Appliance Energy Prediction with Supervised Machine Learning & Forecasting

Abdoulfatah Abdillahi

May 2025

1. Executive Summary

This study analyzes residential energy consumption using the Appliance Energy Prediction dataset, which captures 10-minute interval data over 4.5 months from a Belgian household. Indoor temperature and humidity were collected via a ZigBee wireless sensor network and aggregated into 10-minute averages, while appliance energy use was recorded using m-bus meters. External weather conditions (e.g., temperature, humidity, pressure) were merged from Chievres Airport, Belgium, via Reliable Prognosis (rp5.ru). To ensure model robustness, two random variables were included to test regression algorithms' ability to filter non-predictive features. This high-resolution dataset enables granular analysis of intraday energy patterns and their relationship to environmental drivers.

The study addresses two critical challenges: (1) forecasting hourly energy use for May 2016 and (2) predicting appliance energy consumption via regression modeling. By optimizing energy forecasts, this work aims to support cost-effective grid management and sustainability efforts.

For the first objective, we decomposed the Appliances energy series into trend, seasonality, and residuals using STL decomposition. This revealed an upward/downward long-term trend, strong daily seasonality peaking at hourly and daily seasonality, and irregular residuals suggesting external factors not captured by temporal patterns. The SARIMAX model captures some of the hourly and daily seasonality in appliance energy usage, particularly through a strong seasonal moving average component. However, there are signs of residual

autocorrelation, non-normality, and heteroskedasticity, suggesting that the model does not fully explain the variability in the data. The model could be refined further, or complemented with other models.

For the second objective, we trained regression models (including Random Forest, XGBoost, and LASSO) to predict appliance energy use using all features. The regression approach outperformed pure time-series forecasting, underscoring the value of integrating environmental data.

Finally, I have learned how to forecast in a time series and classification task using regression techniques.

2. Introduction

Energy consumption in residential buildings is significantly influenced by environmental factors such as temperature, humidity, and weather conditions. Understanding these relationships is essential for improving energy efficiency and promoting sustainable living. This project focuses on the Appliance Energy Prediction dataset, which was collected from a low-energy residential building using a ZigBee wireless sensor network. The dataset includes a variety of environmental measurements such as indoor and outdoor temperature, humidity, and weather data alongside time-stamped appliance energy usage.

The motivation for this project is both educational and practical. I chose this dataset because I had not previously worked with time series data, and this project gave me the opportunity to explore it in depth and strengthen my understanding of regression tasks. On the practical side, developing accurate models to predict energy consumption can support the design of more efficient and intelligent energy management systems. On the educational side, this project allowed me to apply a range of techniques including supervised machine learning,

time series forecasting, and regression modeling building on my background in classification while expanding into less familiar territory.

3. Data Description, Cleaning & Preparation

The data originates from the UC-Irvine Machine Learning Repository and consists of 19,735 observations with 29 features, with the target variable being appliance energy consumption.

Variable Name	Role	Type	Description	Missing Values
Date	Feature	Date	Date(YY-DD-HH-SS)	No
Appliance	Target	Integer	Energy consumption (Wh) of appliances	No
Lights	Feature	Integer	Energy consumption of lights (in Wh)	No
T1	Feature	Continuous	Temperature in living room	No
RH_1	Feature	Continuous	Relative Humidity (%) in living room	No
T2	Feature	Continuous	Temperature in kitchen	No
RH_2	Feature	Continuous	Relative Humidity (%) in kitchen	No
T3	Feature	Continuous	Temperature in laundry room	No
RH_3	Feature	Continuous	Relative Humidity (%) in laundry room	No
T4	Feature	Continuous	Temperature in office room	No

RH_4	Feature	Continuous	Relative Humidity (%) in office room	No
T5	Feature	Continuous	Temperature in bathroom	No
RH_5	Feature	Continuous	Relative Humidity (%) in bathroom	No
T6	Feature	Continuous	Temperature (°C) in north-facing rooms	No
RH_6	Feature	Continuous	Relative Humidity (%) in north-facing rooms	No
T7	Feature	Continuous	Temperature (°C) in ironing room	No
RH_7	Feature	Continuous	Relative Humidity (%) in ironing room	No
T8	Feature	Continuous	Temperature (°C) in teen room 1	No
RH_8	Feature	Continuous	Relative Humidity (%) in teen room 1	No
T9	Feature	Continuous	Temperature (°C) in parents' room	No
RH_9	Feature	Continuous	Relative Humidity (%) in parents' room	No
T_out	Feature	Continuous	Outdoor temperature (°C)	No
Press_mg	Feature	Continuous	Atmospheric pressure (mm Hg)	No
RH_out	Feature	Continuous	Outdoor relative humidity (%)	No
Windspeed	Feature	Continuous	Wind speed (m/s)	No
Visibility	Feature	Continuous	Visibility (km)	No
Tdewpoint	Feature	Continuous	Dew point temperature (°C)	No

Rv1 & rv2	Feature	Continuous	Random variables with no specific meaning (for testing)	No

Table 1: *Variable Description Table*

3.1 Exploratory Data Analysis (EDA)

I examined the dataset through four stages of analysis: univariate non-graphical, univariate graphical, multivariate non-graphical, and multivariate graphical methods.

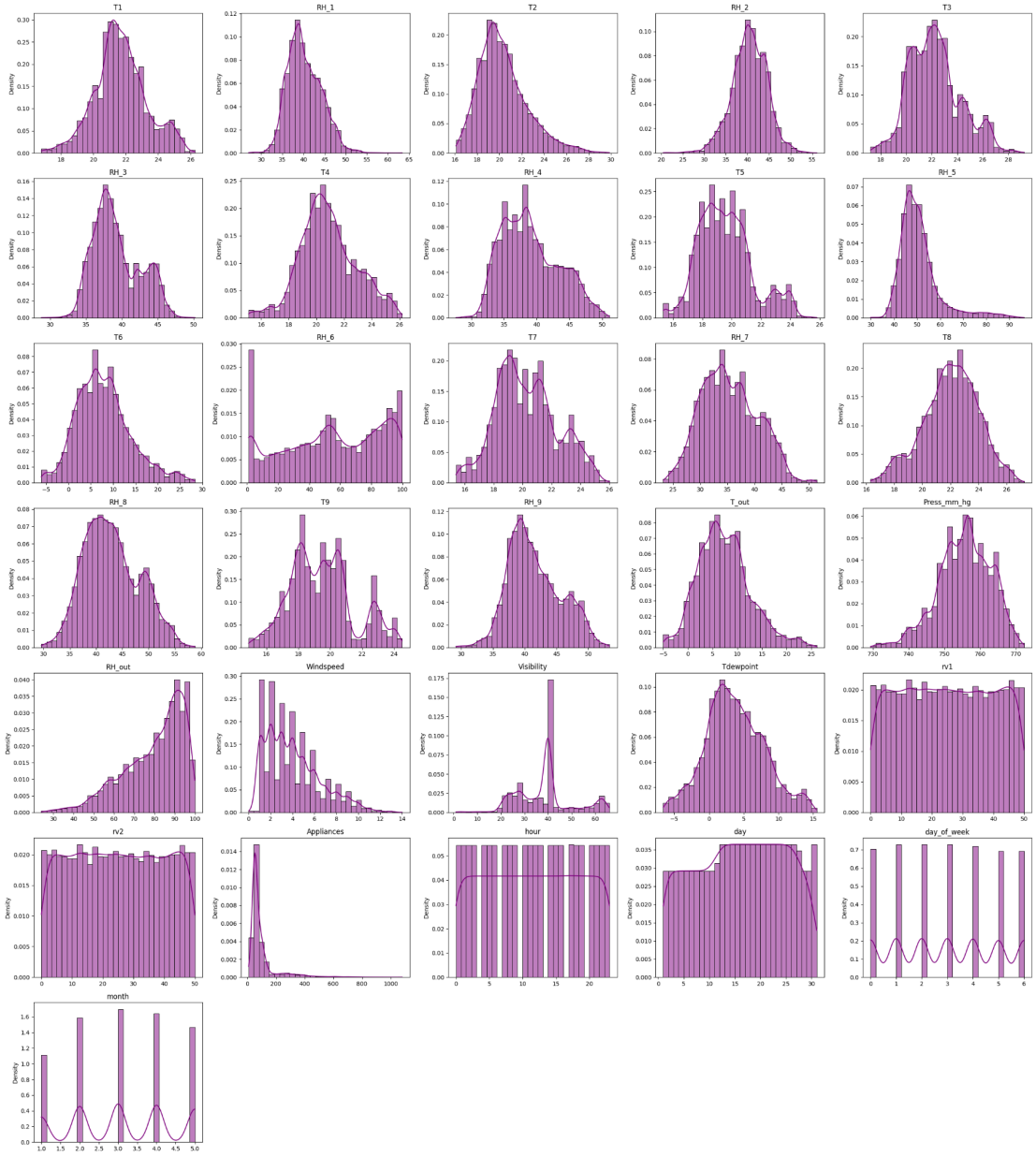

Figure 1: *Description of Features Distributions*

The temperature and humidity features generally follow a Gaussian-like distribution. The target variable, ‘Appliances’, is a right-skewed distribution, suggesting the presence of outliers with high energy consumption. Based on the figure 1, features such as hour, day, day of the month, month, visibility, RH_6, and windspeed do not follow a normal distribution, while the remaining features show distributions that are approximately normal.

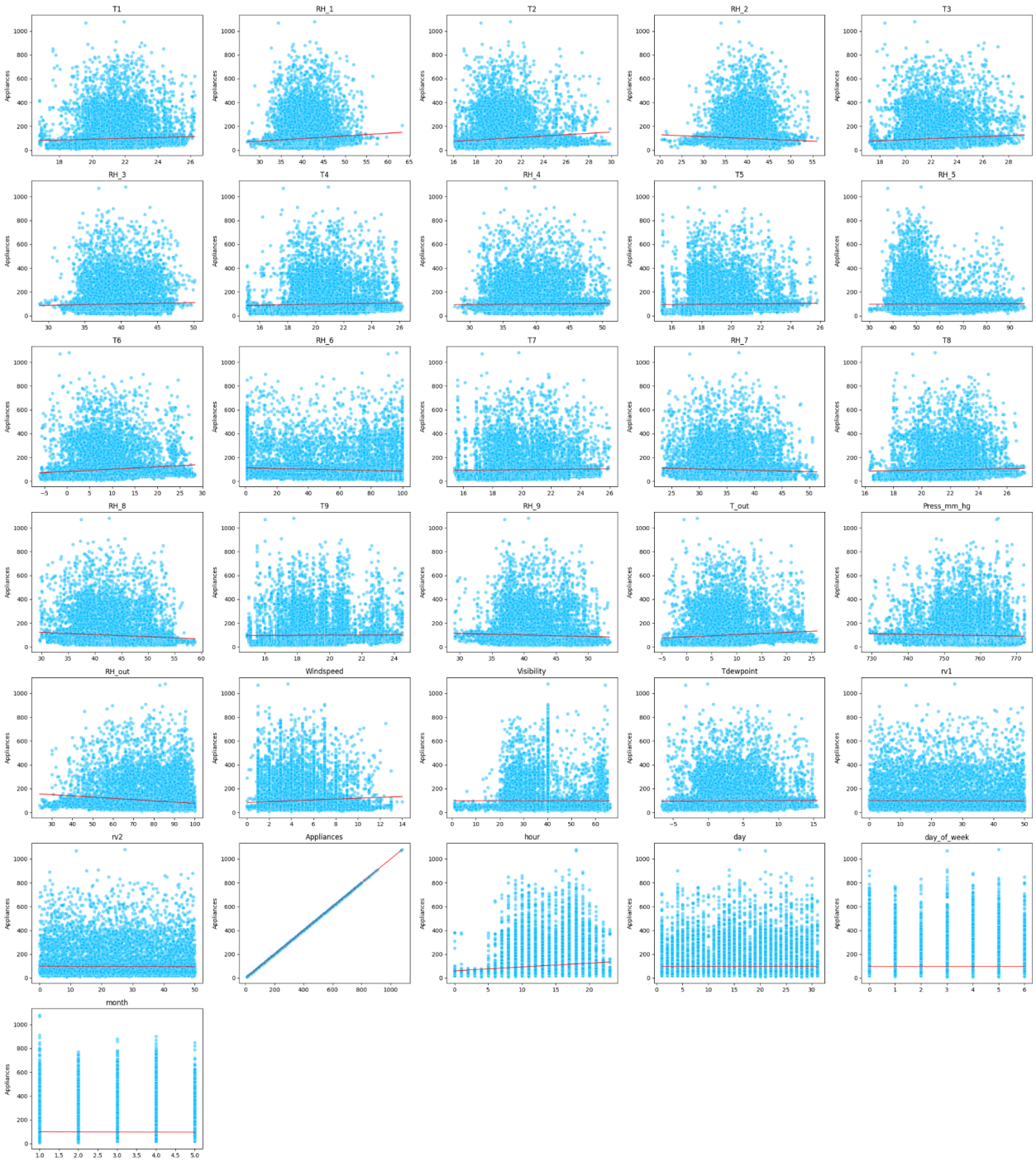

Figure 2: *Scatter Relationships Between Target and Predictors*

The target variable, ‘Appliances’, was analyzed for its relationship with the independent features using scatter plots. Most variables showed weak or no clear linear correlation with the target

data points were widely dispersed, and the trend lines were mostly flat or had only a slight slope. No strong predictive relationship was observed, except for the target's own distribution.

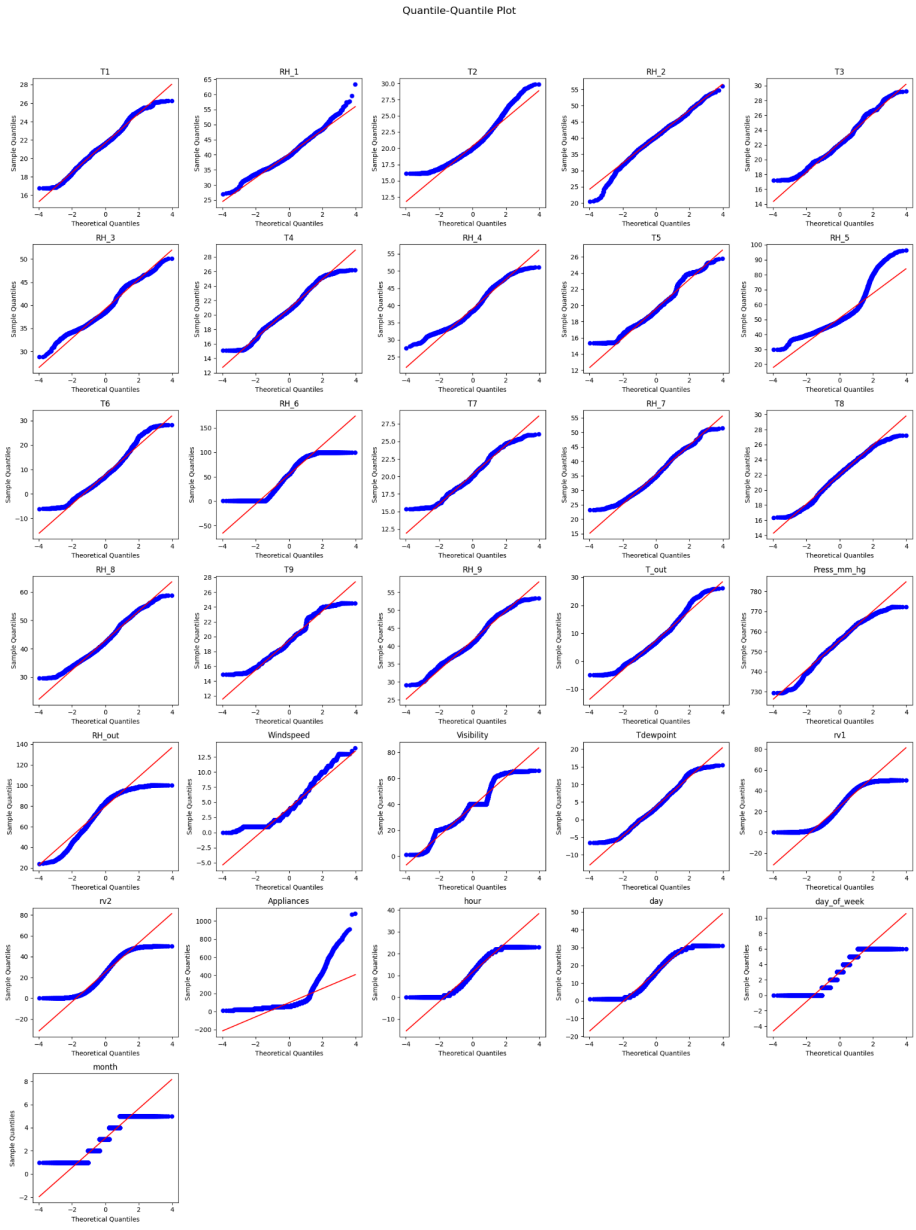

Figure 3: *Qantile-Quantile plot of the Features*

The Q-Q plots help assess if each feature follows a normal distribution. Features like T1 align well with the diagonal line, suggesting normality, while the target variable Appliances deviates strongly, indicating right-skewness. This is important because models like linear regression assume normality, so skewed features may need transformation or alternative modeling approaches.

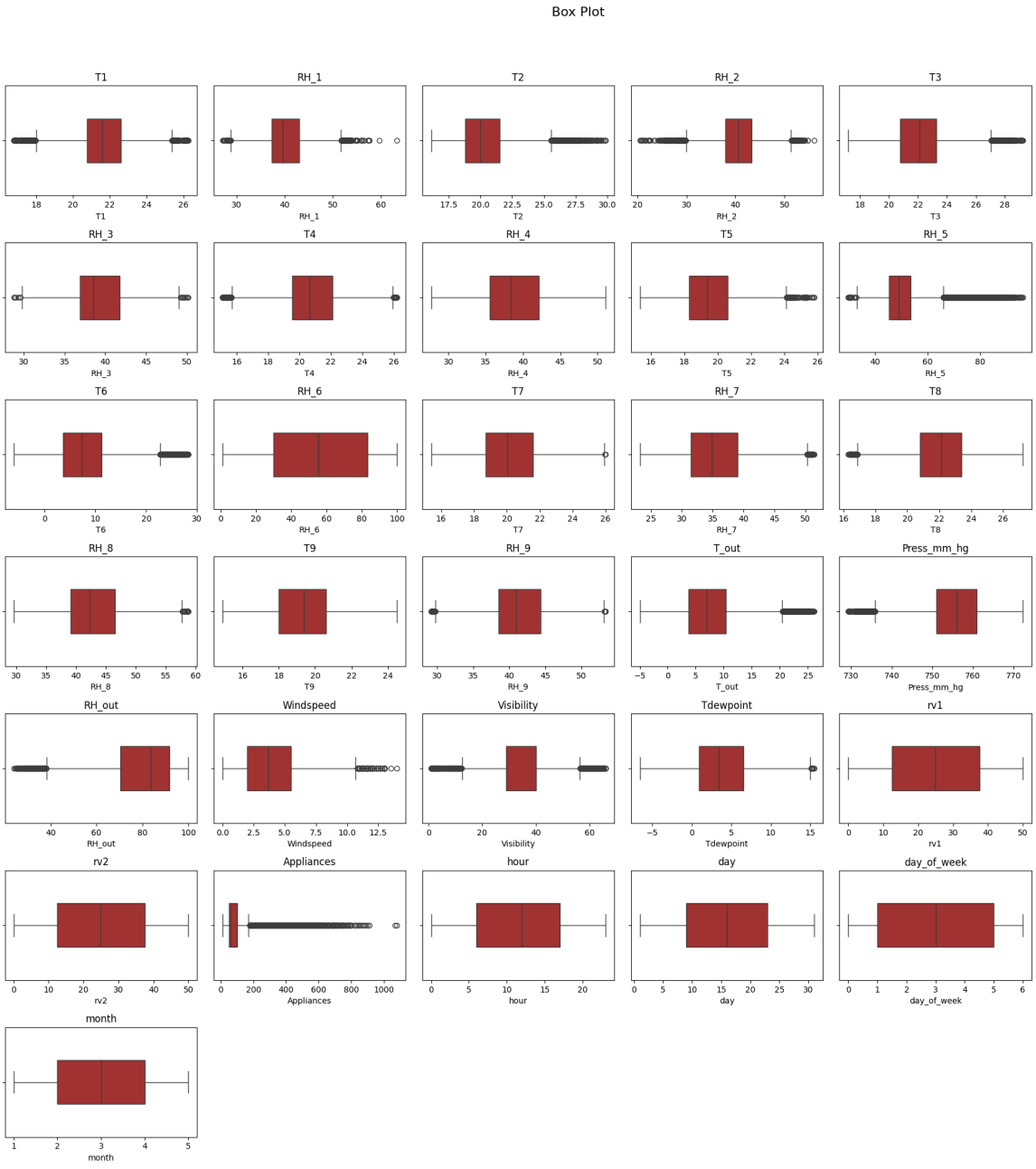

Figure 4: *The Spread and Central Tendency of the Features*

The box plot reveals that most features contain outliers, indicating variability in the data that could potentially affect model performance. These outliers may result from measurement noise, real-world variation, or anomalies in sensor data. For example, the feature `T1` shows clear outliers, while `hour` does not contain any, suggesting that some features are more stable than others. Identifying such characteristics helps inform data preprocessing steps, such as outlier handling or robust scaling.

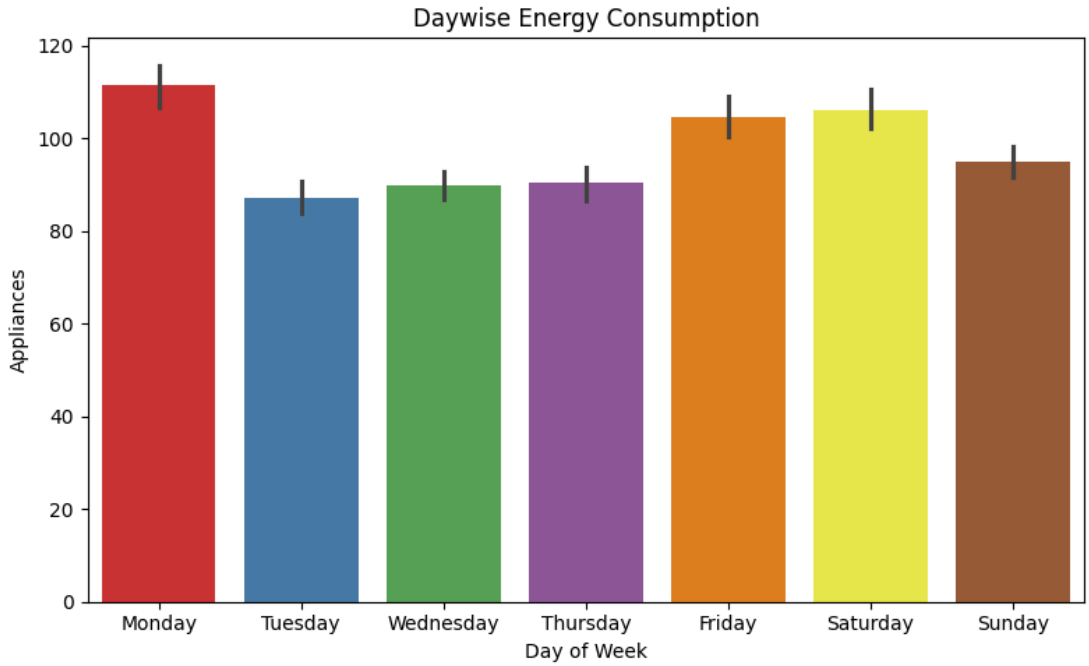

Figure 5: *Day of the Week Energy Usage*

The bar plot illustrates the average appliance energy usage across each day of the week. Energy consumption peaks on Monday, followed by Friday and Saturday, suggesting that people tend to spend more time at home on these days. In contrast, lower usage during other weekdays may indicate that residents are typically away from home during those times.

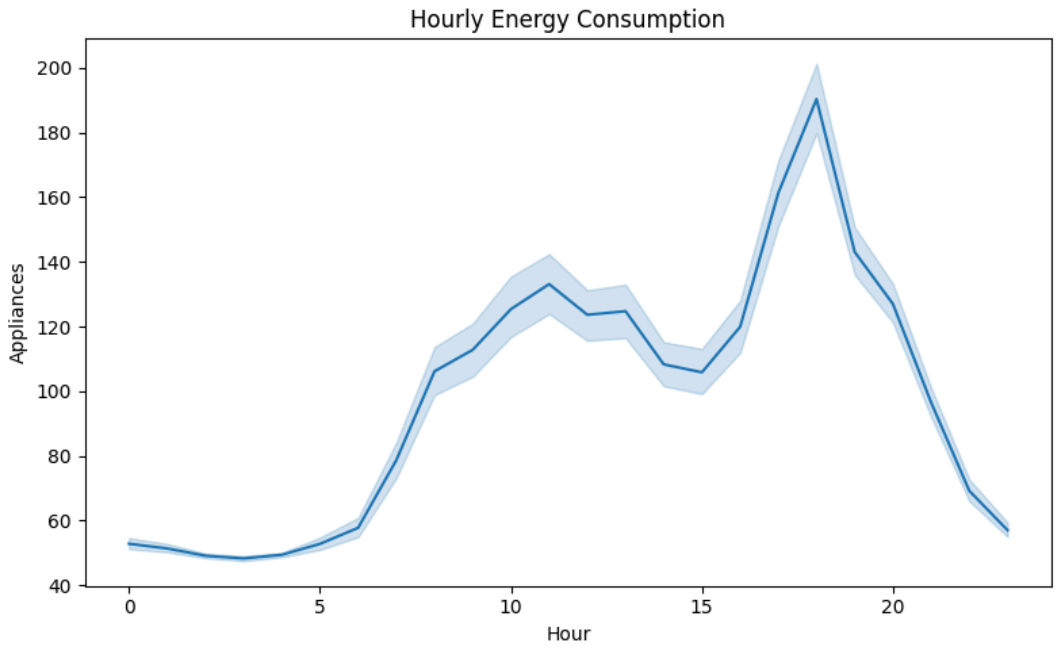

Figure 6: *Hourly Energy Consumption*

The line plot displays average hourly appliance energy consumption throughout the day. Energy usage starts to rise around 6 AM, with a noticeable peak between 10–11 AM, followed by fluctuations during the afternoon. The highest spike occurs around 6–7 PM, likely reflecting evening routines when more appliances are used at home. This pattern suggests typical daily activity cycles, with higher energy demand in the morning and evening hours.

3.2 Data Preprocessing

Part I : Forecasting

Initially, the date column contained formatting issues that prevented proper parsing and resampling. The date strings used non-standard dash characters and lacked a separator between the date and time components. I replaced en dashes with standard hyphens for consistency, and I inserted a ‘T’ character between the data and tme to maich the datetime

format and finally converted the cleaned strings into proper datetime objects using pandas. This allowed me to resample the data at an hourly frequency for time series analysis.

The second step I took was to resampled to hourly intervals using mean energy consumption within each hour. I applied seasonal decomposition using an additive model with a period of 24, reflecting daily seasonality. This process separates the series into three components:

- Trend: Captures the long-term movement in appliance energy usage over time.
- Seasonality: Reflects the repeating daily patterns of consumption, such as higher usage during specific hours.
- Residuals: Represents the random noise or irregular variations not explained by trend or seasonality.

This decomposition provided a clearer understanding of temporal patterns in the data, which helped inform modeling decisions and improved interpretability.

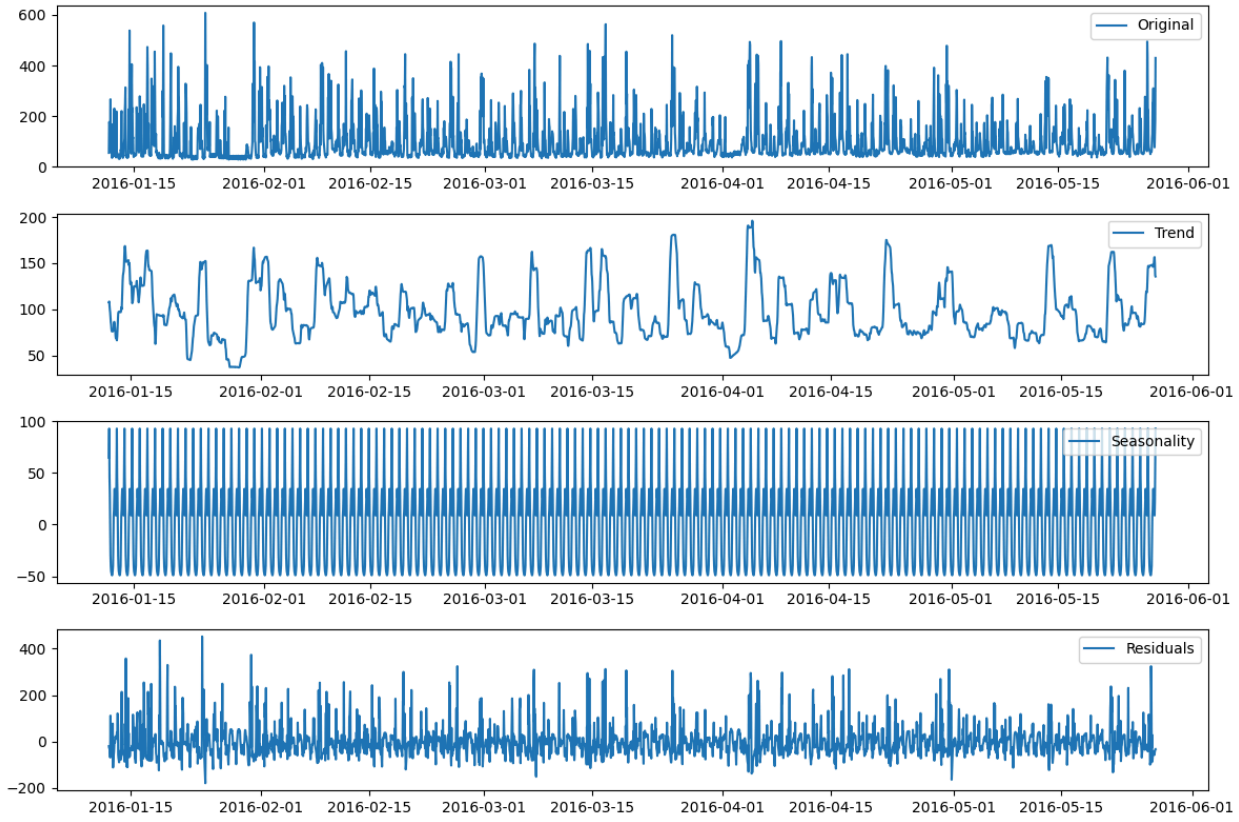

Figure 7: *Seasonal Decomposition*

In the figure as you can see the Trend shows the long-term pattern of energy usage across the dataset. It does show periods of gradually increasing and decreasing energy consumption. For instance you can observe a slight upward trend in early April and late May, indicating increased appliance use during those times, possibly due to seasonal changes like warming weather.

The seasonality component reflects repeating daily patterns with a period of 24 hours. This consistent high-frequency oscillation shows that energy usage follows a strong daily cycle.

The residuals capture random noise that are not explained by the trend and seasonality. While mostly centered around zero, there are some large spikes, indicating occasional unusual energy usage events.

Part II : Regression Preprocessing and Feature Engineering =

To prepare the modeling the dataset, I have used different methods to preprocessed the dataset:

1. Categorical Features Encoding
    - Categorical variables were one-hot encoded using `pd.get_dummies()` with `drop_first=True` to avoid multicollinearity.
    - Boolean features were also converted to binary format (0/1).
2. Outlier Capping (IQR Method)
    - The Interquartile Range (IQR) method was used to cap outliers in continuous variables. Values beyond 1.5 times the IQR from the first and third quartiles were replaced with the respective upper or lower bound.
3. Datetime Feature Extraction
    - From the original datetime column, additional features were derived, including hour, day, day of the week, month, and their respective names (e.g., month name, day name) to capture potential temporal patterns.
4. Face of the Day Classification
    - A new feature, ‘face_of_day’, was created to categorize hours into Morning, Afternoon, Evening, and Night, enhancing time-based pattern detection.
5. Low-Variance Feature Removal
    - Features with very low variance (below a specified threshold, default = 0.01) were removed using VarianceThreshold, as they contribute little to model performance.
6. F-value Feature Selection (ANOVA F-test)

- To retain the most predictive features, I applied univariate feature selection using the F-test for regression. Features with F-values above a threshold (e.g., 30) were selected, indicating a stronger linear relationship with the target variable.

3.3 Model Training & Evaluation

To assess model performance and identify the most suitable algorithm for predicting energy consumption, I implemented and compared multiple regression models: Linear Regression, Decision Tree, and Random Forest.

Linear Regression

- Used as a baseline model to understand general trends in energy usage.
- Applied Ridge and Lasso regularization to improve generalization and reduce overfitting.
- Although performance across linear, Ridge, and Lasso models was similar, regularized models offered greater stability.
- All models were validated using 10-fold cross-validation to ensure robustness.

Decision Tree

- Implemented to explore non-linear relationships between features and the target.
- Used GridSearchCV to tune key hyperparameters, including:
  - `max_depth`: [3, 5, 10, None]
  - `min_samples_split`: [2, 5, 10]
  - `min_samples_leaf`: [1, 2, 4]
- This model demonstrated improved performance over linear models and effectively captured interactions and non-linear patterns.
- To avoid overfitting, tree depth was restricted to a maximum of 5.

Random Forest

- Deployed to further enhance accuracy by aggregating multiple decision trees.
- Configured with 100 estimators and parallel computation (`n_jobs = -1`) for efficiency.
- Delivered the best overall performance among all tested models.
- Random Forest proved to be the most reliable and accurate model for predicting energy usage.

Model	RMSE	MAE	R^2	MAPE (%)
Linear Regression	92.75	54.45	0.14	66.96
Ridge Regression	92.75	54.45	0.14	66.95
Lasso Regression	92.74	54.44	0.14	66.95
Decision Tree	90.72	50.0	0.18	58.58
Grid Search Tree	85.04	37.94	0.28	35.65
Random Forest	63.1	29.55	0.6	29.54

Table 2: *Model Performance Comparisons*

4. Conclusion

- I explored five regression models, Linear Regression, Ridge, Lasso, Decision Tree, and Random Forest to predict energy consumption.
- Comprehensive data preprocessing was performed, including outlier treatment, feature engineering, encoding, and feature selection.
- Models were evaluated using key performance metrics: RMSE, MAE, R^2, and MAPE.

- The linear-based models (Linear, Ridge, and Lasso) assigned negligible weights to most predictors, revealing limited predictive power for this problem.
- The Random Forest Regressor achieved the highest performance, with an R^2 score of 0.64, indicating a strong ability to capture complex relationships in the data and predict energy use effectively.

5. References:

- Abdoulfatah Abdillahi . (n.d.). *Appliance Power ML [GitHub repository]*. GitHub. Retrieved from : https://github.com/Abdoul1996/appliance-power-ml
- UCI Machine Learning Repository. (n.d.). *Appliances energy prediction data set*. Retrieved from: https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction
- George, T. (n.d.). *Appliance Energy Prediction [GitHub repository]*. GitHub. Retrieved from : https://github.com/Thomas-George-T/Appliance-Energy-Prediction
- Ali, M., Junaid, M., & Aslam, S. (2023). *Appliance Energy Consumption Prediction Using Ensemble Learning*. Insights in Computer Research, 2(2), 69–76. Retrieved from https://journals.umt.edu.pk/index.php/icr/article/view/2856