The Quality of Wine

Data Analysis and Mining

Rúben Carvalho & Sofia Begonha Morgado

Faculdade de Ciências e Tecnologia

Universidade Nova de Lisboa

June 27, 2022

Contents

1 Introduction

Wine is a significant component of the Portuguese economy and culture, and its excellence is praised across the world. Portugal shipped more than 3 million hectoliters of wine in 2020, putting it in tenth place globally in terms of wine exports [1].

The dataset "Red Wine Quality" is made available through the open-source website Kaggle. It is a result of a study of red and white portuguese wine characteristics [2]. It consists of 11 variables resulting from psicochemical tests, including acidity, alcohol percentage and density, among others; and 1 variable corresponding to the expert-assigned quality score based on sensory data.

Wine certification and quality assessment are critical components of wine quality improvement and stratification [2].

Numerous methodologies were used in this study to conduct an exploratory data analysis. After selecting two characteristics, a linear regression is implemented and its adequacy is determined. We run a Principal Component analysis and determine the best way to normalize/standardize our data. Finally, Anomalous Pattern Clustering Algorithm is implemented in order to find the best number of clusters and cluster centers for the implementation of the Fuzzy c-Means algorithm. Validation indexes are applied in order to evaluate the clustering.

2 Experimental study

2.1 Linear Regression

To perform a linear regression on a pair of variables, an independent and a dependent variable, one must suppose that a linear function illustrating their connection exists, plus a normally distributed error. The objective is to define the parameters of the function in such a way that it is feasible to make a forecast for each independent variable value [3] [4] [5].

We perform a linear regression in two previously selected variables, we validate the linear regression assumptions, test the quality of the regression and compute the confidence intervals for the values obtained.

2.1.1 Selection of features for implementation

A matrix composed of the scatter plots between all the features was computed, as is shown in Figure 25 in Appendix 1, and was utilized to observe the relationship between all the features in our dataset. In order to perform a linear regression, two features with a linear-like scatter plot were selected.

The features "Fixed Acidity" and "pH" were selected and a scatter plot is shown in Figure 1.

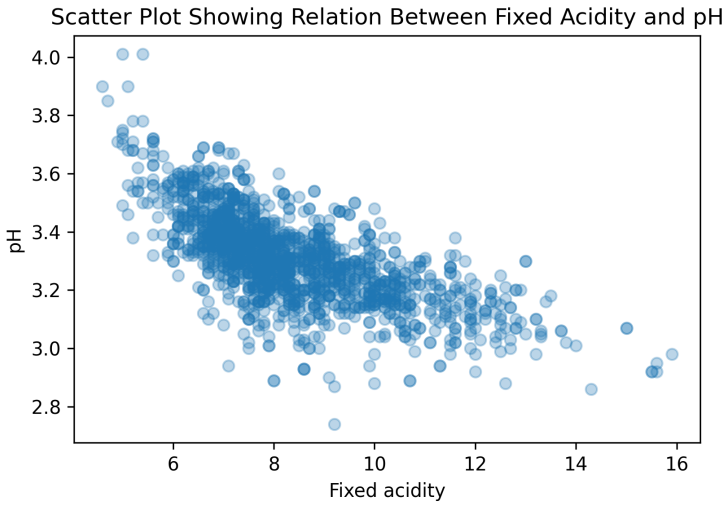

Figure 1: Scatter plot between Fixed Acidity and pH

2.1.2 Linear Regression without Removal of outliers

A linear regression was computed by determining the coefficients defining the line that would minimize the sum of squared errors (Figure 2).

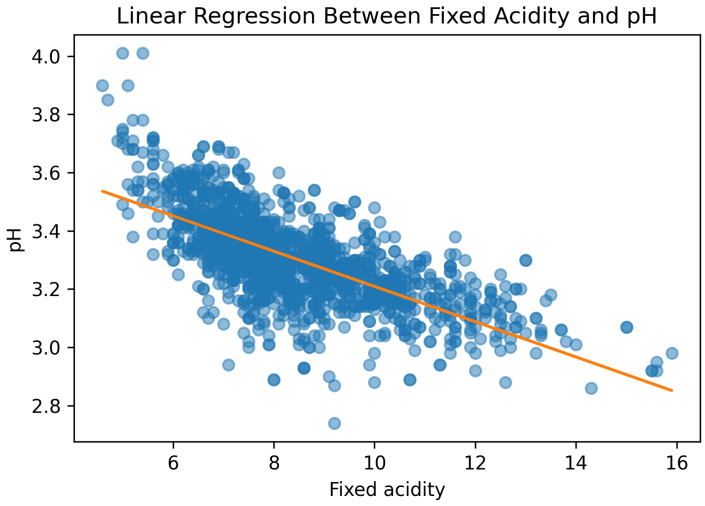

Figure 2: Linear Regression over Scatterplot between Fixed Acidity and pH

The population regression equation obtained for our model is expressed in equation 2, as follows:

$$y = -0.061x + 3.8150 \tag{1}$$

The parameter β_1 corresponds to the slope of our linear model. In fact, β_1 = -0.061 means that as fixed acidity increases by 1 unit, the pH will decrease 0.061 units. The parameter β_0 corresponds to the value of pH for which fixed acidity is 0. This value has no meaning as we cannot interpret a value outside the range of our dataset.

The errors, alternatively referred to as residuals, are the difference between the obtained value and the value existing in our sample. One of the assumptions underlying running a linear model is that these residuals are normally distributed. Hence, the value of the residuals for each data point was calculated and a QQ-plot was used to infere the normality of their distribution (Figure 3).

A QQ-plot compares two sets of quantiles in order to demonstrate their association. The x-axis of the left graphic of Figure 3 depicts the theoretical quantiles of a normal distribution, whereas the y-axis represents the distribution of our residuals. If they have a normal distribution, the plot will show a collection of points above the red line. As this scatter plot demonstrates, there is some skewness. There are fewer samples on the distribution's periphery on the left, whereas there are exceeding on the right, showing there is skewness to the higher values. This may be due to the presence of outliers.

On the right figure, we may observe the standardized residuals versus the fitted values plot. It is a scatter plot with the y axis representing residuals and the x axis representing fitted values (estimated responses). Non-linearity, uneven error variances, and outliers are all detected using

the plot. As we can observe, there is a slightly funnel pattern, which violates the constant variance assumption.

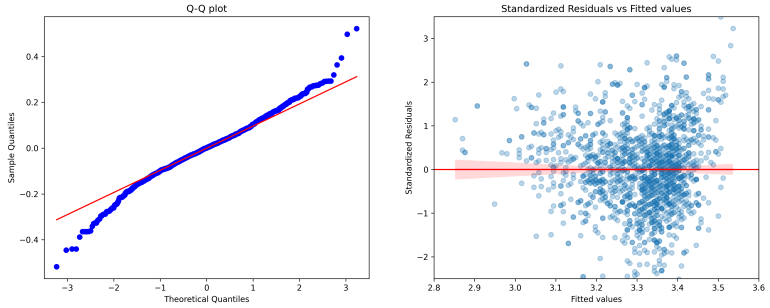

Figure 3: QQ_plot and Std residuals vs fitted values

2.1.3 Removal of outliers

Observing the previous plots, we realized that there were outliers in our dataset that should be removed. We may observe them in Figure 4.

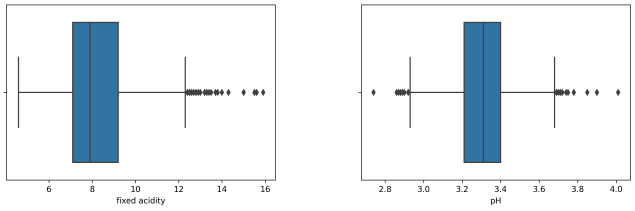

Figure 4: Boxplot for Fixed Acidity (left) and pH (right) before Removal of Outliers

The interquantile range was used to remove the outliers. We started with 1599 examples and, after we removed the outliers concerning the variable "fixed acidity" and "pH", we retained 1505 examples, removing 5.88% of the dataset.

The boxplots after the removal of our outliers are shown below, in Figure 5.

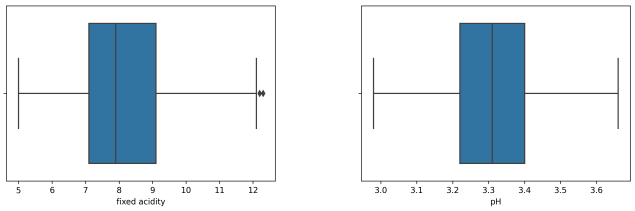

Figure 5: Boxplot for Fixed Acidity (left) and pH (right) after Removal of Outliers

Finally, after the removal of the outliers, a new regression was performed.

2.1.4 Linear Regression without outliers

Finally, after the removal of the outliers, a new regression was performed (Figure 6).

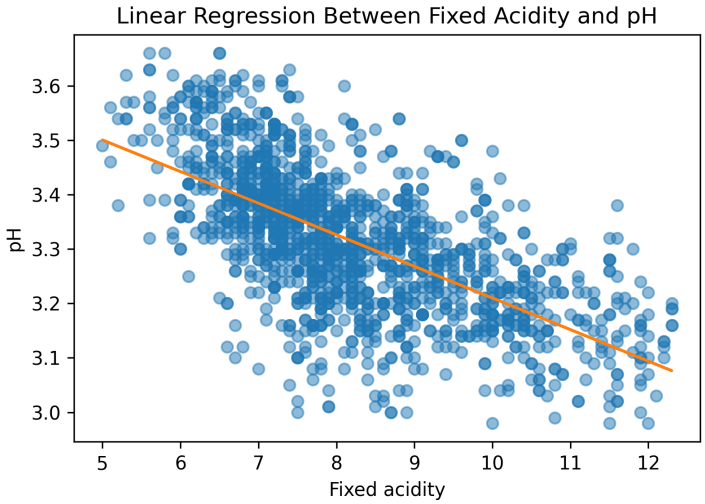

Figure 6: Linear Regression after the Removal of Outliers

The population regression equation obtained for our model is expressed in equation 2, as follows:

$$y = -0.0581x + 3.7909 \tag{2}$$

A new QQ-plot and a standardized residuals versus the fitted values plot were performed (Figure 7).

This QQ-plot shows a much better fit, with only some light skewness on the left side. Additionally, there is no visible pattern in the standardized residuals versus fitted values plot. The plot is not curved, indicating that linearity is maintained and the residuals have a mean of 0. Additionally, the residuals are normally distributed, as is possible to conclude due to their even dispersion. The constant variance assumption appear to be respected as well, given there is no pattern in the residuals' vertical dispersion.

With this plots, it is not possible to determine the independence of the values, but there is no reason to believe this assumption is violated taking into account our knowledge of the way the dataset was constructed.

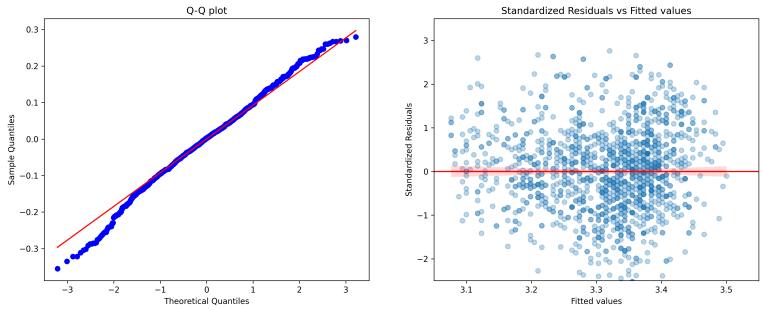

Figure 7: QQ_plot and Std residuals vs fitted values after the Removal of Outliers

2.1.5 Logarithmic transformation of the dependent feature

Some non-linear transformations may help increase the relationship between the variables and hence the performance of linear regressions. The optimal transformation is unique to each dataset and is determined by the variables selected.

The logarithmic transformation is one type of transformation that may be used to strengthen this association. In this work, we chose to apply the natural log to the dependent variable, the "pH". A new linear regression was generated using the altered features, as well as a QQ-plot to examine the residuals' normality distribution (Figures 8 and 9).

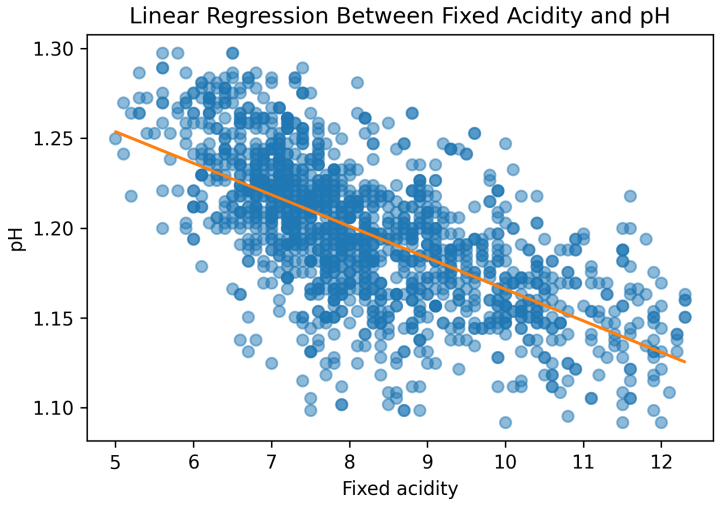

Figure 8: Linear Regression after the Removal of Outliers

As we can observe, this transformation does not seem to improve the fit of our data. Furthermore, observing the QQ-plot, it is possible to see that the residuals are even less close to a normal distribution.

As a result, the dataset without the logarithmic treatment will be utilized in the remainder of this project.

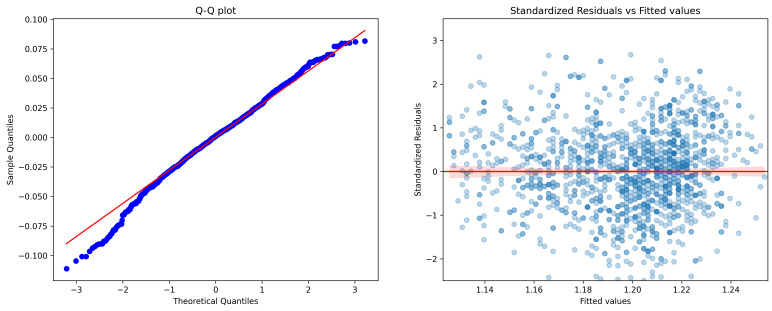

Figure 9: QQ_plot and Std residuals vs fitted values for the logarithmically transformed dataset

2.1.6 Goodness of fit

Other way of measuring the goodness of our linear regression is to calculate the coefficient of determination. The coefficient of determination, also referred to as R^2, represents the overall measure of the estimated regression equation error.

The value of the coefficient of determination ranges from 0 to 1, and, for a good linear regression we need a value close to 1 [8]. The value obtained for our linear regression was 0.42, showing that we were not able to make a good fit.

We also determined the coefficient of correlation, obtained as $\rho = \pm r^2$. In this case, since the slope of the regression is negative, the value obtained was -0.64.

2.1.7 Inference in Regression

Confidence intervals (CI) may be constructed around acquired values, such as the slope of our regression or even the coefficient of determination, in order to assess the linear correlation's significance. This is applicable to our dataset because we were able to confirm the linear regression's assumptions.

CI for the slope Firstly, we calculated the 95% confidence interval for the slope of the linear regression. The value obtained was from -0.0541 to -0.0620. This interval excludes the value 0, implying that a linear connection between the fixed acidity and the pH exists with 95% confidence.

CI for the correlation coefficient We have estimated 95% CI for our correlation coefficient of -0.6866 to -0.6095. Since both endpoints of the CI are negative, we can say there is a negative correlation between these two variables. Also, we can say they are mildly negatively correlated, as this values are between -0.7 and -0.33.

CI for the mean of the y-variable at a fixed x value The value obtained for the CI of the mean of y-variable at a fixed $x = 8$ value was 3.3262 to 3.3264.

CI for a randomly chosen value of the y-variable at a fixed x value The result obtained for the CI of a randomly chosen value of the y variable at a fixed $x = 8$ value was 3.3223 to 3.3302. As may be seen, the variability associated with the mean of y is smaller than the variability of a random individual observation of the same variable, as it is simpler to anticipate the mean of observations than a random single observation.

2.1.8 Conclusions on the linear regression

To summarize, we were able to do a linear regression that respected the assumptions used to design it. Additionally, the coefficient of determination was calculated, which revealed a modest linear correlation between our variables. Nonetheless, distinct approaches must always be used to evaluate a model's performance. As a result, inference techniques were applied. This resulted in the discovery of a weakly, nevertheless present, linear negative association between the properties fixed acidity and pH. Also, a data non-linear transformation was applied to the dataset, which included a logarithmic transformation of the dependent variable. Nevertheless, this showed no improvement in respect to the assumptions and correlation of determination of the regression.

2.2 Principal Component Analysis

The aim of dimensionality reduction is to reduce the dimension of a high-dimensional data set, preserving the essential properties of the full data matrix [5]. This can be performed because: some features on the data set may be irrelevant for our data mining and/or analysis; reduce the dimensionality of the data set can help us to visualize high dimensional data; the "intrinsic" dimensionality may be smaller than the number of features; and it helps us to deal with the curse of dimensionality [8].

One of those methods is the Principal Component Analysis (PCA). This is a multivariate technique which entails a sequence of data transformations in order to acquire fresh uncorrelated data variables. The principal objective of this technique is to provide dimensionality reduction, although preserving the data variability [6] [7].

The direction with the largest projected variance is called the first principal component. The orthogonal direction that captures the second largest projected variance is called the second principal component, and so on. One important note is that, the direction that maximizes the variance is also the one that minimizes the mean squared error [5].

Regarding our data set, and taking in account that it is related to wine quality, we decided to choose five features (fixed acidity, residual sugar, density, pH and alcohol), after a brief study on the main characteristics and measures related to wine production, and analyse how the would relate to the six targets from the data set.

2.2.1 Visualization in 2D/3D PC plane and types of normalization

When we want to perform an analysis of a data set, most of the times we need to normalize the values contained on it, specially on those case where the values do not correspond to the same measure or when they are vastly different in scale [5]. There are several techniques for normalization, but for this PCA analysis we utilized only two: the Range Normalization (also called Min-Max Normalization) and Standard Score Normalization (also called Z-Score Standardization).

The range normalization works by seeing how much greater the field value (X) is than the minimum value min(X), and scaling this difference by the range [9],

$$Range = \frac{X - min(X)}{range(X)} = \frac{X - min(X)}{max(X) - min(X)} \tag{3}$$

The z-score, otherwise, uses the difference between the field value and the field mean value, and scaling this difference by the standard deviation (SD) of the field values [9],

$$Z - Score = \frac{X - mean(X)}{SD(X)} \tag{4}$$

Before we show the results of PCA for each of normalization technique, it is useful to see how variance is affected for our features, before and after data normalization. As we can see, the

variation is much more stabilized after each data normalization:

Table 1: Variation before and after Range and Z-Score Normalization.

Feature	Before	Range	Z-Score
Fixed Acidity	3.030	0.024	1
Residual Sugar	1.987	0.009	1
Density	0.000	0.019	1
pH	0.024	0.015	1
Alcohol	1.135	0.027	1

The 2D and 3D PC planes for each type of normalization are shown below:

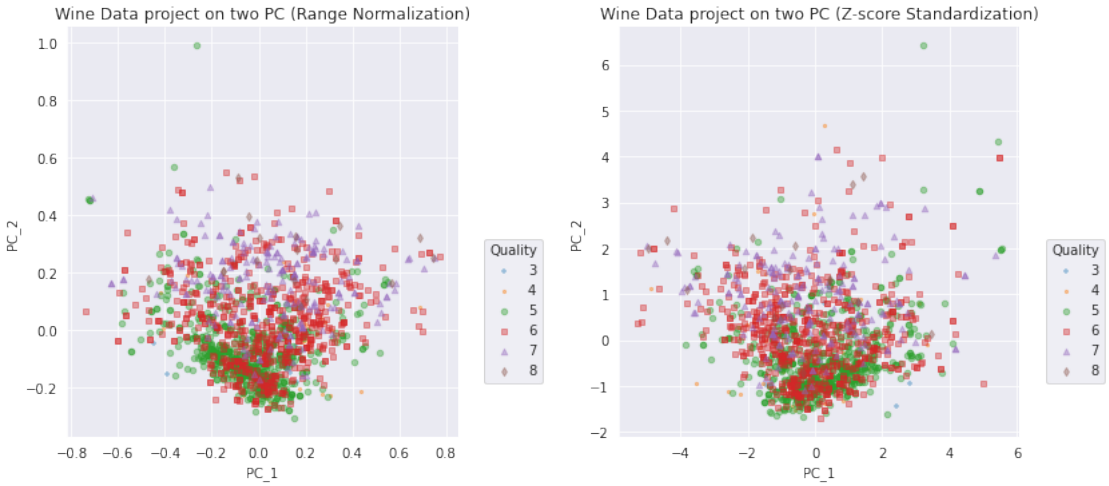

Figure 10: 2D planes for Range and Z-Score Normalization

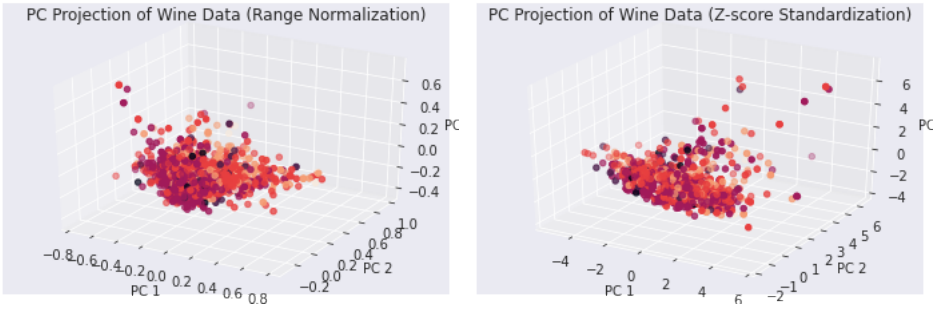

Figure 11: 3D planes for Range and Z-Score Normalization

Looking for the two pairs of planes, we can conclude that they are slightly different to each other. This can be explained by understanding the PCA linear combinations. To implement this method calculate the eigenvalues and eigenvectors of the data set covariance matrix. The eigenvectors correspond to the PC and the eigenvalues to the proportion of the retained variance. The number of PC retained is determined in a variety of ways, including the variance retained and the dimensionality reduction.

The p linear combinations $y = (y_1, y_2, ..., y_p)$ of the p original variables $x = (x_1, x_2, ..., x_p)$ are the Principal Components (sample):

$$
\begin{aligned}
y_1 &= a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p \\
y_2 &= a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p \\
&... \\
y_p &= a_{p1}x_1 + a_{p2}x_2 + ... + a_{pp}x_p
\end{aligned}
$$

which can be interpreted mathematically by $Y = \mathbf{A}X$. The coefficients $a_{kj}(k, j = 1, ..., p)$, by row, are the values of matrix $\mathbf{A}$, and they are called the eigenvectors of the variance-covariance matrix S_x of the original data. It is common to call these values PC loadings (or weights).

For our data set, the Principal Components calculated can be described as the following linear combinations of X_1, X_2, X_3, X_4 and X_5:

$$
\begin{aligned}
PC1_{range} &= -0.584x_1 - 0.099x_2 - 0.563x_3 + 0.386x_4 + 0.428x_5 \\
PC1_{z-score} &= 0.556x_1 + 0.217x_2 + 0.565x_3 - 0.484x_4 - 0.300x_5
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
PC2_{range} &= 0.497x_1 + 0.080x_2 - 0.068x_3 - 0.243x_4 + 0.826x_4 \\
PC2_{z-score} &= 0.227x_1 + 0.559x_2 - 0.106x_3 - 0.098x_4 + 0.785x_5
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
PC3_{range} &= -0.102x_1 + 0.654x_2 + 0.498x_3 + 0.526x_4 + 0.193x_5 \\
PC3_{z-score} &= -0.348x_1 + 0.698x_2 + 0.307x_3 + 0.457x_4 - 0.298x_5
\end{aligned} \tag{7}
$$

where X_1 corresponds to fixed acidity, X_2 to residuals sugar, X_3 to density, X_4 to pH and X_5 to alcohol. For PC1 (an example), as we can see, there is similar weights given for the values of fixed acidity and density.

2.2.2 PCA and SVD visualization. Choosing between Range Normalization and Z-Score Standardization

There is another method for dimensionality reduction called Singular Value Decomposition (SVD). The SVD of a matrix D is the factorization of D into the product of three matrices,

$$D = U\Lambda R^T \tag{8}$$

where U is a orthogonal nn matrix, R is an orthogonal dd matrix, and Λ is an nd diagonal matrix. The columns of U are called the left singular vectors, and the columns of R (or rows of R^T) are called the right singular vectors [5]. This technique allows us to decompose a rectangular matrix, unlike eigen decomposition where the matrix we want to decompose has to be a square matrix.

We now present the results of SVD for our data set, for the two types of normalization already shown for PCA technique, Range Normalization and Z-Score Standardization:

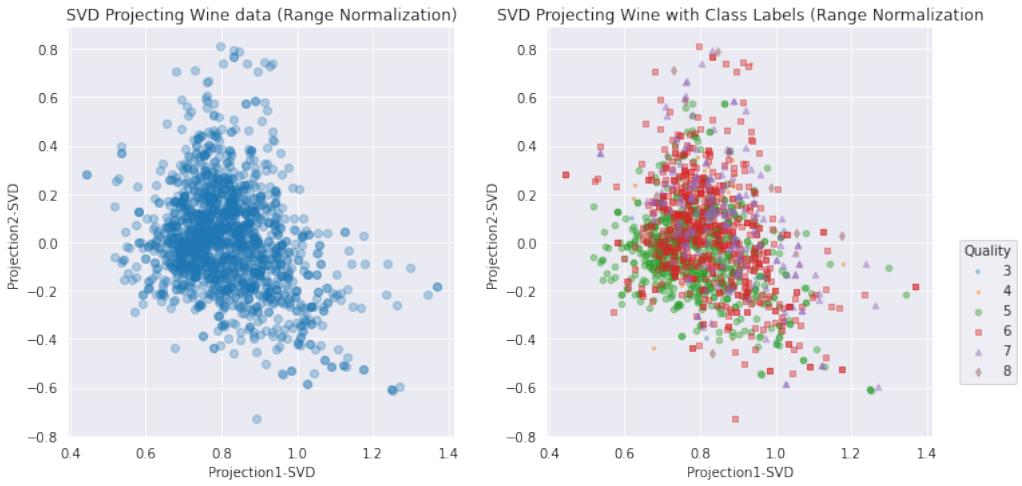

Figure 12: SVD for Range Normalization

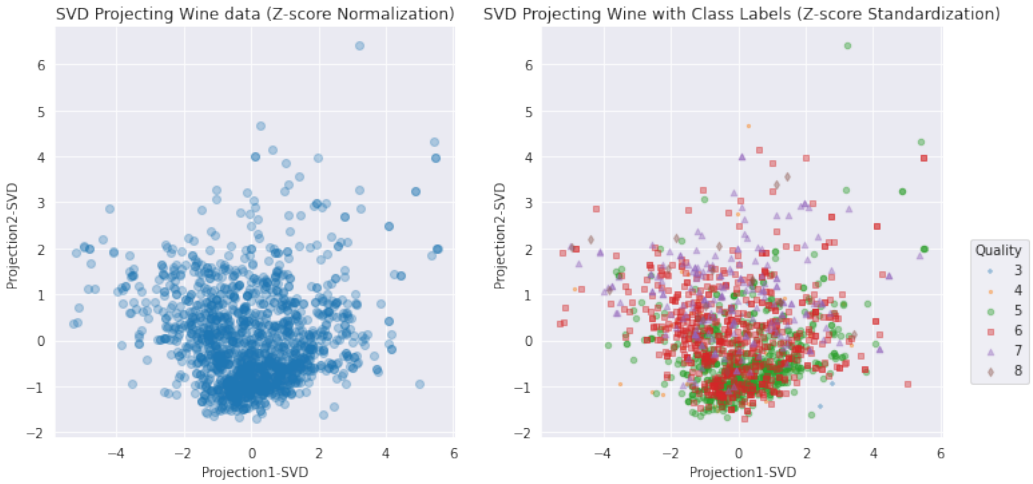

Figure 13: SVD for Z-Score Standardization

After performing SVD of our data and calculating the explained variance of the various PC, we can conclude that using PCA, with range normalization, the first three PC explain around 90.7% of variance and the groups are better separated.

Table 2: Explained variance for PCA and SVD, with Range Normalization and Z-Score Standardization

Technique	PC1	PC1+PC2	PC1+PC2+PC3
SVD Range Norm.	0.15046796	0.63680366	0.88819509
SVD Z-Score Stand.	0.47008132	0.68075632	0.88574702
PCA Range Norm.	0.50425178	0.77818952	0.9066675
PCA Z-Score Stand.	0.47008132	0.68075632	0.88574702

2.2.3 Define different classes for our features

The data set that we worked with has six target groups, related to the *quality* of the wine (from number three to eight). To respond to this question, we decided to divide them equally, into three groups. We inserted a new column with the name $wine_quality$, with the groups *low* (related to *quality* values three and four), *medium* (related to *quality* values five and six) and *high* (related to *quality* values seven and eight).

After this selection we decided to perform a PCA 2D and 3D planes, with the data normalized by range, in order to examine and assess if we can understand better the distribution of the data. The planes are shown bellow:

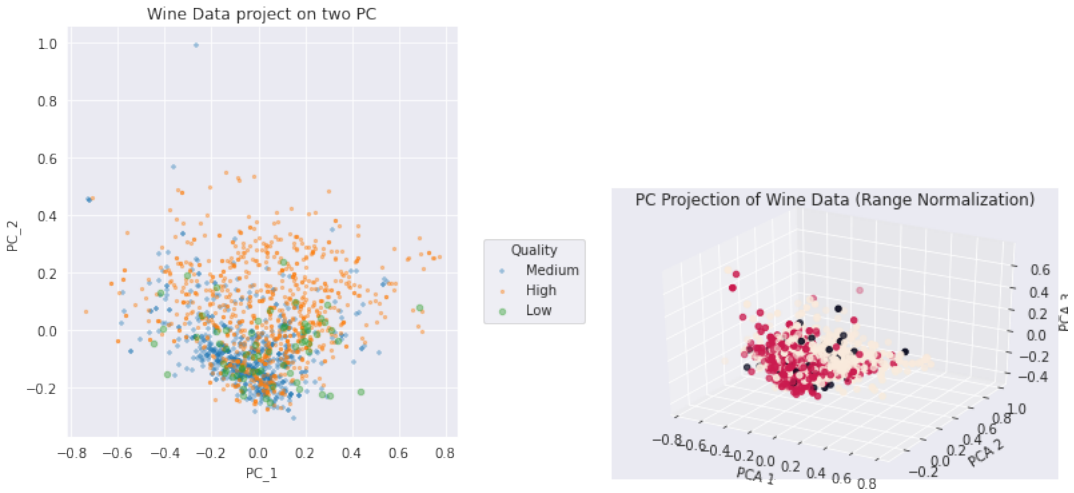

Figure 14: 2D and 3D planes for groups groups *low*, *medium* and *high* (Range Normalization)

With this visualization and partition, we can better assess and understand the data distribution regarding the *quality* of the wine.

2.2.4 Quality and suitability of PCA analysis

The main goal of PCA is to replace the original p variables by a number of new variables q ¡¡ p, retaining as much statistical information as possible. It is then crucial to know how small q can be without the loss of "relevant" information. There are several criteria for deciding on the number of components to retain [10]. For our study, we chose three, as follow:

Proportion of explained variance: This criteria retain a number q of components that can explain 90% cumulative percentage of data set variability (usually). In our analysis, the first three PC explain 90,67% of total variability (it is also explain by the values on Table 2). To graphically show this, we present the followings figures:

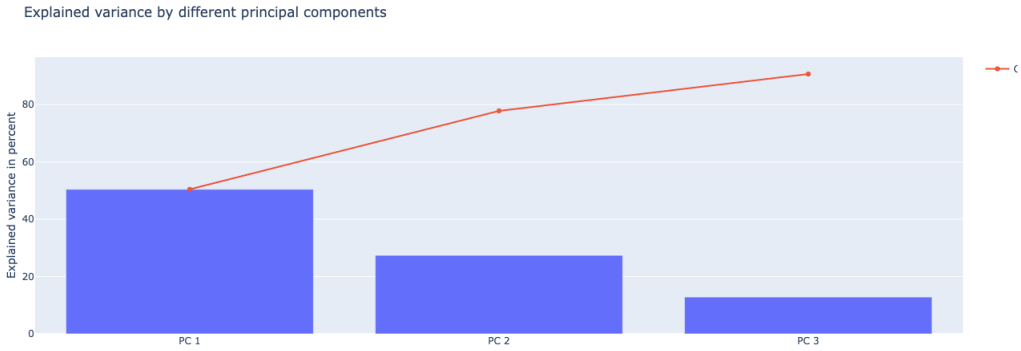

Figure 15: Explained variance by different PC

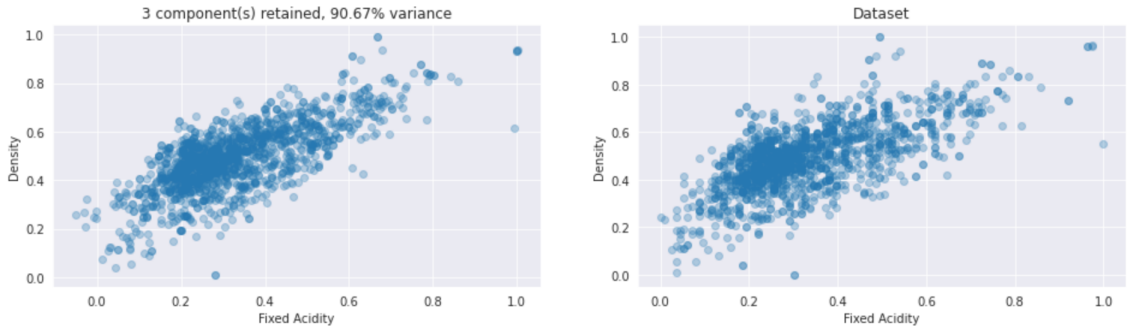

Figure 16: Explained variance by first three PC

Scree Plot: A scree plot relates the eigenvalues of PCs to the number of PCs that can be retained. Scree plots are useful for finding an upper bound (maximum) for the number of components that should be retained. The number of PCs is dictated by the point where the drawn line sharply decreases in slope, becoming horizontal [10]. As we can see, the figure does not show a sharp elbow, but we can PC3 as a breaking point between the steep slope and the flattening of the line. And that's why we should retain the three first PCs.

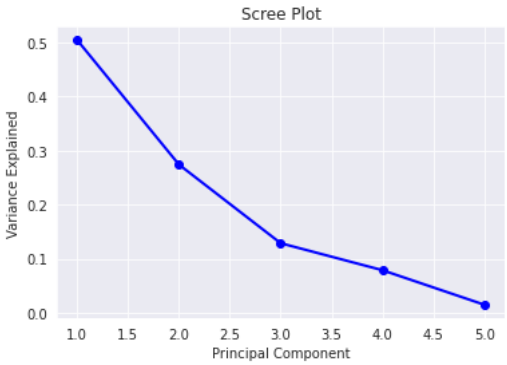

Figure 17: Scree Plot

Kaiser Criterion: According to this criterion, in the case of standardised variables, PCs with an eigenvalue greater than 1 are retained. With standardized variables, a component with eigenvalue less than 1 indicates that it does not retain even the information (variability) equivalent to one of the original variables (with variance 1, for being standardized) [10]. For our study, we would not retain any of the component, because all of them are less than 1.

2.3 Clustering

When initializing C-means clustering, two primary issues must be addressed: the number of clusters and the initial prototypes, which may be determined using different methods.

To begin, we shall discuss the Elbow Method and its application to a plot of the JM value for a variable number of clusters.

JM is the objetive function for the python built-in Fuzzy C-mean algorithm. This function monotonously decreases with increasing number of clusters, as shown in Figure 18. For this reason, we must not use the Elbow method applyed to this function to determine the optimal number of clusters.

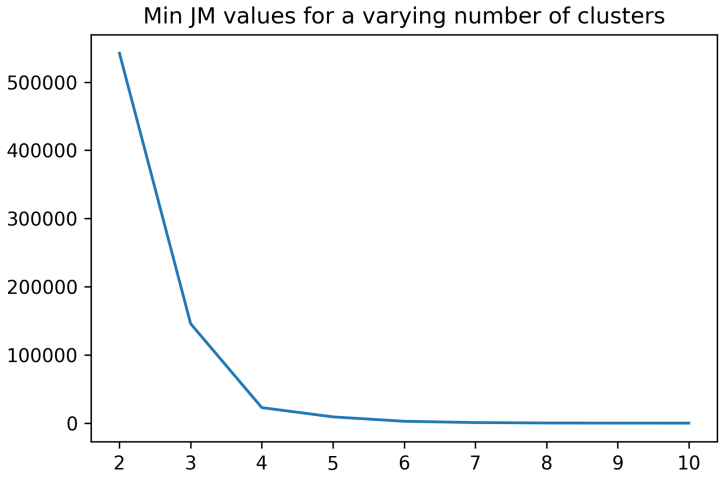

Figure 18: Minimum JM values for a varying number of clusters

On the other hand, other measure we must consider is the the fuzzy partition coefficient(FPC). The FPC is defined on a scale of 0 to 1 and the objetive is to maximize it. It is a metric that tells us how well a given model describes our data. (Figure 19). As we can see, FPC is maximum for 2 clusters. However, this value alone is insufficient to determine the optimal number of clusters.

It is typical to use a random sample of prototypes, since this ensures the iterative algorithm's convergence. Likewise, the Iterative Anomalous Pattern method is investigated as a possible initial configuration scheme for the FCM that also serves as a measure of the number of clusters in the dataset [11], and may as well be used to select better initialization centroids.

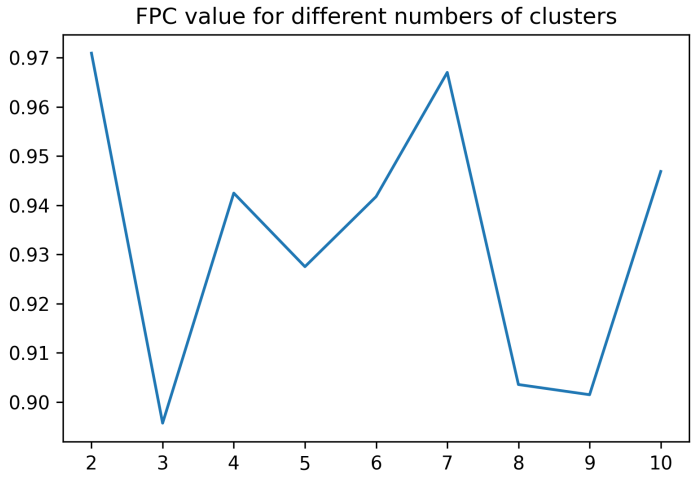

Figure 19: FPC values for a varying number of clusters

2.3.1 Anomalous Pattern Clustering Algorithm

The Iterative Anomalous Pattern (IAP) algorithm extracts clusters progressively from previously standardized data. It operates by shifting the origin of the data to the grand mean, which is used as a reference point for selecting the seed, that corresponds to the farthest location from this reference. By aggregating the points that are closer to the seed than to the reference point, a cluster is formed. The reference point is then moved to the center of the remaining data points and the process is repeated until a stop condition is met, which may be that all points have been assigned to a cluster, the contribution of the new clusters to the data scatter is too small, or a pre-defined maximum number of clusters has been reached [11].

To implement the IAP algorithm, the dataset was normalized using the performance-optimized approach described in the PCA chapter. The algorithm was computed, and the clusters were created. As a consequence, 24 clusters were generated, each with the following number of elements:

$$[361, 534, 237, 146, 7, 195, 7, 40, 28, 1, 9, 6, 2, 4, 2, 1, 4, 5, 2, 1, 1, 1, 1, 4] \tag{9}$$

The algorithm stopped when all points in the dataset where assigned to a cluster. Additionally, a measure of data scatter (dD) was computed to determine the optimal number of clusters. Five clusters had a dD of more than one (1.87 to 17.03), whereas the other clusters all had a value less than one, indicating that they contributed less to the data scatter (Figure 20).

Additionally, 5 is the number of labels in the dataset, as we can verify in Figure 21. We can also see there is an imbalance in the number of points assigned to each class.

However, we must keep in mind that this is an empirical decision, and determining the optimal number of clusters is difficult. For example, when considering the FPC plot results (Figure 19), the best result would be two clusters. When we consider the value of data scatter, we see that

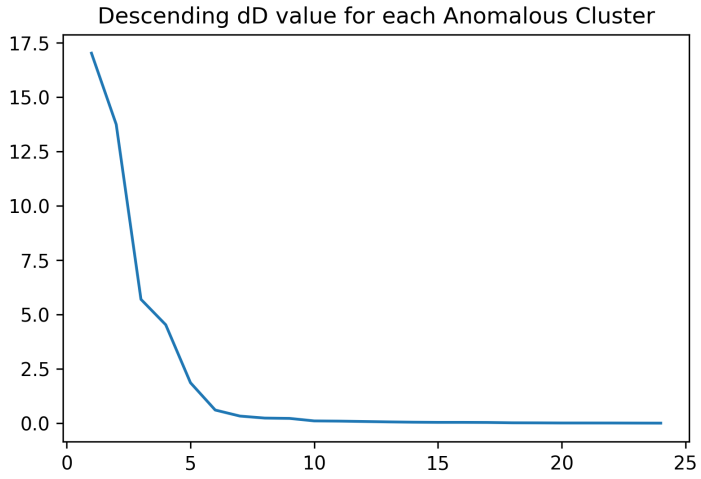

Figure 20: Sorted data scatter values for each cluster obtained by the Anomalous Pattern Clustering Algorithm

most clusters do not contribute to it (Figure 20); and when we consider the data labels, we should select 5 clusters.

For these reasons, we decided not to limit the number of clusters at this time, and instead used the Anomalous Pattern Clustering Algorithm to compute only the initial cluster centers for a range of 2 to 10 centers. The prototype coordinates for these clusters were saved to be used as the initialization of the Fuzzy C-Means algorithm. As a result, we will be able to compute the validation indexes for different numbers of clusters and make a more informed decision.

2.3.2 Fuzzy C-Means Algorithm

Fuzzy C-means (FCM) is an unsupervised clustering approach that is applicable to a broad variety of feature analysis, clustering, and classifier creation challenges [12].

In FCM, a data sample can be assigned to numerous clusters simultaneously. The membership value indicates the degree of resemblance, based on its Euclidean distance to the cluster center. Membership values range from 0 to 1, with the stronger the resemblance, the higher the membership value.

At the conclusion of the clustering procedure, defuzzification is used to determine the group of each sample. FCM is a repetitive algorithm, and the solution is obtained by changing the cluster center and membership value in a repetitive manner [13].

After normalization, the fuzzy C-means algorithm was applied to our dataset, and the centroids' initial coordinates were obtained from the Anomalous Pattern Clustering Algorithm.

We used the FCM for selecting a variety of number of centroids ranging from 2 to 10.

We examined the quality of the obtained clusters. More than one method should always be

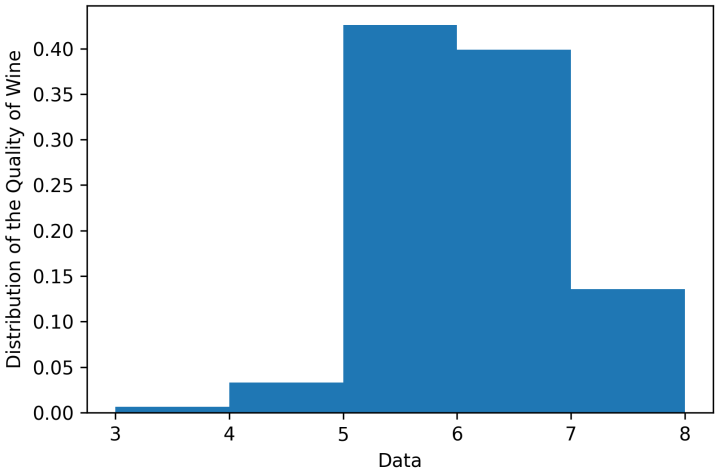

Figure 21: Distribution of Original Classes in Our Dataset

used, and a combination of many methods is usually a better method for determining the number of clusters.

Firstly, we applied the Xie Beni Index. It is one of the most famous validation indexes and it is associated with the proximity of data in one cluster and the distance between cluster centers. The optimal number of clusters is indicated by the smallest value of this index [14]. As we may verify in Figure 22, this corresponds to 2 clusters.

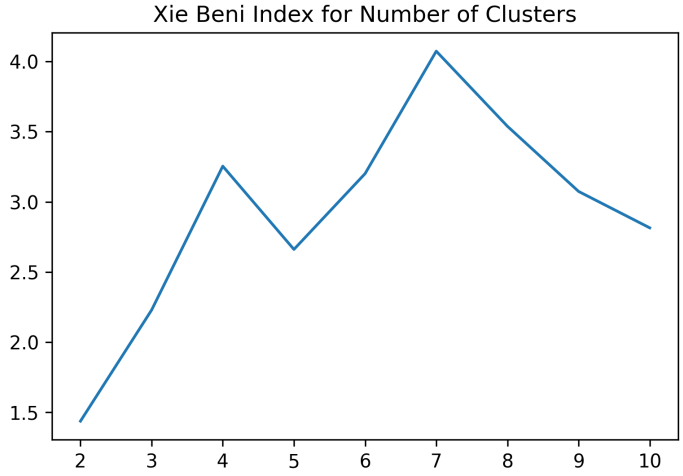

Figure 22: Xie Beni Index for Different Number of Clusters

On the other hand, we computed the partition coefficient index. For this index, the fuzzy partition matrix is used to calculate the fuzzy degree of final divided clusters; the higher the value, the better the partition result [15]. As illustrated in Figure 23, this value decreased with

the increasing number of clusters, being optimal for 2 clusters.

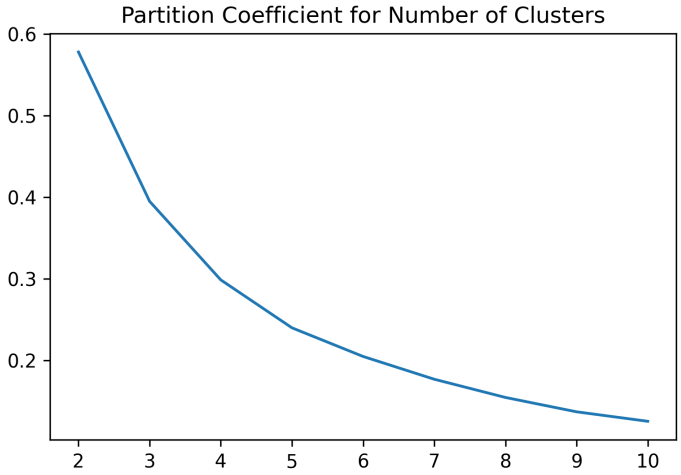

Figure 23: Particion Coefficient Index for Different Number of Clusters

Taking into account all of the values obtained, including the FPC, data scatter, and the two validation indexes computed, we concluded that, despite the presence of five classes, the best number of clusters for our dataset is two. This could be due to class continuity (resulting from the fact that is corresponds to a scale, but also due to the subjectivity is atributing a class), but it could also be due to an imbalance in the number of points in each class.

Figure 24 shows the results for the clusters obtained by the Fuzzy C-means algorithm after defuzzification, after the selection of two centroids.

The plot was created as a means of representation, displaying the data points using the first two principal components for all variables in our dataset as axis (PC1 corresponding to Dim1, and PC2 corresponding to Dim2).

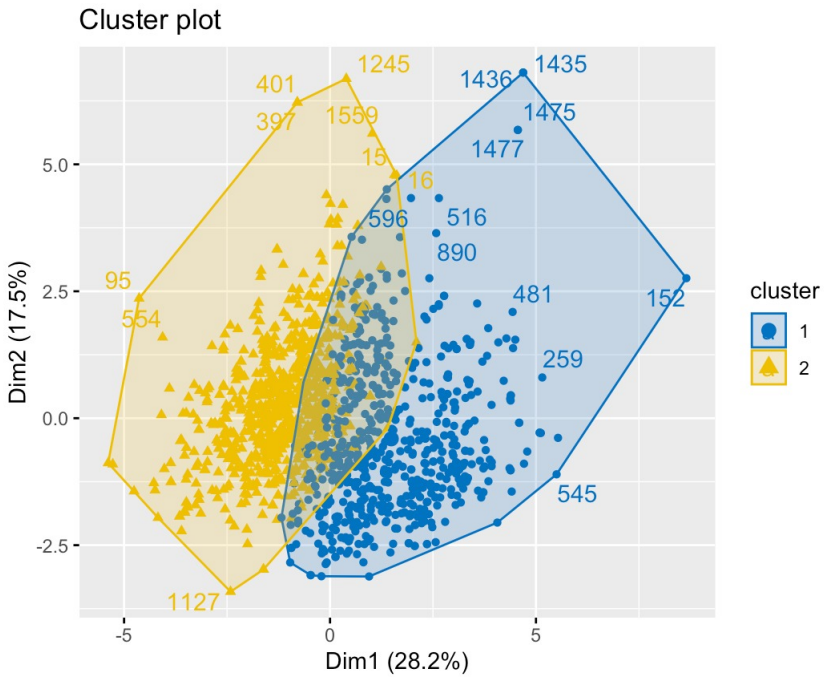

Figure 24: Cluster Results by applying Fuzzy C-means with 2 Centroids

3 Conclusion

To conclude, we performed a linear regression on two selected variables from our dataset, respecting the necessary assumptions. We evaluated the performance of the model with various tehniques and concluded a weakly negative linear association. Also, we concluded that the logarithmic transformation did not help improve the linear regression.

With the PCA, we assessed which technique, PCA or SVD, was the best to perform a better dimensionality redution of the data set. We also conclude that for our data set the range normalization was a the best standardization method, comparing with the z-score. For the features that we chose to study, the number of PC that we should retain was three.

Concerning the application of Fuzzy C-Means, various methods were used to determine the best number of clusters and the Anomalous Pattern Algorithm was used to determine the best initialization centroids.

4 Appendix

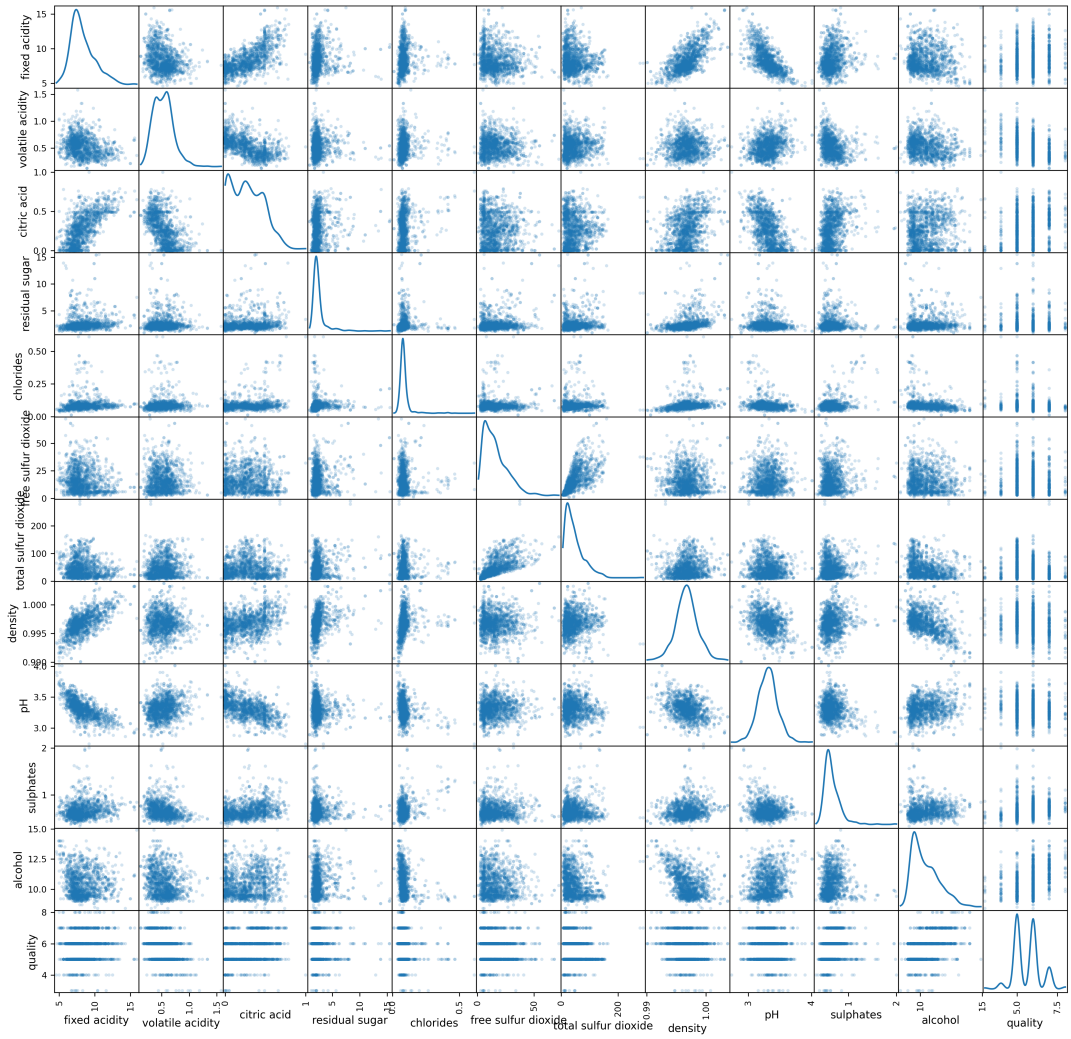

Figure 25: Scatter Plot Matrix for All Features

References

[1] https://www.statista.com/statistics/240649/top-wine-exporting-countries-since-2007/

Last accessed April.2022

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier (2009)

[3] Marsland, S. Machine Learning: An Algorithmic Perspective. First edition. Chapman Hall/CRC (2009)

[4] Ringnér, M. What is principal component analysis? Nature (2008)

[5] Zaki M., Meira Jr, W. Data Mining and Machine Learning: Fundamental Concepts and Algorithms. Second Edition. Cambridge University Press (2020)

[6] Hongyu, K., Sandanielo, V., Junior, G. Análise de Componentes Principais: resumo teórico, aplicação e interpretação. Engineering and Science (2015)

[7] Ringnér, M. What is principal component analysis? Nature (2008)

[8] Alpaydin E. Introduction to Machine Learning Second Edition. The MIT Press (2010)

[9] Larose, D. T., Larose, C. D. Data Mining and Predictive Analytics. Second Edition. Wiley (2015)

[10] Bispo, R., Marques, F. Estatística Multivariada. Departamento de Matemática. Faculdade de Ciências e Tecnologia. Universidade Nova de Lisboa (2021)

[11] Gama J., Costa v., Jorge A. et al. Discover Science: 12th International Conference. Porto (2009)

[12] Ghosh S., Dubey, S. Comparative Analysis of K-Means and Fuzzy CMeans Algorithms. International Journal of Advanced Computer Science and Applications. (2013)

[13] Choudhry, M., Kapoor, R. Performance Analysis of Fuzzy C-Means Clustering Methods for MRI Image Segmentation. Twelfth International Multi-Conference on Information Processing (2016)

[14] Mota, V. C., Damasceno, F. A., Soares, E. A., et al. Fuzzy clustering methods applied to the evaluation of compost bedded pack barns.IEEE International Conference on Fuzzy Systems. (2017)

[15] https://www.hindawi.com/journals/jece/2019/2719617/tab1/ (Last accessed April 2022)