UNSUPERVISED LEARNING

CREDIT CARD PROJECT- CODED

By: BENITA MERLIN.E

PGP-Data Science and Business Analytics.

BATCH: PGP DSBA. O. MAY24.A

Contents

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 Project Overview

All Life 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

1.2 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.

1.3 Problem Definition

- The goal is to segment the existing customer base based on spending patterns and past interactions with the bank. This segmentation will allow the bank to:
- Run personalized marketing campaigns to attract new customers and increase spending by existing ones.
- Improve customer support by understanding which segments are more likely to perceive the service negatively and upgrade the service delivery model accordingly.
- Use clustering techniques (K-Means, Hierarchical Clustering) to group customers.
- Apply Elbow Method and Silhouette Scores to determine optimal clusters.
- Targeted campaigns for better customer acquisition and retention.
- Improved customer service tailored to specific segments.
- Increased revenue through personalized marketing and better customer support.

2. 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).

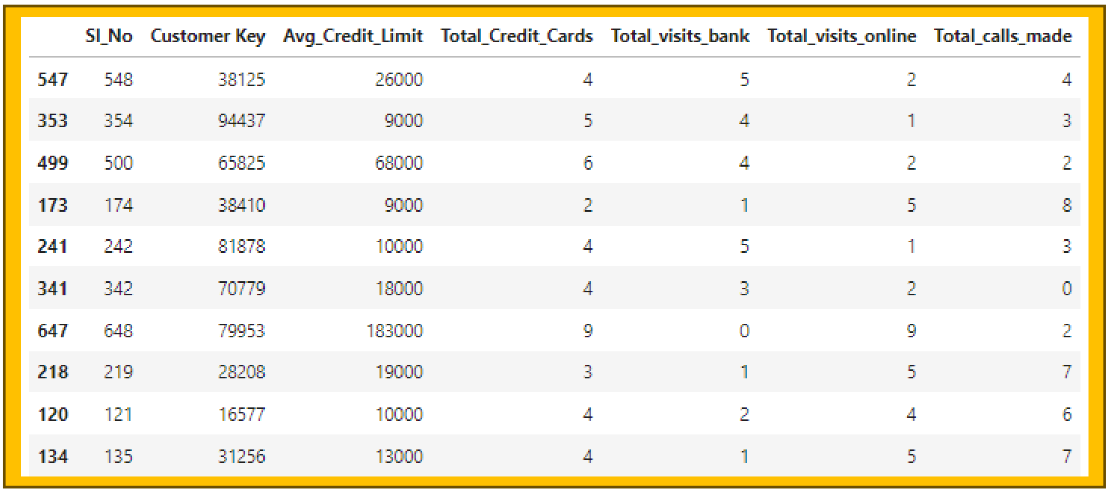
2.1 Data Dictionary

SI.NO	VARIABLE	DESCRIPTION
1	SI_NO	Primary key of the records
2	Customer Key	Customer identification number
3	Average Credit Limit	Average credit limit of each customer for all credit cards
4	Total credit cards	Total number of credit cards possessed by the customer
5	Total visits bank	Total number of visits that the customer made (yearly) personally to the bank
6	Total visits online	Total number of visits or online logins made by the customer (yearly)
7	Total calls made	Total number of calls made by the customer to the bank or its customer service department (yearly)

Table 1 Data Description

2.2 Sample Dataset

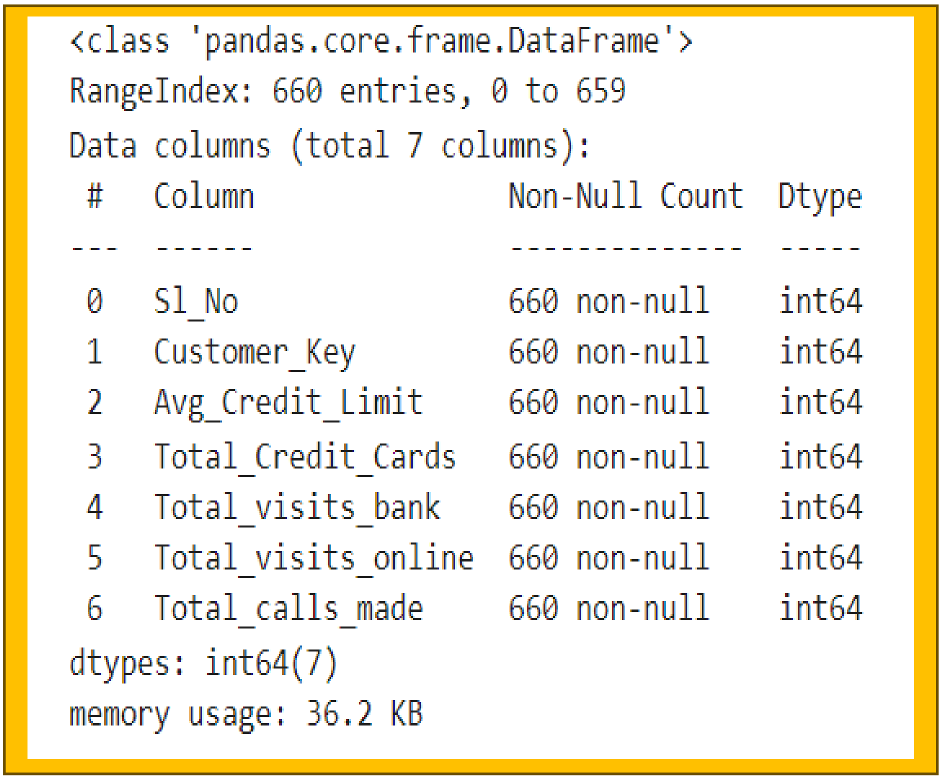
These are the random sample dataset.



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

Table 2 Sample Dataset

2.3 Data Information



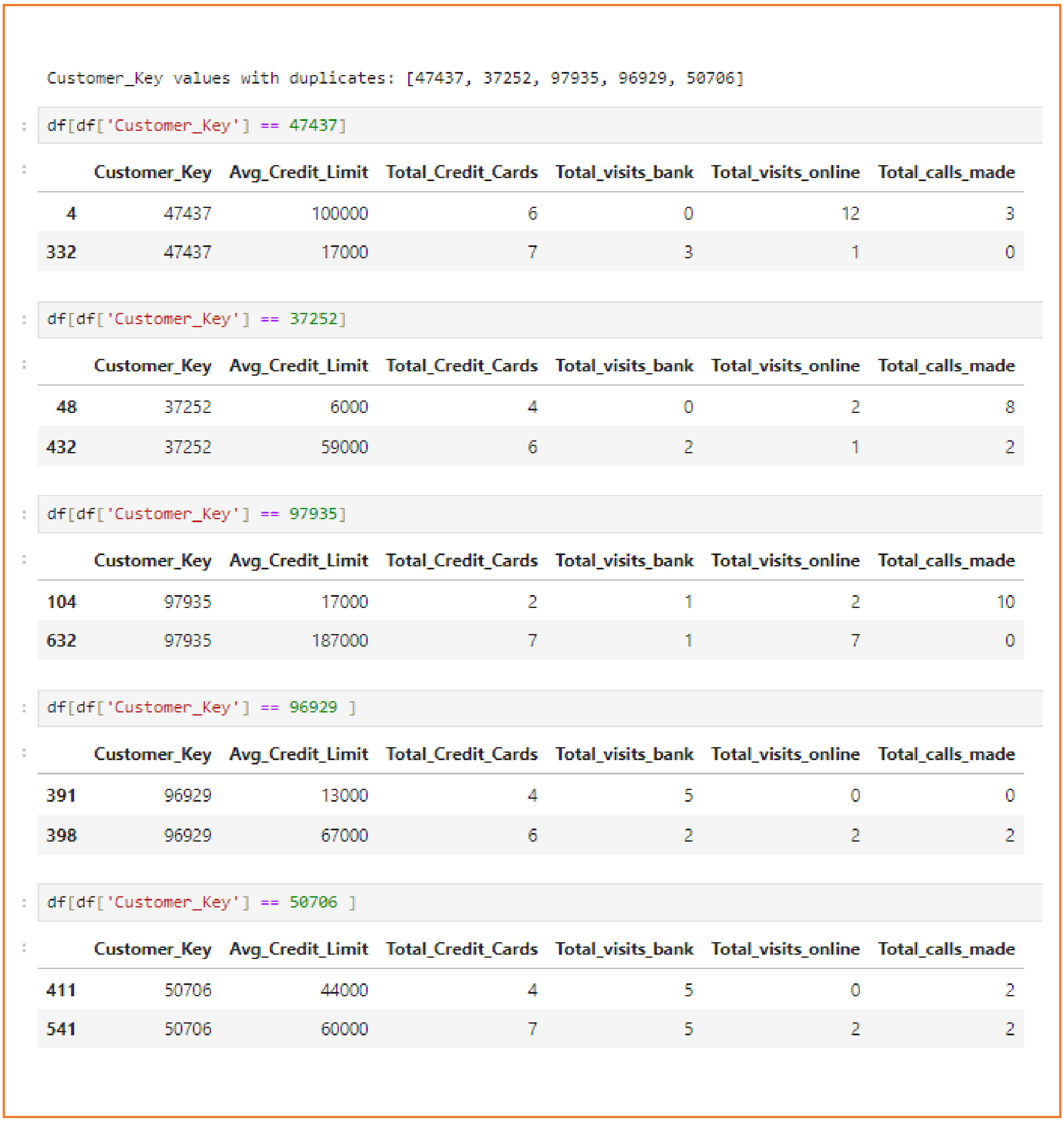
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 660 entries, 0 to 659
Data columns (total 7 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Sl_No                660 non-null    int64
 1   Customer_Key         660 non-null    int64
 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
```

Table 3 Data Information

Observation:

- There are 7 attributes which are int64.
- There are no missing values.
- The Shape of the dataset is 660 rows and 7 columns.
- There are no duplicated rows.

Duplicate Data Observation:



```
Customer_Key values with duplicates: [47437, 37252, 97935, 96929, 50706]
```

```
df[df['Customer_Key'] == 47437]
```

	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
4	47437	100000	6	0	12	3
332	47437	17000	7	3	1	0

```
df[df['Customer_Key'] == 37252]
```

	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
48	37252	6000	4	0	2	8
432	37252	59000	6	2	1	2

```
df[df['Customer_Key'] == 97935]
```

	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
104	97935	17000	2	1	2	10
632	97935	187000	7	1	7	0

```
df[df['Customer_Key'] == 96929 ]
```

	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
391	96929	13000	4	5	0	0
398	96929	67000	6	2	2	2

```
df[df['Customer_Key'] == 50706 ]
```

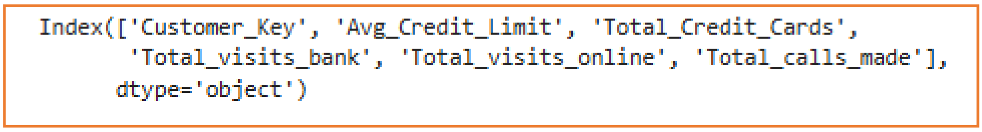
	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
411	50706	44000	4	5	0	2
541	50706	60000	7	5	2	2

Table 4 Finding Duplicate Values in Customer Key

Observation:

- Dataset Check: Upon checking the entire dataset, no duplicate rows were found.
- Customer key Check:
- However, upon further inspection, we found that there are duplicate values specifically in the Customer key column.
- This indicates that while the overall data rows are unique. The 'Customer Key' is a unique ID given to each customer in the database.

- The duplicate values might correspond to customer profile changes, and as such, there is no need to delete these records as these are actual occurrences at some point in the time.
- The Customer key column can be removed during the analysis.
- We won't need Serial. No. for analysis, so let's drop these columns.

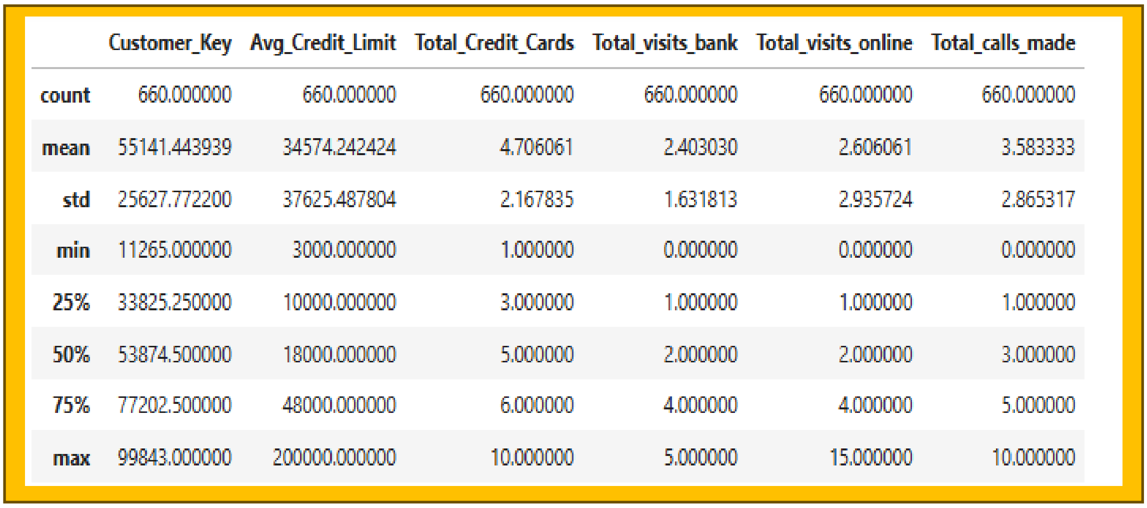


```
Index(['Customer_Key', 'Avg_Credit_Limit', 'Total_Credit_Cards',
       'Total_visits_bank', 'Total_visits_online', 'Total_calls_made'],
      dtype='object')
```

Table 5 After Dropping Serial Number in Dataset

3. STATISTICAL ANALYSIS

The statistical analysis provides a summary of the key metrics for the numerical columns in the dataset. This includes measures such as the count, mean, standard deviation, minimum, 25th percentile (Q1), median (50th percentile, Q2), 75th percentile (Q3), and maximum values for each column.



	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
mean	55141.443939	34574.242424	4.706061	2.403030	2.606061	3.583333
std	25627.772200	37625.487804	2.167835	1.631813	2.935724	2.865317
min	11265.000000	3000.000000	1.000000	0.000000	0.000000	0.000000
25%	33825.250000	10000.000000	3.000000	1.000000	1.000000	1.000000
50%	53874.500000	18000.000000	5.000000	2.000000	2.000000	3.000000
75%	77202.500000	48000.000000	6.000000	4.000000	4.000000	5.000000
max	99843.000000	200000.000000	10.000000	5.000000	15.000000	10.000000

Table 6 Statistical Summary

Observation:

Customer Key:

- As you noted, this column serves as a unique identifier for each customer, similar to an ID, and will not be useful as a feature for clustering.

Average Credit Limit:

- The average credit limit across customers is about 34,574, with a minimum of 3,000 and a maximum of 200,000.
- There's a wide range in credit limits, which could indicate different levels of financial capacity or spending power among customers.

Total Credit Cards:

- Customers have, on average, 4.7 credit cards, with a range of 1 to 10.
- The distribution here may help identify customers with higher engagement or loyalty to the bank.

Total Visits to Bank:

- The mean number of visits is 2.4, with a maximum of 5.
- Some customers do not visit the bank at all (minimum = 0), which might indicate a preference for digital or remote interactions.

Total Online Visits:

- The average is 2.6, with a maximum of 15.
- The large range here may signify a group of customers who are more digitally active and comfortable with online interactions.

Total Calls Made:

- The average number of calls made is 3.58, with some customers never calling (minimum = 0) and others making up to 10 calls.
- This could help identify customers with higher support needs.

Customer Engagement Analysis:

- There are 100 customers who haven’t visited banks once
- There are 144 customers who haven’t done online logins made by the customer (yearly)
- There are 97 customers who haven’t done calls made by the customer to the bank or its customer service department (yearly)
- Total number of customers who didn't visit the bank and didn't make any phone calls: 7
- Total number of customers who didn't visit the bank and didn't visit online: 0
- Total number of customers who didn't call to the bank and didn't visit online: 30

4. EXPLORATORY DATA ANALYSIS

4.1 UNIVARIATE ANALYSIS

4.1.1 Avg_Credit_Limit

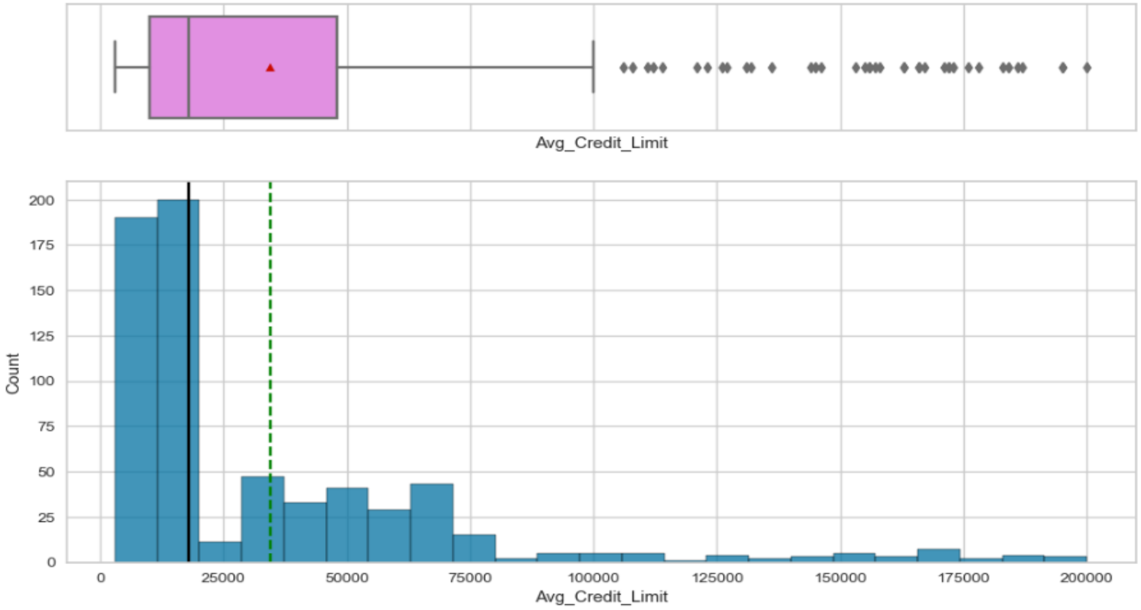


Figure 1 Distribution Plot of Average Credit Limit

Observation:

The image displays two plots that provide insights into the distribution of the Average Credit Limit:

1. Boxplot (Top)

- The boxplot shows that the median credit limit is relatively low (around 20,000).
- There are outliers on the higher end of the distribution, as evidenced by the dots beyond the upper whisker.
- The interquartile range (IQR) is also quite large, indicating high variability in the data.
- A red triangle marks the mean credit limit, which is higher than the median, suggesting a right-skewed distribution due to the presence of high-value outliers.

2. Histogram (Bottom)

- The histogram confirms the right-skewed nature of the data, with a large number of customers having low average credit limits, mostly around the 0 to 25,000 range.
- A black line indicates the median, closely aligned with the first bar of the histogram, reinforcing that most customers have lower credit limits.
- The green dashed line marks the mean, positioned higher than the median due to the influence of outliers with much larger credit limits.

- There is a long tail towards the higher credit limit values, with very few customers having limits above 100,000.

Overall Observations:

- The distribution is heavily skewed to the right, with most customers having relatively low credit limits and a few customers having significantly higher ones.
- Outliers in the higher credit range impact the mean, pulling it up above the median.

4.1.2 Total_Credit_Cards

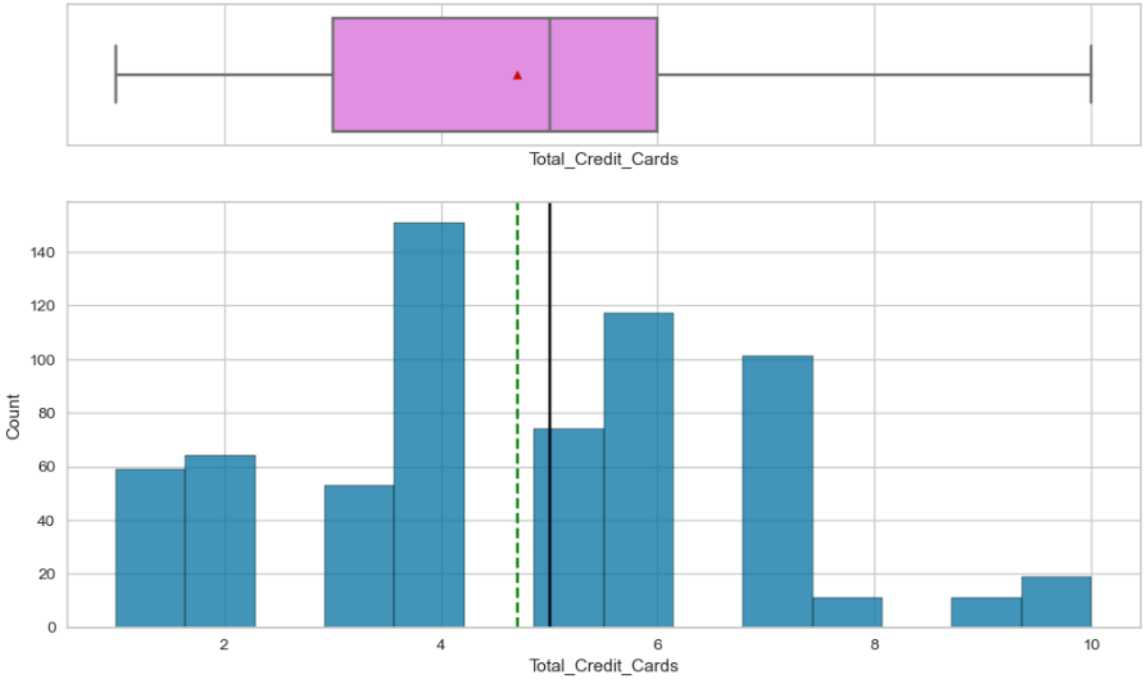


Figure 2 Distribution Plot of Total Credit Cards

Observation:

The image shows two plots representing the distribution of Total Credit Cards held by customers:

1. Boxplot (Top)

- The boxplot indicates that the median number of credit cards is around 5.
- The interquartile range (IQR) spans from approximately 3 to 7, showing moderate variability in the number of credit cards.
- There are no significant outliers, as the whiskers capture most of the data, and there are no distant points.
- The mean, marked by the red triangle, is slightly below the median, suggesting a relatively symmetric distribution, though with a possible slight left skew.

2. Histogram (Bottom)

- The histogram reveals a multimodal distribution. Peaks can be observed at 4 and 6 total credit cards, indicating that these are the most common values among customers.
- The black vertical line indicates the median (around 5), while the green dashed line marks the mean (slightly below 5), showing they are fairly close.
- There are distinct groups of customers with 2, 4, 6, and 8 credit cards, indicating clusters of behavior, while a small number of customers hold 9 or 10 cards.

Overall Observations:

- The distribution is fairly symmetrical, with most customers having between 3 to 7 credit cards.
- The data suggests multiple popular values, with clear peaks at 4 and 6 credit cards, possibly reflecting common customer behavior or product offerings.
- There are no extreme outliers, as most customers fall within a reasonable range of credit cards.

4.1.3 Total_Visits_Bank

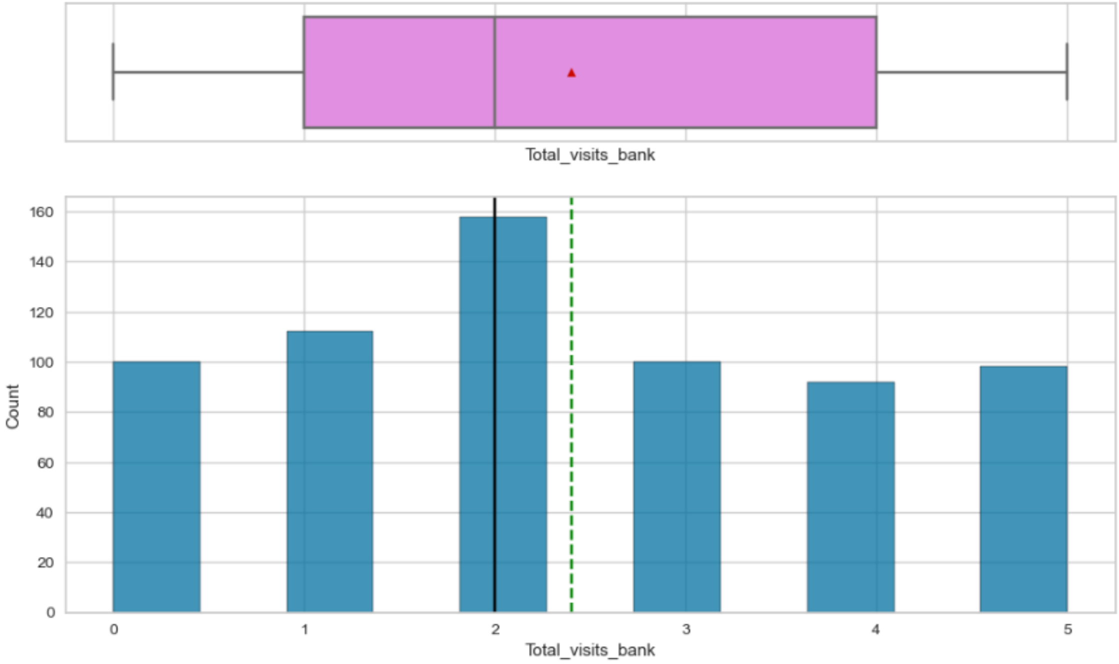


Figure 3 Distribution Plot of Total Visits Bank

Observation:

The image provides two plots representing the distribution of Total Visits to the Bank:

1. Boxplot (Top)

- The boxplot shows that the median number of bank visits is around 2.

- The interquartile range (IQR) spans from 1 to 4, indicating that the majority of customers visit the bank between 1 to 4 times.
- The whiskers extend to 0 and 5 visits, showing that most of the data falls within this range without any outliers.
- The mean (marked by the red triangle) is slightly above 2, indicating that the distribution is slightly right-skewed.

2. Histogram (Bottom)

- The histogram indicates a relatively uniform distribution with peaks at 0, 2, and 4 visits.
- The highest frequency of customers made 2 visits, as indicated by the tallest bar.
- The black line marks the median at 2, showing that half the customers visit 2 or fewer times.
- The green dashed line marks the mean, slightly above the median, suggesting a small positive skew.
- A notable portion of customers (around 100) did not visit the bank at all (0 visits), while others have a fairly spread out distribution of up to 5 visits.

Overall Observations:

- The distribution of bank visits is slightly right-skewed, with 2 visits being the most common.
- No extreme outliers exist, and the data is spread across a reasonable range of 0 to 5 visits.
- A significant number of customers do not visit the bank, but others visit 1-5 times, with no pronounced trend beyond 2 visits.

4.1.4 Total_Visits_Online

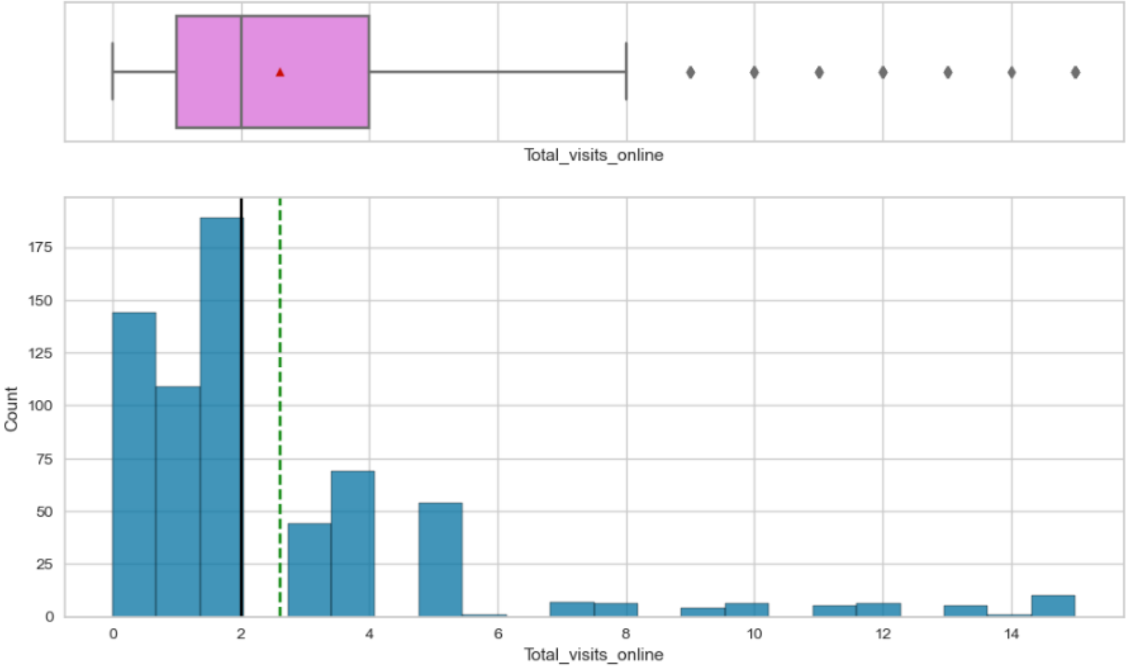


Figure 4 Distribution Plot of Total Visits Online

Observation for Total Visits Online:

Boxplot:

- The median number of online visits is around 2.
- The interquartile range (IQR) spans from around 0 to 4 online visits, indicating that half of the customers visit the online banking system 0 to 4 times.
- There are several outliers beyond 8 online visits, with some customers making up to 14 or 15 visits.
- The whiskers extend from 0 to 6 visits, showing that most customers fall within this range, but outliers stretch further.

Histogram:

- There is a high concentration of customers who have made 0 to 2 online visits, with a significant peak at 2 visits.
- Beyond 4 visits, the frequency drops considerably, indicating that only a small portion of customers make frequent online visits.
- A small but noticeable group of customers make 8 or more online visits, but they are a minority.

Summary:

- Most customers make fewer than 5 online visits per year, with the majority clustering around 0 to 2 visits.
- There are a few customers who are highly active online, with more than 8 visits, but they represent a minority.
- The presence of outliers suggests that while online engagement is low for most customers, some have very high interaction levels.

4.1.5 Total_Calls_Made

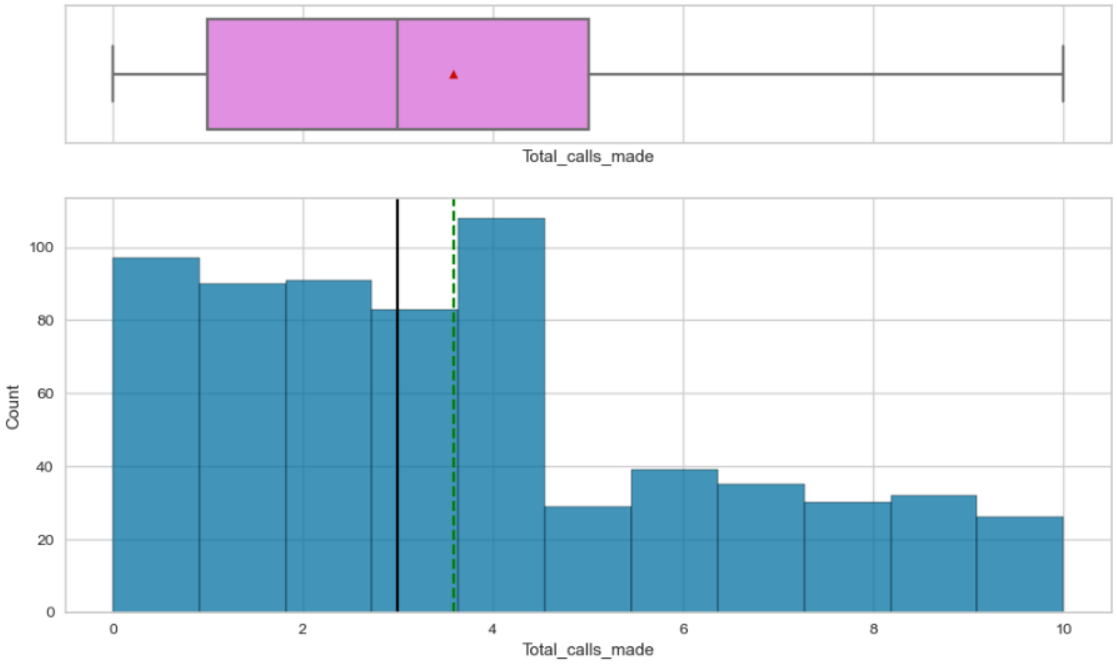


Figure 5 Distribution Plot of Total Calls Made

Observation for Total Calls Made:

Boxplot:

- The median number of calls made is around 3.5, with the IQR (Interquartile Range) spanning from 1 to 6 calls.
- There are no significant outliers, meaning the total calls made by customers are fairly evenly distributed without extreme values.
- The mean is slightly higher than the median, indicating a slight positive skew (right-skewed distribution).
- The whiskers extend from 0 to 10, showing that the entire range of calls made falls within this limit.

Histogram:

- The distribution shows a relatively uniform spread between 0 and 5 calls, with frequencies around 80-100 for each call range.
- There is a small drop-off in customers making 6 to 10 calls, with the frequency declining progressively after 5 calls.
- A notable peak is seen around 1 to 4 calls, with slightly fewer customers making more than 5 calls.
- The frequency of 0 calls (no calls made) is still considerable, indicating some customers are not making calls to the bank.

Summary:

- The total number of calls made by customers is distributed relatively evenly up to 5 calls, with fewer customers making more than 5 calls.
- There is a slight positive skew, as most customers make fewer than 5 calls.
- There are no outliers, indicating a consistent calling pattern across customers.

4.1.6 Cumulative Distribution Function (CDF) Plot of Numerical Variables:

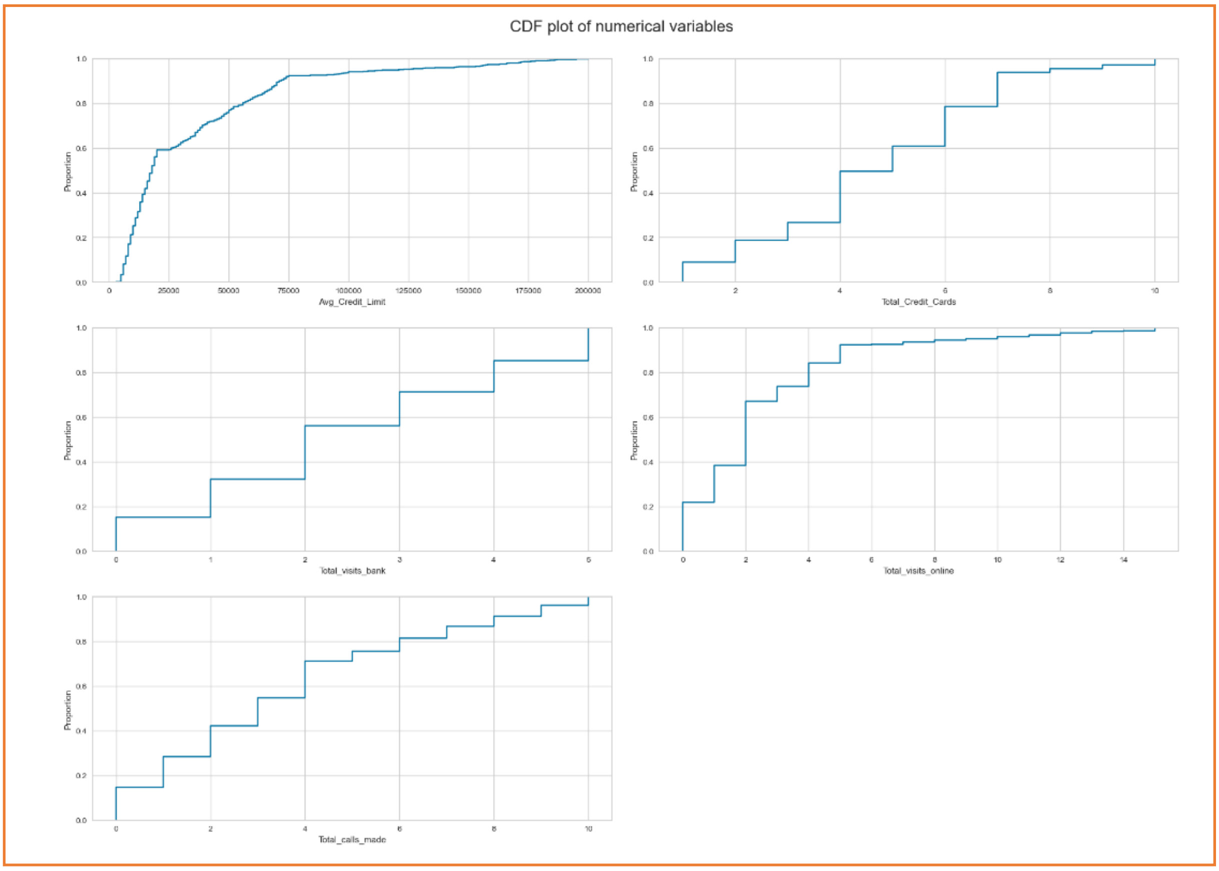


Figure 6 Cumulative Distribution Function (CDF) Plot

Observations from the CDF Plots:

1. **Avg_Credit_Limit (Top Left)**
   - The CDF shows a steep increase between 0 and around 50,000, indicating that a large proportion of customers have an average credit limit below this value.
   - After 50,000, the curve flattens, suggesting fewer customers have credit limits above this range.
2. **Total_Credit_Cards (Top Right)**
   - The CDF has a step-like pattern with major jumps at 2, 4, and 6 credit cards, implying that most customers hold between 2 and 6 cards.
   - There is a small fraction of customers with more than 6 credit cards.
3. **Total_visits_bank (Middle Left)**
   - The CDF indicates that a significant proportion of customers (around 40%) visit the bank 2 or fewer times.
   - Fewer customers make more than 3 bank visits, as the curve becomes less steep after 2.
4. **Total_visits_online (Middle Right)**
   - The CDF shows that the majority of customers (over 50%) make 3 or fewer online visits.
   - After 3 visits, the curve rises more gradually, meaning fewer customers make higher numbers of online visits.
5. **Total_calls_made (Bottom Left)**
   - The CDF indicates that a large portion of customers (about 40%) make between 0 to 4 calls.
   - The curve rises steadily, showing a distribution of customers making between 0 to 10 calls, with no major concentration at any single number of calls.

4.1.7 Bar Plot of Total Credit Cards

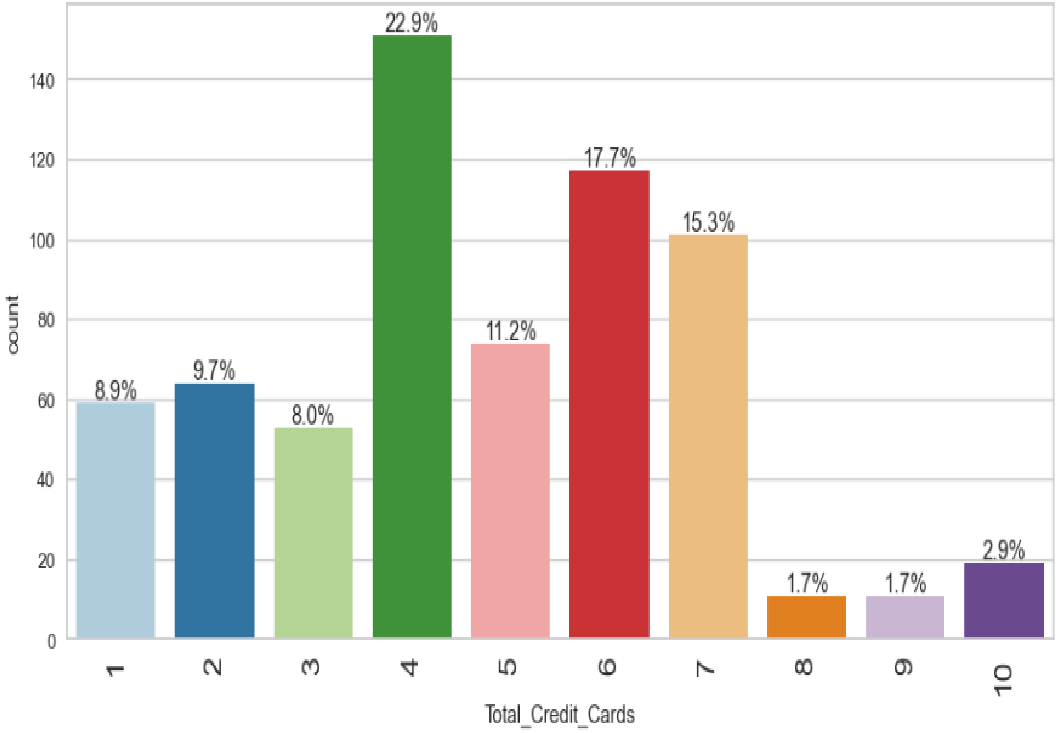


Figure 7 Bar Plot of Total Credit Cards

Observation:

1. The highest percentage of customers (22.9%) own 4 credit cards.
2. The second-highest percentage of customers (17.7%) have 6 credit cards.
3. There is a significant drop in the number of customers with 8, 9, or 10 credit cards, each constituting a small portion of the dataset.
4. 1, 2, and 3 credit cards are less common than 4, 5, 6, or 7 cards, with percentages ranging between 8.0% to 9.7%.
5. A pattern seems to suggest that most customers tend to have between 4 to 7 credit cards, making up the bulk of the distribution.

This distribution indicates that the majority of customers have moderate credit card holdings, with fewer customers having very few or very many cards.

4.1.8 Bar Plot of Total Visit Bank

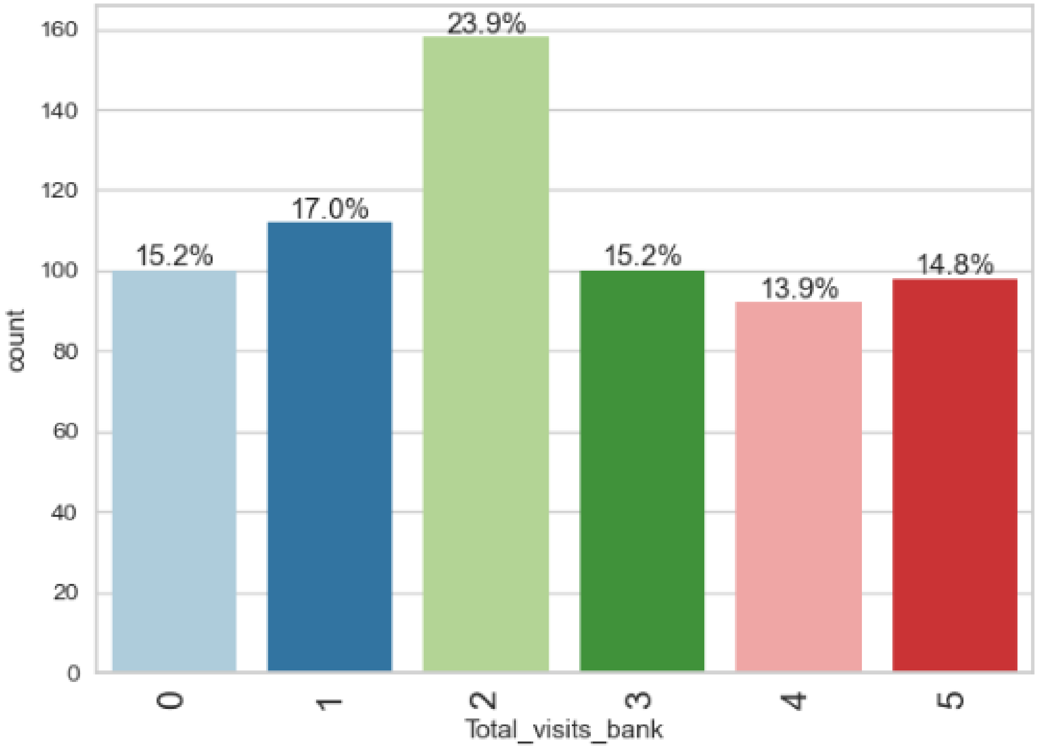


Figure 8 Bar Plot of Total Visits Bank

Observation:

1. The highest percentage of customers (23.9%) visit the bank 2 times.
2. 17.0% of customers visit the bank once, which is the second most frequent number of visits.
3. 15.2% of customers either visit 0 times or 3 times.
4. 14.8% of customers visit the bank 5 times, and 13.9% visit 4 times.
5. Overall, most customers visit the bank 2 or 1 times, with a fairly even distribution across 0, 3, 4, and 5 visits.

This suggests that the majority of customers tend to visit the bank infrequently, with 2 visits being the most common.

4.1.9 Bar Plot of Total Visit Online

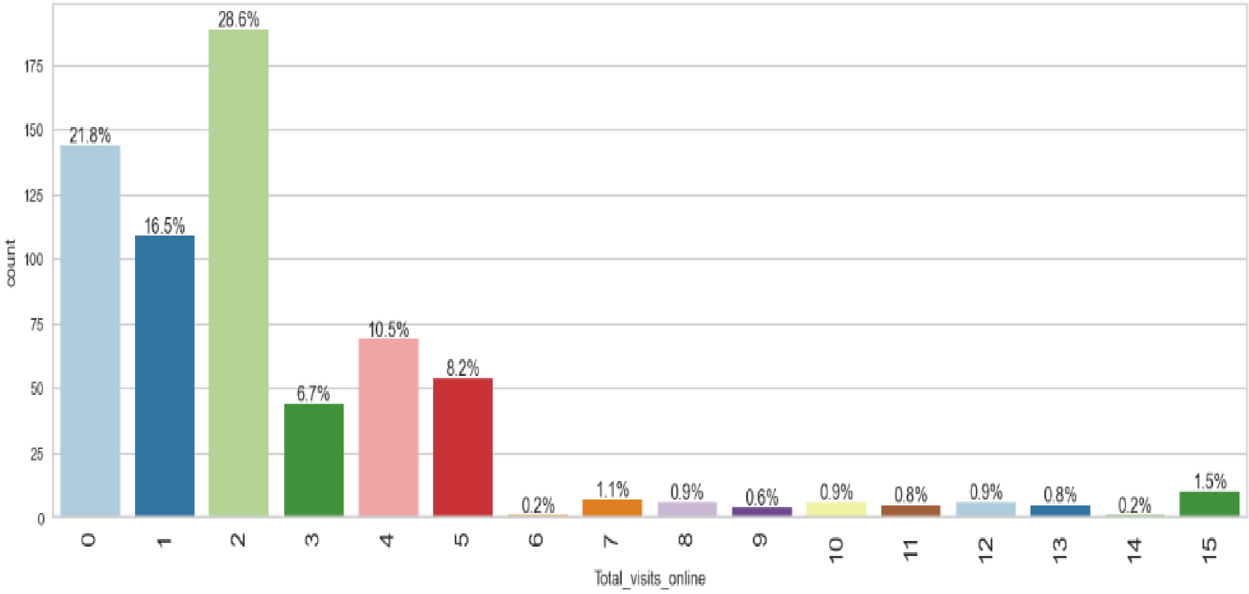


Figure 9 Bar Plot of Total Visits Online

Observation:

1. The majority of customers (28.6%) visit the bank online 2 times.
2. 21.8% of customers make 0 online visits, suggesting they don't use online services.
3. 16.5% of customers visit online once, while 10.5% visit 4 times.
4. The percentage of customers who visit online 3 or 5 times is 6.7% and 8.2%, respectively.
5. Online visits above 5 are rare, with minimal customer percentages for visits ranging from 6 to 15. The percentage is especially low for visits between 6 and 13, with values around 1% or lower.

This indicates that most customers visit online 0 to 5 times, with 2 being the most frequent. Very few customers exceed 5 online visits, indicating limited frequent use of online services.

4.1.10 Bar Plot of Total Calls Made

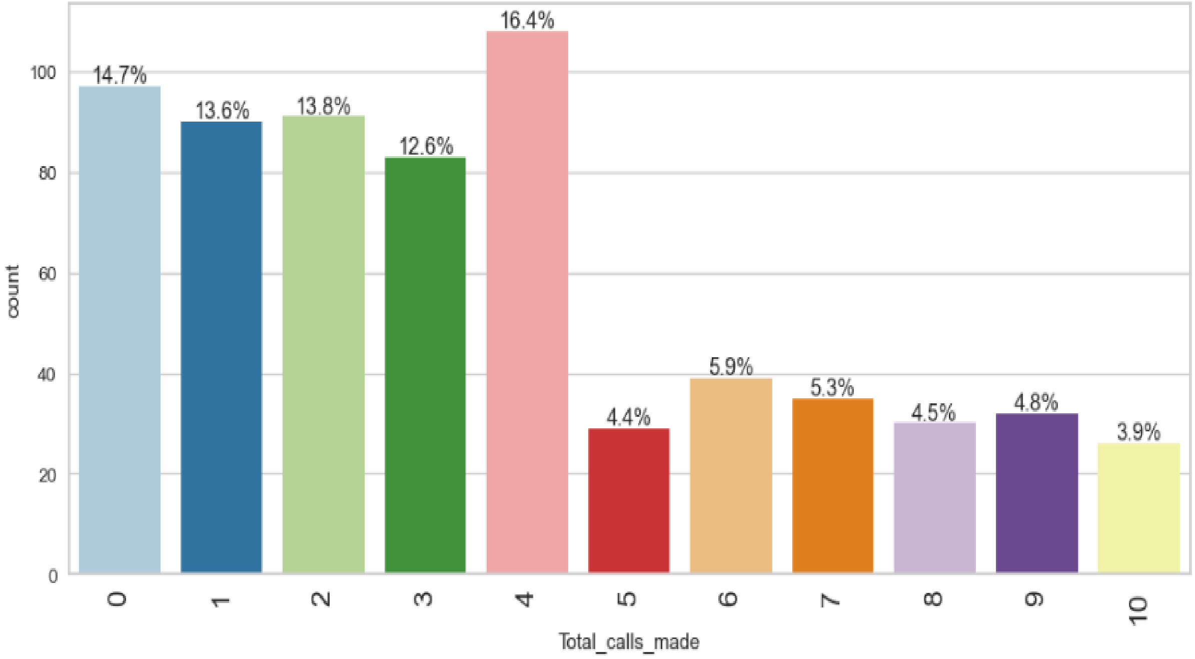


Figure 10 Bar Plot of Total Calls Made

Observation

- High Concentration at 4 Calls: The majority of customers make around 4 calls, with 16.4% of users in this category. This might indicate that a typical customer interaction or resolution might require multiple calls, with 4 being the most frequent.
- Gradual Decline After 4: After the 4-call mark, the number of calls significantly drops, which could imply that after a certain number of calls, the issue is generally resolved, or customer service fatigue sets in, reducing further interaction.
- Not Many Customers Call More Than 6 Times: Beyond 6 calls, the proportion of customers decreases notably (less than 6% for each category). This suggests that issues are mostly resolved in fewer than 5 or 6 calls, and prolonged interactions are less common.
- 0 Calls (14.7%): A significant portion of customers (14.7%) haven’t made any calls, which could indicate either no issues or a preference for self-service options like mobile apps, websites, or chatbots.

4.1.11 Bar Plot of Avg Credit Limit

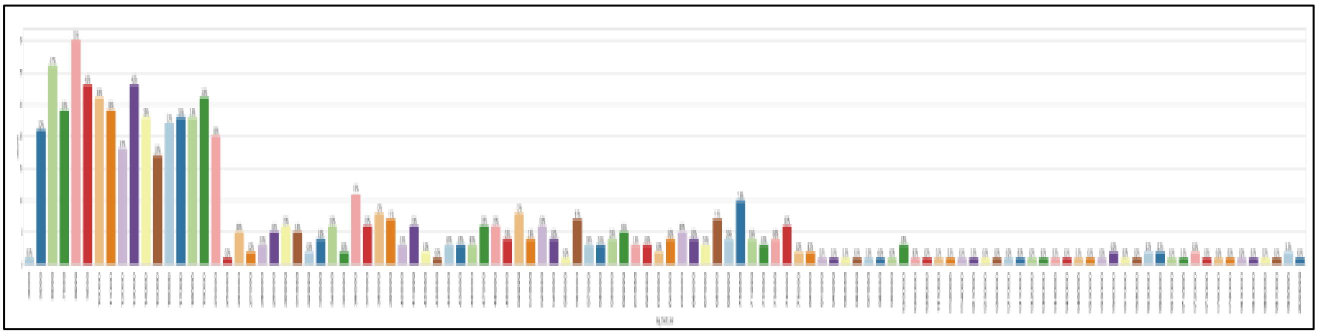


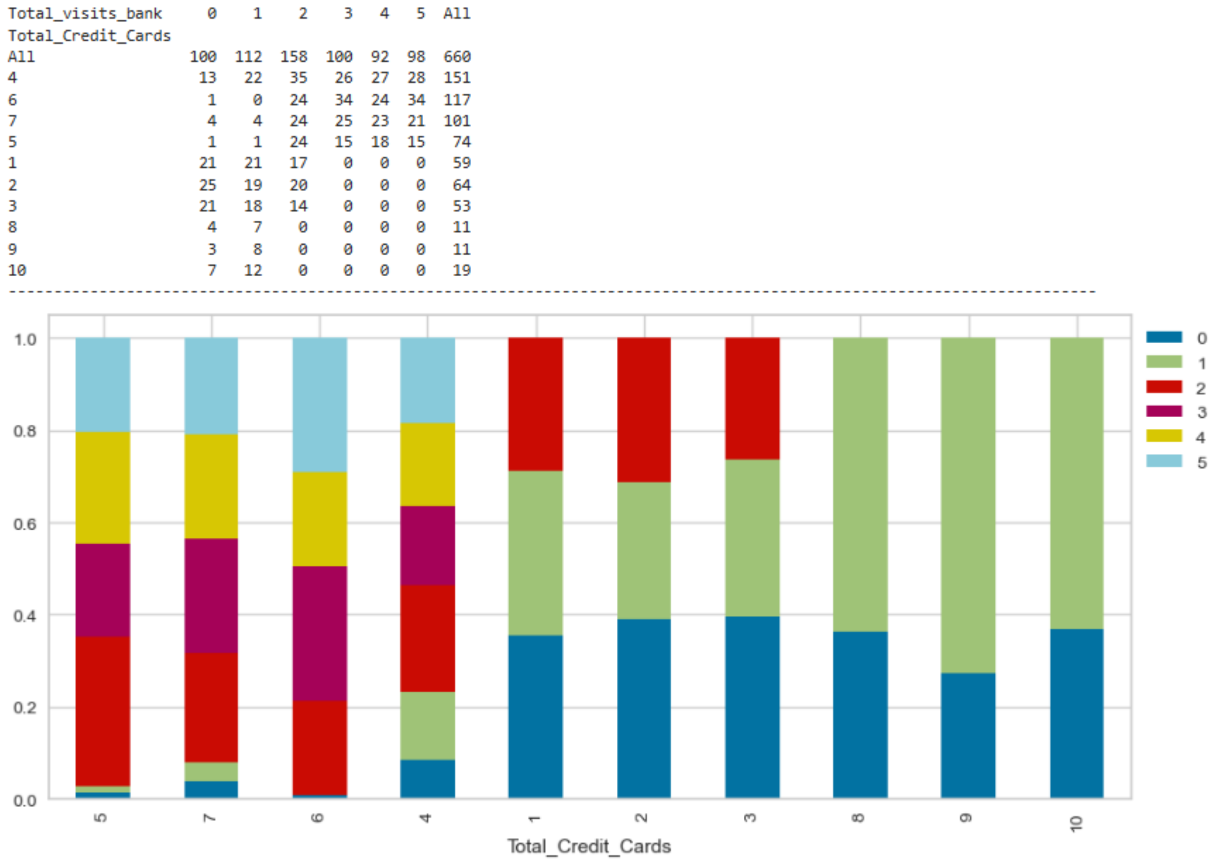
Figure 11 Bar plot of Average Credit Limit

Observation:

- Concentration of Activity: Most of the important or frequent events happen in the earlier categories. If this represents customer behavior, transaction amounts, or other similar data, the first few categories are where the focus should be.
- Tail-End Events: The long tail indicates there are still a few occurrences far out on the x-axis, but these are much rarer compared to the frequent activity seen in the early categories.

4.2 BIVARIATE ANALYSIS

4.2.1 Total_Credit_Cards Vs Total_Visits_Bank



```
Total_visits_bank     0    1    2    3   4   5  All
Total_Credit_Cards
All                 100  112  158  100  92  98  660
4                    13   22   35   26  27  28  151
6                     1    0   24   34  24  34  117
7                     4    4   24   25  23  21  101
5                     1    1   24   15  18  15   74
1                    21   21   17    0   0   0   59
2                    25   19   20    0   0   0   64
3                    21   18   14    0   0   0   53
8                     4    7    0    0   0   0   11
9                     3    8    0    0   0   0   11
10                    7   12    0    0   0   0   19
-------------------------------------------------------------------------------------------------------------------
```

Figure 12 Stacked Bar Plot of Total Credit Cards Vs Total Visit Bank

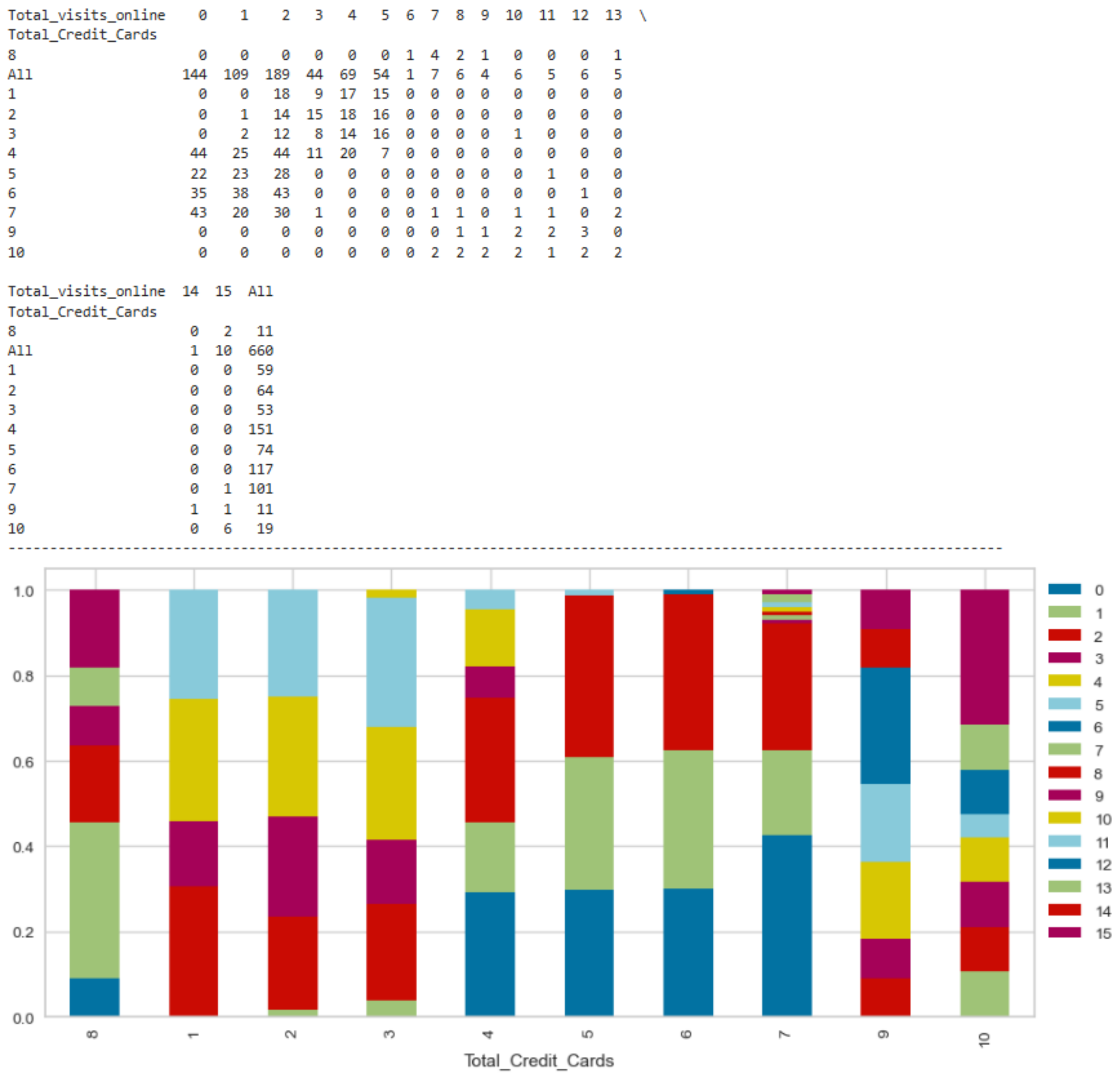
Observation

- The image provides a stacked bar chart that depicts the relationship between the Total Credit Cards a person has and the Total Visits to the Bank. Here are the key observations:
- Variation Across Credit Card Holders:
- Individuals with 5 or more credit cards (represented on the left side of the chart) tend to have a more varied number of bank visits. This group has a more spread-out distribution, with a larger share of people making anywhere from 1 to 5 visits.
- Lower Bank Visits for Fewer Credit Cards:
- Individuals with 1 or 2 credit cards have fewer visits overall, as indicated by the dominance of the blue (0 visits) and green (1 visit) sections. This suggests that people with fewer credit cards are less likely to visit the bank frequently.
- High Number of Visits for Certain Groups:
- For people with 5 or more credit cards, there are noticeable bars representing 4 and 5 visits to the bank, which indicates more frequent banking activity. This group tends to visit the bank more often, possibly due to managing more complex financial needs.
- The group with 1 credit card has a substantial portion of individuals making 2 or more visits (represented by red and purple bars), which might indicate that certain customers with fewer credit cards still visit the bank frequently, likely due to other banking needs.
- Overall Trend:
- As the number of credit cards increases, there is a trend towards a higher diversity in the number of bank visits, with a broader range of visit counts (0 through 5).
- The zero visits (blue) segment is quite prominent across most categories, suggesting that many individuals across all credit card ownership levels rarely visit the bank.

Summary:

- Individuals with more credit cards generally make more frequent visits to the bank, with a more diverse distribution of visit counts.
- People with fewer credit cards (especially 1 or 2) tend to make fewer bank visits, and the majority have either 0 or 1 visits.

4.2.2 Total_Credit_Cards Vs Total_Visits_Online



```
Total_visits_online    0    1    2   3   4   5  6  7  8  9  10  11  12  13  \
Total_Credit_Cards
8                      0    0    0   0   0   0  1  4  2  1   0   0   0   1
All                  144  109  189  44  69  54  1  7  6  4   6   5   6   5
1                      0    0   18   9  17  15  0  0  0  0   0   0   0   0
2                      0    1   14  15  18  16  0  0  0  0   0   0   0   0
3                      0    2   12   8  14  16  0  0  0  0   1   0   0   0
4                     44   25   44  11  20   7  0  0  0  0   0   0   0   0
5                     22   23   28   0   0   0  0  0  0  0   0   1   0   0
6                     35   38   43   0   0   0  0  0  0  0   0   0   1   0
7                     43   20   30   1   0   0  0  1  1  0   1   1   0   2
9                      0    0    0   0   0   0  0  0  1  1   2   2   3   0
10                     0    0    0   0   0   0  0  2  2  2   2   1   2   2

Total_visits_online  14  15  All
Total_Credit_Cards
8                     0   2   11
All                   1  10  660
1                     0   0   59
2                     0   0   64
3                     0   0   53
4                     0   0  151
5                     0   0   74
6                     0   0  117
7                     0   1  101
9                     1   1   11
10                    0   6   19
------------------------------------------------------------------------------------------------------------------------
```

Figure 13 Stacked Bar Plot of Total Credit Cards Vs Total Visits Online

Observation

• The image provides a stacked bar chart that depicts the relationship between the Total Credit Cards a person has and their Total Visits Online. Here are the key observations:

Variation Across Credit Card Holders:

- Individuals with 5 or more credit cards have a more varied pattern of online visits. This group shows a more spread-out distribution of online visits, with a larger share of people making between 1 to 5 visits, represented by a variety of colors across the bar.

Lower Online Visits for Fewer Credit Cards:

- Individuals with 1 or 2 credit cards show a dominance of 0 visits (represented by the dark blue color) and a significant proportion of 1 visit (light blue). This indicates that people with fewer credit cards tend to make fewer online visits, with a large portion of them not visiting online platforms at all.

High Online Visits for Certain Groups:

- Individuals with 5 or more credit cards have a noticeable share of bars representing higher online visits (e.g., 4 or 5 visits, represented by red and yellow bars). This suggests that people with more credit cards are more engaged online, possibly due to more complex financial management needs.
- The group with 2 credit cards has a significant portion of individuals making 2 or more visits (seen in the red, purple, and yellow sections), suggesting that even among customers with fewer credit cards, some are still more active online, potentially due to other factors such as online banking preferences or transactional needs.

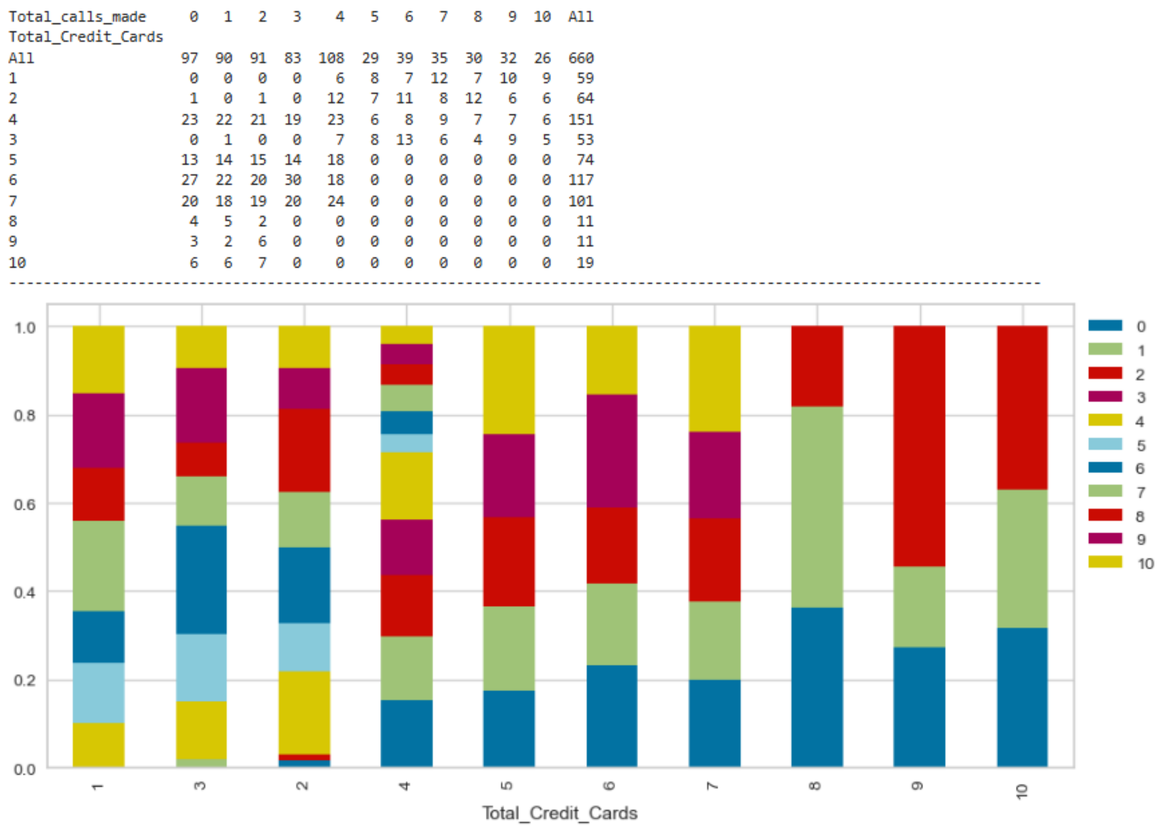
Overall Trend:

- As the number of credit cards increases, the range of online visits becomes more diverse, with a broader range of visit counts (0 through 5). This suggests that customers with more credit cards are likely to be more digitally active, potentially due to managing multiple accounts or conducting more frequent online transactions.
- Zero online visits (dark blue) is prominent across most categories, indicating that a significant proportion of individuals, regardless of the number of credit cards they own, do not engage in online visits at all.

Summary:

- Individuals with more credit cards tend to make more online visits, with a broader and more varied range of visit counts.
- Those with fewer credit cards (especially 1 or 2) are more likely to have lower online activity, with the majority either not making online visits or making only 1 visit.

4.2.3 Total_Credit_Cards Vs Total_calls_made



```
Total_calls_made     0   1   2   3    4   5   6   7   8   9  10  All
Total_Credit_Cards
All                 97  90  91  83  108  29  39  35  30  32  26  660
1                    0   0   0   0    6   8   7  12   7  10   9   59
2                    1   0   1   0   12   7  11   8  12   6   6   64
4                   23  22  21  19   23   6   8   9   7   7   6  151
3                    0   1   0   0    7   8  13   6   4   9   5   53
5                   13  14  15  14   18   0   0   0   0   0   0   74
6                   27  22  20  30   18   0   0   0   0   0   0  117
7                   20  18  19  20   24   0   0   0   0   0   0  101
8                    4   5   2   0    0   0   0   0   0   0   0   11
9                    3   2   6   0    0   0   0   0   0   0   0   11
10                   6   6   7   0    0   0   0   0   0   0   0   19
------------------------------------------------------------------------------------------
```

Figure 14 Stacked Bar Plot of Total Credit Cards Vs Total Calls Made

Observation

Variation Across Credit Card Holders:

- Individuals with 5 or more credit cards show a more varied distribution of calls made, as indicated by the spread of colors across the bars. These individuals tend to make more frequent calls, as seen by the noticeable sections for 3 or more calls in red and yellow.

Lower Call Frequency for Fewer Credit Cards:

- Individuals with 1 to 3 credit cards tend to have lower call frequencies overall. The majority of these groups show a significant proportion of people making 0 calls (dark blue) and 1 call (light blue), indicating that people with fewer credit cards are less likely to contact the bank frequently.

High Number of Calls for Certain Groups:

- For individuals with 5 or more credit cards, there is a noticeable increase in the number of people making 3 or more calls (seen in the red, purple, and yellow bars). This suggests that customers with more credit cards tend to engage with the bank more often, potentially due to more complex financial management needs.

- The group with 7 credit cards also shows a significant number of people making 2 or more calls, indicating that even among this group, there is increased contact with the bank.

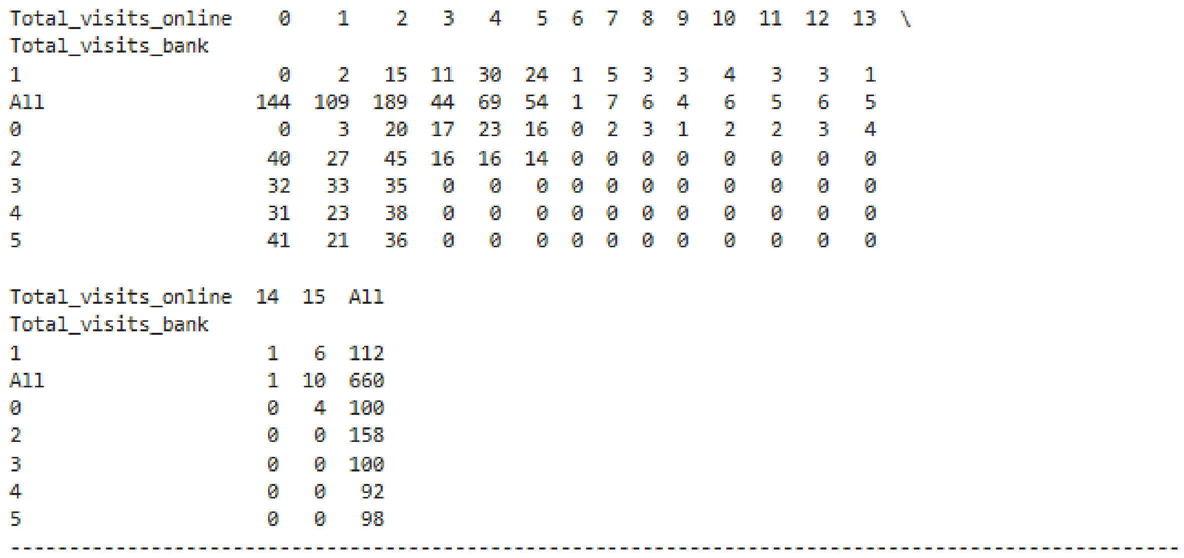
Overall Trend:

- As the number of credit cards increases, the distribution of calls made becomes more varied. Those with more credit cards tend to make more calls to the bank, and the number of calls ranges from 0 to as high as 10.
- Zero calls (dark blue) is still quite prominent across most categories, especially for those with fewer credit cards, indicating that a large portion of customers do not make any calls at all, regardless of how many credit cards they own.

Summary:

- Individuals with more credit cards (especially 5 or more) tend to make more frequent calls to the bank, with a more varied range of call counts.
- Customers with fewer credit cards (1 to 3) are more likely to have lower call frequencies, and the majority either make 0 or 1 call.

4.2.4 Total_Visits_Bank Vs Total_Visits_Online



```
Total_visits_online    0    1    2   3   4   5  6  7  8  9  10  11  12  13  \
Total_visits_bank
1                      0    2   15  11  30  24  1  5  3  3   4   3   3   1
All                  144  109  189  44  69  54  1  7  6  4   6   5   6   5
0                      0    3   20  17  23  16  0  2  3  1   2   2   3   4
2                     40   27   45  16  16  14  0  0  0  0   0   0   0   0
3                     32   33   35   0   0   0  0  0  0  0   0   0   0   0
4                     31   23   38   0   0   0  0  0  0  0   0   0   0   0
5                     41   21   36   0   0   0  0  0  0  0   0   0   0   0

Total_visits_online  14  15  All
Total_visits_bank
1                     1   6  112
All                   1  10  660
0                     0   4  100
2                     0   0  158
3                     0   0  100
4                     0   0   92
5                     0   0   98
------------------------------------------------------------------------------------------------
```

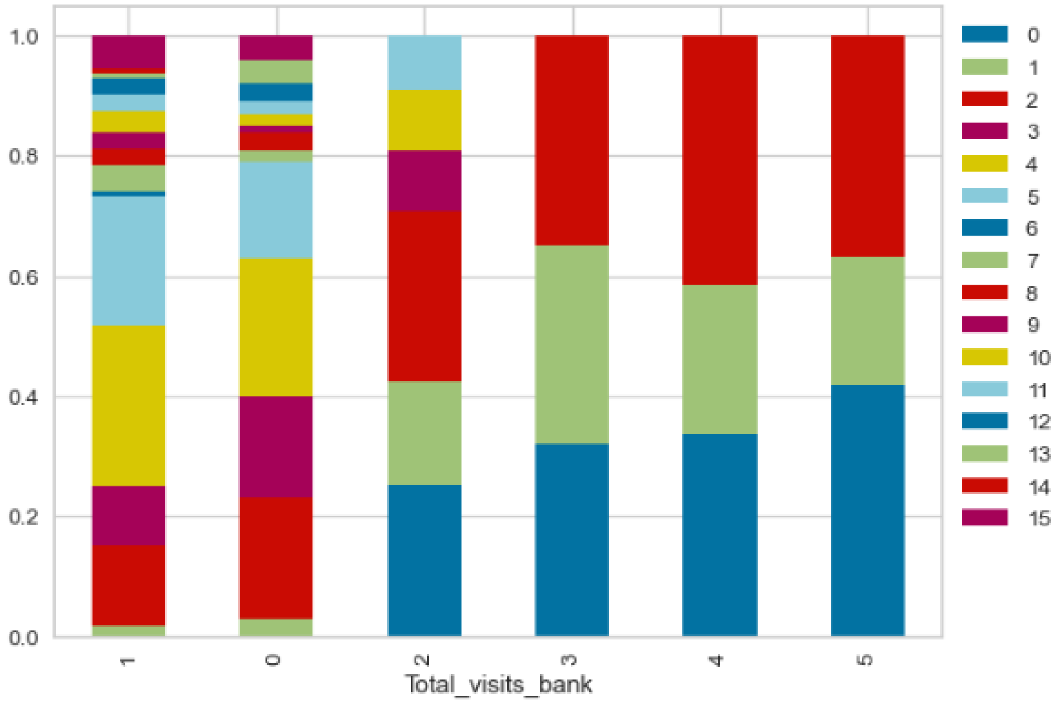


Figure 15 Stacked Bar Plot of Total Visits Bank Vs Total Visits Online

Distribution of Individuals Across Total Visits Online:

- 0 visits online:
  - The largest group of individuals (144) falls into this category.
  - Out of these, 40 individuals made 2 visits to the bank, and 41 individuals made 5 visits to the bank.
  - This indicates that individuals who do not visit online tend to rely more on in-person visits.

- 1 visit online:
    - A significant number of individuals (109) made 1 visit online.
    - Among these, 33 individuals made 3 visits to the bank, suggesting a mixed engagement across both online and bank visits.
- 2 visits online:
    - 189 individuals made 2 visits online, the largest in terms of online visits.
    - Among these, 45 individuals also made 2 visits to the bank, while 36 individuals made 5 visits to the bank. This shows a higher overlap in engagement across both online and physical visits for this group.

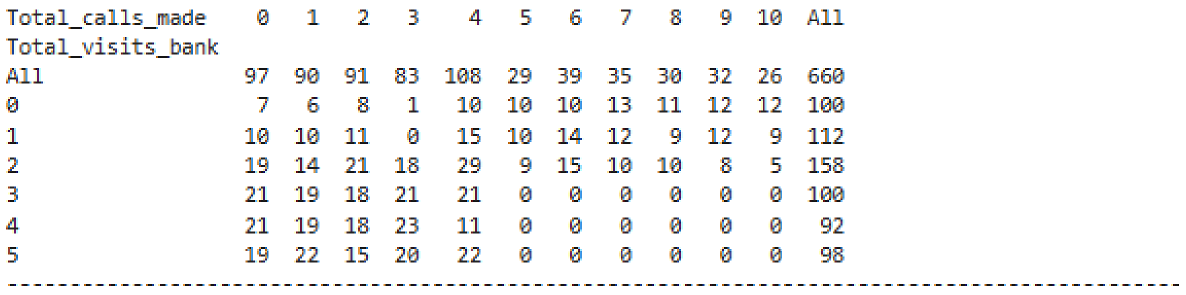
Total Visits to the Bank:

- 1 visit to the bank:
    - A large number of individuals (112) made 1 visit to the bank, but there's a wide spread in their online activity.
    - While 0 individuals made 0 visits online, 30 individuals made 4 online visits, showing a pattern of more engagement online but only limited in-person visits.
- 0 visits to the bank:
    - 100 individuals did not visit the bank at all.
    - Out of these, 41 individuals made 5 online visits, suggesting that these individuals primarily rely on online platforms for their banking needs.
- 2 visits to the bank:
    - 158 individuals made 2 visits to the bank, and most of them (45) also made 2 visits online, which indicates a consistent engagement across both platforms.

Summary:

- The largest group of individuals seems to prefer a low engagement across both Total Visits Online and Total Visits to the Bank (e.g., 0 or 1 visit in both categories).
- However, there are notable segments of individuals with high online engagement (e.g., 2 or more visits online), some of whom still rely on the bank for in-person visits, while others exclusively engage online.
- A significant portion of individuals who do not visit the bank are frequent online users, which could indicate a shift towards online services for a portion of the population.

4.2.5 Total_Visits_Bank Vs Total_Calls_Made



```
Total_calls_made    0   1   2   3    4   5   6   7   8   9  10  All
Total_visits_bank
All                97  90  91  83  108  29  39  35  30  32  26  660
0                   7   6   8   1   10  10  10  13  11  12  12  100
1                  10  10  11   0   15  10  14  12   9  12   9  112
2                  19  14  21  18   29   9  15  10  10   8   5  158
3                  21  19  18  21   21   0   0   0   0   0   0  100
4                  21  19  18  23   11   0   0   0   0   0   0   92
5                  19  22  15  20   22   0   0   0   0   0   0   98
--------------------------------------------------------------------------------------------
```

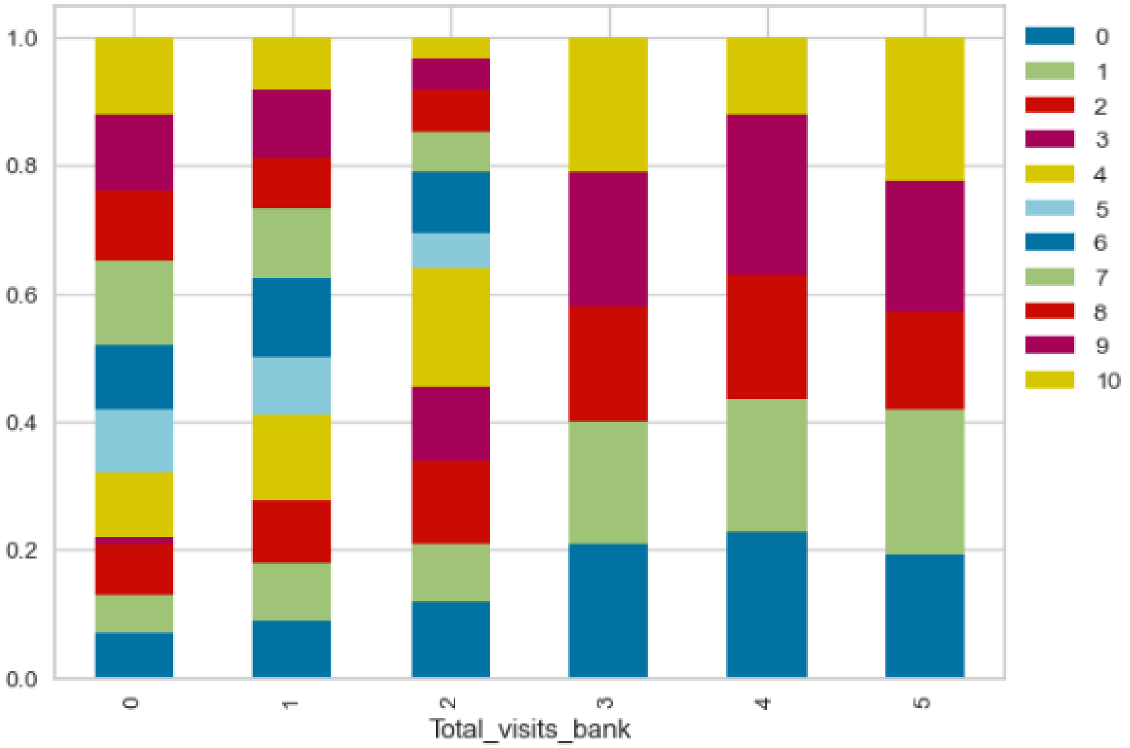


Figure 16 Stacked Bar Plot of Total Visits Bank Vs Total Calls Made

Observation:

1. Distribution of Calls vs. Visits:
   - Each bar represents a different number of bank visits (from 0 to 5).
   - The colored segments within each bar represent the number of calls made (from 0 to 10).
2. Fewer Visits Tend to Have More Variation in Calls:
   - For lower numbers of bank visits (0–2), the distribution of calls made is more varied. We see calls ranging from 0 to 10.
   - As the number of bank visits increases (3–5), the number of calls made appears to narrow down. For instance, with 5 bank visits, most calls made are 0 or 1.
3. High Call Counts Are Less Common:

- The higher call counts (e.g., 9 and 10) appear to be relatively rare across all visit categories, as indicated by smaller colored segments in the bars.

4. Total Calls for Each Visit Group:

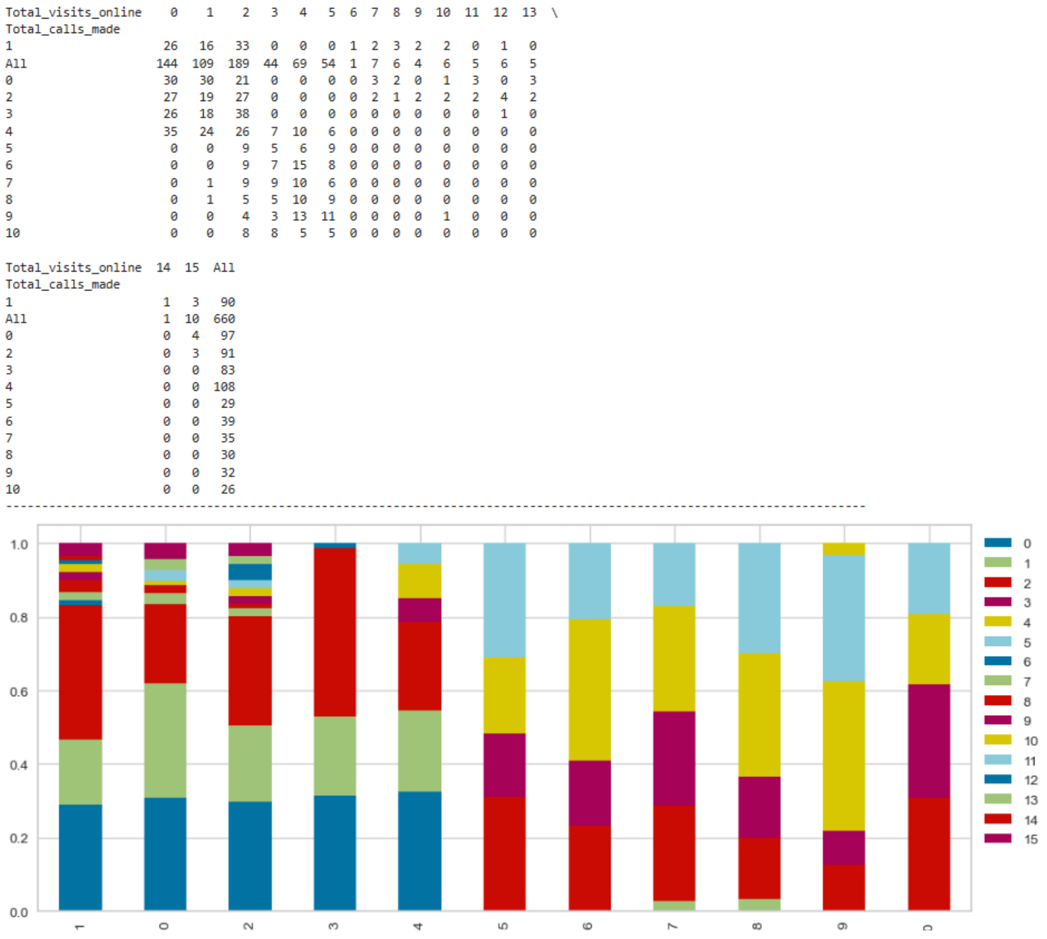
   - The table shows the total number of calls made for each bank visit group. For instance, there are 97 people who made no bank visits and no calls, while there are 19 who made 2 visits and no calls. These totals are reflected in the stacked bar chart.

5. General Trends:

   - There seems to be a pattern where fewer visits are associated with a wider range of calls, while higher visits are associated with fewer or no calls.

This plot suggests a potential relationship where people who visit the bank more frequently may rely less on phone calls and vice versa.

4.2.6 Total_Calls_Made Vs Total_Visits_Online



```
Total_visits_online    0    1    2   3   4   5  6  7  8  9  10  11  12  13  \
Total_calls_made
1                     26   16   33   0   0   0  1  2  3  2   2   0   1   0
All                  144  109  189  44  69  54  1  7  6  4   6   5   6   5
0                     30   30   21   0   0   0  0  3  2  0   1   3   0   3
2                     27   19   27   0   0   0  0  2  1  2   2   2   4   2
3                     26   18   38   0   0   0  0  0  0  0   0   0   1   0
4                     35   24   26   7  10   6  0  0  0  0   0   0   0   0
5                      0    0    9   5   6   9  0  0  0  0   0   0   0   0
6                      0    0    9   7  15   8  0  0  0  0   0   0   0   0
7                      0    1    9   9  10   6  0  0  0  0   0   0   0   0
8                      0    1    5   5  10   9  0  0  0  0   0   0   0   0
9                      0    0    4   3  13  11  0  0  0  0   1   0   0   0
10                     0    0    8   8   5   5  0  0  0  0   0   0   0   0

Total_visits_online  14  15  All
Total_calls_made
1                     1   3   90
All                   1  10  660
0                     0   4   97
2                     0   3   91
3                     0   0   83
4                     0   0  108
5                     0   0   29
6                     0   0   39
7                     0   0   35
8                     0   0   30
9                     0   0   32
10                    0   0   26
------------------------------------------------------------------------------------------
```

Figure 17 Stacked Bar Plot of Total Calls Made Vs Total Visits Online

Observations:

1. Distribution of Calls vs. Online Visits:
   - The bars represent different numbers of online visits (from 0 to 15), while the colored segments in each bar show the distribution of calls made (from 0 to 13+).
2. Fewer Online Visits Have More Variability in Calls Made:
   - For users with fewer online visits (0–3 visits), there is a wider variation in the number of calls made. The segments include a large range of calls from 0 to 13.
   - Users with higher online visits (e.g., 4+ visits) tend to have fewer calls, with some bars entirely dominated by 0–3 calls made.
3. High Online Visits Correspond with Low Calls Made:
   - Users with more than 4 online visits have a large proportion of 0 calls, as shown by the dominant blue segment (representing 0 calls). This suggests that the more online visits a customer makes, the less they tend to rely on phone calls.
4. High Call Counts (10–13) Are Rare:
   - Just like in the previous chart, higher call counts (e.g., 10 or more calls) are quite rare, occurring mostly among users with fewer online visits.
5. Total Calls and Visits (Table):
   - The table above the chart quantifies the number of calls made for each online visit group. For example, 144 people with 0 online visits made only 1 call, while only 1 person with 13 online visits made more than 12 calls.
6. General Trend:
   - There seems to be a pattern where people who visit online frequently are less likely to make calls. Lower online visit numbers are associated with a higher variance in the number of calls made, while higher visit numbers lead to fewer calls.

This suggests that as customers engage more through online platforms, their reliance on phone support diminishes.

4.3 MULTIVARIATE ANALYSIS

4.3.1 Heatmap

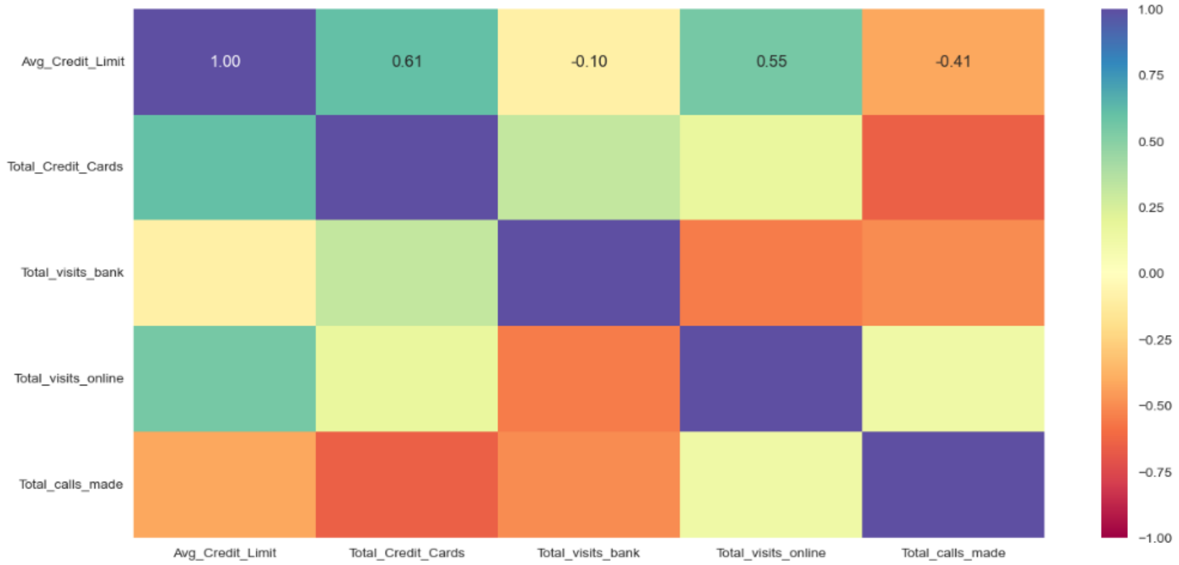


Figure 18 Heat Map of Numerical Variables

Observation:

- Total credit cards and total visit online has medium positive correlation with average credit limit: 0.61,0.55 respectively.
- Total credit cards and total visit bank has medium negative correlation with total calls made: -0.65, -0.5 respectively.
- Total visit online has medium negative correlation with total visit bank: -0.55.

4.3.2 Pairplot

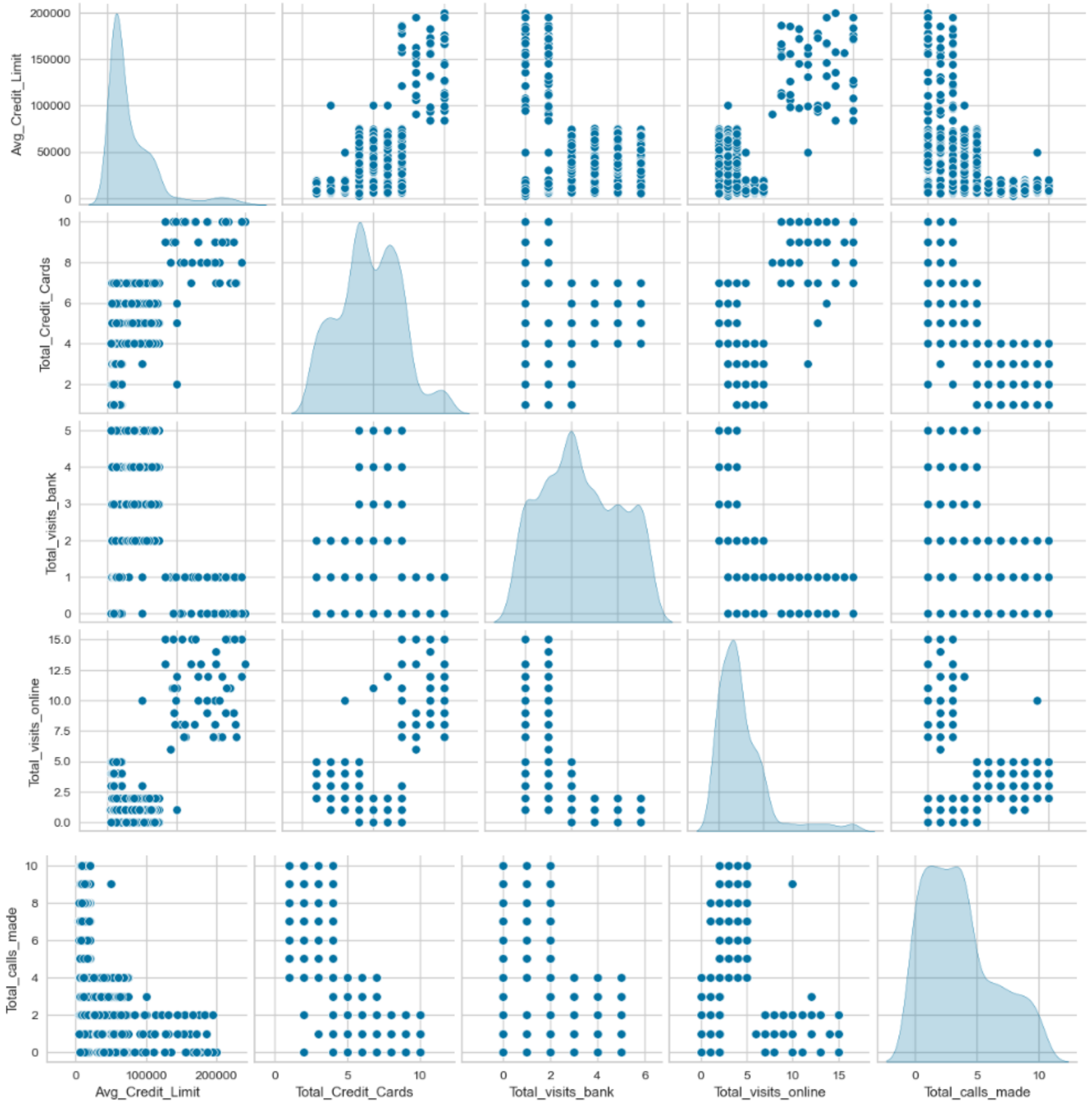


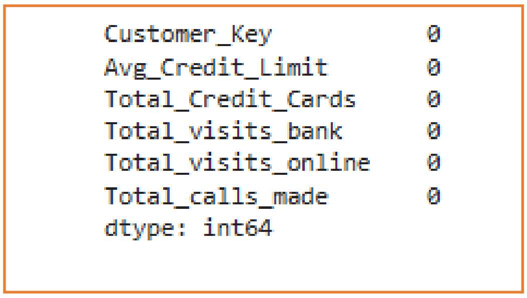
Figure 19 Pair Plot of Numerical Variables

5. DATA PREPROCESSING

5.1 Feature Engineering

Feature engineering involves creating, transforming, or selecting input features that make machine learning models more effective. It is a crucial step to ensure that the data is prepared in a form that maximizes the performance of algorithms.

5.1.1 Missing value Treatment

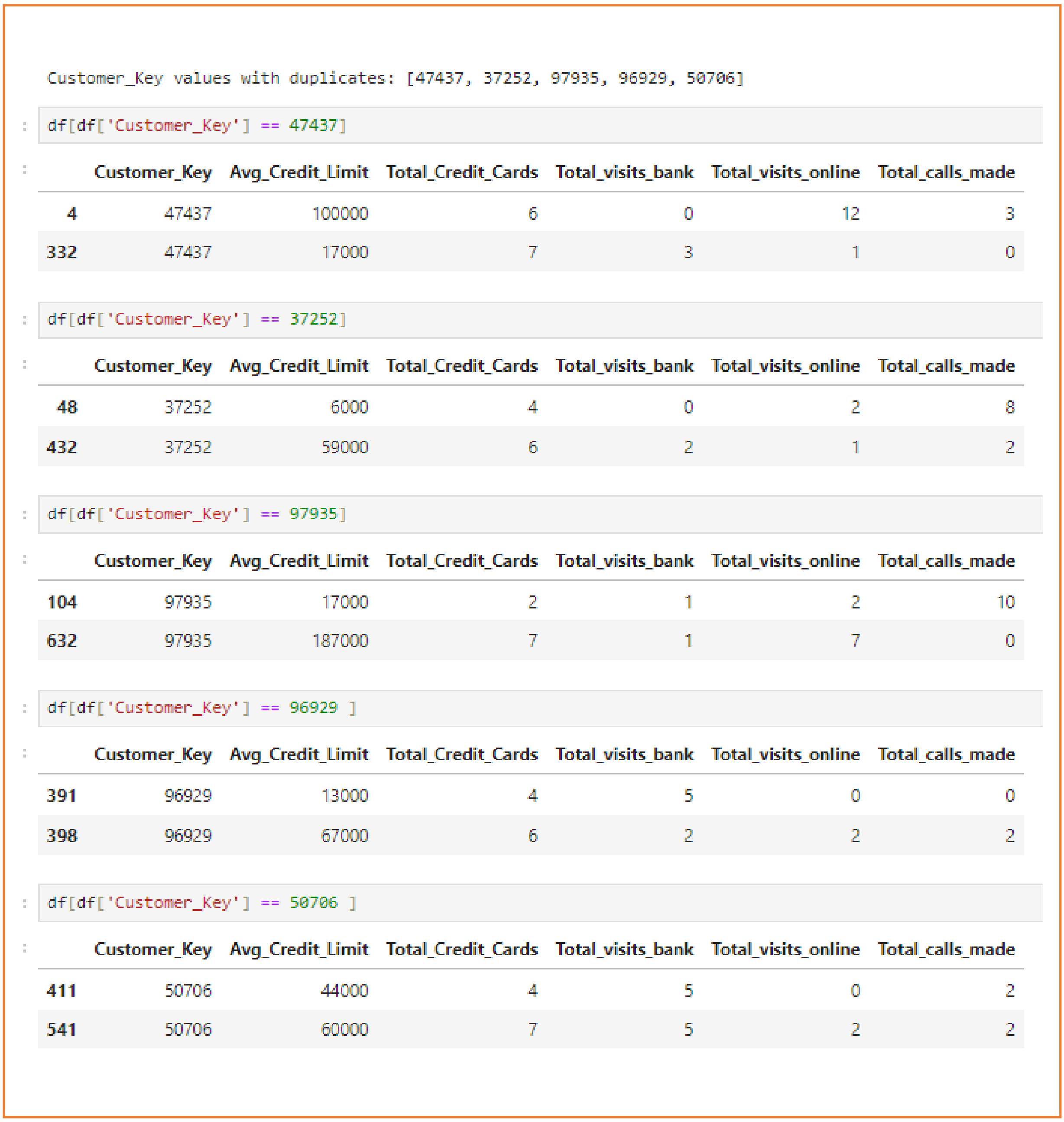


```
Customer_Key           0
Avg_Credit_Limit       0
Total_Credit_Cards     0
Total_visits_bank      0
Total_visits_online    0
Total_calls_made       0
dtype: int64
```

Table 7 Checking Missing Values

- There are no missing values in the dataset.

5.1.2 Duplicate value Treatment



```
Customer_Key values with duplicates: [47437, 37252, 97935, 96929, 50706]
```

```
df[df['Customer_Key'] == 47437]
```

	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
4	47437	100000	6	0	12	3
332	47437	17000	7	3	1	0

```
df[df['Customer_Key'] == 37252]
```

	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
48	37252	6000	4	0	2	8
432	37252	59000	6	2	1	2

```
df[df['Customer_Key'] == 97935]
```

	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
104	97935	17000	2	1	2	10
632	97935	187000	7	1	7	0

```
df[df['Customer_Key'] == 96929 ]
```

	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
391	96929	13000	4	5	0	0
398	96929	67000	6	2	2	2

```
df[df['Customer_Key'] == 50706 ]
```

	Customer_Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
411	50706	44000	4	5	0	2
541	50706	60000	7	5	2	2

Table 8 Checking Duplicate Values

- Dataset Check: Upon checking the entire dataset, no duplicate rows were found.
- Customer key Check:
- However, upon further inspection, we found that there are duplicate values specifically in the Customer key column.
- This indicates that while the overall data rows are unique. The 'Customer Key' is a unique ID given to each customer in the database.

- The duplicate values might correspond to customer profile changes, and as such, there is no need to delete these records as these are actual occurrences at some point in the time.
- The Customer key column can be removed during the analysis.
- We won't need Serial. No. for analysis. Serial no is dropped.

5.1.3 Outlier Detection

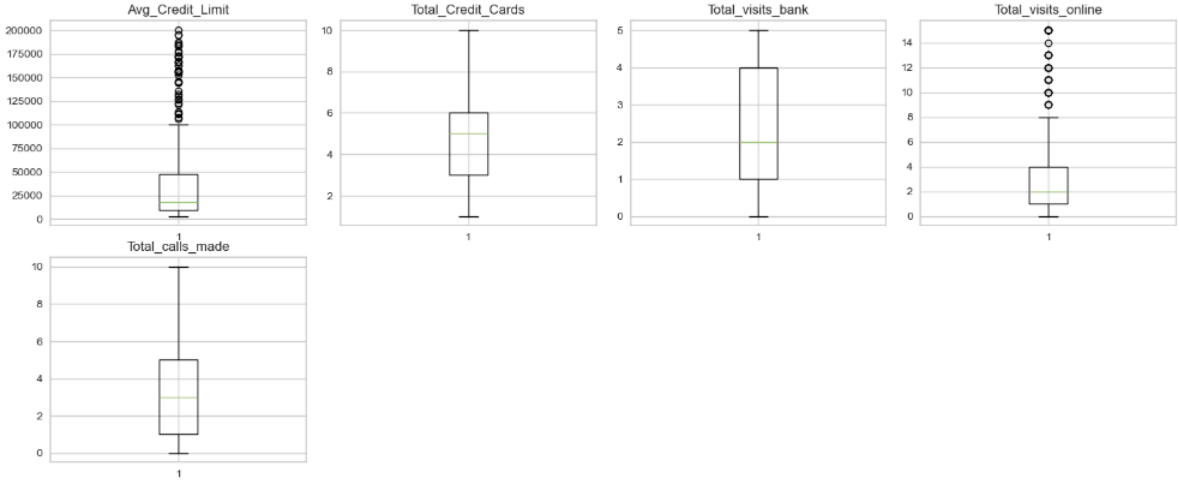
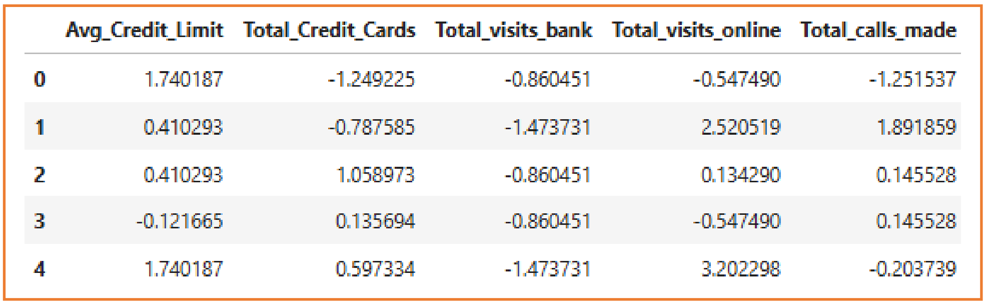


Table 9 Outlier Detection

- There are some outliers in Total visit online and Average credit limit. But these values are genuine, so will need not to treat the outliers.

5.1.4 Data Scaling



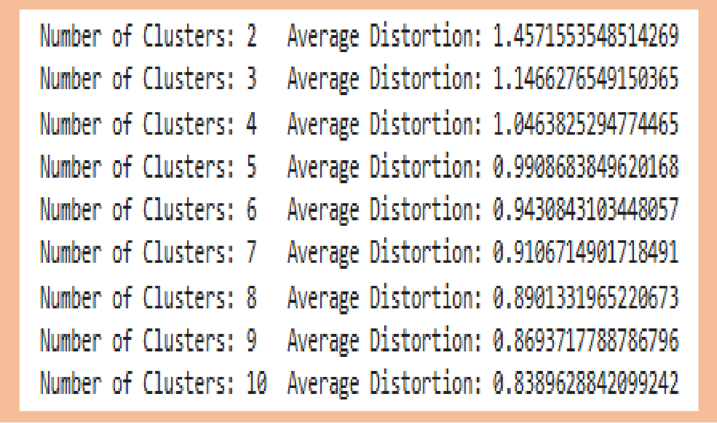
	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	1.740187	-1.249225	-0.860451	-0.547490	-1.251537
1	0.410293	-0.787585	-1.473731	2.520519	1.891859
2	0.410293	1.058973	-0.860451	0.134290	0.145528
3	-0.121665	0.135694	-0.860451	-0.547490	0.145528
4	1.740187	0.597334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

Table 10 Data Scaling

6. MODEL BUILDING

6.1 K - Means Clustering

Applying K-means clustering algorithms



```
Number of Clusters: 2   Average Distortion: 1.4571553548514269
Number of Clusters: 3   Average Distortion: 1.1466276549150365
Number of Clusters: 4   Average Distortion: 1.0463825294774465
Number of Clusters: 5   Average Distortion: 0.9908683849620168
Number of Clusters: 6   Average Distortion: 0.9430843103448057
Number of Clusters: 7   Average Distortion: 0.9106714901718491
Number of Clusters: 8   Average Distortion: 0.8901331965220673
Number of Clusters: 9   Average Distortion: 0.8693717788786796
Number of Clusters: 10  Average Distortion: 0.8389628842099242
```

Table 11 Average Distortion of Clusters

6.2 Apply K-means - Elbow curve

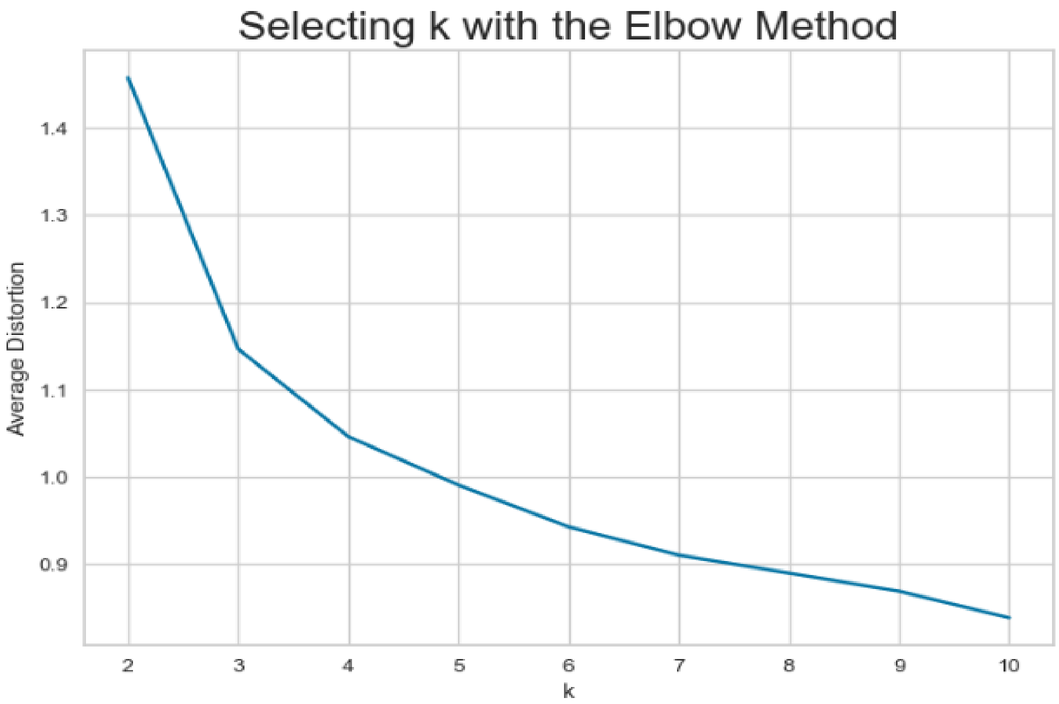


Figure 20 Elbow Curve

6.2.1 Distortion Score Elbow for K – Means Clustering

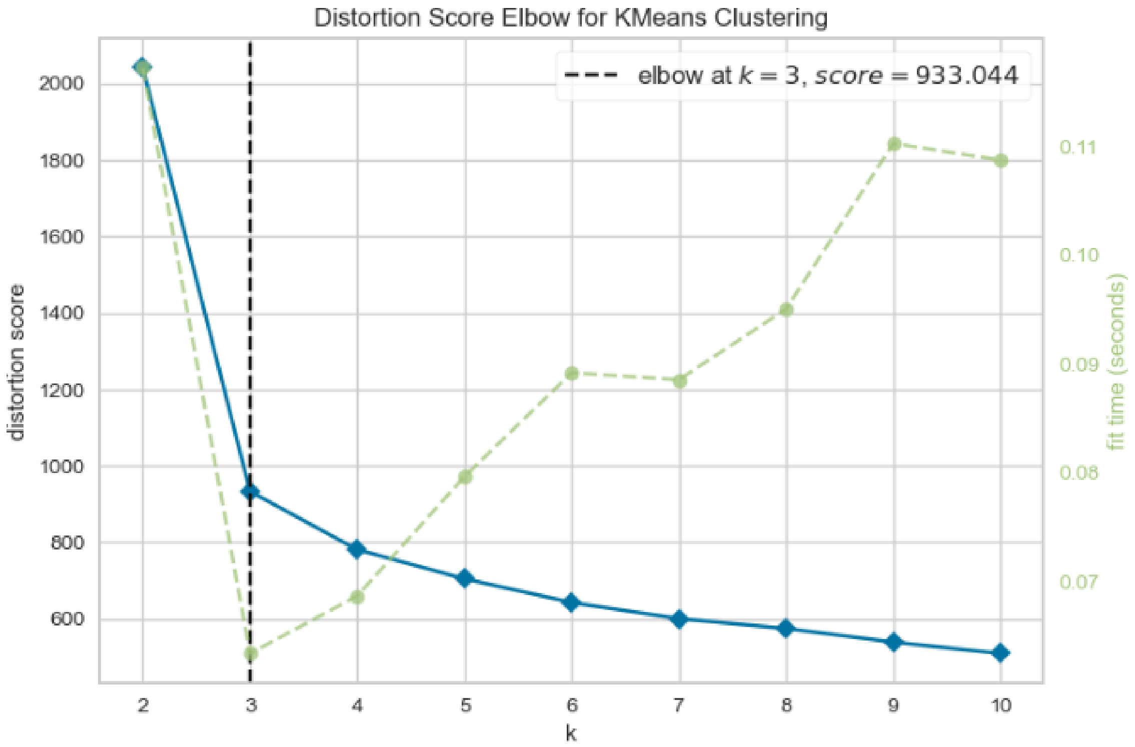
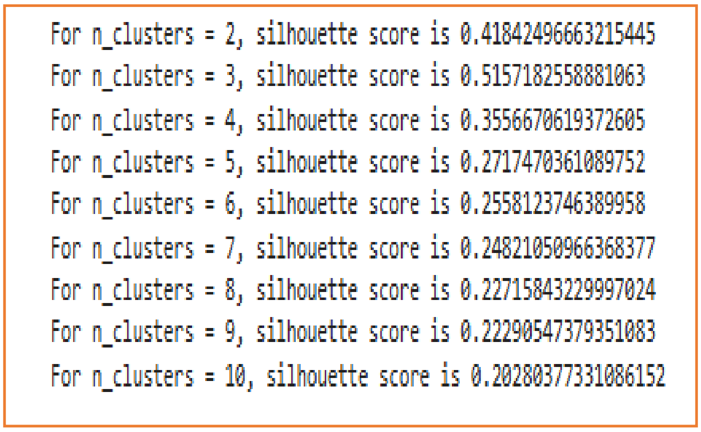


Figure 21 Distortion Score Elbow Curve for K – Means Clustering

The appropriate value of k from the elbow curve seems to be 3

Let's check the silhouette scores

6.3 Silhouette Scores



```
For n_clusters = 2, silhouette score is 0.41842496663215445
For n_clusters = 3, silhouette score is 0.5157182558881063
For n_clusters = 4, silhouette score is 0.3556670619372605
For n_clusters = 5, silhouette score is 0.2717470361089752
For n_clusters = 6, silhouette score is 0.2558123746389958
For n_clusters = 7, silhouette score is 0.24821050966368377
For n_clusters = 8, silhouette score is 0.22715843229997024
For n_clusters = 9, silhouette score is 0.22290547379351083
For n_clusters = 10, silhouette score is 0.20280377331086152
```

Table 12 Silhouette Score

Silhouette score for 3 is the highest which is 0.5157182558881063

6.3.1 Silhouette Scores Plot For K- Means Clustering

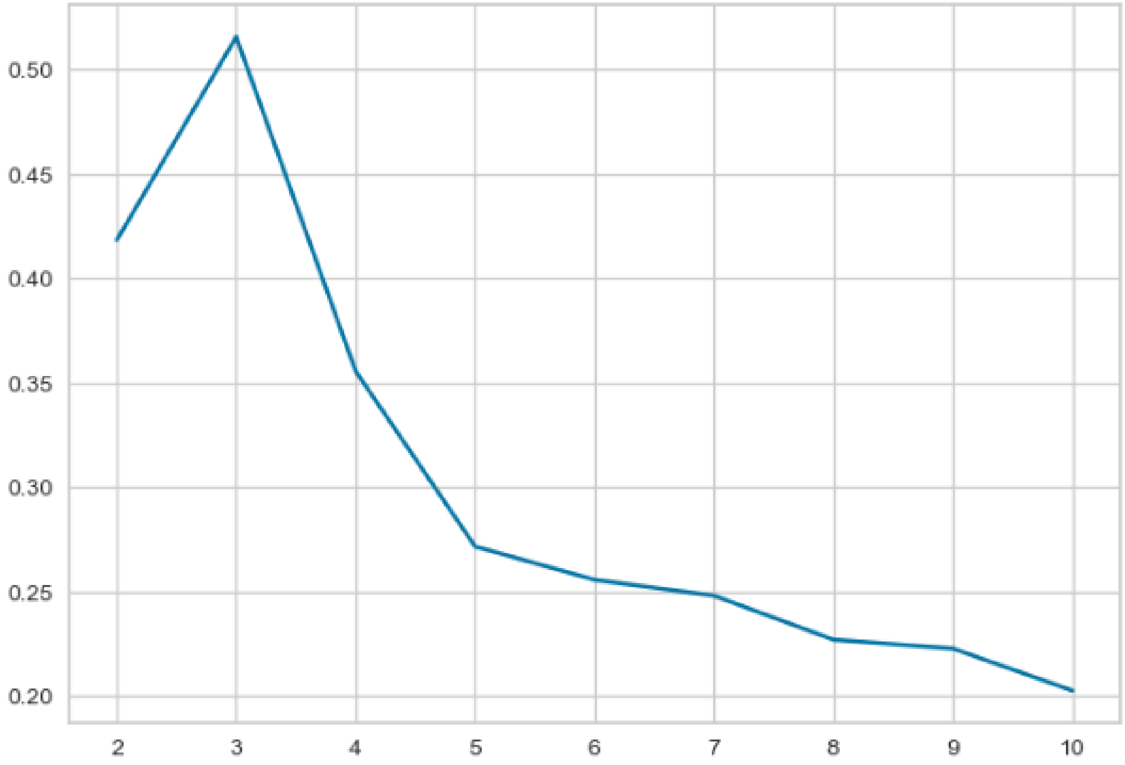


Figure 22 Silhouette Scores Plot For K- Means Clustering

6.3.2 Silhouette Scores Elbow Plot For K- Means Clustering

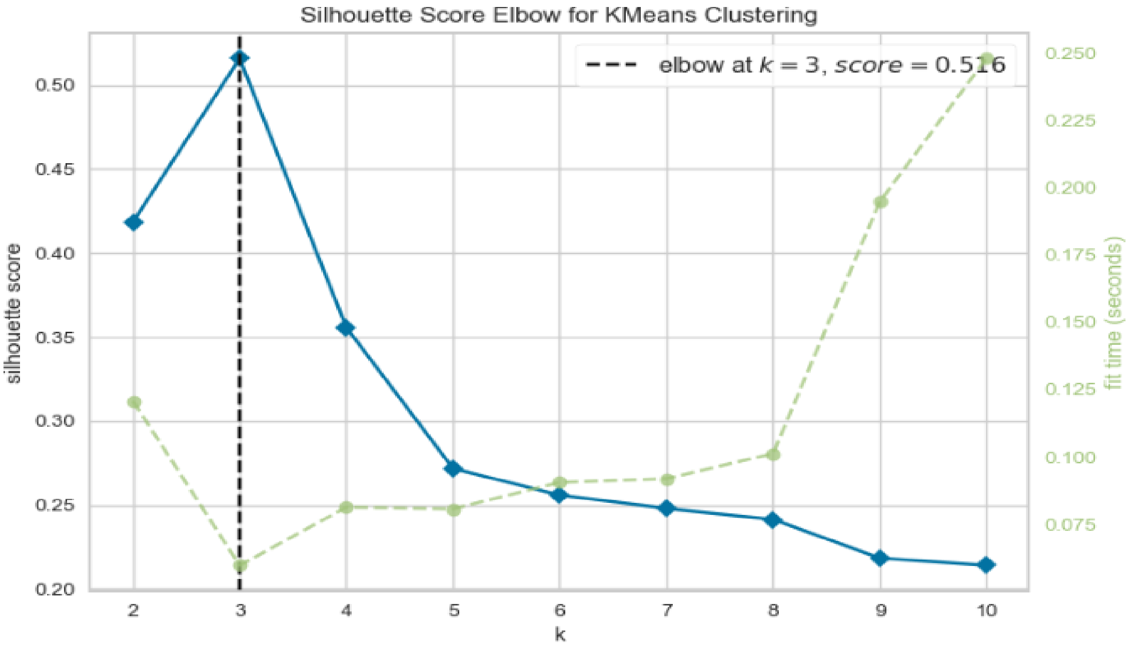


Figure 23 Silhouette Scores Elbow Plot For K- Means Clustering

6.4 Finding Optimal No. Of Clusters with Silhouette Coefficients

6.4.1 Silhouette plot of k means clustering for 660 samples in 2 centres

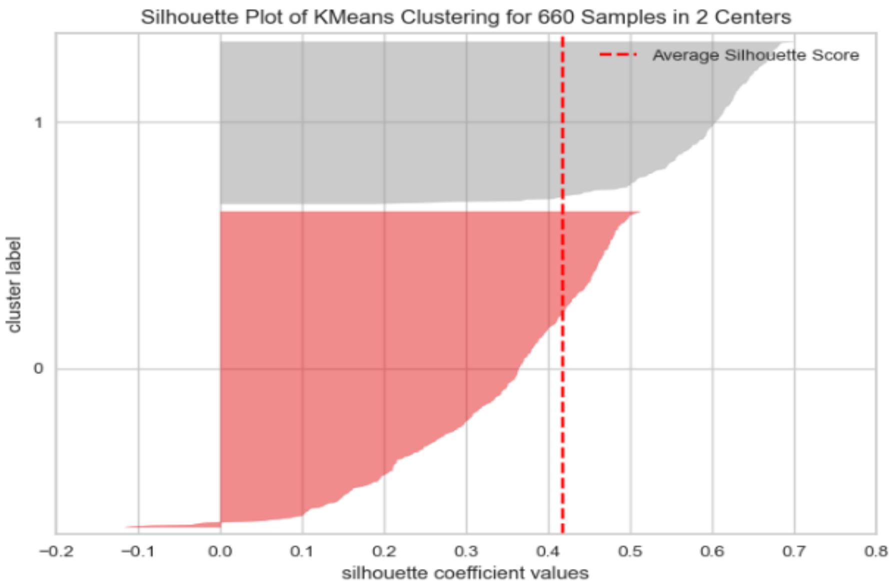


Figure 24 Silhouette plot of k means clustering for 660 samples in 2 centres

6.4.2 Silhouette plot of k means clustering for 660 samples in 3 centres

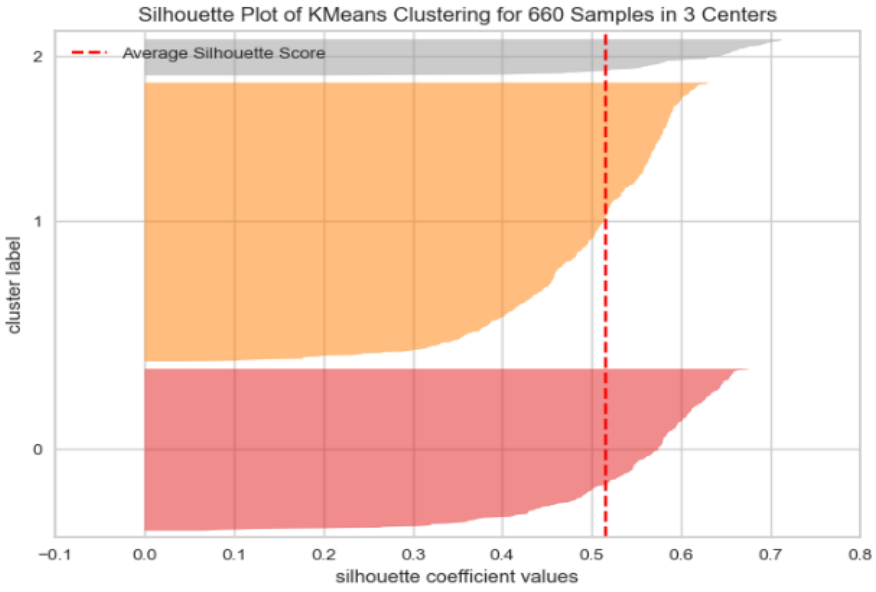


Figure 25 Silhouette plot of k means clustering for 660 samples in 3 centres

6.4.3 Silhouette plot of k means clustering for 660 samples in 4 centres

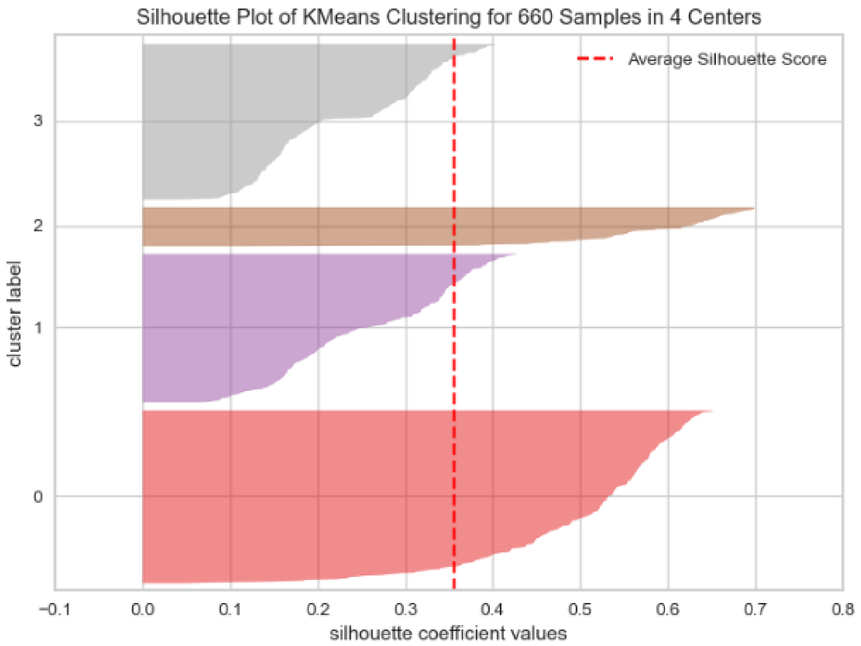


Figure 26 Silhouette plot of k means clustering for 660 samples in 4 centres

6.4.4 Silhouette plot of k means clustering for 660 samples in 5 centres

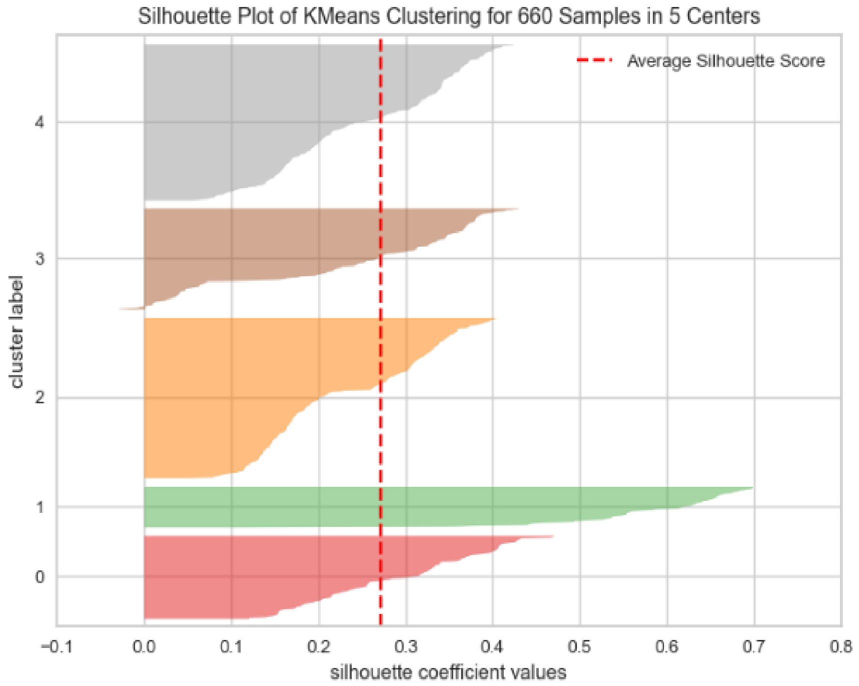


Figure 27 Silhouette plot of k means clustering for 660 samples in 5 centres

6.4.5 Silhouette plot of k means clustering for 660 samples in 6 centres

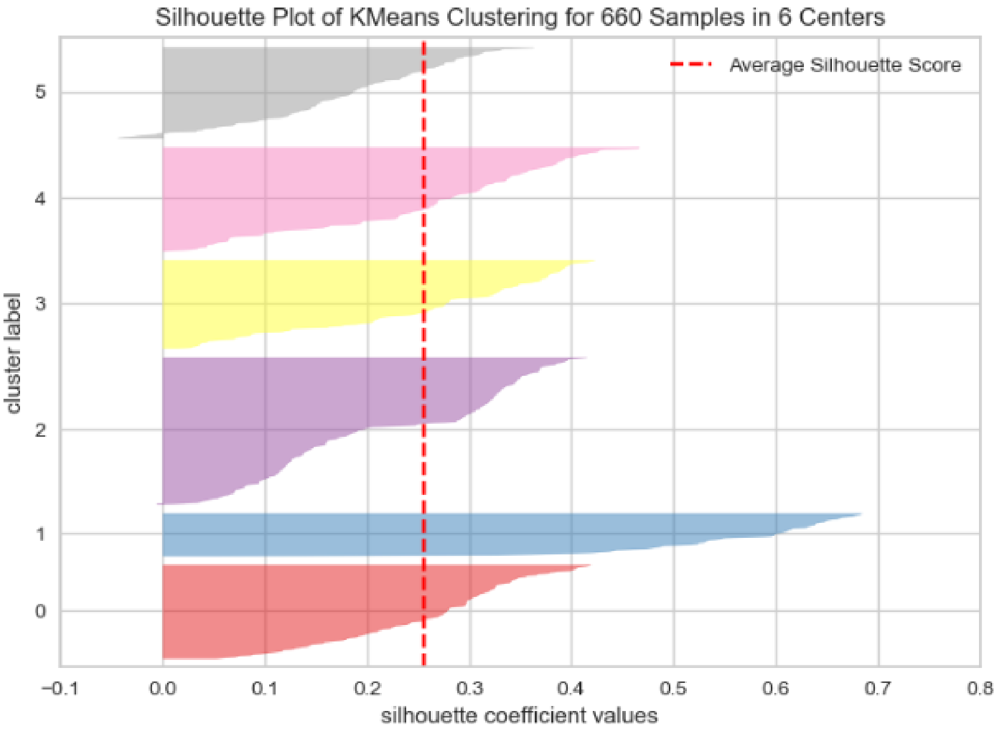


Figure 28 Silhouette plot of k means clustering for 660 samples in 6 centres

6.4.6 Silhouette plot of k means clustering for 660 samples in 7 centres

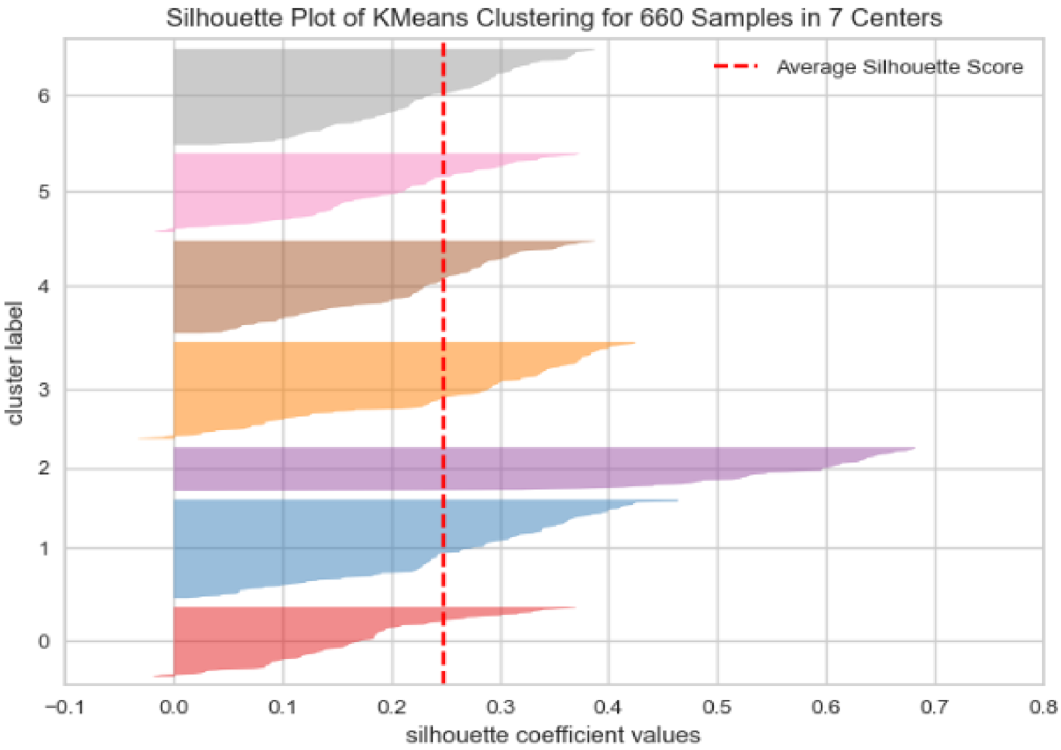
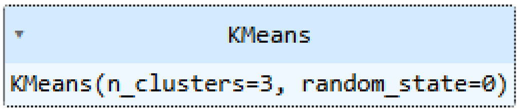


Figure 29 Silhouette plot of k means clustering for 660 samples in 7 centres

Let us take 3 as appropriate no. of clusters as silhoutte score is high enough and there is knick at 3 in elbow curve and there is no outlier.

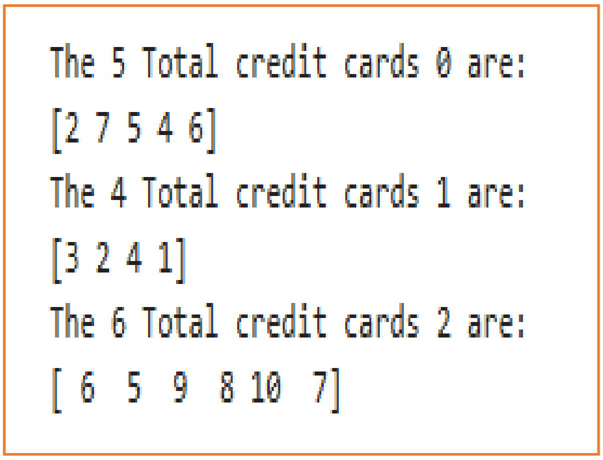
6.5 Selecting final model



```
KMeans(n_clusters=3, random_state=0)
```

6.6 Cluster Profiling

6.6.1 Credit Card Distribution by Cluster



```
The 5 Total credit cards 0 are:
[2 7 5 4 6]
The 4 Total credit cards 1 are:
[3 2 4 1]
The 6 Total credit cards 2 are:
[ 6  5  9  8 10  7]
```

Table 13 Credit Card Distribution by Cluster in K - Means

Observation

Segment 0 (Total Credit Cards = 1)

Customers: [2, 7, 5, 4, 6]

Count: 5 customers

Observation: This segment consists of moderate credit card users, with an average of 1 credit card per customer. These customers may prefer minimal credit exposure or find it easier to manage a single card effectively. They may benefit from targeted offers to encourage broader engagement.

Segment 1 (Total Credit Cards = 0)

Customers: [3, 2, 4, 1]

Count: 4 customers

Observation: This segment has slightly fewer customers and no credit card ownership. These may include newly acquired clients, individuals who have opted out of credit cards, or those with lower credit eligibility. Financial education or incentive programs could help this segment explore credit offerings and build engagement.

Segment 2 (Total Credit Cards = 2)

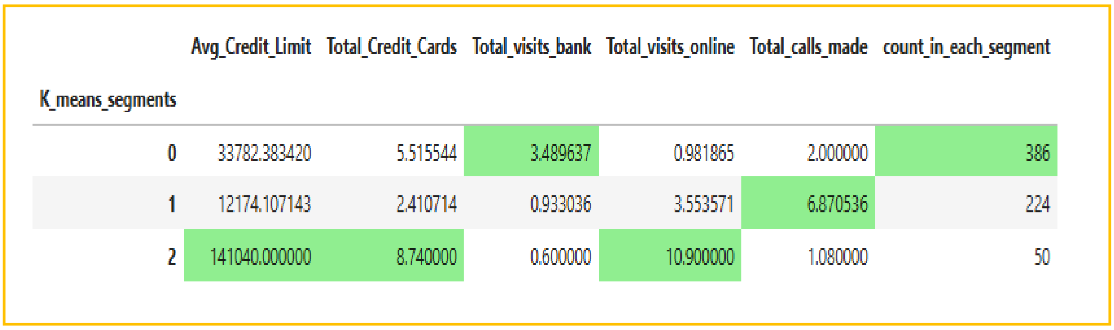
Customers: [6, 5, 9, 8, 10, 7]

Count: 6 customers

Observation: With each customer holding multiple credit cards, this segment displays higher credit engagement. This likely indicates a greater willingness to leverage credit, potentially reflecting higher spending capacity. These customers might appreciate offers centered on rewards, benefits, or enhanced credit services.

Each segment's distinct credit card behavior highlights opportunities for targeted strategies to enhance customer engagement and optimize credit offerings.

6.6.2 Displaying Cluster Profile



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

Table 14 K - Means Cluster Profiling

Observation

- segment 2 consists of high-value customers who are comfortable with online banking and managing multiple credit cards. They are likely to seek premium offerings that match their lifestyle.

6.6.3 Boxplot of Numerical Variables for Each Cluster

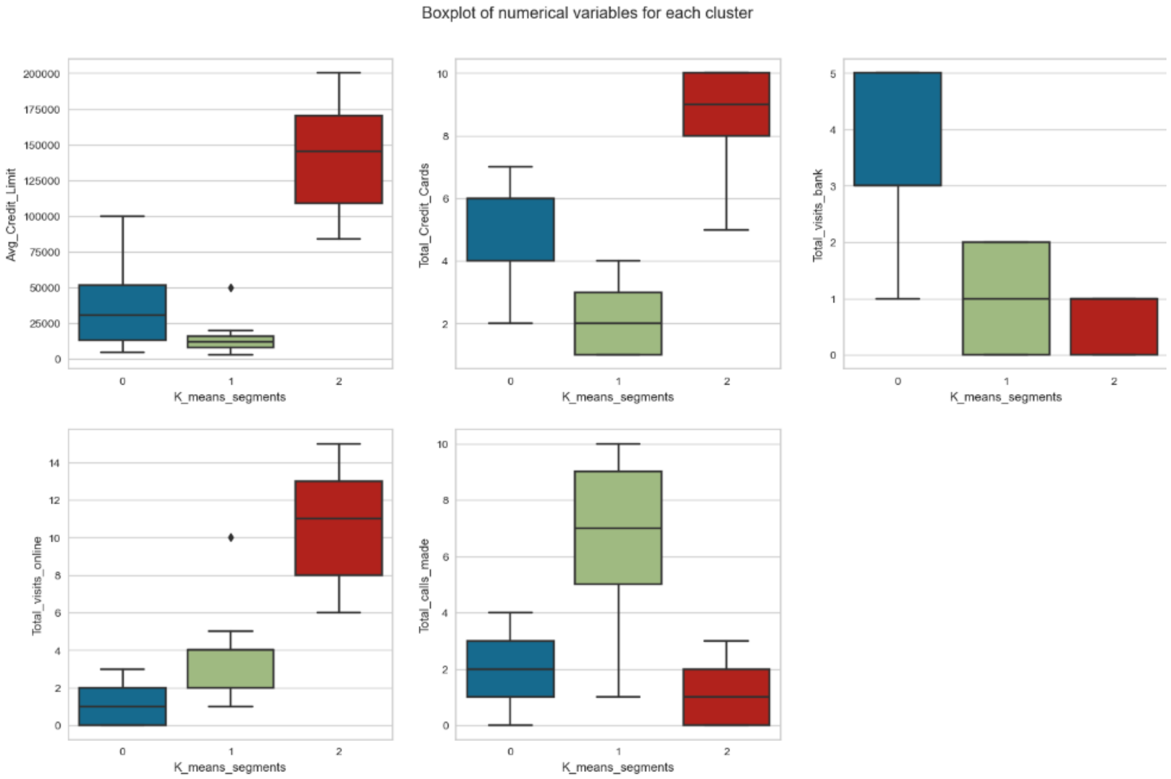


Figure 30 Boxplot for K - Means

6.6.4 Checking the groups for Avg_Credit_Limit

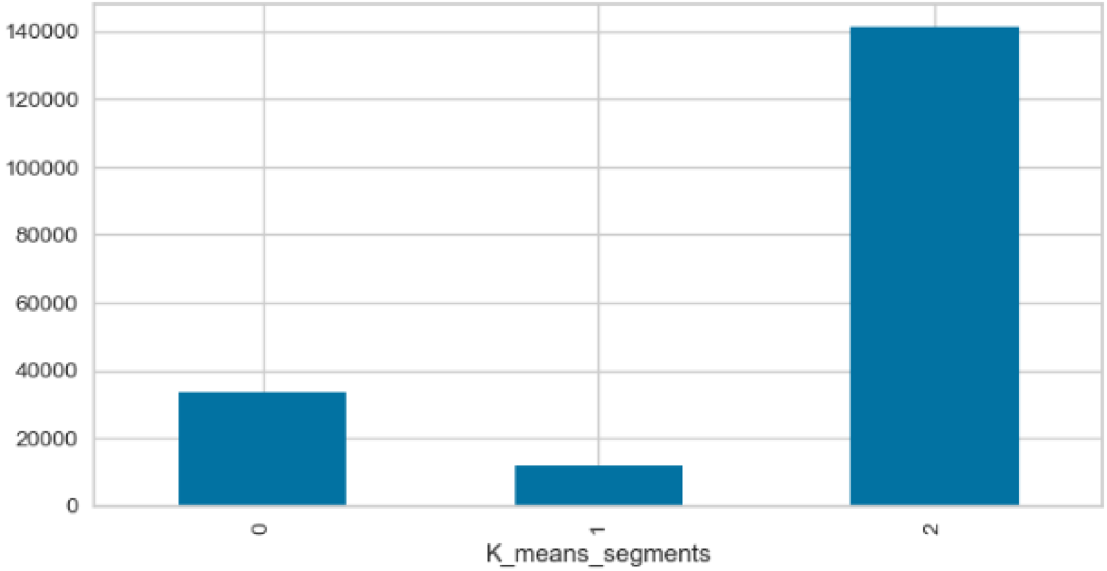


Figure 31 Bar Plot of Average Credit Limit In K - Means Segments

6.6.5 Checking the groups for the remaining features

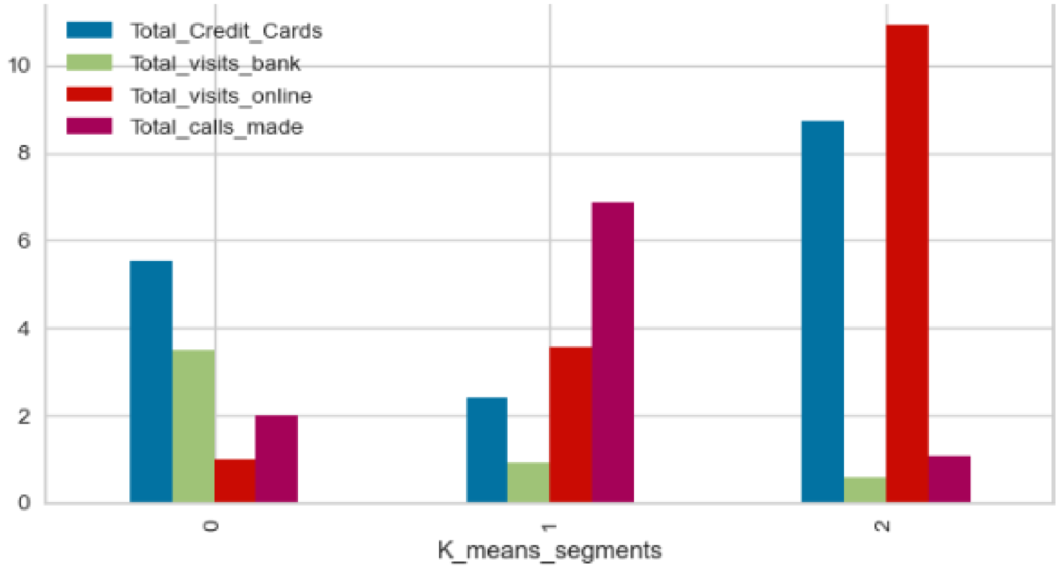


Figure 32 Bar Plot of Discrete Numeric Columns In K Means Segment

6.7 Insight for K – Means

Cluster 0

- Average Credit Limit: $33,782.38 (highest among lower segments)
- Total Credit Cards: 5.52 (moderate)
- Total Visits to Bank: 3.49 (indicates some preference for traditional banking)
- Total Visits Online: 0.98 (low online engagement)
- Total Calls Made: 2.00 (average level of customer support interaction)
- Count in Segment: 386 (largest segment)
- Overall Assessment: Represents a substantial customer base with balanced engagement; reliable revenue source due to size and moderate financial capability.

Cluster 1

- Average Credit Limit: $12,174.11 (lowest)
- Total Credit Cards: 2.41 (lowest)
- Total Visits to Bank: 0.93 (very low traditional banking usage)
- Total Visits Online: 3.55 (highest online engagement)
- Total Calls Made: 6.87 (high call volume indicating need for assistance)
- Count in Segment: 224

- Overall Assessment: Engaged but financially weaker; high reliance on customer support suggests potential areas for improvement in service delivery.

Cluster 2

- Average Credit Limit: $141,040.00 (highest)
- Total Credit Cards: 8.74 (highest)
- Total Visits to Bank: 0.60 (minimal in-person banking)
- Total Visits Online: 10.90 (very high online engagement)
- Total Calls Made: 1.08 (very low, indicating confidence in managing finances)
- Count in Segment: 50 (smallest segment)
- Overall Assessment: High-value customers with significant financial power; small size but high revenue potential; tech-savvy and predominantly prefer digital banking.

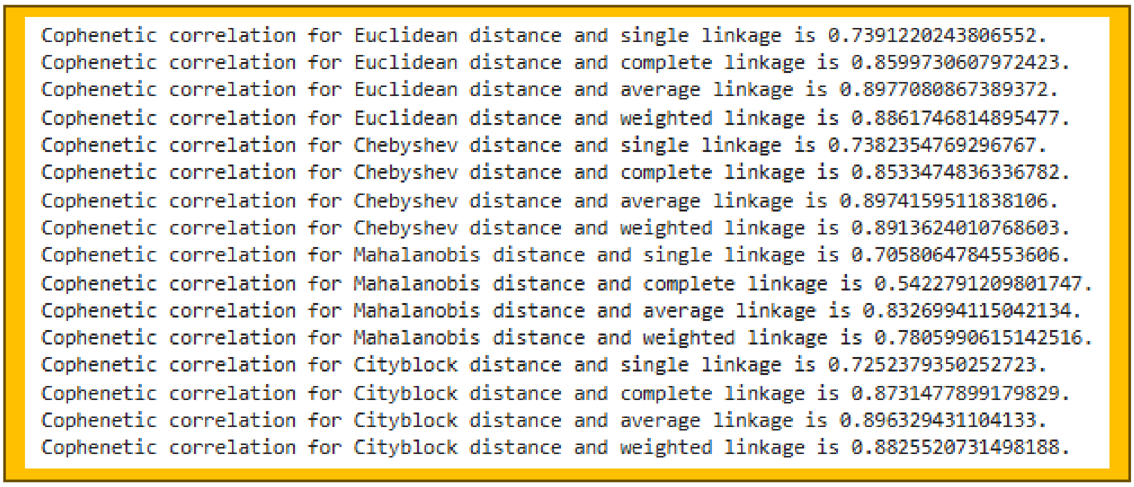
Summary

- Cluster 0: Strong and reliable revenue base, good mix of engagement.
- Cluster 1: Needs support and resources; opportunity to improve financial health.
- Cluster 2: Highly lucrative, tech-savvy, and comfortable with digital banking; focus on retention strategies.

7. HIERARCHICAL CLUSTERING

Apply Hierarchical clustering with different linkage methods and plot dendrograms for each linkage methods

7.1 Computing Cophenetic Correlation

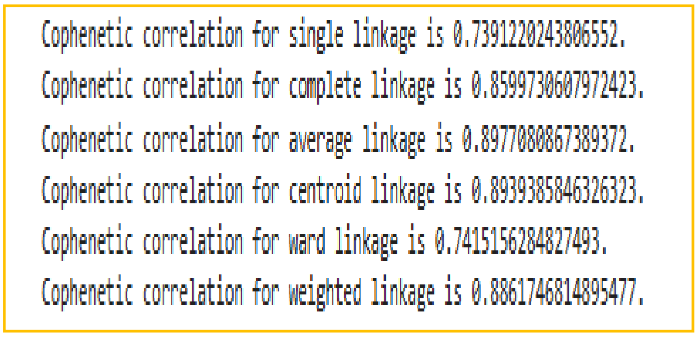


```
Cophenetic correlation for Euclidean distance and single linkage is 0.7391220243806552.
Cophenetic correlation for Euclidean distance and complete linkage is 0.8599730607972423.
Cophenetic correlation for Euclidean distance and average linkage is 0.8977080867389372.
Cophenetic correlation for Euclidean distance and weighted linkage is 0.8861746814895477.
Cophenetic correlation for Chebyshev distance and single linkage is 0.7382354769296767.
Cophenetic correlation for Chebyshev distance and complete linkage is 0.8533474836336782.
Cophenetic correlation for Chebyshev distance and average linkage is 0.8974159511838106.
Cophenetic correlation for Chebyshev distance and weighted linkage is 0.8913624010768603.
Cophenetic correlation for Mahalanobis distance and single linkage is 0.7058064784553606.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.5422791209801747.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.8326994115042134.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.7805990615142516.
Cophenetic correlation for Cityblock distance and single linkage is 0.7252379350252723.
Cophenetic correlation for Cityblock distance and complete linkage is 0.8731477899179829.
Cophenetic correlation for Cityblock distance and average linkage is 0.896329431104133.
Cophenetic correlation for Cityblock distance and weighted linkage is 0.8825520731498188.
```

Table 15 Cophenetic Correlation for different distance and different linkage

Highest cophenetic correlation is 0.8977080867389372, which is obtained with Euclidean distance and average linkage.

Let's explore different linkage methods with Euclidean distance only.



```
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.
```

Table 16 Cophenetic Correlation for Different Linkage Methods With Euclidean Distance

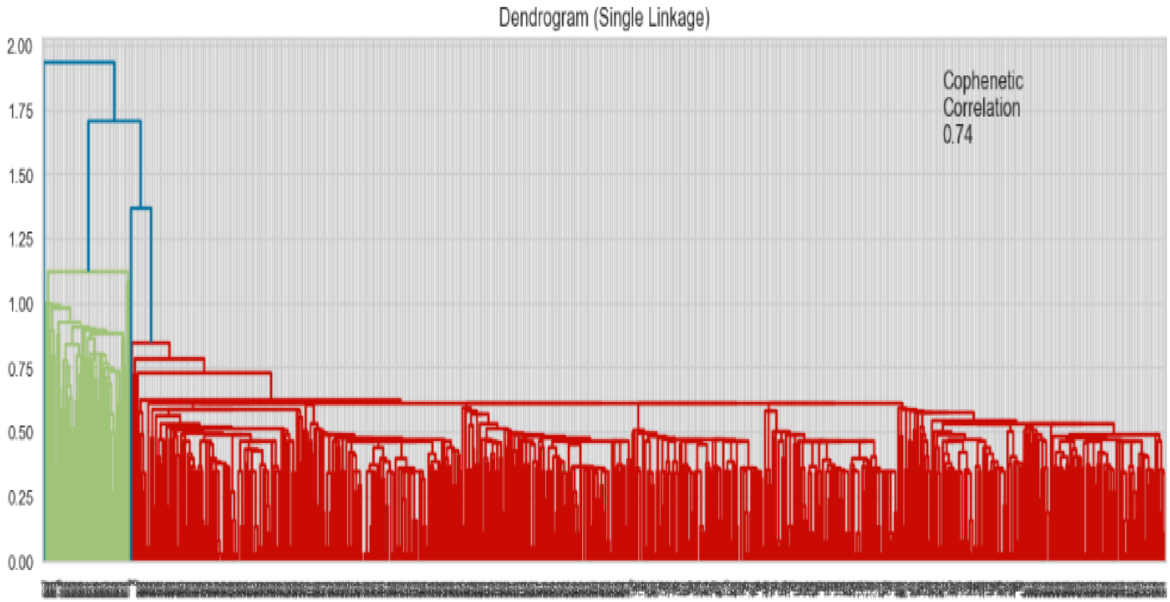
Highest cophenetic correlation is 0.8977080867389372, which is obtained with average linkage.

Observations

- We see that the cophenetic correlation is maximum with Euclidean distance and average linkage.

7.2 Checking Dendrograms

7.2.1. Single Linkage Dendrogram

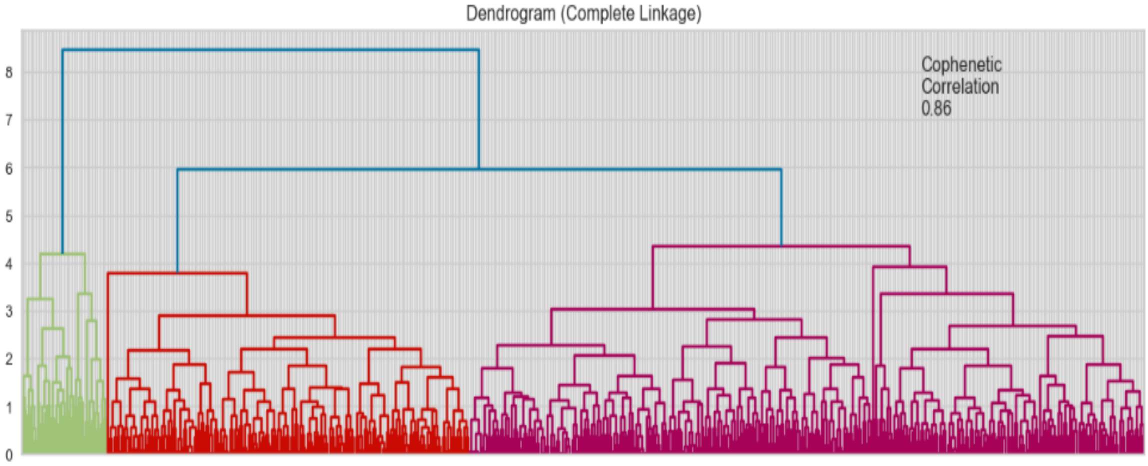


Observation

- Structure: This dendrogram is more condensed with tighter clusters in some regions, especially near the bottom, with outliers causing the larger branches at the top.
- Cophenetic Correlation: 0.74 indicates how well the dendrogram preserves the pairwise distances between original data points.

- The clusters form rapidly at the lower levels of the dendrogram, with a large number of smaller clusters.
- Outliers cause the larger, more isolated branches at the top.
- Some clusters merge early, meaning they have lower similarity with neighboring clusters, resulting in long branches for outliers.

7.2.2. Complete Linkage Dendrogram



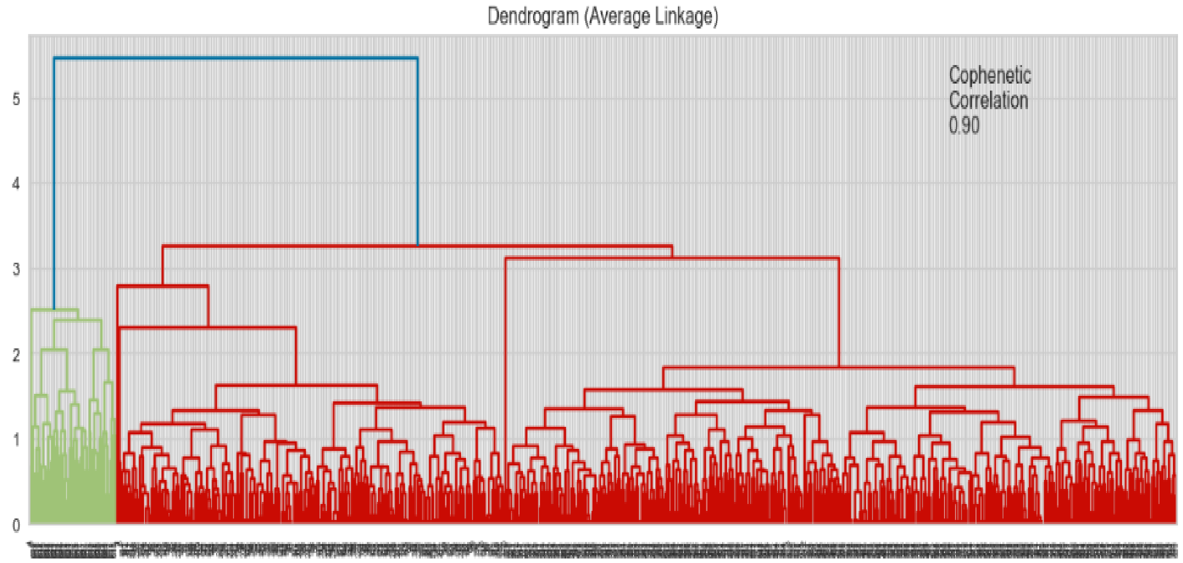
Observation

- Structure: The complete linkage method results in more distinct, evenly distributed clusters compared to single linkage. Larger groups of objects are being merged more progressively.

- Cophenetic Correlation: 0.86, higher than the single linkage dendrogram, indicates better preservation of pairwise distances.

- The clusters are more evenly distributed and merge more gradually.
- The upper branches are spaced further apart, showing that the clusters remain separate for a longer time before merging.
- Fewer large clusters compared to single linkage, indicating that complete linkage tends to avoid chaining effects seen in single linkage.

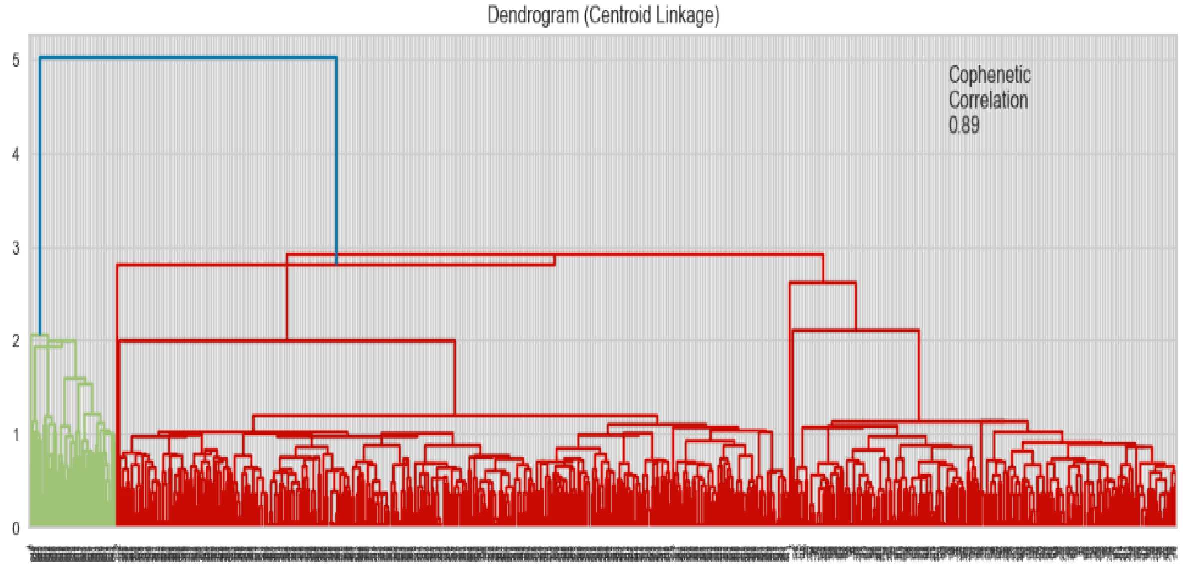
7.2.3. Average Linkage Dendrogram



Observations:

- Structure: This dendrogram shows a more balanced and gradual merging of clusters, with some larger groups merging at higher levels.
- Cophenetic Correlation: 0.90, the highest seen so far, indicates strong preservation of pairwise distances, suggesting a good fit for the data.
- The clustering process is more evenly spread out, with branches merging progressively as one moves higher up the tree.
- There are fewer sudden large jumps between clusters, which means that the clustering is more stable across levels.
- Clusters are merging at moderate heights (around 2-3), suggesting that clusters are more similar in size and composition compared to methods like single or complete linkage.

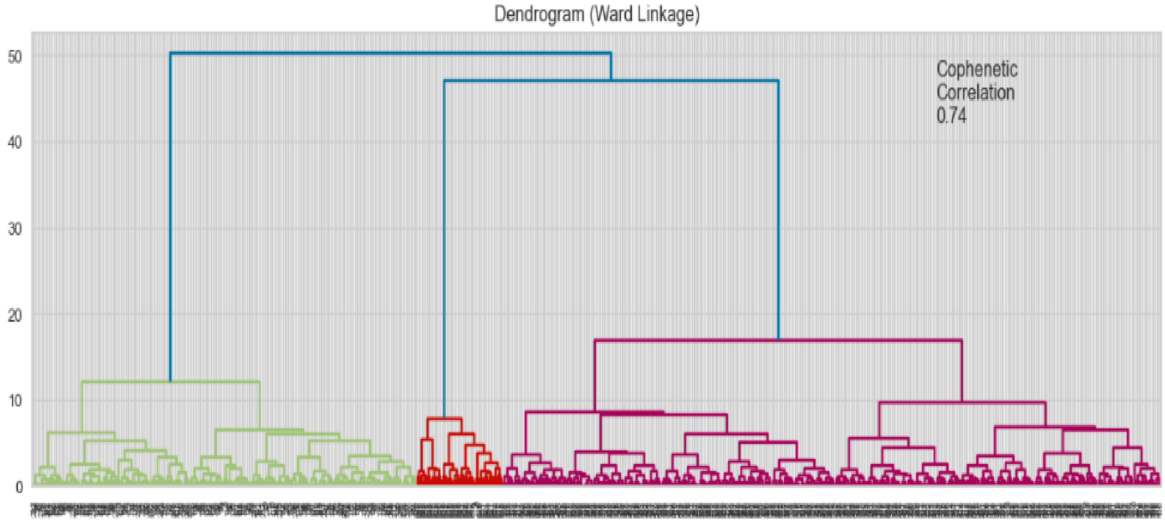
7.2.4. Centroid Linkage Dendrogram



Observations:

- Structure: The centroid linkage dendrogram shows a pattern similar to average linkage but is slightly more compact.
- Cophenetic Correlation: 0.89, very close to average linkage, indicating a strong preservation of pairwise distances.
- The clusters form gradually, and large groupings occur around the same height (around 2-3) as in the average linkage dendrogram.
- There are larger clusters forming earlier compared to average linkage, indicating that centroid linkage might slightly favor creating larger groups earlier.
- The pattern of branching is similar to average linkage but is slightly tighter, meaning clusters merge more quickly.

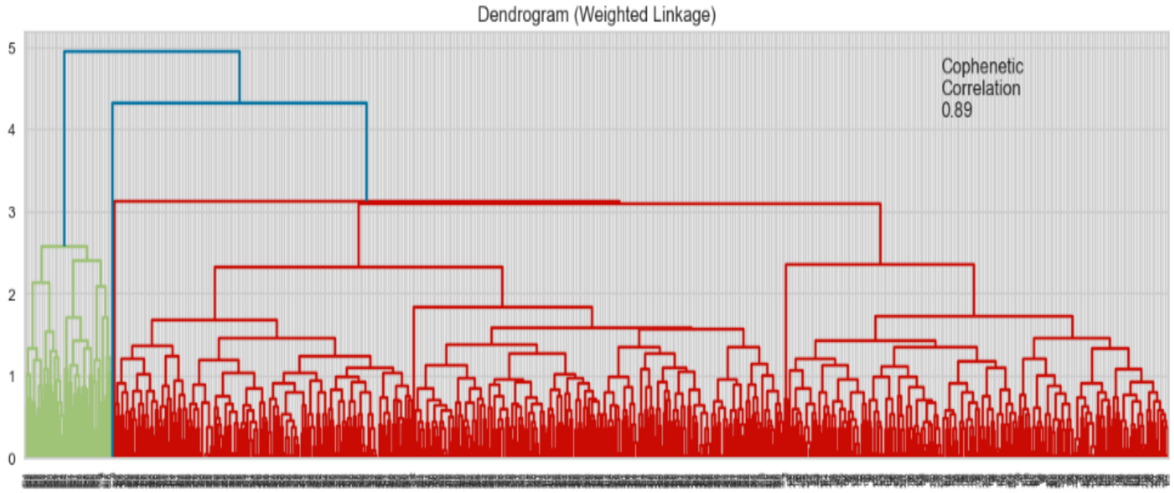
7.2.5. Ward Linkage Dendrogram



Observation

- Structure: The Ward linkage dendrogram shows larger clusters forming at very high levels, with fewer branches compared to other linkage methods.
- Cophenetic Correlation: 0.74, relatively lower than the other methods seen, indicating less effective preservation of pairwise distances.
- Clusters are merging at much higher heights, suggesting Ward's method prioritizes minimizing variance within clusters, thus forming fewer but larger clusters.
- There are distinct large branches, with the height difference between major merges quite prominent. This shows Ward's linkage emphasizes compact, spherical clusters.
- The larger distances at which clusters merge (up to 50) indicate a tendency to merge data points into larger groups only when necessary.

7.2.6 Weighted Linkage Dendrogram



Observation

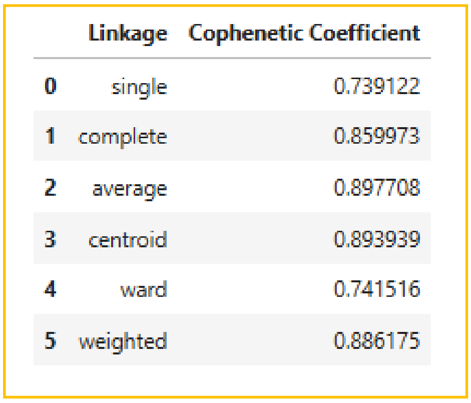
- Structure: The weighted linkage dendrogram shows more gradual merging of clusters, with a cophenetic correlation of 0.89, indicating a strong ability to preserve pairwise distances.
- Cophenetic Correlation: 0.89, which is close to centroid and average linkage, indicating better representation of the true structure of the data.
- The clustering is more evenly distributed, with gradual and progressive merging of smaller clusters.
- The weighted linkage provides a balanced approach between forming large clusters early and maintaining small clusters for longer, showing a better mix of clusters sizes.
- The structure resembles centroid or average linkage in that the clusters are smaller and more cohesive, merging at moderate heights (around 2-3).

Summary:

From the six dendrogram plots,

- Ward linkage (top plot of the last image) shows the most distinct and separate clusters.
- This is because Ward's method tends to minimize the variance within clusters, leading to more compact and clearly defined clusters.
- In the Ward linkage dendrogram, clusters are formed at much higher levels, and the separation between them is more pronounced.

let's create a dataframe to compare cophenetic correlations for each linkage method



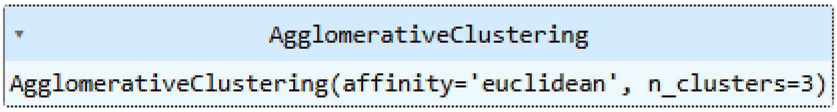
	Linkage	Cophenetic Coefficient
0	single	0.739122
1	complete	0.859973
2	average	0.897708
3	centroid	0.893939
4	ward	0.741516
5	weighted	0.886175

Table 17 cophenetic correlations for each linkage method

Observations

- Out of all the dendrograms we saw, it is clear that the dendrogram with ward linkage gave us separate and distinct clusters.
- 3 would be the appropriate number of the clusters from the dendrogram with ward linkage method.

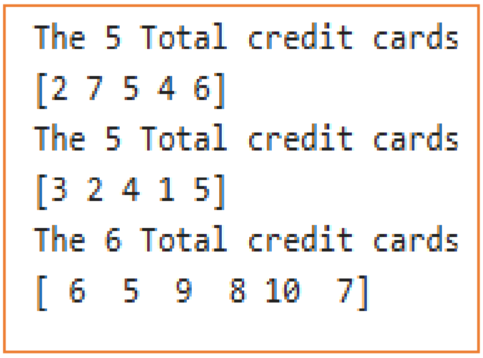
7.3 Creating Final Model



```
AgglomerativeClustering
AgglomerativeClustering(affinity='euclidean', n_clusters=3)
```

7.4 Cluster Profiling

7.4.1 Credit Card Distribution by Cluster



```
The 5 Total credit cards
[2 7 5 4 6]
The 5 Total credit cards
[3 2 4 1 5]
The 6 Total credit cards
[ 6  5  9  8 10  7]
```

Table 18 Credit Card Distribution by Cluster in HC

Cluster 0

Total Credit Cards: 5

Customers: [2, 7, 5, 4, 6]

Observation: Customers in Cluster 0 hold multiple credit cards, indicating moderate engagement and a willingness to explore various credit options. They may be interested in offers that maximize the benefits of their credit usage.

Cluster 1

Total Credit Cards: 5

Customers: [3, 2, 4, 1, 5]

Observation: While customers in Cluster 1 also have multiple credit cards, they show a more conservative approach to credit. This segment may benefit from educational resources to promote responsible credit use and further engagement with credit products.

Cluster 2

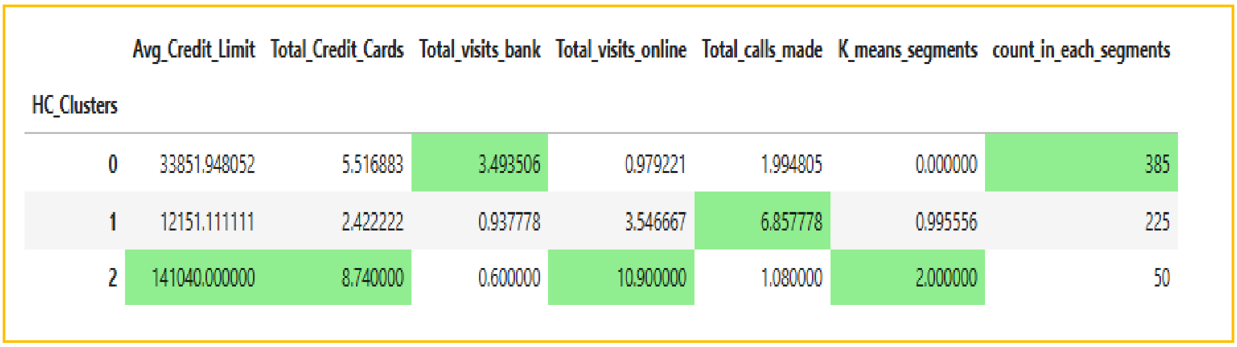
Total Credit Cards: 6

Customers: [6, 5, 9, 8, 10, 7]

Observation: Cluster 2 includes the highest number of credit card holders, reflecting strong engagement and experience with credit products. These customers may value premium services and exclusive offerings tailored to their active credit management habits.

This version highlights each cluster's credit behavior more succinctly, emphasizing targeted opportunities.

7.4.2 Displaying Cluster Profile



	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	K_means_segments	count_in_each_segments
HC_Clusters							
0	33851.948052	5.516883	3.493506	0.979221	1.994805	0.000000	385
1	12151.111111	2.422222	0.937778	3.546667	6.857778	0.995556	225
2	141040.000000	8.740000	0.600000	10.900000	1.080000	2.000000	50

Table 19 Displaying Hierarchical Cluster Profile

Observation

The analysis of each cluster reveals distinct patterns in customer behavior and preferences for credit usage and banking interactions:

- Cluster 0: Customers in this cluster show a balanced approach, using both in-person and online banking channels. They may benefit from services that integrate these experiences, such as hybrid advisory options.
- Cluster 1: These customers prefer digital services but may need additional support to increase confidence in online interactions. Targeted educational resources and user-friendly digital tools could improve their engagement.

Cluster 2: Highly engaged and confident in online banking, these customers present strong opportunities for premium services and customized financial products that align with their digital-first lifestyle.

7.4.3 Boxplot of Numerical Variables for Each Cluster

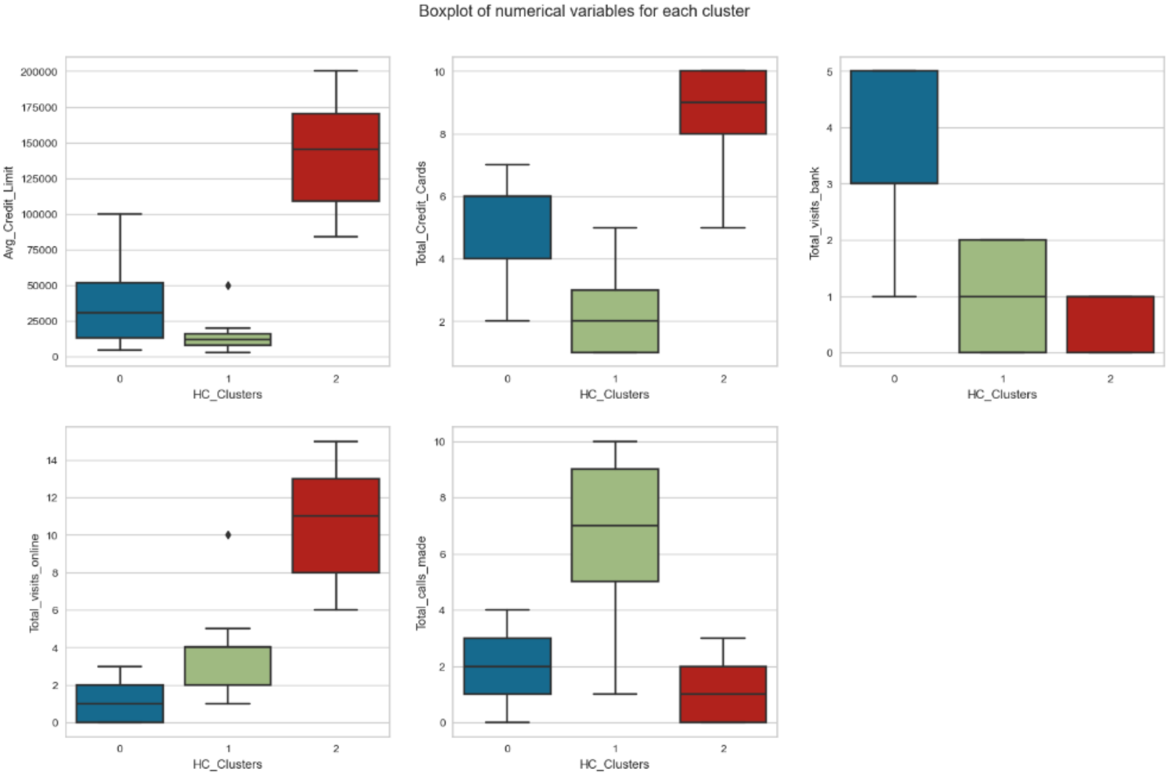


Figure 33 Boxplot for HC

7.4.4 Checking the groups for Avg_Credit_Limit

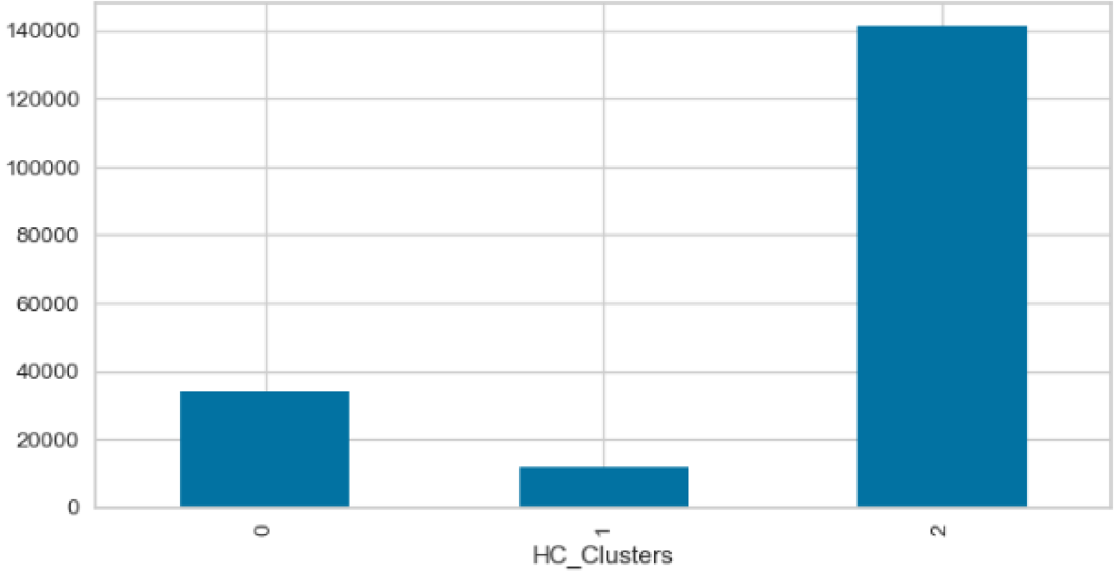


Figure 34 Bar Plot of Average Credit Limit In HC

7.4.5 Checking the groups for the remaining features

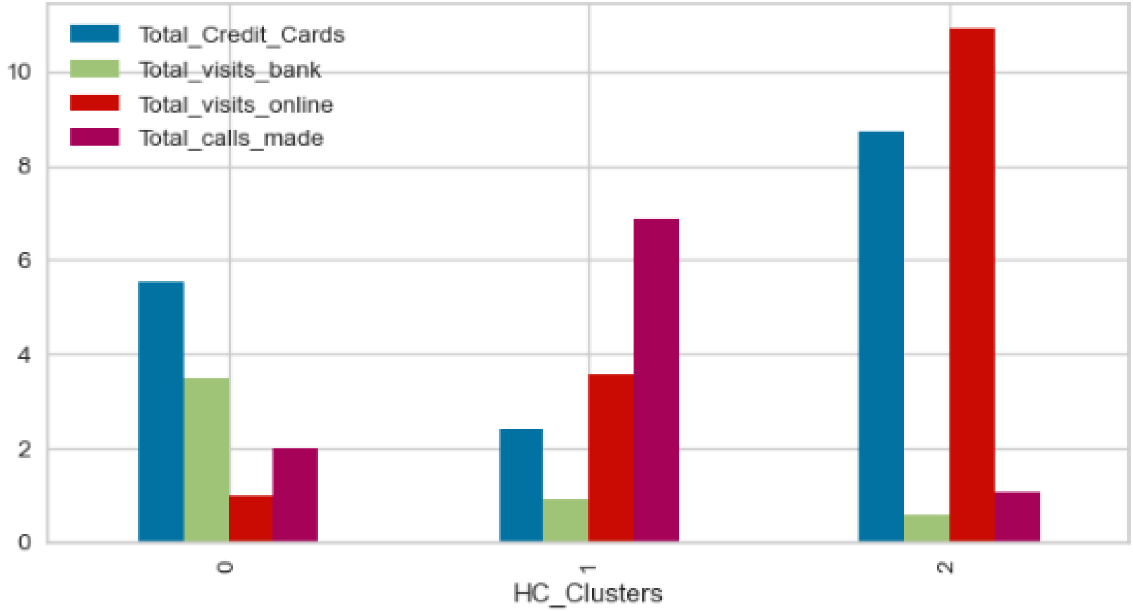


Figure 35 Bar Plot of Discrete Numeric Columns In HC

7.5 Insight for Hierarchial Clustering

HC Cluster Observations

HC Cluster 0

- Average Credit Limit: $33,851.95
- Total Credit Cards: 5.52 (moderate)
- Total Visits to Bank: 3.49 (indicates some preference for in-person banking)
- Total Visits Online: 0.98 (low online engagement)
- Total Calls Made: 1.99 (average level of customer support interaction)
- Count in Segment: 385 (largest cluster)

Observation:

This segment consists of a large customer base with moderate financial capability and balanced banking habits, favoring in-person visits over online interactions. It's a reliable source of revenue due to its size and consistent engagement.

HC Cluster 1

- Average Credit Limit: $12,151.11 (lowest)
- Total Credit Cards: 2.42 (fewest cards)
- Total Visits to Bank: 0.94 (low traditional banking usage)
- Total Visits Online: 3.55 (high online engagement)

- Total Calls Made: 6.86 (high, indicating frequent need for assistance)
- Count in Segment: 225

Observation:

This segment has the lowest financial capability but shows high customer support engagement through calls and online visits. They may require additional resources or support, suggesting an opportunity to enhance service and increase their financial empowerment over time.

HC Cluster 2

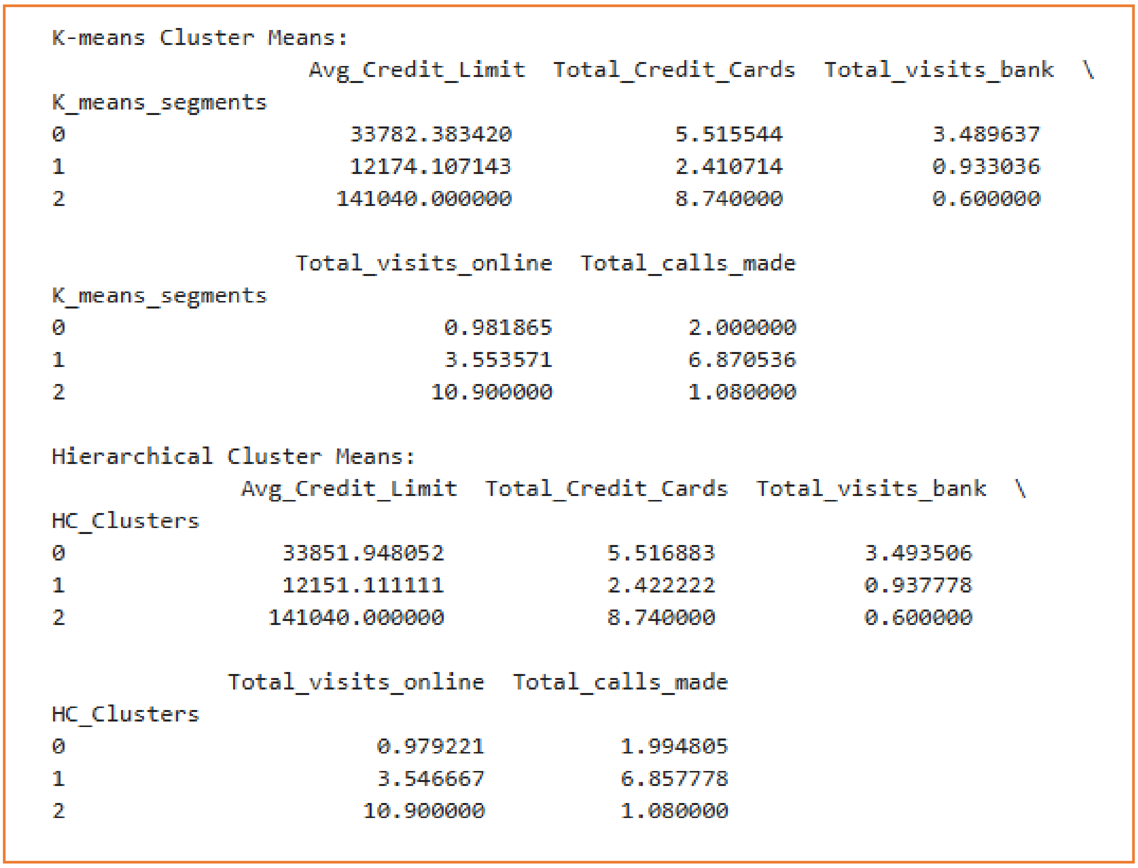
- Average Credit Limit: $141,040.00 (highest)
- Total Credit Cards: 8.74 (highest number of cards)
- Total Visits to Bank: 0.60 (minimal in-person banking)
- Total Visits Online: 10.90 (very high online engagement)
- Total Calls Made: 1.08 (very low, indicating confidence in digital banking)
- Count in Segment: 50 (smallest cluster)
- Observation: This cluster includes high-value, financially empowered customers who rely heavily on online banking. They require minimal support and are likely independent in managing their finances digitally. Despite their small size, they have significant revenue potential and would benefit from targeted loyalty and retention programs.

Summary of HC Clusters

- HC Cluster 0: A large segment with a balanced approach to banking, providing steady revenue.
- HC Cluster 1: Engaged but financially less capable, with high support needs. Potential for growth through increased support and tailored financial products.
- HC Cluster 2: The most financially valuable and independent segment; strong potential for profitability through digital services and loyalty incentives.

8. COMPARE K-MEANS CLUSTER AND HIERARCHICAL CLUSTERS

8.1 Summary of K – Means Cluster Means And Hirarchial Cluster Means



```
K-means Cluster Means:
                   Avg_Credit_Limit  Total_Credit_Cards  Total_visits_bank  \
K_means_segments
0                     33782.383420            5.515544           3.489637
1                     12174.107143            2.410714           0.933036
2                    141040.000000            8.740000           0.600000

                  Total_visits_online  Total_calls_made
K_means_segments
0                            0.981865          2.000000
1                            3.553571          6.870536
2                           10.900000          1.080000

Hierarchical Cluster Means:
              Avg_Credit_Limit  Total_Credit_Cards  Total_visits_bank  \
HC_Clusters
0                33851.948052            5.516883           3.493506
1                12151.111111            2.422222           0.937778
2               141040.000000            8.740000           0.600000

             Total_visits_online  Total_calls_made
HC_Clusters
0                       0.979221          1.994805
1                       3.546667          6.857778
2                      10.900000          1.080000
```

Table 20 Summary of K – Means Cluster Means And Hierarchial Cluster Means

8.2 Boxplot of Numerical Variables for Each Cluster in K -Means

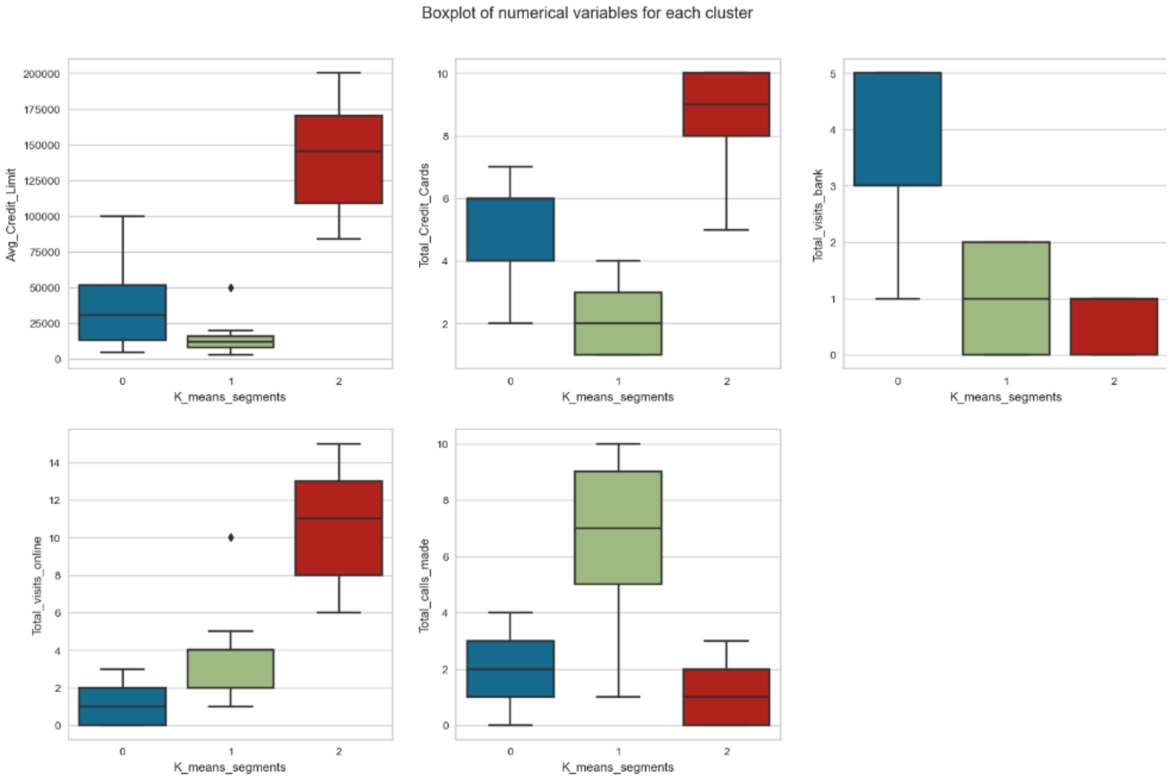


Figure 36 Boxplot for K - Means

8.3 Conclusion for K-Means Cluster

- Diverse Segments: The K-means method successfully segments customers into distinct groups based on financial capability and engagement preferences.
- Targeted Strategies: Each segment has unique needs, enabling the development of targeted strategies:
- Segment 0 for consistent revenue.
- Segment 1 for growth opportunities through support.
- Segment 2 for maximizing high-value customer retention.

8.4 Boxplot of Numerical Variables for Each Cluster in HC

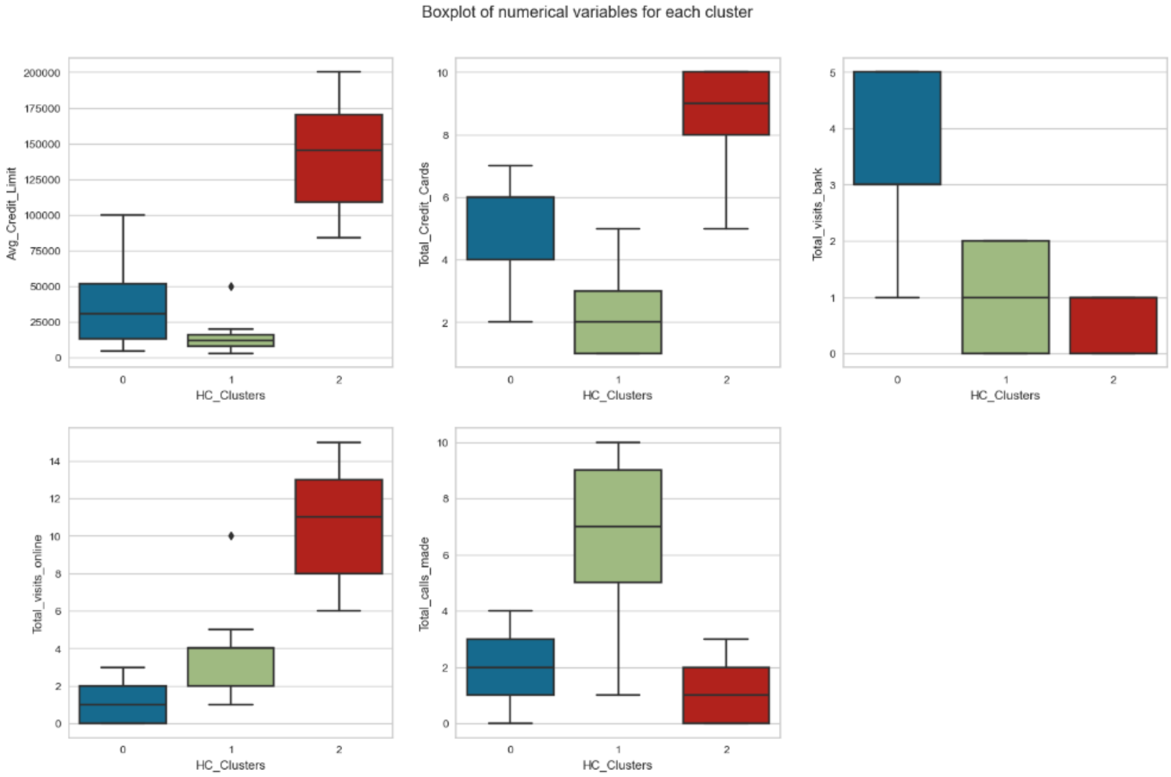


Figure 37 Boxplot for HC

8.5 Conclusion for Hierarchial Cluster

- Distinct Segmentation: The hierarchical clustering successfully identifies three distinct customer groups, each with unique financial capabilities and engagement behaviors.
- Strategic Recommendations:
- Cluster 0 for stable, mixed engagement strategies.
- Cluster 1 for growth and support-focused initiatives.
- Cluster 2 for retention and high-value digital engagement.

8.6 Overall Comparison of K-means and HC Clusters

Segment 0

- K-means Segment 0:
  - Average Credit Limit: $33,782.38
  - Total Credit Cards: 5.52
  - Total Visits to Bank: 3.49
  - Total Visits Online: 0.98
  - Total Calls Made: 2.00
  - Count in Segment: 386
- HC Cluster 0:
  - Average Credit Limit: $33,851.95 (slightly higher)

    - Total Credit Cards: 5.52 (same)
    - Total Visits to Bank: 3.49 (same)
    - Total Visits Online: 0.98 (same)
    - Total Calls Made: 1.99 (slightly lower)
    - Count in Segment: 385
- Observation: Both K-means Segment 0 and HC Cluster 0 have almost identical characteristics, with very slight differences in credit limits and calls made. This segment shows consistency across both clustering methods, representing a large group with moderate financial capacity and a balanced approach to traditional banking.

Segment 1

- K-means Segment 1:
    - Average Credit Limit: $12,174.11
    - Total Credit Cards: 2.41
    - Total Visits to Bank: 0.93
    - Total Visits Online: 3.55
    - Total Calls Made: 6.87
    - Count in Segment: 224
- HC Cluster 1:
    - Average Credit Limit: $12,151.11 (slightly lower)
    - Total Credit Cards: 2.42 (slightly higher)
    - Total Visits to Bank: 0.94 (slightly higher)
    - Total Visits Online: 3.55 (same)
    - Total Calls Made: 6.86 (slightly lower)
    - Count in Segment: 225

- Observation: K-means Segment 1 and HC Cluster 1 are also almost identical, with only minor variations in average credit limit, total credit cards, and visits to the bank. This segment represents a financially weaker but highly engaged group, suggesting a potential need for support or tailored financial products.

Segment 2

- K-means Segment 2:
    - Average Credit Limit: $141,040.00
    - Total Credit Cards: 8.74
    - Total Visits to Bank: 0.60
    - Total Visits Online: 10.90
    - Total Calls Made: 1.08
    - Count in Segment: 50
- HC Cluster 2:
    - Average Credit Limit: $141,040.00 (same)
    - Total Credit Cards: 8.74 (same)
    - Total Visits to Bank: 0.60 (same)
    - Total Visits Online: 10.90 (same)
    - Total Calls Made: 1.08 (same)
    - Count in Segment: 50

- Observation: Both methods yield identical results for this segment, representing high-value, tech-savvy customers who prefer online banking and require minimal customer support. This segment has the highest financial power and revenue potential across both clustering techniques.

Conclusion

- Consistency: Both K-means and Hierarchical Clustering methods produced nearly identical results across all segments, showing consistency in grouping customers based on similar attributes.
- Segment Definitions: Both clustering methods reveal three distinct customer types:
    1. Segment 0/Cluster 0: Large, stable customer base with moderate financial capability and balanced engagement.
    2. Segment 1/Cluster 1: Financially weaker but highly engaged customers, likely needing more support.
    3. Segment 2/Cluster 2: High-value customers with strong digital engagement, low support needs, and high profitability.

Summary:

K-means Clustering:

- Ideal for larger datasets, providing clear and easily interpretable clusters.
- Sensitive to outliers, which can skew results.
- Requires the number of clusters to be predetermined, which can affect outcomes if the choice is not optimal.

Hierarchical Clustering:

- Offers greater flexibility in terms of cluster shape and structure.
- Does not necessitate prior specification of the number of clusters, allowing for more exploratory analysis.
- May be less efficient for large datasets, as its computational complexity increases with size.
- Can result in complex interpretations when applied to extensive data, making it challenging to derive actionable insights.

9. ADD-ON: PCA FOR VISUALIZATION

Let's use PCA to reduce the data to two dimensions and visualize it to see how well-separated the clusters are.

The first two principal components explain 83.41% of the variance in the data.

9.1 Scatterplot of First Two Principal Components

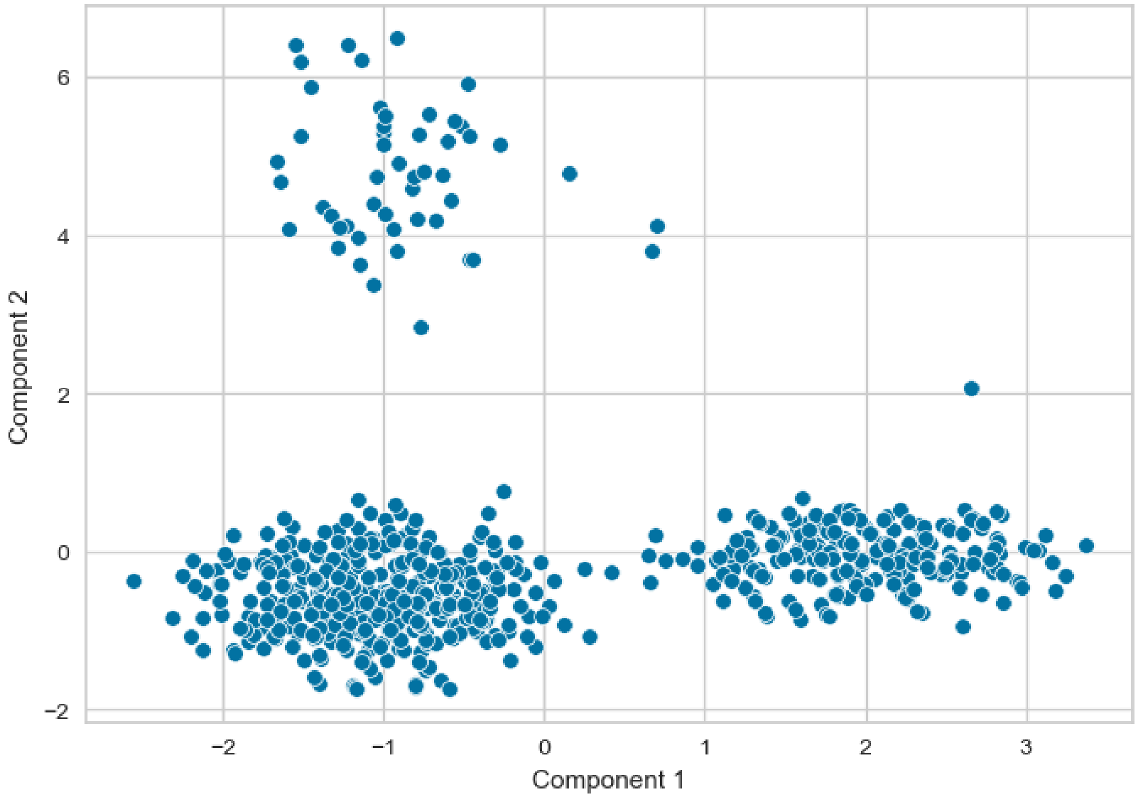


Figure 38 Scatterplot of First Two Principal Components

Observation:

- The data points are clustered into three distinct groups: one large cluster in the bottom-left, another smaller cluster in the top-right, and a more spread-out group in between.
- There is also an isolated point in the middle of the plot, above the main cluster.

9.2 K - Means Clustering

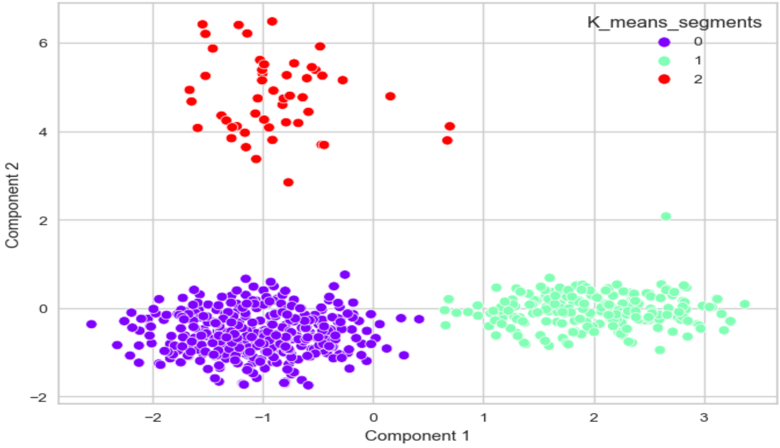


Figure 39 Scatterplot of First Two Principal Components with k – Means Cluster

Observation

- Cluster 0 (purple): This group is positioned in the bottom-left area of the plot.
- Cluster 1 (light green): This group is centered near the bottom-middle of the plot.
- Cluster 2 (red): This group is found in the top-right section of the plot.
- The K-means algorithm has successfully divided the data into three clusters, corresponding to the visible groups from the previous plot.
- The isolated points from the earlier plot are now included in the appropriate clusters, with one point in the purple segment (Cluster 0).

9.3 Hierarchial Clustering

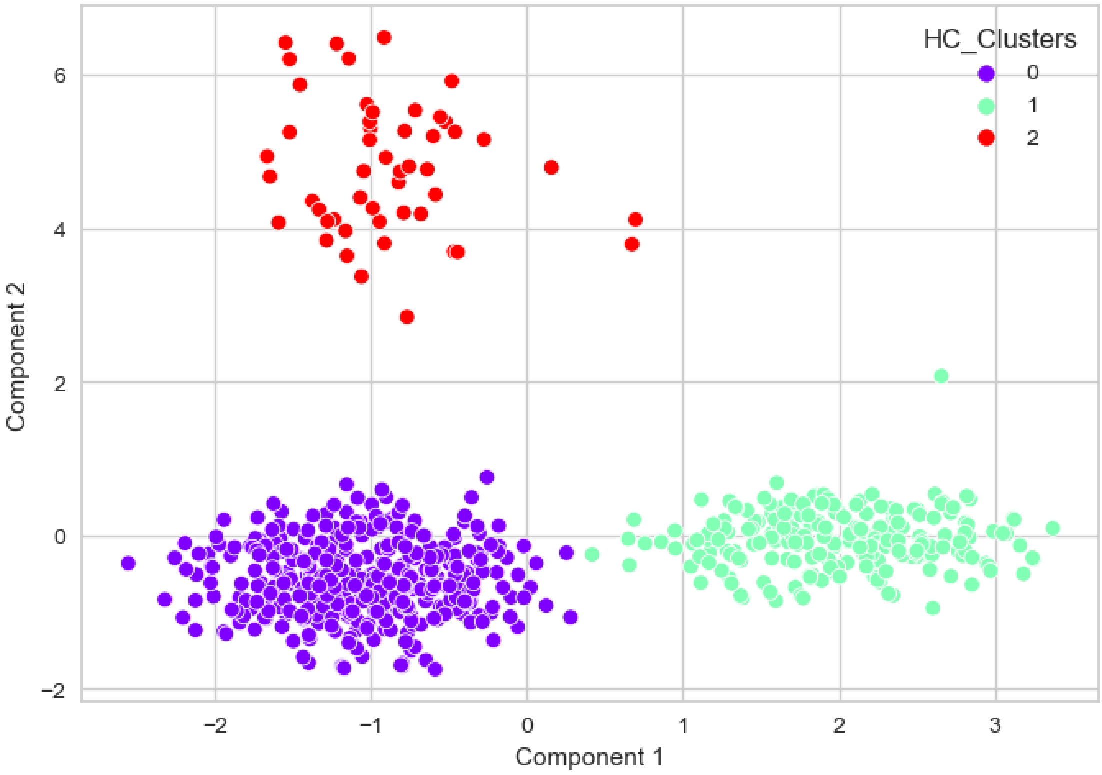


Figure 40 Scatterplot of First Two Principal Components with HC

Observation:

- Cluster 0 (purple): Positioned in the bottom-right area of the plot.
- Cluster 1 (light green): Located in the bottom-left and middle-left regions.
- Cluster 2 (red): Found in the top-right section of the plot.

9.4 Comparison of K- Means and HC Clustering For PCA

When comparing K-means and hierarchical clustering, both methods generally identified the same three groups; however, there are slight differences in partitioning:

- Cluster 0 (purple): Contains a few additional points on the left compared to K-means.
- Cluster 1 (green): Includes more points in the upper-middle region.

While the overall groups are similar, hierarchical clustering captures a slightly different structure in the transition areas between clusters, indicating a nuanced separation in these regions.

10. INSIGHTS AND RECOMMENDATION

10.1 Insights

Segment 2 (High-Value, Digital-Engaged Customers)

Profile:

- Highest financial activity in terms of credit limits and credit card ownership.
- Strong preference for online interactions; minimal in-branch visits and phone calls.

Actionable Insights:

- Enhance Digital Offerings: Provide premium digital services tailored to high-net-worth customers, such as virtual financial advisory, personalized financial insights, or exclusive online offers.
- Leverage Convenience: Introduce advanced self-service options like instant messaging with advisors, online investment management, and priority customer service for digital transactions.
- Build Loyalty Through Exclusivity: Offer digital-only rewards, early access to investment opportunities, or exclusive online events to strengthen engagement.

Segment 1 (Moderate-Value, Physically Engaged Customers)

Profile:

- Moderate financial activity, balancing between online and offline engagement.
- Frequent in-branch visits with steady online activity.

Actionable Insights:

- Promote Hybrid Services: Offer services that integrate both in-person and online experiences, such as scheduled in-branch consultations followed by digital follow-ups.
- Targeted Promotions: Incentivize digital use for convenience while maintaining in-branch benefits. For example, discounts for in-person advisory sessions booked online.
- Personalized Communication: Use a multi-channel approach to reach this segment with reminders, offers, and updates on both online and offline channels.

Segment 0 (Low-Value, High-Contact Customers)

Profile:

- Lower financial activity; primarily engages through phone calls and in-branch visits.
- Minimal use of online services.

Actionable Insights:

- Encourage Digital Adoption: Introduce incentives for trying online services, such as discounts, free online banking tools, or simplified onboarding to mobile banking.
- Enhance In-Person Experience: Ensure a seamless, customer-friendly experience in branches with well-trained staff ready to assist, especially in handling routine transactions.
- Personalized Outreach: Provide tailored communications and follow-ups to encourage digital transitions, offering guidance for those less comfortable with digital platforms.

10.2 Recommendation

Increase Digital Engagement for High-Value Segments (Segment 2)

- Expand Digital Offerings: Introduce features like personalized app dashboards, AI-driven financial advisors, and high-limit credit products with rewards for online spending.
- Enhance Online Loyalty Programs: Offer loyalty benefits tied to digital usage, such as cash-back for online transactions, early access to digital investment products, or premium digital content.
- Promote Self-Service: Enable quick access to services like instant support chat and online investment tracking, reducing their need for in-branch visits.

Provide Hybrid Service Models for Moderate-Value Customers (Segment 1)

- Cross-Channel Marketing: Highlight the benefits of digital convenience for routine tasks (e.g., transfers, bill payments) while reserving branch visits for complex needs.
- Complementary Tools: Encourage app use with features like appointment scheduling, transaction summaries, and alerts that enhance in-branch service experiences.
- Hybrid Service Options: Promote services that integrate digital convenience with in-person reassurance, such as setting up online consultations for follow-ups after in-branch meetings.
- Focus on Personalization for Low-Value, High-Contact Customers (Segment 0)
- Personalized Support: Assign dedicated relationship managers and offer robust phone-based customer support to meet their preference for human interaction.
- Introduce Digital Incentives: Offer onboarding programs with incentives like waived fees or discounts on online transactions to encourage mobile or web banking use.
- Streamline Digital Transitions: Simplify processes like online registration, making it easier for those hesitant about digital channels to get started.

Enhance Customer Education (All Segments, with Emphasis on Segment 0)

- Educational Content: Provide resources on digital banking benefits, such as webinars, video tutorials, and one-on-one sessions to build comfort and familiarity.
- Targeted Outreach: Emphasize digital security, ease of access, and the efficiency of online tools in communications with segments less familiar with digital banking.
- Monitor Behavior for Future Segmentation
- Behavior Tracking: Regularly monitor engagement levels to detect shifts, especially in Segment 0, where digital adoption may increase over time.
- Dynamic Marketing Adjustments: Adjust marketing strategies to reflect changes in segment behavior and channel preferences, ensuring satisfaction across both digital and physical interactions.
- Optimize Resource Allocation: Use behavioral data to allocate resources effectively, prioritizing digital tool development for high-engagement segments and personalized service for high-contact customers.

This approach should enhance customer satisfaction, boost digital service adoption, and optimize bank resources by aligning services with each segment’s preferences.