Clustering report

Machine learning and Data mining II

Nguyễn Mạnh Hưng - 22BI13183

Nguyễn Trọng Minh - 22BI13304

**TABLE OF CONTENT**

1. **K-means clustering**
2. Experiment protocol
3. Centroid initialization
4. Analyze the result
5. Clustering quality
6. **Subspace clustering**
7. Visualize the dataset
8. Apply K-means
9. Analyze result
10. Vary subspaces
11. **K-means clustering**
12. **Experiment protocol**

In this lab work, we chose 2 data sets, one is data about the energy efficiency of the heating load and cooling load requirements of buildings, and the other is CPUs performance data of different machines.



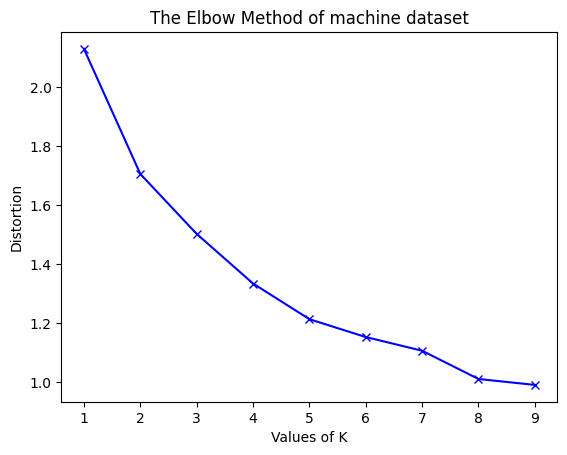
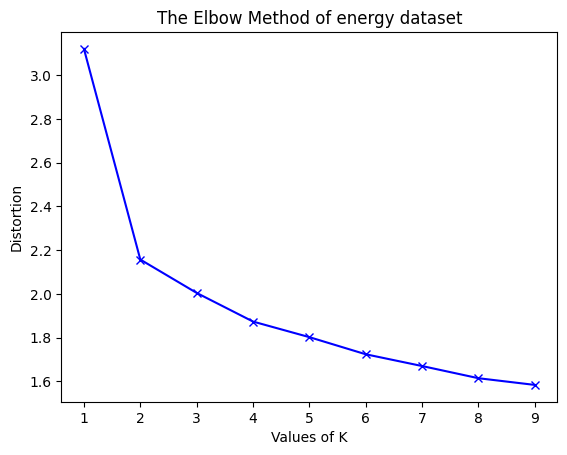
* 1. Data about CPUs performance



* 1. Data about energy efficiency

After a quick look into the data set, we can start applying K-means clustering. Because K-means clustering measures the Euclidian distance between a data point and centroids, we need to standardize data into an appropriate scale. We simply do this by divide each category by their standard deviation. And for later visualization using PCA, I also want to center the data to the origin.

The next step is to determine how many clusters we want to find from this data set. Normally, we can set the number of clusters to any positive integer, but there is a method to find which number give the most efficient. This method is called “Elbow method”. In this method, we continuously iterate through every k from 1 to n, which is set to be 10 in this article. For each iteration, we calculate the within-cluster sum of squares, or cluster’s variance. Then, we plot a graph and surprisingly, this graph looks like an elbow.



As we can see from the graph, we need to pick k when the line starts to look like a straight line, in both cases, I would choose number 4.

Now the data is ready to be separated into 4 clusters with K-means algorithm. It starts by initialize the first 4 centroids, and assigns every data point into its nearest centroid. After that, it stops and calculate a new centroid which is the mean value of every data point inside a cluster. And again, it comes back to step 2, assign data point into a cluster. This loop occurs for several times to find the best cluster centroid with the least data variance.

At the end, we get a clustered data set with labels for each data point.

1. **Centroid initialization**

In the above session, we talked about an overall protocol of K-means clustering, I want to go deeper into the centroid initialization in this part of the report.

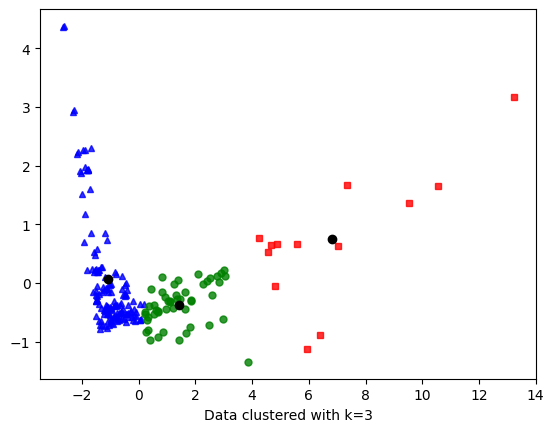
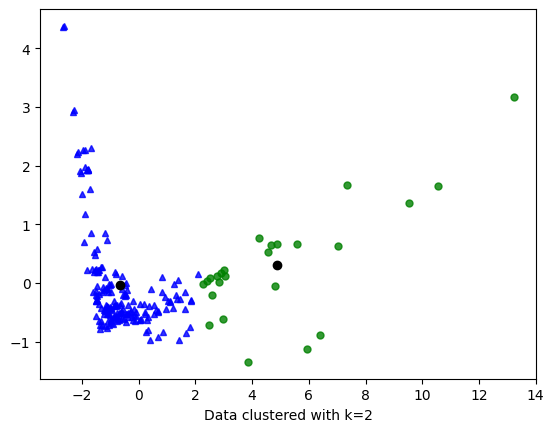
Normally, the first centroids of the algorithm are generated randomly inside data set. But in this experiment, I applied a method to determine those centroids, called “greedy k-means++” published by Arthur and Vassilvitskii in 2007. The greedy k-means++ algorithm samples k initial centers by adaptive sampling, where in each step, 𝑙 possible centroids are chosen, and then among these 𝑙 centers, the one set that decreases the k-mean cost the most is chosen to be initial centroids.

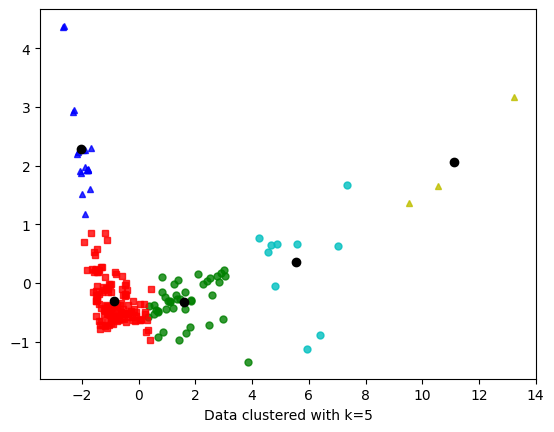
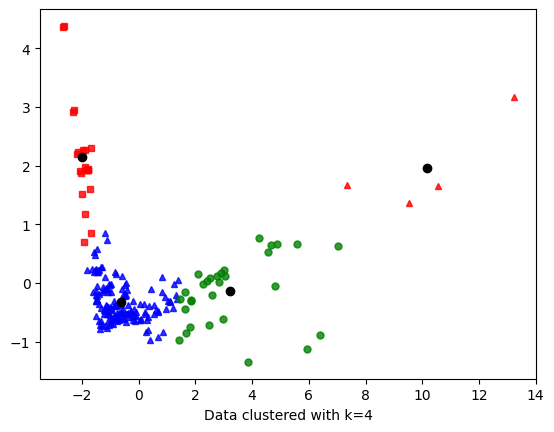
This technique is believed to speed up convergence that provides an

O(log k)-approximation in expectation.

1. **Analyze the result**

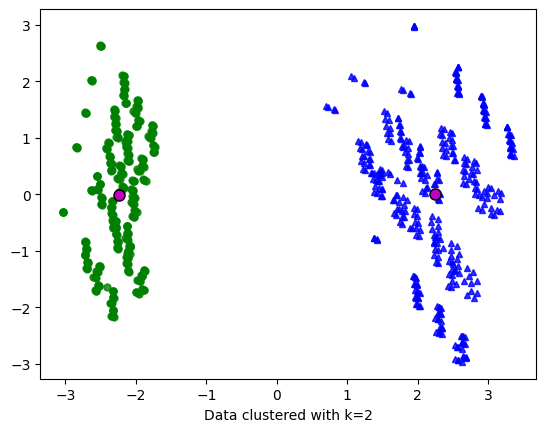
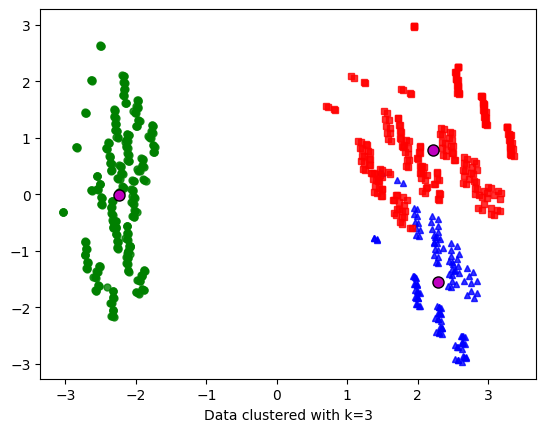
For the CPUs performance dataset, here are the results with different values of k:

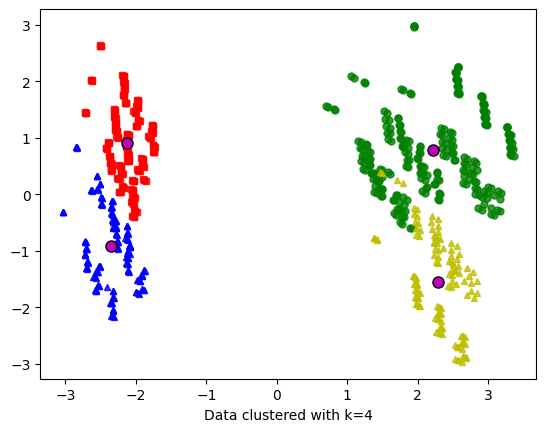
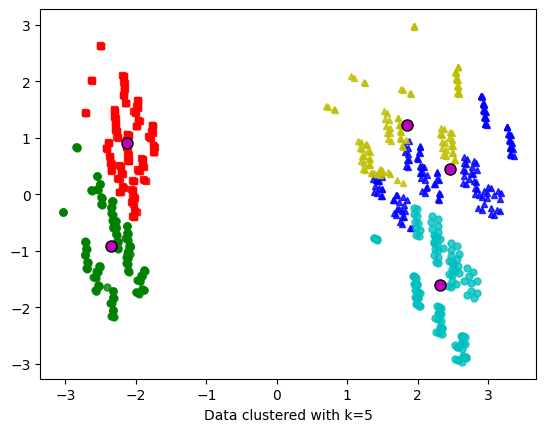




As we can see in case k = 4 and 5, there is a cluster with tiny amount of data points, 4 and 3. These 2 clusters should be merged into bigger cluster for better performance.

For the energy dataset here are the results with different values of k:

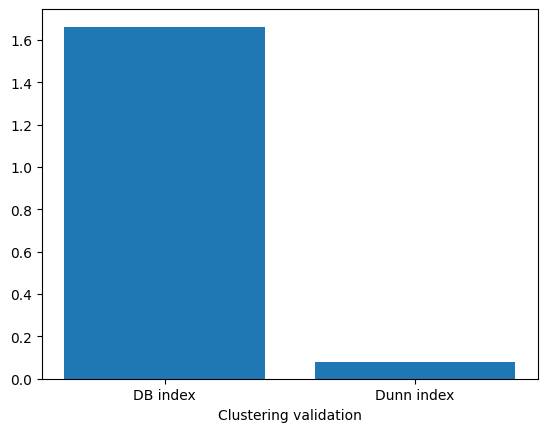
 

The plotted data show that k=2 is the best number of clusters for this dataset. We can also see some data points might be clustered into wrong cluster, but this phenomenal is just data error due to PCA.

1. **Clustering quality**

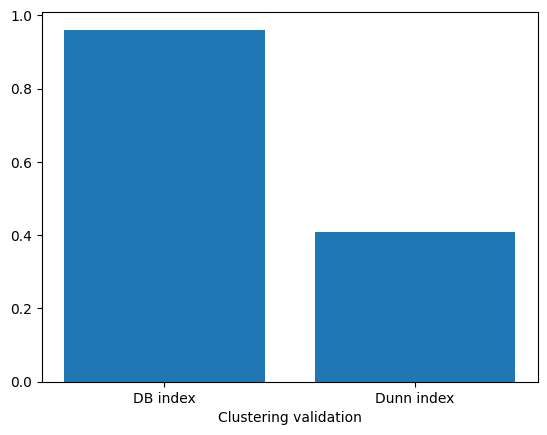
To validate the quality of the cluster, we apply 2 internal validation methods, Davies-Bouldin index and Dunn index.

Here is the validation result for energy dataset with k=4:



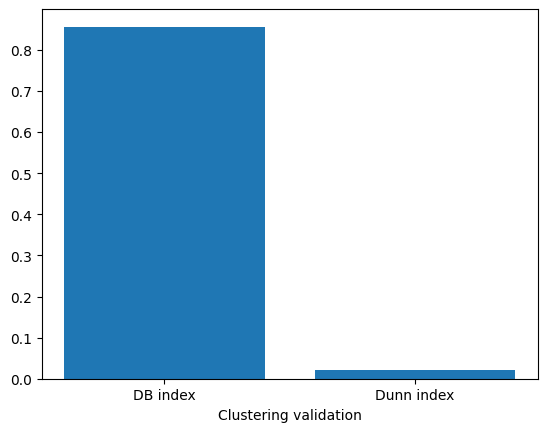
We can see that DB index is quite high, that means the data is not well-separated. This is true according to the plot we have above. Take a look at Dunn index, Dunn index is opposite from DB index, since it need to be large to indicate a good clustering. And the Dunn index in this part is also very small, show us the same result as DB index.

Here is the same energy dataset but with k=2 validation result:



It’s clearly that the DB index and Dunn index got closer with k=2. That means the clusters are now well-separated.

Now we apply the same validation techniques on machine dataset, and got this result with k=4:

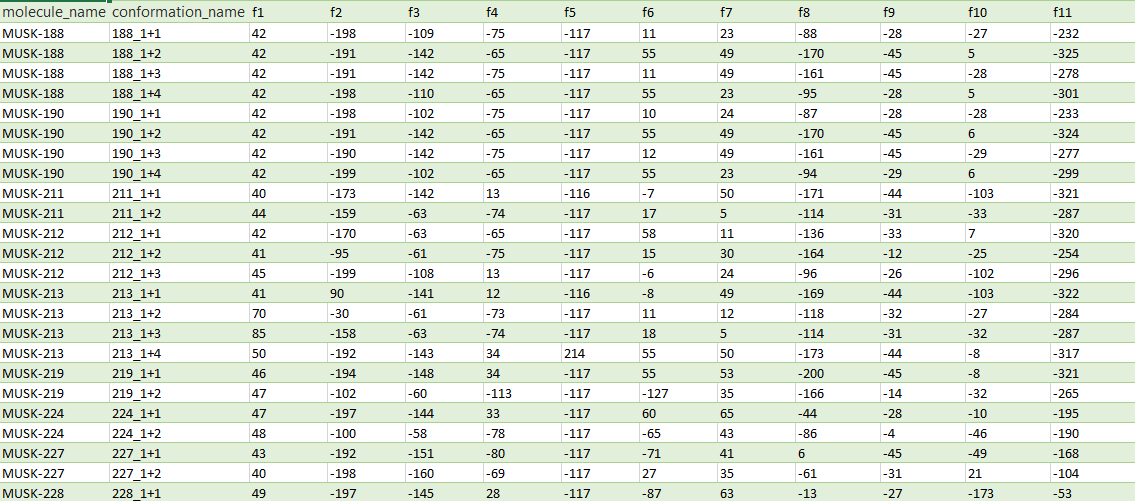


The DB index is smaller than 1 but the Dunn index is still very small, this means our clustering attempt failed to create well-separated clusters.

1. **PCA – Principal Component Analysis**
2. Visualize the dataset
3. Apply K-means
4. Analyze result
5. Vary subspaces

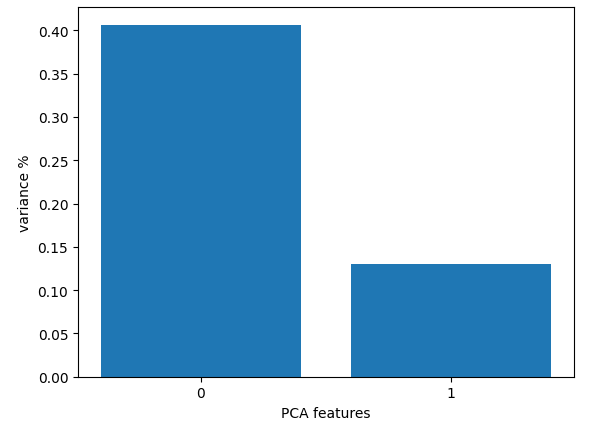
**1.visualize the dataset**

For the dataset that has more than 100 features, we chose a dataset about the musk and non-musk molecules.



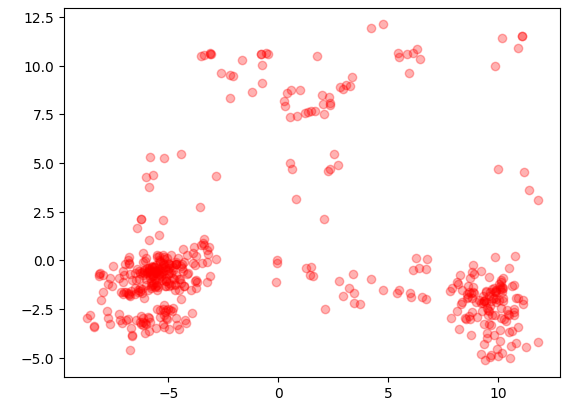
* 1. data about musk and non-musk molecules.

Looking at the dataset, first we need to drop 2 first columns since they are category variable. And then after that, we scale the data so that the variable with large number doesn’t account for majority of the ratio between category by dividing all the variables with their standard deviation. The next step is to center the data by subtract them with the mean value. Following that is to compute the eigenvalues and eigenvectors of the covariance matrix, from that we can plot the scree graph that represents the variance explained by eigenvalue.



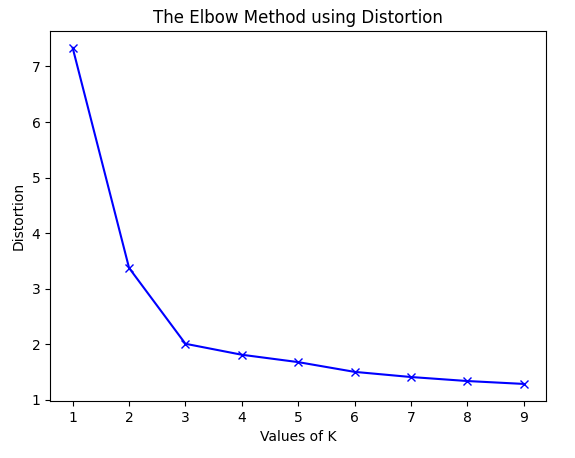
* 1. variance explained by eigenvalue

Look at this graph we can only choose these two components to form the new space. Finally, we got the graph that represent the data after reducing dimension.

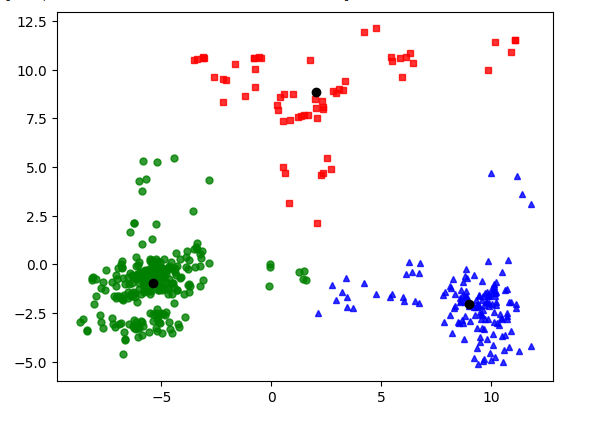


**2.Apply K-means into the dataset.**

Using the same step as the two previous datasets, first we will use the “elbow method” to find the suitable number of centroids.



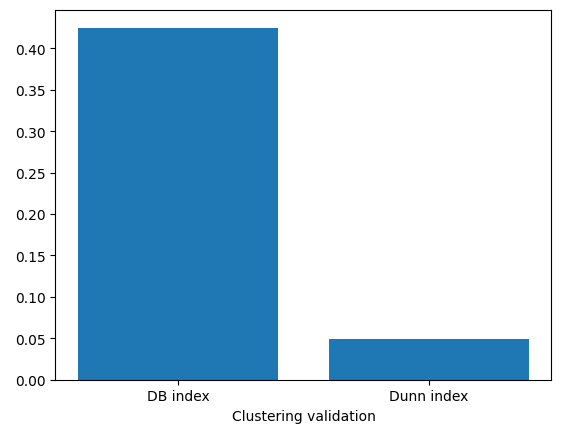
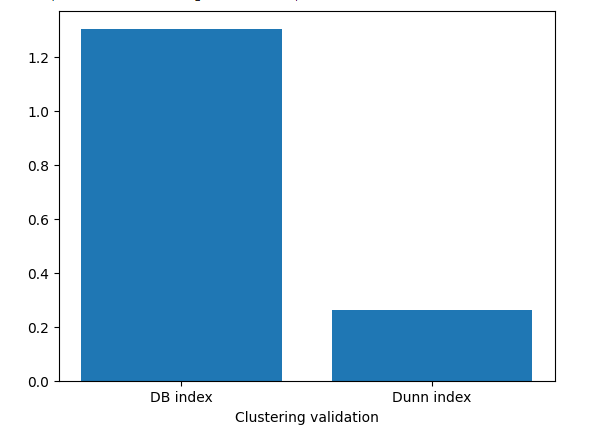
As we see from the graph, we can choose 3 is the number of centroids. After that we assign the data point to the suitable cluster and we got the final result.



4.3 dataset clustering after PCA

**3.Compare the results**

To compare the performance between two clusters before and after PCA. We can compare the clustering quality using 2 internal validation methods, Davies-Bouldin index and Dunn index

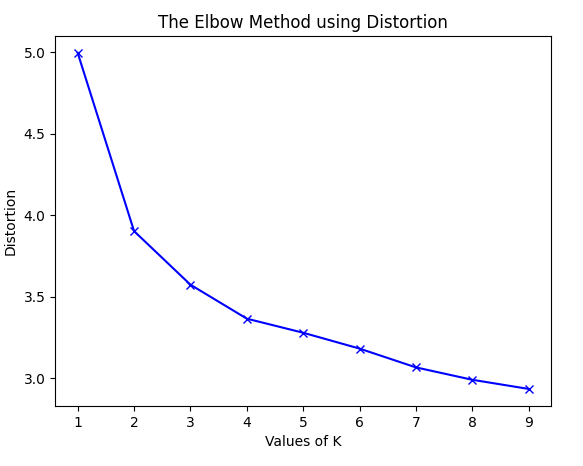


4.4 cluster quality before and after PCA

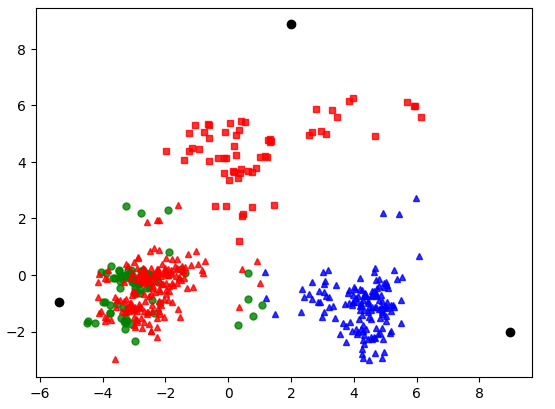
As we can see from the two graphs, the dataset before PCA got a result of high DB index and low Dunn index, which meant that the it is not well clustered. For the dataset after PCA, it received the result of low DB index but still a little low Dunn index. But overall, the dataset after PCA got a better cluster performance.

**4.subspace**

In this part, we will turn the origin dataset into subspace with 40 features by using “numpy.random.choice” to randomly choose the features. Using the same step as the previous part. We can compute and plot the “elbow method”.

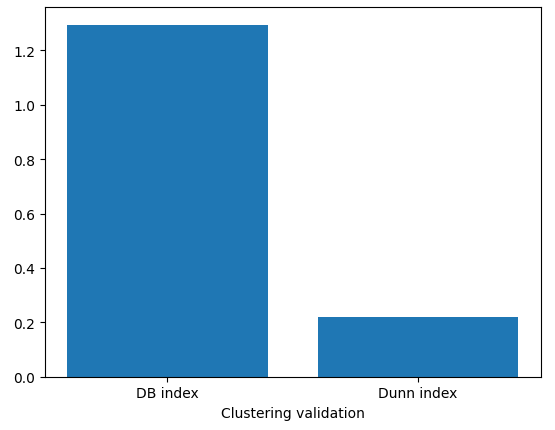


From the plot, we can choose k=4 is the number of clusters. After we assign all the data into clusters, we can visual the data in 2D dimension.



4.5 the cluster visualization in 2D

Finally, we find the cluster quality by using “Davies-Bouldin” index and “Dunn” index and have a graph as below:



4.6 Db index and Dunn index of the dataset after clustering

Compare to the origin dataset’s cluster quality. We can see that there is no difference in the clustering quality of the subspace and the origin dataset.