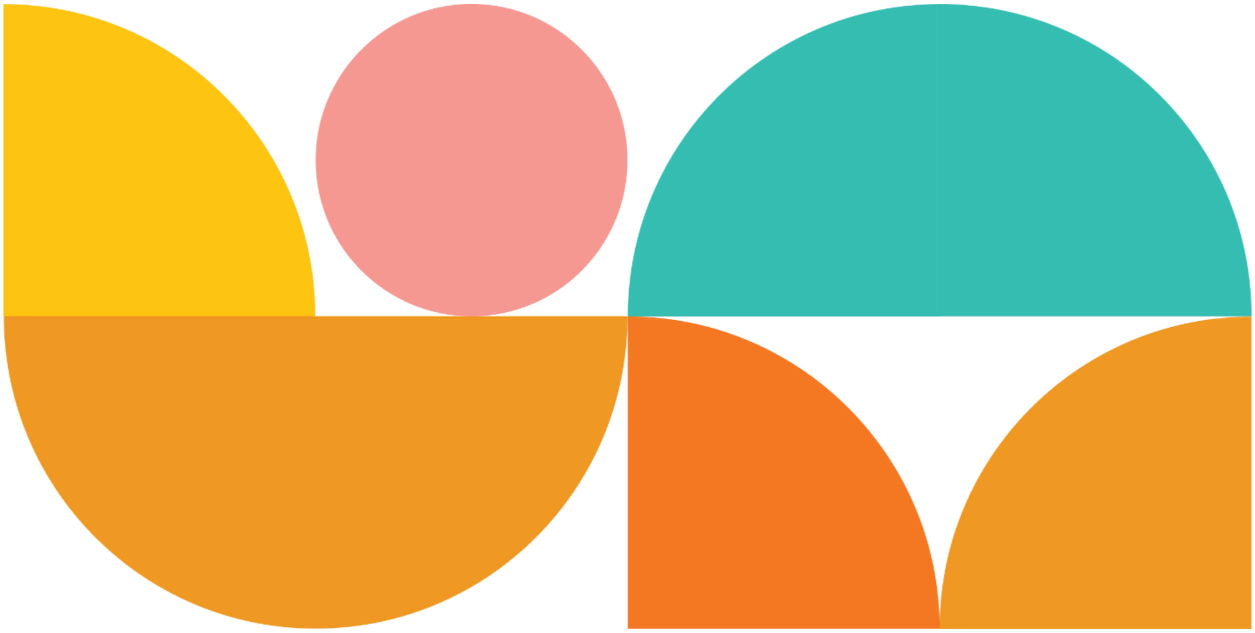
03/11/2024

Unsupervised Learning

Business Report (Coded - Project)



NABANKUR RAY
PGP-DSBA

Contents

List of figures

List of Tables

Problem Statement - UL Project - Coded

Business Context

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

Objective

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

Data 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).

Data Dictionary

- **Sl_No:** Primary key of the records
- **Customer Key:** Customer identification number
- **Average Credit Limit:** Average credit limit of each customer for all credit cards
- **Total credit cards:** Total number of credit cards possessed by the customer
- **Total visits bank:** Total number of visits that the customer made (yearly) personally to the bank
- **Total visits online:** Total number of visits or online logins made by the customer (yearly)

- **Total calls made:** Total number of calls made by the customer to the bank or its customer service department (yearly)

Executive Summary:

This analysis focuses on understanding AllLife Bank’s credit card customers through data exploration, aiming to assist the Marketing and Operations teams in crafting tailored campaigns and enhancing customer service. The univariate and bivariate analyses of customers' credit usage and interaction patterns reveal distinct behavioral segments, which could guide the bank's strategy for upselling and improving support services.

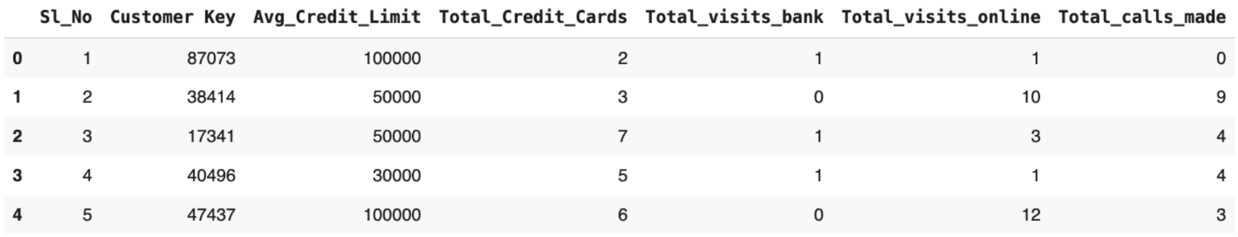
Deliverables:

- **Exploratory Data Analysis (EDA)**: In-depth analysis of credit usage and interaction preferences.
- **Actionable Insights**: Identification of patterns in spending and service usage.
- **Business Recommendations**: Suggested strategies for personalized marketing and enhanced service delivery.
- **Conclusion**: Summary of findings and final recommendations.

Understanding the Data

Data Overview

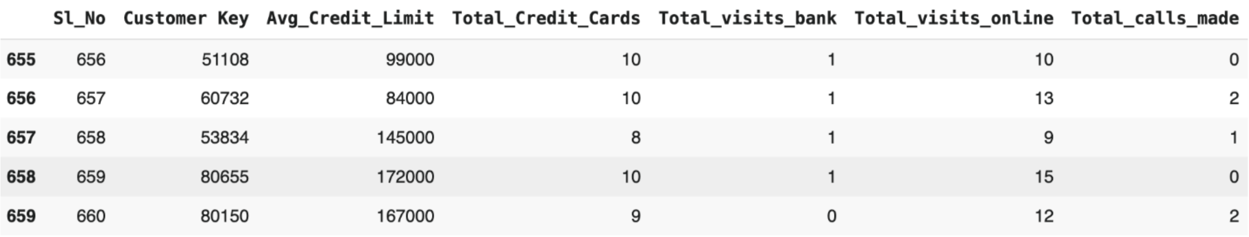
Displaying the first 5 rows:



	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	1	87073	100000	2	1	1	0
1	2	38414	50000	3	0	10	9
2	3	17341	50000	7	1	3	4
3	4	40496	30000	5	1	1	4
4	5	47437	100000	6	0	12	3

Table 1: First 5 rows of the given dataset

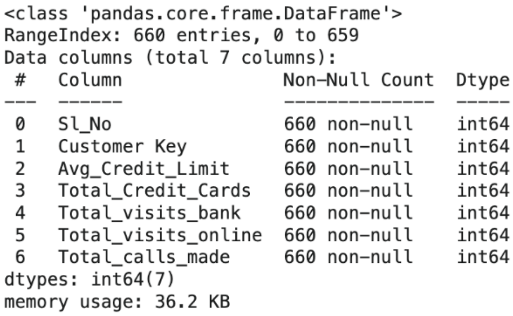
Displaying the last 5 rows:



	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
655	656	51108	99000	10	1	10	0
656	657	60732	84000	10	1	13	2
657	658	53834	145000	8	1	9	1
658	659	80655	172000	10	1	15	0
659	660	80150	167000	9	0	12	2

Table 2: Last 5 rows of the given dataset

Structure and Types of Data



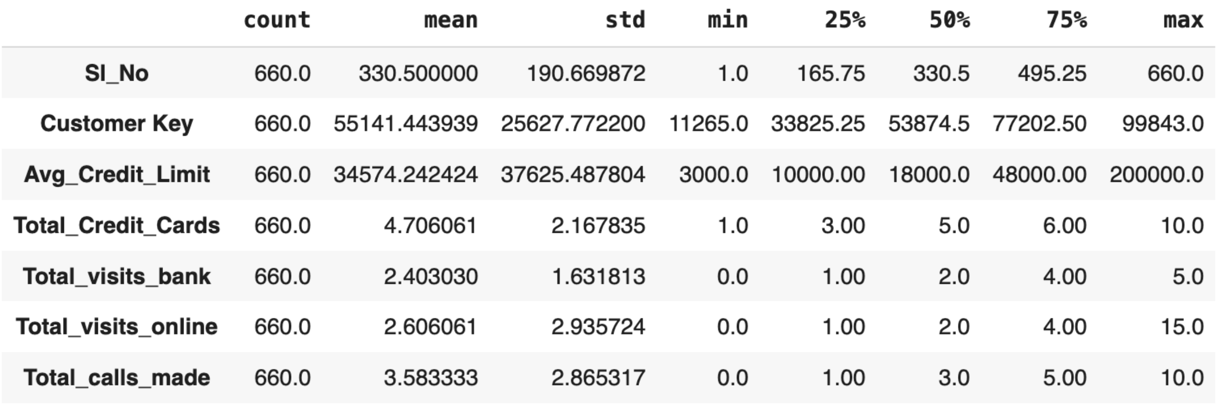
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 660 entries, 0 to 659
Data columns (total 7 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   Sl_No                660 non-null    int64
 1   Customer Key         660 non-null    int64
 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: Checking the structure and type of data

OBSERVATIONS:

- There are **660 rows** and **7 Columns** are present in the given datasets.
- It can be observed that no columns have less entries (less than 660 rows) which indicates that there are no missing values in the given dataset.
- Here, all the columns are numerical (int data type).
- Dependent variable is the "Total calls made" which is of numerical type.

Statistical summary of the Numerical Data



	count	mean	std	min	25%	50%	75%	max
Sl_No	660.0	330.500000	190.669872	1.0	165.75	330.5	495.25	660.0
Customer Key	660.0	55141.443939	25627.772200	11265.0	33825.25	53874.5	77202.50	99843.0
Avg_Credit_Limit	660.0	34574.242424	37625.487804	3000.0	10000.00	18000.0	48000.00	200000.0
Total_Credit_Cards	660.0	4.706061	2.167835	1.0	3.00	5.0	6.00	10.0
Total_visits_bank	660.0	2.403030	1.631813	0.0	1.00	2.0	4.00	5.0
Total_visits_online	660.0	2.606061	2.935724	0.0	1.00	2.0	4.00	15.0
Total_calls_made	660.0	3.583333	2.865317	0.0	1.00	3.0	5.00	10.0

Table 4: Statistical summary of the numerical data

OBSERVATIONS:

- The **Customer Key** values range widely from 11,265 to 99,843, with a mean of 55,141. This suggests a diverse set of customer identifiers across the dataset.

- The **Avg_Credit_Limit** shows significant variability, with a mean of 34,574, but ranging from as low as 3,000 to as high as 200,000. This could indicate a wide range of credit profiles among the customers, with a notable difference in credit limits.

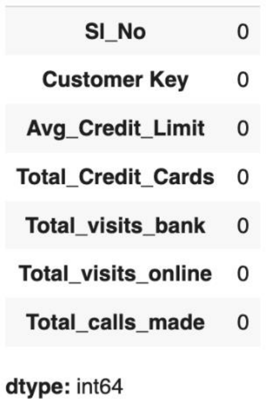
- The average number of credit cards held by customers is approximately 4.7, with a standard deviation of about 2.2. Most customers have between 3 and 6 credit cards, but there are some customers with up to 10 cards.

- The **Total_visits_bank** and **Total_visits_online** variables show similar averages, around 2.4 and 2.6, respectively. However, the online visits have a higher standard deviation, indicating more variability in online interactions. This could suggest that while many customers interact similarly in person and online, some rely much more heavily on online channels.

- The **Total_calls_made** metric has a mean of about 3.6, with a fairly large spread (standard deviation of 2.87). The maximum value is 10, indicating that some customers make frequent calls, perhaps for support or inquiries.

- For most metrics (like credit limit, visits, and calls), the median (50th percentile) is closer to the lower end than the upper end, suggesting a right-skewed distribution. For example, in Avg_Credit_Limit, the median is 18,000, which is lower than the mean of 34,574, indicating that a few high-value customers drive up the average.

Checking for missing values



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

dtype: int64
```

Table 5: No. of Missing Values in each column

INFERENCE:

- There is no missing values.
- There is no duplicate values.
- There is no null values.

Exploratory Data Analysis

Univariate Analysis

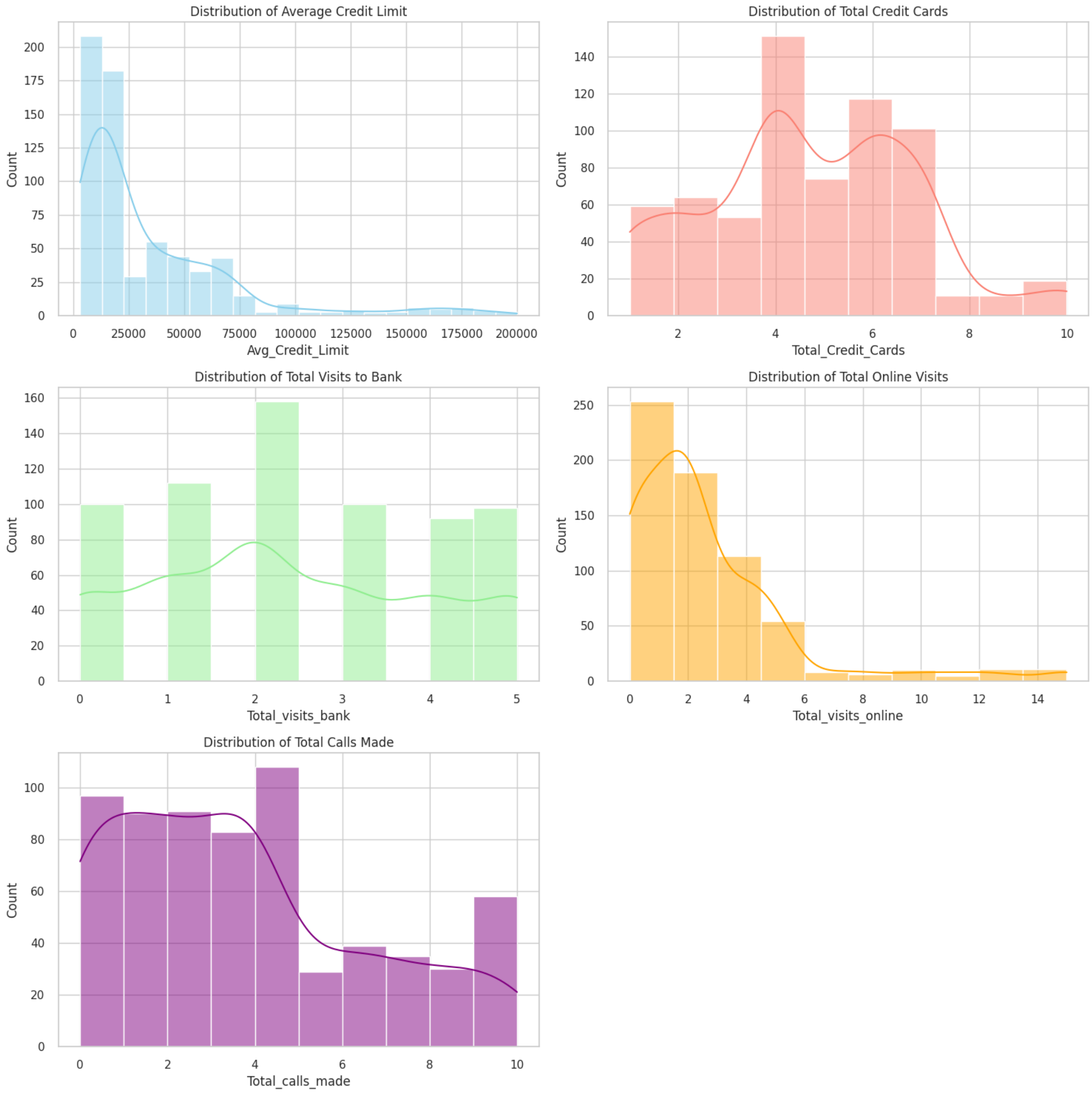


Fig 1: Histogram distribution for the given datasets

OBSERVATIONS:

Avg_Credit_Limit: The distribution is right-skewed, indicating that most customers have lower credit limits, with only a few having high credit limits.

Total_Credit_Cards: The distribution suggests that customers typically have between 2 to 6 credit cards, with fewer customers holding an unusually high number.

Total_visits_bank: The majority of customers make very few in-person visits to the bank, with most values clustering around zero, indicating a preference for other service channels.

Total_visits_online: Many customers frequently use online channels, with some making more than 10 online visits per year.

Total_calls_made: Similar to online visits, customers make a substantial number of calls, though the frequency varies, suggesting that calls may be a secondary channel after online visits.

Bivariate Analysis

Correlation Check

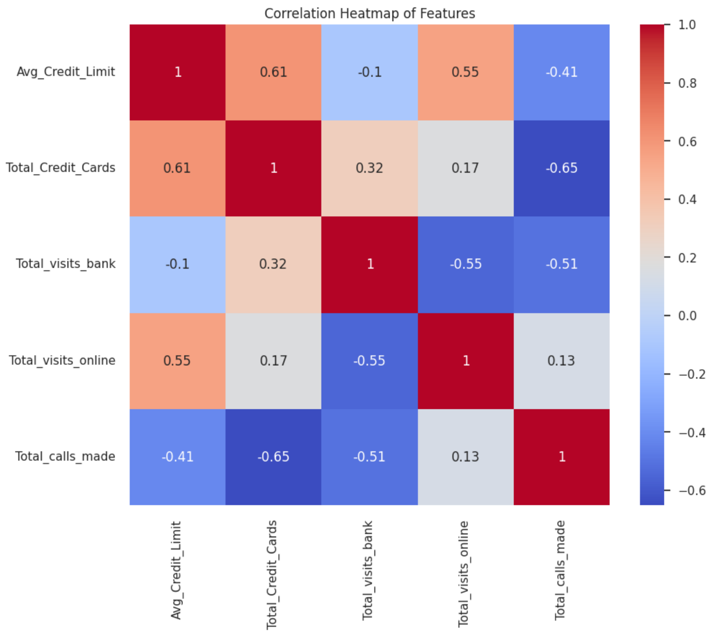


Fig 2: Heatmap for the Numerical Columns

OBSERVATIONS:

- **Avg_Credit_Limit and Total_Credit_Cards**: There is a moderate positive correlation, suggesting that customers with more credit cards tend to have higher average credit limits.

- **Total_visits_online and Total_calls_made**: A positive correlation exists, indicating that customers who frequently visit online channels also tend to make more calls, possibly reflecting a segment that prefers remote service channels.

- **Total_visits_bank**: This variable has weak correlations with other variables, implying that in-person visits do not strongly align with spending behavior or remote service usage.

Pair plot

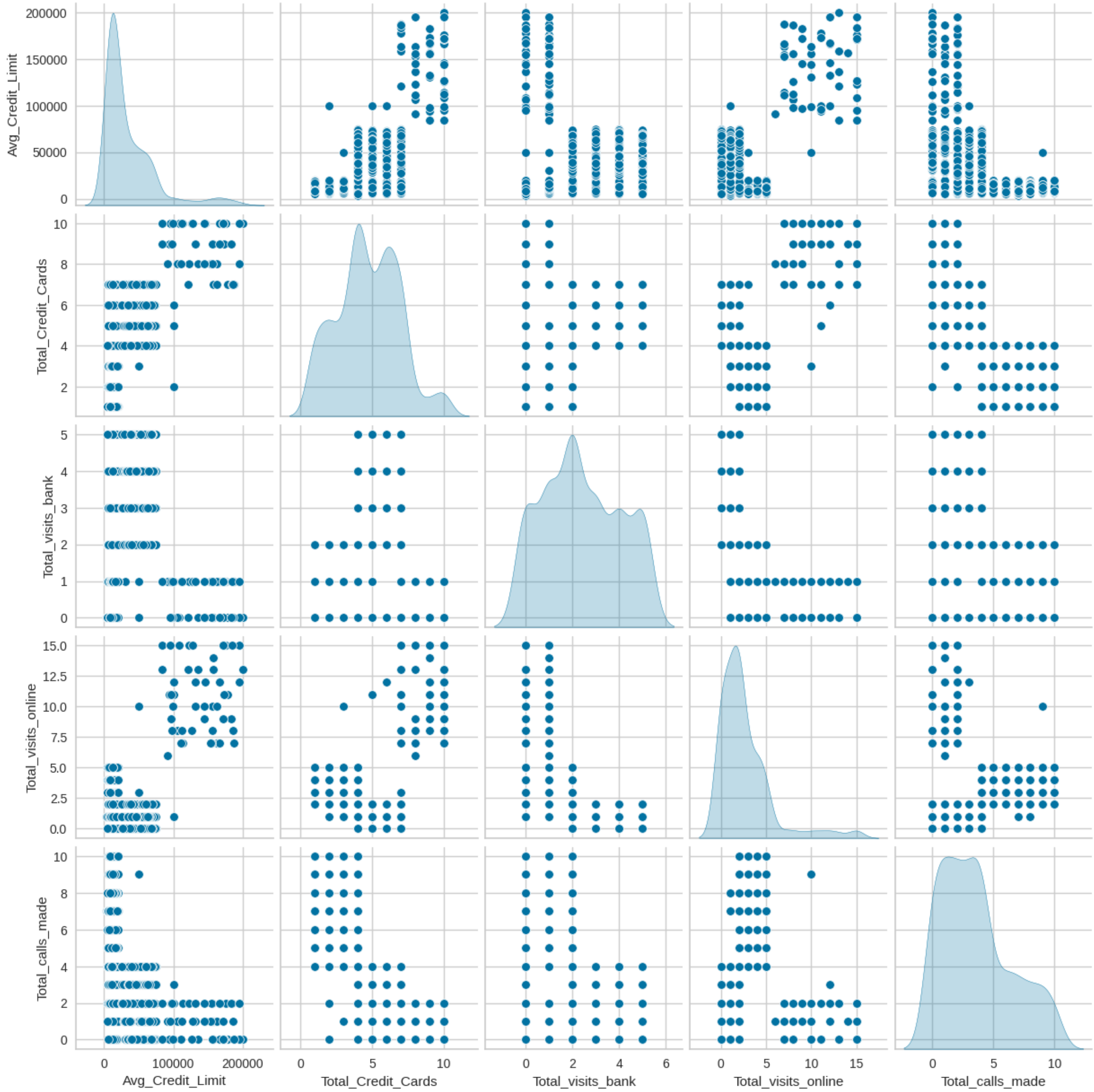


Fig 3: Pair Plot for the Numerical Columns

OBSERVATIONS:

Avg_Credit_Limit Distribution:

- The Avg_Credit_Limit feature shows a right-skewed distribution with a high concentration of customers having lower credit limits and only a few with very high limits.
- This aligns with the summary statistics and indicates that high credit limits are rare within this customer group.

Total_Credit_Cards:

- The distribution of Total_Credit_Cards shows distinct levels, suggesting that customers tend to have a specific number of cards (e.g., 3, 5, or 6) rather than values that vary continuously.
- This could imply some customer segmentation or thresholds based on certain credit profiles.

Total_visits_bank and Total_visits_online:

- Total_visits_bank has a relatively normal distribution, centered around 2-3 visits. Most customers make only a few in-person bank visits.
- Total_visits_online, however, is more spread out, with some customers making over 10 online visits. This suggests that while in-person visits are moderate, online activity varies more widely across customers.

Total_calls_made:

- Similar to Total_visits_online, Total_calls_made shows considerable variability, with some customers making many calls, though the majority seem to make 3-5 calls.
- This variation may indicate differences in support needs or preferences in communication channels.

Relationships Between Variables:

- **Avg_Credit_Limit and Total_Credit_Cards**: There is a positive trend between Avg_Credit_Limit and Total_Credit_Cards, suggesting that customers with higher credit limits tend to hold more cards.
- **Avg_Credit_Limit and Total_calls_made**: Customers with very high credit limits appear to make fewer calls overall, indicating they may require less frequent customer support.
- **Total_visits_online and Total_calls_made**: There is a noticeable cluster of customers with low online visits and fewer calls, as well as another group with high online engagement and more calls, hinting at different customer engagement styles.

Exploratory Data Analysis – Insights

1. **Customer Spending Patterns**:
   a. Most customers have lower average credit limits, with a select few holding high limits. Higher credit limits generally align with customers who possess multiple credit cards.
   b. The moderate correlation between credit limits and the number of credit cards suggests that customers with diversified credit profiles (more cards) may represent higher-value segments for upselling.
2. **Customer Interaction Channels**:
   a. In-person visits are infrequent, indicating a preference for remote service options among the customer base.
   b. Online visits and call frequency show positive correlation, highlighting a segment that relies heavily on digital and phone channels. This group may benefit most from enhanced online services and prompt call support.
3. **Channel Segmentation Potential**:
   a. Customers exhibit distinct preferences for certain interaction channels, suggesting opportunities to create segmented service models: digital-focused, call-focused, and high-touch in-person segments.

Data Preprocessing

Outlier Check

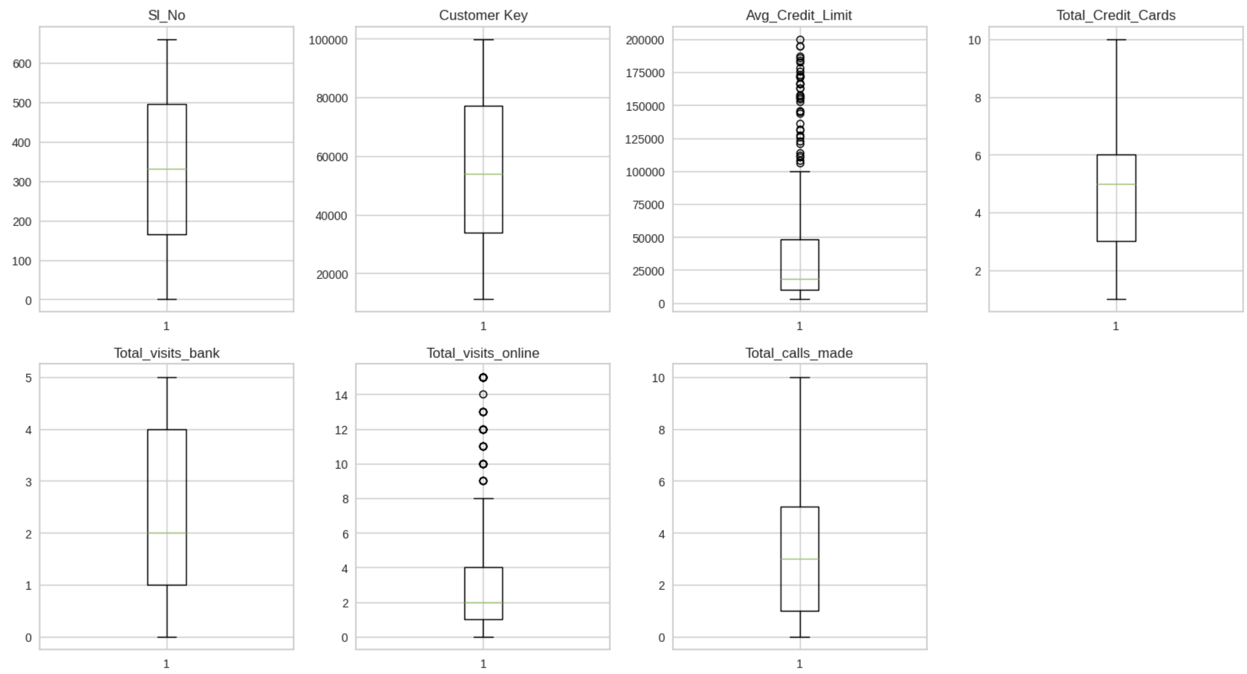


Fig 4: Boxplot of all numerical values

- **Total_credit_cards, Total_visits_bank, Total_calls_made** don't have visible outliers. Therefore, no need to outlier treatment.

- **Avg_Credit_Limit:** This variable has a significant number of outliers on the higher end. The distribution is skewed, with many data points lying above the upper quartile. This indicates that while most customers have an average credit limit within a lower range, a few customers have high credit limits. The outliers in this case are normal. Therefore, no need to outlier treatment.

- **Total_visits_online:** This variable has a few outliers above the upper quartile, indicating that while most customers visit the online portal a relatively low number of times, some customers have a significantly higher frequency of online visits. The outliers in this case are also normal. Therefore, no need to outlier treatment.

INFERENCE

- There are no outliers that require treatment.
- The data is scaled for clustering.

K-means Clustering

- Finding optimal no. of clusters (value of k) with -

Elbow Method

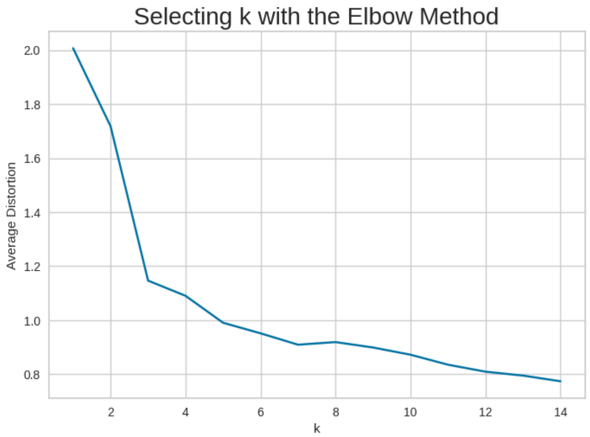


Fig 5: Elbow Curve

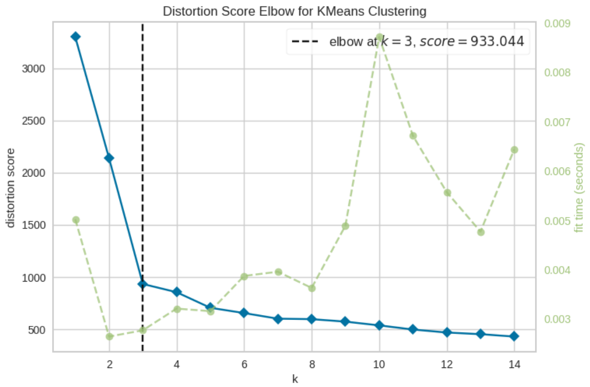


Fig 6: Elbow Curve with optimal value of K

- The elbow curve shows a significant drop in avg distortion to around **3 clusters**, suggesting that this might be an optimal point.
- **The appropriate value of k from the elbow curve seems to be 3.**

Silhouette scores

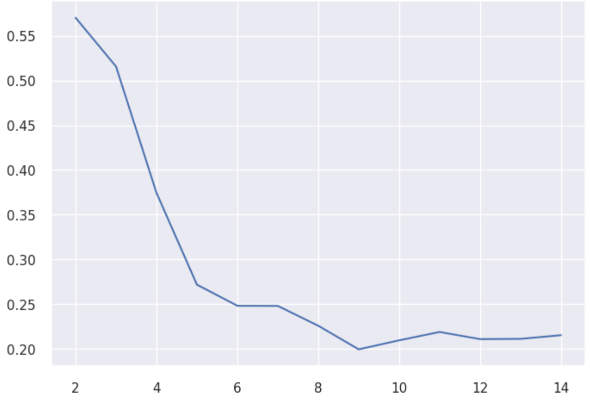


Fig: 7: Silhouette scores

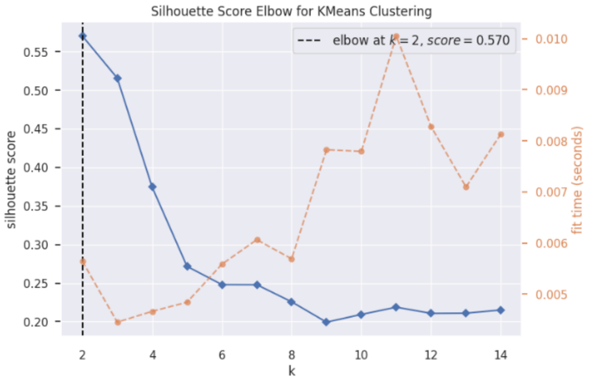


Fig 8: Silhouette scores with optimal value of K

- The silhouette scores are highest at **2 clusters** (0.5157) but remain reasonably high for 3 clusters (0.3557), after which the scores decrease steadily.
- **The appropriate value of k from the Silhouette scores seems to be 2.**

silhouette coefficients

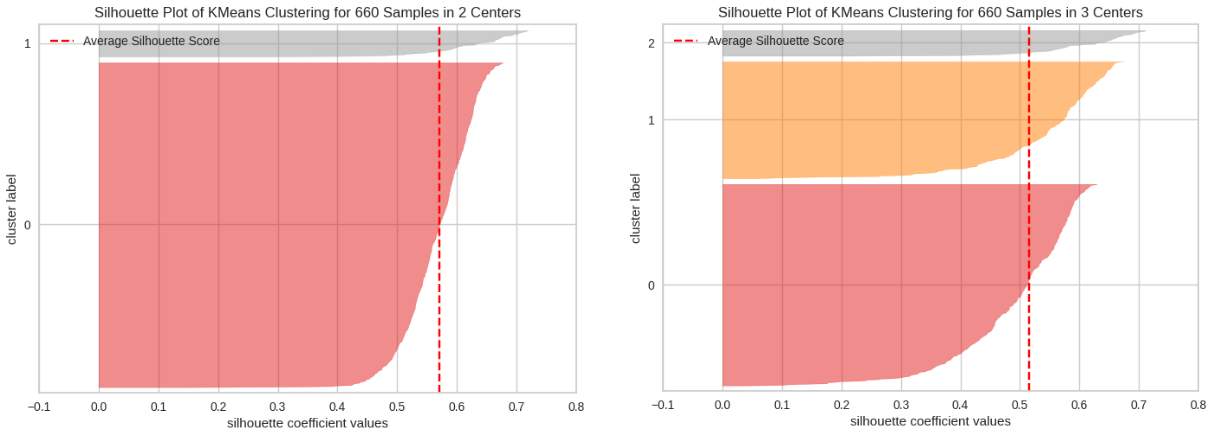
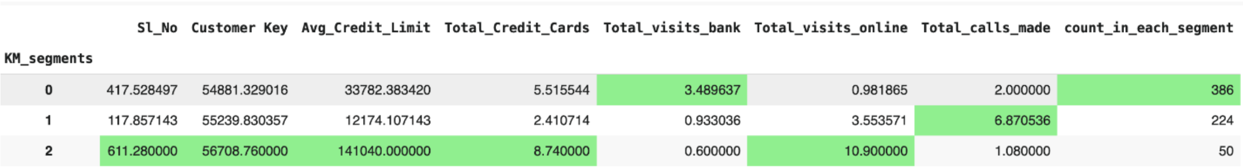


Fig 9: Silhouette plot of K-means Clustering for 660 samples in 2 & 3 centers

Based on this, Let’s proceed with clustering using **3 clusters** to balance interpretability and potential insights as the silhouette score is high enough and there is knick at 3 in the elbow curve.

Cluster Profiling



	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
KM_segments								
0	417.528497	54881.329016	33782.383420	5.515544	3.489637	0.981865	2.000000	386
1	117.857143	55239.830357	12174.107143	2.410714	0.933036	3.553571	6.870536	224
2	611.280000	56708.760000	141040.000000	8.740000	0.600000	10.900000	1.080000	50

Fig 10: Average of Numerical attributes based on different K-means cluster

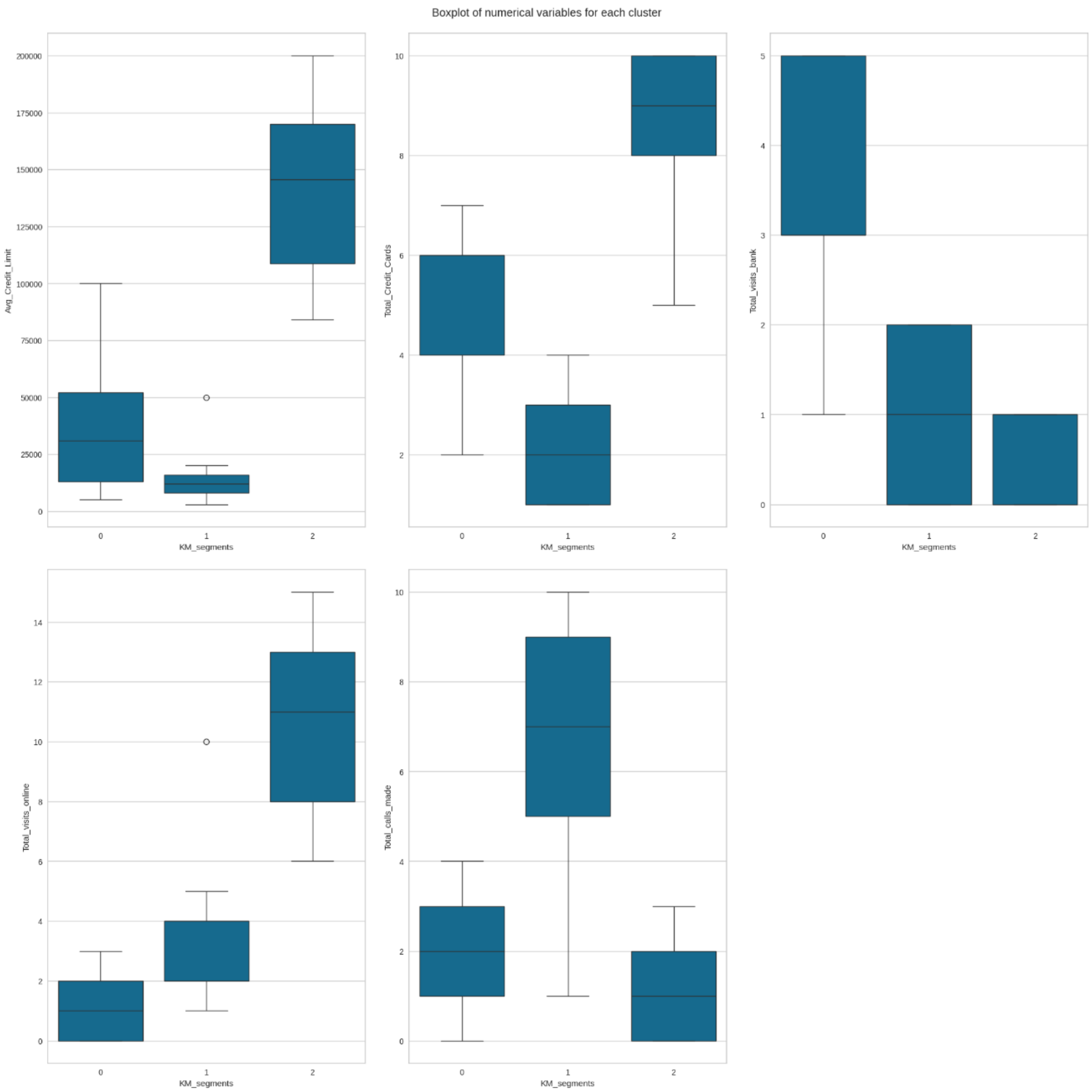


Fig 11: Boxplot of Numerical attributes vs K-means cluster

Insights

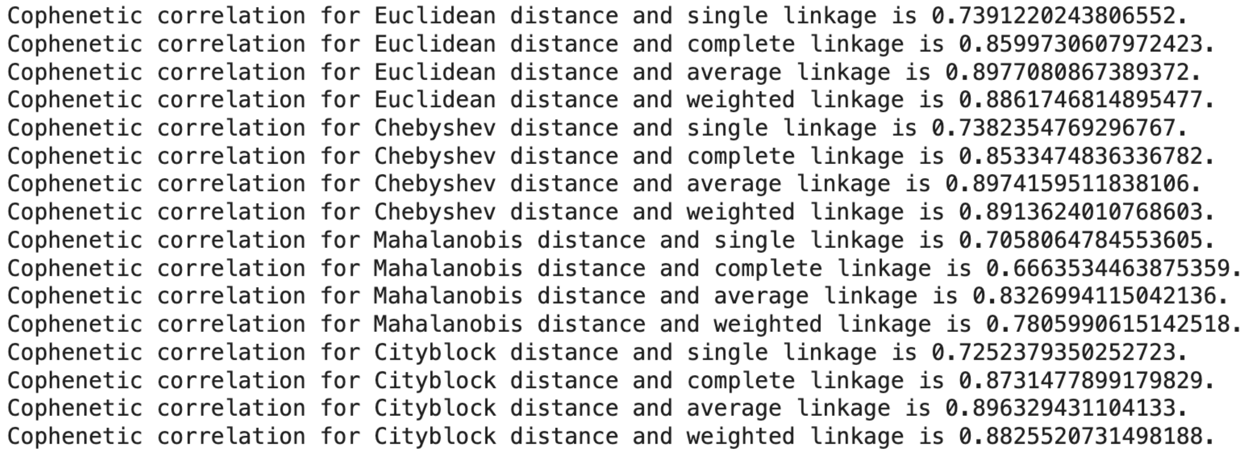
Cluster 0: These are medium credit limit customers with a moderate number of credit cards. They tend to visit the bank more than using online services or calling.

Cluster 1: These customers have low credit limits and fewer credit cards. They prefer online channels and make frequent calls to customer service, possibly due to service or support needs.

Cluster 2: High-value customers with large credit limits and many cards. They predominantly use online channels and rarely call or visit the bank, likely indicating high digital engagement.

Hierarchical Clustering

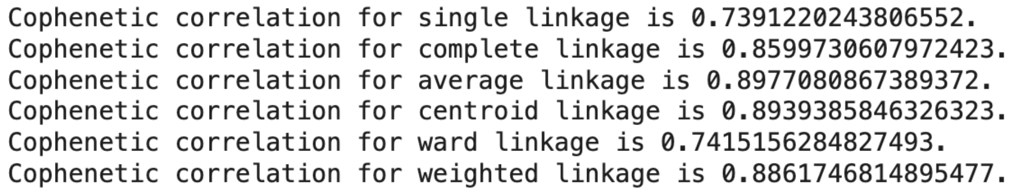
The cophenetic correlations for each linkage method with different distance metrics are as follows:



```
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 0.7058064784553605.
Cophenetic correlation for Mahalanobis distance and complete linkage is 0.6663534463875359.
Cophenetic correlation for Mahalanobis distance and average linkage is 0.8326994115042136.
Cophenetic correlation for Mahalanobis distance and weighted linkage is 0.7805990615142518.
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.
```

- **We can Observe that the cophenetic correlation is maximum with Euclidean distance and average linkage.**

The cophenetic correlations for each linkage method with Euclidean distance are as follows:



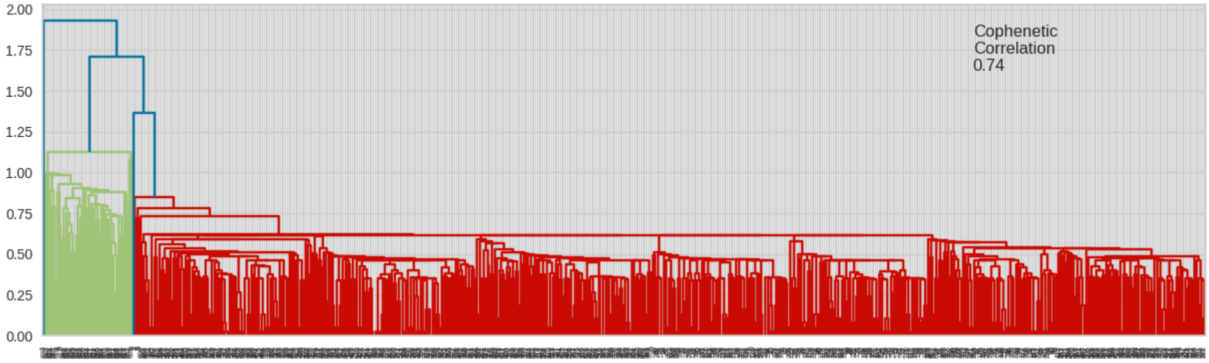
```
Cophenetic correlation for single linkage is 0.7391220243806552.
Cophenetic correlation for complete linkage is 0.8599730607972423.
Cophenetic correlation for average linkage is 0.8977080867389372.
Cophenetic correlation for centroid linkage is 0.8939385846326323.
Cophenetic correlation for ward linkage is 0.7415156284827493.
Cophenetic correlation for weighted linkage is 0.8861746814895477.
```

- The **average linkage** method has the highest cophenetic correlation (0.898), suggesting that it preserves the hierarchical clustering structure most effectively for this dataset.

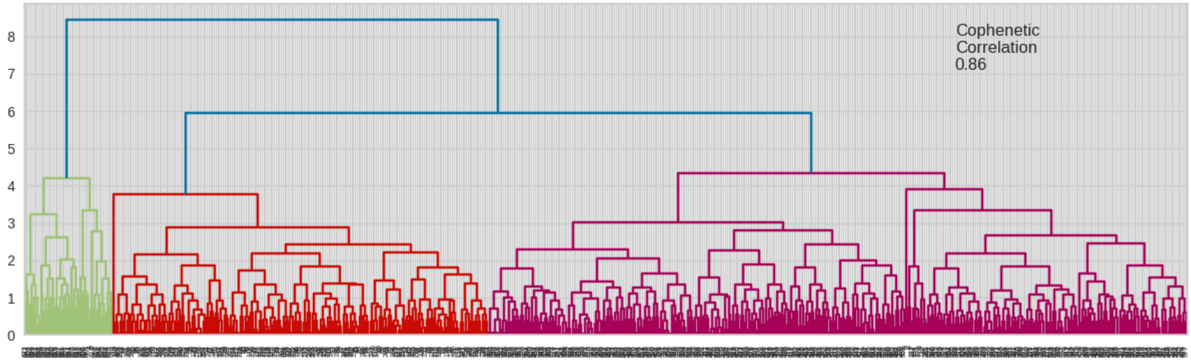
Next, Let’s determine the optimal number of clusters based on the dendrograms and perform cluster profiling.

Dendrograms

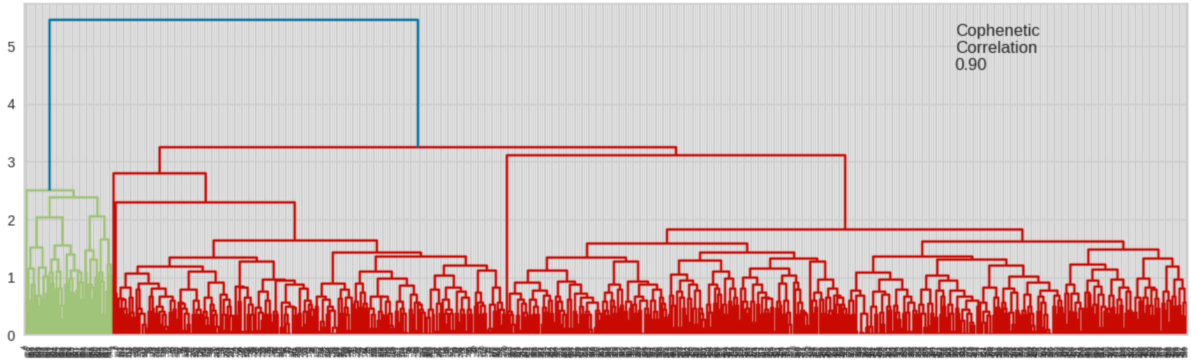
Dendrogram (Single Linkage)



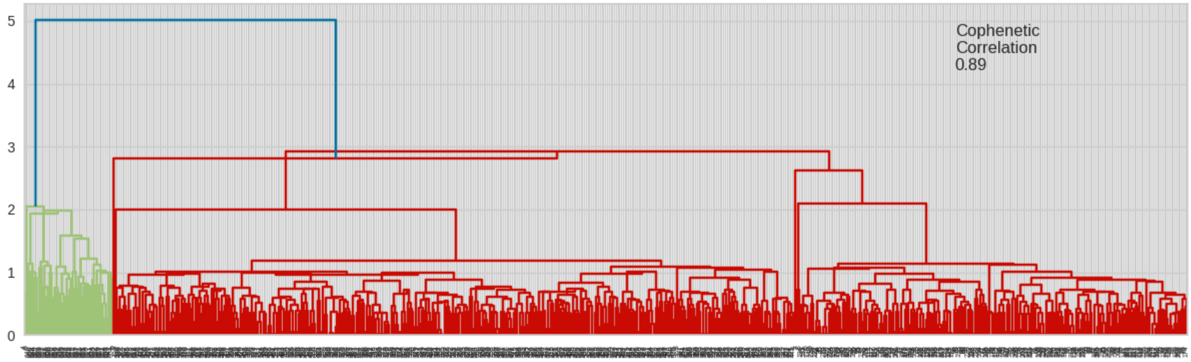
Dendrogram (Complete Linkage)



Dendrogram (Average Linkage)



Dendrogram (Centroid Linkage)



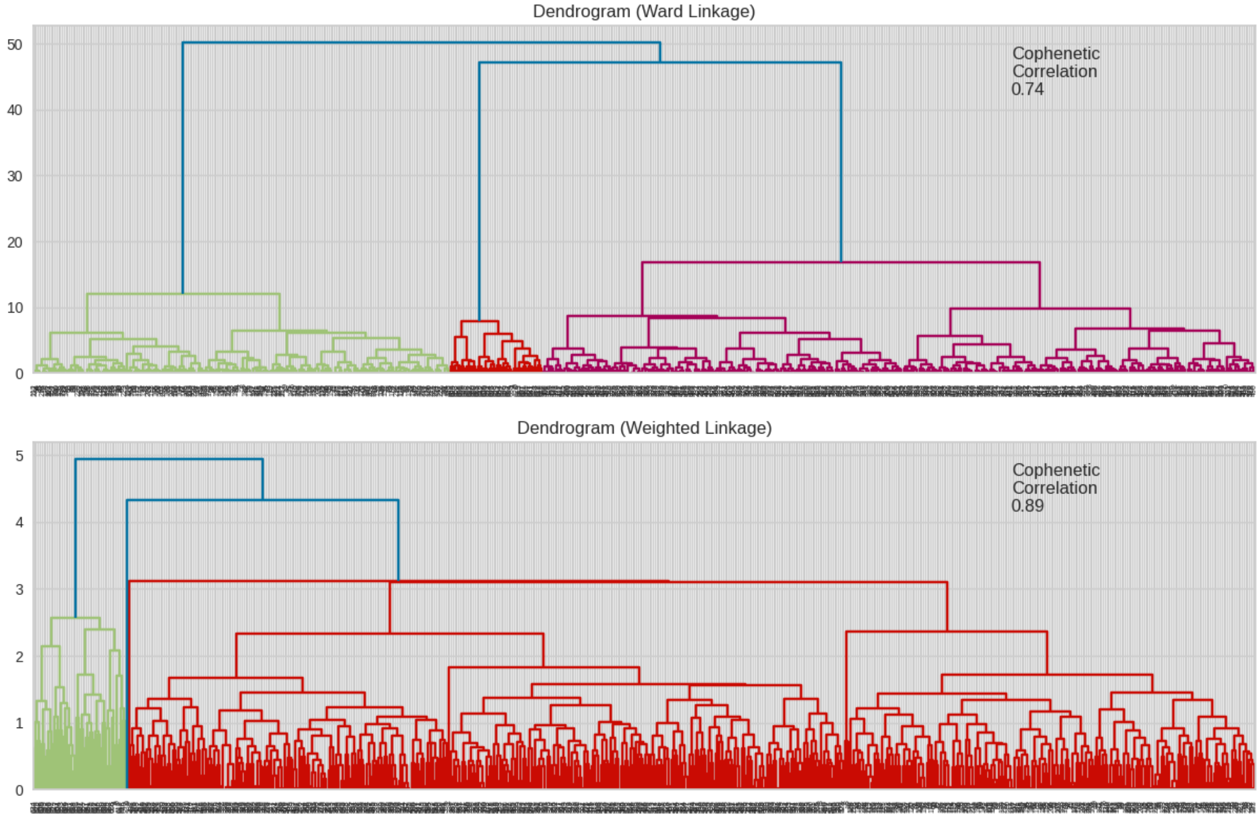
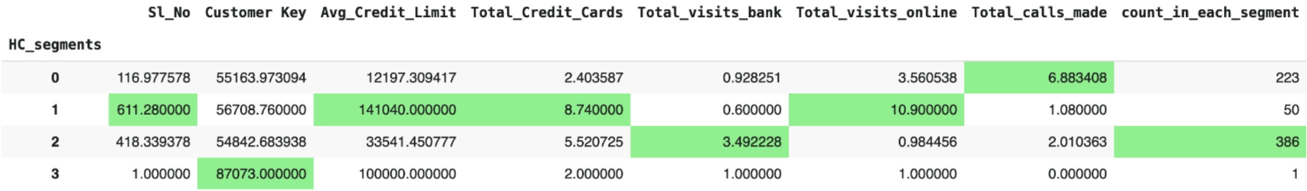


Fig 12: Dendrograms for different types of linkage

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

Cluster Profiling



	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
HC_segments								
0	116.977578	55163.973094	12197.309417	2.403587	0.928251	3.560538	6.883408	223
1	611.280000	56708.760000	141040.000000	8.740000	0.600000	10.900000	1.080000	50
2	418.339378	54842.683938	33541.450777	5.520725	3.492228	0.984456	2.010363	386
3	1.000000	87073.000000	100000.000000	2.000000	1.000000	1.000000	0.000000	1

Fig 13: Average of Numerical attributes based on different Hierarchical cluster

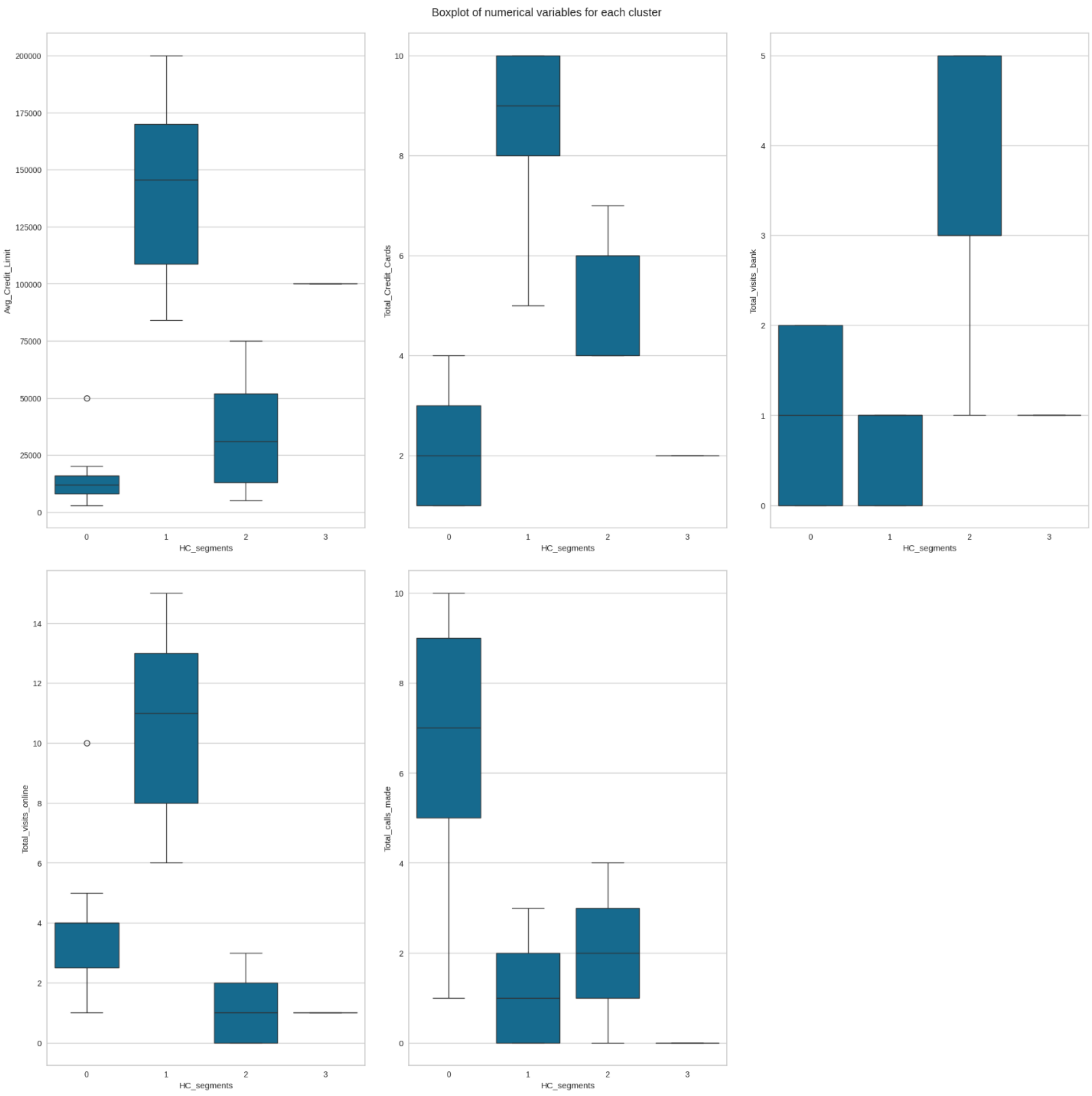


Fig 14: Boxplot of Numerical attributes vs Hierarchical cluster

Insights

- **Cluster 0**: Low-credit-limit customers with fewer cards, moderate online activity, and higher call frequency, suggesting a need for more support.
- **Cluster 1:** High-spending, multi-card customers with high credit limits and infrequent interactions (infrequent bank visits and calls but more online interactions).

- **Cluster 2**: Mid-range customers with moderate credit limits, high in-person bank visits, and limited online interactions.
- **Cluster 3**: Customers with high credit limits but limited interactions across all channels.

K-means vs Hierarchical Clustering

Metrics	K-means Clustering:	Hierarchical Clustering:
Optimal number of clusters	3	4
Silhouette score/ Cophenetic correlation	Silhouette score: Relatively good, with a peak around 3 clusters.	Cophenetic correlation: Shows that the Ward linkage with Euclidean distance produced the most suitable clusters.
Insights:	It grouped customers based on their average credit limit, credit card usage, and bank/online/call interaction patterns.	It offers a hierarchical structure of customer segments, highlighting different levels of relationships

- K-means execution time is 0.0103 seconds and Hierarchical clustering execution time is 0.0137 seconds which means K-means clustering took less time for execution.

- While both K-means and Hierarchical clustering techniques provided meaningful insights, the Hierarchical clustering, specifically with Ward linkage, appears to generate more distinct and interpretable clusters.

- This can be seen by visualizing the dendrogram and profiling the clusters. The Ward linkage method effectively identifies the subgroups within the data, leading to clearer cluster separation compared to K-means.

- Therefore, the Hierarchical clustering using Ward linkage method provides more distinct clusters in this specific case.

Business Recommendations & Insights

- **Targeted Marketing Campaigns**:
    - **High Credit Limit & Multiple Cards Segment**: Offer exclusive benefits, higher credit limits, or premium card options to this high-value segment.
    - **Frequent Online and Call Users**: Promote digital financial products or app-based services to customers with high online and call interaction frequencies, enhancing convenience and engagement.

- **Service Delivery Enhancements**:
    - **Digital-first Strategy**: Given the low frequency of bank visits, strengthen the online portal and mobile app to provide comprehensive support. Include features such as live chat or AI-driven help to reduce call volumes.
    - **Improved Call Center Support**: For frequent callers, consider expanding call center capabilities to reduce wait times and improve satisfaction. Offering callbacks or priority queues for frequent callers could enhance the customer experience.

K-Means Clusters insights -

- **Marketing Strategy**: Cluster 2 represents high-value customers who prefer online services, making them prime candidates for digital marketing campaigns and exclusive online services. Clusters 0 and 1 may benefit from targeted campaigns that promote in-person or personalized support options.
- **Service Improvements**: Cluster 1 has a high frequency of calls, indicating possible dissatisfaction or high support needs. Enhancing support channels or proactively addressing common issues for these customers could improve satisfaction and reduce call volumes.
- **Upsell Opportunities**: Cluster 0, with moderate credit limits and credit card holdings, could be targeted for cross-selling additional credit products. Their higher bank visit frequency suggests potential interest in personalized in-branch or high-touch services.

Hierarchical Clusters insights -

- **Cluster 0** may need enhanced support services through digital channels and call centers.
- **Cluster 1** could benefit from exclusive loyalty programs and high-value incentives due to their high spending and online presence.

- **Cluster 2** might respond well to incentives for in-person engagement, perhaps via personalized consultations or special in-branch services.
- **Cluster 3** could benefit from incentives to increase engagement, such as online self-service programme and targeted product recommendations.

Conclusion

The analysis identifies clear segments within the customer base, primarily differing in credit usage and preferred interaction channels. By focusing on digital engagement and tailored service enhancements, AllLife Bank can effectively improve customer satisfaction and loyalty, thereby achieving its goals for deeper market penetration and a better service model.

THE END