**IST 687**

**Introduction To Data Science**

**Energy Consumption Analysis Report**

**Swetha, Arsh, Ruchir, Revanth**

**Prof. Christopher Dunham**

**&**

**Instructor: Akit Kumar**

**Introduction**

The primary objective of our study is to comprehensively understand the fundamental factors influencing energy consumption across various counties in South Carolina and North Carolina, specifically throughout the month of July. This period, marked by its peak summer temperatures, presents a unique opportunity to analyze patterns and drivers of energy usage in these regions.

Energy consumption is a critical issue that not only affects economic stability and environmental sustainability but also has profound implications for social equity. In light of this, our analysis employs a combination of Exploratory Data Analysis (EDA) and advanced modeling techniques to delve into the data, uncover significant relationships, and provide predictive insights that could inform future energy management strategies.

Our approach begins with a detailed data collection phase, where we gather comprehensive datasets from multiple trusted sources including local government databases and national energy reports. The datasets encompass a wide range of variables such as household income levels, building insulation quality, and historical weather conditions, which are crucial for understanding the nuances of energy usage in the studied areas.

Following data collection, we engage in rigorous data cleaning processes to ensure accuracy and relevance in our analysis. This involves removing redundant or irrelevant features, handling missing values, and ensuring that the data is appropriately scaled and normalized for the modeling phase.

The EDA phase is pivotal as it allows us to visualize the data and identify patterns or outliers that could skew our models or insights. During this phase, we generate a variety of visualizations such as heat maps, scatter plots, and bar charts that reveal the relationships between energy consumption and factors like economic status, weather conditions, and building characteristics.

With a solid understanding of the data, we proceed to the modeling stage, where we apply several statistical and machine learning models to predict energy consumption trends. These models help us simulate future scenarios and assess the impact of various factors on energy demand. For instance, we explore how a rise in temperature affects energy consumption and how this can vary across different income groups or housing types.

An essential part of our study is the focus on sustainability and cost-effectiveness. By identifying the key drivers of energy consumption, we aim to propose actionable solutions that can reduce the need for additional energy production facilities. These solutions may include promoting energy-efficient appliances, enhancing building insulation, or implementing renewable energy systems.

Moreover, our study is not just about reducing energy usage but also about promoting energy equity. By understanding how different factors affect energy consumption across various demographics, we can help formulate policies that ensure fair energy distribution and support for underprivileged communities.

In conclusion, our research is a comprehensive effort to decode the complex dynamics of energy consumption during a critical time of the year. Through meticulous data analysis and modeling, we aim to contribute valuable insights that could lead to more sustainable energy management practices in South Carolina and North Carolina, ultimately leading to a reduction in energy costs and an improvement in quality of life for all residents.

**Research Question Explanation**

This study is grounded in a critical examination of the dynamics between economic status and household energy consumption within the states of South Carolina and North Carolina. Our research question seeks to unravel the complex interactions between socio-economic factors and energy utilization, with an eye towards identifying sustainable solutions that can reduce energy demand and bridge the gap in energy inequity among different communities.

The premise of our inquiry is based on the hypothesis that economic status significantly influences a household's energy consumption patterns. This relationship is multifaceted, involving various elements such as the affordability of energy-efficient technologies, the quality of housing insulation, and general awareness of energy conservation practices. By exploring these factors, we aim to shed light on how disparities in income levels contribute to differences in energy usage and how these differences can be mitigated to promote a more equitable distribution of energy resources.

Our approach involves a detailed analysis of household energy consumption data, segmented by economic brackets. We will examine this data alongside variables that typically impact energy usage, such as the age and condition of housing, the presence of energy-efficient appliances, and the typical behaviors and lifestyle choices of residents. This analysis will help us understand the baseline energy consumption trends across different economic groups and the potential for reducing unnecessary energy expenditure through targeted interventions.

Furthermore, the study will explore the effectiveness of current energy efficiency and weatherization programs. We are particularly interested in how these programs are accessed by households of varying economic statuses and how their implementation can be optimized to extend their reach and effectiveness. This aspect of the research will involve gathering qualitative data from program participants and administrators, as well as analyzing program outcomes to assess their impact on reducing energy consumption.

In addition to analyzing existing data, our study aims to propose new strategies that could enhance energy sustainability across economically diverse communities. These strategies will focus on scalable solutions such as the widespread adoption of advanced insulation materials, the promotion of solar energy systems, and the implementation of smart energy management technologies. We will also consider policy recommendations that could support these initiatives, such as financial incentives for low-income households to upgrade their energy systems and educational campaigns aimed at raising awareness about energy conservation.

Ultimately, the goal of our research is to propose a comprehensive framework that not only addresses the immediate challenges of reducing household energy consumption but also contributes to the broader goal of energy sustainability. By linking economic status with energy usage and identifying actionable solutions, we aim to foster a more sustainable and equitable energy landscape in South Carolina and North Carolina.

**Overview**

This project examines the effects of rising temperatures on energy consumption, analyzing how various factors influence these changes.

**1. Data Cleaning and Exploratory Data Analysis (EDA)**

**Data Cleaning:**

Data cleaning is a critical first step in any data analysis process as it prepares the raw data for meaningful analysis. In your project, you've mentioned removing all unique values present in the columns. This typically involves eliminating columns that have a high degree of uniqueness which do not contribute to predictive modeling because they are either identifiers (like personal IDs, which are unique to each entry) or contain a high variance which might not be useful in analysis due to overfitting concerns.

Other common data cleaning steps might include:

- Removing or imputing missing values: Deciding how to handle missing data through imputation techniques or by omitting rows/columns with too many missing values.

- Standardizing formats: Ensuring that all data is in the same format or scale, which is crucial for subsequent analysis.

- Filtering out noise or errors: Identifying and correcting data points that do not make sense or are likely to be errors (e.g., ages that are too high to be plausible).

**Exploratory Data Analysis (EDA):**

EDA is the process of performing initial investigations on data to discover patterns, spot anomalies, test hypotheses, or check assumptions with the help of summary statistics and graphical representations. It is a practice aimed at understanding the dynamics underlying the data.

In the context of your project, EDA could involve:

- Visualizing the distribution of key variables to understand the central tendency and spread.

- Examining the relationships between variables using tools like scatter plots or correlation matrices.

- Identifying patterns or trends that could inform further analysis or feature engineering.

**2. Feature Engineering and Modeling**

Feature Engineering:

This step involves creating new variables from existing data to improve the predictive power of the machine learning algorithms. This could include:

- Aggregating data: Summarizing multiple data points into a single feature, like the average monthly energy consumption.

- Decomposing features: Breaking down data entries into simpler components, such as extracting parts of a date (year, month, day) or transforming a categorical variable into binary variables.

- Interactions between features: Creating new features by combining two or more variables, such as an interaction term between household size and income for predicting energy consumption.

**Modeling:**

Modeling involves selecting and applying appropriate algorithms to analyze data and make predictions. This includes:

- Choosing a model based on the hypothesis and the type of data. For instance, regression models for continuous outcomes or classification models for categorical outcomes.

- Training the model using historical data, with an aim to optimize its parameters for the best predictive performance.

- Validating the model using a separate validation set or using techniques like cross-validation to ensure that the model generalizes well to new data.

**3. Prediction and Impact Analysis**

**Prediction:**

In this phase, you use the trained model to make predictions about future or unseen data. This helps in making informed decisions based on those predictions.

**Impact Analysis:**

This involves assessing the effects and effectiveness of the predictions. For instance, understanding how changes in temperature could influence energy consumption and what this means for future energy requirements and conservation strategies. It often includes:

- Evaluating the potential savings from suggested changes.

- Analyzing the broader impact on communities, particularly focusing on different socioeconomic groups.

- Recommending policy or practical changes based on the model’s outcomes.

Each of these phases is crucial for a thorough analysis and effective application of findings in real-world scenarios, helping to ensure that the project not only addresses the initial research questions but also provides actionable insights and solutions.

**Data Cleaning:**

In our study, we initially sourced datasets from the United States Government, focusing on three key areas: Static House data, Energy Usage, and Weather Data. These datasets provided a comprehensive foundation for analyzing the intricacies of energy consumption across various metrics.

During the data cleaning phase, we meticulously extracted essential parameters from each dataset that were crucial for our analysis. From the Static House dataset, attributes such as building size (measured in square feet) and occupants' income were prioritized, as these factors are fundamental in determining energy consumption levels. This dataset was particularly expansive, containing over 170 columns. To refine our analysis and enhance the dataset's manageability, we eliminated columns with singular values that contributed minimal variance or insight into energy usage patterns. This step was crucial to focus our study on meaningful data without the clutter of redundant information.

Additionally, we specifically retained and focused on variables like square footage and insulation attributes of residences. These selections were informed by our research objectives, which aim to understand the relationship between household characteristics and energy consumption, particularly across different income levels. This streamlined approach allowed us to target our analysis more effectively, ensuring relevance and precision in our findings.

In our analysis of the Energy Usage dataset, we identified four key columns that are highly influential on energy consumption. These were then classified based on the climatic conditions they represented—specifically into hot and cold climate categories.

Particularly notable is the hot-humid climate zone, which we expect to show increased energy consumption. This is due to the typical weather conditions in these areas, such as higher temperatures and extended daylight hours, which naturally drive up energy usage, especially during peak summer months. A critical aspect of our study involves predicting future energy consumption under the assumption of a temperature increase by 5 degrees Celsius during the summer months. The focus on the hot-humid climate zone is directly aligned with this aspect of our research, providing a targeted framework for understanding and modeling energy consumption under exacerbated heat conditions.

Additionally, the Weather dataset has been integral to our approach. We are developing models that incorporate specific variables from the Energy Usage dataset to test their predictive accuracy and relevance. This modeling is essential for validating our hypotheses about climate impact on energy consumption and for enhancing the robustness of our predictions, ensuring they are well-grounded in empirical data and real-world scenarios.

**Data Preparation**

Initially, our strategy was to select the top three counties based on income diversity for a comprehensive study. However, after reconsidering the climatic factors, specifically the hot-humid conditions prevalent in our target region, we narrowed our focus to just two counties. This decision was influenced by the heightened energy usage typical of such climates, making them more relevant to our research objectives.

In these counties, we compiled a substantial dataset from numerous buildings, reflecting a broad spectrum of energy usage data. To streamline our analysis, we consolidated the energy data from all these buildings into a single data frame. This amalgamation was guided by previously determined parameters such as building size and insulation characteristics, ensuring consistency and relevance in our dataset.

Recognizing the need to explore energy consumption disparities more acutely, we decided to shift our analytical focus from building size to income levels. This adjustment aims to provide deeper insights into how economic factors influence energy demands across different demographics.

The refined data frame, enriched with targeted data on building energy use, will next be integrated with weather data from the same counties. This integration is crucial for our subsequent modeling efforts, as it allows us to assess the interplay between weather conditions and energy consumption, enhancing the predictive accuracy of our study.

**Climate Zone and Economic Diversity in South and North Carolina**

South Carolina and North Carolina are predominantly influenced by a subtropical climate, especially notable in the coastal and eastern regions. This climate is characterized by hot, humid summers and mild winters. The interior counties may experience slightly more variation with somewhat cooler temperatures due to higher elevations, particularly in the Appalachian Mountains in North Carolina.

The "hot-humid" climate zone, specifically, is critical to consider in energy consumption studies because it typically requires extensive cooling needs during long, sweltering summers. Increased daylight hours during these months further exacerbate energy demands as air conditioning systems run longer to maintain comfortable indoor environments. Additionally, occasional extreme weather events like hurricanes can impact energy infrastructure and usage patterns significantly.

This graph displays the total energy consumption for different income levels in the year 2020. The x-axis represents income levels, divided into categories such as "20000", "20000-39999", "40000-59999", and so on, with the last category being "150000+". The y-axis represents the total energy consumption.

A graph of a bar chart

Description automatically generated with medium confidence

**Economic Diversity**

Both states exhibit considerable economic diversity, which influences energy consumption patterns. This diversity is not only evident between rural and urban areas but also within regions varying in economic development levels. More affluent urban areas, such as Charlotte and Raleigh in North Carolina and Charleston in South Carolina, typically have higher energy consumption due to denser populations and greater commercial activity. However, they also tend to have better access to energy-efficient technologies and infrastructure.

In contrast, rural areas may face challenges like older housing stock and less investment in energy efficiency, which can lead to disproportionately higher energy costs relative to income. These areas often have lower average incomes and less economic activity, which impacts the ability to invest in modern, energy-efficient appliances and home improvements.

A graph of a number of bars

Description automatically generated

A graph of different colored lines

Description automatically generated

**Interplay Between Climate and Economy**

The interplay between climate conditions and economic diversity is crucial. In economically disadvantaged areas, particularly in hot-humid zones, residents may struggle with higher energy bills due to inefficient energy use and poor insulation.

This scenario underlines the importance of targeted energy efficiency programs and weatherization assistance to help mitigate energy inequity and reduce overall consumption.

Understanding this dynamic is vital for developing effective policies and interventions aimed at reducing energy consumption while ensuring that all communities, regardless of economic status, can achieve sustainable energy usage. Such strategies include incentivizing energy-efficient upgrades, enhancing public awareness about energy conservation, and implementing pricing strategies that encourage lower energy consumption without disproportionately impacting lower-income households.

**Models Used & Output**

**Linear Regression Summary**

Along with this, we did some cross-validation to conclude that the model would be overfitting on adding any more features and moved to a different approach. In our project, the linear regression model is applied to predict energy consumption (heating and cooling) based on independent variables such as climate zone, income level, and building size.

The model estimates the coefficients of the linear equation that best fits the relationship between the independent and dependent variables, allowing for predictions of energy consumption for new data points.

**Heating Model:- 94.5% , RMSE Value: 0.08, R-squared value: 0.945  
Cooling Model:- 98.8% , RMSE value: 0.04, R-squared: 0.988**

**Decision Tree using rpart package Summary.**

The decision rpart algorithm mitigates overfitting through built-in regularization techniques.   
It selects the best attribute to split the data based on criteria like information gain and creating decision paths represented by branches. Performance is evaluated using accuracy, RMSE, and R-squared values, akin to linear regression. Accuracy reflects the percentage of correctly classified instances, RMSE measures average deviation between predicted and observed values, and a higher R-squared value signifies better model explanatory power.

**Heating Model: 92.4%, RMSE Value: 0.09, R-squared value: 0.92   
Cooling Model: 95.44%, RMSE value: 0.08, R-squared: 0.954**

**Decision Tree using tree package Summary.**

It helps mitigate the problem of overfitting - because of its inbuilt regularization techniques.  
The tree package provides an alternative implementation of the decision tree algorithm, which may incorporate differences in splitting criteria, pruning techniques, or other algorithmic aspects. Nonetheless, the fundamental concept of recursively partitioning data based on significant attributes remains consistent. Like the r-part package, the decision model using the tree package is assessed for predictive performance using accuracy, RMSE, and R-squared values.

This are the values we got with the Decision model with tree package   
  
**Heating Model: 92.4% , RMSE Value: 0.09, R-squared value: 0.92   
  
Cooling Model : 95.4%, RMSE value: 0.08, R-squared: 0.954**

Both models produced comparable outcomes; therefore, we opted for the rpart package model to forecast our target variable.

**Energy Consumption by Income Level**

The graph categorizes households into six income brackets, ranging from less than $20,000 to more than $60,000 annually. For each income level, energy consumption is further divided based on ceiling insulation levels, from uninsulated to R-38, which indicates a high level of insulation.

From the graph, we can infer that households with better insulation levels (higher R-values) generally show a trend of lower energy consumption across most income brackets. This is consistent with the principle that well-insulated homes require less energy for heating and cooling, leading to reduced energy usage.

However, the pattern is not uniform across all income levels, indicating that factors other than insulation—perhaps those related to economic status—are influencing energy consumption. For instance, higher-income households may afford more energy-efficient appliances or may live in larger homes, both of which could affect energy usage regardless of insulation quality.

**Observations**

The highest energy consumption is in the less than $20,000 income bracket with poor insulation (uninsulated), which aligns with the project's focus on the interplay between economic status and energy efficiency.

Higher-income brackets appear to have a lesser variation in energy consumption between different insulation levels, suggesting that other energy-saving measures or lifestyle choices could be mitigating factors.

In the middle-income brackets ($20,000-$40,000 and $40,000-$60,000), there is a noticeable decrease in energy consumption with improved insulation, indicating the effectiveness of insulation upgrades.

The graph provides empirical support for the project’s premise that economic status and structural features of a home (like insulation) are significant determinants of energy consumption. The disparities observed suggest targeted interventions, such as subsidizing insulation improvements for low-income households, could be an effective strategy to address energy inequity.

**Feature Engineering and Modelling**

This bar graph presents energy consumption categorized by diverse income levels and the count of occupants.

**A graph showing the number of occupancy

Description automatically generated**

This bar graph illustrates energy consumption across varying income levels and the number of bedrooms.

**A graph showing a number of bedrooms and a number of bedrooms

Description automatically generated**

**Predicting Energy Usage if the summer was 5 degrees Warmer**

Enhanced insulation and a substantial PV system can significantly improve a building's energy efficiency, potentially reducing energy consumption despite rising temperatures.  
Occupants may adapt to warmer conditions by adopting more energy-efficient practices, such as using fans instead of air conditioning or adjusting thermostat settings.

Building characteristics, such as the number of bedrooms and occupants, offer insights into energy usage patterns. A well-insulated home with a large PV system and fewer occupants may consume less energy, even with more bedrooms.

The decrease in energy consumption despite warmer temperatures may be attributed to improved building insulation and the contribution of PV systems to energy generation. Changes in occupant behaviour or other energy-saving measures could offset the anticipated increase in energy usage from higher cooling demands, resulting in an overall reduction in net energy consumption.

**Impact of 5 degrees Increase in Temperature**

The average of energy consumption before temperature increase is 0.6892606 and average after 5-degree increase is 0.6801906

The percentage change of average energy consumption is 1.31% in next July

A graph of green and orange bars

Description automatically generated

We used the combined heating and cooling data for predicting the energy. First, we created a subset of data for July month and then used the model created for combined data using rpart() package for prediction. This graph shows the Total Energy Consumption for July 2018 versus the Predicted Energy Consumption for next year's July month after increasing 5 degrees in temperature.

**Providing Actionable insights into how to reduce energy Consumption.**

1. Energy Efficiency Incentives: Instituting a structured incentive program for households to adopt energy-efficient appliances and implement home energy-saving upgrades is crucial. These incentives should be designed with tiered benefits based on income levels, ensuring equitable participation and impact. By incentivizing energy efficiency, households across all income brackets can be encouraged to actively reduce their energy consumption, leading to long-term savings and environmental benefits.

2. Renewable Energy Solutions: Promoting the installation of renewable energy systems, such as solar panels, particularly in affluent households, can significantly contribute to reducing overall energy consumption. Offering financial incentives or tax credits for renewable energy investments can make these solutions more accessible and attractive to high-income households. This not only reduces reliance on traditional energy sources but also fosters a culture of sustainable energy practices in communities.

3. Awareness Campaigns: Developing and implementing educational campaigns targeting higher-income households is essential. These campaigns should focus on raising awareness about the environmental impact of excessive energy consumption and providing practical tips on how to reduce energy usage. By increasing awareness and understanding, households can make informed decisions to lower their energy consumption, leading to a more sustainable energy future.

4. Tiered Energy Pricing: Implementing a tiered energy pricing system can effectively incentivize households to reduce their energy consumption. By increasing the unit cost of energy with higher consumption levels, households are encouraged to adopt energy-saving measures to avoid higher costs. This approach not only reduces energy consumption but also promotes a more equitable distribution of energy costs among consumers.

In conclusion, a multifaceted approach that combines incentives, renewable energy solutions, awareness campaigns, and tiered pricing can effectively reduce energy consumption. By addressing energy usage patterns across different income brackets, these strategies can lead to a more sustainable and equitable energy future.

**Conclusion:**

In conclusion, the project focusing on energy consumption in South Carolina and North Carolina has revealed several key insights and strategies for reducing energy usage effectively. Through meticulous data cleaning, exploratory data analysis (EDA), and feature engineering, the project has identified significant factors influencing energy consumption, such as economic status, climate zones, and building characteristics.

The analysis has highlighted the importance of targeted interventions to promote energy efficiency, particularly in lower-income households. Strategies such as offering incentives for energy-efficient appliances, promoting renewable energy solutions, and implementing tiered energy pricing can play a crucial role in reducing energy consumption across all income brackets.

Furthermore, the project emphasizes the need for awareness campaigns to educate households, especially higher-income ones, about the environmental impact of excessive energy consumption and ways to reduce it. By encouraging behavioral changes and promoting sustainable energy practices, households can contribute to a more sustainable and equitable energy future.

Overall, the project underscores the importance of a comprehensive approach to energy management that considers economic, environmental, and social factors. By implementing the insights and strategies outlined in this project, policymakers and energy providers can work towards a more sustainable and energy-efficient future for South Carolina and North Carolina.

**References:**

[1] Debs L. and Metzinger J. (January 2022). *A Comparison of Energy Consumption in American Homes by Climate Region.* MDPI.

https://www.mdpi.com/2075-5309/12/1/82

[2] Osser R. and Kerrigan P. (February 2012). *Performance Evaluation of a Hot-Humid Climate Community.* NREL.

https://nrel.gov/docs/fy12osti/53183.pdf