IST 687 final project technical report

Group 5

1. Introduction

This report undertakes a comprehensive analysis focused on managing and predicting energy consumption, with a primary emphasis on forecasting the impact of increasing temperatures on electricity demand, specifically during the peak month of July. The main objective is to proactively address potential energy shortages and blackouts that may arise due to an overburdened electrical grid during exceptionally hot summer conditions.

To tackle this challenge, we have adopted a multi-faceted approach, utilizing a rich dataset that includes static house information, detailed hourly energy usage, and corresponding weather data. By engaging in meticulous data preparation, exploratory analysis, and advanced modeling techniques, our goal is to uncover the primary drivers of energy consumption. This understanding will empower us to propose data-driven strategies not only for predicting future energy demand in high-temperature scenarios but also for exploring opportunities to conserve energy.

- **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

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 Linear model 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 Linear 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.

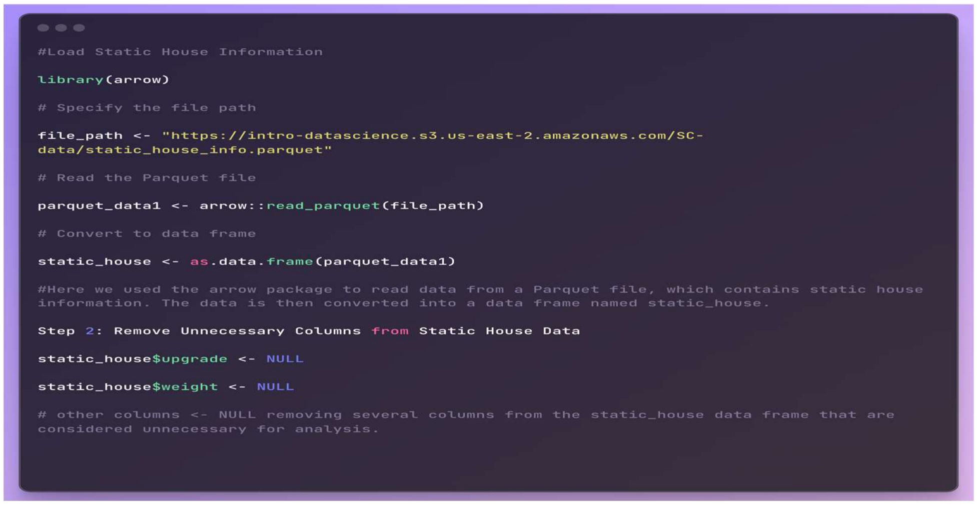
3. Data reading and merging approach

In the data preparation phase, we focus on consolidating and preparing three key datasets for our analysis of household energy consumption. These datasets include house static data, electricity consumption data, and weather data.

House Static Data Loading

We begin by loading the static information about the houses. This dataset, obtained in Parquet format, contains essential identifiers such as building IDs (bldg_id) and county information. These identifiers are crucial as they link the static house data with the dynamic electricity consumption and weather data.

In the code snippet as shown below, we are utilizing the Arrow package in R to load data from a Parquet file containing static house information. The specified file path points to the location of the Parquet file on Amazon S3.

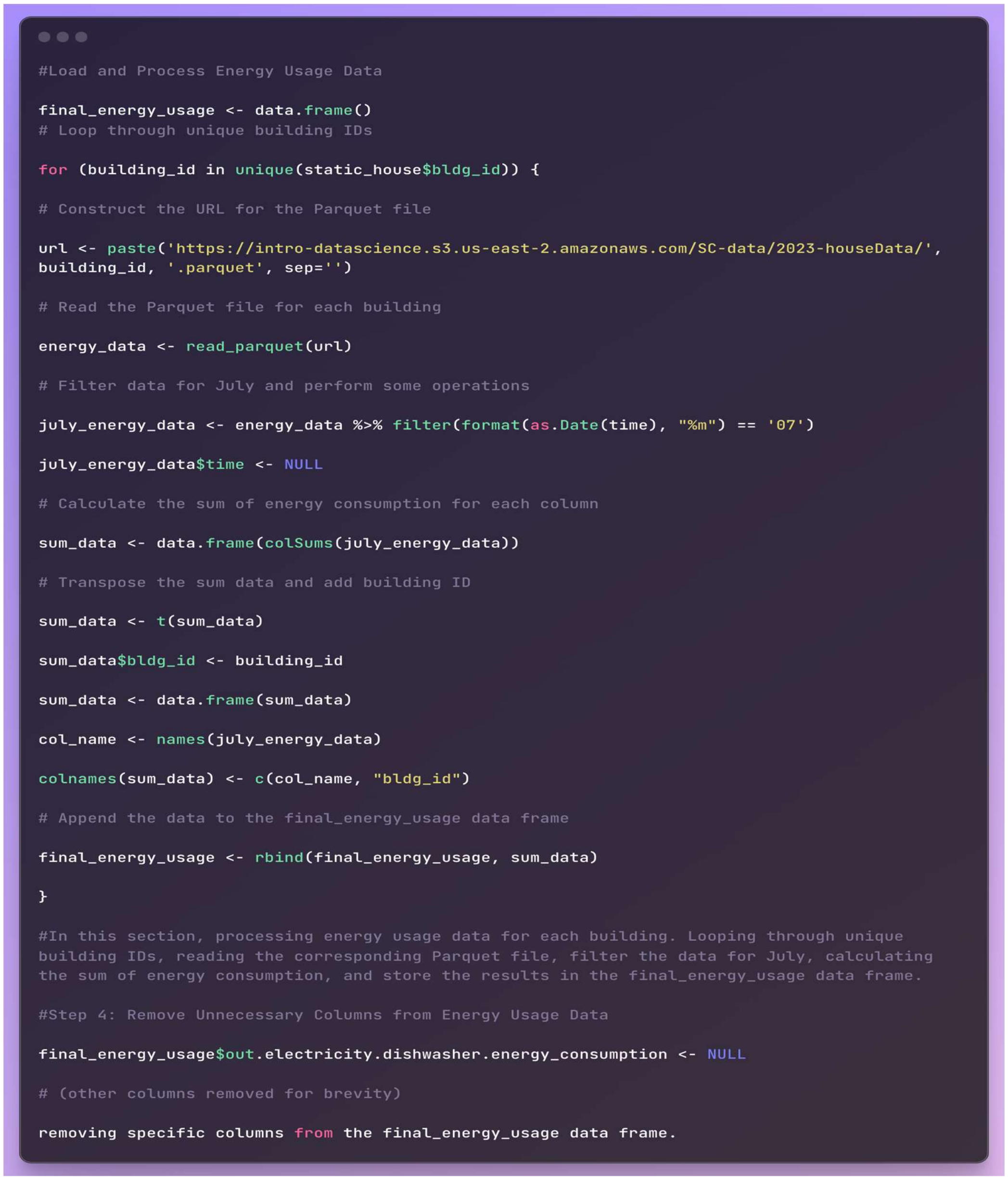


```
#Load Static House Information

library(arrow)

# Specify the file path

file_path <- "https://intro-datascience.s3.us-east-2.amazonaws.com/SC-
data/static_house_info.parquet"

# Read the Parquet file

parquet_data1 <- arrow::read_parquet(file_path)

# Convert to data frame

static_house <- as.data.frame(parquet_data1)

#Here we used the arrow package to read data from a Parquet file, which contains static house
information. The data is then converted into a data frame named static_house.

Step 2: Remove Unnecessary Columns from Static House Data

static_house$upgrade <- NULL

static_house$weight <- NULL

# other columns <- NULL removing several columns from the static_house data frame that are
considered unnecessary for analysis.
```

Electricity Consumption Pipeline

The electricity consumption data is processed in a pipeline designed to handle data for each building individually. This involves iterating over each building ID, fetching electricity consumption data from corresponding Parquet files. The pipeline includes error handling to manage potential issues in data loading. For each building, we calculate the total electricity consumption and retain essential timestamps.

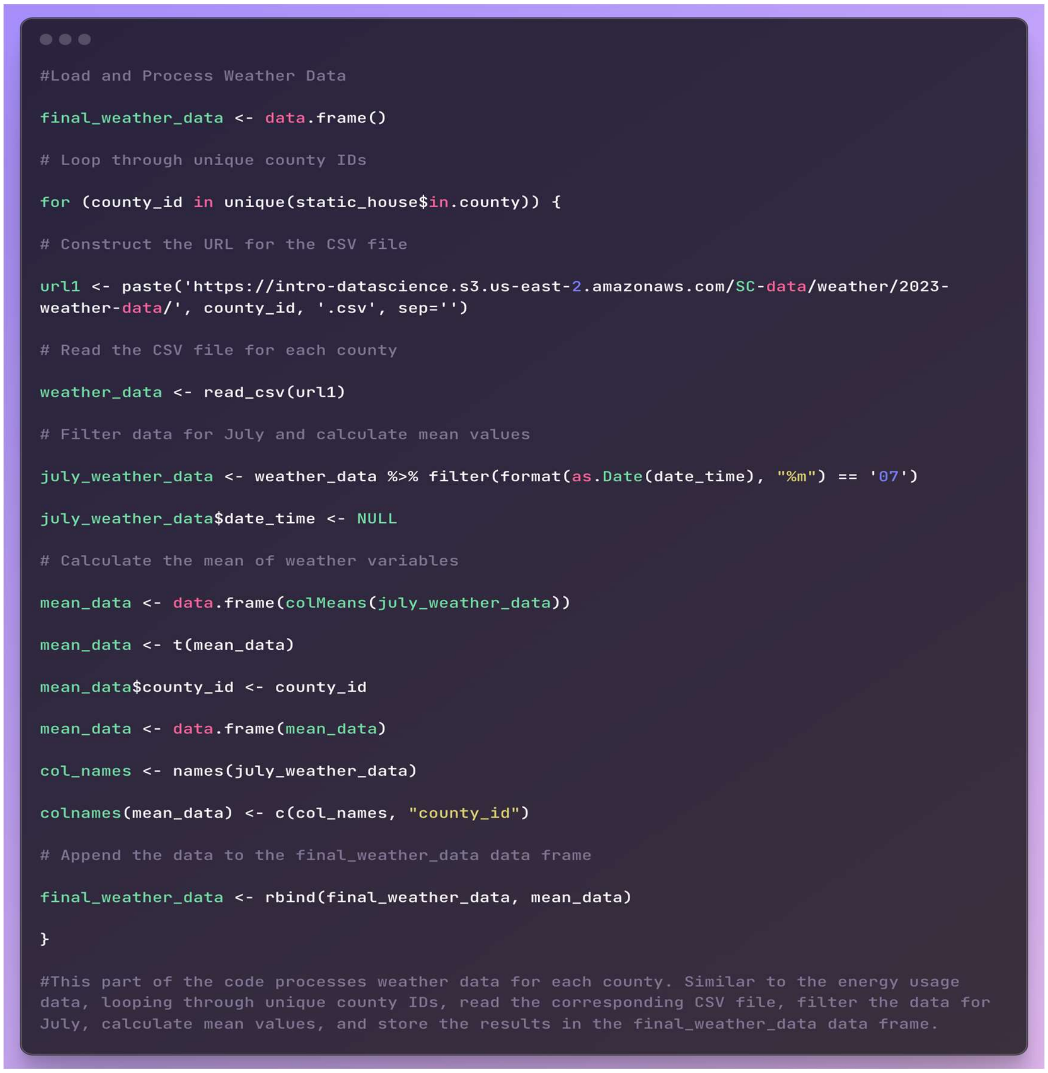
This systematic process ensures that the final_energy_usage dataframe contains consolidated and relevant information for further analysis, with each row representing a unique building and its corresponding energy consumption characteristics during the month of July.



```
#Load and Process Energy Usage Data

final_energy_usage <- data.frame()
# Loop through unique building IDs

for (building_id in unique(static_house$bldg_id)) {

# Construct the URL for the Parquet file

url <- paste('https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/2023-houseData/',
building_id, '.parquet', sep='')

# Read the Parquet file for each building

energy_data <- read_parquet(url)

# Filter data for July and perform some operations

july_energy_data <- energy_data %>% filter(format(as.Date(time), "%m") == '07')

july_energy_data$time <- NULL

# Calculate the sum of energy consumption for each column

sum_data <- data.frame(colSums(july_energy_data))

# Transpose the sum data and add building ID

sum_data <- t(sum_data)

sum_data$bldg_id <- building_id

sum_data <- data.frame(sum_data)

col_name <- names(july_energy_data)

colnames(sum_data) <- c(col_name, "bldg_id")

# Append the data to the final_energy_usage data frame

final_energy_usage <- rbind(final_energy_usage, sum_data)

}

#In this section, processing energy usage data for each building. Looping through unique
building IDs, reading the corresponding Parquet file, filter the data for July, calculating
the sum of energy consumption, and store the results in the final_energy_usage data frame.

#Step 4: Remove Unnecessary Columns from Energy Usage Data

final_energy_usage$out.electricity.dishwasher.energy_consumption <- NULL

# (other columns removed for brevity)

removing specific columns from the final_energy_usage data frame.
```

Weather Data Pipeline

Parallel to processing electricity data, we also manage weather data, specifically focusing on temperature metrics across different counties. The weather data, sourced in CSV format, includes detailed temperature readings. We process this data by iterating over each county, aligning the weather data with our other datasets based on the county identifier. This step is crucial for analyzing the impact of weather on electricity consumption patterns.



```
#Load and Process Weather Data

final_weather_data <- data.frame()

# Loop through unique county IDs

for (county_id in unique(static_house$in.county)) {

# Construct the URL for the CSV file

url1 <- paste('https://intro-datascience.s3.us-east-2.amazonaws.com/SC-data/weather/2023-
weather-data/', county_id, '.csv', sep='')

# Read the CSV file for each county

weather_data <- read_csv(url1)

# Filter data for July and calculate mean values

july_weather_data <- weather_data %>% filter(format(as.Date(date_time), "%m") == '07')

july_weather_data$date_time <- NULL

# Calculate the mean of weather variables

mean_data <- data.frame(colMeans(july_weather_data))

mean_data <- t(mean_data)

mean_data$county_id <- county_id

mean_data <- data.frame(mean_data)

col_names <- names(july_weather_data)

colnames(mean_data) <- c(col_names, "county_id")

# Append the data to the final_weather_data data frame

final_weather_data <- rbind(final_weather_data, mean_data)

}

#This part of the code processes weather data for each county. Similar to the energy usage
data, looping through unique county IDs, read the corresponding CSV file, filter the data for
July, calculate mean values, and store the results in the final_weather_data data frame.
```

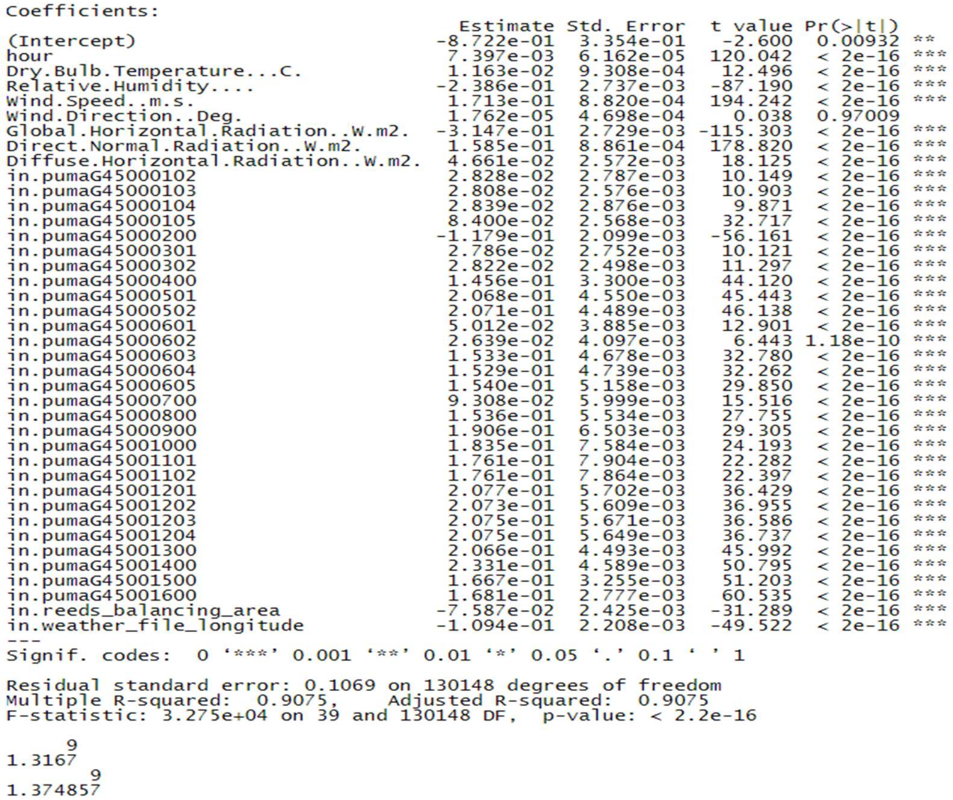
4. Model Selection

We changed the way we look at energy consumption data by using logarithms. This was important because our initial observations showed that the relationship between energy use and temperature is exponential. By using logarithms, we make it easier to understand this relationship as linear and also allow the model's coefficients to be interpreted in terms of percentage changes. This transformation helps us measure the percentage increase in energy consumption for a one-unit change in each variable, making the analysis more effective.

Initially, we wanted to create a detailed model considering various household features to predict electricity usage for each house every hour. However, this approach became impractical due to computer limitations. It would have required a massive amount of data replication for each house across all hours and days, making it computationally unfeasible. So, we switched to a simpler model that looks at energy consumption at the county level, which is more manageable and practical.

Linear Regression Model

The next approach involved modeling a linear regression, where we extracted the hour and the day of the week as independent features. The regression model is as follows:
y = a + bX + e
where y represents the electricity consumption, and X includes the temperature,the county, the hour, and the day of the week. Notably, ‘county’, ‘hour’, and ‘day of the week’ are treated as categorical variables. The regression model is structured to provide insights into the impact of these variables on electricity consumption. Upon running the linear regression, we derived key summary statistics, which are presented in the table below:

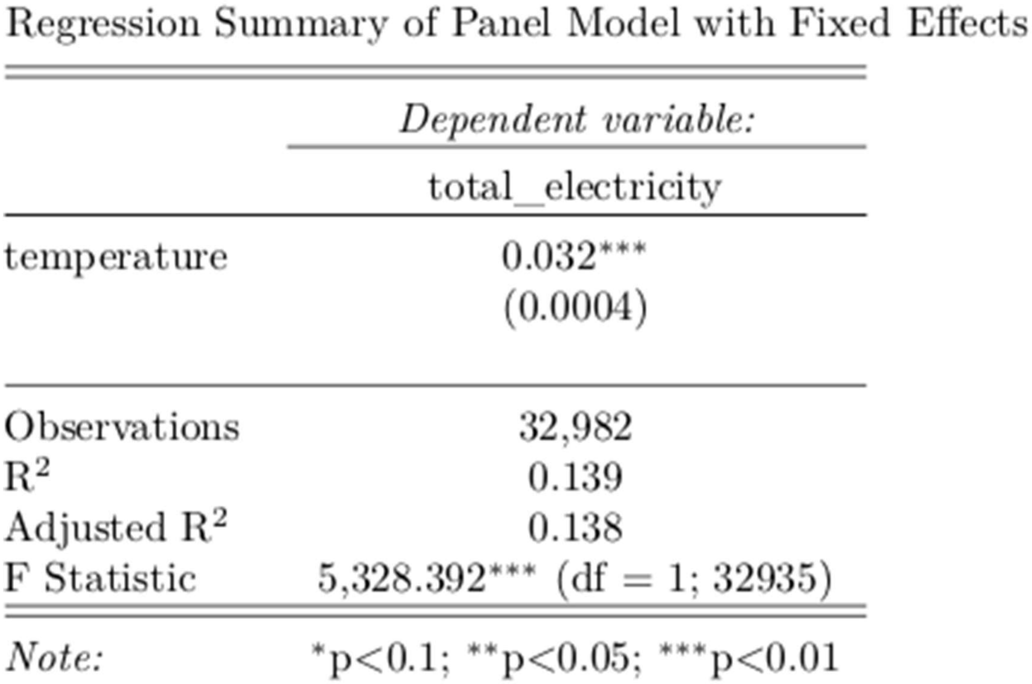


```
Coefficients:
                                     Estimate Std. Error  t value Pr(>|t|)    
(Intercept)                        -8.722e-01  3.354e-01   -2.600  0.00932 ** 
hour                                7.397e-03  6.162e-05  120.042  < 2e-16 ***
Dry.Bulb.Temperature...C.           1.163e-02  9.308e-04   12.496  < 2e-16 ***
Relative.Humidity....              -2.386e-01  2.737e-03  -87.190  < 2e-16 ***
Wind.Speed..m.s.                    1.713e-01  8.820e-04  194.242  < 2e-16 ***
Wind.Direction..Deg.                1.762e-05  4.698e-04    0.038  0.97009    
Global.Horizontal.Radiation..W.m2. -3.147e-01  2.729e-03 -115.303  < 2e-16 ***
Direct.Normal.Radiation..W.m2.      1.585e-01  8.861e-04  178.820  < 2e-16 ***
Diffuse.Horizontal.Radiation..W.m2. 4.661e-02  2.572e-03   18.125  < 2e-16 ***
in.pumaG45000102                    2.828e-02  2.787e-03   10.149  < 2e-16 ***
in.pumaG45000103                    2.808e-02  2.576e-03   10.903  < 2e-16 ***
in.pumaG45000104                    2.839e-02  2.876e-03    9.871  < 2e-16 ***
in.pumaG45000105                    8.400e-02  2.568e-03   32.717  < 2e-16 ***
in.pumaG45000200                   -1.179e-01  2.099e-03  -56.161  < 2e-16 ***
in.pumaG45000301                    2.786e-02  2.752e-03   10.121  < 2e-16 ***
in.pumaG45000302                    2.822e-02  2.498e-03   11.297  < 2e-16 ***
in.pumaG45000400                    1.456e-01  3.300e-03   44.120  < 2e-16 ***
in.pumaG45000501                    2.068e-01  4.550e-03   45.443  < 2e-16 ***
in.pumaG45000502                    2.071e-01  4.489e-03   46.138  < 2e-16 ***
in.pumaG45000601                    5.012e-02  3.885e-03   12.901  < 2e-16 ***
in.pumaG45000602                    2.639e-02  4.097e-03    6.443 1.18e-10 ***
in.pumaG45000603                    1.533e-01  4.678e-03   32.780  < 2e-16 ***
in.pumaG45000604                    1.529e-01  4.739e-03   32.262  < 2e-16 ***
in.pumaG45000605                    1.540e-01  5.158e-03   29.850  < 2e-16 ***
in.pumaG45000700                    9.308e-02  5.999e-03   15.516  < 2e-16 ***
in.pumaG45000800                    1.536e-01  5.534e-03   27.755  < 2e-16 ***
in.pumaG45000900                    1.906e-01  6.503e-03   29.305  < 2e-16 ***
in.pumaG45001000                    1.835e-01  7.584e-03   24.193  < 2e-16 ***
in.pumaG45001101                    1.761e-01  7.904e-03   22.282  < 2e-16 ***
in.pumaG45001102                    1.761e-01  7.864e-03   22.397  < 2e-16 ***
in.pumaG45001201                    2.077e-01  5.702e-03   36.429  < 2e-16 ***
in.pumaG45001202                    2.073e-01  5.609e-03   36.955  < 2e-16 ***
in.pumaG45001203                    2.075e-01  5.671e-03   36.586  < 2e-16 ***
in.pumaG45001204                    2.075e-01  5.649e-03   36.737  < 2e-16 ***
in.pumaG45001300                    2.066e-01  4.493e-03   45.992  < 2e-16 ***
in.pumaG45001400                    2.331e-01  4.589e-03   50.795  < 2e-16 ***
in.pumaG45001500                    1.667e-01  3.255e-03   51.203  < 2e-16 ***
in.pumaG45001600                    1.681e-01  2.777e-03   60.535  < 2e-16 ***
in.reeds_balancing_area            -7.587e-02  2.425e-03  -31.289  < 2e-16 ***
in.weather_file_longitude          -1.094e-01  2.208e-03  -49.522  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1069 on 130148 degrees of freedom
Multiple R-squared:  0.9075,	Adjusted R-squared:  0.9075 
F-statistic: 3.275e+04 on 39 and 130148 DF,  p-value: < 2.2e-16

      9
1.3167
        9
1.374857
```

The output of the regression model indicates a strong fit, as evidenced by the significance of all coefficients for each iteration of the categorical variables, and an adjusted R-squared value of 0.91. This high R-squared value suggests that the model explains a substantial portion of the variance in electricity consumption based on the selected variables.

Panel Data Regression with Fixed Effects

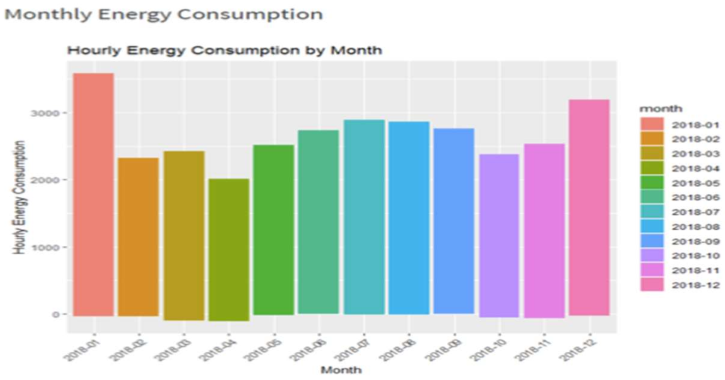
This subsequent approach treated this problem as a panel data regression with fixed effects, considering counties as the cross-sectional variable and the hour as the time variable. However, this approach yielded an R-squared of 0.13, resulting in an ineffective model.



Regression Summary of Panel Model with Fixed Effects

	Dependent variable:
	total_electricity
temperature	0.032***
	(0.0004)
Observations	32,982
R^2	0.139
Adjusted R^2	0.138
F Statistic	5,328.392*** (df = 1; 32935)
Note:	*p<0.1; **p<0.05; ***p<0.01

5. Graph explanation and conclusion:



The bar graph named "Monthly Energy Consumption" displays the variation in energy usage throughout the year 2018 on an hourly basis.

Each month is represented by differently colored bars, illustrating the fluctuations in energy consumption. The vertical axis indicates the amount of energy consumed per hour, while the horizontal axis represents the months from January to December in the year 2018.

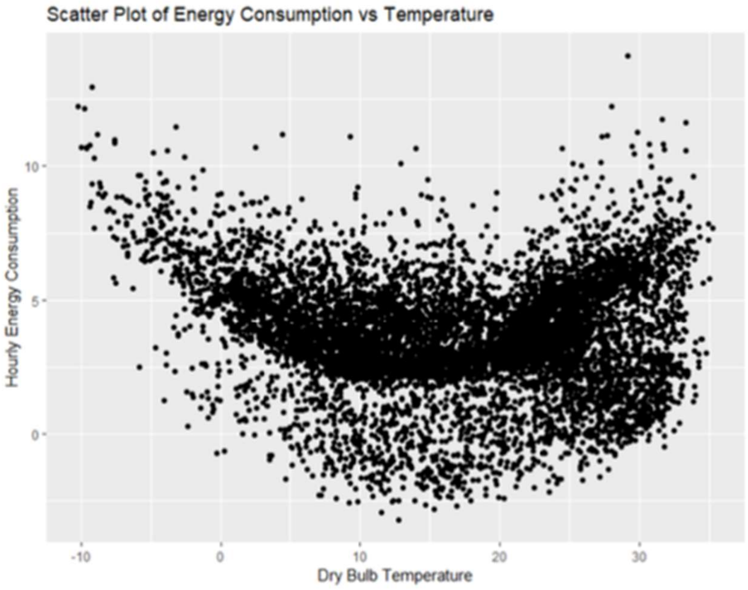
Observations:

- The highest energy usage is observed in January 2018, followed by a notable decrease in February.
- There is a steady rise in energy consumption from February to July.
- Peak energy usage is noticed in the summer months, followed by a slight decline as the year progresses towards winter.
- December 2018 records the lowest energy consumption.

Conclusion:

This observed pattern suggests a potential connection between energy consumption and seasonal temperature fluctuations. It implies higher energy use for heating in colder months and cooling during warmer months. The dip in December could be attributed to milder temperatures or efforts to conserve energy during the holiday season.

Scatter Plot



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

Observations:

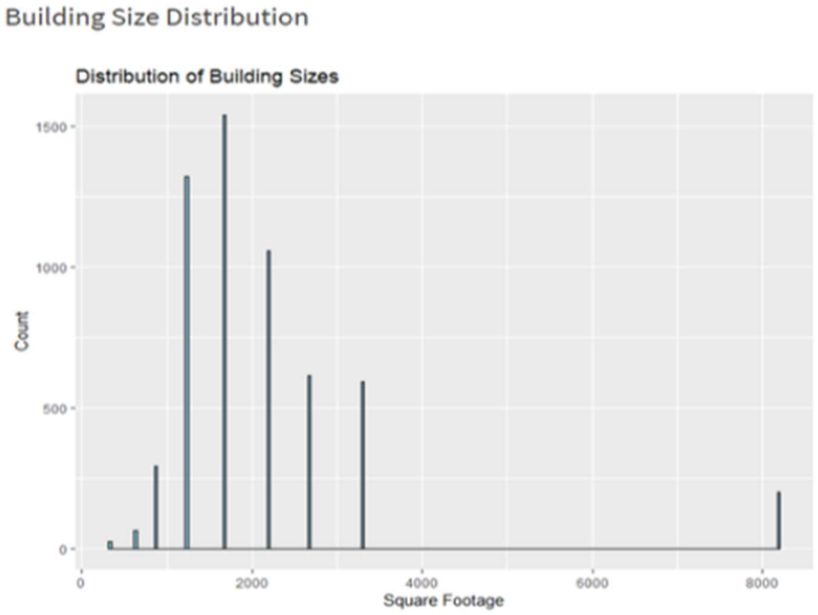
- The graph exhibits a broad distribution of data points, indicating variability in how temperature influences energy consumption.

- A parabolic relationship is evident, with higher energy use observed at both low and high temperatures, and a dip in consumption within the mid-range temperatures. This pattern may suggest increased energy needs for heating in colder temperatures and cooling in hotter temperatures.

- The least energy consumption occurs in the middle-temperature range, hinting at a comfort zone where neither heating nor cooling is necessary.

Conclusion:

The data implies that energy consumption for heating or cooling is linked to temperature extremes. There exists a temperature range where energy usage is minimized, representing a natural comfort zone in the environment from which the data was collected. This pattern is common in climates with cold winters and hot summers, where significant energy is used for temperature control.



The image shows a histogram titled "Building Size Distribution" with the subtitle "Distribution of Building Sizes."

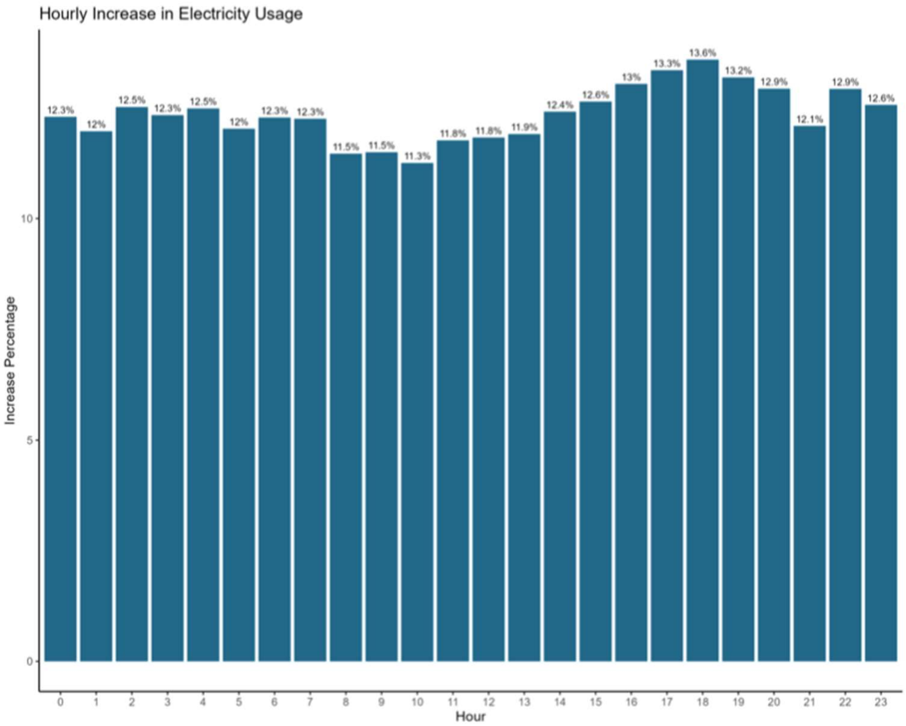
It represents the frequency of buildings according to their sizes in square footage.

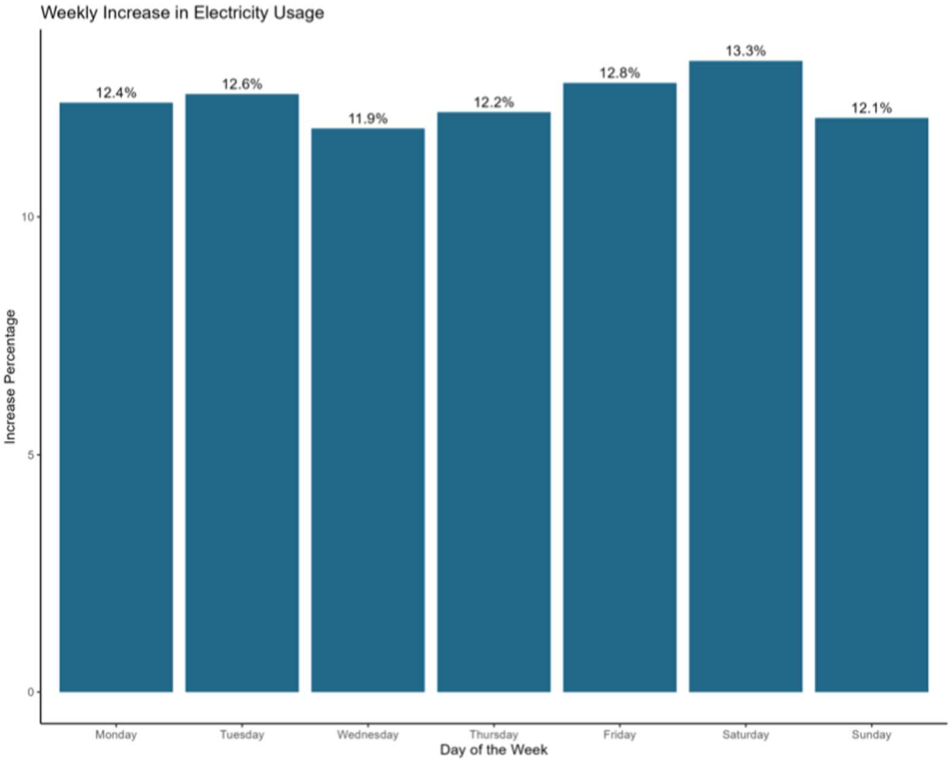
Observations:

- The histogram displays multiple peaks, signifying that specific building sizes are more prevalent than others.
- A prominent peak is noticeable around the 1,000 square footage range, indicating a significant number of buildings falling within this size bracket.
- The occurrence of buildings decreases as the square footage increases, with a decline in the number of larger-sized buildings.
- Minimal buildings are observed in the smallest size category (near 0 square footage) and the largest size category (over 8,000 square footage).

Conclusion:

This distribution implies that the majority of buildings in the dataset are of moderate size, with a scarcity of both very small and very large buildings. The pattern suggests that buildings around 1,000 square feet are the most prevalent, possibly representing a standard size for specific types of residential or commercial properties in the area where this data was collected.





The bar graph named "Monthly Energy Consumption" displays the variation in energy usage throughout the year 2018 on an hourly basis.

Each month is represented by differently colored bars, illustrating the fluctuations in energy consumption. The vertical axis indicates the amount of energy consumed per hour, while the horizontal axis represents the months from January to December in the year 2018.

The analysis based on days of the week offers a more detailed perspective. Anticipated demand increases are expected to be more significant on Saturdays and Fridays, likely due to heightened activities and occupancy during these days. Conversely, the smallest increase is projected on Wednesdays, possibly due to more stable routines and potentially lower occupancy or activity levels midweek. In summary, there's a higher demand increase on Saturdays and Fridays and a lower increase on Wednesdays.

Further insights are provided by the hourly analysis. The peak surge in electricity demand is predicted around 6:00 PM, aligning with the typical return of occupants to their homes and the subsequent rise in the use of electrical appliances and air conditioning systems. Conversely, the least increase is expected around 10:00 AM, a time characterized by lower occupancy and reduced electrical usage in residential settings.

6. Shiny Application:

The shiny application acts a dynamic interface for examining trends in energy use and forecasting upcoming energy needs. Stakeholders can utilize this application to delve into the elements that affect energy usage and to strategize appropriately for peak demand management. The application can process complex datasets and present them in a understandable format, allowing users to identify peak usage times and adjust strategies accordingly. It also provides a predictive analysis based on historical data, which can aid in the development of more efficient energy distribution plans and the reduction of waste through targeted demand response initiatives. This proactive approach not only enhances operational efficiency but also promotes sustainable energy practices.

Shiny App URL: https://citibiketracker.shinyapps.io/IST_687/

Features:

1. Linear Regression:

- This function permits users to choose a particular data column to focus their examination on.

- It then generates a detailed summary of the linear regression analysis, highlighting the connections between the chosen data points and their impact on energy use.

- In a more detailed view, this feature enables a deeper dive into the data by allowing a user to isolate a single variable and understand its correlation with energy consumption. The provided summary includes statistical measures such as coefficients, p-values, and R-squared values, which collectively elucidate the strength and nature of the relationship within the data.

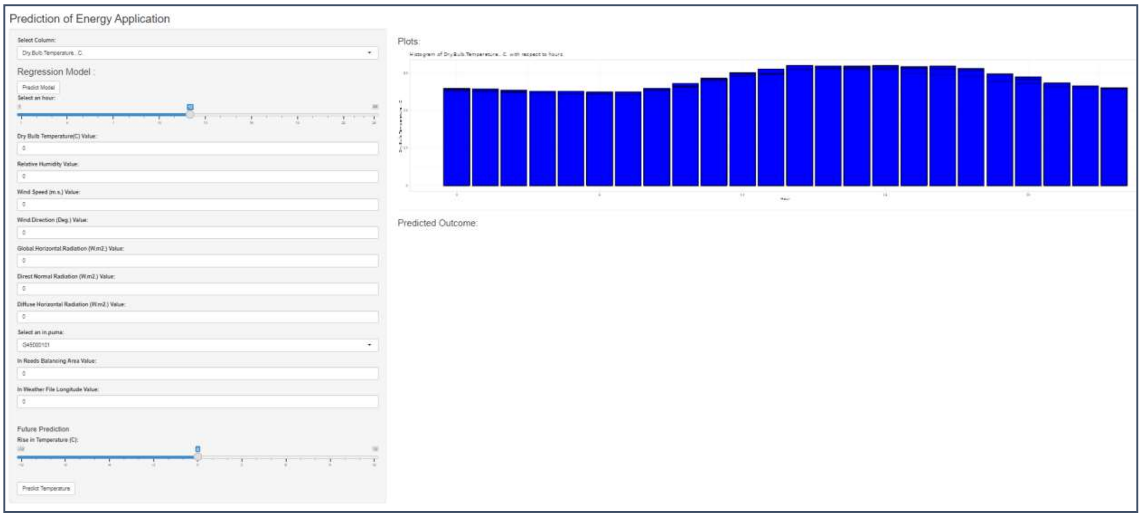
2. Creating Visualizations:

- With this tool, users have the capability to construct and interact with charts that map out how a selected variable is distributed over different hours.

- These charts not only visualize the frequency of data points but also help in detecting patterns over time, such as peak periods. By dragging along the time axis, users can explore variations in data distribution, which can be pivotal for identifying temporal trends that may influence energy usage.

3. Forecasting Function for Energy Use:

- The application provides a feature where users can enter variables like temperature, humidity, and wind speed to forecast future energy demands.
- It takes into account the possibility of changing environmental conditions to enhance the accuracy of its energy consumption forecasts.
- Users can leverage this forecasting function to stimulate different scenarios, effectively assessing how changes in weather conditions could affect energy needs. This predictive insight is particularly beneficial for energy providers and planners to adjust their supply or for business to optimize their energy strategies for future conditions.



Prediction of Energy Application

Select Column: Dry.Bulb.Temperature...C.

Select an hour:

Dry Bulb Temperature(C) Value: 0

Relative Humidity Value: 0

Wind Speed (m.s.) Value: 0

Wind Direction (Deg.) Value: 0

Global Horizontal Radiation (W.m2.) Value: 0

Direct Normal Radiation (W.m2.) Value: 0

Diffuse Horizontal Radiation (W.m2.) Value: 0

Select an in.puma: G45000101

In Reeds Balancing Area Value: 0

In Weather File Longitude Value: 0

Rise in Temperature (C):

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:

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:
   - Use the Linear Model: This model achieved the highest accuracy (RMSE of 0.1069) in the study.

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.

Model Performance: Predicting Consumption with Precision

Through rigorous model selection and evaluation, we identified the Linear model as the best performing tool for predicting hourly energy consumption. This model achieved an impressive RMSE of 0.1069 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 goes beyond the surface, pinpointing the peak periods of energy demand during the hottest hours of July. This critical understanding identifies specific times when energy consumption reaches its zenith, offering valuable insights 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 Linear 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.

Contribution by Individual Members in the Research

This research project was a collaborative effort where all 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:

Swapnil ,Sahil and Kenny 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:

Swapnil spearheaded the data analysis and manipulation efforts. He meticulously investigated the data, identifying patterns and relationships. Based on his 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:

Sahil took the lead on model selection and training. He researched various modeling techniques and proposed the use of Linear and Panel Data Regression, 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.