UNSUPERVISED LEARNING BUSINESS REPORT

Krishnan CS

GREAT LEARNING DSBA

Contents

1. Problem Statement

1.1 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.

1.2 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.

1.3 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).

2 Exploratory Data Analysis

2.1 Problem Definition

AllLife Bank wants to focus on its credit card customer base in the next financial year and improve market penetration. The different segments in the existing customers would be identified, based on their spending patterns as well as past interaction with the bank, using clustering algorithms. Recommendations will be provided to the bank, on how to better market to and service these customers.

2.2 Univariate Analysis

The data contains 660 records and 7 columns. All are integer columns.

```
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
```

Figure 1 Information on the columns in the dataset

A description of all the columns is shown in the figure.

	count	mean	std	min	25%	50%	75%	max
Sl_No	660.0	330.50	190.67	1.0	165.75	330.5	495.25	660.0
Customer Key	660.0	55141.44	25627.77	11265.0	33825.25	53874.5	77202.50	99843.0
Avg_Credit_Limit	660.0	34574.24	37625.49	3000.0	10000.00	18000.0	48000.00	200000.0
Total_Credit_Cards	660.0	4.71	2.17	1.0	3.00	5.0	6.00	10.0
Total_visits_bank	660.0	2.40	1.63	0.0	1.00	2.0	4.00	5.0
Total_visits_online	660.0	2.61	2.94	0.0	1.00	2.0	4.00	15.0
Total_calls_made	660.0	3.58	2.87	0.0	1.00	3.0	5.00	10.0

Figure 2 Description of numerical columns

Sl_No: This is just the primary key of the records. It can be dropped as it is not required for the clustering analysis.

Customer Key: This helps to identify the customers. There are two records for 5 of the customers.

Number of unique customers: 655

	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
48	49	37252	6000	4	0	2	8
432	433	37252	59000	6	2	1	2
332	333	47437	17000	7	3	1	0
4	5	47437	100000	6	0	12	3
411	412	50706	44000	4	5	0	2
541	542	50706	60000	7	5	2	2
391	392	96929	13000	4	5	0	0
398	399	96929	67000	6	2	2	2
104	105	97935	17000	2	1	2	10
632	633	97935	187000	7	1	7	0

Figure 3 Customer Keys which are repeated

However, this column can be dropped as it is not required for the clustering analysis.

Average Credit Limit: This indicates the average credit limit of each customer for all credit cards.

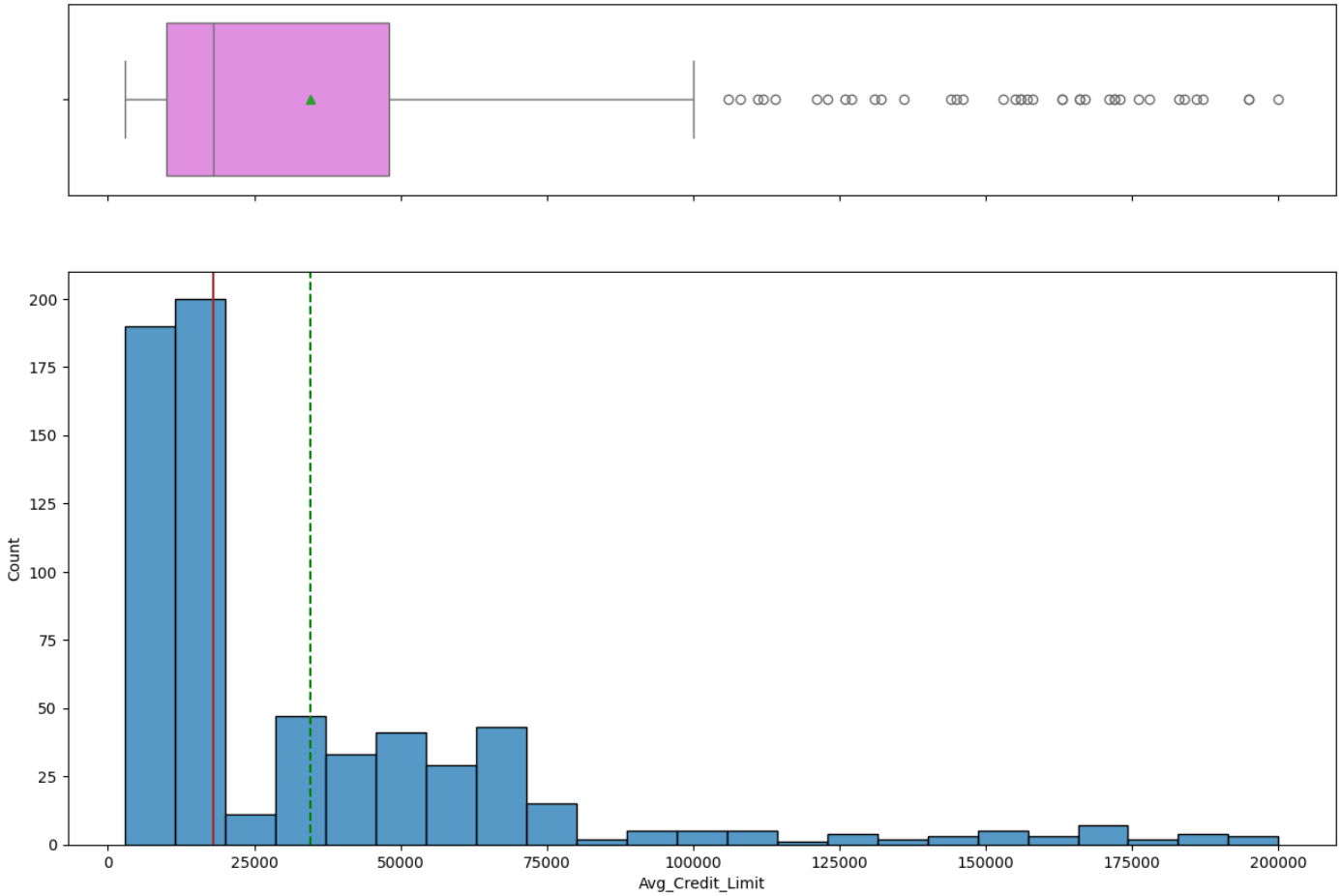

Figure 4 Box plot and histogram for average credit limit

- Based on the plot, the distribution is right skewed.
- There are a lot of outliers above the mean.
- The values are all reasonable and could be retained for analysis.

Total credit cards: It indicates the total number of credit cards possessed by the customer.

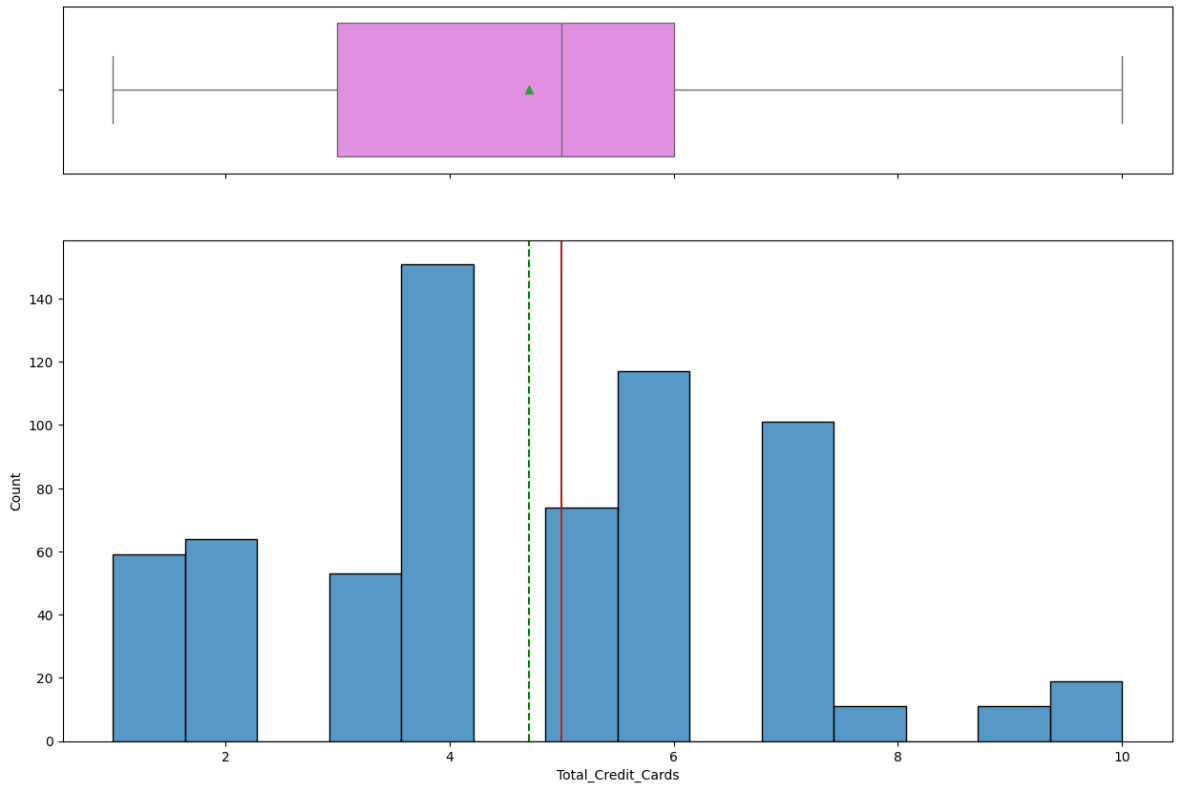

Figure 5 Box plot and histogram for total credit cards

- The number of credit cards vary from 1 to 10.
- There are only a few customers who have more than 7 credit cards.

Total visits bank: This column indicates the total number of visits that the customer made to the bank during the year.

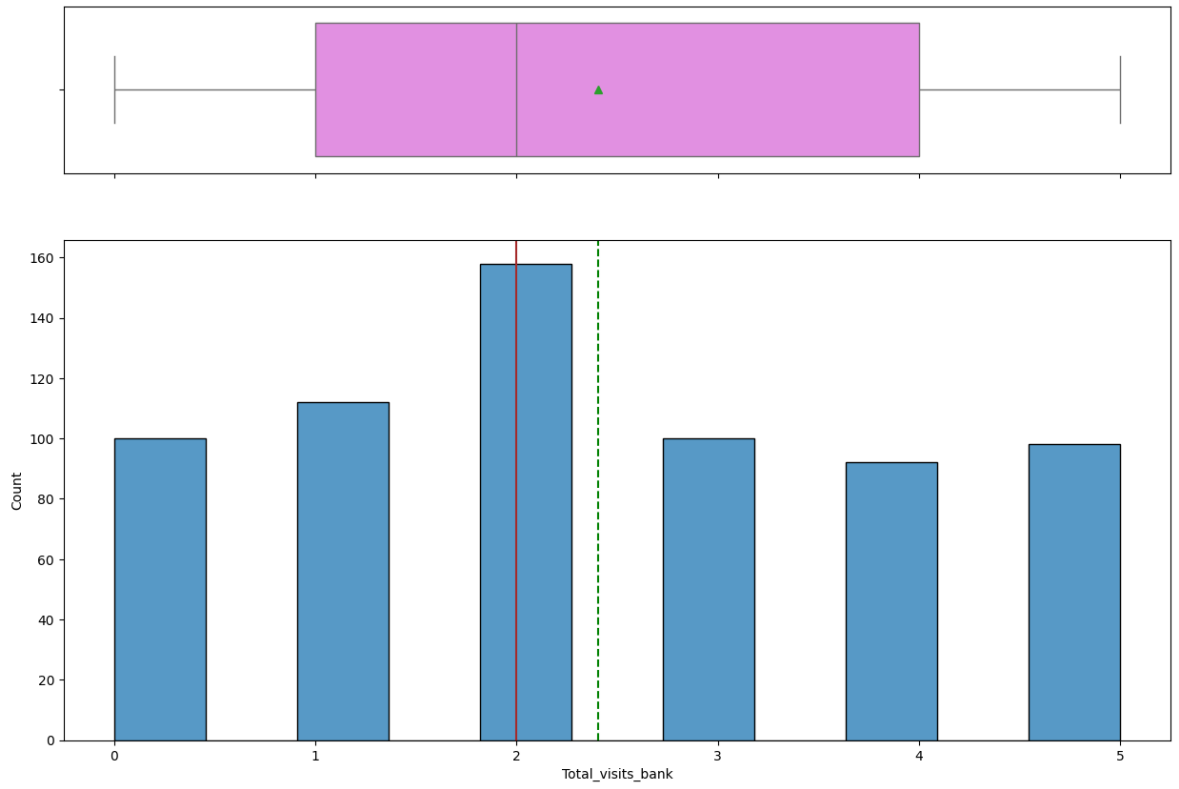

Figure 6 Box plot and histogram for total visits bank

- This is roughly a uniform distribution ranging from 0 to 5.

Total visits online: This column indicates the total number of visits or online logins made by the customer on a yearly basis.

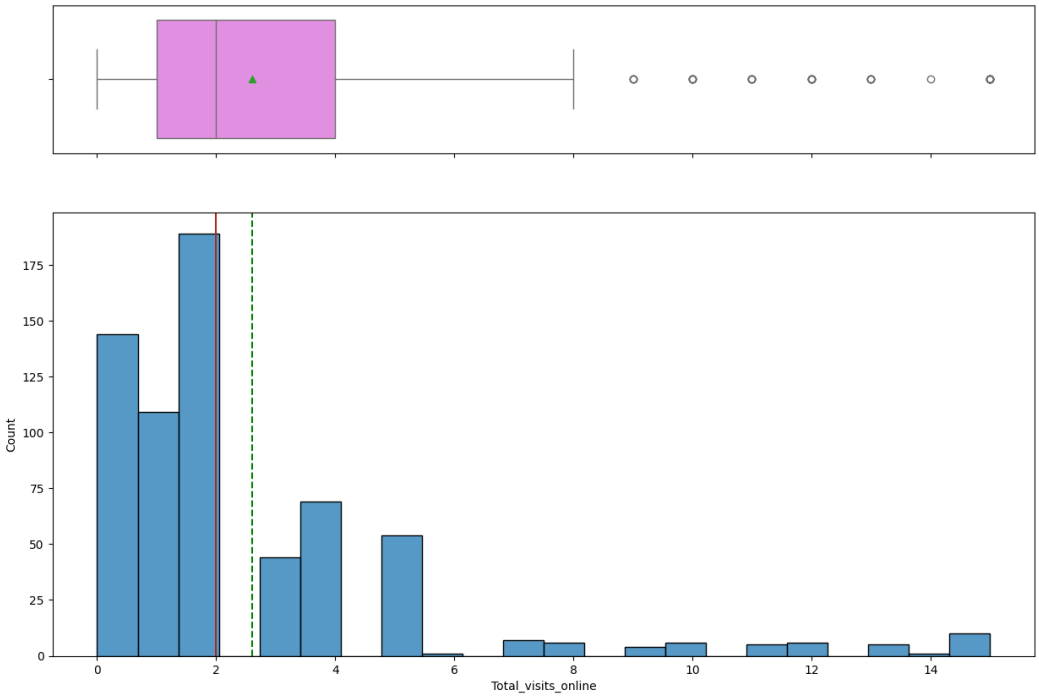

Figure 7 Box plot and histogram for total visits online

- Based on the plot, the distribution is right skewed.
- There are a lot of outliers above the mean.
- The values are all reasonable and could be retained for analysis.

Total calls made: This column indicates the total number of calls made by the customer to the bank or its customer service department during the year.

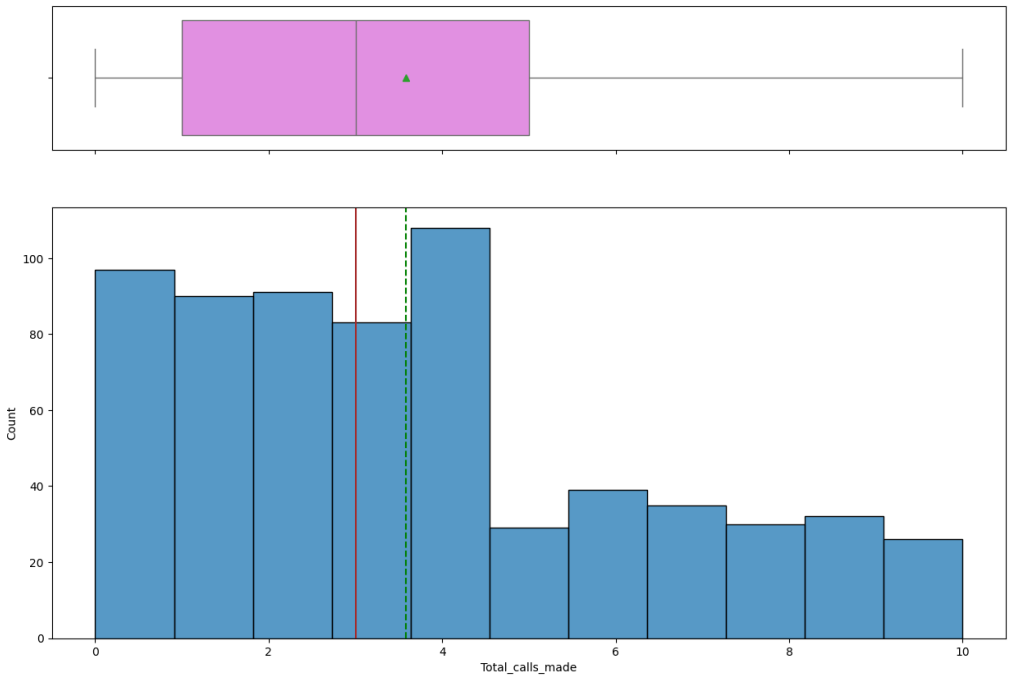

Figure 8 Box plot and histogram for total calls made

- Based on the plot, the distribution is right skewed.
- There are no outliers in the data.

2.3 Bivariate Analysis

A quick way to start the bivariate analysis is a heatmap, using which all the numerical columns could be studied.

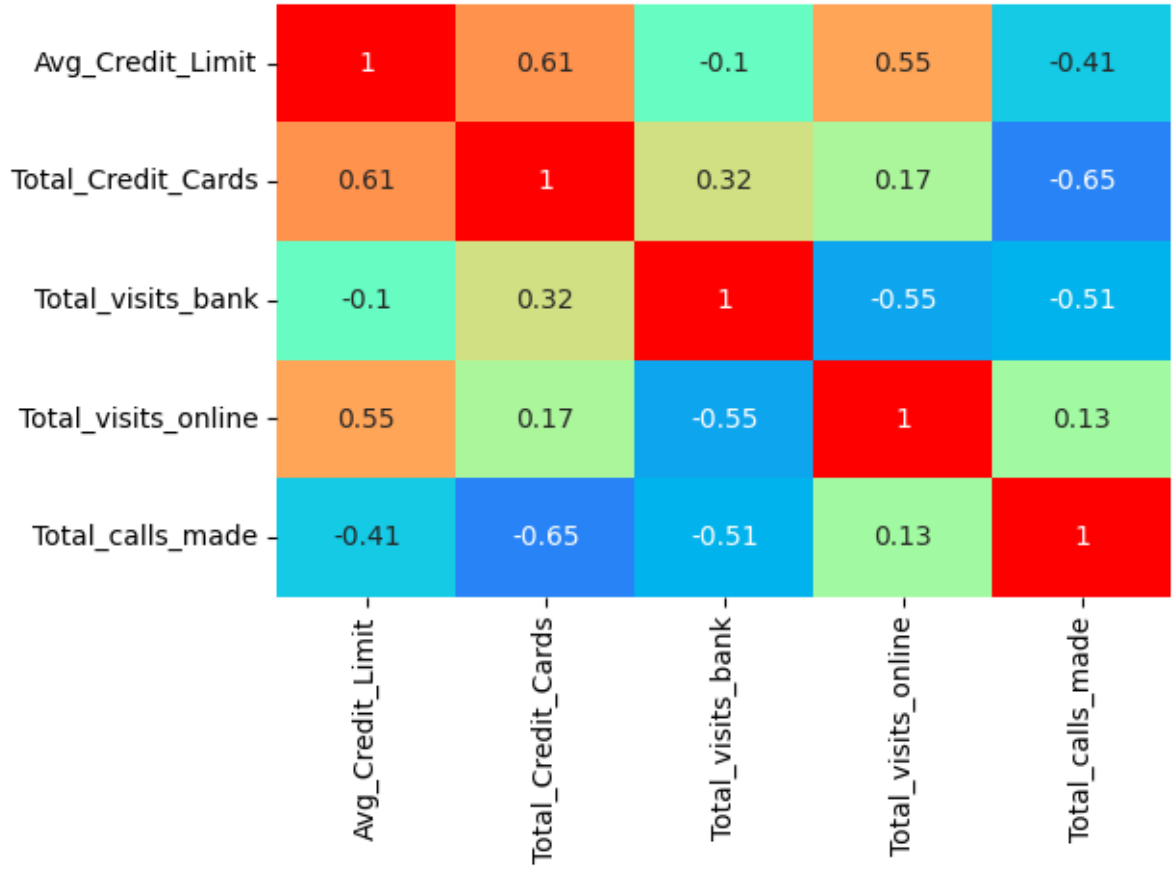

Figure 9 Heatmap for numerical columns

- Average credit limit has a moderate positive correlation with total credit cards and total visits online and a moderate negative correlation with total calls made.
- Total credit cards has a moderate negative correlation with total calls made and a moderate positive correlation with total visits to the bank.
- Total visits to the bank is negatively correlated with total visits online and total calls made.

A pair-wise plot for the numerical columns is shown below.

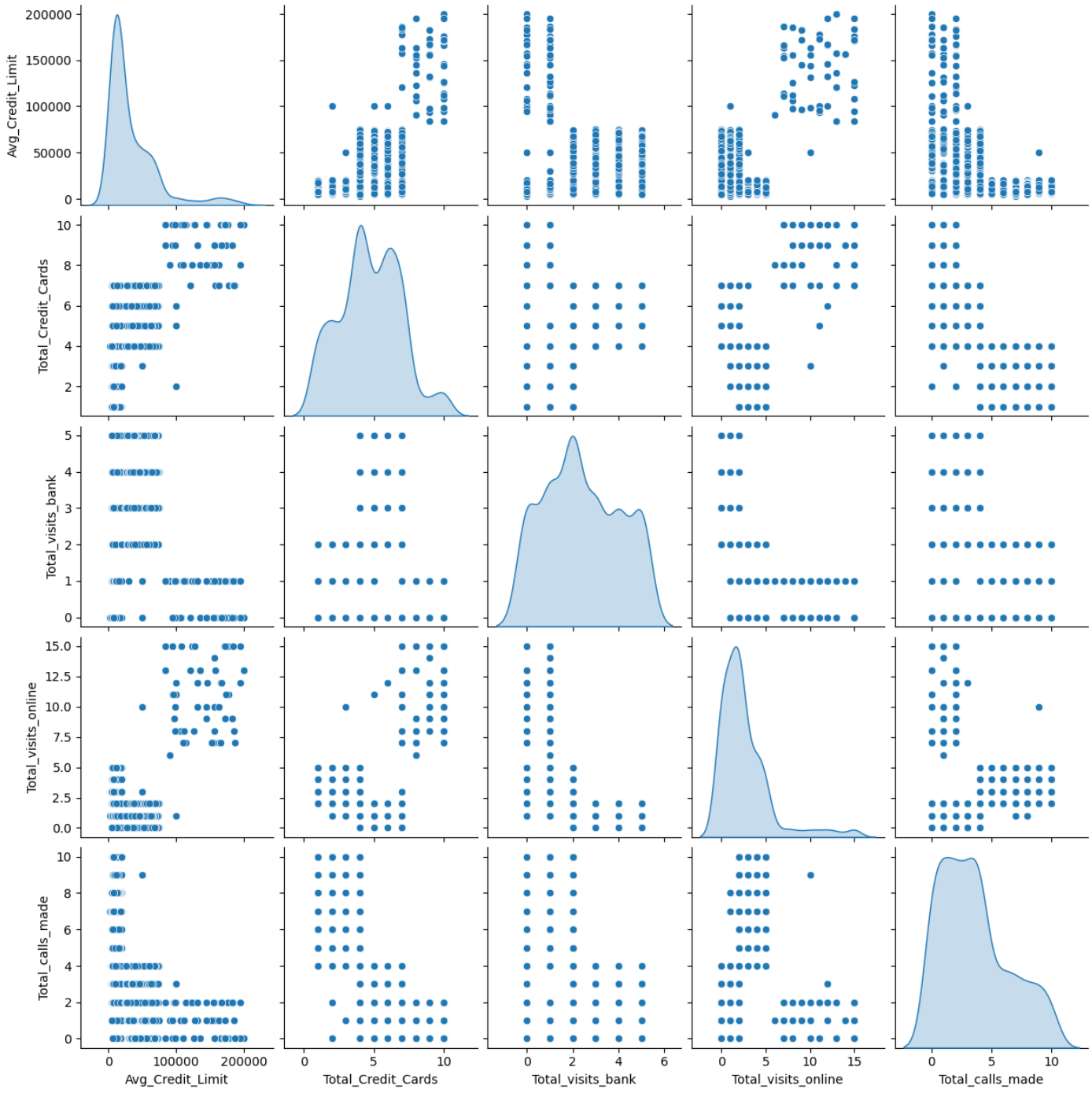

Figure 10 Pair plot for numerical columns

2.4 Observations on individual variables and relationship between variables

- People with better average credit limit are the ones with a good financial history. They maintain more credit cards and tend to use the online method for transactions.
- Customers with more credit cards make fewer calls. They either visit the bank or get their problems resolved online.
- The customers who visit the bank do not make a lot of calls or do online transactions. This is not surprising and these are the customers who prefer visiting the bank directly for their transactions.
- The number of local maxima and local minima in the plots along the diagonal of the pair-wise plot give an indicator of the number of clusters. For example, the total credit cards plot suggests 4 clusters.

3 Data Preprocessing

The data contains 660 records and 7 columns. The columns have been described in the data description section.

3.1 Missing value treatment

There are no missing values in the data.

```
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
```

Figure 11 Information on columns indicates no missing values

3.2 Outlier Detection and Treatment

- The box plots show a few outliers in Total_visits_online and Avg_Credit_Limit.
- The values are reasonable and they need not been treated.

3.3 Feature Engineering

The numerical columns can be analyzed directly. There is no need for feature engineering.

3.4 Data Scaling

- Data will be scaled using Z-scores.
- The values for Average Credit Limit are much higher than those in other columns. Scaling ensures that all columns are given equal importance.
- Z-score scaling ensures that the outlier information is also retained.

4 K-Means Clustering

4.1 Applying K-Means Clustering

After dropping Serial Number and Customer Key, the data contains 5 columns - Average Credit Limit, Total credit cards, Total visits bank, Total visits online and Total calls made. The K-Means clustering technique is applied on the data containing these columns.

4.2 Elbow Curve

In the elbow method, the number of clusters (k) is varied from 1 to 9. For each value in that range, the average distortion is calculated. The plot between the average distortion and the number of clusters is the elbow curve.

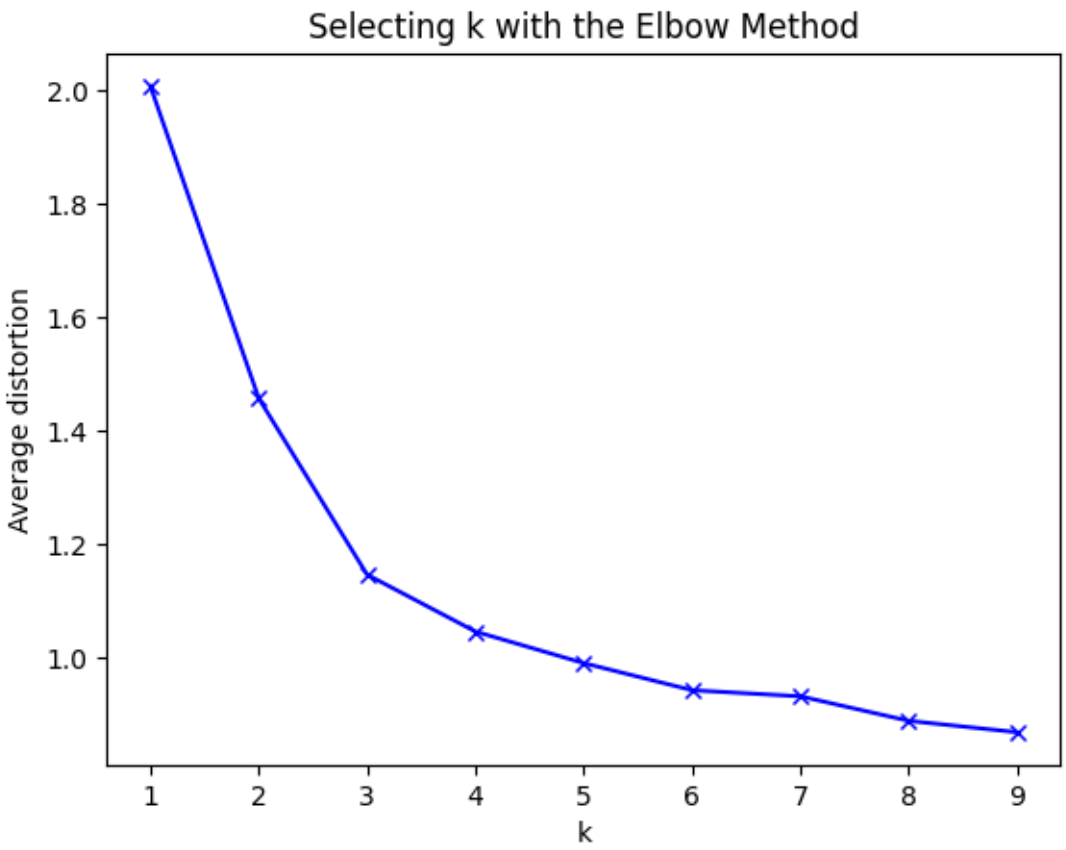

Figure 12 Elbow curve for K-Means clustering

The optimal number of clusters is the point at which the drop in average distortion slows down. Based on the above graph, the best value for the number of clusters is 3. The optimal number is fixed based the silhouette score.

4.3 Silhouette Scores

The silhouette score is used as a metric for testing the quality of clusters. The silhouette scores for the different number of clusters is shown below:

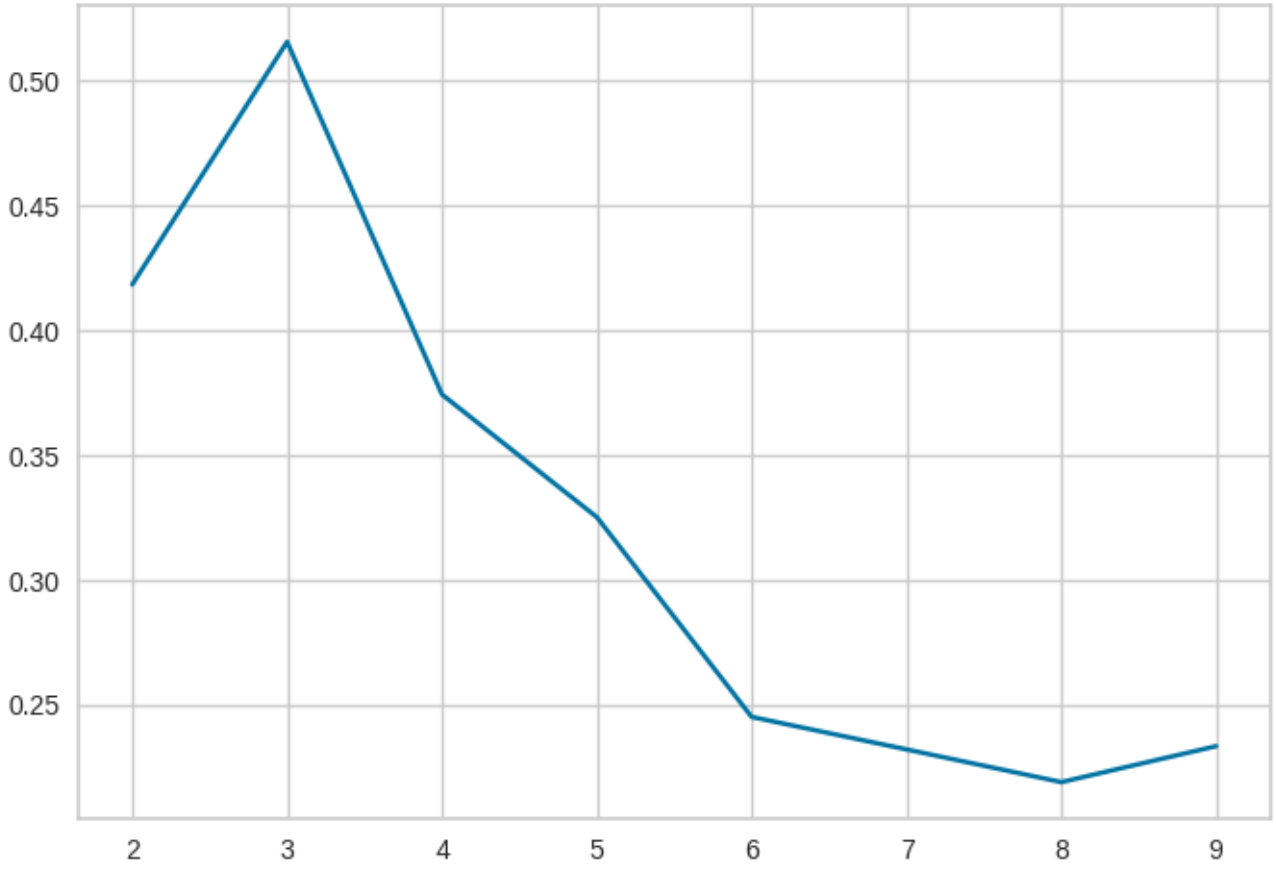

Figure 13 Silhouette score vs Number of clusters

The highest score is observed for k = 3 clusters.

The silhouette plot can also be drawn for each value of k. The silhouette plots for k = 3, 4, 5, 6 is shown below:

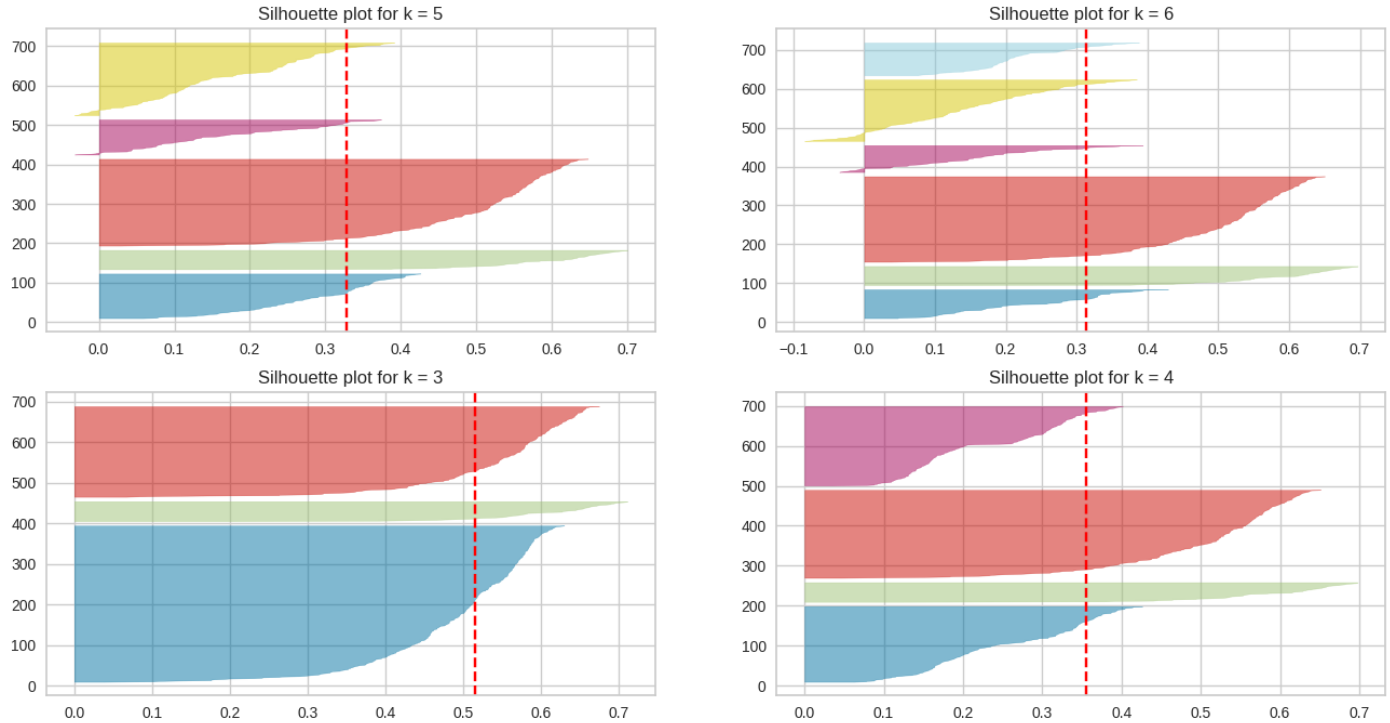

Figure 14 Silhouette visualizer for 3, 4, 5, 6 clusters

4.4 Appropriate number of clusters

Based on the silhouette scores, the appropriate number of clusters is 3.

4.5 Cluster profiling

The cluster profiling is performed by calculating the average values for each numerical column. This highlights the significance of each cluster. Box plots and bar charts are also drawn for additional insights.

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

Figure 15 Table highlighting maximum means of data columns for each K-Means cluster

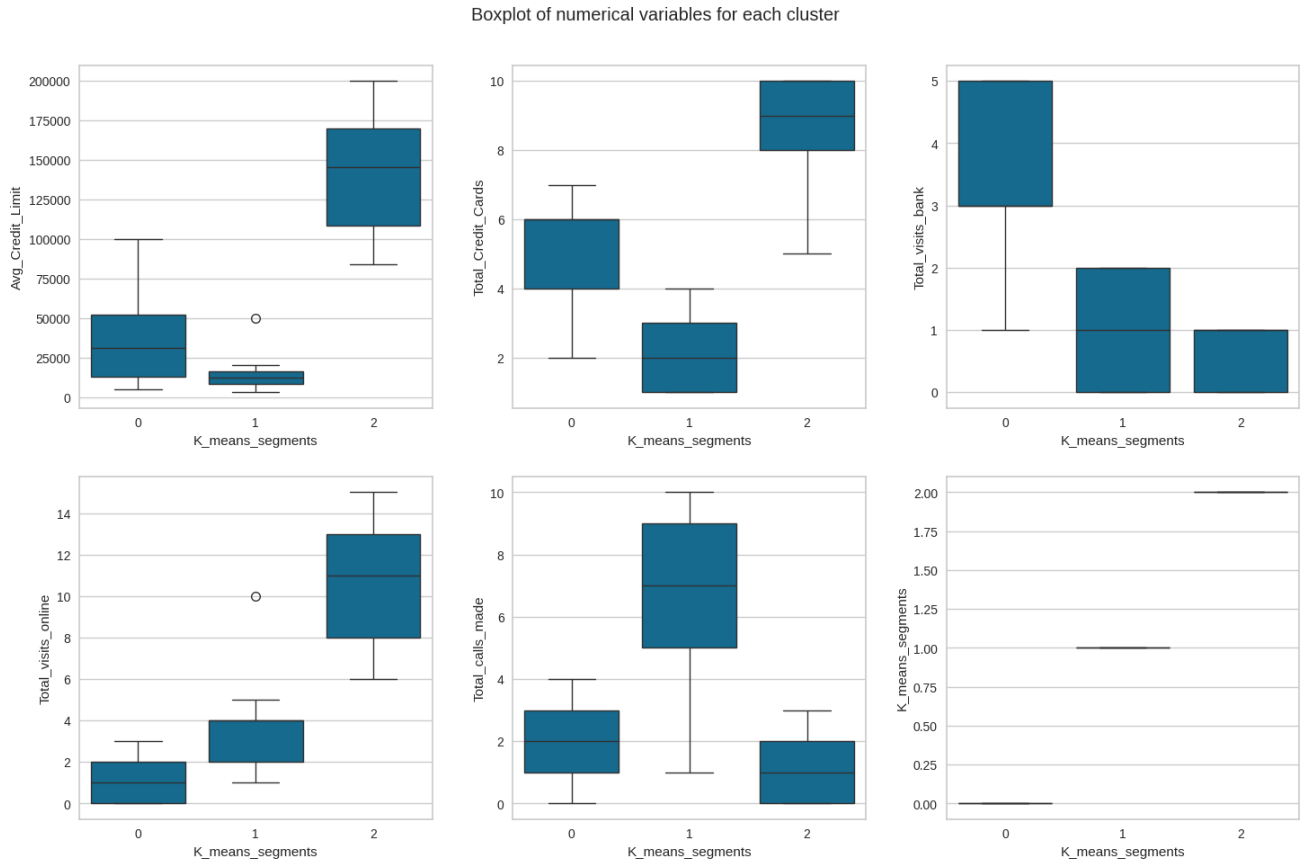

Figure 16 Box plot for numerical columns grouped by K-Means clusters

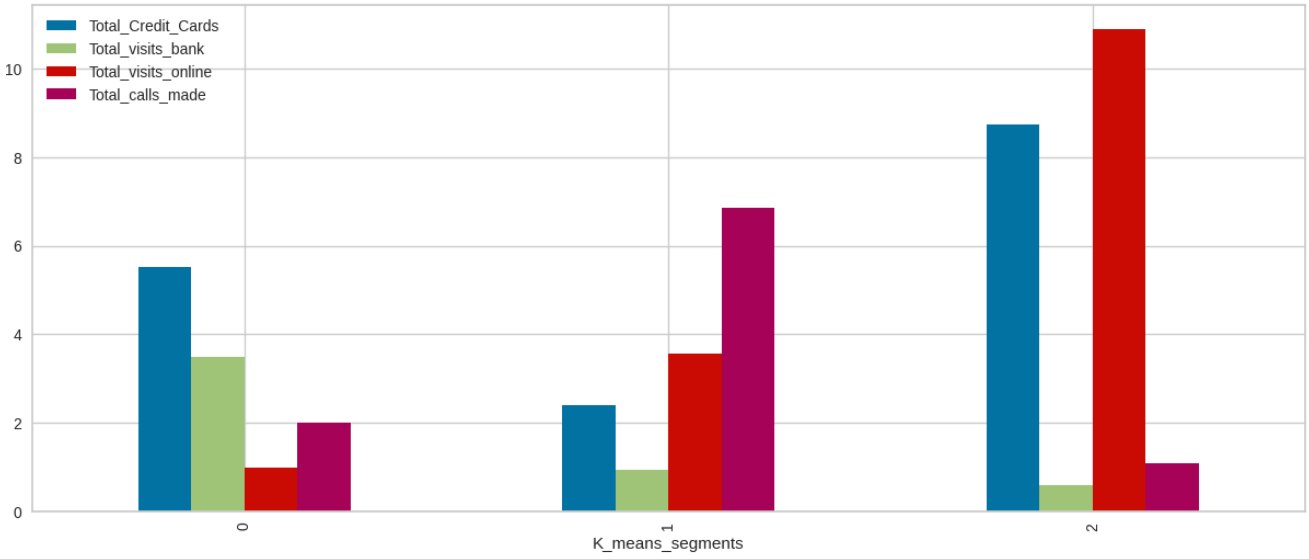

Figure 17 Bar chart for numerical columns for each K-Means cluster

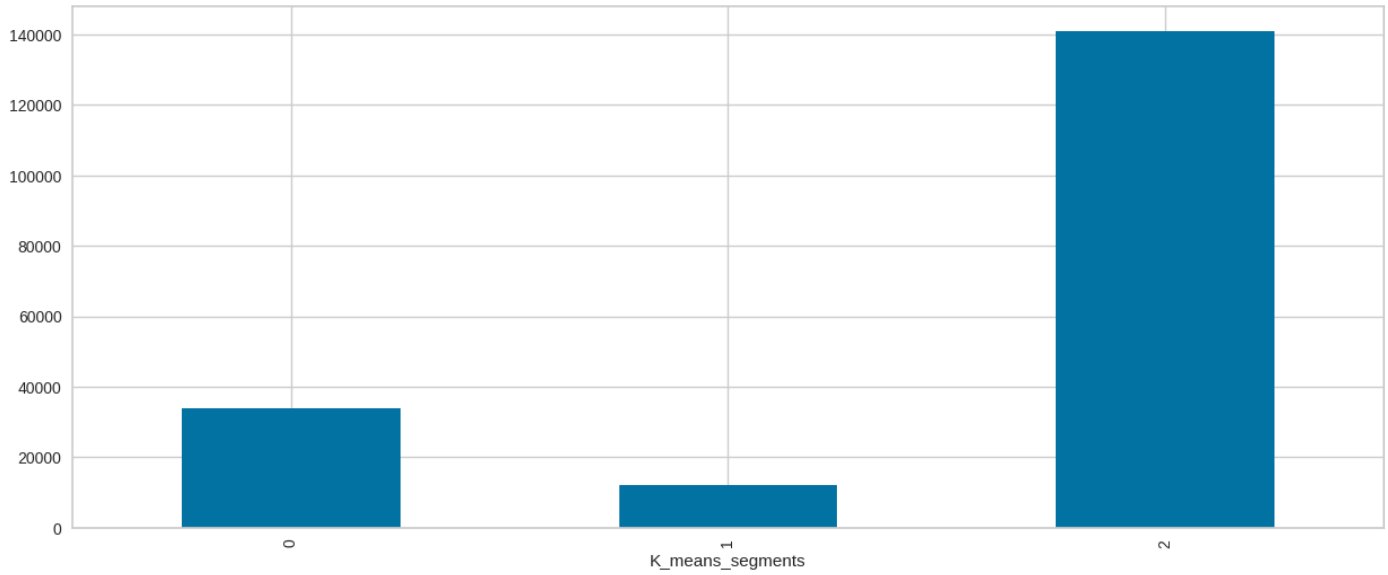

Figure 18 Bar chart of average credit limit for each K-Means cluster

Cluster 0:

- Maximum number of customers belong to this cluster.
- These customers have a high number of visits to the bank.
- The total online visits are low for these customers.
- Average credit limit, the number of calls made and the number of credit cards are moderate.
- This cluster will be renamed as High_Bank.

Cluster 1:

- About one-third of the customers belong to this cluster.
- These customers have a high number of calls made.
- The average credit limit, total credit cards and visits to the bank are low for these customers.
- The number of visits online is moderate.
- This cluster will be renamed as High_Calls.

Cluster 2:

- Only a few customers belong to this cluster.
- These customers have a high average credit limit, total credit cards and number of visits online.
- The total bank visits and total calls made are low for these customers.
- This cluster will be renamed as High_CL.

5 Hierarchical Clustering

Hierarchical clustering is an alternate approach to clustering the records. In this method, the pairwise distances between the rows are analysed and clustering is performed based on the distances. Different distances like Euclidean, Chebyshev, Mahalanobis, and Cityblock would be used.

5.1 Hierarchical clustering with different linkage methods

There are 6 linkage methods available - single, complete, average, centroid, ward, weighted. The cophenetic correlations and dendrograms would be checked for the different combinations of distances and linkages to finalize the model.

5.2 Dendrograms for each linkage method

The dendrograms for the different linkage methods are shown below:

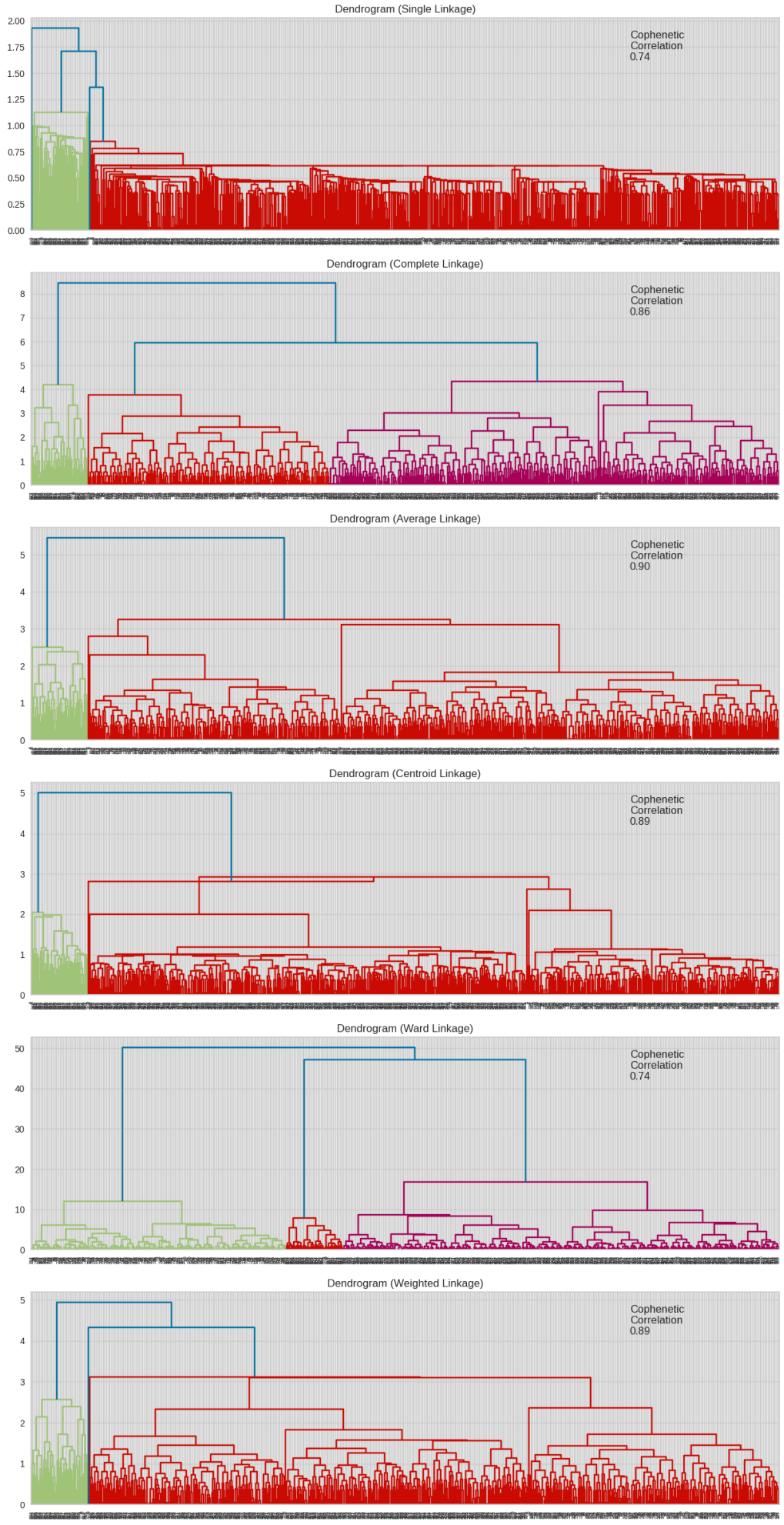

Figure 19 Dendrogram for different linkage methods

5.3 Cophenetic correlation for each linkage method

The cophenetic correlation for the different linkage methods is shown below:

```
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.
```

Figure 20 Cophenetic correlation scores for each linkage method

The best score of 0.898 is obtained for a combination of Euclidean distance and average linkage.

5.4 Appropriate number of clusters

- The best cophenetic correlation score was obtained with average linkage.
- In the dendrogram for average linkage, a big difference in value is seen at the split of 3 clusters.
- The number of clusters can be fixed at 3.

5.5 Cluster Profiling

The cluster profiling is performed by calculating the average values for each numerical column. This highlights the significance of each cluster. Box plots and bar charts are also drawn for additional insights.

HC_Clusters	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
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

Figure 21 Table highlighting maximum means of data columns for each Hierarchical cluster

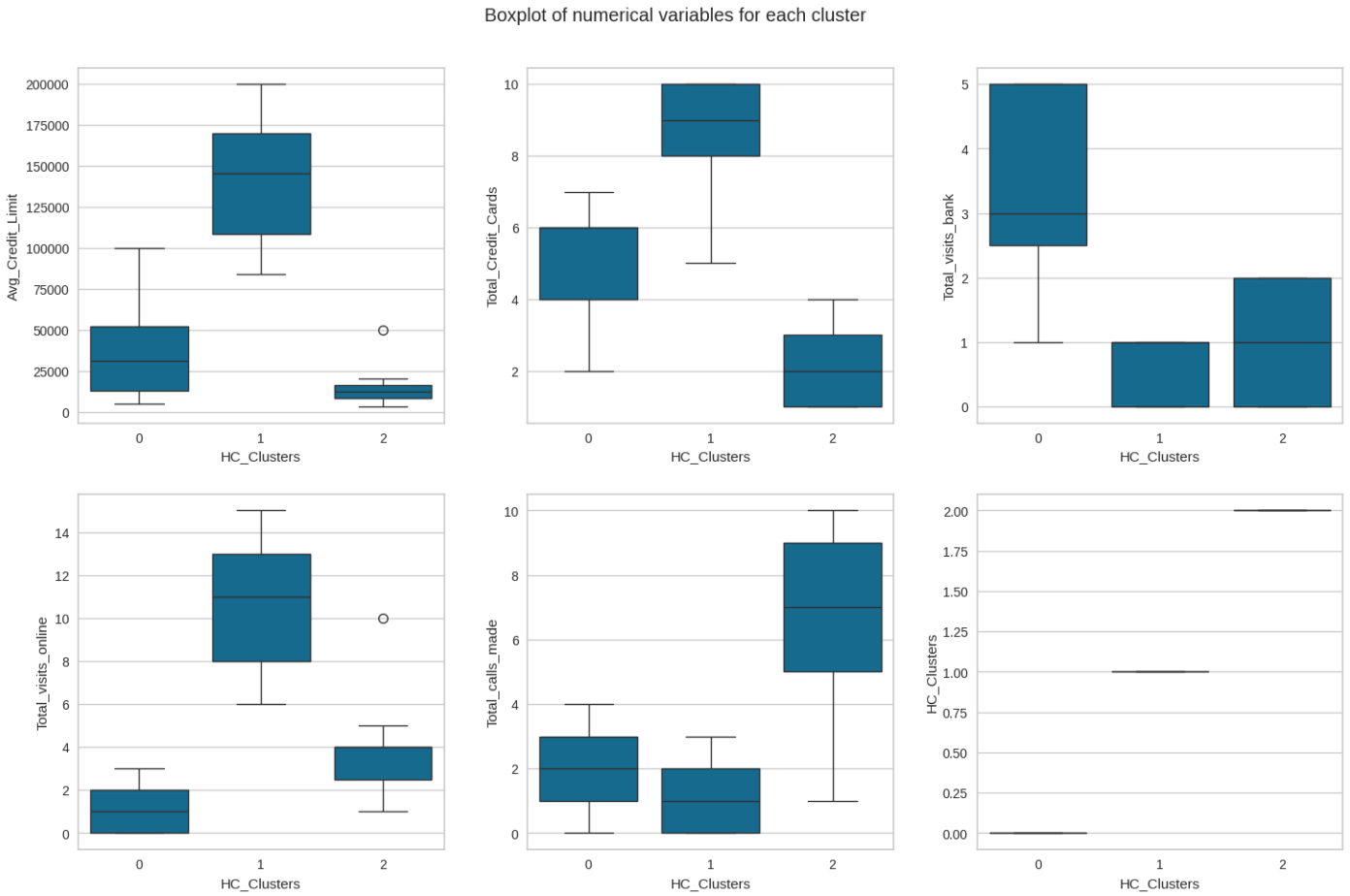

Figure 22 Box plot of numerical columns for each hierarchical cluster

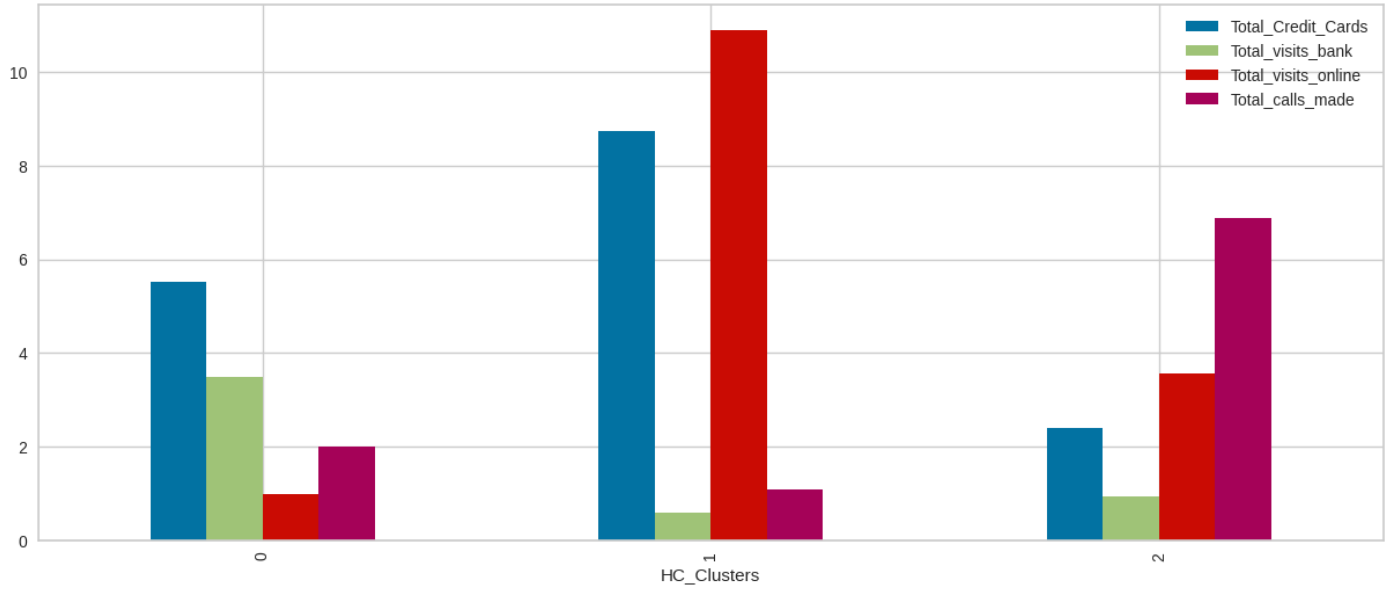

Figure 23 Bar chart for numerical columns for each hierarchical cluster

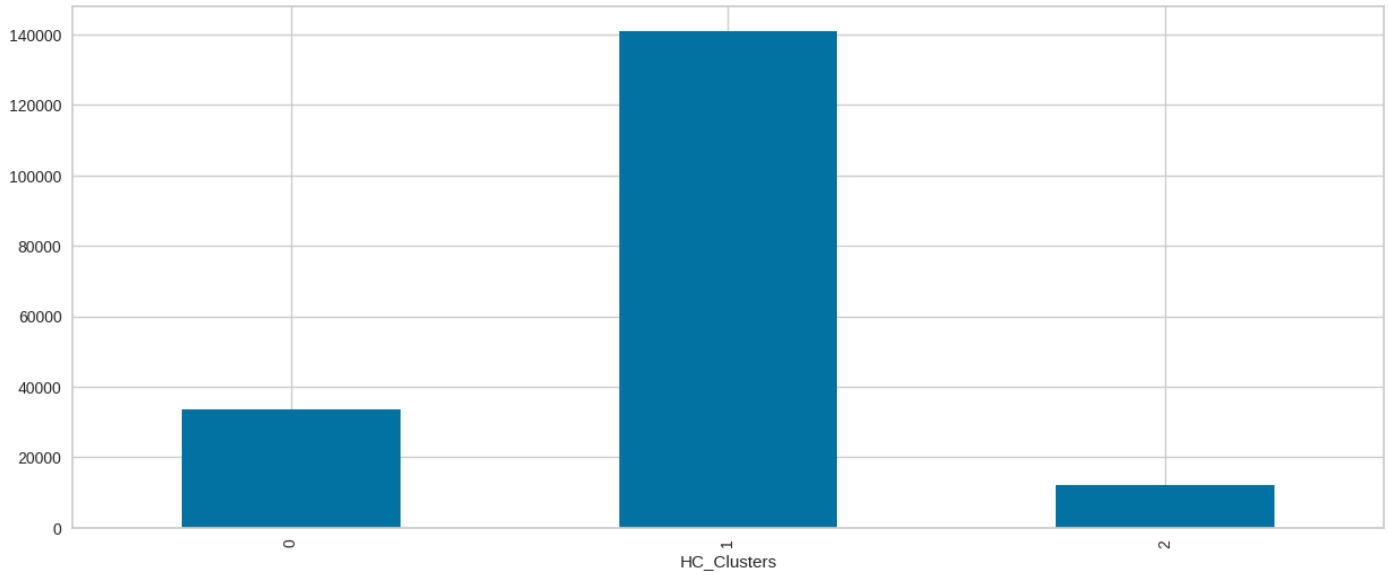

Figure 24 Bar chart for average credit limit for each hierarchical cluster

- The cluster profiles are very similar to that obtained from K-means clustering.
- Here, Cluster 0 has high bank visits.
- Cluster 1 has high average credit limit.
- Cluster 2 has a high number of calls.

6 K-Means Clustering vs Hierarchical Clustering

- K-means clustering and hierarchical clustering have produced very similar groups.
- Only for 1 record, the classification is different.
- Based on the univariate analysis charts, this customer can be kept in High_Calls category.

	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	K_means_segments	HC_Clusters
313	7000	4	2	2	4	High_Calls	High_Bank

Figure 25 K-Means Clustering vs Hierarchical Clustering

7 Actionable Insights & Recommendations

- Cluster High_Bank:
  - Roughly 59% of the customers fall in this cluster.
  - These customers rely on visiting the bank for their services.
  - The bank should be staffed accordingly, to cater to this group.

- Cluster High_Calls:
  - Roughly 33% of the customers fall in this cluster.
  - These customers rely on calls to the bank for their services.
  - The bank should ensure that there are enough call centre employees to cater to this group.

- Cluster High_CL:
  - Roughly 8% of the customers fall in this cluster.
  - These customers have a high credit limit.
  - They generally operate online.
  - The bank should provide reliable online services to cater to this group.
  - These customers also have many credit cards. The banks can offer credit cards for customers with a high credit limit.