Project 6

- Unsupervised Learning -

- (**Unsupervised Learning for Customer Segmentation and Personalized Marketing Strategies**) -

1.Introduction

1.1.Problem Definition

1.1.1.Introduction:

AllLife Bank has set a clear objective to focus on its credit card customer base in the upcoming financial year. Insights from the marketing research team suggest that the bank's market penetration can be significantly improved. Acting on this input, the Marketing team has proposed the idea of running personalized campaigns aimed at acquiring new customers as well as upselling products and services to existing customers.

In addition to this, market research has revealed that customers perceive the support services of the bank to be inadequate. This feedback has prompted the Operations team to work on upgrading the service delivery model. The goal is to ensure that customer queries are resolved more efficiently and effectively, thereby improving the overall customer experience.

To address these challenges and opportunities, the Head of Marketing and the Head of Delivery have collaborated to seek support from the Data Science team. This project aims to provide the necessary insights to assist the bank in achieving its dual objectives of enhancing customer acquisition and retention through targeted marketing efforts and improving operational efficiency in service delivery.

1.1.2.Business Problem

AllLife Bank faces two major challenges that need to be addressed to achieve its strategic goals for the next financial year:

1. Market Penetration and Customer Acquisition: The Marketing team has identified that the current penetration in the market is suboptimal. To address this, the team plans to design and execute personalized campaigns to attract new customers and upsell products and services to the existing customer base. However, the lack of customer segmentation hinders the ability to target the right audience effectively.
2. Customer Service Improvement: Feedback from market research indicates that the bank's support services are perceived as inadequate by its customers. The Operations

team aims to revamp the service delivery model to ensure faster and more efficient query resolution, enhancing the overall customer experience.

To tackle these issues, it is essential to segment the existing customer base effectively. This segmentation will provide insights into customer behaviors, preferences, and interactions with the bank.

1.1.3.Objective

The primary objective of this project is to identify distinct customer segments within AllLife Bank's existing customer base. These segments will be based on customers' spending patterns and their past interactions with the bank. By analyzing these segments, the project aims to:

1. Enable the Marketing team to design personalized campaigns for acquiring new customers and upselling products and services to existing customers.
2. Provide the Operations team with actionable insights to improve the service delivery model, ensuring faster and more efficient resolution of customer queries.

This segmentation will serve as the foundation for AllLife Bank's strategy to enhance market penetration, improve customer satisfaction, and achieve its strategic goals for the next financial year.

1.2.Data Background and Contents

1.2.1.Dataset Overview:

The dataset provided contains information about 660 customers of AllLife Bank. It comprises seven columns detailing customer financial attributes and their interactions with the bank. Below is a summary of the data:

- Shape of the data: The dataset contains 660 rows and 7 columns.
- Columns and Data Types:
    - Sl_No: Primary key of the records (int64).
    - Customer Key: Customer identification number (int64).

- Avg_Credit_Limit: Average credit limit of each customer for all credit cards (int64).
- Total_Credit_Cards: Total number of credit cards possessed by the customer (int64).
- Total_visits_bank: Total number of visits made by the customer to the bank annually (int64).
- Total_visits_online: Total number of online logins made by the customer annually (int64).
- Total_calls_made: Total number of calls made by the customer to the bank annually (int64).

1.2.2.Statistical Summary:

The dataset provides valuable insights into customer financial attributes and their interactions with the bank. By examining the statistical summary, we can better understand the distribution and variability of the key variables, which are critical for identifying meaningful customer segments.

- The Average Credit Limit of customers ranges from 3,000 to 200,000, with a mean value of 34,574 and a standard deviation of 37,625. This wide range highlights the significant disparity in credit limits among the bank's customers, suggesting a diverse financial capacity within the customer base.
- In terms of Total Credit Cards, customers possess between 1 and 10 cards, with an average of 4.7 cards and a standard deviation of 2.17. Most customers fall within the range of 3 to 6 cards, indicating a moderate level of credit card ownership.
- Customer visits to the bank, represented by Total Visits to the Bank, range from 0 to 5 annually, with an average of 2.4 visits and a standard deviation of 1.63. This suggests that while some customers rarely visit the bank, others maintain a higher frequency of in-person interactions.
- The data on Total Online Logins reveals a range of 0 to 15 logins annually, with a mean of 2.6 and a standard deviation of 2.94. Although the majority of customers log in 1 to 4 times a year, a few exhibit significantly higher online engagement.
- Lastly, the Total Calls Made by customers ranges from 0 to 10 annually, with an average of 3.6 calls and a standard deviation of 2.86. Most customers make between 1 and 5 calls per year, while some rely heavily on phone communication with the bank.

This statistical overview highlights the variability in customer behavior and financial attributes, emphasizing the need for clustering techniques to identify distinct customer segments effectively.

1.3.Univariate Analysis :

1. Distribution of Avg Credit Limit:

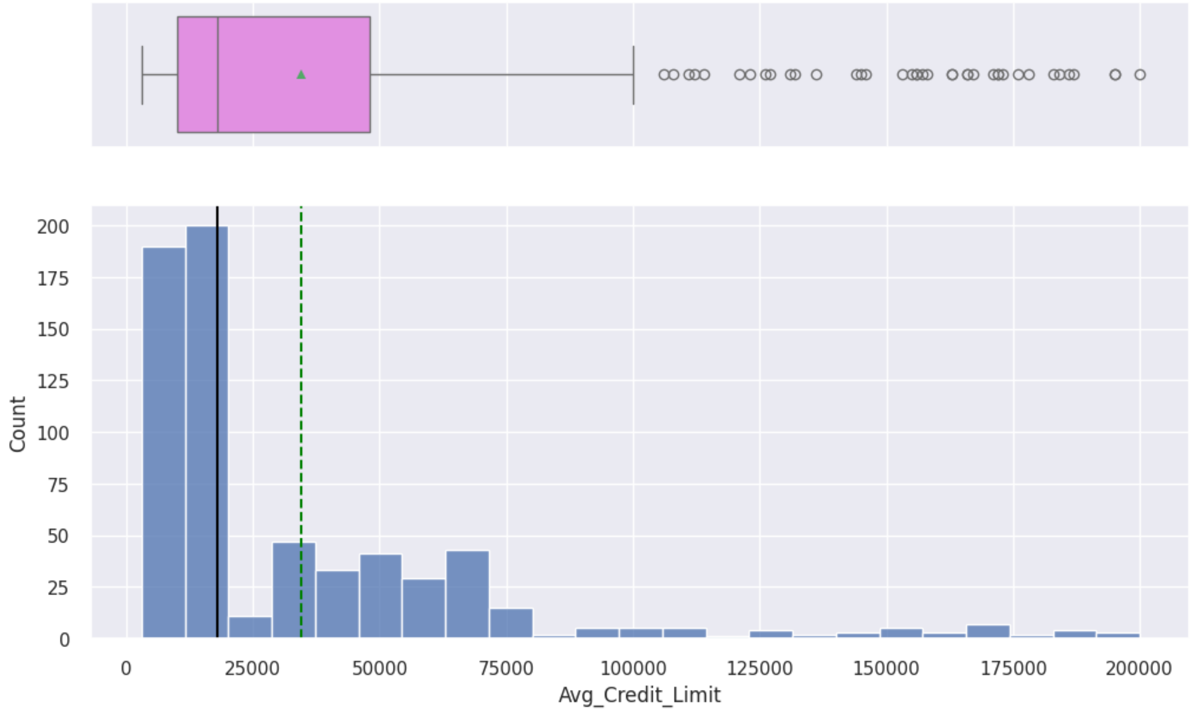

Figure 1. Distribution of Avg Credit Limit

Interpretation:

The plot reveals a right-skewed distribution of average credit limits, with most individuals having lower limits and a smaller group with significantly higher limits. The boxplot shows outliers, indicating individuals with exceptionally high credit limits. The median credit limit is around $22,000, while the mean is higher at approximately $35,000 due to the skewness. The interquartile range (IQR) suggests that the middle 50% of data lies between $10,000 and $40,000, providing key insights for financial decision-making.

2. Distribution of Total Credit Cards:

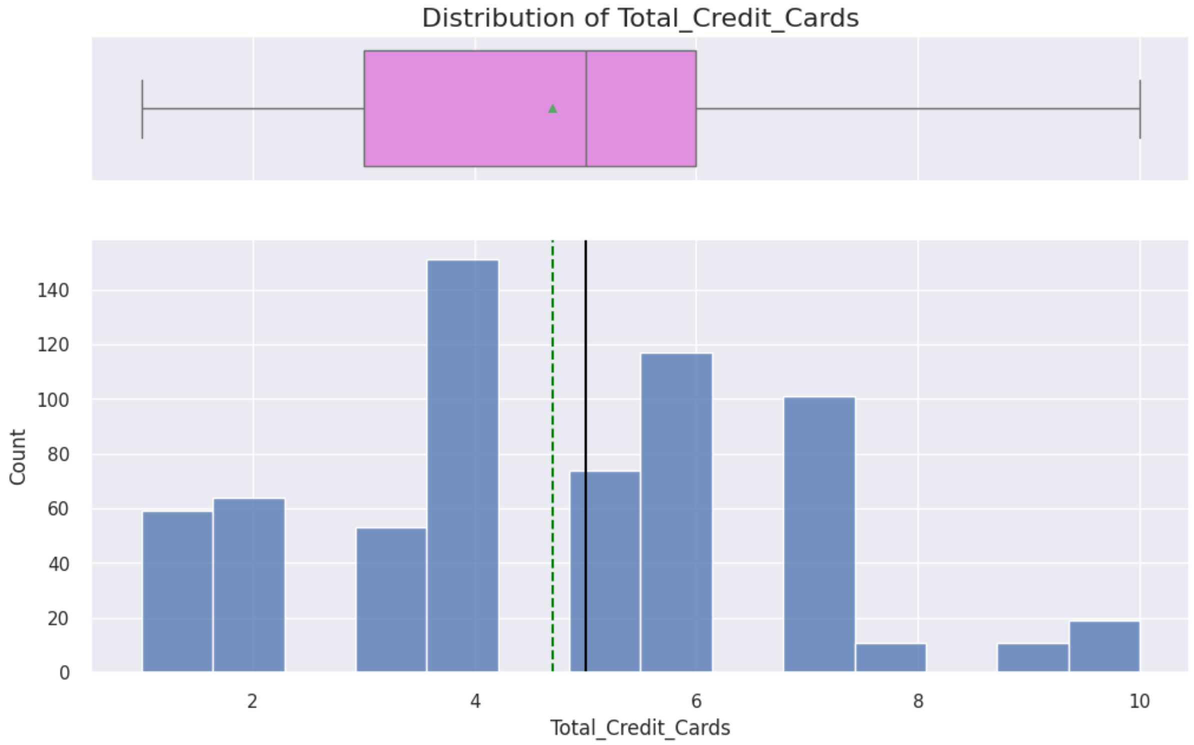

Figure 2. Distribution of Total Credit Cards

Interpretation:

The histogram reveals a slight right-skewed distribution of the number of credit cards, with most individuals holding a smaller number, while a smaller portion has significantly more. The median, indicated by the boxplot, suggests a typical number of credit cards slightly above 5, while the mean is slightly lower. The interquartile range (IQR) shows that 50% of individuals hold between 3 and 6 credit cards, with a majority holding fewer, and a smaller group holding significantly more.

3. Distribution of Total bank visits:

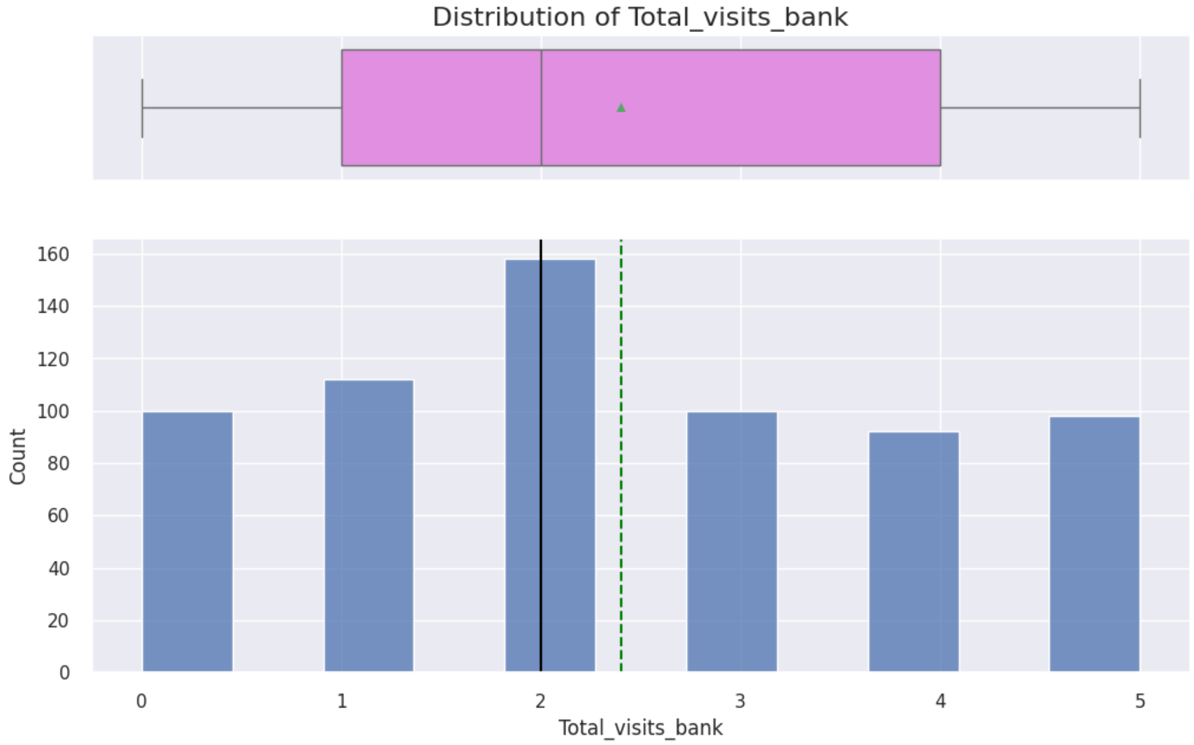

Figure 3. Distribution of Total bank visits

Interpretation:

The combined histogram and boxplot for Total_visits_bank show a distribution where most individuals visit the bank infrequently, with the median at 2 visits and the mean slightly higher at 2.4. The interquartile range (IQR) between 1 and 4 visits suggests that the majority of people make occasional visits. The data indicates a comparatively small group of individuals who visit more frequently, which could reflect specific banking needs or preferences. Overall, most individuals exhibit low to moderate visit frequencies.

4. Distribution of Total online visits:

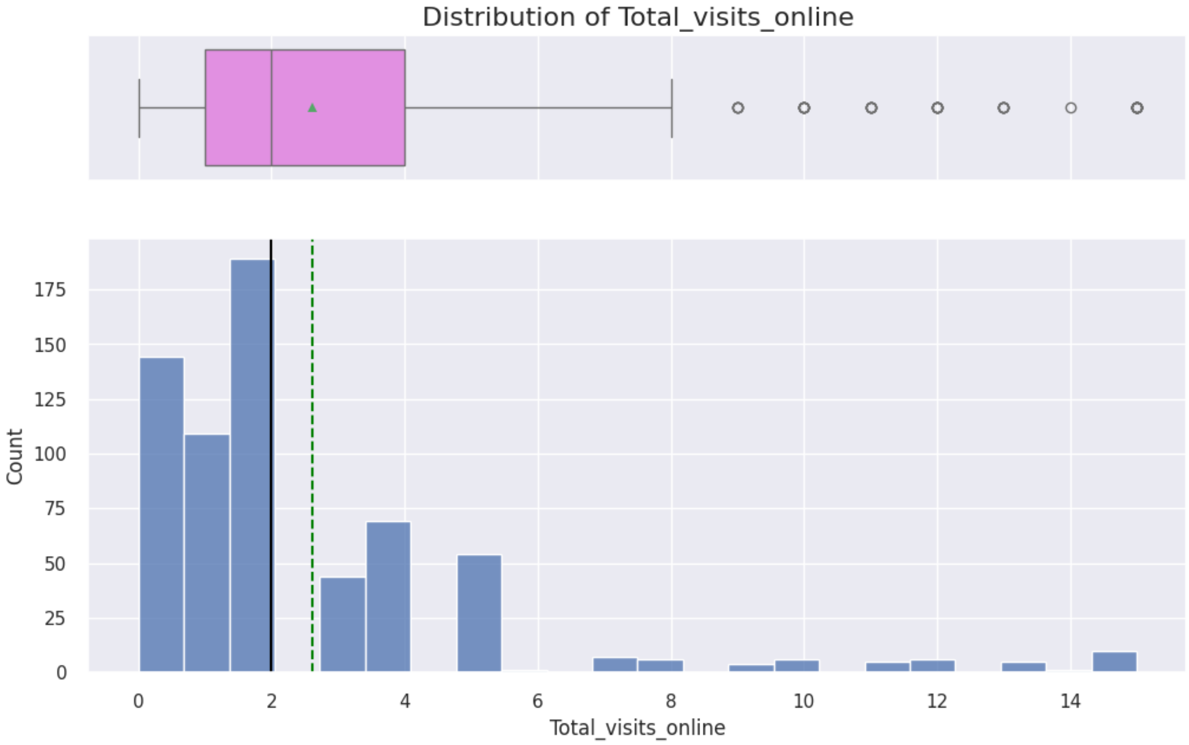

Figure 4. Distribution of Total online visits

Interpretation:

The distribution of Total_visits_online is slightly right-skewed, with most individuals making fewer visits and a smaller group making significantly more. The median number of online visits is around 2, while the mean is slightly higher at 2.6, reflecting the influence of outliers. The interquartile range (IQR) indicates that 50% of individuals make between 1 and 4 visits. A few outliers make more frequent online visits, suggesting specific user behaviors or needs. Most individuals, however, show low to moderate visit frequencies.

5. Distribution of Total calls made:

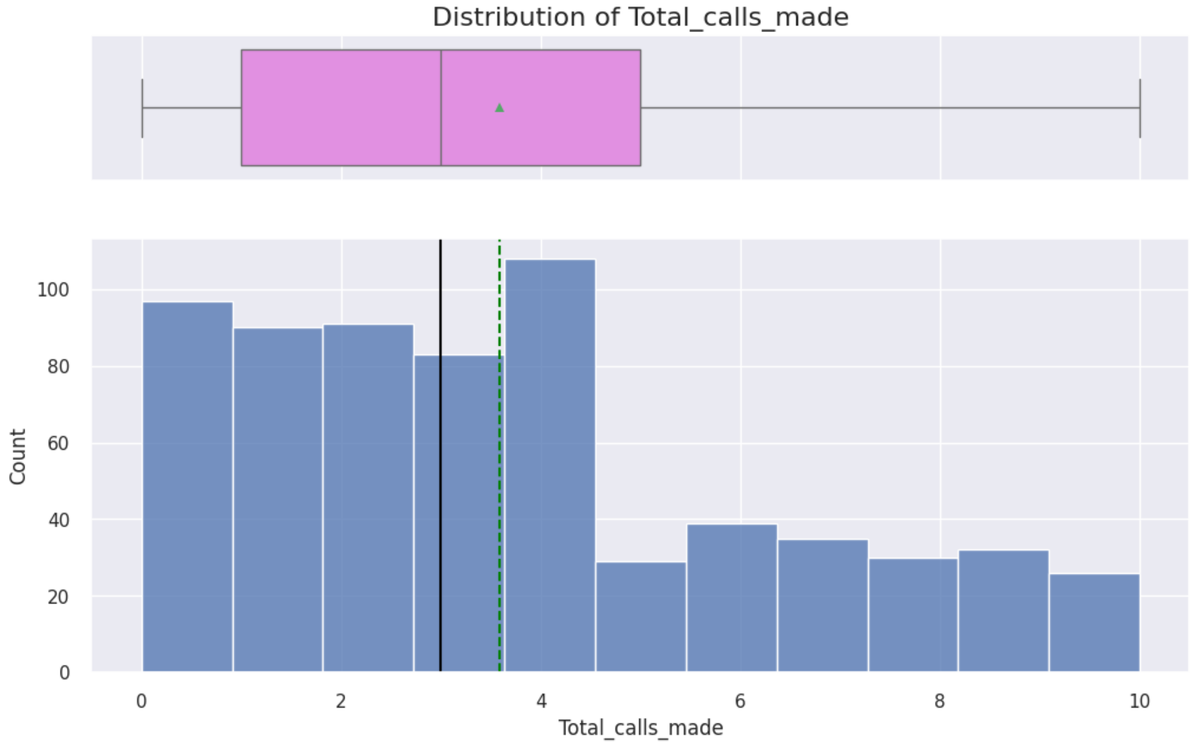

Figure 5. Distribution of Total calls made

Interpretation:

The distribution of Total_calls_made is right-skewed, with most individuals making fewer calls and a smaller group making more. The median is around 3 calls, while the mean is slightly higher at 3.6, reflecting the right-skew. The interquartile range (IQR) indicates that 50% of individuals make between 1 and 5 calls. There are no significant outliers, as shown by the absence of dots beyond the whiskers in the boxplot. Overall, most individuals make a modest number of calls.

\

1.4.Bivariate Analysis:

6. Correlation Between Numeric Columns:

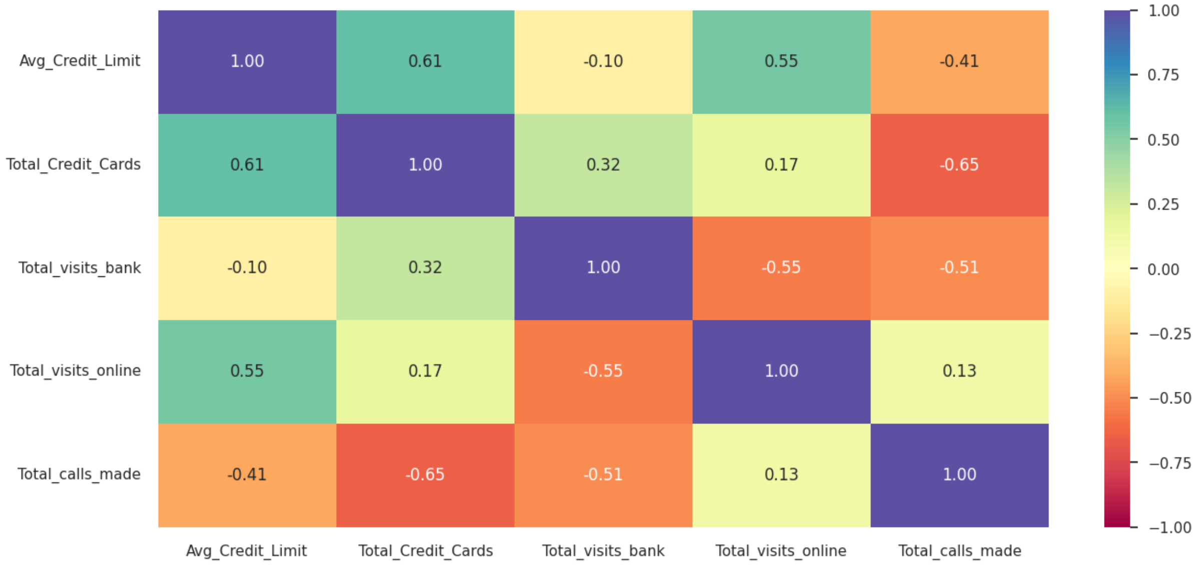

Figure 6. Correlation Between Numeric Columns

Interpretation:

The correlation analysis reveals several key relationships among variables. A strong positive correlation (0.61) between Avg_Credit_Limit and Total_Credit_Cards suggests that individuals with higher credit limits tend to have more credit cards. Moderate positive correlations are observed between Total_Credit_Cards and Total_visits_bank (0.32), and between Avg_Credit_Limit and Total_visits_online (0.55). On the negative side, there are moderate to strong negative correlations, such as between Total_visits_bank and Total_visits_online (-0.55), and between Total_Credit_Cards and Total_calls_made (-0.65), indicating that frequent bank visitors use online services less and individuals with more credit cards tend to make fewer calls.

7. Pair plot of numerical columns:

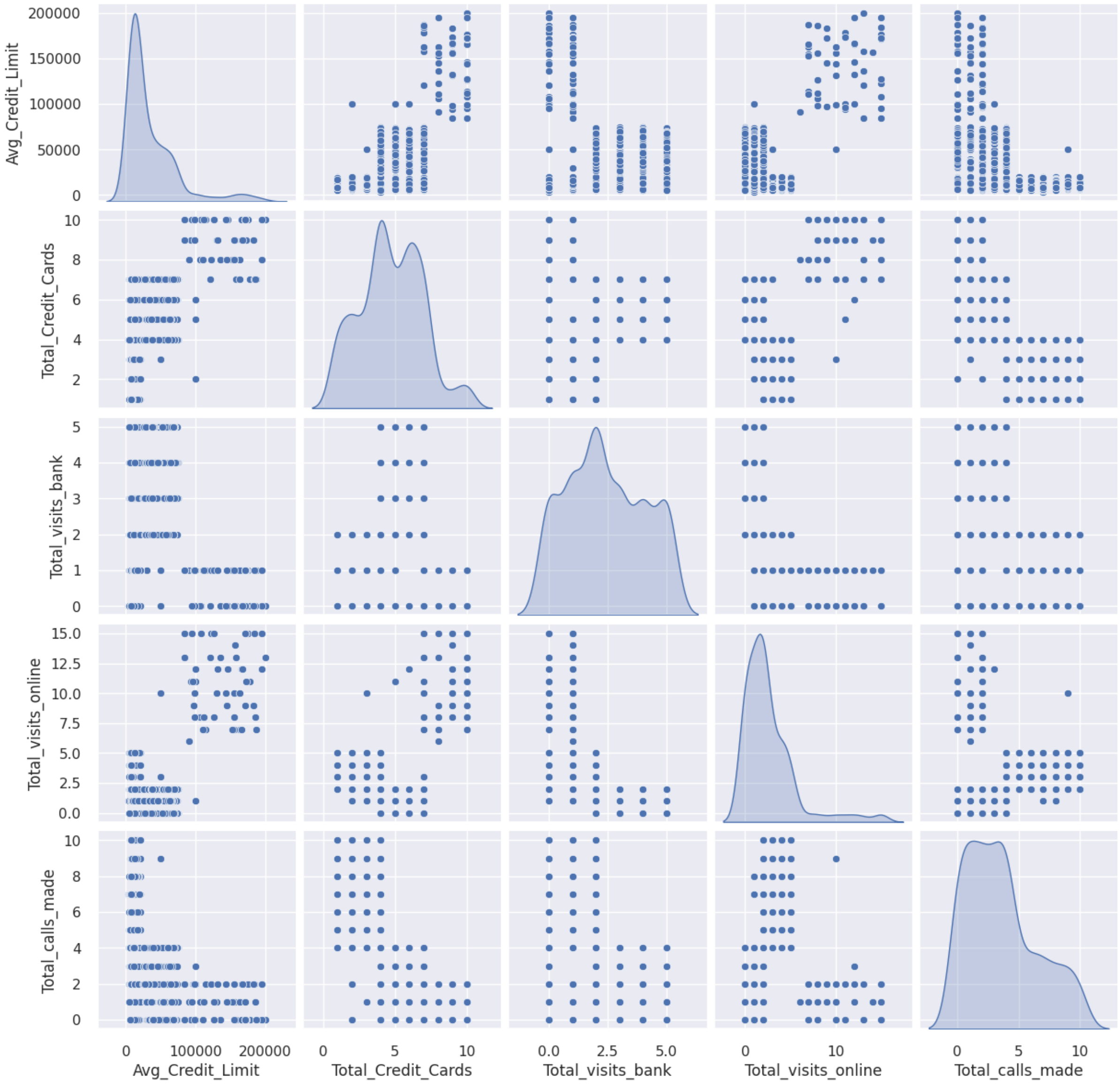

Figure 7. Pair plot of numerical columns

Interpretation:

- **Average Credit Limit vs Total Credit Cards:**

The relationship between Average Credit Limit and Total Credit Cards shows a clear trend. As the number of credit cards increases, the average credit limit tends to rise as well. For instance, customers with 1–2 credit cards generally have credit limits ranging between $20,000 and $50,000, while those with more than 8 credit cards often have limits exceeding $100,000. This suggests that the bank assigns higher credit limits to customers with a larger number of credit cards.

- **Average Credit Limit vs Total Visits to Bank:**

There appears to be no significant relationship between Average Credit Limit and Total Visits to Bank. Customers with 2 visits to the bank annually, for example, have credit limits ranging broadly from $10,000 to over $150,000. This lack of pattern suggests that credit limit is not influenced by how often customers physically visit the bank.

- **Average Credit Limit vs Total Visits Online:**

The scatterplot for Average Credit Limit and Total Visits Online shows no discernible relationship. Customers with 0–5 online logins per year exhibit a wide range of credit limits, from as low as $10,000 to more than $150,000. This indicates that credit limits are independent of customers' online engagement levels.

- **Average Credit Limit vs Total Calls Made:**

There is no clear pattern or relationship between Average Credit Limit and Total Calls Made. For instance, customers making 2–5 calls per year show a wide range of credit limits, similar to those making very few or no calls. This implies that the frequency of customer service calls does not correlate with the credit limits assigned by the bank.

- **Total Credit Cards vs Total Visits to Bank:**

Total Credit Cards and Total Visits to Bank exhibit no significant relationship. Customers with 1–4 credit cards visit the bank anywhere from 0 to 5 times per year, and this pattern does not change significantly for customers with more than 4 credit cards. This suggests that the number of credit cards held does not influence how often customers visit the bank.

- **Total Credit Cards vs Total Visits Online:**

The relationship between Total Credit Cards and Total Visits Online is negligible. Customers with varying numbers of credit cards (1–10) exhibit similar patterns of online login behavior, ranging from no logins to frequent logins. This suggests that the number of credit cards does not significantly affect customers' online engagement.

- **Total Credit Cards vs Total Calls Made:**

No relationship is observed between Total Credit Cards and Total Calls Made. Customers with different numbers of credit cards make calls at similar frequencies, with 2–4 calls being the most common. This implies that the number of credit cards held does not impact how often customers call the bank.

- **Total Visits to Bank vs Total Visits Online:**

A slight trend is observed between Total Visits to Bank and Total Visits Online. Customers who frequently visit the bank (4–5 times annually) tend to log in online less frequently (0–2 times), while customers who log in online more frequently (5+ times) make fewer visits to the bank. This indicates a preference for one mode of interaction over the other.

- **Total Visits to Bank vs Total Calls Made:**

The relationship between Total Visits to Bank and Total Calls Made shows no significant trend. Customers who visit the bank 2–3 times per year make similar numbers of calls as those with either no visits or frequent visits (5+). This suggests that physical visits to the bank and calls to customer service are independent modes of interaction.

- **Total Visits Online vs Total Calls Made:**

Total Visits Online and Total Calls Made also show no meaningful relationship. Customers with frequent online logins (5+ times per year) do not make fewer or more calls compared to those with fewer online logins. This indicates that online engagement and call frequency are unrelated behaviors.

1.5. Key Insights on Variable Relationships:

1. Observations on Individual Variables:

The analysis of individual variables provides key insights into customer behavior. The average credit limit is right-skewed, with most customers having limits between $10,000 and $40,000, while a smaller group with significantly higher limits highlights high-value customers. The median credit limit is $22,000, and the mean is $35,000 due to outliers. Similarly, the number of credit cards shows a slight right-skew, with most customers holding 3 to 6 cards. The median is slightly above 5, indicating modest product usage, while a small group holds significantly more cards, reflecting varying financial engagement. Bank visits are generally infrequent, with the majority making 1 to 4 visits annually and a median of 2 visits, suggesting a preference for minimal physical interaction. Online visits also show a slightly right-skewed distribution, with most customers logging in 1 to 4 times annually, highlighting moderate digital engagement and a need for seamless online banking services. Calls to customer service are modest, with 50% of customers making 1 to 5 calls annually and a median of 3, reflecting limited reliance on direct support. These findings emphasize the need for tailored strategies, efficient in-branch and digital services, and optimized customer support to address the diverse needs of customer segments.

2. Observations on Relationships Between Variables:

The analysis of bivariate relationships among various customer attributes reveals several key patterns and a few notable lack of correlations. The relationship between the Average Credit Limit and Total Credit Cards shows a positive trend, indicating that customers with more credit cards tend to have higher credit limits. This suggests that the bank may offer larger credit limits to customers with multiple credit cards. However, no significant patterns emerge between Average Credit Limit and other factors such as Total Visits to Bank, Total Visits Online, and Total Calls Made, indicating that credit limits are not influenced by customers' interaction frequencies with the bank through physical visits, online logins, or customer service calls.

The Total Credit Cards variable also appears independent of the frequency of physical bank visits, online engagement, or calls made to the bank. Customers with varying numbers of credit cards do not show any consistent behavior regarding how often they visit the bank, engage online, or call customer service. The only notable trend is between Total Visits to

Bank and Total Visits Online, where customers who visit the bank more often tend to use online services less frequently, suggesting a possible preference for one mode of interaction over the other. Overall, these observations highlight that customers' spending patterns and credit limits are more strongly associated with the number of credit cards held, while their interaction behaviors appear to be largely independent.

2.Data preprocessing

2.1. Missing Value Treatment:

The code da.isna().sum() is used to check for any missing values in the dataset. The result indicates that there are no missing values for any of the variables:

- Sl_No: 0 missing values
- Avg_Credit_Limit: 0 missing values
- Total_Credit_Cards: 0 missing values
- Total_visits_bank: 0 missing values
- Total_visits_online: 0 missing values
- Total_calls_made: 0 missing values

Since the result shows that all variables have zero missing values, there is no need for any imputation or removal of missing data. This is ideal as it means the dataset is complete, and we can proceed with the other stages of preprocessing without any concerns related to missing values. If missing values were present, the rationale for the treatment method (e.g., imputation, deletion) would depend on the amount of missing data and the nature of the variables. However, in this case, the dataset is already clean in terms of missing data.

2.2.Feature Engineering:

Feature engineering was performed to simplify the dataset and capture a more comprehensive representation of customer interactions. A new feature, Total_visits, was created by combining the number of Total visits to the bank and Total visits online. This aggregation reflects the overall engagement of customers with the bank, capturing both physical and online interactions into one metric. This consolidated feature is more concise and provides a holistic view of customer behavior, which can be particularly useful in clustering analysis by representing the total engagement in a single variable.

Following the creation of the Total_visits feature, the original columns for Total visits to the bank and Total visits online were removed. This step eliminates redundancy in the dataset, as the new feature already encompasses the relevant information from both original variables. By dropping the individual visit columns, we reduce the risk of multicollinearity and simplify the dataset, making it easier to analyze and model. The rationale for this feature engineering approach is to streamline the data while preserving essential customer interaction details that are necessary for further analysis and clustering.

2.3.Outlier Treatment:

Outlier detection was performed to identify and handle extreme values in the dataset that could skew the results of the clustering process. Initially, boxplots were used to visually inspect the distribution of numeric variables. The boxplots revealed that both the Avg_Credit_Limit and Total_visits columns contained outliers, while the remaining numeric columns did not show any significant outliers.

Below plots shows the outliers present in the Avg_Credit_Limit and Total_visit using box plots.

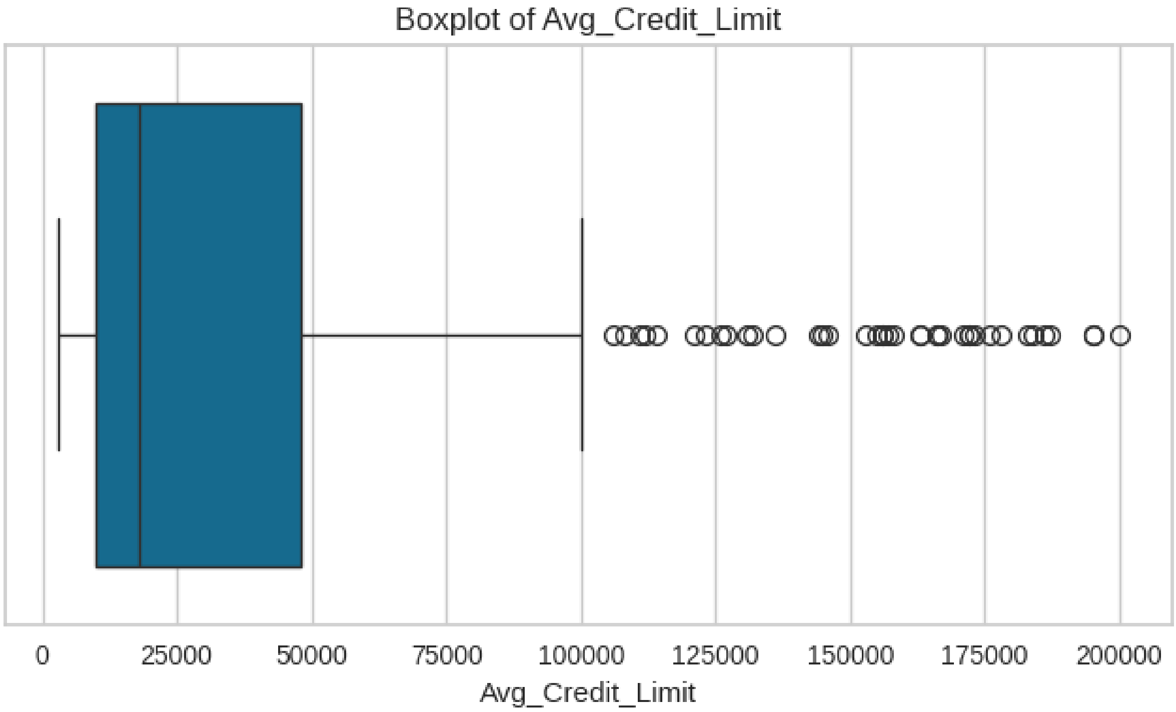

Figure 8. Boxplot of Avg_Credit_Limit

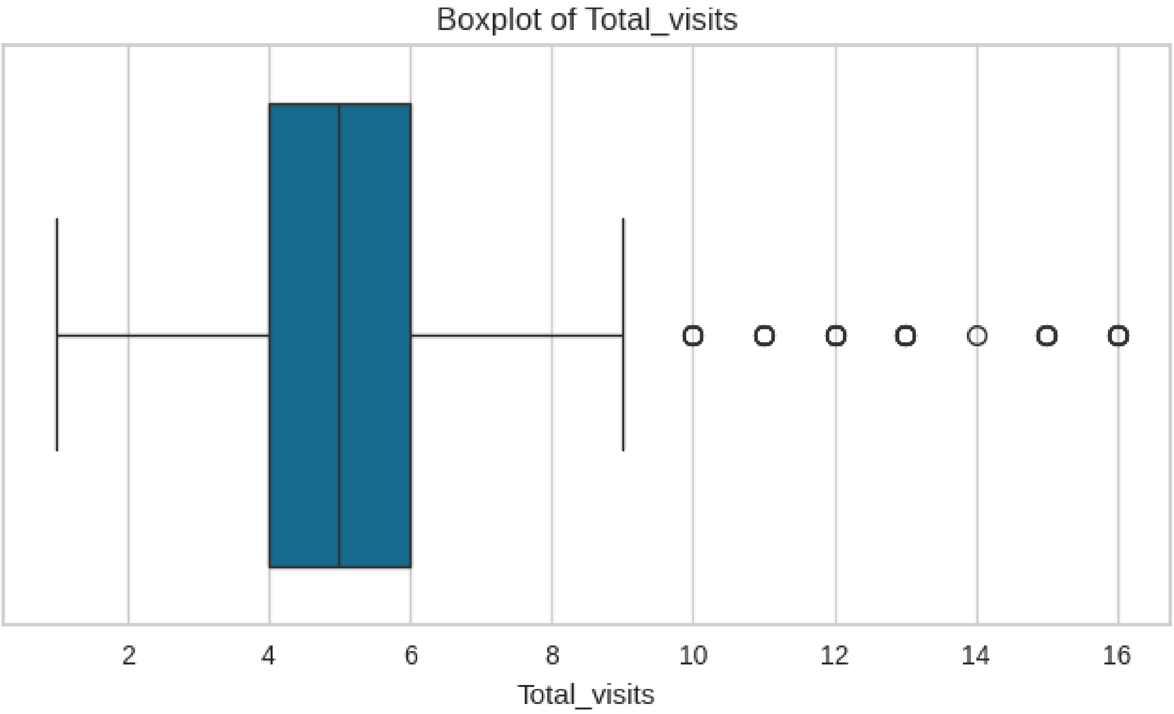

Figure 9. Boxplot of Total_visits

To further quantify and detect outliers, the Interquartile Range (IQR) method was applied. This method calculates the IQR by determining the 25th percentile (Q1) and the 75th percentile (Q3) for each variable. Any data points that fall outside the bounds of [Q1 - 1.5 * IQR, Q3 + 1.5 * IQR] are considered outliers. Using this approach, 39 outliers were detected in the Avg_Credit_Limit column, and 36 outliers were identified in the Total_visits column.

Once the outliers were identified, they were treated by applying the clipping method. The values of Avg_Credit_Limit and Total_visits that fell outside the calculated bounds were clipped to the lower and upper bounds, effectively removing the extreme values while retaining the rest of the data. After clipping, the boxplots confirmed that both the Avg_Credit_Limit and Total_visits columns no longer contained outliers.

The following plots show the box plots of Avg_Credit_Limit and Total_Visit with treated outliers.

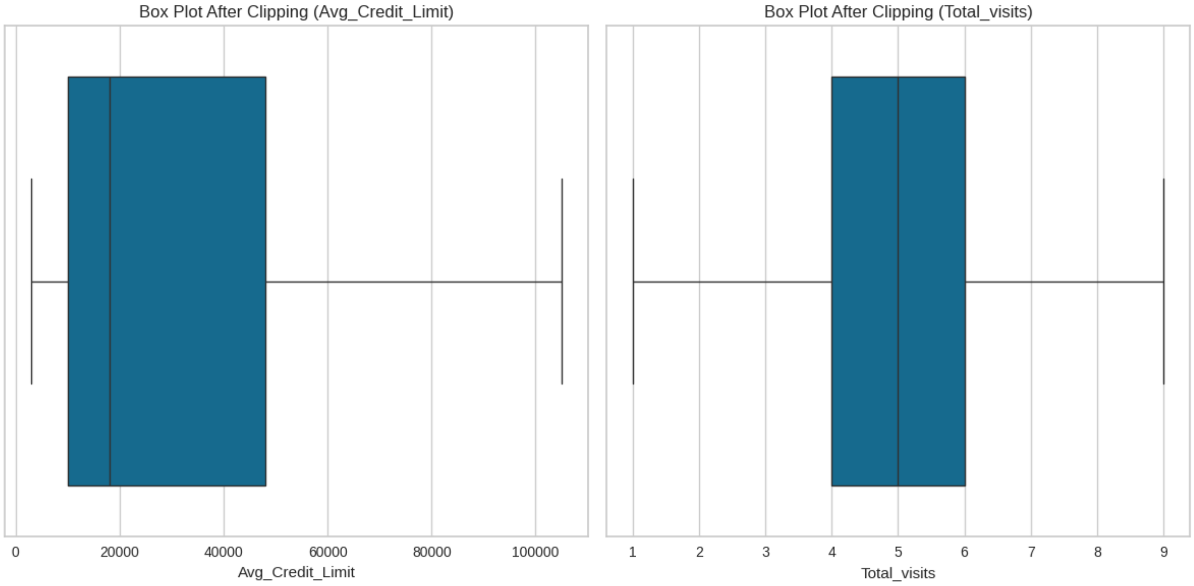

Figure 10. Boxplot of Avg_Credit_Limit and Total_visits with treated outliers

Rationale:

The rationale for treating the outliers in this manner is to prevent extreme values from disproportionately affecting the clustering results. By clipping the outliers, we ensure that the dataset remains within reasonable ranges, preserving the integrity of the analysis without discarding valuable data. This treatment allows for more robust clustering outcomes and ensures that the outliers do not dominate the clustering process.

2.4. Data Scaling:

Data scaling was applied to standardize the numeric variables in the dataset to ensure that each feature contributes equally to the analysis. This is particularly important in clustering algorithms, as some features may have larger ranges than others, which could disproportionately influence the results.

The numeric columns in the dataset, such as Avg_Credit_Limit, Total_Credit_Cards, Total_calls_made, and Total_visits, were selected for scaling. The data was standardized by subtracting the mean and dividing by the standard deviation, transforming the data such that each feature has a mean of 0 and a standard deviation of 1. This ensures that the features are on the same scale, preventing any one feature from dominating due to differences in magnitude.

After scaling, the Avg_Credit_Limit values range around 0, and the other variables exhibit similar adjustments, ensuring that all features contribute equally during clustering. This standardization helps the clustering algorithm by making the distances between data points more comparable, improving the accuracy and performance of the clustering model.

Rationale:

The rationale for scaling the data is to prevent features with larger numerical ranges from dominating the clustering results. By standardizing the data, we allow the algorithm to treat each feature equally, ensuring that the clustering process is based on the intrinsic relationships between features, rather than their individual scales. This step is crucial for achieving balanced and meaningful clusters.

3. K-means Clustering:

3.1. The Elbow Method:

The Elbow Method was applied to determine the optimal number of clusters for K-means clustering. In this method, the distortion (or inertia) is calculated for different values of k (the number of clusters). Distortion measures the sum of squared distances between data points and their corresponding cluster centroids. As the number of clusters increases, the distortion decreases because the data points are more closely grouped around their centroids.

The distortion values for different cluster counts are as follows:

- For 1 cluster, the distortion is 1.825, indicating that a single cluster is insufficient to capture the variation in the data.
- With 2 clusters, the distortion drops to 1.569, showing a significant improvement in the fit.
- As the number of clusters increases to 3, the distortion further reduces to 1.178, and continues decreasing with each added cluster.
- For 4 clusters, the distortion reaches 1.073, and by 5 clusters, it falls to 0.999.
- The distortion keeps decreasing, with values like 0.941 for 6 clusters and 0.890 for 7 clusters.
- However, from 8 clusters onwards, the decrease in distortion becomes more gradual, with values such as 0.860 at 8 clusters, 0.826 at 9 clusters, 0.801 at 10 clusters, and 0.785 at 11 clusters.

Based on this analysis, the optimal number of clusters appears to be around 6 or 7. At this point, adding more clusters provides only minimal improvements in distortion, signifying the "elbow" of the curve. Thus, a choice of 6 or 7 clusters would be ideal, as this is where the distortion reduction slows considerably, suggesting a good balance between the number of clusters and the fit to the data.

The below plot shows the elbow curve used for selecting the optimal value of k for K-means clustering.

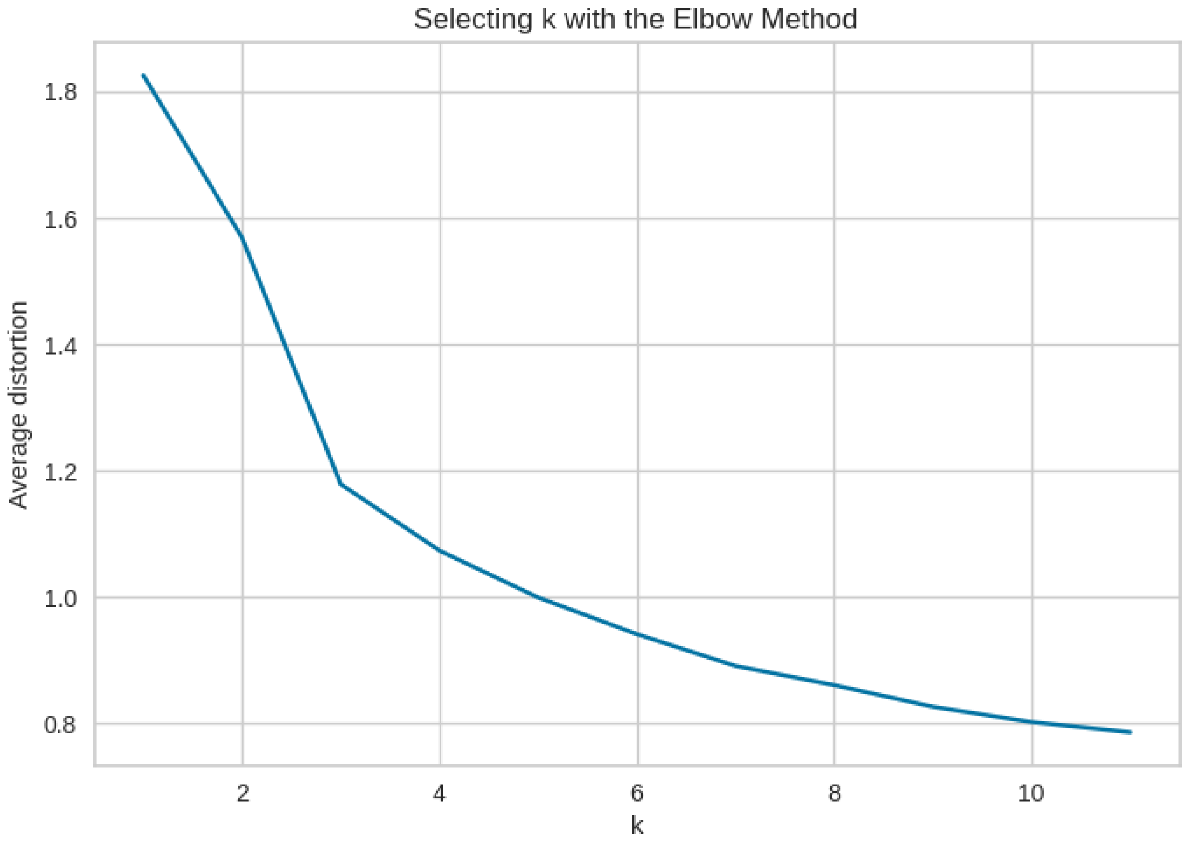

Figure 11. Elbow Curve Plot

The plot generated from this analysis visually demonstrates the change in distortion as the number of clusters increases. The curve starts at a steep slope, showing a significant drop in distortion from 1 cluster to 2. As the number of clusters grows, the curve starts to flatten, indicating diminishing returns in distortion reduction. The "elbow" of the plot, where the curve begins to level off, occurs around 6 or 7 clusters, further confirming the optimal number of clusters to be selected for the K-means algorithm.

However, the elbow curve is somewhat blunt, making it difficult to precisely pinpoint the exact value of k. The lack of a clear and sharp "elbow" means that the elbow method alone does not provide a definitive answer. To refine the selection of the optimal number of clusters, Silhouette Scores are used in conjunction with the elbow method. Silhouette scores help assess the quality of clustering and provide more certainty in determining the best value for k.

3.2. Silhouette Scores:

To further evaluate the optimal number of clusters, we computed the silhouette scores for different values of k. The silhouette score provides a measure of how similar each point is to its own cluster compared to other clusters, with higher values indicating better-defined clusters.

The silhouette scores for various cluster counts are as follows:

- For n_clusters = 2, the silhouette score is 0.4376.
- For n_clusters = 3, the silhouette score is 0.4067.
- For n_clusters = 4, the silhouette score is 0.3328.
- For n_clusters = 5, the silhouette score is 0.3293.
- For n_clusters = 6, the silhouette score is 0.3124.
- For n_clusters = 7, the silhouette score is 0.2806.
- For n_clusters = 8, the silhouette score is 0.2810.
- For n_clusters = 9, the silhouette score is 0.2723.
- For n_clusters = 1 0, the silhouette score is 0.2621.
- For n_clusters = 11, the silhouette score is 0.2555.

The below plot shows the Silhouette Score Plot used for selecting the optimal value of k for K-means clustering.

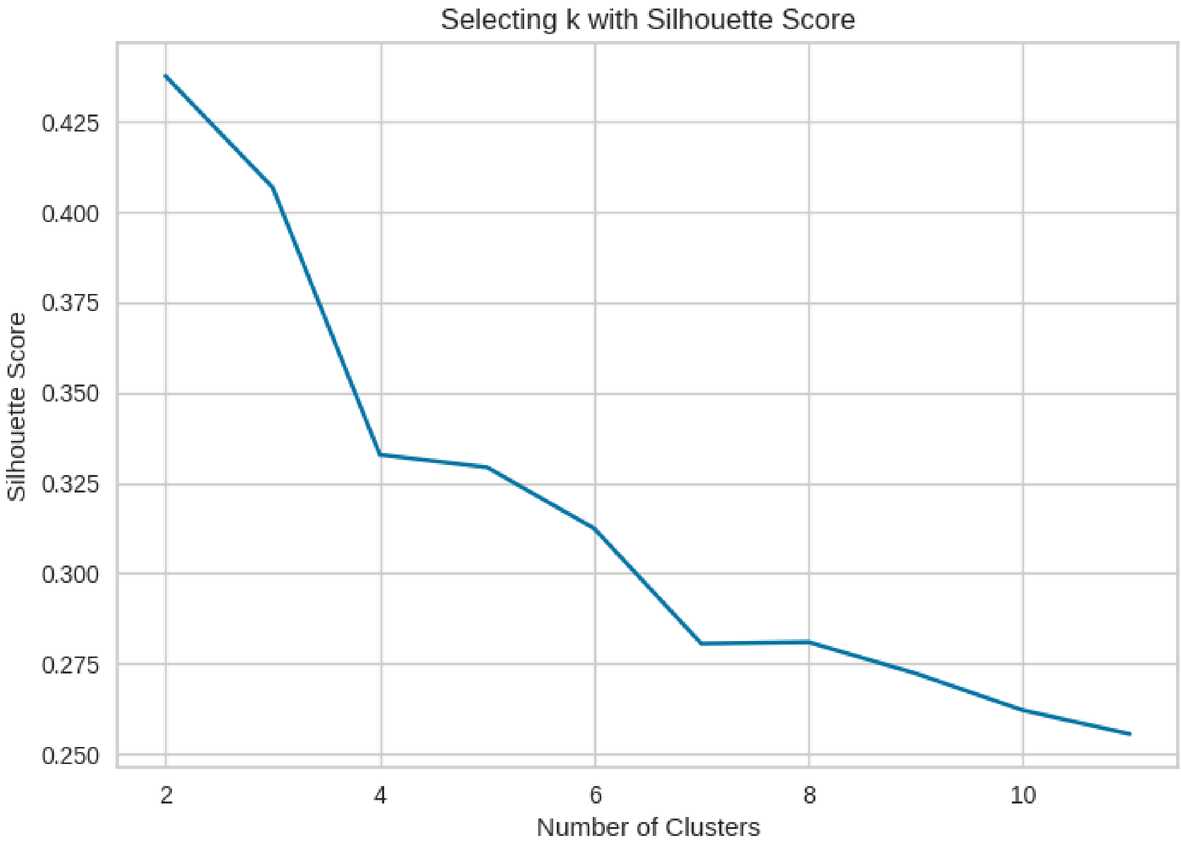

Figure 12. Silhouette Score Plot

The plot illustrates that the silhouette scores for different values of k. As seen in the silhouette score plot, the score starts high for two clusters and gradually decreases as the number of clusters increases. This trend further supports the observation that a smaller number of clusters may yield better-defined groups.

Based on these results, we notice that the silhouette score decreases as the number of clusters increases, and there is no clear upward trend throughout the plot. However, there is a slight upward trend in the silhouette score from 7 clusters to 8 clusters. Despite this, the highest silhouette score is observed for k = 2, suggesting that two clusters may provide a more distinct separation of data points compared to higher values of k.

However, since the silhouette score does not provide a clear-cut answer in this case, we will use the silhouette plot to explore further and finalize the optimal k.

Silhouette Plot for Exploring Optimal k:

The silhouette plot provides a detailed visual representation of the silhouette coefficients for each cluster when k = 7, as identified as the most optimal choice after exploring different values of kkk. Each cluster is represented by a unique color, and the width of each silhouette bar corresponds to the silhouette coefficient of individual data points within that cluster. The red dashed line indicates the average silhouette score for the clustering.

The below plot shows the Silhouette Plot used for selecting the optimal value of k for K-means clustering.

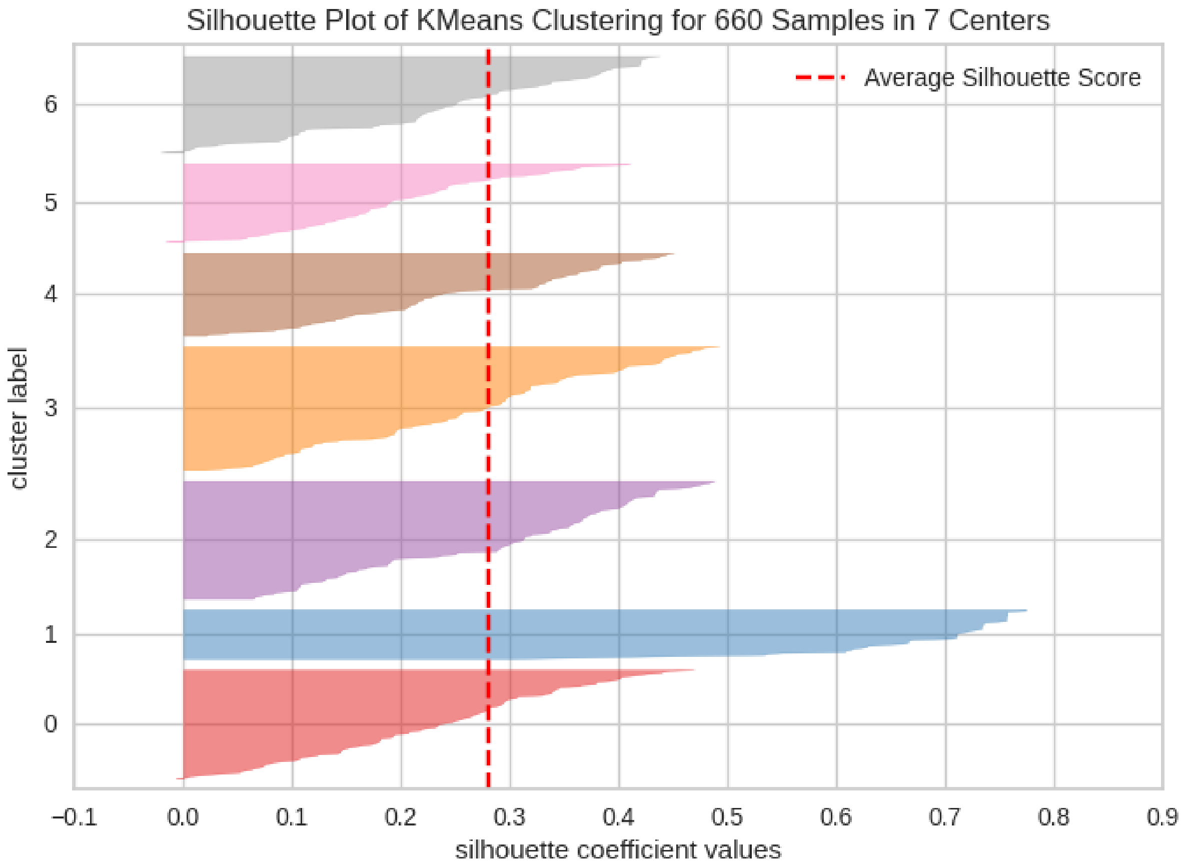

Figure 13. Silhouette Plot

From the plot, we can observe the following:

- Most clusters show a consistent silhouette width with minimal negative values, indicating well-separated clusters.

- The variation in silhouette widths between clusters suggests some degree of unevenness in how data points are grouped, but the overall score is reasonable.
- While smaller clusters exhibit tighter groupings (higher silhouette coefficients), larger clusters show more variability.

The choice of k = 7 aligns with the slight upward trend noticed in the silhouette scores between k = 7 and k = 8. Although the silhouette score decreases after k = 2, the silhouette plot reinforces that k = 7 provides a practical trade-off between well-separated clusters and meaningful segmentation.

3.3.Determining the Optimal Number of Clusters for K-means Clustering:

To determine the optimal number of clusters for K-means clustering, we used the Elbow Method and Silhouette Scores. The Elbow Method revealed a significant reduction in distortion between 1 and 2 clusters, with minimal improvements after 7 clusters, suggesting that 6 or 7 clusters were ideal. Silhouette Scores, which measure cluster separation, peaked at 2 clusters but gradually declined as the number of clusters increased. A slight improvement was observed between 7 and 8 clusters, but 7 clusters offered the best balance. The silhouette plot for k = 7 showed well-separated clusters despite the declining score. Based on these insights, 7 clusters were selected as the optimal number, balancing both cluster separation and minimal distortion, and will be used for further analysis.

3.4.Cluster Profiling for K-means Clustering:

Cluster profiling involves analyzing the distinct characteristics of each segment formed through clustering. The table and boxplots illustrate key patterns across the seven clusters identified during the K-means analysis.

Summary of Key Characteristics for Each Cluster (K-means Clustering)

	Avg_Credit_Limit	Total_Credit_Cards	Total_calls_made	Total_visits	count_in_each_segment
KM_segments					
0	19100.91743	5.752294	1.798165	3.330275	109
1	102660	8.74	1.08	8.72	50
2	12525.42373	2.372881	7.635593	5.432203	118
3	15588.70968	5.064516	2.290323	5.741935	124
4	58807.22892	5.39759	2.26506	3.168675	83
5	53924.05063	5.78481	1.772152	5.56962	79
6	11680.41237	2.360825	6.206186	3.206186	97

Table 1. Key Characteristics for Each Cluster (K-means Clustering)

From the cluster summary table the following is observed:

1. Cluster 0: This segment has an average credit limit of approximately $19,100 and a moderately high average number of total credit cards (5.75). Customers in this segment also exhibit relatively low engagement through calls and visits.

2. Cluster 1: Customers in this segment show the highest average credit limit ($102,660) and hold the highest average number of credit cards (8.74). They have the lowest average calls made (1.08) and total visits (8.72).

3. Cluster 2: This segment has a significantly lower average credit limit ($12,525) but stands out for the highest average number of calls made (7.63).

4. Cluster 3: Customers in this cluster have an average credit limit of $15,588 and a balanced interaction pattern, with moderately high calls and visits.
5. Cluster 4: This segment exhibits a higher-than-average credit limit ($58,807) but lower engagement in terms of calls and visits.
6. Cluster 5: Similar to Cluster 4, this segment has a high average credit limit ($53,924) but slightly higher engagement in calls and visits.
7. Cluster 6: Customers in this cluster have the lowest average credit limit ($11,680) and a low number of credit cards. They show relatively high engagement through calls and moderate visits.

The below plot shows the cluster-wise Distribution of Scaled Numerical Variables of K-means Clustering

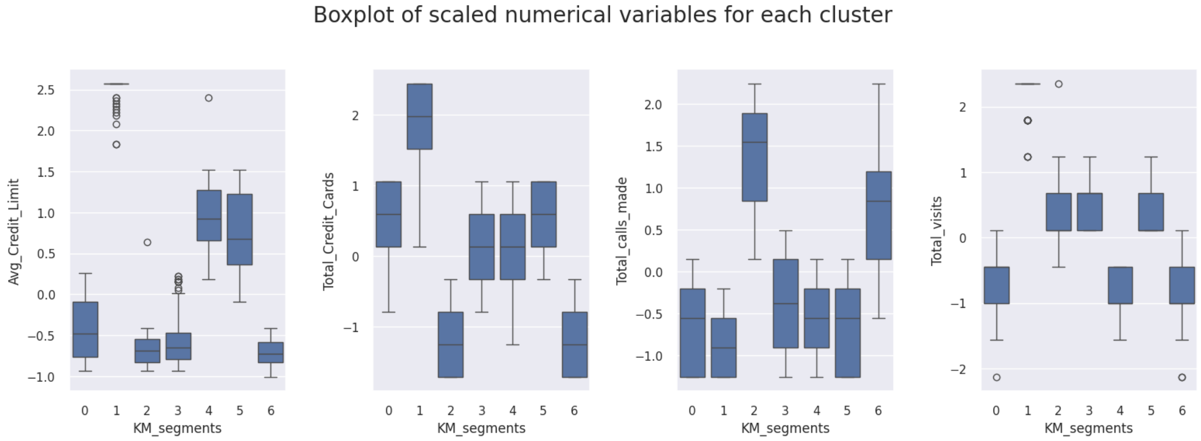

Figure 14. Box Plot of Cluster-wise Scaled Variables in K-means

Key Observations from Boxplots:

1. Avg_Credit_Limit:
   - Cluster 1 has the highest average credit limit, while Cluster 6 has the lowest average credit limit.
   - Cluster 0 and Cluster 3 show moderate average credit limits.

2. Total_Credit_Cards:

   - Cluster 1 holds the highest average number of credit cards (8.74), while Cluster 6 has the fewest (2.36).
   - Cluster 0 and Cluster 5 have slightly higher numbers of credit cards than the others.

3. Total_calls_made:

   - Cluster 2 shows the highest number of calls made on average (7.64), indicating a potential need for more support services.
   - Cluster 1 and Cluster 5 have the lowest call volume (1.08 and 1.77, respectively), suggesting lower customer service needs.

4. Total_visits:

   - Cluster 5 has the highest number of visits on average (5.57), followed by Cluster 2 (5.74), indicating frequent interaction with the bank.
   - Cluster 1 has the lowest number of visits (8.72), indicating less frequent engagement.

4. Hierarchical Clustering:

4.1. Cophenetic Correlation for Different Linkage Methods:

In hierarchical clustering, the cophenetic correlation coefficient measures the degree to which the hierarchical clustering structure preserves the original pairwise distances between data points. Higher cophenetic correlation values indicate better preservation of the data structure.

Highest Cophenetic Correlation:

- The highest cophenetic correlation is 0.8247, which was obtained using Euclidean distance and average linkage.

The cophenetic correlations for various distance metrics and linkage methods are as follows:

- For Euclidean Distance, the highest cophenetic correlation is observed with Average Linkage (0.8247), followed by Complete Linkage (0.7881), Single Linkage (0.7871), and Weighted Linkage (0.7527).
- For Chebyshev Distance, the highest cophenetic correlation is again with Average Linkage (0.8052), with Single Linkage at 0.7276, Weighted Linkage at 0.7083, and Complete Linkage showing the lowest correlation at 0.6668.
- For Mahalanobis Distance, Average Linkage also achieves the highest cophenetic correlation (0.7563), followed by Single Linkage at 0.7092, Cityblock Distance at 0.5170, and Weighted Linkage with the lowest value of 0.5778.
- For Cityblock Distance, the cophenetic correlations are relatively close, with Complete Linkage having the highest value at 0.7741, followed by Single Linkage at 0.7738, Average Linkage at 0.7848, and Weighted Linkage at 0.7631.

Insights:

- Euclidean Distance with Average Linkage shows the highest cophenetic correlation, suggesting that this combination preserves the pairwise distances most effectively.
- Other combinations such as Cityblock Distance with Complete Linkage and Chebyshev Distance with Average Linkage also show strong performance, but they do not surpass the highest correlation achieved.

4.2. Plot Dendrograms for Each Linkage Method:

Dendrograms visually represent the hierarchical clustering process, showing how individual data points are merged into clusters at various linkage distances. By plotting dendrograms for different linkage methods, we can observe the structure and separation of clusters for each method.

The dendrograms were plotted for six linkage methods, which include **Single**, **Complete**, **Average**, **Centroid**, **Ward**, and **Weighted**. For each linkage method, a dendrogram was generated using the **Euclidean distance** metric. Additionally, the cophenetic correlation coefficient was calculated and annotated on each respective dendrogram.

The below dendrogram shows the cluster-wise Distribution of Scaled Numerical Variables.

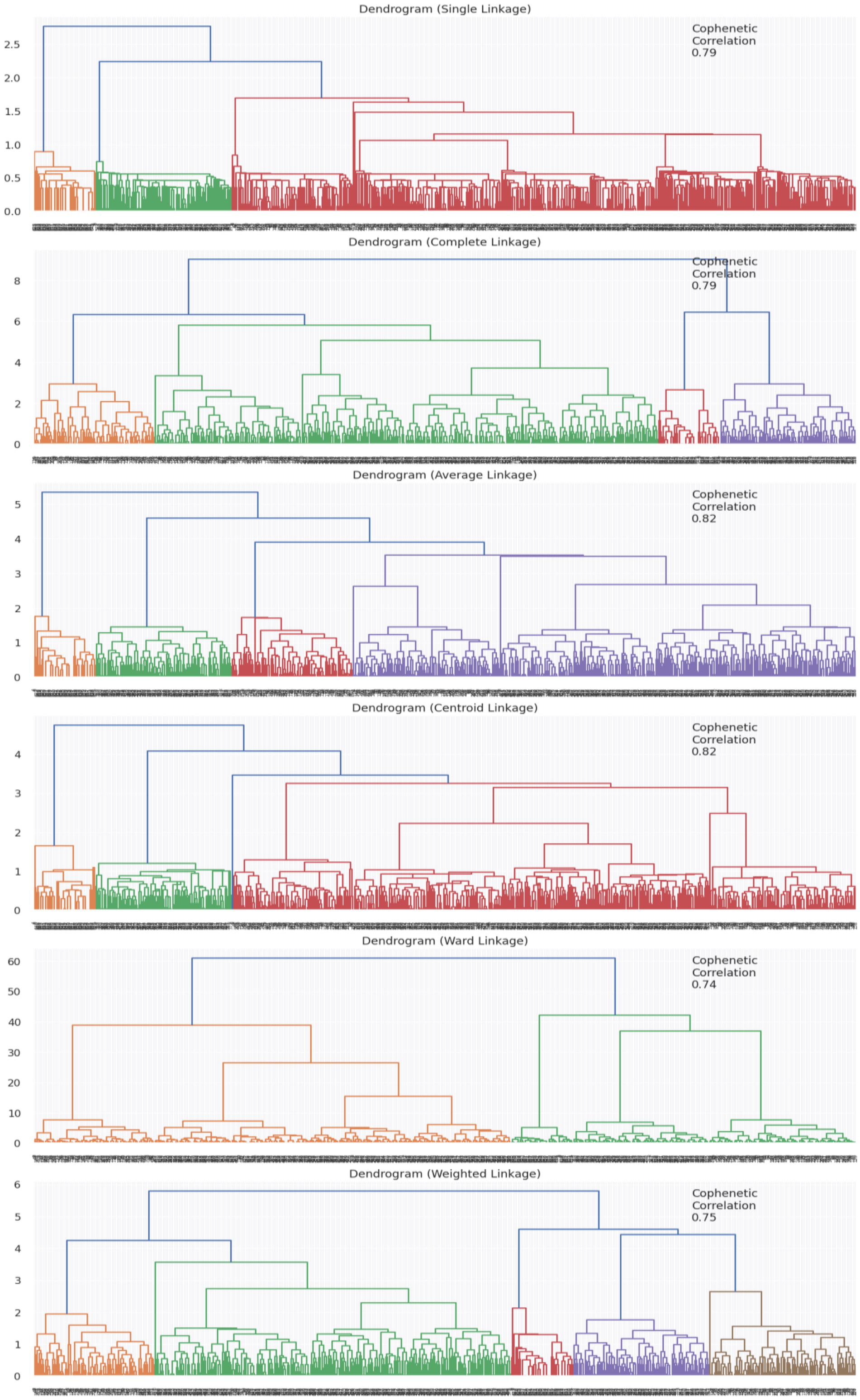

Figure 15. The dendrogram compares clustering across linkage methods.

Observations from Dendrograms:

- The Ward method produced the best visual separation between clusters, even though its cophenetic correlation was lower than other methods.
- From the dendrogram generated using the Ward method, 7 clusters appeared to be a good choice based on the clear separation of branches.

Insights:

- Dendrograms provide a visual basis for determining the number of clusters and evaluating the quality of separation for different linkage methods.
- The choice of linkage method can significantly impact the interpretation of clusters, as observed with the Ward method, which shows better-defined cluster separation despite a lower cophenetic correlation.

4.3.Determining the Optimal Number of Clusters for Hierarchical Clustering:

To determine the optimal number of clusters for hierarchical clustering, we analyzed cophenetic correlations and dendrograms for various linkage methods. The highest cophenetic correlation was achieved with Euclidean Distance and Average Linkage (0.8247), indicating the best preservation of pairwise distances. Dendrogram analysis, particularly using the Ward method, showed clear separation of clusters, with 7 clusters emerging as the most suitable choice due to distinct branch separations. Despite Ward's lower cophenetic correlation compared to other methods, its visual clarity in separating clusters made it the preferred choice. Thus, based on both cophenetic correlation and dendrogram observations, 7 clusters were selected as the optimal configuration for hierarchical clustering, balancing both cluster quality and separation.

4.4. Cluster Profiling for Hierarchical Clustering:

Cluster profiling involves analyzing the characteristics of each customer segment to differentiate them based on their behaviors and interactions with the bank. Using the Agglomerative Clustering algorithm with 7 clusters, Ward linkage, and Euclidean distance, the following insights were derived.

Summary of Key Characteristics for Each Cluster (Hierarchical Clustering)

	Avg_Credit_Limit	Total_Credit_Cards	Total_calls_made	Total_visits	count_in_each_segment
HC_segments					
0	12525.42373	2.372881	7.635593	5.432203	118
1	11680.41237	2.360825	6.206186	3.206186	97
2	15588.70968	5.064516	2.290323	5.741935	124
3	53924.05063	5.78481	1.772152	5.56962	79
4	19100.91743	5.752294	1.798165	3.330275	109
5	102660	8.74	1.08	8.72	50
6	58807.22892	5.39759	2.26506	3.168675	83

Table 2. Key Characteristics for Each Cluster (Hierarchical Clustering)

From the cluster summary table, we observe the following:

1. Cluster 0:
   This segment has an average credit limit of $12,525 and a relatively low number of credit cards (2.37). Customers in this segment exhibit high engagement with the bank, particularly through calls (7.64) and visits (5.43). This suggests that these customers may be seeking frequent assistance.

2. Cluster 1:

   Customers in this cluster have a similar credit limit to Cluster 0, with an average of $11,680 and 2.36 credit cards. However, they show slightly lower engagement, with 6.21 calls and 3.21 visits, indicating a moderate level of interaction with the bank.

3. Cluster 2:

   This segment stands out with a higher average credit limit of $15,588 and 5.06 credit cards. These customers also exhibit relatively low engagement with only 2.29 calls and 5.74 visits. This group seems to be less reliant on customer service.

4. Cluster 3:

   Customers in Cluster 3 have a substantially higher credit limit of $53,924 and 5.78 credit cards. Despite these financial attributes, they show moderate engagement with 1.77 calls and 5.57 visits. This suggests that, although they are high-value customers, they do not require frequent interaction with the bank.

5. Cluster 4:

   This group has an average credit limit of $19,100 and 5.75 credit cards. Engagement levels are somewhat low with 1.80 calls and 3.33 visits. The cluster might represent customers who have moderate spending power but tend to handle most issues independently.

6. Cluster 5:

   Cluster 5 features customers with the highest credit limit of $102,660 and the most credit cards (8.74). Despite their high financial capacity, they have the lowest engagement with only 1.08 calls and 8.72 visits, which suggests that they might be more self-sufficient or prefer digital interactions.

7. Cluster 6:

   This segment has a moderate credit limit of $58,807 and 5.40 credit cards. Customers in this segment show moderate engagement with 2.27 calls and 3.17 visits, indicating a balanced interaction with the bank.

The below plot shows the cluster-wise Distribution of Scaled Numerical Variables of Hierarchical Clustering

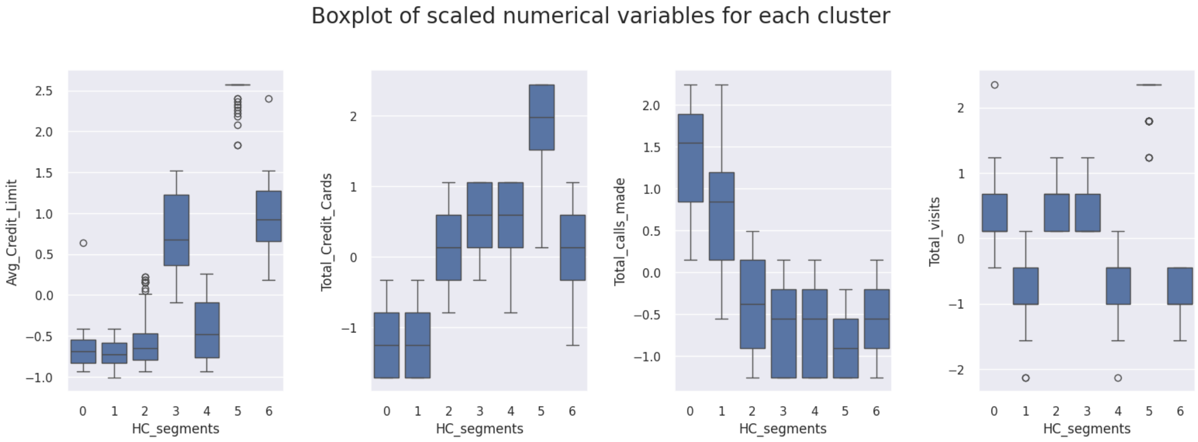

Figure 16. Box Plot of Cluster-wise Scaled Variables

Key Observations from Boxplots:

- Avg_Credit_Limit:
  - Cluster 5 has the highest average credit limit, while Clusters 0 and 1 have the lowest.
- Total_Credit_Cards:
  - Cluster 5 also has the most credit cards on average, followed by Clusters 3 and 4.
  - Clusters 0 and 1 have the fewest credit cards.
- Total_calls_made:
  - Clusters 0 and 1 show the highest number of calls, indicating potential service needs.
  - Cluster 5 has the fewest calls, which may suggest higher satisfaction or better self-service options.

- Total_visits:
    - Cluster 5 leads in the number of visits, followed by Cluster 2, indicating frequent bank engagement.
    - Clusters 1 and 4 have the lowest number of visits on average.

This analysis provides a comprehensive understanding of the behavioral and financial attributes of each customer segment.

5.K-means vs Hierarchical Clustering:

5.1.Cluster Comparison:

5.1.1.K-means Clustering Insight:

The K-means clustering results show distinct customer segments. Cluster 1 (50 customers) stands out with the highest Average Credit Limit of 102,660, suggesting a segment of high-value customers who also engage significantly with the bank (high visits at 8.72). Cluster 2, with 118 customers, has low credit limits (12,525) and high Total Calls Made (7.64), indicating customers who may require more service or have lower spending power. Cluster 0 has a moderate credit limit (19,100) and is the largest cluster (109 customers), with relatively few calls and visits. Cluster 3 (124 customers) shows similar patterns to Cluster 0, but with a slightly higher Average Credit Limit (15,588). Other clusters such as Cluster 4, Cluster 5, and Cluster 6 represent varying levels of engagement and credit limits, with Cluster 4 having a high credit limit (58,807) but fewer visits and calls compared to Cluster 5 (53,924).

5.1.2.Hierarchical Clustering Insight:

As previously mentioned, Hierarchical clustering produces similar insights. Cluster 5 (50 customers) has the highest Average Credit Limit (102,660), with fewer calls and high engagement (8.72 visits). Cluster 0 (118 customers) has the lowest credit limits (12,525) and the highest Total Calls Made (7.64), representing customers with potentially higher service needs. Other clusters such as Cluster 2 and Cluster 3 show moderate levels of engagement and credit limits, with Cluster 3 having the second-highest credit limit (58,807). The general customer segmentation follows similar patterns to K-means, indicating consistency in the clustering results across both algorithms.

5.1.3.Clusters Comparison from K-means and Hierarchical clustering techniques:

K-means and Hierarchical clustering produce similar customer segments. Cluster 1 in K-means aligns with Cluster 5 in Hierarchical clustering, both representing high-credit, highly engaged customers. Cluster 0 in both methods represents customers with lower credit limits and high service interactions. Despite slight differences in cluster labels, the underlying characteristics, such as credit limits and call frequency, are consistent, indicating both algorithms capture similar customer segments.

5.2 PCA for Visualization:

Principal Component Analysis (PCA) is a dimensionality reduction technique used to simplify datasets while retaining their most significant information. For this analysis, PCA was applied to reduce the dataset's multiple features to two principal components. This reduction helps in visualizing the underlying structure of the data in a two-dimensional space without losing much critical information.

The first two principal components explain 80.71% of the variance in the data, indicating that most of the dataset's original variability is preserved in this reduced form. This makes PCA an effective method for identifying patterns, relationships, and groupings that might otherwise be hidden in higher dimensions.

By transforming the data into two components, PCA not only simplifies the dataset but also makes it easier to interpret and analyze. This visualization enables an intuitive understanding of the data structure, which is especially useful when performing clustering analysis to identify distinct groups or customer segments.

The below plot shows the PCA Scatter Plot

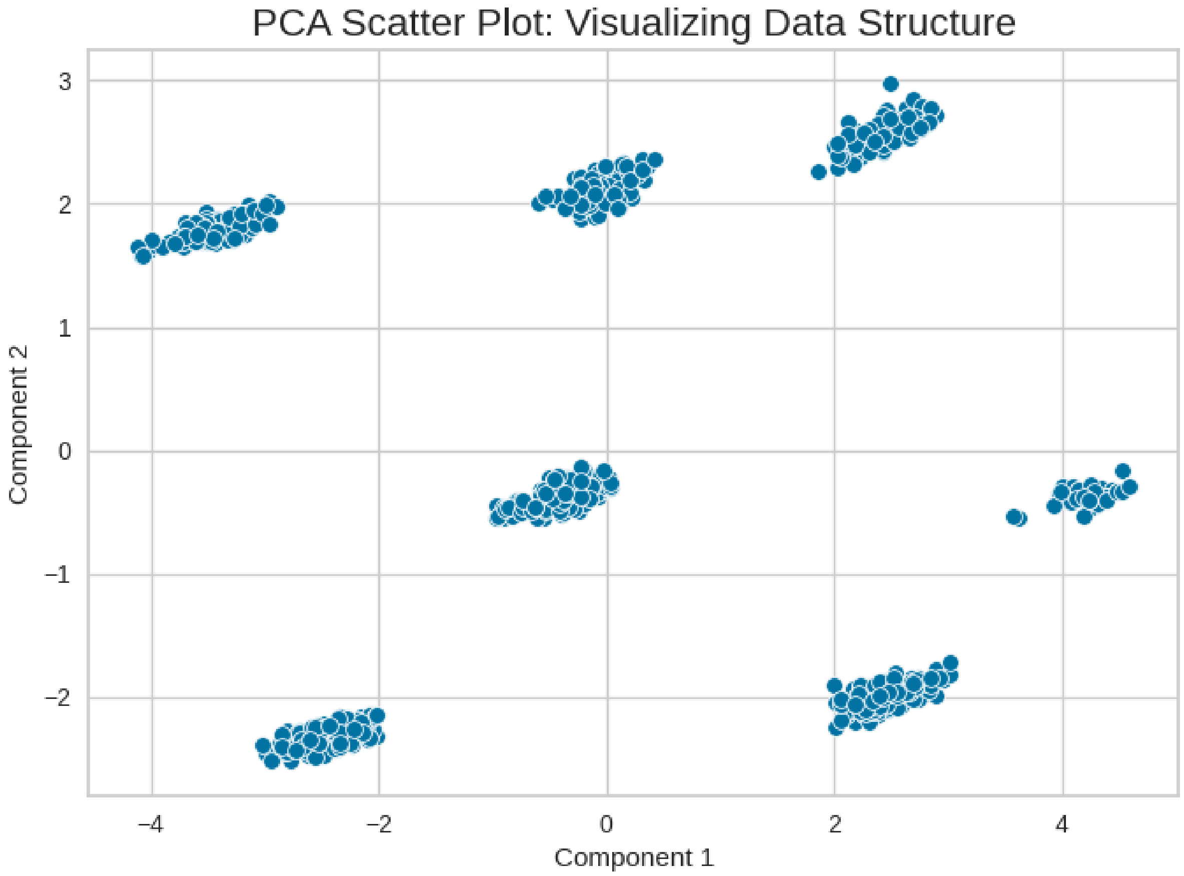

Figure 17. PCA Scatter Plot: Visualizing Data Structure

This scatter plot visualizes the dataset in two dimensions using Principal Component Analysis (PCA). The x-axis represents the first principal component, and the y-axis represents the second principal component. Each point in the plot corresponds to a data record. The clusters observed in this two-dimensional space highlight distinct groupings within the dataset, indicating underlying patterns or structures.

The data points are distributed in distinct clusters with minimal overlap, showing clear separation between groups. Each cluster represents a unique segment of the dataset with distinct characteristics. Some clusters are tightly packed, indicating high similarity among data points within those groups, while others are more spread out, suggesting variability within those clusters. The scatter plot effectively captures the inherent structure of the data, making it easier to identify patterns and groupings.

The below plot shows the PCA Scatter Plot with Hierarchical Clustering Labels

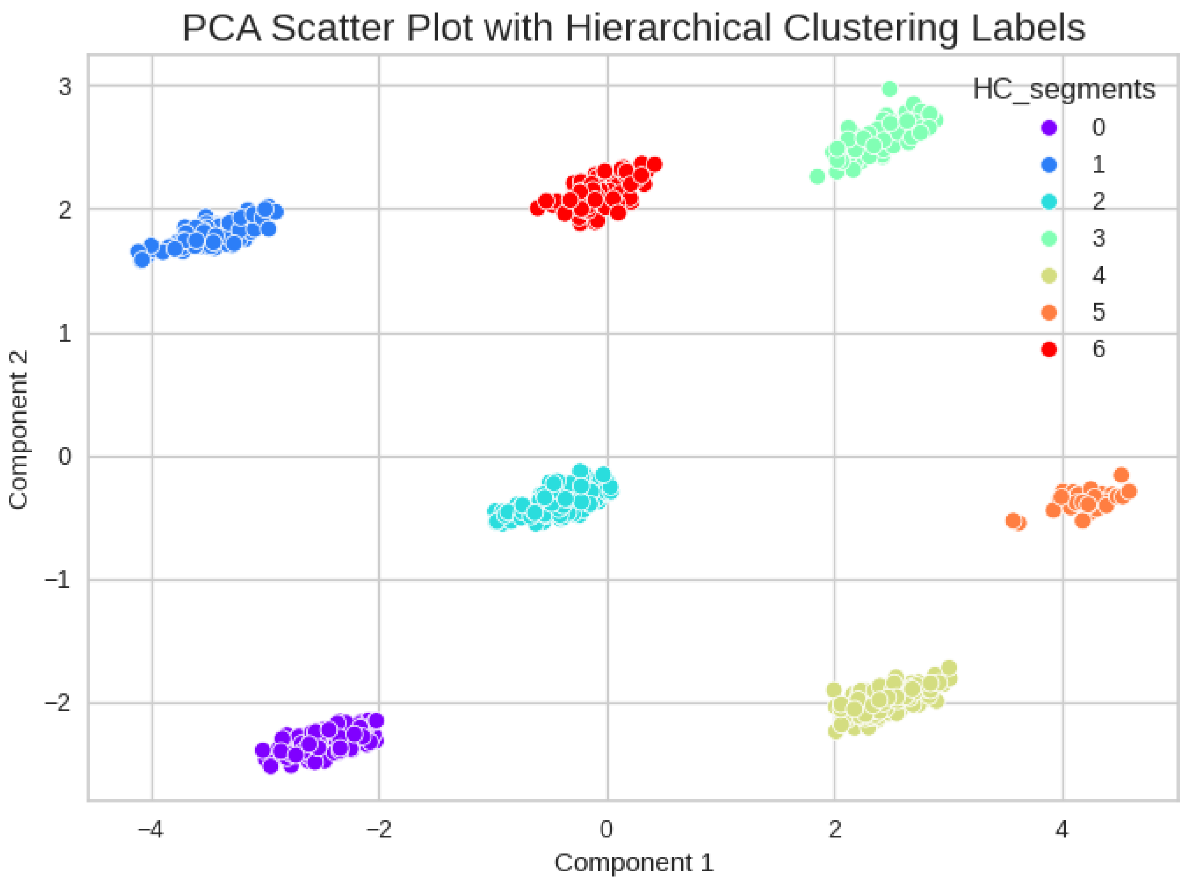

Figure 18. PCA Scatter Plot with Hierarchical Clustering Labels

This scatter plot presents the PCA-reduced data, color-coded to indicate the clusters identified using Hierarchical Clustering. The x-axis and y-axis represent the first and second principal components, respectively. Each cluster is assigned a unique color, making it easier to differentiate between the groups.

The clusters are well-separated, demonstrating the effectiveness of Hierarchical Clustering in grouping the data. Each cluster represents a distinct group of data points, with unique characteristics that differentiate it from other clusters. The compact and well-defined clusters, such as those near the top-left and bottom-left, suggest high internal similarity, while the slightly more dispersed clusters, such as the ones near the center, indicate some variability within those groups.

5.2.1.Choosing Hierarchical Clustering:

Although Hierarchical Clustering was chosen for this analysis, K-means Clustering also produced similar groupings. Both clustering methods effectively reveal distinct customer segments in the dataset. The choice between them can be made based on personal preference or specific use-case requirements. In this case, Hierarchical Clustering highlights the dataset's structure effectively, as evidenced by the clear separation of clusters in the colored plot. Each color-coded cluster represents a distinct group within the dataset.

6. Actionable Insights & Recommendations

6.1.Insights from the Analysis Conducted:

Based on the clustering analysis of customer spending patterns and past interactions with AllLife Bank, the following actionable insights and recommendations have been derived:

1. Customer Segmentation:

Insight:

The clustering analysis reveals distinct customer segments characterized by their spending behavior and interaction patterns. For instance, some customers exhibit high credit limits but low interaction with the bank, while others have frequent interactions but lower credit limits.

Recommendation:

- High-value customers (low interaction, high spending): Focus on retention strategies such as loyalty programs, premium offers, or exclusive promotions.
- Low-value customers (high interaction, low spending): Develop targeted upselling strategies, personalized offers, and incentives to boost engagement and spending.

2. Personalized Marketing Campaigns:

Insight:

Different customer segments are likely to respond differently to marketing messages based on their interaction and spending behaviors. Customers with high interaction frequency may prefer in-branch promotions, while those with strong online engagement may favor digital communication.

Recommendation:

- For high-engagement customers (frequent visits/calls): Deploy personalized campaigns through both in-branch promotions and digital channels to maintain strong relationships and increase product uptake.

- For digital-focused customers (high online logins): Focus on personalized digital marketing campaigns via email, SMS, or app-based notifications to drive engagement.

3. Service Improvements:

Insight:

Frequent customer service interactions, such as high call volume, could indicate dissatisfaction or areas where service processes need improvement. Understanding this behavior allows the bank to proactively address support issues.

Recommendation:

- For customers with high call frequency: Enhance customer service by reducing response times, introducing more self-service options like chatbots, and ensuring faster resolution of common queries.
- For low-interaction customers: Engage them proactively through emails or notifications to enhance their support experience and encourage deeper engagement with the bank.

4. Credit Limit Adjustments:

Insight:

The analysis of customer spending and credit usage patterns indicates that some customers are underutilizing their credit limits, while others are nearing their credit limits. This behavior can inform potential credit policy adjustments.

Recommendation:

- For low-usage, high-limit customers: Consider adjusting credit limits to optimize risk management, reduce exposure, and improve profitability.
- For high-usage, low-limit customers: Offer credit limit increases to encourage higher spending, provide added value, and strengthen customer loyalty.

5. Upselling and Cross-Selling Opportunities:

Insight:

Customers with multiple credit cards or high interaction frequency present strong potential for upselling and cross-selling additional financial products such as insurance or premium services.

Recommendation:

- For multi-card holders: Promote additional products such as insurance or investment services that align with their financial behavior.
- For low-engagement but high-potential customers: Reach out with targeted offers based on their spending patterns and customer profile to increase engagement and expand product usage.

6.2. Conclusion:

The customer segmentation analysis using K-means and Hierarchical clustering has revealed distinct customer segments based on spending patterns and interactions with AllLife Bank. These insights provide a foundation for targeted marketing campaigns, personalized service improvements, and strategic credit limit adjustments. By focusing on high-value customers with tailored offers, addressing frequent service interactions, and exploring upselling opportunities, the bank can enhance customer engagement and satisfaction. Additionally, optimizing credit limits based on customer behavior can improve profitability and reduce risk. Adopting these recommendations will help AllLife Bank strengthen relationships with existing customers, attract new ones, and achieve its business objectives, ultimately fostering long-term growth and success.