Unsupervised Learning
Business report

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

Contents

Objective

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

Data Dictionary:

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)

Rubric:

Exploratory Data Analysis

- Problem definition - Univariate analysis - Bivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables

Data preprocessing

- Missing Value Treatment (with rationale if needed) - Outlier Detection and Treatment (with rationale if needed) - Feature Engineering (with rationale if needed) - Data Scaling (with rationale if needed).

K-means Clustering

- Apply K-means Clustering - Plot the Elbow curve - Check Silhouette Scores - Figure out the appropriate number of clusters - Cluster Profiling.

Hierarchical Clustering

- Apply Hierarchical clustering with different linkage methods - Plot dendrograms for each linkage method - Check cophenetic correlation for each linkage method - Figure out the appropriate number of clusters - Cluster Profiling.

K-means vs Hierarchical Clustering

Compare clusters obtained from K-means and Hierarchical clustering techniques.

Actionable Insights & Recommendations

- Write down insights from the analysis conducted - Provide actionable business recommendations

Business Report Quality

- Adhere to the business report checklist

Exploratory Data Analysis:

– The Name of the dataset is “Credit Card Customer Data.xlsx”.

– The Dataset contain a total of 660 rows and 7 columns.

– There are no Duplicates present in the Data.

– There are no missing values present in the dataset.

– There some amount of outliers present in the dataset. In no_of_employees, yr_of_estab, prevailing_wage. But those outliers seem to be legit and we can leave it as it is.

– There are 4 Categorical Columns ‘Total_Credit_Cards’, ‘Total_visits_bank’, ‘Total_visits_online’, ‘Total_calls_made’. These are actually numerical values but we can also consider them as categorical values.

– And there are 7 Numerical column: Customer Key, Avg_Credit_Limit, Total_Credit_Cards, Total_visits_bank, Total_visits_online, Total_calls_made

– The Datatypes present are :

- Int64.
- Float64.

Univariate Analysis:

Numerical Variables:

1. **Avg_Credit_Limit:**

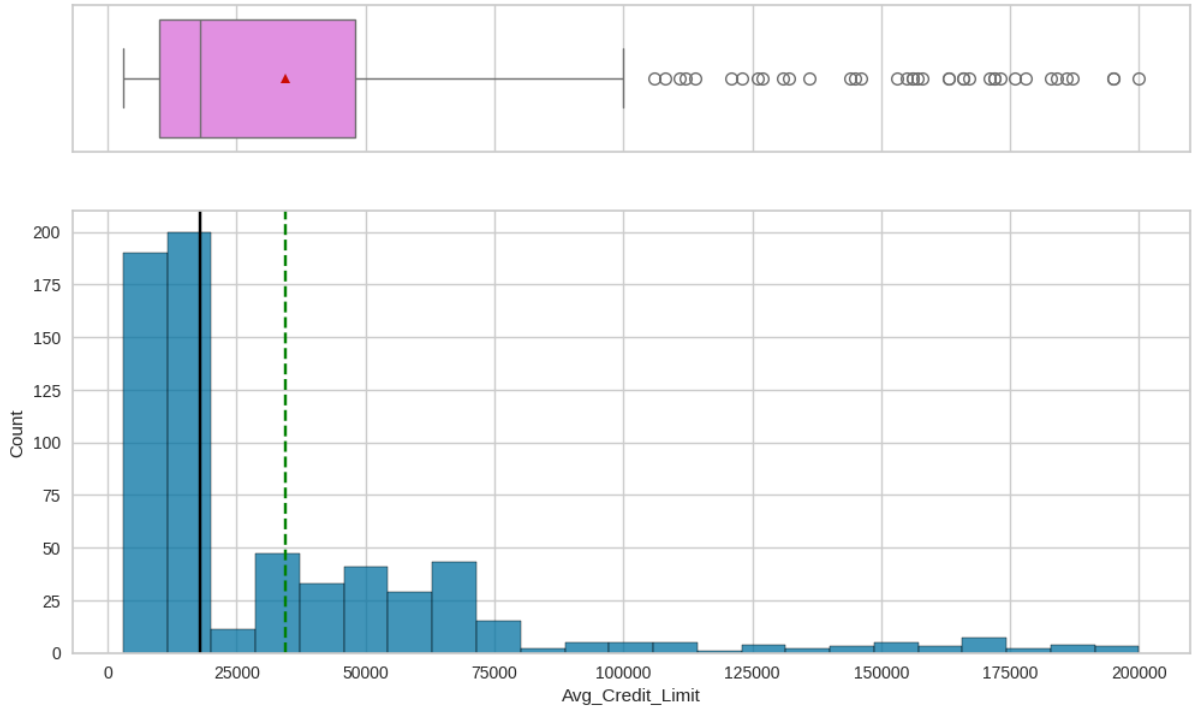


Figure 1: no_of_employees

- From the above plot we an clearly see that the Plot for `Avg_Credit_Limit` is right skewed.

- The mean value lies around 35000 and the median is 15000.

- Most of the values lie between 10000 and 45000.

- We can clearly see that there are many outliers present in the dataset.

- Most of the outliers seems to be legit and we can leave it as it is.

2. **Total_Credit_Cards:**

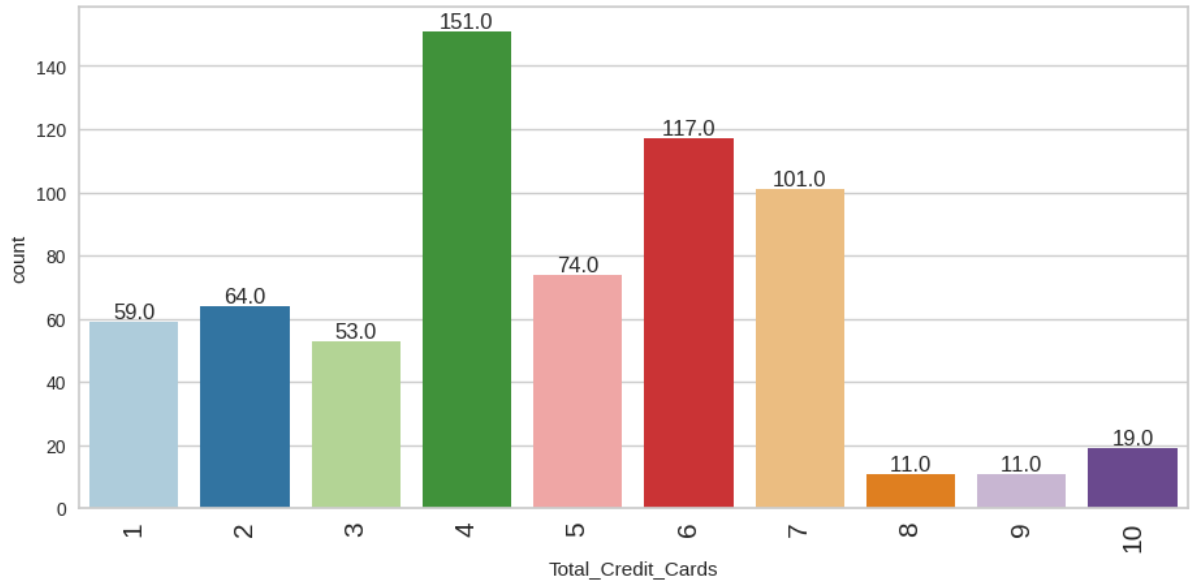


Figure 2:Total_Credit_Cards

- From the above graph we can conclude that 151 peoples have 4 credit cards.

- Next to it we have people with 6 credit cards with a count of 117.

- There are 101 People having 7 credit cards which is next.

- There are 74 peoples who have 5 credit cards.

- People having 2 credit cards are 64, 1 credit card are 59, 3 credit cards are 53 in count.

3. Total_visits_bank:

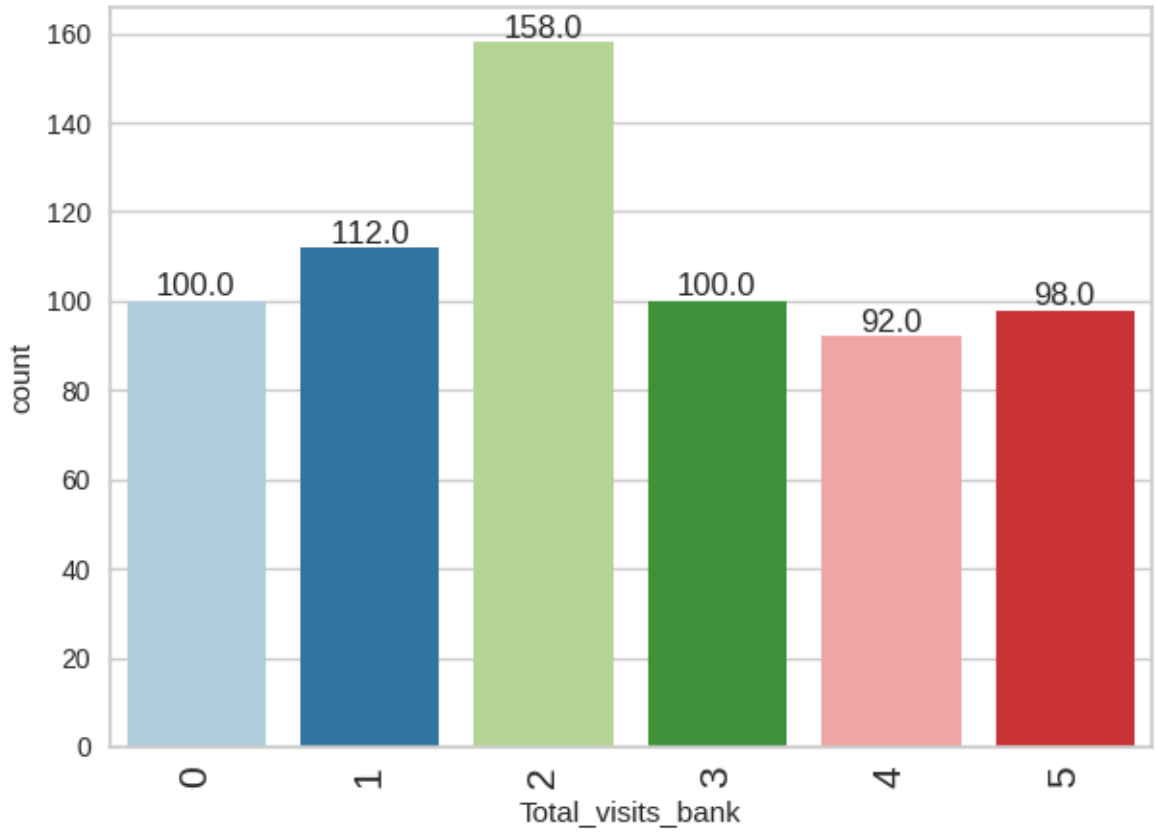


Figure 3: Total_visits_bank

- We have 158 people who totally visited the bank only 2 times.
- There are 112 people wo visited the bank only 1 time.
- People who never visited the bank and and people who visited the bank 3 times are similar in count of 100.
- Next to these values have `Total_visit_bank` are 4 with a count of 92 and 5 with a count of 98.

4. Total_visits_online:

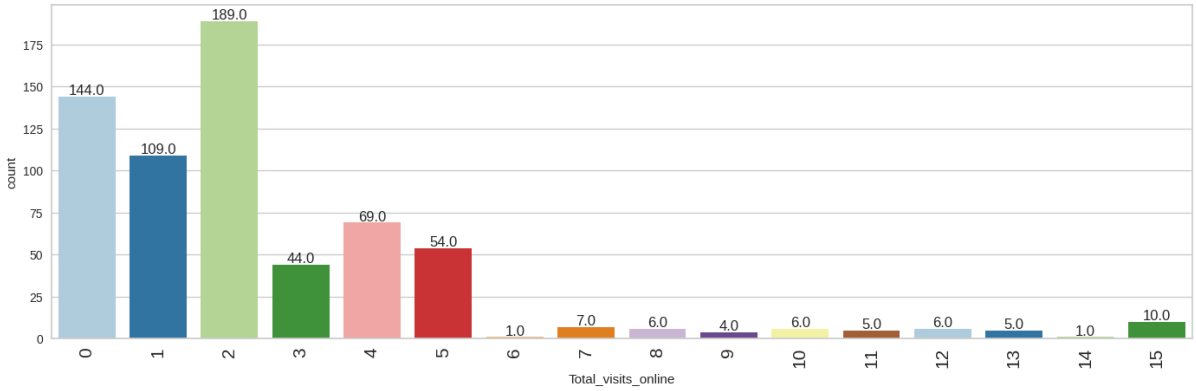


Figure 4: Total_visits_online.

- A total of 189 people visit bank online only 2 times a year.
- There are 144 people who never visit bank online.
- Next we have customers who visit banks online only once a year.
- Number of customers who visit banks online 4 times are 69, who visit 5 times are 54, who visit banks online 3 times are 44.
- Remaining values are very less compared to these values, so we can conclude that people visiting the bank online is not that common.

5. Total_calls_made:

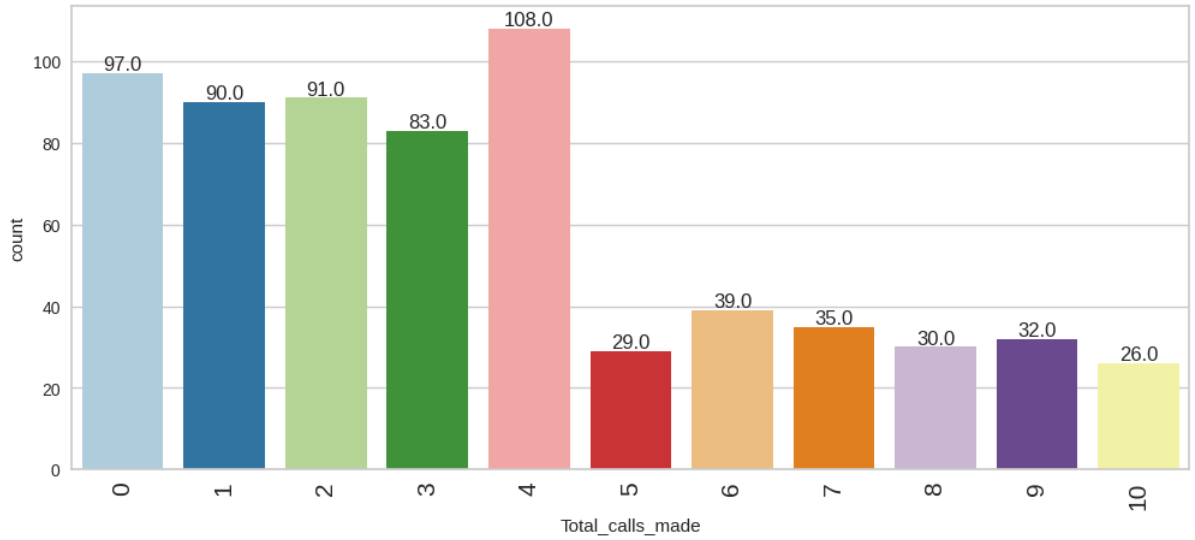


Figure 5: Total_calls_made

- From the above plot we can conclude that there are 108 customers who contact their bank 4 times a year.

- There are 97 customers who never contacted their bank in a year.

- There are 91 customers who contact their bank 2 times a year.

- There are 90 customers who contact their bank once a year.

- There are 83 customers who contact their bank 3 times a year.

- The remaining counts are 29 customers contact 5 times, 39 customers contact 6 times, 35 customers contact 7 times, 30 customers contact 8 times, 32 customers contact 9 times, 26 customers contact 10 times.

Bivariate Analysis:

Heatmap:

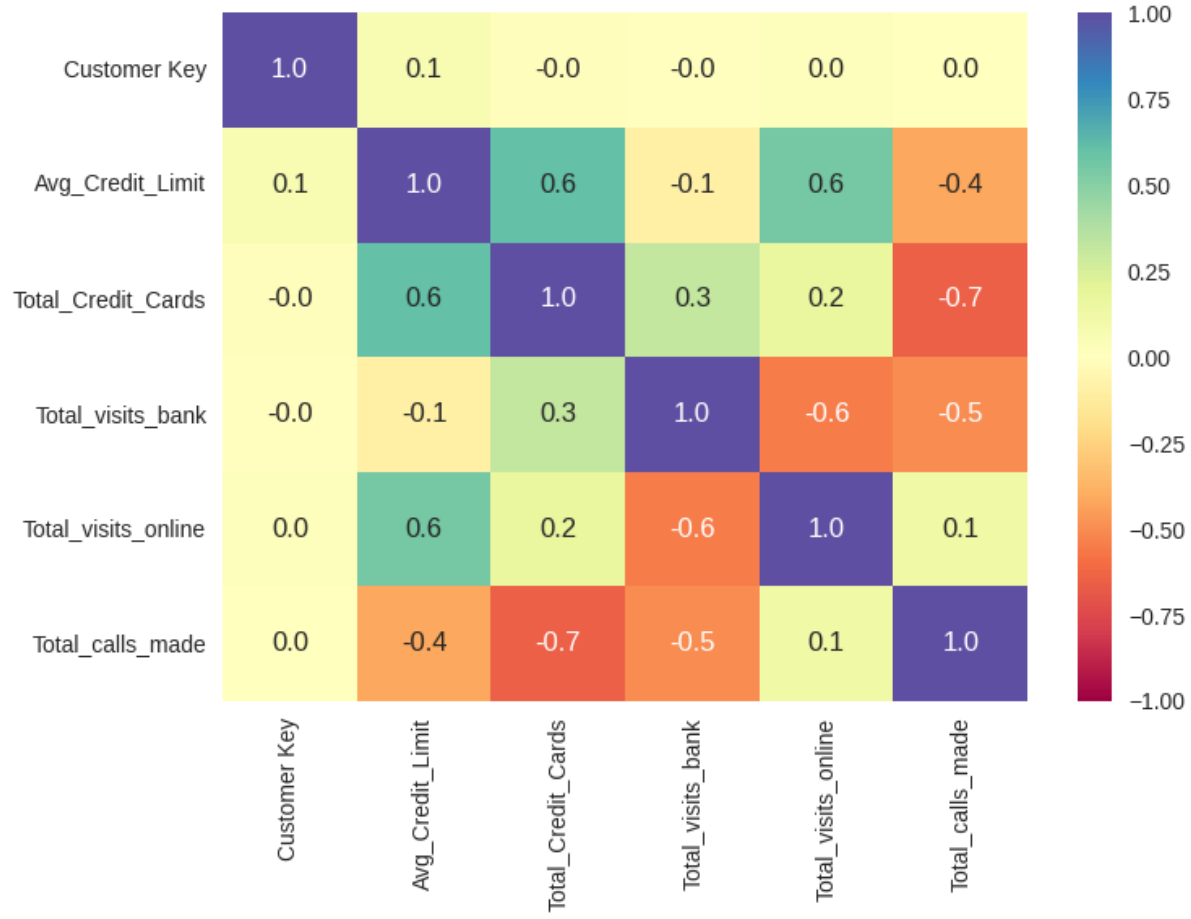
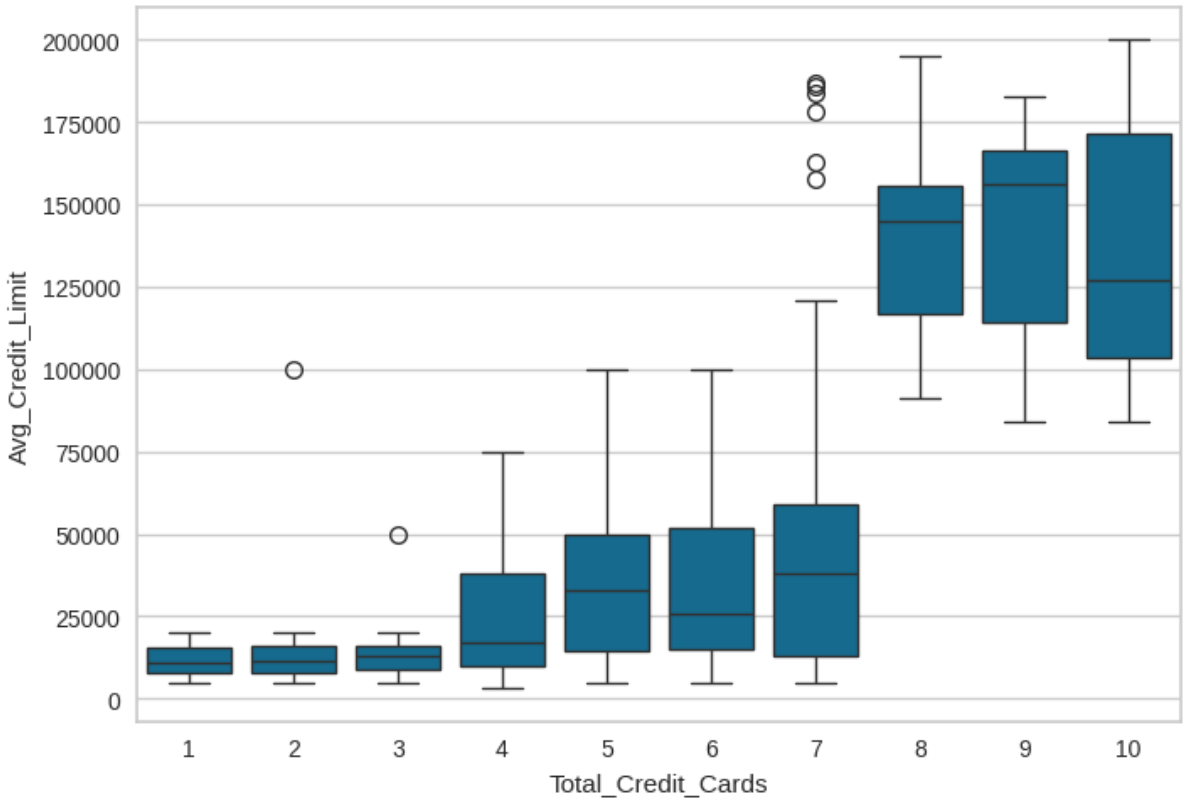


Figure 6: Heatmap

- From the above heatmap we can clearly see that the `Customer Key` column doesn't have any relationship with any of the column, since it is unique value.
- We can see high relationship between `Avg_Credit_Limit` have `Total_Credit_Cards`.
- We can see a high positive relationship between `Avg_Credit_Limit` and `Total_visits_online`.
- There is a high negative relationships between `Total_call_made` and `Total_credit_cards`
- We can see that the `Avg_Credit_Limit` and `Total_calls_made` also have a negativ correlation.
- Column `Total_calls_made` and `Total_visits_bank` also have a good negative correlation.

Avg_Credit_Limit VS Total_Credit_Cards:

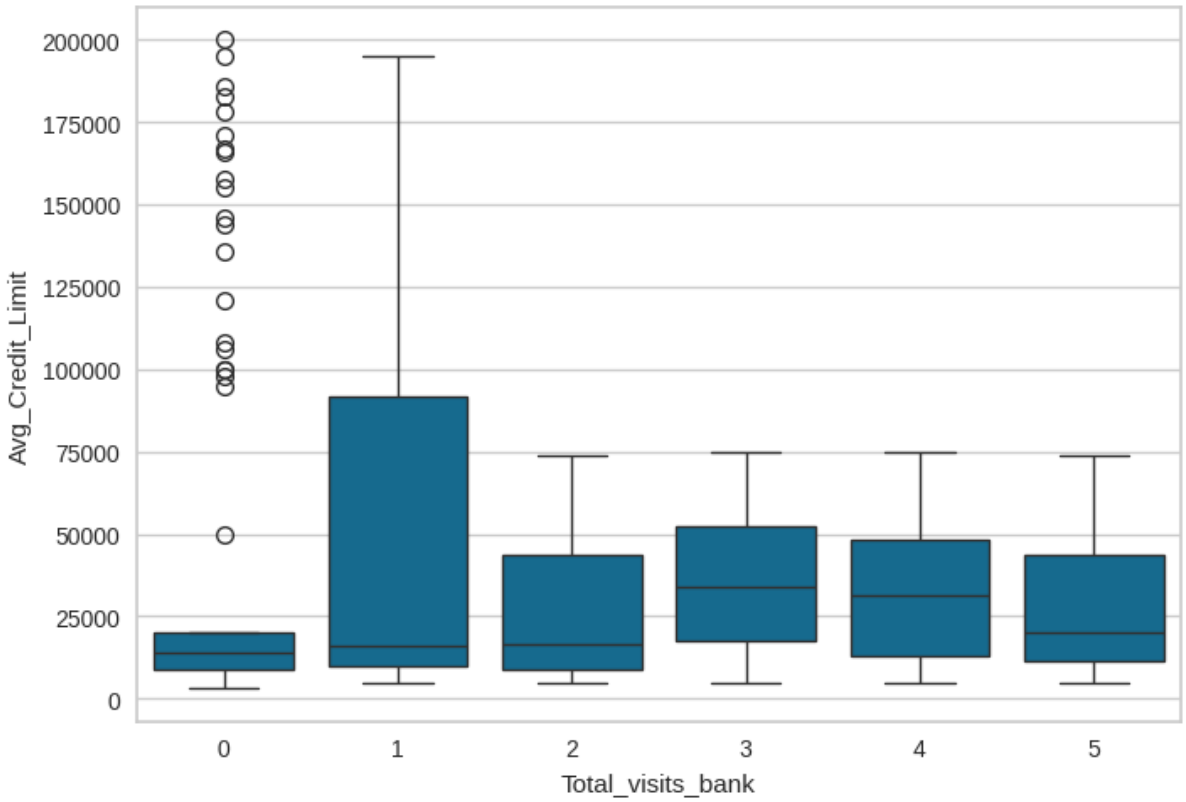


- The are some outliers values in some `Total_Credit_cards` values.

- The distribution is mostly rigt skwed for all plots.

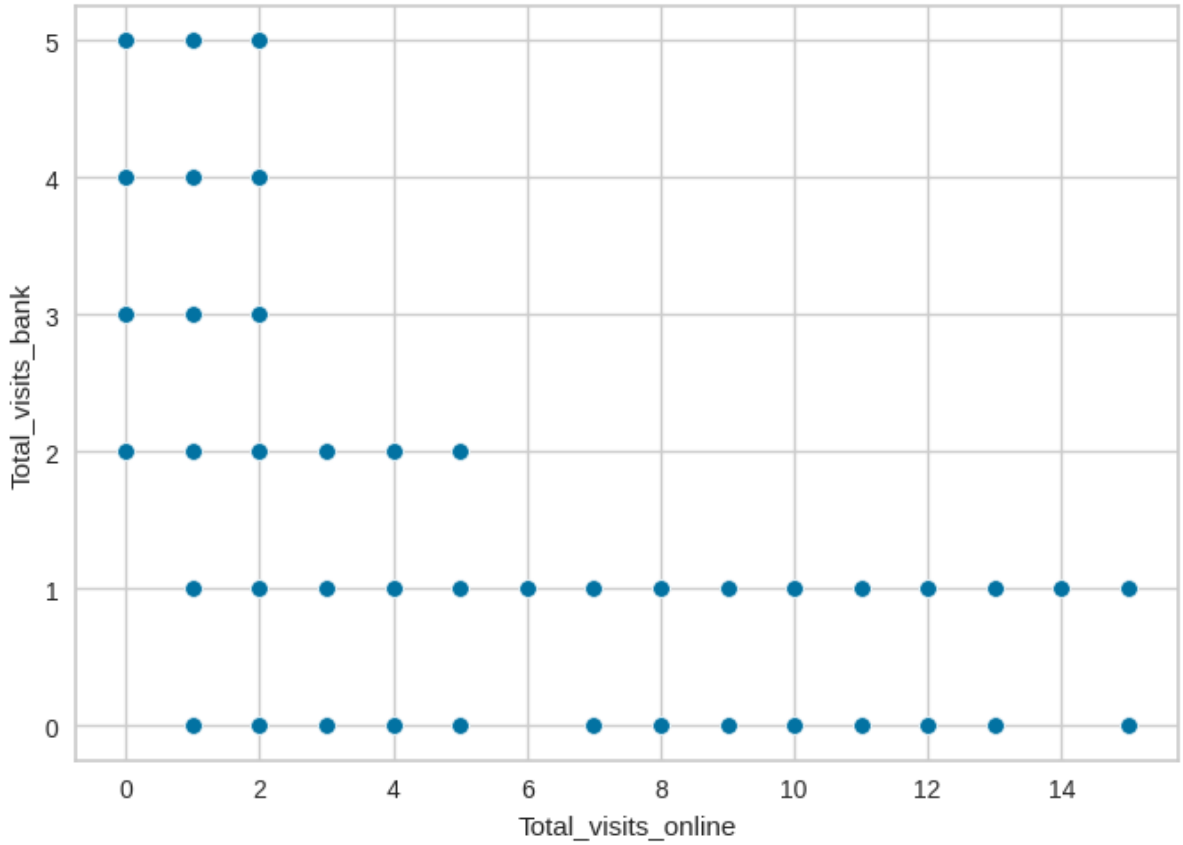
- Some values are left skewed for values 8, 9 and 10.

Total_visits_bank VS Total_visits_bank:



- All the values are Right skewed here.
- There are some Outliers for 0 `Total_visits_bank` with `Avg_Credit_Limit`

Total_visits_bank VS Total_visits_online:



- This plot clearly defines that `Total_visits_bank` and `Total_visits_bank` is negatively correlated.
- If number of people visit the bank is more the number of people visiting online is less.
- They are inversely proportional.

Data Preprocessing:

Outlier Treatment:

- We have some outlier present in `Avg_Credit_Limit` but the values seems to okay and we don't neeed to do any modifications.
- we also have outliers in `Total_visits_online` column but the values 14 and 12 are okay since people visiting online portal is not a outlier. So we can leave it as it is.

The outliers present in the data seems to be fine, we don’t need to remove or impute those values we are leaving it as it is. We are not treating them.

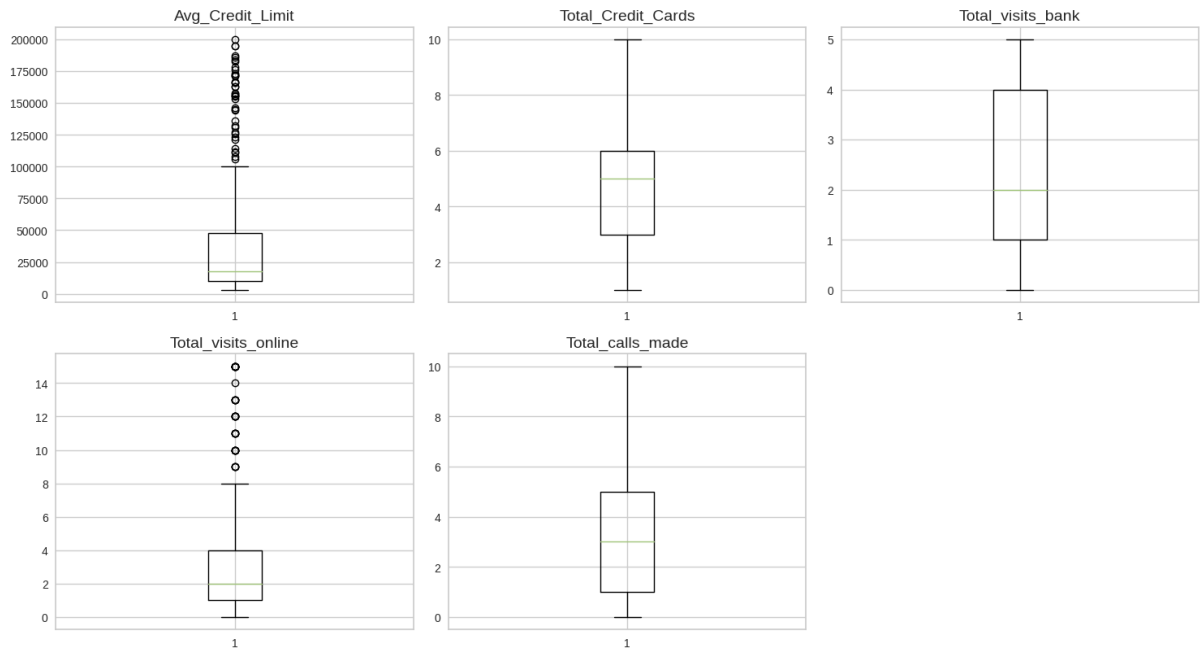


Figure 7: Outlier Treatment

- We have some outlier present in `Avg_Credit_Limit` but the values seems to okay and we don't neeed to do any modifications.
- we also have outliers in `Total_visits_online` column but the values 14 and 12 are okay since people visiting online portal is not a outlier. So we can leave it as it is.

Feature Engineering:

Dropping ‘Sl.no’ and ’Customer Key’ column:

	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	100000	2	1	1	0
1	50000	3	0	10	9
2	50000	7	1	3	4
3	30000	5	1	1	4
4	100000	6	0	12	3

- We have removed both the columns since the slno. Column is not creating any impact and the ‘Customer Key’ column is having only unique values which also won’t create any impact.
-

Missing value Treatment:

- We don’t have any missing values or null values present in the data.

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

dtype: int64

Duplicate value:

```
The number of duplicates present: 0
```

- Since, we don't see any duplicate values we are good to go without any treatment for the duplicate.

Scaling the data:

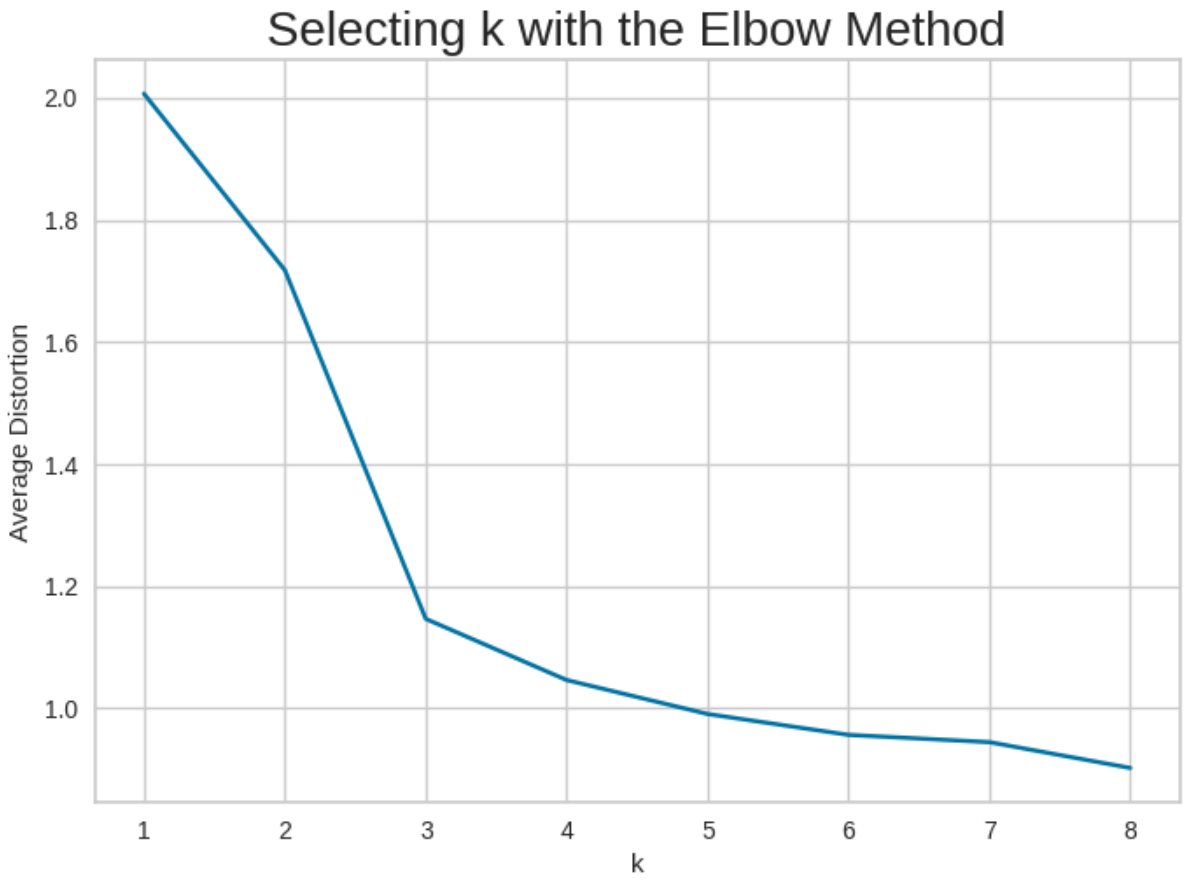
	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
0	1.740187	-1.249225	-0.860451	-0.547490	-1.251537
1	0.410293	-0.787585	-1.473731	2.520519	1.891859
2	0.410293	1.058973	-0.860451	0.134290	0.145528
3	-0.121665	0.135694	-0.860451	-0.547490	0.145528
4	1.740187	0.597334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

- The data is scaled with respect to z-score.

K Means Clustering:

- For K means we are checking the right value K using Elbow plot.

```
Number of Clusters: 1   Average Distortion: 2.006922226250361
Number of Clusters: 2   Average Distortion: 1.7178787250175898
Number of Clusters: 3   Average Distortion: 1.1466276549150365
Number of Clusters: 4   Average Distortion: 1.0463825294774463
Number of Clusters: 5   Average Distortion: 0.990778179643084
Number of Clusters: 6   Average Distortion: 0.9566047496610922
Number of Clusters: 7   Average Distortion: 0.9446794420340151
Number of Clusters: 8   Average Distortion: 0.9023869555204838
```

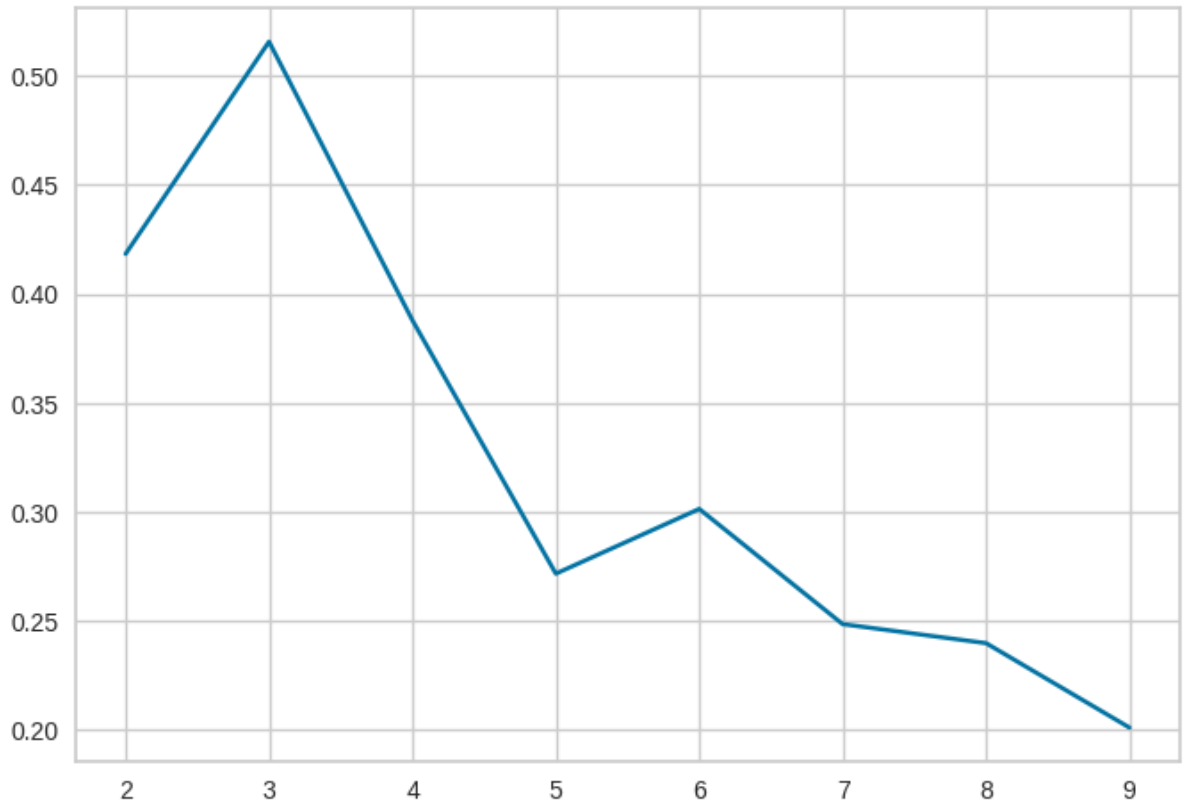


- From the above elbow plot we can conclude that the appropriate values for K can be 3 or 4.

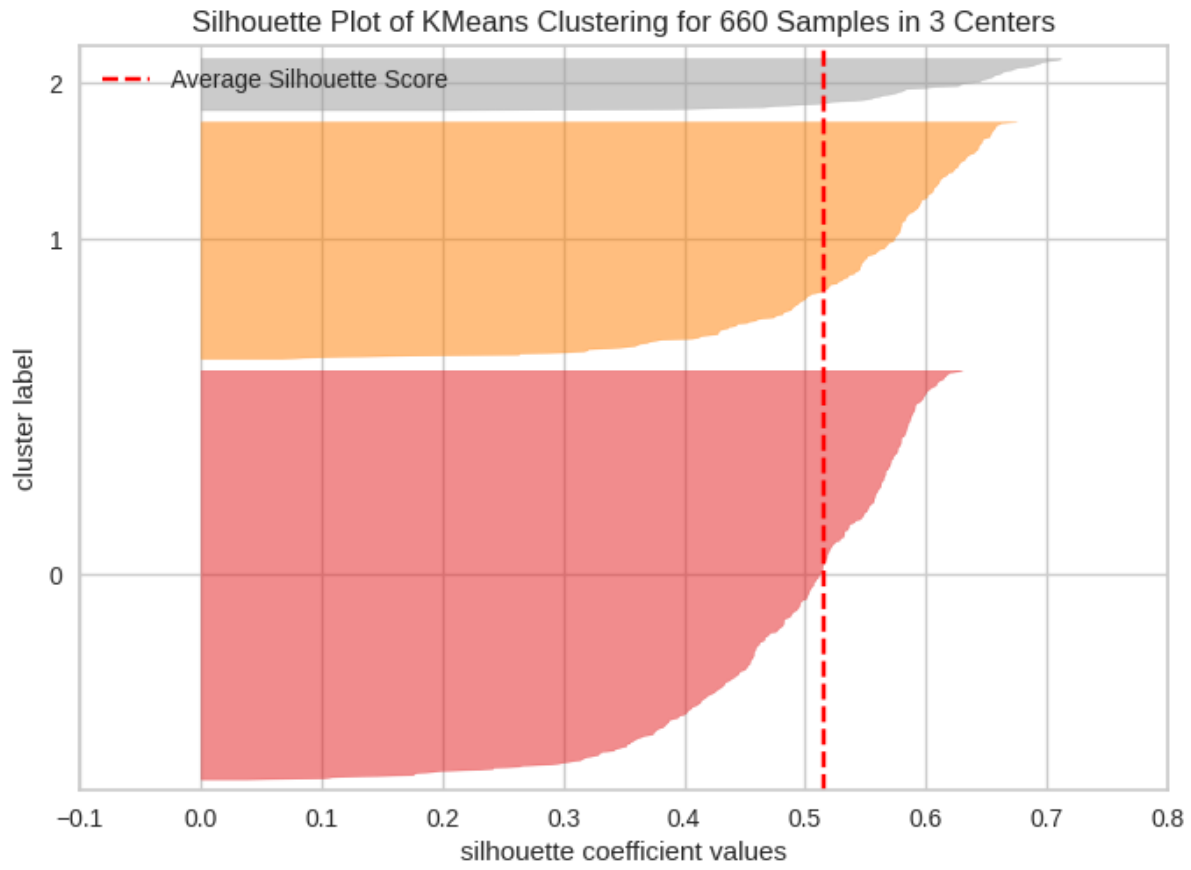
Checking for Silhoutte value:

The Silhoutte values for different clusters are,

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 0.38796493017642897
For n_clusters = 5, silhouette score is 0.2717835112303572
For n_clusters = 6, silhouette score is 0.3014652925105362
For n_clusters = 7, silhouette score is 0.2486484349998537
For n_clusters = 8, silhouette score is 0.23991891027451814
For n_clusters = 9, silhouette score is 0.20119159100339573



- We can clearly conclude that the Cluster point we can use in 3 which has the highest silhoutte's score of 0.5157182558881063.



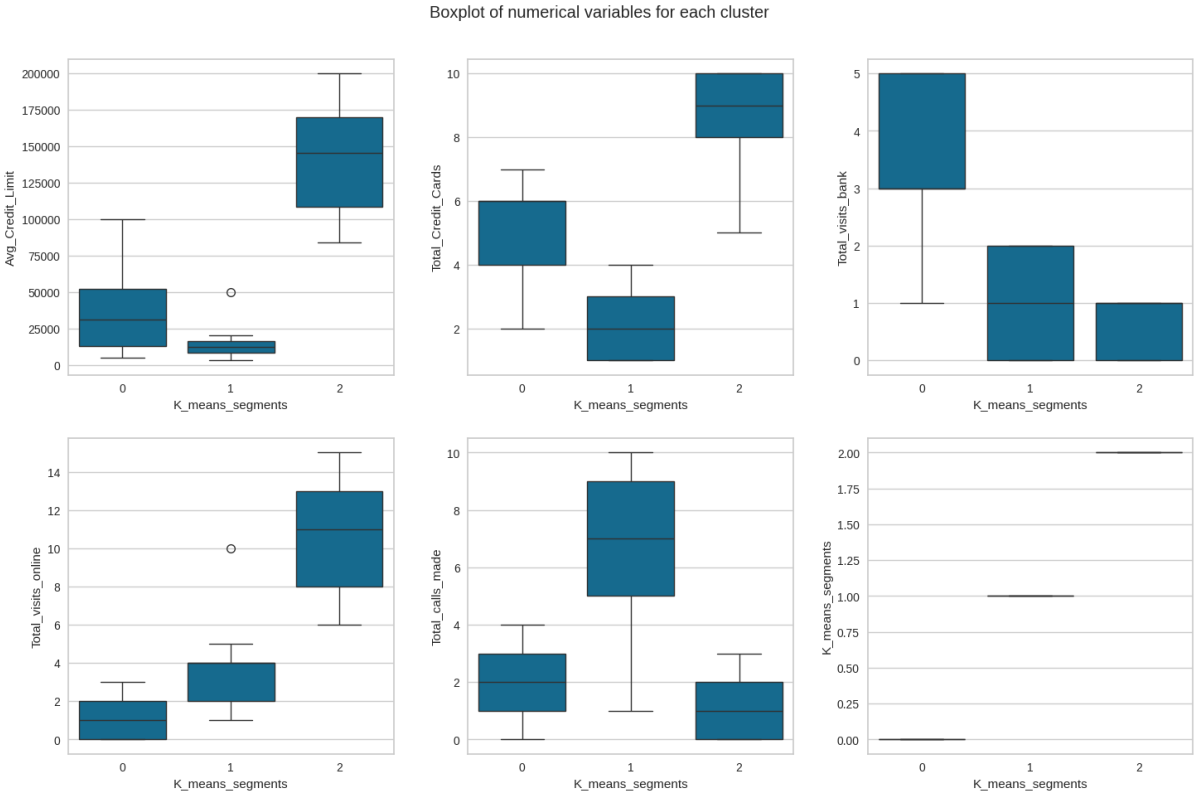
- Let's take 3 as the appropriate no. of clusters as the silhouette score is high enough and there is knick at 3 in the elbow curve.

Cluster Profiling:

K_means_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	Customer_Segement
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

- From the above cluster profile we can say that there are 3 clusters created.
- The first cluster has 386 values, second cluster has 224 values and the third cluster has 50 values.

Boxplots for checking the clusters:



- We can see that the clusters split for each feature in the dataset.
- It gives the distribution of the data in the datasets with respect to the Clusters our model has predicted.
- Cluster 2 has an obvious distinction compare to group 0 and 1.
    - Customers who have average credit limit of 75k and above are in Cluster 2.
    - Customers who have more than 7 credit cards are in Cluster 2.
    - Customers who have visited online more than 6 times are in Cluster 2.
- Cluster 1 has less credit cards compare to other groups - 1-4 credit cards.
- Cluster 1 has one attribute that is distinct from other group namely, customers who make calls more than 4 times.
- Cluster 0 visited bank more than other groups - more than 2 times and up to 5 times.
- Cluster 0 also less visited bank online compare to other Cluster.

Hierarchical Clustering:

- Checking Cophenetic Correlation for the scaled datasets

```
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.
```

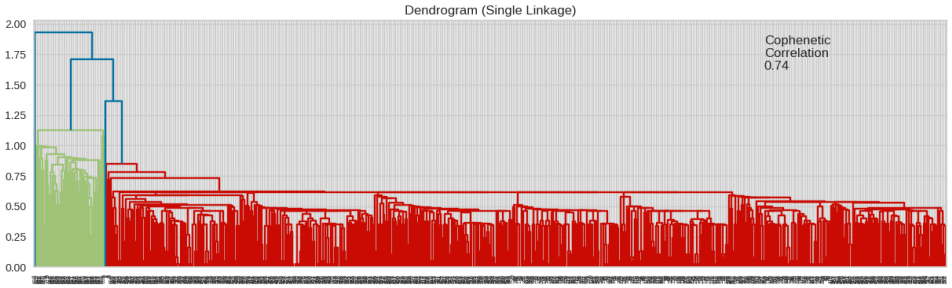
- From the above cophenetic correlation values for different distances and linkage we are selecting the highest value.
- Highest cophenetic correlation is 0.8977080867389372, which is obtained with Euclidean distance and average linkage.

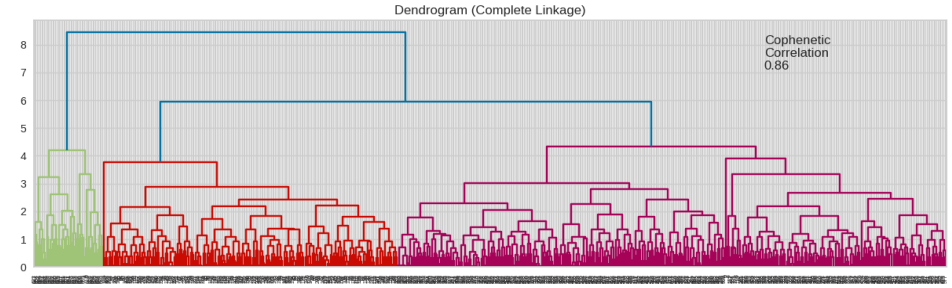
Exploring different linkage for Euclidean distance:

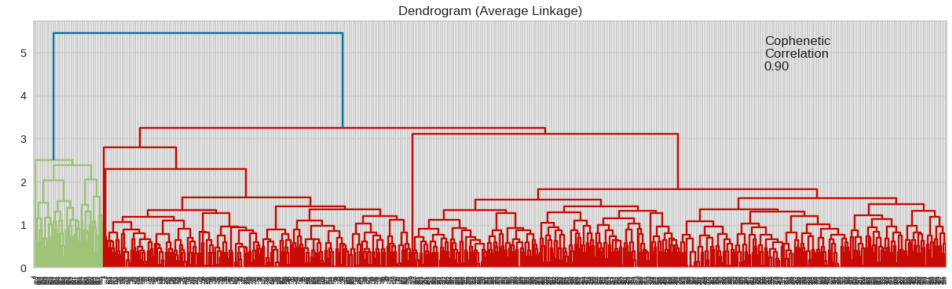
```
Cophenetic correlation for single linkage is 0.7391220243806552.
Cophenetic correlation for complete linkage is 0.8599730607972423.
Cophenetic correlation for average linkage is 0.8977080867389372.
Cophenetic correlation for centroid linkage is 0.8939385846326323.
Cophenetic correlation for ward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

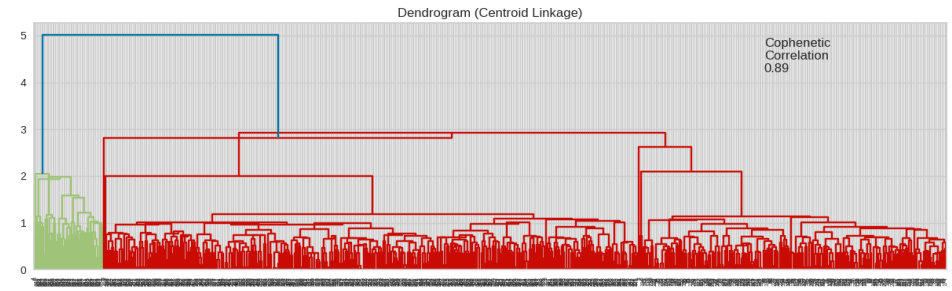
- Highest cophenetic correlation is 0.8977080867389372, which is obtained with average linkage.

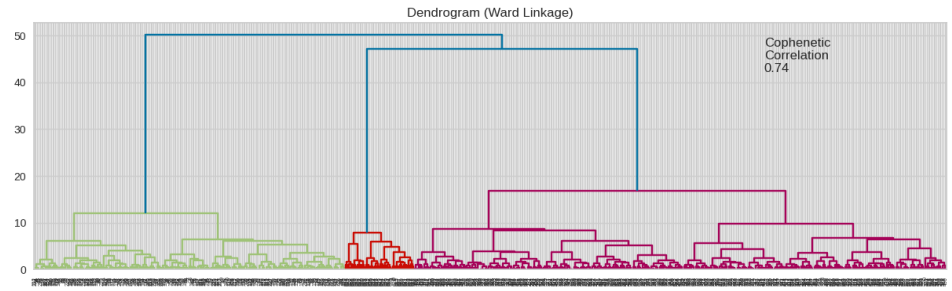
Checking Dendrograms:

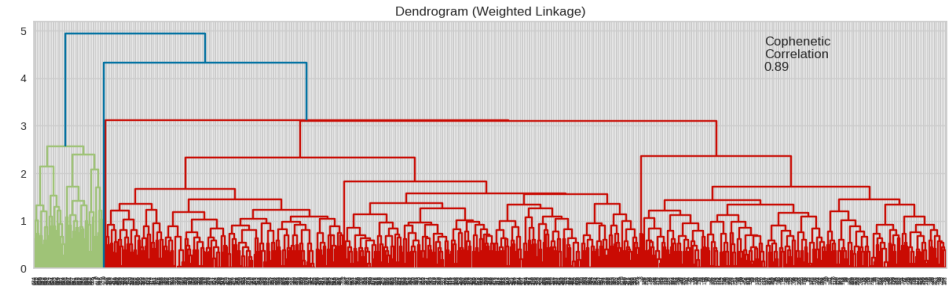












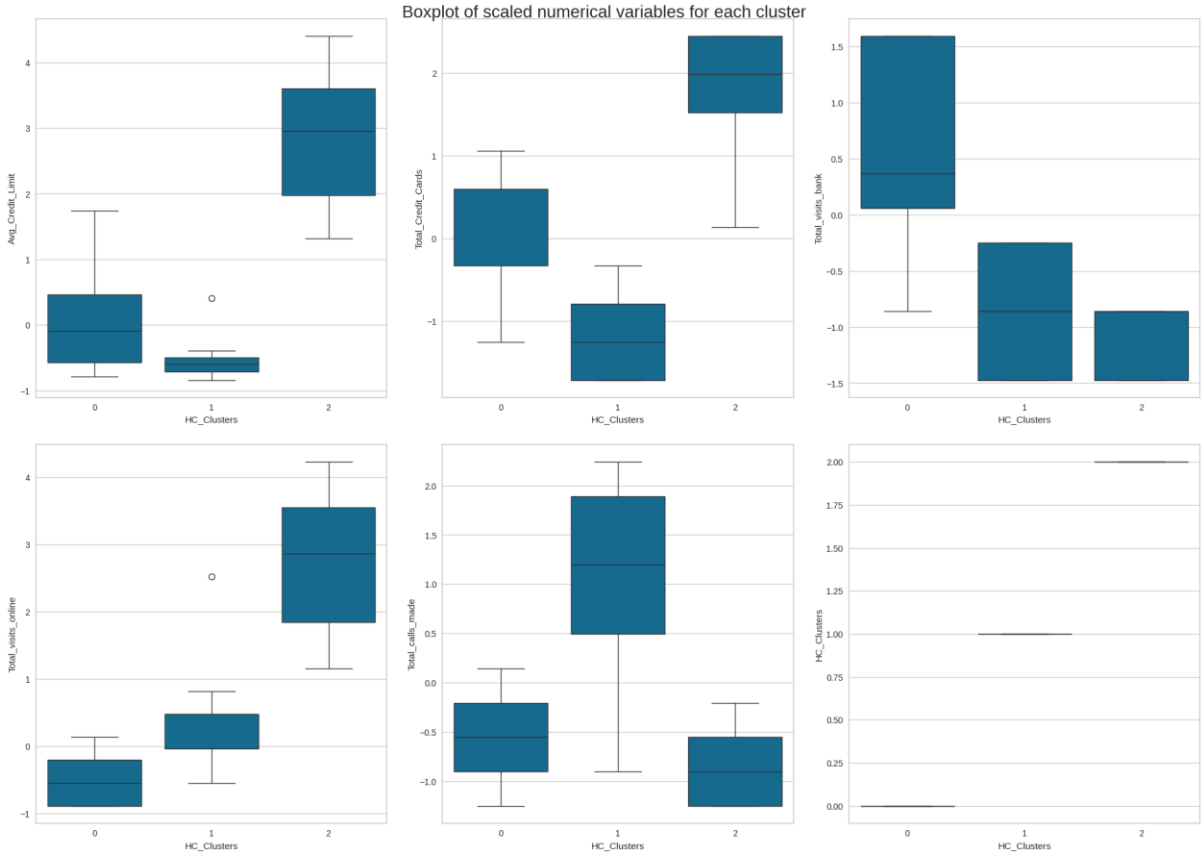
- From the above dendogram values we can take the cluster values as 3 which is suitable for our data.
- The cophenatic correlation is higher for Euclidean distance and Average linkage.
- 3 appears to be the appropriate number of clusters from the dendrogram for average linkage.

Final model value for Hierarchical clustering:

HC_Clusters	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	K_means_segments	Customer_Segement
0	33713.178295	5.511628	3.485788	0.984496	2.005168	0.002584	387
1	141040.000000	8.740000	0.600000	10.900000	1.080000	2.000000	50
2	12197.309417	2.403587	0.928251	3.560538	6.883408	1.000000	223

- We chose Distance as ‘Euclidean’ and Linkage as ‘Average’
- We have the 387 value in cluster 0, and 50 values in cluster 1 and 223 values in cluster 2.

Boxplot for Hierarchical Clustering:



- Cluster 0 has an obvious distinction compare to Cluster 1 and 2.
  - Customers who have average credit limit of 75k and above are in Cluster 0.
  - Customers who have more than 7 credit cards are in Cluster 0.
  - Customers who have visited online more than 6 times are in Cluster 0.
- Cluster 1 has less credit cards compare to other labels - 1-4 credit cards.
- Cluster 1 has one attribute that is distinct from other Labels namely, customers who make calls more than 4 times.

- Cluster 2 visited bank more than other labels - more than 2 times and up to 5 times.
- Cluster 2 also less visited bank online compare to other labels.

Comparision between K-means and Hierarchical:

From the clusters and numbers we can see that both K means and Hierarchical gives similar results. The results are as below:

K- means:

K_means_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	Customer_Segement
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

- From the above cluster profile we can say that there are 3 clusters created.
- The first cluster has 386 values, second cluster has 224 values and the third cluster has 50 values.
- K-means is more efficient type of clustering, it is more widely used.

Hierarchical Clustering:

HC_Clusters	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	K_means_segments	Customer_Segement
0	33713.178295	5.511628	3.485788	0.984496	2.005168	0.002584	387
1	12197.309417	2.403587	0.928251	3.560538	6.883408	1.000000	223
2	141040.000000	8.740000	0.600000	10.900000	1.080000	2.000000	50

- From the above cluster profile we can say that there are 3 clusters created.
- The first cluster has 387 values, second cluster has 50 values and the third cluster has 223 values.
- The distance we choose is ‘Euclidean’ and linkage we choose is ‘Average’.
- The only drawback we have in Hierarchical clustering is it is computationally very expensive and consumes more time.

Actionable Insights

Customer Segmentation:

Customers fall into distinct groups based on credit usage, visits, and online activity. Each group shows different behaviors and preferences.

Credit Limit Outliers:

Some customers have unusually high or low credit limits, indicating opportunities for personalized offers.

Engagement Patterns:

Some customers prefer online banking, while others visit branches more frequently. Tailor communication and services accordingly.

At-Risk Customers:

Low-engagement clusters (few visits or calls) might be at risk of churn.

Cross-Sell Opportunities:

Certain high-usage segments offer potential for upselling other products, such as loans or premium services.

Business Report: Credit Card Customer Analysis

Summary:

Unsupervised learning identified key customer segments based on behaviors like credit card usage, bank visits, and online activity. These insights enable targeted marketing, product development, and customer retention strategies.

Key Insights:

Customer Segmentation: Distinct customer groups emerged, providing clarity on high and low credit users, frequent bank visitors, and online users.

Outliers: Some customers with extreme credit limits present opportunities for customized service offerings.

Engagement Preferences: Clear division between customers who prefer digital vs. in-person banking.

Retention Risks: Lower engagement clusters may need targeted retention efforts.

Cross-Selling Potential: High-usage customers are primed for cross-selling financial products.

Recommendations:

Personalized Marketing: Tailor campaigns to each segment's needs (e.g., rewards for high-credit users, incentives for low-credit users).

Retention Strategy: Focus on re-engaging low-engagement customers with special offers or improved service.

Optimize Digital Channels: Invest in digital platforms for online-focused customers; enhance in-branch experiences for others.

Outlier Opportunities: Offer personalized financial services to high-credit limit customers; consider credit increases for others.

Product Customization: Create tailored credit card products based on customer usage patterns.