Bank Customer Segmentation

Table of Contents

Contents

Bank Customer Segmentation

Summary:

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.

Sample of the Dataset:

	spending	advance_payments	probability_of_full_payment	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	19.94	16.92	0.8752	6.675	3.763	3.252	6.550
1	15.99	14.89	0.9064	5.363	3.582	3.336	5.144
2	18.95	16.42	0.8829	6.248	3.755	3.368	6.148
3	10.83	12.96	0.8099	5.278	2.641	5.182	5.185
4	17.99	15.86	0.8992	5.890	3.694	2.068	5.837

Exploratory Data Analysis:

Let us check the basic info of the data frame

```
<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
```

Questions

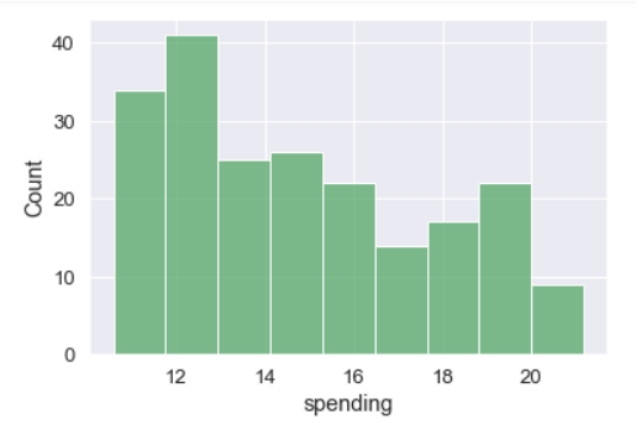
1.1 Read the data, do the necessary initial steps, and exploratory data analysis (Univariate, Bi-variate, and multivariate analysis).

Univariate Analysis

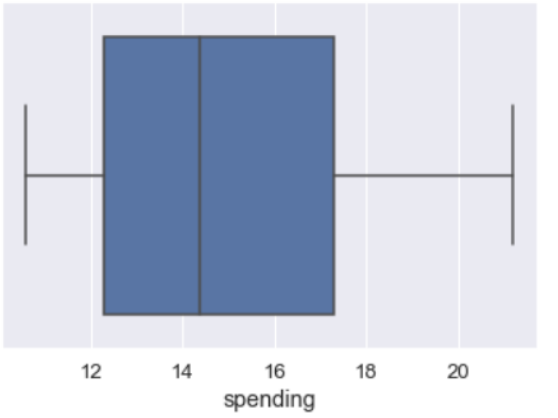
This analysis will display the statistical description of the numeric variable to view 5-point summary, histogram or distplot to view the distribution and the box plot to view outliers if any

- spending

```
count    210.000000
mean      14.847524
std        2.909699
min       10.590000
25%       12.270000
50%       14.355000
75%       17.305000
max       21.180000
Name: spending, dtype: float64 Distribution of spending
----------------------------------------------------------------
```

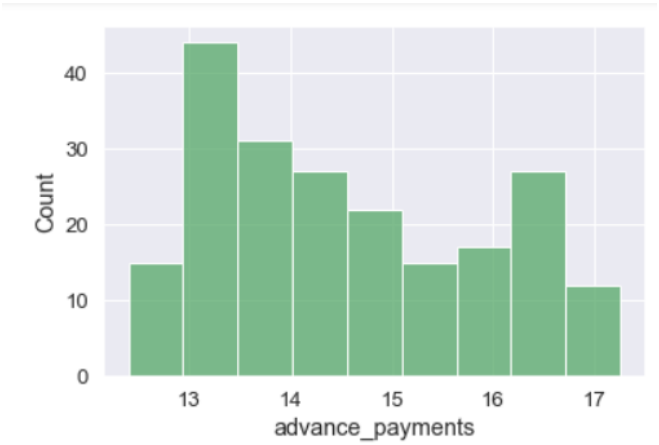


BoxPlot of spending

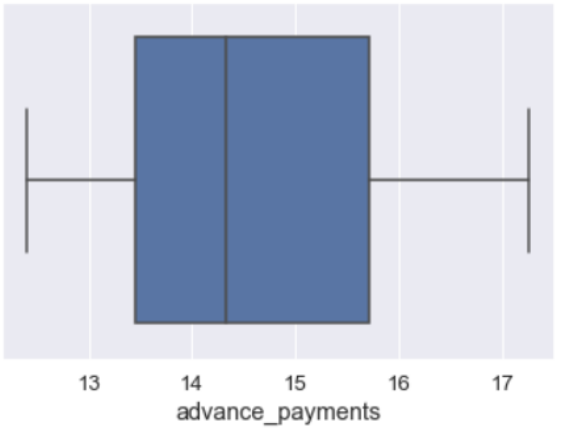


- advance_payments

```
Description of advance_payments
-------------------------------------------------------------------------------
count    210.000000
mean      14.559286
std        1.305959
min       12.410000
25%       13.450000
50%       14.320000
75%       15.715000
max       17.250000
Name: advance_payments, dtype: float64 Distribution of advance_payments
-------------------------------------------------------------------------------
```

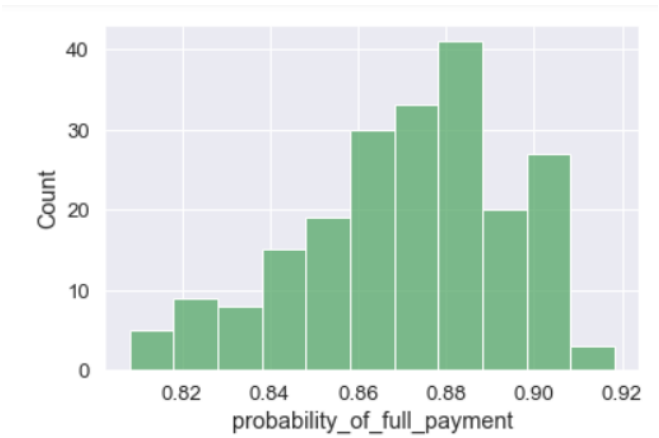



```
BoxPlot of advance_payments
-------------------------------------------
```

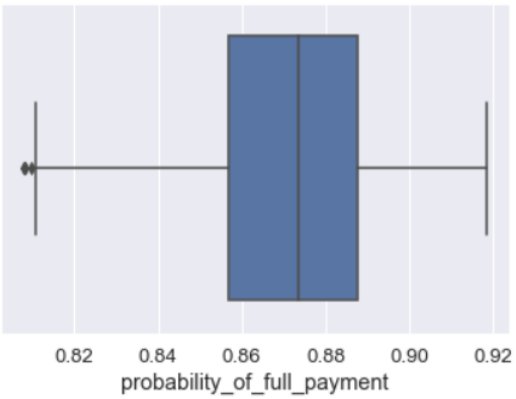


- probability_of_full_payment

```
Description of probability_of_full_payment
-----------------------------------------------------------------------------
count    210.000000
mean       0.870999
std        0.023629
min        0.808100
25%        0.856900
50%        0.873450
75%        0.887775
max        0.918300
Name: probability_of_full_payment, dtype: float64 Distribution of probability_of_full_payment
-----------------------------------------------------------------------------
```



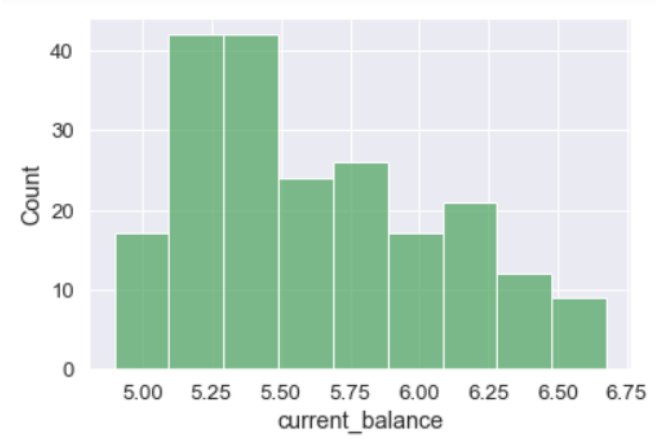
BoxPlot of probability_of_full_payment

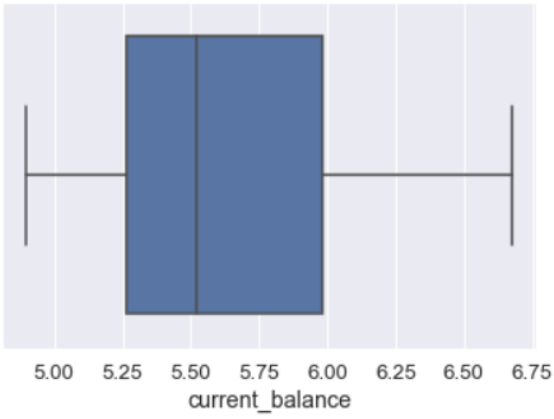


- current_balance

```
Description of current_balance
-------------------------------------------------------------------------
count    210.000000
mean       5.628533
std        0.443063
min        4.899000
25%        5.262250
50%        5.523500
75%        5.979750
max        6.675000
Name: current_balance, dtype: float64 Distribution of current_balance
-------------------------------------------------------------------------
```

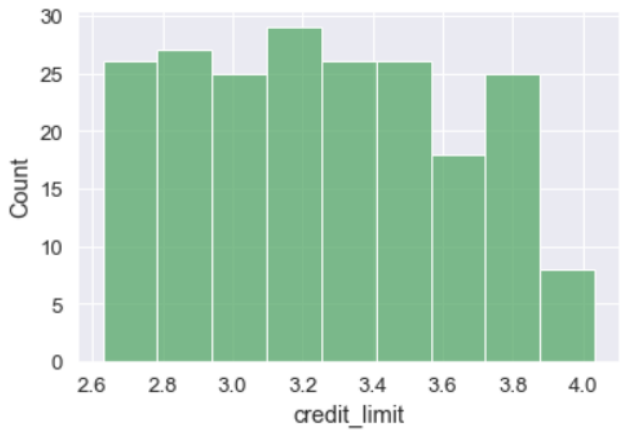
BoxPlot of current_balance



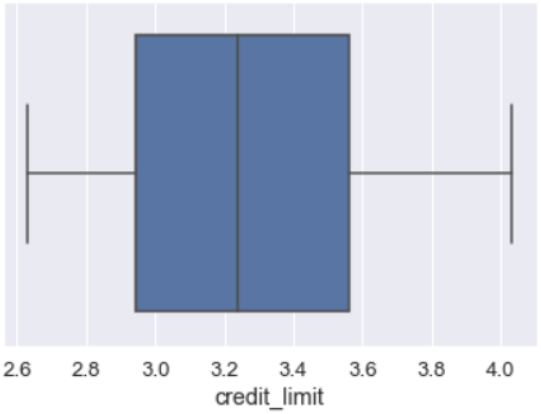


- credit_limit

```
Description of credit_limit
-----------------------------------------------------------------
count    210.000000
mean       3.258605
std        0.377714
min        2.630000
25%        2.944000
50%        3.237000
75%        3.561750
max        4.033000
Name: credit_limit, dtype: float64 Distribution of credit_limit
-----------------------------------------------------------------
```



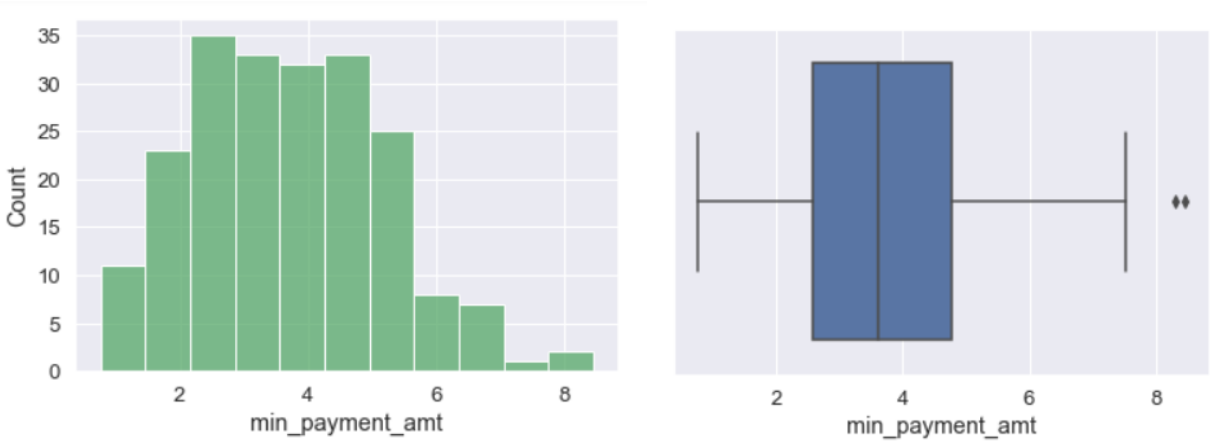
BoxPlot of credit_limit



- min_payment_amount

```
Description of min_payment_amt
-----------------------------------------------------------------------
count    210.000000
mean       3.700201
std        1.503557
min        0.765100
25%        2.561500
50%        3.599000
75%        4.768750
max        8.456000
Name: min_payment_amt, dtype: float64 Distribution of min_payment_amt
-----------------------------------------------------------------------
```

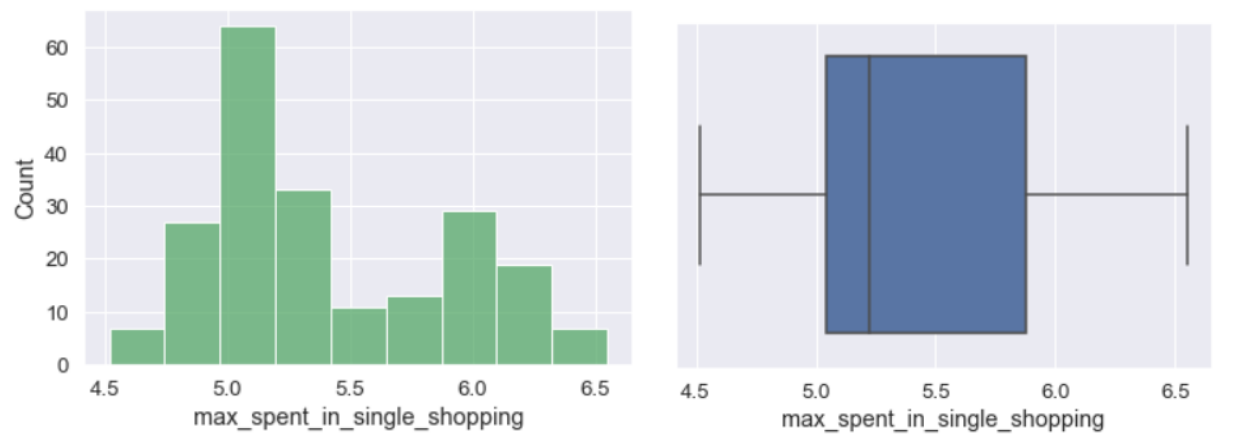
BoxPlot of min_payment_amt



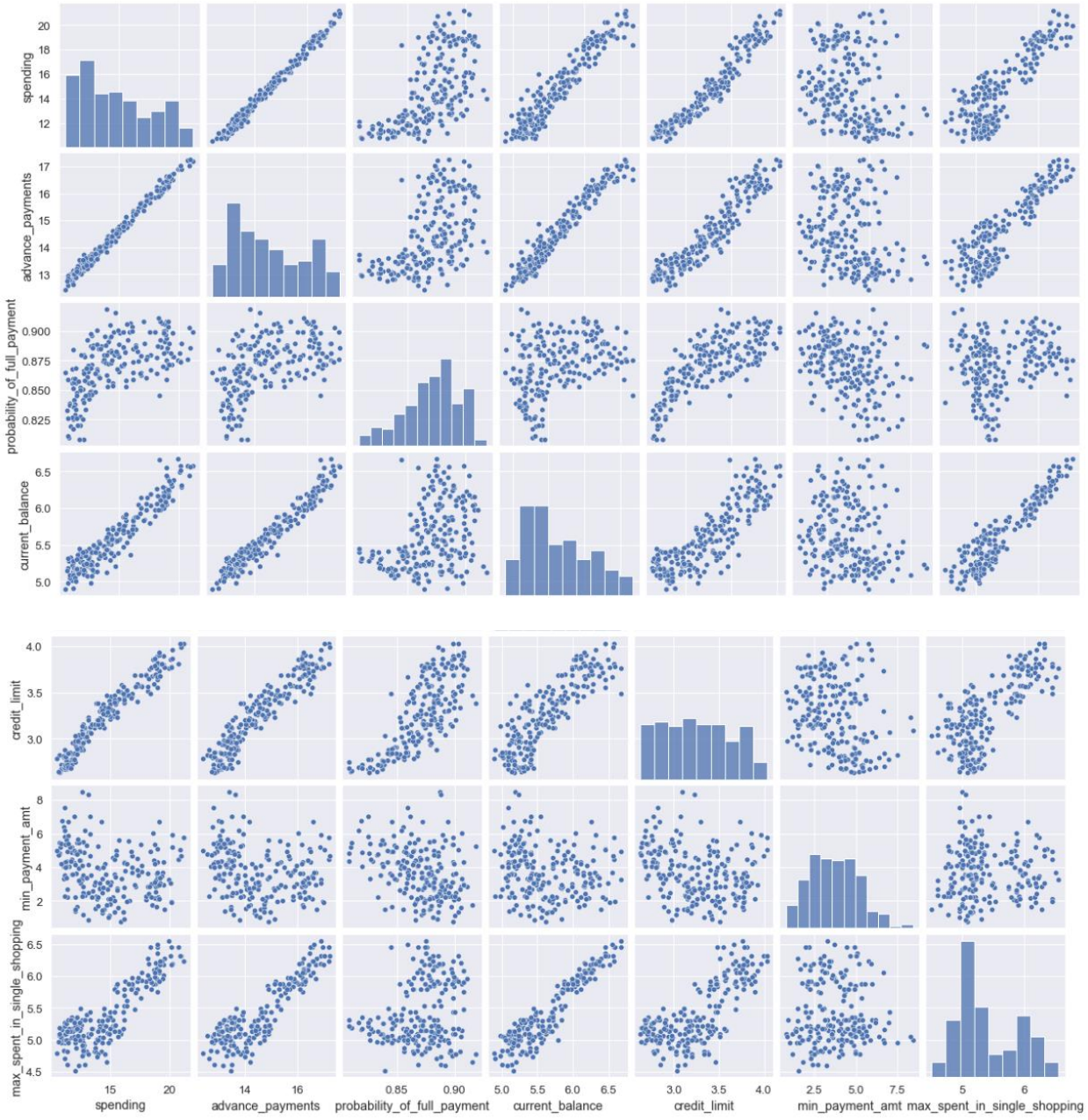
- max_spent_in_single_shopping

```
Description of max_spent_in_single_shopping
-------------------------------------------------------------------------------
count    210.000000
mean       5.408071
std        0.491480
min        4.519000
25%        5.045000
50%        5.223000
75%        5.877000
max        6.550000
Name: max_spent_in_single_shopping, dtype: float64 Distribution of max_spent_in_single_shopping
-------------------------------------------------------------------------------
```

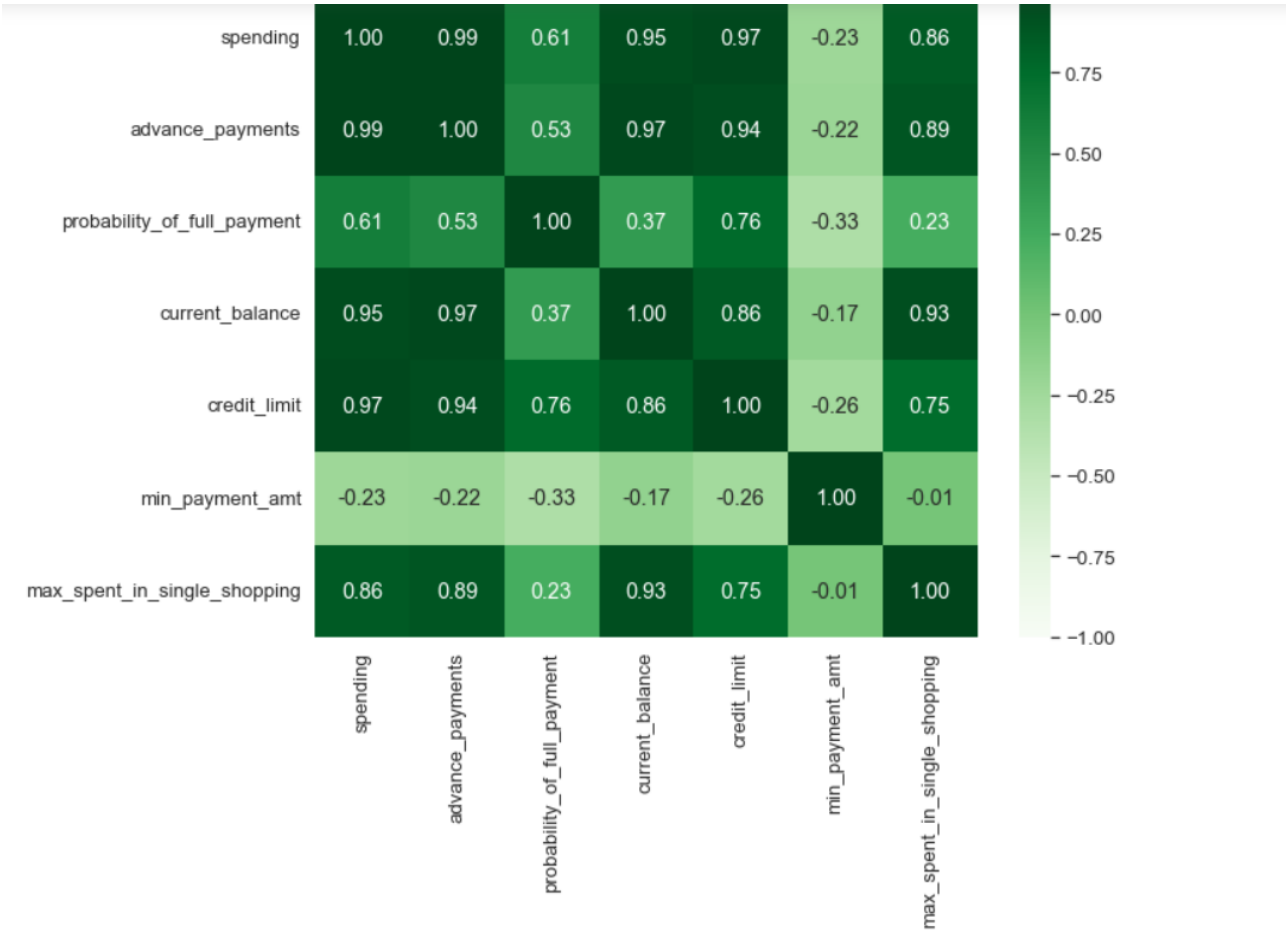
BoxPlot of max_spent_in_single_shopping



Multivariate Analysis



Correlation Heatmap



Insights: --

- The dataset has 7 columns and 210 rows
- There are no Null values as indicated by non-null values
- All columns are of the float data type
- There are no duplicate rows in the dataset
- A few features are right skewed (spending, current_balance, max_spent_in_single_shopping)
- A few features are left skewed (probability_of_full_payment,)
- A few features seem normally distributed (credit_limit, min_payment_amt)
- Only the min_payment_amt feature has Outliers as demonstrated by the box plots and as there are only 2 of them, we'll not be removing them
- We have plotted scatter diagrams for all the numerical columns in the dataset. A scatter plot is a visual representation of the degree of correlation between any two columns

- We've also plotted a heatmap to display the numerical values of the degree of correlation between any two columns; apart from probability_of_full_payment, the rest of the features are highly correlated with each other
- This could make Clustering quite difficult; however, lets how our model makes the segregation

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

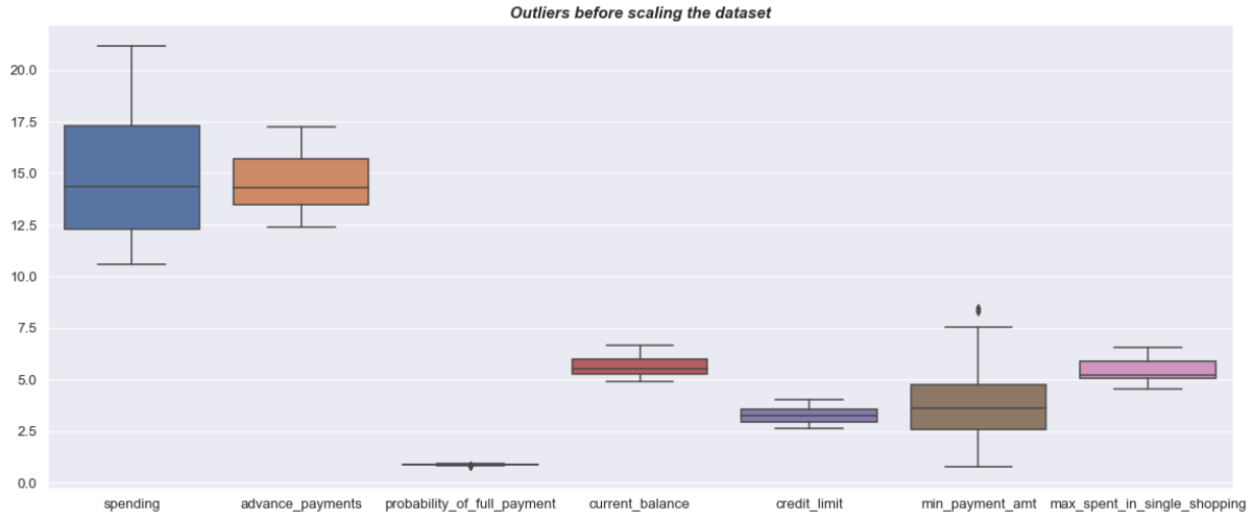
- Often the variables of the data set are of different scales i.e., one variable is in millions and other in only 100.
- In our data set, the values range from a minimum of 0.77 in min_payment_amt to a maximum of 21.18 in spending. Since the variation in data is huge, it is tough to compare these variables
- Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing
- In this method, we convert variables with different scales of measurements into a single scale and will be doing this only for the numerical variables
- StandardScaler normalizes the data using the formula (x-mean)/standard deviation
- We can either use StandardScaler for each and every feature or apply the z-score (both
- methods will give us the same result)
- Once scaled, all values will be in the range of -3 to +3 in our dataset (see boxplot below)
- Also, we will not be scaling the probability column as we do not want the feature to be in the above range and this will also help us make better inferences
- I've added the probability column with the original values at the end in the scaled dataset
- As we mentioned earlier, we will not be removing the 2 outliers in the min_payment_amt column

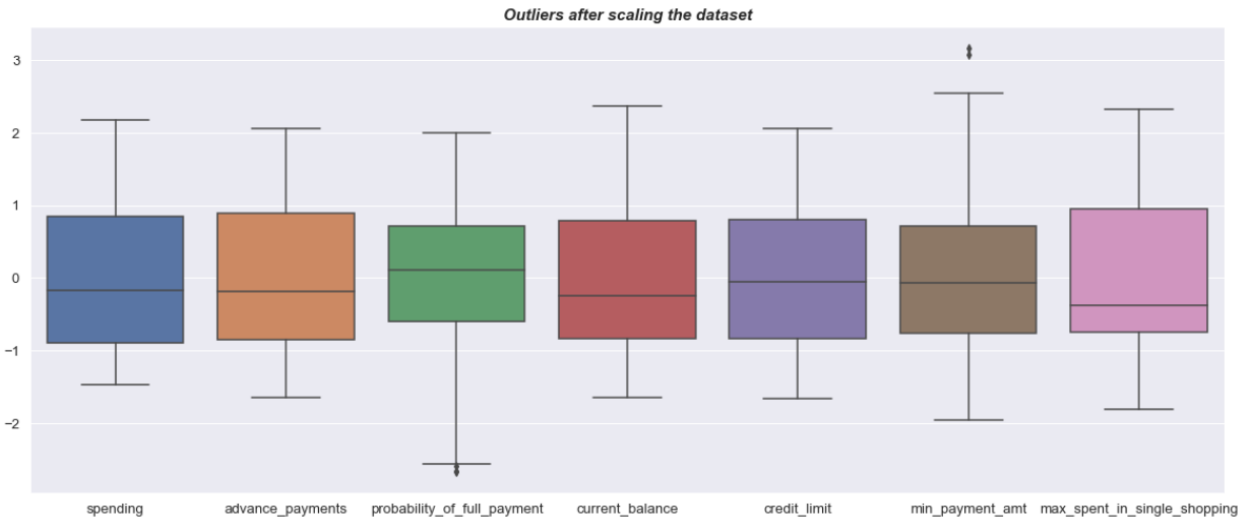
Below is the statistical summary of the dataset before Scaling

	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
25%	12.27	13.45	0.86	5.26	2.94	2.56	5.04
50%	14.36	14.32	0.87	5.52	3.24	3.60	5.22
75%	17.30	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

Below is the statistical summary of the scaled dataset

	spending	advance_payments	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping
0	1.754355	1.811968	2.367533	1.338579	-0.298806	2.328998
1	0.393582	0.253840	-0.600744	0.858236	-0.242805	-0.538582
2	1.413300	1.428192	1.401485	1.317348	-0.221471	1.509107
3	-1.384034	-1.227533	-0.793049	-1.639017	0.987884	-0.454961
4	1.082581	0.998364	0.591544	1.155464	-1.088154	0.874813





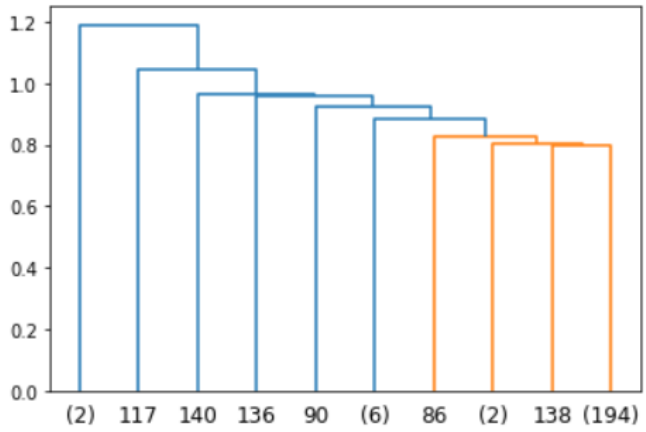
Introduction to Clustering

- Clustering is an Unsupervised Learning Technique for finding similar groups in data, called clusters
- Clustering helps simplify data by reducing many data points into a few segments
- To define **"Similar"** objects you need measure the distance
- Distance between points (smaller the difference = similar, higher the difference = dissimilar)

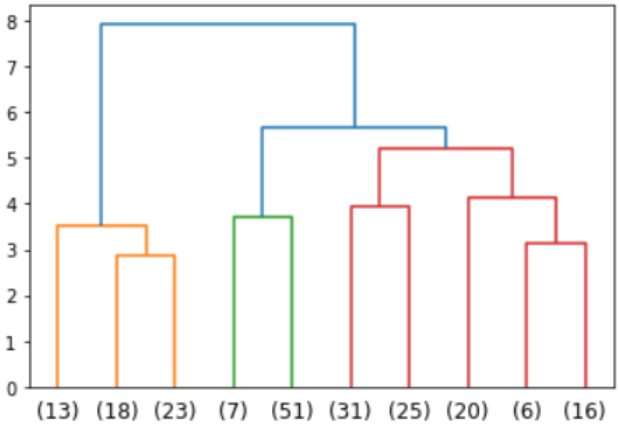
1.3 Apply hierarchical clustering to scaled data. Identify the number of optimum clusters using Dendrogram and briefly describe them

- The linkage methods work by calculating the distances or similarities between all objects.
- Then the closest pair of clusters are combined into a single cluster, reducing the number of clusters remaining.
- There are different linkage methods namely

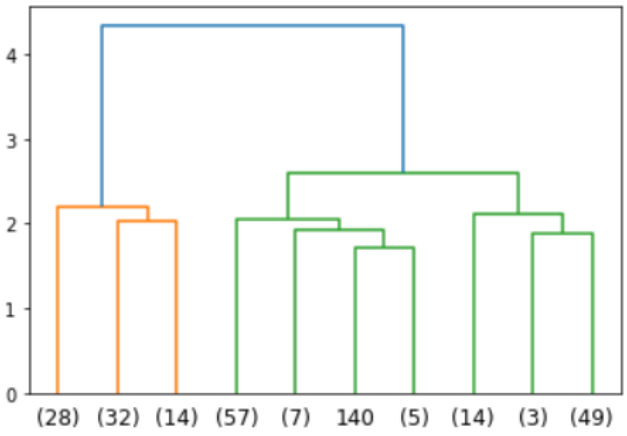
- **Single linkage** – Minimum distance or Nearest neighbor



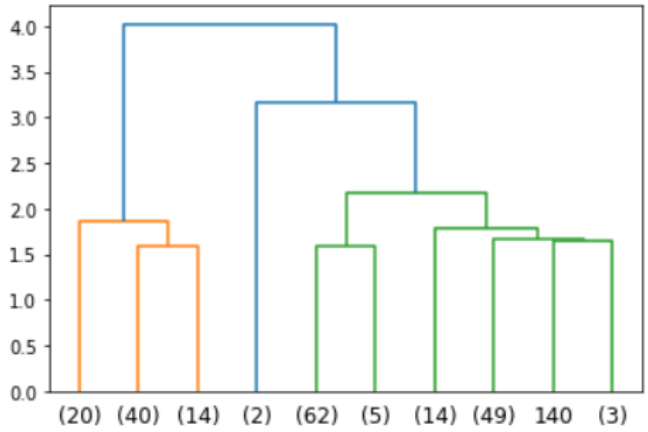
- **Complete linkage** – Maximum distance or Farthest distance



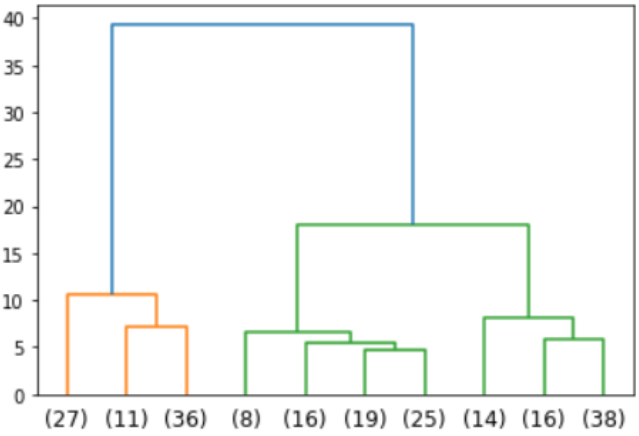
- **Average linkage** – Average of the distances between all pairs



- **Centroid method** – combine cluster with minimum distance between the centroids of the two clusters



- **Ward's method** – Combine clusters with which the increase in within cluster variance is to the smallest degree



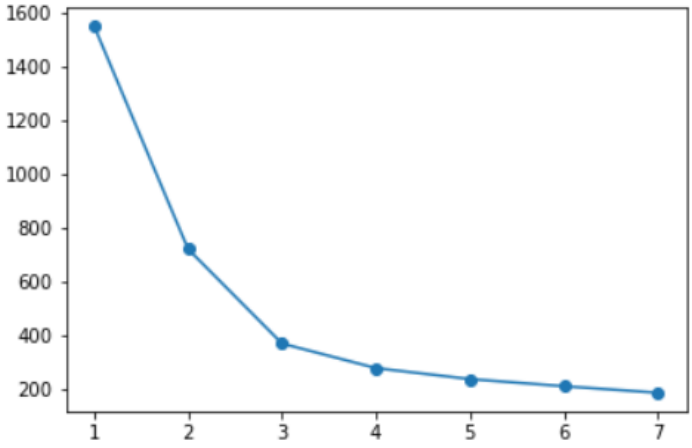
- A Dendrogram is a diagram that shows the hierarchical relationship between objects
- Its main function is to work out the best way to allocate the observations in to clusters
- Based on the above dendrograms, I'm selecting the "Ward" linkage method as it gives us 2 distinct clusters
- The advantage of the Ward method is it keeps the similar customers together and also takes the maximum distance between the clusters to give us the optimum output
- As we can see from the below, we have divided the bank's customers in to 2 tiers with 74 and 136 observations each
- This is where not scaling the probability feature comes in handy as it ranges from 0 to 1 and easier to make inferences

clusters	spending	advance_payments	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	probability_of_full_payment	Freq
1	1.169823	1.180191	1.167632	1.084241	-0.079694	1.214174	0.883335	74
2	-0.636521	-0.642163	-0.635329	-0.589955	0.043363	-0.660654	0.864286	136

1.4 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.

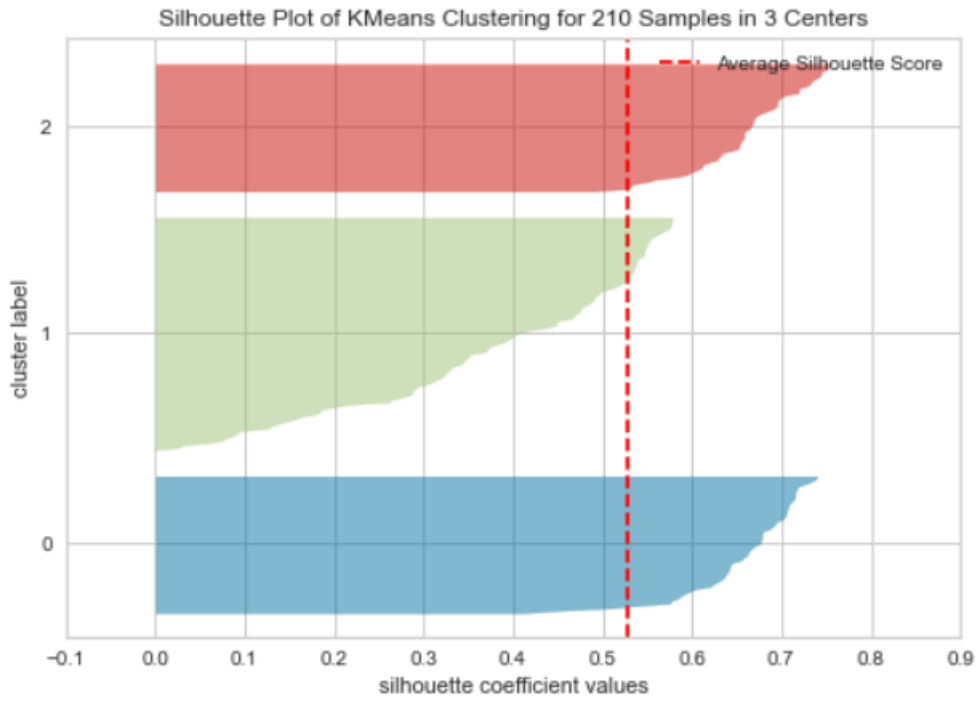
- In K-Means clustering, we assume the number of clusters to be formed and determine the optimum value through iteration
- If K = 3, it means we're assuming there are 3 Centroids and the algorithm computes Euclidean distance of each observation with these Centroids and then assigns the observations to each Centroid (clusters) with shortest distance
- We use Within Sum of Squares (WSS), which is the sum of distances between the points and the corresponding centroids for each cluster to arrive at the optimum k-value
- We've computed the WSS through the Elbow method from 1 to 7 clusters and found that as the K-values increases, the WSS decreases
- From 1 to 2, and then 3, there is a significant drop hence 2 and 3 is a valuable addition in k-means algorithm. The below Elbow-curve confirms the same
- Hence, K = 3 is the optimum value as the decrease in WSS on adding additional clusters isn't substantial

```
[1551.2785997714768,
 721.6727517280757,
 370.1026268096892,
 278.63134210774933,
 237.79237957433884,
 211.2792182777221,
 186.83122033149942]
```



- Another method to determine the appropriate k-value is the Silhouette score
- Silhouette score is used to evaluate the quality of clusters created using clustering algorithms such as K-Means in terms of how well samples are clustered with other samples that are similar to each other. The Silhouette score is calculated for each sample of different clusters
- Sil-width = (b-a)/max(a,b) where
    - a = distance between observation and its own cluster centroid(c1)
    - b = distance between observation and the neighbour cluster centroid (c2)
- If sil-width is a positive value, then we say the mapping of the observation is correct to its current centroid.
- If the silhouette score is close to +1 then we can say the clusters are well separated from each other on an average
- As we can see from the below K = 3 is the nearest to 1 with a score of 0.528 and we can conclude that 3 is the optimum number of clusters
- We have visualized the samples being split into 3 clusters with the optimal silhouette score

```
Silhouette Score for k =  2 is 0.478
Silhouette Score for k =  3 is 0.528
Silhouette Score for k =  4 is 0.478
Silhouette Score for k =  5 is 0.457
Silhouette Score for k =  6 is 0.395
Silhouette Score for k =  7 is 0.398
```



Clus_kmeans4	spending	advance_payments	current_balance	credit_limit	min_payment_amt	max_spent_in_single_shopping	probability_of_full_payment	freq
0	-1.069352	-1.050864	-0.942446	-1.127652	0.875744	-0.638695	0.847505	58
1	-0.144399	-0.153353	-0.216828	-0.042008	-0.444569	-0.407885	0.876990	98
2	1.410621	1.407014	1.405759	1.287419	-0.133803	1.426242	0.885359	54

- **Hence, we can conclude that the K-Means gives us a better output compared to Hierarchical method as the customers are split into 3 distinct categories and we can make inferences from each**

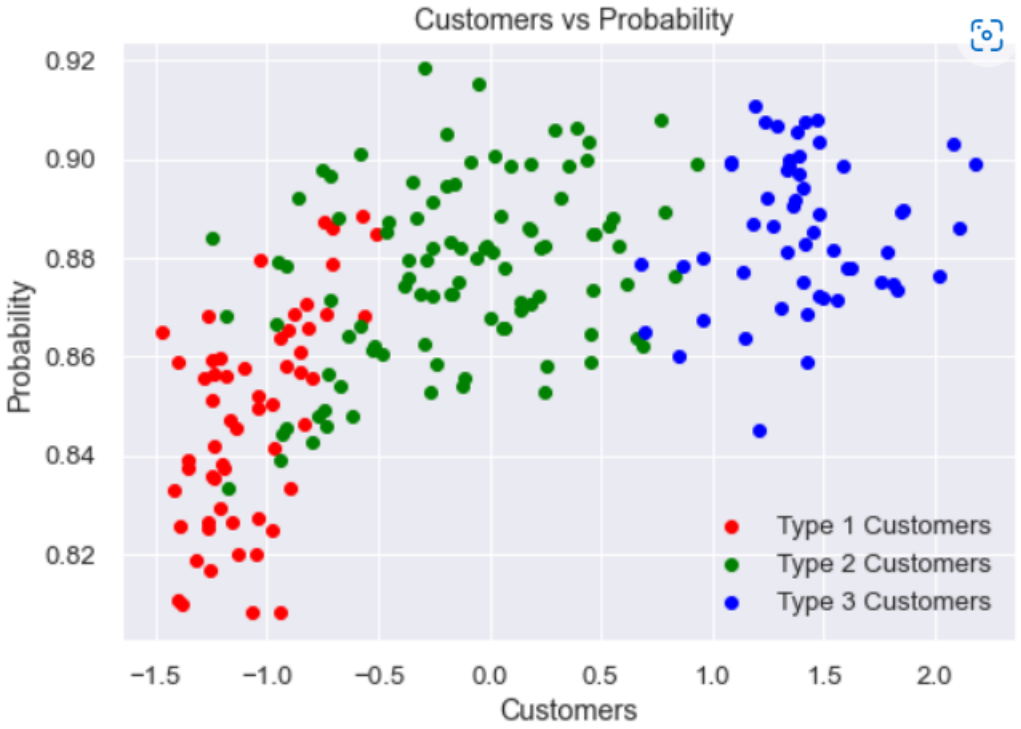
1.5 Describe cluster profiles for the clusters defined. Recommend different promotional strategies for different clusters.

- We were able to segregate the customers into 3 different tiers using the k-means technique
- All customers have a very high probability of more than 0.8 which is a very good sign for the bank
- We can summarize our 3 groups as below
    - Tier 1 – There are 54 customers and all features are positive and have the highest probability of full payment
    - Tier 2 – There are 98 customers and have the 2nd highest probability of full payment
    - Tier 3 – There are 58 customers and have the lowest probability of full payment

- Tier 1 is your premium set of customers with the important features (advance payment, current balance and credit limit) being positive and the highest
    - These customers seem to be high earners and high spenders

    - The bank should make an effort to increase these types of customers by providing personalized offers and memberships to encourage them to spend more (they have the highest minimum payment per purchase) – this would increase the revenue for the bank as more cash is paid during a purchase
    - They also have the highest amount spent in a single purchase; the bank could get more details on the purchases to identify trends and could make offers that are customized to each customer
    - The bank should try to increase the credit limit for them as they are capable of repaying on time (they have the highest advance payments); perhaps the limit can be increased for international trips or a foreign education loan
    - As they may not be too keen on increasing their savings balance, the bank could approach them to open a trading account with them and also offer a personal advisor to educate them on their financial services
    - If these customers have their own business', which is a high possibility, the bank can have them open salary account for their employees with them
- Tier 2 seems to be the upper middle-class as their probability of payment is quite close to Tier 1; however, they lag them in all the other parameters
    - The amount spent in a single purchase is quite moderate; the offer could include having the customers make a minimum of 5-10 purchases a month to qualify for additional benefits
    - The bank should also try to target their friends and family by providing good offers for referring
    - The discounts and offers should be targeted to their desires like vehicle, marriage, house
    - The bank can help them plan for their kid's future by providing long term options for saving
- Tier 3 ranks the lowest in all features and the bank could do with a little caution among these customers
    - These could be the Gen Z or millennial crowd wherein although the probability of payment is quite high, they rank the lowest in all the other features
    - Email marketing – the bank can send them regular emails and videos educating them about the various financial jargons
    - The amount spent in a single purchase is the least; the offer could include having the customers make a minimum of 15-20 purchases a month to qualify for additional benefits
    - One advantage of this segment is their large friends circle; the bank can encourage them to get their friends along to provide a group discount at the time of opening new accounts
    - The banks can request them to provide ratings and reviews in social media as they are very active in such platforms
    - The banks can encourage them to invest in its mutual funds to save and plan for the future

- We've plotted a scatter plot of the 3 tiers of customers against their probability of full payment



- We've plotted a scatter plot of the 3 tiers of customers and their current account balance