PCA Analysis Report

Introduction

This report presents the results of a Principal Component Analysis (PCA) performed on a dataset representing red wine quality. The dataset was loaded, visualized, scaled, and subjected to PCA to uncover underlying patterns and relationships among the variables.

Step 1: Data Visualization

The initial step involved visualizing the dataset using boxplots to gain insights into the distribution of variables.

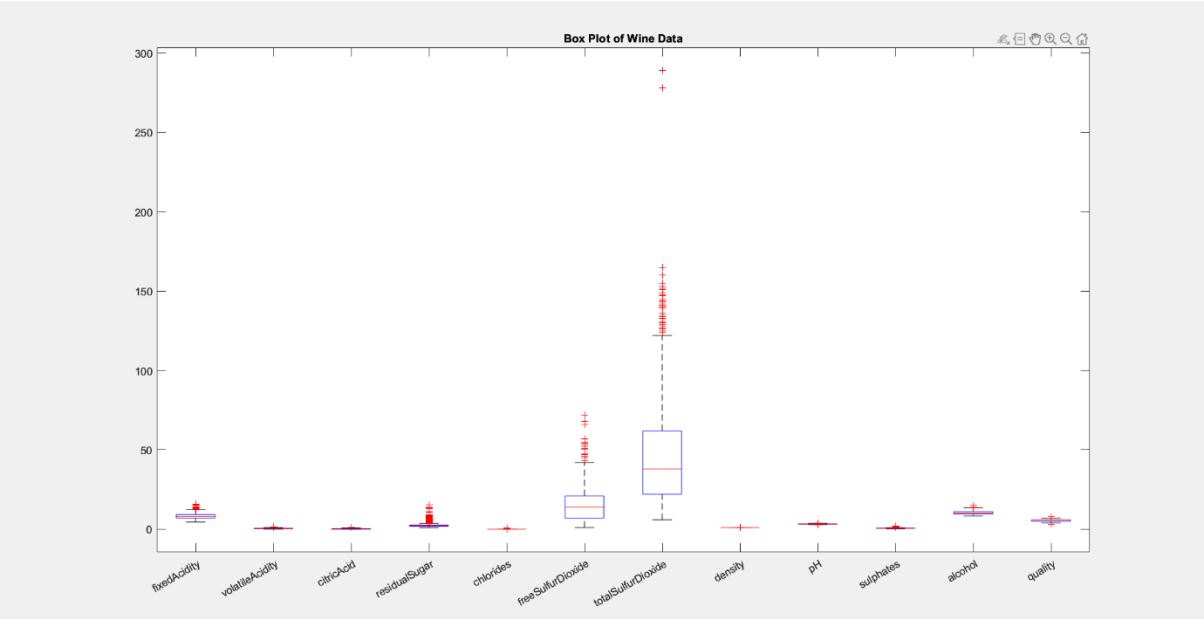

Figure 1: Box Plot of Wine Data

Figure 1 displays boxplots for each variable in the dataset. These boxplots reveal the distribution of data, including median, quartiles, and potential outliers. It seems like the variables are distributed over a large area.

Step 2: Data Scaling and Centering

To prepare the data for PCA, it was scaled and centered using the Z-score method.

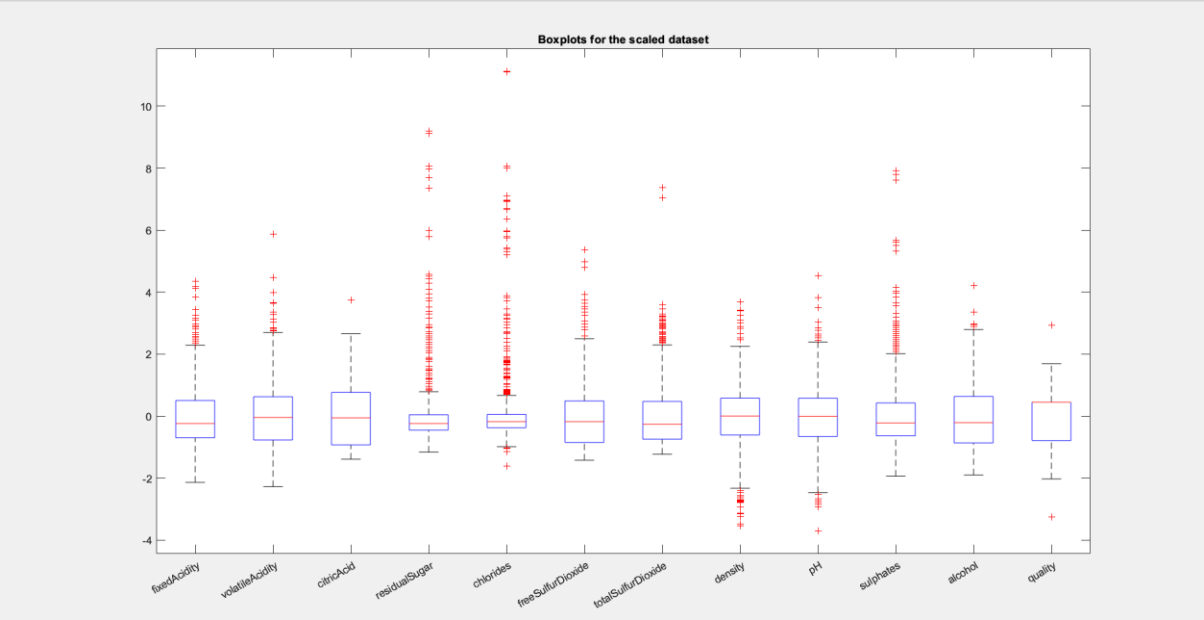

Figure 2: Boxplots for Scaled Dataset

Figure 2 presents boxplots for the scaled and centered dataset. Scaling ensures that variables have comparable ranges and units, facilitating meaningful PCA results.

Step 3: Principal Component Analysis

PCA was applied to the scaled dataset to identify principal components that capture the most significant variance in the data.

Step 4: Explained Variance

The explained variance for each principal component is shown below:

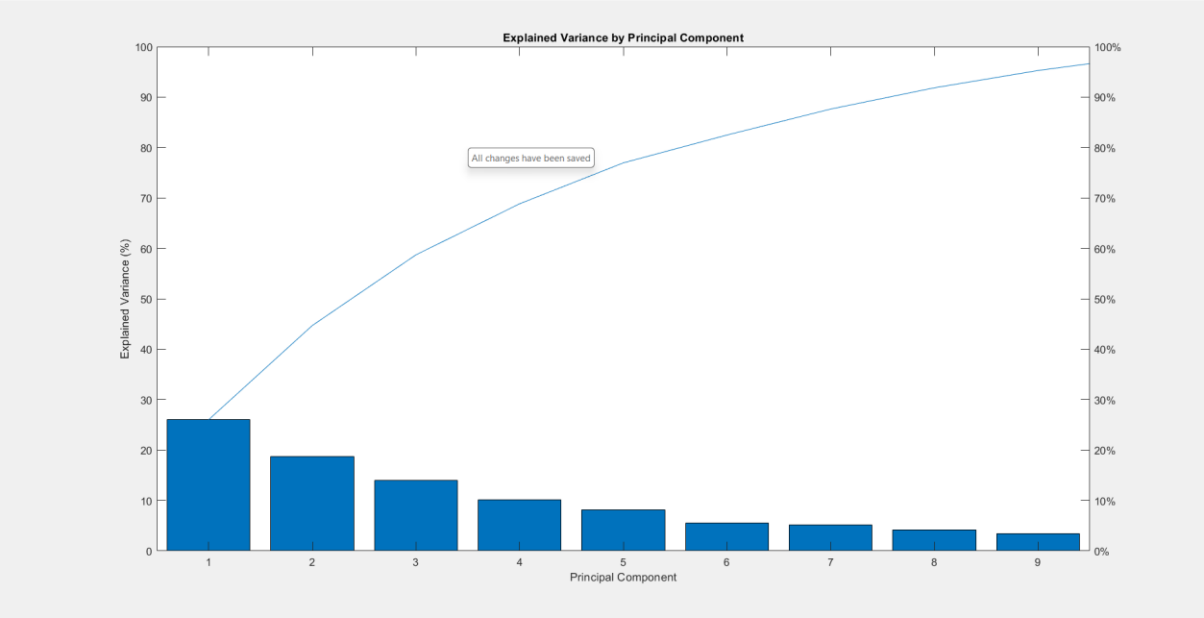

Figure 3: Explained Variance by Principal Component

Figure 3 illustrates the cumulative explained variance by principal component. The first few components capture the majority of the data's variance, making them the most informative. 2nd and 3rd component also share fairly substantial amount of variance in the data which makes them important as well.

Step 5: Biplot of First Two Principal Components

A biplot was created to visualize how variables load onto the first two principal components.

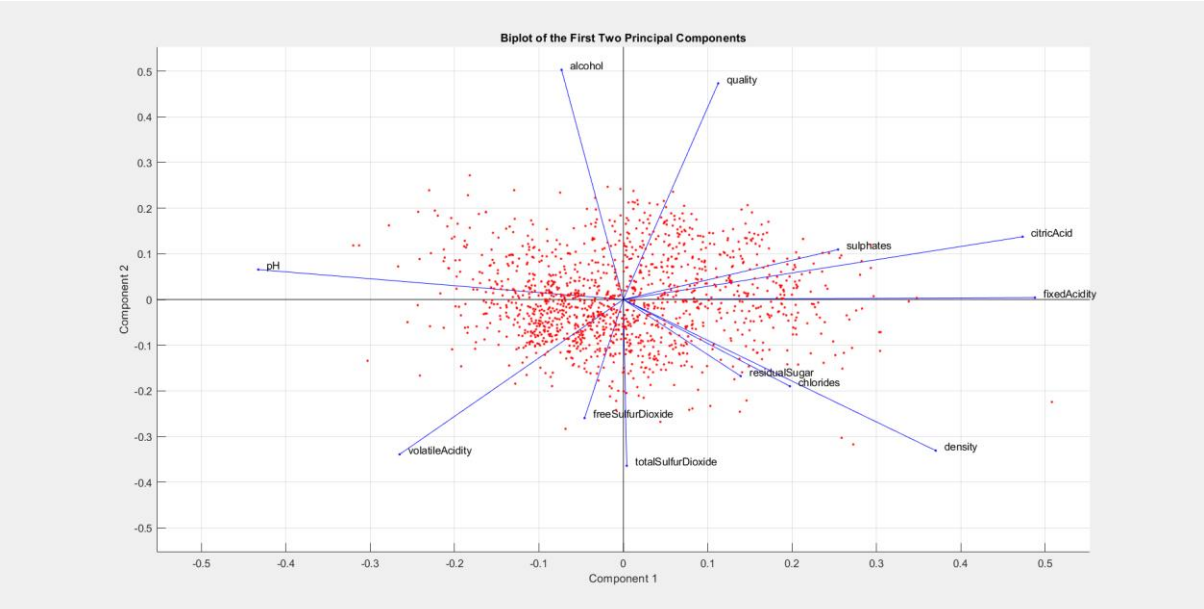

Figure 4: Biplot of the First Two Principal Components

Now, let's explore the covariances in the biplot. When angles between variables exceed 90 degrees (approaching 180), we observe negative correlations, as seen in citric acid and pH, citric acid and volatile acidity, quality and volatile acidity, quality and free sulphur dioxide, fixed acidity and pH, density and pH, and density and alcohol. Conversely, angles around 90 degrees indicate no significant correlation, seen in quality and residual sugar chlorides, volatile acidity and density, pH and alcohol, pH and free sulphur dioxide, and others.

Regarding high covariances, angles below 90 degrees suggest positive correlations, such as citric acid and sulphates, and density and residual chlorides.

Step 6: Quality Indicator Covariations

Next, we examine covariations between the quality indicator variable and the rest. Alcohol shows the highest positive correlation with quality, followed by sulphates, citric acid, and fixed acidity. No significant covariation exists between quality and pH, while negative correlations are seen with free sulphur dioxide, volatile acidity, and total sulphur dioxide, indicated by angles approaching 180 degrees.

Step 7: Loading Coefficients

Loading coefficients for the first principal component were visualized using a bar plot.

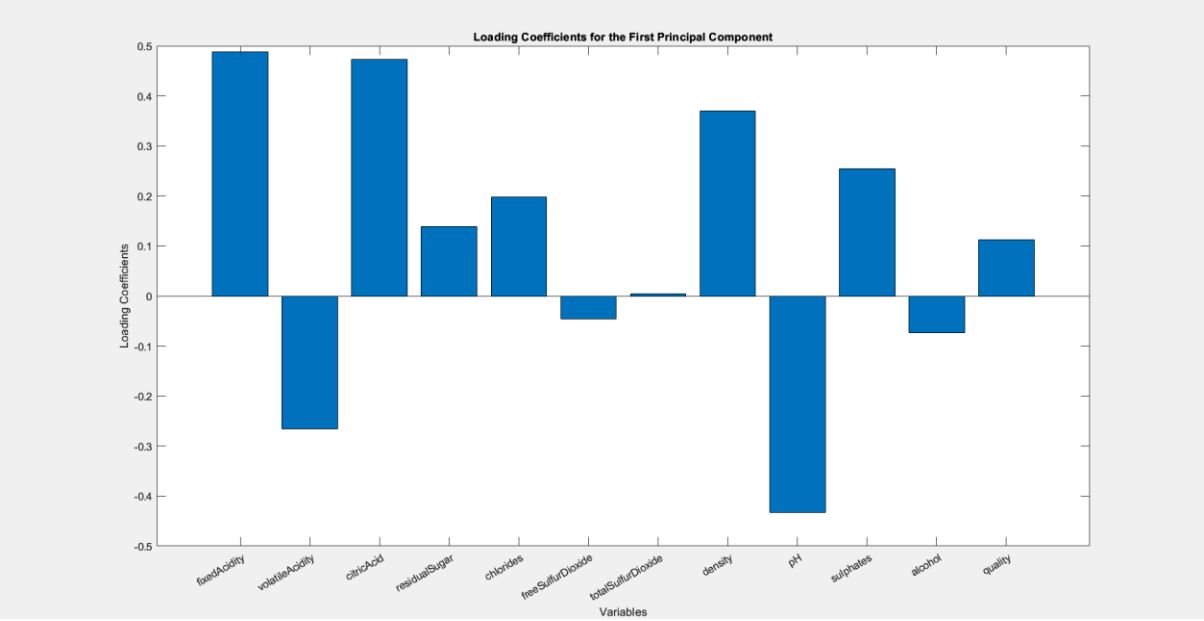

Figure 5: Loading Coefficients for the First Principal Component

Let's clarify how to interpret this figure. Variables with positive loading coefficients positively influence the principal component, while negative loading coefficients have a negative impact. In simple terms, an increase in a positively loaded variable leads to a higher principal component score,

and vice versa. The magnitude of this effect corresponds to the height of the bar. Larger or smaller values result in more significant or lesser contributions to the principal component.

In this analysis, we observe that fixed acidity and citric acidity have the most substantial positive impact on the PCA score, while pH and volatile acidity exhibit the strongest negative correlation with the PCA score. Variables with values close to zero, such as total sulfur dioxide and free sulfur dioxide, make minimal contributions to the PCA score.

Moving on, let's explore the T2 score. This score is a positive number representing the distance from the hyperplane's center to the observation's projection onto the hyperplane. The T-square plot is presented below:

Step 8: T2 Square Score Control Chart

A control chart for T2 square scores was generated, with control limits set at three standard deviations from the mean.

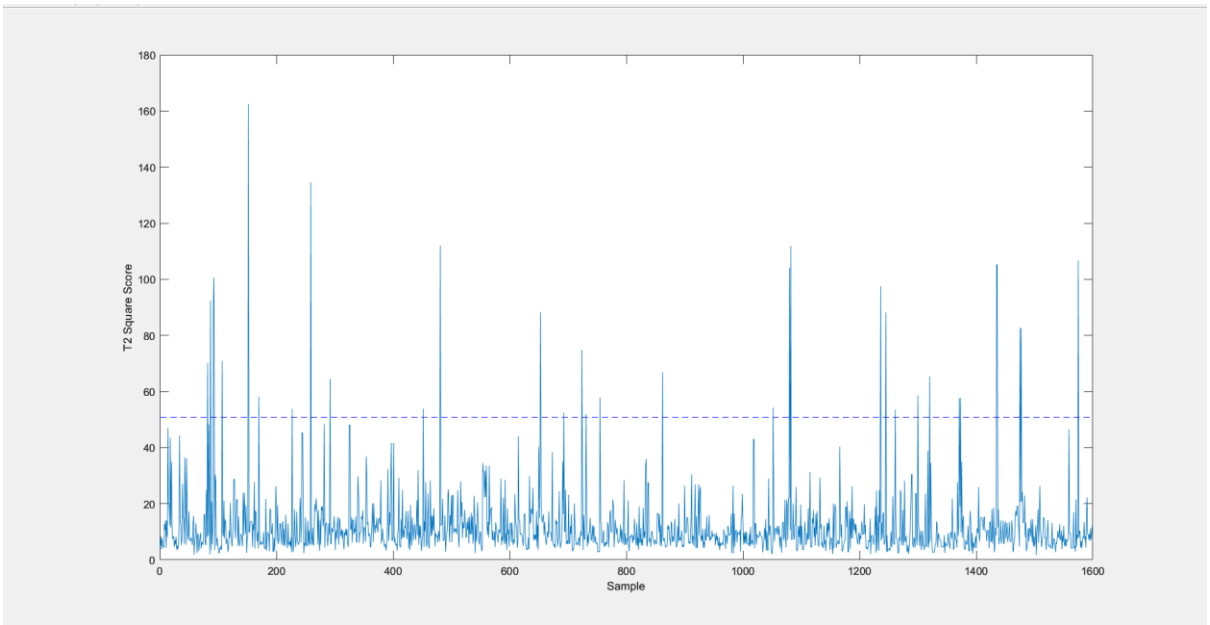

Figure 6: T2 Square Score Control Chart

Figure 6 depicts the T2 square scores along with the upper control limit (dashed blue line). Points exceeding this limit may indicate outliers or anomalies in the data.

Conclusion

This PCA analysis provided valuable insights into the dataset's structure and relationships among variables. The first few principal components captured significant variance, and their loadings were visualized to understand variable contributions. Additionally, a T2 square score control chart was

used to identify potential outliers. Further analysis, interpretation, and exploration can be conducted based on these results, depending on the specific objectives of the analysis.