Credit Card
Customer Segmentation

BUSINESS REPORT

PGP DSA 28 SEP 2025

ANJU SAINI

INDEX

Graphs

1) Business 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

2) 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.

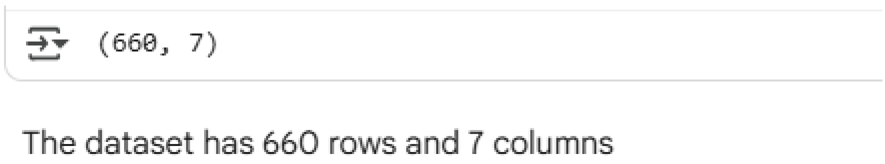
3) Data Description :-

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 center.

4) 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)

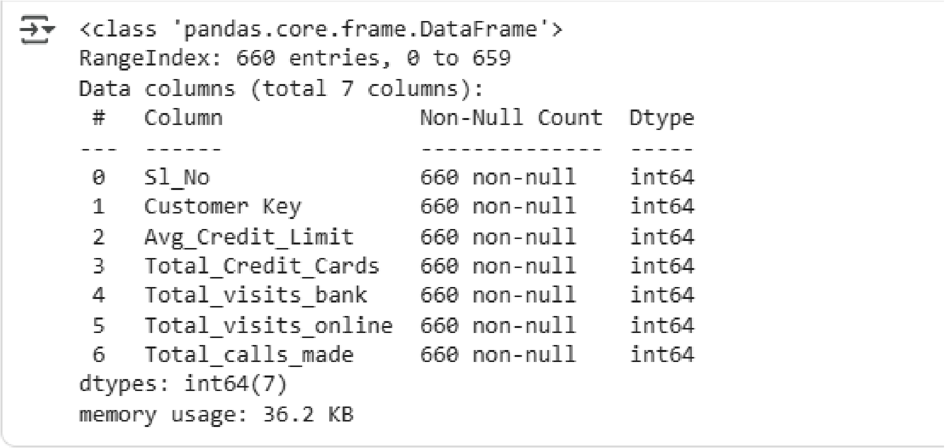
5) Checking the shape of the Dataset



```
(660, 7)
```

The dataset has 660 rows and 7 columns

6) Data information



```
<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
```

This dataset contains 7 columns, and all of the columns are numerical.

7) Exploratory Data Analysis

In data Analysis the following steps have been completed:

- Problem definition

- Univariate analysis

- Bivariate analysis

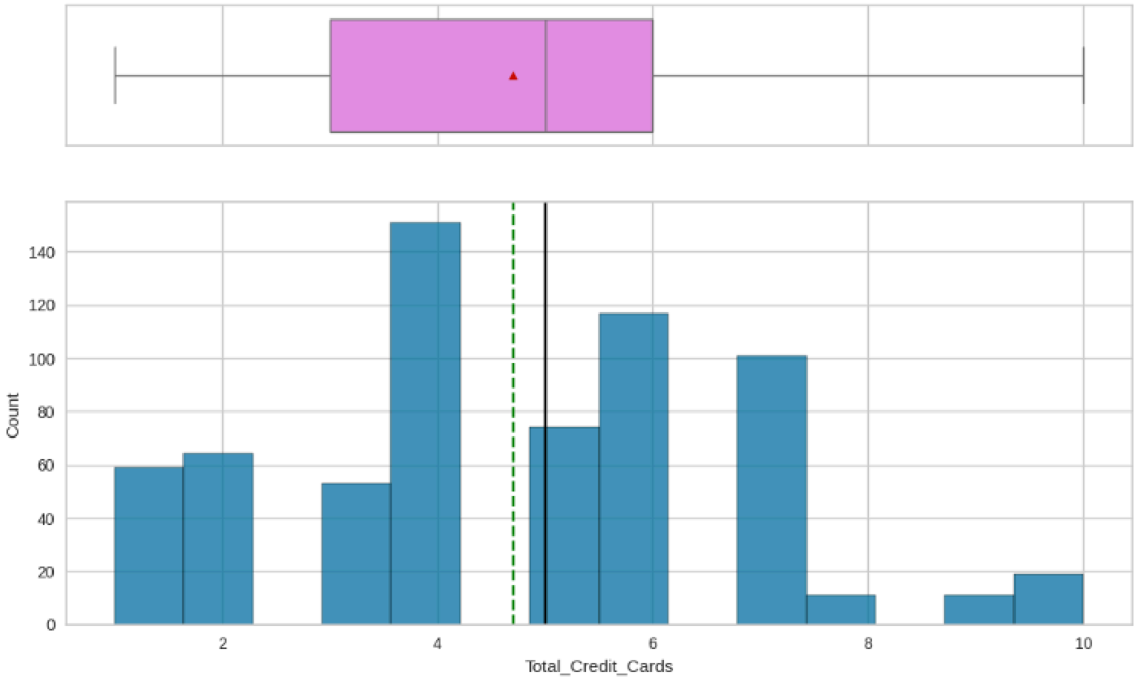
- Use appropriate visualizations to identify the patterns and insights

- Key meaningful observations on individual variables and the relationship between variables

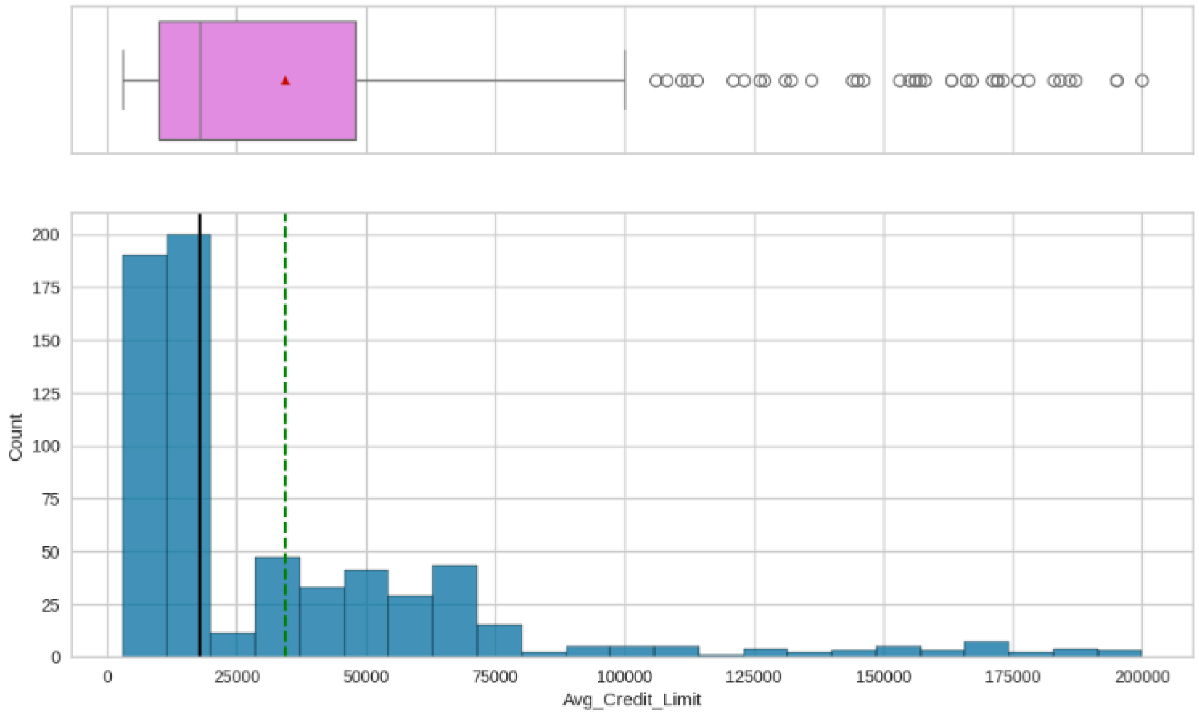
6.1) Problem Definition:-

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

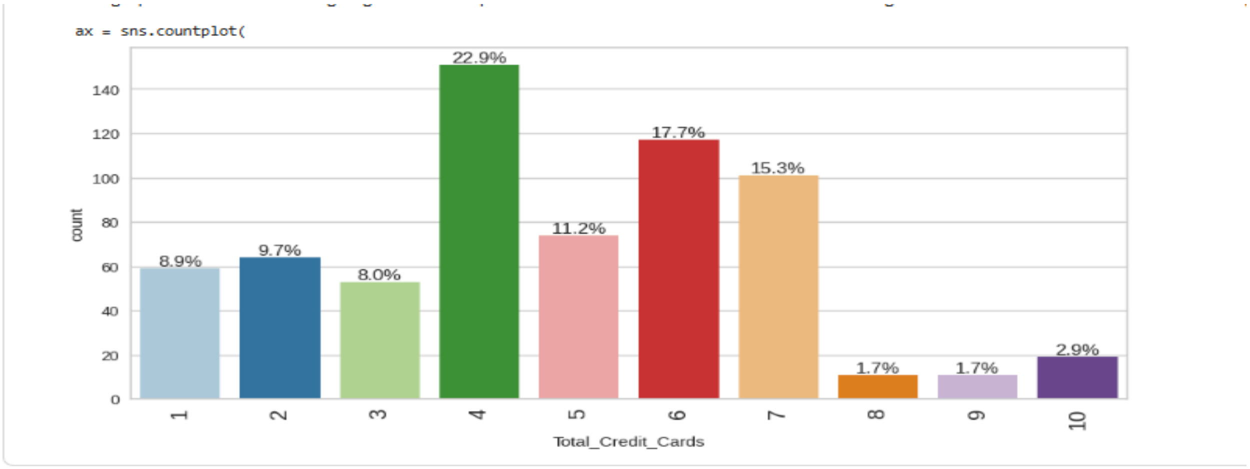
6.2) Univariate analysis :-



Total Credit Cards Customers typically hold 3–6 cards, with a few outliers managing up to 10.



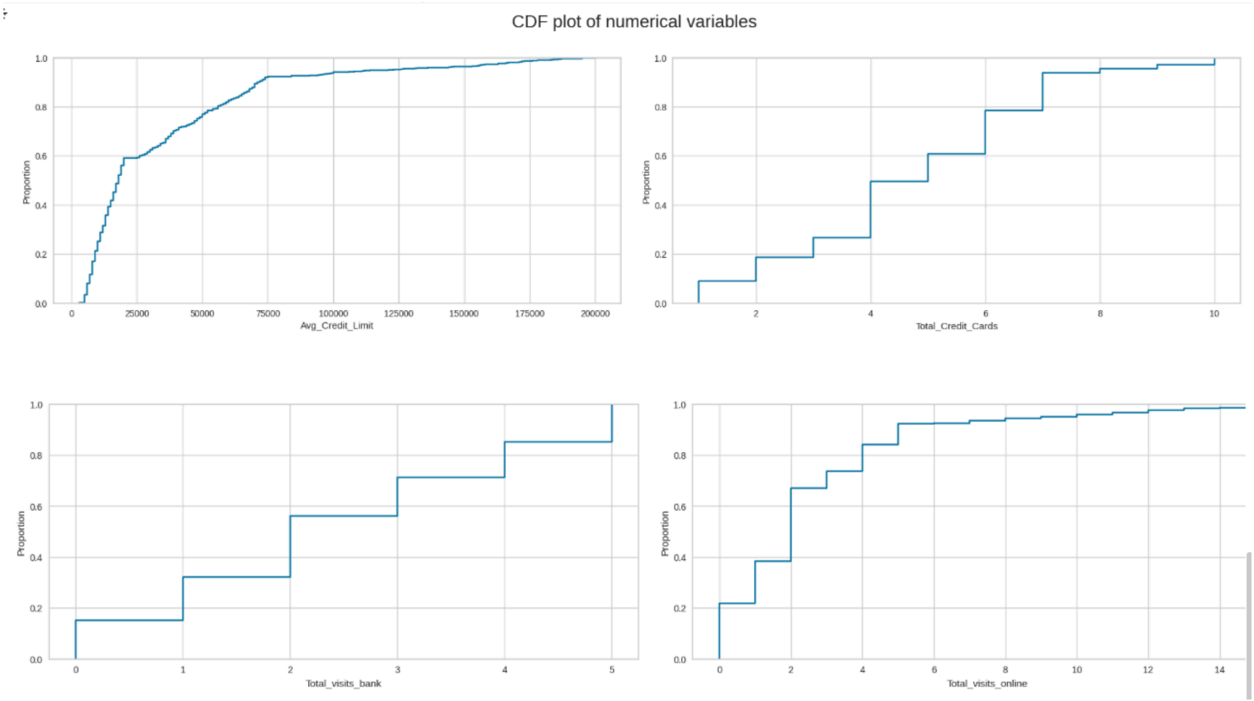
Most customers have modest credit limits, but a small premium segment holds significantly higher limits.

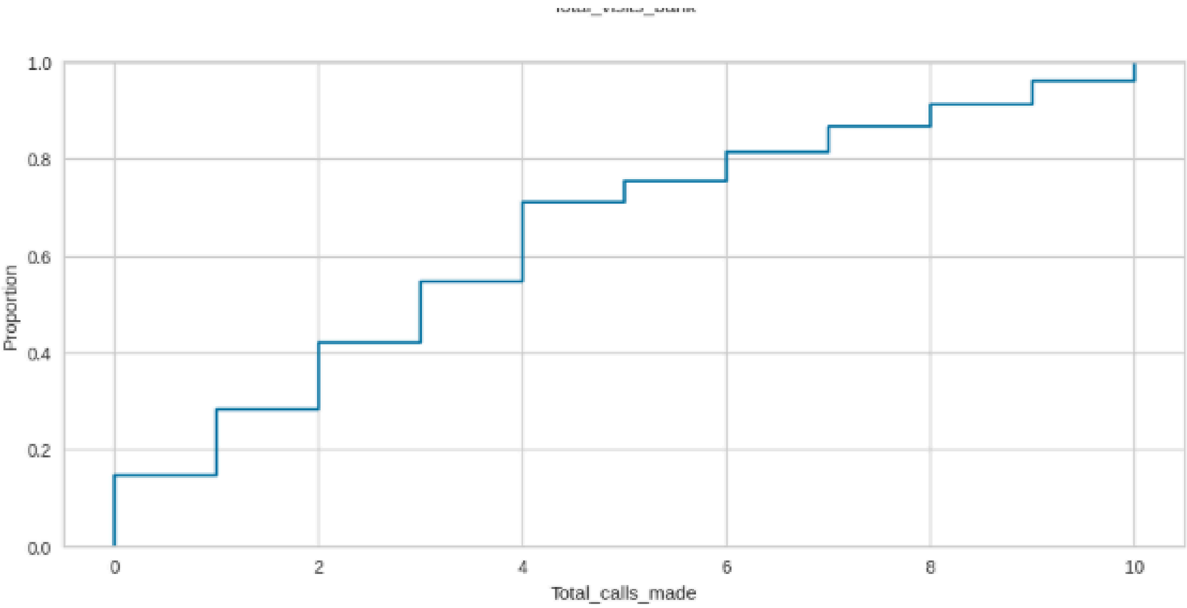


```
ax = sns.countplot(
```

In this barplot show no of customer that get a card

There is 22% of customer who has credit card count is 4


CDF plot of numerical variables



Avg_Credit_Limit Most customers have fairly low credit limits

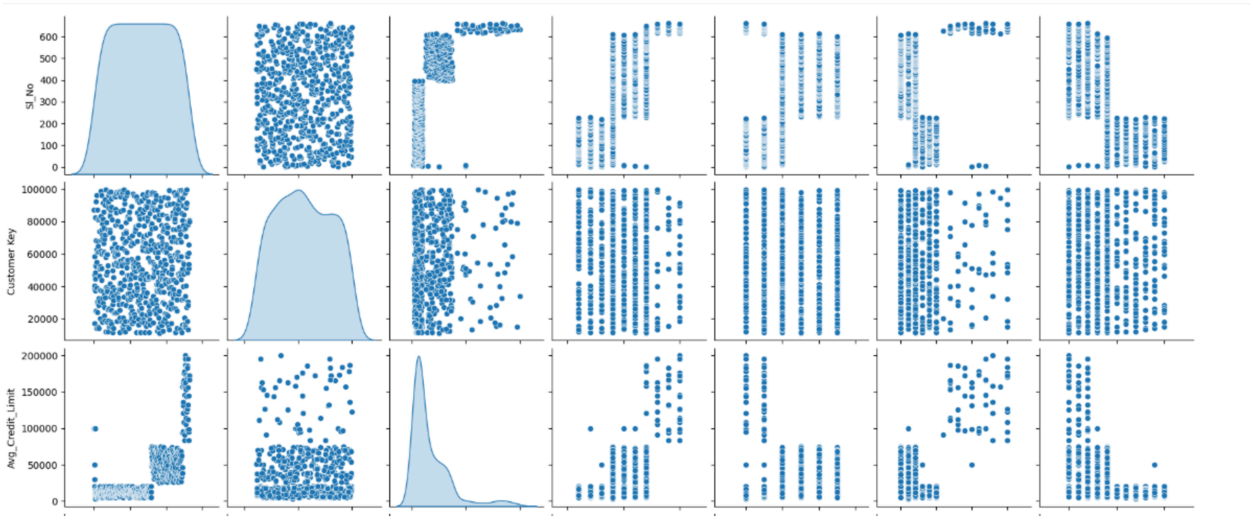
Total_Credit_Cards Since you can’t have half a credit card, the graph looks like steps

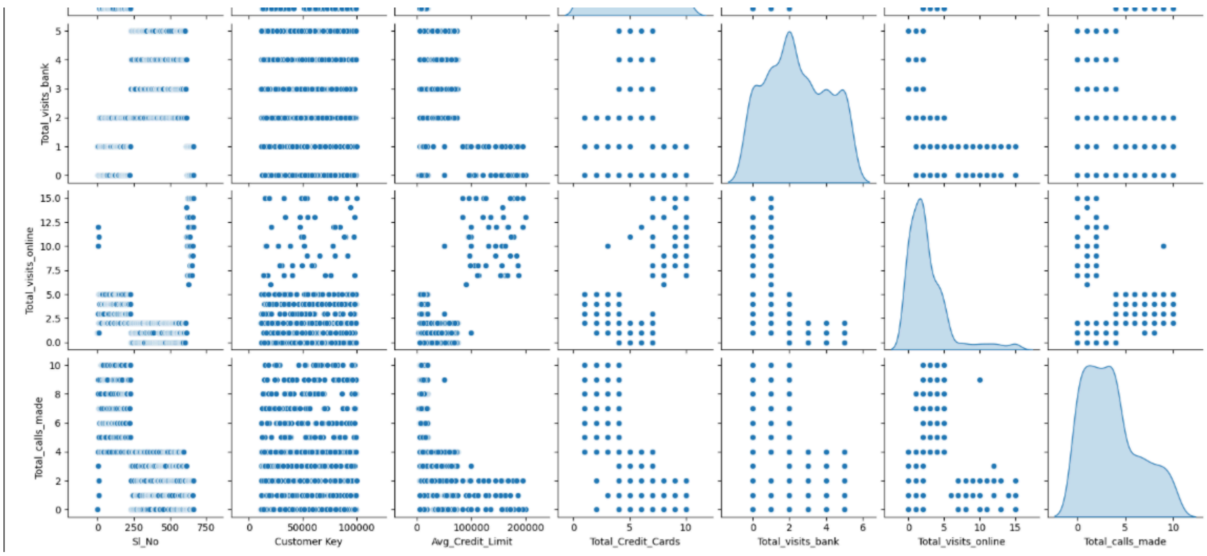
Total_visits_bank The curve shoots up early, meaning the vast majority of customers rarely step into a branch. Most are probably handling things remotely — or just don’t need to visit often. Only a handful are frequent visitors.

Total_visits_online Similar story here — most customers don’t log in that often. But unlike bank visits, there’s a longer tail, meaning we do have a small group of power users who are super active online.

Total_calls_made This one’s interesting — the curve climbs steadily up to around 5 calls, then plateaus. So a lot of customers make a few calls but very few go beyond that.

6.3) Bivariate Analysis



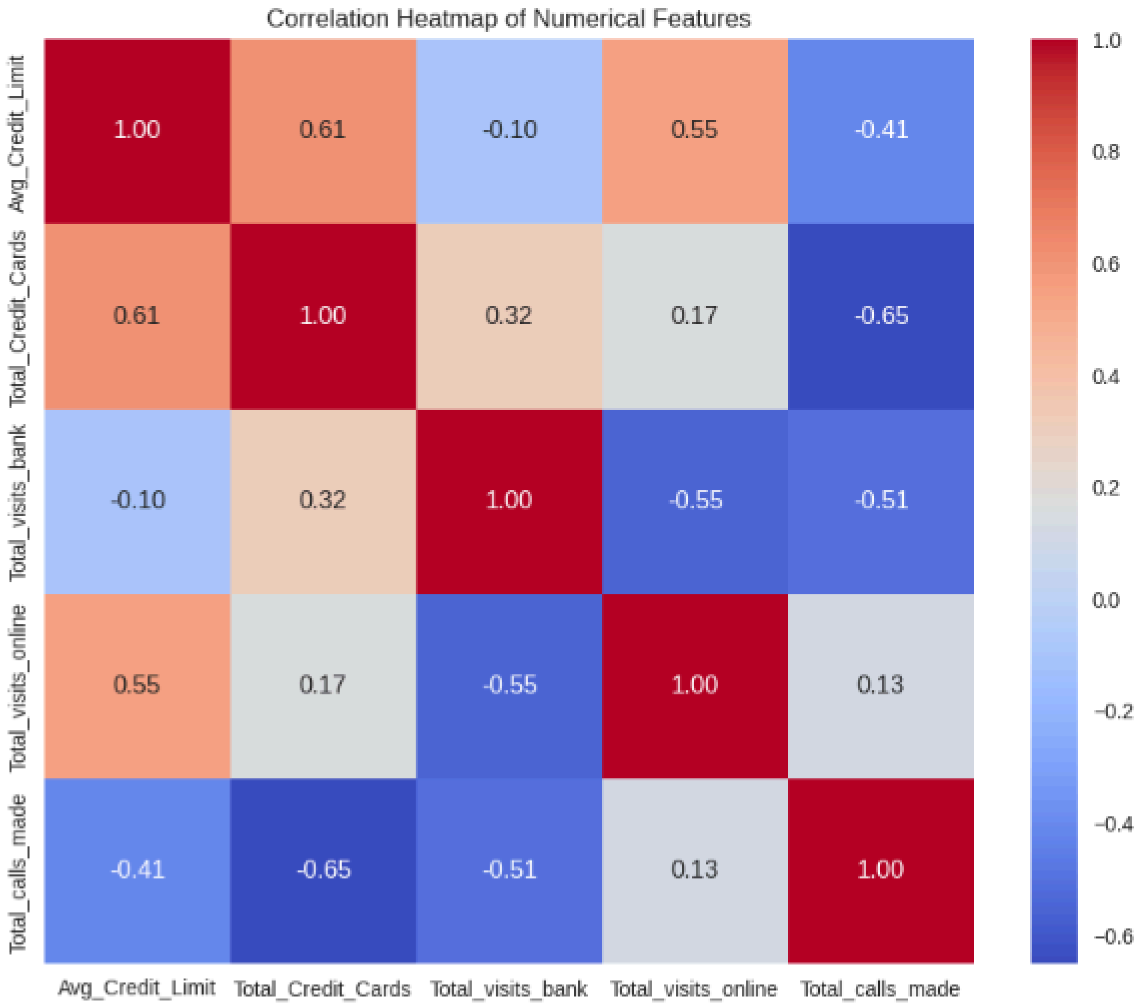


Most customers have low bank visits and calls, but a few customers are highly active.

Online visits are more common, but still skewed towards lower counts.

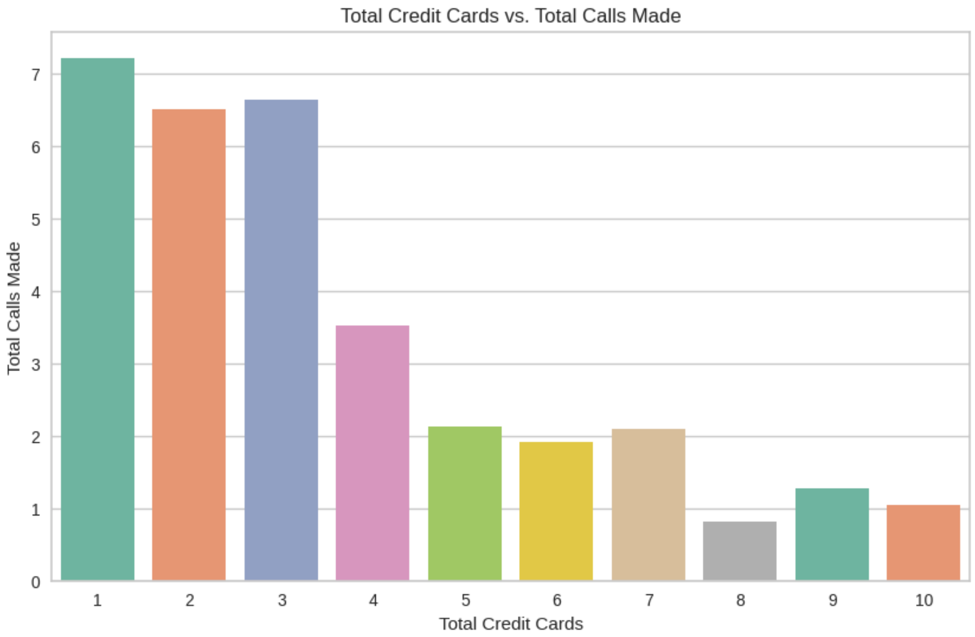
Avg Credit Limit is highly skewed: most customers have modest limits, but a small group has very high credit limits.

Customers with more cards usually have higher credit limits.

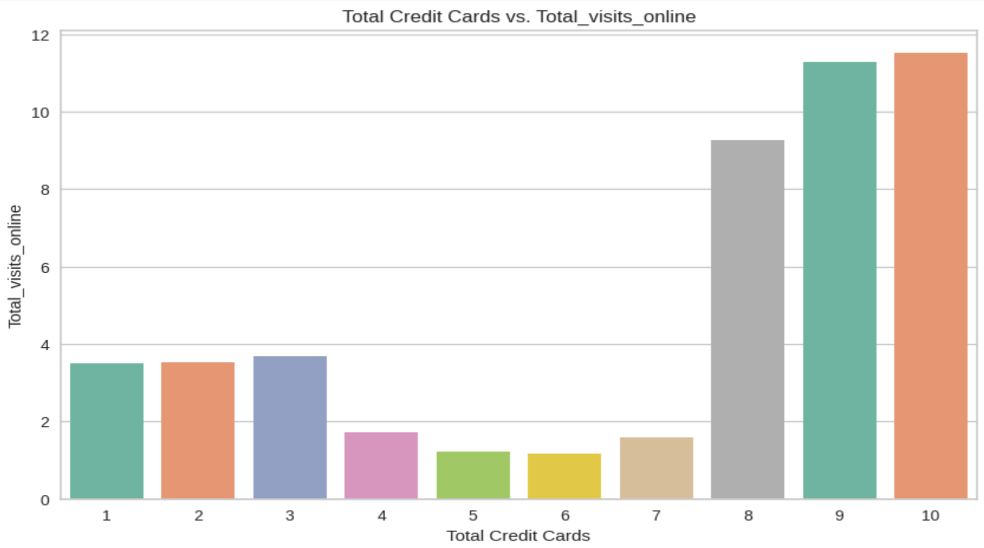


Avg_Credit_Limit is positively correlated with Total_Credit_Cards Total_visits_online which can make sense. Avg_Credit_Limit is negatively correlated with Total_calls_made and Total_visits_bank. Total_visits_bank, Total_visits_online, Total_calls_made are negatively correlated which implies that the majority of customers use only one of these channels to contact the bank.

6.4) Visualizations to identify the patterns and insights

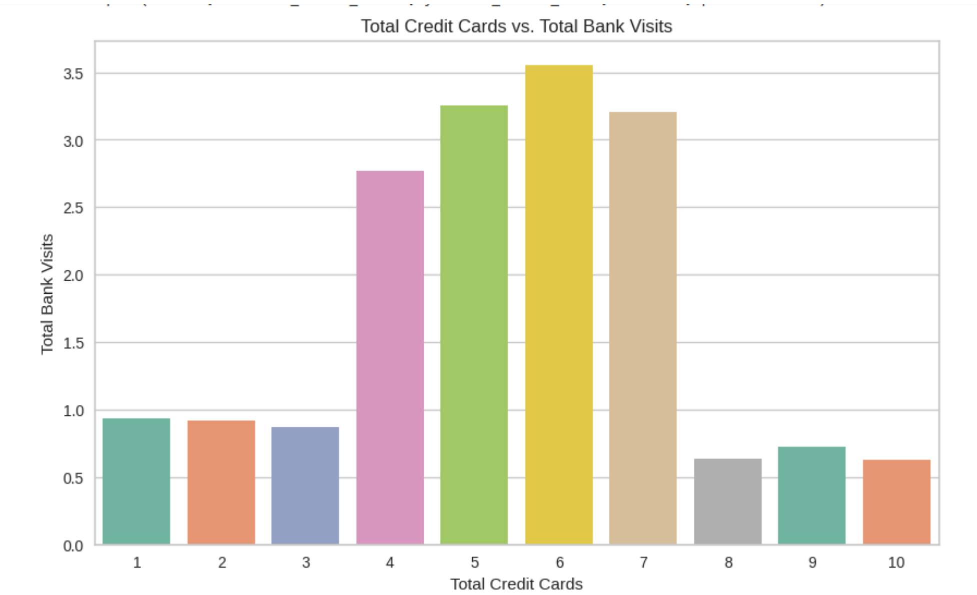


Customers with 1, 2, and 3 credit cards seem to have a relatively higher number of calls made compared to those with more credit cards.As the number of credit cards increases beyond 3, the average number of calls made generally decreases.Customers with a very high number of credit cards (7 or more) appear to make fewer calls on average compared to those with a moderate number of cards.



The customer has high credit card has high online visit

If the customer has medium credit card has low online visit



Customer who has medium credit card like 4,5,6,7 they have high bank visit

Other low and high credit card has low bank visit

6.5) Key meaningful observations on individual variables and the relationship between variables

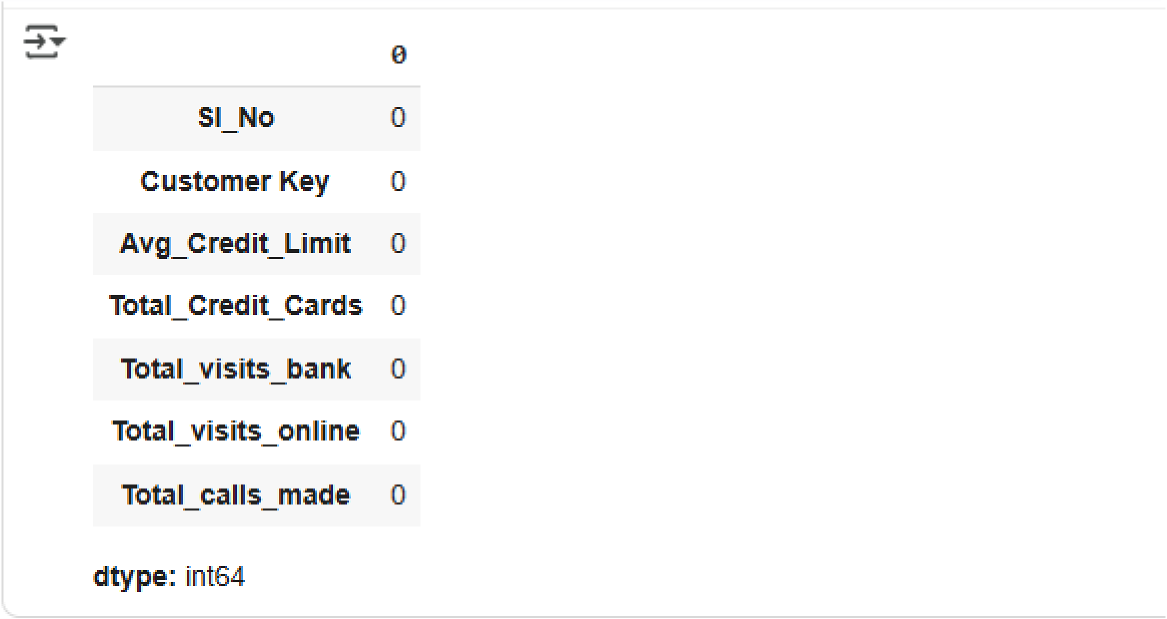
1. Low Card Holders (1–3 cards) These customers tend to prefer calling customer service to resolve queries or seek assistance.
2. Medium Card Holders (4–7 cards) This group shows a stronger inclination to visit bank branches physically.
3. High Card Holders (8–10+ cards) These customers are more likely to use online channels for support or transactions.

7) Data Preprocessing

In data preprocessing, the following steps have been completed:

- Checking for missing values
- Checking for duplicate values
- Outlier detection and treatment
- Feature engineering
- Data scaling

7.1) Checking Missing Values

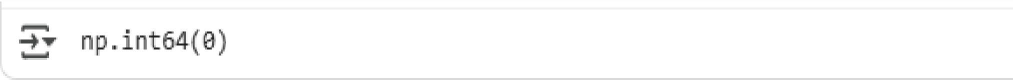


	0
Sl_No	0
Customer Key	0
Avg_Credit_Limit	0
Total_Credit_Cards	0
Total_visits_bank	0
Total_visits_online	0
Total_calls_made	0

dtype: int64

No missing values were found, so no imputation was needed.

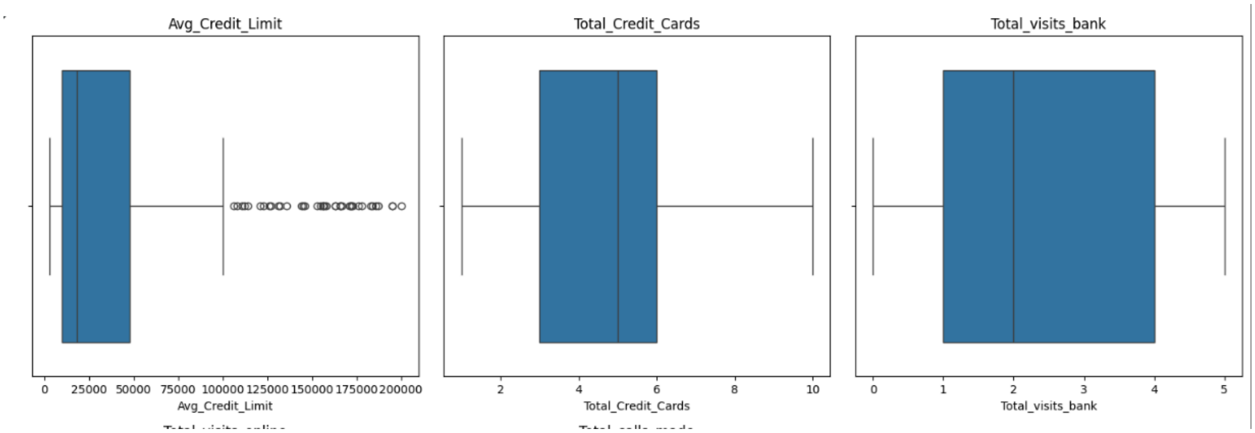
7.2) Checking Duplicate Values

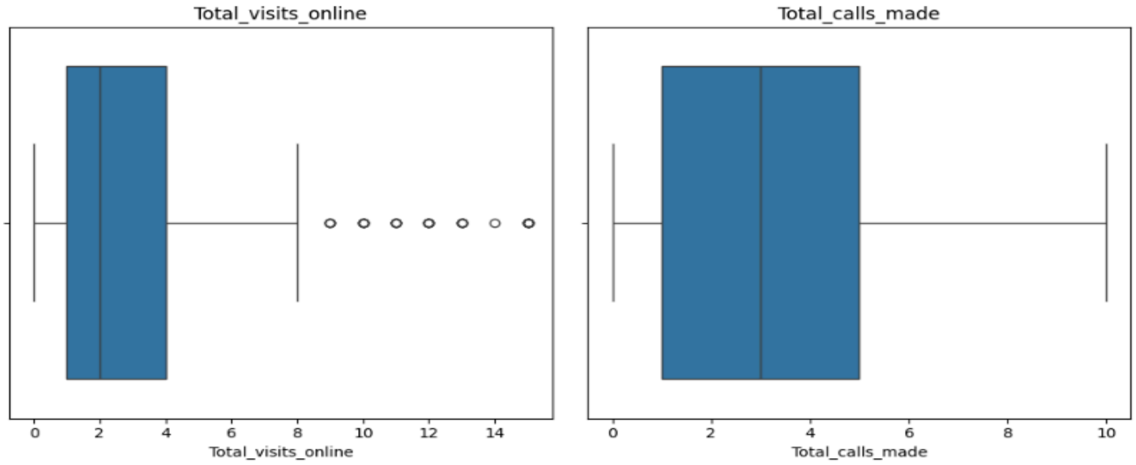


```
np.int64(0)
```

There is no any Duplicate Value in the given data set

7.3) Outlier Detection And Treatment





During the exploratory analysis, I observed outliers in features such as Average Credit Limit and Total Online Visits. However, I chose not to remove these outliers. In this context, high values may represent legitimate customer behavior rather than errors—for example, some customers naturally have higher credit limits or significantly more online interactions. Removing such data could risk losing valuable insights about high-value or highly engaged customers.

7.4) Feature Engineering

To better capture customer behavior and credit patterns, I created several derived features:

- Total Activity: Combined the number of bank visits, online visits, and calls to measure overall engagement.

- Online Ratio: Ratio of online visits to bank visits (with a small adjustment to avoid division errors), highlighting customers' preference for digital banking.

- Credit Exposure: Product of average credit limit and total credit cards, providing a view of overall credit capacity.
- Credit Level: Categorized average credit limits into three groups like Low, Medium, and High for easier segmentation.
- Log Credit: Applied logarithmic transformation to average credit limit to reduce skewness.
- Digital Score: Difference between online visits and offline interactions (bank visits and calls), showing how digitally inclined a customer is.

7.5) Data Scaling

To ensure that variables with different ranges (credit limit in thousands vs. visit counts in single digits do not dominate clustering, I standardized all the selected features using StandardScaler. This step brought every feature onto a comparable scale while preserving their distribution. The scaled data was then used for clustering analysis.

8) K-means Clustering

In data K-means Clustering, the following steps have been completed:

- Apply K-means Clustering

- Plot the Elbow curve

- Check Silhouette Scores

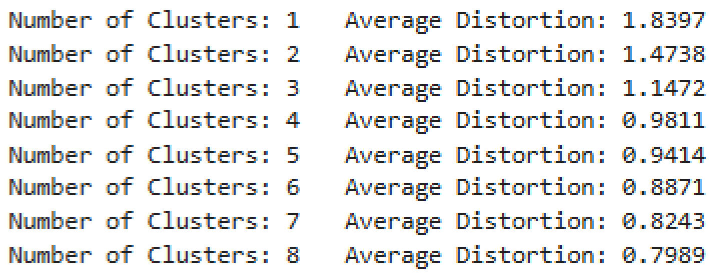
- Figure out the appropriate number of clusters

- Cluster Profiling

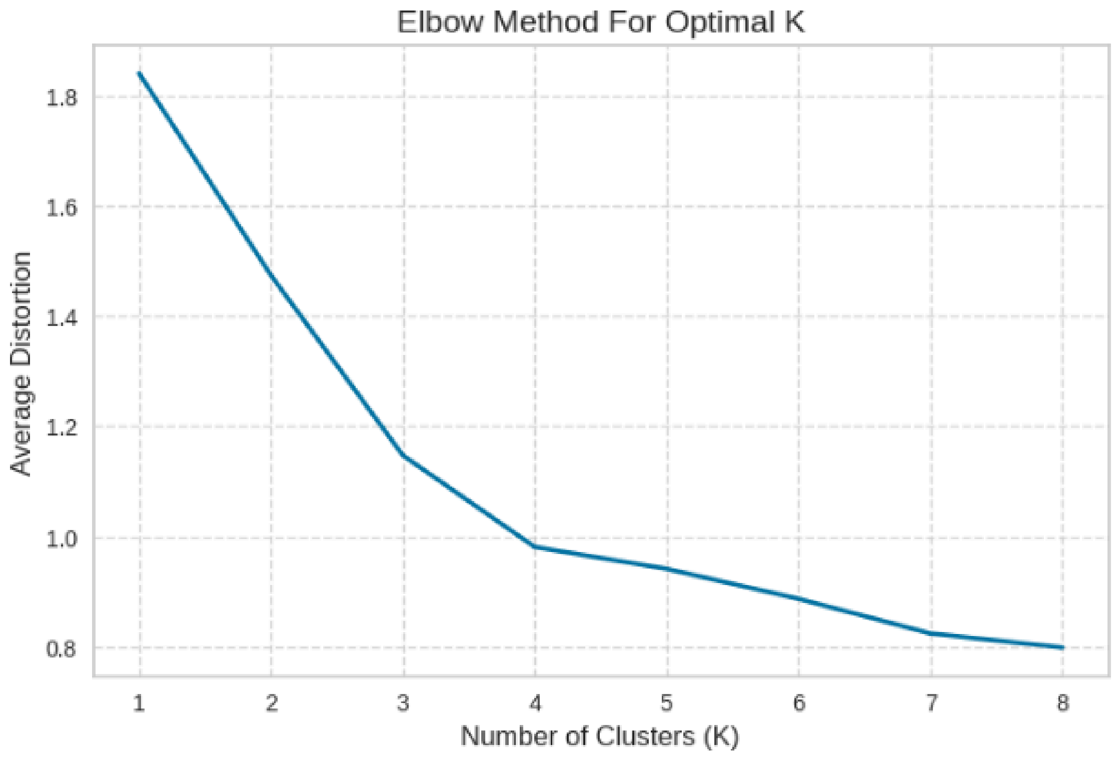
8.1) Apply K-means Clustering

Tested different values of K (number of clusters) and assigned cluster labels to each customer**.**

8.2) Plot the Elbow curve

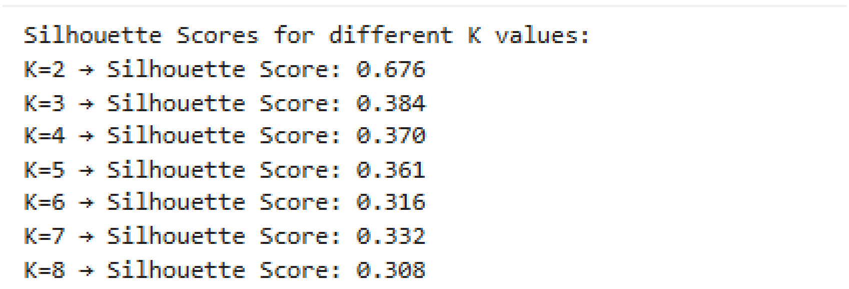


```
Number of Clusters: 1   Average Distortion: 1.8397
Number of Clusters: 2   Average Distortion: 1.4738
Number of Clusters: 3   Average Distortion: 1.1472
Number of Clusters: 4   Average Distortion: 0.9811
Number of Clusters: 5   Average Distortion: 0.9414
Number of Clusters: 6   Average Distortion: 0.8871
Number of Clusters: 7   Average Distortion: 0.8243
Number of Clusters: 8   Average Distortion: 0.7989
```

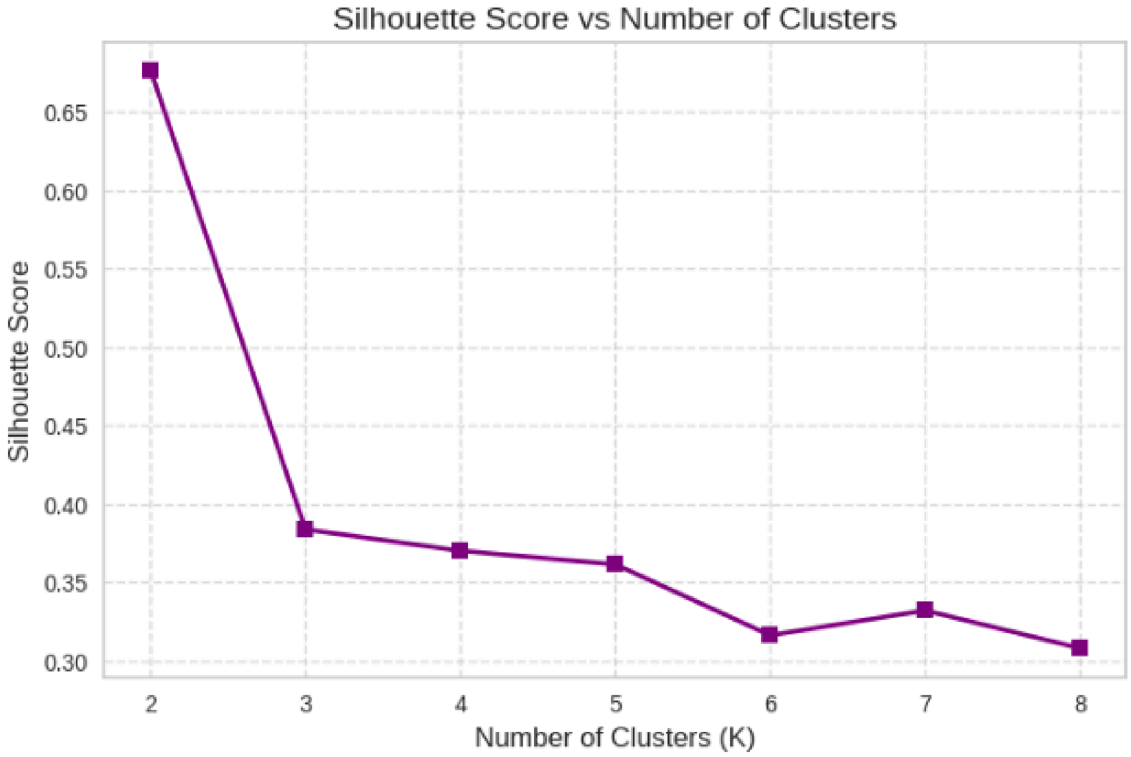


Using the Elbow method, I checked how the average distortion decreases with more clusters. The curve started flattening around K=2 to K=3, suggesting that too many clusters wouldn’t add much value.

8.3) Check Silhouette Scores

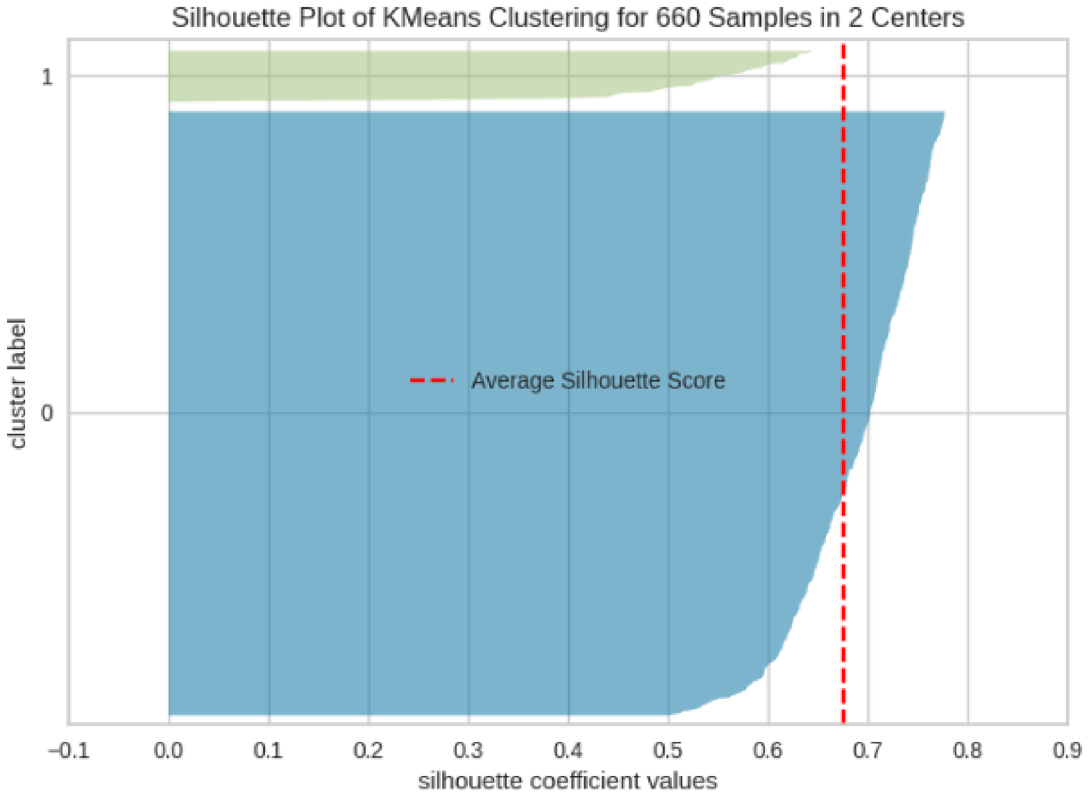


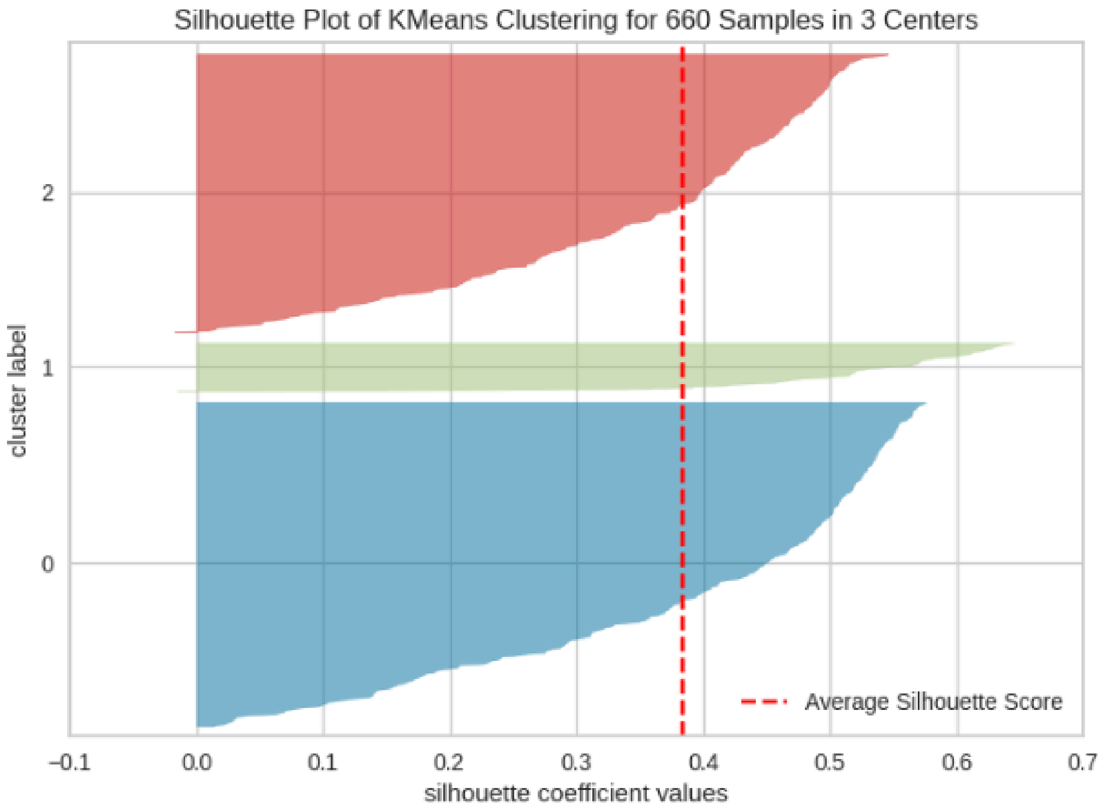
```
Silhouette Scores for different K values:
K=2 → Silhouette Score: 0.676
K=3 → Silhouette Score: 0.384
K=4 → Silhouette Score: 0.370
K=5 → Silhouette Score: 0.361
K=6 → Silhouette Score: 0.316
K=7 → Silhouette Score: 0.332
K=8 → Silhouette Score: 0.308
```



To further validate, I calculated silhouette scores for different values of K. The highest score (0.676) was observed at K=2, which indicates that 2 clusters give the most meaningful separation of customers.

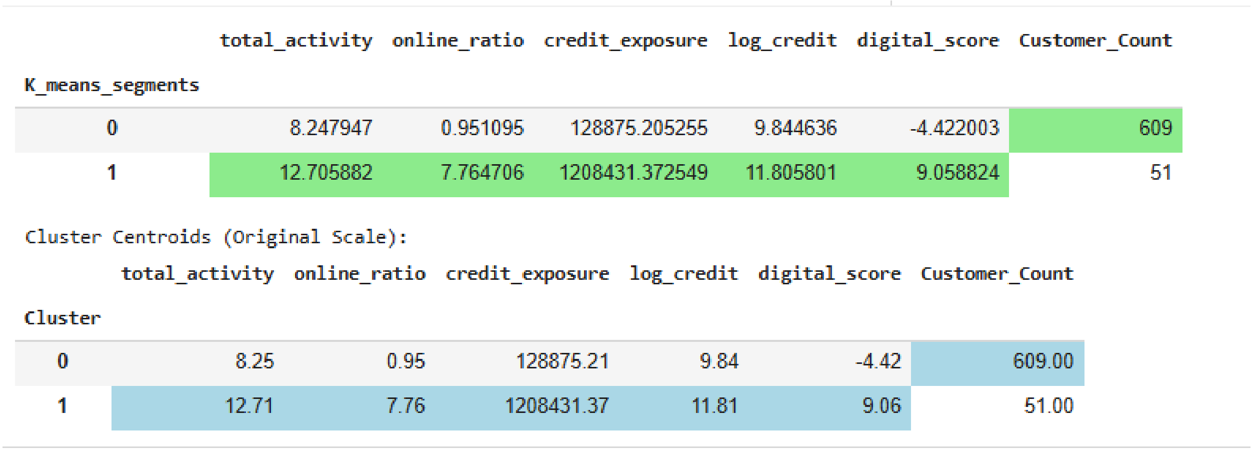
8.4) Figure out the appropriate number of clusters





Based on both the Elbow curve and Silhouette scores, I finalized K=2 as the optimal number of clusters.

8.5) Cluster Profiling



K_means_segments	total_activity	online_ratio	credit_exposure	log_credit	digital_score	Customer_Count
0	8.247947	0.951095	128875.205255	9.844636	-4.422003	609
1	12.705882	7.764706	1208431.372549	11.805801	9.058824	51

Cluster Centroids (Original Scale):

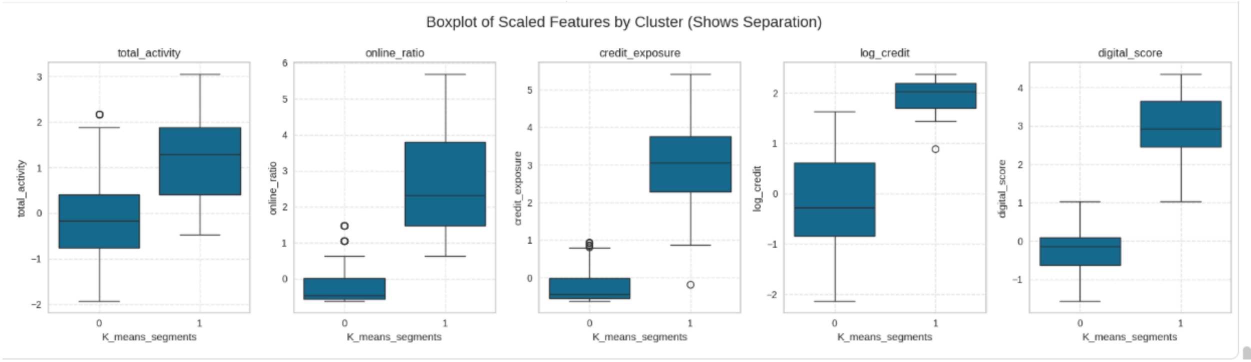
Cluster	total_activity	online_ratio	credit_exposure	log_credit	digital_score	Customer_Count
0	8.25	0.95	128875.21	9.84	-4.42	609.00
1	12.71	7.76	1208431.37	11.81	9.06	51.00

After assigning customers to the two clusters, I analyzed their average values across features.

- Cluster 0 (609 customers): Moderate total activity, lower online ratio, lower digital score, and smaller overall credit exposure. This group seems to represent traditional or moderate users.

- Cluster 1 (51 customers): Much higher activity, significantly higher online ratio and digital score, and very large credit exposure. This group represents digitally active, high-credit customers.

8.6) Boxplot of Scaled Features by Cluster



Total activity: Cluster 1 (Premium & Digital) is the most active overall. Clusters 0 (Moderate Engagement) and 2 (Low Engagement & High Callers) are less active and quite similar.

Online ratio: Cluster 1 does much more online, while Clusters 0 and 2 rely more on branches and calls.

Credit exposure: Highest in Cluster 1, moderate in Cluster 0, and lowest in Cluster 2.

Log credit: Follows the same pattern—Cluster 1 highest, then 0, then 2.

Digital score: Cluster 1 is very digital-first. Clusters 0 and 2 are less so, with Cluster 2 leaning more toward calls and visits.

9)Hierarchical Clustering

- Apply Hierarchical clustering with different linkage methods

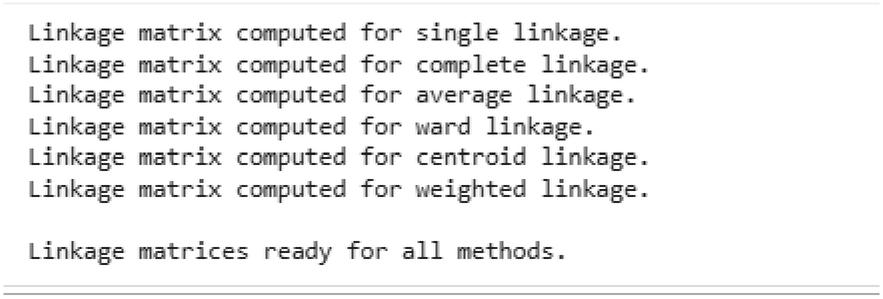
- Plot dendrograms for each linkage method

- Check cophenetic correlation for each linkage method

- Figure out the appropriate number of clusters

- Cluster Profiling

9.1) Apply Hierarchical clustering with different linkage methods



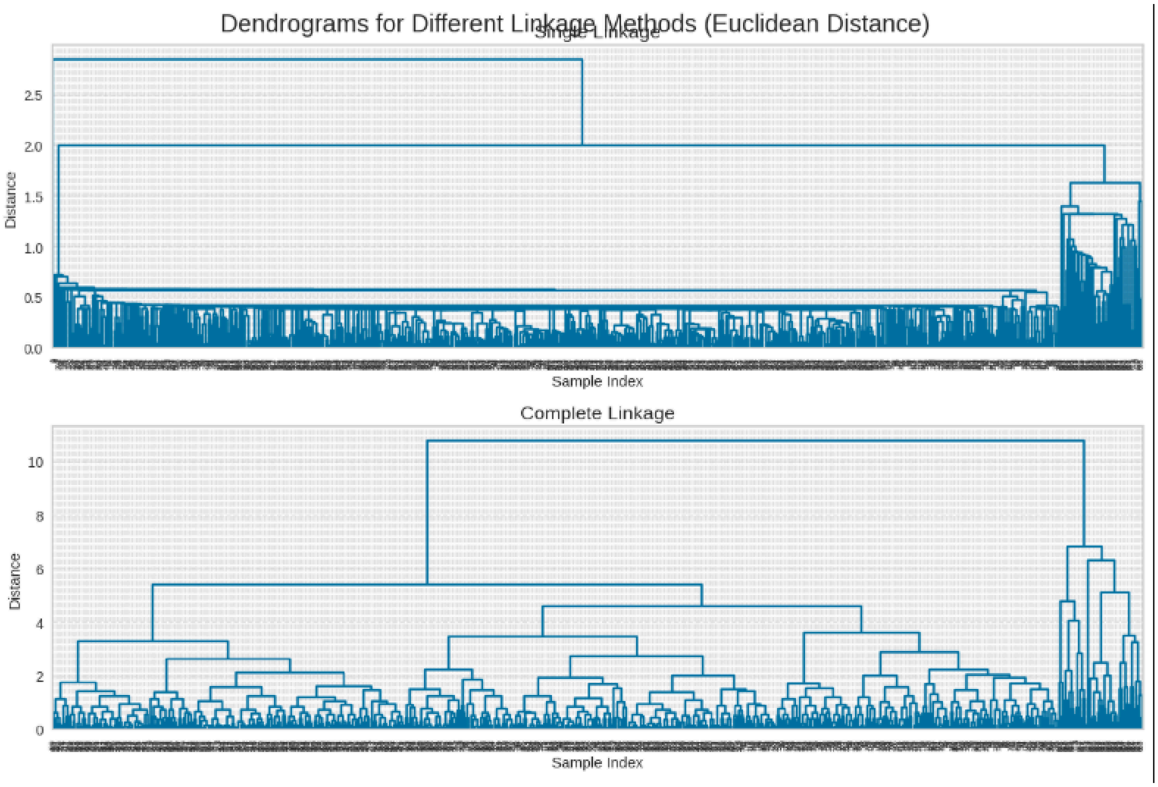
```
Linkage matrix computed for single linkage.
Linkage matrix computed for complete linkage.
Linkage matrix computed for average linkage.
Linkage matrix computed for ward linkage.
Linkage matrix computed for centroid linkage.
Linkage matrix computed for weighted linkage.

Linkage matrices ready for all methods.
```

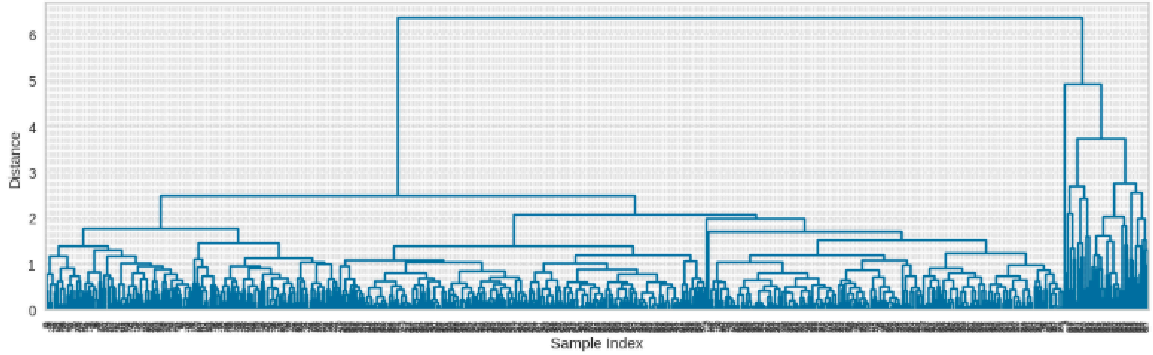
I tested several linkage methods, including single, complete, average, ward, centroid, and weighted. Each method represents a different way of calculating the distance between clusters.

9.2) Plot dendrograms for each linkage method

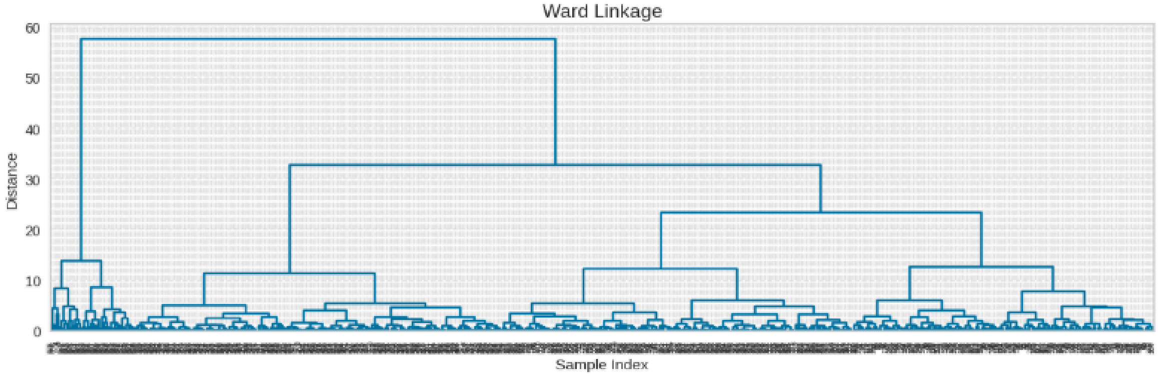
Dendrograms were plotted for each method to visually assess how customers merge into clusters. This provided insights into possible cut-off points for deciding the number of clusters.



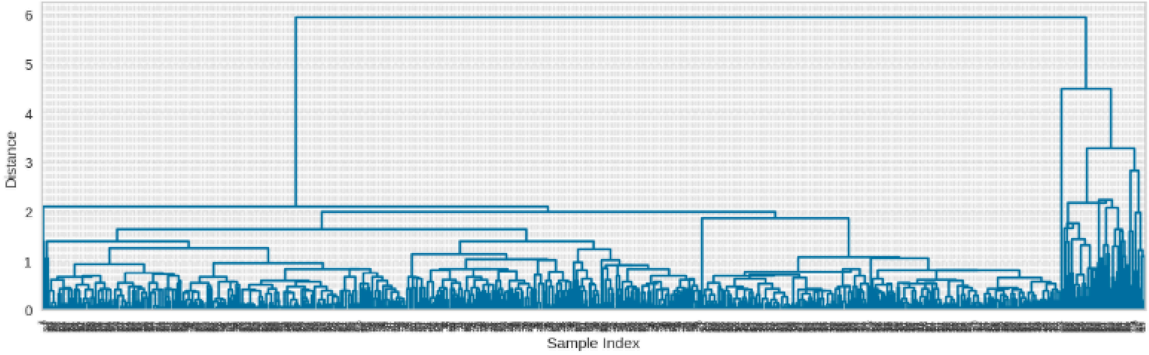
Average Linkage



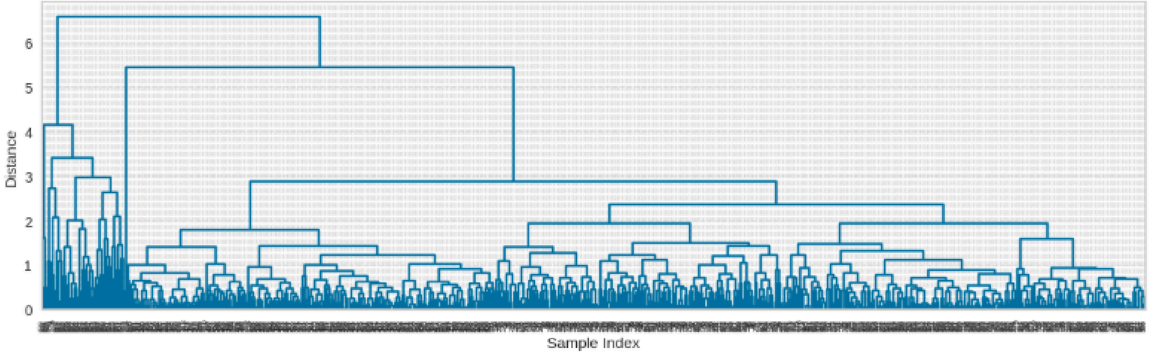
Ward Linkage



Centroid Linkage

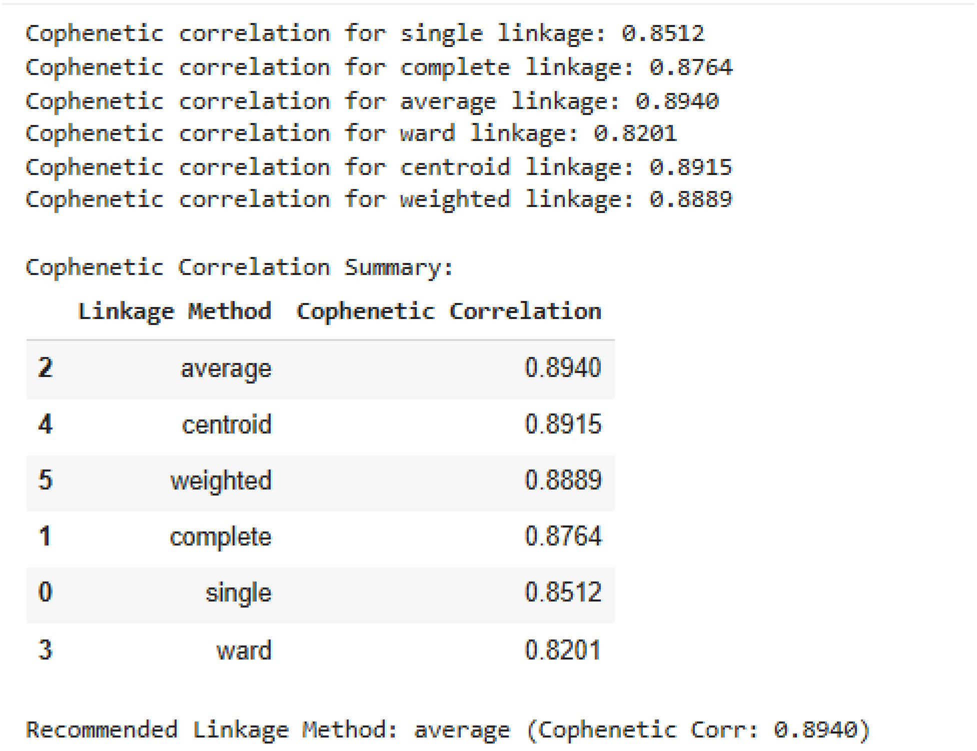


Weighted Linkage



The dendrograms show how different methods group customers based on similarity. Think of each line as a merge where two groups of customers are combined because they are close to each other in terms of their features. Taller lines mean those groups were less similar before being merged. By looking at these, we can get a feel for how distinct the potential customer groups are with each clustering method.

9.3) Check cophenetic correlation for each linkage method



```
Cophenetic correlation for single linkage: 0.8512
Cophenetic correlation for complete linkage: 0.8764
Cophenetic correlation for average linkage: 0.8940
Cophenetic correlation for ward linkage: 0.8201
Cophenetic correlation for centroid linkage: 0.8915
Cophenetic correlation for weighted linkage: 0.8889

Cophenetic Correlation Summary:
```

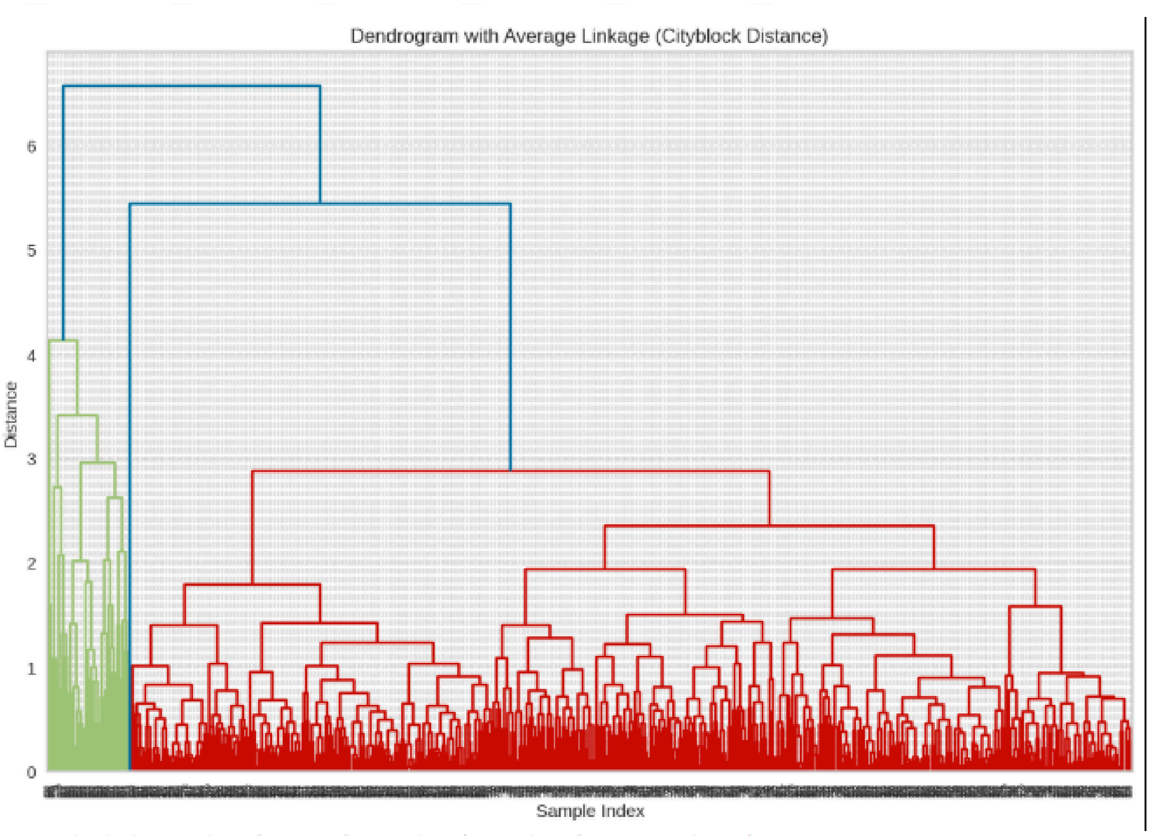
	Linkage Method	Cophenetic Correlation
2	average	0.8940
4	centroid	0.8915
5	weighted	0.8889
1	complete	0.8764
0	single	0.8512
3	ward	0.8201

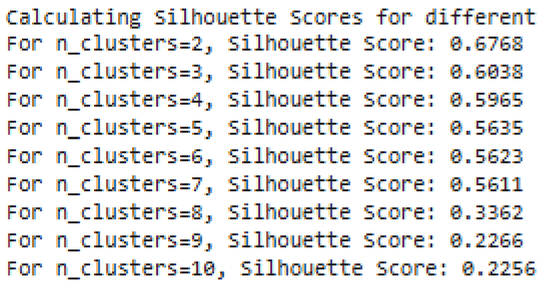
```
Recommended Linkage Method: average (Cophenetic Corr: 0.8940)
```

To evaluate which linkage method best preserved the original distances, I calculated the cophenetic correlation coefficient. Among all methods, average linkage performed the best with a score of 0.894, meaning it represented the data structure more accurately.

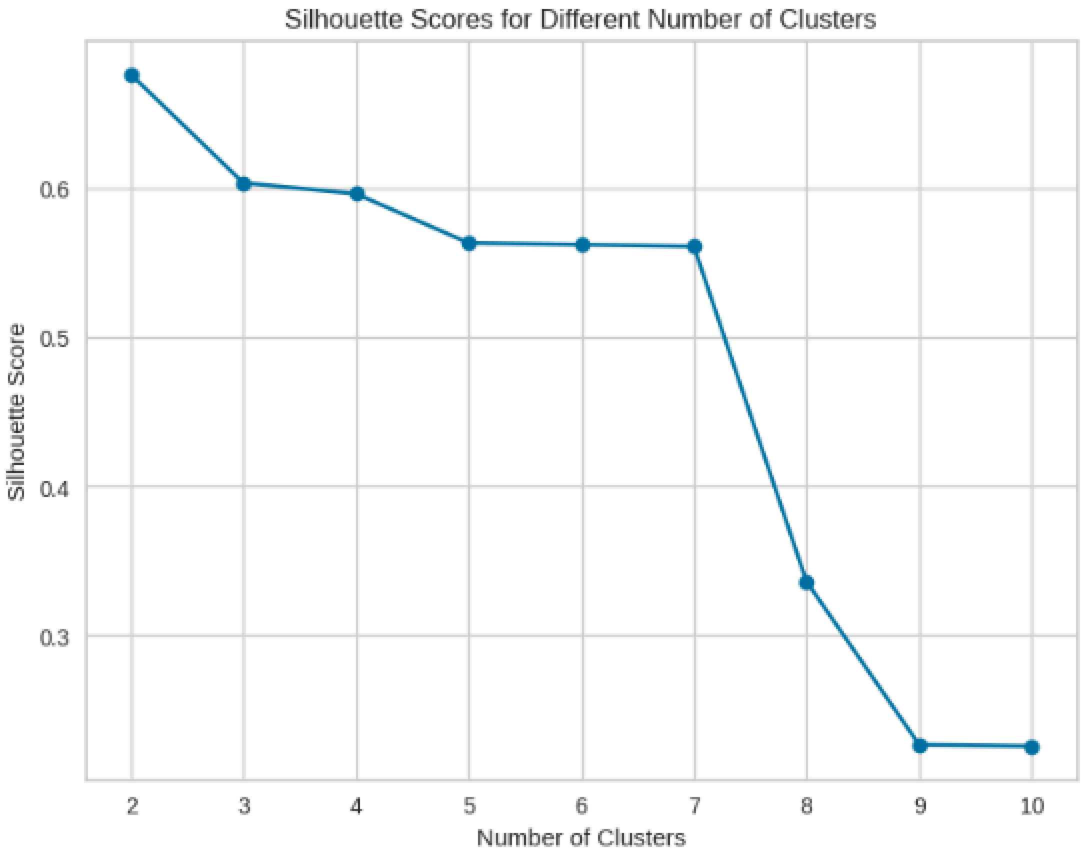
9.4)Figure out the appropriate number of clusters

9.4.1) Calculating Silhouette Scores for different number of clusters





```
Calculating Silhouette Scores for different number of clusters:
For n_clusters=2, Silhouette Score: 0.6768
For n_clusters=3, Silhouette Score: 0.6038
For n_clusters=4, Silhouette Score: 0.5965
For n_clusters=5, Silhouette Score: 0.5635
For n_clusters=6, Silhouette Score: 0.5623
For n_clusters=7, Silhouette Score: 0.5611
For n_clusters=8, Silhouette Score: 0.3362
For n_clusters=9, Silhouette Score: 0.2266
For n_clusters=10, Silhouette Score: 0.2256
```

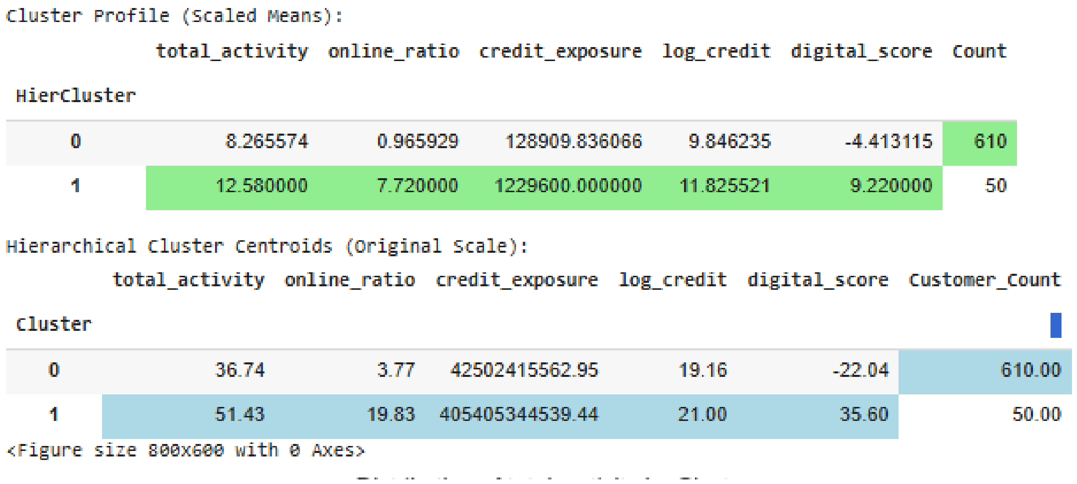


The dendrogram helps us see how to group customers. We look for big gaps in the tree, which suggest natural places to cut and form clusters.

To double-check, we also looked at Silhouette Scores. This score tells us how well each customer fits into their assigned cluster compared to others. A higher score is better.

Based on both the dendrogram and the silhouette scores, having 2 clusters seems to be the best way to group our customers with this method.

9.5) Cluster Profiling



Cluster Profile (Scaled Means):

HierCluster	total_activity	online_ratio	credit_exposure	log_credit	digital_score	Count
0	8.265574	0.965929	128909.836066	9.846235	-4.413115	610
1	12.580000	7.720000	1229600.000000	11.825521	9.220000	50

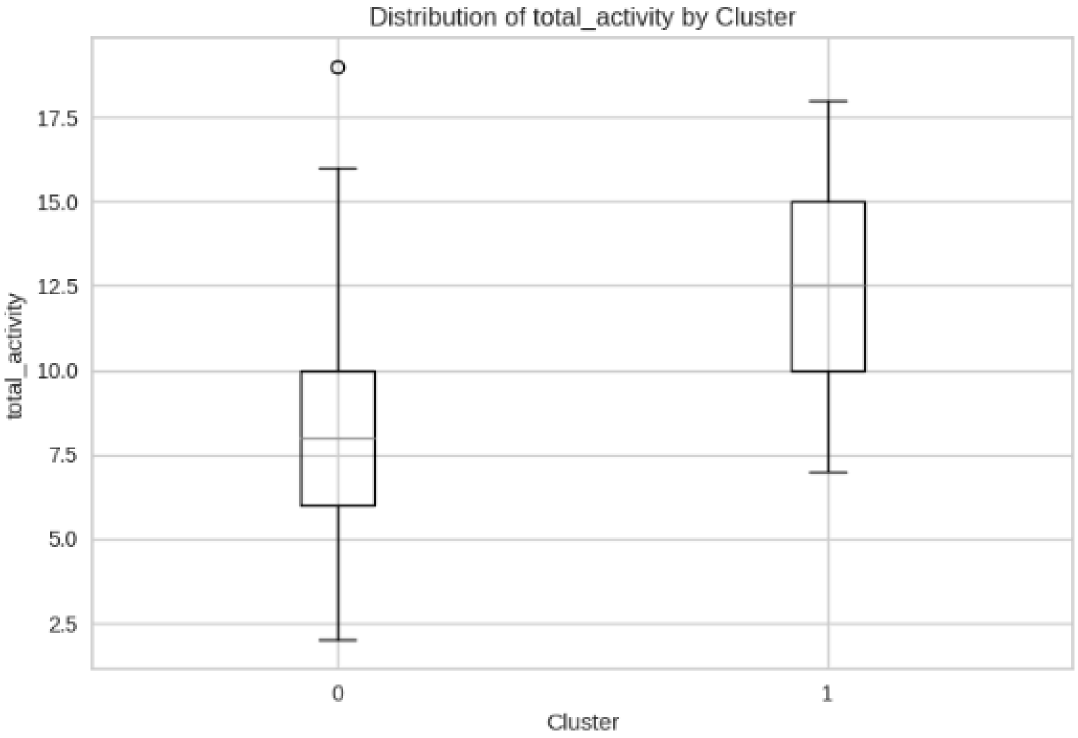
Hierarchical Cluster Centroids (Original Scale):

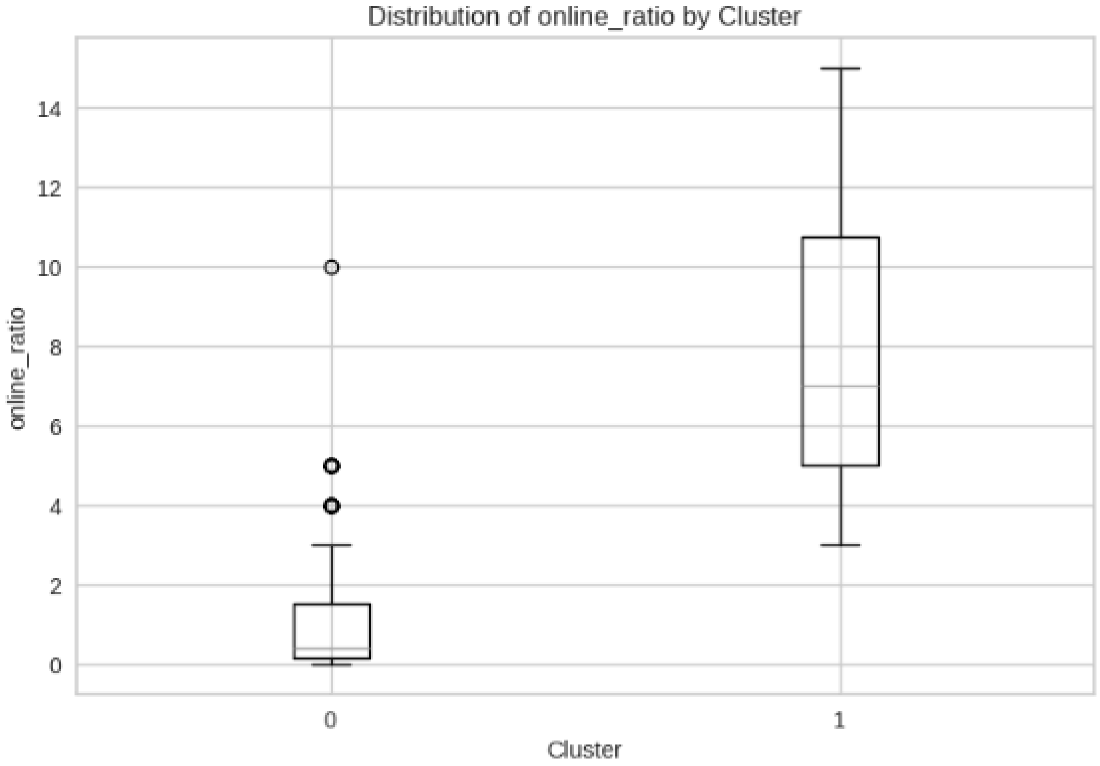
Cluster	total_activity	online_ratio	credit_exposure	log_credit	digital_score	Customer_Count
0	36.74	3.77	42502415562.95	19.16	-22.04	610.00
1	51.43	19.83	405405344539.44	21.00	35.60	50.00

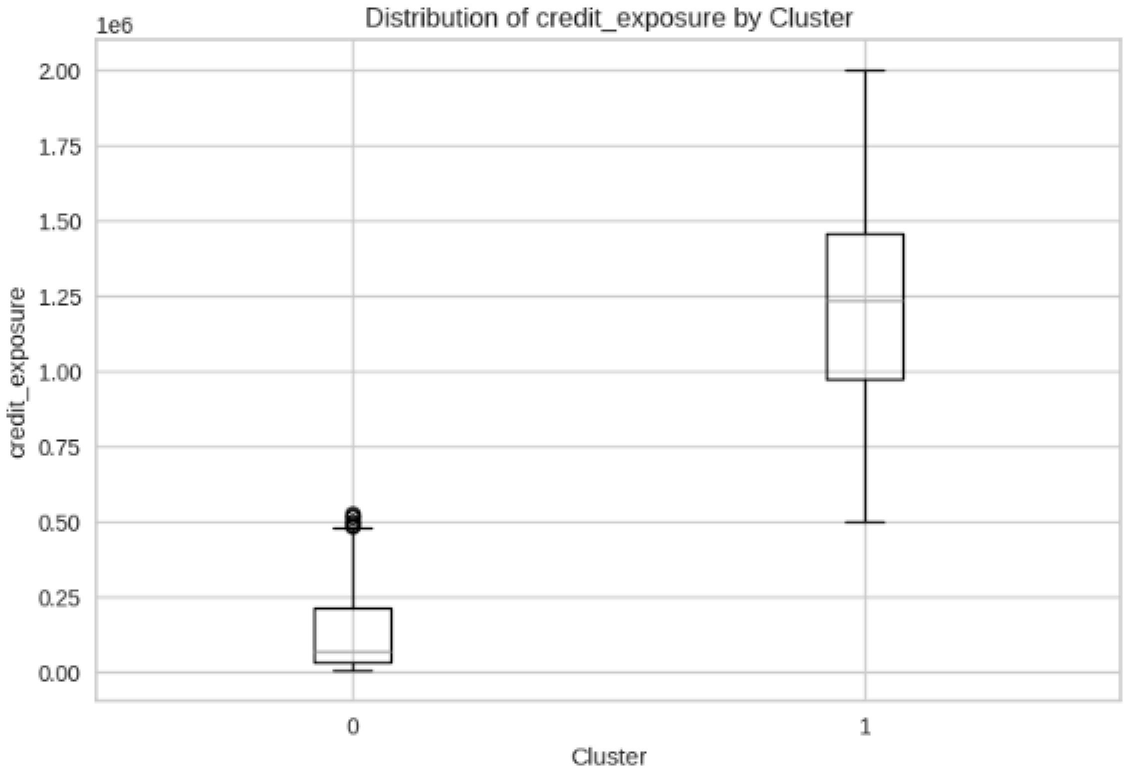
```
<Figure size 800x600 with 0 Axes>
```

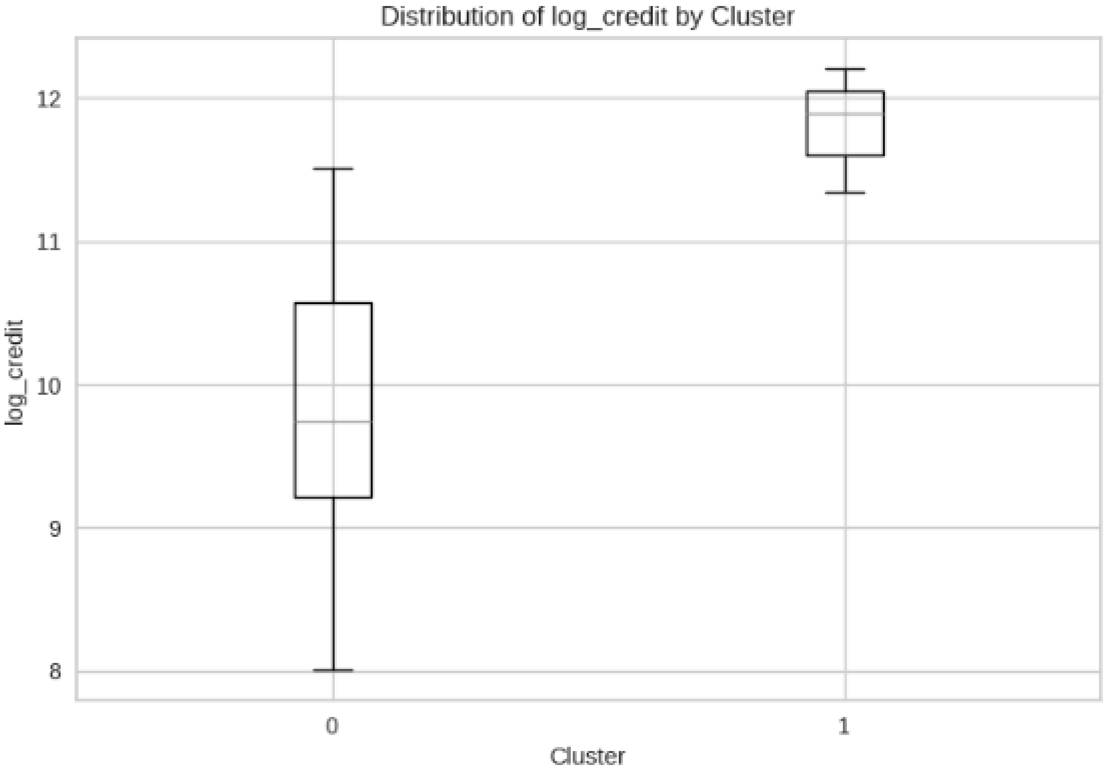
Cluster 0 (610 customers): Lower total activity, lower online ratio, smaller credit exposure, and negative digital score. This segment appears to consist of traditional or less digitally active customers.

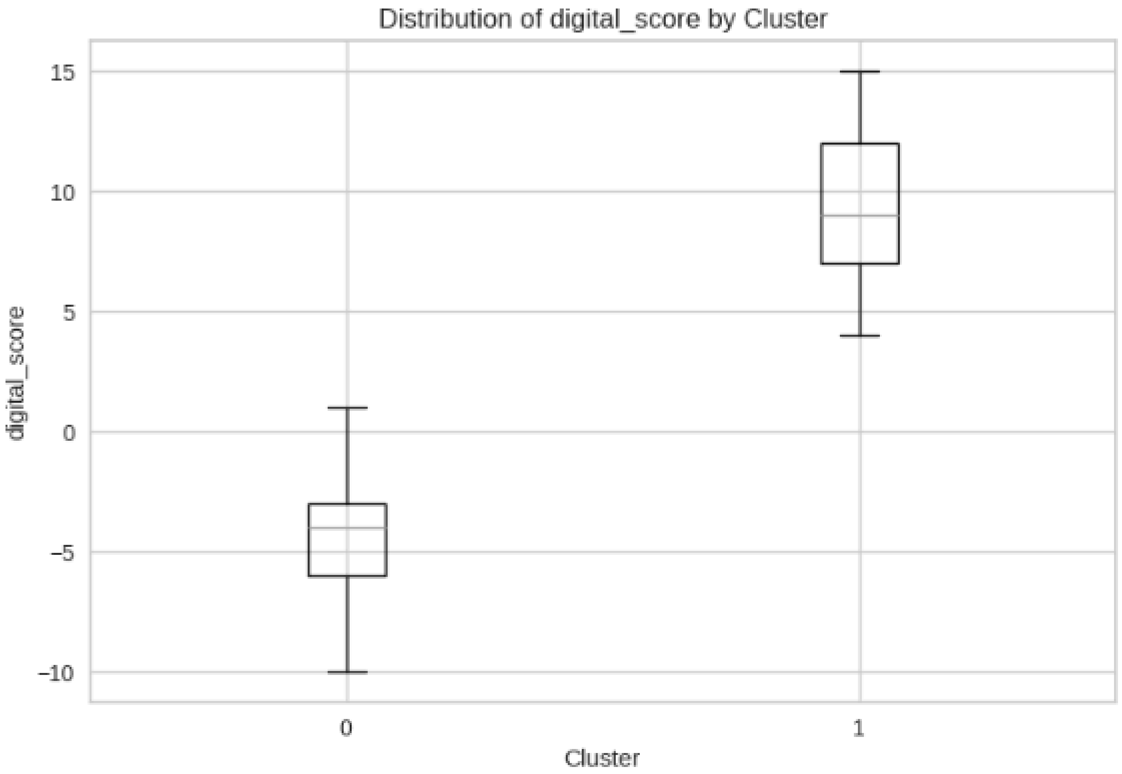
Cluster 1 (50 customers): Much higher activity, high online ratio, very large credit exposure, and strong positive digital score. This represents digitally active, high-value customers.


Distribution of total_activity by Cluster
total_activity
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0
1
Cluster


Distribution of online_ratio by Cluster
online_ratio
14
12
10
8
6
4
2
0
0
1
Cluster


Distribution of credit_exposure by Cluster
1e6
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
credit_exposure
0
1
Cluster


Distribution of log_credit by Cluster
12
11
10
9
8
log_credit
0
1
Cluster

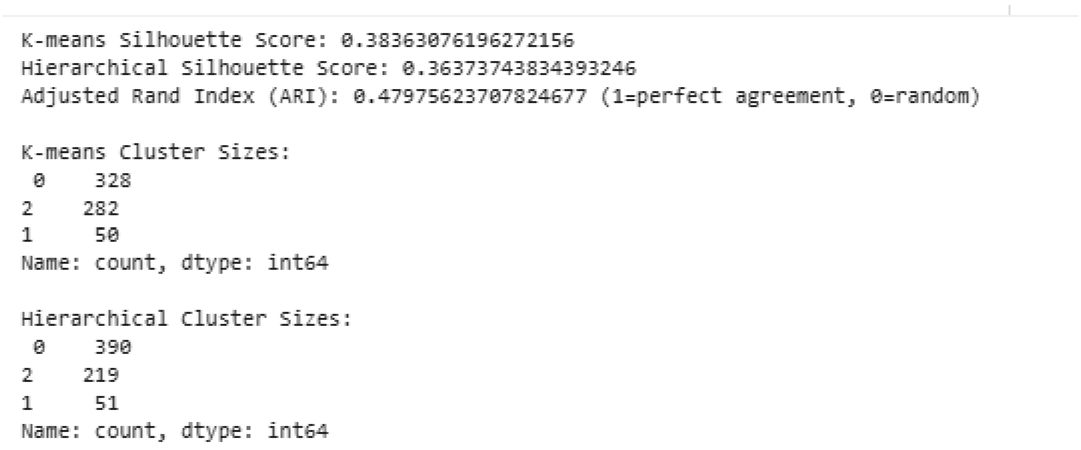


Cluster 0 : This group is much bigger. They have lower total activity, online ratio, credit exposure, log credit, and digital score compared to Cluster 1. They likely have lower credit limits and prefer less digital interaction.

Cluster 1 : This group is smaller but stands out with higher values across all the analyzed features (total activity, online ratio, credit exposure, log credit, and digital score). These are likely the customers with higher credit limits who are very active online.

10) K-means vs Hierarchical Clustering

10.1)Compare clusters obtained from K-means and Hierarchical clustering techniques



```
K-means Silhouette Score: 0.38363076196272156
Hierarchical Silhouette Score: 0.36373743834393246
Adjusted Rand Index (ARI): 0.47975623707824677 (1=perfect agreement, 0=random)

K-means Cluster Sizes:
 0    328
2    282
1     50
Name: count, dtype: int64

Hierarchical Cluster Sizes:
 0    390
2    219
1     51
Name: count, dtype: int64
```

Cluster Quality: K-means performed slightly better (Silhouette 0.384) than Hierarchical (0.364).

Cluster Sizes: Both methods found a small cluster (~50 customers) and two larger ones, with minor differences in counts.

Agreement: ARI score of 0.48 shows moderate overlap between the two methods.

Insight: K-means gave cleaner clusters, but both methods confirmed a similar segmentation pattern, adding confidence to the results.

11) Actionable Insights & Recommendations

Customer Segments (K-means, 3 Clusters)

Segment 0 – The Regulars (Largest group) Moderate credit, few cards, prefer branch visits.

Tip: Focus on branch perks and friendly service. Gently push digital adoption with demos or small rewards.

Segment 1 – The High Rollers (Smallest, high value) Very high credit, many cards, heavy online users.
Tip: Offer premium digital perks, smooth app/website experience, and exclusive wealth products.

Segment 2 – The Callers Lower credit, fewer cards, rely on phone support.
Tip: Strengthen call centers, ensure fast resolutions, and guide them toward simple online tools.

Conclusion:
Tailor outreach by segment—branch for Regulars, digital for High Rollers, and phone support for Callers. Push digital adoption gradually while respecting customer preferences.