Technical Explanation Document

IST 687 M005 Group 4

**Note:**

1. Work done by every person is reported (and self-evaluated) only by himself
2. This technical document aims to provide working details and information to the technical manager. We assume potential readers know the background and don’t care about the code we used but the process, which should make sense.
3. Since this work is co-worked by a group processing massive data, every R markdown file is associated with multiple stages. For each stage, data frames used are extracted from saved R workspaces. Thus, there isn’t any knitted PDF file for code due to multiple stages within one file. Please refer to this GitHub link for full code access: “*https://github.com/ht6631/IST687\_Project*”.

**Description**

The project mainly focuses on energy consumption measured for each house in different counties for several months. The project aims to accurately predict and manage future energy usage for single-family houses, particularly in the context of rising temperatures, to optimize energy efficiency and reduce costs.

**Objectives**

1. **Data Preprocessing & Merging**
   1. Data preprocessing & merging by Hang Tian by 10.30:

I retrieved static house information from AWS and named the data frame “static\_house.” All character columns were converted as factors for later usage. Later, I removed columns with only one unique value. After that, I calculated the correlation coefficients between columns in the data frame, then removed columns with correlation coefficients higher than 0.8 with other columns.

After preprocessing the static house table, I retrieved energy and weather data using “for” loops. In this loop, every house building ID and its county number are provided at the beginning, then the energy usage information for July is read from AWS. After that, weather data in July is retrieved after converting it to the same time zone as energy data (EST or EDT). By the end of each loop, static house information, energy usage, and weather data are joined together by building ID, county, and time columns. This results in a final table for modeling, having more than 4 million rows and 100+ columns.

* 1. Data preprocessing and merging by Aadit Malikayil

Using the 4 million dataset produced by Hang, I tried to view all the building energy consumptions for each county. Then I sampled the buildings per county and their energy consumption using the building energy dataset. I found that the data for the buildings is skewed in a way that per county the number of buildings that used 16,17 and 18 electrical appliances outnumbered the number of buildings that consumed energy using lesser appliances.

Using this information I calculated the mean of the count of the appliances that show energy consumption of buildings in the county. This number was used to reduce the number of buildings and sample the building IDs per county effectively for further analysis. A for loop was created to go through each and every county and use the calculated mean to slice the number of building IDs per county. Every loop created a subset dataframe which was appended to the master dataframe for future use. The result of this preprocessing reduced the row count from roughly 4.8 million to 2.8 million rows.

1. **Descriptive Analysis & Visualization**
   1. Plotting by Hang Tian by 11.30:

I made a boxplot for all the hourly energy usage amounts and another one for the by-hour energy usage boxplots. Negative values are noticed among hourly total energy usage records; we cannot understand the meaning of those values, so we decided to exclude them.

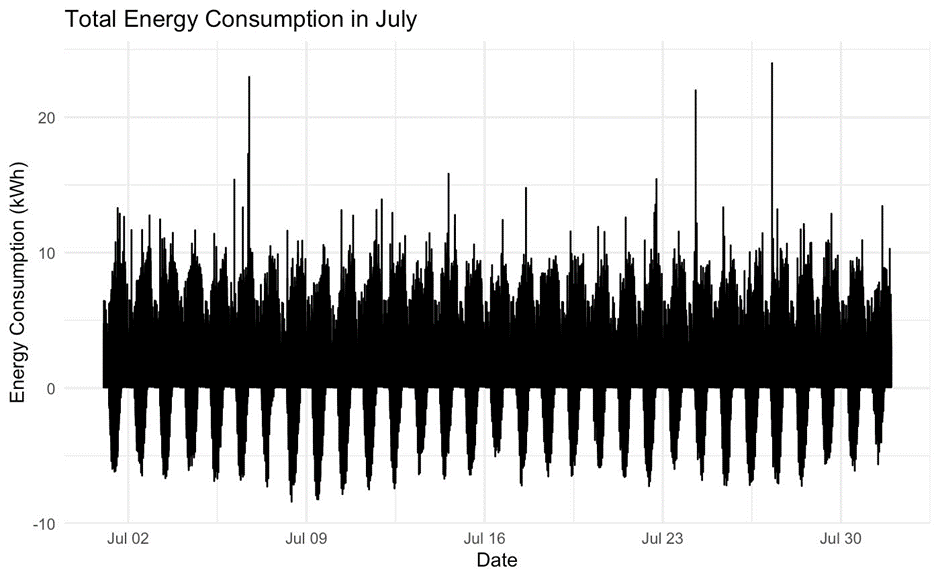
* 1. Plotting by Abhishek Namana

I have explored the dataset provided by my teammates to identify the variables which have dependencies, relations and effects on the energy consumptions and selected the suitable variables for the plot to provide the best insights on the data which would be helpful in the modeling phase.

1. Reduced the dataset to include just the energy consumptions in the month of july as per the problem statement.This dataset includes the total energy consumption in the month of july for all the counties along with the different energy sources available and the energy consumptions as per appliance.

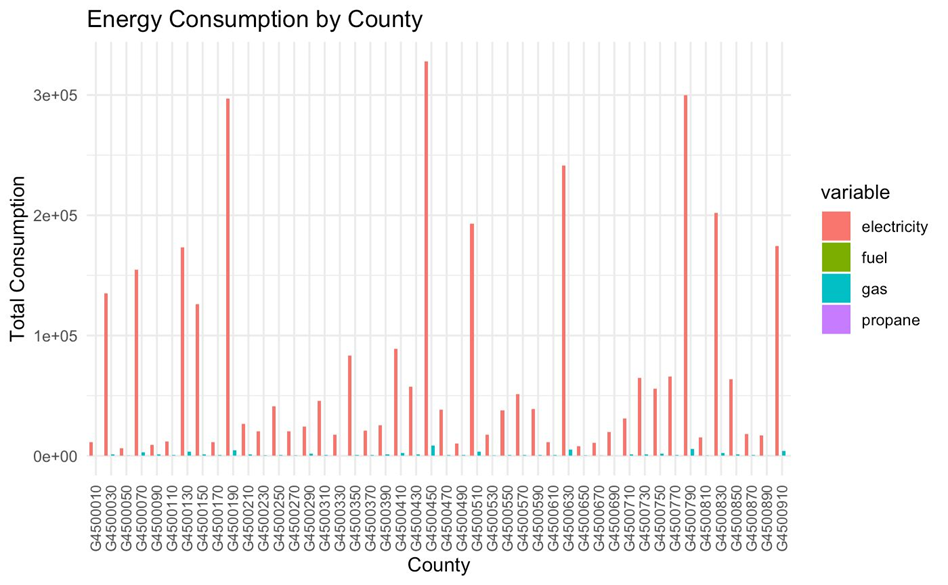
c. Total Energy Consumption in July (Time Series Plot):

1. I made a Time Series plot to display the daily energy consumption in kWh for the month of july by considering the Total energy consumption. The plot provides an insight on which days have the highest consumption of energy and which days have the lowest.It was also observed that the dataset has negative values which helped us to identify and work on them, to facilitate on developing a model with best possible accuracy.



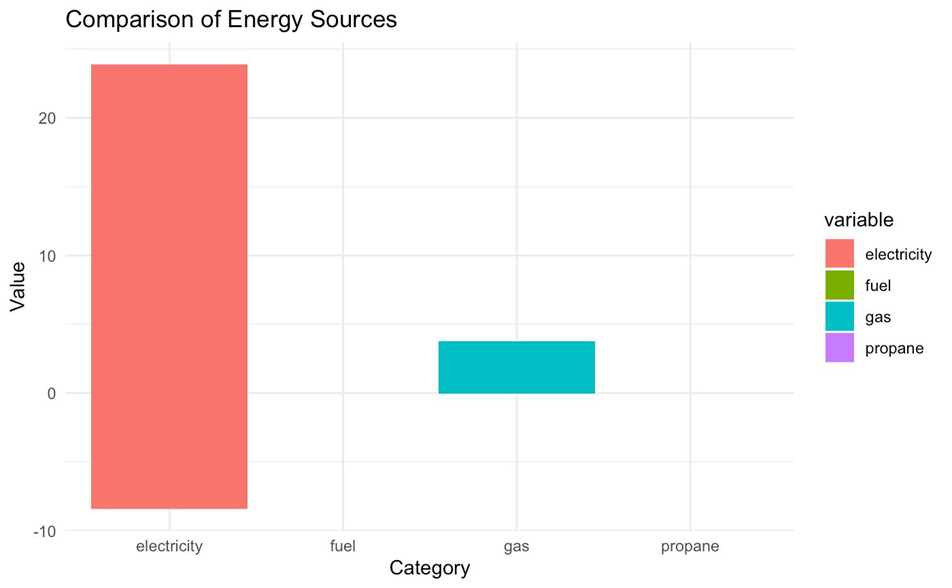
d. Total Energy Consumption by County (Stacked Bar Chart):

1. Plotted a stacked bar chart to observe the energy consumption county wise.Along with the total energy consumption,various energy sources have been considered to know their contribution in each county.Through this plot, we were able to determine the counties with highest energy usage and county with least energy usage. There were few counties with highest usage which made them our focus points to examine more and make suggestions to reduce the usage of energy.



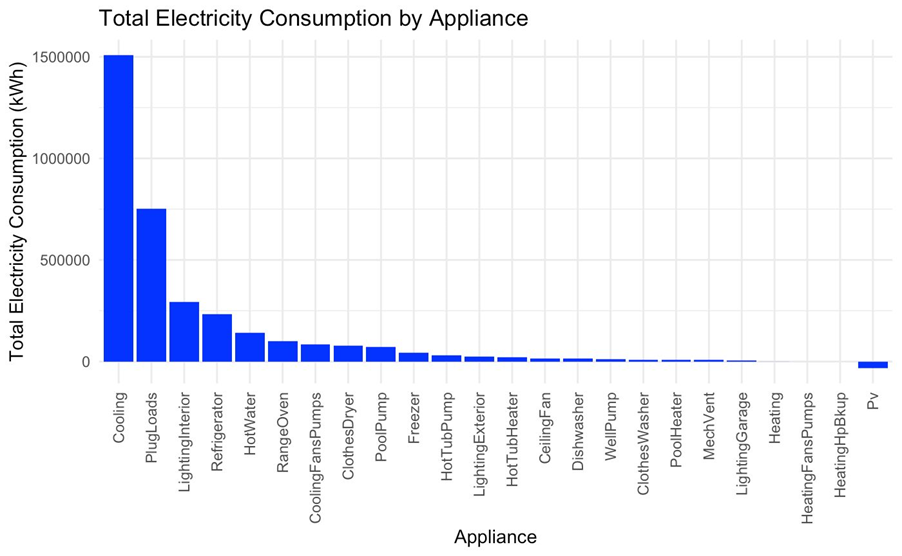
e. Comparison Of Energy Source (Scatter Plot):

1. To have a more detailed information on the different types of energy resources and their consumption,I chose a scatter plot on the energy sources and their contribution.It was visibelm that only electricity and gas are the energy sources used in all the households ,leaving fuel and propane with null values



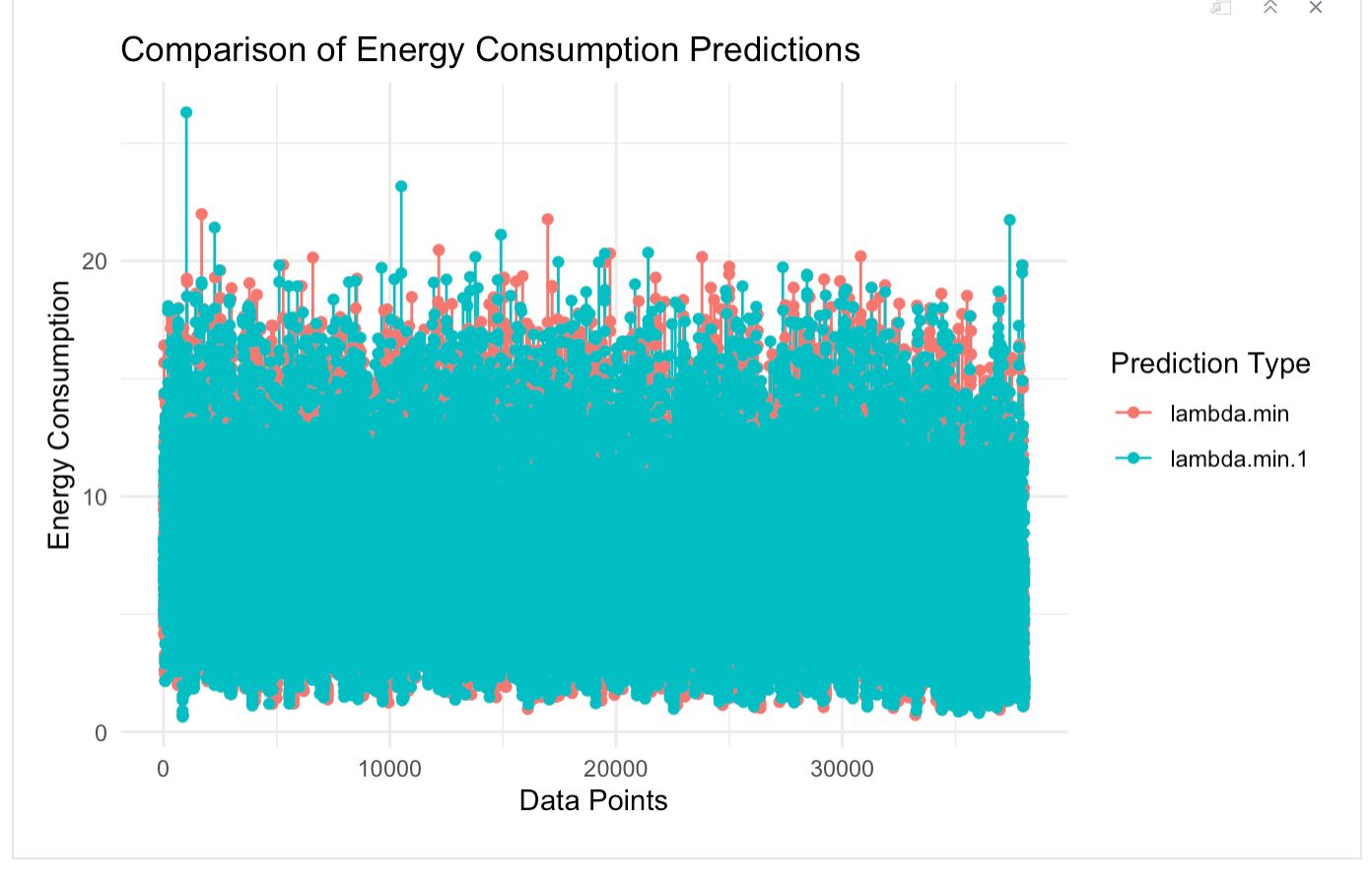
f. Total Electricity Consumption by Appliance (Horizontal Bar Chart):

1. I used a horizontal bar chart to understand how each appliance is consuming the energy in kWh. From the plot it was found that the cooling appliance is using the most energy followed by plug loads,lighting interior.Whereas, the lighting garage has the least energy consumption.Based on the observations made we can make few suggestions on how to reduce the energy consumption.



g. Comparison of Energy Consumption Predictions(Scatter Plot):

1. After the Linear modeling phase, the predictions have been made with the dataset which has the temperature 5 degree warmer for the month of july. A scatter plot was used to show the energy consumption difference between the old dataset with normal temperature and new dataset with 5 degree warmer temperatures.



1. **Modeling Preparation & Attribute Selection**
2. Attribute selection by Hang Tian by 11.23:

I ran a linear model on the table for modeling created by 10.30. However, the vector memory of R was exceeded due to the number of records. As a result, we decided to work on attribute selection and other ways to cut off the dimensions of this table. First, I calculated the average value of hourly total energy usage data for each building in July. Then, I figured out both Pearson's and Spearman's rank-based correlation coefficients between energy average and other static house attributes. Those having a coefficient of less than 0.1 in both results are excluded.

1. Attribute selection by Hang Tian by 12.1:

After the work done by my groupmates, running a linear regression model on the modeling table is now feasible. I ran linear regression twice, each time excluding those columns having P-values higher than 0.01, considering them not statistically significant in affecting the dependent variable - hourly total energy usage. This list of excluded columns is passed to my group mates for later modeling.

1. Attribute selection by Aadit and Biswadip:  
   Me and Biswadip worked on a linear model on the cleaned dataset and got rid of columns using the p-values where we considered variables having p-values < 0.05. We created a new dataframe, this time the temperatures were 5 degrees higher. We used this linear model to predict the future energy consumptions. This data was used to present data in the form of graphs on the shiny dashboard.
2. **Grouping Rows based on bldg\_id, plain\_date, time\_period by Revanth Shahukaru**
   1. Data Grouping Strategy

Revanth implemented a grouping strategy based on three primary criteria: 'bldg\_id' (building ID), 'plain\_date' (date without timestamp), and 'time\_period'. This involved organizing the data into distinct groups based on these criteria.

* 1. Time Period Segmentation

Revanth specifically segmented time periods to capture different parts of the day. This segmentation likely involved categorizing hours into morning, afternoon, evening, and night. This step is crucial for understanding how energy consumption varies throughout the day.

* 1. Aggregation

The data was aggregated within each group to derive meaningful insights. For example, total energy consumption has been calculated for each time period within a day for each building. This specific aggregation strategy facilitates a more detailed analysis of energy usage patterns over time.

These points highlight Revanth's contribution in structuring the data based on building ID, date, and time periods, with a focus on time-of-day segmentation for a more nuanced understanding of energy consumption patterns. Grouping the rows based on the time period in a day also helped in reducing the overall rows from around 2.8 Million data points to about 480K data points.

1. **ML Pipeline - Data Cleaning by Revanth Shahukaru**
   1. Data Preparation:
      1. Utilized regular expressions to transform specific columns like 'in.cooling\_setpoint\_mode', 'in.cooling\_setpoint\_offset\_magnitude\_mode', etc., from non-numeric to numeric values.
   2. Data Transformation - Numerical:
      1. Performed Data Transformation on Numerical columns using a standardization technique called as Z-Score Normalization.
   3. Categorical Column Extraction:
      1. Extracted information from columns such as 'in.clothes\_dryer\_mode', 'in.clothes\_washer\_mode', and 'in.cooking\_range\_mode'.
      2. Created new columns like 'in.clothes\_dryer\_type', 'in.clothes\_dryer\_usage', 'in.clothes\_washer\_type', 'in.clothes\_washer\_usage', 'in.cooking\_range\_type', and 'in.cooking\_range\_usage'.
      3. Replaced "None" strings with null values where applicable.
   4. Column Removal:
      1. Removed specific columns like 'in.insulation\_rim\_joist\_mode' based on the presence of too many null values.
      2. Removed columns like 'upgrade.ducts\_mode', 'upgrade.infiltration\_reduction\_mode', 'upgrade.insulation\_ceiling\_mode', 'upgrade.insulation\_wall\_mode' due to having only one true level.
   5. Numeric Value Extraction:
      1. Extracted numeric values from columns like 'in.heating\_setpoint\_mode', 'in.heating\_setpoint\_offset\_magnitude\_mode', etc.
   6. Null Value Handling:
      1. Checked and handled null values appropriately, dropping rows or columns when necessary.
2. **ML Pipeline - Model Building by Revanth Shahukaru**
   1. Linear Regression
      1. Built a linear regression model ('lm\_model') on the cleaned dataset.
      2. Made predictions on the test set and evaluated model performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-Squared.
   2. Lasso Regression
      1. Built a LASSO regression model ('lasso\_model') using cross-validation.
      2. Made predictions on the test set and evaluated model performance.
   3. Ridge Regression
      1. Built a ridge regression model ('ridge\_model') using cross-validation.
      2. Made predictions on the test set and evaluated model performance.
   4. Elastic Net Model
      1. Built an elastic net regression model ('elastic\_net\_model').
      2. Made predictions on the test set and evaluated model performance.
   5. Temperature Increase Experiment
      1. Data Preparation
         1. Created a new data frame ('model\_df\_5') by increasing the values in the "Dry.Bulb.Temperature...C.\_mean" column by +5 units.
      2. Model Building with Increased Temperature:
         1. Performed a train-test split on the new data.
         2. Built an elastic net regression model ('elastic\_net\_model\_5') on the data with increased temperature.
         3. Made predictions on the test set and evaluated model performance.
   6. Overall Contribution:

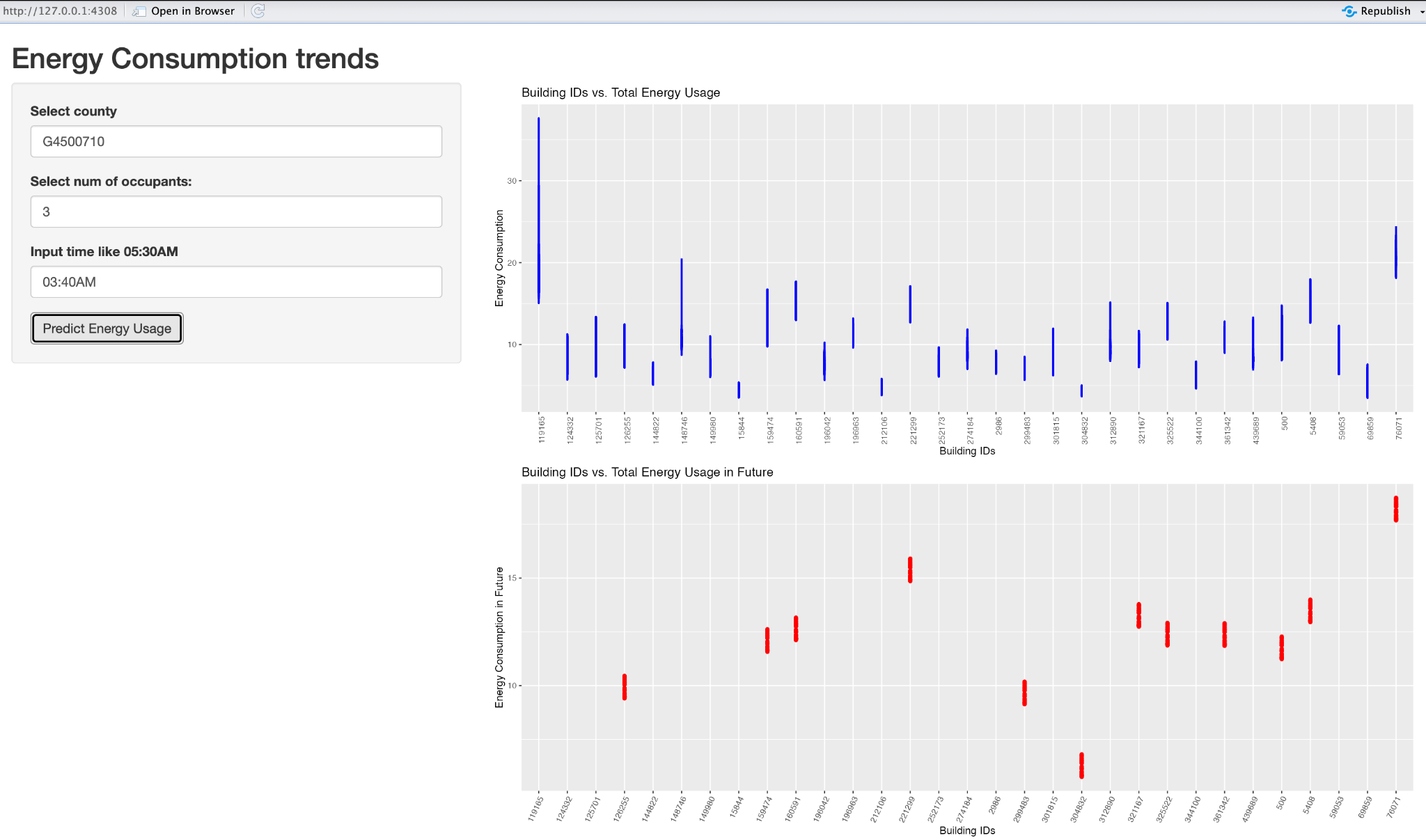
Revanth's work significantly contributed to data preprocessing, cleaning, and the development of multiple regression models. His involvement ensured that the data was prepared appropriately for modeling, and he played a crucial role in assessing the performance of different regression techniques.

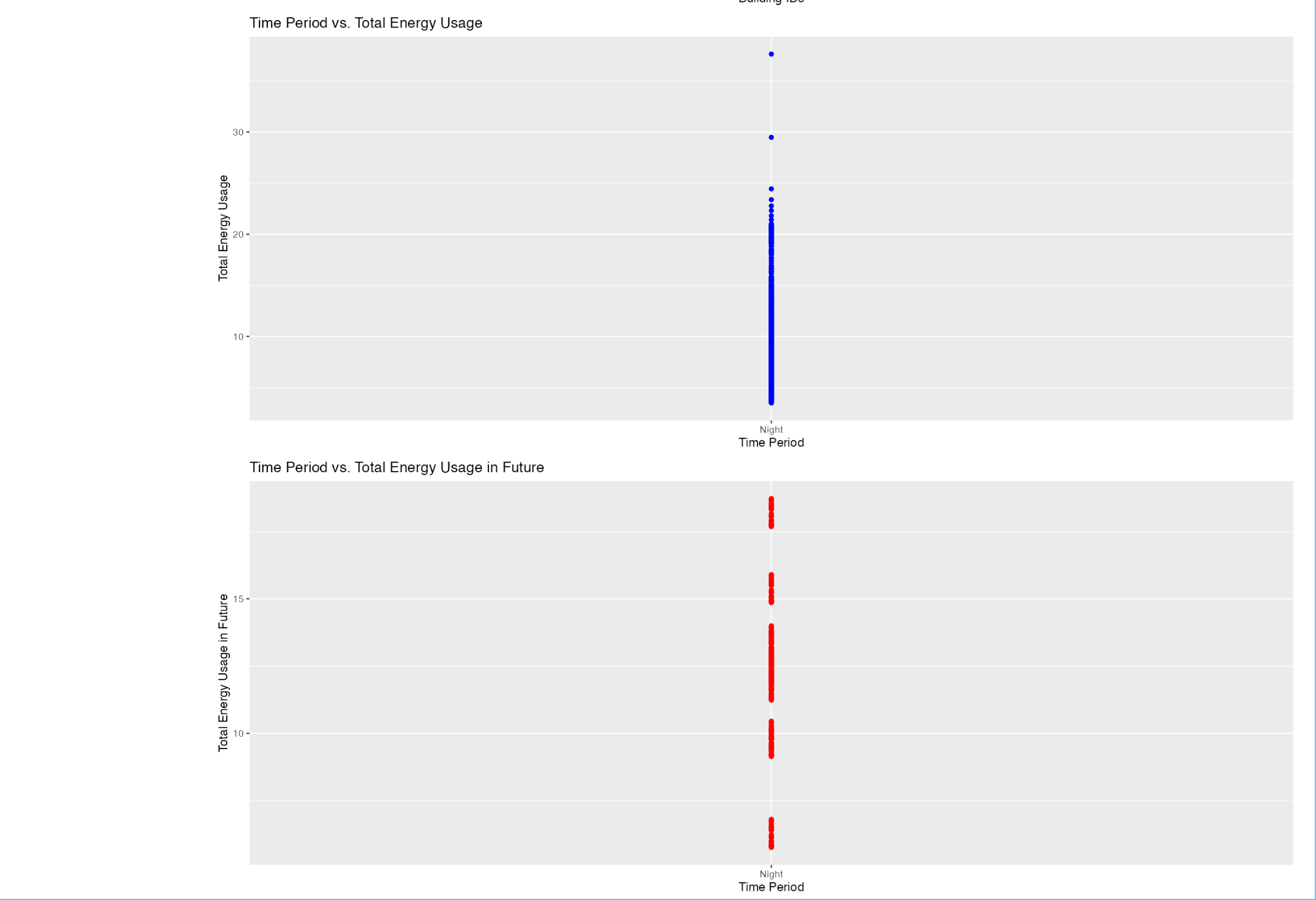
1. **Shiny App:**

Build by Biswadip and Aadit:

Me and Aadit have developed a Shiny web application designed to predict and visualize energy consumption trends for different counties. The application features three user-input parameters: time, number of occupants, and county ID. Leveraging present and future datasets, the app generates insightful graphs depicting energy consumption patterns. Users can explore how varying time periods, including morning, evening, afternoon, and night, impact energy usage. The tool serves as a valuable resource for analyzing and understanding the dynamic nature of energy usage across different counties and timeframes.

The graphs generated help produce the current and future energy consumption for different building IDs in the county. Also, they show the peak usage for different sets of building IDs based on the number of occupants in different building IDs. Based on the time input, eSC can view the energy consumption over a small range of hours for a county.

***Screenshots for the Shiny Dashboard:***  
 



1. **Conclusion:**
2. If the temperature increases by 5 degrees, our model predicts that the energy consumption will decrease. This may come from the fact that coefficients for temperature are primarily negative for our regression models. We may dig into it and determine if that could be converted into feasible suggestions.
3. There are a few houses with the most energy consumption, which can be considered outliers and can be examined separately and taken care of to reduce the energy consumption.
4. **Suggestions:**
5. Raise awareness on using the cooler in households.
6. In colder regions – take good care of the insulation of the home.
7. Replacing old appliances which consume a large amount of energy
8. Examining the houses in counties with the most energy consumption and taking measures to reduce the energy usage based on the issues found.