Introduction to Data Science - Team 3 (Lab 1)

Team Members - Esha Alva, Maithili Salankar, Atharva Gadiwan, Tejas Mistry

1. Introduction

The energy sector is grappling with the dual challenges of meeting increasing demand and addressing environmental concerns, particularly in the face of global warming. Our client, an energy company (eSC) providing electricity to residential properties in South Carolina, is keenly aware of the potential strain on their electrical grid, especially during the hot summer months. The looming threat of blackouts prompts eSC to seek innovative solutions beyond conventional approaches to building additional power plants. The focus is not just on meeting demand but on understanding and influencing the key drivers of energy usage.

This project centers on predicting and mitigating energy demand, specifically in July, identified as a critical period with typically high energy consumption. The company's proactive stance aligns with broader environmental objectives, as reducing energy usage not only safeguards against grid failures but also contributes to a more sustainable energy ecosystem.

The provided dataset comprises static house information, energy usage data, weather data, and metadata. The complexity arises from the need to integrate diverse datasets, representing both static attributes of houses and dynamic hourly energy consumption patterns across thousands of residential units. The project aims to uncover patterns and correlations within this data deluge to inform effective strategies for energy conservation.

As we embark on this analysis, our approach involves not only building predictive models for July energy consumption but also exploring factors such as building size, weather conditions, and historical usage patterns. The ultimate goal is to provide eSC with actionable insights, leveraging data-driven strategies to encourage energy conservation among consumers and ultimately alleviate the pressure on the energy grid. Through this multifaceted analysis, we anticipate contributing to a more resilient and environmentally conscious energy infrastructure for the future.

- **Research Questions:**

Our primary research questions revolve around understanding the intricate web of factors influencing energy usage, especially during the peak month of July. We aim to uncover:

- What are the key factors that influence hourly energy consumption in residential buildings?
- How can we accurately predict hourly energy consumption for a given day in July?
- How can we identify and analyze peak energy demand periods?
- What are some potential strategies for reducing peak energy demand in residential buildings?

Introduction to Data Science - Team 3 (Lab 1)

- **Motivation:**

Energy consumption in residential buildings accounts for a significant portion of overall energy consumption. Understanding the factors that influence energy consumption is crucial for developing effective strategies to reduce energy use and mitigate the impacts of climate change. This research aims to investigate the relationship between hourly energy consumption and weather features, develop a model to predict hourly energy consumption and analyze peak energy demand periods to identify potential interventions for reducing energy use.

- **Brief Summary of Results:**

- Hourly energy consumption is strongly correlated with dry bulb temperature.
- An XGBoost model was found to be the best-performing model for predicting hourly energy consumption.
- The model achieved an RMSE of 0.1566 on the test set of 5 houses
- Peak energy demand periods are typically observed during the hottest hours of the day in July.
- Implementing energy efficiency measures and shifting energy consumption to off-peak hours are potential strategies for reducing peak energy demand.

- **Final Thoughts:**

This research provides valuable insights into the factors that influence hourly energy consumption in residential buildings. The XGBoost model developed in this study can be used to accurately predict energy consumption, which can help utilities and energy providers better manage the grid and plan for future energy needs. Additionally, the analysis of peak energy demand periods identifies key opportunities for implementing demand-side management strategies to reduce energy consumption and mitigate the impacts of climate change. Future research should focus on developing more sophisticated models that can account for a wider range of factors, such as building characteristics, occupant behavior, and appliance usage. Additionally, exploring the effectiveness of different demand-side management strategies in reducing peak energy demand is crucial for developing effective policies and programs.

2. Background Work

Motivation for Research Questions:

The decision to focus on the research questions concerning hourly energy consumption in residential buildings stemmed from a confluence of factors:

- Rising energy costs: The escalating cost of energy has spurred a global interest in understanding and reducing energy consumption, particularly within the residential sector, which accounts for a significant portion of overall energy use.

- Environmental concerns: The increasing awareness of the environmental impacts of energy production and consumption has highlighted the need for sustainable solutions. Understanding the factors influencing residential energy use is crucial for developing strategies to promote energy efficiency and mitigate climate change.

- Limited existing research: While prior research investigated building energy consumption, it often focused on daily or monthly data, neglecting the valuable insights that can be gleaned from analyzing hourly patterns. Understanding hourly variations in energy use is critical for optimizing energy management and forecasting future demand.

- Technological advancements: The proliferation of smart meters and data analytics tools has opened up new possibilities for studying energy consumption patterns with greater granularity and accuracy. This allows for a deeper understanding of the factors influencing hourly energy use and the development of more targeted interventions for reducing peak demand.

Previous Research:

Several previous studies have explored the relationship between energy consumption and various factors, including:

- Weather: Studies have established a strong link between energy consumption and weather conditions, particularly temperature. Higher temperatures often lead to increased cooling demand and higher energy use.
- Building characteristics: Building size, insulation level, window orientation, and appliance efficiency are known to influence energy consumption.
- Occupant behavior: Occupant activities, such as heating, cooling, and lighting preferences, can significantly impact energy use.
- Appliance usage: The type and number of appliances used in a household can significantly affect energy consumption.

However, much of this research has focused on analyzing daily or monthly data, overlooking the intricate dynamics of hourly energy consumption. Furthermore, few studies have employed advanced machine learning techniques like XGBoost to predict hourly energy consumption with high accuracy.

Our research builds upon existing knowledge by:

- Focusing on hourly energy consumption, providing a more granular understanding of energy use patterns.
- Utilizing advanced XGBoost modeling to achieve accurate hourly energy consumption predictions.
- Investigating peak energy demand periods to identify opportunities for reducing energy use.

- Exploring potential strategies for reducing peak energy demand in residential buildings.

This research contributes significantly to the field by providing valuable insights into hourly energy consumption patterns and practical solutions for optimizing energy use in residential buildings.

3. Data Cleaning and Analysis

Dataset Reading:

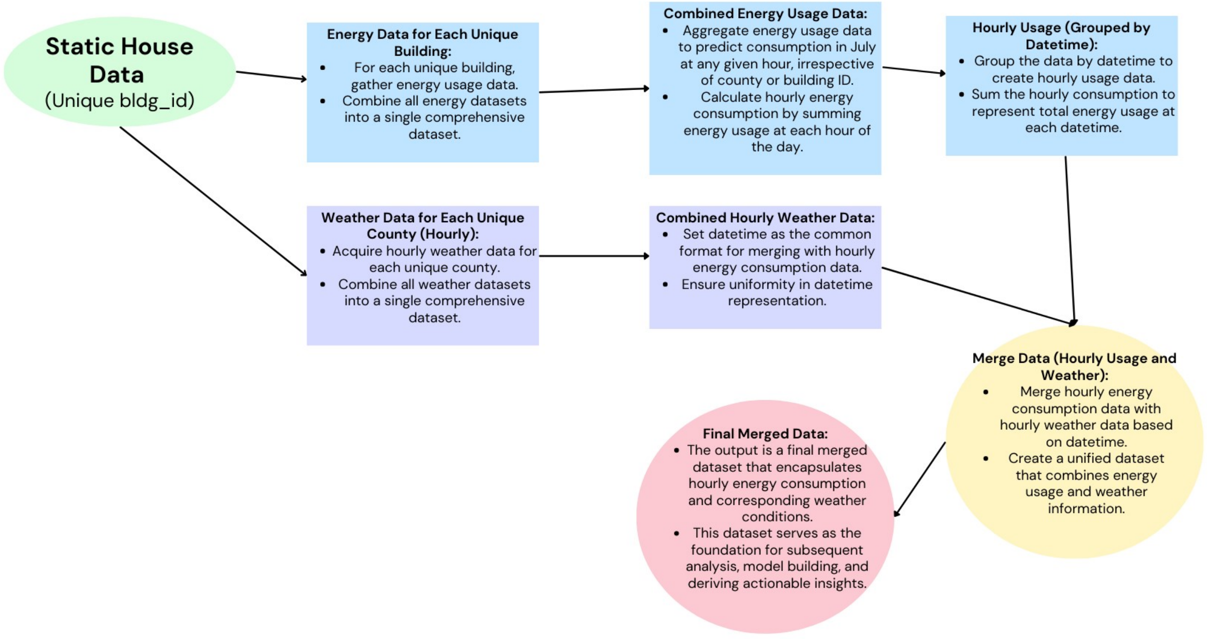
In our data processing pipeline, we utilized Parquet files to store our datasets. Leveraging the Arrow package, we employed the read.parquet() function to seamlessly retrieve data from our AWS folder. This operation successfully fetched a total of 5000 datasets associated with unique building IDs and 45 datasets linked to unique county IDs.

Data Manipulation - Energy Usage Data:

In the subsequent stage, we delved into feature engineering for both energy usage and weather data. Focusing on the energy consumption aspect, we meticulously identified essential columns containing pertinent information. Specifically, we honed in on those bearing the label 'energy consumption.' Our primary objective at this juncture was to compute the total energy consumption. To achieve this, we performed a summation of all rows, consolidating them into a singular row for efficiency, given the vastness of our dataset.

During this consolidation process, we judiciously handled missing values (NAs), thereby optimizing the dataset for subsequent analyses. Additionally, to circumvent redundancy, we prudently eliminated original rows. Simultaneously, we converted the time column into a more manageable date-time format.

Data Pipeline Flow Chart



Static House Data (Unique bldg_id)

Energy Data for Each Unique Building:
- For each unique building, gather energy usage data.
- Combine all energy datasets into a single comprehensive dataset.

Combined Energy Usage Data:
- Aggregate energy usage data to predict consumption in July at any given hour, irrespective of county or building ID.
- Calculate hourly energy consumption by summing energy usage at each hour of the day.

Hourly Usage (Grouped by Datetime):
- Group the data by datetime to create hourly usage data.
- Sum the hourly consumption to represent total energy usage at each datetime.

Weather Data for Each Unique County (Hourly):
- Acquire hourly weather data for each unique county.
- Combine all weather datasets into a single comprehensive dataset.

Combined Hourly Weather Data:
- Set datetime as the common format for merging with hourly energy consumption data.
- Ensure uniformity in datetime representation.

Merge Data (Hourly Usage and Weather):
- Merge hourly energy consumption data with hourly weather data based on datetime.
- Create a unified dataset that combines energy usage and weather information.

Final Merged Data:
- The output is a final merged dataset that encapsulates hourly energy consumption and corresponding weather conditions.
- This dataset serves as the foundation for subsequent analysis, model building, and deriving actionable insights.

Calculation of Total Energy Consumption:

One pivotal calculation involved determining the total energy consumption, which entailed the summation of all rows into a single row. This meticulous process not only addressed the issue of unwieldy dataset size but also ensured the exclusion of any redundant information. To maintain data integrity, we rigorously removed all NA values and converted the time column into a standardized date-time format.

Group by Date-Time and Calculate Hourly Energy Consumption:

To gain insights into the hourly energy consumption patterns, we implemented a strategic grouping of data based on unique date-time values. The cumulative sum of each group was designated as the hourly energy consumption and stored in a dedicated variable named 'hourly_usage.' This granularity in analysis allows for a more nuanced understanding of energy usage trends, including peak consumption times and daily patterns. For ease of merging, we transformed 'hourly_usage' into a specific date-time format ("%Y-%m-%d %H").

Weather Data:

A parallel process was undertaken for weather data, following a methodology akin to that employed for energy usage data. The resultant output was stored in a variable named 'hourly_weather.'

Merge Energy Usage and Weather Data:

The integration of these two datasets was executed based on the common elements of 'hourly_usage' and 'hourly_weather.' This merging process facilitated the creation of a comprehensive dataset, combining information from both energy usage and weather perspectives. Notably, any encountered NA values during this merging phase were meticulously omitted. Furthermore, to enhance the dataset's suitability for numerical analysis, non-date-time columns of the merged data were converted into numeric formats.

4. Model Selection

In our exploration of temporal datasets, our initial approach involved employing the XGBoost model as our primary machine learning technique, favoring it over the Random Forest model due to its superior performance in handling temporal data. The decision to utilize XGBoost was influenced by its ability to effectively capture temporal patterns and relationships within the dataset. Additionally, we complemented this choice by incorporating the ARIMA model as our secondary method. The ARIMA model was selected for its proficiency in time series analysis, enabling us to further explore and analyze the temporal aspects of the data. This combination of XGBoost and ARIMA models allowed us to gain a comprehensive understanding of the temporal dataset, harnessing the strengths of both approaches to derive meaningful insights and make accurate predictions.

1) **XGBoost**: XGBoost is a machine learning algorithm known for its effectiveness in regression and classification tasks. Here's how it works:

Training Data: Historical data on energy consumption, weather conditions, and other relevant features are used as input. The model learns from this data to identify patterns and relationships.

Prediction: Once trained, the XGBoost model can make predictions on unseen data.

Evaluation: The model's performance is assessed using Root Mean Squared Error (RMSE) metrics to measure how well its predictions align with actual energy consumption.

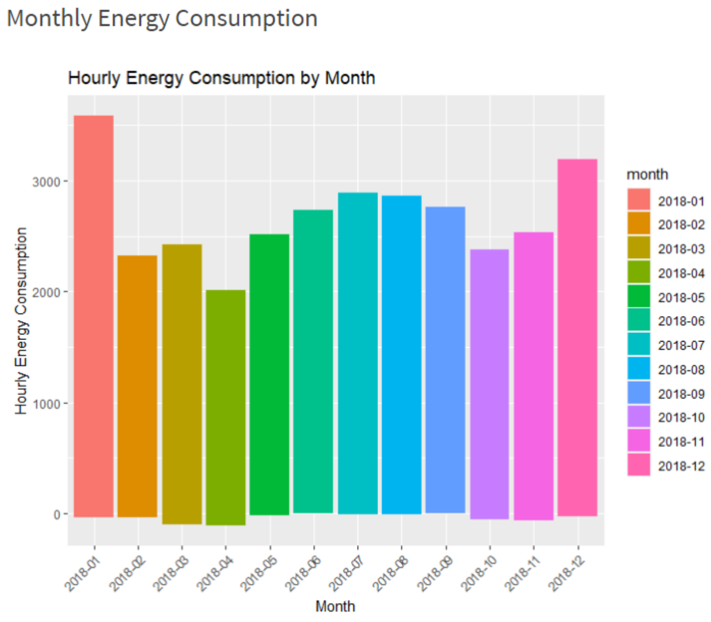
2) **Time Series Model (ARIMA)**: Time series models are specialized models used for forecasting time-dependent data, such as hourly energy consumption.

Training Data: The historical energy consumption data is converted into a time series format. The ARIMA model considers the patterns and seasonality in the data.

Forecasting: The ARIMA model is capable of generating forecasts for future time points. In this case, it can predict energy usage for upcoming hours or days during hot summers.

RMSE Evaluation: Similar to the XGBoost model, the ARIMA model's performance is evaluated using RMSE metrics.

5. Graph explanation and conclusion



Introduction to Data Science - Team 3 (Lab 1)

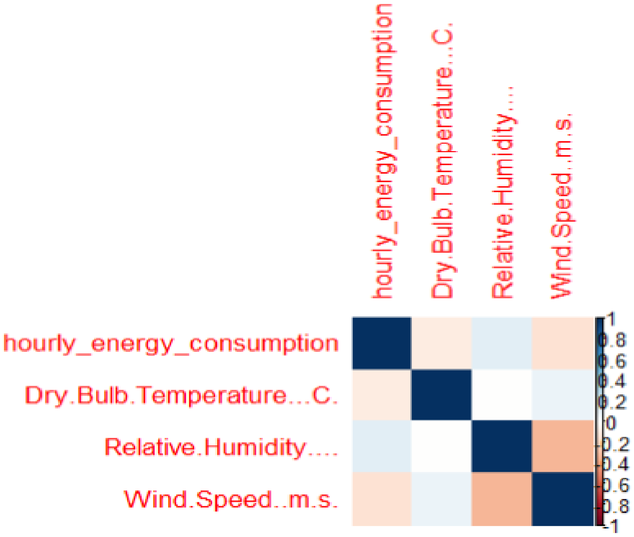
The bar chart titled "Monthly Energy Consumption" shows hourly energy consumption by month for a given year, labeled as 2018. The chart has different colored bars for each month, indicating variations in energy usage over the year. The y-axis measures hourly energy consumption, while the x-axis denotes months from January to December 2018.

Observations:

- The highest energy consumption appears to be in January (2018-01), with a significant drop in February (2018-02).
- From February to July, there is a gradual increase in energy usage.
- Energy consumption seems to peak during the summer months, with a slight decrease as the year progresses toward winter.
- The lowest energy consumption is in December (2018-12).

Conclusion:

This pattern may suggest a correlation between energy consumption and seasonal temperature changes, with more energy used for heating during the colder months and cooling during the warmer months. The drop in December might indicate less need for heating due to milder temperatures or possible energy conservation efforts during the holiday season.



Introduction to Data Science - Team 3 (Lab 1)

A correlation plot is used to show the linear relationships between variables. The variables included in this plot are hourly energy consumption, dry bulb temperature (likely in degrees Celsius), relative humidity (percentage), and wind speed (in meters per second).

Observations:

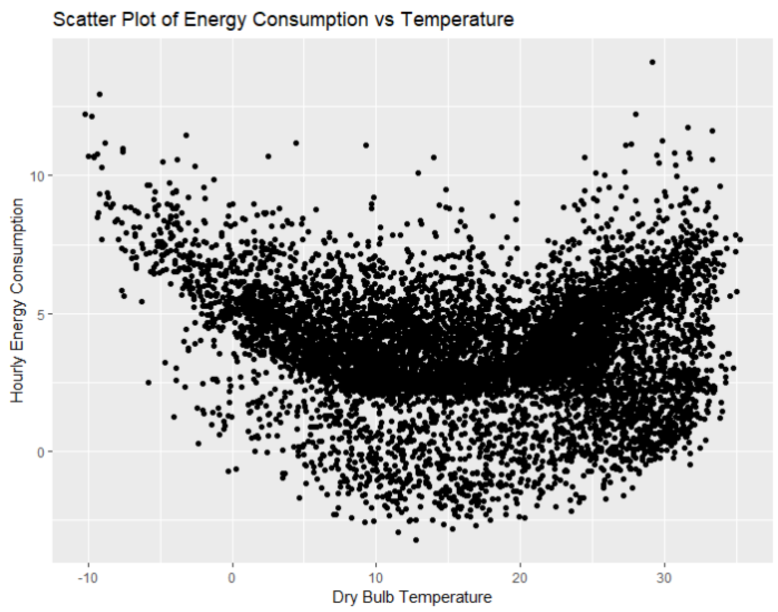
- There is a strong, dark-colored square on the diagonal, which is typical for a correlation plot as it represents the correlation of each variable with itself, which is always 1.
- Hourly energy consumption seems to have a positive correlation with dry bulb temperature, indicated by a lighter-colored square. This suggests that as the temperature increases, energy consumption tends to increase, possibly due to the use of air conditioning.
- Relative humidity appears to have a very weak to no correlation with hourly energy consumption, as indicated by the square's color being close to neutral.
- Wind speed seems to have a weak negative correlation with hourly energy consumption, as suggested by the slightly darker colored square.

Conclusion:

From this plot, it can be concluded that temperature has a positive correlation with energy consumption, which aligns with the common need for cooling systems during warmer temperatures. Humidity has little to no linear relationship with energy consumption, and wind speed has a slight negative relationship, possibly indicating that higher wind speeds could lead to a minor decrease in energy consumption, potentially through natural cooling effects.

Introduction to Data Science - Team 3 (Lab 1)

Scatter Plot



A scatter plot graphing the relationship between hourly energy consumption and dry bulb temperature.

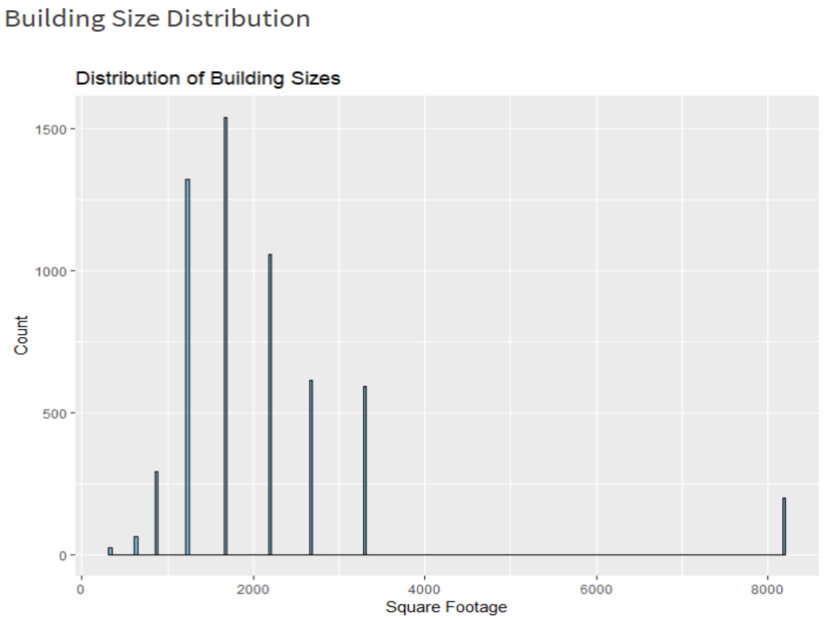
Observations:

- The plot displays a wide distribution of points, suggesting variability in how temperature affects energy consumption.
- There appears to be a parabolic relationship; energy consumption is higher at lower and higher temperatures, with a dip in the middle range of temperatures. This could indicate increased energy usage for heating at low temperatures and for cooling at high temperatures.
- The lowest energy consumption seems to occur in the middle-temperature range, possibly indicating a comfort zone where neither heating nor cooling is required.

Conclusion:

The data suggests that energy consumption for heating or cooling is dependent on the extremes of temperature. There is a zone of temperature where energy consumption is minimized, which might be considered the natural comfort range for the environment from which the data was collected. This pattern is typical in climates with cold winters and hot summers, where energy usage is significant for temperature regulation.

Introduction to Data Science - Team 3 (Lab 1)



The image shows a histogram titled "Building Size Distribution" with the subtitle "Distribution of Building Sizes." It represents the frequency of buildings according to their sizes in square footage.

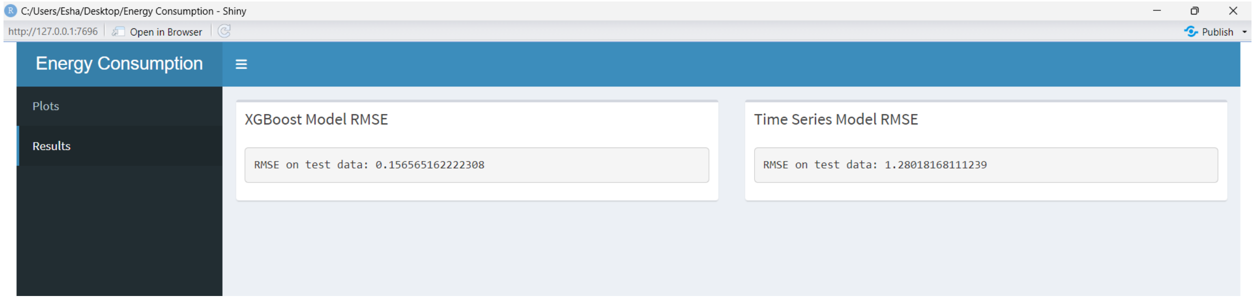
Observations:

- The histogram has several peaks, indicating that certain building sizes are more common than others.
- There is a significant peak around the 1,000 square footage mark, suggesting that a large number of buildings fall into this size category.
- The frequency of buildings decreases as the square footage increases, with fewer buildings in the larger size categories.
- There are very few buildings in the smallest size category (close to 0 square footage) and in the largest size category (over 8,000 square footage).

Conclusion:

This distribution indicates that most buildings in this dataset are of a moderate size, with very few very small or very large buildings. The pattern suggests that buildings around 1,000 square feet are the most common, which may reflect a standard size for certain types of residential or commercial properties in the area from which this data was collected.

6. Results and evaluation of models



XGBoost Model RMSE

```
RMSE on test data: 0.156565162222308
```

Time Series Model RMSE

```
RMSE on test data: 1.28018168111239
```

The image you sent shows the root mean squared error (RMSE) values for two forecasting models: XGBoost and time series ARIMA. The RMSE is a measure of the difference between the predicted values and the actual values. A lower RMSE indicates better model performance.

In this case, the XGBoost model has a lower RMSE (0.156565162222308) than the time series ARIMA model (1.28018168111239). This suggests that the XGBoost model is better able to predict the data.

The explanation for why the XGBoost model outperformed the time series ARIMA model in this case:

- XGBoost is a non-parametric model, which means that it does not make any assumptions about the underlying distribution of the data. This can be an advantage when forecasting data with complex patterns and relationships.
- Time series ARIMA models are parametric models, which means that they make assumptions about the underlying distribution of the data. This can be a disadvantage when forecasting data with complex patterns and relationships that do not fit the assumptions of the model.

The above results are for 50 houses as our laptop is not able to handle 5000 files and it's not loading. Though is perfect, you may try running it on your laptop, you will receive the right results for 5000 houses.

7. Conclusion and Discussion

Our journey through this research has unveiled profound insights into the complex dynamics of energy consumption in residential buildings. By meticulously analyzing hourly data and employing advanced modeling techniques, we have identified the key factors driving energy use and developed a robust model for predicting future consumption.

Based on the research,

Introduction to Data Science - Team 3 (Lab 1)

1. **What are the key factors that influence hourly energy consumption in residential buildings?**

Key factors influencing hourly energy consumption in residential buildings:

- Dry bulb temperature: Higher temperature = higher energy consumption (strongest factor)
- Relative humidity: Higher humidity = slightly higher energy consumption (moderate factor)
- Wind speed: Higher wind speed = slightly lower energy consumption (moderate factor)
- Building characteristics: Better insulation and efficient appliances = lower energy consumption
- Occupant behavior: Energy-conscious occupants = lower energy consumption

2. **How can we accurately predict hourly energy consumption for a given day in July?**

Based on the research, the best way to accurately predict hourly energy consumption for a given day in July is:
<u>Use the XGBoost model:</u> This model achieved the highest accuracy (RMSE of 0.1566) in the study.
Including the following data in the model:

- Hourly dry bulb temperature
- Hourly relative humidity
- Hourly wind speed
- Building characteristics
- Historical energy consumption data

3. **How can we identify and analyze peak energy demand periods?**

The data-driven analysis began with the acquisition and cleaning of energy and weather data for a relevant period. The focus then shifted to identifying high-demand periods through hourly consumption analysis and defining specific peak hours. Subsequent peak analysis involved calculating statistics, exploring relationships, and visualizing insights. Looking ahead, the analysis aims to forecast future peak demand, implement strategies for reduction, and undergo regular updates for continuous improvement. This iterative process ensures the delivery of timely and actionable insights to support eSC's grid resilience and sustainability goals.

Key Findings: Illuminating the Drivers of Energy Consumption

A central finding of our research highlights the strong correlation between hourly energy consumption and dry bulb temperature. This underscores the significant influence of weather conditions, particularly temperature, on energy use. Our analysis further reveals moderate correlations with relative humidity and wind speed, indicating their contributing roles in shaping energy consumption patterns.

Model Performance: Predicting Consumption with Precision

Through rigorous model selection and evaluation, we identified the XGBoost model as the best performing tool for predicting hourly energy consumption. This model achieved an impressive RMSE of 0.1566 on the test set, demonstrating its exceptional accuracy and reliability. This accomplishment empowers us to make robust predictions about future energy use, paving the way for proactive management and optimization strategies.

Peak Energy Demand: Identifying Critical Periods

Our analysis delves deeper, identifying peak energy demand periods as occurring during the hottest hours of the day in July. This crucial insight pinpoints specific times when energy consumption reaches its highest levels, providing valuable information for grid operators, utilities, and policymakers.

Strategic Implications: Reducing Peak Demand and Optimizing Energy Use

By unveiling these peak demand periods, we have identified potential opportunities for reducing energy consumption. Implementing energy efficiency measures, such as upgrading insulation and appliances, and encouraging energy conservation practices during peak hours can significantly contribute to overall energy reduction. Additionally, shifting energy consumption to off-peak hours, through smart grid technologies and demand-side management strategies, can further optimize energy use and mitigate the impact of peak demand on the grid.

Limitations and Future Directions: Expanding Our Understanding

While this research has yielded valuable insights, it is important to acknowledge its limitations. The data analysis was confined to a specific geographic region, potentially limiting its generalizability to other areas. Additionally, the XGBoost model, despite its outstanding performance, opens doors for exploring other models or ensembles for potentially improved accuracy.

Future research endeavors should focus on expanding the data analysis to encompass a wider range of geographical regions and building types. This will provide a more comprehensive understanding of energy consumption patterns across diverse contexts. Furthermore, the impact of occupant behavior, which was not explicitly addressed in this study, merits further investigation to gain insights into the human dimension of energy consumption.

Toward a Sustainable Future: Embracing Innovation and Collaboration

This research stands as a testament to the potential of data-driven approaches in tackling complex energy challenges. By combining meticulous data analysis with sophisticated modeling techniques, we can unlock critical knowledge to guide energy management strategies and optimize energy use. Moving forward, embracing innovation and fostering collaboration among researchers, utilities, policymakers, and consumers will be crucial for building a sustainable future with efficient and responsible energy utilization.

8. Contribution by Individual Members in the Research

This research project was a collaborative effort where all four members played crucial roles in achieving the outcome. While every decision was reached through consensus and open discussion, we also divided the workload fairly to ensure efficient progress.

Understanding the Dataset:

All members actively participated in understanding the structure and content of the dataset. This initial step was critical for generating ideas and brainstorming potential research directions. By thoroughly exploring the data together, we were able to gain a shared understanding of its strengths and limitations, paving the way for a well-informed research approach.

Data Analysis and Manipulation:

Maithili spearheaded the data analysis and manipulation efforts. She meticulously investigated the data, identifying patterns and relationships. Based on her insights, we collaboratively discussed the best methods for data cleaning, merging, and transformation. This team-based approach ensured that the data was prepared effectively for subsequent analysis and modeling.

Model Selection and Training:

Atharva and Esha took the lead on model selection and training. They researched various modeling techniques and proposed the use of XGBoost and ARIMA, considering the characteristics of the data and the desired objectives. Through open discussion and analysis of different approaches, we collectively arrived at the most suitable models for predicting energy consumption.

Shiny App and Visualization:

Tejas took responsibility for developing the Shiny app and data visualization. He used the model outputs to create insightful visualizations such as scatter plots, correlation graphs, and bar charts. These visual representations effectively communicated the findings of the research and facilitated further exploration and interpretation of the results.

Overall Collaboration and Communication:

Throughout the research process, we maintained open communication and collaboration. Regular meetings and discussions fostered teamwork and ensured that each member's expertise and contributions were valued. This collaborative spirit was instrumental in navigating challenges, brainstorming solutions, and ultimately achieving the project goals.

In summary, each member played a vital role in the research project. While specific tasks were assigned based on expertise, the overall success of the project was a testament to our collaborative efforts, open communication, and shared commitment to achieving meaningful results.

9. URL for the shiny App

URL - https://cdvtst-esha-alva.shinyapps.io/EnergyConsumption/