AllLife Bank

Project : Customer Segmentation
Course : UnSupervised Learning

Document Version : 1.0
Document Owner : Rahul Kulkarni
Document ID : Project 4 - USL - AllLife Bank (Customer Segmentation).pdf
Submission Date : 16th September 2023

Contents

Executive Summary

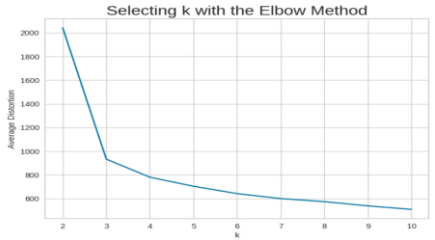
Executive Summary – Business Context

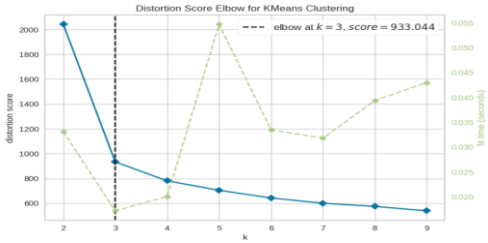
- **Business Context:** AllLife Bank wants to focus on its credit card customer base for the next financial year. The Marketing Research team has advised that market penetration can be improved, and hence the Marketing team is planning to run personalized campaigns to target new customers and upsell to existing customers. Moreover, the research team has provided insight that the customers perceive the Bank's Support Services to be of poor standards. Hence the Operations team wants to upgrade the service delivery model, to ensure that customer queries are resolved faster
- **The Problem Statement :** Provide recommendations to the bank on how to better market and service the customers
- **Solution Approach:** In order to resolve the above problem, we will undertake the following key tasks:
  - Perform a deep-dive on the previous Bank's dataset using libraries such as numpy and pandas for data manipulation, and seaborn and matplotlib for data visualisation
  - Perform exploratory data analysis on the dataset to deliver key findings and insights
  - Identify different segments in the existing customer dataset based on their spending patterns and past interaction
  - Build a model using K-Means & Hierarchical Clustering techniques that will categorise customers into distinct segments
  - Identify the key services that would need improvisation
  - Recommend opportunities for improvement that will help the bank to enhance customer services and boost potential customer acquisition

Executive Summary – K-Means Clustering, Profiling & Application

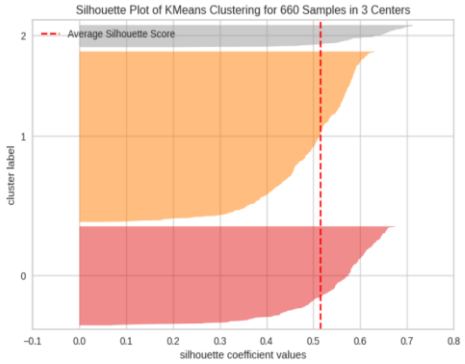
Following are the results of K-Means clustering on the dataset.

- **Elbow Curve & Visuals :** Based on the 9 clusters plotted for K-Means, the elbows are formed at clusters 3,4,5,6 and 7 with distortion scores of 933, 800, 700, 650, 600 respectively.
- **Silhoutte Scores:** The Silhouette score of 0.516 indicates that using 3 clusters will provide the best results
- **Silhoutte Coefficient:** Using k=3, all 3 clusters cross the average silhouette score, and are of different widths and have distinct silhouette scores
- **Optimal Number of Clusters :** 3 – Based on the above, 3 seems to be the ideal number of clusters for K-Means clustering.





No. Of Clusters	Silhouette Score
2	0.41842496663215445
3	0.5157182558881063
4	0.3556670619372605
5	0.2717470361089752
6	0.255906765297388
7	0.24798644656011146
8	0.2414240144760896
9	0.2184645050755029



Executive Summary – K-Means Clustering, Profiling & Application

- **K-Means Cluster Profiling:** K-Means has profiled the customers into 3 segments – 0, 1 & 2 with counts of 224, 50 and 386 customers respectively

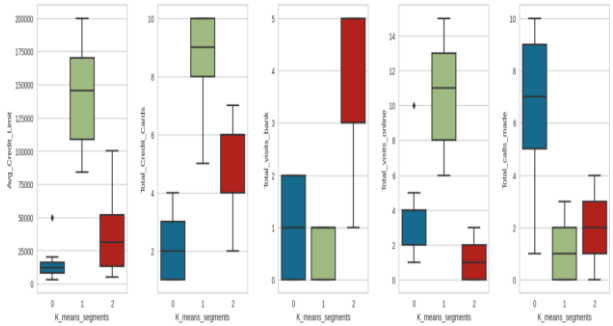
K_means_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	12174.107143	2.410714	0.933036	3.553571	6.870536	224
1	141040.000000	8.740000	0.600000	10.900000	1.080000	50
2	33782.383420	5.515544	3.489637	0.981865	2.000000	386

- **Customer Segmenation using K-Means Clustering on Dataset :**

- **Segment 0** - Customers who prefer phone-banking, have the lowest total number of credit cards and average credit card limit. However, they do often visit the bank and use the bank's online services

- **Segment 1** – Customers who prefer online interactions with their bank, have a much higher average credit limit and have a higher number of credit cards. They rarely visit the bank and/or make phone calls than compared to other segments

- **Segment 2** – Customer who prefer in-person visits to the bank, have a mid-range number of total credit cards and an average credit card limit. They are the lowest in their online presence than compared to the other segment of customers



Executive Summary – Hierarchical Clustering, Profiling & Application

Following are the results of Hierarchical clustering on the dataset.

- **Distance Metrics Used:** Euclidean, Chebyshev, Mahalanobis and Cityblock
- **Linkage Methods Used:** Single, Complete, Average and Weighted
- **Highest Cophentic Correlation:** Euclidean Average linkage yielded a maximum correlation of 0.8977080867389372
- **Best Dendogram:** Euclidean Average linkage Dendogram with a Cophenetic Coefficient of 0.897708 shows distinct and separate clusters
- **Optimal Number of Clusters :** 3 – Based on the dendogram, 3 seems to be the optimal number of clusters for Hierarchical clustering.
- **Hierarchical Clustering Model Used:** Agglomerative Model using 3 clusters with Euclidean Average Linkage to yield customer segmentation

#	Distance	Linkage	Cophenetic Correlation
1	Euclidean	single	0.7391220243806552
2	Euclidean	complete	0.8599730607972423
3	Euclidean	average	0.8977080867389372
4	Euclidean	weighted	0.8861746814895477
5	Euclidean	centroid	0.8939385846326323
6	Euclidean	ward	0.7415156284827493
7	Chebyshev	single	0.7382354769296767
8	Chebyshev	complete	0.8533474836336782
9	Chebyshev	average	0.8974159511838106
10	Chebyshev	weighted	0.8913624010768603
11	Mahalanobis	single	0.7058064784553605
12	Mahalanobis	complete	0.6663534463875359
13	Mahalanobis	average	0.8326994115042136
14	Mahalanobis	weighted	0.7805990615142518
15	Cityblock	single	0.7252379350252723
16	Cityblock	complete	0.8731477899179829
17	Cityblock	average	0.896329431104133
18	Cityblock	weighted	0.8825520731498188

AgglomerativeClustering

```
AgglomerativeClustering(affinity='euclidean', linkage='average', n_clusters=3)
```

Executive Summary – Hierarchical Clustering, Profiling & Application

- **Hierarchical Cluster Profiling:** Hierarchical Clustering has profiled the customers into 3 segments – 0, 1 & 2 with counts of 387, 50 and 223 customers respectively

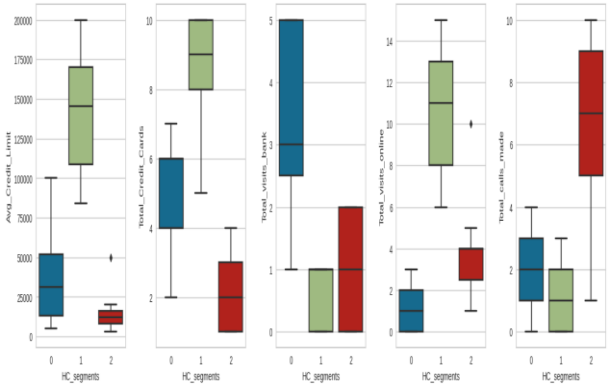
HC_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	33713.178295	5.511628	3.485788	0.984496	2.005168	387
1	141040.000000	8.740000	0.600000	10.900000	1.080000	50
2	12197.309417	2.403587	0.928251	3.560538	6.883408	223

- **Customer Segmenation using Hierarchical Clustering on Dataset :**

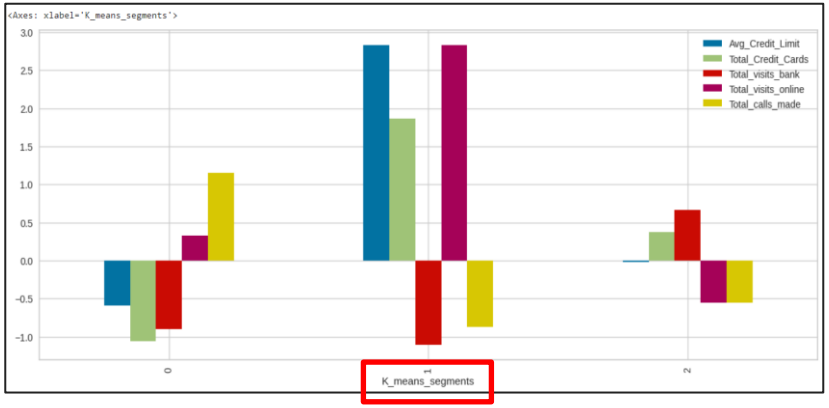
- **Segment 0** - Customer prefer in-person visits to the bank, have a mid-range number of total credit cards and a mid-range average credit card limit. They are the lowest in their online presence than compared to the other segment of customers

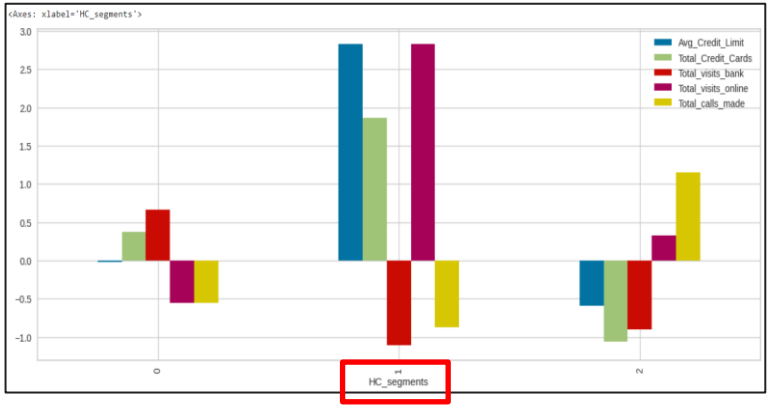
- **Segment 1** – Customers prefer online interactions with their bank, have a much higher average credit limit and have a higher number of credit cards. They rarely visit the bank and/or make fewer phone calls than compared to other segments

- **Segment 2** – Customers prefer phone-banking, have the lowest total number of credit cards and the lowest average credit card limit. However, they do often visit the bank and use the bank's online services



Executive Summary – K-Means vs Hierarchical Clustering





- **Segment 0:** For K-Means, customers prefer phone-banking and have the lowest total number of credit cards and lowest average credit limit. Whereas, for Hierarchical, the customers prefer in-person visits to the bank, have a mid-range number of total credit cards and a mid-range average credit card limit.
- **Segment 1:** For both, K-Means & Hierarchical Clustering, customers prefer online interactions with their bank, have a much higher average credit limit and have a higher number of credit cards. They rarely visit the bank and/or make fewer phone calls than compared to other segments
- **Segment 2:** For K-Means, customers prefer in-person visits to the bank, have a mid-range number of total credit cards and a mid-range of average credit card limit. Whereas, for Hierarchical, the customers prefer phone-banking, have the lowest total number of credit cards and the lowest average credit card limit

Executive Summary – K-Means vs Hierarchical Clustering

- **Appropriate Number of Clusters:** 3 – For both, K-Means & Hierarchical Clustering, 3 clusters were identified as the optimal number of clusters to yield appropriate customer segmentation results
- **Number Of Observations in similar Clusters:** For both, K-Means & Hierarchical Clustering, Segment 1 had 50 identical observations – higher number of online visits, higher average credit limit and have a higher total number of credit cards

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

HC_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	33713.178295	5.511628	3.485788	0.984496	2.005168	387
1	141040.000000	8.740000	0.600000	10.900000	1.080000	50
2	12197.309417	2.403587	0.928251	3.560538	6.883408	223

- **Execution Time:** K-Means took 15.3 ms than compared to Agglomerative Clustering that took 37.7 ms

```
CPU times: user 24.2 ms, sys: 794 µs, total: 25 ms
Wall time: 15.3 ms
KMeans
KMeans(n_clusters=3, random_state=1)
```

```
CPU times: user 19.5 ms, sys: 0 ns, total: 19.5 ms
Wall time: 37.7 ms
AgglomerativeClustering
AgglomerativeClustering(affinity='euclidean', linkage='average', n_clusters=3)
```

- **Distinct Clusters:** Both, K-Means & Hierarchical Clustering yielded distinct clusters

Executive Summary - Conclusion

- Both K-Means & Hierarchical Clustering Models have produced 3 distinct customer segments
    - Phone Banking Customer Segment - customers who prefer to transact via the telephony channel
    - Online Banking Customer Segment - customers who prefer online transactions with their bank
    - In-Person Customer Segment - customers who prefer physically visiting the bank to perform transactions
- The above segmentation can be used by the Marketing team to deliver personalised marketing campaigns based on the customer's banking behaviours and preference
- The above segmentation can be used by the Operations team to improve their service delivery model by optimising the underlying IT Applications & Infrastructure to ensure that customer queries are resolved faster based on the customer's banking behaviours and preference

Executive Summary - Key Actionable Business Insights

- **Summarised Key Actionable Business Insights:**
  - There are 3 distinct segments of customers produced by K-Means & Hierarchical Clustering Methods:
    - Phone Banking Customers: These customers prefer to transact via the telephony channel, and have the lowest total number of credit cards and the lowest average credit card limit
    - Online Customers: These customers prefer online transactions with their bank, have a much higher average credit limit and have a higher number of credit cards. They rarely visit the bank and/or make fewer phone calls
    - In-Person Customers: These customers prefer in-person visits to the bank, and have a mid-range number of total credit cards and a mid-range of average credit card limit
  - A Personalised Preference Campaign Strategy should be put in place to cater to the above 3 segments of customers:
    - Phone Banking & Online Customers: Email, SMS, Instant Messaging Services / Social Media Applications should be used to target these segment of customers
    - In-Person Customers: Mail notifications, Promotion Flyers, Marketing Leaflets, Posters at Bank branches etc should be used to target this segment of customers.

Executive Summary - Our Recommendation

Based on our key observations and insights, we recommend the following areas of improvement / opportunities that will drive business growth and lead to a better customer experience

- **Implement Customer Satisfaction Survey:** The Bank should initiate a targeted Customer Satisfaction Survey to understand customer pain points for the current services and implement the findings to improve retention ratio of customers
- **Implement Customer Incentivisation Scheme:** Incentivising customers by offering them cashback schemes and discounts / vouchers on credit card purchases will encourage frequent spending and will drive customer gowth and increase revenue
- **Implement Tier based Rewards:** The Bank should introduce a Tier based Loyalty & Rewards Scheme for credit card purchases. Cumulative loyalty points above a certain threshold will promote the customer to a new tier, that will offer specific rewards such as First-Class Lounge access at Airports, Spa & Well-Being discounts etc. This will enable the Marketing team to upsell new products to existing customers and drive customer retention

Business Problem Overview & Solution Approach

Business Problem Overview and Solution Approach

- **Business Context:** AllLife Bank wants to focus on its credit card customer base for the next financial year. The Marketing Research team has advised that market penetration can be improved, and hence the Marketing team is planning to run personalized campaigns to target new customers and upsell to existing customers. Moreover, the research team has provided insight that the customers perceive the Bank's Support Services to be of poor standards. Hence the Operations team wants to upgrade the service delivery model, to ensure that customer queries are resolved faster
- **The Problem Statement :** Provide recommendations to the bank on how to better market and service the customers
- **Solution Approach:** In order to resolve the above problem, we will undertake the following key tasks:
  - Perform a deep-dive on the previous Bank's dataset using libraries such as numpy and pandas for data manipulation, and seaborn and matplotlib for data visualisation
  - Perform exploratory data analysis on the dataset to deliver key findings and insights
  - Identify different segments in the existing customer dataset based on their spending patterns and past interaction
  - Build a model using K-Means & Hierarchical Clustering techniques that will categorise customers into distinct segments
  - Identify the key services that would need improvisation
  - Recommend opportunities for improvement that will help the bank to enhance customer services and boost potential customer acquisition

Data Overview & Analysis

Data Overview & Analysis

- The dataset has the following Data-Structure:

#	Columns	Data-type	Total Rows	Description
1	Sl_No	Integer 64	10127	Primary key of the records
2	Customer_Key	Integer 64	10127	Customer identification number
3	Average_Credit_Limit	Integer 64	10127	Average credit limit of each customer for all credit cards
4	Total_Credit_Cards	Integer 64	10127	Total number of credit cards possessed by the customer
5	Total_visits_bank	Integer 64	10127	Total number of Visits that customer made (yearly) personally to the bank
6	Total_visits_online	Integer 64	8608	Total number of visits or online logins made by the customer (yearly)
7	Total_calls_made	Integer 64	9378	Total number of calls made by the customer to the bank or its customer service department (yearly)

- Shape of the Dataset: Total No. Of Columns - 7 | Total No. Of Rows - 660
- Column Data-types: All 7columns are of numerical data types (Integer64)
- Missing & Junk Values: There are no missing values in the dataset

Data Overview & Analysis (Cont'd)

- Duplicate Values: There are 5 Customer_Key values (37252, 47437, 50706, 96929 & 97935) that are duplicated in the dataset. However, each duplicated Customer_Key has unique corresponding attribute values. This could imply that the same customer holds multiple credit cards with varying average credit limits and has a different personal preferential behaviour whilst banking with those cards. Thus these duplicate values cannot be treated as redundant and cannot be eliminated

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

Data Overview & Analysis (Cont'd)

- Statistical Summary: Following is the statistical summary of the dataset

	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
Total_Credit_Cards	660.0	4.706061	2.167835	1.0	3.0	5.0	6.0	10.0
Total_visits_bank	660.0	2.403030	1.631813	0.0	1.0	2.0	4.0	5.0
Total_visits_online	660.0	2.606061	2.935724	0.0	1.0	2.0	4.0	15.0
Total_calls_made	660.0	3.583333	2.865317	0.0	1.0	3.0	5.0	10.0

Data Overview & Analysis – Key Observations & Insights

- **Key Observations:**
  - There are no missing and junk values in the dataset
  - There are 5 Customer_Key values (37252, 47437, 50706, 96929 & 97935) that are duplicated in the dataset.
  - 24.84% of Customers (164 Customers) have a Credit Limit between $48K - $200K
  - The minimum and maximum average credit card limit is $3K and $200K respectively, whereas the mean is circa $34.5K
  - The minimum and maximum total credit cards is 1 and 10 cards respectively, whereas the mean is approx. 4 cards
  - The minimum and maximum number of visits to the bank is 0 and 65 respectively, whereas the mean number of bank visits is approx. 5 visits
  - The minimum and maximum online banking activity period is 0 and 15 times respectively, whereas the mean online activity period is approx. 2.6
  - The minimum and maximum customer calls with the bank is 0 and 10 calls respectively, whereas the mean number of calls is approx. 3.5

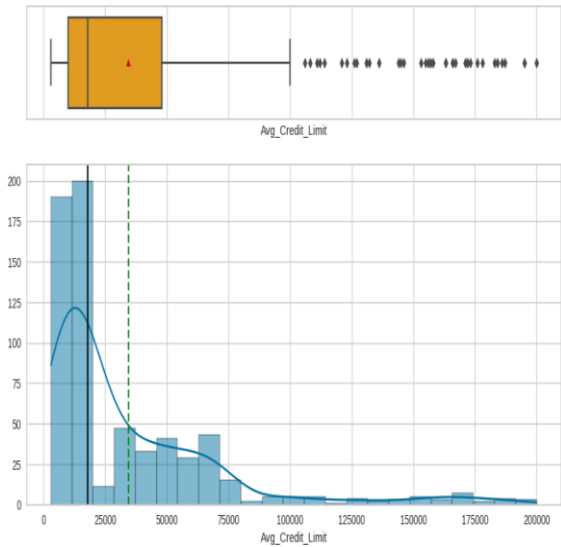
Data Overview & Analysis – Key Observations & Insights

- **Key Insights:**
  - There are 5 duplicate Customer_Key values in the dataset. However, each duplicated Customer_Key has unique corresponding attribute values, which implies that the same customer holds multiple credit cards with varying average credit limits and has a different baking persona whilst banking with those cards. Hence these duplicate values cannot be treated as redundant and cannot be eliminated.
  - The duplicated customer values have an extreme range of Avg_Credit_Card_Limit:
    - 97935 has an Avg_Credit_Limit of $17K and $187K, 96929 has an Avg_Credit_Limit of $13K and $67K
    - 50706 has an Avg_Credit_Limit of $44K and $60K, 47437 has an Avg_Credit_Limit of $17K and $100K
    - 37252 has an Avg_Credit_Limit of $6K and $59K

EDA - Univariate Analysis

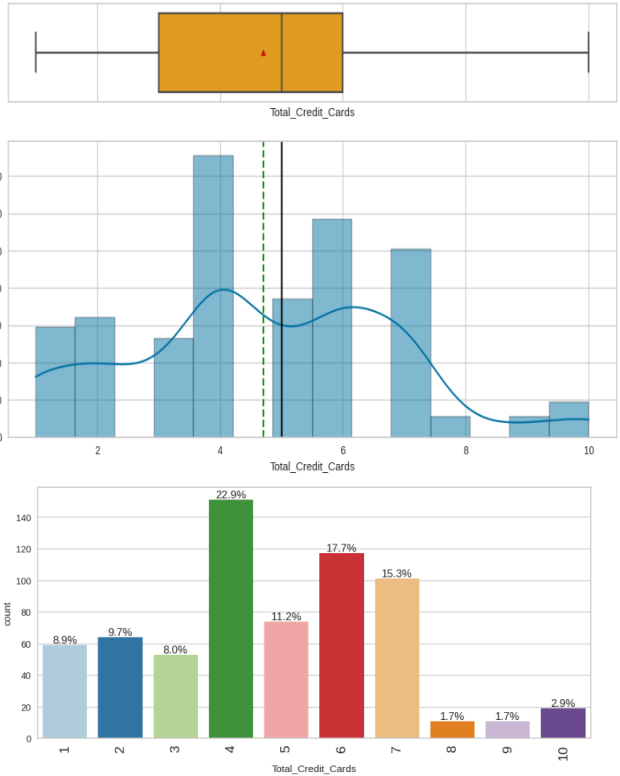
EDA - Univariate Analysis – Average Credit Limit

- **Avg_Credit_Limit:**
  - Highest: There are circa 200 customers that have an average credit limit of circa $22K
  - Minimum: The minimum average credit limit is circa $3K
  - Q1: 25% of the customers have an average credit limit of less than $10K
  - Q3: 75% of the customers have an average credit limit of less than $48K
  - Upper Fence: The upper fence for the average credit limit is circa $100K
  - Maximum: The maximum average credit limit is circa $200K
  - Mean: On average, customers have an average credit limit is circa $34,574
  - Median: The median average credit limit is circa $18K
  - Outliers: There are circa 6% - 39 outliers ranging $106 to $200K
  - Skewness: From the plot, it can be observed that the graph has a right skewed distribution



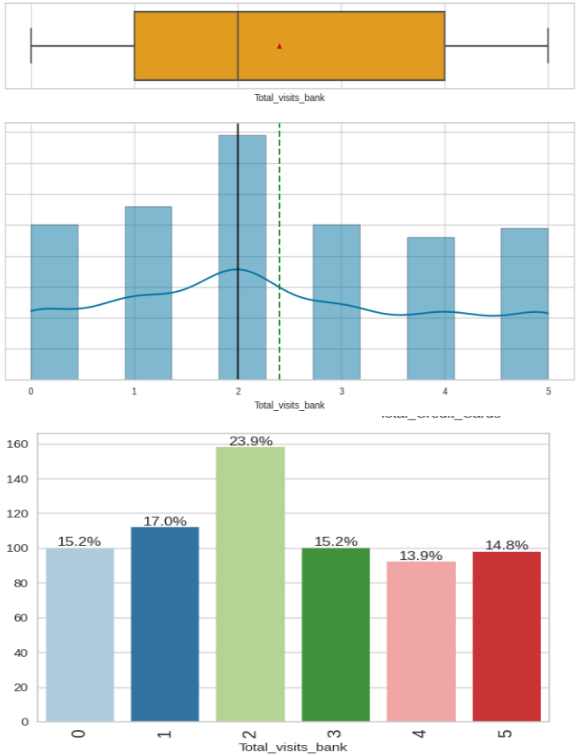
EDA - Univariate Analysis – Total Credit Cards

- **Total_Credit_Cards:**
  - Highest: There are circa 150 customers that have approx 4 credit cards
  - Minimum: The minimum number of credit cards held by a customer is 1
  - Q1: 25% of the customers have less than 3 credit cards
  - Q3: 75% of the customers have less than 6 credit cards
  - Minimum: The minimum number of credit cards held by a customer is 1
  - Maximum: The maximum number of credit cards held by a customer is 10
  - Mean: On average, customers have circa 4 credit cards
  - Median: The median number of credit cards is 5
  - Outliers: There are no outliers
  - Skewness: From the plot, it can be observed that the graph has a near to normal distribution



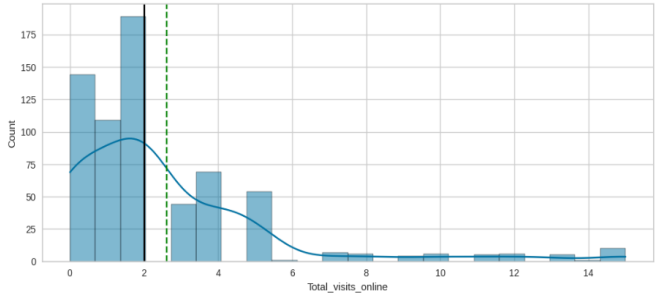
EDA - Univariate Analysis – Total Bank Visits

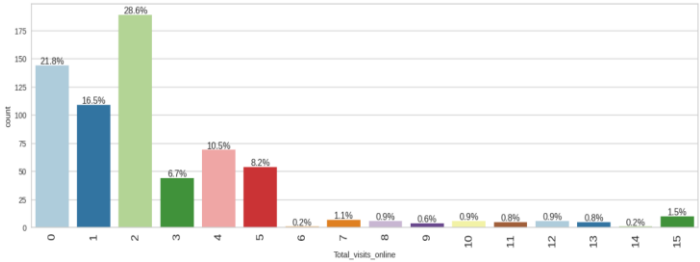
- **Total_Visits_Bank:**
  - Highest: There are circa 170 customers that have visited the bank twice
  - Minimum: The minimum visits to the bank is 0
  - Q1: 25% of the customers have visited the bank at least 1 time
  - Q3: 75% of the customers have visited the bank less than 4 times
  - Maximum: The maximum visits to the bank is 5 times
  - Mean: On average, customers have been to the bank 2.5 times
  - Median: The median visits to the bank is 2
  - Outliers: There are no outliers
  - Skewness: From the plot, it can be observed that the graph has a near to normal distribution



EDA - Univariate Analysis – Total Visits Online

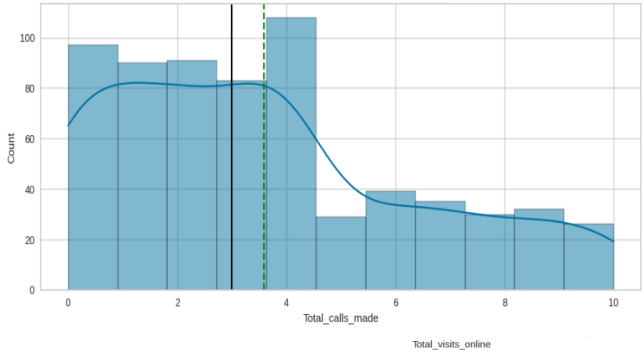
- **Total_Visits_Online**
  - Highest: There are circa 185 customers that have been online to the bank 2 times
  - Minimum: The minimum online visits to the bank is 0
  - Q1: 25% of the customers have been online to the bank atleast once
  - Q3: 75% of the customers have been online to the bank less than 4 times
  - Maximum: The maximum online visits to the bank is 15 times
  - Mean: On average, customers have been online circa 2.6 times
  - Median: The median online visits to the bank is 2 times
  - Outliers: There are several outliers ranging from 12 to 15 online visits
  - Skewness: From the plot, it can be observed that the graph has is right skewed

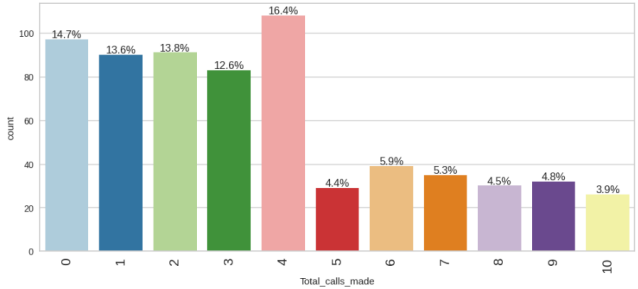




EDA - Univariate Analysis – Total Calls Made

- **Total_Calls_Made:**
  - Highest: There are circa 90 customers that called the bank 4 times
  - Minimum: The minimum calls made to the bank are 0
  - Q1: 25% of the customers have called the bank at least once
  - Q3: 75% of the customers have called the bank 4 times
  - Maximum: The maximum number of calls made to the bank are 10
  - Mean: On average, customers have made circa made 3.58 calls
  - Median: The median calls made to the bank are 2
  - Outliers: There are no outliers
  - Skewness: From the plot, it can be observed that the graph is right skewed

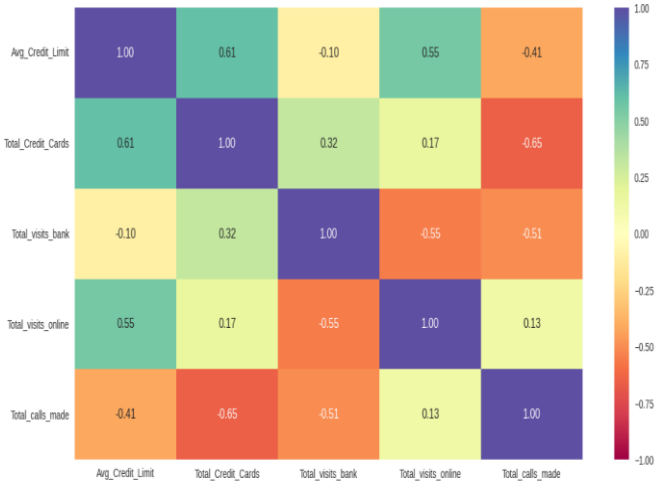




EDA - Bivariate Analysis

EDA - Bivariate Analysis – Correlation Check

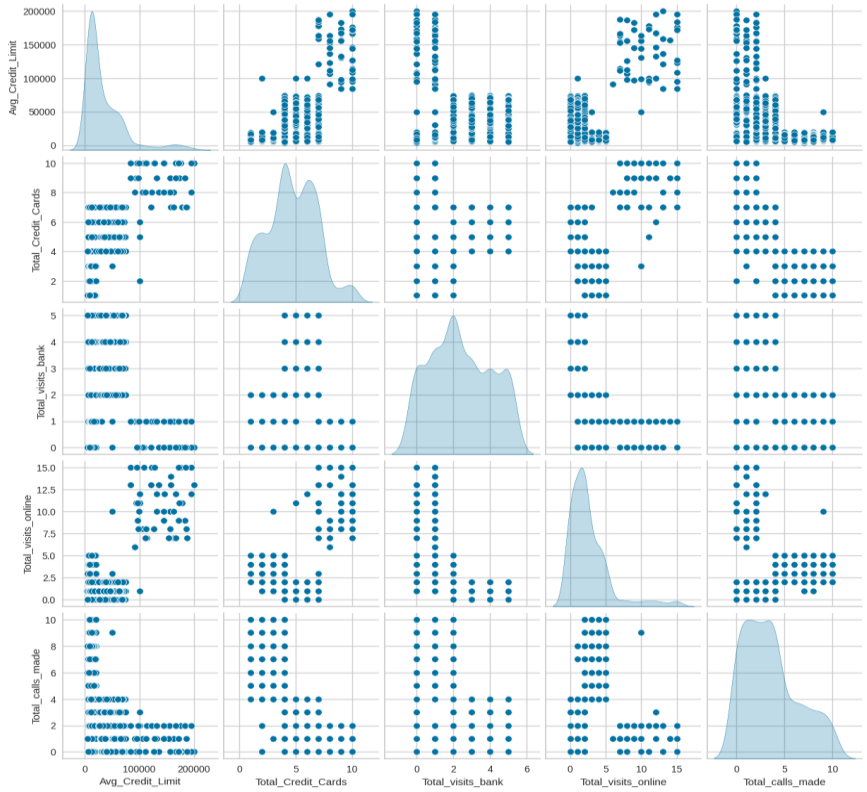
- **Correlation Amongst Variables :**
  - There is a strong positive correlation of 0.61 between Total_Credit_Cards and Avg_Credit_Limit, which implies that with increasing number of credit cards there is an increase in the credit limit for the customer
  - There is a strong positive correlation of 0.55 between Total_Visits_Online and Avg_Credit_Limit, which could imply that customers could be often transacting online to check their credit balance, credit usage or history or paying off their outstanding credit. This could also imply that customers are having issues with their cards and often use the banks online services to raise queries and / or resolve it.



  - There is a weak positive correlation of 0.32 between Total_Visits_Bank and Total_Credit_Cards, which could imply that with the increasing number of credit cards customers could be often visiting the bank to either pay off their consolidated outstanding credit, or check their total credit balance, credit usage or history across all the cards. This could also imply that customers are having issues with their cards and often visit the banks to raise queries and / or resolve oustanding issues.

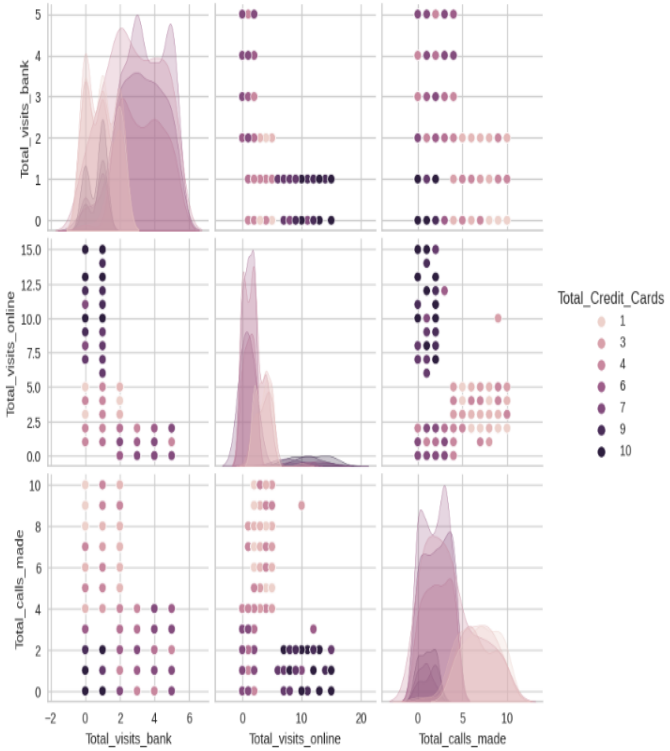
EDA - Bivariate Analysis – Correlation Check (Cont'd)

- **Correlation Amongst Variables (Cont'd) :**
  - There is an insignificant positive correlation of 0.17 between Total_Visits_Online and Total_Credit_Cards, which could imply that with the increasing number of credit cards customers could be often using the bank's online services to either pay off their consolidated outstanding credit, or check their total credit balance, credit usage or history across all the cards. This could also imply that customers are having issues with their cards and often use the banks online services to raise queries and / or resolve outstanding issues
  - There is an insignificant positive correlation of 0.13 between Total_Visits_Online and Total_Calls_Made, which could imply that customers have been trying to use both channels (digital and telephony) to raise and resolve outstanding issues



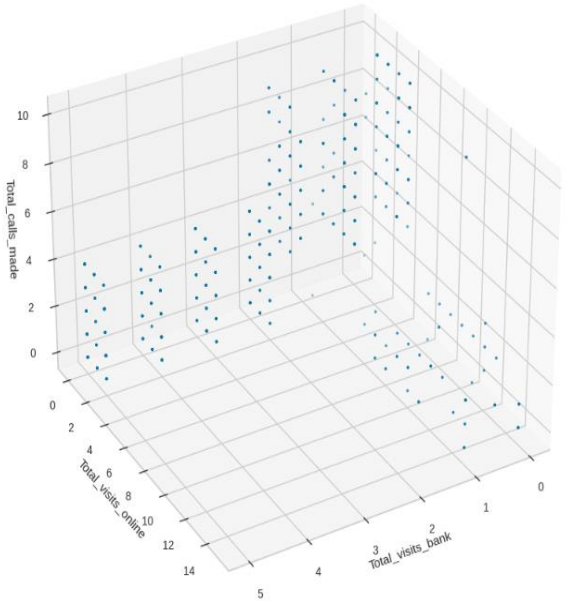
EDA - Bivariate Analysis – Correlation Check (Cont'd)

- **Correlation Amongst Variables (Cont'd) :**
  - There is a strong negative correlation of -0.65 between Total_Credit_Cards and Total_Calls_Made, which implies that higher the number of credit cards lesser are the total number of calls made by the customer. This could imply that the customers having higher number of credit cards are proficient in using the bank's online services and hence make fewer calls to the bank to resolve issues or queries. On the other hand, customers having fewer credit cards often call the bank to resolve issues or queries
  - There is a strong negative correlation of -0.55 between Total_Visits_Bank and Total_Visits_Online, which implies that the if the customer uses the bank's online services then he rarely visits the bank personally and vice-versa. i.e. If the customer visits the bank then the bank responds to the customers issues, queries or services, which therefore leads to lower usage of bank's online services. On the contrary, it could also be implied that the higher usage of bank's online services leads to lower number of bank visits i.e. The bank's online services resolves customer issues and queries effectively and hence the number of visits to the bank are minimised.



EDA - Bivariate Analysis – Correlation Check (Cont'd)

- **Correlation Amongst Variables (Cont'd) :**
  - There is a weak negative correlation of -0.41 between Total_Calls_Made and Avg_Credit_Limit, which implies that customers are more likely to call regarding queries on their average credit limit i.e. As, average credit limit is increased the total calls made decreases
  - There is an insignificant negative correlation of -0.10 between Total_Visits_Bank and Avg_Credit_Limit, which implies that customers are more likely to visit the bank regarding queries on their average credit limit i.e. As, average credit limit is increased the total number of visits to the bank decreases



EDA – Uni & Bivariate Analysis – Key Observations & Insights

- **Key Observations:**
  - 22.9% (151) Customers have 4 credit cards, 17.7% (117) have 6 credit cards, 15.3% (101) have 7 credit cards and 11.2% (74) have 5 credit cards
  - 9.7% (64) customers have 2 credit cards, 8.9% (59) have 1 credit card, 8% (53) have 3 credit cards, 2.9% (19) have 10 credit cards, 1.7%(11) have 8 credit cards and the remaining 1.7% (11) have 9 credit cards
  - 23.9% (158) Customers have visited the bank twice, 17.0% (112) have visited the bank only once, 15.2% (100) have visited the bank thrice, 13.9% (92) have visited the bank 4 times and 14.8% (98) have visited the bank 5 times
  - 28.6% (189) Customers have visited been online twice, 16.5% (109) have been online only once, 6.7% (44) have been online 3 times, 10.5% (69) have been online 4 times, and 8.2% (54) have been online 5 times. 21.8% (144) have never transacted online with the bank
  - 16.4% (108) Customers have called the bank 4 times, 13.8% (90) have called the bank twice, 13.6% (90) have called the bank only once, and 12.6% (83) have called the bank thrice. Approx. 28% (191) customers have called the bank between 5-10 times and 14.7% (97) of customers have never called the bank

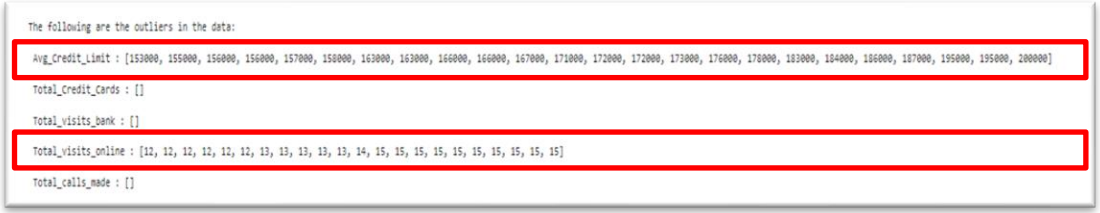
EDA – Uni & Bivariate Analysis – Key Observations & Insights

- **Key Insights:**
  - Majority of customers circa 67% have 4 to 7 credit cards, 26.6 % of customer have 1 to 3 credit cards, whereas 2.9% of customers have 10 credit cards. Approx 3.4% of customers have 8-9 credit cards
  - 15% of customers have never visited the bank, whereas the other 15% have visited the bank 5 times. Most customers have visited the bank only twice
  - Circa 62% of the customers have been only between 1-4 times, whereas 21% have never been banking online
  - Circa 54% of the customers have made calls to the bank between 1-4 times, whereas 14.7% have never used phone-banking
  - There is a strong positive correlation between Total_Credit_Cards, Total_Visits_Online and Avg_Credit_Limit which implies that having higher number of credit cards often use the banks' online banking services to either pay off their consolidated outstanding credit, or check their total credit balance, credit usage or history across all the cards. This could also imply that customers are having issues with their cards and often visit the banks to raise queries and / or resolve oustanding issues.

Data Preprocessing

Data Preprocessing – Key Observations & Insights

- **Outlier Detection:**
  - Avg_Credit_Limit has 24 outliers
  - Total_Visits_Online has 22 outliers



- **Columns:** We have dropped columns Sl_No & Customer_Key since they do not add any value for model building
- **Data Scaling:** Following is the output of the normalised data using standard scaler

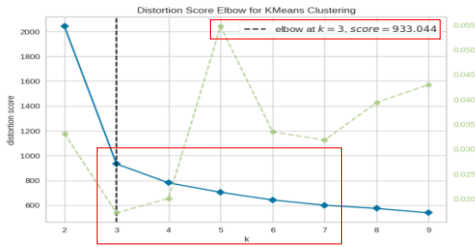
	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	1.740187	-1.249225	-0.860451	-0.547490	-1.251537
1	0.410293	-0.787585	-1.473731	2.520519	1.891859
2	0.410293	1.058973	-0.860451	0.134290	0.145528
3	-0.121665	0.135694	-0.860451	-0.547490	0.145528
4	1.740187	0.597334	-1.473731	3.202298	-0.203739

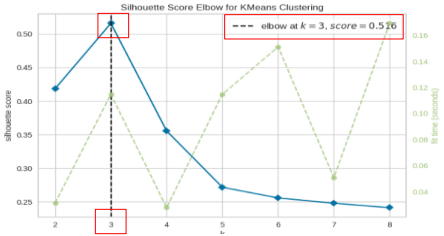
K-Means Clustering

K-Means Clustering, Profiling & Application

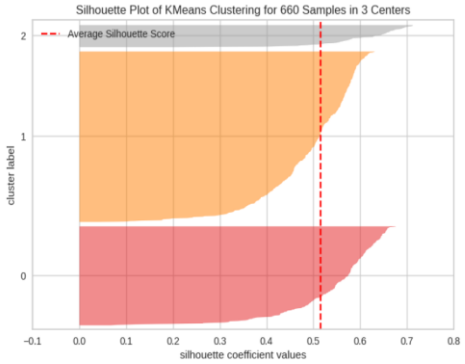
We have used 9 clusters (2 to 10) for K-Means clustering on the dataset.

- **Elbow Curve & Visuals :** Based on the 9 clusters plotted for K-Means, the elbows are formed at clusters 3,4,5,6 and 7 with distortion scores of 933, 800, 700, 650, 600 respectively.
- **Silhoutte Scores:** Using the Silhouette method, the silhouette score of 0.516 indicates that using 3 clusters will provide the best results
- **Silhoutte Coefficient:** Of 3,4,5,6 & 7 clusters, for k=3, all 3 clusters crossed the average silhouette score, are of different widths and have distinct silhouette scores
- **Optimal Number of Clusters :** 3 – Based on the above, 3 seems to be the ideal number of clusters for K-Means clustering
- **K-Means Final Model:** 3 clusters & Wall time = 15.9 ms





No. Of Clusters	Silhouette Score
2	0.41842496663215445
3	0.5157182558881063
4	0.3556670619372605
5	0.2717470361089752
6	0.255906765297388
7	0.24798644656011146
8	0.2414240144760896
9	0.2184645050755029



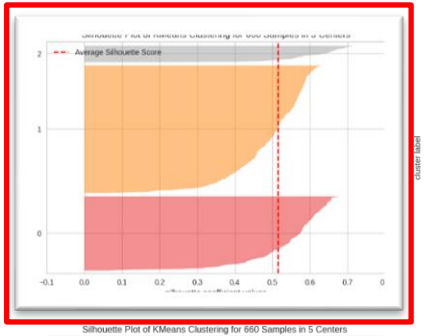

```
CPU times: user 26.6 ms, sys: 0 ns, total: 26.6 ms
Wall time: 15.9 ms
```

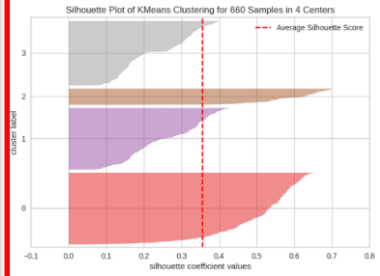
```
KMeans
KMeans(n_clusters=3, random_state=1)
```

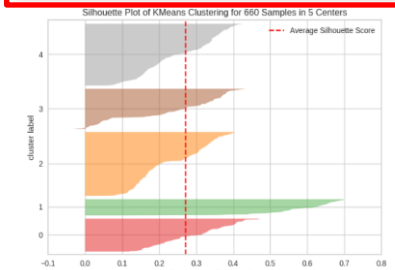
K-Means Clustering, Profiling & Application (Cont'd)

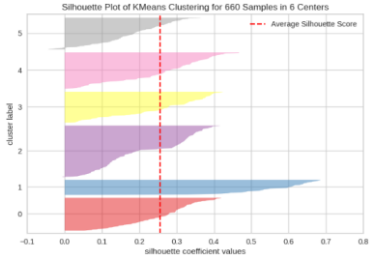
Based on the deep dive analysis below, k=3 seems to be the optimal number of clusters since the silhouette score is high above the average silhouette score for all 3 clusters and there is a knick at 3 in the elbow curve.

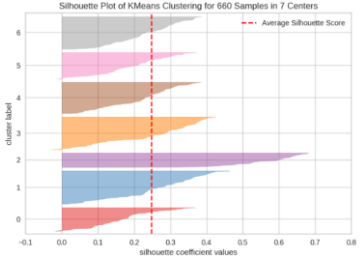
- **k=3:** All clusters (0,1,2) cross the average silhouette score, they are of different widths and there have distinct silhouette scores
- **k=4:** All clusters (0,1,2,3) cross the average silhouette score, and they are of different widths. However, clusters 1 & 2 have very similar silhouette scores
- **k=5:** All clusters (0,1,2,3,4) cross the average silhouette score, and they are of different widths. However, clusters 2,3 & 4 have very similar silhouette scores
- **k=6:** All clusters (0,1,2,3,4,5) cross the average silhouette scores. However, clusters 3 & 4 have very similar width, and cluster 5 has a negative silhouette coefficient value
- **k=7:** All clusters (0,1,2,3,4,5,6) cross the average silhouette scores. However, clusters 1 & 6 have very similar width, and cluster 1,3 & 5 has a negative silhouette coefficient value











K-Means Clustering, Profiling & Application (Cont'd)

- **K-Means Cluster Profiling:** K-Means final model with k=3 has profiled the customers into 3 segments – 0, 1 & 2 with counts of 224, 50 and 386 customers respectively

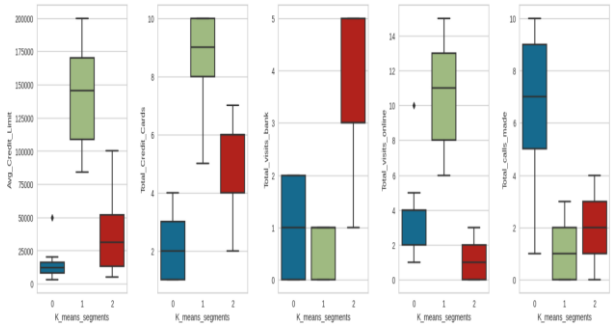
K_means_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	12174.107143	2.410714	0.933036	3.553571	6.870536	224
1	141040.000000	8.740000	0.600000	10.900000	1.080000	50
2	33782.383420	5.515544	3.489637	0.981865	2.000000	386

- **Customer Segmenation Results using K-Means Clustering on the Dataset :**

- **Segment 0** - Customers that prefer phone-banking, have the lowest total number of credit cards and average credit card limit. However, they do often visit the bank and use the bank's online services

- **Segment 1** – Customers that prefer online interactions with their bank, have a much higher average credit limit and have a higher number of credit cards. They rarely visit the bank and/or make phone calls than compared to other segments

- **Segment 2** – Customer that prefer in-person visits to the bank, have a mid-range number of total credit cards and an average credit card limit. They are the lowest in their online presence than compared to the other segment of customers



Hierarchical Clustering

Hierarchical Clustering, Profiling & Application

Following are the results of Hierarchical clustering on the dataset.

- **Distance Metrics Used:** Euclidean, Chebyshev, Mahalanobis and Cityblock
- **Linkage Methods Used:** Single, Complete, Average and Weighted
- **Highest Cophentic Correlation:** Euclidean Average linkage yielded a maximum correlation of 0.8977080867389372
- **Best Dendogram:** Euclidean Average linkage Dendogram with a Cophenetic Coefficient of 0.897708 shows distinct and separate clusters
- **Optimal Number of Clusters :** 3 – Based on the Euclidean Average linkage dendogram, 3 seems to be the optimal number of clusters for Hierarchical clustering.
- **Hierarchical Clustering Model Used:** Agglomerative Model using 3 clusters with Euclidean Average Linkage to yield customer segmentation

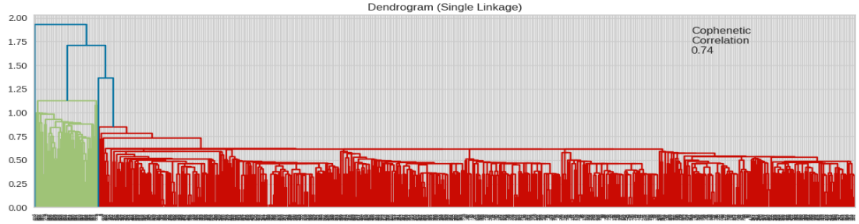
#	Distance	Linkage	Cophenetic Correlation
1	Euclidean	single	0.7391220243806552
2	Euclidean	complete	0.8599730607972423
3	Euclidean	average	0.8977080867389372
4	Euclidean	weighted	0.8861746814895477
5	Euclidean	centroid	0.8939385846326323
6	Euclidean	ward	0.7415156284827493
7	Chebyshev	single	0.7382354769296767
8	Chebyshev	complete	0.8533474836336782
9	Chebyshev	average	0.8974159511838106
10	Chebyshev	weighted	0.8913624010768603
11	Mahalanobis	single	0.7058064784553605
12	Mahalanobis	complete	0.6663534463875359
13	Mahalanobis	average	0.8326994115042136
14	Mahalanobis	weighted	0.7805990615142518
15	Cityblock	single	0.7252379350252723
16	Cityblock	complete	0.8731477899179829
17	Cityblock	average	0.896329431104133
18	Cityblock	weighted	0.8825520731498188

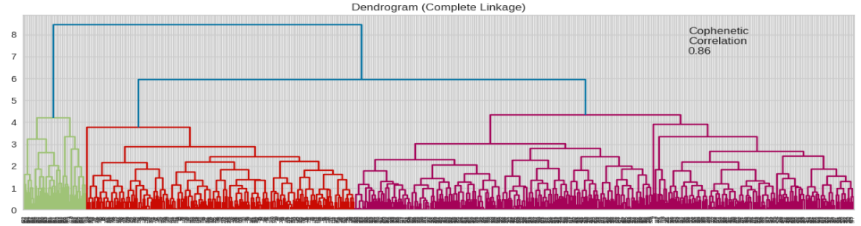
AgglomerativeClustering

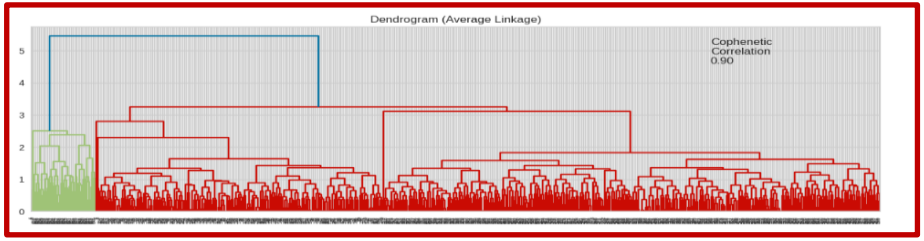
```
AgglomerativeClustering(affinity='euclidean', linkage='average', n_clusters=3)
```

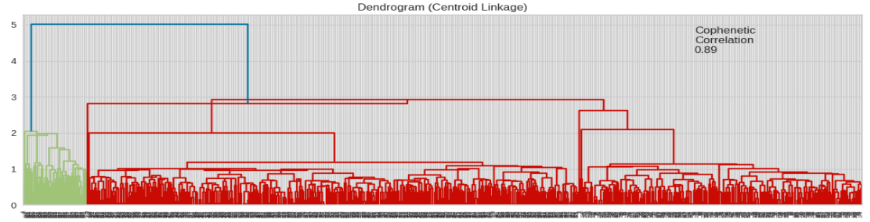
Hierarchical Clustering, Profiling & Application (Cont'd)

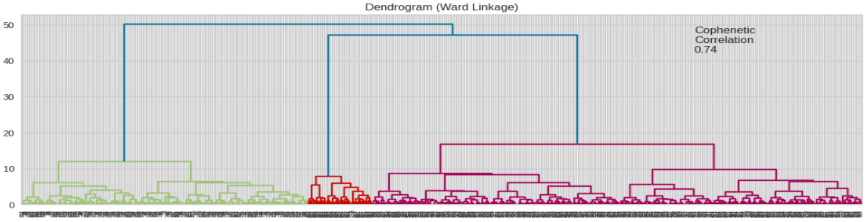
Based on the deep dive analysis on the dendograms below, the Euclidean Average Linkage with a Cophenetic Coefficient of 0.897708 (~0.90) seems to yield the optimal results

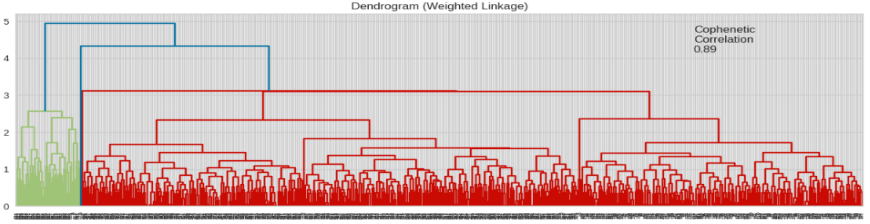












Hierarchical Clustering, Profiling & Application (Cont'd)

- **Hierarchical Cluster Profiling:** Hierarchical Clustering has profiled the customers into 3 segments – 0, 1 & 2 with counts of 387, 50 and 223 customers respectively

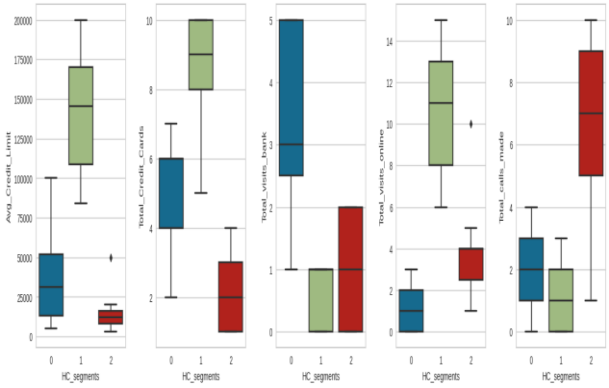
HC_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	33713.178295	5.511628	3.485788	0.984496	2.005168	387
1	141040.000000	8.740000	0.600000	10.900000	1.080000	50
2	12197.309417	2.403587	0.928251	3.560538	6.883408	223

- **Customer Segmenation Results using Hierarchical Clustering on Dataset :**

- **Segment 0** - Customer prefer in-person visits to the bank, have a mid-range number of total credit cards and a mid-range average credit card limit. They are the lowest in their online presence than compared to the other segment of customers

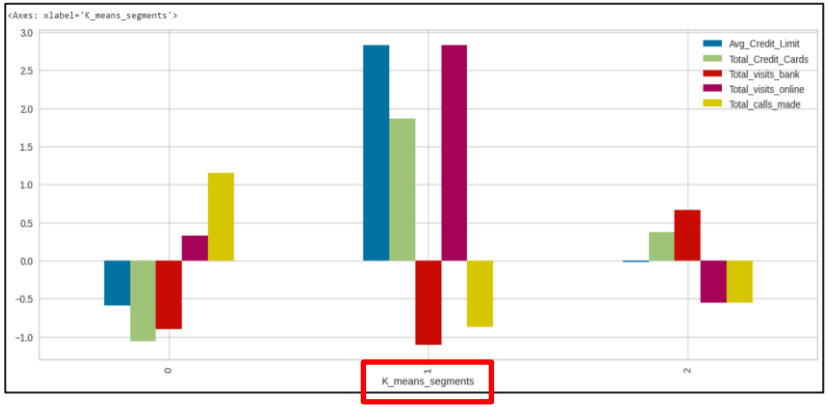
- **Segment 1** – Customers prefer online interactions with their bank, have a much higher average credit limit and have a higher number of credit cards. They rarely visit the bank and/or make fewer phone calls than compared to other segments

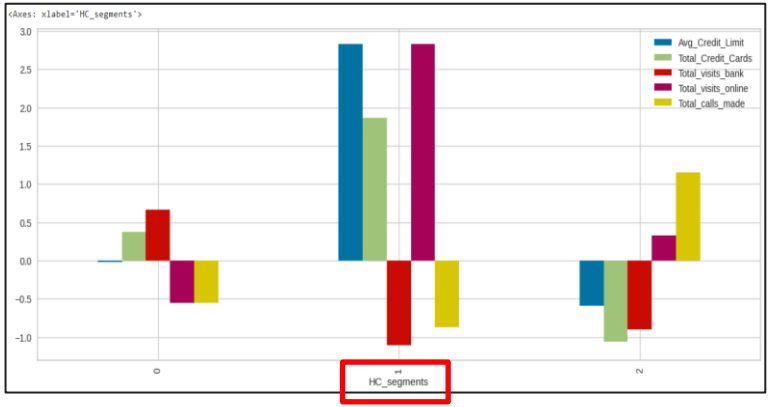
- **Segment 2** – Customers prefer phone-banking, have the lowest total number of credit cards and the lowest average credit card limit. However, they do often visit the bank and use the bank's online services



K-Means vs Hierarchical Clustering

K-Means vs Hierarchical Clustering Comparison





- **Segment 0:** For K-Means, customers prefer phone-banking and have the lowest total number of credit cards and lowest average credit limit. Whereas, for Hierarchical, the customers prefer in-person visits to the bank, have a mid-range number of total credit cards and a mid-range average credit card limit.

- **Segment 1:** For both, K-Means & Hierarchical Clustering, customers prefer online interactions with their bank, have a much higher average credit limit and have a higher number of credit cards. They rarely visit the bank and/or make fewer phone calls than compared to other segments

- **Segment 2:** For K-Means, customers prefer in-person visits to the bank, have a mid-range number of total credit cards and a mid-range of average credit card limit. Whereas, for Hierarchical, the customers prefer phone-banking, have the lowest total number of credit cards and the lowest average credit card limit

K-Means vs Hierarchical Clustering Comparison (Cont'd)

- **Appropriate Number of Clusters:** 3 – For both, K-Means & Hierarchical Clustering, 3 clusters were identified as the optimal number of clusters to yield appropriate customer segmentation results
- **Number Of Observations in similar Clusters:** For both, K-Means & Hierarchical Clustering, Segment 1 had 50 identical observations – higher number of online visits, higher average credit limit and have a higher total number of credit cards

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

HC_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	33713.178295	5.511628	3.485788	0.984496	2.005168	387
1	141040.000000	8.740000	0.600000	10.900000	1.080000	50
2	12197.309417	2.403587	0.928251	3.560538	6.883408	223

- **Execution Time:** K-Means took 15.3 ms than compared to Agglomerative Clustering that took 37.7 ms

```
CPU times: user 24.2 ms, sys: 794 µs, total: 25 ms
Wall time: 15.3 ms
KMeans
KMeans(n_clusters=3, random_state=1)
```

```
CPU times: user 19.5 ms, sys: 0 ns, total: 19.5 ms
Wall time: 37.7 ms
AgglomerativeClustering
AgglomerativeClustering(affinity='euclidean', linkage='average', n_clusters=3)
```

- **Distinct Clusters:** Both, K-Means & Hierarchical Clustering yielded distinct clusters

Conclusion

- Both K-Means & Hierarchical Clustering Models have produced 3 distinct customer segments
    - Phone Banking Customer Segment - customers who prefer to transact via the telephony channel
    - Online Banking Customer Segment - customers who prefer online transactions with their bank
    - In-Person Customer Segment - customers who prefer physically visiting the bank to perform transactions
- The above segmentation can be used by the Marketing team to deliver personalised marketing campaigns based on the customer's banking behaviours and preference
- The above segmentation can be used by the Operations team to improve their service delivery model by optimising the underlying IT Applications & Infrastructure to ensure that customer queries are resolved faster based on the customer's banking behaviours and preference

Key Actionable Business Insights

- **Summarised Key Actionable Business Insights:**
  - There are 3 distinct segments of customers produced by K-Means & Hierarchical Clustering Methods:
    - Phone Banking Customers: These customers prefer to transact via the telephony channel, and have the lowest total number of credit cards and the lowest average credit card limit
    - Online Customers: These customers prefer online transactions with their bank, have a much higher average credit limit and have a higher number of credit cards. They rarely visit the bank and/or make fewer phone calls
    - In-Person Customers: These customers prefer in-person visits to the bank, and have a mid-range number of total credit cards and a mid-range of average credit card limit
  - A Personalised Preference Campaign Strategy should be put in place to cater to the above 3 segments of customers:
    - Phone Banking & Online Customers: Email, SMS, Instant Messaging Services / Social Media Applications should be used to target these segment of customers
    - In-Person Customers: Mail notifications, Promotion Flyers, Marketing Leaflets, Posters at Bank branches etc should be used to target this segment of customers.

Our Recommendation

Based on our key observations and insights, we recommend the following areas of improvement / opportunities that will drive business growth and lead to a better customer experience

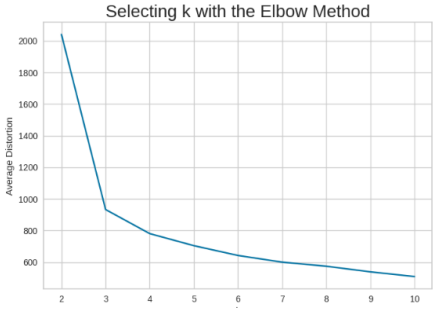
- **Implement Customer Satisfaction Survey:** The Bank should initiate a targeted Customer Satisfaction Survey to understand customer pain points for the current services and implement the findings to improve retention ratio of customers
- **Implement Customer Incentivisation Scheme:** Incentivising customers by offering them cashback schemes and discounts / vouchers on credit card purchases will encourage frequent spending and will drive customer gowth and increase revenue
- **Implement Tier based Rewards:** The Bank should introduce a Tier based Loyalty & Rewards Scheme for credit card purchases. Cumulative loyalty points above a certain threshold will promote the customer to a new tier, that will offer specific rewards such as First-Class Lounge access at Airports, Spa & Well-Being discounts etc. This will enable the Marketing team to upsell new products to existing customers and drive customer retention

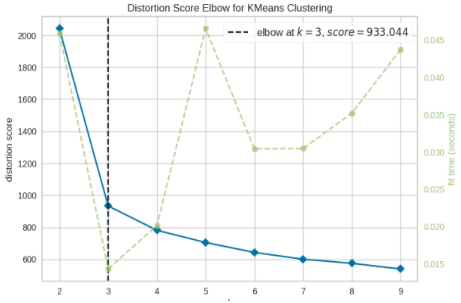
APPENDIX

K-Means Clustering - Elbow Curve Observation & Visuals

- Elbow Curve Observation & Visuals

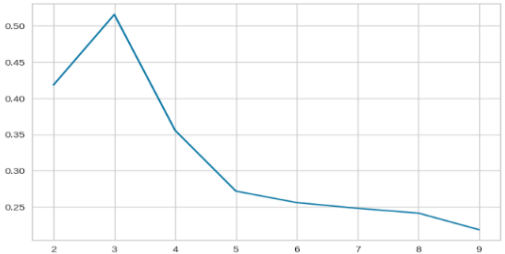
```
Number of Clusters: 2   Average Distortion: 2040.9898164784947
Number of Clusters: 3   Average Distortion: 933.0437490000531
Number of Clusters: 4   Average Distortion: 780.7736895551769
Number of Clusters: 5   Average Distortion: 704.4759188657513
Number of Clusters: 6   Average Distortion: 642.4285451423211
Number of Clusters: 7   Average Distortion: 600.2238778375963
Number of Clusters: 8   Average Distortion: 574.4418958177623
Number of Clusters: 9   Average Distortion: 538.8269188945014
Number of Clusters: 10  Average Distortion: 509.16313788912544
```

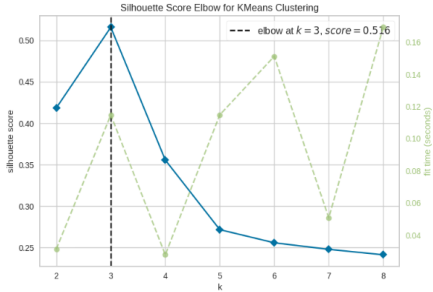




- Silhouette scores for 2, 3, 4, 5, 6, 7, 8, & 9 clusters

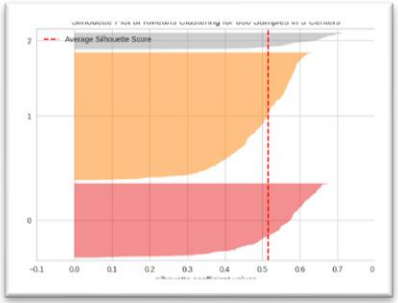
```
For n_clusters = 2, the silhouette score is 0.41842496663215445)
For n_clusters = 3, the silhouette score is 0.5157182558881063)
For n_clusters = 4, the silhouette score is 0.3556670619372605)
For n_clusters = 5, the silhouette score is 0.2717470361089752)
For n_clusters = 6, the silhouette score is 0.255906765297388)
For n_clusters = 7, the silhouette score is 0.24798644656011146)
For n_clusters = 8, the silhouette score is 0.2414240144760896)
For n_clusters = 9, the silhouette score is 0.2184645050755029)
```

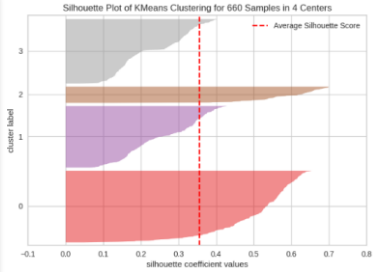


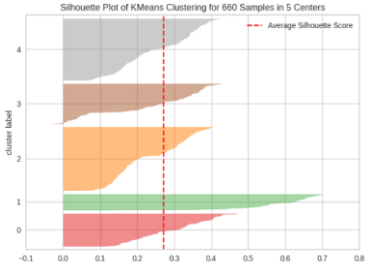


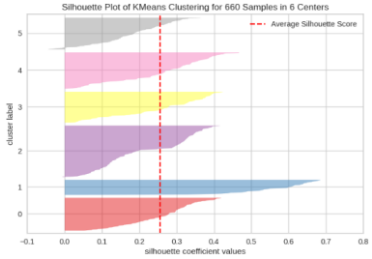
K-Means Clustering – Silhouette Score Visualiser

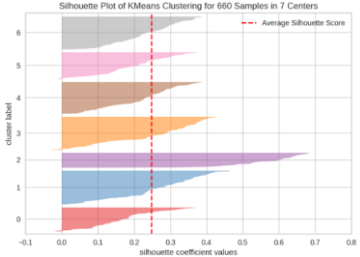
- **k=3:** All clusters (0,1,2) cross the average silhouette score, they are of different widths and there have distinct silhouette scores
- **k=4:** All clusters (0,1,2,3) cross the average silhouette score, and they are of different widths. However, clusters 1 & 2 have very similar silhouette scores
- **k=5:** All clusters (0,1,2,3,4) cross the average silhouette score, and they are of different widths. However, clusters 2,3 & 4 have very similar silhouette scores
- **k=6:** All clusters (0,1,2,3,4,5) cross the average silhouette scores. However, clusters 3 & 4 have very similar width, and cluster 5 has a negative silhouette coefficient value
- **k=7:** All clusters (0,1,2,3,4,5,6) cross the average silhouette scores. However, clusters 1 & 6 have very similar width, and cluster 1,3 & 5 has a negative silhouette coefficient value











Hierarchical Clustering – Cophenetic Correlations

- Cophenetic Correlations for different Linkage Methods

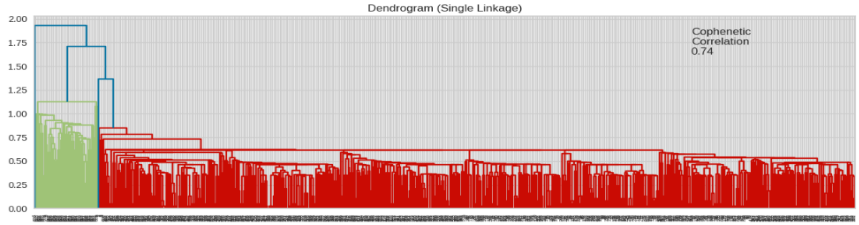
```
Cophenetic correlation for Euclidean distance and single linkage is 0.7391220243806552.
Cophenetic correlation for Euclidean distance and complete linkage is 0.8599730607972423.
Cophenetic correlation for Euclidean distance and average linkage is 0.8977080867389372.
Cophenetic correlation for Euclidean 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.
```

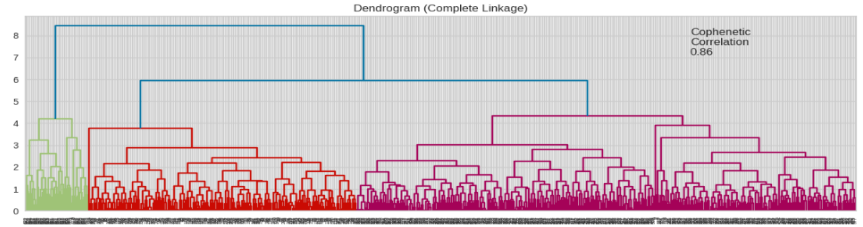
- Cophenetic Correlations for Euclidean Distance Method

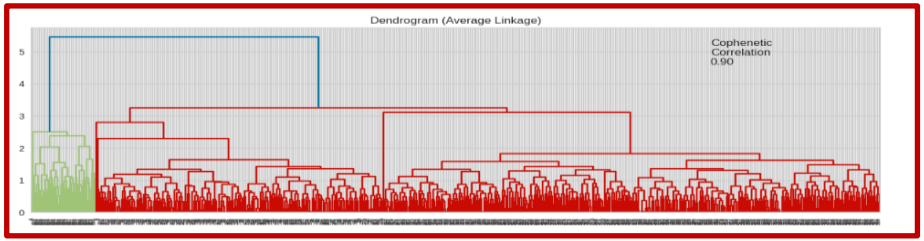
```
Cophenetic correlation for single linkage is 0.7391220243806552.
Cophenetic correlation for complete linkage is 0.8599730607972423.
Cophenetic correlation for average linkage is 0.8977080867389372.
Cophenetic correlation for centroid linkage is 0.8939385846326323.
Cophenetic correlation for ward linkage is 0.7415156284827493.
Cophenetic correlation for weighted linkage is 0.8861746814895477.
```

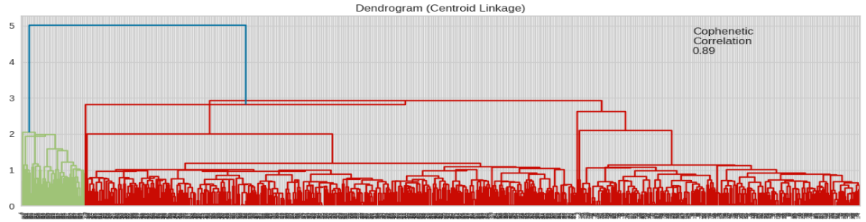
Hierarchical Clustering – Euclidean Method Dendograms

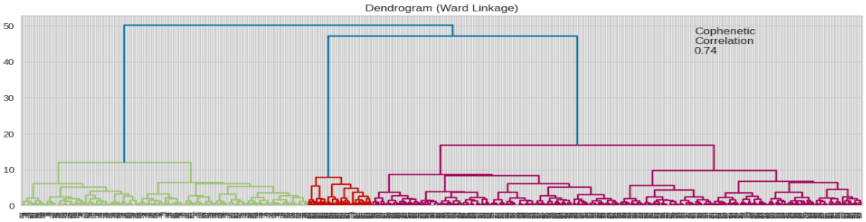
Based on the deep dive analysis on the dendograms below, the Euclidean Average Linkage with a Cophenetic Coefficient of 0.897708 (~0.90) seems to yield the optimal results

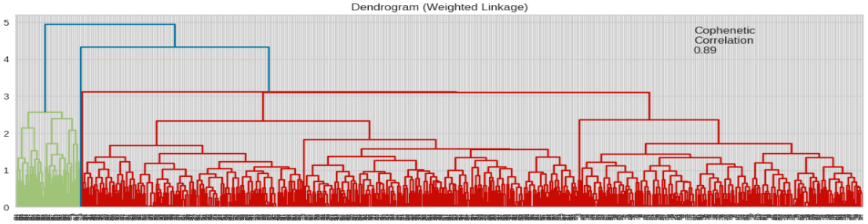




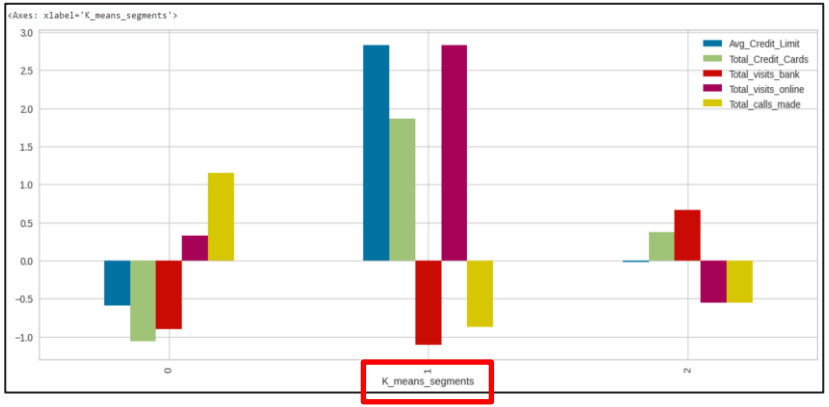


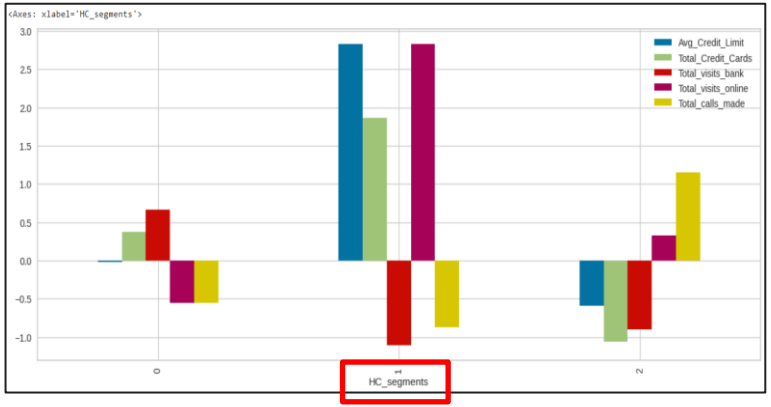






K-Means vs Hierarchical Clustering Comparison





- **Segment 0:** For K-Means, customers prefer phone-banking and have the lowest total number of credit cards and lowest average credit limit. Whereas, for Hierarchical, the customers prefer in-person visits to the bank, have a mid-range number of total credit cards and a mid-range average credit card limit.

- **Segment 1:** For both, K-Means & Hierarchical Clustering, customers prefer online interactions with their bank, have a much higher average credit limit and have a higher number of credit cards. They rarely visit the bank and/or make fewer phone calls than compared to other segments

- **Segment 2:** For K-Means, customers prefer in-person visits to the bank, have a mid-range number of total credit cards and a mid-range of average credit card limit. Whereas, for Hierarchical, the customers prefer phone-banking, have the lowest total number of credit cards and the lowest average credit card limit

Great Learning

Happy Learning !