Business Report

AllLife Bank Credit Card Customer Segmentation

– Ruchi Rikta

Contents

List of Figures

Context

AllLife 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

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.

Data Description

The data 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).

Data Dictionary

- 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).

Data Overview

- **1st 5 rows of the Dataset**

	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	1	87073	100000	2	1	1	0
1	2	38414	50000	3	0	10	9
2	3	17341	50000	7	1	3	4
3	4	40496	30000	5	1	1	4
4	5	47437	100000	6	0	12	3

- **Data Type**

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 660 entries, 0 to 659
Data columns (total 7 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Sl_No                660 non-null    int64
 1   Customer Key         660 non-null    int64
 2   Avg_Credit_Limit     660 non-null    int64
 3   Total_Credit_Cards   660 non-null    int64
 4   Total_visits_bank    660 non-null    int64
 5   Total_visits_online  660 non-null    int64
 6   Total_calls_made     660 non-null    int64
dtypes: int64(7)
memory usage: 36.2 KB
```

Observations:

- There are 660 observations and 7 columns in the dataset.
- All columns have 660 non-null values, i.e., there are no missing values.
- All columns are of int64 data type.
- There are no missing values.

- **Statistical Summary**

	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
count	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000
mean	330.500000	55141.443939	34574.242424	4.706061	2.403030	2.606061	3.583333
std	190.669872	25627.772200	37625.487804	2.167835	1.631813	2.935724	2.865317
min	1.000000	11265.000000	3000.000000	1.000000	0.000000	0.000000	0.000000
25%	165.750000	33825.250000	10000.000000	3.000000	1.000000	1.000000	1.000000
50%	330.500000	53874.500000	18000.000000	5.000000	2.000000	2.000000	3.000000
75%	495.250000	77202.500000	48000.000000	6.000000	4.000000	4.000000	5.000000
max	660.000000	99843.000000	200000.000000	10.000000	5.000000	15.000000	10.000000

Observations:

- The Avg_Credit_Limit has a high range from 3000 to 200000 in money. The average is approximately 34543, but with a high standard deviation of 37428.
- The max amount on the Total_Credit_Cards is 10, and the lowest is 1, with a mean of 4 or 5 credit cards.
- There is a low range on Total_visits_bank from 0 to 5, and a mean of 2 visits.
- Total_visits_online has the highest range, from 0 to 15, for online visits. This is logical because people prefer to resolve problems through a website rather than calling or going to the bank, and wasting time.
- On Total_calls_made, have an average of between 3 and 4 calls.

Exploratory Data Analysis

1. Univariate Data Analysis

1.1. Observations on Average Credit Limit

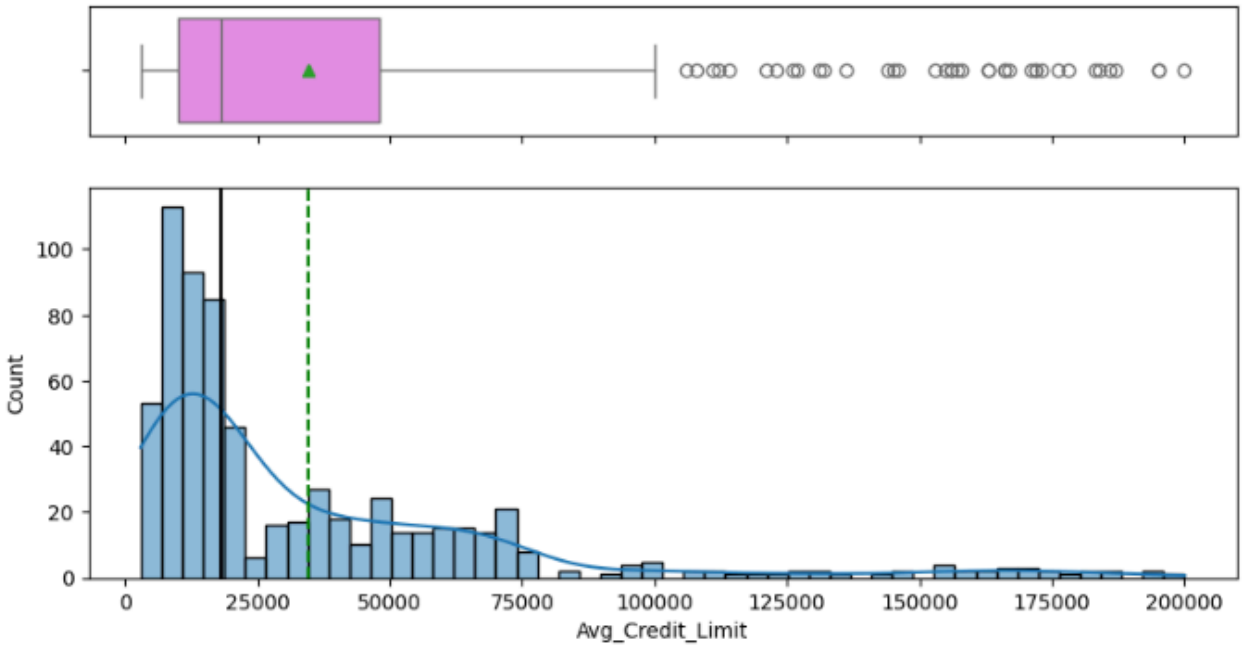


Fig.1.1: Univariate Analysis Average Credit Limit

- It has outliers, there are a few people with a lot of credit limit on their credit cards. So those values are acceptable.

1.2 Observations on Total Credit Cards

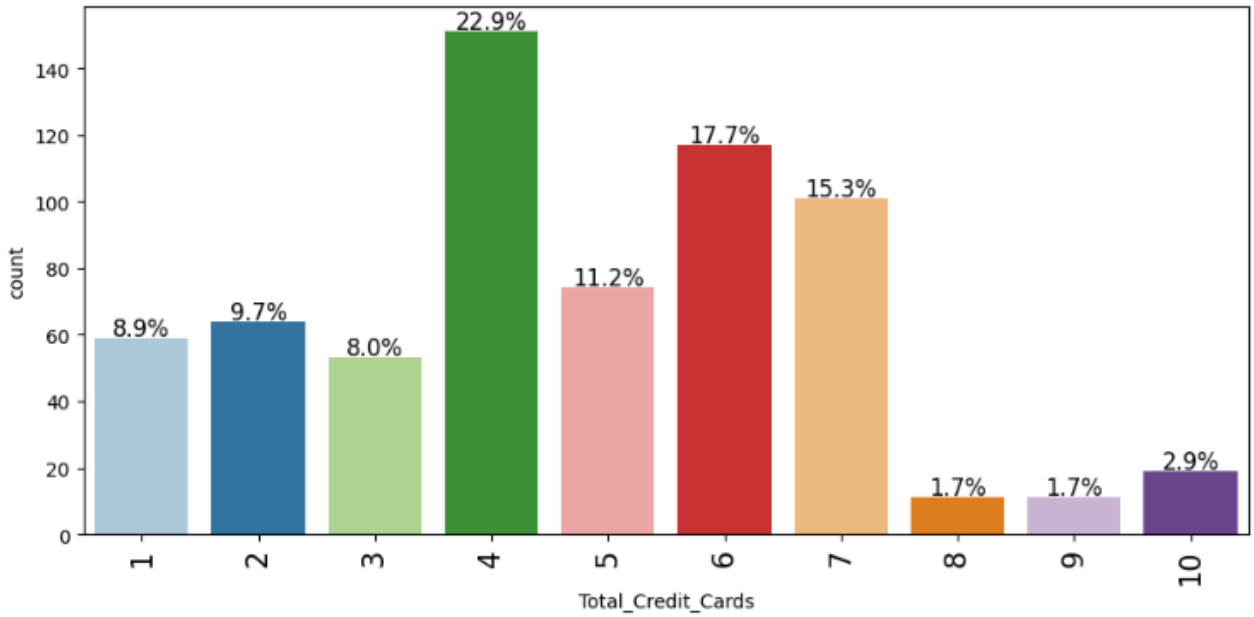


Fig.1.2: Univariate Analysis Total Credit Cards

- 22.9% people prefer having 4 credit cards, 17.7% prefer 6, and 15.3% prefer 7.

1.3 Observations on Total Number of Bank Visits

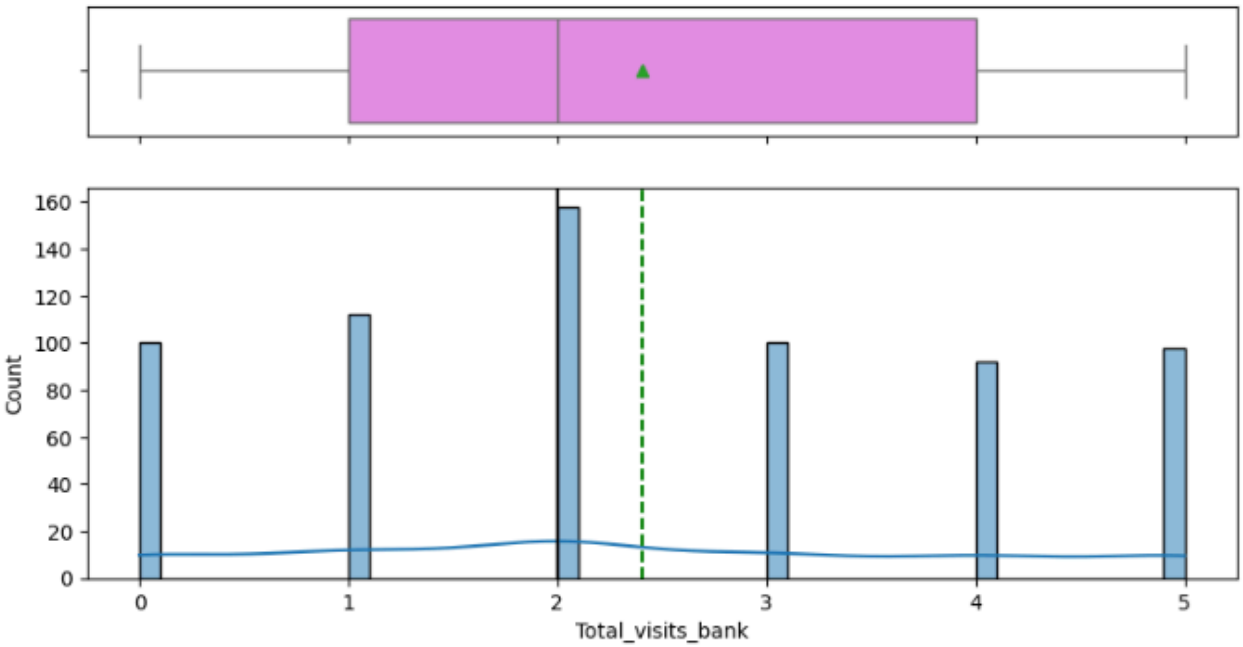


Fig.1.3: Univariate Analysis Total Number of Bank Visits

- The majority of customers have made between 1 to 3 bank visits.
- The distribution is slightly left-skewed.

1.4 Observations on Total Number of Online Visits

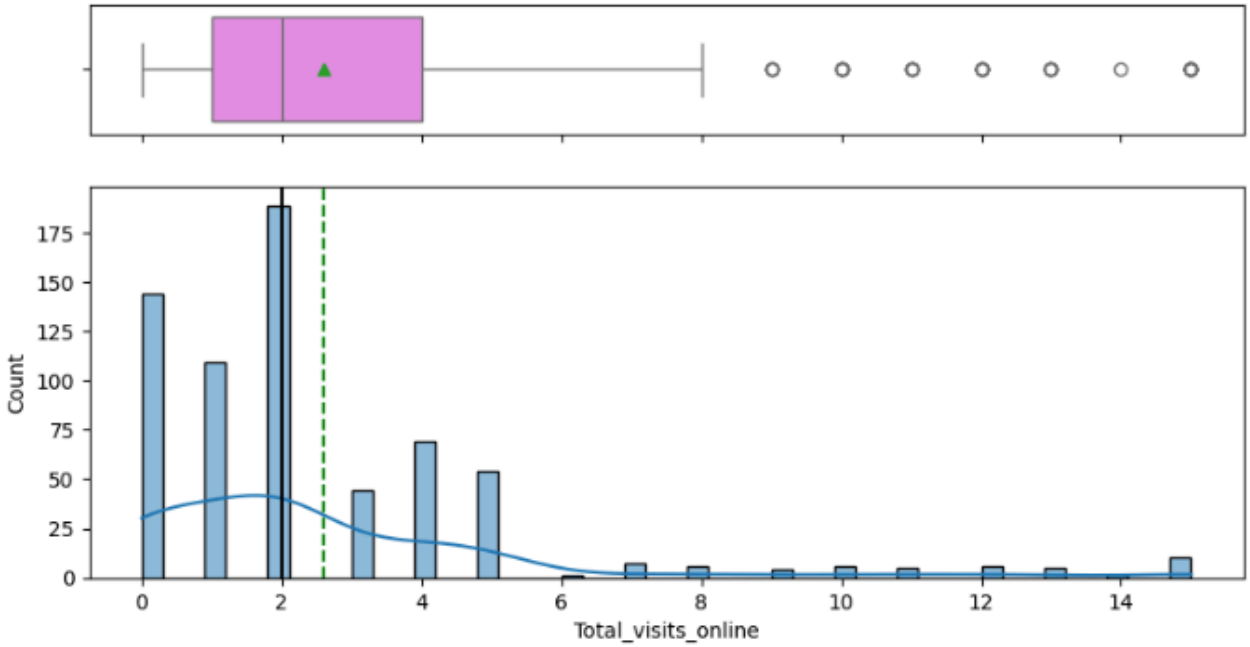


Fig.1.4: Univariate Analysis Total Number of Online Visits

- There are outliers.
- The majority of customers made between 1 to 4 online visits.
- The distribution is right-skewed.

1.5 Observations on Total Number of Calls Made

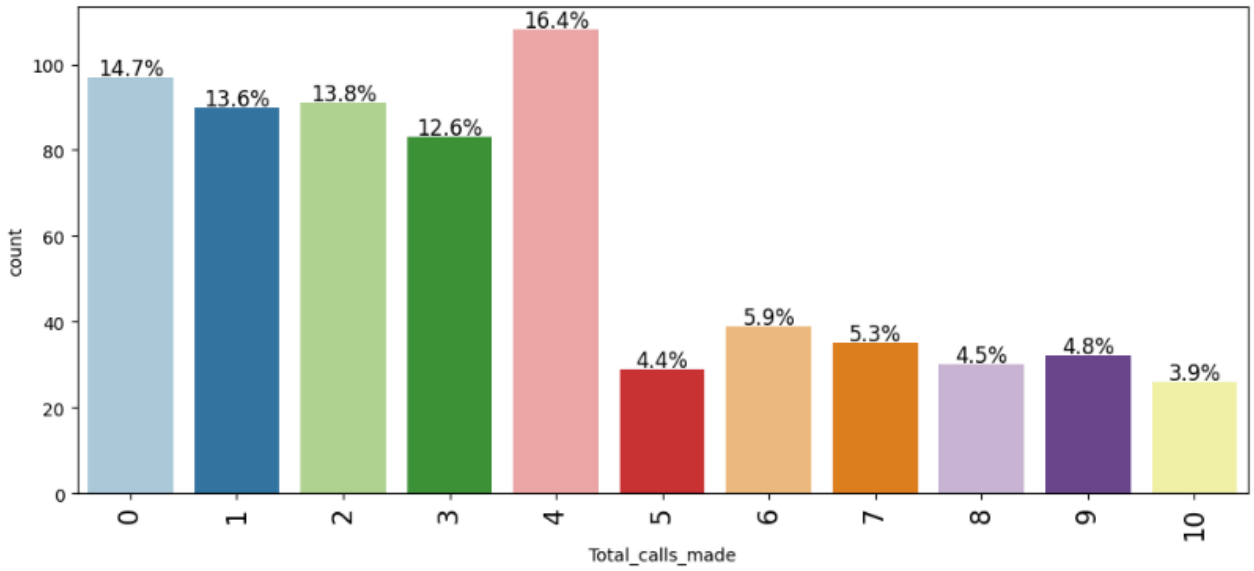


Fig.1.5: Univariate Analysis Total Number of Calls Made

- The majority of customers made 4 calls.

1. Bivariate Data Analysis

2.1 Correlation between numerical values

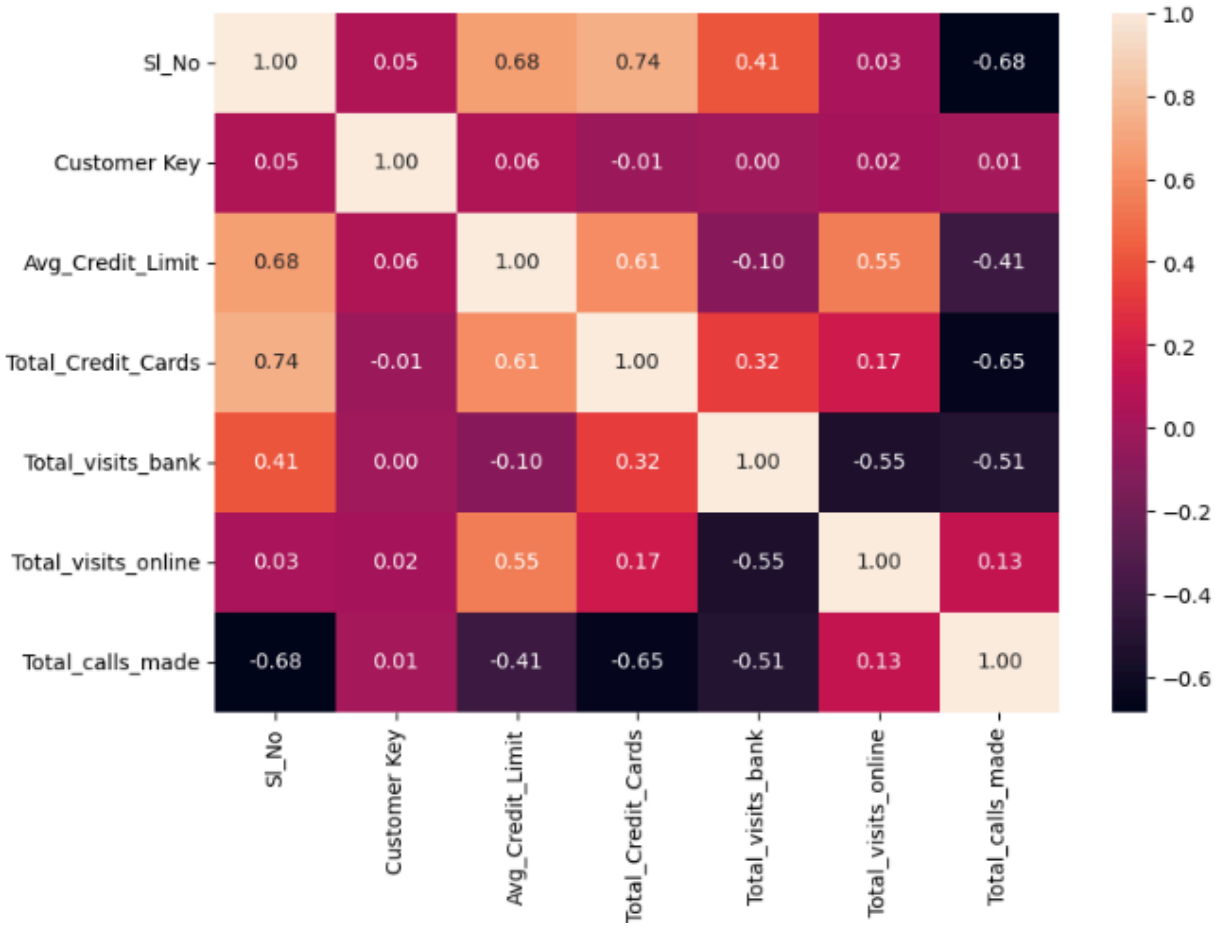


Fig.2.1: Correlation Between Numerical Values

Observation:

- Avg_Credit_Limit is positively correlated with Total_Credit_Cards and Total_visits_online, which makes sense.
- Avg_Credit_Limit is negatively correlated with Total_calls_made and Total_visits_bank.

- Total_visits_bank, Total_visits_online, and Total_calls_made are negatively correlated, which implies that the majority of customers use only one of these channels to contact the bank.

2.2 Pairplot

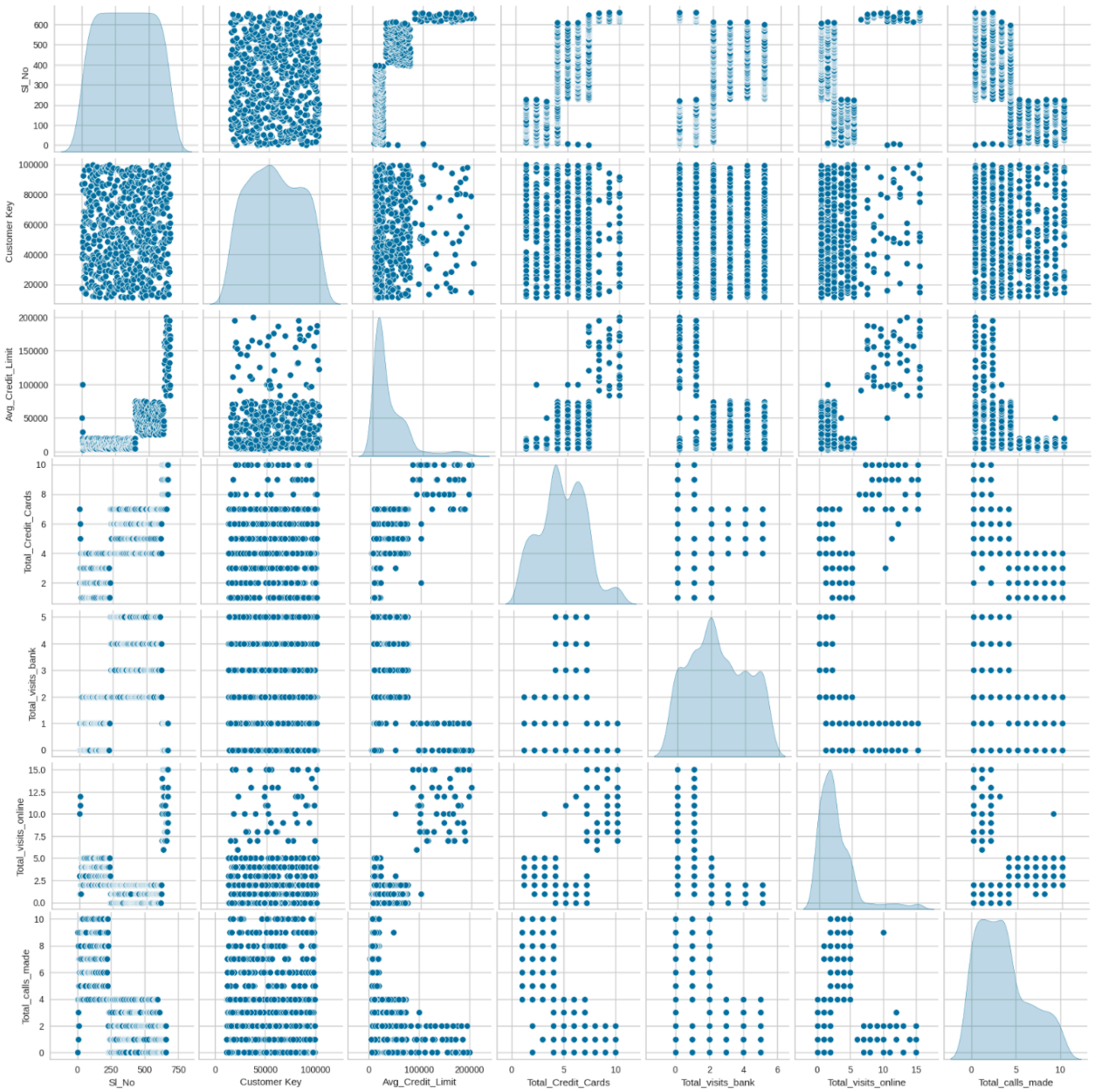


Fig.2.2 : Pairplot

Observation:

- Avg_Credit_Limit and Total_visits_online are rightly skewed.
- Total_Credit_Cards is distributed multi-modal.

Data Processing

- **Checking Missing Values**

	0
Avg_Credit_Limit	0
Total_Credit_Cards	0
Total_visits_bank	0
Total_visits_online	0
Total_calls_made	0

dtype: int64

- No missing values found.

- **Outlier Detection**

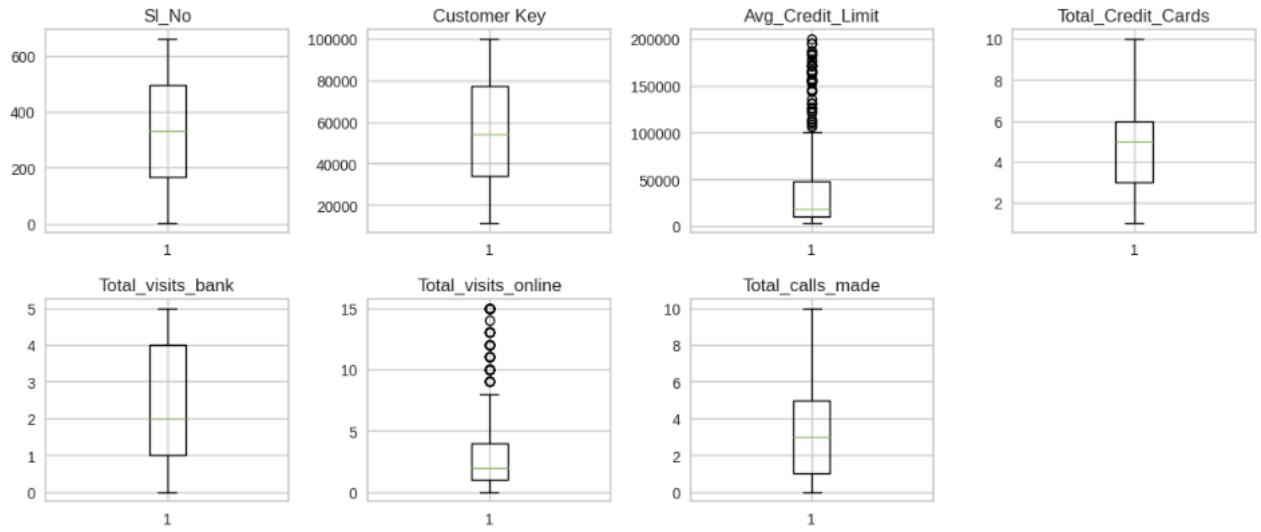


Fig.3: Boxplot Outlier Detection

- Avg_Credit_Limit and Total_visits_online have outliers, but we will keep them as they may contain some valuable input.

- **Data Scaling**

	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	-1.729428	1.246920	1.740187	-1.249225	-0.860451	-0.547490	-1.251537
1	-1.724180	-0.653203	0.410293	-0.787585	-1.473731	2.520519	1.891859
2	-1.718931	-1.476098	0.410293	1.058973	-0.860451	0.134290	0.145528
3	-1.713683	-0.571901	-0.121665	0.135694	-0.860451	-0.547490	0.145528
4	-1.708434	-0.300857	1.740187	0.597334	-1.473731	3.202298	-0.203739

- The first 5 rows of the Scaled data.

K-means Clustering

- **Elbow Curve**

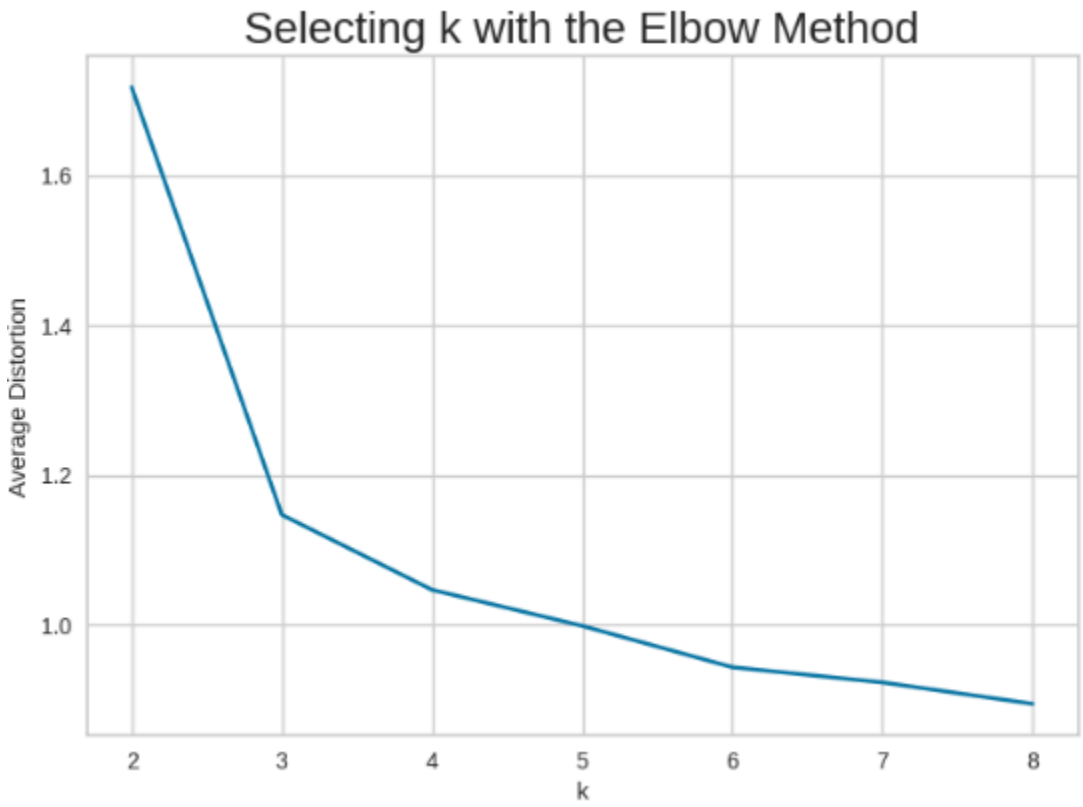


Fig.4: Elbow Curve Plot

- The appropriate value of k from the elbow curve seems to be 3 or 4.

- **Silhouette Score**

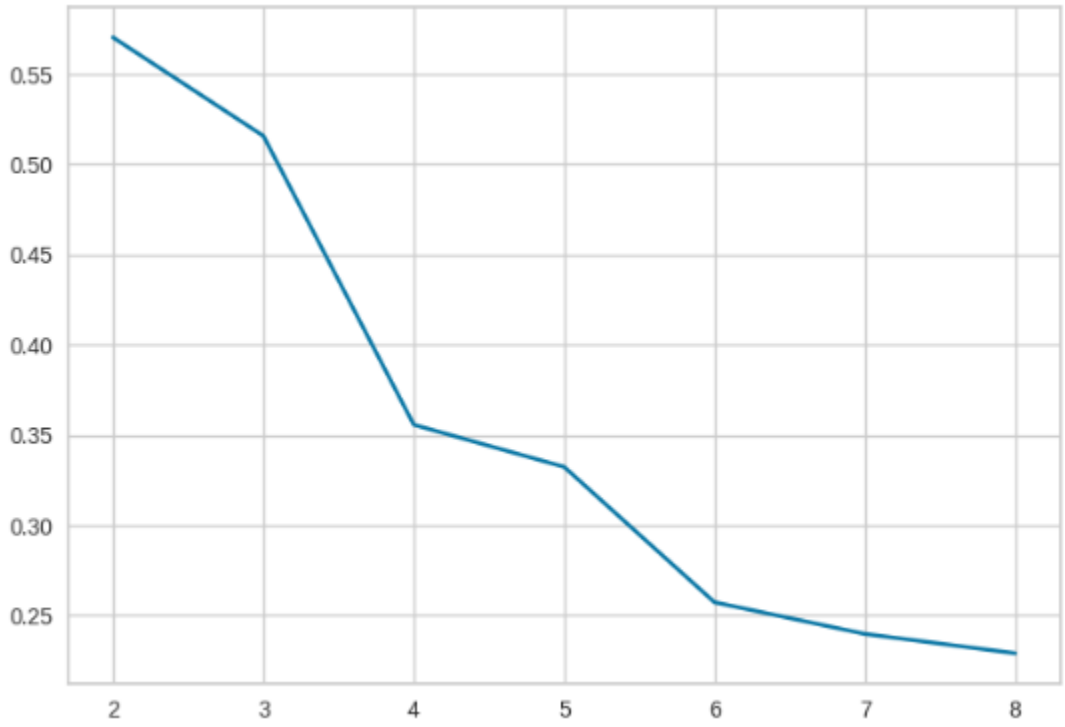


Fig.5: Silhouette Score (K-Means)

- From the silhouette scores, it seems that 3 is a good value of k.

- **Silhouette Plot of KMeans Clustering**

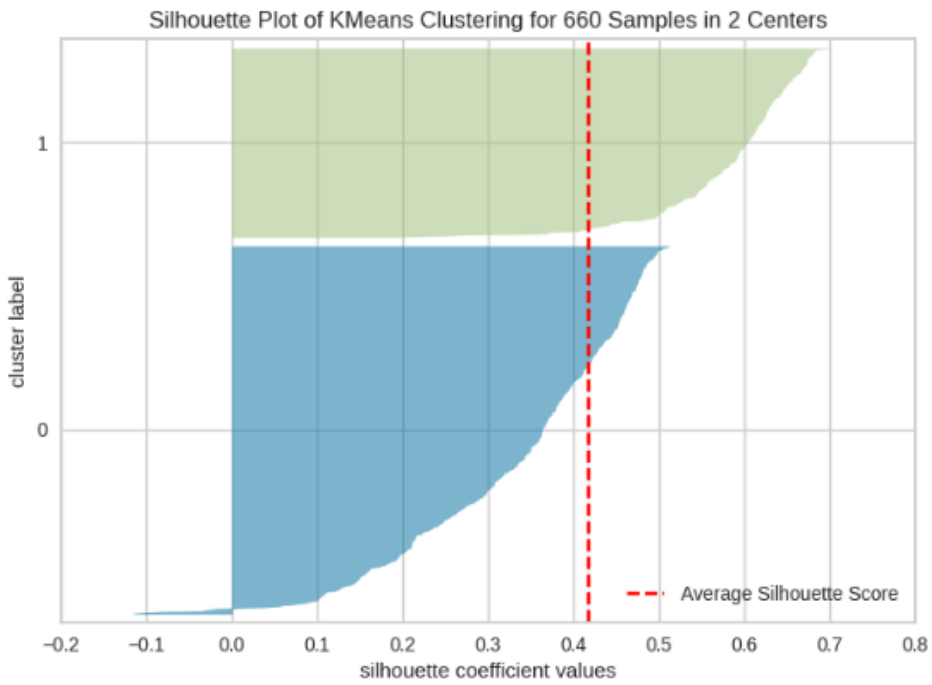


Fig.6: Silhouette Plot in 2 centers

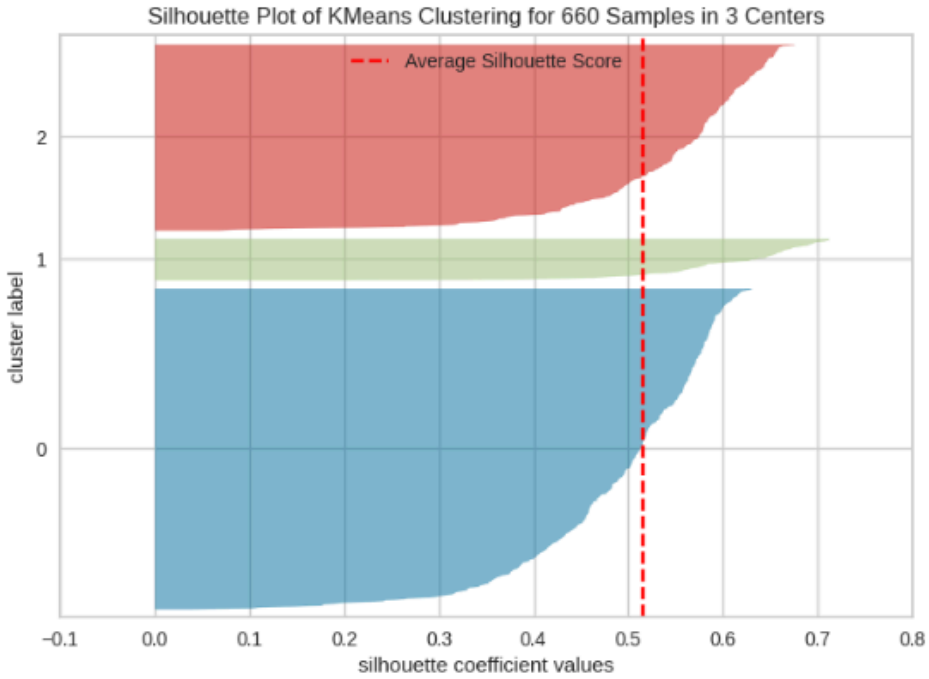


Fig.7: Silhouette Plot in 3 centers

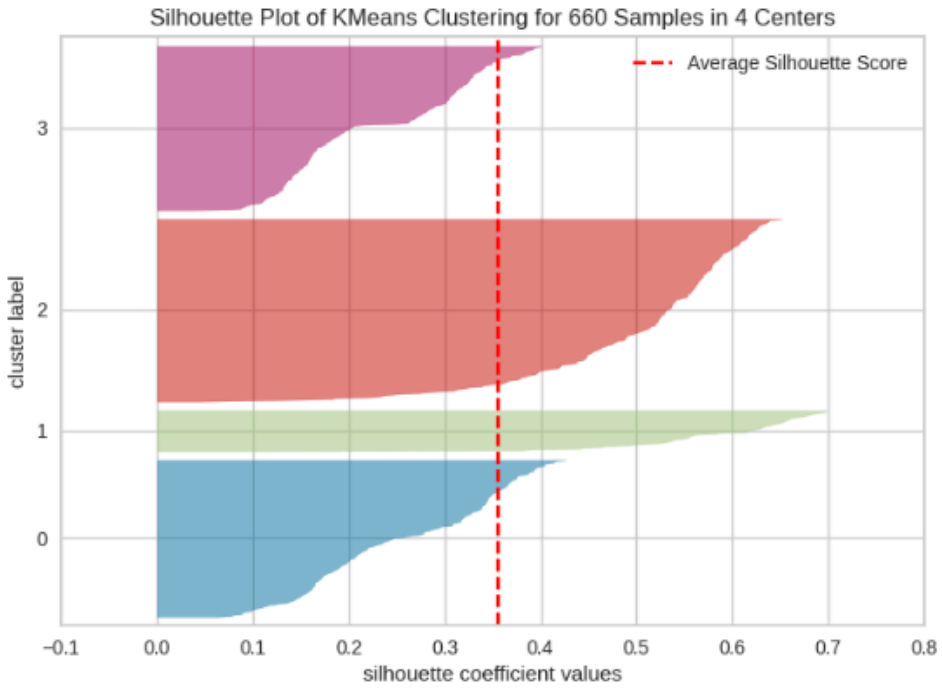


Fig.8: Silhouette Plot in 4 centers

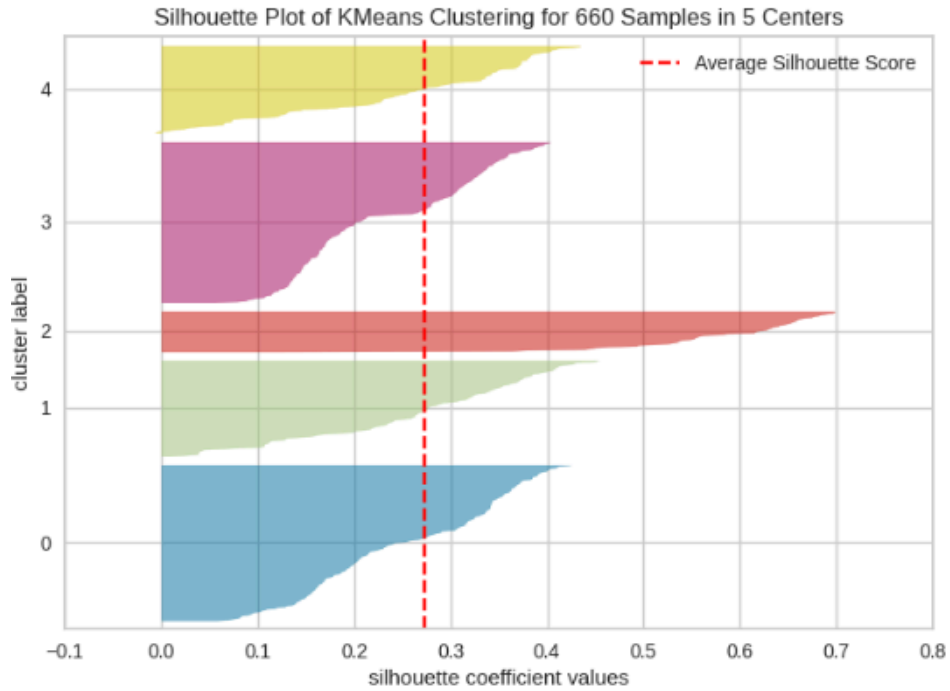


Fig.9: Silhouette Plot in 5 centers

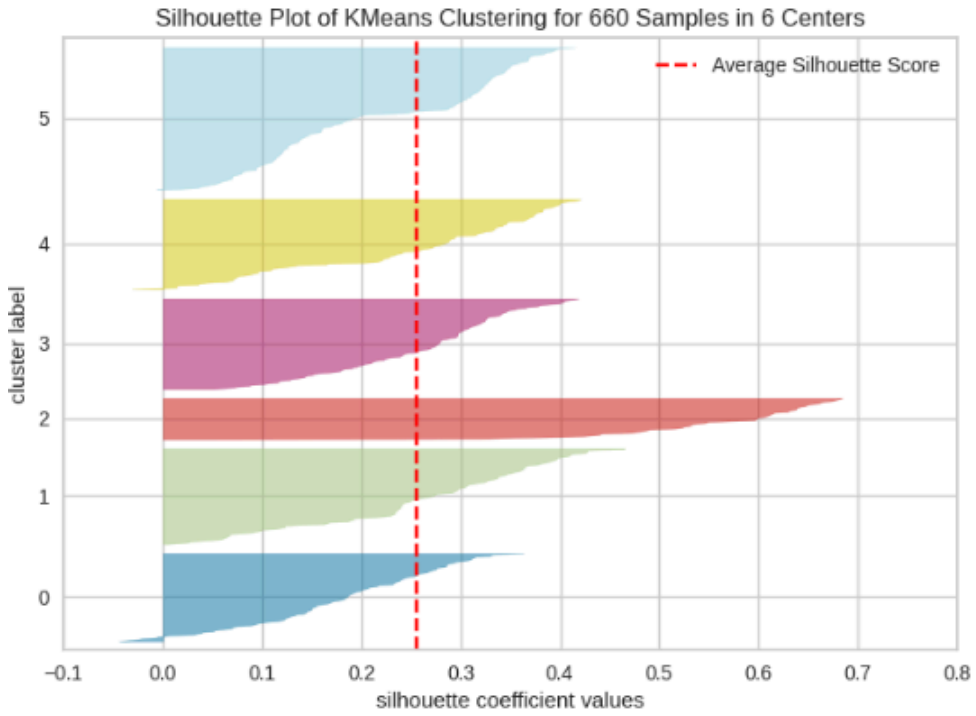


Fig.10: Silhouette Plot in 6 centers

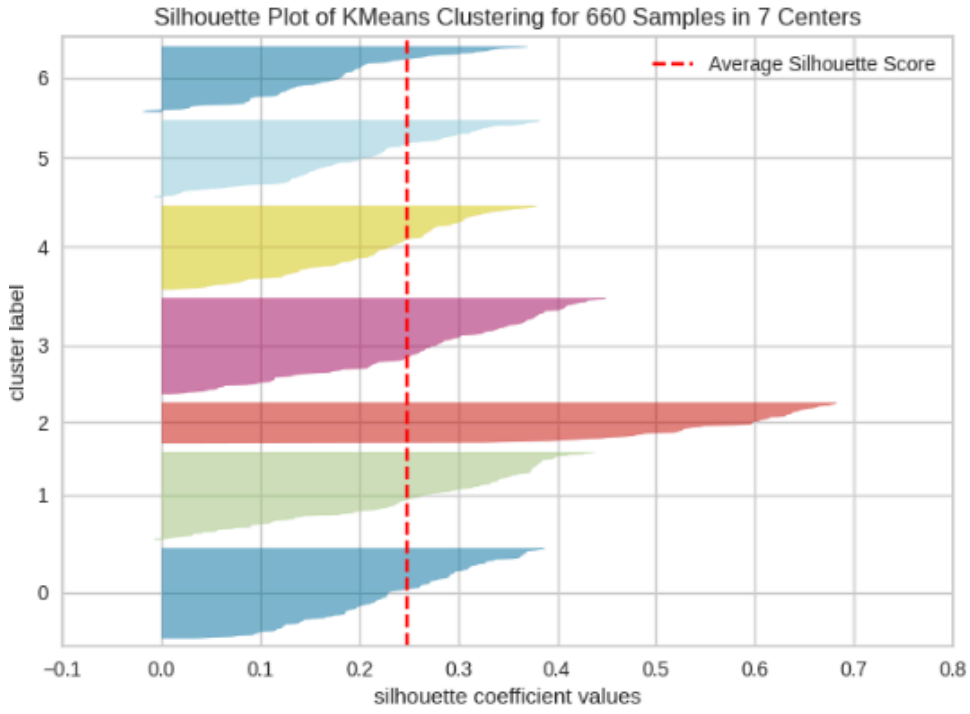


Fig.11: Silhouette Plot in 7 centers

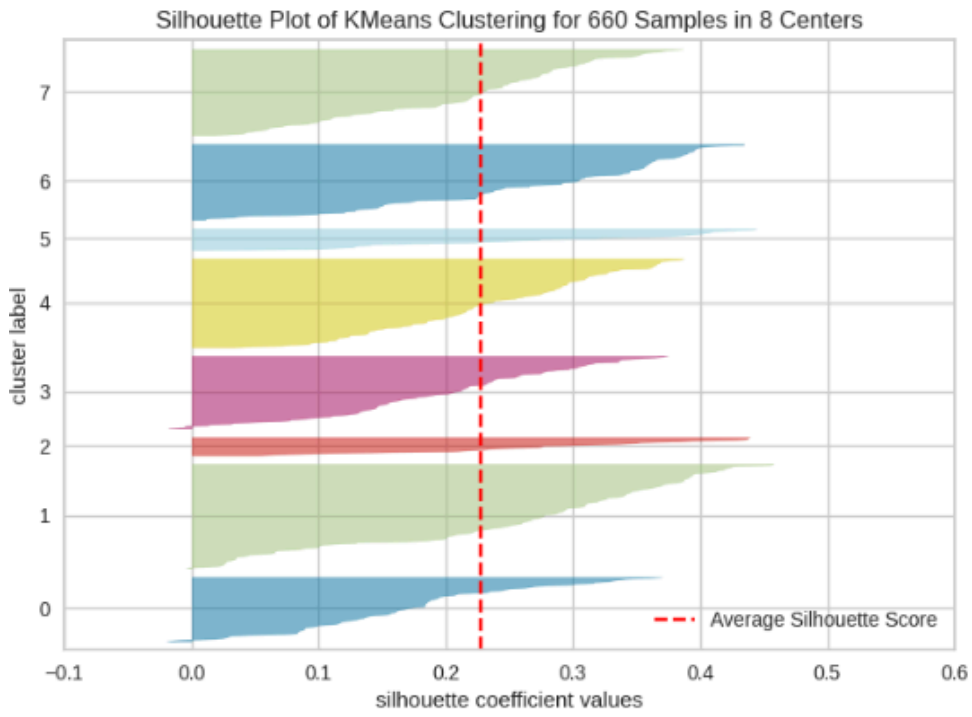


Fig.12: Silhouette Plot in 8 centers

- **Cluster Profiling**

K_means_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
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

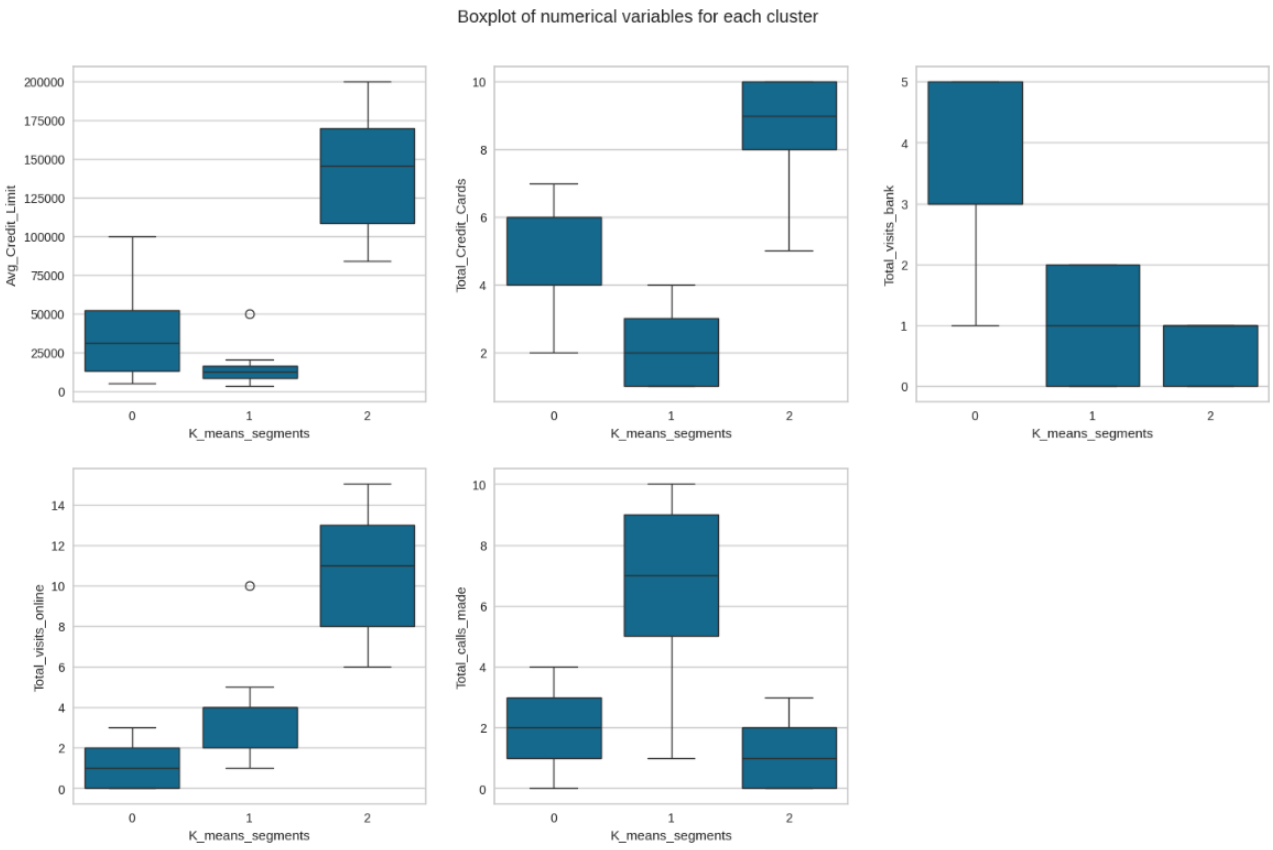


Fig.13: Boxplot of numerical variables for each cluster (KMeans)

- **Insights :**

    - Cluster 0
        - Customers with middle credit limits (~ 34K on average).
        - They also have the middle average number of credit cards(~ 6 cards each).
        - They tend to visit the bank more often rather than making calls and online transactions.

    - Cluster 1

        - Customers with minimum credit limits (~ 12K on average).
        - They also have the lowest average number of credit cards (~ 2 cards each).
        - They tend to make phone calls rather than online or bank visits.

    - Cluster 2

        - Customers with maximum credit limits (~ 140K on average).
        - They also have the maximum average number of credit cards(~ 9 cards each).
        - They tend to make online transactions rather than phone calls and bank visits.

Hierarchical Clustering

- **Hierarchical clustering with different linkage methods**

```
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 0.7058064784553605.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.6663534463875359.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.8326994115042136.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.7805990615142518.
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.
```

  - The highest cophenetic correlation is 0.8977080867389372, which is obtained with Euclidean distance and average linkage.

- **Different linkage methods with Euclidean distance**

```
Cophenetic correlation for single linkage is 0.7391220243806552.
Cophenetic correlation for complete linkage is 0.8599730607972423.
Cophenetic correlation for average linkage is 0.8977080867389372.
Cophenetic correlation for centroid linkage is 0.8939385846326323.
Cophenetic correlation for ward linkage is 0.7415156284827493.
Cophenetic correlation for weighted linkage is 0.8861746814895477.
```

  - The highest cophenetic correlation is 0.8977080867389372, which is obtained with average linkage.

- **Silhouette Score**

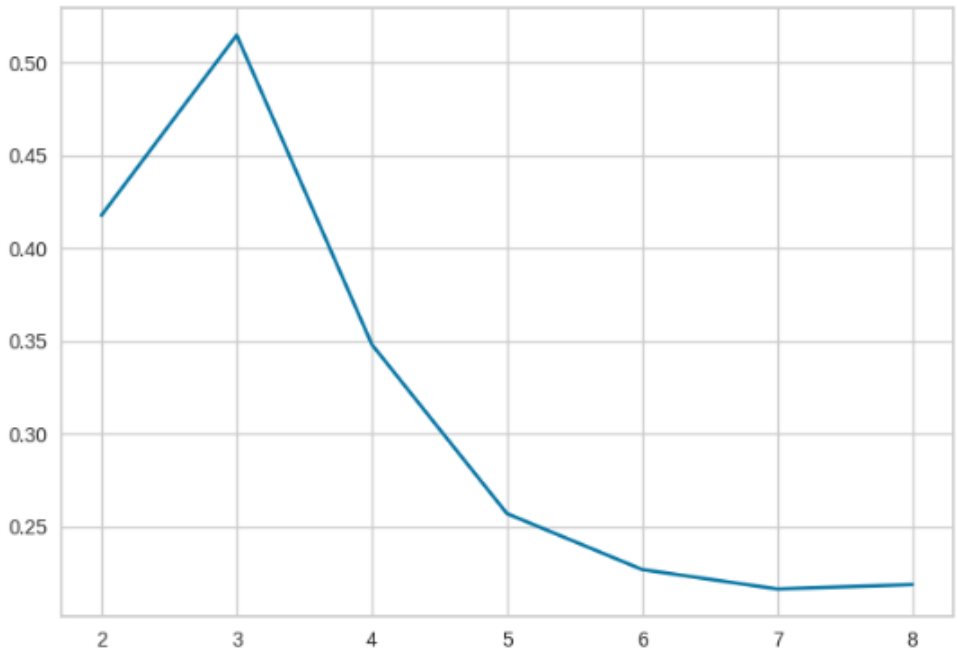


Fig.14: Silhouette Score (HC Cluster)

- From the silhouette scores, it seems that 3 is a good value of k.

- **Dendrograms for each linkage method**

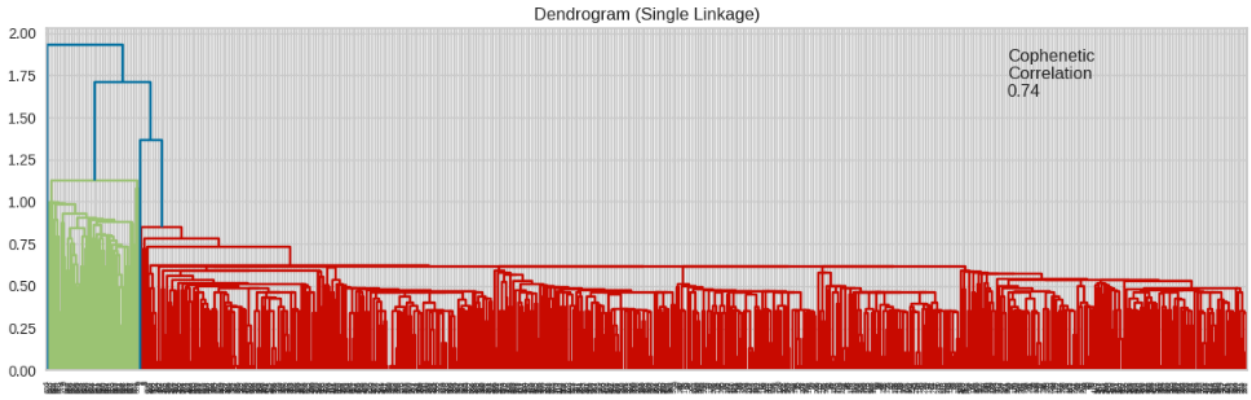


Fig.15: Dendrogram for Single Linkage

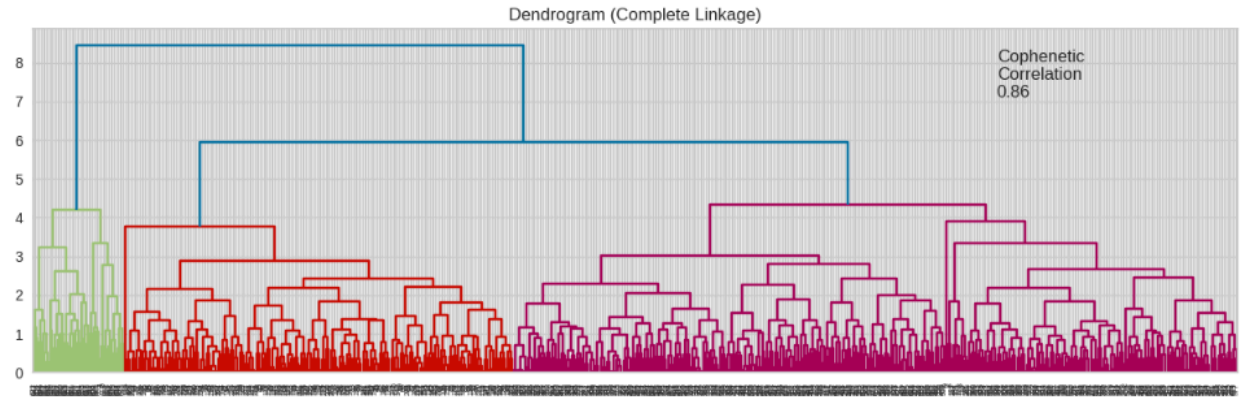


Fig.16: Dendrogram for Complete Linkage

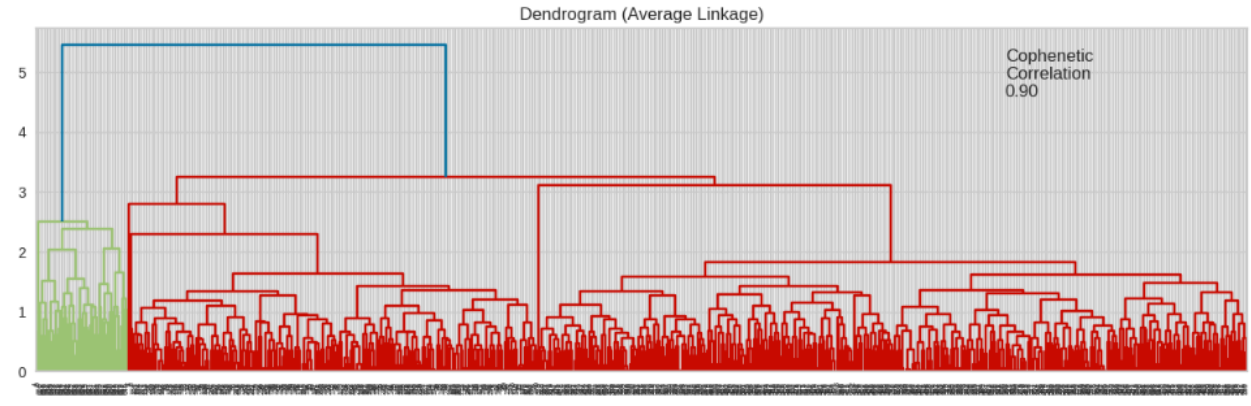


Fig.17: Dendrogram for Average Linkage

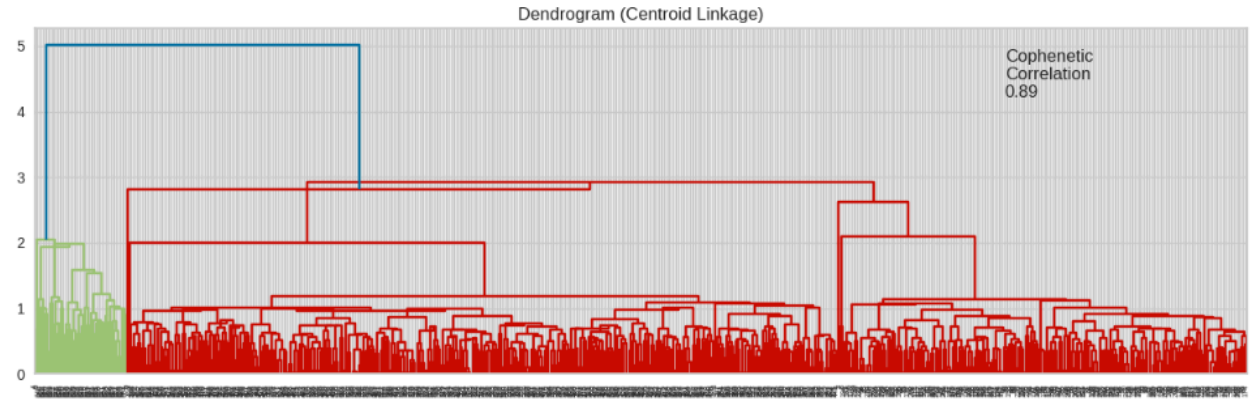


Fig.18: Dendrogram for Centroid Linkage

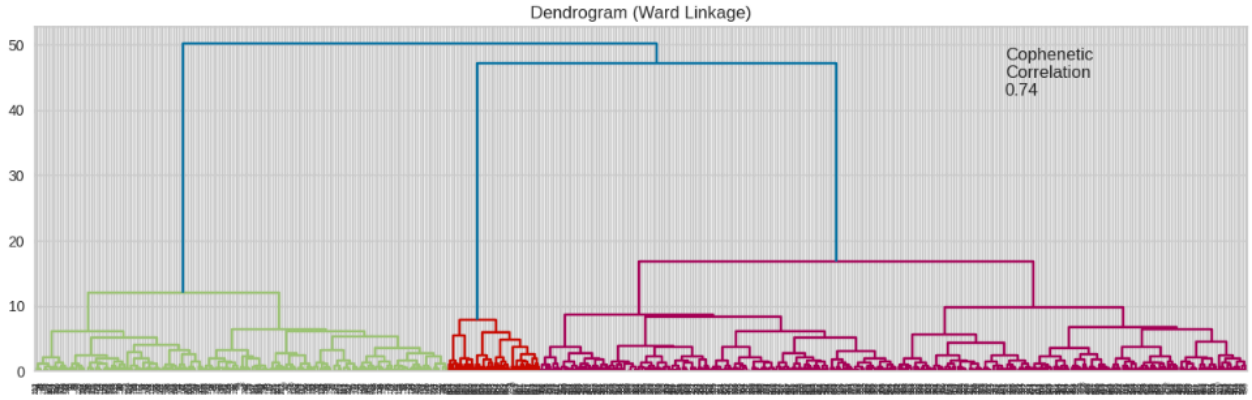


Fig.19: Dendrogram for Ward Linkage

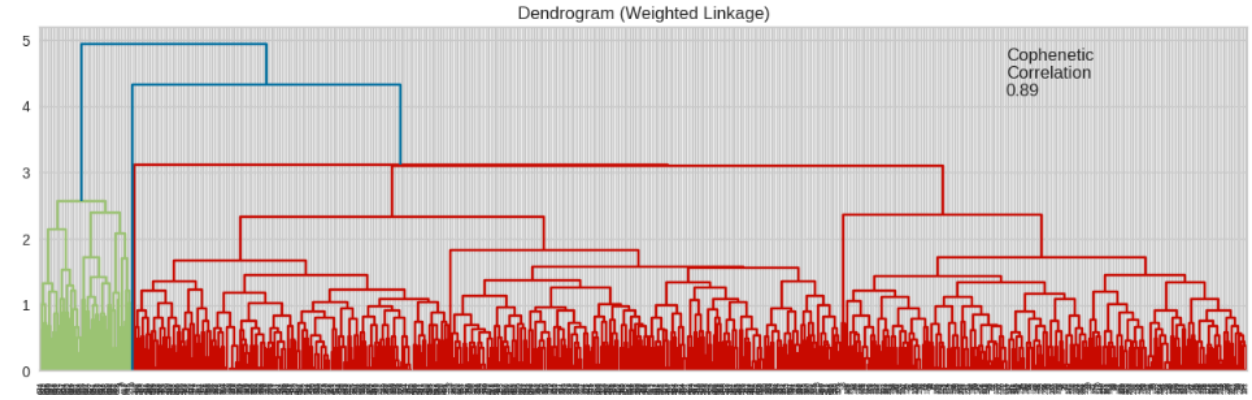


Fig.20: Dendrogram for Weighted Linkage

Observation:

- The cophenetic correlation is highest for average linkage methods.
- We will move ahead with average linkage.
- 3 is the appropriate number of clusters from the dendrogram for average linkage.

• Cluster Profiling

HC_Clusters	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segments
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

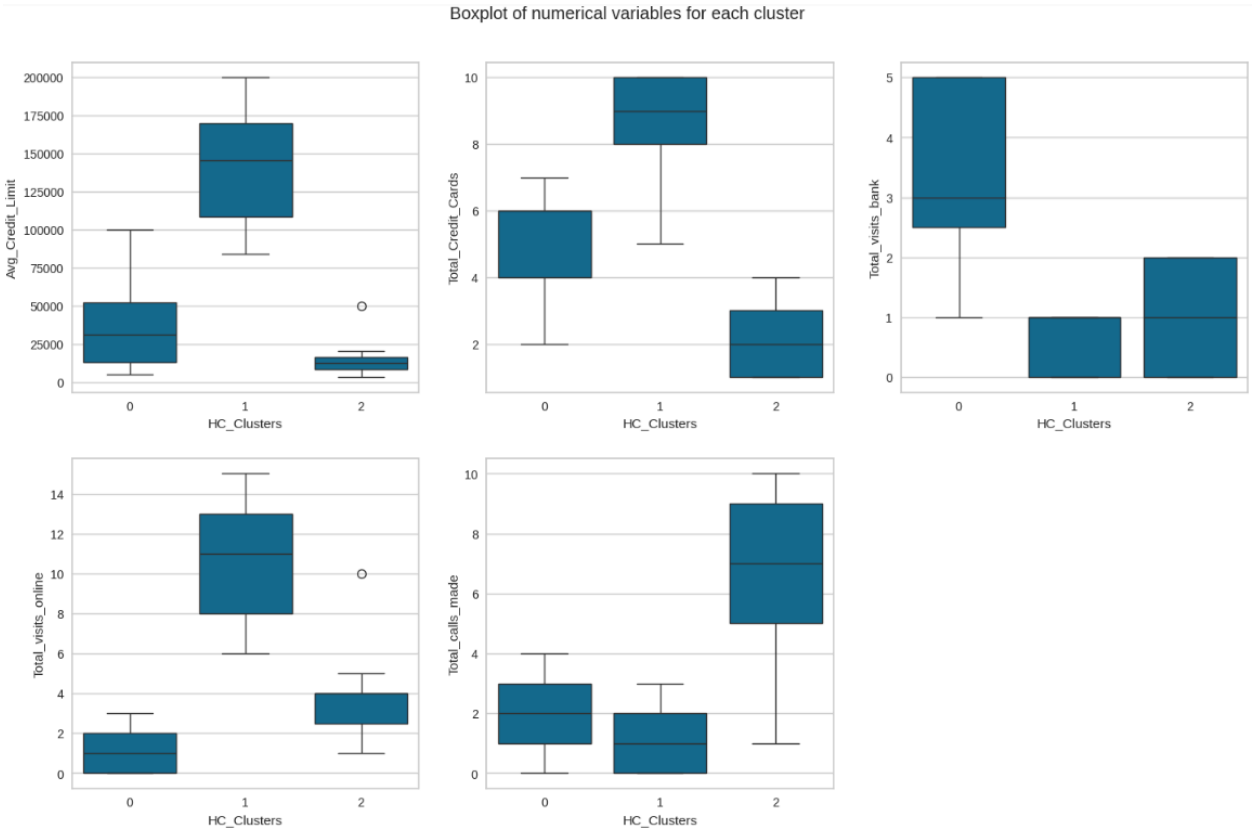


Fig.21: Boxplot of numerical variables for each cluster (HC)

- **Insights :**

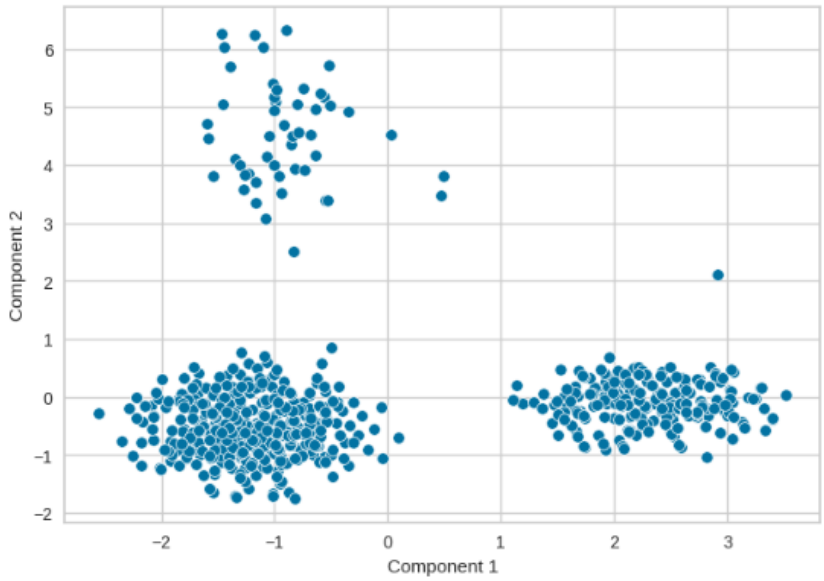
- Cluster 0 (Majority Cluster — 367 members):
    - Lowest average credit limit (~33K)
    - Moderate total credit cards and visits to the bank
    - Low online visits and total calls made

- Cluster 1 (Smallest Cluster — 50 members):
    - Highest average credit limit (~110K)
    - The highest visits to both the bank and online platforms
    - Lowest number of credit cards
    - Fewest calls made

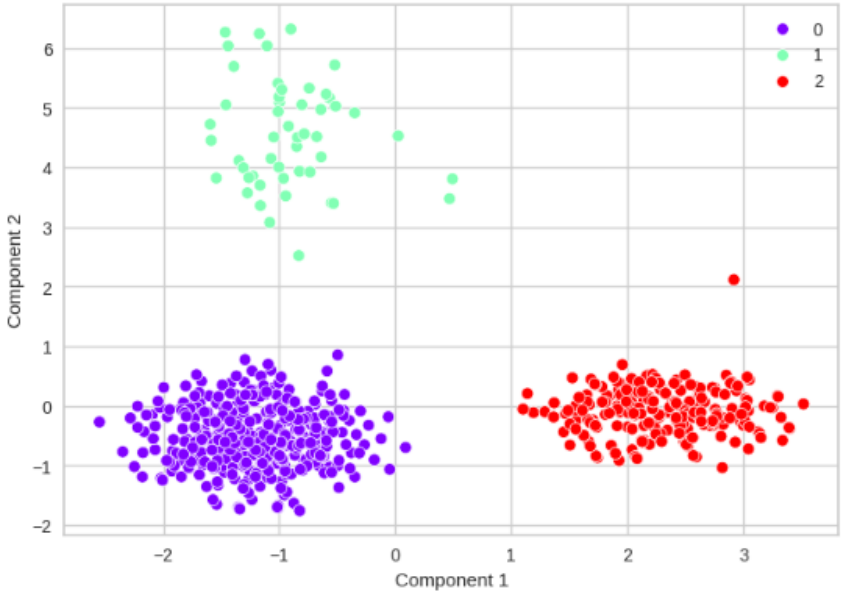
- Cluster 2 (Mid-size Cluster — 223 members):
    - Moderate credit limit (~122K)
    - The highest total of credit cards
    - Very high total calls made
    - Low visits to both online and bank

PCA for visualization

- The first two principal components explain 84% of the variance in the data.



- We can kind of see 3 broad clusters.



- All three clusters are the major clusters.

K-means vs Hierarchical Clustering

❖ **Which clustering technique took less time for execution?**

K-means took less time than HC took more time because of dendrograms.

❖ **Which clustering technique gave you more distinct clusters, or are they the same?**

They both gave the same.

❖ **How do the silhouette scores vary?**

For K-Means-

```
For n_clusters = 2, silhouette score is 0.5703183487340514
For n_clusters = 3, silhouette score is 0.5157182558881063
For n_clusters = 4, silhouette score is 0.3556670619372605
For n_clusters = 5, silhouette score is 0.332312434784495
For n_clusters = 6, silhouette score is 0.2572826107002575
For n_clusters = 7, silhouette score is 0.2397404411036028
For n_clusters = 8, silhouette score is 0.2290267171721957
```

For Hierarchical Clustering -

```
For n_clusters = 2, silhouette score is 0.417704147620949
For n_clusters = 3, silhouette score is 0.5147639589977819
For n_clusters = 4, silhouette score is 0.3480822581261928
For n_clusters = 5, silhouette score is 0.2569177732675831
For n_clusters = 6, silhouette score is 0.22677849725544041
For n_clusters = 7, silhouette score is 0.2162968685485734
For n_clusters = 8, silhouette score is 0.2186949061936046
```

- K-Means performs best at 2 clusters, while Hierarchical Clustering is optimal at 3 clusters.
- Both methods show declining silhouette scores with increasing clusters, implying that over-clustering reduces cluster quality.

- K-Means gives better scores overall than hierarchical clustering for most cluster counts in this case.

❖ **How many observations are there in the similar clusters of both Algorithms?**

- K-Means-
  - Cluster 0 - 386
  - Cluster 1- 224
  - Cluster 2- 50

- Hierarchical Clustering-
  - Cluster 0 - 387
  - Cluster 1- 50
  - Cluster 2- 223

❖ **How many clusters are obtained as the appropriate number of clusters from both algorithms?**

3 was obtained as appropriate number of clusters from both algorithms.

Comparison:

- The clusters profiles from both the clustering techniques are very similar to each other.
- Users with a high average credit limit and a high number of credit cards were clustered together with both methods. They also appear to make the most online visits.
- Users who made the most calls to the bank tended to be those with the lowest credit limits, and were clustered together.

Actionable Insights & Recommendations

- Since both algorithms gave very similar results, it is recommended to use the K-Means method with any future data due to its faster compute time, and the cluster numbers from the K-Means method will be referred to here.

- Cluster 0:

    - This appears to be the average customer, with mid-range scores on all accounts.
    - Track visit frequency to identify high-touch customers and prioritize premium service offers.
    - Provide incentives for online banking to reduce the need for in-branch visits.
    - Provide rewards or cashback for increasing digital usage or referrals.
    - Target this group with SMS/WhatsApp campaigns introducing app features and online tools.

- Cluster 1:

    - These are customers with a lower credit limit and the least number of credit cards. They make the most calls to the bank.
    - Target with offers for financially immature customers and those with lower credit scores.
    - Provide educational resources to improve the customer's financial knowledge.

    - Introduce basic card upgrades, savings boosters, or EMI options over outbound calls.
    - Offer agent-based digital onboarding through calls to ease transition into mobile banking.

- Cluster 2:

    - These appear to be the high net-worth customers with higher average credit limits and the most credit cards.
    - Target with exclusive offers, premium credit cards, and investment offers.
    - Offer concierge services or a premium support line, focused on online channels.
    - Target with mobile app notifications and personalized offers based on spending trends.
    - Provide self-service dashboards with transaction analytics, credit tracking, and spending insights.