

Machine Learning Project Part1 - Clustering

By Ramprasad Mohan



Bank Customer Segmentation

Table of Contents

1	Project Objective.....	3
2	Data Dictionary	3
3	Step by step approach	3
4	Solution.....	5
4.1	Qn 1.....	4
4.2	Qn 2.....	7
4.3	Qn 3.....	7
4.4	Qn 4.....	13
4.5	Qn 5.....	18
5	Conclusion.....	23
6	Appendix A – Source Code.....	23

1 Project Objective

The project objective is to solve the business problem of a leading bank, who wants to develop a customer segmentation to give promotional offers to its customers. The bank has collected a sample that summarizes the activities of users during the past few months. We are given the task to identify the segments based on credit card usage.

Managerial Report is to be prepared as below.

2 Data Dictionary

Below are the variables in the seven variables in the data set and its explanation:

- 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 (in 100's)
- max spent in single shopping: Maximum amount spent in one purchase (in 1000s)

3 Step by Step approach

We shall follow step by step approach to arrive to the conclusion as follows:

- Exploratory Data Analysis
- Scaling/ Normalizing the data
- Build Hierarchical clustering model
- Interpret optimal number of clusters with dendrogram , Nbclust and other methods
- Calculate silhouette score
- Visualize the clusters
- Do profiling to get the aggregate of the clusters
- Determine optimal number of clusters for K-means using WSS & Silhouette scores
- Perform K-means clustering with the determined number of clusters
- Visualize the clusters
- Calculate the silhouette score, profile the clusters based on the aggregate

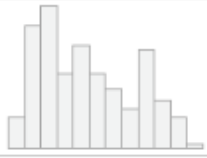
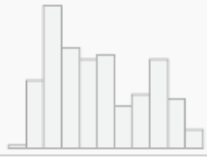
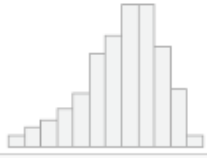
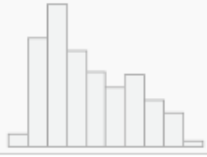
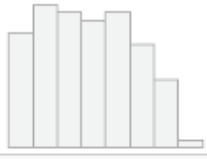
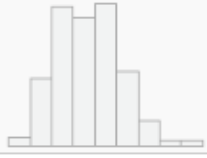
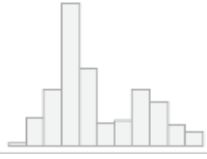
4 Solution

4.1 Qn 1 : Read the data and do exploratory data analysis. Describe the data briefly.

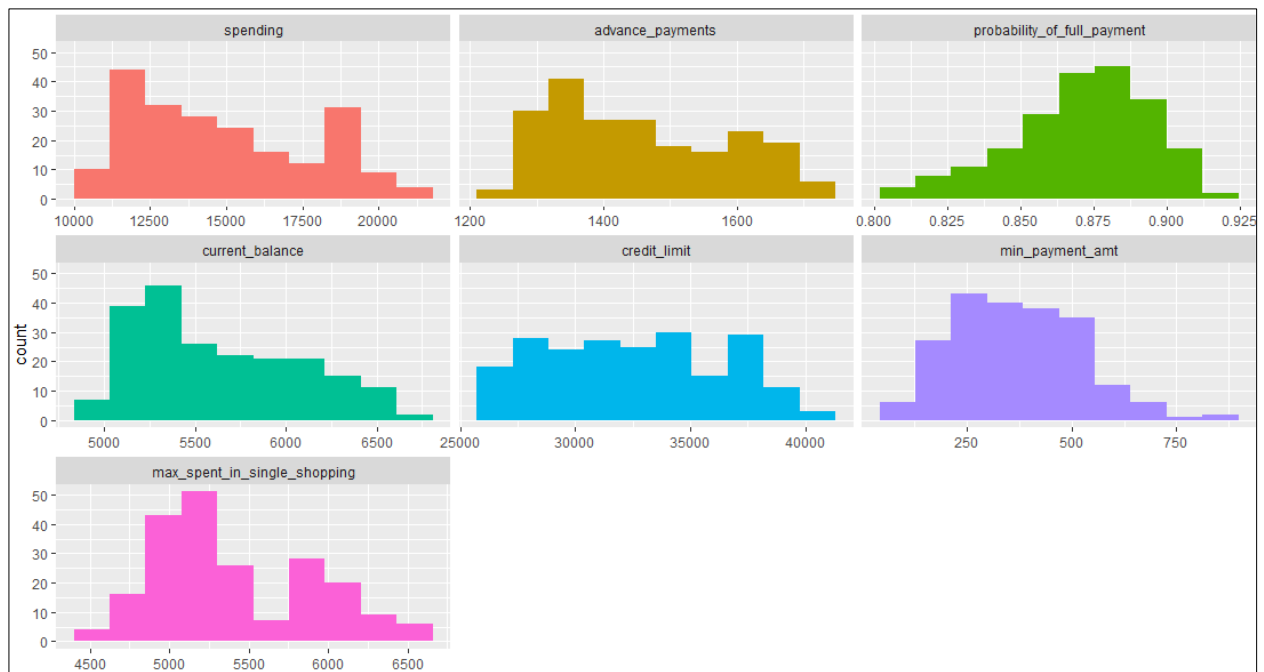
After importing the data, we are checking the top five rows of the bank data. The bank data has 210 observations and 7 variables (The variable names are available in Data dictionary section)

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
1	19940	1692	0.8752	6675	37630	325.2	6550
2	15990	1489	0.9064	5363	35820	333.6	5144
3	18950	1642	0.8829	6248	37550	336.8	6148
4	10830	1296	0.8099	5278	26410	518.2	5185
5	17990	1586	0.8992	5890	36940	206.8	5837
6	12700	1341	0.8874	5183	30910	845.6	5000

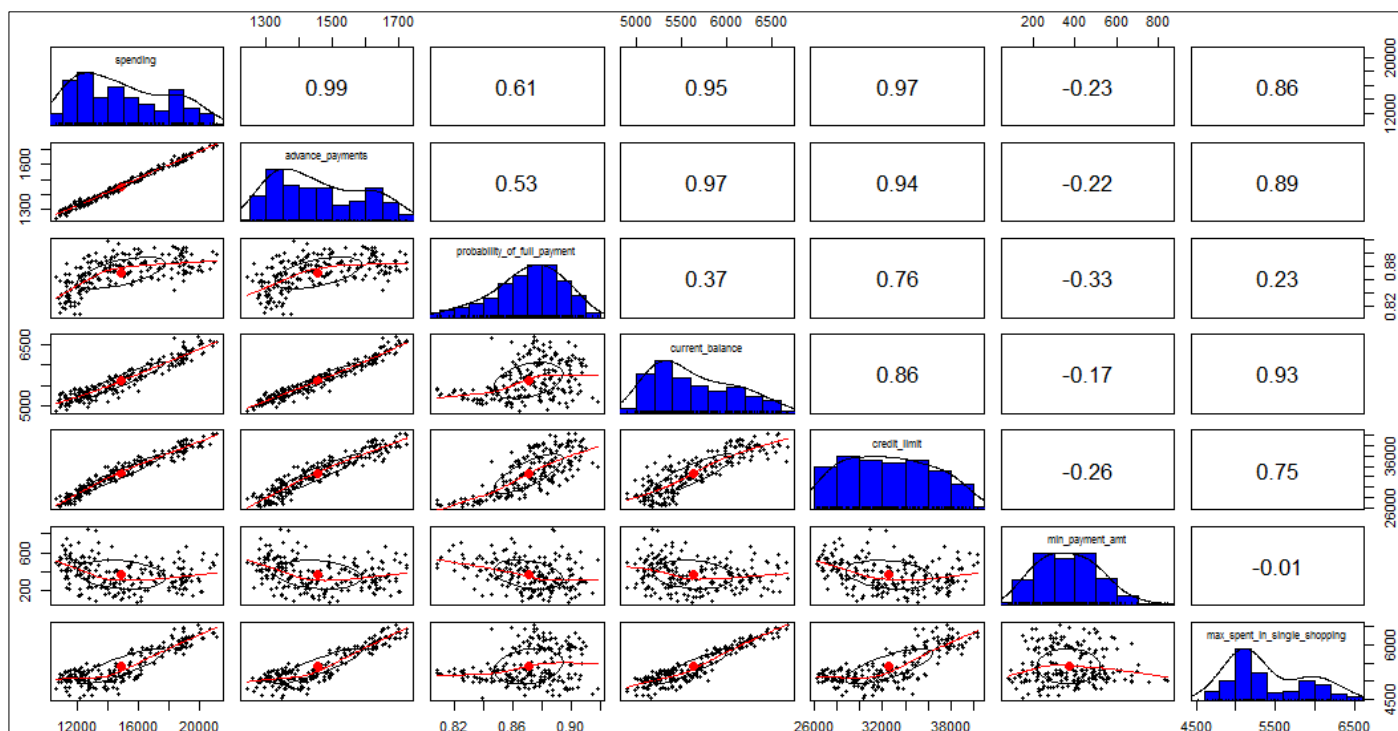
Using summarytools package, we can get the below report for our dataset. It provides the datatype, mean, standard deviation, minimum, maximum and other values for our dataset.

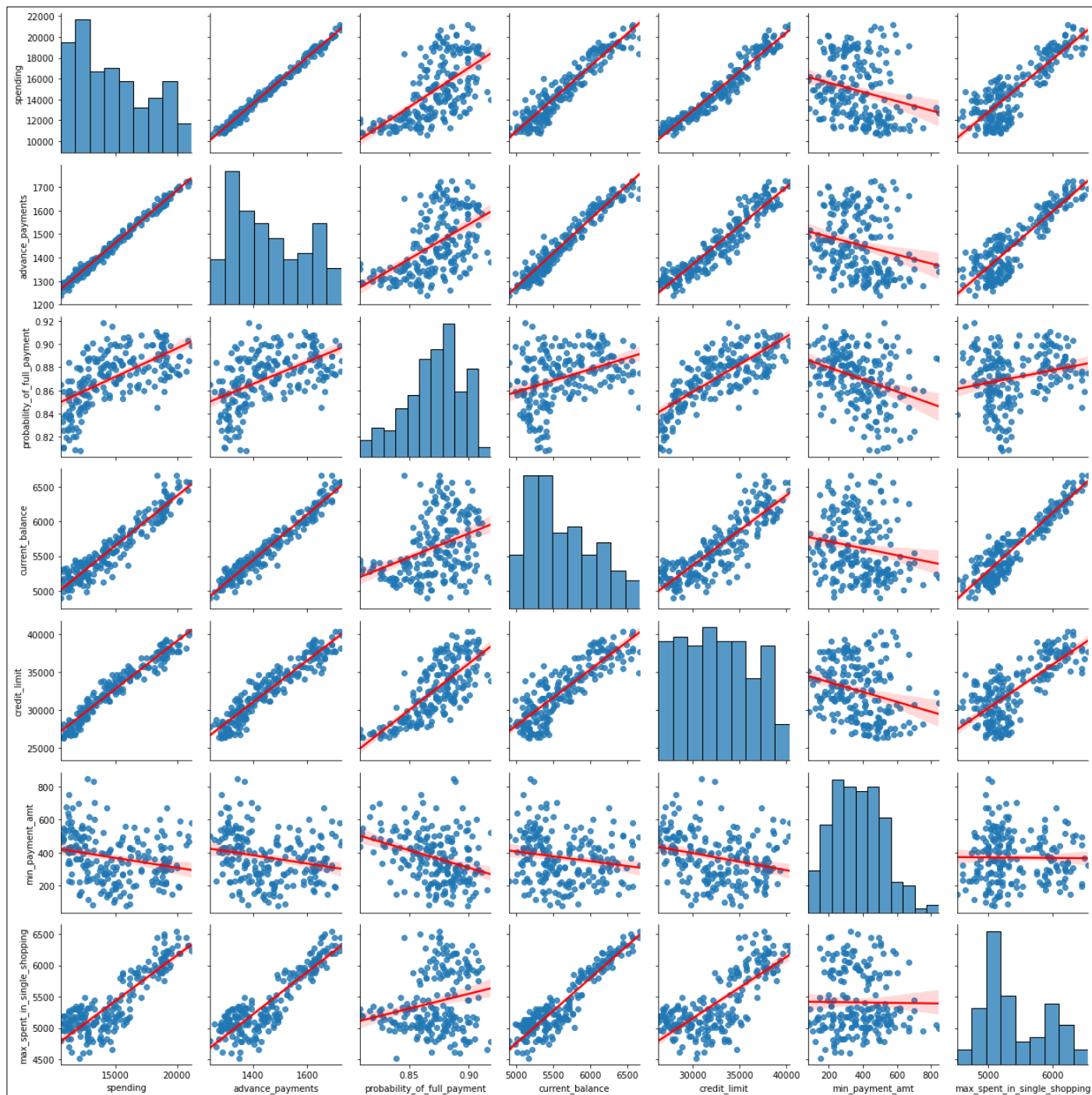
1	spending [numeric]	Mean (sd) : 14847.5 (2909.7) min < med < max: 10590 < 14355 < 21180 IQR (CV) : 5035 (0.2)	193 distinct values		210 (100%)	0 (0%)
2	advance_payments [numeric]	Mean (sd) : 1455.9 (130.6) min < med < max: 1241 < 1432 < 1725 IQR (CV) : 226.5 (0.1)	170 distinct values		210 (100%)	0 (0%)
3	probability_of_full_payment [numeric]	Mean (sd) : 0.9 (0) min < med < max: 0.8 < 0.9 < 0.9 IQR (CV) : 0 (0)	186 distinct values		210 (100%)	0 (0%)
4	current_balance [numeric]	Mean (sd) : 5628.5 (443.1) min < med < max: 4899 < 5523.5 < 6675 IQR (CV) : 717.5 (0.1)	188 distinct values		210 (100%)	0 (0%)
5	credit_limit [numeric]	Mean (sd) : 32586 (3777.1) min < med < max: 26300 < 32370 < 40330 IQR (CV) : 6177.5 (0.1)	184 distinct values		210 (100%)	0 (0%)
6	min_payment_amt [numeric]	Mean (sd) : 370 (150.4) min < med < max: 76.5 < 359.9 < 845.6 IQR (CV) : 220.7 (0.4)	207 distinct values		210 (100%)	0 (0%)
7	max_spent_in_single_shopping [numeric]	Mean (sd) : 5408.1 (491.5) min < med < max: 4519 < 5223 < 6550 IQR (CV) : 832 (0.1)	148 distinct values		210 (100%)	0 (0%)

The histograms for each of the continuous variables are plotted below using FunModelling package. Most of the variables are slightly left skewed while some are right skewed

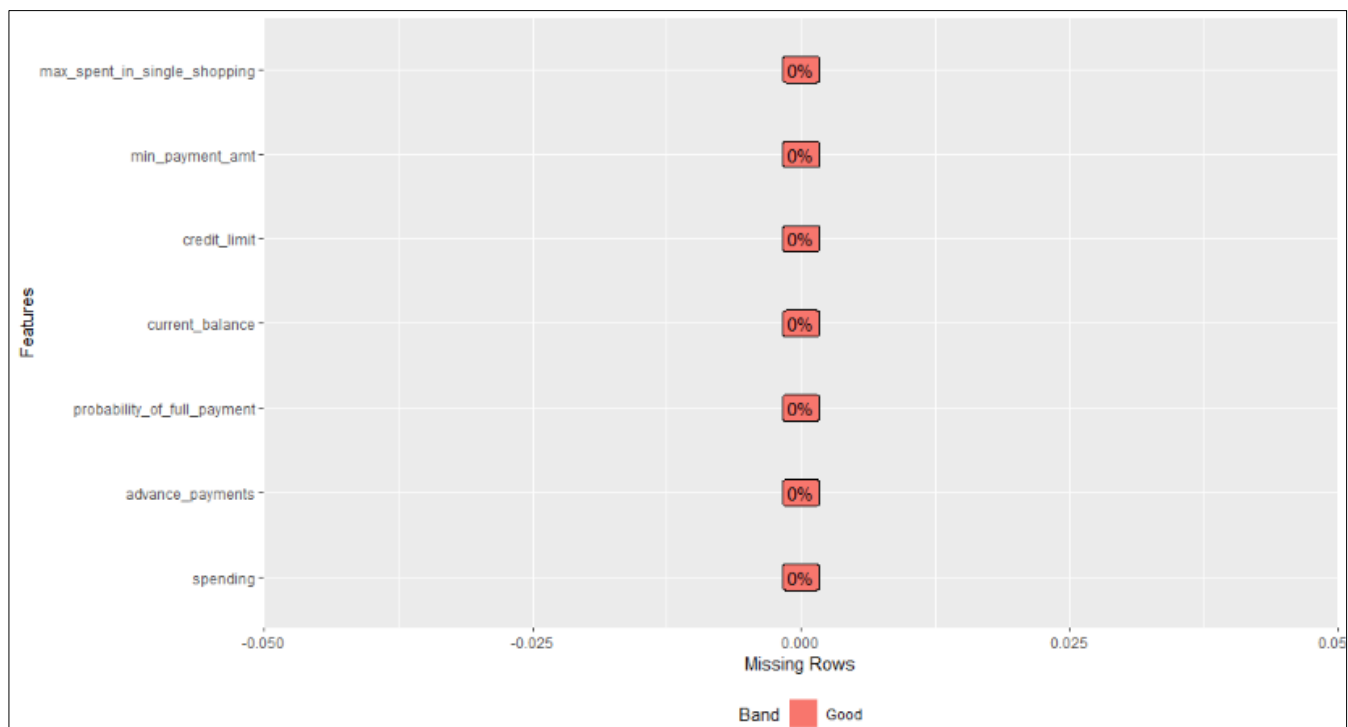


Below is the correlation plot for each combination and the density plot for the continuous variables done using `pairs.panels()` function. Few of the variables have strong correlation of 0.97, 0.95 and the same is also visualized in scatter plot. Spending, advance payment, probability, current balance, credit limit are strongly correlated for various combinations.



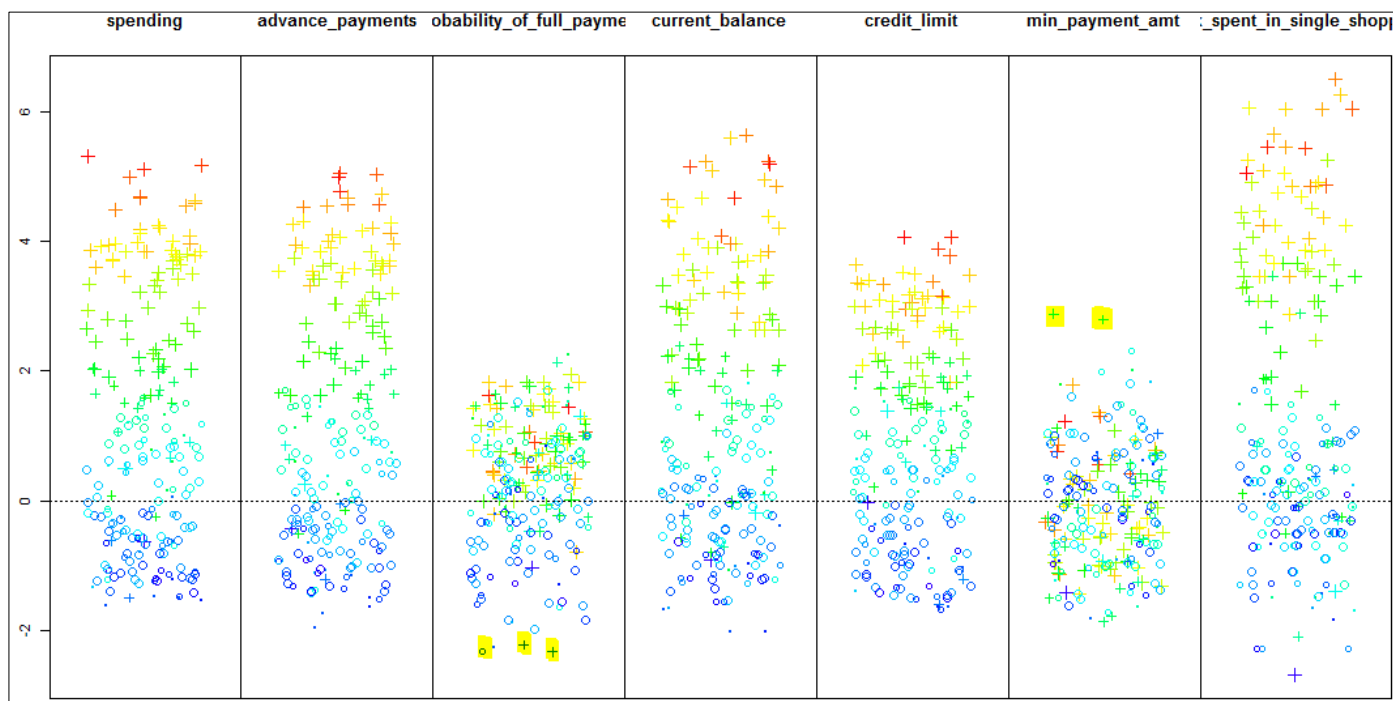


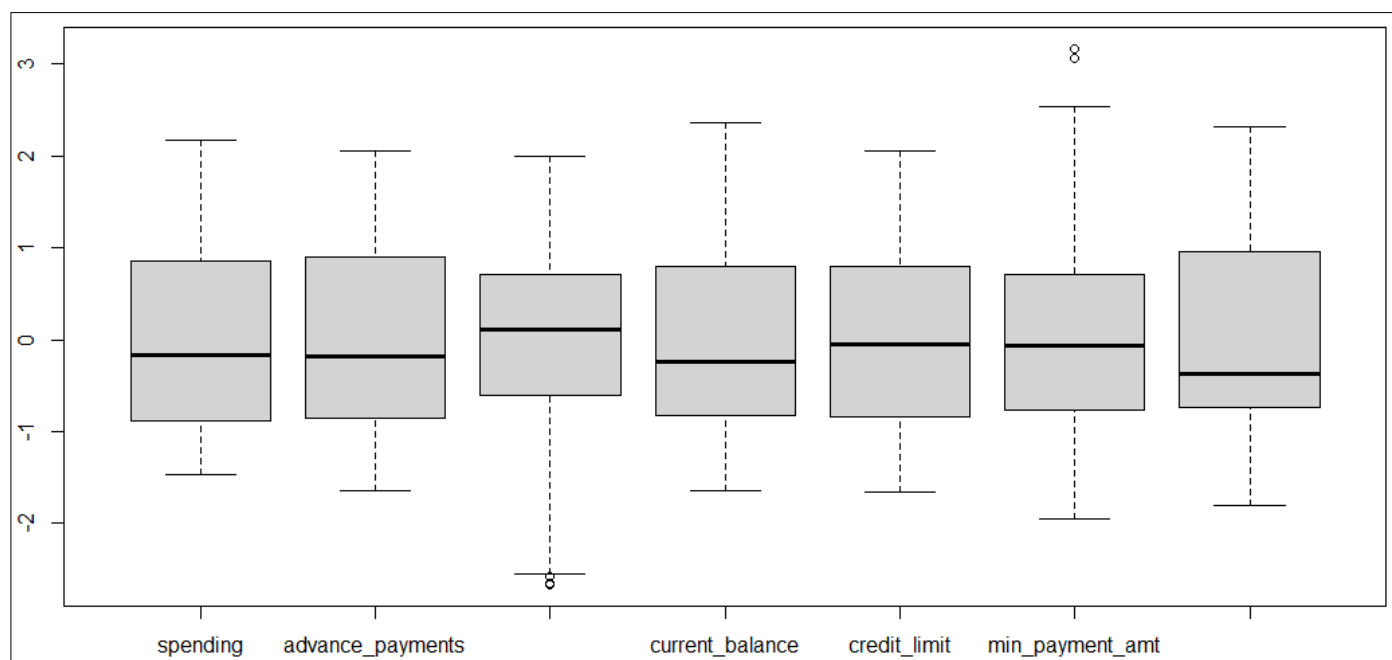
Next we check for missing values in the data set using the Dataexplorer package.



There are no missing values or null values in the dataset so our data is fine.

Checking for outliers: From the below plot we could see that the minimum payment amount has outlier values and rest of the variable looks fine with no or few outliers. Since it's not an extreme outlier which will affect our clustering model we are ignoring the outliers and will proceed with analysis including the outlier values.





4.2 Qn 2 : Do you think scaling is necessary for clustering in this case

For clustering algorithms, scaling data is a must to avoid models being prone to outlier and huge values. Clustering algorithms such as K-means require scaling before they are run on the clustering model. Since, clustering techniques use distance method to form the groups, it will be wise to scale the variables before calculating the distance. Unscaled data will impact the performance of all distance based model as it will give higher weightage to variables which have higher magnitude. One of the most common technique to do so is normalization where we calculate the mean and standard deviation of the variable.

Since our data variables are in different scales such as 1000's , 100's and in decimal's like probability values, we would perform normalization scaling below, is the output of the scaled data.

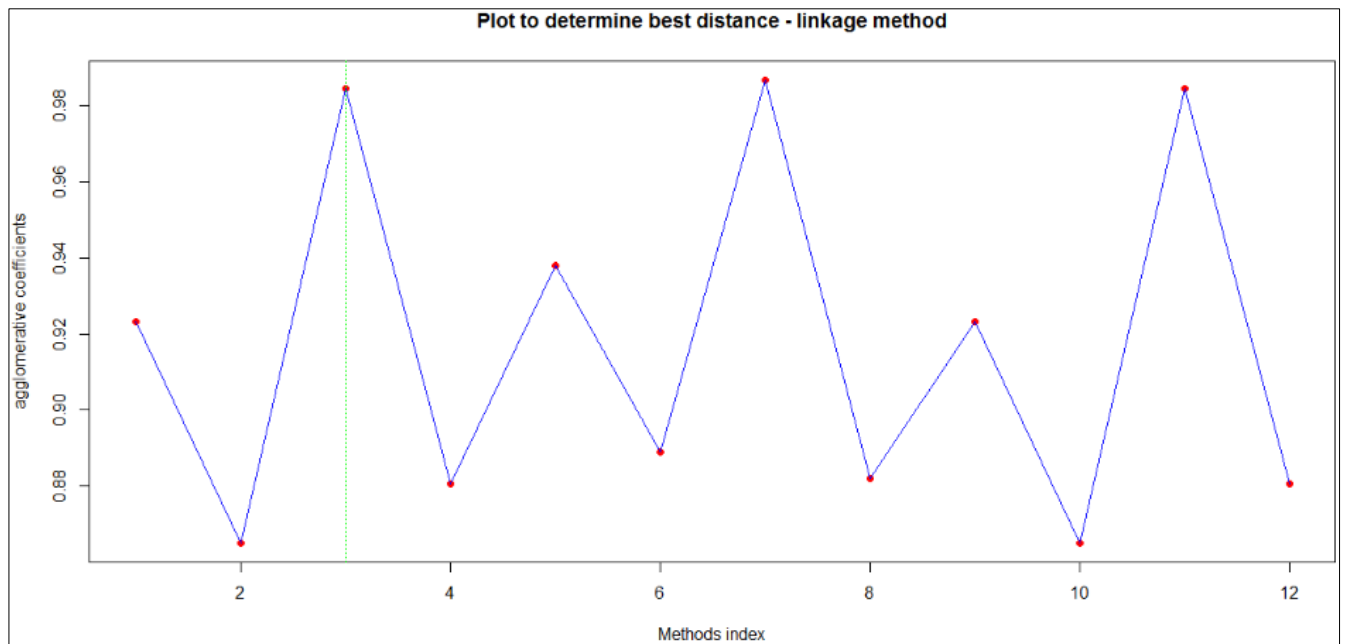
spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
1.7501726	1.8076485	0.1778050	2.3618888	1.3353877	-0.2980937	2.3234463
0.3926441	0.2532349	1.4981931	-0.5993122	0.8561898	-0.2422262	-0.5372979
1.4099313	1.4247880	0.5036700	1.3981443	1.3142077	-0.2209434	1.5055095
-1.3807350	-1.2246066	-2.5856995	-0.7911583	-1.6351103	0.9855289	-0.4538765
1.0800003	0.9959842	1.1934881	0.5901336	1.1527101	-1.0855596	0.8727275
-0.7380569	-0.8800322	0.6941106	-1.0055745	-0.4437341	3.1630318	-0.8302902

4.3 Qn 3 : Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them

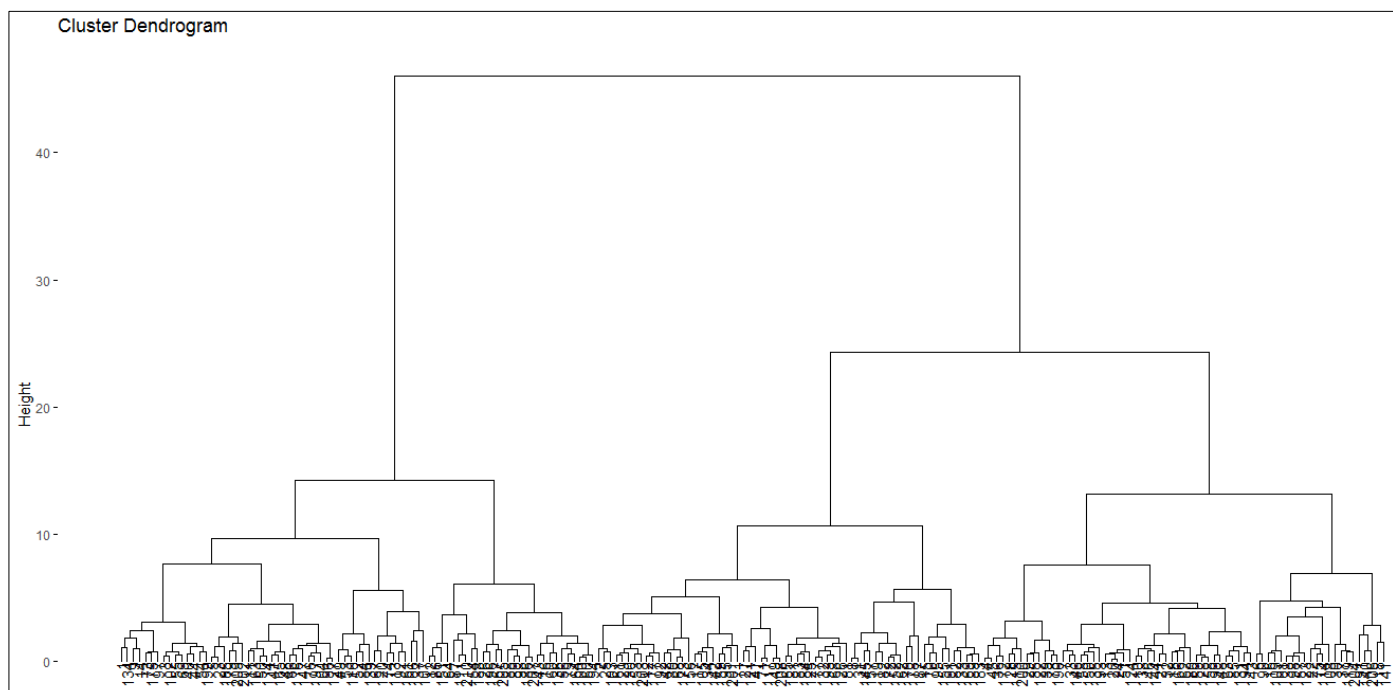
To Apply hierarchical clustering, we need to first find out the best distance method from commonly used such as "Euclidean", "manhattan", "minkowski" and the best linkage method such as "complete", "ward", "average", "centroid" to build the hierarchical clustering model based on the methods.

The method used is by using agnes() function from the cluster package and finding the agglomerative

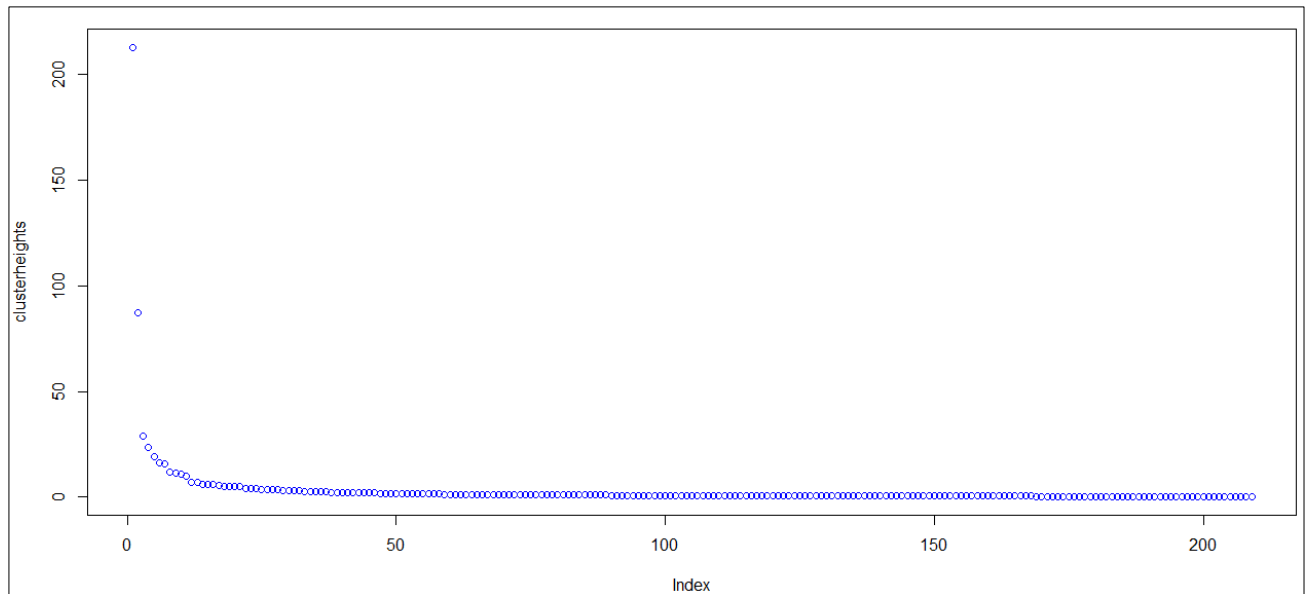
coefficients for various combinations of distance method and linkage method. The resulting coefficients are plotted below. We could see that the distance “Euclidean” and the linkage method “ward” combination has high agglomerative coefficients hence we are proceeding with this combination. (refer code for combinations).



Next we are building the hierarchical clustering model using cluster package with both `agnes()` and `hclust()` function and plotting the dendrogram output. Below is the obtained dendrogram:



Next step is to determine the optimal number of clusters from the dendrogram output. We would first try to interpret the optimal cluster using cluster height merging plot.



We could see that the cluster heights begin to disperse after 10 clusters. Though this plot does not provide us much information in determining the optimal number of clusters hence we are proceeding with a special method called Nbclust method to determine the optimal clusters.

Special Method : Nbclust Package

Nbclust provides 30 indices for determining the number of clusters and proposes to use the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods. It uses voting method to determine the best number of clusters from various combinations. This method is recommended only for small data sets as large datasets will take longer time to execute and produce output.

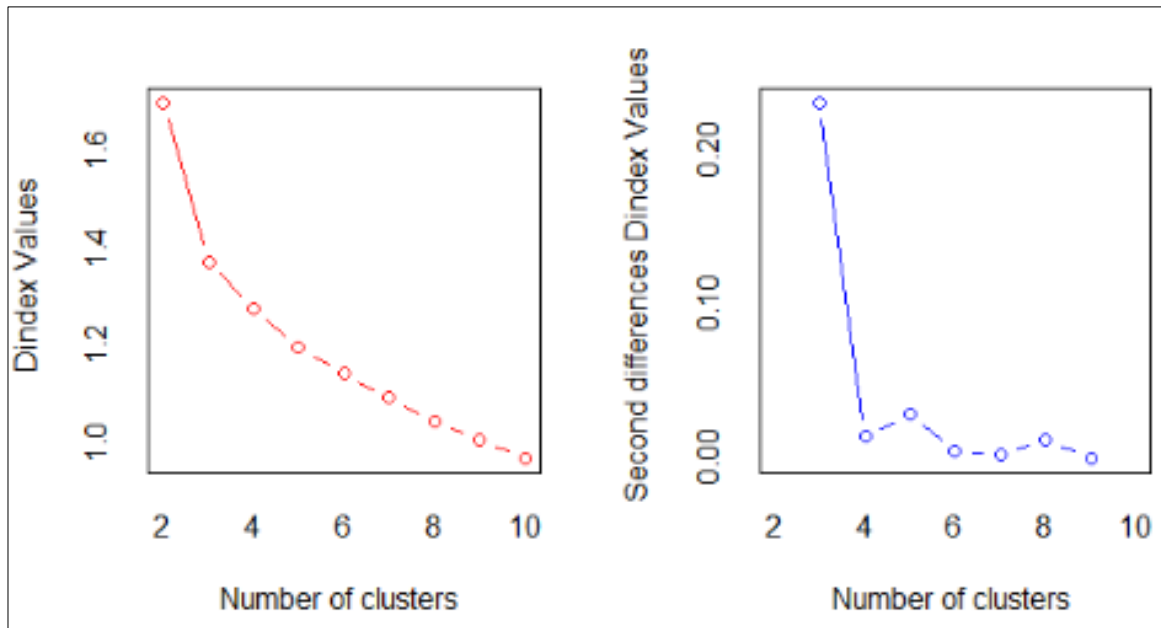
The Hubert index is a graphical method of determining the number of clusters.

In the plot of Hubert index, we seek a significant knee that corresponds to a significant increase of the value of the measure i.e the significant peak in Hubert index second differences plot.

The D index is a graphical method of determining the number of clusters.

In the plot of D index, we seek a significant knee (the significant peak in Dindex second differences plot) that corresponds to a significant increase of the value of the measure. Below is the output of the respective plots.

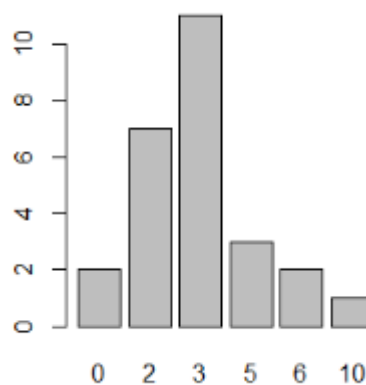
Below are the output graphs of both the plots and conclusion of the Nbclust voting method.



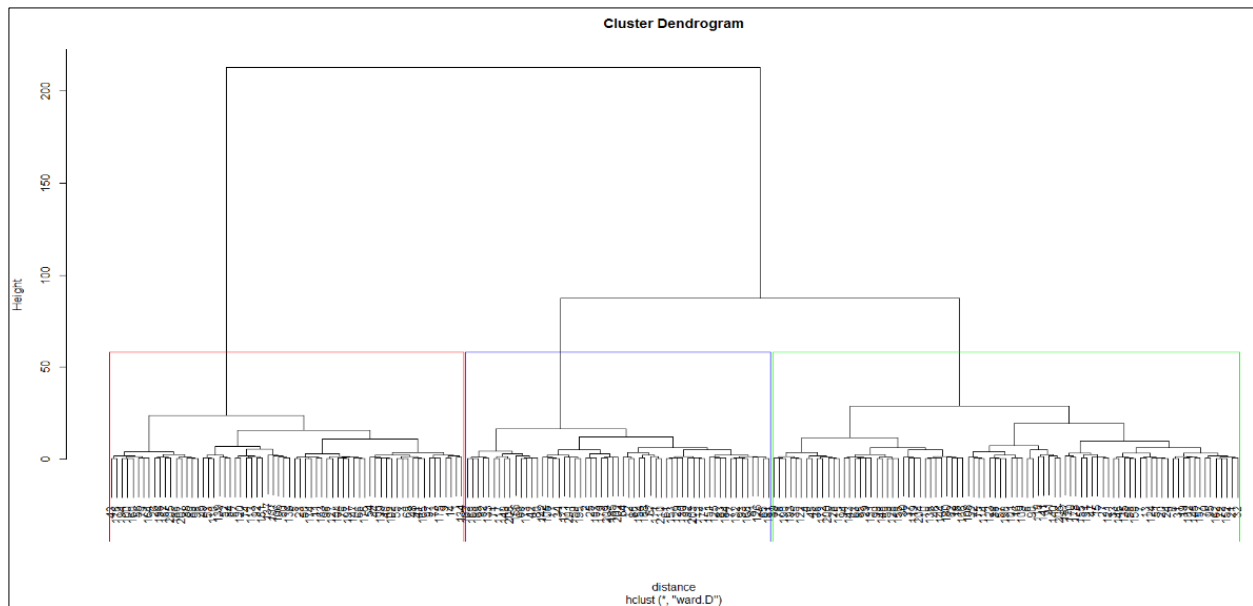
- * Among all indices:
- * 7 proposed 2 as the best number of clusters
- * 11 proposed 3 as the best number of clusters
- * 3 proposed 5 as the best number of clusters
- * 2 proposed 6 as the best number of clusters
- * 1 proposed 10 as the best number of clusters

***** Conclusion *****

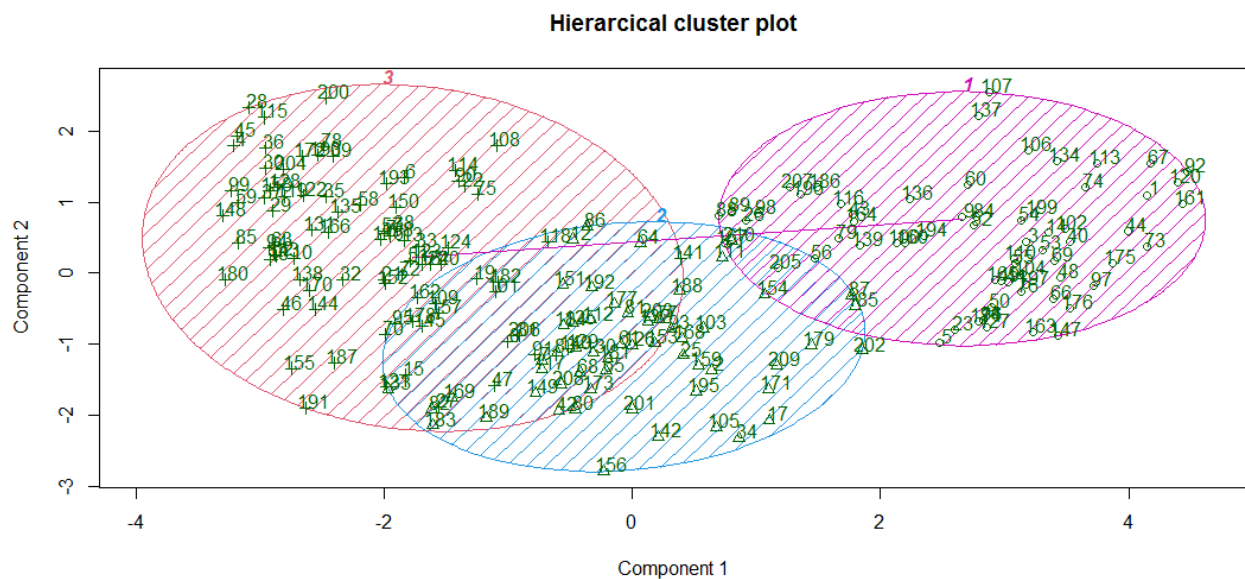
- * According to the majority rule, the best number of clusters is 3



As per this special method we have concluded that **three** is the optimal number of clusters. We are plotting the hierarchal clustering again with three as the optimal number of cluster and using rectangular plots. Below is the output of the three clusters.



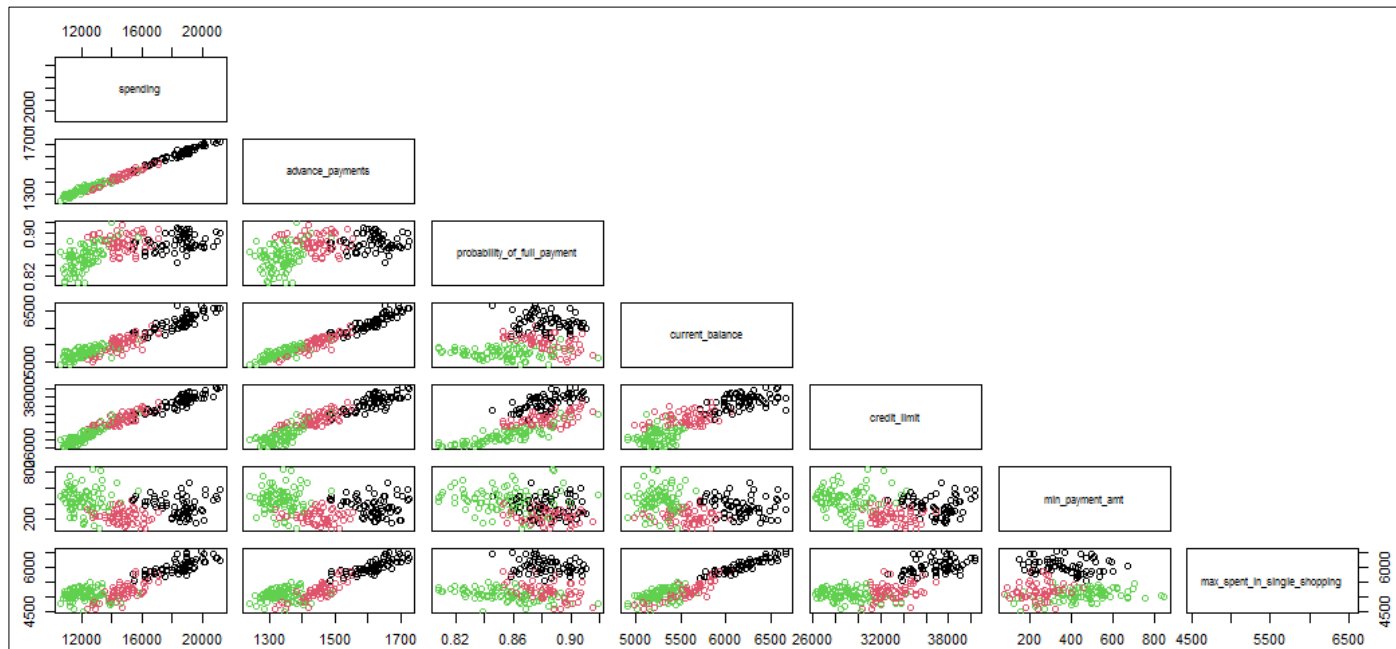
Visualizing the clusters can be done using cluster plot and below are the three clusters which is plotted in two axis which are two principal components.



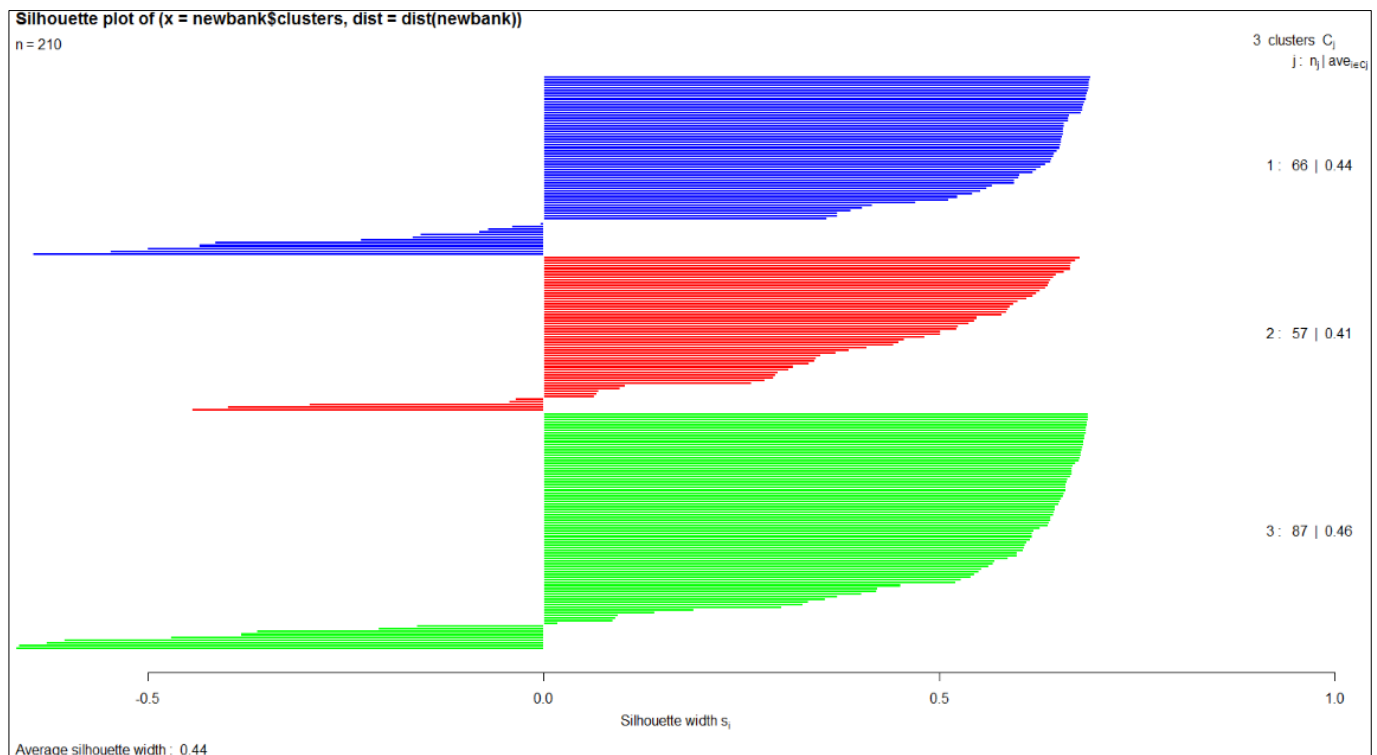
Next we aggregate the obtained cluster mapping to determine the number of clusters, below are the aggregate mean of the clusters after the required aggregation.

Group	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
1	18473.33	1619.697	0.8836803	6178.742	36921.97	370.7394	6045.606
2	14777.02	1449.632	0.8828193	5561.789	33067.54	244.6072	5184.035
3	12143.10	1335.816	0.8536333	5254.862	28981.26	451.6415	5071.207

Through the dimensionality reduced cluster plot, we will not be able to make out which variable is reduced dimensions. Hence we are plotting the below scatter cluster plot for each variable combination to visualize the clusters



Next we are plotting the calculated silhouette width for the three clusters and plotting the same below. We could see that the average silhouette width is high for three clusters compared to 1 and 2.



4.4 Qn 4 : Apply K-Means clustering on scaled data and determine optimum clusters.

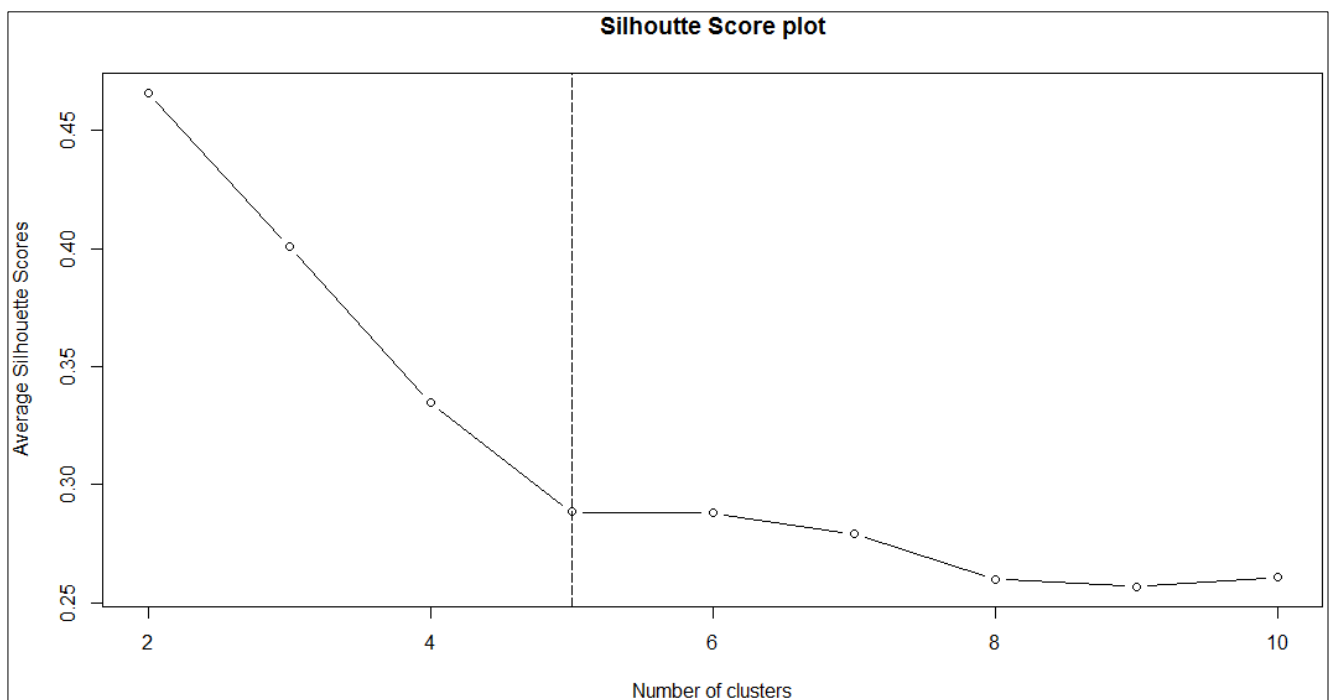
In case of the K – Means clustering model , we would require the optimal number of clusters to be provided/calculated before proceeding with building the model. The K – means algorithm generally works in the below order

- Select K centroids (K rows chosen at random).
- Assign each data point to its closest centroid.
- Recalculate the centroids as the average of all data points in a cluster (that is, the centroids are p-length mean vectors, where p is the number of variables).
- Assign data points to their closest centroids.
- Continue steps 3 and 4 until the observations aren't reassigned or until the maximum number of iterations is reached.

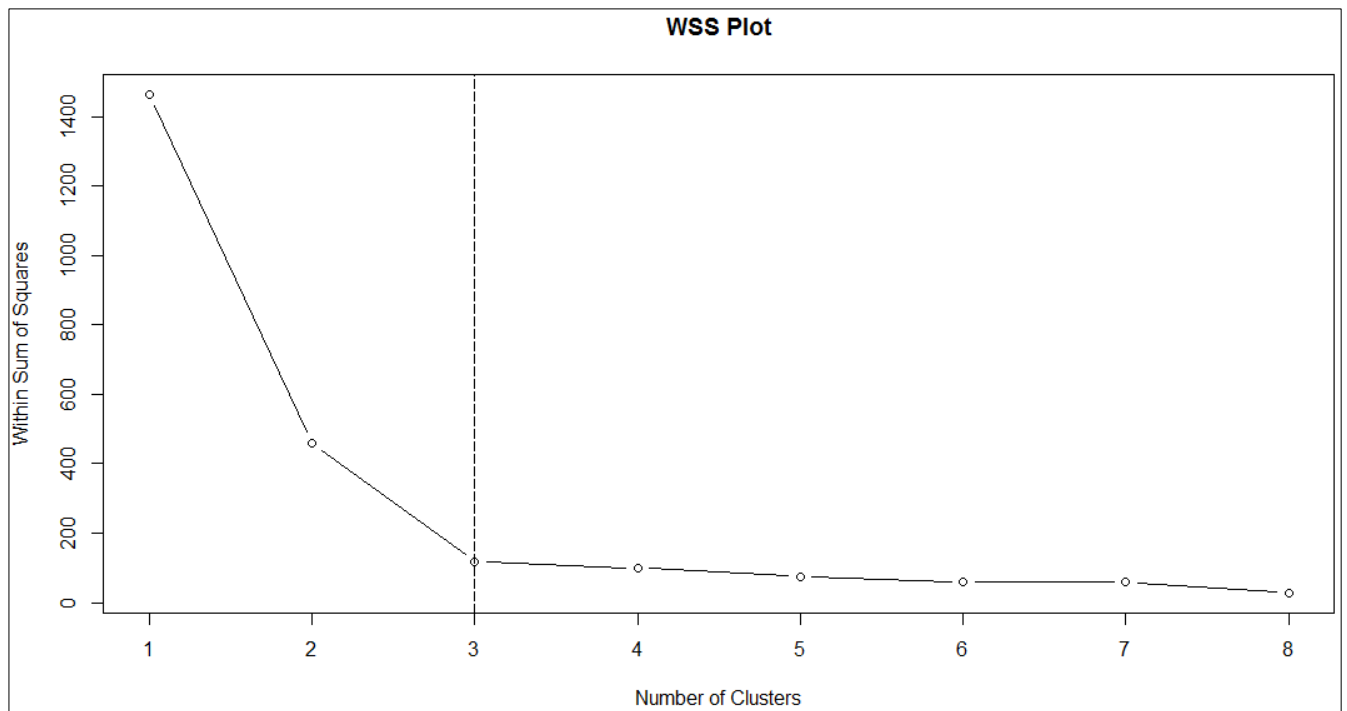
There are many methods to find the optimal number of clusters before proceeding with K – means which are discussed one by one below :

1) Silhoutte score plot method :

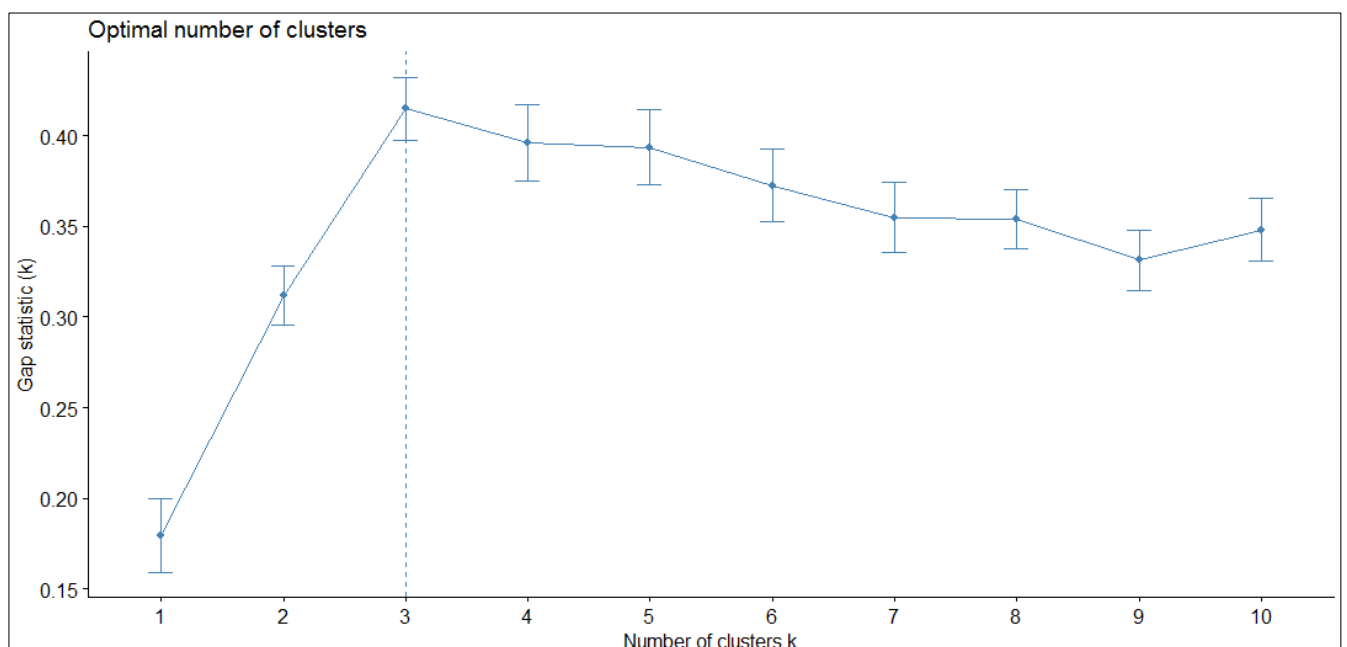
In this method the silhouette width scores are computed in a loop for the various number of clusters and then by means of elbow method we determine the number of clusters where the drastic change in silhouette score reduces. The Silhoutte is shown below, we could see that at number of clusters equal to five the elbow bend of the graph is seen. However this method does not confirm on the exact value of the number of clusters.



WSS plot method : In this method the within sum of squares values are computed in a loop for the various number of clusters and then by means of elbow method we determine the number of clusters .We could see that the cluster value of three shows an elbow bend which best states that the number of clusters must be three.



Gap Stat Method : This technique compares the change in the within cluster dispersion and it measures the deviation of the observed value from its expected value under the null hypothesis. The estimate of the optimal clusters will be the value that maximizes gap statistic values. From the below plot we could see that the optimal number of clusters chosen is three as per gap statistics method.



Confirming with Nbclust method :

The Hubert index is a graphical method of determining the number of clusters. In the **plot** of Hubert index, we **seek** a significant knee that corresponds to a Significant increase of the value of the measure i.e the significant peak **in** Hubert index second differences plot.

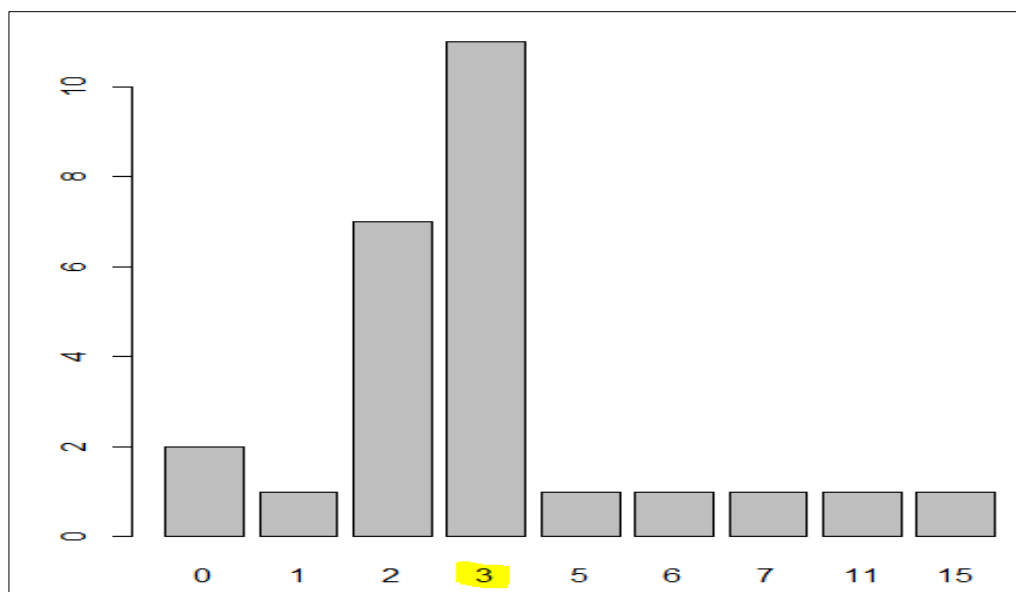
******* : The D index is a graphical method of determining the number of clusters. In the **plot** of D index, we **seek** a significant knee (the significant peak **in** Dindex second differences **plot**) that corresponds to a significant increase of the value of the measure.

* Among **all** indices:

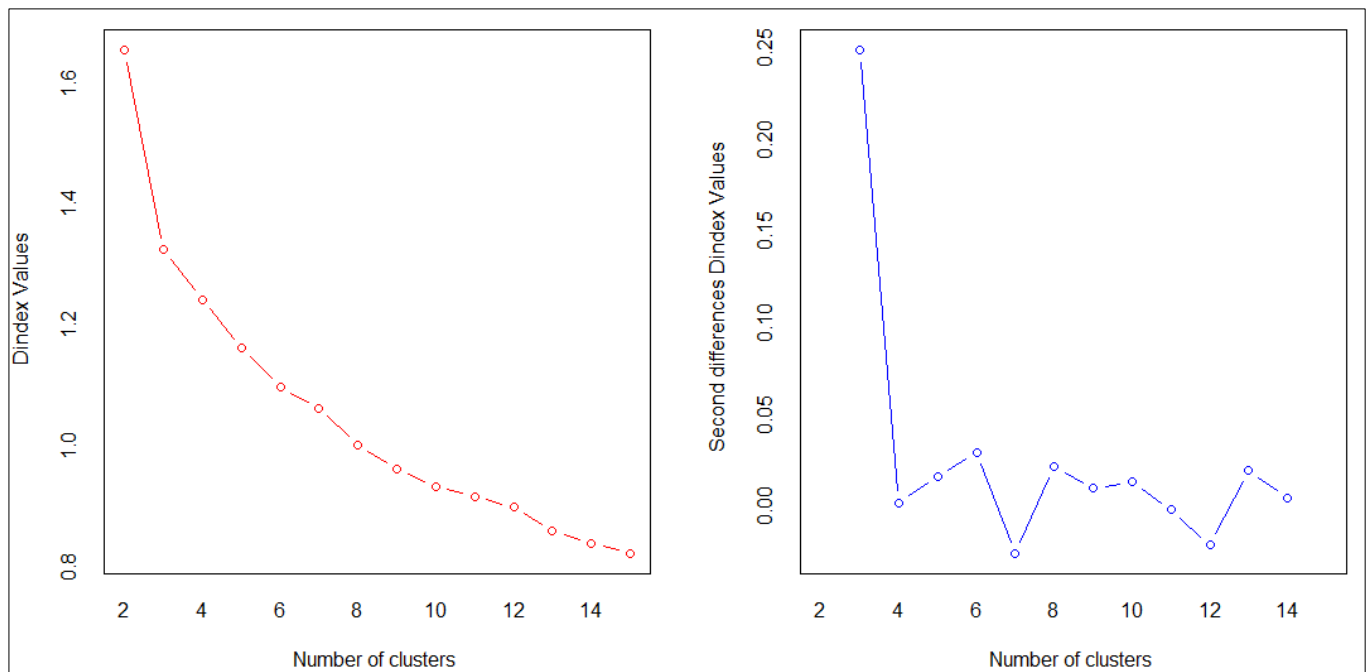
- * **7** proposed **2** as the best number of clusters
- * **11** proposed **3** as the best number of clusters
- * **1** proposed **5** as the best number of clusters
- * **1** proposed **6** as the best number of clusters
- * **1** proposed **7** as the best number of clusters
- * **1** proposed **11** as the best number of clusters
- * **1** proposed **15** as the best number of clusters

***** Conclusion *****

* According to the majority rule, the best number of clusters is **3**

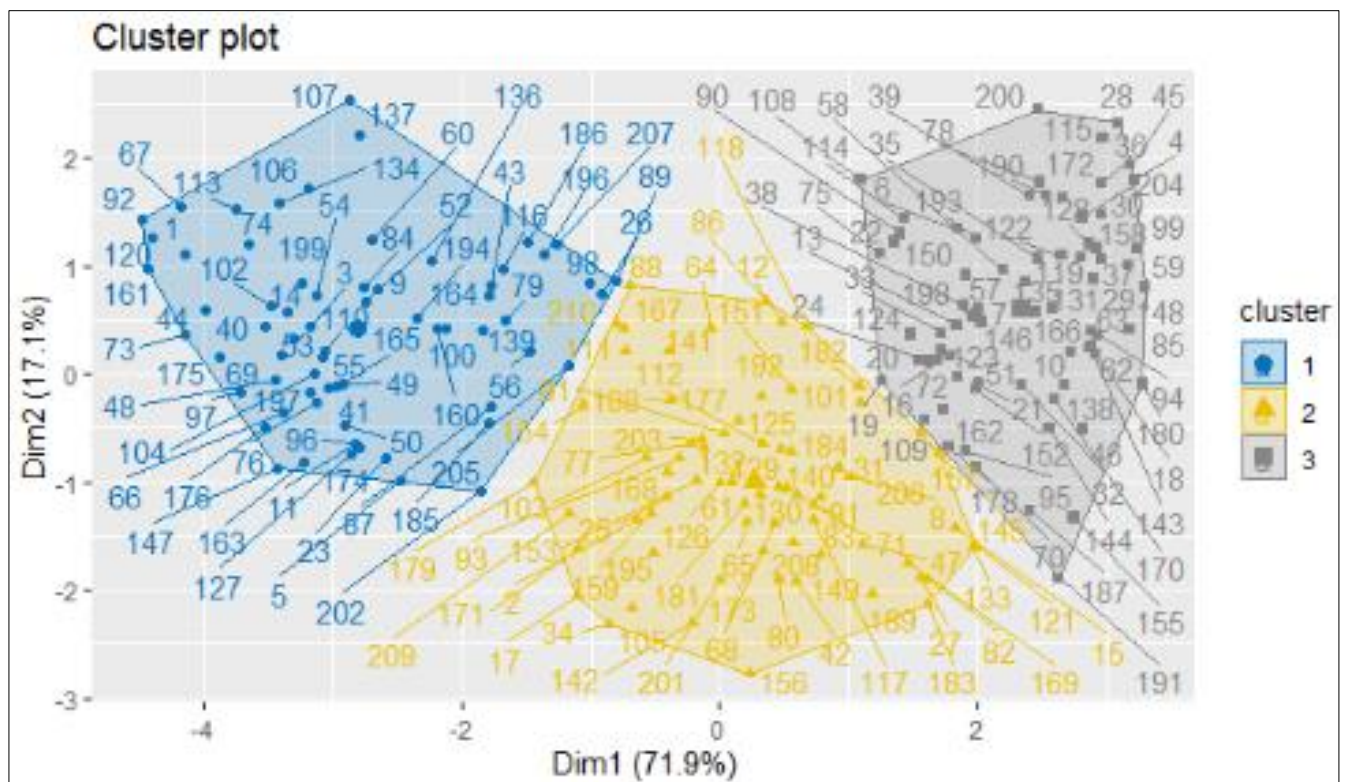


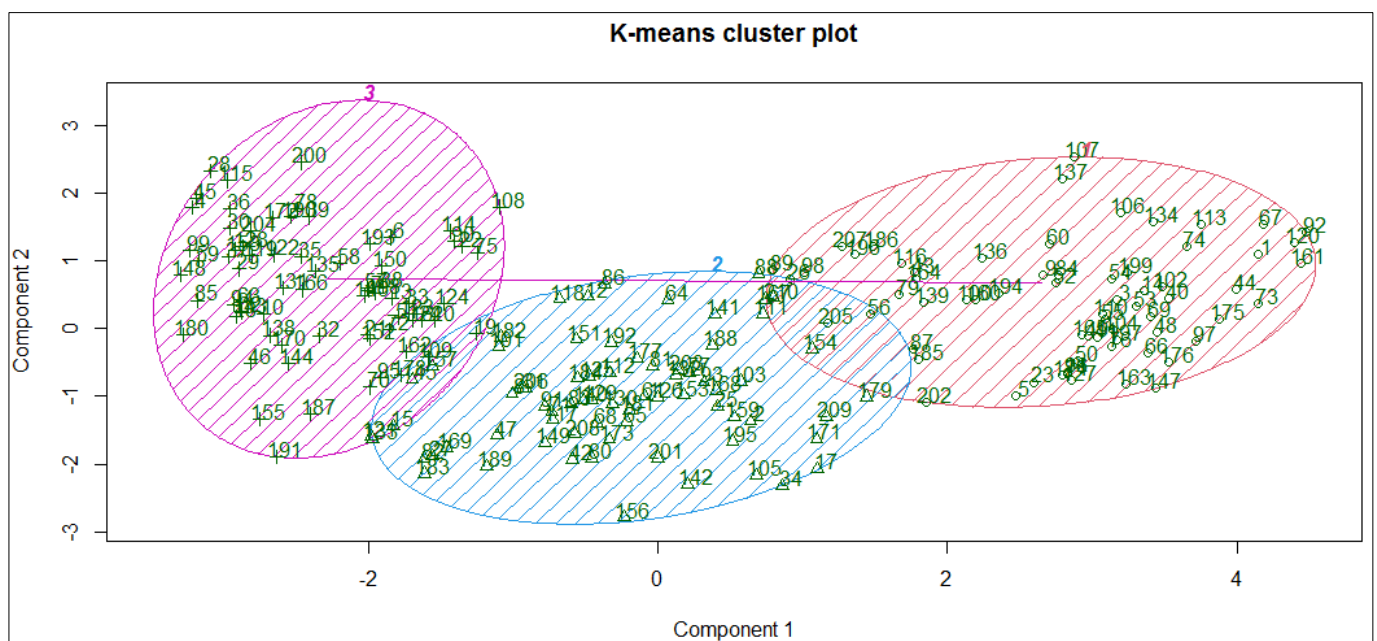
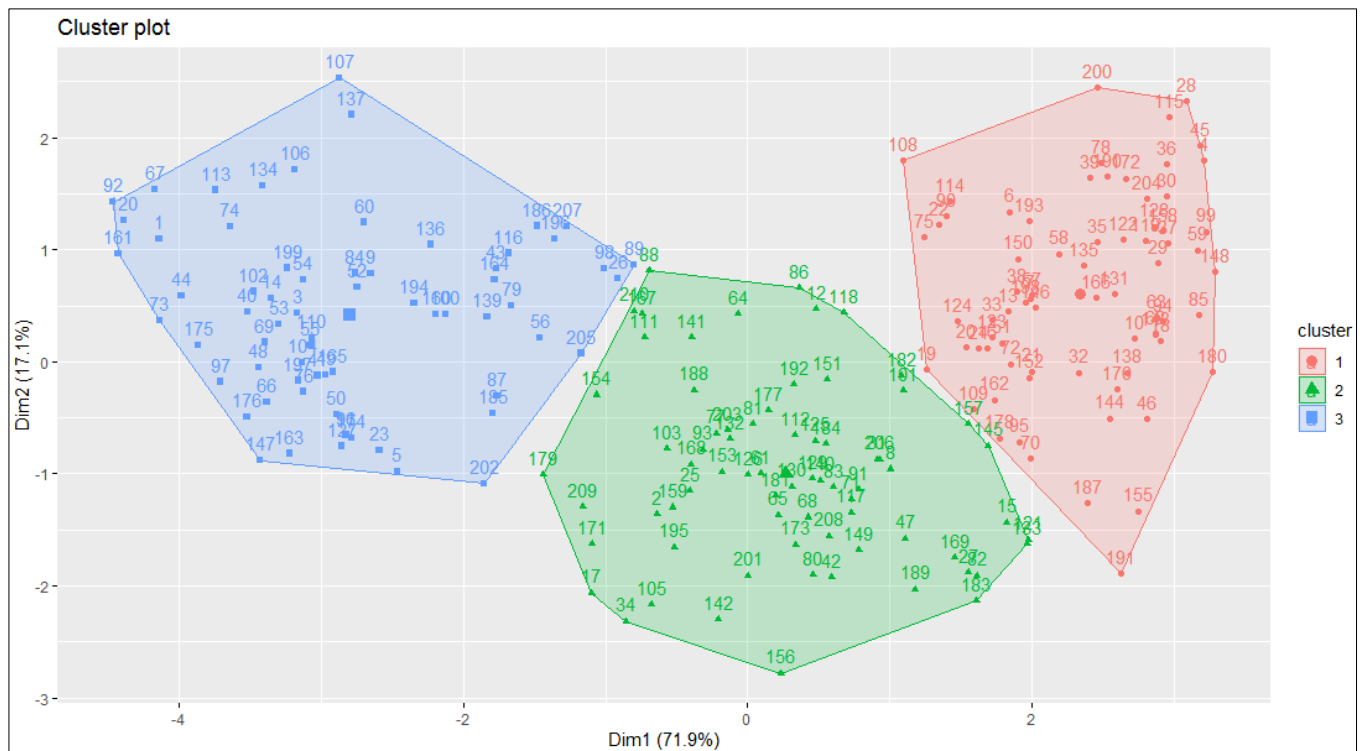
By Nbclust voting method, maximum votes is for three clusters hence the same has been confirmed with other methods such as gap stat and wss method.



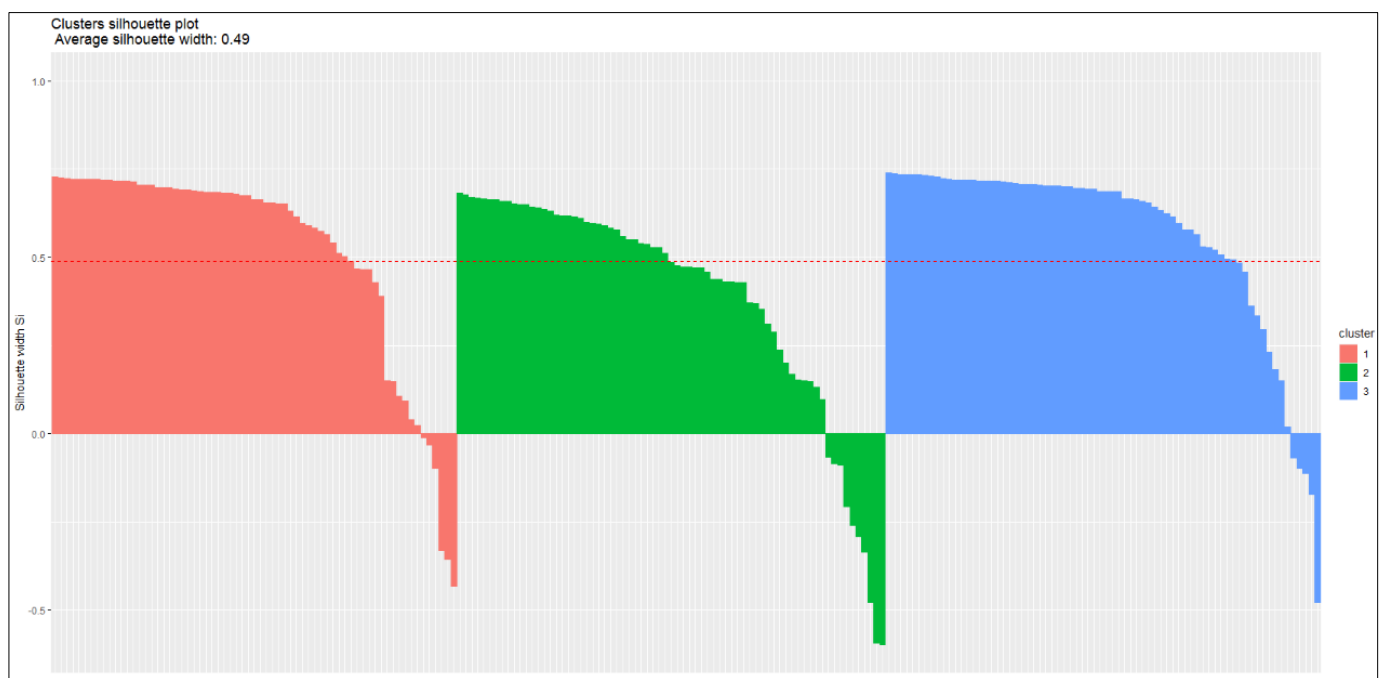
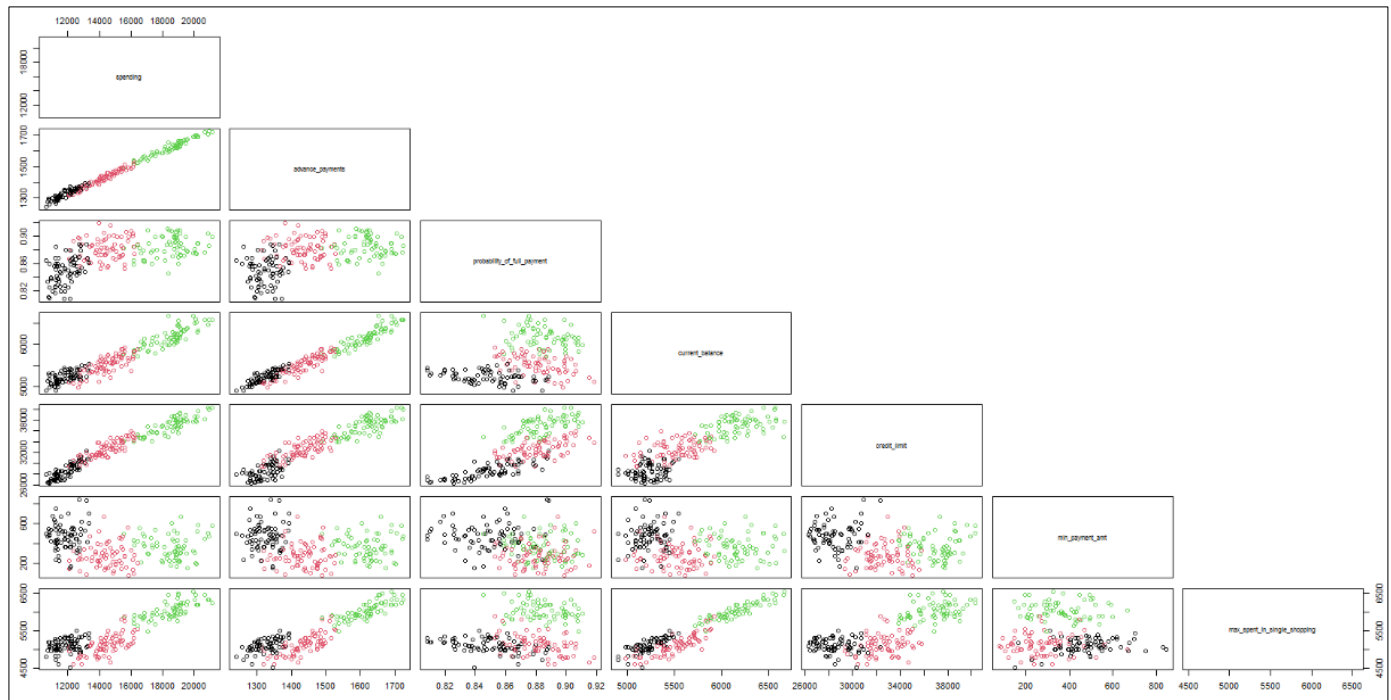
The optimal number of clusters as per this method is determined to be three hence we now perform the K – means cluster with three centers.

We now perform the K – means cluster with three centers and below is the output of the K – means model.





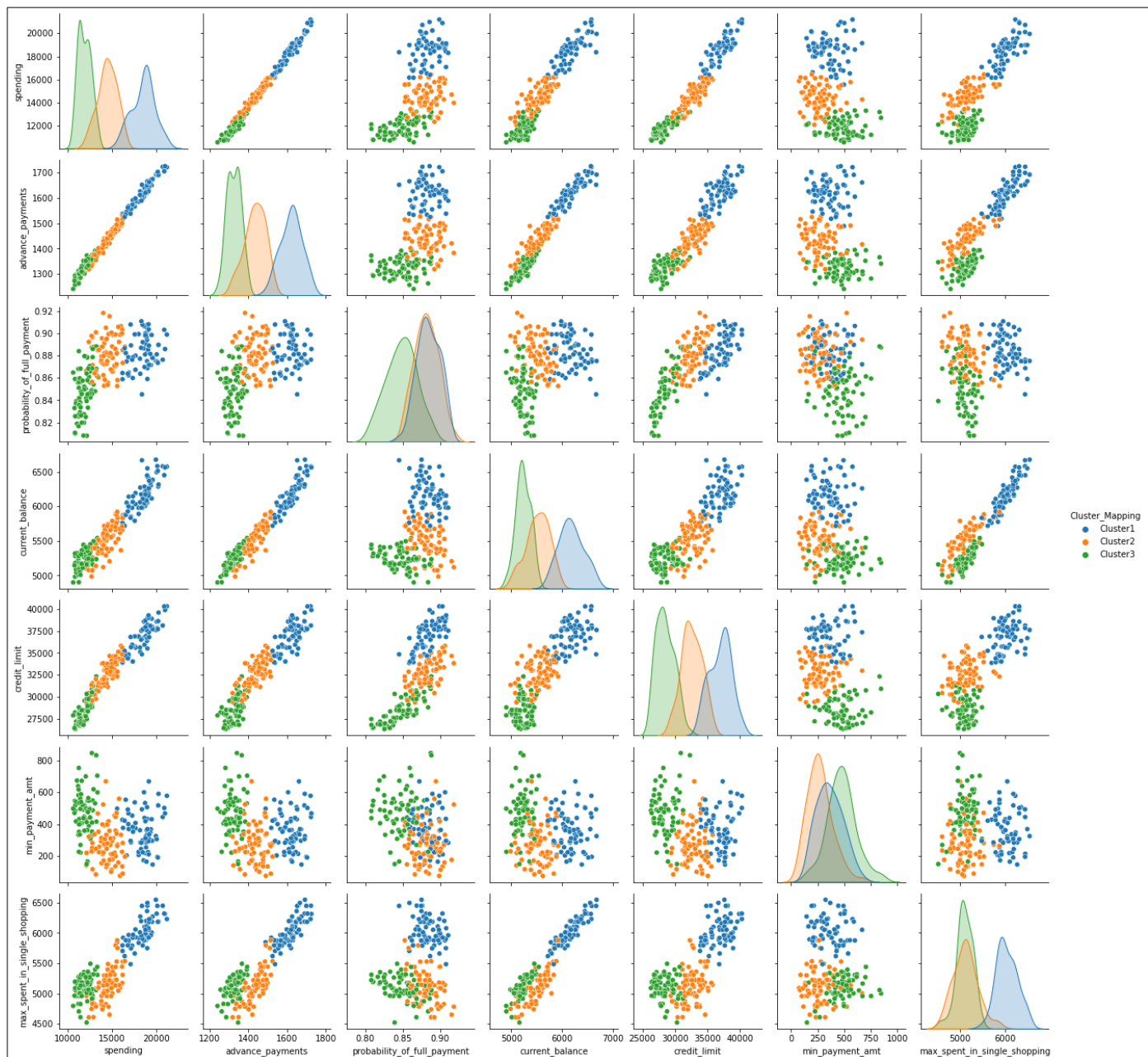
The three clusters are further split into individual variables and then plotted using special functions. It does not take into account the principal components as in case of above plot.



The aggregate mapping is done and the centroid of the three clusters are shown below.

Group	spending	advance_p ayments	probability_of_ full_payment	current_bala nce	credit_limit	min_payment_ amt	max_spent_in_ single_shoppin g	Comments
1	18495.37	1620.343	0.8842104	6175.687	36975.37	363.2373	6041.701	High spenders
2	14437.89	1433.775	0.8815972	5514.577	32592.25	270.7341	5120.803	Moderate spenders
3	11856.94	1324.778	0.8482528	5231.75	28495.42	474.2389	5101.722	Least spenders

The above profile and below plots best summarizes the three cluster profiles and the patterns.



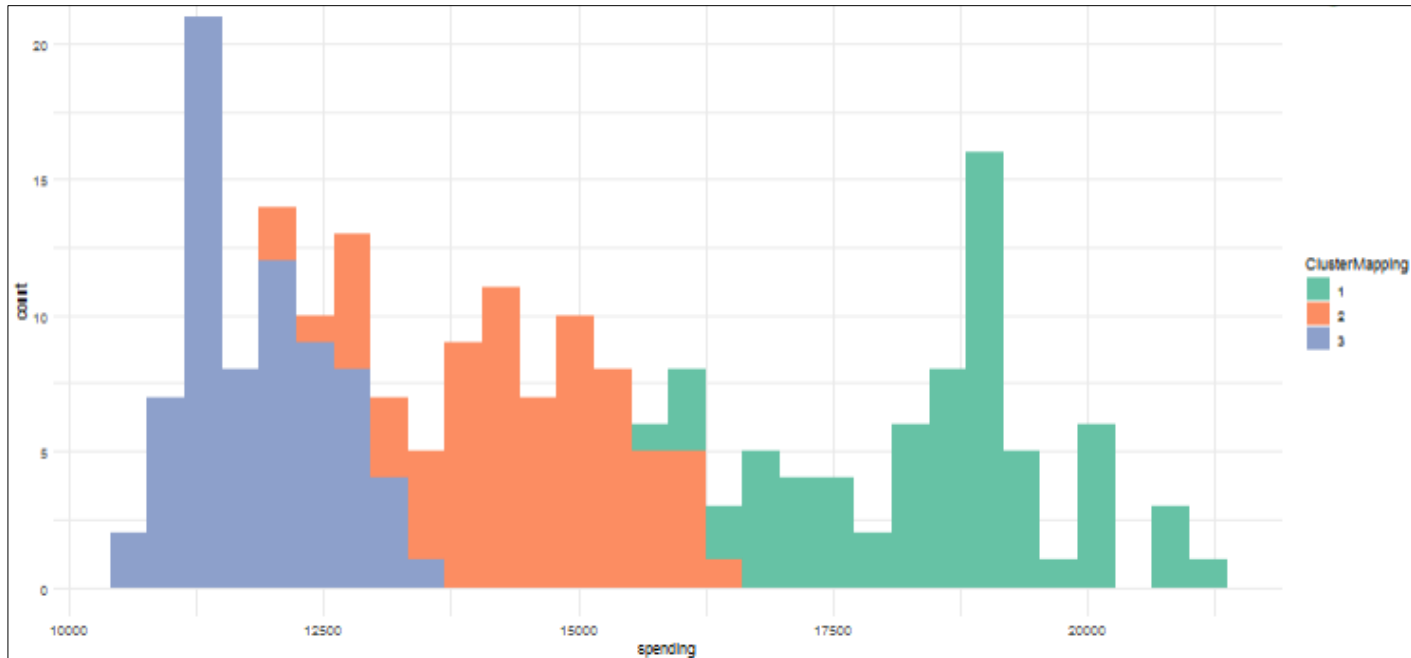
4.6 Qn 5 : Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters

The three cluster profiles are defined below with the help of histograms:

- Cluster 1: High spenders
- Cluster 2: Moderate spenders
- Cluster 3: Low spenders

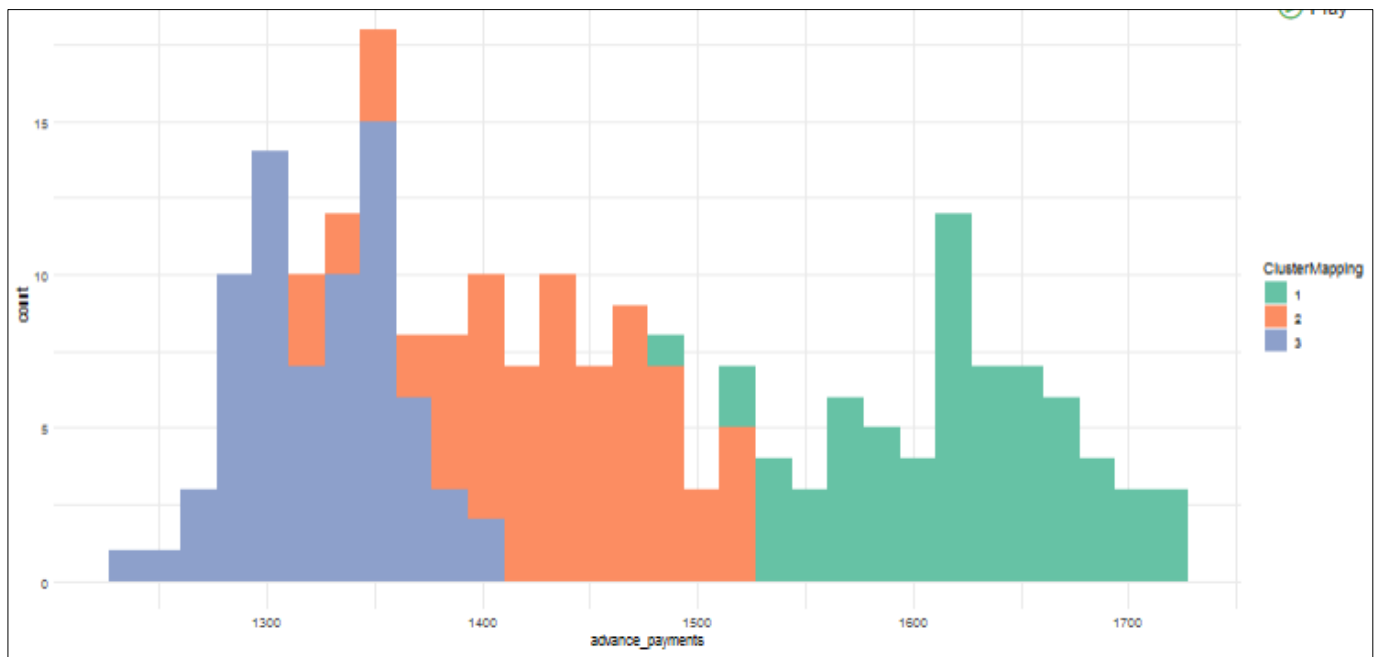
Spending

The cluster 1 are high spending customers who spend between Rs 16,000 to Rs 22,500, cluster 2 are moderate spenders and they spend between Rs 13,000 to Rs 16,000 and cluster 3 are those who spend low on credit card and they spend between Rs 11,000 to Rs 12,500



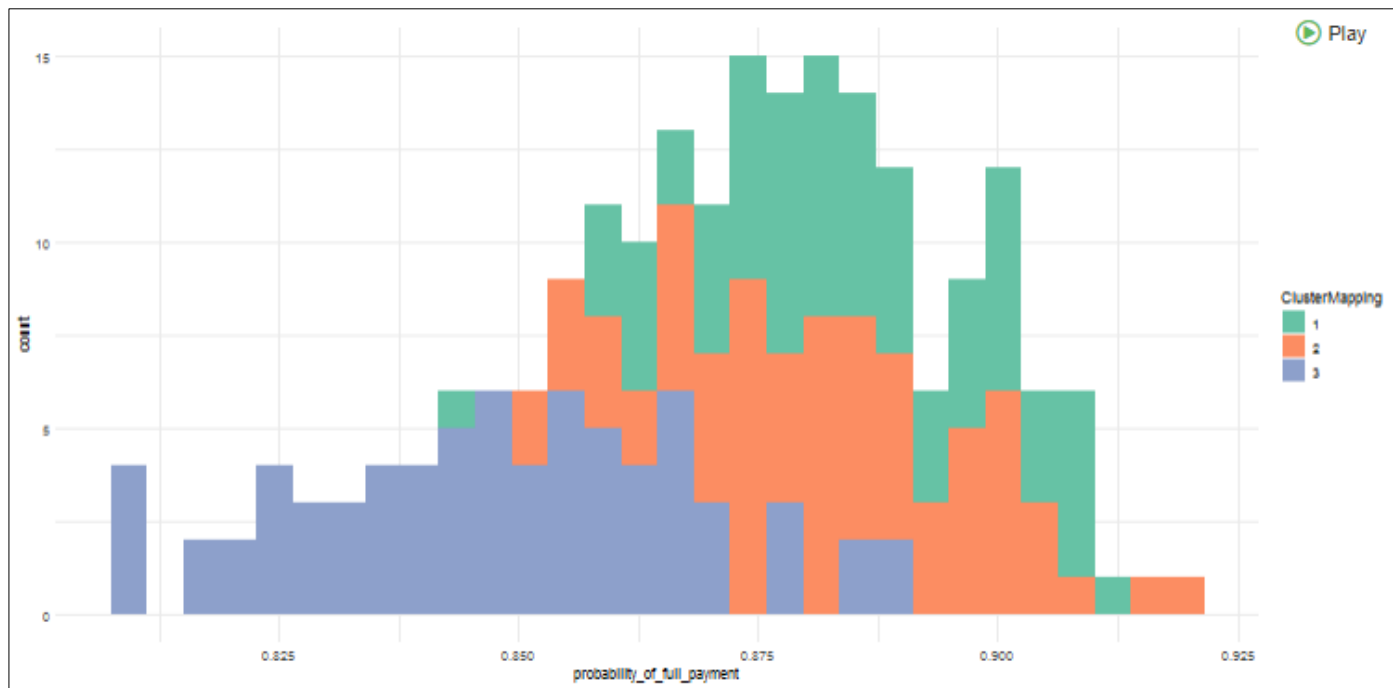
Advance Payments:

The cluster 1 are high advance payers who pay advance amount between Rs 1525 to Rs 1750 , cluster 2 are moderate advance payers and they pay advance amount between Rs 1375 to Rs 1525 and cluster 3 are those who pay low advance amount on credit card and its between Rs 1200 to Rs 1350



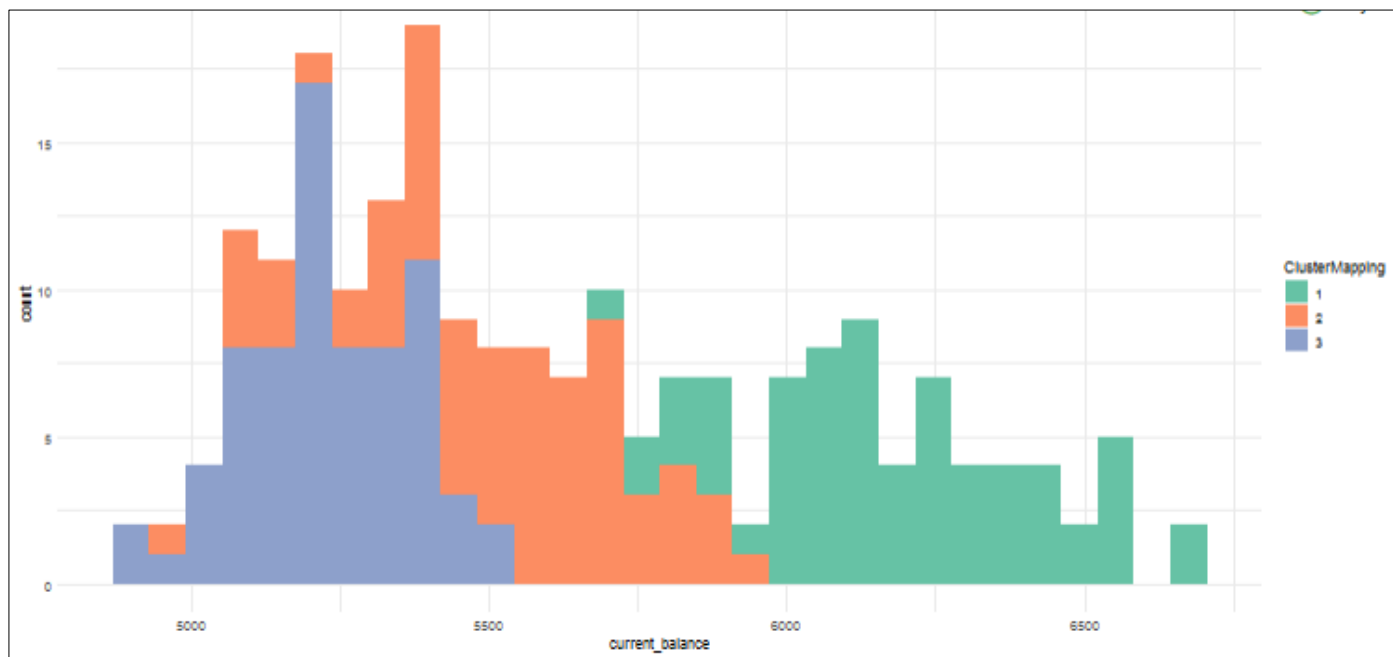
Probability of Full Payment:

This is a normally distributed histogram with cluster 1 and cluster 2 showing probability of paying the full amount to bank between 86 % to 92 % . The cluster 3 shows low probability of paying compared to other two which is between 78 % to 86 % .



Current Balance

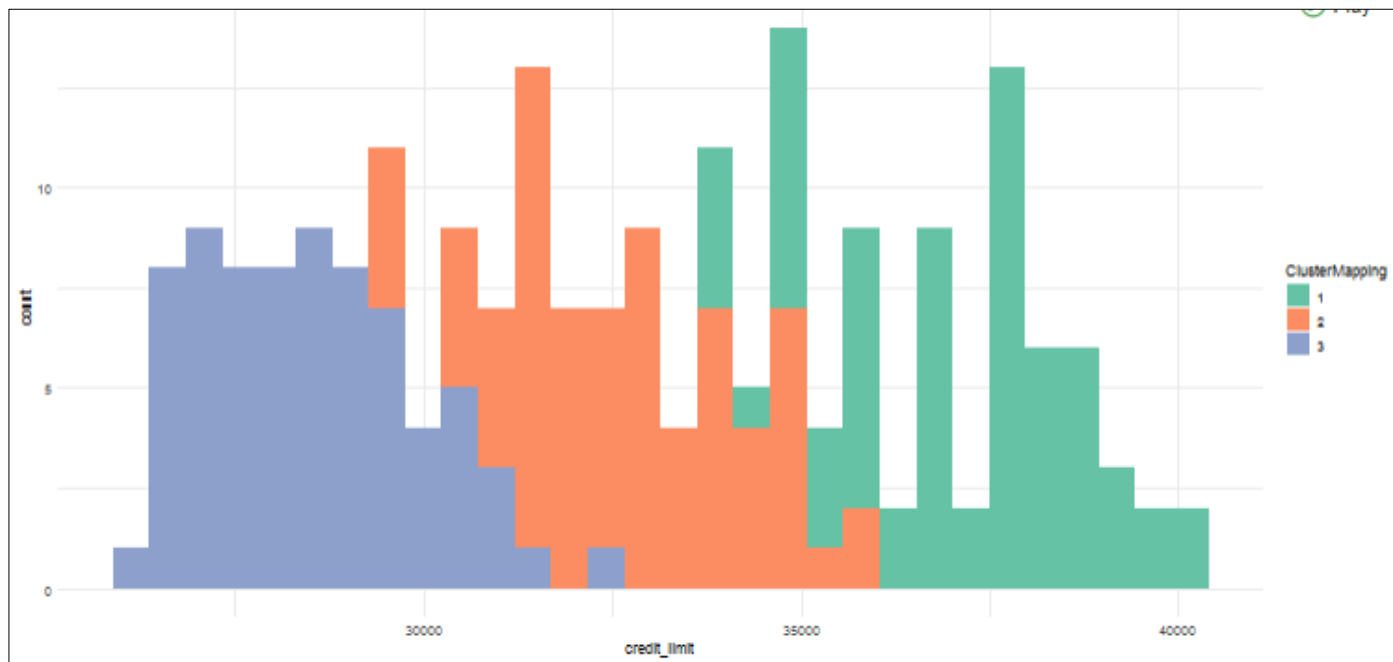
The cluster 1 have high current balance amount in the account between Rs 5750 to Rs 6750 , cluster 2 and 3 are moderate & low current balance amount maintainers between Rs 4000 to Rs 5600.



Credit Limit

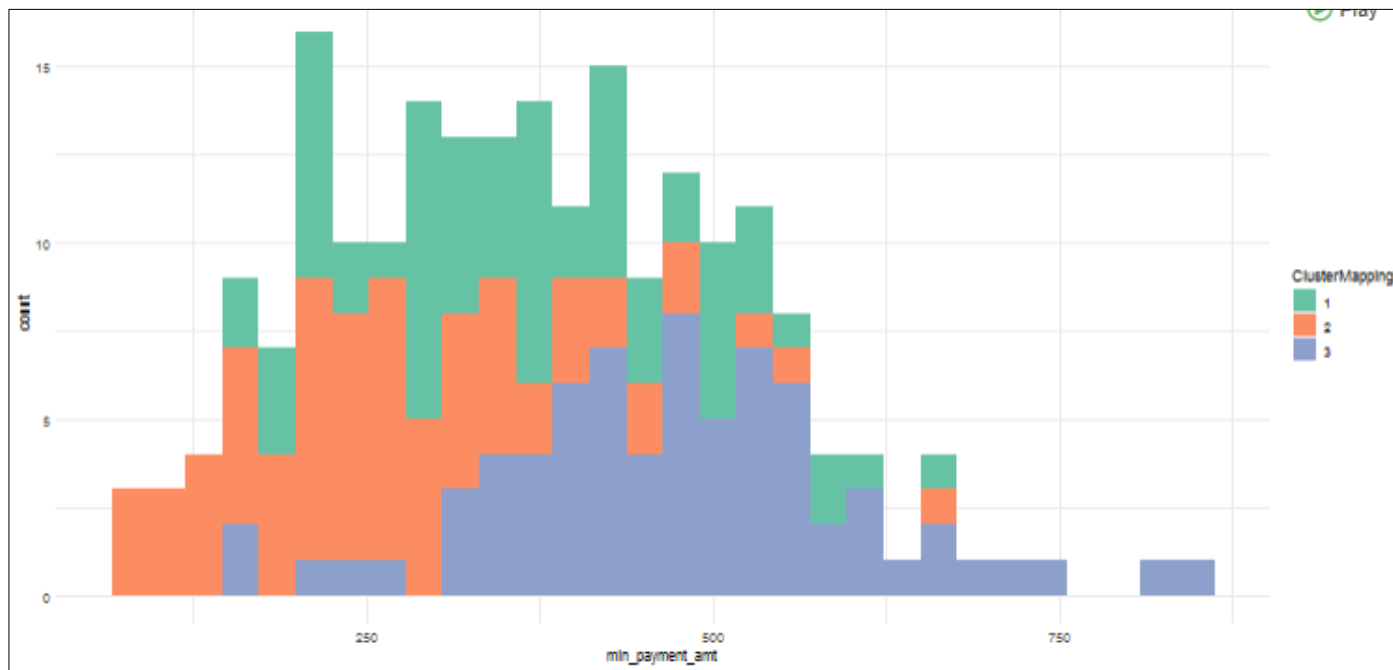
The cluster 1 have high credit limit for credit card between Rs 35,000 to Rs 40,500, cluster 2 have moderate credit limit between Rs 31,000 to Rs 34,000 and cluster 3 are those who spend low and hence

have low credit limit between Rs 25,000 to Rs 30,000.



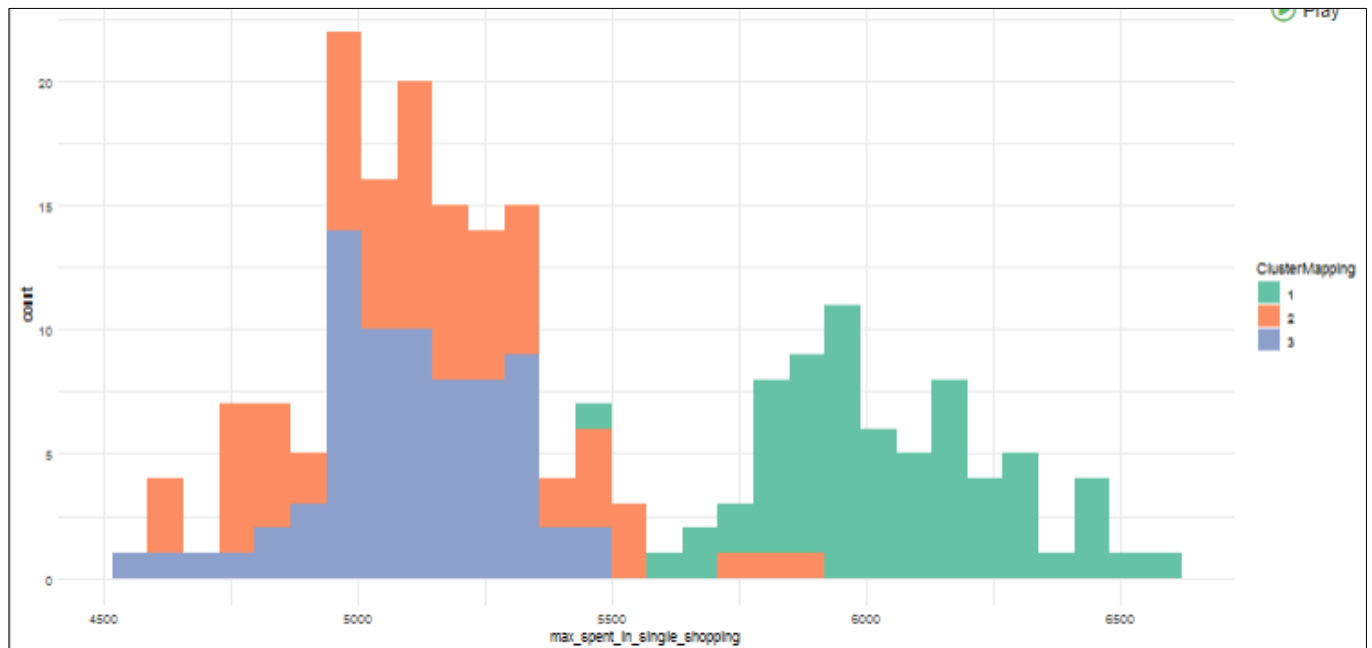
Minimum Payment Amount

The cluster 3 pay higher minimum amount between Rs 400 to Rs 800 since they have less probability of repayment and spend less. The repayment due amount will be high. Cluster 2 and 1 have low minimum payment amount between Rs 100 to Rs 500 since they have high probability of repayment.



Maximum Amount Spent in a single shopping

It is evident below that the Cluster 1 spend maximum in a single shopping between Rs 5700 to Rs 6700 while the cluster 2 & 3 are thrifty and spend moderate or less amount in the range Rs 4500 to Rs 5500.



Promotional Strategies:

- Spending group (Cluster 1): This group has large monthly spending as well as they spend large amount in single shopping. As they have a larger current balance their tendency to spend is more and they pay in advance for their shopping and have larger probability of completing payment. Hence the promotion strategy is to provide cashback and discount offers for shopping with credit card at lucrative percentage rates. Since they spend large amount in single shopping such cashback on e-commerce or others will.
- Saving group(Cluster 2): This group are average spenders and average account balance but the reason the group was named saving group because of the min and max spent payment as the min payment is the least and maximum payment is near about third group. These groups are ambiguous as they spend at reasonable times and are conservative at challenging times. Hence during festive seasons or at the time when this group spends cashback and. Since this group focuses more on saving offers like credit points for fuel, utilities, insurance and cash advance payments. Fuel surcharge waiver, grocery and supermarket save offers can be provided.
- Least spenders (Cluster 3): These group have least spending habits, maintain less account balance, have less probability of repayment and pay high minimum payment amount. The promotional strategy is to increase advance payment times which does not lead to penalty or fine, provide promotion to increase the credit limit if they spend more than specific amount in a month. The least spenders can be made to spend more by cross selling and providing more offers to move from thriftiness to moderate spending. Another way is to provide redeem points for them and offer more credit limit, by providing points for travel or shopping or to redeem gifts can be a promotion strategy. Also to see where these low spenders are best interested in initial small promotional offers such as discount on travel, restaurants, holiday resorts etc can be provided and tested to see which area of interest these customers try to spend more due to promotional offers.

5 Conclusion

We studied how unsupervised machine learning technique can be used to get patterns from the data set. We studied how K – means and hierarchical clustering are most prominent clustering methods. We have seen various methods to determine optimal number of clusters. Using the business problem of a bank, we identified the three clusters and the behavior of customers on the individual clusters based on the data variables. Thus segmentation of customers were performed using both hierarchical and k-means clustering methods that will further enable us for targeted marketing through promotional offers on the credit card customers. We also saw various methods in determining the optimal number of clusters with special reference to Nbclust which is based on voting method on various cluster method combinations. We learnt how to build clustering models and evaluated its accuracy.

6 Appendix A – Source Code

Attached separately with file named “ML main proj clustering (updated)”