Clustering: Individual Assignment

**Problem 1:**

**A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.**

1. **Read the data and do exploratory data analysis. Describe the data briefly.**
2. **Do you think scaling is necessary for clustering in this case? Justify**
3. **Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them**
4. **Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score.**
5. **Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.**

**Data** **Dictionary** **for** **Market** **Segmentation:**

* spending: Amount spent by the customer per month (in 1000s)
* advance payments: Amount paid by the customer in advance by cash (in 100s)
* probability of full payment: Probability of payment done in full by the customer to the bank
* current balance: Balance amount left in the account to make purchases (in 1000s)
* credit limit: Limit of the amount in credit card (10000s)
* min payment amt: minimum paid by the customer while making payments for purchases made monthly (in 100s)
* max spent in single shopping: Maximum amount spent in one purchase (in 1000s)

**1) Read the data and do exploratory data analysis. Describe the data briefly**

1. Loading the data set- We will be loading the” bank\_marketing\_part1\_Data.csv” file using pandas library in python. For this, we will be using read\_csv file.

2. The head function will tell you the top head records in the data set. By default, python shows you only the top 5 records, but we will check top 10 records.

3. The tail function will tell you the top tail records in the data set. By default, python shows you only the top 5 records, but we will check top 10 records for the totals/subtotals if any. The bank\_marketing\_part1\_Data dataset doesn’t contain any total/subtotals.

4. info() is used to check the Information about the data and the datatypes of each respective attribute.

5. The shape attribute tells us a number of observations and variables we have in the data set. It is used to check the dimension of data. The bank\_marketing\_part1\_Data data set has 210 observations and 7 variables in the data set.

Looking at the data in the head function and in info, we come to know that the all the variables comprise of float which doesn’t requires conversion. Further there are no null values in dataset and the count is 210 for all the variables.

6. df.isnull().sum() is used to check the Null values. It was identified that there are Zero Null Values in above dataset.

7. duplicate = df[df.duplicated()] is used to check the duplicated values. It was identified that there are Zero Duplicate Values in above dataset.

8. The described method will help to see how data has been spread for numerical values. We can clearly see the minimum value, mean values, different percentile values, and maximum values for the bank\_marketing\_part1\_Data data set.

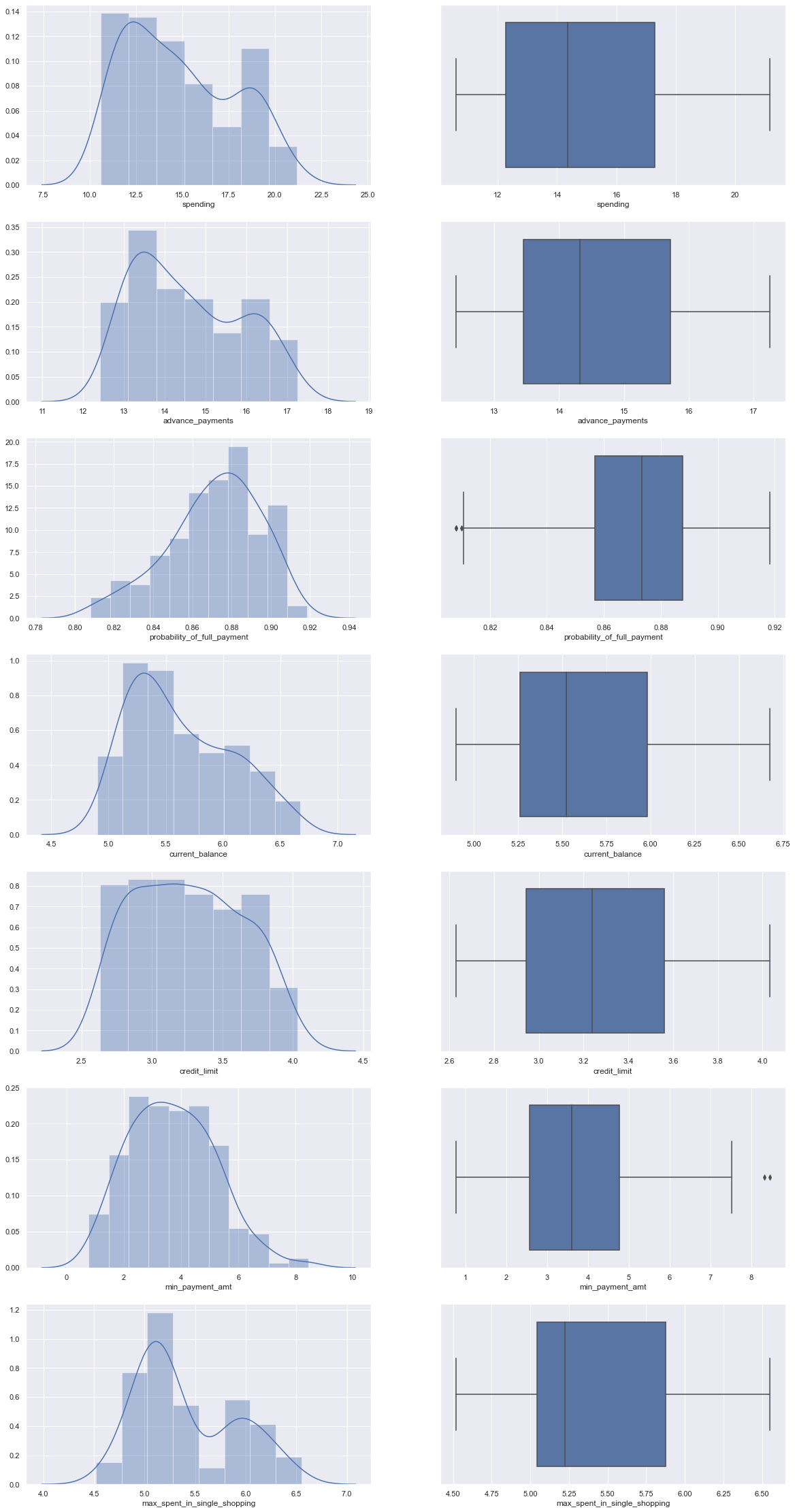


7. from pylab import rcParams Check whether the variables are normally distributed or not.

**A picture containing text, scoreboard, cabinet, clipart

Description automatically generated**

**Visual Representation of data by Dist. plot & Boxplot.**

****

**Insights -**

* The distribution for Spending Variable seems to be not Normal. The distribution tends to be right skewed.
* The distribution for Advance Payments also Variable seems to be not Normal. The distribution tends to be right skewed.
* The distribution for Probability of full payment also Variable seems to be not Normal. The distribution tends to be left skewed.
* The distribution for Current Balance also Variable seems to be not Normal. The distribution tends to be Right skewed.
* The distribution for Credit limit Variable seems to be Normal. The distribution tends to be Slight Right skewed.
* The distribution for Min Payment Amount Variable seems to be Normal. The distribution tends to be Slight Right skewed.
* The distribution for Max spent in single shopping Variable seems to be not Normal. The distribution tends to be Right skewed.

**2) Do you think scaling is necessary for clustering in this case? Justify**

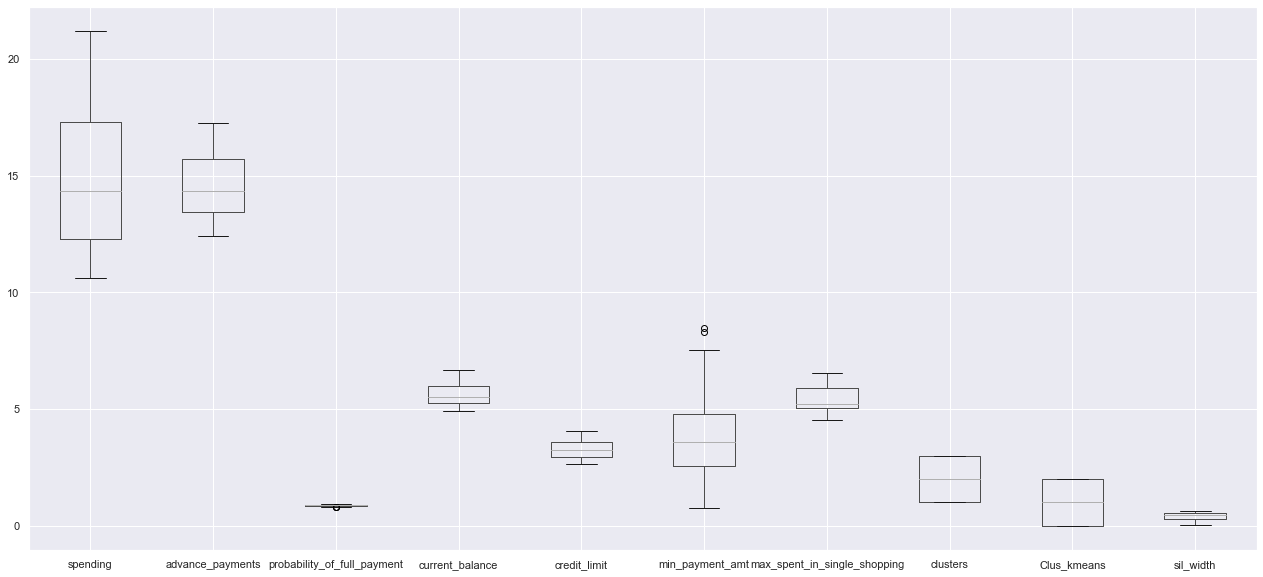
**Inference** - Data standardization is about making sure that data is internally consistent; that is, each data type has the same content and format.

In statistics, standardization is the process of putting different variables on the same scale. This process allows you to compare scores between different types of variables. Typically, to standardize variables, you calculate the mean and standard deviation for a variable.

Different of types of Scaling are

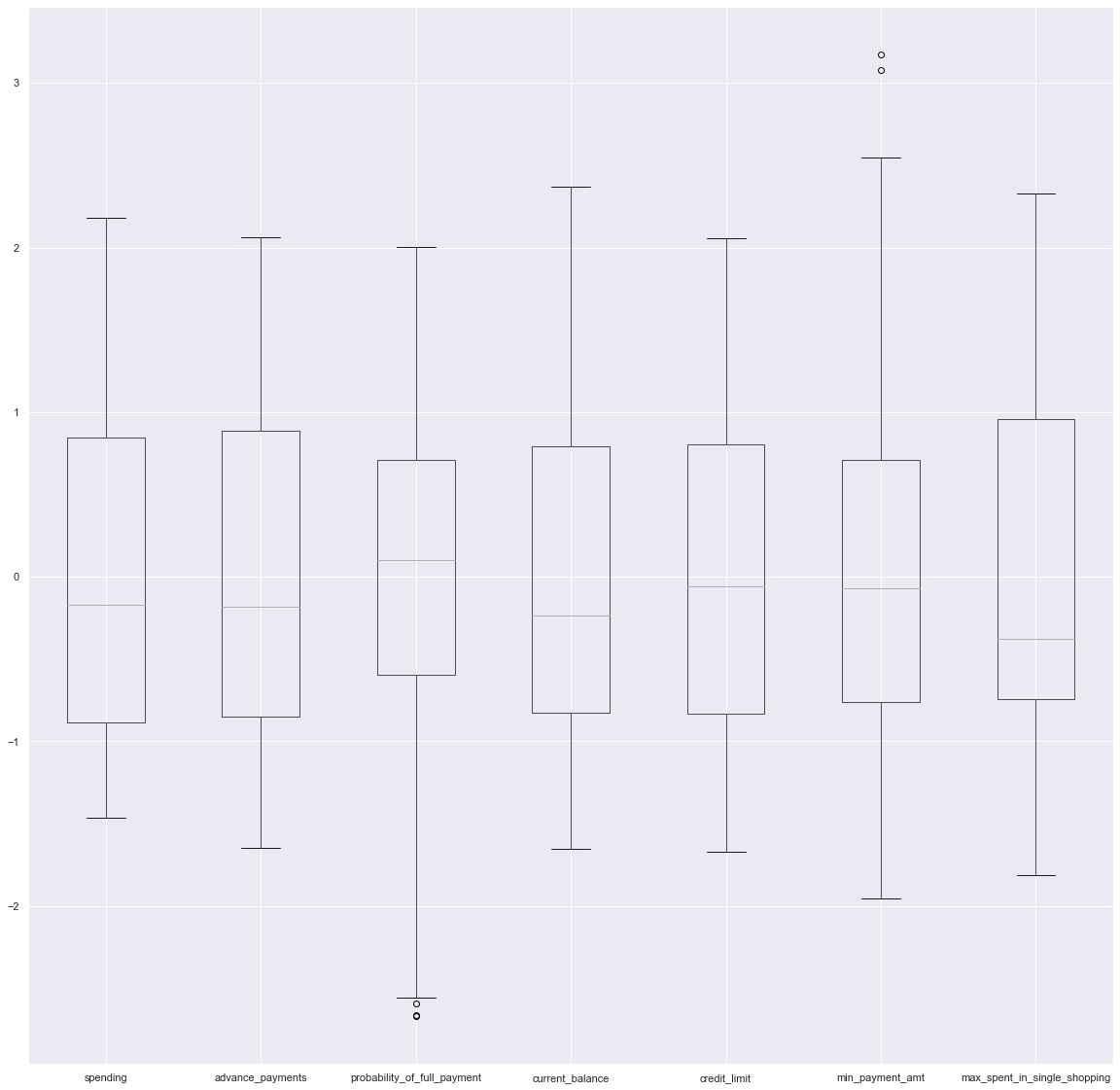
* Z scaling - Z-score is a variation of scaling that represents the number of standard deviations away from the mean. You would use z-score to ensure your feature distributions have mean = 0 and std = 1.
* Min-max scaling - min-max scaling or min-max normalization, is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data.

Before scaling the data looks like.



**Insights** - Post viewing the data it was identified that Variables like Spending, Current balance & max spent in single shopping are in 1000’s. Whereas the Variables like Advance payments & Min payment amt & are in 100’s, Credit limit Variable is in 10000’s. Since all the variables are not on the same scale. We need to Scale the data.

Scaled data looks like.



* 1. **Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them**

**Inference** - Hierarchical clustering method is based on hierarchy representation of clusters where parent cluster node is connected to further to child cluster node. Anode represents collection of data points to one cluster. It is further divided into two types:

* **Agglomerative Clustering** – The agglomerative clustering is the most popular and common hierarchical clustering also known as Agglomerative Nesting (AGNES). The methods start by considering each data point as a single cluster. In the next step the singleton clusters are merged into a bog cluster based on the similarity between them. The procedure is repeated until all the datapoints are merged into one big cluster. The procedure can be represented as hierarchy/tree of clusters. Divisive: This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

**Note**-It should be noted that agglomerative clustering is good choice to identify small sized clusters whereas divisive clustering is more effective in case of big size clusters

* **Divisive Clustering** – The divisive clustering works completely opposite to agglomerative clustering and also known as Divisive Analysis (DIANA). The method starts from one big cluster considering all data points within it. In the next the big cluster is divided into the most heterogeneous two clusters. The procedure is repeated until each data point is in its own cluster.

**Dendrogram** - A dendrogram is a pictorial way to visualize hierarchical clustering. It is mainly used to show the outcome of hierarchical clustering a tree like diagram that records the sequences of merges and splits.

**Steps to Perform Hierarchical Clustering**

We merge the most similar points or clusters in hierarchical clustering – we know this. Now the question is – how do we decide which points are similar and which are not? It’s one of the most important questions in clustering!

Here’s one way to calculate similarity – Take the distance between the centroids of these clusters. The points having the least distance are referred to as similar points and we can merge them. We can refer to this as a distance-based algorithm as well (since we are calculating the distances between the clusters).

Measuring similarity -Distances

* Eucledian distance
* Manhattan distance
* Chebyshev distance
* Minkowski Distance

Hierarchical clustering – Distance between clusters Linkage types

• Single linkage – Distance between two clusters is defined as the shortest distance between two points in each cluster

• Complete linkage – Distance between two clusters is defined as the longest distance between two points in each cluster.

• Average linkage - Distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.

• Centroid linkage - Based on centroid distance. clusters are represented by their mean values for each variable, which forms a vector of means. Distance between 2 clusters is distance between the 2 vectors

• Ward’s Method Proprietary - Similar to group average and centroid distance. joins records and clusters together progressively to produce larger and larger clusters but operates slightly differently from the general approach.

#### Method 1: Performing Hierarchical Clustering with the 'scipy' package

* First step is looking at the head of the dataset and checking which all Variables need to be considered for the model. From data set its clear that there are no any categorical data which we need to exclude/ drop. So, we will go ahead will all the Seven variables.



* Next step is to check whether scaling is required by seeing the dataset. Post viewing the data it was identified that Variables like Spending, Current balance & max spent in single shopping are in 1000’s. Whereas the Variables like Advance payments & Min payment amt & are in 100’s, Credit limit Variable is in 10000’s. Since all the variables are not on the same scale. We need to Scale the data.



* Post Scaling the data Will import the dendrogram & linkage from scipy.cluster.hierarchy.
* Let us now try to cluster the data with the Euclidean distance and Ward's method for linkage.
* Next is will plot the dendrogram and will check for number of clusters.

Dendrogram for above data set looks like.

Chart, histogram

Description automatically generated

* Now that we have visualized the number of clusters, we need to cluster the data according to their similarity metrics.
* Now Let’s Import fcluster from scipy.cluster.hierarchy. From fcluster library helps us to extract the cluster numbers by looking at the dendrogram.
* By choosing a cut off distance on the y-axis, a set of clusters is created.
* From the dendrogram, we see that 3 clusters are optimum. Thus, we are going to form 3 clusters based on the 'Distance' criterion (i.e. Distance on Y axis is 9.7) is in the fcluster package.

**4) Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.**

**Inference** - K-means clustering is a unsupervised learning algorithm whose goal is to find groups or assign the data points to clusters on the basis of their similarity. Which means the points in same cluster are similar to each other and in different clusters are dissimilar with each other. It was developed by researcher named James Macqueen in 1967.Here K means the number of clusters.

Working of K-means Clustering - The working of k-means clustering can be summarized as :

**Step 1**-Initialize the K random centroids or k points. (There canbe two strategy for it.)

1. Pick random data points and consider those as starting points.
2. Choose K random values for each particular variable.

**Step 2** - For each data point calculate the distance of it from randomly chosen K centroid 𝐶'and assign each point to minimum distance cluster.

**Step 3** - Update the centroid by using newly assigned data points to the cluster by calculating the average of data points.

**Step 4** - Repeat the above process for a given no. of iterations or until the centroid allocation no longer changes.

There are some other techniques from which can be used to find the approximate or optimal value of k.

**Elbow method** - It is most popular and well-known method to find the optimal no. of clusters or the value of k in the process of clustering. This method is based of plotting the value of cost function against different values of k. As the number of clusters (k) increase lesser number of points fall within clusters or around the centroids. Hence the average distortion decreases with the increase of number of clusters. The point where the distortion declines most is said to be the elbow point and define the optimal number of clusters for dataset.

**Silhouette Method** - Silhouette is a different method to determine optimal number of clusters for given dataset. It defines as a coefficient of measure of how similar an observation to its own cluster compared to that of other clusters. The range of silhouette coefficient varies between -1 to 1.1 value indicate that an observation is far from its neighbouring cluster and close to its own whereas-1 denotes that an observation is close to neighbouring cluster than its own cluster. The 0 value indicate the presence of observation on boundary of two clusters

**Let’s run the K-Means Clustering**.

* Since we have already scaled the data above, let us go ahead and perform the K-Means Clustering.
* Since we do not know the value of 'K' i.e. the optimum number of clusters we will start with 2 clusters and check the Within Sum of Squares (WSS).
* The 'inertia' gives us the Within Sum of Squares (WSS) for the number of clusters defined in the KMeans function inside the 'sklearn' library.
* import KMeans from sklearn.cluster.
* Let us now check the WSS for 2 clusters. The WSS for 659.1717544870407.
* Let us now check the WSS for 3 clusters. The WSS for 430.65897315130053.
* Now, we see that the WSS is decreasing. But it is very cumbersome to manually compute for each value. So, we are going to pass the KMeans function through a loop to automate this process of manually calculating the 'inertia'.
* Let us define an empty list to begin the process of automating the calculation of 'inertia'.
* And calculate the WSS for the range for 1 to 11.
* Now, let us print 'wss' and check the values.
* The wss value for 2 clusters is 1470.0
* The wss value for 3 clusters is 659.1717544870407
* The wss value for 4 clusters is 430.65897315130053
* The wss value for 5 clusters is 371.65314399951626
* The wss value for 6 clusters is 326.71692151404477
* The wss value for 7 clusters is 290.0504842894782
* The wss value for 8 clusters is 263.60586850020917
* In K-Means Clustering we have to choose the value of 'K' very accurately by looking at the Within Sum of Squares (WSS). Basis the wss scores from above table we are good to go with 3 clusters.
* There are some other techniques from which can be used to find the approximate or optimal value of k.

**Elbow method**

In cluster analysis, the **elbow** method is a heuristic used in determining the number of clusters in a data set. The method consists of **plotting** the explained variation as a function of the number of clusters and picking the **elbow** of the curve as the number of clusters to use.

Let’s plot the elbow curve.

A picture containing shape

Description automatically generated

The ideal 'WSS' plot has to have a sharp elbow like structure. The number of clusters corresponding to that elbow-like graph is considered to be the most optimum.

As it is clear from above figure, the distortion declines most at 3. Hence the optimal value of k will be 3 for performing the clustering. In other words, the plot looks as an arm with anelbow at k = 3. Having said that, here we will go for 3 clusters.

Let us now store the values of the clusters into a variable and we will attach the particular variable to the data set.

* **Silhouette Method** - Silhouette is a different method to determine optimal number of clusters for given dataset. It defines as a coefficient of measure of how similar an observation to its own cluster compared to that of other clusters. The range of silhouette coefficient varies between -1 to 1.1 value indicate that an observation is far from its neighbouring cluster and close to its own whereas-1 denotes that an observation is close to neighbouring cluster than its own cluster. The 0 value indicate the presence of observation on boundary of two clusters.

Import silhouette\_score from sklearn.metrics

Let’s calculate the silhouette scores for range of 2 to 11.

* silhouette = 0.46577247686580914 for i= 2
* silhouette = 0.4007270552751299 for i= 3
* silhouette = 0.3291966792017613 for i= 4
* silhouette = 0.2865045701398138 for i= 5
* silhouette = 0.29127768970444345 for i= 6
* silhouette = 0.27393907672148887 for i= 7
* silhouette = 0.25022538653640086 for i= 8
* silhouette = 0.24491366410664225 for i= 9
* silhouette = 0.2502746699759954 for i= 10

silhouette Plot Basis on above scores.

Chart, line chart

Description automatically generated

From above Silhouette Method it was identified that 0.46577247686580914 is highest Silhouette score for K=2. Whereas as per Elbow method the distortion declines most at 3. Hence, we will go ahead with 3 clusters.

* Let us now go ahead and attach these clusters with the original data frame and try to interpret it from a business perspective.
* We will try to profile the clusters with the mean of the spending on each category. This will give us an idea about the various clusters thus built.



* Let us check the frequency of the occurrence of the clusters for each individual cluster.

|  |  |
| --- | --- |
| 0 | 67 |
| 1 | 71 |
| 2 | 72 |

**5) Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.**

**Basis the Hierarchical Clustering & K means Clustering we are good to go with 3 Clusters.**



**Insights for different promotional strategies for different clusters in context to the business problem in-hand**

**High Spenders**

High spenders can be easily distinguished based on the high spending history. As they are highest spenders in the room the bank / Lending institution can offer luxurious and exclusive privileges and discounts with top dining, hotel and nightlife establishments. Other privileges like higher welcome miles bonuses, higher miles per dollar earn rates, special concierge services like airport limousine transfers, expedited immigration clearance, complimentary access to airport lounges and even complimentary hotel room stays as they are considered as high-income earners.

**Mid Spenders**

Basis data it was identified that in spite of being having the current balance equivalent to the High Spenders they have low credit limits. Financial institution can increase their Credit Limits and offer them offers based on giving them a target amount and if achieved given them Instant cash backs & Rewards on each penny spent. Also, they can give offers them EMI offers on purchase of online Electronics gadgets as it was also identified that They are the ones who do lower minimum payment.

**Low Spenders**

Low Spenders are those who come under Middle Class. Who can be classified as low average spenders basis on their specific needs. As the they use credit cards for their specific needs like groceries purchases, Petrol pumps, Booking Movie Tickets, Booking Air Tickets. Lending Institution can give they offers on groceries purchases, Petrol pumps, Booking Movie Tickets, Booking Air Tickets, Online food Ordering Apps like Zomato, Swiggy, Food Panda, Faaso's, etc. Also, in spite of having the credit limit same as Mid Spenders they are spending the lowest so giving ample offers for them would make sense. Also, data says that probability if full payment is 100%.