Table of Contents

Introduction

The Global Appliance Energy prediction data set from the repository of machine learning presents high-resolution to household energy usage, the data arriving at intervals of 10 minutes from one house in Belgium over an about 4.5-month time span. The data comprises 29 features, from indoor and outdoor climatic conditions (ambient temperature, humidity), weather conditions (wind speed, visibility, pressure), and time variables (hour, weekday/day of the week). The response variable would be appliances, which accounts for the home appliances' energy usage, making this data set an excellent model for real-time energy usage analysis and model building.

This data was selected to analyse because it can have real-world application in the field of energy sustainability and smart homes. As there is an increased demand for IoT devices and rising costs of electricity, it is now necessary for households and utilities to know how much they are consuming. This data can be utilized by smart home and smart metering to plan devices effectively, manage peak loads, and make personalized suggestions for conserving energy (Chen et al., 2023). Such concepts such as predictive models plotted from data also have the potential to inform prices, load forecasting, and infrastructure planning (Ga et al., 2024).

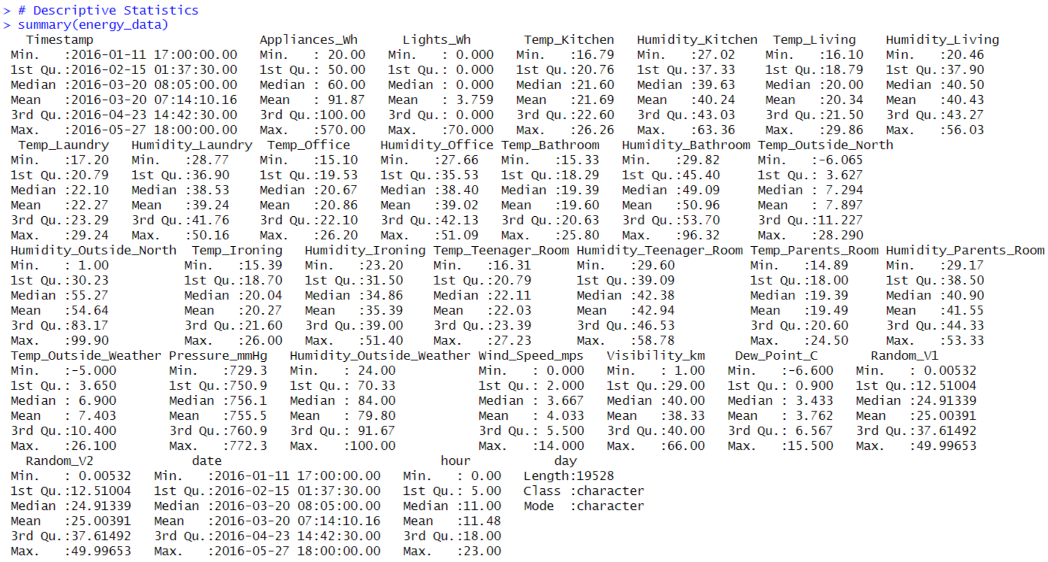
Business value is critical, and that value is realized here through potential to enable demand-side energy efficiency. Through the identification of electricity energy usage determinants of significance—such as time of day, room-by-room temperatures, and weather—dwellings and institutions can design more responsive control systems and conserving smart systems to ultimately promote and support consumer behaviour transformation. The purpose of this study is to provide robust statistical models and derive actionable conclusions with the potential to inform policy and technology innovation in home energy (Himeur et al., 2021).

Descriptive Statistics and Preliminary Correlation Analysis

Descriptive statistics is the business analytics technique that summarizes and consolidates complex data sets allowing us to identify trends, uncover outliers and form hypotheses to explore and reveal insights with the support of measures like mean, median, standard deviation and inter-quartile ranges. The insights can provide data which can inform operations as well as strategic decision-making.

Summary Statistics

Descriptive summary statistics were generated to summarise variables (Figure 1), to determine relationships and explore each variable and its diversity. This dataset comprises of the electricity usage within different spaces in home along with the corresponding data on environmental conditions.



```
> # Descriptive Statistics
> summary(energy_data)
   Timestamp                       Appliances_Wh      Lights_Wh       Temp_Kitchen   Humidity_Kitchen  Temp_Living     Humidity_Living
 Min.   :2016-01-11 17:00:00.00   Min.   : 20.00   Min.   : 0.000   Min.   :16.79   Min.   :27.02    Min.   :16.10   Min.   :20.46  
 1st Qu.:2016-02-15 01:37:30.00   1st Qu.: 50.00   1st Qu.: 0.000   1st Qu.:20.76   1st Qu.:37.33    1st Qu.:18.79   1st Qu.:37.90  
 Median :2016-03-20 08:05:00.00   Median : 60.00   Median : 0.000   Median :21.60   Median :39.63    Median :20.00   Median :40.50  
 Mean   :2016-03-20 07:14:10.16   Mean   : 91.87   Mean   : 3.759   Mean   :21.69   Mean   :40.24    Mean   :20.34   Mean   :40.43  
 3rd Qu.:2016-04-23 14:42:30.00   3rd Qu.:100.00   3rd Qu.: 0.000   3rd Qu.:22.60   3rd Qu.:43.03    3rd Qu.:21.50   3rd Qu.:43.27  
 Max.   :2016-05-27 18:00:00.00   Max.   :570.00   Max.   :70.000   Max.   :26.26   Max.   :63.36    Max.   :29.86   Max.   :56.03  
  Temp_Laundry   Humidity_Laundry  Temp_Office     Humidity_Office Temp_Bathroom    Humidity_Bathroom Temp_Outside_North
 Min.   :17.20   Min.   :28.77    Min.   :15.10   Min.   :27.66   Min.   :15.33   Min.   :29.82     Min.   :-6.065    
 1st Qu.:20.79   1st Qu.:36.90    1st Qu.:19.53   1st Qu.:35.53   1st Qu.:18.29   1st Qu.:45.40     1st Qu.: 3.627    
 Median :22.10   Median :38.53    Median :20.67   Median :38.40   Median :19.39   Median :49.09     Median : 7.294    
 Mean   :22.27   Mean   :39.24    Mean   :20.86   Mean   :39.02   Mean   :19.60   Mean   :50.96     Mean   : 7.897    
 3rd Qu.:23.29   3rd Qu.:41.76    3rd Qu.:22.10   3rd Qu.:42.13   3rd Qu.:20.63   3rd Qu.:53.70     3rd Qu.:11.227    
 Max.   :29.24   Max.   :50.16    Max.   :26.20   Max.   :51.09   Max.   :25.80   Max.   :96.32     Max.   :28.290    
 Humidity_Outside_North  Temp_Ironing    Humidity_Ironing Temp_Teenager_Room Humidity_Teenager_Room Temp_Parents_Room Humidity_Parents_Room
 Min.   : 1.00          Min.   :15.39   Min.   :23.20    Min.   :16.31      Min.   :29.60          Min.   :14.89     Min.   :29.17        
 1st Qu.:30.23          1st Qu.:18.70   1st Qu.:31.50    1st Qu.:20.79      1st Qu.:39.09          1st Qu.:18.00     1st Qu.:38.50        
 Median :55.27          Median :20.04   Median :34.86    Median :22.11      Median :42.38          Median :19.39     Median :40.90        
 Mean   :54.64          Mean   :20.27   Mean   :35.39    Mean   :22.03      Mean   :42.94          Mean   :19.49     Mean   :41.55        
 3rd Qu.:83.17          3rd Qu.:21.60   3rd Qu.:39.00    3rd Qu.:23.39      3rd Qu.:46.53          3rd Qu.:20.60     3rd Qu.:44.33        
 Max.   :99.90          Max.   :26.00   Max.   :51.40    Max.   :27.23      Max.   :58.78          Max.   :24.50     Max.   :53.33        
 Temp_Outside_Weather Pressure_mmHg    Humidity_Outside_Weather Wind_Speed_mps    Visibility_km     Dew_Point_C        Random_V1       
 Min.   :-5.000       Min.   :729.3   Min.   : 24.00           Min.   : 0.000   Min.   : 1.00   Min.   :-6.600   Min.   : 0.00532  
 1st Qu.: 3.650       1st Qu.:750.9   1st Qu.: 70.33           1st Qu.: 2.000   1st Qu.:29.00   1st Qu.: 0.900   1st Qu.:12.51004  
 Median : 6.900       Median :756.1   Median : 84.00           Median : 3.667   Median :40.00   Median : 3.433   Median :24.91339  
 Mean   : 7.403       Mean   :755.5   Mean   : 79.80           Mean   : 4.033   Mean   :38.33   Mean   : 3.762   Mean   :25.00391  
 3rd Qu.:10.400       3rd Qu.:760.9   3rd Qu.: 91.67           3rd Qu.: 5.500   3rd Qu.:40.00   3rd Qu.: 6.567   3rd Qu.:37.61492  
 Max.   :26.100       Max.   :772.3   Max.   :100.00           Max.   :14.000   Max.   :66.00   Max.   :15.500   Max.   :49.99653  
   Random_V2             date                               hour            day           
 Min.   : 0.00532   Min.   :2016-01-11 17:00:00.00   Min.   : 0.00   Length:19528      
 1st Qu.:12.51004   1st Qu.:2016-02-15 01:37:30.00   1st Qu.: 5.00   Class :character  
 Median :24.91339   Median :2016-03-20 08:05:00.00   Median :11.00   Mode  :character  
 Mean   :25.00391   Mean   :2016-03-20 07:14:10.16   Mean   :11.48                     
 3rd Qu.:37.61492   3rd Qu.:2016-04-23 14:42:30.00   3rd Qu.:18.00                     
 Max.   :49.99653   Max.   :2016-05-27 18:00:00.00   Max.   :23.00                     
```

Figure 1.1 Overall Summary Statistics with Renamed Columns

Histogram analysis (Appliances of Energy use)

The Appliance energy usage was observed using a histogram overlapped by a right-skewed curve (figure 2). The approximate mean value interpreted was 92.0*wh* which portrays the deviation from normality, validating the transformations and the outlier solving procedures. The skewness is influenced by the extreme values also denoting the majority energy usage lies below 200 Wh.

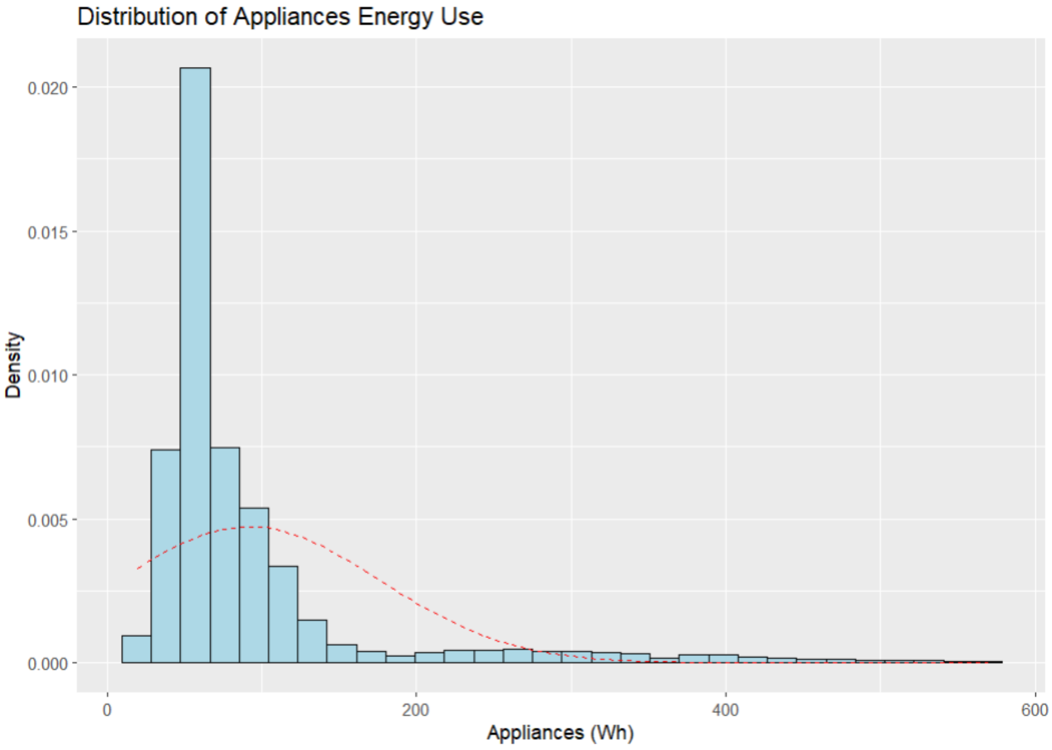


Figure 2 – Histogram Analysis signifying right skewness

Correlation analysis

The correlation matrix (figure 3) signifies every relationship between every variable that is represented by a colour palette as a legend. This helps with the multicollinearity risks and the strong predictors for our primary variable.

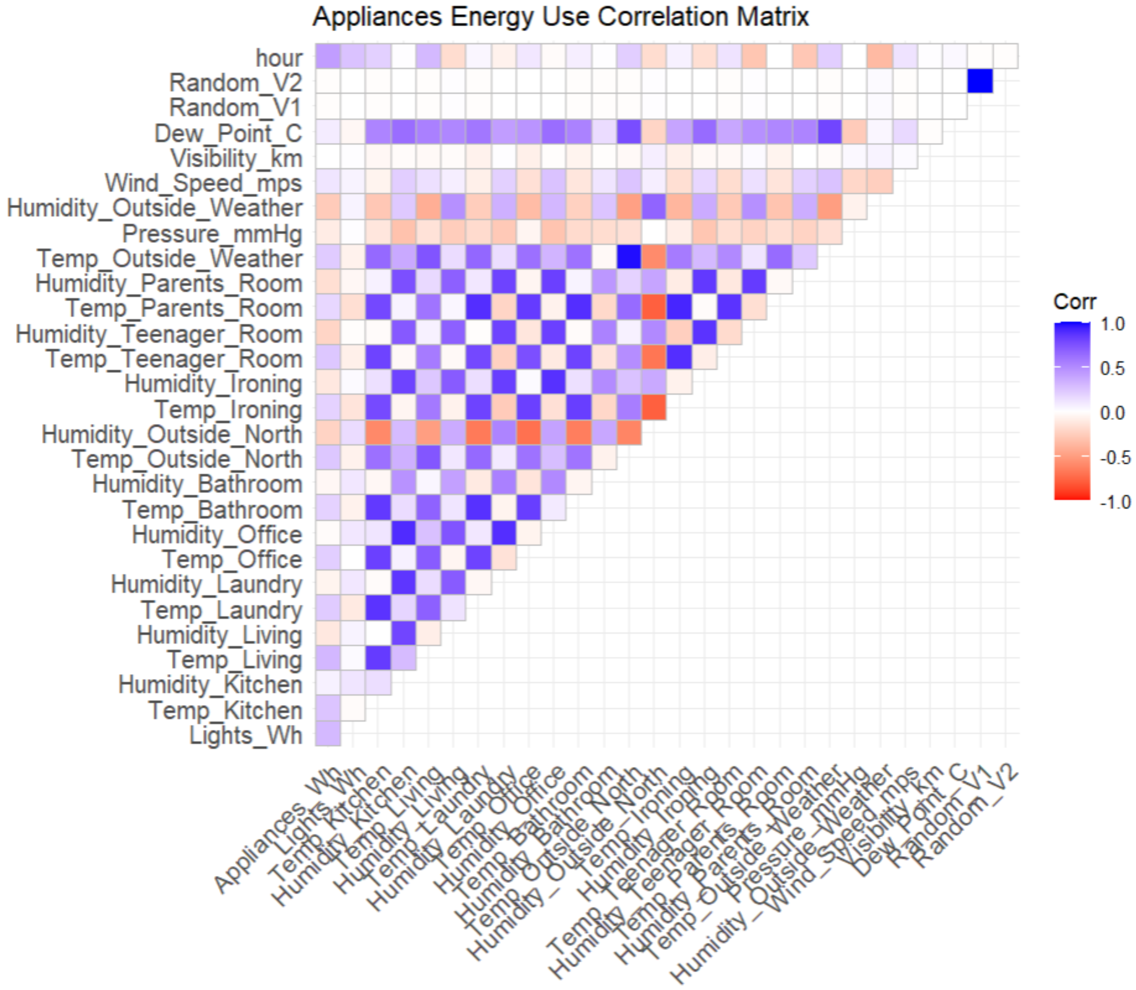


Figure 3 – Correlation Spearman Matrix

Boxplot analysis (Appliances by Hour)

The boxplots are the best for outlier understanding. Here (figure 4), it explains the appliance usage patterns, clearly showing how the energy usage is higher during the day hours.

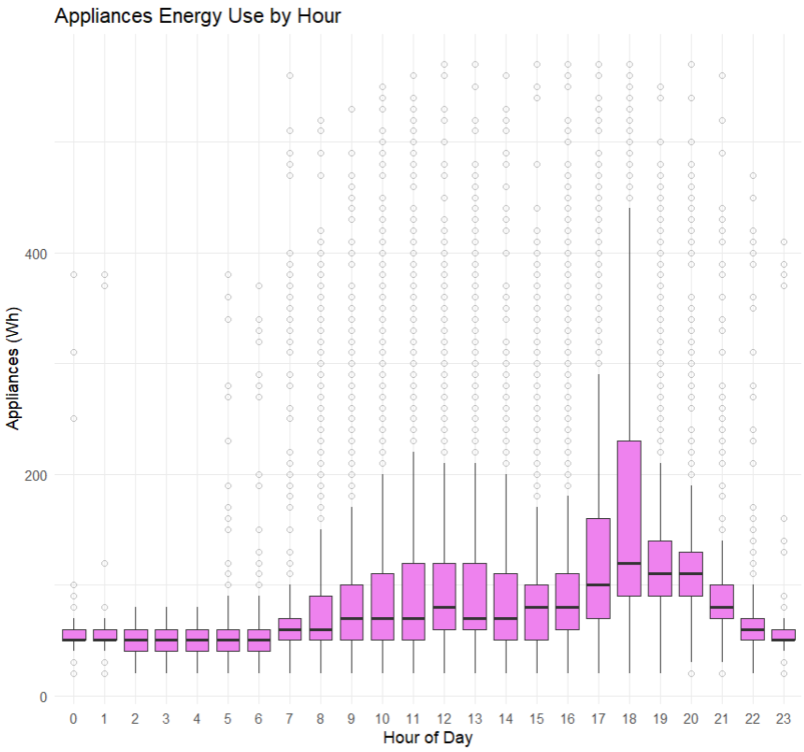


Figure 4 – Boxplot Analysis to understand Interquartile Range

Density plot analysis (Appliances by Hour)

The density plot (figure 5) illustrates the smoothened distribution with a superimposed normal curve of the appliance usage, supporting the skewness.

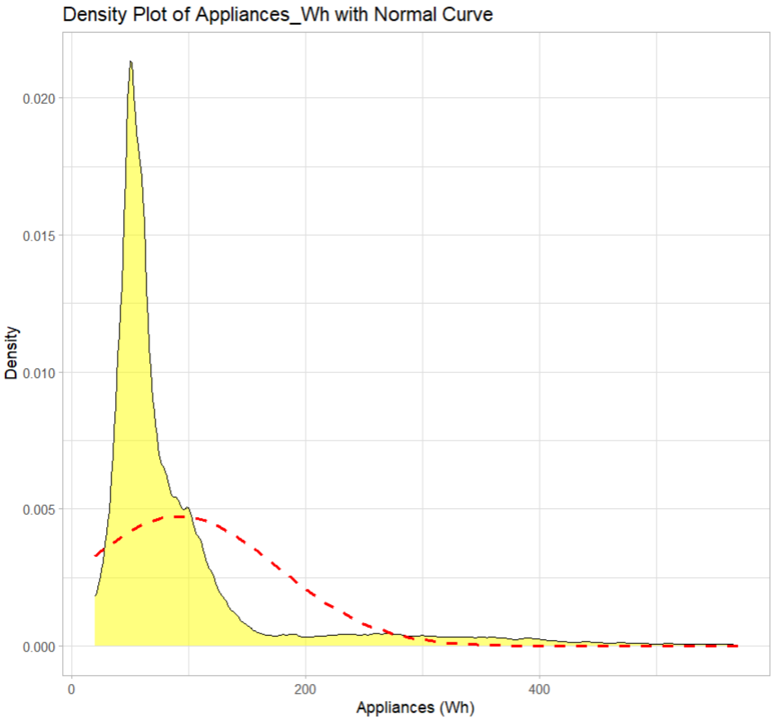


Figure 5 – Density Plot of Appliances usage

Overall, the descriptive statistics says that the household energy usage has high variability and influenced by multiple environmental and temporal factors. This variability emphasizes the need for optimised strategies in smart home systems.

Methodology

Multiple linear regression model is utilized for this practice to predict the usage of energy by household appliances, and if they are subject to any external factors. This dataset is high-level, with the duration being long and multi-variate, with independent factors such as temperature and time factors. Assumptions of the model called L.I.N.E. are also applied – Linearity, Independence of errors, homoscedasticity and Normality, which enable meaningful interpretation of the output.

The first operation carried out in the dataset is cleaning and reshaping to determine the statistically relevant model to use. Both **glimpse()** and **str()** were used for getting accustomed to the data. Correlation, confidence intervals, and descriptive statistics were carried out at initial explorations with the aim of approximating the significant predictors. Top and bottom 1% scores are trimmed out to eliminate outliers in the dependent variable. Missing values were checked absent with **vis_miss()** to maintain data cleanliness. The **groupby()** function was considered to put the temperatures and humidities together; but as the data suggests distinct locations, singularity would cause errors.

The initial MLR models comprised all the numeric variables with the exception of the timestamp and randomised variables. The model was subsequently simplified with the consideration of the p-value and the VIF value. To address the presence of autocorrelation in the energy data, an additional lag variable for appliances was incorporated, as in the scenario of time-series-dependent energy forecasting models (Fan, Sun, & Wang, 2020). Timestamp, date, and day columns were excluded because they are identifiers or duplicate time features already captured through the hour variable, which better represents daily usage patterns. Lights_Wh was excluded since we are attempting to isolate appliance-related energy use, and lighting can distort total household usage patterns.

The ANOVA analysis revealed statistically significant differences among the energy consumption over different parts of the day. Along with model assumption (L.I.N.E.), plots were created. Multicollinearity remained in bounds.

The best model with the lagged consumption, selected environmental variables, and time variables exhibited a greater R^2 of 0.541, illustrating moderate explanatory power in making future energy consumption forecasts. The methodology followed here is consistent with earlier empirical research in the field of intelligent home analysis and includes understandable, data-driven suggestions for energy-efficiency planning.

Analytics

The best prediction model was selected after refinement and diagnostics for improved R-squared value, yet simpler model with improved performance. In the selected best model, the baseline appliance energy estimate failed to be significant indicating the desired result of non-zero scale being relevant for real variables. All the model's significant predictors, namely, Temp_Living, Humidity_Living, Temp_Kitchen, Humidity_Kitchen, hour, lag_Appliances, and selected interaction terms, were statistically significant at the 5% significance level, except Temp_Kitchen: Humidity_Kitchen which were marginal.

The predictor variables account for approximately 54.1% of the appliance energy usage. The very low p-value (< 2.2e-16) shows the model's statistical significance. The model's assumption values validate that the residuals are normally distributed but skewed (Shapiro-Wilk test p-value > 0.05), there is not much evidence of heteroscedasticity (Breusch-Pagan test p-value > 0.05), and all the Variance Inflation Factor values are below 5 with no multicollinearity issues.

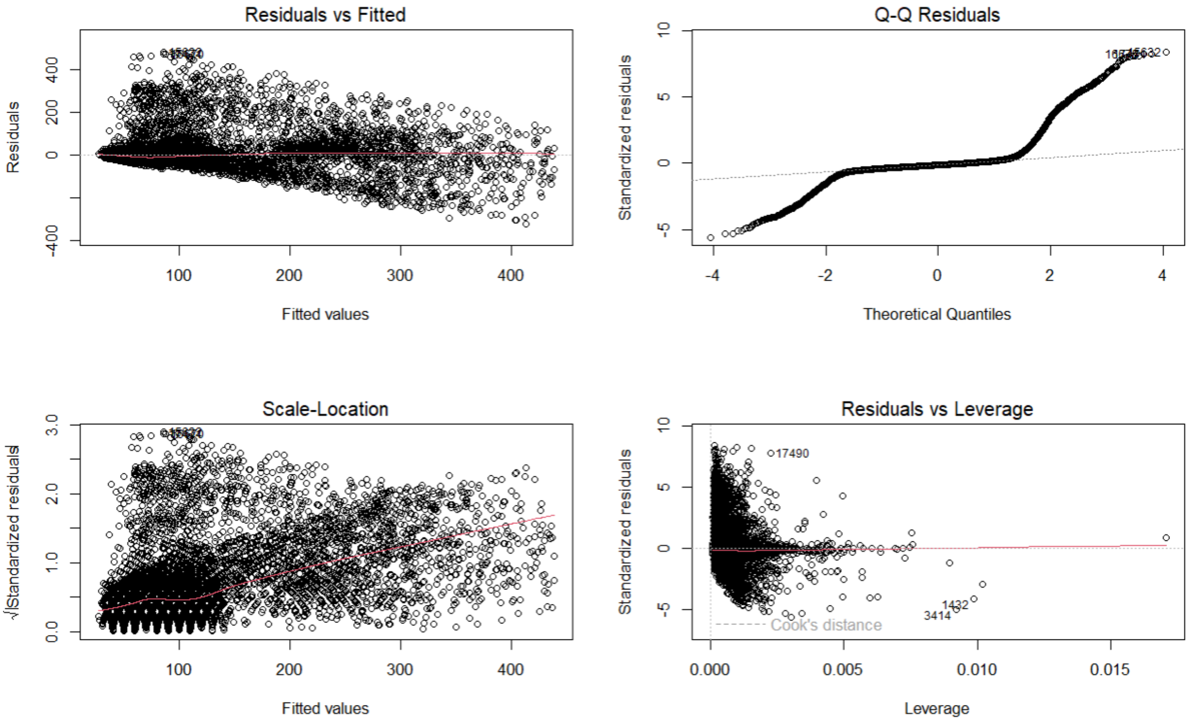


Figure 6 – Model Analysis Plots

The significant predictors are the appliances past energy usage, the time of the day, and indoor climatic conditions (Temp_Living and Humidity_Living), confirming the significance of temporal as well as thermal factors to house energy profiles. The interactions enhance the predictability, capturing the nuance of the effect of humidity over temperature-driven energy demand. Energy usage can be well-foreseen with an accurately specified linear model with high interpretability.

Time of day and lagged energy usage were highly predictive, affirming the utility of time-aware and behaviour-aware energy forecasting. The results are of direct utility to utilities for peak load forecasting, real-time pricing, and optimization of smart home automation systems.

Appliances_Wh = 16.14 - 8.63 * Temp_Living - 5.34 * Humidity_Living + 8.52 * Temp_Kitchen + 4.26 * Humidity_Kitchen + 0.70 * lag_Appliances + 4.37 * hour + 0.25 * Temp_Living:Humidity_Living - 0.20 * Temp_Kitchen:Humidity_Kitchen - 0.019 * Temp_Living^2 - 0.16 * hour^2

This model shows the energy use rises with prior usage and time; kitchen factors increase it, while living room conditions reduce consumption. Most data points show acceptable leverage; a few high-leverage cases exist but remain below Cook’s distance threshold, ensuring model stability.

Recommendations & Conclusions

The model developed here shows promising potential for forecasting appliance-level energy usage in intelligent homes. With the determination of the best predictor variables such as time of day, indoor conditions, and past energy usage, the model addresses the stakeholders' various questions.

These findings can be used by the Utility Providers for deploying Time-of-Use (ToU) pricing policies, which would persuade users to move their energy consumption to off-peak hours, thereby improving grid stability and lowering the operational cost. Predictability of the energy usage patterns using the model assists in optimizing energy distribution and load balancing.

The model can be integrated into home automation developers' smart home systems to support predictive control of appliances. With the model's ability to predict energy demand, systems can anticipate and adjust maintain comfort while minimizing energy usage, hence, realizing energy efficiency and sustainability goals. (Malik et al., 2024)

It can be utilized by policy makers to design energy conservation plans aimed at peak demand times and to raise consumer consciousness about energy conservation procedures. The model's findings can guide policy initiatives to reduce total energy demand and environmental impacts.

Finally, the regression model not just improves the understanding of energy use behaviour but also serves as an efficient decision-support tool for strategic energy management. It can lead to improved energy efficiency, cost-effectiveness, and promote the transition to sustainable energy systems.

References

Chen, H., Liu, Z., Yu, H., Wang, B., & Zhang, Z. (2023). Modeling residential energy consumption with fine-grained IoT data: A regression-based approach. *Energy and Buildings, 287*, 113242. https://doi.org/10.1016/j.enbuild.2023.113242

Ga, K., Edigaa, P., Sa, A., Pa, A., Ta, S., & Mittal, A. (2024). Smart energy management: Real-time prediction and optimization for IoT-enabled smart homes. *Cogent Engineering, 11*(1), 2390674. https://doi.org/10.1080/23311916.2024.2390674

Himeur, Y., Alsalemi, A., Bensaali, F., & Amira, A. (2021). Data-driven approach for energy optimization in smart buildings: Ten years review. *Renewable and Sustainable Energy Reviews, 132*, 110112. https://doi.org/10.1016/j.rser.2020.110112

Fan, C., Sun, Y., & Wang, J. (2020). Short-term load forecasting based on an improved LSTM neural network. IEEE Access, 8, 162972–162981. https://doi.org/10.1109/ACCESS.2020.3021494

Malik.S., Singh, I., Gupta, H. V., Prakash, S., Jain, R., Acharya, B., & Hu, Y.-C. (2024). Deep learning based predictive analysis of energy consumption for smart homes. *Multimedia Tools and Applications, 84*, 10665–10686. https://doi.org/10.1007/s11042-024-18758-z