Student Name: Zhou Chen

Student ID: 112214117

Source of Dataset: [UCI Machine Learning Repository: Wine Quality Data Set](https://archive.ics.uci.edu/ml/datasets/Wine+Quality)

Lab 2 Report

The dataset being used in this lab is the dataset that collects the variants of the Portuguese “Vinho Verde” red wine. In this dataset, there are 12 attributes, the first 11 attributes are input variables, which are based on physicochemical tests. The last attribute quality is the output variable, which is based on the sensory test by the specialist that assesses the appearance, aroma, taste, body, and persistence in mouth, then rate the quality of wine from score 1 to 10. There are 1599 instances of the dataset being used in this lab, only including the portion of red wine of this dataset, the quality of red wine is scored from 3 to7.

Table 1. Description of Attributes in Wine Quality

| Feature Name | Description |
| --- | --- |
| fixed acidity | The scale of tartaric, malic, citric, and succinic in wine |
| volatile acidity | a measure of the wine's volatile (or gaseous) acids |
| citric acid | The percentage of citric acid added to wine |
| residual sugar | The grams per liter (g/L) in wine |
| chlorides | The percentage of chlorides added to wine |
| free sulfur dioxide | The concentrations of free SO2 |
| total sulfur dioxide | The concentrations of total sulfur dioxide |
| density | Density of wine |
| pH | pH of Wine |
| sulfates | The concentration of sulfates found in wine |
| alcohol | Percentage of alcohol |
| Quality | Quality of Wine from scale 1 to 10 |

In this lab, the principal component analysis is done on the dataset. By using the result of PCA on the dataset, the eigenvalue and eigenvector are derived from it. The scree plot of explained variance are plotted. The biplot of the dataset including the projected dimension axes. Displaying the PCA in a table based on the intrinsic dimensionality index. With the scatter matrix that demonstrates the top four attributes with highest sum squared of loadings. Perform a K-mean clustering algorithm on the dataset, using the value of k from 1 to 10 as the number of clusters, then use the elbow method to identify the elbow from the inertia plot. Then use the elbow point, k = 3, as the optimal number cluster to display the k-mean clustering graph.

The output of PCA shows the attribute fixed acidity has the highest explained variance rate, and volatile acidity is the second. When the intrinsic dimensionality index is two, the top four attributes with highest sum square of loadings are alcohol, total sulfur dioxide, fixed acidity, and citric acid. According to the scatter matrix created based on these four attributes, the fixed acidity and citric acid are positively correlated. Other attributes in the scatter matrix show zero correlation. But depending on their clusters, the graph based on the different attributes shows different layers, each layer consists of only one cluster.