|  |
| --- |
|  |

|  |
| --- |
| **Project - Data Mining** |
| **Project-4 DSBA - Data Science & Business Analytics** |
|  |
| **The Answer Book contains answers to two Problem statements. Problem 1 deals with credit usage of credit usage of customer segments & Problem 2 is about the claim frequency of a tour insurance company. Solution of Problem 1 is based on Clustering while in Problem 2, it’s based on concepts of CART-Random Forest-Artificial Neural Network.** |
|  |
| **Gunjar Fuley** |
| **11th April 2021** |
|  |

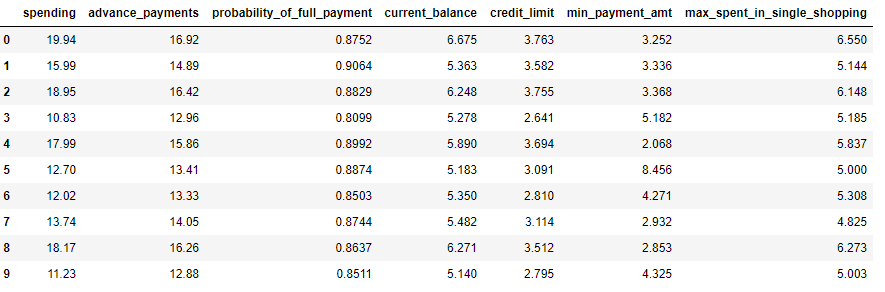
|  |  |  |  |
| --- | --- | --- | --- |

**Problem 1: Clustering**

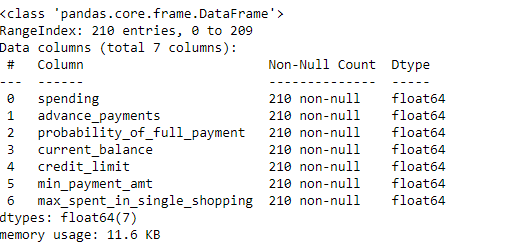
A leading bank wants to develop a customer segmentation to give promotional offers to its customers. They collected a sample that summarizes the activities of users during the past few months. You are given the task to identify the segments based on credit card usage.

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

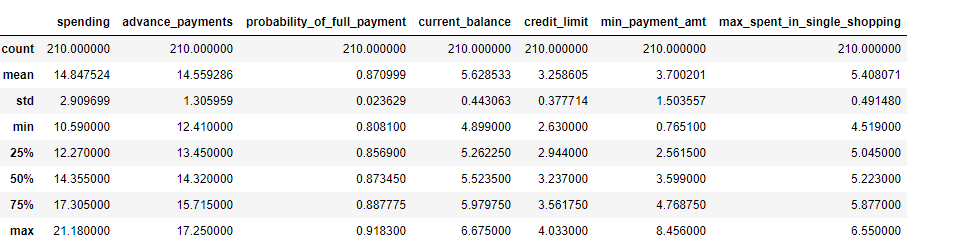
1. spending: Amount spent by the customer per month (in 1000s)
2. advance\_payments: Amount paid by the customer in advance by cash (in 100s)
3. probability\_of\_full\_payment: Probability of payment done in full by the customer to the bank
4. current\_balance: Balance amount left in the account to make purchases (in 1000s)
5. credit\_limit: Limit of the amount in credit card (10000s)
6. min\_payment\_amt : minimum paid by the customer while making payments for purchases made monthly (in 100s)
7. max\_spent\_in\_single\_shopping: Maximum amount spent in one purchase (in 1000s)
   1. Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).



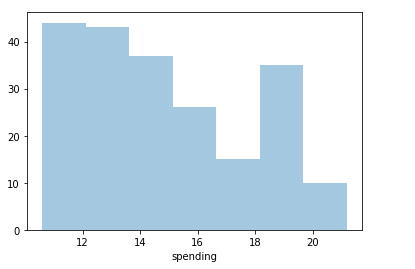
By checking the initial 10 data points we can understand the data file is read & uploaded well into the python jupyter notebook.

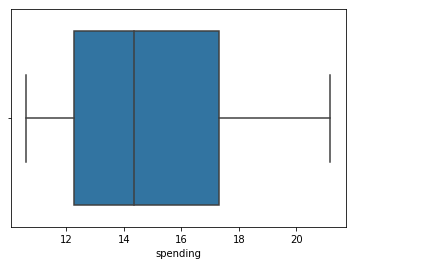


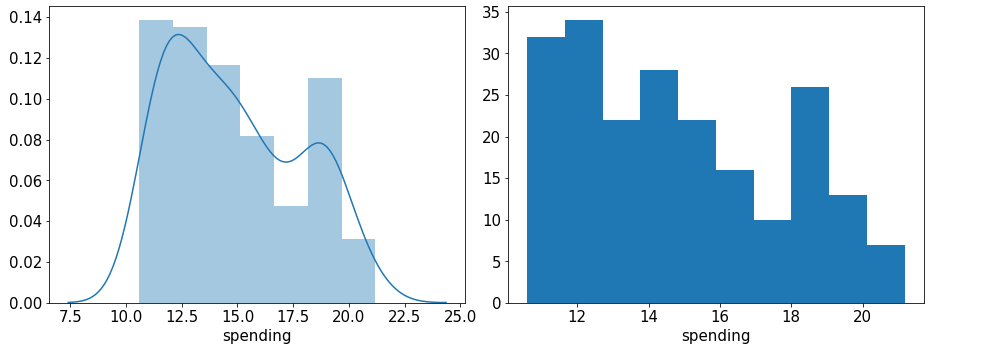
Info clearly tells us that there’s no null data in the dataset.



**Evaluating 'Spending'**

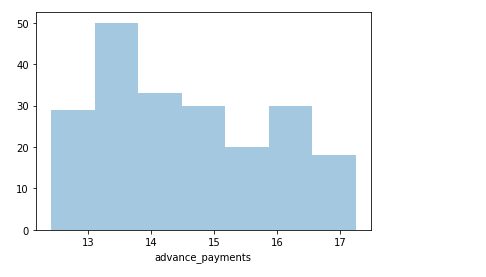


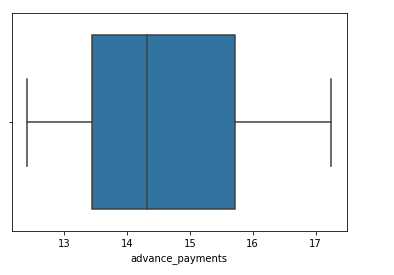


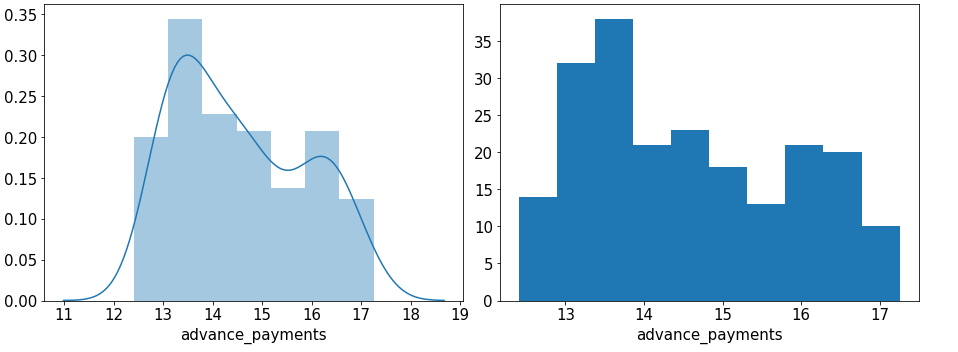


We can see that spending variable is right skewed in nature. There’s no outlier.

# Evaluating 'advance\_payments'

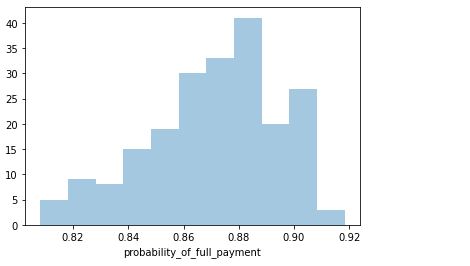


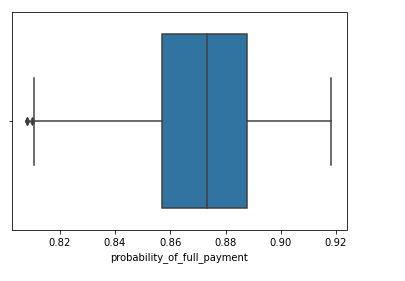


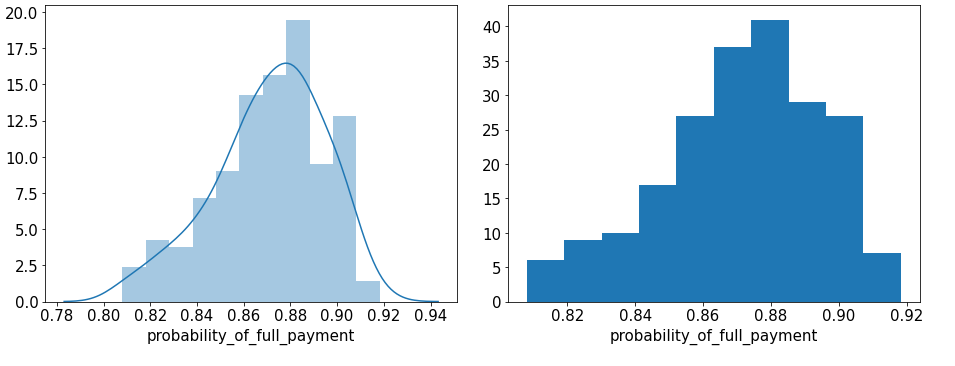


We can see that advance\_payments variable is right skewed in nature. There’s no outlier.

# Evaluating 'probability\_of\_full\_payment'

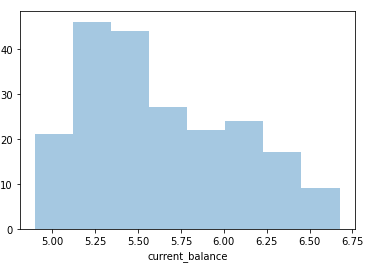


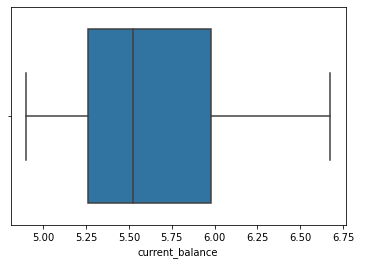


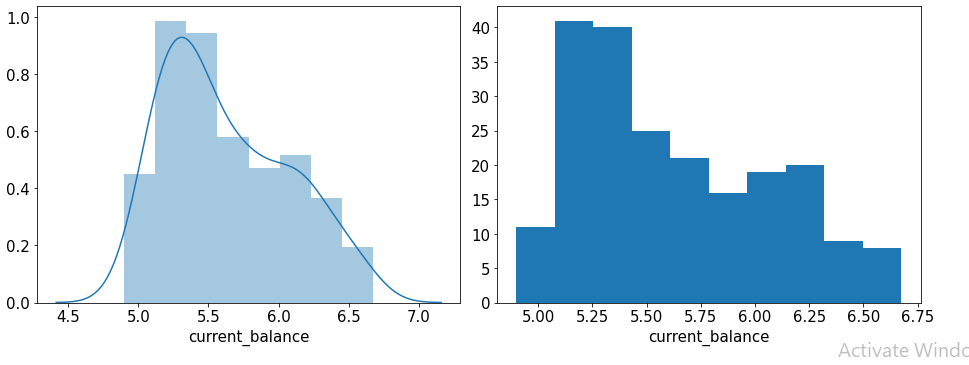


We can see that spending variable is left skewed in nature. Outlier present.

# Evaluating 'current\_balance'

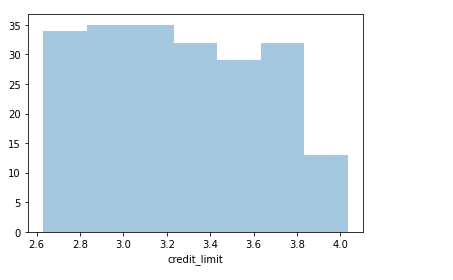


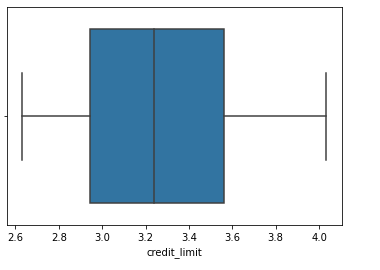


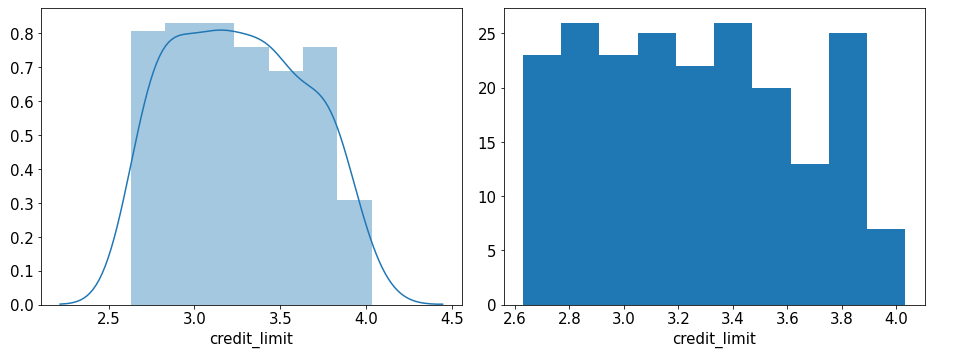


We can see that current\_balance variable is right skewed in nature. There’s no outlier.

# Evaluating 'credit\_limit'

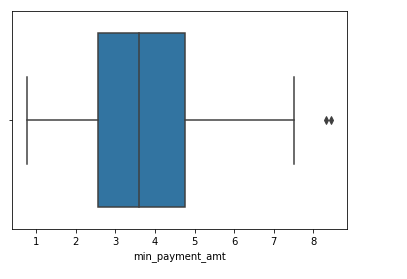


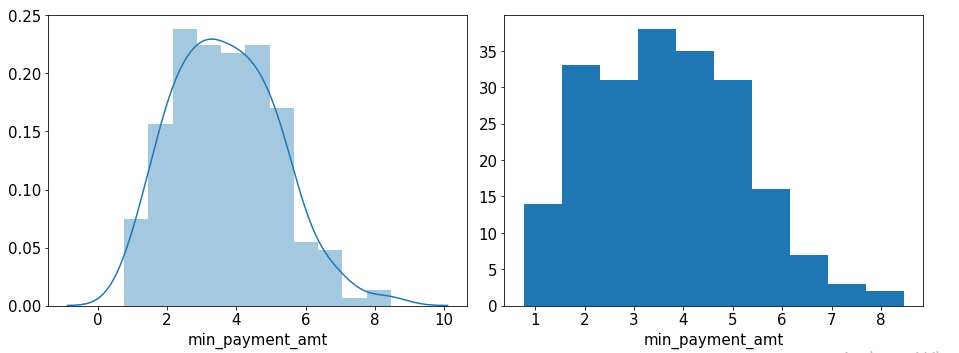




We can see that credit\_limit variable is right skewed in nature. There’s no outlier.

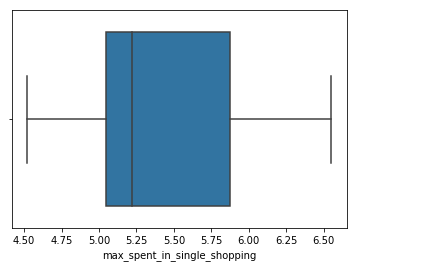
# Evaluating 'min\_payment\_amt'

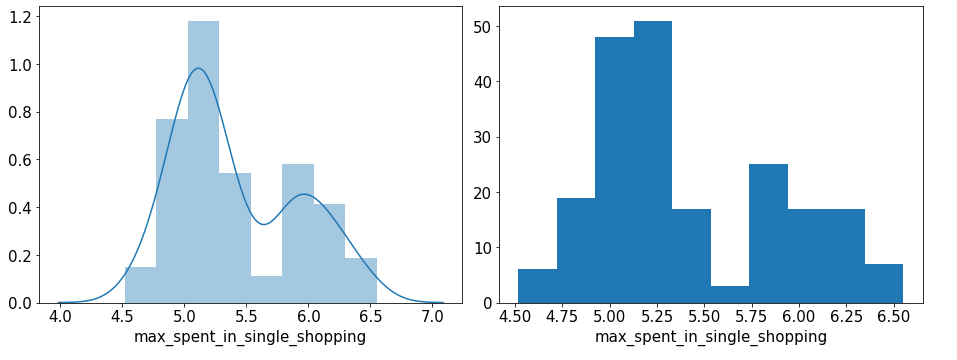




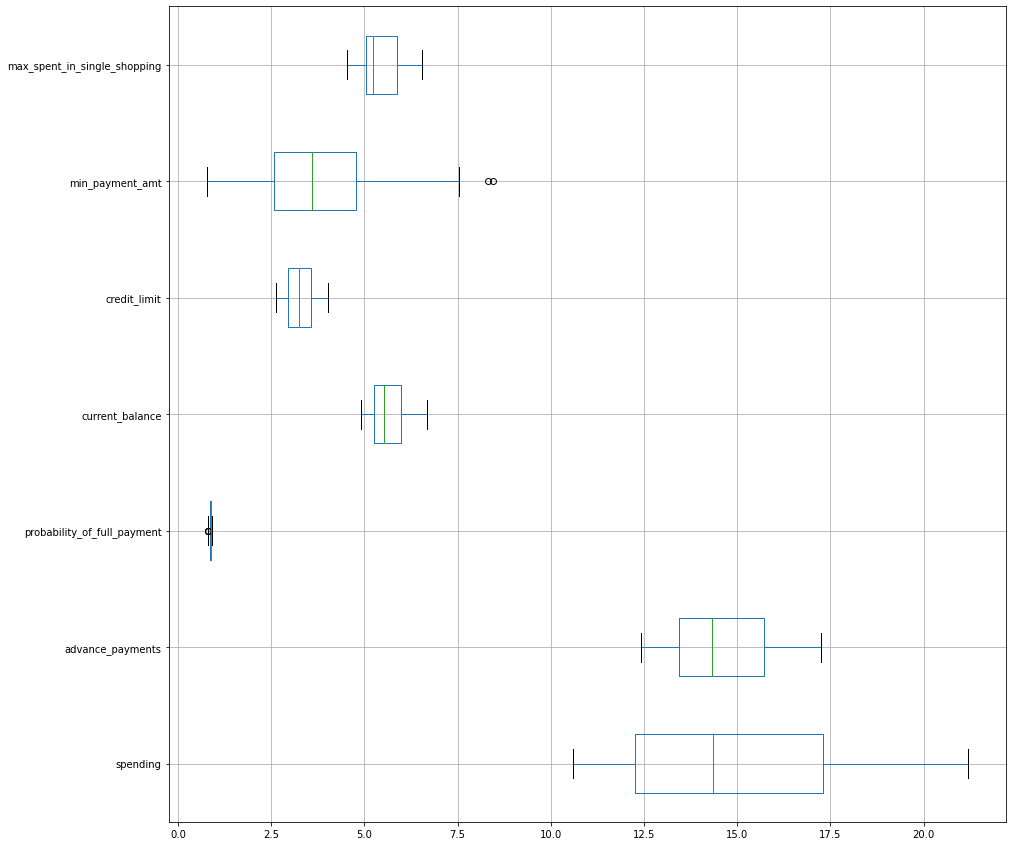
We can see that min\_payment\_amt variable is right skewed in nature. Outlier present.

# Evaluating 'max\_spent\_in\_single\_shopping'



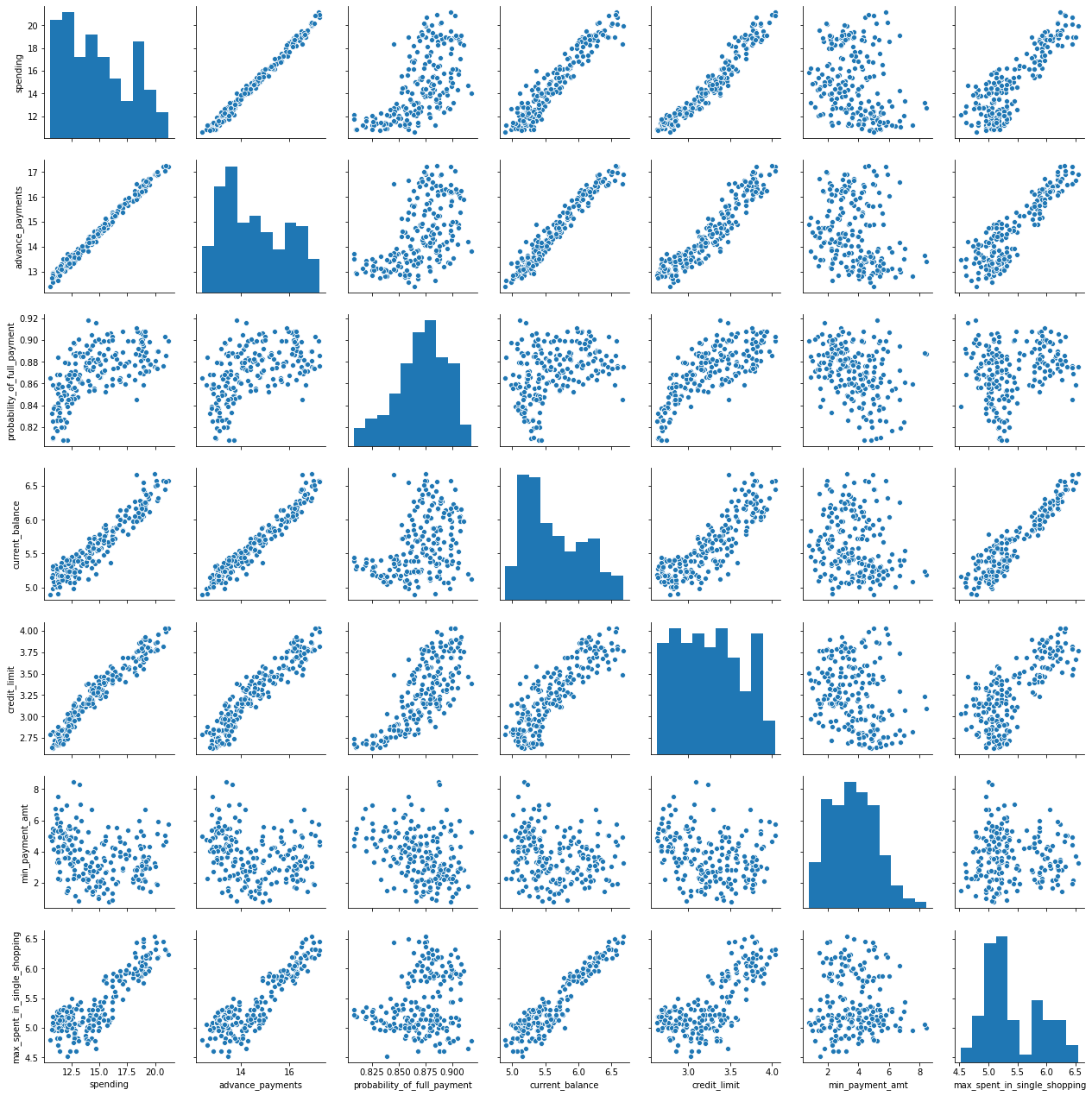


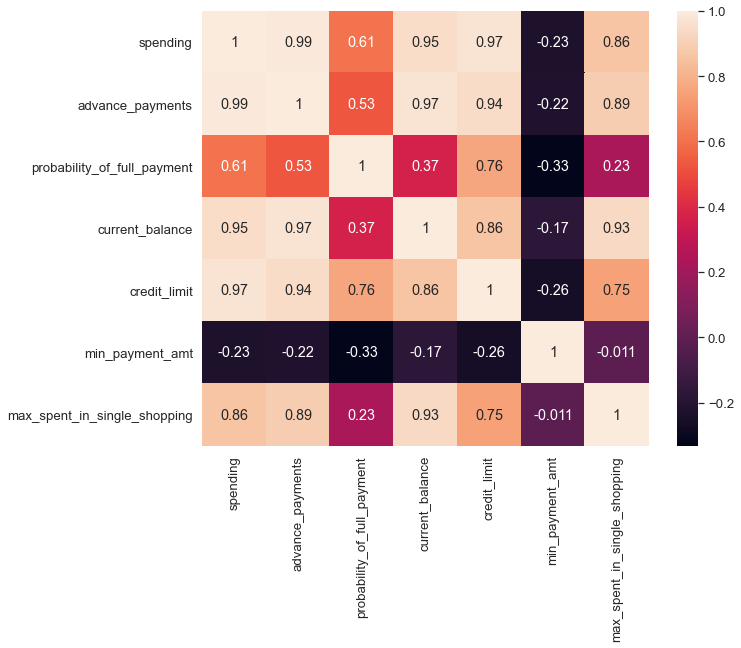
We can see that max\_spent\_in\_single\_shopping variable is right skewed in nature. There’s no outlier.



From the above analysis we can see outliers are present in probability\_of\_full\_payment & min\_payment\_amt

# Multivariate analysis





We can clearly see there is correlation between the following variables-

- spending & current\_balance

- spending & credit\_limit

- spending & advance\_payment

- advance\_payments & credit\_limit

- advance\_payments & current\_balance

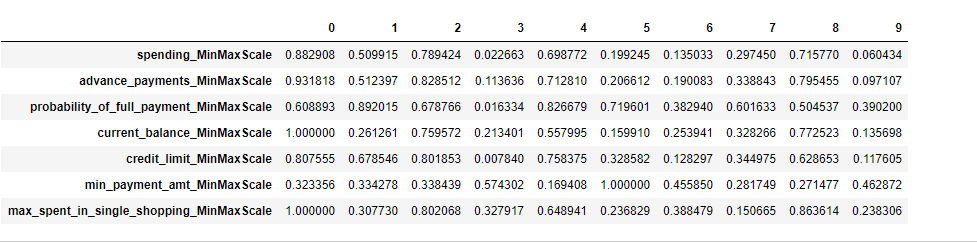
- current\_balance & max\_spent\_in\_single\_shopping

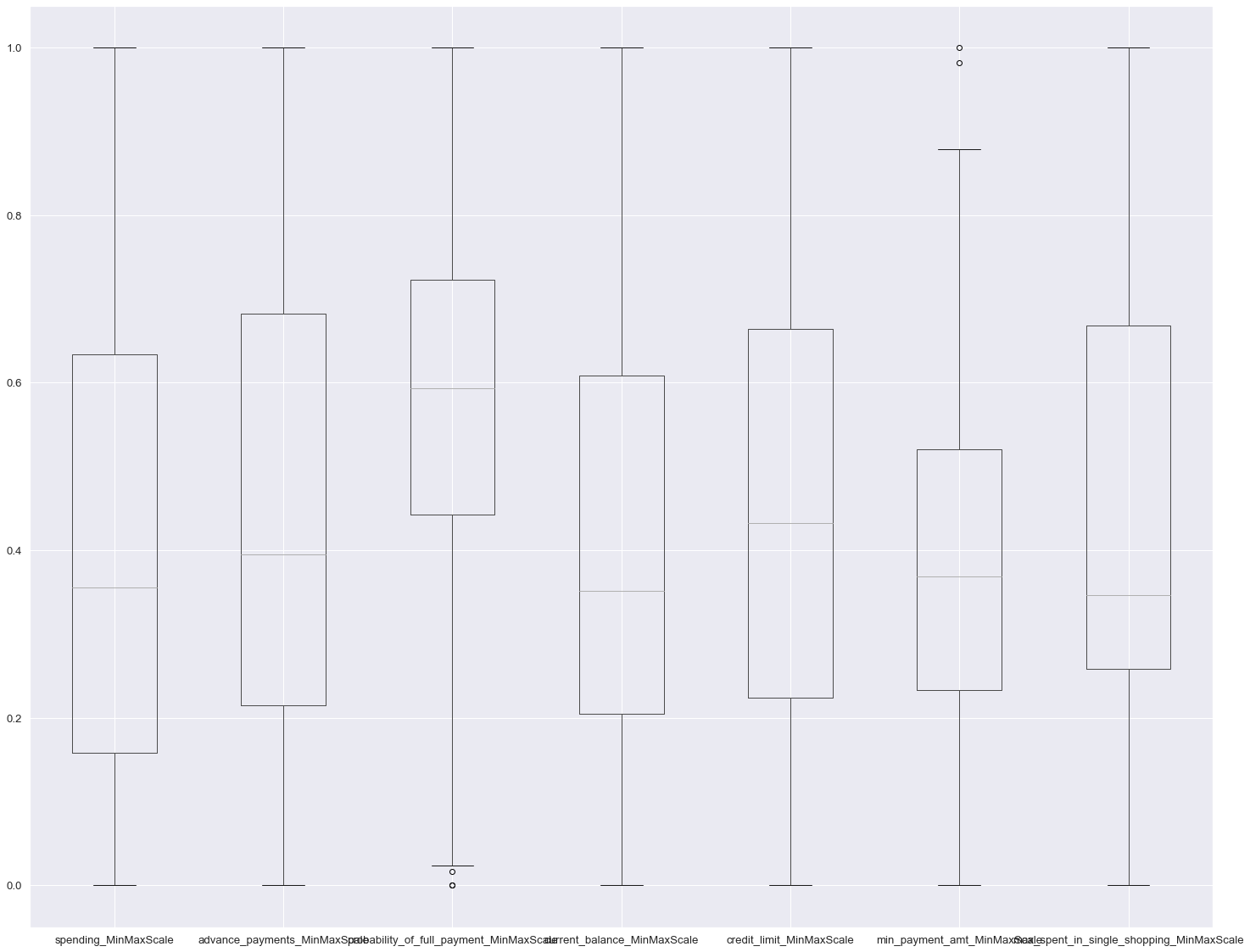
As we can see there are outliers therefore we will remove them.

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

Yes clustering is necessary in this case because the data values are in different value formats. Therefore to analyze them we need to bring them under common Z-score value. k-nearest neighbors with an Euclidean distance measure is sensitive to magnitudes and hence should be scaled for all features to weigh in equally.

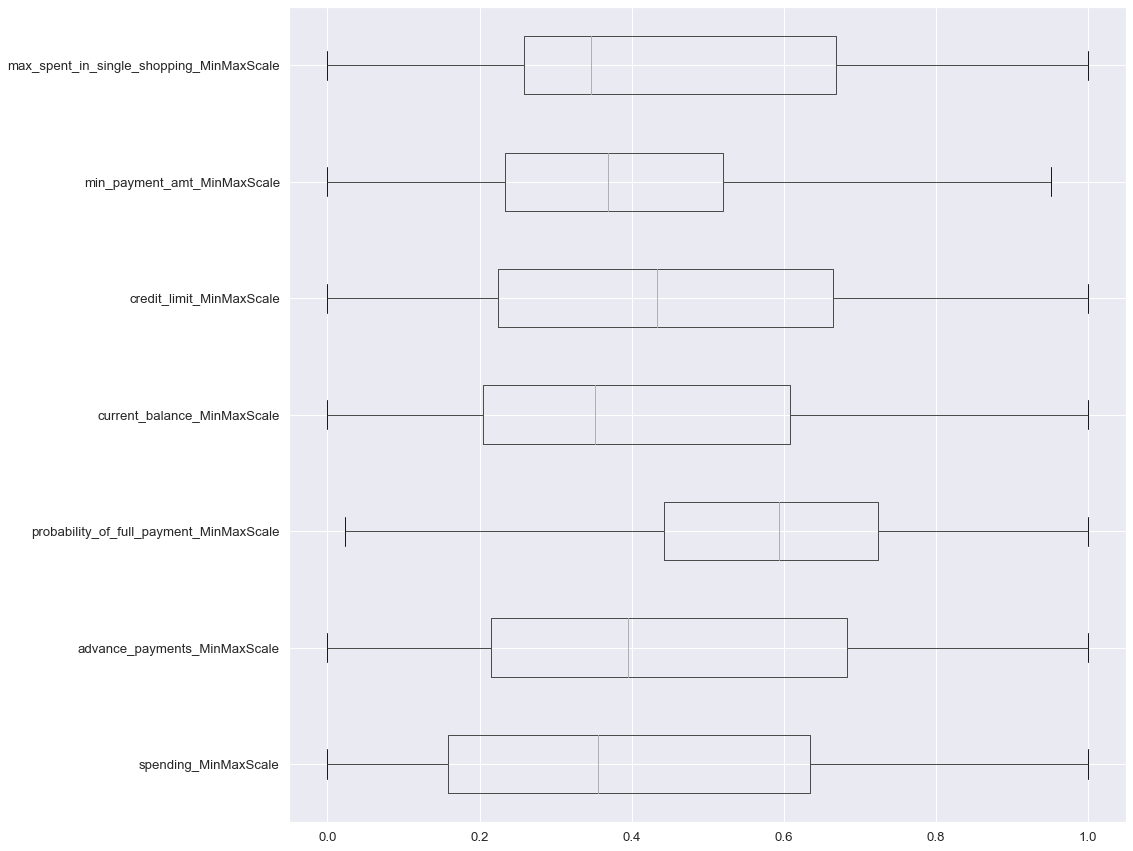
In below head(10) statement you can see how our data looks after scaling. In box plot you can see how it looks. It looks same as it was earlier. However, representation values have been changed.



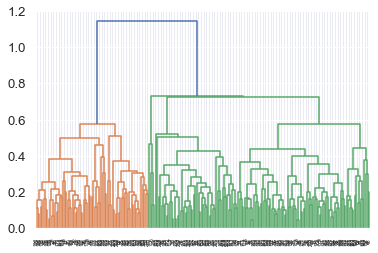


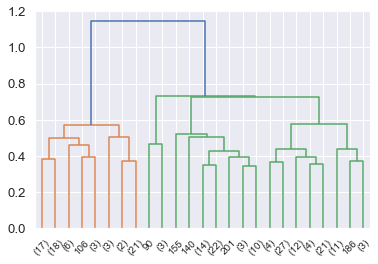
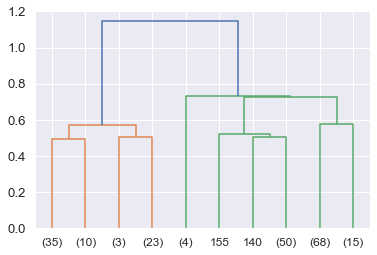
After this step I have removed the outliers.

After the removal of the outliers. This is how our boxplot looks.



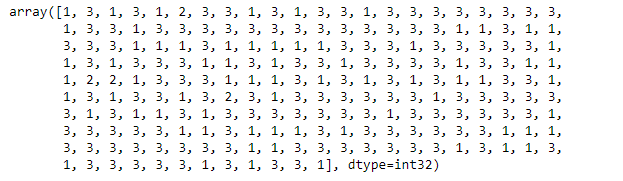
* 1. Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them

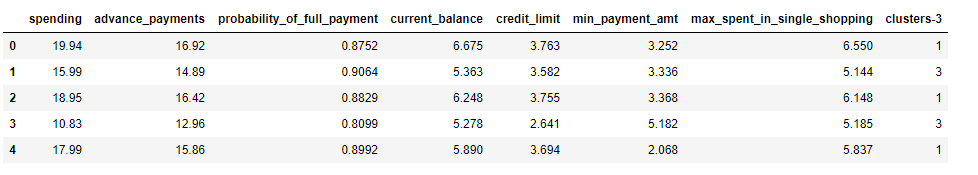




After applying the hierarchal clustering to scaled data we can clearly see that the scaled data roughly has data in 3 different clusters

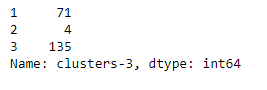
After this I am going to set criterion as maxclust, then I shall create 3 clusters, and store the result in another object 'cluster3'



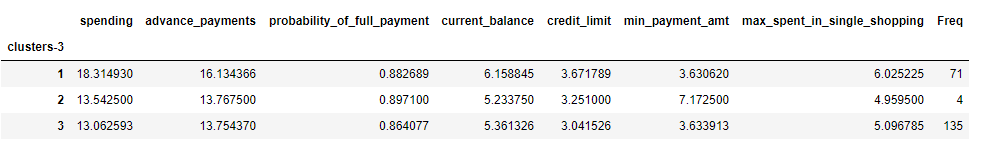


## Cluster Frequency

Here we can see what is frequency for each cluster.



## Cluster Profiles



* 1. Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.

Applied K-Means when n=1,

88.83560425994767

n=2

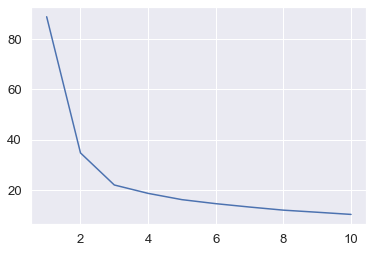
34.67235327560583

n=3

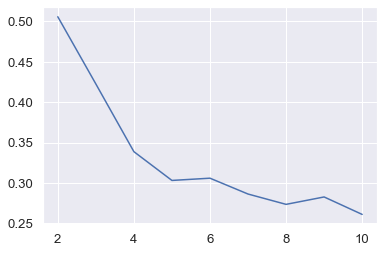
21.912014534778592

n=4

18.60109004012824



Above is the elbow curve x axis is number of clusters & y axis is numbers under each cluster.

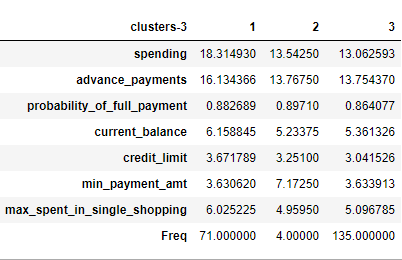


In the above plot I have plotted SC scores.

We can clearly see from scores & graph above, the number of optimal clusters could be 2 or 3

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

3 group cluster via hierarchical clustering



Based on above analysis we can see that we have 2 types of people

1. People who spend more. They do more advance payments. The probability of paying completely is also very high. Therefore, such customers should be offered big deals with high revenue expectation. Because they has capacity to do that.
2. Less/ medium spening customers. Strategy for then should be somewhere not like that the case- they should be given better offers on small deals.