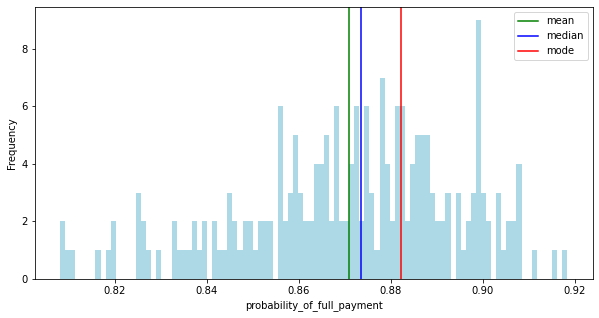
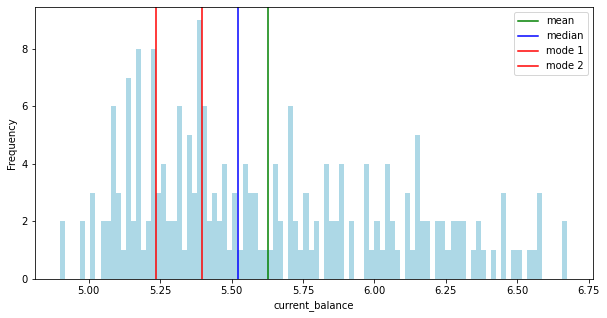
**Problem 1: Clustering**

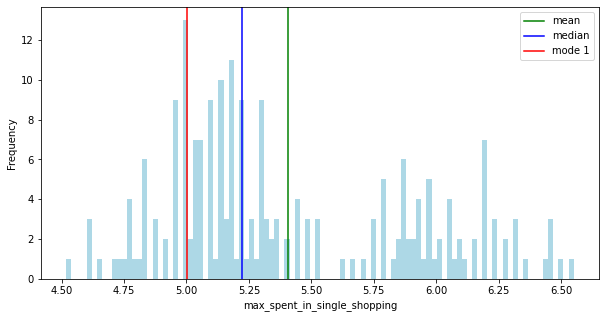
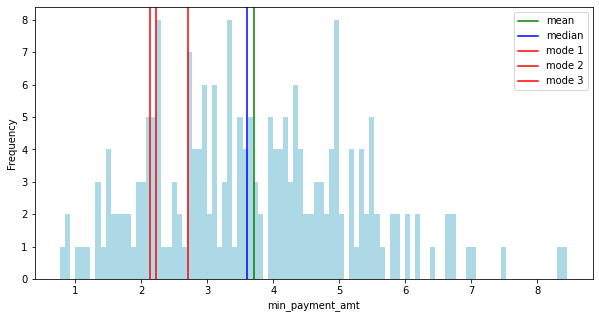
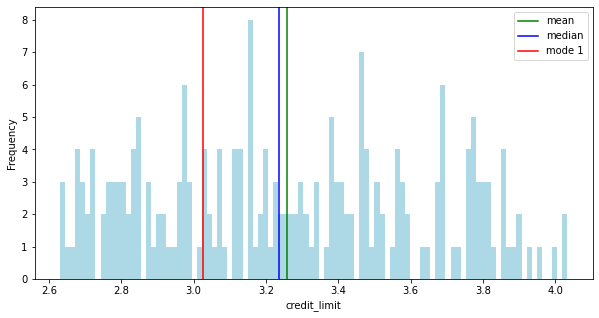
**A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.**

**1.1 Read the data and do exploratory data analysis. Describe the data briefly.**

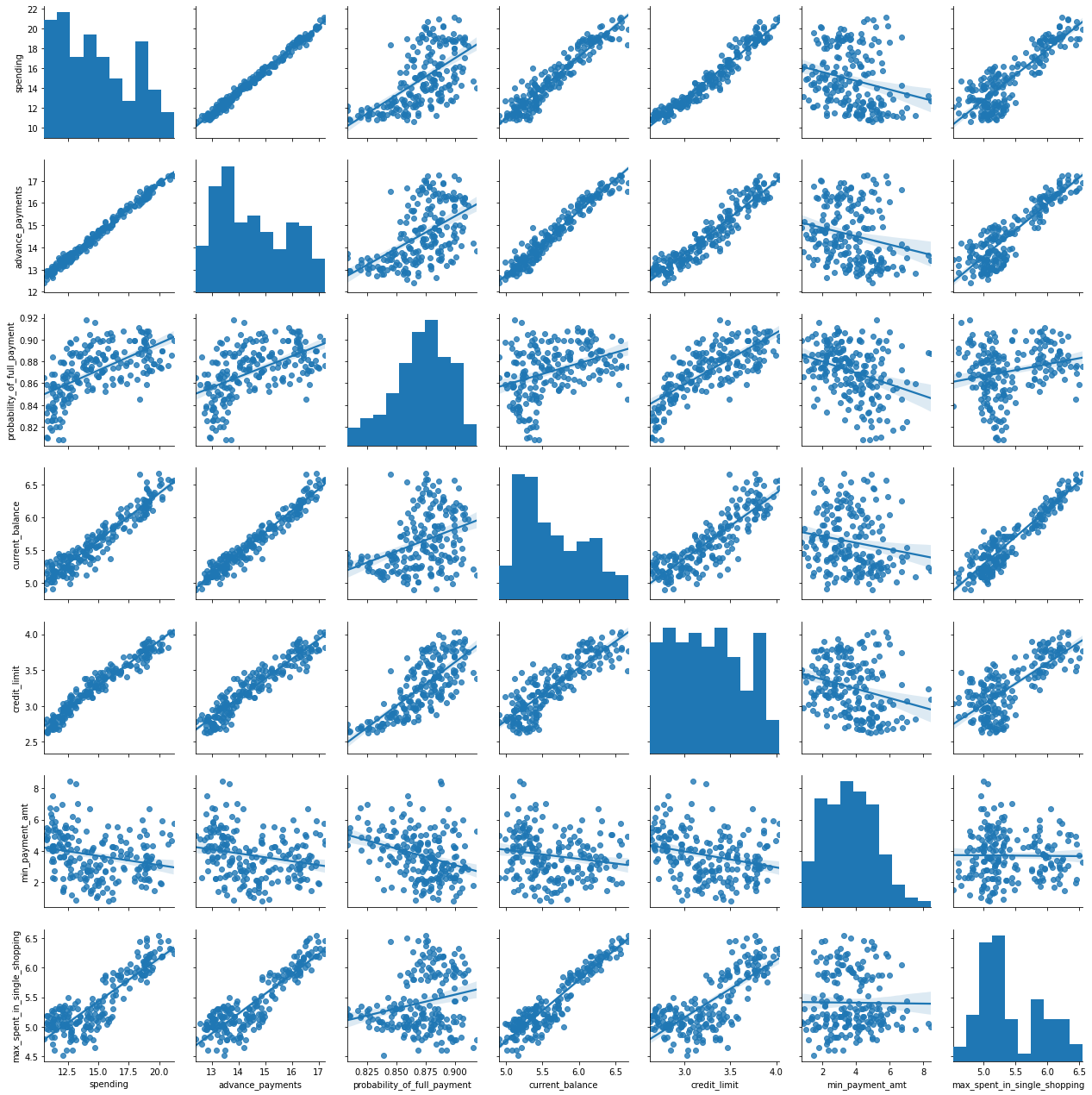
There are 210 Observations and 7 Attributes. All are of float data type. Mean and median are almost the same for all the variables. Data dispersion is more in spending. The summary of mean, mode, median using histogram.

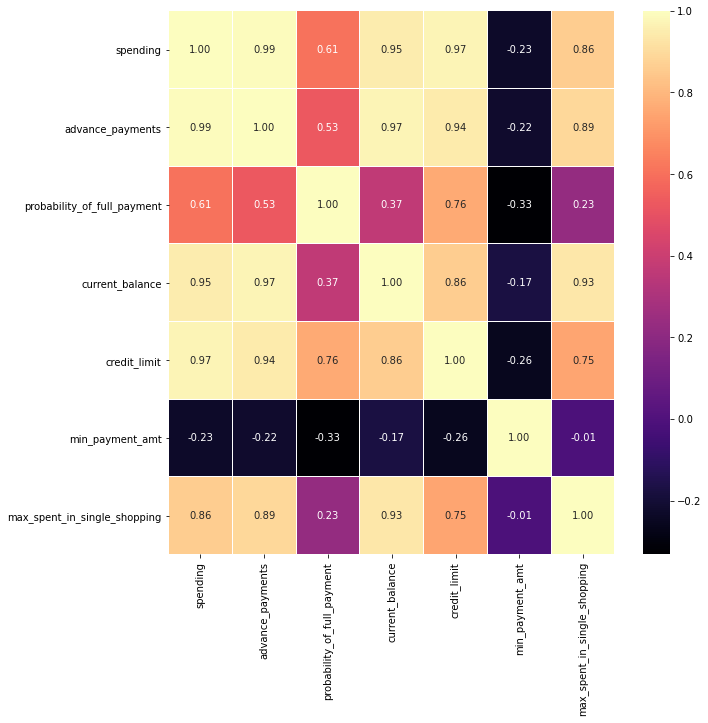
****

****

****

For all the variables except probability\_of\_full\_payment, data seems to be positively distributed.

Correlation between variables is observed by using pair plot and heat map.

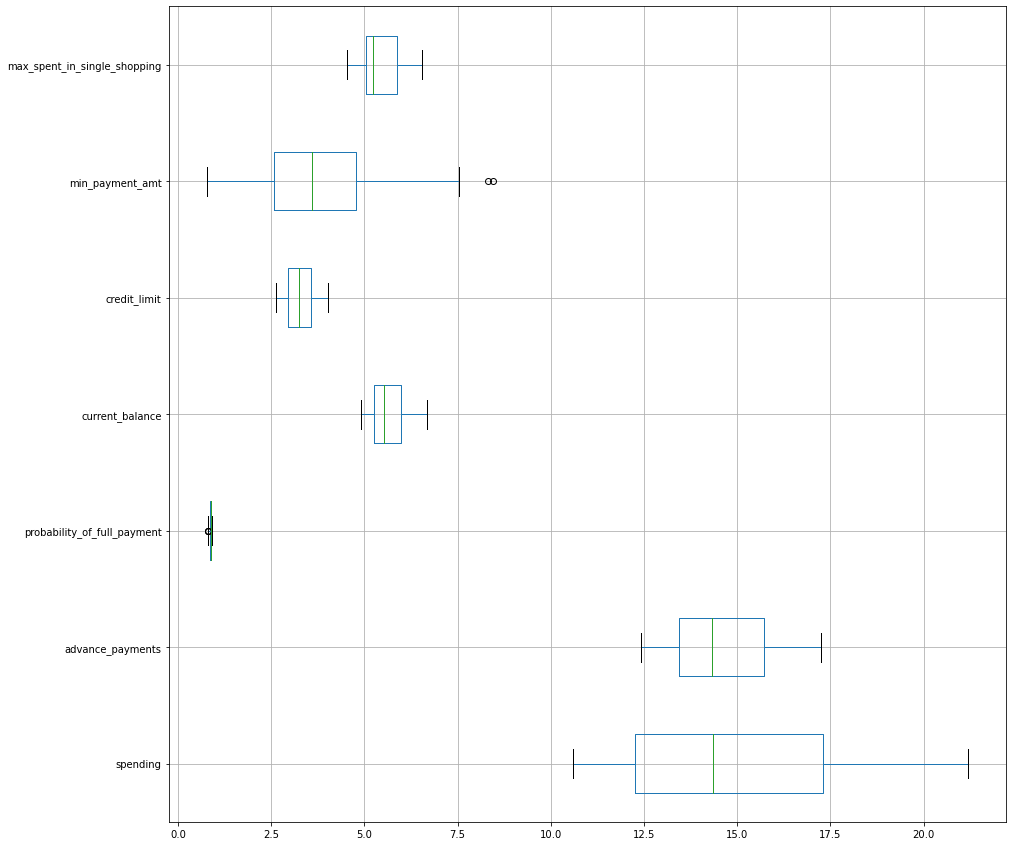
****

There is some negative correlation between variables. spending and advance\_payments are highly correlated. spending and credit\_limit are also highly correlated. advance\_payments and current\_balance are also strongly correlated. probability\_of\_payment and min\_payment\_amt are negatively and least correlated.

Missing values are checked using isnull() function and there are no missing values.

Duplicate values are checked using the duplicated() function and observed that there are no duplicate values.

Outliers are checked using boxplot and there seems to be outliers in min\_payment\_amt and probability\_of\_full\_payment.

****

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

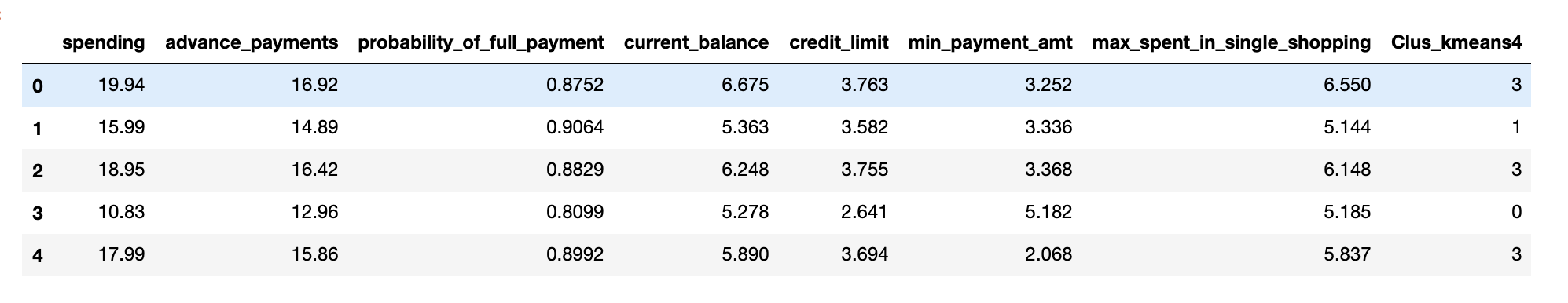
Scaling is necessary because By checking the summary, we can say that min, max and mean are not equal for all the columns. Data is scaled using MinMaxScaler()

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

p value shows the last 10 merges. Green cluster has maximum products, the orange cluster has minimum products. We created the linkage, that is the distances. Now obtaining the clusters and products belong to each cluster for final verification. We will use these fcluster to form the clusters by using distance as criterion.

|  |
| --- |
| array([1, 3, 1, 2, 1, 2, 2, 3, 1, 2, 1, 3, 2, 1, 3, 2, 3, 2, 2, 2, 2, 2,  1, 2, 3, 3, 3, 2, 2, 2, 3, 2, 2, 3, 2, 2, 2, 2, 2, 1, 1, 3, 1, 1,  2, 2, 3, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 2, 1, 3, 2, 2, 3, 3, 1,  1, 3, 1, 2, 3, 2, 1, 1, 2, 1, 3, 2, 1, 3, 3, 3, 3, 1, 2, 3, 3, 3,  1, 2, 2, 1, 3, 2, 2, 1, 1, 1, 2, 1, 2, 1, 3, 1, 3, 1, 1, 2, 2, 1,  3, 3, 1, 2, 2, 1, 3, 2, 2, 1, 3, 2, 2, 2, 3, 3, 1, 2, 3, 3, 2, 3,  2, 1, 2, 1, 1, 2, 1, 3, 3, 3, 2, 2, 3, 2, 1, 2, 3, 2, 3, 2, 3, 3,  3, 3, 2, 2, 3, 1, 1, 2, 1, 1, 1, 2, 3, 3, 3, 3, 3, 2, 3, 1, 1, 1,  3, 3, 3, 2, 3, 3, 3, 3, 3, 1, 3, 3, 3, 2, 3, 3, 2, 1, 3, 1, 1, 2,  1, 2, 3, 3, 3, 2, 1, 3, 1, 3, 3, 3], dtype=int32) |

The clusters for the first five rows are as follows in the output below

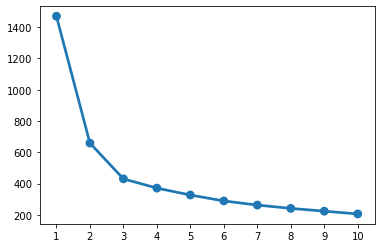
****

**1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score.**

Data is scaled using StandardScaler() function. Within the cluster sum of squares reduces as K keeps increasing. `Wss` for `k value` 1 to 11 are

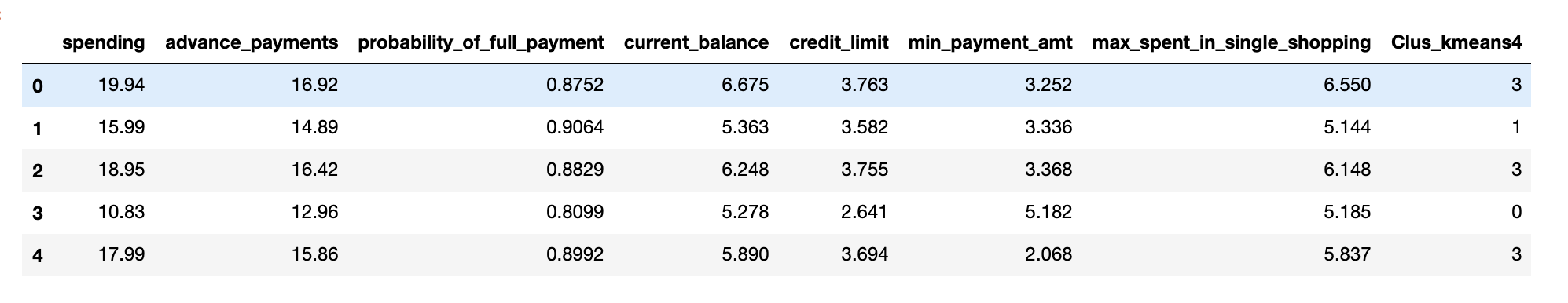
|  |
| --- |
| [1469.9999999999998,  659.171754487041,  430.6589731513006,  371.38509060801096,  327.21278165661346,  289.31599538959495,  262.98186570162267,  241.81894656086033,  223.91254221002725,  206.39612184786694] |

From the plot, we can say that after 3 clusters there isn't much difference. To evaluate clusters, silhouette score is used here.



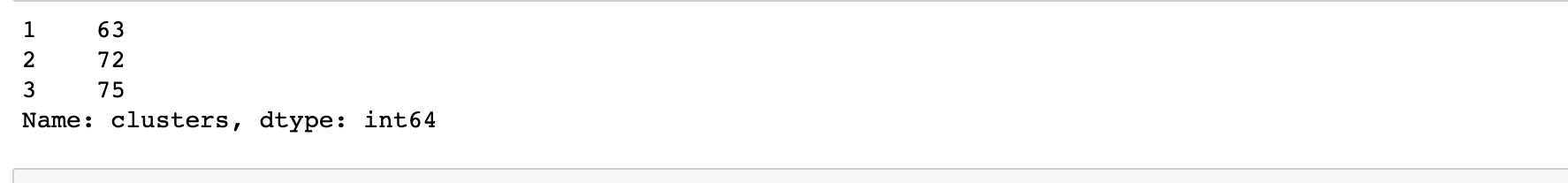
Silhouette score is better for 3 clusters than for 4 clusters. So, final clusters will be 3.

After appending Clusters to the original dataset, the dataset looks like this



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

Cluster frequency for Hierarchical Cluster

****

*Cluster profiles:*

Cluster 1: Tier 3 (contains high spending customers per month and also Probability of payment done in full by the customer to the bank is high, Balance amount left in the account to make purchases is also high, Limit of the amount in credit card is high, Amount paid by the customer in advance by cash is high, minimum paid by the customer while making payments for purchases made monthly is low compared to Tier 2, Maximum amount spent in one purchase is high)

Cluster 2: Tier 2 (contains low spending customers per month and Probability of payment done by the customer to the bank is low , Balance amount left in the account to make purchases is low, Limit of the amount in credit card is low, Amount paid by the customer in advance by cash is low, minimum paid by the customer while making payments for purchases made monthly is high, Maximum amount spent in one purchase is low)

Cluster 3: Tier 1 (contains low spending customers per month compared to Tier 3 and Probability of payment done by the customer to the bank is low compared to Tier 1 , Balance amount left in the account to make purchases is low compared to Tier 3, Limit of the amount in credit card is low compared to Tier 3, Amount paid by the customer in advance by cash is low compared to segment3, minimum paid by the customer while making payments for purchases made monthly is low, Maximum amount spent in one purchase is low compared to Tier 3)

Recommendations:

1. Bank can give high preference to the people in Tier 3 for giving promotional offers because they seems to be spending high and their credit limit is also high and the Probability of payment done in full by the customer to the bank is high

2. Bank can give low preference to the people in Tier 2 for giving promotional offers because they seem to be spending low but their credit limit is high and the Probability of payment done in full by the customer to the bank is low and also their balance in the account is low.

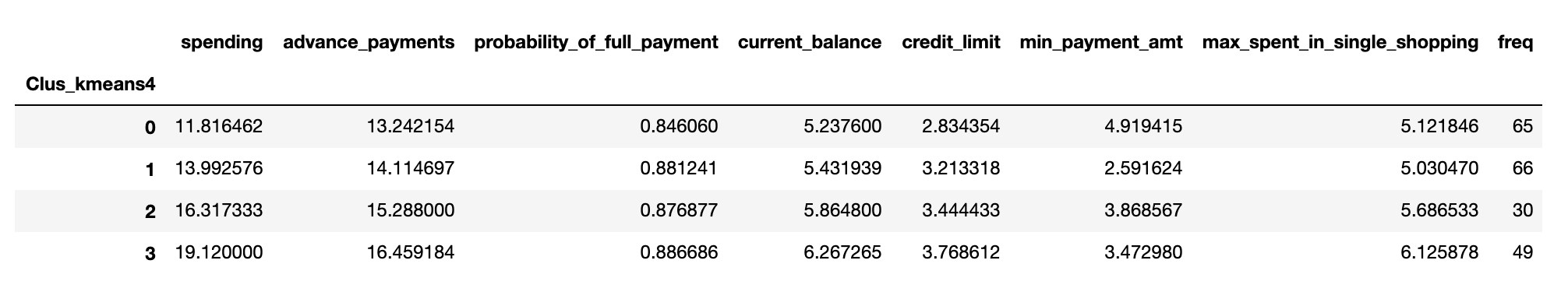
3. Bank can give second preference to the people Tier1. They seems to be spending high compared to Tier 2 but low compared to Tier 3 and their credit limit is low compared to Tier 3 but higher than Tier 2, and the Probability of payment done in full by the customer to the bank is high compared to Tier 3 but low than compared to Tier 2

4. Overall,Banks can give priority to Tier 3 over Tier 2 and 1

*Cluster frequency for K Means Cluster:*

****

*Cluster profile:*

****

Cluster 0: Medium frequency with less spending, advance payments, probability of full payment, balance in the account, credit limit but highest in minimum amount paid by the customer while making payments for purchases made monthly

Cluster 1: Large frequency with less in minimum amount paid by the customer while making payments for purchases made monthly and Maximum amount spent in one purchase

Cluster 2: Small frequency with medium spending,advance\_payments and all other attributes

Cluster 3: Medium frequency with highest spending, advance payments, probability of full payment, balance in the account, credit limit, Maximum amount spent in one purchase.

*Recommendations:*

1. People in Cluster 0 has less spending, advance payments, probability of full payment, balance in the account, credit limit but highest in minimum amount paid by the customer while making payments for purchases made monthly So if bank gives some promotional offers to this segment of people, they may spend more using credit card but still it's risky because the probability of full payment is low.

2. Customers in Cluster 3 seem to be spending more and also are good in paying advance payments and also the probability of full payment is high. Banks can consider this segment of people to give promotional offers. They may spend more and also usage of credit card will be high.