Faculdade de Ciências e Tecnologia Nova de Lisboa

Masters in Analysis and Engineering of Big Data

Data Analytics and Mining

A Framework for Predicting and Categorizing Energy Yield from Gas Turbines: Exploring Patterns and Relationships in the Data

Professor: Susana Nascimento
Student: João Samarão - N.º 66378
Miguel Santos - N.º 64474

April - 2023

Abstract

Data analysis is a crucial part of any decision-making process in organizations and businesses. The primary goal of this project is to explore a real-world dataset and apply regression analysis, principal component analysis (PCA), and fuzzy clustering techniques to gain insights into the data.

The dataset chosen was the *The Gas Turbine CO and NOx Emission Data Set* since it is a crucial dataset for academics and decision-makers who are concerned about global warming and other contemporary issues.

Regression analysis was applied to establish relationships between variables, more concretely, a regression model was created with a Mean Square Error of 0.00034. In addition, some inference was done.

PCA was used to reduce the dimensionality of the data and identify underlying patterns. From a more specific point of view, the top two principal components are sufficient to explain a vast proportion of the variance and capable of establishing a pattern to identify the Turbine Energy Yield by the hour.

Fuzzy clustering was implemented to enable the identification of groups within the dataset based on similarity. The algorithm couldn't identify the proposed categories, nonetheless, was able to identify some structure between other variables.

Keywords: Data Analytics, Regression Analysis, Principal Component Analysis (PCA), Fuzzy Clustering

Table of contents

1 Introduction

As the world's population continues to grow, so does the energy demand, leading to an increase in the use of fossil fuels. The burning of fossil fuels is a significant contributor to global warming and air pollution, which have severe consequences for the environment and human health. In response, countries worldwide have set ambitious goals to reduce carbon emissions and promote sustainable development. Achieving these goals requires accurate and comprehensive data on energy consumption and emissions, as well as a thorough understanding of the sources and impacts of pollution.

The Gas Turbine CO and NOx Emission Data Set is a crucial dataset for researchers and policymakers who are concerned about global warming and other modern problems. This dataset provides valuable information about the emissions of carbon monoxide (CO) and nitrogen oxides (NOx) from gas turbines, which are significant sources of air pollution.

This dataset is also essential for achieving specific Sustainable Development Goals (SDGs) related to sustainable development and environmental protection. For instance, the dataset can contribute to achieving

- **SDG 3 (Good Health and Well-being)** by reducing the number of deaths and illnesses caused by air pollution.
- **SDG 7 (Affordable and Clean Energy)** by providing information to improve efficiency and reduce the emissions of gas turbines which are a significant source of energy worldwide.
- **SDG 11 (Sustainable Cities and Communities)** by helping policymakers to develop effective strategies to improve air quality in cities and reduce the environmental impact of gas turbines.
- **SDG 13 (Climate Action)** by providing accurate and comprehensive data on emissions from gas turbines, which can inform policies and actions to mitigate the effects of climate change.

In this paper, we present an analytical exploration of the dataset using regression analysis, principal component analysis (PCA), and fuzzy clustering techniques. Regression analysis will be applied to establish relationships between variables, while PCA will be used to reduce the dimensionality of the data and identify underlying patterns. Fuzzy clustering will enable the identification of groups within the dataset based on similarity.

The paper is organized as follows. In the next section, the dataset is introduced with a brief statistical analysis of its features. The methods used are presented in Section 3. Section 4 discusses our results and Section 5 concludes with observations for the future.

2 Description and Analysis of the Dataset

The dataset under consideration is a comprehensive collection of hourly average sensor measurements of eleven variables, which were gathered over five years. The data is comprised of 36,733 instances, providing a detailed record of various environmental and process parameters.

Out of the eleven variables, nine are independent input variables, while the remaining two are target variables. These nine input variables can be further categorized into two groups based on their characteristics. The first group is made up of ambient variables, such as temperature, humidity, and pressure, which represent the atmospheric conditions in which the sensors were located. The second group is composed of process parameters, such as turbine energy yield and air filter difference pressure, which are directly related to the operational processes.

The study presents a comprehensive summary of the variables used, including their names, abbreviations, and basic statistics, which are summarized in Table 1. Supplement 1 displays the histograms for all features. This information offers a quick overview of the dataset, which can be useful for identifying outliers or unusual observations that require further investigation and analysis.

Variable	**Abbreviaton**	**Unit**	**Range**	**Mean**	**Variance**
Ambient temperature	AT	ºC	[6.23; 37.10]	17.71	55.46
Ambient pressure	AP	mbar	[985.85; 1036.56]	1013.07	41.77
Ambient humidity	AH	%	[24.08; 100.20]	77.87	209.13
Air filter difference pressure	AFDP	mbar	[2.09; 7.61]	3.93	0.60
Gas turbine exhaust pressure	GTEP	mbar	[17.70; 40.72]	25.56	17.61
Turbine inlet temperature	TIT	ºC	[1000.85; 1100.89]	1081.43	307.52
Turbine after temperature	TAT	ºC	[511.04; 550.61]	546.16	46.82
Compressor discharge pressure	CDP	mbar	[9.85; 15.16]	12.06	1.19
Turbine energy yield	TEY	MWH	[100.02; 179.50]	133.51	243.94
Carbon monoxide	CO	mg/m^3	[0.00; 44.10]	2.37	5.12
Nitrogen oxides	NOx	mg/m^3	[25.90; 119.91]	65.29	136.38

Table 1: Summary of Variables Used in the Study

Since the data is divided into multiple files, each containing information for a specific year between 2011 and 2015, we decided to add a variable 'YEAR' to the dataset to keep track of the year to which each row of data corresponds. Hence there is no existing timestamp in the data, this variable will provide useful information for temporal analysis and can help identify trends or patterns that may vary over time. With this additional variable, we can easily group and analyze the data based on the year of collection, which may provide insights into how the performance of the turbine has changed over the years.

The correlation matrix in Figure 1 provides a comprehensive overview of the pairwise correlations between the variables using the method of Pearson. The Pearson correlation coefficient formula is:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- r_{xy} is the Pearson correlation coefficient between variables x and y
- x_i and y_i are the ith values of x and y, respectively
- $\bar{x}$ and $\bar{y}$ are the mean values of x and y, respectively
- n is the number of observations in the dataset.

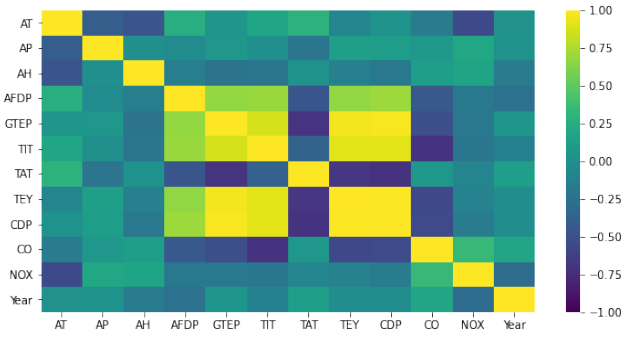


Figure 1: Correlation matrix

Understanding the correlation between variables is important in many areas of research, including data analysis, modeling, and prediction. Through the correlation between variables, researchers can gain insights into the relationships between different factors and how they might affect outcomes of interest. The strength of a correlation is measured by the correlation coefficient, which ranges from -1 to +1. A correlation coefficient of +1 indicates a perfect positive correlation, while a coefficient of -1 indicates a perfect negative correlation. A coefficient of 0 indicates no correlation between the variables.

One notable finding in the correlation matrix is the strongest correlation between TEY (turbine energy yield) and GTEP (gas turbine exhaust pressure) with a correlation coefficient of 0.98. This suggests that GTEP could be a crucial factor in determining the energy yield of the turbine. Additionally, the positive correlation between CDP and GTEP suggests that monitoring and optimizing the compressor's discharge pressure and the gas turbine's exhaust pressure can improve the overall performance and efficiency of the gas turbine system. Such strong correlations indicate that certain features may convey duplicate or overlapping information. Therefore, it may be reasonable to consider excluding some of these variables during model learning to avoid potential issues such as overfitting.

In this paper, we focus on using the Turbine Energy Yield (TEY) as the target variable for our studies. TEY is a measure of the useful work that can be obtained from a system, and it provides a more comprehensive measure of the performance of a gas turbine than other measures such as thermal efficiency.

A set of 2D KDE plots (Supplement 2) was generated to explore the correlation between our target variable and the other input variables. The results showed a clear positive linear relationship between TEY and compressor discharge pressure (CDP), gas turbine exhaust pressure (GTEP), and turbine inlet temperature (TIT) with varying degrees of correlation strength. In contrast, the correlation between TEY and the remaining input variables, including ambient temperature (AT), ambient pressure (AP), and ambient humidity (AH), was less pronounced. These findings suggest that CDP, GTEP, and TIT are key factors to consider when predicting turbine energy yield and that other input variables may have a smaller impact on energy output.

Using TEY as the target variable, we aim to gain a better understanding of the factors that affect the performance of gas turbines and their environmental impact. We will use regression analysis, PCA, and fuzzy clustering to identify patterns and relationships between TEY and other variables such as ambient temperature. This analysis will provide valuable insights into the design and operation of gas turbines, which can inform policies and strategies to reduce their environmental impact and promote sustainable development.

Overall, this dataset provides a valuable resource for researchers and practitioners interested in studying the relationship between environmental and process parameters in a real-world setting.

3 Methods

3.1 Linear Regression

Regression modeling is a statistical approach that enables us to understand the relationship between variables and make predictions based on this understanding. It is a powerful tool that is widely used in various fields, including economics, engineering, social sciences, and healthcare. In this paper, we discuss regression modeling through simple linear regression, which is a technique used to estimate the relationship between two variables: a single predictor (independent) variable and a single response (dependent) variable.

The goal of simple linear regression is to find a straight line that best describes the relationship between the predictor variable and the response variable. The line is represented by a mathematical equation, the Estimated Regression Equation[1], in the form of

$$\hat{y} = b_0 + b_1 x$$

where

- $\hat{y}$ is the predicted response variable
- x is the predictor variable
- b_0 is the y-intercept of regression line
- b_1 is the slope of the line

While b_0 and b_1 are called the regression coefficients, the slope represents the change in the predicted response variable for every one-unit change in the predictor variable, and the y-intercept represents the predicted value of the response variable when the predictor variable is zero. However, this equation only provides an average estimate of the relationship between the two variables and does not take into account the whole population. The **Regression Equation** represents the true linear relationship between the variables for the whole population[1]:

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon$$

This model takes the same form as the previous one with the addition of an error term ϵ. The error term accounts for the randomness in the relationship between the variables. Its value is assumed to be normally distributed with a mean of zero and a constant variance. The line of best fit is determined using the least squares method, which involves minimizing the sum of the squared residuals between the actual response values and the predicted response values. This results in the least squares estimates of the intercept and slope parameters in the regression equation, represented by the formulas:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

where

- n is the number of observations
- x and y are the predictor and response variables
- $\bar{x}$ and $\bar{y}$ are the sample means of the predictor and response variables, respectively.
- b_0 and b_1 are estimates.

To evaluate the performance and analyze the model we can use metrics like

- Mean Squared Error (MSE)

$$MSE = \frac{1}{n-m-1}\sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

- Sum of Squared Errors (SSE)

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y_i})^2$$

- Sum of Squares Total (SST)

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

- Sum of Squared Residuals (SSR)

$$SSR = \sum_{i=1}^{n}(\hat{y_i} - \bar{y})^2$$

Additionally, the regression equation allows for inference and for assessing the overall fit of the model. The T-test is a commonly used statistical test for assessing the significance of the relationship between the predictor and response variables. This test is based on the slope coefficient (β_1) in the regression equation, which indicates the strength and direction of the relationship between the variables and is based on t-distribution with $n-2$ degrees of freedom[16]

$$t = \frac{(b1 - \beta_1)}{s_{b_1}}$$

where s_{b_1} is a point estimate of σ_{b_1} (standard error of the estimate) and

- H0: asserts $\beta_1 = 0$ (no linear relationship exists)
- Ha: asserts $\beta_1 \neq 0$ (linear relationship exists)

A significant T-test indicates that the slope coefficient is different from zero, which means that there is a significant relationship between the predictor and response variables. This method also provides a p-value, which is used for deciding whether to reject or not the null hypothesis (i.e., usually one uses a threshold of 0.05 and rejects if the given value is lower than that).

Confidence intervals are another important tool for making inferences about the population parameters in regression analysis. A confidence interval for the slope coefficient [16] provides a range of values that the population parameter is likely to fall within, based on the sample data. With $100(1-\alpha)\%$ of confidence this interval is given by

$$b_1 \pm (t_{n-2})(s_{b_1})$$

Similarly, for a $100(1-\alpha)\%$ confidence interval for the mean of the response variable (given the predictor)[16], we can be $100(1-\alpha)\%$ confident that the population correlation coefficient ρ lies between:

$$\hat{y_p} \pm (t_{n-2})(s)\sqrt{\frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

where

- x_p is the given value of x, for which prediction being made
- y_p is the point estimate of y, for the given value of x
- t_{n-2} is the multiplier associated with sample size and confidence level
- s is the standard error of estimate

It is also important to mention the confidence interval for the population correlation coefficient ρ[1]. One tells us that we can be $100-\alpha\%$ where ρ lies between:

$$r \pm t_{\alpha/2,n-2}\sqrt{\frac{1-r^2}{n-2}}$$

Overall, confidence intervals are useful for understanding the precision of the estimated population parameter and for making predictions about the response variable. On the other hand, a prediction interval provides a range of values that a random response value is likely to fall within, given a specific value of the predictor variable. For example, a prediction interval for a randomly chosen value of y, given x is defined by

$$\hat{y_p} \pm (t_{n-2})(s)\sqrt{1 + \frac{1}{n} + \frac{(x_p - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

Prediction intervals are always wider than confidence intervals because they account for the variability in the response variable that is not explained by the regression model. These intervals are useful for making individual predictions about the response variable and for understanding the uncertainty associated with those predictions. Globally, these tests and intervals are important tools for making inferences in regression analysis since they provide a way to estimate the population parameters (based on the sampled data) and to understand the precision and uncertainty associated with those estimates.

Finally, verifying regression assumptions is an important step in ensuring the accuracy and validity of the regression model. There are several key assumptions that must be met for simple linear regression to be valid:

- Linearity: The relationship between X and the mean of Y is linear.
- Homoscedasticity(Constant Variance): The variance of residual is the same for any value of X.
- Independence: Observations are independent of each other.
- Normality: For any fixed value of X, Y is normally distributed.

These assumptions can be checked using residual plots like a Normal Probability Plot or ploting Standardized Residuals Against Fits(Predicted Values), and other diagnostic tools (hypothesis tests like Durbin-Watson or Anderson-Darling, for example)

3.2 Principal Component Analysis

The Principal Component Analysis (PCA) is a technique that transforms high-dimension data into lower-dimension while retaining as much information as possible. This technique is widespread when there is a dataset with many features (usually more than thousands of features)[2]. Despite the current dataset only having 11 features, a PCA is always a good methodology to test as it can reduce the variables and still bring better results. Although it is important before applying a PCA to the dataset to verify if we are in the "right" conditions to apply it. In order to do that, there are some "rules"/methods to guarantee that the use of the principal components will be benefic.

- As PCA is based on the existence of a statistical association pattern between variables, in other words, the existence of high bivariate correlations, we should look to the correlation matrix.
- Another way to verify if the PCA is adequate is by using the Sphericity test ([2,13]), which tests the hypothesis that the variables $(x_1, ..., x_p)$ are independent and have the same variance, i.e.:

$$H_0 : \rho = \begin{bmatrix} 1 & \rho_{12} & . & . & . & \rho_{1p} \\ \rho_{21} & 1 & . & . & . & \rho_{2p} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ \rho_{p1} & \rho_{p2} & . & . & . & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & . & . & . & 0 \\ 0 & 1 & . & . & . & 0 \\ . & . & . & & & . \\ . & . & & . & & . \\ 0 & 0 & & & . & 1 \\ \rho_{p1} & \rho_{p2} & . & . & . & 1 \end{bmatrix}$$

Therefore, if all the points above are verified it indicates that a PCA should be appropriate to be applied within our dataset.

Now that we know when should we apply a PCA let's see how to perform one.

Firstly, we need to decide whether we use the dataset with the features as they are or if we apply any kind of scaling. To decide that, we can look if the measurements are done in the same metrics and if those are being measured on the same axis (e.g., vertical and horizontal, different axis). Also, if the variance of each variable is very different, we should scale the data.

Hence the dataset is scaled, it should be noted that if it is standardized the covariance-variance matrix is equal to the correlation matrix. Thus, using the dataset in this way, we are resorting to the correlation matrix to perform the PCA.

Moreover, the objective of a PCA is to find a set of orthogonal vectors that define the most interesting directions, from a statistical point of view, the most interesting are the ones that minimize the loss of information - preserve as much variability as possible. Thus, resorting to the matrix of covariances-variances of our dataset its possible to obtain the vectors that will define the most interesting directions, which are obtained by calculating the eigenvectors, and each eigenvalue will give the information total variance retained by each component (eigenvector). These vectors are

linear combinations that are represented by the following equations:

$$\begin{aligned} PC_1 &= a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p \\ PC_2 &= a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p \\ ... &= ... \\ PC_k &= a_{k1}x_1 + a_{k2}x_2 + ... + a_{kp}x_p \end{aligned}$$

where a_{kp} is the coefficient of the eigenvector k in the p position. Note that the number of principal components (PC) is equal to the number of the p features of the dataset.

After getting our components, it is important to know how much information they will retain, as the objective is to select the minimum components possible. To do that analysis, we resort to the eigenvalues associated with each component. Thus, we can calculate the total sample variance as:

$$\sum_{k=1}^{p} l_k$$

where l_k is the eigenvalue associated with the k component.

Also, the proportion of the total variance explained by the k-th PC is:

$$\frac{l_k}{\sum_{k=1}^{p} l_k}$$

and the proportion of the total variance explained by the first m PC is:

$$\frac{\sum_{k=1}^{m} l_k}{\sum_{k=1}^{p} l_k}$$

Despite acknowledging this information, how do we know how many components should we choose? Well, some techniques can be used to decide, however, in this project only three were used:

1. **Percentage of explained variance**: Retain enough components to explain some *a priori* specified percentage of the total variation of the original variables, typically between 70 and 90%.

2. **Kaiser's criterion**: Exclude principal components whose eigenvalues are less than the average $\frac{\sum_{k=1}^{p} l_k}{p}$, or less than 1 if a correlation matrix has been used.

3. **Scree plot**: Plot of l_k versus k, i.e., the magnitude of an eigenvalue versus its number. To determine the appropriate number of components, look for an elbow (bend) in the scree plot. The number of components is the point at which the remaining eigenvalues are relatively small and all about the same size.

Selected the components that retain most of the total variance, for each observation we will calculate new values that will represent each selected component.

3.3 Fuzzy and Anomalous Clustering

Clustering is a powerful technique in machine learning and data analysis that involves grouping observations into mutually exclusive clusters based on similarities/dissimilarities or distances between objects (without the need for predefined labels). The goal is to find an optimal grouping where the observations or objects within each cluster are similar, while the clusters themselves are dissimilar to each other. This technique is widely used in various fields, such as image recognition, customer segmentation, anomaly detection, and recommendation systems.

Clustering algorithms utilize various mathematical and statistical techniques, such as distance metrics, density estimation, centroid-based approaches, and membership functions, to partition data points into clusters. More specifically, these can be divided, among others, into [2]:

- Hierarchical: hierarchical clustering techniques are designed to group the n items into k clusters based an initial matrix with size $n \times n$, whose elements quantify either similarities or dissimilarities between all pairs of objects. Example: Agglomerative Hierarchical Clustering
- Nonhierarchical: nonhierarchical clustering techniques are designed to group items into k clusters. The number of clusters, k, may be specified in advance (or determined as part of the clustering procedure). It does not require an initial matrix of distances (similarities). Example: K-means
- Fuzzy: fuzzy clustering techniques allow data points to belong to multiple clusters with varying degrees of membership. It assigns membership degrees to each data point indicating the degree to which the point belongs to different clusters. Example: Fuzzy C-Means

Fuzzy clustering algorithms, in particular, introduce fuzziness or uncertainty into the clustering process, allowing data points to belong to multiple clusters. These algorithms do not rely on a matrix of similarity but rather assign membership degrees to each data point which represent the fuzzy or probabilistic nature of cluster assignments, indicating the level of similarity or closeness of a data point to each cluster.

One example of a fuzzy clustering algorithm is the fuzzy C-means (FCM). The algorithm iteratively updates the membership degrees of data points and the centroids of the clusters until convergence. The membership degrees are usually represented as a fuzzy partition matrix nxc, where each entry represents the membership degree of a data point to a particular cluster. These membership degrees are updated based on a fuzzy membership function, which computes the similarity between data points and cluster centroids, taking into account the distances between these. From a more specific point of view, the algorithm can be seen as an optimization of an objective function/clustering criterion and described as [17]:

Given the dataset X, choose the number of clusters $1 < c < n$, the weighting exponent m, the termination tolerance ϵ, the maximum number of iterations l and the norm-inducing matrix A. Initialize the partion matrix randomly, such that $U^{(0)} \in M_{fc}$ and for $l = 1, 2, ...$ repeat

1. Compute the cluster prototypes (means)

$$v_i^{(l)} = \frac{\sum_{k=1}^{n} (\mu_{i,k}^{(l-1)})^m x_k}{\sum_{k=1}^{n} (\mu_{i,k}^{(l-1)})^m}, 1 \leq i \leq c$$

where

- $\mu_{i,k}^{(l-1)}$: The membership degree of the data point 'k' to the cluster 'i' at iteration 'l-1'.

- x_k: The feature vector of the data point 'k'.
- n: The total number of data points.

2. Compute the (euclidean) distances

$$D^2_{i,kA} = (x_k - v_i)^T A(x_k - v_i)\ ,\ 1 \leq i \leq c\ ,\ 1 \leq k \leq n$$

3. Update the partition matrix

$$u^{(l)}_{i,k} = \frac{1}{\sum_{j=1}^{c}(D_{i,kA}/D_{j,kA})^{(2/(m-1))}} \tag{1}$$

- $D_{i,kA}$: The Euclidean distance between the data point 'k' and the centroid of the cluster 'i'.
- $D_{j,kA}$: The Euclidean distance between the data point 'k' and the centroid of the cluster 'j'.

4. Repeat the previous steps until $\|U^{(l)} - U^{(l-1)}\| < \epsilon$ (or the maximum number of iterations is reached)

This flexibility of fuzzy clustering enables it to capture more complex and overlapping patterns in data, making it suitable for handling real-world scenarios with ambiguous data boundaries.

However, this kind of clustering can be very sensitive to outliers and noise within the dataset. To address this issue, pre-processing clustering is used to give initial prototypes to the traditional clusters methodologies implemented, namely, Anomalous Clustering. To apply this methodology one should standardize the data by centering on the mean [1]. Moreover, an iterative process starts by finding the point farthest away from 0, which will be defined as the initial center. Further, we search for a point, at which the distance to the centroid will be smaller than to the origin, assigning that point to the cluster. After, the new centroid is calculated, and if the new centroid is equal to the previous we stop, exclude the points within the cluster from the dataset, and apply again the same methodology, otherwise, we repeat the cluster update until the stop condition is verified (note that one of the stop conditions is running out of points, *i.e.*, the process continues until the dataset is empty). Algorithmically, the method can be described as (just for finding one Anomalous cluster)[1]:

1. Initial center c is the entity farthest away from 0;
2. Cluster update: if $d(y_i, c) < d(y_i, 0)$, assign y_i to S, where y_i is an observation of the standardized dataset, $i = 0, ..., n$;
3. Centroid update: Within-S mean, c';
4. Stop condition: if c' $\neq$ c, go to 2. with c <= c'; Otherwise, halt.

Hence we got all the anomalous clusters it is possible to estimate their contribution to the data scatter using the following formula:

$$DS_k = \frac{|S_c|\langle c_k, c_k\rangle}{\sum_{v=1}^{V} y^2_{vk}}$$

where V is the number of entities of the cluster k, $\langle c_k, c_k \rangle$ is the euclidean squared distance between 0 and c_k, and k = 1, ..., K number of clusters.

Furthermore, it is possible to establish beforehand how many clusters you want by defining a threshold of minimum entities that a cluster needs to have to be considered as a cluster. The last observations (the set of observations that could not be considered as clusters) will join that last cluster created. This methodology tends to converge as usually, the points are closer to the centroids than the mean.

Given the centroids, we can use them as initializers of the Fuzzy C-Means (FCM) and give the initial membership by applying the formula cited above (1).

Despite having a methodology to cluster by being sensitive to noise and outliers (Anomalous clustering), a methodology that allows the observations to belong with a degree of membership to distinct clusters, how can we measure the optimum number of clusters that we should use? This introduces another question, how to measure the 'quality' of clustering structures?

Ordinarily, cluster validation has to do with the "right" number of clusters, c, to be found in a cluster structure however treatment of cluster validity would also include validation of clustering methods and validation of individual clusters. Solution:

- Cluster data for different values of c and apply validity measures *a posteriori* to assess the goodness of the obtained partitions

For this, we enhance the following measures [1]

- Partition Coefficient (PC): measures the amount of overlap between clusters[1,3]. Defined as

$$PC(c) = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} (\mu_{ij})^2$$

- Classification Entropy (CE)[1,3]: measures the fuzziness of the cluster partition only (also known as Partition Entropy). Defined as

$$CE(c) = -\frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{ij} log(\mu_{ij})$$

- Xie and Beni's index (XB)[1,3,4,5]: aims to quantify the ratio of the total variation within clusters and the separation of clusters. Defined as

$$XB(c) = \frac{\sum_{i=1}^{n} \sum_{k=1}^{M} (\mu_{ik})^2 \|x_i - C_k\|^2}{n min_{t \neq s} (\|C_t - C_k\|^2)}$$

Other metrics were used resorting to the library sklearn, those can be checked in the references section [1,3,6,7,8,9,10,11,12,13]. However, no validation index is reliable only by itself and the optimum c can only be detected by comparing all the results. Therefore, we consider that partitions with fewer clusters are better when the differences between the values of a validation index are minor. Thus, we can say that the procedure to find the optimum number of clusters is to find the best c for each validation index in the study, compare the results and make a "fair" decision.

4 Results and Discussion

4.1 Regression Analysis

As mentioned in Section 2, we will use the Turbine Energy Yield (TEY) as our target variable for our analysis. Based on Figure 1 and Supplement 2, we can see there are two main contenders for the independent variable: Compression Discharge Pressure (CDP) and Gas Turbine Exhaust Pressure (GTEP). From these, we selected CDP to be our predictor variable due to its correlation with the response variable.

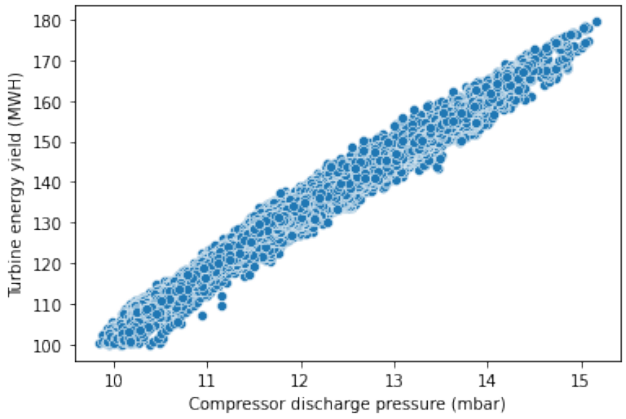


Figure 2: TEY vs CDP: Scatter Plot

As we can see, the scatter plot (Figure 2) displays a pattern that appears to be linear-like, suggesting that there is a relationship between the two variables. Furthermore, the plot indicates a positive association between the two variables, implying that as the independent variable increases, the dependent variable also tends to increase. With all set, the first step was to build the regression model, and the following results were obtained

$$\hat{y} = \beta_0 + \beta_1 x <=> \hat{y} = -37.56 + 14.18x$$

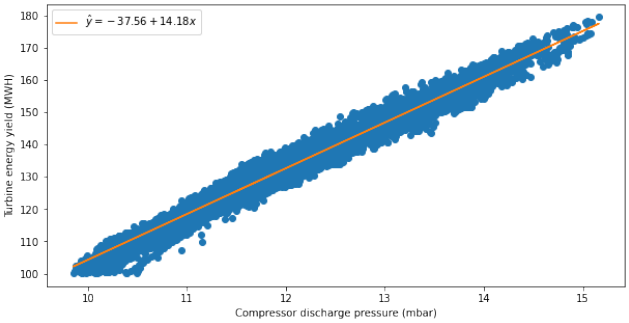


Figure 3: TEY vs CDP: Scatter Plot with Regression Line

Considering the obtained values for the parameters β_0 and β_1, we can interpret

- $\beta_1 = 14.18$: per unit increase in CDP, the estimated TEY increases by 14.18 units (on average). This also confirms the presence of a positive relationship between the two variables
- $\beta_0 = -37.56$: the interpretation of this value is meaningless since the predictor variable is unlikely to be zero in the context of the dataset

Before proceeding with the regression analysis, we need to ensure that the regression assumptions are satisfied. To verify these assumptions, we used residual plots and normal probability plots.

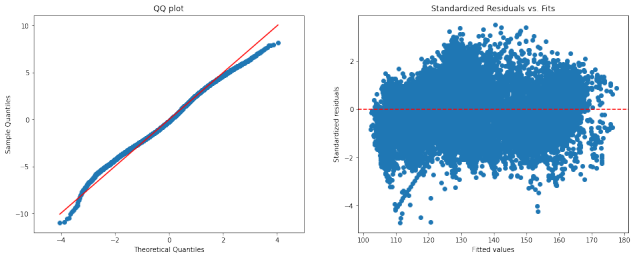


Figure 4: Regression Assumptions Validation: Linear Regression

Analyzing Figure 4, more concretely, the Quantile-Quantile plot of the quantiles of residuals against quantiles of standard normal distribution, we can confirm the Normality of the residuals since the majority of the points lie on the straight line. Focusing on the plot *Standardized Residuals Against Fits*, no discernible patterns are detected (the data points form an overall rectangular shape) so the regression assumptions remain intact.

However, another experiment was made: Logarithmic Transformation, i.e., the natural log of both of the variables was taken and a linear regression on the transformed features was performed. This generated the following results

$$log(\hat{y}) = \beta_0 + \beta_1 log(x) <=> log(\hat{y}) = 1.67 + 1.29 log(x)$$

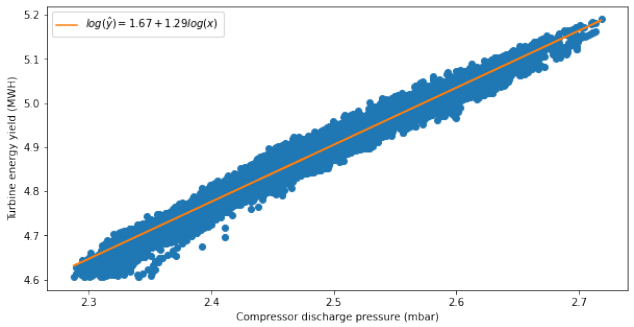


Figure 5: log(TEY) vs log(CDP): Scatter Plot with Regression Line

Which can be interpreted as

- $\beta_1 = 1.29$: per unit increase in log(CDP), the estimated log(TEY) increases by 1.29 units (on average).

- $\beta_0 = 1.67$: when log(CDP) is zero (i.e. when CDP is equal to 1), the expected value of log(TEY) is 1.67

To ensure the regression assumptions are satisfied the same method was used and we obtain the following plots

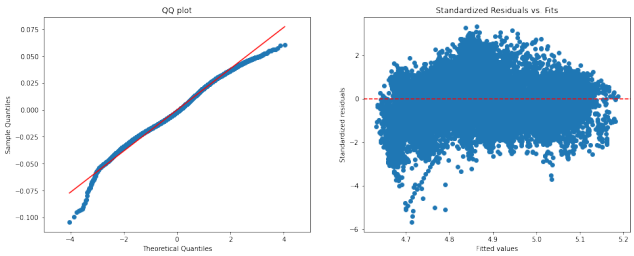


Figure 6: Regression Assumptions Validation: Logarithmic Transformation

In similarity with previous results, the plots in Figure 5 show no discernible patterns, therefore the regression assumptions also remain intact. In some cases, transforming the variables using logarithmic transformation can improve the model's performance. Thus, comparing the results of the regression analysis with and without applying the logarithmic transformation to the variables.

	SSR	SST	r^2 (%)	r (%)
Normal Variables	8760483.616	8960467.979	97.77	98.88
Log Scaled Variables	497.418	509.831	97.57	98.78

Table 2: Linear Regression Results: With and Without Logarithmic Transformation

Inspecting Table 2, we can see a major decrease in SSR and SST where the scale of the data is the primary cause of this reduction. Nonetheless, we can comment on the correlation and determinacy coefficients

- Coefficient of Determination: $r^2 = 97\%$. This coefficient measures the proportion of variability in the response variable explained by ERE and the obtained value can be considered very good
- Coefficient of Correlation: $r = 98\%$. This coefficient measures the strength of the linear relationship between the two variables and the obtained value indicates that these are positively correlated

	SSE	MSE	s	Sb1	T	P-value
Normal Variables	199984.362	5.445	2.333	0.011	1268.476	0.000
Log Scaled Variables	12.413	0.00034	0.018	0.00107	1213.218	0.000

Table 3: Linear Regression Results: With and Without Logarithmic Transformation cont.

Analyzing Table 3, we can see that major improvements (in the model's performance) were detected. This is explained by the fact that applying this kind of transformation helps to address issues related to nonlinearity, heteroscedasticity, and outliers. In fact, in Table 4, we have some

concrete examples of the model's performance where $Diff1$ corresponds to the difference between the real value and the predicted (without the logarithmic transformation) and $Diff2$ to the difference between the real value and the predicted (with the logarithmic transformation). Note that, for the logarithmic values we have that $\hat{y} = e^{1.67+1.29log(x)}$

Predict	Log Predict	True Value	Diff1	Diff2
131.2	130.97	134.67	3.47	3.7
131.12	130.88	134.67	3.55	3.79
133.24	133.02	135.1	1.86	2.08
132.51	132.28	135.03	2.52	2.75
131.37	131.14	134.67	3.3	3.53

Table 4: Predicted values vs Real values

It's important to notice these are just a few examples and the measures presented in Table 3 are calculated on the whole dataset. With all experiments done, we can do some inferences. From now on we will use a rejection threshold, α, of 0.05.

First, let's confirm the existence of a linear relationship between the two variables and for this, we can use the T-test mentioned above. In Table 3 we have all values that we need, more precisely the p-values, that are lower than our α, so H0 is rejected. This confirms the existence of a linear relationship between the variables in the analysis.

Moreover, we built a 95% confidence interval for the unknown true slope of the regression line, using the formulas mentioned in section 3.1, and obtained

$$CI\ for\ \beta_1 [14.16197853; 14.20581197]$$

$$Log\ CI\ for\ \beta_1 [1.29089046; 1.29506824]$$

Since zero (our hypothesis was $\beta_1 = 0$) is not contained in the confidence intervals, we are 95% confident in a linear relationship between the Turbine Energy Yield and Compression Discharge Pressure. Likewise, a 95% confidence interval for the population correlation coefficient was constructed

$$Correlation\ Coef.\ CI\ [0.98724993; 0.99030561]$$

$$Log\ Correlation\ Coef.\ CI\ [0.98615558; 0.98934713]$$

With this, we can determine the range of values that the true correlation coefficient lies within with a 95% level of confidence. This also tells us the relation between the two variables, i.e., since both endpoints of the confidence interval are positive, then TEY and CDP are positively correlated, with a confidence level of 95%.

Additionally, a 95% confidence interval for the mean of TEY at a fixed value of CDP was produced. For the fixed value of CDP, we chose the observation with index = 13 (no specific reason for the choice) and value = 13.929 mbar, we obtained:

$$CI : [159.97502683; 160.04228339]\ for\ x_{13}$$

This informs us that we can be 95% confident that the mean TEY by all elements with CDP = 13.929 mbar, lies between 159.975 and 160.042 MWH. A wider interval implies greater uncertainty in the prediction, while a narrower interval implies greater precision since we can consider this one as narrow, it can be useful for making predictions about future observations.

Finally, we calculated the prediction interval for a randomly chosen value of y given the same x chosen in the last step which generated

$$Prediction\ Interval : [155.43508032; 164.58222989]\ for\ x_{13}$$

Interpreting it, we can be 95% confident that the Turbine Energy Yield by a randomly selected element with a Compressor Discharge Pressure of 13.929 mbar, lies between 155.435 and 164.582 MWH (which we can consider useful).

4.2 Principal Component Analysis

In this section, six variables were selected to study the target variable (Turbine Energy Yield). These variables were selected based on their correlation with the response variable (Figure 1 and Supplement 2).

The first step was checking if it was necessary to scale this partition of the dataset. From Section 2, we already know that these variables have different metrics (Table 1) and by looking at Table 5, the variances are very distinct from variable to variable (even the ones with the same metrics).

Variable	**Variance**
CDP	1.185443
TIT	307.516004
GTEP	17.605580
AFDP	0.598960
TAT	46.816622
AT	55.463020

Table 5: Variance of the variables.

Thus, two different scales were applied (Supplements 3 and 4) and compared with the original data in 2D and 3D plots (Figures 7, 8, and Supplement 6). For the sake of brevity, only one 3D plot was shown as the interaction between the variables in the different scales has identical behaviors.

The variable Compressor discharge pressure (CDP) has a direct relationship with the Turbine energy yield (TEY), as the discharge increases, the energy increases too. On the other hand, Ambient temperature (AT) showed to not contribute to an increase of the energy independently of the temperature. In the 3D plot, it was evaluated one more variable - Gas turbine exhaust pressure (GTEP) which has an identical relationship with the energy as the discharge.

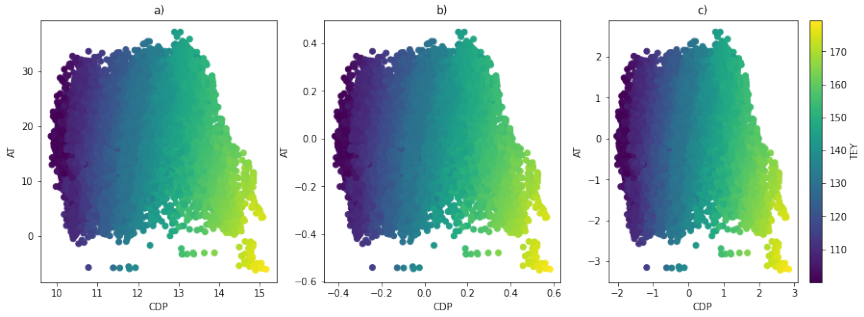


Figure 7: a) Non-Scaled visualization; b) RangeScaler visualization; c) StandardScaler visualization

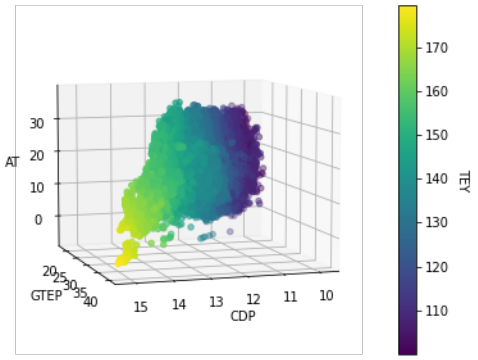


Figure 8: 3D Visualization of the non-scaled data

Despite applying the PCA on the different scaled and non-scaled datasets, it is always a good routine to analyze if the Principal Component Analysis will benefit our study. For this purpose, we looked at the correlation matrix (Figure 1) where it is essential for the variables to have a high bivariate correlation. Also, we applied the sphericity test, p-value $< 10e^{-6}$ (Supplement 5) rejecting the null hypothesis and concluding that a PCA would benefit our analysis to find underlying patterns within our data.

Moreover, by applying a PCA on the different scaled datasets, through the criteria cited in Section 3.2, Table 6 represents the total variances explained by the selected number of components and Figure 9 illustrates the screeplots for each Data Scale. For the first criterion, a threshold of at least 90% of the total explained variance was defined.

Hence all the results were produced it is important to analyze what was the impact of the Data Scaling on dimensionality reduction. The objective is to have fewer features (components) and retain most of the variance. Table 6, shows that scaling the original dataset does not bring many improvements as the original data with two components retains 93% of the total variance which is identical to the scaled datasets, but with three components. However, using the RangeScaler

brought better results than Standardization.

Moreover, Figure 9 represents the total variance retained by the components, whereas b) and c) tend to have a lower difference between variance retained from components 3 to 4, despite c) having a bigger difference in variance retained from components 4 to 5, which could indicate that we should retain more than 3 components. Figure 9 a) illustrates that would be more beneficial to retain the first four components.

Data Scaling		**Total Variance Explained**	**Number of Components**
Non-Scaled	Total	93.12%	2
	Kaiser	77.48%	1
Range Scaler	Total	94.98%	3
	Kaiser	87.76%	2
Standard Scaler	Total	93.45%	3
	Kaiser	85.82%	2

Table 6: Summary of Principal Component Analysis (PCA) results with different data scaling methods.

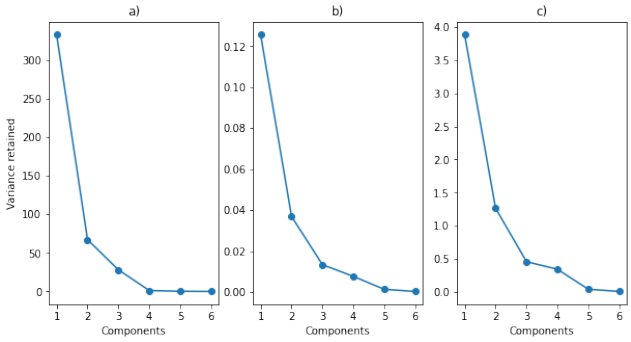
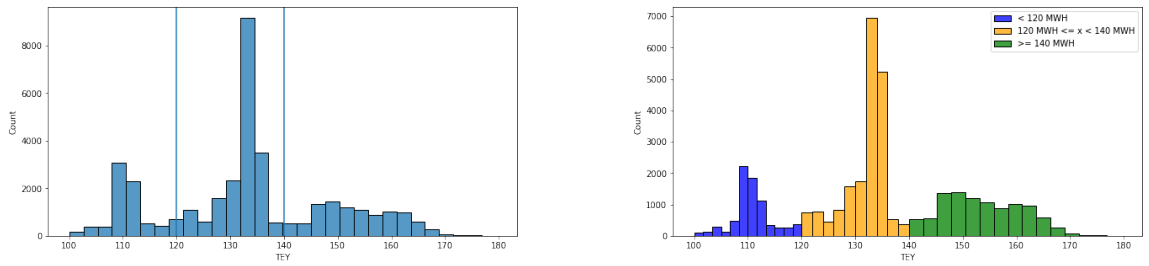


Figure 9: Screeplots a) Non-Scaled Data; b) Range-Scaled Data c) Standard-Scaled Data

Besides analyzing analytically the results obtained by the component analysis, let's compare visually the results obtained. As our main goal is to study the energy yield by the turbine, instead of analyzing it as a continuous variable, let's group the values into three subgroups (Figure 10). The first group (from left to right in Figure 10 b)) represents 20.58% of the points on the dataset, the second group 52.33%, and the third one 27.06%.



(a) (b)

Figure 10: (a) Thresholds to select the energy subgroups; (b) Energy subgroups

For the sake of brevity, only one of the PCA visualization is shown, the others can be found in supplement 7. From Figure 11, the first component can differentiate the three subgroups as for the other components, when plotting PC2 vs PC3 it is clear that they cannot distinguish the different subgroups. Therefore, in order to retain at least more than 85% of the total variance independently of the data scale used, the optimum number of components would be two.

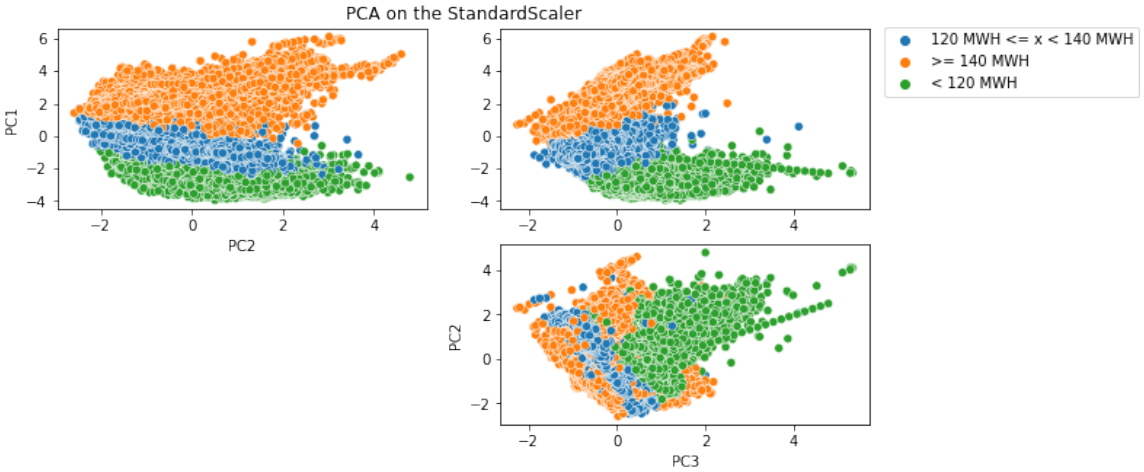


Figure 11: Principal Component Visualization on the StandardScale Data

Once we have used PCA and selected the optimum number of principal components for our data, it is important to examine the meaning of each PC and how it correlates with the variables of interest. The PCs themselves are linear combinations of the original variables and do not have any inherent meaning until we interpret them. Therefore, analyzing the meaning of the PCs is crucial to understanding the underlying patterns in the data.

In (2) we have the equations for our components, where usually the first component usually is related to the mean of the variables [2], however, in this case, we already have a contrast between TAT and the other variables. As for the second component, it is possible to analyze a contrast between CDP, GTEP and TIT, AFDP, TAT, and AT. Although we have contrasts in the second component, it is very complicated to interpret them in the context of the energy yield by the turbine. As for the first one, when the positive coefficients have higher values from the observations it is most

likely that the turbine is yielding energies above 140 MWH. If the observations have lower values then there is a higher chance of the turbine yielding energies between 120 MWH and 140 MWH. Lastly, if the temperature after the turbine working is high (TAT) and the rest of the observations have lower values it is more likely to have energies below 120 MHW.

$$PC1_i = 0.50 \times CDP_i + 0.46 \times TIT_i + 0.49 \times GTEP_i + 0.41 \times AFDP_i - 0.36 \times TAT_i + 0.04 \times AT_i$$
$$PC2_i = 0.46 \times CDP_i - 0.19 \times TIT_i + 0.04 \times GTEP_i - 0.23 \times AFDP_i - 0.46 \times TAT_i - 0.83 \times AT_i \quad (2)$$

Additionally, we can investigate how each PC correlates with the variables of interest by examining the correlation coefficients (Figure 12 a)). This analysis can provide insights into the underlying structure of the data and help us identify important variables that may be driving the observed patterns [2]. Furthermore, using the correlation and powering it to two, we obtain how much a variable is explained by a component (Figure 12 b)). Figure 12 a), illustrates that the first component is influenced positively by the variables CDP, TIT, GTEP, and AFDP; Negatively by TAT and it has a poor correlation with AT. The second component has a correlation between -0.5 and 0.25 with most of the variables, but it has a strong negative correlation with AT. Concurring on the same outcomes of analyzing the principal components' coefficients.

Relatively to Figure 12 b), most of the variables are explained by the first component, as the second component only explains the variable AT.

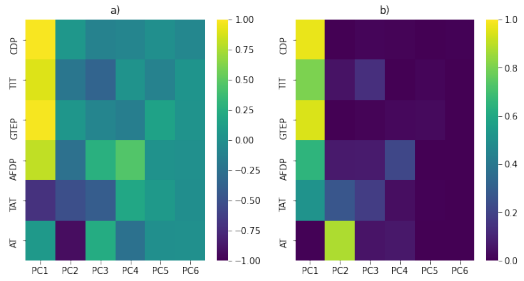


Figure 12: Correlation and Explainability between the variables and the PCs

4.3 Fuzzy Clustering and Anomalous Patterns

The previous sections had more focus on predicting or discriminating the target variable TEY. However, in the cluster algorithms, despite comparing the results with the target variables, we have more details about the data structure as these provide information about the centroids and membership of the variables within the clusters.

Figure 13 illustrates the metrics for the clusters obtained by two algorithms implemented. The Fuzzy C-Means (FCM) without any specific initialization (only how many clusters we want to produce) and the Anomalous Pattern Fuzzy C-Means (AP FCM). As cited earlier, to decide which is the optimum number of clusters we should not answer on just one metric.

Thus, analyzing the silhouette score which reveals the goodness of a clustering technique, whereas positive values close to one indicate that clusters are well apart from each other and clearly distinguished, otherwise they are assigned in the wrong way [6,11]. The best scores were obtained for clusters equal to two and four.

Calinski Harabasz measures the sum of between-cluster dispersion against the sum of within-cluster dispersion [10], therefore higher values are translated in well-defined clusters. Once again,

higher values were obtained for the cluster mentioned above.

Contrary, lower values on Davies Bouldin indicate that the clusters are not similar to each other [12]. When using four clusters we obtain fewer similar clusters.

For Xie-Beni, which was explained before, lower values illustrate that there is a good separation between the clusters. Using two and four clusters produced a better separation.

Partition Coefficient measures the amount of overlap between clusters, whereas values closer to one the better the fuzzy partition will be. The closest partition is with two clusters.

Adjusted Rand Score takes some ground of truth, in this case, the ground of truth was the categories defined in the Principal Component Analysis. Values closer to one mean the cluster is identical to the ground of truth established [13]. Thus, the best scores obtained using this metric were for three and four clusters.

Lastly, partition entropy, or classification entropy, can be interpreted as measuring the ambiguity associated with a fuzzy partition. Contrary to Partition Coefficient, a good cluster identification would be indicated by an entropy value close to 0. Therefore, the best cluster partition would be using two clusters.

Also, it should be noted, that independently of the cluster methodology used they both converged to the same result, Figure 13 illustrates an overlap between the lines.

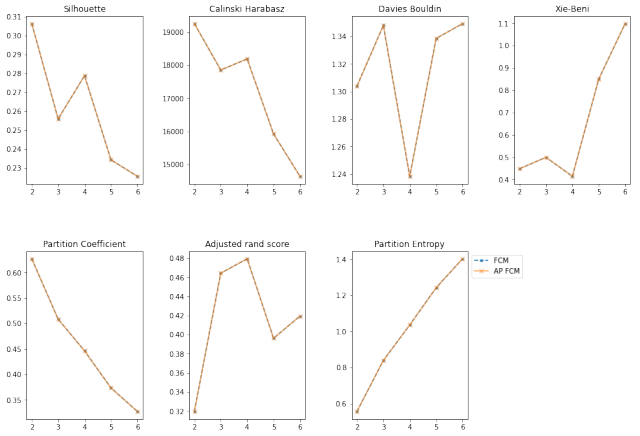


Figure 13: Measurements of the quality of the clusters

It is possible to have better separation and fewer similarities between clusters using 4 partitions (Xie-Beni and Davies Bouldin). Also, if we want to consider the ground of truth that was established, once again, a partition of four clusters would be the best. Overall, using two or four clusters produced better results. However, for the reasons explained, we will proceed with this study with four clusters.

Figure 14 illustrates a projection of the clusters on the first two principal components. In a) red crosses represent the initial centers obtained using the Anomalous clusters and blue squares as the final centers. b) only shows the final centers without using any initialization in FCM. c) Illustrate how should be the clusters using the ground of the truth. The clusters produced by the algorithms were really different from the expected clusters Figure 14 - c). Thus, we should go deeper into the analysis of the cluster structure. However, which method should we choose to analyze?

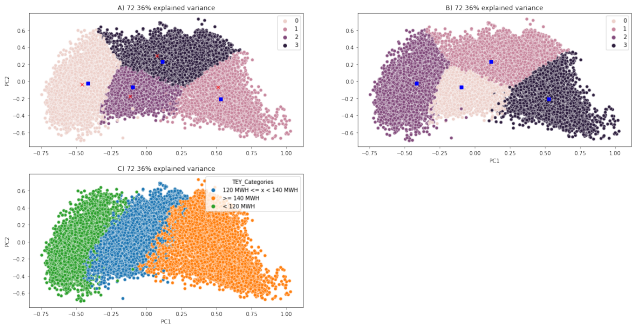


Figure 14: a) AP FCM Clusters and Centroids; b) FCM Clusters and Centroids; c) Clusters using the ground of truth

Previously it was cited that both converged for the same solution, reaching the same measurement scores. So to decide which one we should analyze we evaluated the time of convergence of each method when applying a partition of four clusters. Using the Anomalous Pattern (AP) the method converged to the solution in 41.67 seconds; without using the initialization parameters given by the AP, the method converged in 56.65 seconds. Moreover, by analyzing Figure 15 it is possible to verify why it is faster to use the Anomalous Pattern, the loss function [15] starts smaller making it converge faster for the optimum solution. Whence, for further analysis, we will use the AP FCM results.

Despite proceeding with the analysis with four clusters, if we were to decide the optimum number of clustering only by looking at the loss, the best fit would be six clusters.

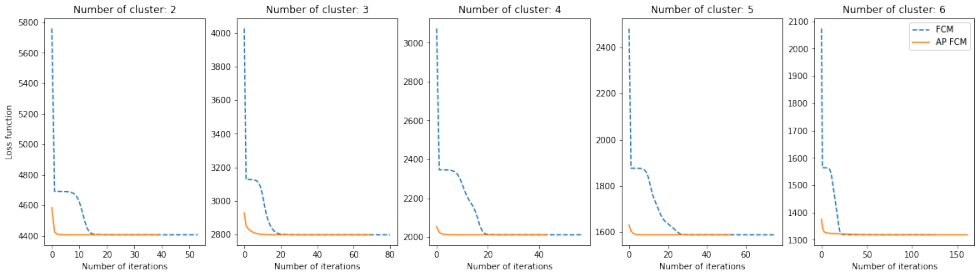


Figure 15: Loss function score on the clustering algorithms for different clusters

Figure 16 illustrates the membership of forty observations. Through this figure, it is possible to analyze that the first ten observations, in a crisp partition, would belong to Cluster 1. However, these observations have a higher membership with Cluster 3. If we look at Table 7, which has the centroids of each cluster, the difference between the two centroids (in Figure 14 a) 0 - Cluster 1 and 2 - Cluster 3) is not that big, only the variables GTEP, TIT, and CO differ with a bigger difference.

For the following ten points, it is very clear that they have a strong membership with Cluster 2. Furthermore, from observations 0 to 9, despite weaker memberships, it is clear that in a crisp partition, they would belong to Cluster 3 (Figure 14 a) - Label 2).

The last ten observations, similarly to the first ten share an identical membership with cluster 3. However, only analyzing these observations and relating them to the centroids is not enough to understand the underlying structure of our dataset.

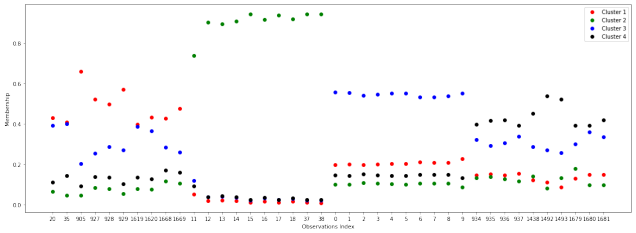


Figure 16: Membership of forty observations

	AT	AP	AH	AFDP	GTEP	TIT	TAT	CDP	CO	NOX
Cluster 1	14.898	1013.641	84.674	3.735	24.556	1081.614	549.254	11.854	2.013	67.425
Cluster 2	24.640	1011.121	66.304	4.195	26.521	1089.802	548.348	12.300	1.645	61.941
Cluster 3	16.586	1012.557	83.556	3.251	20.827	1058.658	548.987	10.784	4.226	65.793
Cluster 4	13.558	1015.891	78.675	4.660	31.819	1098.627	533.578	13.707	1.294	65.119
Grand Mean	17.713	1013.070	77.867	3.926	25.564	1081.428	546.159	12.061	2.372	65.293

Table 7: Centroids and Grand Mean (the centroids were converted to their normal units)

Table 8 illustrates the difference between the centroids and the grand mean (the mean of all features). Starting with variable AT, cluster 2 holds greater values (than the mean) for this variable, as for the other clusters they all contain smaller values, however, Cluster 3 is closer to the mean. The variable AP does not show a big difference from the mean, yet we can still highlight cluster 2 for holding smaller values and cluster 4 for larger values.

Lower values for AH are most likely to be in cluster 2, higher values have a tendency to be in cluster 1 or 3. Relatively to AFDP, we can divide it into categories, values below and above the mean, but close to it (Cluster 1 and Cluster 2 respectively) and values below and above the mean, but away from it (Cluster 3 and Cluster 4 respectively). The same interpretation can be retrieved for the variable GTEP.

The variable TIT in most of the centroids is really close to the mean, although, we can emphasize that Cluster 3 will have smaller values and Cluster 4 higher values. Looking at the TAT variable we can only interpret that Cluster 4 will have smaller values.

CDP has the same similarity division as AFDP and GTEP. CO has higher values in cluster 3 and smaller values in the other three clusters, highlighting cluster 4 for retaining the smallest ones. Lastly, NOX values are really close to the mean too, however, in cluster 1 we can find higher values. Lower values are most likely to be in cluster 2.

Moreover, if we gather all the information and select only the differences greater than ten, Cluster 1 will have smaller values for AT and CO, comparing Figure 14 a) and Figure 14 c) this cluster has most of its values where the energy yield by the hour is less than 120 MWH. Cluster 2 has higher values for AT and lower for AH and CO. Once again, looking at the same figures cluster 2 will have more points where the energy is greater than 140 MHW. Note that, CO levels are low and similar despite some clusters having CO levels greater than the mean. Clusters 3 and 4 are the opposite, whereas the first one, is most likely to have lower values for AFDP, GTEP, and CDP, and Cluster 4 is presumably to have higher values. These two clusters when compared with Figure 14 c) represent most of the points that have an energy yield between 120 MWH and 140 MWH. In this case, we cannot clearly use this information to decide which variables will contribute more to the energy

yield, however, the cluster structure compares the turbine attributes (Cluster 3 against Cluster 4), and Ambient Variables and Nox are compared in Clusters 1 and 2.

	AT	AP	AH	AFDP	GTEP	TIT	TAT	CDP	CO	NOX
Cluster 1	-15.89%	0.06%	8.74%	-4.86%	-3.94%	0.02%	0.57%	-1.71%	-15.13%	3.26%
Cluster 2	39.11%	-0.19%	-14.85%	6.87%	3.74%	0.77%	0.40%	1.99%	-30.64%	-5.13%
Cluster 3	-6.36%	-0.05%	7.31%	-17.19%	-18.53%	-2.11%	0.52%	-10.58%	78.14%	0.77%
Cluster 4	-23.46%	0.28%	1.04%	18.70%	24.47%	1.59%	-2.30%	13.65%	-45.47%	-0.27%

Table 8: Difference to the Grand Mean

5 Conclusion

The Gas Turbine CO and NOx Emission Data Set is a valuable resource for researchers and engineers working to improve efficiency and reduce the environmental impact of gas turbines. By applying regression analysis, principal component analysis (PCA), and fuzzy clustering, we were able to explore the relationship between various features and demonstrate that the dataset contains valuable information that can be used to optimize gas turbine performance and minimize harmful emissions.

The regression analysis revealed a strong positive correlation between TEY and several features, but the linear correlation between TEY and CDP ($r = 98.88$) highlights the importance of optimizing compressor operation to maximize energy yield. Applying a Logarithmic Transformation to the model resulted in great improvements, more precisely, a reduction of the Mean Squared Error from 5.445 to 0.00034. However, as we are not experts on turbines and how they work, a unidimensional model may be too irrealistic, therefore adding more features to the model, or even using other methodologies such as Polynomial Regression, Neural Networks, or Regression Trees should be considered in future work.

The PCA analysis was performed under a subset of features, namely, CDP, TIT, GTEP, AFDP, TAT, and AT. Applying this methodology revealed that the top two principal components explained a huge portion of the variance (85.82% to 93.12%). Those were capable of establishing a pattern for identifying the energy yield by the hour. Also, applying the PCA using different scales did not bring significant improvements, however, as the variables have different metrics and variances, we should scale the subset used. From the scales used, we decide that using the Standard Scaler would be better because it provides an easier interpretably. These results suggest that feature selection and dimensionality reduction techniques can be useful tools for improving model accuracy and interoperability. Future work should be applied using the Principal Components to do predictive models or even cluster analysis.

Lastly, the fuzzy clustering analysis could not clearly identify the energy categories proposed, however, it could identify some structure between the turbine attributes and the ambient attributes. Relatively to the emissions of CO and NOX, independently of the other features, the analysis of the clusters' centroids demonstrated that they are always constant.

Further investigation into the characteristics of these clusters may reveal additional insights into the system's behavior and inform the development of more effective control strategies.

Future work should include a deep understanding of which variables we can use to apply clustering algorithms, as some of the features showed not to be so relevant in the clustering process, namely,

AP, TIT, TAT, and depending on what you consider close to the mean, NOX emissions. Besides analyzing the variables, other algorithms should be tested such as Spectral Cluster. Resorting to this algorithm, it is possible to cluster the variables using similarities matrixes, which could bring better results as it could establish a pattern for the energy categories proposed.

6 References

1. Susana Nascimento, Data Analysis and Mining Slides, 2^{nd} semester 2022/2023 - Master in Analysis and Engineering of Big Data

2. Regina Bispo, Multivariate Statistics Slides, 1^{st} semester 2022/2023 - Master in Analysis and Engineering of Big Data

3. Yongli Liu, Xiaoyang Zhang, Jingli Chen, Hao Chao, "A Validity Index for Fuzzy Clustering Based on Bipartite Modularity", Journal of Electrical and Computer Engineering, vol. 2019, Article ID 2719617, 9 pages, 2019. https://doi.org/10.1155/2019/2719617

4. xb: Xie-Beni Index In zcebeci/fcvalid: Internal Validity Indexes for Fuzzy and Possibilistic Clustering, RPackage formula

5. XB: Xie and Beni index In fclust: Fuzzy Clustering

6. Ashutosh Bhardwaj, May 26, 2020 - Silhouette Coefficient, Validating clustering techniques

7. Kay Jan Wong, Dec 9, 2022 - 7 Evaluation Metrics for Clustering Algorithms, In-depth explanation with Python examples of unsupervised learning evaluation metrics

8. DAN-DUMITRU DUMITRESCU, 9 - Fuzzy Hierarchical Classification Methods in Analytical Chemistry, Editor(s): DENNIS H. ROUVRAY, Fuzzy Logic in Chemistry, Academic Press, 1997, Pages 321-356, ISBN 9780125989107, https://doi.org/10.1016/B978-012598910-7/50011-1.

9. API design for machine learning software: experiences from the scikit-learn project Lars Buitinck (ILPS), Gilles Louppe, Mathieu Blondel, Fabian Pedregosa (INRIA Saclay - Ile de France), Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort (INRIA Saclay - Ile de France, LTCI), Jaques Grobler (INRIA Saclay - Ile de France), Robert Layton, Jake Vanderplas, Arnaud Joly, Brian Holt, Gaël Varoquaux (INRIA Saclay - Ile de France), 1 Sep 2013, https://doi.org/10.48550/arXiv.1309.0238

10. T. Calinski and J. Harabasz, 1974. "A dendrite method for cluster analysis". Communications in Statistics

11. Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics 20: 53-65.

12. Davies, David L.; Bouldin, Donald W. (1979). "A Cluster Separation Measure". IEEE Transactions on Pattern Analysis and Machine Intelligence. PAMI-1 (2): 224-227

13. L. Hubert and P. Arabie, Comparing Partitions, Journal of Classification 1985

14. Sphericity test, wikipedia

15. Ross, Timothy J. Fuzzy Logic With Engineering Applications, 3rd ed. Wiley. 2010. ISBN 978-0-470-74376-8 pp 352-353, eq 10.28 - 10.35.

16. Larose, T. Larose, C. (2015). Data Mining and Predictive Analytics, Wiley Series on Methods and Applications in Data Mining, Wiley (2nd edition), Chapter 8

17. Abonyi, J., Feil, B. (2007), Cluster Analysis for Data Mining and System Identification, ISBN: 978-3-7643-7987-2, Birkhäuser Verlag AG.

7 Supplement

Supplement 1

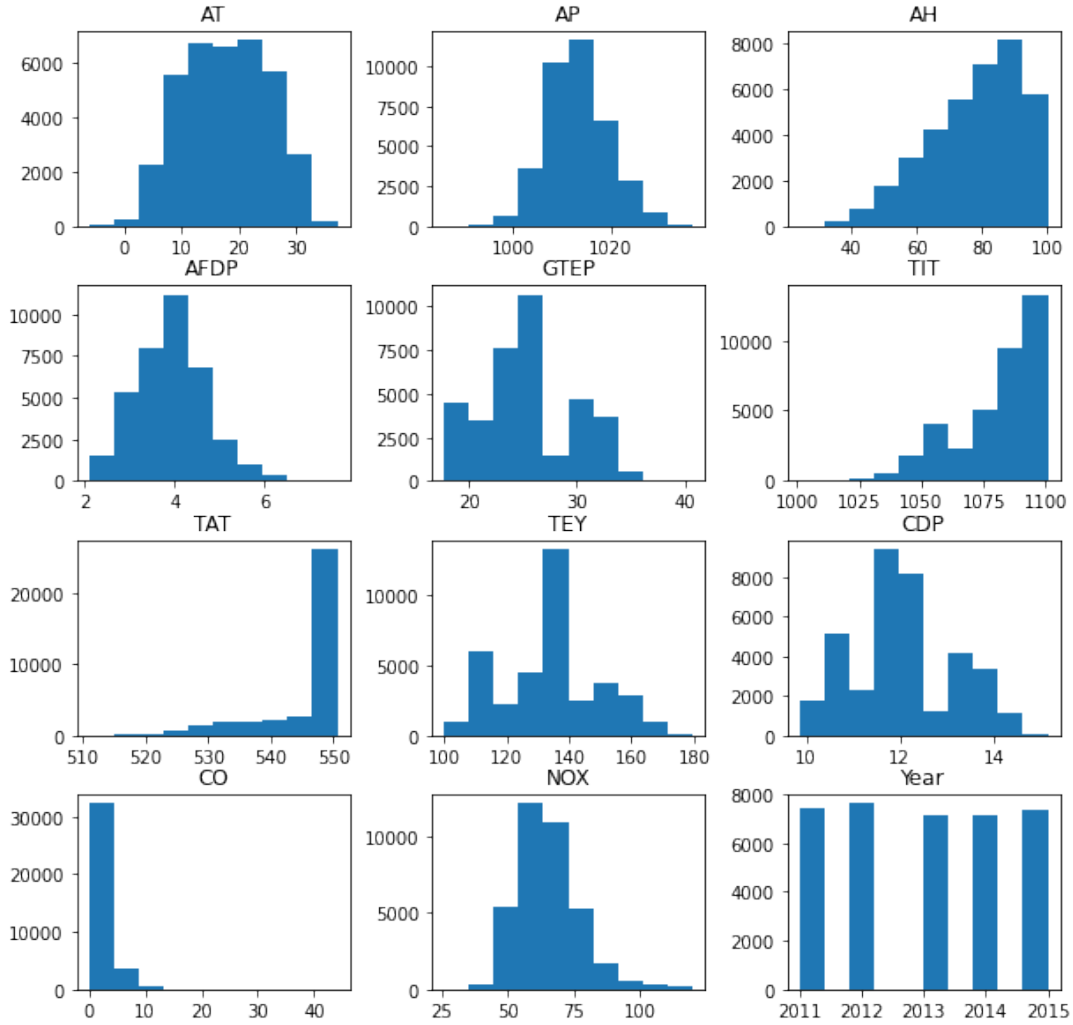


Figure 17: Histograms for all features

Supplement 2

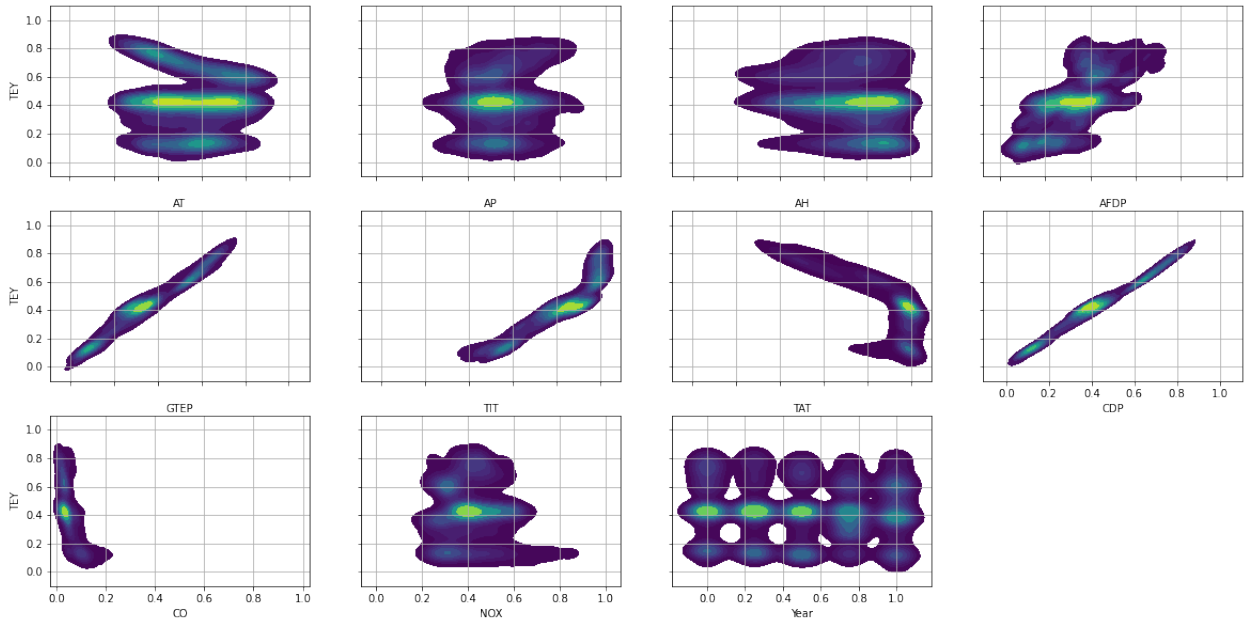


Figure 18: 2D Kernel Plot

Supplement 3

$$X_{std} = \frac{X - X_{mean}}{\sqrt{\sigma^2}}$$

Supplement 4

$$X_{range} = \frac{X - X_{mean}}{max(X) - min(X)}$$

Supplement 5

```
Sphericity test results:

SpherResults(spher=False, W=2.102034506407241e-05, chi2=395583.9033987725, dof=14, pval=0.0)
```

Figure 19: Sphericity Test

Supplement 6

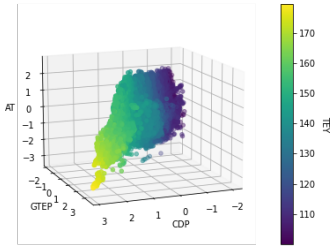


Figure 20: 3D Visualization of the StandardScaled Data

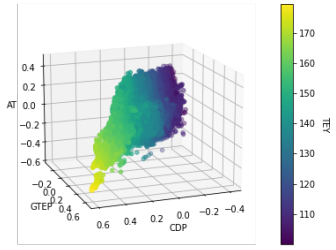


Figure 21: 3D Visualization of the RangeScaled Data

Supplement 7

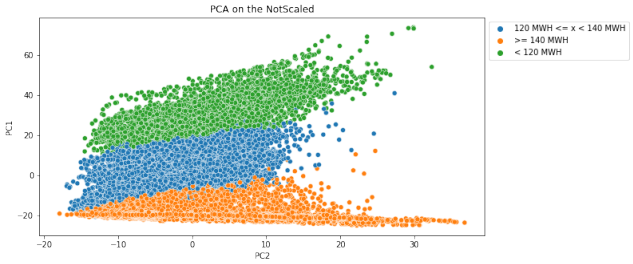


Figure 22: Principal Component Visualization on the Original Data

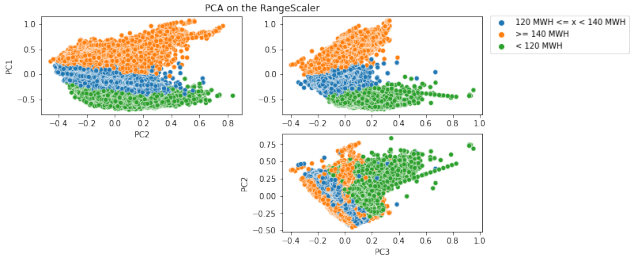


Figure 23: Principal Component Visualization on the RangeScaled Data