Unsupervised Learning Report

Presented by :
Sanjay Rajan J

TABLE OF CONTENTS

CHAPTER NO	CONTENT	PAGE NO
	LIST OF FIGURES	3
1.	DATA OVERVIEW	5
2.	EXPLORATORY DATA ANALYSIS	8
	2.1 Univariate Analysis	8
	2.2 Bivariate Analysis	13
3.	DATA PREPROCESSING	17
4.	K-MEANS CLUSTERING	19
5.	HIERARCHICAL CLUSTERING	22
6.	CLUSTER COMPARISON	25
7.	ACTIONABLE INSIGHTS & RECOMMENDATIONS	26

LIST OF FIGURES

FIG NO.	NAME	PAGE
1.	Data Info	6
2.	Null values check	6
3.	Numerical Statistics	7
4.	First 10 rows	7
5.	Avg_Credit_Limit distribution	8
6.	Total_Credit_Cards distribution	8
7.	Total_visits_bank distribution	9
8.	Total_visits_online distribution	9
9.	Total_calls_made distribution	10
10.	Total_visits_bank distribution 2	10
11.	Total_Credit_cards distribution 2	11
12.	Total_calls_made distribution 2	11
13.	Total_visits_online distribution 2	12
14.	Pairplot	13
15.	Heatmap	14
16.	Avg_credit_limit vs Total_calls_made	14
17.	Avg_credit_limit vs Total_visits_online	15
18.	Avg_credit_limit vs Total_Credit_cards	15
19.	Avg_credit_limit vs Total_visits_bank	16
20.	Null value checks	17
21.	Outlier checks	17
22.	No. of Clusters vs Avg. Distortion	19
23.	Silhouette Scores	20
24.	Silhouette Plot	20
25.	KM Cluster Profile	21

26.	KM Cluster Distribution	21
27.	Cophenetic correlations for all Combinations	22
28.	Cophenetic correlations for diff linkage	22
29.	Dendrograms	23
30.	HC Cluster Profile	24
31.	HC Cluster Distribution	24
32.	K-means Cluster Profile	25
33.	Hierarchical Cluster Profile	25

1.DATA OVERVIEW

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 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)

➢ Shape:

There are 660 rows and 7 columns in this dataset.

➢ Duplicates:

There are no duplicate entries in this dataset.

➢ Basic Info:

Fig.1 Data Info

```
<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
```

➢ Null values Check:

Fig.2 Null values check

```
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
```

- Numerical Statistics:

Fig.3 Numerical Statistics

	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

- Data Sample:

Fig.4 First 10 rows

	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	1	87073	100000	2	1	1	0
1	2	38414	50000	3	0	10	9
2	3	17341	50000	7	1	3	4
3	4	40496	30000	5	1	1	4
4	5	47437	100000	6	0	12	3
5	6	58634	20000	3	0	1	8
6	7	48370	100000	5	0	11	2
7	8	37376	15000	3	0	1	1
8	9	82490	5000	2	0	2	2
9	10	44770	3000	4	0	1	7

2. EXPLORATORY DATA ANALYSIS

2.1 UNIVARIATE ANALYSIS:

Fig.5 Avg_Credit_Limit distribution

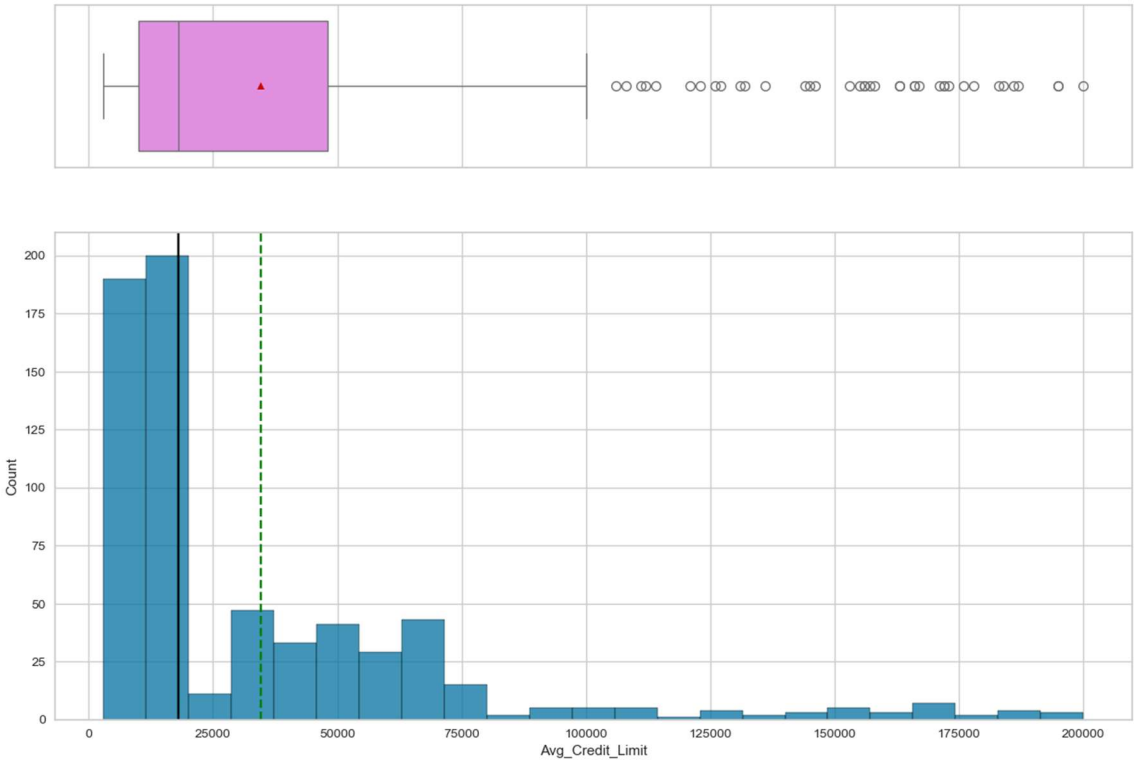


Fig.6 Total_Credit_Cards distribution

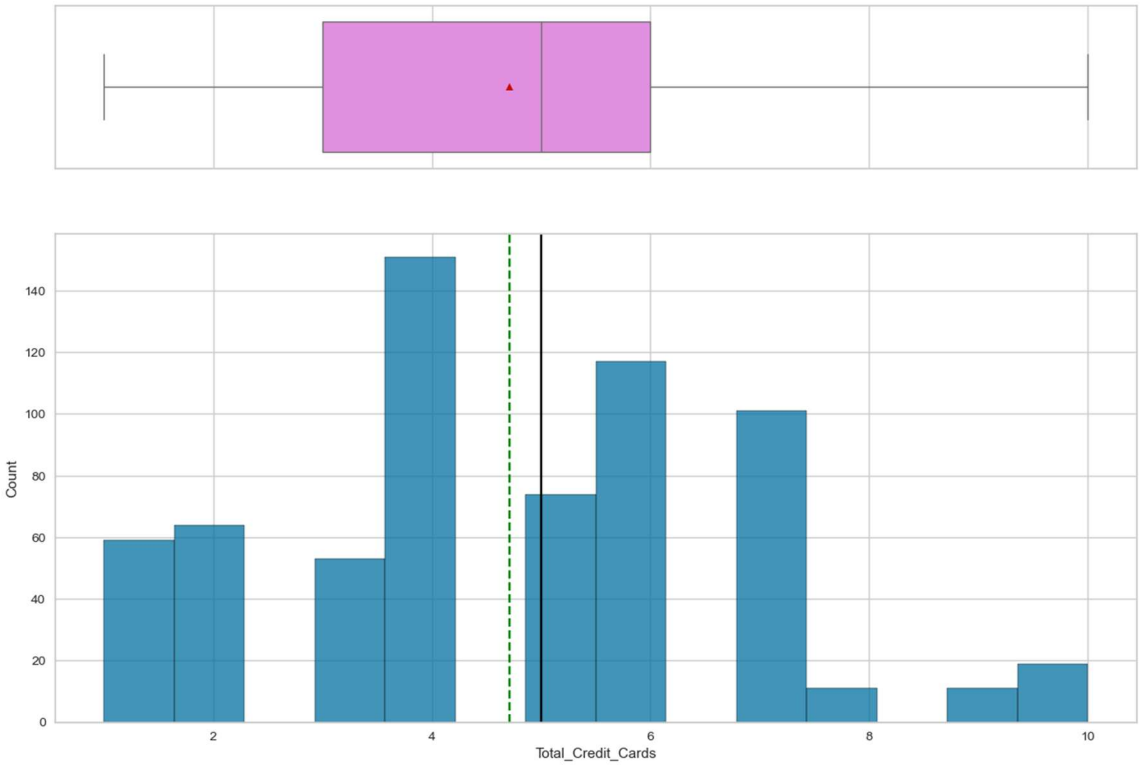


Fig.7 Total_visits_bank distribution

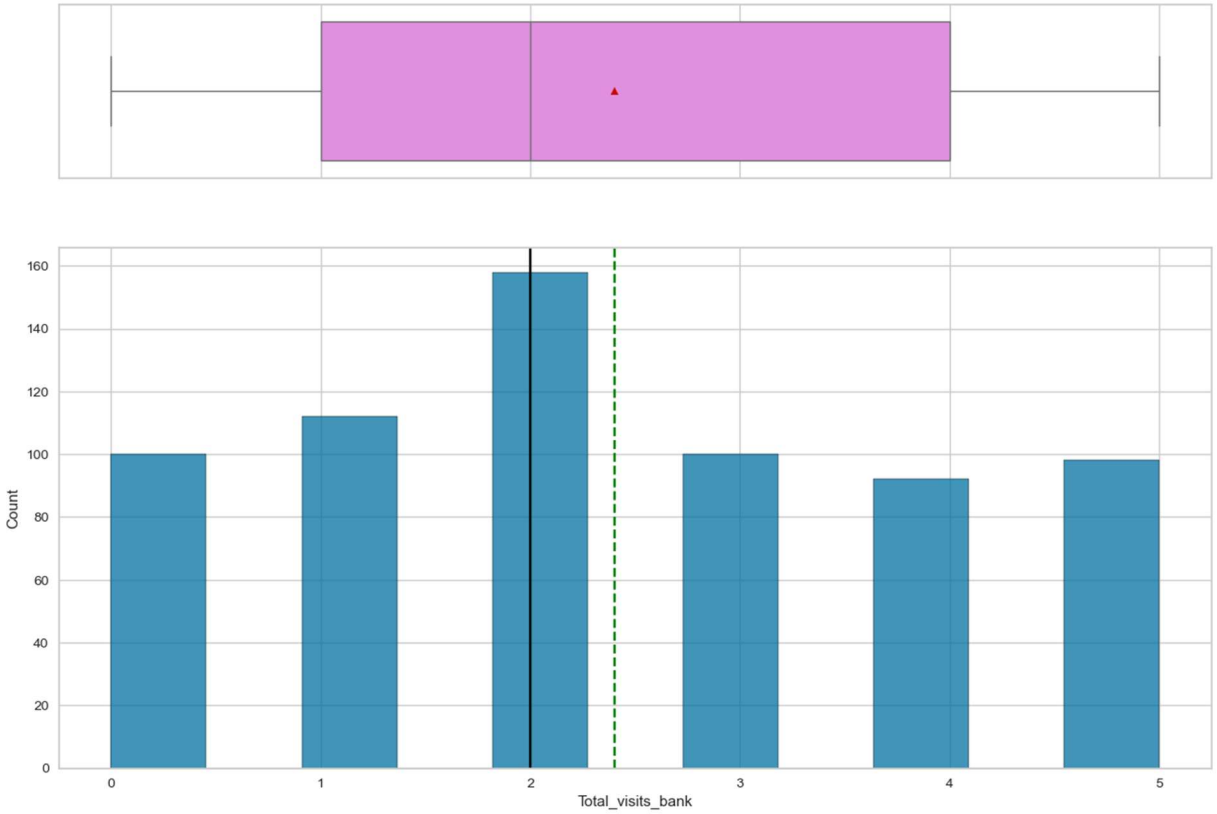


Fig.8 Total_visits_online distribution

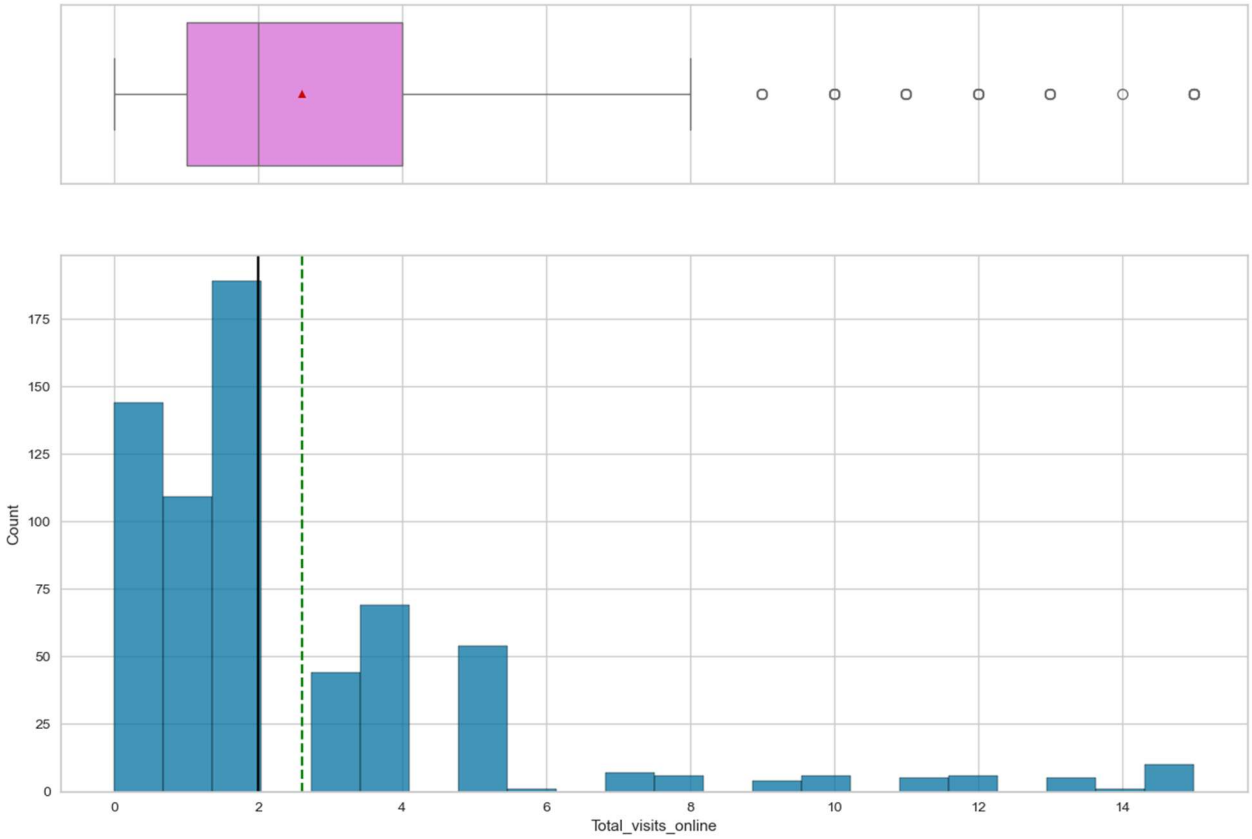


Fig.9 Total_calls_made distribution

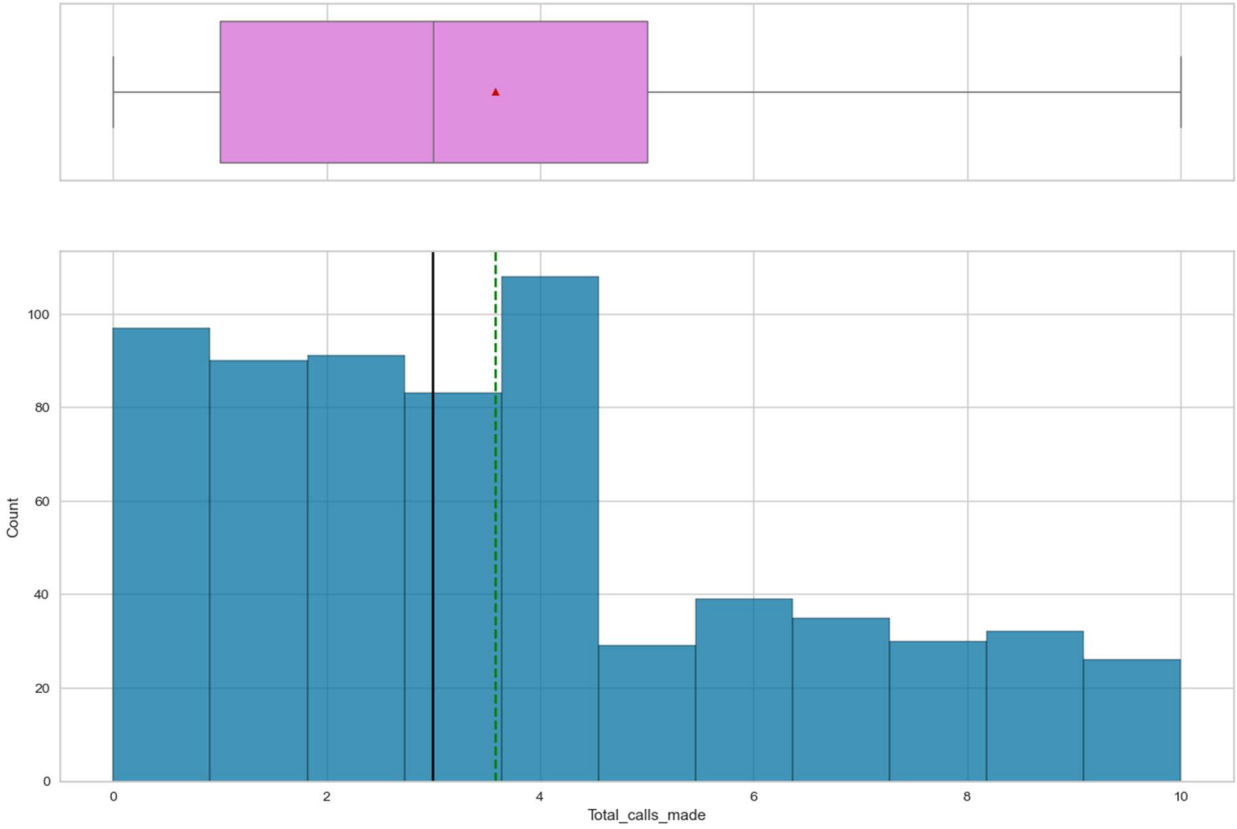


Fig.10 Total_visits_bank distribution 2

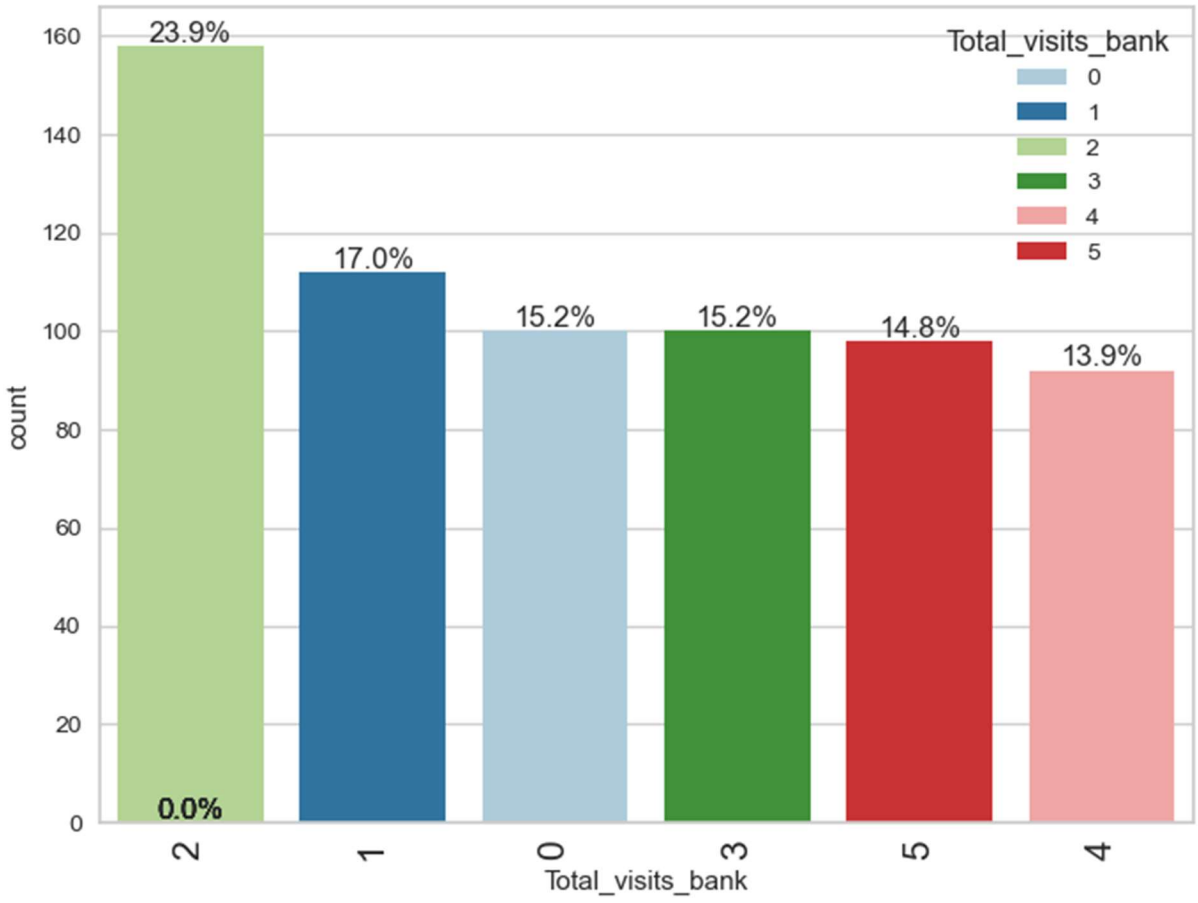


Fig.11 Total_Credit_cards distribution 2

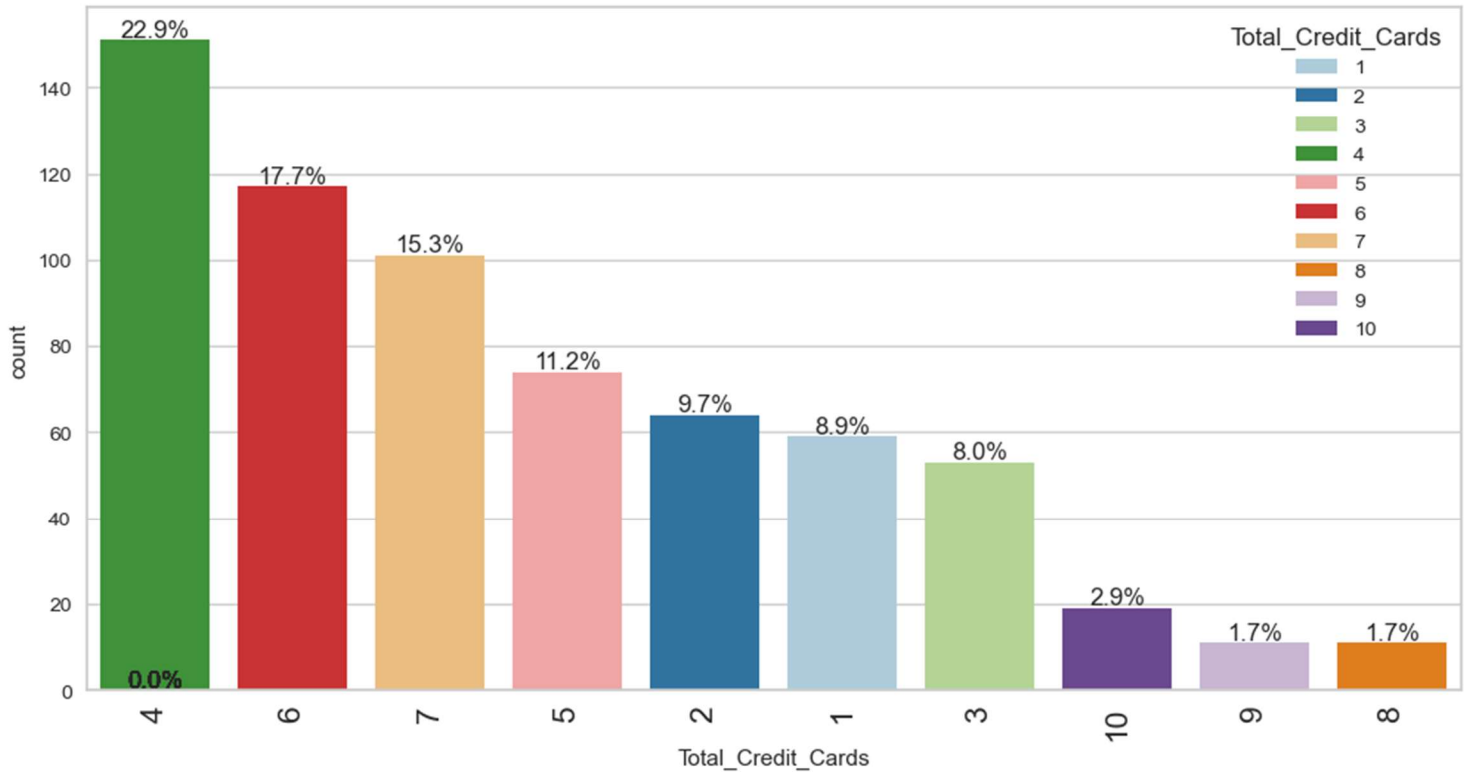


Fig.12 Total_calls_made distribution 2

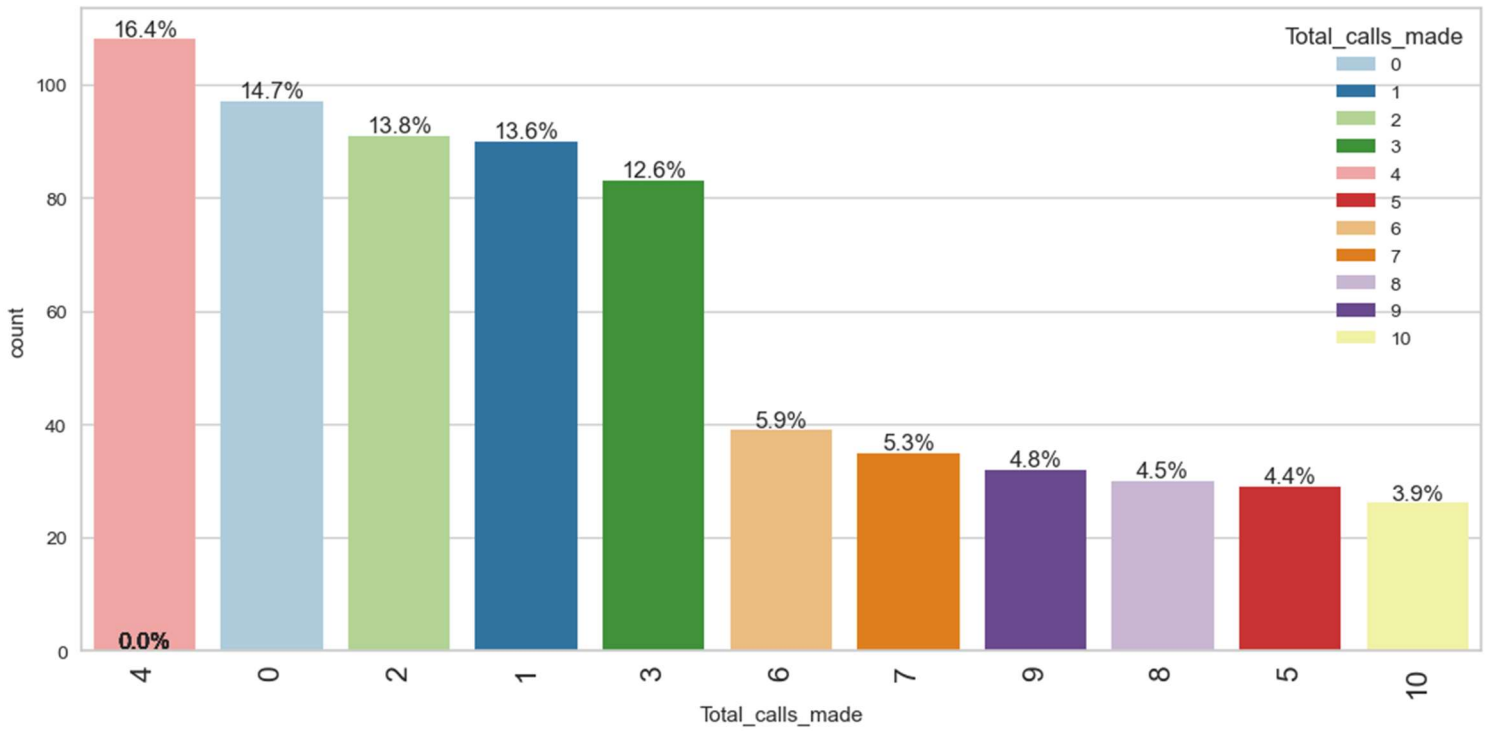
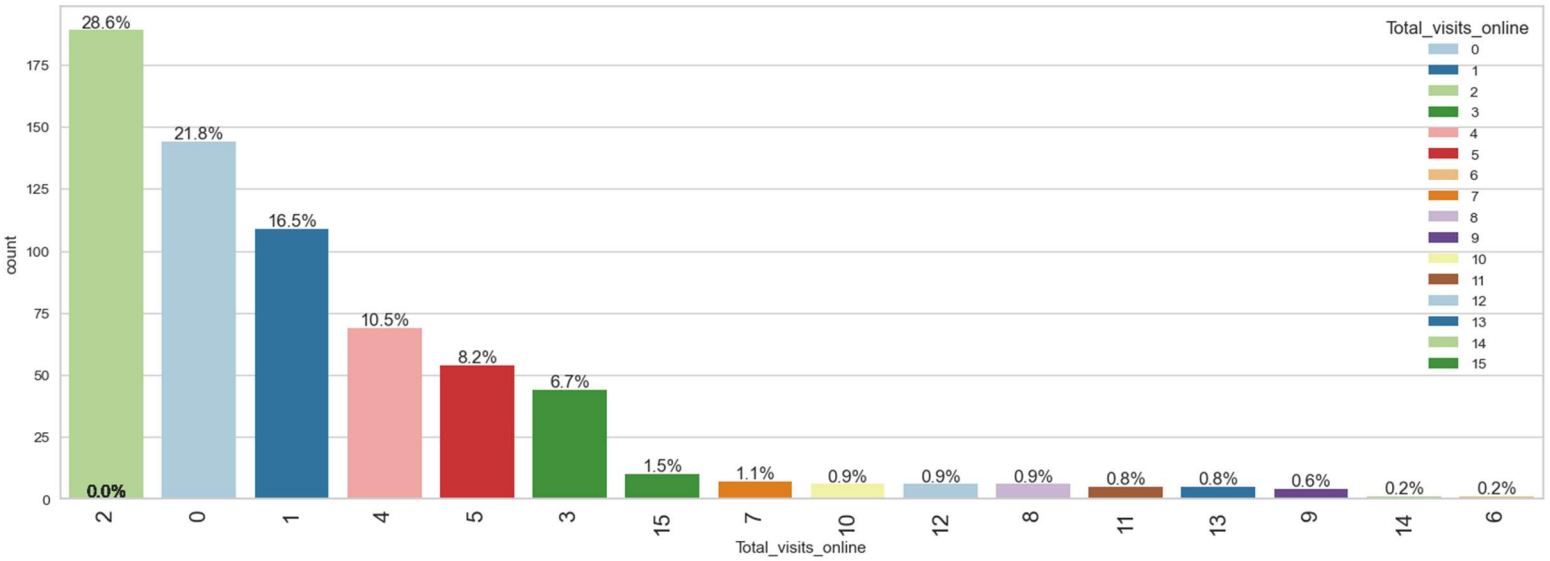


Fig.13 Total_visits_online distribution 2



COMMENTS:

- The distribution of the Avg_credit_limit is heavily right-skewed indicating a very few customers having higher average credit limit.
- The distribution of the Total_visits_online is heavily right-skewed indicating that the most common total visits are 0 to 2, with a significant proportion of customers making these few visits.
- Majority of the customers hold 4 credit cards (22.9%) followed by the count of 6 (17.7%).
- Majority of the customers have 2 bank visits per year, accounting for 23.9% of the total.
- The highest proportion of customers made 4 calls to the bank or its customer service department per year (16.4%).
- Only 3.9% of the customers made 10 calls in a year.
- Majority of the customers (28.6%) made 2 online visits in a year.

2.2 BIVARIATE ANALYSIS:

Fig.14 Pairplot

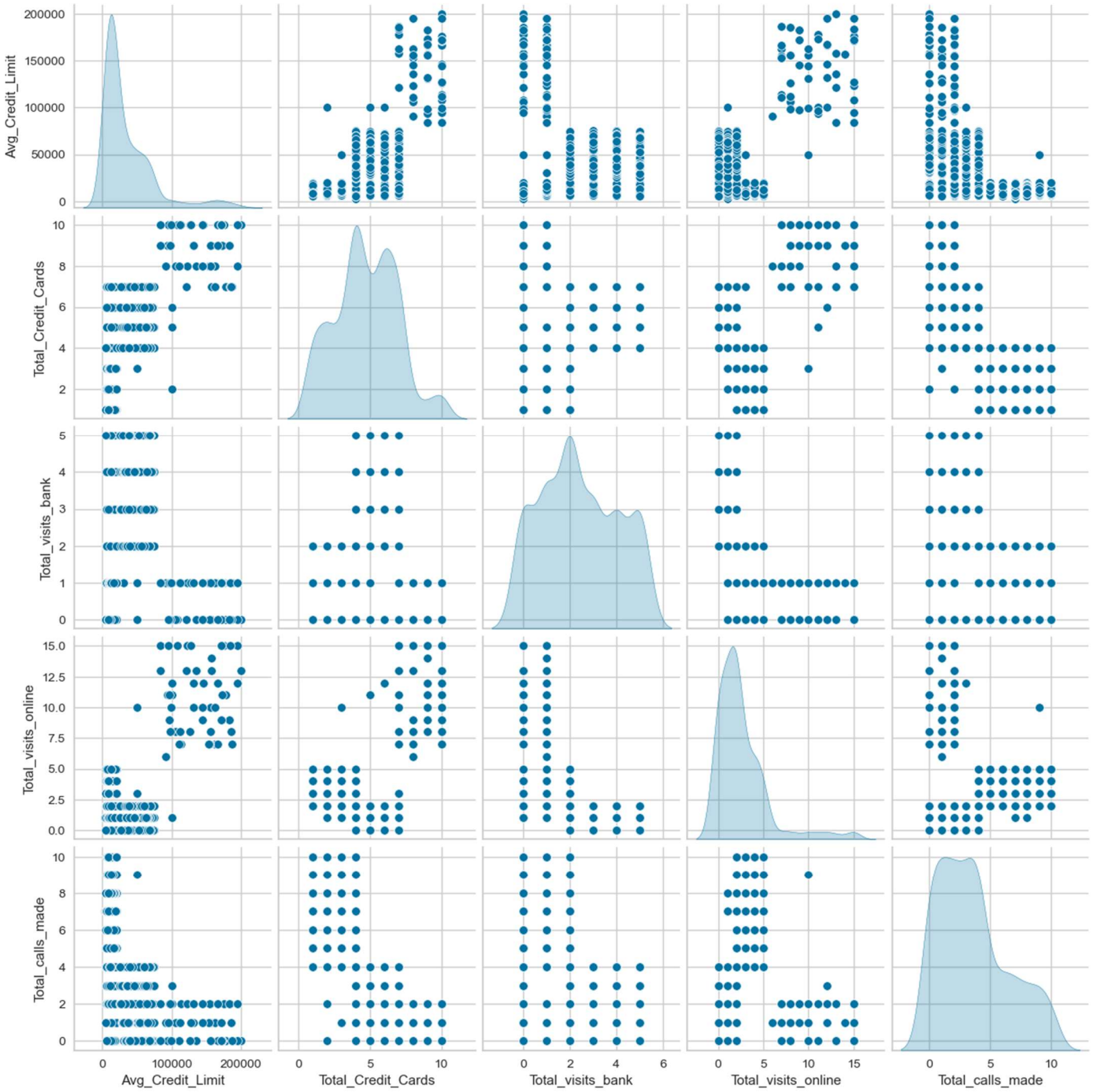


Fig.15 Heatmap

Fig.16 Avg_credit_limit vs Total_calls_made

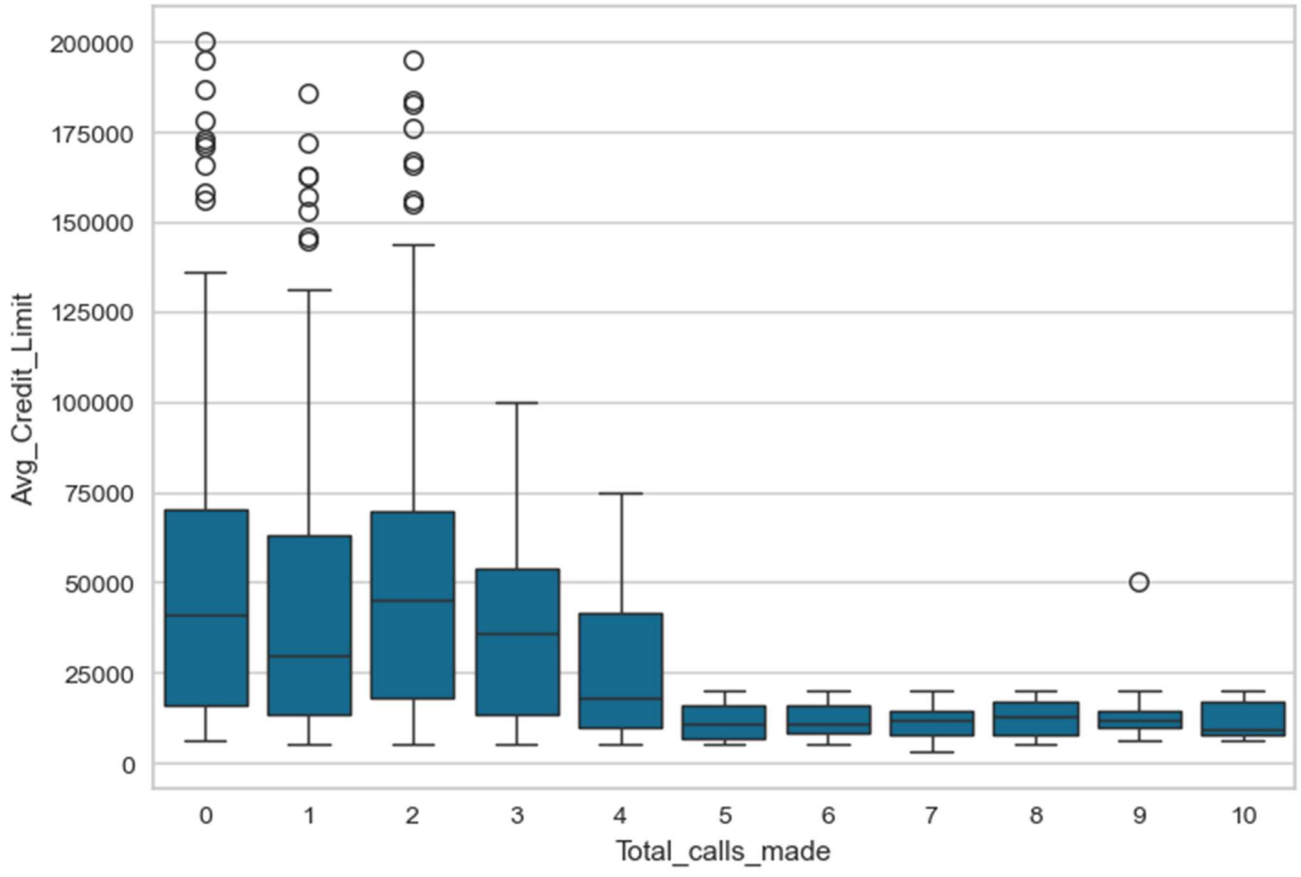


Fig.17 Avg_credit_limit vs Total_visits_online

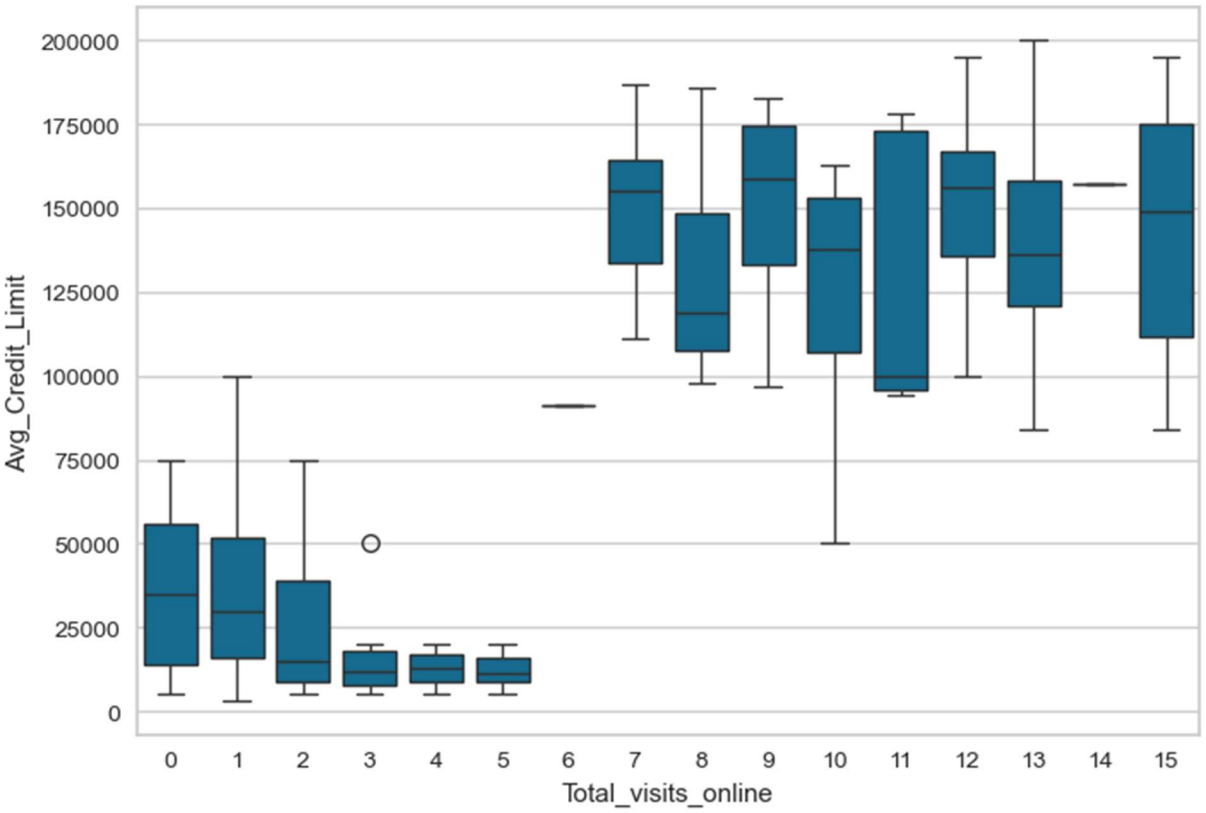


Fig.18 Avg_credit_limit vs Total_Credit_cards

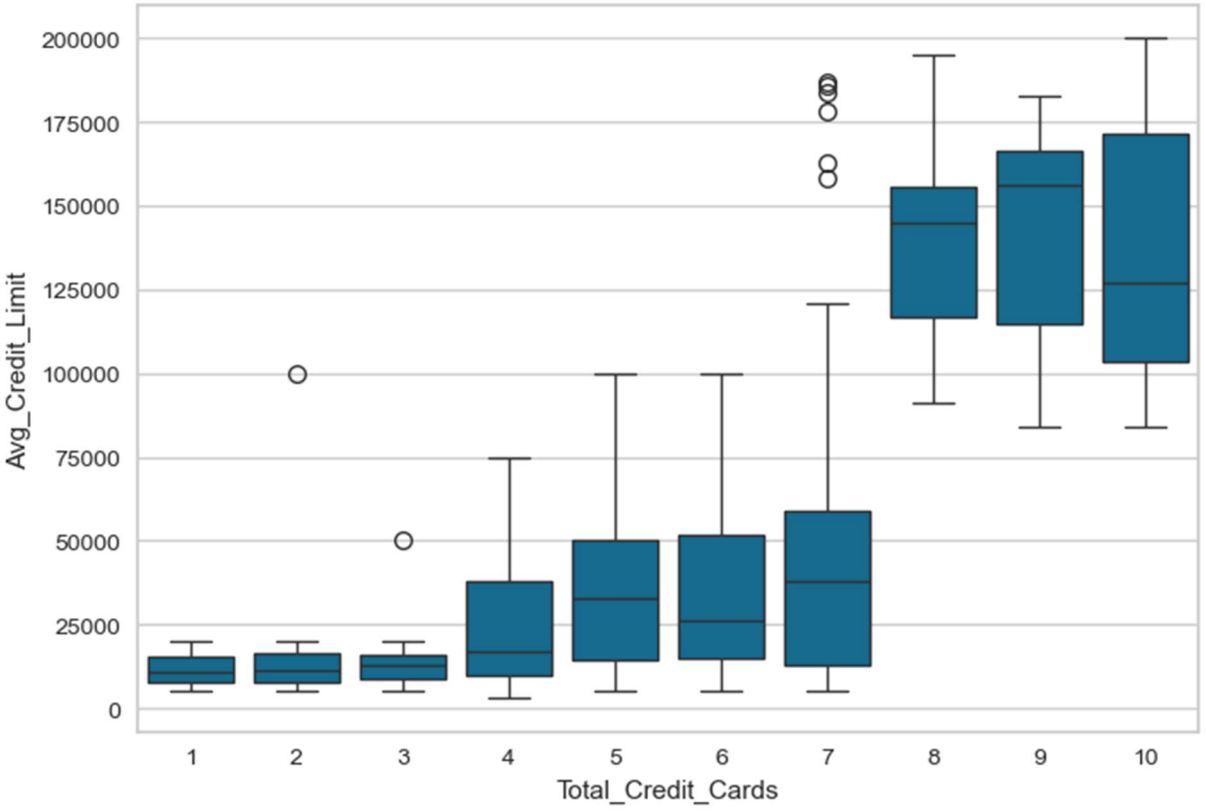
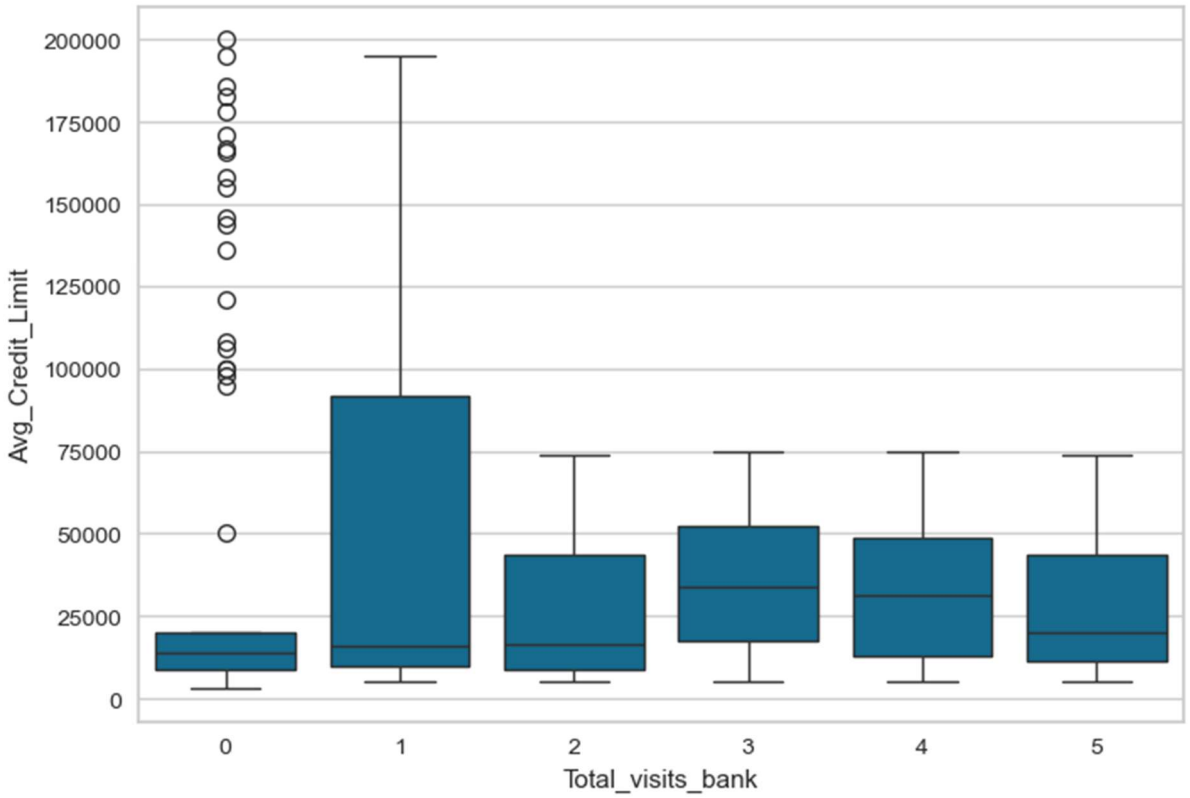


Fig.19 Avg_credit_limit vs Total_visits_bank



COMMENTS:

- A moderate positive correlation (0.61) indicates that the customers with a higher avg. credit limit tend to have more credit cards.
- A weak negative correlation (-0.41) indicates that customers with higher average credit limits make fewer calls to the company
- A strong negative correlation (-0.65) suggests that customers having more credit cards are less likely to contact customer service department or bank.
- Customers with higher average credit limits are most likely to be contact online and less likely to make calls.
- Customers with less online visits have lower credit limits.
- As the number of calls increases beyond 4, the average credit limit decreases, stabilizing at lower values with less variation which indicates that customers having high credit limit may rely less on calls.

3. DATA PREPROCESSING

Missing Value Treatment:

Fig.20 Null value checks

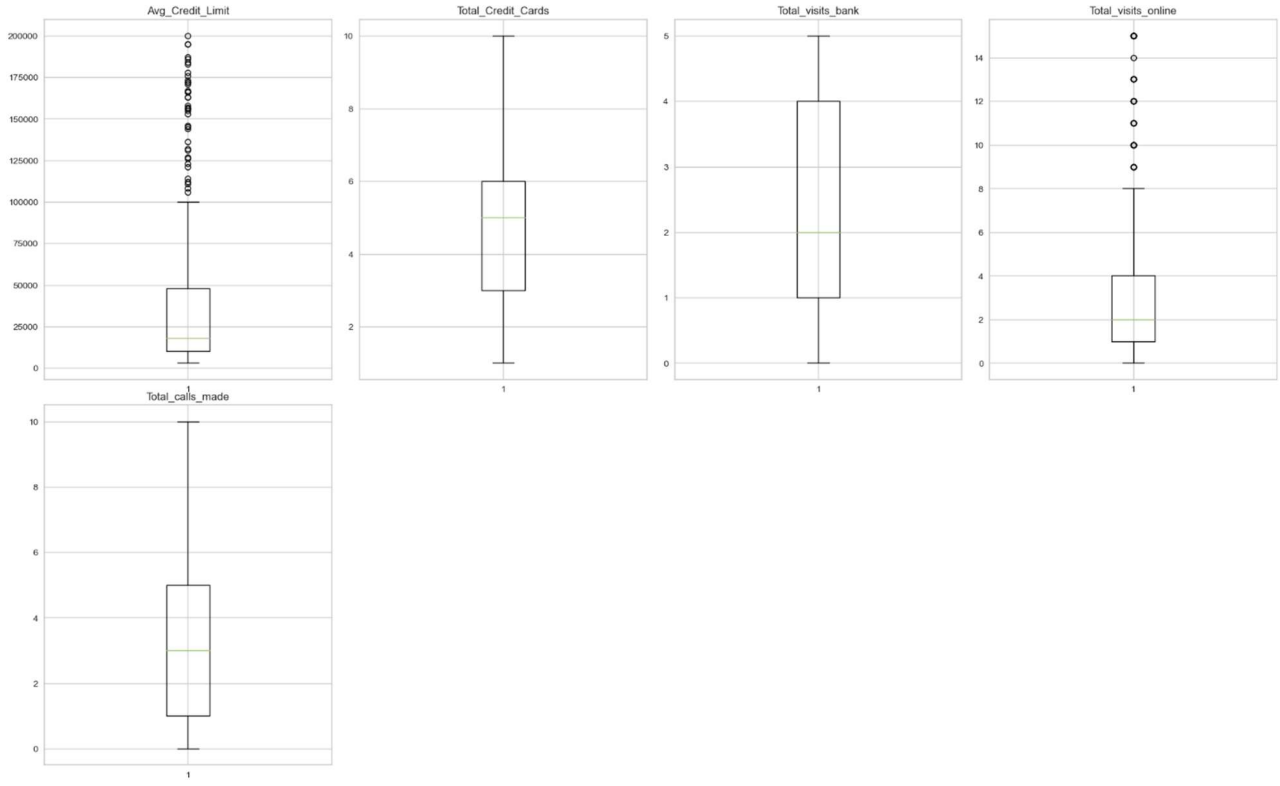
```
Sl_No                  0
Customer Key           0
Avg_Credit_Limit       0
Total_Credit_Cards     0
Total_visits_bank      0
Total_visits_online    0
Total_calls_made       0
dtype: int64
```

There are no null values or missing values in this dataset.

Outlier Treatment:

- As we can clearly see, there are many outliers in the "Avg_Credit_Limit" and "Total_visits_online" columns.
- These outliers are meaningful and are not treated.

Fig.21 Outlier checks



Duplicates Check:

There are no duplicates in this dataset.

Feature Engineering:

- The columns “Sl_No” and “Customer Key” are dropped, as they are not useful in analysis.
- The dataset is duplicated and are scaled for both K-means and Hierarchical clustering methods.

4. K-MEANS CLUSTERING

- For K-Means Clustering method, “Euclidean” is chosen as the distance metric.
- After trying different number of clusters, k value is selected as 3 by using Elbow method.
- From the Silhouette score, 3 seems to be a good value of k.

Fig.22 No. of Clusters vs Avg. Distortion

```
Number of Clusters: 1   Average Distortion: 2.2337229406380987
Number of Clusters: 2   Average Distortion: 1.4670657209967661
Number of Clusters: 3   Average Distortion: 1.146701238427549
Number of Clusters: 4   Average Distortion: 1.0468854739869728
Number of Clusters: 5   Average Distortion: 0.9988946856686043
Number of Clusters: 6   Average Distortion: 0.9678831299202312
Number of Clusters: 7   Average Distortion: 0.9202339383367122
Number of Clusters: 8   Average Distortion: 0.9027045698067332
```

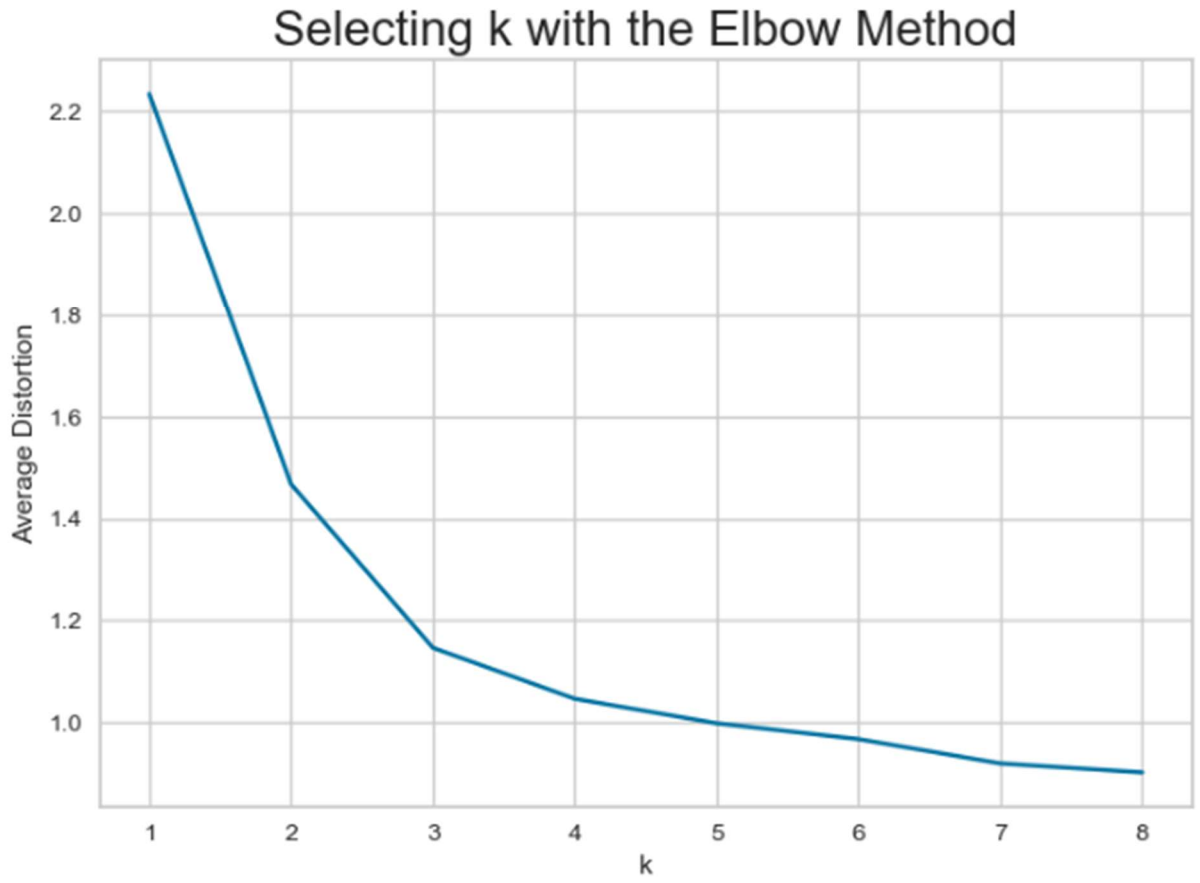


Fig.23 Silhouette Scores

```
For n_clusters = 2, silhouette score is 0.4975887345071433
For n_clusters = 3, silhouette score is 0.590990473596407
For n_clusters = 4, silhouette score is 0.38895131642270275
For n_clusters = 5, silhouette score is 0.35394953628284137
For n_clusters = 6, silhouette score is 0.25116177153756997
For n_clusters = 7, silhouette score is 0.25329681108745083
For n_clusters = 8, silhouette score is 0.2284146552864042
For n_clusters = 9, silhouette score is 0.22067565428657318
```

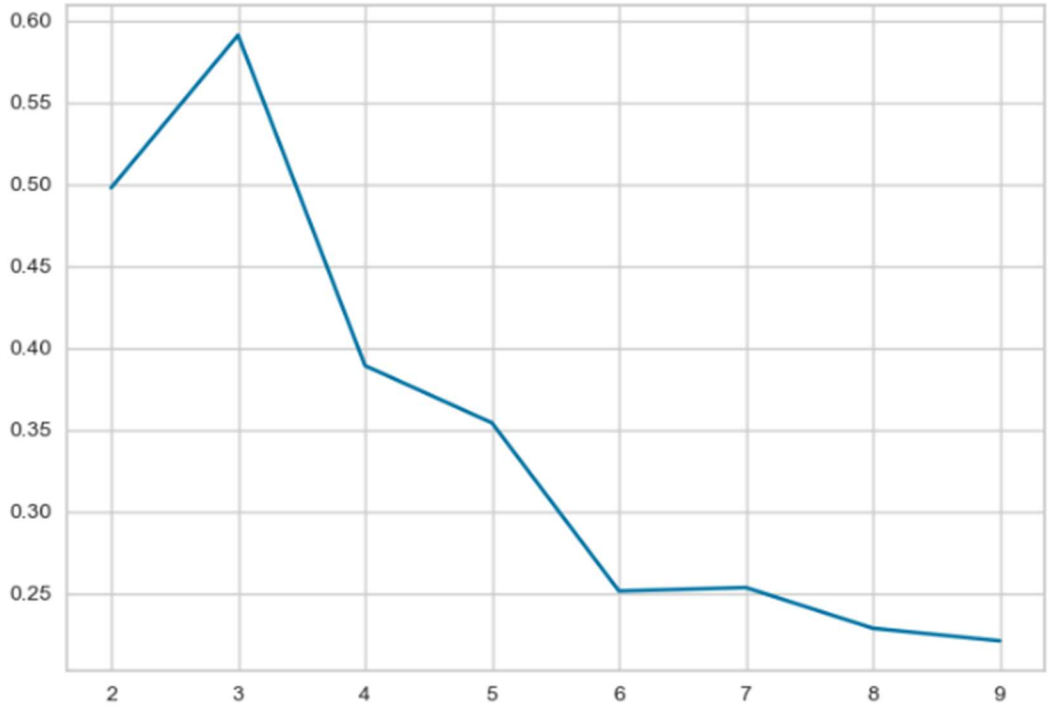


Fig.24 Silhouette Plot

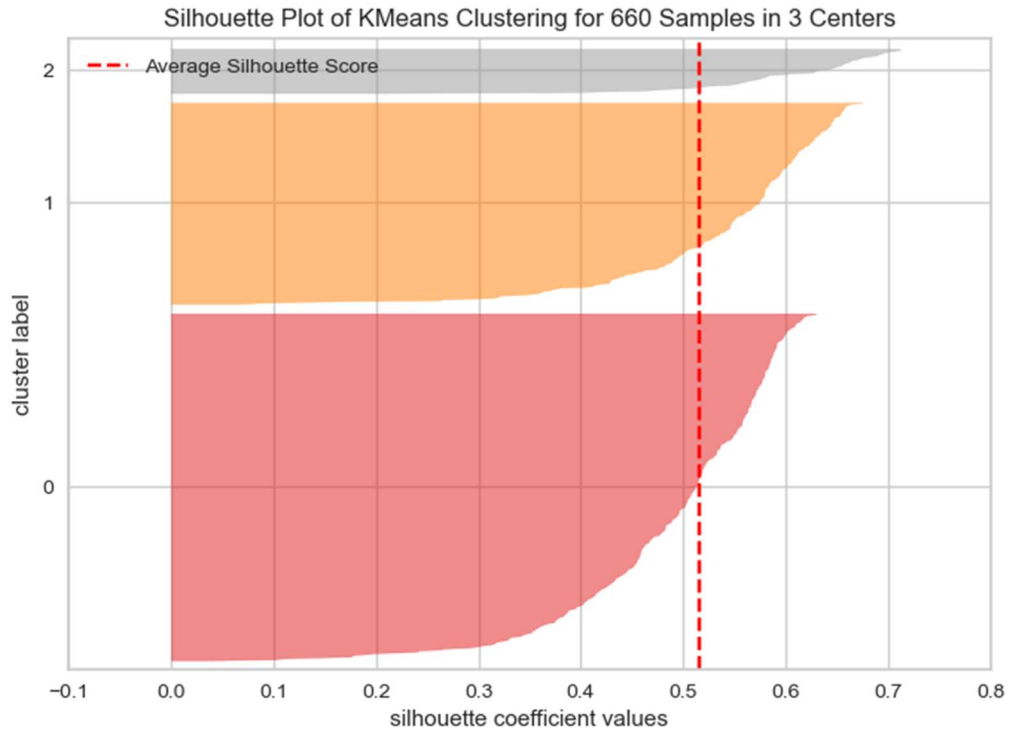
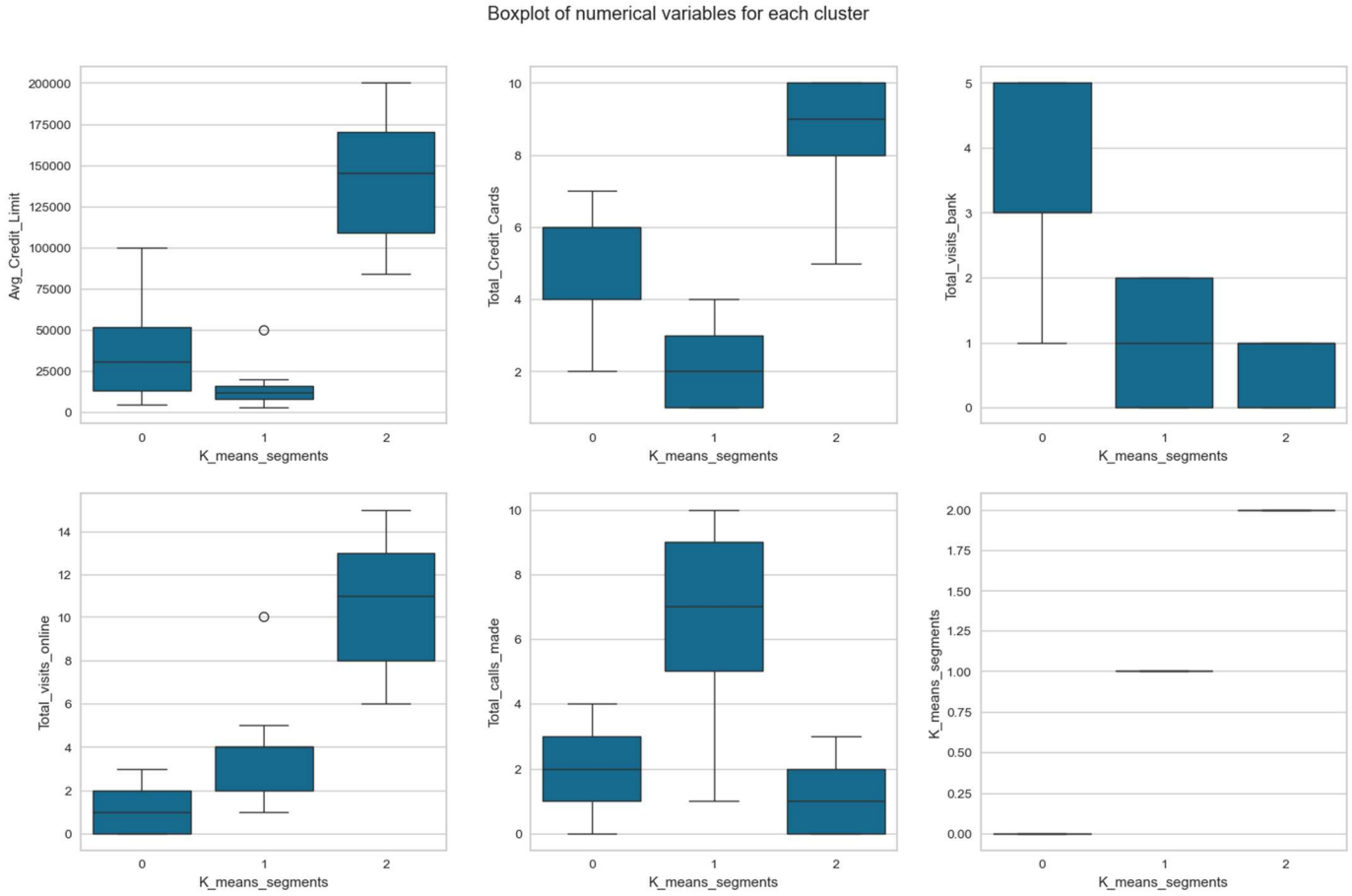


Fig.25 KM Cluster Profile

K_means_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_of_customers
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

Fig.26 KM Cluster Distribution



5. HIERARCHICAL CLUSTERING

- For Hierarchical clustering, various distance metrics and linkage methods are used.
- By checking the cophenetic correlation for all combinations, the highest cophenetic correlation is 0.8977, which is obtained with Euclidean distance and average linkage.
- 3 appears to be the appropriate number of clusters from the dendrogram.
- The Silhouette score for Agglomerative Clustering of n_clusters = 3 is 0.5159.

Fig.27 Cophenetic correlations for all Combinations

```
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 0.5422791209801746.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.8326994115042136.
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.
```

Fig.28 Cophenetic correlations for diff linkage

```
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.
```

Fig.29 Dendrograms

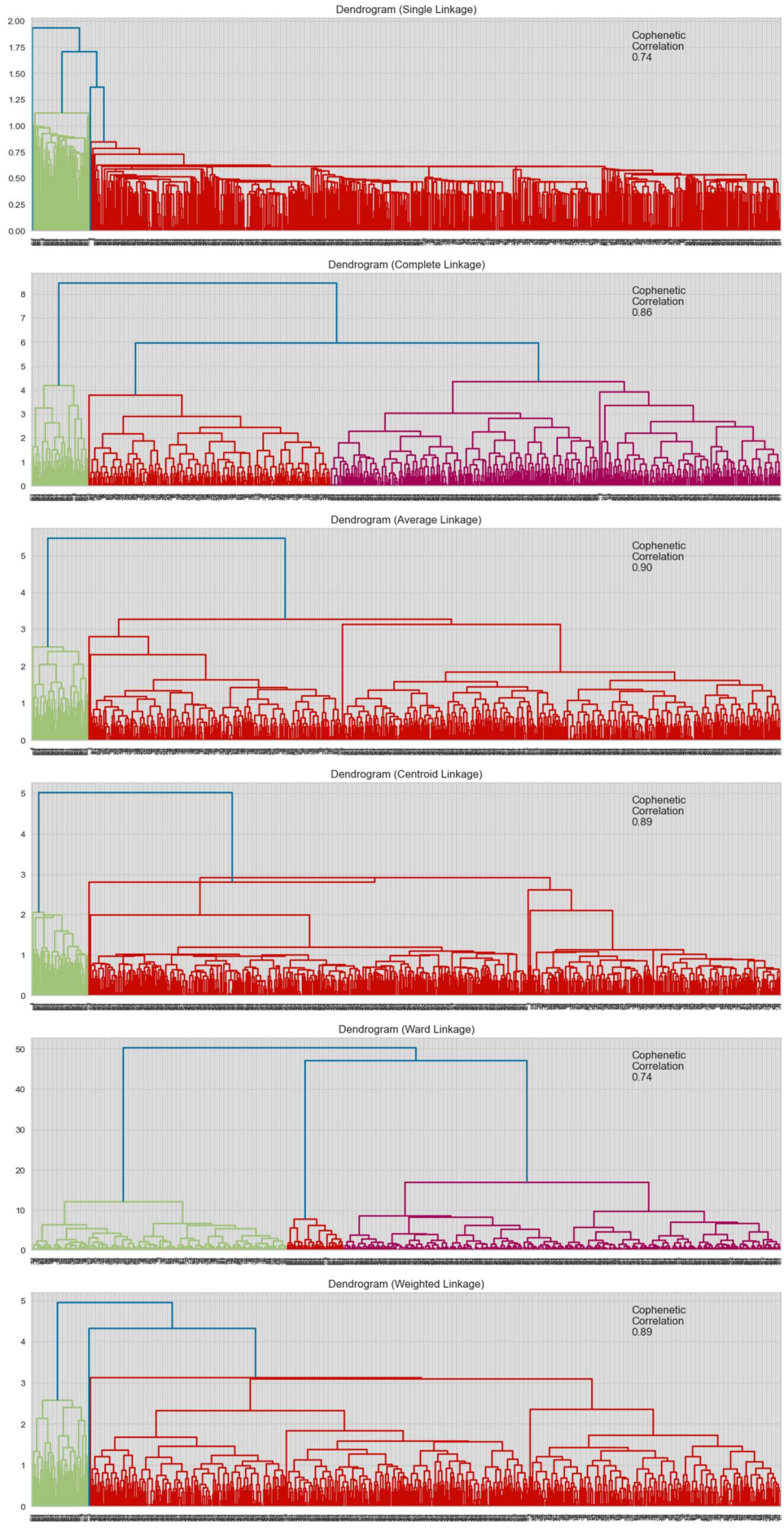
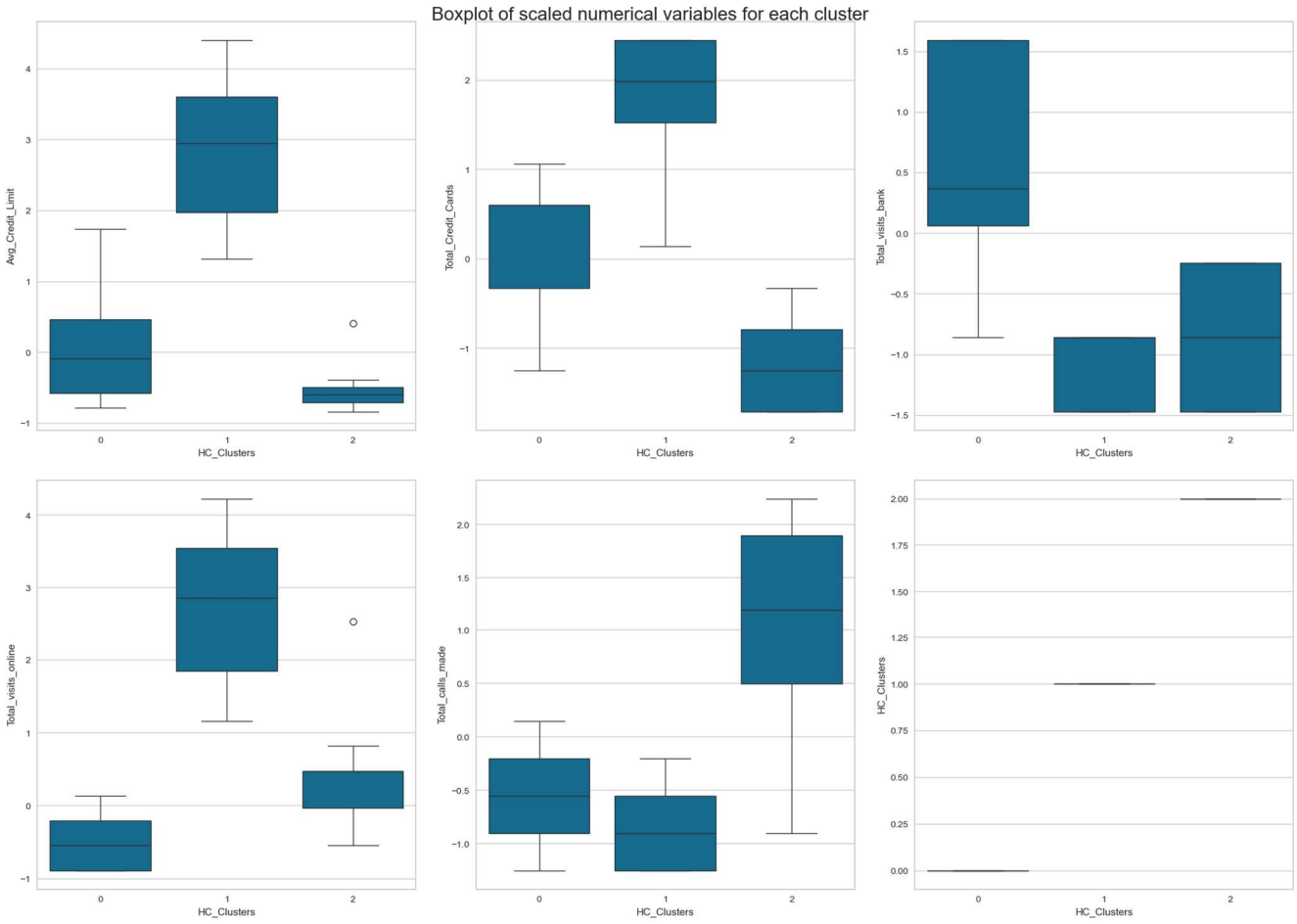


Fig.30 HC Cluster Profile

HC_Clusters	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segments
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

Fig.31 HC Cluster Distribution



6. CLUSTER COMPARISON

Silhouette Scores:

- K-means Clustering = 0.5157182558881063
- Hierarchical Clustering = 0.515922432650965

Cluster Profiles:

Fig.32 K-means Cluster Profile:

K_means_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_of_customers
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

Fig.33 Hierarchical Cluster Profile:

HC_Clusters	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segments
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

COMMENTS:

- Both clustering methods perform nearly identical by identifying **three** distinct customer segments with approximately same count.
- With respect to Silhouette scores, the **Hierarchical clustering method** slightly outperforms the K-Means Clustering method.

7. ACTIONABLE INSIGHTS & RECOMMENDATIONS

INSIGHTS:

Based on both clustering methods, the clustering groups are approximately **similar** and **identical.**

- **Group 1:**
  - Moderate Financial group
  - Avg. Credit limit: ~ $33700
  - Largest cluster with 386 or 387 customers based on clustering type.
  - This group has balanced approach across calls and bank visits.
  - This group has lowest online visits.
- **Group 2:**
  - Low financial Customer group
  - Avg. Credit limit: ~ $12200
  - This group has the lowest credit limit and lowest credit card usage.
  - This group is highly relied on call support
- **Group 3:**
  - High-Value Customer Group
  - Avg. Credit limit: $141,040
  - Smallest cluster with only 50 customers.
  - This group has the highest credit card usage (8.7) and highest credit limit.
  - Mostly involved in online banking visits with no need for in-person contacts.

RECOMMENDATIONS:

- The **High value customer group** can be retained by offering premium online services and discounted memberships. This group can be satisfied by offering effective and clean online support.
- For **Moderate financial group**, the online visits can be improved by educating the customers on the advantages of online banking and app support.
- The **Moderate financial group** can be retained by offering investing ideas and plans for their development.
- The **Low financial group** are mainly relied on call support and it can be reduced by educating the online services and support.
- The **Low financial group** has the lowest credit card usage and it can be boosted by offering low-fee services and entry-level credit cards with lesser interest rates.
- Improving the digital services for more user-friendly interfaces and lag free support.
- Online and offline services can be improved by collecting regular feedbacks from all groups.
- Promote campaigns to leverage the importance of investing and online banking.