# Energy Consumption Analysis - Data Society

Analysis of energy consumption patterns in the steel industry to predict power usage.

## Variable Definitions

**Aim:** Predict Real Power Consumption using a Linear Regression Model.

**Definition of Active Power:** Usable or consumable energy in an AC Circuit (in the form of KW).

The **Usage_kWh** field contains the industry energy consumption in kWh, reflecting the real power consumption and is directly associated with the active current component.

**Assumption:** Variables associated with date and time affect power consumption patterns. We assume that energy consumption will vary depending on the weekday or time of the day.

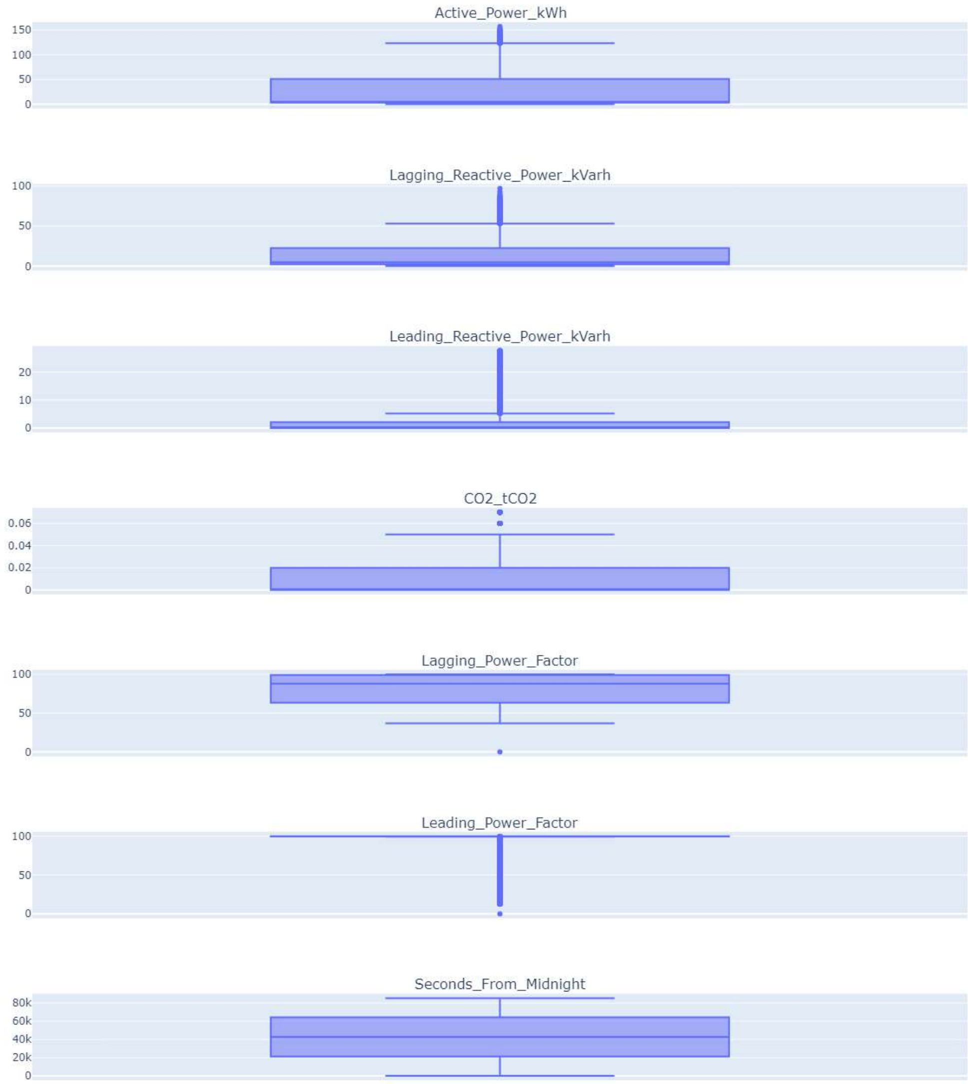
## Data Pre-Processing Steps

1. Data Cleaning: Import data and print columns to understand the dataset.
2. Column Renaming: Rename columns for better clarity.
3. Outlier Detection: Identify and remove outliers using the IQR method.
4. Feature Engineering: Create new features such as day, month, year, and time from the date.
5. Handling Missing Values: Check and handle any missing values.
6. Data Transformation: Apply log transformation to normalize skewed data.
7. Encoding Categorical Variables: Convert categorical variables to numerical format.
8. Splitting Data: Separate the dataset into training and testing sets.

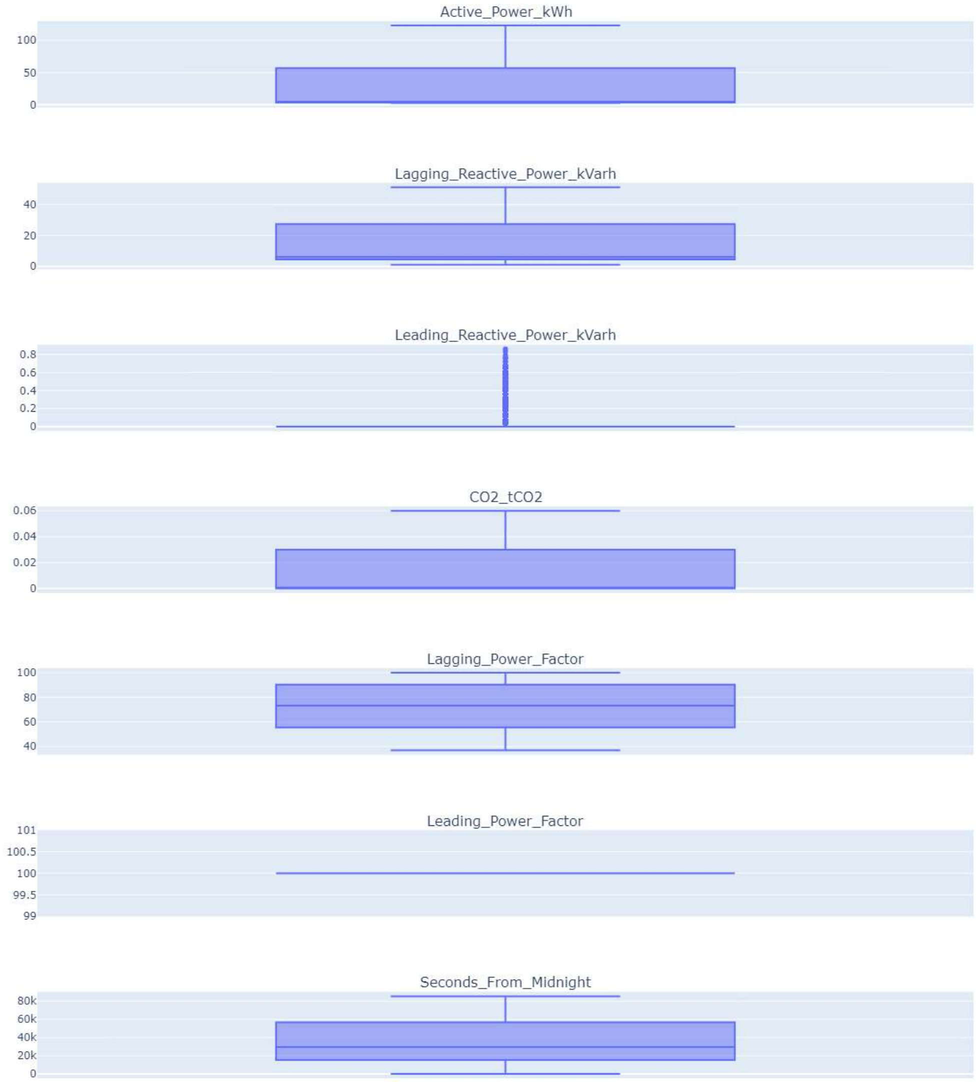
## High-Level Summary

This analysis focuses on understanding the patterns of energy consumption in the steel industry. Key findings include the identification of outliers, correlation between CO2 emissions and power usage, and variations in consumption based on time and load type. The analysis also evaluates the impact of holidays and weekdays on energy usage. The final model evaluation compares Linear Regression, Lasso, and Ridge models to determine the most accurate prediction method.

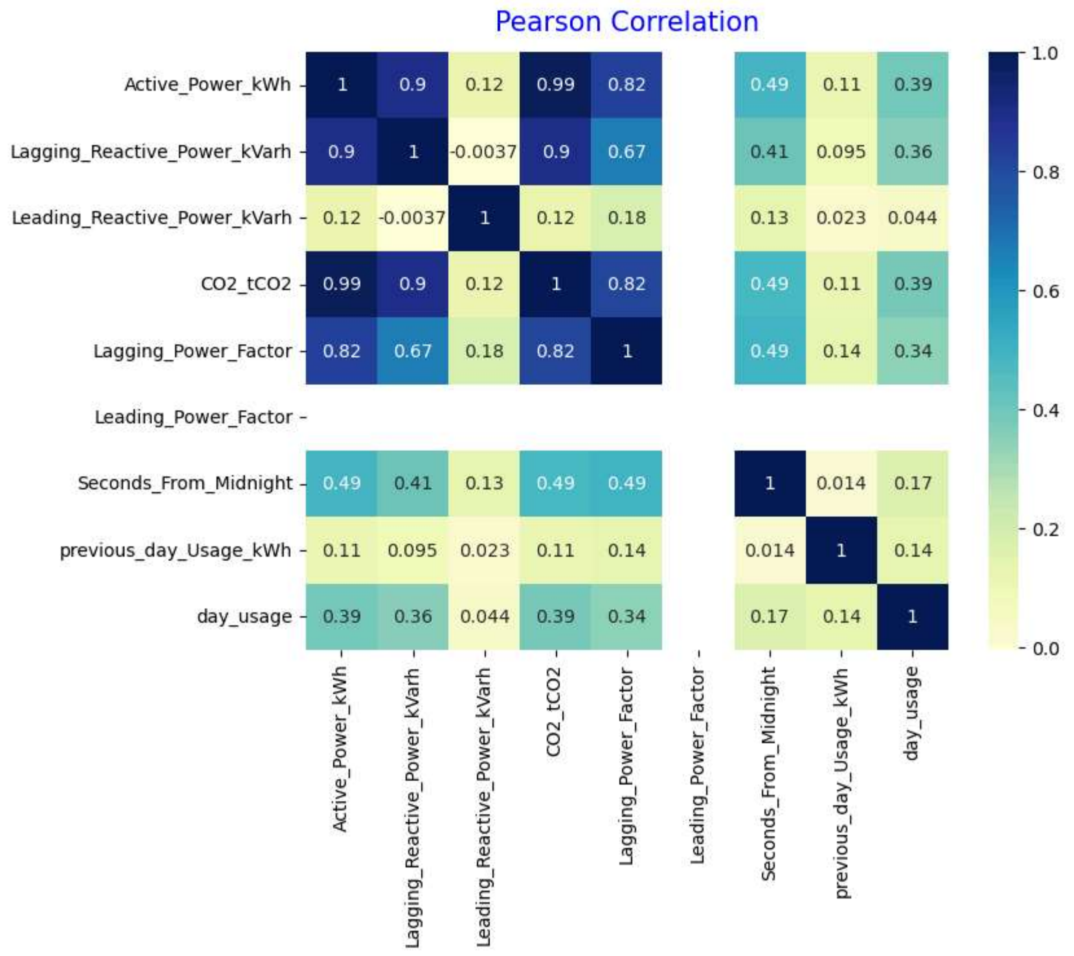
## Outlier Detection (Plot 1)



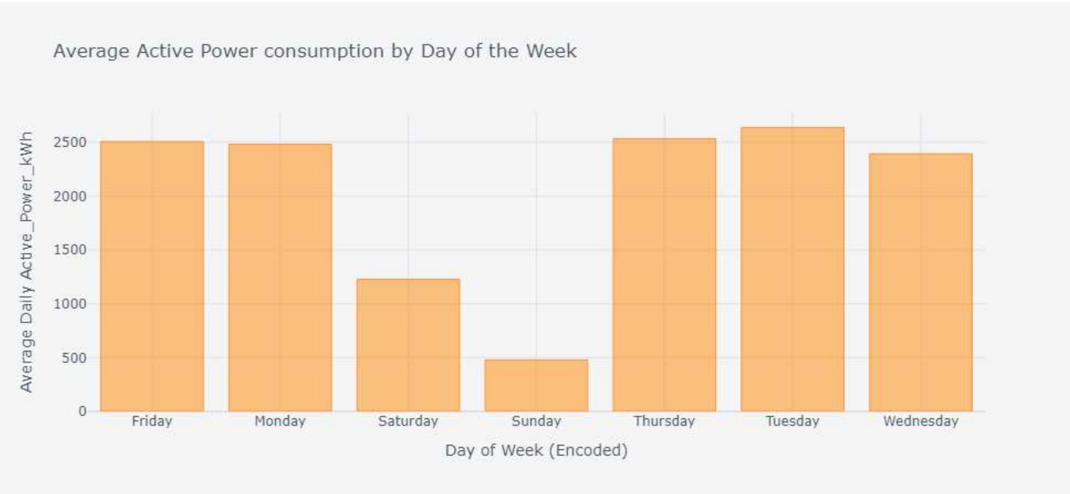

## Correlation Analysis (Plot 2)




## Average Power Consumption by Day (Plot 3)

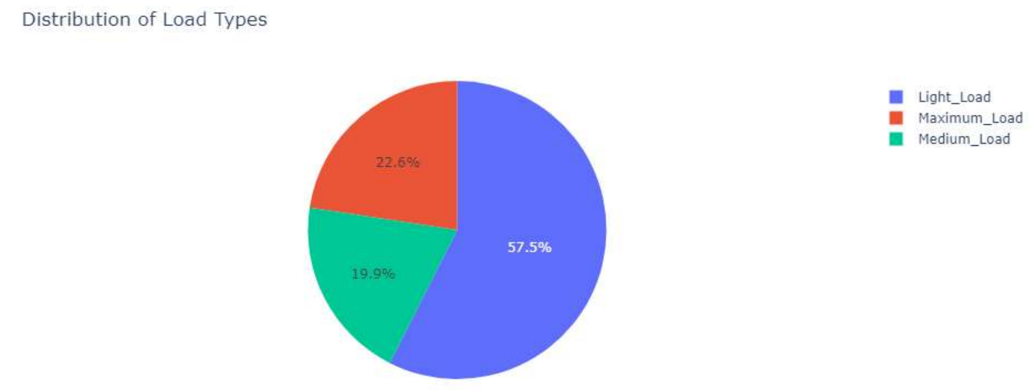



## Average Power Consumption by Week Status (Plot 4)

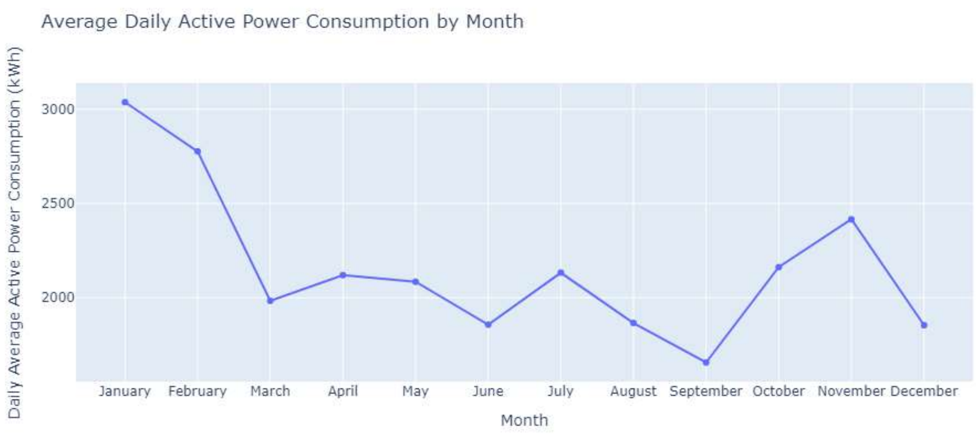



## Load Type Distribution (Plot 5)


## Monthly Power Consumption (Plot 6)




## Holiday vs Non-Holiday Consumption (Plot 7)



## Plot 8

Description for plot 8.


## Plot 9

Description for plot 9.


## Plot 10

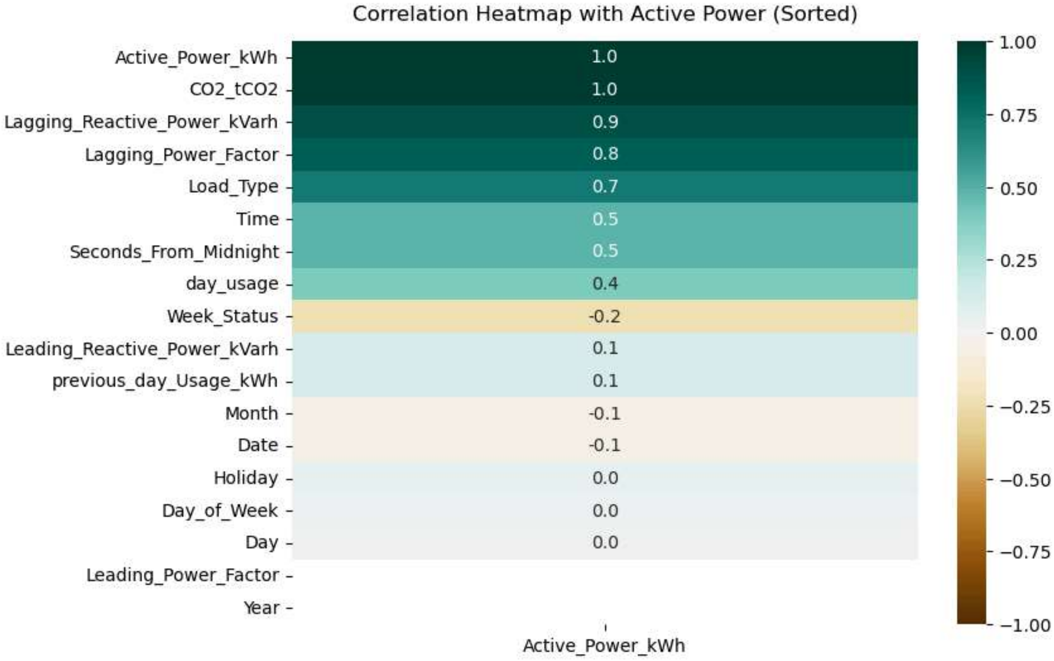
Description for plot 10.


# Plot 11

Description for plot 11.

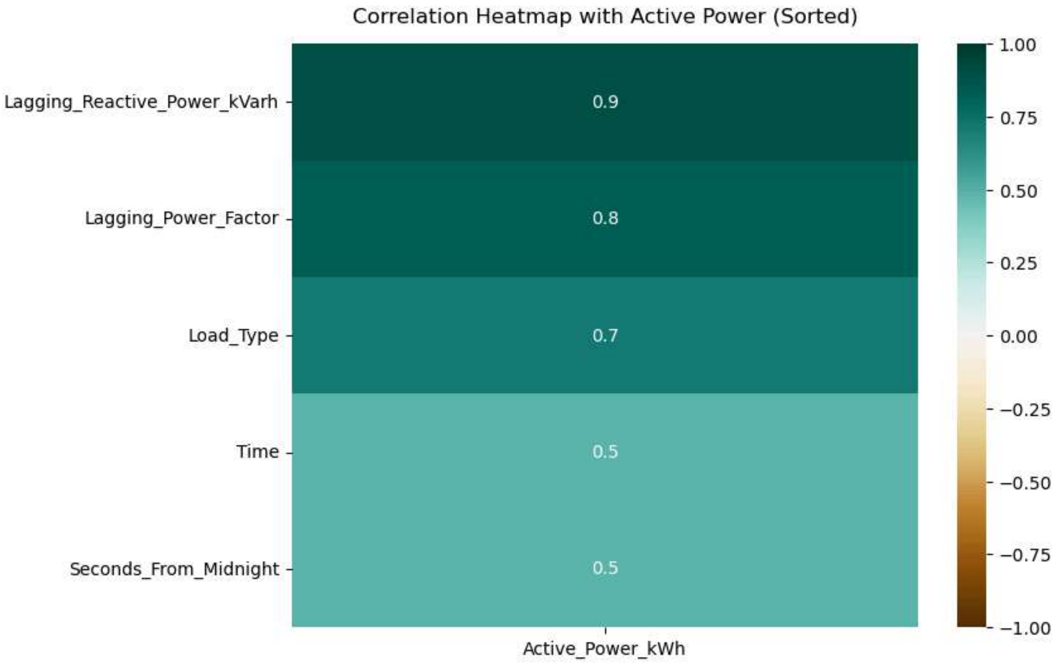

# Plot 12

Description for plot 12.



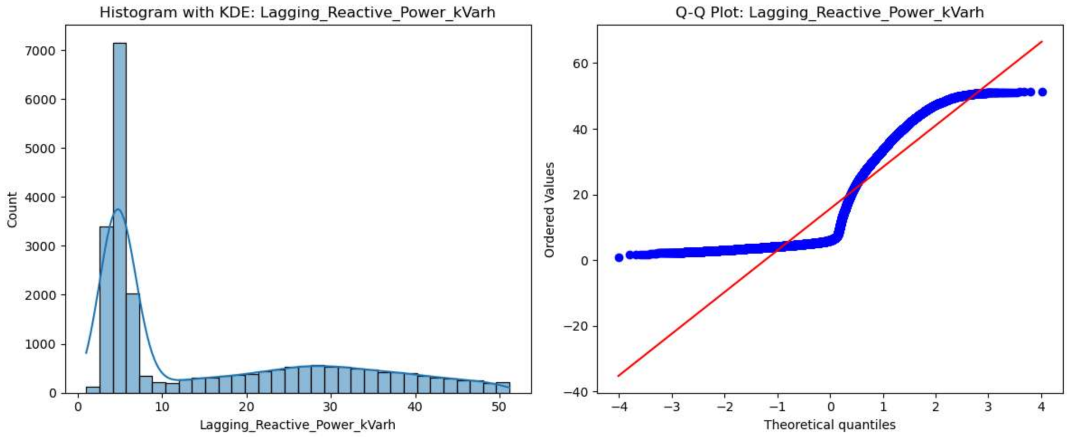

# Plot 13

Description for plot 13.



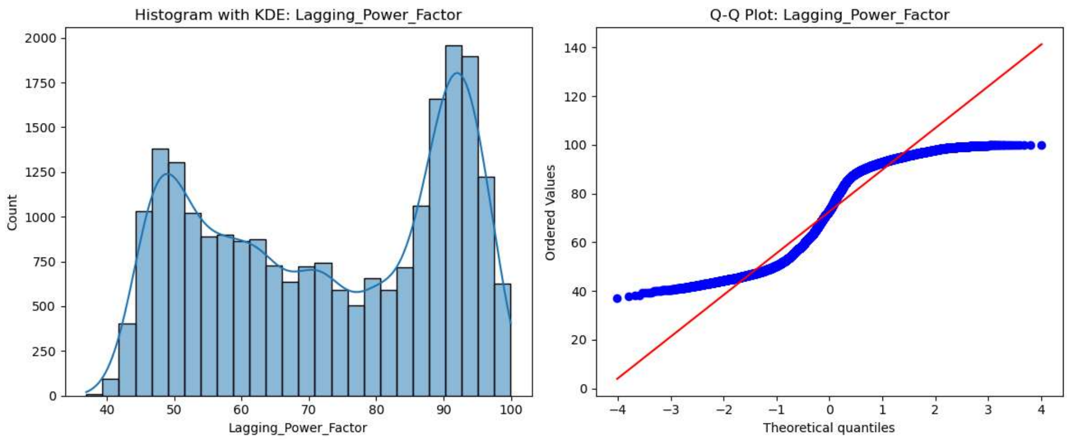

# Plot 14

Description for plot 14.



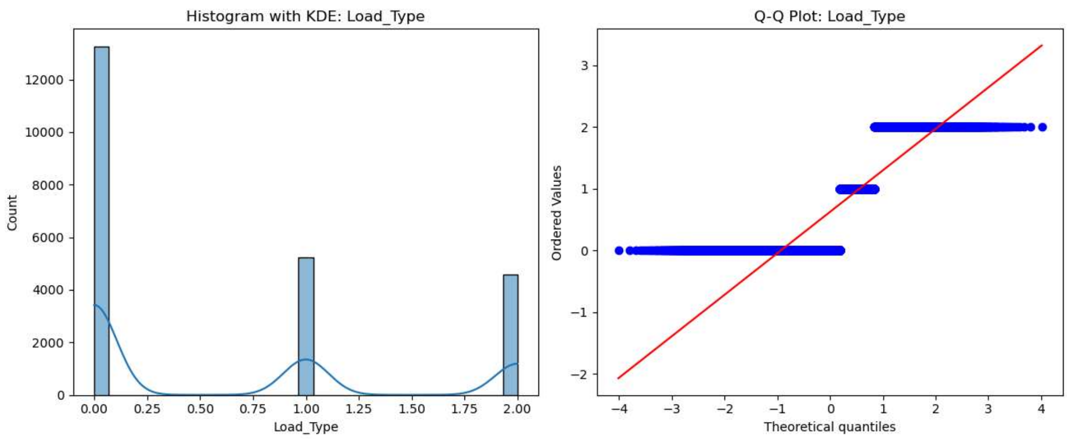

## Plot 15

Description for plot 15.



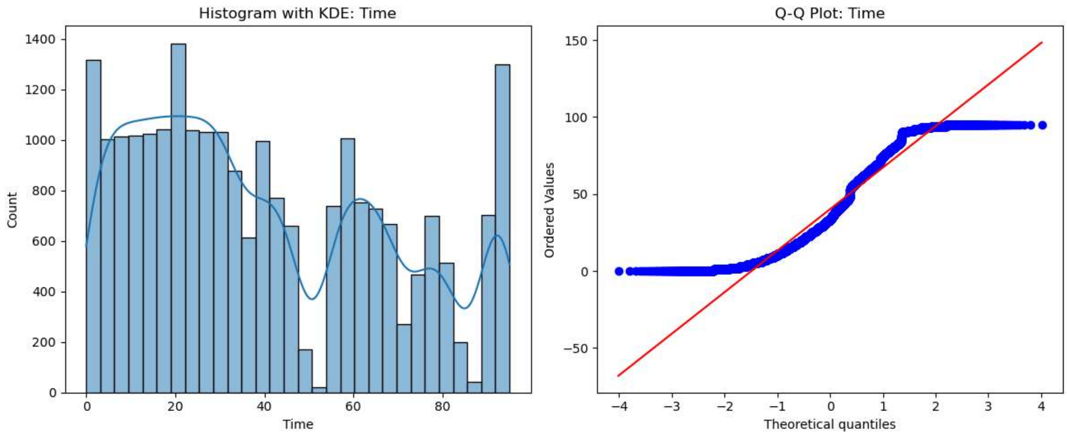

## Plot 16

Description for plot 16.



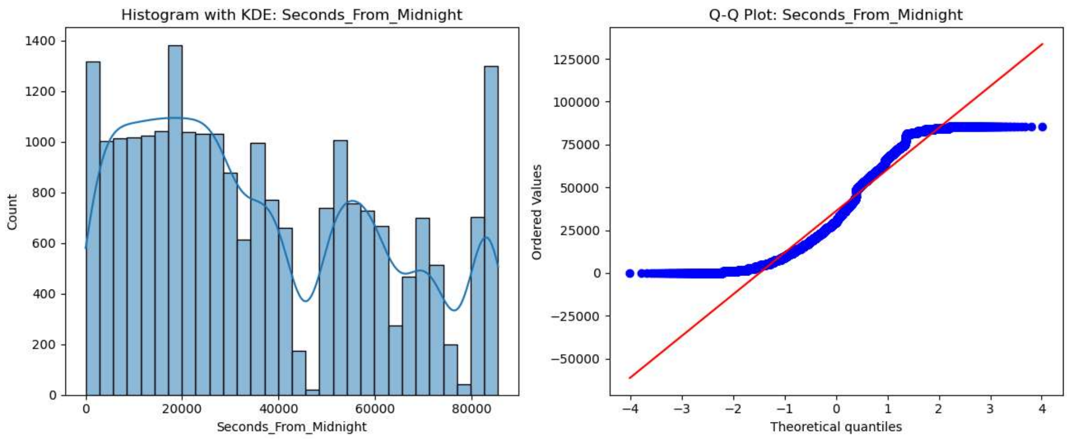
# Plot 17

Description for plot 17.



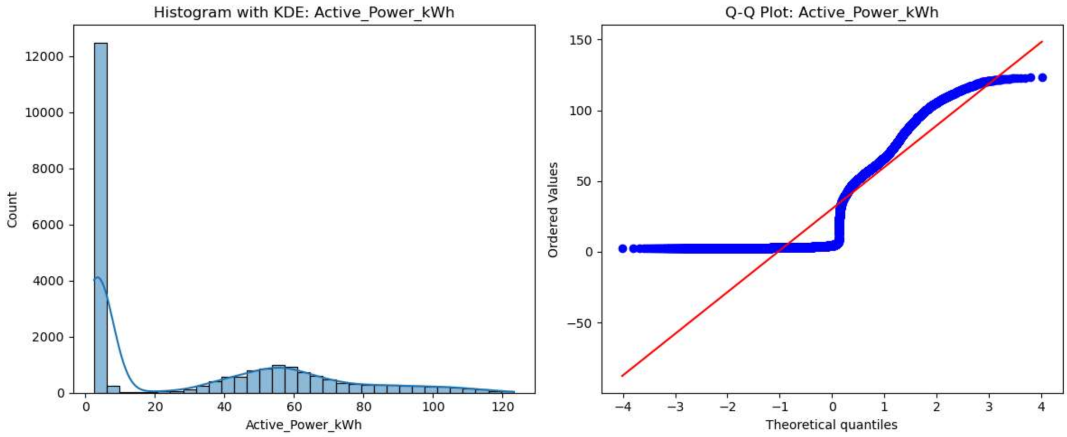

# Plot 18

Description for plot 18.



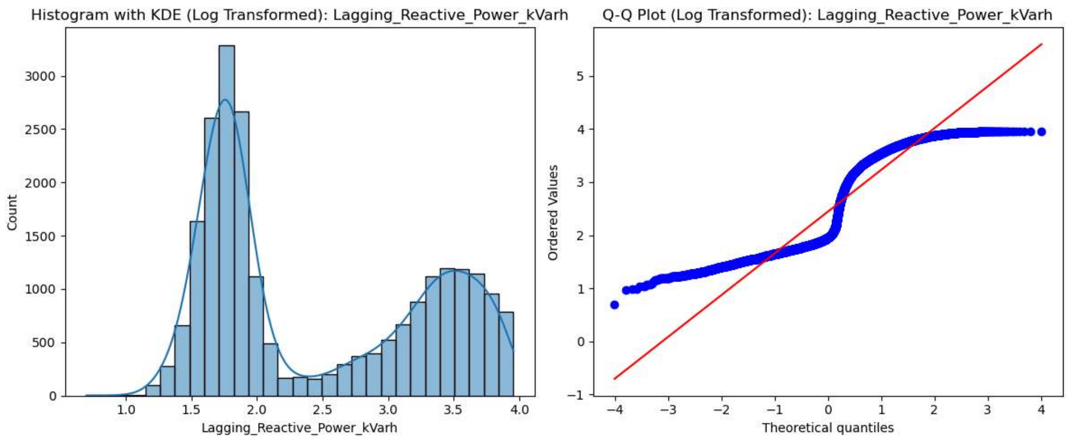

# Plot 19

Description for plot 19.



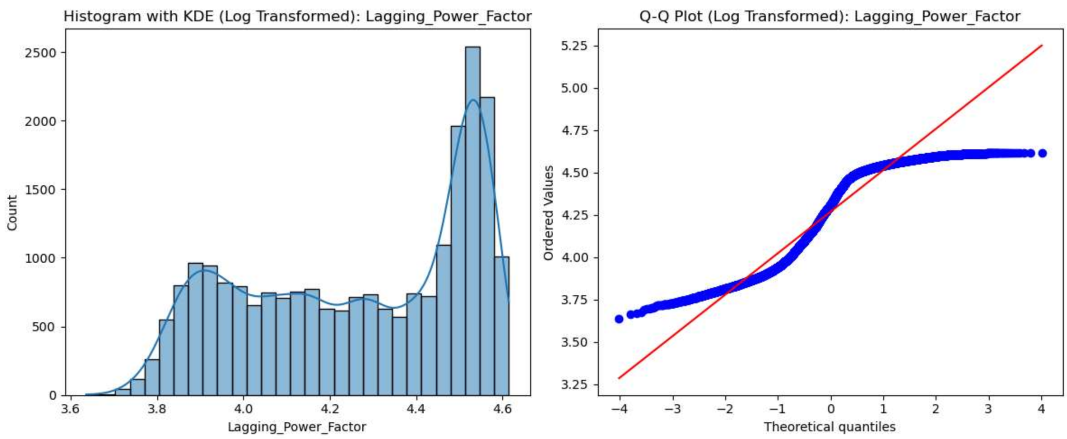

## Plot 20

Description for plot 20.



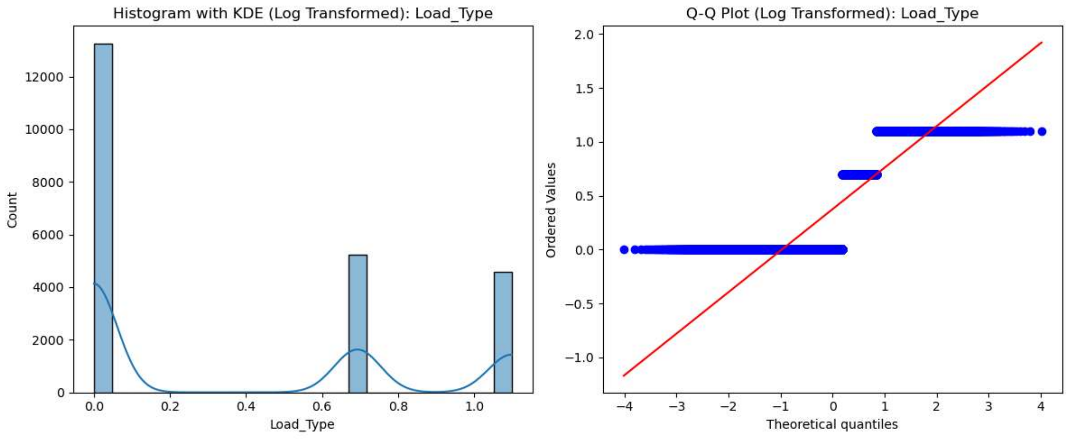

## Plot 21

Description for plot 21.



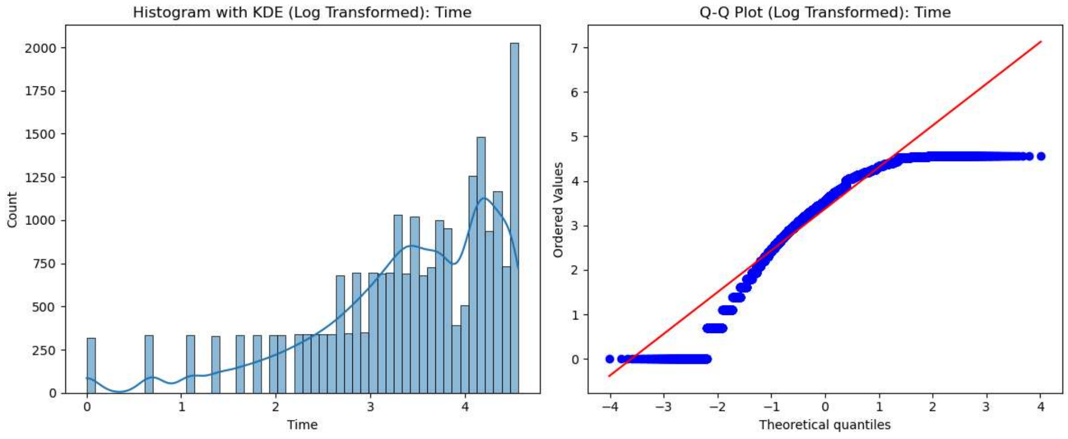
# Plot 22

Description for plot 22.



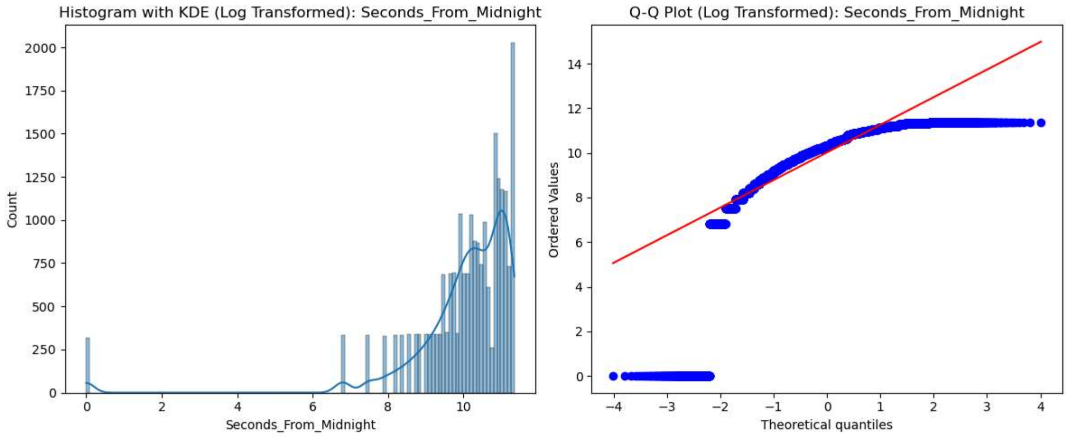

# Plot 23

Description for plot 23.



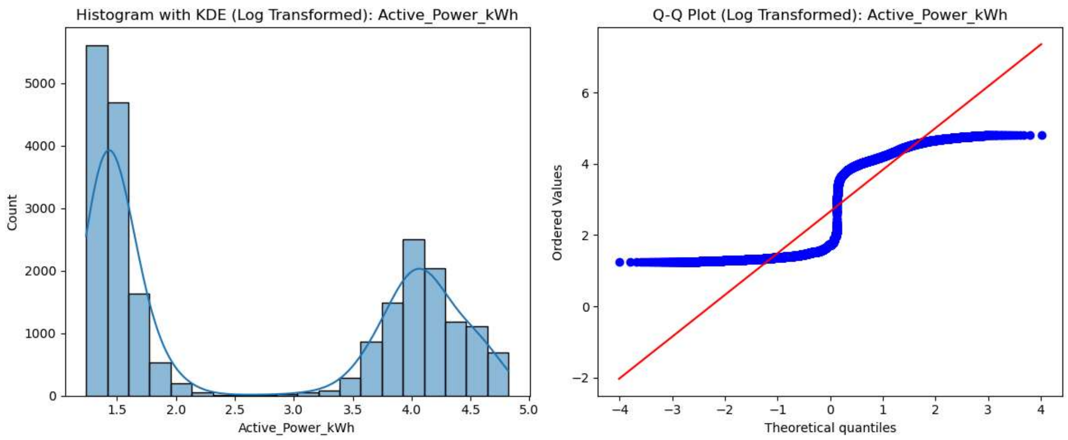

# Plot 24

Description for plot 24.



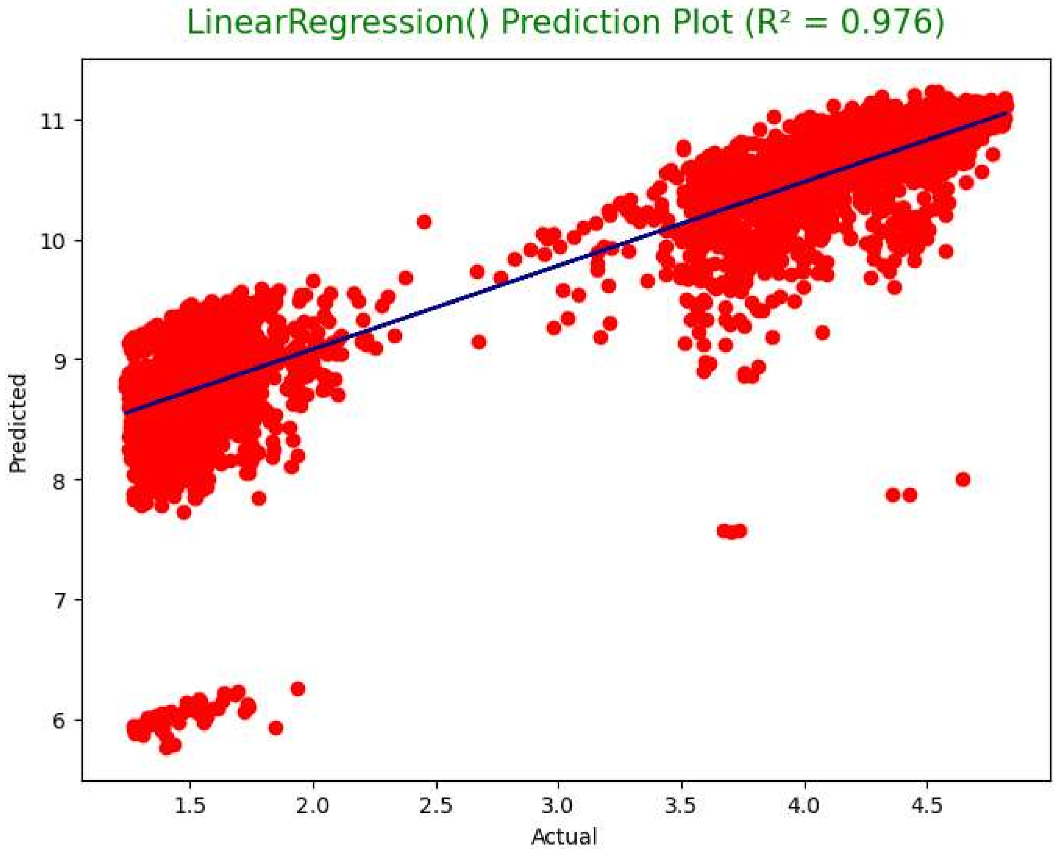

## Plot 25

Description for plot 25.



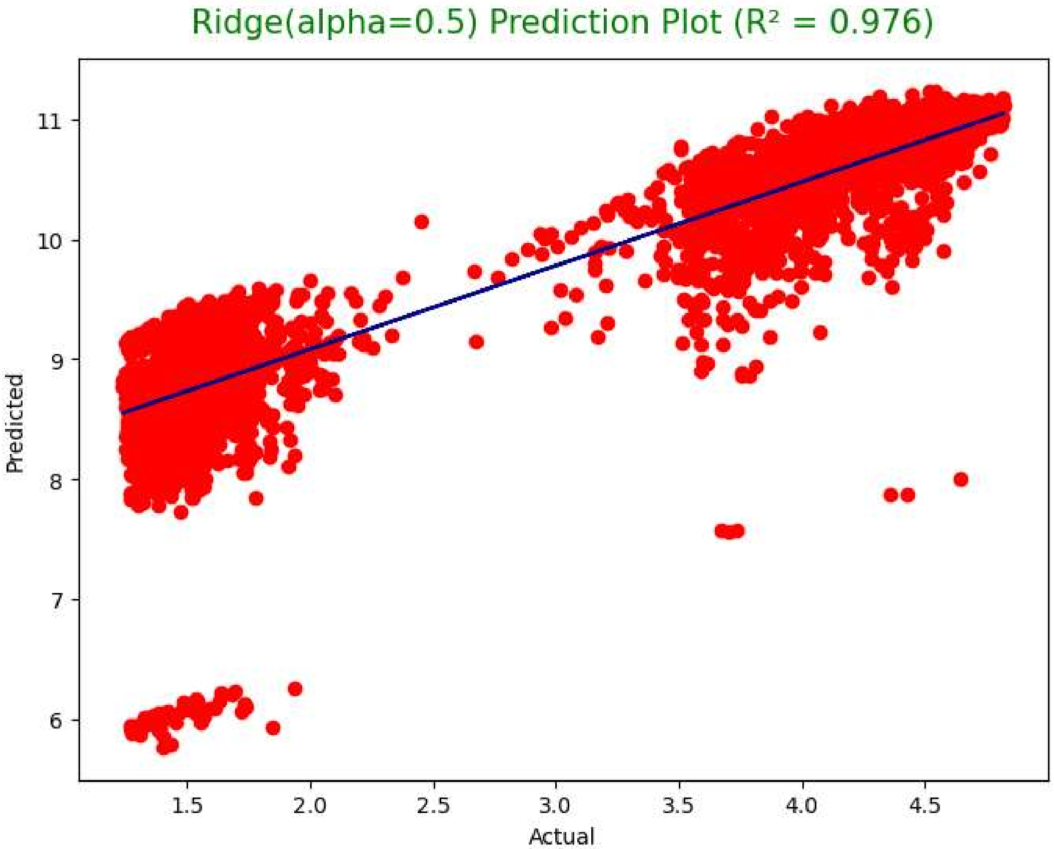

## Linear Regression Model (Plot 26)

The Linear Regression model achieved a high accuracy with a score of 0.976. The Mean Absolute Error (MAE) was 6.89, and the Root Mean Squared Error (RMSE) was 6.92, indicating a strong predictive capability. This model is straightforward and interpretable, making it suitable for stakeholders who need clear insights into energy consumption patterns.



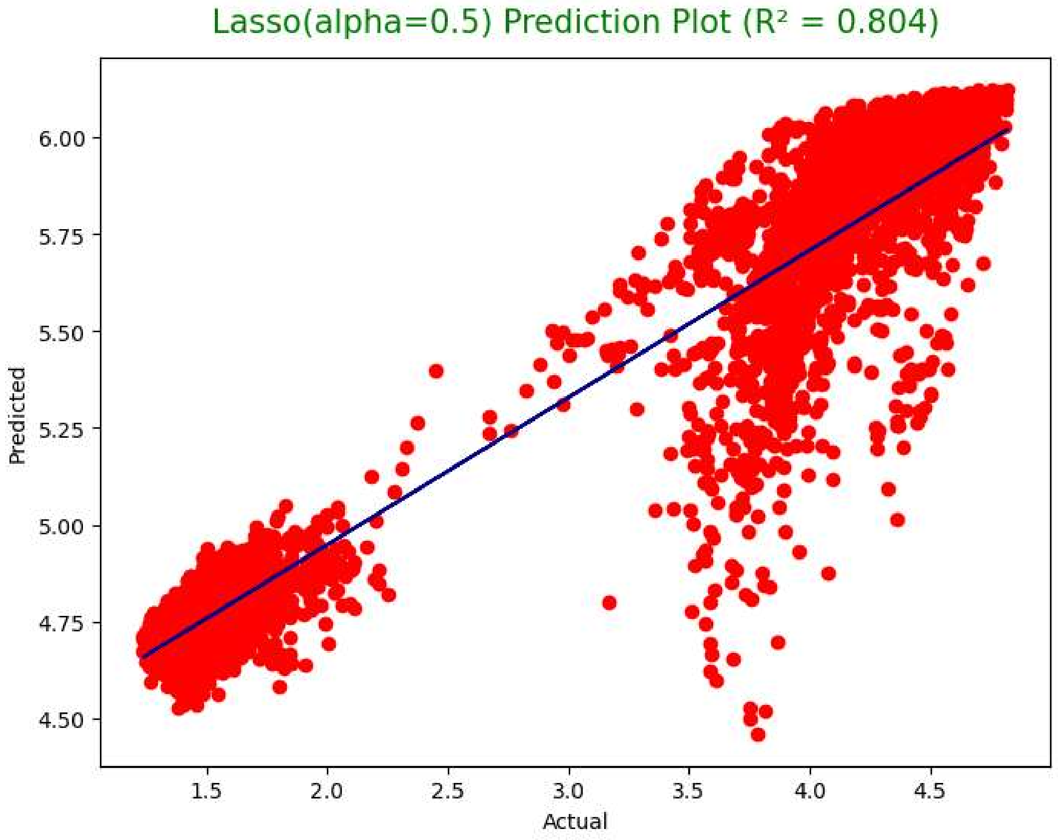

## Ridge Model (Plot 27)

The Ridge model also performed well with a score of 0.976. It had a slightly lower MAE of 6.89 and RMSE of 6.92, similar to the Linear Regression model. Ridge regression is beneficial when dealing with multicollinearity, as it adds a penalty to the coefficients, reducing their variance. This makes it a robust choice for predicting energy consumption in complex systems.




## Lasso Model (Plot 28)

The Lasso model had a lower score of 0.804, with an MAE of 2.56 and RMSE of 2.69. While it is less accurate than the other models, Lasso is advantageous for feature selection due to its ability to shrink some coefficients to zero, effectively selecting a simpler model. This can be useful for identifying the most critical factors affecting energy consumption.




## Executive Summary

In evaluating the cost implications of using each model, it's important to consider both the accuracy and the interpretability of the models:

- **Linear Regression:** This model is cost-effective due to its simplicity and ease of implementation. It provides clear insights, making it suitable for quick decision-making. However, it may not handle complex relationships as effectively as other models.
- **Ridge Regression:** While slightly more complex, Ridge regression offers robustness against multicollinearity, which can be crucial in systems with interrelated variables. The cost of implementation is slightly higher, but it provides a balance between accuracy and complexity.
- **Lasso Regression:** Lasso is beneficial for feature selection, potentially reducing costs by identifying the most impactful variables. However, its lower accuracy compared to the other models might lead to less precise predictions, which could affect decision-making.

Overall, the choice of model should align with the specific needs of the organization, balancing the trade-offs between cost, accuracy, and interpretability.