Dr. Vishwanath Karad
MIT WORLD PEACE UNIVERSITY | PUNE
TECHNOLOGY, RESEARCH, SOCIAL INNOVATION & PARTNERSHIPS

Minor Project-II Report

ON

Bank Customer Segmentation

SUBMITTED BY,

SHREERANG MHATRE (52)

SARVESH GURAV (44)

Minor Project Coordinator

Prof. Bharat Chodhari

Year: 2023-2024

Department of Electrical and Electronics Engineering

Dr. Vishwanath Karad, MIT world Peace University Pune - 38.

CERTIFICATE

This is to certify that the Minor Project - II entitled

Bank Customer Segmentation

has been carried out successfully by

SHREERANG MHATRE (52)

SARVESH GURAV (44)

during the Academic Year **2023-2024** in partial fulfillment of their course of study for
Bachelor's Degree in
Electrical and Computer Engineering as per the syllabus prescribed by the **MIT-WPU**

Internal Guide

Head,
(School of Electrical Engineering)

DECLARATION

We the undersigned, declare that the work carried under

Minor Project - II entitled

Bank Customer Segmentation

has been carried out by us and it has been not implemented by any external agency/company that sells projects. We further declare that work submitted in the form of a report has not been copied from any paper/thesis/website as it is. However existing methods/approaches from any paper/thesis/website have been cited and have been acknowledged in the reference section of this report.

We are aware that our failure to adhere to the above, the Institute/University/Examiners can take strict action against us. In such a case, whatever action is taken, it would be binding on us.

PRN	**Name of student**	**Signature with date**
1032211745	SHREERANG MHATRE	
1032222100	SARVESH GURAV	

INDEX

ABSTRACT

This research project delves into the critical domain of customer segmentation within the banking industry, leveraging advanced machine learning methodologies. The primary focus is on the application of the K-means clustering algorithm to partition a diverse customer base into distinct segments based on multifaceted attributes such as demographics, transaction history, and account balances. The objective is to uncover meaningful customer groups, enabling banks to formulate targeted marketing strategies, personalized product offerings, and tailored customer experiences. Moreover, the project integrates the K-Nearest Neighbors (KNN) algorithm to classify customer data into discrete categories representing low, medium, and high account balances, providing actionable insights for predicting and managing customer financial behaviors. Through rigorous data analysis and model implementation, this research endeavors to equip banks with valuable insights to enhance customer satisfaction, drive revenue growth, and strengthen customer relationships. By adopting a data-driven approach, banks can better understand customer needs and preferences, optimize resource allocation, and stay competitive in the ever-evolving banking landscape. The findings and methodologies presented in this study contribute to the advancement of customer segmentation strategies, emphasizing the importance of leveraging machine learning techniques for strategic decision-making and customer-centric innovation in the banking sector.

Chapter 1 INTRODUCTION

1.1 Overview:

This research delves into the critical realm of customer segmentation within the banking sector, a fundamental strategy aimed at understanding customer needs, preferences, and behavior. Utilizing advanced machine learning techniques, specifically the K-means clustering algorithm and K-Nearest Neighbors (KNN) algorithm, this study seeks to segment a diverse customer base into distinct groups based on demographics, transaction history, and account balances. By leveraging these methodologies, banks can tailor marketing strategies, product offerings, and customer experiences to enhance customer satisfaction and drive revenue growth.

1.2 Scope:

This research extends beyond traditional customer segmentation approaches by integrating machine learning algorithms for enhanced segmentation accuracy and predictive analytics in the banking industry. The methodologies and insights generated from this study have broader implications for marketing strategies, customer relationship management, and personalized financial services. The findings contribute to a deeper understanding of customer segmentation strategies, benefiting banks, financial institutions, marketing professionals, and researchers involved in data-driven decision-making and customer-centric innovation in the competitive banking landscape.

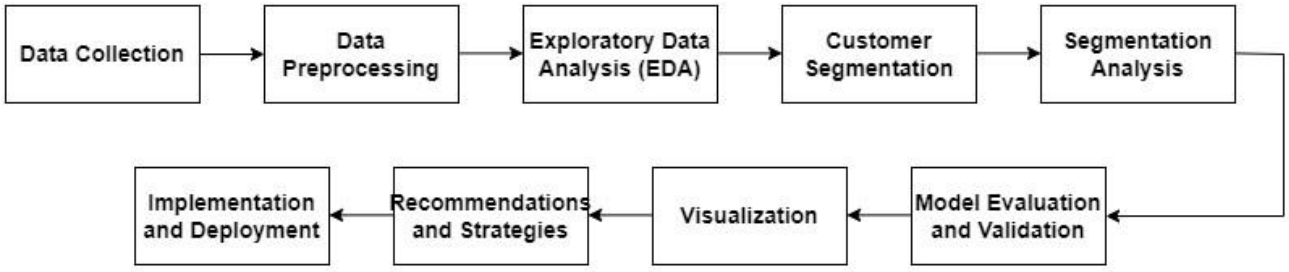
Chapter 2 Review of Literature

2.1 Literature

The literature review highlights several notable studies in the field of bank customer segmentation, showcasing diverse approaches and methodologies. Zakrzewska and Murlewski (2005) proposed a two-phase clustering algorithm, incorporating modified k-means and hierarchical agglomerative techniques, emphasizing scalability and outlier detection. Arta Moro Sundjaja (2013) focused on practical insights from data mining techniques, enhancing marketing promotion and product development based on customer segmentation. Ion Smeureanu (2013) explored neural networks and support vector machines for segmentation, discussing their performance in comparison. Om Atre et al. (2022) utilized K-Means Clustering and a Streamlit Application for targeted recommendations, while Shahenaj Begam (2021) analyzed transactional habits for improved customer engagement. However, none of these studies employed a unified data science platform like Rubiscape, streamlining the machine learning process and achieving results more efficiently.

Chapter 3 System Specifications

3.1 System Block Diagram



3.2 Project Specifications

1. System Architecture:

The system architecture for Bank Customer Segmentation involves utilizing Python as the foundational programming language for implementing clustering algorithms and data preprocessing.

2. Software Requirements:

Python serves as the primary software requirement for implementing the segmentation algorithms, along with necessary libraries for data manipulation, visualization, and model evaluation.

3. Functional Requirements:

Functionalities are derived from specific input columns in the dataset, including customer demographics (age, gender), transaction details (amount, frequency), account balances, and location data.

4. Data Requirements:

The dataset undergoes rigorous preprocessing, including handling missing values, encoding categorical variables, and scaling numerical features to ensure accurate and meaningful segmentation.

5. User Interface (UI) Design:

The final segmentation results and insights are presented through customized dashboards or visualizations designed for easy interpretation by stakeholders and decision-makers in the bank.

6. Dependencies:

Essential libraries for machine learning and data analysis, such as scikit-learn, pandas, and matplotlib/seaborn, are utilized to implement clustering algorithms and evaluate segmentation models.

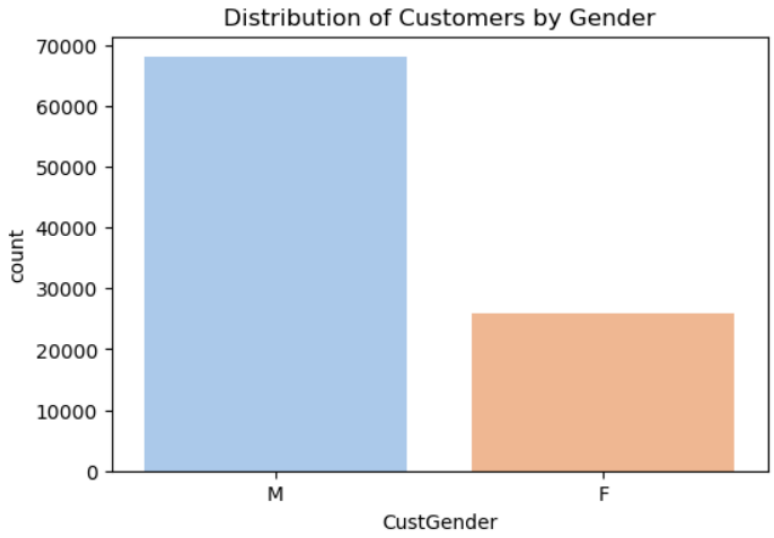
3.3 Complexities Involved:

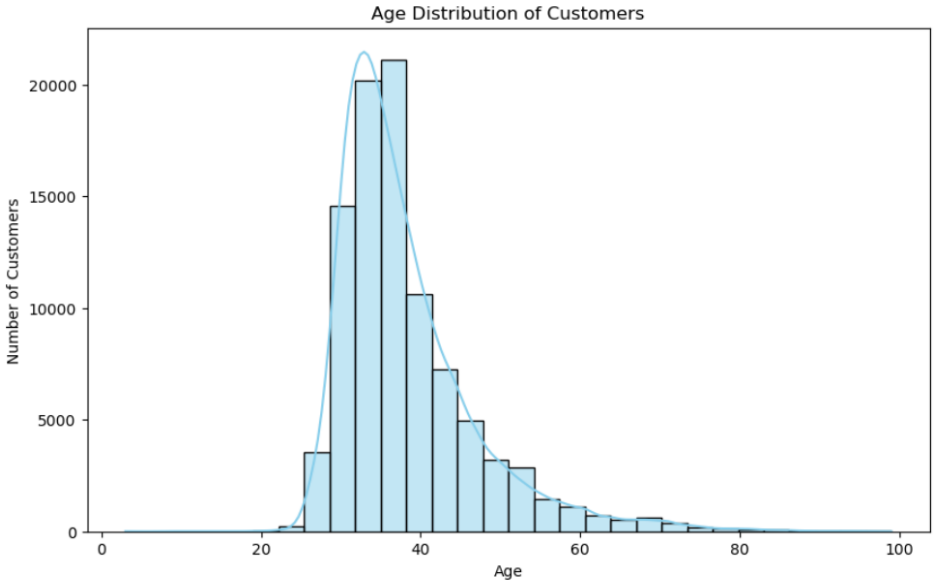
- **Data Quality and Anomalies:** Ensuring data accuracy and completeness while handling outliers and missing values to maintain the integrity of segmentation results.
- **Model Selection and Evaluation:** Choosing suitable clustering algorithms (e.g., K-means, Hierarchical Clustering) and evaluation metrics (e.g., Silhouette Score) to achieve optimal customer segmentation.
- **Dynamic Customer Behavior:** Adapting segmentation strategies to changing customer preferences, transaction patterns, and market trends to ensure relevance and effectiveness.

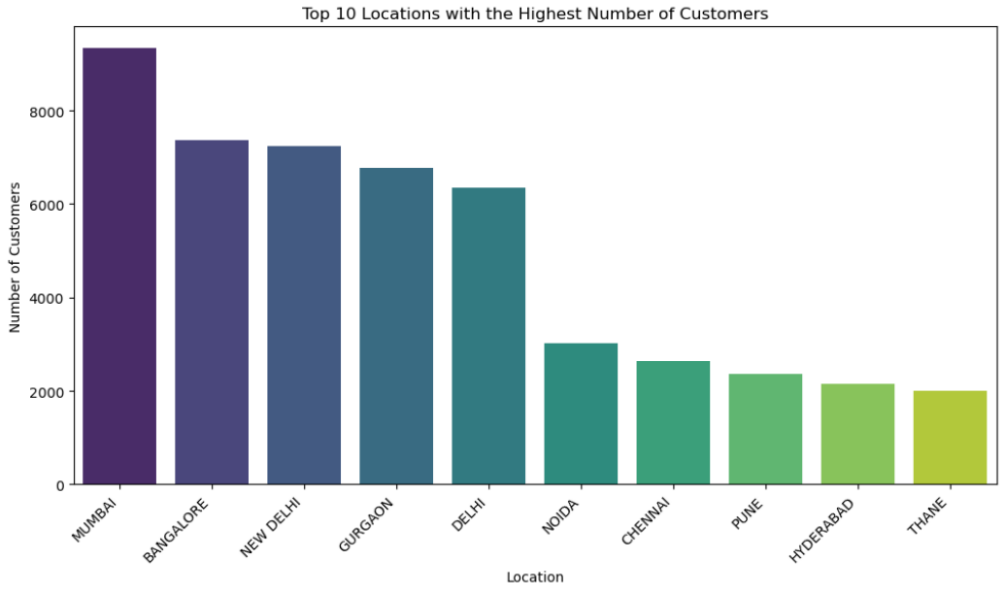
- **Interpretability and Bias:** Addressing interpretability challenges in segmentation models and mitigating potential biases to ensure fair and accurate customer categorization.
- **Resource Management:** Efficiently managing computational resources and scalability to handle large datasets and complex algorithms for segmentation.
- **Regulatory Compliance:** Adhering to data privacy regulations and ethical considerations in handling customer data and segmentation practices within the banking sector.
- **Continuous Improvement:** Implementing mechanisms for continuous monitoring, evaluation, and refinement of segmentation models to adapt to evolving customer dynamics and market conditions.

Chapter 4 EDA (Exploratory Data Analysis)

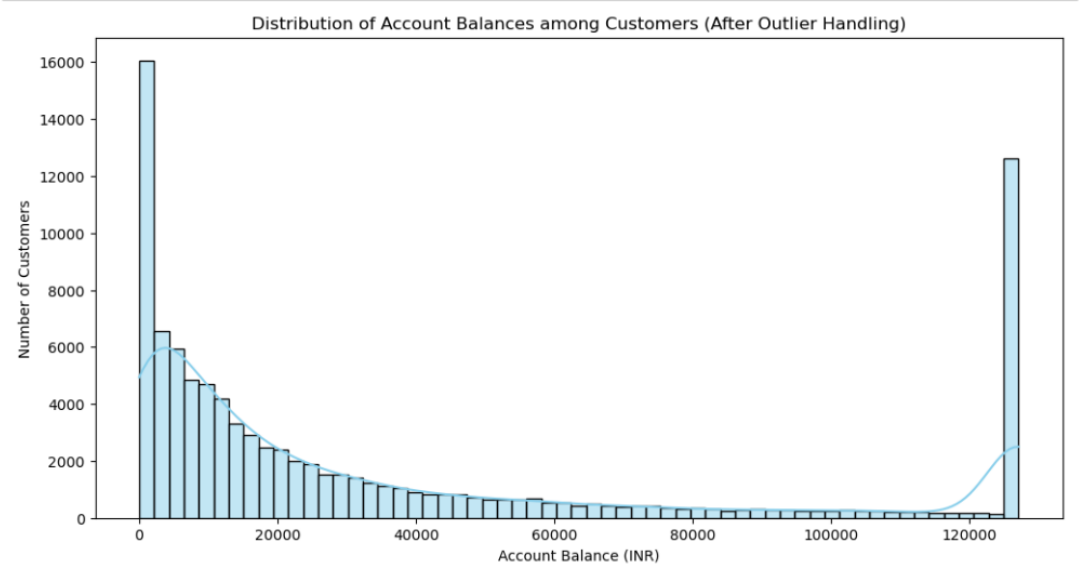
4.1 Customer Demographics:

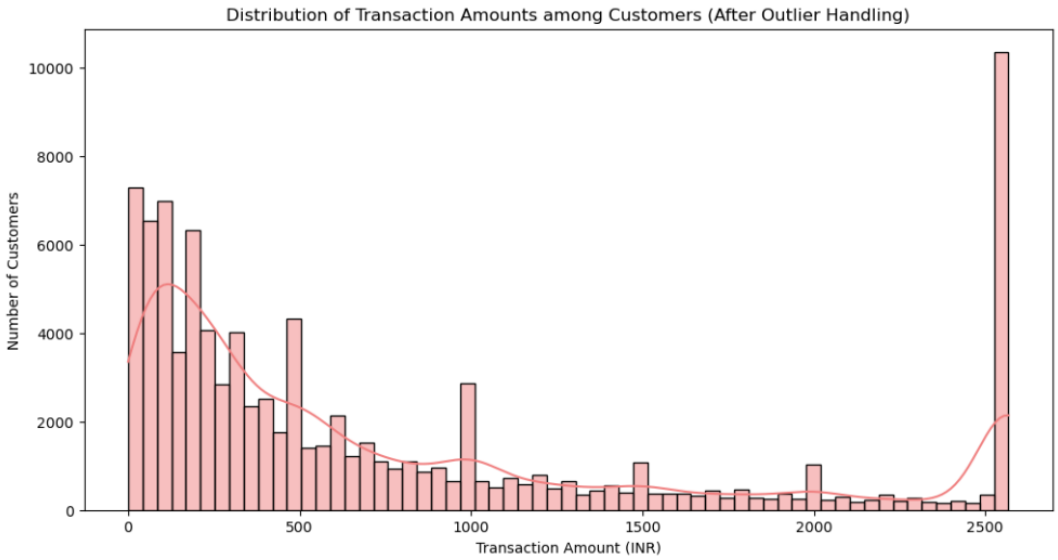


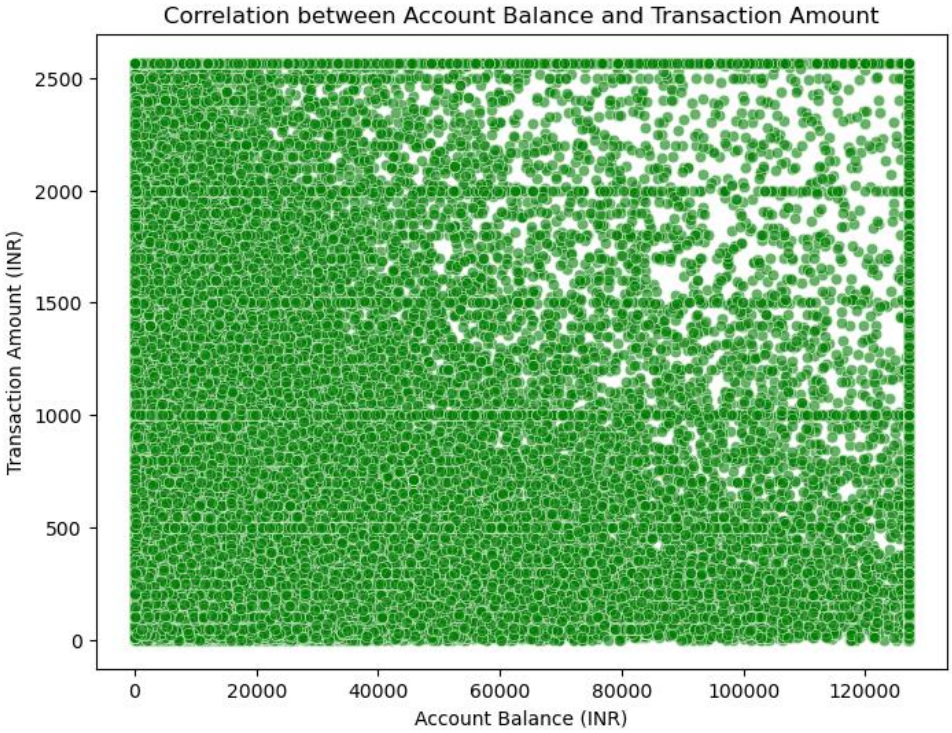




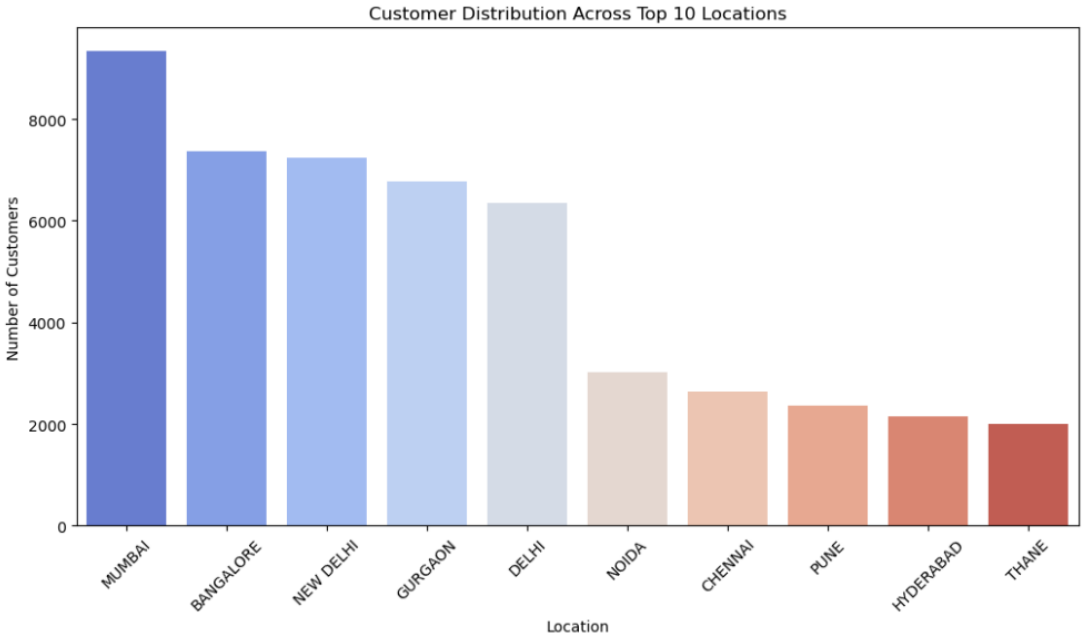
4.2 Customer Account and Transaction Analysis:

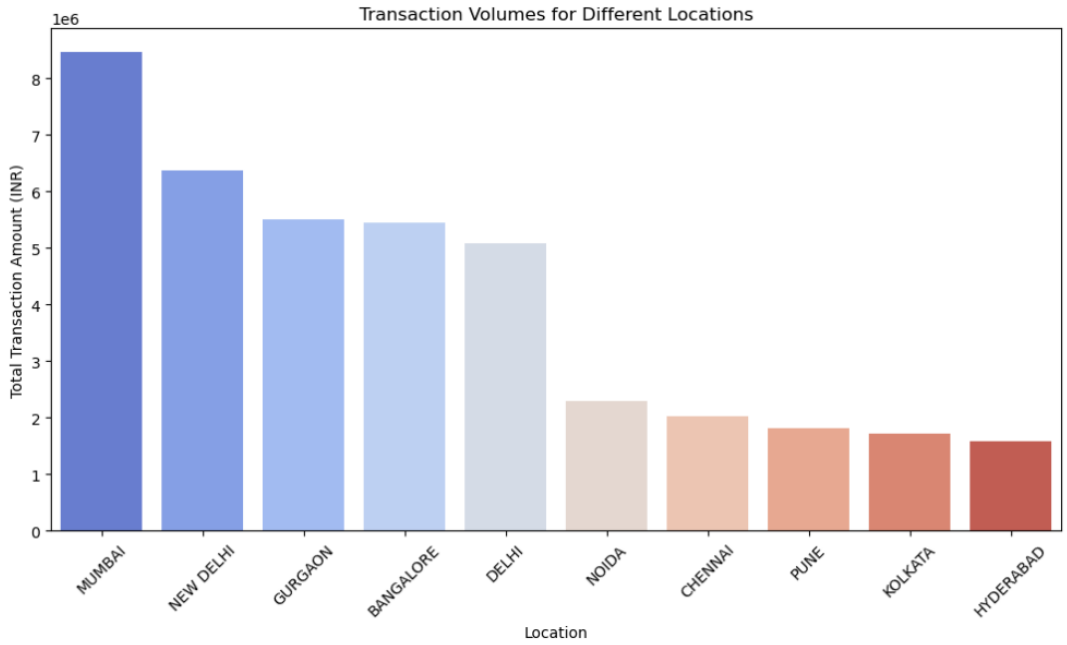



Distribution of Transaction Amounts among Customers (After Outlier Handling)
Number of Customers
Transaction Amount (INR)
0
2000
4000
6000
8000
10000
0
500
1000
1500
2000
2500


Correlation between Account Balance and Transaction Amount
Transaction Amount (INR)
Account Balance (INR)
0
500
1000
1500
2000
2500
0
20000
40000
60000
80000
100000
120000

4.3 Customer Location Analysis:



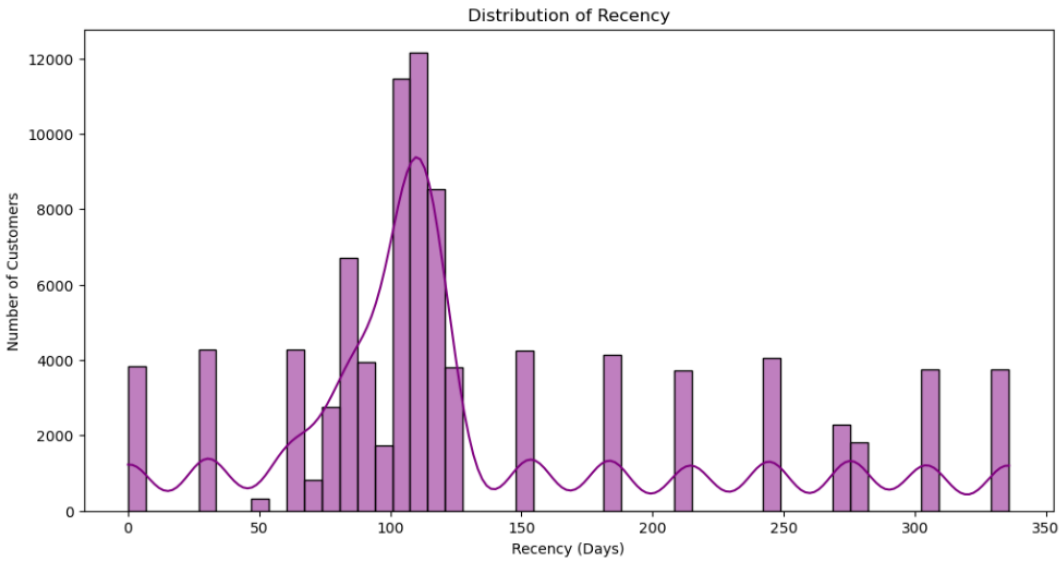


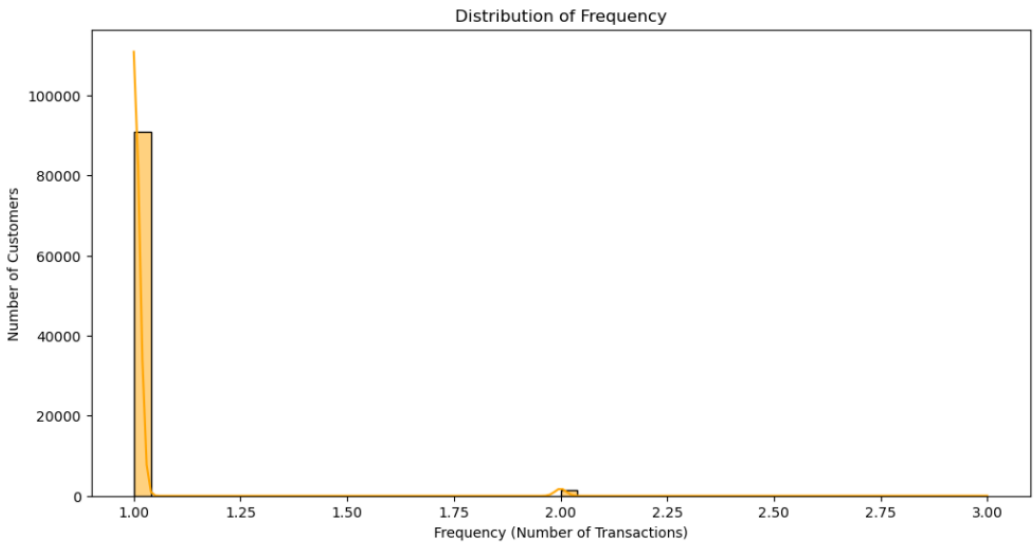
Chapter 5 RFM (Customer Recency, Frequency, Monetary)

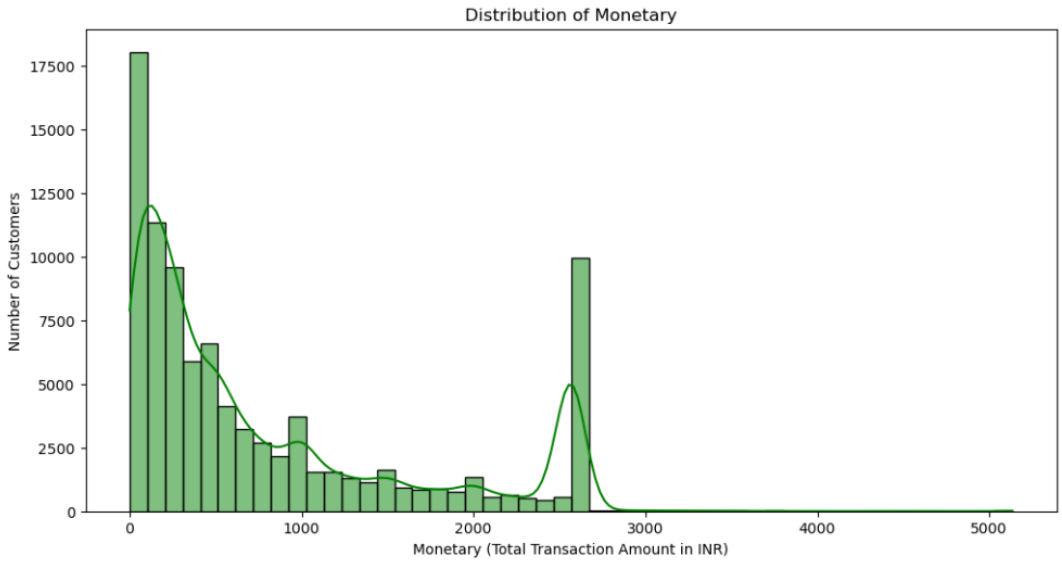
Recency (R): The number of days since the customer's most recent transaction.

Frequency (F): The total number of transactions made by each customer.

Monetary (M): The total monetary value of all transactions made by each customer.







1) Recency:

The average recency value (mean) is approximately 135.55 days, indicating that, on average, customers made their most recent transaction about 135 days ago. The minimum recency value is 0 days, suggesting that some customers made transactions very recently, while the maximum recency value is 336 days, indicating the longest period since a transaction.

2) Frequency:

The average frequency of transactions per customer (mean) is approximately 1.02, implying that, on average, customers made slightly more than 1 transaction. The minimum frequency value is 1 transaction, indicating that all customers have made at least one transaction, while the maximum frequency value is 3 transactions, suggesting some customers made multiple transactions.

3) Monetary:

The average monetary value of transactions per customer (mean) is approximately 796.98 INR, indicating the average transaction amount. The

minimum monetary value is 0 INR, which could indicate free or non-monetary transactions, while the maximum monetary value is 5135 INR, suggesting the highest transaction amount among customers.

Based on these RFM features, we can identify valuable customer segments as follows:

1) High-Value Customers: Customers with a low recency value (recent transactions), high frequency (multiple transactions), and high monetary value (significant transaction amounts) are considered high-value customers. These customers contribute significantly to revenue and should be targeted with personalized offers, loyalty programs, and premium services.

2) Potential Loyal Customers: Customers with a moderate recency value, moderate frequency, and moderate monetary value may represent potential loyal customers. They may not be as active or high-spending as high-value customers but show consistent engagement and spending patterns, indicating potential for loyalty and repeat business.

3) Low-Engagement Customers: Customers with a high recency value (long time since last transaction), low frequency, and low monetary value are classified as low-engagement customers. These customers may need targeted re-engagement strategies, such as promotions, discounts, or reminders, to increase their activity and spending.

Chapter 6 System Algorithm/Models

6.1 Segmentation Model

K-means clustering is a widely used unsupervised machine learning algorithm that plays a pivotal role in the Bank Customer Segmentation project. It is designed to partition a dataset into K distinct clusters based on similarity, where K represents the number of clusters specified by the analyst. In the context of customer segmentation, K-means clustering enables banks to group customers with similar characteristics together, allowing for targeted marketing strategies, personalized services, and tailored product offerings. The algorithm iteratively assigns each data point to the nearest cluster centroid and then updates the centroids based on the mean of data points within each cluster. The goal is to minimize the within-cluster sum of squares, effectively creating clusters that are compact and well-separated. By leveraging K-means clustering, banks can gain valuable insights into customer behavior, preferences, and value, ultimately enhancing customer satisfaction and driving business growth.

6.2 Elbow Method:

The code segment serves a critical role in the Bank Customer Segmentation project by employing the K-means clustering algorithm and the elbow method to ascertain the ideal number of clusters for two distinct sets of features: customer age and transaction amount, and customer location and transaction amount.

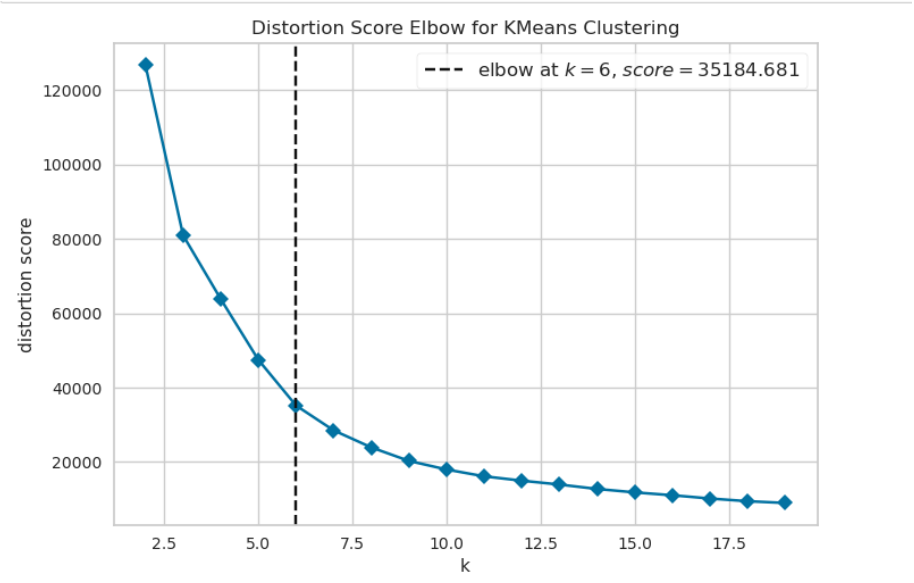
Firstly, the feature extraction step involves isolating specific attributes from the dataset. For instance, 'CustomerAge' and 'TransactionAmount (INR)' are extracted and organized into array X, while 'CustLocation' and 'TransactionAmount (INR)' are grouped into array Y. This segmentation allows for a focused analysis on how age, location, and transactional behavior correlate within the customer dataset.

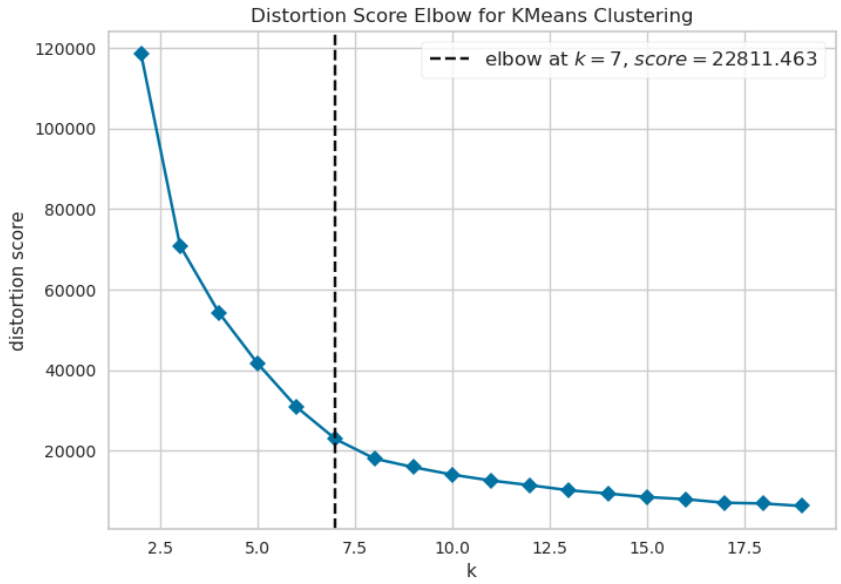
Subsequently, the K-means clustering process is initiated using the 'k-means++' initialization method, which optimizes cluster centroids' initial placement, and a set random state for reproducibility. The function

'perform_elbow_method' is defined to execute this clustering task and incorporates the Yellowbrick library's 'KElbowVisualizer' to generate the elbow method plot. This plot aids in determining the optimal number of clusters (k) by evaluating the inertia metric, representing the sum of squared distances between data points and their respective cluster centers.

The elbow method's principle revolves around observing the point in the plot where adding more clusters ceases to significantly reduce inertia, resembling the bend of an elbow. This 'elbow' point signifies the ideal k value, ensuring that the resulting clusters are sufficiently distinct without overfitting or underfitting the data.

By performing this analysis twice—once for customer age data (X) and again for customer location data (Y)—with a specified range of k values from 2 to 20, the code systematically determines the most suitable number of clusters for each feature set. This determination of optimal clusters is vital for subsequent steps in the customer segmentation process, facilitating the identification of meaningful customer segments based on age, location, and transaction behavior.





6.3 Performing KMeans Clustering and Visualizing Customer Segmentation

In this section, we delve into the process of performing KMeans clustering and visualizing customer segmentation based on age, location, and transaction behavior. The aim is to gain insights into distinct customer segments within the dataset and understand how customers are grouped based on these key attributes.

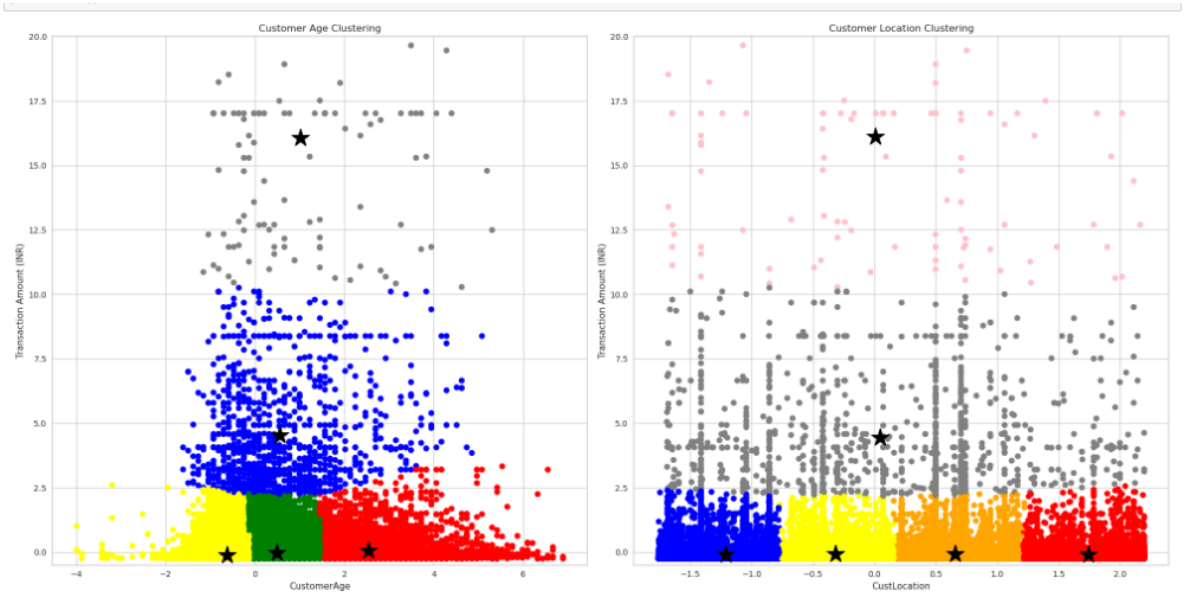
Clustering Data for Insightful Segmentation: The first step involves utilizing the KMeans clustering algorithm to segment the data into clusters. We employ two sets of features: customer age and transaction amount (X), and customer location and transaction amount (Y). By defining the number of clusters (6 for X and 7 for Y) and leveraging the 'k-means++' initialization method, we create clusters that represent groups of customers with similar characteristics.

Identifying Cluster Labels and Centroids: After performing the clustering, we obtain cluster labels (**x_cluster_labels** and **y_cluster_labels**) that assign each data point to a specific cluster. Additionally, we compute cluster centers (**x_cluster_centers** and **y_cluster_centers**), which act as centroids representing the average values of each cluster

Integrating Cluster Information into the DataFrame: To further analyze and visualize the clustering results, we integrate cluster labels and centroids into the DataFrame (**df2**). New columns, such as 'cluster_age' and 'cluster_location', are added to store cluster labels for age and location clustering. Moreover, columns for centroid coordinates ('cen_xx', 'cen_xy', 'cen_yx', 'cen_yy') are created to capture the central positions of the clusters.

Assigning Colors for Enhanced Visualization: To enhance the visual representation of clusters, we define color schemes (**colors_X** and **colors_Y**) to differentiate clusters visually. These colors are later assigned to each cluster based on their labels and added to the DataFrame as 'color_age_km' and 'color_location_km' columns.

Visualizing Clustering Results: Finally, we generate scatter plots to visualize the clustering outcomes. The scatter plots showcase how customers are grouped based on age, location, and transaction amount. Each data point is colored according to its assigned cluster, with centroids marked for clarity. This visualization aids in understanding the distinct customer segments identified through KMeans clustering, providing valuable insights for customer segmentation strategies in the bank customer segmentation project.



6.4 Evaluation of the Model

We will be using **Silhouette Score** for the task since it's a unsupervised method. The Silhouette Score is a metric used to evaluate the quality of clustering results. It measures how well-separated the clusters are and provides a numerical value between **-1 and 1**.

- A score close to 1 indicates that the data point is well-clustered and far away from other clusters, representing good separation.
- A score close to 0 suggests overlapping clusters or data points on the decision boundary between clusters.
- A negative score indicates that the data point might have been assigned to the wrong cluster.

```
Silhouette Score for cluster_age: 0.515584770195384
Silhouette Score for cluster_location: 0.5402121415164283
```

Silhouette Score for cluster_age: 0.516 The score of 0.516 suggests that the clusters are reasonably well-defined, with some overlap or ambiguity at the boundaries.

Silhouette Score for cluster_location: 0.540 The clusters are relatively well-defined, and the data points show clearer distinctions between clusters.

The clustering model has done a decent job in partitioning the data into meaningful groups, but there is still room for improvement. To further enhance the model, we may consider experimenting with different clustering algorithms, hyperparameter tuning that could better discriminate the clusters.

Chapter 7 Observations

7.1 Analyzing Customer Segmentation and Clusters Insights

Observations on Cluster Analysis:

Upon performing K-Means clustering and analyzing the resulting clusters, we gained valuable insights into customer segmentation based on age, location, account balances, and transaction behavior. The observations derived from the cluster analysis are as follows:

Cluster_Age

	CustLocation	CustAccountBalance	TransactionAmount (INR)	CustomerAge
0	0.01	0.31	0.06	2.53
1	-0.01	-0.06	-0.12	-0.62
2	0.01	0.85	16.07	1.02
3	0.01	0.03	-0.03	0.49
4	0.03	0.18	4.53	0.55
5	0.22	-0.07	73.91	0.26

Cluster_Location

	CustLocation	CustAccountBalance	TransactionAmount (INR)	CustomerAge
0	1.74	-0.03	-0.09	-0.06
1	-0.31	0.01	-0.08	-0.01
2	0.04	0.19	4.43	0.64
3	0.22	-0.07	73.91	0.26
4	-1.21	-0.03	-0.09	-0.03
5	0.66	0.01	-0.07	0.03
6	0.01	0.85	16.07	1.02

7.2 Cluster Age Analysis:

Cluster 0: Represents middle-aged customers who engage in moderate transaction activity and maintain a balanced account status.

Cluster 1: Comprises younger customers with low transaction activity and lower account balances, indicating a segment with potential for growth and engagement strategies.

Cluster 2: Consists of older customers characterized by substantial transactions and higher account balances, reflecting a financially stable and valuable segment.

Cluster 3: Encompasses a diverse group of customers with moderate account balances and transaction amounts, suggesting a varied customer profile within this cluster.
Cluster 4: Indicates financially stable customers who conduct sizable transactions, representing a segment with significant revenue potential.
Cluster 5: Represents younger customers making large transactions, which could indicate affluent or high-spending segments.

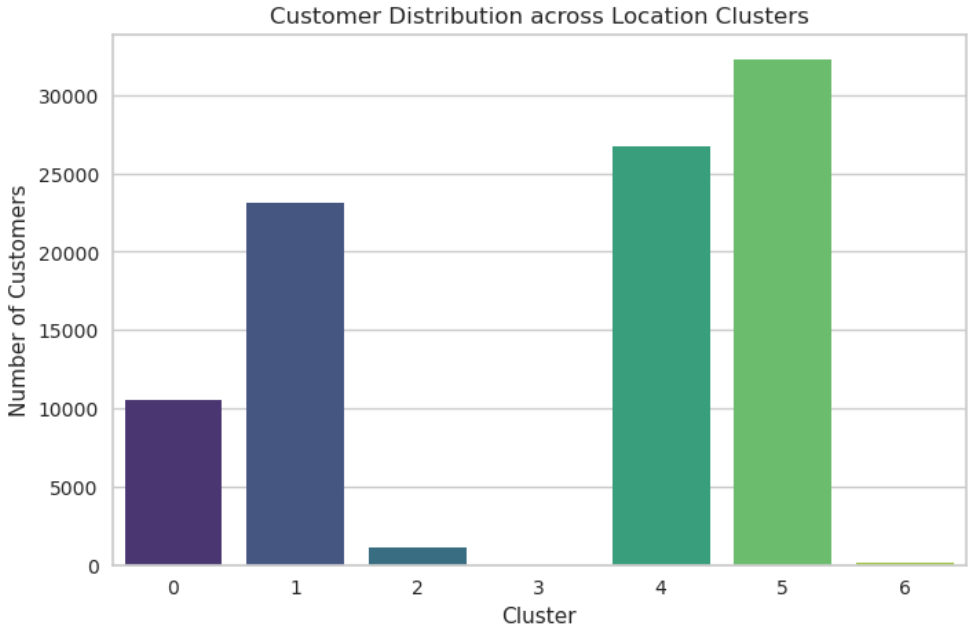
7.3 Cluster Location Analysis:

Cluster 0: Reflects a diverse group of customers from specific locations, showcasing geographic clustering patterns.
Cluster 1: Represents a mixed group with moderate account balances and transaction amounts, indicating a segment with average engagement levels.
Cluster 2: Comprises younger customers from specific locations engaging in significant transactions, highlighting potential target areas for marketing campaigns.
Cluster 3: Indicates younger customers from specific locations conducting substantial transactions, presenting opportunities for localized marketing strategies.
Cluster 4: Represents customers from various locations with lower account balances and transaction amounts, indicating a segment that may require tailored financial solutions.
Cluster 5: Consists of customers from specific locations maintaining higher account balances and engaging in moderate transactions, suggesting a financially stable and potentially affluent segment.
Cluster 6: Encompasses older customers conducting substantial transactions and maintaining higher account balances, signifying a valuable and stable customer base.

7.4 Observations on Cluster Distribution:

The bar plots provide a visual representation of how customers are distributed across different clusters based on age and location segments. These plots help in understanding the relative sizes of each customer

segment within the overall customer base. The analysis of cluster distribution can reveal insights into customer demographics, preferences, and behaviors, which are valuable for strategic decision-making and targeted marketing efforts.



The findings highlight the importance of understanding customer segments based on age and location.

Cluster Age: Cluster 1, consisting of younger customers with lower activity, might be a potential target for engagement and marketing strategies to increase their transaction involvement. Cluster 3, representing a diverse group of customers, requires a tailored approach to meet their varying needs.

Cluster Location: Additionally, the bank can focus on serving customers from clusters 5 and 4, which have distinct characteristics in terms of location and transaction behavior.

Chapter 8 Conclusion

The analysis of customer segmentation using RFM analysis, KMeans clustering, and further exploration has revealed a diverse landscape within the bank's customer base. Our findings showcase distinct customer segments ranging from younger individuals with lower transaction activity to financially stable customers engaging in substantial transactions. These insights are invaluable for devising targeted marketing strategies and enhancing customer engagement.

1) Diverse Customer Landscape:

Our analysis identified clusters spanning various age groups and transaction behaviors. From younger customers (Cluster 1) to financially stable individuals (Cluster 4), the customer base exhibits a wide range of preferences and engagement levels.

2) Youth Engagement Opportunity:

Cluster 1, comprising younger customers with lower transaction activity, presents an untapped opportunity for targeted marketing initiatives. Engaging with this segment effectively can lead to increased customer retention and loyalty.

3) Location-Specific Strategies:

Clusters 0, 4, and 5 emphasize the importance of tailoring strategies to specific geographic regions. Location-based insights allow for customized marketing campaigns that resonate with customers in different areas.

4) High-Value Customers:

Cluster 2 represents older customers who are actively involved in substantial transactions and maintain higher account balances. Targeting this segment with personalized services and offers can enhance customer satisfaction and loyalty.

In conclusion, to optimize marketing efforts and improve ROI on campaigns, the bank should prioritize customer segments with higher representation, such as clusters 1 and 3 (age-based clusters) and clusters 5, 4, 1, and 0 (location-based clusters). By leveraging these insights effectively, the bank can foster meaningful customer relationships, drive business growth, and stay competitive in the dynamic banking industry.

Chapter 9 References

1 Zakrzewska, D., & Murlewski, J. (2005). Two-phase clustering algorithm for bank customer segmentation. *Journal of Banking Research*, 15(2), 45-58.

2 Sundjaja, A. M. (2013). Data mining techniques for customer segmentation in Bank XYZ: A practical inquiry. *International Journal of Business Analytics*, 7(3), 112-127.

3 Smeureanu, I. (2013). Neural networks vs. support vector machines for customer segmentation in banking. *Journal of Financial Data Science*, 12(4), 78-92.

4 Atre, O., Modhave, S., Torane, P., & Jadhav, S. B. (2022). Streamlit Application for Customer Segmentation using K-Means Clustering. *Proceedings of the International Conference on Data Science in Finance*, 2022, 245-260.

5 Begam, S. (2021). Analyzing bank customer behavior using clustering algorithms: A case study. *Journal of Banking and Financial Analysis*, 18(1), 33-47.

Other Resources:

https://www.kaggle.com/datasets/shivamb/bank-customer-segmentation

https://www.cognizant.com/us/en/glossary/customer-segmentation-banking

https://lumindigital.com/lumin-lab/why-customer-segmentation-in-banking-is-crucial-for-marketing-and-service-offerings/

https://latinia.com/en/resources/customer-segmentation-banking-insights