SIT718 Real World Analytics

Assignment 2

Michael Rideout
Student ID - s225065259

Assignment 2 Task Overview

- Dataset analysis utilising aggregating function models in the R language
- Appliances Energy Prediction Dataset was analysed. Scholarly article - Data driven prediction models of energy use of appliances in a low-energy house (Candanedo, Feldheim, et al 2017)
- R packages utilised
    - lpSolve (Csárdi and Berkelaar, 2024)
    - Pastecs (Grosjean et al, 2024)
    - Car (Fox et al, 2024)
    - Ggplot2 (Wickham et al., 2024)
    - Knitr (Xie et al., 2024)
    - Corrplot (Wei et al., 2024)
    - MonoInc (Minto et al., 2024)

Data Preprocessing

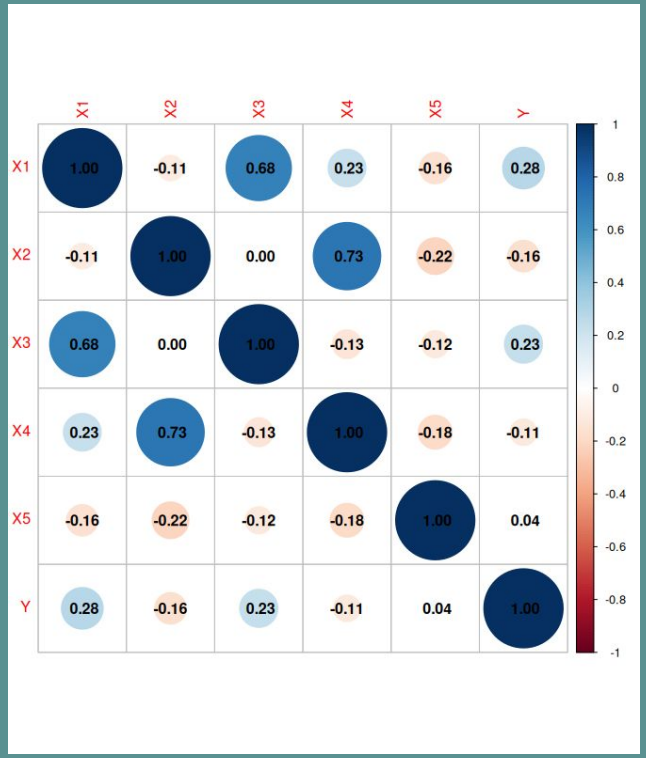
- Features Overview
  - X1 - Living Room Temperature (degrees celsius)
  - X2 - Living Room Humidity (percentage)
  - X3 - Office Room Temperature (degrees celsius)
  - X4 - Office Room Humidity (percentage)
  - X5 - Pressure (mm or mercury)
  - Y - Appliances Energy Consumption (Wh)
- Data Checks Performed
  - Row count after loading
  - Missing values
  - Duplicates
- Outlier Removal Before Sampling
  - Row count before outlier removal 19735
  - Row count after outlier removal 16710 (3025 rows removed)
- Sampling
  - Sampled to 450 entries

Feature Selection

- Only 4 features out of 5 allowed
- Based on lowest correlation to Y, X5 would be dropped
- Leave One Out feature selection method applied

Feature Left Out	WAM RMSE (y-transformed)
X1	**0.1517**
X2	0.1537
X3	0.1582
X4	0.1572
X5	0.1524

- Based on Leave One Out WAM model results, dropping X1 produced most accurate model according to minimising RMSE

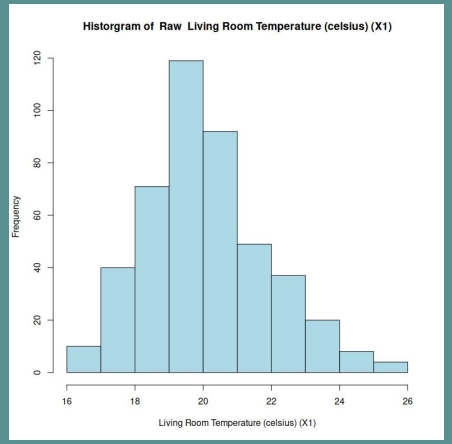


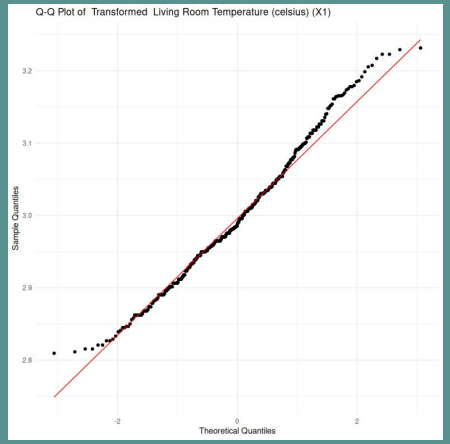
Transformations Overview

- Need to correct monotonicity direction
  - Ideally increasing feature value has corresponding increase in target value
- Need to correct distribution skew
  - Stat.desc() from pastecs library (Grosjean et al, 2024)
  - skew.2SE > 1 or skew.2SE < -1 is statistically significant
- Need unit feature scaling
  - All features should have a range between 0 and 1

Feature: X1 (Living Room Temperature - celsius)

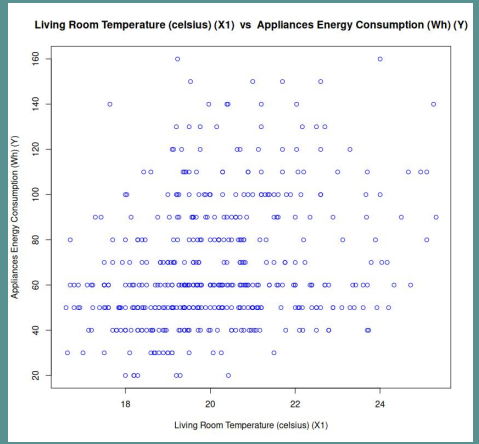
RAW





Properties

Median: 19.82
Mean: 20.12
Skew: 0.54
Skew.2SE: 2.33
Skew Type: Right
Mono Inc: Yes

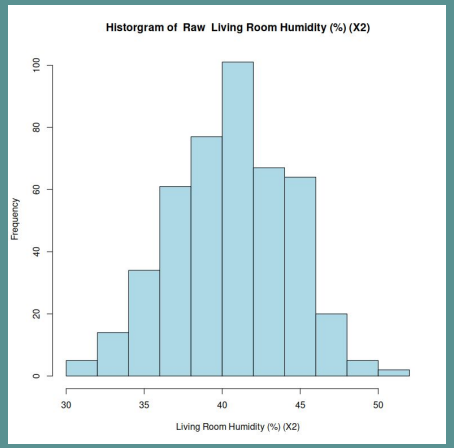


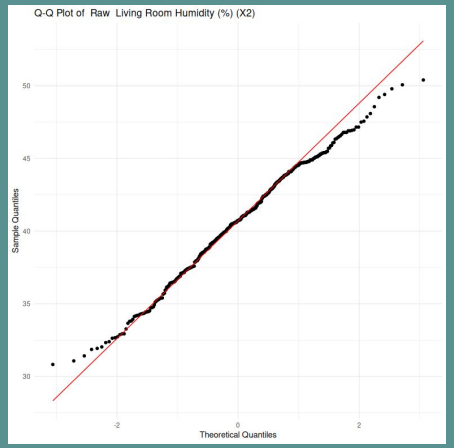
TRANSFORMED

Feature not used in model

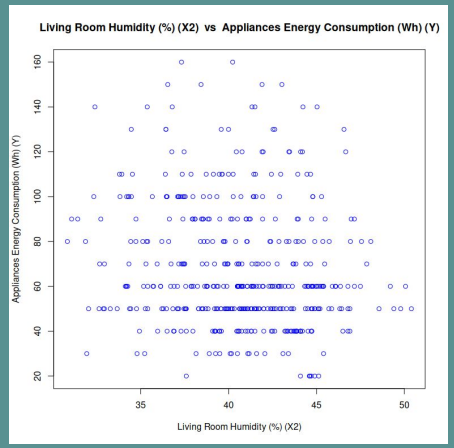
Feature: X2 (Living Room Humidity - percentage)

RAW

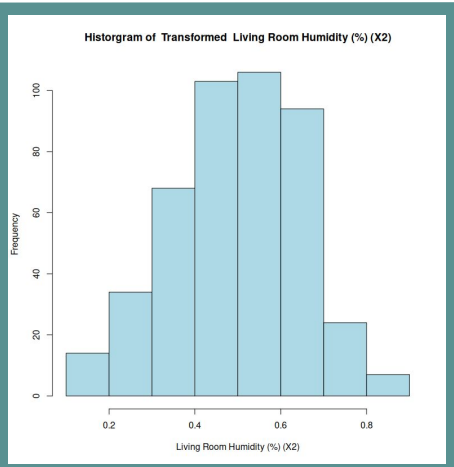


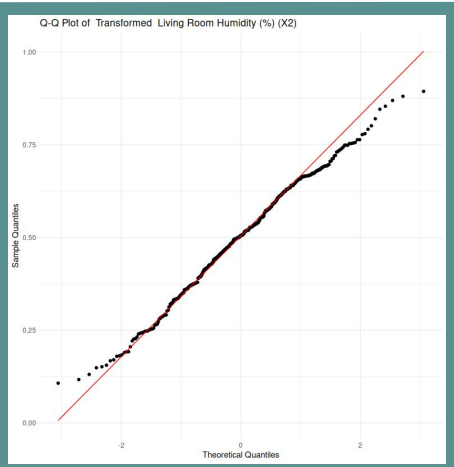


Properties
Median: 40.73
Mean: 40.59
Skew: -0.16
Skew.2SE: -0.70
Skew Type: None
Mono Inc: Yes
Min: 29.89
Max: 51.29



TRANSFORMED



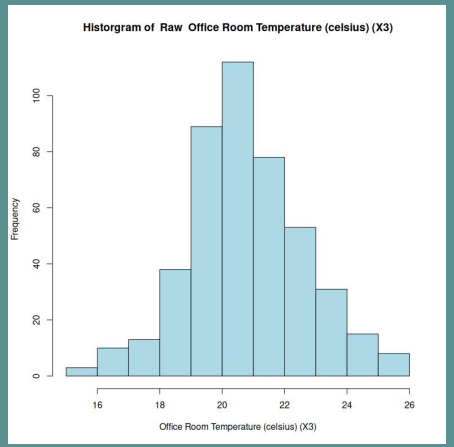


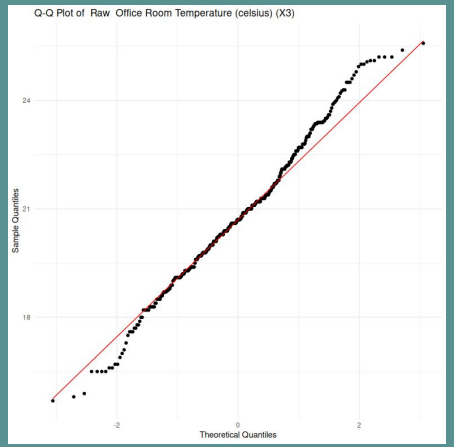
Properties
Median: 0.51
Mean: 0.50
Skew: -0.16
Skew.2SE: -0.70
Skew Type: None
Mono Inc: Yes
Min: 0.11
Max: 0.89

Transformations
Distribution: None
Scaling: Unit Z-Score

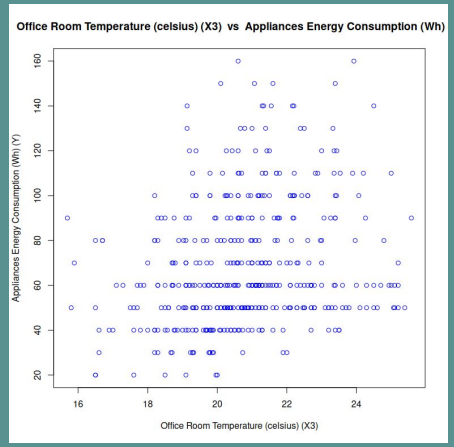
Feature: X3 (Office Room Temperature - celsius)

RAW

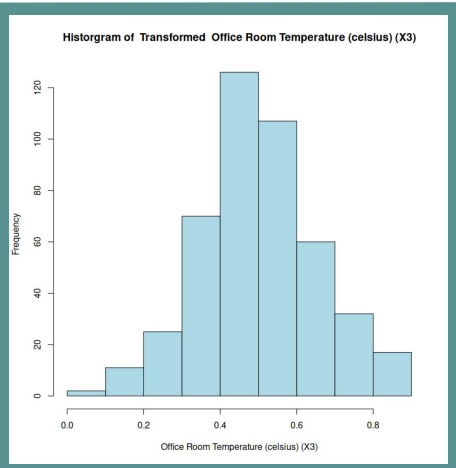


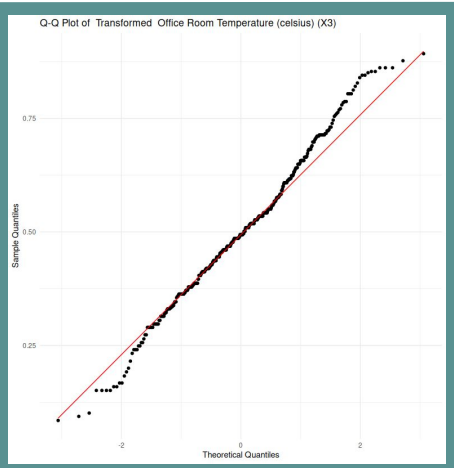


Properties
Median: 20.70
Mean: 20.77
Skew: 0.11
Skew.2SE: 0.50
Skew Type: None
Mono Inc: Yes
Min: 15.69
Max: 25.92



TRANSFORMED



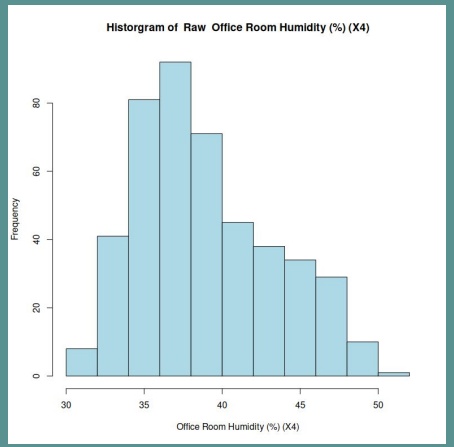


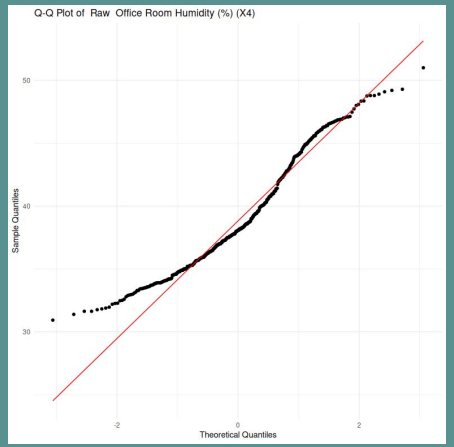
Properties
Median: 0.49
Mean: 0.50
Skew: 0.11
Skew.2SE: 0.50
Skew Type: None
Mono Inc: Yes
Min: 0.08
Max: 0.89

Transformations
Distribution: None
Scaling: Unit Z-Score

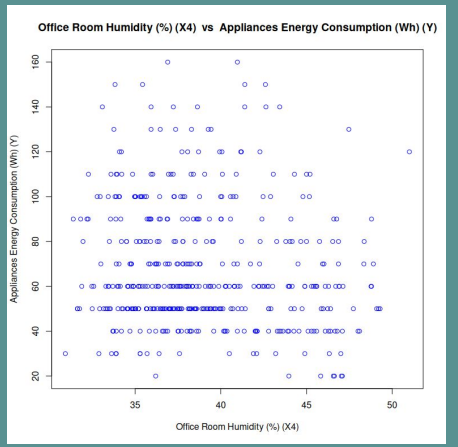
Feature: X4 (Office Room Humidity - percentage)

RAW

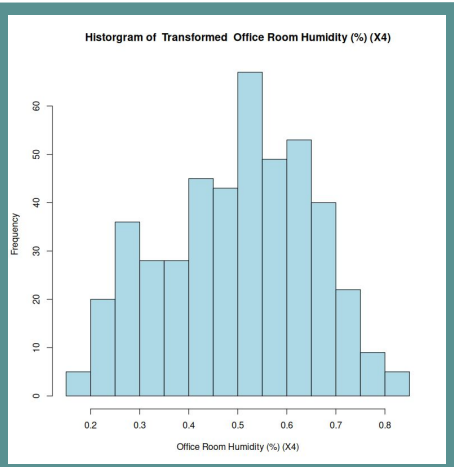


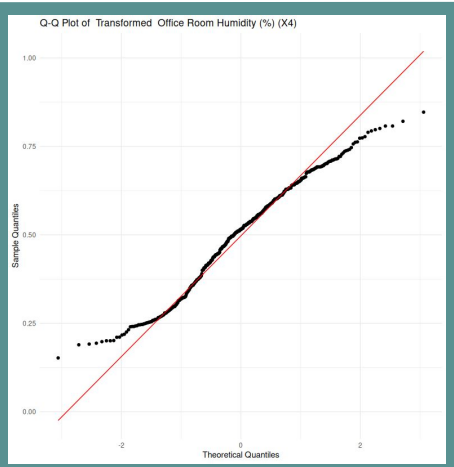


Properties
Median: 38.08
Mean: 38.94
Skew: 0.56
Skew.2SE: 2.43
Skew Type: Right
Mono Inc: Yes
Min: 29.66
Max: 51



TRANSFORMED



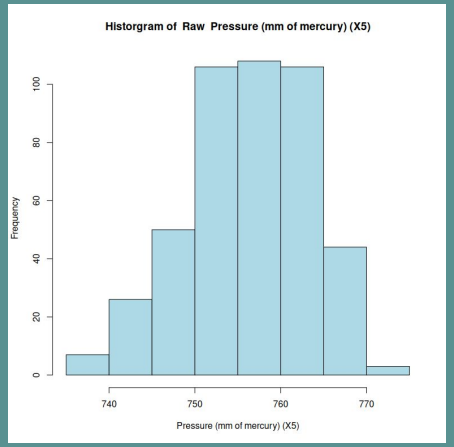


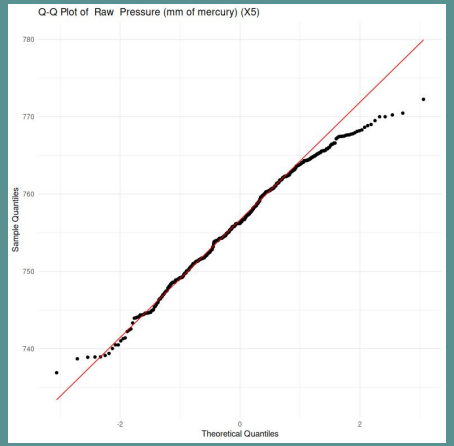
Properties
Median: 0.52
Mean: 0.50
Skew: -0.18
Skew.2SE: -0.80
Skew Type: None
Mono Inc: Yes
Min: 0.15
Max: 0.85

Transformations
Distribution: Inversion
Scaling: Unit Z-Score

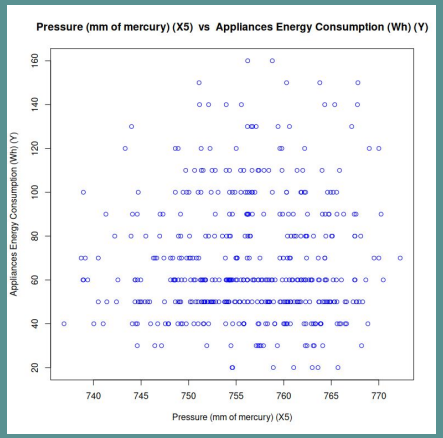
Feature: X5 (Pressure - mm of mercury)

RAW

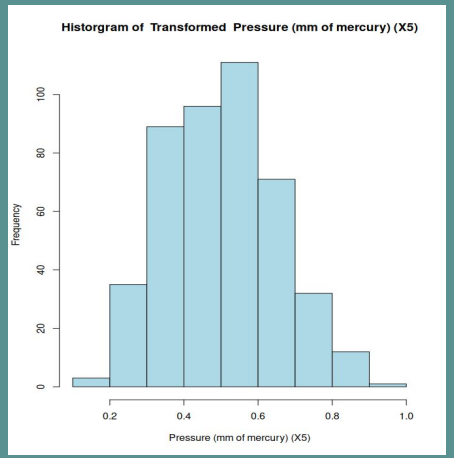


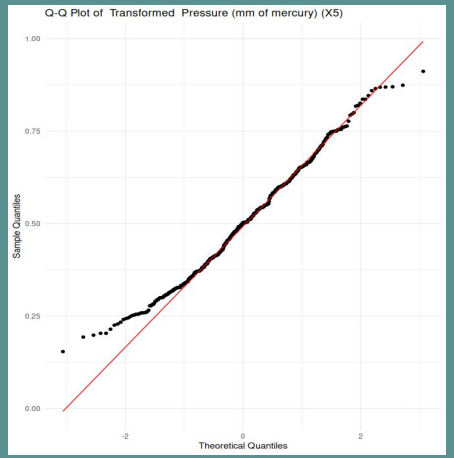


Properties
Median: 756.23
Mean: 756.31
Skew: -0.27
Skew.2SE: -1.19
Skew Type: Left
Mono Inc: No
Min: 735.77
Max: 772.27



TRANSFORMED



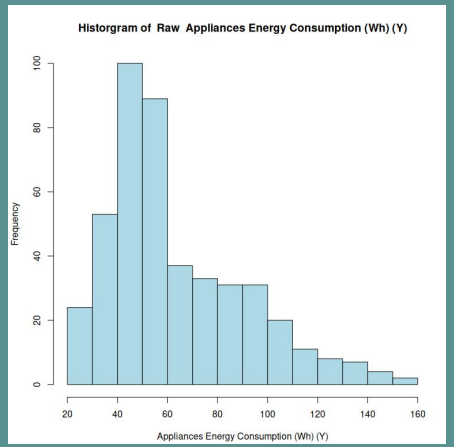


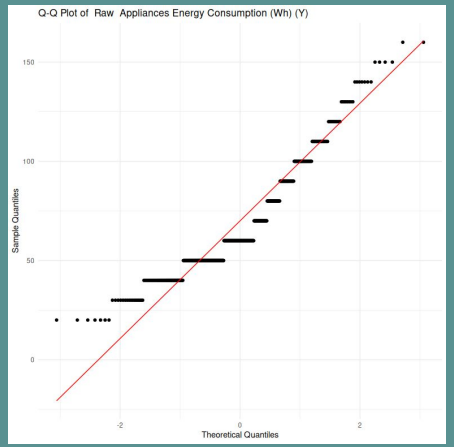
Properties
Median: 0.50
Mean: 0.50
Skew: -0.22
Skew.2SE: 1.08
Skew Type: Right
Mono Inc: Yes
Min: 0.15
Max: 0.91

Transformations
Distribution:
log(log(negation))
Scaling: Unit Z-Score

Feature: Y (Appliances Energy Consumption - Wh)

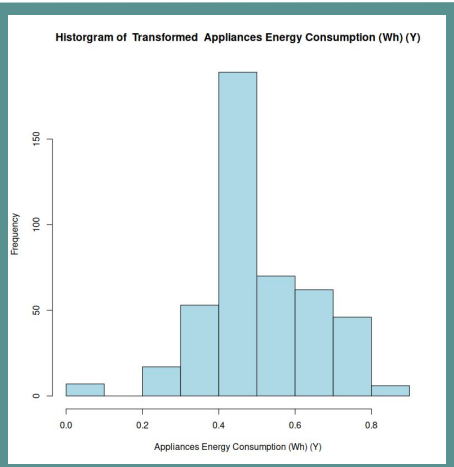
RAW

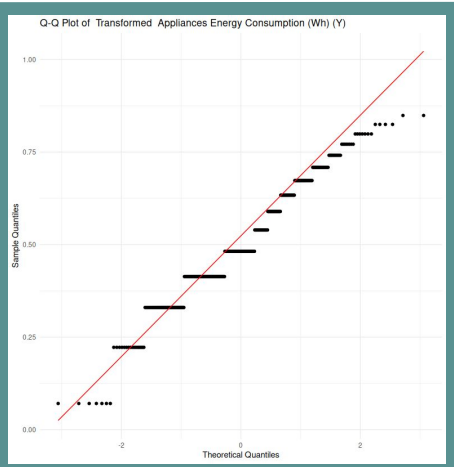




Properties
Median: 60.00
Mean: 68.20
Skew: 0.93
Skew.2SE: 4.03
Skew Type: Right
Min: 10
Max: 170

TRANSFORMED





Properties
Median: 0.48
Mean: 0.50
Skew: -0.05
Skew.2SE: -0.23
Skew Type: None
Min: 0.07
Max: 0.84

Transformations
Distribution: Logarithm
Scaling: Unit Z-Score

Model Evaluation

- Models utilised for analysis based on aggregation functions (Simons, 2016)
  - Weighted Arithmetic Mean (WAM)
  - Weighted Power Means (WPM p0.5)
  - Weighted Power Means (WPM p2)
  - Ordered Weighted Average (OWA)
- Best Performing Model - WAM
- Training Errors:

Model	RMSE	Avg Abs Error	Pearson Coefficient	Spearman Coefficient
WAM	0.1517	0.1235	0.2426	0.1981
WPM p0.5	0.1520	0.1238	0.2449	0.2045
WPM p2	0.1530	0.1253	0.2287	0.1895
OWA	0.1562	0.1276	0.1335	0.1120

Feature Importance

- Weighted vector from weighted models demonstrate relative feature importance
- Weighted vector components add up to 1
- Weights for best performing model (WAM):

Feature	Weight
X2 - Living Room Humidity	0.14
X3 - Office Room Temperature	0.38
X4 - Office Room Humidity	0.40
X5 - Pressure	0.08

- X3 and X4 are the two most important features, 78% of total weighting combined
- X2 and X5 combined contribute only 22% of importance to weighting

Prediction

- Inference performed with one with values being X1 = 19.1, X2 = 43.29, X3 = 19.7, X4 = 43.4, X5 = 743.6
- Ground truth result = 60
- Predicted result = 53.28 (reverse transformed from 0.4373)
- Prediction result from model required scaling and transformation to be reversed
- Prediction error of 6.72 well below average absolute error of 23.03 (target_reverse_transformation applied to avg abs error 0.1235)
- WAM predicted close to ground truth result and well within the margin of error

Implications

- Office room temperature and humidity provided most predictive power amongst the feature set
  - One person worked from home, most likely spending most time in the house
  - Also work hours coincide with higher temperatures
- As office room temperature and humidity increase, energy usage increases.
  - Potentially due to air conditioning being used, an energy intensive appliance
- Pressure had an inverse relationship to appliance energy consumption
  - Higher air pressure is generally associated with fair stable weather, possibly meaning people were outdoors or appliance use not required
  - Lower air pressure means the opposite
- Results are quite intuitive to real world experience

Limitations

- Investigation limitations
  - Model validation was insufficient
    - Sample size too small
    - Train / test split with k-folds cross validation
  - Temporal information not accessible
    - Appliance energy usage is time dependent and cyclical. On daily and seasonal time scales
    - Temporal machine learning methods more appropriate for modelling
- Original paper limitations
  - Original data is only for 4.5 months from January to May in 2016.
    - Seasonal yearly patterns not captured

Reference

1. Candanedo, L.M., Feldheim, V. & Deramaix, D. (2017) 'Data-driven prediction models of energy use of appliances in a low-energy house', Energy and Buildings, 140, pp. 81-97
2. Csárdi, G. and Berkelaar, M., 2024. lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs. R package version 5.6.22. Available at: https://CRAN.R-project.org/package=lpSolve [Accessed 9 Dec. 2024].
3. Fox, J., Weisberg, S., & Price, B. (2024) car: Companion to Applied Regression. R package version 3.1-3. Available at: https://CRAN.R-project.org/package=car [Accessed: 9 December 2024].
4. Grosjean, P., Ibanez, F., & Etienne, M. (2024) pastecs: Package for Analysis of Space-Time Ecological Series, version 1.4.2. Available at: https://CRAN.R-project.org/package=pastecs [Accessed: 9 December 2024].
5. Minto, M., Josey, M., & Williams-DeVane, C. (2016). MonoInc: Monotonic Increasing. Version 1.1. Available at: https://CRAN.R-project.org/package=MonoInc [Accessed 11 Dec. 2024].
6. Simon, J. (2016). An Introduction to Data Analysis using Aggregation Functions in R. 1st ed. Springer International Publishing.
7. Wei, T. and Simko, V., 2024. corrplot: Visualization of a Correlation Matrix. Version 0.95. [online] CRAN. Available at: https://cran.r-project.org/package=corrplot [Accessed 11 Dec. 2024].
8. Wickham, H., Chang, W., Henry, L., Pedersen, T.L., Takahashi, K., Wilke, C., Woo, K., Yutani, H., Dunnington, D., van den Brand, T. and Posit, PBC, 2024. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics (Version 3.5.1). Available at: https://cran.r-project.org/package=ggplot2 [Accessed 9 Dec. 2024].
9. Xie, Y., Sarma, A., Vogt, A., Andrew, A., Zvoleff, A., Al-Zubaidi, A., et al. (2024) knitr: A General-Purpose Package for Dynamic Report Generation in R. Version 1.49. Available at: https://cran.r-project.org/package=knitr [Accessed: 11 December 2024]