Multivariate Analysis of Energy Consumption in Steel Industry

Dhruv Kushwaha - 2021MT10235
Vatsal Varshney - 2021MT10700
IIT Delhi

November 5, 2024

Abstract

This project involves a comprehensive multivariate analysis of energy consumption data from the steel industry, specifically from DAEWOO Steel Co. Ltd., Gwangyang, South Korea. We performed several statistical tests, including MANOVA, Profile Analysis, and PCA, to study the seasonal variations and other patterns in the dataset.

1 Introduction

The dataset contains time-series energy consumption data recorded at 15-minute intervals throughout 2018, totaling 35,040 rows with 7 columns:

- date: date and time of the observations
- Usage_kWh: energy consumption in the industry
- CO2
- Lagging_Current_Reactive_Power_kVarh
- Leading_Current_Reactive_Power_kVarh
- Lagging_Current_Power_Factor
- Leading_Current_Power_Factor

2 Preprocessing

The following steps were undertaken to preprocess the data:

- Checked for NaN values (none were found).
- Indexed the data by date
- Resampled to daily averages, resulting in 365 observations with 6 features.
- Normalized the data to prepare for multivariate analysis.

3 Exploratory Data Analysis (EDA)

Summary statistics for key features before normalization are listed in Table 1

Feature	Mean	Std Dev
Usage_kWh	27.38	33.44
CO_2	0.01152	0.01615
Lagging_Current_Reactive_Power_kVarh	13.03	16.31
Leading_Current_Reactive_Power_kVarh	3.87	7.42
Lagging_Current_Power_Factor	80.57	18.92
Leading_Current_Power_Factor	84.36	30.45

Table 1: Summary Statistics

The daily average usage was plotted with time, depicted in Fig 1

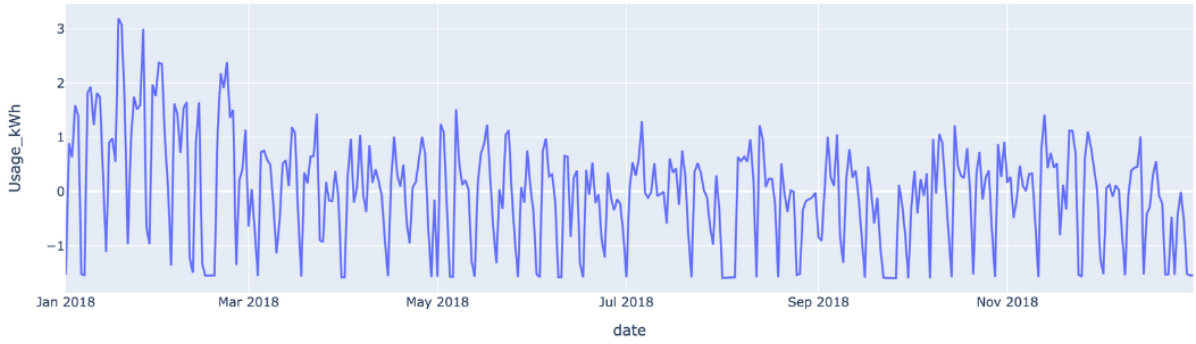

Figure 1: Daily average value of Usage_kWh with time

Correlation matrix among the features is depicted in Fig 2

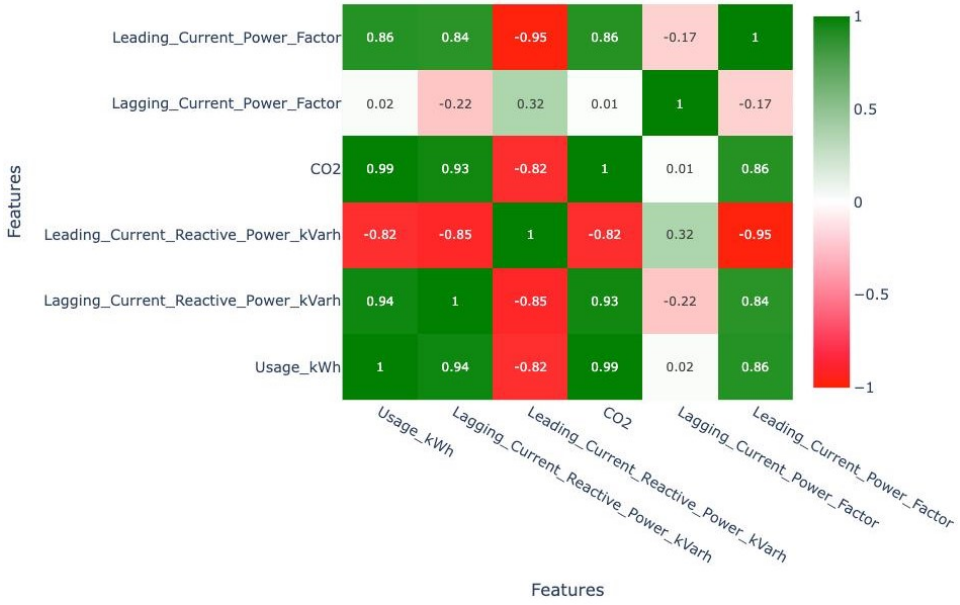

Figure 2: Correlation Matrix among features

4 Normality Test

Normality was assessed to determine the suitability of parametric tests. The univariate normality for each feature after daily resampling has been plotted in Fig 3.

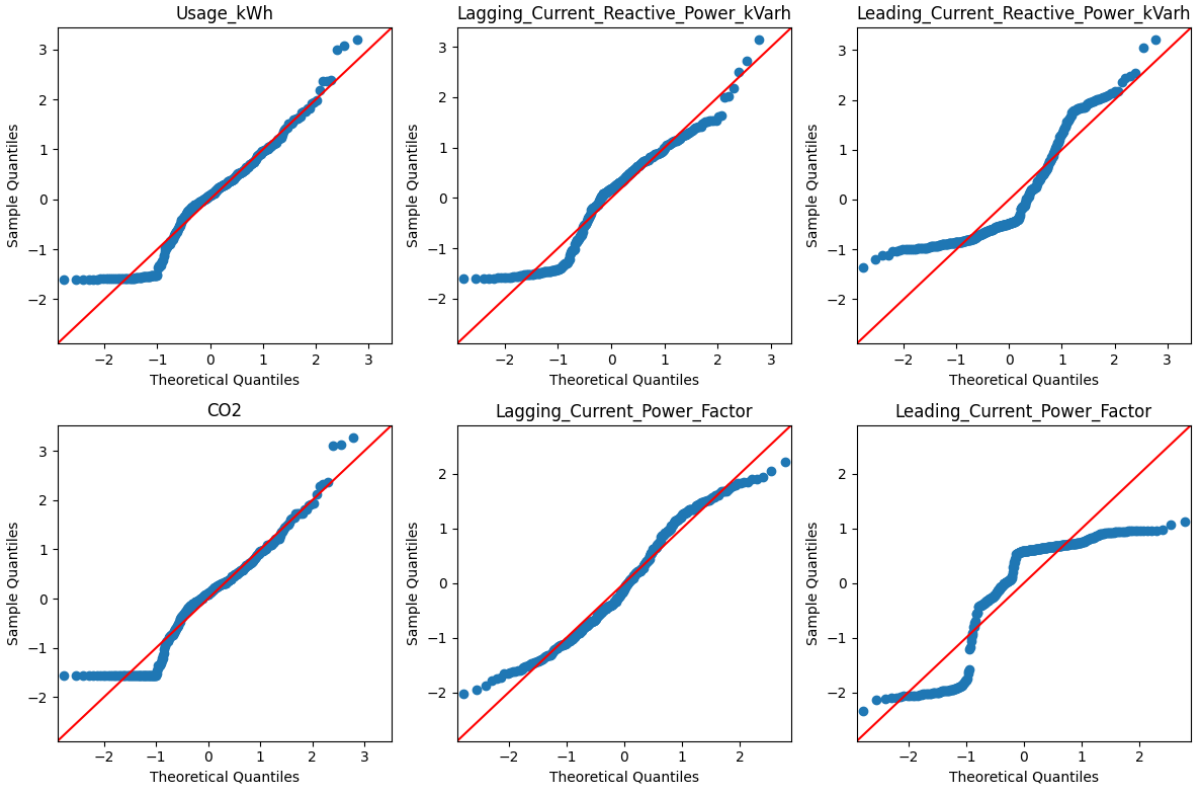

Figure 3: QQ Plots for Univariate Normality

The QQ plot for multivariate normality test is in Fig 4

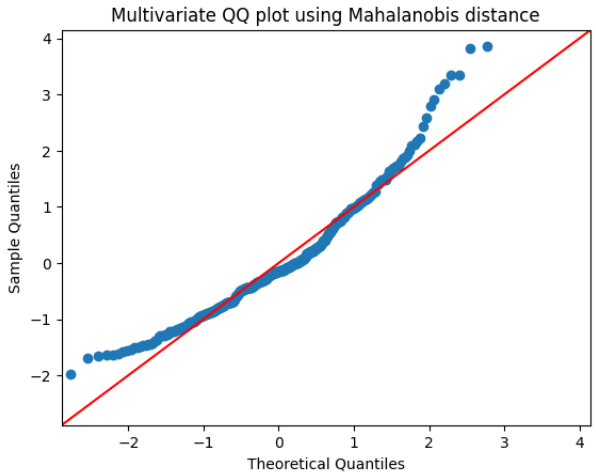

Figure 4: QQ Plot for Multivariate Normality

5 Data Splitting

For splitting the data into two populations, we considered many different splits The data was split into two populations: Summer (April-September) and Winter (October-March) for seasonal analysis, with nearly equal durations for both populations.

6 Covariance Comparison

Using Wilks' Lambda, we tested for the equivalence of covariance matrices between Summer and Winter data:

- Test statistic (-2 log Λ): 544.714
- Critical value at $\alpha = 0.05$: 32.671
- p-value: 0.000
- **Conclusion**: Null hypothesis rejected, indicating a difference in covariance.

7 Mean Comparison

The means of the two populations were compared using Hotelling's T-squared test:

- Hotelling's T-squared: 138.072
- F-statistic: 22.695
- Critical value at $\alpha = 0.05$: 2.124
- p-value: 0.000
- **Conclusion**: Null hypothesis rejected, indicating a significant difference in means.

8 MANOVA

Multivariate ANOVA (MANOVA) was performed taking the original 6 features as dependent variables and the column for season (winter/summer) as the independent variable, to compare means of the variables between the populations. The results obtained are as follows:

- Test statistic: 22.695
- Critical value at $\alpha = 0.05$: 2.124
- p-value: 0.000
- **Conclusion**: Null hypothesis rejected, confirming a significant difference in multivariate means.

9 Profile Analysis

For Profile analysis, we considered a different split than before: we split it between AM and PM data and resampled by monthly mean, so we have two populations with 12 observations each.

9.1 Test 1: Parallel Profiles

- Hotelling's T-squared: 977.834
- Critical value: 16.945
- **Conclusion**: Null hypothesis rejected, indicating differences in profiles across AM and PM data.

Since the test for parallel profiles has failed, we can't perform the tests for coincident profiles or level profiles.

10 Principal Component Analysis (PCA)

After PCA, two principal components were sufficient to capture most of the dataset's variance, providing a compact representation for further analysis.

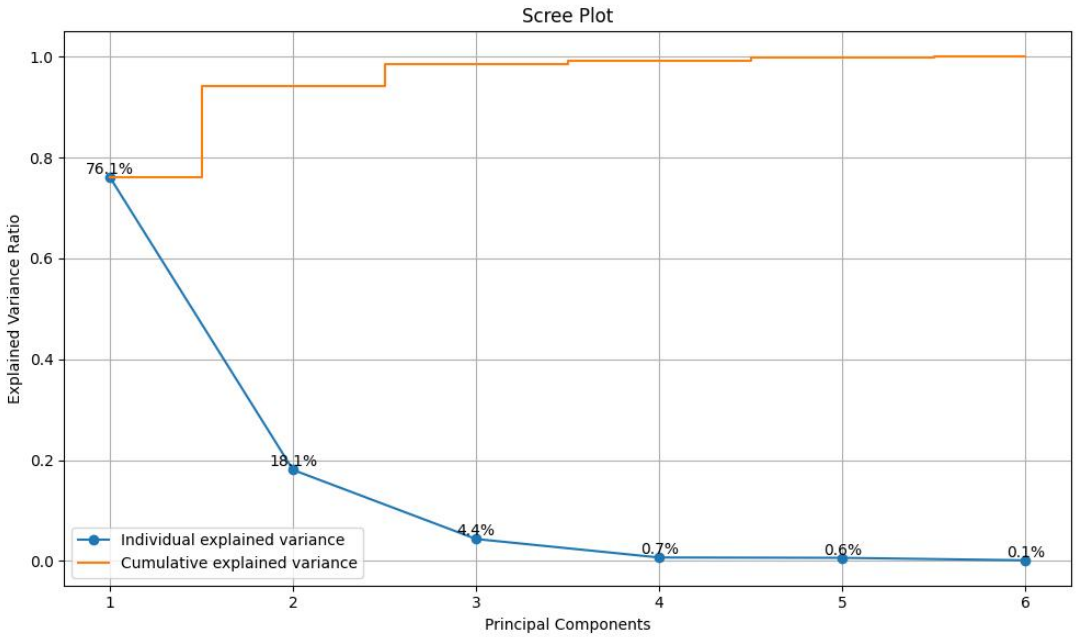

Figure 5: Scree Plot for PCA

11 LDA

We trained an LDA model on PC1 and PC2 to classify the data on weekend or weekday. The data was split into 80% training data and 20% testing data.

The results obtained are as follows:

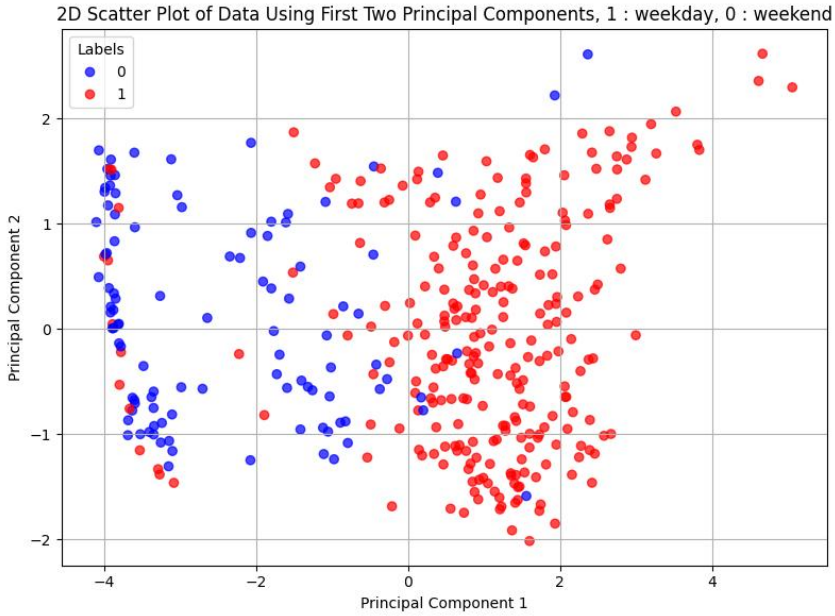

Figure 6: Plotting PC1 and PC2 against each other, separated by weekday or weekend

- Training accuracy: 88%
- Test accuracy: 86%

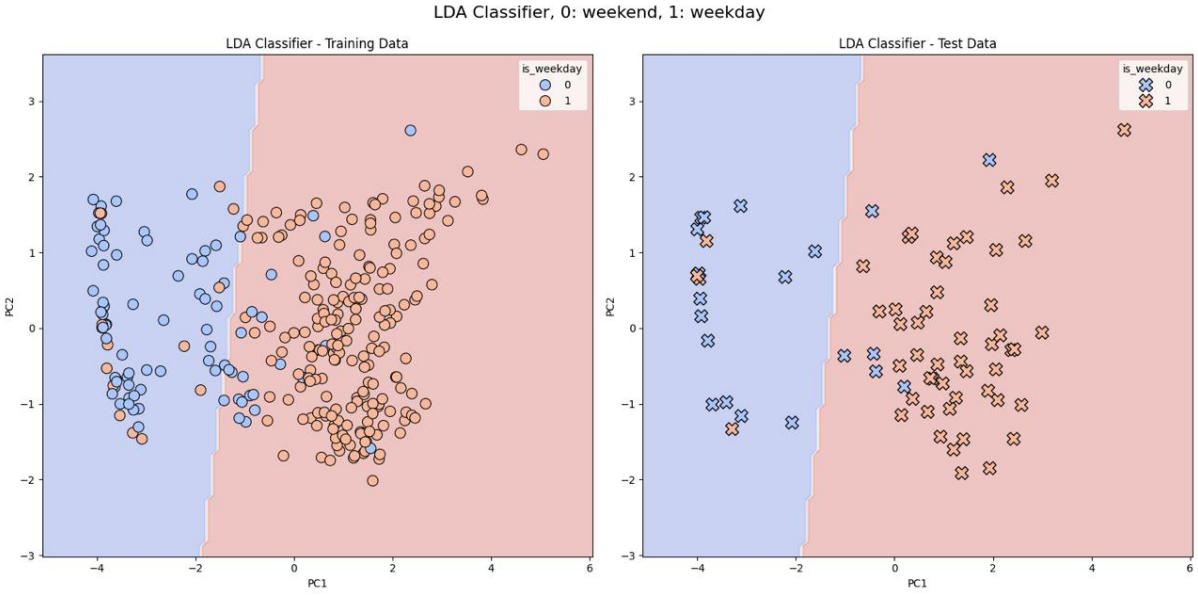

Figure 7: LDA Result

12 Conclusion

The analysis revealed seasonal variations and diurnal patterns in energy consumption. Despite the non-normality of the raw data, preprocessing enabled effective statistical testing, and PCA helped reduce dimensionality while retaining key information.