Clustering Analysis Report

Introduction

This report details the results of a clustering analysis performed on customer transaction and signup data. The goal was to segment customers based on their behavior and characteristics. Three clustering algorithms were evaluated: K-Means, Agglomerative Clustering, and DBSCAN.

Data Preparation

Before clustering, the data underwent the following preprocessing steps:

1. **Feature Engineering:**
   - RFM (Recency, Frequency, Monetary) values were calculated from transaction data.
   - Customer tenure was calculated based on signup and transaction dates.
   - Categorical variables ('Region') were one-hot encoded.
2. **Outlier Handling:** Z-scores were calculated for all features and data points with Z-scores greater than 3 were removed to reduce the influence of outliers
3. **Feature Scaling:** The data was standardized using StandardScaler to bring all features to a similar scale.

Clustering Algorithm Evaluation

The following clustering algorithms were tested and evaluated using silhouette score and Davies-Bouldin index.

1. K-Means Clustering

- Number of Clusters: The optimal number of clusters for K-Means was determined by analyzing silhouette scores and distortions for various cluster counts (2-10). A silhouette score plot indicated 4 clusters to be the optimum number of clusters.

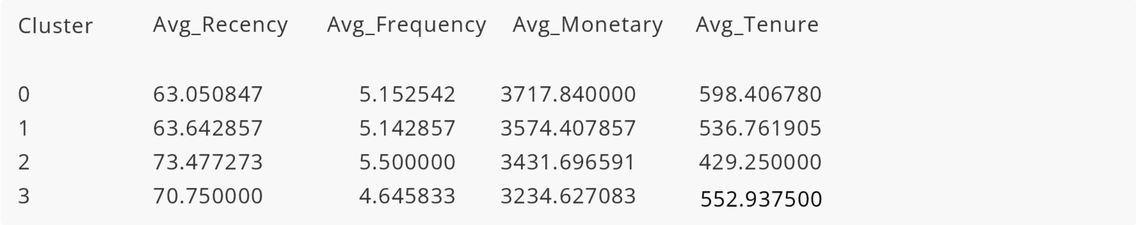
  The silhouette score for each number of clusters is as follows:

  - n_clusters = 2, Silhouette Score = 0.20
  - n_clusters = 3, Silhouette Score = 0.29
  - n_clusters = 4, Silhouette Score = 0.38
  - n_clusters = 5, Silhouette Score = 0.32
  - n_clusters = 6, Silhouette Score = 0.34
  - n_clusters = 7, Silhouette Score = 0.32
  - n_clusters = 8, Silhouette Score = 0.30
  - n_clusters = 9, Silhouette Score = 0.30
  - n_clusters = 10, Silhouette Score = 0.30

  The elbow plot was also used as another method to determine the optimal number of clusters.

- Final Number of Clusters: Based on the silhouette scores, 4 clusters were selected.
- DB Index: 1.15
- Silhouette Score: 0.38

- **Cluster Summary:** The following table shows average values of the RFM and Tenure for each cluster.



Cluster	Avg_Recency	Avg_Frequency	Avg_Monetary	Avg_Tenure
0	63.050847	5.152542	3717.840000	598.406780
1	63.642857	5.142857	3574.407857	536.761905
2	73.477273	5.500000	3431.696591	429.250000
3	70.750000	4.645833	3234.627083	552.937500

2. Agglomerative Clustering

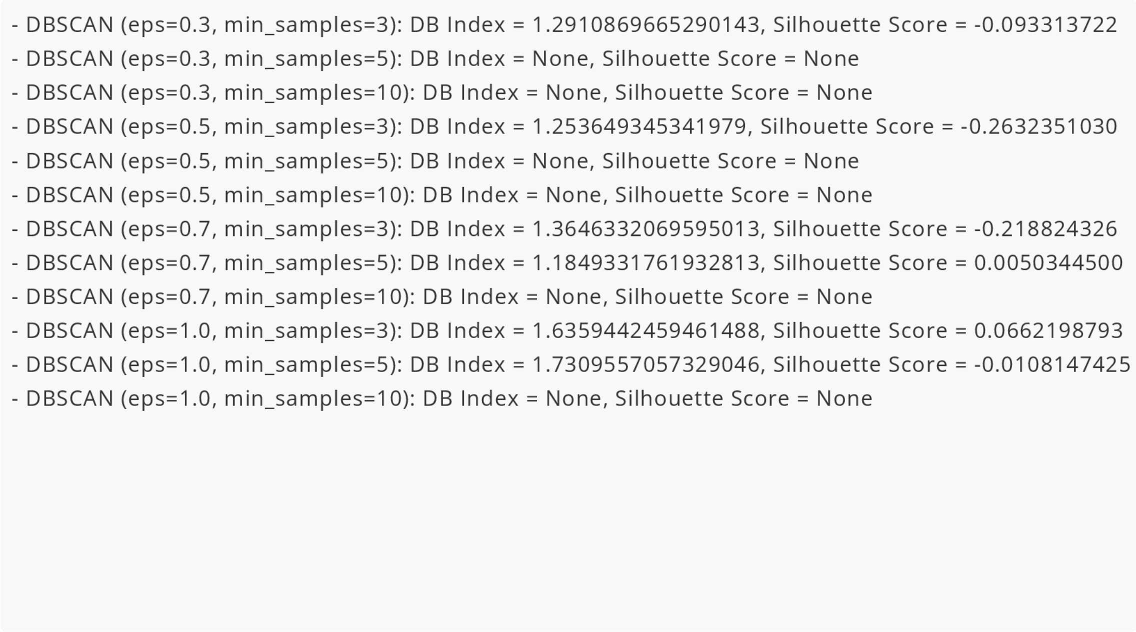
- Number of Clusters: 4 clusters (same as K-Means for comparison)
- DB Index: 1.24
- Silhouette Score: 0.37

3. DBSCAN Clustering

- **Parameter Tuning:** The algorithm was tested with different combinations of eps (neighborhood radius) and min_samples (minimum number of points required to form a dense region) values. The optimal parameters were chosen based on the lowest DB Index and a non-None Silhouette score.

- Optimal Parameters: eps=0.7, min_samples=5.
- DB Index: 1.18
- Silhouette Score: 0.005
- ***Note:*** *Several parameter combinations produced only a single cluster or noisy data, resulting in invalid DB Index and Silhouette score. DBSCAN was not chosen as the optimal model, as its Silhouette score and DB Index was significantly worse than those of the other two models.*

The silhouette scores and DB indices for different parameters for DBSCAN are as follows:



```
- DBSCAN (eps=0.3, min_samples=3): DB Index = 1.2910869665290143, Silhouette Score = -0.093313722
- DBSCAN (eps=0.3, min_samples=5): DB Index = None, Silhouette Score = None
- DBSCAN (eps=0.3, min_samples=10): DB Index = None, Silhouette Score = None
- DBSCAN (eps=0.5, min_samples=3): DB Index = 1.253649345341979, Silhouette Score = -0.2632351030
- DBSCAN (eps=0.5, min_samples=5): DB Index = None, Silhouette Score = None
- DBSCAN (eps=0.5, min_samples=10): DB Index = None, Silhouette Score = None
- DBSCAN (eps=0.7, min_samples=3): DB Index = 1.3646332069595013, Silhouette Score = -0.218824326
- DBSCAN (eps=0.7, min_samples=5): DB Index = 1.1849331761932813, Silhouette Score = 0.0050344500
- DBSCAN (eps=0.7, min_samples=10): DB Index = None, Silhouette Score = None
- DBSCAN (eps=1.0, min_samples=3): DB Index = 1.6359442459461488, Silhouette Score = 0.0662198793
- DBSCAN (eps=1.0, min_samples=5): DB Index = 1.7309557057329046, Silhouette Score = -0.0108147425
- DBSCAN (eps=1.0, min_samples=10): DB Index = None, Silhouette Score = None
```

Visualization

t-SNE (t-distributed Stochastic Neighbor Embedding) was used to reduce the dimensionality of the scaled features for visualization purposes. The 2D t-SNE representations were then plotted, colored by cluster assignments from each of the three clustering algorithms.

Conclusion

Based on this analysis, K-Means clustering provides a slightly better segmentation based on Silhouette scores and Davies-Bouldin indices. Agglomerative clustering performance was very close to K-means, but was ultimately not selected for this analysis. DBSCAN's low Silhouette score and high DB Index shows that it is not an adequate algorithm to use in this instance.