influence of these environmental factors on energy consumption.

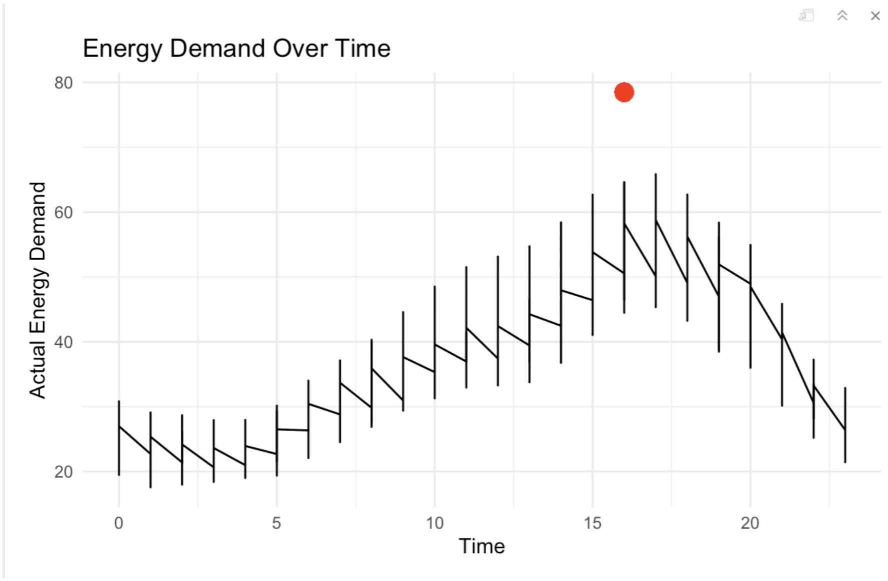
PREDICTIONS

Future Peak Energy Demand

The analysis of future peak energy demand is a critical component of proactive energy management, especially in the face of global warming and its potential to increase energy loads during hot summer months. In our study, we calculated the predicted electricity demand for July, identifying the maximum projected demand and the specific time it is expected to occur. This peak demand time, pinpointed to 16:00 (4 PM), is likely when energy usage will hit its highest point during the day. Understanding this peak is essential for eSC to strategize effectively in balancing the load and preventing blackouts.

The graph created to analyze current energy demands over time provides a clear visual representation of the patterns of electricity consumption throughout the day. The line plot, marked in black, indicates the trend of actual energy demand with time, while the red dot highlights the peak demand. This visualization serves a crucial function by enabling us to pinpoint the hours at which energy demand is the highest, thus allowing for targeted interventions aimed at energy conservation during these critical hours.

Knowing the time of peak demand is invaluable for eSC, as it can drive the implementation of demand response programs, targeted pricing strategies, and customer education on energy use. By anticipating when the demand is likely to spike, eSC can work with customers to shift or reduce load during these times, thereby mitigating the risk of overloading the grid without the need for costly infrastructure expansion. The findings from this analysis will be instrumental in guiding both short-term operational decisions and long-term planning for a more resilient and environmentally friendly energy system.

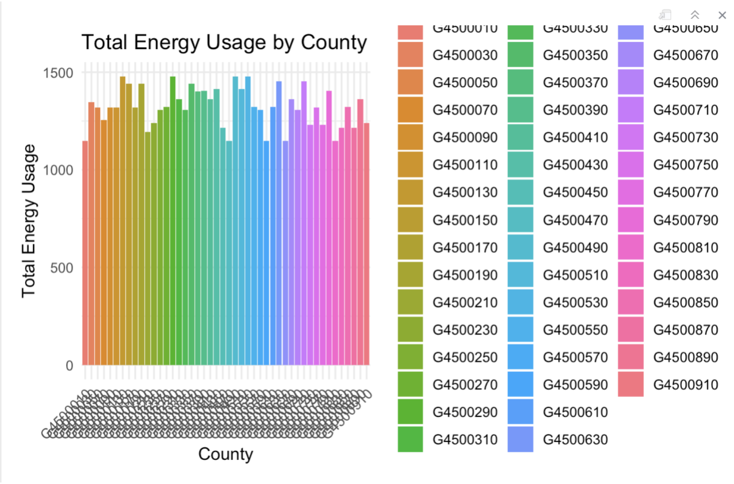


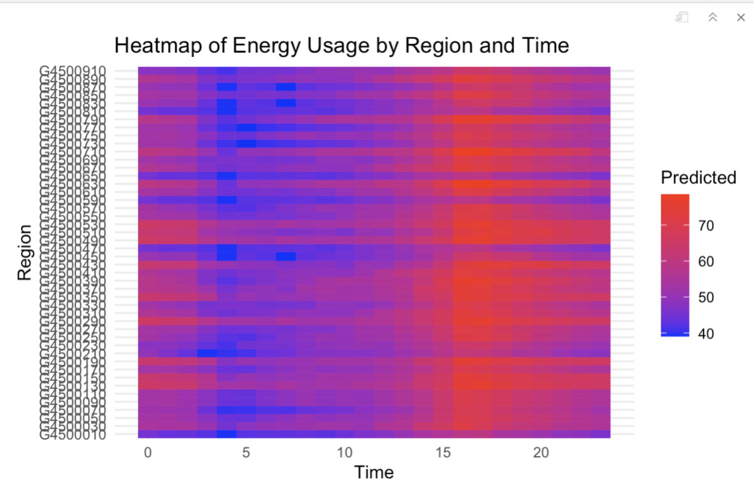
Analysis of Energy Usage

The visualizations at hand offer a striking analysis of energy usage across various counties and times, providing invaluable insights into consumption patterns. The bar graph titled "Total Energy Usage by County" presents a vibrant display of the energy consumption within each county. Each bar, distinguished by color, represents a unique county, with its height reflecting the total energy usage within that specific area. This graph not only serves as a tool for identifying the regions with the highest demand but also visually emphasizes the disparity in energy usage across different counties. Such disparities may be attributed to a range of factors including population density, industrial activity, and efficiency of buildings in terms of energy use.

Meanwhile, the heatmap titled "Heatmap of Energy Usage by Region and Time" paints a more dynamic picture, showcasing how energy consumption varies not only by region but also across different times of the day. The intensity of color corresponds to the level of predicted energy usage, with warmer colors indicating higher usage and cooler colors denoting lower usage. This heatmap is particularly telling as it highlights the times when energy demand peaks, which typically coincide with late afternoon and early evening hours. This pattern suggests a correlation with times when people are generally returning home from work or school, indicating a potential for peak load management and energy-saving opportunities.

Together, these graphs underscore the necessity for targeted energy management strategies that take into account regional and temporal variations in energy demand. By analyzing these patterns, energy providers can develop more effective demand response initiatives, tailored energy-saving programs, and better informed infrastructure development plans. The data-driven insights gleaned from these visualizations can also inform policy decisions, promote energy conservation measures, and ultimately contribute to the creation of a more sustainable and resilient energy system.



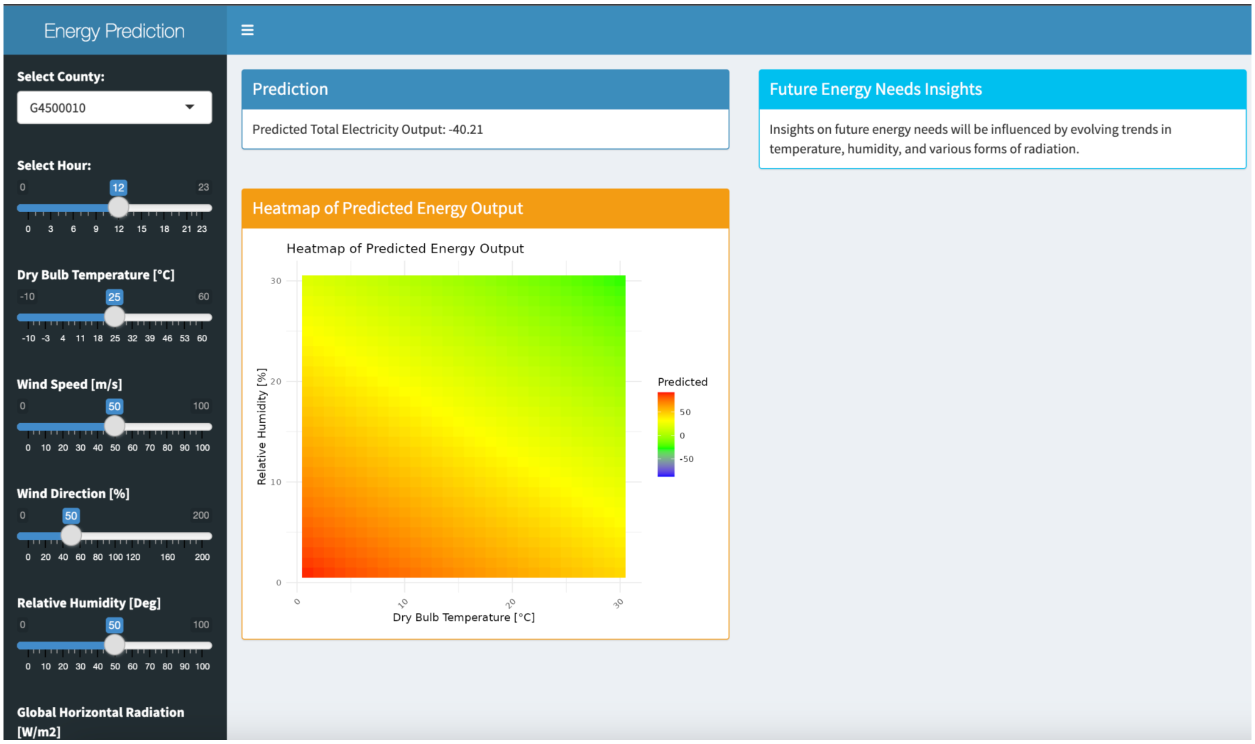


Visualisation & Reporting (SHINY)

URL for shiny app: https://lmqz952567.shinyapps.io/energypredict/

We have developed a user-friendly Shiny application that serves as an analytical tool to predict and visualize energy consumption. This application is tailored to provide custom forecasts of energy demand based on a variety of environmental inputs. Users can select a specific county and time of day, then adjust parameters like temperature, humidity, and wind conditions through an intuitive series of sliders and inputs. The application instantaneously processes these inputs to deliver a real-time prediction of total electricity output for the chosen settings.

The core feature of the application is its interactive heatmap, which vividly displays the relationship between temperature and humidity against the predicted energy output. This visual aid, with its gradient color scheme, makes it immediately apparent how slight variations in weather conditions could potentially impact energy demand across different regions. By offering this level of interactivity and real-time feedback, our Shiny app empowers users with the ability to explore and understand complex data relationships in an accessible format, facilitating informed decision-making for energy management and planning, for both those who live in a given area and those considering living in that area.



Energy Prediction

Select County: G4500010

Select Hour: 12

Dry Bulb Temperature [°C] 25

Wind Speed [m/s] 50

Wind Direction [%] 50

Relative Humidity [Deg] 50

Global Horizontal Radiation [W/m2]

Prediction

Predicted Total Electricity Output: -40.21

Future Energy Needs Insights

Insights on future energy needs will be influenced by evolving trends in temperature, humidity, and various forms of radiation.

Heatmap of Predicted Energy Output

ACTIONABLE INSIGHTS

Peak Demand Management: Our model can help in predicting peak energy demand periods, especially during summer months, due to the relationship between temperature and energy consumption. This insight is crucial for energy providers in planning and distributing resources efficiently. It also encourages homeowners and others to plan around peak energy usage time periods (example: doing activities outside the home) to minimize energy usage and costs.

Energy Efficiency Programs: By understanding the impact of various factors on energy consumption, utility companies can design targeted energy efficiency programs. For instance, in areas with high direct normal radiation, solar panels could be more effective. These programs may cost taxpayer money, so a solid campaign would be needed to convince the public, but with the right steps, there is definitely potential.

Customized Energy Solutions: The Shiny app developed by Mengqi allows clients to receive tailored energy consumption forecasts based on their specific county and adjusted weather conditions. This tool can be invaluable for both households and energy providers in planning and optimizing energy usage. As we mentioned in our presentation, while the linear model gives insights into the general usage, it is not really actionable. However, with the Shiny app, anyone can input their specifications, and have a much better sense of where they, in particular, stand.

Policy Recommendations: The insights from our model can inform policy-making, particularly in the areas of building codes (e.g., insulation standards) and urban planning (e.g., integrating green spaces to modulate urban microclimates). Putting people in positions of impact to incorporate these changes and whose platforms include energy consumption is paramount to making this into a reality.

Climate Adaptation Strategies: Our findings underline the need for adaptive strategies in the face of climate change, especially in regions prone to extreme temperature fluctuations. Many of the potential solutions are on a larger scale, but one can look to minimize their own energy output, particularly in ways that are harmful to the environment, to ensure that they are contributing. Similar to our policy recommendations, people in power being willing to take on these issues with realistic and proven-to-be-effective solutions plays a huge role in if this insight can be acted upon.

Consumer Awareness: The model's insights can be used to develop educational campaigns to raise awareness about the impact of weather and environmental factors on energy consumption, guiding consumers towards more sustainable energy usage patterns. Many people likely have no idea on how to think about energy usage, what their consumption looks like, and what consumption looks like for one area versus another. By being more aware of these things and more, they are much more likely to take action on the strategies recommended above, in large part due to the fact they will have a clear link between how their actions can affect their bills and the environment.

CONCLUSION

This study represents just one step in increasing awareness of energy usage and its factors. The integration of house-specific energy use data, weather variations, and diverse radiation measures into our predictive model has enabled us to not only model energy consumption but also offer strategic insights for energy management. These insights highlight the importance of this project, as two key issues today are money and climate change, both of which energy usage has a part in.

CONTRIBUTIONS

Data Identification and Cleaning - Vaishnavi Meka, Hayden Wasserman

Final Dataset Preparation - Abhi Chakraborty , Mengqi Li

Exploratory Analysis and Visualisation - Abhi Chakraborty, Mengqi Li

Building Initial Model - David Gold, Hayden Wasserman, Abhi Chakraborty

Accuracy Improvement & Results - Abhi Chakraborty

Model Interpretation - Abhi Chakraborty , Mengqi Li, Hayden Wasserman

Future Peak Energy Demand - David Gold, Mengqi Li

Analysis of Energy Usage - Abhi Chakraborty , Mengqi Li, Hayden Wasserman

Visualisation & Reporting (SHINY) - Mengqi Li

ACTIONABLE INSIGHTS & CONCLUSION - Abhi Chakraborty, Vaishnavi Meka, Hayden Wasserman**,** David Gold