# CENG 499 - Introduction to Machine Learning

Homework 2

Ahmet Eren Çolak
e2587921@ceng.metu.edu.tr

December 6, 2023

---

## Contents



---

# 1 Part 1

## 1.1 KNN

| ID | K | Similarity Metric |
|---|---|---|
| 1 | 3 | Cosine Distance |
| 2 | 3 | Minkowski Distance (p=2) |
| 3 | 5 | Cosine Distance |
| 4 | 5 | Minkowski Distance (p=2) |
| 5 | 5 | Mahalonobis Distance |

Table 1: Hyperparameter configurations

I applied 10-fold cross validation 5 times for each hyperparameter configuration and plotted their accuracy's 95% confidence intervals.

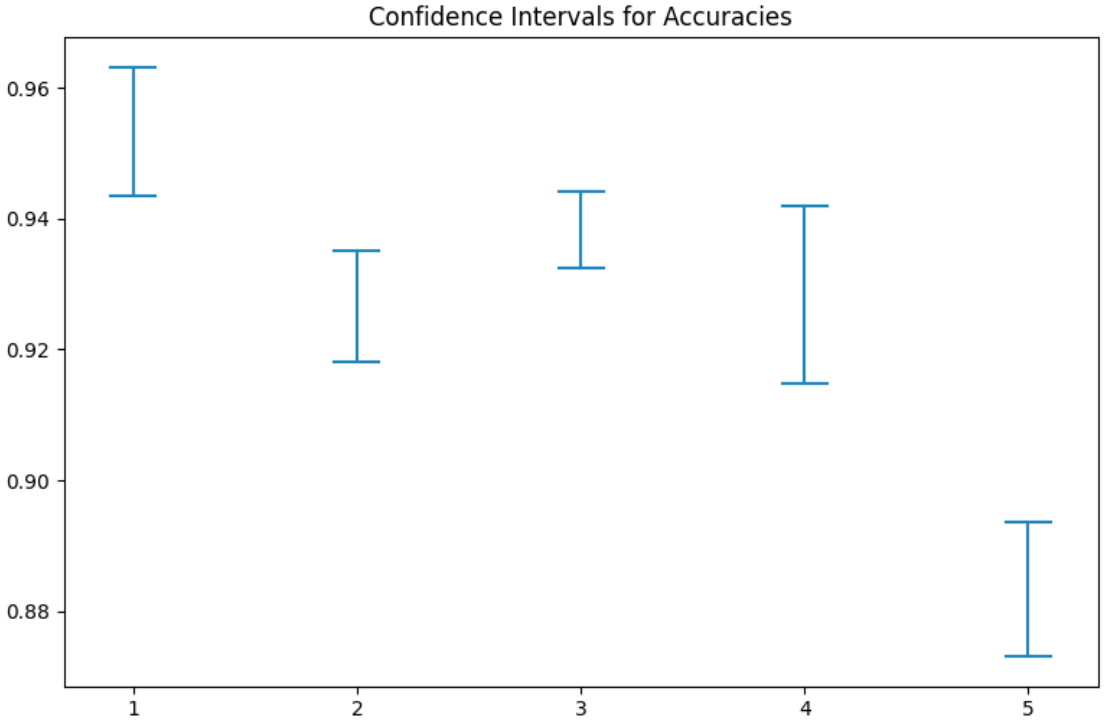


Figure 1: Accuracy confidence intervals for each hyperparameter

I tested the best performing model which is the first one with $K = 3$ and cosine distance metric, with the test set which I set apart before cross validation. Best performing model has the %96.67 accuracy on the test set.

# 2 Part 2

## 2.1 K-Means

I picked the initial cluster centers with K-means++ algorithm and ran K-means algorithm 10 times for each $K$ value and picked the smallest loss values for each $K$ value.

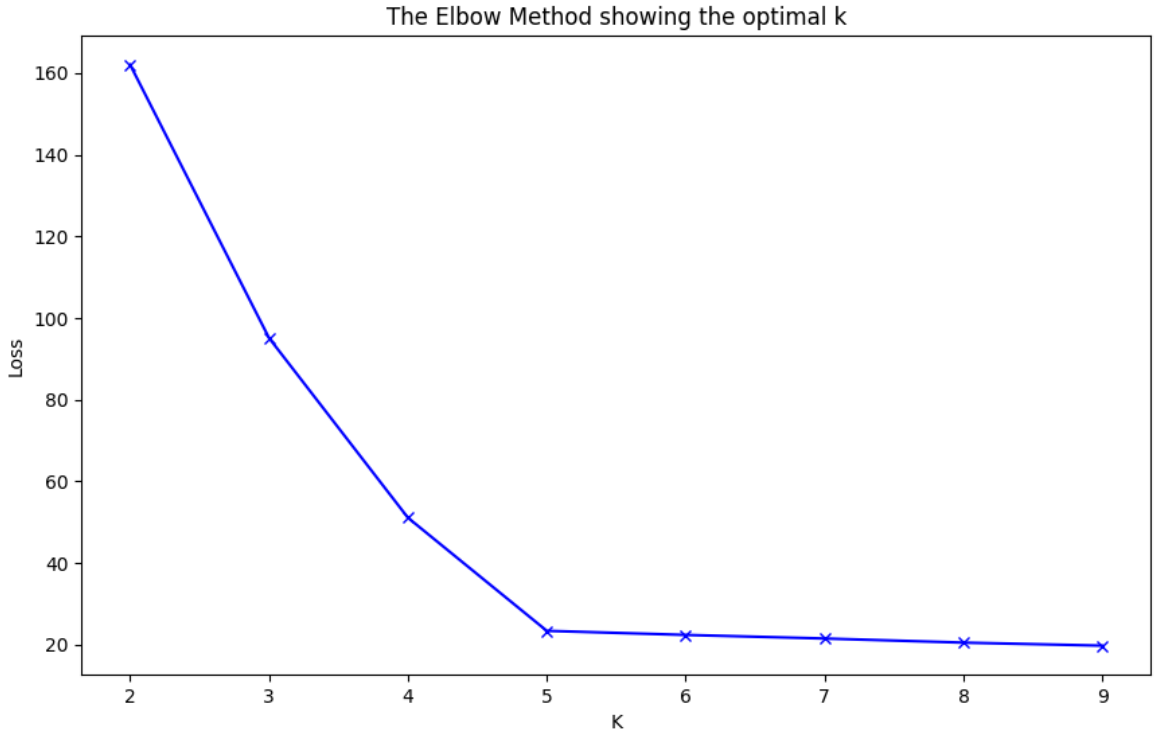


Figure 2: K vs. Loss for dataset 1

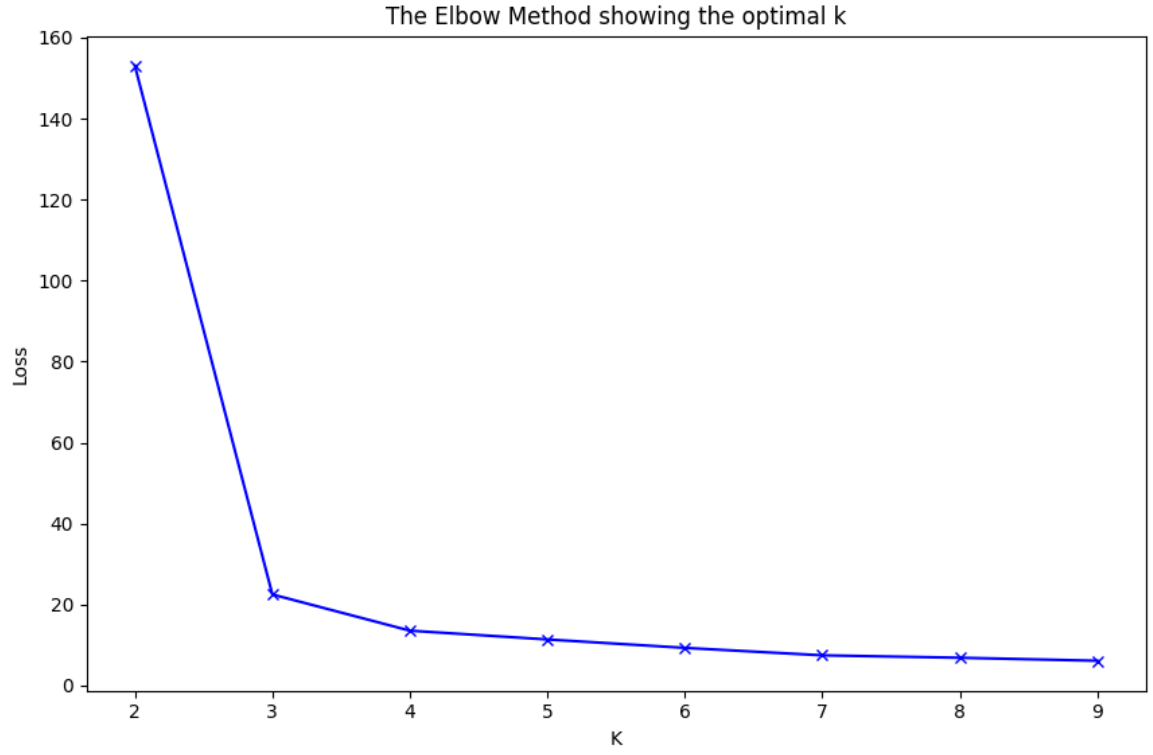


Figure 3: K vs. Loss for dataset 2

Best $K$ value for dataset 2 is 3 and the best $K$ value for dataset 1 is 5.

## 2.2 K-Medoids

I ran one K-medoids algorithm with cosine distance and one K-medoids algorithm with euclidean distance 10 times for each K value. Then I picked the smallest loss values for each K value.

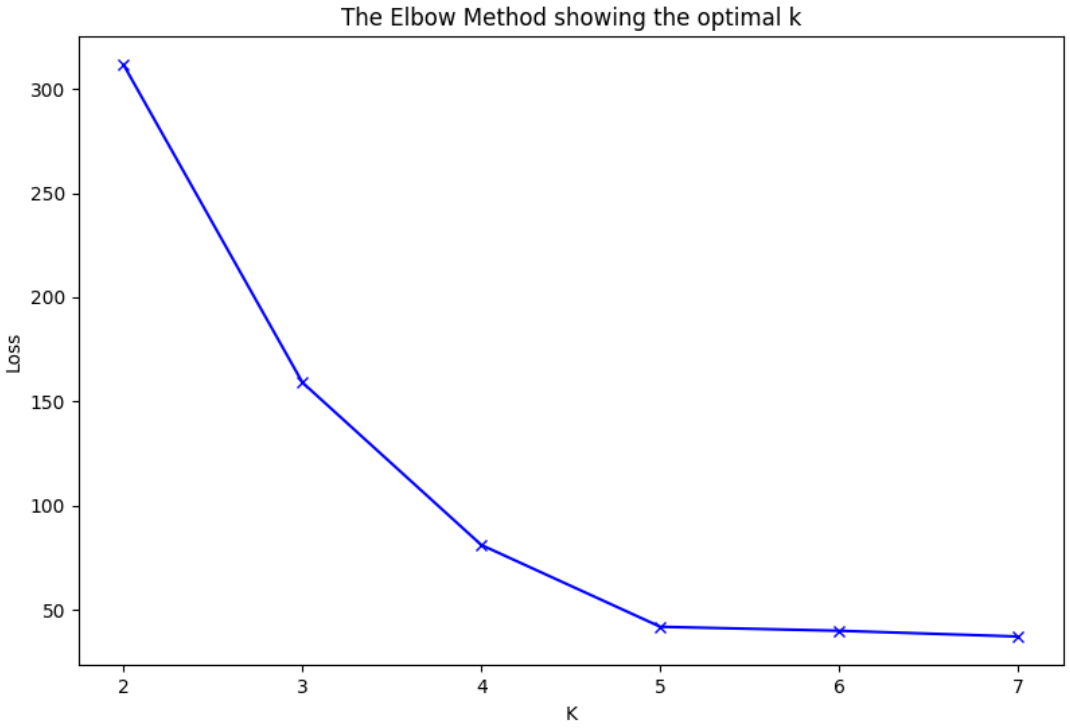


Figure 4: K vs. Loss for dataset 1 (cosine)

Best $K$ value for dataset 1 with cosine distance is 5.

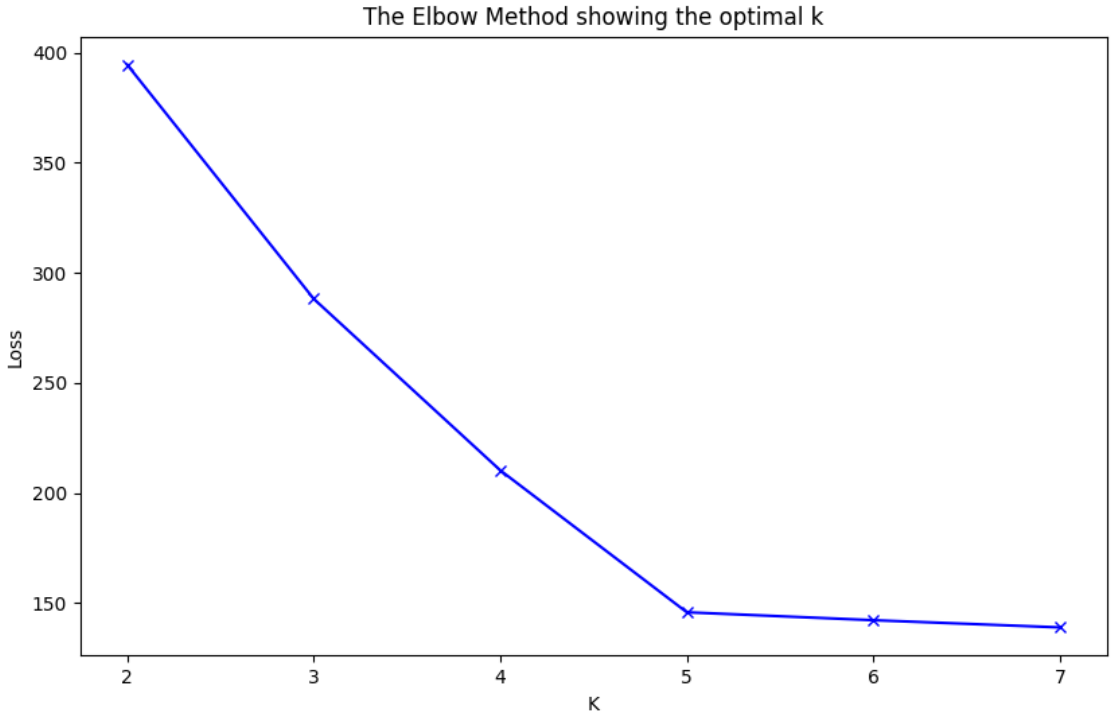


Figure 5: K vs. Loss for dataset 1 (euclidean)

Best $K$ value for dataset 1 with euclidean distance is 5.

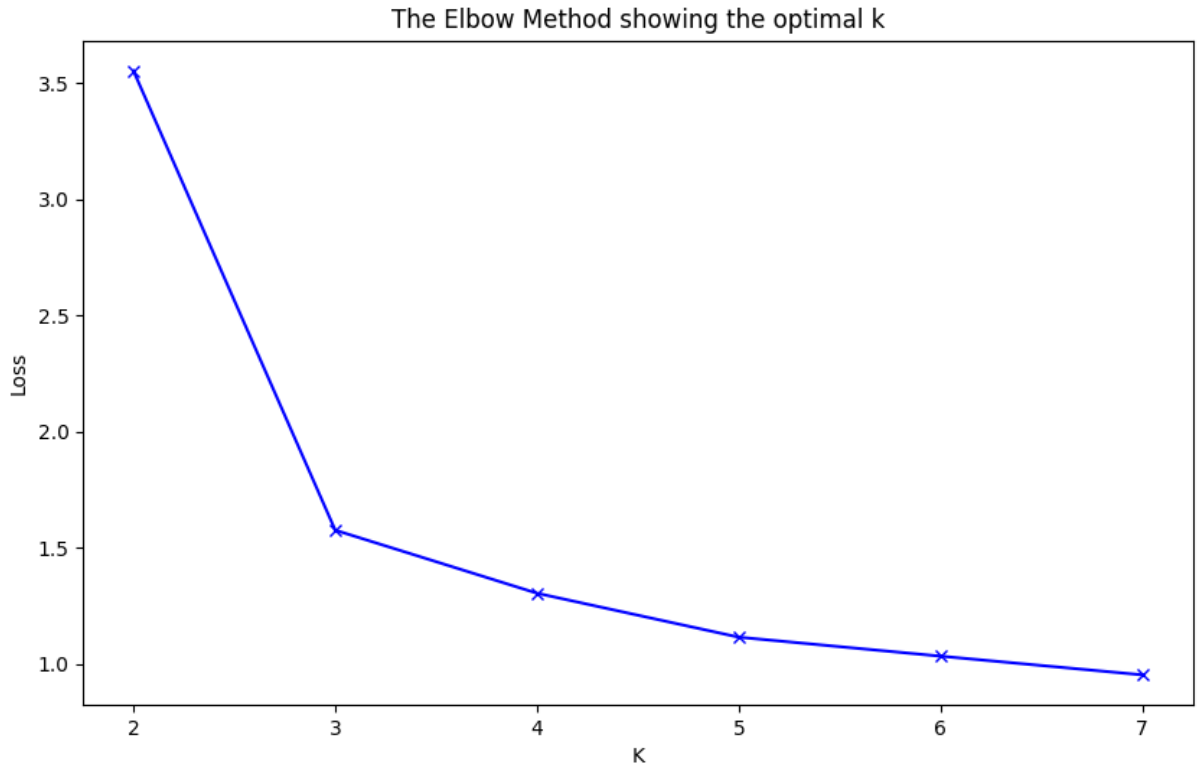


Figure 6: K vs. Loss for dataset 2 (cosine)

Best $K$ value for dataset 2 with cosine distance is 3.

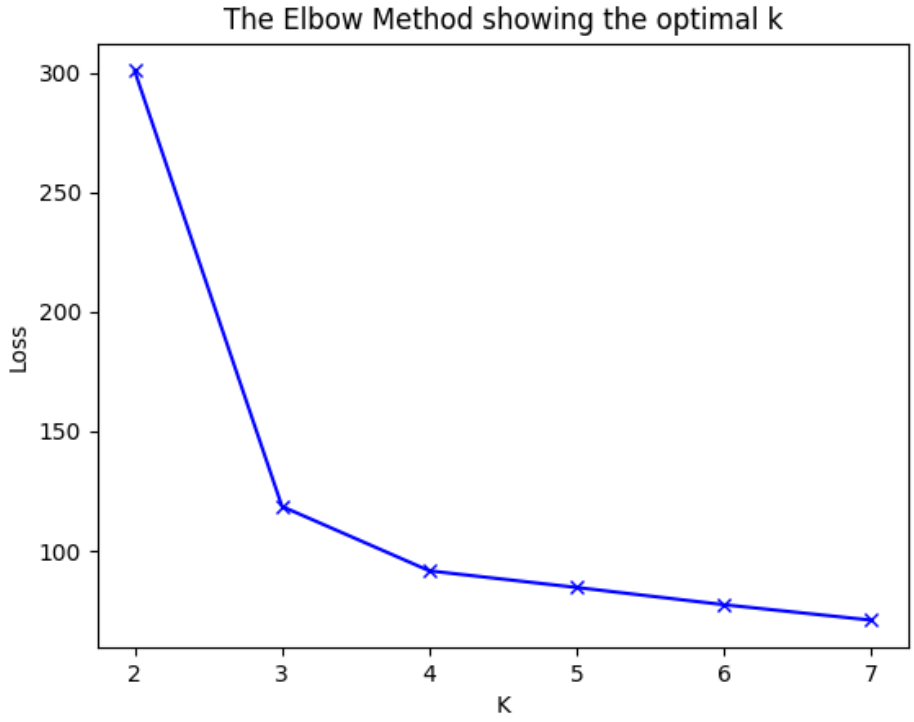


Figure 7: K vs. Loss for dataset 2 (cosine)

Best $K$ value for dataset 2 with euclidean distance is 3.

## 2.3 2D Visualisation

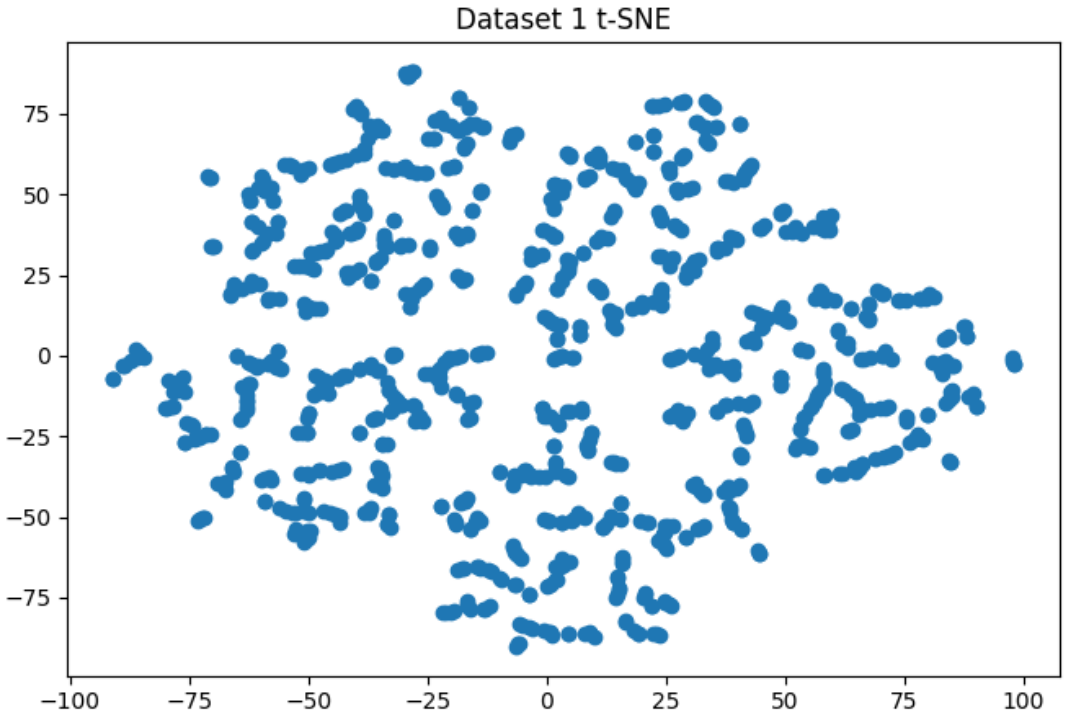


Figure 8: Dataset 1 dimensionality reduction with t-SNE

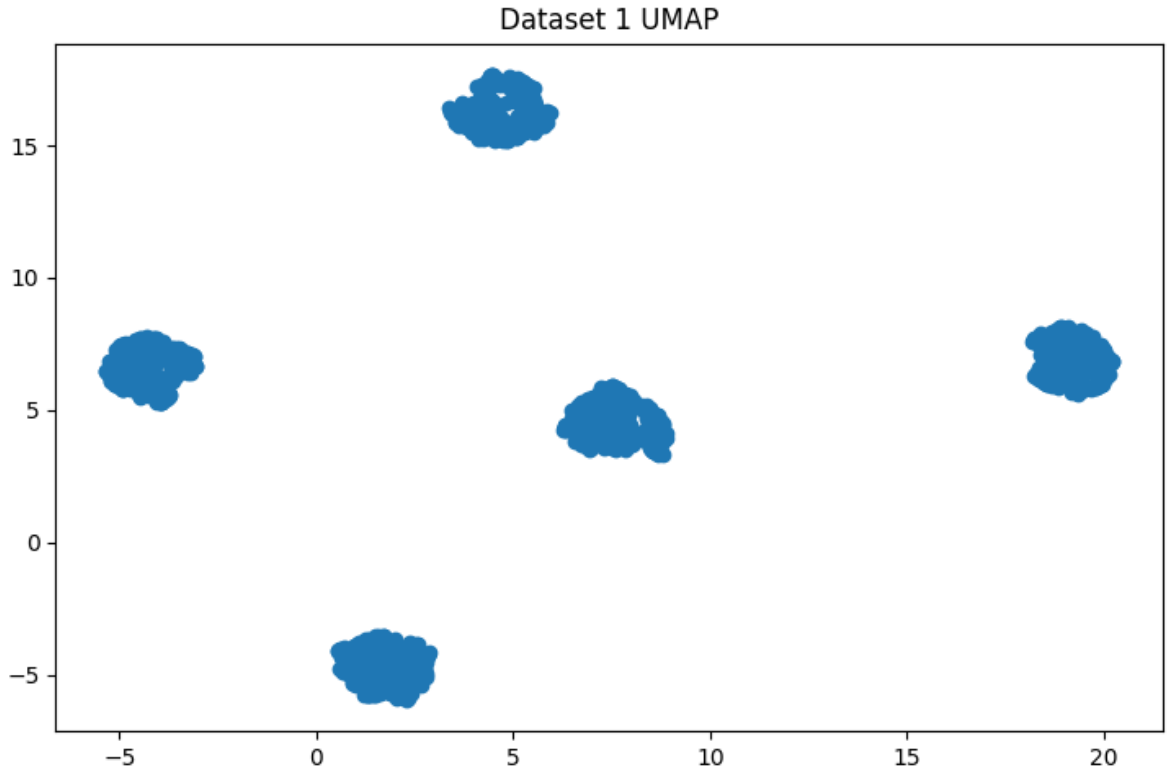


Figure 9: Dataset 1 dimensionality reduction with UMAP

For dataset 1, optimum $K$ value obtained from elbow method and visualizations match with each other and is equal to 5.

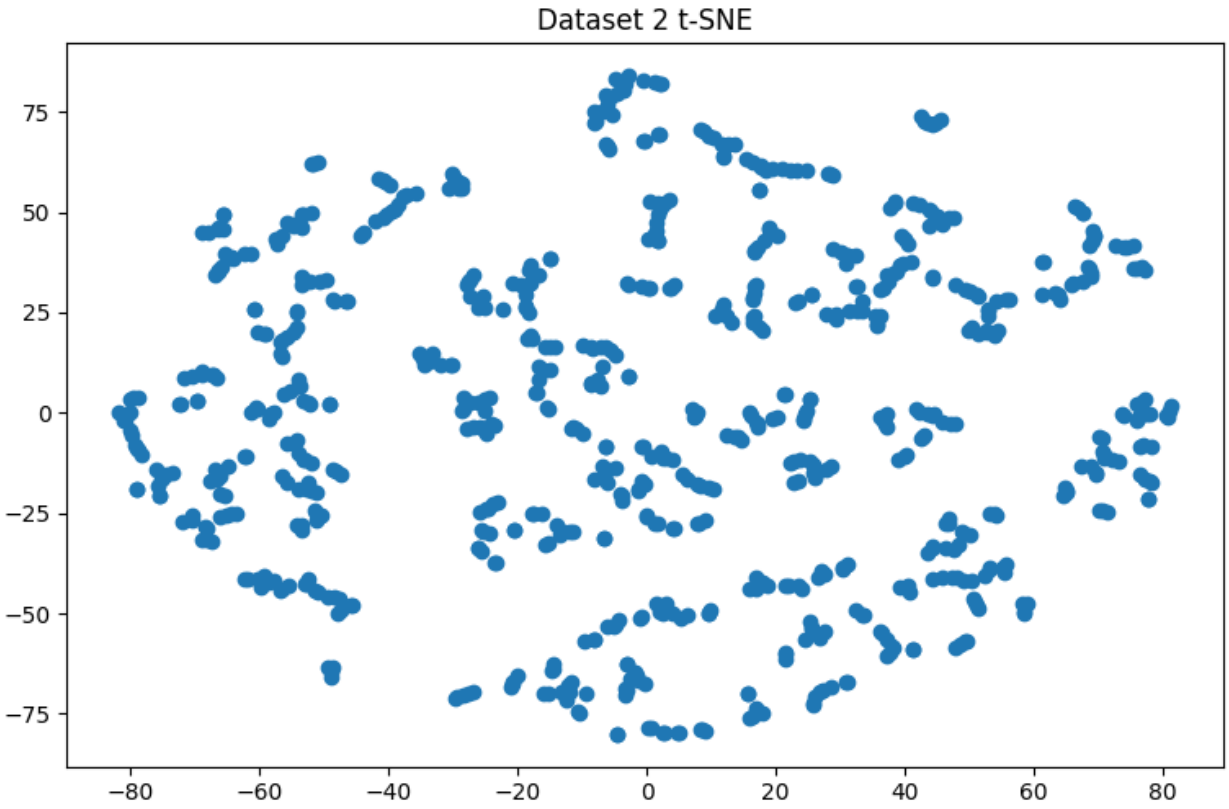


Figure 10: Dataset 2 dimensionality reduction with t-SNE

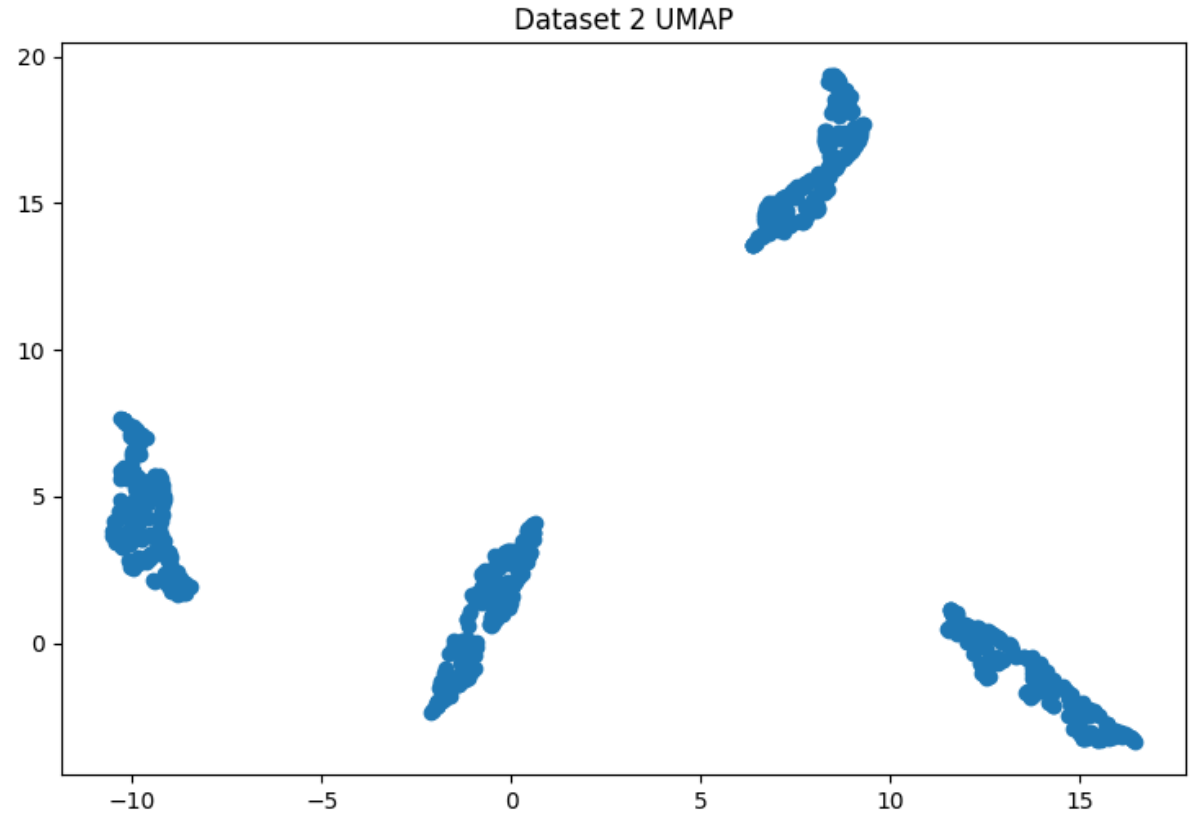


Figure 11: Dataset 2 dimensionality reduction with UMAP

For dataset 2 however, optimum $K$ value obtained from elbow method and visualizations does not match with each other. According to t-SNE and UMAP visualizations there are 4 clusters. In contrary to 2D visualisations, elbow method suggest that optimum $K$ value is 3.

### 2.4 Time Complexity Analysis

Worst case time complexity of K-means is $O(kndi)$, where $k$ is the number of clusters, $n$ is the number of data samples, $d$ is the number of attributes, $i$ is the maximum number of iterations till convergence. Worst case time complexity of K-medoid is worse than K-means since we have to iterate whole dataset once more to find data examples which minimizes loss. That means worst case time complexity of K-medoids is $O(kn^2i)$.

## 3 Part 3

### 3.1 HAC

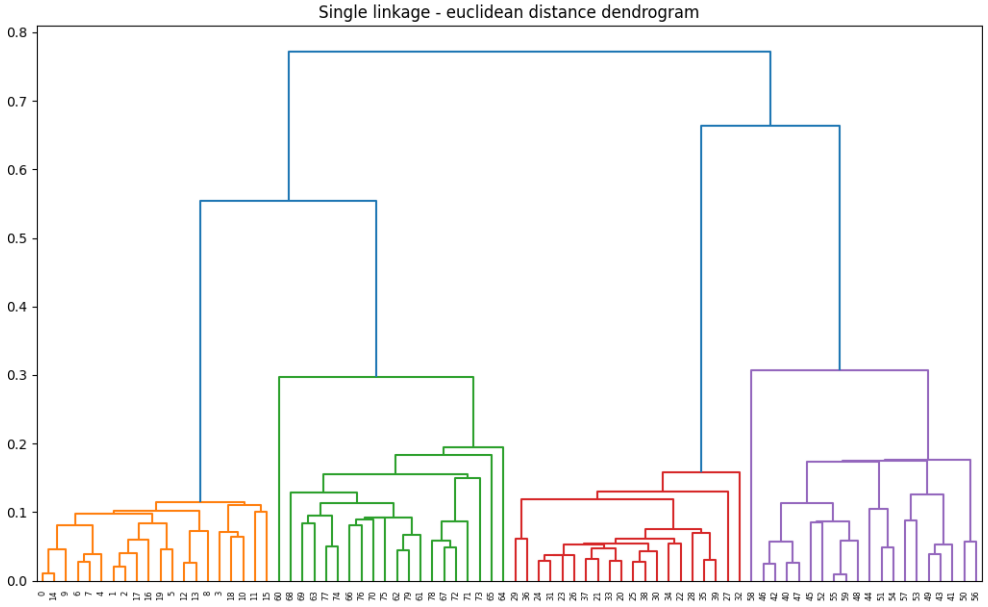


Figure 12: Dendrogram for single linkage - euclidean distance HAC

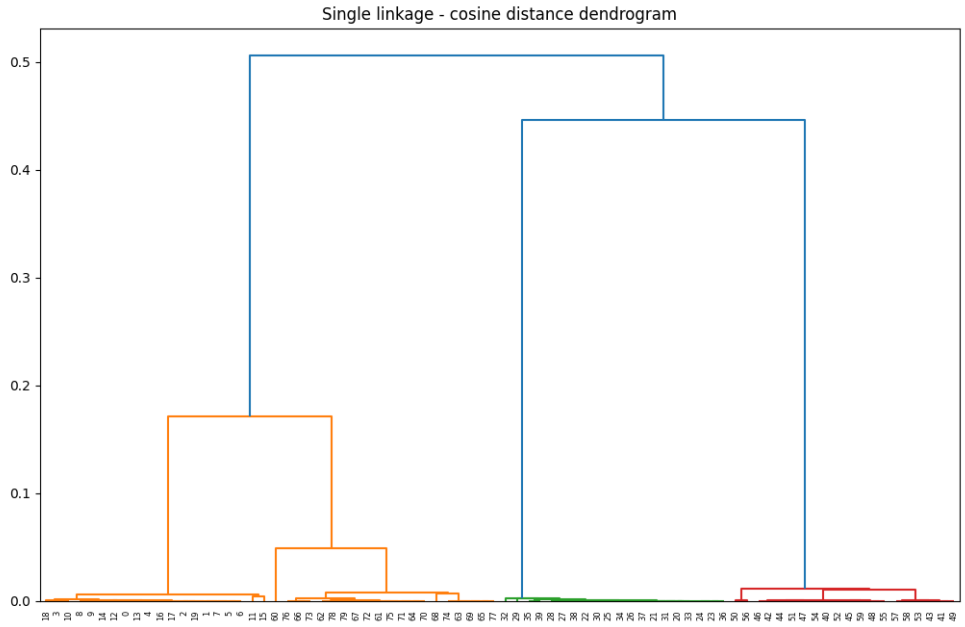


Figure 13: Dendrogram for single linkage - cosine distance HAC

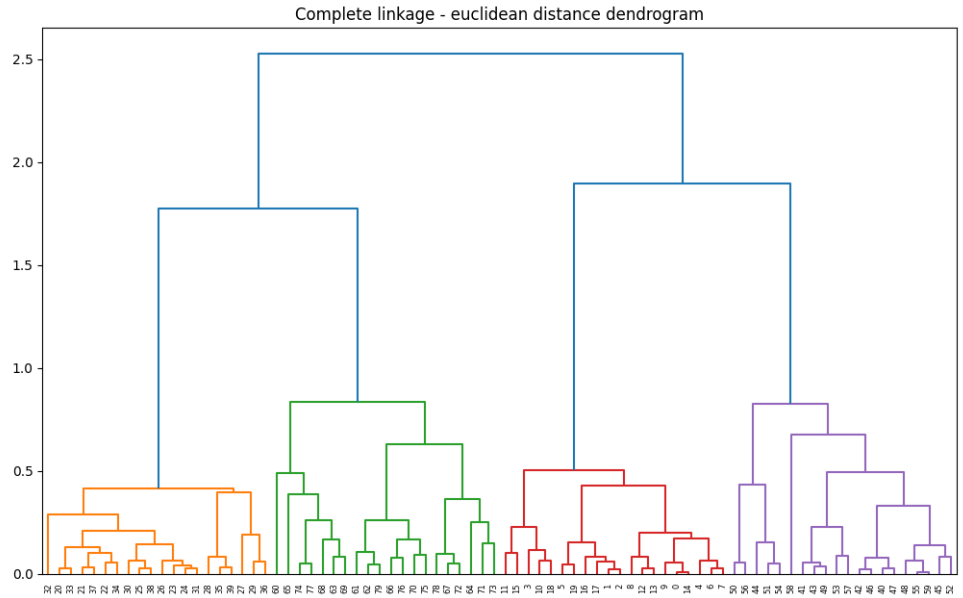


Figure 14: Dendrogram for complete linkage - euclidean distance HAC

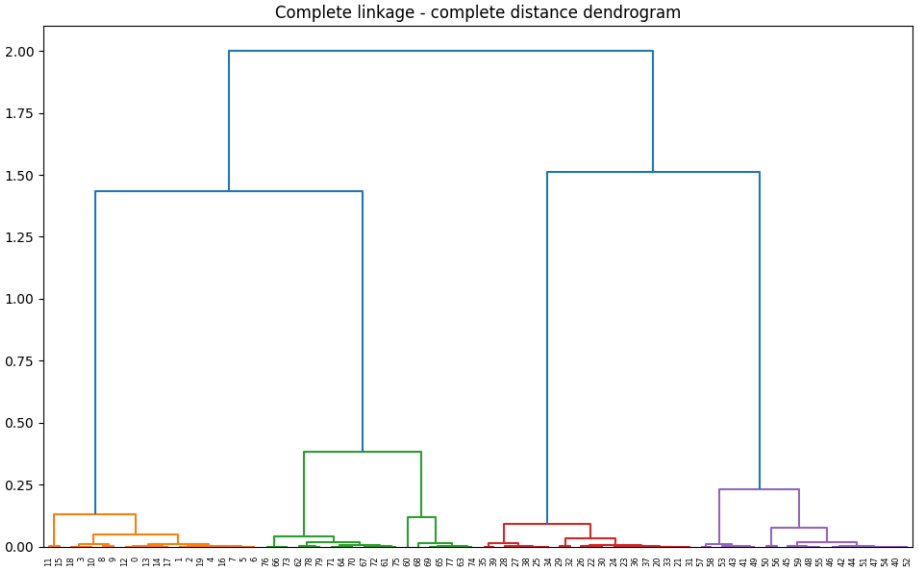


Figure 15: Dendrogram for complete linkage - cosine distance HAC

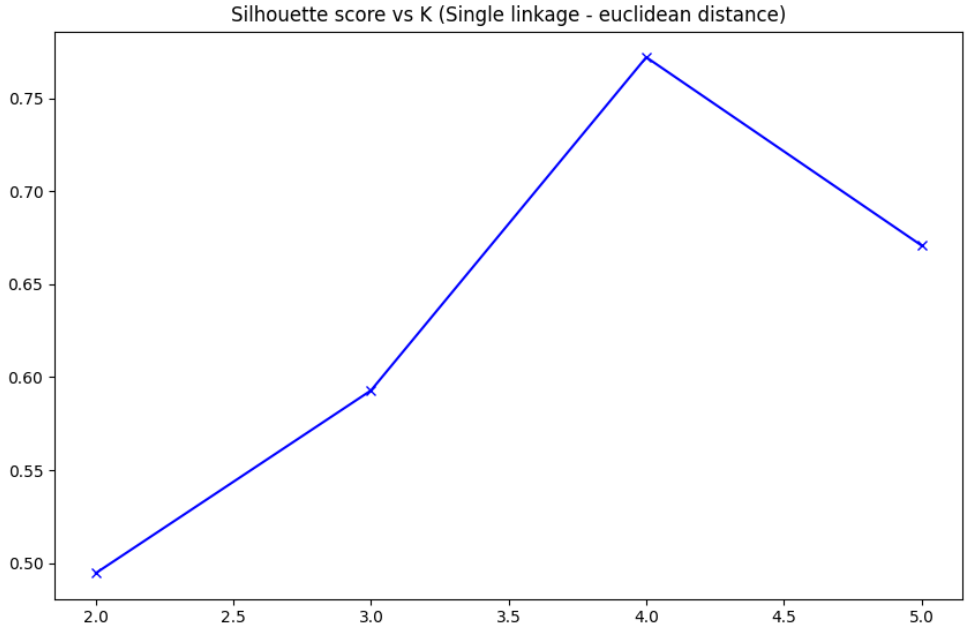


Figure 16: Silhouette scores for single linkage - euclidean distance HAC

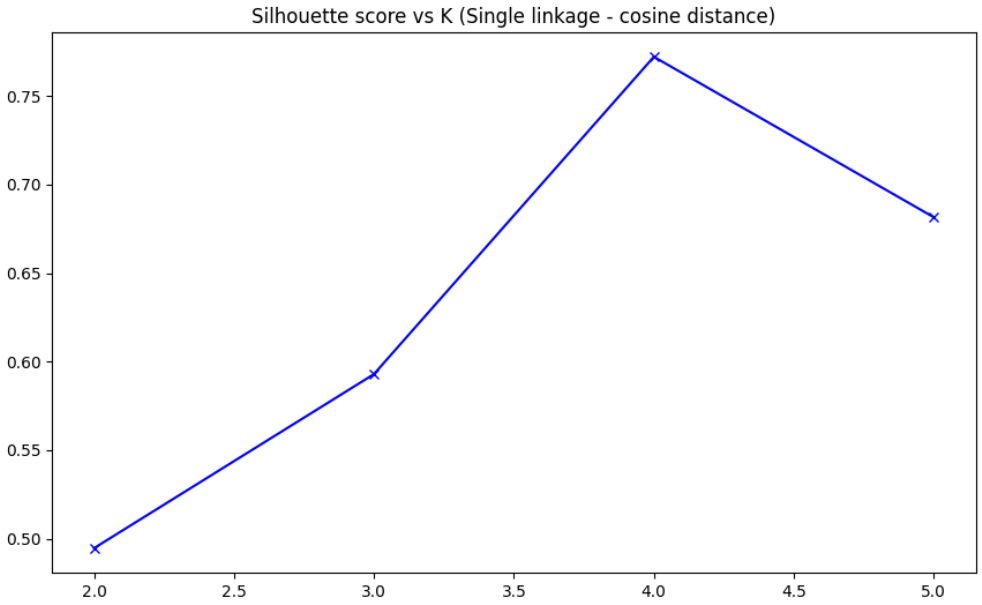


Figure 17: Silhouette scores for single linkage - cosine distance HAC

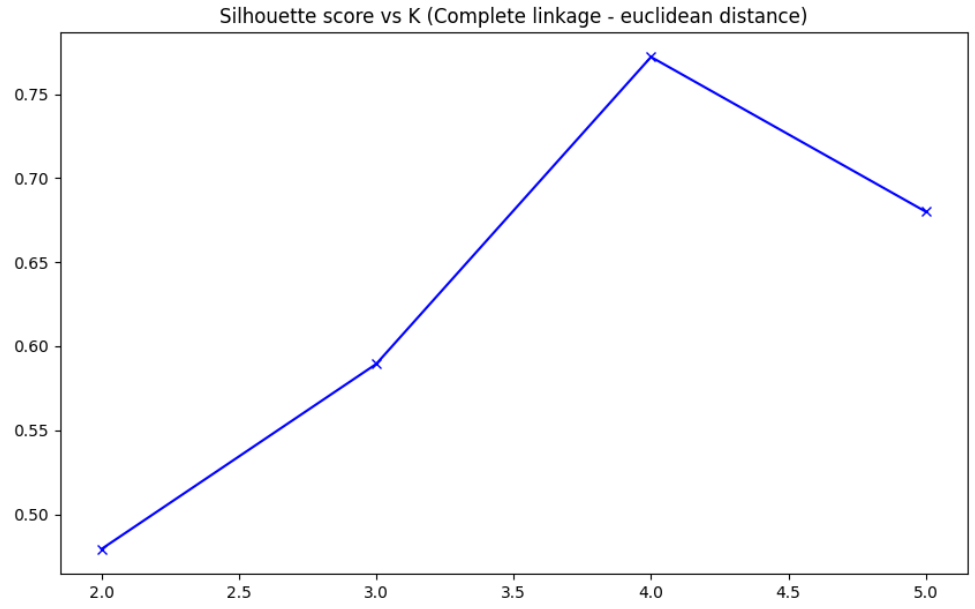


Figure 18: Silhouette scores for complete linkage - euclidean distance HAC

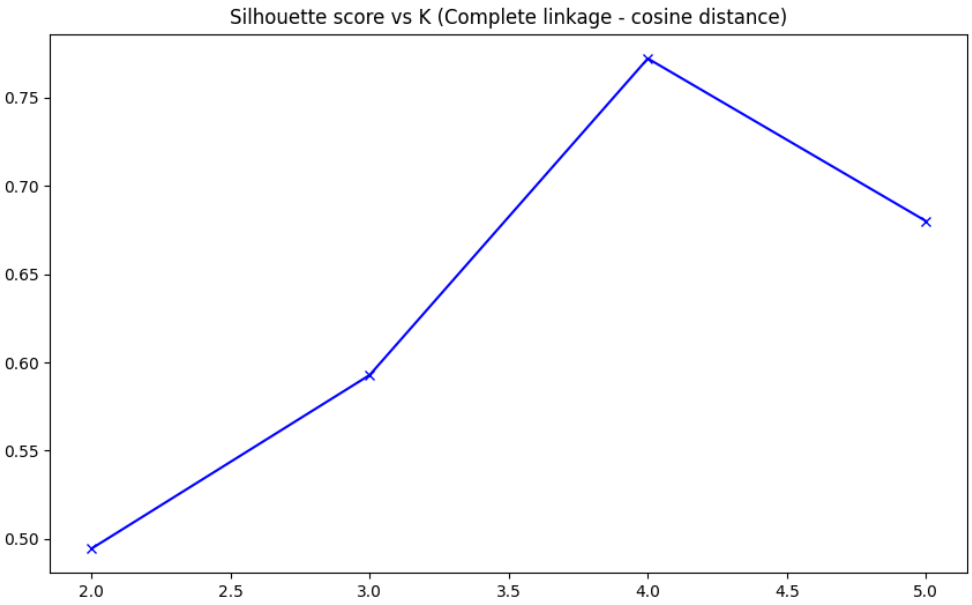


Figure 19: Silhouette scores for complete linkage - cosine distance HAC

Best K value is 4 for all of the configurations of HAC since it has the highest silhouette score for all of them.

## 3.2 2D Visualisation

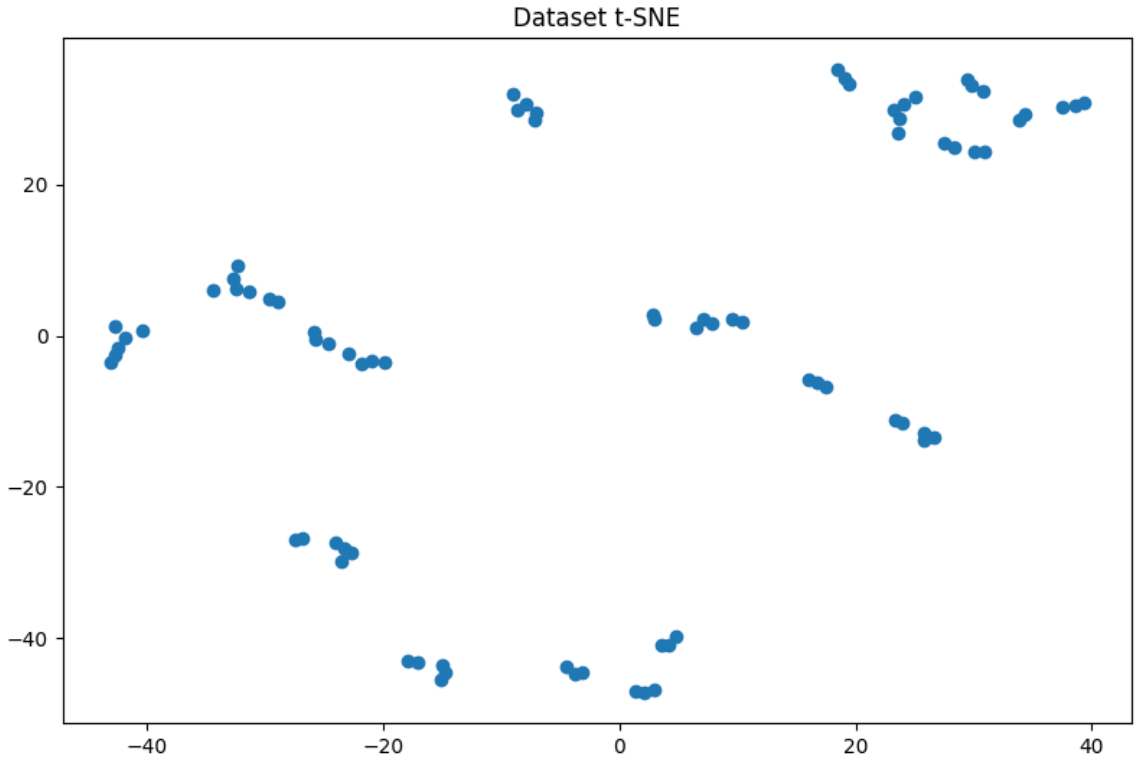


Figure 20: Dataset t-SNE reduction

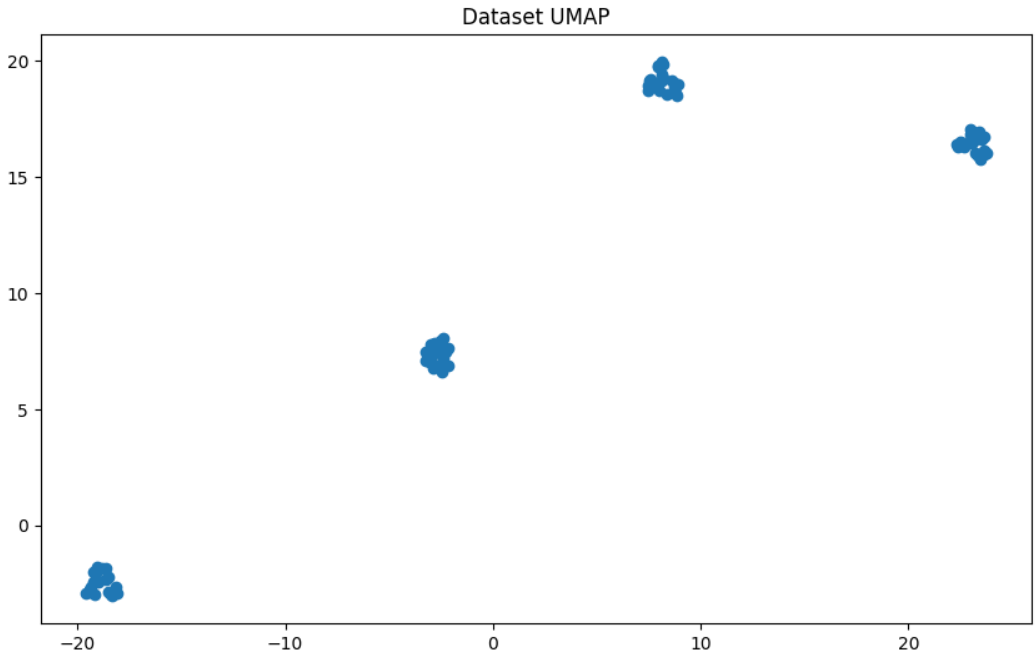


Figure 21: Dataset UMAP reduction

Although clusters are not quite obvious in the t-SNE reduction graph, result of the UMAP reduction aligns with the silhouette score analysis. In the UMAP graph, it is obvious that there are 4 clusters.

## 3.3 Time Complexity Analysis

Worst case time complexity of the HAC is $O(n^3 d)$, where $n$ is the number of data samples and $d$ is the number of attributes. Assuming that number of iterations $i$, is not too large for K-means, I would prefer K-means to cluster dataset with large amount of samples.