AllLife Bank

CREDIT CARD CUSTOMER SEGMENTATION

September 15th, 2023

Contents / Agenda

- Executive Summary
- Business Problem Overview and Solution Approach
- EDA Results
- Data Preprocessing
- K-Means Clustering
- Hierarchical Clustering
- Appendix

Executive Summary

Insights

We found **3** segments of different sizes:

The largest group is segment 0 and has a moderate credit limit, makes the most visits to the bank, has the fewest online visits, a moderate number of credit cards, and a moderate number of calls to the bank.

Group 1 being the smallest of the three. It has the highest credit limit and the most online visits. It also has the fewest visits and calls to the bank.

Group 2 has few credit cards, makes the most frequent calls, and has a low credit limit

Executive Summary

Marketing Strategy

Cluster 0 and Cluster 2 represent opportunities for upselling higher credit limit cards, thru more and better credit cards and promoting online services.

Cluster 1 customers, having the highest average credit limits, could be targeted with exclusive offers and privileges to enhance loyalty.

Service Delivery Strategy

Cluster 0: Focus on enhancing in-bank experiences as these customers frequently visit the bank.

Cluster 1: Invest in improving the online platform, considering their high usage. Unlock benefits for using online platforms to promote their use.

Cluster 2: Strengthen the call center services to cater to these customers who prefer reaching out through calls.

Executive Summary

Outlier Handling

In the preliminary stages of our data analysis, we identified a subset of customers who significantly surpassed the average credit limit norm, categorizing them as outliers. However, upon meticulous consideration, we recognized the latent potential in this unique customer segment, choosing to not treat these outliers in the conventional analytical manner.

Rationale

- High Value, High Potential: This segment, characterized by a considerably higher credit limit, naturally gravitates towards a premium category, representing a high-value, high-potential customer base.
- Customized Services: By delineating this group, we pave the way for bespoke services and products, essentially offering a ‘concierge’ service line, designed to cater to their sophisticated financial needs and preferences.
- Business Growth: Cultivating a specialized approach towards this segment can potentially fuel business growth, fostering a relationship marked by exclusivity and premium service delivery.

Executive Summary

The Strategy

- Sub-Clustering: We propose to create a sub-cluster, essentially a ‘Super Premium’ category, where we delve deeper to understand their behavioral patterns, preferences, and financial portfolio to design a product and service line that mirrors their expectations.

- Personalized Campaigns: Leveraging data analytics, we envisage crafting personalized marketing campaigns, offering products that resonate with their financial stature, and services that echo their preferences, thus enhancing customer satisfaction and brand loyalty.

- Feedback Loop: Establishing a continuous feedback loop with this segment to understand their evolving needs and preferences, thus ensuring a dynamic and responsive service framework.

By choosing to not treat these outliers, we are not only preserving the integrity of our data but also unlocking a vista of opportunities, eyeing a trajectory of growth marked by premium service delivery and customer satisfaction.

This approach fosters a strategy rooted in personalization and exclusivity, aiming to carve a niche in the competitive market landscape not to mention the creation of an aspirational status that people that do not belong to the segment may start gravitating towards to thus making the outlier data the start of a pattern and a new market that can be beneficial to the whole customer base by being part of a company that caters this level of service and the customer can access that status.

Business Problem Overview and Solution Approach

Problem

AllLife Bank intends to focus on it's credit card customers in the upcoming financial year. The market research team has advised that market penetration can be improved. Based on this, the marketing team proposes a personalized campaign targeting both new and existing customers. It was found in the market research that customers perceive support services as inadequate. Consequently, the operations team aims to enhance the service delivery model to ensure quicker resolution of customer queries and requests. The Director of Marketing and the Director of Service Delivery have decided to turn to the Data Science team for assistance.

Objective

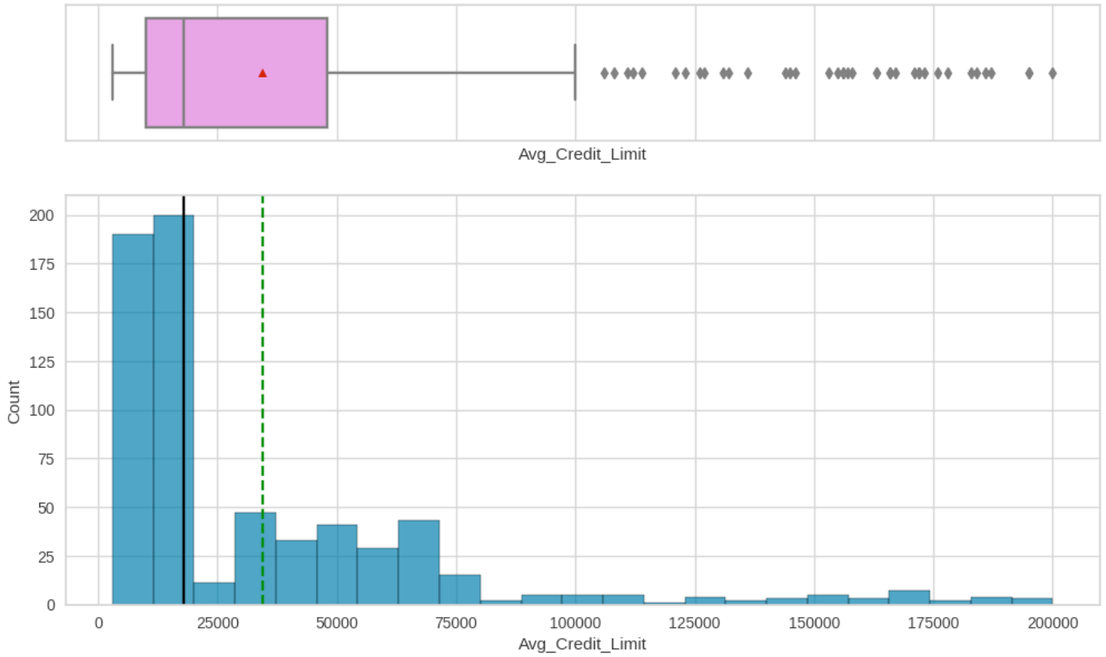
The objective is to identify different segments among existing consumers based on spending patterns and past interactions with the bank using clustering algorithms and provide recommendations to the bank on how to market to and serve these consumers in the market.

EDA Results Univariate Analysis

Average Credit Limit

	count	mean	std	min	25%	50%	75%	max
Avg_Credit_Limit	660.0	34574.242424	37625.487804	3000.0	10000.0	18000.0	48000.0	200000.0

The mean average credit limit is considerably higher than the median, reinforcing the right-skewed distribution observed in the histogram.



The majority of customers have a credit limit below 50,000, with a steep decline as the credit limit increases, indicating a right-skewed distribution.

We see outliers in the boxplot on a credit limit starting at 108,000

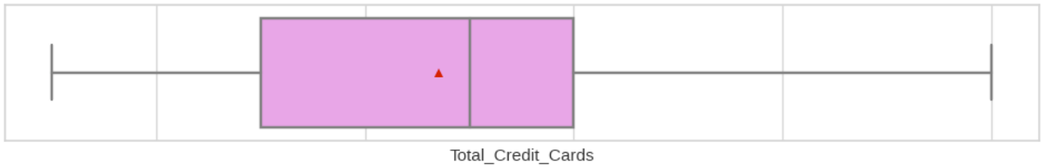
LINK TO APPENDIX SLIDE ON DATA BACKGROUND CHECK

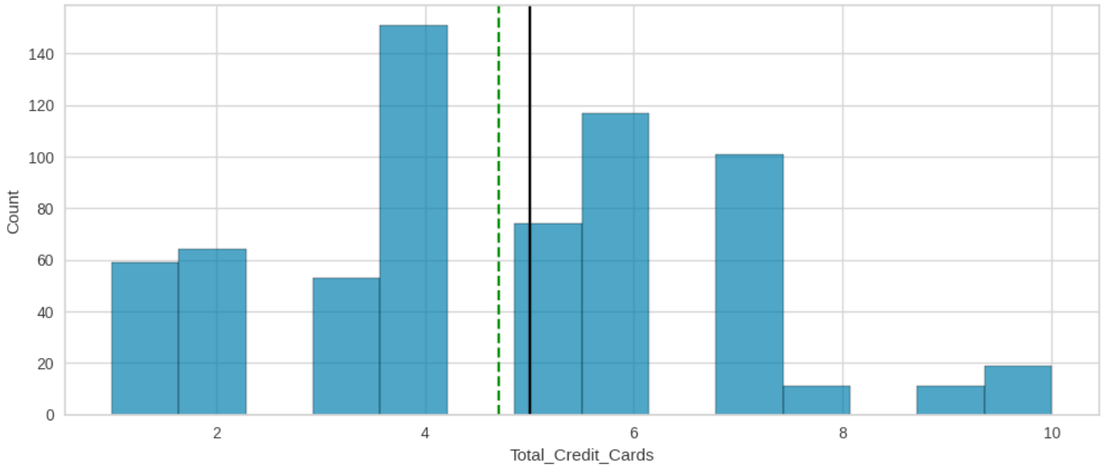
EDA Results Univariate Analysis

Total Credit Cards

	count	mean	std	min	25%	50%	75%	max
Total_Credit_Cards	660.0	4.706061	2.167835	1.0	3.0	5.0	6.0	10.0

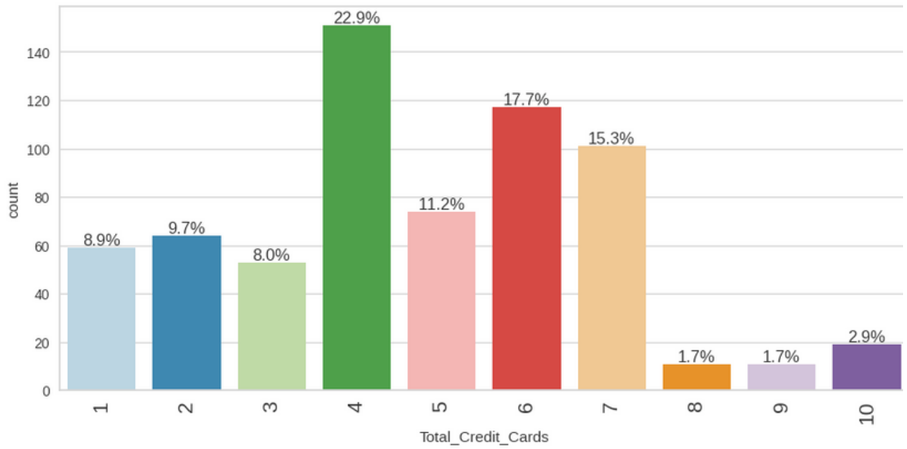
The average number of credit cards held is near 5, which is also the median value, suggesting a relatively symmetrical distribution around the mean.





Customers most commonly have between 4 and 7 credit cards.

We see no outliers in this data.

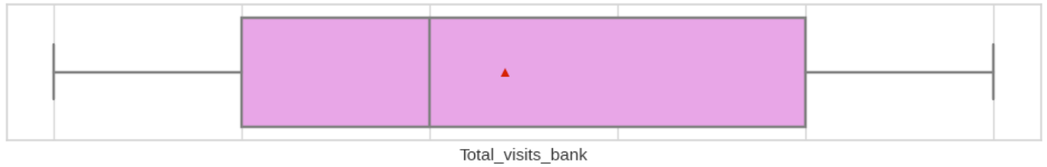


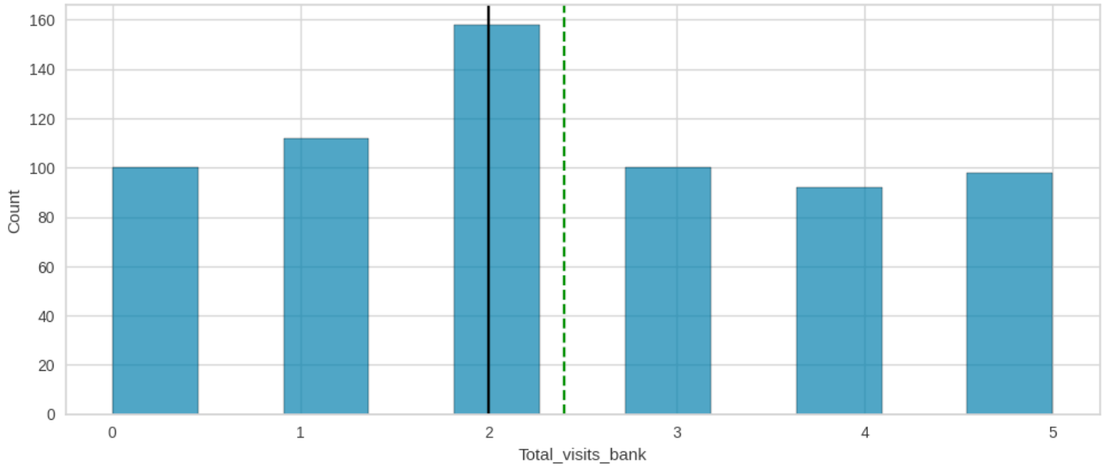
EDA Results Univariate Analysis

Total Visits to Bank

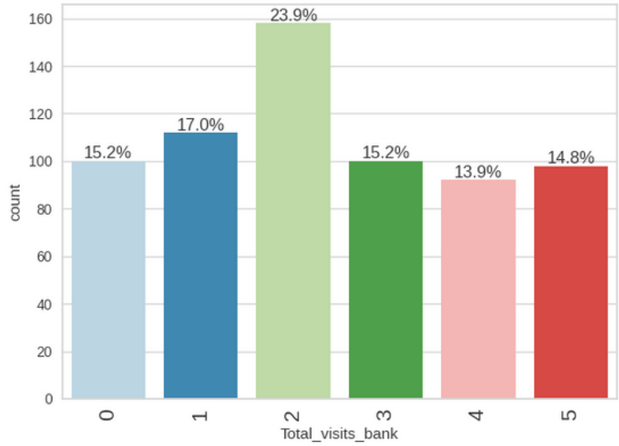
	count	mean	std	min	25%	50%	75%	max
Total_visits_bank	660.0	2.403030	1.631813	0.0	1.0	2.0	4.0	5.0

Customers visit the bank on average about 2.4 times a year, with a median of 2 visits.





A significant number of customers visit the bank between 0 and 2 times a year. However, there is also a considerable number who visit 3-5 times.



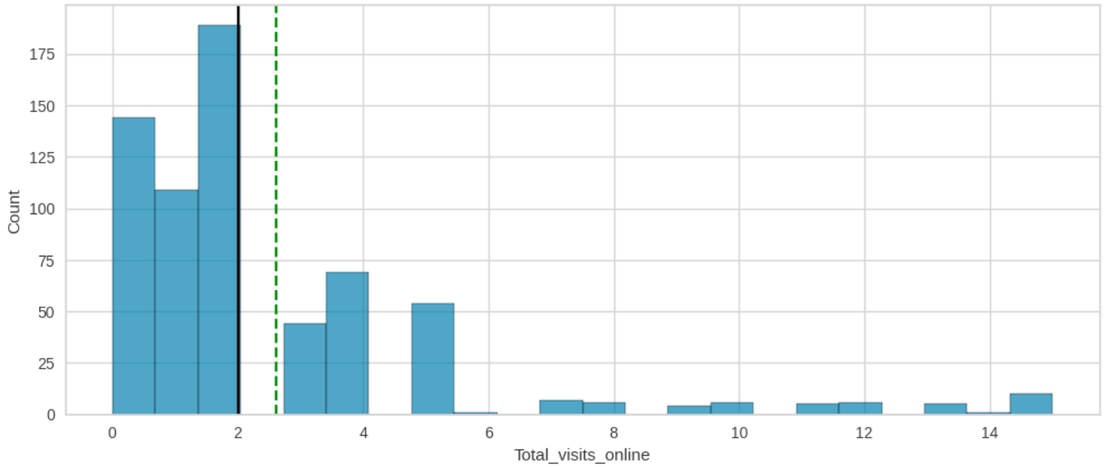
We see no outliers in this data.

EDA Results Univariate Analysis

Total Visits Online

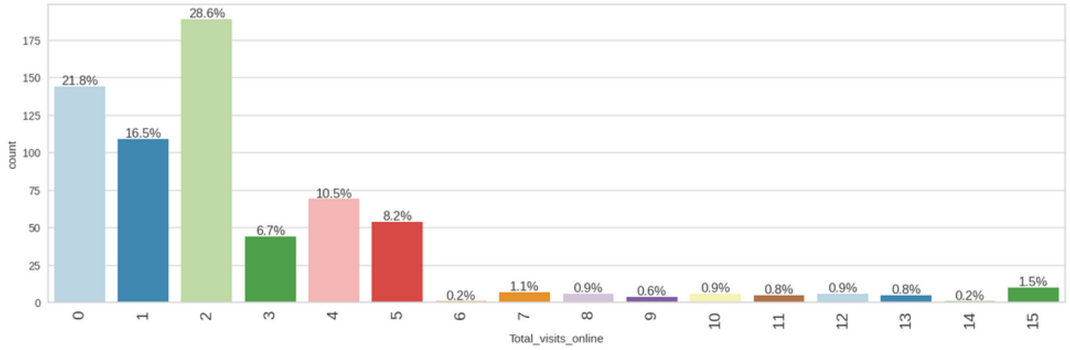
	count	mean	std	min	25%	50%	75%	max
Total_visits_online	660.0	2.606061	2.935724	0.0	1.0	2.0	4.0	15.0

The average number of online visits is higher than the median, indicating a right-skewed distribution.



Many customers have less than 2 online visits a year. We observe a gradual decrease as the number of visits increases, indicating a right-skewed distribution.

We see outliers in the boxplot beyond 9 online visits.

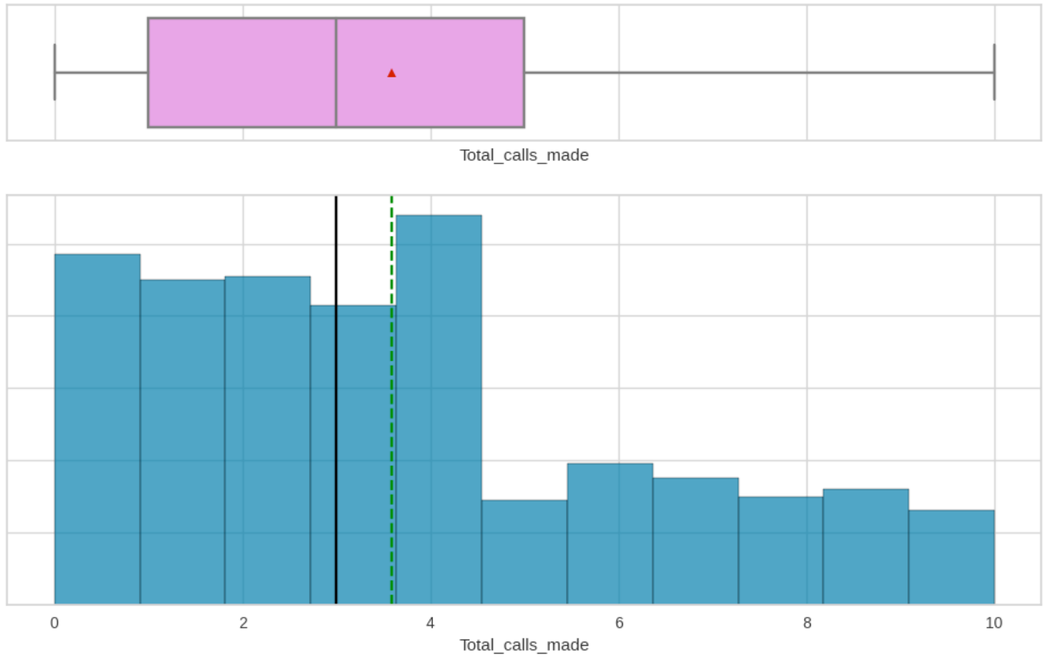


EDA Results Univariate Analysis

Total Calls Made

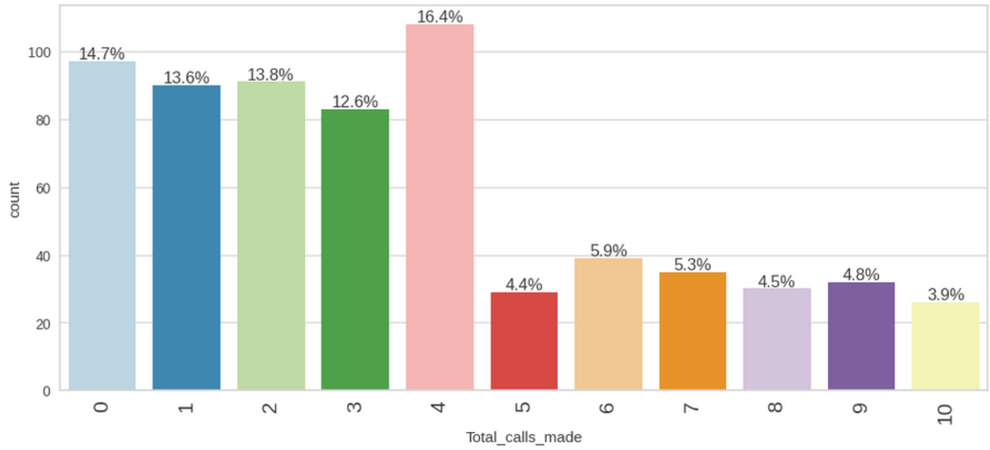
	count	mean	std	min	25%	50%	75%	max
Total_calls_made	660.0	3.583333	2.865317	0.0	1.0	3.0	5.0	10.0

On average, customers made around 3.6 calls a year, with a median of 3 calls.



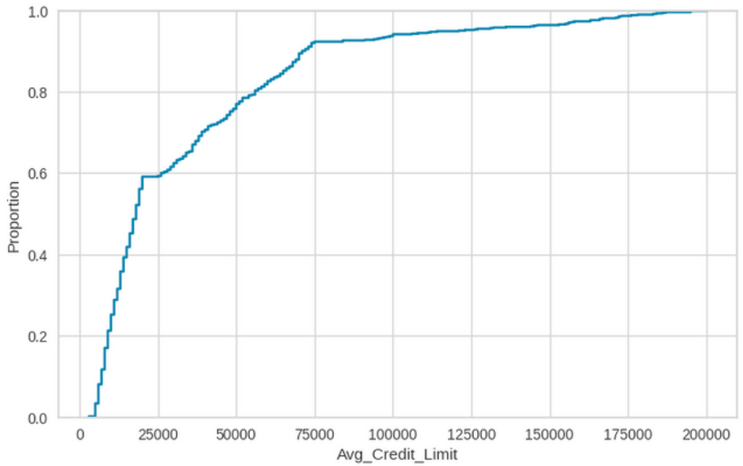
The distribution shows that a high number of customers make between 0 and 4 calls a year.

We see no outliers in this data.

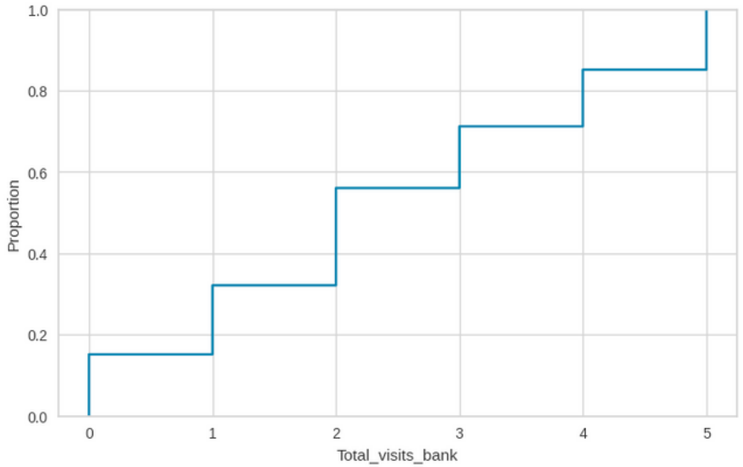


EDA Results Univariate Analysis

Cumulative Distribution Function (CDF)



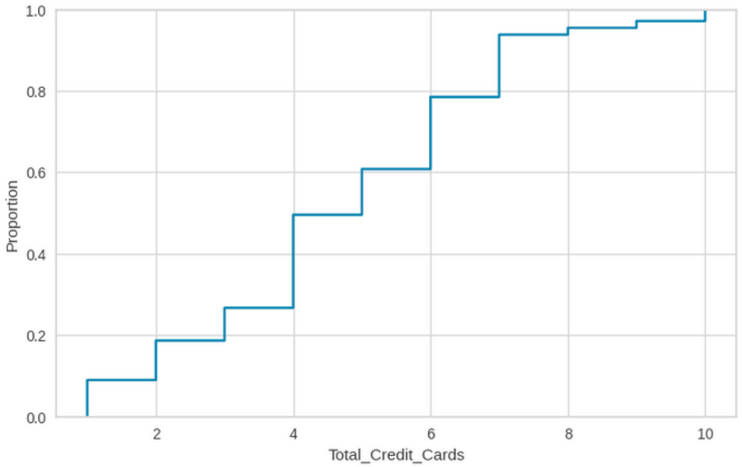
- About 80% of customers have a credit limit below 50,000, showcasing a significant portion with a moderate to low credit limit.
- The curve steepens significantly beyond the 100,000 mark, indicating the presence of outliers with exceptionally high credit limits.



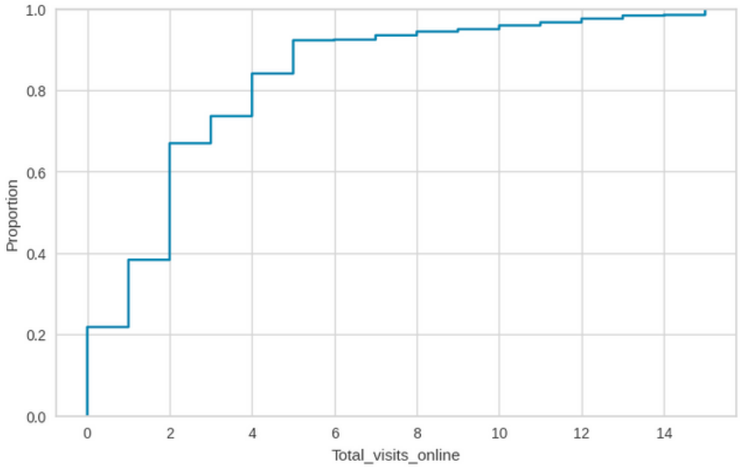
- More than 55% of customers visited the bank 2 or less a year.
- Nearly 80% visited the bank 5 or fewer times a year, indicating infrequent bank visits by a majority of the customers.

EDA Results Univariate Analysis

Cumulative Distribution Function (CDF)



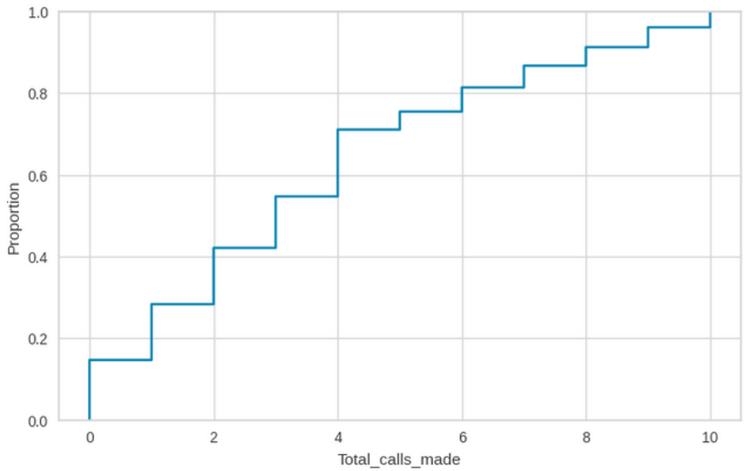
- Nearly 20% of customers hold 3 or fewer credit cards.
- Around 50% of customers have 4 or less credit cards, representing a majority with a moderate number of cards.



- More than 50% of customers made less than 2 online visits per year, illustrating a low preference for online banking.
- A small fraction of customers exhibit a very high frequency of online visits, which is reflected in the steep slope towards the end of the curve.

EDA Results Univariate Analysis

Cumulative Distribution Function (CDF)



- About 25% of customers made 1 or fewer calls per year to the bank.
- Around 90% made up to 7 calls a year, indicating a moderate call frequency with the bank.

EDA Results Bivariate Analysis

Correlation Matrix

- Average Credit Limit has a strong positive correlation with Total Credit Cards (0.61) and Total Visits Online (0.55), and a moderate negative correlation with Total Calls Made (-0.41) suggesting that customers with higher credit limits tend to have more credit cards and visit the bank's website more frequently but call significantly less.
- Total Visits Bank has a moderate negative correlation with Total Visits Online (-0.55) and Total Calls Made (-0.51), indicating that customers who visit the bank more often tend to make fewer calls and online visits.
- Total Credit Cards has a significant negative correlation with Total Calls Made (-0.65), suggesting that customers with more credit cards tend to make fewer calls to the bank.

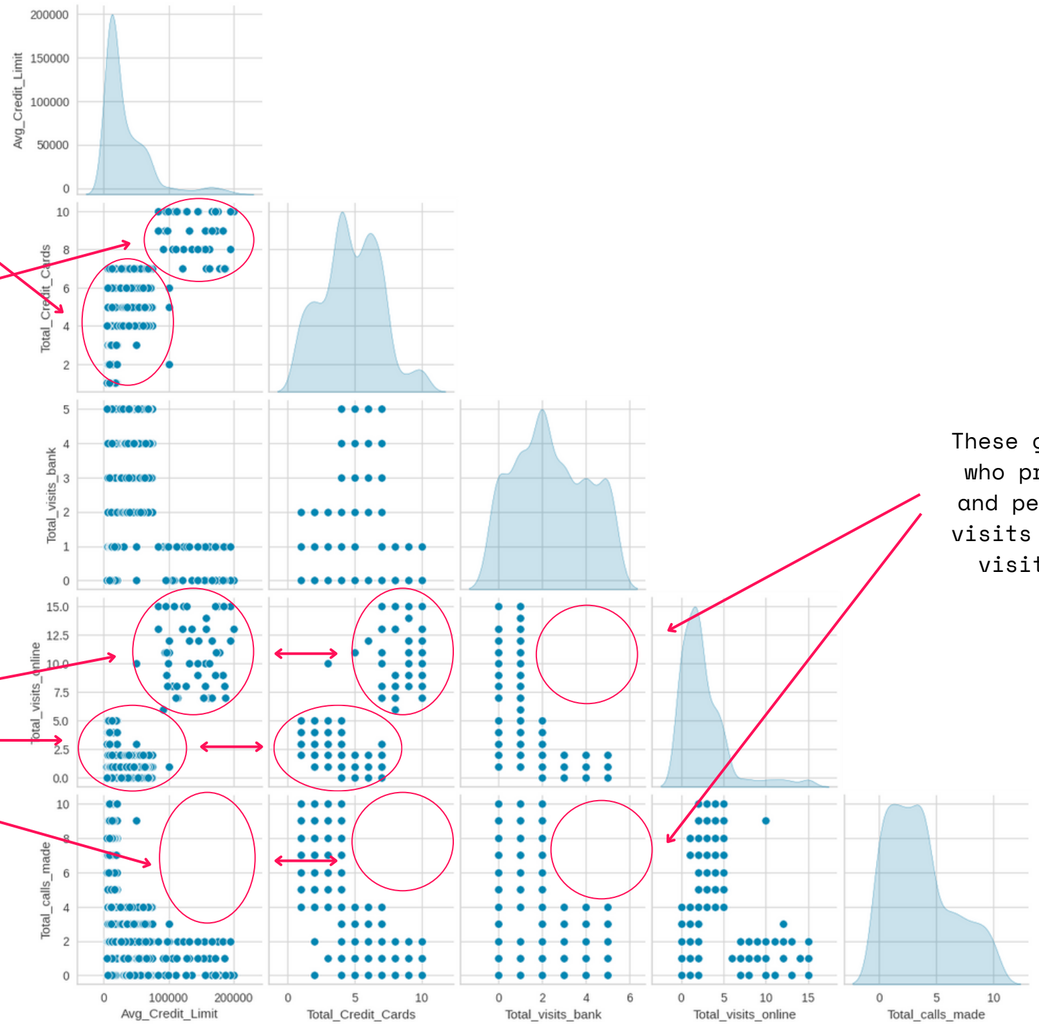
EDA Results Bivariate Analysis

Pairplot

This graph shows that customers with fewer than seven credit cards do not have access to high credit limits. This can be a selling point for offering more credit cards with the incentive of accessing higher credit limits.
In contrast, there are no customers with high credit limits who have fewer than six credit cards.

This suggests that it is the customers with high credit limits who really use the online service platform and call 2 or less times a year, while customers with low credit limits access the platform five or fewer times a year.

And, as the credit limit and the total of credit cards share a high correlation, the behavior corresponds in the total of calls made.



These graphs show that customers who prefer to make phone calls and people who engage in online visits do not make more than two visits to the bank per year.

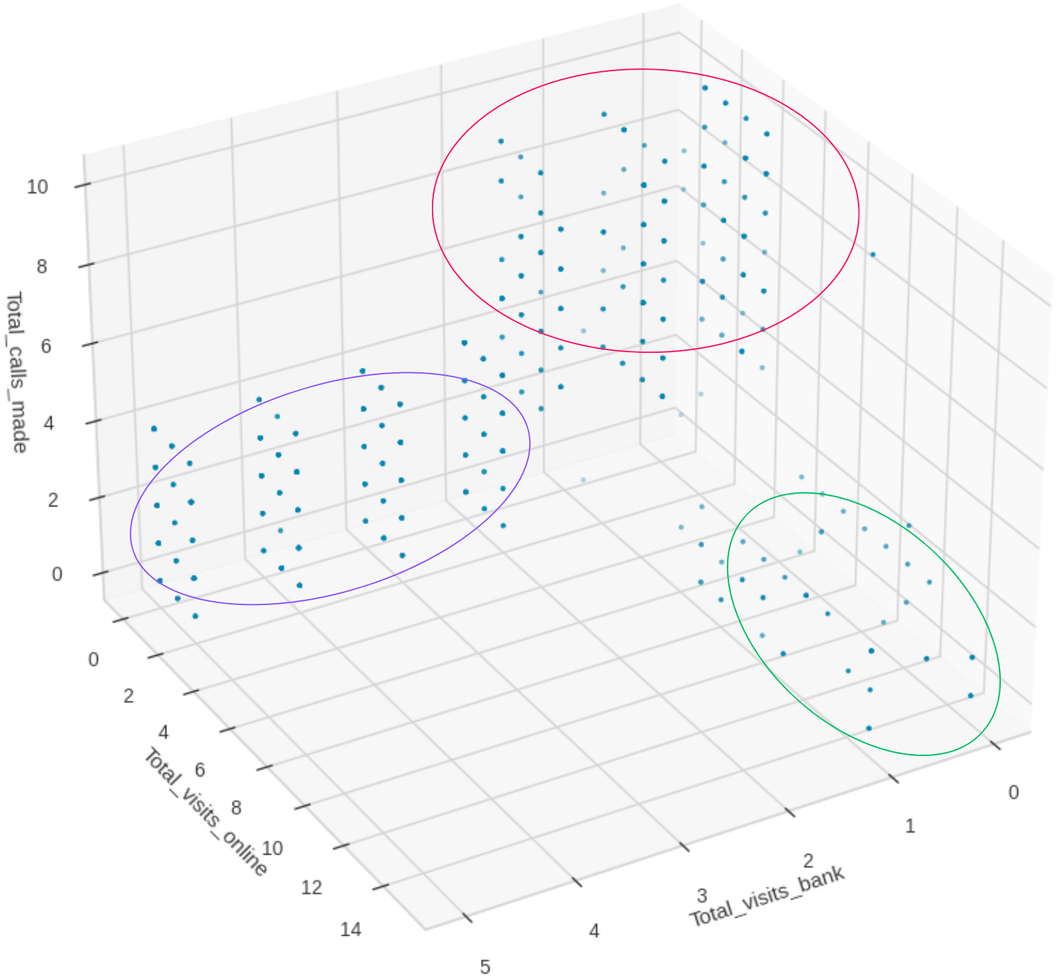
EDA Results Bivariate Analysis

Pairplot

Some of the conclusions that can be drawn from this graph are that customers with a high number of credit cards have a high number of online visits, a low number of visits to the bank, and a low number of calls made to the bank. It is worth noting that individuals with many credit cards who visit the website frequently are people who do not go to the bank and hardly make phone calls.

EDA Results Bivariate Analysis

Pairplot



Potential Customer Segments:

- Online Preferential Group: Customers who predominantly interact with the bank online.
- Bank Visit Preferential Group: Customers with a higher number of bank visits compared to other channels.
- Call Preferential Group: Customers who prefer calling the bank for their queries.

Data Preprocessing

Missing values

There are no missing values in the data.

Duplicate Values

There are five duplicated data points in the Customer Key column. This correspond to real serial number changes in the records and behaviors in customer accounts. Therefore, we will not remove these rows since they only have duplicate customer numbers; beyond that, the rows are different.

We will drop both Sl_No and Customer Key for the analysis.

Data Preprocessing

Outlier Detection

The outlier detection using the Interquartile Range (IQR) method reveals the following:

- Average Credit Limit: Values ranging from 108,000 to 200,000.
- Total Visits Online: Values ranging from 9 to 15.
- Total Credit Cards, Total Visits Bank, and Total Calls Made: No outliers

The outliers identified using the Z-score method with a threshold of 3 are as follows:

- Average Credit Limit: Values ranging from 153,000 to 200,000.
- Total Visits Online: Values ranging from 12 to 15.
- Total Credit Cards, Total Visits Bank, and Total Calls Made: No outliers were identified using this method.

Outlier Handling

After conducting a comparative analysis on the clusters using both techniques, handling outliers, and not handling them, it was concluded to retain them in the dataset. This decision was made because they do not significantly influence the results but maybe refine them at best, and they actually provide valuable information that has been observed in the study and is mentioned in the recommendations.

Comparative Analysis: Z-score Method vs. IQR Method

1. Identification Criterion
   - Z-score: It leverages the mean and standard deviation to identify outliers. A data point is considered an outlier if it is more than a certain number of standard deviations away from the mean.
   - IQR: It uses quartiles to identify outliers. Outliers are data points that lie below Q1 - 1.5 × IQR or above Q3 + 1.5 × IQR.

2. Effect of Skewness
   - Z-score: Sensitive to skewness. In a skewed distribution, the mean is influenced by extreme values, potentially misidentifying outliers.
   - IQR: Less sensitive to skewness since it uses median-based quartiles, which are robust to extreme values.

3. Outliers Identified
   - In this specific case, the Z-score method identified fewer outliers for "Total Visits Online" and more distinct outlier values for "Average Credit Limit" compared to the IQR method.

4. Applicability:
   - Z-score: Best suited for symmetric distributions or when the data follows a Gaussian distribution.
   - IQR: Can be used for both symmetric and asymmetric distributions, providing a more general approach to identifying outliers.

K-Means Clustering Summary

We employed the Elbow method to determine the optimal number of clusters and validated it using the silhouette score, ensuring a balanced and distinct segmentation.

The K-Means clustering algorithm delineated **3** distinct customer groups, each exhibiting unique characteristics:

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	$33,782.38	5.515544	3.489637	0.981865	2.000000	386
1	$141,040.00	8.740000	0.600000	10.900000	1.080000	50
2	$12,174.11	2.410714	0.933036	3.553571	6.870536	224

Cluster 0: Customers with moderate credit limits and a medium number of credit cards, preferring in-bank services.

Cluster 1: High-value customers characterized by high credit limits and a preference for online services.

Cluster 2: Customers with lower credit limits and fewer credit cards, predominantly utilizing call services for their queries.

Link to Appendix slide on K-Means Clustering

Hierarchical Clustering Summary

Hierarchical clustering effectively isolated 3 distinct customer segments with differing behavioral patterns and preferences

Cluster Profiling

HC_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	$33,713.18	5.511628	3.485788	0.984496	2.005168	387
1	$141,040.00	8.740000	0.600000	10.900000	1.080000	50
2	$12,197.31	2.403587	0.928251	3.560538	6.883408	223

Cluster 0: Customers more inclined to visit the bank physically, holding a moderate number of credit cards and showcasing a moderate credit limit.

Cluster 1: Customers who prefer online interactions, characterized by high credit limits and a high number of credit cards.

Cluster 2: Customers primarily using call services, with lower credit limits and fewer credit cards.

Link to Appendix slide on Hierarchical Clustering

APPENDIX

Data Background and Contents

The dataset has 660 rows and seven columns.
All the columns in the data are numeric.

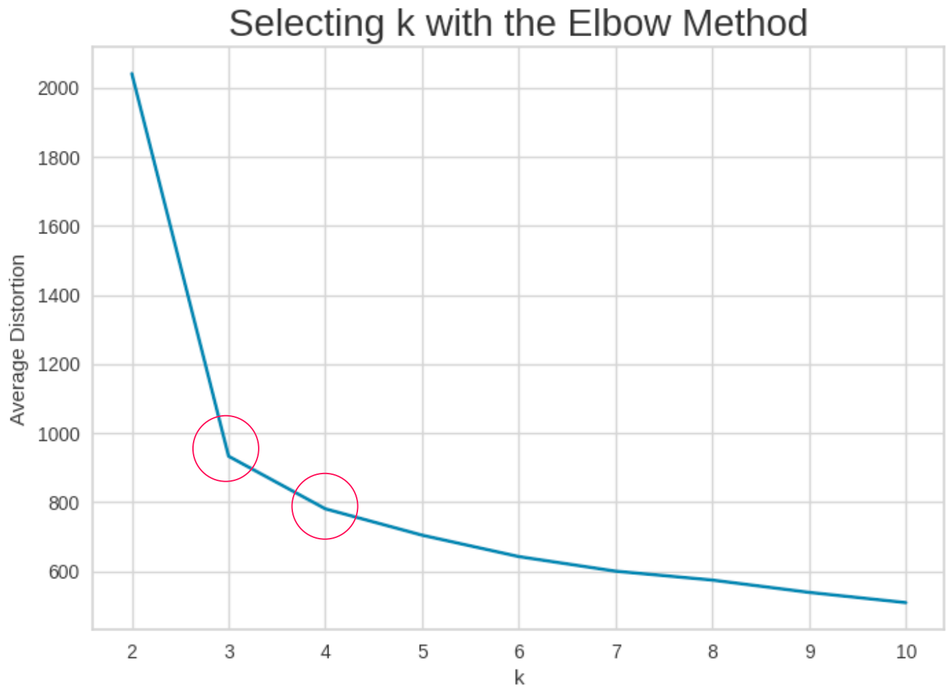
Data Dictionary

- Sl_No: A primary key for record identification.
- Customer Key: A unique identifier for customers.
- Average Credit Limit: The average credit limit across all credit cards for each customer.
- Total Credit Cards: The total number of credit cards held by each customer.
- Total Visits Bank: The total number of in-person visits made by each customer to the bank on a yearly basis.
- Total Visits Online: The total number of online visits or logins made by each customer on a yearly basis.
- Total Calls Made: The total number of calls made by each customer to the bank or its customer service department on a yearly basis.

K-Means Clustering Technique

We employed the Elbow method to determine the optimal number of clusters and validated it using the silhouette score, ensuring a balanced and distinct segmentation.

Elbow Curve

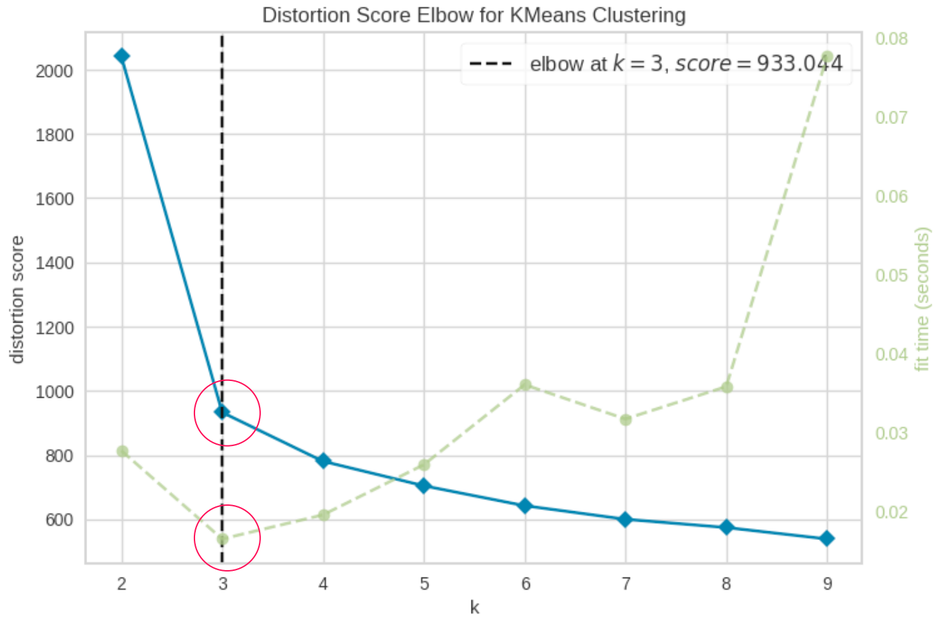


Analyzing the plot, we can observe that the "elbow" is formed at around 3 or 4 clusters. This suggests that choosing 3 or 4 as the number of clusters might be a good choice, as increasing the number of clusters beyond this point results in a smaller reduction in Average Distortion, indicating diminishing returns.

K-Means Clustering Technique

We employed the Elbow method to determine the optimal number of clusters and validated it using the silhouette score, ensuring a balanced and distinct segmentation.

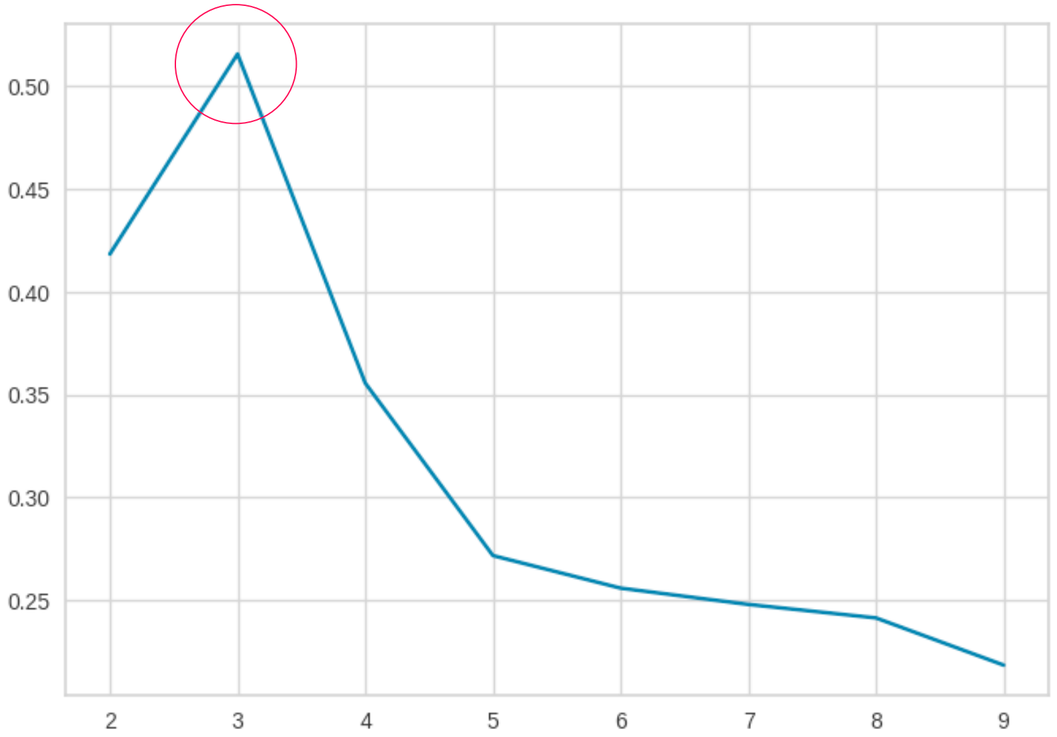
Elbow Curve

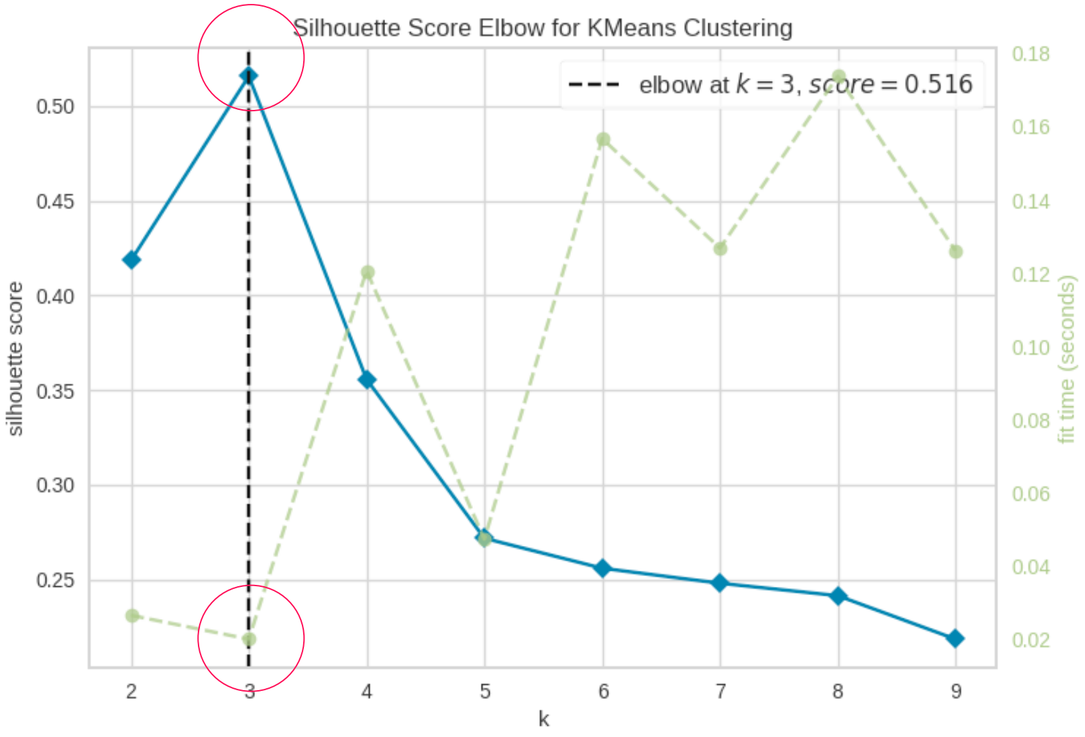


In this graph, we not only observe that the Distortion Score has a sharp drop from 2041 to 933 from K=2 to K=3 but also has the lowest fit time level.

K-Means Clustering Technique

Silhouette scores

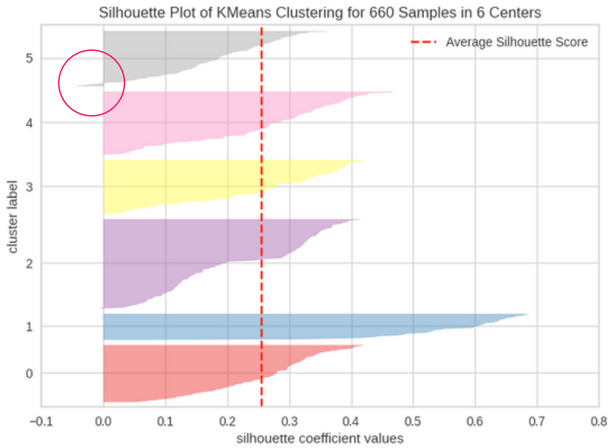


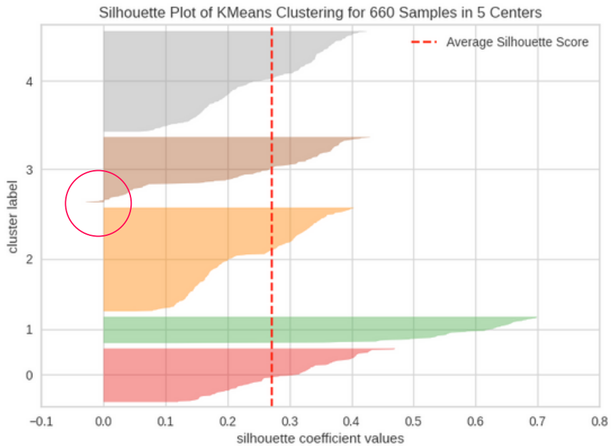


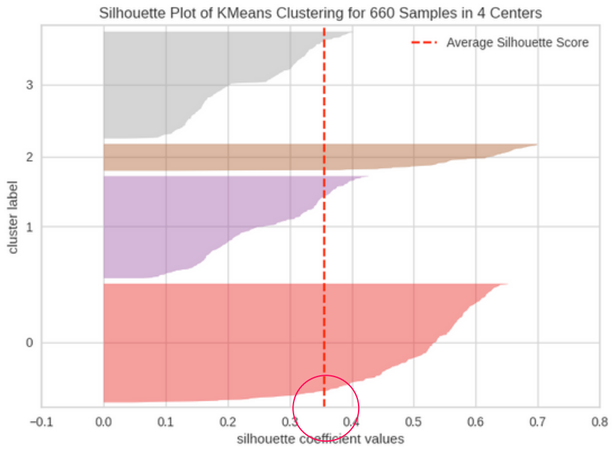
In this Silhouette scores graph, we see the highest score at 0.516, and we also observe the lowest fit time on the graph.

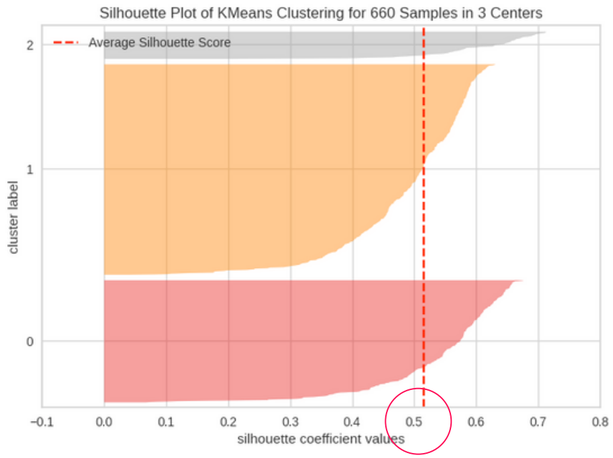
K-Means Clustering Technique

Silhouette scores









In these Silhouette plots of K-means, we analyze the coefficient values with 6, 5, 4, and 3 as values of K or cluster centers.

We observe that when we use 6 and 5 centers, we obtain negative coefficients, indicating that those values for K are not suitable.

When we use 4 centers, the coefficient value is 0.355, whereas the coefficient when we use 3 centers is 0.515.

In both cases all clusters have different Silhouette scores, widths, and cross the average scores.

Hierarchical Clustering Technique

To delineate distinct customer segments through hierarchical clustering, which offers a granular view of the customer base, facilitating a deeper understanding of the various customer groups and their preferences.

Distance Metric and Linkage Method	Cophenetic Correlation
Euclidean single linkage	0.7391220243806552.
Euclidean complete linkage	0.8599730607972423.
Euclidean average linkage	0.8977080867389372.
Euclidean weighted linkage	0.8861746814895477.
Chebyshev single linkage	0.7382354769296767.
Chebyshev complete linkage	0.8533474836336782.
Chebyshev average linkage	0.8974159511838106.
Chebyshev weighted linkage	0.8913624010768603.
Mahalanobis single linkage	0.7058064784553605.
Mahalanobis complete linkage	0.6663534463875359.
Mahalanobis average linkage	0.8326994115042136.
Mahalanobis weighted linkage	0.7805990615142518.
Cityblock single linkage	0.7252379350252723.
Cityblock complete linkage	0.8731477899179829.
Cityblock average linkage	0.896329431104133.
Cityblock weighted linkage	0.8825520731498188.

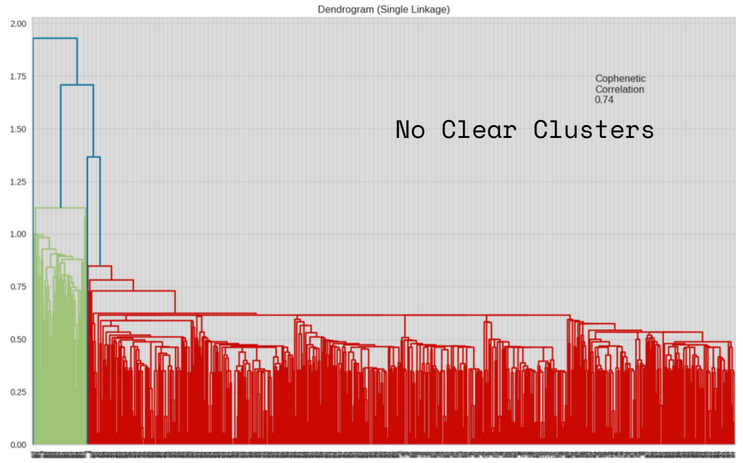
We adopted the Euclidean distance metric and average linkage criteria to create the hierarchical clustering dendrogram

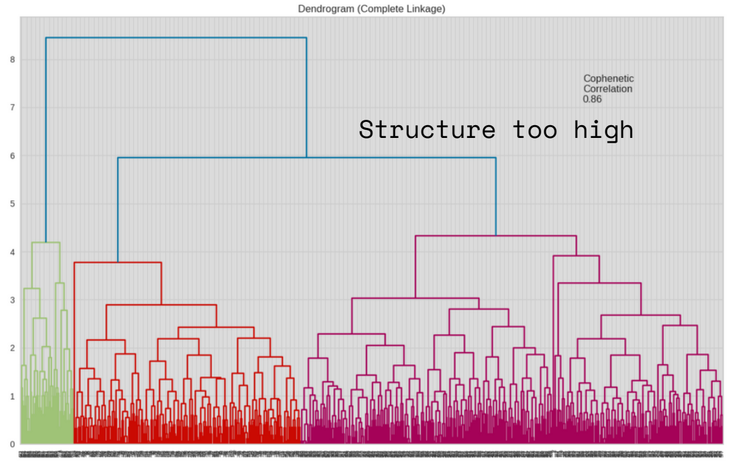
Distance Metric and Linkage Method	Cophenetic Correlation
Euclidean single linkage	0.739122024380655
Euclidean complete linkage	0.859973060797242
Euclidean average linkage	0.897708086738937
Euclidean centroid linkage	0.893938584632632
Euclidean median linkage	0.889379953701672
Euclidean ward linkage	0.741515628482749
Euclidean weighted linkage	0.886174681489548

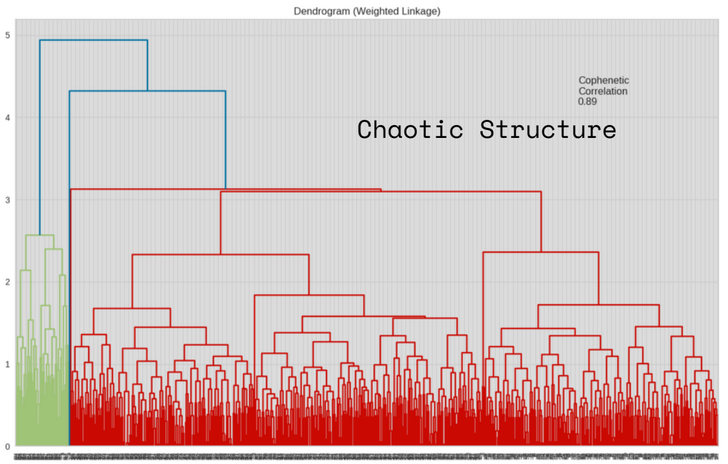
Even with different linkage methods with Euclidean distance only, the average linkage has the best Cophenetic Correlation.

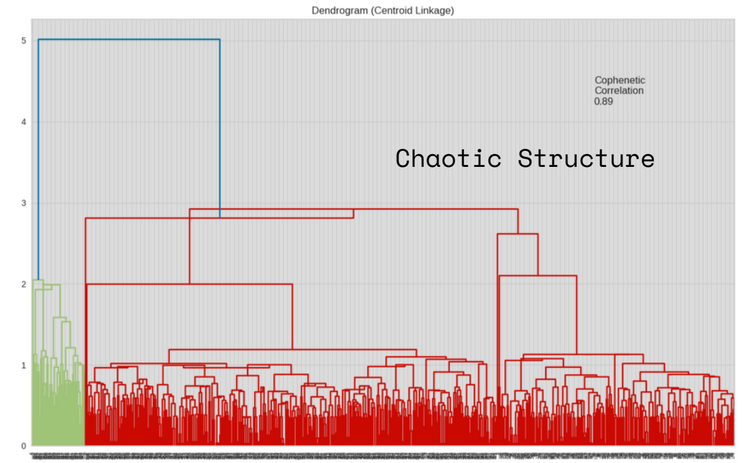
Hierarchical Clustering Technique

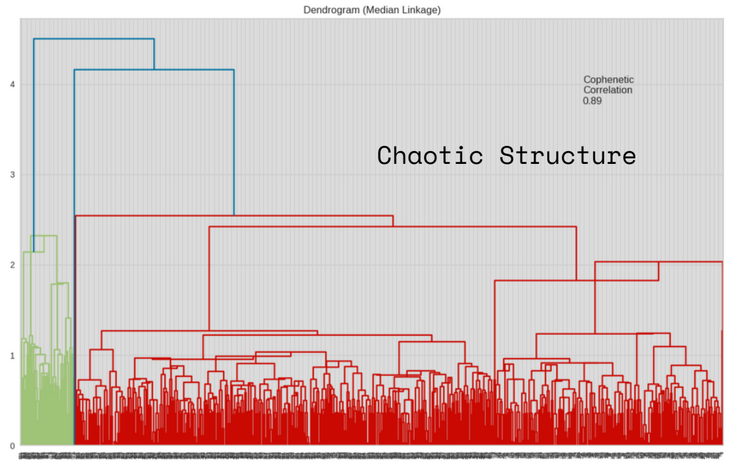
Dendrograms

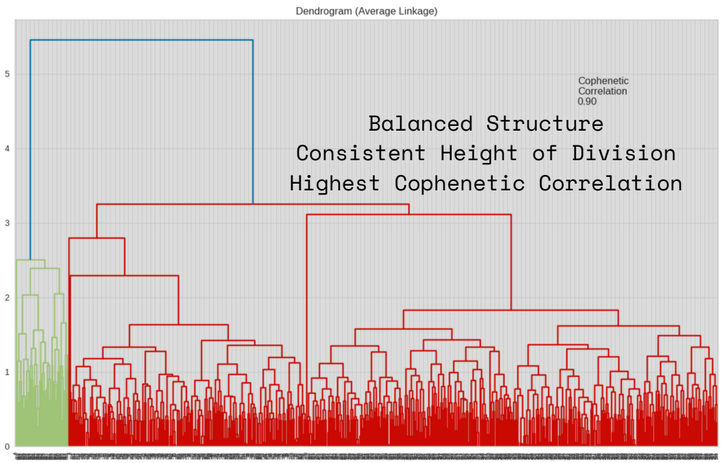






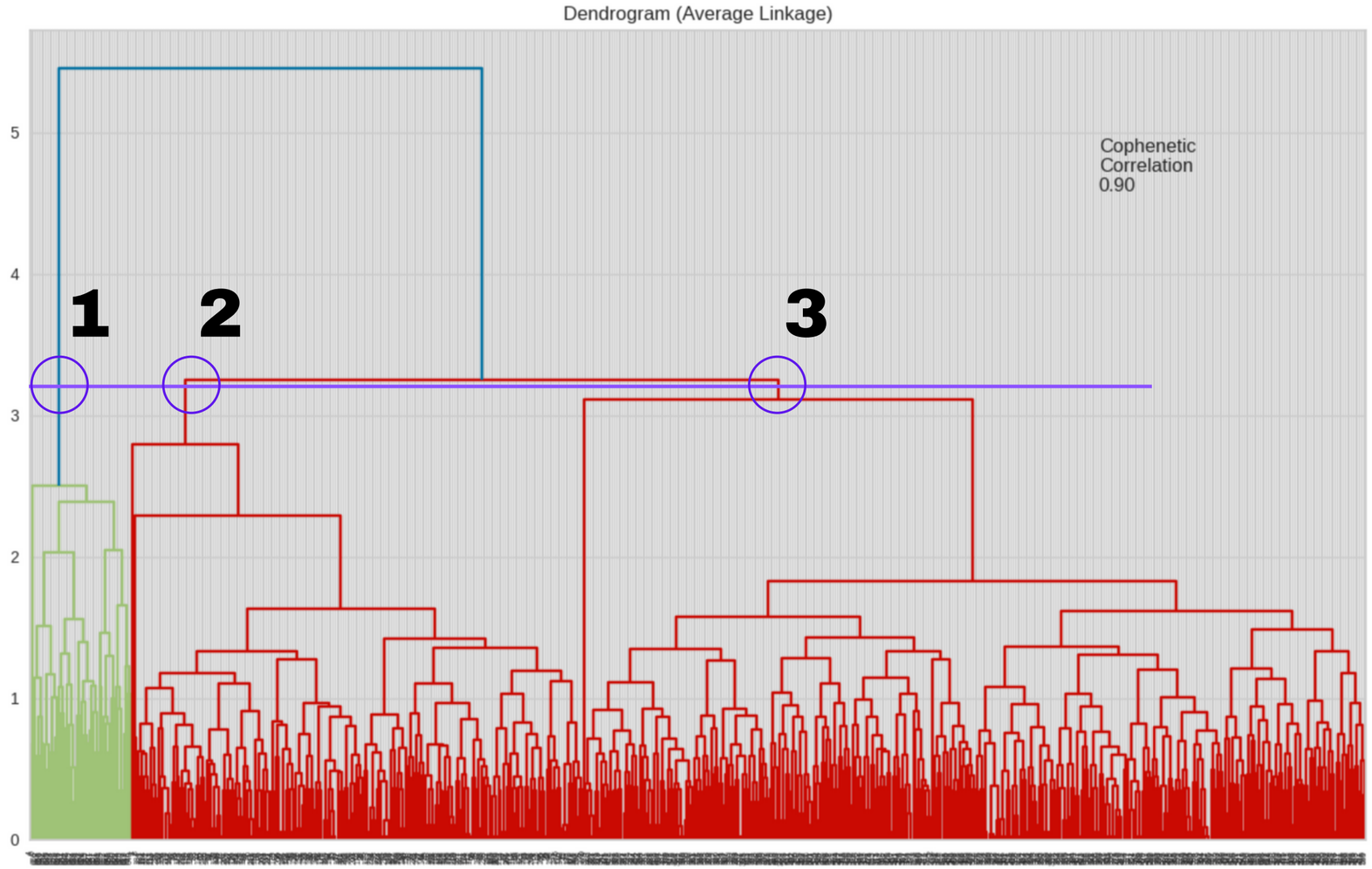






Hierarchical Clustering Technique

Dendrograms



K-Means vs Hierarchical Clustering

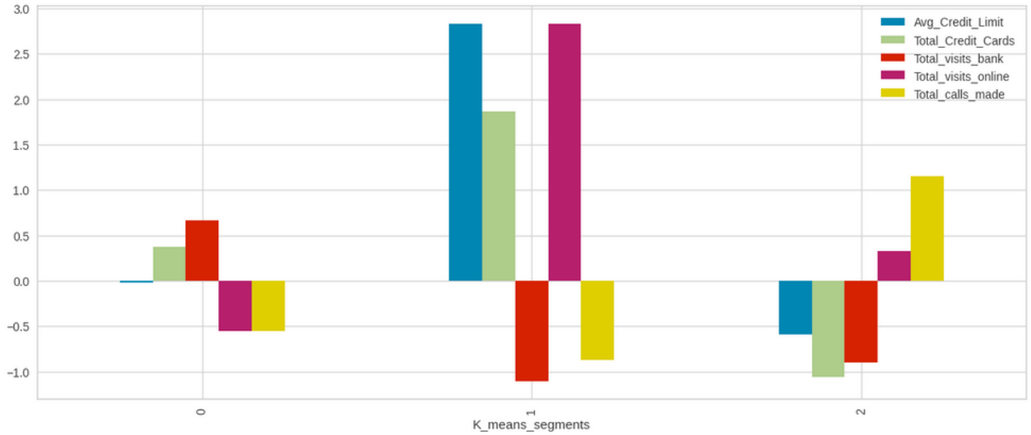
The K-means and hierarchical clustering (average linkage) methods have identified similar clusters (almost identical), indicating a consistency in the segmentation of customers.

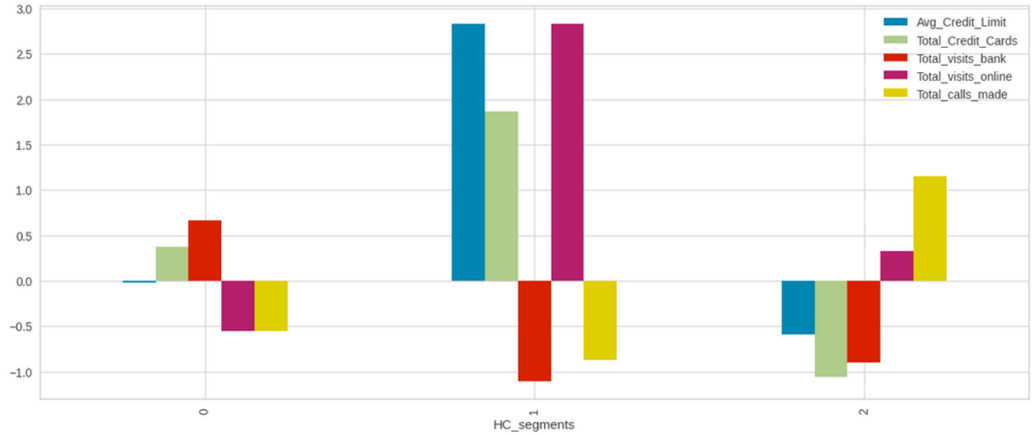
K_means_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	$33,782.38	5.515544	3.489637	0.981865	2.000000	386
1	$141,040.00	8.740000	0.600000	10.900000	1.080000	50
2	$12,174.11	2.410714	0.933036	3.553571	6.870536	224

HC_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	$33,713.18	5.511628	3.485788	0.984496	2.005168	387
1	$141,040.00	8.740000	0.600000	10.900000	1.080000	50
2	$12,197.31	2.403587	0.928251	3.560538	6.883408	223

K-Means vs Hierarchical Clustering

Both clustering graphs display similar characteristics:

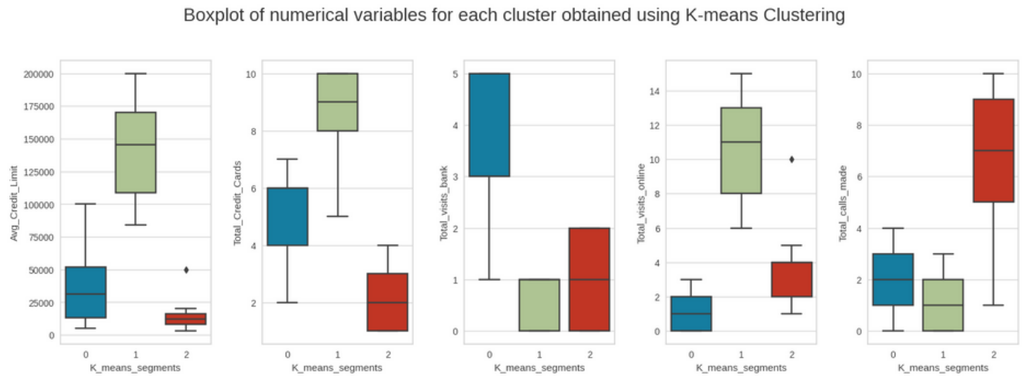


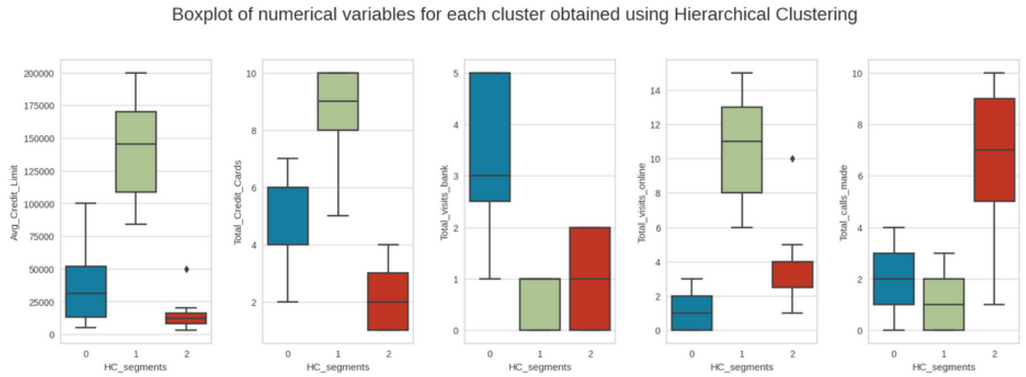


- In Segment 0, we see low Credit Limit, many visits to the bank with no online visits or calls, and a moderate total number of credit cards.
- In Segment 1, there is a high credit limit, many credit cards, many online visits with no bank visits or calls.
- In Segment 2, we observe a very low credit limit, very few credit cards, very few bank visits, many calls, and some online visits.

K-Means vs Hierarchical Clustering

Both clustering graphs display graphs almost indistinguishable:





In these boxplots, we observe distinctive clusters that have little or no overlap in the numeric variables, specifically in the variables of the total number of credit cards, total online visits, and the average credit line.

There are some overlaps in the total visits and total calls.

Great Learning

Happy Learning !