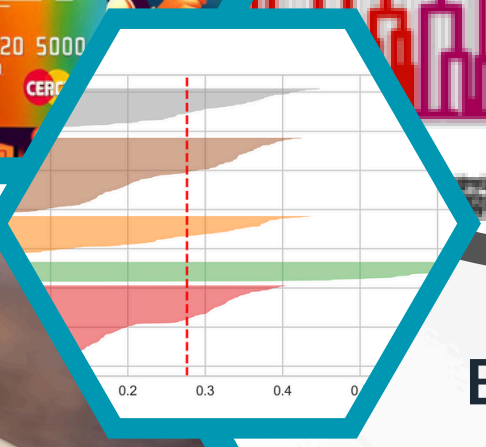
UNSUPERVISED LEARNIG



BUSINESS REPORT

-- HARAPRASAD DHAL

ALLLIFE BANK CREDIT CARD CUSTOMER SERVICE CASE STUDY

Great Learning
POWER AHEAD

TEXAS McCombs
The University of Texas at Austin
McCombs School of Business

Dhal analytics.

dhalharaprasad@gmail.com

Contents

List of Figures

AllLife Bank Credit Card Customer Service

1 Business context

AllLife Bank aims to enhance its credit card customer base in the upcoming financial year. Insights from the marketing research team suggest that market penetration can be improved. Consequently, the Marketing team plans to launch personalized campaigns targeting new customers and upselling to existing ones. Additionally, market research revealed that customers perceive the bank's support services poorly. In response, the Operations team intends to upgrade the service delivery model to ensure faster resolution of customer queries. Both the Head of Marketing and Head of Delivery have decided to seek assistance from the Data Science team to achieve these goals.
As a data scientist at AllLife Bank, my objective is to identify distinct customer segments based on their spending patterns and past interactions with the bank. By employing clustering algorithms, I aim to uncover meaningful insights that will enable the bank to tailor its marketing strategies and enhance customer service. My recommendations will focus on how to effectively target new customers and up sell to existing ones, ensuring a more personalized and efficient approach to customer engagement.

2 Exploratory data analysis

2.1 Data set description

The dataset provided includes information on various bank customers and their financial attributes, such as credit limits and the total number of credit cards they hold. It also details the different channels through which customers have contacted the bank for queries, including in-person visits, online interactions, and calls to the call center. The data set consists of 660 records and 7 columns. I dropped

	Sl_No	Customer Key	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made
547	548	38125	26000	4	5	2	4
353	354	94437	9000	5	4	1	3
499	500	65825	68000	6	4	2	2
173	174	38410	9000	2	1	5	8
241	242	81878	10000	4	5	1	3

Figure 1: A snapshot of data set used for analysis.

two columns which represented unique customers and Serial number. Not useful for the analysis. After this we have 5 columns and all are numerical. No missing value or irregular values were found. All the records were unique. My motivation is to cluster these customers based on this five dimensional data set.

2.2 Univariate and Bivariate Analysis

The figure 2 shows the distribution of all 5 numerical variables in the dataset. It shows the Histogram on the top of which box plot is aligned with the same scale. Figure 2a shows the distribution of Average Credit Limit of customers. On average customers have a limit of 34,574. The Mean is much higher than median which shows that the distribution is highly right skewed. There are many outliers on the higher end of limit. Figure 2b the distribution of total number of credit card owned by the customers from AllLife Bank. 75% of customers have 6 or less number of credit cards. The maximum number owned is 10. Figure 2c shows the distribution of number of visits from the customers to bank. Maximum number of times visited is 5. Figure 2d shows the distribution for online visits to banks websites and services. 75% visited 5 times or less. But there are many outliers up to 15 times. Similarly from figure 2e we can see the distribution for number of phone calls made by the customers.
I tried to investigate how number of cards and credit limit owned by the customers varied with respect to how many times they contacted or visited bank through different channels. Figure 3 shows the Credit cards and average credit limits distribution with respect to online visits. We observe that customer visits higher than 6 have significantly higher number of credit cards and limit. In fact from this plot we

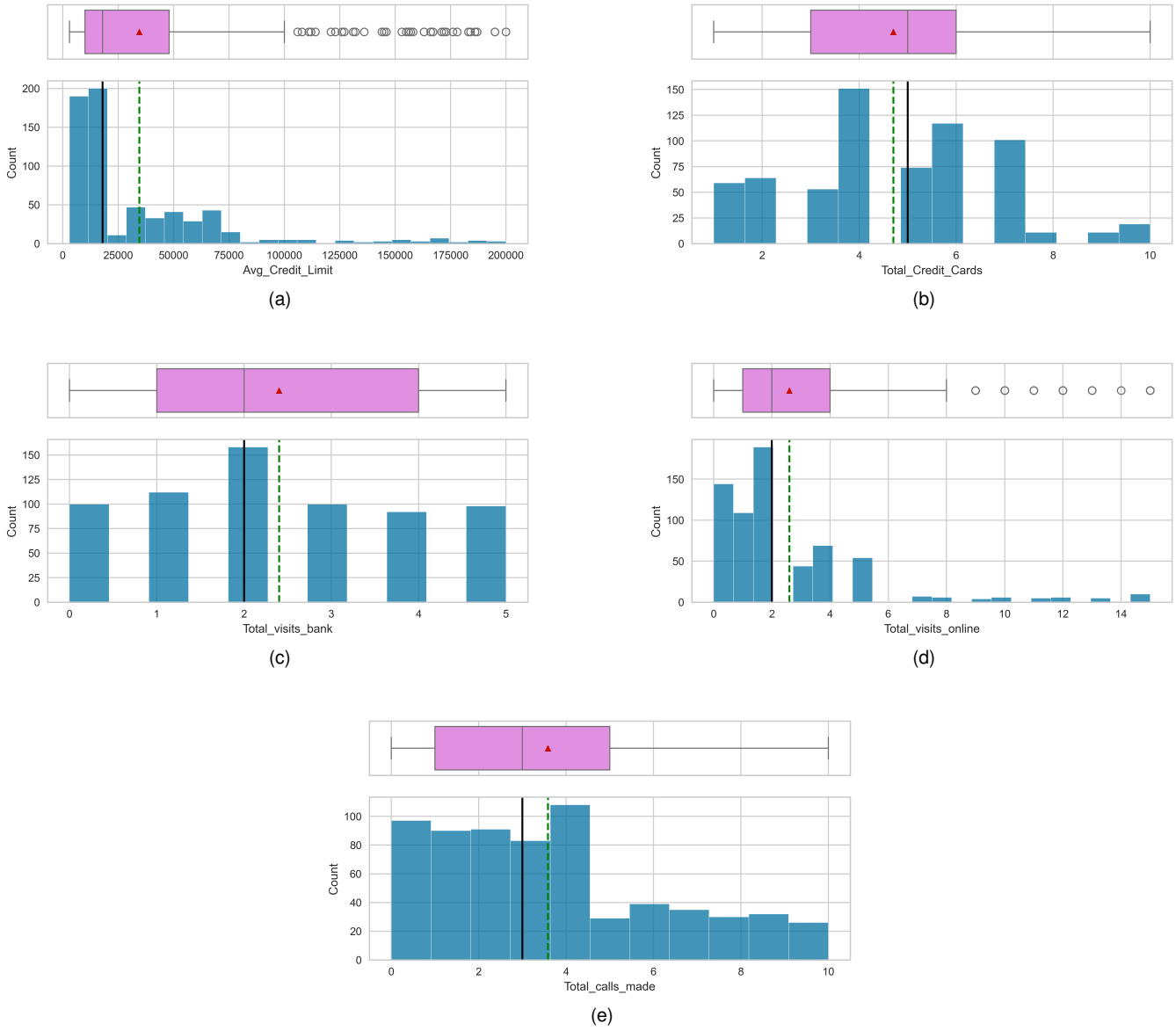
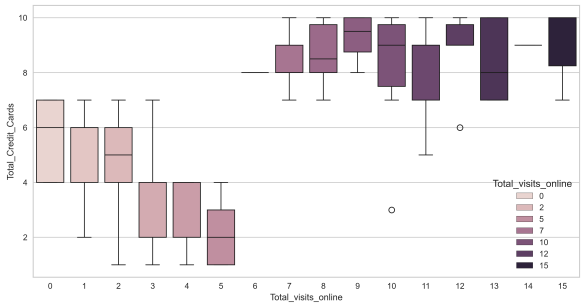
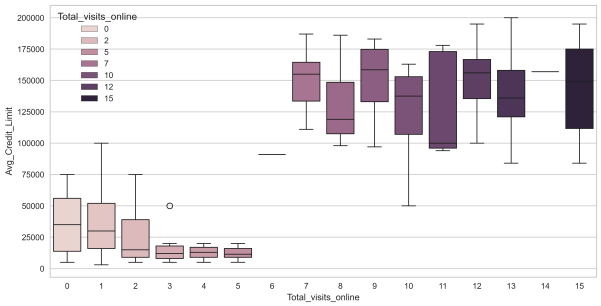


Figure 2: Numerical variables distribution

sort of get a glimpse into two major clusters. Therefore online visit is a good segregate variable for different customers group. Figure 4 shows the Credit cards and average credit limits distribution with respect to Bank visits. We observe that the variation is more pronounced for Total number of credit cards. Customers with higher number of cards made more visits to Banks which is expected since more the cards the more maintenance and services one might need. Similar story is also observed for the Total calls made in case of Number of credit cards as shown in figure 5 which depicts Credit cards and average credit limits distribution with respect to Number of calls made.
Figure 6 shows the heat map for the correlation among five variables. Bank visits and Online visits are negatively correlated with -0.55. This shows that customers using online services/modes are less likely to visit the bank in person. we also observe that number of credit cards and number of phone calls are negatively correlated with -0.65. In contrast to this number of credit cards is positively correlated to online visits. Customers with higher credit cards and limit are more likely to use online methods.

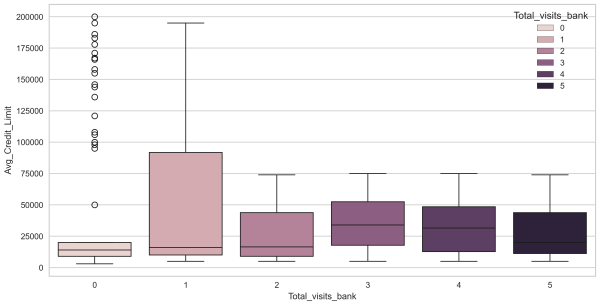


(a) Total credit cards vs total online visits.

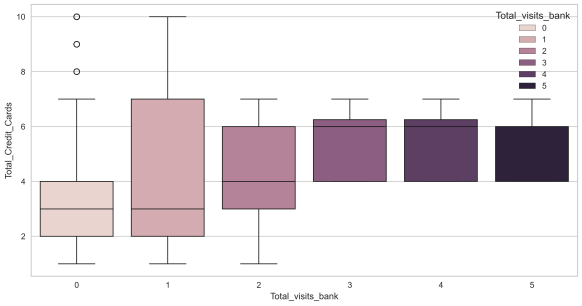


(b) Average credit limit vs Total online visits.

Figure 3: Credit cards and average credit limits distribution with respect to online visits.

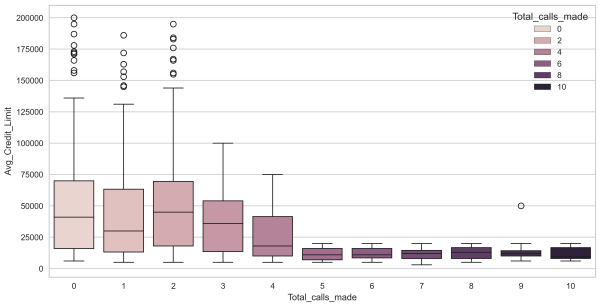


(a) Average credit limit vs Total Bank visits.

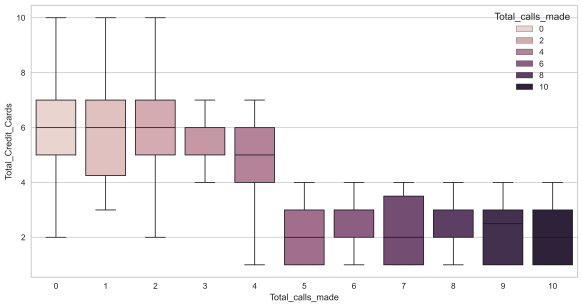


(b) Total credit cards vs total Bank visits.

Figure 4: Credit cards and average credit limits distribution with respect to Bank visits.



(a) Average credit limit vs Total calls made.



(b) Total credit cards vs total calls made.

Figure 5: Credit cards and average credit limits distribution with respect to Number of calls made.

Figure 6: Correlation among the variables.

3 Data Preprocessing

Data preprocessing is a crucial step in clustering to ensure that the data is clean, consistent, and suitable for analysis. The data set did not have any irregular or missing values. As for the outliers that we found in credit limit and online visits, it would not have been a good choice to remove these, since all of these records are valid and also no instability was found during clustering implementation due to these outliers which off course if described posterior to implementation. Normalization of the data ensures that all features contribute equally to the distance calculations. Common methods include Min-Max scaling and Z-score normalization. I carried out the Z-score on all the five features since the range of values for credit limit is significantly higher than all other columns. This was necessary to avoid bias to any particular feature. After the transformation the normalized dataset is shown in distribution plot figure 8. We see that all the features are now centered around the mean zero.

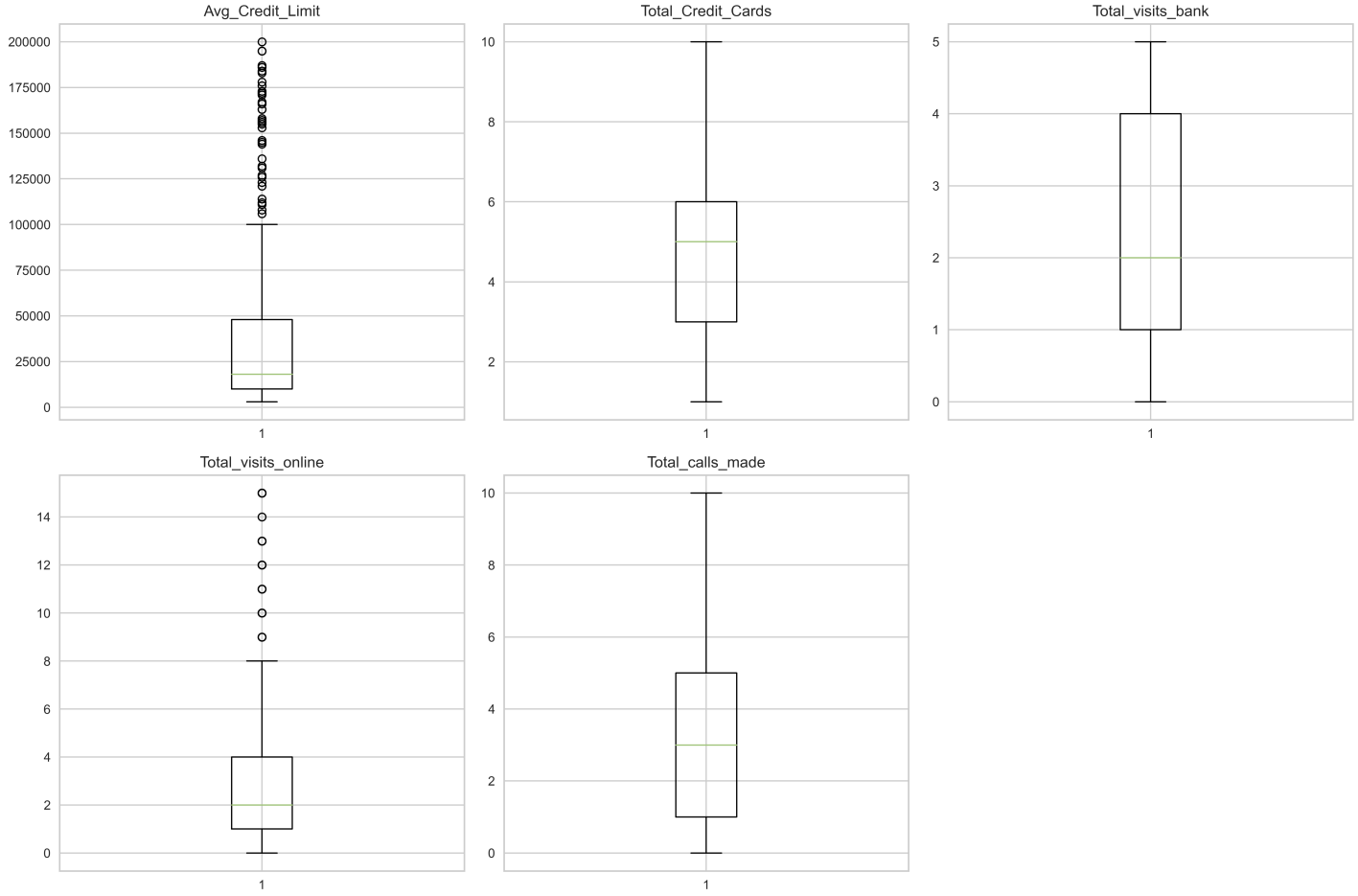


Figure 7: Plot for outlier detection

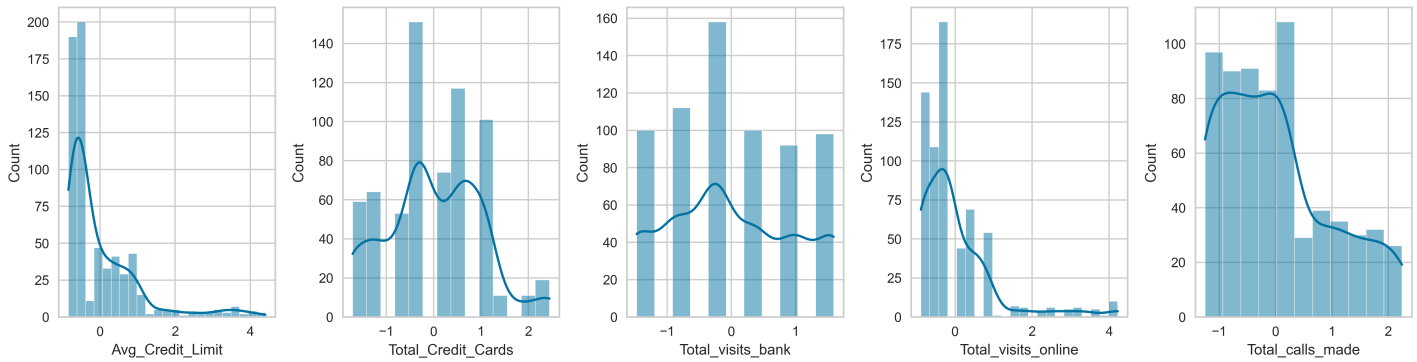
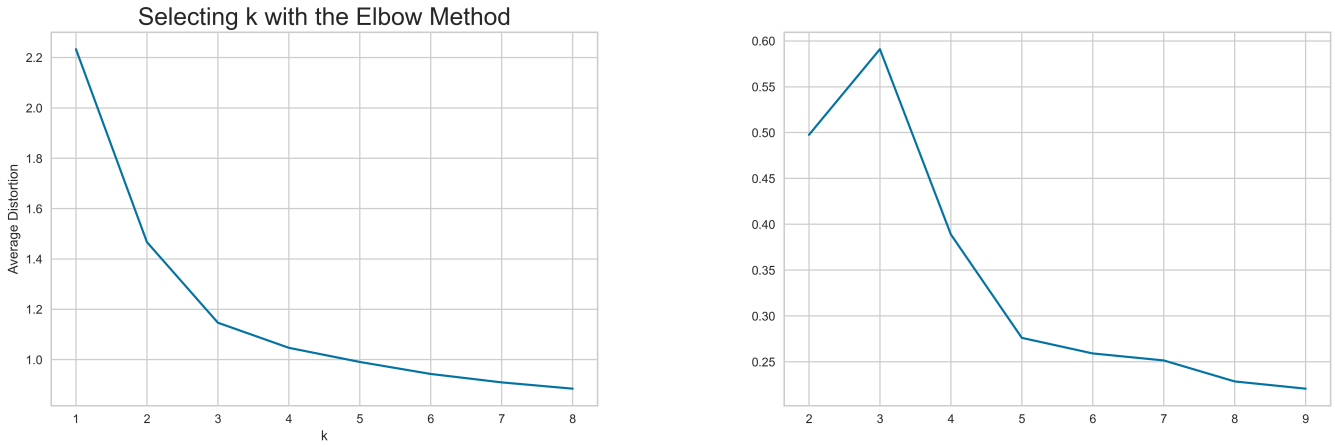


Figure 8: scaled data

4 K-means Clustering

K-Means clustering is a popular unsupervised machine learning algorithm used to partition a dataset into distinct groups or clusters. The goal is to divide the data points into k clusters, where each data point belongs to the cluster with the nearest mean. The steps involved are as follows :

- Initialization: Choose the number of clusters k and randomly initialize k centroids.
- Assignment: Assign each data point to the nearest centroid based on the Euclidean distance.
- Update: Recalculate the centroids as the mean of all data points assigned to each cluster.
- Repeat: Repeat the assignment and update steps until the centroids no longer change significantly or a maximum number of iterations is reached.



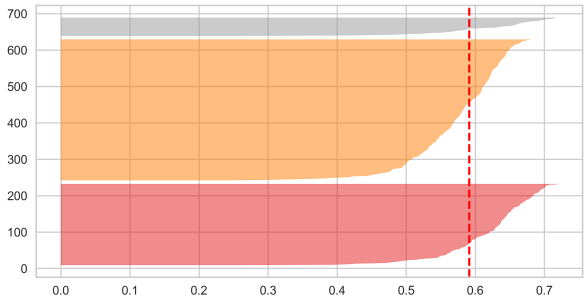
(a) Average distortion across K to find elbow point.

(b) Silhouette score with K

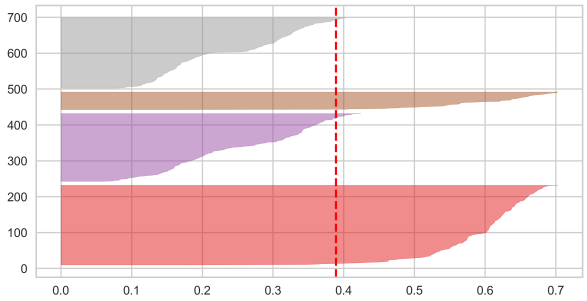
Figure 9: Average distortion and Silhouette score across K.

The metric used here for the KMeans distance is Euclidean. To figure out suitable K for the given data set I checked the Elbow method and Silhouette score for different K. The Plot of the sum of squared distances from each point to its assigned centroid (within-cluster sum of squares) for different values of k is shown in figure 9a. The optimal k is often found at the ”elbow” point where the rate of decrease slows down. we observe that K =3. Also along side we calculate the silhouette score for different K as shown in figure 9b and 9. Silhouette score measures how similar a data point is to its own cluster compared to other clusters. A higher silhouette score indicates better-defined clusters. From this analysis we find that the optimal k for this data set is 3. The Silhouette score is highest for k = 3 : 0.59. we chose the final KMeans model with 3 clusters.
Figure 11 shows the cluster profile of 3 clusters obtained from the final KMeans model. It shows the mean values of each field corresponding to all three clusters. Figure 12 shows box plot for all the numerical variables in each cluster categories.
From these two figures, 11 and 12 a summary for each cluster properties is generated.

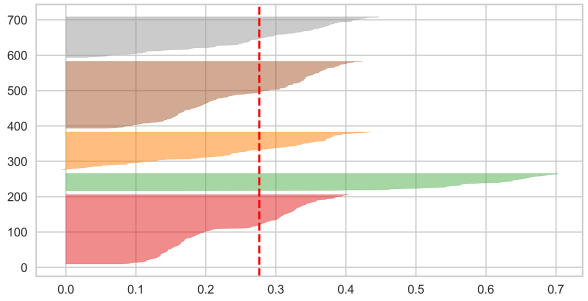
- Cluster 0:
  - There are 386 customers in this segment out of 660.
  - 75% of customers in this group made less than 3 calls.
  - Atmost three online visits are made from these customers.
  - These customers have relatively higher number of in person Bank visits than other two.
  - 75 % have 6 or less number of credit cards.
  - Median credit limit is around 30,000.
- Cluster 1:



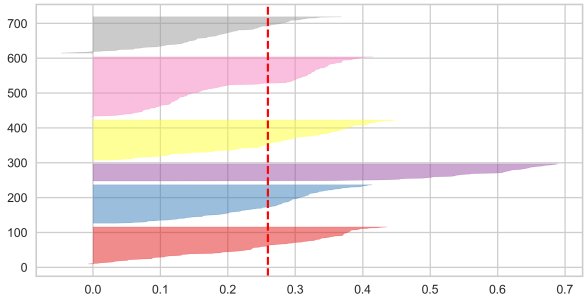
(a) Silhouette plot of KMeans Clustering for 660 samples in 3 centers.



(b) Silhouette plot of KMeans Clustering for 660 samples in 4 centers.



(c) Silhouette plot of KMeans Clustering for 660 samples in 5 centers.



(d) Silhouette plot of KMeans Clustering for 660 samples in 6 centers.

Figure 10: Silhouette Score visualization with varied K

K_means_segments	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	count_in_each_segment
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

Figure 11: Cluster profile from KMeans. K = 3

  - There are 224 customers in this segment.
  - This segment have made higher number of phone calls. At least half of them made more than 6 calls.
  - 75 % of them made less than 4 online visits.
  - All of them had at most two visits to the bank in person.
  - They have at most 4 credit cards.
  - Median credit limit is very less around 12,000.

- Cluster 2:

  - There are only 50 customers in this segment out of 660.
  - 75 % of them made less than two phone calls.
  - They made relatively a very high number of online visits.
  - These customers don't visit to bank in person. At most they made 1 visit personally.
  - They hold higher number of credit cards. Up to 10.
  - Credit limit is also very high reaching up to 200,000.

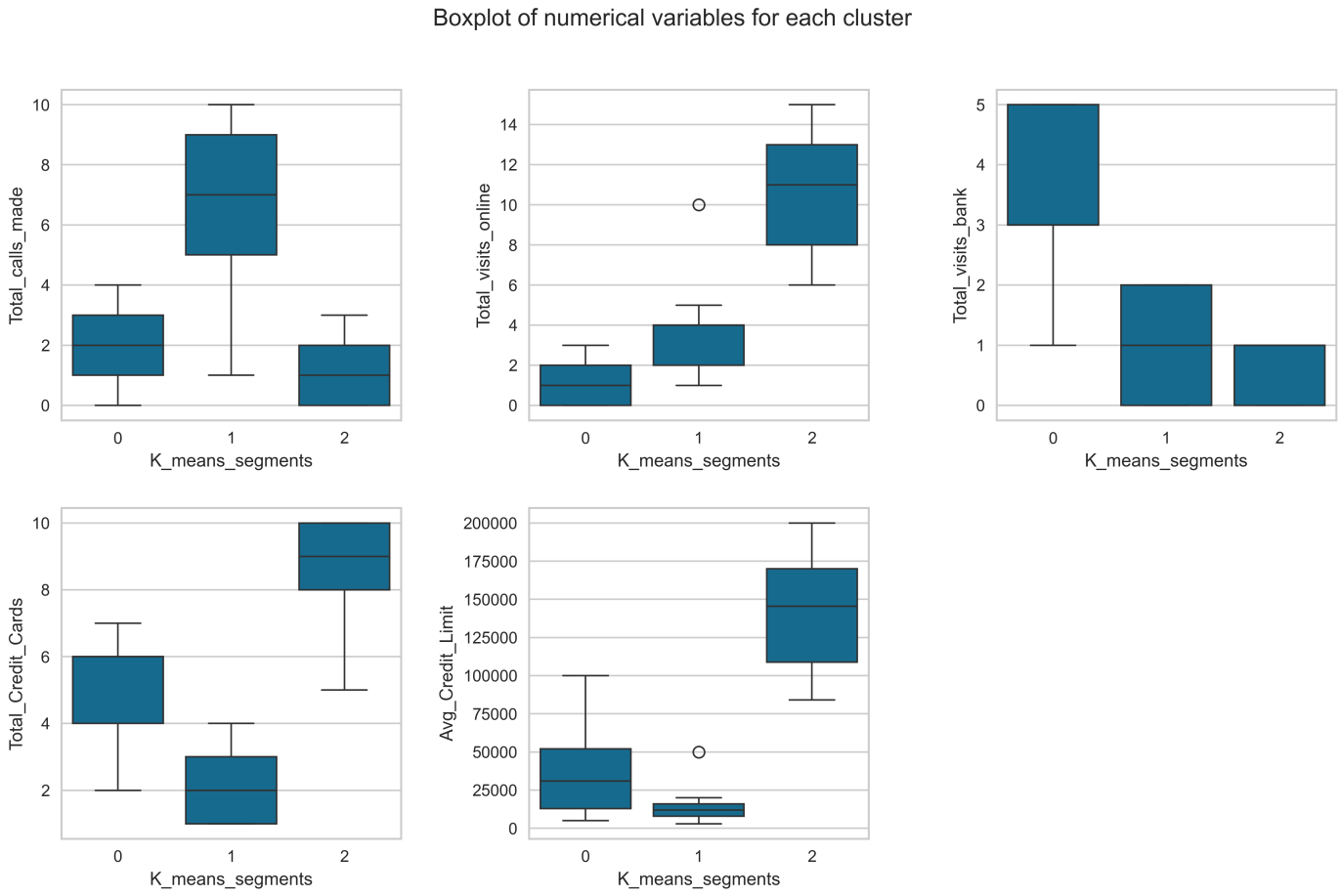


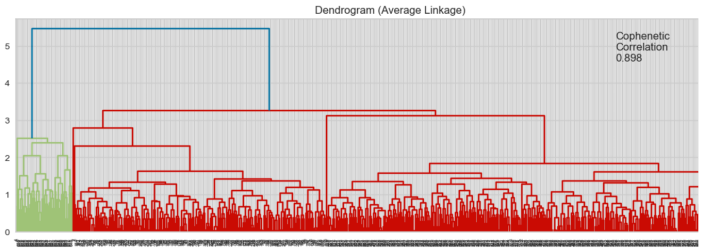
Figure 12: Box plot of different clusters obtained from KMeans.

5 Hierarchical Clustering

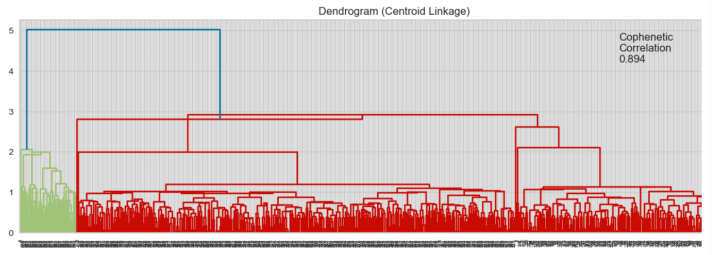
Hierarchical clustering is a method of cluster analysis that seeks to build a hierarchy of clusters. It is particularly useful for identifying nested clusters and understanding the structure of the data. There are two main types of hierarchical clustering:

- Agglomerative (Bottom-Up) Clustering: This approach starts with each data point as its own cluster and iteratively merges the closest pairs of clusters until all points are in a single cluster or a desired number of clusters is reached.
- Divisive (Top-Down) Clustering: This approach starts with all data points in a single cluster and iteratively splits the clusters into smaller clusters until each data point is its own cluster or a desired number of clusters is reached.

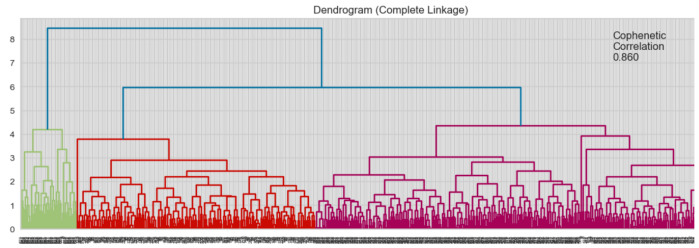
In this analysis, I have used Agglomerative Clustering approach. To figure out appropriate number of cluster for the data set I have used the dendrogram which is a tree-like diagram that shows the arrangement of the clusters produced by hierarchical clustering. Along side this I have calculated The Cophonetic correlation coefficients for various combination of metric and Linkage Methods. The cophenetic correlation coefficient measures how faithfully the dendrogram preserves the pairwise distances between the original data points. A higher cophenetic correlation indicates that the dendrogram accurately reflects the original data structure.



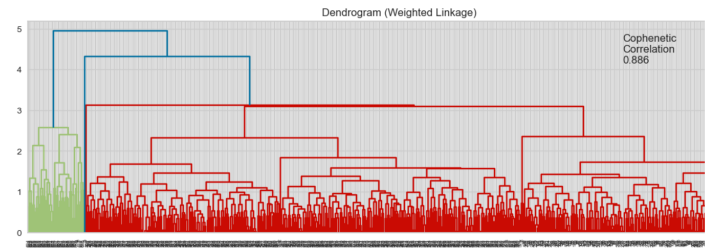
(a) Dendrogram for Average linkage. Cophenetic correlation = 0.898



(b) Dendrogram for centroid linkage. Cophenetic correlation = 0.894.



(c) Dendrogram for Complete linkage. Cophenetic correlation = 0.860 .



(d) Dendrogram for Weighted linkage. Cophenetic correlation = 0.886.

Figure 13: Dendogram with different linkage methods for Euclidean metric.

I found that among different metric and linkage methods, Euclidean metric with Average linkage gave best correlation of 0.898. Figure 13 shows Dendrograms with different linkage methods for Euclidean metric with at least 0.8 of cophenetic correlation. From these reliable dendrograms it is observed that 3 clusters would have been appropriate for the dataset as beyond this the branches grew rapidly. So we chose the final Hierarchical(HC) model with 3 clusters.

HC_Clusters	Avg_Credit_Limit	Total_Credit_Cards	Total_visits_bank	Total_visits_online	Total_calls_made	K_means_segments	count_in_each_segments_HC_cluster
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

Figure 14: Cluster profile from Hierarchical clustering. No. of clusters = 3

Figure 14 shows the cluster profile of 3 clusters obtained from the final HC model. It shows the mean values of each field corresponding to all three clusters. Figure 15 shows box plot for all the numerical variables in each cluster categories.

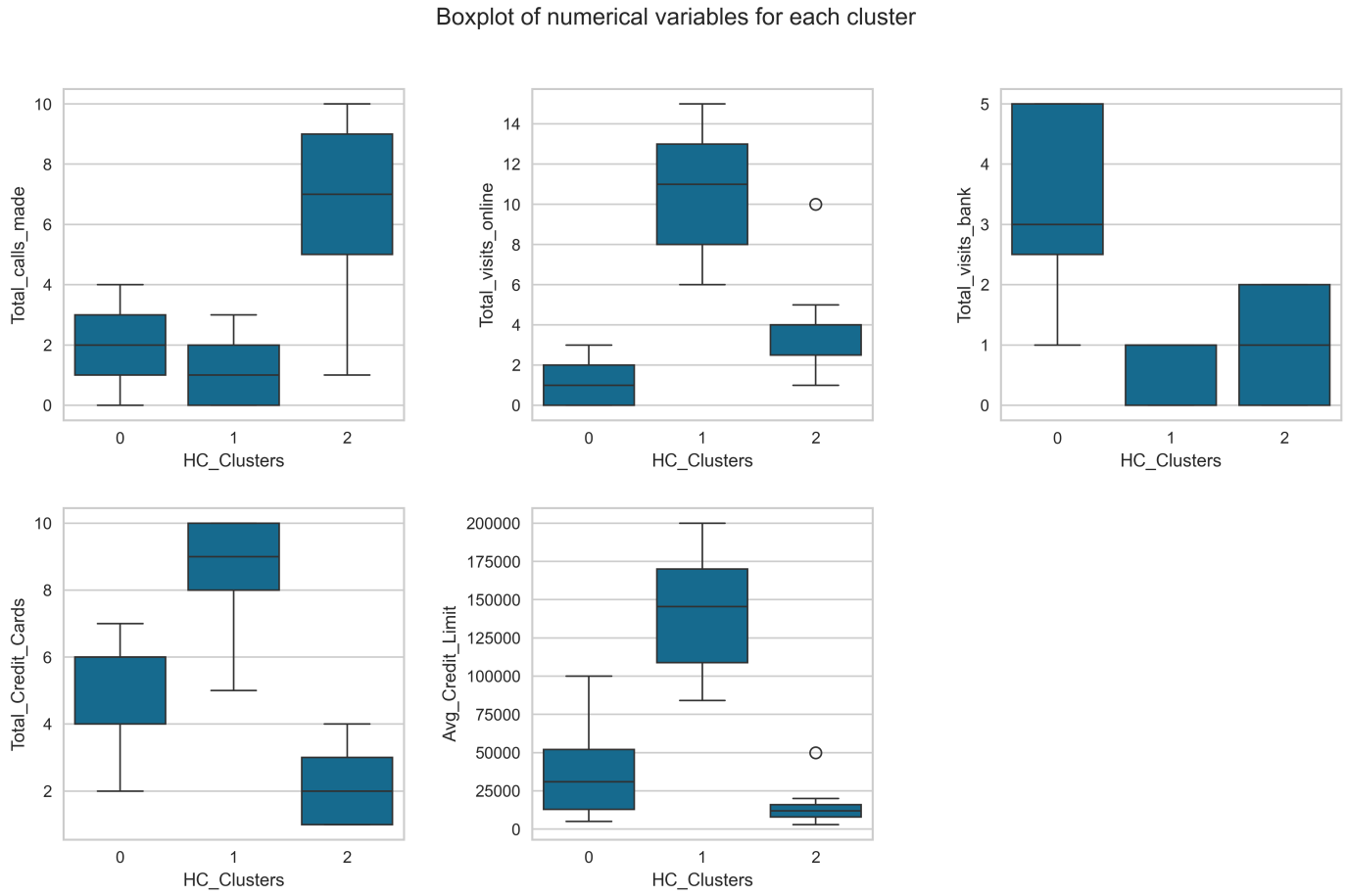


Figure 15: Box plot of different clusters obtained from Hierarchical clustering.

From these two figures, 14 and 15 a summary for each cluster properties is generated.

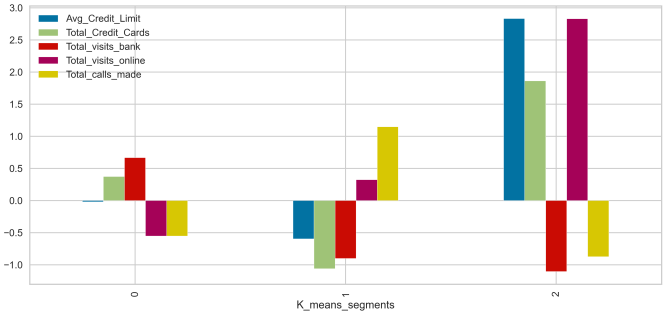
- Cluster 0:
  - There are 387 customers in this segment out of 660.
  - 75% of customers in this group made less than 3 calls.
  - At most three online visits are made from these customers.
  - These customers have relatively higher number of in person Bank visits than other two.
  - 75 % have 6 or less number of credit cards.
  - Median credit limit is around 30,000.
- Cluster 1:
  - There are only 50 customers in this segment out of 660.
  - 75 % of them made less than two phone calls.
  - They made relatively a very high number of online visits.
  - These customers don't visit to bank in person. At most they made 1 visit personally.
  - They hold higher number of credit cards. Up to 10.
  - Credit limit is also very high reaching up to 200,000.
- Cluster 2:
  - There are 223 customers in this segment.

- This segment have made higher number of phone calls. At least half of them made more than 6 calls.
- 75 % of them made less than 4 online visits.
- All of them had at most two visits to the bank in person.
- They have atmost 4 credit cards.
- Median credit limit is very less around 12,000.

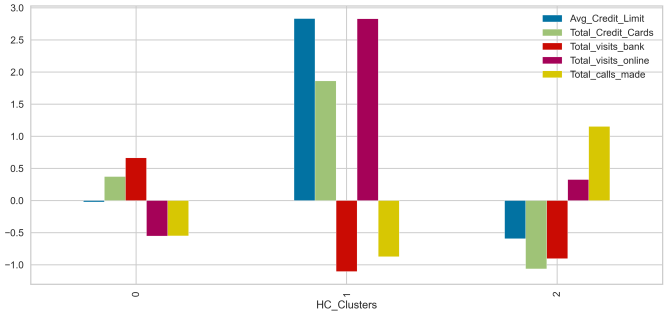
6 Comparison between two techniques

The results from both clustering techniques KMeans and Hierarchical(HC) are same both qualitatively and quantitatively. Both of them gave 3 clusters with almost exact identity/properties, only cluster 1 and 2 were interchanged. This can be observed from Figure 16. The bar plots are identical with cluster 1 and 2 interchanged. Also this is evident from cluster visualization with 2 features in figure 17, Only colors of cluster 1 and 2 are interchanged. The number of customers were also almost same for each segments in both methods.

In terms of time complexity HC took significantly higher time than KMeans since HC is connectivity based clustering. It grows as O(n*n).

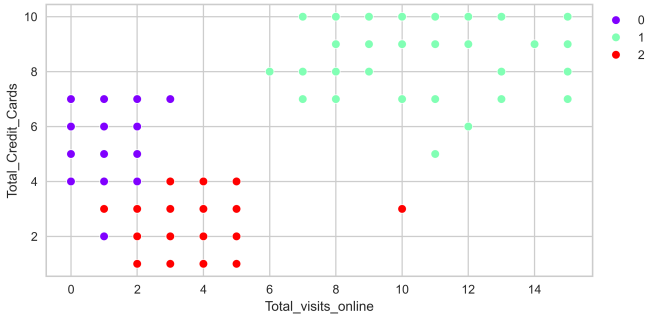


(a) Bar plot of means of all the fields across 3 KMeans segment.

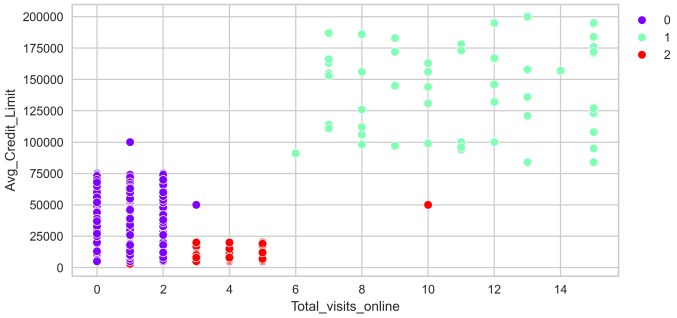


(b) Bar plot of means of all the fields across 3 HC segment.

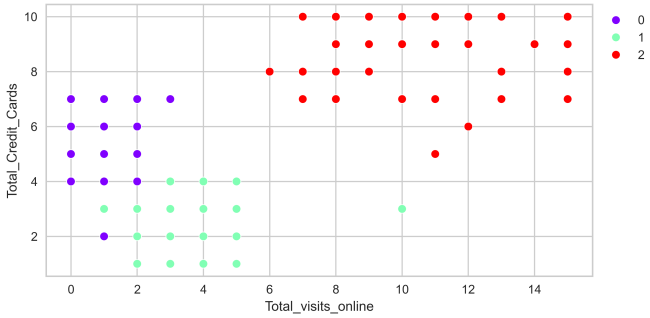
Figure 16: Cluster Method Comparison.



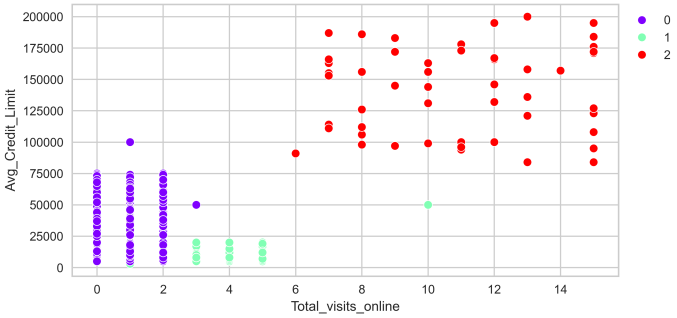
(a) Scatterplot between total Cards and online visits with cluster annoted from HC.



(b) Scatterplot between Credit limit and online visits with cluster annoted from HC.



(c) Scatterplot between total Cards and online visits with cluster annoted from Kmeans



(d) Scatterplot between Credit limit and online visits with cluster annoted from Kmeans.

Figure 17: Cluster Visualization.

7 Actionable insights and Recommendations

7.1 Actionable insights

- There are three customers segment for the credit service at AllLife Bank. Let's call them A, B and C.
- Segment A comprise more than 50 % of customers. They are more likely to visit Bank in person and hold up to 6 or 7 credit cards. Their median credit limit is round 30,000.
- Segemt B consists of 7.5 % of total customers. They make very less phone calls or visit in person. They make higher activity in online mode. Hold higher number of cards and credit limits are usually very high reaching up to 200,000.
- Segment C consists of around 34 % of total customers. They make higher number of phone calls for the services. They have at most 4 credit cards in possession and median credit limit is very less around 12,000.

7.2 Business Recommendations

- The mode of communication for their queries is very distinct in three customer segments we found. Customer service should be fine tailored to adapt to these communication modes (phones, in person and online).
- Segment A is the largest. Yet most of them visit bank in person. If we could improve online mode or encourage them to use online mode, we could profit from this segment as higher usage of online mode is correlated with more credit cards.
- Section C has 34 % of total yet they have comparatively few credit cards and limits. Incentives could be offered to customers in these segment to promote credit service. Since they prefer phone calls for query, customer care department should be trained to understand this segment better during calls and could also make some promotions towards call end.
- Segment B only makes 7.5 % yet they are biggest users in terms of amount. Since they prefer online modes we need to make sure that their user experience is seamless during website or app visit. They have usually high number of credit cards. We need to make sure that apps or websites make access to their card portfolios easily.
- Robust Feed backs mechanism should be developed in all the above modes to address all the three segments and understand if they were satisfied in the customer support.
- The three cluster could be used to identify new incoming customers and serve them accordingly. Simply record their prefer mode of communication, credit limit and number of cards they start to acquire.

End of Report
Submitted by : Haraprasad Dhal