**1. Research about the Spectral Clustering method, and answer the following questions:**

**a. In which cases might it be more useful to apply?**

R/ When the data is high dimensional, has a non-linear structure, has a network structure and when fine control over clustering is needed.

**b. What are the mathematical fundamentals of it?**

R/ First, a graph describing the relationships between the data is constructed. Each data point is represented as a node in the graph, and is connected to its nearest neighbors by edges, the Laplacian matrix is calculated, which is a symmetric matrix describing the properties of the graph. The Laplacian matrix is used to find the eigenvectors and eigenvalues with the spectral decomposition, The eigenvectors corresponding to the smallest eigenvalues are used to reduce the dimensionality of the data to a lower dimensional space, Finally, a clustering algorithm, such as k-means, is applied to the reduced eigenvectors to group the data into clusters.

**c. What is the algorithm to compute it?**

R/ Construction of the affinity matrix, A, which measures the similarity between each pair of points in the data. The affinity matrix can be constructed using different similarity measures, such as Euclidean distance or cosine similarity.

Calculation of the Laplacian matrix from A, which measures the similarity between each pair of points in the data.

The k smallest eigenvalues of the Laplacian matrix L, which are known as Fiedler eigenvalues (smallest eigenvalues), are calculated. The eigenvectors corresponding to these eigenvalues are used to construct a k-column V matrix.

A clustering matrix Y is constructed from the matrix V by taking the k rows corresponding to the eigenvectors associated with the k smallest eigenvalues. The matrix Y is normalized and used as input for a clustering algorithm, such as k-means, to obtain the final k clusters.

**d. Does it hold any relation to some of the concepts previously mentioned in class? Which, and how?**

R/ If it is related to the search of the vectors and the eigenvalues of the data, with the objective of obtaining the most relevant characteristics of the information (variance) and performing a dimensionality reduction, and with the use of a clustering method such as kmeans to group the reduction data.

**2. Research about the DBSCAN method, and answer the following questions:**

1. **In which cases might it be more useful to apply?**

R/ High dimensional data, with noise, with different densities, also for anomaly analysis and nonlinear data.

1. **What are the mathematical fundamentals of it?**

R/ It works based on the density of points in the feature space. The goal of DBSCAN is to group points that are close to each other and separate points that are far apart into different clusters. The algorithm starts by selecting a random point and determining if this point is a center point. A center point is one that has at least minPts points within its epsilon radius neighborhood.

1. **Is there any relation between DBSCAN and Spectral Clustering? If so, what is it?**

R/ Both are good with nonlinear data, although Spectral Clustering is based on the similarity matrix of the data, it can be used with different similarity measures, including the Euclidean distance used in DBSCAN. Both methods try to extract features from the data.

**3. What is the elbow method in clustering? And which flaws does it pose to assess quality?**

The name of the method comes from the fact that the plot of the relationship between the number of clusters and the clustering quality measure (such as the sum of intra-cluster distances or intra-cluster variance) often resembles an elbow.

The elbow method involves plotting the clustering quality measure as a function of the number of clusters. As the number of clusters increases, the quality measure generally decreases. This is because the more clusters there are, the smaller they are and the more dispersed the clusters are. However, at some point the increase in clustering quality will level off and resemble an elbow in the graph.

The point at which the graph begins to form a kink is the optimal number of clusters. This point can be determined visually or using more advanced techniques, such as the silhouette method.

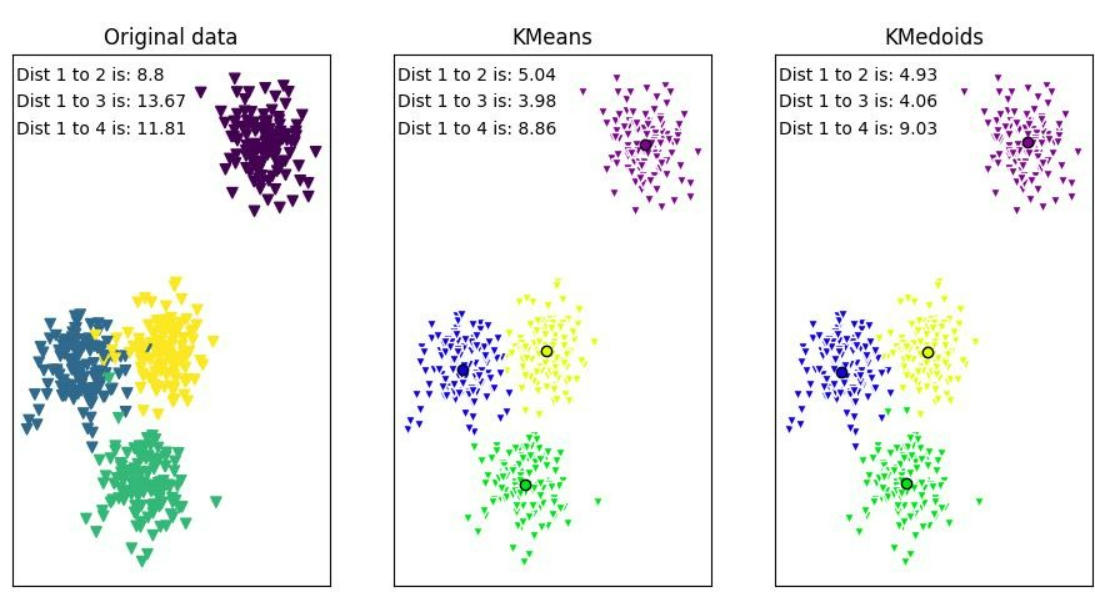
Inter-cluster: refers to the similarity or proximity between points belonging to the same cluster.

Shortcomings: it is not always infallible and can be difficult to apply in data sets with complex distribution shapes or with clusters of different sizes and densities.

**5. Let’s use the newly created modules in unsupervised to cluster some toy data**

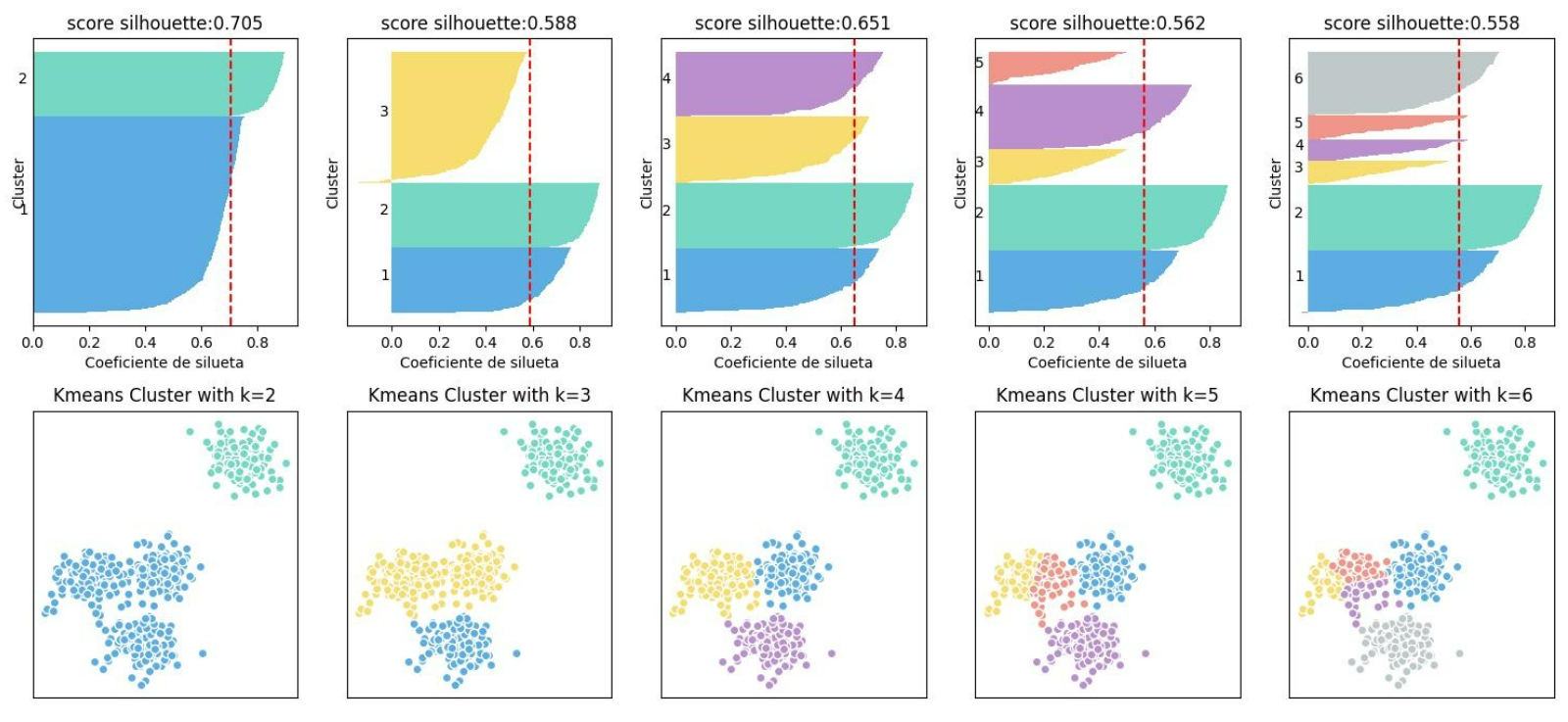
**b. Plot the resulting dataset. How many clusters are there? How far are they from one another?**

R/ Three clusters. The distance was calculated using L2 (Euclidean distance) starting from the original centers, then with k-means and k-medoids.

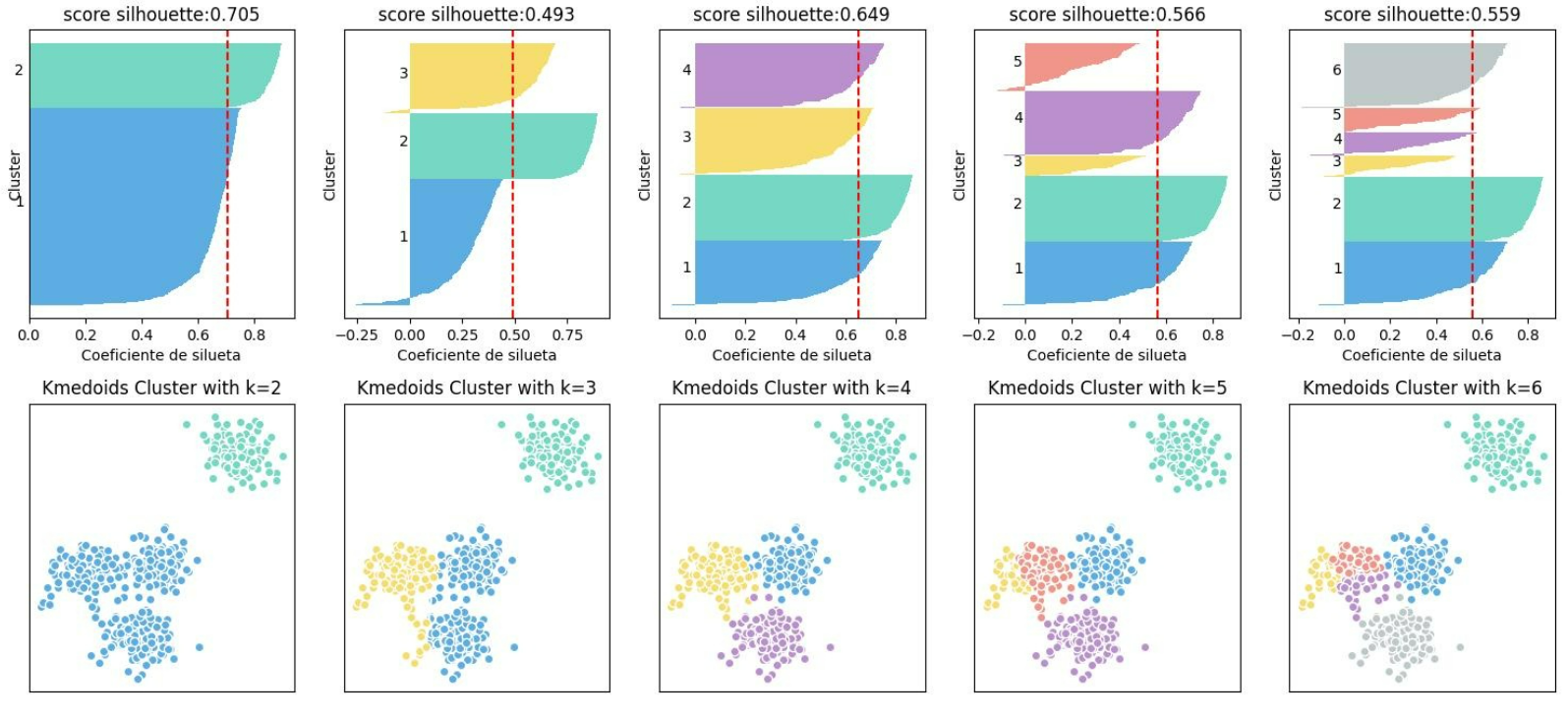


**c. For both k-means and k-medoids (your implementations), calculate the silhouette plots and coefficients for each run, iterating K from 1 to 5 clusters.**

k-means



k-medoids



1. **What number of K got the best silhouette score?**

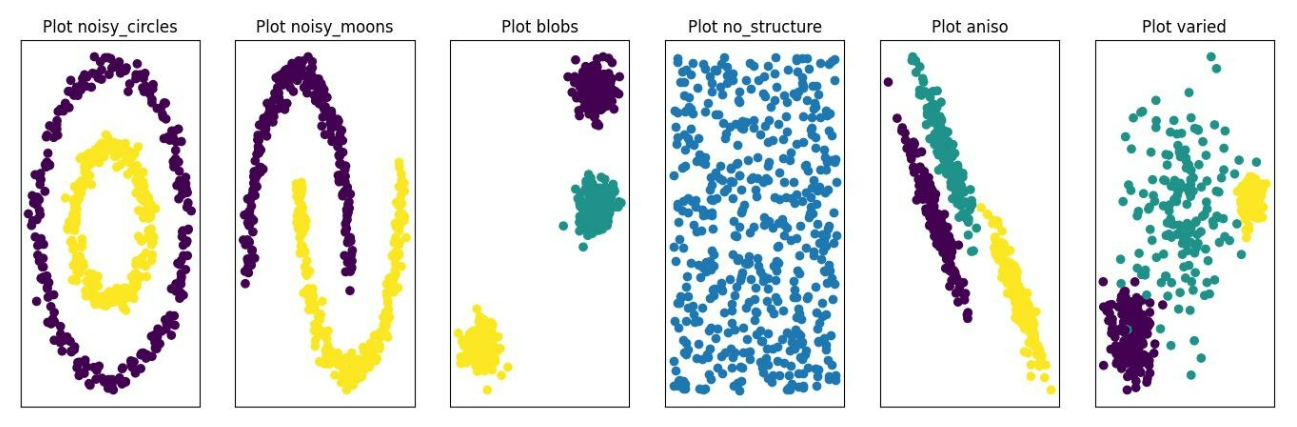
R/ K=2 obtained the best result in both k-means and k-medoids with 0.705.

**What can you say about the figures? Is this the expected result?**

R/ On the figures it can be observed that the best results were obtained with k equal to 2 and 4, the latter being the original group. In general the grouping of the clusters is good in almost all the graphs the silhouette passes the threshold and there are few erroneous assignments. Yes it was the result I expected especially for k equals 2 since in the graph the points on the left with those on the right are well separated so a good clustering is obtained. The results between k-means and k-medoids are very similar, but k-means did a bit better.

**6. Use the following code snippet to create different types of scattered data:**

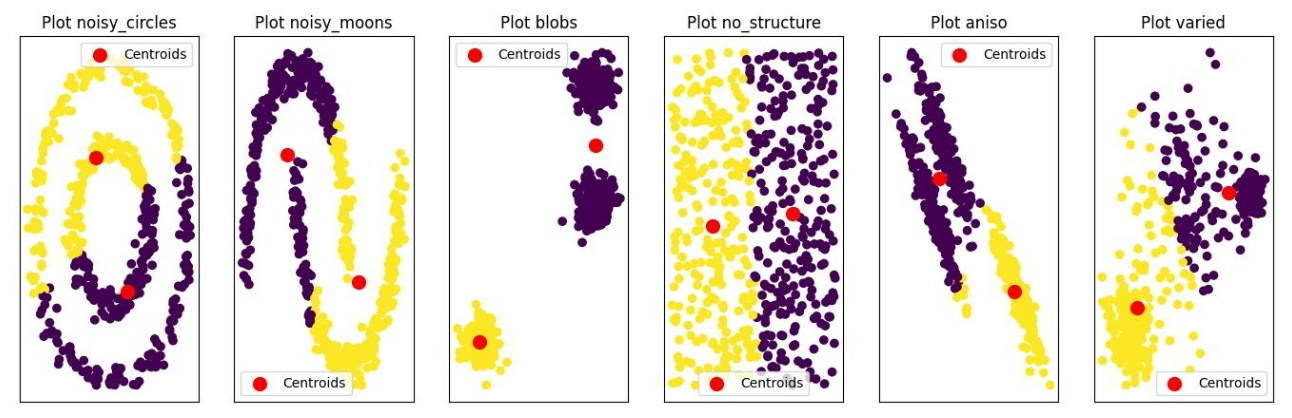
**a. Plot the different datasets in separate figures. What can you say about them?**

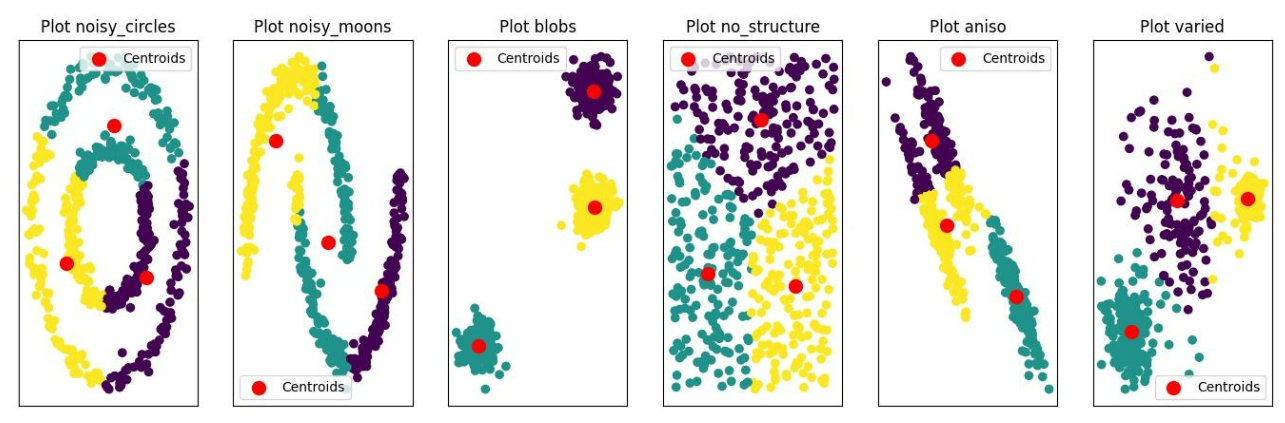


R/ Different graphs, the first two noisy circles and noisy moons with nonlinear data, blobs and aniso more structured data and with a marked separation, the unstructured and varied graphs have scattered data.

**b. Apply k-means, k-medoids, DBSCAN and Spectral Clustering from Scikit-Learn over each dataset and compare the results of each algorithm with respect to each dataset.**

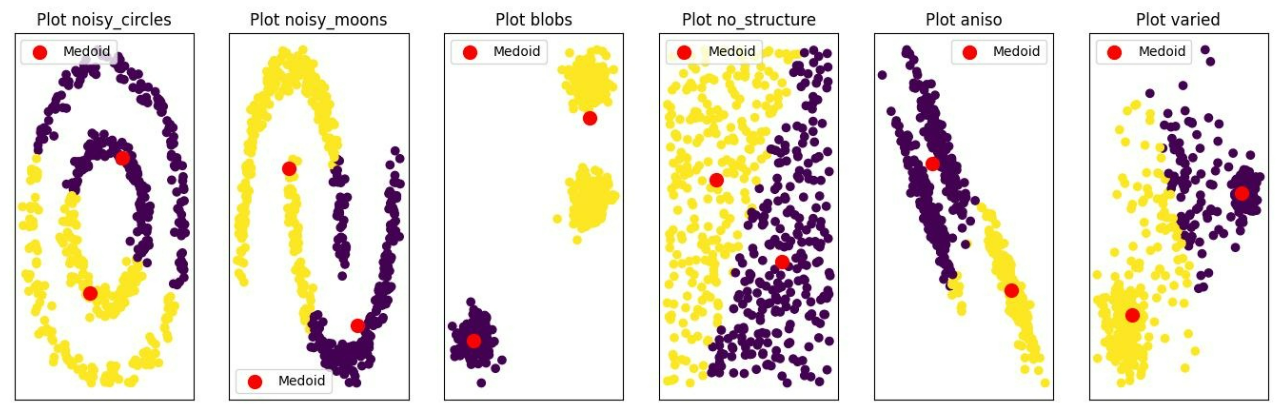
K-means cluster for k=2

K-means cluster for k=3

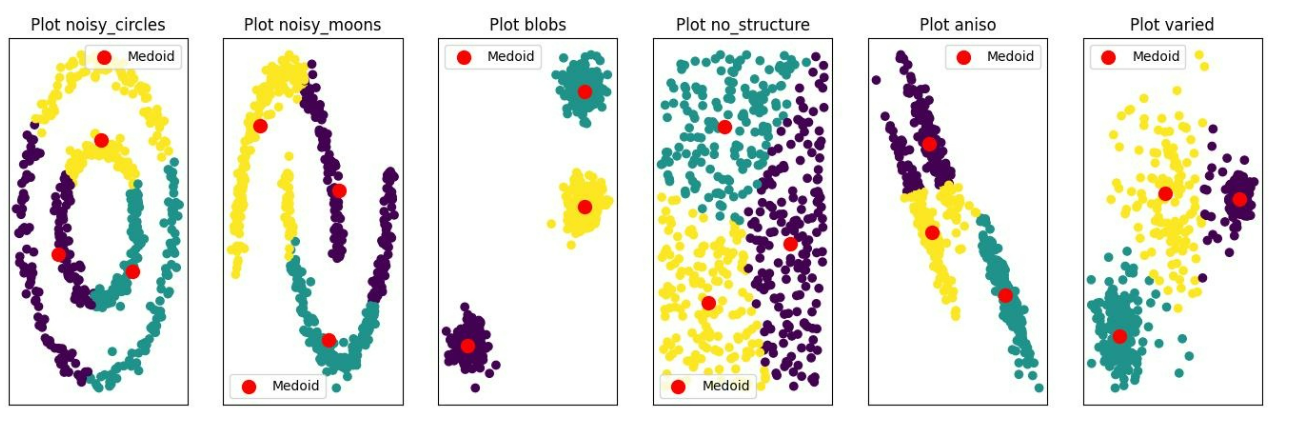


Observations: k-means had good clustering in the blobs plots, unstructured and varied, where the groups are defined and the centroids located in the average of the data. For the other plots the clusters are not very good and errors in the assignments are observed.

K-medoids cluster for k=2

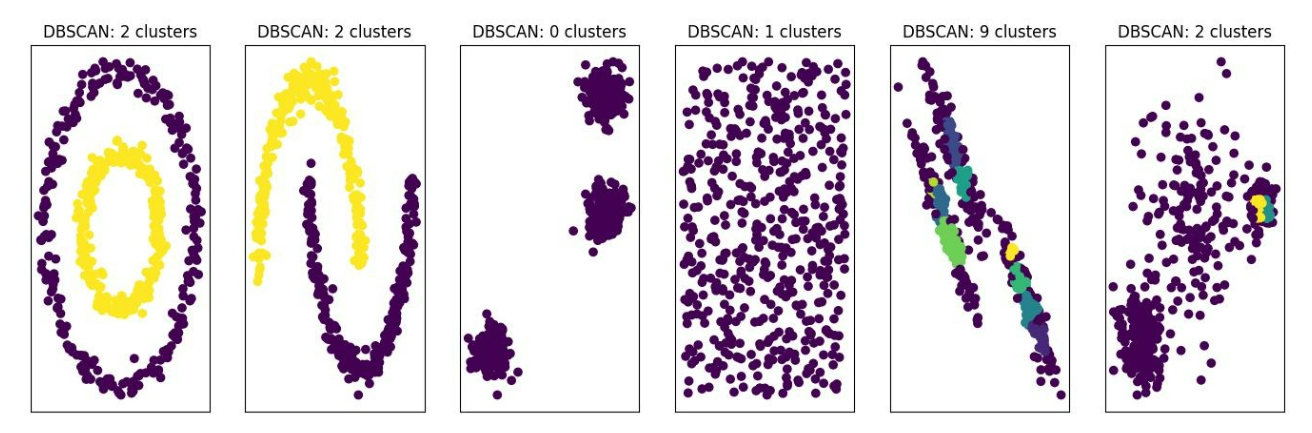


K-medoids cluster for k=3



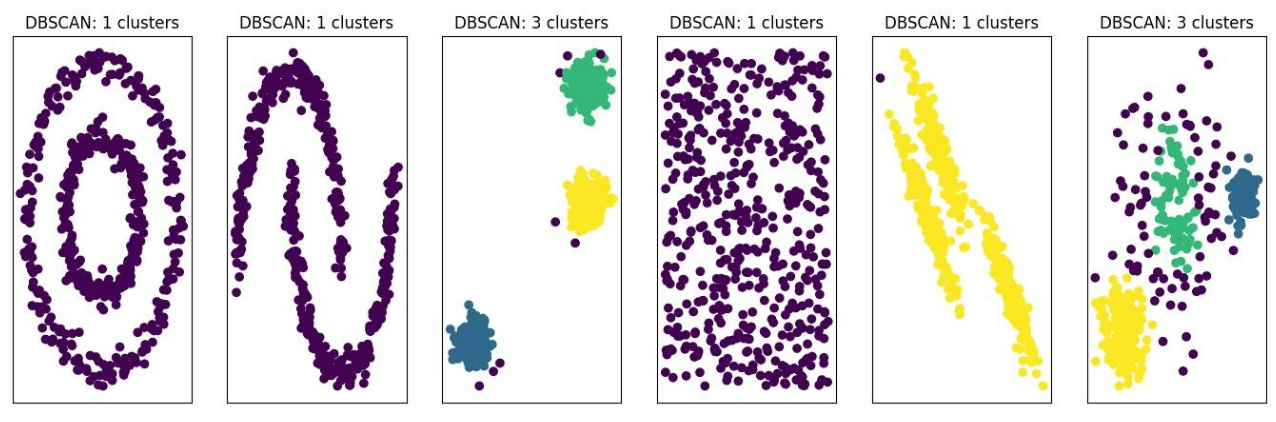
Observations: Similar to k-means, k-medoids obtained good clustering in the blobs, unstructured and variational plots, where the clusters look defined and the medoids located at the mean of the data. For the other plots the clusters are not very good and errors in the assignments are observed.

DBSCAN cluster with eps=0.2, min\_samples=10



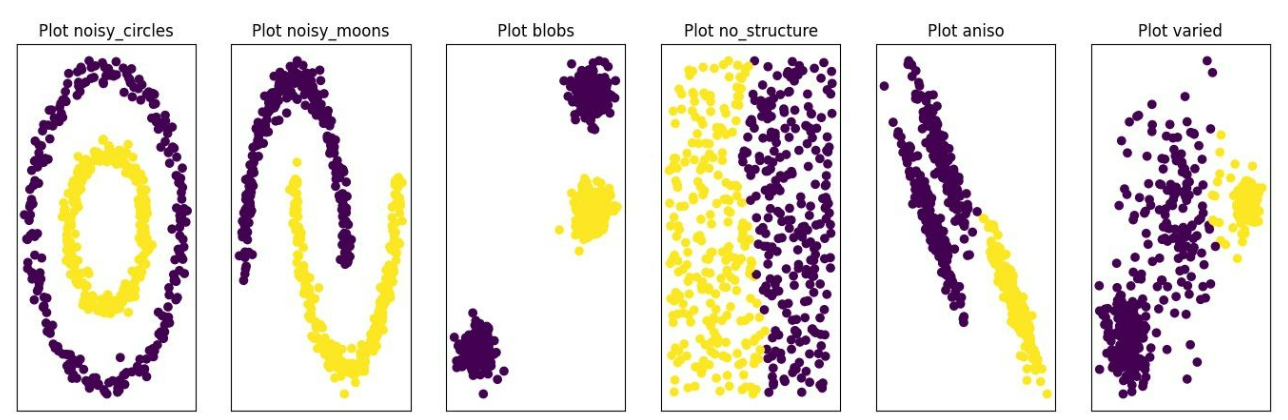
Observations: DBSCAN had very good results in the noisy circles and noisy moons plots, properly clustering the data. In the rest of the plots the results are not good and the clusters cannot be appreciated.

DBSCAN cluster with eps=1, min\_samples=10



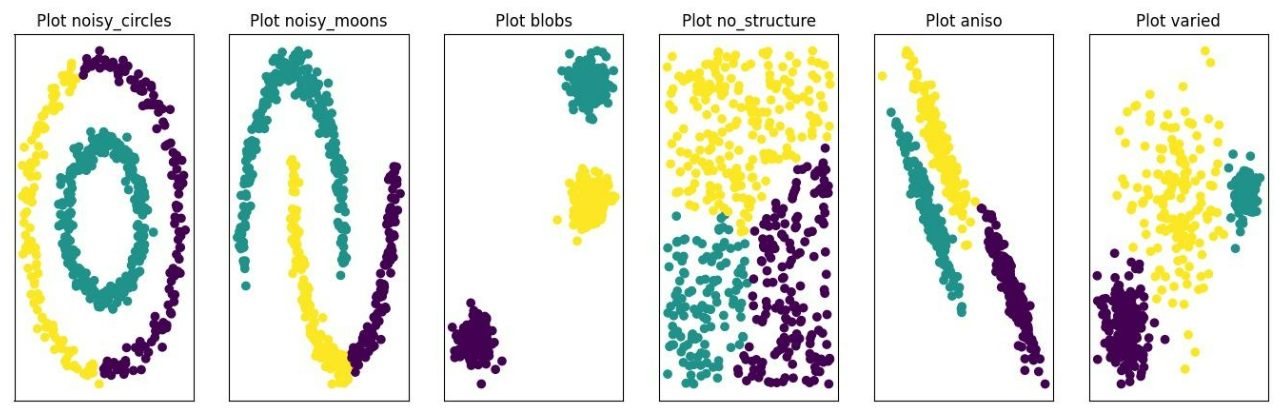
Observations: By modifying the epsilon parameter to 1, better results were obtained for the blobs and variance plots, better clustering and the elimination of "noise" in the clusters can be observed.

Spectral clustering k=2



Observations: Spectral clustering obtained very good result in general, with K=2 it can be seen that almost in all graphs it performed an acceptable clustering. Especially for the noisy circles and noisy moons data.

Spectral clustering k=3



Observations: Spectral clustering with K=3 performed a better clustering of the blobs, nested and varied data.