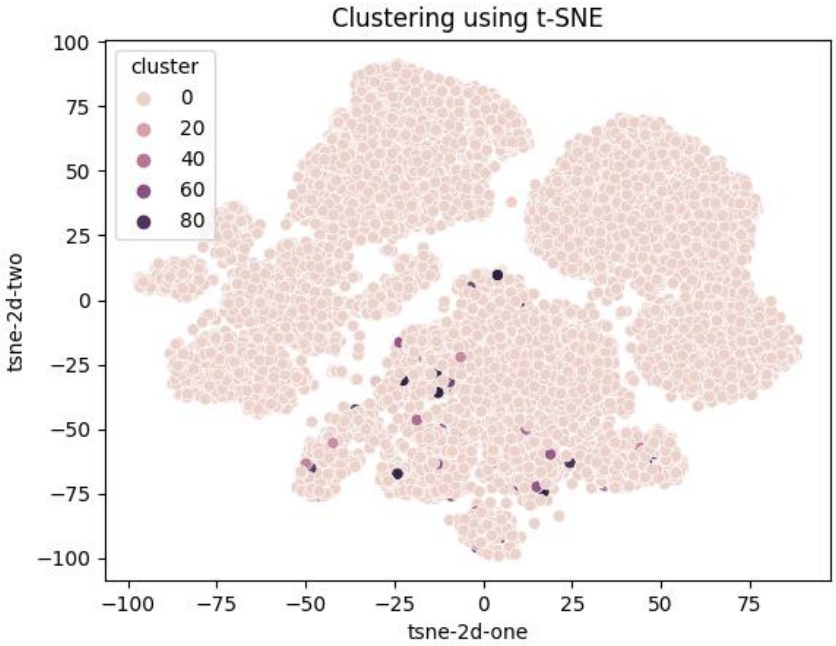
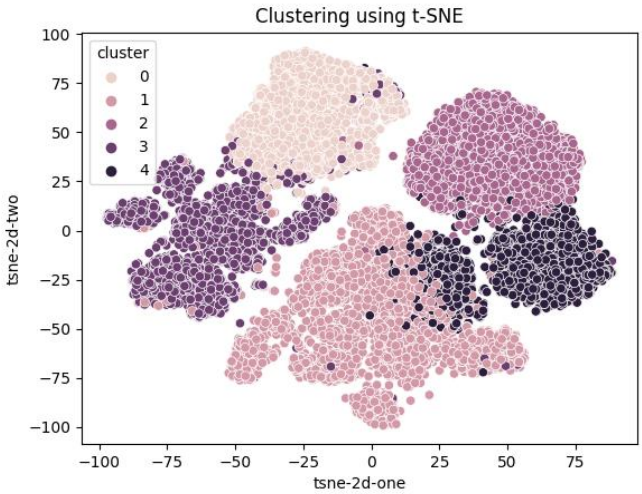
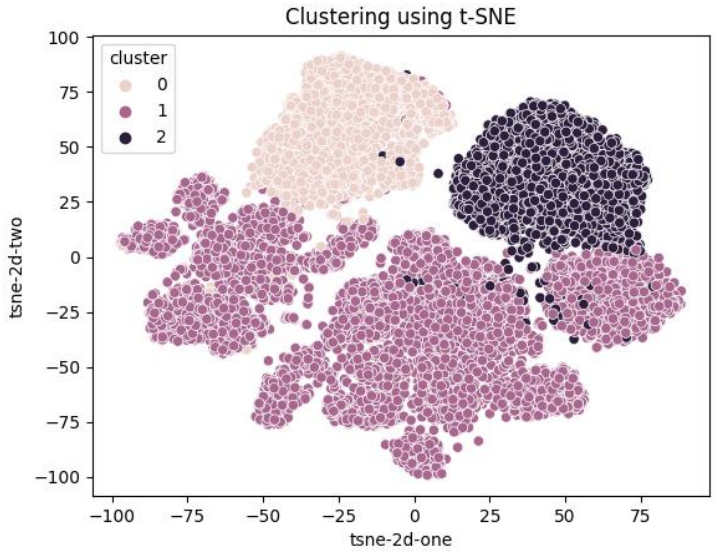
Initially, we used the DBSCAN algorithm to cluster the data. However, apart from the time it takes, DBSCAN does not effectively cluster the data because it doesn't know the number of clusters beforehand, which prevents it from correctly identifying clusters within our dataset due to its structure. Thus, this algorithm is not suitable for our data.



Alternatively, using the K-means algorithm, we can reach results faster and determine an optimal number of clusters by comparing metrics. Since we can specify the number of clusters beforehand, K-means provides better results.



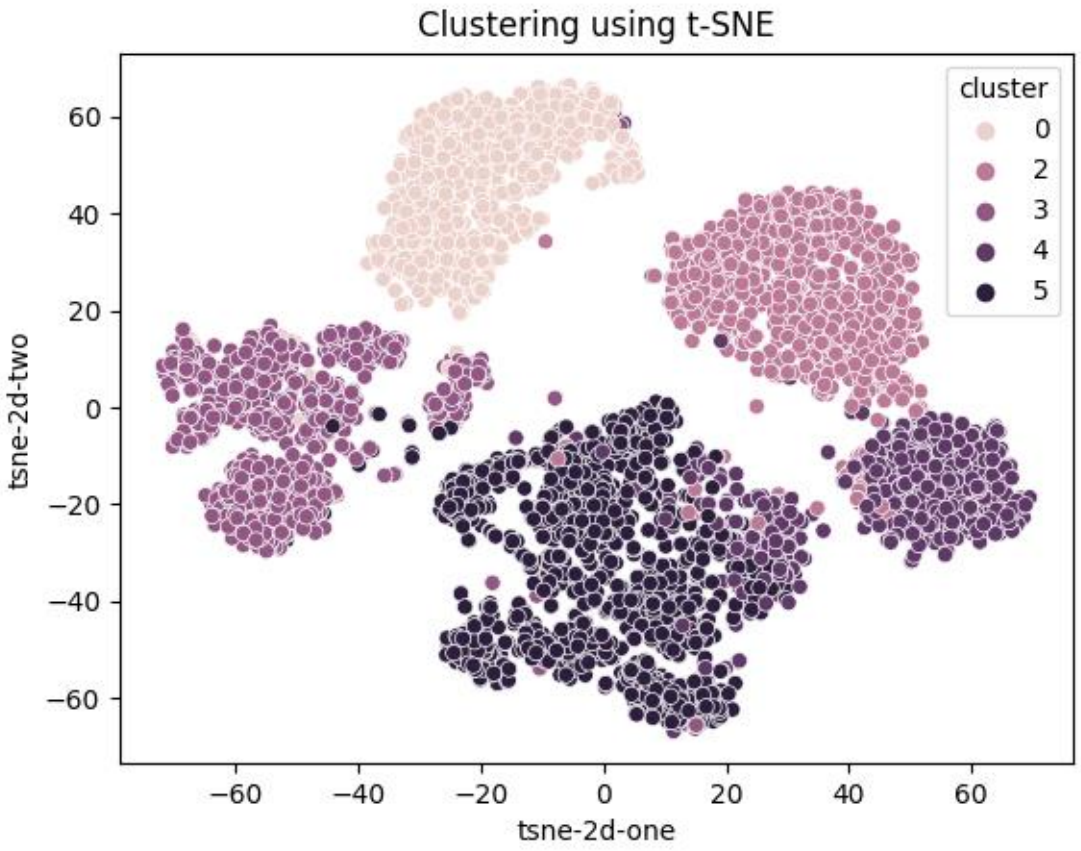


Finally, by applying K-means with various cluster numbers and comparing silhouette and Calinski-Harabasz Index values, we found that three clusters yield the best result. Higher values of these indices indicate better clustering performance.

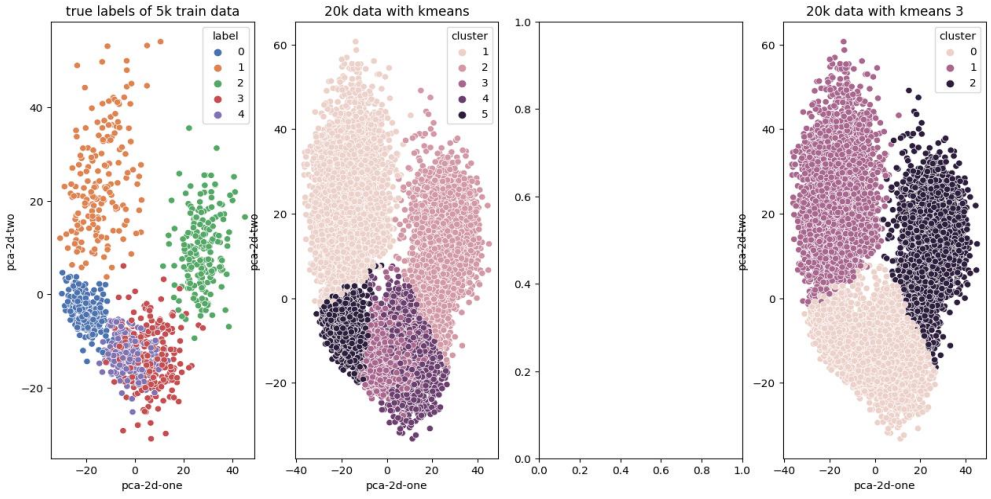
From the visualizations, we observed that the data could potentially be divided into five clusters based on density and spacing. However, K-means with five clusters does not give the desired results, so heuristic methods are considered.

The first approach increases the number of clusters, then merges clusters with nearby centers until we reach the desired number.

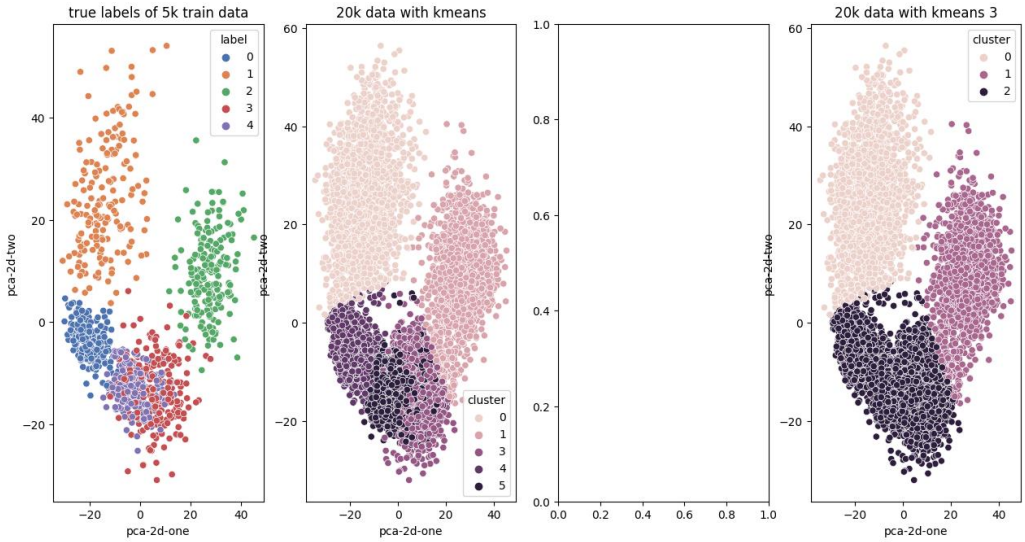
Our chosen approach first applies K-means with three clusters, revealing that cluster 1 (which contains more data points) can be further subdivided. Since we aim for good intra-cluster density and sufficient separation from other clusters, we reapply K-means with three clusters to cluster 1. The three new density centers, combined with the previous two, result in five final centers, achieving optimal clustering.



By comparing with 5% of labeled data, we find our results closely match the actual labels.



The left figure shows actual labels, and the adjacent figure shows our clustering after the above steps, which aligns well with the labeled data.



Our final clustering on the test data is shown above. Since this is an unsupervised problem, there's no definitive way to verify clustering accuracy. However, based on the silhouette and Calinski-Harabasz metrics and comparison with 5,000 correctly labeled data points, we achieved better results.