Business Report - 6

PG Program in Data Science and Business Analytics

submitted by

Sangram Keshari Patro

BATCH:PGPDSBA.O.AUG24.B

Contents

List of Figures

1 Objective

AllLife Bank wants to focus on its credit card customer base in the next financial year. They have been advised by their marketing research team, that the penetration in the market can be improved. Based on this input, the Marketing team proposes to run personalized campaigns to target new customers as well as upsell to existing customers. Another insight from the market research was that the customers perceive the support services of the back poorly. Based on this, the Operations team wants to upgrade the service delivery model, to ensure that customer queries are resolved faster.

To identify different segments in the existing customers, based on their spending patterns as well as past interaction with the bank, using clustering algorithms, and provide recommendations to the bank on how to better market to and service these customers.

2 Data 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). The detailed data dictionary is given below.

2.1 Data dictionary

Data Dictionary

Attribute	Description
Sl_No	Primary key of the records.
Customer Key	Unique identification number assigned to each customer.
Average Credit Limit	Average credit limit of each customer across all credit cards.
Total Credit Cards	Total number of credit cards possessed by the customer.
Total Visits Bank	Yearly total number of in-person visits the customer made to the bank.
Total Visits Online	Yearly total number of online logins or interactions by the customer.
Total Calls Made	Yearly total number of calls made by the customer to the bank or customer service.

3 Data Overview

3.1 Importing necessary libraries and the dataset

The dataset is printed. It has 660 rows & 7 columns.

3.2 Structure and type of data

Data is explored further. The dataset is free from duplicate rows and contains no null values.

```
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
```

Figure 1: Table depicting the datatype and Non-Null values in each column.

3.3 Statistical summary

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

Figure 2: Statistical summary of the data

4 Exploratory Data Analysis

4.1 Univariate Analysis

4.1.1 Numerical columns

- **'Avg_Credit_Limit'**

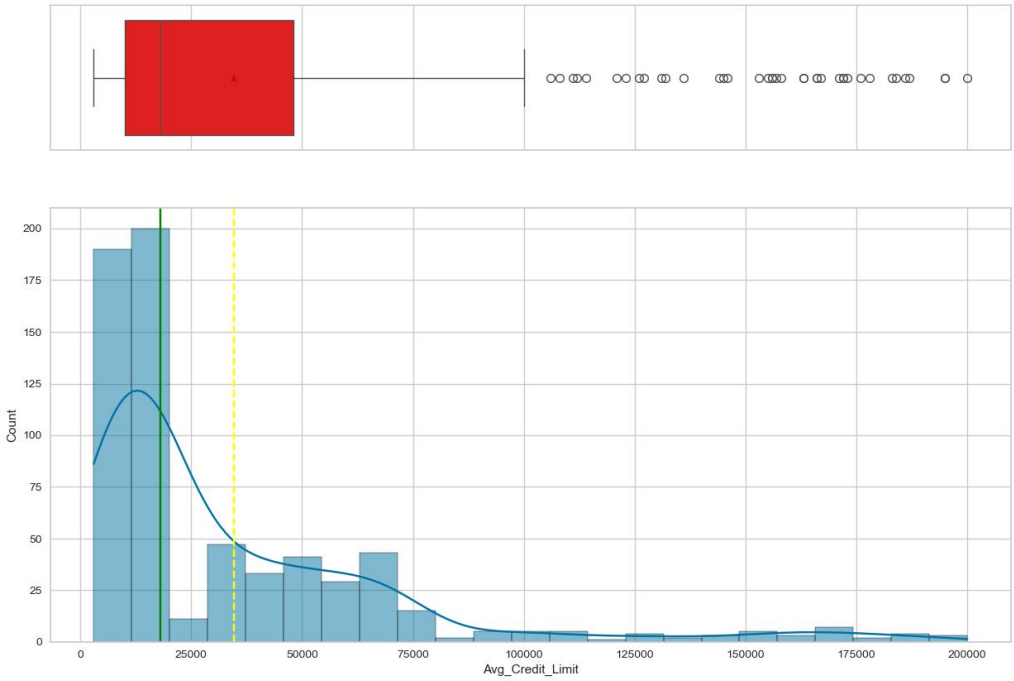


Figure 3: Histogram and boxplot of **'Avg_Credit_Limit'** column

Observations

- **Histogram**: The distribution of average credit limits is right-skewed, with most customers having a lower credit limit. The density decreases as the credit limit increases, showing that fewer customers have high credit limits.
- **Box Plot**: The median credit limit lies within the interquartile range (IQR), with a significant number of outliers at the higher end. This indicates that while most customers have lower credit limits, a few have exceptionally high limits.

Business Recommendations

- **Customized Credit Offerings**: Since a few customers have high credit limits, targeted premium credit products should be designed for high-value customers.
- **Risk Management**: The presence of high-value outliers suggests the need for careful credit risk assessment for customers with exceptionally high limits.
- **Market Expansion**: Since most customers fall in the lower credit limit range, banks should focus on financial products that cater to this majority segment.
- **'Total_Credit_Cards'**

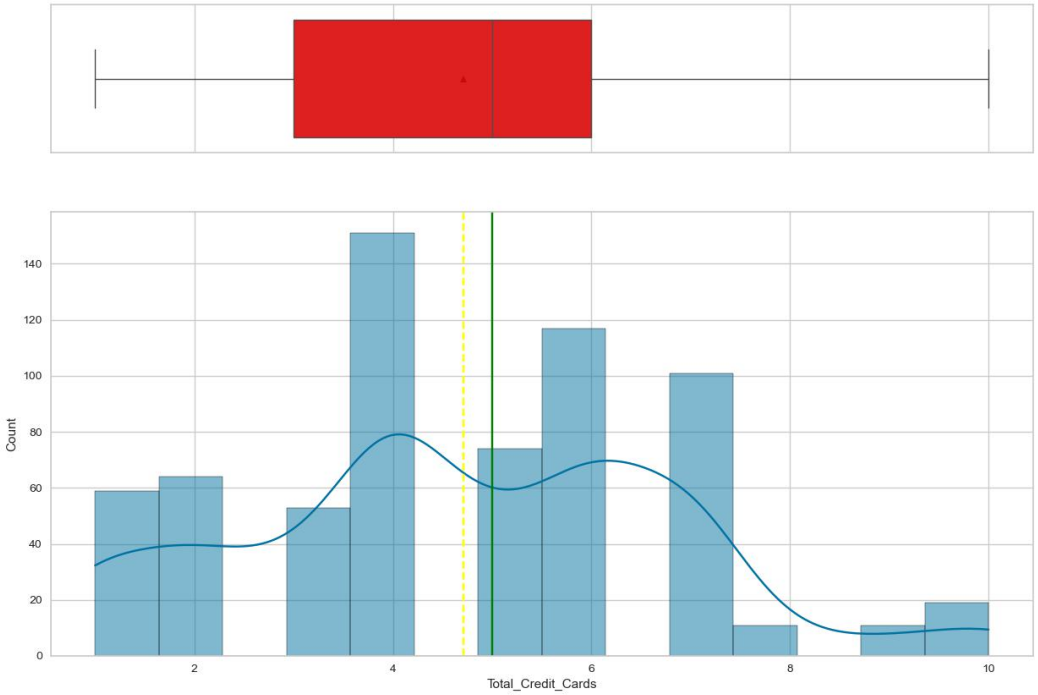


Figure 4: Histogram and boxplot of **'Total_Credit_Cards'** column

Observations

- **Histogram**: The number of credit cards owned by customers is distributed in distinct clusters, indicating customer segments with different credit needs. The distribution has multiple peaks, suggesting that specific numbers of cards are more common.

- **Box Plot**: The median number of credit cards falls within a typical range, but there are some customers who own an exceptionally high number of credit cards.

Business Recommendations

- **Tailored Credit Card Offerings**: The presence of multiple peaks suggests different customer segments. Banks should create specific marketing strategies targeting each segment.
- **Loyalty and Retention Programs**: Customers with multiple credit cards may be valuable for retention efforts through exclusive rewards and benefits.
- **Credit Utilization Monitoring**: Customers with numerous credit cards may pose a higher risk in terms of debt accumulation, requiring more refined credit monitoring policies.
- **'Total_visits_bank'**

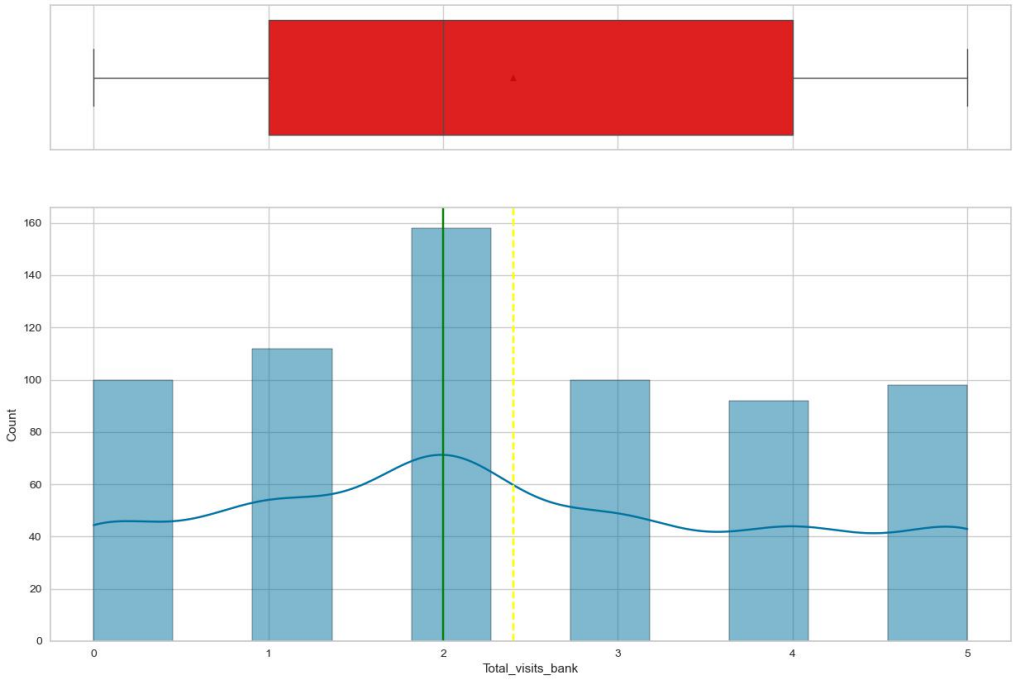


Figure 5: **'Total_visits_bank'** column

Observations

- **Histogram**: The number of bank branch visits follows a slightly right-skewed distribution, with most customers making fewer visits, while a smaller proportion visits frequently.
- **Box Plot**: The median number of visits is low, indicating that a majority of customers prefer fewer in-person interactions. However, some customers visit significantly more, suggesting specific needs.
- **Outliers**: A few customers visit the bank far more than the average, which could indicate special service needs or a lack of digital adoption.

Business Recommendations

- **Promote Digital Banking**: Since most customers make fewer visits, encourage further digital adoption by offering incentives for online and mobile banking usage.
- **Optimize Branch Services**: For high-frequency visitors, analyze their needs and provide personalized branch services or hybrid support models.
- **Reduce Operational Costs**: With low in-person engagement, consider streamlining branch operations, optimizing staff allocation, and reallocating resources to digital customer support.
- **Targeted Customer Education**: Identify frequent branch visitors who might not be comfortable with digital banking and offer training sessions to enhance their online banking experience.
- **'Total_visits_online'**

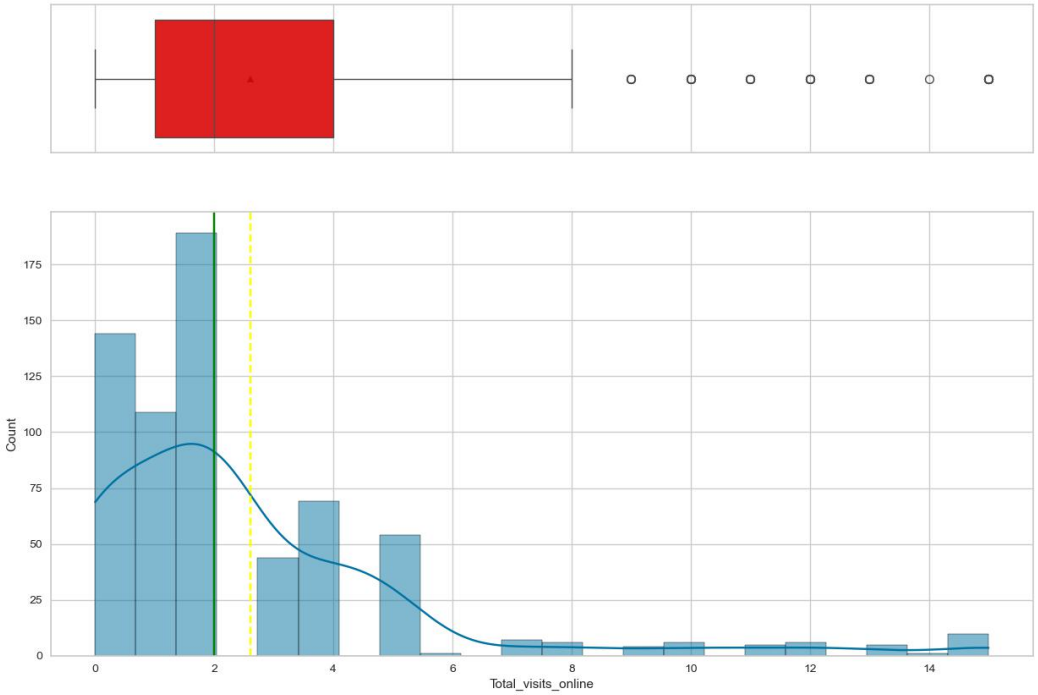


Figure 6: **'Total_visits_online'** column

Observations

- **Histogram**: Online visits are right-skewed, with most customers making a few visits while some engage frequently.
- **Box Plot**: The median is low, showing limited online interactions, but a few customers visit very often.
- **Outliers**: Some users have exceptionally high visits, indicating strong digital engagement.

Business Recommendations

- **Increase Digital Engagement**: Encourage low-frequency users to explore online banking with promotions.
- **Enhance User Experience**: Improve website/app usability for frequent visitors.
- **Optimize Digital Support**: Provide chatbot assistance for high-traffic users.
- **'Total_calls_made'**

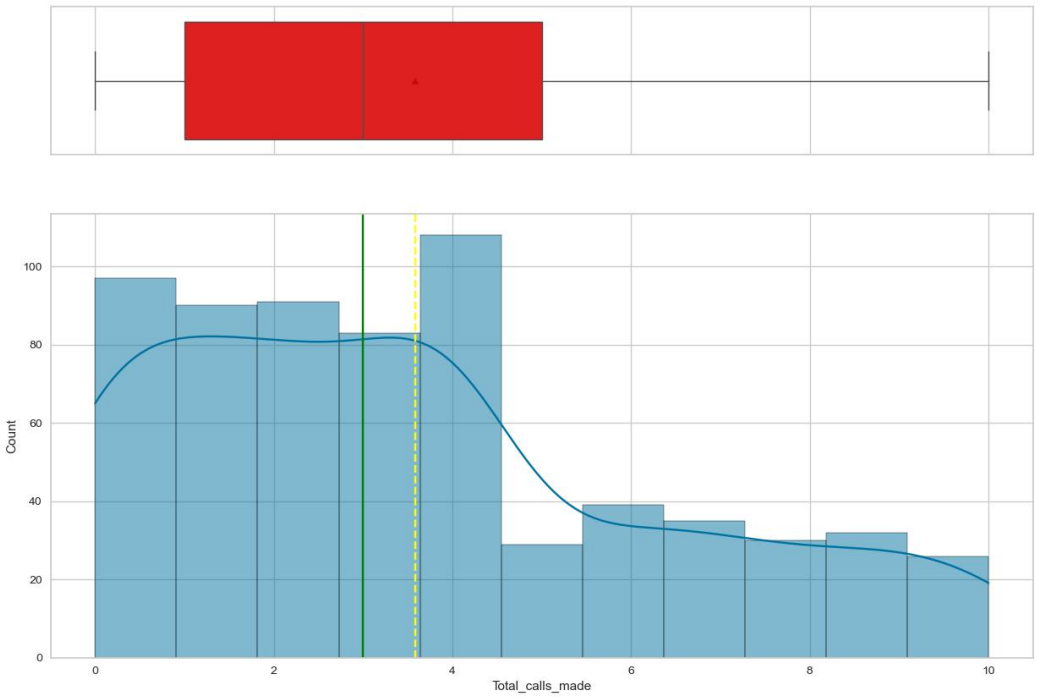


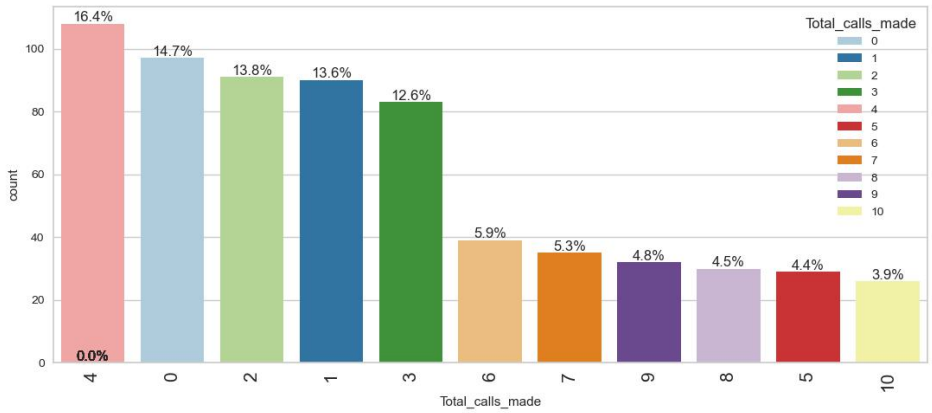
Figure 7: **'Total_calls_made'** column

Observations

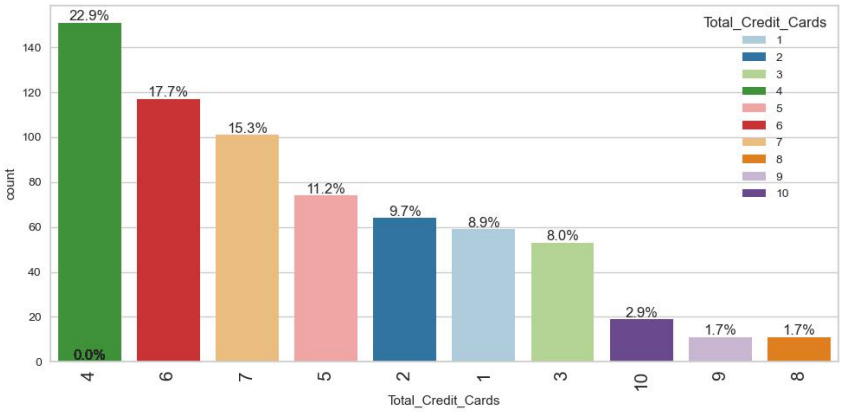
- **Histogram**: The number of calls made by customers follows a slightly right-skewed distribution, with most customers making a low number of calls. The frequency of high call volumes decreases gradually.
- **Box Plot**: The median number of calls is within the IQR, showing a balanced distribution, but a few customers have made a significantly higher number of calls.

Business Recommendations

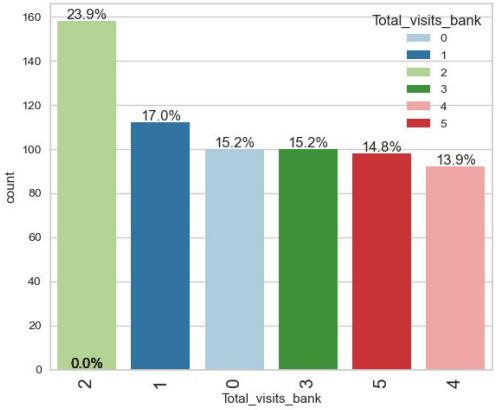
- **Enhance Self-Service Options**: Since most customers make fewer calls, investing in digital and self-service banking solutions can further reduce dependency on customer support.
- **Improve Call Center Efficiency**: A segment of customers makes frequent calls, indicating potential dissatisfaction or complex queries that need to be addressed with better FAQs and AI chat support.
- **Segment-Based Service Models**: Offer premium support services for high-frequency callers and encourage digital interaction for low-frequency callers.



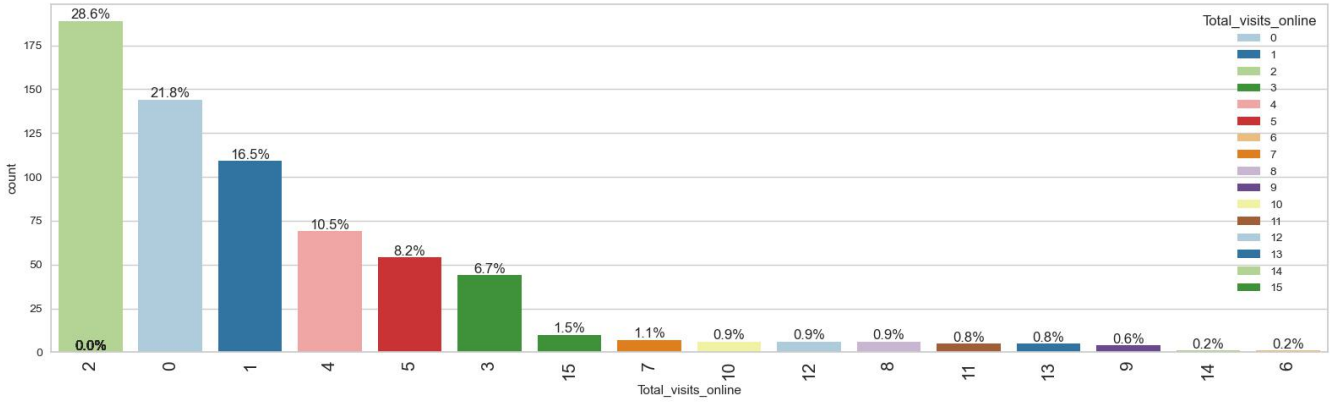
(a) '**Total_calls_made**'



(b) '**Total_Credit_Cards**'



(c) '**Total_visits_bank**'



(d) '**Total_visits_online**'

Observations

- **Total Calls Made**: The distribution is right-skewed, with most customers making a limited number of calls. A small segment makes frequent calls, indicating a need for assistance or unresolved issues.
- **Total Credit Cards**: Most customers hold a low number of credit cards, while a small portion owns multiple cards, possibly indicating high credit usage or loyalty to the bank.
- **Total Bank Visits**: A significant number of customers visit the bank rarely, but a subset makes frequent visits, likely for complex transactions or lack of digital adoption.
- **Total Online Visits**: Online banking is widely used, though visit frequency varies, suggesting different levels of digital engagement among customers.

Business Recommendations

- **Enhance Self-Service and Digital Support**: Since call volumes are low for most but high for some, improving AI chatbots and FAQs can reduce reliance on call centers.
- **Targeted Credit Card Strategies**: Offer personalized promotions to high-credit users while encouraging others to explore additional banking products.
- **Reduce In-Branch Dependency**: Educate frequent branch visitors on digital banking options, ensuring smoother transitions for routine transactions.
- **Boost Online Engagement**: Incentivize low-frequency digital users with promotions, tutorials, or exclusive online banking benefits.

4.2 Bivariate Analysis

4.2.1 Numerical variables

- Heatmap

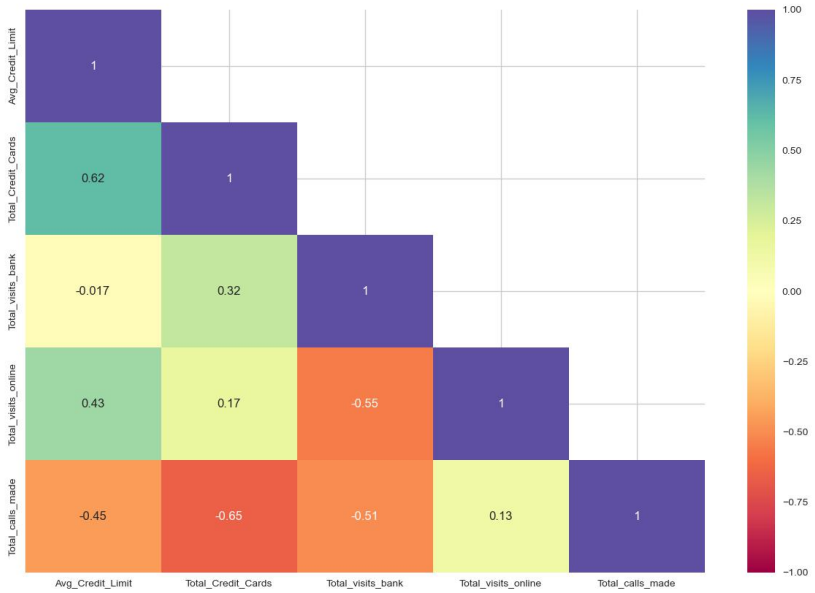


Figure 9: Heatmap of all numerical variables

Observations

- **Credit Limit and Total Credit Cards**: A moderate positive correlation (0.62) suggests that customers with more credit cards tend to have higher credit limits.
- **Total Visits Bank vs. Total Visits Online**: A strong negative correlation (-0.55) implies that customers visiting the bank frequently are less likely to engage in online banking.
- **Total Calls Made vs. Total Credit Cards**: A significant negative correlation (-0.65) indicates that customers with more credit cards tend to make fewer calls.

Business Recommendations

- **Promote Digital Banking**: Since higher bank visits correlate negatively with online visits, targeted incentives for digital banking adoption can help reduce branch congestion.
- **Optimize Call Center Services**: Customers with fewer credit cards make more calls, indicating they may need better onboarding or self-help resources.
- **Tailored Credit Strategies**: As credit limit correlates positively with the number of credit cards, segmenting customers for personalized credit limit offers can enhance engagement.
- **Pairplot**

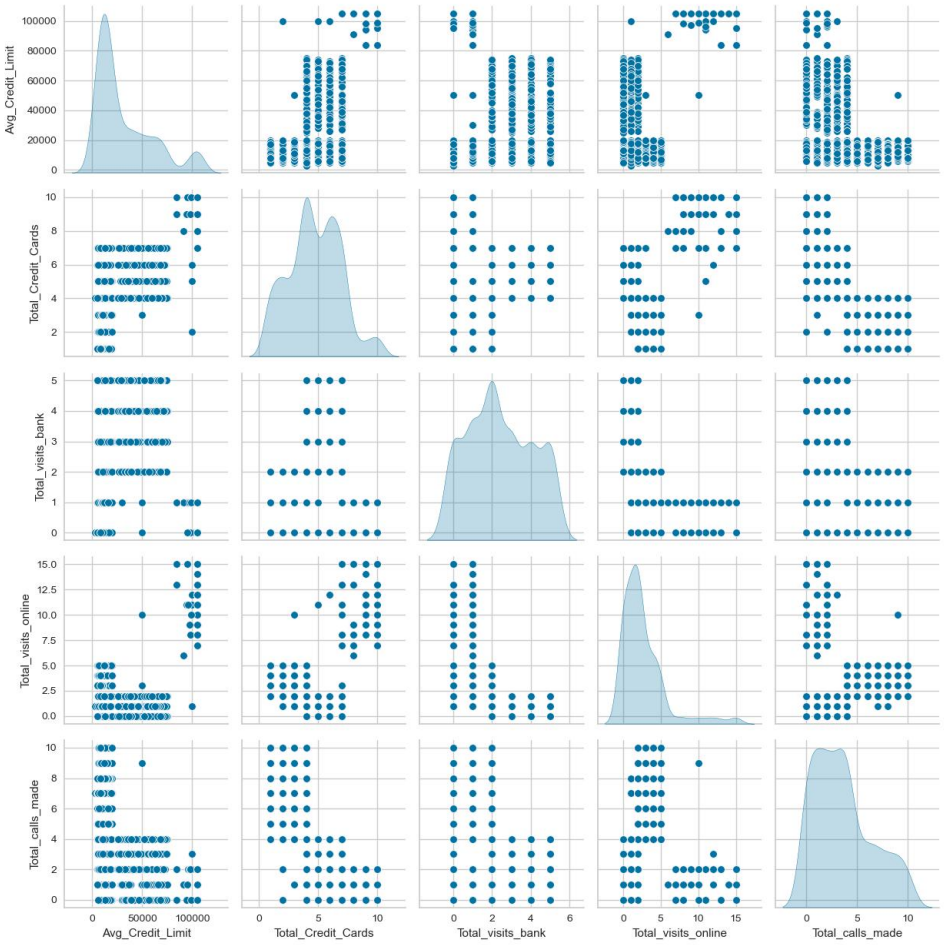


Figure 10: Pairplot of all numerical variables

Observations

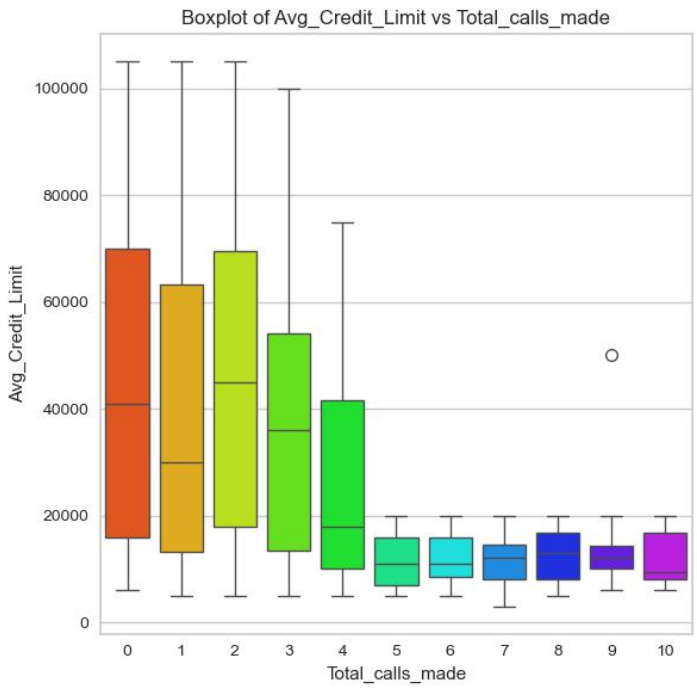
- **Avg. Credit Limit Distribution**: Right-skewed, indicating most customers have lower credit limits, while a few have significantly higher limits.
- **Total Credit Cards**: A bimodal distribution suggests distinct groups–those with fewer cards and those with multiple.
- **Total Visits (Bank vs. Online)**: Customers who visit banks more tend to have fewer online interactions, reinforcing an inverse relationship.
- **Total Calls Made**: Most customers make a limited number of calls, but a small group makes significantly more, possibly indicating service issues.

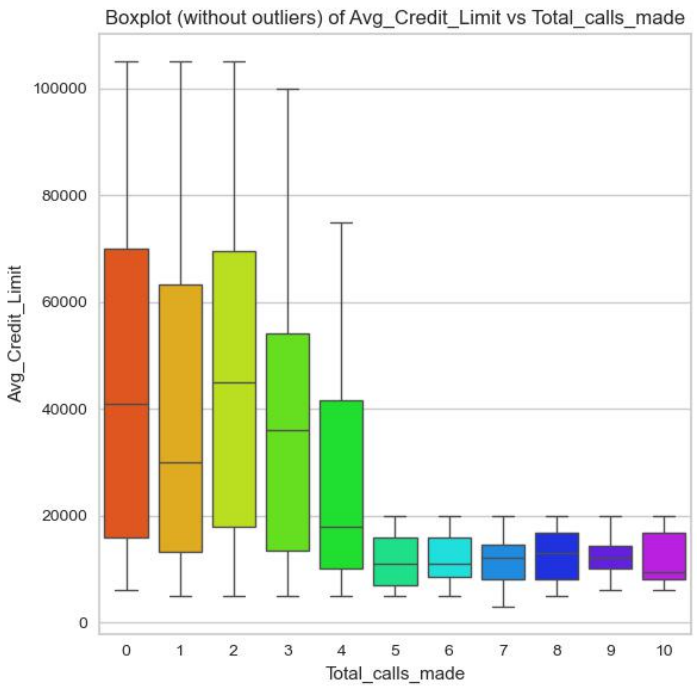
Business Recommendations

- **Segment-Based Credit Offers**: Identify high-credit customers separately to offer exclusive financial products.
- **Encourage Digital Adoption**: Customers with frequent branch visits should be incentivized to shift to digital services.
- **Call Center Enhancements**: Address high-call-frequency customers with better self-service resources and chatbot support.
- **Cluster-Specific Engagement**: Use distinct behavioral groups to personalize marketing strategies and enhance customer experience.

4.2.2 Categorical vs numerical variables

- **'Avg_Credit_Limit' vs all columns**





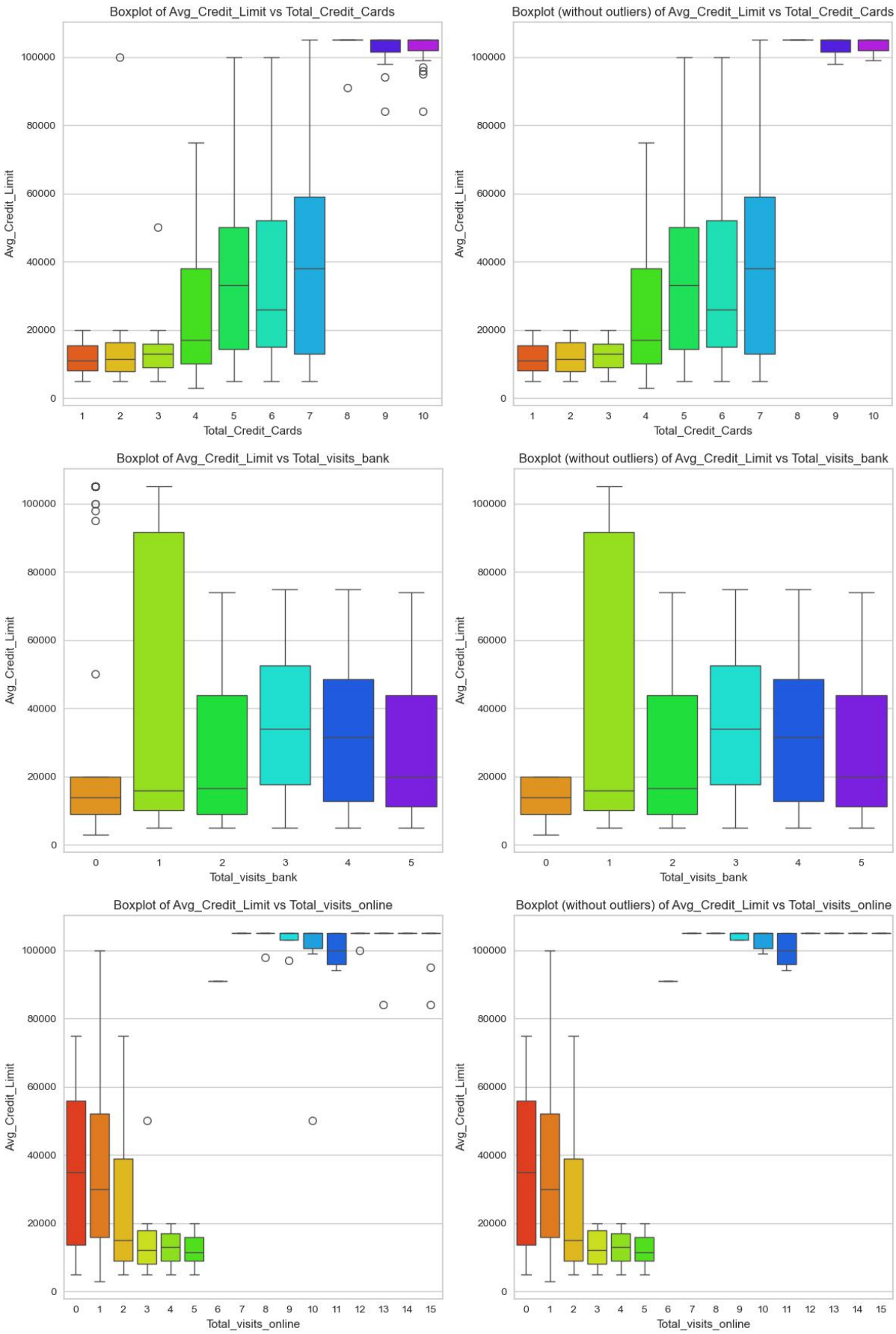


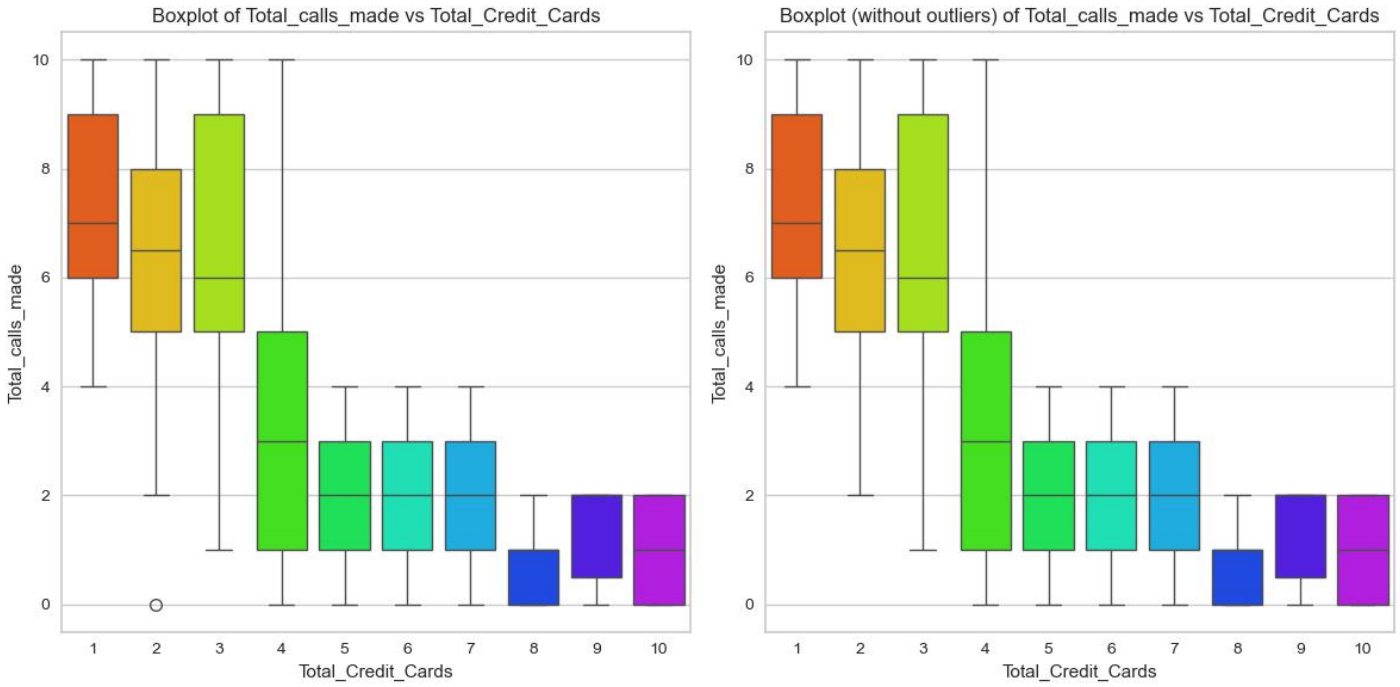
Figure 11: **'Avg_Credit_Limit' vs all columns**

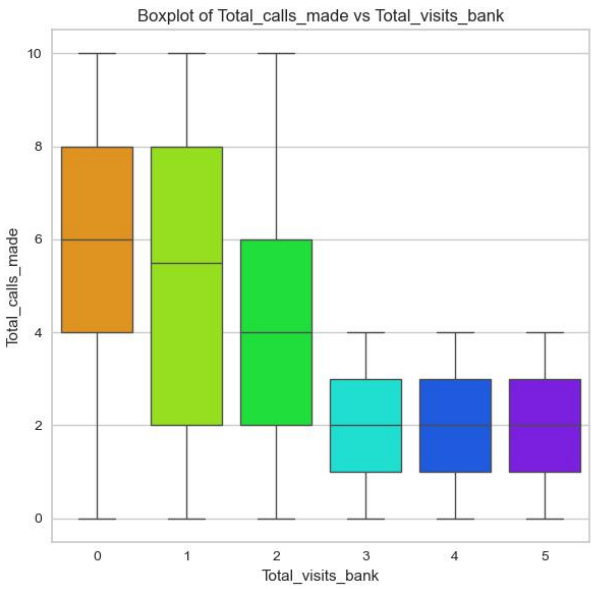
Observations

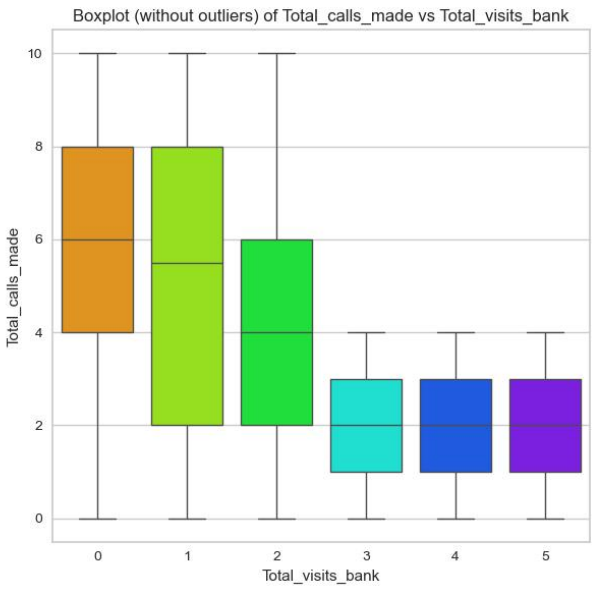
- **Average Credit Limit vs. Total Calls Made**: Customers with lower credit limits tend to make more calls, possibly due to inquiries or service-related issues, whereas those with higher credit limits make fewer calls, suggesting they require less assistance or have access to premium support channels.
- **Average Credit Limit vs. Total Credit Cards**: The distribution suggests that customers with higher credit limits tend to have more credit cards, indicating a correlation between higher financial trust and multiple credit lines.
- **Average Credit Limit vs. Total Visits to the Bank**: Customers with higher credit limits tend to visit the bank more frequently, possibly due to their engagement in more complex financial transactions or seeking premium services.
- **Average Credit Limit vs. Total Online Visits**: Higher credit limit customers tend to engage more in online banking, showing a preference for digital financial management over physical branch visits.

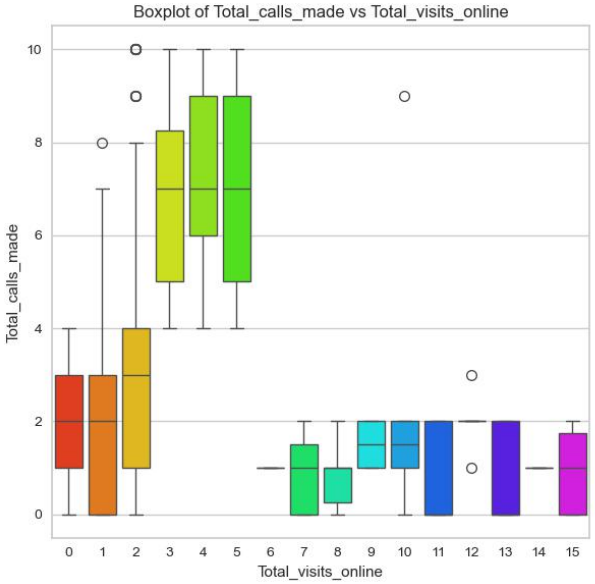
Business Recommendations

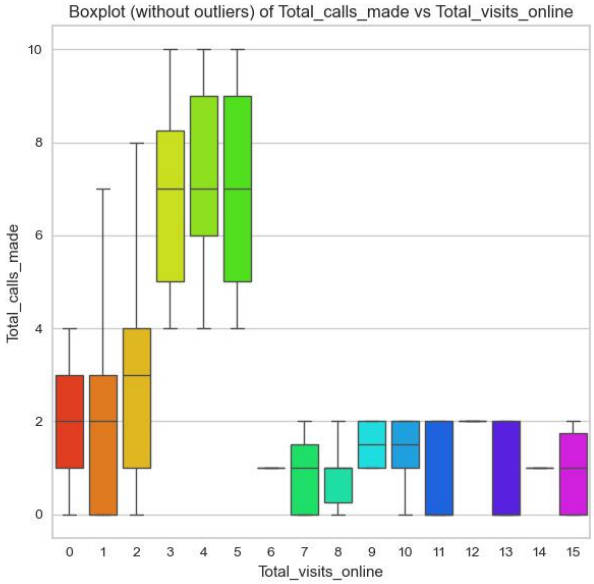
- **Enhanced Support for Low-Credit Customers**: Implement better self-service resources and proactive customer support to reduce the need for excessive call center interactions.
- **Exclusive Offers for High-Credit Customers**: Provide tailored benefits such as premium cards and financial products to encourage customer loyalty.
- **Encouraging Digital Banking Adoption**: Develop targeted campaigns to educate and incentivize low-credit customers to shift towards online banking, reducing their dependency on branch visits.
- **Optimizing Credit Card Strategies**: Identify customers who can benefit from additional credit options and promote appropriate products based on their financial behavior.
- **'Total_calls_made' vs 'Total_Credit_Cards' vs 'Total_visits_bank' vs 'Total_visits_online'**

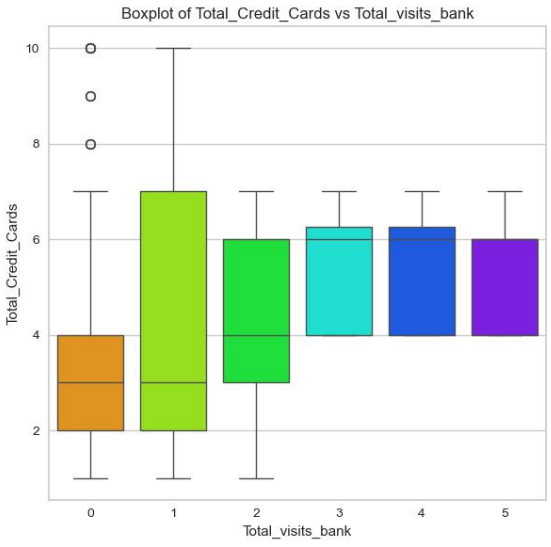


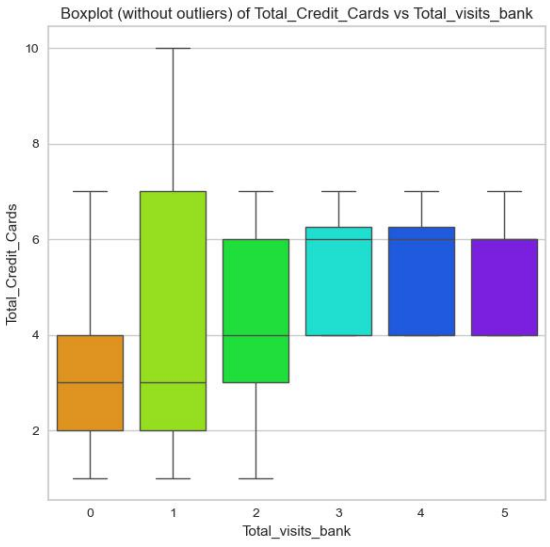

Boxplot of Total_calls_made vs Total_visits_bank
Total_calls_made
Total_visits_bank


Boxplot (without outliers) of Total_calls_made vs Total_visits_bank
Total_calls_made
Total_visits_bank


Boxplot of Total_calls_made vs Total_visits_online
Total_calls_made
Total_visits_online


Boxplot (without outliers) of Total_calls_made vs Total_visits_online
Total_calls_made
Total_visits_online


Boxplot of Total_Credit_Cards vs Total_visits_bank
Total_Credit_Cards
Total_visits_bank


Boxplot (without outliers) of Total_Credit_Cards vs Total_visits_bank
Total_Credit_Cards
Total_visits_bank

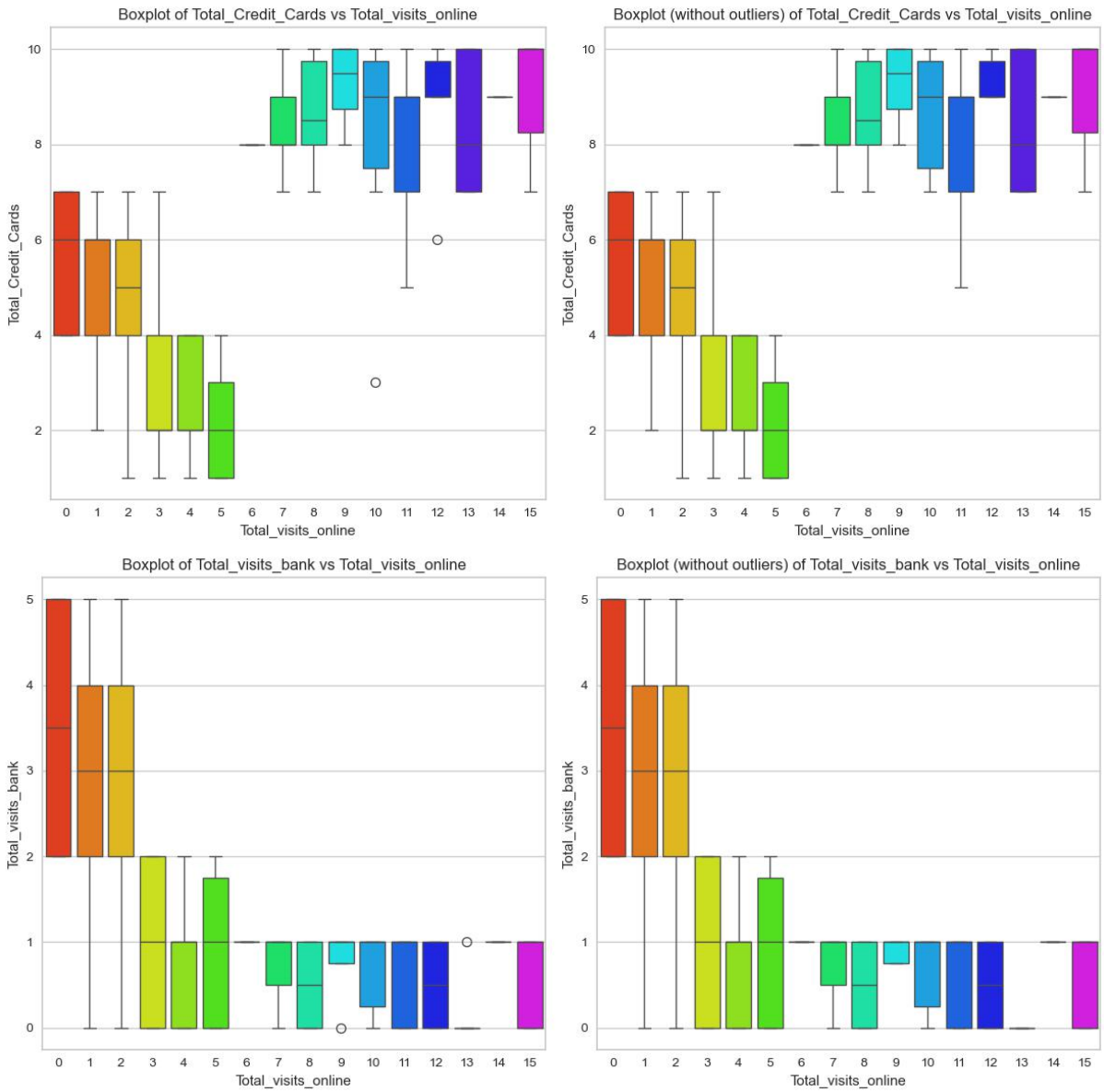


Figure 12: **'Total_calls_made' vs 'Total_Credit_Cards' vs 'Total_visits_bank' vs 'Total_visits_online'**

Observations

Total Calls Made vs. Total Visits to Bank

Customers who visit the bank more frequently make fewer calls, suggesting that in-person visits help resolve issues.

Total Calls Made vs. Total Credit Cards

The median number of calls decreases as the number of credit cards increases, indicating that multi-card holders require less support.

Total Calls Made vs. Total Visits Online

Customers with higher online activity make fewer calls, indicating reliance on digital channels for issue resolution.

Total Credit Cards vs. Total Visits to Bank

Customers with more credit cards visit the bank more frequently, suggesting higher service needs.

Total Credit Cards vs. Total Visits Online

Online visits initially decrease with more credit cards but increase for customers with multiple cards, indicating varied digital adoption.

Total Visits to Bank vs. Total Visits Online

Customers with higher online engagement visit branches less, showing a shift towards digital banking.

Business Recommendations

Total Calls Made vs. Total Visits to Bank

- Improve in-branch service to reduce follow-up calls.
- Offer remote support for customers who visit less but call frequently.
- Optimize resource allocation based on visit and call patterns.

Total Calls Made vs. Total Credit Cards

- Enhance self-service options for independent issue resolution.
- Provide dedicated support for customers with fewer credit cards.
- Re-engage high-credit-card holders with targeted promotions.

Total Calls Made vs. Total Visits Online

- Promote digital banking through tutorials and incentives.
- Enhance self-service tools such as chatbots and FAQs.
- Provide digital onboarding for customers with low online engagement.

Total Credit Cards vs. Total Visits to Bank

- Introduce priority service for high-card customers.
- Cross-sell financial products during in-branch visits.
- Implement an appointment-based system to manage high-traffic customers.

Total Credit Cards vs. Total Visits Online

- Promote online banking to low-card customers.
- Enhance digital services for multi-card users.
- Personalize outreach to mid-level credit card holders.

Total Visits to Bank vs. Total Visits Online

- Strengthen digital banking capabilities.
- Offer exclusive digital promotions to low-visit customers.
- Reduce branch dependency through digital education initiatives.

5 Data preprocessing

The dataset contains no missing or duplicate values. The outliers are significant for the data so we don't require to remove them except for the column **'Avg_Credit_Limit'**. The records for the same Customer Key appear to be significantly different from each other. This could be due to an error in Customer Key assignment or the absence of a current_version_indicator in the dataset. For now, I will treat these as separate customers. After clustering, I will analyze the groups associated with these sets of records.

The data is scaled with standard scaler function for clustering.

6 Clustering Methods

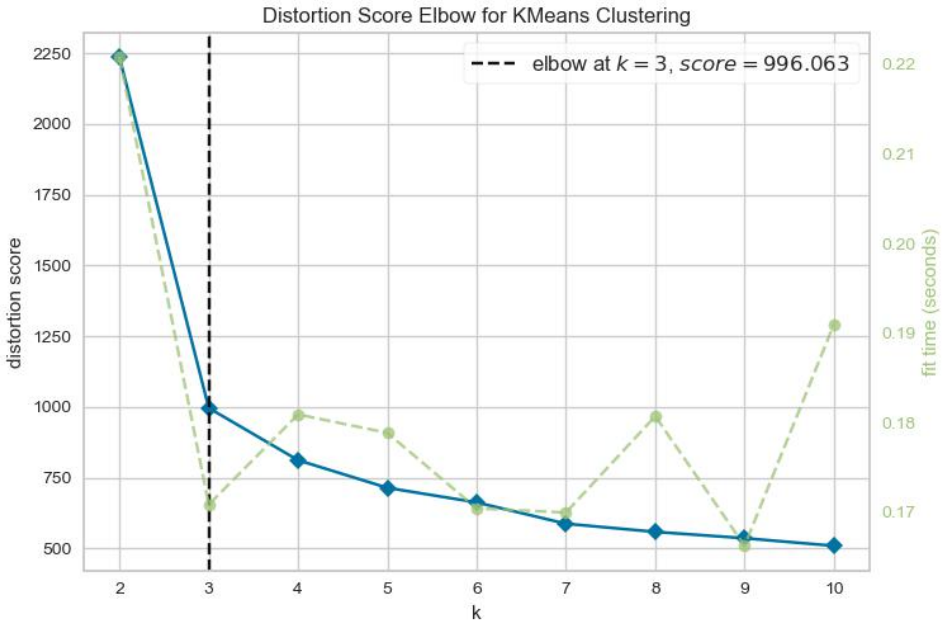
6.1 K-means Clustering

K-Means is an unsupervised machine learning algorithm used for clustering. It partitions data into K clusters based on similarity. The algorithm works as follows:

- Choose K cluster centroids randomly.
- Assign each data point to the nearest centroid.
- Update centroids based on the mean of assigned points.
- Repeat until centroids stabilize.

It is efficient for large datasets but sensitive to the choice of K and outliers.

6.1.1 Checking Elbow Plot



```
Number of Clusters: 2   Average Distortion: 1.755501055389377
Number of Clusters: 3   Average Distortion: 1.1836293794561177
Number of Clusters: 4   Average Distortion: 1.0700524605928141
Number of Clusters: 5   Average Distortion: 1.0024110842154574
Number of Clusters: 6   Average Distortion: 0.9642295040919028
Number of Clusters: 7   Average Distortion: 0.9092212293058839
Number of Clusters: 8   Average Distortion: 0.8832543493966902
Number of Clusters: 9   Average Distortion: 0.8645268683858071
Number of Clusters: 10  Average Distortion: 0.8419992835184149
```

Figure 13: Distortion score Elbow for KMeans Clustering

- We can clearly observe that the change in slope is observed at k=3,4 and 5 out of which k=3 is best as it takes less time to fit.

6.1.2 Check Silhouette Scores

Silhouette Score

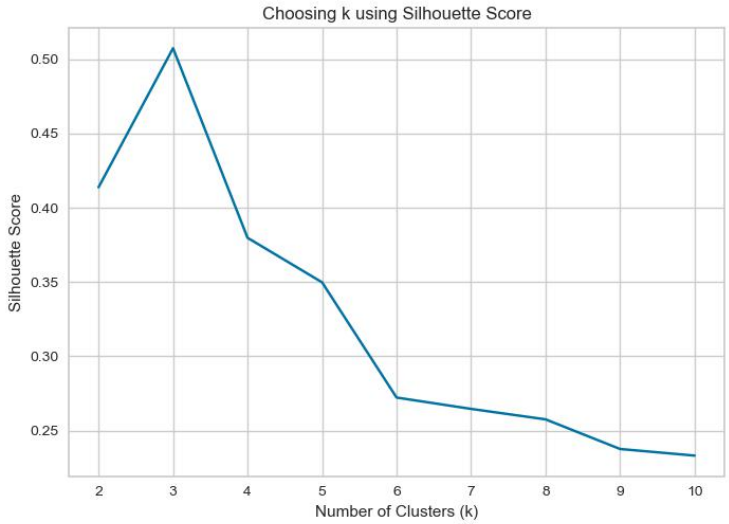
Silhouette Score measures clustering quality by evaluating how well data points fit within their assigned clusters. It is calculated as:

$$S = \frac{b - a}{\max(a, b)}$$

where:

- a = average intra-cluster distance.
- b = average nearest-cluster distance.

Values range from -1 to 1. Higher values indicate well-separated clusters, while negative values suggest misclassification.




```
For n_clusters = 2, the silhouette score is 0.4138293431016039)
For n_clusters = 3, the silhouette score is 0.5075309885053054)
For n_clusters = 4, the silhouette score is 0.37980395444812504)
For n_clusters = 5, the silhouette score is 0.3498370635383446)
For n_clusters = 6, the silhouette score is 0.27220958947721213)
For n_clusters = 7, the silhouette score is 0.26456292258119385)
For n_clusters = 8, the silhouette score is 0.2575123443993521)
For n_clusters = 9, the silhouette score is 0.23755436819502862)
For n_clusters = 10, the silhouette score is 0.23309787384423367)
```

Figure 14: Silhouette scores for different k

- We can observe that silhouette score for k=3 is the highest which indicates well-separated clusters. So, we will choose 3 as value of k.

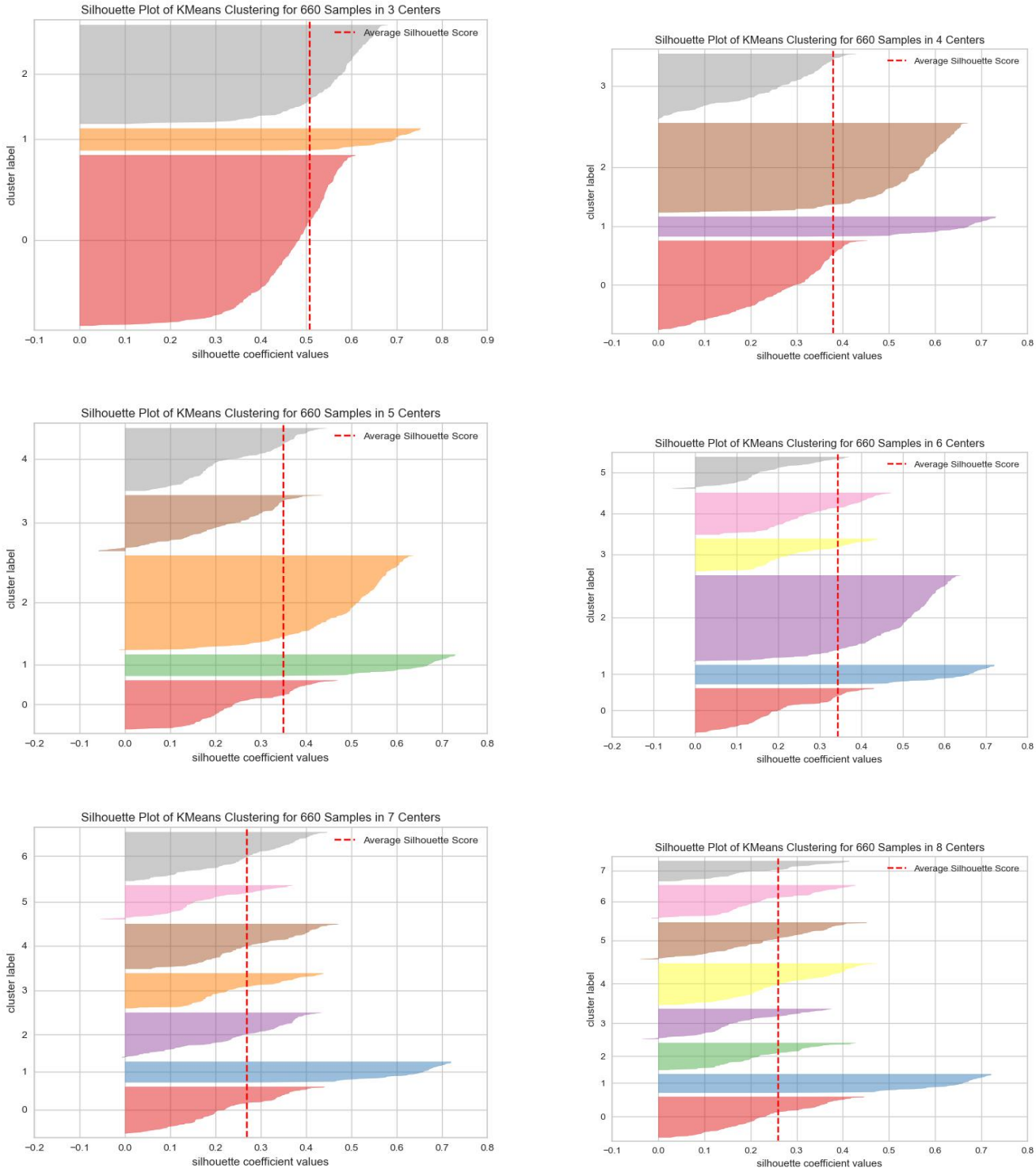


Figure 15: Silhouette plots for different k

6.1.3 Cluster Profiling

After dividing into clusters the data is explored further to get insights about the formed clusters.

KMeans_group	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count
0	33782.383420	5.515544	3.489637	0.981865	2.000000	386
1	102660.000000	8.740000	0.600000	10.900000	1.080000	50
2	12174.107143	2.410714	0.933036	3.553571	6.870536	224

Total_visits_online	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
KMeans_group																
0	144	106	135	1	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	1	7	6	4	5	5	6	5	1	10
2	0	3	54	43	69	54	0	0	0	0	1	0	0	0	0	0

Total_Credit_Cards	1	2	3	4	5	6	7	8	9	10
KMeans_group										
0	0	1	0	102	73	116	94	0	0	0
1	0	0	0	0	1	1	7	11	11	19
2	59	63	53	49	0	0	0	0	0	0

Total_visits_bank	0	1	2	3	4	5
KMeans_group						
0	0	3	93	100	92	98
1	20	30	0	0	0	0
2	80	79	65	0	0	0

Total_calls_made	0	1	2	3	4	5	6	7	8	9	10
KMeans_group											
0	81	74	72	82	77	0	0	0	0	0	0
1	16	15	18	1	0	0	0	0	0	0	0
2	0	1	1	0	31	29	39	35	30	32	26

Figure 16: Cluster Profiling of KMeans group

Observations

- **Cluster 0: Moderate Credit Limit, Mixed Engagement** Customers have a moderate credit limit (3378) and credit cards (5.5). They visit the bank (3.49) but engage less online (0.98) and via calls (2).
- **Cluster 1: High Credit Limit, Digital-Savvy** This segment has the highest credit limit (10266) and credit cards (8.74). They prefer online banking (10.9 visits) and rarely visit the bank (0.6) or call (1.08).
- **Cluster 2: Low Credit Limit, High Call Volume** Customers have the lowest credit limit (2174) and credit cards (2.4). They make the most calls (6.87) but visit the bank (0.93) and engage online (3.55) moderately.
- **Online vs. Bank Visits** Higher online visits reduce physical visits, showing a clear shift in customer preferences.
- **Call Frequency** Cluster 2 requires more support, likely indicating service issues or digital unfamiliarity.

Business Recommendations

- **Enhance Online Banking for Digital-Savvy Customers (Cluster 1)** Since this group prefers online banking, the bank should invest in improving the digital experience, offering premium online services, and ensuring seamless mobile banking.
- **Improve Support Services for High-Call Customers (Cluster 2)** Customers in this segment require frequent assistance. The bank should provide AI-driven chatbots, better FAQ sections, and proactive customer education to reduce the call volume.
- **Strengthen In-Person Customer Engagement for Cluster 0** Since these customers prefer visiting the bank, the bank can introduce appointment-based services, dedicated relationship managers, and personalized assistance to enhance their experience.
- **Targeted Marketing for Upselling Credit Cards** Cluster 1 already has a high number of credit cards, so upselling additional cards may not be effective. Instead, focus on Cluster 0 customers, who hold around 5.5 cards, and could be encouraged to upgrade.
- **Channel Optimization Strategy** The bank should implement an omnichannel strategy, ensuring smooth transitions between digital and physical services to cater to all customer segments efficiently.

- Box plot of different columns vs KMeans groups

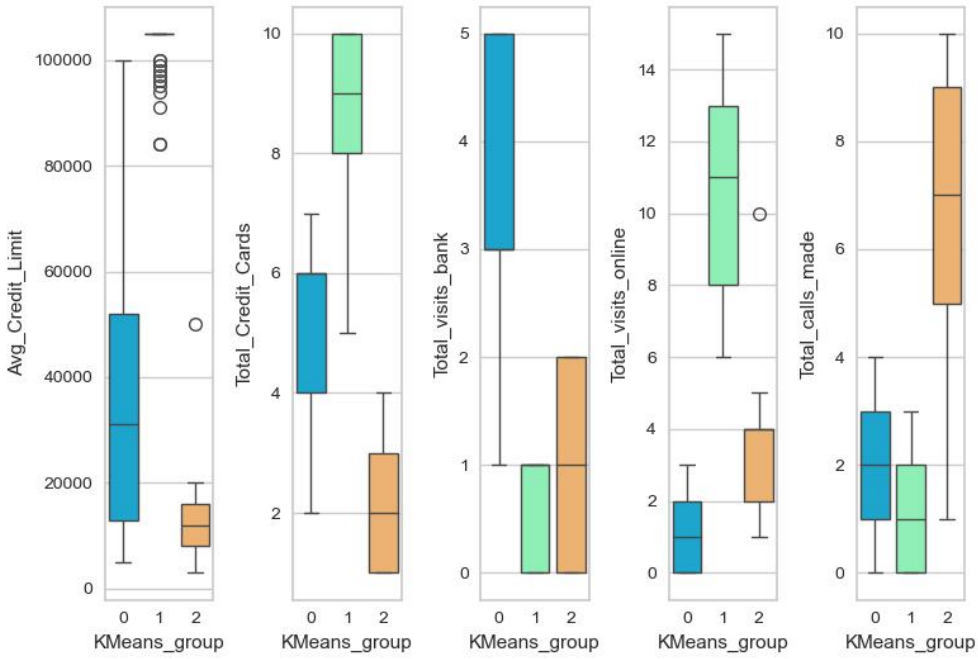


Figure 17: Box plot of different columns vs KMeans groups

Observations

- **Cluster 0: Mid-Range Users** - Moderate **credit limit**, **credit cards**, and **balanced** engagement across channels.
- **Cluster 1: High Credit, Digital Users** - **Highest credit limit**, **most cards**, prefer **online banking**, rarely visit the bank.
- **Cluster 2: Low Credit, High Support Users** - **Lowest credit limit**, **few cards**, **frequent calls**, indicating service issues or low digital use.
- **Service Channels** - Higher **online activity reduces** bank visits. **Cluster 2** relies more on **calls**.

Business Recommendations

- **Improve Digital Services** - Enhance **self-service tools** for Cluster 1.
- **Targeted Credit Offers** - Upsell Cluster 0, encourage **credit use** for Cluster 2.
- **Optimize Support** - AI chatbots and proactive help for Cluster 2.
- **Omnichannel Strategy** - Integrate **digital and traditional** services for a seamless experience.
- Pairplot of different columns vs KMeans groups

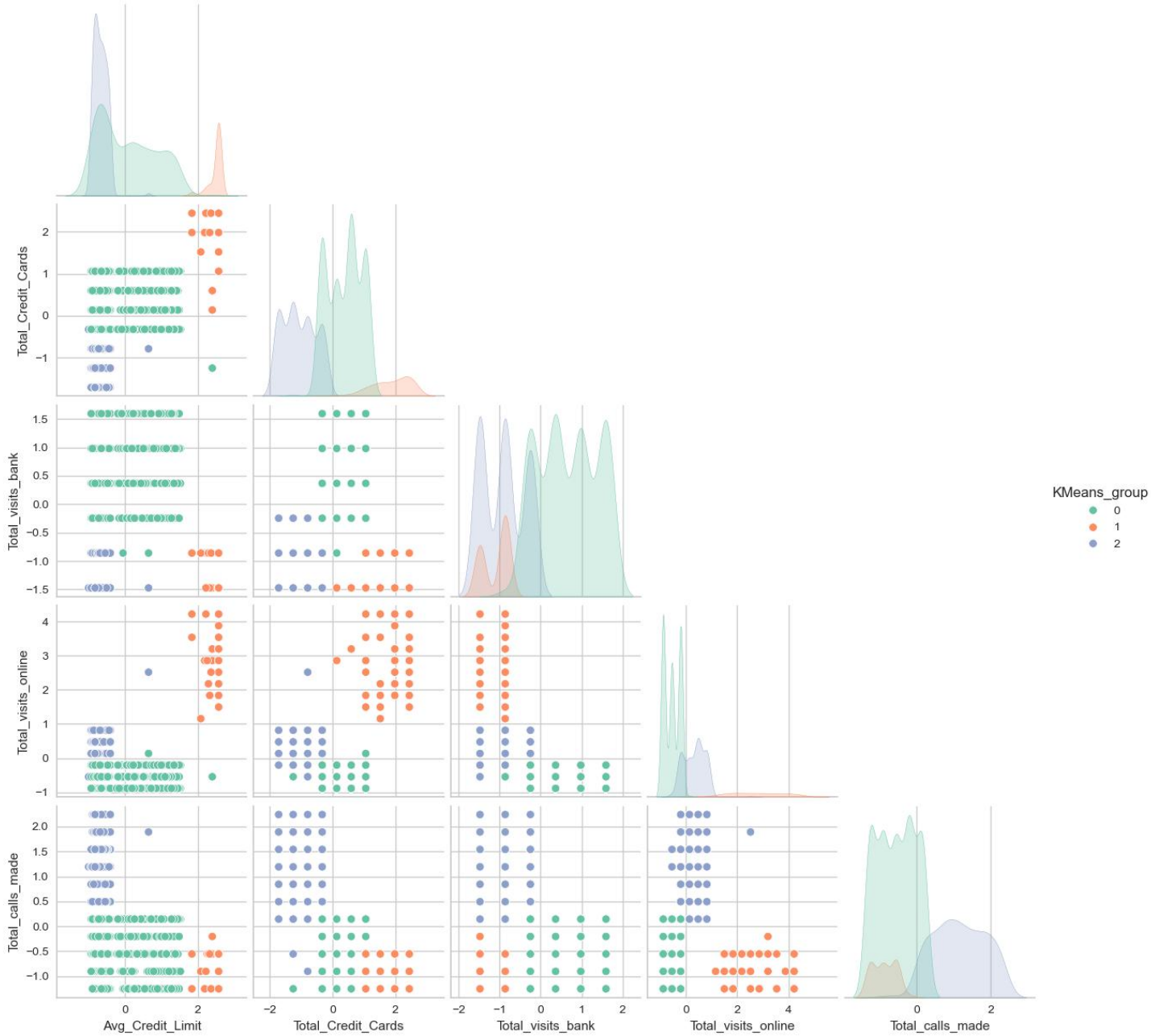


Figure 18: Pairplot of different columns vs KMeans groups

Observations

- **Cluster 0: Mid-Level Users** - Moderate **credit limit**, **balanced visits across channels**, and **few calls**.
- **Cluster 1: High Credit, Digital Users** - **Highest credit limit**, **many credit cards**, prefer **online banking**, minimal bank visits.
- **Cluster 2: Low Credit, High Support Users** - **Lowest credit limit**, **fewer cards**, **high call volume**, suggesting more service needs.
- **Channel Preference** - **Online visits reduce** bank visits. **Cluster 2** shows **higher dependency** on calls.

Business Recommendations

- **Enhance Digital Services** - Cluster 1 needs **improved self-service options** to strengthen engagement.
- **Credit Upsell Strategies** - Cluster 0 can be targeted for **credit expansion offers**.
- **Support Optimization** - AI-based **chatbots and proactive help** for Cluster 2 to **reduce call dependency**.
- **Integrated Experience** - Align **digital and traditional banking** to ensure **seamless service** across all clusters.
- 3D plot of different columns vs KMeans groups

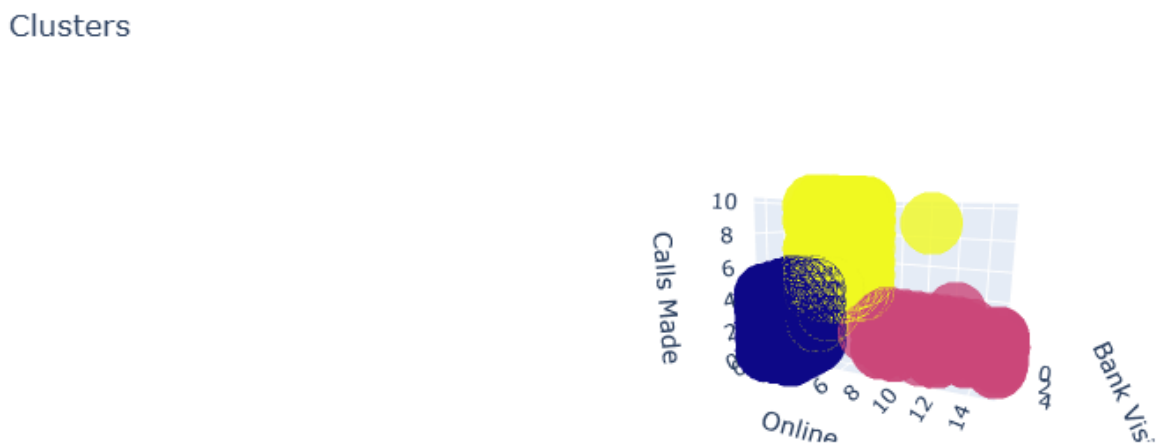


Figure 19: 3D plot of different columns vs KMeans groups

The plot gives a great way to visualize 3 different clusters in 3d space.

6.2 Hierarchical Clustering

Hierarchical clustering builds a hierarchy of clusters using a **dendrogram**. It can be:

- **Agglomerative** - Starts with **each point as a cluster** and merges them iteratively.
- **Divisive** - Starts with **one cluster** and splits it recursively.

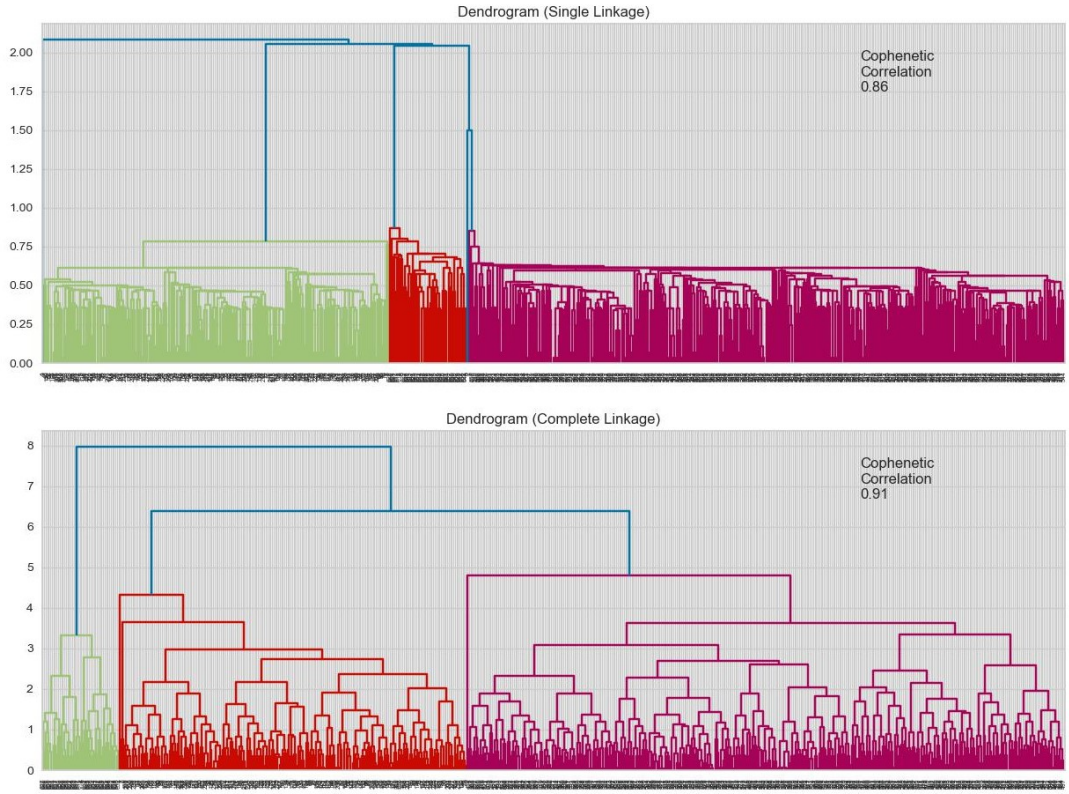
The choice of **linkage (single, complete, average, ward's)** affects clustering. It is useful for **visualizing cluster relationships** but can be computationally expensive.

6.2.1 Hierarchical clustering with different linkage methods

```
Cophenetic correlation for Braycurtis distance and single linkage is 0.8375544614800204.
Cophenetic correlation for Braycurtis distance and complete linkage is 0.5267857784120373.
Cophenetic correlation for Braycurtis distance and average linkage is 0.7880599893601058.
Cophenetic correlation for Braycurtis distance and weighted linkage is 0.6934546603183509.
Cophenetic correlation for Canberra distance and single linkage is 0.880520623317417.
Cophenetic correlation for Canberra distance and complete linkage is 0.7135324030750694.
Cophenetic correlation for Canberra distance and average linkage is 0.8158829456697183.
Cophenetic correlation for Canberra distance and weighted linkage is 0.7381468205101213.
Cophenetic correlation for Chebyshev distance and single linkage is 0.7853902350966868.
Cophenetic correlation for Chebyshev distance and complete linkage is 0.8318695955800918.
Cophenetic correlation for Chebyshev distance and average linkage is 0.9101241814670633.
Cophenetic correlation for Chebyshev distance and weighted linkage is 0.8971112755943227.
Cophenetic correlation for Cityblock distance and single linkage is 0.8937692082241637.
Cophenetic correlation for Cityblock distance and complete linkage is 0.8971781820037298.
Cophenetic correlation for Cityblock distance and average linkage is 0.9174837479004229.
Cophenetic correlation for Cityblock distance and weighted linkage is 0.8441151996862417.
Cophenetic correlation for Correlation distance and single linkage is 0.7434147385694589.
Cophenetic correlation for Correlation distance and complete linkage is 0.5671627831783883.
Cophenetic correlation for Correlation distance and average linkage is 0.808833971374701.
Cophenetic correlation for Correlation distance and weighted linkage is 0.6618470786522401.
Cophenetic correlation for Cosine distance and single linkage is 0.7582813557031118.
Cophenetic correlation for Cosine distance and complete linkage is 0.5520842533770532.
Cophenetic correlation for Cosine distance and average linkage is 0.794116433670034.
Cophenetic correlation for Cosine distance and weighted linkage is 0.6849926879442223.
Cophenetic correlation for Euclidean distance and single linkage is 0.8585604997359698.
Cophenetic correlation for Euclidean distance and complete linkage is 0.9142304605297017.
Cophenetic correlation for Euclidean distance and average linkage is 0.9200007623324095.
Cophenetic correlation for Euclidean distance and weighted linkage is 0.8751837821272724.
Cophenetic correlation for Hamming distance and single linkage is 0.6093628746891797.
Cophenetic correlation for Hamming distance and complete linkage is 0.7443538840983682.
Cophenetic correlation for Hamming distance and average linkage is 0.8158249696740025.
Cophenetic correlation for Hamming distance and weighted linkage is 0.8099847757371119.
Cophenetic correlation for Jaccard distance and complete linkage is 0.3292966588757202.
Cophenetic correlation for Jaccard distance and average linkage is 0.7268715596141124.
Cophenetic correlation for Jaccard distance and weighted linkage is 0.6963525193253491.
Cophenetic correlation for Mahalanobis distance and single linkage is 0.8483050422923873.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.4830878724188039.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.8116044877305987.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.8048941468069815.
Cophenetic correlation for Matching distance and single linkage is 0.6093628746891797.
Cophenetic correlation for Matching distance and complete linkage is 0.7443538840983682.
Cophenetic correlation for Matching distance and average linkage is 0.8158249696740025.
Cophenetic correlation for Matching distance and weighted linkage is 0.8099847757371119.
Cophenetic correlation for Minkowski distance and single linkage is 0.8585604997359698.
Cophenetic correlation for Minkowski distance and complete linkage is 0.9142304605297017.
Cophenetic correlation for Minkowski distance and average linkage is 0.9200007623324095.
Cophenetic correlation for Minkowski distance and weighted linkage is 0.8751837821272724.
Cophenetic correlation for Seuclidean distance and single linkage is 0.8508735593038139.
Cophenetic correlation for Seuclidean distance and complete linkage is 0.9136881896345714.
Cophenetic correlation for Seuclidean distance and average linkage is 0.919349046846001.
Cophenetic correlation for Seuclidean distance and weighted linkage is 0.8730649308803604.
Cophenetic correlation for Sqeuclidean distance and single linkage is 0.8558383323391816.
Cophenetic correlation for Sqeuclidean distance and complete linkage is 0.9177339466062053.
Cophenetic correlation for Sqeuclidean distance and average linkage is 0.9051607392311324.
Cophenetic correlation for Sqeuclidean distance and weighted linkage is 0.8542999408265486.
Highest cophenetic correlation is 0.9200007623324095, which is obtained with Euclidean distance and average linkage.
```

Figure 20: Among different distance and linkage methods, the highest cophenetic correlation is obtained using Euclidean distance and average linkage.

Let's view the dendrograms for the different linkage methods. A dendrogram, in general, is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering. The main use of a dendrogram is to work out the best way to allocate objects to clusters.



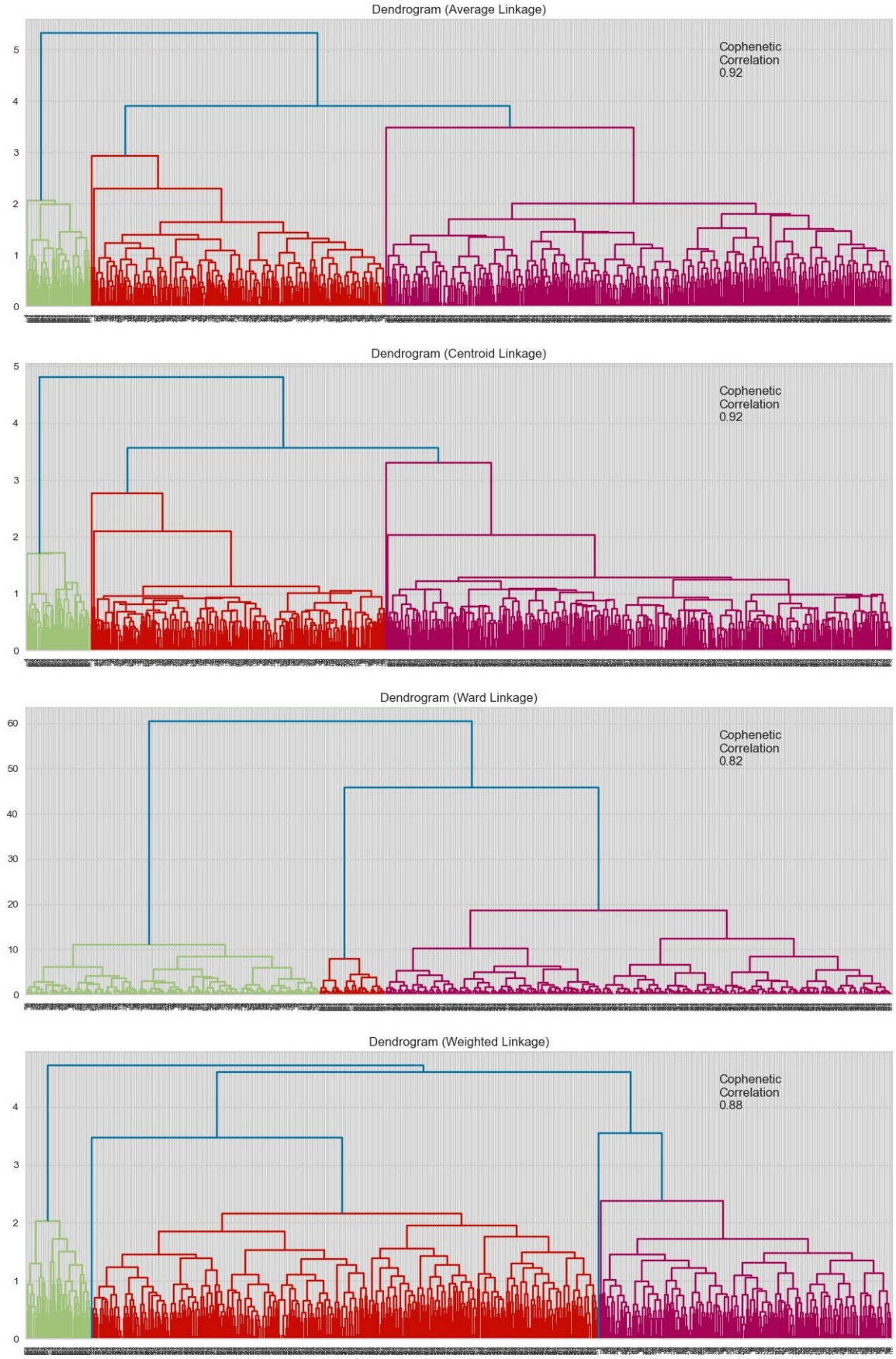


Figure 21: Dendrograms for the different linkage methods

Looking the the above dendrograms, the average linkage seems to result in the best separation between clusters, and its cophenetic correlation is lower than the other linkages. 3 looks to be a good choice for no. of clusters. Then we used hierarchical clustering with average linkage and euclidean distance and performed clustering only to find that both the clusters are identical (i.e. **we get the same set of clusters by different methods**). Hence cluster profiling is same for both methods.

6.2.2 Cluster Profiling

The cluster profiling remains the same as the K-Means grouping since both clustering methods yield similar results. The region for both the cluster resulting in same clusters is because the data has distinct, well-separated clusters which can be seen from the 3d plot.

Hierarchical_group	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	KMeans_group	count_of_customers
0	5.515544	3.489637	0.981865	2.000000	0.000000	386
1	8.740000	0.600000	10.900000	1.080000	1.000000	50
2	2.410714	0.933036	3.553571	6.870536	2.000000	224

Cluster Analysis and Recommendations

- **Cluster 0: Balanced Users** - Moderate credit cards, bank visits, and calls, but low online activity. **Action:** Encourage online banking to reduce call dependency.
- **Cluster 1: High Credit, Digital Users** - Most credit cards, high online usage, few bank visits, smallest group. **Action:** Offer premium services and credit perks to retain them.
- **Cluster 2: Low Credit, High Support Users** - Fewest cards, low bank visits, high call volume. **Action:** Use AI chatbots and credit-building programs for better service.

Strategy: Integrate digital and traditional channels for a seamless experience.

6.3 K-means vs Hierarchical Clustering

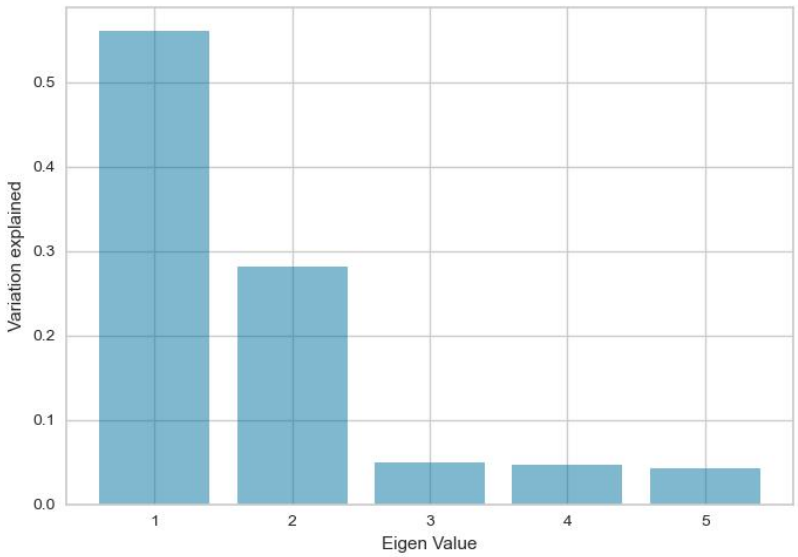
- The cluster profiling remains the same as the K-Means grouping since both clustering methods yield similar results. The region for both the cluster resulting in same clusters is because the data has distinct, well-separated clusters which can be seen from the 3d plot.
- Both methods obtained 3 clusters.
- The time taken by the **Hierarchical Clustering** is 0.34 seconds whereas the time taken by **K-means** is 1.14 seconds.
- In both the methods clusters 0,1,2 contained 386, 50 and 224 observations (data points) respectively.
- Silhouette Score for both the methods is same and is equal to 0.5967.

6.4 PCA for Visualization

PCA reduces dimensionality while preserving variance by transforming correlated features into orthogonal principal components (PCs).

Although there are only 5 dimensions, it'll be really cool to be able to visualize the clusters at 3 dimensional space without loosing much of the information. Let's use PCA to reduce the dimensions so that 90% of the variance in the data is explained.

```
Eigen Values:
 [3.77957988 1.90124379 0.33208971 0.31243381 0.28371041]
Eigen Vectors:
 [[-0.2165888  -0.37787563 -0.39119639  0.18601838  0.44169265  0.4623686
   0.4623686 ]
 [ 0.57570449  0.37713071 -0.32518818  0.61838124 -0.14210955  0.09486552
   0.09486552]
 [ 0.50422941 -0.4243117  -0.52105071 -0.49657115 -0.17497878 -0.09224411
  -0.09224411]
 [ 0.56135995 -0.0594997   0.44978562 -0.08823727  0.68623983 -0.01258524
  -0.01258524]
 [-0.16937766  0.63282797 -0.47008686 -0.33111273  0.45418194 -0.13027171
  -0.13027171]]
Percentage of variance explained by each eigen vector:
 [0.56091846 0.28215907 0.04928464 0.04636756 0.04210479]
```



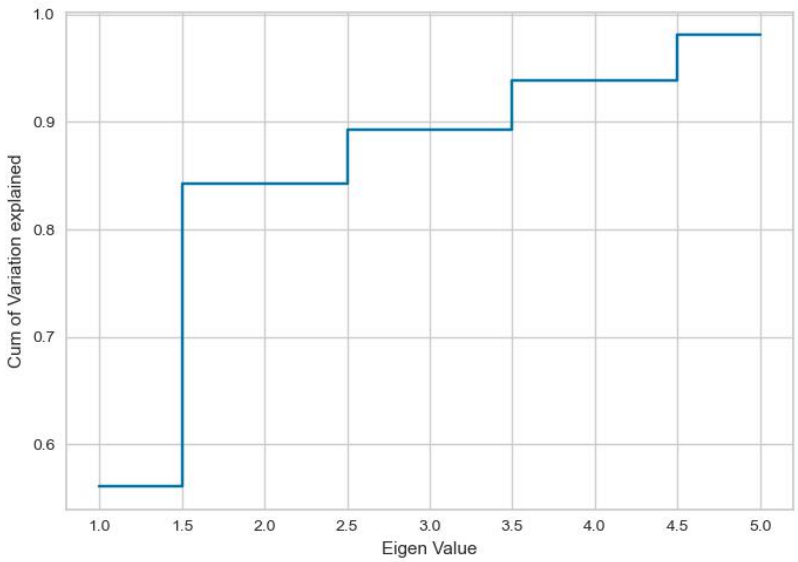


Figure 22: Xgboost Classifier performance

We can visualize data in 2 dimensions and also data in 3 dimensions (using 3-D plots). In some cases, we can also visualize data in 4 dimensions by using different hues for the 4th dimension in a 3-D plot. But it's impossible for us to visualize and interpret data in 5 dimensions.

So using PCA, we scaled down to 2 dimensions, and now it's easy for us to visualize the data.

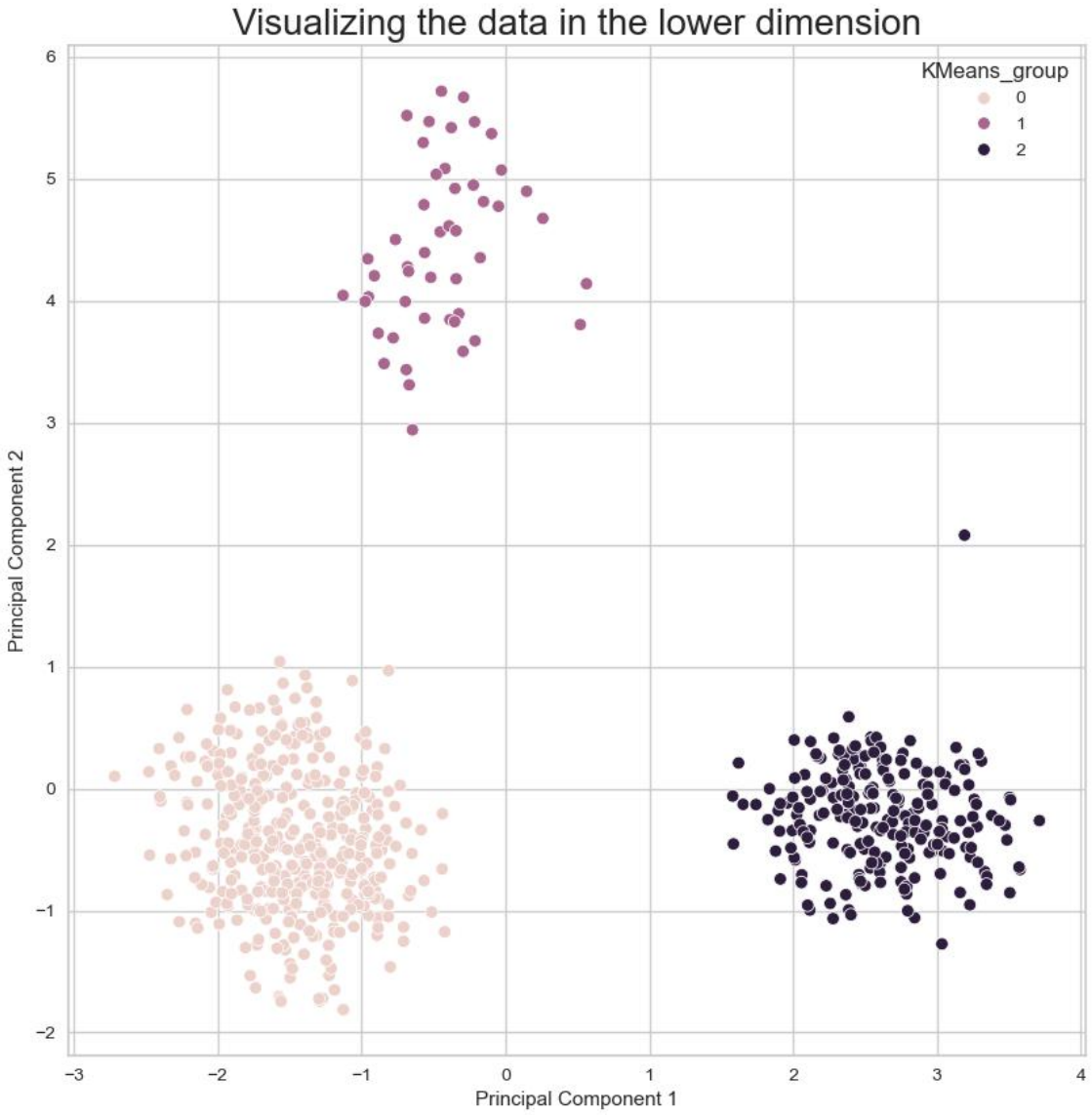


Figure 23: Visualizing data in 2 dimensions

In the above result, the explained variance is shown.

- The first principal component explains **56.1%** of the total variance in the data.
- The second principal component explains **28.2%** of the total variance in the data.
- Pairplot of different columns vs Hierarchical groups

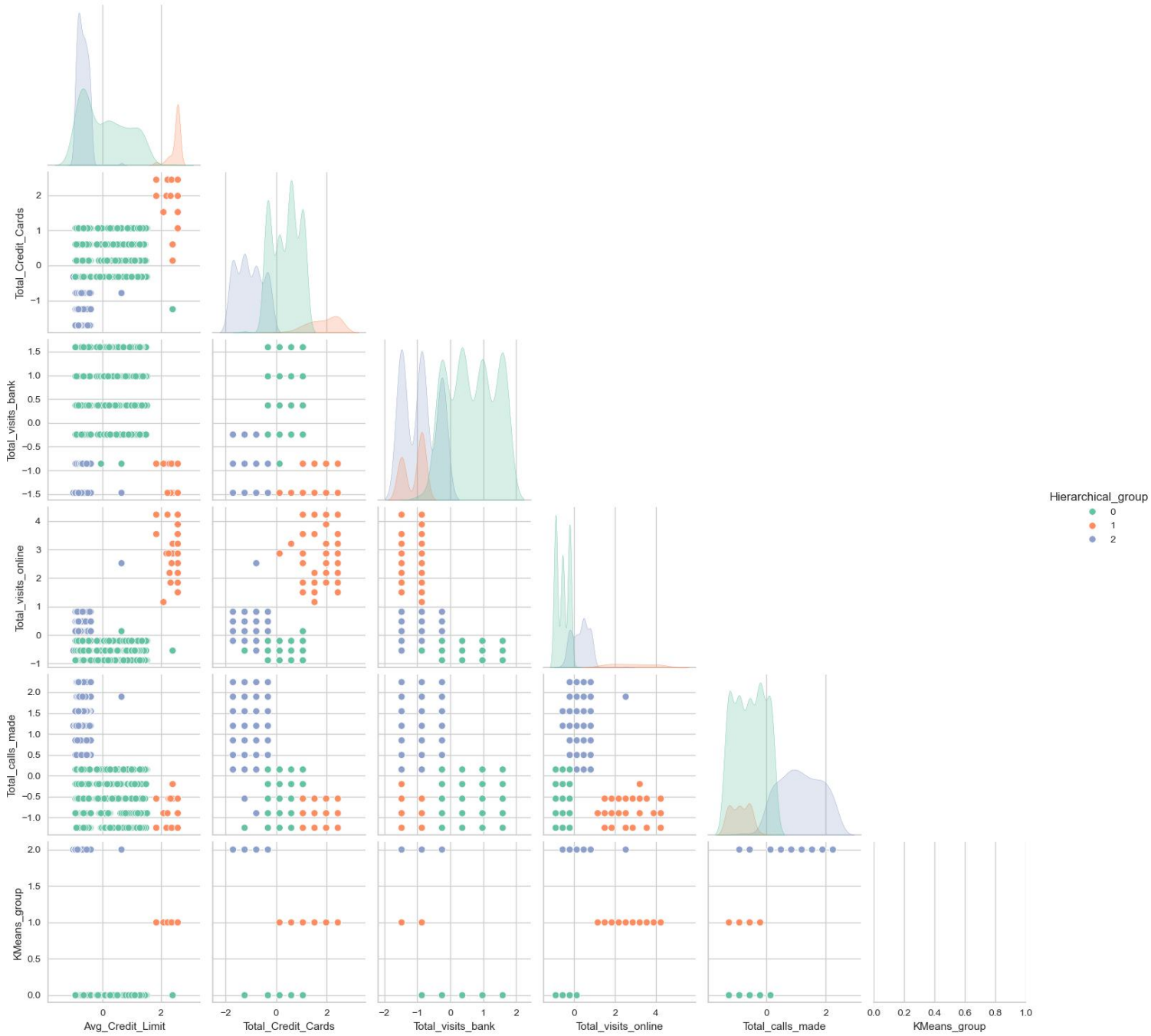


Figure 24: Pairplot of different columns vs Hierarchical groups

As expected the pairplots for both the groups are exactly identical.

6.5 PCA in 3 dimension

- Pairplot of PCA columns

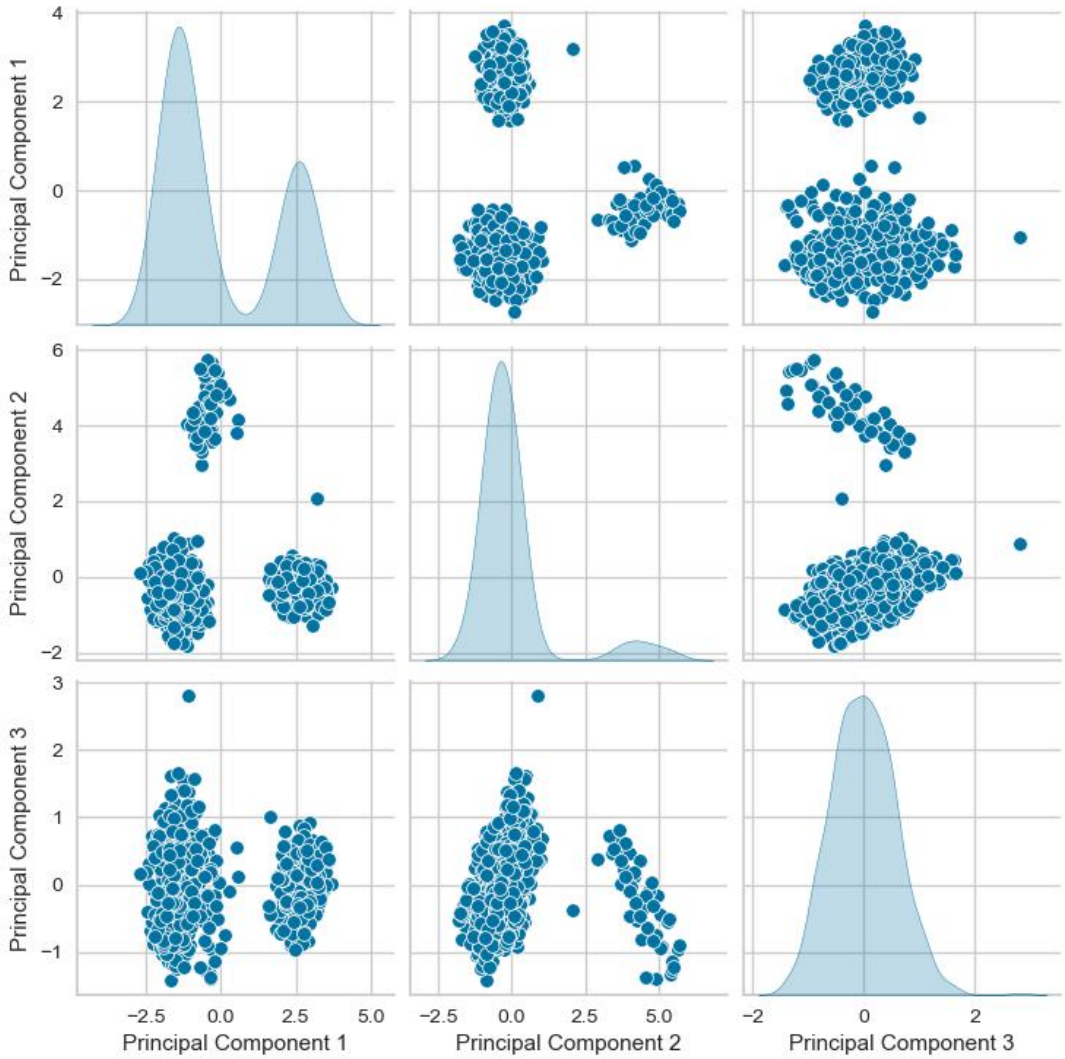


Figure 25: Pairplot of PCA columns

- **Cluster Separation:** Principal Components 1 and 2 show distinct clusters, indicating well-defined groups in the data.
- **Principal Component 3:** Less variance and more overlap, meaning it contributes less to differentiation.
- **Variance Distribution:** PC1 captures the highest variance, followed by PC2, while PC3 shows minimal spread.
- **Recommendations:**
  - Focus on PC1 and PC2 for customer segmentation and decision-making.
  - Reduce dimensionality by ignoring PC3 to simplify analysis.
  - Tailor marketing strategies based on the well-separated clusters in PC1 and PC2.

6.5.1 Hierarchical Clustering on lower-dimensional data

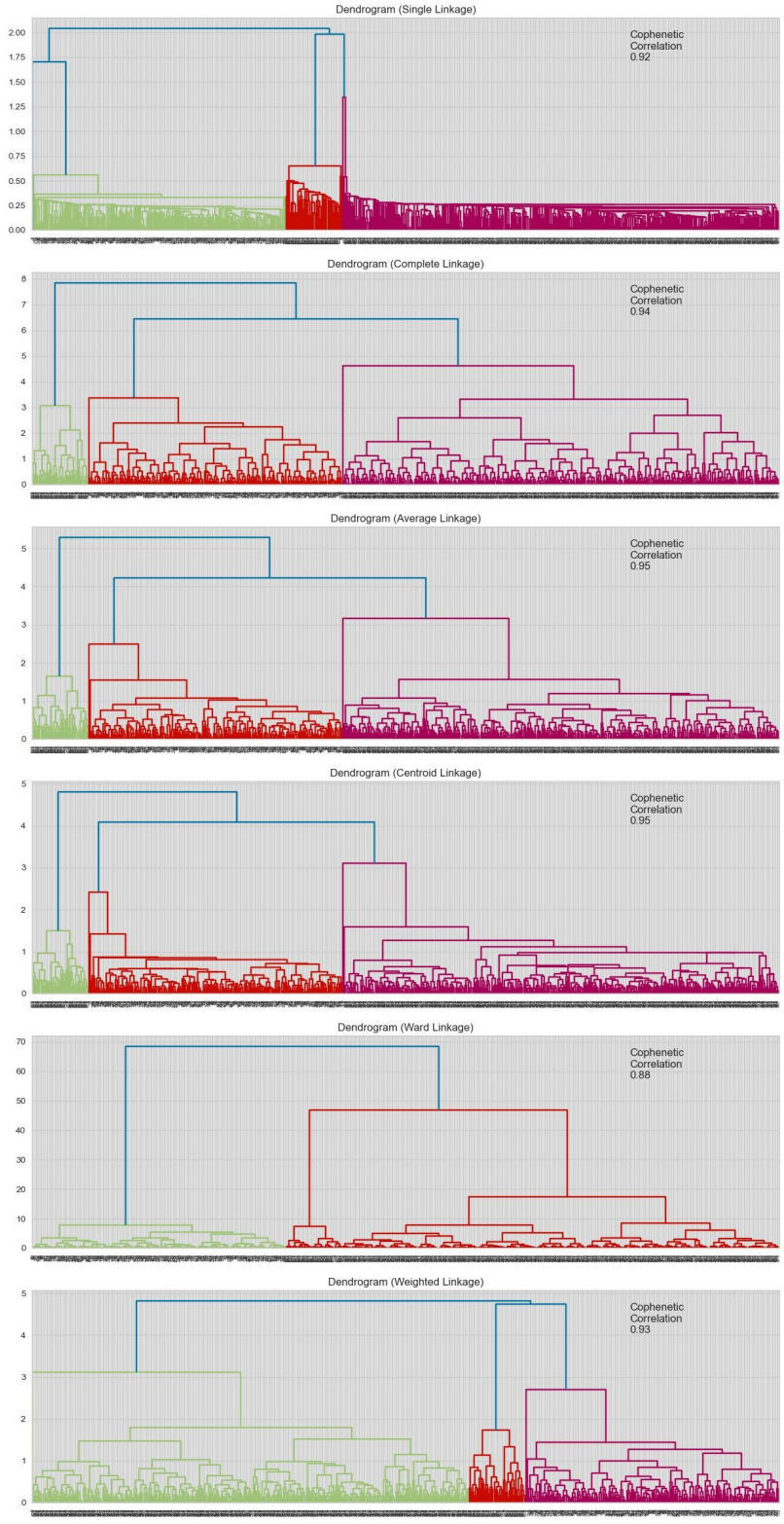


Figure 26: Dendrograms for the different linkage methods (Hierarchical Clustering on lower-dimensional data)

Observations

- The cophenetic correlation is highest for **average** and **centroid linkage**.
- **Average linkage** is preferred due to more distinct clusters, with a cophenetic correlation of **0.95**

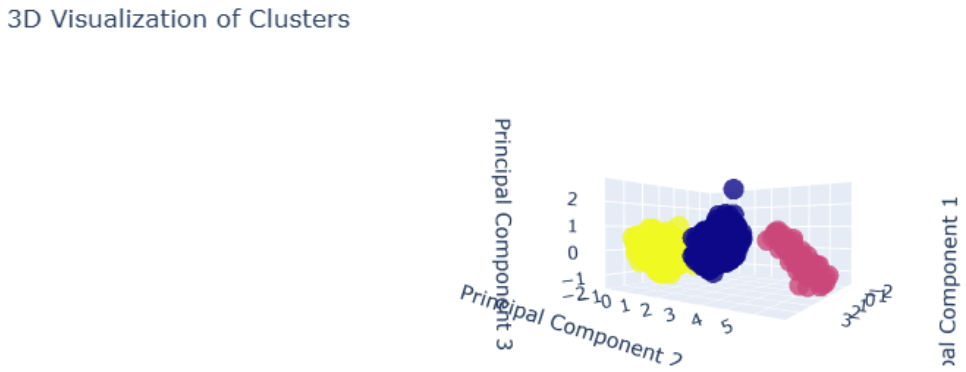


Figure 27: 3D plot of PCA columns

In the above result, the explained variance is shown.

- The first principal component explains **56.1%** of the total variance in the data.
- The second principal component explains **28.2%** of the total variance in the data.
- The third principal component explains **4.92%** of the total variance in the data.

7 Actionable Insights and Business Recommendations

1. Customer Segmentation Based on Banking Interaction

Insight: Customers exhibit different banking interaction patterns:

- **Cluster 0: Balanced Users** - Moderate credit limits, balanced engagement across online, calls, and in-person visits.
- **Cluster 1: High Credit, Digital Users** - High credit limits, frequent online banking usage, minimal bank visits.
- **Cluster 2: Low Credit, High Support Users** - Low credit limits, frequent calls, low online engagement.

Recommendation: Implement an omnichannel approach to ensure seamless transitions between digital and traditional banking services.

2. Enhancing Digital Banking for High-Credit Users

Insight: Cluster 1 customers have the highest credit limits and prefer digital banking, minimizing in-branch visits.
Recommendation:

- Enhance self-service features like chatbots and digital advisors.
- Offer exclusive online promotions and premium services to retain high-value customers.

3. Optimizing Support for High-Call Volume Customers

Insight: Cluster 2 customers make frequent calls, indicating high service dependency.
Recommendation:

- Introduce AI-driven chatbots and enhanced FAQ sections to reduce call volumes.
- Proactively educate customers on digital banking tools to improve self-service capabilities.

4. Targeted Credit Upsell Strategies

Insight: Cluster 0 customers have a mid-range credit limit and hold an average of 5.5 credit cards, making them ideal for credit expansion.
Recommendation:

- Offer personalized credit limit increases and additional credit card options.
- Use behavioral analytics to identify the right moment for credit upsell campaigns.

5. Relationship-Based In-Person Banking for Balanced Users

Insight: Cluster 0 customers still visit the bank but have moderate online engagement.
Recommendation:

- Introduce appointment-based services for personalized banking experiences.
- Assign dedicated relationship managers to high-value customers preferring in-branch interactions.

6. Reducing Branch Dependency Through Digital Adoption

Insight: A strong negative correlation exists between online visits and in-person banking, highlighting a shift towards digital preferences.
Recommendation:

- Promote digital literacy campaigns to encourage online banking adoption.
- Provide incentives for first-time digital banking users.

7. Improving Customer Retention with Personalized Engagement

Insight: Customers with multiple credit cards tend to require less support and show higher retention rates.
Recommendation:

- Develop targeted loyalty programs for multi-card holders.
- Provide exclusive financial planning assistance to high-credit customers.

Conclusion

By leveraging clustering analysis, AllLife Bank can implement a data-driven strategy to enhance customer experience, optimize service delivery, and drive credit card growth. Integrating digital and traditional banking solutions will create a seamless experience across all customer segments.