CSE 454 DATA MINING 2021-2022 FALL SEMESTER

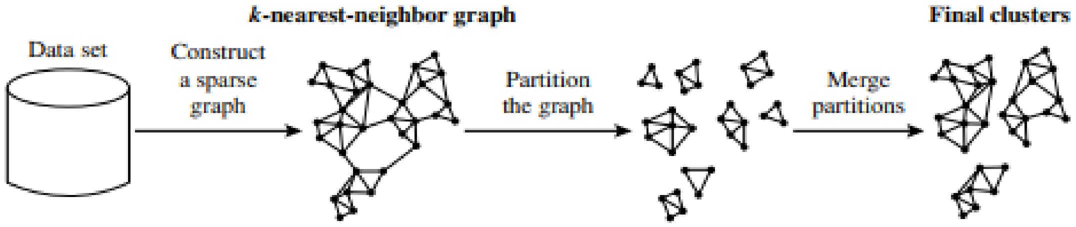
ASSIGMENT 1 REPORT

OZAN GEÇKİN

1801042103

Assignment Explanation:

My homework is the implementation of Chameleon clustering technique. I implemented it using the python programming language. I created the 2D dataset myself and bought the dataset from the internet, and I will send the datasets I found with my homework.
Chameleon clustering scheme in general:



The CHAMELEON algorithm takes place in two stages. In the first stage, by creating subsets; Creates a graph of the form G =(V , E). Here, each node represents an object (v ∈V) and if vj is one of vi's k-nearest neighbors, there is a weighted boundary e(vi ,vj) between nodes vi and vj. The weight of each border on the graph is represented by the closeness between the two objects, and the closer the two objects are to each other, the more weight the border will gain. Thus, CHAMELEON uses graph segmentation algorithm to create many small repetitive splits within the graph, and each repetition is divided into sub graphs by making minimum cut (min-cut) in the graph. This separation process is repeated until a certain criterion is reached. In the second stage, the algorithm follows a bottom-up process and uses the aggregator hierarchical clustering method. CHAMELEON judges the similarity between pairs of Ci and Cj clusters by looking at the RI (Ci ,Cj) relative connectivity and the RC (Ci ,Cj) relative affinities.

The relative interconnectivity, RI(Ci ,Cj), between two clusters, Ci and Cj , is defined as the absolute interconnectivity between Ci and Cj , normalized with respect to the internal interconnectivity of the two clusters, Ci and Cj . That is,

$$RI(C_i, C_j) = \frac{|EC_{\{C_i,C_j\}}|}{\frac{1}{2}(|EC_{C_i}| + |EC_{C_j}|)},$$

where EC{Ci ,Cj} is the edge cut as previously defined for a cluster containing both Ci and Cj . Similarly, ECCi (or ECCj) is the minimum sum of the cut edges that partition Ci (or Cj) into two roughly equal parts

The relative closeness, RC(Ci ,Cj), between a pair of clusters, Ci and Cj , is the absolute closeness between Ci and Cj , normalized with respect to the internal closeness of the two clusters, Ci and Cj . It is defined as

$$RC(C_i, C_j) = \frac{\overline{S}_{EC_{\{C_i,C_j\}}}}{\frac{|C_i|}{|C_i|+|C_j|}\overline{S}_{EC_{C_i}} + \frac{|C_j|}{|C_i|+|C_j|}\overline{S}_{EC_{C_j}}},$$

where SEC{Ci ,Cj } is the average weight of the edges that connect vertices in Ci to vertices in Cj , and SECCi (or SECCj) is the average weight of the edges that belong to the mincut bisector of cluster Ci (or Cj).

2. Prepare an assignment report showing extracted clusters for at least 3 values of each parameter. You should show clusters with figures.

```
def chameleonCluster(dataFrame, k,k_neighbor_number, subCluster):
```

Parameter "k" = Number of cluster.

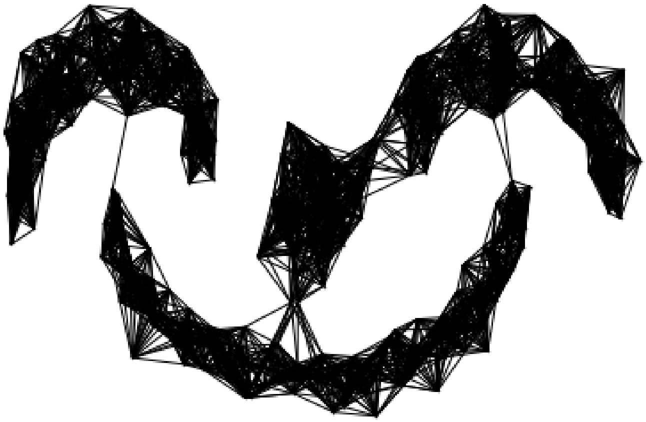
Parameter "k_neighbor_number" = K_nearest neighbor graph. Represents the number of nearest vertices to which it is connected.

Parameter "subCluster" = The minimum size of the initial cluster. Initially, all elements belong to a set. After partitioning starts, it continues until all clusters are smaller than the subCluster.

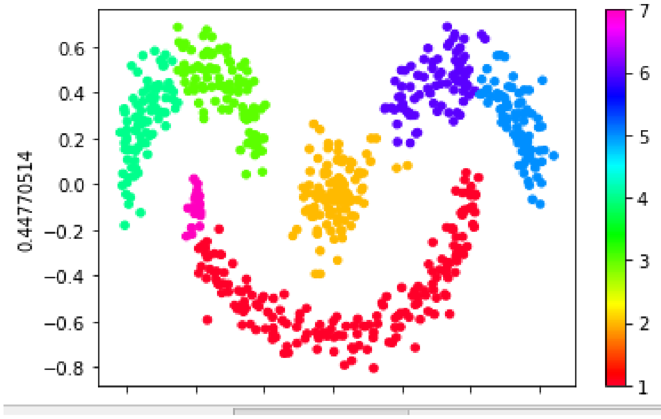
Effect of "k" parameter:

1) For k = 7, k_neighbor_number=20, subCluster=40

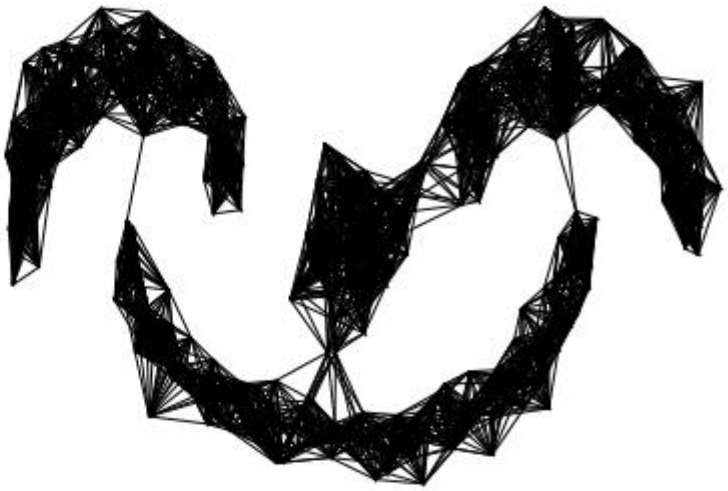
```
Create k neighbor number graph
```



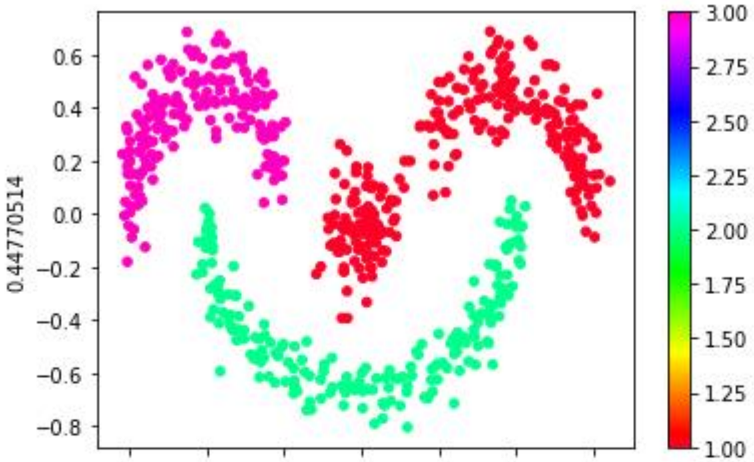
```
Stat Clustering...
```



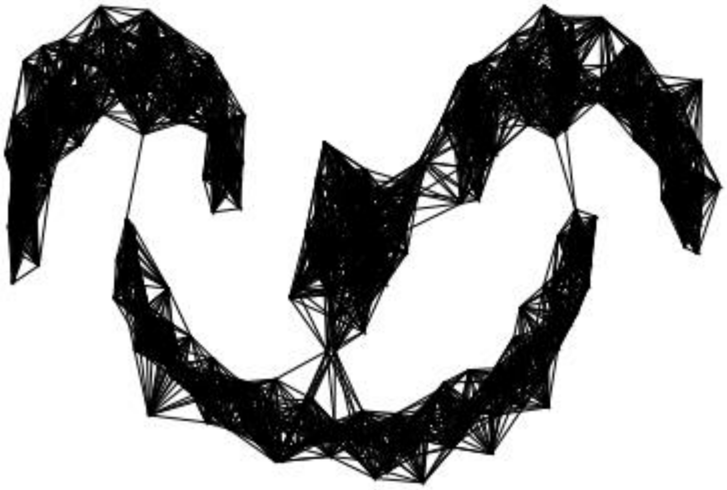
2) For k = 3, k_neighbor_number=20, subCluster=40



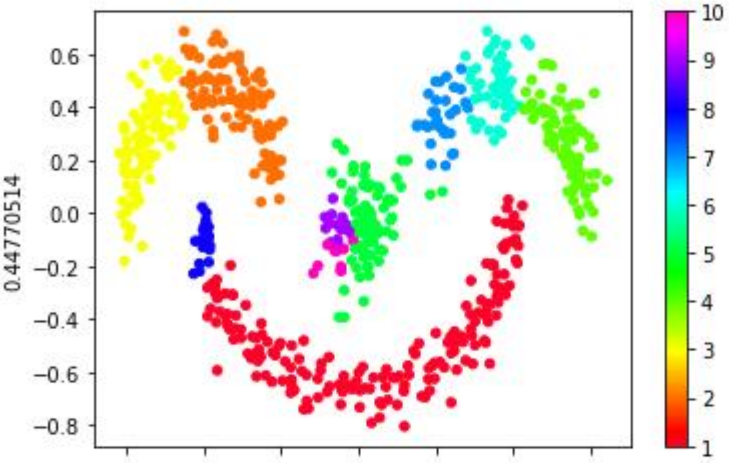
```
Stat Clustering...
```



3) For k = 10, k_neighbor_number=20, subCluster=40

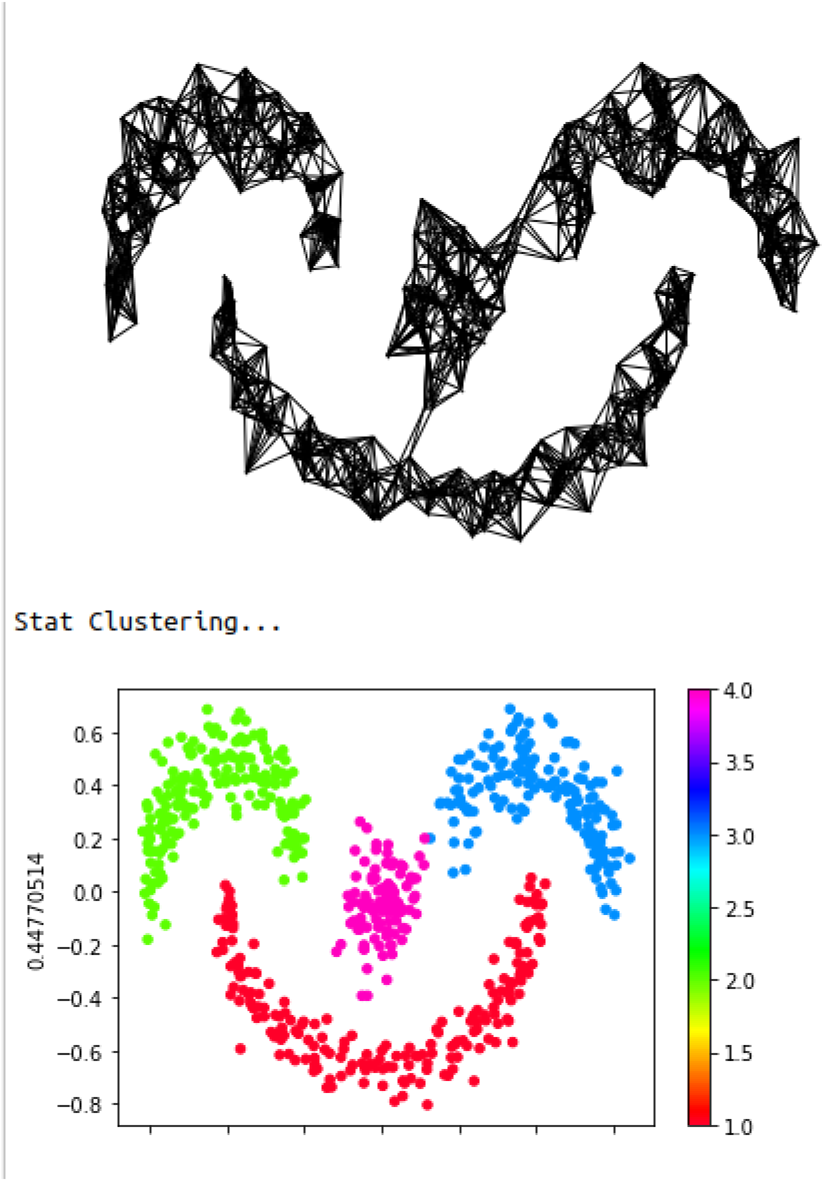


```
Stat Clustering...
```



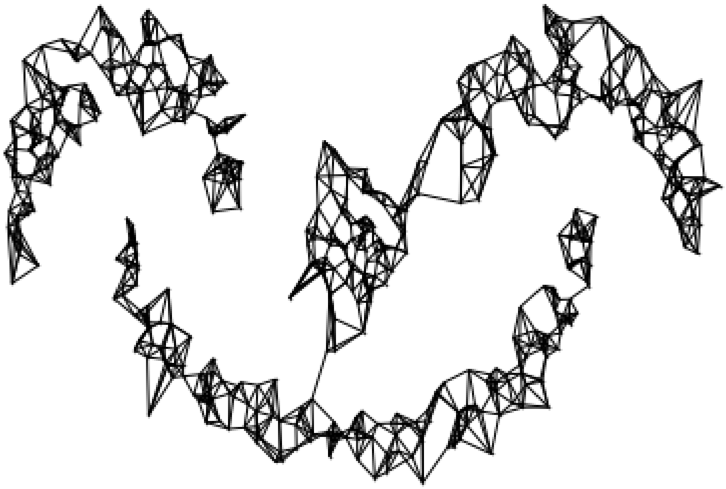
Effect of “k_neighbor_number” parameter:

1) For k = 4, k_neighbor_number=10, subCluster=40

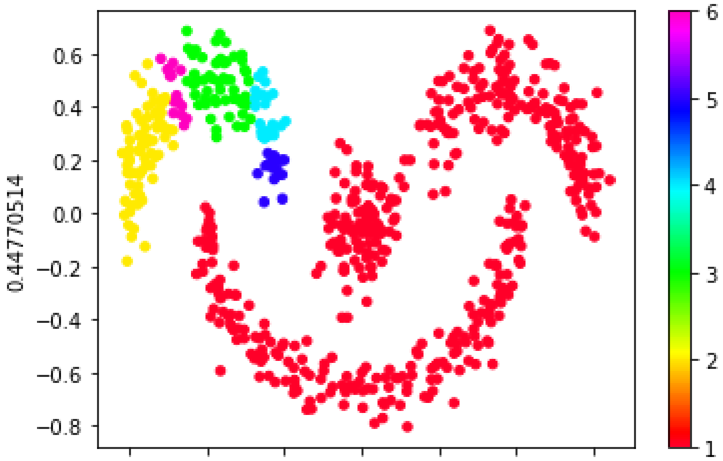


```
Stat Clustering...
```

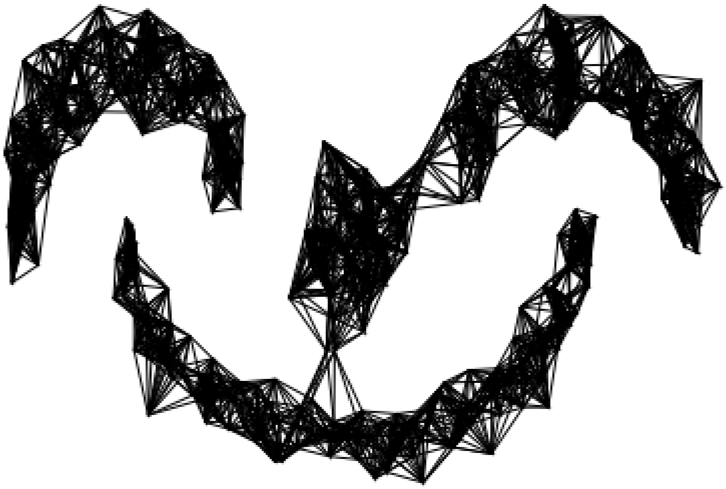
2) For k = 4, k_neighbor_number=5, subCluster=40



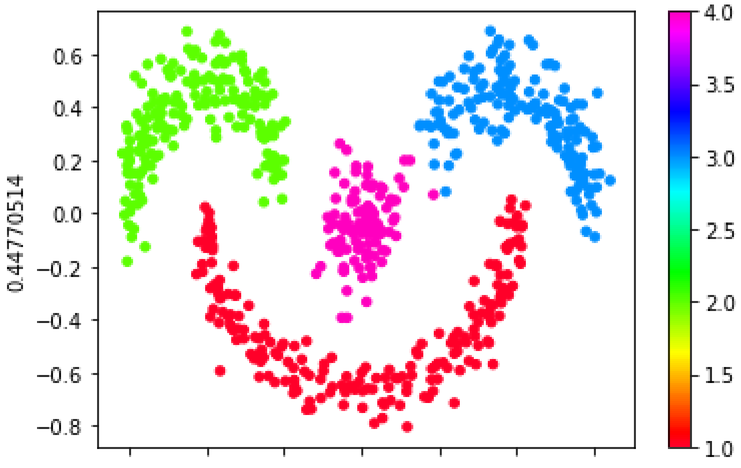
```
Stat Clustering...
```



3)For k = 4 , k_neighbor_number=15, subCluster=40

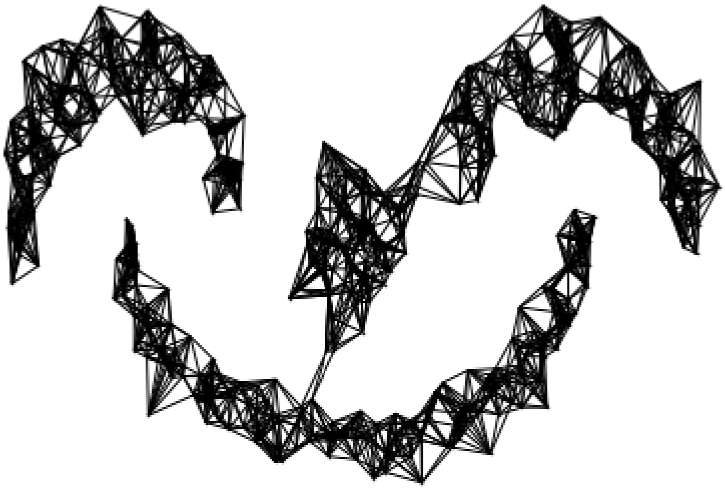


```
Stat Clustering...
```

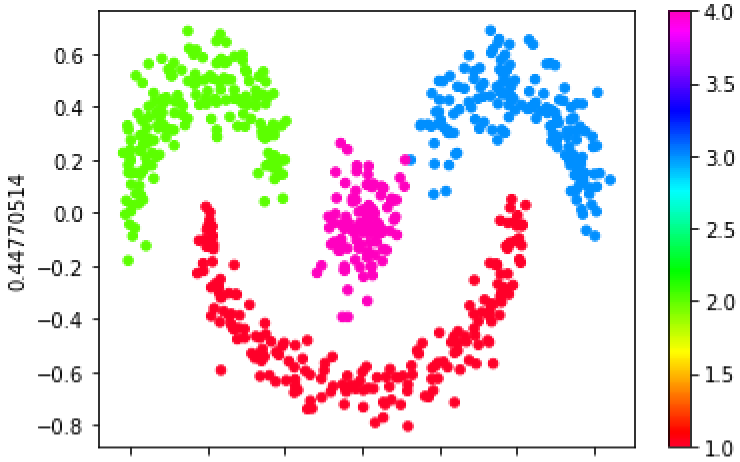


Effect of “subCluster” parameter:

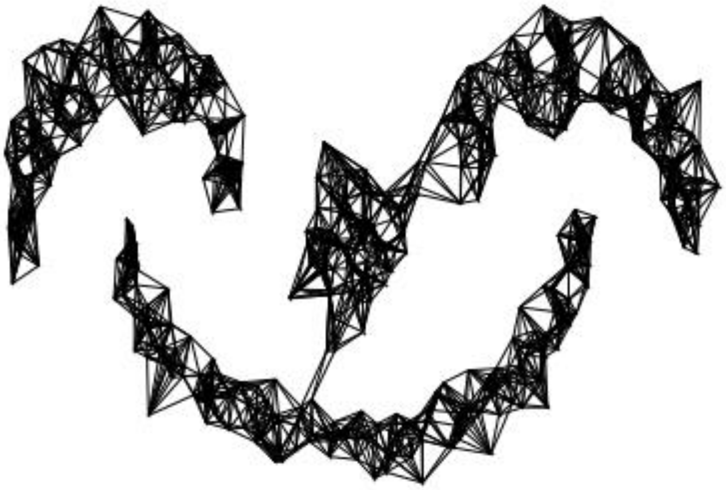
1)For k = 4 , k_neighbor_number=10, subCluster=20



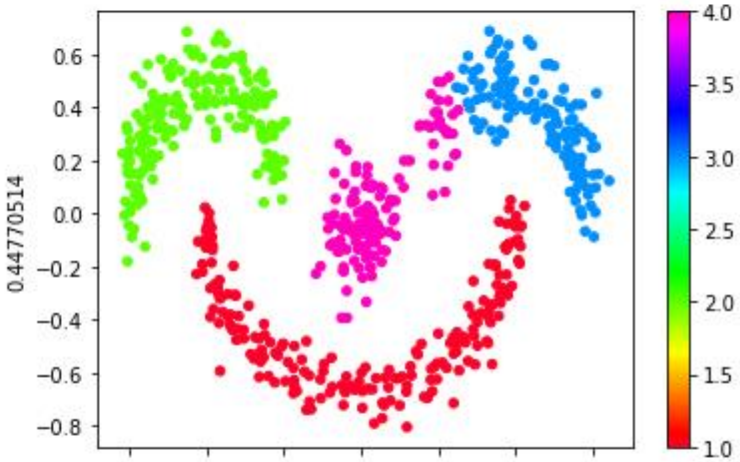
```
Stat Clustering...
```



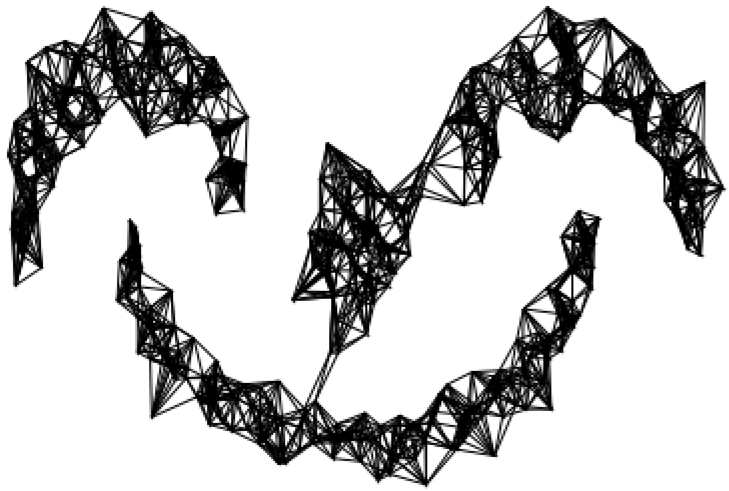
2)For k = 4 , k_neighbor_number=10, subCluster=10



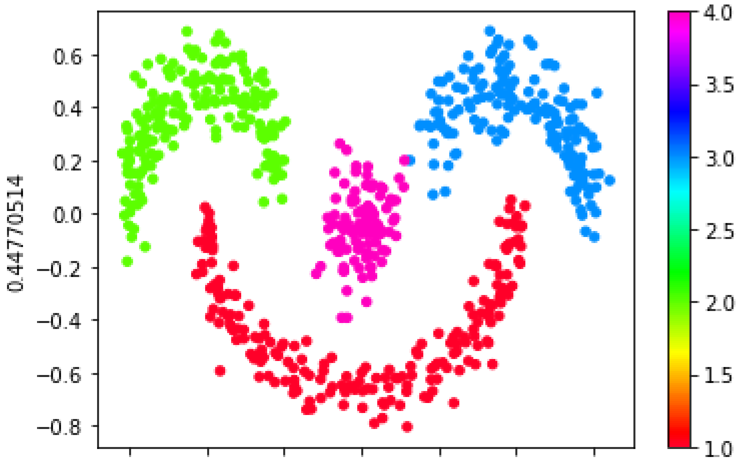
```
Stat Clustering...
```



3)For k = 4 , k_neighbor_number=10, subCluster =40



```
Stat Clustering...
```



3.Write a discussion about how the parameter(s) effect the results.

The number of K determines how many clusters I will have. If I change the number of K_neighbor_number, it determines the nearest point number, which changes the number of clusters. I use the Euclidean distance to determine the nearest point number. When partitioning, all points initially belong to the same cluster. After that, we continue to partition until all clusters are smaller than the subCluster parameter.

4. What are the advantages and disadvantages of the algorithm. Write a discussion about it while comparing it with other clustering techniques.

4.1 Advantages

One of the most important advantages of the chameleon clustering algorithm is that it can produce all kinds of shaping with clustering. In this way it takes into account the special properties of individual clusters. The dynamic modeling methodology of clusters in agglomerative hierarchical methods can be applied to any type of data. It is a dynamic clustering algorithm, so it overcomes the limitations of static clustering algorithms. It also overcomes the cluster similarity constraint as it joins clusters according to their interdependence and proximity.

4.2 Disadvantages

This algorithm doesn't handle noise at all, so data needs to be preprocessed and cleaned to avoid distusrbance. CHAMELEON is known for low dimensional spaces and has not been applied to higher dimensions. It is not very fast and should be used for special occasions. Well known algorithms such as BIRCH and density based DBSCAN. However, the processing cost of high-dimensional data may require O(n^2) time for n objects in the worst case.

5. What is the time complexity of chameleon. How it is comparing to the other clustering techniques. Write a discussion about it.

Chameleon clustering technique depends on 2 parameters. N number of item data items and M number of initial sub-clusters produced by the graph partitioning algorithm. For low-dimensional data set it would need in O(nlogn) . But for high-dimensional data set it would need O(n^2). It is more costly than order models as it will cost more time in high-dimensional data.

In general, there are 4 methods, these are Hierarchical, Partitioning, Density-based, Grid-based.
Hierarchical has 4 algorithms.

Birch -> Time complexity O(n)
Cure -> Time complexity O(n^2logn)
Rock -> Time complexity O(n^2logn)
Chameleon -> Time complexity O(n^2)

Partitioning has 2 algorithms.

K-means -> Time complexity O(n k d)

K-medoids -> Time complexity O(k(n-k)^2)

Density-based has 2 algorithms.

DBSCAN -> Time complexity O(nlogn)

OPTICS -> Time complexity O(nlogn)

Grid-based has 2 algorithms.

STING -> Time complexity O(n)

CLIQUE -> Time complexity O(n+d^2)