PGP-DSBA PROJECT REPORT

UL – Coded Project

BY

ISHAAN SHAKTI JAYARAMAN

PGPDSBA.O.JULY24.A

Contents

LIST OF FIGURES

INTRODUCTION

All Life Bank wants to focus on its credit card customer base in the next financial year. They have been advised by their marketing research team, that the penetration in the market can be improved. Based on this input, the Marketing team proposes to run personalized campaigns to target new customers as well as upsell to existing customers. Another insight from the market research was that the customers perceive the support services of the back poorly. Based on this, the Operations team wants to upgrade the service delivery model, to ensure that customer queries are resolved faster. The Head of Marketing and Head of Delivery both decide to reach out to the Data Science team for help.

1.1 Objective

To identify different segments in the existing customers, based on their spending patterns as well as past interaction with the bank, using clustering algorithms, and provide recommendations to the bank on how to better market to and service these customers.

EXPLORATORY DATA ANALYSIS

2.1 Problem Definition

All Life Bank plans to focus on its credit card customer base in the upcoming financial year. The marketing research team has advised that market penetration can be improved. In response, the Marketing team intends to launch personalized campaigns targeting both new customers and existing customers for upselling. Research shows that customers have a poor perception of the bank's support services. To address this issue, the Operations team aims to enhance the service delivery model to resolve customer queries more efficiently. The Head of Marketing and Head of Delivery have jointly decided to collaborate with the Data Science team for support.

The primary objective is to identify distinct customer segments based on spending behavior and past interactions with the bank. Clustering algorithms will be used for this purpose, with the goal of providing actionable recommendations on how to better market to and serve these customers. The insights will help improve both marketing strategies and customer service efforts.

2.2 Data Contents

The data (Credit Card Customer Data.xlsx) provided is of various customers of a bank and their financial attributes like credit limit, the total number of credit cards the customer has, and different channels through which customers have contacted the bank for any queries (including visiting the bank, online, and through a call center)

- There are 660 observations in the dataset.
- There are 7 columns in the dataset containing a detailed view of each financial attribute of customers.
- There are 7 numerical variables, all are integer data types.
- “Sl_No” consists of entirely unique values, This column will be dropped from the dataset.
- There are no null values.
- There are no duplicate values.

2.3 Data Dictionary

The data contains the different financial attributes of various customers of a bank, the detailed data dictionary is given below.

- Sl_No: Primary key of the records
- Customer Key: Customer identification number
- Average Credit Limit: Average credit limit of each customer for all credit cards
- Total credit cards: Total number of credit cards possessed by the customer
- Total visits bank: Total number of visits that the customer made (yearly) personally to the bank
- Total visits online: Total number of visits or online logins made by the customer (yearly)
- Total calls made: Total number of calls made by the customer to the bank or its customer service department (yearly)

2.4 Statistical Summary

Figure 1 - Statistical Summary of the Dataset

	count	mean	std	min	25%	50%	75%	max
Sl_No	660.0	330.500000	190.669872	1.0	165.75	330.5	495.25	660.0
Customer Key	660.0	55141.443939	25627.772200	11265.0	33825.25	53874.5	77202.50	99843.0
Avg_Credit_Limit	660.0	34574.242424	37625.487804	3000.0	10000.00	18000.0	48000.00	200000.0
Total_Credit_Cards	660.0	4.706061	2.167835	1.0	3.00	5.0	6.00	10.0
Total_visits_bank	660.0	2.403030	1.631813	0.0	1.00	2.0	4.00	5.0
Total_visits_online	660.0	2.606061	2.935724	0.0	1.00	2.0	4.00	15.0
Total_calls_made	660.0	3.583333	2.865317	0.0	1.00	3.0	5.00	10.0

Insights:

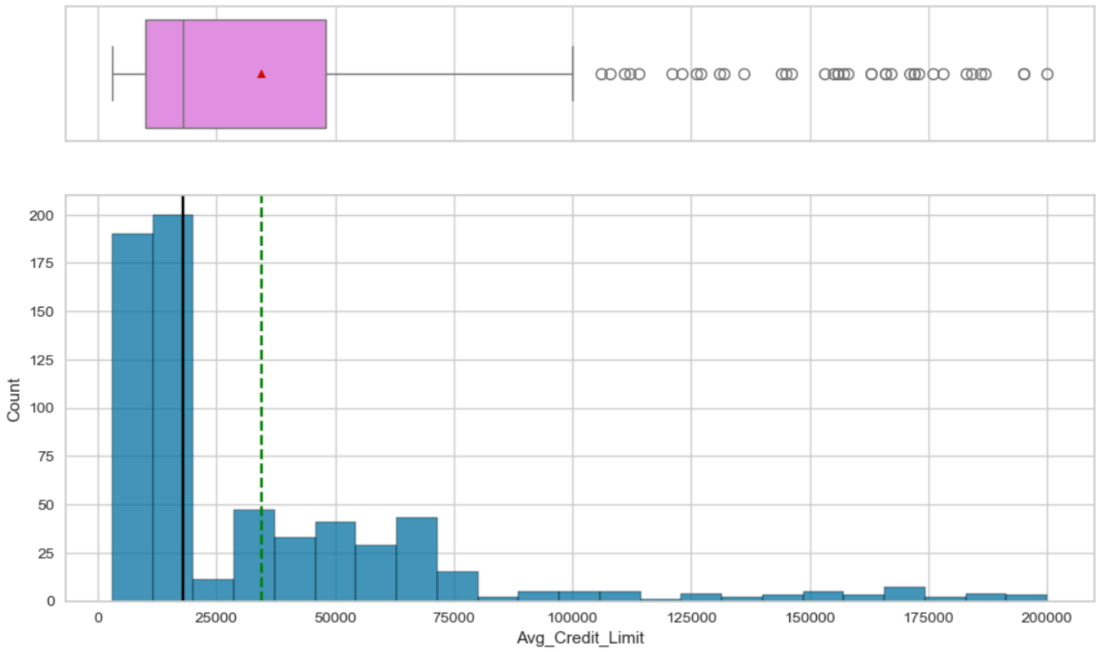
- The Sl No column are unique values which specifies the serial number of the data. This column can be ignored.
- The Customer Key is a unique customer identification number. This column can be ignored.
- The Average Credit limit is approximately 34,574 for all customers, less than 50% of the customers have less than 18,000 average credit limits.
- An average of 5 credit cards are issued to all customers, with the highest number of cards issued to a person being 10. Less than 50% of the customers are issue 5 and lesser credit cards
- Customers on average personally visit the bank 2-3 days yearly, the highest number of visits a customer has made to the bank is 5
- Customers on average made online logins to the bank's website a total of 2-3 days yearly. 75% of the customer base use online bank portal for less than 4 days yearly.
- Customers make 3-4 calls to the bank or customer service department yearly. Less than 75% of the customer base make 5 calls in a year and the highest number of calls the bank has received by a single customer is 10 in a year.

2.5 Univariate Analysis

Univariate analysis is a form of statistical analysis that involves examining the distribution and characteristics of a single variable. It focuses on summarizing and visualizing the data for one variable to understand its central tendency, spread, and distribution.

2.5.1 Average Credit Limit

Figure 2 - Histogram Boxplot of Average Credit Limit

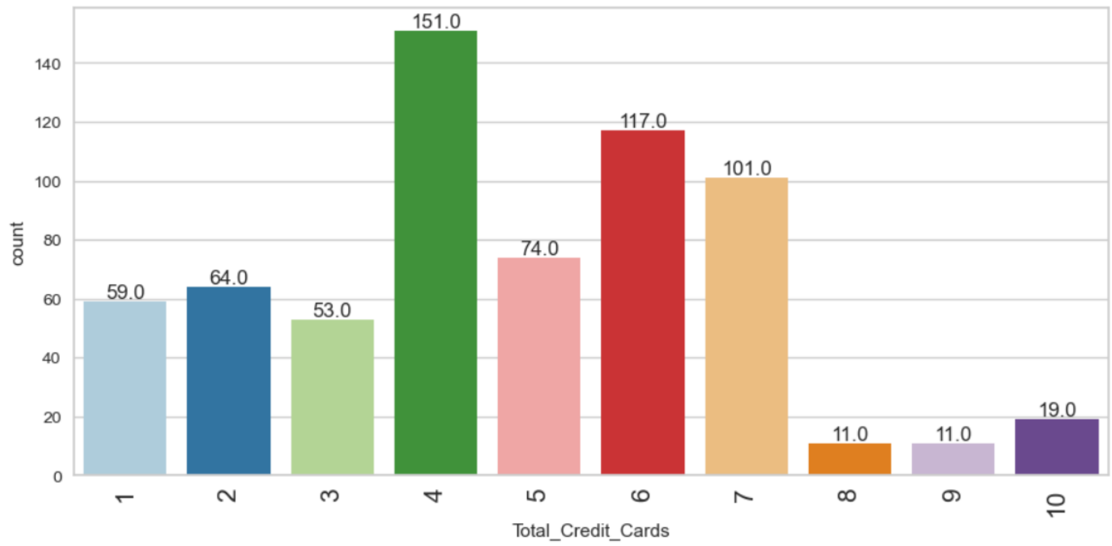


Insights:

- The distribution is skewed right
- The majority of the observations have credit limit less than the mean.
- 75% of the total data has credit limit below ~ 45,000.
- There are many outliers present in this variable.

2.5.2 Total Credit Cards

Figure 3 - Labeled Barplot of Total Credit Cards

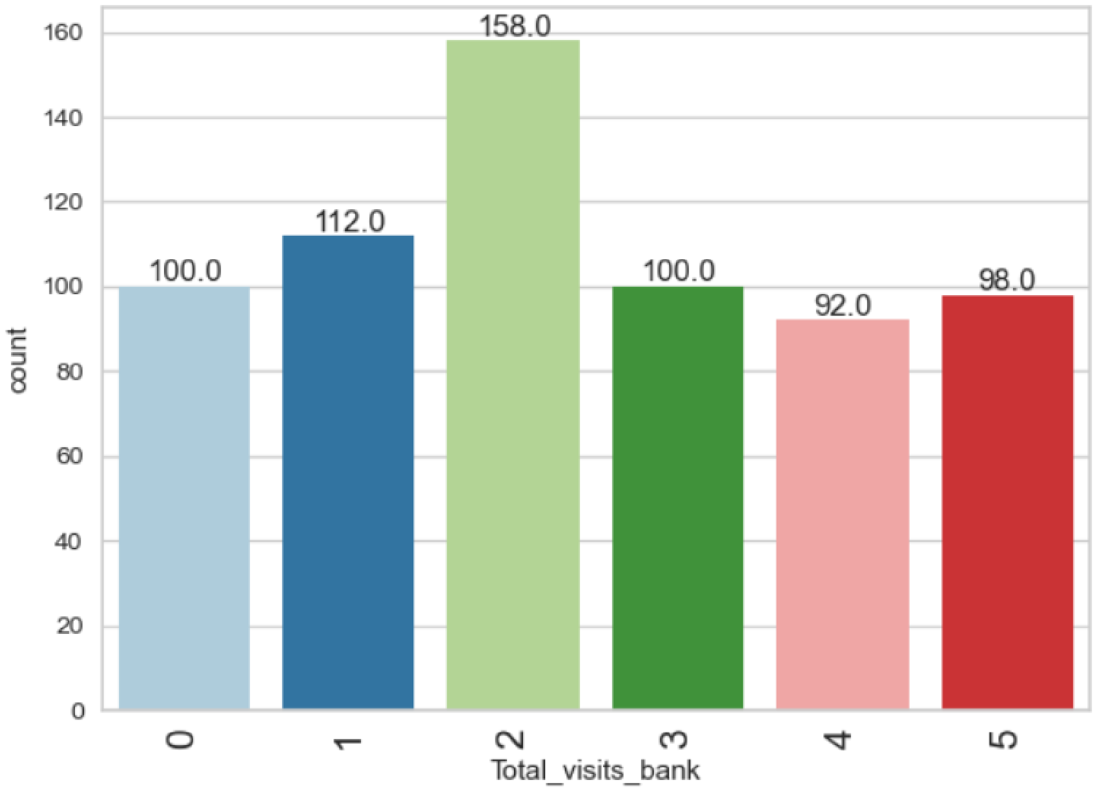


Insights:

- A maximum of 10 credit cards have been issued to few customers.
- Approximately 22.8% of the customers were issued 4 credit cards.
- 4-7 credit cards are issued to the majority of the customers.
- Very few customers are issued more than 7 credit cards.
- 1-3 credit cards are issued to a moderate amount of customers.
- The distribution is fairly normal.

2.5.3 Total Visits Bank

Figure 4 - Labeled Barplot of Total Visits Bank

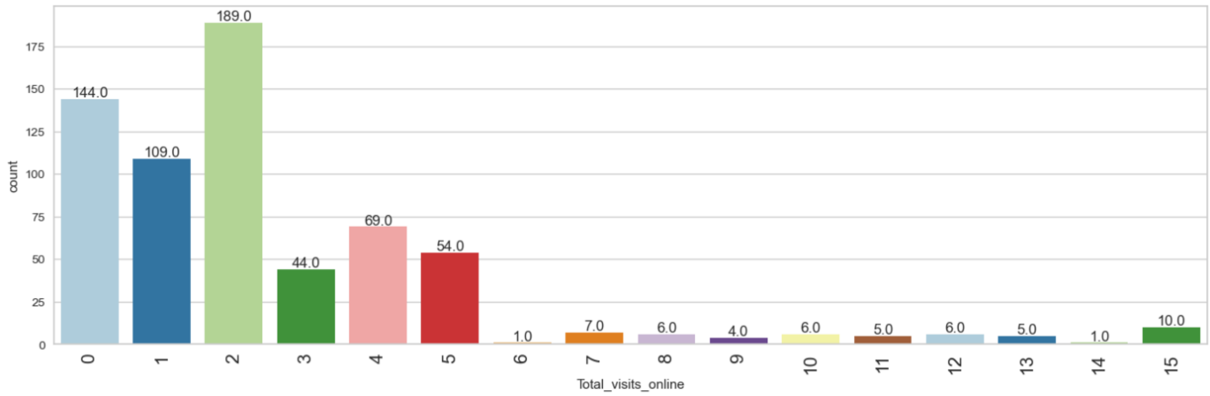


Insights:

- The distribution is fairly normal
- Approximately 24% of the customers visit the bank at least 2 times in a year.
- Approximately 15% of the customers do not visit the bank at all.
- Approximately 44% of the customers visit the bank more than 3 times in a year.

2.5.4 Total Visits Online

Figure 5 - Labeled Barplot of Total Visits Online

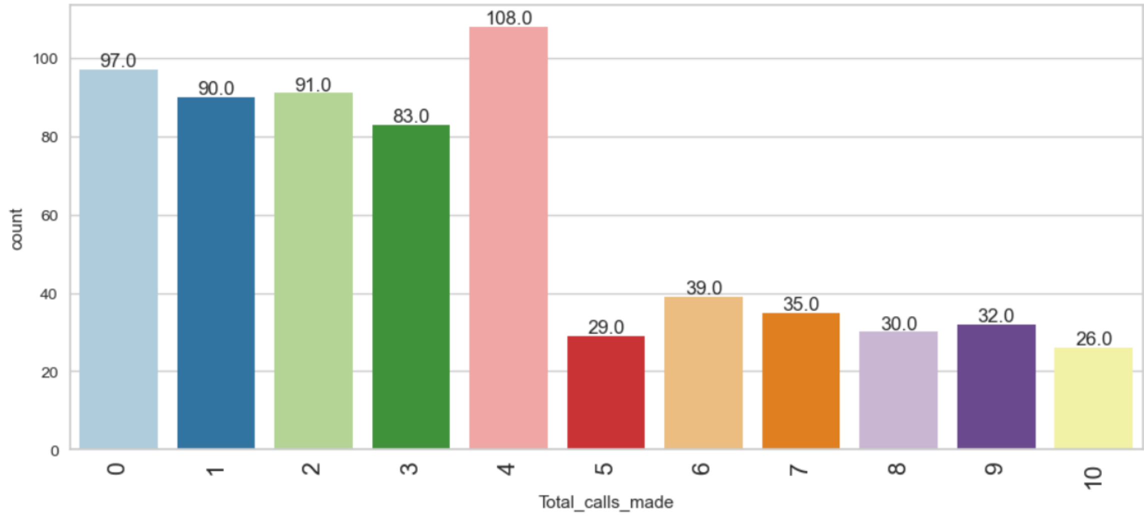


Insights:

- The distribution is skewed right.
- Approximately 22% of the customers do not use bank facilities online.
- Approximately 45.15% of the customers use online bank facilities at least 2 times.
- Approximately 25.30% of the customers use the bank login online between 3-5 times.
- Approximately 7.27% of the customers use the bank facilities online more than 5 times.

2.5.5 Total Calls Made

Figure 6 - Labeled Barplot of Total Calls Made



Insights:

- The distribution is skewed right.
- Approximately 14.69% of the customers do not make service calls to the bank.
- Approximately 57.72% of the customers make service calls to the bank at least 1-4 times.
- Approximately 28.93% of the customers make calls to the bank more than 4 times.

2.6 Bivariate Analysis

Bivariate analysis examines the relationship between two variables to understand how they interact or correlate with each other. It can identify patterns, trends, and dependencies between the variables, such as whether changes in one variable correspond to changes in another.

Utilizing various charts such as heatmap etc. will help us determine if a relationship exists and provide insights into cause-and-effect dynamics or associations between variables.

2.6.1 Heatmap

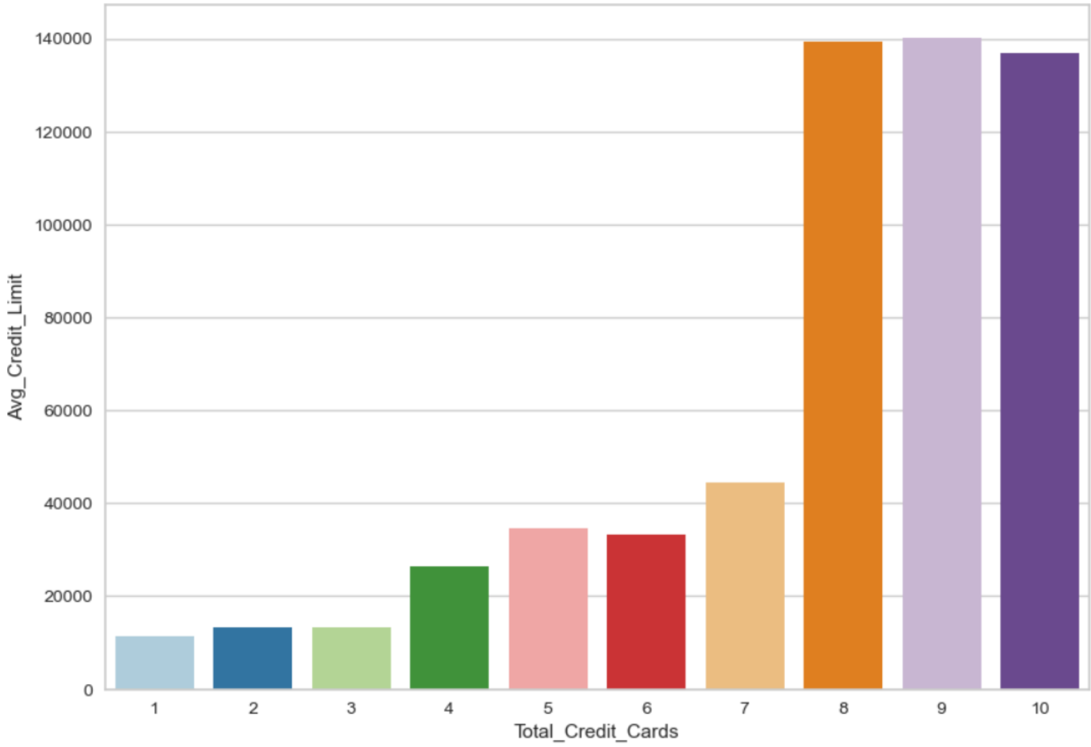
Figure 7 - Heatmap of all Numerical Variables

Insights:

- Total credit cards issued have a high correlation with average credit limit, logically as a customer has a higher limit more cards can be issued to them.
- Total visits online and average credit limit are positively correlated
- Total visits bank and total credit cards are positively correlated.
- The other variables are either negatively correlated or have a very weak correlation.

2.6.2 Average Credit Limit vs Total Credit Cards

Figure 8 - Barplot of Average Credit Card vs Total Credit Cards

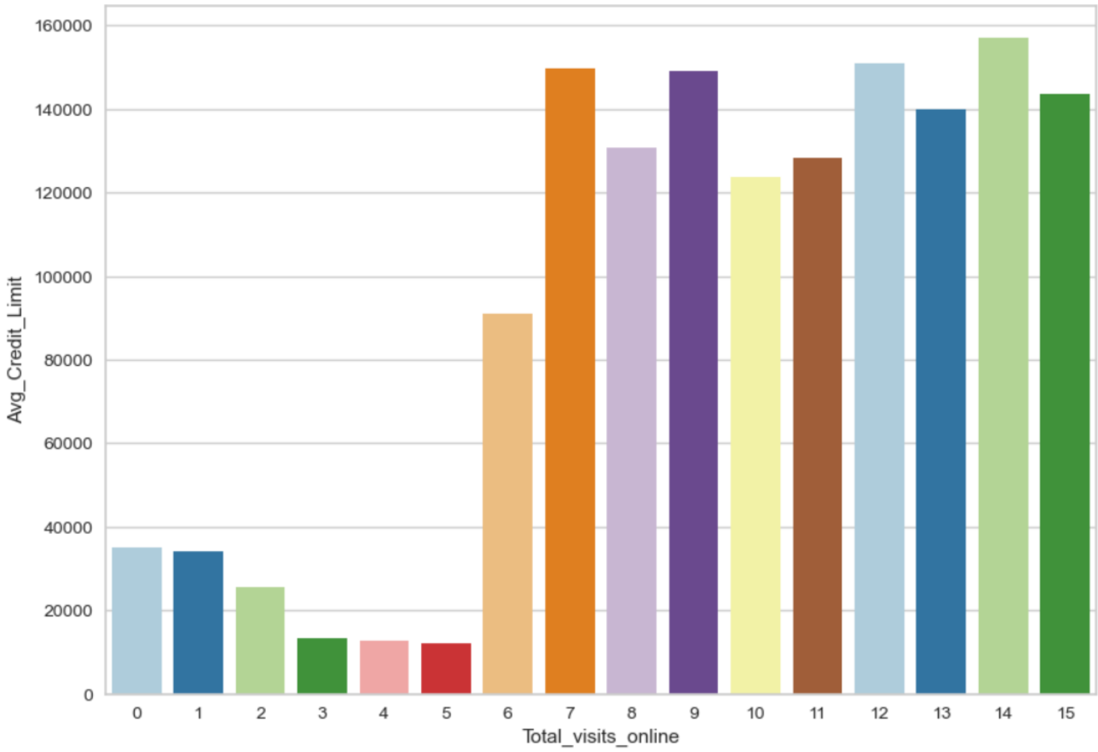


Insights:

- The distribution is skewed left.
- More credit cards are issued to customers with higher average credit limit.

2.6.3 Average Credit Limit vs Total Visits Online

Figure 9 - Barplot of Average Credit Limit vs Total Visits Online

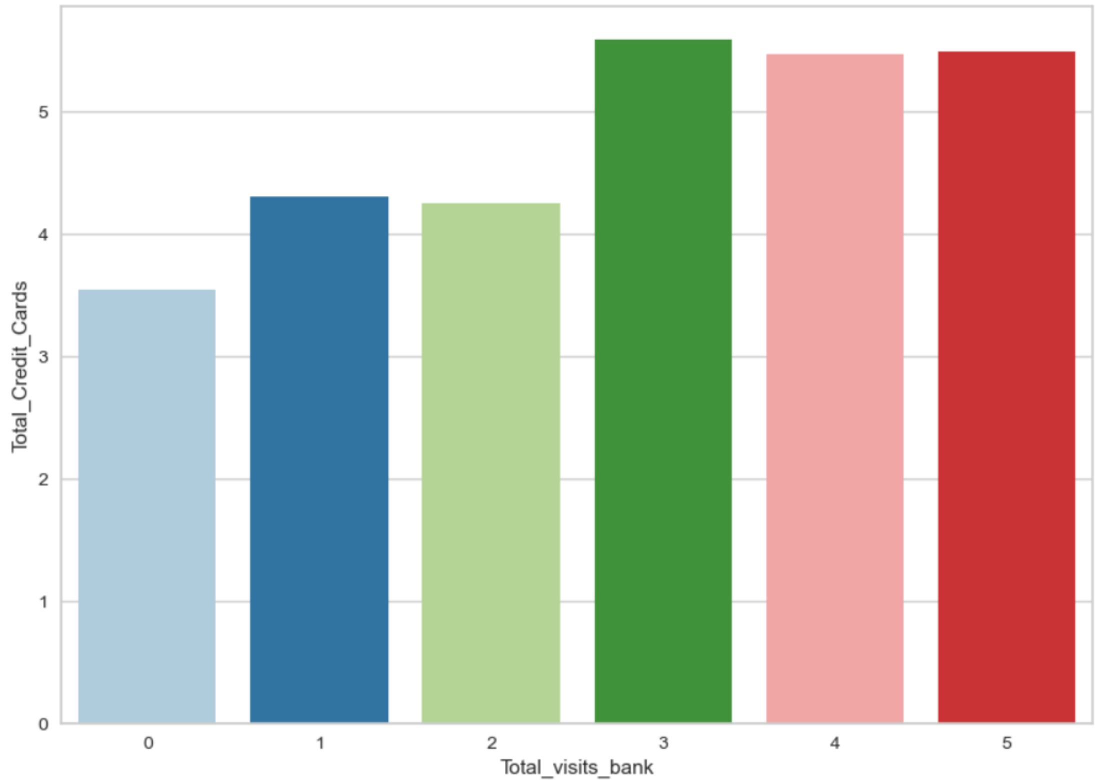


Insights:

- The distribution is skewed left.
- Customers with higher credit limit utilize the banks online net banking services more in a year compared to customers with lower credit limit.

2.6.4 Total Credit Cards vs Total Visits Bank

Figure 10 - Barplot of Total Credit Cards vs Total Visits Bank



Insights:

- The distribution is slightly skewed left.
- Customers who have been issued more credit cards tend to visit the bank in person more often in a year.

DATA PREPROCESSING

3.1 Missing Value Check

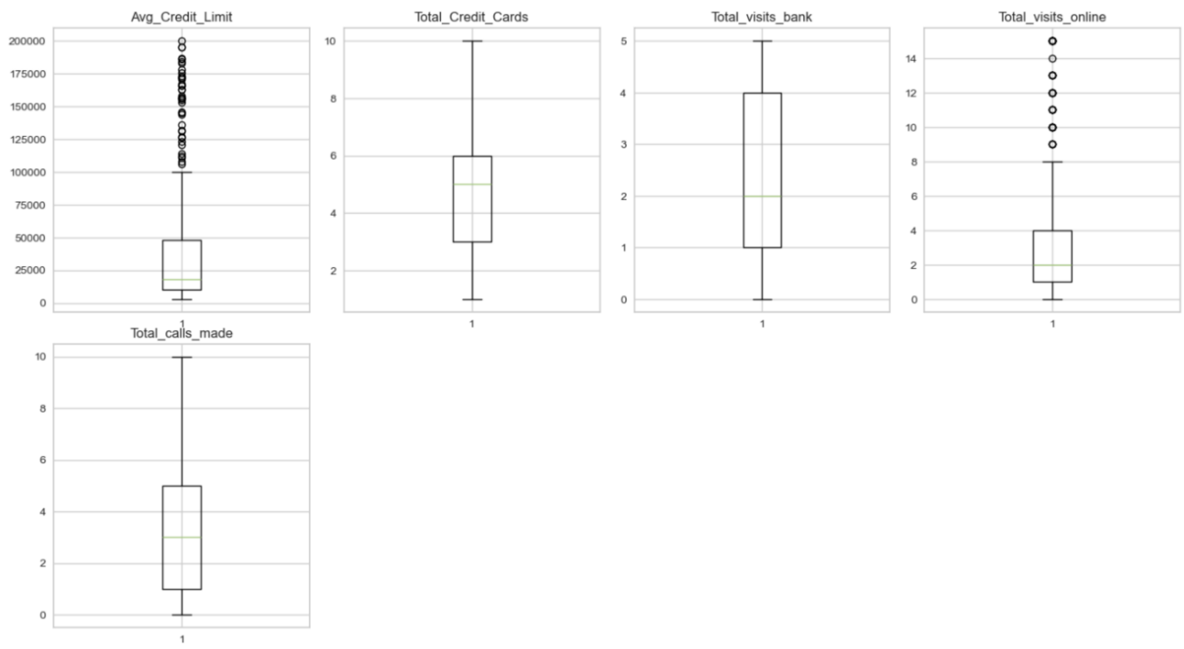
There are no missing values in the dataset therefore no treatment is required

3.2 Duplicate Value Check

There are no duplicate values in the dataset therefore no treatment is required.

3.3 Outlier Check

Figure 11 - Boxplot of Numerical Variables



There are large number of outliers present in “Avg_Credit_Limit” and “Total_visits_online” columns, these values will be considered as genuine values as they can contribute to providing insights so these outliers will not be needing treatment.

3.4 Data Preparation

The columns of “Sl_No” and “Customer Key” are dropped as they do not contribute to identifying segments as “Sl_No” are unique values and “Customer Key” are customer identification number which is a unique number associated to a customer used for identification purposes.

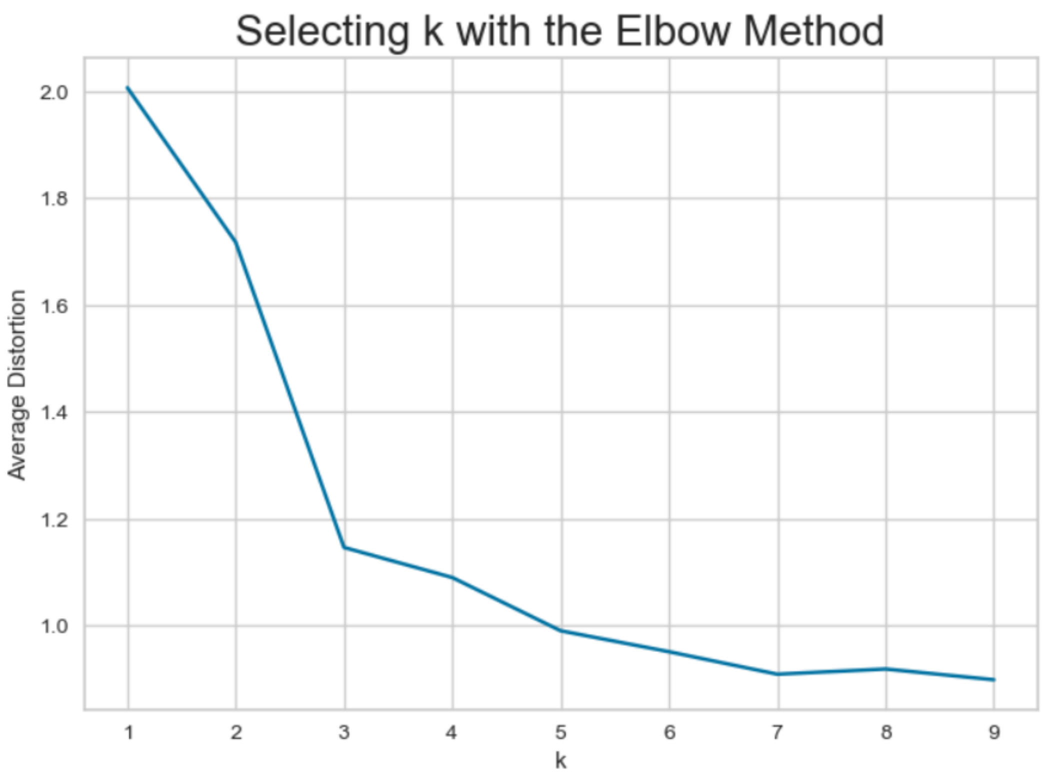
The dataset has been scaled using standard scaler as the values in average credit limit column are larger compared to the values in the other columns, this is to ensure that all features contribute equally to the model to prevent one larger variable from overshadowing the smaller variables.

K-MEANS CLUSTERING

K-Means Clustering is an unsupervised machine learning algorithm used to partition data into K distinct clusters based on feature similarity. The algorithm works by initializing K centroids and iteratively assigning data points to the nearest centroid, then recalculating the centroids based on the mean of the points assigned to them. This process repeats until the centroids no longer change significantly, indicating convergence. K-Means is commonly used for grouping data into clusters for further analysis, segmentation, or pattern recognition.

4.1 Elbow Method Curve Plot

Figure 12 – K-means Elbow Curve Plot Average Distortion

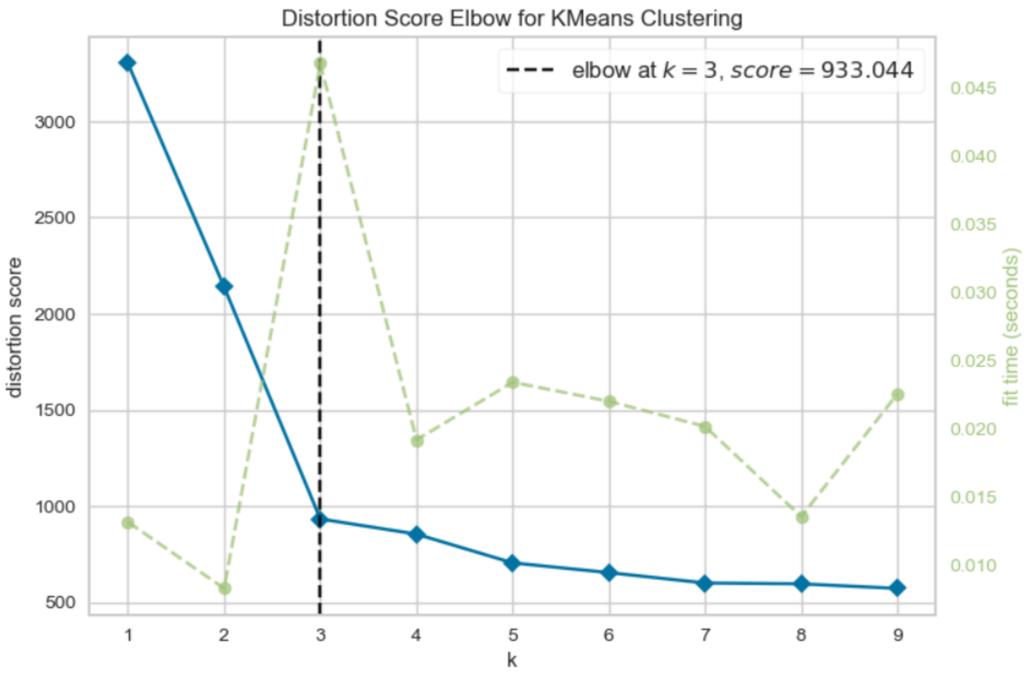


The elbow curve method is a technique used to determine the optimal number of clusters in K-means clustering. The point where the distortion declines most is said to be the elbow point and defines the optimal number of clusters for the dataset. From the above graph, The elbow bends around point 3 with an average distortion score of 1.71.

Figure 13 - Average Distortion Scores for Various K-values

```
Number of Clusters: 1   Average Distortion: 2.0069222262503614
Number of Clusters: 2   Average Distortion: 1.7178787250175893
Number of Clusters: 3   Average Distortion: 1.1466276549150365
Number of Clusters: 4   Average Distortion: 1.0902973540817664
Number of Clusters: 5   Average Distortion: 0.9906853650098948
Number of Clusters: 6   Average Distortion: 0.9515009282361341
Number of Clusters: 7   Average Distortion: 0.9094119827472316
Number of Clusters: 8   Average Distortion: 0.9191292344244387
Number of Clusters: 9   Average Distortion: 0.8990131857179275
```

Figure 14 - K-means Elbow Curve Plot Distortion Score



4.2 Silhouette Score

The silhouette score is a metric in K-means clustering which is used to evaluate the quality of the clusters. It provides a way to assess how well each data point fits within its assigned cluster, and how distinct the clusters are from each other. The silhouette score ranges from -1 to 1. 1 indicating that the data points are well clustered and -1 being the data points are misclassified, a score close to 0 indicates that data points lie close to the boundaries which indicates uncertainty.

Figure 15 – K-means Silhouette Score

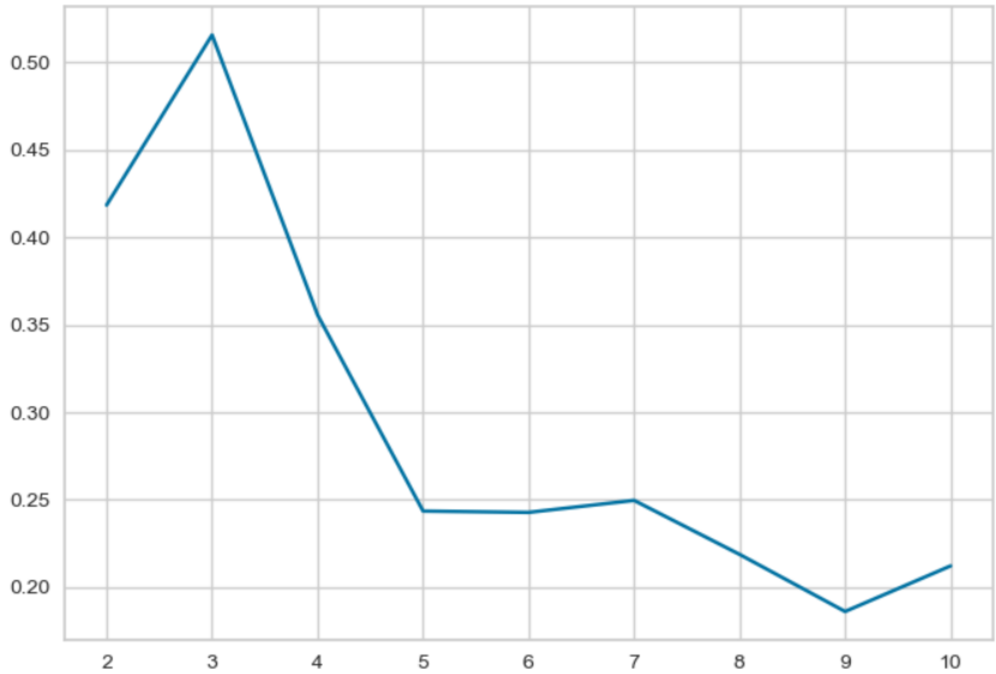
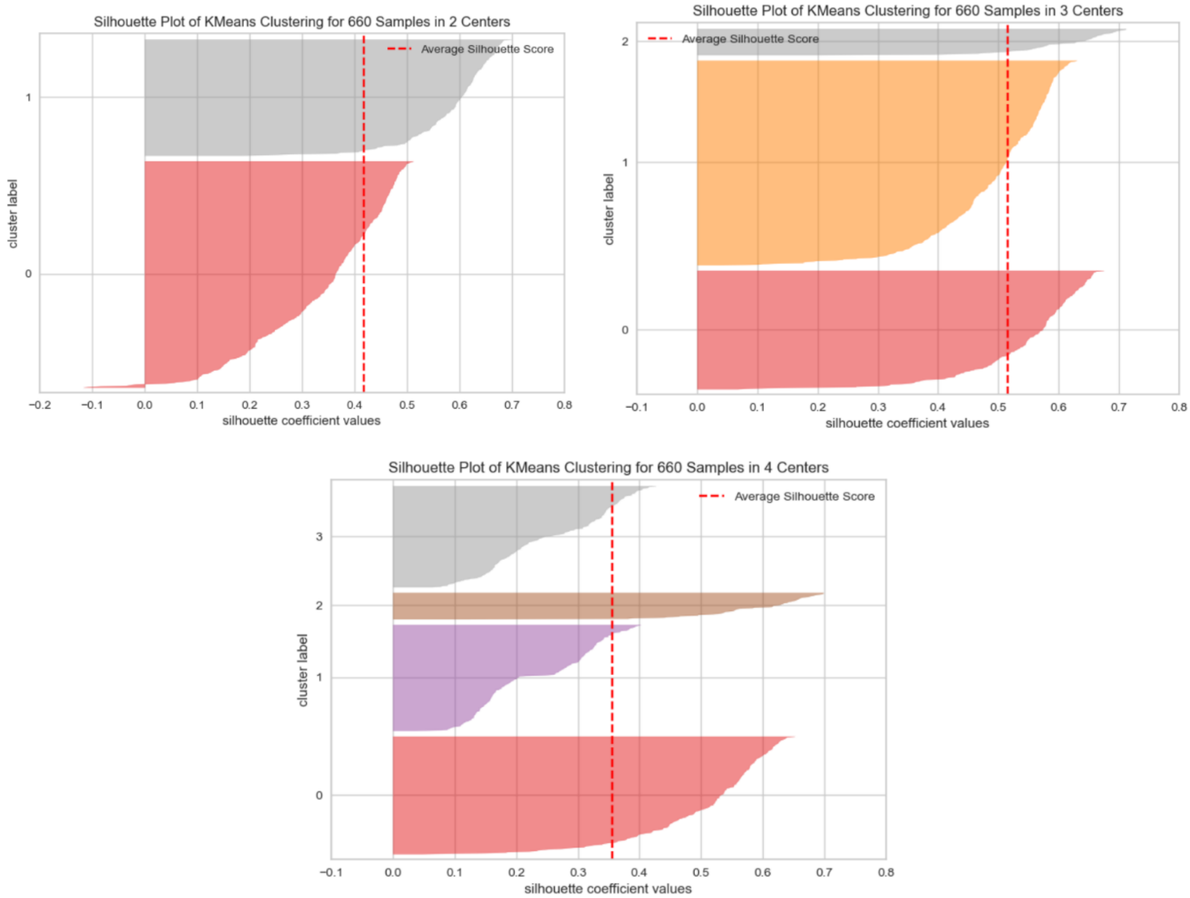


Figure 16 - Silhouette Scores of Various Clusters

```
For n_clusters = 2, the silhouette score is 0.41842496663215445)
For n_clusters = 3, the silhouette score is 0.5157182558881063)
For n_clusters = 4, the silhouette score is 0.3556670619372605)
For n_clusters = 5, the silhouette score is 0.2434505689196401)
For n_clusters = 6, the silhouette score is 0.24269461746124507)
For n_clusters = 7, the silhouette score is 0.24956289249330887)
For n_clusters = 8, the silhouette score is 0.21866686700173374)
For n_clusters = 9, the silhouette score is 0.18602042975434668)
For n_clusters = 10, the silhouette score is 0.21210329269212566)
```

From the above graph, the highest score is 0.51, indicating that the best value of k is 3 so we will choose k=3.

Figure 17 - Visualization of Silhouette Coefficients



The graphs above give us a visual representation of the different clusters and their silhouette coefficients. when k=3, it has the highest coefficient score of around 0.5 and it shows a good distribution of data points

We can consider the final number of clusters (k) as 3.

4.3 Cluster Profiling

Figure 18 - K-means Cluster Profile

KM_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	Freq
0	33782.383420	5.515544	3.489637	0.981865	2.000000	386
1	12174.107143	2.410714	0.933036	3.553571	6.870536	224
2	141040.000000	8.740000	0.600000	10.900000	1.080000	50

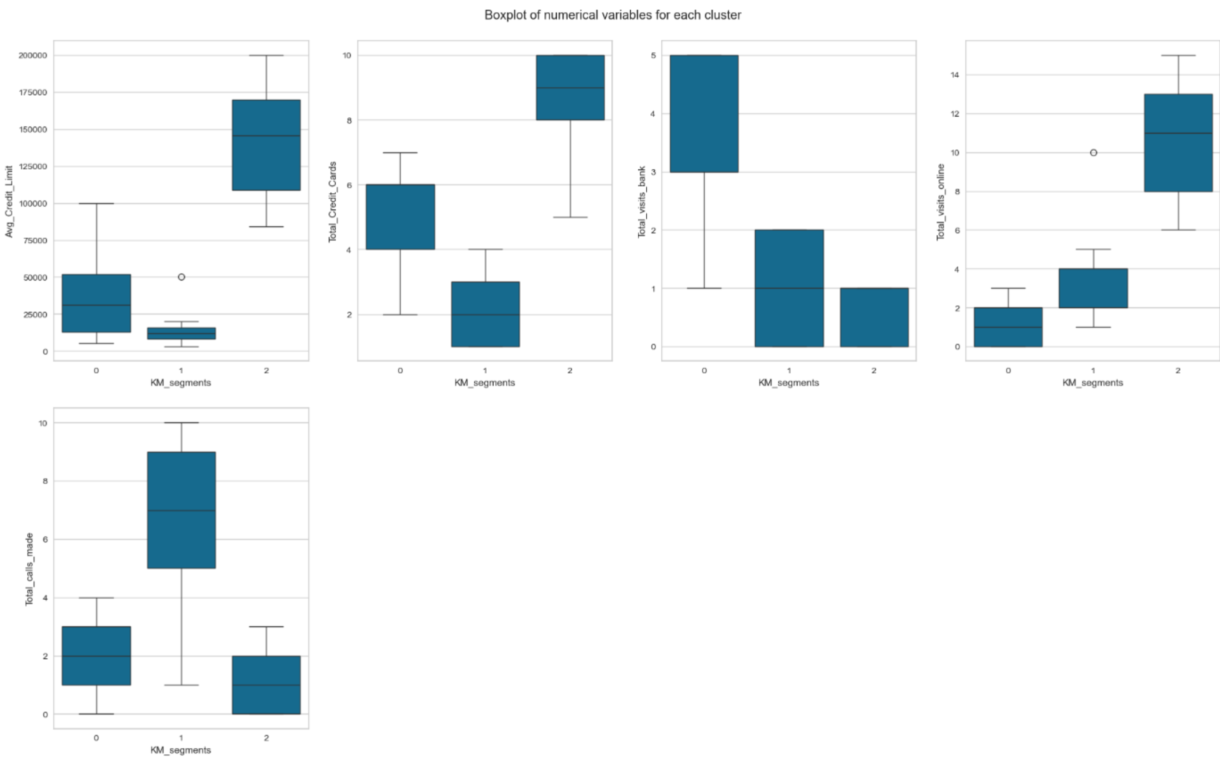
From the above cluster profiles, we can see that

- Cluster 0 has 386 customers (~58.5%)
- Cluster 1 has 224 customers (~34%)

- Cluster 2 has 50 customers (~7.5%)

The highest mean values of each column are highlighted in green. Cluster 2 has the highest mean for Average Credit Limit (141040.00), Total Credit Cards (8.74), Total Visits Online (10.9). Cluster 1 has the highest mean for Total Calls Made (6.87). Cluster 0 has the highest mean for Total Visits Bank (3.48).

Figure 19 - Boxplot of All Clusters - K-means



Insights:

- **Cluster 0**
    - There are 386 customers in this cluster.
    - This cluster of customers have a medium level of average credit limit up to 100,000.
    - Total credit cards issued to the customers in cluster range from 2-7.
    - The customers in this cluster have the greatest number of bank visits up to 5.
    - The customers in this cluster have the least number of online visits up to 3.
    - Total calls per year made by customers in this cluster range from 0-4.
    - Overall, this cluster consists of the mid-end customers.

- **Cluster 1**
  - There are 224 customers in this cluster.
  - This cluster of customers have the lowest average credit limit below 25,000.
  - Total credit cards issued to the customers range from 0-4.
  - The customers in this cluster visit the bank maximum of 2 times.
  - The customers in this cluster have used online banking 1-5 times per year.
  - This cluster of customers make the most calls per year ranging from 1-10.
  - Overall, this cluster consists of the low-end customers.
- **Cluster 2**
  - There are 50 customers in this cluster.
  - This cluster of customers have the highest average credit limit of up to 200,000.
  - Total credit cards issued to the customers range from 5-10.
  - The customers in this cluster visit the bank least number of times up to 1 per year.
  - The customers in this cluster have used online banking the most of up 6-15 times per year.
  - This cluster of customers make the least calls per year ranging from 0-3.
  - Overall, this cluster consists of the high-end customers.

HIERARCHICAL CLUSTERING

Hierarchical Clustering is an unsupervised machine learning algorithm that builds a hierarchy of clusters either by iteratively merging smaller clusters (agglomerative) or by splitting larger clusters (divisive). In agglomerative hierarchical clustering, each data point starts as its own cluster, and the algorithm progressively merges the closest pairs of clusters based on a distance metric. The result is a dendrogram which shows the merging process. The algorithm continues until all points belong to one single cluster or until a desired number of clusters is achieved.

5.1 Cophenetic Correlation

Cophenetic correlation is a measure used to assess the accuracy of a hierarchical clustering algorithm. It quantifies how well the clustering solution preserves the original distances between data points. Different types of distance measures can be used depending on the nature of the data and the goals of the analysis, for this analysis we will analyse the following distance measures and choose the best result

- Euclidean Distance
- Chebyshev Distance
- Mahalanobis Distance
- Cityblock Distance

Linkage methods are techniques used in hierarchical clustering to determine the distance between clusters during the clustering process. We will calculate the cophenetic correlation for the following linkage methods.

- Single Linkage
- Complete Linkage
- Average Linkage
- Weighted Linkage

Figure 20 - Cophenetic Correlation of Various Distance Measures

```
Cophenetic correlation for Euclidean distance and single linkage is 0.7391220243806552.
Cophenetic correlation for Euclidean distance and complete linkage is 0.8599730607972423.
Cophenetic correlation for Euclidean distance and average linkage is 0.8977080867389372.
Cophenetic correlation for Euclidean distance and weighted linkage is 0.8861746814895477.
Cophenetic correlation for Chebyshev distance and single linkage is 0.7382354769296767.
Cophenetic correlation for Chebyshev distance and complete linkage is 0.8533474836336782.
Cophenetic correlation for Chebyshev distance and average linkage is 0.8974159511838106.
Cophenetic correlation for Chebyshev distance and weighted linkage is 0.8913624010768603.
Cophenetic correlation for Mahalanobis distance and single linkage is 0.7058064784553606.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.5422791209801747.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.8326994115042134.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.7805990615142516.
Cophenetic correlation for Cityblock distance and single linkage is 0.7252379350252723.
Cophenetic correlation for Cityblock distance and complete linkage is 0.8731477899179829.
Cophenetic correlation for Cityblock distance and average linkage is 0.896329431104133.
Cophenetic correlation for Cityblock distance and weighted linkage is 0.8825520731498188.
```

From the above figure, we can see the various cophenetic correlation scores for various distance measures and linkage methods. The Highest cophenetic correlation is ~0.898, which is obtained with Euclidean distance and average linkage.

5.2 Plotting Dendrograms

A dendrogram is a tree-like diagram used to visualize the results of hierarchical clustering, which is a method of clustering objects based on their similarity. The dendrogram illustrates how clusters are merged or split, reflecting the hierarchical structure of the data.

Figure 21 - Single Linkage Dendrogram

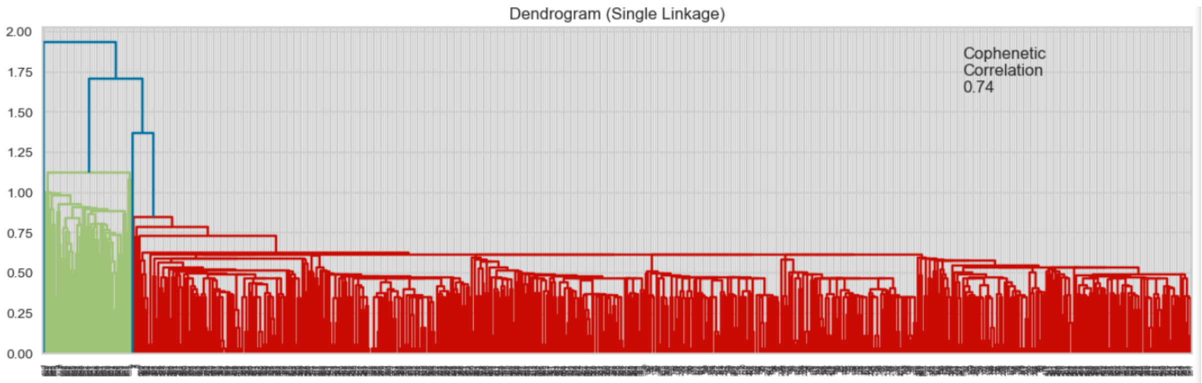


Figure 22 - Complete Linkage Dendrogram

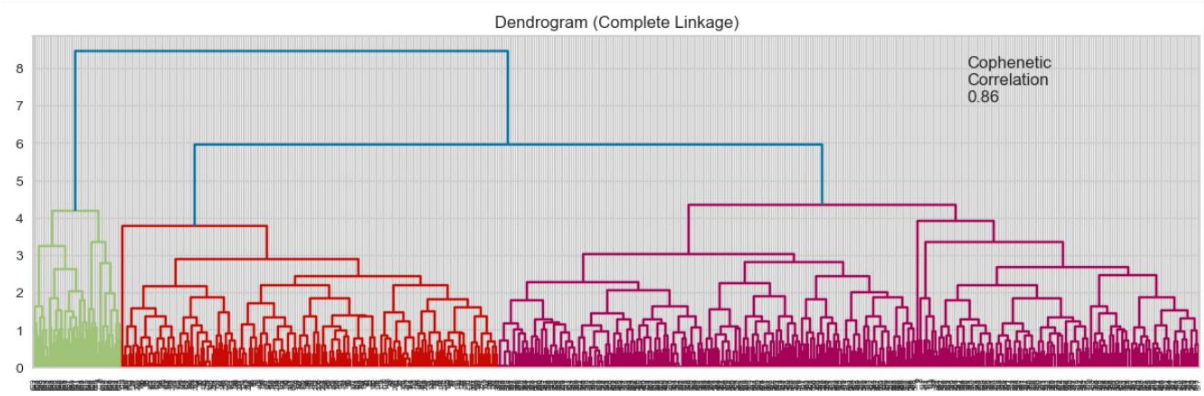


Figure 23 - Average Linkage Dendrogram

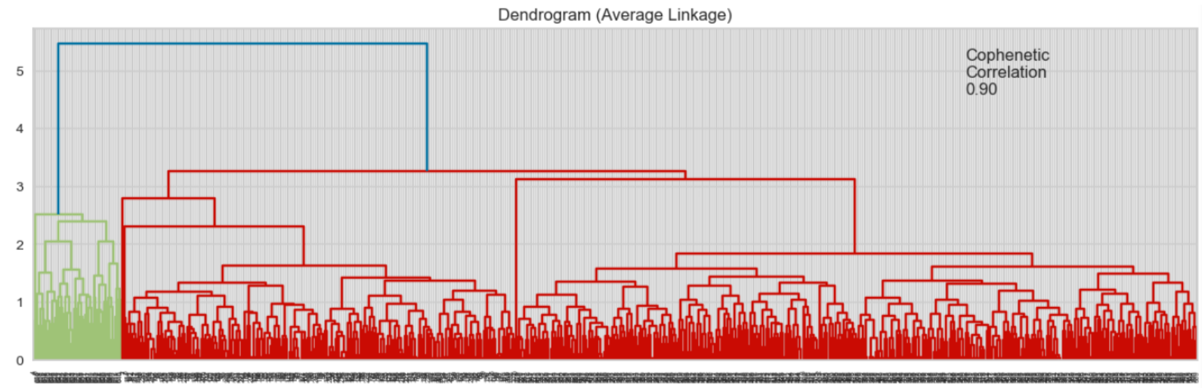


Figure 24 - Centroid Linkage Dendrogram

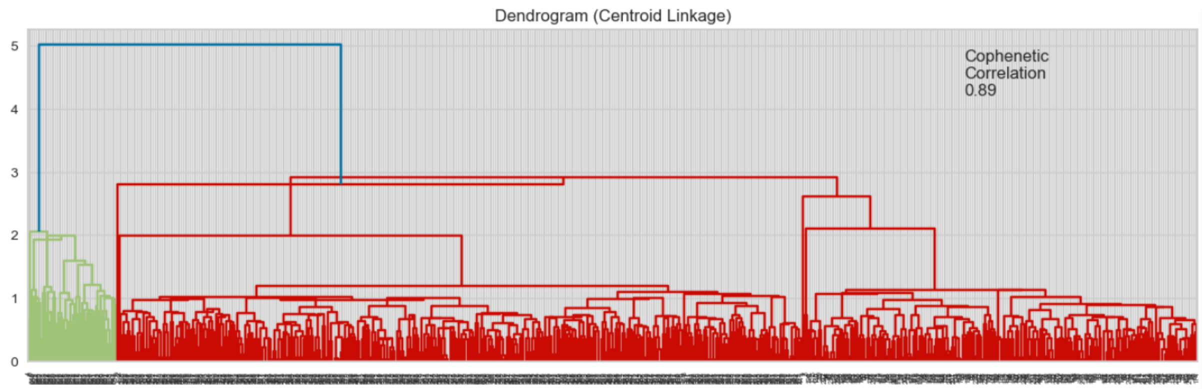


Figure 25 - Ward Linkage Dendrogram

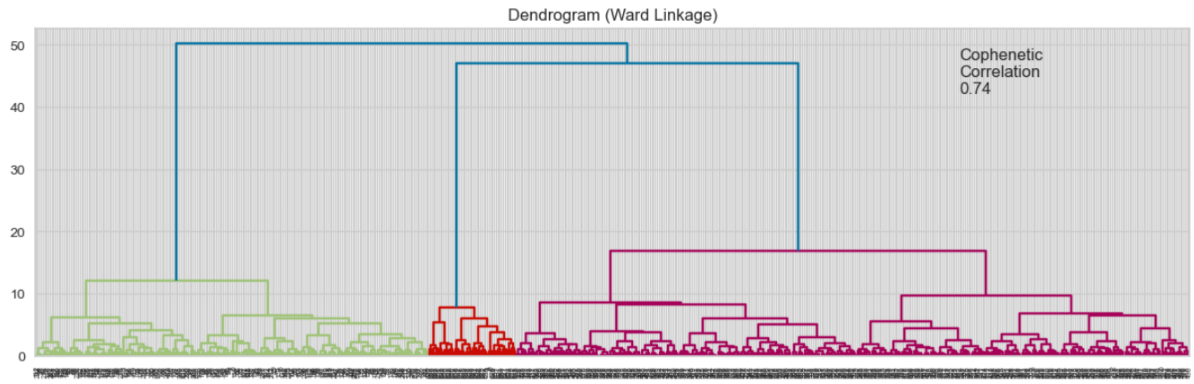
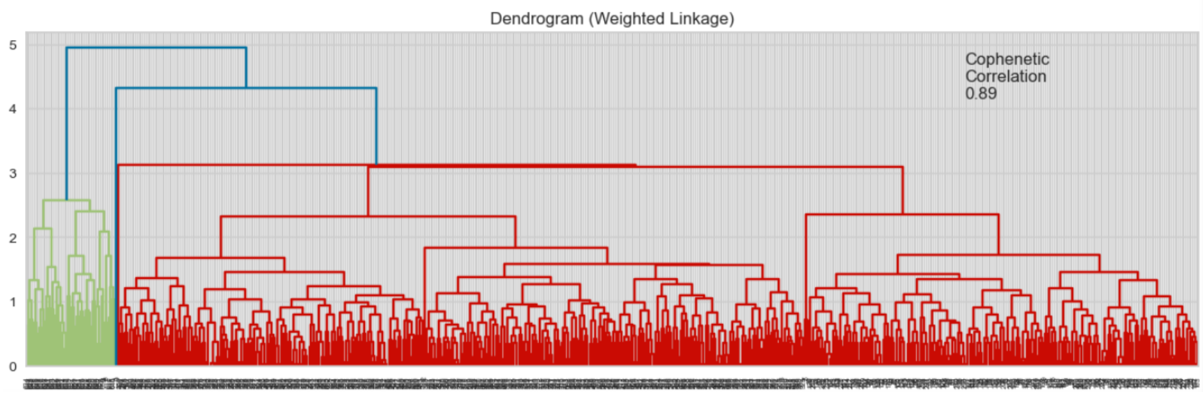


Figure 26 - Weighted Linkage Dendrogram



Visualizing the dendrograms, we can see that average linkage method with Euclidian distance has the highest cophenetic correlation at 0.90. The closer the score is to 1, the better is the clustering process. We can build the final model based on Euclidean distance and average linkage method.

Figure 27 - Cophenetic Coefficient Comparison

	Linkage	Cophenetic Coefficient
0	single	0.739122
4	ward	0.741516
1	complete	0.859973
5	weighted	0.886175
3	centroid	0.893939
2	average	0.897708

5.3 Cluster Profile

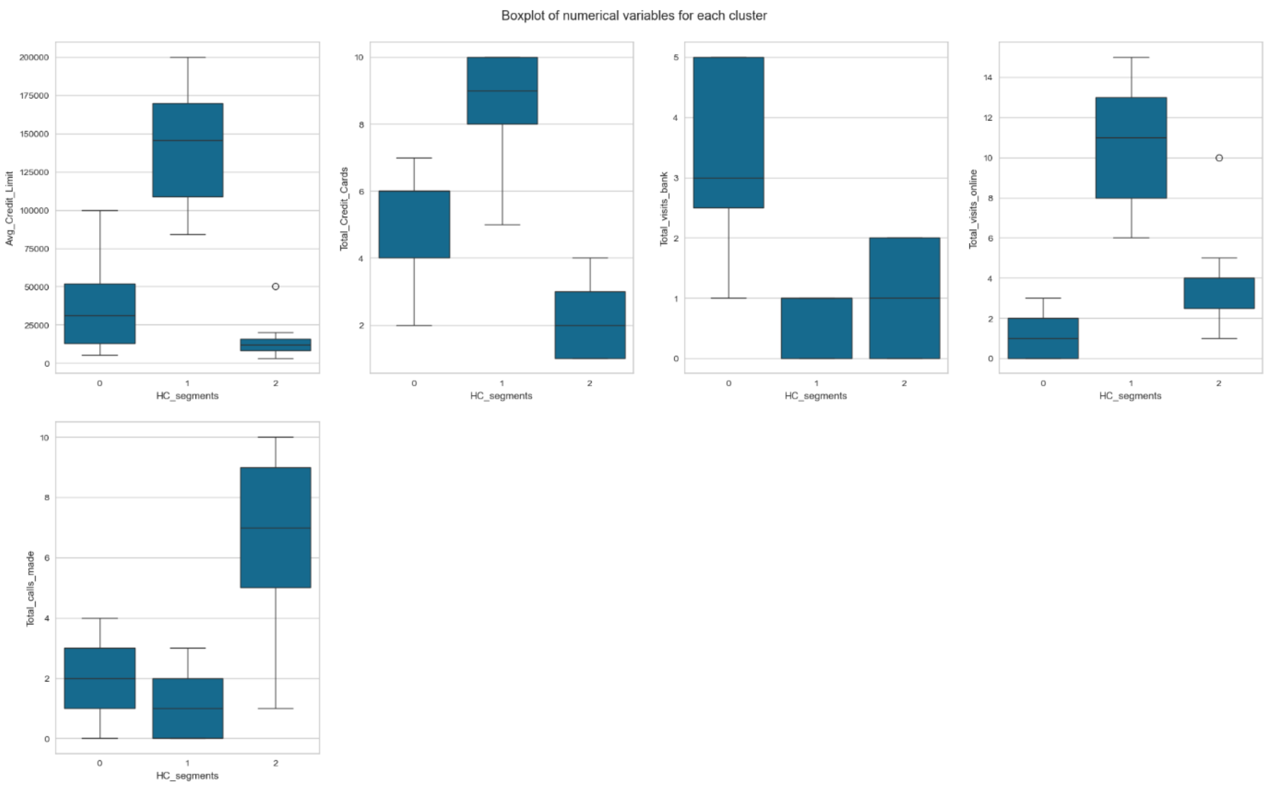
Figure 28 - Hierarchichal Clustering Cluster Profile

HC_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	Freq
0	33713.178295	5.511628	3.485788	0.984496	2.005168	387
1	141040.000000	8.740000	0.600000	10.900000	1.080000	50
2	12197.309417	2.403587	0.928251	3.560538	6.883408	223

From the above cluster profile, we can see that

- Cluster 0 has 387 customers (59%)
- Cluster 1 has 50 customers (8%)
- Cluster 2 has 223 customers (33%)

Figure 29 - Boxplot of All Clusters - Hierarchical Clustering



Insights:

- **Cluster 0**
    - There are 387 customers in this cluster.
    - This cluster of customers have a medium level of average credit limit up to 100,000.

- Total credit cards issued to the customers in cluster range from 2-7.
- The customers in this cluster have the greatest number of bank visits up to 5.
- The customers in this cluster have the least number of online visits up to 3.
- Total calls per year made by customers in this cluster range from 0-4.
- Overall, this cluster consists of the mid-end customers.

- **Cluster 1**
    - There are 50 customers in this cluster.
    - This cluster of customers have the highest average credit limit of up to 200,000.
    - Total credit cards issued to the customers range from 5-10.
    - The customers in this cluster visit the bank least number of times up to 1 per year.
    - The customers in this cluster have used online banking the most of up 6-15 times per year.
    - This cluster of customers make the least calls per year ranging from 0-3.
    - Overall, this cluster consists of the high-end customers.
- **Cluster 2**
    - There are 223 customers in this cluster.
    - This cluster of customers have the lowest average credit limit below 25,000.
    - Total credit cards issued to the customers range from 1-4.
    - The customers in this cluster visit the bank within the range of 0-2 times.
    - The customers in this cluster have used online banking 1-5 times per year.
    - This cluster of customers make the most calls per year ranging from 1-10.
    - Overall, this cluster consists of the low-end customers.

K-MEANS VS HIERARCHICHAL CLUSTERING

Both clustering algorithms can be compared to determine which method provides more meaningful, reliable, and interpretable results based on the data.

- The Clusters formed in K-means is very similar to the clusters in Hierarchical clustering.
- The average means of cluster 0 from K-means is identical to the average means of cluster 0 from hierarchical cluster. The average means of Cluster 1 from K-means is identical to Cluster 2 from hierarchical cluster. The average means of Cluster 2 from K-means is identical to cluster 1 from hierarchical cluster.
- Both the clustering algorithms obtain a total 3 clusters from the dataset, a total of 3 customer segments can be identified.
- The K-means cluster 0 has 386 observations compared to the hierarchical cluster 0 with 387 observations. The k-means cluster 2 has 224 observations compared to hierarchical cluster 1 with 223 observations. The observations of k-means cluster 2 and hierarchical cluster 1 are the same at 50.

As both the cluster algorithms produce very similar clusters, it is more suitable to use hierarchical clustering technique as it is more stable and offers more flexibility.

ACTIONABLE INSIGHTS AND RECOMMENDATIONS

We have identified 3 different customer segments based on the existing customer data using clustering algorithms.

- Group 1 (Credit limit up to 20,000) can be classified as Low-end Segment
- Group 2 (Credit limit up to 100,000) can be classified as Mid-end Segment
- Group 3 (Credit limit up to 200,000) can be classified as High-end Segment

Here are some recommendations to better target these segments.

Group 1 (Low-end):

- Market small loans, credit-building products, or basic credit cards designed to help improve their credit score and provide easier access to credit.
- Providing enhanced customer support through call centers or online chat to resolve frequent queries and simplify their banking experience.
- As this segment uses both phone calls and personal bank visits, market hybrid banking services like online loan applications that can be completed in-branch for added support.

Group 2 (Mid-end):

- This segment visits the bank frequently, focus marketing efforts on in-branch promotions, such as personalized financial advisory or consultations.
- Market products like savings accounts, personal loans, and credit cards which align with their credit habits and preference for physical banking.
- Create a loyalty program rewarding frequent bank visits, offering perks like exclusive bank promotions.
- Marketing insurance products and wealth management services as they are likely to value face-to-face interactions for financial decisions.

Group 3 (High-end):

- Target this segment with online banking services with mobile apps, online loan applications, and digital wealth management tools, to match their preference for convenience of online engagement.

- Given their high credit limit and relatively low engagement with physical banking, market premium financial products like investment accounts, credit cards with higher limits, or personalized wealth management services.
- Develop digital-only promotions, such as better interest rates or cashback offers for transactions made online, to cater to their digital-first approach.
- As this segment has low bank visits, offer virtual consultations with financial advisors to help them with portfolio management or investment planning.
- Specific email marketing campaigns offering exclusive digital banking products, credit line increases, and benefits designed for high-net-worth individuals can be marketed to this segment.