Customer Segmentation

Younes Mahdavi

Sep 24, 2024

DATA

8950 unique records, all unlabeled, from customers of a credit card company.

kaggle

Features	
1. CUST_ID	10. PURCHASES_INSTALLMENTS_FREQUENCY
2. BALANCE	11. CASHADVANCE_FREQUENCY
3. BALANCE_FREQUENCY	12. CASH_ADVANCE_TRX
4. PURCHASES	13. PURCHASES_TRX
5. ONEOFF_PURCHASES	14. CREDIT_LIMIT
6. INSTALLMENTS_PURCHASES	15. PAYMENTS
7. CASH_ADVANCE	16. MINIMUM_PAYMENTS
8. PURCHASES_FREQUENCY	17. PRC_FULL_PAYMEN
9. ONEOFF_PURCHASES_FREQUENCY	18. TENURE

Exploratory Data Analysis

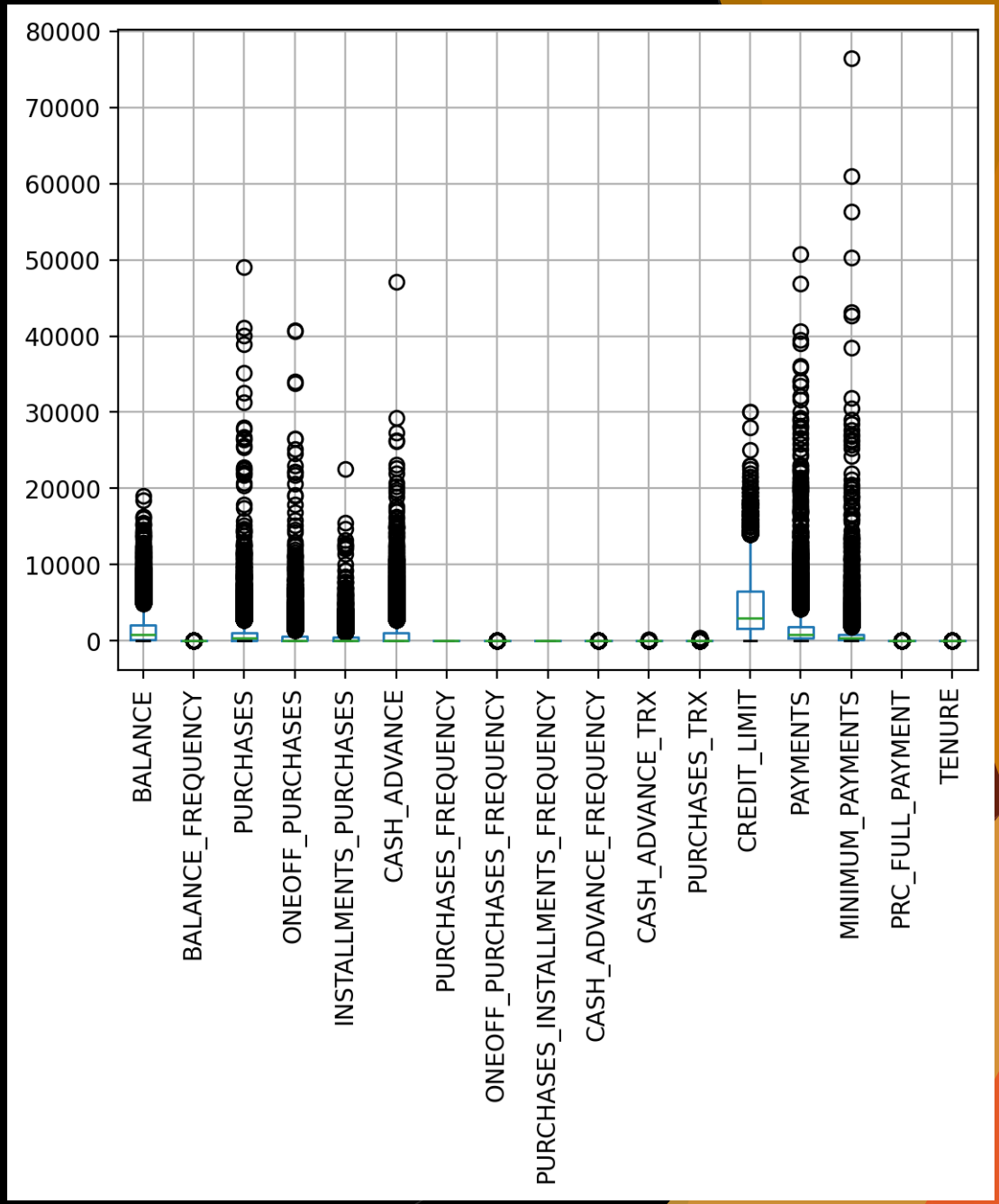
The dataset had a few missing values in the following features, which have been filled with the mean of each feature:

- CREDIT_LIMIT
- MINIMUM_PAYMENTS

Outliers

Methods used to detect outliers:

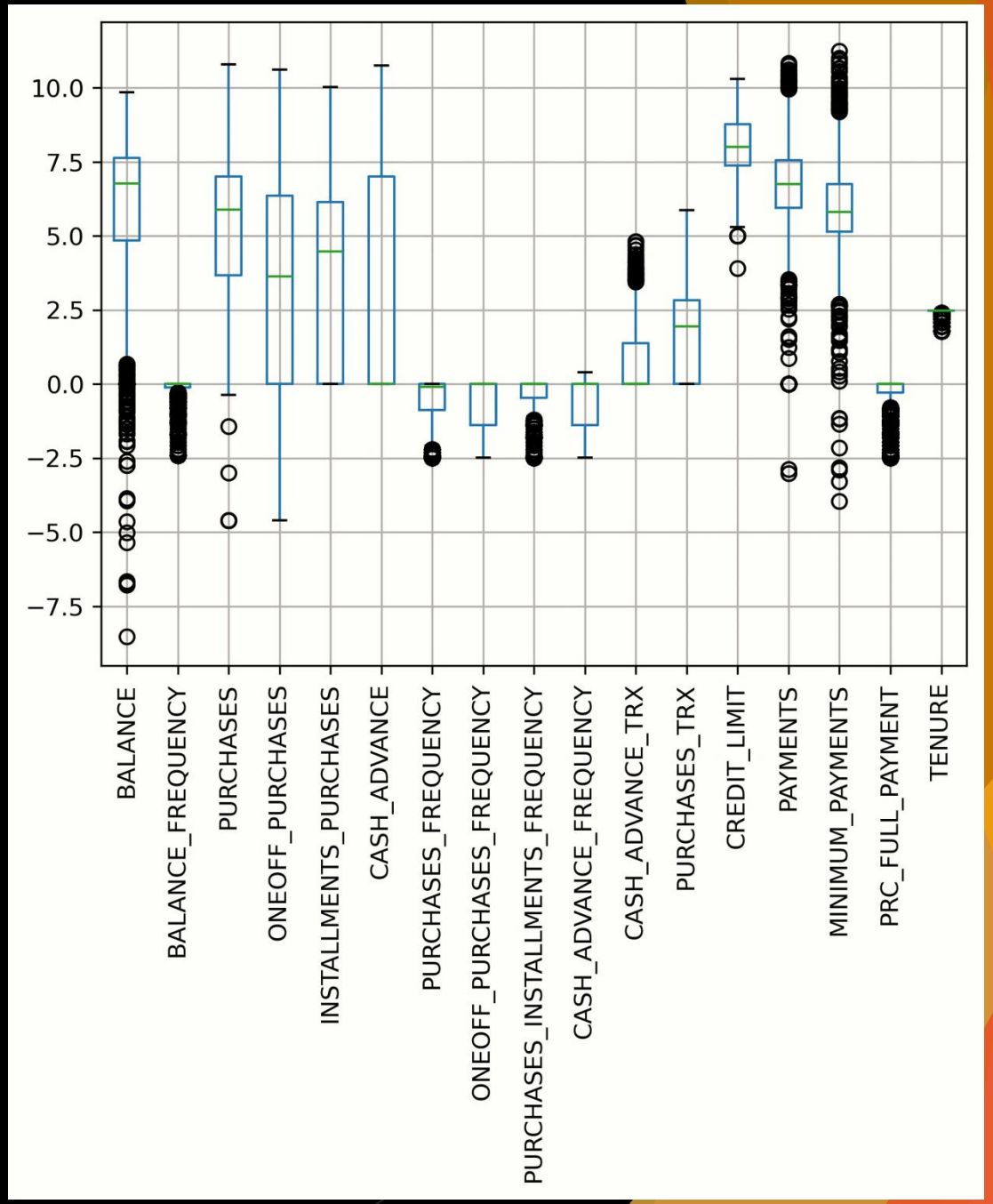
- Visualization (Boxplot)
- Tukey's test



Outliers capping and transforming

Methods:

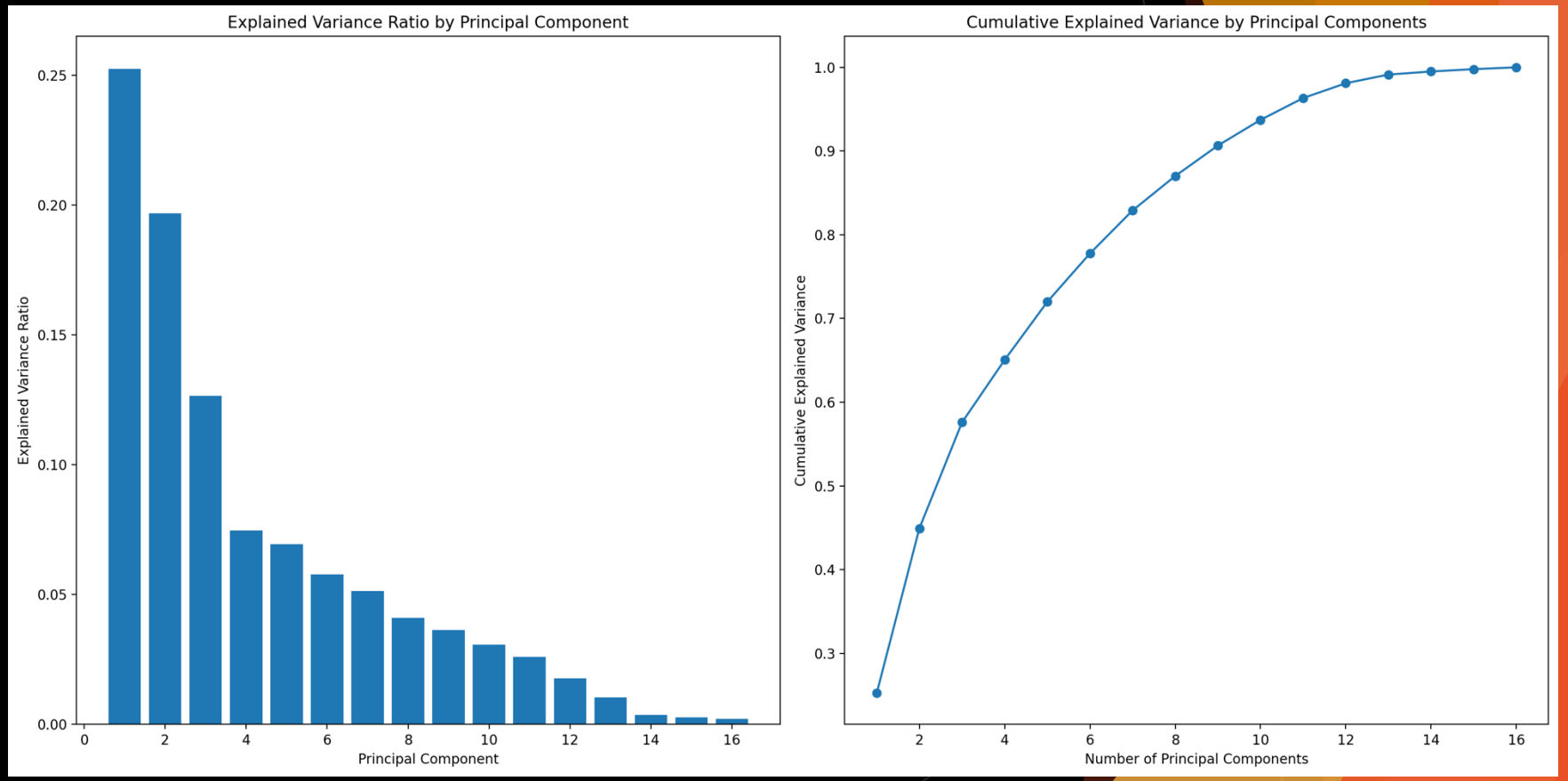
- Winsorization
- Log_transform



Feature Selection

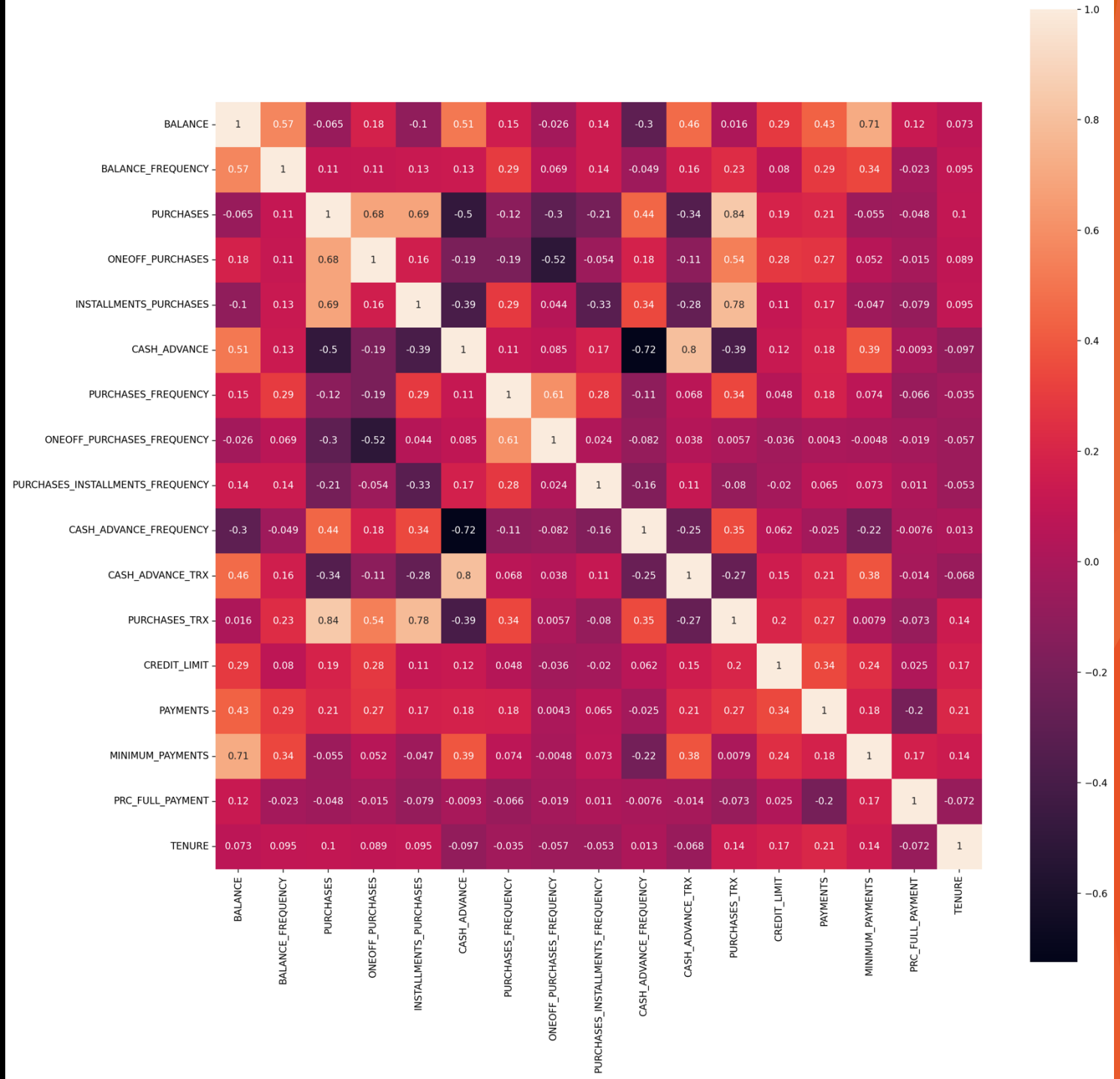
Since the data is unlabeled, it is not easy to decide about the features. But I have run some tests both visual a numeric to find the features with less variance in the dataset.

`TENURE` turned out to have the lowest variance of `0.001` and I dropped it.

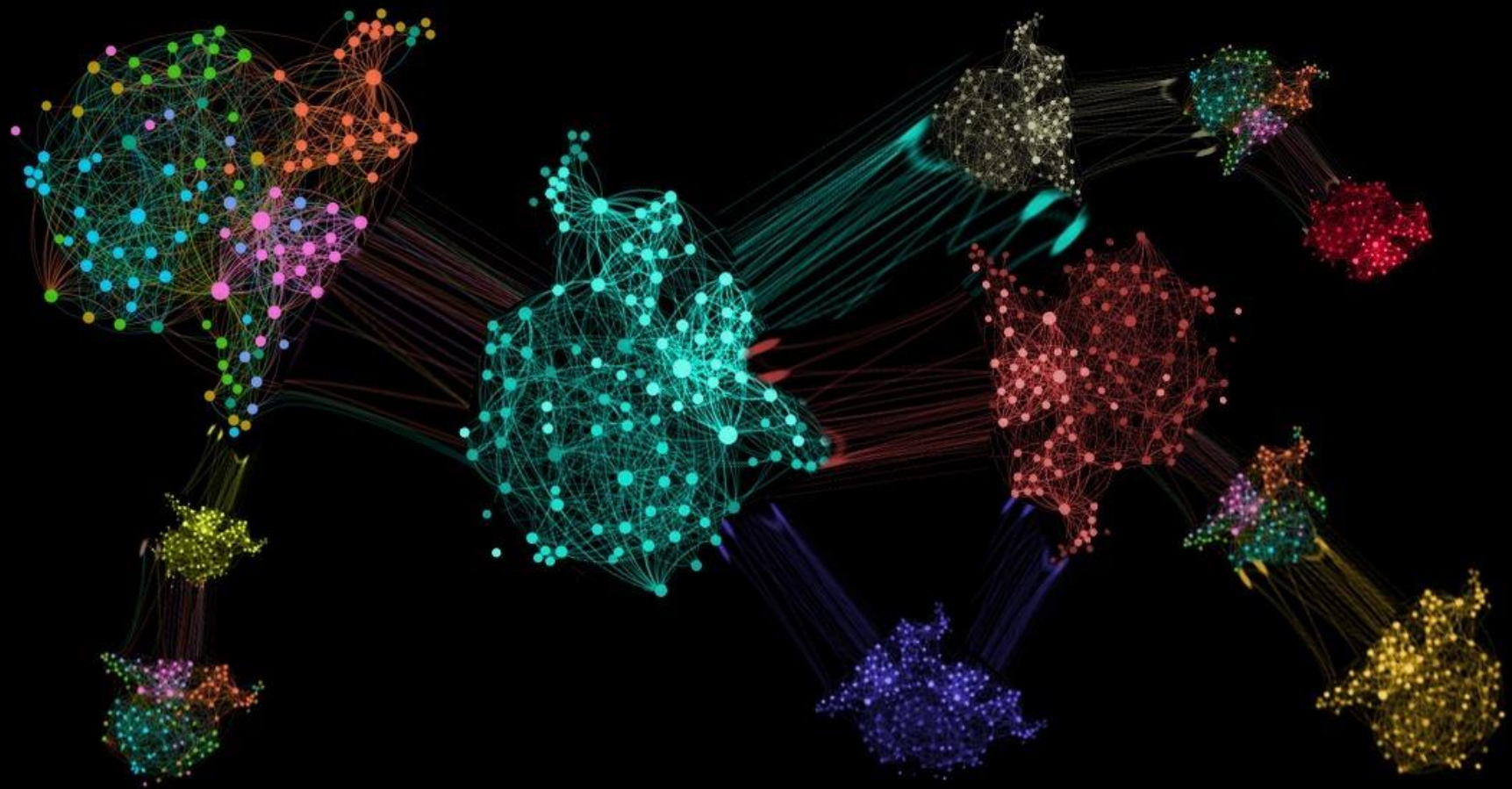


Feature Selection

From this heatmap, it is also obvious that TENURE has a very low variance.

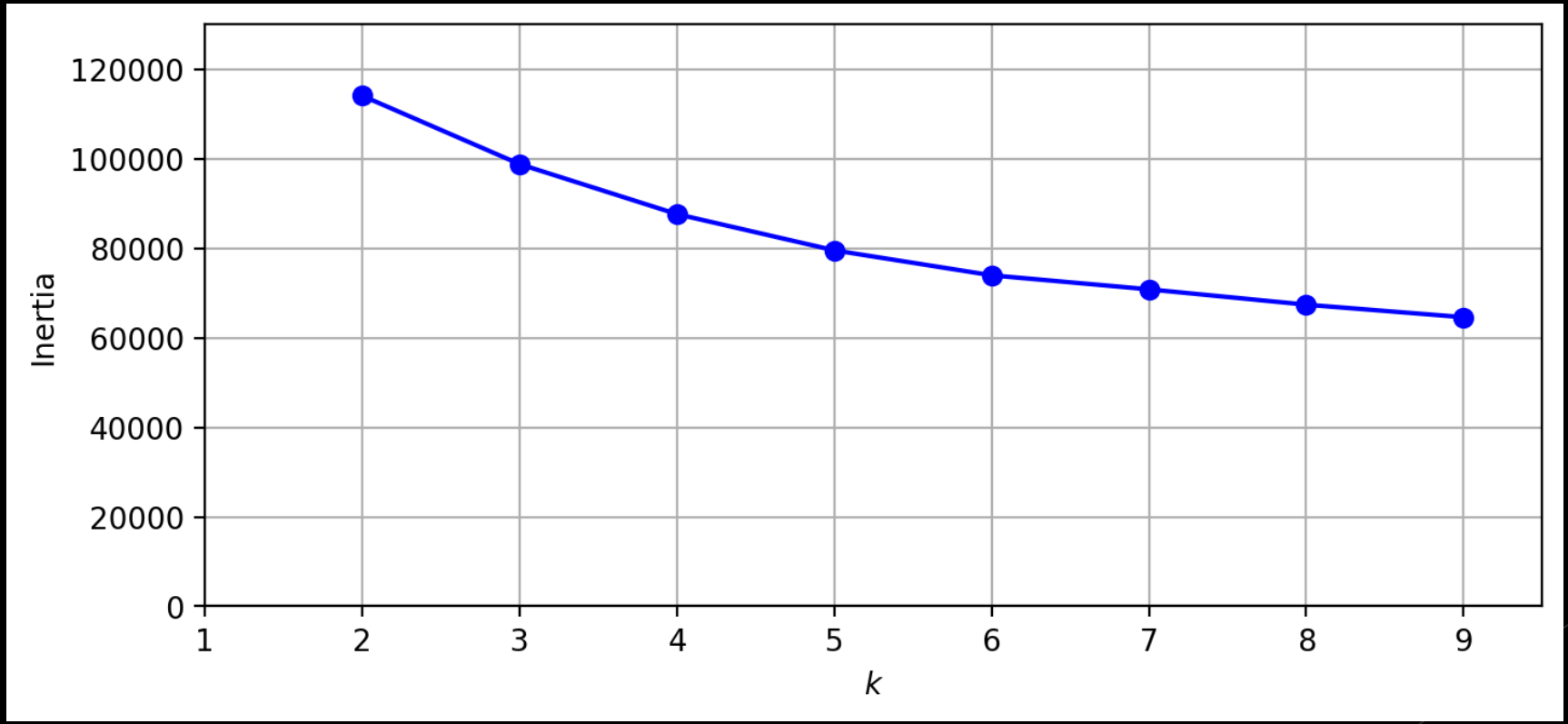


CLUSTERING



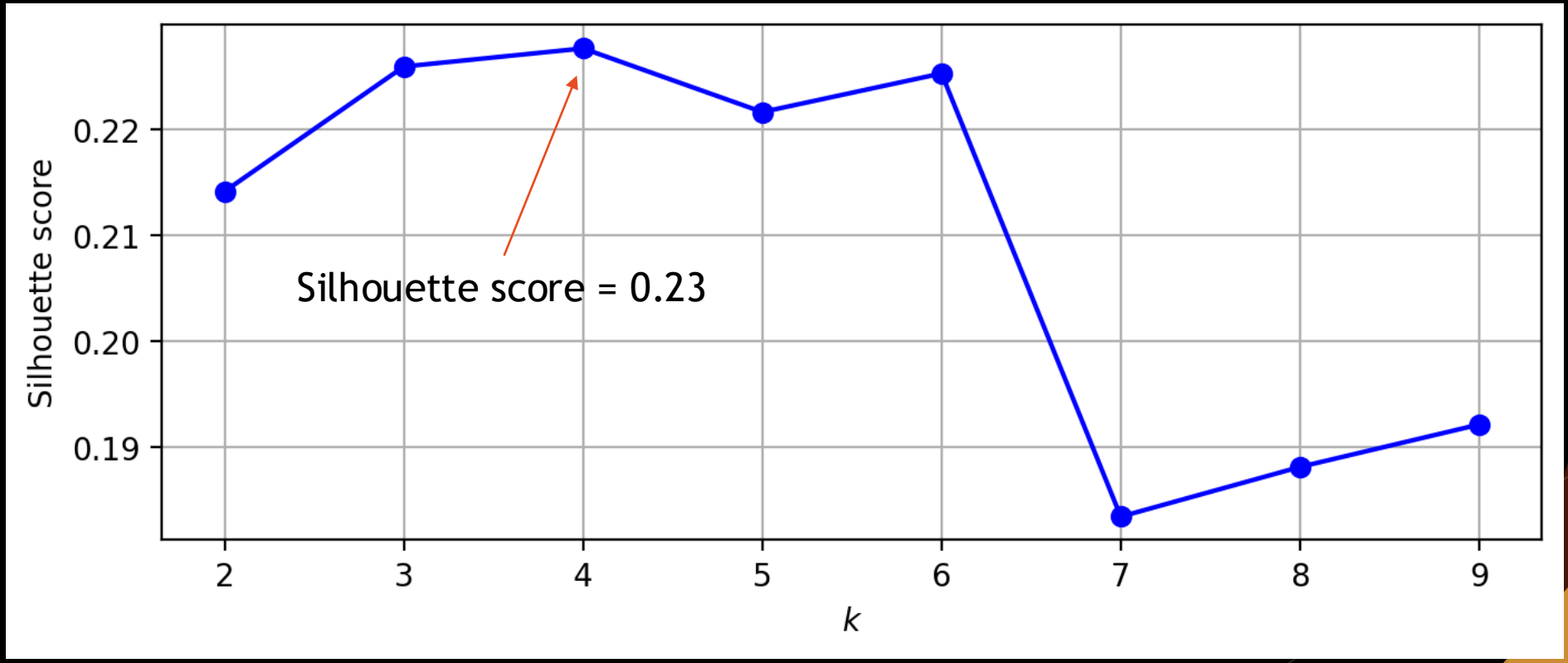
K-Means

An iteration of the k-mean with cluster values of 2 to 9 resulted in the following diagram. However, there can be no elbow detected on the line.

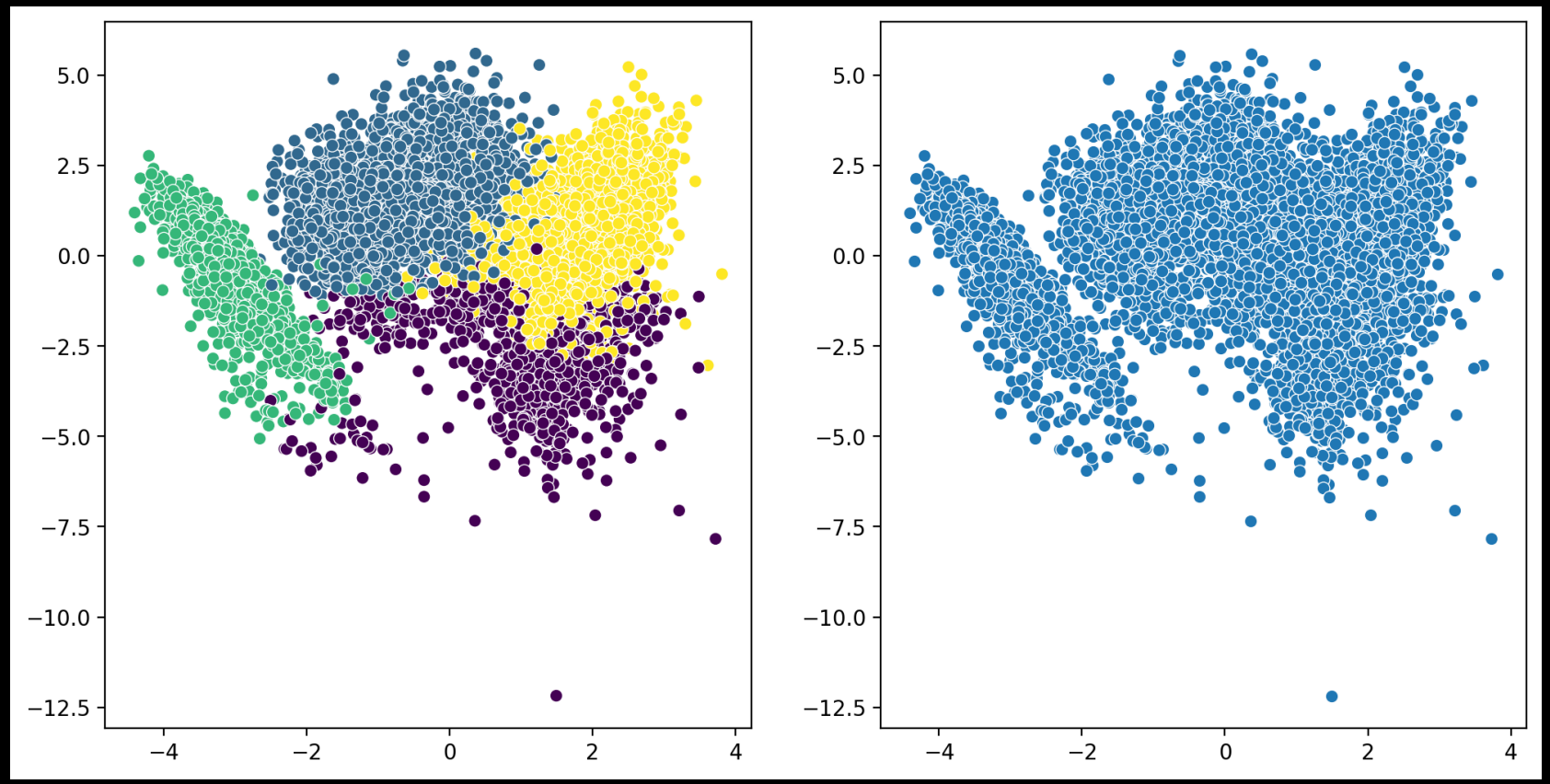


K-mean

Measuring the Silhouette score better represents the most effective number of clusters. According to this diagram, k = 4 and k = 6 are good choices.

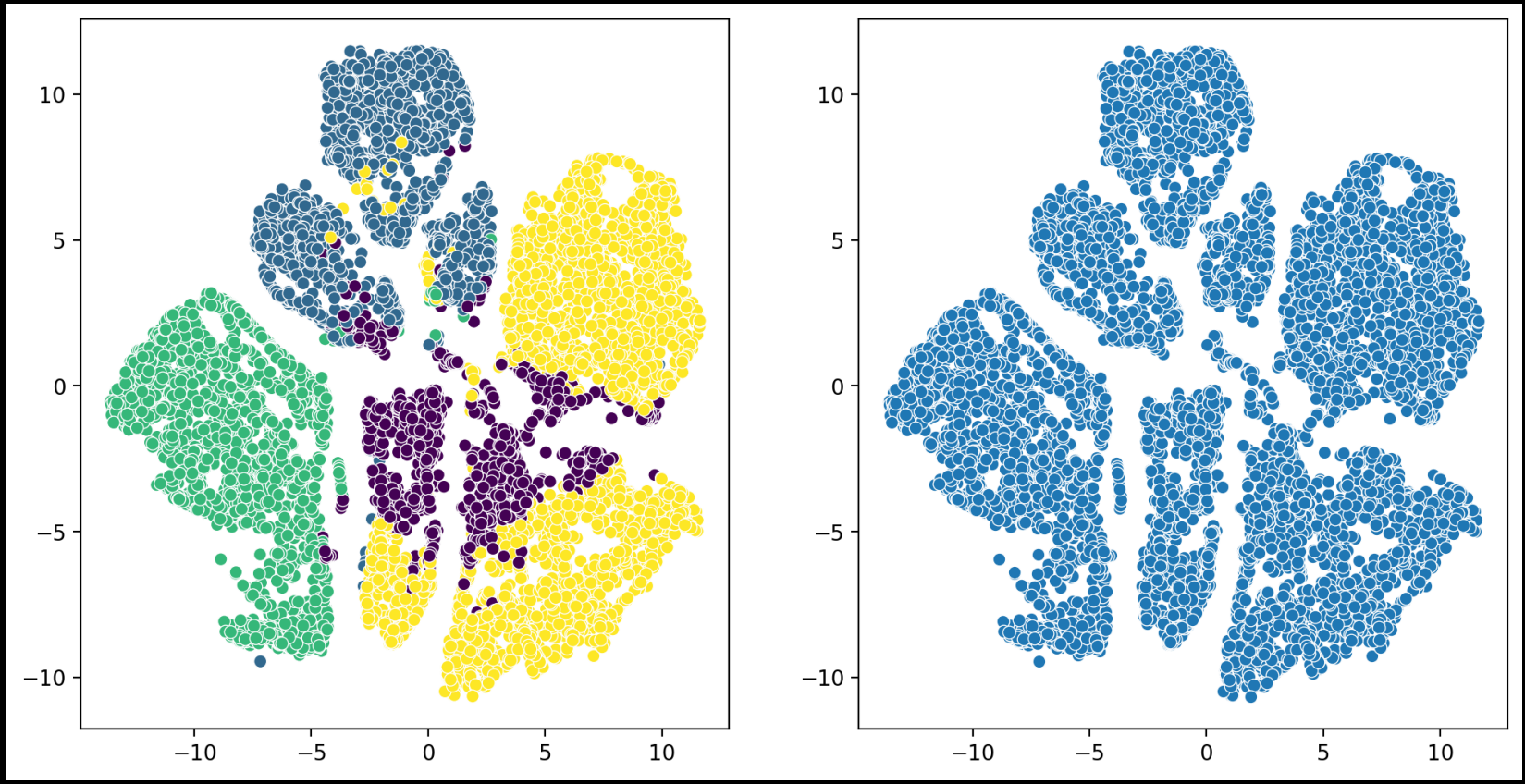


K-mean: PCA reduction



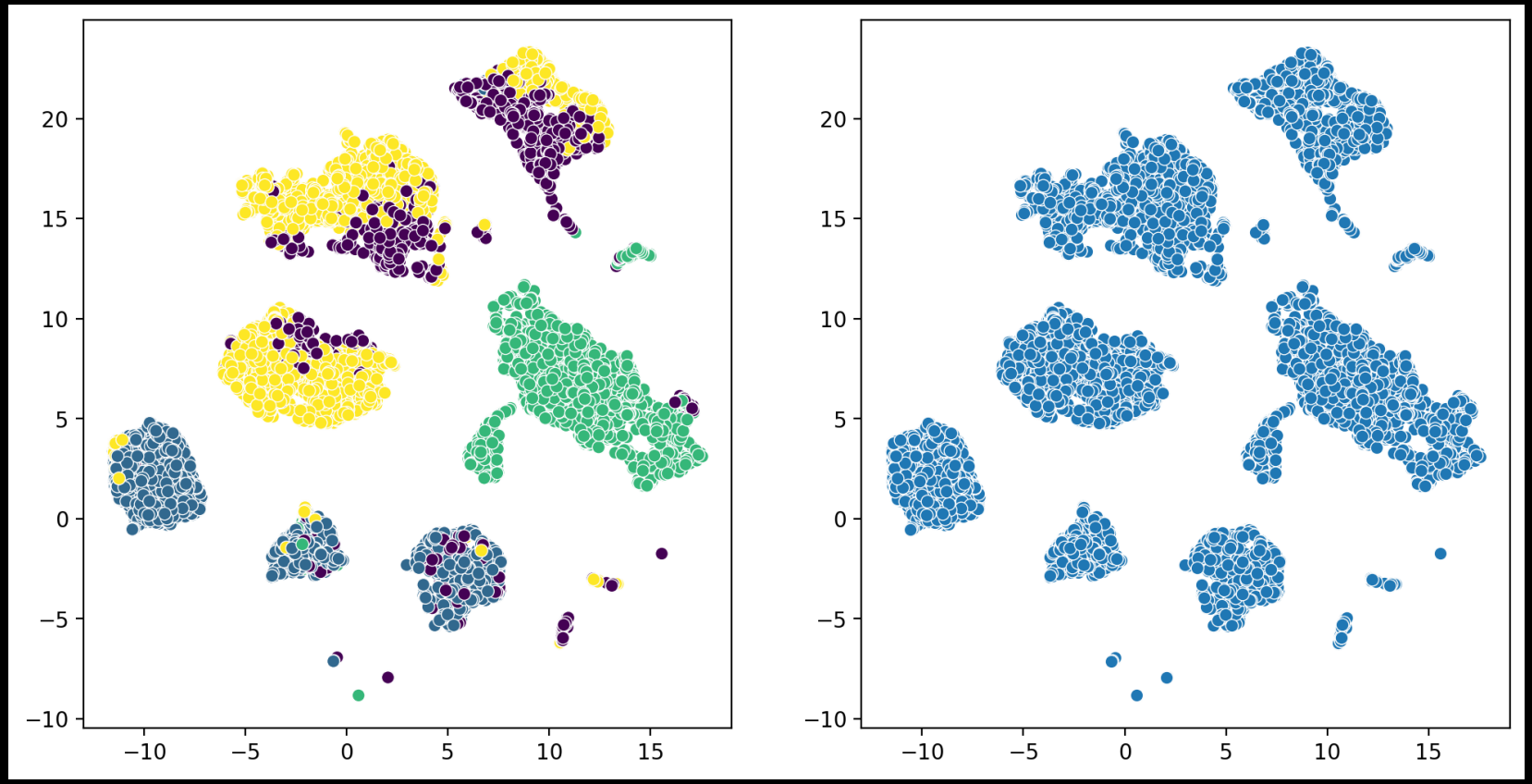
The PCA dimensionality reduction method shows four clusters, but only two of them are distinguishable from the scatter plot.

K-mean: t-SNE reduction



The t-SNE dimensionality reduction method shows 7 clusters, all distinguishable from the scatter plot, and confirmed by the UMAP method.

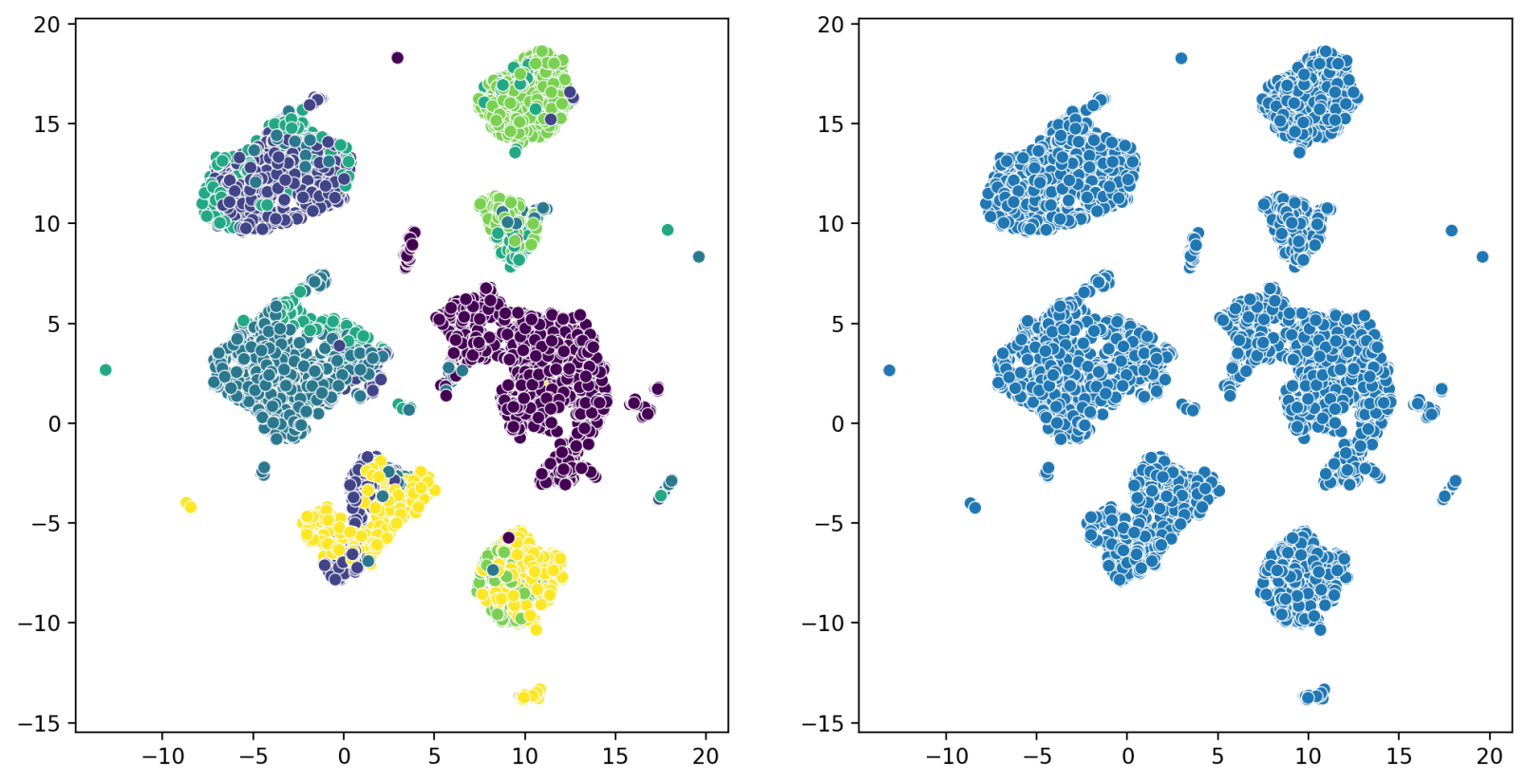
K-mean: UMAP reduction



The UMAP dimensionality reduction method also shows 7 clusters.

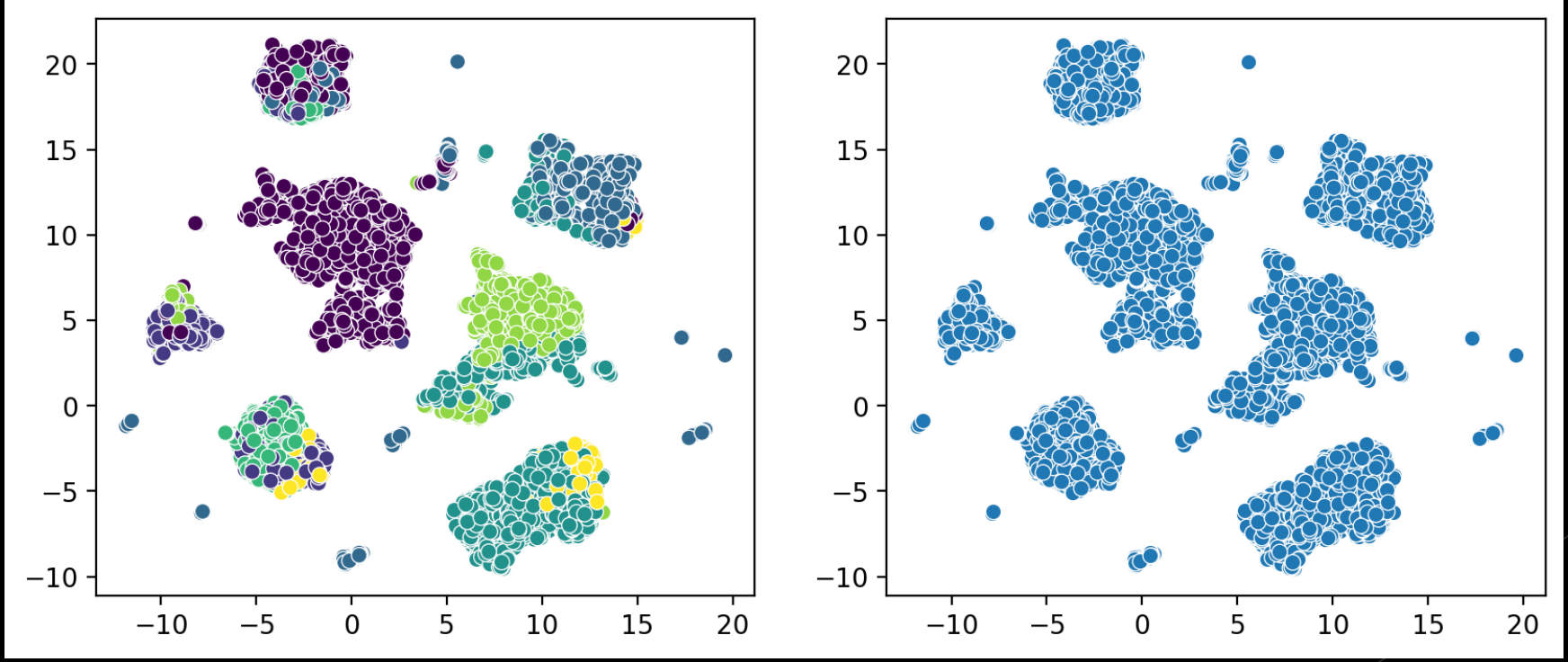
Mini Batch K-means

- The highest Silhouette score of 0.22 is obtained by k = 6 clusters.
- However, the UMAP dimensionality reduction again shows 7 clusters.



Hierarchical Clustering/ Agglomerative Clustering

- In this method, k = 2 gives the highest Silhouette score as 0.19, but the UMAP demonstration obviously shows 7 distinct clusters.
- K = 7 obtains Silhouette = 0.13.

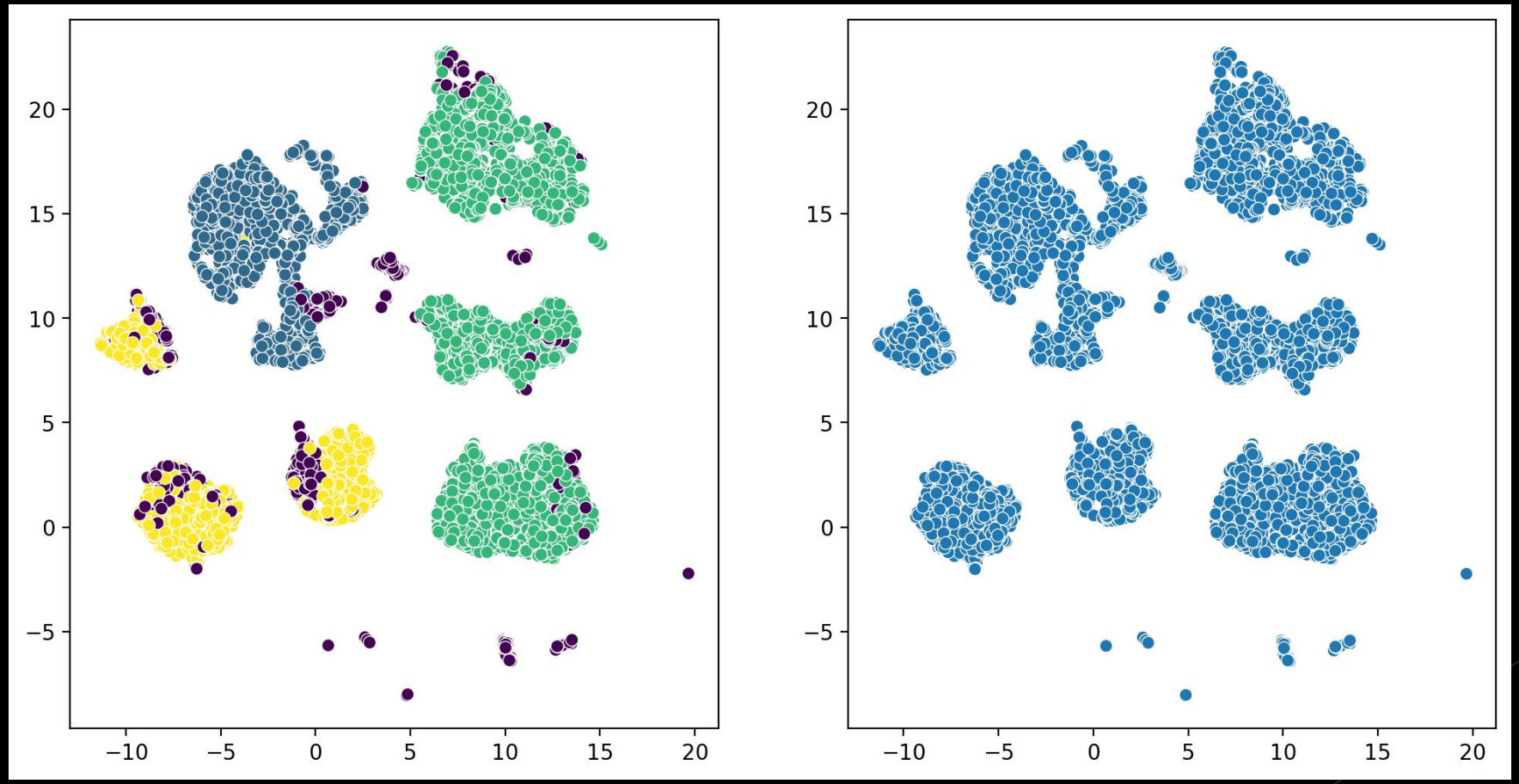


DBSCAN Clustering

- Implemented several times with two hyperparameter values of:
  - eps = 1.0 to 6.5 by 0.5 step and min_sample = 1.
  - min_sample = 2 to 50 and eps = 1.
- The Silhouette scores were calculated for all the iterations. In the first loop, the highest score of 0.62 belongs to eps = 4.5 with 2 clusters. In the second loop, the highest score of -0.05 belongs to min_sample = 5 with 5 clusters. However, when putting these values in the model, only one cluster is distinguishable by the dimensionality reduction methods.

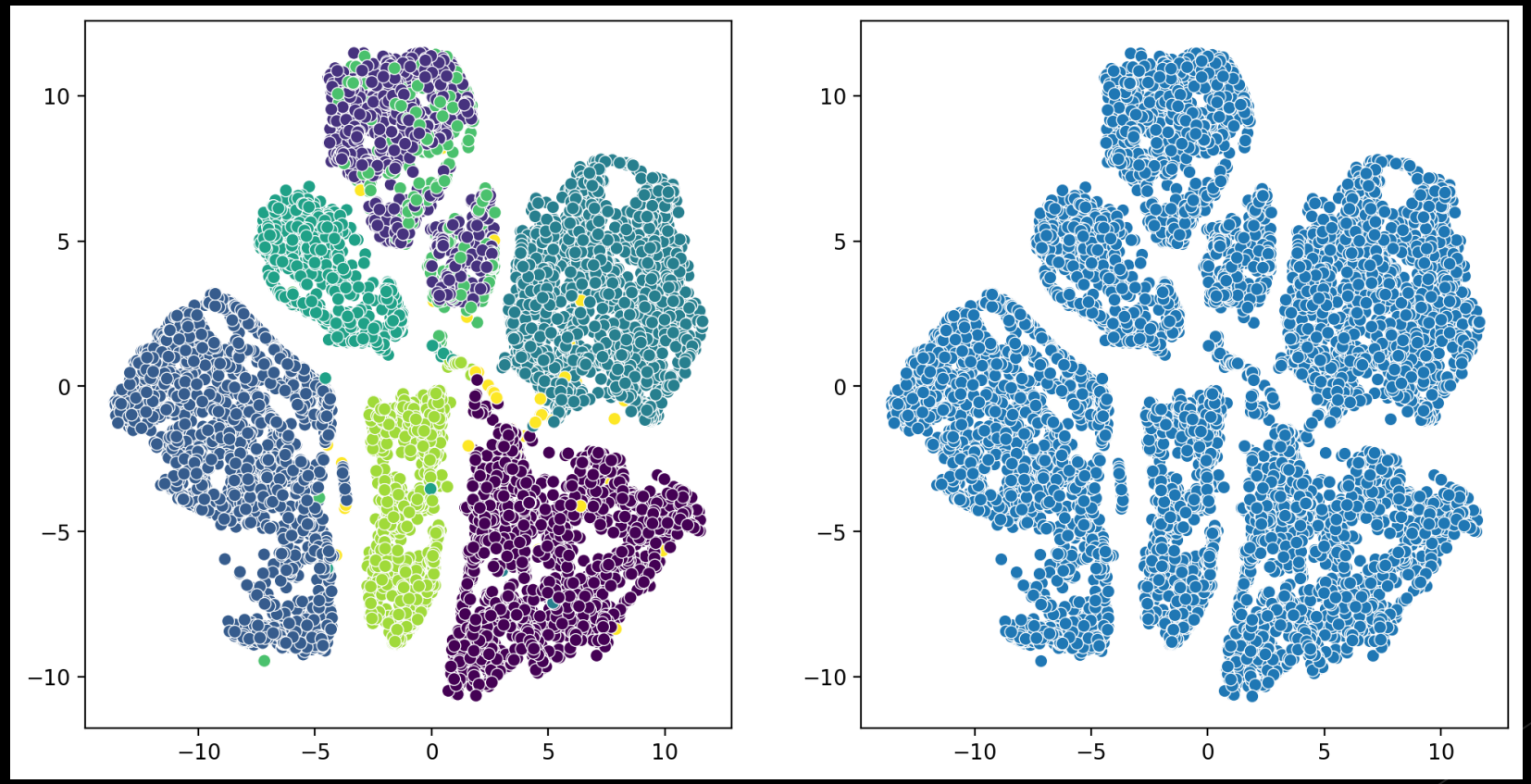
DBSCAN Clustering: UMAP

- Implemented with two hyperparameter values of:
  - eps = 2 and min_sample = 30.



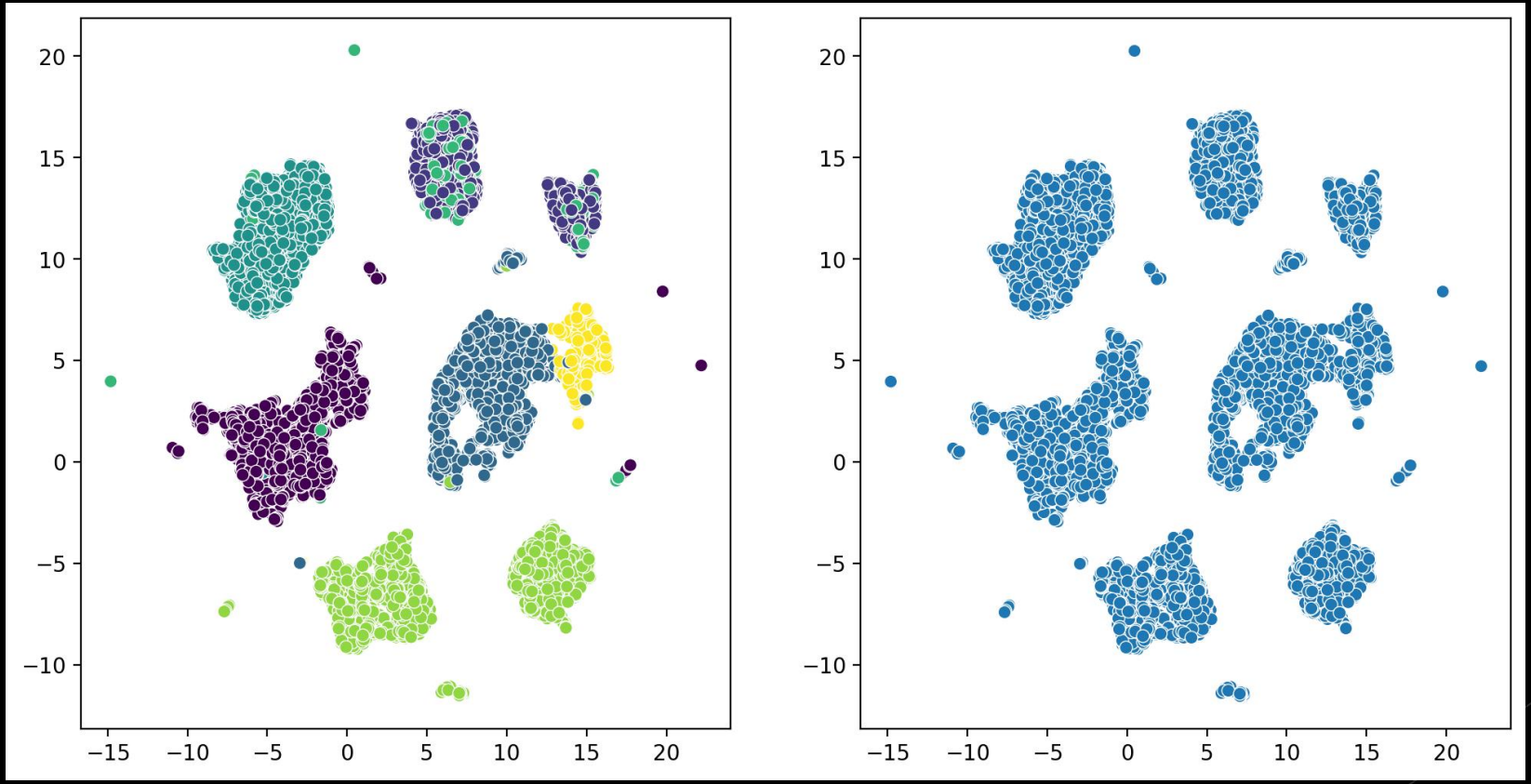
Gaussian Mixture Model

- n_components = 7 gives Silhouette score = 0.18. The t-SNE shows also 7 clusters.



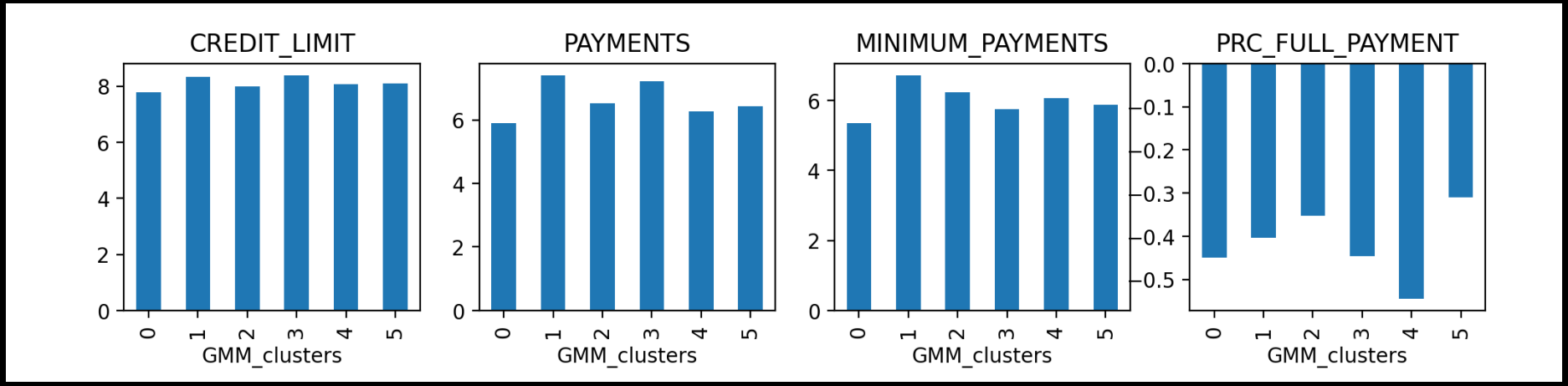
Gaussian Mixture Model

- n_components = 7 gives Silhouette score = 0.18.
- The UMAP also shows also 7 clusters.



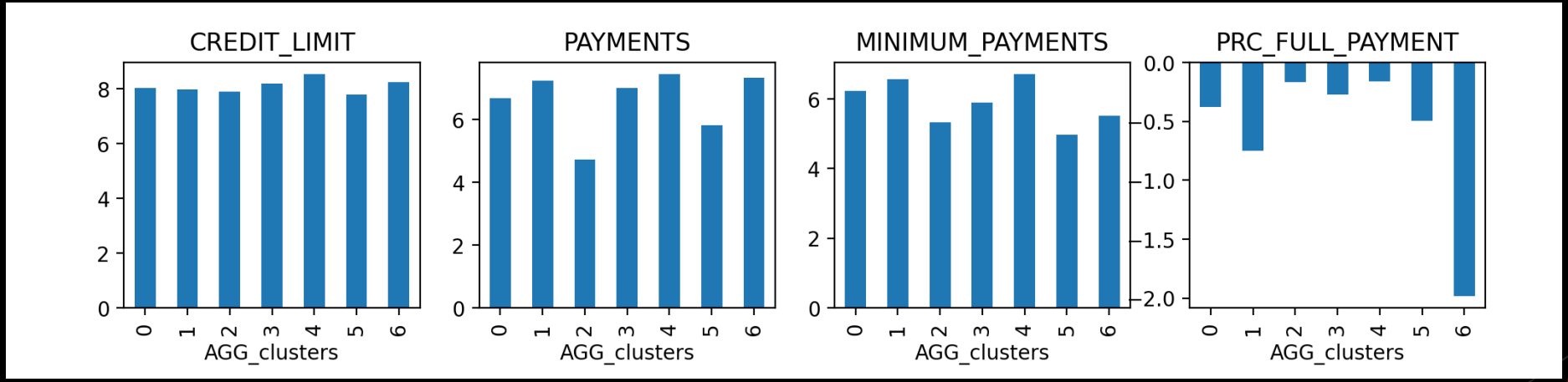
Inference from Clustering: GMM

- By merging the GMM clustering results with the dataset we see that most of the features have close average values in each cluster.



Inference from Clustering: Agglomerative

- By merging the AGG clustering results with the dataset we see that features like Purchases, Credit_limit, Payments, and Minimum_payments, have close average values in each cluster.



CONCLUSION

- According to the results of each clustering method, both numeric and visually demonstrated, it seems that the best-performing models are:
  - Agglomerative Clustering
  - Gaussian Mixture Model
  - DBSCAN