2. Clustering

2.2. Second dataset

Scatter Plot

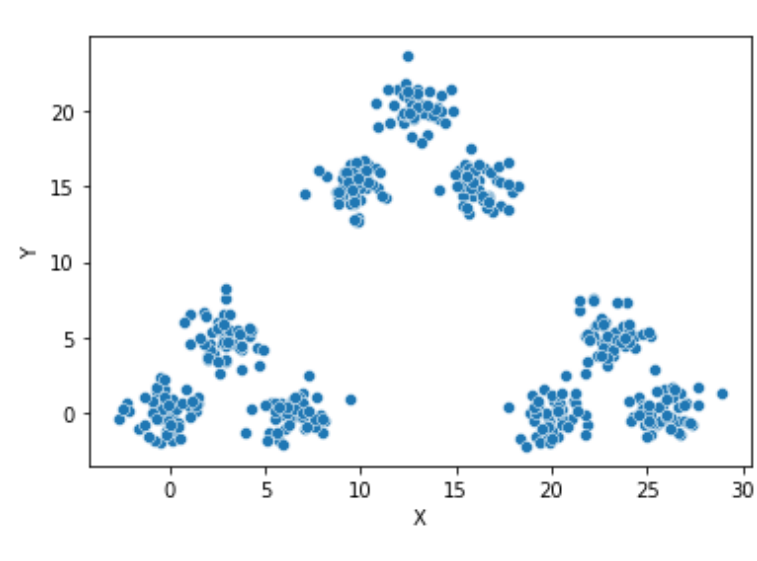


Figure 1: Scatter plot for the second dataset

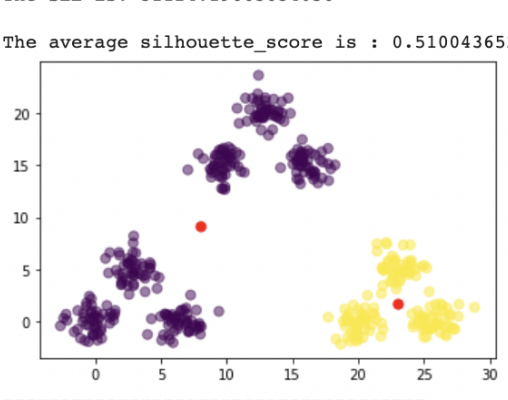
Type of clusters: Density-based Clusters
- A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
- Used when the clusters are irregular or intertwined, and when noise and outliers are present.

Predicted number of clusters: $K = 3$ or $K = 9$
This solely depends on the nature of our problem and if we want sparse or dense clusters.

Lloyd's k-means Clustering

Clusters are separated by colors and the **centroids** are colored **red**.

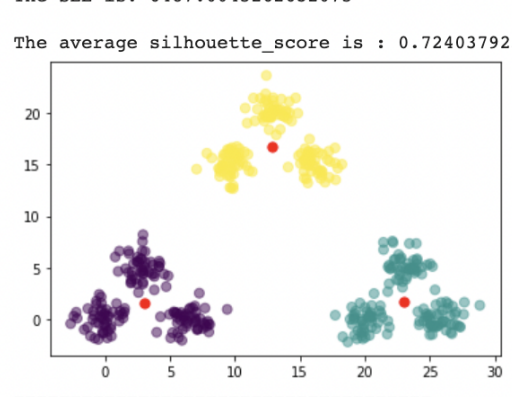
```
====================================
========== Results for k=2 ==========

The centers are: [[ 7.96561081  9.21502853]
 [22.98988192  1.62558656]]

The SEE is: 31124.19663634056

The average silhouette_score is : 0.510043652072989
```



```
====================================
```

Figure 2: Number of clusters = 2

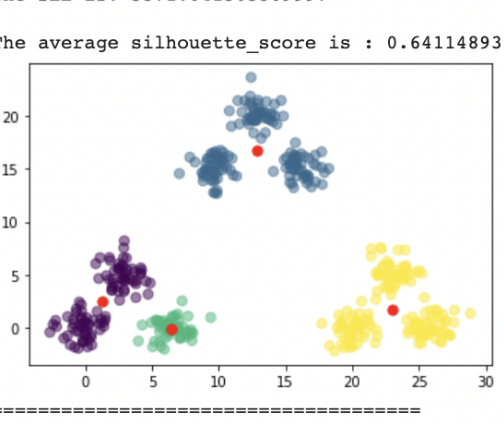
```
====================================
========== Results for k=3 ==========

The centers are: [[ 3.03019658  1.61465616]
 [22.98988192  1.62558656]
 [12.90102504 16.8154009 ]]

The SEE is: 6487.0045262632075

The average silhouette_score is : 0.7240379233766401
```



```
====================================
```

Figure 3: Number of clusters = 3

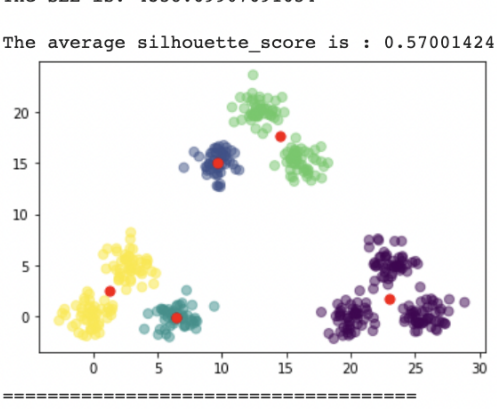
```
====================================
========== Results for k=4 ==========

The centers are: [[ 1.30948949  2.48588738]
 [12.90102504 16.8154009 ]
 [ 6.47161078 -0.12780626]
 [22.98988192  1.62558656]]

The SEE is: 5371.041505309997

The average silhouette_score is : 0.641148939502236
```



```
====================================
```

Figure 4: Number of clusters = 4

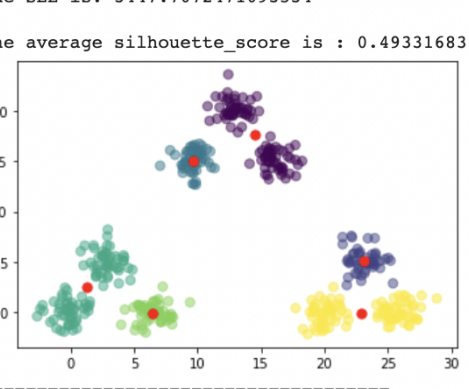
```
====================================
========== Results for k=5 ==========

The centers are: [[22.98988192  1.62558656]
 [ 9.65835456 15.08826824]
 [ 6.47161078 -0.12780626]
 [14.52236028 17.67896722]
 [ 1.30948949  2.48588738]]

The SEE is: 4358.69907691654

The average silhouette_score is : 0.5700142407916128
```



```
====================================
```

Figure 5: Number of clusters = 5

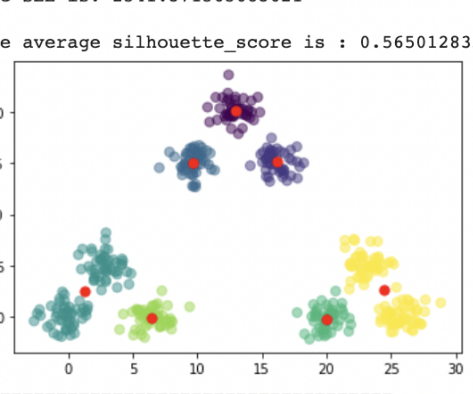
```
====================================
========== Results for k=6 ==========

The centers are: [[14.52236028 17.67896722]
 [23.11473144  5.10854081]
 [ 9.65835456 15.08826824]
 [ 1.30948949  2.48588738]
 [ 6.47161078 -0.12780626]
 [22.92745716 -0.11589057]]

The SEE is: 3447.7072471093534

The average silhouette_score is : 0.49331683106610574
```



```
====================================
```

Figure 6: Number of clusters = 6

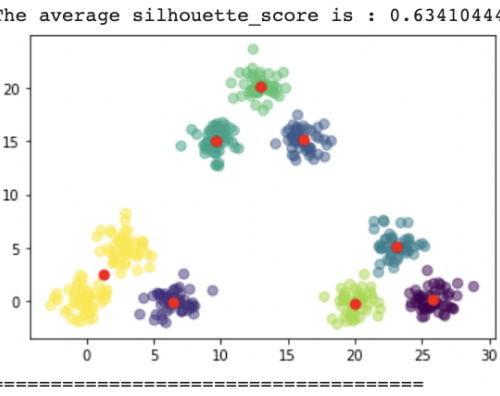
```
====================================
========== Results for k=7 ==========

The centers are: [[12.90964167 20.21423962]
 [16.13507889 15.14369483]
 [ 9.65835456 15.08826824]
 [ 1.30948949  2.48588738]
 [20.02738714 -0.22284562]
 [ 6.47161078 -0.12780626]
 [24.47112931  2.54980265]]

The SEE is: 2541.371565683621

The average silhouette_score is : 0.5650128357224193
```



```
====================================
```

Figure 7: Number of clusters = 7

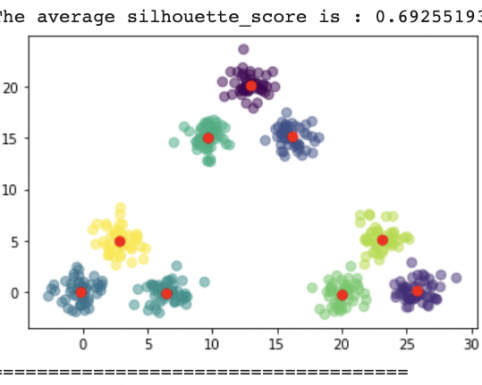
```
====================================
========== Results for k=8 ==========

The centers are: [[25.81913993  0.04673579]
 [ 6.47161078 -0.12780626]
 [16.13507889 15.14369483]
 [23.06809785  5.1550355 ]
 [ 9.65835456 15.08826824]
 [12.90964167 20.21423962]
 [20.02738714 -0.22284562]
 [ 1.30948949  2.48588738]]

The SEE is: 1700.1342337556189

The average silhouette_score is : 0.6341044460116986
```



```
====================================
```

Figure 8: Number of clusters = 8

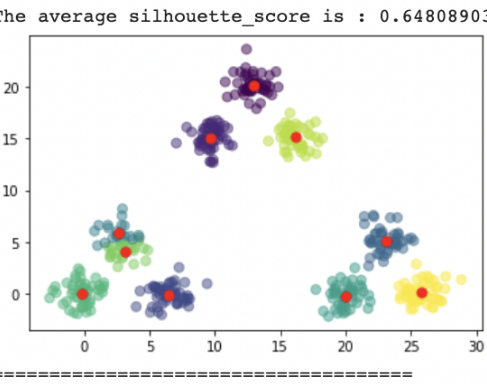
```
====================================
========== Results for k=9 ==========

The centers are: [[12.90964167 20.21423962]
 [25.81913993  0.04673579]
 [16.13507889 15.14369483]
 [-0.23091099  0.02934982]
 [ 6.47161078 -0.12780626]
 [ 9.65835456 15.08826824]
 [20.02738714 -0.22284562]
 [23.06809785  5.1550355 ]
 [ 2.84988997  4.94242493]]

The SEE is: 859.3931951407192

The average silhouette_score is : 0.6925519321164127
```



```
====================================
```

Figure 9: Number of clusters = 9

```
====================================
========== Results for k=10 ==========

The centers are: [[12.90964167 20.21423962]
 [ 9.65835456 15.08826824]
 [ 6.47161078 -0.12780626]
 [23.06809785  5.1550355 ]
 [ 2.58729553  5.88169118]
 [20.02738714 -0.22284562]
 [-0.23091099  0.02934982]
 [ 3.09228483  4.07540993]
 [16.13507889 15.14369483]
 [25.81913993  0.04673579]]

The SEE is: 815.4927216725604

The average silhouette_score is : 0.6480890332990465
```



```
====================================
```

Figure 10: Number of clusters = 10

Minimum SSE: 815
Number of clusters with minimum SSE: 10
** Be careful! a small number of SSE does not guarantee a good clustering! Elbow method is used to solve this in the next part.**

Maximum silhouette score: 0.72
Number of clusters with maximum silhouette score: 3

K-Means Elbow Method

```
Elbow point: 3
SSE for this point is:  5371.041505309997
<Figure size 360x360 with 0 Axes>
```

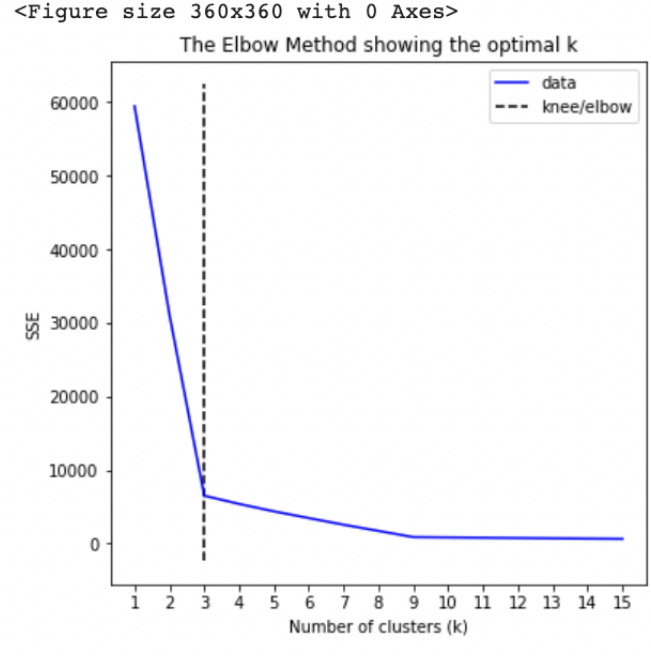
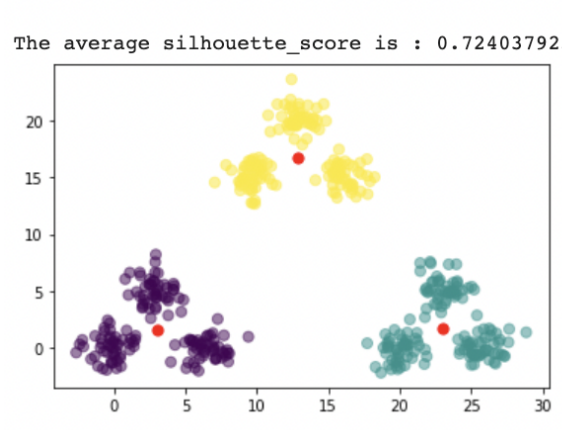


Figure 11: Elbow Method for the second dataset

Thus we conclude that the **best clustering using k-means algorithm and validating by both SSE, elbow method** and **silhouette score** is:

```
====================================
========== Results for k=3 ==========

The centers are: [[ 3.03019658  1.61465616]
 [22.98988192  1.62558656]
 [12.90102504 16.8154009 ]]

The SEE is: 6487.0045262632075

The average silhouette_score is : 0.7240379233766401
```

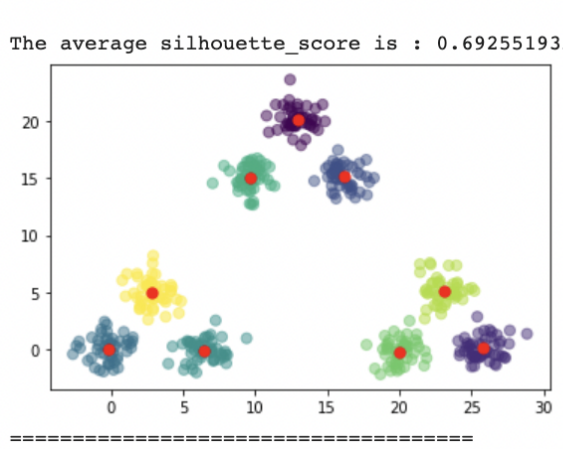


```
====================================
```

Figure 12: Best clustering for the second dataset using SSE, elbow method and silhouette score

It is worth mentioning we have **another elbow point at** $k = 9$:

```
====================================
========== Results for k=9 ==========

The centers are: [[12.90964167 20.21423962]
 [25.81913993  0.04673579]
 [16.13507889 15.14369483]
 [-0.23091099  0.02934982]
 [ 6.47161078 -0.12780626]
 [ 9.65835456 15.08826824]
 [20.02738714 -0.22284562]
 [23.06809785  5.1550355 ]
 [ 2.84988997  4.94242493]]

The SEE is: 859.3931951407192

The average silhouette_score is : 0.6925519321164127
```



```
====================================
```

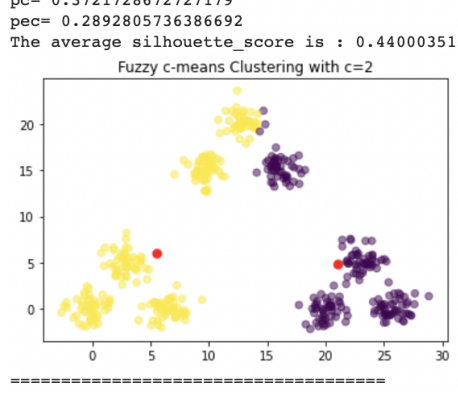
Figure 13: Another elbow point with close silhouette score

Fuzzy C-Means Clustering

Clusters are separated by colors, the **centroids** are colored **red** and the data points that are a member of **more than two clusters** are colored **purple**. (except for the first one)

cutoff_coefficient = 0.001

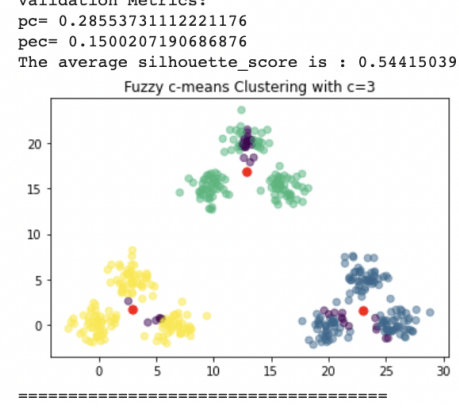
```
====================================
========== Results for k=2 ==========

The centers are:
[[ 5.52174262  6.01312285]
 [21.02361847  4.82320456]]

Number of datapoints belonging to more than one cluster= 0

Validation Metrics:
pc= 0.3721728672727179
pec= 0.2892805736386692
The average silhouette_score is : 0.4400035130819573
```




```
====================================
```

Figure 14: Number of clusters = 2

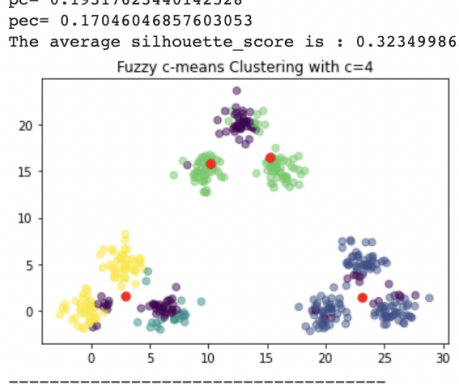
```
====================================
========== Results for k=3 ==========

The centers are:
[[ 2.94692748  1.63257103]
 [23.01244543  1.58299326]
 [12.90384001 16.89543729]]

Number of datapoints belonging to more than one cluster= 40

Validation Metrics:
pc= 0.28553731112221176
pec= 0.1500207190686876
The average silhouette_score is : 0.5441503938639852
```




```
====================================
```

Figure 15: Number of clusters = 3

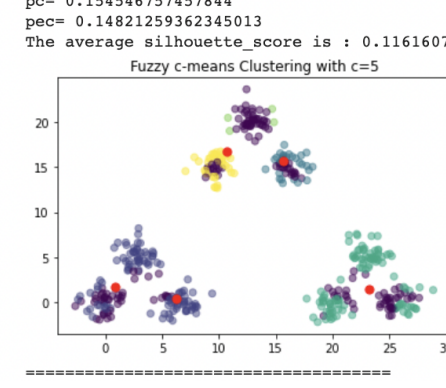
```
====================================
========== Results for k=4 ==========

The centers are:
[[23.06760887  1.45407281]
 [ 2.86911185  1.52330525]
 [10.2107555  15.8391191 ]
 [15.2553809  16.45509729]]

Number of datapoints belonging to more than one cluster= 99

Validation Metrics:
pc= 0.19317623440142528
pec= 0.17046046857603053
The average silhouette_score is : 0.3234998654628582
```




```
====================================
```

Figure 16: Number of clusters = 4

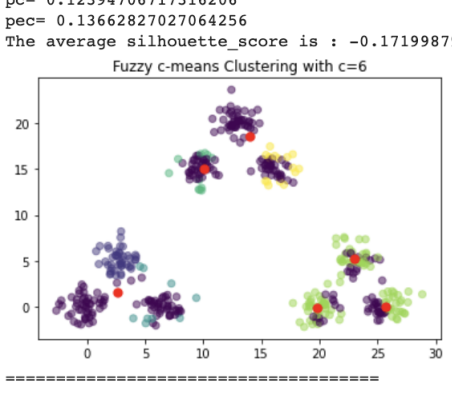
```
====================================
========== Results for k=5 ==========

The centers are:
[[23.19437623  1.45824426]
 [ 0.83091353  1.72749663]
 [15.66936134 15.73572958]
 [10.70029593 16.77432647]
 [ 6.20302115  0.40694677]]

Number of datapoints belonging to more than one cluster= 170

Validation Metrics:
pc= 0.154546757457844
pec= 0.14821259362345013
The average silhouette_score is : 0.11616071879586662
```




```
====================================
```

Figure 17: Number of clusters = 5

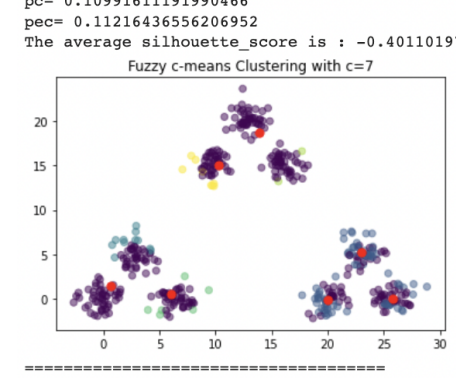
```
====================================
========== Results for k=6 ==========

The centers are:
[[ 1.39688162e+01  1.85392113e+01]
 [ 2.60938190e+00  1.57929870e+00]
 [ 1.98041341e+01 -1.88861441e-01]
 [ 1.01095649e+01  1.50341268e+01]
 [ 2.29581767e+01  5.20392286e+00]
 [ 2.56788967e+01  3.55970801e-03]]

Number of datapoints belonging to more than one cluster= 270

Validation Metrics:
pc= 0.12394706717316206
pec= 0.13662827027064256
The average silhouette_score is : -0.17199879812800842
```




```
====================================
```

Figure 18: Number of clusters = 6

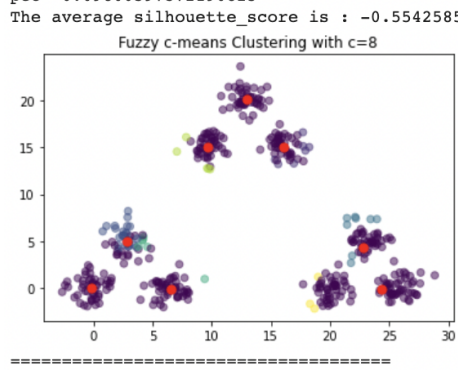
```
====================================
========== Results for k=7 ==========

The centers are:
[[ 1.02163943e+01  1.51201283e+01]
 [ 2.57412235e+01  1.55899461e-03]
 [ 6.02989573e+00  5.11321992e-01]
 [ 1.38490137e+01  1.87901975e+01]
 [ 2.30369604e+01  5.23014609e+00]
 [ 5.87968954e-01  1.38251616e+00]
 [ 1.99819083e+01 -1.77631125e-01]]

Number of datapoints belonging to more than one cluster= 345

Validation Metrics:
pc= 0.10991611191990466
pec= 0.11216436556206952
The average silhouette_score is : -0.40110197868135894
```




```
====================================
```

Figure 19: Number of clusters = 7

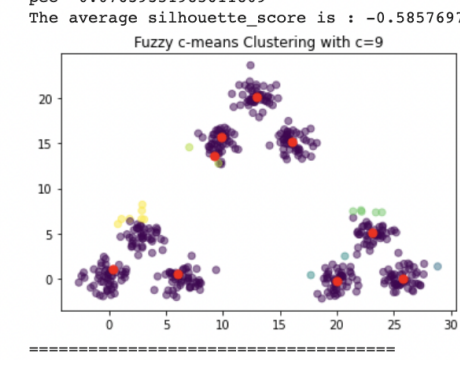
```
====================================
========== Results for k=8 ==========

The centers are:
[[ 2.42974249e+01 -1.32982225e-01]
 [ 9.68479297e+00  1.51163823e+01]
 [ 6.55333249e+00 -1.05747298e-01]
 [ 2.27548446e+01  4.35874612e+00]
 [-1.89277124e-01 -1.06888304e-02]
 [ 1.61111839e+01  1.51044593e+01]
 [ 2.85866734e+00  4.94524529e+00]
 [ 1.29327376e+01  2.01972452e+01]]

Number of datapoints belonging to more than one cluster= 395

Validation Metrics:
pc= 0.09540394369801365
pec= 0.09800597372190825
The average silhouette_score is : -0.5542585802440843
```




```
====================================
```

Figure 20: Number of clusters = 8

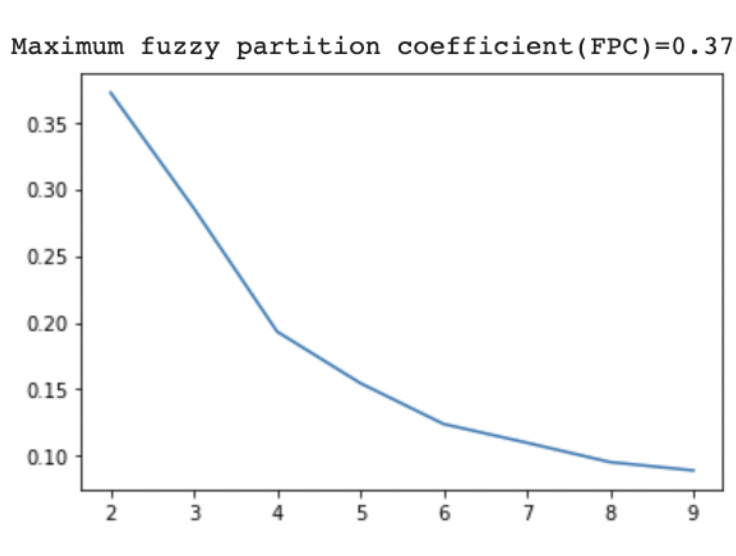
```
====================================
========== Results for k=9 ==========

The centers are:
[[12.9341823  20.2101901 ]
 [25.7571188  -0.03376178]
 [16.08222523 15.13567547]
 [ 9.22156232 13.6185359 ]
 [23.11299765  5.11054723]
 [ 9.80379585 15.65495411]
 [ 6.05228873  0.45419799]
 [19.99542992 -0.22652947]
 [ 0.37044967  1.00049141]]

Number of datapoints belonging to more than one cluster= 433

Validation Metrics:
pc= 0.08917283615930664
pec= 0.07639351985611809
The average silhouette_score is : -0.5857697571111562
```




```
====================================
```

Figure 21: Number of clusters = 9

```
Maximum fuzzy partition coefficient(FPC)=0.3721728672727179 for c=2
```



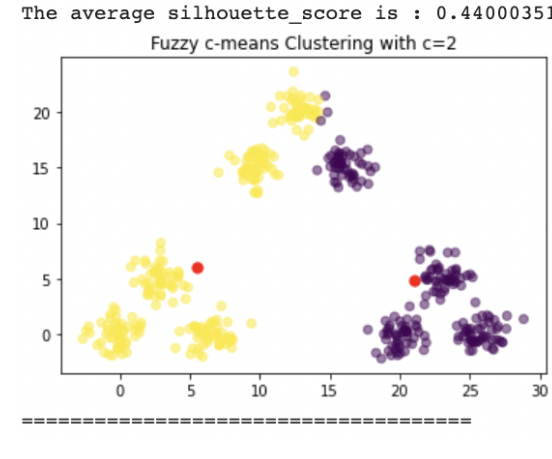
Minimum PEC: 0.076
Number of clusters with minimum PEC: 9
** Sounds good but one of the centers is placed in an inappropriate place.**

Maximum PC: 0.372
Number of clusters with Maximum PC: 2

Maximum silhouette score: 0.54
Number of clusters with maximum silhouette score: 3
Note that for the clustering with $k = 6,\ 7,\ 8$, and 9, the silhouette score becomes negative which means the datapoints are in the wrong clusters!

Thus we conclude that the **best clustering using fuzzy c-means algorithm and validating by partition coefficient** is:

```
====================================
========== Results for k=2 ==========

The centers are:
[[ 5.52174262  6.01312285]
 [21.02361847  4.82320456]]

Number of datapoints belonging to more than one cluster= 0

Validation Metrics:
pc= 0.3721728672727179
pec= 0.2892805736386692
The average silhouette_score is : 0.4400035130819573
```

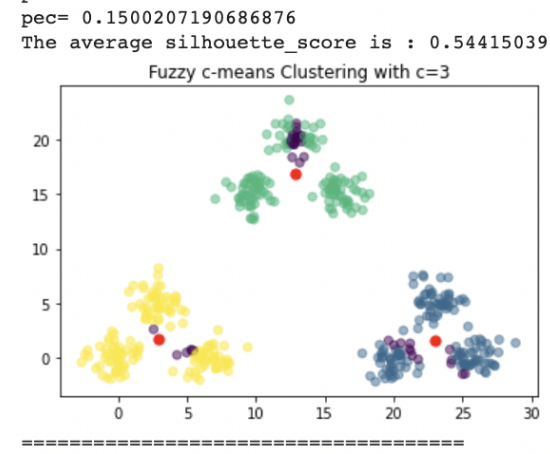



```
====================================
```

Figure 22: Best clustering for the second dataset using partition coefficient

Thus we conclude that the **best clustering using fuzzy c-means algorithm and validating by silhouette score** is:

```
====================================
========== Results for k=3 ==========

The centers are:
[[ 2.94692748  1.63257103]
 [23.01244543  1.58299326]
 [12.90384001 16.89543729]]

Number of datapoints belonging to more than one cluster= 40

Validation Metrics:
pc= 0.28553731112221176
pec= 0.1500207190686876
The average silhouette_score is : 0.5441503938639852
```




```
====================================
```

Figure 23: Best clustering for the second dataset using silhouette score

DBSCAN Clustering

```
Knee point: 1.3242966586518992
<Figure size 360x360 with 0 Axes>
```

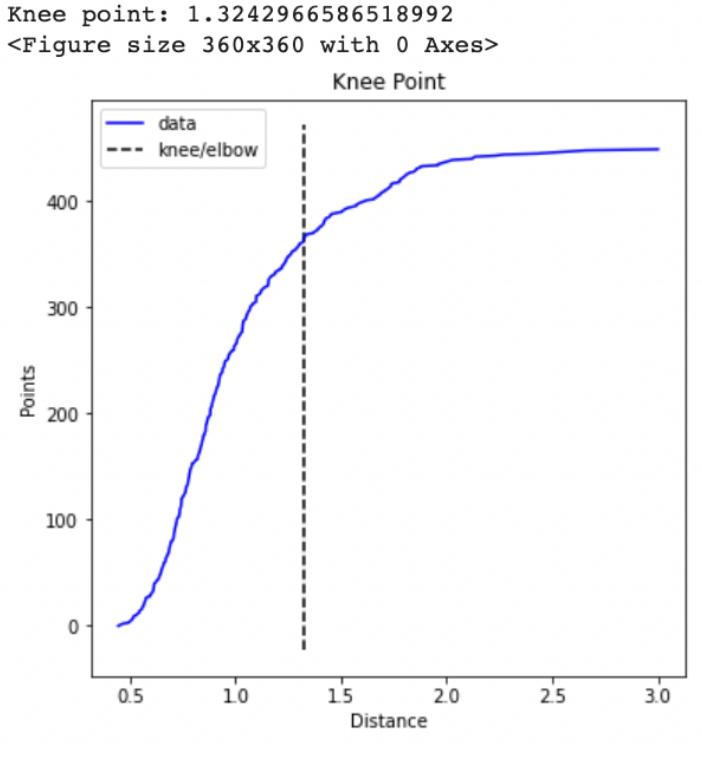
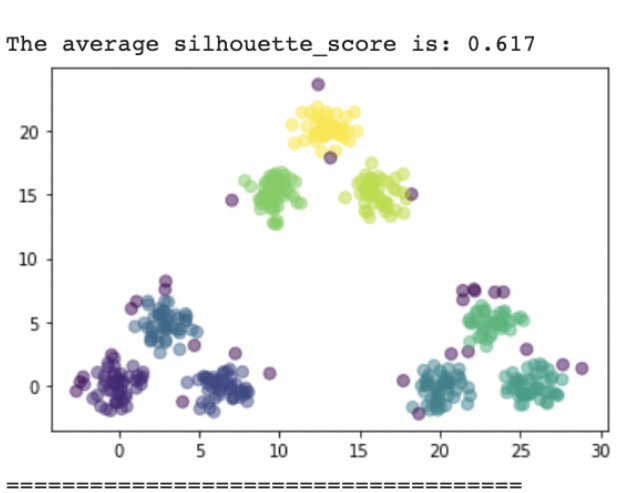


Figure 24: Best eps for DBSCAN using kNN algorithm

As explained in the previous section, we find the best *eps* parameter for the DBSCAN using the kNN algorithm. Thus the **best eps is 1.324**.

Then we try a range of numbers on the *min_samples* parameter and maximize the silhouette score. Doing so we achieve the below clustering:
Purple dots are the **noise**.

```
====================================
========== Results for eps=1.3242966586518992 ==========

Estimated number of clusters: 9
Estimated number of noise points: 29

The average silhouette_score is: 0.617
```



```
====================================
```

Thus we conclude that the **DBSCAN** algorithm **succeeds** wonderfully to cluster this dataset. Which is probably because it is density-based.