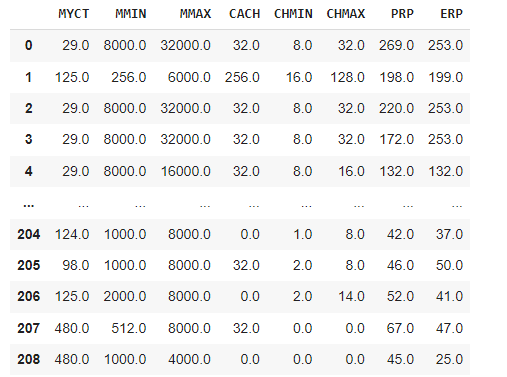
I.

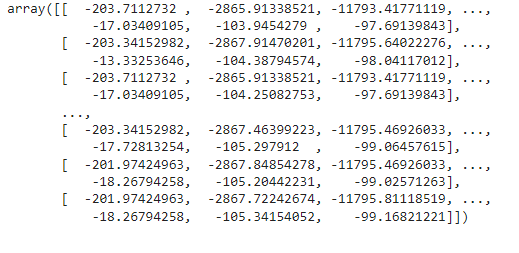
\*Machine data

1. Explain the experimental protocol.

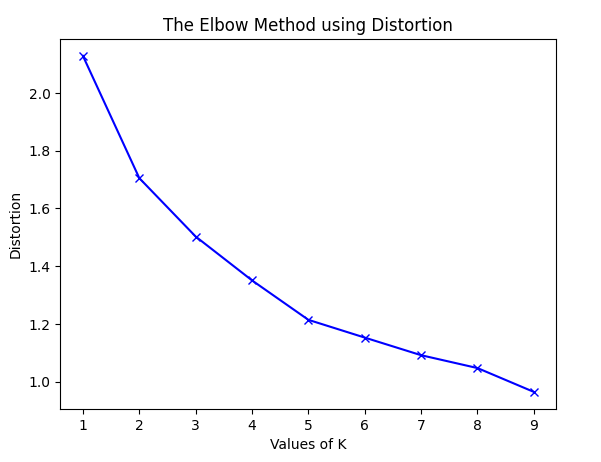
With the machine dataset, first we drop the category columns which are “Vendor Name” and “Model Name” since the category variable cannot distribute in the clustering. After dropping the two columns, we have the dataset below (which has 209 rows and 8 columns):



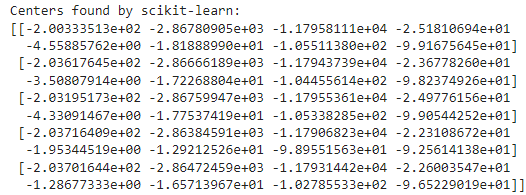
after that we will do the preprocessing the data to find the array with adjust data:



The array that has been adjusted

In order to find the best number of centroids for the k-means clustering, we will run the “KMeans” module in “sklearn.cluster” library with the number of clusters run from 1 to a specific number. After that we will plot the elbow method to find the suitable number of clusters as follow: 

According to the plot, we can choose 5 is the number of the clusters for this algorithm, which are:



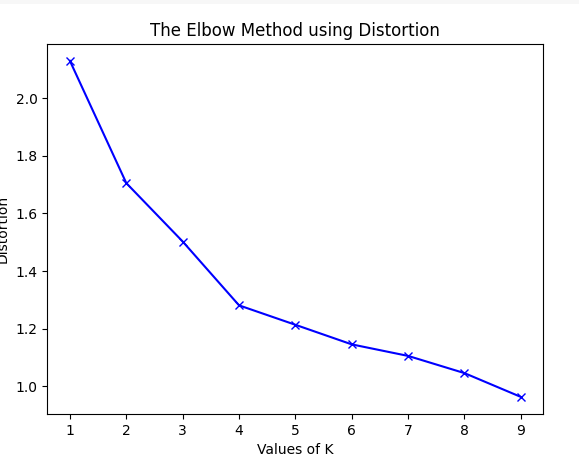
5 centroids of the k-means clustering alg.

2. Centroid initialization.

Firstly, we will choose randomly n prototypes with n is the number of centroids we want to generate. After that we calculate the distance between all the objects to the centroids, and then assign the objects to the closest centroid. The next step is to calculate the new centroid which is the average value of all the objects that belong to 1 cluster. And we repeat the assign step for all the objects until the centroids are convergent.

3.analyze and compare the result.

We still use the elbow method:



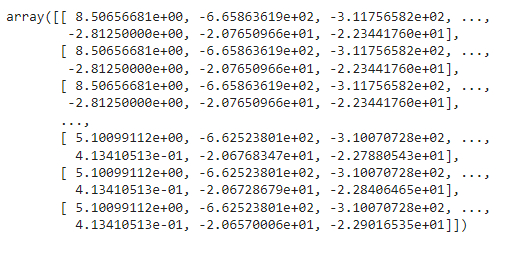
As we can we with k = 1, 2 or 3. The inertia drop significantly means that they are not the good number of clusters to choose. With the k from 4 to 9 we can see that the sum of squared distance reduces slightly after k=4 which means that it may not provide significant improvement in clustering quality as adding more clusters beyond 4 and 4 is a good number of clusters that we need to choose.

4. Clustering quality.

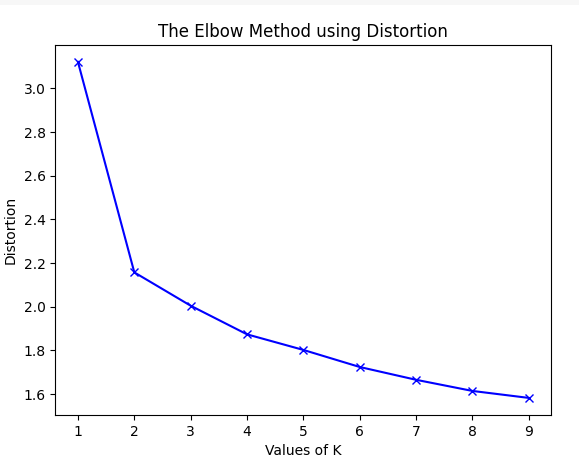
\*Energy\_efficiency dataset

1.run with k-means

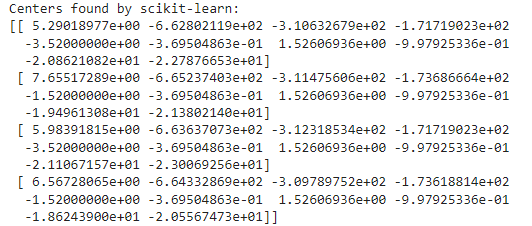
The same as machine dataset but this dataset does not have category data so we only have to adjust the data:



After that we will find the number of clusters for the dataset that is suitable by using the elbow method:



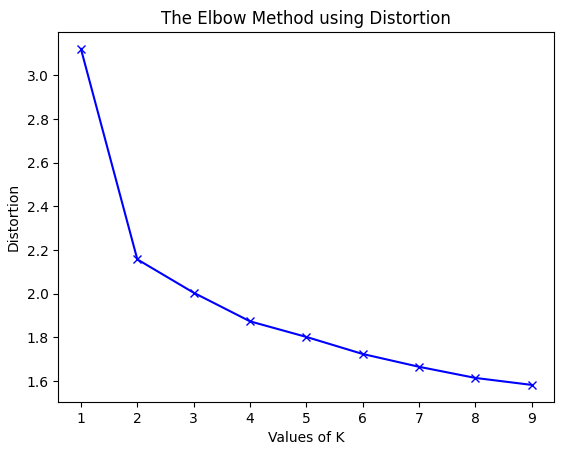
Looking at the graph we can see that the graph starts to decline rapidly at k=4 so 4 will be the number of clusters that we will choose for this dataset and these are the centroids for this dataset.



2.explain (the same as machine data)

3.analyze and compare the results with different k

The same as the machine dataset, we will use the elbow method for the analyze:



Firstly, we can see that with k=1, the line for the inertia decreases significantly from over 3.0 to under 2.2 which is about 1.3. The graph continues to decrease with k=2 and 3 and reduce for about 0.1 each. And starts from k=4 it only decreases slightly that if we increase the k value more than 4, the result may stay unchanged.

4.evaluate cluster quality

II. Subspace clustering

1. Data with 100+ features.

We chose a dataset which has 169 features and 477 objects

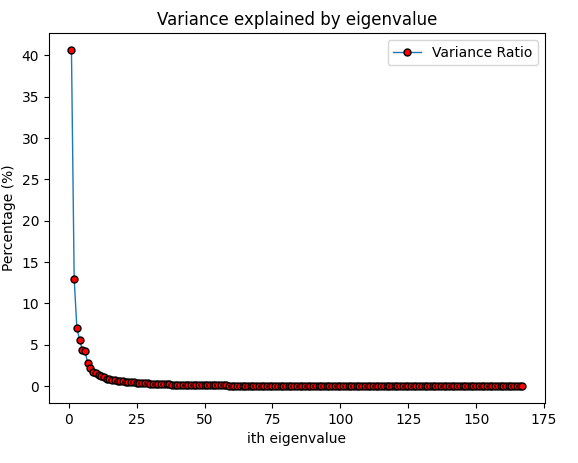
1. PCA

For this dataset, we will load the csv file as pandas dataframe

Then drop 2 first columns since they are category variables

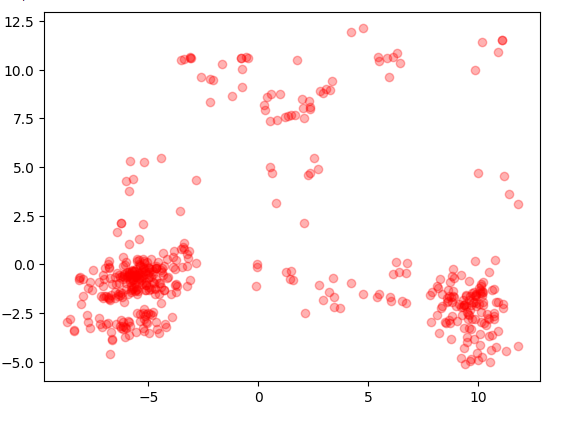
After that we will process the data and then calculate the eigenvalues and eigenvectors of the covariance matrix.

From that we can plot a scree graph that represents the variance explained by eigenvalue.



we will choose the first 2 components and use these to form a new space.

The last step is to plot the objects:

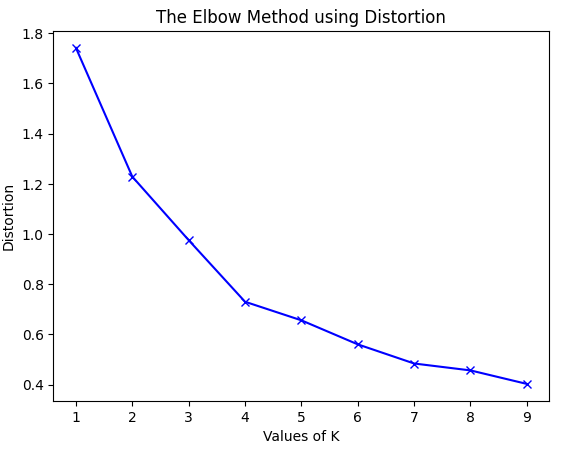


The dataset after PCA

3 kmeans after PCA

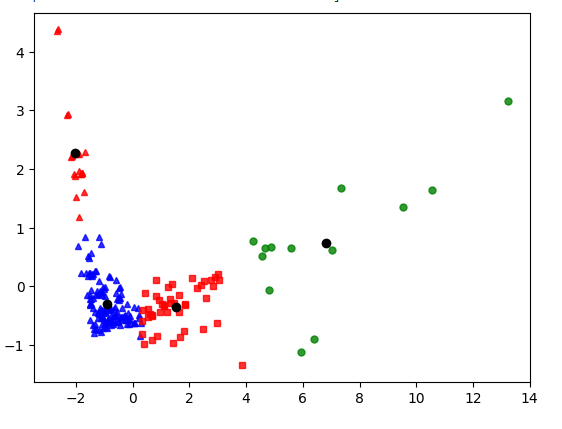
* machine dataset:

we apply PCA into the dataset with the same way as we did to the previous dataset. After reduce the dimension of the dataset, we use elbow method to find the number of clusters for the dataset:



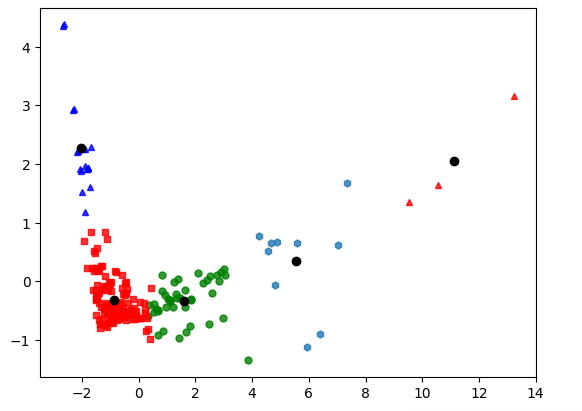
Elbow method after PCA

We can see that the number of centroids reduce from 5 to 4 according to the graph, and then we plot it out to see the objects of the dataset:



Machine dataset after clustering.

Compare to before PCA:

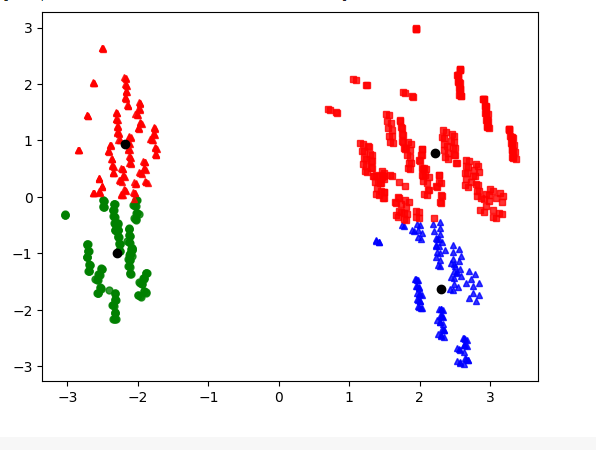


Machine dataset before PCA

We can see that the performance of the cluster before PCA perform much better than after PCA

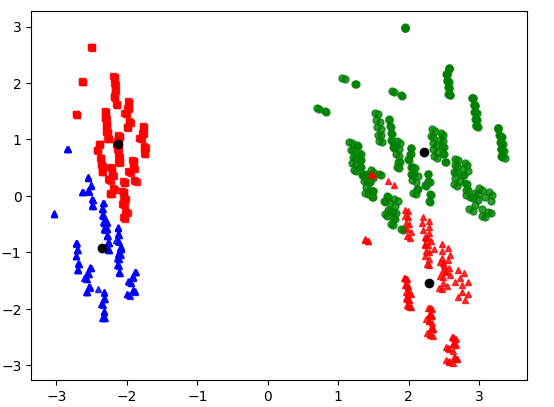
* energy dataset

doing the same as machine dataset we have the plot for the clustering of the dataset after the PCA:



Dataset after PCA

Compare to before PCA:

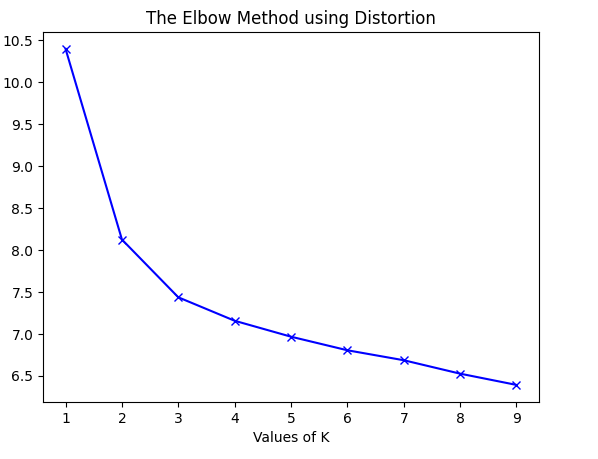
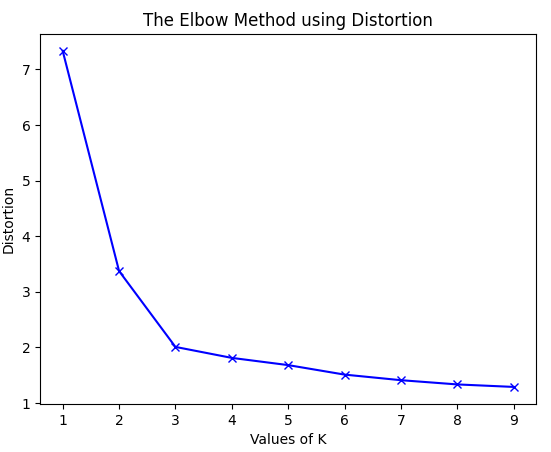


Dataset before PCA

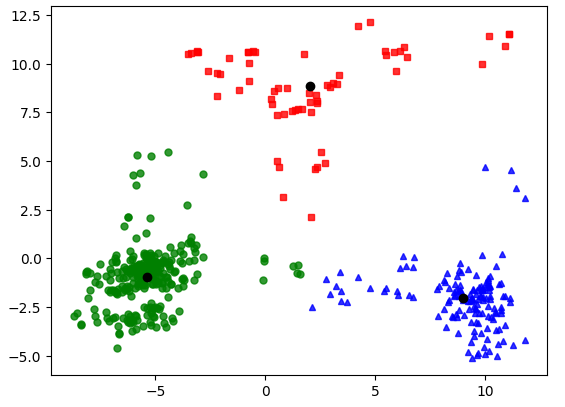
We can see a little difference between two plots but it still tells us that the clustering after the PCA perform better.

\*data100features:

With this dataset, after elbow method between before PCA and after PCA, we both have the same number of clusters which is 3:

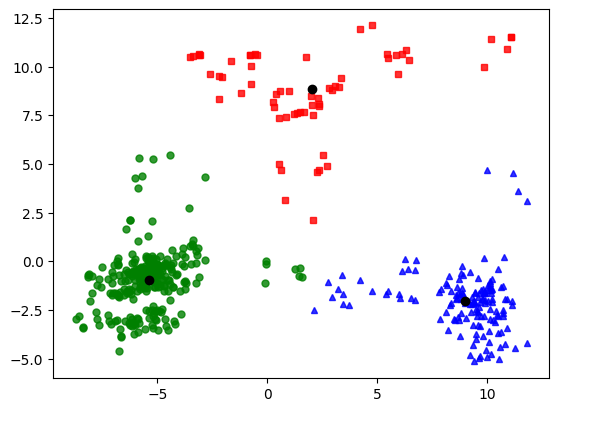
 

Dataset before PCA dataset after PCA

We can see the plot for the clustering before PCA:  


Clustering before PCA

And we can see the plot for the clustering after PCA:



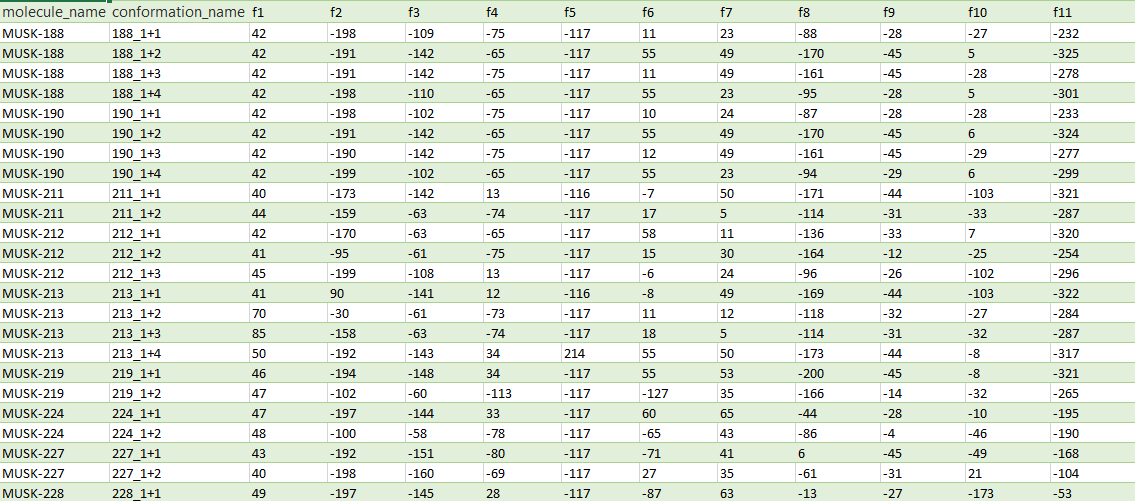
Clustering after PCA

We don’t see any difference between two plots so we can conclude that they perform the same.

4.sub-space

1.dataset

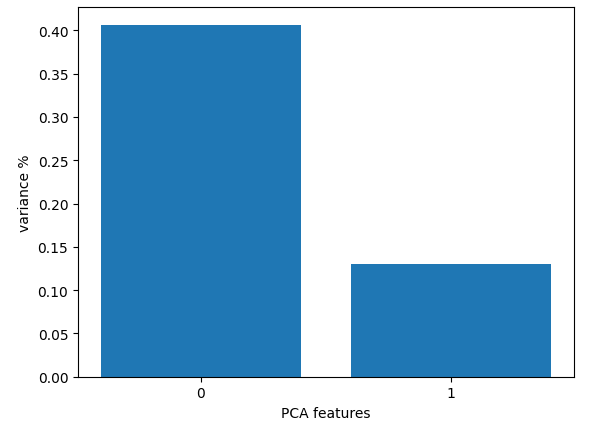
For the dataset that has more than 100 features, we chose a dataset about the musk and non-musk molecules.



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

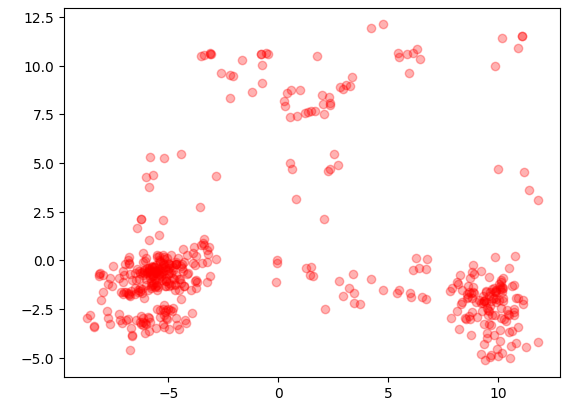
2.apply PCA

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.



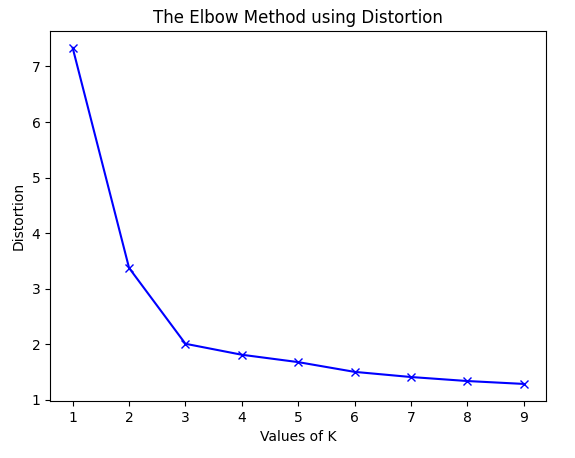
* 1. variance explained by eigenvalue

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.

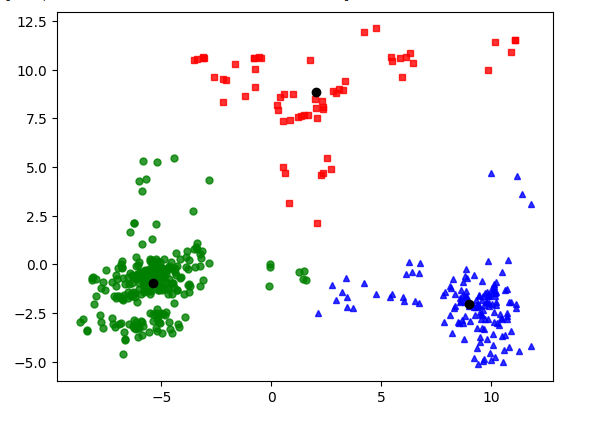


3.Apply K-means into the dataset.

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

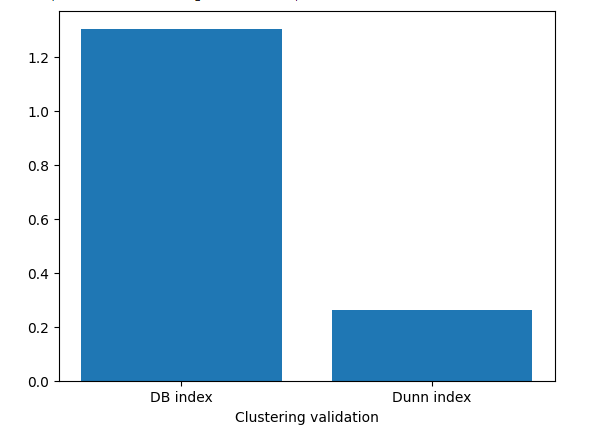
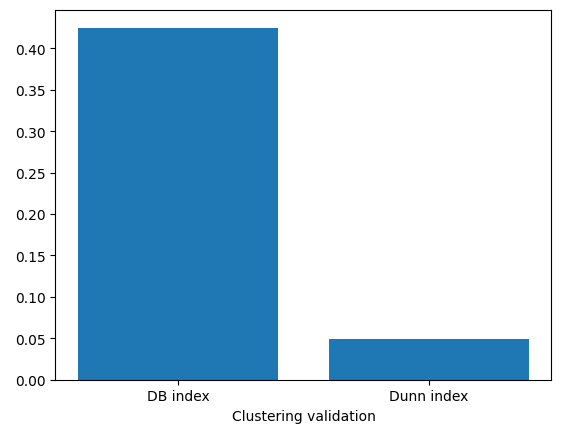


As we see from the graph, we can choose 3 is the number of centroids. After that we assign the data point to the suitable cluster and we got the final result.



4.3 dataset clustering after PCA

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.

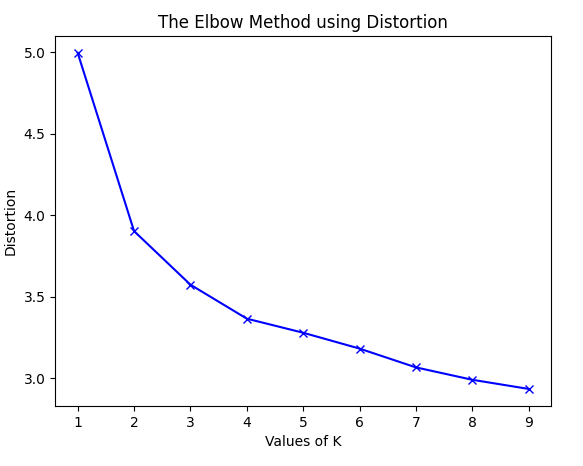
 

4.4 cluster quality before and after PCA

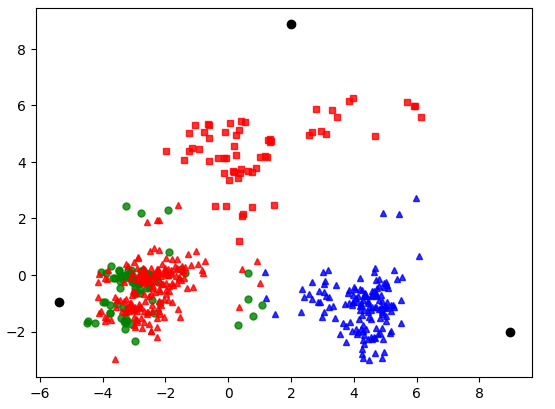
As we can see from the two graphs, the dataset before PCA got a result of high DB index and low Dunn index, which meant that the it is not well clustered. For the dataset after PCA, it received the result of low DB index but still a little low Dunn index. But overall, the dataset after PCA got a better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

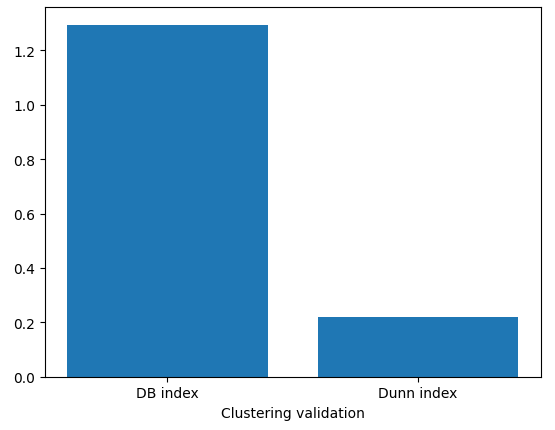


From the plot, we can choose k=4 is the number of clusters. After we assign all the data into clusters, we can visual the data in 2D dimension.



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 subpace and the origin dataset.