CLUSTERING
A Subjective overview

Presented by : Noorzad, Khoshdel, AliakbariKhoie

01. **Cluster Analysis**

A Machine Learning technique

02. **Applications of Clustering**

Social Network Analysis, Costumer Segmentation, ...

03. **Clustering Methods**

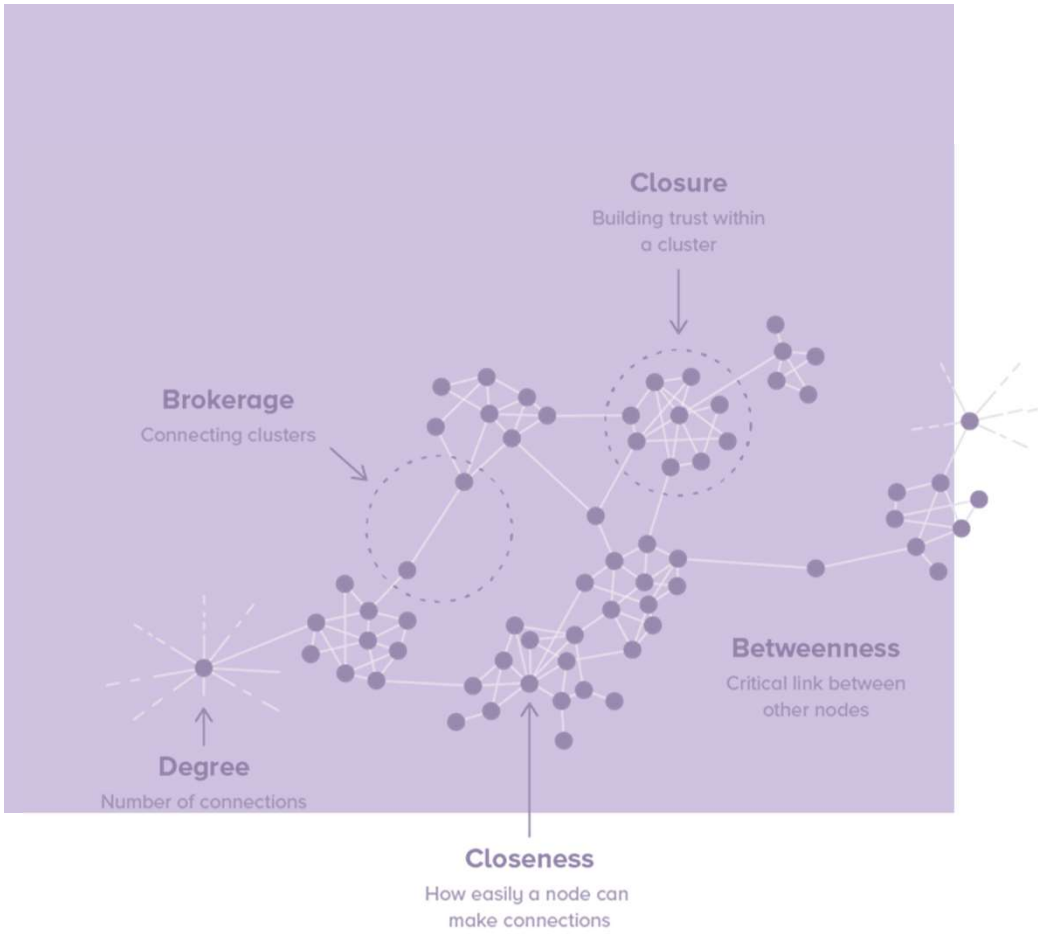
Based on Partition, Density, Hierarchy, ...

04. **Tuning Clustering**

Sillhoutte method, AIC, BIC, ...

05. **Clustering Metrics**

Separation Index – FD index



01
Cluster Analysis

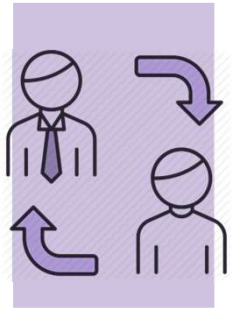
Clustering is a Machine Learning technique that involves the grouping of data points.

Given a set of data points, we can use a clustering algorithm to classify each data point into a specific group.

02

Applications of Clustering

Applications of Clustering

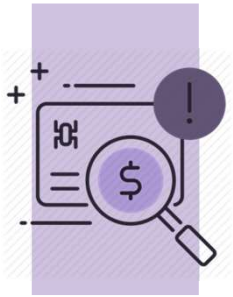


Social Network Analysis

The clustering in social network requires grouping objects into classes based on their links as well as their attributes.

Costumer Segmentation

Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics.



Identifying fraudulent activity

Based on the characteristics of the groups you are then able to classify them into those that are real and which are fraudulent.

03 Clustering Methods

K-means – Based on Partiton

Agglomerative Clustering – based on Hierarchy

Mean shift – Based on Density

Affinity Propagation

Affinity Propagation

Steps

1	Compute *Similarity matrix* using negative pairwise distance of data points.
2	Compute *Adjacency Matrix*, mathematical modeling of how appropriate is to be an exemplar for other data points
3	Compute *Availability Matrix*, mathematical modeling of likeliness of a point to be part of other point's clusters
4	Compute *Criterion Matrix* which is choosing which points to be in a cluster based on common exemplars

DBSCAN – Based on Density

DBSCAN

Steps

1	Checks the possibility of an unvisited point being a sample of estimated PDF or being noise, by measuring density of vicinity points.
2	for all unvisited vicinity numbers until reaching the border of estimated PDF, then checking other points to estimate other potential PDFs.

OPTICS – Based on Density

Spectral Clustering

Spectral Clustering

Steps

1	Compute *Similarity matrix* using pairwise distance of data points. (nodes)
2	Compute *Responsibility Matrix*, where the row and column indices represent the nodes, and the entries represent the absence or presence of an edge between the nodes
3	Compute *Degree matrix* by counting the number of edges connected to each node.
4	Compute *Laplacian Matrix*. Its diagonal is the degree of our nodes, and the off diagonal is the negative edge weights
5	Compute Eigenvalues and Eigen*vectors* of Laplacian Matrix

GMM – Based on Density

GMM

Steps

1	E – step : use current best knowledge of the center and shapes of each cluster to calculate the probability of belonging to cluster
2	M – step : use current best knowledge of which class each point belongs to update and improve our estimates for the center and shape of each cluster

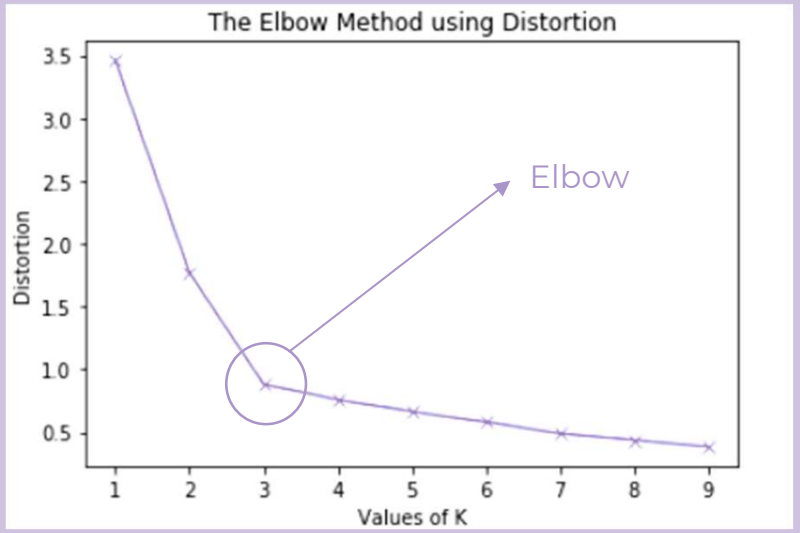
04

Tuning Clustering Methods

Elbow Method

In cluster analysis, the elbow method is a heuristic used in determining the number of clusters in a data set.

Silhouette coefficient and BIC score are better alternatives to the elbow method for visually discerning the optimal number of clusters.



Silhouette Method

The silhouette method computes silhouette coefficients of each point that measure how much a point is similar to its own cluster compared to other clusters.

The value of the silhouette ranges between [1, -1], where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters

Information Criterion

a method for scoring and selecting a model

The Bayesian Information Criterion:

$$BIC: -2 \times LL + \log(N) \times k$$

BIC penalizes the model more for its complexity, meaning that more complex models have worse scores

The Akaike Information Criterion:

$$AIC: -\frac{2}{N} \times LL + \log(N) \times \frac{k}{N}$$

AIC penalizes complex model less, meaning that it may put more emphasis on model performance

05

Clustering Metrics

Separation Index

Separation index is based on the ratio of the inter cluster distance to intra cluster distance, and then minimizing it across all the possible pairs.

$$d(S_i, S_j) = \min_{x,y}\{d(x, y | x \epsilon S_i, y \epsilon S_j)\}$$

$$d(S_l, S_l) = \max_{x,y}\{d(x, y | x, y \epsilon S_l)\}$$

$$SI = \min_{j} \left\{ \min_{i(i \neq j)} \left\{ \frac{d(S_i, S_j)}{\max_l d(S_l, S_l)} \right\} \right\}$$

Fisher's Discrimination Index

like Separation, it maximizes a function by using covariance matrices of within and between cluster data points.

Between variability is calculated among centroids

Within variability is simply the sum of covariance matrices of the data points in a cluster.

$$S_w = \sum S_i$$

$$S_B = \sum Q_i(\widehat{\mu_i} - \hat{\mu})(\widehat{\mu_i} - \hat{\mu})^T$$

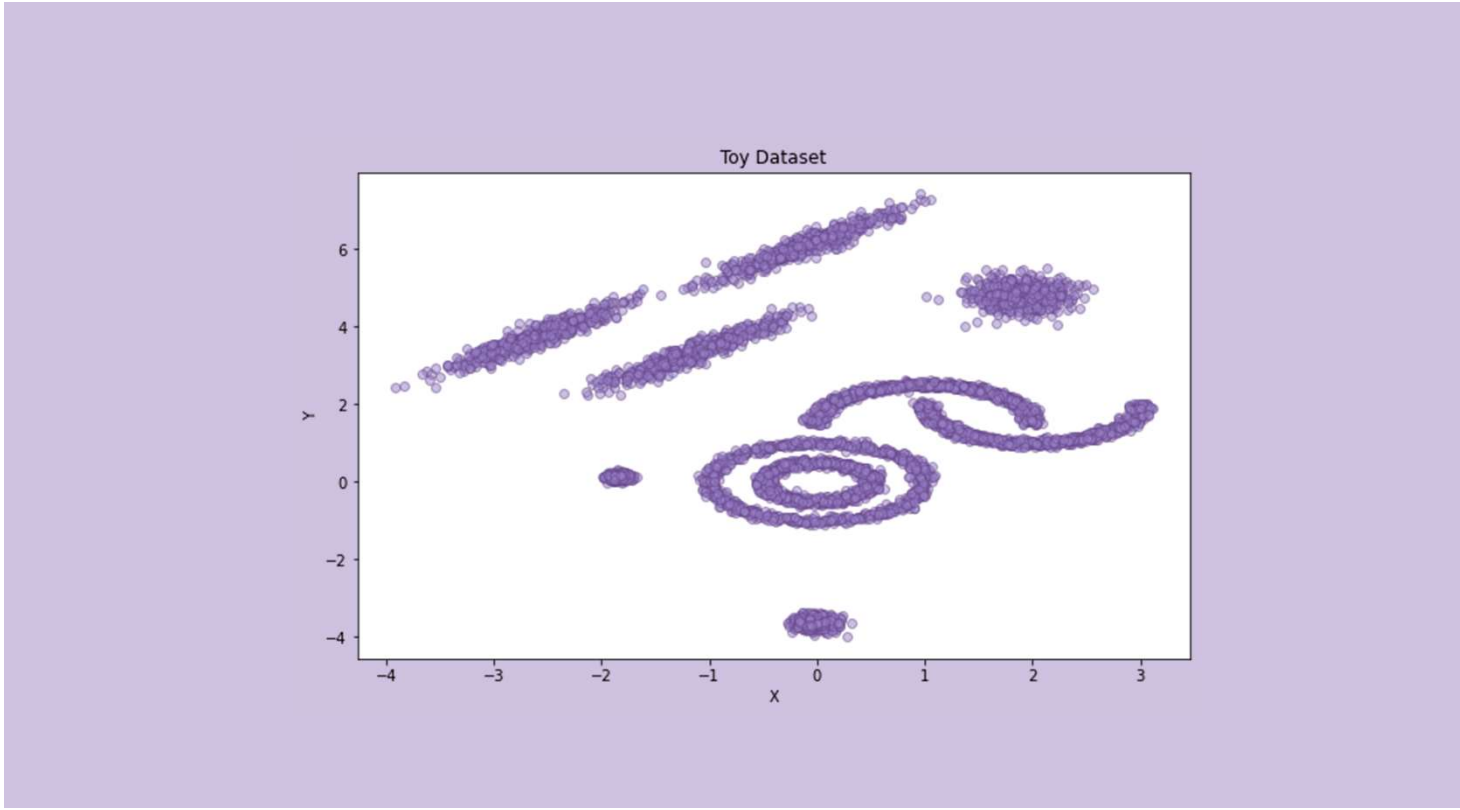
$$S_i = \sum (\widehat{\mu_i} - \hat{\mu})(\widehat{\mu_i} - \hat{\mu})^T$$

$$FDI = trace(S_W^{-1} S_B)$$

06

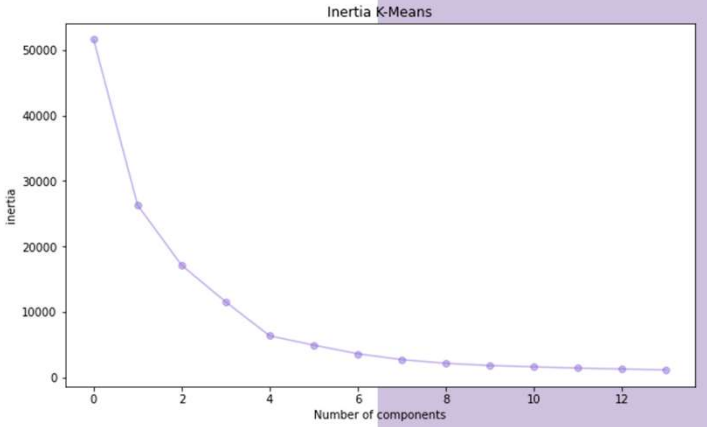
Study

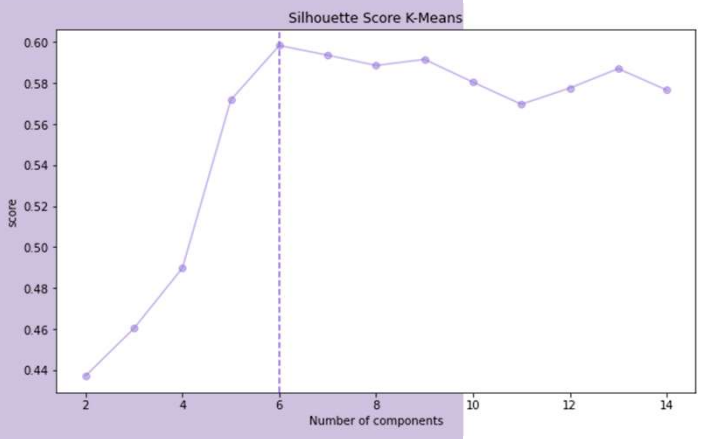
TOY dataset



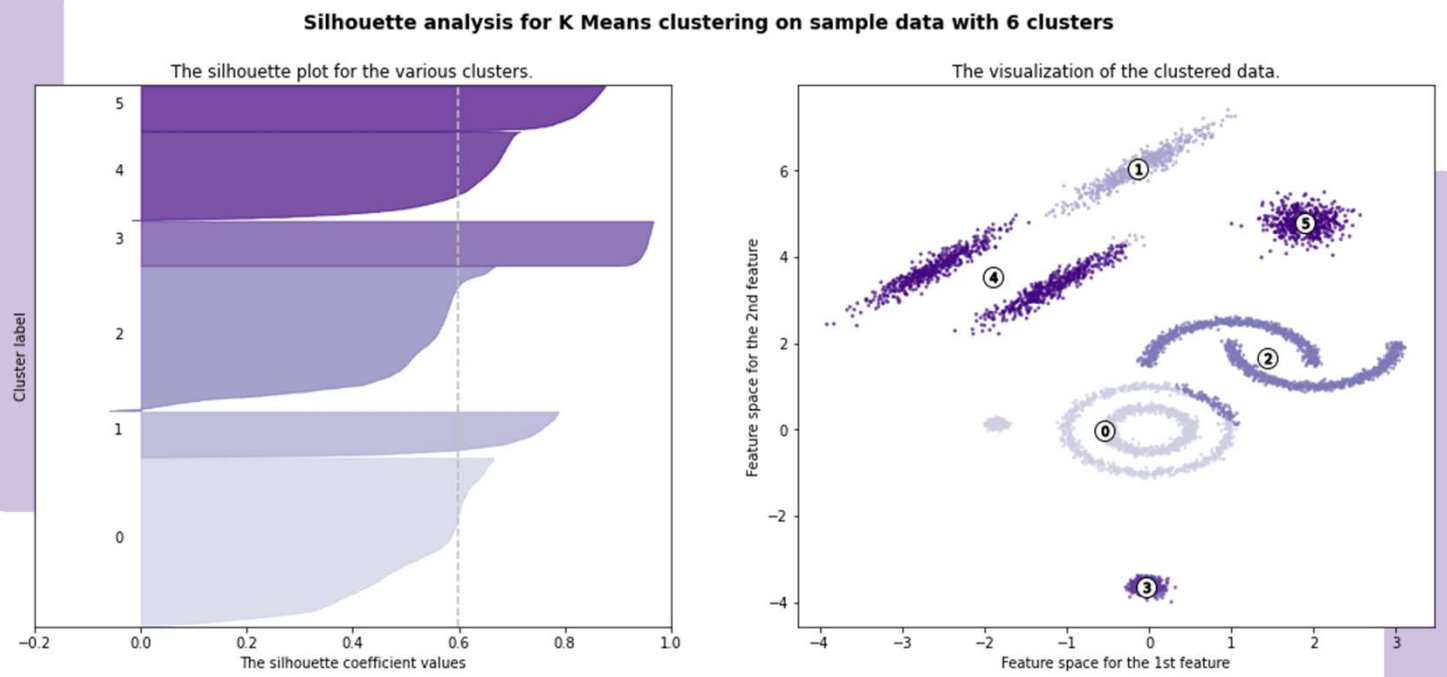
K-Means

Elbow Method : Doesn’t help with the clusters



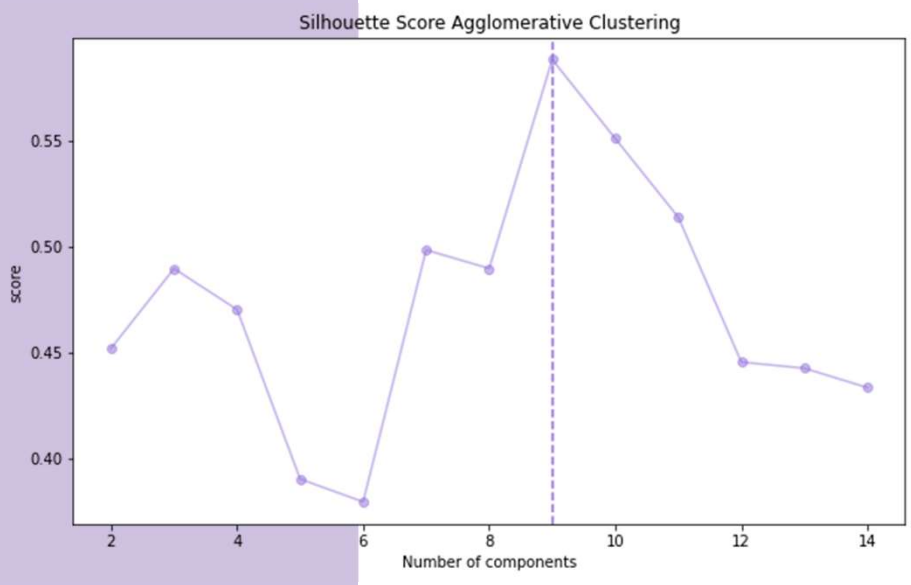


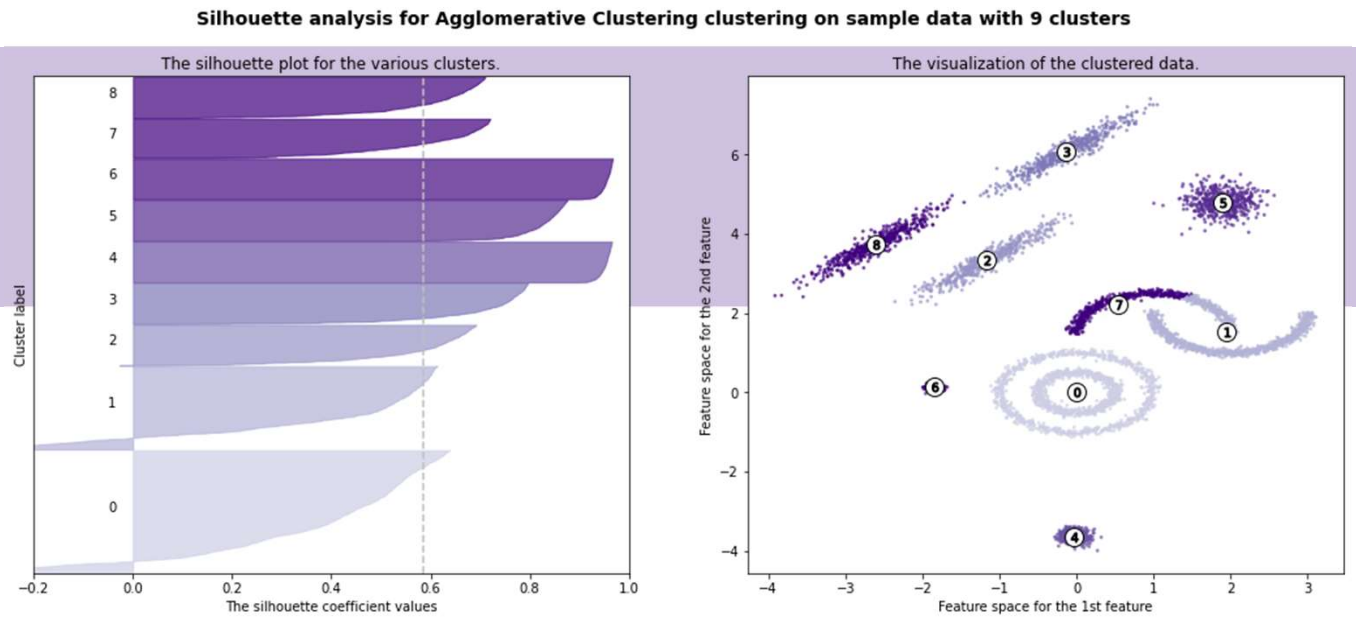
Sillhoutte Score : 6 clusters gives out the best score


Silhouette analysis for K Means clustering on sample data with 6 clusters
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature

Agglomerative Clustering

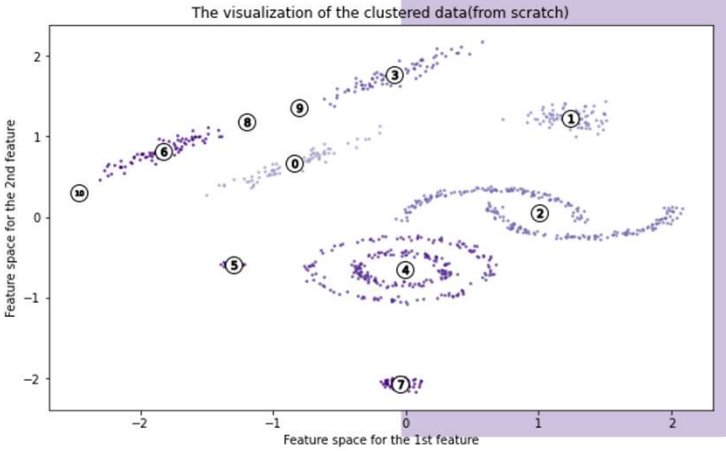
According to Sillhouette score, 9 clusters best represent our dataset.

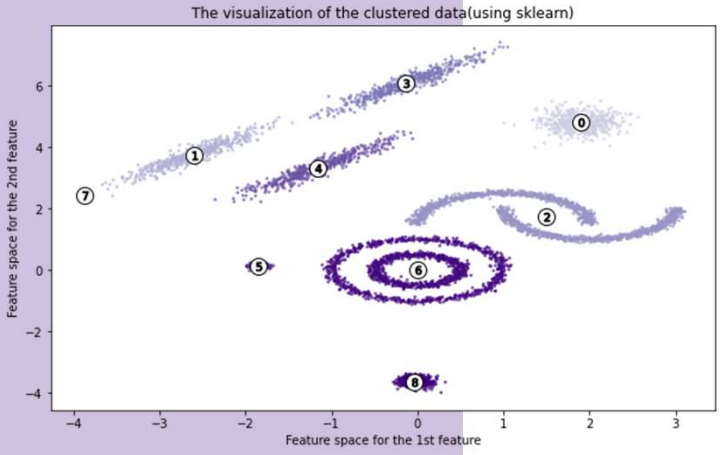



Silhouette analysis for Agglomerative Clustering clustering on sample data with 9 clusters
The silhouette plot for the various clusters.
Cluster label
8
7
6
5
4
3
2
1
0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
6
4
2
0
−2
−4
−4
−3
−2
−1
0
1
2
3
Feature space for the 1st feature

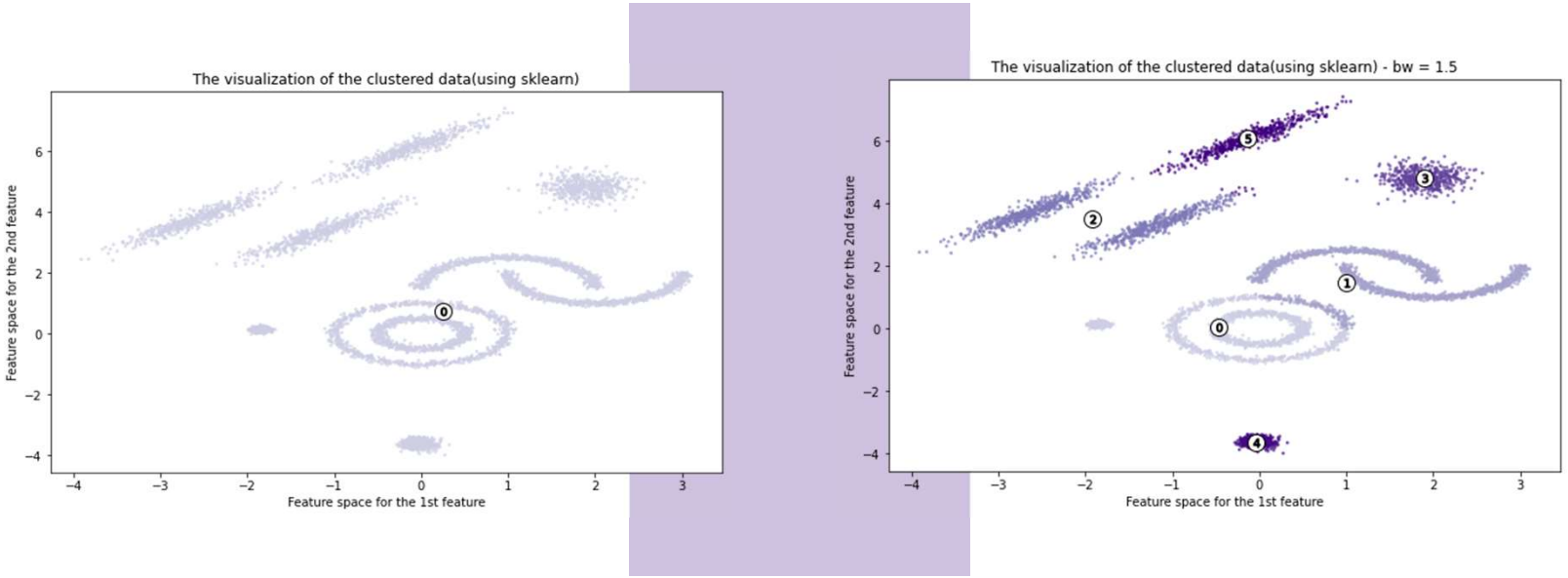
Agglomerative Clustering

clustered using samples



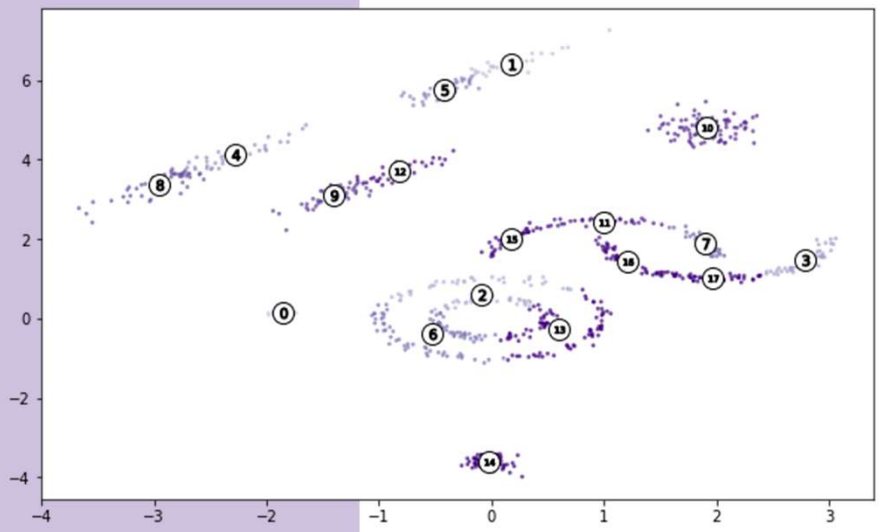


Mean Shift

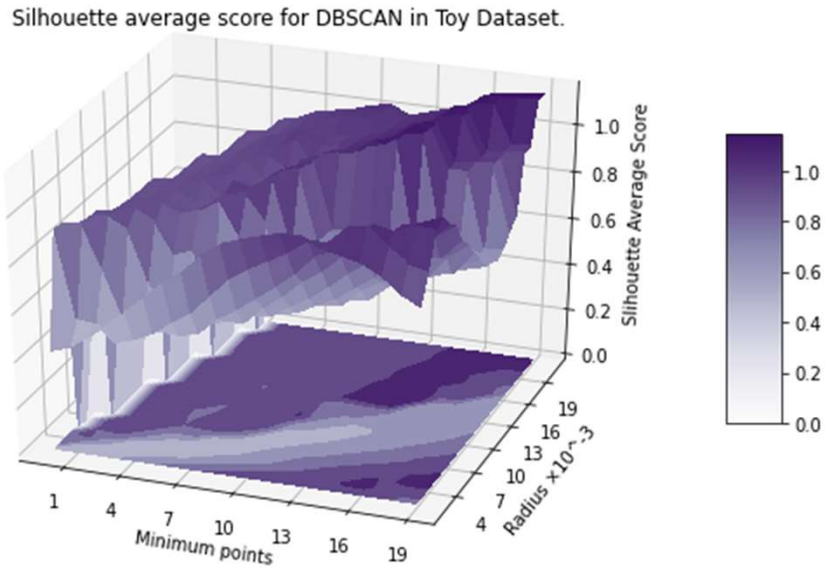

The visualization of the clustered data(using sklearn)
Feature space for the 2nd feature
Feature space for the 1st feature
0
The visualization of the clustered data(using sklearn) - bw = 1.5
Feature space for the 2nd feature
Feature space for the 1st feature
0
1
2
3
4
5

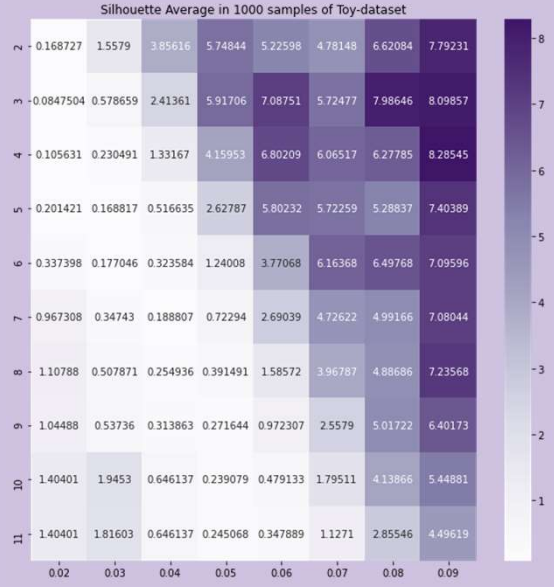
Affinity Clustering

clustered using samples and 0.8 damping factor !

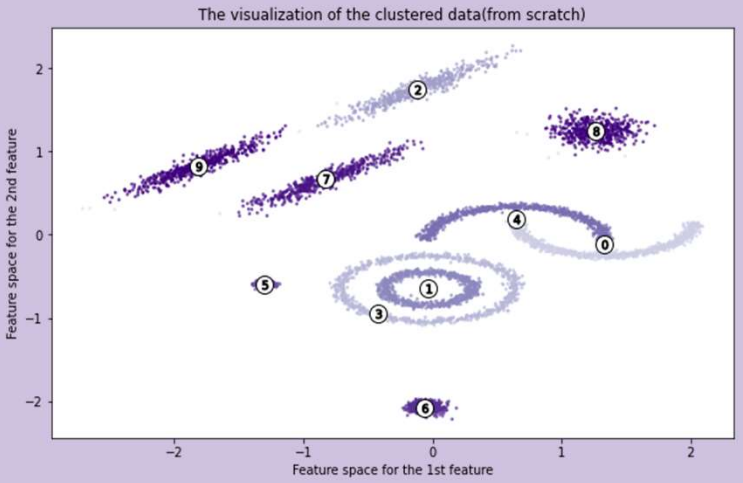


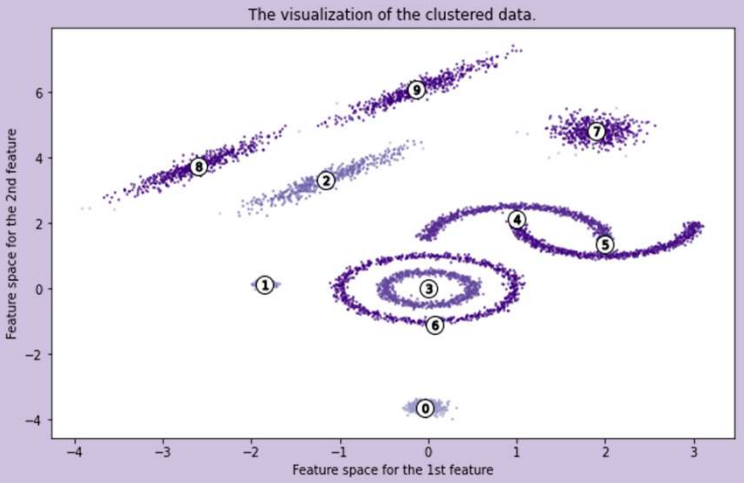
DBSCAN


Silhouette average score for DBSCAN in Toy Dataset.
Silhouette Average Score
1.0
0.8
0.6
0.4
0.2
0.0
Minimum points
1
4
7
10
13
16
19
Radius x10^-3
4
7
10
13
16
19


Silhouette Average in 1000 samples of Toy-dataset
2 0.168727 1.5579 3.85616 5.74844 5.22598 4.78148 6.62084 7.79231
3 0.0847504 0.578659 2.41361 5.91706 7.08751 5.72477 7.98646 8.09857
4 0.105631 0.230491 1.33167 4.15953 6.80209 6.06517 6.27785 8.28545
5 0.201421 0.168817 0.516635 2.62787 5.80232 5.72259 5.28837 7.40389
6 0.337398 0.177046 0.323584 1.24008 3.77068 6.16368 6.49768 7.09596
7 0.967308 0.34743 0.188807 0.72294 2.69039 4.72622 4.99166 7.08044
8 1.10788 0.507871 0.254936 0.391491 1.58572 3.96787 4.88686 7.23568
9 1.04488 0.53736 0.313863 0.271644 0.972307 2.5579 5.01722 6.40173
10 1.40401 1.9453 0.646137 0.239079 0.479133 1.79511 4.13866 5.44881
11 1.40401 1.81603 0.646137 0.245068 0.347889 1.1271 2.85546 4.49619
0.02 0.03 0.04 0.05 0.06 0.07 0.08 0.09
8
7
6
5
4
3
2
1

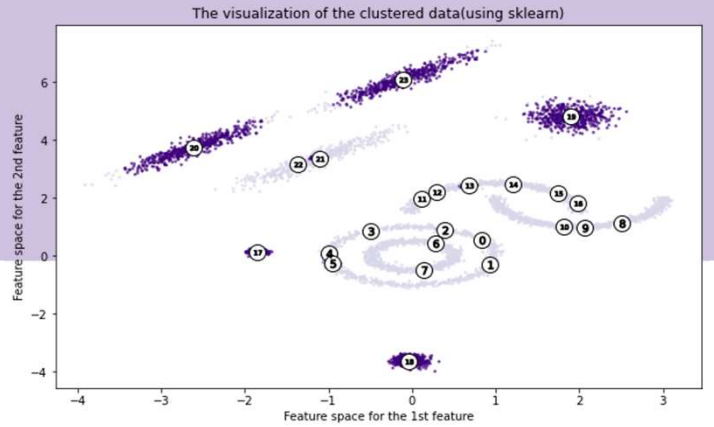
DBSCAN

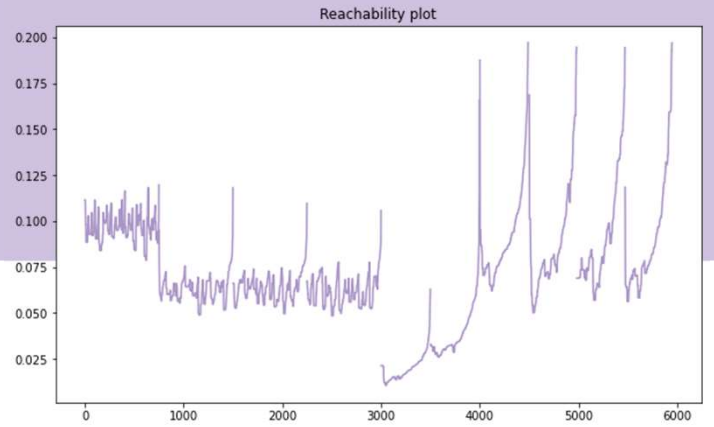

The visualization of the clustered data(from scratch)
Feature space for the 2nd feature
Feature space for the 1st feature


The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature

OPTICS

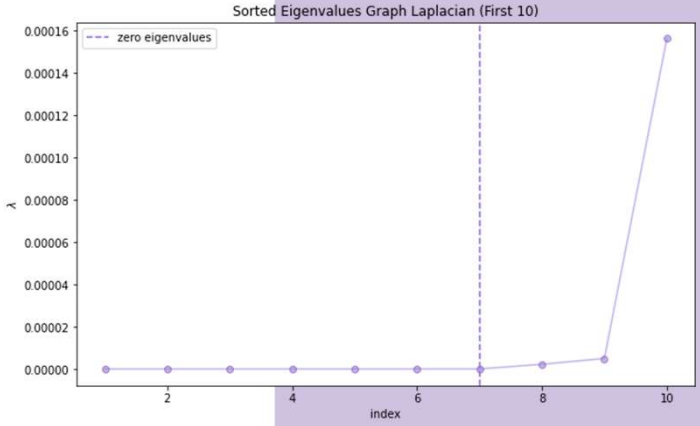
Estimated Number of Clusters : 25
Estimated Number of Noise points : 2903



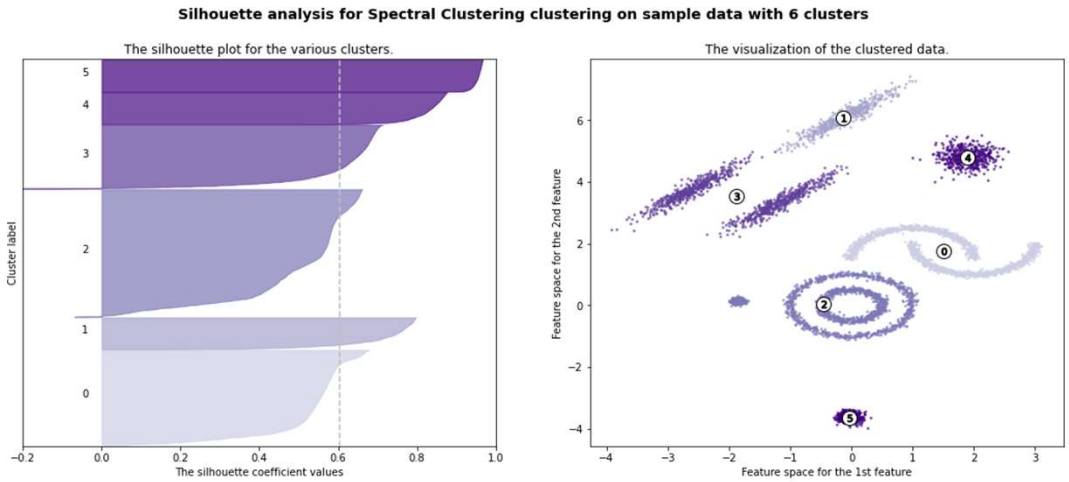


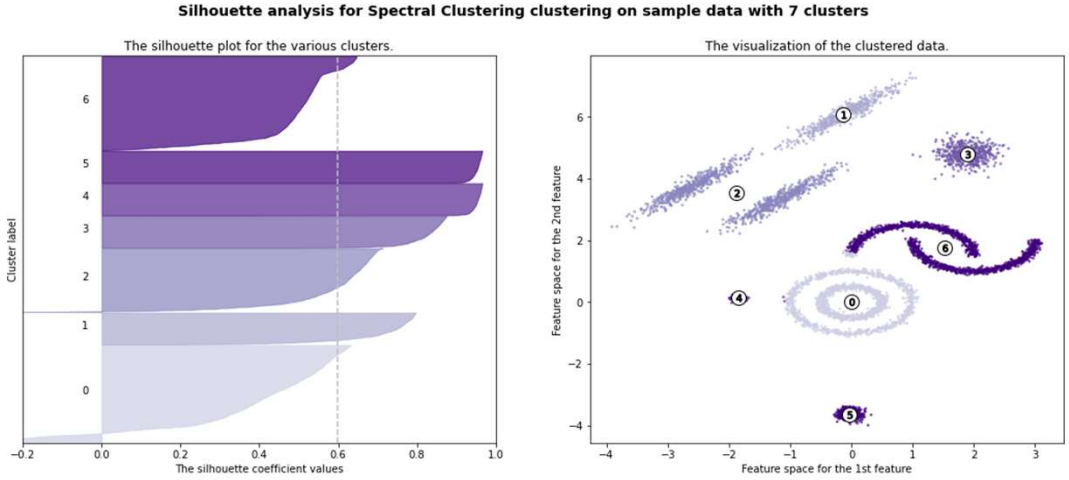
Spectral Clustering

Optimal Number of Components turned out to be 7 according to spectral gap

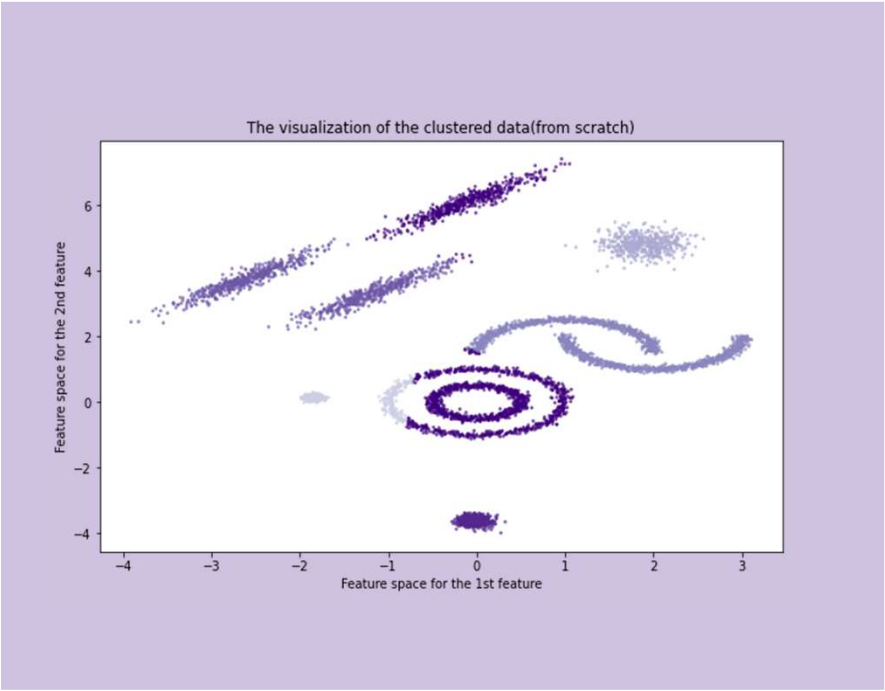


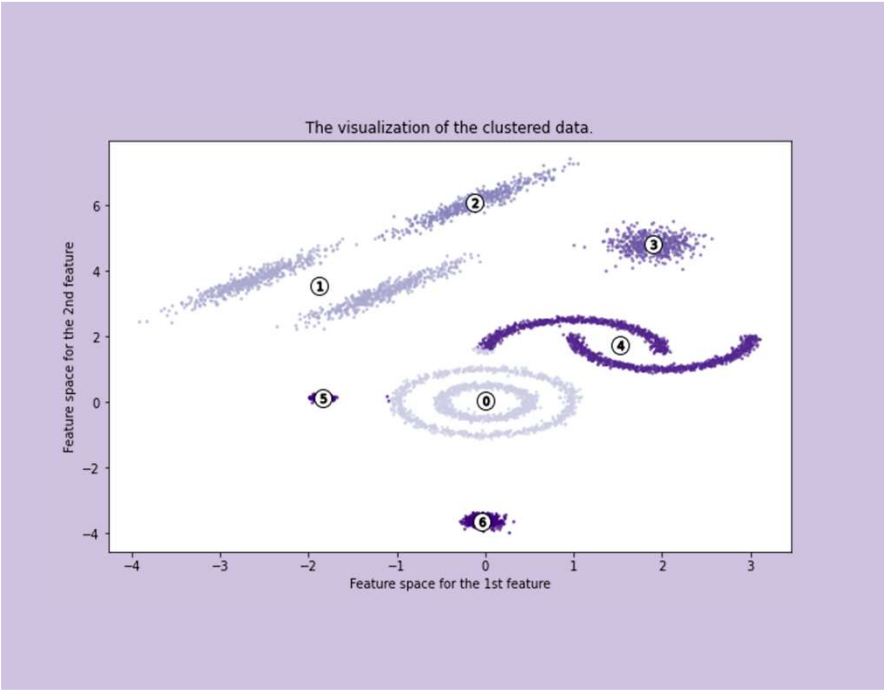
Optimal Number of Components turned out to be 7 according to Sillhoutte score


Silhouette analysis for Spectral Clustering clustering on sample data with 6 clusters
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature


Silhouette analysis for Spectral Clustering clustering on sample data with 7 clusters
The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature

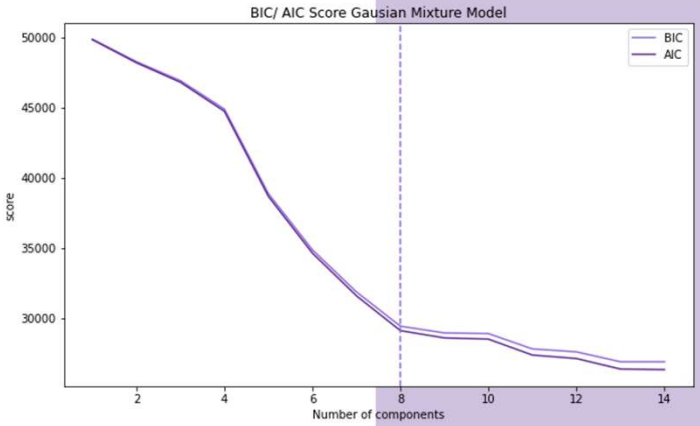
Spectral Clustering

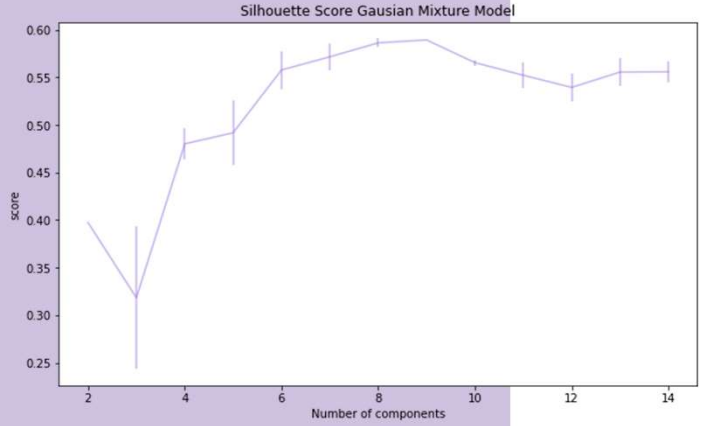




GMM

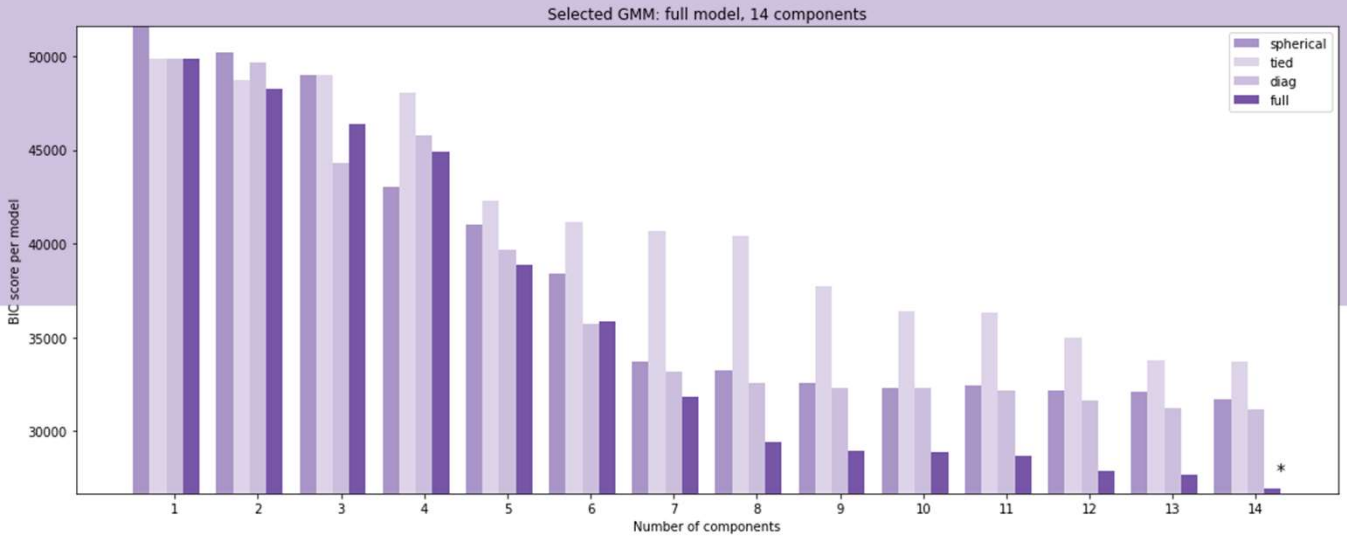
BIC/AIC score : can’t really tell …



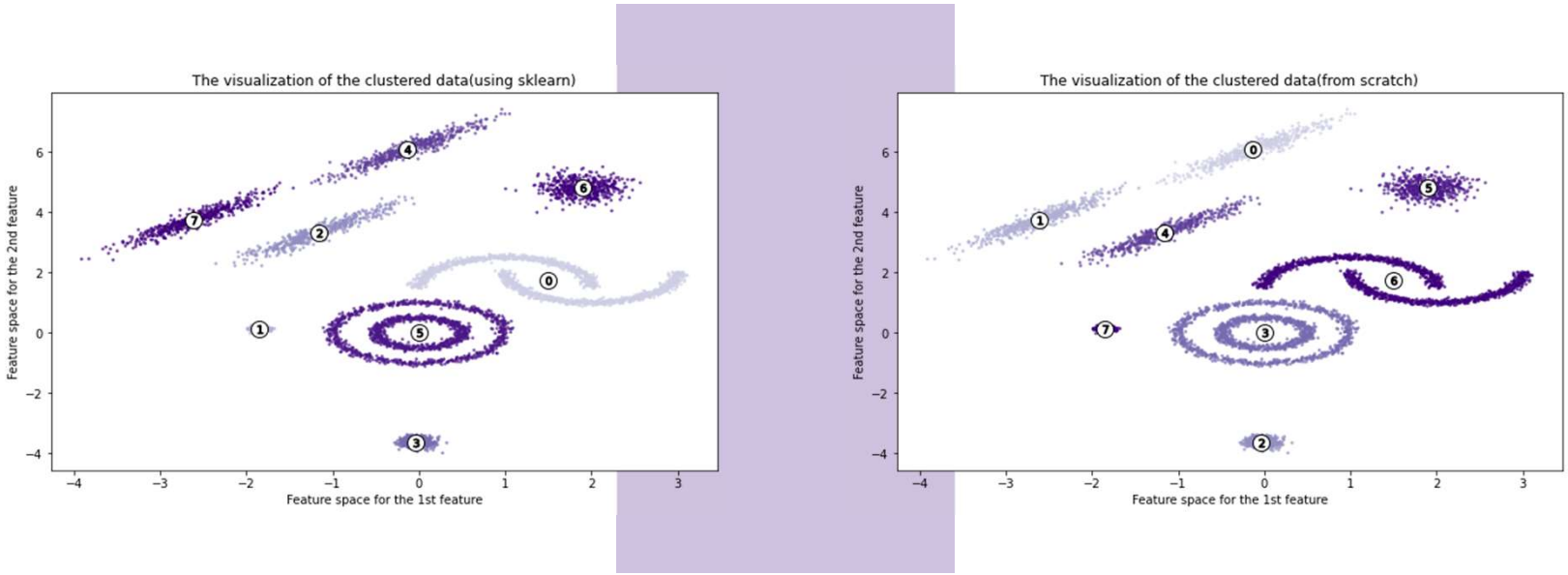


Sillhoutte Score : 8 clusters gives out the best score and lease amount of stochasticity

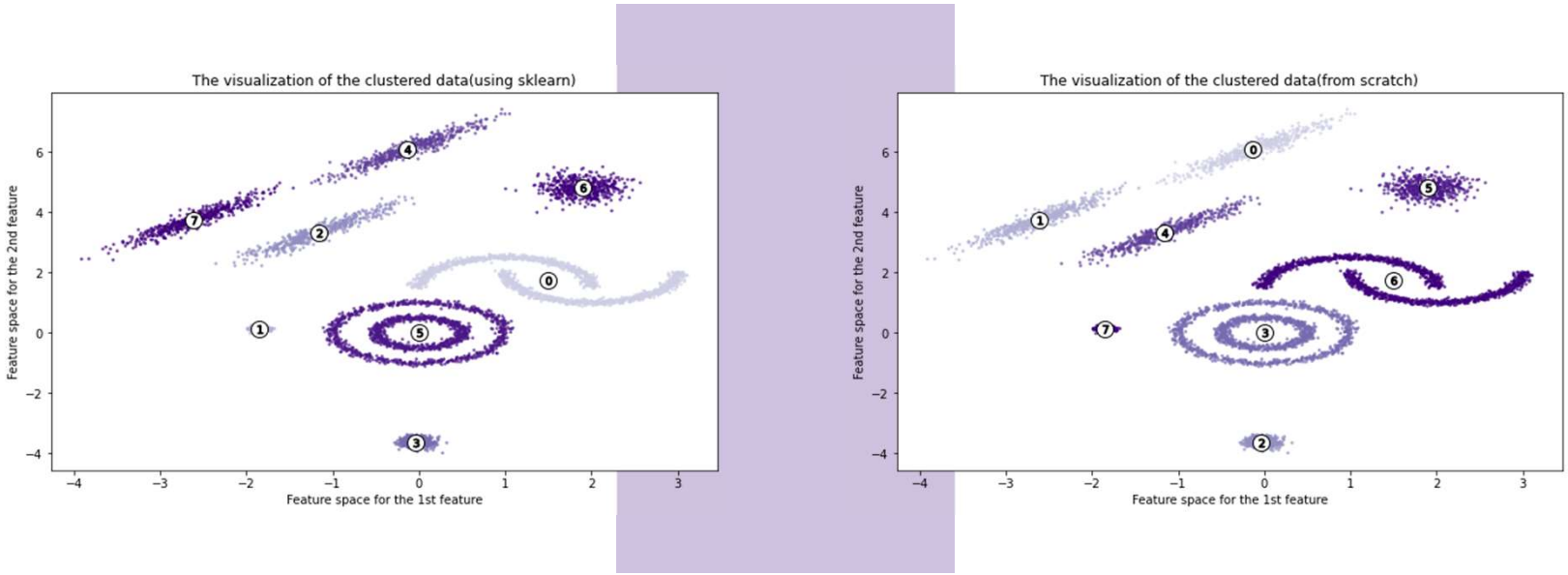
GMM – BIC score – Different covariance types



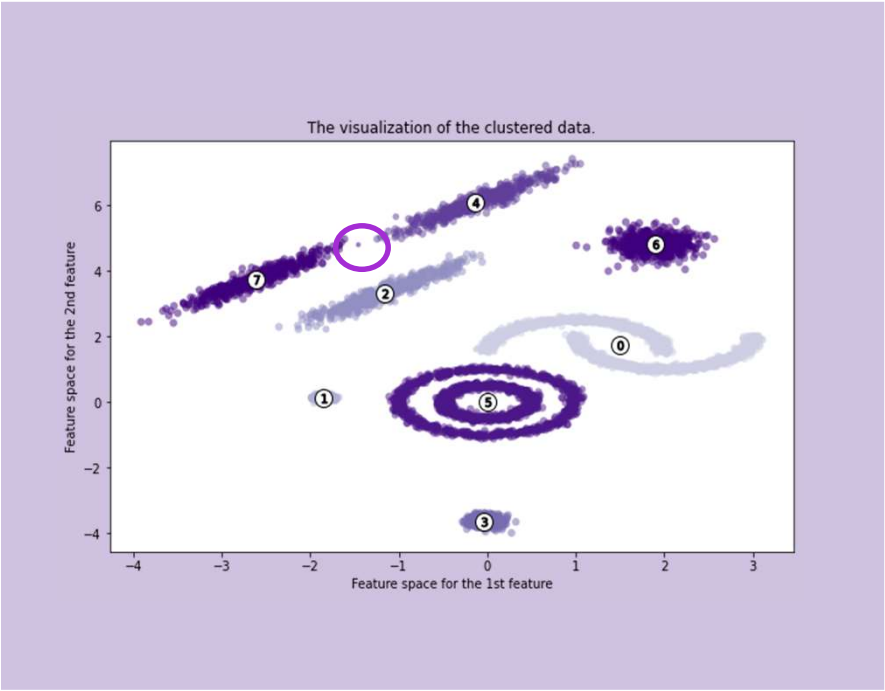
GMM

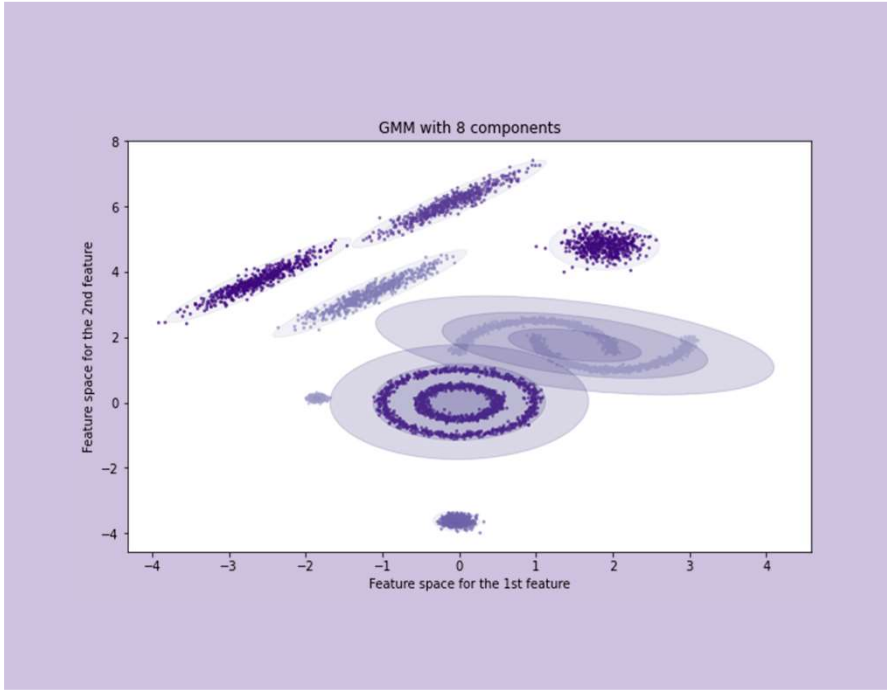

The visualization of the clustered data(using sklearn)
Feature space for the 2nd feature
Feature space for the 1st feature
The visualization of the clustered data(from scratch)
Feature space for the 2nd feature
Feature space for the 1st feature

GMM

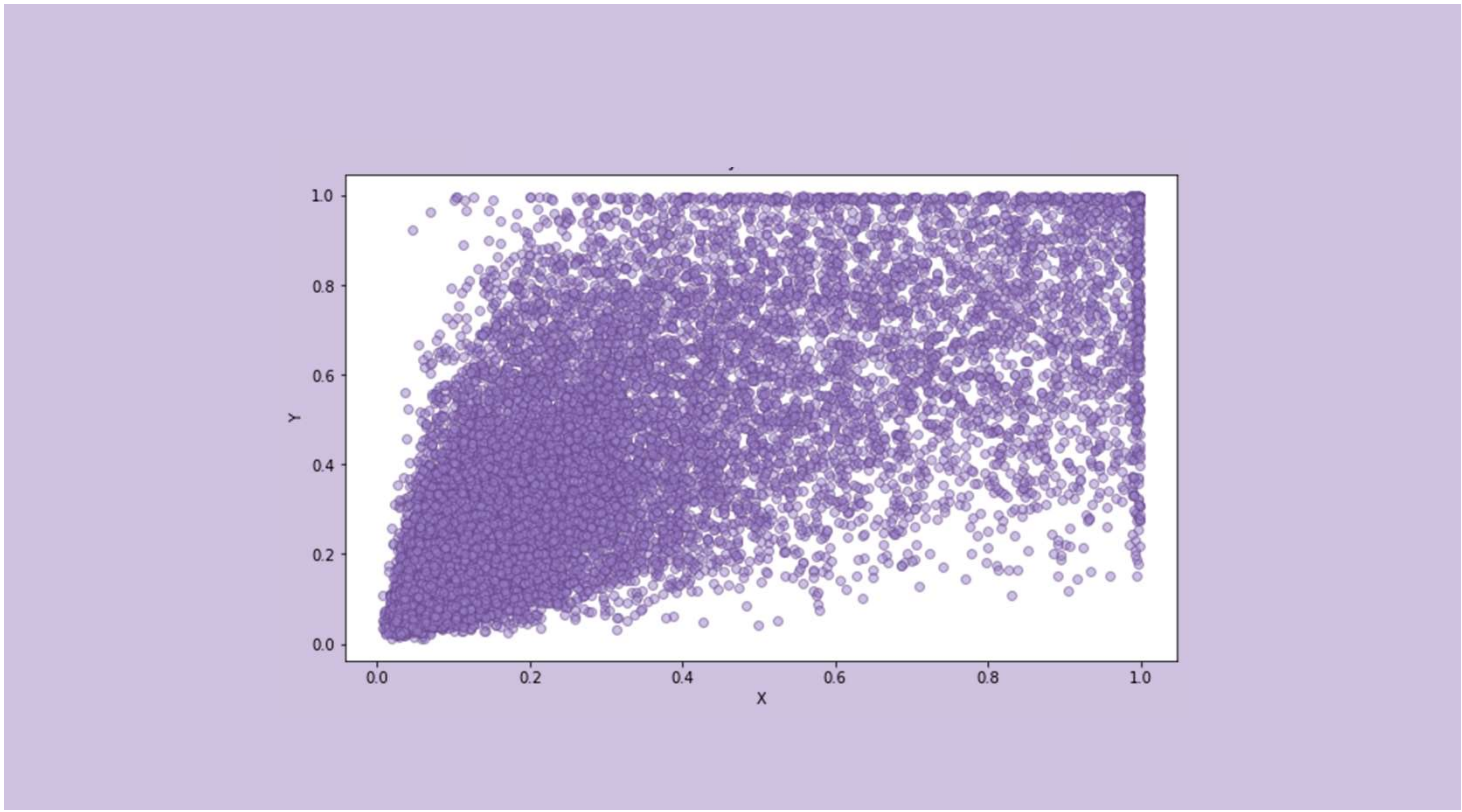

The visualization of the clustered data(using sklearn)
Feature space for the 2nd feature
Feature space for the 1st feature
The visualization of the clustered data(from scratch)
Feature space for the 2nd feature
Feature space for the 1st feature

GMM


The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature
7
4
6
2
0
1
5
3

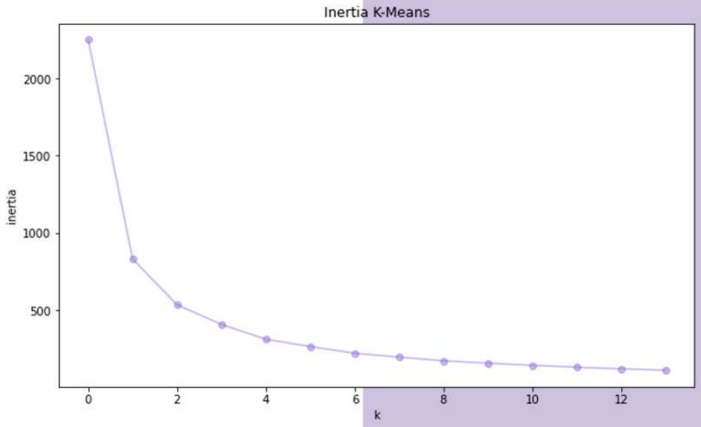

GMM with 8 components
Feature space for the 2nd feature
Feature space for the 1st feature

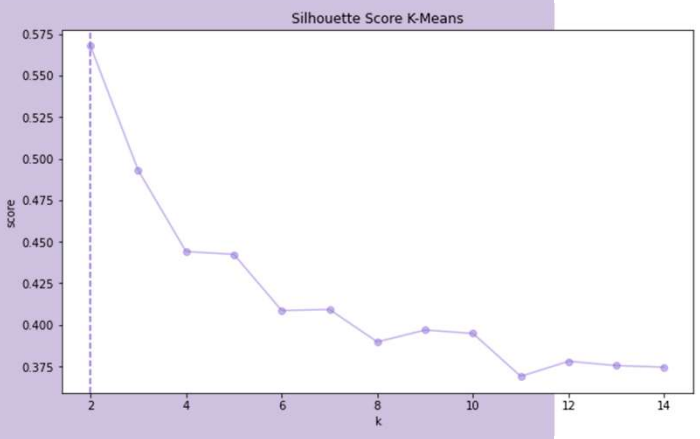
BBOX dataset



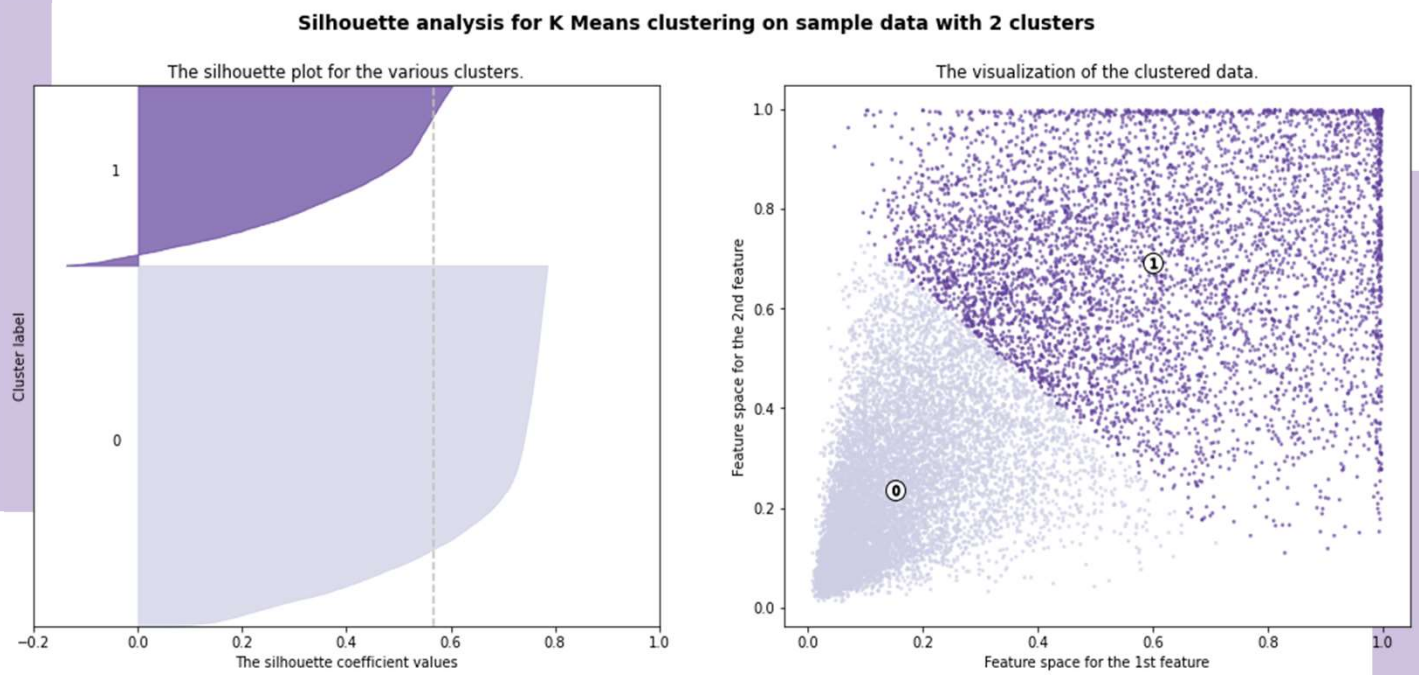
K-Means

Elbow Method : Doesn’t help with the clusters



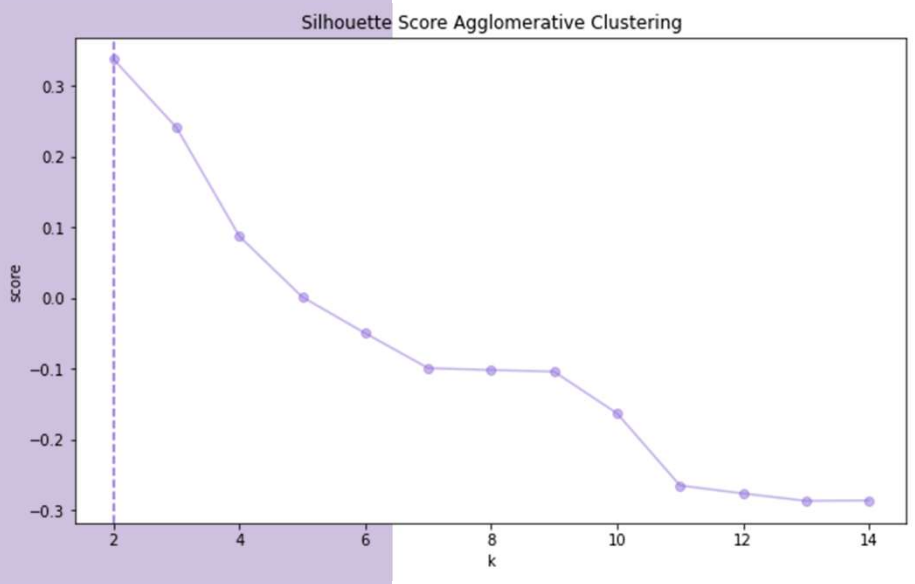


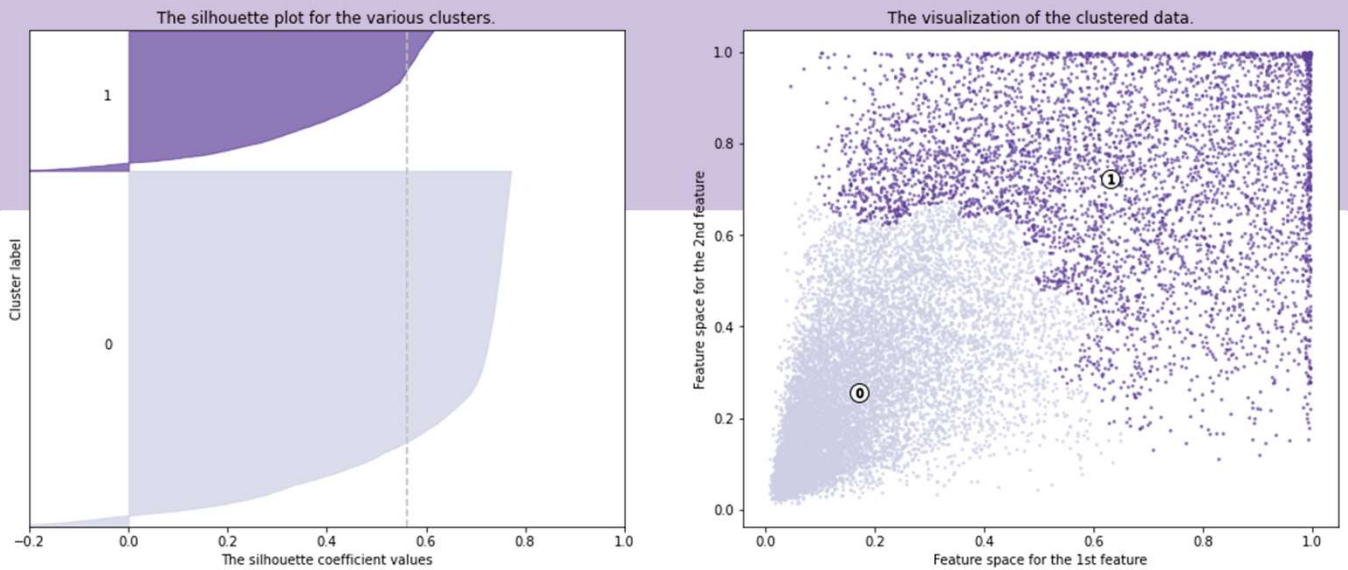
Sillhoutte Score : 2 clusters gives out the best score


Silhouette analysis for K Means clustering on sample data with 2 clusters
The silhouette plot for the various clusters.
Cluster label
1
0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature
0.0
0.2
0.4
0.6
0.8
1.0
1
0

Agglomerative Clustering

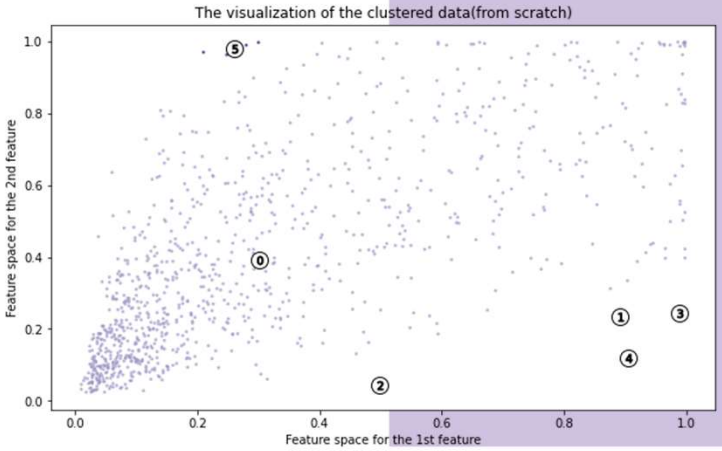
According to Sillhouette score, 2 clusters best represent our dataset.

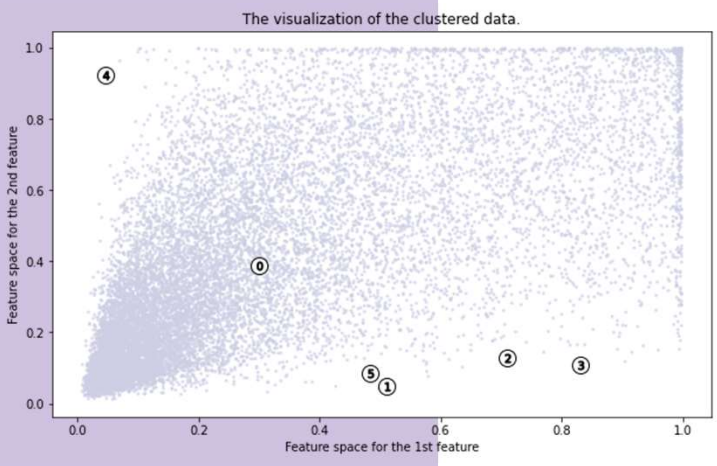



Silhouette analysis for Agglomerative Clustering clustering on sample data with 2 clusters
The silhouette plot for the various clusters.
Cluster label
1
0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature
1
0

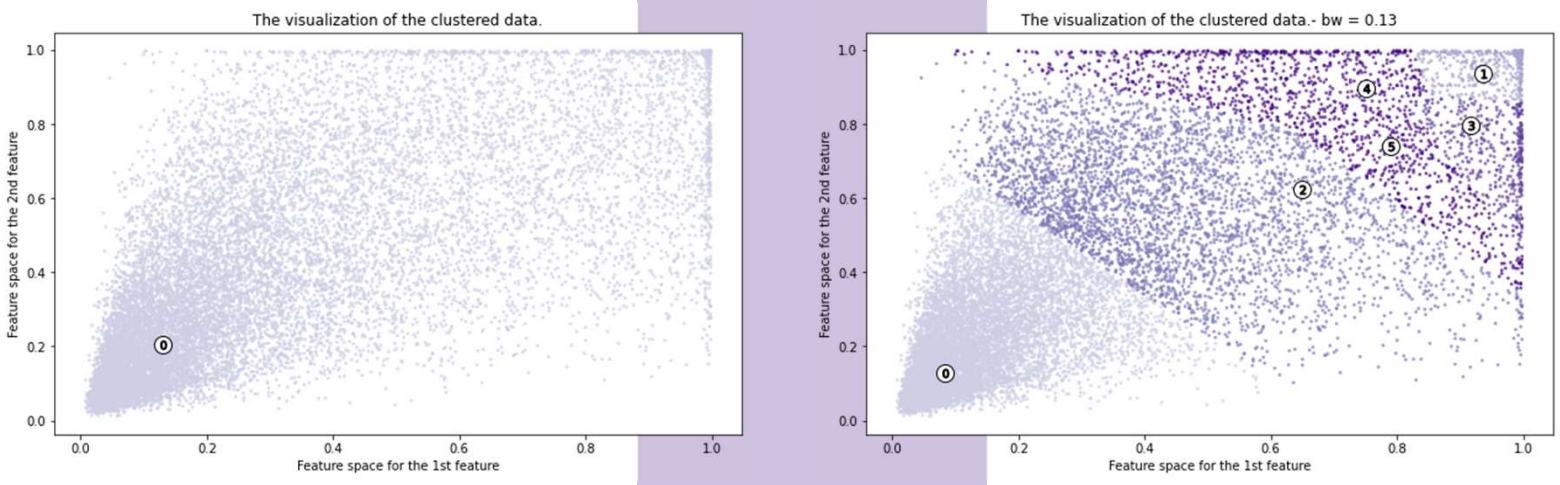
Agglomerative Clustering

clustered using samples



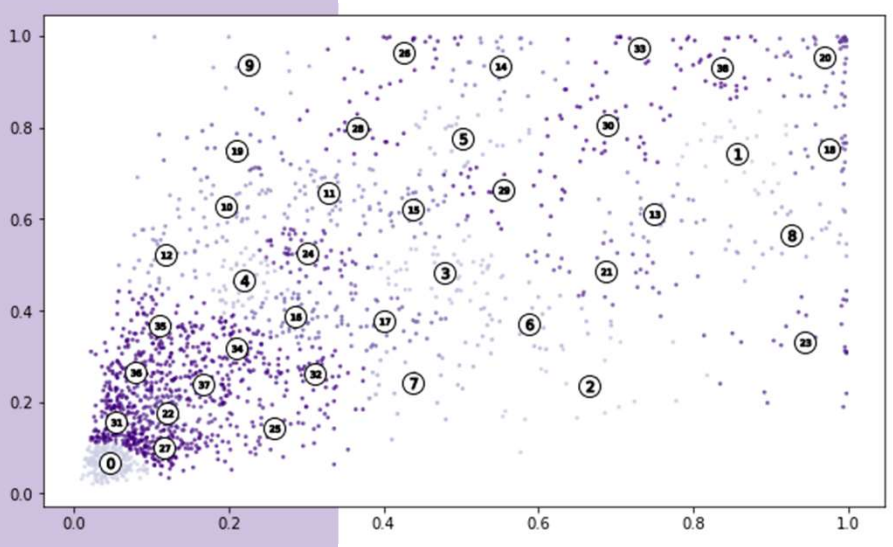


Mean Shift

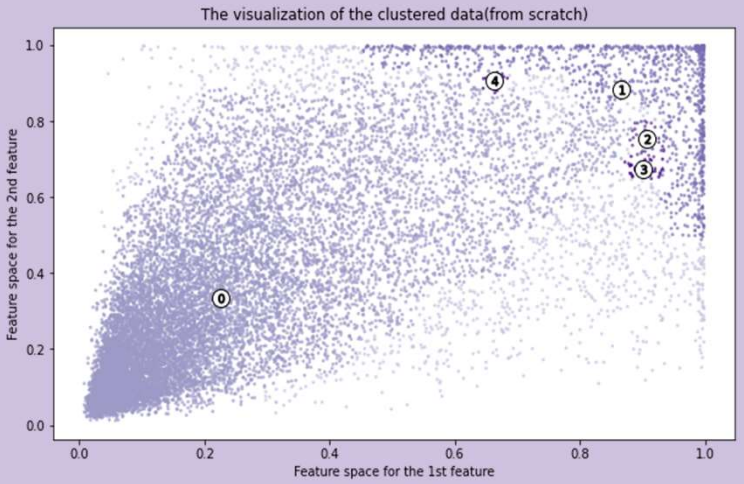

The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature
0
The visualization of the clustered data.- bw = 0.13
Feature space for the 2nd feature
Feature space for the 1st feature
0
1
2
3
4
5

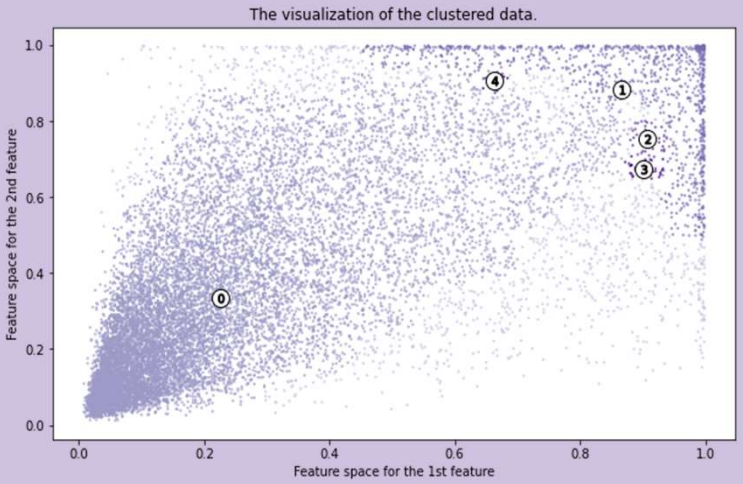
Affinity Clustering

clustered using samples and 0.8 damping factor !



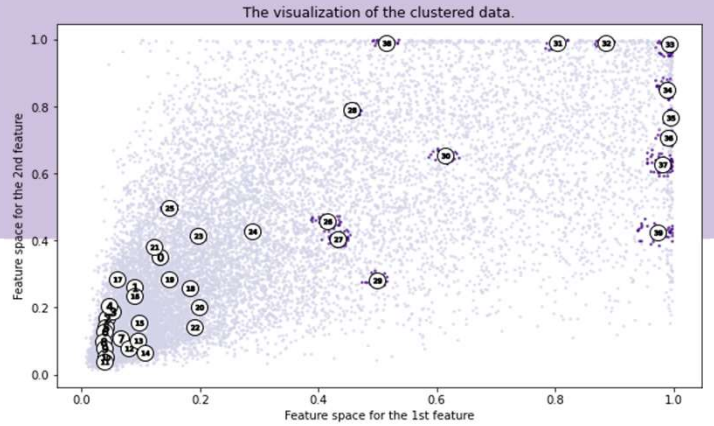
DBSCAN

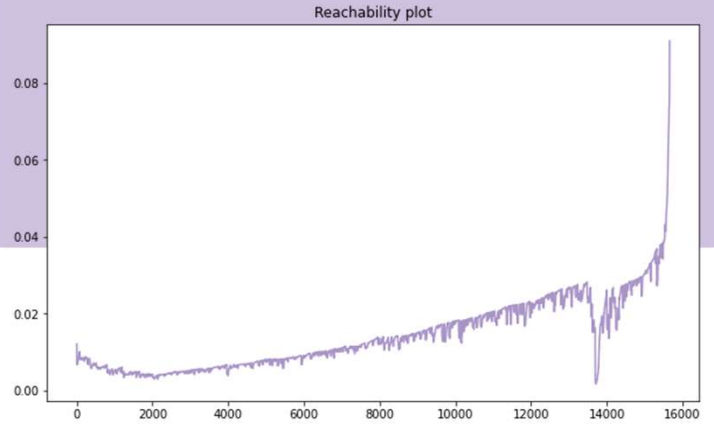

The visualization of the clustered data(from scratch)
Feature space for the 2nd feature
Feature space for the 1st feature
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4


The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4

OPTICS

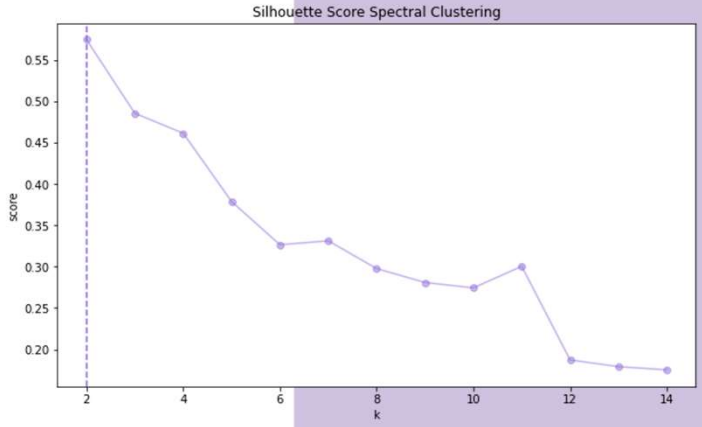
Estimated Number of Clusters : 41
Estimated Number of Noise points : 14386

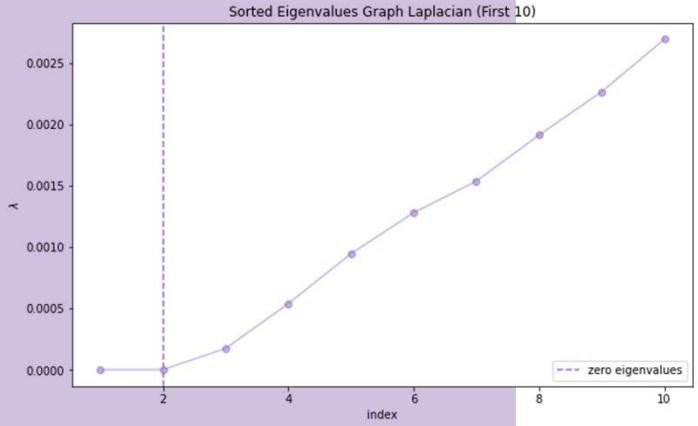




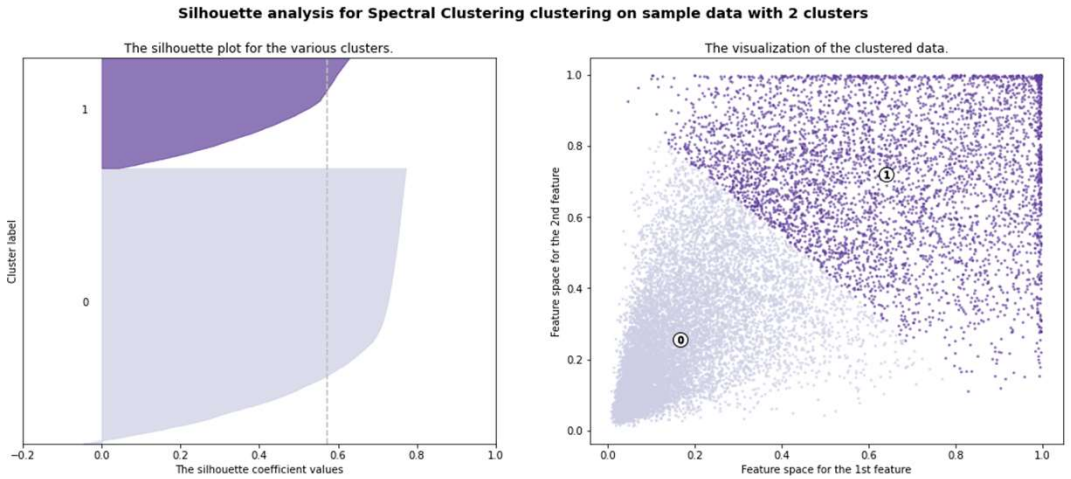
Spectral Clustering

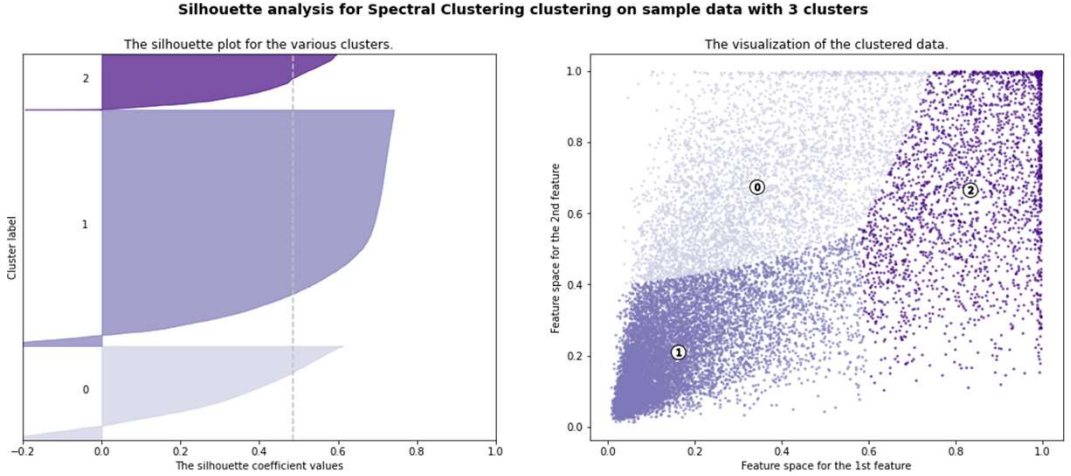
Optimal Number of Components turned out to be 2 according to spectral gap



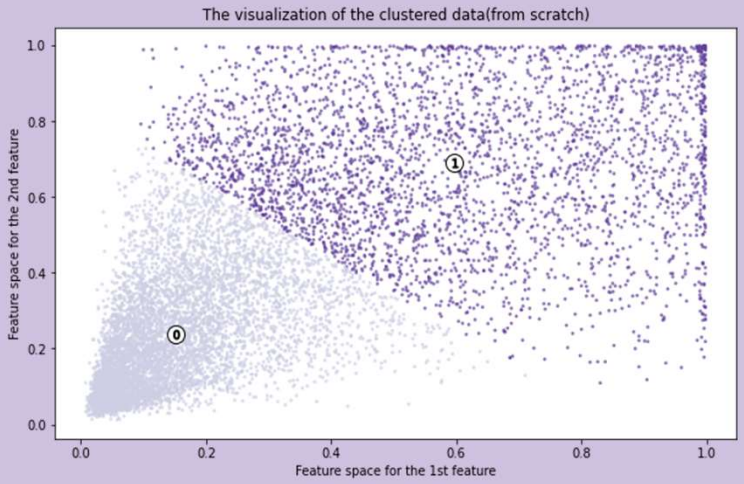


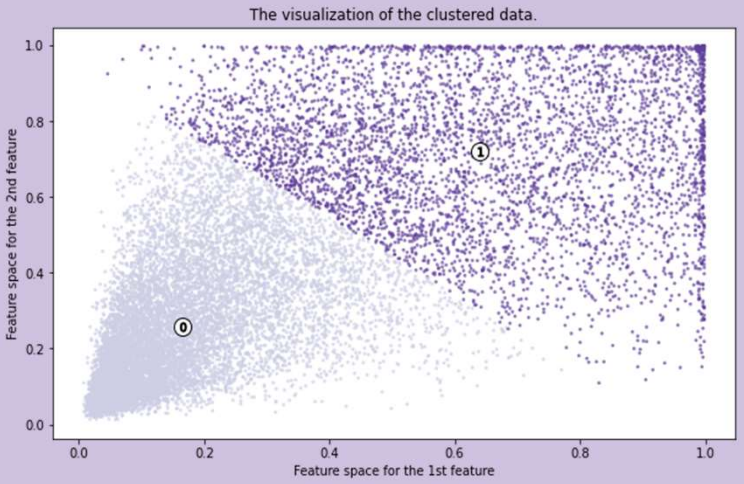
Optimal Number of Components turned out to be 2 according to Sillhoutte score


Silhouette analysis for Spectral Clustering clustering on sample data with 2 clusters
The silhouette plot for the various clusters.
The visualization of the clustered data.
Cluster label
The silhouette coefficient values
Feature space for the 2nd feature
Feature space for the 1st feature


Silhouette analysis for Spectral Clustering clustering on sample data with 3 clusters
The silhouette plot for the various clusters.
The visualization of the clustered data.
Cluster label
The silhouette coefficient values
Feature space for the 2nd feature
Feature space for the 1st feature

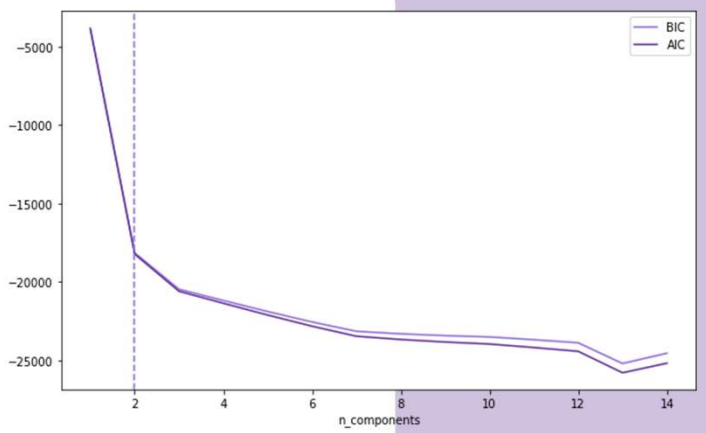
Spectral Clustering

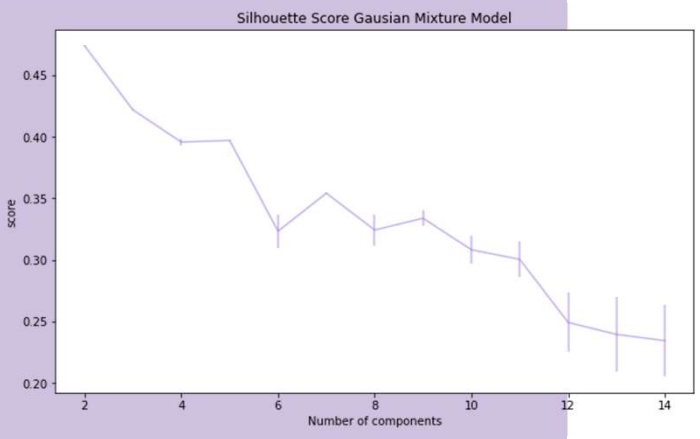




GMM

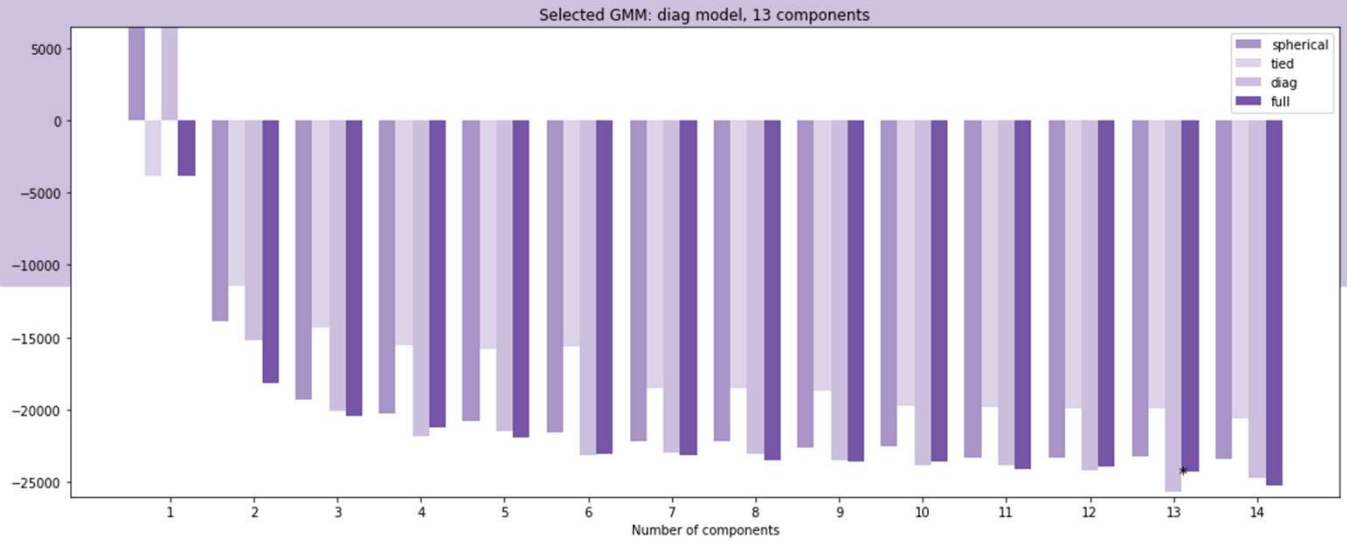
BIC/AIC score : 2 or 12 !

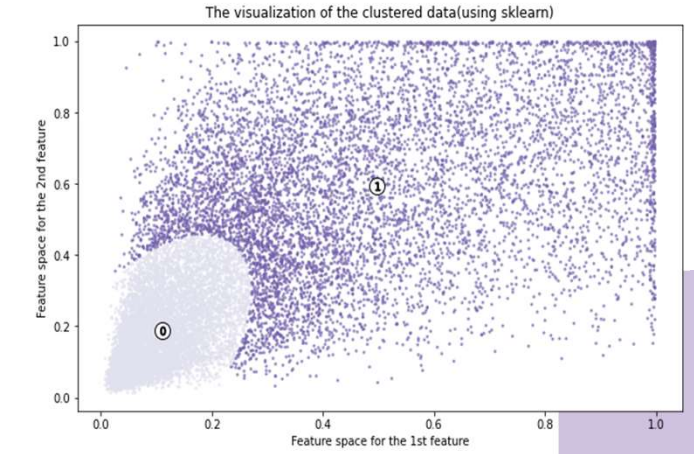


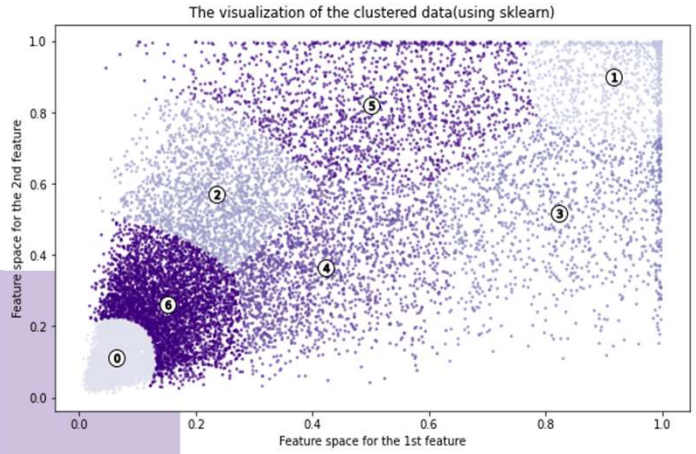


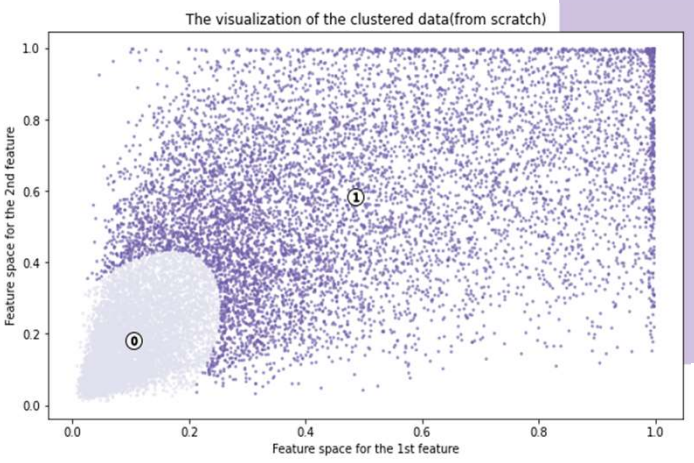
Sillhoutte Score : 2, 3, 5, 7 clusters all give out the very good scores and lease amount of stochasticity

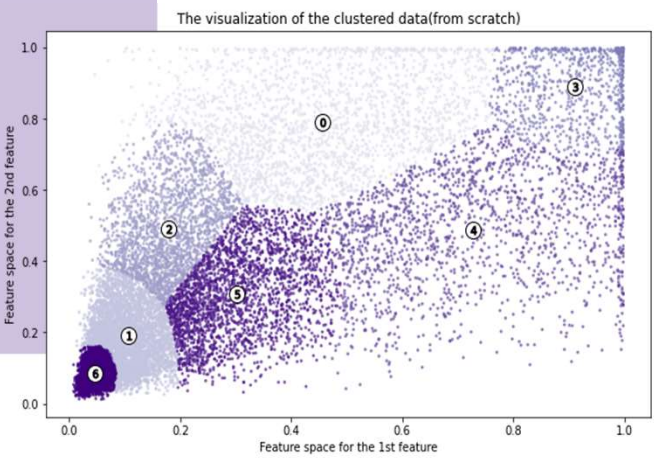
GMM – BIC score – Different covariance types




The visualization of the clustered data(using sklearn)
Feature space for the 2nd feature
Feature space for the 1st feature
0
1


The visualization of the clustered data(using sklearn)
Feature space for the 2nd feature
Feature space for the 1st feature
0
1
2
3
4
5
6

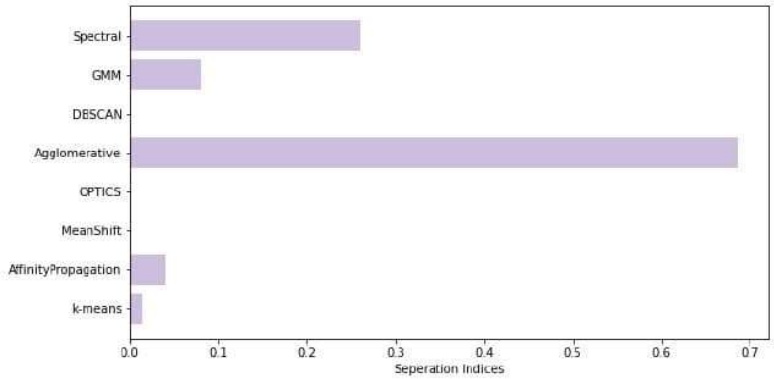

The visualization of the clustered data(from scratch)
Feature space for the 2nd feature
Feature space for the 1st feature
0
1

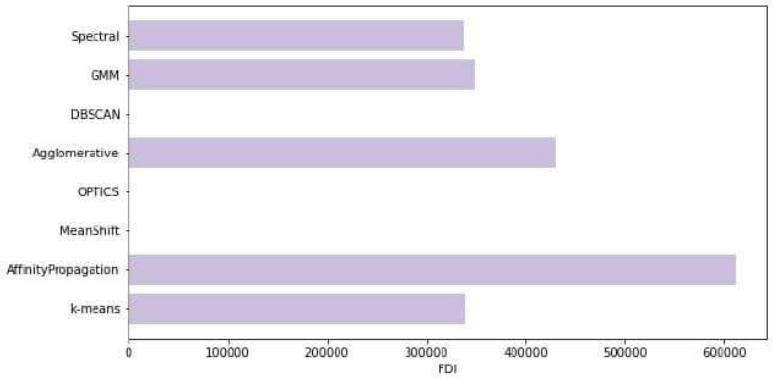

The visualization of the clustered data(from scratch)
Feature space for the 2nd feature
Feature space for the 1st feature
0
1
2
3
4
5
6

Comparison and Evaluation - Problems
(sk-learn, Toy DataSet)

OPTICS, DBSCAN fail in evaluation due to Auto-Noise Detection

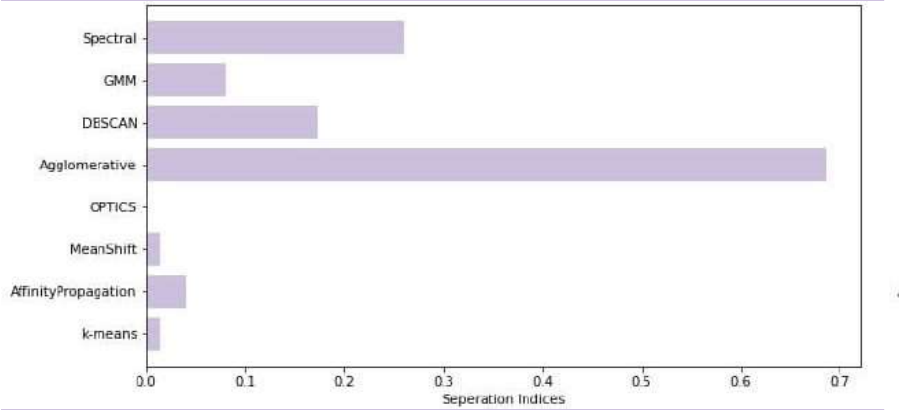
Data was not appropriate for mean-shift's estimated BW. Single Cluster



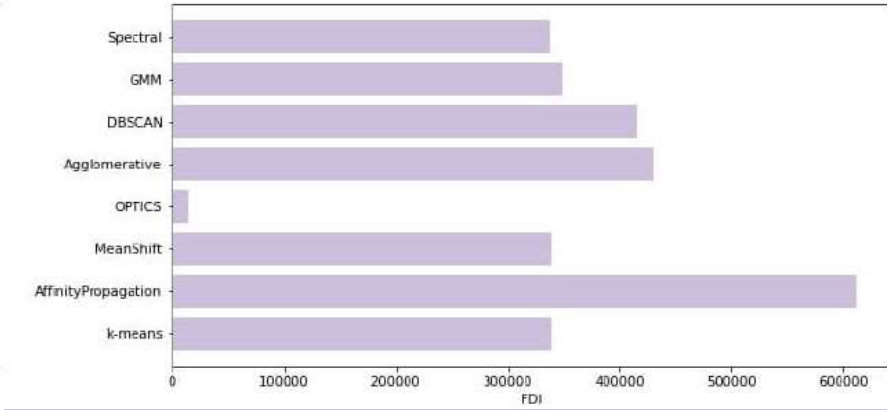


Comparison and Evaluation - Solutions
(sk-learn, Toy DataSet)

1. Clean Data by throwing away detected noisy points.

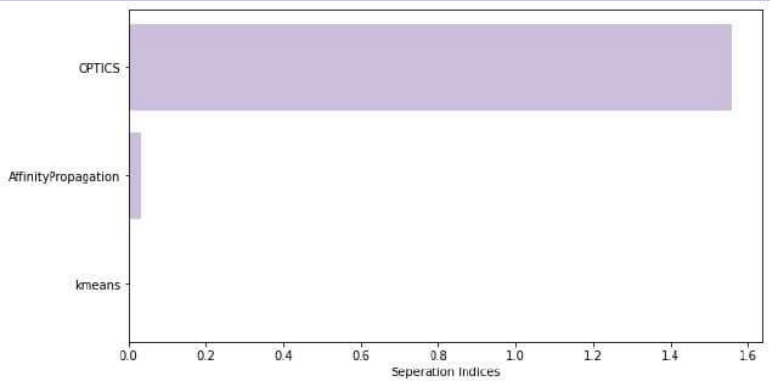


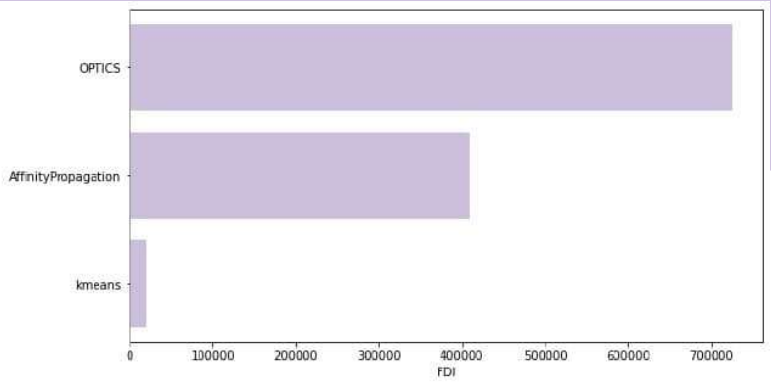
2. Set Mean-Shift BW Manually



Comparison and Evaluation
(sk-learn, bboxes DataSet)

Due to high time and computation Complexity, 3 Algorithms that had fast convergence were evaluated on this dataset

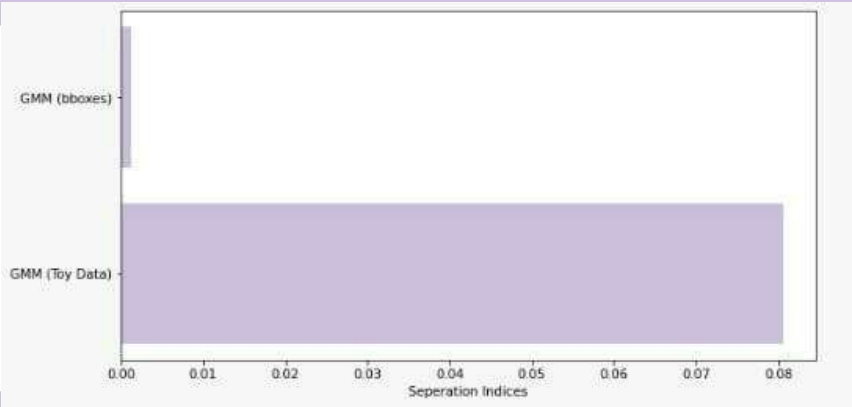


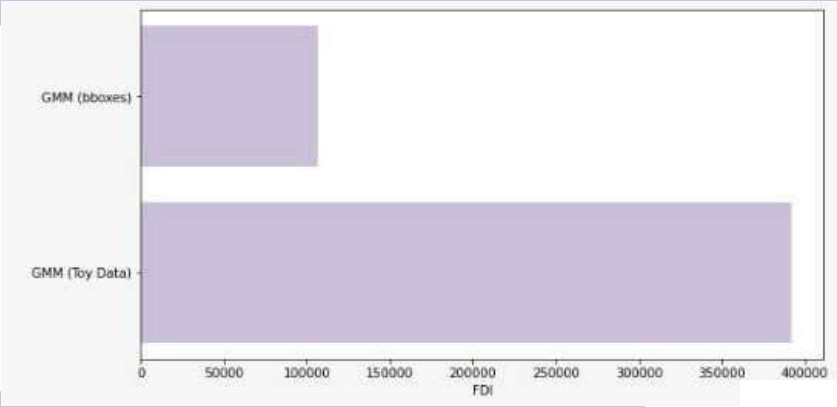


GMM – BIC score – Different covariance types

Scratch Algorithms we implemented usually performed better on Toy Dataset

One of the reasons can be the high dimensional structure and data complexity of ‘bboxes’





THANKS!