UNSUPERVISED LEARNING

K-MEANS

&

HIERARCHICAL CLUSTERING

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. PROBLEM DEFINITION

1.1. Context:

Credit cards have evolved to meet the changing financial habits and technological advancements of society.Credit cards are increasingly integrated with digital wallets allowing users to make contactless payments with their smartphones or smartwatches.For younger generations or those with limited credit history, some cards are designed to help users build credit responsibly, with features like lower credit limits and educational resources.Credit cards have become more globally accessible, with lower foreign transaction fees and wider acceptance internationally, catering to the increasing trend of global travel and e-commerce.Using a credit card responsibly can help individuals build a positive credit history, which is crucial for securing loans or mortgages in the future.

Credit card providers operate in a highly competitive environment, and they face several challenges as they strive to differentiate themselves and capture market share.Competitors who leverage data analytics to provide personalized services, such as tailored credit limits or customized rewards, can gain a competitive edge. Providers need to enhance their data analytics capabilities to offer similar or better services.High-quality customer service, including 24/7 support and fast issue resolution, is a key differentiator. Providers with less responsive or efficient customer service risk losing customers to those who excel in this area.

1.2. Objective:

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.

AllLife Bank wants 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.

2.DATA BACKGROUND AND CONTENTS

2.1.Purpose:

The purpose of collecting the data was to improve market penetration for AllLife Bank credit card.

2.2.Data Description and Dictionary:

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

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).

2.3.Data Types:

```
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
```

Table 1: Data Types

- All the variables from the dataset are numerical .
- All columns have 660 non-null values.

2.4.Data Summary:

	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
count	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000	660.000000
mean	330.500000	55141.443939	34574.242424	4.706061	2.403030	2.606061	3.583333
std	190.669872	25627.772200	37625.487804	2.167835	1.631813	2.935724	2.865317
min	1.000000	11265.000000	3000.000000	1.000000	0.000000	0.000000	0.000000
25%	165.750000	33825.250000	10000.000000	3.000000	1.000000	1.000000	1.000000
50%	330.500000	53874.500000	18000.000000	5.000000	2.000000	2.000000	3.000000
75%	495.250000	77202.500000	48000.000000	6.000000	4.000000	4.000000	5.000000
max	660.000000	99843.000000	200000.000000	10.000000	5.000000	15.000000	10.000000

Table 2: Data Summary

- 50% of the customers individually own 5 credit cards .
- The number of customer queries raised online and through calls is higher compared to those raised through bank visits.

3.FEATURE ENGINEERING

- The features **Total_visits_bank, Total_visits_online, and Total_calls_made** are related to the same subject, which is analyzing customer queries regarding their transactions and rewards. Therefore, combining these three features into a single new feature will retain the same information and could make it easier for the algorithm to group and analyze the data.
- **Total_visits_bank, Total_visits_online, and Total_calls_made** have been combined into a new feature named **Total_Interactions**. Therefore, the original columns have been deleted since **Total_Interactions** contains the same information.
- The features **SI_No** and **Customer Key** have been dropped since they serve as unique identifiers and do not add value for segmenting customers.

4.UNIVARIATE ANALYSIS

For performing Univariate analysis we will take a look at the Boxplots and Histograms to get a better understanding of the distributions.

4.1.Observation on Total_Interactions

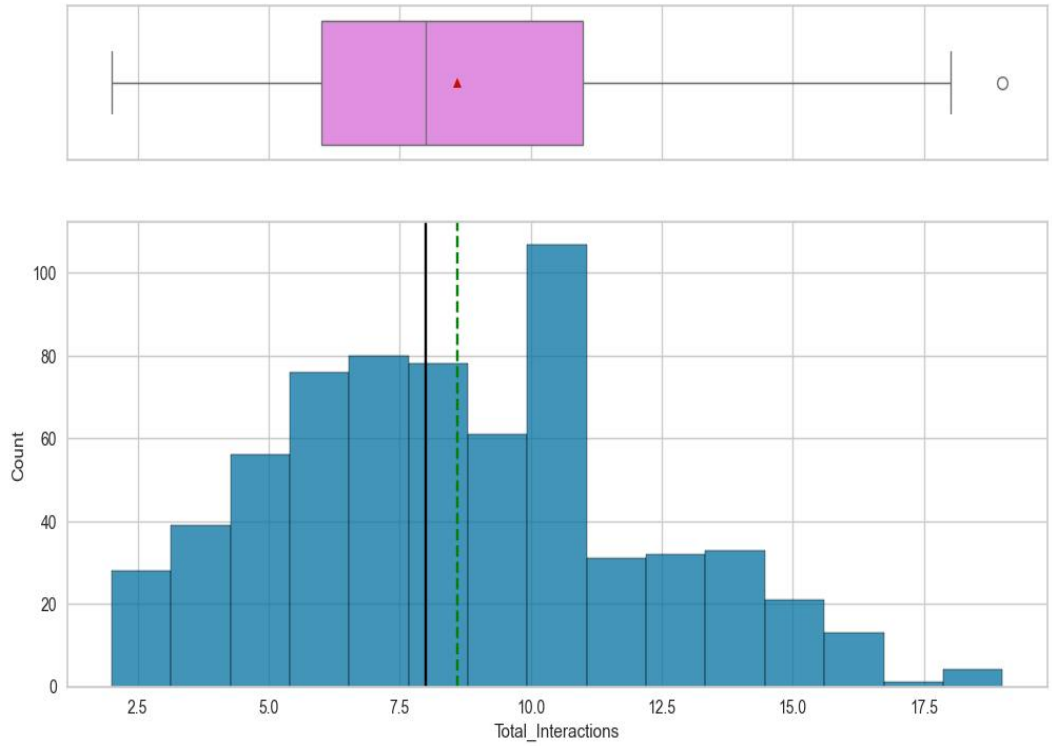

Fig 1: Total_Interactions

Observation:

- The distribution of the variable Total_Interactions is slightly right skewed.
- With an average Total_Interactions being around 8.

4.2.Observation on Avg_Credit_Limit

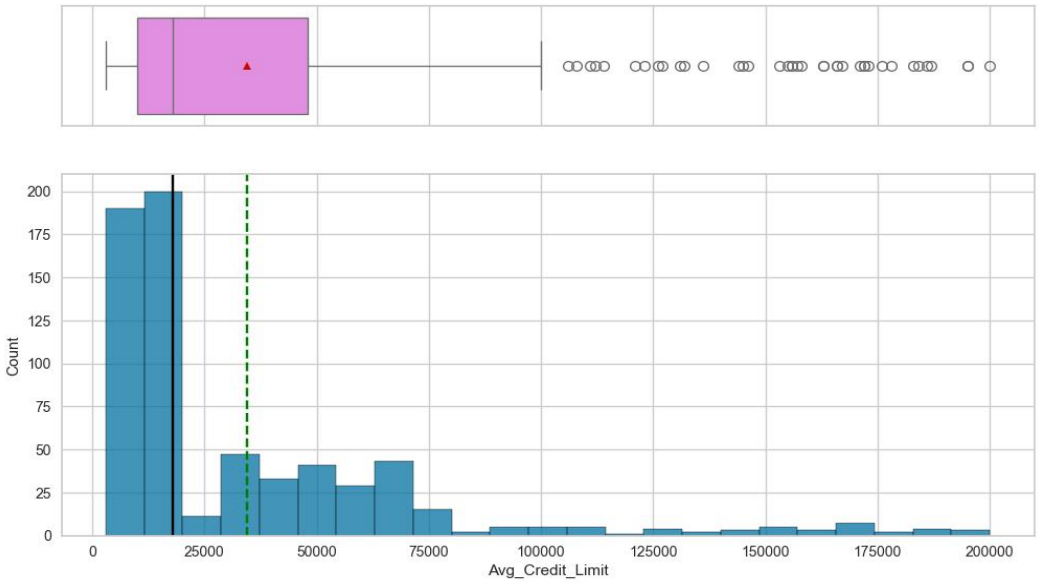

Fig 2:Avg_Credit_Limit

Observation:

- The distribution of the variable Avg_Credit_Limit is highly right skewed.
- With an average Credit_Limit being around $34574.

4.3.Observation on Total_Credit_Cards

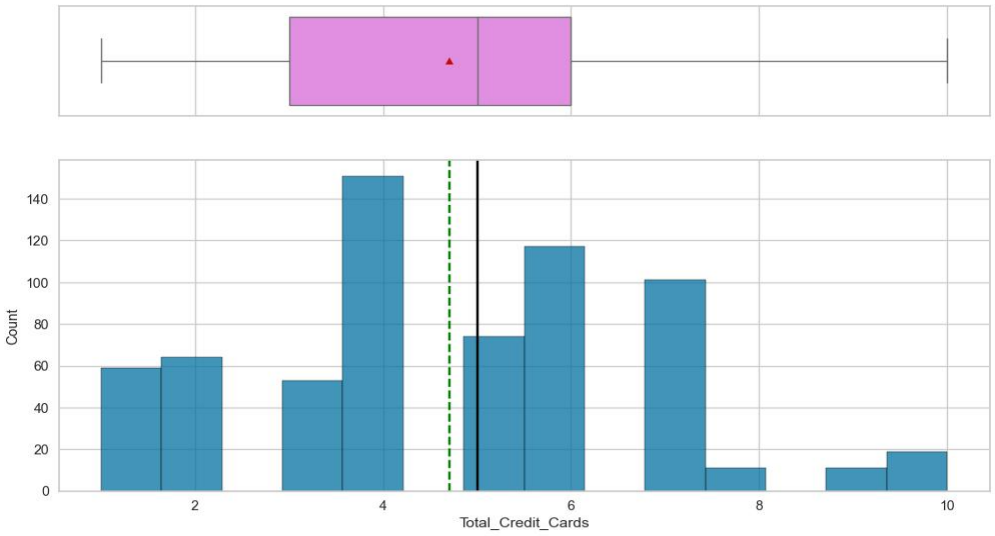

Fig 3:Total_Credit_Cards

Observation:

- The variable Total_Credit_Cards forms a multimodal distribution.
- With an average Total_Credit_Cards being around 5.

5. MULTIVARIATE ANALYSIS

5.1.Correlation of Numerical Variables

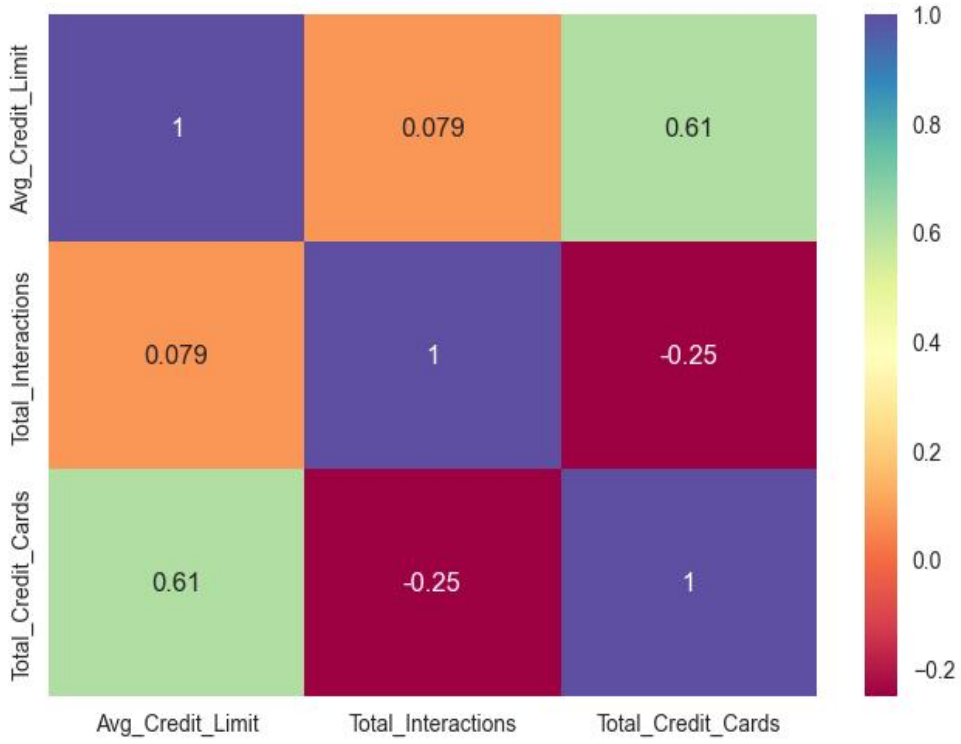

Fig 4:Correlation of numerical variables

Observation:

- There is a moderate correlation between the variable Avg_Credit_Limit and Total_Credit_Cards,it is understanble that more credit cards results in more average credit limits.

5.2.Relationship between numercial variables:

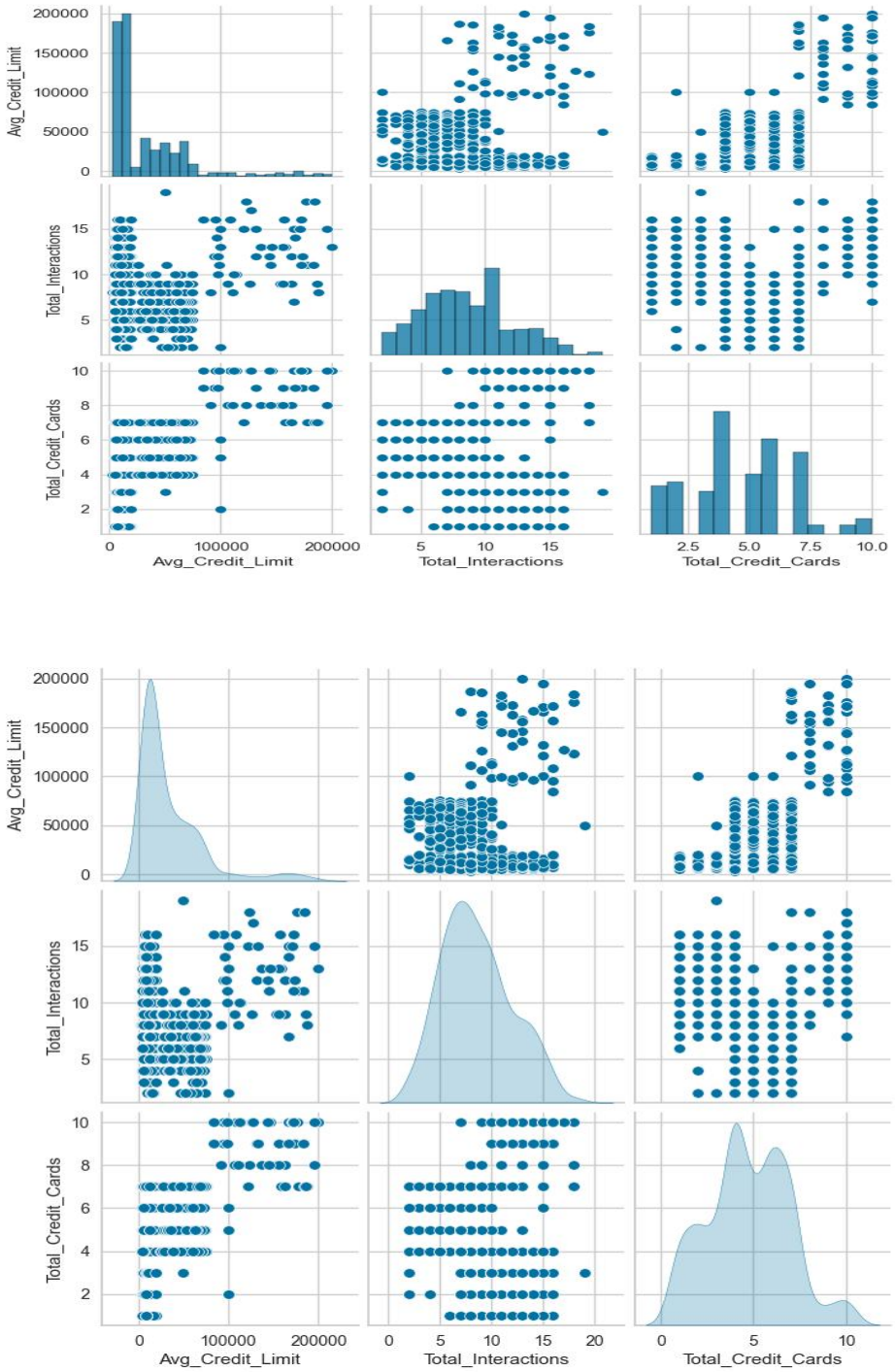

Fig 5:Relationship between numerical variables

Observation:

- Without relationships between variables, it might indeed be challenging for algorithms to form meaningful clusters because clusters typically rely on some form of similarity or pattern in the data.

5.3.Avg_Credit_Limit vs Total_Interactions

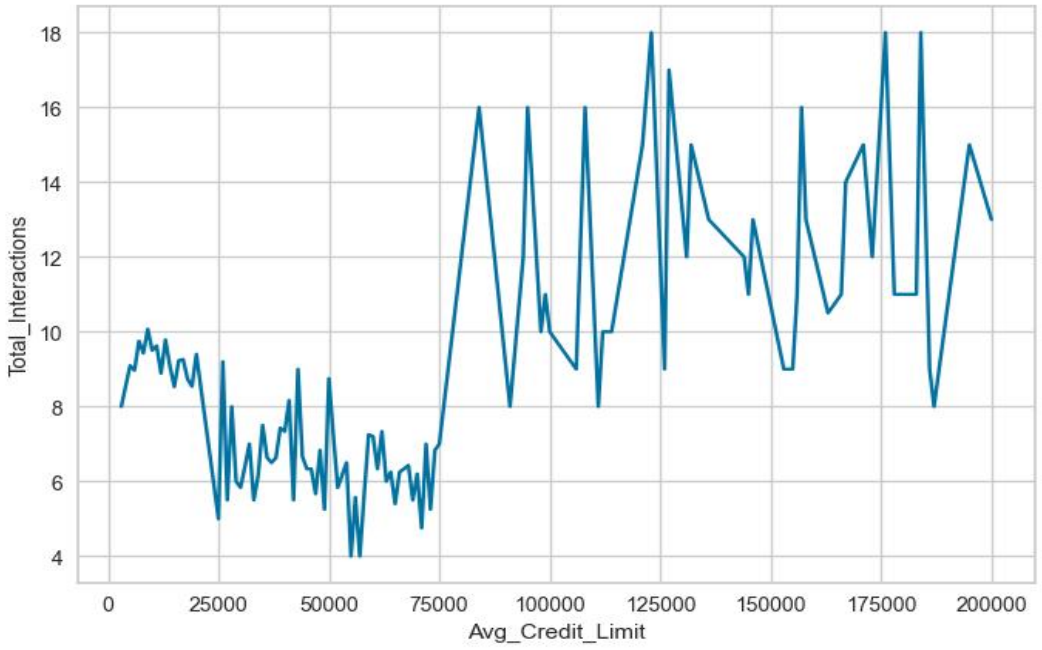

Fig 6: Avg_Credit_Limit vs Total_Interactions

Observation:

- The total interactions by customers with average credit limit above $75000 is higher compared to customer with below $75000.

5.4.Total_Credit_Cards vs Total_Interactions

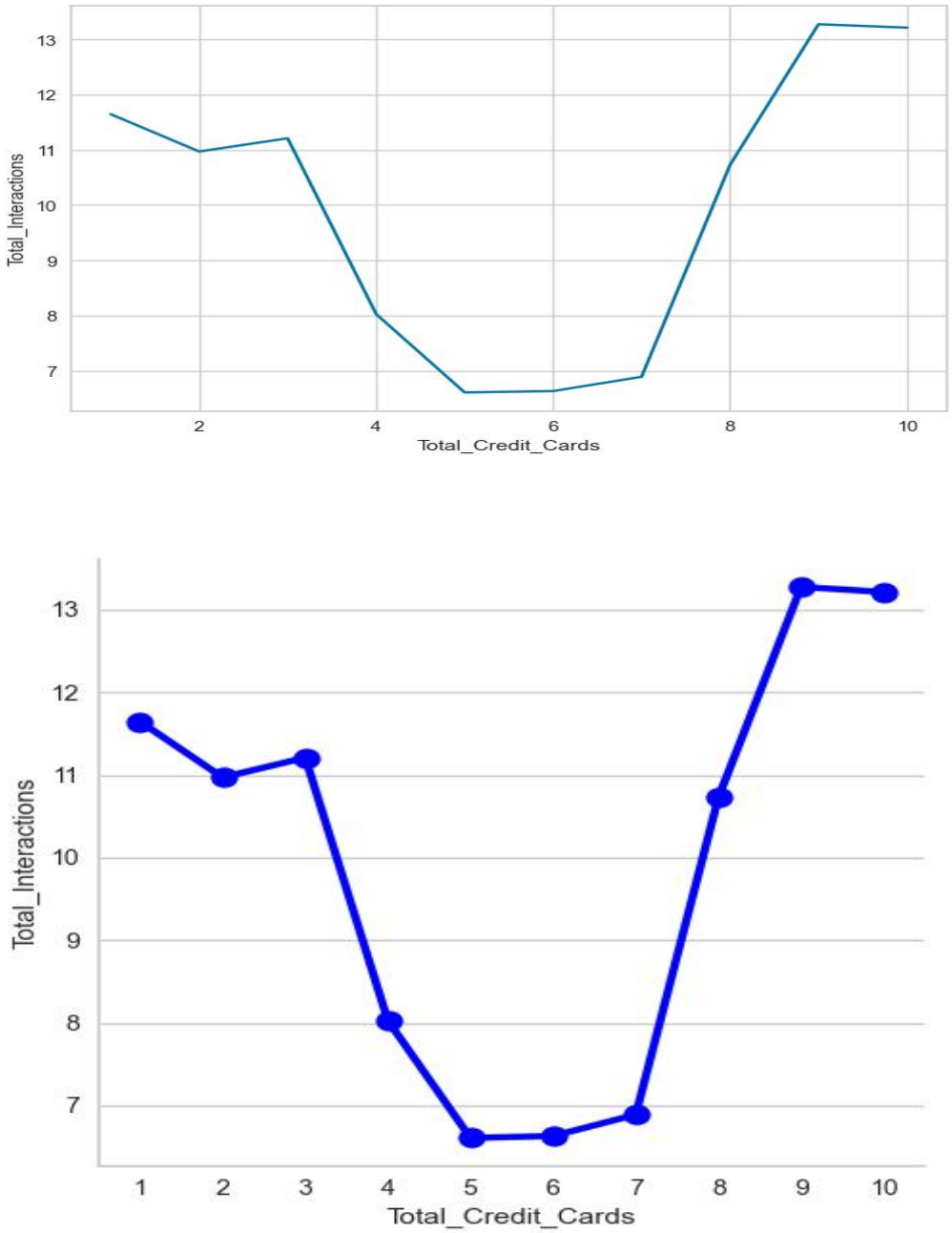

Fig 7: Total_Credit_Cards vs Total_Interactions

Observation:

- The total number of interactions by customers who have between 1 to 3 credit cards and those who have between 8 to 10 credit cards is higher compared to the interactions from customers who have between 4 to 7 credit cards.

6. INSIGHTS BASED ON UNIVARIATE AND MULTIVARAITE

- As the average credit limit increases, customer interactions with the bank also increase, suggesting that individuals with higher credit limits tend to make more transactions.

- Customers with 1 or 2 credit cards have approached the bank frequently, indicating that they are making more transactions. Since they have fewer alternative credit cards, they rely on their AllLife Bank credit card for most of their transactions.

- Even though customers with 8, 9, or 10 credit cards have other alternatives, their interactions with AllLife Bank remain high. This suggests that customers prefer using AllLife Bank's credit card, indicating that the bank is providing excellent service in the credit card business.

- Since most customers have multiple credit cards, encouraging them to use AllLife Bank credit card can be challenging. AllLife should focus on addressing customers recurring queries effectively to ensure these issues are not raised again. Providing exceptional customer service is crucial, as poor experiences can drive customers away.

7.OUTLIER DETECTION AND HANDLING

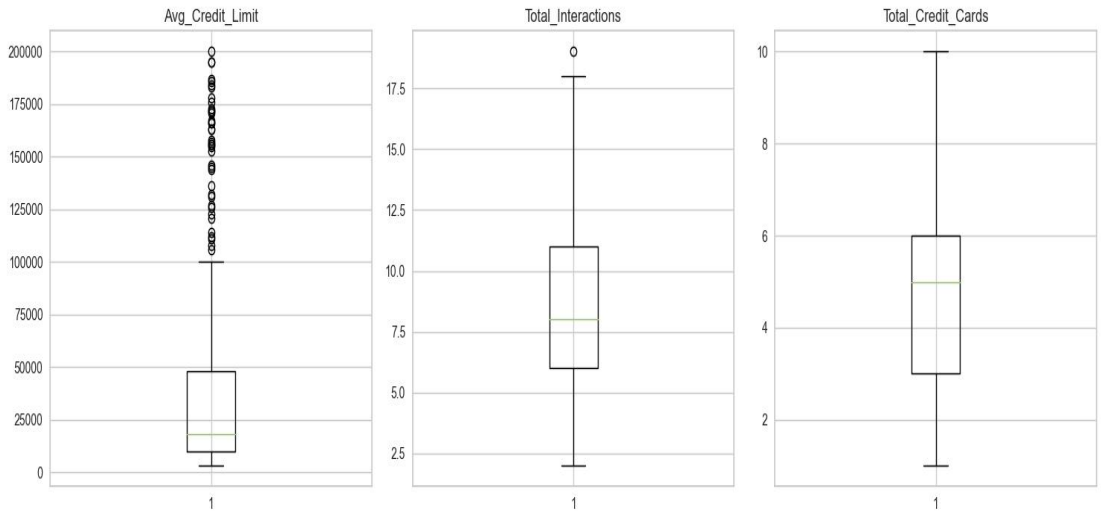

Fig 8: Outlier Detection

- There are outliers found in the data.
- However,the outliers are not treated as they are proper values.

8. MISSING VALUE CHECK

```
Avg_Credit_Limit      0
Total_Credit_Cards    0
Total_Interactions    0
dtype: int64
```

- The dataset has no missing values to be handled.

9. DATA SCALING

- Scaling the data is essential when using algorithms like K-means and Hierarchical clustering in unsupervised learning to segment customers based on their interaction and spending patterns. Since these algorithms rely on distance metrics, features with larger ranges can dominate the distance calculations, leading to biased groupings. Therefore, scaling ensures that all features contribute equally to the segmentation process.Scaled data are stored in the attribute **subset_scaled_df.**

10. K-MEANS CLUSTERING

- When applying K-means clustering to a dataset, choosing the appropriate number of clusters, denoted as K, is a crucial step. The selection of K determines how the data will be grouped, which directly impacts the effectiveness of the clustering. An inappropriate choice of K can lead to poor clustering results, where either too many or too few clusters fail to accurately represent the underlying data structure.
- To help determine the optimal value of K, we often use the Elbow method. This technique involves running the K-means algorithm with different values of KK and plotting the average distortion against K. Average distortion represents the average distance of points to their assigned cluster centroids.
- As K increases, the average distortion typically decreases because adding more clusters reduces the distance of points to their nearest cluster centroid. However, after a certain point, the rate of decrease sharply diminishes. The plot of average distortion versus K will start to form an 'elbow,' and the value of K at this elbow point is considered a good choice for the number of clusters.
- We have used **Euclidean** distance metric for calculating average distortion calculation.

Elbow Method for selecting K

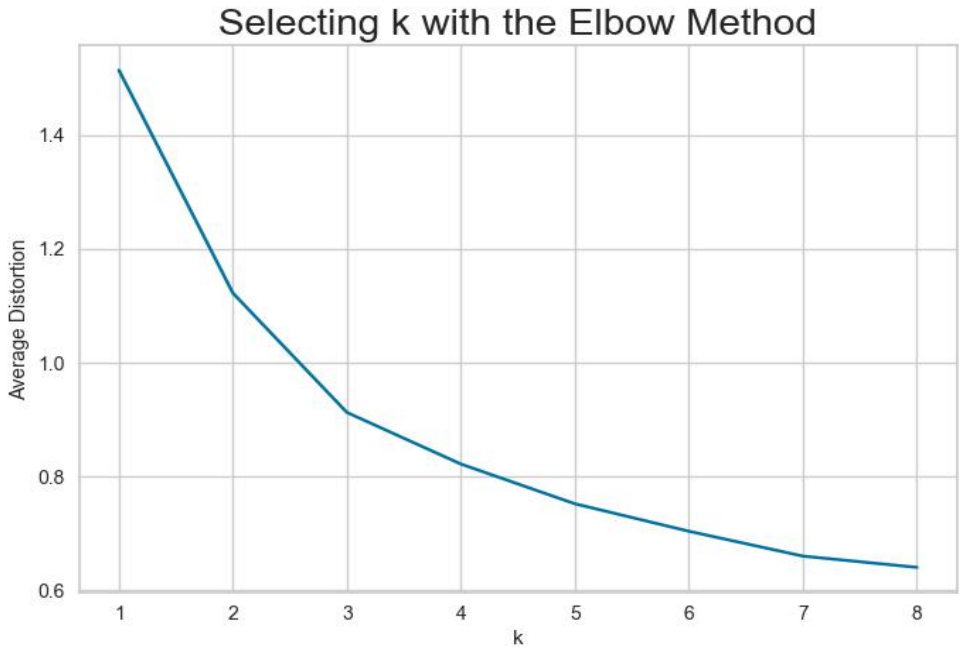

Fig 9: Elbow Method

```
Number of Clusters: 1   Average Distortion: 1.5141640026755379
Number of Clusters: 2   Average Distortion: 1.1219512779067473
Number of Clusters: 3   Average Distortion: 0.9120122427769022
Number of Clusters: 4   Average Distortion: 0.8211642302941319
Number of Clusters: 5   Average Distortion: 0.7512629484890923
Number of Clusters: 6   Average Distortion: 0.7029395305049557
Number of Clusters: 7   Average Distortion: 0.6591807942499348
Number of Clusters: 8   Average Distortion: 0.6394166007536708
```

- The range of K is valued form (1 to 8), from the scores of average distortion K at 3 has started to form an 'elbow,' and the value of K at this elbow point is considered a good choice for the number of clusters.
- To assess the quality of clusters at different values of K, we calculate the **Silhouette Scores.** This provides a better understanding of which value of K results in better-segmented clusters, as higher Silhouette Scores indicate that clusters are more cohesive and well-separated from each other.

Silhouette Scores

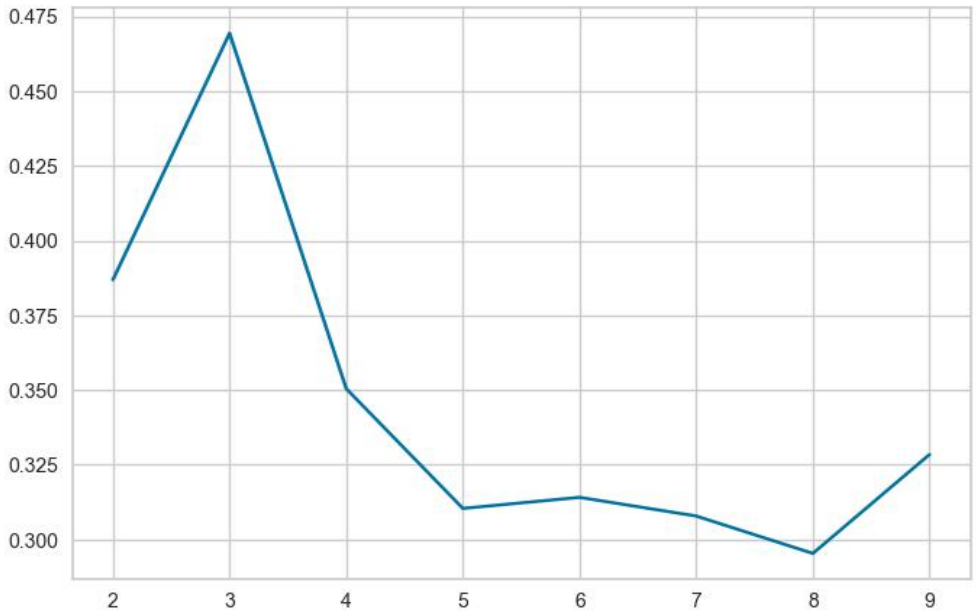

Fig 10: Silhouette Scores

```
For n_clusters = 2, silhouette score is 0.3869348345296014
For n_clusters = 3, silhouette score is 0.4694219335271744
For n_clusters = 4, silhouette score is 0.3504882491661683
For n_clusters = 5, silhouette score is 0.31042612544496284
For n_clusters = 6, silhouette score is 0.31412677140479983
For n_clusters = 7, silhouette score is 0.3078785402593218
For n_clusters = 8, silhouette score is 0.2953675246922671
For n_clusters = 9, silhouette score is 0.3284765224402659
```

- Since the silhouette score for n_cluster =3 stands high with 0.4694 which shows fair level of cohesion and separation.
- Let us perform K-means at n_cluster=3.

Cluster Profiling

We perform cluster profiling which involves summarizing the characteristics of each cluster generated by a clustering algorithm (K-means) to understand the groups formed within the data.

- A new dataframe **cluster_profile** is cereated,which conatins the average values of the features for each cluster.

- The count of occurances of data points in each segement based on the **Total_Credit_Cards** column is created.
- The result is then assigned to a new column in cluster_profile called "count_in_each_segment", which shows how many data points belong to each cluster.

Result of cluster profiling:

K_means_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_Interactions	count_in_each_segment
0	141040.000000	8.740000	12.580000	50
1	33992.105263	5.523684	6.347368	380
2	12391.304348	2.478261	11.434783	230

Table 3: Cluster profiling - K-means

Cluster 0:

Average Credit Limit: $141,040.00

This cluster has the highest average credit limit among the three clusters, suggesting that the customers in this segment are likely to have higher financial standing or creditworthiness.

Total Credit Cards: 8.74

On average, customers in this cluster hold about 7 to 10 credit cards, which is the highest among the three clusters. This indicates that these customers have multiple credit card options, likely due to their higher credit limits and spending power.

Total Interactions: 12.58

Customers in this cluster have an average of 12.58 interactions with the bank. This shows a relatively high level of engagement with the bank, possibly due to their multiple accounts and higher credit activity.

Count in Each Segment: 50

There are 50 customers in this cluster, making it the smallest of the three clusters, but it includes the most affluent customers based on the credit limit.

Cluster 1:

Average Credit Limit: $33,992.11

This cluster has a moderate average credit limit, significantly lower than Cluster 0 but higher than Cluster 2. These customers likely have moderate spending power and creditworthiness.

Total Credit Cards: 5.52

Customers in this cluster hold around 4 to 6 credit cards on average. This is fewer than Cluster 0 but more than Cluster 2, suggesting moderate access to credit.

Total Interactions: 6.35

With 6.35 average interactions, this cluster shows moderate engagement with the bank. These customers are somewhat active, possibly due to their need to manage multiple credit cards.

Count in Each Segment: 380

This is the largest cluster, with 380 customers. This suggests that a substantial proportion of the customer base falls into this middle range of credit limits and engagement levels.

Cluster 2:

Average Credit Limit: $12,391.30

This cluster has the lowest average credit limit, indicating that these customers have the least credit availability among the three clusters.

Total Credit Cards: 2.48

Customers in this cluster have, on average, about 1 to 3 credit cards, which is the lowest among the clusters. These customers have limited credit options compared to those in the other segments.

Total Interactions: 11.43

Interestingly, despite having lower credit limits and fewer cards, this cluster has a relatively high number of interactions (11.43), close to that of Cluster 0. This might suggest that these customers rely more on their existing accounts for their banking needs.

Count in Each Segment: 230

There are 230 customers in this cluster, making it the second-largest segment. These customers have lower financial profiles but still show considerable interaction with the bank.

Visualization of the silhouette scores of the clusters generated by the K-means algorithm

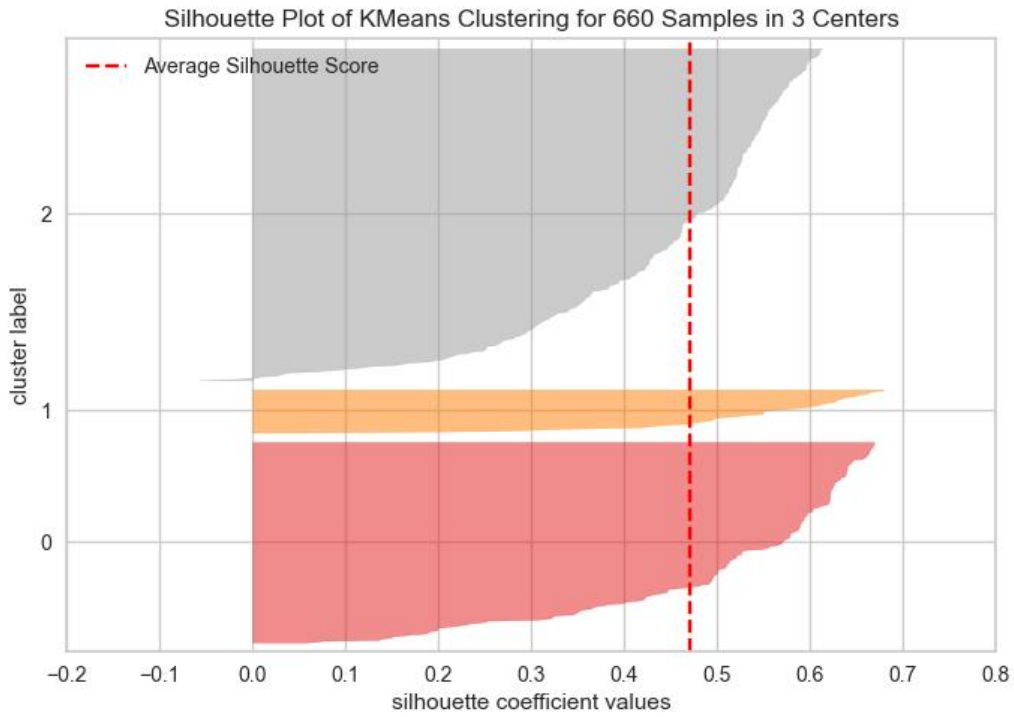

Fig 11: Silhouette Plot of KMeans Clustering

- From the plot, we can see that each cluster has a silhouette score above the average silhouette score. Additionally, each cluster has a silhouette score above 0.6, which indicates a good level of cohesion within clusters and separation between clusters.

Boxplot for each cluster

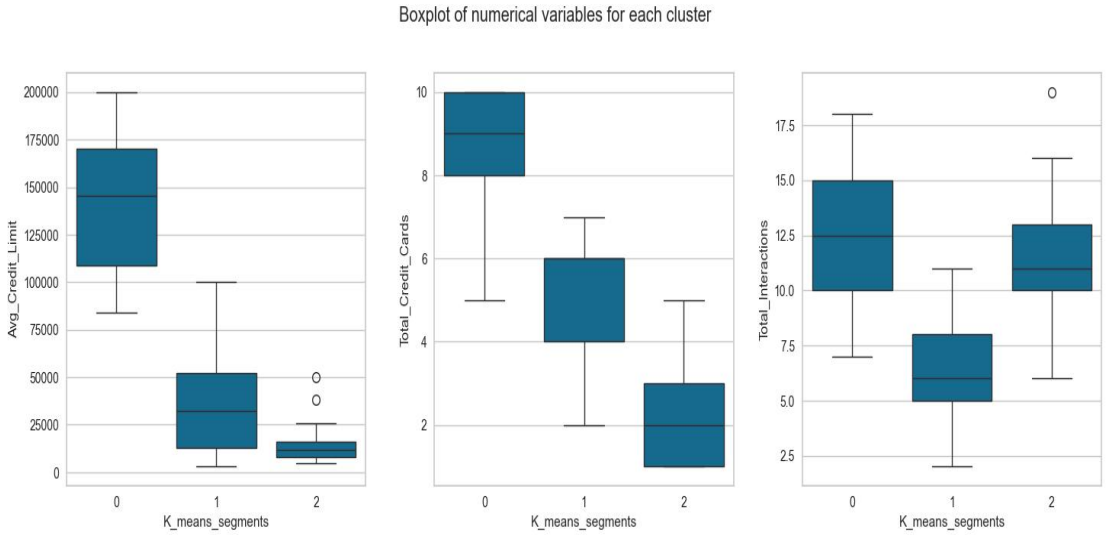

Fig 12: Boxplot for K-means clusters

- The boxplot for the variable **Total_Credit_Limit** shows good separation between clusters, indicating that the clusters are well-differentiated based on this variable.

11.HIERARCHICAL CLUSTERING

- Intially,let us perform cophenetic correlation for different distance metrics like **euclidean,chebyshev,mahalanobis and cityblock** with linkage methods like **single, complete, average, weighted.**
- This will be give us an idea of which distance matric to move forward.

```
Cophenetic correlation for Euclidean distance and single linkage is 0.6654277276946772.
Cophenetic correlation for Euclidean distance and complete linkage is 0.7671286292885412.
Cophenetic correlation for Euclidean distance and average linkage is 0.8375407403410419.
Cophenetic correlation for Euclidean distance and weighted linkage is 0.731656222128375.
Cophenetic correlation for Chebyshev distance and single linkage is 0.5890411537211859.
Cophenetic correlation for Chebyshev distance and complete linkage is 0.6805980414830478.
Cophenetic correlation for Chebyshev distance and average linkage is 0.8404102563578326.
Cophenetic correlation for Chebyshev distance and weighted linkage is 0.6699668638700003.
Cophenetic correlation for Mahalanobis distance and single linkage is 0.67293899219701.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.6271555995941986.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.8063869497392993.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.6827308903893113.
Cophenetic correlation for Cityblock distance and single linkage is 0.6707416771120457.
Cophenetic correlation for Cityblock distance and complete linkage is 0.7767809973910594.
Cophenetic correlation for Cityblock distance and average linkage is 0.8356790646291422.
Cophenetic correlation for Cityblock distance and weighted linkage is 0.7718157347761637.
```

- Highest cophenetic correlation is 0.8404102563578326, which is obtained with Chebyshev distance and average linkage.

- We will check the cophenetic correlation for Chebyshev distance with linkage methods like single, complete, average, weighted.

```
Cophenetic correlation for single linkage is 0.5890411537211859.
Cophenetic correlation for complete linkage is 0.6805980414830478.
Cophenetic correlation for average linkage is 0.8404102563578326.
Cophenetic correlation for weighted linkage is 0.6699668638700003.
```

- Highest cophenetic correlation is 0.8404102563578326, which is obtained with average linkage.

- Let us perform dendrogram for cophenetic correlation

Dendrogram for cophenetic correlation for Chebyshev distance

Fig 13: Dendrogram-Chebyshev distance

- One of the main reasons to check the dendrogram is to determine the optimal number of clusters.
- By cutting the dendrogram at a specific level (horizontal line), we can see how many clusters we would have. The height at which we cut corresponds to the distance or dissimilarity at which clusters are merged.
- From the dendrogram we have picked up optimal number of cluster as 2 and perform agglomerative methiod under Hierarchical clustering.
- Reason for picking agglomerative method ,as method starts by treating each data point as its own cluster **(bottom up approach)** and then successively merges the closest pairs of clusters, making it intuitive to understand.

Cluster Profiling

We perform cluster profiling, which involves summarizing the characteristics of each cluster generated by hierarchical agglomerative clustering using the model stored in the **HCmodel** attribute with preferred linkage **average.**

```
In cluster 0, there are 610 customers.
The average number of interactions in cluster 0 is 8.27
The average number of cards in cluster 0 is 4.38
The total number of card held by the customers in cluster 0 are: [2 3 7 5 4 1 6]

In cluster 1, there are 50 customers.
The average number of interactions in cluster 1 is 12.58
The average number of cards in cluster 1 is 8.74
The total number of card held by the customers in cluster 1 are: [ 6  5  9  8 10  7]
```

Table 4: Cluster profiling - Chebyshev average linkage

Cluster 0

Number of Customers:

There are 610 customers in this cluster. This indicates that this is a relatively large segment of customers.

Interactions :

The average number of interactions for customers in this cluster is 8.27. This metric suggests that, on average, each customer in this cluster interacts with the bank (or relevant service) approximately 8 times, which might include transactions, inquiries or other forms of engagement.

Average Number of Cards:

The average number of cards held by customers in this cluster is 4.38. This indicates that customers in this cluster, on average, possess a little over 4 credit cards, suggesting a moderate level of credit card ownership among them.

Total Number of Cards Held:

The total number of cards held by customers in this cluster is represented by the array [2, 3, 7, 5, 4, 1, 6]. This array lists the different quantities of credit cards that customers in this cluster hold.The values show the variety of card ownership, ranging from 1 to 7 cards, indicating some diversity in card ownership within this group.

Cluster 1

Number of Customers:

There are 50 customers in this cluster. This is a smaller segment compared to Cluster 0.

Average Number of Interactions:

The average number of interactions for customers in this cluster is 12.58. This suggests that customers in Cluster 1 are more engaged, interacting with the bank (or service) about 12 times on average.

Number of Cards:

The average number of cards held by customers in this cluster is 8.74. This indicates that, on average, customers in this cluster own about 9 credit cards, which suggests a higher level of credit card ownership compared to Cluster 0.

Total Number of Cards Held:

The total number of cards held by customers in this cluster is represented by the array [6, 5, 9, 8, 10, 7]. This array shows the different quantities of credit cards held by customers in this cluster, ranging from 5 to 10 cards.The higher average number of cards in this cluster compared to Cluster 0 suggests that these customers are likely more financially engaged or may have a stronger preference for credit options.

By comparing the two clusters we can understand the clustering within these clusters are not fair enough to give enough recoomendations.Hence we prefer to move with **euclidean metric** which normally perform better compared to other distance metric .

- We will check the cophenetic correlation for **Euclidean** distance with linkage methods like single, complete, average, weighted,ward and centroid.

```
Cophenetic correlation for single linkage is 0.6654277276946772.
Cophenetic correlation for complete linkage is 0.7671286292885412.
Cophenetic correlation for average linkage is 0.8375407403410419.
Cophenetic correlation for centroid linkage is 0.8355787650417409.
Cophenetic correlation for ward linkage is 0.6052249486961842.
Cophenetic correlation for weighted linkage is 0.731656222128375.
```

Dendrogram for cophenetic correlation for Euclidean distance

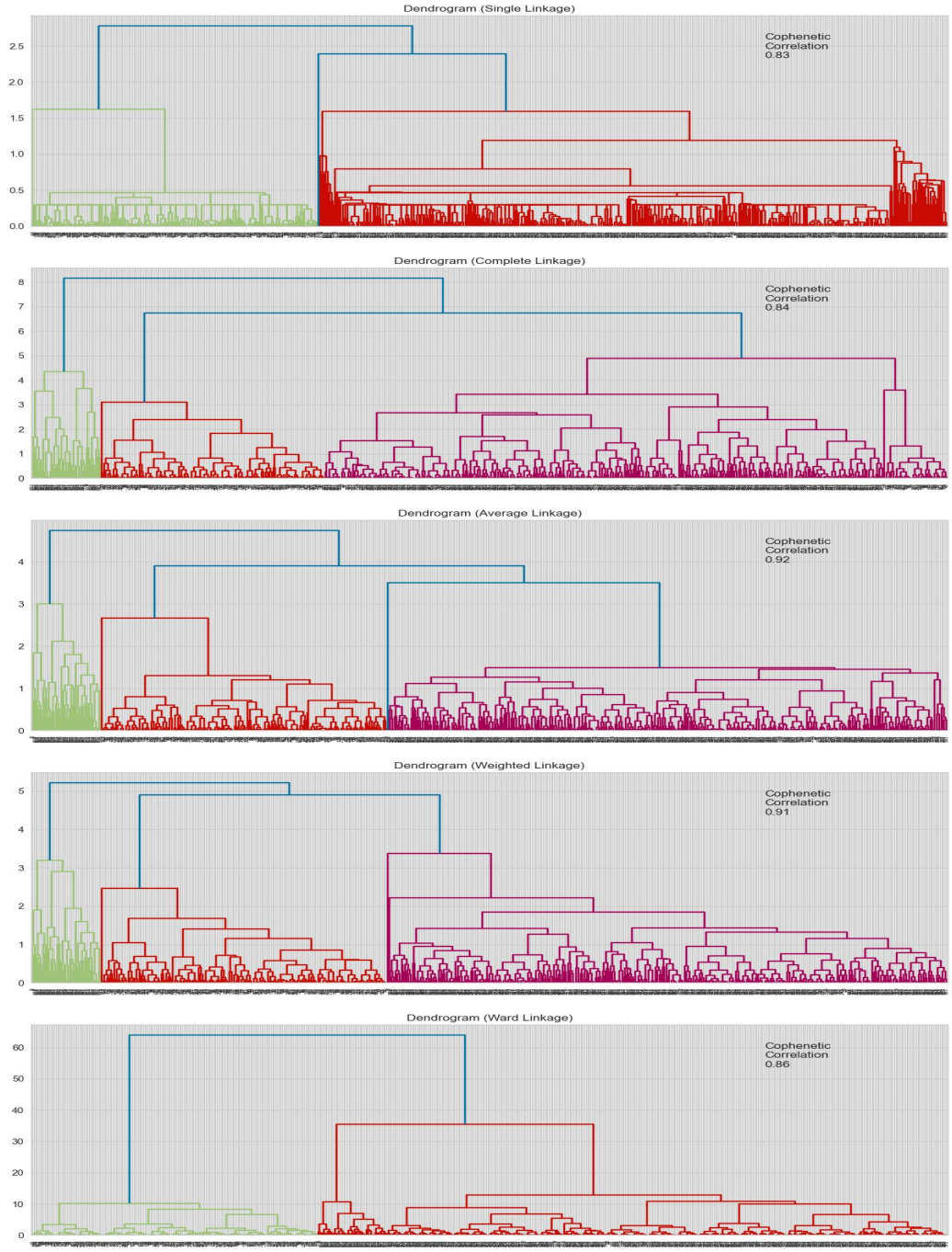

Fig 14: Dendrogram-Euclidean distance

Cluster Profiling

We perform cluster profiling, which involves summarizing the characteristics of each cluster generated by hierarchical agglomerative clustering using the model stored in the **HCmodel1 and HCmodel2** attribute with preferred linkage **average and ward.** From the dendrogram we have picked up **optimal number of cluster as 4 for average and 3 for ward .**

Euclidean(ward linkage):

```
In cluster 2, there are 336 customers.
The average number of interactions in cluster 2 is 6.24
The average number of cards in cluster 2 is 5.68
The total number of card held by the customers in cluster 2 are: [2 7 5 3 6 4]

In cluster 0, there are 274 customers.
The average number of interactions in cluster 0 is 10.74
The average number of cards in cluster 0 is 2.78
The total number of card held by the customers in cluster 0 are: [3 4 1 2 5]

In cluster 1, there are 50 customers.
The average number of interactions in cluster 1 is 12.58
The average number of cards in cluster 1 is 8.74
The total number of card held by the customers in cluster 1 are: [ 6  5  9  8 10  7]
```

Table 5: Cluster profiling - Euclidean ward linkage

Cluster 2:

Interactions: The customers in Cluster 2 are moderately engaged with their credit cards, averaging 6.24 interactions. This suggests that these customers are regular users of credit cards but do not engage in high-frequency transactions.

Number of Cards: On average, these customers hold 5.68 credit cards, which is relatively high. This indicates that they might be leveraging multiple cards for different purposes, such as using one card for travel rewards, another for cashback on groceries, etc.

Total number of cards held : The distribution of the total number of cards held [2, 7, 5, 3, 6, 4] shows variability, with some customers holding as few as 2 cards and others up to 7. This suggests a mix of moderate and heavy users within this cluster, possibly indicating varied levels of engagement with different card types.

Cluster 0:

Interactions : Customers in Cluster 0 show a high level of engagement, with an average of 10.74 interactions. This suggests these customers are frequent users of their credit cards, possibly using them for everyday transactions or business expenses.

Number of Cards: Despite their high usage, these customers hold fewer cards on average (2.78). This indicates that they prefer to concentrate their spending on a smaller, more selective portfolio of cards. They might be very loyal to specific card issuers or prefer to use cards that offer the best rewards or the lowest fees.

Total number of cards held : The card distribution [3, 4, 1, 2, 5] shows that most customers in this cluster hold between 1 and 5 cards. This reinforces the idea that they are selective and may be using cards that offer significant benefits for their specific spending habits.

Cluster 1:

Interactions : Cluster 1 customers are highly engaged, with the highest average number of interactions at 12.58. This indicates that these customers are power users of credit cards, possibly using them for a wide range of transactions, including large purchases, recurring bills, and online shopping.

Number of Cards: On average, these customers hold 8.74 cards, which is significantly higher than the other clusters. This suggests that they manage a diverse portfolio of credit cards, possibly optimizing their use across different rewards programs, interest rates, and benefits.

Total number of cards held : The distribution [6, 5, 9, 8, 10, 7] indicates that most customers in this cluster hold between 5 and 10 cards, which reflects a very diverse and extensive use of credit cards. These customers likely enjoy managing multiple cards to maximize their rewards and benefits.

Euclidean(average linkage):

```
In cluster 3, there are 1 customers.
The average number of interactions in cluster 3 is 2.00
The average number of cards in cluster 3 is 2.00
The total number of card held by the customers in cluster 3 are: [2]

In cluster 0, there are 274 customers.
The average number of interactions in cluster 0 is 10.74
The average number of cards in cluster 0 is 2.78
The total number of card held by the customers in cluster 0 are: [3 4 1 2 5]

In cluster 2, there are 335 customers.
The average number of interactions in cluster 2 is 6.26
The average number of cards in cluster 2 is 5.69
The total number of card held by the customers in cluster 2 are: [7 5 3 2 6 4]

In cluster 1, there are 50 customers.
The average number of interactions in cluster 1 is 12.58
The average number of cards in cluster 1 is 8.74
The total number of card held by the customers in cluster 1 are: [ 6  5  9  8 10  7]
```

Table 6: Cluster profiling - Euclidean average linkage

Cluster 3

Interactions : The sole customer in Cluster 3 has low engagement with their credit cards, with only 2 interactions. This suggests minimal usage, possibly indicating that the customer uses their cards only for essential or infrequent purchases.

Number of Cards: Holding only 2 cards, this customer maintains a modest portfolio. They might be using these cards for specific purposes, such as one for everyday spending and another for a specific reward program.

Cluster 0

Interactions Customers in Cluster 0 are highly engaged, with an average of 10.74 interactions. This suggests that they are regular users of their credit cards, likely utilizing them for a wide range of purchases, possibly including both everyday expenses and larger transactions.

Number of Cards:: On average, these customers hold 2.78 cards, indicating a preference for a focused portfolio. They might have one or two primary cards that they use frequently, with additional cards for specific purposes or as backups.

Total number of cards held: The total number of cards held [3, 4, 1, 2, 5] shows variability, with most customers holding between 1 and 5 cards. This distribution suggests that while some customers might be focused on maximizing the benefits of a few cards, others may be experimenting with additional cards for varied benefits.

Cluster 2

Moderate Engagement with Diverse Portfolio: Customers in Cluster 2 have a moderate level of engagement, with an average of 6.26 interactions. This suggests regular but not excessive credit card use, likely for both everyday expenses and occasional larger purchases.

Higher Card Holdings: These customers hold an average of 5.69 cards, indicating a diverse portfolio. They may be leveraging different cards for different benefits, such as rewards programs, low interest rates, or specific perks like travel insurance.

Total number of cards held: The total number of cards held [7, 5, 3, 2, 6, 4] shows a broad range, with some customers holding as few as 2 cards and others holding up to 7. This variety suggests that while some customers prefer to keep their card portfolio lean, others diversify extensively to maximize benefits.

Cluster 1

Very High Engagement, Extensive Portfolio: Cluster 1 customers are the most engaged, with an average of 12.58 interactions, indicating frequent and possibly varied use of their credit cards. These customers are likely well-versed in maximizing the benefits of their cards and may use them for a wide range of transactions.

Extensive Card Holdings: Holding an average of 8.74 cards, these customers manage a broad portfolio, possibly to optimize different rewards programs, interest rates, or other benefits. They are likely to be sophisticated users who strategically use different cards for specific types of spending.

Total number of cards held: The total number of cards held [6, 5, 9, 8, 10, 7] shows that most customers in this cluster hold between 5 and 10 cards. This extensive card portfolio suggests that these customers might be highly engaged with the credit card market, seeking to maximize rewards, benefits, and financial flexibility.

- By comparing the **Euclidean Ward and Average linkage** clustering methods, Ward's method has clustered the data points more effectively, resulting in a more meaningful segmentation of customers.

12.K-means vs Hierarchical Clustering

K-means

K_means_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_Interactions	count_in_each_segment
0	141040.000000	8.740000	12.580000	50
1	33992.105263	5.523684	6.347368	380
2	12391.304348	2.478261	11.434783	230

Hierarchical - Ward

```
In cluster 2, there are 336 customers.
The average number of interactions in cluster 2 is 6.24
The average number of cards in cluster 2 is 5.68
The total number of card held by the customers in cluster 2 are: [2 7 5 3 6 4]

In cluster 0, there are 274 customers.
The average number of interactions in cluster 0 is 10.74
The average number of cards in cluster 0 is 2.78
The total number of card held by the customers in cluster 0 are: [3 4 1 2 5]

In cluster 1, there are 50 customers.
The average number of interactions in cluster 1 is 12.58
The average number of cards in cluster 1 is 8.74
The total number of card held by the customers in cluster 1 are: [ 6  5  9  8 10  7]
```

- By comparing the K-means and Hierarchical Ward methods, both have performed similarly and provide valuable insights. However, Ward's method offers a more detailed and nuanced view of customer segmentation, while K-means might be better suited for creating clear, high-level segments based on specific criteria like credit limit and engagement.

13. ACTIONABLE INSIGHTS AND RECOMMENDATIONS

- Customers with 1-3 credit cards have raised more queries, as have those with 8-10 credit cards, indicating that these groups are likely making more transactions. Understanding the repetitive queries raised by these customers and providing tailored support could enhance customer satisfaction. Offering lower interest rates and cashback incentives to customers with one or two credit cards could encourage more frequent transactions. For customers with 8-10 cards, travel allowances, airport lounge access, and offers on hotel bookings could increase their engagement.

- From both clustering methods, we observe that customers with 4-7 credit cards have approached the bank less frequently for transaction issues, which suggests they may be making fewer transactions. This likely indicates they could be seasonal spenders. Offering them special seasonal promotions, such as on their birthdays, their children's birthdays, or during festive occasions, could encourage more consistent credit card use.

- Customers with 1-3 credit cards have raised a higher number of queries, which suggests they may be relatively new to using credit cards or are more cautious with their transactions.Providing these customers with proactive support, such as personalized assistance or quick response times to resolve their transaction issues and offer guides or tutorials on how to maximize the benefits of their cards, reducing confusion and queries.Also introduce loyalty programs with incremental rewards based on the number of transactions, encouraging them to use their cards more frequently.Offer cashback or discounts on everyday spending categories like groceries or fuel to drive up engagement.

- Customers with 8-10 credit cards also raise a lot of queries, likely because they are high-volume transaction users or have complex financial needs,providing them with premium services with decidated service lines to address their needs quickly and efficiently.Since these customers are likely to use multiple cards, offering credit limit increases could enhance their spending power.Providing them with advanced transaction alerts and fraud protection to ease concerns over multiple card use.

- Customers with 4-7 may be more responsive to offers that introduce new products like personal loans or mortgage options linked to their credit cards, given their stable nature.Offer benefits tied to subscriptions like streaming services, fitness memberships, or meal deliveries to align with their likely interest in convenience.

- The strategies mentioned can also be applied to attract new customers by categorizing them into similar groups based on their potential credit card usage, such as 1-3, 4-7, and 8 or more credit card holders. By identifying which group new customers are likely to fall into, the bank can implement tailored approaches to meet their specific needs and preferences.