ADVANCED DATA ANALYSIS AND MACHINE LEARNING

Practical assignment, PCA

Lappeenranta–Lahti University of Technology LUT

BM20A6100 Principal component analysis on red wine quality

September 16, 2024

Subin Khatiwada

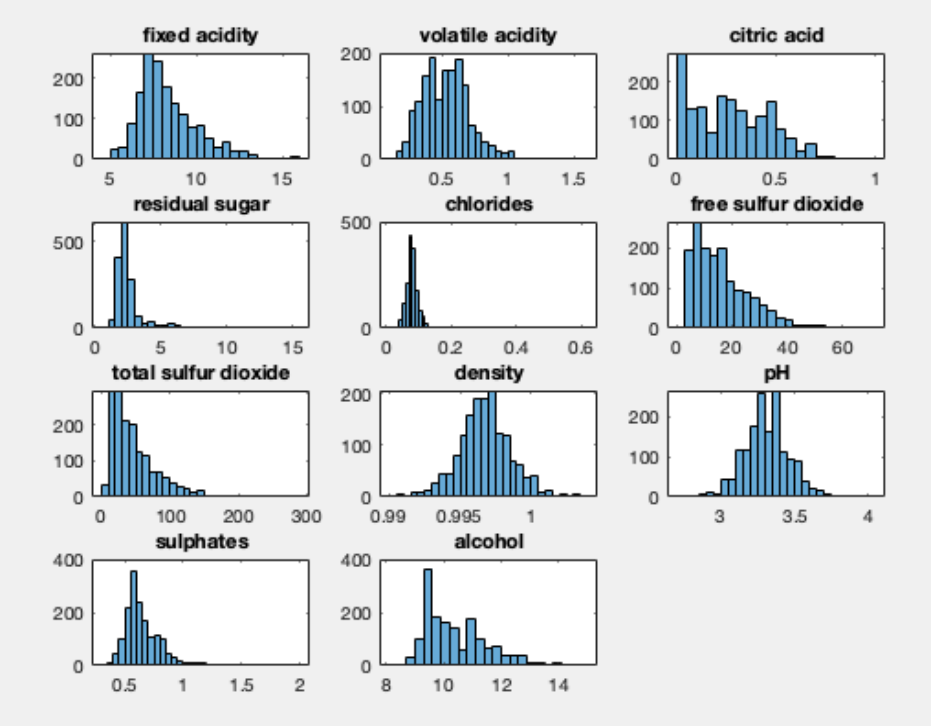
Introduction

The dataset used for this analysis is related to red variants of Portuguese "Vinho Verde" wine, with physicochemical attributes and sensory data available. The dataset contains the following input variables:

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH
10. Sulphates
11. Alcohol

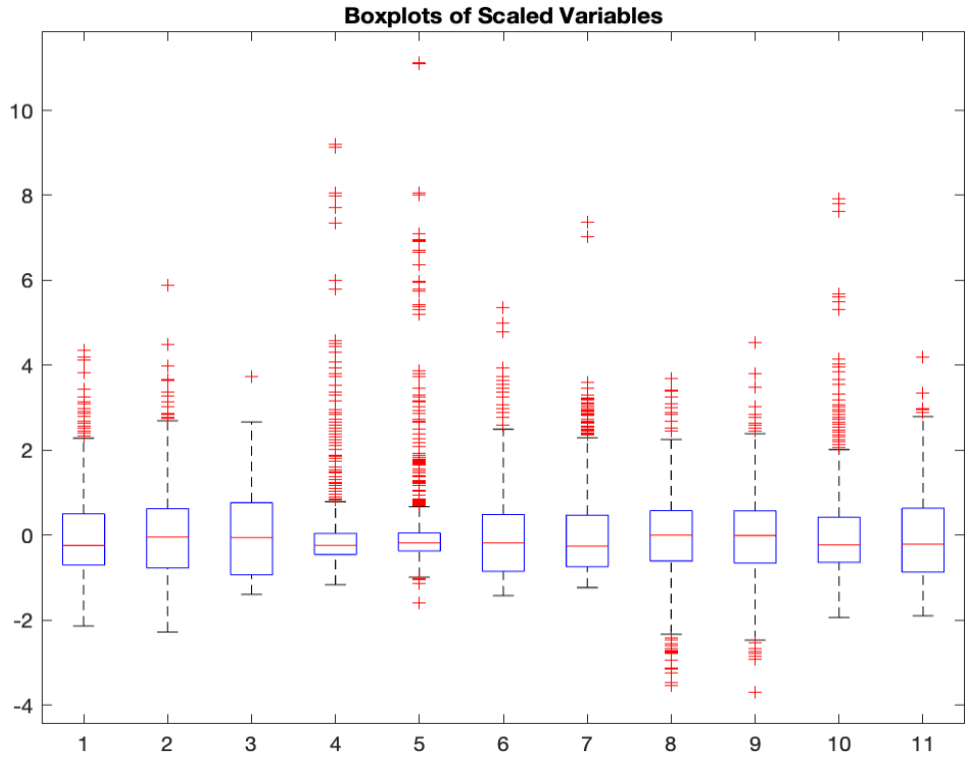
The output variable is wine quality, scored between 0 and 10. This dataset allows for classification and regression tasks, with the quality indicator representing the sensory evaluation of the wine.

1. Visualize the Dataset



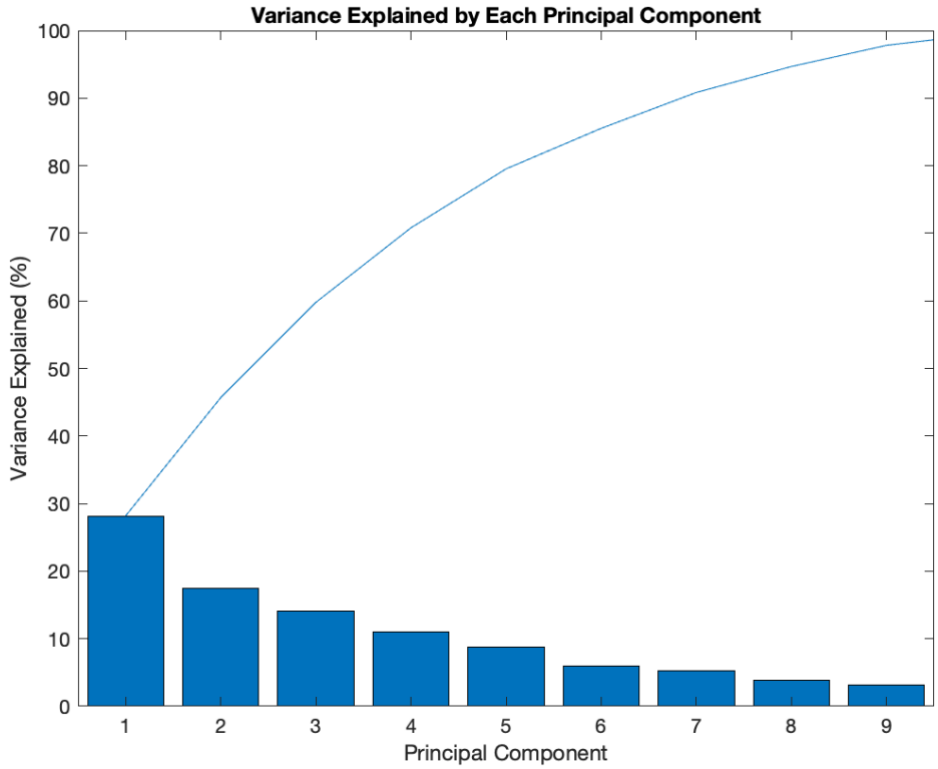
The histogram visualizations represent the distribution of each physicochemical variable in the red wine quality dataset, indicating variations in factors like acidity, sugar content, pH, and alcohol levels across the samples.

2. Scale and Centre the Dataset



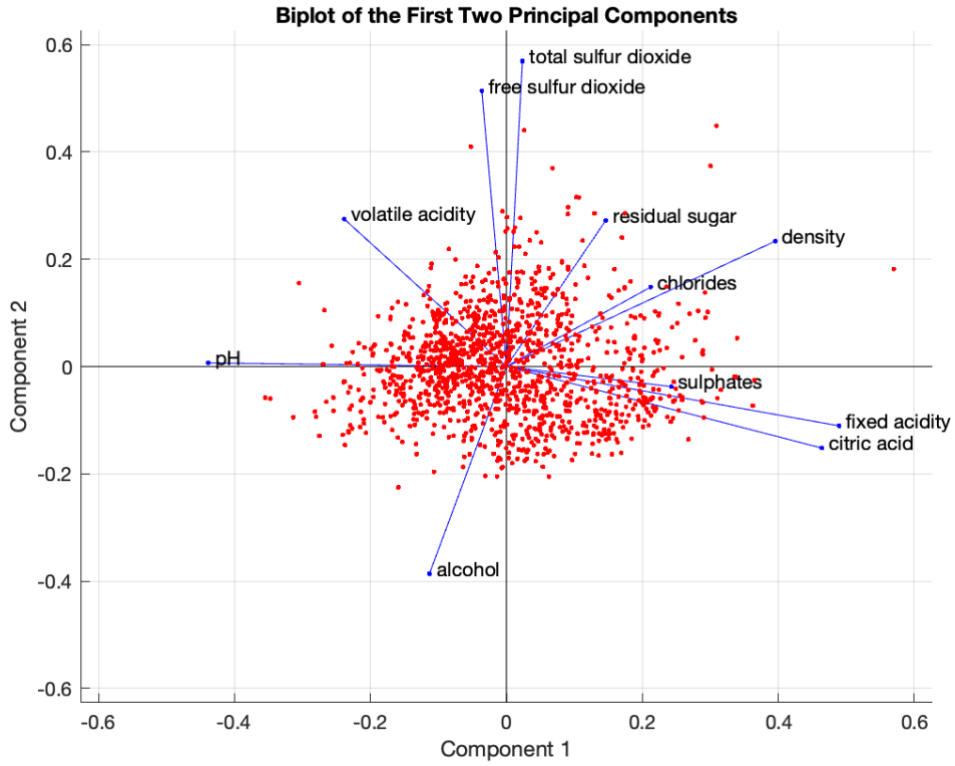
This boxplot shows the distribution of the scaled and centered variables in the red wine quality dataset. The values are centered around zero, and each variable's spread and outliers are displayed. Some variables, like residual sugar and total sulfur dioxide, have notable outliers. Scaling ensures all variables contribute equally to the PCA analysis.

3. Visualize and Comment on the Variation Explained



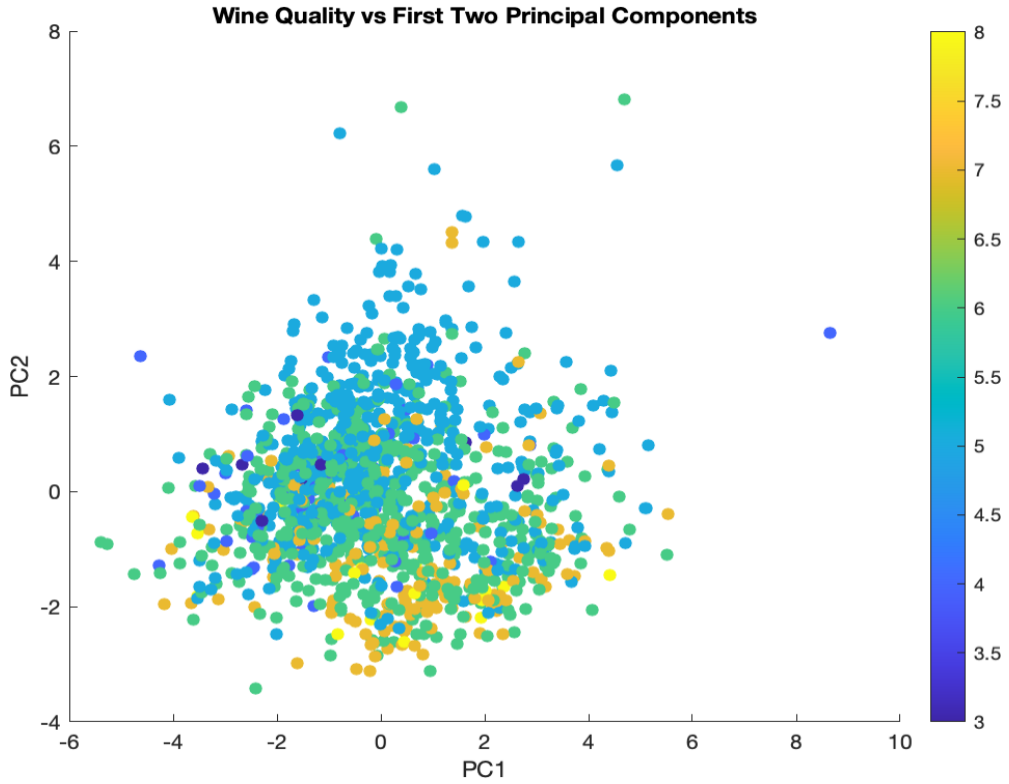
This scree plot shows the variance explained by each principal component (PC). The first few PCs capture the majority of the variance, with the first PC explaining around 30%. The cumulative variance reaches about 80% by the fifth PC, indicating that these components contain most of the information in the dataset.

4. Compute and Comment on the Biplot



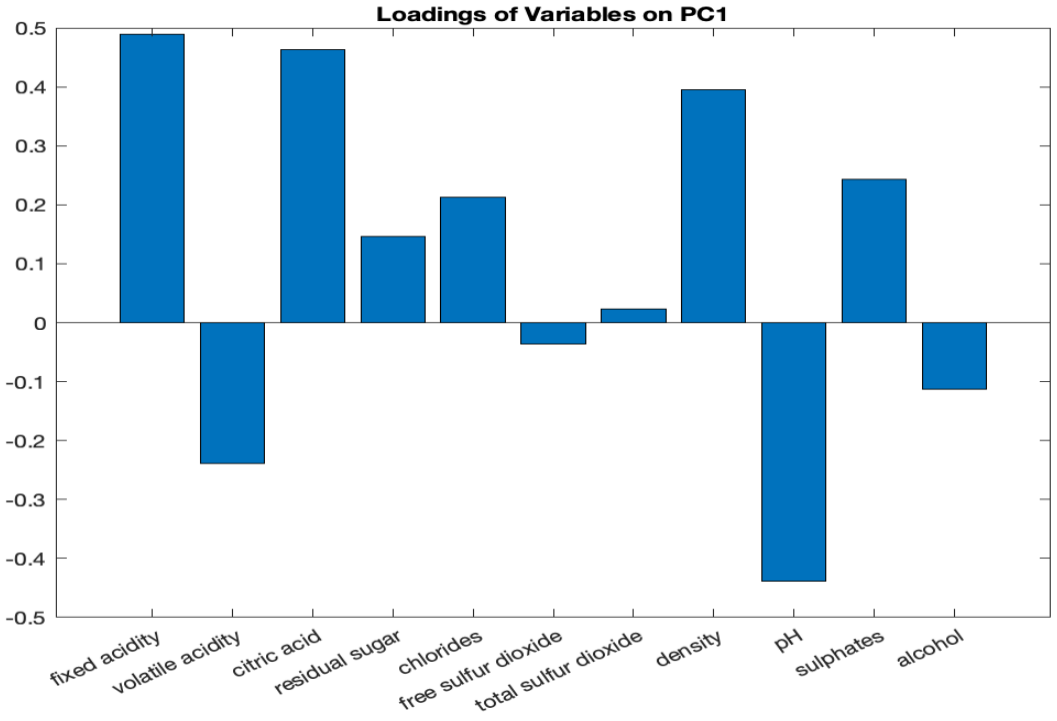
This biplot illustrates the first two principal components from the PCA analysis, showing how each physicochemical variable contributes to the components. Longer vectors, like those for total sulfur dioxide and free sulfur dioxide, indicate a stronger influence, while the angles between vectors reflect correlations between variables. The red points represent individual wine samples, with their distribution revealing the variance in wine characteristics.

5. Explain Co-variations Related to Quality Indicator



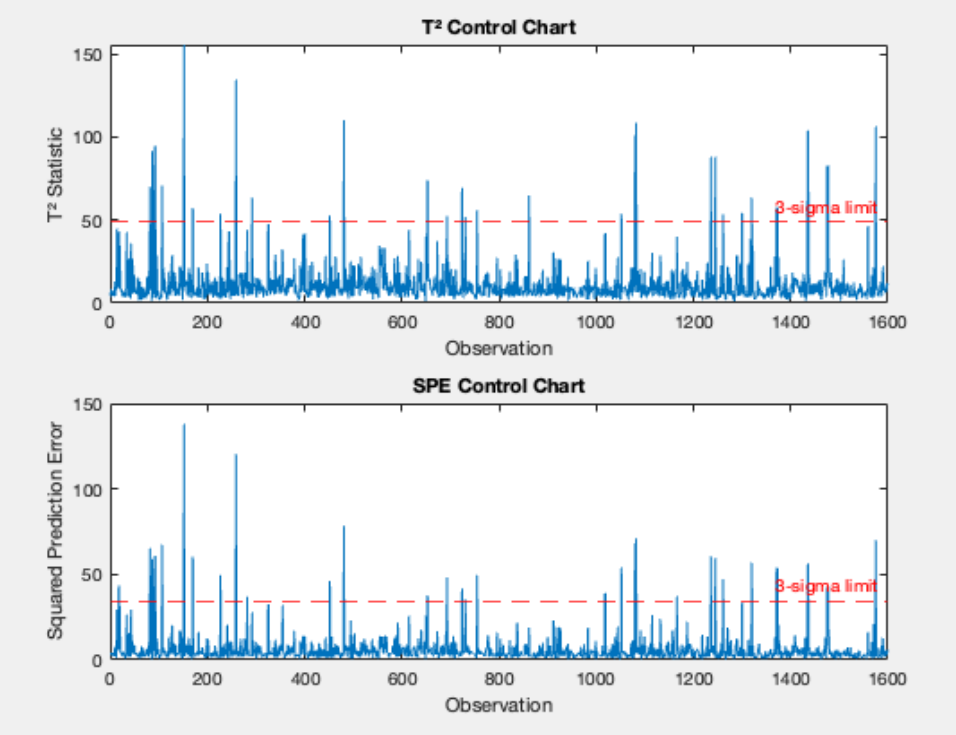
This scatter plot shows wine quality mapped against the first two principal components (PC1 and PC2) from the PCA. The points represent individual wines, coloured based on their quality score, as indicated by the colour bar. The plot reveals how the wines cluster based on these principal components, with quality variations spread throughout the PCA space, showing no clear separation between different quality levels.

6. Loading Bar Plots



This bar plot displays the loadings of each variable on the first principal component (PC1). Variables like fixed acidity, volatile acidity, and density contribute positively to PC1, while pH, sulphates, and alcohol have negative contributions, indicating their influence in separating the data along this component.

7. T^2 and SPE Control Charts



Conclusion

The PCA analysis on the red wine quality dataset reveals that the first few principal components capture most of the data's variance, with the first two components showing strong contributions from variables like acidity and sulfur dioxide. Despite the distribution of wine quality across these components, no clear separation of quality levels is observed. The analysis helps identify key physicochemical factors influencing wine characteristics but shows that wine quality is influenced by a complex interplay of multiple variables.