12963 76651

ENERGY REPORT.docx

Assignment

Class

University

Document Details

Submission ID

trn:oid:::1:2927237515

Submission Date

May 21, 2024, 7:17 PM UTC

Download Date

May 21, 2024, 7:19 PM UTC

File Name

AgADmBIAAjdiaFI

File Size

506.6 KB

70 Pages

8,077 Words

58,378 Characters

How much of this submission has been generated by AI?

0%

of qualifying text in this submission has been determined to be generated by AI.

Caution: Percentage may not indicate academic misconduct. Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

Frequently Asked Questions

What does the percentage mean?

The percentage shown in the AI writing detection indicator and in the AI writing report is the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was generated by AI.

Our testing has found that there is a higher incidence of false positives when the percentage is less than 20. In order to reduce the likelihood of misinterpretation, the AI indicator will display an asterisk for percentages less than 20 to call attention to the fact that the score is less reliable.

However, the final decision on whether any misconduct has occurred rests with the reviewer/instructor. They should use the percentage as a means to start a formative conversation with their student and/or use it to examine the submitted assignment in greater detail according to their school's policies.

How does Turnitin's indicator address false positives?

Our model only processes qualifying text in the form of long-form writing. Long-form writing means individual sentences contained in paragraphs that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. Qualifying text that has been determined to be AI-generated will be highlighted blue on the submission text.

Non-qualifying text, such as bullet points, annotated bibliographies, etc., will not be processed and can create disparity between the submission highlights and the percentage shown.

What does 'qualifying text' mean?

Sometimes false positives (incorrectly flagging human-written text as AI-generated), can include lists without a lot of structural variation, text that literally repeats itself, or text that has been paraphrased without developing new ideas. If our indicator shows a higher amount of AI writing in such text, we advise you to take that into consideration when looking at the percentage indicated.

In a longer document with a mix of authentic writing and AI generated text, it can be difficult to exactly determine where the AI writing begins and original writing ends, but our model should give you a reliable guide to start conversations with the submitting student.

Table of Contents

Executive Summary

Energy use by appliances and lights fittings have been investigated in this study. Further, this study intended to develop models that predict energy usage based on environmental parameters and variables such as temperature, wind speed and visibility. A multivariate time series data set from Appliances Energy Predictions (an online repository) with 28 variables has been utilized. The data consisted of aggregating environmental measurements and conditions and recorded data on energy usage for a 4.5-month period.

A number of analytical methods, including Principal Components Analysis, Canonical Correlation Analysis, and Partial Least Squares (PLS) Regression, were performed. Findings revealed significant associations between environmental factors, weather conditions, and energy usage, validating hypotheses on the impact of these factors on energy consumption. The PLS regression model captured 91.11% of the variance in predictor variables and 73.03% of the variance in energy usage, demonstrating robust predictive capabilities. Cross-validation analysis confirmed the model's reliability, with a Root Mean PRESS of 0.6702.

1.0 Introduction

1.1 Background

Energy usage in buildings constitutes a significant portion of total energy usage globally. Optimizing energy consumption is paramount as the world increasingly focuses on sustainability (Cao et al., 2016). Low-energy buildings are designed to minimize energy use while maintaining comfort. Therefore, effective energy management in such buildings relies on understanding energy consumption patterns and identifying their influencing factors. In US, residential and commercial buildings explains for approximately 40% of total energy consumption (U.S. Energy Information Administration, 2018).

Developments in technologies, coupled with an ever-growing number of energy-efficiency regulatory initiatives, have furthered the promotion of low-energy building practices around the world. According to (Du et al., 2022), these buildings implement all types of design strategies and technologies that lower the demand for energy and increase comfort among their occupants. From passive solar design principles to integrating energy-efficient appliances and HVAC systems today, low-energy buildings clearly show the potential for high-impact sustainable living without compromising the quality of life. But achieving energy performance should go beyond technologies; it requires good knowledge of how different factors interact and affect energy consumption patterns

1.2 Purpose and Rationale

This paper analyses the energy use in appliances and light fittings in a relatively low-energy building and how the indoor and outdoor environmental variables affect energy consumption. A low-energy building is designed to use very little energy to sustain comfort for its users. This study is intended to develop models that predict energy consumption from environmental parameters and weather conditions.

The study sheds more light on the complex dynamic in energy management. Additionally, the idea of energy management in low-energy buildings finds space for discussion based on its total setting. Indoor temperatures, humidity, and whether it is hot or cold from the outside bear together on energy use. Such approaches often depend on simple heuristics that are employed without the ability to capture the full complexity of the dynamics in energy consumption.

1.3 Hypothesis

H1: Outdoor weather conditions (temperature, wind speed and visibility) significantly influences energy usage for the appliances and lights.

H2: A relationship exists between indoor temperatures and humidity levels with energy usage.

H3: Past energy usage and environmental conditions possess predictive power for forecasting future energy use.

1.4 Objectives

This study's purpose is to:

1. Analyze the energy usage patterns in a low-energy building.

2. Investigate the relationship between indoor temperatures, humidity levels, outdoor weather conditions, and energy consumption.

3. Develop a predictive models for future based on past energy usage and environmental conditions.

2.0 Literature

Previous literature has explored the study of energy consumption in buildings and the factors responsible for driving it. Some studies have examined energy consumption patterns in buildings, focusing on predicting energy used by appliances in a low-energy houses. Studies by Candanedo et al. (2017) and Guo et al. (2015) used the data from wireless sensors installed to assess the environment inside buildings and outside spaces and data from smart electric meters for demand load studies to describe the outline of energy demand loads.

Concurrently, research efforts have examined occupant behaviours in homes and offices to rate appliance efficiency. Works by Hong et al. (2016) and Kavousian et al. (2015) analyzed occupant behaviours during their stay, employing regression model and probabilistic models to analyse and identify patterns. Additionally, studies have attempted to precisely predict occupancy numbers by investigating appliance use behaviours (Candanedo, 2016).

Taken together, these studies emphasized the nature of appliance energy use in homes or office spaces, which is basically driven by the number of occupants, internal and external environmental conditions, building architecture, and geographical location. The knowledge and incorporation of these become vital for the derivation of appliance energy consumption in Inhabited Environment Buildings.

This study utilizes weather data (temperature, humidity, wind speed, visibility and dew points) to understand the complex dynamic relationship between internal and external environment factors in the building as well as energy usage. Research in this domain has the potential to reveal new insights and strategies toward higher energy efficiency. Additionally, the data in this research on energy use by appliances are very comprehensive and set the stage for understanding the dynamics likely to emerge in great detail.

3.0 Methods

3.1 Data Source

The study uses multivariate time series data sourced from the Appliances Energy dataset (oline repository). The source has been considered reputable because of its reliability, wide use, and thorough documentation.

3.2 Data Collection

The dataset contained comprehensive data on the energy usage of appliances and light fixtures in a low-energy building. It also included various environmental readings inside the house and weather conditions from the nearest airport weather station. Data was collected at regular 10-minute intervals over 4.5 months. Data collection was facilitated using a ZigBee module and m-bus energy meters. Temperature and humidity conditions were monitored using a ZigBee module. Each module recorded and transmitted the data every 3.3 minutes. The data was then averaged over 10-minute periods. Energy usage data for appliances and light fixtures was recorded in every 10 minutes using m-bus energy meters.

Variables such as wind speed, dew points and visibility accounts for the weather data. This data was obtained from a public dataset and merged with the experimental data using the date and time columns. 19, 735 samples consitituted the data. Each sample constituted the data, representing a unique 10-minute data collection interval. There were 28 variables, including appliance energy use in *Wh*, light energy use in *Wh*, indoor temperatures and humidity percentages for various rooms, outdoor temperature and humidity percentages, and weather-related variables such as wind speed and visibility.

Table 1

Variable Information

Variable Name	***Unit***
Date time year-month-day	hour: minute: second
Appliances, energy use	Wh
lights, energy use of lights fixtures in the house	Wh
T1, Temperature in kitchen	^{o}C
T2, Temperature in living room	^{o}C
T3, Temperature in laundry room	^{o}C
T4, Temperature in office room	^{o}C
T5, Temperature in bathroom	^{o}C
T6, Temperature outside the building	^{o}C
T7, Temperature in ironing room	^{o}C
T8, Temperature in teenager room	^{o}C
T9, Temperature in parents room	^{o}C
To Temperature outside (from Chievres weather station)	^{o}C
RH_1, Humidity in kitchen	%
RH_2, Humidity in living room	%
RH_3, Humidity in laundry room	%
RH_4, Humidity in office room	%
RH_5, Humidity in bathroom	%
RH_6, Humidity outside the building	%
RH_7, Humidity in ironing room	%
RH_8, Humidity in teenager room	%
RH_9, Humidity in parents room	%
RH_out, Humidity outside (from Chievres weather station)	%
Pressure (from Chievres weather station)	mm Hg
Wind speed (from weather station)	m/s
Visibility (From Chievres weather station)	km
Tdewpoint (from Chievres weather station)	A°C
rv1, Random variable 1	non-dimensional
rv2, Random variable 2	non-dimensional

3.4 Plotting the Data

Energy usage, shown by figure 1 and 2 indicates fluctuatuation over time. The enegy used by lights is relatively lower than that of appliances because lights are often turned off during the day. In contrast, appliances require more energy because electronic household items are used frequently throughout the day.

Figure 1

Daily average energy use by appliance

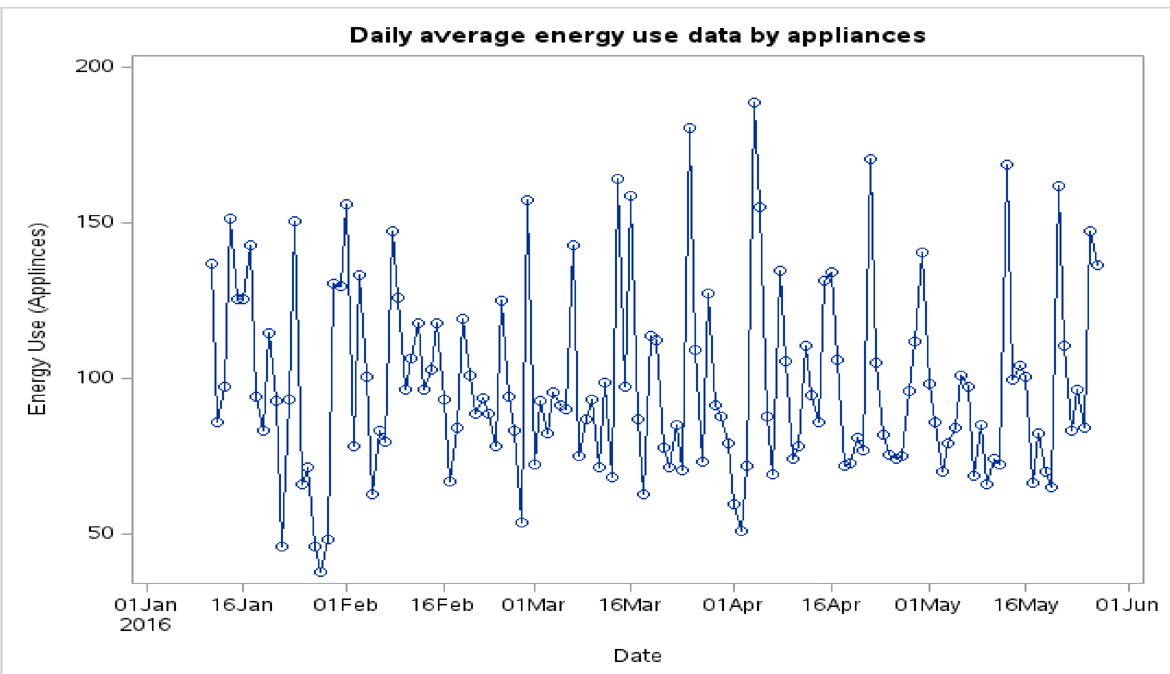
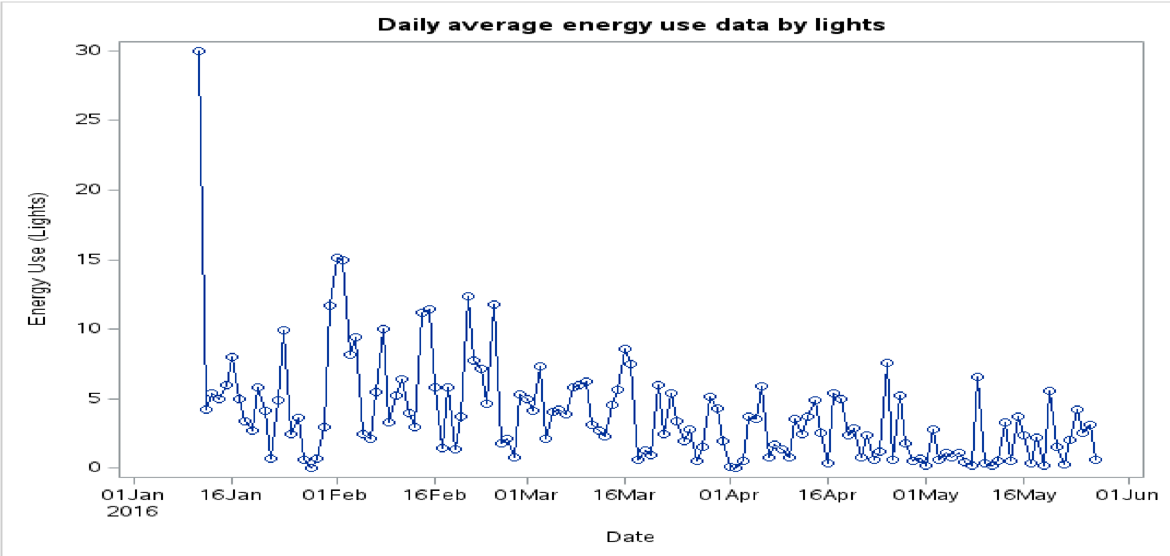


Figure 2: Daily average energy use by lights



3.5 Data Preprocessing

Data preprocessing was conducted before analysis to prepare the data for statistical modelling. Since there was no explicit group structure within the dataset, it was divided into indoor and outdoor data based on the captured environmental readings and weather conditions. Random variables (rv1 and rv2) were used to test regression models by filtering out non-predictive attributes, thus, ensuring the robustness of the predictive models. The classification into indoor and outdoor variables was primarily based on the measurement location and the nature of the data.

In addition, since the dataset contained energy usage time series data for each 10-minute interval, several attributes were generated, including daily average, daily minimum, daily maximum, morning, afternoon, evening, and night. All these attributes were prepared to identify different energy use patterns at other times of the day. These analyses helped identify relationships among the predictor and target variables (Appliances).

3.6 Analytical Methods Used

The SAS Statistics Analysis System was used broadly in the research through key analytical methods. It is a robust statistical software package designed to run under large datasets and complex analyses, with an enormous set of procedures and tools for data manipulation, visualization, advanced statistical modelling, and more. SAS comprises a wide-ranging set of procedures and tools useful for manipulating data, visualizing data, and performing highly advanced statistical modeling.

The following methods were employed to analyze the dataset and address the research aims:

3.6.1 Principal Components Analysis (PCA) and its Visualization

PCA is used to reduce dimensionality in the data set and to detect patterns in the data. Visualization techniques such as scatter and scree plots were generated to explore energy usage patterns and relationships between variables.

3.6.2 Eigenvalues

As part of PCA, the eigenvalues were obtained to determine the variance accounted for by each principal component. It assists in measuring how much important information regarding the scatter of the data is captured by each principal component (Allee et al., 2022).

3.6.3 Factor Analysis & MDS

MDS helps visualize dissimilarities or similarities between samples or variables in a low-dimensional space and explore relationships between indoor conditions, outdoor weather, and energy consumption patterns. Together with factor analysis, identification of the latent factors underlying observed variables was carried out, revealing the structures under the data (Tucker-Drob & Salthouse, 2009)

3.6.4 Correspondence Analysis

Correspondence analysis explores associations between categorical variables (Riani et al., 2022). It helped understand the relationships between indoor environmental conditions and energy consumption behaviour, shedding light on factors influencing energy use within the building. The results will be visualized using biplots to display the associations between categories.

3.6.5 Canonical Correlation Analysis with PROC CANCORR

The relationship in the data variables was explored using the canonical correlation analysis. It is hypothesized that there exists connection between indoor conditions, energy usage and outdoor weather conditions. It identifies the most potent linear combinations of variables from different sets, providing insights into the associations between indoor and outdoor factors influencing energy consumption.

3.6.6 Canonical Discriminant Analysis

Observation in the data were clasified into unique groups based on predictor variables. The discriminant functions will be visualized to understand how well the groups are separated. It identifies the most discriminant variables that differentiate between groups (Ariza et al., 2021). It helped understand the factors contributing to variations in energy consumption patterns, facilitating the identification of key predictors for energy efficiency.

3.6.7 Clustering

Obsevations in the data with similar claracteristics were group together to form data clusters. K-Means Cluster analysis is a commonly applied method of data clustering and therefore it was employed (Oti et al., 2021). It helped identify subgroups in the data having similar energy consumption profiles, providing insights into factors influencing energy use variations.

3.6.8 PLS Regression

PLS regression was employed to model the relationships between predictor variables (e.g., indoor conditions, outdoor weather) and a response variable (Appliances). It helped develop predictive models for future energy use based on past energy usage and environmental conditions, contributing to understanding and forecasting energy consumption.

4.0 Results

4.1 Principal Components Analysis (PCA) and its Visualization

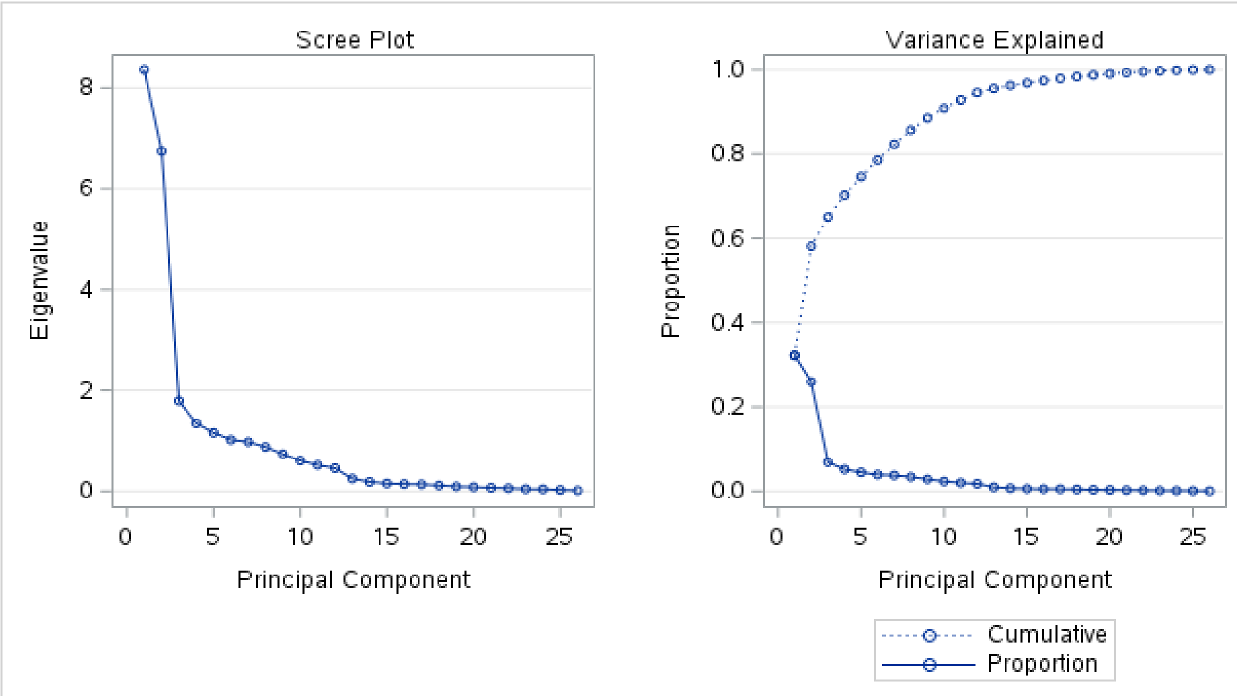
Principal Components Analysis (PCA) applied to the dataset reduced its dimensionality, uncovering underlying patterns. This technique transformed the original variables into a new set of uncorrelated variables, known as principal components, which capture the maximum variance in the data. The dataset was simplified by focusing on the first few principal components while retaining most of its original information.

4.1.1 Scree plot

The scree plot in Figure 1 visualizes the eigenvalues and determines the number of significant principal components. The plot identifies the elbow point, where the explained variance starts to level off, indicating the optimal number of components to retain.

Figure 3

Scree plot

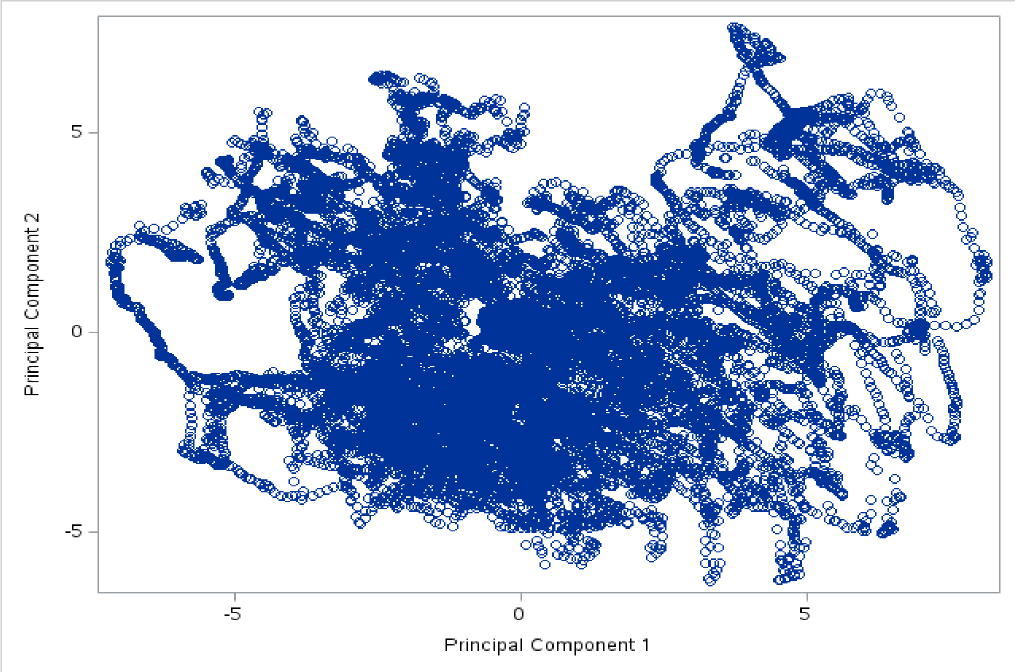


4.1.2 Scatter Plot of Principal Components

A scatter plot of the first two principal components visualizes the data in reduced dimension space. This plot identifies clusters and patterns in the data (see Figure 2).

Figure 4

Scatter Plot of Principal Components



4.2 Eigenvalues

Eigenvalues of the correlation matrix in Table 2 reveal the variance explained by each principal component, shedding light on their significance in understanding energy usage patterns. The scree plot in Figure 1 illustrates the diminishing magnitude of eigenvalues, with the first few eigenvalues contributing substantially to the variance. The first principal component accounts for 29.87% of the total variance, followed by the second component at 24.10%. The first two components explain over half (53.97%) of the variance. As the number of principal components increases, their contribution to the cumulative variance decreases gradually, with subsequent components capturing diminishing proportions of variability.

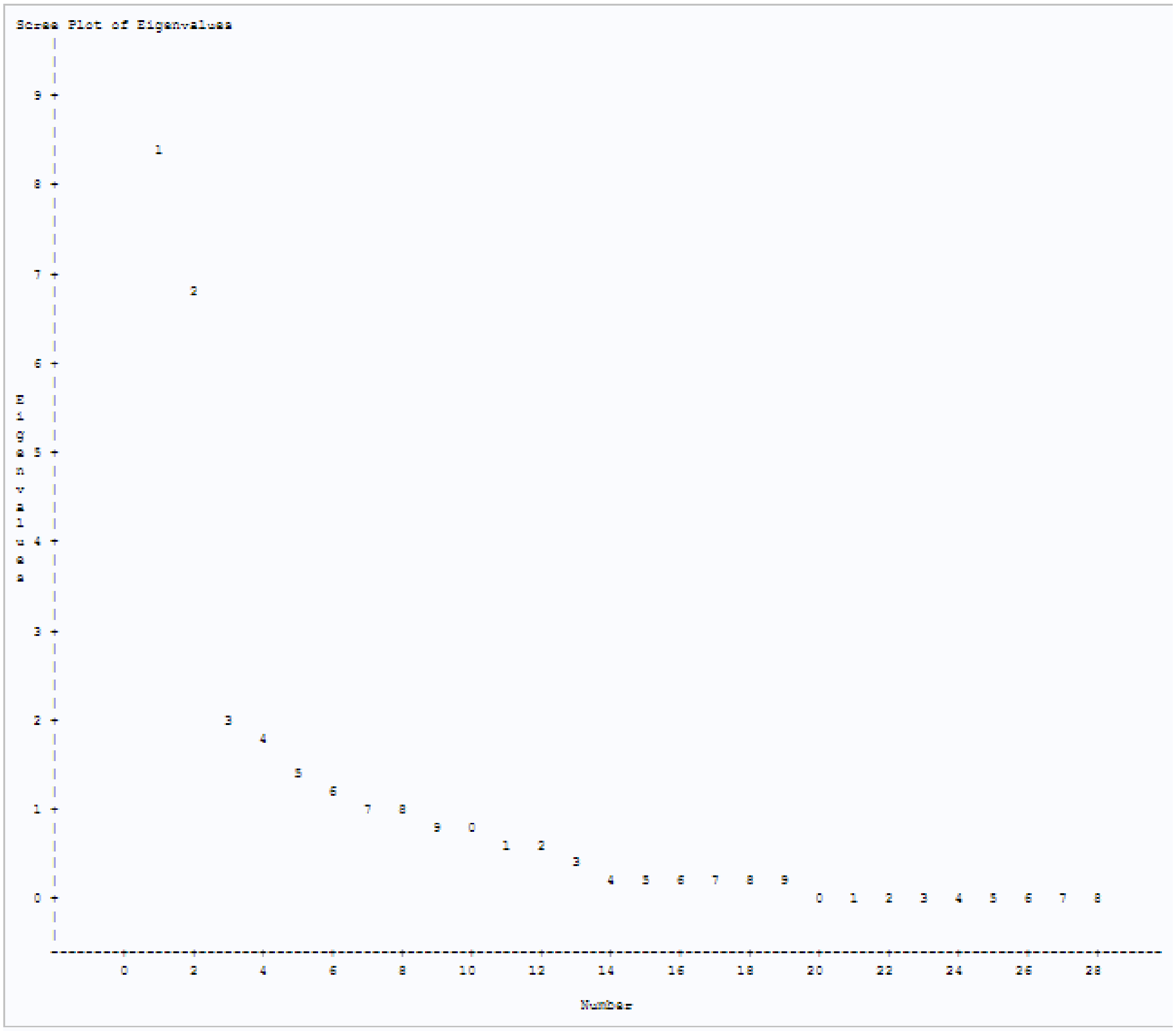
Table 2

Eigenvalues

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	8.36391608	1.61713651	0.2987	0.2987
2	6.74677957	4.73952193	0.2410	0.5397
3	2.00725764	0.22203118	0.0717	0.6114
4	1.78522646	0.43955484	0.0638	0.6751
5	1.34567163	0.18793895	0.0481	0.7232
6	1.15773268	0.14088556	0.0413	0.7645
7	1.01684712	0.04049376	0.0363	0.8008
8	0.97635336	0.10142099	0.0349	0.8357
9	0.87493237	0.13986452	0.0312	0.8670
10	0.73506785	0.12889218	0.0263	0.8932
11	0.60617567	0.08771918	0.0216	0.9149
12	0.51845649	0.06067677	0.0185	0.9334
13	0.45777972	0.20873528	0.0163	0.9497
14	0.24904443	0.06416787	0.0089	0.9586
15	0.18487656	0.03057107	0.0066	0.9652
16	0.15430549	0.01333131	0.0055	0.9707
17	0.14097418	0.00653152	0.0050	0.9758
18	0.13444267	0.01930781	0.0048	0.9806
19	0.11513485	0.02148678	0.0041	0.9847
20	0.09364807	0.01055418	0.0033	0.9880

4.3 Factor Analysis & MDS

The factor analysis aimed to identify the underlying structure in the data by reducing the number of observed variables into a smaller set of latent factors. The initial step involved extracting eigenvalues from the correlation matrix. The eigenvalues for the first two factors were significantly higher, with Factor 1 at 8.364 and Factor 2 at 6.747(see factor Analysis in Appendix A). Together, these two factors explain 53.97% of the total variance. The scree plot in Figure below supported the retention of two factors, showing a steep decline after the second factor.



The initial factor loadings showed that Factor 1 had strong positive correlations with variables related to temperature (e.g., T1_num at 0.906, T2_num at 0.802, T3_num at 0.898). In contrast, Factor 2 strongly correlated with relative humidity variables (e.g., RH_1_num at 0.913, RH_2_num at 0.792, RH_3_num at 0.890) (see factor Analysis in Appendix A). This indicated that temperature and humidity were the data structure's primary dimensions.

Varimax rotation improved the interpretability of the factors; the rotated factor loadings reinforced the initial findings. Factor 1 continued to have high loadings on temperature-related variables (e.g., T1_num at 0.930, T2_num at 0.838, T3_num at 0.931), while Factor 2 maintained high loadings on humidity-related variables (e.g., RH_1_num at 0.897, RH_2_num at 0.807, RH_3_num at 0.922) (see factor Analysis in Appendix A). This rotation clarified the distinction between the two factors.

The commonalities, representing the proportion of each variable's variance explained by the retained factors, were high for most temperature and humidity variables, indicating they were well-represented by the two factors. T1_num and RH_1_num had commonalities of 0.868 and 0.834, respectively. Conversely, variables like Appliances_num and Visibility_num had low communalities, suggesting they could have been more effectively captured.

The orthogonal transformation matrix revealed linear solid relationships between the original and rotated factors, reinforcing the reliability of the factor structure. The standardized scoring coefficients used to compute factor scores showed that temperature variables heavily influenced Factor 1, while humidity variables predominantly affected Factor 2.

MDS

The MDS analysis's iterative process involved several iterations to minimize the badness-of-fit criterion, which indicates the degree of dissimilarity between observed and predicted distances. The convergence criteria were ultimately satisfied, indicating the solution's stability. The final badness-of-fit criterion of 0.1408 highlighted this convergence (see table 3)

Table 3

Multidimensonal Scaling

Multidimensional Scaling: Data=WORK.DISTANCE_LONG.DATA Shape=TRIANGLE Condition=MATRIX Level=ORDINAL Coef=IDENTITY Dimension=2 Formula=1 Fit=1

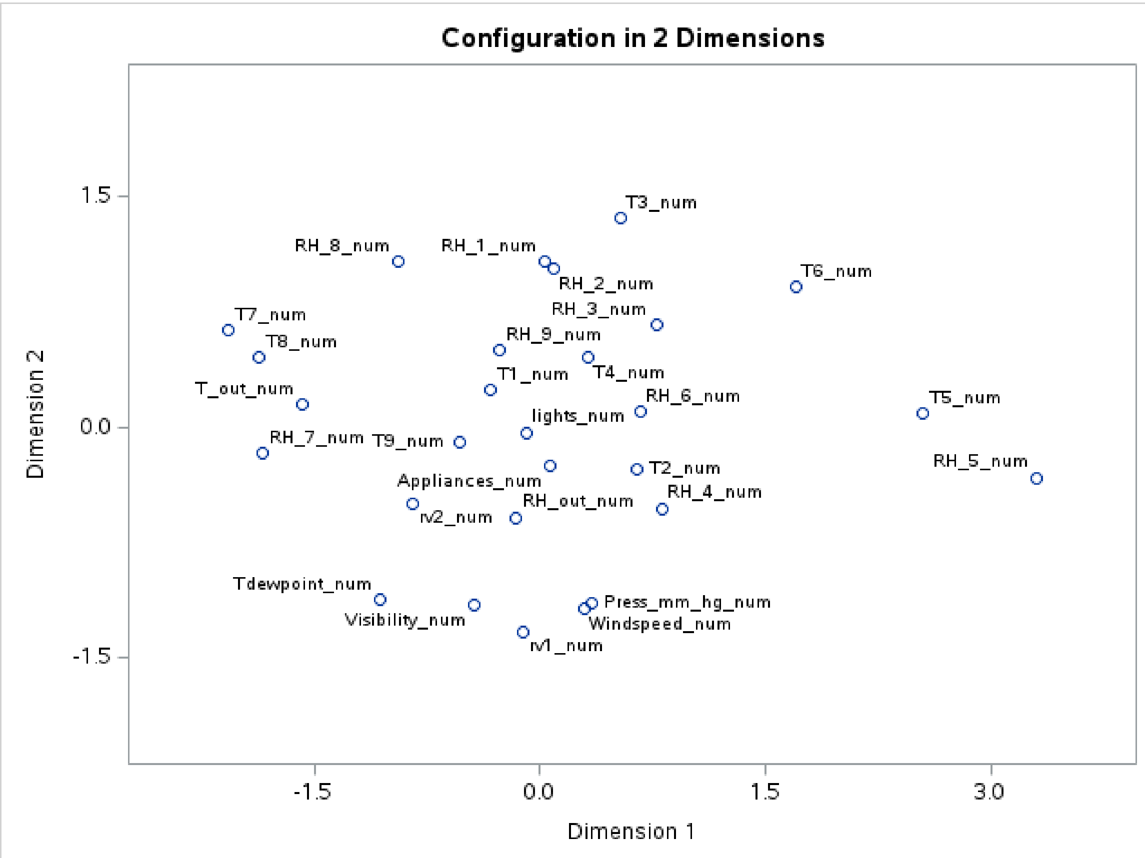
Mconverge=0.01 Gconverge=0.01 Maxiter=100 Over=2 Ridge=0.0001

Iteration	Type	Badness-of-Fit Criterion	Change in Criterion	Convergence Measures	
				Monotone	Gradient
0	Initial	0.2156	.	.	.
1	Monotone	0.1793	0.0363	0.1108	0.4981
2	Gau-New	0.1552	0.0241	.	.
3	Monotone	0.1495	0.005656	0.0383	0.2254
4	Gau-New	0.1484	0.001172	.	.
5	Monotone	0.1421	0.006230	0.0415	0.1024
6	Gau-New	0.1418	0.000297	.	.
7	Monotone	0.1414	0.000404	0.0105	0.0811
8	Gau-New	0.1414	0.0000888	.	.
9	Monotone	0.1412	0.000127	0.005844	0.0774
10	Gau-New	0.1408	0.000417	.	0.0182
11	Gau-New	0.1408	0.0000351	.	0.0116
12	Gau-New	0.1408	0.0000168	.	0.008932

Convergence criteria are satisfied.

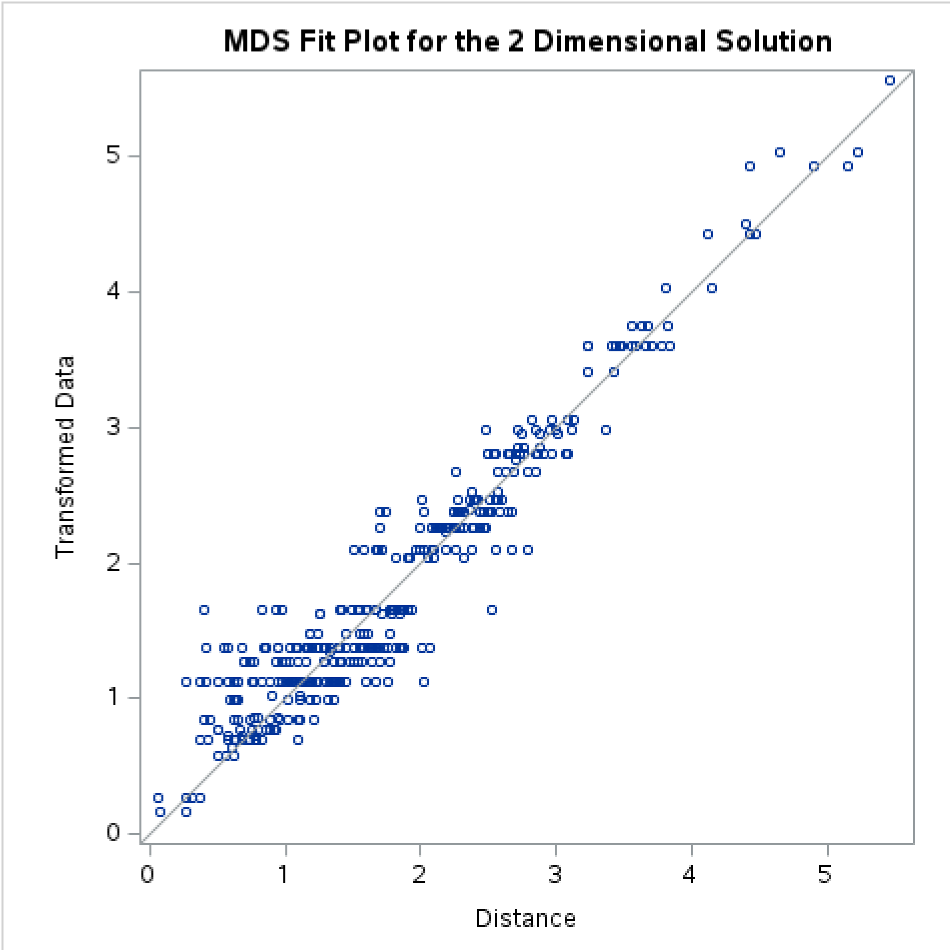
The MDS plot revealed how observations are spatially arranged based on their dissimilarities. In this plot, each point represents an observation, and the proximity of points signifies their similarity. Observations that are close together are more similar, while far-apart ones are more dissimilar.

Figure 5: MDS Plot



The MDS fit plot (see Figure 6) for the 2-dimensional solution demonstrates a high degree of fit, as evidenced by the alignment of transformed data points with the distance metric. This indicates that the two-dimensional representation effectively captures the structure of the original high-dimensional data.

Figure 5 Continued



4.4 Correspondence Analysis

Table 4

Inertia and Chi-Square Decomposition

Inertia and Chi-Square Decomposition

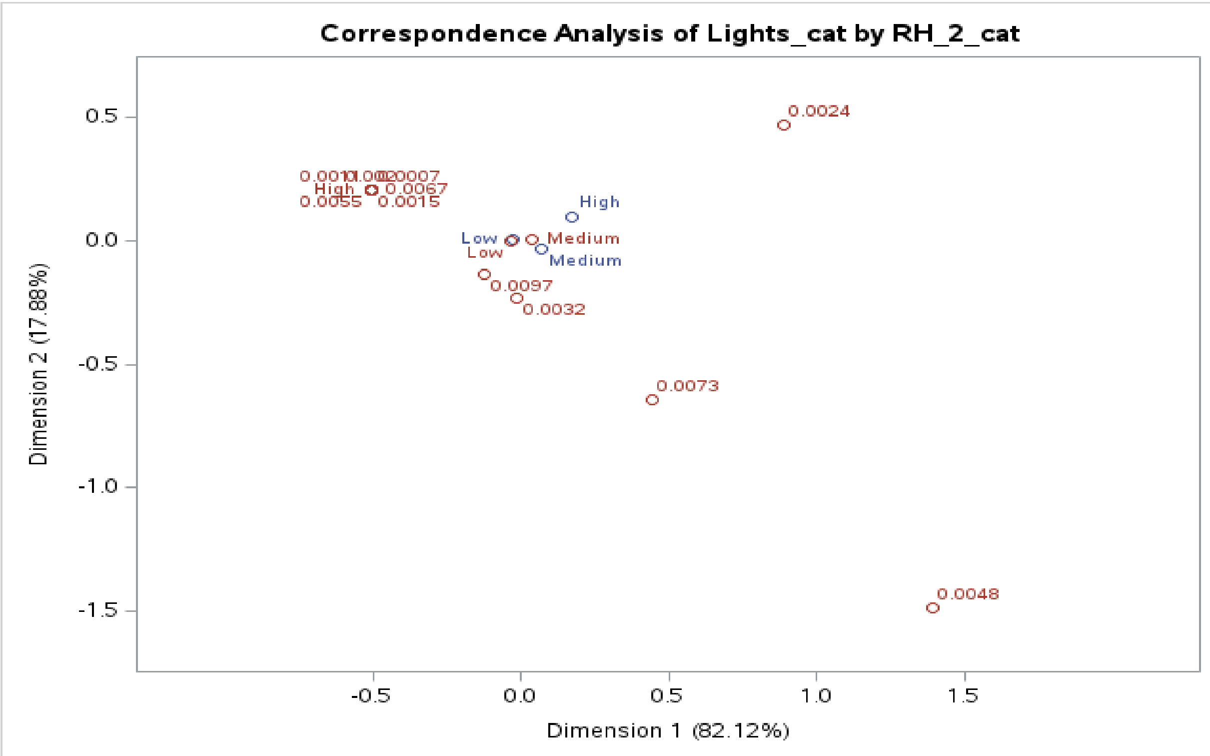
Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent
0.04833	0.00234	46.0991	82.12	82.12
0.02255	0.00051	10.0375	17.88	100.00
	0.00284	56.1366	100.00	

Degrees of Freedom = 26

Correspondence analysis explored the relationships between categorical variables (Lights_cat) and (RH_2_cat). It identified the most significant associations between these sets of variables, unveiling how these factors influence energy consumption patterns. The row and column coordinates provide a visual representation (see Figure 6) of these relationships in a reduced-dimensional space, where each category's position reflects its association strength.

The analysis provides insights into how these variables interact to affect energy usage by exploring relationships between indoor conditions (represented by Lights_cat) and outdoor weather conditions (indicated by RH_2_cat). Specifically, it elaborates how weather conditions, such as temperature and humidity (RH_2_cat), relate to factors like lighting usage (Lights_cat). The correspondence analysis facilitated the identification of key associations between past energy usage and environmental conditions,

Figure 6: Correspondence Analysis of Lights and Humidity in living room (RH_2_cat)



4.5 Canonical Correlation Analysis with PROC CANCORR

Canonical correlation analysis revealed significant relationships between environmental factors (such as temperature and humidity in various rooms) and energy usage (appliances and lights). The analysis extracted two canonical functions, each highlighting different aspects of the relationship between these variables.

Table 5

Canonical Correlation Analysis

The CANCORR Procedure

Canonical Correlation Analysis

Note: The correlation matrix for the Environmental Factors is less than full rank. wTherefore, some canonical coefficients will be zero.

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				Test of H0: The canonical correlations in the current row and all that follow are zero				
					Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.482280	0.481219	0.005463	0.232594	0.3031	0.1716	0.6974	0.6974	0.67821195	168.92	50	39416	<.0001
2	0.340923	0.339520	0.006291	0.116228	0.1315		0.3026	1.0000	0.88377182	108.00	24	19709	<.0001

The first canonical correlation is 0.482, with an adjusted canonical correlation of 0.481 and a squared canonical correlation of 0.233. The second canonical correlation is 0.341, with an adjusted canonical correlation of 0.340 and a squared canonical correlation of 0.116(see table 5). These values indicate moderate relationships between the sets of variables, with the first function being stronger than the second.

Table 6

Canonical Structure

The CANCORR Procedure

Canonical Structure

Correlations Between the Environmental Factors and Their Canonical Variables		
	Env1	**Env2**
T1_num	-0.0079	0.1923

Correlations Between the Environmental Factors and Their Canonical Variables		
	Env1	Env2
RH_1_num	0.2530	0.0898
T2_num	0.0667	0.3505
RH_2_num	0.0554	-0.2465
T3_num	-0.1254	0.3840
RH_3_num	0.2658	-0.0871
T4_num	0.0095	0.1277
RH_4_num	0.2234	-0.1187
T5_num	-0.1329	0.1708
RH_5_num	0.2656	-0.1854
T6_num	-0.0705	0.4502
RH_6_num	0.2308	-0.4606
T7_num	-0.2337	0.2702
RH_7_num	0.0291	-0.2096
T8_num	-0.1067	0.2166
RH_8_num	-0.0365	-0.2868
T9_num	-0.2849	0.2575
RH_9_num	-0.0493	-0.1340
T_out_num	0.0200	0.0048
Press_mm_hg_num	-0.0420	-0.0841
RH_out_num	0.0289	-0.5337

Correlations Between the Environmental Factors and Their Canonical Variables		
	Env1	Env2
Windspeed_num	0.1674	0.1608
Visibility_num	0.0372	-0.0285
Tdewpoint_num	0.0060	0.0048
rv1_num	-0.0062	-0.0325
rv2_num	-0.0062	-0.0325

Table 7

Multivariate Statistics

Multivariate Statistics and F Approximations					
S=2 M=11 N=9853					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					
Wilks' Lambda	0.67821195	168.92	50	39416	<.0001
Pillai's Trace	0.34882200	166.55	50	39418	<.0001
Hotelling-Lawley Trace	0.43460468	171.30	50	36596	<.0001
Roy's Greatest Root	0.30309089	238.94	25	19709	<.0001

The multivariate tests in table 7 yielded significant results ($p < 0.0001$), indicating statistical significance of the canonical correlations. Table 6 shows correlations between the original variables and their respective canonical variables, highlighting the strength of the relationship. The correlation between humidity in the kitchen (RH_1) and the first canonical variable (Env1) is 0.253, indicating a moderate association.

The first canonical function demonstrated a moderate relationship, primarily driven by variables such as humidity in the kitchen (RH_1) and temperature in the living room (T2). The second function, while weaker, showed significant relationships, particularly with variables like temperature in the laundry room (T3) and appliance usage. These findings suggest that specific environmental conditions, especially humidity and temperature measures, are crucial in predicting energy usage patterns.

Table 8

Canonical Correlation Analysis

The CANCORR Procedure

Canonical Correlation Analysis

Raw Canonical Coefficients for the Environmental Factors		
	Env1	**Env2**
T1_num	0.0956054206	-0.137900759
RH_1_num	0.2066215687	0.3652293397
T2_num	-0.416956337	-0.277915658
RH_2_num	-0.199260468	-0.308716092
T3_num	0.1580900442	0.7109108285
RH_3_num	0.0461940068	0.1259574776

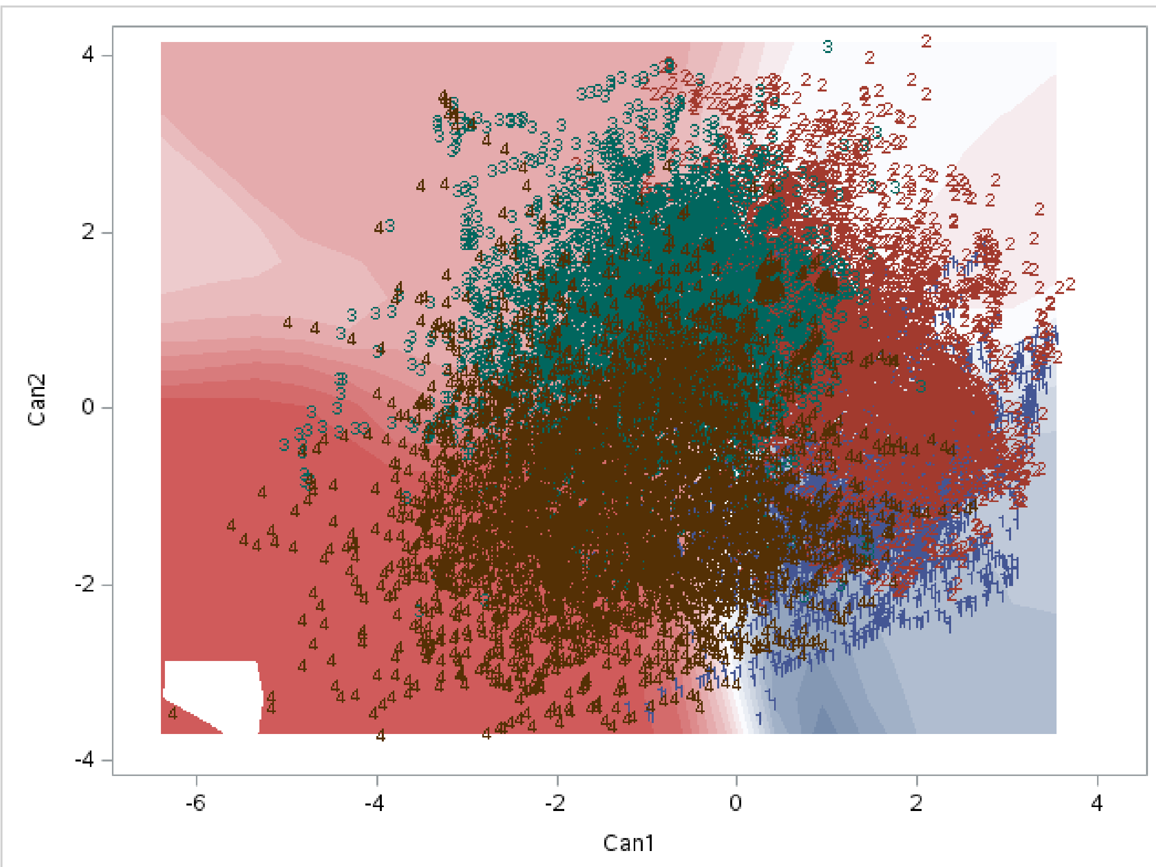
Raw Canonical Coefficients for the Environmental Factors		
	Env1	Env2
T4_num	0.7747072019	-0.51254824
RH_4_num	0.2616524122	-0.14176947
T5_num	-0.016679897	-0.028503461
RH_5_num	0.0135520859	-0.001251812
T6_num	0.0568879567	0.0681374454
RH_6_num	0.0073788409	0.0004019398
T7_num	-0.097062237	0.0719442753
RH_7_num	-0.047005702	-0.026650221
T8_num	0.2972142214	0.093286855
RH_8_num	-0.193642635	-0.063661759
T9_num	-0.857933297	-0.034875889
RH_9_num	-0.07758793	0.0145926643
T_out_num	-0.014172122	-0.000510982
Press_mm_hg_num	-0.001827977	0.0039567672
RH_out_num	0.0272994622	-0.000579518
Windspeed_num	0.0691798321	0.0269105451
Visibility_num	0.0031396084	0.002316258
Tdewpoint_num	-0.013215857	-0.015785915
rv1_num	-0.000222195	-0.001246338
rv2_num	0	0

Raw Canonical Coefficients for the Energy Usage		
	Energy1	Energy2
Appliances_num	0.0030215841	0.0094793304
lights_num	0.1123557349	-0.062427914

4.6 Canonical Discriminant Analysis

Figure 7

Discriminant Analysis Visualization



The scatter plot from the Canonical Discriminant Analysis (CDA) depicts the distribution of Can1 and Can2. Each point represents an observation, and the numbers indicate the group to which each observation belongs.

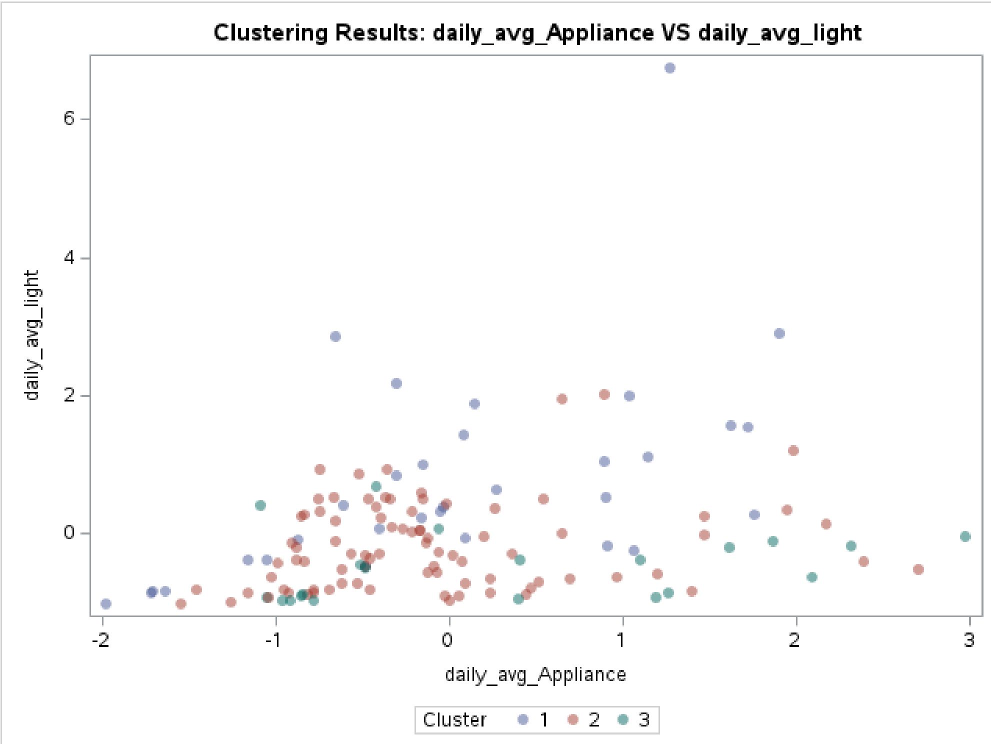
There is a noticeable overlap between the groups, especially in the central region of the plot, indicating that while the canonical variables provide some separation, there is still a significant degree of similarity between the groups. Despite some overlap, the analysis highlights significant differences between the groups, with particular effectiveness in distinguishing Group 2.

4.7 Clustering

The K-means clustering analysis revealed distinct groupings in energy consumption patterns for daily average appliances and daily average lights. Clusters represent varying usage levels, from high to low, suggesting diverse energy usage. This segmentation aids in understanding factors influencing energy usage variability.

Figure 8

K - means Clustering for Daily average Appliance vs daily average light.



4.8 PLS Regression

The PLS regression analysis effectively elaborated the energy usage patterns in a low-energy building, highlighting the significant relationships between indoor temperatures, humidity levels, outdoor weather conditions, and energy usage. The model, employing five optimal factors, captured 91.11% of the variance in predictor variables and 73.03% in energy usage, indicating a robust fit and comprehensive explanatory power. The Correlation Loading Plot (see Figure 10) emphasises the importance of indoor temperature and humidity, which show strong correlations with energy usage.

Table 9

PLS Procedure

The PLS Procedure

Percent Variation Accounted for by Partial Least Squares Factors				
	Model Effects		Dependent Variables	
Number of Extracted Factors	Current	Total	Current	Total
1	54.7978	54.7978	5.7739	5.7739
2	27.8571	82.6548	12.1742	17.9481
3	5.2053	87.8601	38.6263	56.5745
4	2.3017	90.1618	12.7405	69.3150
5	0.9517	91.1135	3.7184	73.0334

The cross-validation analysis confirms the model's reliability, with a minimum Root Mean PRESS of 0.6702 achieved using five factors. Additionally, including two random variables demonstrates the model's ability to filter out non-predictive attributes, further validating its predictive accuracy.

Minimum root mean PRESS	0.6702
Minimizing number of factors	5

Figure 9

Cross Validation

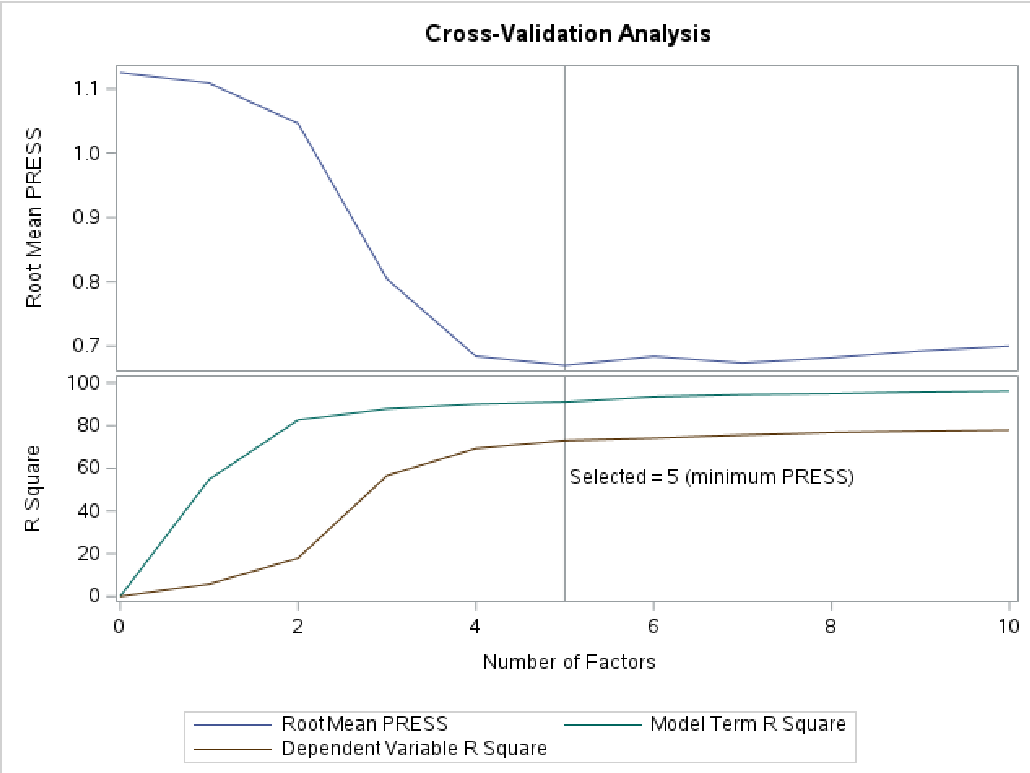
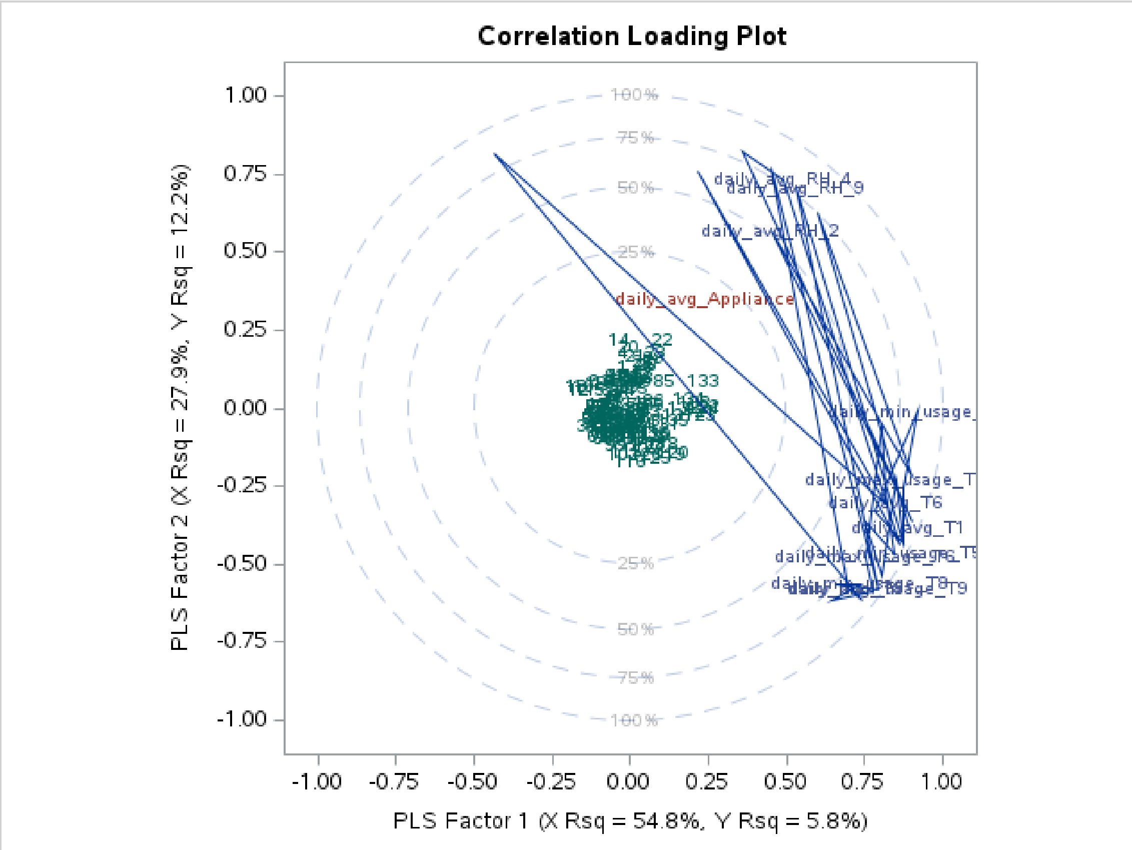


Figure 10

Correlation loading plot


Correlation Loading Plot
PLS Factor 2 (X Rsq = 27.9%, Y Rsq = 12.2%)
PLS Factor 1 (X Rsq = 54.8%, Y Rsq = 5.8%)
daily_avg_RH_4
daily_avg_RH_9
daily_avg_RH_2
daily_avg_Appliance
daily_min_usage_
daily_avg_T6
daily_avg_T1

5.0 Discussion

Findings from the study provide helpful knowledge of the relationships between environmental factors, weather conditions, and energy usage patterns in low-energy buildings. Using a multidimensional analysis method, the study presented vital relationships and trends concerning energy use. It supported the hypotheses that outdoor weather conditions, indoor temperature, humidity levels, and past energy use are critical in predicting energy consumption within low-energy buildings. PLS regression accounted for variances up to 91.11% of the predictor variables' variance and up to 73.03% of energy usage variance, indicating good predictive properties of the model.

Limitations

However, the study's limitations are the possible confounding variables—occupants' behaviour and building design features- that have not been considered when performing the analyses. The general identification of critical factors that determine energy consumption, mainly based on this study, permits its use for a decision-making process related to energy management strategies in low-energy buildings with the objective of an optimal energy supply that makes the operation of buildings economically and ecologically sustainable. In addition, considering indoor environmental conditions, outdoor weather factors, and past energy usage allows stakeholders to develop more effective strategies to reduce energy consumption and carbon footprints.

Future Directions

Future studies should also consider additional research variables, such as occupant behaviour and building design features, to delineate predictive models of energy use further. Other longitudinal studies, which observe energy consumption over time, will enable taking notice of seasonal variations and tendencies from the longer view, thus working out more robust strategies. Possible strategies to optimize energy performance in low-energy buildings could be defined from research on implementing real-time monitoring with advanced control strategies. Overall, this study emphasized holistic approaches to energy management and reveal potential data-driven techniques for driving sustainable practices towards operations in buildings.

6.0 References

Allee, K. D., Do, C., & Raymundo, F. G. (2022). Principal component analysis and factor analysis in accounting research. *Journal of Financial Reporting /Journal of Financial Reporting*, *7*(2), 1–39. https://doi.org/10.2308/jfr-2021-005

Allouhi, A., Fouih, Y. E., Kousksou, T., Jamil, A., Zeraouli, Y., & Mourad, Y. (2015). Energy consumption and efficiency in buildings: current status and future trends. *Journal of Cleaner Production*, *109*, 118–130. https://doi.org/10.1016/j.jclepro.2015.05.139

Arghira, N., Hawarah, L., Ploix, S., & Jacomino, M. (2012). Prediction of appliances energy use in smart homes. *Energy*, *48*(1), 128–134. https://doi.org/10.1016/j.energy.2012.04.010

Ariza, A. G., Arbulu, A. A., González, F. J. N., Bermejo, J. V. D., & Vallejo, M. E. C. (2021). Discriminant Canonical Analysis as a validation tool for multi-variety native breed egg commercial quality classification. *Foods*, *10*(3), 632. https://doi.org/10.3390/foods10030632

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*, 81–97. https://doi.org/10.1016/j.enbuild.2017.01.083

Cao, X., Dai, X., & Liu, J. (2016). Building energy-consumption status worldwide and the state-of-the-art technologies for zero-energy buildings during the past decade. *Energy and Buildings*, *128*, 198–213. https://doi.org/10.1016/j.enbuild.2016.06.089

Cetin, K. S. (2016). Characterizing large residential appliance peak load reduction potential utilizing a probabilistic approach. *Science and Technology for the Built Environment*, *22*(6), 720-732.

Cetin, K., Tabares-Velasco, P., & Novoselac, A. (2014). Appliance daily energy use in new residential buildings: Use profiles and variation in time-of-use. *Energy and Buildings*, *84*, 716–726. https://doi.org/10.1016/j.enbuild.2014.07.045

Du, Y., Li, F., Kurte, K., Munk, J., & Zandi, H. (2022). Demonstration of intelligent HVAC load management with deep reinforcement learning: Real-World experience of Machine Learning in demand control. *IEEE Power & Energy Magazine*, *20*(3), 42–53. https://doi.org/10.1109/mpe.2022.3150825

Guo, Z., Wang, Z. J., & Kashani, A. (2014). Home appliance load modelling from aggregated smart meter data. *IEEE Transactions on power systems*, *30*(1), 254-262.

Hong, T., Taylor-Lange, S. C., D'Oca, S., Yan, D., & Corgnati, S. P. (2016). Advances in research and applications of energy-related occupant behavior in buildings. *Energy and Buildings*, *116*, 694–702. https://doi.org/10.1016/j.enbuild.2015.11.052

Kavousian, A., Rajagopal, R., & Fischer, M. (2013). Determinants of residential electricity consumption: Using smart meter data to examine the effect of climate, building characteristics, appliance stock, and occupants' behavior. *Energy*, *55*, 184–194. https://doi.org/10.1016/j.energy.2013.03.086

Kavousian, A., Rajagopal, R., & Fischer, M. (2015). Ranking appliance energy efficiency in households: Utilizing smart meter data and energy efficiency frontiers to estimate and identify the determinants of appliance energy efficiency in residential buildings. *Energy and Buildings*, *99*, 220–230. https://doi.org/10.1016/j.enbuild.2015.03.052

Kodinariya, T. M., & Makwana, P. (2013). Review on Determining of Cluster in K-Means Clustering. *ResearchGate*.

https://www.researchgate.net/publication/313554124_Review_on_Determining_of_Cluster_in_K-means_Clustering

Oti, E. U., Olusola, M. O., Eze, F. C., & Enogwe, S. U. (2021). Comprehensive review of K-Means Clustering Algorithms. *International Journal of Advances in Scientific Research and Engineering*, *07*(08), 64–69. https://doi.org/10.31695/ijasre.2021.34050

Ramli, N. A., & Shapi, M. K. M. (2022). Building Energy Management. In *Studies in Infrastructure and Control* (pp. 37–73). https://doi.org/10.1007/978-981-19-0375-5_3

Reeves, J. B., & Delwiche, S. R. (2008). SAS® Partial Least Squares for Discriminant analysis. *Journal of Near Infrared Spectroscopy*, *16*(1), 31–38. https://doi.org/10.1255/jnirs.757

Riani, M., Atkinson, A. C., Torti, F., & Corbellini, A. (2022). Robust correspondence analysis. *Applied Statistics/Journal of the Royal Statistical Society. Series C, Applied Statistics*, *71*(5), 1381–1401. https://doi.org/10.1111/rssc.12580

Tucker-Drob, E. M., & Salthouse, T. A. (2009). METHODS AND MEASURES: Confirmatory factor analysis and multidimensional scaling for construct validation of cognitive abilities. *International Journal of Behavioral Development*, *33*(3), 277–285. https://doi.org/10.1177/0165025409104489

UCI Machine Learning Repository. (2017). https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction

Use of energy in commercial buildings - U.S. Energy Information Administration (EIA). (2018). https://www.eia.gov/energyexplained/use-of-energy/commercial-buildings.php

7.0: Appendix

Appendix A

Factor analysis

The FACTOR Procedure

Input Data Type	Raw Data
Number of Records Read	19735
Number of Records Used	19735
N for Significance Tests	19735

The FACTOR Procedure
Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 28 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	8.36391608	1.61713651	0.2987	0.2987
2	6.74677957	4.73952193	0.2410	0.5397
3	2.00725764	0.22203118	0.0717	0.6114
4	1.78522646	0.43955484	0.0638	0.6751
5	1.34567163	0.18793895	0.0481	0.7232
6	1.15773268	0.14088556	0.0413	0.7645
7	1.01684712	0.04049376	0.0363	0.8008
8	0.97635336	0.10142099	0.0349	0.8357
9	0.87493237	0.13986452	0.0312	0.8670

Eigenvalues of the Correlation Matrix: Total = 28 Average = 1				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
10	0.73506785	0.12889218	0.0263	0.8932
11	0.60617567	0.08771918	0.0216	0.9149
12	0.51845649	0.06067677	0.0185	0.9334
13	0.45777972	0.20873528	0.0163	0.9497
14	0.24904443	0.06416787	0.0089	0.9586
15	0.18487656	0.03057107	0.0066	0.9652
16	0.15430549	0.01333131	0.0055	0.9707
17	0.14097418	0.00653152	0.0050	0.9758
18	0.13444267	0.01930781	0.0048	0.9806
19	0.11513485	0.02148678	0.0041	0.9847
20	0.09364807	0.01055418	0.0033	0.9880
21	0.08309389	0.01447625	0.0030	0.9910
22	0.06861764	0.00954427	0.0025	0.9934
23	0.05907338	0.01516863	0.0021	0.9956
24	0.04390475	0.00319515	0.0016	0.9971
25	0.04070960	0.01542731	0.0015	0.9986
26	0.02528228	0.01058673	0.0009	0.9995
27	0.01469555	0.01469555	0.0005	1.0000
28	0.00000000		0.0000	1.0000

Factor Pattern

	Factor1	Factor2
Appliances_num	0.07987	-0.00323
lights_num	-0.11302	0.07922
T1_num	0.90577	0.21910
RH_1_num	0.01355	0.91292
T2_num	0.80219	0.27945
RH_2_num	-0.15916	0.79189
T3_num	0.89780	0.27475
RH_3_num	-0.26483	0.88953
T4_num	0.92269	0.10597
RH_4_num	-0.12455	0.94937
T5_num	0.89783	0.23736
RH_5_num	-0.15430	0.38240
T6_num	0.75448	0.30092
RH_6_num	-0.84837	0.32723
T7_num	0.93801	0.01488
RH_7_num	-0.06984	0.93415
T8_num	0.87268	-0.02439
RH_8_num	-0.23089	0.88686
T9_num	0.94061	0.11331
RH_9_num	-0.12661	0.89242

Factor Pattern		
	Factor1	Factor2
T_out_num	-0.06216	0.07162
Press_mm_hg_num	-0.10585	-0.34410
RH_out_num	-0.54182	0.37741
Windspeed_num	-0.13330	0.25754
Visibility_num	-0.11692	-0.01086
Tdewpoint_num	0.22949	0.41562
rv1_num	-0.01050	-0.00154
rv2_num	-0.01050	-0.00154

Variance Explained by Each Factor	
Factor1	Factor2
8.3639161	6.7467796

Final Communality Estimates: Total = 15.110696

Appliances_num	lights_num	T1_num	RH_1_num	T2_num	RH_2_num	T3_num	RH_3_num	T4_num	RH_4_num	T5_num	RH_5_num	T6_num	RH_6_num	T7_num	RH_7_num	T8_num	RH_8_num	T9_num	RH_9_num	T_out_num	Press_mm_hg_num	RH_out_num	Windspeed_num	Visibility_num	Tdewpoint_num	rv1_num	rv2_num
0.006390	0.01904	0.86842	0.83360	0.72161	0.65242	0.88153	0.86140	0.86259	0.91681	0.86243	0.17003	0.65980	0.82681	0.88008	0.87752	0.76216	0.83983	0.89758	0.81245	0.00899	0.12960622	0.43600	0.084097	0.013787	0.225405	0.00011	0.00011

Final Communality Estimates: Total = 15.110696																											
Appliances_num	lights_num	T1_num	RH_1_num	T2_num	RH_2_num	T3_num	RH_3_num	T4_num	RH_4_num	T5_num	RH_5_num	T6_num	RH_6_num	T7_num	RH_7_num	T8_num	RH_8_num	T9_num	RH_9_num	T_out_num	Press_mm_hg_num	RH_out_num	Windspeed_num	Visibility_num	Tdewpoint_num	rv1_num	rv2_num
05	840	740	252	007	112	547	357	410	347	421	733	038	672	919	275	490	319	255	184	351		195	28	41	00	252	252

The FACTOR Procedure
Rotation Method: Varimax

Orthogonal Transformation Matrix		
	1	2
1	0.98550	-0.16965
2	0.16965	0.98550

Rotated Factor Pattern		
	Factor1	Factor2
Appliances_num	0.07817	-0.01674
lights_num	-0.09794	0.09724
T1_num	0.92981	0.06226
RH_1_num	0.16823	0.89739
T2_num	0.83797	0.13931
RH_2_num	-0.02251	0.80741

Rotated Factor Pattern		
	Factor1	**Factor2**
T3_num	0.93140	0.11846
RH_3_num	-0.11009	0.92157
T4_num	0.92730	-0.05210
RH_4_num	0.03831	0.95674
T5_num	0.92508	0.08161
RH_5_num	-0.08719	0.40303
T6_num	0.79460	0.16856
RH_6_num	-0.78056	0.46641
T7_num	0.92694	-0.14447
RH_7_num	0.08965	0.93246
T8_num	0.85589	-0.17208
RH_8_num	-0.07710	0.91318
T9_num	0.94620	-0.04791
RH_9_num	0.02662	0.90097
T_out_num	-0.04911	0.08113
Press_mm_hg_num	-0.16269	-0.32115
RH_out_num	-0.46994	0.46385
Windspeed_num	-0.08767	0.27642
Visibility_num	-0.11706	0.00913
Tdewpoint_num	0.29667	0.37066

Rotated Factor Pattern		
	Factor1	Factor2
rv1_num	-0.01060	0.00027
rv2_num	-0.01060	0.00027

Variance Explained by Each Factor	
Factor1	Factor2
8.3173750	6.7933207

Final Communality Estimates: Total = 15.110696

Appliances_num	lights_num	T1_num	RH_1_num	T2_num	RH_2_num	T3_num	RH_3_num	T4_num	RH_4_num	T5_num	RH_5_num	T6_num	RH_6_num
0.00639005	0.01904840	0.86842740	0.83360252	0.72161007	0.65242112	0.88153547	0.86140357	0.86259410	0.91681347	0.86243421	0.17003733	0.65980038	0.82681672

T7_num	RH_7_num	T8_num	RH_8_num	T9_num	RH_9_num	T_out_num	Press_mm_hg_num	RH_out_num	Windspeed_num	Visibility_num	Tdewpoint_num	rv1_num	rv2_num
0.88008919	0.87752275	0.76216490	0.83983319	0.89758255	0.81245184	0.00899351	0.12960622	0.43600195	0.08409728	0.01378741	0.22540500	0.00011252	0.00011252

The FACTOR Procedure
Rotation Method: Varimax

Scoring Coefficients Estimated by Regression

Squared Multiple Correlations of the Variables with Each Factor	
Factor1	**Factor2**
1.0000000	1.0000000

Standardized Scoring Coefficients

	Factor1	Factor2
Appliances_num	0.00933	-0.00209
lights_num	-0.01132	0.01386
T1_num	0.11223	0.01363
RH_1_num	0.02455	0.13308
T2_num	0.10155	0.02455
RH_2_num	0.00116	0.11890
T3_num	0.11269	0.02192
RH_3_num	-0.00884	0.13531
T4_num	0.11138	-0.00324
RH_4_num	0.00920	0.14120
T5_num	0.11176	0.01646
RH_5_num	-0.00856	0.05899
T6_num	0.09647	0.02865
RH_6_num	-0.09173	0.06501
T7_num	0.11090	-0.01685
RH_7_num	0.01526	0.13787
T8_num	0.10221	-0.02126

Standardized Scoring Coefficients		
	Factor1	Factor2
RH_8_num	-0.00491	0.13423
T9_num	0.11368	-0.00253
RH_9_num	0.00752	0.13292
T_out_num	-0.00552	0.01172
Press_mm_hg_num	-0.02112	-0.04812
RH_out_num	-0.05435	0.06612
Windspeed_num	-0.00923	0.04032
Visibility_num	-0.01405	0.00078
Tdewpoint_num	0.03749	0.05605
rv1_num	-0.00255	-0.00002
rv2_num	0.00000	0.00000

Appendix B

Sas Code

```
/* IMPORT THE energy_complete.csv data*/

/* Convert character variables to numeric type */
data WORK.IMPORT_NUMERIC;
       set WORK.IMPORT;

       /* Convert the date string to SAS datetime and date values */
    datetime = input(date, anydtdtm19.);
    format datetime datetime20.;
    date_num = datepart(datetime); /* Extract the date part */
    days = DAY(date_num);

    /* Convert character variables to numeric */
    Appliances_num = input(Appliances, best12.);
    lights_num = input(lights, best12.);
    T1_num = input(T1, best12.);
    RH_1_num = input(RH_1, best12.);
    T2_num = input(T2, best12.);
    RH_2_num = input(RH_2, best12.);
    T3_num = input(T3, best12.);
    RH_3_num = input(RH_3, best12.);
    T4_num = input(T4, best12.);
    RH_4_num = input(RH_4, best12.);
```

```
    T5_num = input(T5, best12.);
    RH_5_num = input(RH_5, best12.);
    T6_num = input(T6, best12.);
    RH_6_num = input(RH_6, best12.);
    T7_num = input(T7, best12.);
    RH_7_num = input(RH_7, best12.);
    T8_num = input(T8, best12.);
    RH_8_num = input(RH_8, best12.);
    T9_num = input(T9, best12.);
    RH_9_num = input(RH_9, best12.);
    T_out_num = input(T_out, best12.);
    Press_mm_hg_num = input(Press_mm_hg, best12.);
    RH_out_num = input(RH_out, best12.);
    Windspeed_num = input(Windspeed, best12.);
    Visibility_num = input(Visibility, best12.);
    Tdewpoint_num = input(Tdewpoint, best12.);
    rv1_num = input(rv1, best12.);
    rv2_num = input(rv2, best12.);

    /* Extract the hour from the datetime */
    hour = hour(datetime);

    /* Create a numeric time segment variable */
    if hour < 6 then time_segment = 1; /* Night */
    else if hour < 12 then time_segment = 2; /* Morning */
    else if hour < 18 then time_segment = 3; /* Afternoon */
    else time_segment = 4; /* Evening */

    /* Apply appropriate formats */
    format datetime datetime20.;

    /* Create weekday and weekend variables */
    day_of_week = weekday(date_num); /* 1=Sunday, 2=Monday, ...,
7=Saturday */
    weekday = (day_of_week in (2, 3, 4, 5, 6)); /* 1=True if Monday-
Friday, 0=False */
    weekend = (day_of_week in (1, 7)); /* 1=True if Sunday or
Saturday, 0=False */
    day_of_month = day(date_num);

    /* Apply appropriate formats */
    format datetime datetime20.;
    format date_num date9.;

    /* Drop original character variables */
    drop T1 RH_1 T2 RH_2 T3 RH_3 T4 RH_4 T5 RH_5 T6 RH_6 T7 RH_7 T8
```

```
RH_8 T9 RH_9 T_out Press_mm_hg RH_out Windspeed Visibility Tdewpoint
rv1 rv2;
run;

/* Calculate the daily averages*/
proc sql;
    create table daily_averages as
    select date_num,
           mean(Appliances_num) as daily_avg_Appliance,
           mean(lights_num) as daily_avg_light,
           mean(T1_num) as daily_avg_T1,
           mean(T2_num) as daily_avg_T2,
           mean(T3_num) as daily_avg_T3,
           mean(T4_num) as daily_avg_T4,
           mean(T5_num) as daily_avg_T5,
           mean(T6_num) as daily_avg_T6,
           mean(T7_num) as daily_avg_T7,
           mean(T8_num) as daily_avg_T8,
           mean(T9_num) as daily_avg_T9,
           mean(RH_1_num) as daily_avg_RH_1,
           mean(RH_2_num) as daily_avg_RH_2,
           mean(RH_3_num) as daily_avg_RH_3,
           mean(RH_4_num) as daily_avg_RH_4,
           mean(RH_5_num) as daily_avg_RH_5,
           mean(RH_6_num) as daily_avg_RH_6,
           mean(RH_7_num) as daily_avg_RH_7,
           mean(RH_8_num) as daily_avg_RH_8,
           mean(RH_9_num) as daily_avg_RH_9,
           mean(T_out_num) as daily_avg_T_out,
           mean(Press_mm_hg_num) as daily_avg_Press_mm_hg,
           mean(RH_out_num) as daily_avg_RH_out,
           mean(Windspeed_num) as daily_avg_Windspeed,
           mean(Visibility_num) as daily_avg_Visibility,
           mean(Tdewpoint_num) as daily_avg_Tdewpoint,
           mean(rv1_num) as daily_avg_rv1,
           mean(rv2_num) as daily_avg_rv2,
           min(Appliances_num) as daily_min_usage_Appliances,
           max(Appliances_num) as daily_max_usage_Appliances,
           min(T1_num) as daily_min_usage_T1,
           max(T1_num) as daily_max_usage_T1,
           min(T2_num) as daily_min_usage_T2,
           max(T2_num) as daily_max_usage_T2,
           min(T3_num) as daily_min_usage_T3,
           max(T3_num) as daily_max_usage_T3,
           min(T4_num) as daily_min_usage_T4,
           max(T4_num) as daily_max_usage_T4,
```

```
           min(T5_num) as daily_min_usage_T5,
           max(T5_num) as daily_max_usage_T5,
           min(T6_num) as daily_min_usage_T6,
           max(T6_num) as daily_max_usage_T6,
           min(T7_num) as daily_min_usage_T7,
           max(T7_num) as daily_max_usage_T7,
           min(T8_num) as daily_min_usage_T8,
           max(T8_num) as daily_max_usage_T8,
           min(T9_num) as daily_min_usage_T9,
           min(T9_num) as daily_max_usage_T9


    from WORK.IMPORT_NUMERIC

    group by date_num;
quit;

PROC CONTENTS DATA=daily_averages; RUN;

/* Merge the daily average energy use back with the original data */
data WORK.IMPORT_NUMERIC;
    merge WORK.IMPORT_NUMERIC(in=a) daily_averages(in=b);
    by date_num;
    if a;
run;

/* check if the conversion has taken place*/
proc contents data=WORK.IMPORT_NUMERIC;
run;

PROC UNIVARIATE DATA = WORK.IMPORT_NUMERIC;
VAR days datetime weekday day_of_week day_of_month T3_num;
RUN;
PROC UNIVARIATE DATA = WORK.IMPORT_NUMERIC;
VAR T2_num T3_num  T4_num T5_num ;
RUN;

PROC UNIVARIATE DATA = WORK.IMPORT_NUMERIC;
VAR time_segment;
RUN;


/* Create a dataset containing only indoor variables */
data indoor_data;
    set WORK.IMPORT_NUMERIC;
    keep date Appliances_num lights_num T1_num RH_1_num T2_num
```

```
RH_2_num T3_num RH_3_num T4_num RH_4_num T5_num RH_5_num T7_num
RH_7_num T8_num RH_8_num T9_num RH_9_num;
run;

/* Create a dataset containing only outdoor variables */
data outdoor_data;
    set WORK.IMPORT_NUMERIC;
    keep date T6_num Press_mm_hg_num RH_6_num T_out_num RH_out_num
RH_out_num Windspeed_num Visibility_num  Tdewpoint_num rv1_num
rv2_num time_segment;
run;

/* check if the conversion has taken place*/
proc contents data=WORK.indoor_data;
run;

/* check if the conversion has taken place*/
proc contents data=WORK.outdoor_data;
run;

/* Merge indoor and outdoor datasets by date */
data merged_data;
    merge indoor_data (in=a) outdoor_data (in=b);
    by date;
    /* Check for missing values */
    if a and b;
run;

PROC CONTENTS DATA=merged_data; RUN;


/*Principal Components Analysis (PCA) and its visualization*/

/* Standardize the dataset */
proc standard data=WORK.IMPORT_NUMERIC mean=0 std=1
out=energy_standardized;
    var Appliances_num lights_num T1_num RH_1_num T2_num RH_2_num
T3_num RH_3_num
    T4_num RH_4_num T5_num RH_5_num T6_num RH_6_num T7_num RH_7_num
T8_num RH_8_num
    T9_num RH_9_num T_out_num Press_mm_hg_num RH_out_num
Windspeed_num Visibility_num
    Tdewpoint_num rv1_num rv2_num;
run;

/*Eigenvalues*/
```

```
/* Perform PCA */
proc princomp data=energy_standardized out=pca_output
outstat=pca_stats plots=all;
    var Appliances_num lights_num T1_num RH_1_num T2_num RH_2_num
    T3_num RH_3_num T4_num RH_4_num T5_num RH_5_num T6_num RH_6_num
    T7_num RH_7_num T8_num RH_8_num T9_num RH_9_num T_out_num
    Press_mm_hg_num RH_out_num Windspeed_num Visibility_num
Tdewpoint_num rv1_num rv2_num;
run;

/* Perform Factor Analysis with nfactor=2 */
proc factor data=energy_standardized method=principal rotate=varimax
scree nfactor=2 out=factor_scores;
    var Appliances_num lights_num T1_num RH_1_num T2_num RH_2_num
        T3_num RH_3_num T4_num RH_4_num T5_num RH_5_num T6_num
RH_6_num
        T7_num RH_7_num T8_num RH_8_num T9_num RH_9_num T_out_num
        Press_mm_hg_num RH_out_num Windspeed_num Visibility_num
Tdewpoint_num rv1_num rv2_num;
run;

/* Print the factor loadings for interpretation */
proc print data=factor_scores(obs=10);
run;


/* Computing MDS*/
/* Step 1: Compute the Distance Matrix */
proc distance data=energy_standardized method=euclid
out=distance_matrix;
    var Appliances_num lights_num T1_num RH_1_num T2_num RH_2_num
        T3_num RH_3_num T4_num RH_4_num T5_num RH_5_num T6_num
RH_6_num
        T7_num RH_7_num T8_num RH_8_num T9_num RH_9_num T_out_num
        Press_mm_hg_num RH_out_num Windspeed_num Visibility_num
Tdewpoint_num rv1_num rv2_num;
run;

/* Step 2: Transpose the Distance Matrix to a long format */
proc transpose data=distance_matrix out=distance_long(drop=_NAME_);
    var Dist1-Dist28;
run;

/* Step 3: Create unique identifiers for each observation */
data distance_long;
```

```
    set distance_long;
    length Subject $50;

/* Define an array of variable names */
array var_names[28] $50 _temporary_ ('Appliances_num' 'lights_num'
'T1_num' 'RH_1_num' 'T2_num' 'RH_2_num'
                                          'T3_num' 'RH_3_num'
'T4_num' 'RH_4_num' 'T5_num' 'RH_5_num' 'T6_num' 'RH_6_num'
                                          'T7_num' 'RH_7_num'
'T8_num' 'RH_8_num' 'T9_num' 'RH_9_num' 'T_out_num'
                                          'Press_mm_hg_num'
'RH_out_num' 'Windspeed_num' 'Visibility_num' 'Tdewpoint_num'
'rv1_num' 'rv2_num');
/* Assign Subject based on observation number */
if _N_ <= dim(var_names) then Subject = var_names[_N_];
run;

/* Step 4: Perform MDS using the reshaped distance matrix */
proc mds data=distance_long out=mds_out level=ordinal;
    id Subject;
    var COL1-COL28;
run;

/* Step 5: Scatter Plot of MDS results */
proc sgplot data=mds_out;
    scatter x=Dim1 y=Dim2 / datalabel=Subject;
    xaxis label='Dimension 1';
    yaxis label='Dimension 2';
    title 'MDS Plot';
run;


/* Scatter Plot of First Two Principal Components */
proc sgplot data=pca_output;
    scatter x=Prin1 y=Prin2;
    xaxis label='Principal Component 1';
    yaxis label='Principal Component 2';
run;

/* Prepare data for biplot */
/* Extract the principal component loadings */
data loadings;
    set pca_stats(where=(_TYPE_='SCORE'));
    keep _NAME_ Prin1 Prin2;
run;
```

```
/* Create the combined dataset for biplot */
data biplot_data;
    set pca_output(in=a) loadings(in=b);
    if a then type='score';
    if b then type='loading';
run;

/* Biplot */
proc sgplot data=biplot_data;
    vector x=Prin1 y=Prin2 / group=type name='Variable
Contributions';
    scatter x=Prin1 y=Prin2 / group=type;
    xaxis label='Principal Component 1';
    yaxis label='Principal Component 2';
    title 'Biplot of Principal Components';
run;

/* Print the first 20 observations to verify the merge */
proc print data=WORK.IMPORT_NUMERIC(obs=20);
    var date weekend weekday Appliances daily_avg_Appliance
daily_avg_light;
run;

title "Daily average energy use data by appliances";
proc sgplot data=WORK.IMPORT_NUMERIC;
   series x=date_num y=daily_avg_appliance / markers;
    xaxis label="Date";
   yaxis label="Energy Use (Applinces)";
run;

title "Daily average energy use data by lights";
proc sgplot data=WORK.IMPORT_NUMERIC;
   series x=date_num y=daily_avg_light / markers;
    xaxis label="Date";
   yaxis label="Energy Use (Lights)";
run;


/*  Correspondence Analysis Method */

/* Check summary statistics */
proc means data=energy_standardized;
    var Appliances_num lights_num T1_num T2_num RH_1_num RH_2_num;
run;

/* Define formats for categorizing continuous variables */
```

```
proc format;
    value energy_fmt
        low - 0 = 'Low'
        0.01 - 3 = 'Medium'
        3.01 - high = 'High';
    value temp_fmt
        low - 0 = 'Low'
        0.01 - 3 = 'Medium'
        3.01 - high = 'High';
    value humidity_fmt
        low - 0 = 'Low'
        0.01 - 3 = 'Medium'
        3.01 - high = 'High';
run;

/* Apply the formats to categorize the continuous variables */
data categorized_data;
    set energy_standardized;
    Appliances_cat = put(Appliances_num, energy_fmt.);
    Lights_cat = put(lights_num, energy_fmt.);
    T1_cat = put(T1_num, temp_fmt.);
    T2_cat = put(T2_num, temp_fmt.);
    RH_1_cat = put(RH_1_num, humidity_fmt.);
    RH_2_cat = put(RH_2_num, humidity_fmt.);
run;

/* Create a contingency table for Correspondence Analysis */
proc freq data=categorized_data;
    tables (Appliances_cat Lights_cat) * (T1_cat T2_cat RH_1_cat
RH_2_cat) / out=contingency_table;
run;

/* Print the contingency table to verify */
proc print data=contingency_table (obs=20);
run;

/* Perform Correspondence Analysis */
proc corresp data=contingency_table outc=coord;
    tables Lights_cat, RH_2_cat;
    weight COUNT;
run;

/* Plot the results */
proc sgplot data=coord;
    scatter x=dim1 y=dim2 / group=_type_
markerattrs=(symbol=circlefilled);
```

```
    text x=dim1 y=dim2 text=_name_ / position=right;
    xaxis label="Dimension 1";
    yaxis label="Dimension 2";
    title "Correspondence Analysis Plot";
run;

/*Canonical Correlation Analysis with PROC CANCORR*/

/* Step 1: Define the variable sets */
%let environmental_vars = T1_num RH_1_num T2_num RH_2_num T3_num
RH_3_num T4_num RH_4_num
                          T5_num RH_5_num T6_num RH_6_num T7_num
RH_7_num T8_num RH_8_num
                          T9_num RH_9_num T_out_num Press_mm_hg_num
RH_out_num Windspeed_num
                          Visibility_num Tdewpoint_num rv1_num
rv2_num;

%let energy_vars = Appliances_num lights_num;
run;

/* Step 2: Perform Canonical Correlation Analysis */
proc cancorr data=WORK.IMPORT_NUMERIC
             vprefix=Env vname="Environmental Factors"
             wprefix=Energy wname="Energy Usage";
    var &environmental_vars;
    with &energy_vars;
run;


/*Canonical Discriminant Analysis*/

/* Check the structure of the merged_data */
proc contents data=WORK.merged_data;
run;

/* Ensure merged_data has the time_segment variable and is ready for
discriminant analysis */
data discriminant_data;
    set WORK.merged_data;
    /* Ensure necessary variables are included */
    keep date Appliances_num lights_num T1_num RH_1_num T2_num
RH_2_num T3_num RH_3_num T4_num RH_4_num
         T5_num RH_5_num T6_num RH_6_num T7_num RH_7_num T8_num
RH_8_num T9_num RH_9_num T_out_num
         Press_mm_hg_num RH_out_num Windspeed_num Visibility_num
```

```
Tdewpoint_num rv1_num rv2_num time_segment;
run;

/* Check the structure and contents of the prepared data */
proc print data=discriminant_data(obs=10);
run;
/* Perform discriminant analysis */
proc discrim data=discriminant_data out=discrim_out canonical;
    class time_segment;
    var Appliances_num lights_num T1_num RH_1_num T2_num RH_2_num
T3_num RH_3_num T4_num RH_4_num
        T5_num RH_5_num T6_num RH_6_num T7_num RH_7_num T8_num
RH_8_num T9_num RH_9_num T_out_num
        Press_mm_hg_num RH_out_num Windspeed_num Visibility_num
Tdewpoint_num rv1_num rv2_num;
run;

/* Step 3: Prepare data for visualization */
/* Sort the discrim_out dataset by time_segment */
proc sort data=discrim_out;
    by time_segment;
run;

/* Merge the sorted dataset for visualization */
data plotclass;
    set discrim_out;
run;

/* Step 4: Define a template for plotting the discriminant analysis
results */
proc template;
    define statgraph classify;
        begingraph;
            layout overlay;
                contourplotparm x=Can1 y=Can2 z=_into_ /
contourtype=fill nhint=30 gridded=false;
                scatterplot x=Can1 y=Can2 / group=time_segment
includemissinggroup=false markercharactergroup=time_segment;
            endlayout;
        endgraph;
    end;
run;

/* Step 5: Render the plot */
proc sgrender data=plotclass template=classify;
run;
```

```
/*Clustering for using daily averages */

/* Step 1: Standardize the dataset for clustering */
proc standard data=daily_averages mean=0 std=1
out=clustering_standardized;
    var daily_avg_Appliance daily_avg_light
    daily_avg_T1 daily_avg_RH_1 daily_avg_T2 daily_avg_RH_2
daily_avg_T3
    daily_avg_RH_3 daily_avg_T4 daily_avg_RH_4 daily_avg_T5
daily_avg_RH_5
    daily_avg_T6 daily_avg_RH_6 daily_avg_T7 daily_avg_RH_7
daily_avg_T8
    daily_avg_RH_8 daily_avg_T9 daily_avg_RH_9 daily_avg_T_out
daily_avg_Press_mm_hg
    daily_avg_RH_out daily_avg_Windspeed daily_avg_Visibility
daily_avg_Tdewpoint
    daily_avg_rv1 daily_avg_rv2;
run;

/* Step 2: Perform Clustering using K-means (PROC FASTCLUS) */
proc fastclus data=clustering_standardized maxclusters=3
out=clus_output;
    var daily_avg_Appliance daily_avg_light
    daily_avg_T1 daily_avg_RH_1 daily_avg_T2 daily_avg_RH_2
daily_avg_T3
    daily_avg_RH_3 daily_avg_T4 daily_avg_RH_4 daily_avg_T5
daily_avg_RH_5
    daily_avg_T6 daily_avg_RH_6 daily_avg_T7 daily_avg_RH_7
daily_avg_T8
    daily_avg_RH_8 daily_avg_T9 daily_avg_RH_9 daily_avg_T_out
daily_avg_Press_mm_hg
    daily_avg_RH_out daily_avg_Windspeed daily_avg_Visibility
daily_avg_Tdewpoint
    daily_avg_rv1 daily_avg_rv2;
run;

/* Step 3: Evaluate Clustering Results */
proc print data=clus_output(obs=10);
    var cluster daily_avg_Appliance daily_avg_light
    daily_avg_T1 daily_avg_RH_1 daily_avg_T2 daily_avg_RH_2
daily_avg_T3
    daily_avg_RH_3 daily_avg_T4 daily_avg_RH_4 daily_avg_T5
daily_avg_RH_5
    daily_avg_T6 daily_avg_RH_6 daily_avg_T7 daily_avg_RH_7
daily_avg_T8
```

```
    daily_avg_RH_8 daily_avg_T9 daily_avg_RH_9 daily_avg_T_out
daily_avg_Press_mm_hg
    daily_avg_RH_out daily_avg_Windspeed daily_avg_Visibility
daily_avg_Tdewpoint
    daily_avg_rv1 daily_avg_rv2;
run;

/* Step 3: Summarize Cluster Characteristics */
proc means data=clus_output n mean std min max;
    class cluster;
    var daily_avg_Appliance daily_avg_light
    daily_avg_T1 daily_avg_RH_1 daily_avg_T2 daily_avg_RH_2
daily_avg_T3
    daily_avg_RH_3 daily_avg_T4 daily_avg_RH_4 daily_avg_T5
daily_avg_RH_5
    daily_avg_T6 daily_avg_RH_6 daily_avg_T7 daily_avg_RH_7
daily_avg_T8
    daily_avg_RH_8 daily_avg_T9 daily_avg_RH_9 daily_avg_T_out
daily_avg_Press_mm_hg
    daily_avg_RH_out daily_avg_Windspeed daily_avg_Visibility
daily_avg_Tdewpoint
    daily_avg_rv1 daily_avg_rv2;
run;


/* Step 4: Visualize the Clusters */
proc sgplot data=clus_output;
    scatter x=daily_avg_Appliance y=daily_avg_light / group=cluster
markerattrs=(symbol=circlefilled) transparency=0.5;
    title 'Clustering Results: daily_avg_Appliance VS daily_avg_light
';
run;


/* Step 5: Hierarchical Clustering (PROC CLUSTER) */
proc cluster data=clustering_standardized method=ward
outtree=clus_tree;
    var daily_avg_Appliance daily_avg_light
    daily_avg_T1 daily_avg_RH_1 daily_avg_T2 daily_avg_RH_2
daily_avg_T3
    daily_avg_RH_3 daily_avg_T4 daily_avg_RH_4 daily_avg_T5
daily_avg_RH_5
    daily_avg_T6 daily_avg_RH_6 daily_avg_T7 daily_avg_RH_7
daily_avg_T8
    daily_avg_RH_8 daily_avg_T9 daily_avg_RH_9 daily_avg_T_out
daily_avg_Press_mm_hg
```

```
    daily_avg_RH_out daily_avg_Windspeed daily_avg_Visibility
daily_avg_Tdewpoint
    daily_avg_rv1 daily_avg_rv2;
run;

/* Step 6: Create Clusters from the Hierarchical Tree using PROC TREE
*/
proc tree data=clus_tree out=tree_clusters nclusters=3;
    id _NAME_; /* Use _NAME_ to identify observations */
run;

/* Step 7: Print the Clusters Created by PROC TREE */
proc print data=tree_clusters;
run;


/*Clustering*/

/* Step 1: Standardize the dataset for clustering */
proc standard data=WORK.IMPORT_NUMERIC mean=0 std=1
out=clustering_standardized;
    var Appliances_num lights_num T1_num RH_1_num T2_num RH_2_num
T3_num RH_3_num
        T4_num RH_4_num T5_num RH_5_num T6_num RH_6_num T7_num
RH_7_num T8_num RH_8_num
        T9_num RH_9_num T_out_num Press_mm_hg_num RH_out_num
Windspeed_num Visibility_num
        Tdewpoint_num rv1_num rv2_num;
run;

/* Step 2: Perform Clustering using K-means (PROC FASTCLUS) */
proc fastclus data=clustering_standardized maxclusters=3
out=clus_output;
    var Appliances_num lights_num T1_num RH_1_num T2_num RH_2_num
T3_num RH_3_num
        T4_num RH_4_num T5_num RH_5_num T6_num RH_6_num T7_num
RH_7_num T8_num RH_8_num
        T9_num RH_9_num T_out_num Press_mm_hg_num RH_out_num
Windspeed_num Visibility_num
        Tdewpoint_num rv1_num rv2_num;
run;

/* Step 3: Evaluate Clustering Results */
proc print data=clus_output(obs=10);
    var cluster Appliances_num lights_num T1_num RH_1_num T2_num
RH_2_num T3_num RH_3_num
```

```
        T4_num RH_4_num T5_num RH_5_num T6_num RH_6_num T7_num
RH_7_num T8_num RH_8_num
        T9_num RH_9_num T_out_num Press_mm_hg_num RH_out_num
Windspeed_num Visibility_num
        Tdewpoint_num rv1_num rv2_num;
run;


/* Step 3: Summarize Cluster Characteristics */
proc means data=clus_output n mean std min max;
    class cluster;
    var Appliances_num lights_num T1_num RH_1_num T2_num RH_2_num
T3_num RH_3_num
        T4_num RH_4_num T5_num RH_5_num T6_num RH_6_num T7_num
RH_7_num T8_num RH_8_num
        T9_num RH_9_num T_out_num Press_mm_hg_num RH_out_num
Windspeed_num Visibility_num
        Tdewpoint_num rv1_num rv2_num;
run;


/* Step 4: Visualize the Clusters */
proc sgplot data=clus_output;
    scatter x=T1_num y=T2_num / group=cluster
markerattrs=(symbol=circlefilled) transparency=0.5;
    title 'Clustering Results: T1_num vs T2_num';
run;

/* Additional scatter plots for other variable pairs */
proc sgplot data=clus_output;
    scatter x=Appliances_num y=lights_num / group=cluster
markerattrs=(symbol=circlefilled) transparency=0.5;
    title 'Clustering Results: Appliances_num vs lights_num';
run;


/* Step 5: Hierarchical Clustering (PROC CLUSTER) */
proc cluster data=clustering_standardized method=ward
outtree=clus_tree;
    var Appliances_num lights_num T1_num RH_1_num T2_num RH_2_num
T3_num RH_3_num
        T4_num RH_4_num T5_num RH_5_num T6_num RH_6_num T7_num
RH_7_num T8_num RH_8_num
        T9_num RH_9_num T_out_num Press_mm_hg_num RH_out_num
Windspeed_num Visibility_num
        Tdewpoint_num rv1_num rv2_num;
```

```
run;

/* Step 6: Create Clusters from the Hierarchical Tree using PROC TREE
*/
proc tree data=clus_tree out=tree_clusters nclusters=3;
    id _NAME_; /* Use _NAME_ to identify observations */
run;

/* Step 7: Print the Clusters Created by PROC TREE */
proc print data=tree_clusters;
run;


/*PLS TRY*/
/* Step 1: Standardize the dataset */
proc standard data= daily_averages mean=0 std=1
out=daily_averages_standardized;
    var daily_avg_Appliance daily_avg_light
    daily_avg_T1 daily_avg_RH_1 daily_avg_T2 daily_avg_RH_2
daily_avg_T3
    daily_avg_RH_3 daily_avg_T4 daily_avg_RH_4 daily_avg_T5
daily_avg_RH_5
    daily_avg_T6 daily_avg_RH_6 daily_avg_T7 daily_avg_RH_7
daily_avg_T8
    daily_avg_RH_8 daily_avg_T9 daily_avg_RH_9 daily_avg_T_out
daily_avg_Press_mm_hg
    daily_avg_RH_out daily_avg_Windspeed daily_avg_Visibility
daily_avg_Tdewpoint
    daily_min_usage_T1 daily_max_usage_T1 daily_min_usage_T2
daily_max_usage_T2
    daily_min_usage_T3 daily_max_usage_T3 daily_min_usage_T4
daily_max_usage_T4
    daily_min_usage_T5 daily_max_usage_T5 daily_min_usage_T6
daily_max_usage_T6
    daily_min_usage_T7 daily_max_usage_T7 daily_min_usage_T8
daily_max_usage_T8
    daily_min_usage_T9 daily_max_usage_T9 ;
run;

proc pls data=daily_averages_standardized;
       model daily_avg_Appliance = daily_avg_light
    daily_avg_T1 daily_avg_RH_1 daily_avg_T2 daily_avg_RH_2
daily_avg_T3
    daily_avg_RH_3 daily_avg_T4 daily_avg_RH_4 daily_avg_T5
daily_avg_RH_5
    daily_avg_T6 daily_avg_RH_6 daily_avg_T7 daily_avg_RH_7
```

```
daily_avg_T8
    daily_avg_RH_8 daily_avg_T9 daily_avg_RH_9 daily_avg_T_out
    daily_avg_Press_mm_hg daily_avg_RH_out daily_avg_Windspeed
    daily_avg_Visibility daily_avg_Tdewpoint daily_min_usage_T1
daily_max_usage_T1
    daily_min_usage_T2 daily_max_usage_T2 daily_min_usage_T3
daily_max_usage_T3 daily_min_usage_T4 daily_max_usage_T4
    daily_min_usage_T5 daily_max_usage_T5 daily_min_usage_T6
daily_max_usage_T6
    daily_min_usage_T7 daily_max_usage_T7 daily_min_usage_T8
daily_max_usage_T8
    daily_min_usage_T9 daily_max_usage_T9;
run;


/* Step 2: Perform PLS Regression */
proc pls data=daily_averages_standardized nfac=10 cv=split(5)
method=pls;
    model daily_avg_Appliance = /*daily_avg_light*/ daily_avg_T1
daily_avg_RH_1 daily_avg_T2 daily_avg_RH_2
                                 daily_avg_T3 daily_avg_RH_3
/*daily_avg_T4*/ daily_avg_RH_4 daily_avg_T5
                                 daily_avg_RH_5 daily_avg_T6
daily_avg_RH_6 daily_avg_T7 /*daily_avg_RH_7*/
                                 daily_avg_T8 daily_avg_RH_8
daily_avg_T9 daily_avg_RH_9 /*daily_avg_T_out*/
                                 /*daily_avg_Tdewpoint*/
daily_min_usage_T1 daily_max_usage_T1
    daily_min_usage_T2 daily_max_usage_T2 daily_min_usage_T3
daily_max_usage_T3 daily_min_usage_T4 daily_max_usage_T4
    daily_min_usage_T5 daily_max_usage_T5 daily_min_usage_T6
daily_max_usage_T6
    daily_min_usage_T7 daily_max_usage_T7 daily_min_usage_T8
daily_max_usage_T8
    daily_min_usage_T9 daily_max_usage_T9;
    output out=pls_pred p=y_pred;
run;

/* Step 3: Generate PLS Scores */
proc score data=daily_averages_standardized score=pls_pred
out=pls_scores(rename=(y_pred=_SCORE_));
run;

/* Step 3: Generate PLS Scores */
proc score data=daily_averages_standardized score=pls_pred type=parms
out=pls_scores(rename=(y_pred=_SCORE_));
```

```
run;



/* Step 3: Generate PLS Scores */
proc score data=daily_averages_standardized score=pls_pred type=parms
out=pls_scores;
    var daily_avg_light daily_avg_T1 daily_avg_RH_1 daily_avg_T2
daily_avg_RH_2
        daily_avg_T3 daily_avg_RH_3 daily_avg_T4 daily_avg_RH_4
daily_avg_T5
        daily_avg_RH_5 daily_avg_T6 daily_avg_RH_6 daily_avg_T7
daily_avg_RH_7
        daily_avg_T8 daily_avg_RH_8 daily_avg_T9 daily_avg_RH_9
daily_avg_T_out
        daily_avg_Press_mm_hg daily_avg_RH_out daily_avg_Windspeed
daily_avg_Visibility
        daily_avg_Tdewpoint daily_avg_rv1 daily_avg_rv2;
run;


/* Step 3: Generate PLS Scores */
proc score data=daily_averages_standardized score=pls_pred
out=pls_pred;
run;

/* Step 3: Assess Variable Importance */
proc sgplot data=pls_scores;
    vbar _NAME_ / response=_VIP_ datalabel;
    xaxis label="Predictor Variables";
    yaxis label="Variable Importance in Projection (VIP)";
    title "PLS Regression: VIP Scores";
run;

/* Step 4: Identify and Filter Non-Predictive Variables */
/* Example: Print VIP Scores to Identify Non-Predictive Variables */
proc print data=pls_scores(where=(_VIP_ < 0.8));
    var _NAME_ _VIP_;
    title "Variables with VIP Scores Less Than 0.8";
run;

/* Step 5: Assess the Model */
proc print data=pls_out(obs=10);
run;

proc sgplot data=pls_pred;
```

```
    scatter x=Appliances_num y=y_pred;
    lineparm x=0 y=0 slope=1 / lineattrs=(color=red);
    xaxis label="Actual Appliance Energy Use";
    yaxis label="Predicted Appliance Energy Use";
    title "PLS Regression: Actual vs Predicted Appliance Energy Use";
run;

proc sgplot data=pls_out;
    series x=_CV_ y=_PRESS_ / markers;
    xaxis label="Number of Components";
    yaxis label="Predictive Residual Sum of Squares (PRESS)";
    title "PLS Regression: Model Selection using PRESS";
run;

/*PLS Regression*/

/* Step 1: Standardize the dataset */
proc standard data= daily_averages mean=0 std=1
out=daily_averages_standardized;
    var daily_avg_Appliance daily_avg_light
    daily_avg_T1 daily_avg_RH_1 daily_avg_T2 daily_avg_RH_2
daily_avg_T3
    daily_avg_RH_3 daily_avg_T4 daily_avg_RH_4 daily_avg_T5
daily_avg_RH_5
    daily_avg_T6 daily_avg_RH_6 daily_avg_T7 daily_avg_RH_7
daily_avg_T8
    daily_avg_RH_8 daily_avg_T9 daily_avg_RH_9 daily_avg_T_out
daily_avg_Press_mm_hg
    daily_avg_RH_out daily_avg_Windspeed daily_avg_Visibility
daily_avg_Tdewpoint
    daily_min_usage_T1 daily_max_usage_T1 daily_min_usage_T2
daily_max_usage_T2
    daily_min_usage_T3 daily_max_usage_T3 daily_min_usage_T4
daily_max_usage_T4
    daily_min_usage_T5 daily_max_usage_T5 daily_min_usage_T6
daily_max_usage_T6
    daily_min_usage_T7 daily_max_usage_T7 daily_min_usage_T8
daily_max_usage_T8
    daily_min_usage_T9 daily_max_usage_T9 ;
run;

/* Partial Least Squares (PLS) Regression */
proc pls data=daily_averages_standardized method=pls nfac=5;
    model daily_avg_Appliance = daily_avg_light
        daily_avg_T1 daily_avg_RH_1 daily_avg_T2 daily_avg_RH_2
daily_avg_T3
```

```
        daily_avg_RH_3 daily_avg_T4 daily_avg_RH_4 daily_avg_T5
daily_avg_RH_5
        daily_avg_T6 daily_avg_RH_6 daily_avg_T7 daily_avg_RH_7
daily_avg_T8
        daily_avg_RH_8 daily_avg_T9 daily_avg_RH_9 daily_avg_T_out
daily_avg_Press_mm_hg
        daily_avg_RH_out daily_avg_Windspeed daily_avg_Visibility
daily_avg_Tdewpoint
        daily_min_usage_T1 daily_max_usage_T1 daily_min_usage_T2
daily_max_usage_T2
        daily_min_usage_T3 daily_max_usage_T3 daily_min_usage_T4
daily_max_usage_T4
        daily_min_usage_T5 daily_max_usage_T5 daily_min_usage_T6
daily_max_usage_T6
        daily_min_usage_T7 daily_max_usage_T7 daily_min_usage_T8
daily_max_usage_T8
        daily_min_usage_T9 daily_max_usage_T9 daily_avg_rv1
daily_avg_rv2;
    output out=pls_output predicted=Predicted_Appliances_num;
run;

/* Step 2: Use PROC REG to obtain coefficients */
proc reg data=pls_output;
    model daily_avg_Appliance = daily_avg_light
        daily_avg_T1 daily_avg_RH_1 daily_avg_T2 daily_avg_RH_2
daily_avg_T3
        daily_avg_RH_3 daily_avg_T4 daily_avg_RH_4 daily_avg_T5
daily_avg_RH_5
        daily_avg_T6 daily_avg_RH_6 daily_avg_T7 daily_avg_RH_7
daily_avg_T8
        daily_avg_RH_8 daily_avg_T9 daily_avg_RH_9 daily_avg_T_out
daily_avg_Press_mm_hg
        daily_avg_RH_out daily_avg_Windspeed daily_avg_Visibility
daily_avg_Tdewpoint
        daily_min_usage_T1 daily_max_usage_T1 daily_min_usage_T2
daily_max_usage_T2
        daily_min_usage_T3 daily_max_usage_T3 daily_min_usage_T4
daily_max_usage_T4
        daily_min_usage_T5 daily_max_usage_T5 daily_min_usage_T6
daily_max_usage_T6
        daily_min_usage_T7 daily_max_usage_T7 daily_min_usage_T8
daily_max_usage_T8
        daily_min_usage_T9 daily_max_usage_T9 daily_avg_rv1
daily_avg_rv2;
    output out=reg_output p=predicted;
run;
```

```
/* Print the first 20 observations to verify the PLS output */
proc print data=pls_output(obs=20);
    var daily_avg_Appliance Predicted_Appliances_num;
run;

/* Scatter plot of Actual vs. Predicted Values */
proc sgplot data=pls_output;
    scatter x=daily_avg_Appliance y=Predicted_Appliances_num;
    lineparm x=0 y=0 slope=1 / lineattrs=(color=red);
    xaxis label="Actual Appliances Energy Consumption";
    yaxis label="Predicted Appliances Energy Consumption";
    title "Actual vs. Predicted Energy Consumption (PLS)";
run;

/* Print the actual and predicted values for all observations */
proc print data=pls_output noobs label;
    var daily_avg_Appliance Predicted_Appliances_num;
    label daily_avg_Appliance = "Actual Appliances Energy
Consumption"
          Predicted_Appliances_num = "Predicted Appliances Energy
Consumption";
    title "Table of Actual vs. Predicted Values";
run;
```