AllLife Bank Study

Unsupervised Learning Project
Credit Card Customer Segmentation

August 4, 2023

By: Yolanda OMalley

Contents / Agenda

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

Executive Summary

Objective:

To identify different segments in the existing customer, 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.

Executive Summary

Conclusions:

We will look into clusters 0, 1, and 2 only because cluster 3 have only 1 custoomer in it. Both Hierarchical Clustering & K-means Clustering are very similar except for a small change in cluster 2 on Total visit Online & Average Credit Limit.

- **Cluster 0**

There are 223 customers in this cluster.

The average credit limit of customers is low with 12156.95

Total credit cards are low with 2 credit cards.

Total visits at the bank are low, but the Total visits online are moderate with 4

Total calls made to the bank are high with 7

Executive Summary

Conclusions:

- **Cluster 1**

There are 50 customers in this cluster.

The average credit limit of customers is high with 141040.00

Total credit cards are high with 9 credit cards.

Total visits at the bank are low, but the Total visits online are high with 11

Total calls made to the bank are low with 1

- **Cluster 2**

There are 387 customers in this cluster.

The average credit limit of customers is moderade with 33744.19

Total credit cards are high with 6 credit cards.

Total visits at the bank are moderade, but the Total visits online are low with 1

Total calls made to the bank are 2

Executive Summary

Recomendations:

AllLife bank have to build a clustering algorithms obtained by K-means Clustering with 3 clusters to provide recommendations to the bank on how to get a better market and give better service to these customers.

Executive Summary

Recomendations:

- Cluters 0 customers are good places for AllLife Bank to focus on marketing campaigns to target customers with low Total credit cards & Total visits at the bank to increse customers service & upsell existing customers at the bank based on cluster profiling done.

- Cluters 1 customers are good places for AllLife Bank to focus on marketing campaigns to target customers with low Total visits at the bank to increse customers service at the bank and to target new customers based on cluster profiling done.

- Cluters 2 customers are good places for AllLife Bank to focus on marketing campaigns to target customers with low Total visits online at the bank to increse customers service at the bank based on cluster profiling done.

Business Problem Overview and Solution Approach

We will be focusing on:

- AllLife Bank credit card customer base in the next financial year by identifying different segments in the existing customer, based on their spending patterns as well as past interaction with the bank.

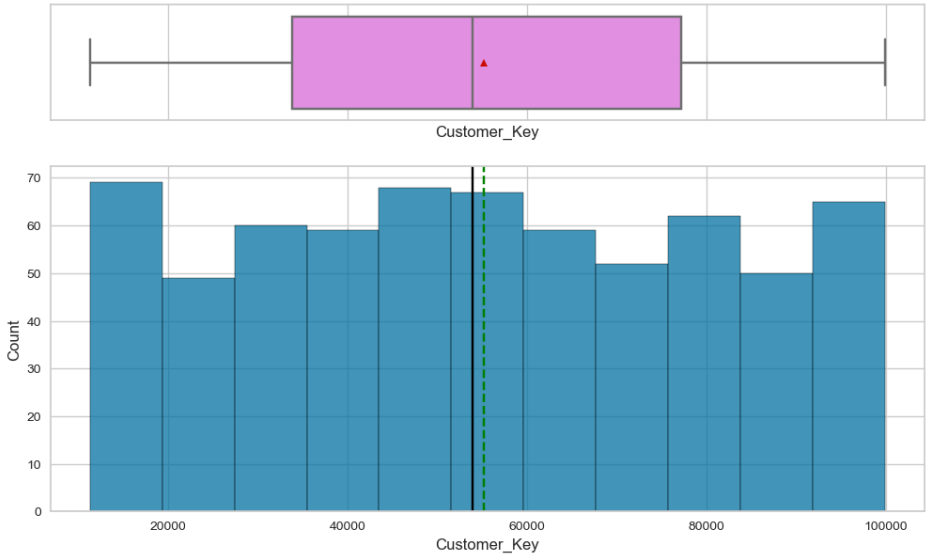
- Way to target new customers as well as upsell existing customers using the marketing department.

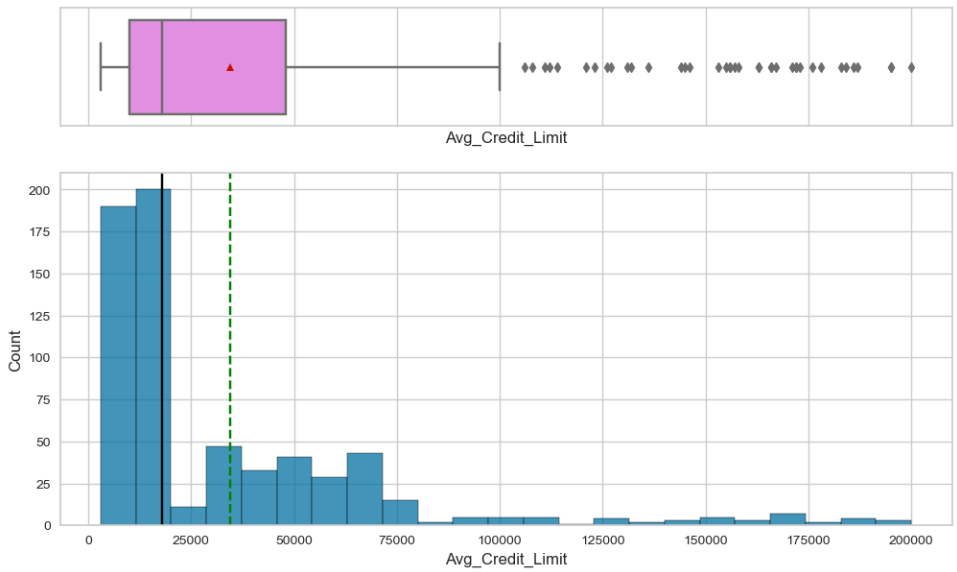
- We will use data preprocessing and EDA using descriptive statistics and visualizations

- We will use clustering algorithms to do customer segmentation and analyze these segments to gain insights.

EDA Results

Univariate Analysis – Customer Key & Avg Credit Limit



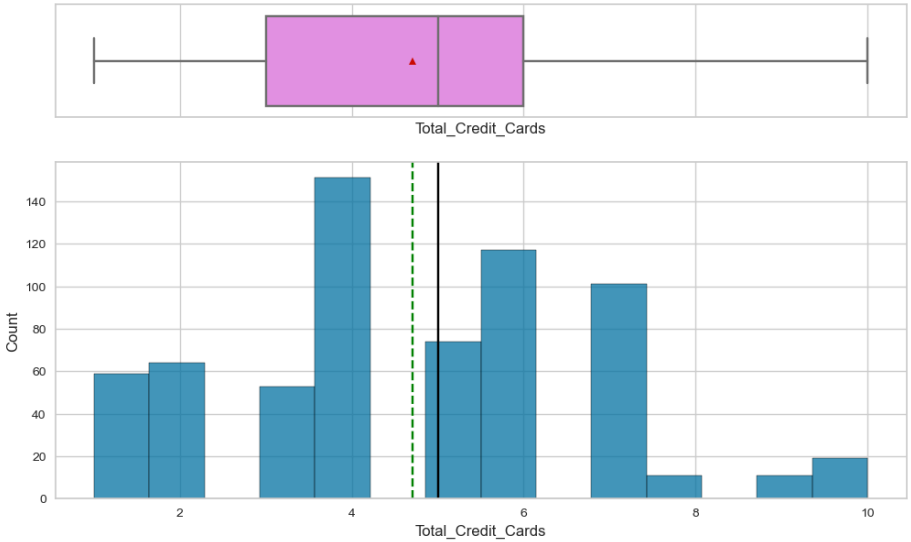


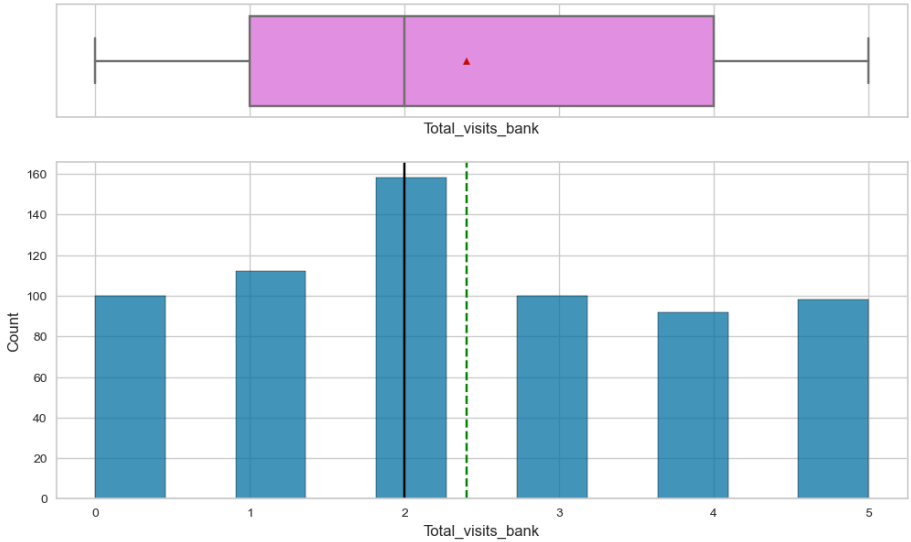
Observations:

Avg credit limit has right-skewed distributions with upper outliers, which indicates the presence of customers with very high credit limit.

EDA Results

Univariate Analysis – Total credit cards & Total visits bank



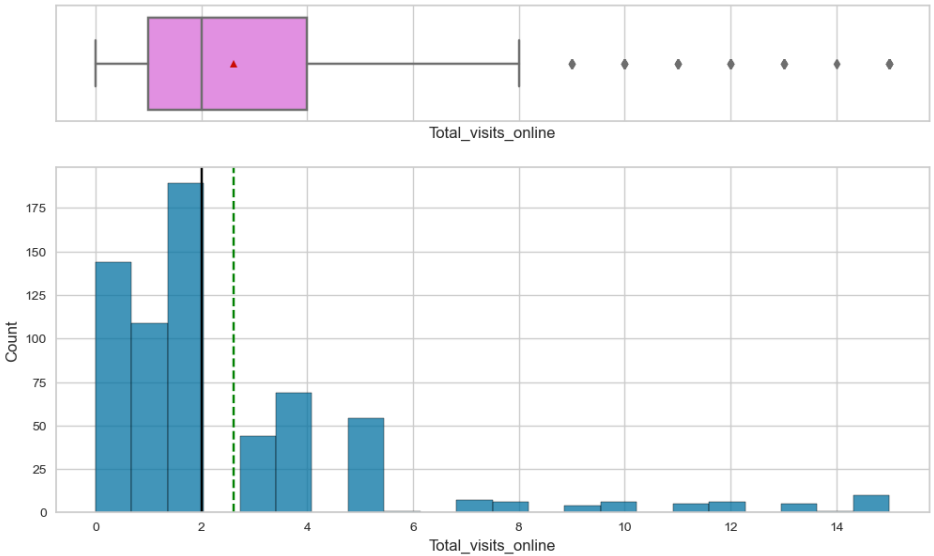


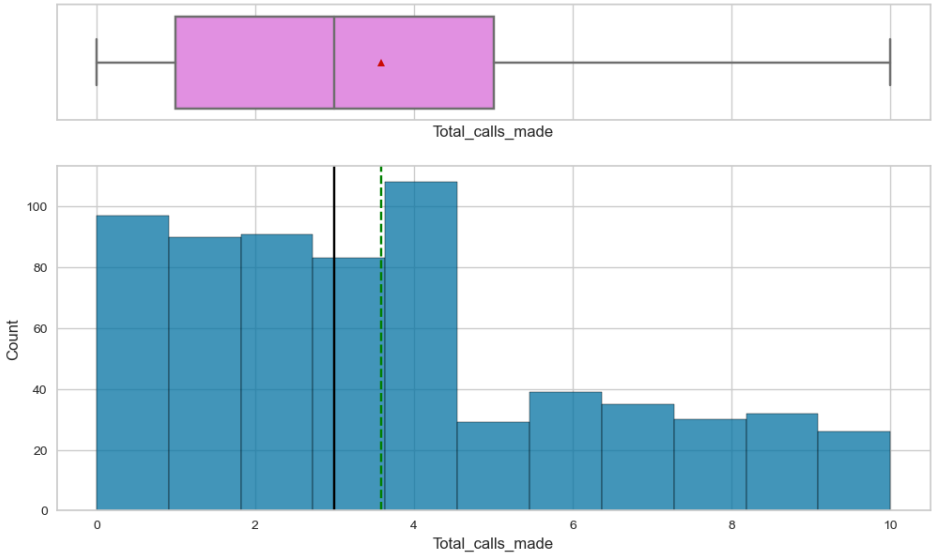
Observations:

The average total credit cards is 5 & the avg total visit bank is 2.

EDA Results

Univariate Analysis – Total visits online& Total calls made



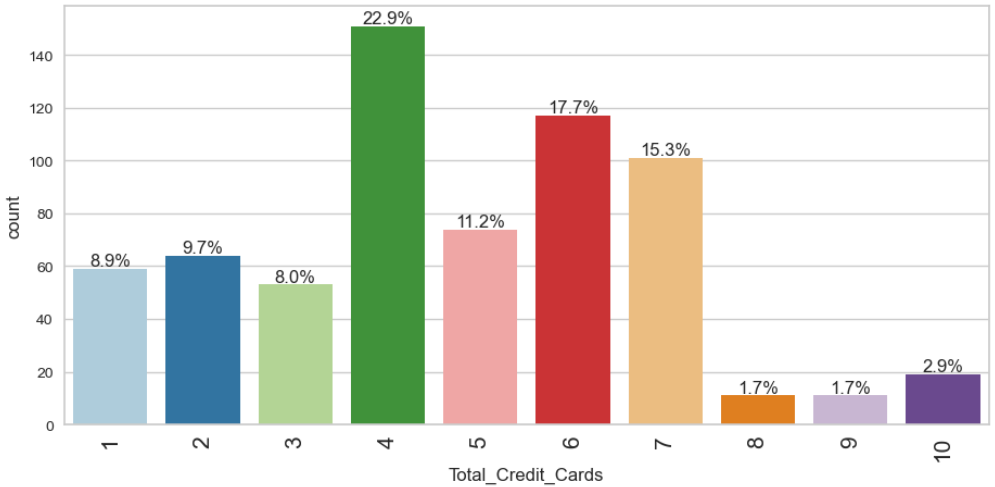


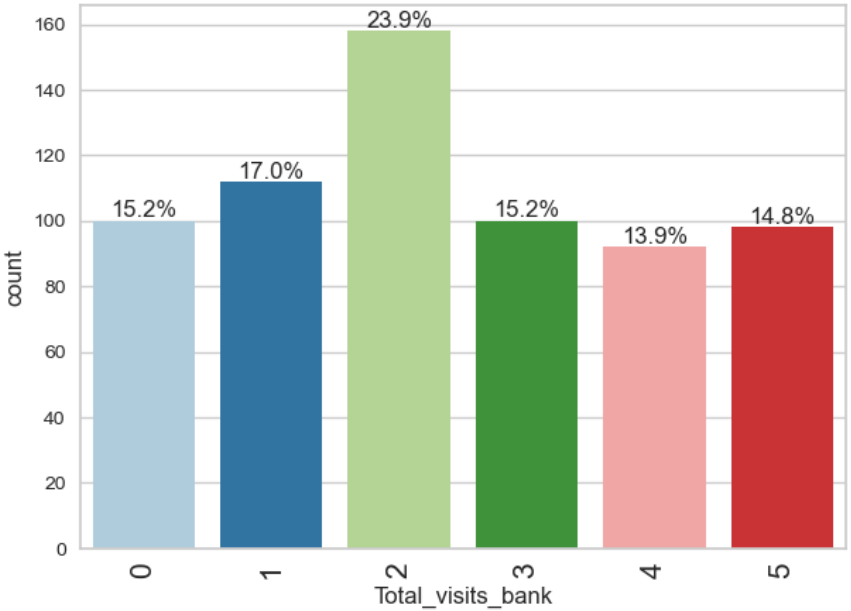
Observations:

Total visit online have right-skewed distributions with upper outliers, which indicates the presence of customers with very high credit limit & visit online.

EDA Results

Univariate Analysis – Total credit cards & Total visits bank





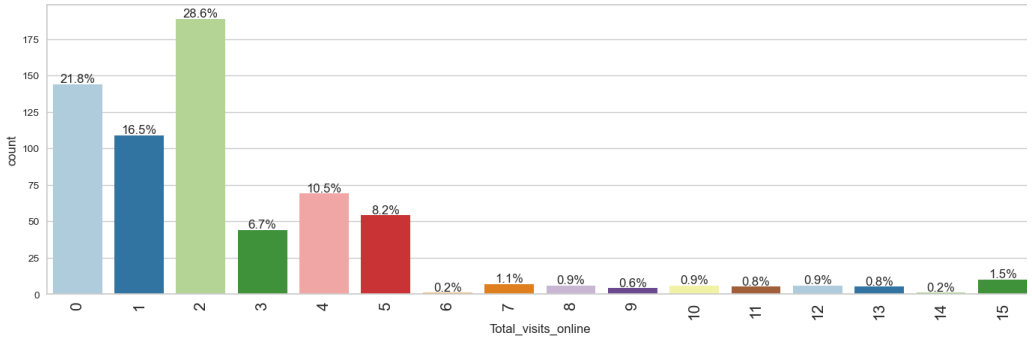
Observations:

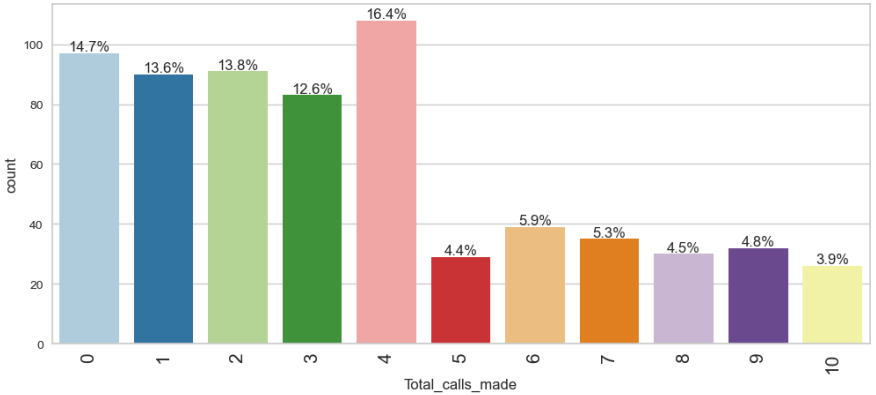
22.9% of the customers of the bank have 4 credit cards

23.9% of the customers of the bank have 2 total visits to the bank

EDA Results

Univariate Analysis – Total visits online & Total calls made





Observations:

28.6% bank customers have 2 total visit online

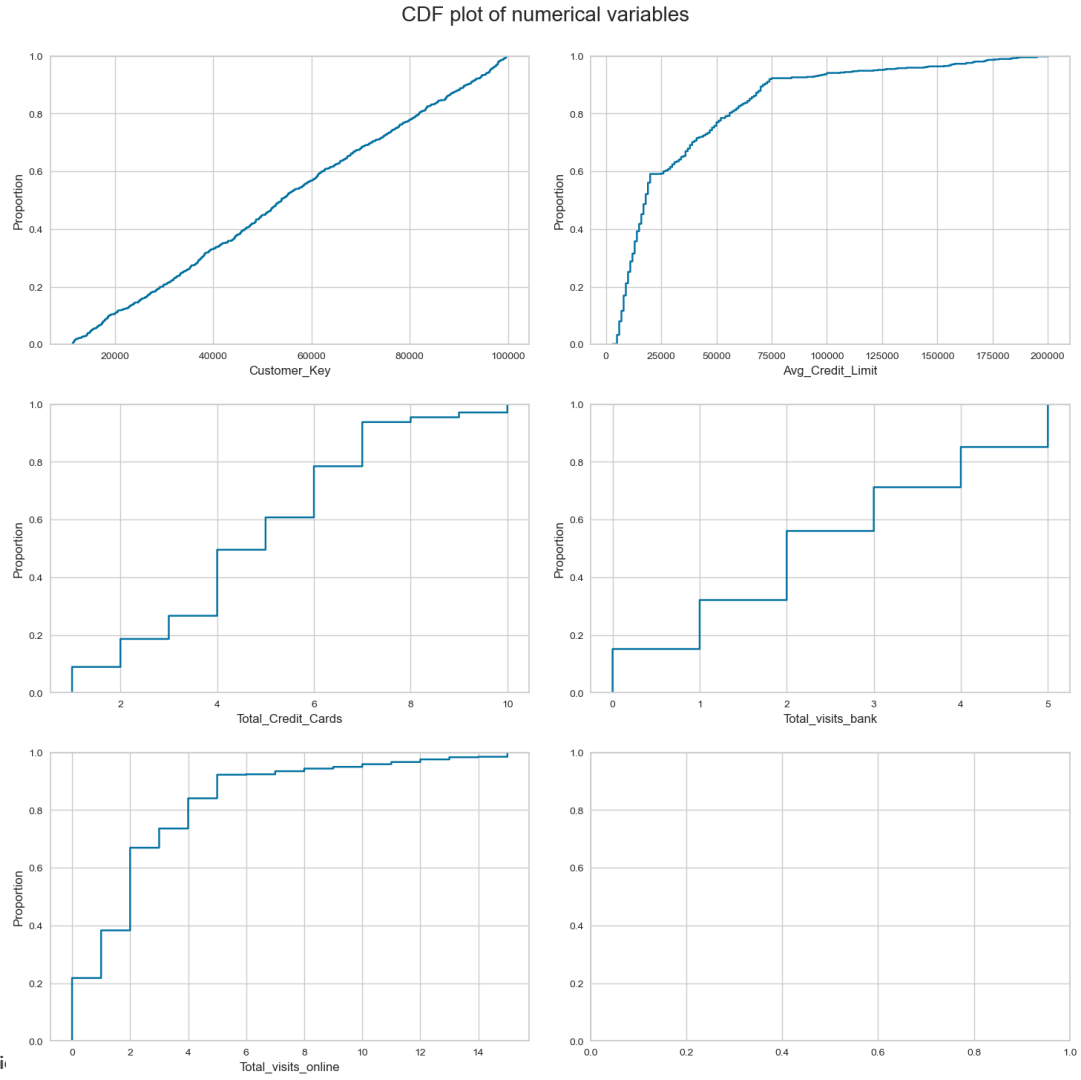
16.4% bank customers have 4 total calls made

EDA Results

Univariate Analysis – Numerical variables

Observations:

- 90% of bank customers have 75000 average credit card limit
- 95% of bank customers have 8 credit cards
- 75% bank customers have 4 visits to the bank
- The maximum visit online is 15



Great Learning
POWER AHEAD

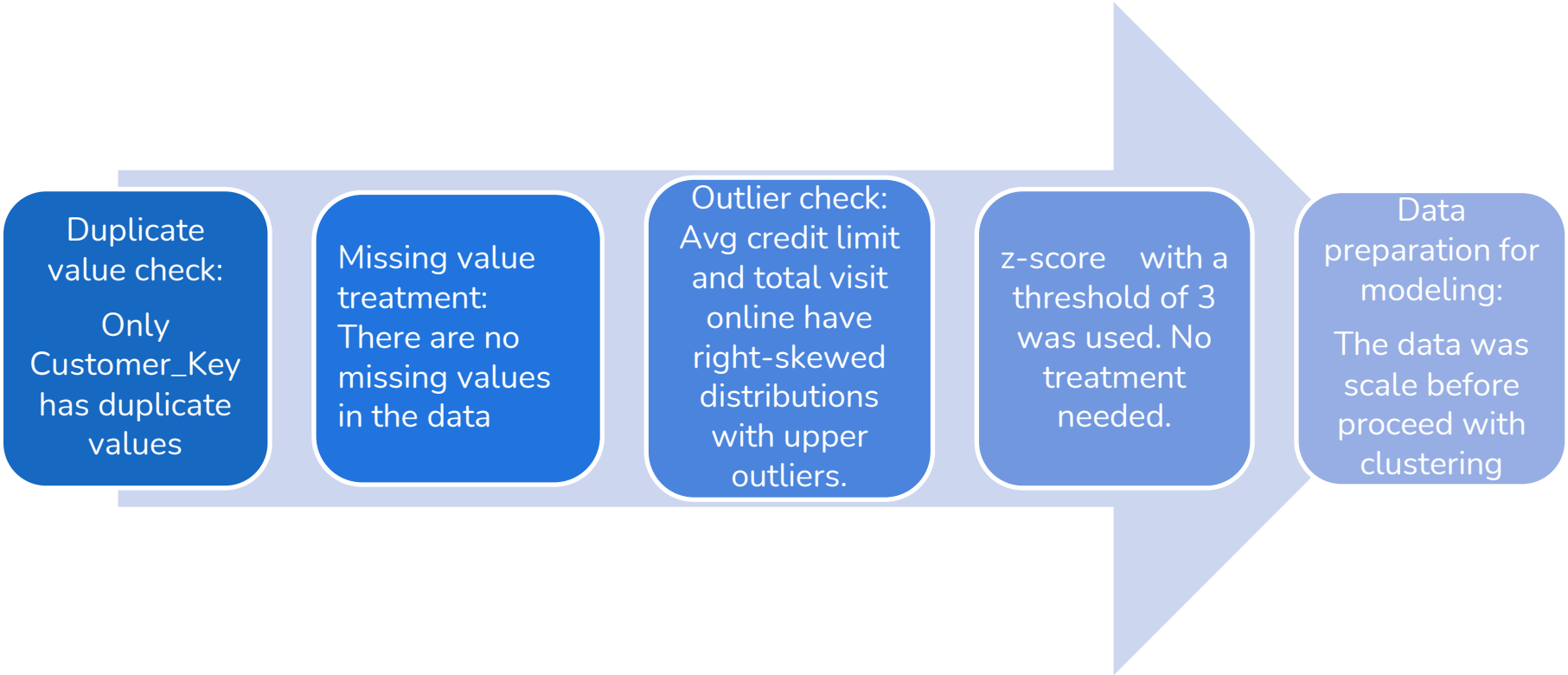
EDA Results

Bivariate Analysis

Observations:

- Average credit limit and total credit cards have a high to moderate correlation
- Total calls made and total credit cards have a negative correlation.

Data Preprocessing

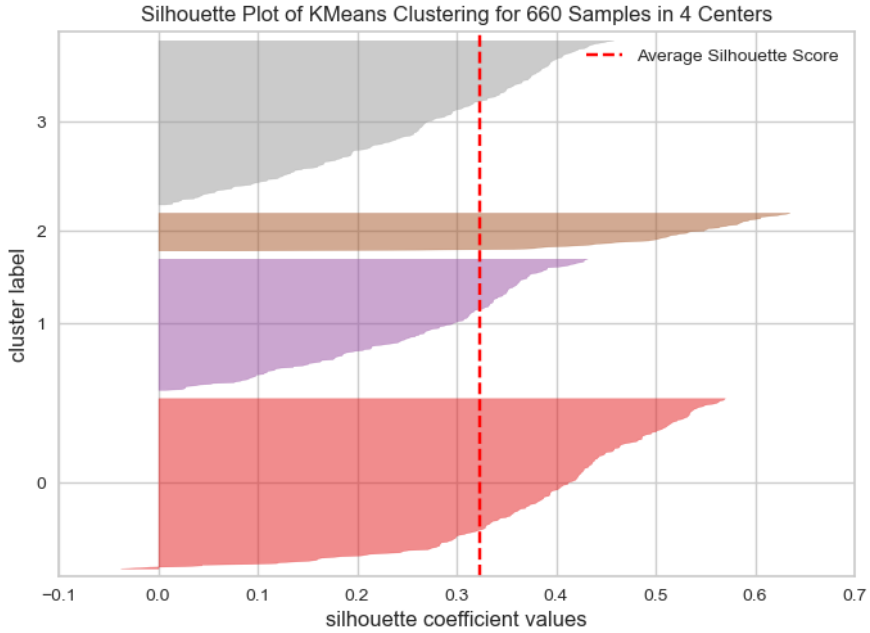


K-Means Clustering Summary

- Optimal Number of clusters using K-Means

The plot shows that all the clusters meet the requirements. All the clusters have crossed Avg silhouette score, all the clusters have different silhouette scores and have different width sizes. So, let's take 4 as the appropriate no. of clusters(K) as the silhouette score is high enough for all the 4-clusters(above average silhouette score), and there is a Knick at 4 in the elbow curve

- Cluster Profiling



Link to Appendix slide on K-Means Clustering

K-Means Clustering Summary - Cluster Profiling

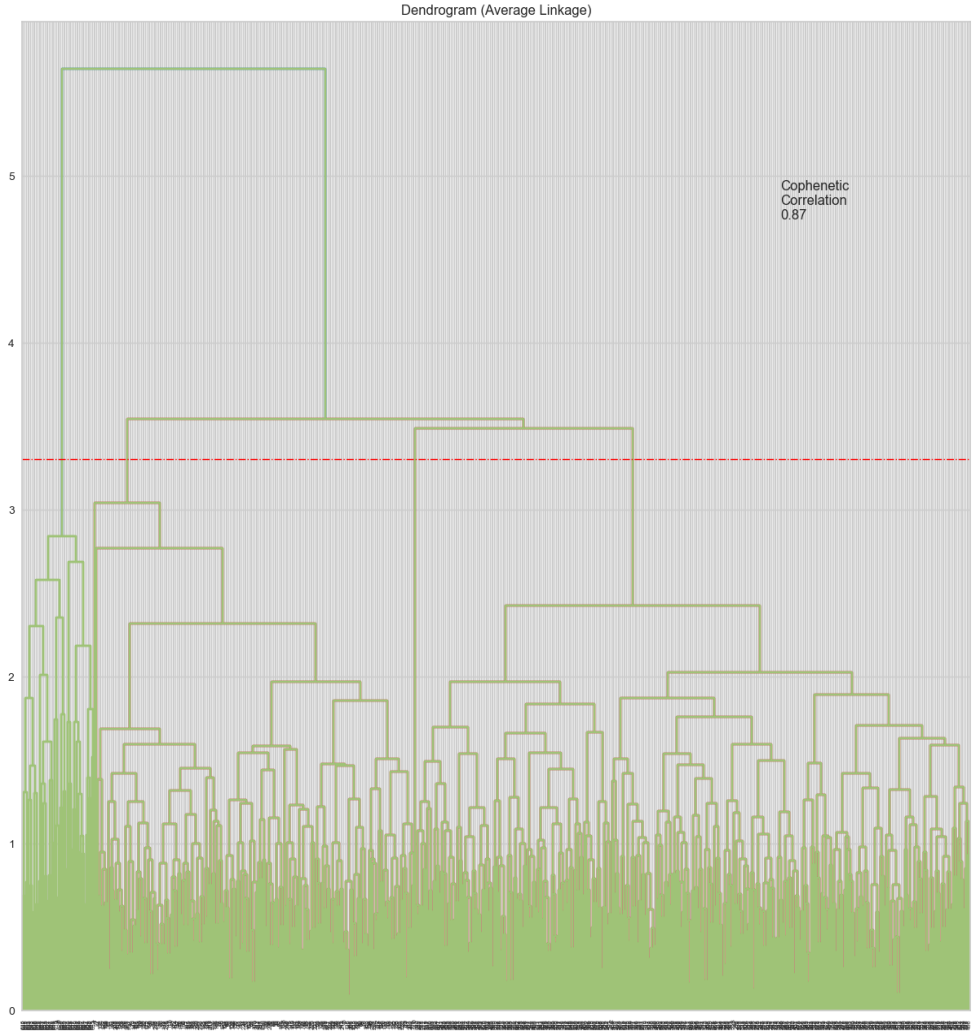
HC_Clusters	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	55252.730	12156.950	2.403587	0.928251	3.556054	6.883408	223
1	56708.760	141040.00	8.740000	0.600000	10.900000	1.080000	172
2	54791.406	33564.766	5.520725	3.492228	0.987047	2.010363	50
3	87073.000	100000.00	2.000000	1.000000	1.000000	0.000000	215

Cluster 0: There are 223 customers in this cluster. Total credit cards are low with 2 credit cards. Total visits at the bank are low, but the Total visits online are moderate with 4. Total calls made to the bank are high with 7.
Cluster 1: There are 172 customers in this cluster. Total credit cards are high with 9 credit cards. Total visits at the bank are low, but the Total visits online are high with 11. Total calls made to the bank are low with 1.
Cluster 2: There are 50 customers in this cluster. Total credit cards are high with 6 credit cards. Total visits at the bank are moderate, but the Total visits online are low with 1. Total calls made to the bank are 2.
Cluster 3: There are 215 customers in this cluster. Total credit cards are low with 2 credit cards. Total visits at the bank & the Total visits online are low with 1 each.

Hierarchical Clustering Summary

- Optimal Number of clusters using Hierarchical Clustering

The cophenetic correlation is highest for average and centroid linkage methods. 4 appears to be the appropriate number of clusters from the dendrogram for Average linkage.



chical Clustering

Hierarchical Clustering Summary - Cluster Profiling

HC_segments	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	55252.730	12156.950	2.403587	0.928251	3.556054	6.883408	223
1	56708.760	141040.00	8.740000	0.600000	10.900000	1.080000	50
2	54791.406	33564.766	5.520725	3.492228	0.987047	2.010363	386
3	87073.000	100000.00	2.000000	1.000000	1.000000	0.000000	1

Cluster 0: There are 223 customers in this cluster. Total credit cards are low with 2 credit cards. Total visits at the bank are low, but the Total visits online are moderate with 4. Total calls made to the bank are high with 7.
Cluster 1: There are 50 customers in this cluster. Total credit cards are high with 9 credit cards. Total visits at the bank are low, but the Total visits online are high with 11. Total calls made to the bank are low with 1.
Cluster 2: There are 386 customers in this cluster. Total credit cards are high with 6 credit cards. Total visits at the bank are moderade, but the Total visits online are low with 1. Total calls made to the bank are 2.
Cluster 3: There are 1 customers in this cluster. Total credit cards are low with 2 credit cards. Total visits at the bank & the Total visits online are low with 1 each.

APPENDIX

Data Background and Contents

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). 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 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)

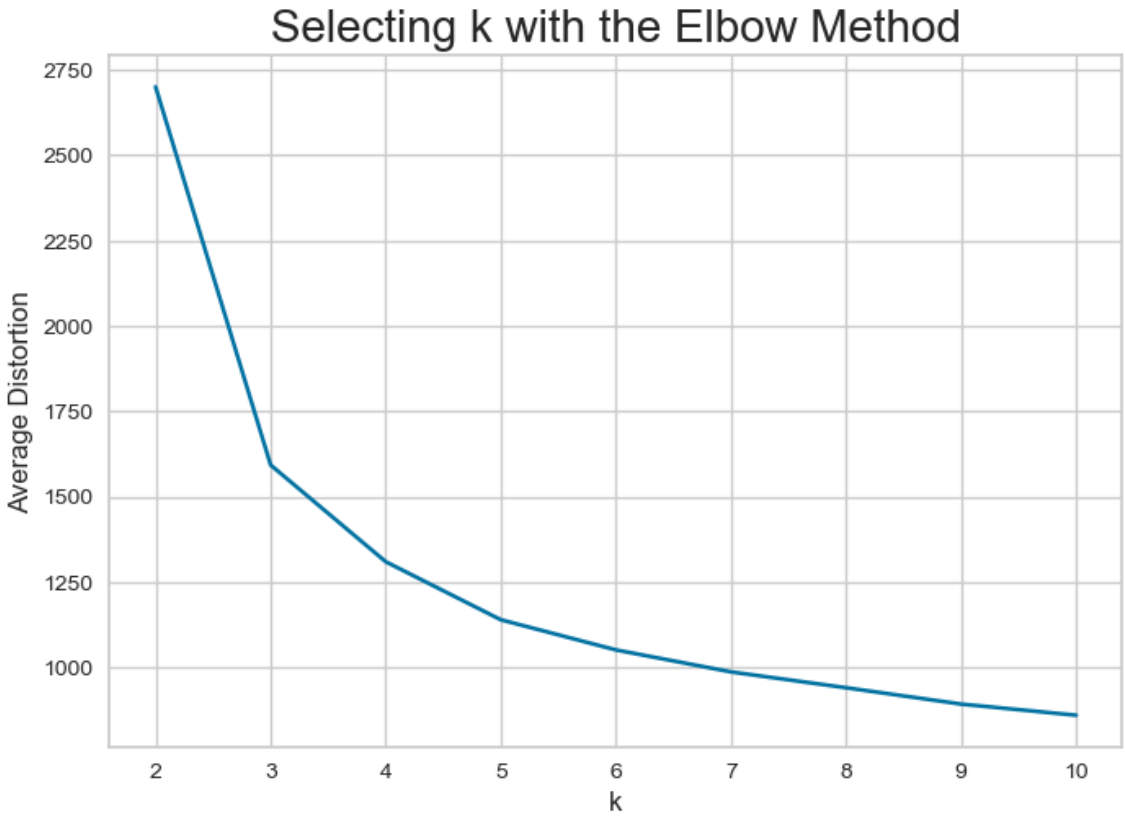
K-Means Clustering Technique

- Please update regarding application of K-Means Clustering:

In Number of Clusters 4 the Average Distortion is 1309.51, which is lower than cluster 3.

- Observations using Elbow Curve along with visuals:

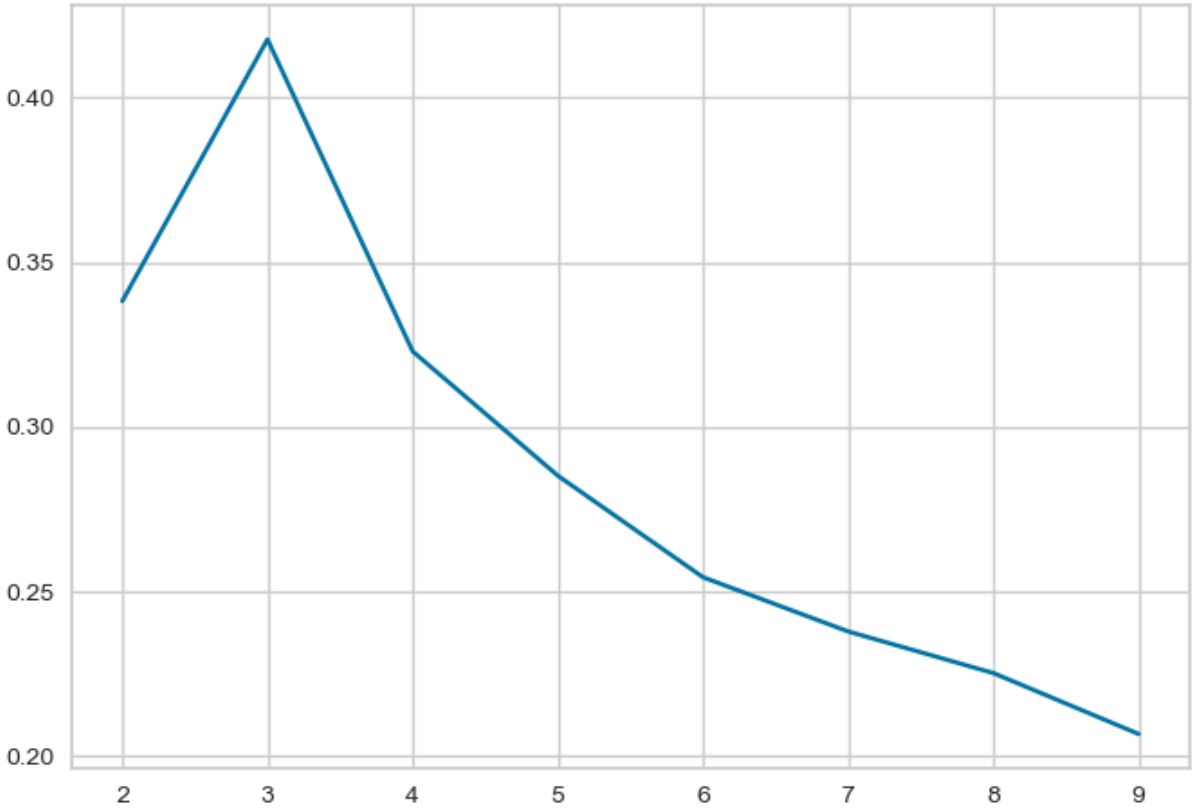
Appropriate value for k seems to be 3, 4, 5 or 6.



K-Means Clustering Technique

- Observations from Silhouette scores for different number of clusters:

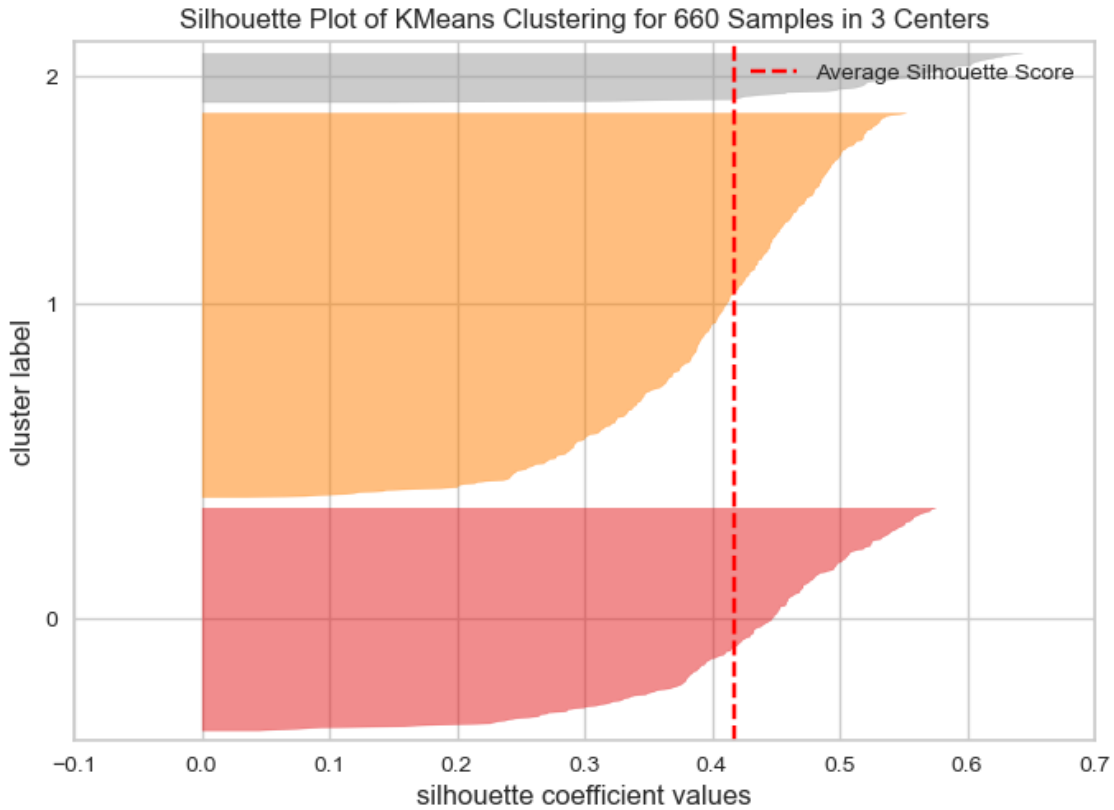
Silhouette score for 3 is higher than that for 4 and 5. So, we can take 3 as value of k, but we can visualize the silhouette scores for different number of clusters to find the optimal no. of clusters.



K-Means Clustering Technique

- Observations from Silhouette scores for different number of clusters:

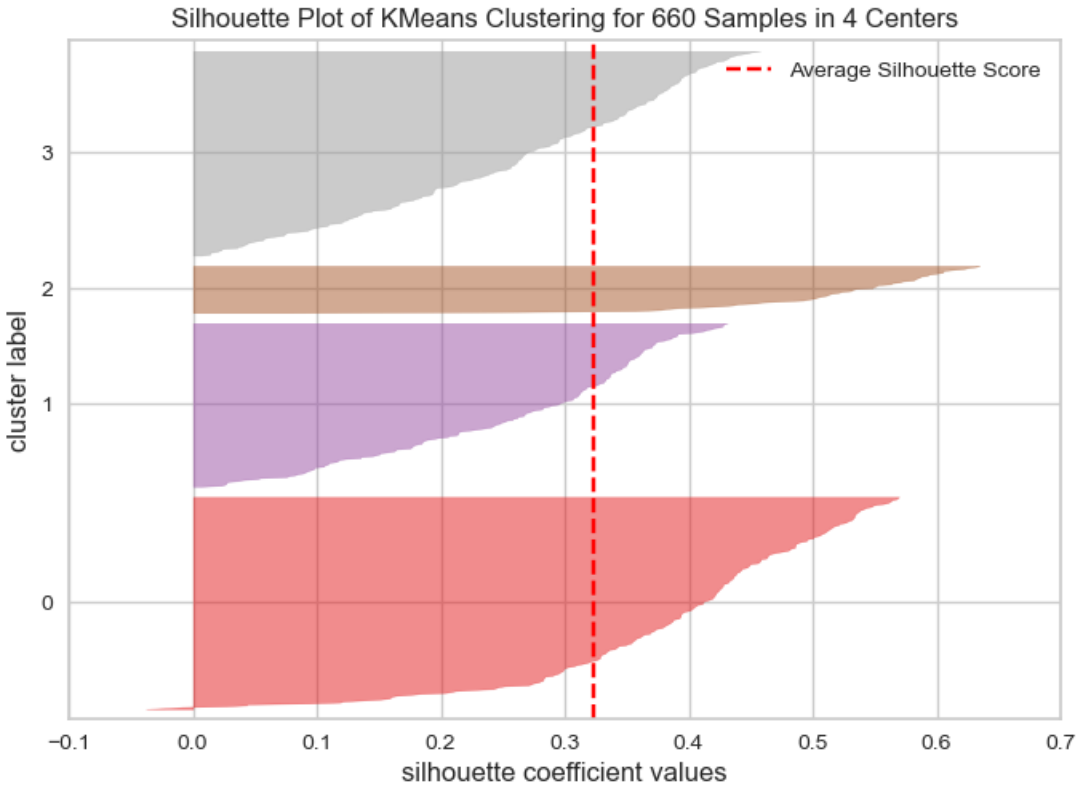
Silhouette score for 3 clusters: cluster with label 1 has a big number of observations (Width), and label 2 has a very small number of observations so k=3 will not be an appropriate value.



K-Means Clustering Technique

- Observations from Silhouette scores for different number of clusters:

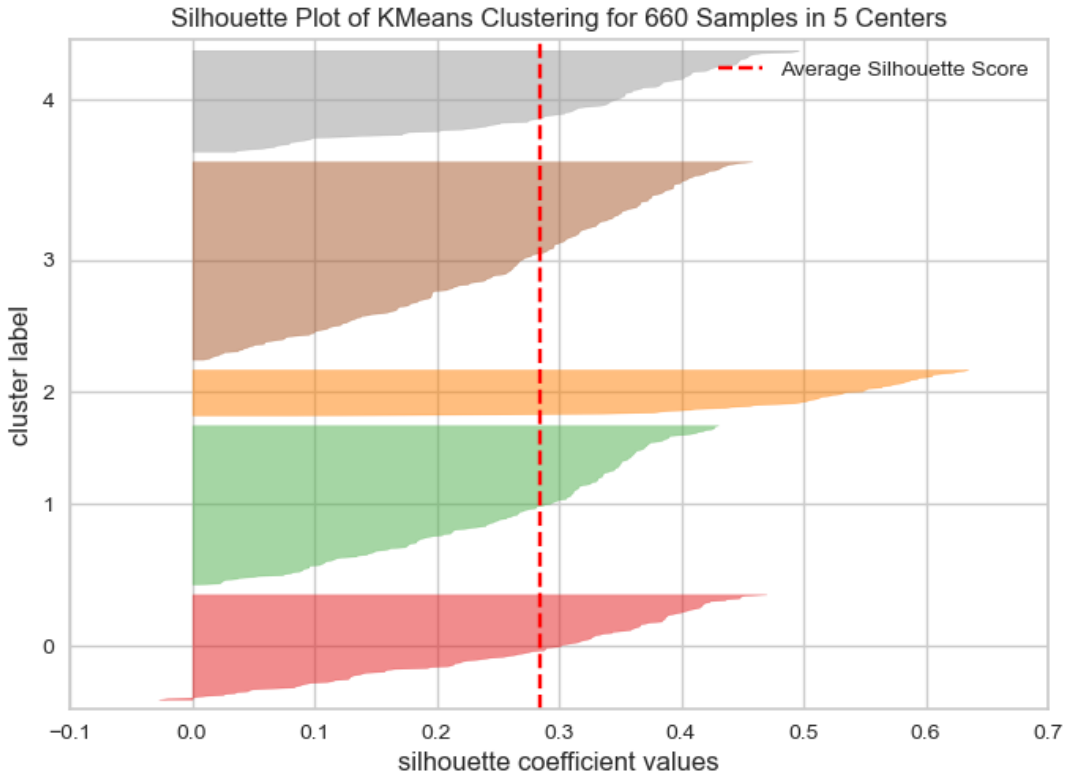
Silhouette score for 4 clusters: clusters meet the requirements. All the clusters have crossed Avg silhouette score, all the clusters have different silhouette scores and have different width sizes. So, let's take 4 as the appropriate no. of clusters(K) as the silhouette score is high enough for all the 4-clusters(above average silhouette score), and there is a knick at 4 in the elbow curve



K-Means Clustering Technique

- Observations from Silhouette scores for different number of clusters:

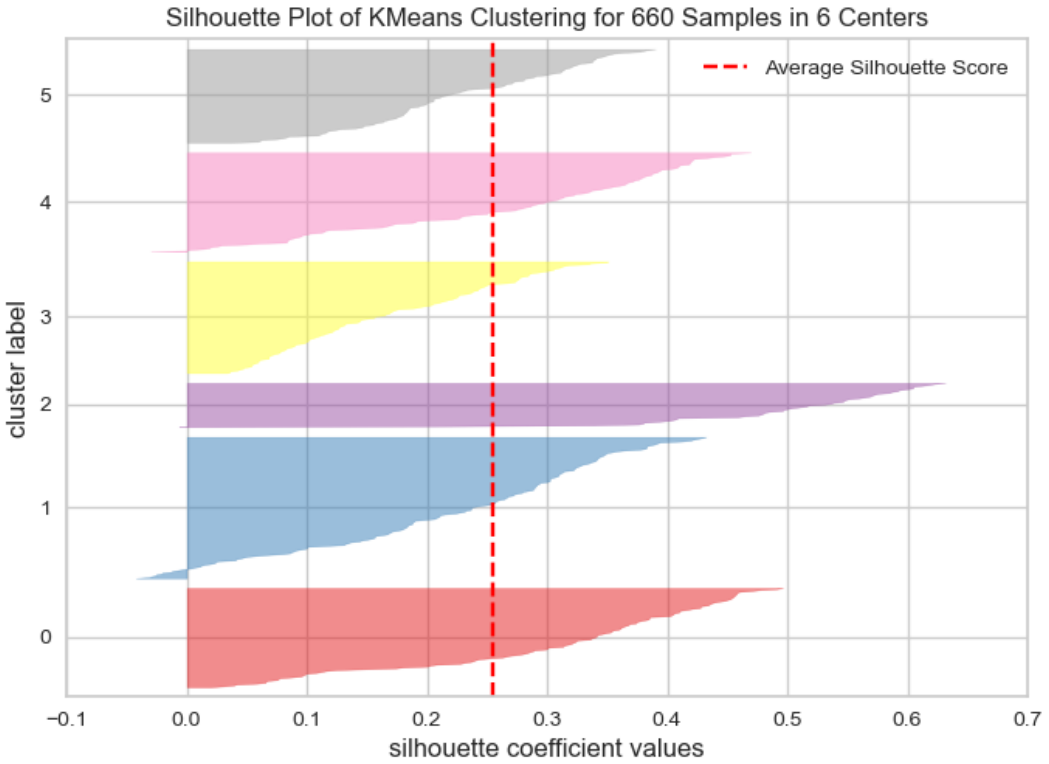
Silhouette score for 5 clusters: all the clusters have crossed Avg silhoutte score, all the clusters have different silhoutte score and have different width size, but the silhoutte score is low.



K-Means Clustering Technique

- Observations from Silhouette scores for different number of clusters:

Silhouette score for 6 clusters: all the clusters have crossed Avg silhouette score, all the clusters have different silhouette scores and have different width sizes, but there is no elbow at K=6, so k=6 will not be an appropriate value.



Hierarchical Clustering Technique

- Please update regarding application of Hierarchical Clustering:

The cophenetic correlation is maximum with Euclidean distance and average linkage with 0.8684

- Observations using different linkage methods:

Highest cophenetic correlation is 0.8684, which is obtained with average linkage and Euclidean distance.

	Linkage	Cophenetic Coefficient
4	ward	0.706719
0	single	0.715826
1	complete	0.833336
5	weighted	0.864225
3	centroid	0.865643
2	average	0.868423

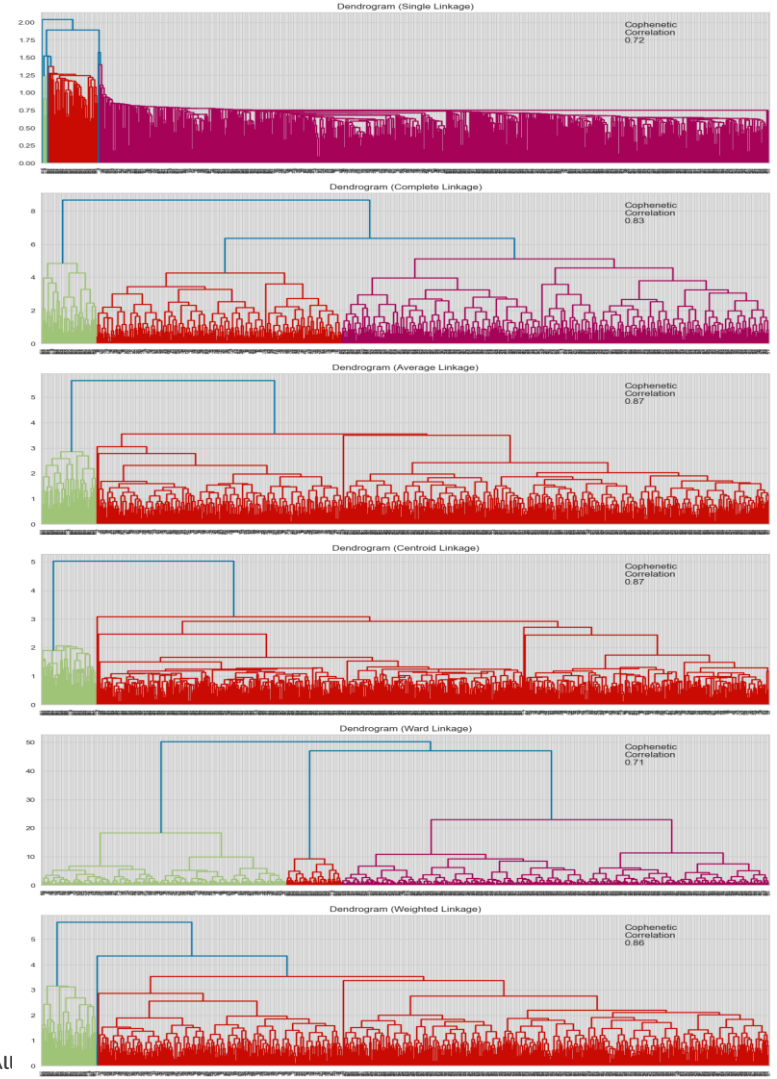
Hierarchical Clustering Technique

- Dendrograms for linkage methods used and their observations: Dendrogram with average linkage shows distinct and separate cluster tree.

The cophenetic correlation is highest for average and centroid linkage methods.

We will move ahead with Average linkage.

4 appears to be the appropriate number of clusters from the dendrogram for Average linkage.



Hierarchical Clustering Technique

Distance	Linkage	Cophenetic correlation
Euclidean	single	0.7158
Euclidean	complete	0.8333
Euclidean	average	0.8684
Euclidean	weighted	0.8642
Chebyshev	single	0.6993
Chebyshev	complete	0.7832
Chebyshev	average	0.8628
Chebyshev	weighted	0.8345

Distance	Linkage	Cophenetic correlation
Cityblock	single	0.7109
Cityblock	complete	0.8375
Cityblock	average	0.8648
Cityblock	weighted	0.8583
Mahalanobis	single	0.6829
Mahalanobis	complete	0.6051
Mahalanobis	average	0.7754
Mahalanobis	weighted	0.7777

- Observations from Cophenetic correlation for different combinations of distance and metrics:

Highest cophenetic correlation is 0.8684, which is obtained with Euclidean distance and average linkage

K-Means vs Hierarchical Clustering

HC_Clusters	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bankne	Total_visits_online	Total_calls_made	count_in_each_segment
0	55252.7309	12156.9506	2.403587	0.928251	3.556054	6.883408	223
1	56708.7600	141040.000	8.740000	0.600000	10.900000	1.080000	172
2	54791.4067	33564.7668	5.520725	3.492228	0.987047	2.010363	50
3	87073.0000	100000.000	2.000000	1.000000	1.000000	0.000000	215

On the Hierarchical cluster profiles, cluster 1 has the lowest customers with 50 customers and on the K-means cluster profiles cluster 2 has the lowest customers with 50 customers.

K-Means vs Hierarchical Clustering

HC_segments	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	55252.730942	12156.950673	2.403587	0.928251	3.556054	6.883408	223
1	56708.760000	141040.000000	8.740000	0.600000	10.900000	1.080000	50
2	54791.406736	33564.766839	5.520725	3.492228	0.987047	2.010363	386
3	87073.000000	100000.000000	2.000000	1.000000	1.000000	0.000000	1

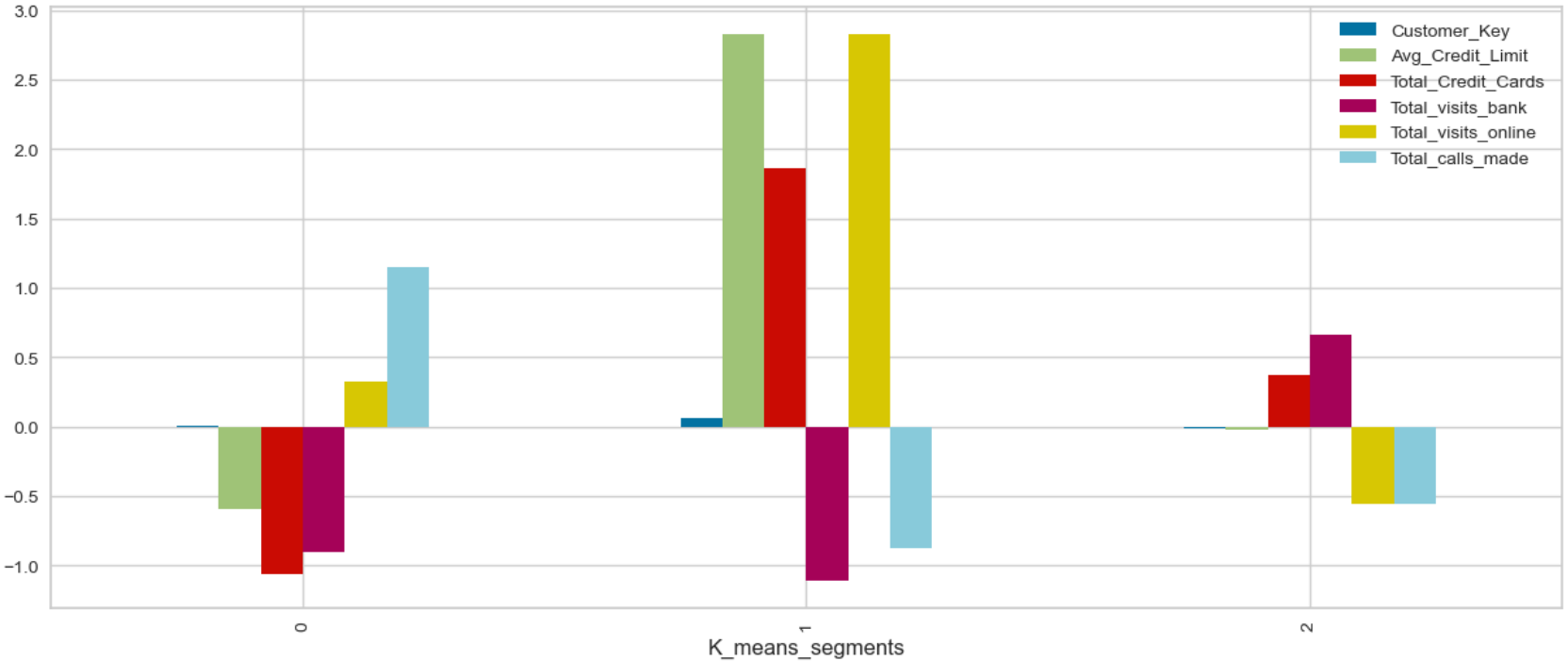
Eventhough on Hierarchical cluster profiles cluster 3 has only one and on the K-means cluster profiles cluster 3 has 215 the average credit limit, total credit card, total visit at the bank, total visit online and total call made to the bank are the same.

K-Means vs Hierarchical Clustering

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	55412.762332	12143.497758	2.403587	0.928251	3.551570	6.883408	223
1	56708.760000	141040.000000	8.740000	0.600000	10.900000	1.080000	50
2	54782.607235	33744.186047	5.511628	3.485788	0.989664	2.005168	387

- After combining cluster 3 & 1 from Kmean segments in cluster 2. It became more homogeneous clusters, with more variability between clusters.
- -
- Both Hierarchical Clustering & K-means Clustering are very similar except for a small change in cluster 2 on Total visit Online & Average Credit Limit.

K-Means vs Hierarchical Clustering



- After combining cluster 3 & 1 from Kmean segments in cluster 2. It became more homogeneous clusters, with more variability between clusters.

- -

- Both Hierarchical Clustering & K-means Clustering are very similar except for a small change in cluster 2 on Total visit Online & Average Credit Limit.

Great Learning

Happy Learning !