Group H - Assignment 2 Report

1. Analysis of the meaning of the first two principal components

PCA (Principal Component Analysis) is a statistical procedure that finds components that explain the most of the variance in the data set. When using PCA, we are assuming that the information contained in the original variables itself is very low. Therefore, we create Principle Components, which are linear combinations of the original variables and therewith explaining the variability of the original variables. The first Principle Component depicts the highest variability while the last Principle component depicts the least variability. That way, we are able to reduce the number of variables while making sure we describe a certain level of variability. In general, a condition to apply the PCA is that we have a quantitative variable. Therefore, we had to exclude the categorical variable “Customer Segment” from our dataset in order to apply the PCA.

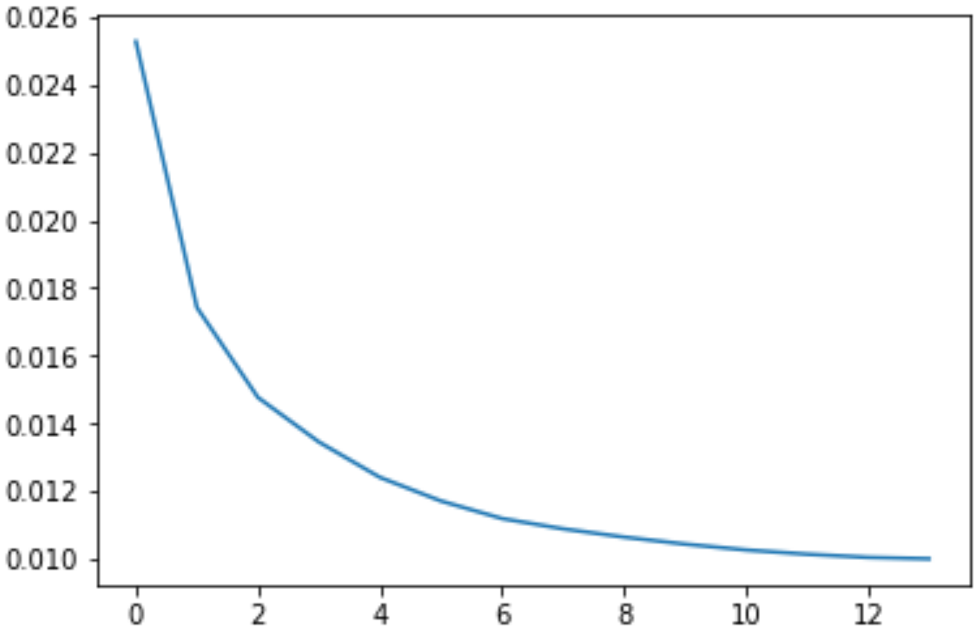


Fig. 1 - Scree Plot

Regarding the question in our assignment, the first two Principal Components explain 57% of the variance in the data, as implied by the scree plot above. This variance can be displayed in a two dimensional space in the circle of correlation (as seen below), where the x-axis describes the first Principal Component and the y-axis the second Principal Component. For the first Principal Component, we can see that consists of several original variables with a relatively small negative effect. Since they point in the same direction, we can say that they are correlated. For the second Principle Component, we can see relatively strong negative effects from Ash_Acanity and Flavanoids and a positive effect from Total Phenols. In practice, this means that Total Phenols is negatively correlated with Ash_Acanity and Flavanoids.

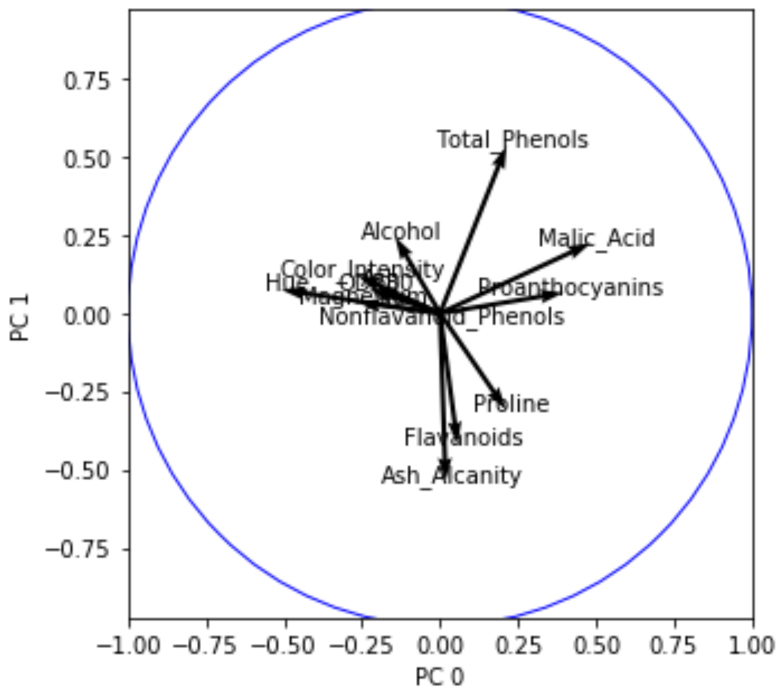


Fig. 2.1 - Correlation Circle with labels

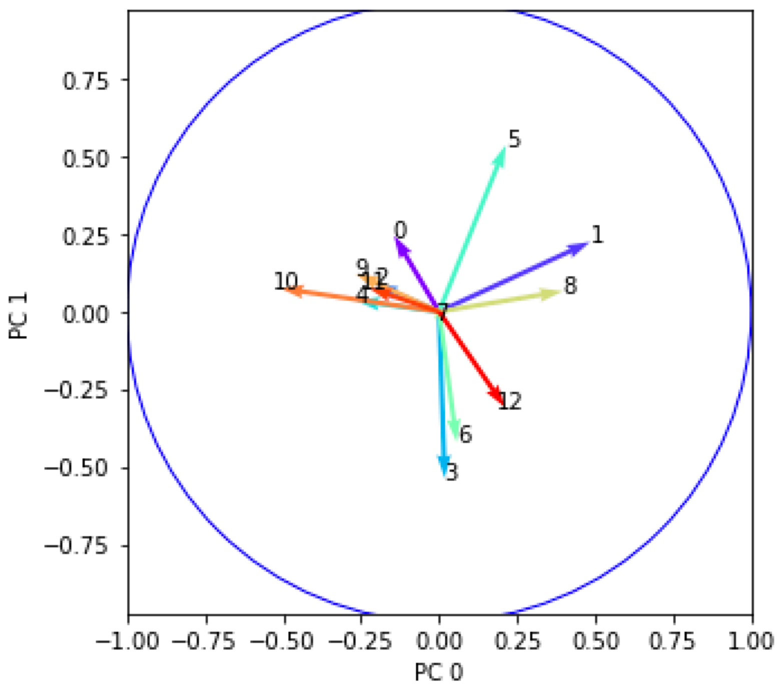
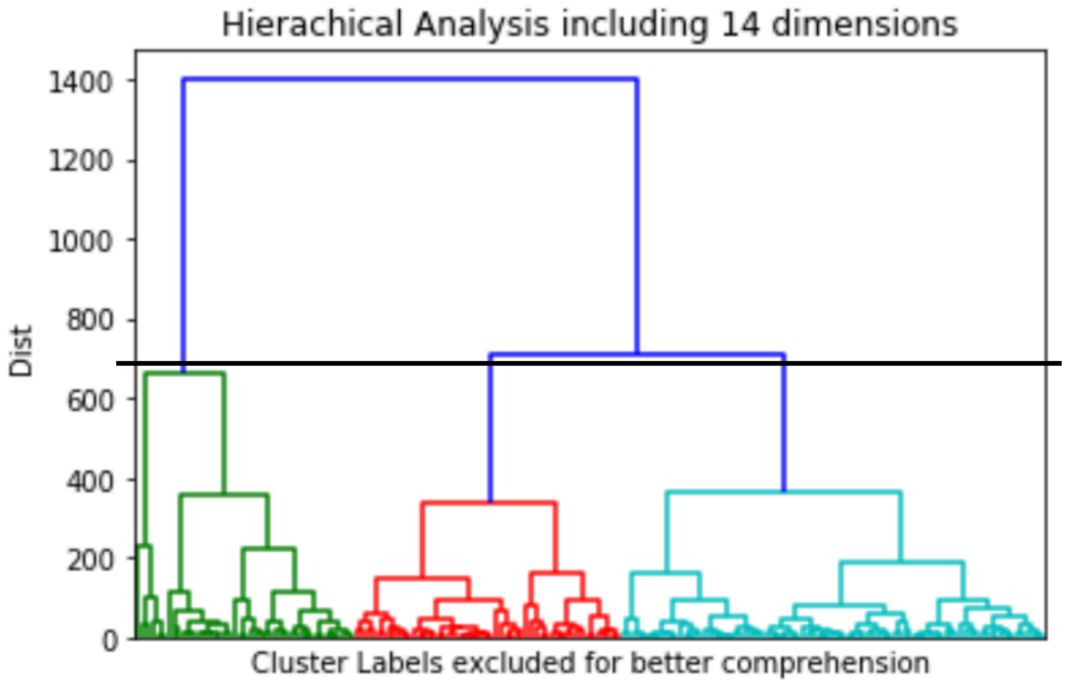


Fig. 2.2 - Correlation Circle

While these are among the most significant inferences we can make from the circle of correlation, the circle of correlation describes many more implications.

2. Use a hierarchical cluster algorithm to guess a likely number of clusters present in the data

The dendrogram below is a tree diagram used to show the way in which the clusters are created by hierarchical clustering. Horizontal lines are drawn when two clusters are joined, therefore, a likely number of clusters is three because it allows to "cut" relatively long vertical lines. These long vertical lines are indicators of increasing heterogeneity within the clusters which, as a consequence, should not be joined.



. *Fig. 3 - Dendogram*

3. Use the previous number of cluster to perform a K-means cluster analysis

   3.1. Analyse the “silhouette” of the clusters

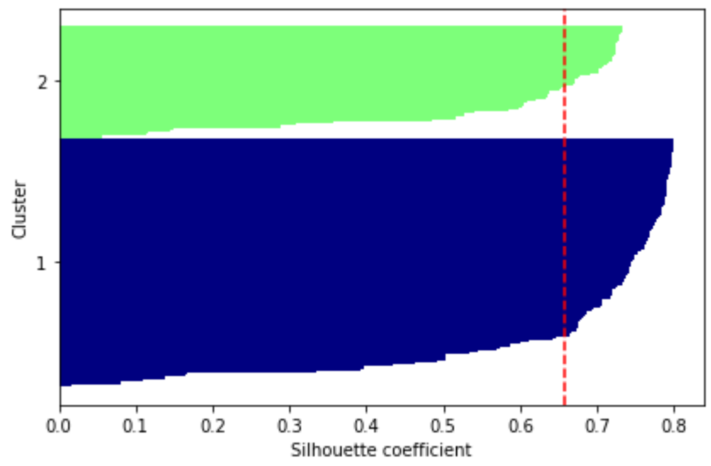


Fig. 4 - silhouette with two clusters

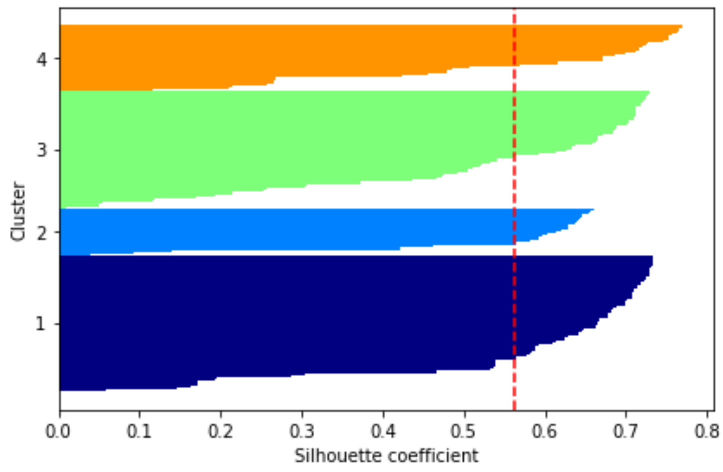


Fig. 5 - silhouette with four clusters

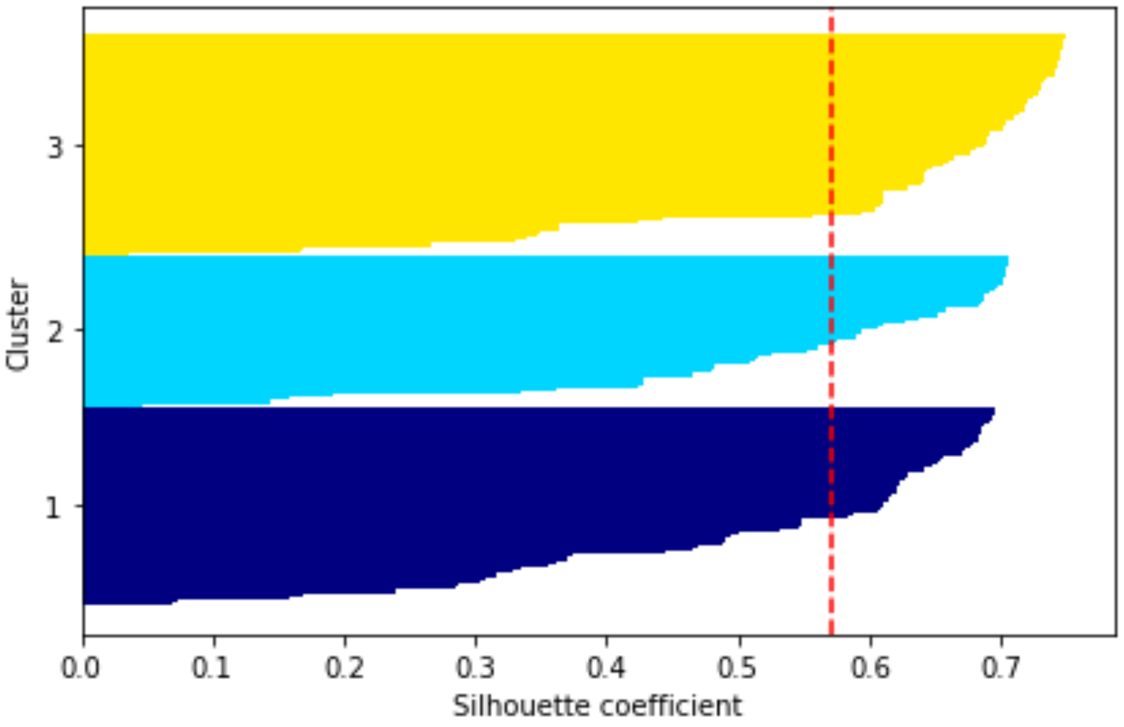


Fig. 6 - silhouette with three clusters

The silhouette analysis is a method to find how much a point is closer to the points in its neighboring clusters. Values of silhouette coefficient near one mean that the point is very different to the neighbor cluster (that is, it is in the right cluster); values of silhouette coefficient near zero mean that the point is very near to the neighbor cluster (in between the two clusters) and negative values mean that the point is in the wrong cluster. Silhouette analysis can be used to find what is the best number of cluster to divide data in.

From the k-mean analysis we find that the best number of cluster is three and the silhouette analysis confirms this number. In fact, we can see in the image above that the three clusters are all above the mean silhouette coefficient line and they are more or less uniform in width, which would not be true if we had only two clusters. However, it is true that two clusters have a higher coefficient mean.

We can compare our result to what the silhouette would look like if we had chosen four clusters. We can understand from the graph that this is not the right number of cluster because:

- the mean coefficient is much lower than the one for three clusters;
- the clusters are all above the average coefficient line, but they have a lot of the observations below that line, which is not ideal;
- the clusters are all very different in width.

If we decide to have just two clusters, we notice that the the mean coefficient is higher than in both the other two cases and also that the number of points below that average is small (similar to the case with three clusters). However, the two clusters are very different in width, which leads us to choose to have three clusters instead of only two.

3.2. Plot on the space of the first two dimensions of the PCA the clusters obtained with K-means, using a different colour for each cluster

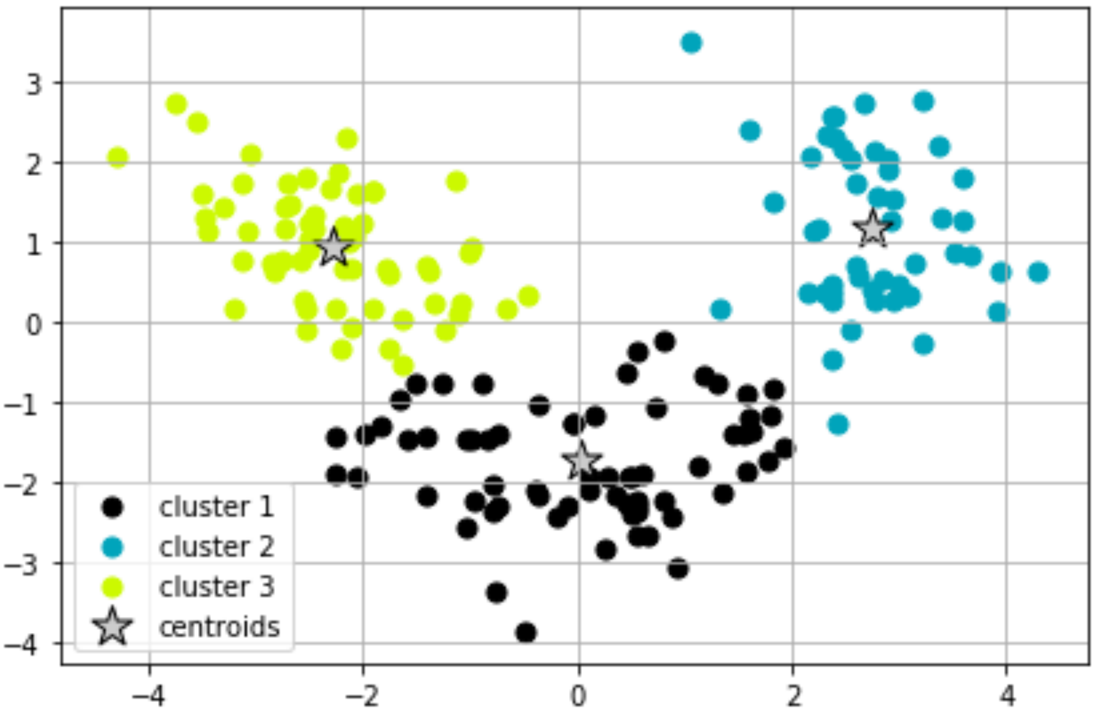


Fig. 7 - clusters obtained with k-means

3.3. For each cluster, which “original” variables (ex ante the PCA) are more important? Consider the barycenter of each cluster (the barycenter is an observation) and its variables values

The most important original variables for each cluster are:

- Cluster 1: Ash_Alcanity and Flavanoids
- Cluster 2: Total_Penols and Malic Acid
- Cluster 3: Color_Intensity and Alcohol

3.4. Using both the information of barycenters and of PCA, give an interpretation to each cluster

- Cluster 1: Wines with high Ash_Alcanity and high Flavanoids are most strongly represented in this cluster. In fact, for this cluster, PC1 is negative and Ash_Alcanity and Flavanoids are negatively correlated with PC1.
- Cluster 2: Wines with high Total_Phenols and high Malic_Acid (which are the two most important original variables for this cluster) can be identified within this cluster. In fact, for this cluster, both PC0 and PC1 are positive and Total_Phenols and high Malic_Acid are positively correlated with them.
- Cluster 3: Wines with high Color_Intensity but also a high level of alchohol would fall within this cluster. At least these original variables are identified to have a fairly strong effect. However, we have to note that this cluster is correlated with several original variables that do not have as strong as an effect as variables in other clusters.

4. Write a function that takes in input the dataset and that returns 1) the value of K (for the K-means) that is associated with the best overall silhouette of the K-means algorithm and 2) the plot of the correspondent clusters on the space of the first two dimensions of the PCA (performed over the same dataset)

Note: We are setting the minimum number of clusters equal to 2 and the maximum number of clusters as 11, which can be easily changed by manipulating the variables called min_cluster and max_cluster.

5. Write a function that takes in input the dataset: the function performs the PCA and returns the circle of correlations of each pair of principal components (1 and 2, 1 and 3, 1 and ..., 2 and 1, 2 and 3, ...). Plot all the circles in the same plot and/or in a series of plots 3x3

Note: In the file called PCA_component_comparison_exercise.py we have two functions:
1. called principal_components_comparison_3by3, which runs the comparison of all the PCAs against each others without repeating the comparison (1vs2 is considered the repetition of 2vs1). This was tested on the wines_properties dataset, which

produced 9 figures with a total of 78 graphs plotted on the aforementioned figures and in a 3x3 grid.

2. principal_components_comparison_given_data runs the comparison against the selected PCA and all the others. It also takes in dataset as an argument. All of the plots are done in the same figure in this case.