Unsupervised Learning on All Life Bank (UL Coded)

Apr B Sunday Onkar 10:30 AM Batch

Arindam Saha

Table of Contents

List of Tables:

List of Figures:

1. 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

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.

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).

4. Data Dictionary

Variables	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)

Table 1: Data Dictionary

5. Data Information

- The dataset has 660 rows and 7 columns
- All the variables are of integer type.
- There are no null values present in the dataset
- The space occupied by the data is 36.2 KB
- There are no missing values in our data

	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made

count	660	660	660	660	660
mean	34574.24242	4.706061	2.40303	2.606061	3.583333
std	37625.4878	2.167835	1.631813	2.935724	2.865317
min	3000	1	0	0	0
25%	10000	3	1	1	1
50%	18000	5	2	2	3
75%	48000	6	4	4	5
max	200000	10	5	15	10

Table 2: Summary of the data

Observations

- For avg credit limit the mean and median seems to be not matching and there is a presence of outliers can be observed.

6. Exploratory Data Analysis

6.1. Univariate analysis

6.1.1. Observation On Total Credit Card

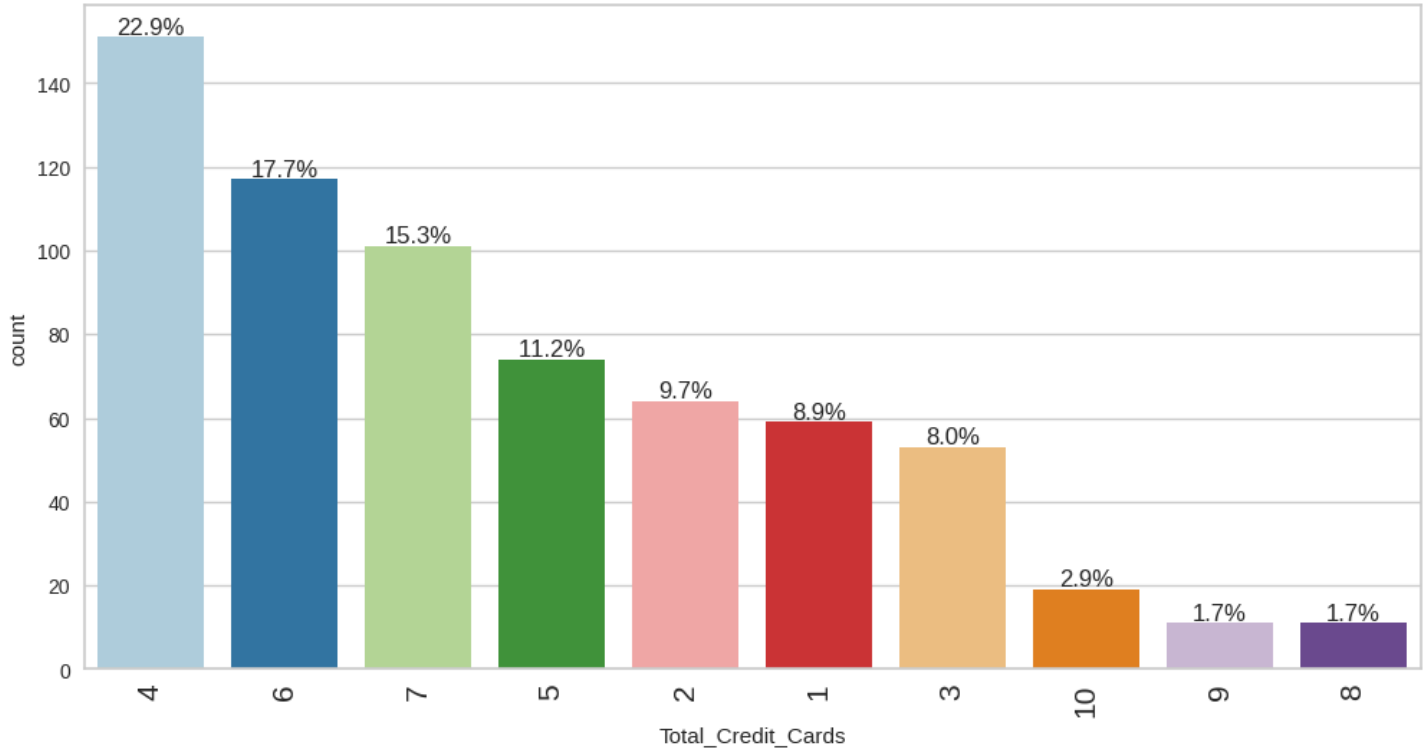


Figure 1: Univariate on Total Credit Card

- Just below one fifth of the customers have at least 4 credit cards, while more than eight credit card holders are at below five percentiles.

6.1.2. Observation on total visits bank

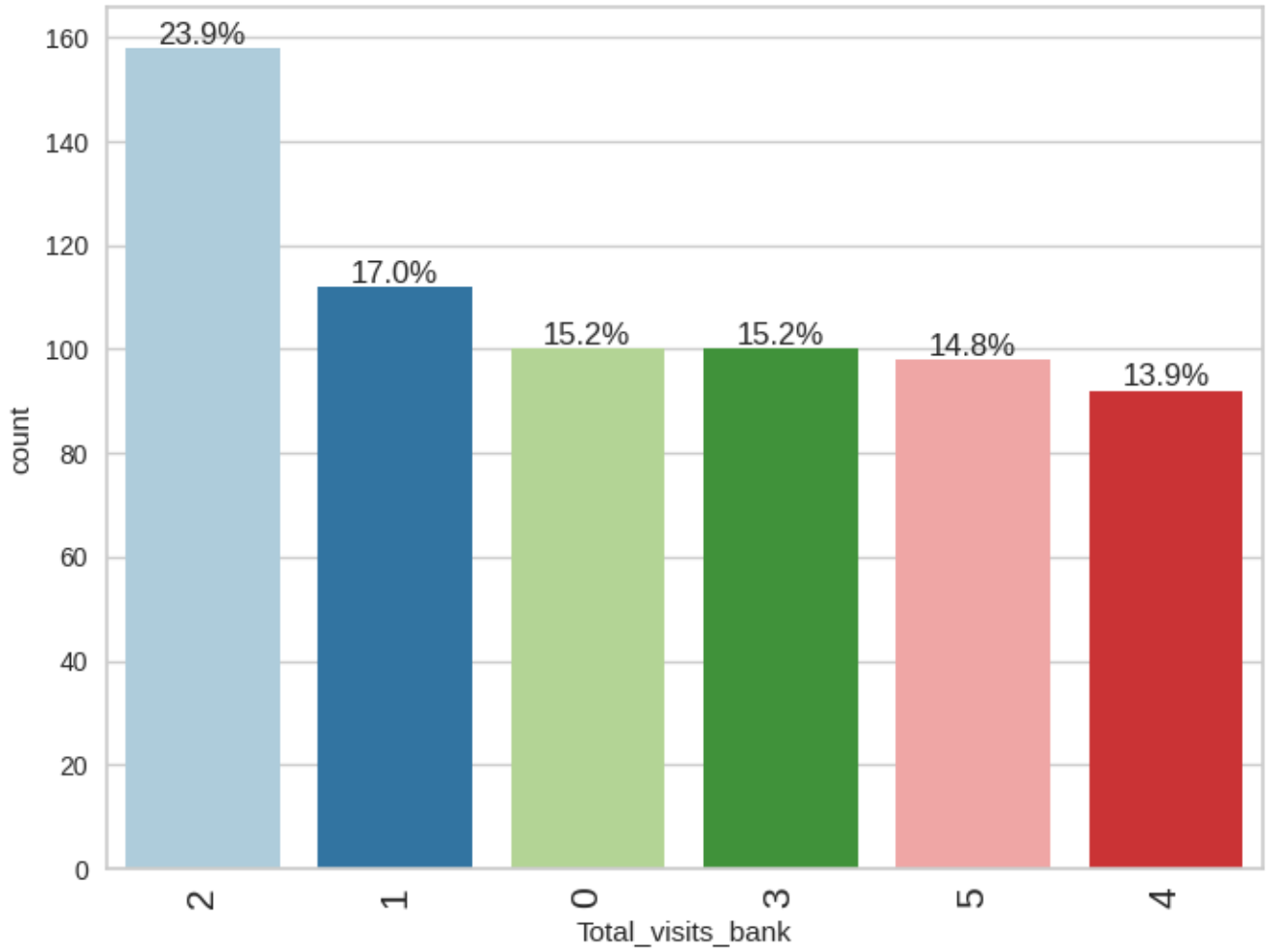


Figure 2: Univariate on total visits bank

- About 24 percentiles of the customers are visiting the bank twice, while other numbers of time visiting are about just below one fifth of the total proportion.

6.1.3. Observation on total visits online

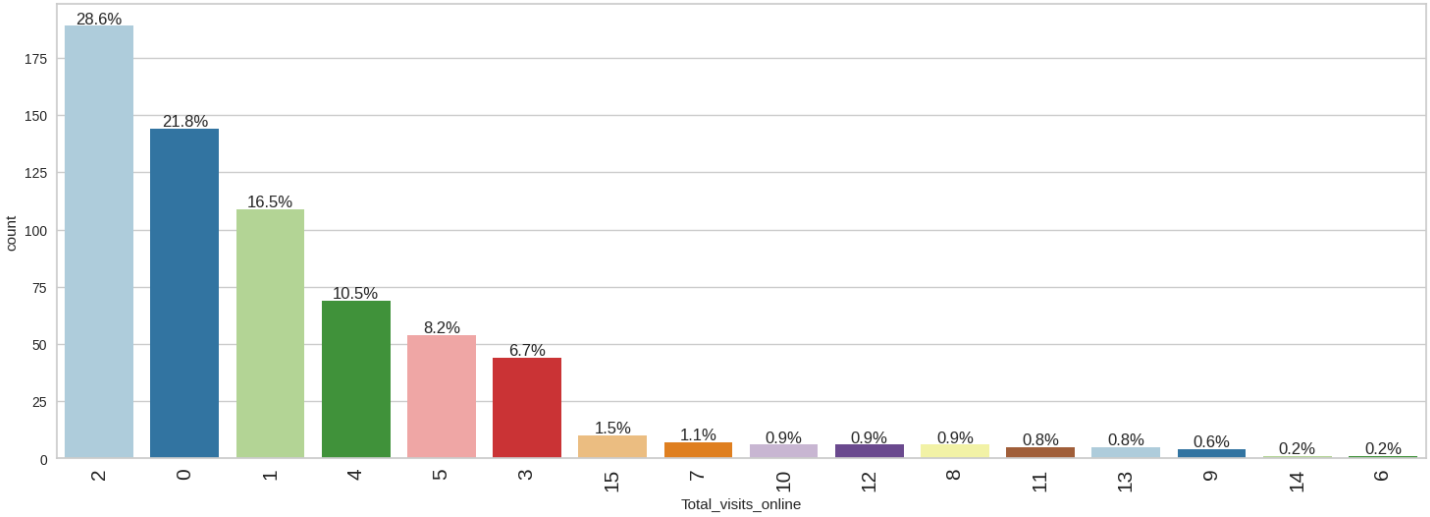


Figure 3: Univariate on total visits online

6.1.4. Observation on total calls made

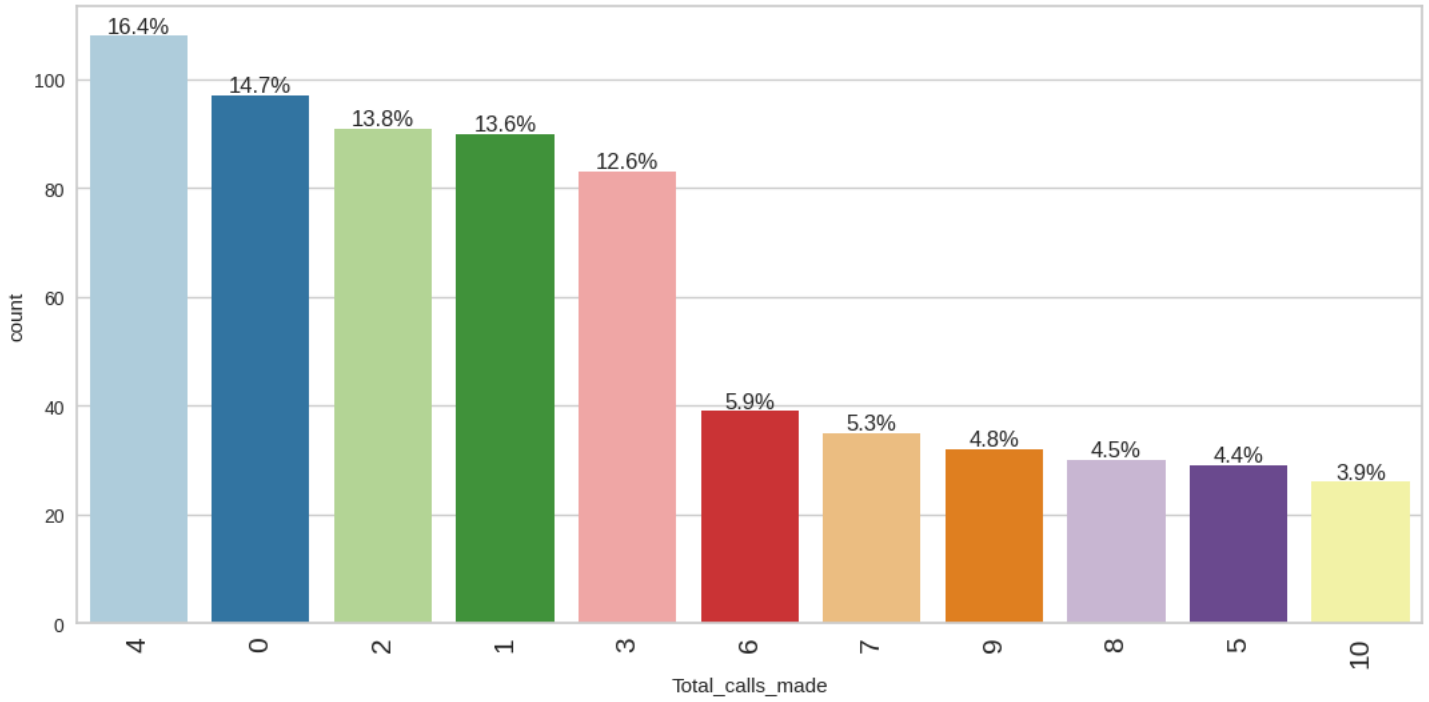


Figure 4: Univariate on total calls made

- More than 65% of the colleges have a rating less than 4 for placements.
- More than 80% of the colleges have a rating of 3 or more for infrastructure.

6.2. Bivariate Analysis

Figure 5: Heat Map

Observation

- Rating for teaching is strongly positively correlated with the rating for placements and internships.
- This is obvious because if teaching quality is high, students are more likely to get placements and internships.

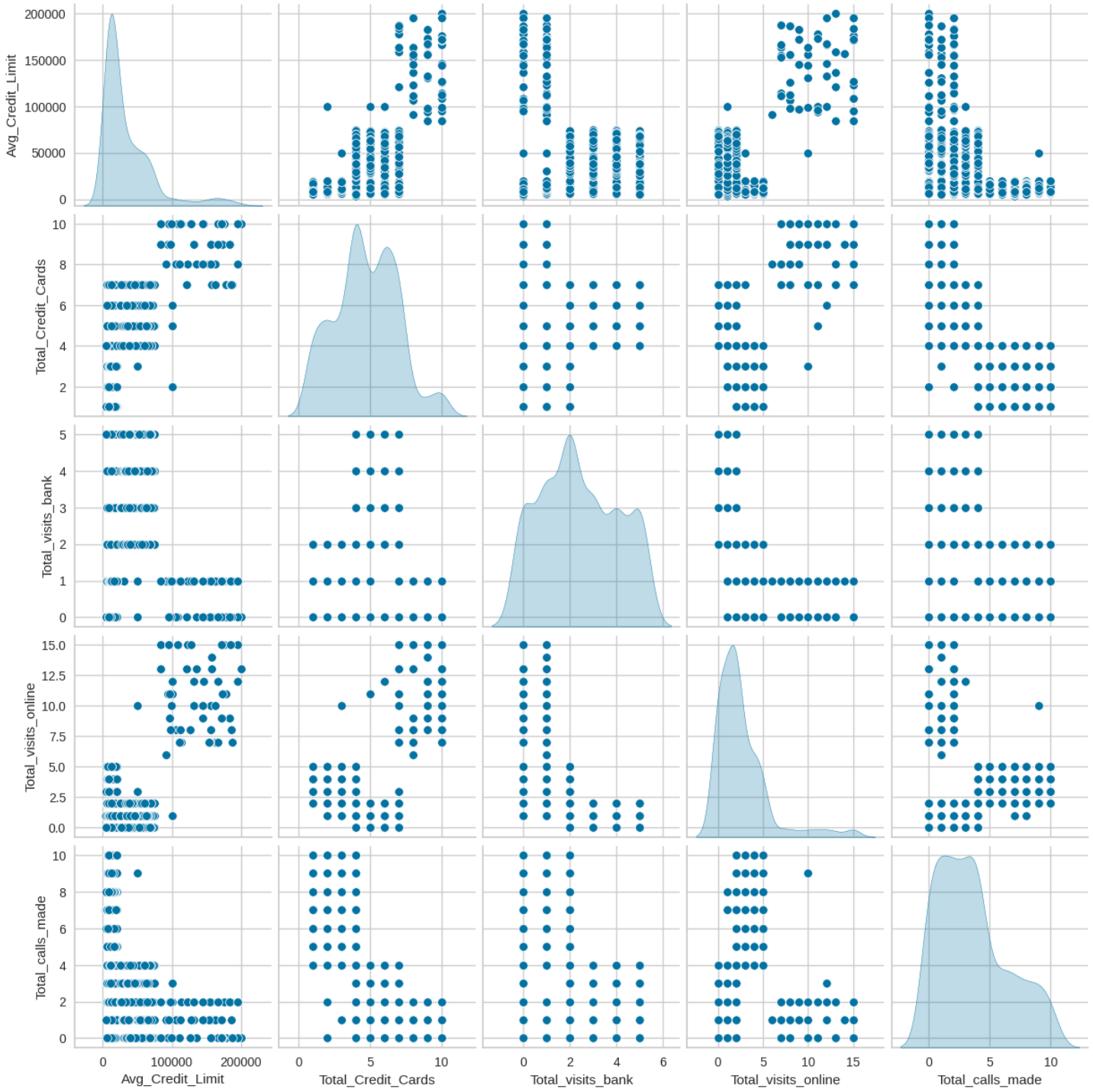


Figure 6: Pair Plot

- The credit limit is showing very high correlation with the total number of credit cards. More the number of credit card more is the limit. We also noticed that more number of credit card a customer has, he/she tends to visit online site more.

- The total number of credit cards is also positively correlated to number of total visits to bank. High number of credit cards is very highly negatively correlated to total calls made.
- There is a negative correlation between the number of physical visits and total online visits and total calls made. May be this is because physical visits resolve issues which is not followed up by online, phone calls or vice-versa.

7. Data Preprocessing

7.1. Outlier Detection and Treatment

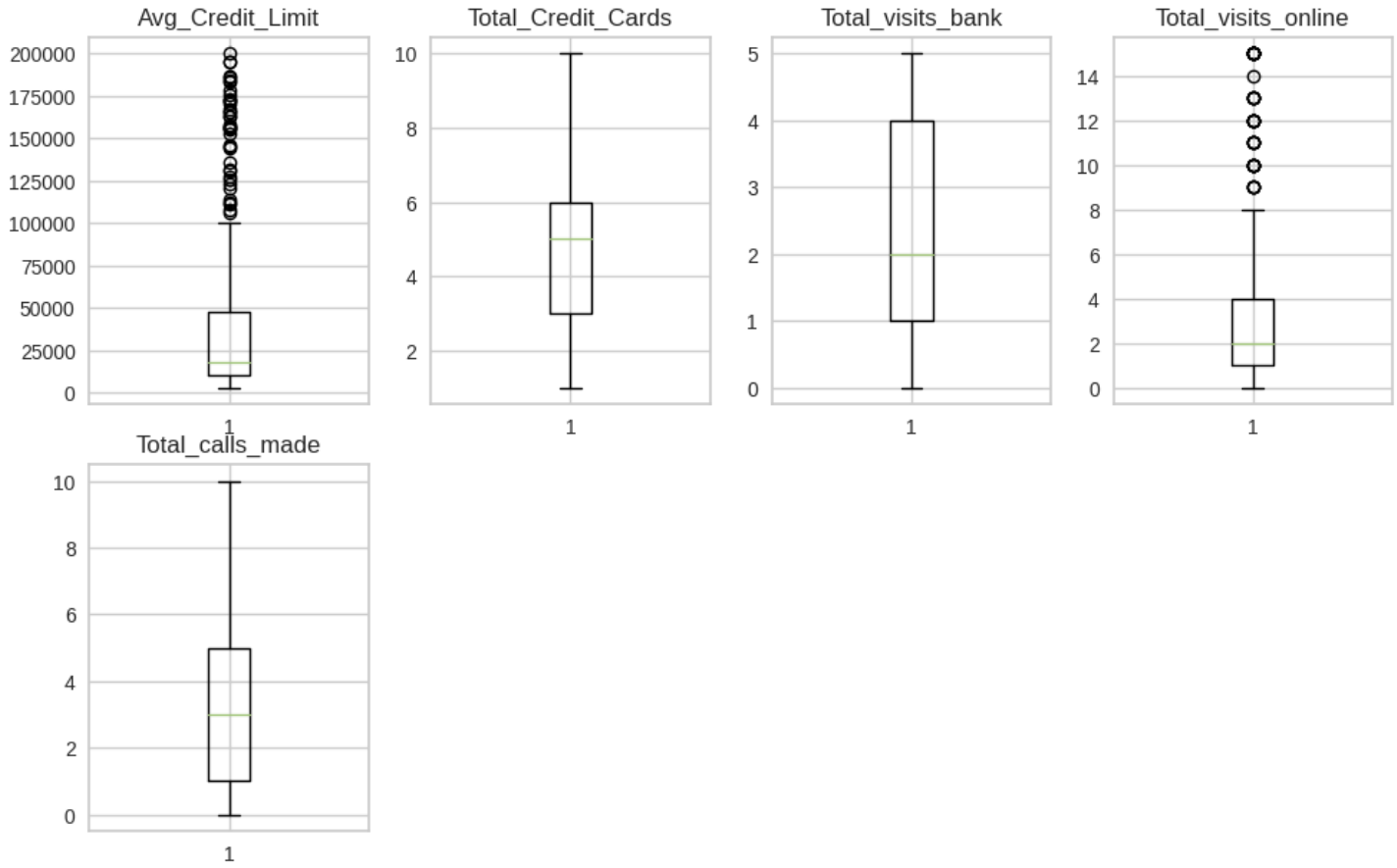


Figure 7: Outliers detection

Observation Since all the values are right we will prefer to not treat the outliers.

7.2. Feature Engineering

The data doesn't require any feature engineering. We will not use the below features as they will not have any say in analysis

Sl_No: Primary key of the records

Customer Key: Customer identification number

7.3. Data Scaling

We shall further scale the data using the standard scaler package in python as all the features might have different scale of measure. It is useful because.

1) Improves Model Performance

2) Improves Accuracy because

The scaling is done by using the Z-score normalization

8. K-Means Clustering

8.1. Apply K-means Clustering

First technique we shall use is the K-Means Clustering

.

K-means clustering is an unsupervised machine learning algorithm used for grouping data points into distinct clusters based on similarity. It divides the data into K predefined non overlapping clusters, where each data point belongs to the cluster with the nearest mean (centroid).

8.2. Plot the Elbow Curve

Distortion Values for different No. of clusters

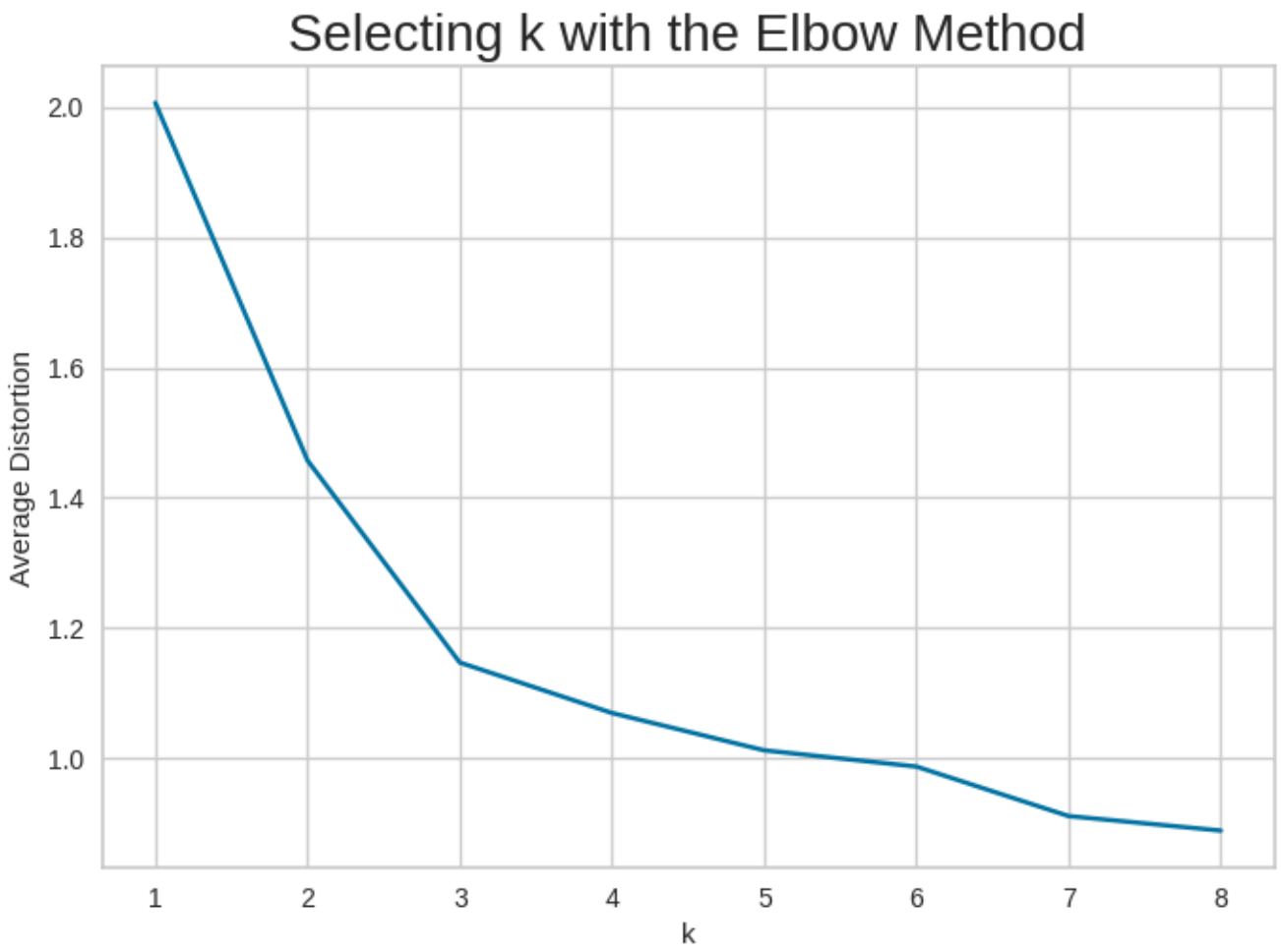


Figure 8: Select K with Elbow Method

- The appropriate value of k from the elbow curve seems to be 3 to 5.

8.3. Plot Silhouette Scores

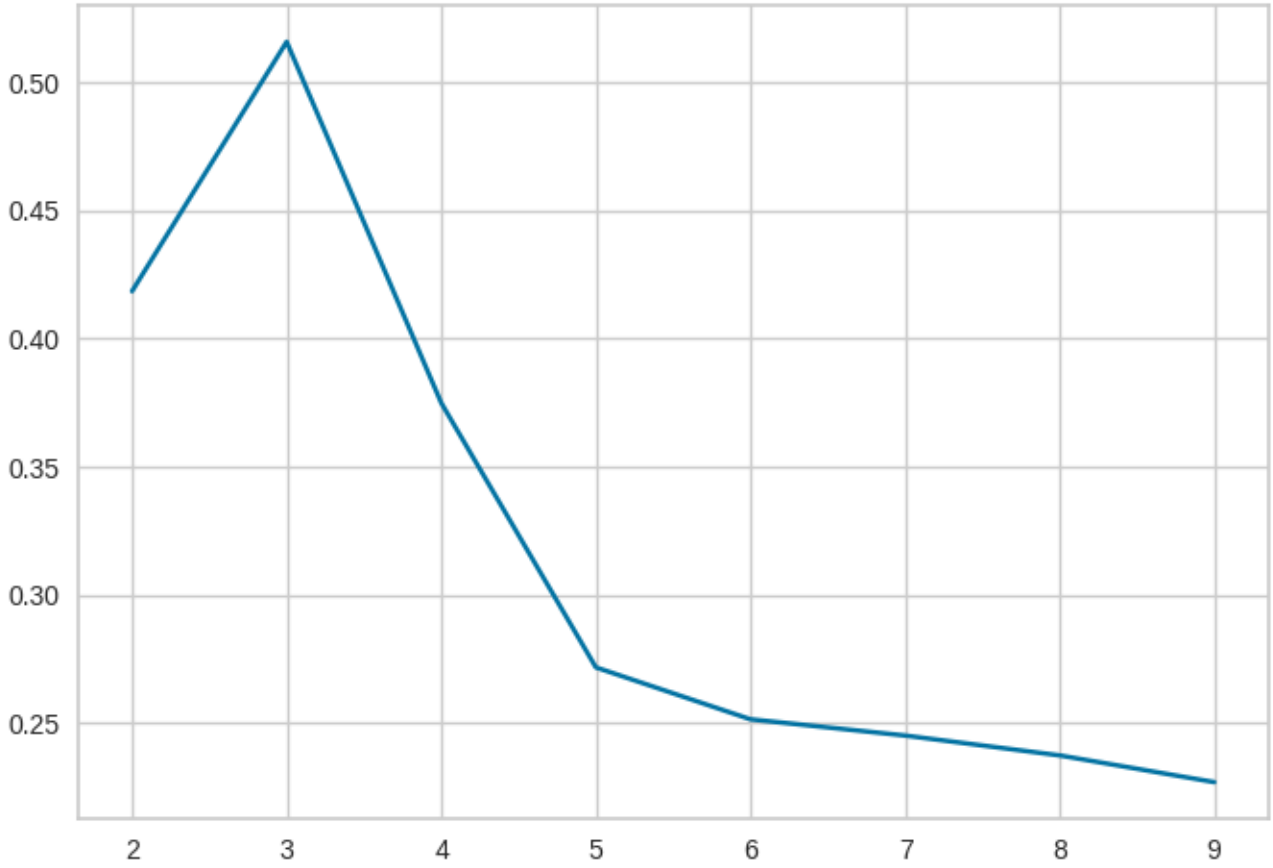


Figure 9:Silhouette Scores

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

2 Clustered

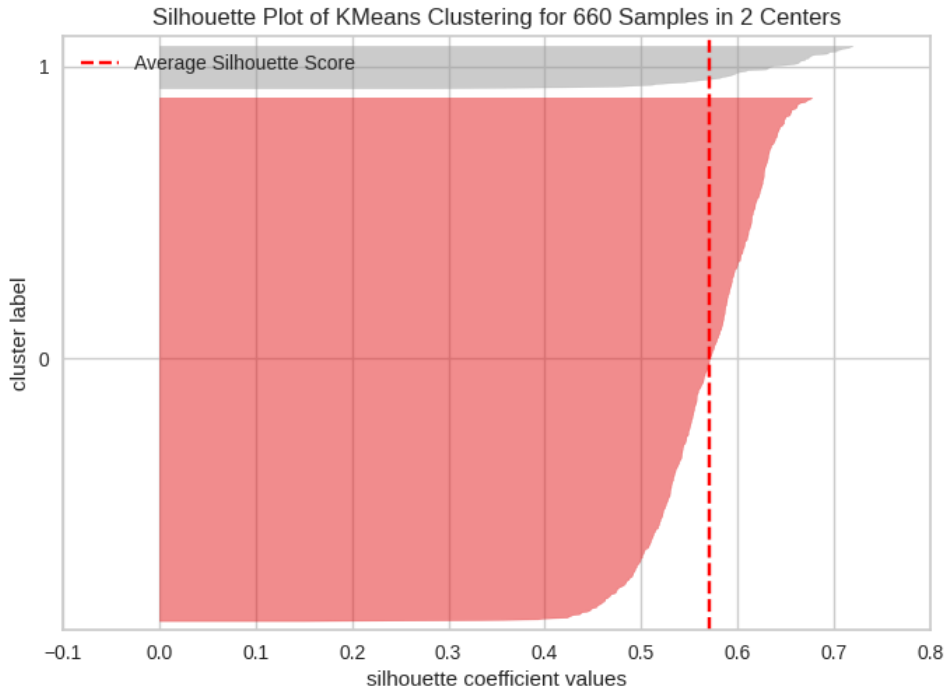


Figure 10: Silhouette coefficient (2 Clustered)

3 Clustered

Figure 11: Silhouette coefficient (3 Clustered)

4 Clustered

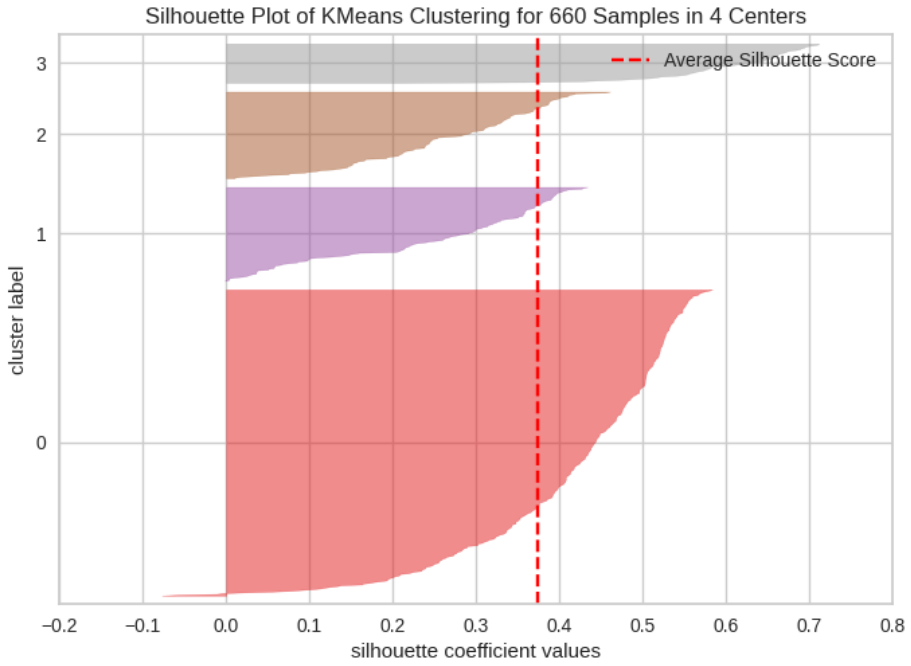


Figure 12: Silhouette coefficient (4 Clustered)

5 Clustered

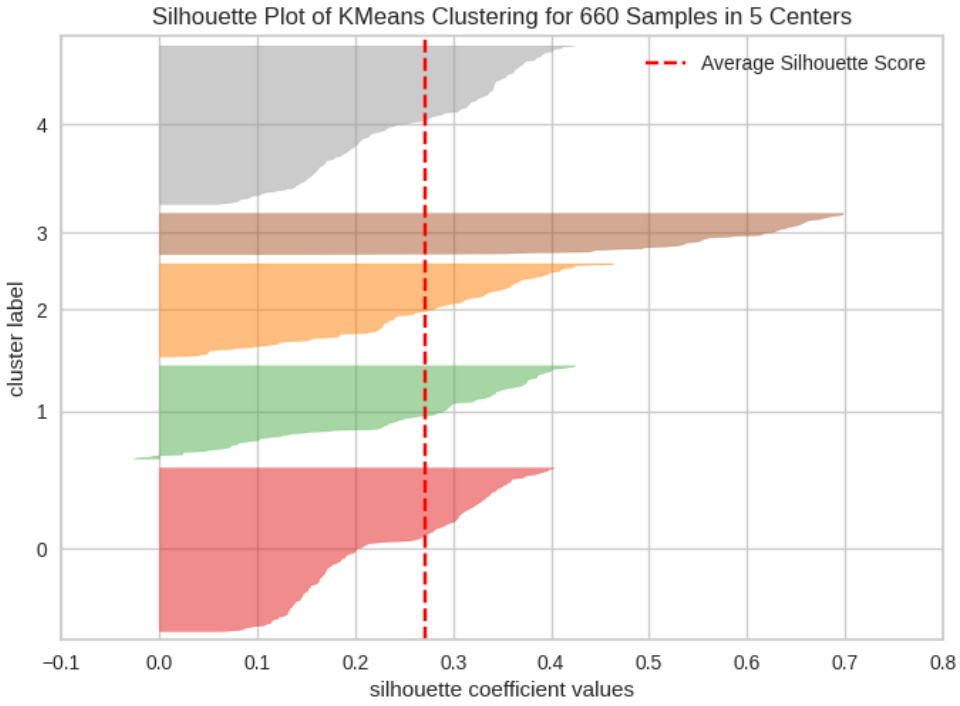


Figure 13: Silhouette coefficient (5 Clustered)

6 Clustered

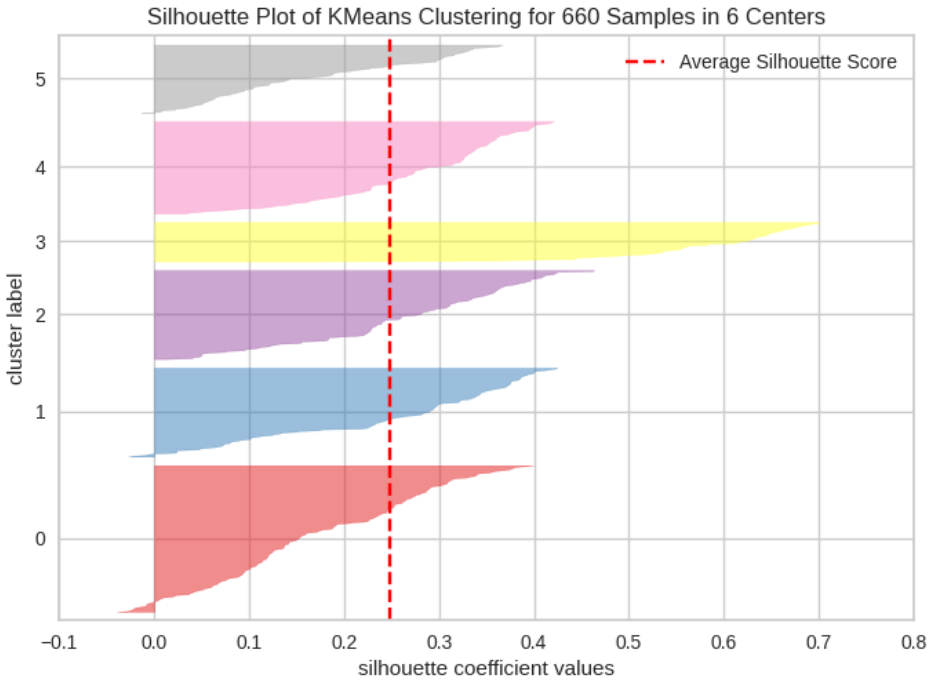


Figure 14: Silhouette coefficient (6 Clustered)

- Silhouette Coefficient is calculated for individual data points, and it tells you how well a single point fits into its cluster.
- Silhouette Score is the average of the Silhouette Coefficients for all points in a dataset, providing an overall evaluation of the clustering quality.
- We notice that 2 and 3 have the best Silhouette Coefficients but the 3 will be the apt no of coefficients as the density of points is well distributed and there is a nick in the plot.

8.4. Cluster profiling

	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
K_means_segments						
0	33782.38342	5.515544	3.489637	0.981865	2	386
1	12174.10714	2.410714	0.933036	3.553571	6.870536	224
2	141040	8.74	0.6	10.9	1.08	50

Table 3: Cluster profile

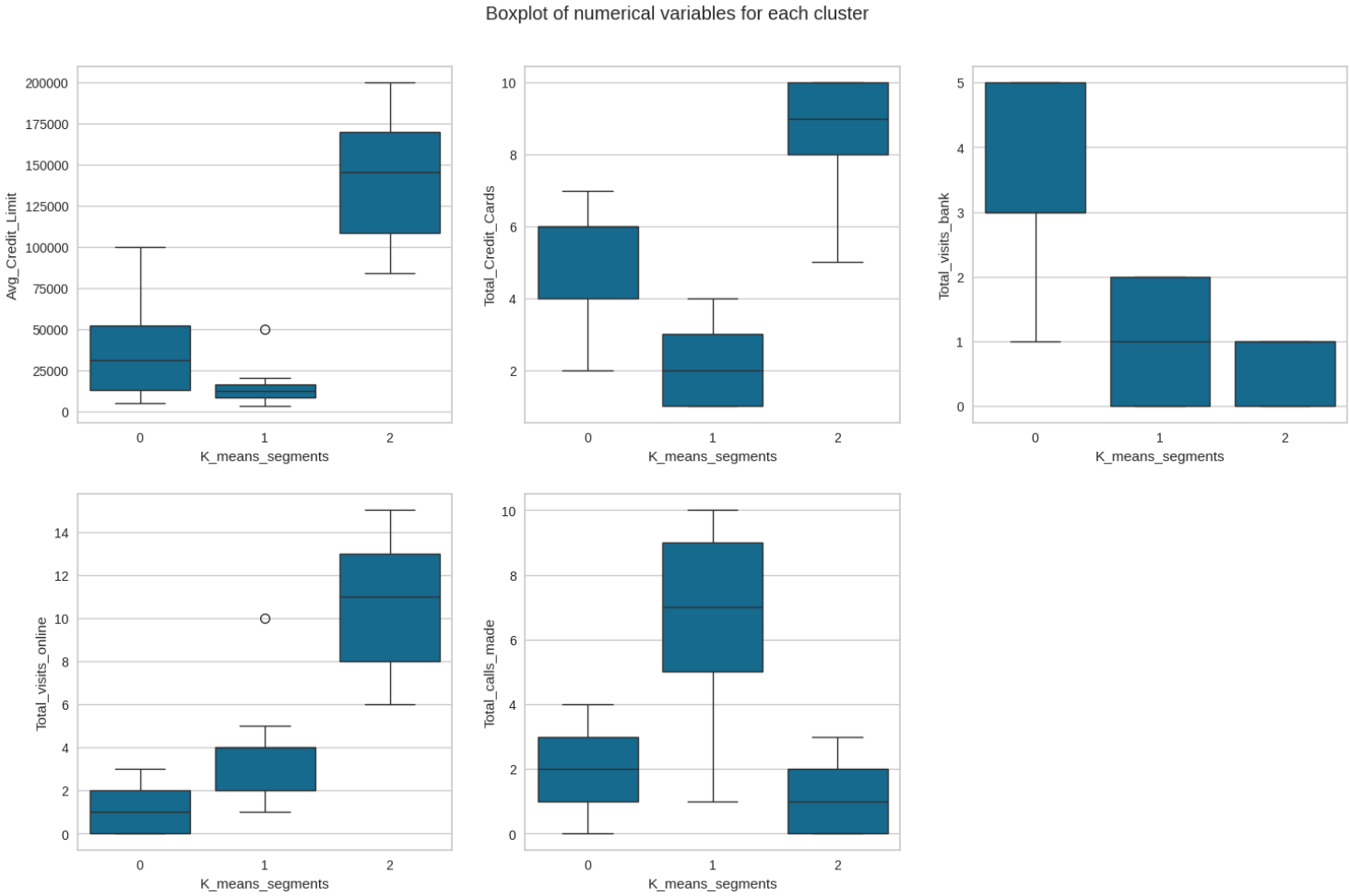


Figure 15: Boxplot of numerical variables for each cluster

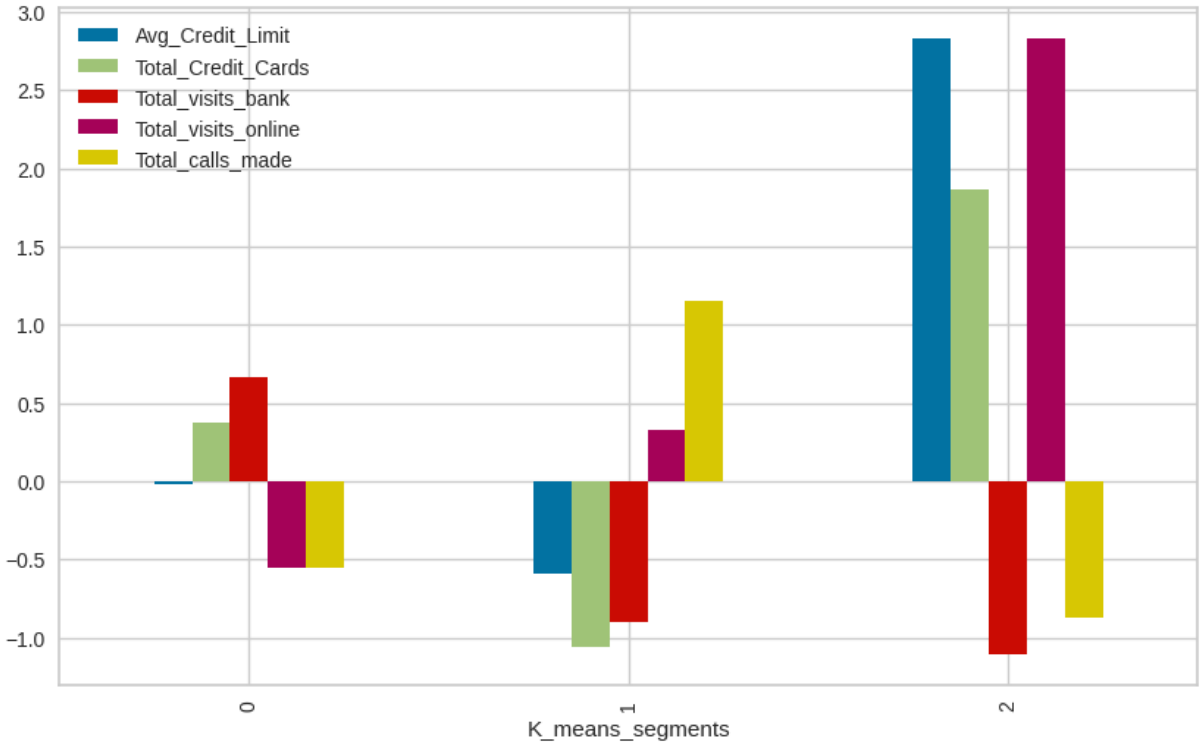


Figure 16: K means segments

Observation:

1) Avg_credtid_limit : As already discussed the cluster number 2 has the best credit limit between 90000 to 200000. CLuster 1 has the lowest which is in between minimum to 20000.

2) Total_credit_cards: The total number of credit cards is also highest in cluster 2 group. Cluster 1 has between 1-4 no of credit cards and cluster 0 has between 2-7.

3) Total_visits_bank : The total visits to the bank was very high in cluster 0. and the lowest in cluster 2. this means people with high credit facility don't prefer visiting banks

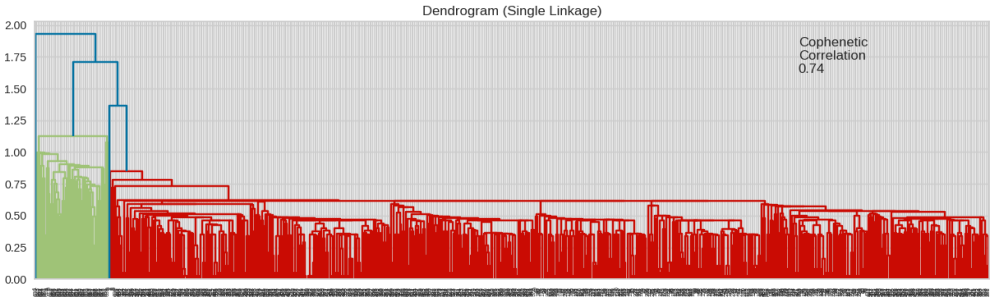
4) Total_visits_online : The case with the this parameter is completely in reverse to the total_visits_bank. We notice that the cluster 2 prefer visiting online more followed by cluster 1 and cluster 0,

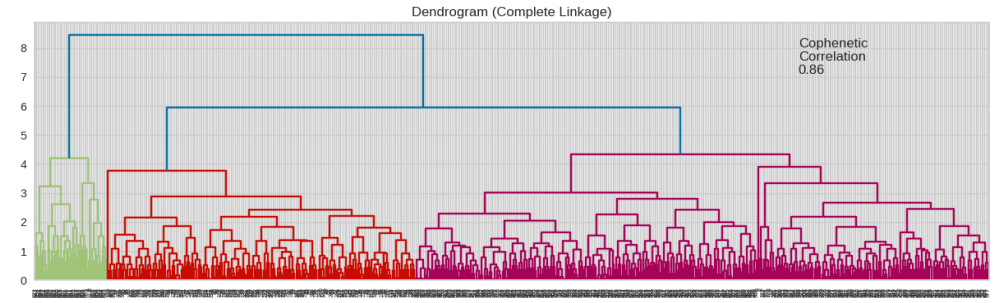
5) The_total_calls_made: the total calls made was highest in the customers who have a low credit limit. They made between 1 to 10 calls per annum.

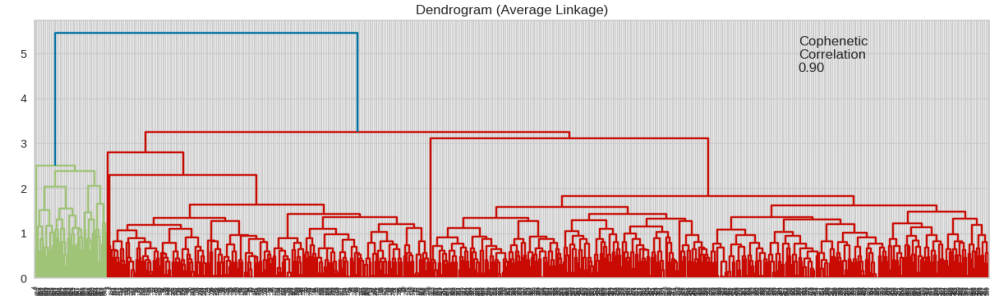
9. Hierarchical Clustering

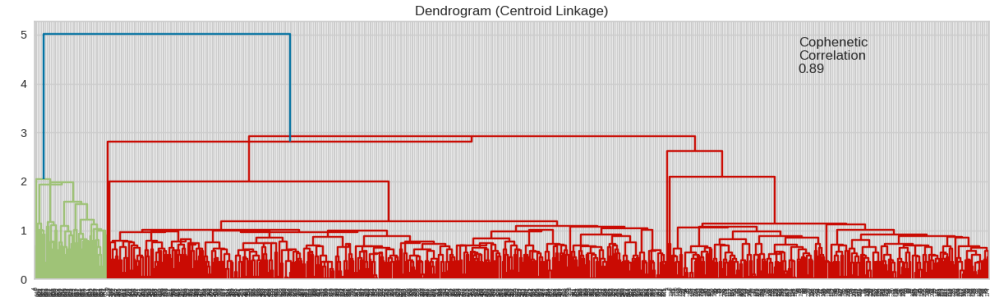
- Highest cophenetic correlation is 0.8977080867389372, which is obtained with Euclidean distance and average linkage.
- Highest cophenetic correlation is 0.8977080867389372, which is obtained with average linkage.

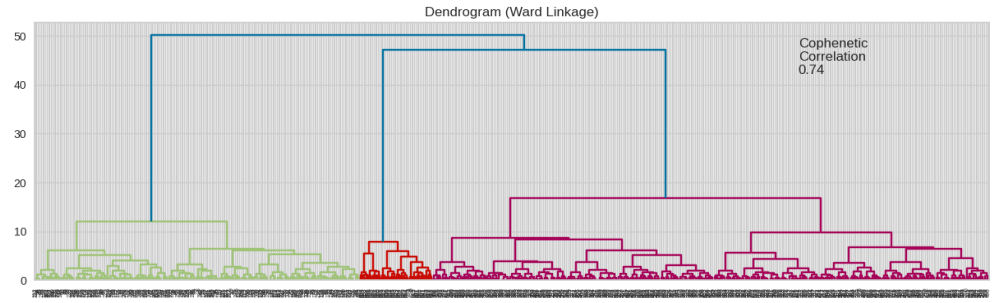
9.1. Checking Dendrograms


Dendrogram (Single Linkage)
Cophenetic Correlation 0.74


Dendrogram (Complete Linkage)
Cophenetic Correlation 0.86


Dendrogram (Average Linkage)
Cophenetic Correlation 0.90


Dendrogram (Centroid Linkage)
Cophenetic Correlation 0.89


Dendrogram (Ward Linkage)
Cophenetic Correlation 0.74

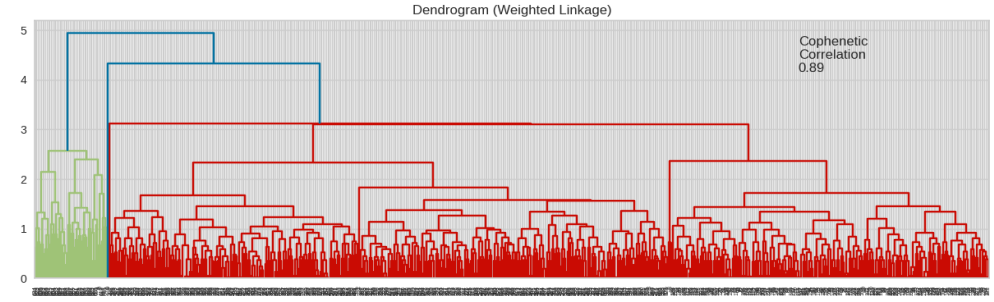

Dendrogram (Weighted Linkage)
Cophenetic Correlation 0.89

Figure 17: Dendrogram

Observations

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

9.2. Creating Model Using SKlearn

After segmenting into 3 clusters, we found the below number of customers per Cluster in 0, 1 and 2 are as below:

	Avg Credit Limit	Total Credit Cards	Total visits bank	Total visits online	Total calls made	K means segments	Count in each segments
HC Clusters							
0	33713.1783	5.511628	3.485788	0.984496	2.005168	0.002584	387
1	141040	8.74	0.6	10.9	1.08	2	50
2	12197.30942	2.403587	0.928251	3.560538	6.883408	1	223

Figure 18: Cluster profile Average

Observation

There are three clusters which have been separated.

Cluster 0: The cluster has been made based on the grouping data points which are based on highest average total visits.
Cluster 1: This cluster is made based on the average highest credit limit, credit cards.

Cluster 2: The last cluster is made from the total highest average visits online and total calls made.

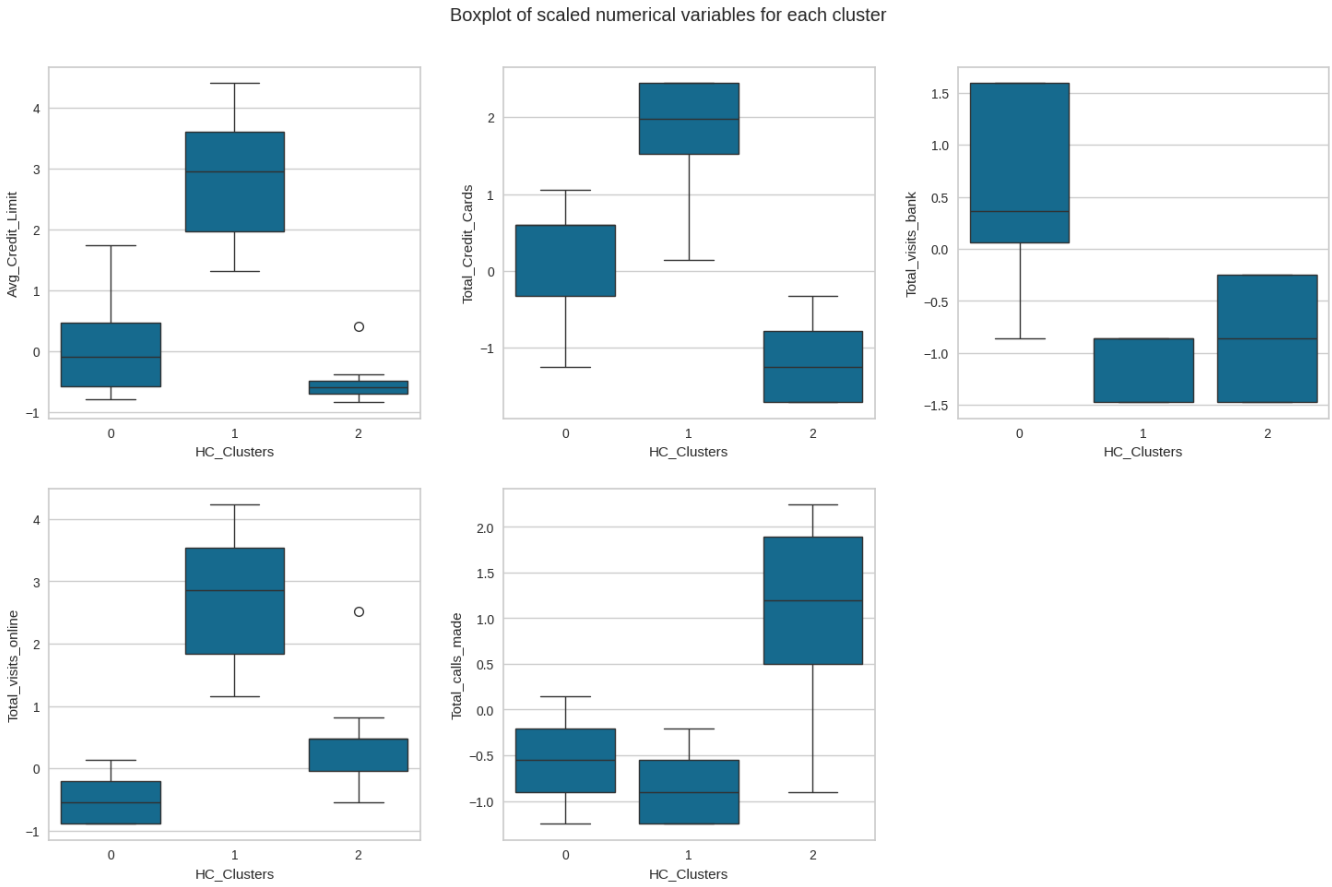


Figure 19:Boxplot 2 of scaled numerical variables for each cluster

Observation on Average Cluster:

1) Avg_credtid_limit : As already discussed the cluster number 1 has the best credit limit between 90000 to 200000. CLuster 2 has the lowest which is in between minimum to 20000.

2) Total_credit_cards: The total number of credit cards is also highest in cluster 1 group. Cluster 2 has between 1-4 no of credit cards and cluster 0 has between 2-7.

3) Total_visits_bank : The total visits to the bank was very high in cluster 0. and the lowest in cluster 1. this means people with high credit facility don't prefer visiting banks

4) Total_visits_online : The case with the this parameter is completely in reverse to the total_visits_bank. We notice that the cluster 1 prefer visiting online more followed by cluster 1 and cluster 0,

5) The_total_calls_made: the total calls made was highest in the customers who have a low credit limit which is cluster 2 . They made between 1 to 10 calls per annum.

Observation

To bring in some more variability we can try using Ward linkage method. Even if the cophenetic correlation is high we can check if ward method is clustering the data differently or not. Hence as per the dendrogram let's try with 4 clusters in Ward Linkage method as the clusters are well separated.

	Avg Credit Limit	Total Credit Cards	Total visits bank	Total visits online	Total calls made	K means segments	count in each segments
HC Clusters							
0	12197.30942	2.403587	0.928251	3.560538	6.883408	1	223
1	37520.40816	5.642857	2.52551	0.97449	2.142857	0.005102	196
2	141040	8.74	0.6	10.9	1.08	2	50
3	29806.28272	5.376963	4.471204	0.994764	1.863874	0	191

Figure 20: Cluster profile ward linkage

Observation

We observe that there doesn't seem to be much variability except

Cluster 0 : The average calls made is the highest in this group.

Cluster 1 : the cluster 1 is not having anything which is high in any parameter. we can consider that this set of customer are having less avg number of credit cards which is 5 and visit almost equally more the bank.

Cluster 2 : This is the cream customer of the bank which have high credit limit and highest avg credit cards. They also tend to visit online more.

Cluster 3 : The average total visit is the highest amongst the customer base.

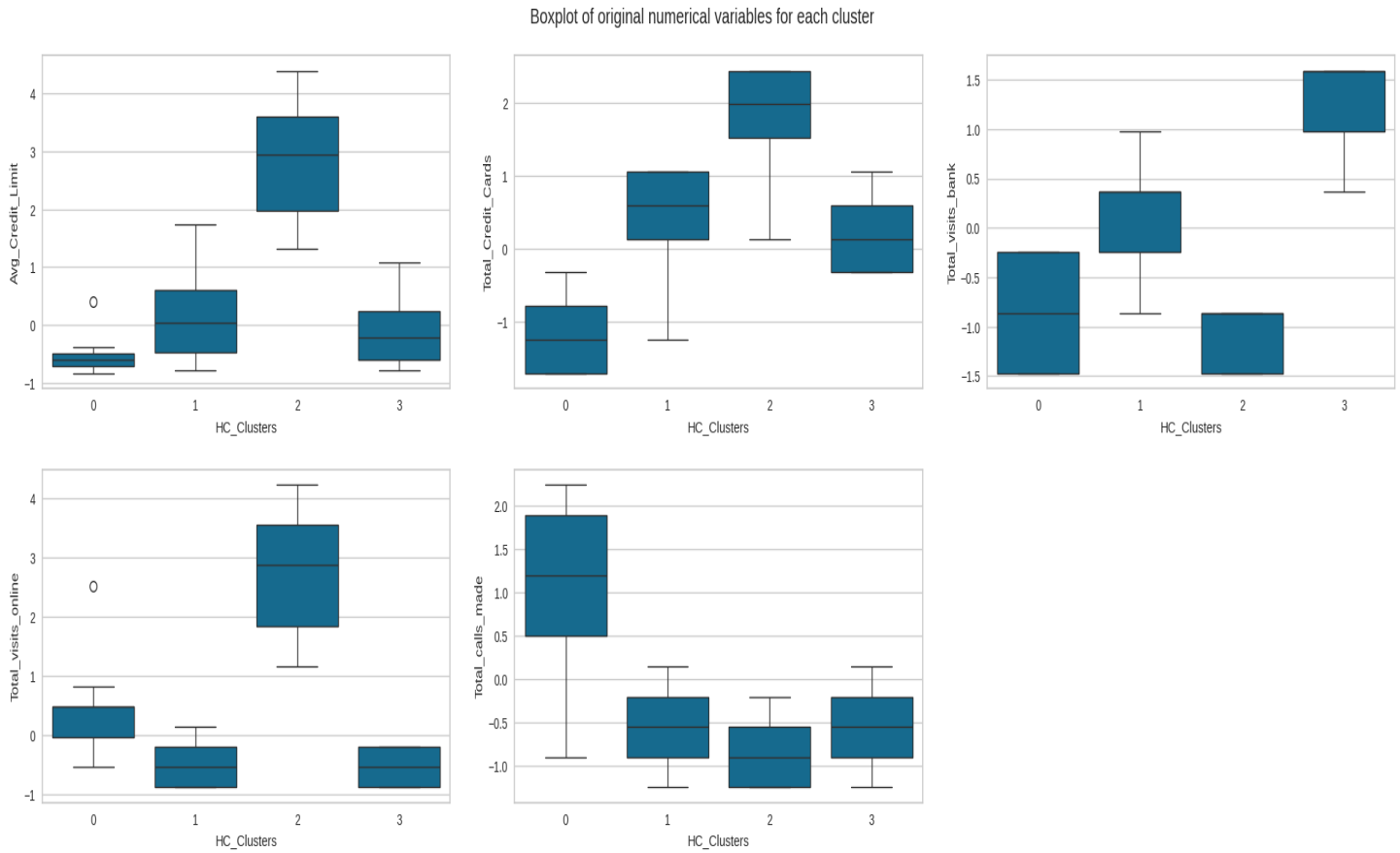


Figure 21: Boxplot 3 of scaled numerical variables for each cluster

Observation

We observe the below

Avg_Credit Limit : As we can observe the average credit limit is the highest in the cluster 2 group and lowest in cluster 0

Total credit cards : The total credit cards also is the highest in the cluster 2 group

Total visits to bank : the cluster 3 and 1 are having a good number of visits to the bank every annum.

Total visits online : Customers who have high credit limit and more number of cards tend to visit online more.

Total calls made : the total calls made is the highest in the cluster 0 group and lowest in the cluster 2 group.

Observation on Ward Linkage

We notice that Ward linkage is not able to cluster the data better than the other methods so we can conclude that average linkage with 3 clusters is the best option.

10. K-means Vs Hierarchical Clustering (Average and Ward linkage)

K-Means segments

	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
K_means_segments						
0	33782.38342	5.515544	3.489637	0.981865	2	386
1	12174.10714	2.410714	0.933036	3.553571	6.870536	224
2	141040	8.74	0.6	10.9	1.08	50

Table 4: Cluster profile K means

Hierarchical Clustering

	Avg Credit Limit	Total Credit Cards	Total visits bank	Total visits online	Total calls made	K means segments	count in each segments
HC Clusters Average							
0	33713.1783	5.511628	3.485788	0.984496	2.005168	0.002584	387
1	141040	8.74	0.6	10.9	1.08	2	50
2	12197.30942	2.403587	0.928251	3.560538	6.883408	1	223
HC Clusters Ward Linkage							
0	12197.30942	2.403587	0.928251	3.560538	6.883408	1	223
1	37520.40816	5.642857	2.52551	0.97449	2.142857	0.005102	196
2	141040	8.74	0.6	10.9	1.08	2	50
3	29806.28272	5.376963	4.471204	0.994764	1.863874	0	191

Table 5: Cluster profile Hierarchical Clustering

Observation on K Means Vs Hierarchical Clustering:

If both Hierarchical and K-Means clustering give the same kind of results, it indicates that the data has a strong, inherent structure that both algorithms can detect. It shows that the clustering of data is well defined in the dataset.

We tried average and Wards linkage in the Hierarchical Clustering but there doesn't seem to be, much difference in clustering pattern.

11. Actionable Insights & Recommendations

The data analysis of the dataset has given is pretty decent insights to drive marketing.

1) People who visit bank offline for queries can be guided to visit their website to get solution thus increasing the overall efficiency of bank. We also notice that people who visit offline have enough number of credit cards but don't have higher credit limit. With correct investigation bank can offer higher credit limit to those customers.

2) Customers who visit online tend to have higher avg credit limit. This is very critical information. New campaigns, offers which need to be targeted on high-net-worth customers can be done online.

3) The customers who had the lowest credit limit are also the ones who make phone calls to the bank the most. We can ask those customers to visit bank to educate about credit facility or as them to visit online for information. The customers who visit bank or go online might avail more cards thus increasing the revenue of bank.