Dheeraj Goud Borlla (dxb160130)

ASSIGNMENT – #2: (PRINCIPAL COMPONENT ANALYSIS)

PURPOSE: In this assignment, we study the application of PCA (Principal Component Analysis). First, we apply classification techniques like K-means clustering and Hierarchal clustering to the data set. Now the dimensionality reduction technique, PCA is used on the data to map it into lower-dimensional space. Then again, the K-means and Hierarchal clustering methods are applied to the transformed data and the results obtained are verified with the results obtained before to check how the PCA has altered the results.

DATASETS: winequality-white.csv, winequality-red.csv <uci/datasets/Wine+Quality>

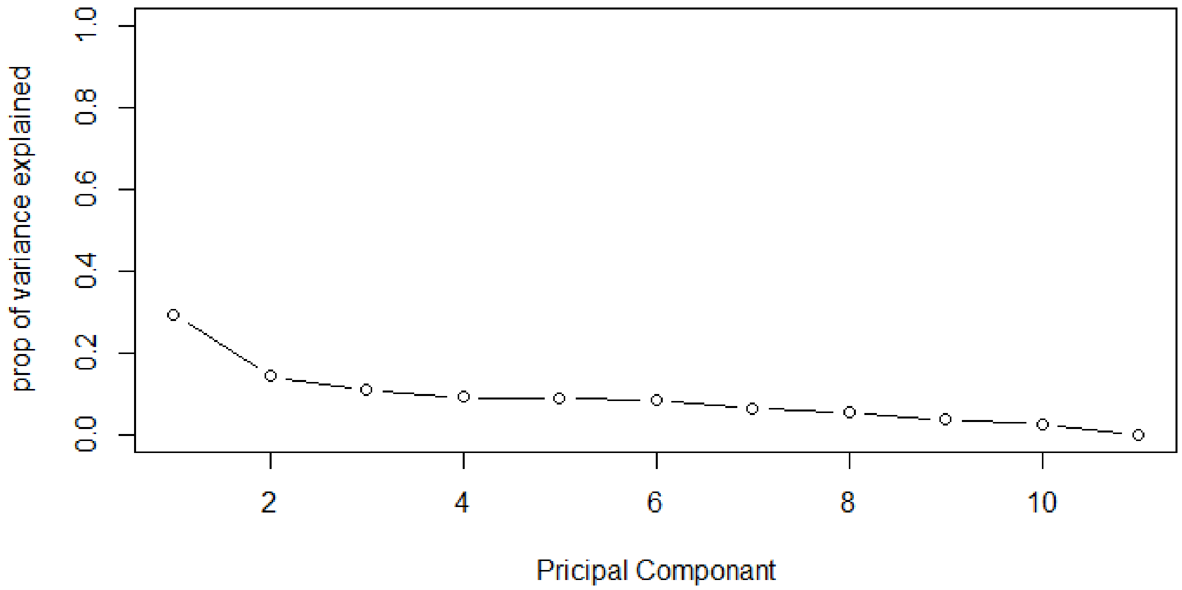
APPROACH 1: In this approach, winequality-white.csv data set is used for analysis.

- First, we applied the clustering techniques to the dataset. When the K-means clustering is applied to the 'scaled' data for different values of k, and k=4 seems to be a reasonable clustering for the data. When the same 'scaled' data is processed under hierarchal clustering with complete, average and ward.D2 methods, and we can determine that four clusters are reasonable.
- Next, we applied Principal Component Analysis to the dataset. And we determined that 2 principal components were sufficient (see graph 1), quality is very much correlated with alcohol content and fixed acidity.
- Next, we applied K-means clustering to the data set transformed under PCA, and found that the ratio, between_ss / total_ss is increased from 31.8% to 69.9%.

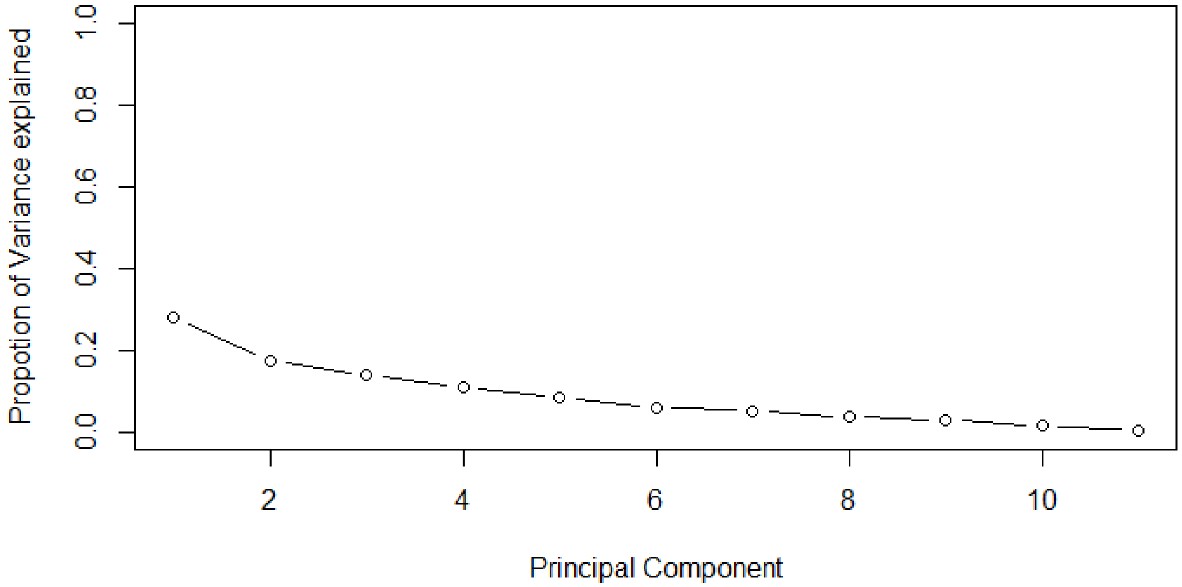
APPROACH 2: In this approach, winequality-red.csv data set is used for analysis.

- First, we applied the K-means clustering to the scaled dataset for different values of k, and k=4 seems to be a reasonable clustering for the data. When the hierarchal clustering is applied to the 'scaled' data with complete, average and ward.D2 methods, we can determine from the dendrogram that four clusters seems reasonable for the data.
- Next, we applied Principal Component analysis to the dataset. We determined that 2 principal components were sufficient (see graph 2).
- Next, we applied K-means clustering to the data that is transformed under PCA, and found that the ratio, between_ss / total_ss is increased from 34.8% to 68.4%.

GRAPHS:



Graph 1



Graph 2

SUMMARY:

- When the given data set is scaled and K-means clustering is applied to it. Four is a reasonable number of clustering but the data seem more complex.
- Next, Principal Component Analysis is applied to the scaled data. The transformed data is defined in a way that the first principal component has the largest possible variance. In both the datasets only two principal components explain the variance of the data.
- From the PCA transformed data, with the maximum variance, K-means can cluster the data more efficiently than before.