# Assignment - 2

1. Question

| tid | Itemset |
|---|---|
| $t_1$ | ABCD |
| $t_2$ | ACDF |
| $t_3$ | ACDEG |
| $t_4$ | ABDF |
| $t_5$ | BCG |
| $t_6$ | DFG |
| $t_7$ | ABG |
| $t_8$ | CDFG |

Transaction Database

(1)

| tid | A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|---|
| $t_1$ | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| $t_2$ | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| $t_3$ | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| $t_4$ | 1 | 1 | 0 | 1 | 0 | 1 | 0 |
| $t_5$ | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| $t_6$ | 0 | 0 | 0 | 1 | 0 | 1 | 1 |
| $t_7$ | 1 | 1 | 0 | 0 | 0 | 0 | 1 |
| $t_8$ | 0 | 0 | 1 | 1 | 0 | 1 | 1 |

Binary Database.

(2.)

$t(x)$

| A | B | C | D | E | F | G |
|---|---|---|---|---|---|---|
| $t_1$ | $t_1$ | $t_1$ | $t_1$ | $t_3$ | $t_2$ | $t_3$ |
| $t_2$ | $t_4$ | $t_2$ | $t_2$ | | $t_4$ | $t_5$ |
| $t_3$ | $t_5$ | $t_3$ | $t_3$ | | $t_6$ | $t_6$ |
| $t_4$ | $t_7$ | $t_5$ | $t_4$ | | $t_8$ | $t_7$ |
| $t_7$ | | $t_8$ | $t_6$ | | | $t_8$ |
| | | | $t_8$ | | | |

Vertical Database

(3.) Using minimum support 3, Apriori Algorithm find F(3).

Sol There are 3 main steps - Count, Filtering and Joining.

Step 1 → Count Items.

$C_1$ =

| Itemsets | Count |
|---|---|
| {A} | 5 |
| {B} | 4 |
| {C} | 5 |
| {D} | 6 |
| {E} | 1. → remove as it is not satisfying min. support of 3. |
| {F} | 4 |
| {G} | 5 |

Step 2 → Filter Items.

$L_1$ →

| Itemset | Count |
|---|---|
| {A} | 5 |
| {B} | 4 |
| {C} | 5 |
| {D} | 6 |
| {F} | 4 |
| {G} | 5 |

Step 3 → Join Items.

$C_2$ →

| Itemsets | Count. |
|---|---|
| {A B} | 3 |
| {A C} | 3 |
| {A D} | 4 |
| {A F} | 2 |
| {A G} | 2 |
| {B C} | 2 |
| {B D} | 2 |
| {BF} | 2 |
| {BG} | 2 |
| {CD} | 4 |
| {CF} | 2 |
| {CG} | 3 |
| {DF} | 4 |
| {DG} | 3 |
| {FG} | 2. |

Remove - because it did not meet the min. support of 3.

$L_2 \rightarrow$

| Itemset | Count |
|---|---|
| {AB} | 3 |
| {AC} | 3 |
| {AD} | 4 |
| {CD} | 4 |
| {CG} | 3 |
| {DF} | 4 |
| {DG} | 3 |

This is F(2).

Now, continue till F(3).

$C_3 \rightarrow$

| Itemset | count |
|---|---|
| {ABC} | 1 |
| {ABD} | 2 |
| {ACD} | 3 |
| {ACG} | 1 |
| {ADF} | 2 |
| {ADG} | 1 |
| {CDG} | 2 |
| {CDF} | 2 |
| {DFG} | 2 |

Removing all the $C_3$ itemsets except {ACD} = 3, because the rest did not meet the min. support of 3.

$L_3 \rightarrow$

| Itemset | Count |
|---|---|
| {ACD} | 3 |

This is F(3).

(4.) FPgrowth using minimum support of 2.

Frequency pattern set:

| Itemset | Count (Frequency) |
|---|---|
| A | 5 |
| B | 4 |
| C | 5 |
| D | 6 |
| E | 1 |
| F | 4 |
| G | 5 |

Arranging the Itemset in decreasing order of its counts.

| Itemset | Count |
|---|---|
| D | 6 |
| A | 5 |
| C | 5 |
| G | 5 |
| B | 4 |
| F | 4 |
| E | 1 |

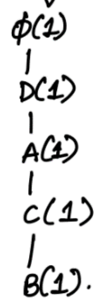
Now as the minimum support is 2, we can eliminate (E) and will not include it in our set. The set L will look like

$L = \{D:6, A:5, C:5, G:5, B:4, F:4\}$

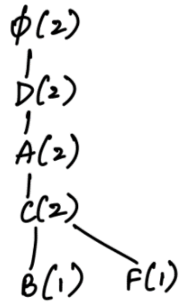
This is frequent pattern set. After this we will create ordered item set.

| tid | Itemset | Ordered Item set |
|---|---|---|
| $t_1$ | {A,B,C,D} | {D,A,C,B} |
| $t_2$ | {A,C,D,F} | {D,A,C,F} |
| $t_3$ | {A,C,D,E,G} | {D,A,C,G} |
| $t_4$ | {A,B,D,F} | {D,A,B,F} |
| $t_5$ | {B,C,G} | {C,G,B} |
| $t_6$ | {D,F,G} | {D,G,F} |
| $t_7$ | {A,B,G} | {A,G,B} |
| $t_8$ | {C,D,F,G} | {D,C,G,F} |

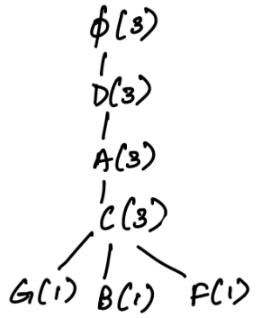
① Inserting set for $t_1$



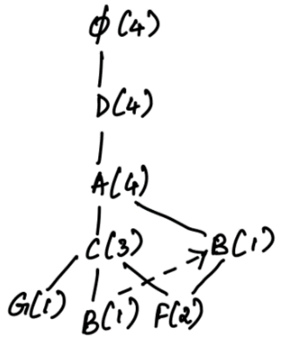

② For $t_2$



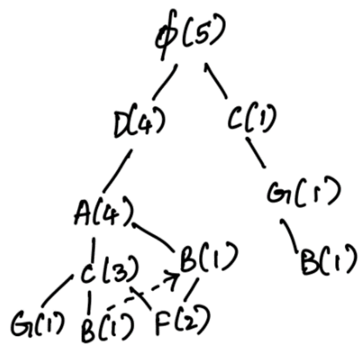

③ for $t_3$



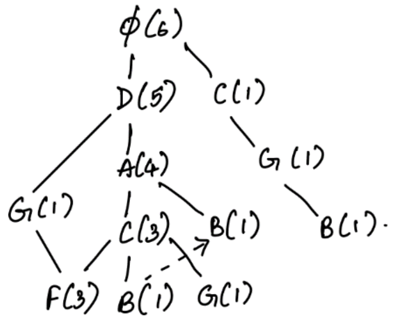

④ for $t_4$.



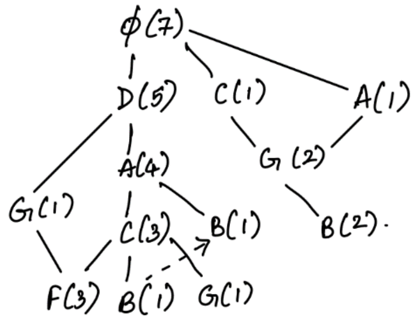

⑤ for $t_5$



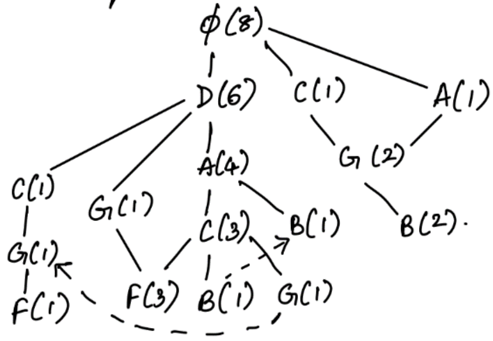
6 for $t_6$




7 for $t_7$




8 for $t_8$




Final FP Tree.

Conditional Pattern Base:

| Items | Conditional Pattern Base | Conditional Frequent Pattern Tree. |
|---|---|---|
| {F} | → { (D,A,C : 1), (D,A,B : 1) (D,C,G : 1), (D,G : 1)} | → {D : 4}. |
| {B} | → { (D,A,C : 1), (D,A : 1), (C,G : 1), (A,G : 1)} | → – |
| {G} | → { (D,A,C : 1), (D,C : 1), (C : 1), (D : 1), (A : 1)} | → – |
| {C} | → { (D,A : 3), (D : 1) } | → {D : 4} |
| {A} | → { (D : 4) } | → {D : 4} |
| {D} | → – | → – |

Question

$\boxed{2.}$ Given $x_1 = (0,3)$ $x_2 = (3,3)$ $x_3 = (0,0)$

Centroid $C_1 = (3.5, -1)$

1 → SSE :

$$\text{for } x_1 = \|x_i - C_i\|^2 = \|(0,3) - (3.5,-1)\|^2$$

$$= \|-3.5, 4\|^2 = 12.25 + 16 = \underline{\underline{28.25}}.$$

$$\text{for } x_2 = \|(3,3) - (3.5,-1)\|^2 = 0.25 + 16 = \underline{\underline{16.25}}$$

$$\text{for } x_3 = \|(0,0) - (3.5,-1)\|^2 = 12.25 + 1 = \underline{\underline{13.25}}$$

$$SSE = SSE(x_1) + SSE(x_2) + SSE(x_3)$$

$$= 28.25 + 16.25 + 13.25$$

$$= 57.75$$

The sum of squared errors for the intial cluster assignment is 57.75.

2 → The location of next centroid can be calculated by taking mean of data points.

$$\text{Centroid} = \frac{x_1 + x_2 + x_3}{3} = \left(\frac{0+3+0}{3}, \frac{3+3+0}{3}\right) = (1,2).$$

The centroid after next iteration is $\underline{(1,2)}$.

## Question 3

### 3.1

$$f_i(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\mu}_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_i|^{\frac{1}{2}}} \exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T\Sigma_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}{2}\right\}$$

Likelihood:

$$P(\mathbf{D}|\boldsymbol{\theta}) = \prod_{j=1}^{n} f(\mathbf{x}_j)$$

Log-likelihood:

$$\ln P(\mathbf{D}|\boldsymbol{\theta}) = \sum_{j=1}^{n} \ln f(\mathbf{x}_j) = \sum_{j=1}^{n} \ln\left(\sum_{i=1}^{k} f(\mathbf{x}_j|\boldsymbol{\mu}_i, \Sigma_i)P(C_i)\right)$$

Log-likelihood value: -66.08363702694996

[Used the above formulas in a python code to calculate the value].

### 3.2 E-Step

```
Posterior Probability is given below for each data point:
```

Out[4]:

| Data Point | w_1_parameter | w_2_parameter |
|---|---|---|
| 1.0 | 0.996776 | 0.003224 |
| 1.3 | 0.995731 | 0.004269 |
| 2.2 | 0.990107 | 0.009893 |
| 2.6 | 0.985656 | 0.014344 |
| 2.8 | 0.982741 | 0.017259 |
| 5.0 | 0.878040 | 0.121960 |
| 7.3 | 0.453138 | 0.546862 |
| 7.4 | 0.429964 | 0.570036 |
| 7.5 | 0.407092 | 0.592908 |
| 7.7 | 0.362622 | 0.637378 |

### 3.3 M-Step

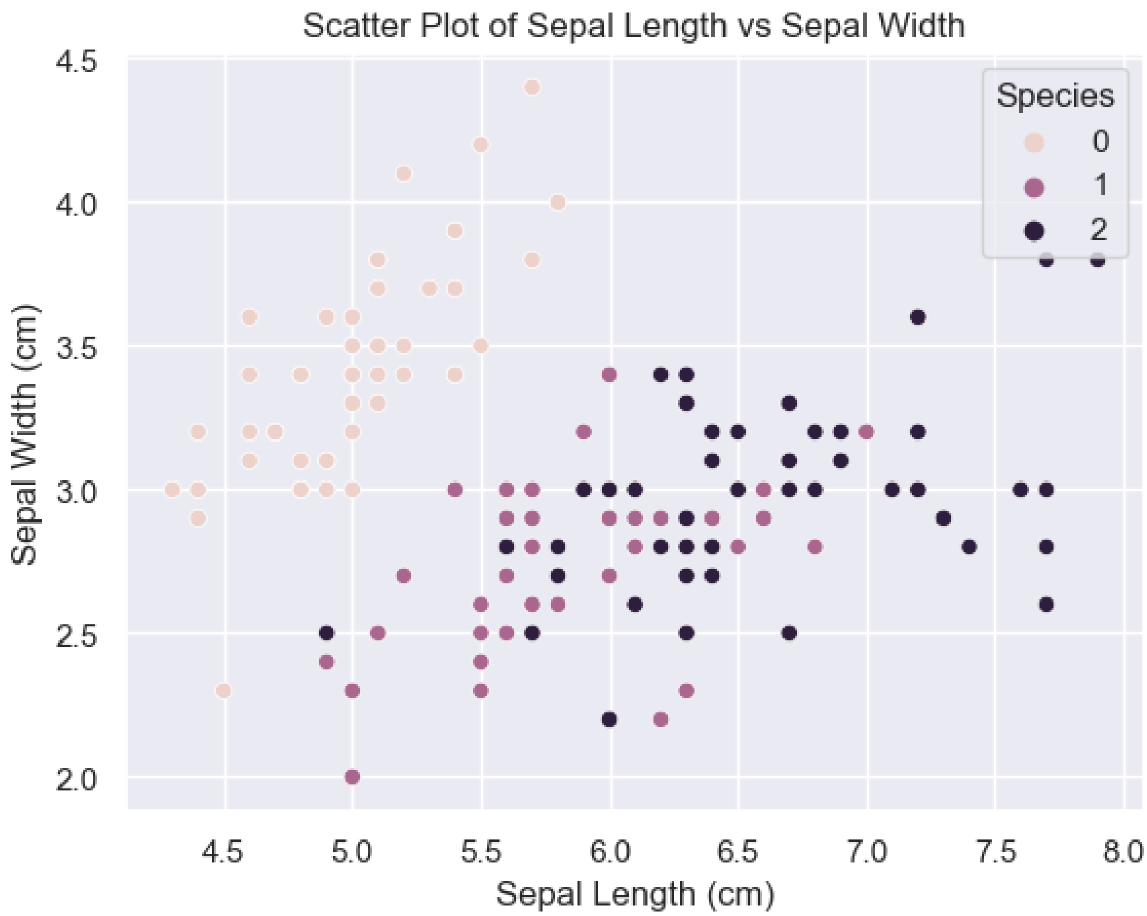
```
The means values are: [3.54310792 7.2636904 ]
The variance values are: [5.60786076 0.78572332]
The prior probability values are: [0.74818672 0.25181328]
```

**Question 4**

| | sepal length (cm) | sepal width (cm) | species |
|---|---|---|---|
| **0** | 5.1 | 3.5 | 0 |
| **1** | 4.9 | 3.0 | 0 |
| **2** | 4.7 | 3.2 | 0 |
| **3** | 4.6 | 3.1 | 0 |
| **4** | 5.0 | 3.6 | 0 |
| **...** | ... | ... | ... |
| **145** | 6.7 | 3.0 | 2 |
| **146** | 6.3 | 2.5 | 2 |
| **147** | 6.5 | 3.0 | 2 |
| **148** | 6.2 | 3.4 | 2 |
| **149** | 5.9 | 3.0 | 2 |

150 rows × 3 columns

**4.1**

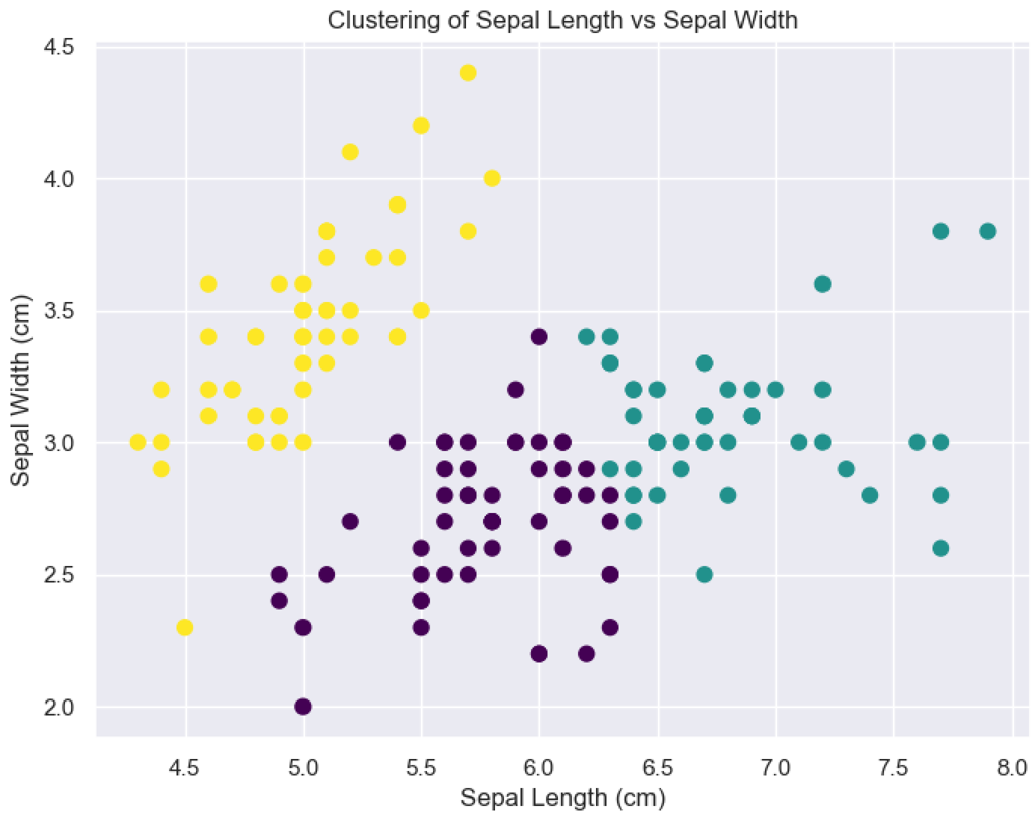


**4.2**

| | sepal length (cm) | sepal width (cm) | species | cluster |
|---|---|---|---|---|
| **0** | 5.1 | 3.5 | 0 | 2 |
| **1** | 4.9 | 3.0 | 0 | 2 |
| **2** | 4.7 | 3.2 | 0 | 2 |
| **3** | 4.6 | 3.1 | 0 | 2 |
| **4** | 5.0 | 3.6 | 0 | 2 |
| **...** | ... | ... | ... | ... |
| **145** | 6.7 | 3.0 | 2 | 1 |
| **146** | 6.3 | 2.5 | 2 | 0 |
| **147** | 6.5 | 3.0 | 2 | 1 |
| **148** | 6.2 | 3.4 | 2 | 1 |
| **149** | 5.9 | 3.0 | 2 | 0 |

150 rows × 4 columns

**4.2 (a)**



**4.2 (b)**

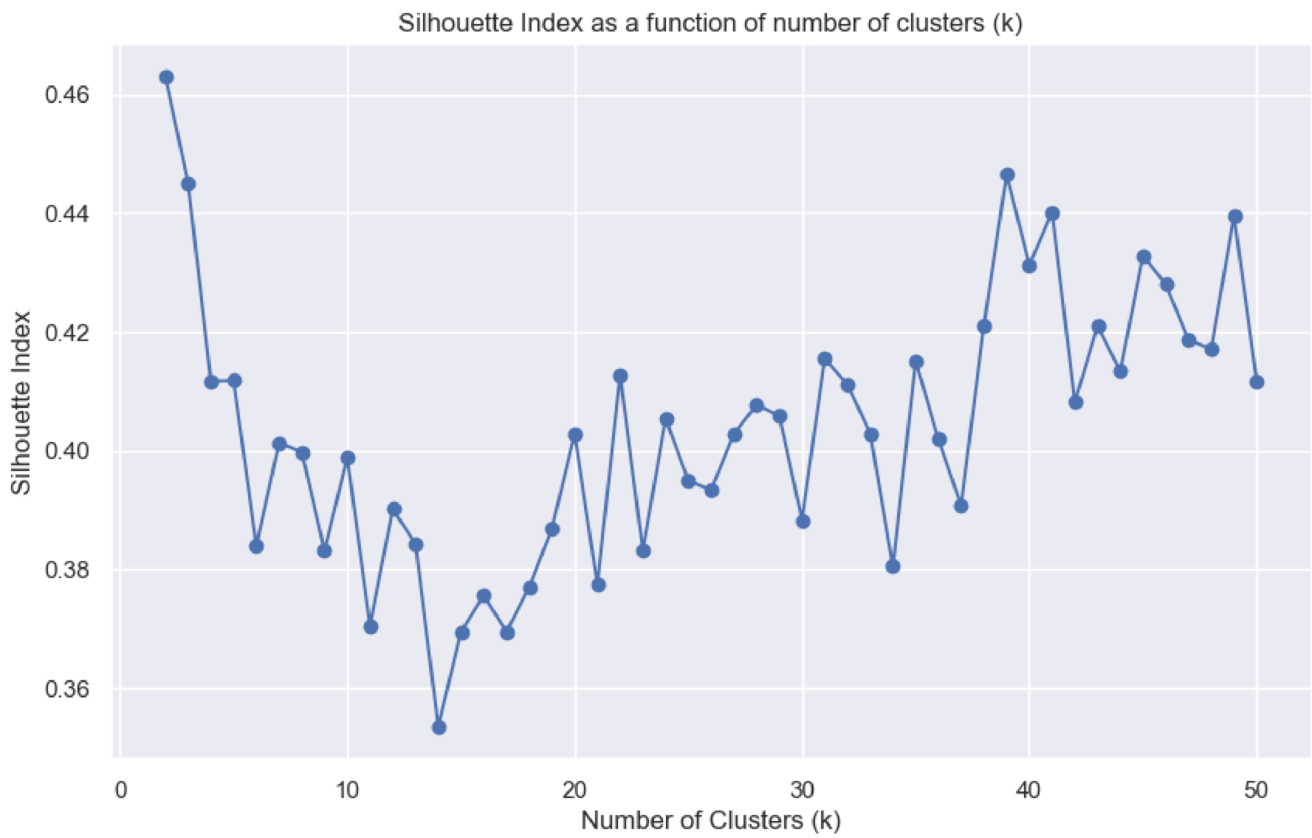
```
Silhouette Index: 0.44505256920836367
```

**4.2(c)**

If the Silhouette index is high(more closer to 1) for a particular clustering, it suggests that the clusters formed by the algorithm are distinct and well-separated based on the chosen features. In this case, the clustering assignment provides valuable insights as the k-Means produced clusters (not fully)partially align with the class labels.

Partial Alignment: The k-Means produces clusters that partially align with the class labels. Which means it may group some species correctly but not others. In this scenario, k-Means captures some of the underlying patterns in the class labels data but not all.

**4.3**

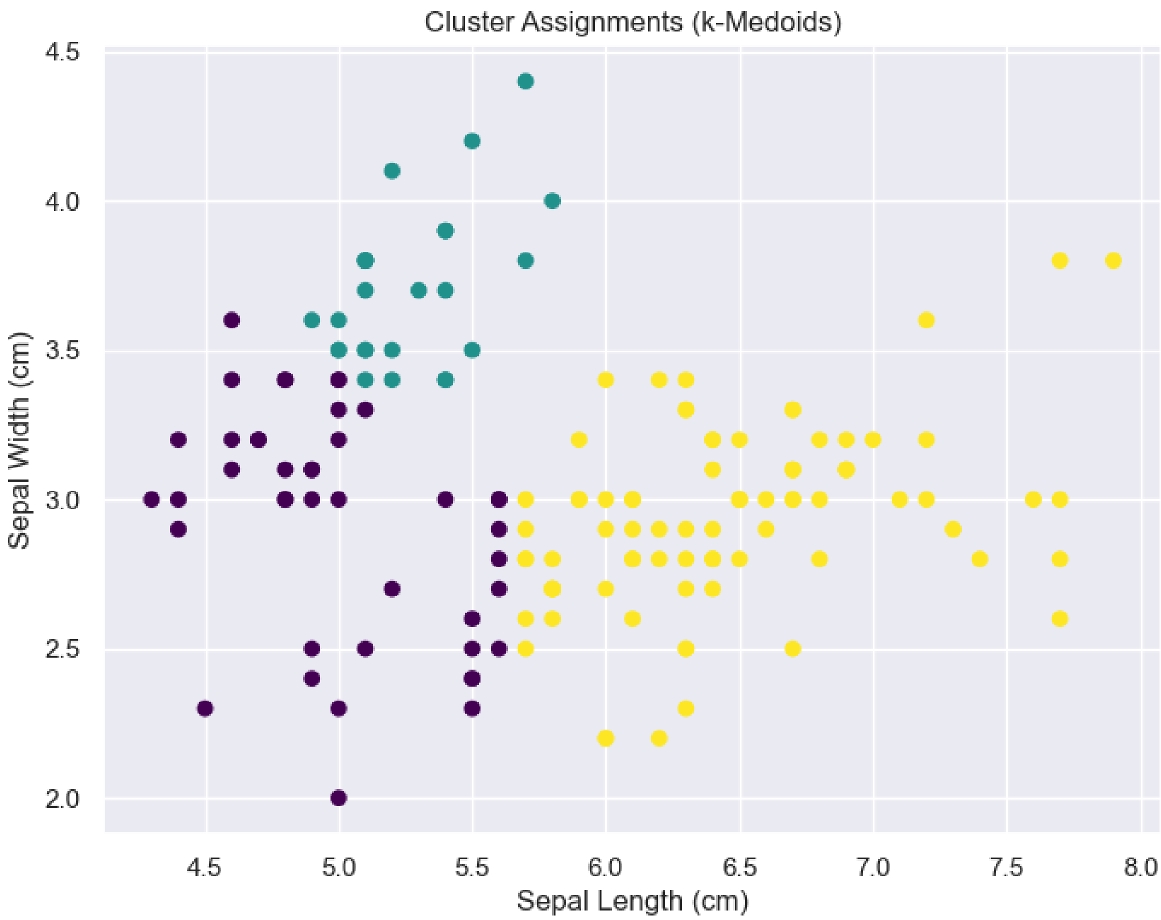


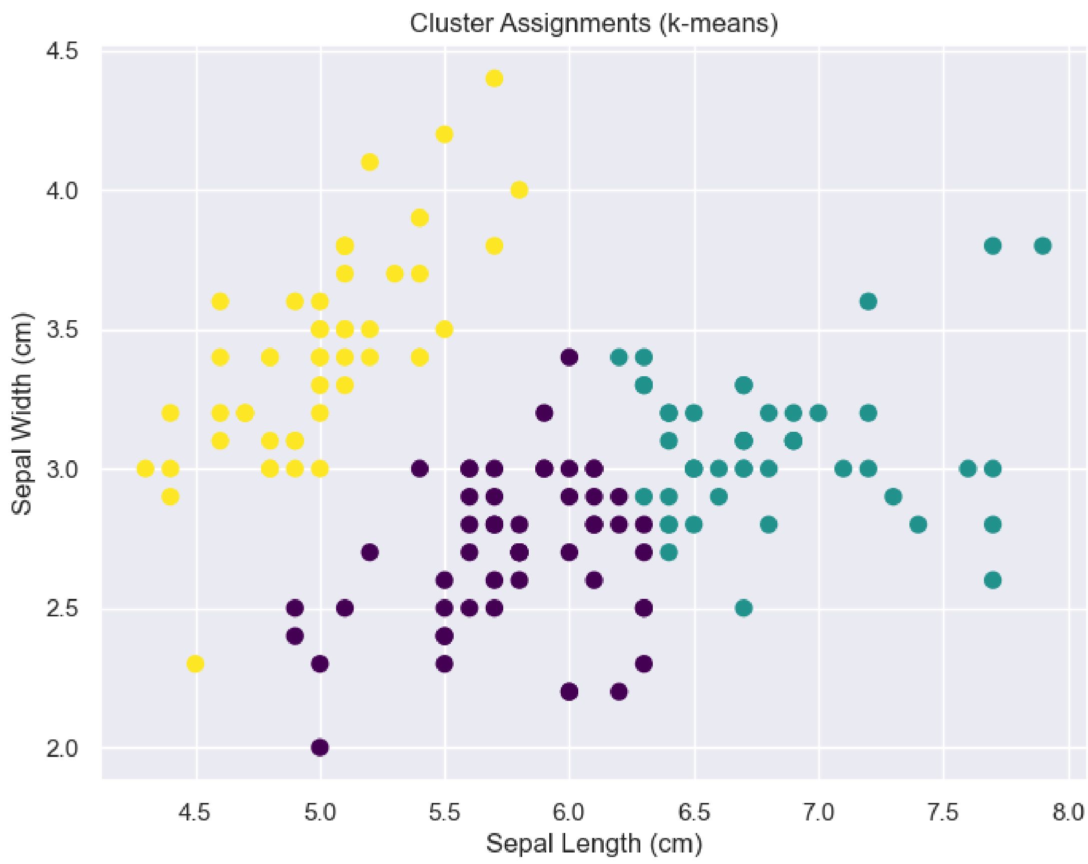
Yes. k=2 and 39 have silhouette index values greater than k=3, but k=2 has the highest value.

Silhouette index is a measure that helps assess the quality of clusters produced by a clustering algorithm. It quantifies how similar each data point is to its own cluster compared to other clusters. A high Silhouette score indicates that the clusters are well-separated and data points are tightly grouped within their respective clusters.

Hence, the one with the higher average Silhouette score is considered to perform better in terms of cluster quality.

**4.4**






The cluster assignment for both K-mediods and k-means are almost the same except very few datapoints. The cluster assignments here for Iris dataset has some outliers and clusters with non-spherical shapes. So K-mediods might be a better choice as compared to k means because it is more robust to outliers and can handle clusters of different shapes.

**References –**


- Gaussian-Mixtures.ipynb (Class Slides)
- https://numpy.org/doc/stable/reference/index.html
- https://pandas.pydata.org/docs/user_guide/10min.html
- https://scikit-learn.org/stable/tutorial/basic/tutorial.html
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- https://scikit-learn-extra.readthedocs.io/en/stable/generated/sklearn_extra.cluster.KMedoids.html
- https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
- https://matplotlib.org/stable/tutorials/introductory/pyplot.html
- https://seaborn.pydata.org/generated/seaborn.scatterplot.html