All Life Bank

Data Analysis and Visualization Business Report

—————————-

Customer Segmentation & Cluster Analysis

9/28/24

Contents / Agenda

Executive Summary

All Life Bank wants to focus on its credit card customer base in the next financial year.

The challenges which need addressing in order to make this campaign a success are:

1. The need to improve market penetration.
2. The customer's perception of the bank's support services

<u>660 observations within 7 categories of data were collected</u>

Sl_no - Customer Serial Number

Customer Key - Customer identification

Avg_Credit_Limit - Average credit limit

Total_Credit_Cards - Total number of credit cards

Total_visits_bank - Total bank visits

Total_visits_online - Total online visits

Total_calls_made - Total calls made

An in-depth analysis into these categories will provide a path for solutions to the 2 challenges, which AllLife faces, as part of their initiative next financial year.

The results will equip the marketing team at All Life to run personalized marketing campaigns aimed at

1. Targeting new customers
2. Up-selling to existing customers

The same analysis will also equip the Operations team in their campaign to upgrade the service delivery model, ensuring that customers' queries are resolved faster.

I took a look at the data from a Univariate perspective to get an understanding of any correlation in behaviors on a linear perspective.

I then took a look at the data using a Bivariate approach to see how any one category could have an impact on another category.

Since the objective at hand is focused on segmenting customers, I used several clustering methods to get a visual correlation on the data and to narrow down segmentation groups.

Distinct categories with

The greatest variance are

1. Number of customers who go online
2. Credit limit among customers

The least variance are

1. Bank Visits
2. Total Credit Cards

To close the gap between these two groups and also accomplish the goal of the Marketing and Operations team, as well as progress forward with AllLife's initiative for the next financial year, I suggest the following strategies:

- Problem - The customers have a poor perception of the bank's support services.
- Goal - To upgrade the service delivery model, ensuring that customers' queries are resolved faster.
- Data based solution - Develop a guaranteed 1 business day turn around for support cases submitted to the bank via the online application.

- Problem - The need to improve market penetration.
- Goal - Target new customers.
- Data based solution - Market a Secure Credit Card offering to customers with a lower credit limit. Run a secondary campaign for a bonus cash savings card for every friend they refer into AllLife.

The data from this report is significant enough to foster several campaigns which can be beneficial to AllLife and support them in their goals. The two mentioned above are the most all encompassing in consideration of the overall objective of all the departments as a whole.

Business Problem Overview and Solution Approach

AllLife Bank is needing to address their need to improve market penetration and the customer's perception of the bank's support services.

The Marketing team at AllLife want to run personalized marketing campaigns aimed at targeting new customers and up-selling to existing customers.

The Operations team wants to upgrade their service delivery model, ensuring that customers' queries are resolved faster.

This analysis will take a look into key categories of existing bank customers and their related behaviors. From this analysis, we will see a trend emerge that will provide a path for solutions to the challenges AllLife faces, as part of their initiative next financial year.

I took a look at the data from a Univariate perspective to get an understanding of any correlation in behaviors on a linear perspective.

I then took a look at the data using a Bivariate approach to see how any one category could have an impact on another category.

Since the objective at hand is focused on segmenting customers, I used several clustering methods to get a visual correlation on the data and to narrow down segmentation group

Data Overview

#	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:(7) int64

Originally, there were 660 observations within 7 columns in the dataset.

<u>The 7 columns were:</u>

- Sl_no - Customer Serial Number
- Customer Key - Customer identification
- Avg_Credit_Limit - Average credit limit
- Total_Credit_Cards - Total number of credit cards
- Total_visits_bank - Total bank visits

- Total_visits_online - Total online visits
- Total_calls_made - Total calls made

All the columns had 660 values

- There were no missing values in the data

All the columns consisted of numbers.

index	Sl No	Customer Key	Avg Credit Limit	Total Credit Cards	Total visits to bank	Total visits online	Total calls made
0	1	87073	$100,000.00	2	1	1	0
1	2	38414	$50,000.00	3	0	10	9
2	3	17341	$50,000.00	7	1	3	4
3	4	40496	$30,000.00	5	1	1	4
4	5	47437	$100,000.00	6	0	12	3

After preprocessing and cleaning the data, I was left with 644 unique observations across 5 columns.

Index	Avg Credit Limit	Total Credit Cards	Total visits to bank	Total visits online	Total calls made
162	$8000.00	2	0	3	4
175	$6000.00	1	0	2	5
215	$8000.00	4	0	4	7
295	$10000.00	6	4	2	3
324	$9000.00	4	5	0	4

361	$18000.00	6	3	1	4
378	$12000.00	6	5	2	1
385	$8000.00	7	4	2	0
395	$5000.00	4	5	0	1
455	$47000.00	6	2	0	4
497	$52000.00	4	2	1	2

EDA and Data Preprocessing

Sl_No	660
Customer Key	655
Avg_Credit_Limit	110
Total_Credit_Cards	10
Total_visits_bank	6
Total_visits_online	16
Total_calls_made	11

Customer key is unique to each customer and should match the number of observations in the dataset. It is reflecting 655 of 660 so it contains duplicate values which need to be addressed before completing the analysis.

I did some data cleaning and preparation by checking for, and removing duplicates.

Index	Sl_No	Customer Key	Avg Credit Limit	Total Credit Cards	Total visits to bank	Total visits online	Total calls made
4	5	47437	$100,000.00	6	0	12	3
48	49	37252	$6,000.00	4	0	2	8
104	105	97935	$17,000.00	2	1	2	10
332	333	47437	$17,000.00	7	3	1	0
391	392	96929	$13,000.00	4	5	0	0
398	399	96929	$67,000.00	6	2	2	2
411	412	50706	$44,000.00	4	5	0	2
432	433	37252	$59,000.00	6	2	1	2
541	542	50706	$60,000.00	7	5	2	2
632	633	97935	$187,000.00	7	1	7	0

There are 5 duplicate customer keys so the first observations were kept and all others were dropped from the overall analysis.

I then decided to remove the columns Sl_No and Customer Key altogether because they are identification columns that do not provide useful information for this analysis.

Another pass over the data returned 11 rows which held the same identical customer features as other rows of data with the same features. These rows were removed because they all represent the same stats.

After removing the duplicate keys, rows, and unnecessary columns, there are now 644 unique observations across 5 categories.

Statistical Summary

index	Avg Credit Limit	Total Credit Cards	Total visits to bank	Total visits online	Total calls made
count	644	644	644	644	644
mean	$34,543.48	4.69	2.4	2.62	3.61
std	$37,428.70	2.18	1.63	2.96	2.88
min	$3,000.00	1.0	0.0	0.0	0.0
25%	$11,000.00	3.0	1.0	1.0	1.0
50%	$18,000.00	5.0	2.0	2.0	3.0
75%	$48,000.00	6.0	4.0	4.0	5.25
max	$200,000.00	10.0	5.0	15.0	10.0

This will now serve as the data by which the remaining analysis is derived from.

Univariate Analysis

This analysis is based on each category separate from any other factors.

The chart above shows the statistical summary of the 5 key categories.
In order to accomplish our end goal we need to lo
Identify different segments of customers
- By looking at

Their spending patterns
Their past interactions with the bank

Average Credit Limit
mean: $34,543.48
std: $37,428.7

* OBSERVATION: The high standard deviation shows a large variation in the credit limit.

* We can conclude that most customers will have either a high or a low credit limit, which is evident in the min and max data values.

* CONCLUSION: This can be a consideration for 2 different customer segments.

Total Credit Cards
mean: 4.69
std: 2.18
25%: 3
75%: 6

* OBSERVATION: This indicates most customers total number of credit cards are close in count to the number of the average.

* We know our data set has a clear group of high credit limit customers and also of low credit limit customers.

* CONCLUSION: The high credit limit customers have the same average amount of credit cards as the low credit limit customer.

Total visits to bank
mean: 2.4
std: 1.6
min: 0
max: 5

* OBSERVATION: The small standard deviation in comparison to the mean indicates most customers visit the bank pretty consistent to the average.

* CONCLUSION: The other categories likely do not have an effect on if a customer comes to bank in person or not.

Total visits online
mean: 2.62
std: 2.96
min: 0
max: 15

* OBSERVATION: The std is nearly the same as the mean itself and the min and max numbers vary a great deal.

* CONCLUSION: We have 2 very defined segments of customers:

1. Those who use online banking services
2. Those who do not.

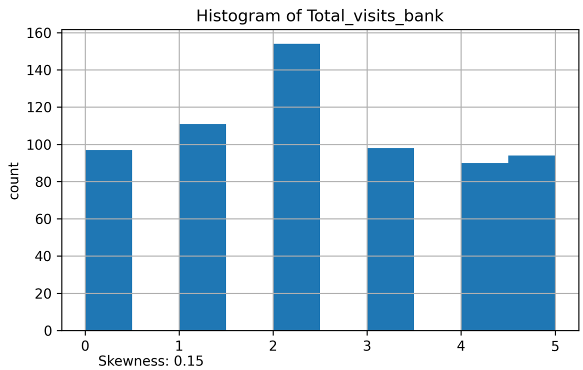
Total calls made
mean: 3.61
std: 2.88
min: 0
max: 10

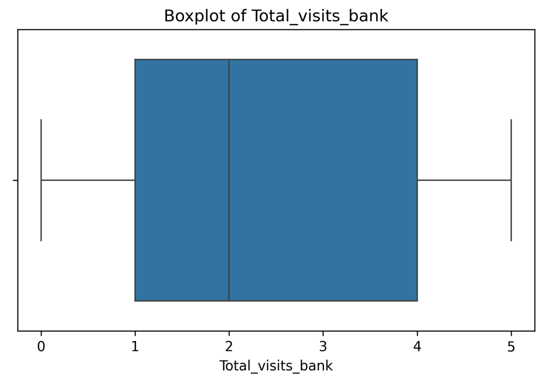
* OBSERVATION: The number of calls made to different groups varies a great deal.

* Given customers perceive the support services of the bank poorly, we see we have 2 different customer segments.

1. Ones who were called often
2. Ones who were not

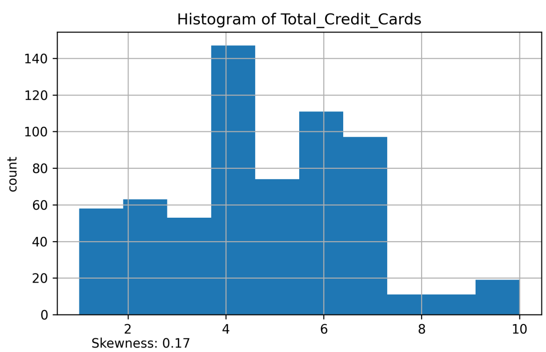
* CONCLUSION: Could there be a correlated behavior of these 2 groups in consideration of the other categories depending on if they fell into one call group or the other?

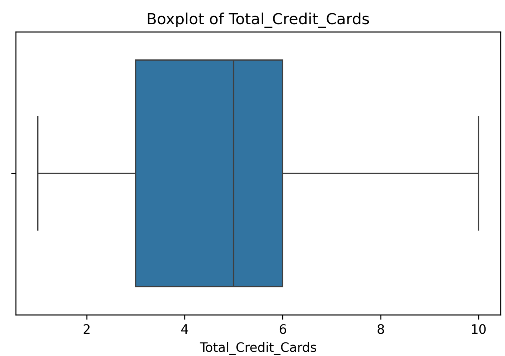




Total Visits to the Bank

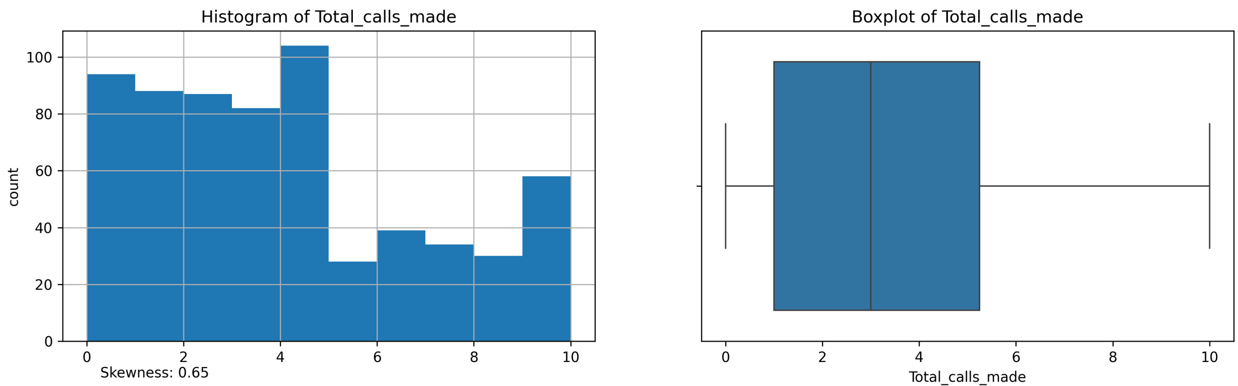
The Histogram on the left and the Boxplot on the right show a balanced occurrence of bank visits within the data.





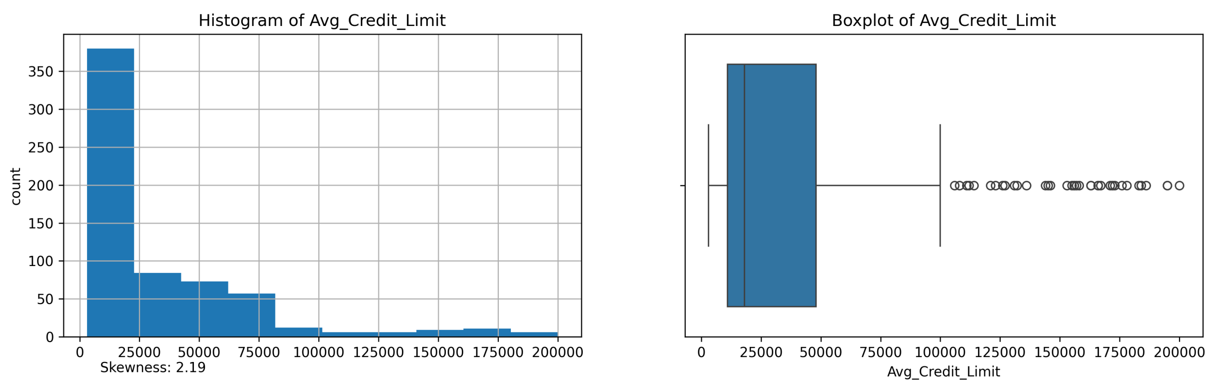
Total Credit Cards

Event hough the Boxplot is a little more skewed to the left, the difference is not significant. The total credit cards across the data is moderately balanced.



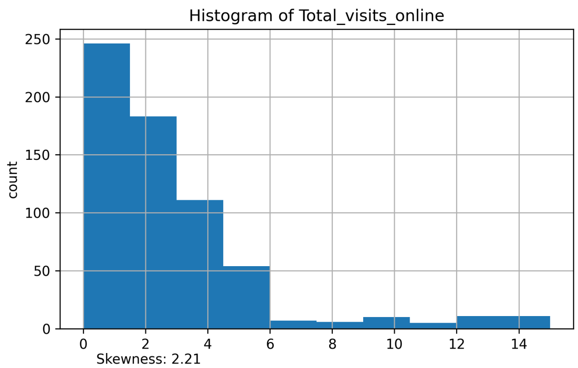
Total Calls Made

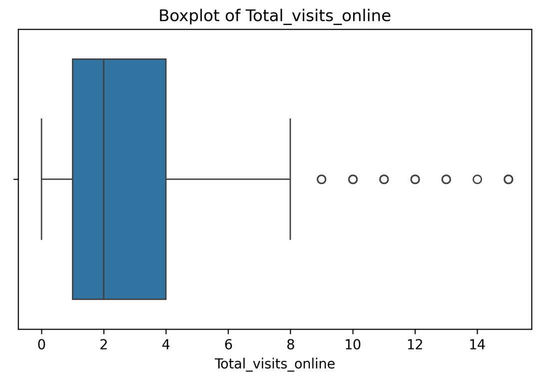
The total number of calls made is skewed negatively to the left showing there is an imbalance of calls.



Average Credit Limit

This data required the largest amount of skewing due to the greatest degree of variability within it’s own range. Even with the skewing, you can see there is a significant amount of outliers. These outliers represent a small portion of customers which have the largest credit available.



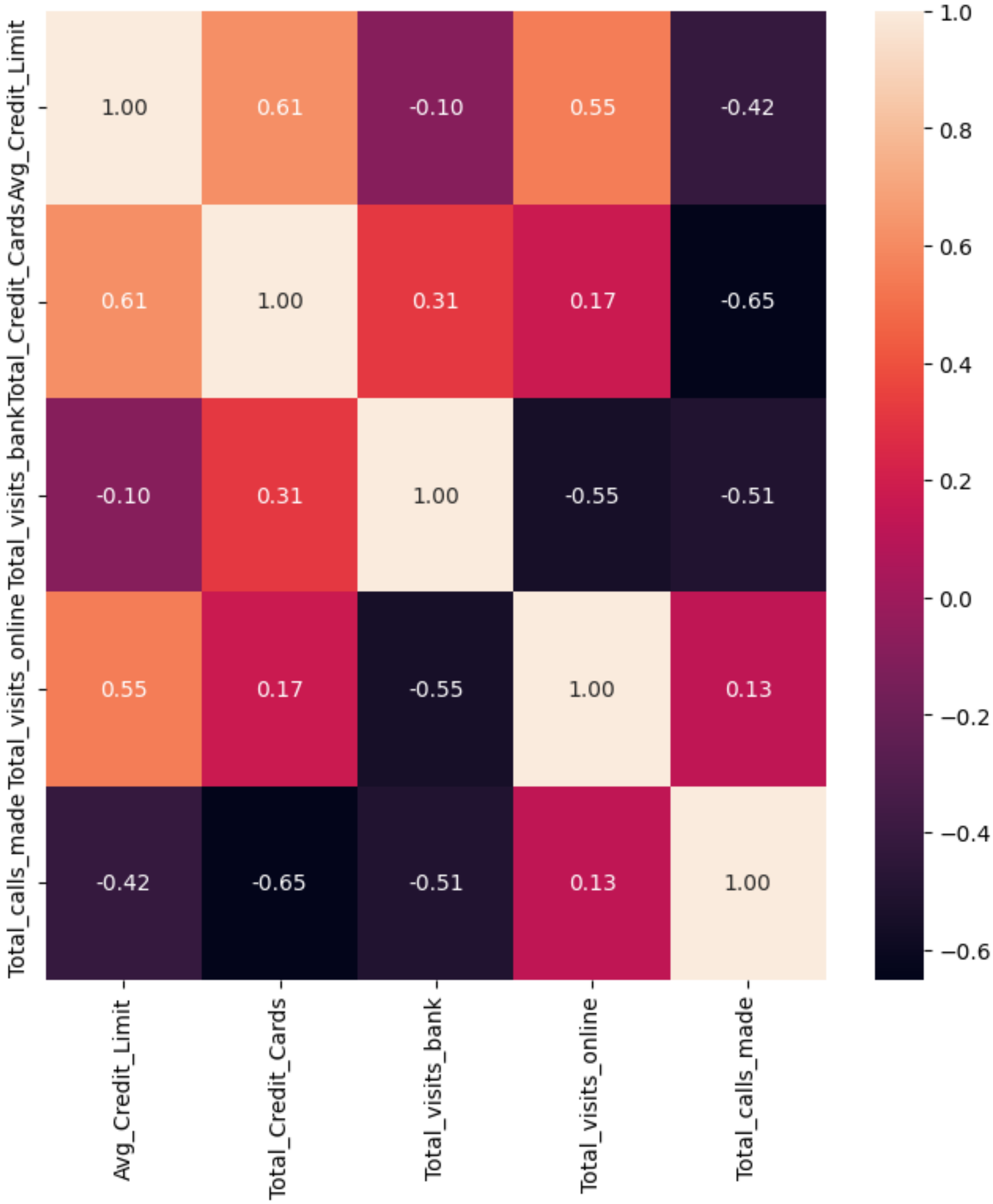


<u>Total Visits Online</u>

Here we see there are a small number of people making a large number of online visits, while the most are making less visits, or none at all.

Bivariate Analysis

This analysis takes the same categories and compares them against one another to see what effects one may have on the other.


Avg_Credit_Limit
Total_Credit_Cards
Total_visits_bank
Total_visits_online
Total_calls_made
1.00
0.61
-0.10
0.55
-0.42
0.61
1.00
0.31
0.17
-0.65
-0.10
0.31
1.00
-0.55
-0.51
0.55
0.17
-0.55
1.00
0.13
-0.42
-0.65
-0.51
0.13
1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Observations based on the Heatmap:

Average Credit Limit is positively correlated with Total Credit Cards and Total Visits Online, which makes sense.

Average Credit Limit is negatively correlated with Total Calls Made and Total Visits to the Bank.

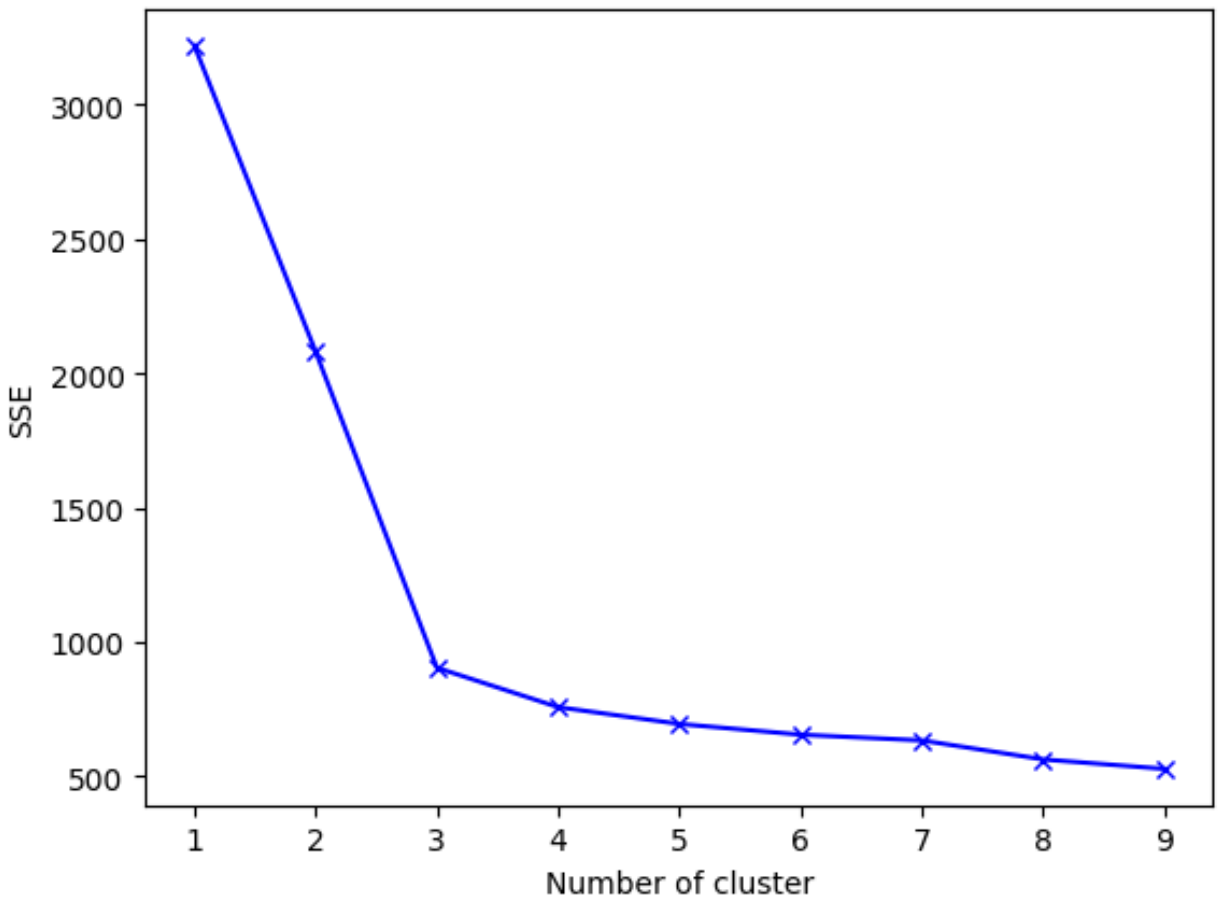
Total Visits to the Bank, Total Visits Online, and Total Calls Made are negatively correlated, which implies that a majority of customers use one or the other.

From here, to prepare the data for further modeling, the data was scaled down using the Standardization Method.

Then, the PCA Technique was applied to reduce the dimensions and create groupings based on the amount they vary or are similar.

Model Building

- K-means clustering was applied to the PCA components. The result was 9 grouped data points, named 0 - 8, which are similar.



	count
Labels	
1	105
5	105
7	86
8	82
3	79
4	78
0	60
6	26
2	23

SUMMARY STATISTICS

	kmeans_group_0 Mean	kmeans_group_1 Mean \
Avg_Credit_Limit	11200.000000	17971.428571
Total_Credit_Cards	2.166667	5.580952
Total_visits_bank	1.850000	2.485714
Total_visits_online	3.333333	0.971429
Total_calls_made	6.200000	2.076190
GmmLabels	1.000000	1.990476
kmedoLabels	0.000000	1.780952

	kmeans_group_2 Mean	kmeans_group_3 Mean \
Avg_Credit_Limit	108043.478261	52822.784810
Total_Credit_Cards	8.826087	5.810127
Total_visits_bank	0.652174	4.506329
Total_visits_online	10.956522	0.810127
Total_calls_made	1.217391	2.000000
GmmLabels	0.000000	2.000000
kmedoLabels	1.000000	1.936709

	kmeans_group_4 Mean	kmeans_group_5 Mean \
Avg_Credit_Limit	12589.743590	15771.428571
Total_Credit_Cards	2.371795	5.219048
Total_visits_bank	0.717949	4.514286
Total_visits_online	3.717949	1.142857
Total_calls_made	9.076923	1.895238
GmmLabels	1.000000	2.000000
kmedoLabels	0.000000	2.000000

	kmeans_group_6 Mean	kmeans_group_7 Mean \
Avg_Credit_Limit	168461.538462	57755.813953
Total_Credit_Cards	8.730769	5.476744

Total_visits_bank	0.538462	2.511628
Total_visits_online	11.000000	0.941860
Total_calls_made	1.000000	2.046512
GmmLabels	0.000000	2.000000
kmedoLabels	1.000000	1.302326

	kmeans_group_8 Mean	kmeans_group_0 Median \
Avg_Credit_Limit	12731.707317	10000.0
Total_Credit_Cards	2.609756	2.0
Total_visits_bank	0.487805	2.0
Total_visits_online	3.597561	3.0
Total_calls_made	5.353659	6.0
GmmLabels	1.000000	1.0
kmedoLabels	0.000000	0.0

	kmeans_group_1 Median	kmeans_group_2 Median \
Avg_Credit_Limit	17000.0	106000.0
Total_Credit_Cards	6.0	9.0
Total_visits_bank	2.0	1.0
Total_visits_online	1.0	11.0
Total_calls_made	2.0	1.0
GmmLabels	2.0	0.0
kmedoLabels	2.0	1.0

	kmeans_group_3 Median	kmeans_group_4 Median \
Avg_Credit_Limit	50000.0	12000.0
Total_Credit_Cards	6.0	2.0
Total_visits_bank	5.0	1.0
Total_visits_online	1.0	4.0
Total_calls_made	2.0	9.0
GmmLabels	2.0	1.0
kmedoLabels	2.0	0.0

	kmeans_group_5 Median	kmeans_group_6 Median \
Avg_Credit_Limit	13000.0	166500.0
Total_Credit_Cards	5.0	9.0
Total_visits_bank	5.0	1.0

Total_visits_online	1.0	11.0
Total_calls_made	2.0	1.0
GmmLabels	2.0	0.0
kmedoLabels	2.0	1.0

	kmeans_group_7 Median	kmeans_group_8 Median
Avg_Credit_Limit	57000.0	13000.0
Total_Credit_Cards	6.0	3.0
Total_visits_bank	3.0	0.0
Total_visits_online	1.0	4.0
Total_calls_made	2.0	5.0
GmmLabels	2.0	1.0
kmedoLabels	1.0	0.0

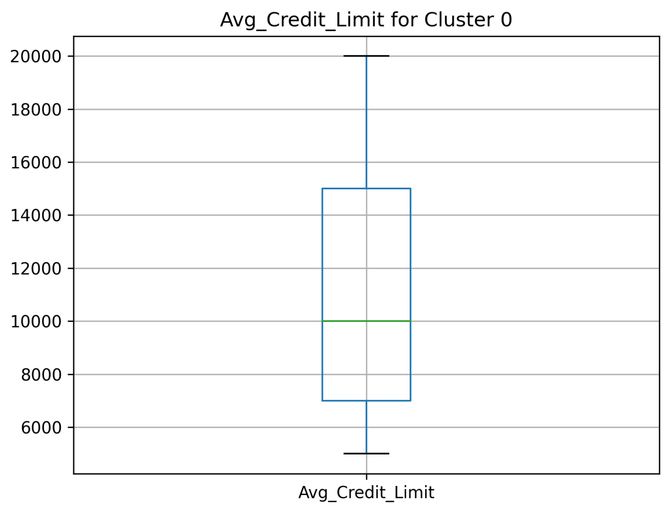
By observing the Summary Statistics, we can segment Customer profiles

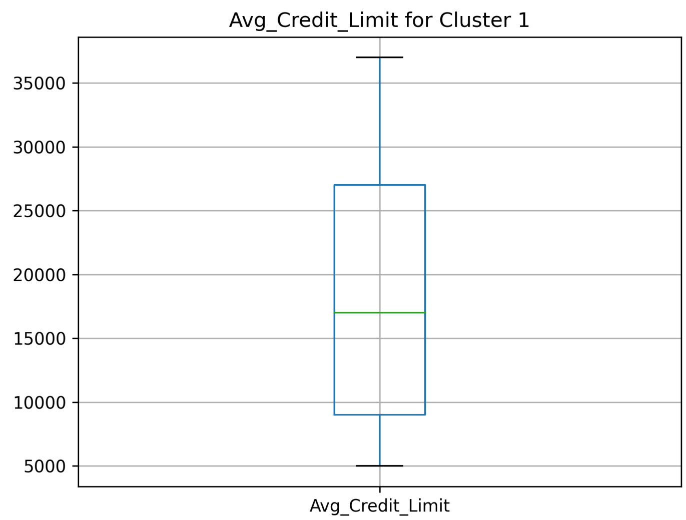
<u>Cluster 0:</u>

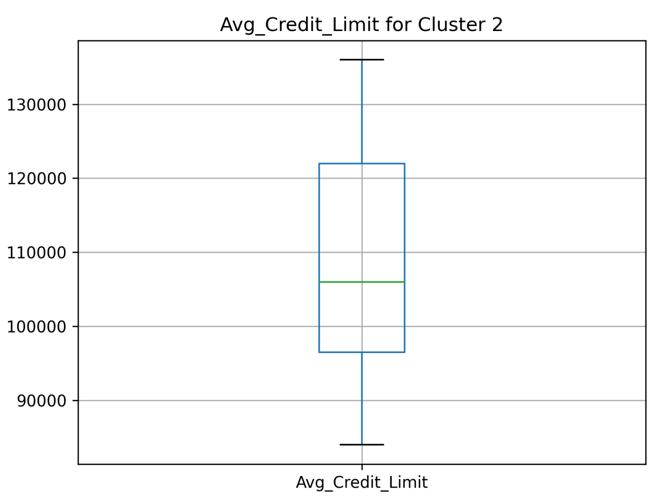
- High average credit limit, high average total spending, mostly active customers
- Most valuable customer segment

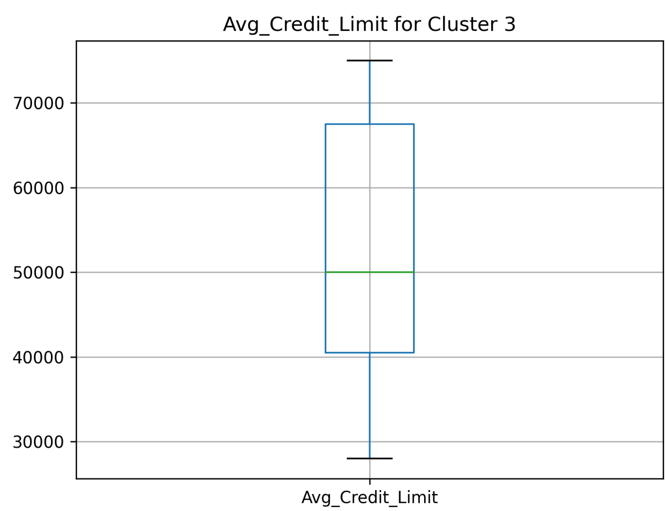
<u>Cluster 1:</u>

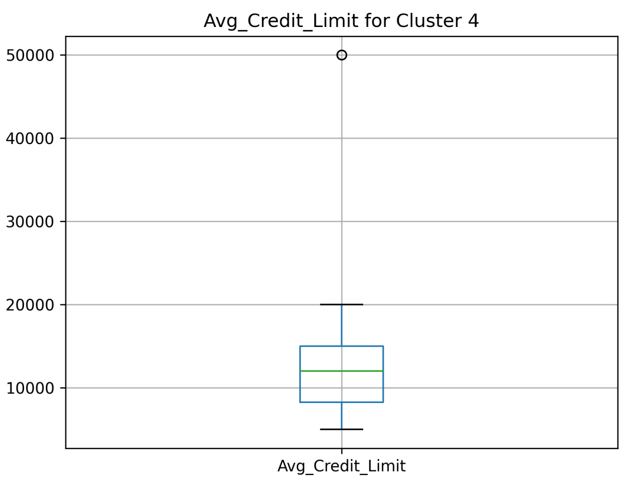
- Low average credit limit, low spending, less active
- Budget-Conscious
- Tend to spend less with moderate activity levels

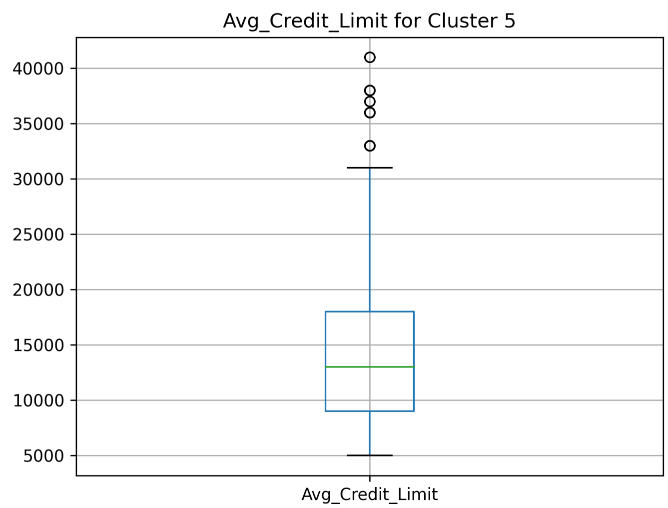

Avg_Credit_Limit for Cluster 0
20000
18000
16000
14000
12000
10000
8000
6000
Avg_Credit_Limit

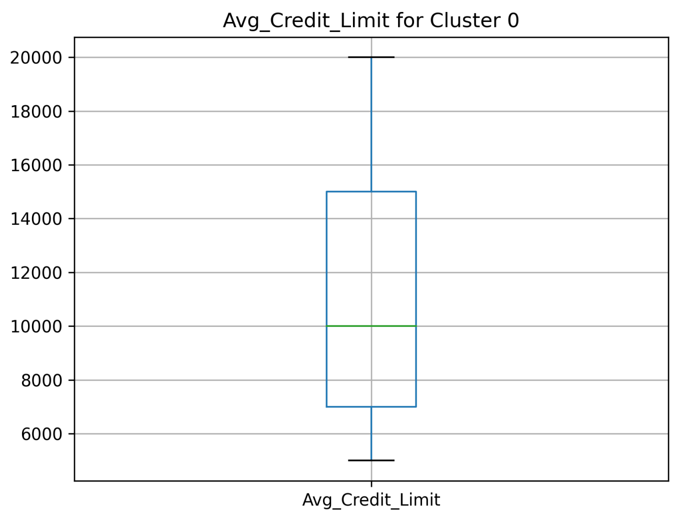

Avg_Credit_Limit for Cluster 1
35000
30000
25000
20000
15000
10000
5000
Avg_Credit_Limit

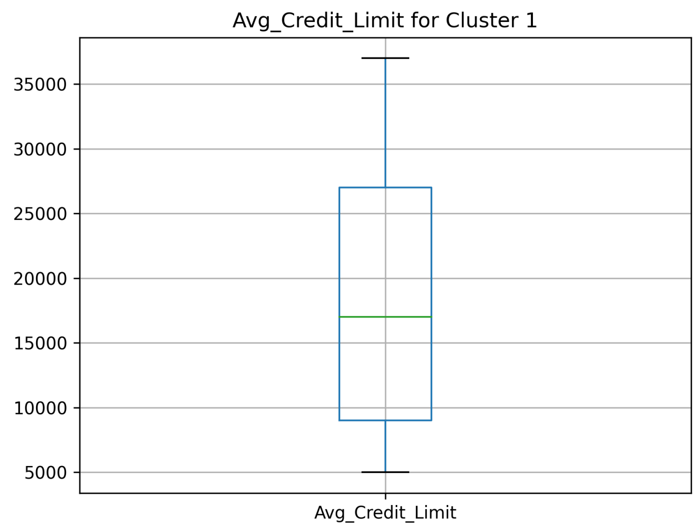

Avg_Credit_Limit for Cluster 2
130000
120000
110000
100000
90000
Avg_Credit_Limit

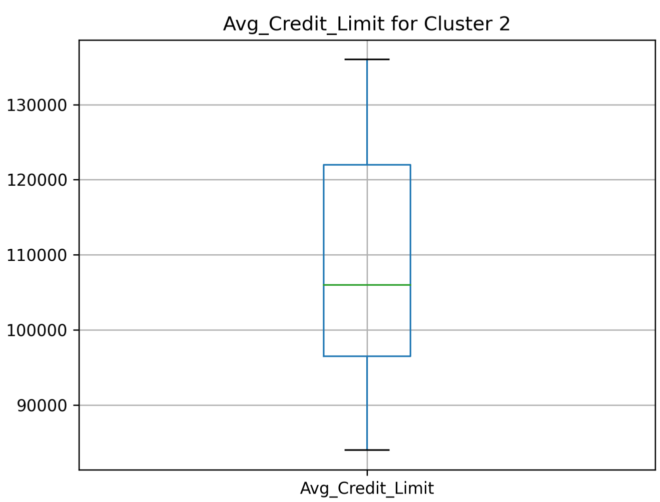

Avg_Credit_Limit for Cluster 3
70000
60000
50000
40000
30000
Avg_Credit_Limit

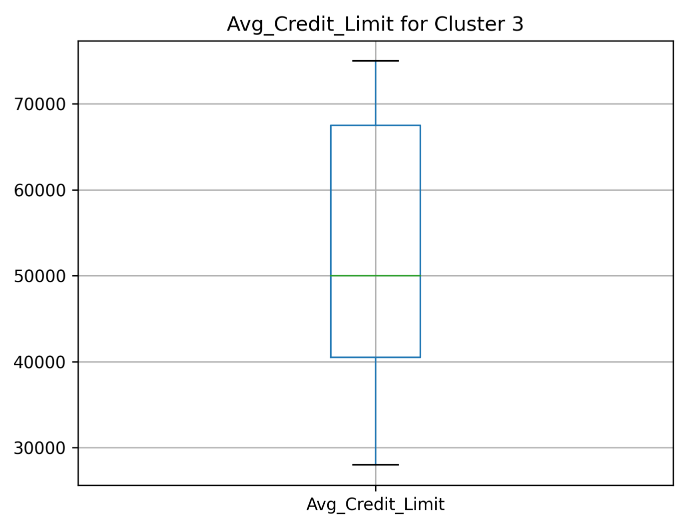

Avg_Credit_Limit for Cluster 4
50000
40000
30000
20000
10000
Avg_Credit_Limit

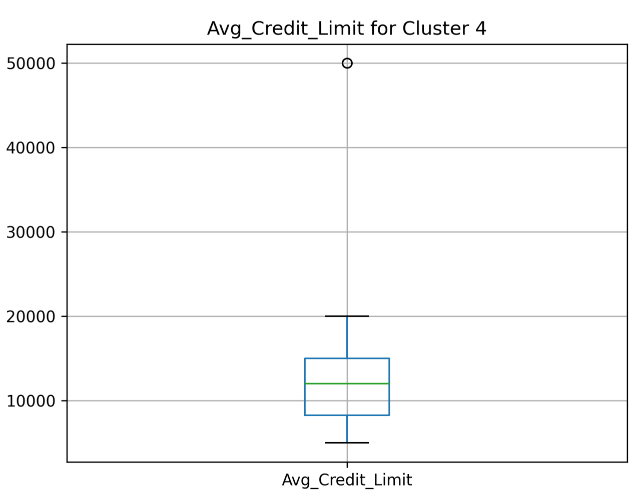

Avg_Credit_Limit for Cluster 5
40000
35000
30000
25000
20000
15000
10000
5000
Avg_Credit_Limit

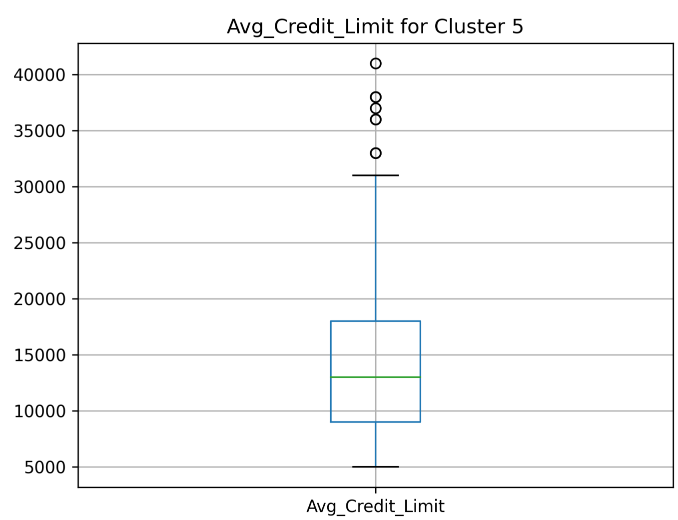

Avg_Credit_Limit for Cluster 0
20000
18000
16000
14000
12000
10000
8000
6000
Avg_Credit_Limit

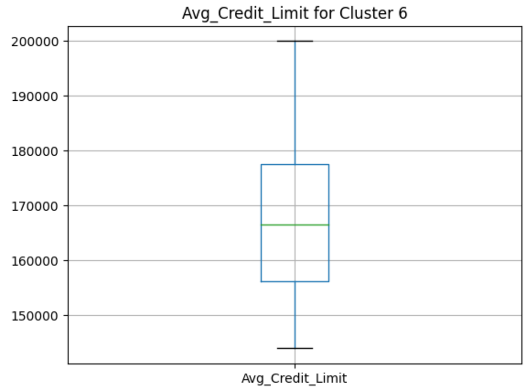

Avg_Credit_Limit for Cluster 1
35000
30000
25000
20000
15000
10000
5000
Avg_Credit_Limit

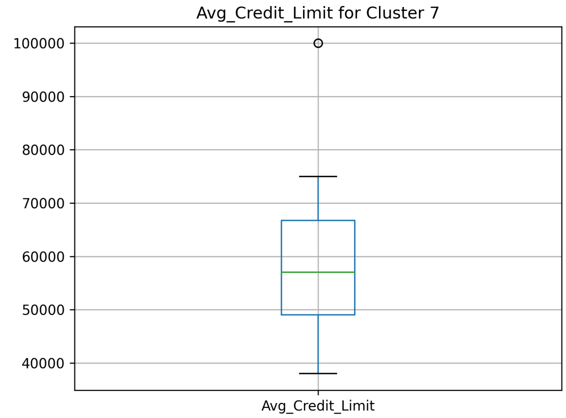

Avg_Credit_Limit for Cluster 2
130000
120000
110000
100000
90000
Avg_Credit_Limit

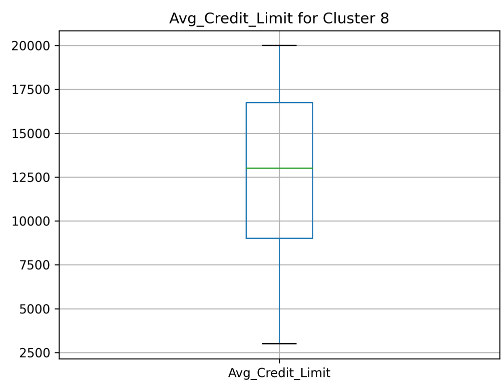

Avg_Credit_Limit for Cluster 3
70000
60000
50000
40000
30000
Avg_Credit_Limit


Avg_Credit_Limit for Cluster 4
50000
40000
30000
20000
10000
Avg_Credit_Limit


Avg_Credit_Limit for Cluster 5
40000
35000
30000
25000
20000
15000
10000
5000
Avg_Credit_Limit





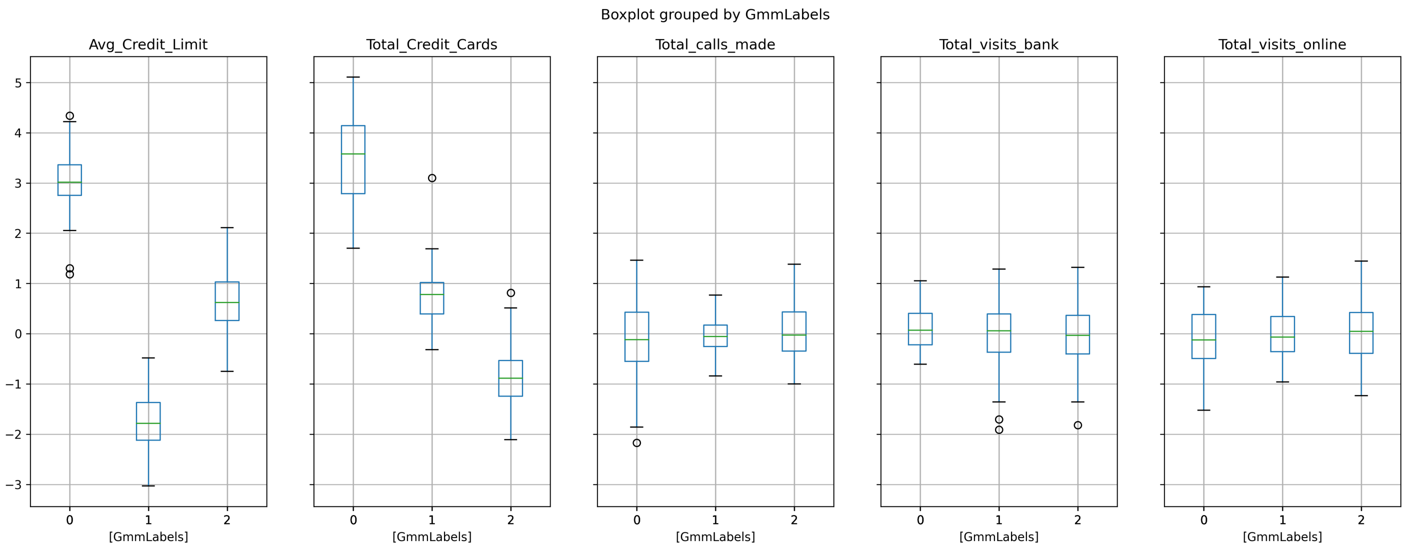


Gaussian Mixture Models (GMMs) assign probabilities of “belonging” to each cluster and are a great application to use when looking at customer segmentation.

This method has grouped 3 clusters together and provided a count within each one.

	count
Gmm Labels	
2	374
1	221
0	49

index	Gmm_group_0 Mean	Gmm_group_1 Mean	Gmm_group_2 Mean	Gmm_group_0 Median	Gmm_group_1 Median	Gmm_group_2 Median
Avg_Credit_Limit	140102.0408	12239.81900	33893.04812	145000.0	12000.0	31500.0
Total_Credit_Cards	8.775510204	2.411764705	5.508021390	9.0	2.0	6.0
Total_visits_bank	0.591836734	0.945701357	3.489304812	1.0	1.0	3.0
Total_visits_online	10.97959183	3.561085972	0.975935828	11.0	4.0	1.0
Total_calls_made	1.102040816	6.891402714	1.997326203	1.0	7.0	2.0



The Boxplots above show the comparison of the groups when the GMM method of clustering was applied.
This method shows the categories which hold strong similarities.

K-Medoids Method uses actual data points, unlike K-Means where cluster centers are the calculated means/ average.

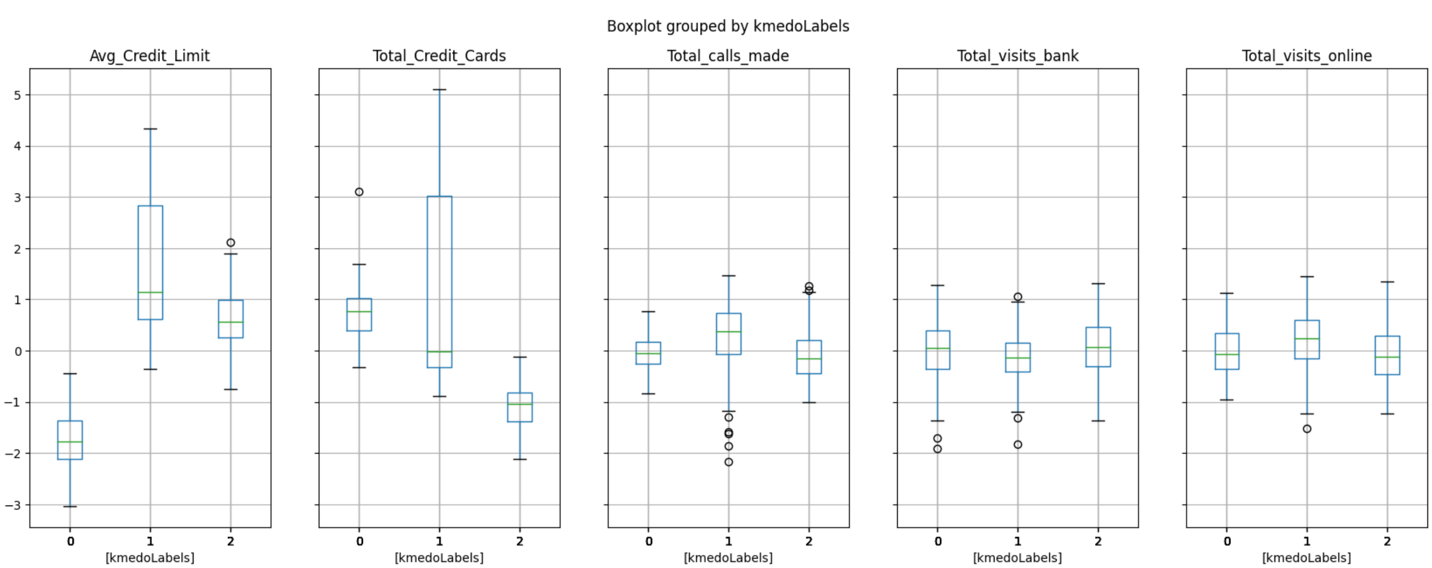
K-Medoids are useful when looking at market segmentation to identifying distinct customer groups, based on their purchasing habits, demographics, or other characteristics.

Here we have another set of clusters with the probability count.

	count
kmedoLabels	
2	289
0	222
1	133

SUMMARY STATISTICS

index	kmedoids_group_0 Mean	kmedoids_group_1 Mean	kmedoids_group_2 Mean	kmedoids_group_0 Median	kmedoids_group_1 Median	kmedoids_group_2 Median
Avg_Credit_Limit	12216.2162162	85052.631578	28449.826	12000.0	68000.0	20000.0
Total_Credit_Cards	2.42342342342	7.0300751879	5.3633217	2.0	7.0	5.0
Total_visits_bank	0.95045045045	1.6917293233	3.8304498	1.0	2.0	4.0
Total_visits_online	3.55405405405	4.6390977443	0.9826989	4.0	2.0	1.0
Total_calls_made	6.87837837837	1.9699248120	1.8512110	7.0	2.0	2.0



The Box Plot of Kmedo Labels looks very similar to the GMM Box Plot, with the exception of credit limit and credit cards. This shows the customers in that grouping fall into different categorical data based on the clustering approach and potentially have a broader range to consider marketing to.

———————————————————————————————————

Cluster profiles created
The comparison profile's labels were all generic based on the original cluster method used, such as 'group_1_median'. There were repeating labels so to differentiate between each of them, I relabeled the indexes for each cluster to clearly define which index value was represented in the comparison.

NaN Values from 0.00000 values
A few NaN values were present. I replaced those with zeros for more uniformity and to maintain the integer structure within the comparison table.

———————————————————————————————————

Conclusions and Recommendations

The goal of this analysis was to identify existing AllLife Bank customer groups, and gain insight into their spending habits and behaviors. Using that information, I needed to identify segmented customer groups which could be marketed to for various goals in line with AllLife's focus in the coming financial year.

The greatest variance in categories are within the number of customers who go online and the credit limit among the customers. The least variance is noted in the ares of bank visits and amount of credit cards.

To close the gap between these two groups and also accomplish the goal of the Marketing and Operations team, as well as progress forward with AllLife's initiative for the next financial year, I suggest the following strategies:

- Problem - The customers have a poor perception of the bank's support services.

- Goal - To upgrade the service delivery model, ensuring that customers' queries are resolved faster.
- Data based solution - Develop a guaranteed 1 business day turn around for support cases submitted to the bank via the online application.

- Problem - The need to improve market penetration.
- Goal - Target new customers.
- Data based solution - Market a Secure Credit Card offering to customers with a lower credit limit. Run a secondary campaign for a bonus cash savings card for every friend they refer into AllLife.

The data from this report is significant enough to foster several campaigns which can be beneficial to AllLife and support them in their goals. The two mentioned above are the most all encompassing in consideration of the overall objective of all the departments as a whole.