Customer Segmentation for Banking

## **Introduction：**

**In today’s competitive financial services landscape, understanding customer behavior is paramount to driving business growth and achieving a sustainable competitive advantage. This assignment applies a comprehensive RFM (Recency, Frequency, Monetary) analysis to a large-scale banking transaction dataset from a South East Asia–India bank. The dataset comprises over one million transactions from approximately 800,000 customers, containing variables such as customer demographics, transaction amounts, dates, and account details.**

**To ensure data quality and reliability, meticulous data cleaning was performed—removing invalid age entries, correcting erroneous transaction records, and harmonizing inconsistent location labels. Key RFM metrics were computed. In this study, the total transaction amount was selected as the Monetary metric, as it offers a more dynamic view of a customer’s transactional value compared to static indicators such as account balance. This aligns with the approach by Anitha and Patil (2022), who emphasized the importance of capturing actual purchase behavior for effective segmentation​.**

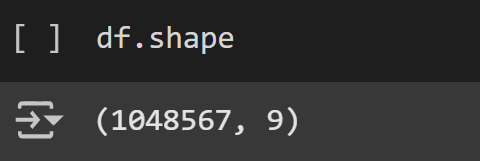
**To prepare the data for clustering, data transformation techniques including the Box-Cox transformation and standardization were applied. The Box-Cox method is particularly advantageous for normalizing skewed data and stabilizing variance, which improves the robustness and interpretability of cluster analysis (Osborne, 2010; Vélez et al., 2015)**

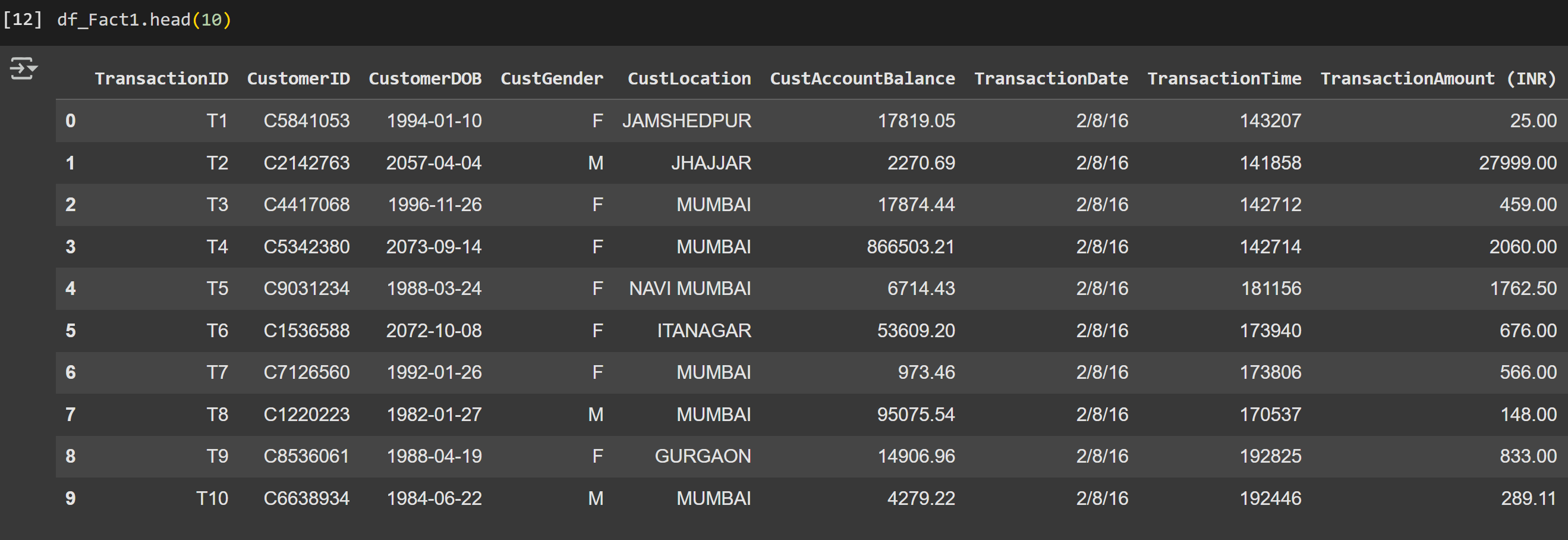
**The goal of this analysis is to generate actionable insights for marketing strategists, allowing for the development of targeted campaigns based on distinct customer profiles. This approach facilitates more effective customer relationship management (CRM), an objective that has been widely endorsed in the literature as a critical success factor for business intelligence applications in banking and retail domains (Safari et al., 2016)​.**

## **Tasks A**

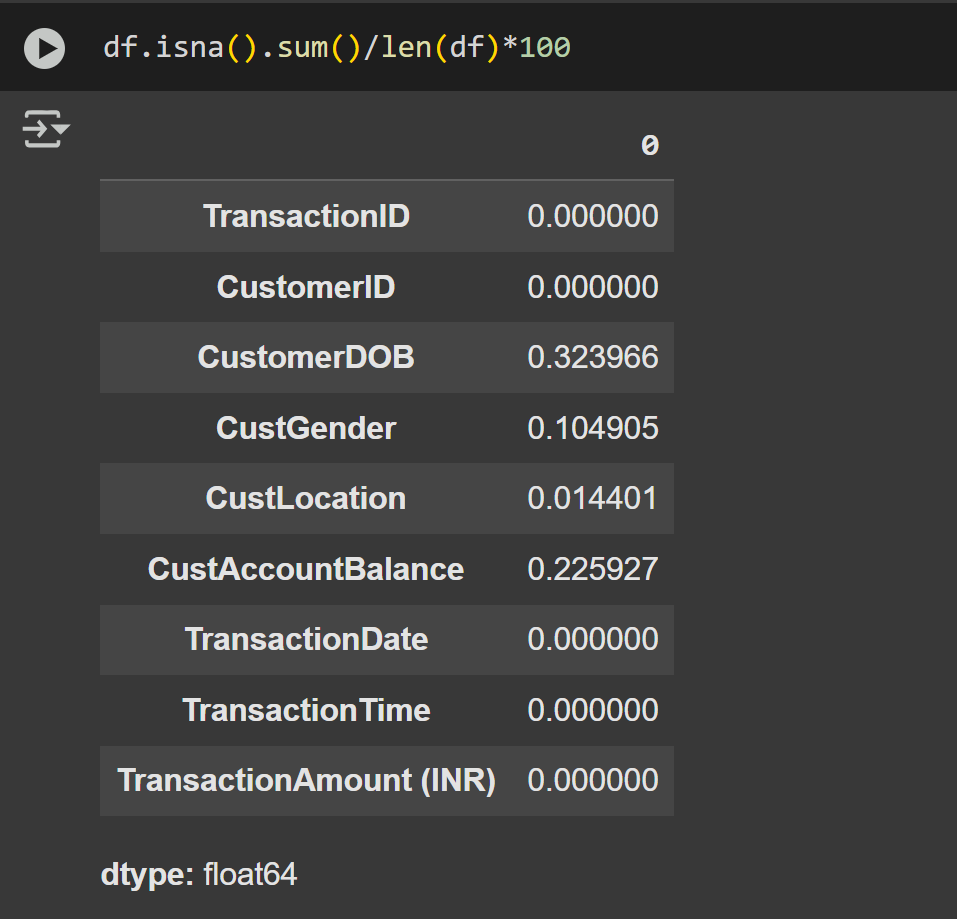
* **Data Understandin**

**The initial phase of the analysis involved a thorough understanding and cleansing of the dataset to ensure data quality and reliability. The dataset consisted of over 1 million transaction records from a bank in South East Asia, comprising features such as customer ID, date of birth (DOB), gender, transaction date, transaction amount, and account location. The data has 9 features and 1,048,567 instances.**

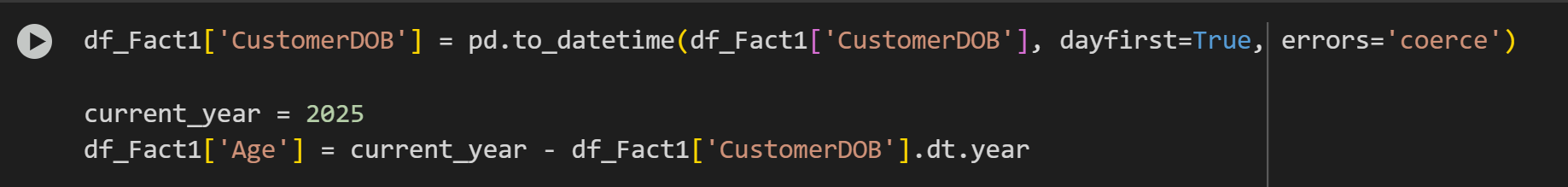




**Missing values were quantified using the df.isna().sum()/len(df)\*100 method. Features such as CustomerDOB, CustGender, and CustLocation were found to have less than 1% missing values, which were dropped accordingly—an acceptable threshold as per industry standards (Rahm & Do, 2000)​.**



**Next, invalid records were addressed. For age calculation, a new column was derived by subtracting the year of birth from 2025 and any entries with ages below 0 or above 100 were deemed invalid and removed. Identifying and filtering such implausible values is fundamental in maintaining dataset integrity, as erroneous data entries can severely distort downstream analytics.**



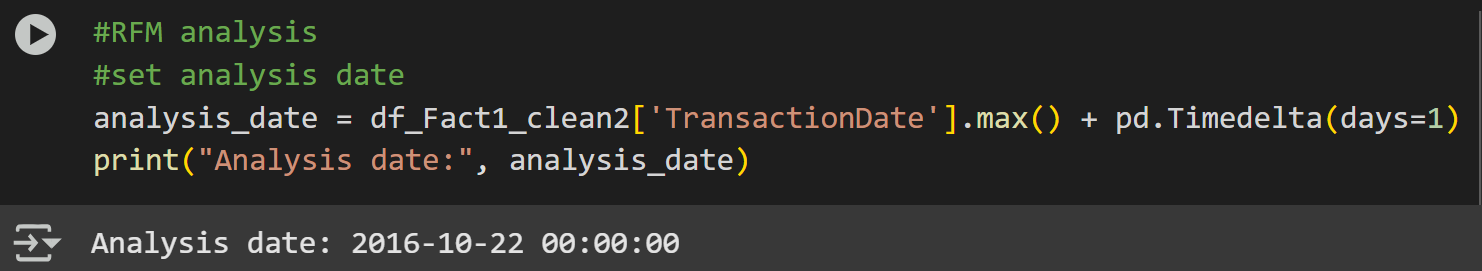
**Since invalid age data accounted for a high proportion, we decided to use the mean age interpolation method to consider data integrity and ensure data size.**

**After cleaning the age data, we get a data dimension as 10 features and 1,041,614 instances.**

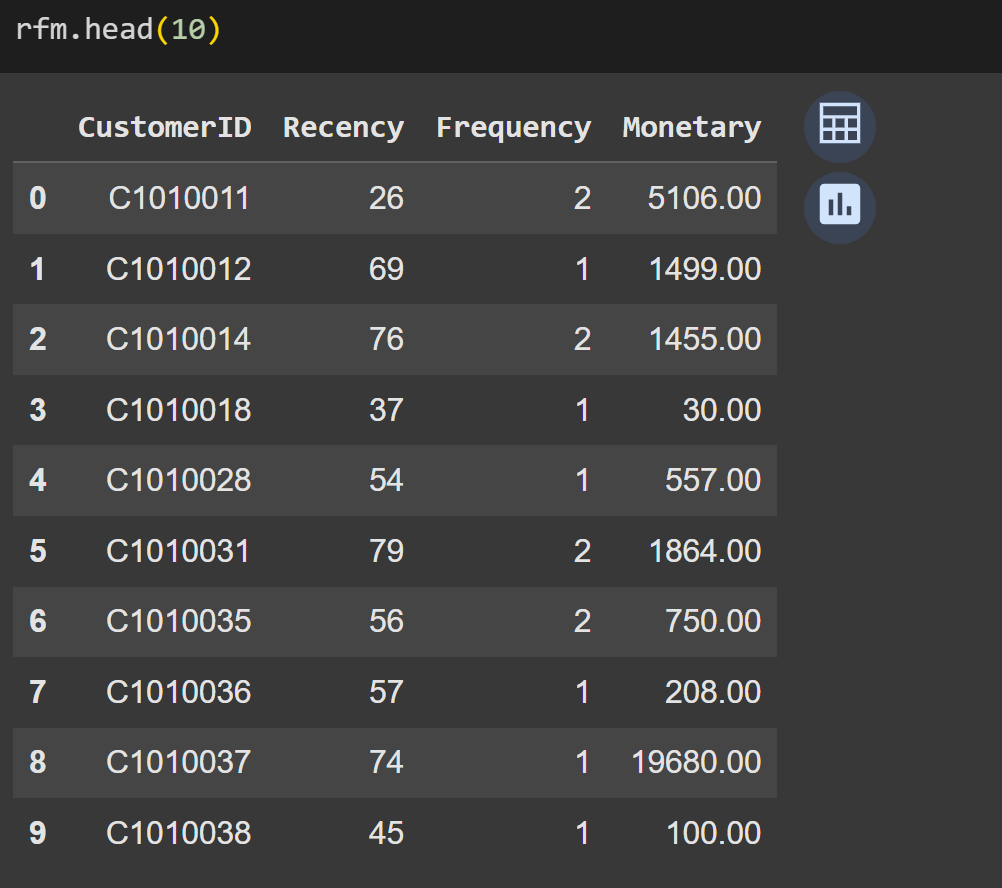
* **RFM Segmentation**

**To perform customer segmentation, we employed the RFM (Recency, Frequency, Monetary) model—an industry-standard framework that classifies customers based on their transactional behavior.**

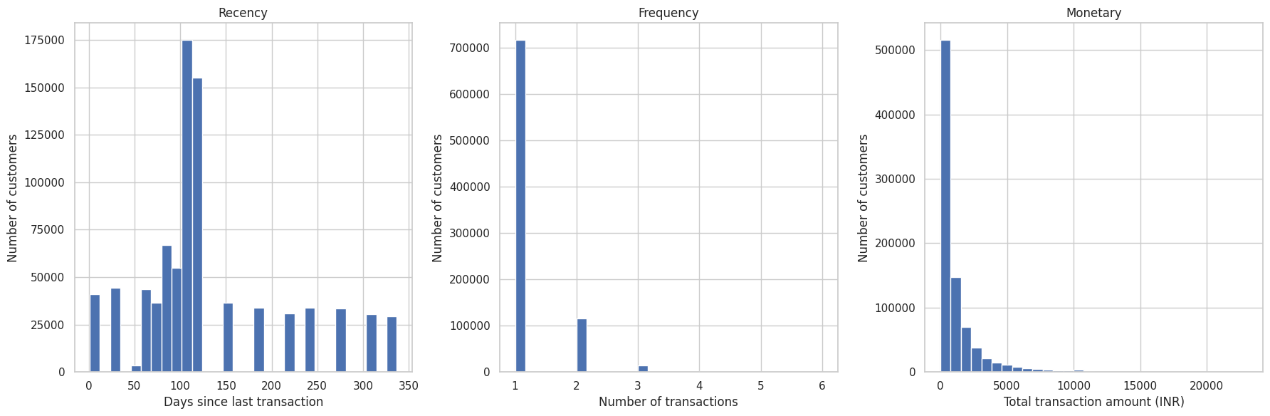
**The Recency was calculated as the number of days between the analysis date and the customer's most recent transaction. We set the analysis date to one day after the latest transaction date in the dataset to avoid zero values and preserve mathematical validity, particularly when applying transformations later on.**



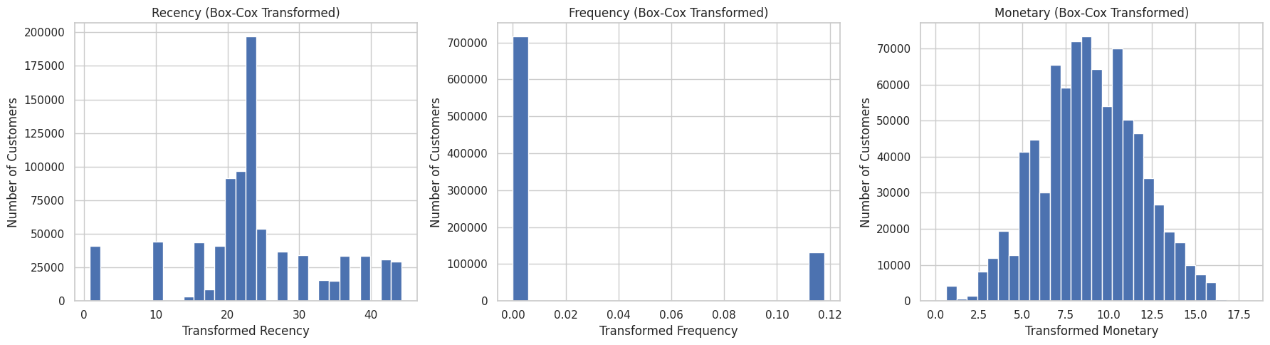
**Frequency was computed as the total number of transactions per customer, while Monetary represented the sum of transaction amounts for each customer, rather than using account balance, which is static and does not accurately reflect behavioral value.**



**An initial exploration of RFM value distributions revealed that both Frequency and Monetary were highly right-skewed. Histograms confirmed that most customers transacted only once, and a few customers contributed disproportionately large transaction amounts.**



**To address this skewness, the Box-Cox transformation was applied to all three RFM variables. Recency and Monetary values were incremented by 1 to meet the requirement of strictly positive inputs. Post-transformation histograms showed significant improvement in normality, particularly for Monetary, which approximated a bell-shaped distribution. The transformation was also effective for Frequency, which, despite its discrete nature, exhibited reduced skewness. The appropriateness of Box-Cox in such use cases is well-documented in literature as an optimal practice for improving normality and variance homogeneity in transactional data (Osborne, 2010; Vélez, Correa and Marmolejo-Ramos, 2015).**



**Following transformation, standardization using StandardScaler was performed to bring all RFM features onto the same scale. This step is critical before implementing distance-based clustering algorithms like K-Means to prevent any single variable from disproportionately influencing the results (Osborne, 2010).**

## **Tasks B**

* **Customer Segmentation Using K-Means Clustering**

**K-Means clustering is a widely used unsupervised learning algorithm for customer segmentation. Compared to other clustering techniques, K-Means performs competitively on numerical datasets and demonstrates strong performance on textual data as well (Gupta et al., 2024). It is considered the second-best among five evaluated clustering algorithms, with agglomerative clustering ranking first. However, K-Means is significantly faster, particularly for large-scale datasets, despite using more memory (Karthikeyan et al., 2020). Therefore, it is considered more suitable for high-volume data environments such as the one used in this study.**

**First, it determines the centers randomly and defines the clusters. Then, for each of these clusters centroids are recalculated until there is no change in the clusters. The efficiency of the algorithm can be measured by examining the clusters or by using a numerical measure called clusters' distortion that is the sum of squared differences between every data point and their correspoding centroids. The lowest distortion value is the best number of clusters. Determining the number of clusters (k) is the main challenge.**

**However, K-Means assumes convex, isotropic clusters with equal variance and is sensitive to initialization. These limitations were mitigated through data preprocessing, including the use of Box-Cox transformations to reduce skewness and standardization to normalize scale. This ensured a fair contribution of each RFM dimension to the clustering process.**

* **Determining the Optimal Number of Clusters (K)**

**Selecting an appropriate number of clusters (K) is a critical step in K-Means clustering, as it directly influences the quality and interpretability of the segmentation results. In this study, we employed two widely accepted internal validation techniques: the Elbow Method and the Silhouette Score.**

**The Elbow Method involves running the K-Means algorithm for a range of K values and plotting the** within-cluster sum of squares (WCSS) **for each. The “elbow” point on the curve, where the marginal gain in WCSS reduction diminishes significantly, suggests the most appropriate number of clusters. However, several researchers caution against over-reliance on this method, as the elbow point can be difficult to interpret, especially when the data does not have a clear natural clustering structure (Schubert, 2023)​.**

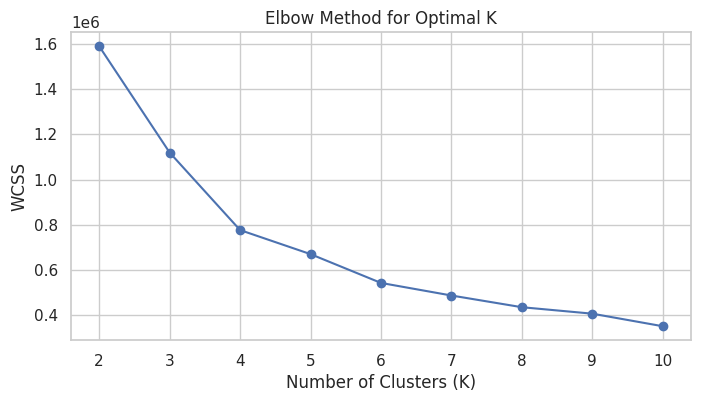
**To complement this, we applied the Silhouette Score** **to verify, which measures how similar a data point is to its own cluster compared to other clusters. It provides a more robust evaluation of clustering quality, particularly when** WCSS **alone is insufficient to determine the optimal K. According to Liu et al. (2010), the Silhouette index is one of the most reliable internal validation metrics and consistently performs well under various data distribution scenarios​.**

**Together, these two methods ensure a more informed and balanced decision on the final value of K. The optimal cluster number selected will be confirmed in the implementation phase based on the convergence of results from both approaches.**

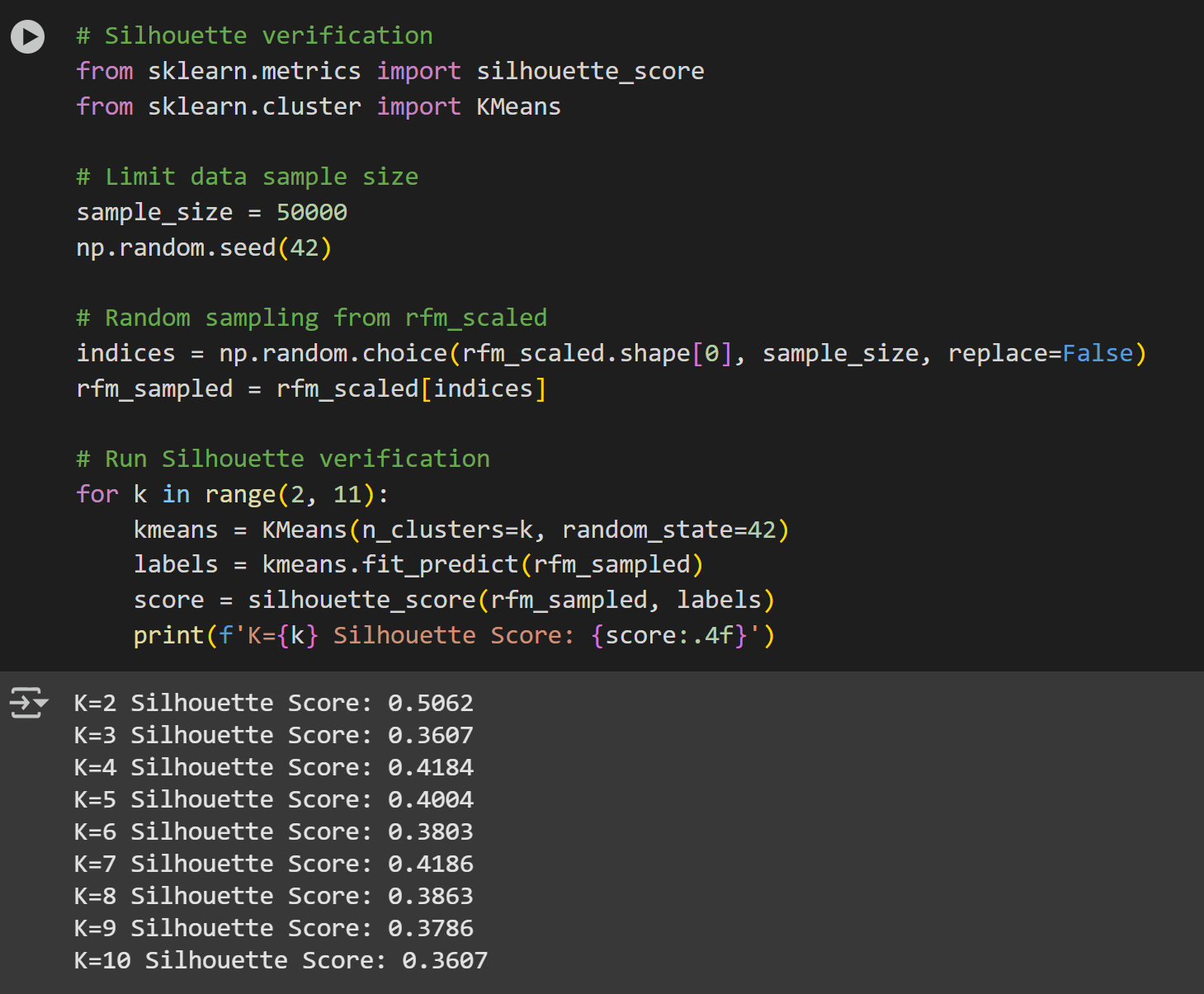
* **Implementation of K-Means**

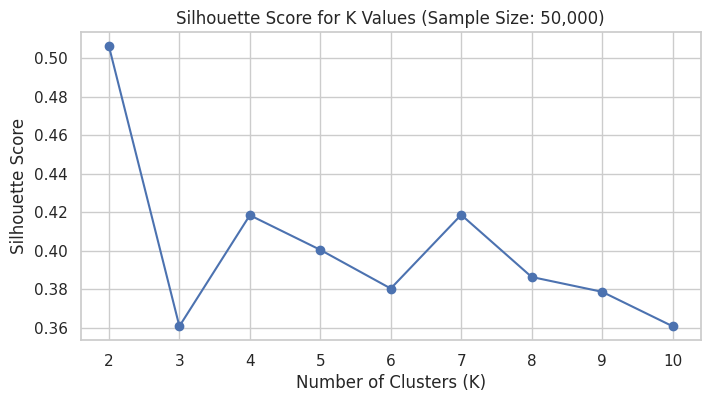
**The standardized RFM dataset was used to implement the K-Means clustering algorithm using scikit-learn. Before applying the algorithm, the optimal number of clusters (K) was determined through a combination of the Elbow Method and Silhouette Analysis, as recommended by Liu et al. (2010) and Gupta et al. (2024).** **Due to the large size of the dataset (~1 million rows), Silhouette score analysis was performed on a stratified random sample of 50,000 records to ensure computational efficiency. This approach is commonly adopted in large-scale clustering tasks (Schubert, 2023).**

**The Elbow Method revealed a clear inflection point at K = 4, indicating a diminishing return in reducing the within-cluster sum of squares (WCSS) beyond this point.**

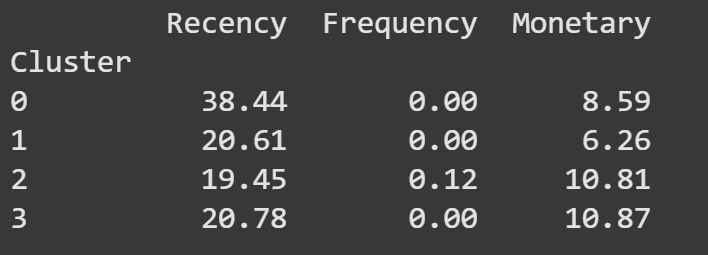


**Meanwhile, the Silhouette Score peaked at K = 2 (0.5062), but the second-highest score was recorded at K = 4 (0.4184), which presented a more meaningful segmentation structure. K = 4 was therefore selected as the optimal cluster number based on its balance between clustering compactness and interpretability.**





**The K-Means algorithm was trained using the scaled RFM features, and each customer was assigned to one of the four resulting clusters. The table below presents the average Recency, Frequency, and Monetary values for each group:**



**The four clusters generated through K-Means reveal distinct transaction behaviors among the bank’s customers:**

**Cluster 0 – Inactive or Lapsing Clients:**

**Customers in this cluster show longer recency, no recent transactions, and moderate historical monetary values. These clients are at risk of churn and may benefit from reactivation campaigns or personalized contact.**

**Cluster 1 – New or Low-Engagement Clients:**

**This segment shows relatively recent engagement but low monetary and frequency levels. They may represent new account holders or clients with minimal transactional activity and should be considered for onboarding or cross-selling strategies.**

**Cluster 2 – Highly Engaged High-Value Clients:**

**These customers demonstrate low recency, frequent transaction activity, and high total transaction amounts, indicating a strong, ongoing relationship with the bank. They are prime candidates for tailored relationship management and value-added services.**

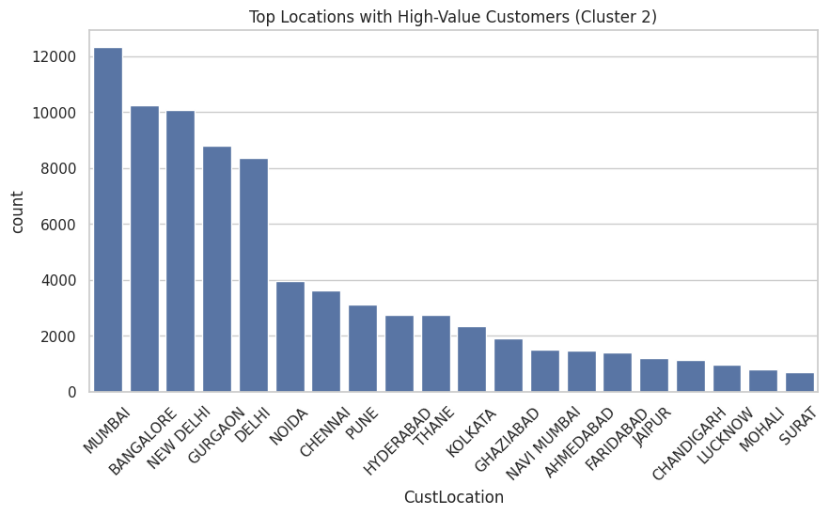
**Cluster 3 – Recent High-Monetary One-Time Transactors:**

**Clients in this group have engaged with the bank recently and exhibit high monetary activity, but with low frequency, suggesting significant one-off transactions. These clients may require personalized outreach to increase their transaction frequency.**

* **Data Mart Design for Marketing Analytics**

**Based on the RFM segmentation and K-Means clustering, Cluster 2 was identified as the most valuable customer group. These clients exhibit high transaction frequency and high monetary activity with recent engagement, making them key targets for relationship management, loyalty programs, or cross-selling initiatives.**

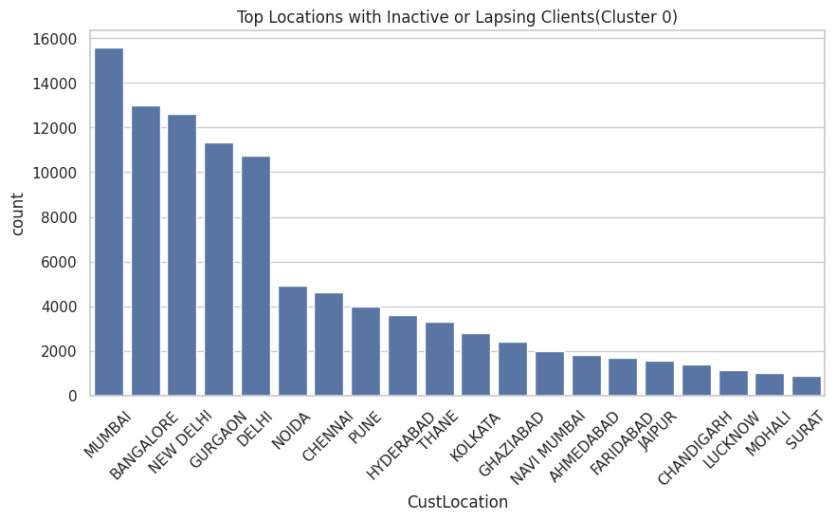
**The top locations of Cluster 2 clients were identified through visual and statistical analysis. The cities with the highest concentrations of high-value customers (Cluster 2) are:**



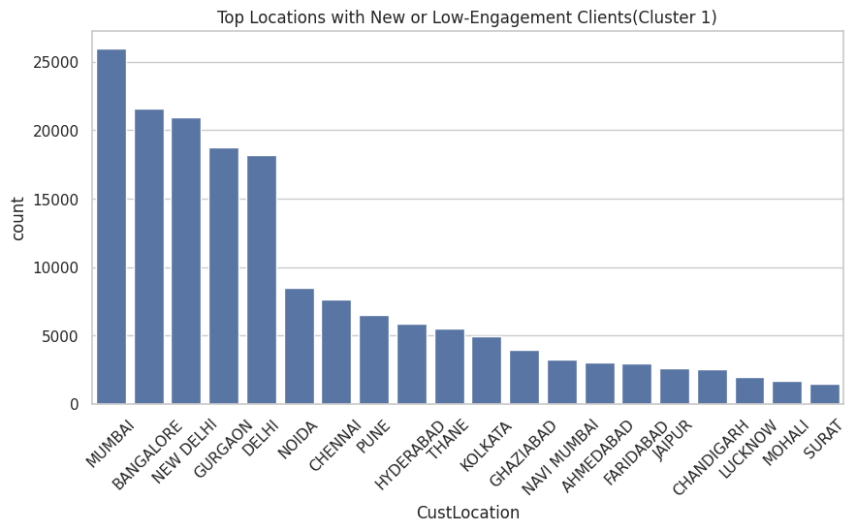
**These urban centers represent the bank’s most financially active and loyal customer base, and thus, strategic investments in personalized services, regional promotions, and high-end product offerings in these areas are strongly recommended.**

**The distribution charts also reveal that:**

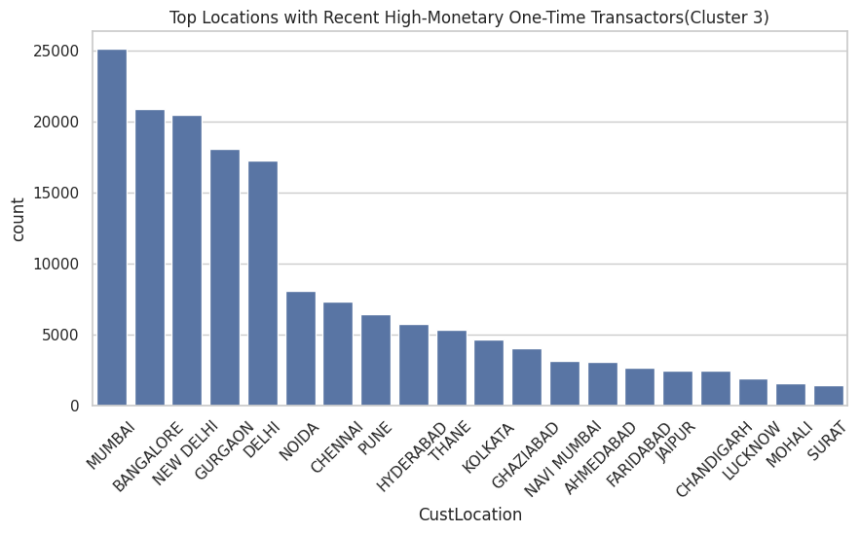
**Cluster 0 (Inactive/Lapsing Customers) are heavily concentrated in Mumbai, Bangalore, and New Delhi. These customers may require reactivation campaigns or incentives to return.**



**Cluster 1 (New or Low-Engagement Clients) are widespread across the same major metros, suggesting opportunities for onboarding or first-time engagement strategies.**



**Cluster 3 (Recent High-Monetary One-Time Transactors) are also prominent in Tier-1 cities, indicating they may be affluent individuals who have yet to develop long-term banking relationships.**



**High-value customers in Cluster 2 not only maintain frequent and recent transactions, but also contribute significantly in terms of transaction volume. It is worth noting that all four customer clusters—ranging from inactive users to high-value transactors—share a similar geographic concentration across major Tier-1 cities such as Mumbai, Bangalore, New Delhi, Gurgaon, and Delhi. This suggests that these cities represent the highest-density banking regions in the dataset.**

**While the overall volume of customers is highest in these locations across all clusters, behavioral segmentation using K-Means remains crucial for identifying quality over quantity. For example, Mumbai hosts both the largest number of high-value clients (Cluster 2) and inactive customers (Cluster 0), indicating might the need for targeted micro-segmentation within high-density areas.**

**Therefore, the marketing strategy should not only focus on “where” customers are, but also on “who” they are within each city, using behavioral attributes to inform campaign design, product targeting, and resource allocation.**

**Reference List**

* Anitha, P. & Patil, M.M. (2022). *RFM model for customer purchase behavior using K-Means algorithm*. Journal of King Saud University – Computer and Information Sciences, 34(2022), pp. 1785–1792.
* Rahm, E. & Do, H.H. (2000). *Data Cleaning: Problems and Current Approaches*. IEEE Data Engineering Bulletin, 23(4), pp. 3–13.
* Safari, F., Safari, N. & Montazer, G.A. (2016). *Customer lifetime value determination based on RFM model*. Marketing Intelligence & Planning, 34(4), pp. 446–461.
* Osborne, J. W. (2010). *Improving your data transformations: Applying the Box-Cox transformation*. Practical Assessment, Research, and Evaluation, 15(12). [Available online](http://pareonline.net/getvn.asp?v=15&n=12)
* Vélez, J. I., Correa, J. C., & Marmolejo-Ramos, F. (2015). *A new approach to the Box–Cox transformation*. *Frontiers in Applied Mathematics and Statistics*, 1:12. <https://doi.org/10.3389/fams.2015.00012>
* Bruce, P.C. (2016) *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner*. Newark: John Wiley & Sons, Incorporated.
* Gupta, S., Kishan, B. and Gulia, P. (2024) ‘Comparative analysis of predictive algorithms for performance measurement’, *IEEE Access*, 12, pp. 33949–33958. doi:10.1109/access.2024.3372082.
* Karthikeyan, B. *et al.* (2020) ‘A Comparative Study on K-means Clustering and Agglomerative Hierarchical Clustering’, *International Journal of Emerging Trends in Engineering Research*, 8(5), pp. 1600–1604. doi:10.30534/ijeter/2020/20852020.
* Liu, Y. *et al.* (2010) *Understanding of Internal Clustering Validation Measures*. Available at: <https://ieeexplore.ieee.org/Xplore/home.jsp> (Accessed: 06 April 2025).
* Schubert, E. (2023) *Stop Using the Elbow Criterion for K-Means and How to Choose the Number of Clusters Instead*. Available at: <https://dl.acm.org/doi/10.1145/3606274.3606278> (Accessed: 06 April 2025).