Assignment 2

Using the dataset “wine_properties.csv” in the folder “Assignment_2” and Python, do the following tasks:

- Perform a PCA, analysing the meaning of the first two principal components using the “circle of correlations”
- Use a hierarchical cluster algorithm to guess a likely number of cluster present in the data
- Use the previous number of cluster to perform a K-means cluster analysis
    - Analyse the “silhouette” of the clusters
    - 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.
    - 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.
    - Using both the information of barycenters and of PCA, give an interpretation to each cluster.
- 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).
- 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.