Customer Segmentation Analysis on An Indian Bank

Prepared by:
Shuai Tan
608-692-9784
stan67@wisc.edu
May 14, 2023

Objective

The mission of the Indian bank is to provide comprehensive financial solutions and exceptional customer service to individuals and businesses, fostering long-term relationships based on trust and mutual success.

Customer segmentation is one of the most significant ways to align with this mission by enabling the bank to deliver targeted marketing, customized financial products, and personalized experiences that meet the specific needs of different customer segments.

The objective of customer segmentation for the Indian bank is to divide the customer base into distinct groups or segments based on various criteria such as demographics, transaction behavior, financial needs, and preferences. This segmentation aims to gain a deeper understanding of the bank's customer base, identify different customer segments with unique characteristics and behaviors, and tailor marketing strategies and product offerings to better meet the specific needs and preferences of each segment. The ultimate goal is to enhance customer satisfaction, improve customer retention, and drive business growth by effectively targeting and serving different customer segments with customized solutions and experiences.

In this analysis, the Indian bank will utilize two machine learning methods: K-Means and Agglomerative Clustering due to their simplicity, scalability, and interpretability. These methods are effective in identifying clusters and patterns within large datasets, allowing the bank to gain insights into customer segments. Additionally, both algorithms provide clear cluster assignments and can handle a variety of data types, making them suitable for analyzing diverse customer attributes.

Data Description

This analysis is based on the dataset titled "Bank Customer Segmentation" available on Kaggle, provided by Shivam Bansal. The dataset offers valuable insights into customer demographics, financial characteristics, and behavior. It consists of over 1 million transaction records by more than 800 thousand customers from an anonymous bank in India. The data spans approximately three months, from August 1st to October 21st, in 2016. Each transaction record includes customer information such as gender, location, date of birth, bank account balance, and transaction details like amount and date.

To ensure data quality, null values were removed from the dataset. Rows with inaccurate information regarding birth dates, specifically those indicating birth in 1800, were dropped. Additionally, undefined gender information denoted as 'T' was excluded from the analysis. Then, the gender column was transformed into a binary variable, where a value of 1 indicates a male customer and 0 indicates a female customer.

Next, the RFM (Recency, Frequency, Monetary) model was employed to gain insights into customer behavior and segment them based on transaction patterns. The RFM model is a widely utilized customer segmentation technique that focuses on three key factors: recency, which measures the number of days since the last transaction; frequency, which represents the average number of transactions within a specific period; and monetary, which reflects the total transaction amount during a specific timeframe. The data was grouped by customer ID, enabling the calculation of RFM values for each customer by tracking their transaction history.

Subsequently, columns related to location and transaction time were removed since I lacked precise geographical data and deemed transaction time as non-essential for this analysis. The date of birth column was transformed into customer age by subtracting the birth year from 2016. Finally, as the data was now aggregated by customer rather than at the transaction level, the transaction ID and customer ID columns were eliminated, and the remaining variables were adjusted accordingly.

Table 1 displays the summary statistics for the final clean dataset.

Table 1: Summary Statistics

	Count	Mean	Std. Dev.	Min	Max
Frequency	838,561	1.174	0.435	1	6
CustGender	838,561	0.723	0.448	0	1
CustAccountBalance	838,561	106,149.6	833,166.1	0	115,035,500
Monetary	838,561	1,706.6	6,689.6	0.01	1,560,035
CustomerAge	838,561	31.03	8.75	0	98
Recency	838,561	55.41	15.22	0	81

Figure 1 illustrates the customer age distribution, indicating most customers fall between 20 and 40, with a peak at 26. ***Figure 2*** provides insights into the customer gender distribution. It reveals that the number of male customers is almost three times higher than that of female customers, indicating a dominance of male customers. ***Figure 3*** and ***Figure 4*** display the frequency histogram and the recency boxplot, respectively. During the three months, it appears that nearly all of the bank customers did not engage in regular transactions with their bank for more than once. ***Figure 5*** presents the relationship between monetary value and recency, with bubble size indicating the frequency. For the majority of customers, the monetary values remain low. Most of the bank customers were low-income customers creating bank accounts for depositing money.

Figure 1: Customer Age Distribution

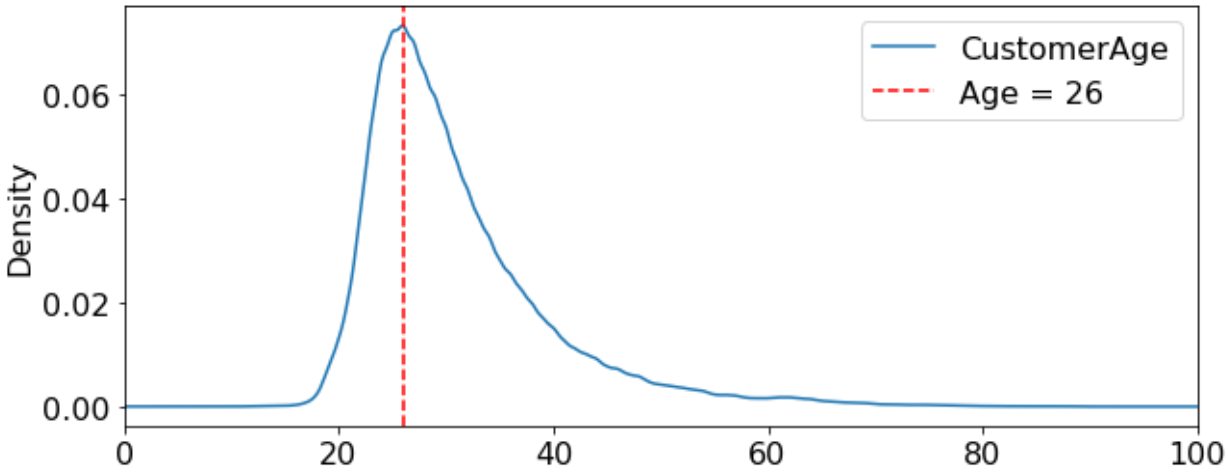


Figure 2: Customer Gender Distribution

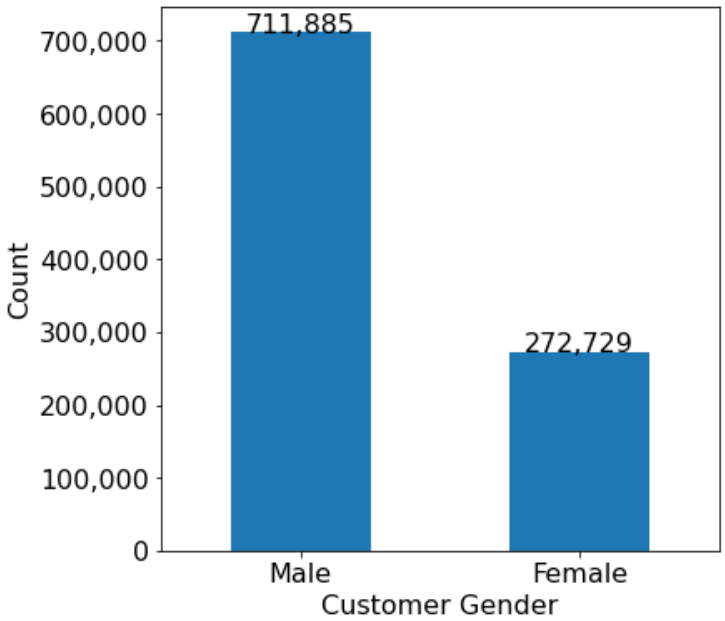


Figure 3: Transaction Frequency Histogram

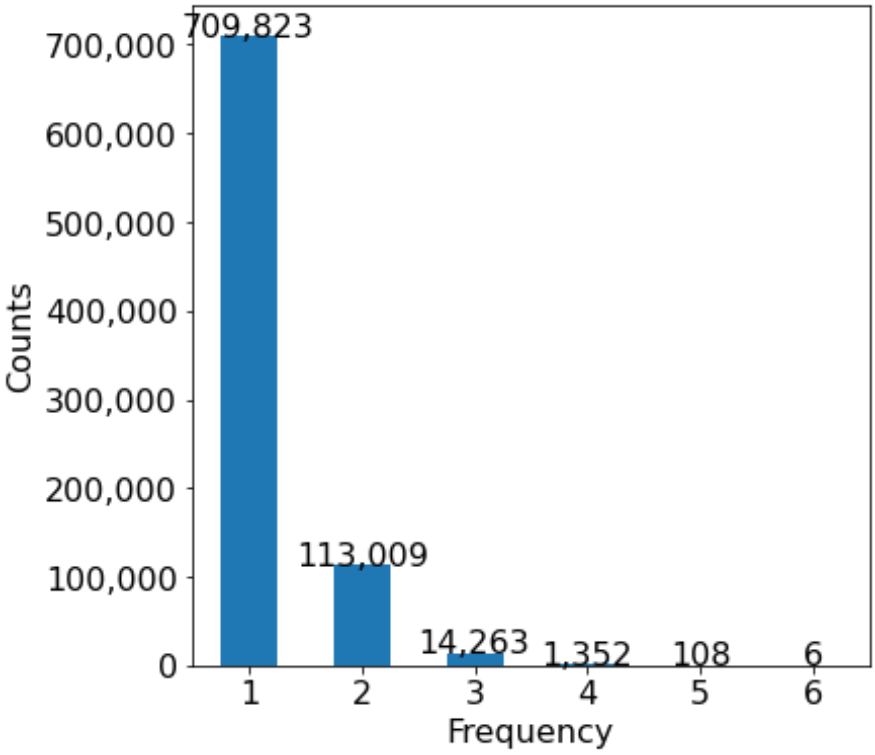


Figure 4: Transaction Recency Boxplot

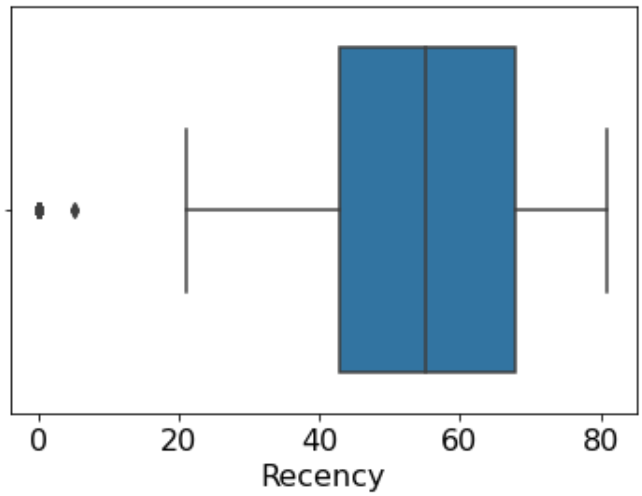
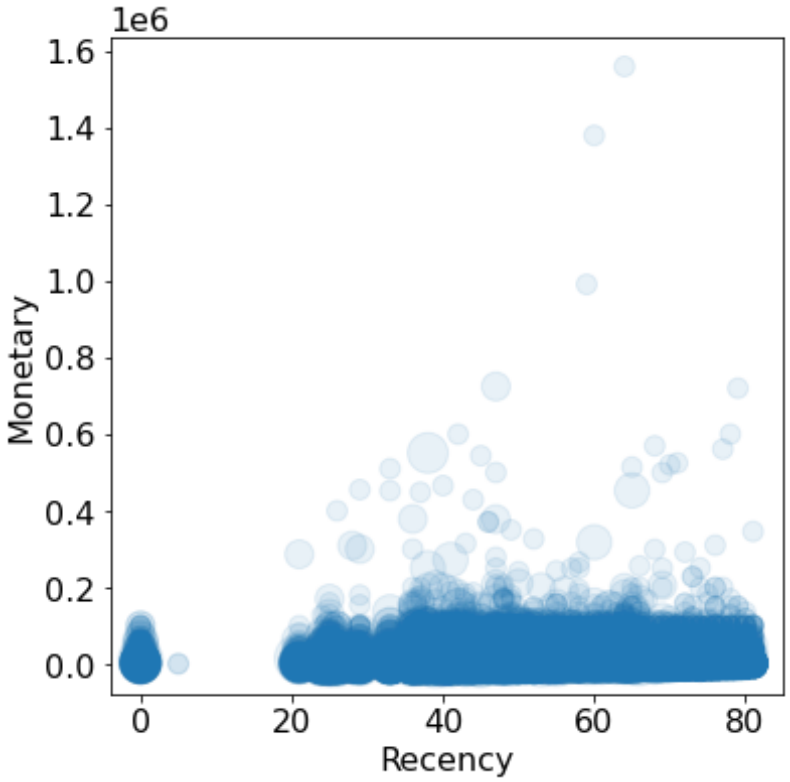


Figure 5: Monetary Value VS Recency



Methodology

1. K-means Clustering

K-means clustering is a widely employed machine learning technique used to partition datasets into distinct groups. In this method, the parameter K represents the desired number of clusters, while "means" refers to the statistical concept of arithmetic means, which are used to define the cluster centroids. These centroids act as representatives of their respective clusters and are not necessarily data points themselves.

Given the dataset X consisting of N (N=838,561) data points, each represented by a 6-dimensional feature vector x_i, where i = 1, 2, 3, ..., N, and a predefined value for K, the K-means clustering algorithm operates as follows:

- **Initialization:** Start by randomly generating K centroids $C = \{c_1, c_2, \ldots, c_k\}$ in the 6-dimensional feature space. Each centroid represents a distinct cluster.

- **Assignment:** For each data point x_i, calculate the Euclidean distance to each centroid c_j, and assign the data point to the cluster with the closest centroid. This is determined by the equation:

  $arg\ min_j\ ||x_i - c_j||^2$, where $||.||$ denotes the Euclidean distance.

- **Recalculation:** After the assignment step, recompute the centroids for each cluster based on the data points assigned to them. Mathematically, for each cluster j, calculate the new centroid c_j as the mean of the assigned data points: $c_j = \frac{1}{|S_j|} \sum x_i$, where S_j represents the set of data points assigned to cluster j.

- **Convergence:** Iterate the assignment and recalculation steps until the centroids become stable and cease to move significantly. It's important to note that the convergence may be influenced by the initial centroid placement and can sometimes lead to biased results.

By following these steps, the K-means algorithm partitions the dataset into K distinct clusters, with each data point assigned to the cluster whose centroid it is closest to.

2. Agglomerative Clustering

Agglomerative clustering is a machine learning algorithm used to identify hierarchical clustering patterns within a dataset.

The algorithm initially treats each data point as an individual cluster. It then iteratively merges the two closest clusters based on a distance metric, such as the Euclidean distance

in my analysis. This merging process continues until either all the clusters are combined into a single large cluster or the predefined number of clusters is reached.

Throughout the process, the algorithm merges clusters and maintains a set of information called linkage, which captures the relationships between the merged clusters. Using this linkage information, the agglomerative clustering method constructs a dendrogram. A dendrogram is a tree-like graph where each node has exactly two children. It illustrates the order and distances at which clusters are merged. The length of the branches in the dendrogram represents the distance calculated by the chosen metric. Longer branches indicate greater dissimilarity between the merged clusters. By cutting the branches at various horizontal lines, different numbers of clusters can be obtained.

3. Elbow Method

The elbow method is a widely employed technique used to determine the optimal number of clusters in the K-means clustering algorithm. It operates on the principle that as the number of clusters increases, the sum of squared distances, also known as inertia, decreases.

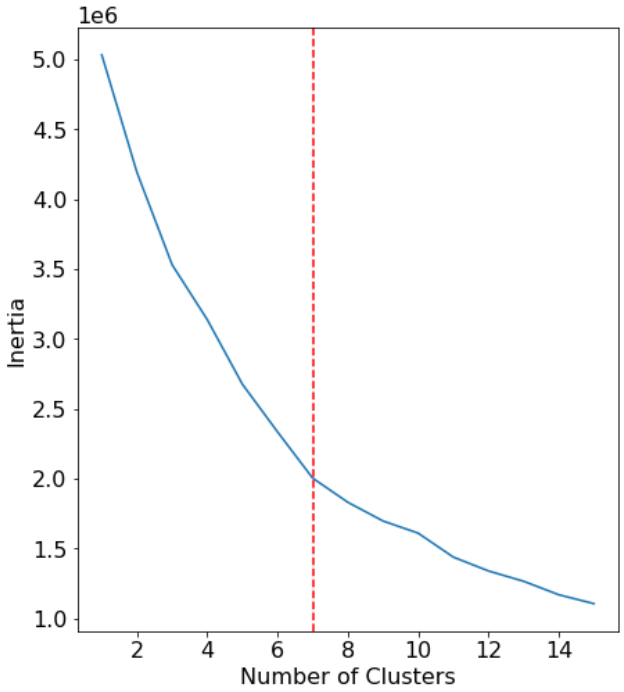
To apply the elbow method, a range of cluster numbers is selected, from 1 to 15 in my analysis. For each K-means clustering iteration, the inertia is computed. A line graph is then generated, plotting the inertia against the number of clusters. The elbow point, characterized by a sharp decline in the graph, is identified. This point corresponds to the optimal number of clusters, providing a trade-off between minimizing inertia and avoiding excessive cluster fragmentation.

Results

Before proceeding with the segmentation analysis, it was necessary to preprocess the dataset by applying standardization. This step was crucial due to the varying magnitudes of the variables of interest. Without standardization, variables with large values, such as bank account balance, would exert a dominant influence on the output compared to binary variables like gender. This could potentially introduce bias and lead to incorrect groupings.

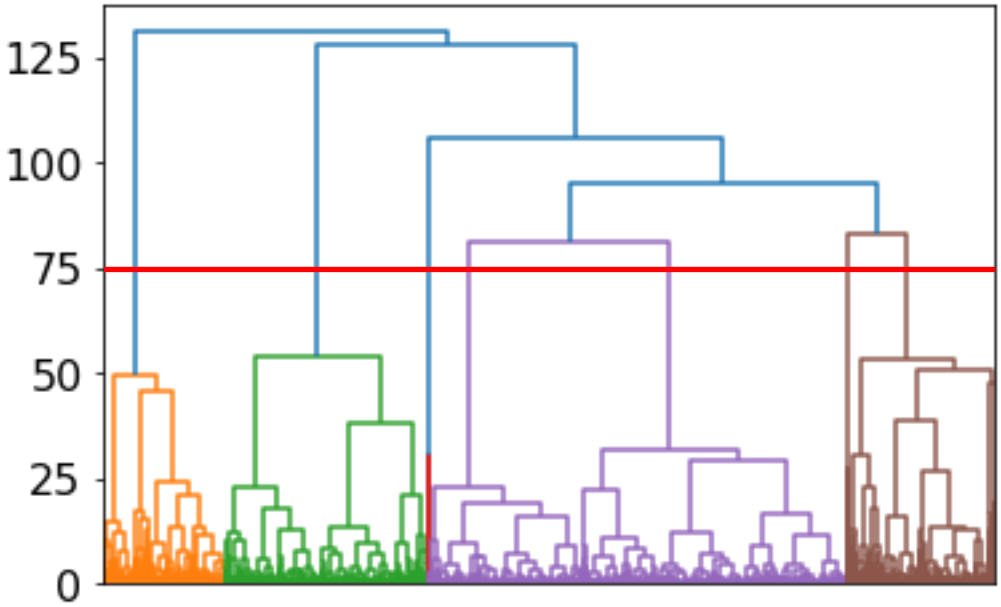
After standardization, I utilized the elbow method to determine the optimal number of clusters for the K-means clustering model within a range of 1 to 15 clusters. ***Figure 6*** showcases the plot generated by the elbow method, indicating that the optimal number of clusters falls within the range of 7 to 10.

Figure 6: Elbow Method



To obtain a more precise and less biased estimation, I also applied agglomerative clustering and generated the dendrogram shown in ***Figure 7***. Due to limitations in computational resources, I randomly selected 10,000 data points for this clustering analysis. By setting a horizontal line at 75, where the branches are relatively sparse but exhibit substantial lengths, I identified 7 distinct clusters.

Figure 7: Agglomerative Clustering Dendrogram

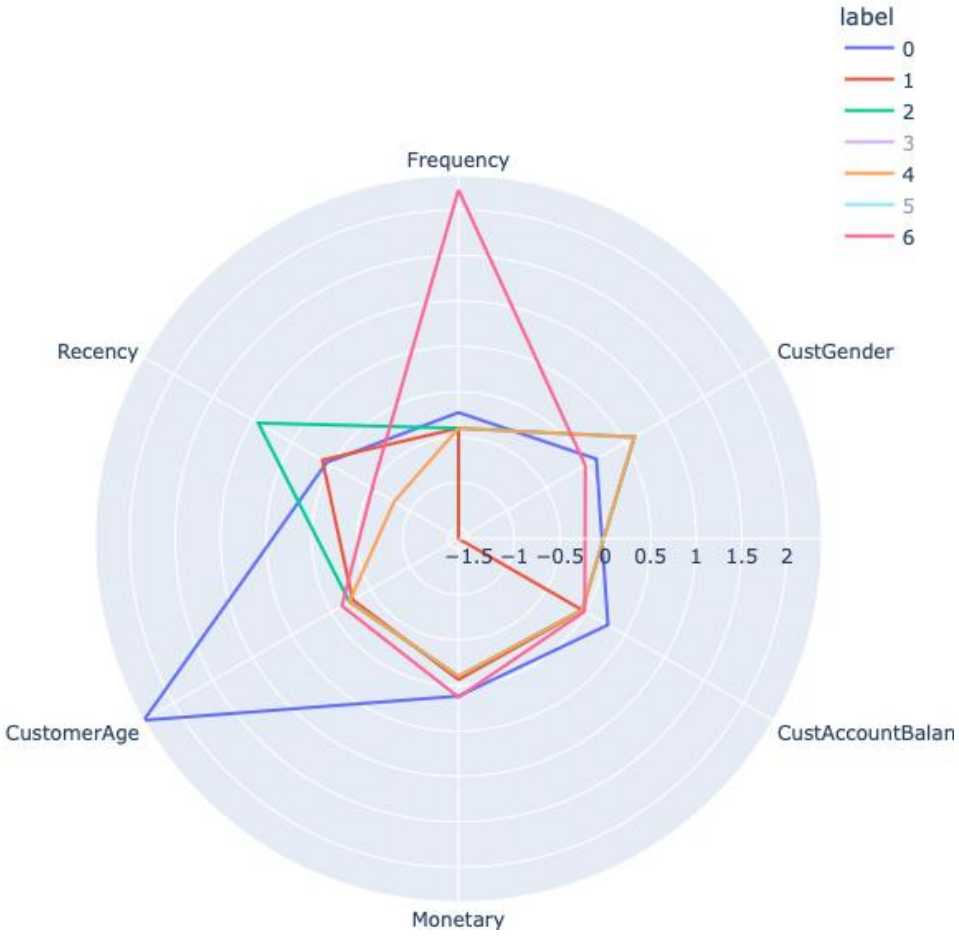


The decision to choose 7 clusters for the customer segmentation analysis was supported by both the results of the K-means clustering and the agglomerative clustering approaches. ***Table 2*** presents the clustering results. Notably, Cluster 3 and Cluster 5 comprise a relatively small percentage of the customer population compared to the other clusters. Furthermore, ***Figure 8*** visually represents the characteristics of each group excluding Cluster 3 and Cluster 5 in a 2-dimensional space.

Table 2: Clustering

Label	Cluster 2	Cluster 4	Cluster 1	Cluster 6	Cluster 0	Cluster 5	Cluster 3
Count	235,782	224,744	181,287	122,568	70,932	3,115	133

Figure 8: Cluster 0, 1, 2, 4, 6



Since the majority of our customers are distributed across Clusters 0, 1, 2, 4, and 6, which will be the primary focus of our analysis. Specifically:

1. Clusters 1, 2, and 4 exhibit similar characteristics, with a frequency value of 1, differing mainly in terms of gender. Consequently, these clusters can be grouped together for the time being.

2. Cluster 6 demonstrates a higher frequency of transactions with the bank compared to Clusters 1, 2, and 4, while sharing similar characteristics.

3. Cluster 0 consists of an older and more affluent population.

Suggestion

Based on the customer segmentation results, the context of banking in India and particularly considering the low account balance of nearly all customers in this bank, I would like to propose the following marketing strategies for the Indian bank:

General suggestions: All these customers are with low bank account balance.

- **Basic Savings Account:** Offer a no-frills savings account with low minimum balance requirements to encourage financial inclusion and cater to the banking needs of customers with low income balance.

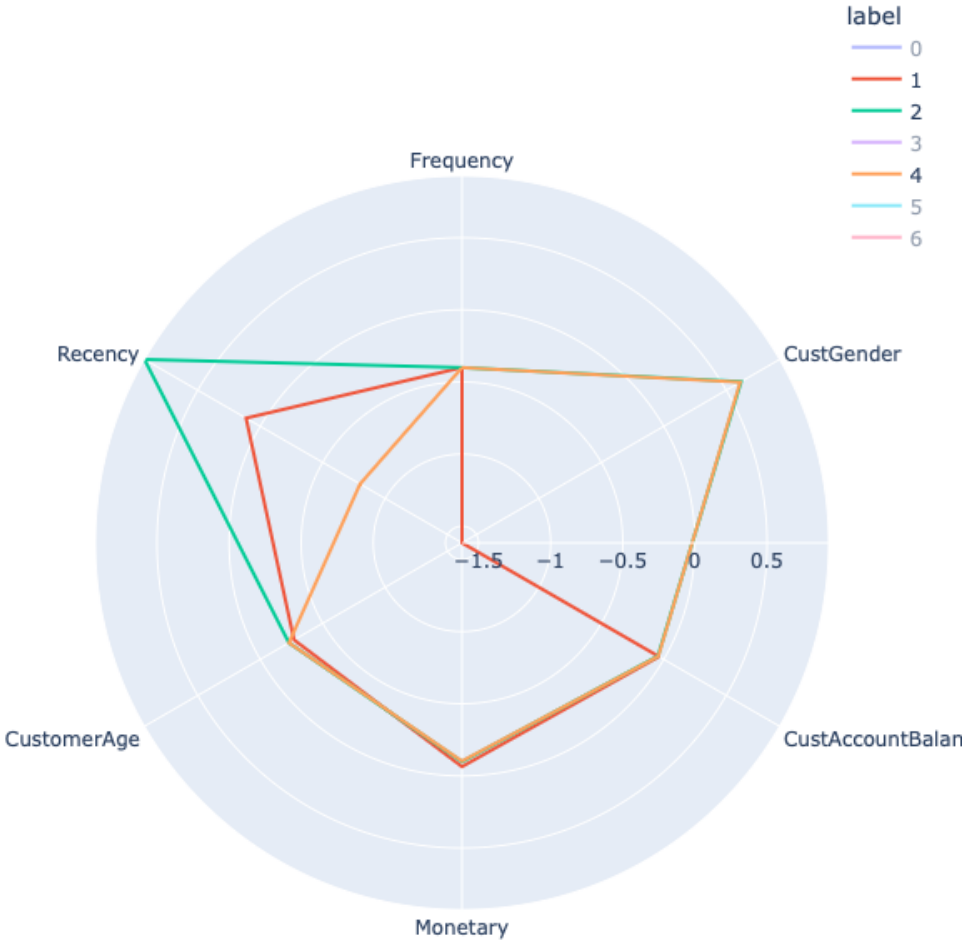
- **Micro Loans:** Provide small-ticket loans with flexible repayment terms and minimal documentation requirements to support the financial needs of customers with low income.

For Clusters 1, 2, and 4: As depicted in ***Figure 9***, the identified clusters share similarities in terms of young age and low transaction frequency, with the primary distinction being gender.

- **Women's Savings Account:** Introduce a specialized savings account exclusively for women in these clusters, offering additional benefits, personalized services, and financial empowerment initiatives.

- **Education Loans:** Design education loan products with competitive interest rates and flexible repayment options to support the educational aspirations of customers in these clusters, particularly targeting students or parents in need of financial assistance.

Figure 9: Cluster 1, 2, 4

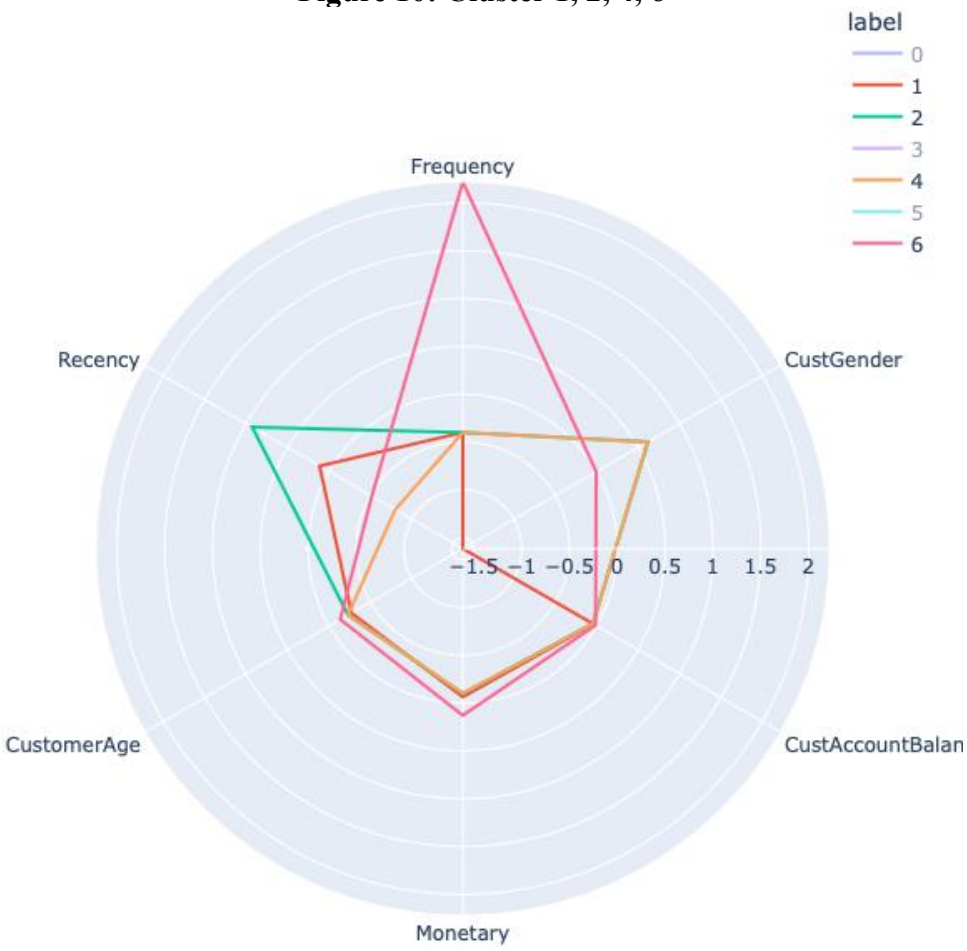


For Cluster 6: As illustrated in ***Figure 10***, Cluster 6 stands out with a higher frequency of transactions with the bank, while still exhibiting similarities with Cluster 1, 2, and 4.

- **Premium Current Account:** Offer a premium current account with enhanced features and benefits, such as higher transaction limits, preferential rates on services, and access to premium customer support, to cater to customers with a higher frequency of transactions.

- **Cashback Credit Cards:** Provide credit cards with cashback rewards or discounts on various spending categories, such as groceries, dining, or fuel, to offer tangible benefits to customers who engage in frequent transactions.

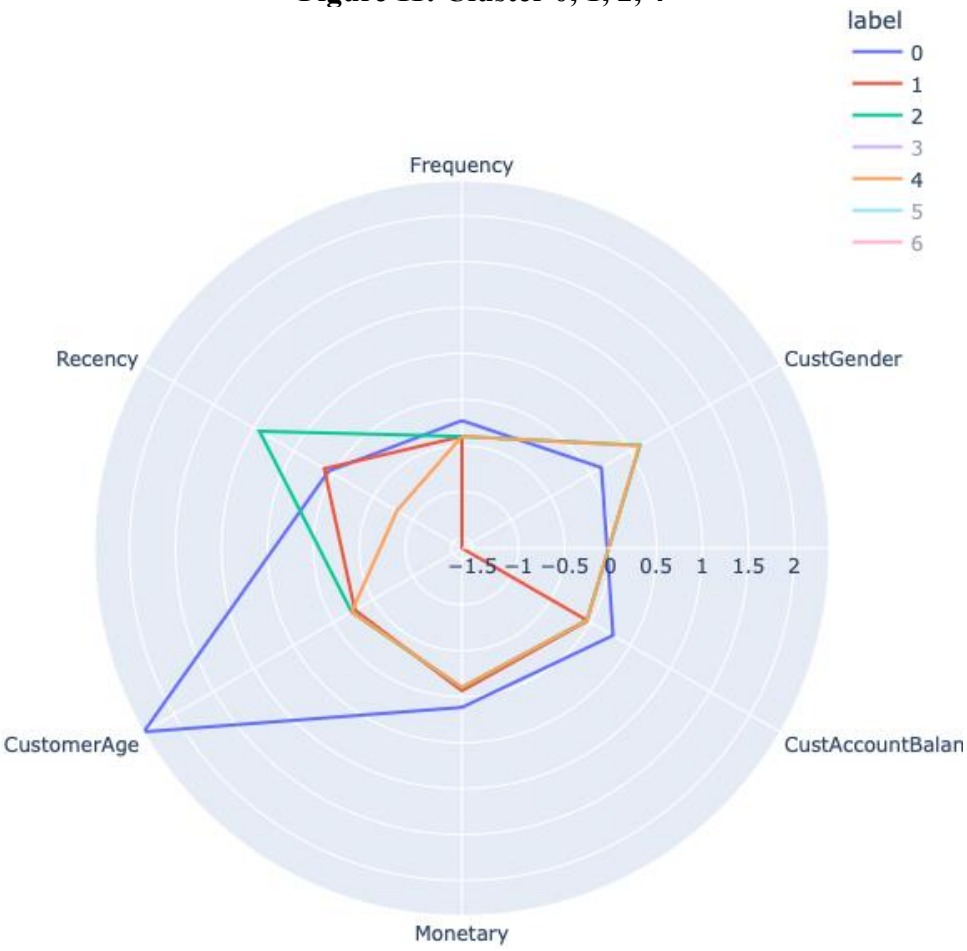
Figure 10: Cluster 1, 2, 4, 6



For Cluster 0: As shown in ***Figure 11***, this particular group consists of older individuals with a relatively higher financial status.

- **Wealth Management Services:** Establish a dedicated wealth management division to cater to the investment and financial planning needs of customers in this affluent cluster, providing personalized advisory services, investment options, and access to exclusive investment opportunities.
- **Senior Citizen Fixed Deposits:** Introduce fixed deposit schemes with preferential interest rates and additional benefits for senior citizens in Cluster 0, recognizing their specific financial needs and providing them with attractive options for secure and stable returns.

Figure 11: Cluster 0, 1, 2, 4



Reference

Shivam Bansal. “Bank Customer Segmentation (1M+ Transactions)” *Kaggle*, https://www.kaggle.com/datasets/shivamb/bank-customer-segmentation. Accessed 14 May 2023.