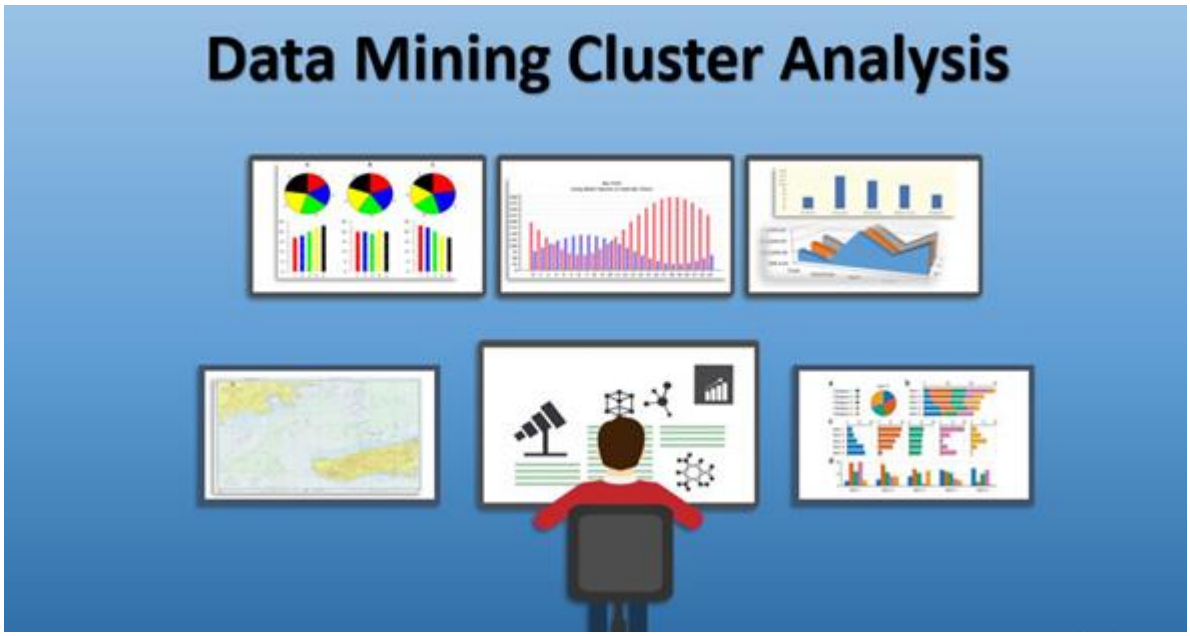


Image Source: - www.educaba.com

Problem 1: Clustering: - Solution

Submitted by :

Kanhaiya Awasthi

PGPDSBA SEP_2020

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.

Dataset for Problem 1: bank_marketing_part1_Data.csv

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)

Problem says that you have to segment the market based on credit card usage, using Clustering methods.

Let's Have a closer Look on the first ten rows of the Data

spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
19.94	16.92	0.88	6.68	3.76	3.25	6.55
15.99	14.89	0.91	5.36	3.58	3.34	5.14
18.95	16.42	0.88	6.25	3.76	3.37	6.15
10.83	12.96	0.81	5.28	2.64	5.18	5.19
17.99	15.86	0.90	5.89	3.69	2.07	5.84
12.70	13.41	0.89	5.18	3.09	8.46	5.00
12.02	13.33	0.85	5.35	2.81	4.27	5.31
13.74	14.05	0.87	5.48	3.11	2.93	4.83
18.17	16.26	0.86	6.27	3.51	2.85	6.27
11.23	12.88	0.85	5.14	2.80	4.33	5.00

All the columns of the Dataset is in numerical form .

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

Since the file is given in Csv format so we use pandas read CSV function to read the file

Checking Info & Null values of the dataset :-

```
<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
```

There are 210 rows and 7 Columns in the Dataset.

Checking Null Values in the dataset:-

```
spending                      0
advance_payments              0
```

```
probability_of_full_payment      0
current_balance                  0
credit_limit                     0
min_payment_amt                  0
max_spent_in_single_shopping     0
dtype: int64
```

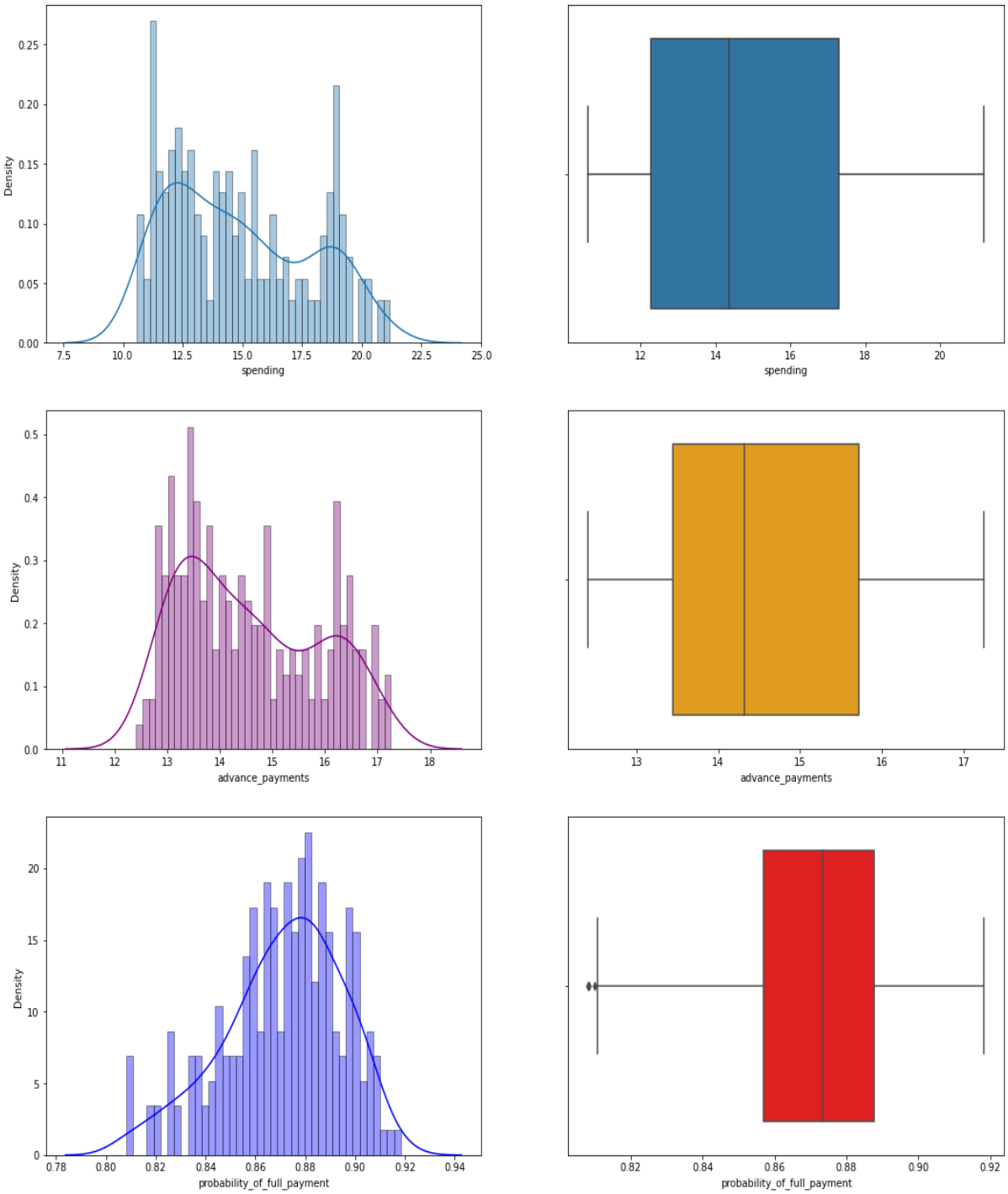
Looking at above results, there are no null values & Duplicate Values present in the Dataset .

Description of Dataset:-

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.00	210.00	210.00	210.00	210.00	210.00	210.00
mean	14.85	14.56	0.87	5.63	3.26	3.70	5.41
std	2.91	1.31	0.02	0.44	0.38	1.50	0.49
min	10.59	12.41	0.81	4.90	2.63	0.77	4.52
0.25	12.27	13.45	0.86	5.26	2.94	2.56	5.05
0.50	14.36	14.32	0.87	5.52	3.24	3.60	5.22
0.75	17.31	15.72	0.89	5.98	3.56	4.77	5.88
max	21.18	17.25	0.92	6.68	4.03	8.46	6.55

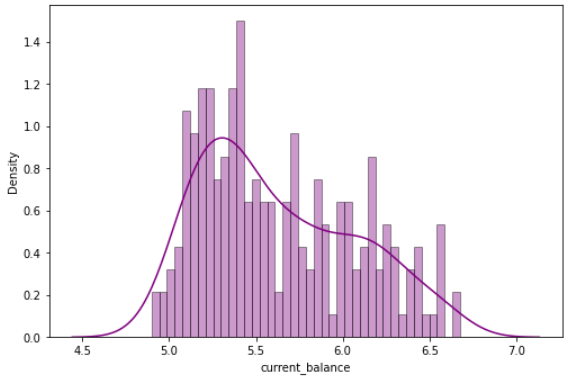
Looking at above table spending and advance payment means are almost similar , Median Values are also Closer the mean. At first glance Data looks clean.

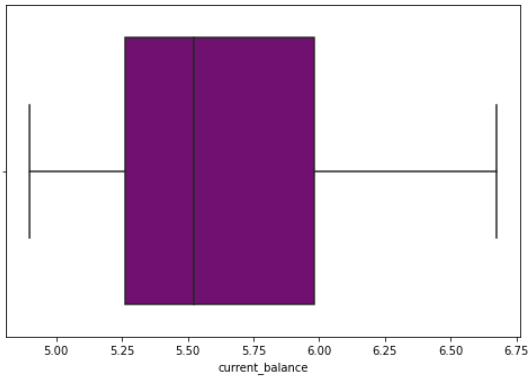
- **Plotting Distribution plot, Histogram and boxplot to check Ouliers & pattern of the Data ('spending', 'advance_payments', 'probability_of_full_payment')**

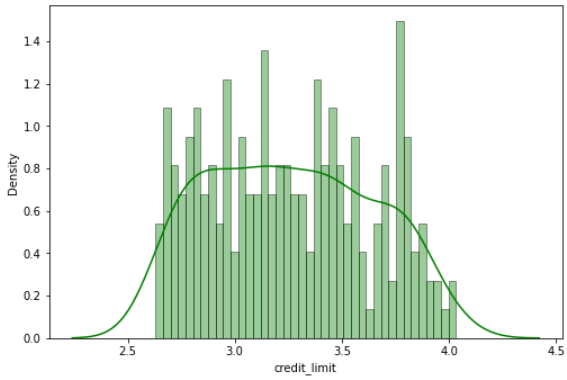


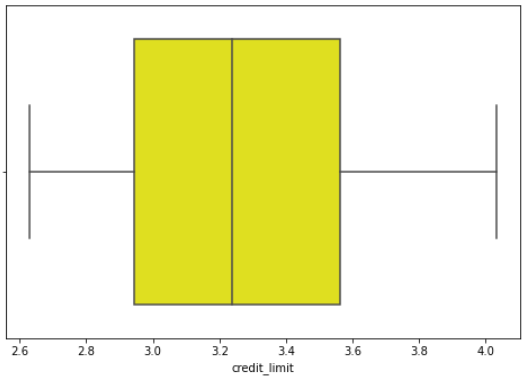
We had noticed from above figure that spending and payment are right skewed & their median values are almost similar.

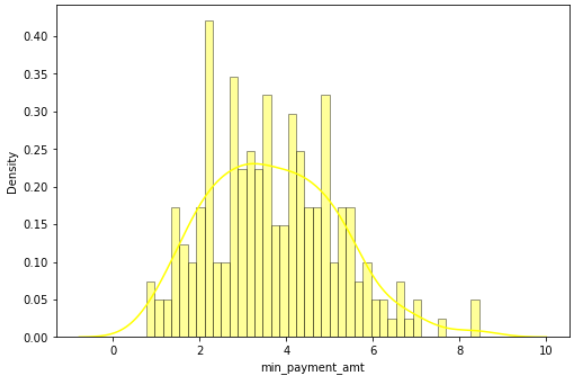
- **Plotting Distribution plot, Histogram and boxplot to check Oul iers & pattern of the Data for ('current_balance', 'credit_limit', ' min_payment_amt', 'max_spent_in_single_shopping')**

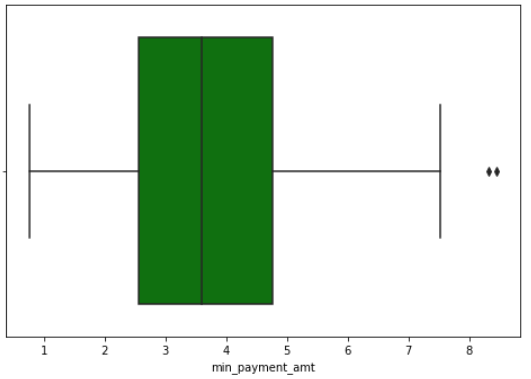


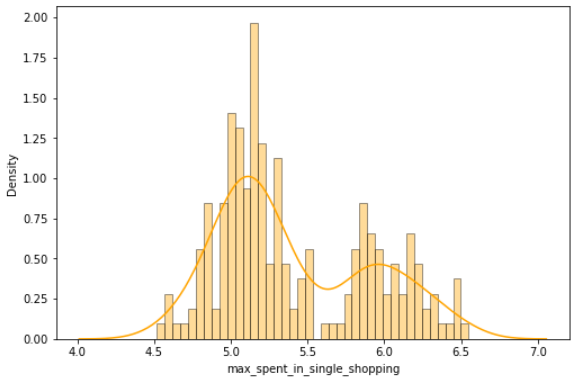


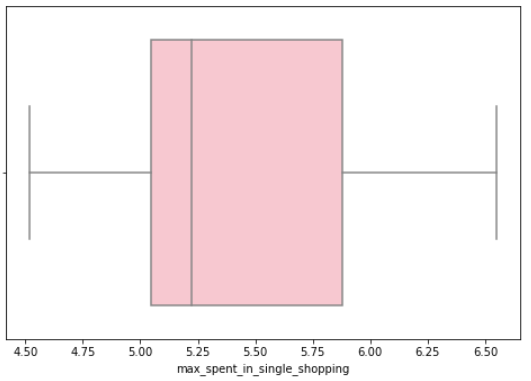








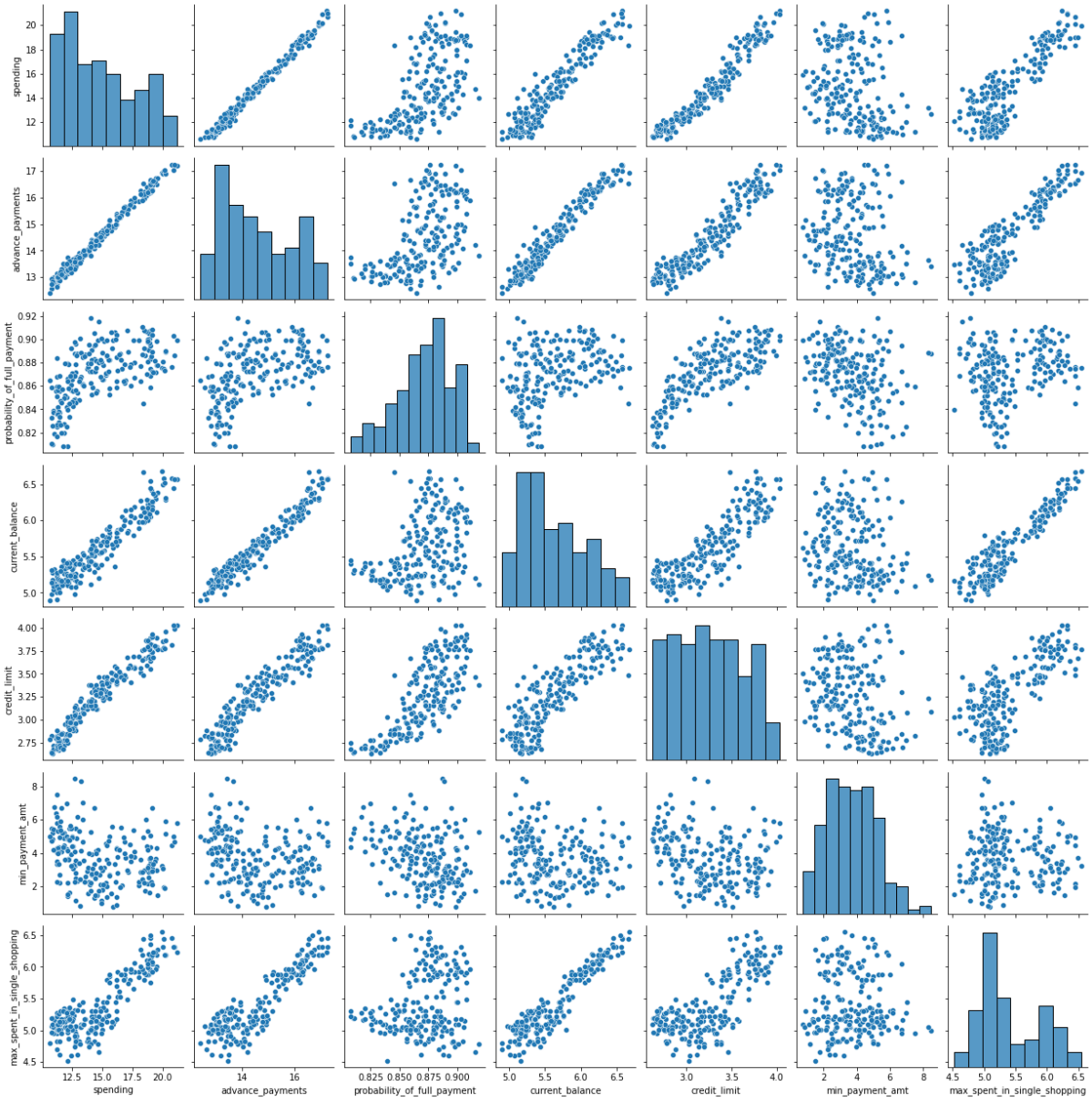




At first glance of above figure we have noticed that ('current_balance', 'credit_limit', ' min_payment_amt', 'max_spent_in_single_shopping') are right skewed and there is very less difference in their median and Average values.

There are very few Outliers in the data so we will proceed without treating the Outli ers

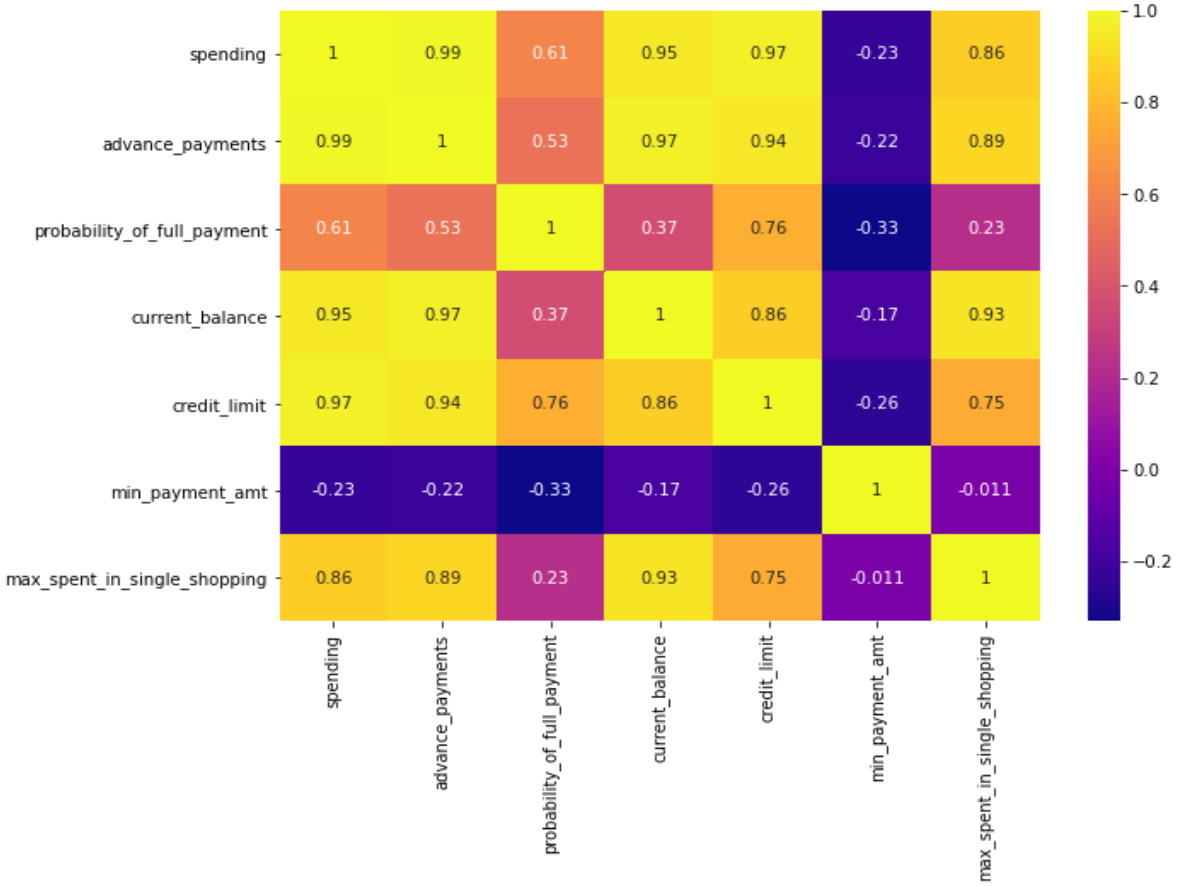
- **Multivariate Analysis:-** We have performed Multivariate analysis as below
  - **Pairplot:-**



▪ Correlation Matrix

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
spending	1.00	0.99	0.61	0.95	0.97	-0.23	0.86
advance_payments	0.99	1.00	0.53	0.97	0.94	-0.22	0.89
probability_of_full_payment	0.61	0.53	1.00	0.37	0.76	-0.33	0.23
current_balance	0.95	0.97	0.37	1.00	0.86	-0.17	0.93
credit_limit	0.97	0.94	0.76	0.86	1.00	-0.26	0.75
min_payment_amt	-0.23	-0.22	-0.33	-0.17	-0.26	1.00	-0.01
max_spent_in_single_shopping	0.86	0.89	0.23	0.93	0.75	-0.01	1.00

▪ Visual Representation using Heat Map: -



Having a glance at above figures we found Strong positive correlation between which are marked as dark green in Correlation matrix.

- credit_limit & advance_payments
- advance_payments & current_balance,
- credit_limit & spending
- max_spent_in_single_shopping current_balance
- spending & current_balance
- spending & advance_payments

since above variables are highly correlated so there will be the problem of Multi collinearity.

What is Multi Collinearity ?

Multicollinearity mostly takes shape when independent variables in a regression model are correlated. This correlation is a problem because independent variables should be *independent*. If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

What will be the possible solutions to deal with Multi collinearity ?

Below Solutions can be effective

- Remove some of the highly correlated independent variables.
- Linearly combine the independent variables, such as adding them together.
- Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.

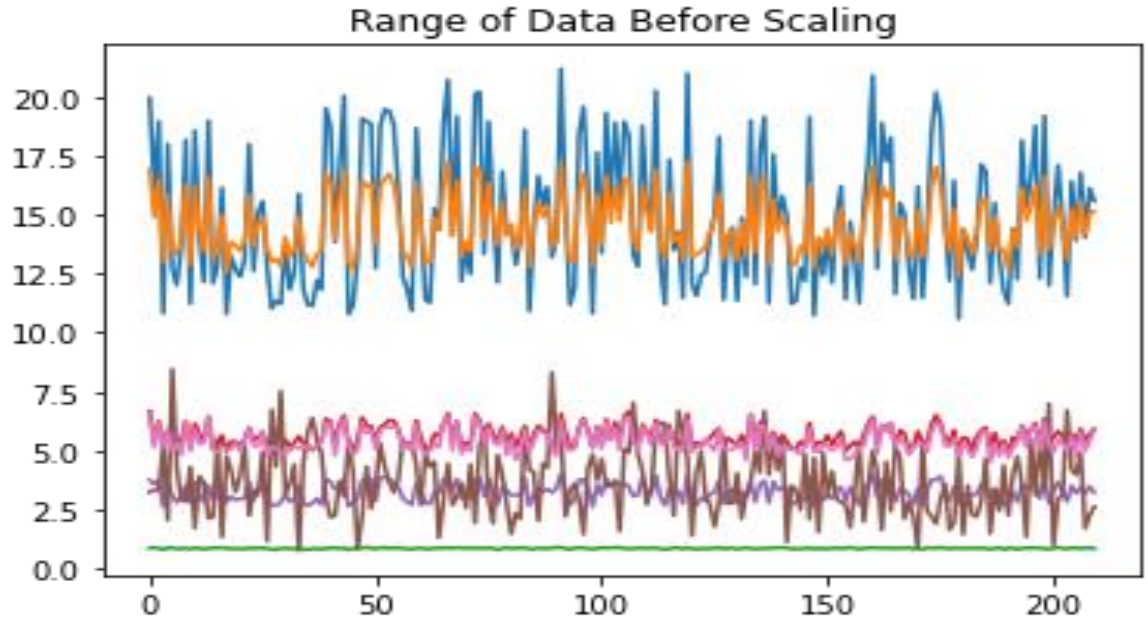
As we know that Multicollinearity generally doesn't impact the clustering process so we are not going to treat multicolinearity in this case.

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

Let's again have a glance at the head values of the dataset

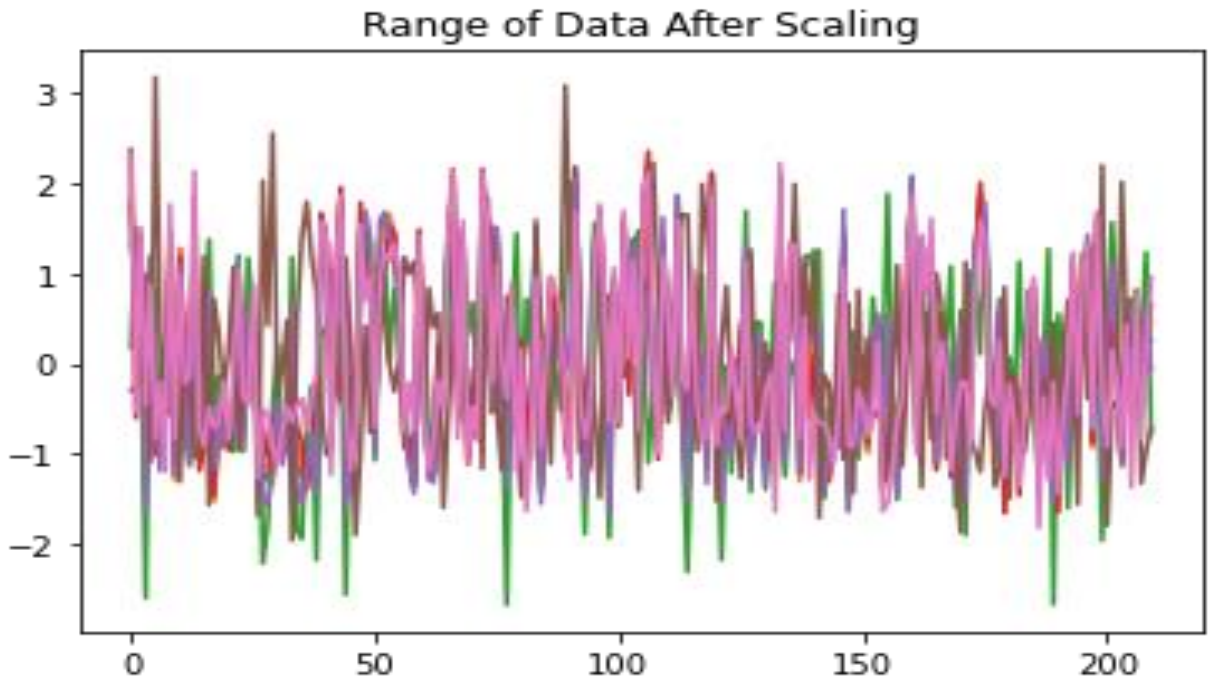
spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
19.94	16.92	0.88	6.68	3.76	3.25	6.55
15.99	14.89	0.91	5.36	3.58	3.34	5.14
18.95	16.42	0.88	6.25	3.76	3.37	6.15
10.83	12.96	0.81	5.28	2.64	5.18	5.19
17.99	15.86	0.90	5.89	3.69	2.07	5.84

Looking at the range of Data Visually



As we can see that from Head of the Data and graph Ranges are different and advance payments (in 100s) & current balance (in 1000s),credit limit (10000s) have different range if we don't do the scaling credit limit will take more weightage while modal fitting . So scaling is necessary in this Scenario.

We will perform Z Score scaling using Standard Scaler Function of Sklearn Module.



- **Scaled Data: -** Below are the first five rows of scaled Data

spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
1.75	1.81	0.18	2.37	1.34	-0.30	2.33
0.39	0.25	1.50	-0.60	0.86	-0.24	-0.54
1.41	1.43	0.50	1.40	1.32	-0.22	1.51
-1.38	-1.23	-2.59	-0.79	-1.64	0.99	-0.45
1.08	1.00	1.20	0.59	1.16	-1.09	0.87

- **Description of scaled Data**

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
count	210.00	210.00	210.00	210.00	210.00	210.00	210.00
mean	0.00	0.00	0.00	0.00	0.00	0.00	0.00
std	1.00	1.00	1.00	1.00	1.00	1.00	1.00
min	-1.47	-1.65	-2.67	-1.65	-1.67	-1.96	-1.81
0.25	-0.89	-0.85	-0.60	-0.83	-0.83	-0.76	-0.74
0.50	-0.17	-0.18	0.10	-0.24	-0.06	-0.07	-0.38
0.75	0.85	0.89	0.71	0.79	0.80	0.71	0.96
max	2.18	2.07	2.01	2.37	2.06	3.17	2.33

Standard Scaler function converted all the data in the range of -2 to +3 , So any other variable can't take more weightage while Modal fitting.

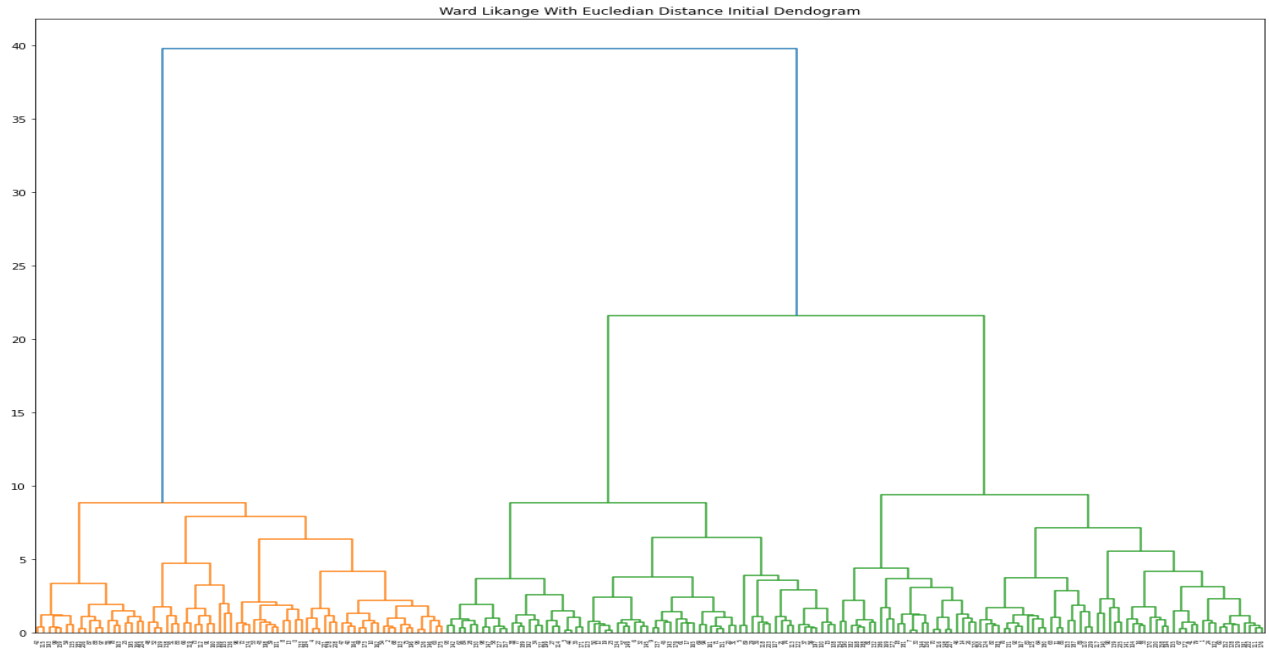
1.3) Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them.

Let's Understand What is Higherical Clustering?

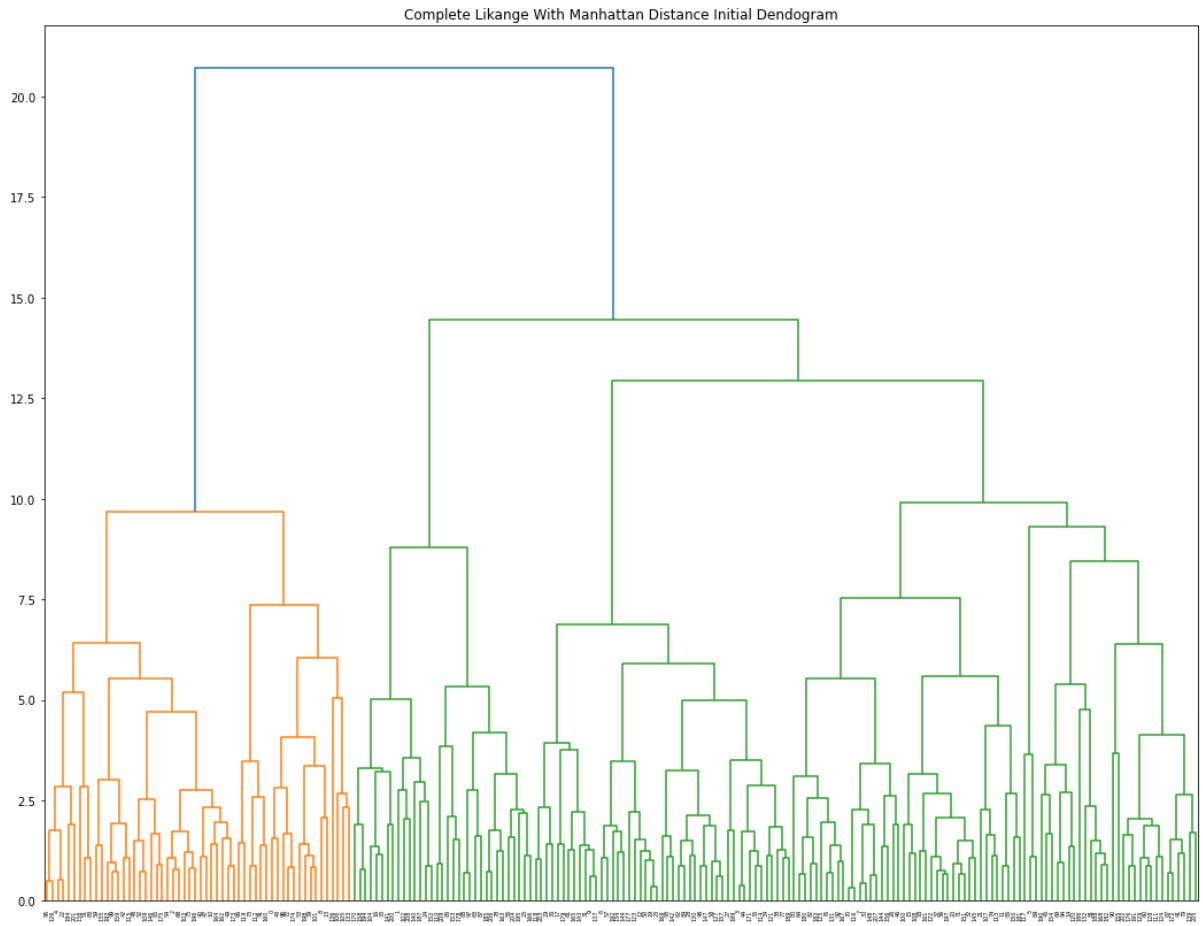
Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom.

We Will construct a Dendrogram for the scaled data and we obtain different cluster patterns using different linkages and distance criterions. They are as follows.

Ward Linkage & Eucledian Distance Dendogram :-



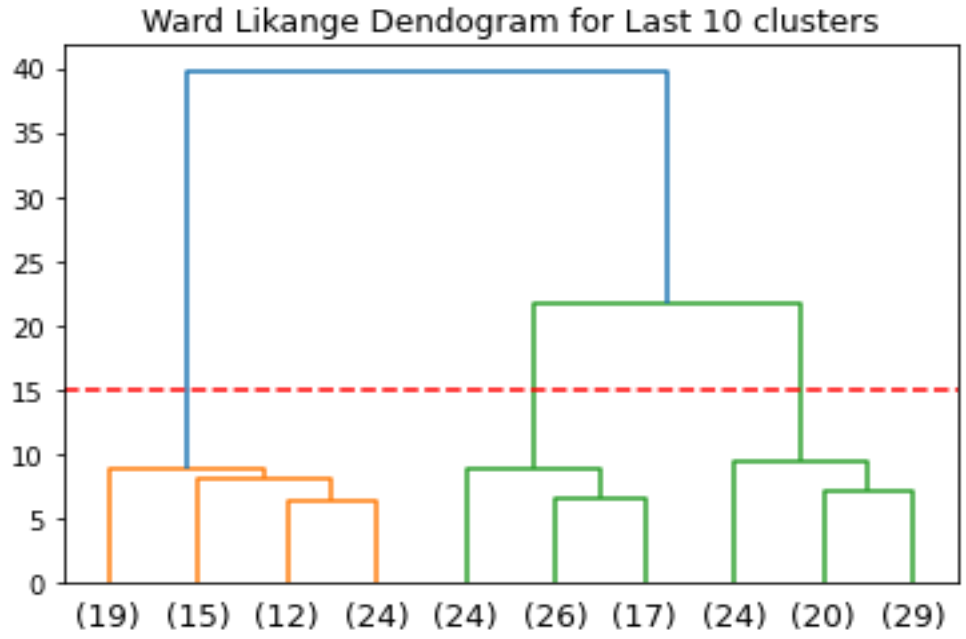
Complete Linkage With Manhattan Distance Dendogram:-



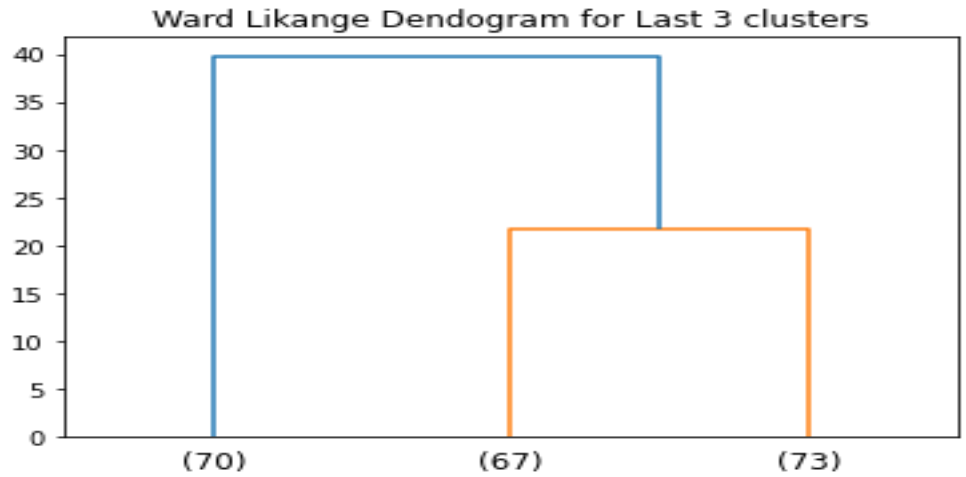
As we can observe from above two dendograms that dendogram with **Ward Linkage & Eucledian Distance** is making more sense so we try out same parameters for further making of dendograms.

As it can have more sense to create 3 clusters for dividing it into High , Low and Medium spending Groups so we will cut dendogram at a specified distance Criterion or maxclust Criterian .

We will be making dendogram for last 10 cluster and we want to Crop it at X=15 to get 3 Clusters



Making dendogram for 3 Clusters using Maxclust Criterion :-

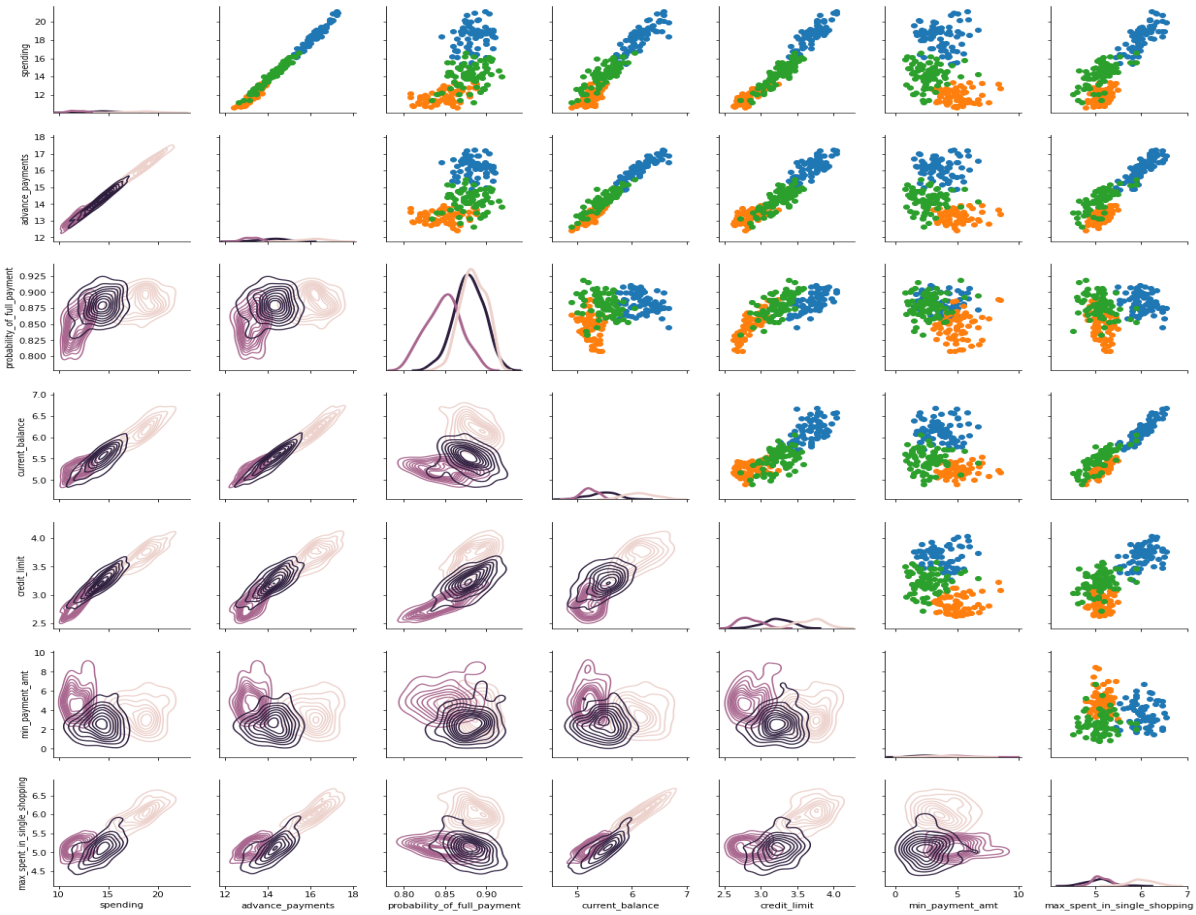


We can clearly see that first Cluster is having 70 numbers of observation and 2nd Cluster having 67 Number of observations & 3rd Cluster consist of 73 observations .

By performing Higherical Clustering on scaled dataset we found below results

clusters	1	2	3
spending	18.37	11.87	14.20
advance_payments	16.15	13.26	14.23
probability_of_full_payment	0.88	0.85	0.88
current_balance	6.16	5.24	5.48
credit_limit	3.68	2.85	3.23
min_payment_amt	3.64	4.95	2.61
max_spent_in_single_shopping	6.02	5.12	5.09
Freq	70.00	67.00	73.00

Looking at Cluster Profiles Visually :- We have plotted clusters as hue in the Original dataset.



We Can also Use Agglomerative Clustering let's first understand

What is Agglomerative Clustering?

Agglomerative clustering uses a bottom-up approach, wherein each data point starts in its own cluster. These clusters are then joined greedily, by taking the two most similar clusters together and merging them. For each cluster, you further divide it down to two clusters until you hit the desired number of clusters.

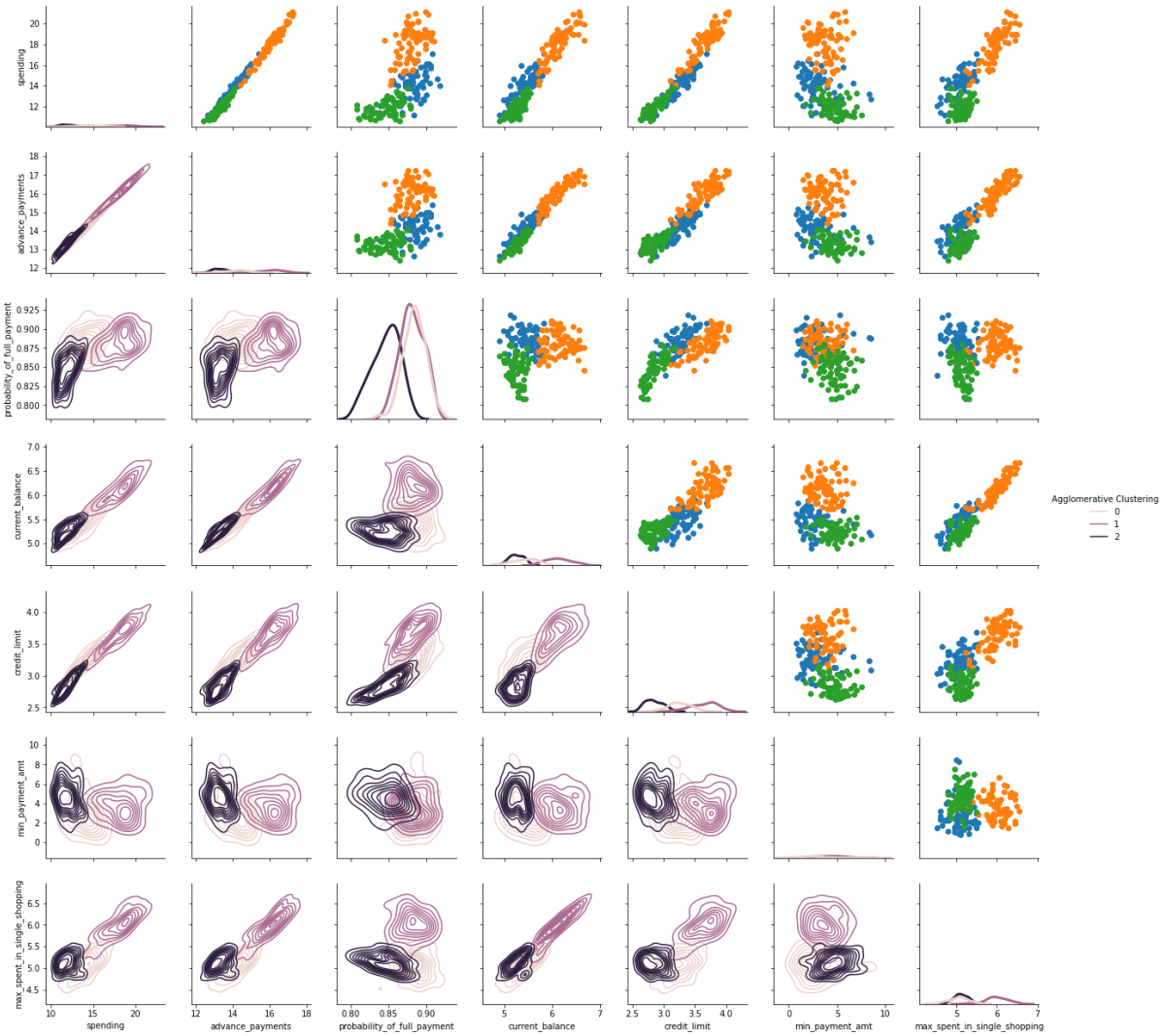
By performing this bottom up approach of Agglomerative clustering we found out below clusters

Agglo_CLusters	0	1	2
spending	14.2	18.1	11.9
advance_payments	14.2	16.1	13.3
probability_of_full_payment	0.9	0.9	0.8
current_balance	5.4	6.1	5.3
credit_limit	3.3	3.6	2.8
min_payment_amt	2.8	3.7	4.6
max_spent_in_single_shopping	5.1	6.0	5.1
Freq	65.0	75.0	70.0

No. of elements in each Cluster are as follows using Agglomerative Clustering

Agglo_CLusters	0	1	2
Freq	65	75	70

Plotting Agglomerative Clusters of whole dataset using pairplot.



Observation: - We can clearly observe that cluster 1 is having high probability of repayment and higher spending habit and their transaction amount in single shopping is also high whereas cluster 2 have lowest spending and probability of full payment is also low so we can give some special discount if customer will pay in one shot.

1.4 Apply K-Means clustering on scaled data and determine optimum clusters .Apply elbow curve and silhouette score Interpret the inferences from the model .

First of all, we have to know that what is the K Means Clustering.

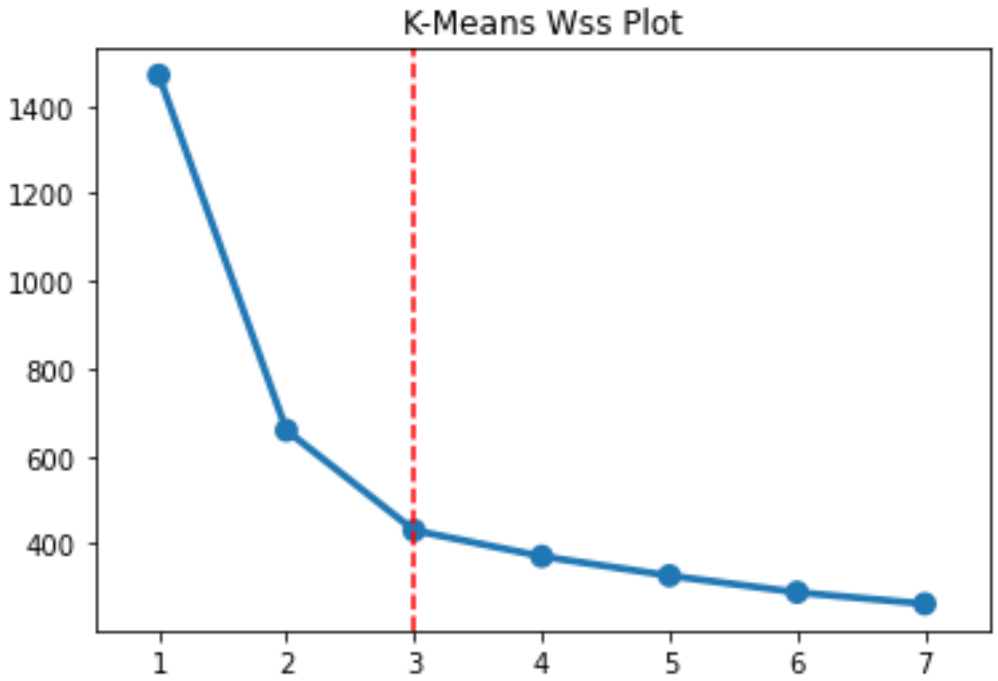
What is K –Means Clustering?
Kmeans algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to **only one group**. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters; the more homogeneous (similar) the data points are within the same cluster. Source:- www.towardsdatascience.com

- **Applying K-Means Clustering to Scaled Data:-**
  We have applied K means clustering by sklearn package, by fitting scaled Data we found various cluster labels(Refer attached Python file) .and Our next aim is to Choose optimum number of clusters.

- Choosing Optimum Number of Clusters: - Optimum number of clusters can be Chosen by looking at Elbow in WSS plot or Silhouette Score plot.

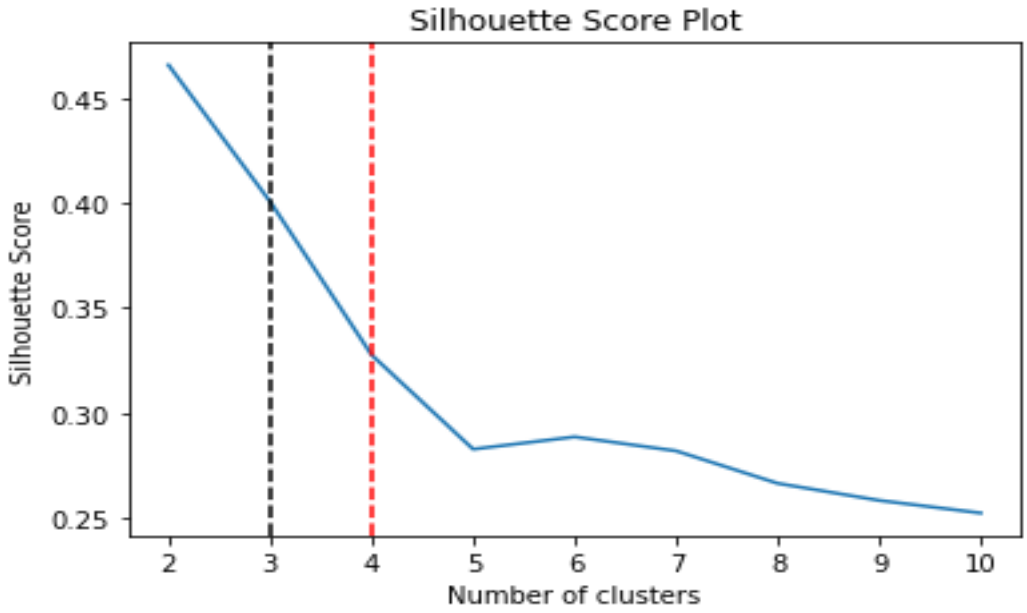
- The wss Score for 1 clusters is 1469.99
- The wss Score for 2 clusters is 659.17
- The wss Score for 3 clusters is 430.65
- The wss Score for 4 clusters is 371.38
- The wss Score for 5 clusters is 327.21

- The wss Score for 6 clusters is 289.31
- The wss Score for 7 clusters is 262.98



Looking at WSS plot there is a sharp decrease from cluster 2 -3 & Elb ow occurs at this point so we use 3 Clusters for Clustering.

- **Silhouette Score Plot**



The Average Silhouette Score for 2 clusters is 0.46577
The Average Silhouette Score for 3 clusters is 0.40073
The Average Silhouette Score for 4 clusters is 0.32765
The Average Silhouette Score for 5 clusters is 0.28273
The Average Silhouette Score for 6 clusters is 0.2886
The Average Silhouette Score for 7 clusters is 0.28191
The Average Silhouette Score for 8 clusters is 0.26644
The Average Silhouette Score for 9 clusters is 0.25831

From python code we have plotted Silhouette Score for Various number of Clusters and this plot is having signifi cant drop from cluster 2 to 3 and 3-4 & 4-5 suggesting th at optimum number of cluster can be 4 or 5 or 3.

- Analysis for N= 4 Number of clusters

❖ Number of Elements in each Clusters

- 0- 65
- 1- 66
- 2- 30
- 3- 49

Clusters at considering N=4 Clusters				
Clus_kmeans4	0	1	2	3
spending	11.82	13.99	16.32	19.12
advance payments	13.24	14.11	15.29	16.46
probability_of_full_payment	0.85	0.88	0.88	0.89
current balance	5.24	5.43	5.86	6.27
credit limit	2.83	3.21	3.44	3.77
min_payment_amt	4.92	2.59	3.87	3.47
max_spent_in_single_shopping	5.12	5.03	5.69	6.13
freq	65.00	66.00	30.00	49.00

As Our idea for Clustering is find out that numbers of clusters o n which can clearly distinguish the data and It can be easily expl ained to Marketing Peoples.

So we are Going with N = 3 numbers of clusters

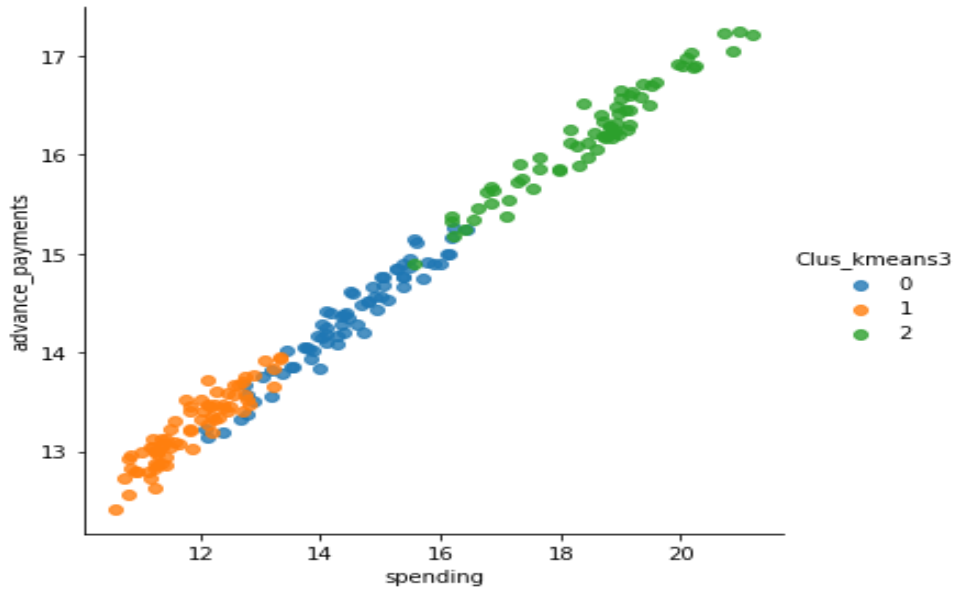
- Analysis for N=3 Number of Clusters Via Kmenas

Clusters at considering N=3 Clusters			
Clus_kmeans3	0	1	2
spending	14.44	11.86	18.50
advance payments	14.34	13.25	16.20
probability_of_full_payment	0.88	0.85	0.88
current balance	5.51	5.23	6.18
credit limit	3.26	2.85	3.70
min_payment_amt	2.71	4.74	3.63
max_spent_in_single_shopping	5.12	5.10	6.04
freq	71.00	72.00	67.00

As we can see from above table that Cluster 0 is having 71 elements i n them cluster 1 is having 72 elements also cluster no. 2 is having 67 elements . After having a closer look on above table we have noticed that cluster No. 2 is having high average spending & their credit limit and Max spent in single shopping is also high .

- **Inference from the Model**

By looking at Wss and Silhouette score plot We found out the optimu m number of clusters, for our easy understanding we have considere d our clusters at N=3. Below graph is showing linear model of clusters between advance payment and spending.



For Our easy understanding we have plotted clusters as below. 3 diffe rent Colours are showing 3 different clusters.

1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters in context to the business problem in-hand .

As per our initial understanding Kmeans algorithm is an iterative algorithm that tries to partition the dataset into Kpre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.

As we have seen from all of above Questions 3 clusters are optimum because we can easily group all of them.

Clusters at considering N=3 Clusters			
Clus_kmeans3	0	1	2
spending	14.44	11.86	18.50
advance payments	14.34	13.25	16.20
probability_of_full_payment	0.88	0.85	0.88
current balance	5.51	5.23	6.18
credit limit	3.26	2.85	3.70
min_payment_amt	2.71	4.74	3.63
max_spent_in_single_shopping	5.12	5.10	6.04
Frequency of elements	71.00	72.00	67.00

As we can see that Group 2 is having high spending, advance payment and high probability of full payment & their Credit limit is also high in this case , Group 2 peoples are also spent maximum amount in their single shopping .

Cluster Group no.1 is having lowest spending, advance payment and lowest probability of full payment & their Credit limit is also lowest among all.

Considering at above two points we can divide it into 3 Groups.

Group 2:- Highest spending Group

Group 0 :- Medium Spending Group

Group 1 :- Lowest Spending Group

❖Promotional strategies for different Groups

▪Group 2:- Highest spending Group

- We have seen from the table that Highest spending group having high current balance and advance payment but the ir minimum payment amount is lowest among all the thre e groups so we can offer a significant discount for making full payment.
- Since high spending group is having high money to spend We can also offer a combination of products to them and give gifts or Cashback for their purchases.
- We can also increase their credit limits and as they are ha ving highest probability of repayment so we can offer loa n on credit card so that we can earn more interest from th em.
- We can also sign contract with online ecommerce repute d brand Like Flipkart , Amazon etc. so that we can increase their spending and maximum spending on their tansaction s.

▪Group 0:- Medium spending Group

- Since these group of customers is having medium spendin g, advance payment and medium probability of full paym

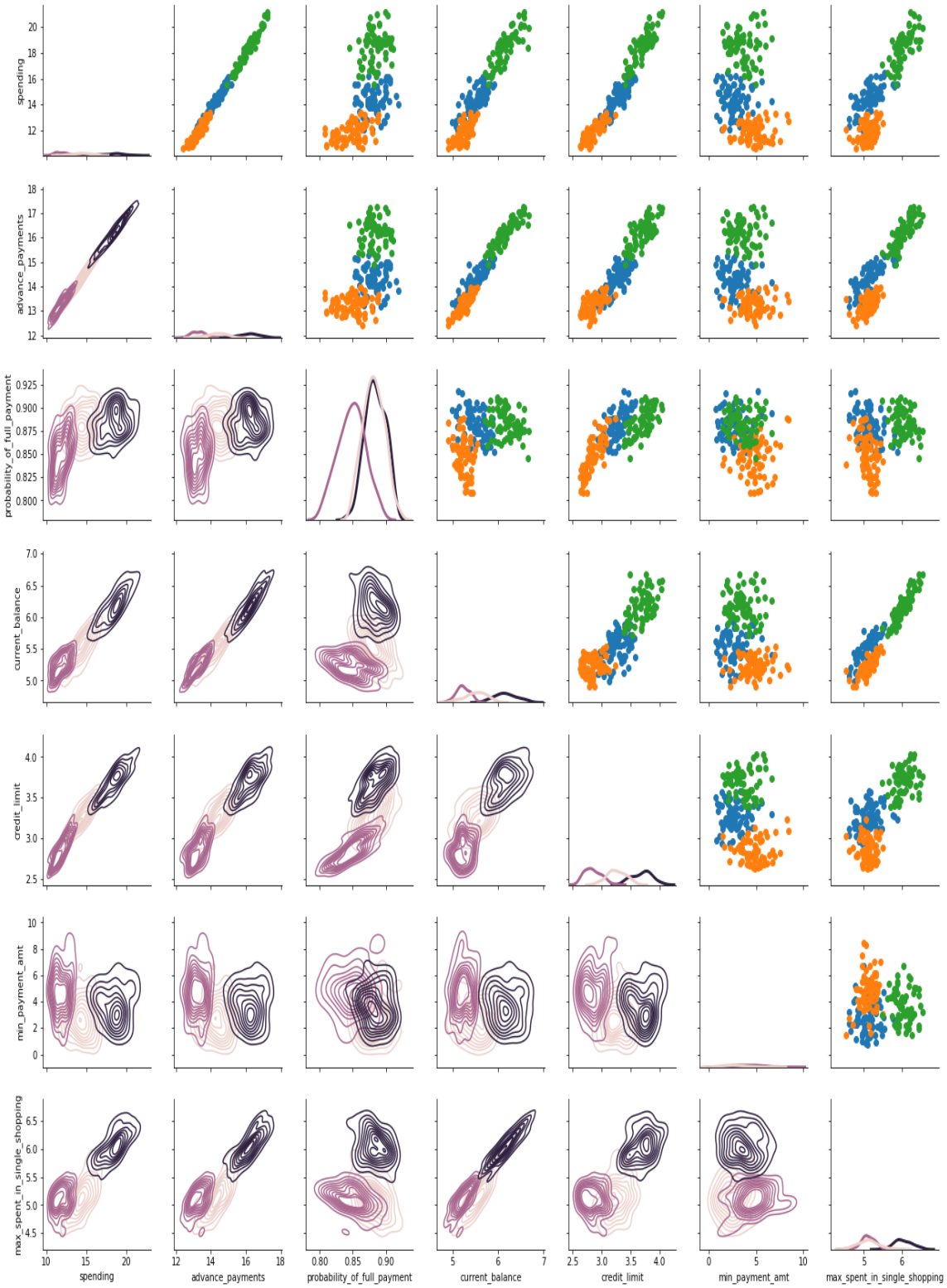
ent & their Credit limit is also medium among all. So these all can be of Our focus for increasing their spending.

- We can also send emails/Message to them for their loyali ty and also include offers for increasing their habbits.
- We can also offer an extra % interest free loans for their h ouse purchases, Ecommerce shopping.
- We can also tie up with online travel brands like yatra.co m, Goibibo.com & payment wallet brands like paytm etc. f or increasing their spending habits.

- **Group 1:- Lowest spending Group**

  - Since these group of customers is having lowest spending, advance payment and lowest probability of full payment & their Credit limit is also lowest among all. So We can give offers or reward points for their loan repayments.
  - We can also send them reminders via various sources for t heir Loan repayments.
  - Since these Customers are having lowest spending habit s o one strategy should be link their municipality tax, electri city bills and we can also offer insurance so that they can s ave their income tax etc.

❖ Let's see Clustering effect visually to validate above written points. All the three Colours shows different clusters.


spending
advance_payments
probability_of_full_payment
current_balance
credit_limit
min_payment_amt
max_spent_in_single_shopping