Table of Contents

[Executive Summary 3](#_Toc167172391)

[1.0 Introduction 4](#_Toc167172392)

[1.1 Background 4](#_Toc167172393)

[1.2 Purpose and Rationale 4](#_Toc167172394)

[1.3 Hypothesis 5](#_Toc167172395)

[1.4 Objectives 5](#_Toc167172396)

[2.0 Literature 6](#_Toc167172397)

[3.0 Methods 7](#_Toc167172398)

[3.1 Data Source 7](#_Toc167172399)

[3.2 Data Collection 7](#_Toc167172400)

[3.4 Plotting the Data 9](#_Toc167172401)

[3.5 Data Preprocessing 10](#_Toc167172402)

[3.6 Analytical Methods Used 11](#_Toc167172403)

[3.6.1 Principal Components Analysis (PCA) and its Visualization 11](#_Toc167172404)

[3.6.2 Eigenvalues 11](#_Toc167172405)

[3.6.3 Factor Analysis & MDS 11](#_Toc167172406)

[3.6.4 Correspondence Analysis 12](#_Toc167172407)

[3.6.5 Canonical Correlation Analysis with PROC CANCORR 12](#_Toc167172408)

[3.6.6 Canonical Discriminant Analysis 12](#_Toc167172409)

[3.6.7 Clustering 12](#_Toc167172410)

[3.6.8 PLS Regression 13](#_Toc167172411)

[4.0 Results 13](#_Toc167172412)

[4.1 Principal Components Analysis (PCA) and its Visualization 13](#_Toc167172413)

[4.1.1 Scree plot 13](#_Toc167172414)

[4.1.2 Scatter Plot of Principal Components 14](#_Toc167172415)

[4.2 Eigenvalues 15](#_Toc167172416)

[4.3 Factor Analysis & MDS 17](#_Toc167172417)

[4.4 Correspondence Analysis 21](#_Toc167172418)

[4.5 Canonical Correlation Analysis with PROC CANCORR 24](#_Toc167172419)

[4.6 Canonical Discriminant Analysis 29](#_Toc167172420)

[4.7 Clustering 30](#_Toc167172421)

[4.8 PLS Regression 31](#_Toc167172422)

[5.0 Discussion 34](#_Toc167172423)

[6.0 References 36](#_Toc167172424)

[7.0 Appendices 39](#_Toc167172425)

# Executive Summary

This study investigated energy usage in appliances and light fittings in a low-energy building. This study intended to develop models that predict energy consumption from environmental parameters and weather conditions. A multivariate time- series data from Appliances Energy Predictions (online repository) with 28 variables was used. It consisted of an aggregation of environmental measurements, conditions and recorded data of energy usage for a 4.5month 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 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. As the world increasingly focuses on sustainability optimizing energy consumption is paramount (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 patterns of energy consumption and identifying the factors that influence it. According to the U.S. Energy Information Administration, residential and commercial buildings account for nearly 40% of total energy consumption in the United States.

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 within their occupants. From passive solar design principles to the integration of 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 presents an analysis of the energy use in appliances and light fittings in a low-energy building and how the indoor and outdoor environmental variables affect energy use. A low-energy building is designed in such a manner that it uses very little energy to sustain comfort to 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—all these bear together on the use of energy. Such approaches often depend on simple heuristics that are employed, probably without the ability to capture the full complexity of the dynamics in energy consumption.

## 1.3 Hypothesis

*H1:* Outdoor weather conditions, such as temperature, wind speed, and visibility, significantly influence the energy usage of appliances and lights.

*H2:* There exists a relationship between indoor temperatures and humidity levels with energy usage.

*H3*: Past energy usage, along with environmental conditions possess predictive power for forecasting future energy use.

## 1.4 Objectives

This study purpose 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 predictive models for future energy use based on past energy usage and environmental conditions.

# 2.0 Literature

Previous literatures have explored study of energy consumption in buildings and the factors responsible for driving it. Some studies have examined patterns of energy consumption in buildings with a focus on predicting appliance energy use in low-energy houses. Studies by Candanedo et al, (2017), Guo et al., (2015), Arghira et al, (2012), Cetin et al, (2014) and Kavousian et al., (2013) 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 behaviors in homes or offices to rank appliance efficiency in household use. Works by Hong et al., (2016), Kavousian et al. (2015), and Cetin (2016) analyzed occupant behaviors during their stay, employing regression and probabilistic analyses to identify load patterns. Additionally, studies have attempted to predict accurate occupancy numbers by analyzing appliance use behaviors (Candanedo, 2016).

Taken together, these studies emphasized nature of appliance energy use in homes or office spaces, 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 humidity, temperature wind speed, dew points, and visibility for understanding the internal and external environment of the building and energy use data from m-bus energy meters. Research in this domain has a 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 in great detail the dynamics likely to emerge.

# 3.0 Methods

## 3.1 Data Source

The study uses a multivariate time series data, sourced from the Appliances Energy dataset, accessible from <https://archive.ics.uci.edu/dataset/374/appliances+energy+prediction> .The source has been considered reputable because of its reliability, wide use, and thorough documentation.

## 3.2 Data Collection

The dataset contained a 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 (Chievres Airport, Belgium).

Data was collected at regular 10-minute intervals over a period of 4.5 months. Data collection was facilitated using a ZigBee wireless sensor network and m-bus energy meters. Temperature and humidity conditions were monitored using a ZigBee wireless sensor network. Each wireless node transmitted data every 3.3 minutes, which was then averaged over 10-minute periods. Energy usage data for appliances and light fixtures were logged every 10 minutes using m-bus energy meters.

Weather data, including wind speed, dew points, and visibility, were downloaded from a public dataset provided by Reliable Prognosis (rp5.ru) and merged with the experimental data using the date and time columns. The dataset contained 19,735 samples, each representing a unique 10-minute interval of data collection. 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 | oC |
| T2, Temperature in living room | oC |
| T3, Temperature in laundry room | oC |
| T4, Temperature in office room | oC |
| T5, Temperature in bathroom | oC |
| T6, Temperature outside the building | oC |
| T7, Temperature in ironing room | oC |
| T8, Temperature in teenager room | oC |
| T9, Temperature in parents room | oC |
| To Temperature outside (from Chievres weather station) | oC |
| 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) | Â°C |
| rv1, Random variable 1 | nondimensional |
| rv2, Random variable 2 | nondimensional |

## 3.4 Plotting the Data

The spikes in the line graphs indicates that power demand fluctuates over time. Light energy use 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*

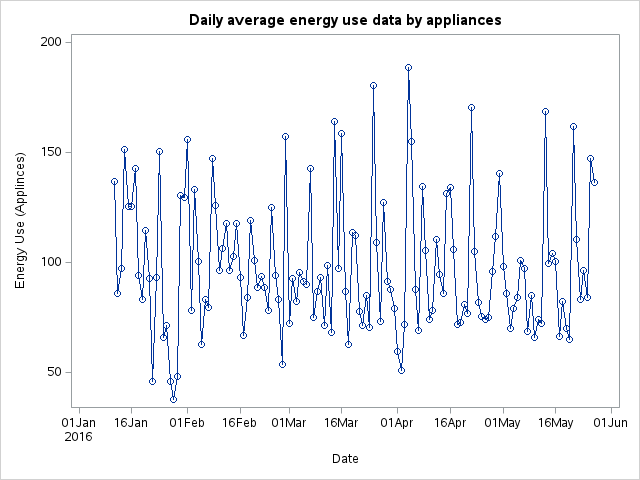
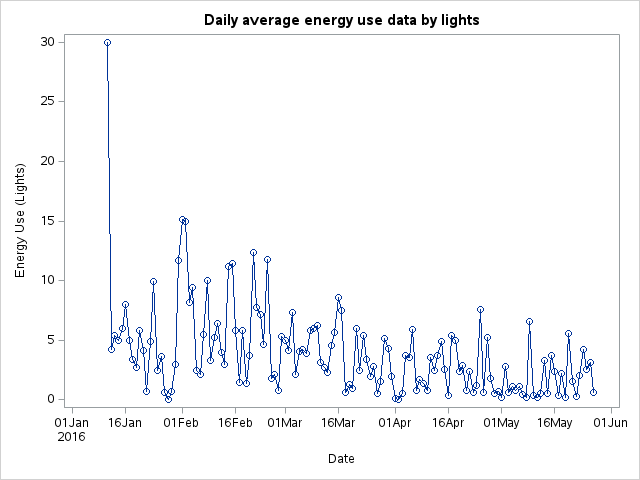


Figure 2: Daily average energy use by lights



## 3.5 Data Preprocessing

Prior to analysis, data preprocessing was conducted 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 environmental readings and weather conditions captured. Two random variables were included in the dataset for testing regression models and filtering out non-predictive attributes 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 time use, afternoon time use, evening time use, and night. All these attributes were prepared to identify different patterns of energy use at different times of the day. These analyses helped identify relationships among the predictor variables and the target variable (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, complex analyses, with an enormous set of procedures and tools for data manipulation, visualization, advanced statistical modeling, and more. SAS comprises a wide-ranging set of procedures and tools useful to manipulate data, visualize data, and perform 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 both to reduce dimensionality in the data set and to detect patterns in the data. Visualization techniques such as scatter plots and scree plot a were generated for the exploration of 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 individual principal component (Allee et al., 2022).

### 3.6.3 Factor Analysis & MDS

MDS helps in visualizing dissimilarities or similarities between samples or variables in a low-dimensional space and, thus, exploring 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 and 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 behavior, 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

Canonical correlation analysis examines the relationships between sets of variables, such as indoor conditions and energy usage, and outdoor weather conditions. It identifies the strongest linear combinations of variables from different sets, providing insights into the associations between indoor and outdoor factors influencing energy consumption. PROC CANCORR in SAS was used to conduct the analysis and visualize the canonical variates.

### 3.6.6 Canonical Discriminant Analysis

Canonical discriminant analysis was used to classify observations into predefined groups based on a set of 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

Cluster analysis was conducted to identify natural groupings or clusters within the dataset. K-means clustering, was conducted. Clustering algorithms group similar samples based on their characteristics (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

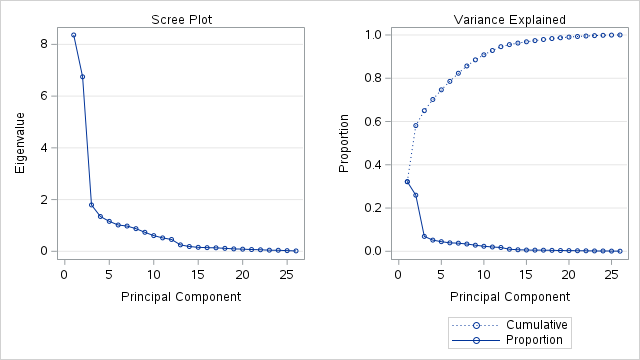
Principal Components Analysis (PCA) applied to the dataset reduced its dimensionality uncovering underlying patterns. This technique transformed the original variables into a new set of uncorrelated variables, known as principal components, which capture the maximum variance in the data. By focusing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

### 4.1.1 Scree plot

Scree plot in figure 1 visualizes the eigenvalues and determine 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*

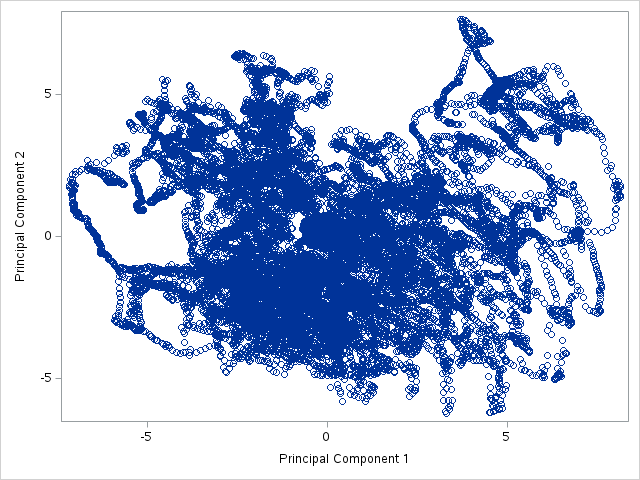


### 4.1.2 Scatter Plot of Principal Components

Scatter plot of the first two principal components visualizes the data in the 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%. Together, 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*



## 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 Appendix A). These two factors together explained 53.97% of the total variance. The scree plot in figure 3 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), while Factor 2 was 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 in Appendix A. This indicated that temperature and humidity were the primary dimensions underlying the data structure.

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 in Appendix A. This rotation clarified the distinction between the two factors.

The communalities, which represent 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 communalities of 0.868 and 0.834, respectively. Conversely, variables like Appliances\_num and Visibility\_num had low communalities, suggesting they were less effectively captured by the two factors.

The orthogonal transformation matrix revealed strong linear 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**



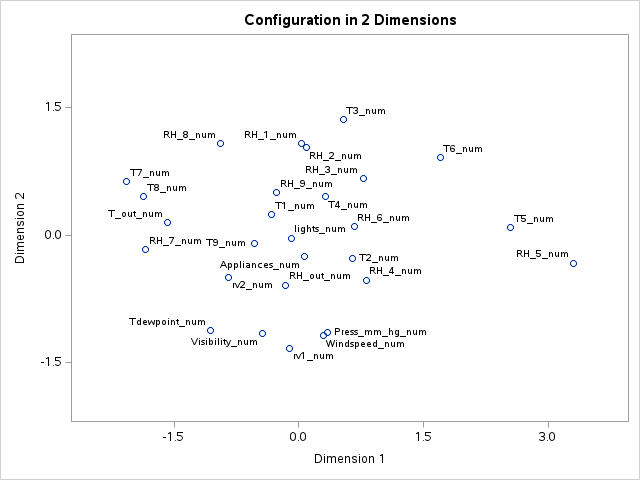


|  |
| --- |
| Convergence criteria are satisfied. |

The iterative process of the MDS analysis involved several iterations to minimize the badness-of-fit criterion, indicating the degree of dissimilarity between observed and predicted distances. Convergence criteria were ultimately satisfied, indicating the stability of the solution. The final badness-of-fit criterion of 0.1408 highlighted this convergence.

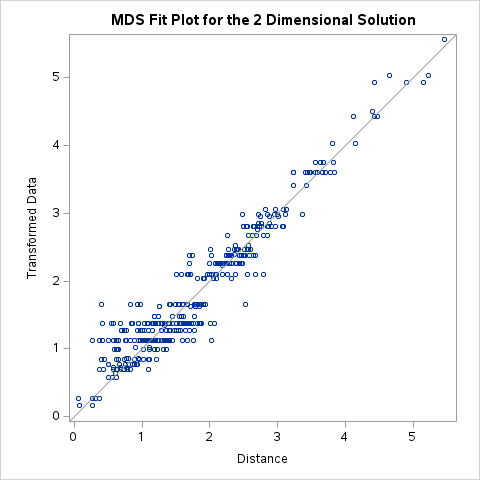
The MDS plot revealed how observations are spatially arranged based on their dissimilarities. In this plot, each point represents an observation, and the proximity of points signifies their similarity. Observations that are close together are more similar, while those that are far apart are more dissimilar.

**Figure 5: MDS Plot**

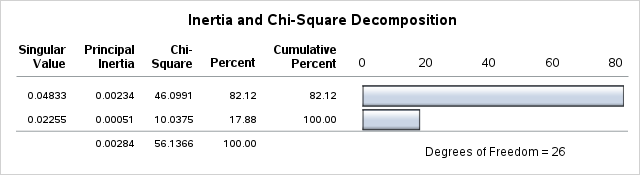


The MDS fit plot (see figure 6) for the 2-dimensional solution demonstrates a high degree of fit, as evidenced by the alignment of transformed data points with the distance metric. This indicates that the two-dimensional representation effectively captures the structure of the original high-dimensional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

**Figure 5 Continued**



## 4.4 Correspondence Analysis



Correspondence analysis explored the relationships between categorical variables (Lights\_cat) and humidity in the living room (RH\_2\_cat). It identified the most significant associations between these sets of variables, unveiling how these factors interact to 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. Additionally, the summary statistics, partial contributions to inertia, and squared cosines offer quantitative insights into the importance and directionality of these associations.

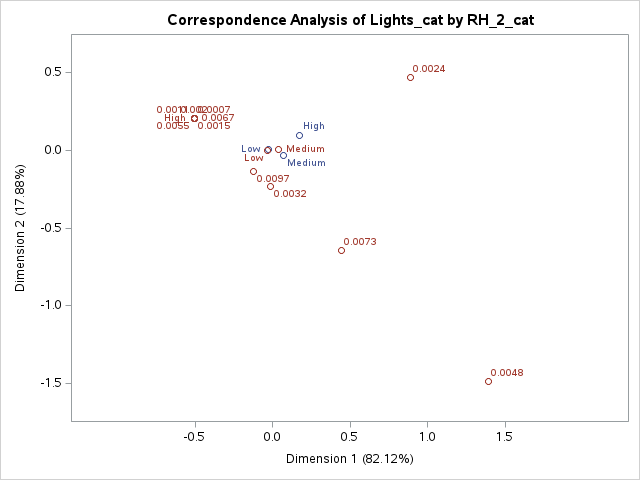
Exploring relationships between indoor conditions (represented by Lights\_cat) and outdoor weather conditions (indicated by RH\_2\_cat), the analysis provides insights into how these variables interact to affect energy usage. Specifically, it elaborates how weather conditions, such as temperature and humidity (RH\_2\_cat), are related 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 sets of variables.

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 canonical correlation above). These values indicate moderate relationships between the sets of variables, with the first function being stronger than the second. The multivariate tests (Wilks' Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Greatest Root) yieldede significant results (p < 0.0001), indicating that the canonical correlations are statistically significant.

The canonical structure tables showed the correlations between the original variables and their respective canonical variables, highlighting the relationship strength. The correlation between humidity in the kitchen (RH\_1) and the first canonical variable (Env1) is 0.253, indicated a moderate association.

**The CANCORR Procedure**

**Canonical Structure**

| **Correlations Between the Environmental Factors and Their Canonical Variables** | | |
| --- | --- | --- |
|  | **Env1** | **Env2** |
| **T1\_num** | -0.0079 | 0.1923 |
| **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 |
| **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 |

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 appliances usage. These findings suggest that specific environmental conditions, especially humidity and temperature measures, are crucial in predicting energy usage patterns.





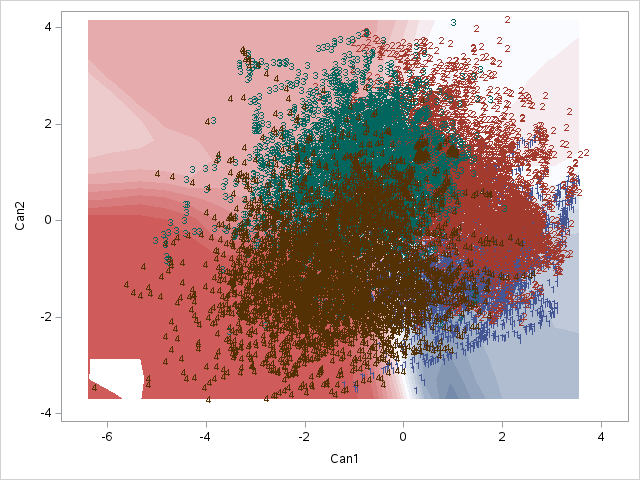
| **Raw Canonical Coefficients for the Environmental Factors** | | |
| --- | --- | --- |
|  | **Env1** | **Env2** |
| **T1\_num** | 0.0956054206 | -0.137900759 |
| **RH\_1\_num** | 0.2066215687 | 0.3652293397 |
| **T2\_num** | -0.416956337 | -0.277915658 |
| **RH\_2\_num** | -0.199260468 | -0.308716092 |
| **T3\_num** | 0.1580900442 | 0.7109108285 |
| **RH\_3\_num** | 0.0461940068 | 0.1259574776 |
| **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 the canonical scores for the first two canonical variables (Can1 and Can2). Each point represents an observation, and the numbers indicate the group to which each observation belongs.

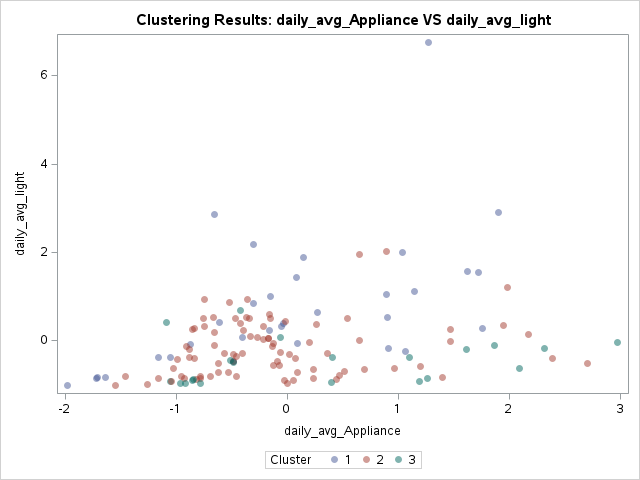
There is noticeable overlap between the groups, especially in the central region of the plot, indicating that while the canonical variables provide some separation, there is still a significant degree of similarity between the groups. Despite some overlap, the analysis highlights significant differences 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 levels of usage, 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% of the variance in energy usage, indicating a robust fit and comprehensive explanatory power. The Correlation Loading Plot (see figure 10) emphasize the importance of variables such as indoor temperature and humidity, which show strong correlations with energy usage.



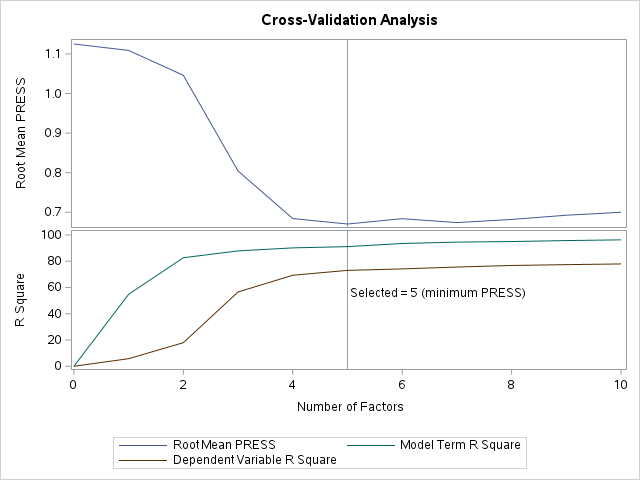


The cross-validation analysis confirms the model's reliability, with a minimum Root Mean PRESS of 0.6702 achieved using five factors. Additionally, the inclusion of two random variables demonstrates the model's ability to filter out non-predictive attributes, further validating its predictive accuracy.



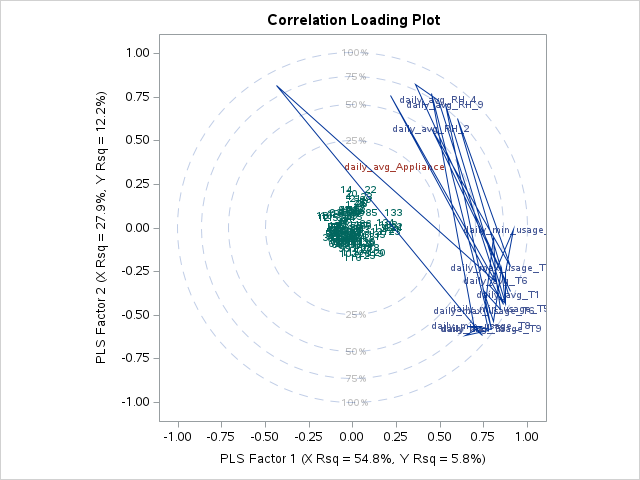
**Figure 9**

*Cross Validation*



**Figure 10**

*Correlation loading plot*



# 5.0 Discussion

Findings from the study would therefore provide useful 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 key relationships and trends with respect to energy use. It supported the hypotheses that outdoor weather conditions, indoor temperature, humidity levels, and past energy use play a most critical role 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 limitations are the possible confounding variables—occupants' behavior and building design features that have not been taken into account when performing the analyses. The general identification of key 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 the carbon footprints.

**Future Directions**

Future studies should also consider additional research variables such as occupant behavior and design features in buildings to further delineate predictive models of energy use. Other maybe longitudinal studies, which observe the 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 that can optimize energy performance in low-energy buildings could be defined from research on how to implement real-time monitoring with advanced control strategies. Overall, this study lay emphasis on 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 multivariety 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 modeling 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 Appendices

**Appendix A**

*Factor Analysis*

**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 |
| **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 |

**2 factors will be retained by the NFACTOR criterion.**

Scree Plot of Eigenvalues

|

|

|

9 +

|

|

| 1

|

8 +

|

|

|

|

7 +

| 2

|

|

|

6 +

|

E |

i |

g |

e 5 +

n |

v |

a |

l |

u 4 +

e |

s |

|

|

3 +

|

|

|

|

2 + 3

| 4

|

| 5

| 6

1 + 7 8

| 9 0

| 1 2

| 3

| 4 5 6 7 8 9

0 + 0 1 2 3 4 5 6 7 8

|

|

--------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------+-------

0 2 4 6 8 10 12 14 16 18 20 22 24 26 28

Number

| **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 |
| **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** |
| 0.00639005 | 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 | 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**

| **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 |
| **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 |
| **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** | **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** |
| 0.00639005 | 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 | 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 |
| **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*

|  |
| --- |
| /\* Generated Code (IMPORT) \*/ /\* Source File: energydata\_complete.csv \*/ /\* Source Path: /home/u58374793/SAS PROJECT \*/ /\* Code generated on: 5/16/24, 7:22 AM \*/  %web\_drop\_table(WORK.IMPORT);   FILENAME REFFILE '/home/u58374793/SAS PROJECT/energydata\_complete.csv';  PROC IMPORT DATAFILE=REFFILE  DBMS=CSV  OUT=WORK.IMPORT;  GETNAMES=YES; RUN;  PROC CONTENTS DATA=WORK.IMPORT; RUN;   %web\_open\_table(WORK.IMPORT);  /\* 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;  /\* 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; |