Assessment Task 2: Problem solving task

SIT-720 Machine Learning

August 29, 2022

T2-2022

Thanh Nguyen

STUDENT ID 218583133

COURSE - Bachelor of IT Honours (S470)

1 Question 1-3: Dataset and Numner of Cluster

1.1 The curse of dimensionality

The curse of dimensionality dictatates that, as the number of dimensions grows, the difference between the minimum distance and the maximum distance approaches 0, the pattern in the dataset hence does not make sense anymore [2]. To illustrate the phenomena, we pick a random datapoint P within the dataset Q. We start with 2 features and increase to 10, for each loop, we compute the distance from P to each point in Q.

```
random_index = int(np.floor(np.random.random()*len(data)))
P = data.iloc[random_index].to_numpy()
Q = data.drop(axis=0, index=random_index).to_numpy()

deltas = []

for N in range(2, 11):

    p = P[:N]
    q = Q[:,:N]


    diffs = [np.linalg.norm(q-p) for q in q]
    mxd = max(diffs)
    mnd = min(diffs)
    delta = math.log10(mxd-mnd)/mnd
    deltas.append( delta )
```

As the number of dimensions increases, the difference between maximum distance and minimum distance became smaller and smaller exponentially as illustrated in Figure 1. Thus the given dataset suffers from the curse of dimensionality.

1.2 Grouping Travellers and Cluster Evaluation Techniques

To find the group of travellers from the given dataset, we can categorize users into groups based on different factors. In this case, we can group users based on their rating in different categories.

The KMeans clustering algorithm to identify such groups of travellers. For example, the scripts belows will partition the data into two groups.

```
kmeans_example = KMeans(n_clusters=2).fit(data.to_numpy())
```

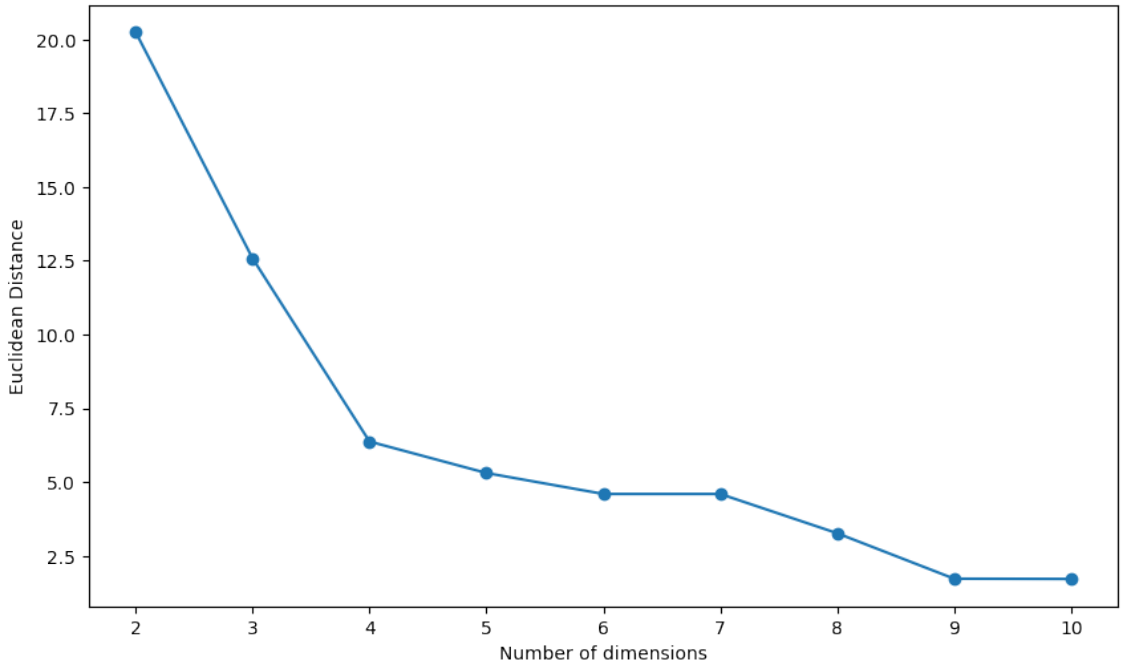

Figure 1: The Euclidean distance of a random value to the rest of the dataset for each feature.

```
2
3 print(f'There are two clusters, their centroids are: \n{
    kmeans_example.cluster_centers_[0]} \nand \n{kmeans_example.
    cluster_centers_[1]}')
4
5 >>> There are two clusters, their centroids are:
6 [0.90599476, 1.41947644, 1.82939791, 0.58280105, 1.18, 2.18356021,
    3.1871466,  2.80489529, 1.52609948, 2.60157068]
7 and
8 [0.88501672, 1.30989967, 0.49198997, 0.50036789, 0.78625418,
    1.62528428, 3.17697324, 2.8543311, 1.59712375, 2.92548495]
```

So far, we know that users can be classified into two groups using KMeans algorithm. However, there must be a number of cluster that is optimal for the dataset. We can to find the optimal number of traveller groups, by using various of cluster validation techniques. *The elbow method* [7] is a widely used for calculate the optimal number of clusters. For different values of k, we calculate the Within-Cluster-Sum of Squared (wss) Errors and chose the first k values for which the error values start to diminish.

We obtain the folloing graph (see Figure 2) for WSS values of our dataset for different k number of clusters. The plot took shape of an arm, and we can chose the optimal at the

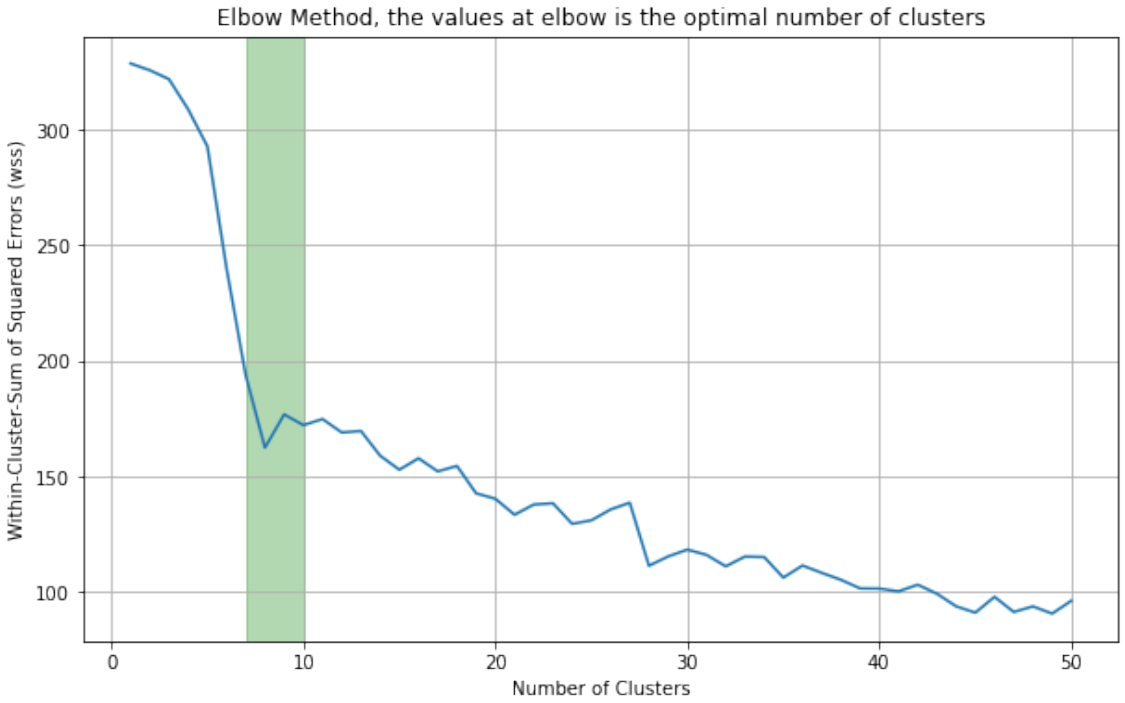

Figure 2: WSS values of the given dataset for different k number of clusters.

elbow. It is suggested that the optimal values k can be chosen within the range $[7, 10]$, as highlighted in the plot.

1.3 Silhouette Method with KMeans Algorithm

Compare to the elbow method, the *silhouette score* measures the similarity of a point to it own cluster compared to the other clusters [6]. The silhouette score is calculated with the function `silhouette_score()` from sklearn library. We aim to choose k values for the highest silhouette scores.

We obtain the silhouette scores of KMeans clustering algorithm for our dataset with different k number of clusters as Figure 3. The result shows that $k = 2$ clusteres would maximize the average silhouette score for the dataset (highlighted as red dot). Otherwise, we can chose the second peak, at $k = 8$ (highlighted as orange dot).

1.4 Silhouette method with Spectral Clustering

Spectral Clustering is another clustering algorithm. Compare to KMeans algorithm wich groups the data based on how close they are toward the cluster center (compactness ap-

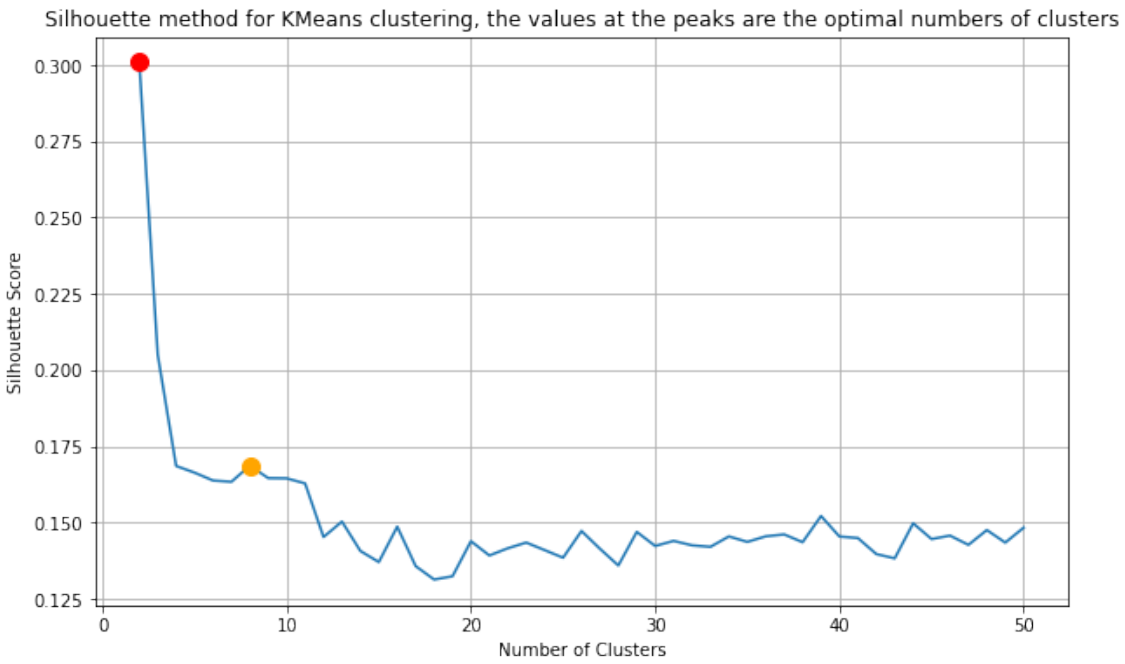

Figure 3: Silhouette score for clusters using KMeans algorithm

Index	Clustering Algorithm	Evaluation Technique	Optimal k
1	KMeans clustering	Elbow WSSE	[7, 10]
2	KMeans clustering	Silhouette Score	2, 8
3	Spectral clustering	Silhouette Score	2, 5

Table 1: Summarised result of Clustering algorithm and evalutaion techniques

proach), Spectral clustering algorithm use the connectivity approach, wich group the datapoint based on their ε distance to each other [3].

Applies the Spectral Clustering to calculate Silhouette score gives the results as Figure 4. The result shows that $k = 2$ clusteres would maximize the average silhouette score for the dataset (highlighted as red dot). Otherwise, we can chose the second peak, at $k = 5$ (highlighted as orange dot).

1.5 The Optimal Number of Cluster

We have obtained results as the optimal number of clusters for the dataset based on different clustering algorithms and cluster evaluation techniques. The findings are summarised as the Table 1.

The first method (elbow) has zoned the optimal values within the range from 7 to 10

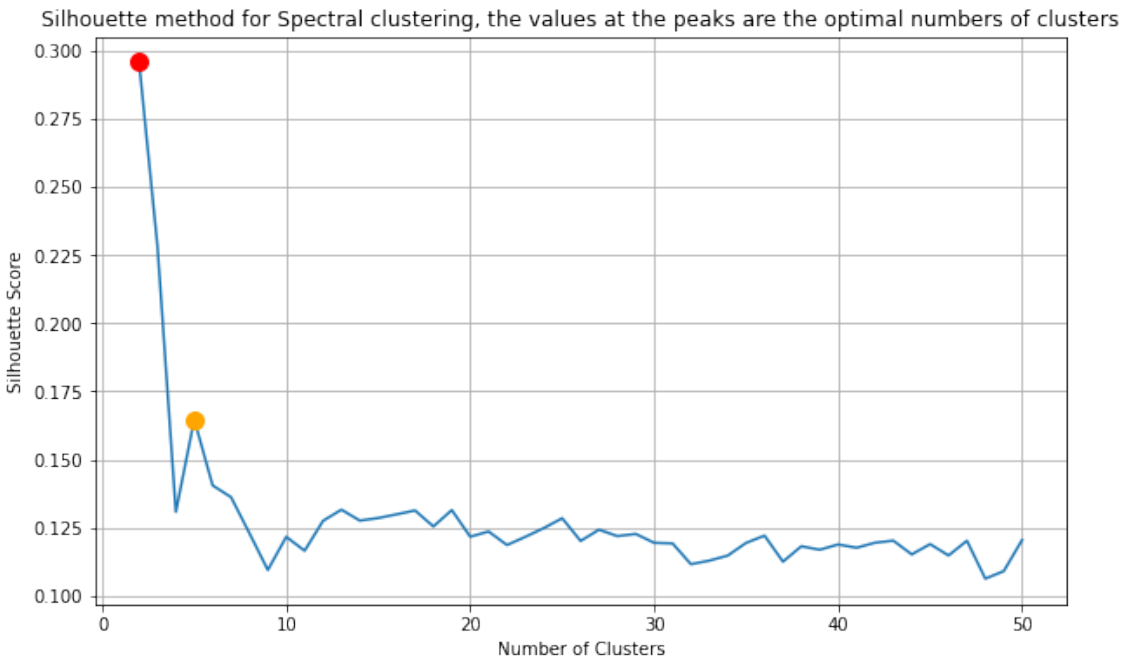

Figure 4: Silhouette score for clusters using Spectral Clustering algorithm

clusters. However, in the second and third method, we can see that the optimal values are 2, 5, and 8. The value $k = 8$ is within the optimal range [7, 10] from the first method and thus having a reasonable Within-Cluster-Sum of Squared Errors value (marked as blue dot). Onthe other hand, the values $k = 2$ and $k = 5$ will result in higher error values (marked as red dot). In conclusion, the optimal number of cluster is $k = 8$ (see Figure 5)

2 Question 4-6: Dimensionality Reduction

2.1 Features Relationship

The *correlation matrix* is a matrix represent the correlation coeficient of different features. The matrix shows the correlation between all pairs of features within the dataset. We use the correlation matrix to summarise the relationship of the given dataset and to visualise patterns within (see Figure 6).

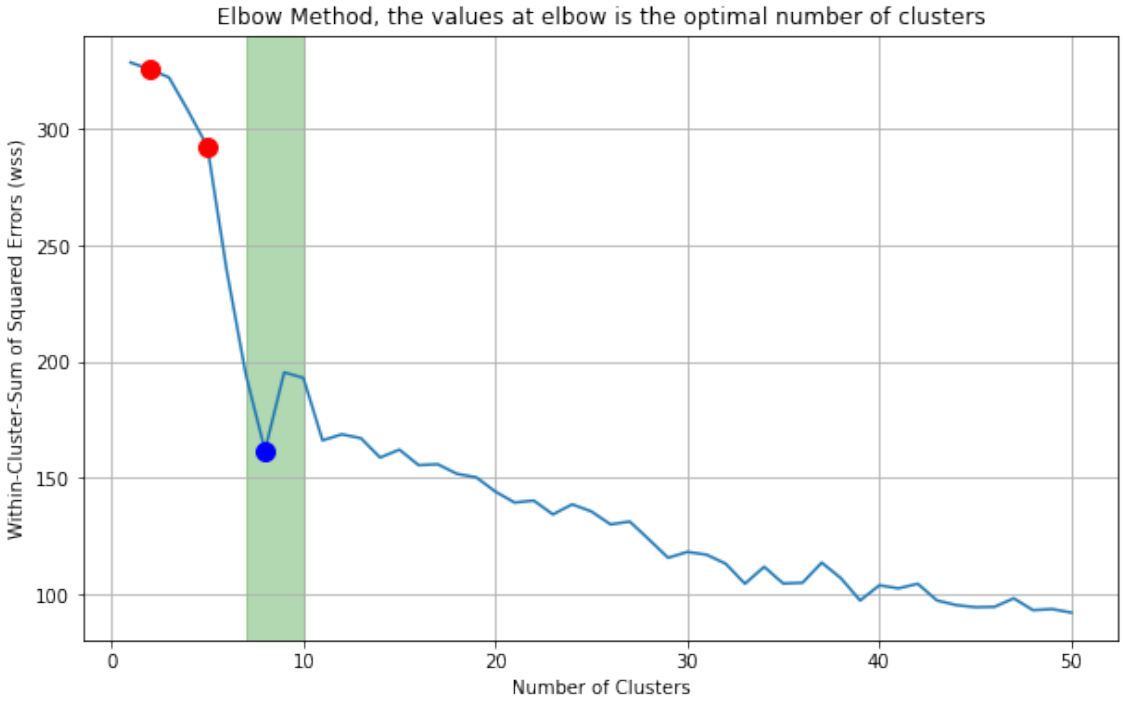

Figure 5: Optimal number of Cluster

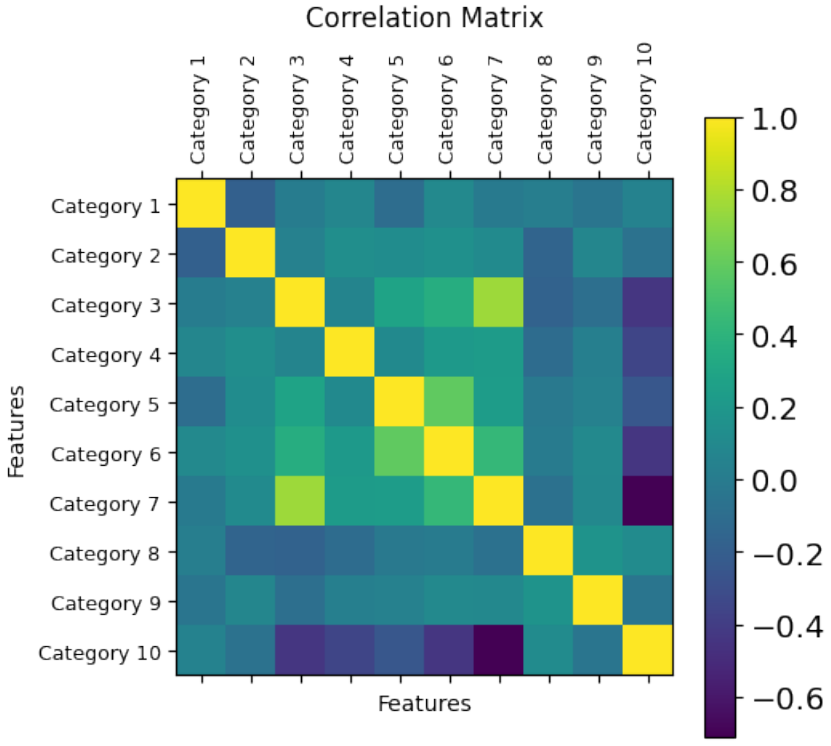

Figure 6: Correlation of each features visualised as a heatmap

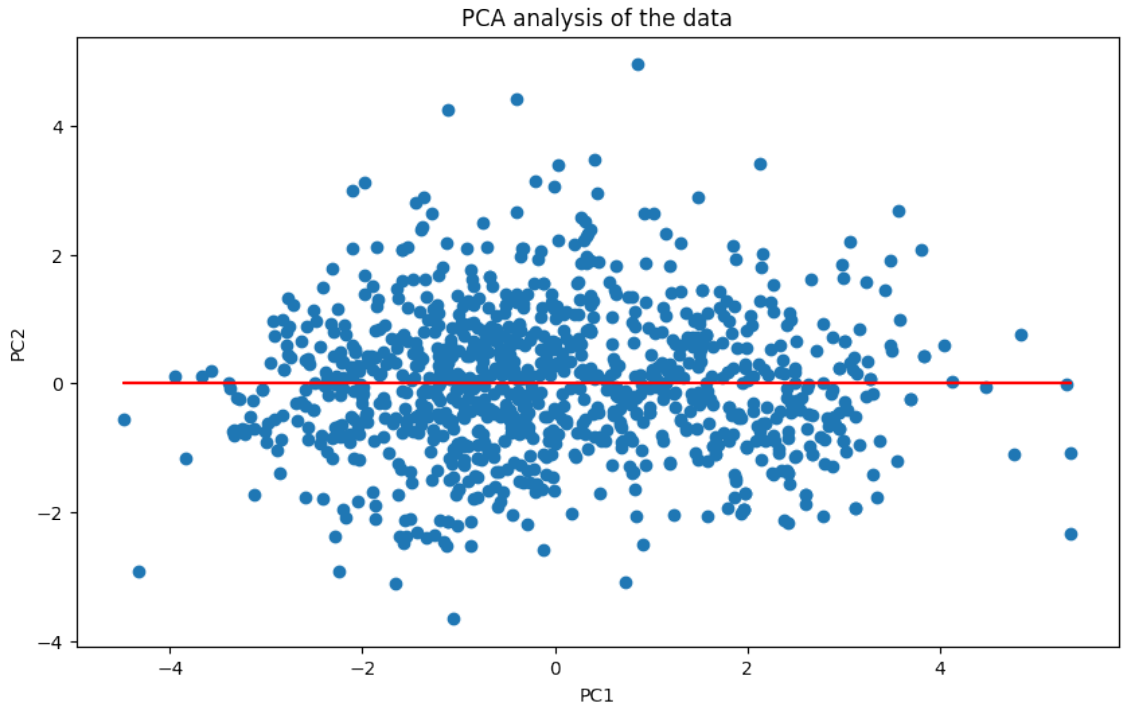

Figure 7: Principal Components analysis reduces 10 dimentsions to 2 dimentsions, the regression line describe the trend of those components

2.2 Dimensional Reduction and Data Loss

The given data of 10 features can be describe by 2 vectors using the Principal Component analysis procedure. The relation between the two vectors can be described in a 2D plot with a linear regression line to describe the trend of the two principal components (see Figure 7).

Dimensional reduction would also introduce information loss. In the given dataset, we have reconstructed the data from 2 dimensions to 10 dimension, results shows that there are 75% of data are mismatched with the original data.

```
# Perform data reconstruction
Xrex = pca.inverse_transform(Zred)
print(f'Reconstructed data shape: {Xrex.shape}')

# Measure the reconstruction error
rec_error = np.linalg.norm(Xnorm - Xrex) / np.linalg.norm(Xnorm)
print(f'Reconstruction error: {rec_error}')

>>>Reconstructed data shape: (980, 10)
```

```
>>>Reconstruction error: 0.7589269810329784
```

2.3 Explained Variance Percentage

Explained variance is a measurement for how much variance can be attributed from the dataset to each of the principal components (also known as eigenvectors) [4]. For example, a dataset with 10 dimension can be explained by two component, the first component can explain 20

We can calculate the explained variance by a function of ratio. For a n eigenvectors, the related eigenvalue λ_i and the sum of all eigenvalues, we calculate the explained variance with the following formula:

$$\frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} \tag{1}$$

3 Question 7

We apply PCA and Isomap algorithms [1] to reduce the dimension of the given dataset from 10 to 3. The KMeans clustering performance based on silhouette score is calculated for the original dataset and reduced dataset accross 50 clusters. The result is presented as Figure 8. As can be see, the original dataset have the lowest Silhouette score overall, thus the datasets with reduced dimension have better performance. It is noticable that the Isomap algorithm can achive better silhouette score compared to PCA.

4 Question 8

Replicate the clusters as given in Figure 9

4.1 Kmean Clustering

KMeans is a clustering algorithm athat divide a dataset into k clusters such that the average euclidean distance from a point to the cluster centers. Hence the KMeans algorithm would not likely to produce the clusters as given in Figure 9. Instead the Kmeans algorithm with $k = 6$ gives the results in Figure 10.

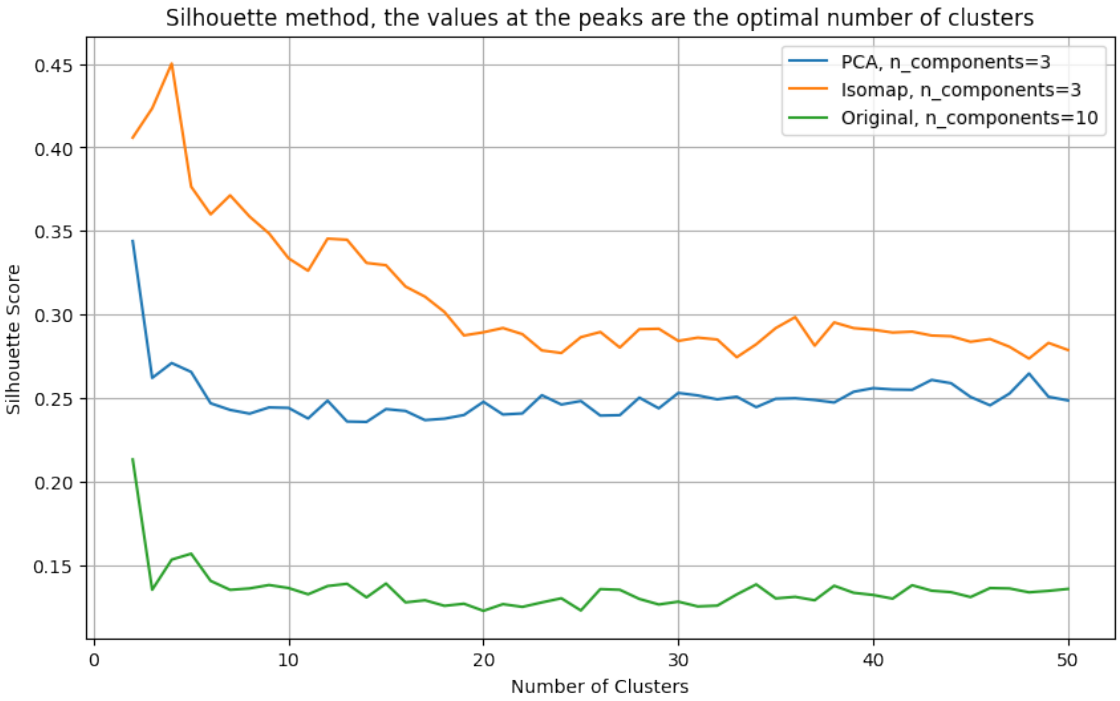

Figure 8: Silhouette score for PCA and Isomap reduced dataset, compared to the original

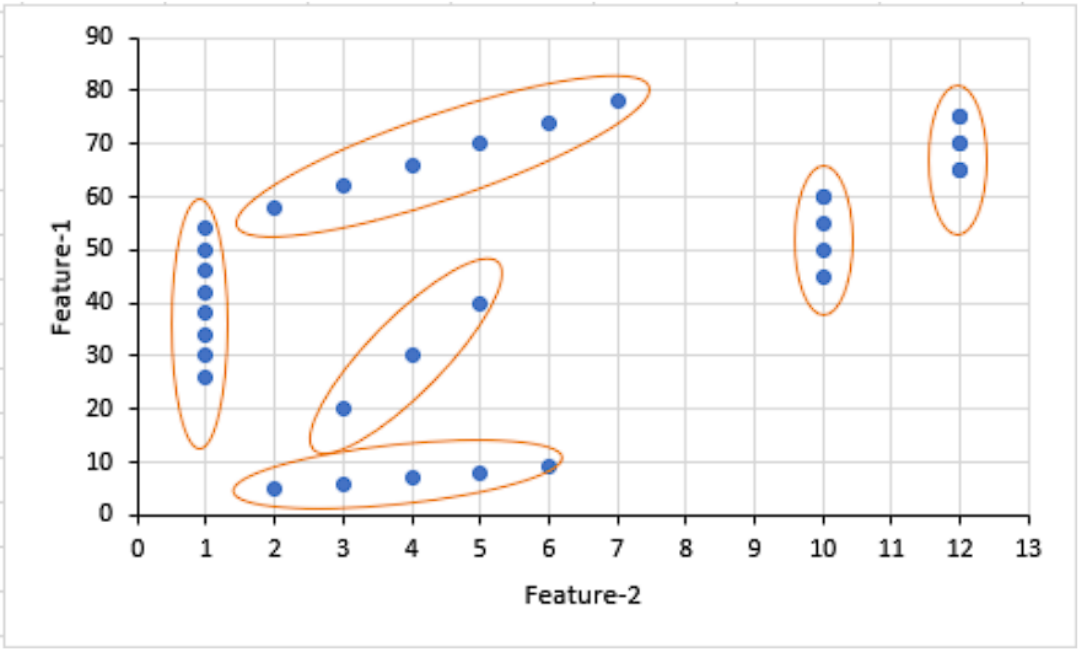

Figure 9: Scatter plot with expected clusters (ecliptical shapes)

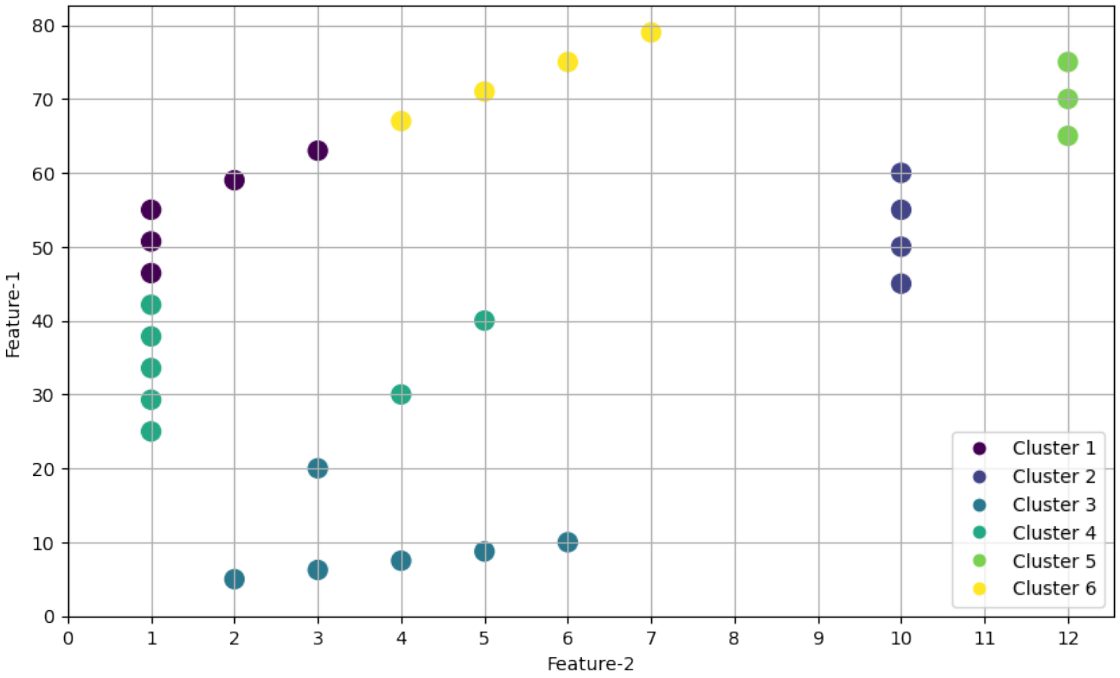

Figure 10: Scatter plot with 6 KMeans clusters

4.2 Gaussian Mixture

Compared to KMeans algorithm, Gaussian Mixture [5] is a model that assumes the distribution of how the data is generated. Data from Gaussian Mixture model are generally distributed within a spherical or eliptical shape. Given the clue that the data are fitted within eliptical clusters, we can use Gaussian Mixtrure to identify clusters as given in Figure 9. The result indeed shows the 6 clusters which matched the desired data (see Figure 11).

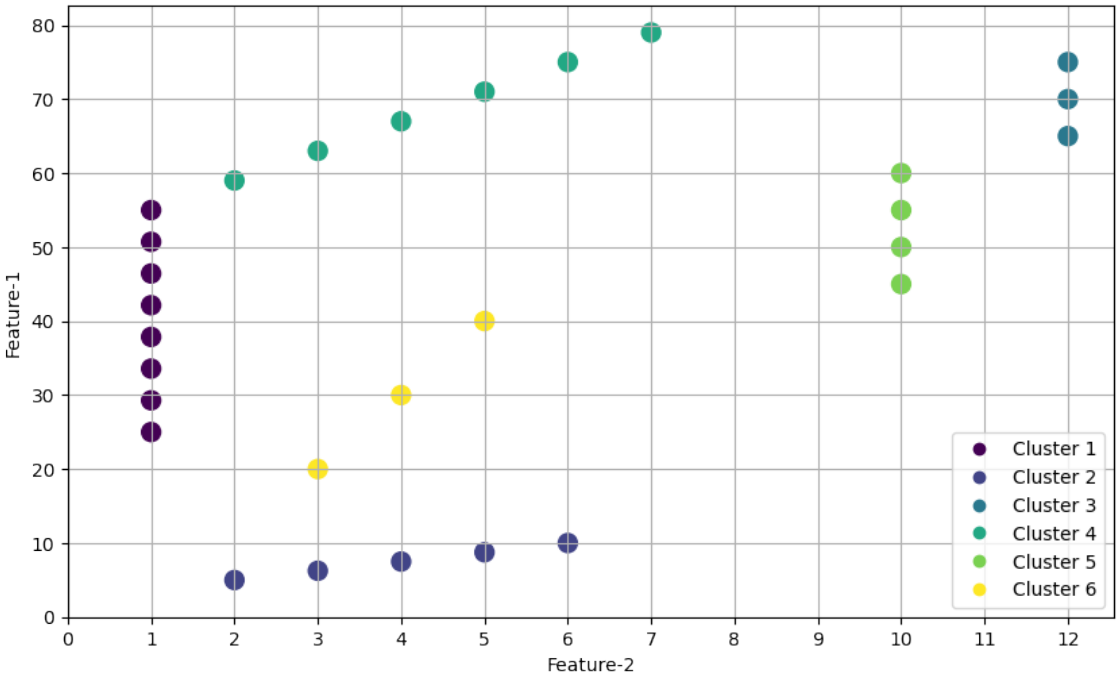

Figure 11: Clustering with Gaussian Mixture model, the data of each cluster matches with Figure 9

References

[1] F. Anowar, S. Sadaoui, and B. Selim. Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review*, 40:100378, 2021. 3

[2] M. Köppen. The curse of dimensionality. In *5th online world conference on soft computing in industrial applications (WSC5)*, volume 1, pages 4–8, 2000. 1.1

[3] J. Liu and J. Han. Spectral clustering. In *Data clustering*, pages 177–200. Chapman and Hall/CRC, 2018. 1.4

[4] K. E. O'Grady. Measures of explained variance: Cautions and limitations. *Psychological Bulletin*, 92(3):766, 1982. 2.3

[5] D. A. Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663), 2009. 4.2

[6] K. R. Shahapure and C. Nicholas. Cluster quality analysis using silhouette score. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 747–748. IEEE, 2020. 1.3

[7] M. Syakur, B. Khotimah, E. Rochman, and B. D. Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP conference series: materials science and engineering*, volume 336, page 012017. IOP Publishing, 2018. 1.2