Question 1. K-Means Clustering

(a) (10 points) Describe the K-means clustering algorithm. An explanation has to include a step-by-step algorithm description together with an explanation of the algorithm convergence.

The answer see on the course site Lecture3_4_Clustering

(b) (10 points) Suppose that the following data is clustered using the K-means method.

	p_1	p_2	p_3	p_4	p_5	p_6
x	2	2	5	7	1	4
y	10	5	8	5	2	10

Assume that number of clusters K = 3 and that the points are initially assigned to clusters as follows:

$$\mathbf{C1 = \{p_1, p_2\},\quad C2 = \{p_3, p_4\},\quad C3 = \{p_5, p_6\}.}$$

Perform two iterations of the K-means algorithm and explain all the operations, which have been done.

Answer:

The first iteration

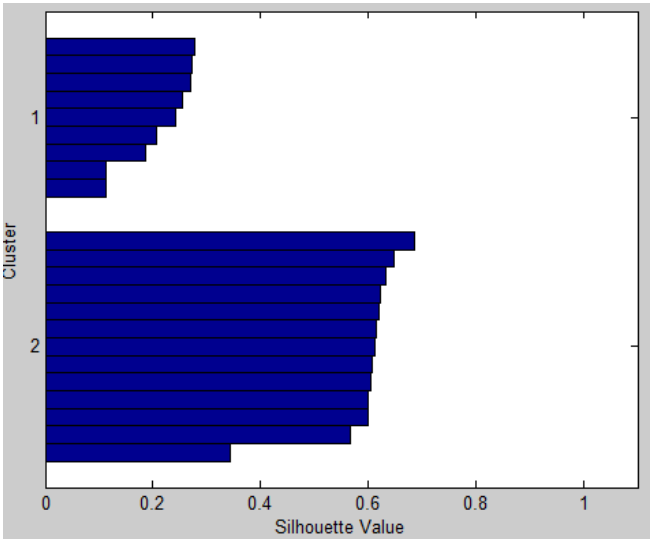
	C1={2; 7.5}	**C2={6; 6.5}**	**C3={2.5; 6}**
p_1	p_1		
p_2	p_1		p_2 C3={2.25; 5.5}
p_3	p_1	p_3	p_2
p_4	p_1	p_3, p_4	p_2
p_5	p_1	p_3, p_4	p_2, p_5
p_6	p_1	p_3, p_4, p_6	p_2, p_5
The class membership	**p_1**	**p_3, p_4, p_6**	**p_2, p_5**
	C1={2;10}	**C2={5; 8.25}**	**C3={1.5; 3,5}**

Second iteration

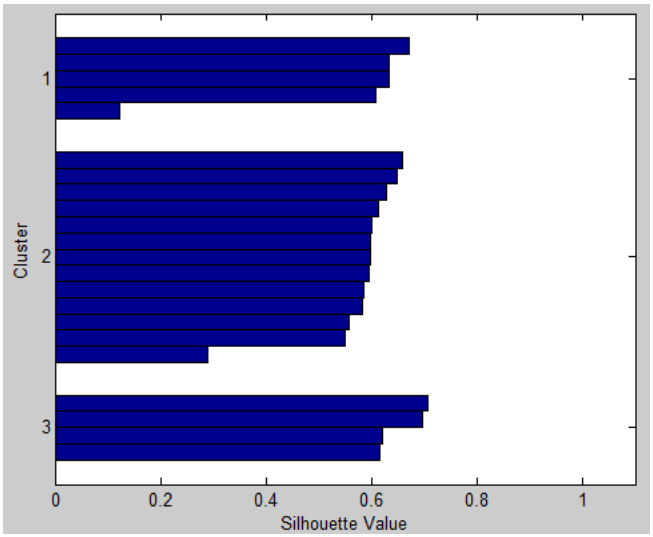
	C1={2;10}	**C2={5; 8.25}**	**C3={1.5; 3,5}**
p_1	p_1		
p_2	p_1		p_2
p_3	p_1	p_3	p_2
p_4	p_1	p_3, p_4	p_2
p_5	p_1	p_3, p_4	p_2, p_5
p_6	p_1, p_6	p_3, p_4	p_2, p_5
The class membership	**p_1, p_6**	**p_3, p_4**	**p_2, p_5**

(c) (5 points) A student wants to estimate the number of clusters in a dataset being classified using the K-means method. To this end, the data were clustered in number of clusters from 2 to 4. The histograms of the silhouette values are given below. Which number of clusters is preferred in this case? Explain your answer.

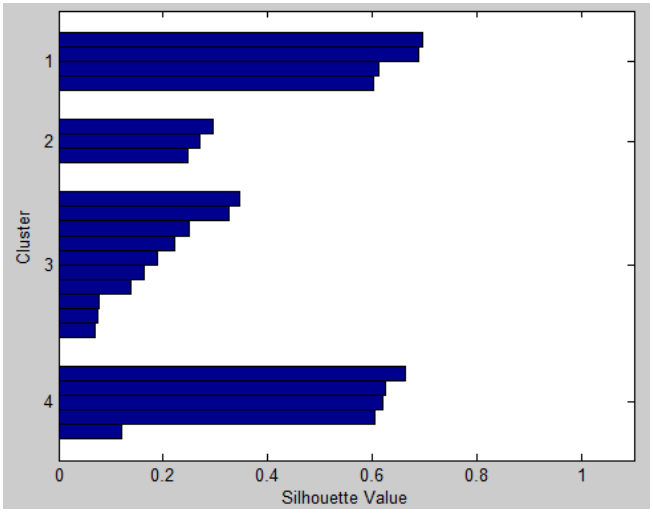
K=2



K=3



K=4



Answer: K=3