UL CODED PROJECT

Business Report

DSBA

Submitted By: Maheep Singh
Batch : PGP-DSBA (PGPDSBA.O.AUG24.A)

Table of Contents

List of Figures

List of Tables

Business Context & Data Dictionary

Context

AllLife Bank wants to focus on its credit card customer base in the next financial year. They have been advised by their marketing research team, that the penetration in the market can be improved. Based on this input, the Marketing team proposes to run personalized campaigns to target new customers as well as upsell to existing customers. Another insight from the market research was that the customers perceive the support services of the back poorly. Based on this, the Operations team wants to upgrade the service delivery model, to ensure that customer queries are resolved faster. The Head of Marketing and Head of Delivery both decide to reach out to the Data Science team for help.

Objective

To identify different segments in the existing customers, based on their spending patterns as well as past interaction with the bank, using clustering algorithms, and provide recommendations to the bank on how to better market to and service these customers.

Data Description

The data provided is of various customers of a bank and their financial attributes like credit limit, the total number of credit cards the customer has, and different channels through which customers have contacted the bank for any queries (including visiting the bank, online, and through a call centre).

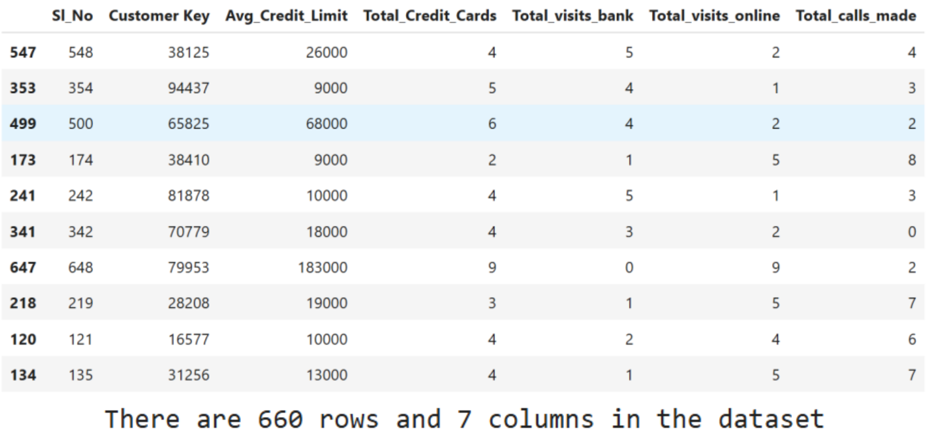
Data Dictionary:

- Sl_No: Primary key of the records
- Customer Key: Customer identification number
- Average Credit Limit: Average credit limit of each customer for all credit cards
- Total credit cards: Total number of credit cards possessed by the customer
- Total visits bank: Total number of visits that the customer made (yearly) personally to the bank
- Total visits online: Total number of visits or online logins made by the customer (yearly)
- Total calls made: Total number of calls made by the customer to the bank or its customer service department (yearly)

Rubric Question 1: Exploratory Data Analysis

Data Overview

- **Load dataset** & display a random sample of 10 rows: -

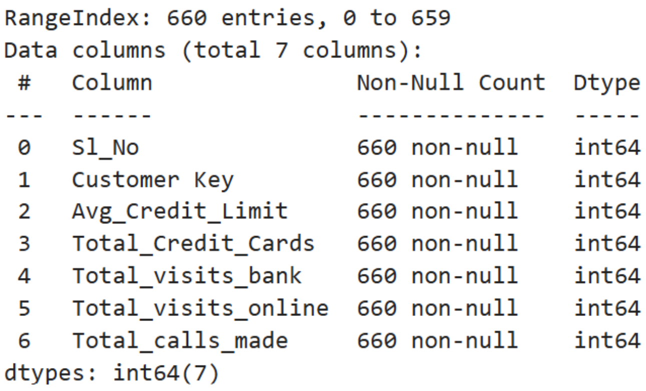


	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
547	548	38125	26000	4	5	2	4
353	354	94437	9000	5	4	1	3
499	500	65825	68000	6	4	2	2
173	174	38410	9000	2	1	5	8
241	242	81878	10000	4	5	1	3
341	342	70779	18000	4	3	2	0
647	648	79953	183000	9	0	9	2
218	219	28208	19000	3	1	5	7
120	121	16577	10000	4	2	4	6
134	135	31256	13000	4	1	5	7

```
There are 660 rows and 7 columns in the dataset
```

Figure 1: Random Sample of 10 rows of the dataset

- There are **660 rows & 7 columns** in the dataset
- **Checking datatypes**: -

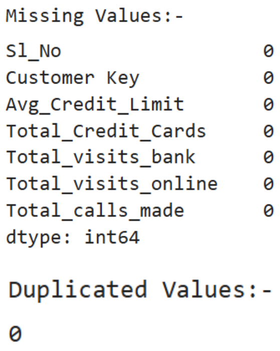


```
RangeIndex: 660 entries, 0 to 659
Data columns (total 7 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Sl_No                660 non-null    int64
 1   Customer Key         660 non-null    int64
 2   Avg_Credit_Limit     660 non-null    int64
 3   Total_Credit_Cards   660 non-null    int64
 4   Total_visits_bank    660 non-null    int64
 5   Total_visits_online  660 non-null    int64
 6   Total_calls_made     660 non-null    int64
dtypes: int64(7)
```

Figure 2: Datatypes in the Dataset

- There are 7 columns of int64 (numeric) datatype in the dataset.

- **Check & Treat Missing/Duplicate Values**: -
  - Upon checking, neither missing nor duplicate values were found. Hence, no treatment required.

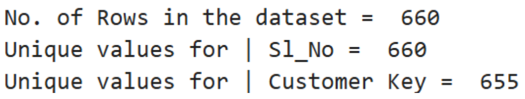


```
Missing Values:-
Sl_No                  0
Customer Key           0
Avg_Credit_Limit       0
Total_Credit_Cards     0
Total_visits_bank      0
Total_visits_online    0
Total_calls_made       0
dtype: int64

Duplicated Values:-
0
```

Figure 3: Missing/Duplicate Value-check

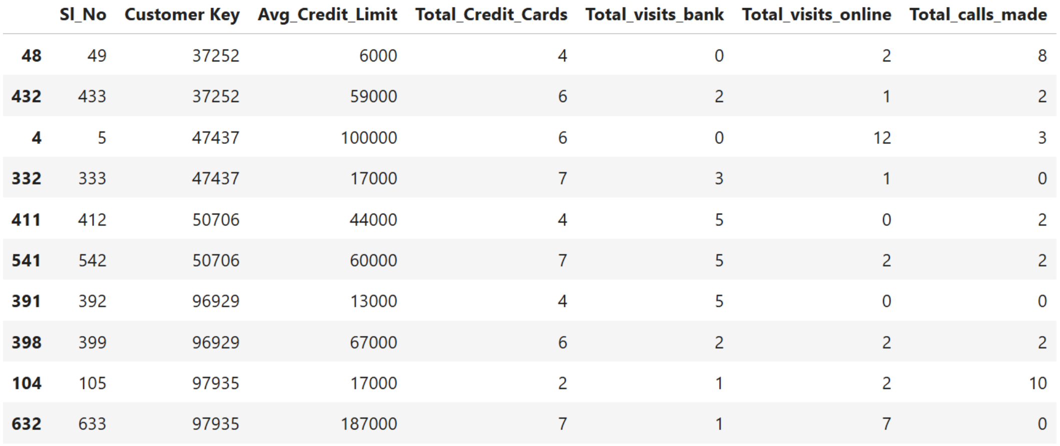
- **Checking Keys**: -
  - We have 2 keys in the table: -
    - ✓ Sl_No: Signifies each line item in the row.
    - ✓ Customer Key: Unique identifier for each customer in the table
  - To check for duplicate values for each key, let's check for their unique values: -



```
No. of Rows in the dataset =  660
Unique values for | Sl_No =  660
Unique values for | Customer Key =  655
```

Figure 4: Checking Keys for Duplicate values

- ✓ As evident above, 'Customer Key' has duplicate values while 'SL_No' doesn't. Clearly, 5 'Customer Keys' are repeated in the dataset. Let's check them: -

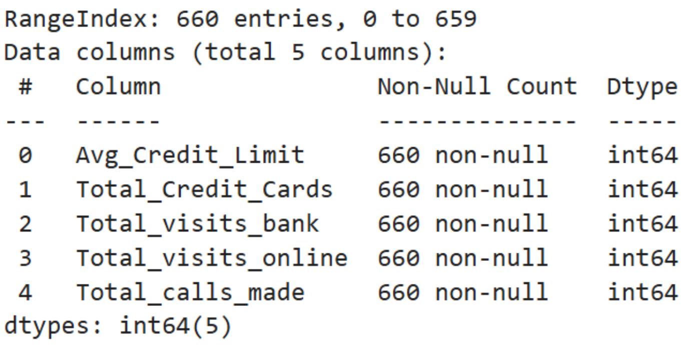


	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
48	49	37252	6000	4	0	2	8
432	433	37252	59000	6	2	1	2
4	5	47437	100000	6	0	12	3
332	333	47437	17000	7	3	1	0
411	412	50706	44000	4	5	0	2
541	542	50706	60000	7	5	2	2
391	392	96929	13000	4	5	0	0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

Figure 5: Dataset with Duplicate Customer Keys

- ✓ By looking at the dataset above, each 'Customer Key' entry has dissimilar values for other attributes, which implies that there may have been an error in updating the Customer Key in the table, otherwise, each represent a different customer.
- With the above observation we will **keep the dataset unchanged as each duplicated entry with the same Customer Key represents a different customer**: -

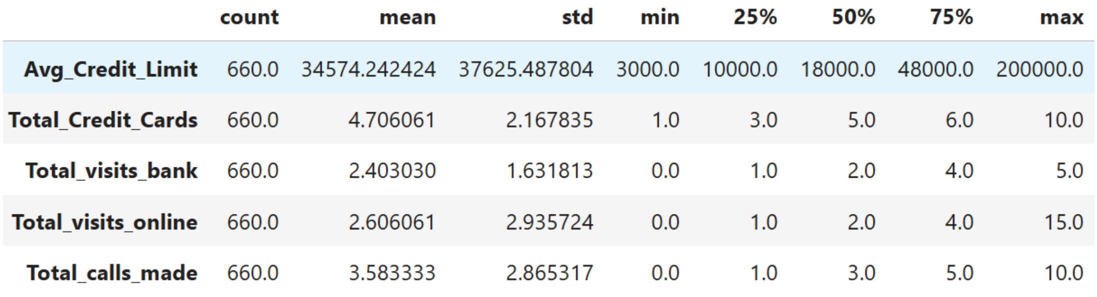
- **Dropping redundant columns:** Both keys ('Sl_No' & 'Customer Key') serve no purpose in the analysis and can be dropped before we proceed with building model for Unsupervised Learning to do Customer Profiling. Below are the final columns post dropping these columns: -



```
RangeIndex: 660 entries, 0 to 659
Data columns (total 5 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Avg_Credit_Limit     660 non-null    int64
 1   Total_Credit_Cards   660 non-null    int64
 2   Total_visits_bank    660 non-null    int64
 3   Total_visits_online  660 non-null    int64
 4   Total_calls_made     660 non-null    int64
dtypes: int64(5)
```

Figure 6: Final list of Columns (with Datatypes) post dropping Redundant columns

- **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

Figure 7: Statistical Summary of the Dataset

Type	Columns	Observations & Insights
Numerical	Avg_Credit_Limit	✓ Average credit limit of each customer for all credit cards range between 3K & 200K ✓ Mean is ~34.5K & median is ~18K ✓ Standard Deviation is ~37K
Numerical	Total_Credit_Cards	✓ Total number of credit cards possessed by the customer range between 1 & 10 ✓ Mean is ~5 & median is also ~5 ✓ Standard Deviation is 2
Numerical	Total_visits_bank	✓ Total number of visits that the customer made (yearly) personally to the bank range between 0 & 5 ✓ Mean is ~2 & median is also ~2 ✓ Standard Deviation is ~2
Numerical	Total_visits_online	✓ Total number of visits or online logins made by the customer (yearly) range between 0 & 15 ✓ Mean is ~3 & median is ~2 ✓ Standard Deviation is ~3
Numerical	Total_calls_made	✓ Total number of calls made by the customer to the bank or its customer service department (yearly) range between 0 & 10 ✓ Mean is ~3 & median is also ~3 ✓ Standard Deviation is ~3

Table 1: Statistical Summary – Observations

Univariate & Bivariate Analysis

- **Perform Univariate Analysis** – Use Histograms & Boxplots to analyse each numerical variable: -

1. Distribution of **Avg_Credit_Limit**: -

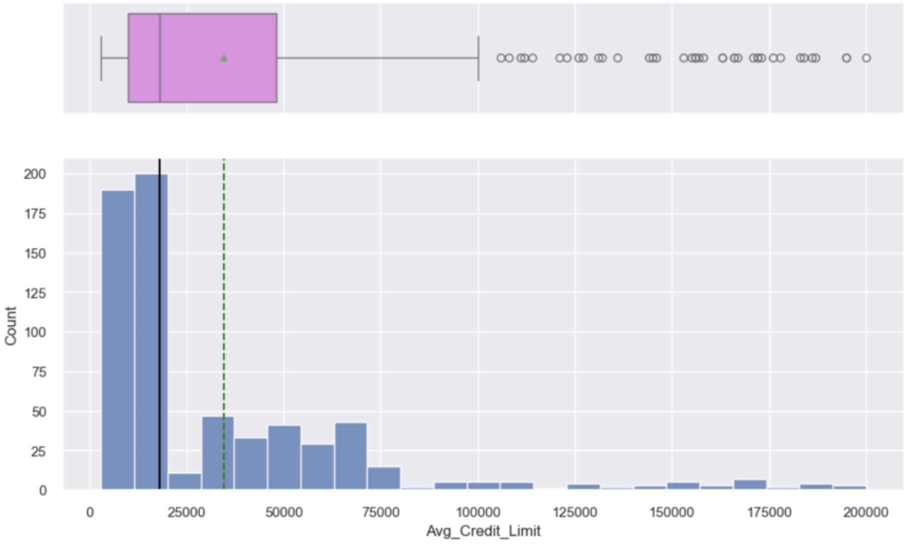


Figure 8: Univariate Analysis – Avg_Credit_Limit

✓ Let's create a new column '**Avg_Credit_Limit_Bins'** to further categorize the Credit Limit – Low (3K to 10K), Mid (10K to 50K), High (50K to 200K) & check distribution: -

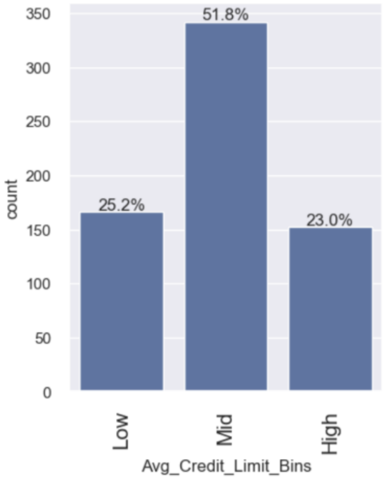


Figure 9: Univariate Analysis – Avg_Credit_Limit_Bins

✓ **Observations & Insights** can be summarized below: -

- There is large variation in the Average Credit Limit of customers.
- Majority of the Customers fall into the 'Mid' category i.e. between 10K to 50K.
- Distribution seems to be highly right-skewed with different mean & median.
- Seems to be unimodal distribution having a single peak.
- Lot of outliers observed, but we would keep the data as-is to avoid any loss of information.

2. Distribution of **Total_Credit_Cards**: -

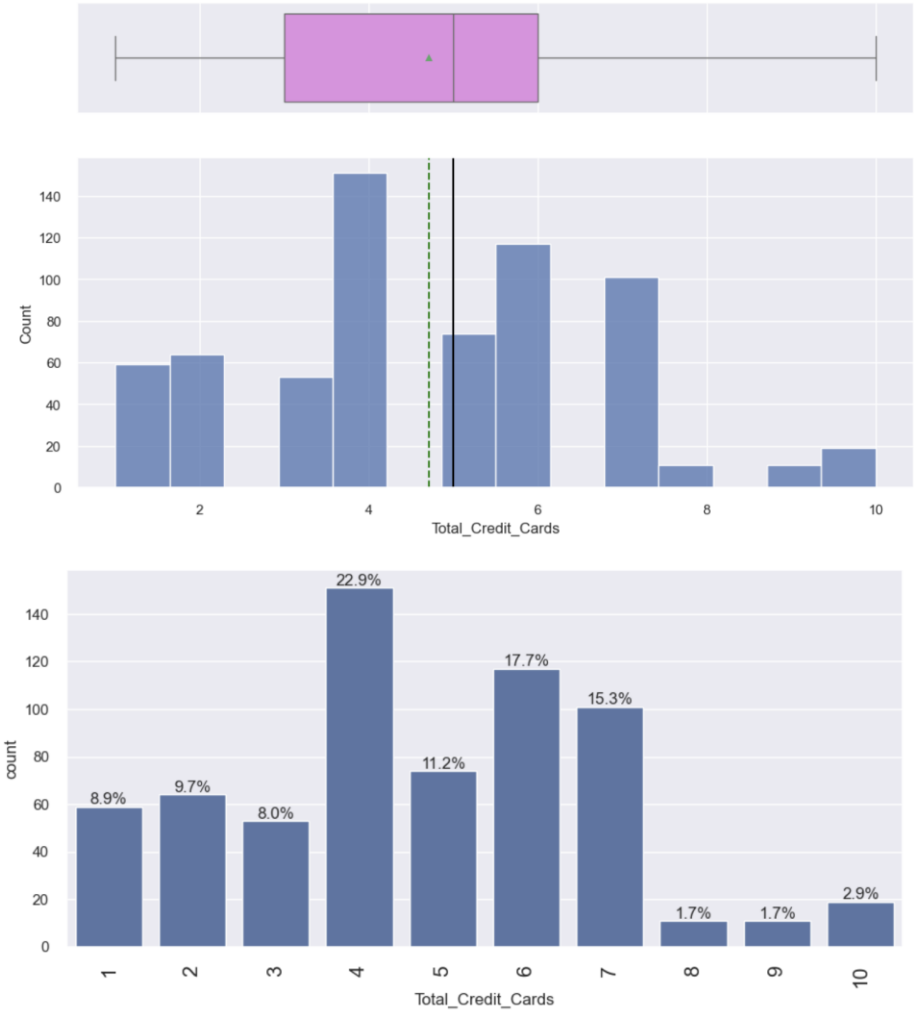


Figure 10: Univariate Analysis – Total_Credit_Cards

✓ **Observations & Insights** can be summarized below: -

- There is moderate variation in total credit cards possessed by customers.
- Majority of the Customers fall hold 4 (23%), followed by 6 (18%) & 7 (15%) credit cards.
- Distribution seems to be nearly symmetric with nearby mean & median.
- Seems to be Multimodal distribution having 2-3 peaks.
- No outliers observed.

3. Distribution of **Total_visits_bank**: -

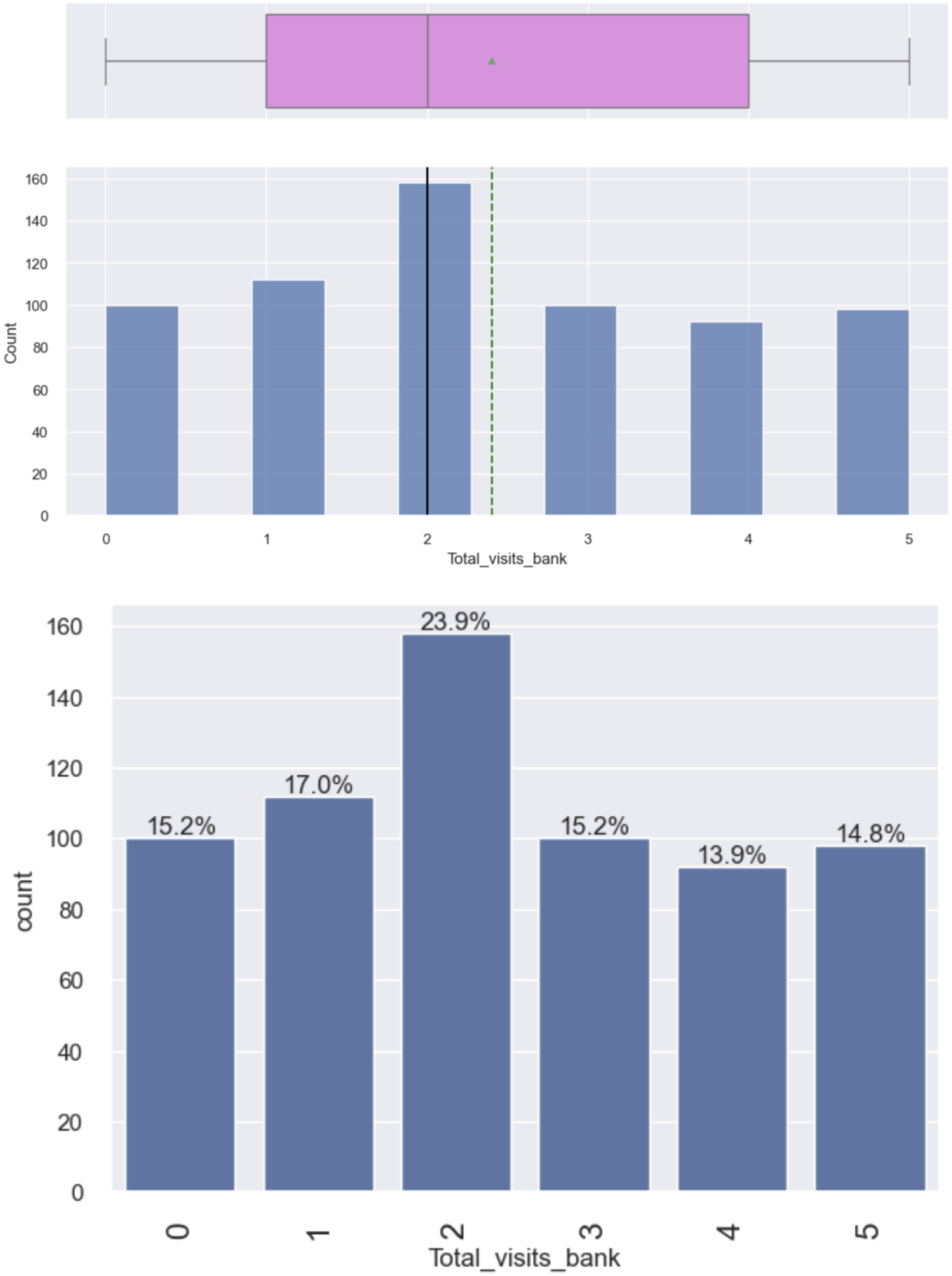


Figure 11: Univariate Analysis – Total_visits_bank

✓ **Observations & Insights** can be summarized below: -

- There is low variation in total bank visits by customers.
- Majority of the customers visited 2 times/year.
- Distribution seems to be nearly symmetric with nearby mean & median.
- Seems to be Unimodal distribution having 1 peak.
- No outliers observed.

4. Distribution of **Total_visits_online**: -

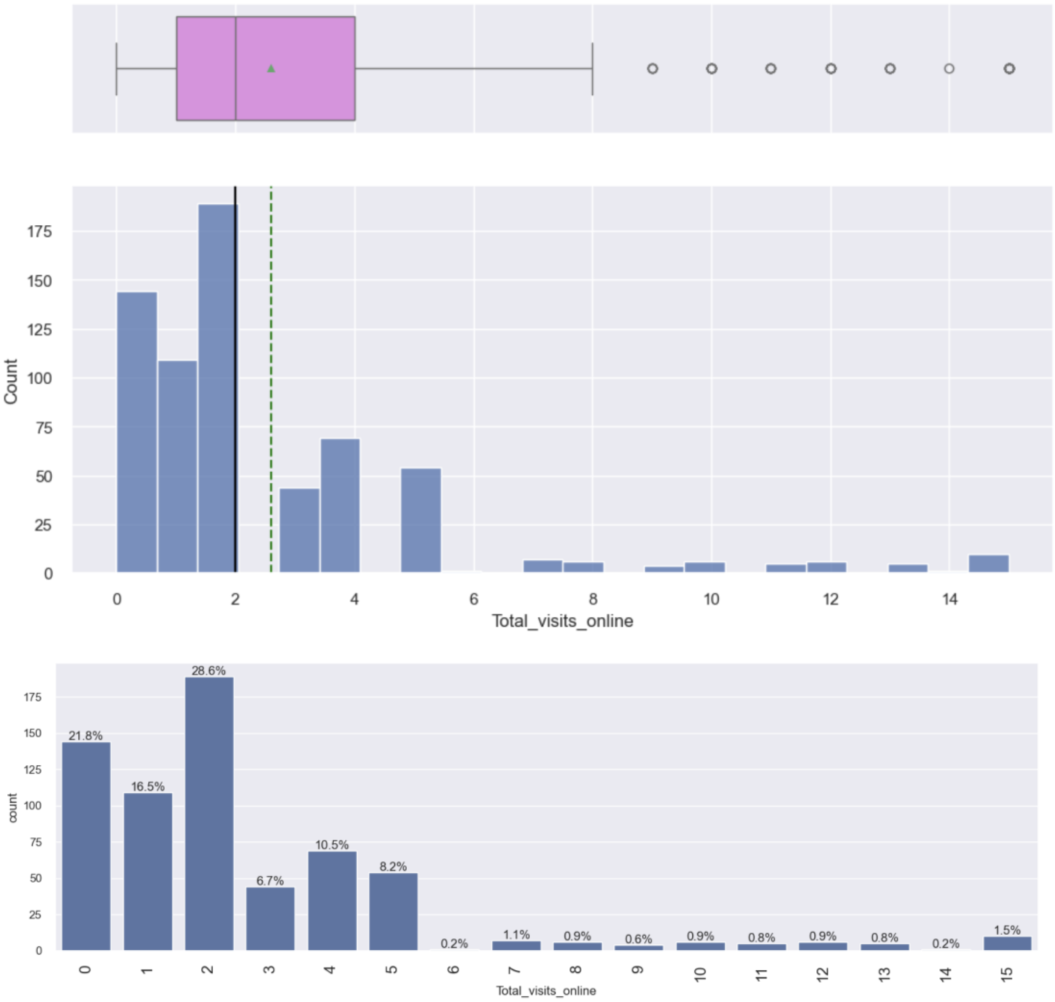


Figure 12: Univariate Analysis – Total_visits_online

✓ **Observations & Insights** can be summarized below: -

- There is low variation in total online visits by customers.
- Majority of the customers visited 2 times/year.
- Distribution seems to be highly right skewed with different mean & median.
- Seems to be Unimodal distribution having 1 peak.
- Lot of outliers observed, but we would keep the data as-is to avoid any loss of information.

5. Distribution of **Total_calls_made**: -

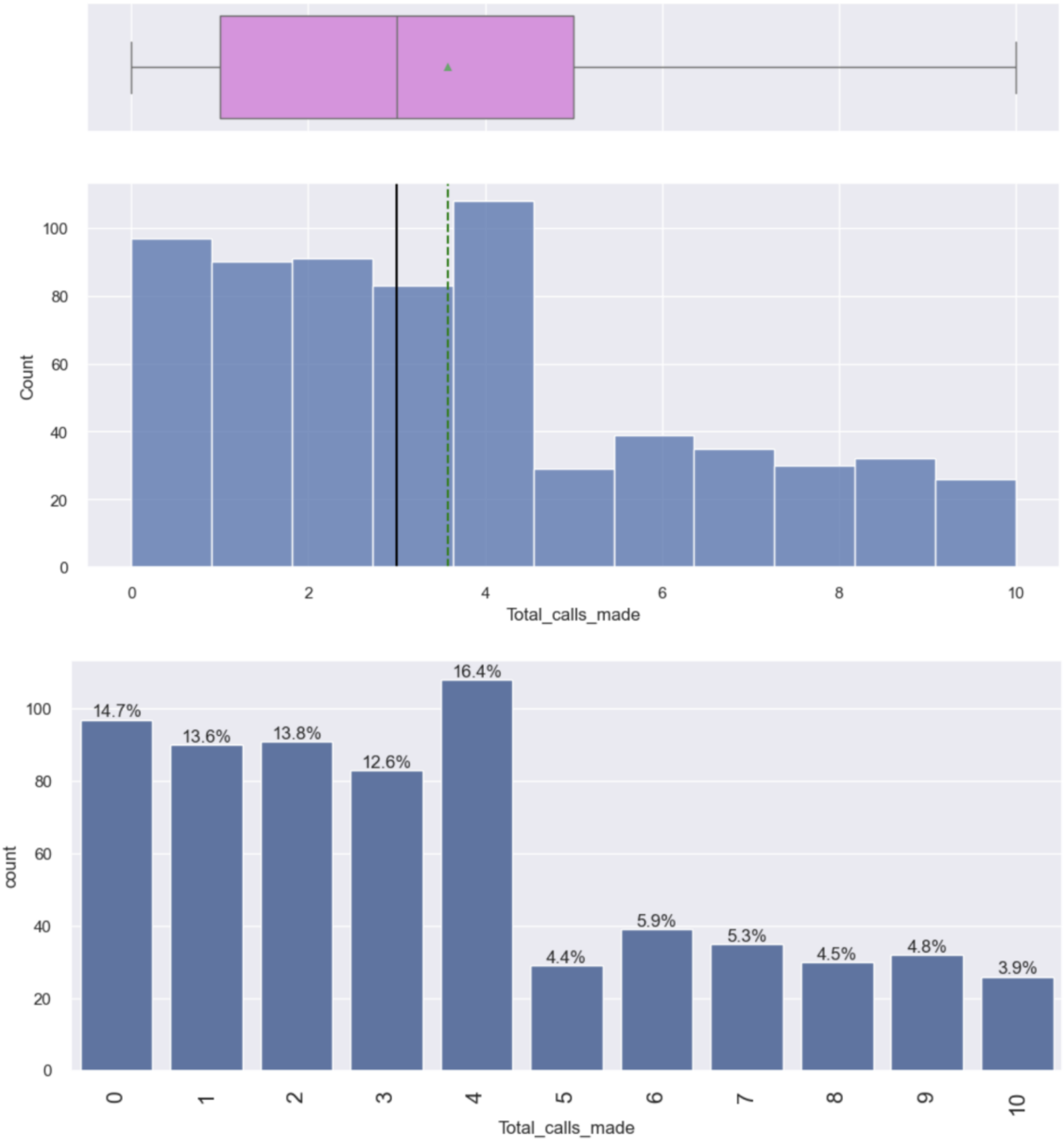


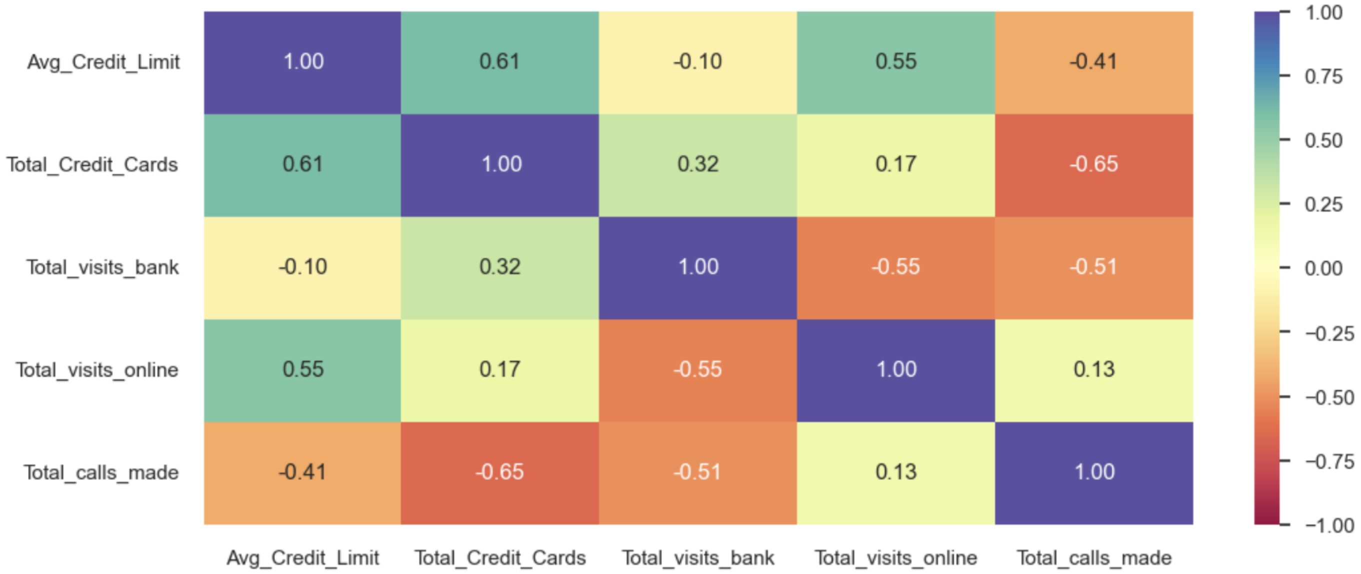
Figure 13: Univariate Analysis – Total_calls_made

✓ **Observations & Insights** can be summarized below: -

- There is low variation in total calls made by the customer.
- Majority of the customers called up to 4 times/year.
- Distribution seems to be highly right skewed with different mean & median.
- Seems to be Multimodal distribution having many peaks.
- No outliers observed.

- **Perform Bivariate Analysis** – Use PairPlot, Heatmap & BoxPlots to carry out bivariate analysis between numerical variables: -

1. **Between Numerical Variables – PairPlot & Heatmap:** -



✓ **Observations & Insights** can be summarized below: -

- Avg_Credit_Limit & Total_Credit_Cards (0.61): There is a moderate positive correlation, suggesting that customers with more credit cards tend to have a higher credit limit.
- Avg_Credit_Limit & Total_visits_online (0.55): A moderate positive correlation, meaning customers with a higher credit limit tend to use online banking more frequently.
- Avg_Credit_Limit & Total_calls_made (-0.41): A moderate negative correlation, indicating that customers with a higher credit limit tend to make fewer calls to the bank.
- Total_Credit_Cards & Total_calls_made (-0.65): A strong negative correlation, meaning that customers with more credit cards tend to make fewer calls to the bank.
- Total_visits_bank & Total_visits_online (-0.55): A moderate negative correlation, suggesting that customers who visit the bank more often tend to use online banking less.
- Total_visits_bank & Total_calls_made (-0.51): A moderate negative correlation, indicating that customers who visit the bank more often also tend to make fewer calls.
- **To summarise:** -
    - Customers with more credit cards and a higher credit limit are more likely to use online banking but make fewer calls to the bank.
    - There is a trade-off between online visits and physical visits, meaning customers who prefer in-person visits are less likely to use online banking.
    - A higher number of calls to the bank is associated with fewer credit cards and lower credit limits, possibly indicating that customers with lower credit access need more support.

Figure 14: Bivariate Analysis - Heatmap

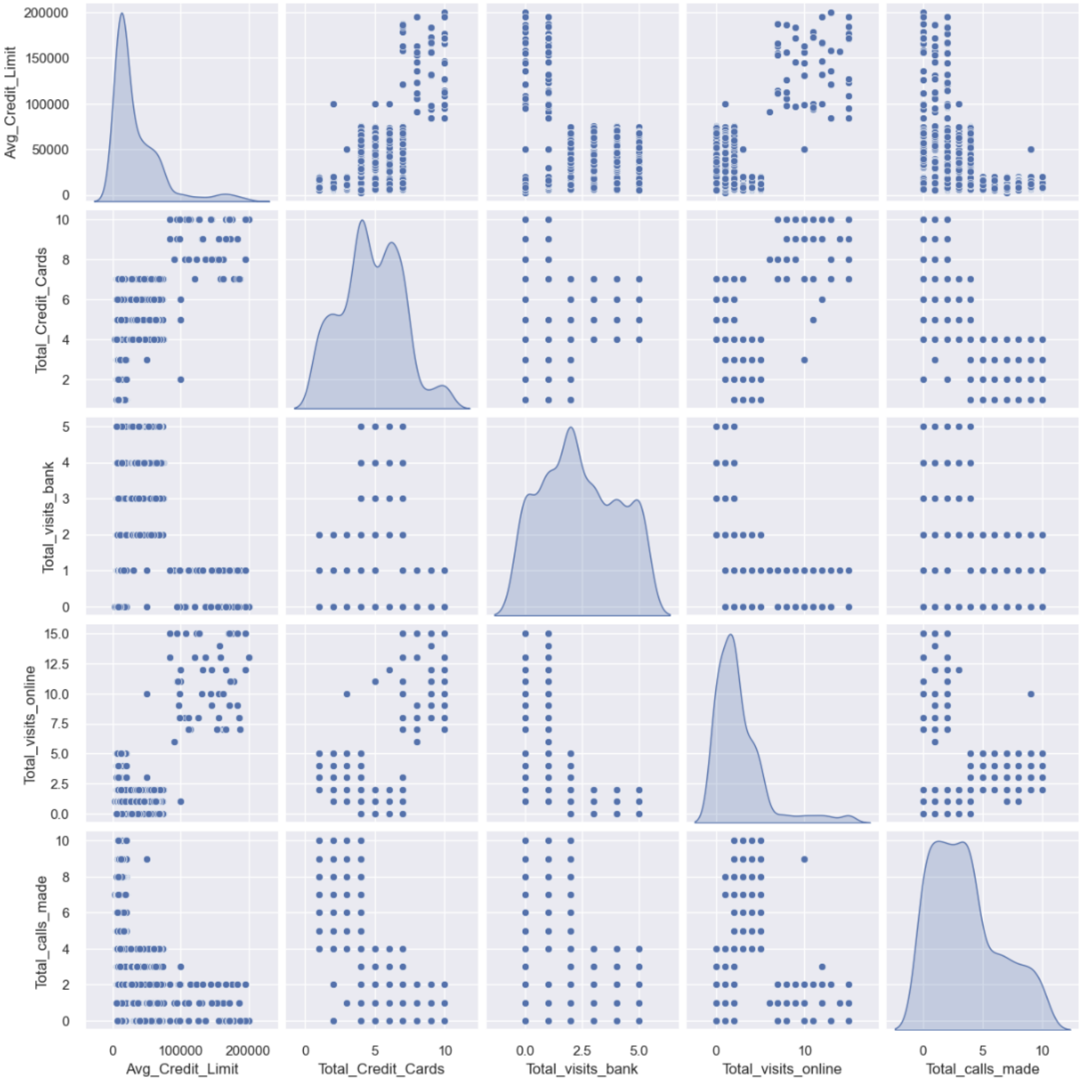


Figure 15: Bivariate Analysis – PairPlot

- ✓ **Observations & Insights** can be summarized below: -
  - ➢ **Diagonal Distributions: -**
    - Avg_Credit_Limit: Positively skewed, with most values concentrated on the lower end and a few very high values.
    - Total_Credit_Cards: Mostly normally distributed but with some skewness.
    - Total_visits_bank: Looks like a bimodal or right-skewed distribution.
    - Total_visits_online: Right-skewed, meaning most users have fewer online visits.
    - Total_calls_made: Most users make only a few calls, forming a right-skewed pattern.
  - ➢ **Scatterplot Insights (Variable Relationships): -**
    - Avg_Credit_Limit vs. Total_Credit_Cards: Shows a slight positive correlation – Higher credit limits are associated with having more credit cards.
    - Avg_Credit_Limit vs. Total_visits_online: A slight positive trend, suggesting that higher credit limit users engage more in online banking.
    - Avg_Credit_Limit vs. Total_calls_made: Negative correlation – Users with a higher credit limit tend to make fewer calls.
    - Total_Credit_Cards vs. Total_calls_made: Strong negative correlation, confirming that users with many credit cards tend to contact customer support less frequently.
    - Total_visits_bank vs. Total_visits_online: Negative correlation, meaning people who visit banks frequently tend to use online banking less.

- **To summarize: -**
    - Customers with higher credit limits tend to have more credit cards, visit the bank less, and rely more on online banking.
    - People who visit the bank frequently are less likely to use online banking.
    - Customers who make frequent calls to the bank generally have lower credit limits and fewer credit cards, possibly indicating more support needs.

2. **Between Numerical Variables – BoxPlots (Avg_Credit_Limit vs. other Key Variables): -**

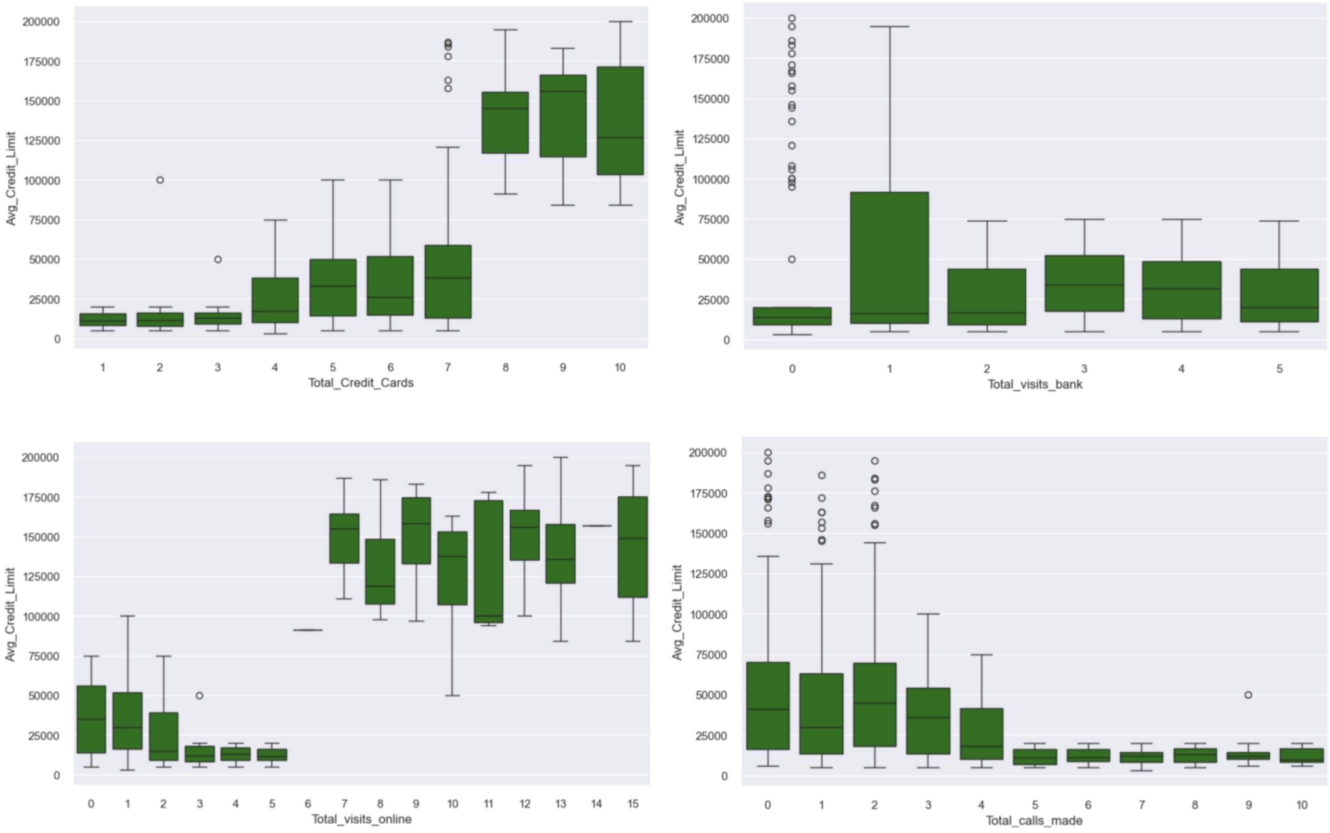


Figure 16: Bivariate Analysis – Avg_Credit_Limit vs. other Key Variables

✓ **Observations & Insights** can be summarized below: -

- **Avg_Credit_Limit vs. Total_Credit_Cards:**
    - As the number of credit cards increases, the average credit limit also increases.
    - Customers with 1–3 credit cards have relatively low credit limits.
    - Those with 7 or more cards tend to have significantly higher credit limits.
    - There are some outliers at higher credit card counts, indicating a few users with exceptionally high limits.
- **Avg_Credit_Limit vs. Total_visits_bank:**
    - Customers who visit the bank only once tend to have a wider range of credit limits, including very high values.
    - The median credit limit decreases slightly for those visiting the bank more frequently.
    - There are many outliers for users with one bank visit, suggesting that some high-credit customers prefer minimal in-person banking interactions.
    - Generally, higher credit limit customers don't visit the bank frequently.
- **Avg_Credit_Limit vs. Total_visits_online:**
    - Customers with higher online visits (6-15 visits) tend to have significantly higher credit limits.
    - Those who rarely use online banking (0-4 visits) have lower median credit limits.
    - This suggests that digitally active customers tend to have higher credit limits, possibly indicating higher financial literacy or better banking engagement.

- **Avg_Credit_Limit vs. Total_calls_made:**
  - The credit limit decreases as the number of calls made increases.
  - Customers who make more calls tend to have lower credit limits.
  - There are several outliers among customers making 0-2 calls, indicating that some high-limit customers do not need to contact customer service often.
  - This suggests that customers with higher credit limits require less direct bank assistance, likely due to better financial stability or understanding of their credit accounts.
- **Summary: -**
  - Customers with more credit cards tend to have higher credit limits.
  - More online visits correlate with higher credit limits, while frequent bank visits or calls correlate with lower credit limits.
  - Digitally active customers tend to have better credit access, while those relying on in-person banking and calls may have lower credit limits.
  - High-credit customers appear to be more self-sufficient and rely less on customer service.

Rubric Question 2: Data Preprocessing

Duplicate & Missing/Error Value-check

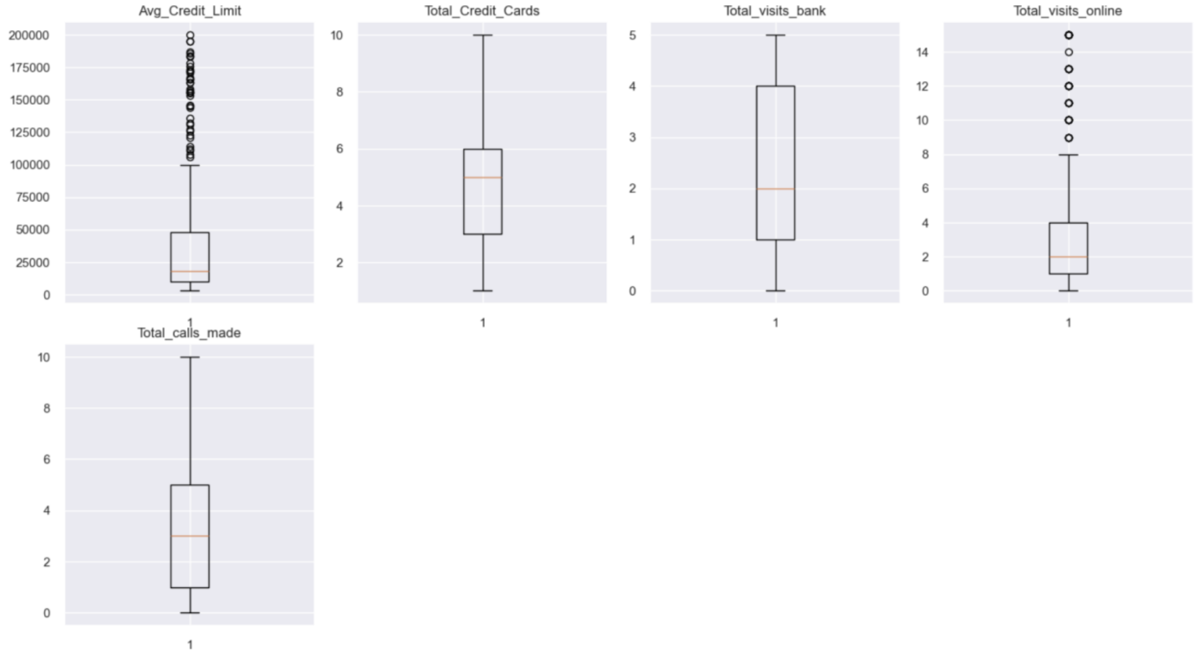
- Please refer Check & Treat Missing/Duplicate Values section.
- No treatment required as explained in the section above.

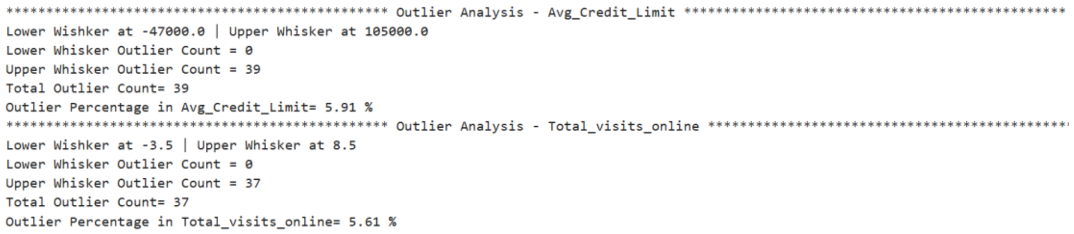
Feature Engineering

- No Action required. The attributes are explicable enough.

Outlier Treatment

- Below is a summary of the boxplots & outlier information for all the numerical variables: -





```
************************************************ Outlier Analysis - Avg_Credit_Limit ************************************************
Lower Wishker at -47000.0 | Upper Whisker at 105000.0
Lower Whisker Outlier Count = 0
Upper Whisker Outlier Count = 39
Total Outlier Count= 39
Outlier Percentage in Avg_Credit_Limit= 5.91 %
************************************************ Outlier Analysis - Total_visits_online ************************************************
Lower Wishker at -3.5 | Upper Whisker at 8.5
Lower Whisker Outlier Count = 0
Upper Whisker Outlier Count = 37
Total Outlier Count= 37
Outlier Percentage in Total_visits_online= 5.61 %
```

Figure 17: Outlier Inspection

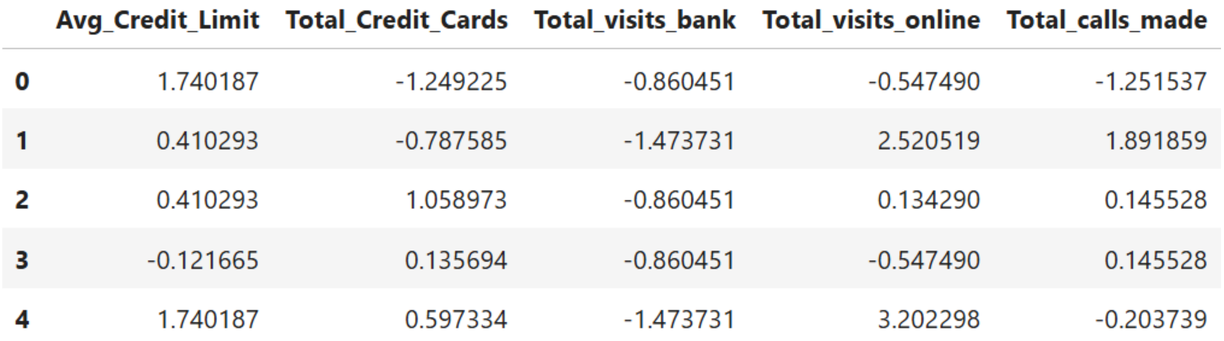
- It is evident that out of all the numerical variables, **‘Avg_Credit_Limit’ & ‘Total_visits_online’ have significant outliers (~6% each)**
- However, we choose **not to treat the outliers to avoid any loss of information**.

Label Encoding

- No action required since there is no categorical variable in the dataset.

Data Preparation for Modelling – Data Scaling

- Unsupervised Learning Models (K-Means & Hierarchical Clustering) are based on distances between data points; hence, we need to **standardize the dataset before building a model**.
- Below are the top 5 rows of the dataset post standardization: -



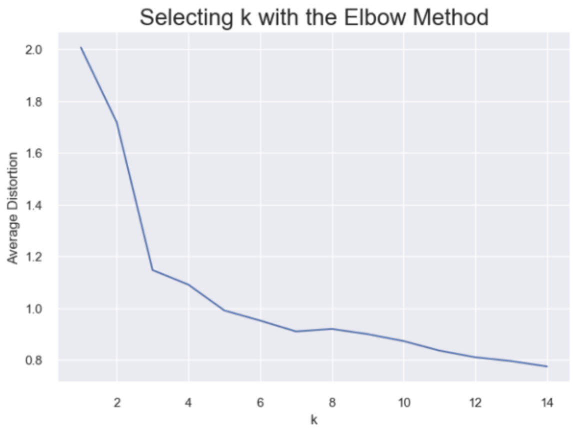
	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	1.740187	-1.249225	-0.860451	-0.547490	-1.251537
1	0.410293	-0.787585	-1.473731	2.520519	1.891859
2	0.410293	1.058973	-0.860451	0.134290	0.145528
3	-0.121665	0.135694	-0.860451	-0.547490	0.145528
4	1.740187	0.597334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

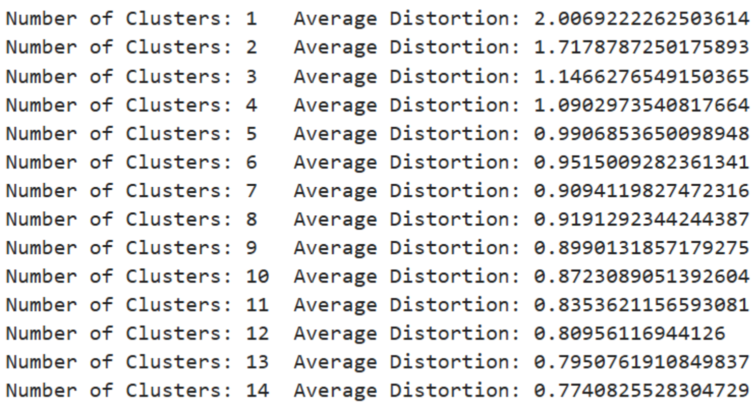
Figure 18: Data Scaling (Standardization of dataset)

Rubric Question 3: Model building – K-means Clustering

Checking Elbow Plot

- Run the model for different values of K (no. of clusters) & plot distribution between Average Distortion & no. of Clusters.





```
Number of Clusters: 1   Average Distortion: 2.0069222262503614
Number of Clusters: 2   Average Distortion: 1.7178787250175893
Number of Clusters: 3   Average Distortion: 1.1466276549150365
Number of Clusters: 4   Average Distortion: 1.0902973540817664
Number of Clusters: 5   Average Distortion: 0.9906853650098948
Number of Clusters: 6   Average Distortion: 0.9515009282361341
Number of Clusters: 7   Average Distortion: 0.9094119827472316
Number of Clusters: 8   Average Distortion: 0.9191292344244387
Number of Clusters: 9   Average Distortion: 0.8990131857179275
Number of Clusters: 10  Average Distortion: 0.8723089051392604
Number of Clusters: 11  Average Distortion: 0.8353621156593081
Number of Clusters: 12  Average Distortion: 0.80956116944126
Number of Clusters: 13  Average Distortion: 0.7950761910849837
Number of Clusters: 14  Average Distortion: 0.7740825528304729
```

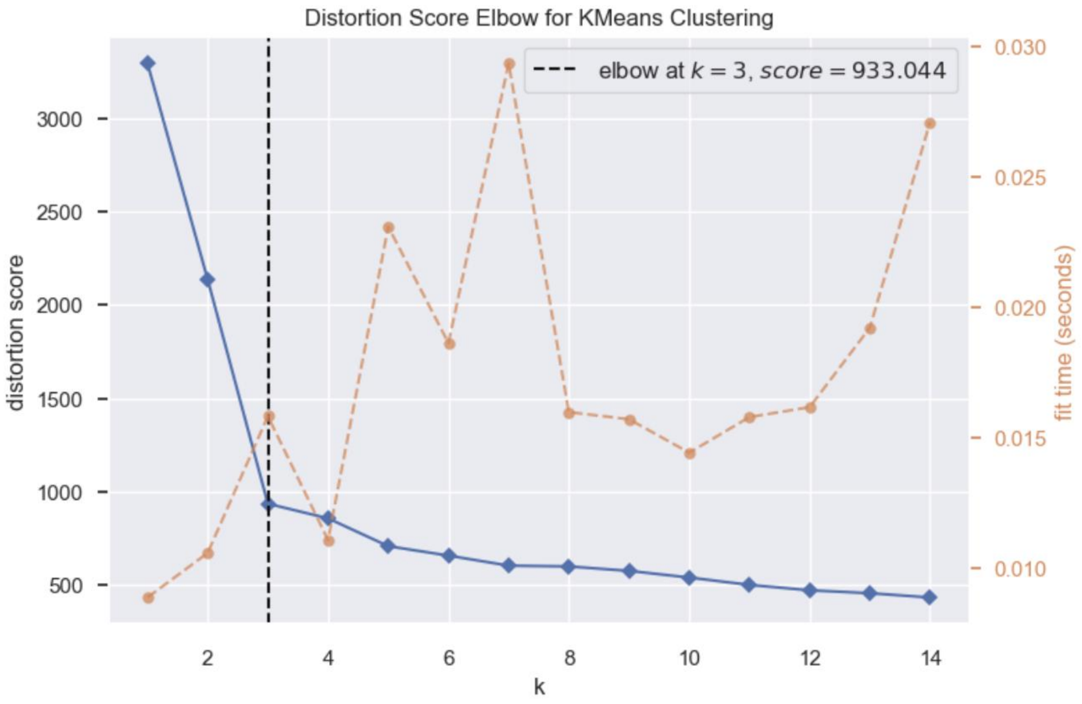
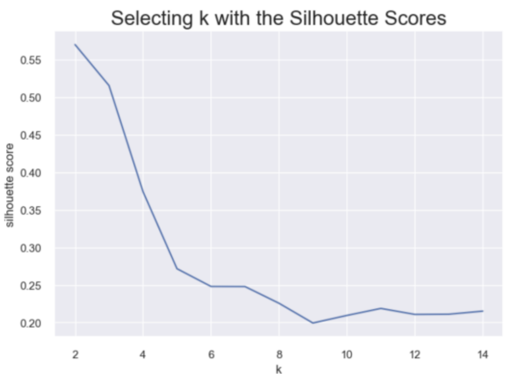


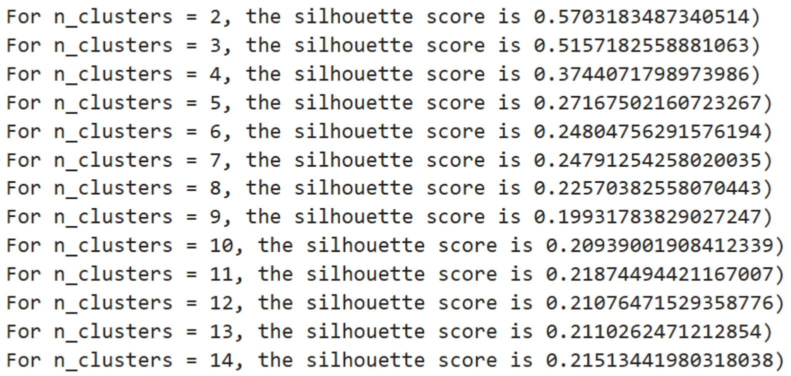
Figure 19: Elbow Plot to determine K (Clusters)

- From the above visualization it is evident that **K = 3 is most appropriate**, **optimizing between distortion score & the time taken to fit the model.**

Checking Silhouette Scores

- Run the model for different values of K (no. of clusters) & its corresponding Silhouette Scores





```
For n_clusters = 2, the silhouette score is 0.5703183487340514)
For n_clusters = 3, the silhouette score is 0.5157182558881063)
For n_clusters = 4, the silhouette score is 0.3744071798973986)
For n_clusters = 5, the silhouette score is 0.27167502160723267)
For n_clusters = 6, the silhouette score is 0.24804756291576194)
For n_clusters = 7, the silhouette score is 0.24791254258020035)
For n_clusters = 8, the silhouette score is 0.22570382558070443)
For n_clusters = 9, the silhouette score is 0.19931783829027247)
For n_clusters = 10, the silhouette score is 0.20939001908412339)
For n_clusters = 11, the silhouette score is 0.21874494421167007)
For n_clusters = 12, the silhouette score is 0.21076471529358776)
For n_clusters = 13, the silhouette score is 0.2110262471212854)
For n_clusters = 14, the silhouette score is 0.21513441980318038)
```

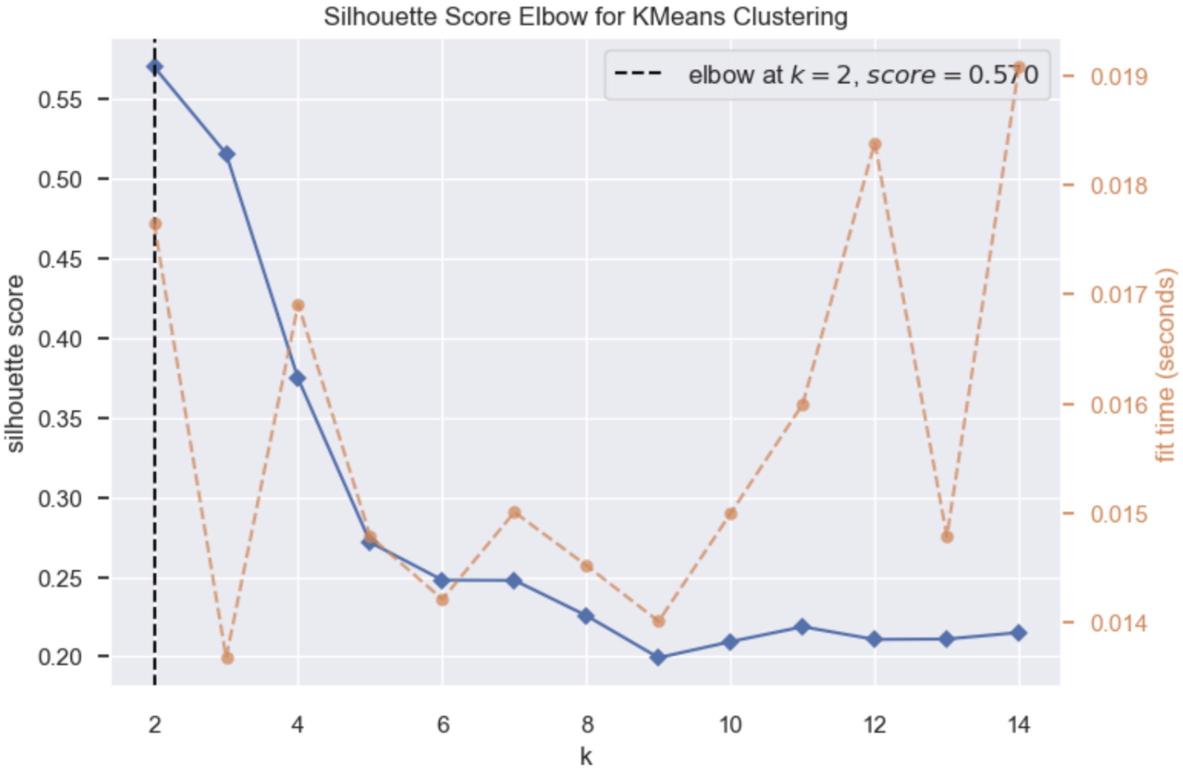
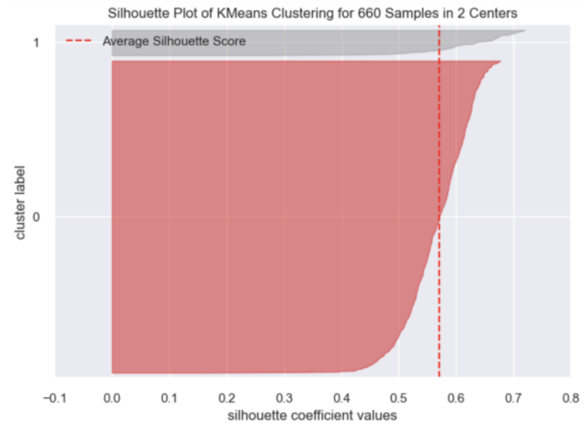
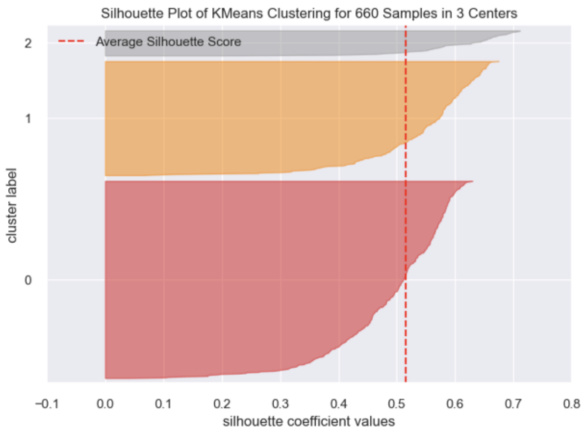


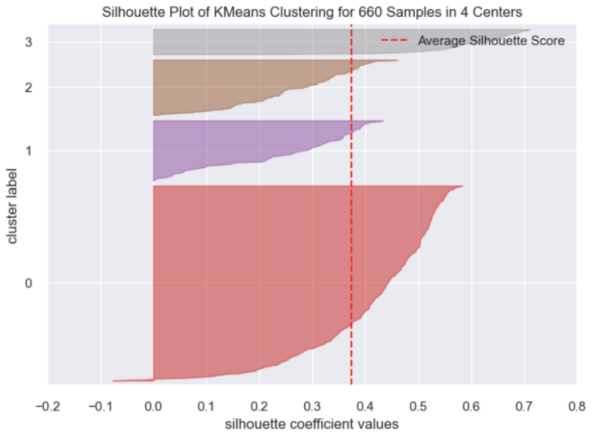
Figure 20: Silhouette Score Plot to determine K (Clusters)

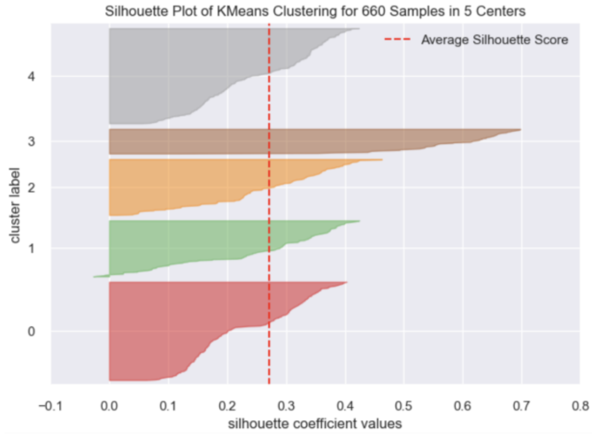
- From the above visualization it is evident that **K = 2 is most appropriate**, **optimizing between Silhouette Score & the time taken to fit the model.**

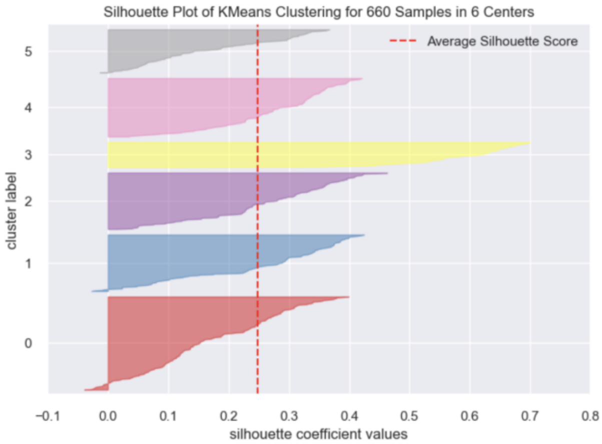
- Let’s further breakdown the visualization to analyse Silhouette Score for each cluster for the given range of K: -

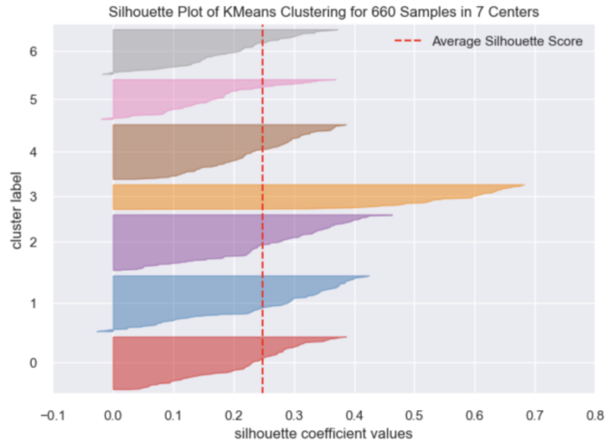












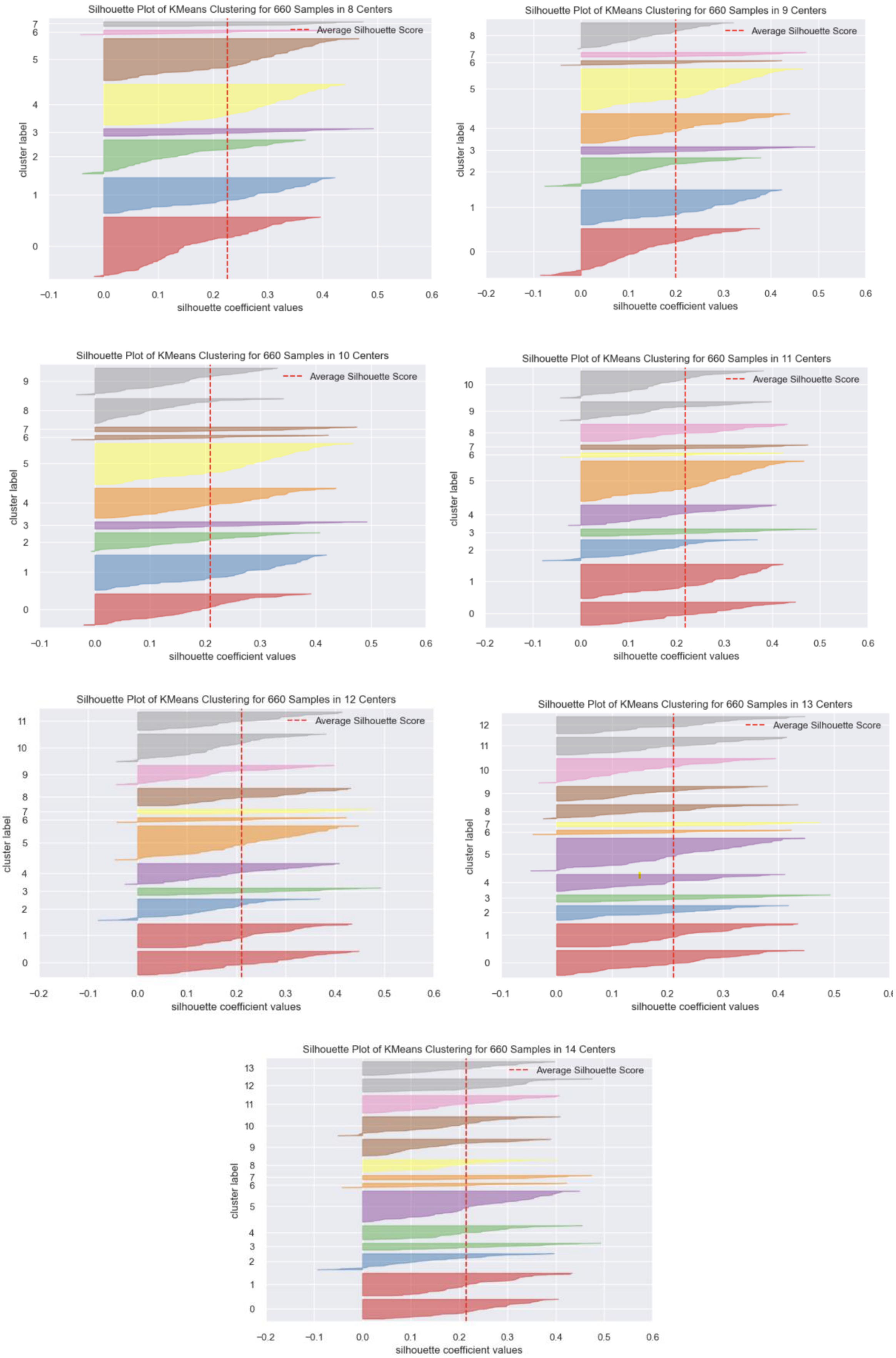
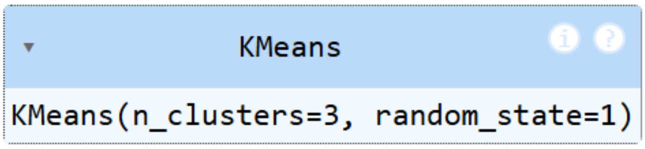


Figure 21: Silhouette Score Plot for different K (Clusters) – Cluster-wise Distribution

- From the above set of visualizations, it is evident that **K = 3 is most appropriate**, **showing distinction between various clusters.**

Creating Final Model – K-Means Clustering

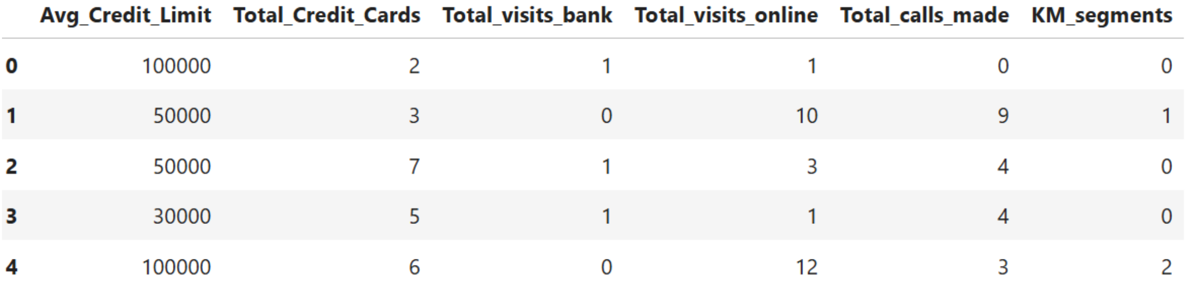
- From all the above visualizations (Elbow Plot & Silhouette Score Plots), it is evident that **K = 3 (no. of clusters) is most appropriate** as it strikes optimal balance: -
  - Between **Distortion Score & Fit Time**
  - Between **Silhouette Score & Fit Time**
  - **Spread of inter-cluster Silhouette Score** to establish proper distinction
- Build K-Means model with 3 clusters (K=3): -



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

Figure 22: K-Means Clustering Model (K=3)

- Adding Cluster Segments to the dataset. Below are the top 5 rows of the dataset post K-Means Clustering: -

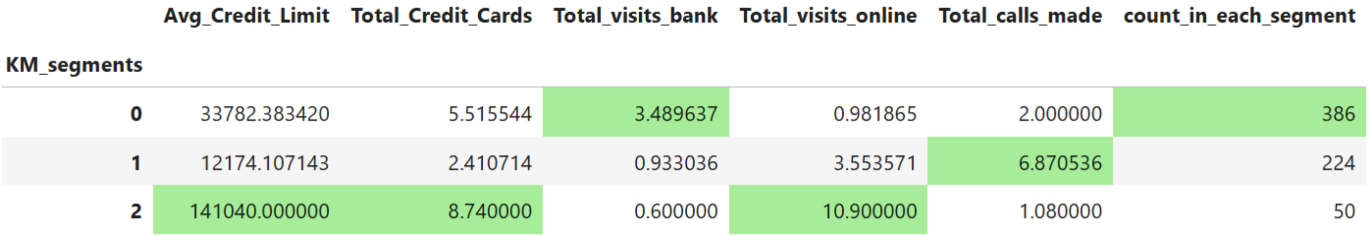


	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	KM_segments
0	100000	2	1	1	0	0
1	50000	3	0	10	9	1
2	50000	7	1	3	4	0
3	30000	5	1	1	4	0
4	100000	6	0	12	3	2

Figure 23: Dataset with Segments post K-Means Clustering

Customer Profiling – K-Means Clustering

- Based on the clusters created, lets analyse the features by taking: -
  - **Mean** of each feature by clusters
  - **Count** of customers in each cluster



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

Figure 24: Customer Profiling (K-Means) – Feature Table

- Visualizing by creating Boxplots of numerical variables for each cluster: -

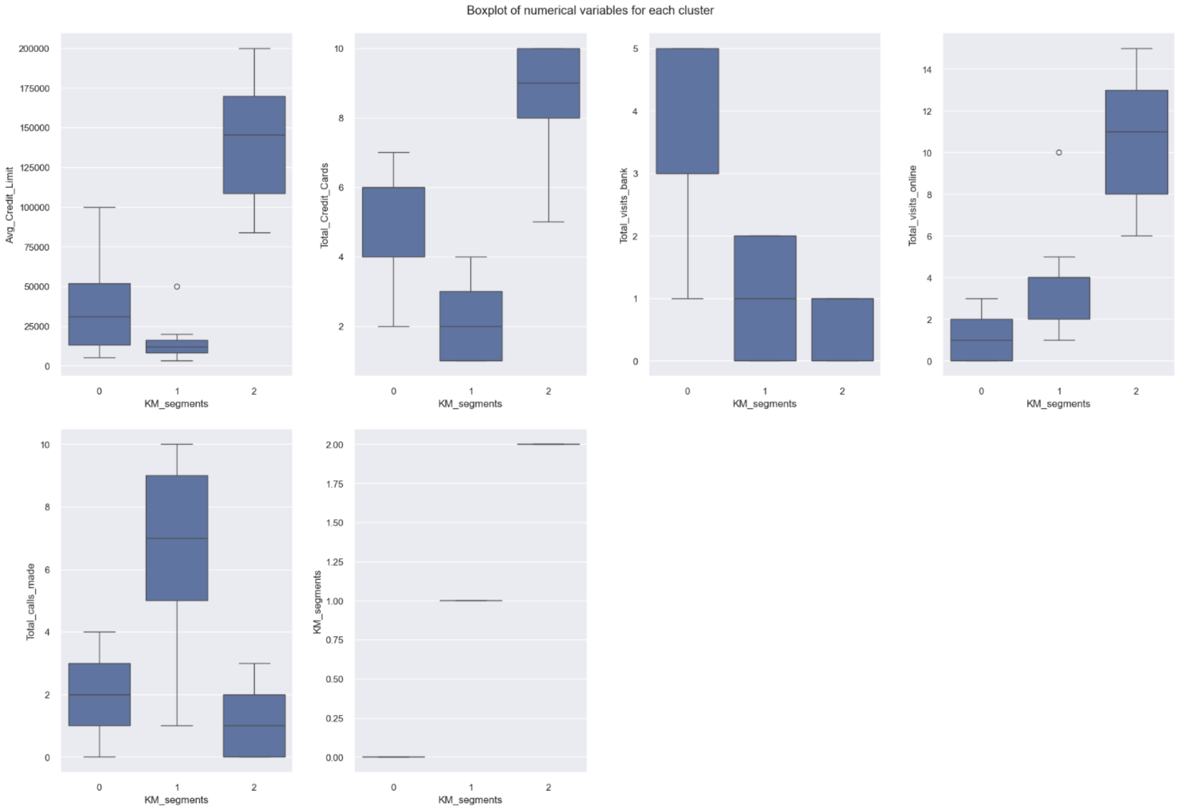


Figure 25: Customer Profiling (K-Means) – Feature BoxPlots

- **Customer Segments Explained: -**

I. **Segment 0: Mid-Range Customers with Balanced Behaviour**
   - Avg. Credit Limit: $33,782
   - Total Credit Cards: ~5.5
   - Total Visits to Bank: ~3.49
   - Total Visits Online: ~0.98
   - Total Calls Made: ~2
   - Segment Size: 386 customers (largest segment)
     - **Customer Behaviour & Insights:**
       - ✓ Customers in this segment have a moderate credit limit and hold an average of 5-6 credit cards.
       - ✓ They visit the bank more frequently than other segments but do not engage much online.
       - ✓ Their call centre interactions are relatively low, indicating they might prefer in-person services.
     - **Marketing & Service Strategy:**
       - ✓ **Enhance physical branch services:** Since these customers prefer branch visits, ensure faster in-person service and exclusive in-branch offers.
       - ✓ **Encourage online engagement:** Promote digital banking services and offer incentives (discounts, cashback) for online transactions.
       - ✓ **Personalized credit offers:** Upsell credit limit enhancements or financial products suited to their spending patterns.

II. **Segment 1: Low Credit, High Support Seekers**
   - Avg. Credit Limit: $12,174
   - Total Credit Cards: ~2.4
   - Total Visits to Bank: ~0.93
   - Total Visits Online: ~3.55
   - Total Calls Made: 6.87 (highest)
   - Segment Size: 224 customers

- **Customer Behaviour & Insights:**
  - ✓ Customers in this segment have the lowest credit limits and the fewest credit cards.
  - ✓ They prefer online banking but also frequently call customer support.
  - ✓ Bank visits are minimal, indicating they prefer remote banking options.
- **Marketing & Service Strategy:**
  - ✓ **Improve call centre support**: Since these customers frequently contact the call centre, automating common queries with chatbots or self-service options can improve efficiency.
  - ✓ **Promote online banking tutorials**: Encourage customers to use online banking tools effectively to reduce dependency on call support.
  - ✓ **Offer financial literacy programs**: Help them understand credit utilization and benefits of increased credit limits, potentially leading to higher product engagement.

III. **Segment 2: High-Value, Digital-First Customers**

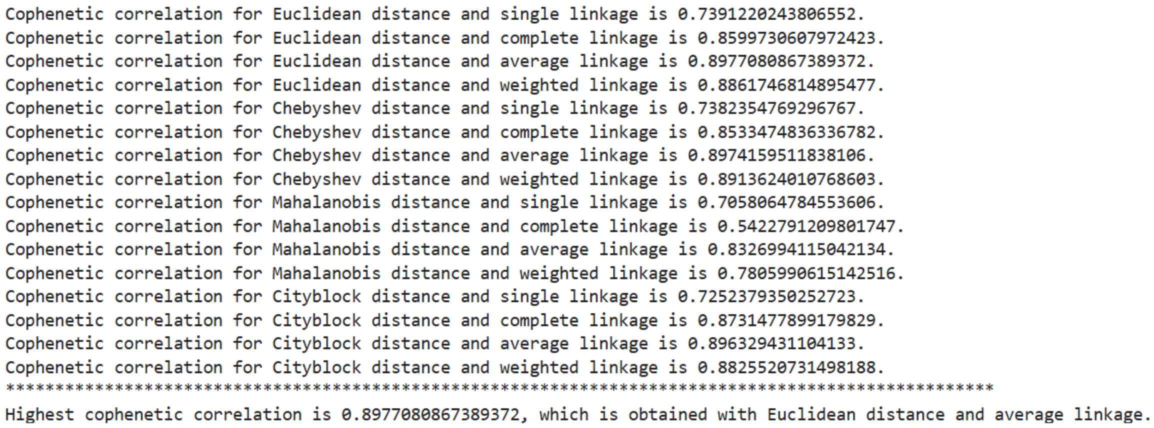
- Avg. Credit Limit: $141,040 (highest)
- Total Credit Cards: ~8.74 (highest)
- Total Visits to Bank: 0.6 (lowest)
- Total Visits Online: 10.9 (highest)
- Total Calls Made: 1.08 (lowest)
- Segment Size: 50 customers (smallest segment)
  - **Customer Behaviour & Insights:**
    - ✓ These are high-value customers with large credit limits and multiple credit cards.
    - ✓ They are primarily digital users, engaging heavily through online banking while avoiding physical branches.
    - ✓ Their call centre interactions are minimal, indicating self-sufficiency in managing their accounts.
  - **Marketing & Service Strategy:**
    - ✓ **Exclusive digital privileges**: Offer premium online banking features, dedicated relationship managers, or priority customer support.
    - ✓ **Personalized rewards & offers**: Provide exclusive cashback, premium card upgrades, and investment advisory services.
    - ✓ **Expand digital banking capabilities**: Ensure that the bank's digital services cater to their needs with advanced security features, seamless transactions, and AI-driven financial insights.

- **Customer Behaviour & Insights:**
  - ✓ Customers in this segment have a moderate credit limit and hold an average of 5-6 credit cards.
  - ✓ They visit the bank more frequently than other segments but do not engage much online.
  - ✓ Their call centre interactions are relatively low, indicating they might prefer in-person services.
- **Marketing & Service Strategy:**
  - ✓ **Enhance physical branch services:** Since these customers prefer branch visits, ensure faster in-person service and exclusive in-branch offers.
  - ✓ **Encourage online engagement:** Promote digital banking services and offer incentives (discounts, cashback) for online transactions.
  - ✓ **Personalized credit offers:** Upsell credit limit enhancements or financial products suited to their spending patterns.

Rubric Question 4: Model building – Hierarchical Clustering

Computing Cophenetic Correlation

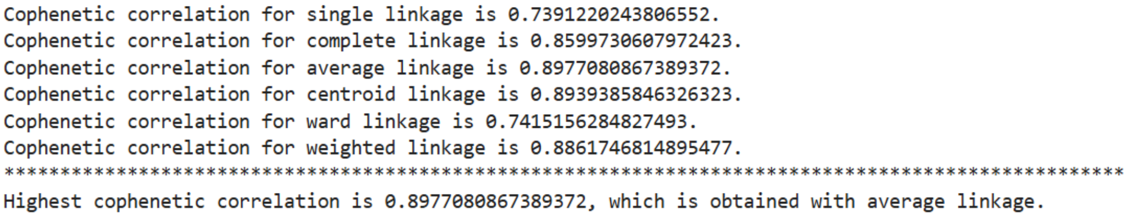
- Let's compute **Cophenetic Correlation** for all possible combinations of **Distance** Metrics (Euclidean, Chebyshev, Mahalanobis, Cityblock) & **Linkage** Methods (Single, Complete, Average, Weighted) to **evaluate the most appropriate combination for building Agglomerative Clustering Model: -**



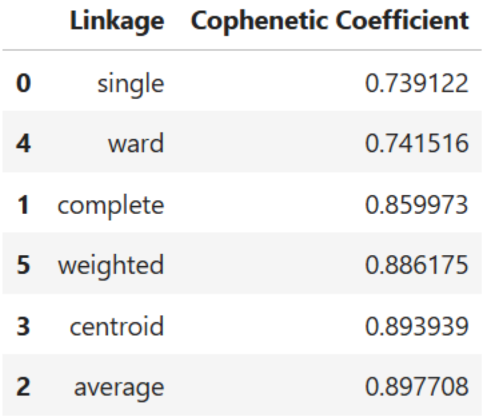
```
Cophenetic correlation for Euclidean distance and single linkage is 0.7391220243806552.
Cophenetic correlation for Euclidean distance and complete linkage is 0.8599730607972423.
Cophenetic correlation for Euclidean distance and average linkage is 0.8977080867389372.
Cophenetic correlation for Euclidean 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 0.7058064784553606.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.5422791209801747.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.8326994115042134.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.7805990615142516.
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.
****************************************************************************************************
Highest cophenetic correlation is 0.8977080867389372, which is obtained with Euclidean distance and average linkage.
```

Figure 26: Cophenetic Correlation for various Distance-Linkage Combination

- Clearly, the **Euclidean distance** with **Average linkage** gives the highest Cophenetic Correlation.
- Let's explore all the **possible linkages for Euclidean distance: -**



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



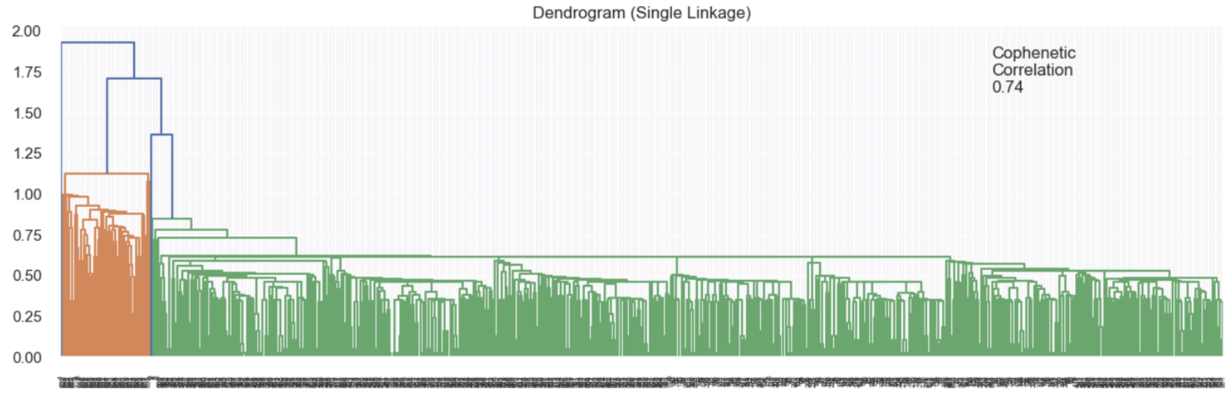
	Linkage	Cophenetic Coefficient
0	single	0.739122
4	ward	0.741516
1	complete	0.859973
5	weighted	0.886175
3	centroid	0.893939
2	average	0.897708

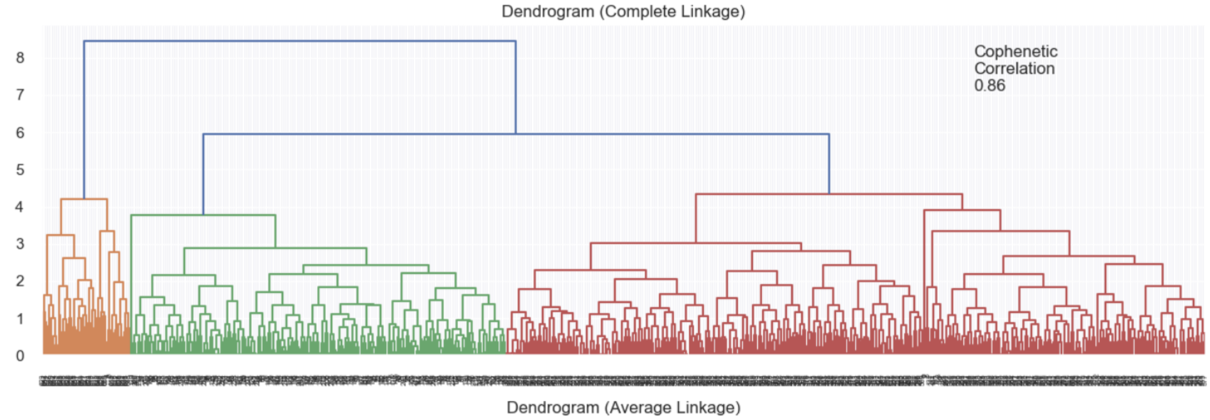
Figure 27: Cophenetic Correlation for all Linkages (Euclidean-distance)

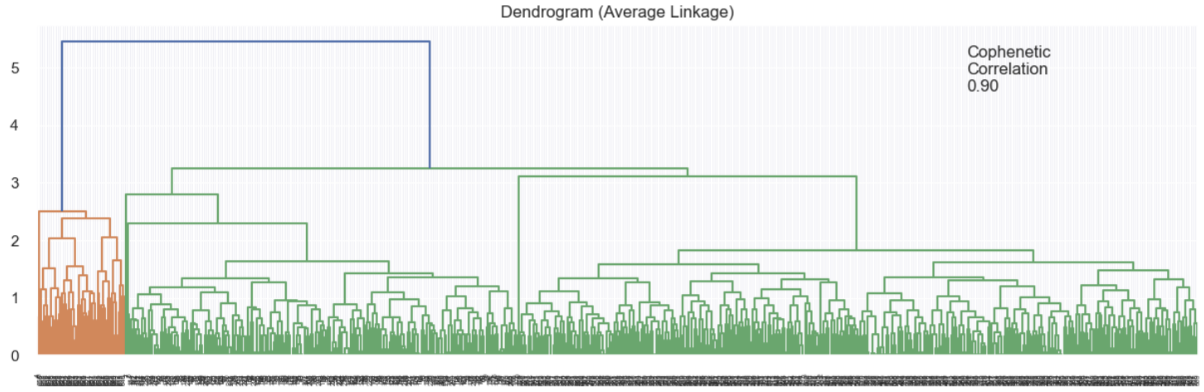
- As evident above, the best combination is with **Euclidean distance** & **Average linkage: -**

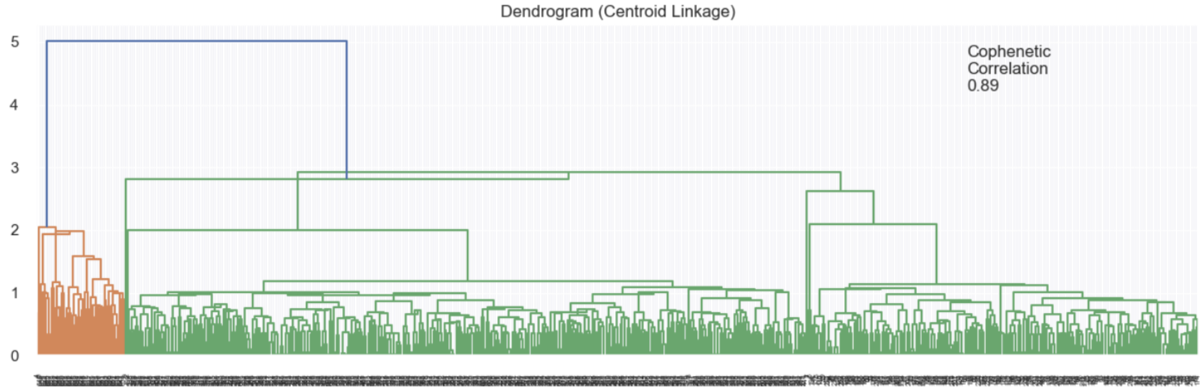
Creating Dendrograms

- Let’s create **Dendograms** for all the Linkage methods with **Euclidean distance: -**









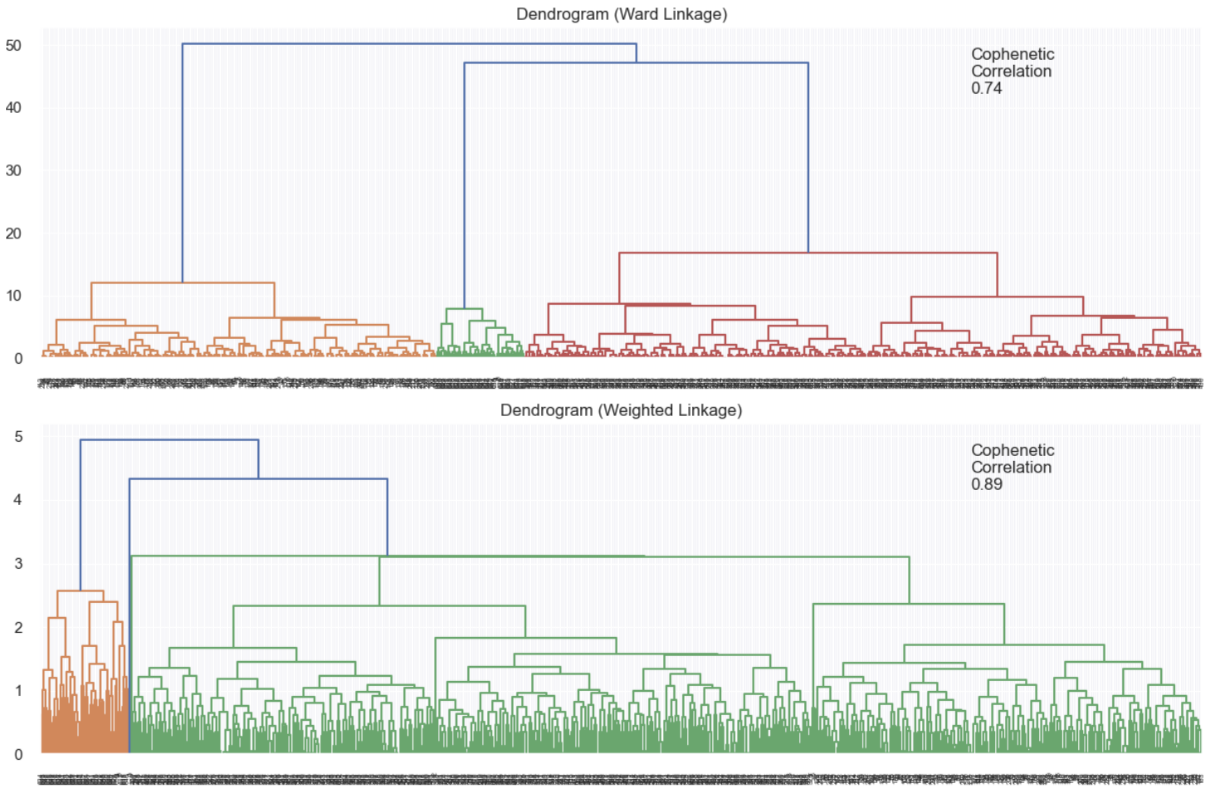
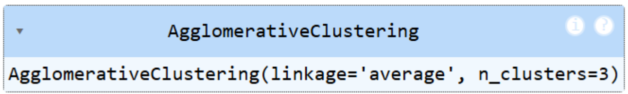


Figure 28: Dendograms with different Linkages

Creating Final Model – Hierarchical Clustering (Agglomerative Clustering)

- To **select appropriate K (clusters) from Dendograms**, the following can be noted: -
  - ✓ Between the Dendograms with various linkages, it is evident that the one with **Average linkage** gives the **highest Cophenetic Correlation.**
  - ✓ **Identify the Largest Vertical Gaps** between horizontal lines in the **Dendrogram with Average linkage.** These gaps indicate distances between clusters that are more distinct. Cutting the dendrogram at these large gaps will result in well-separated clusters.
  - ✓ Select a **height on the dendrogram where the clusters below the cut are meaningful**, and the number of clusters is manageable.
- **3 clusters seem appropriate** because the dendrogram likely shows a clear separation at that level.
- There is likely a noticeable gap between the merges forming 3 clusters and the merges forming 2 clusters. Selecting 3 clusters aligns with the natural structure in the data while keeping the clusters meaningful.
- Build **Agglomerative Clustering** model with **3 clusters (K=3)**: -



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

Figure 29: Agglomerative Clustering Model (K=3)

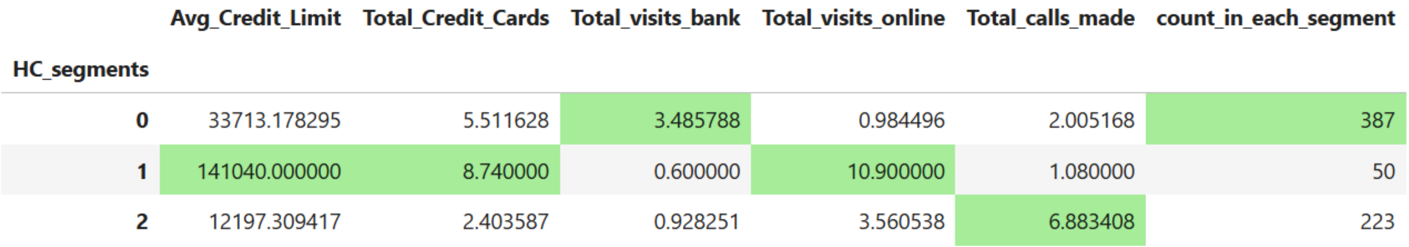
- Adding Cluster Segments to the dataset. Below are the top 5 rows of the dataset post Agglomerative Clustering: -

	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	HC_segments
0	100000	2	1	1	0	0
1	50000	3	0	10	9	2
2	50000	7	1	3	4	0
3	30000	5	1	1	4	0
4	100000	6	0	12	3	1

Figure 30: Dataset with Segments post Agglomerative Clustering

Customer Profiling – Hierarchical Clustering (Agglomerative Clustering)

- Based on the clusters created, lets analyse the features by taking: -
  - **Mean** of each feature by clusters
  - **Count** of customers in each cluster



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

Figure 31: Customer Profiling (Agglomerative) – Feature Table

- Visualizing by creating Boxplots of numerical variables for each cluster: -

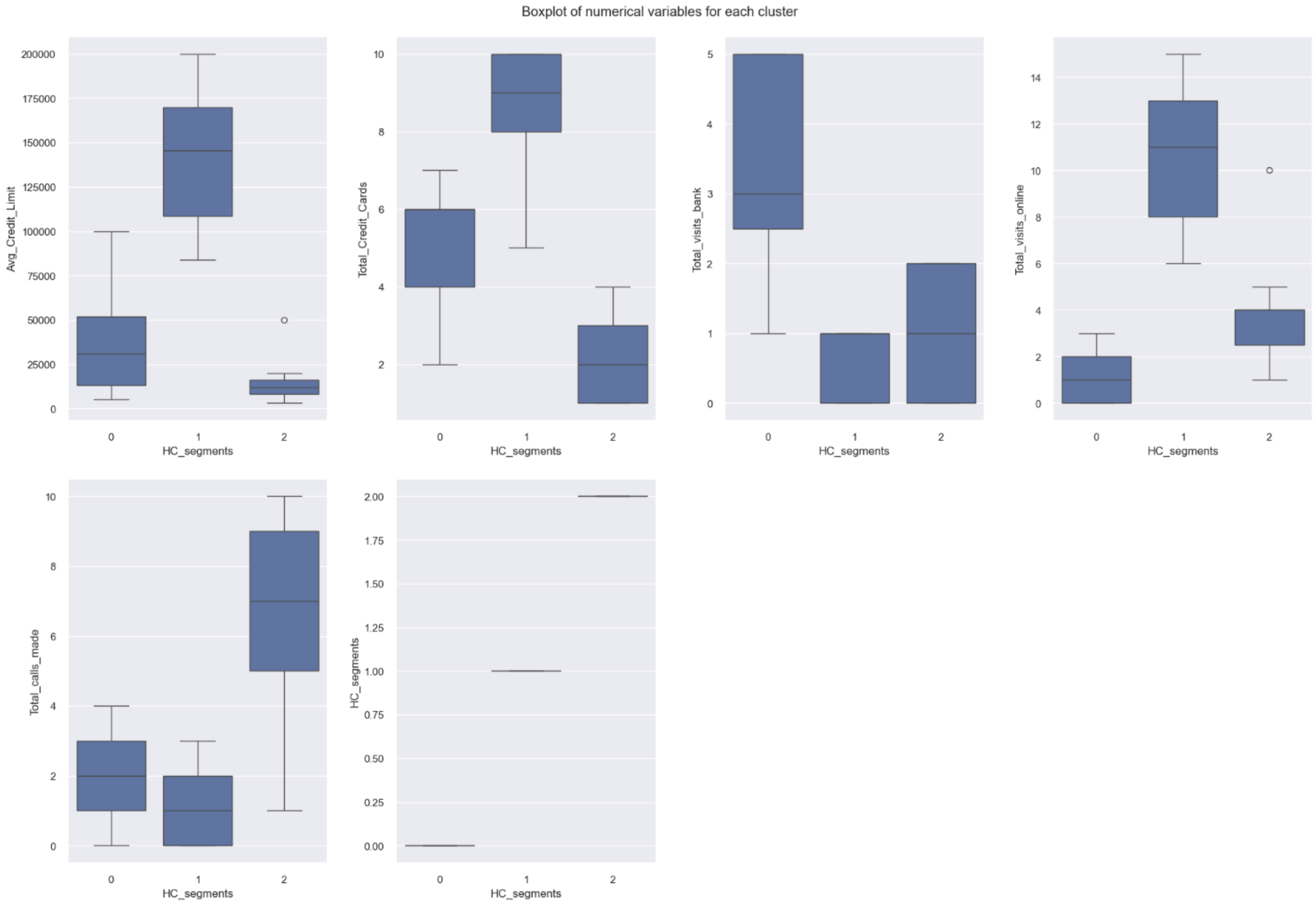


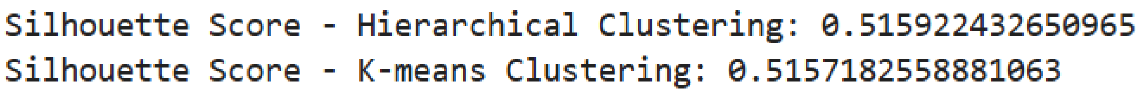
Figure 32: Customer Profiling (Agglomerative) – Feature BoxPlots

- **Customer Segments Explained: -**
  - I. **Segment 0: Mid-Range Customers with Balanced Banking Behavior**
    - Avg. Credit Limit: $33,713
    - Total Credit Cards: ~5.5
    - Total Visits to Bank: ~3.49
    - Total Visits Online: ~0.98
    - Total Calls Made: ~2.01
    - Segment Size: 387 customers (largest group)
      - ➢ **Customer Behaviour & Insights:**
        - ✓ These customers have a moderate credit limit and hold around 5-6 credit cards.
        - ✓ They visit bank branches frequently, showing a preference for face-to-face interactions.
        - ✓ Minimal online activity and low call centre engagement indicate traditional banking behaviour.
      - ➢ **Marketing & Service Strategy:**
        - ✓ **Enhance in-branch services**: Improve staff training and ensure faster resolution times at physical branches.
        - ✓ **Increase digital adoption**: Educate customers on online banking benefits, offering special incentives for digital usage.
        - ✓ **Upsell premium products**: Promote additional credit cards, loans, or investment products based on their spending behaviour.
  - II. **Segment 1: High-Value, Digital-First Customers**
    - Avg. Credit Limit: $141,040 (highest)
    - Total Credit Cards: ~8.74 (highest)
    - Total Visits to Bank: 0.6 (lowest)
    - Total Visits Online: 10.9 (highest)
    - Total Calls Made: 1.08 (lowest)
    - Segment Size: 50 customers (smallest group)
      - ➢ **Customer Behaviour & Insights:**
        - ✓ These high-net-worth customers have significant credit limits and multiple credit cards.
        - ✓ They rarely visit branches and instead rely on digital banking.
        - ✓ Their low call centre interactions suggest a high level of financial independence and digital literacy.
      - ➢ **Marketing & Service Strategy:**
        - ✓ **Exclusive digital banking services**: Provide premium digital features like personalized dashboards, AI-driven investment insights, and seamless fund transfers.
        - ✓ **Loyalty and premium rewards programs**: Offer higher cashback, exclusive airport lounge access, concierge services, and premium credit card upgrades.
        - ✓ **AI-powered relationship management**: Introduce a dedicated virtual assistant or high-net-worth relationship managers for instant problem resolution.
  - III. **Segment 2: Low Credit, High Support-Seekers**
    - Avg. Credit Limit: $12,197 (lowest)
    - Total Credit Cards: ~2.4
    - Total Visits to Bank: ~0.93
    - Total Visits Online: ~3.56
    - Total Calls Made: 6.88 (highest)
    - Segment Size: 223 customers
      - ➢ **Customer Behaviour & Insights:**
        - ✓ These customers have low credit limits and few credit cards.
        - ✓ They engage more via online banking but also frequently call customer support.
        - ✓ Their low in-person visits suggest they prefer remote banking, but they require a lot of assistance.
      - ➢ **Marketing & Service Strategy:**
        - ✓ **Optimize customer support**: Introduce AI chatbots for self-service, reducing the load on call centres.
        - ✓ **Educate customers on digital banking**: Offer video tutorials, personalized onboarding, and step-by-step guidance to help them navigate digital banking.
        - ✓ **Encourage credit engagement**: Provide pre-approved credit limit increases, secured credit cards, and financial literacy content.

Rubric Question 5: K-Means vs Hierarchical Clustering

Silhouette Scores – K-Means & Hierarchical Compared

- Comparing the Silhouette Scores for both Models: -



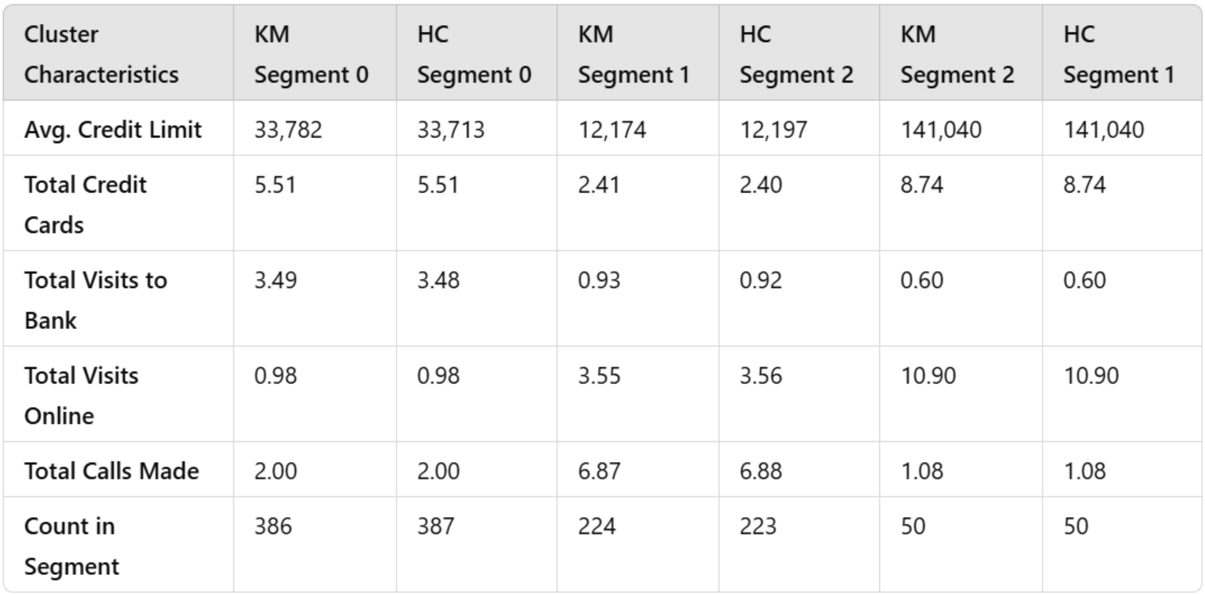
```
Silhouette Score - Hierarchical Clustering: 0.515922432650965
Silhouette Score - K-means Clustering: 0.5157182558881063
```

Figure 33: Silhouette Scores | K-Means vs. Hierarchical

- Clearly, the Silhouette Scores for both models are almost identical.

Customer Segment Comparison between K-Means & Hierarchical Clustering

- Both **K-Means (KM)** and **Hierarchical Clustering (HC)** produced **three customer segments** with similar behavioural patterns, but there are almost negligible variations in the cluster assignments: -



Cluster Characteristics	KM Segment 0	HC Segment 0	KM Segment 1	HC Segment 2	KM Segment 2	HC Segment 1
Avg. Credit Limit	33,782	33,713	12,174	12,197	141,040	141,040
Total Credit Cards	5.51	5.51	2.41	2.40	8.74	8.74
Total Visits to Bank	3.49	3.48	0.93	0.92	0.60	0.60
Total Visits Online	0.98	0.98	3.55	3.56	10.90	10.90
Total Calls Made	2.00	2.00	6.87	6.88	1.08	1.08
Count in Segment	386	387	224	223	50	50

Table 2: Customer Segment Comparison | K-Means & Hierarchical

- The segment structure is **nearly identical** in both models. So, the Customer Profiling remains the same for both. Please refer Actionable Insights & Recommendations Section for the Final Segmentation.
- **Hierarchical Clustering assigned one extra customer to Segment 0** compared to K-Means.
- The **characteristics of each cluster remain consistent**.

Rubric Question 6: Actionable Insights & Recommendations

- Combining results from both the models (which are almost identical), below is the **final Customer Segmentation**: -
  I. **Segment 0: Mid-Tier, Traditional Banking Customers:**
     - Avg. Credit Limit: ~$33,700
     - Total Credit Cards: ~5.5
     - **Preference**: More in-person banking, minimal online usage, low call centre dependency
     - **Marketing Strategy**:
       - ✓ Improve in-branch service efficiency.
       - ✓ Educate customers on digital banking benefits.
       - ✓ Offer personalized financial products (e.g., home loans, savings accounts).

  II. **Segment 1: Low Credit, High Support Seekers:**
     - Avg. Credit Limit: ~$12,200
     - Total Credit Cards: ~2.4
     - **Preference**: Moderate online engagement, high call centre interactions
     - **Marketing Strategy**:
       - ✓ Reduce call centre dependency with AI-powered chatbots.
       - ✓ Offer credit-building programs & financial literacy campaigns.
       - ✓ Provide specialized support teams for frequent queries.

  III. **Segment 2: High-Value, Digital-First Customers:**
     - Avg. Credit Limit: $141,040
     - Total Credit Cards: 8.74
     - **Preference**: Heavy online engagement, rarely visits banks, low call center dependency
     - **Marketing Strategy**:
       - ✓ Provide VIP digital services, exclusive investment opportunities.
       - ✓ Offer personalized relationship managers.
       - ✓ Incentivize premium financial products, such as high-limit credit cards & wealth management tools.

- **Insights on the Model: -**
  - Both clustering models yielded nearly identical segments, confirming the robustness of the results.
  - K-Means is computationally faster, making it more suitable for large datasets.
  - Hierarchical Clustering provides better interpretability by showing how customers are linked hierarchically.
  - **Final Recommendation**: Adopt K-Means for ongoing customer segmentation, but use Hierarchical Clustering for deeper customer behaviour insights.

- **Business Recommendations Based on Customer Profiling: -**

Recommendation	Identified Issue	Actionable Insights
Enhancing Customer Service Model **Expected Impact:** Improved customer satisfaction, reduced call centre costs, and faster issue resolution.	Customers perceive AllLife Bank's support services poorly. Segment 1 has the highest call centre interactions, indicating dissatisfaction or a need for better self-service solutions.	✓ **AI-Powered Chatbots & Self-Service Portals:** Reduce call centre dependency for Segment 1 (Low Credit, High Support Seekers). ✓ **Dedicated Relationship Managers**: Assign to Segment 2 (High-Value Customers) for premium customer service. ✓ **Omnichannel Support**: Ensure seamless experience between branch visits, online banking, and call centres.
Personalized Marketing & Product Offerings **Expected Impact:** Increased customer engagement, higher cross-selling success, and improved brand loyalty.	The bank wants to expand market penetration through personalized marketing campaigns.	**Segment 0: Mid-Tier, Traditional Banking Customers:** ✓ **Promote Digital Banking**: Offer incentives (e.g., cashback on online transactions) to increase online banking adoption. ✓ **Upsell Credit Products**: Offer higher credit limits and additional credit cards. **Segment 1: Low Credit, High Support Seekers:** ✓ **Credit Education & Financial Literacy**: Provide interactive tools to help them improve credit scores and responsible credit utilization. ✓ **Pre-Approved Loan Offers**: Encourage increased spending by offering personal loans or secured credit cards. **Segment 2: High-Value, Digital-First Customers:** ✓ **Exclusive VIP Services**: Introduce premium credit cards, wealth management advisory, and tailored investment products. ✓ **Loyalty & Rewards Program**: Offer cashback, travel perks, and personalized financial insights.
Optimizing Branch & Digital Banking Services **Expected Impact:** Reduced operational costs, increased digital banking adoption, and improved service accessibility.	Different customer segments have varying banking preferences – some prefer in-branch visits, while others are digital-first.	✓ **For Segment 0: Mid-Tier, Traditional Banking Customers:** Improve in-branch experience by enhancing service efficiency and reducing wait times. ✓ **For Segment 1: Low Credit, High Support Seekers**: Enhance mobile banking UX to make self-service easier and reduce reliance on customer support. ✓ **For Segment 2**: **High-Value, Digital-First Customers**: Invest in AI-powered financial assistants to provide real-time insights and better financial decision-making.
Customer Retention & Upselling Strategy **Expected Impact:** Increased credit card adoption, higher retention rates, and stronger customer loyalty.	The bank wants to upsell to existing customers and improve customer lifetime value (CLV).	✓ **Personalized Credit Limit Increases**: Use machine learning models to identify customers eligible for higher limits. ✓ **Loyalty & Referral Programs**: Encourage high-value customers (Segment 2) to refer new clients by offering exclusive rewards. ✓ **Gamification of Banking Services**: Introduce spending challenges & rewards for Segment 1 & 0 to increase credit card usage.
Competitor Benchmarking & Market Expansion **Expected Impact**: Better market positioning, higher brand awareness, and expanded customer base.	The bank aims to increase market share and improve competitiveness.	✓ **Benchmark Against Competitors**: Analyse what competitors offer in terms of credit limits, customer support, and digital features. ✓ **Expand Digital-Only Banking Services**: Launch a dedicated digital banking platform targeting Segment 2. ✓ **Target Younger Demographics**: Leverage social media marketing to attract tech-savvy millennials & Gen Z customers.

Table 3: Business Recommendations & Actionable Insights

- **Final Takeaway: Data-Driven Banking Transformation**
  - Short-Term Goals (0-6 months):
    - ✓ Improve call centre & branch efficiency
    - ✓ Implement AI-powered self-service options
    - ✓ Personalize marketing campaigns
  - **Mid-Term Goals (6-12 months):**
    - ✓ Launch premium banking services for high-value customers
    - ✓ Strengthen loyalty programs & referral incentives
    - ✓ Enhance digital banking UI/UX
  - **Long-Term Goals (12+ months):**
    - ✓ Expand market reach through digital banking innovation
    - ✓ Adopt predictive analytics for financial insights
    - ✓ Continuously refine customer segmentation models