UNSUPERVISED LEARNING

PROJECT WORK - 5

Submitted by
Ms. Ann Raichel John

Enhancing Market Penetration and Customer Service at AllLife Bank

AllLife Bank wants to increase credit card penetration and improve customer service. The goal is to identify customer segments for targeted marketing and analyze service issues to enhance customer satisfaction.

Objective:

To segment existing customers based on spending patterns and past interactions using clustering algorithms and provide recommendations for better marketing and service strategies.

Exploratory Data Analysis

Data Preprocessing

K-Means Clustering

Hierarchical Clustering

Data description

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

Data Overview

```
<class 'pandas.core.frame.DataFrame'>
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)
memory usage: 36.2 KB
```

- There are total **660 entries** of Range Index and total **7 Data columns**
- All data types in the given dataset were **Integer** type
- Memory usage was 36.2KB
- There were **no missing values** in the dataset
- There were **11 duplicate entries** in the dataset, which was removed
- The columns **`Sl_No`** and **`Customer Key`** removed for further analysis

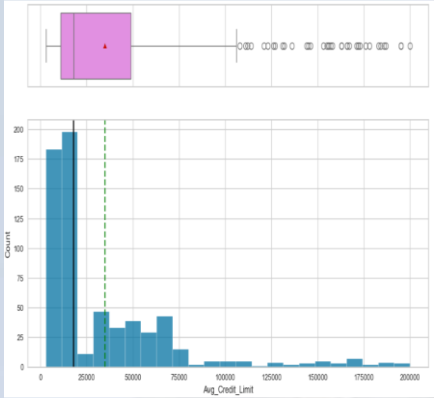
Statistical Summary

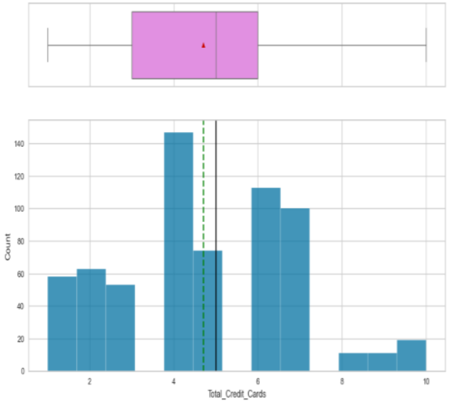
	count	mean	std	min	25%	50%	75%	max
Avg_Credit_Limit	649.0	34878.274268	37813.736638	3000.0	11000.0	18000.0	49000.0	200000.0
Total_Credit_Cards	649.0	4.708783	2.173763	1.0	3.0	5.0	6.0	10.0
Total_visits_bank	649.0	2.397535	1.625148	0.0	1.0	2.0	4.0	5.0
Total_visits_online	649.0	2.624037	2.952888	0.0	1.0	2.0	4.0	15.0
Total_calls_made	649.0	3.590139	2.877911	0.0	1.0	3.0	5.0	10.0

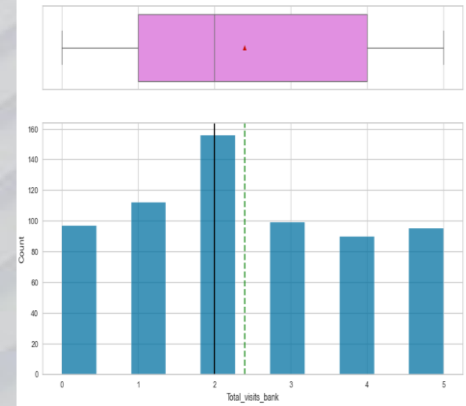
- All columns have **649** observations.
- There is a **wide variation** in Avg_Credit_Limit(3000.0 to 200000.0), indicating diverse spending capacities among customers.
- In Total_Credit_Cards most customers hold between 1 and 10 credit cards, with an average of around 5 cards, suggesting **moderate financial engagement.**
- Total_visits_bank insightes while some customers **do not visit** the bank at all, others visit up to 5 times per year.
- Many customers **prefer online services**, although there are also customers who do not make any online visits.
- On average, Total_calls_made by the customers about **3 times a year**, with some customers making up to 10 calls.

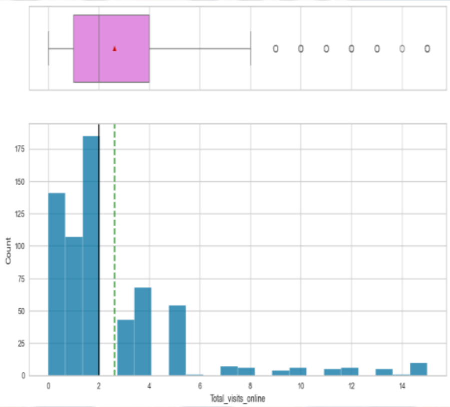
Exploratory Data Analysis(EDA)

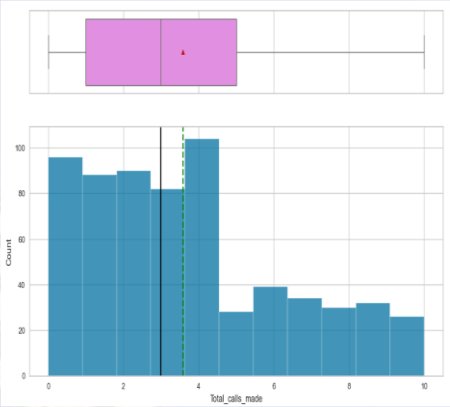
Univariate Analysis









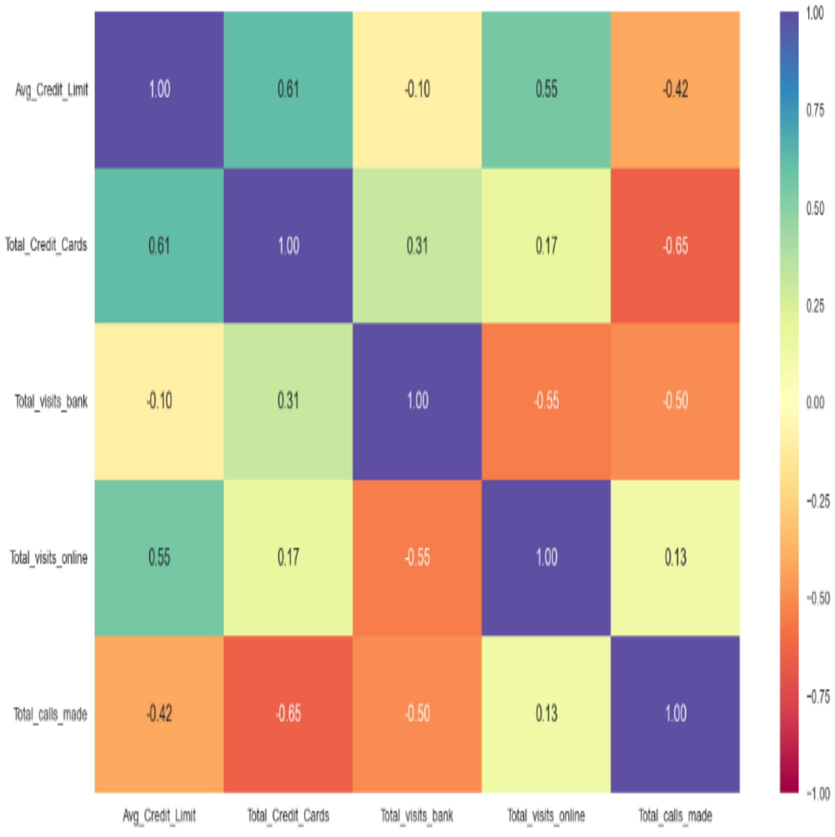


- Outliers are present in 'Avg_Credit_Limit' and 'Total_calls_made'. They are right skewed
- IQR is 3 to 6, most of the customers have credit cards more than 3 and the distribution is not even for 'Total_credit_cards'
- Slightly normal distributed, and no outliers are present for 'Total_visits_bank'
- When compared to visiting bank directly many of them opt to use online visits
- Some customers make zero calls, while others call 10 times a year.

Exploratory Data Analysis(EDA)

Bivariate analysis

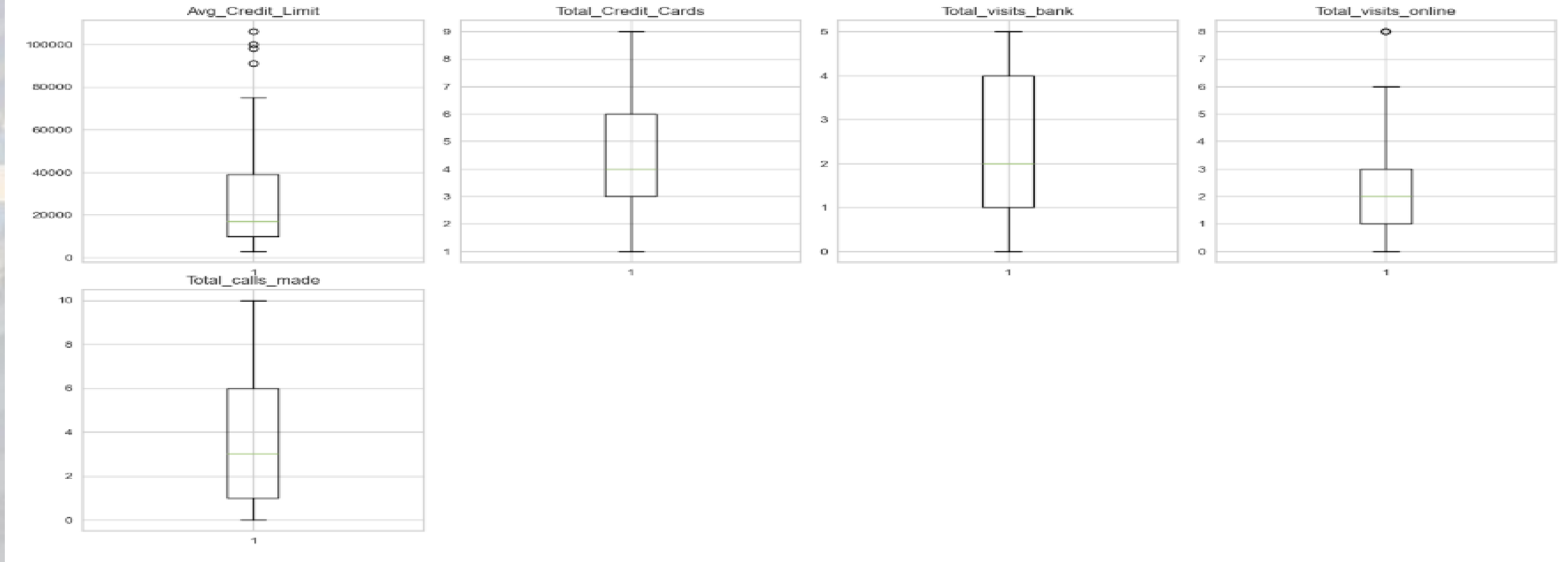
- ‘Total_Credit_Cards’ and ‘Avg_Credit_Limit’ **are highly correlated** with value **0.61**. Customers with more credit cards tend to have a higher credit limit.
- ‘Total_visits_online’ and ‘Avg_Credit_Limit’ are **correlaed** (0.55). Customers with a higher credit limit tend to use online banking services more frequently.
- ‘Total_Credit_Cards’ and ‘Total_calls_made’ are **highly negatively correlated**(-0.65). Customers with more credit cards tend to make fewer calls to the bank.
- ‘Total_visits_online’ and ‘Total_visits_bank’ are **negatively correlated**(-0.55). Customers who visit the bank more often tend to use online banking less, and vice versa.
- ‘Total_calls_made’ and ‘Total_visits_bank’ are **negatively correlated** (-0.50). Customers who visit the bank more frequently tend to make fewer calls.



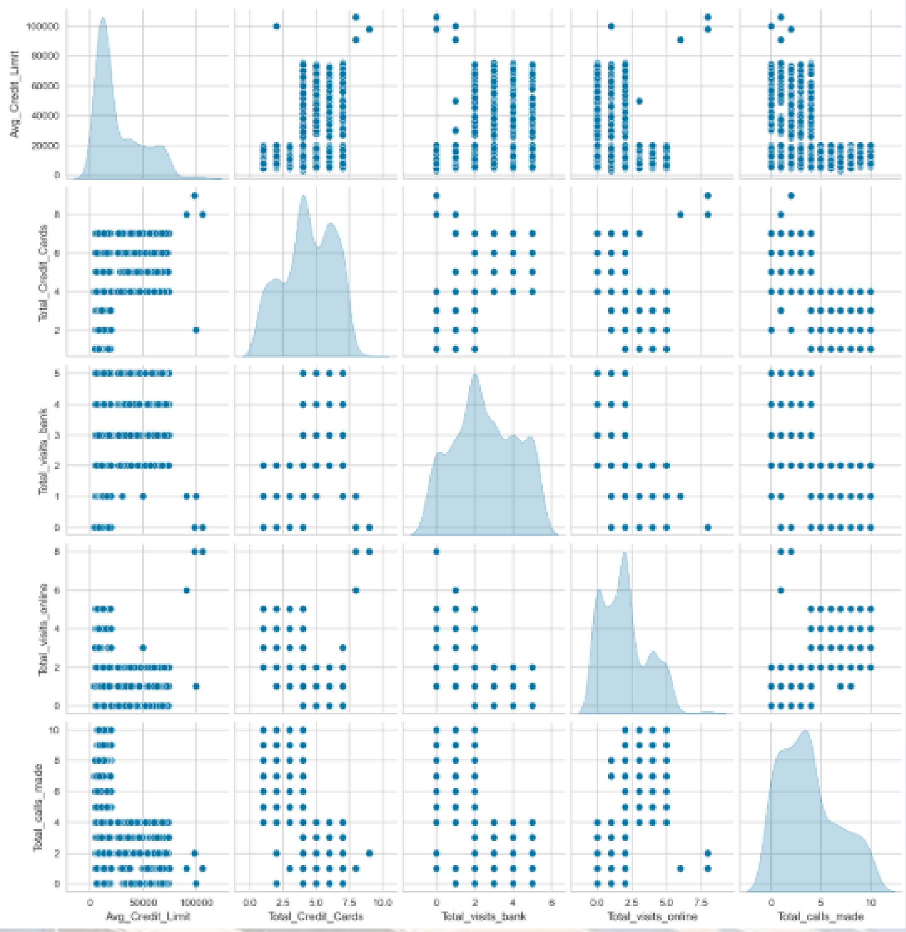
Data preprocessing

Outlier treatment, Pair plot analysis and Scaling

- The Avg_Credit_Limit and Total_visits_online have a lot of outliers, meaning a small subset of high-value customers behaves very differently.
- No major outliers in other categories suggest most customers behave consistently.
- **Treated the outliers using IQR method, Scaling also done.**



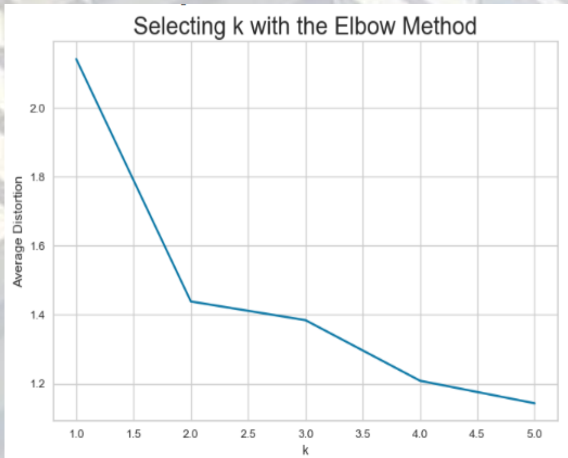
- **Avg_Credit_Limit:**
  - *Right-skewed distribution*
  - *Strong correlation with Total_Credit_Cards. Some outliers visible.*
- **Total_Credit_Cards:**
  - *Shows a bimodal-like distribution.*
  - *Positively correlated with Avg_Credit_Limit and negatively correlated with Total_calls_made*
- **Total_visits_bank vs. Total_visits_online:**
  - *Negative correlation. The distribution of bank visits is more spread out, while online visits are concentrated at lower values.*
- **Total_calls_made:**
  - *Shows negative correlation with both Total_Credit_Cards and Total_visits_bank.*

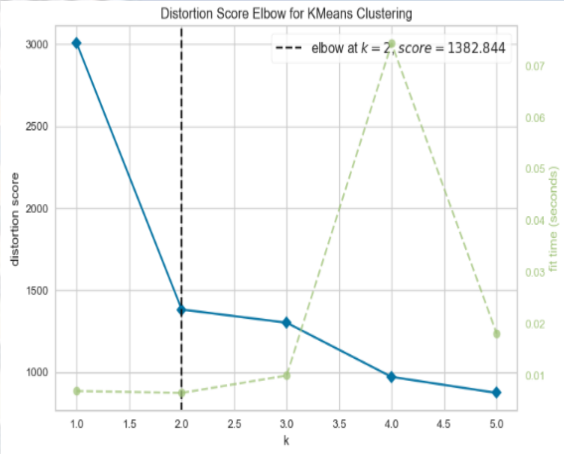


K-Means Clustering

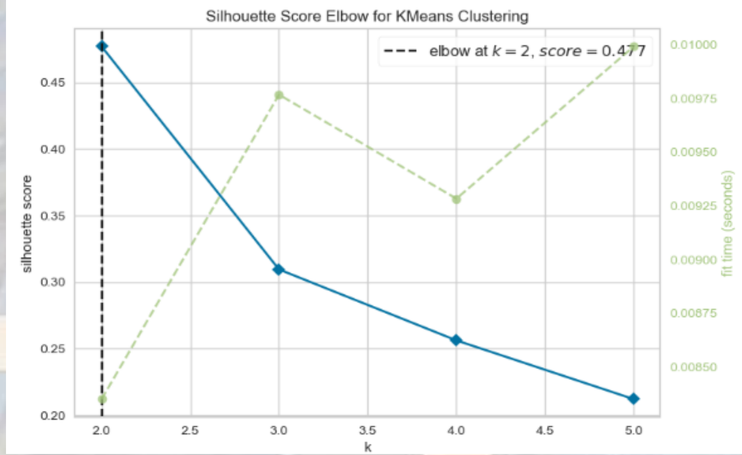
Elbow plot, Silhouette Score

```
Number of Clusters: 1    Average Distortion: 2.1408682583341108
Number of Clusters: 2    Average Distortion: 1.4383953103295255
Number of Clusters: 3    Average Distortion: 1.3841879770659893
Number of Clusters: 4    Average Distortion: 1.2086374412735184
Number of Clusters: 5    Average Distortion: 1.1433851381930007
```






```
For n_clusters = 2, the silhouette score is 0.4770537959184115)
For n_clusters = 3, the silhouette score is 0.3095458163470457)
For n_clusters = 4, the silhouette score is 0.2562215004495415)
For n_clusters = 5, the silhouette score is 0.21206308444900046)
```

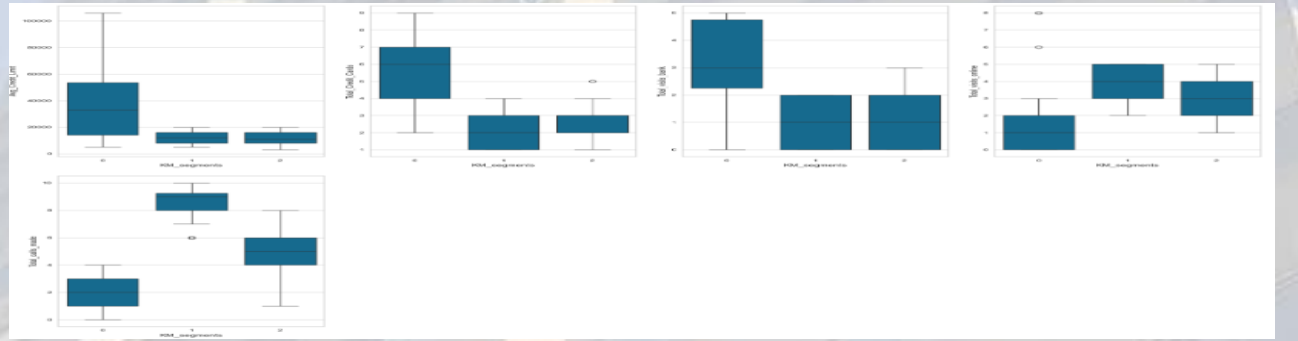


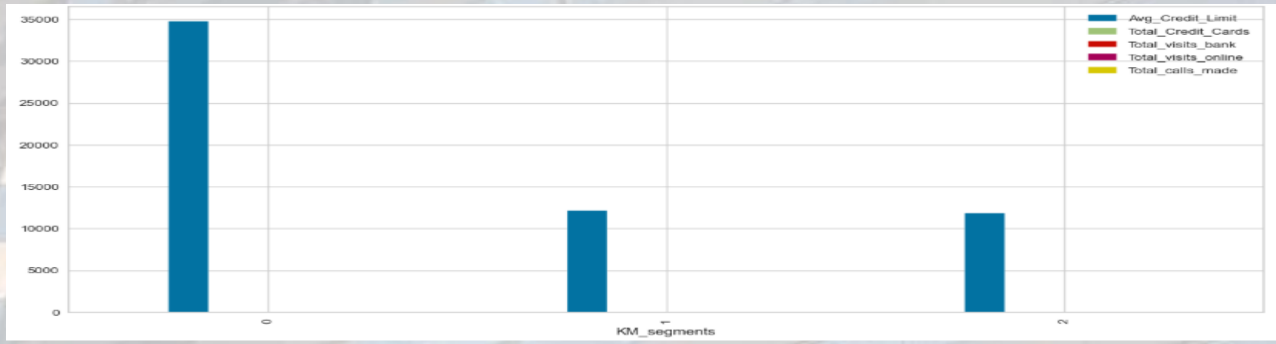
- Appropriate value for k seems to be 3.
- The Silhouette score selected for n-clusters = 3 (0.309)

Final Model: K-Means Clustering

Cluster profiling

KM_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	34796.296296	5.550265	3.468254	1.023810	1.976190	378
1	12134.615385	2.307692	0.923077	3.990385	8.500000	104
2	11882.352941	2.546218	1.008403	3.092437	5.378151	119





- **Cluster 0** is having higher count of **378**
- Cluster profiling visualization shows the average values of different features
- The scale of **Avg_Credit_Limit** is much **higher** than the other variables, making them **almost invisible.**
- This model is sensitive to Outliers and sensitive to initialize the cluster centers'

Hierarchical Clustering

Cophenetic correlation, Linkage methods, Dendograms

```
Cophenetic correlation for Euclidean distance and single linkage is 0.3043225749635933.
Cophenetic correlation for Euclidean distance and complete linkage is 0.8155054171914784.
Cophenetic correlation for Euclidean distance and average linkage is 0.8351177997762187.
Cophenetic correlation for Euclidean distance and weighted linkage is 0.82998699560955.
Cophenetic correlation for Chebyshev distance and single linkage is 0.2876861331056504.
Cophenetic correlation for Chebyshev distance and complete linkage is 0.7300454653181311.
Cophenetic correlation for Chebyshev distance and average linkage is 0.8241458734435407.
Cophenetic correlation for Chebyshev distance and weighted linkage is 0.8092174823416615.
Cophenetic correlation for Mahalanobis distance and single linkage is 0.2732684096233452.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.3481854596826671
Cophenetic correlation for Mahalanobis distance and average linkage is 0.6233920543095369.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.5058108938418997
Cophenetic correlation for Cityblock distance and single linkage is 0.2844099748964493.
Cophenetic correlation for Cityblock distance and complete linkage is 0.8030988771989755.
Cophenetic correlation for Cityblock distance and average linkage is 0.8302526149922482.
Cophenetic correlation for Cityblock distance and weighted linkage is 0.8232951663286479.
```

```
Cophenetic correlation for single linkage is 0.3043225749635933.
Cophenetic correlation for complete linkage is 0.8155054171914784.
Cophenetic correlation for average linkage is 0.8351177997762187.
Cophenetic correlation for centroid linkage is 0.8261994060158777.
Cophenetic correlation for ward linkage is 0.7832495616128994.
Cophenetic correlation for weighted linkage is 0.82998699560955.
```

	Linkage	Cophenetic Coefficient
0	single	0.304323
4	ward	0.783250
1	complete	0.815505
3	centroid	0.826199
5	weighted	0.829987
2	average	0.835118

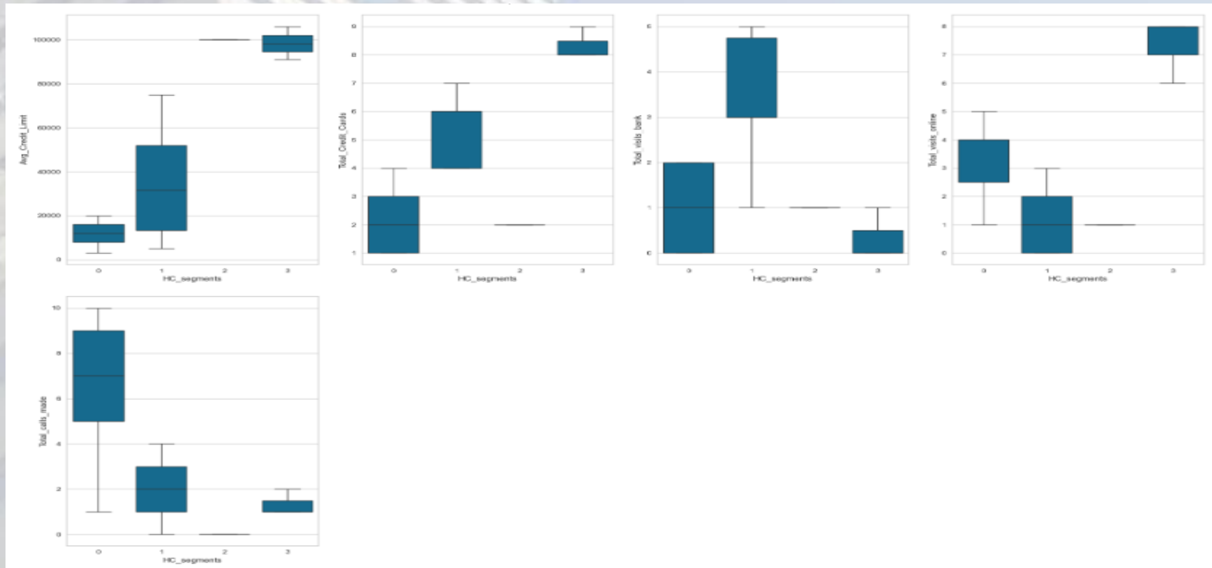
- Highest cophenetic correlation is **0.835**, which is obtained with **Euclidean distance** and **average linkage.**
- After exploring different linkage methods – Observed that cophenetic correlation is maximum with Euclidean distance and average linkage.
- The **Average Linkage** (0.835) is the best choice because it has the highest cophenetic coefficient.

- The cophenetic correlation is highest for average linkage method(0.84).
- We will move ahead with average linkage.
- 4 appears to be the appropriate number of clusters from the dendrogram for average linkage.

Final Model: Hierarchical Clustering

Cluster profiling

HC_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
0	12091.324201	2.401826	0.945205	3.538813	6.894977	219
1	33825.396825	5.523810	3.486772	0.984127	2.002646	378
2	100000.000000	2.000000	1.000000	1.000000	0.000000	1
3	98333.333333	8.333333	0.333333	7.333333	1.333333	3



- We will look into clusters 0 and 1 only
- Cluster 0
  - *There are 219 users in this cluster*
  - *They have made highest number of calls to bank and low bank visits*
  - *They have very low average credit limit ~ 12,000 rupees*
  - *There online visit is moderate*
- Cluster 1
  - *They are active users with Moderate Credit Limit ~33,000 rupees*
  - *They are the largest group with 378 users*
  - *They have the highest bank visit, lowest online visit and few calls made to the bank*
  - *The customers in this cluster have highest number of cards (4 to 6)*

Actionable Insights and Recommendations

Key Observations	K-Means Clustering	Hierarchical Clustering
Execution time	K-Means was significantly faster, making it better for large datasets.	Hierarchical clustering took longer due to pairwise distance calculations.
Number of Clusters	3 clusters	4 clusters
Cluster Similarity	The first two clusters (0 & 1) were identical in both methods. K-Means merged high-credit-limit customers into a single cluster, whereas Hierarchical clustering further separated them.	
Appropriate Number of Clusters	K-Means used the Elbow Method (k=3)	Hierarchical clustering used Cophenetic Coefficient & Dendrogram (k=4)

Actionable Insights

- **Customer Segments Identified**
  - *Frequent bank visits/calls but low credit limits.*
  - *Moderate credit limits with minimal customer service interactions.*
  - *High credit limits, fewer bank visits, and online preference.*
  - *Exclusive segment with minimal need for support.*

- **Spending & Credit Card Usage Trends**
  - *Higher credit limits correlate with more credit cards. Customers with multiple credit cards make fewer customer service calls.*

- **Customer Service Trends**
  - *More online interactions than in-person visits, showing a shift to digital banking. High call volume among low-spending customers, indicating dissatisfaction or confusion.*

Recommendations

- Offer cashback or rewards to encourage more card usage.
- Provide personalized credit limit enhancements and loyalty rewards for high spending customers
- Improve Customer Support Services by enhance online self-service tools, implement AI chatbots for faster issue resolution and prioritize elite customers for premium support.
- Promote mobile banking and online services to reduce branch visits.
- Optimize Credit Card Offerings by introducing tier-based credit cards with tailored benefits. Analyze spending patterns to suggest relevant financial products.