Running clusture analysis on dataset:

We are using wine quality data to apply the clustering method. To determine the appropriate number of clusters we evaluated how the sum of square varies by clusters.

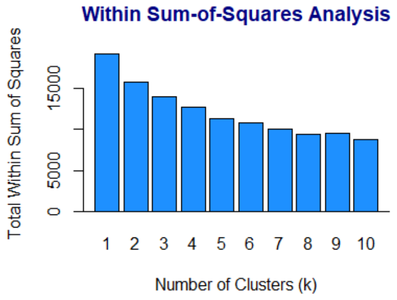


Figure 1: Within Sum-of-Square Analysis

As we can see from the sum-of-square plot the sum-of-square falls continuously till the 8th clusture and then become consistent or has vey less decrease from the previous value. So, we take 8 as the number of clusters for our analysis. Illustrating the cluster with the help of dendrogram. This illustrates the hierarchical relationship between objects.

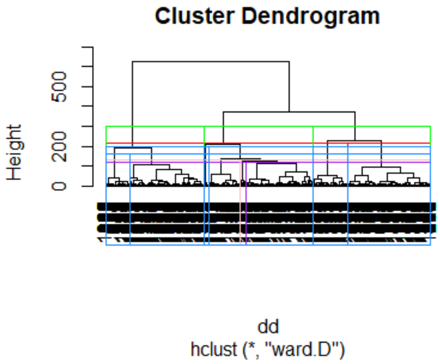


Figure 2: Clusture Dendrogram

Dimensionality Reduction Approach using PCA:

Initially, plotted a correlational plot to see the relation between different components.

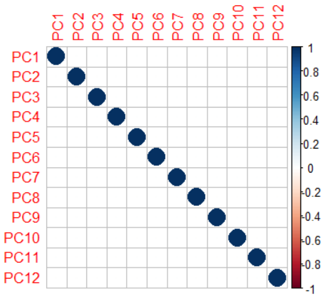


Figure 3: Correlation between different components.

As we can observe from the correlation plot the components are uncorrelated. Plotting a scree plot for the principal components to observe the variance.

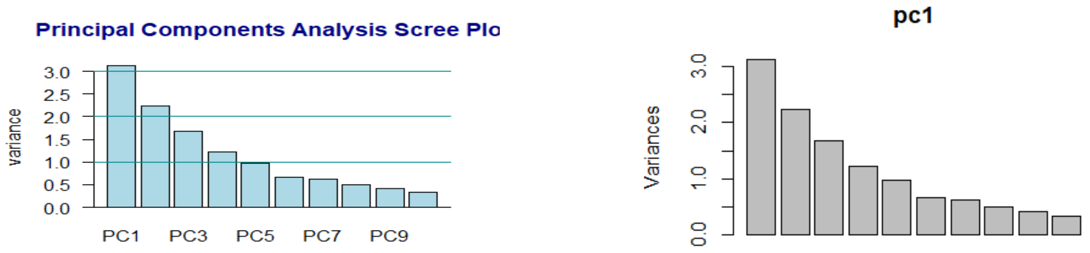


Figure 4: Scree plot for PCs.

So, the first PC seems to explain a lot of variation which can be further illustrated by a Bi-plot.

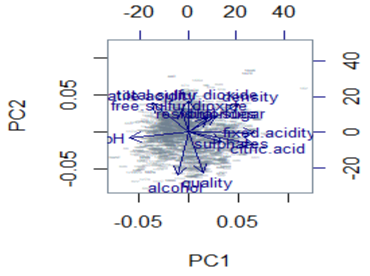


Figure 5: Bi-plot for PC.

So, the first principal component explains around 26% of variation in data. The first and second principal component explains the around 45% of the total variation in the data. By looking at the cumulative proportions the top 6 components are sufficient to explain the variation in data because the cumulative proportion for these components is around 82%.

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
0.26	0.447	0.587	0.688	0.769	0.825	0.876	0.918	0.953	0.98	0.995	1

Table 1: Cumulative proportions of different Principal Components

The final cluster distribution using k-mean (here 8) clearly shows the formation of different clusters.

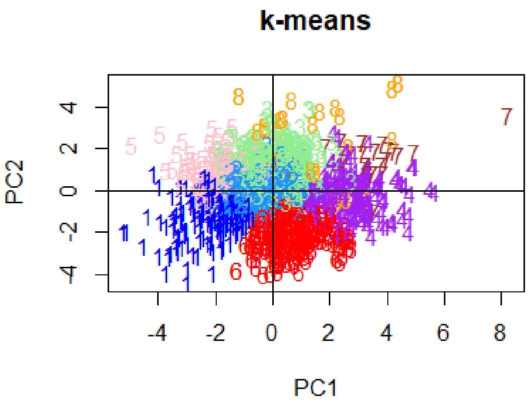
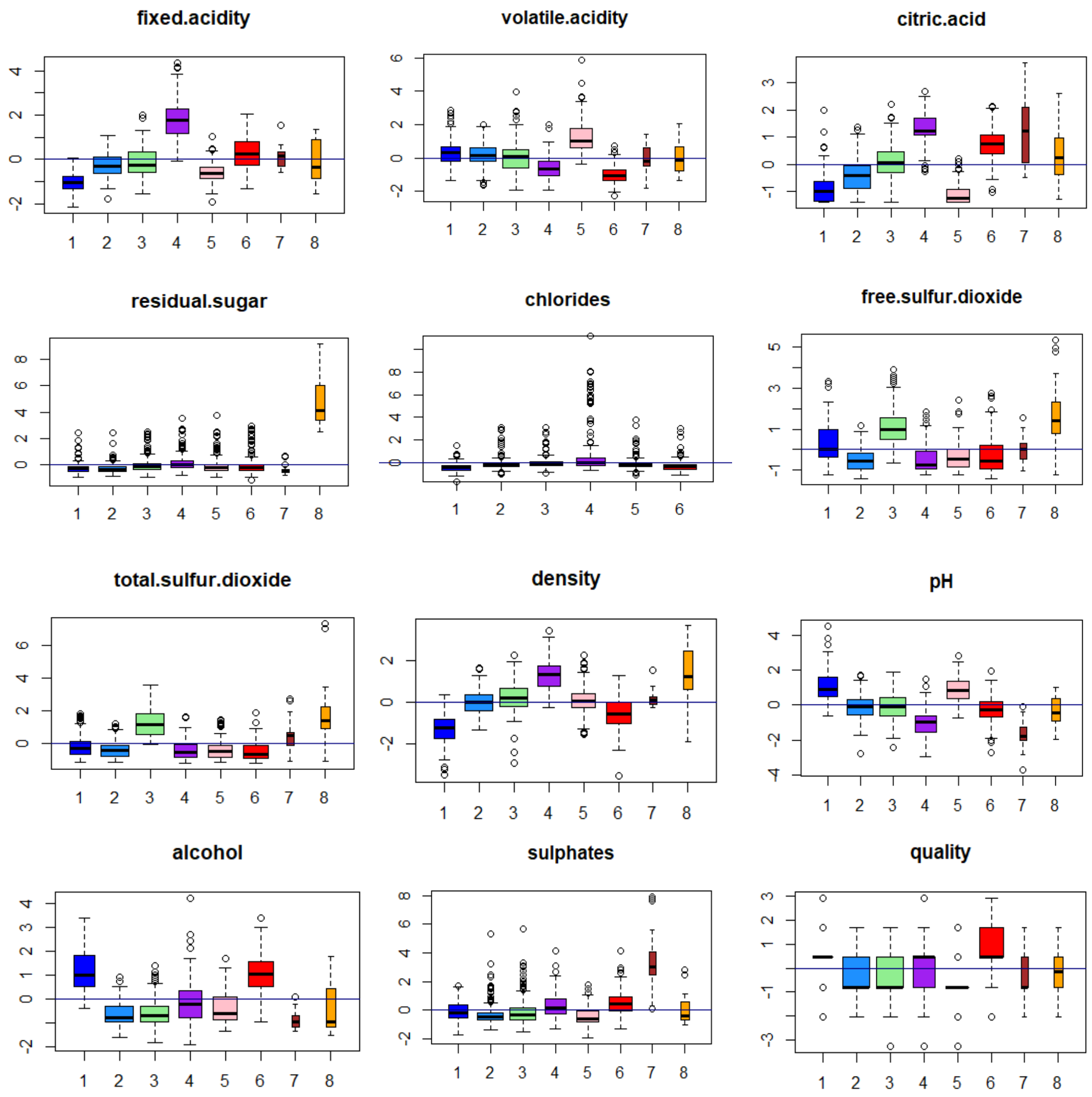


Figure 6: K (8)-Means clusture Formation



To visualize the clusture there are the plots for feature distribution by clusters. We can compare the different features based on their centroid variation in different clusters. For example, in citric acid the centroid varies a lot in different clusters but for features such as chloride the centroid does not vary a lot in different clusters. We can clearly see the variation across the variables for each of the clusters found.

Comparing both the dimensionality reduction and the cluster analysis we can say that in clusture analysis we try to make clusters as minimum as possible which also explains the dataset completely (here 8 clusters are enough to explain the dataset) while in PCA we try to select the minimum principal components which can explain the variation in the dataset (here 6 components are enough to explain about 80% of the variation in the dataset.)

Conclusion:

According to the cluster analysis using sum-of-square plot we concluded that it is best to divide the dataset into 8 clusters. With PCA dimensionality reduction we concluded that 6 principal components could be enough to explain the cumulative variation of about 80% in the dataset.