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.

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

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

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

1.5 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)

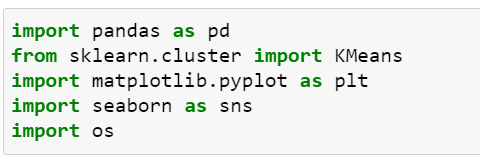
**credit\_limit**: Limit of the amount in credit card (10000s)

**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.**

Import all the necessary libraries.



Specially KMeans for K Means Clustering from sklearn.cluster

Seaborn and Matplotlib for visualization

os to set the home directory or file path.

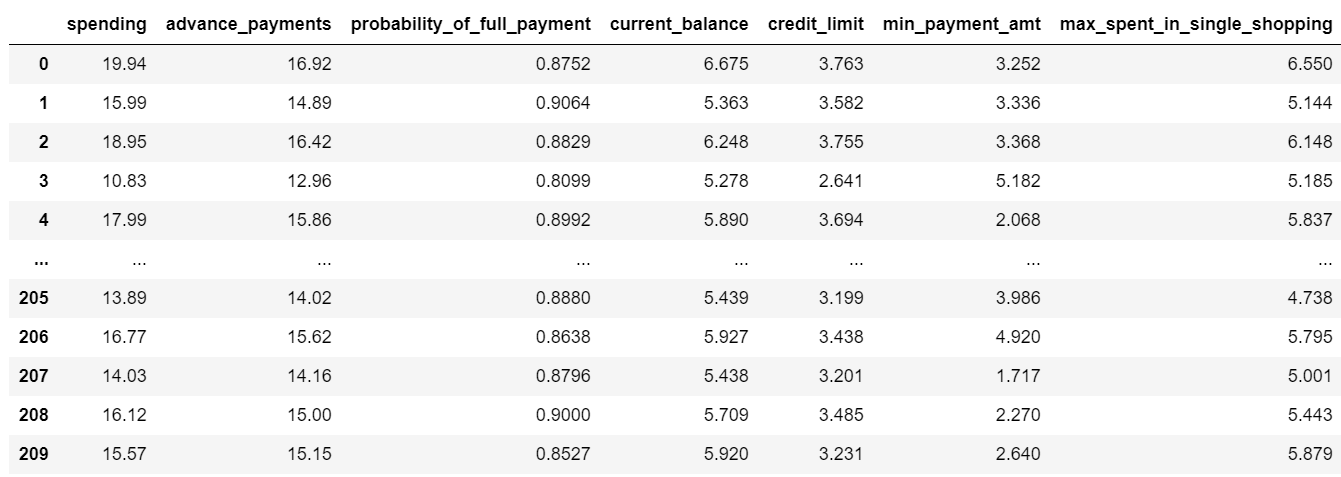
Chang the path :



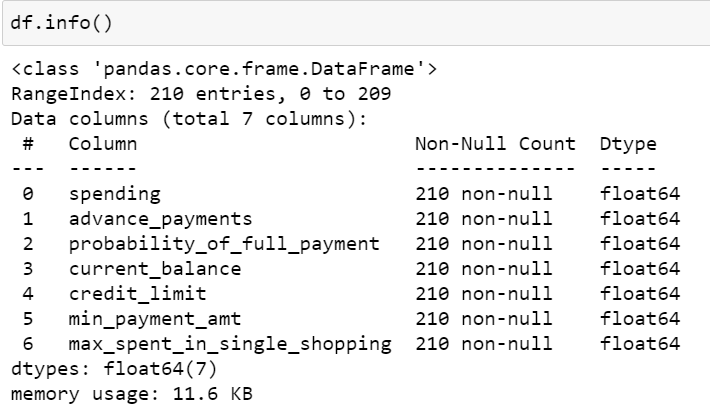
Read the dataset:



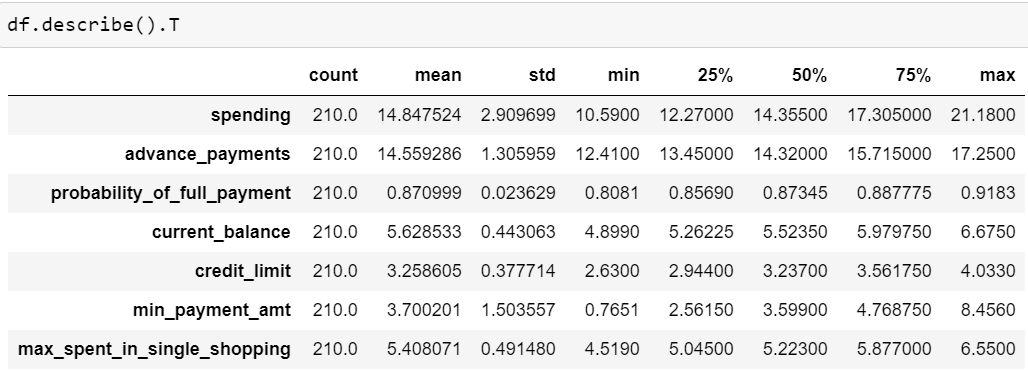
And save it in dataframe.



To get the general idea about dataset we will use command df.info()



It will give us a little idea in simpler form. Like Total 7 columns are there, No Null Values, All having data type Float, so ready to apply clustering. Seems to be 210 rows.



df.describe command will give us idea about values its distribution of Features.

Mean and median are almost same for all variables. And Standard deviation for Spending seems to be high for “Spending” variable. For all other variables it is normal.

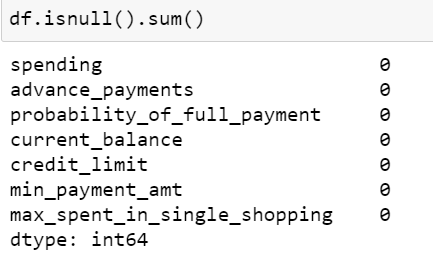
by looking at above dataframe, we can see here distribution of data is in Different forms like 100s,1000s. so we need scaling to bring all data on same level.

Lets Check for Duplicate Values: by below command



We can see that output 0 means no null values are present in the dataset.

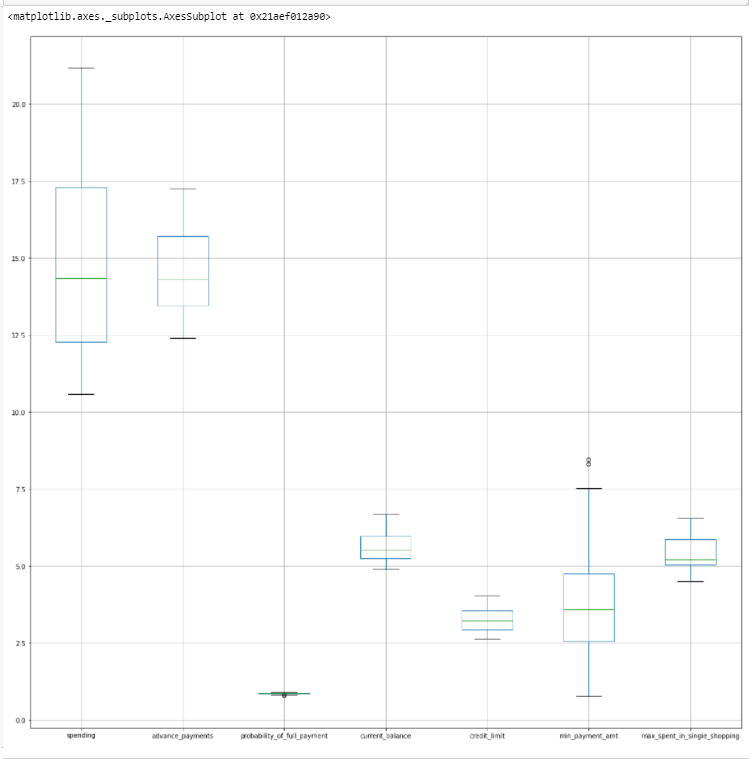
Null Value check by variables:



No Null Values in dataset.

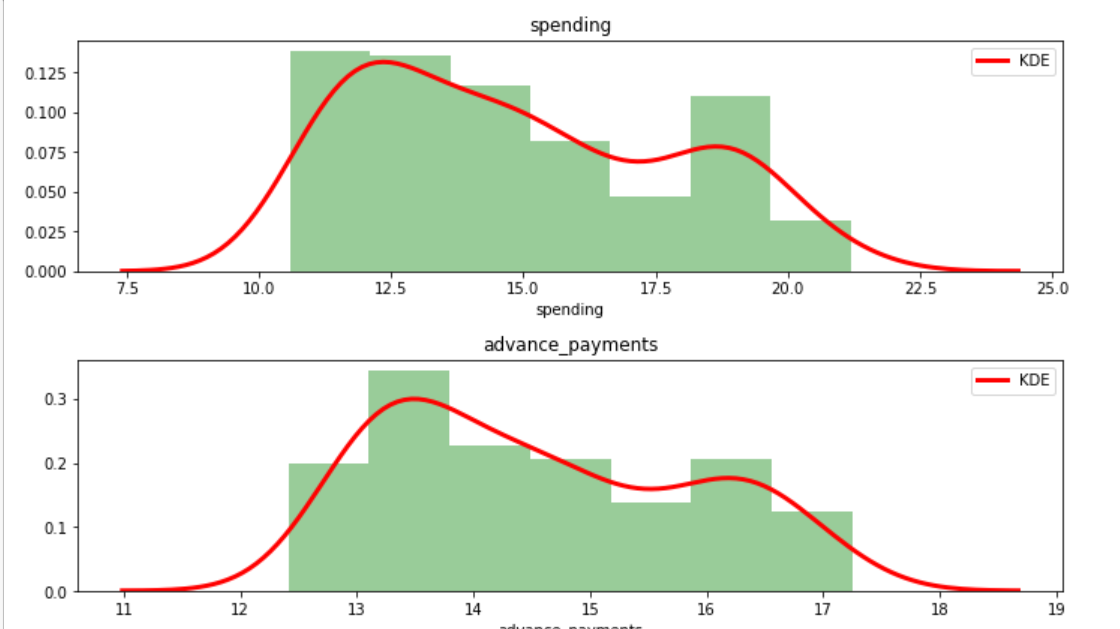
Lets plot the Boxplot for each variable, so we will get rough idea about Outliers by applying below command line.



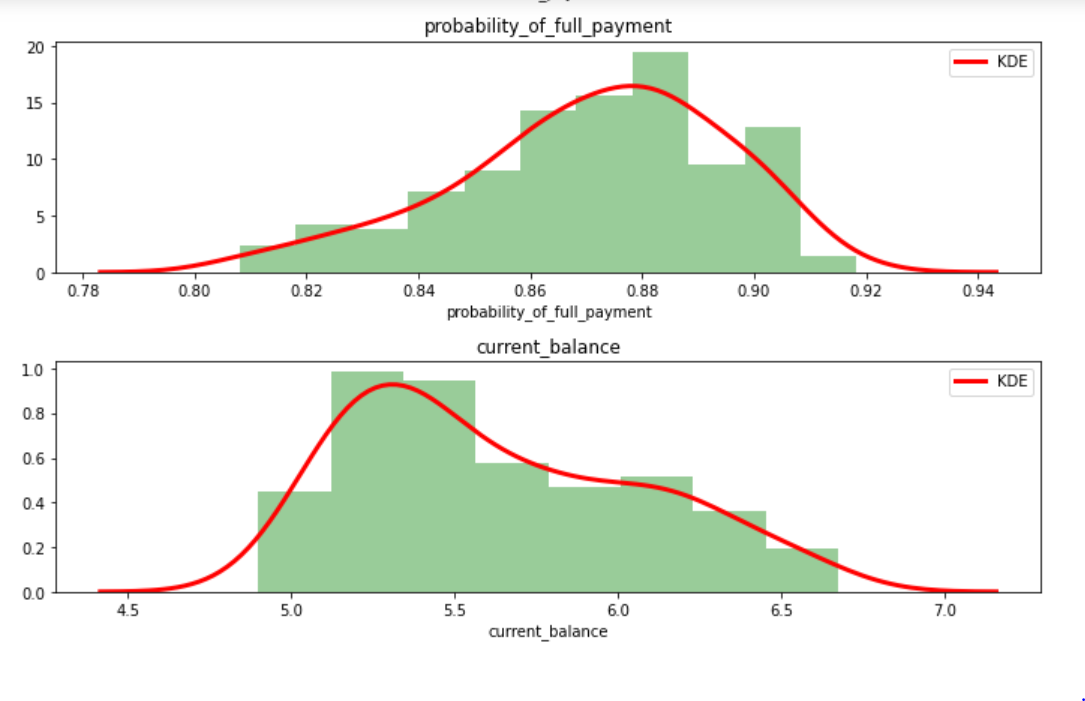


Lets do Univariate Analysis, to analyse the each variable specifically. By plotting distribution plot along with Kde.

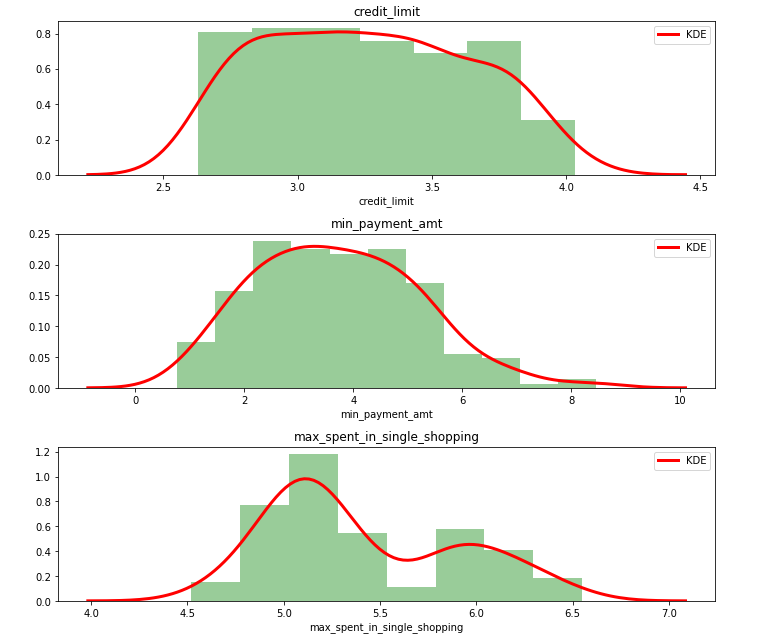




For Variable Spending and advance\_payments we can catch a thing that both are Right Skewed, Not exactly normally distributed around mean.



For Variable “Probability of Full payment” seems to be left skewd whereas current balance is right skewd.



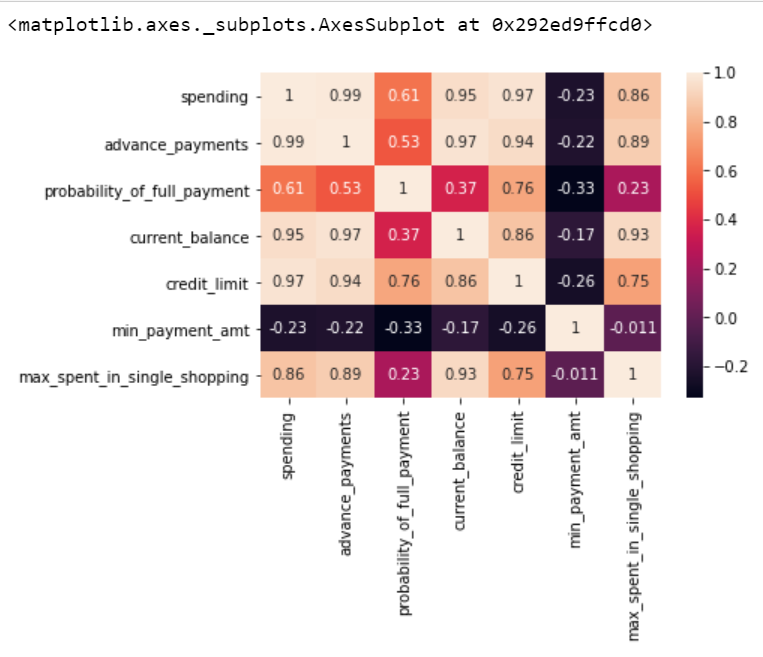
Credit Limit and mon payment amount are seems to be somehow Normally distributed, while

Max spent in single shopping shows right Skewed behavior.

In Overall scenario, Distrubtion is skewed to right tail for all the variable execpt probability\_of\_full\_payment variable, which has left tail.

Lets plot heatmap for correlation:

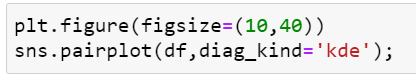


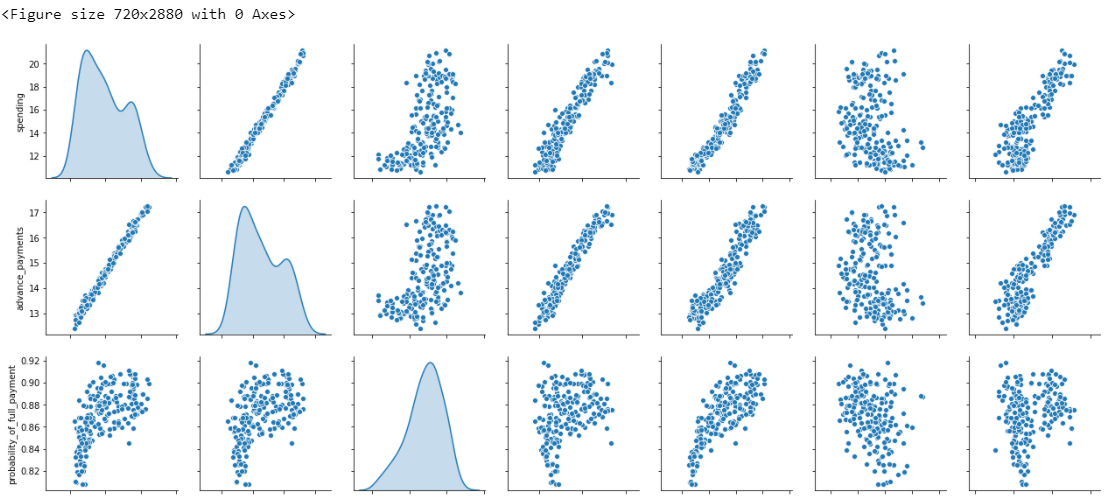


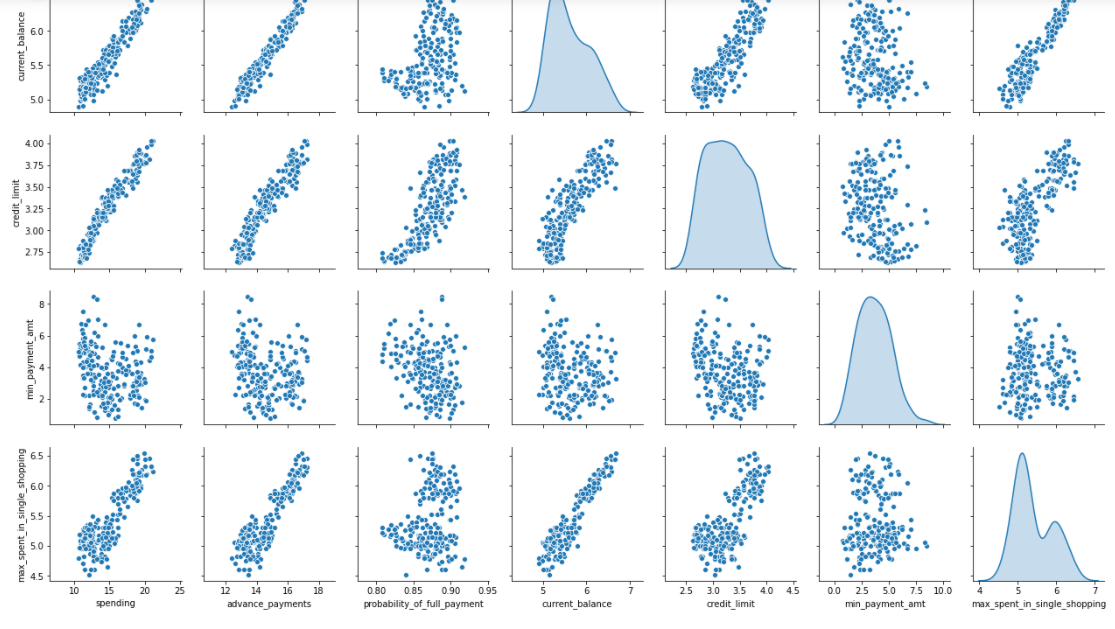
As we can see above data we can infer that

* It does not contain missing values, so no missing value treatment is needed.
* No Outlier treatment required as no invalid outlier can be detected.
* From Describe-Summary table, we can see that Standard Deviation for Spending Feature is high as compared to others and For almost all values Mean and Median are same. and things are normally distributed.

Lets plot a pairplot to perform Multivariate Analysis,







**Few overall observations:**

Strong positive correlation between Spending is highly Correlated with Advance Payment, Prob of Full payment, Current Balance, Credit Limit, Max Spent in Single Shipping.

Advance Payment and Spending, Prob of Full payment, Current Balance, Credit Limit, Max Spent in Single Shipping are highly correlated.

Probability of Full payment is highly correlated with Spending, Advance Payment, Current balance, Credit Limit.

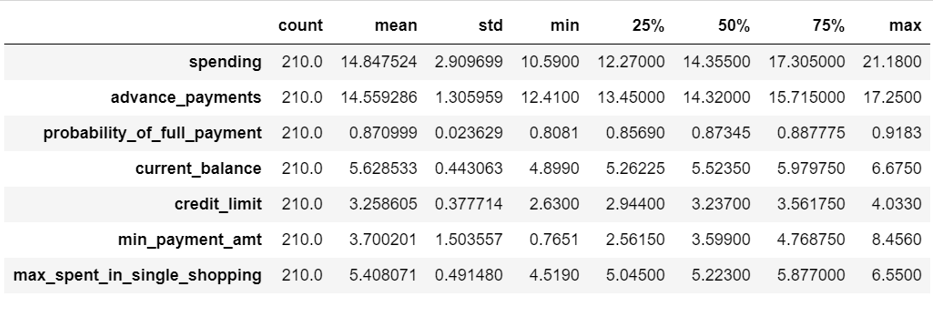
Maximum amt spent in Single shopping is highly correlated with Spending, Advance Payment, Current balance, Credit Limit.

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

If we see the variables closely, we can get a idea about values distribution the 0% 25 % 50% 100%

Are having different values. Some values are in 100’s and some in 1000’s.

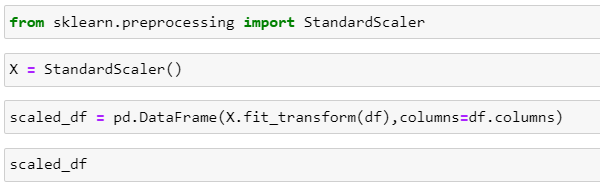
So we need to scale this data and measure it in same scaling. To perform all the operation on it.



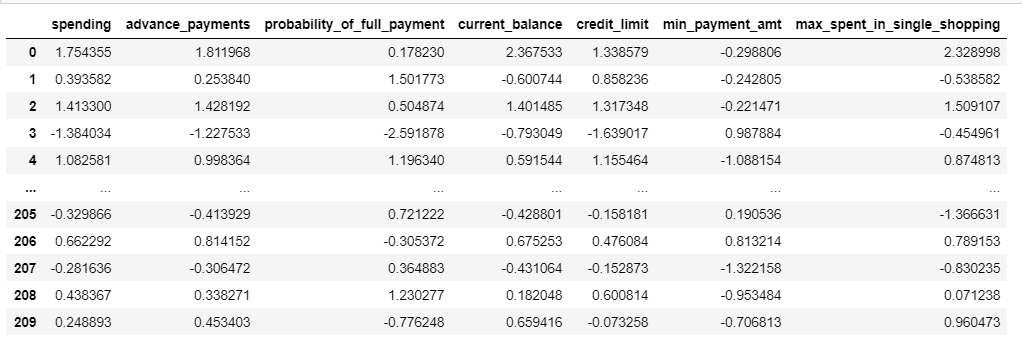
If we again check this table, we will get to know that Ranges are varied for all the variables.hence scaling is necessary.

Lets start scaling by using standardscaler() function.

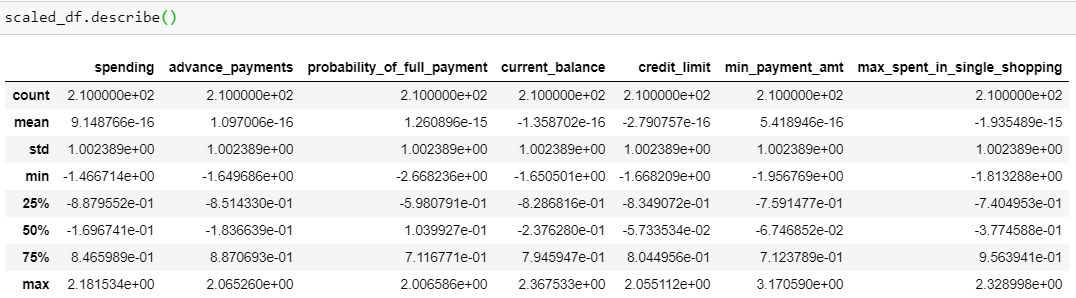
For that first Import StandardScaler from sklearn.preprocessing Library.



Scaled dataframe can be seen like below:

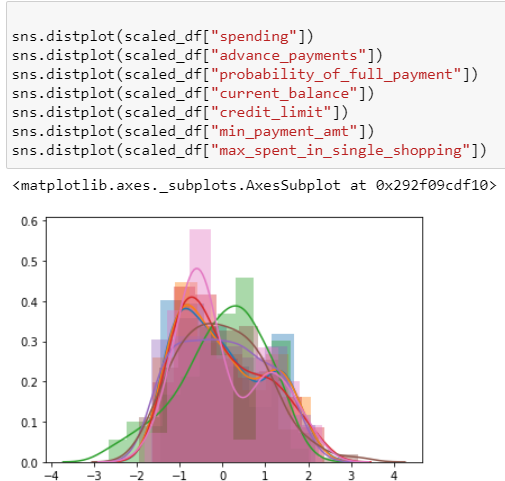


All values seems to be plotted on same table after scaling.



Mean values and max,min values all are brought on same page.

After Scaling the distribution can be seen as below:



By looking at Above graph, it seems that all values are in same platform, and shows scaled data with almost Normal Distribution.The given data is arranged in 100's 1000's so it is mandatory to bring all features on same table. So it is right to apply various clustering techniques.

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

### From scipy.cluster.hierarchy import dendorogram and Linkage, to perform Hierarchical Clustering.

### 

### Here we are using average method to draw a dendrogram,

### We can use ward method as well. It will give un pruned tree.

### 

### 

### Lets cut the tree at p=10.

### 

### 

### Lets cut the tree again at p=15.

### 

Lets take p=25,

### 

### From cluster.hierarchy import fcluster, to perform the clustering.

### 

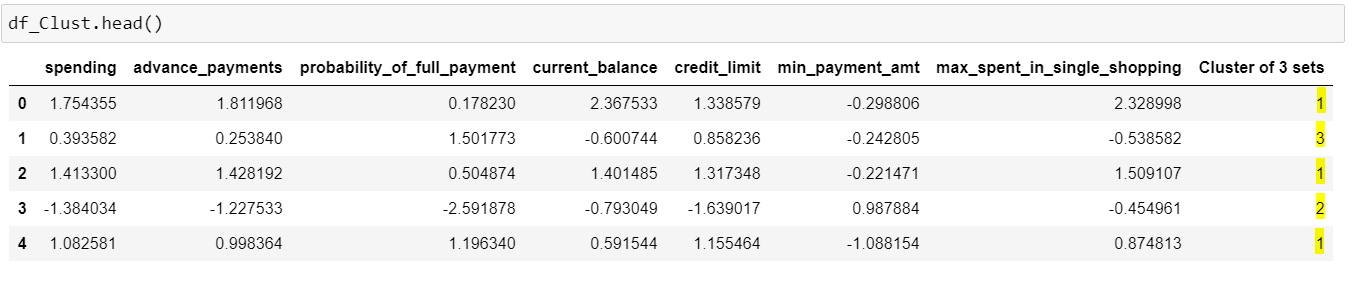
### We are selecting here criterion as maxclust, then we will create 3 clusters and all the results can be seen.

### 

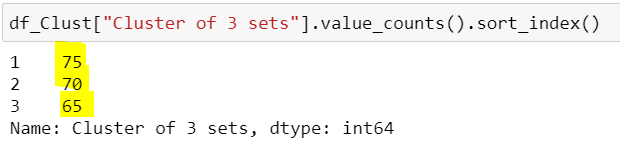
### Now lets copy the original data in df\_Clust, Add one new column “Cluster of 3 sets”.

### 

If we can see the first 5 rows, we can see the 3 Clusters. As 1,2,3.



If we do value\_counts() and then sorting we will get 3 clusters and value count within it as 75,70,65.

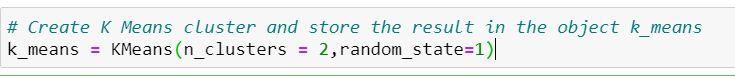


By above Classification and little domain expertise we can differentiate these data between 3 clusters High/Low/medium. That will be perfect for Customer Segmentation.

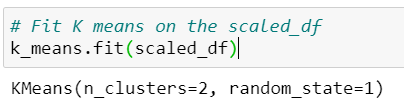
### 1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score.

Applying Kmeans clustering on Scaled data,

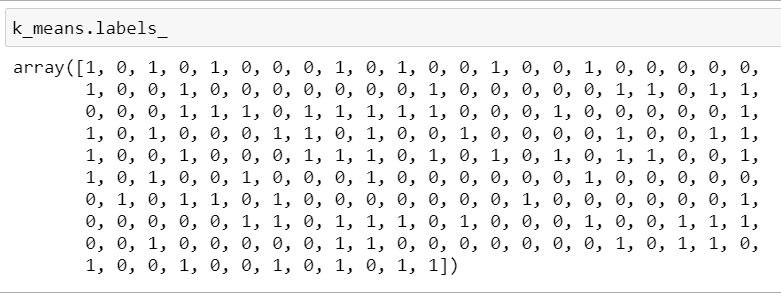
Create K Means cluster and store the result in the object k\_means:



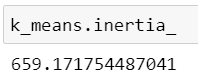
Fit Kmeans on scaled data



Labelling can be done if we take 2 clusters, below labels can be seen.



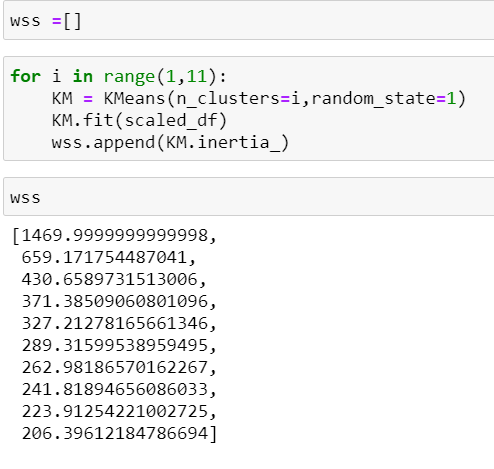
The value of K means Inertia is 659.1717



WSS Score:

It is the sum of distances between points and corresponding centroids for each cluster.

Lets create a WSS array, and calculate for number of Clusters along with WSS score.



i value varies from 1 to 10. And after append all the values to array.

By observing all the values, after first three values, values are becoming less effected.

1469 to 659: difference is 700+

659 to 430:difference is 200+

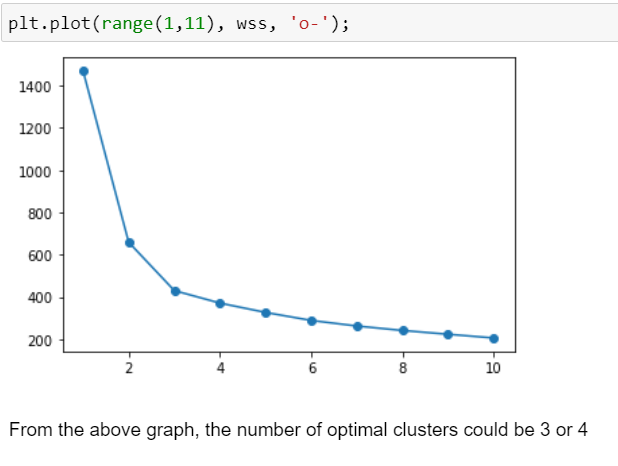
430 to 371:difference is around 60

371 to 327: difference is around 40

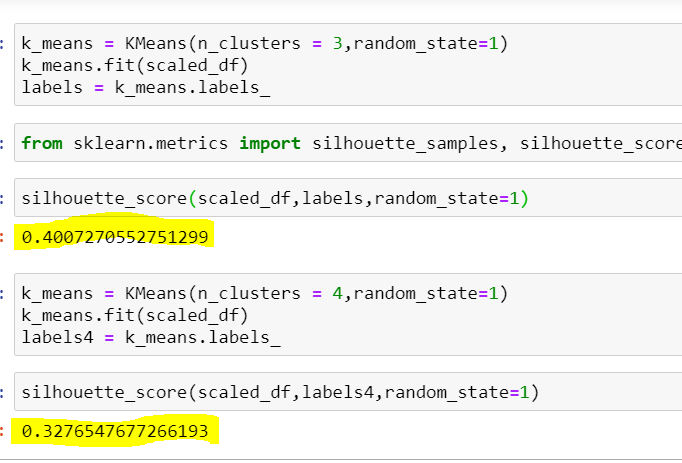
So we can select 2 Clusters, but again 2 clusters cant give perfect segmentation.

So optimal clusters we can select is 3.

Lets see visually, the elbow curve plot by using WSS.

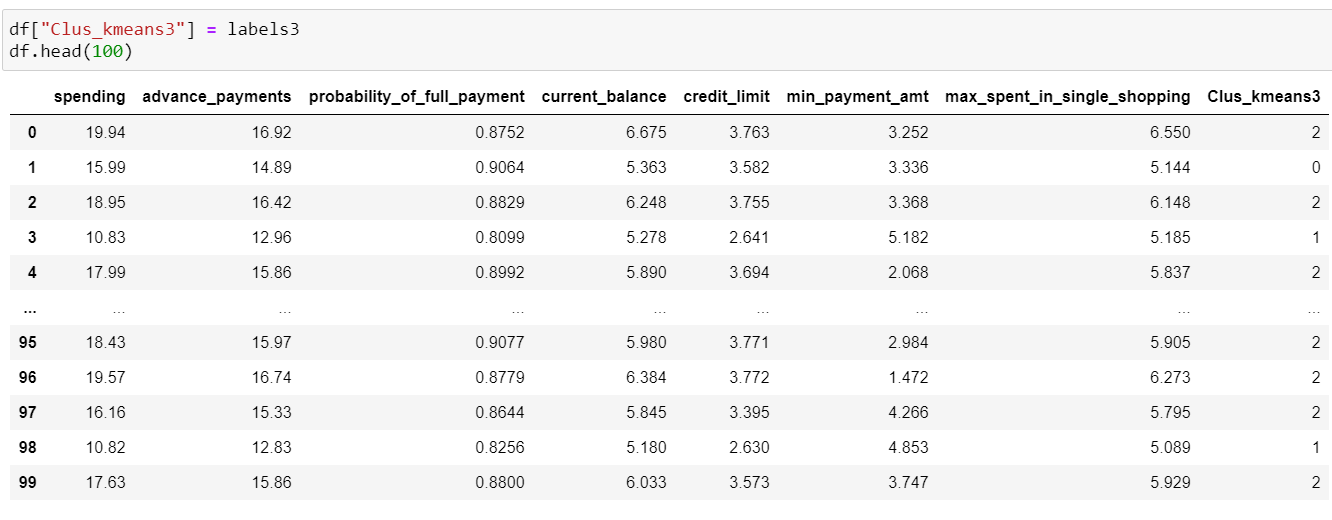


We can also Calculate Silhoutte score, to set the Optimal number of Clusters.

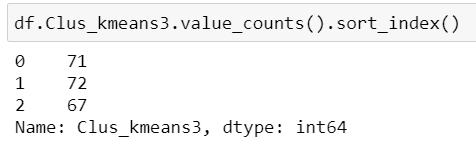


For 3 Clusters, Silhoutte score is 0.400 and 4 Clusters Silhoutte score is 0.3276

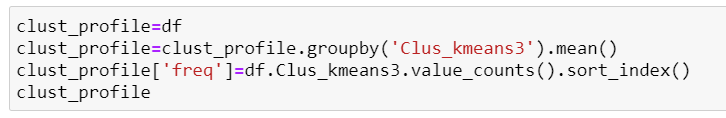
Divide the data then into 3 Clusters-> 0,1,2



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



Lets do Cluster Profiling. We can see above 3



We can see three Clusters with elements 71,72,67.

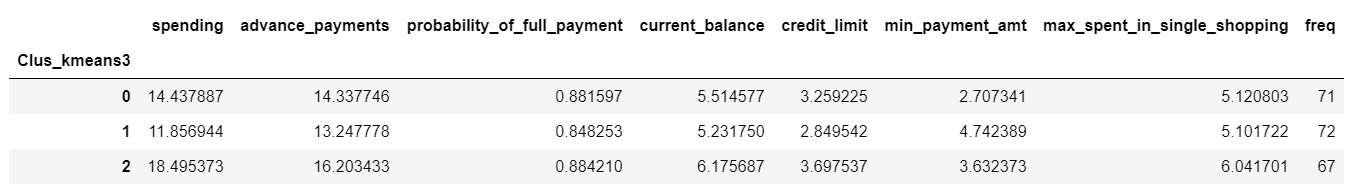
If we take Average values of clusters Column Spending, we can get the Three Clusters, First Cluster Spends average 14000 per month, Second Cluster Spends 11000 per month and the third one 18000 per month.

So we can list the 3 Clusters as

Cluster 0 : Medium

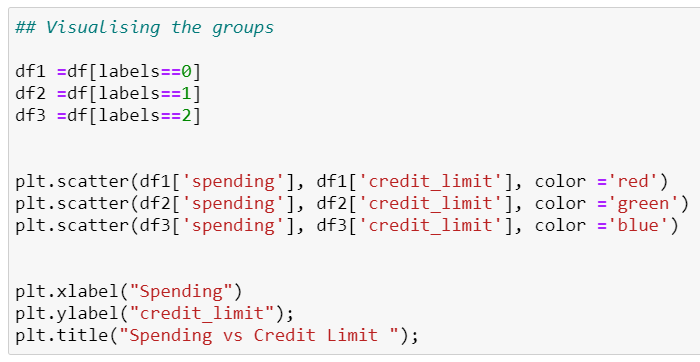
Cluster 1 : Low

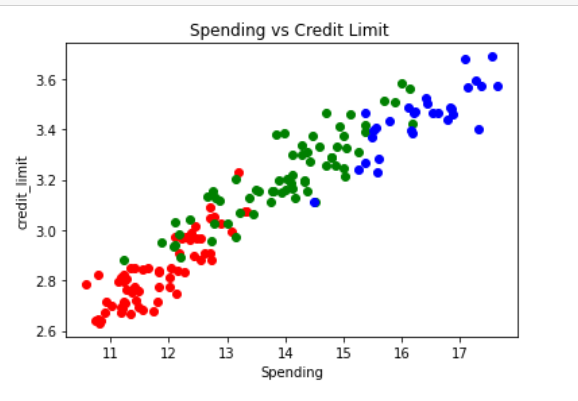
Cluster 2 : High

.

Now Visualize the All 3 groups:

1. Spending vs Credit Limit

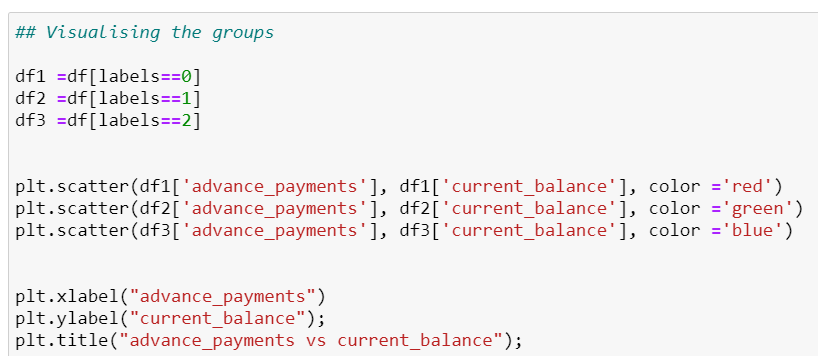


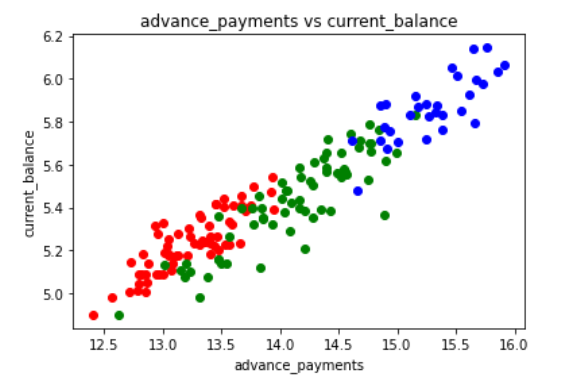


We can clearly see that Spending is increasing as Credit Limit is Increasing.

So we can Increase the Credit Limit of People to Increase the Spending per month.

Advance Payments vs Current Balance:

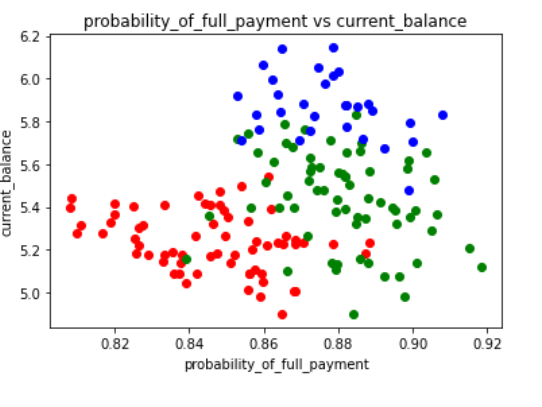




In the above plot, One Thing we can surely say that, Advance payment is done higher by people those are having High current balance in their account. Practically the same thing happens in day to day life.

Probability of Full Payment VS Current balance





Probability of payment seems to be High in “Medium” and “High Class” group.

So we can provide them with some promotional offers, Discount offers to Medium and High

Class, this will ensure Returns to Bank.

##### *Group 1 : High Spending Group*

* For this group the we can Improve spending habits of people by changing offers.
* maximum max\_spent\_in\_single\_shopping is high for this group, so can be offered discount/offer on next transactions upon full payment
* Give loan against the credit card, as they are customers with good repayment record.Tie up with luxary brands, which will drive more one\_time\_maximun spending
* To Increase the Spending we can Tie up with multiple brands, accordingly discounts can be given to Improve monthly spending
* For this group we can give Credit points for Purchase and Timely Repayment. So as per points they can get further discounts in future. they will use more credit cards, that will benefit the bank.

##### *Group 2 : Medium Spending Group*

* The target we can set here is We can convert this group to High Spending group by providing more offers than High Spending Group
* Bank can give Lucrative offers to this group to spend More, as this group repayment is good as compared to Other two groups.
* They are potential target customers who are paying bills and doing purchases and maintaining comparatively good credit score. So we can increase credit limit or can lower down interest rate.
* Promote premium cards/loyality cars to increase transcations.

##### *Group 3 : Low Spending Group*

* customers should be given remainders for payments. Offers can be provided on early payments to improve their payment rate.
* Increase there spending habits by tieing up with grocery stores, utlities (electircity, phone, gas, others)
* This group needs to be worked upon, to Increase their Spendings, their Lifestyle, usage of credit cards. More Marketing/Sales strategy needs to be apply.
* This group seems to be a lesser earning, so we can give them good offers to Initiate the payments.