Introduction to Machine Learning
Assignment 5

Group 06
Lucas Pereira (s4507983) & Matteo Wohlrapp (s4974921)

October 21, 2021

Introduction

In this practical, we experiment with Hierarchical clustering. Hierarchical clustering is a method of cluster analysis, seeking to build a hierarchy of clusters. In general, there exist two types, a 'bottom up' approach, the Agglomerative clustering and a 'top down' approach called Divisive clustering. This assignment focuses only the Agglomerative clustering. The method tries to combine clusters through different linkage measurements and build up a hierarchy to analyze these clusters. Throughout this practical, we apply different techniques to find closest clusters and then merge them. The results are presented in different graphs, including a dendrogram, bar graphs and scatter plots.

Methods

To implement the Hierarchical clustering, we use MATLAB. First we load the given data which contains one set of $200x2$ double values, representing (x, y) coordinates in a $2D$ plane. The general algorithm of hierarchical clustering can be described as follows:

1. Start with each data point as a separate cluster
2. Find and merge the two closest clusters based on the linkage function
3. Repeat 2. until the desired amount of clusters is reached.

To find the two closest clusters C_1 and C_2, different linkage functions exist. In this practical we focus on the following ones:

- Single linkage: $D(C_1, C_2) = min_{x_i \in C_1, x_j \in C_2} D(x_i, x_j)$, using the closest distance between two points in C_1 and C_2
- Complete linkage: $D(C_1, C_2) = max_{x_i \in C_1, x_j \in C_2} D(x_i, x_j)$, using the furthest distance between two points in C_1 and C_2
- Average linkage: $D(C_1, C_2) = \frac{1}{|C_1|} \frac{1}{|C_2|} \Sigma_{x_i \in C_1} \Sigma_{x_j \in C_2} D(x_i, x_j)$, using the average distance between every point in C_1 and every point in C_2
- Ward's linkage:

  $$\Delta D(C_1 \bigcup C_2) = \Sigma_{x_j \in C_1 \cup C_2} D(x_j, m_{C_1 \cup C_2})^2 - \Sigma_{x_j \in C_1} D(x_j, m_{C_1})^2 - \Sigma_{x_j \in C_2} D(x_j, m_{C_2})^2$$

  , where m_i is the center for cluster i, combining cluster when the error in the sum of squares increases

To implement the functionality, we use MATLAB's built-in functions for linkage and cluster. To show the results of the linkage, 'dendrogram' is used, which depicts which clusters are merged together. After the clusters are formed, several measurements can be taken. We use the built-in function silhouette to calculate the silhouette score: $S_i = \frac{b_i - a_i}{max\{a_i, b_i\}}$, where $a_i = \frac{1}{|C_i|-1}\Sigma_{j \in C_i, i \neq j} d(i,j)$ and $b_i = min_{k \neq i} \frac{1}{|C_k|}\Sigma_{j \in C_k} d(i,j)$. As a bonus, we furthermore calculated the within clutster sum of sqares (WSS) and the between cluster sum of sqauares (BSS) according to the following formulas:

- $WSS = \Sigma_i \Sigma_{x \in C_i} (x - m_i)^2$, where m_i is the representative point for cluster C_i
- $BSS = \Sigma_i |C_i| (m - m_i)^2$, where m_i is the representative point for cluster C_i, $|C_i|$ the size of the cluster and m the mean of the cluster centers

Results

In this section we present the results of our program performing Hierarchical clustering. First we describe the dendrograms for the different linkage measures, depicting which clusters are merged. Afterwards a plot about the original data vectors is described before we explain the clustering for every linkage method for different amounts of prototypes, $K = 2, 3, 4$. In the end, we describe the silhouette scores, again for the four different linkage measures and different amounts of clusters, $K = 2, 3, 4$, in a table and a bar graph showing the relation between BSS and WSS for the 'ward' linkage measure and different amounts of prototypes $K = 2, 3, 4$.

0.1. Dendrograms

In this section the dendrograms for the four different linkage functions introduced before are described. When looking at the graphs, *Figure* 1 to *Figure* 4, the 'single' linkage function looks unbalanced, with only one bigger cluster emerging, while two other clusters do not connect with others at all. In comparison, the dendrograms for the 'complete', 'average' and 'ward' linkage function are better balanced. However, the main difference is that with the 'ward' linkage measurement, four clusters have significantly lower proximity than two clusters, while with 'complete' and 'average', the difference is smaller.

0.2. Clusters

In the following section, plots concerning the clustering of the data vectors, shown in *Figure* 5, are described. We created graphs for each linkage method and furthermore vary the number of clusters from $K = 2$ to $K = 4$.

0.2.1. Single linkage

Figure 6 to *Figure* 8 show how the original data vectors are clustered using the 'single' linkage measurement. You can see, that using this measurement, the clusters are not matching the actual clusters we see from observing the data, but rather form one big cluster and several other single values, depending on the number of clusters K.

0.2.2. Complete linkage

Figure 9 to *Figure* 11 show how the data vectors are clustered using the 'complete' linkage measurement. It is observable, that even with 4 clusters, the amount of data vectors in one cluster is always greater than one. You can also see that, especially for lower vales of K, the vectors in the center and the lower right part of the plane are building one group. Only when four different clusters are introduced, this area gets split up.

0.2.3. Average linkage

Figure 12 to *Figure* 14 show how the data vectors are clustered using the 'average' linkage measurement. As with the 'complete' linkage, the group sizes are distributed more equally. Also, the area in the middle is bundled with vectors from another area, just like in the complete linkage, however, this time the center data vectors form a group with the vectors of the upper left until this group is split up at $K = 4$.

0.2.4. Ward's linkage

Figure 15 to *Figure* 17 show how the data vectors are clustered using the 'ward' linkage measurement. This linkage function has a similar effect on our data set, with the center first being assigned to the upper left and then splitting into separate groups for $K = 4$. Overall, the distribution and accuracy of 'ward' is equally good as the one of 'average', with a clustering similar to the observable grouping of the data.

0.3. Silhouette score

In the following section, a table concerning the average silhouette scores of the data clustering is described. *Table* 1 is comprised of the average silhouette score for every introduced linkage measure, with $K = 2, 3, 4$. You can see, that for 'single' linkage the average silhouette score drops for higher values of K, indicating bad matching which is also observed in *Figure* 6 to *Figure* 8. In comparison, the scores for the 'average' and 'ward' linkage measure are higher for greater K. This is also observed in *Figure* 12 to *Figure* 17, where the two linkage functions provide good clustering with increasing K. 'Complete' gets a slightly lower score for higher K but remains mostly stable.

0.4. WSS and BSS

As a bonus, we included a graph, *Figure* 18, depicting the WSS and the BSS for different numbers of clusters, $K = 2, 3, 4$. The graph shows, that for higher numbers of prototypes, the WSS, meaning the within cluster sum of squares decreases, while the between cluster sum of squares, the BSS, increases. Also, the BSS is overall higher than the WSS.

Discussion

Given the data, we observe a lot of noise, as seen in *Figure* 5, which makes 'single' linkage suffer to do proper classification. When analysing the dendrogram, *Figure* 1, and the resulting clusters, *Figure* $6 - 8$, it is made clear that the 'single' method, for this data, does not generate any reasonable clustering. This is also indicated by a decreasing silhouette score for higher K shown in *Table* 1, which generally suggests bad matching.

For 'complete' linkage the dendrogram, *Figure* 2, is much more promising, showing reasonably, and somewhat evenly, sized clusters for any value of K. Looking over at its graphs, *Figure* $9 - 11$, we find some of the problems of the linkage measurement, in that it shows somewhat in-satisfactory behavior when used in areas with mixed data. In *Figure* 9 and *Figure* 10, with $K = 2$ and $K = 3$ respectively, the 'mixed' area seems to be split in a way that does not make much sense, as it seems more logical, that the points in the 'mixed' area are clustered together with adjacent groups, at e.g. the upper left. Still, it performs better than 'single', which is also indicated by the silhouette score which is higher for every K when comparing 'complete' to 'single'.

The same cannot be said for 'average' linkage in our data set. Firstly, it's dendrogram, *Figure* 3, seems to present a much clearer image of how the clustering can be better represented, that being either with $K = 3$ since the distance between it and $K = 2$ is more pronounced, meaning that the fusion of two clusters C from $K = 3$, would include considerable higher distance measures. When

looking at the graph, *Figure* $13-14$, it confirms the hypothesis, since 1 and 2 are very distinct clusters. In the graph of $K=3$, *Figure* 13, the clustering seems to follow exactly what would be expected when looking at the point distribution, having a cluster for each dense area of points and a third cluster for the more scattered portion. When looking at an additional cluster $K=4$, *Figure* 14, as the dendrogram suggested by its short distance between 3 and 4, the clustering in the mixed area has not much influence on the overall performance.

Finally, 'ward's' linkage presented a dendrogram, *Figure* 4, that seems to favor both $K=3$ and $K=4$ as a good numbers of clusters. When it comes to the results on the graph, *Figure* $15-17$, they seem to behave similar to 'average', dividing the graph into reasonable clusters. In general, 'ward' is better than 'average' when it comes to dealing with more scattered data, but, to our understanding, since this data set had both a scattered area and more denser areas, the performance of both became somewhat similar, which is also indicated by similar silhouette scores.

When it comes to the silhouette values *Table* 1, one can argue, that $K=3$ is the optimum amount of clusters for both 'average' and 'ward', since in both linkages, the highest average silhouette value is found at $K=3$, meaning that, on average, the points in the clusters are much closer to members of their own cluster than to those outside. For the values for 'single' linkage it is even more evident how bad the clustering for this method performs in this data set with a sharp decline in value when adding more clusters, since every cluster other than the first has only one member, thus severely skewing the results. A similar problem can be seen in 'complete' linkage, since it tends to split the sparse portion when clusters are added, many more points are closer to points of other clusters than their fellow cluster points, thus diminishing the value.

As a bonus we calculated the BSS and WSS values, *Figure* 18, with the 'ward' linkage. As one can imagine, when we add more well distributed clusters, the value of WSS decreases while the one for BSS increases. WSS, withing cluster sum of squares, measures the cohesion of the points in a certain cluster. When we add more clusters the tendency is that each cluster will focus more and more on a smaller area of points, increasing the cohesion and as a result decresing the WSS. At the same time BSS, between clusters sum of squares, will then increase, as the increase in the number of clusters increases the isolation of points and therefore increases the BSS.

Workload

The workload was split equally, with Lucas Pereira and Matteo Wohlrapp working both on code and report design. Issues and problems were solved in a collaborative manner, with both team partners contributing evenly.

A. Assignment 4

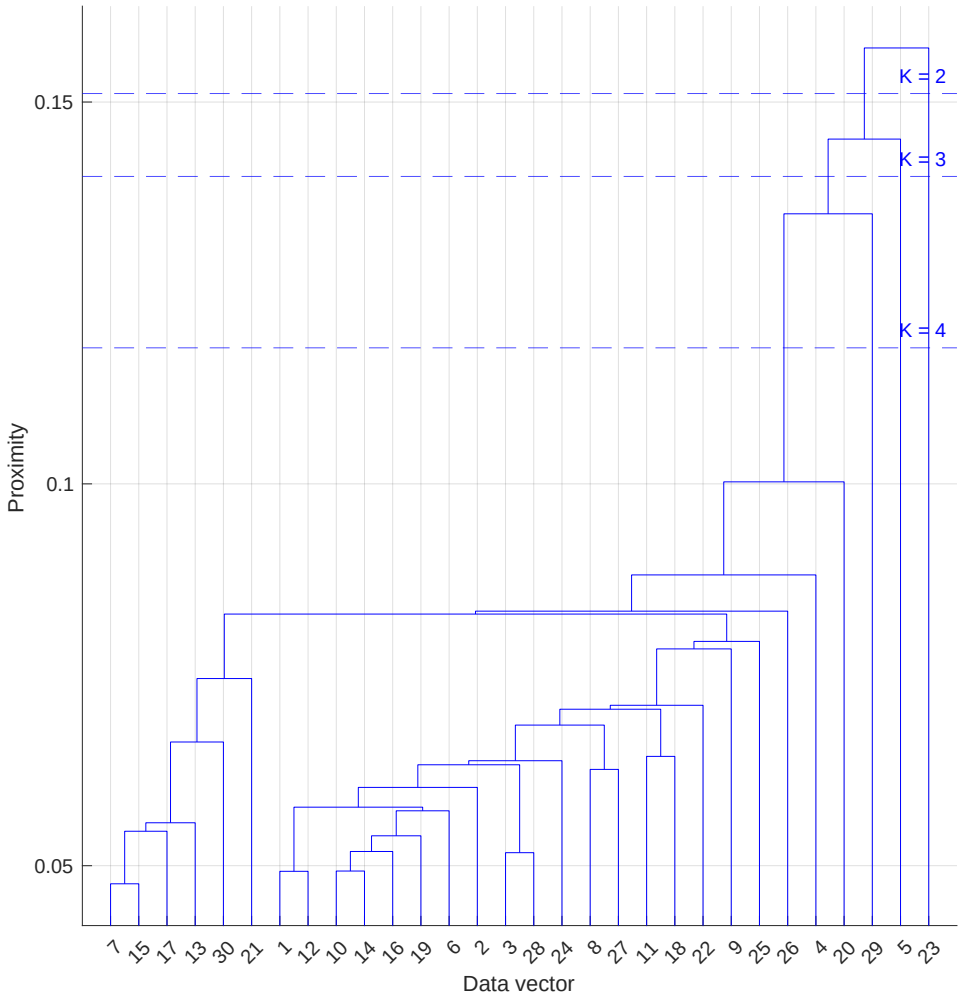

Figure 1: *Dendrogram of the 'single' linkage function*

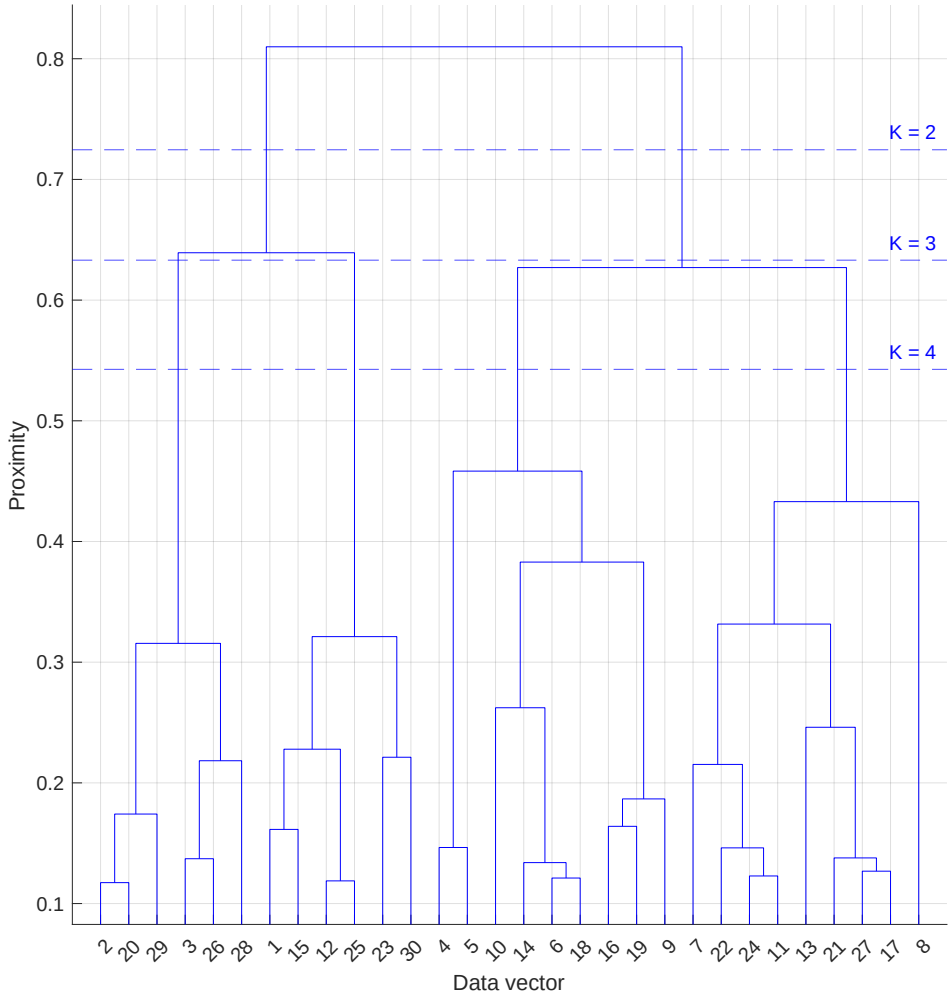

Figure 2: *Dendrogram of the 'complete' linkage function*

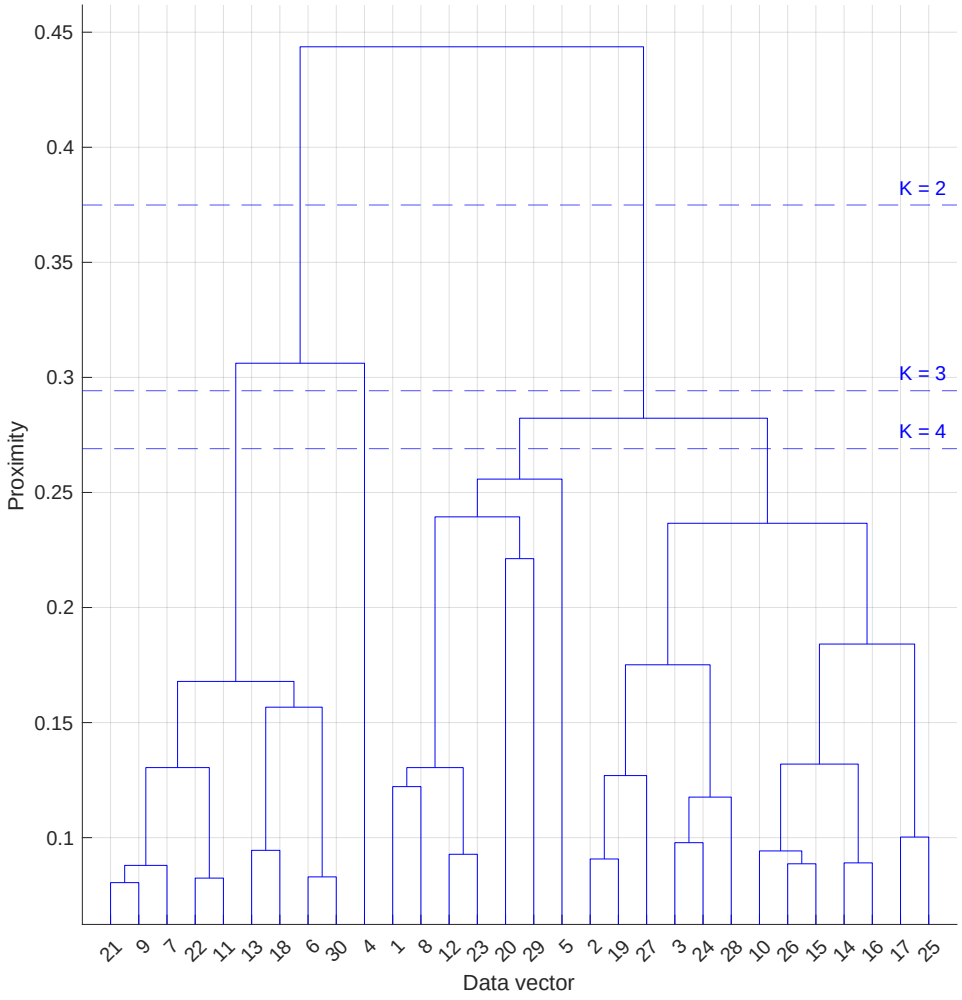

Figure 3: *Dendrogram of the 'average' linkage function*

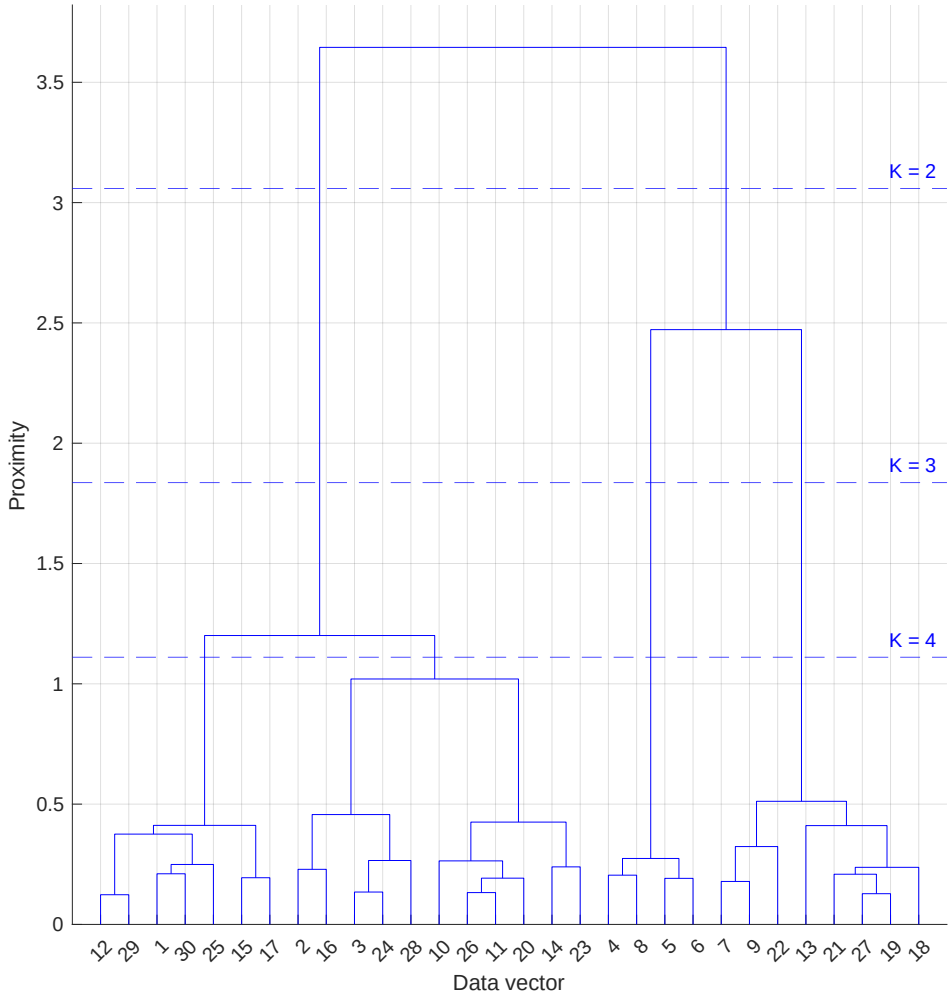

Figure 4: *Dendrogram of the 'ward' linkage function*

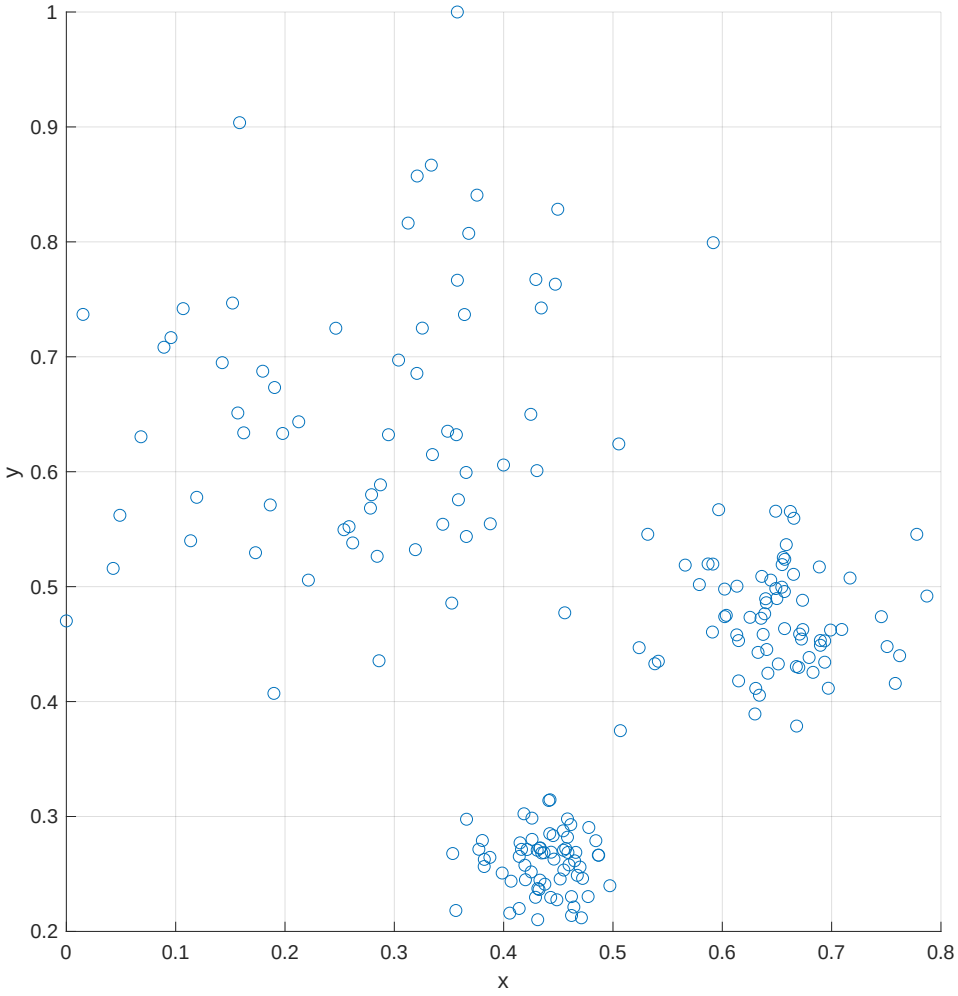

Figure 5: *Graph showing the distribution of the original data*

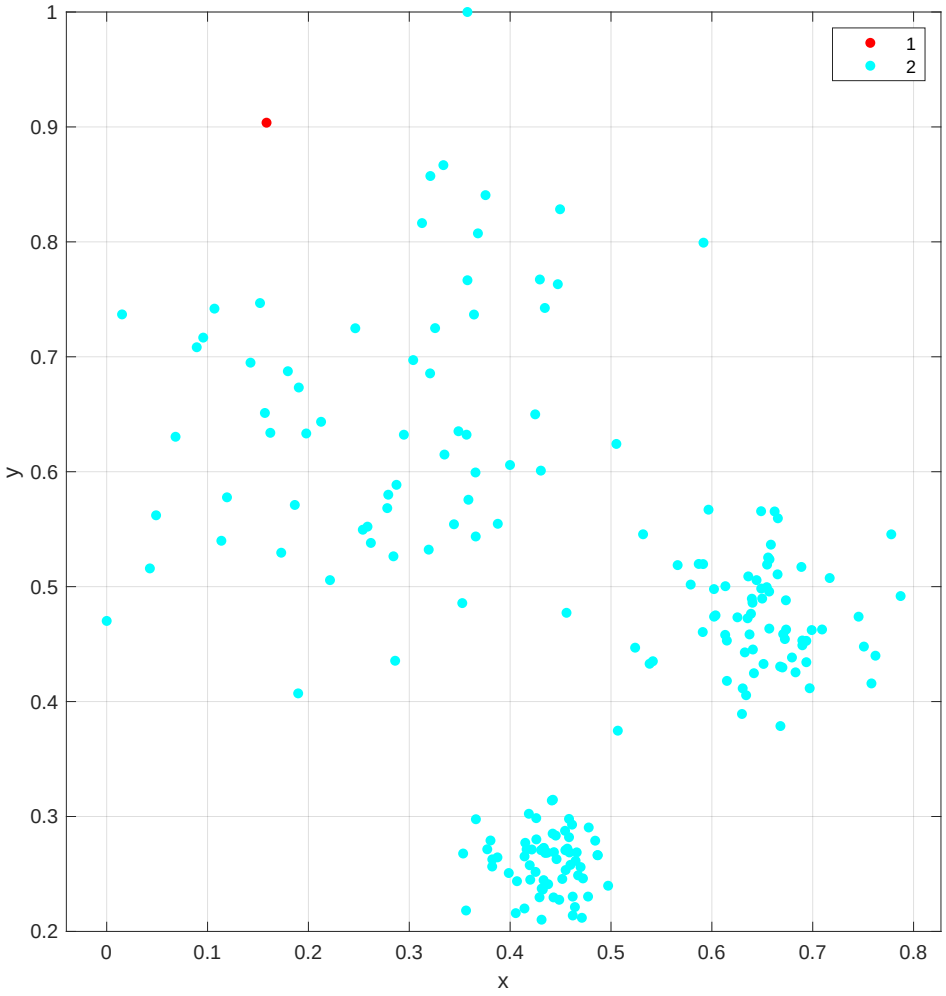

Figure 6: *Graph of the cluster for the 'single' linkage function with* $K = 2$

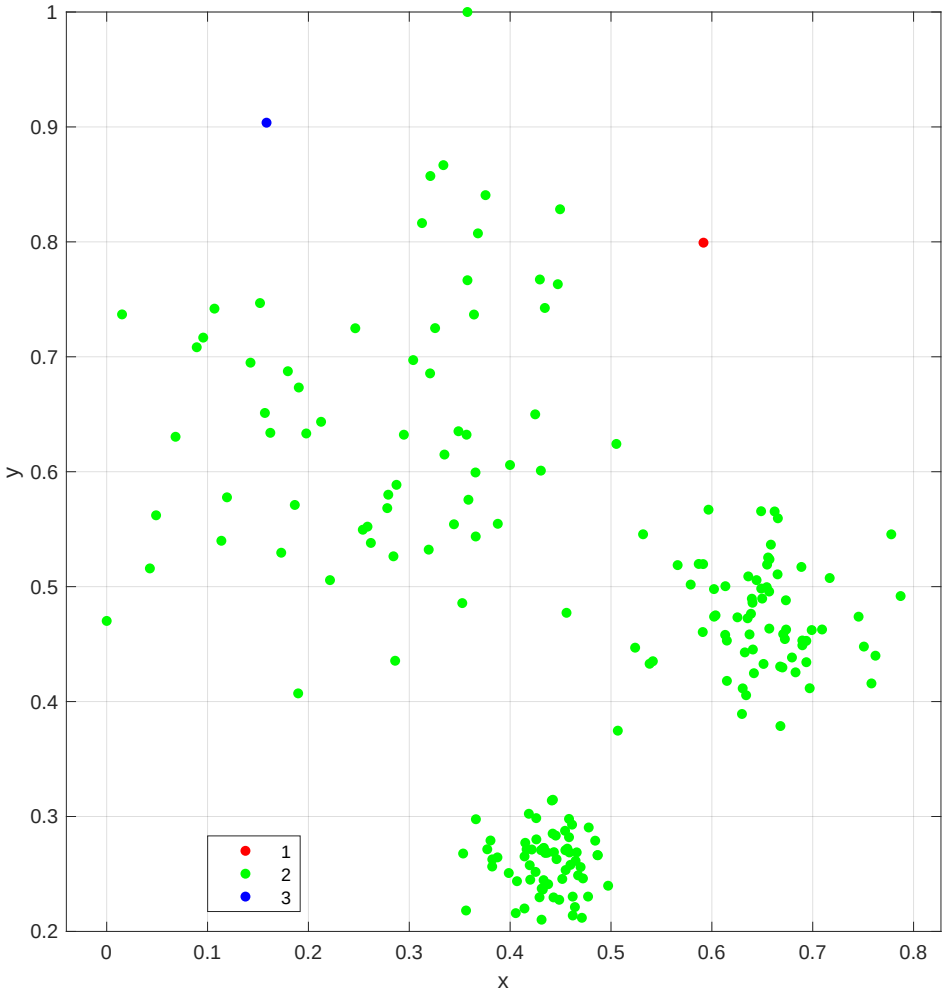

Figure 7: *Graph of the cluster for the 'single' linkage function with* $K = 3$

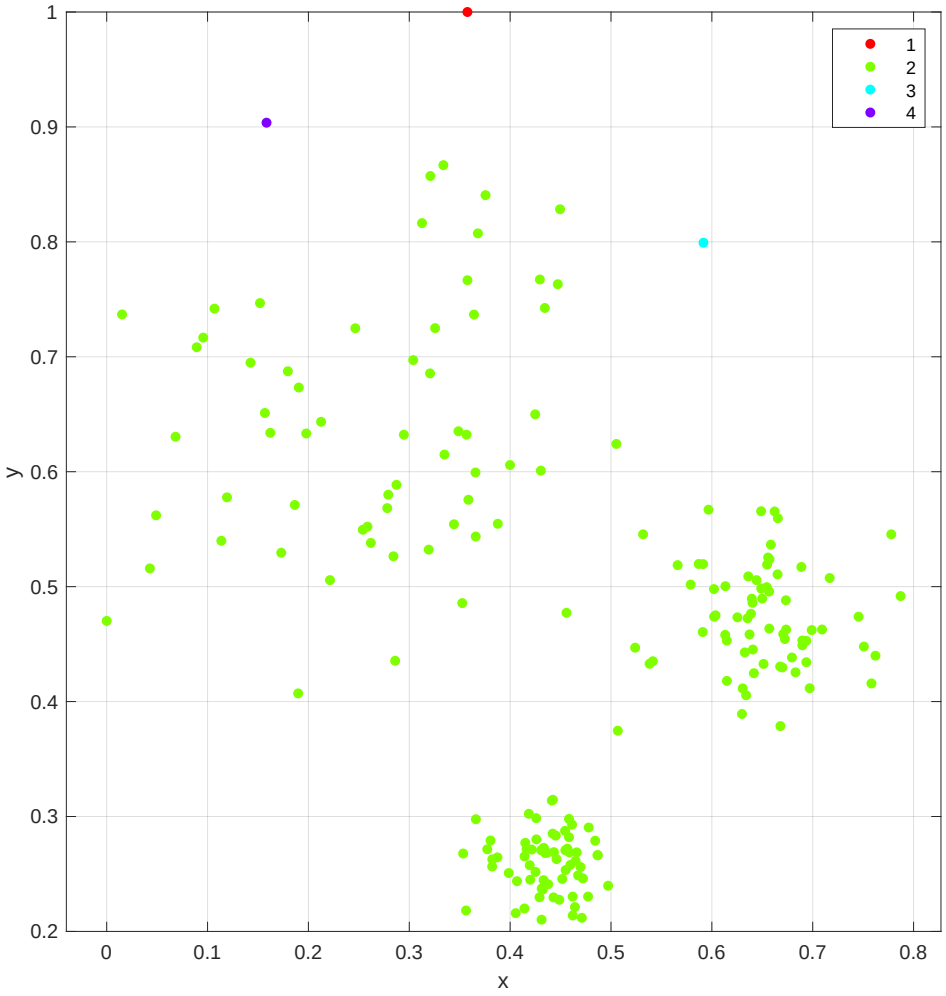

Figure 8: *Graph of the cluster for the 'single' linkage function with* $K = 4$

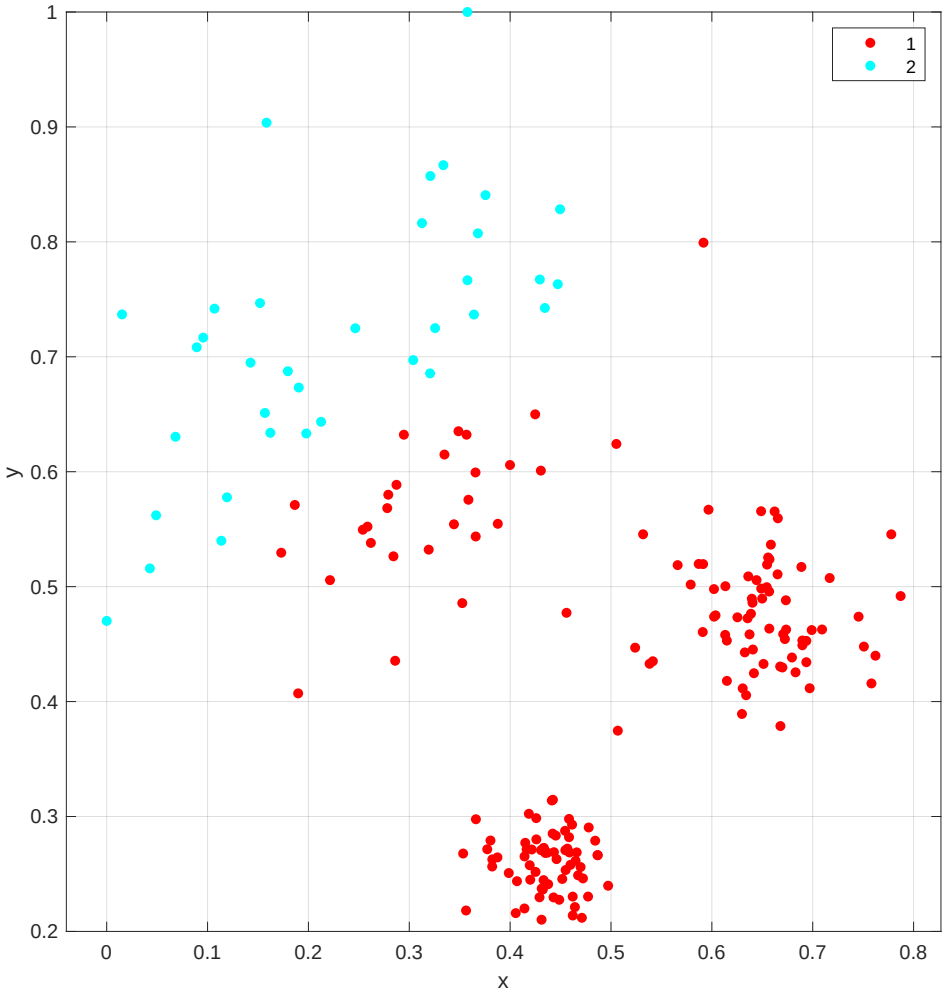

Figure 9: *Graph of the cluster for the 'complete' linkage function with* $K = 2$

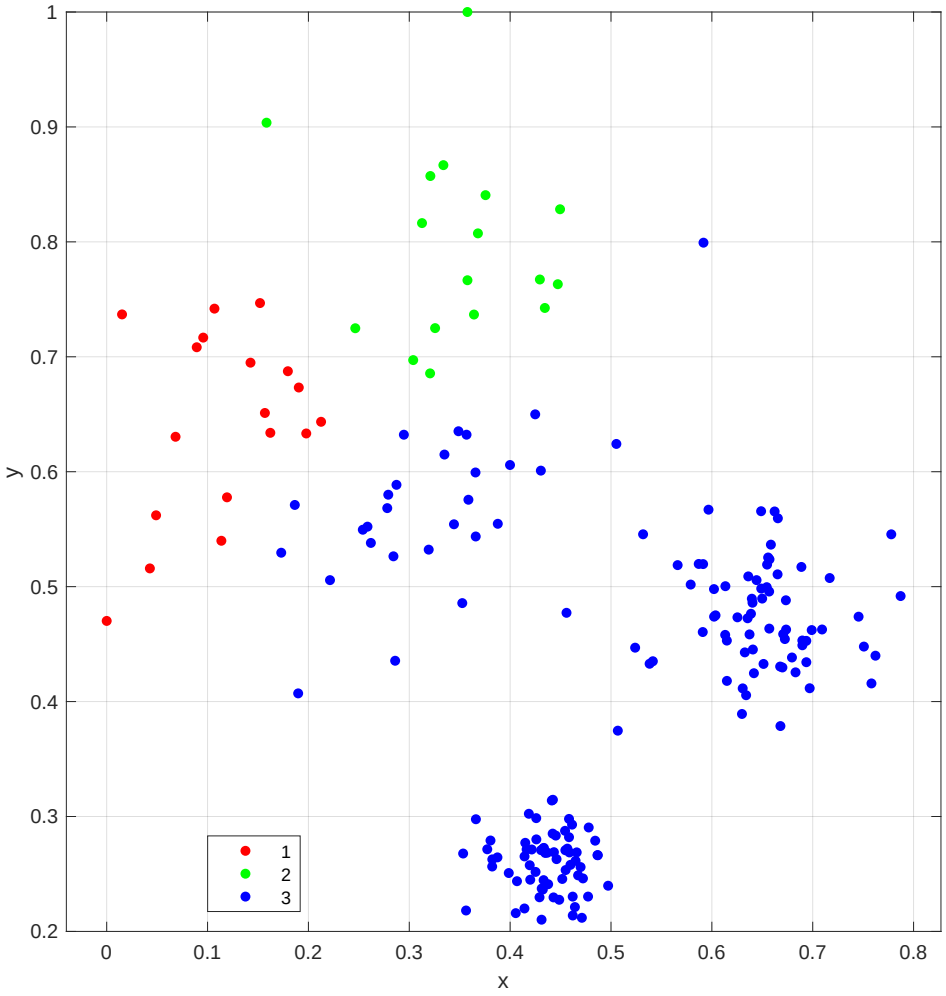

Figure 10: *Graph of the cluster for the 'complete' linkage function with* $K = 3$

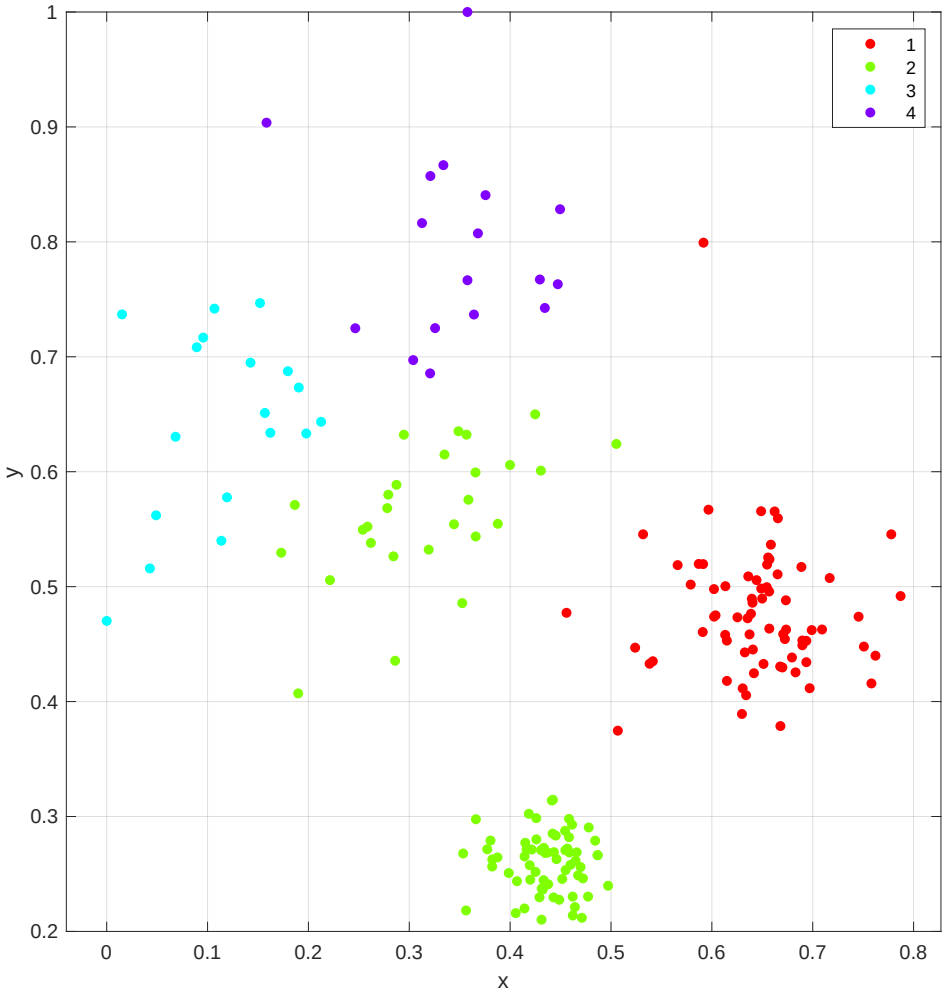

Figure 11: *Graph of the cluster for the 'complete' linkage function with* $K = 4$

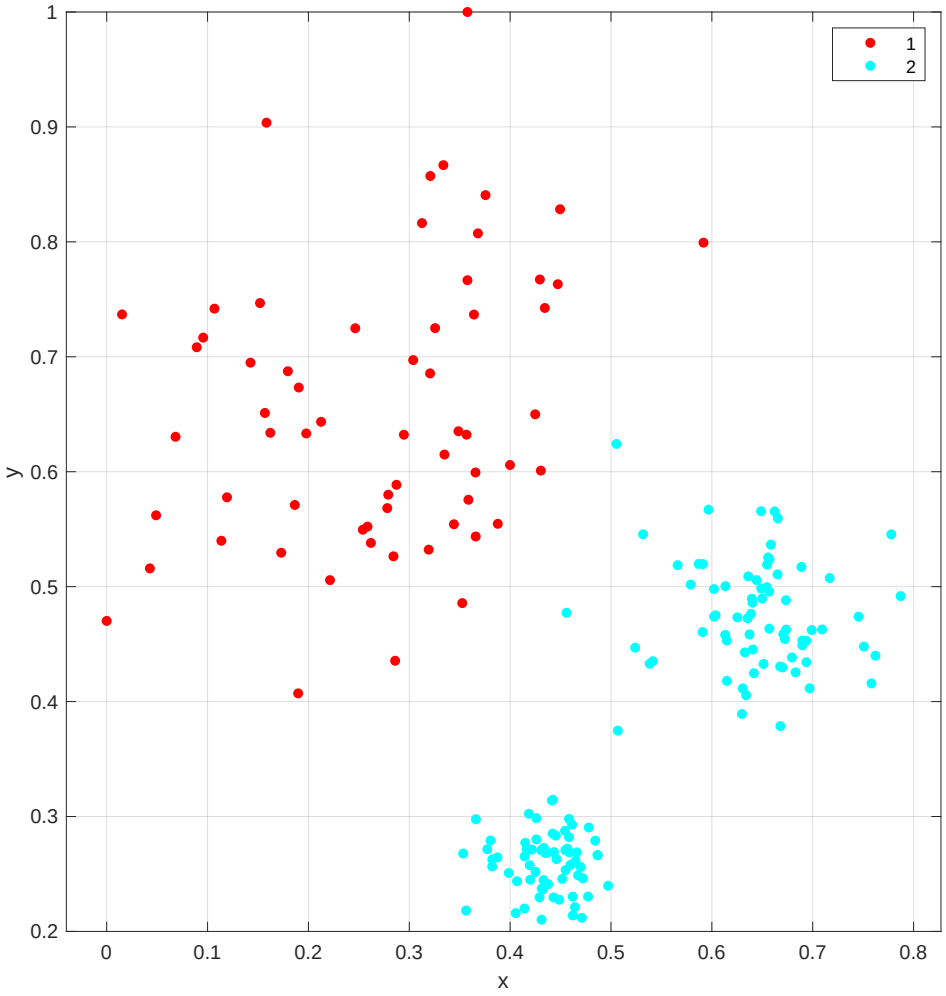

Figure 12: *Graph of the cluster for the 'average' linkage function with* $K = 2$

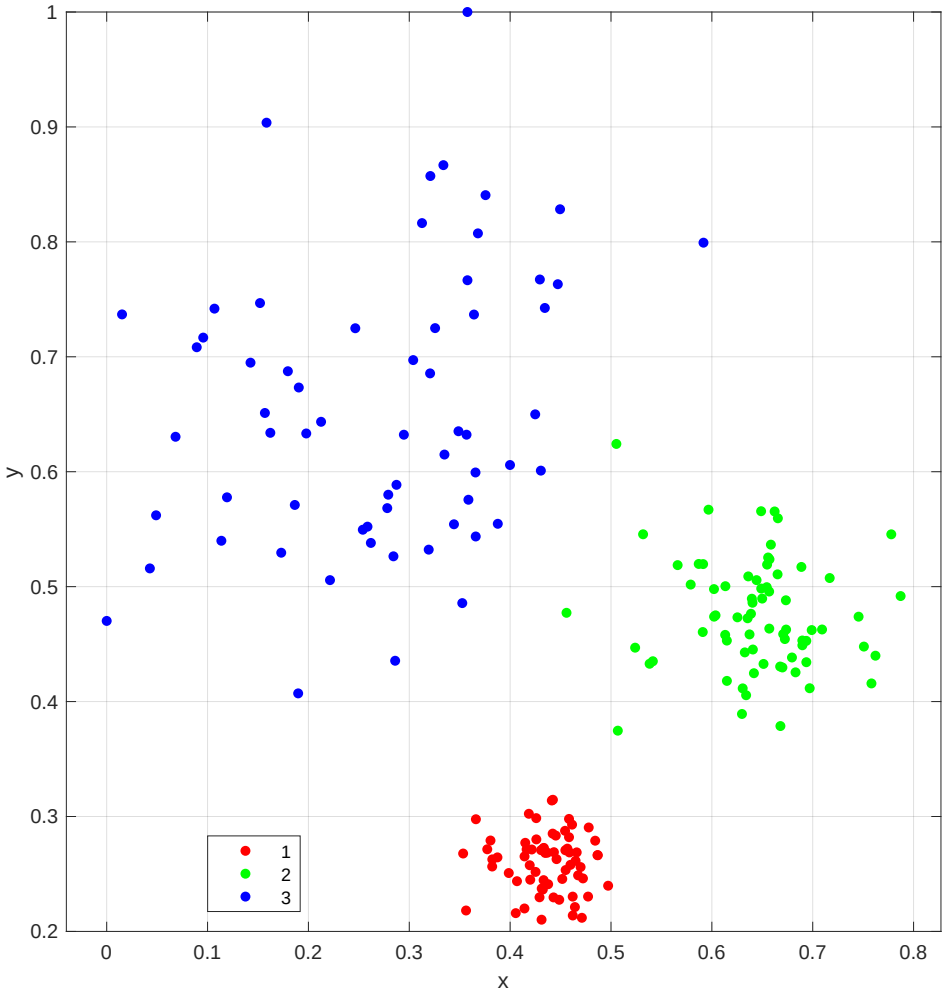

Figure 13: *Graph of the cluster for the 'average' linkage function with* $K = 3$

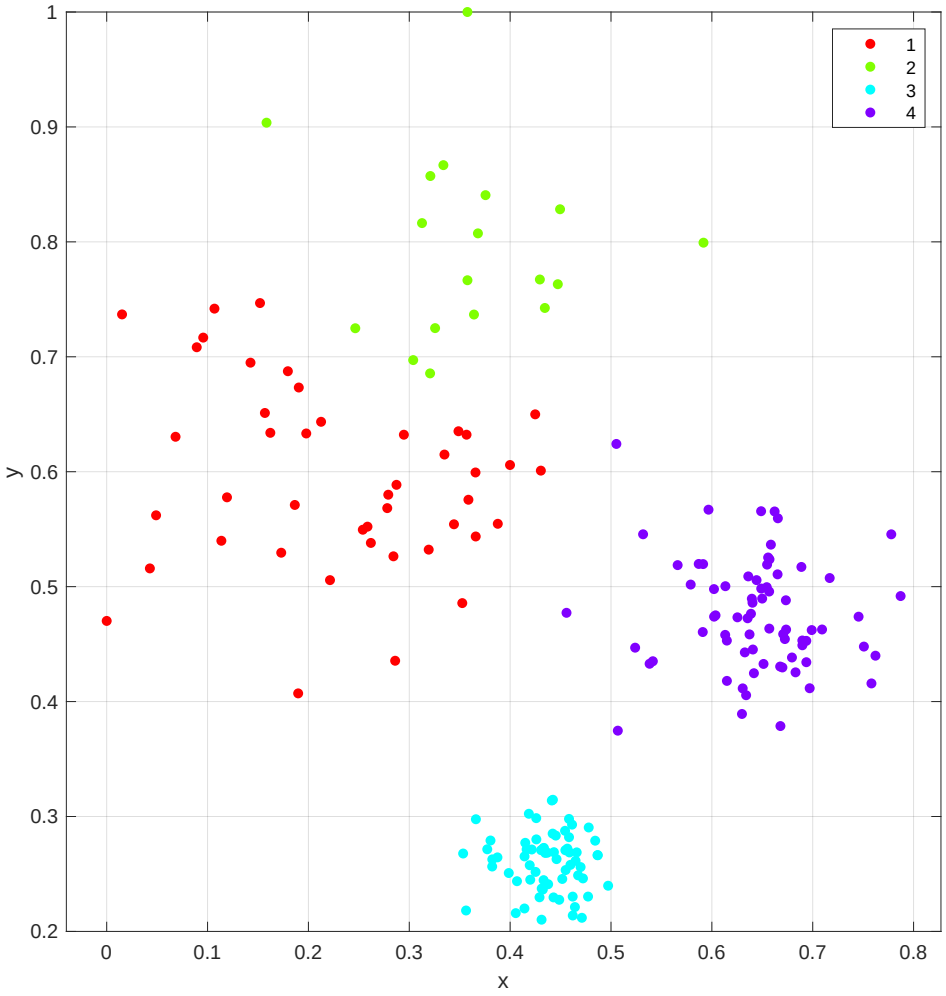

Figure 14: *Graph of the cluster for the 'average' linkage function with* $K = 4$

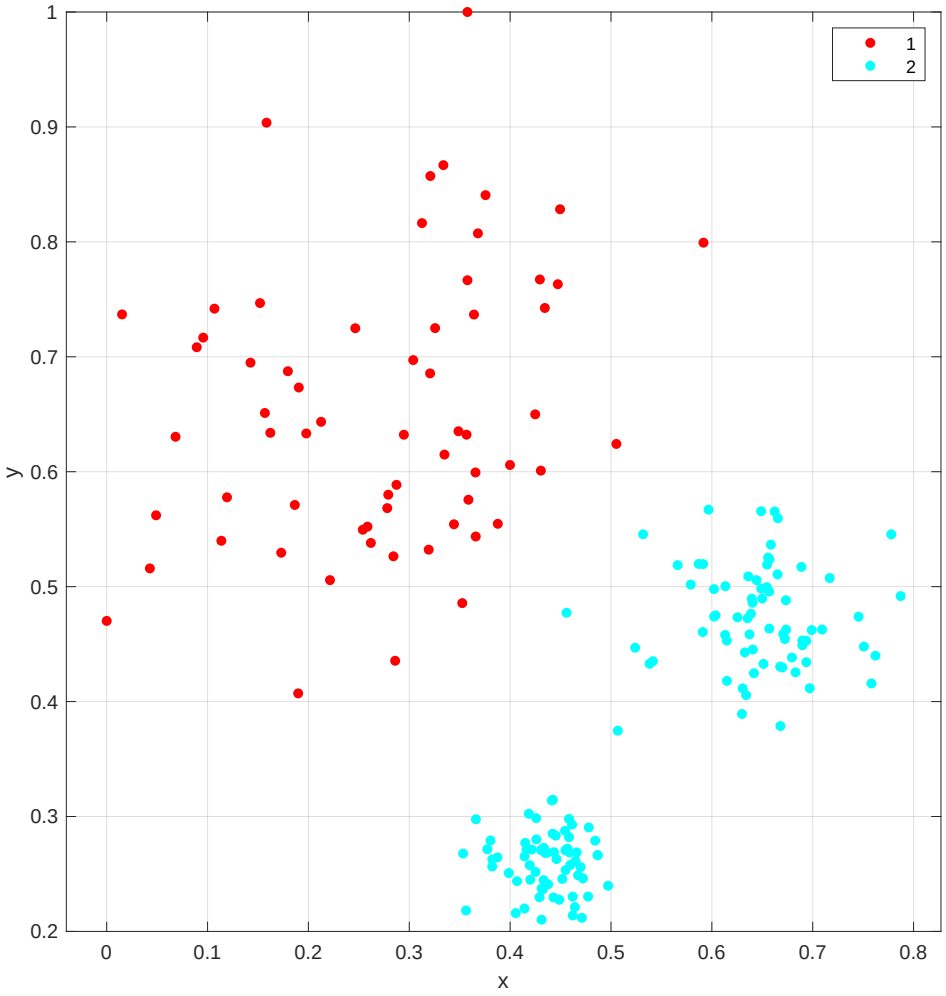

Figure 15: *Graph of the cluster for the 'ward' linkage function with* $K = 2$

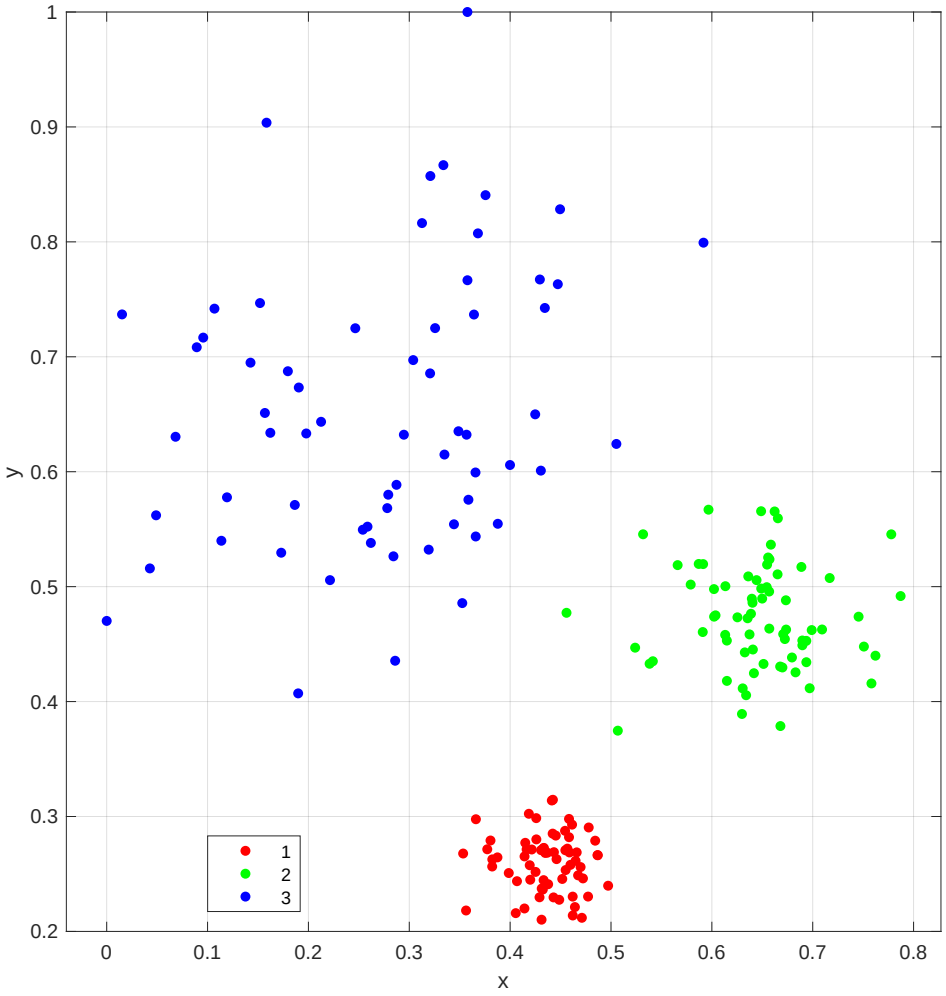

Figure 16: *Graph of the cluster for the 'ward' linkage function with* $K = 3$

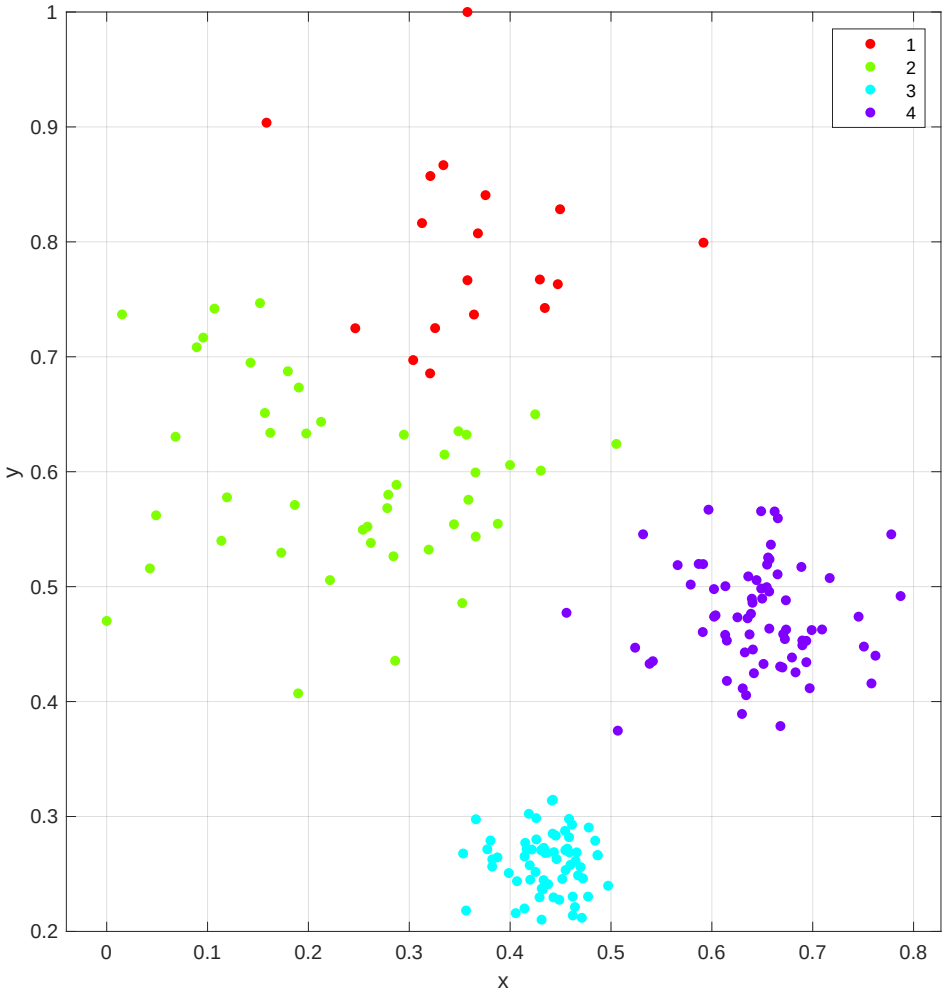

Figure 17: *Graph of the cluster for the 'ward' linkage function with* $K = 4$

K x Linkage Type	Single	Complete	Average	Ward
K = 2	0.4711	0.6101	0.7207	0.7207
K = 3	0.1669	0.5500	0.8049	0.8027
K = 4	0.1816	0.5852	0.7789	0.7744

Table 1: *Table of Silhouette values for each Linkage Type and* K *value*

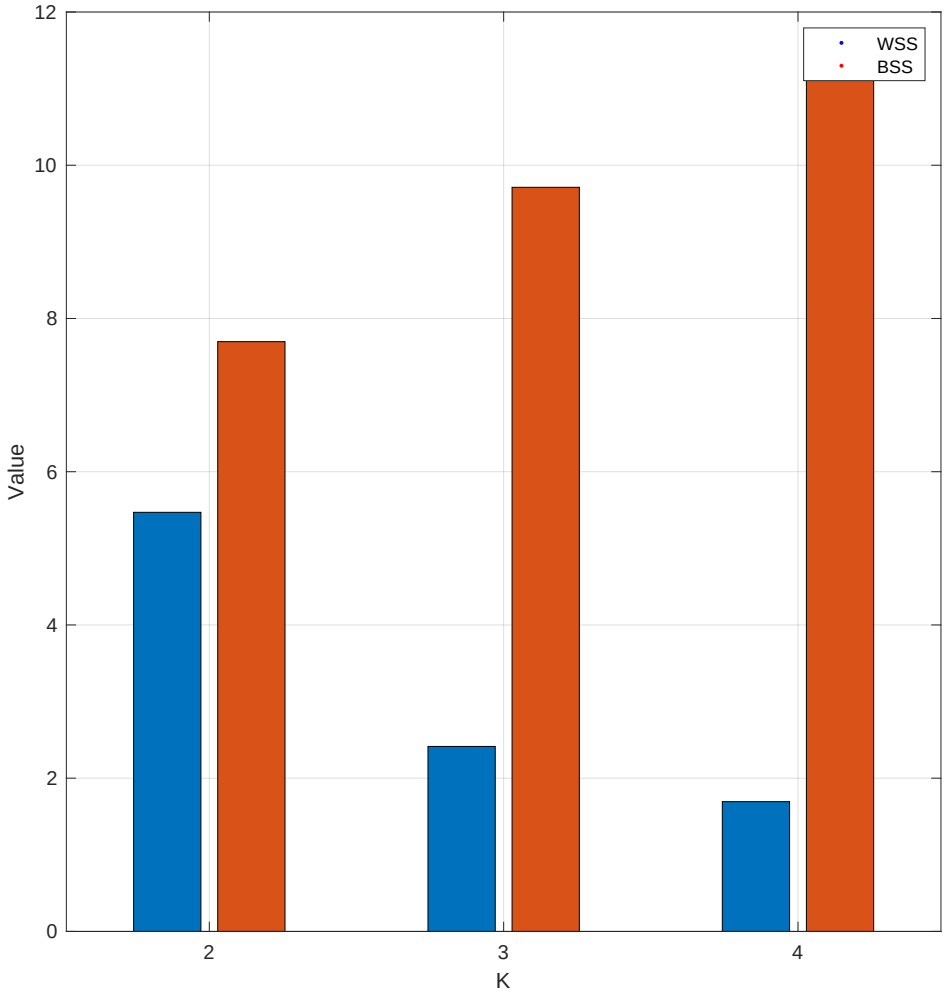

Figure 18: *Bar graph of the WSS and BSS for the 'ward' linkage function with $K = 2, 3, 4$*