All Life Bank

(Business Report)

<u>Report By:</u>
Naman Srivastava
02 Nov, 2025

TABLE OF CONTENT

List of Figures

Context

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.

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).

Variable	*Description*
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)

Exploratory Data Analysis

Exploratory Data Analysis was done on the provided data set using Python tools on google colab. The objective of entire process was done to make data more understandable, reliable for meaningful decision making.

Data Import and Cleaning

After successfully loading the data on the google colab notebook, and importing all the required libraries, we found out that on initial checking, that the data consists of **660 entries, and 6 features**.

- Loading head of the data

	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

fig 1: Head

Since we've no use of Sl_No column, so we have dropped it.

- Getting info of the data
  Info of the data after Sl_No column being dropped.

```
Data columns (total 6 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Customer Key         660 non-null    int64
 1   Avg_Credit_Limit     660 non-null    int64
 2   Total_Credit_Cards   660 non-null    int64
 3   Total_visits_bank    660 non-null    int64
 4   Total_visits_online  660 non-null    int64
 5   Total_calls_made     660 non-null    int64
```

fig 2: Info

- Checking for duplicated values

```
np.int64(0)
```

fig 3: Duplicated values

➢ Investigating null values

	0
Customer Key	0
Avg_Credit_Limit	0
Total_Credit_Cards	0
Total_visits_bank	0
Total_visits_online	0
Total_calls_made	0

fig 4: Null values

➢ Getting description of the data

	count	mean	std	min	25%	50%	75%	max
Customer Key	660.0	55141.443939	25627.772200	11265.0	33825.25	53874.5	77202.5	99843.0
Avg_Credit_Limit	660.0	34574.242424	37625.487804	3000.0	10000.00	18000.0	48000.0	200000.0
Total_Credit_Cards	660.0	4.706061	2.167835	1.0	3.00	5.0	6.0	10.0
Total_visits_bank	660.0	2.403030	1.631813	0.0	1.00	2.0	4.0	5.0
Total_visits_online	660.0	2.606061	2.935724	0.0	1.00	2.0	4.0	15.0
Total_calls_made	660.0	3.583333	2.865317	0.0	1.00	3.0	5.0	10.0

fig 5: Description of numerical data

➢ Data only has numerical data and no categorical data.
➢ Copying the data to another variable, and dropping customer key feature from it, for the **Univariate and Bivariate Analysis.**

Now we have sufficient idea about our data and we can proceed towards further analysis.

Univariate Analysis

In univariate analysis we analyse single variable individually.

- Average Credit Limit:
  We can see that this particular feature is rightly skewed and contains outliers.

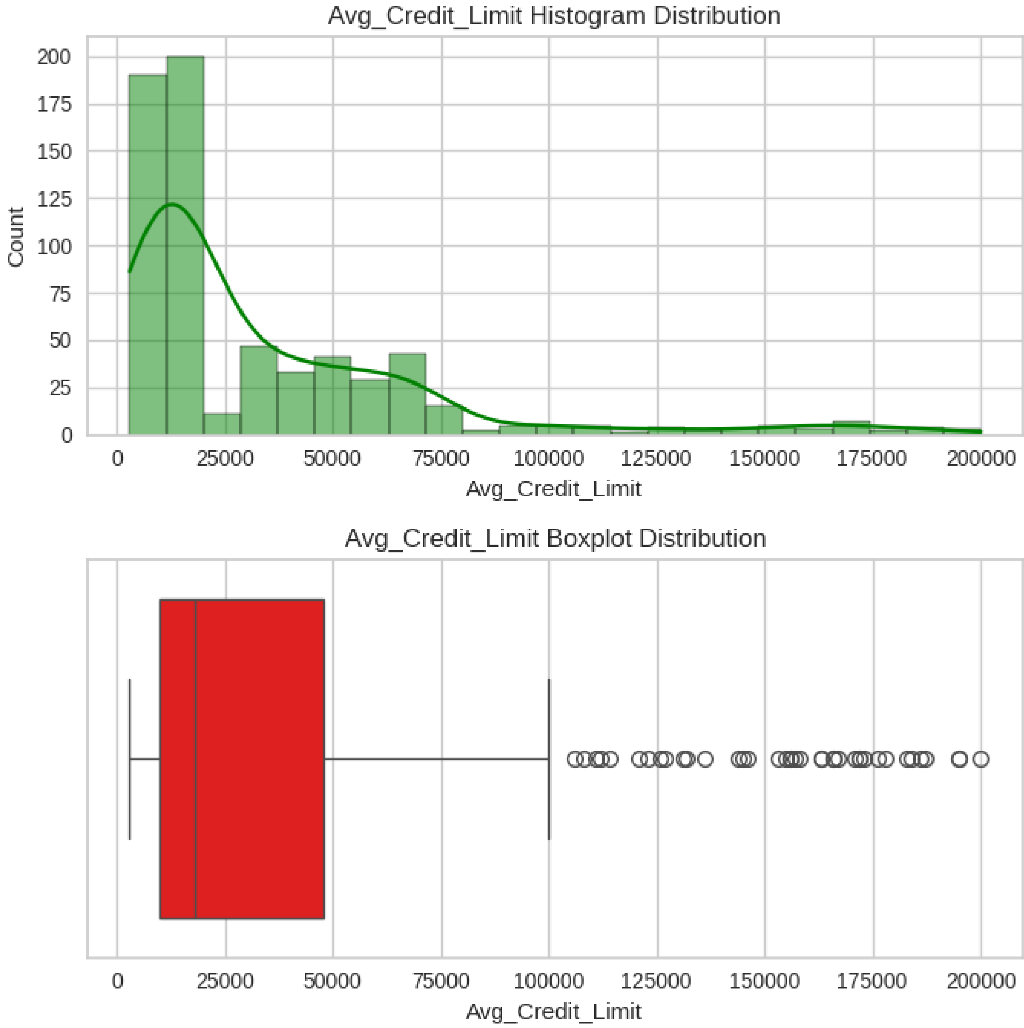

fig 6: Histogram and Boxplot for Average Credit Limit

- Total Credit Cards:
  This feature looks like bimodal distribution.

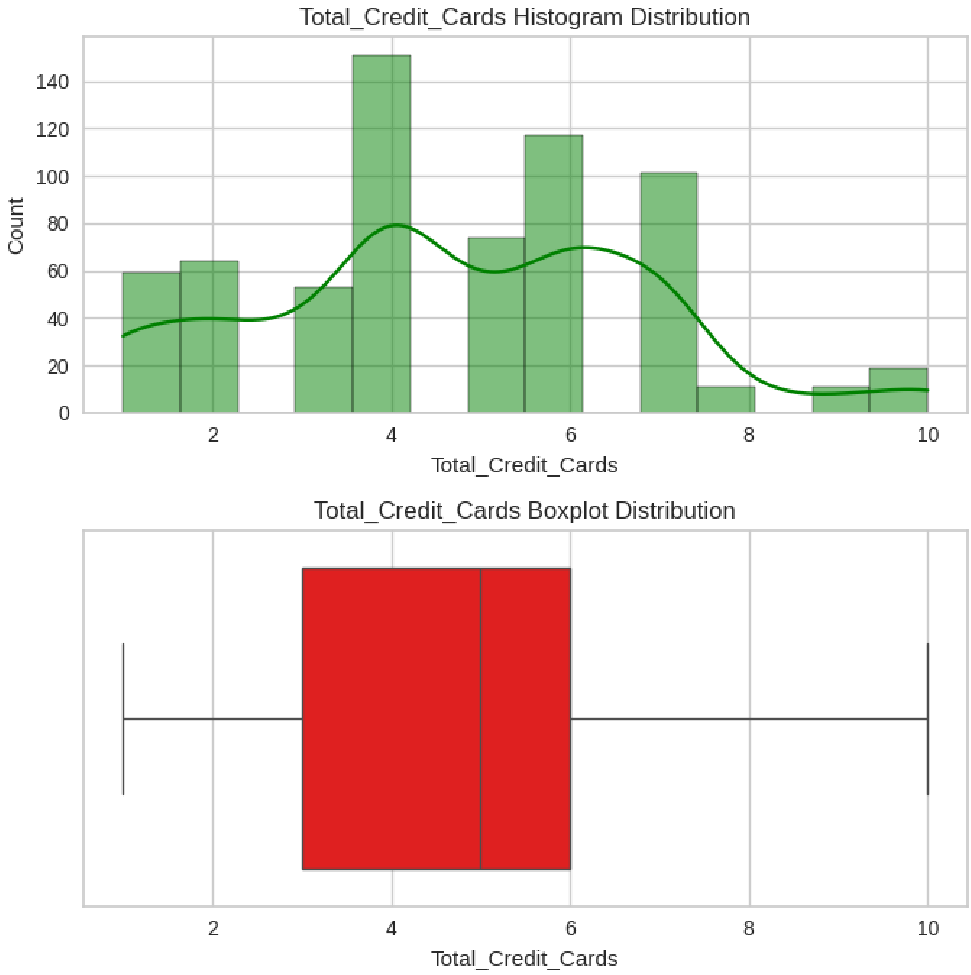

fig 7: Histogram and boxplot for Total Credit Cards

- Total Visits Bank:
  This feature is somewhat normally distributed.

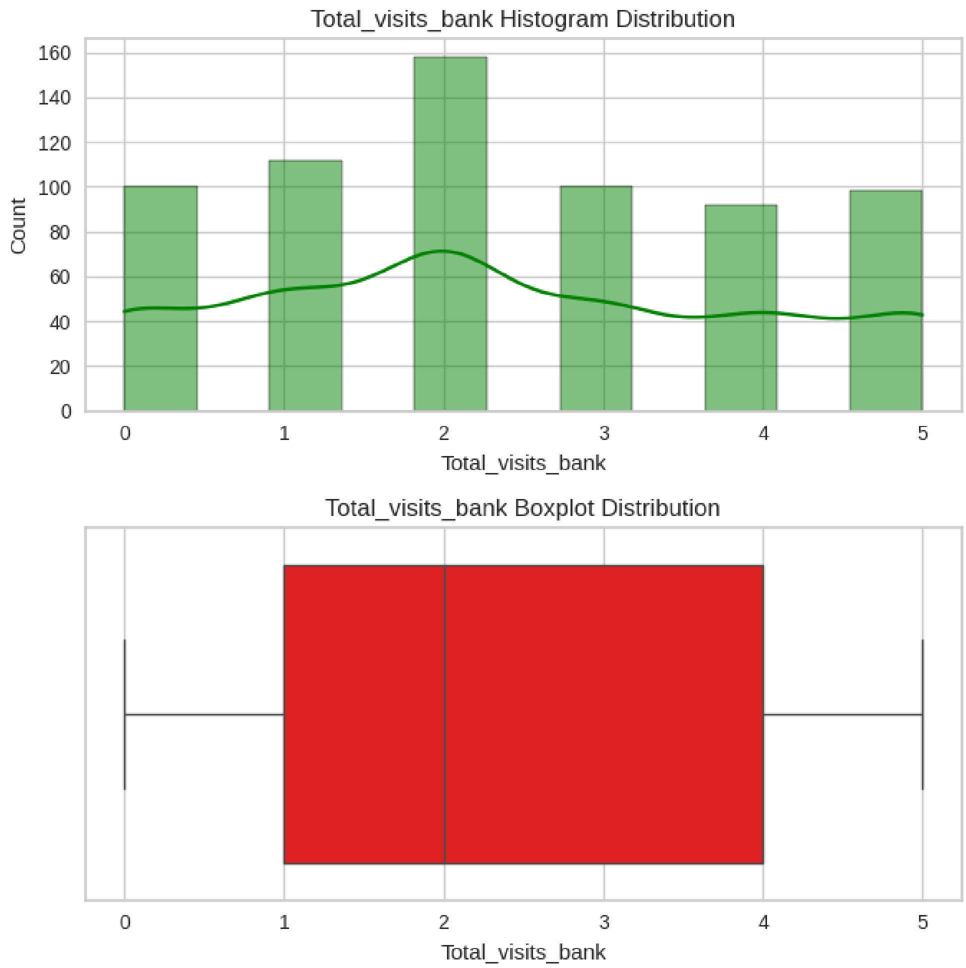

fig 8: Histogram and boxplot for Total visits bank

- Total visits online:
  This feature is rightly skewed and contains some outliers.

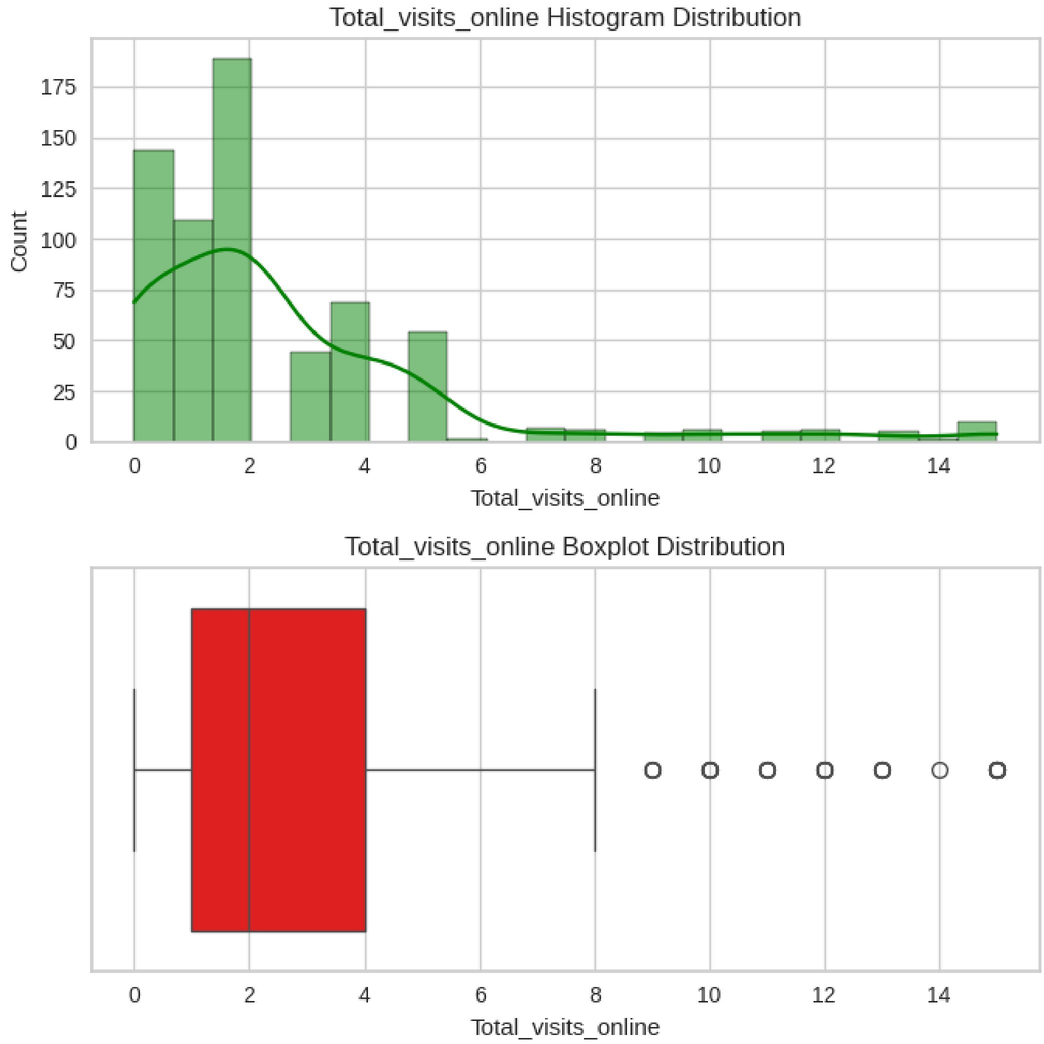

fig 9: Histogram and boxplot for Total visits online

- Total calls made:
  Below is the histogram and boxplot distribution of the feature.

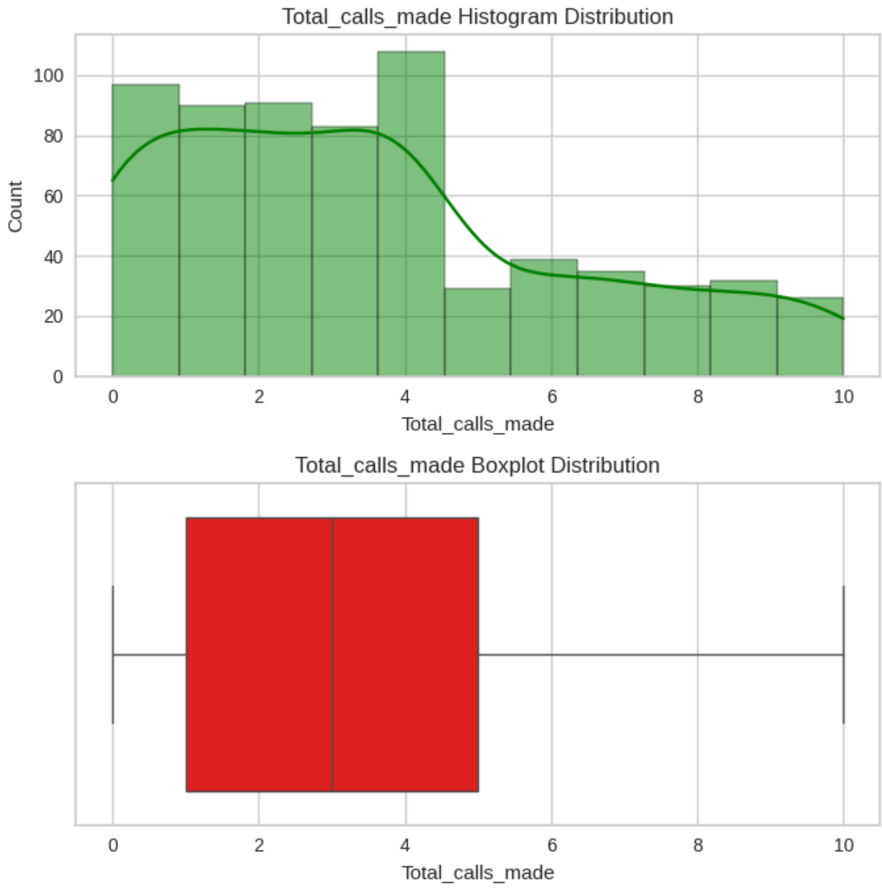

fig 10: Histogram and boxplot Total calls made

Bivariate Analysis

Here, we analyse the relationship between 2 variables.

- Average Credit Limit vs Total Credit Cards:
  Higher the number of cards, higher is the card limit.

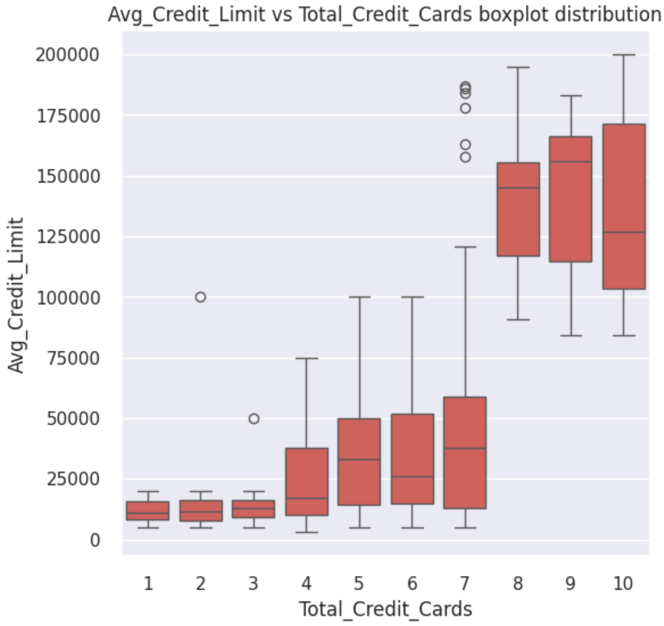

fig 11: Average Credit Limit vs Total Credit Cards

- Average Credit Limit vs Total Visits Banks:
  People who visit bank, on a average have similar card limits.

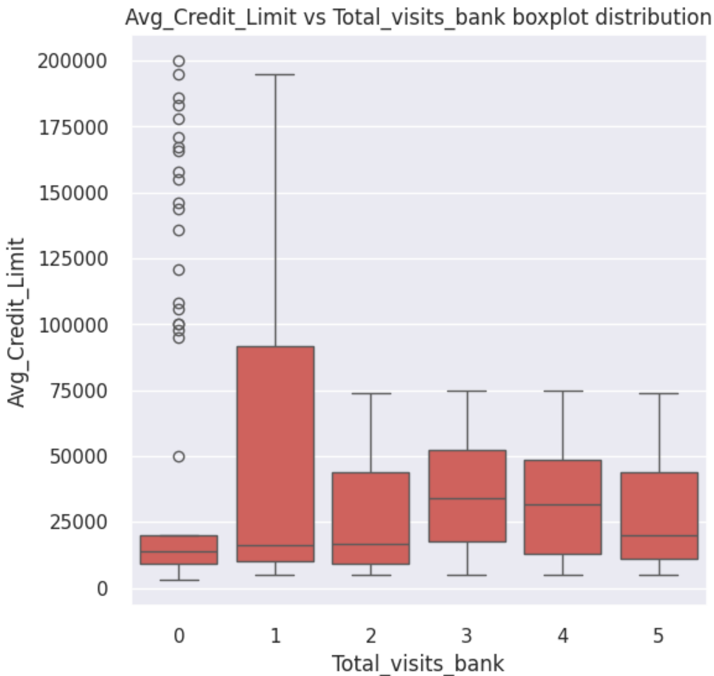

fig 12: Average Credit Limit vs Total visits banks

- Average Credit Limit vs Total Visits Online:
  People who visit online tend to have higher credit limits.

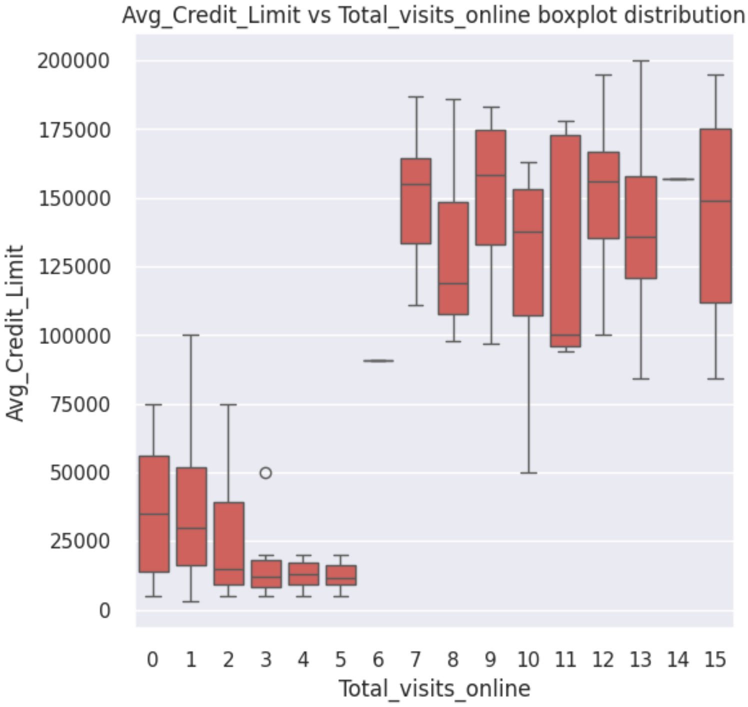

fig 13: Average Credit Limit vs Total Visits Online

- Average Credit Limit vs Total Calls Made:
  People who engage more on calls tend to have less card limits.

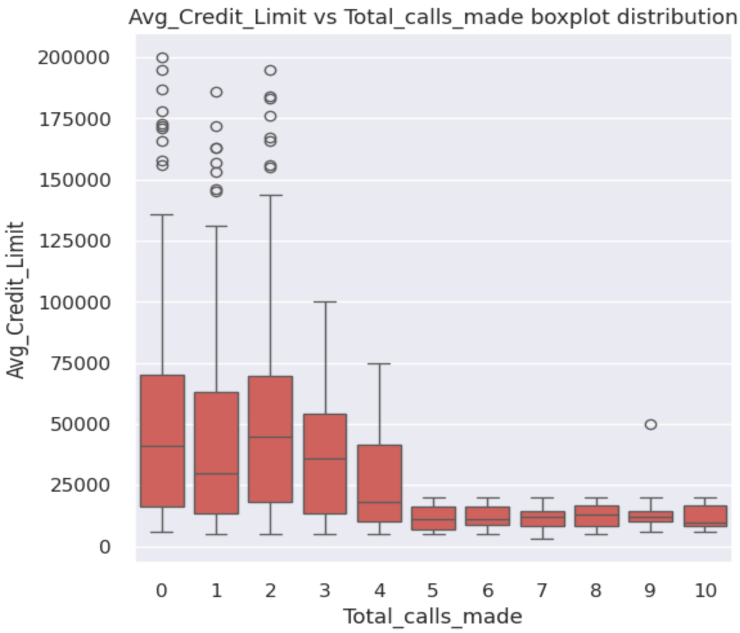

fig 14: Average Credit Limit vs Total Calls Made

- Total Credit Cards vs Total Visits Online:
  People visiting online, have more number of cards than the others.

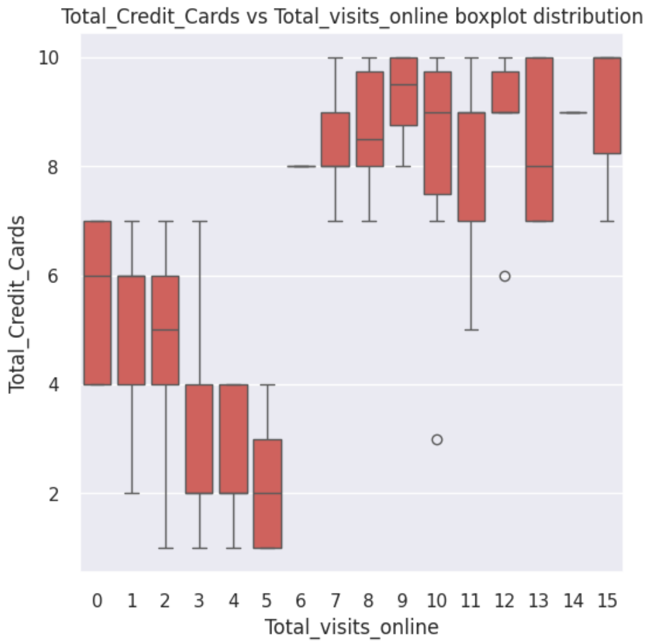

fig 15: Total Credit Cards vs Total Visits Online

- Total Credit Cards vs Total Call Made:
  People engaging via calls have lesser number of the cards.

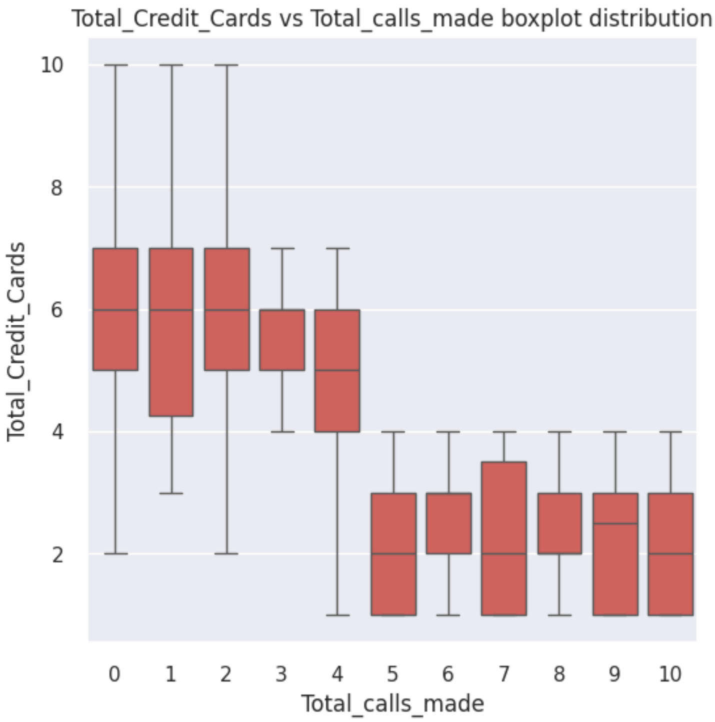

fig 16: Total Credit Cards vs Total Call Made

- Total Credit Cards vs Total Visits Bank:
  People visiting bank have average number of cards, not too less like the ones who engage on call, not too many like the ones who visit online.

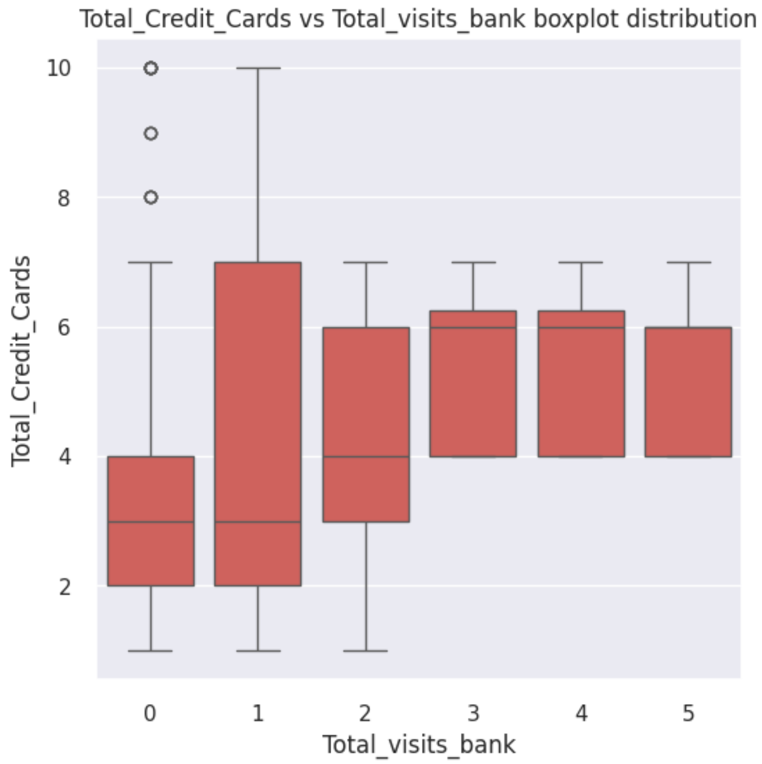

fig 17: Total Credit Cards vs Total Visit Bank

Heatmap

For better understanding of the numerical vs numerical data, we created Heatmap.

In the heatmap we can see the corelation between different numerical data.

Colour more towards blue, means positive corelation is high.

Colour more towards red, means negative corelation is high.

fig 18: Heatmap

Pair Plot

Similar to heatmap, pair plot is the graphical representation of corelation among the numerical features of the data.

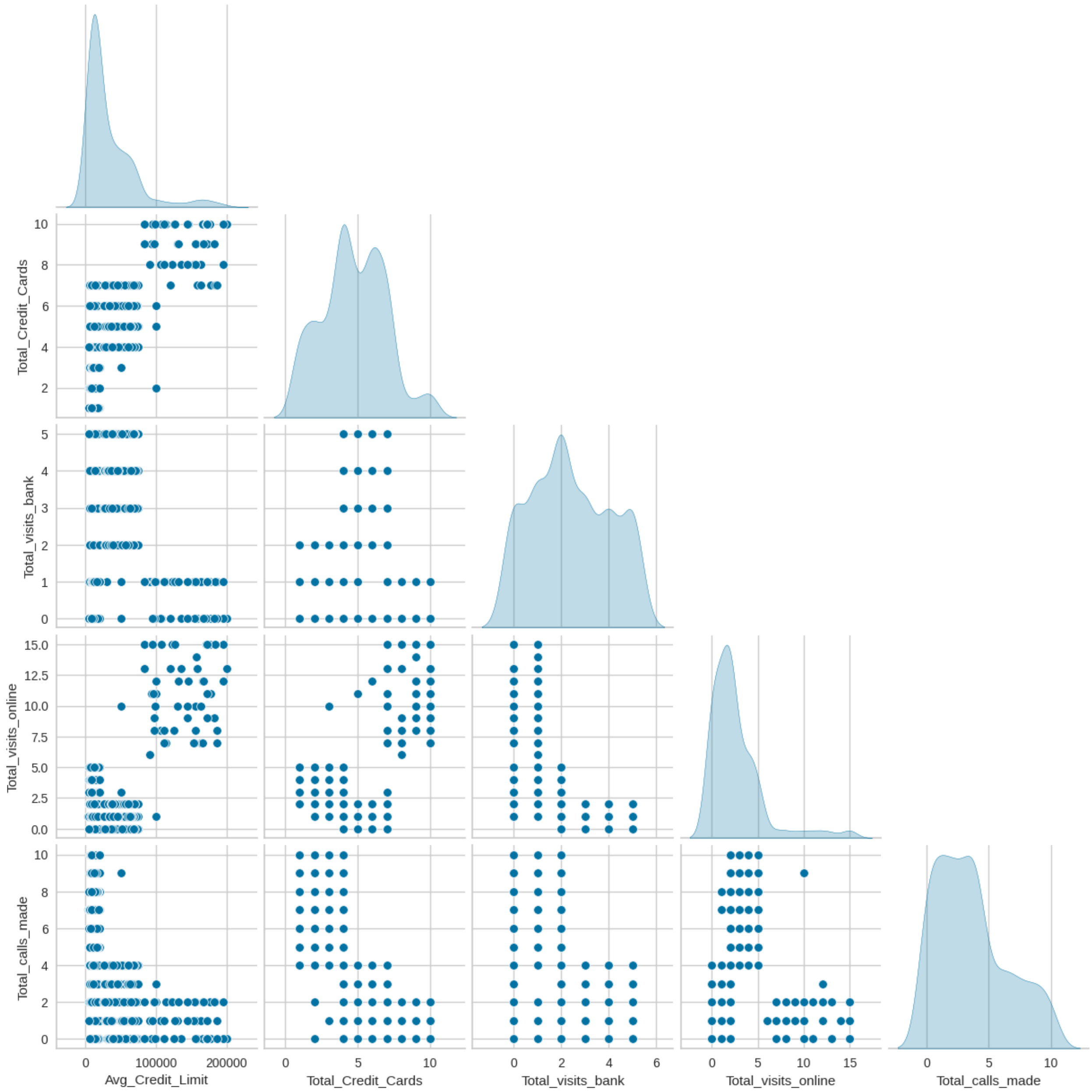

fig 19: Pair plot

KEY INSIGHTS AS PER EDA

- People who visit bank website online are having higher number of cards and also higher is their card limit.
- People who call to bank are seen to have lower number of cards.
- More the person has number of cards, higher is their card limit.
- Less the person has number of cards, lower is their card limit.
- People who visit bank more, their online activity is less.
- People who call to the bank are less likely to visit bank.

OUTLIERS TREATMENT

We have detected outliers in the data set, but those outliers seem to be genuine, therefore, we are not going to treat the outliers. We will keep the outliers in the data and will move ahead with the same.

FEATURE ENGINEERING

Feature Engineering is not specifically required here.

DATA SCALING

Data has been scaled, as K means clustering and Hierarchical clustering both require data scaling. Also, there is a significant difference in the values of average card limit and rest of the features.

K-MEANS CLUSTERING

K-Means Clustering is an unsupervised machine learning algorithm that helps group data points into clusters based on their inherent similarity.

Elbow Method

```
Number of Clusters: 1   Average Distortion: 2.006922226250361
Number of Clusters: 2   Average Distortion: 1.4571553548514269
Number of Clusters: 3   Average Distortion: 1.1466276549150365
Number of Clusters: 4   Average Distortion: 1.0463825294774463
Number of Clusters: 5   Average Distortion: 1.052013445015247
Number of Clusters: 6   Average Distortion: 0.9429600194368428
Number of Clusters: 7   Average Distortion: 0.9104808769756559
Number of Clusters: 8   Average Distortion: 0.9211671231933618
Number of Clusters: 9   Average Distortion: 0.8686088356532263
Text(0.5, 1.0, 'Selecting k with the Elbow Method')
```

fig 20: Number of clusters and their average distortion

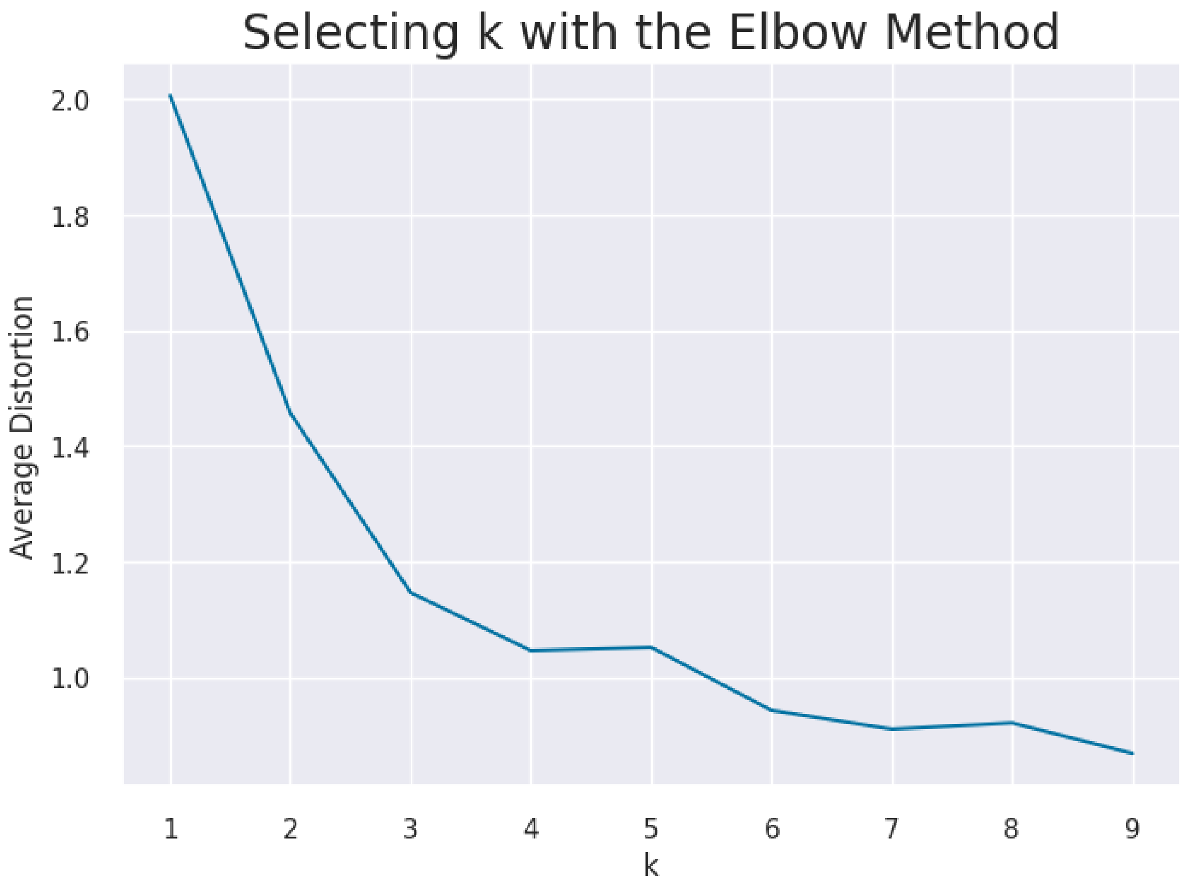

fig 21: Elbow Method for selecting K value

As per the above graph our k value could be 3 or 4.

Silhouette Score

It helps us measure how well each data point fits into its assigned cluster and how far it is from other clusters.

Below are the values of Silhouette Scores:

```
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 0.2726684472506887
For n_clusters = 6, silhouette score is 0.22746263373740702
For n_clusters = 7, silhouette score is 0.2471011696944927
For n_clusters = 8, silhouette score is 0.20677623826520972
For n_clusters = 9, silhouette score is 0.2241954769365604
```

fig 22: Silhouette Scores

fig 23: Silhouette Scores Graph

Silhouette Plot for K Means Clustering

- **7 Centres**

fig 24: Silhouette Plot for K Means Clustering 7 Centres

- **6 Centres**

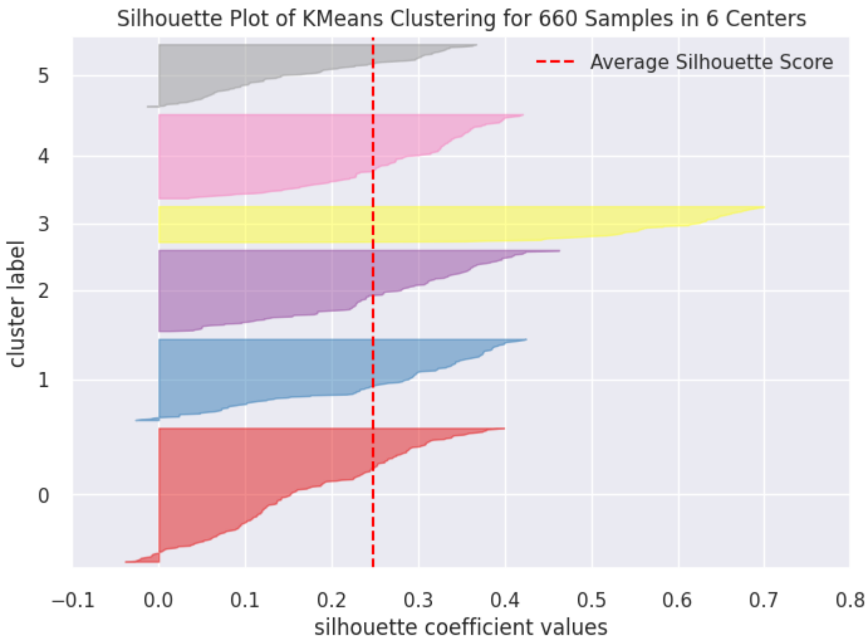

fig 25: Silhouette Plot for K Means Clustering 6 Centres

- **5 Centres**

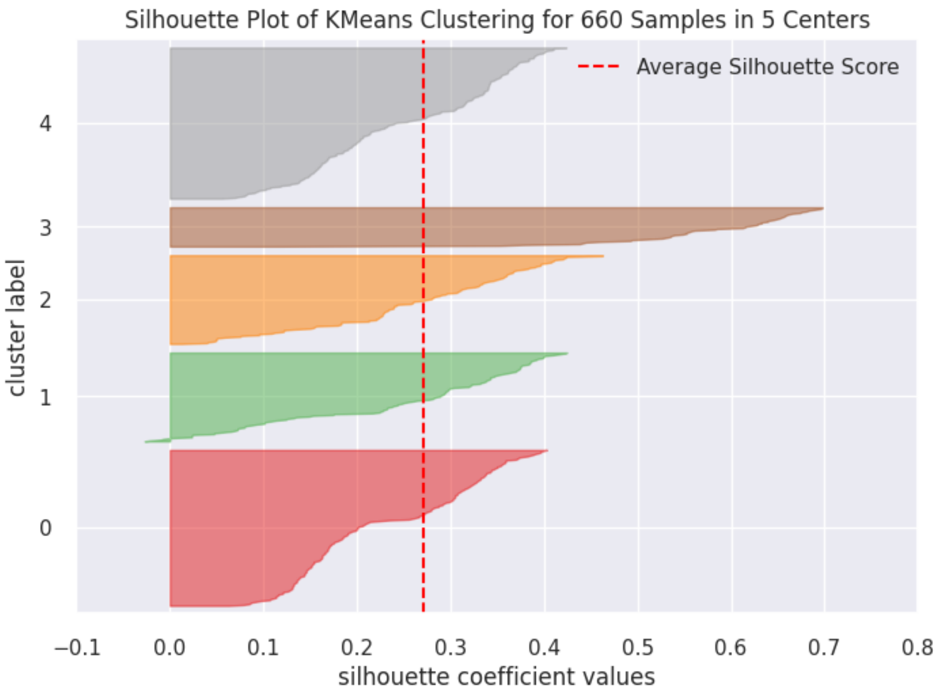

fig 26: Silhouette Plot for K Means Clustering 5 Centres

- **4 Centres**

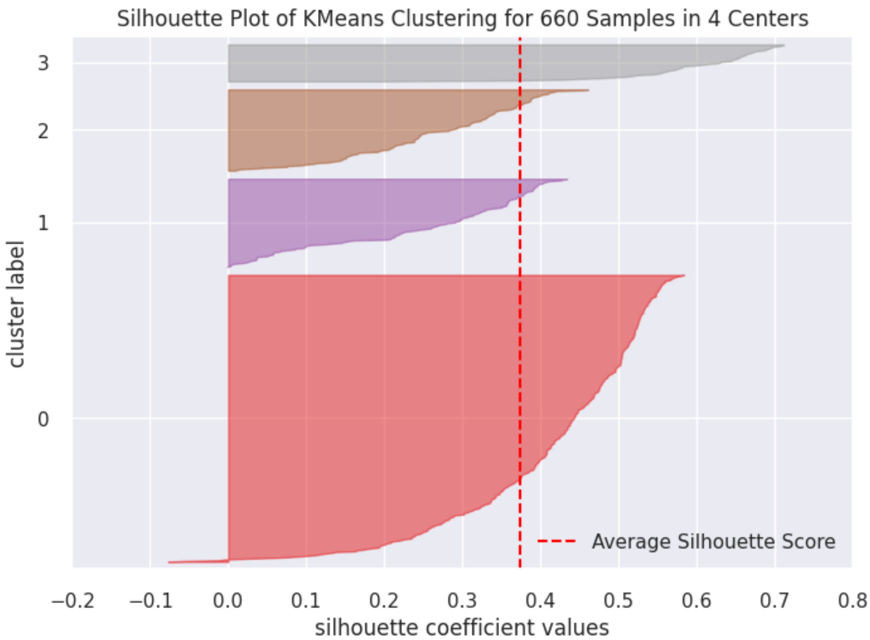

fig 27: Silhouette Plot for K Means Clustering 4 Centres

- **3 Centres**

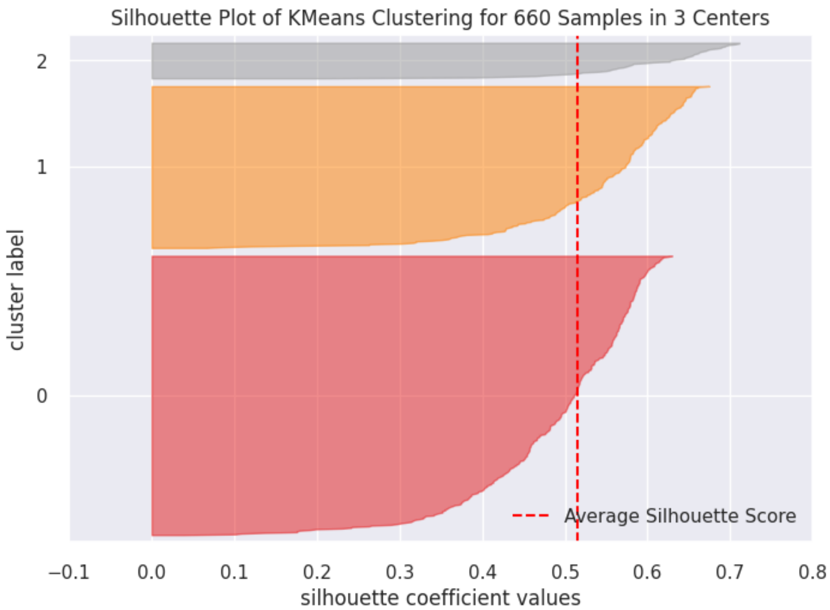

fig 28: Silhouette Plot for K Means Clustering 3 Centres

➢ **2 Centres**

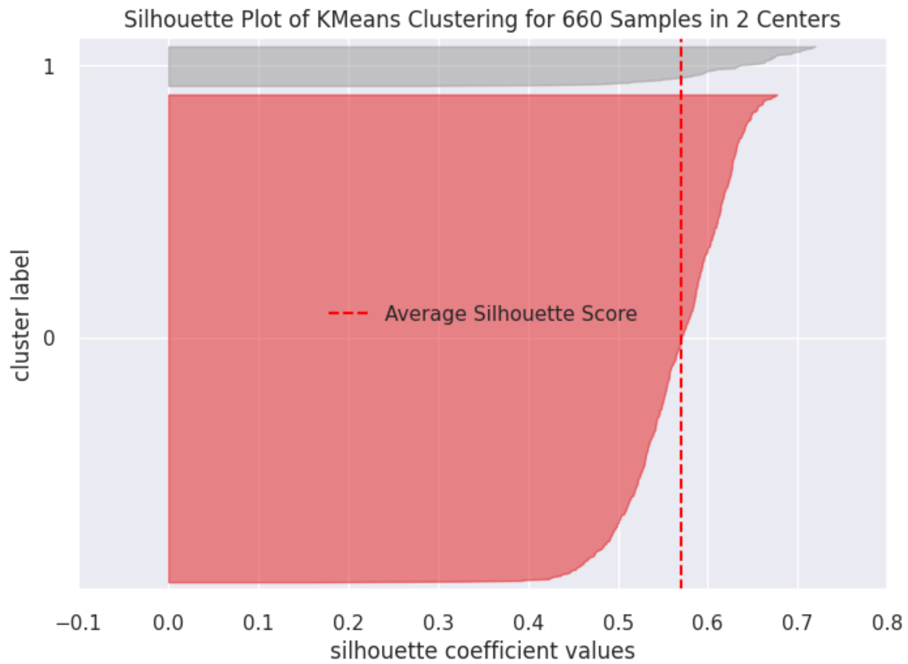

fig 29: Silhouette Plot for K Means Clustering 2 Centres

After going through silhouette plots, it looks like value of k would be optimal at 3.

This is because at values more than 3, the plot is showing negative values also.

And below 3, the algorithm becomes overly simplified.

Thus, we are considering optimal number of clusters at 3.

Also, the same was seen through, elbow method and silhouette score methods.

Cluster Profiling

K_means_segments	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	54881.329016	33782.383420	5.515544	3.489637	0.981865	2.000000	386
1	55239.830357	12174.107143	2.410714	0.933036	3.553571	6.870536	224
2	56708.760000	141040.000000	8.740000	0.600000	10.900000	1.080000	50

fig 30: Cluster Profiling

Cluster 0

- It carries the greatest number of people.
- People in this cluster are the ones who visit bank the most.
- Rest of the features in this clusters are more like mid segment ones.

Cluster 1

- This cluster people are not the highest and also not the lowest.
- They are the ones who phone call the bank most.
- Also, they have the least number of credit cards.
- Their card limits are also the least.

Cluster 2

- It has the least number of people.
- People in the segment are the ones who use online portal most.
- They also have highest average card limit.
- They have the highest number of credit cards.

Boxplot of numerical variables for each cluster

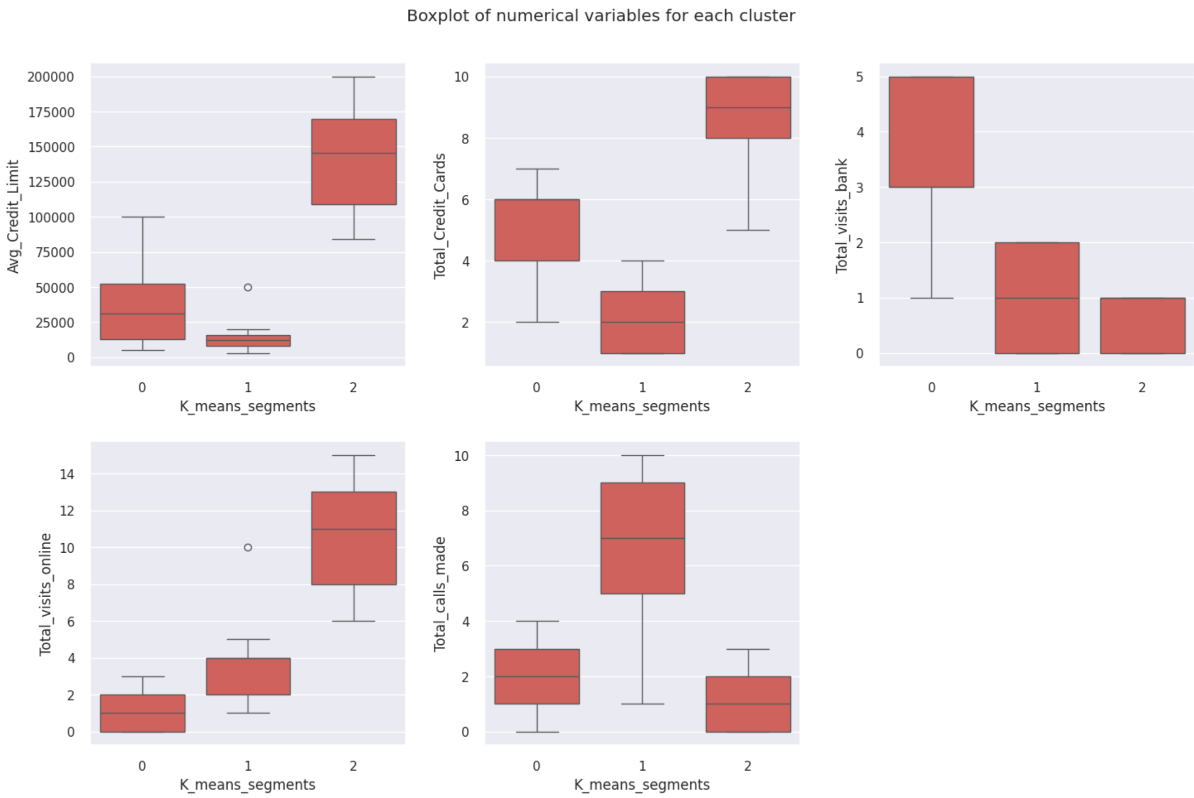

fig 31: Boxplot of the numerical variables of each cluster

Barplot of K means segment

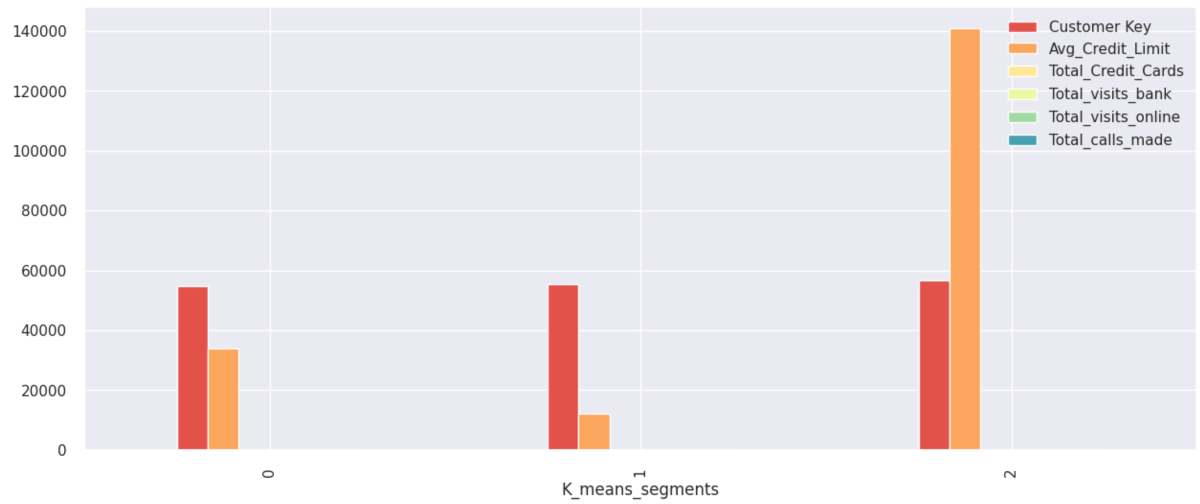

fig 32: Bar plot of K means segment

HIERARCHICAL CLUSTERING

Hierarchical clustering is an unsupervised learning technique used to group similar data points into clusters by building a hierarchy (tree-like structure). Unlike flat clustering like k-means hierarchical clustering does not require specifying the number of clusters in advance.

```
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.
```

fig 33: Cophenetic correlation for different distances and different linkages

Highest cophenetic correlation is 0.8977080867389372, which is obtained with **Euclidean distance and average linkage**.

```
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.
```

fig 34: Cophenetic correlation for Euclidean distances and different linkages

Highest cophenetic correlation is 0.8977080867389372, which is obtained with **average linkage.**

Dendrograms

Dendrogram for different linkages.

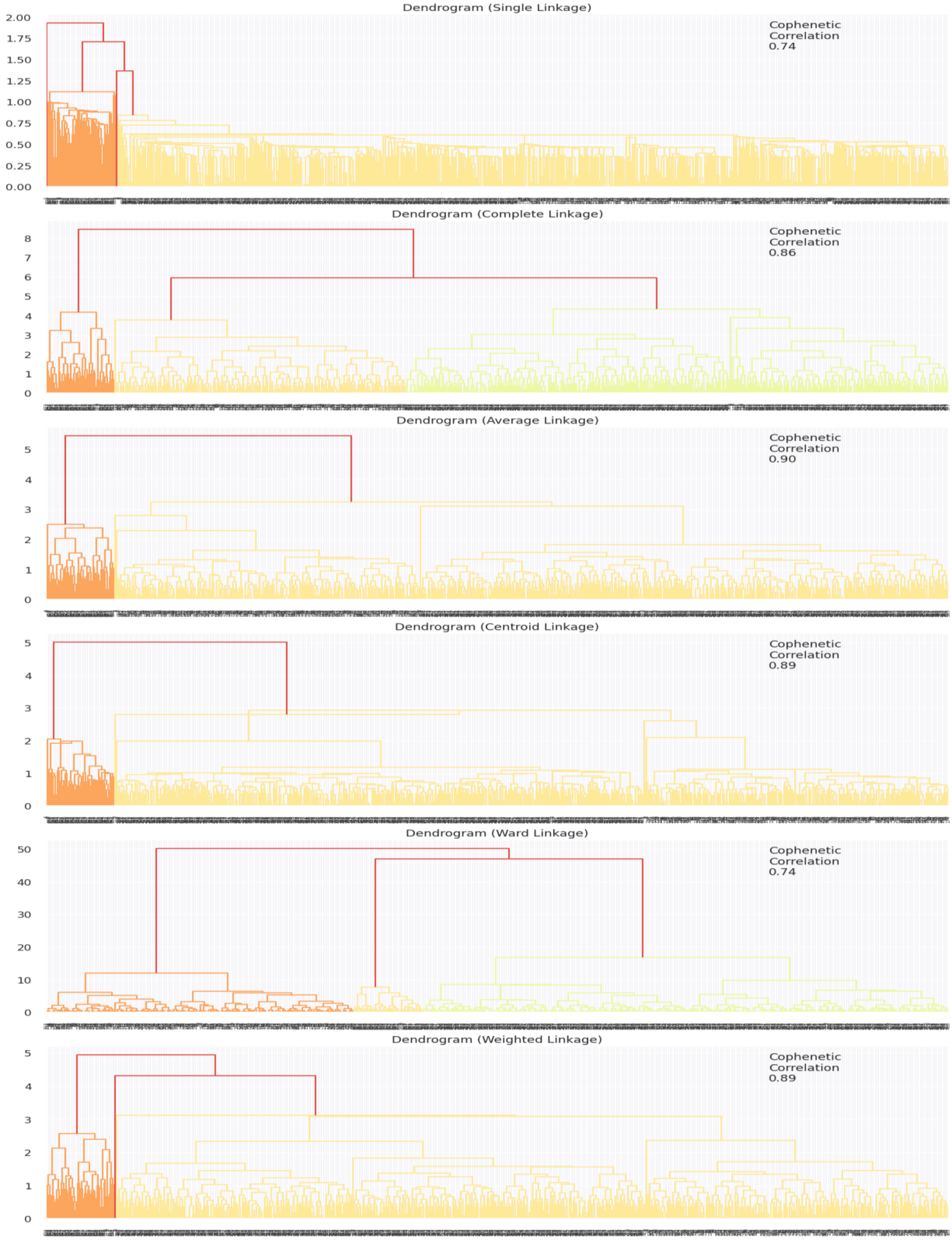

fig 35: Dendrograms

Considering Average linkage at total number of cluster equal to 3 for the further analysis, as it has highest cophenetic correlation.

Cluster Profiling

HC_Clusters	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	K_means_segments	count_in_each_segments
0	54925.966408	33713.178295	5.511628	3.485788	0.984496	2.005168	0.002584	387
1	56708.760000	141040.000000	8.740000	0.600000	10.900000	1.080000	2.000000	50
2	55163.973094	12197.309417	2.403587	0.928251	3.560538	6.883408	1.000000	223

fig 36: Cluster Profiling

Cluster 0

- It carries the greatest number of people.
- People in this cluster are the ones who visit bank the most.
- Rest of the features in this clusters are more like mid segment ones.

Cluster 1

- It has the least number of people.
- People in the segment are the ones who use online portal most.
- They also have highest average card limit.
- They have the highest number of credit cards.

Cluster 2

- This cluster people are not the highest and also not the lowest.
- They are the ones who phone call the bank most.
- Also, they have the least number of credit cards.
- Their card limits are also the least.

Boxplot of numerical variables for each cluster

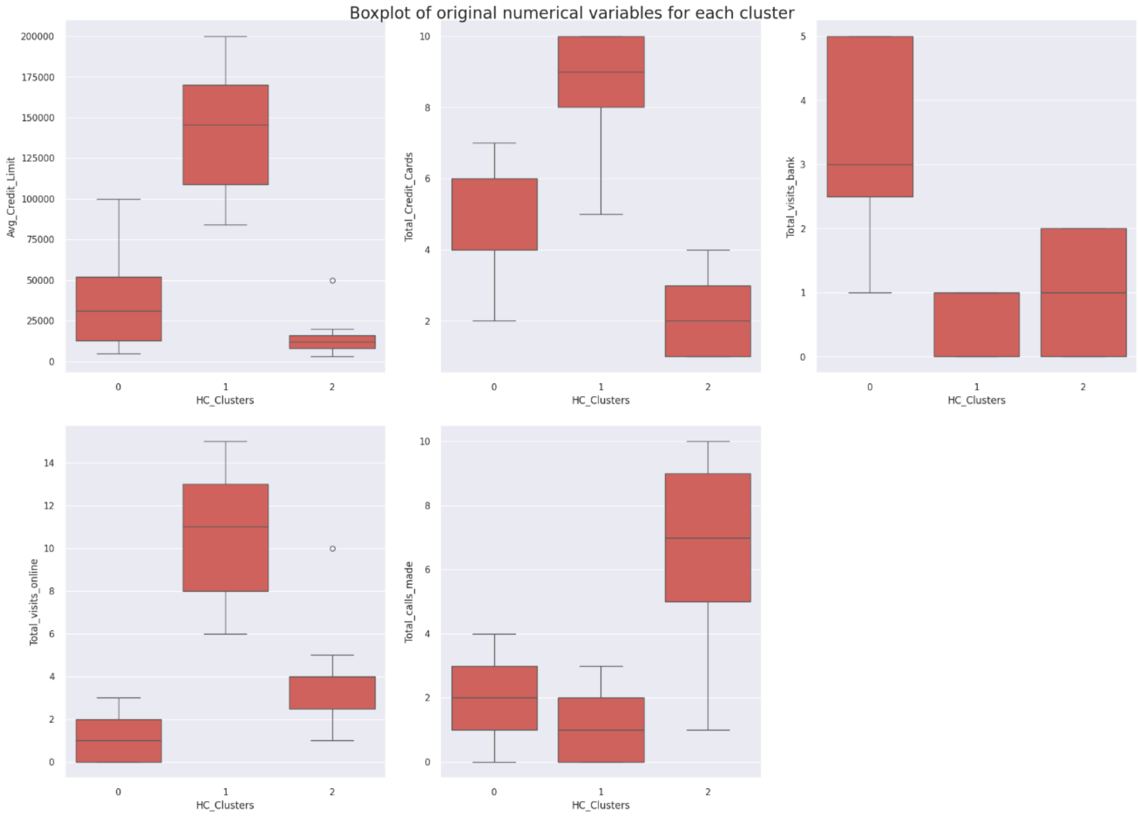

fig 37: Boxplot of numerical variables for each other

COMPARISON K MEANS CLUSTERING AND HIERARCHICAL CLUSTERING

Below is the K means Clustering and Hierarchical Clustering comparison of their cluster profiling

K-Means Cluster Profile:

K_means_segments	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	54881.329016	33782.383420	5.515544	3.489637	0.981865	2.000000	386
1	55239.830357	12174.107143	2.410714	0.933036	3.553571	6.870536	224
2	56708.760000	141040.000000	8.740000	0.600000	10.900000	1.080000	50

Hierarchical Clustering Cluster Profile:

HC_Clusters	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	K_means_segments	count_in_each_segment
0	54925.966408	33713.178295	5.511628	3.485788	0.984496	2.005168	0.002584	387
1	56708.760000	141040.000000	8.740000	0.600000	10.900000	1.080000	2.000000	50
2	55163.973094	12197.309417	2.403587	0.928251	3.560538	6.883408	1.000000	223

fig 38: Comparison

KEY TAKEAWAYS FOR THE BUSINESS

- Bank should focus more on the people who visit online website often, as they avail the highest credit limit and also possess highest number of cards. But they are only 50 in numbers out of the total population of 660 people. Increasing number of these people will directly improve the market for the bank.
- Bank should ensure good ambiance in the bank itself, as the people who visit bank are significant in number almost 60%, these people also have good card limit and decent number of credit cards.
- Bank should strategize to onboard people, who call to the bank, on the internet banking or online banking. Because, these people do visit branch often and calls are not helping them in getting more number credit cards. This will make people to move for this cluster and increase size of the most marketable cluster.