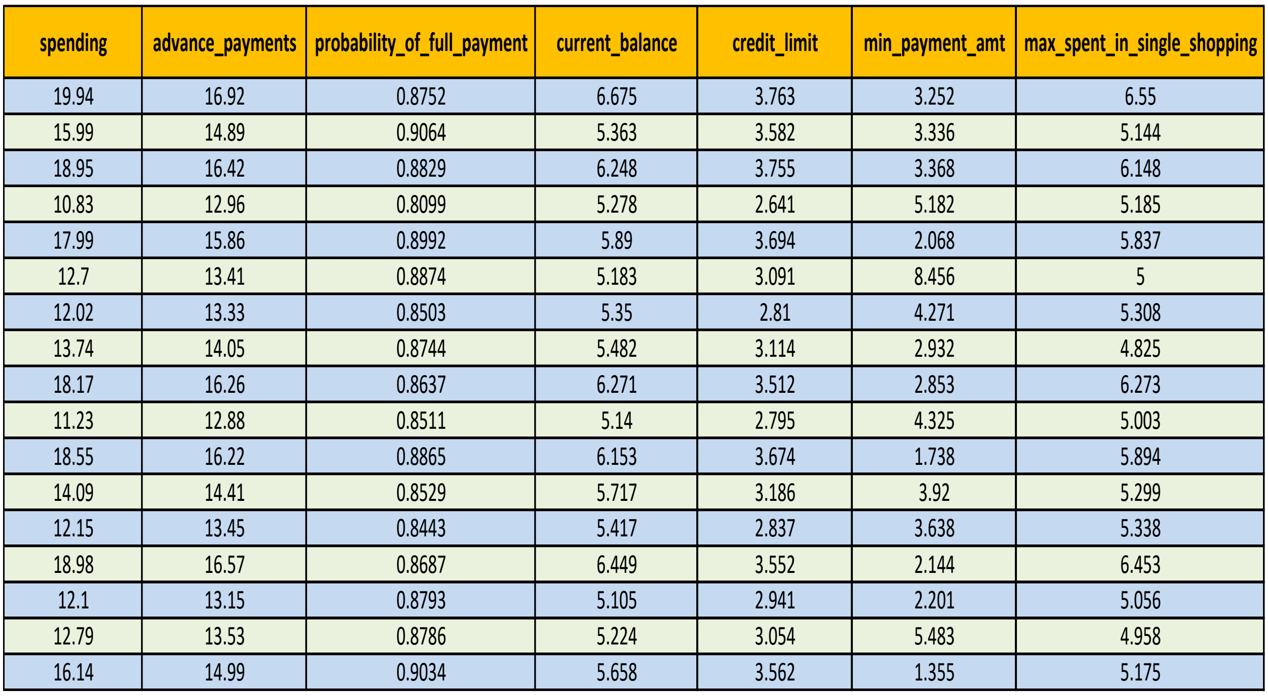
Problem 1 Statement

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.

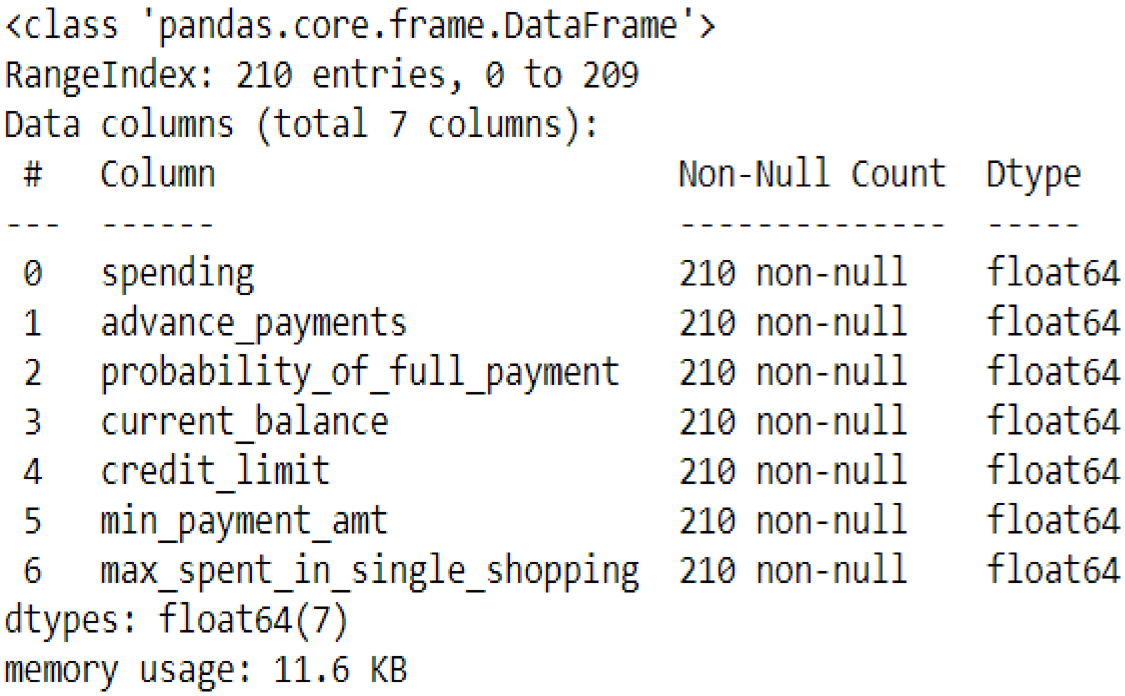
Exploratory Data Analysis :



spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
19.94	16.92	0.8752	6.675	3.763	3.252	6.55
15.99	14.89	0.9064	5.363	3.582	3.336	5.144
18.95	16.42	0.8829	6.248	3.755	3.368	6.148
10.83	12.96	0.8099	5.278	2.641	5.182	5.185
17.99	15.86	0.8992	5.89	3.694	2.068	5.837
12.7	13.41	0.8874	5.183	3.091	8.456	5
12.02	13.33	0.8503	5.35	2.81	4.271	5.308
13.74	14.05	0.8744	5.482	3.114	2.932	4.825
18.17	16.26	0.8637	6.271	3.512	2.853	6.273
11.23	12.88	0.8511	5.14	2.795	4.325	5.003
18.55	16.22	0.8865	6.153	3.674	1.738	5.894
14.09	14.41	0.8529	5.717	3.186	3.92	5.299
12.15	13.45	0.8443	5.417	2.837	3.638	5.338
18.98	16.57	0.8687	6.449	3.552	2.144	6.453
12.1	13.15	0.8793	5.105	2.941	2.201	5.056
12.79	13.53	0.8786	5.224	3.054	5.483	4.958
16.14	14.99	0.9034	5.658	3.562	1.355	5.175

We are provided with the above data set of 210 rows and 7 columns.

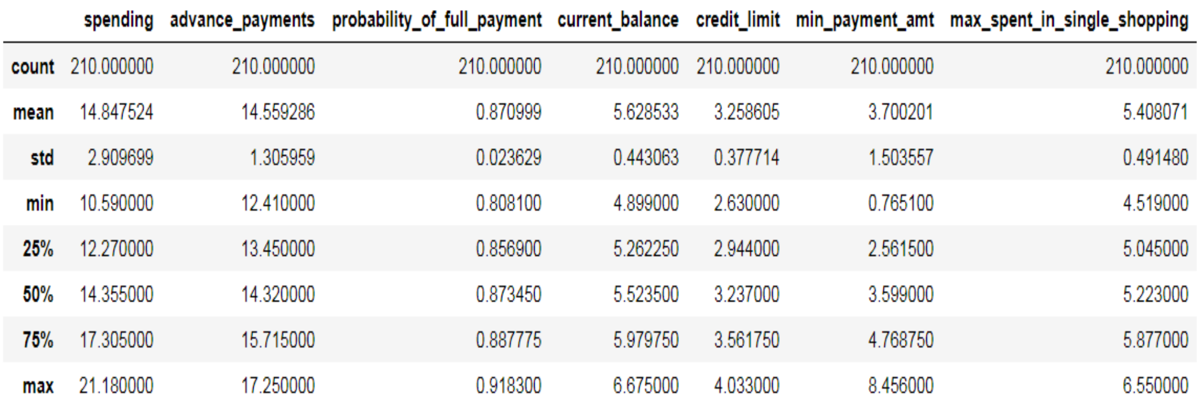
All the available columns are float type date. The data has no Null values.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 210 entries, 0 to 209
Data columns (total 7 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   spending                      210 non-null    float64
 1   advance_payments              210 non-null    float64
 2   probability_of_full_payment   210 non-null    float64
 3   current_balance               210 non-null    float64
 4   credit_limit                  210 non-null    float64
 5   min_payment_amt               210 non-null    float64
 6   max_spent_in_single_shopping  210 non-null    float64
dtypes: float64(7)
memory usage: 11.6 KB
```

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

Descriptive statistics for the dataset:



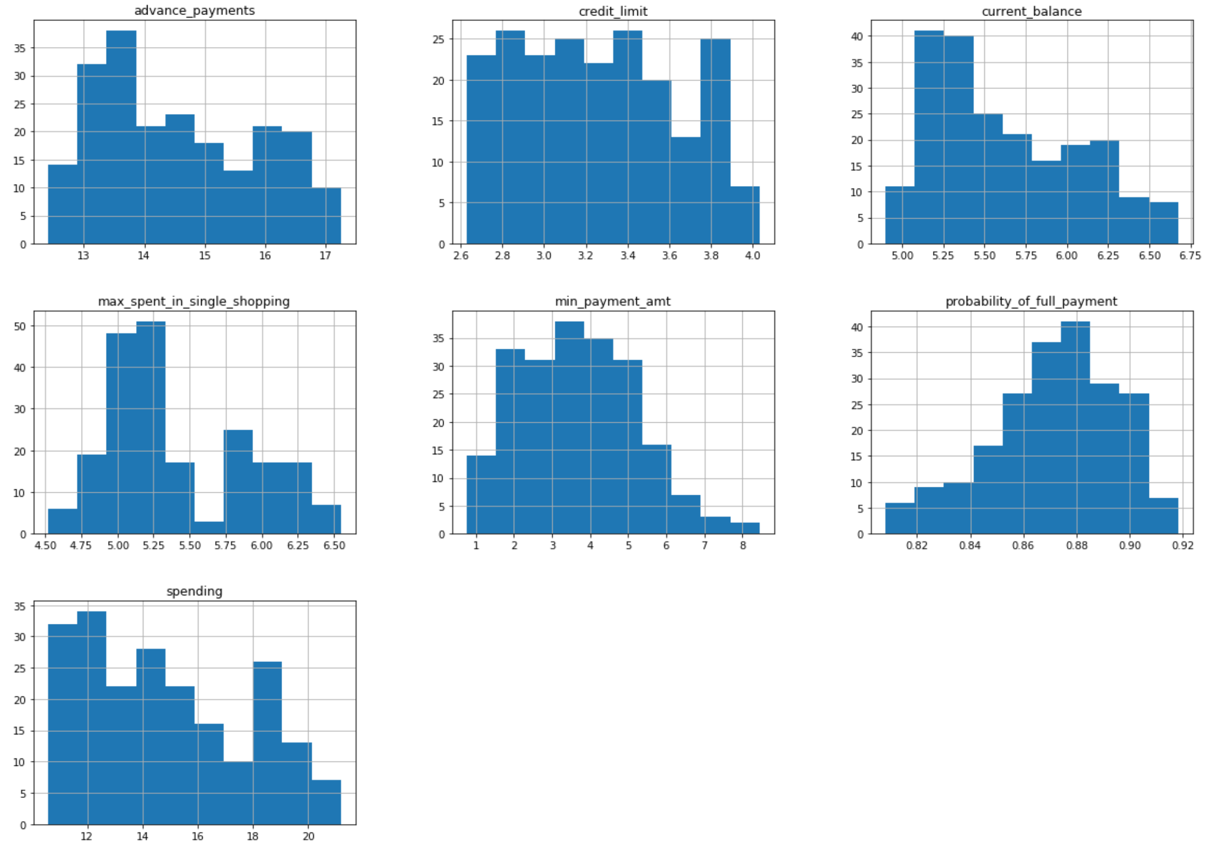
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000	210.000000
mean	14.847524	14.559286	0.870999	5.628533	3.258605	3.700201	5.408071
std	2.909699	1.305959	0.023629	0.443063	0.377714	1.503557	0.491480
min	10.590000	12.410000	0.808100	4.899000	2.630000	0.765100	4.519000
25%	12.270000	13.450000	0.856900	5.262250	2.944000	2.561500	5.045000
50%	14.355000	14.320000	0.873450	5.523500	3.237000	3.599000	5.223000
75%	17.305000	15.715000	0.887775	5.979750	3.561750	4.768750	5.877000
max	21.180000	17.250000	0.918300	6.675000	4.033000	8.456000	6.550000

As per the details resulted from the descriptive statistics of the dataset, we can find that:

All the variables have a **value count** of **210** with **no null values**.

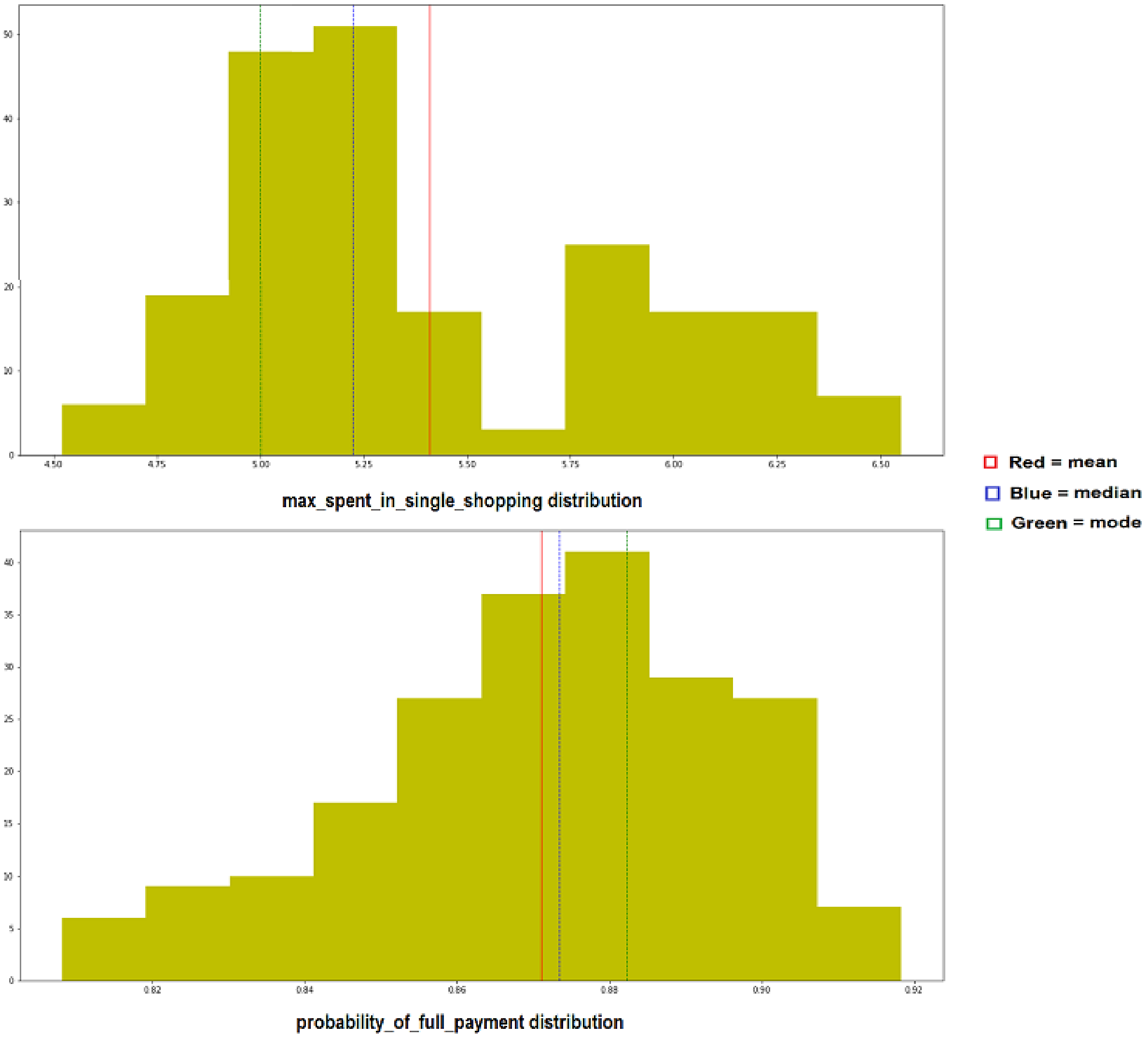
Of the entire dataset, column **'spending'** has **highest max** value of 21.18 and column **'min_payment_amt'** has **least min** value of 0.76510

Columns **'spending'** and **'probability_of_full_payment'** have highest mean value – 14.847524 and least mean value – 0.870999 respectively.



From the above histograms of the variables , we can see that majority of the variables are not symmetrical.

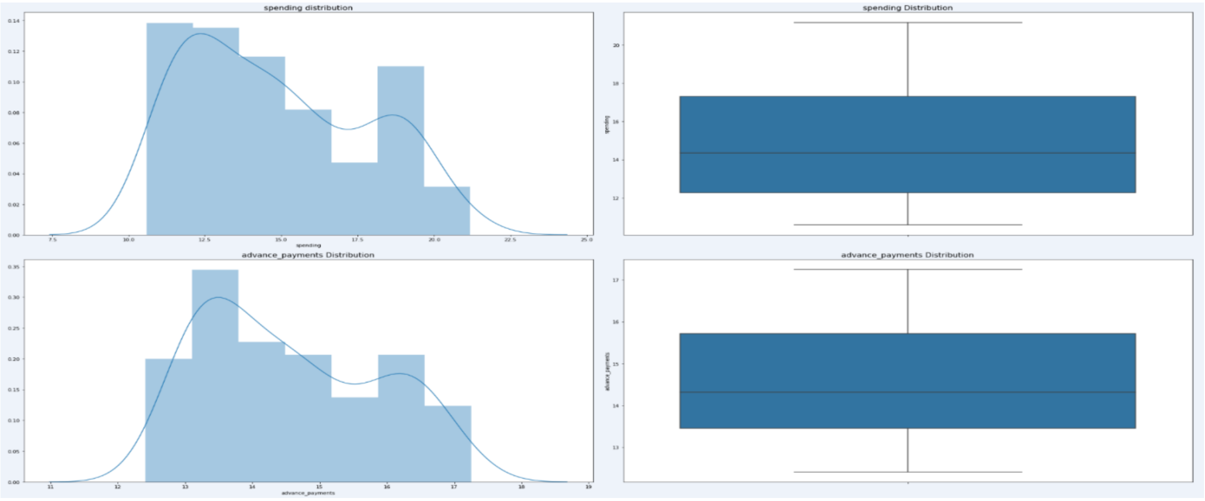
Among all the variables , Variable **‘max_spent_in_single_shopping’** is highly **right skewed** (skew = 0.561897) and Variable **‘probability_of_full_payment’** is highly **left skewed** (skew = -0.537954).

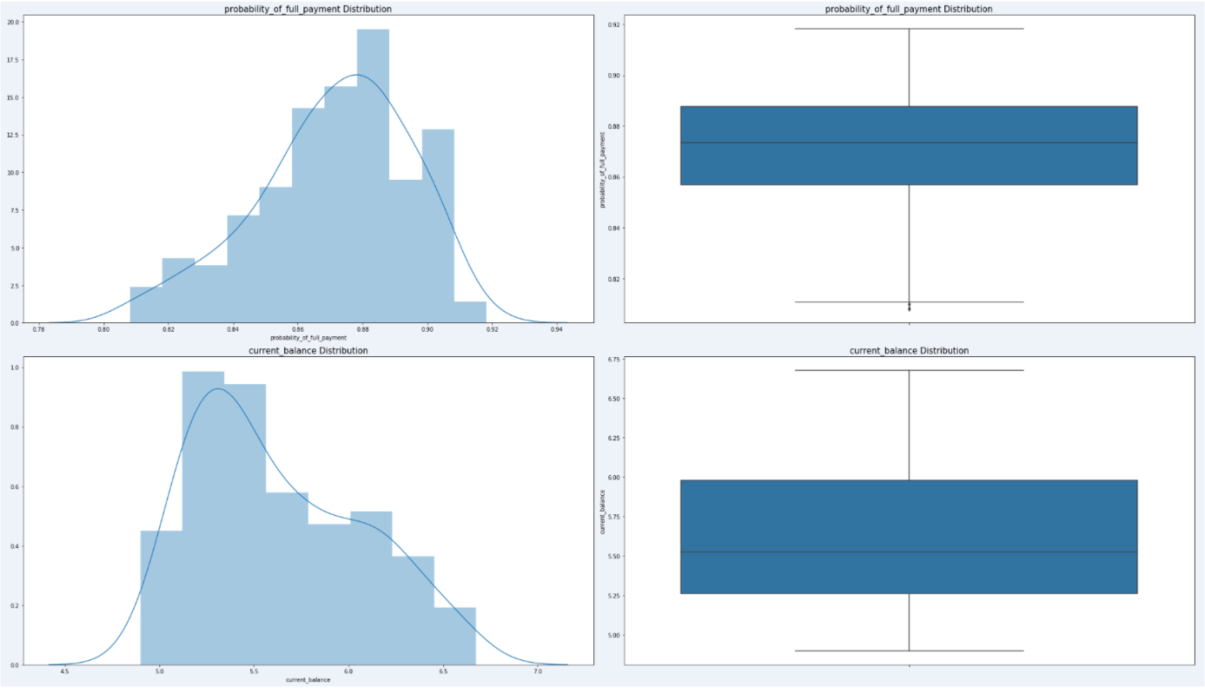


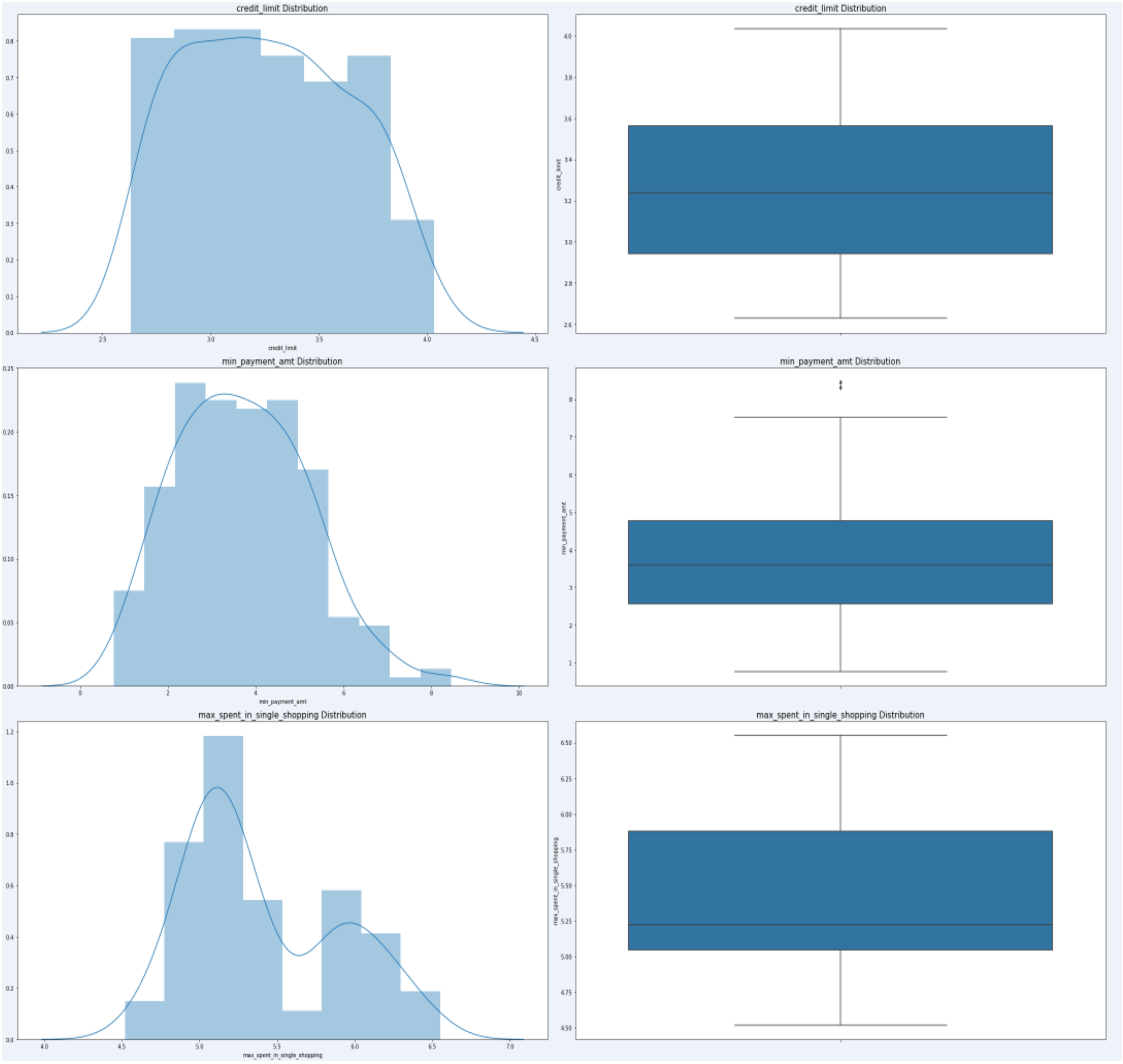
max_spent_in_single_shopping distribution

probability_of_full_payment distribution

Among the given variables, ‘**probability_of_full_payment**’ and ‘**min_payment_amt**’ are the only variables with outliers.

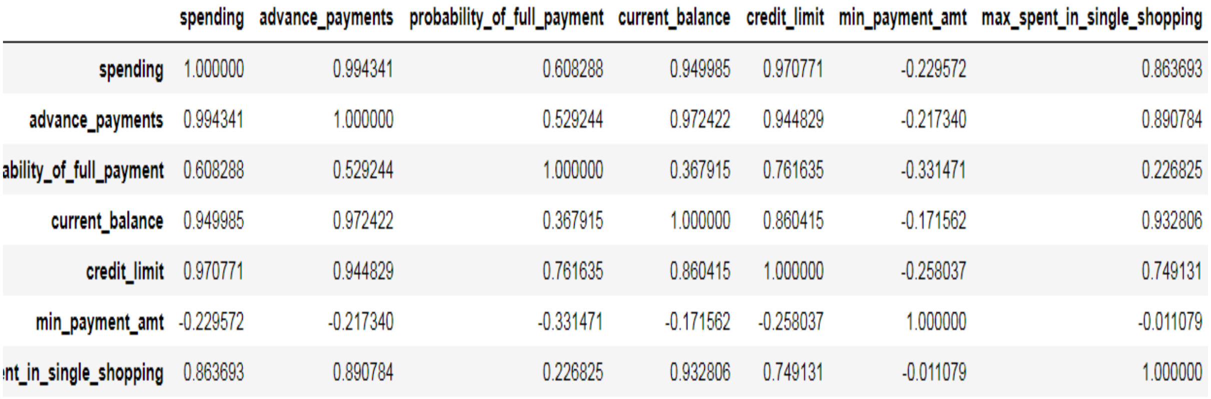



probability_of_full_payment Distribution
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
0.78
0.80
0.82
0.84
0.86
0.88
0.90
0.92
0.94
probability_of_full_payment
probability_of_full_payment Distribution
0.92
0.90
0.88
0.86
0.84
0.82
probability_of_full_payment
current_balance Distribution
1.0
0.8
0.6
0.4
0.2
0.0
4.5
5.0
5.5
6.0
6.5
7.0
current_balance
current_balance Distribution
6.75
6.50
6.25
6.00
5.75
5.50
5.25
5.00
current_balance


credit_limit Distribution
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
2.5
3.0
3.5
4.0
4.5
credit_limit
credit_limit Distribution
4.0
3.8
3.6
3.4
3.2
3.0
2.8
2.6
credit_limit
min_payment_amt Distribution
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
min_payment_amt
min_payment_amt Distribution
8
7
6
5
4
3
2
1
min_payment_amt
max_spent_in_single_shopping Distribution
1.2
1.0
0.8
0.6
0.4
0.2
0.0
4.0
4.5
5.0
5.5
6.0
6.5
7.0
max_spent_in_single_shopping
max_spent_in_single_shopping Distribution
6.50
6.25
6.00
5.75
5.50
5.25
5.00
4.75
4.50
max_spent_in_single_shopping

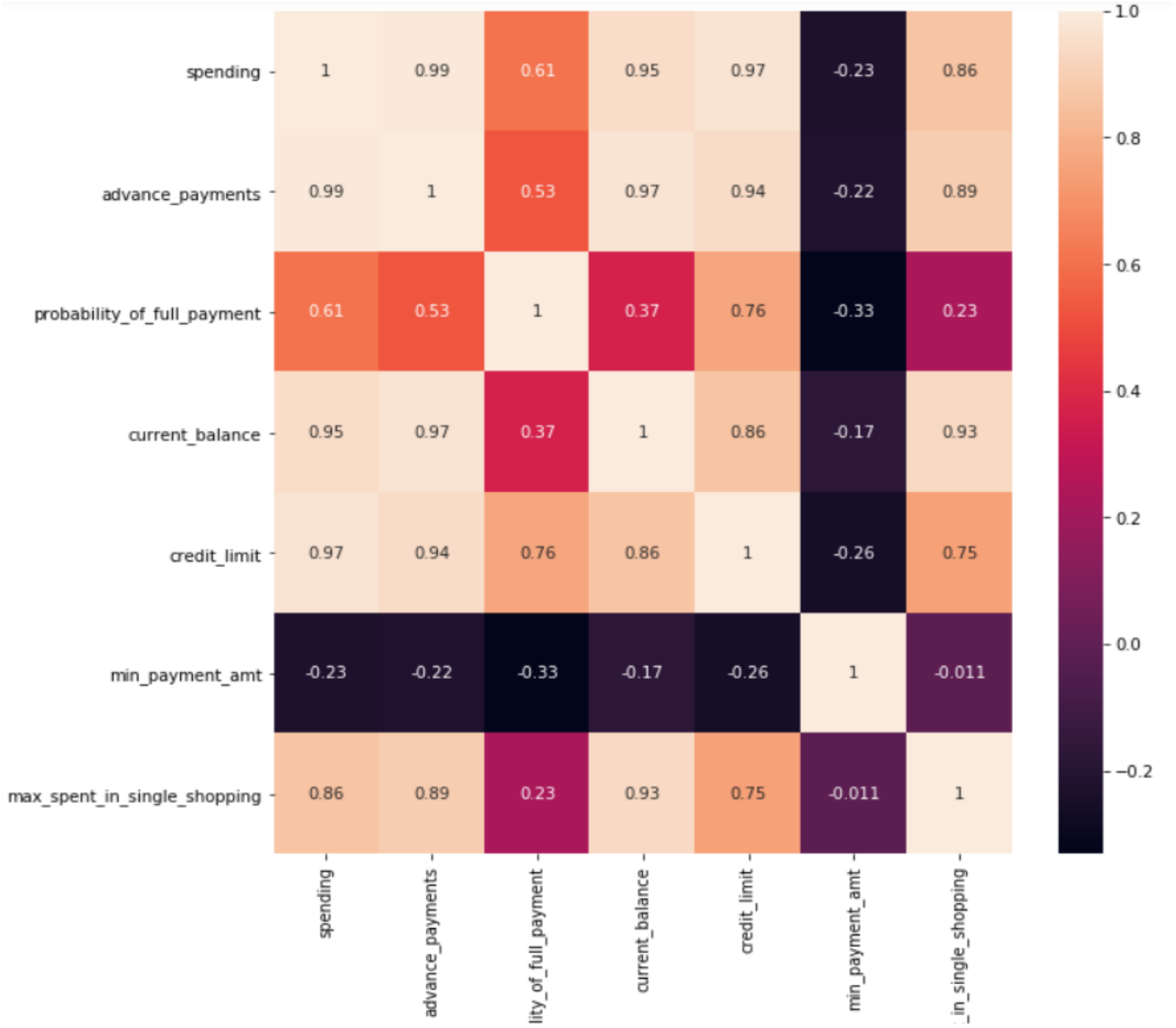
Multivariate Analysis:

We have the following correlation among the different variables given in the dataset.



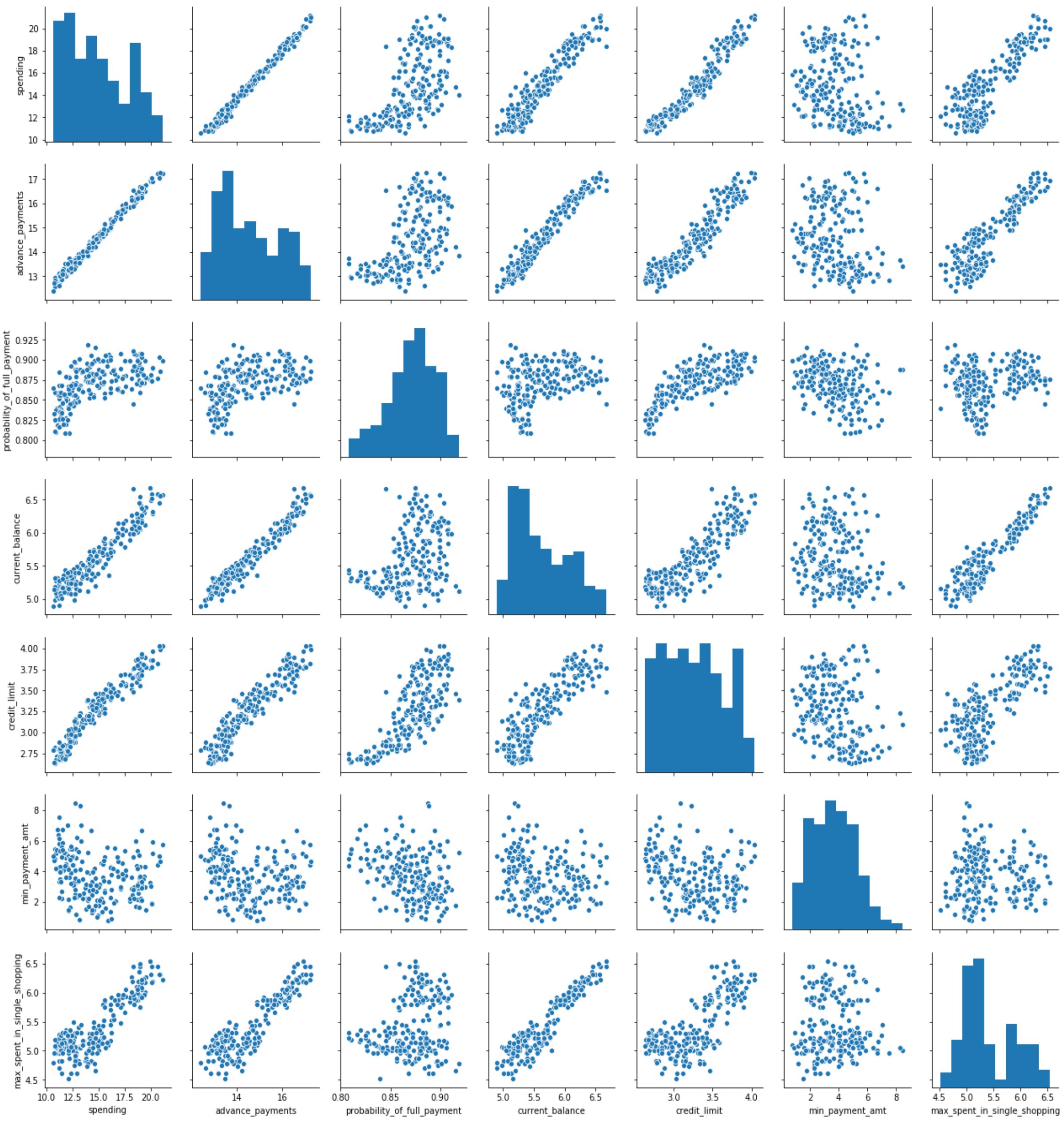
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
spending	1.000000	0.994341	0.608288	0.949985	0.970771	-0.229572	0.863693
advance_payments	0.994341	1.000000	0.529244	0.972422	0.944829	-0.217340	0.890784
ability_of_full_payment	0.608288	0.529244	1.000000	0.367915	0.761635	-0.331471	0.226825
current_balance	0.949985	0.972422	0.367915	1.000000	0.860415	-0.171562	0.932806
credit_limit	0.970771	0.944829	0.761635	0.860415	1.000000	-0.258037	0.749131
min_payment_amt	-0.229572	-0.217340	-0.331471	-0.171562	-0.258037	1.000000	-0.011079
nt_in_single_shopping	0.863693	0.890784	0.226825	0.932806	0.749131	-0.011079	1.000000

HeatMap:



From the above map , we can see that many columns are co-related to each other and there is **highest positive correlation** (0.99) between variables **'advance_payments'** and **'spending'** . Also there is **highest negative correlation**(- 0.33) between variables **'min_payment_amt'** and **'probability_of_full_payment'.**

Pairplot:



In the above plot scatter diagrams are plotted for all the numerical columns in the dataset. From the visual representation , we can understand the degree of correlation between any two columns of the given dataset.

Variables **‘advance_payments’ , ‘current_balance’ , ‘credit_limit’ , ‘max_spent_in_single_shopping’** show high positive linear correlation with Variable **‘spending’**.

Variables **‘current_balance’ , ‘credit_limit’ , ‘max_spent_in_single_shopping’** show high positive linear correlation with Variable **‘advance_payments’**.

Variables **‘credit_limit’** shows high positive linear correlation with Variable **‘probability_of_full_payment’**.

Variables **‘credit_limit’ , ‘max_spent_in_single_shopping’** show high positive linear correlation with Variable **‘current_balance’**.

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

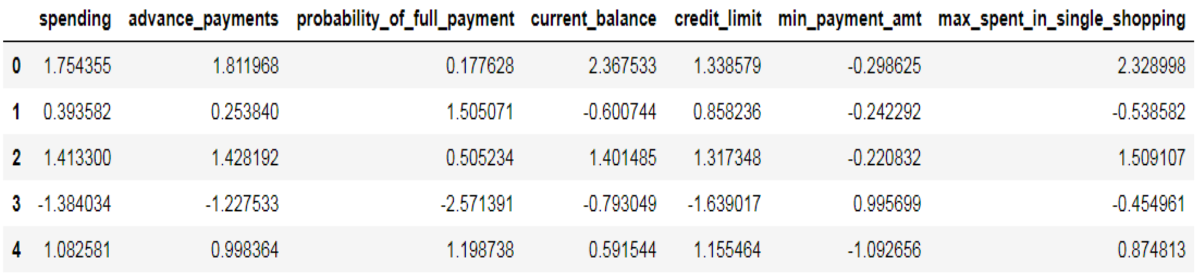
We know that Clustering techniques use Euclidean distance to group different rows of the given dataset. If there is high difference in the magnitudes of the data in the given columns , then the calculated distances can be affected by the high magnitude values i.e, model might become biased towards the variables with higher magnitude. Even sometimes scaling helps in speeding up the calculations in an algorithm.

Moreover scaling helps in controlling the variability of the dataset and will be helpful in improving the accuracy of clustering algorithms.

As most of the variables are skewed towards either right or left and also there is much difference in the magnitude of value in the dataset between values among different variables. So it will be wise to scale the variables before data is fed to the algorithms.

We prefer to use z scale method to minimize the effect of both skewness and variability due to difference between high and low magnitude value of the data.

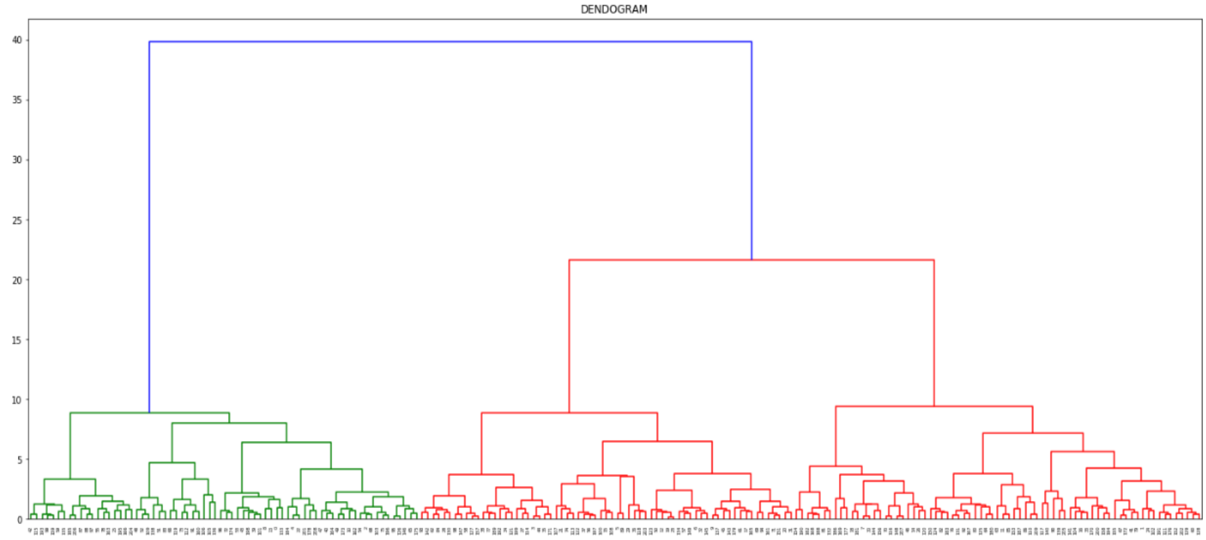
Scaled Variables :



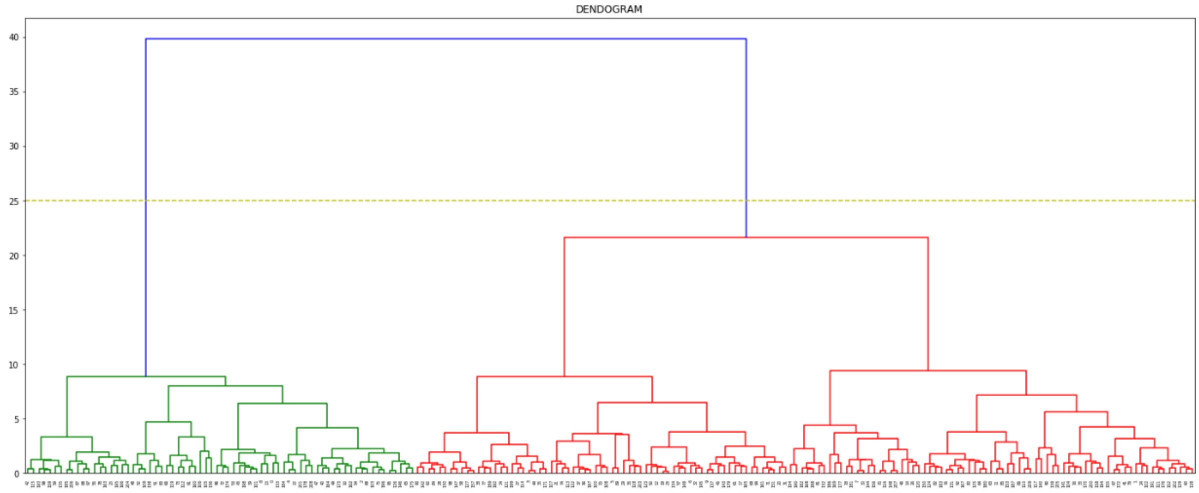
	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1.754355	1.811968	0.177628	2.367533	1.338579	-0.298625	2.328998
1	0.393582	0.253840	1.505071	-0.600744	0.858236	-0.242292	-0.538582
2	1.413300	1.428192	0.505234	1.401485	1.317348	-0.220832	1.509107
3	-1.384034	-1.227533	-2.571391	-0.793049	-1.639017	0.995699	-0.454961
4	1.082581	0.998364	1.198738	0.591544	1.155464	-1.092656	0.874813

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

We have got the following dendrogram for the after performing the Hierarchical Clustering.

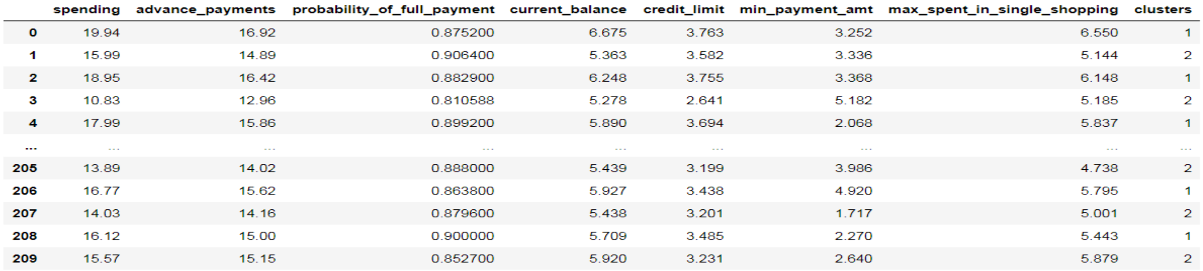


The x-axis contains the samples and y-axis represents the distance between these samples. The vertical line with **maximum distance** is the **blue line** and hence we can decide a threshold of 25 and cut the dendrogram.



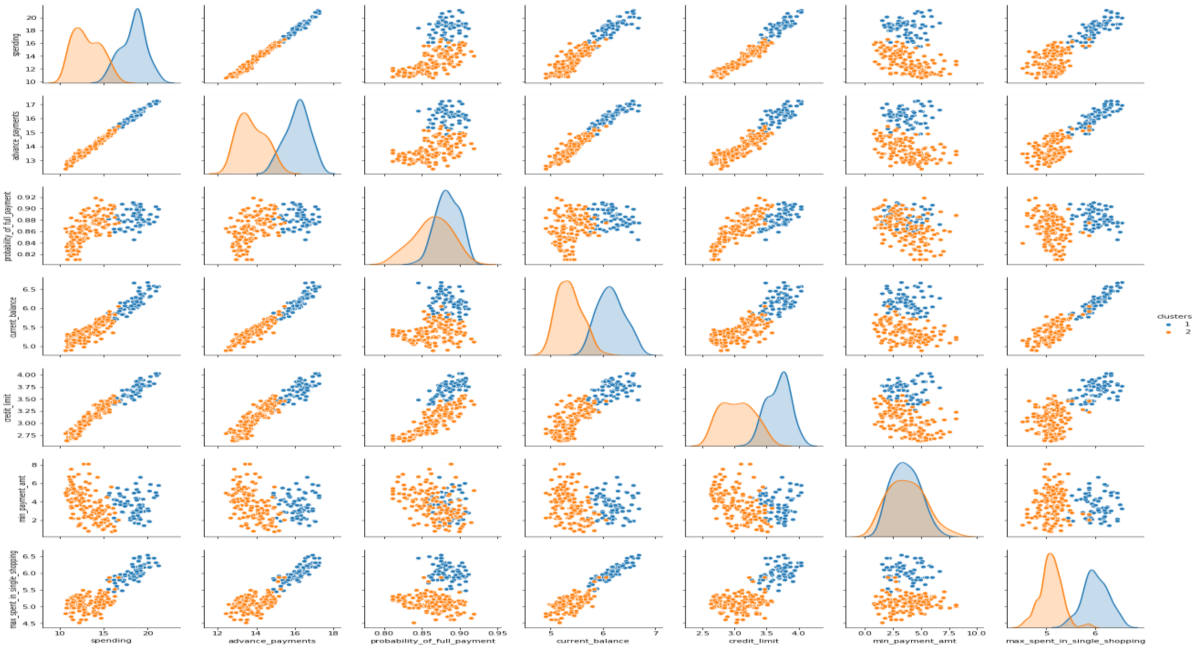
We have **two clusters** as this line cuts the dendrogram at **two points**.

Finally ,below is the dataset showing cluster associated with each row:



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	clusters
0	19.94	16.92	0.875200	6.675	3.763	3.252	6.550	1
1	15.99	14.89	0.906400	5.363	3.582	3.336	5.144	2
2	18.95	16.42	0.882900	6.248	3.755	3.368	6.148	1
3	10.83	12.96	0.810588	5.278	2.641	5.182	5.185	2
4	17.99	15.86	0.899200	5.890	3.694	2.068	5.837	1
...	...	...	...	...	...	...	...	...
205	13.89	14.02	0.888000	5.439	3.199	3.986	4.738	2
206	16.77	15.62	0.863800	5.927	3.438	4.920	5.795	1
207	14.03	14.16	0.879600	5.438	3.201	1.717	5.001	2
208	16.12	15.00	0.900000	5.709	3.485	2.270	5.443	1
209	15.57	15.15	0.852700	5.920	3.231	2.640	5.879	2

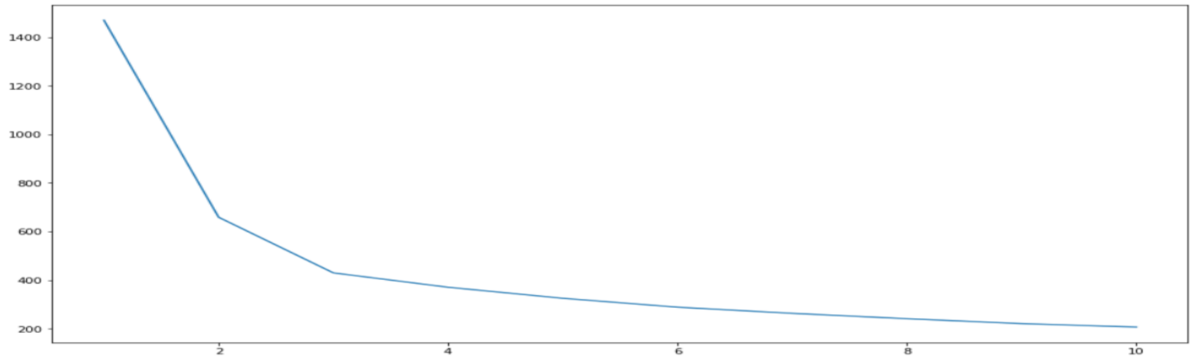
Pairplot of clustered data:



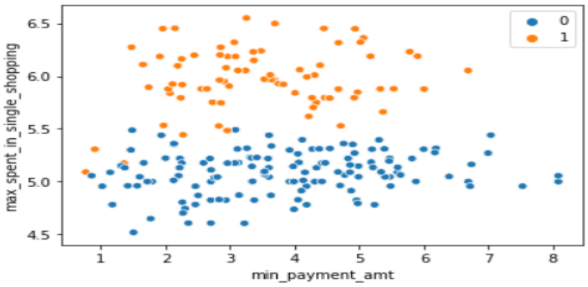
We can see that there is significant difference among the clusters 1 & 2 for all the variables except for the variable **'min_payment_amt'**. For the given data set, we can see that 66.66% of the rows belong to cluster 2 and remaining 33.33% belongs to cluster 01.

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

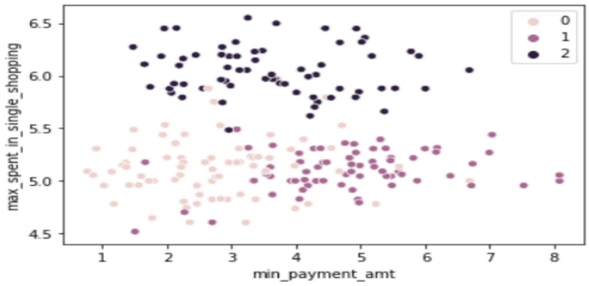
From the below plot, we can notice that the within sum of squares(wss) is not significantly dropping after value of cluster =3 , we can consider **three optimum clusters** by the **elbow curve method**.



But when we calculate the Silhouette Score to check the optimal number of clusters. We can see that the optimal number of clusters is **2(0.46560100442748986)** as its silhouette score is greater than that of **3 clusters (0.40080 59221522216).**



2 Clusters

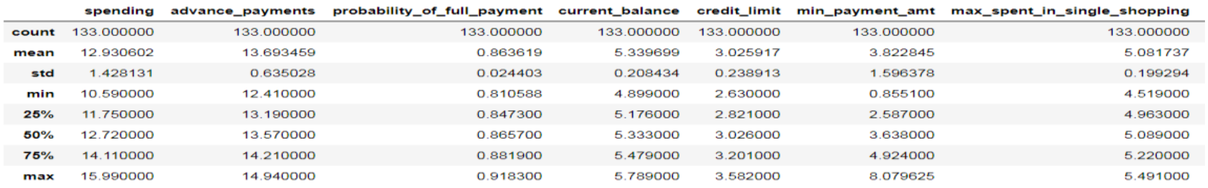


3 Clusters

Even from the above figures ,we can conclude that clustering is done with high accuracy into 2 rather than into 3.

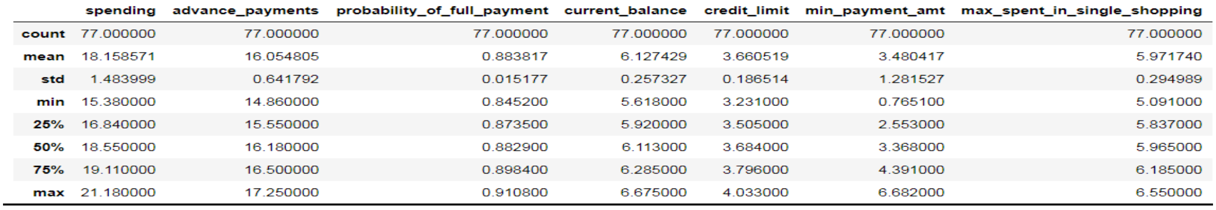
1.5) Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

Cluster 01:



	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	133.000000	133.000000	133.000000	133.000000	133.000000	133.000000	133.000000
mean	12.930602	13.693459	0.863619	5.339699	3.025917	3.822845	5.081737
std	1.428131	0.635028	0.024403	0.208434	0.238913	1.596378	0.199294
min	10.590000	12.410000	0.810588	4.899000	2.630000	0.855100	4.519000
25%	11.750000	13.190000	0.847300	5.176000	2.821000	2.587000	4.963000
50%	12.720000	13.570000	0.865700	5.333000	3.026000	3.638000	5.089000
75%	14.110000	14.210000	0.881900	5.479000	3.201000	4.924000	5.220000
max	15.990000	14.940000	0.918300	5.789000	3.582000	8.079625	5.491000

Cluster 02:

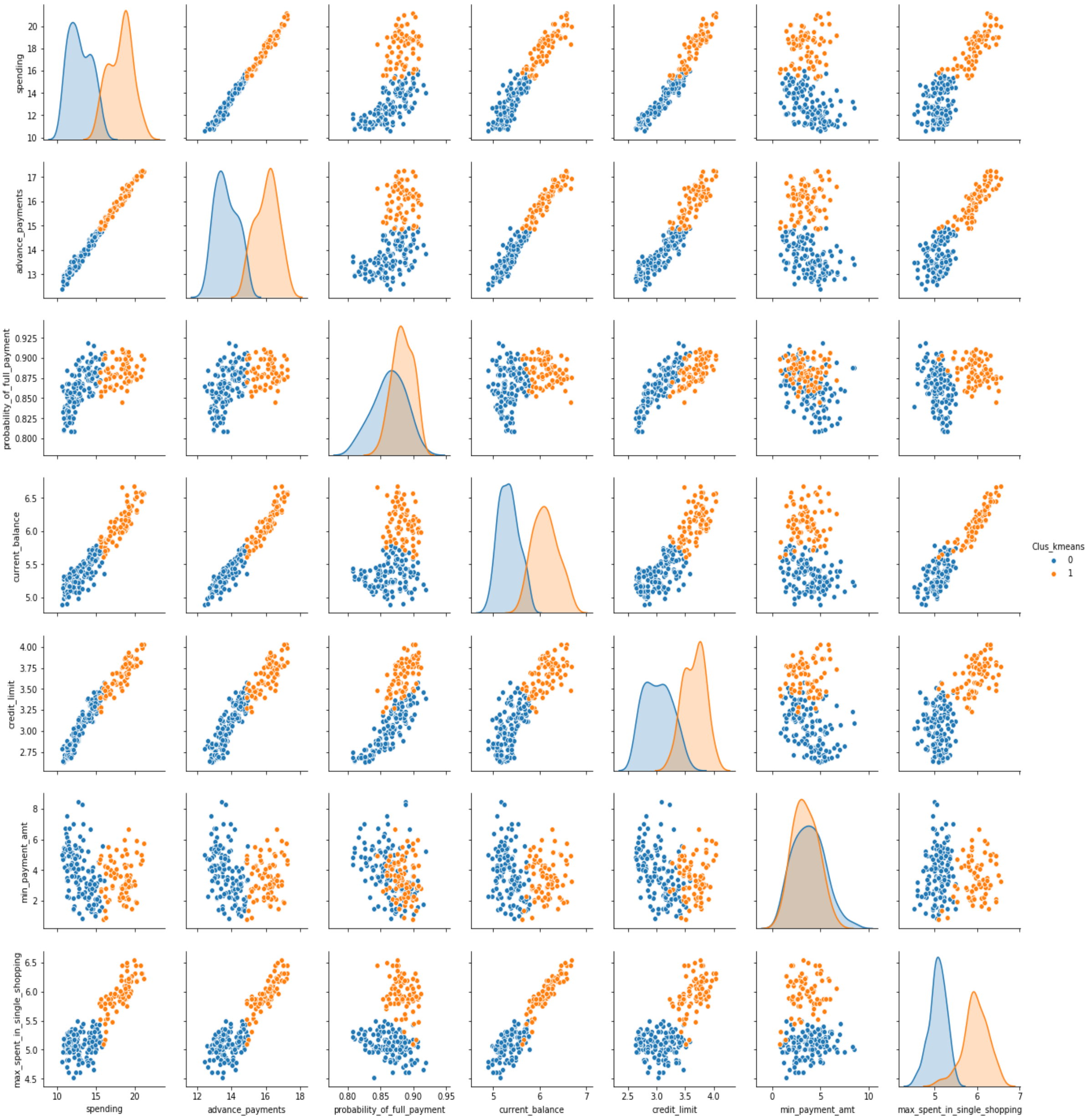


	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000	77.000000
mean	18.158571	16.054805	0.883817	6.127429	3.660519	3.480417	5.971740
std	1.483999	0.641792	0.015177	0.257327	0.186514	1.281527	0.294989
min	15.380000	14.860000	0.845200	5.618000	3.231000	0.765100	5.091000
25%	16.840000	15.550000	0.873500	5.920000	3.505000	2.553000	5.837000
50%	18.550000	16.180000	0.882900	6.113000	3.684000	3.368000	5.965000
75%	19.110000	16.500000	0.898400	6.285000	3.796000	4.391000	6.185000
max	21.180000	17.250000	0.910800	6.675000	4.033000	6.682000	6.550000

From the above descriptive stats of the cluster01 & cluster 02, we have following insights ,

Description	Cluster 01	Cluster 02
spending: Amount spent by the customer per month	Lower (12.93)	Higher (18.15)
advance_payments: Amount paid by the customer in advance by cash	Lower (13.69)	Higher (16.05)
probability_of_full_payment: Probability of payment done in full by the customer to the bank	Lower (0.86)	Higher (0.88)
current_balance: Balance amount left in the account to make purchases	Lower (5.33)	Higher (6.12)
credit_limit	Lower (3.02)	Higher (3.66)
min_payment_amt : minimum paid by the customer while making payments for purchases made monthly	Higher (3.82)	Lower (3.48)
max_spent_in_single_shopping: Maximum amount spent in one purchase	Lower (5.08)	Higher (5.97)

From the Pair plot, we can see the difference between Cluster1 & Cluster2 with respect to different variables.



There is no significant difference in minimum paid by the customer while making payments for purchases made monthly for both Clusters.

Cluster 02 values for most of the variables are on higher side compared to cluster 01

So, we can conclude that Cluster 02 is High spending group compared to Cluster 01

Cluster 02 group has high balance amount left in the account to make purchases compared to cluster01.

Probability of payment done in full by the customer to the bank is almost at same average level for both the clusters.

Maximum amount spent in one purchase is in high linear correlation with balance amount left in the account to make purchases and is on higher side for cluster 02.

Amount spent by the customer per month is in high linear correlation with credit limit, so increasing credit limit would help in increasing spending for both clusters.

Therefore, different strategies have to be adopted to push cluster 01 group towards high spending without neglecting the cluster 02.