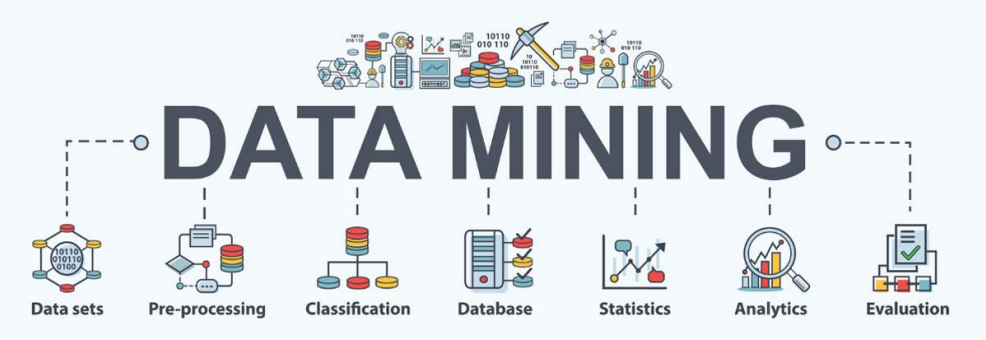
# Data Mining Project

**Submitted by**

**Aruneema**



# Table of Contents

## PLOTS



## Tables

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

The data provided by lending bank looks as follows-

| | 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 |

**Table 1: Sample data for lending bank customer**

Data Dictionary for Market Segmentation:

- 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)

#### 1.1.1 EDA

The data has 210 entries, with 7 columns. These 7 columns can be seen in Table 1 and are all float type. There are no null or duplicated values. The description of these 7 variables are as follows-

| | 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 |

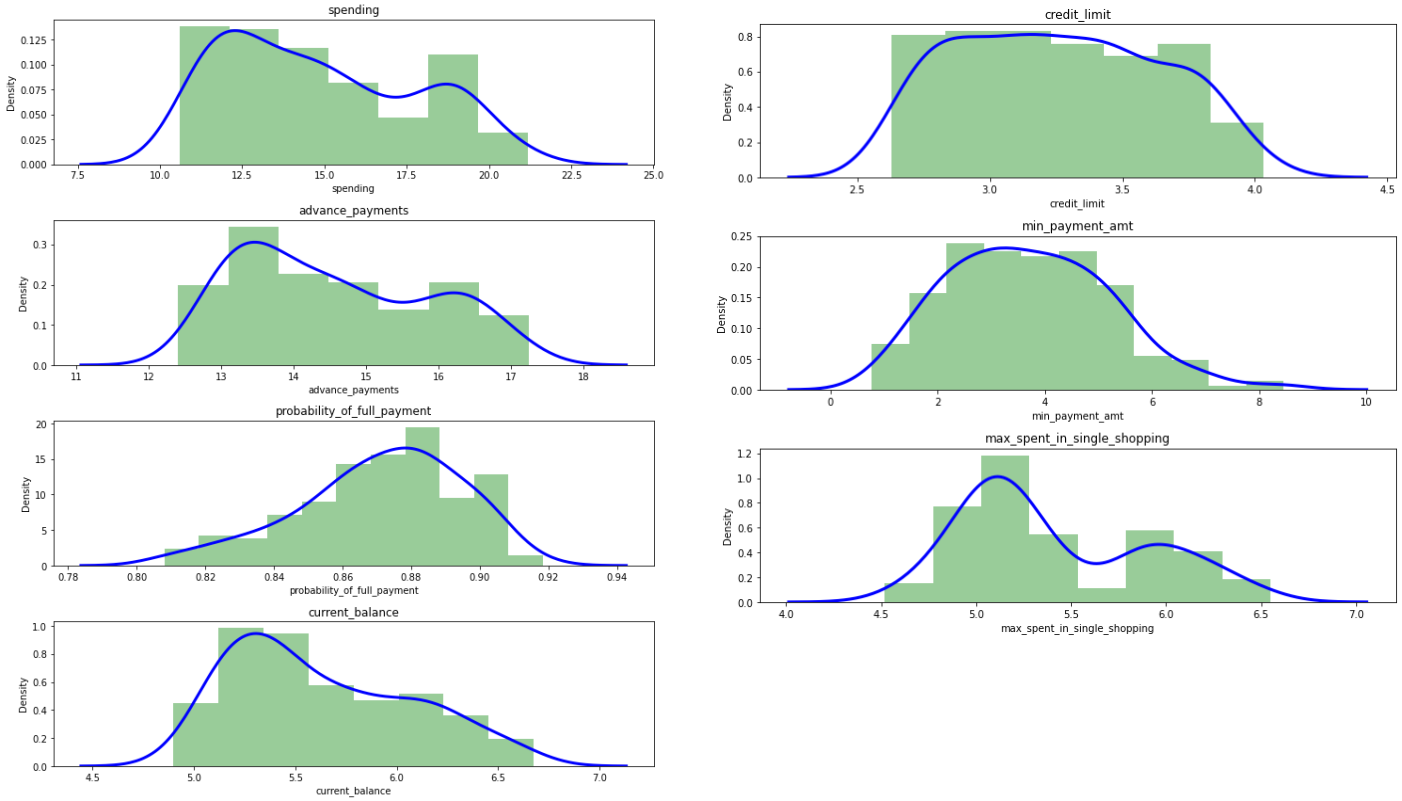
**Table 2: Description table for the data**

Insights:

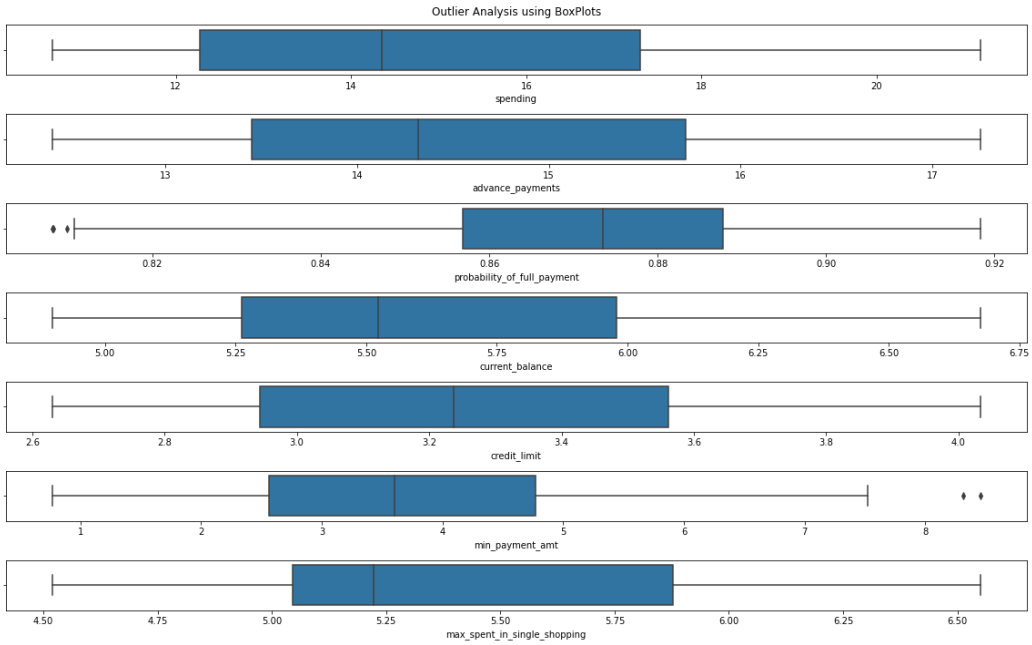
1. We see for most of the variable, mean/medium are nearly equal
2. The standard deviation of spending is higher as compared to other variables

#### 1.1.2 Univariate analysis

For univariate analysis of each variable, a distribution and boxplot was plotted, to visualize the skewness, spread and outlines in the data.

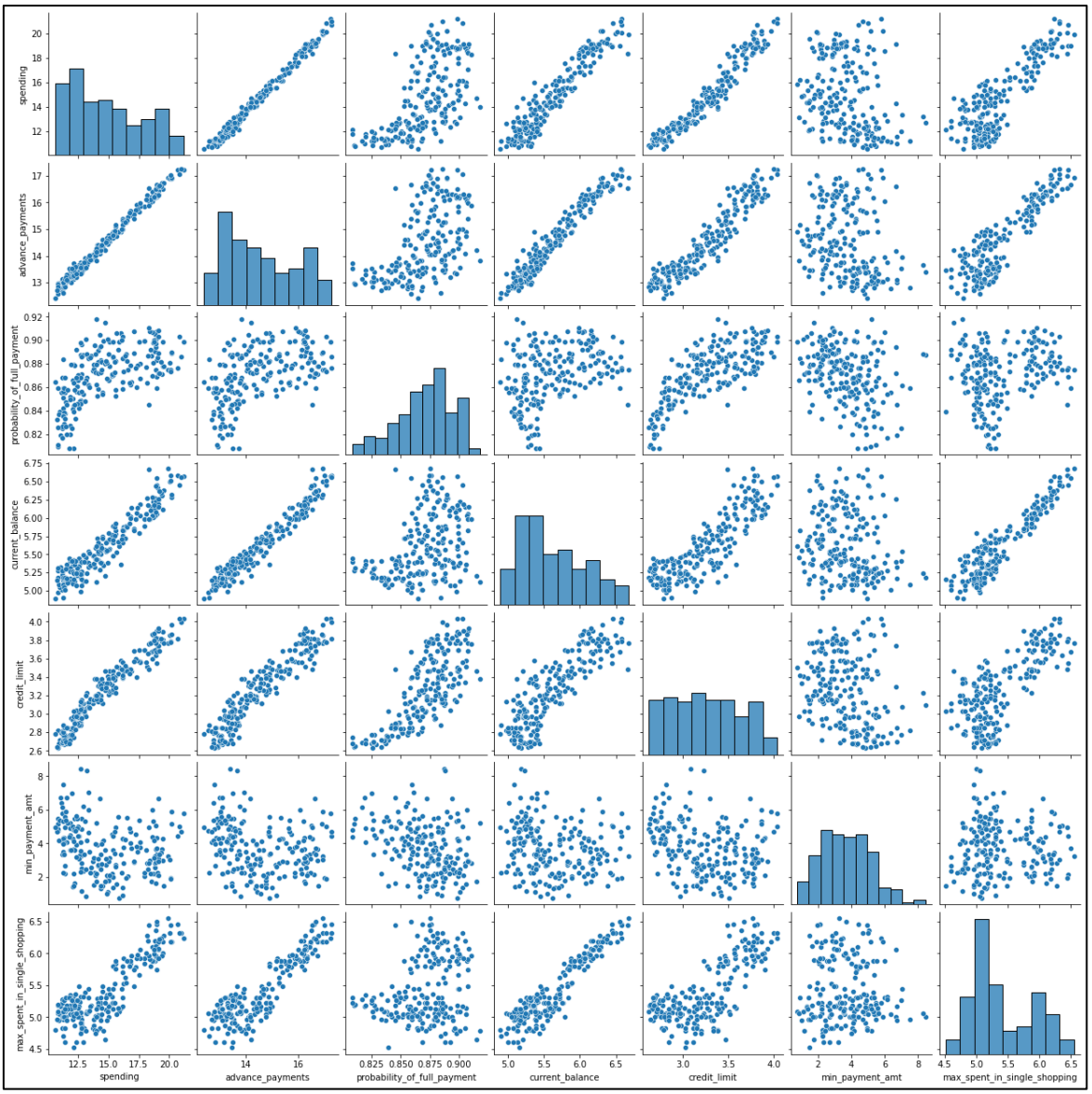



**Plot 1: distplot with skewness of the 7 variables**

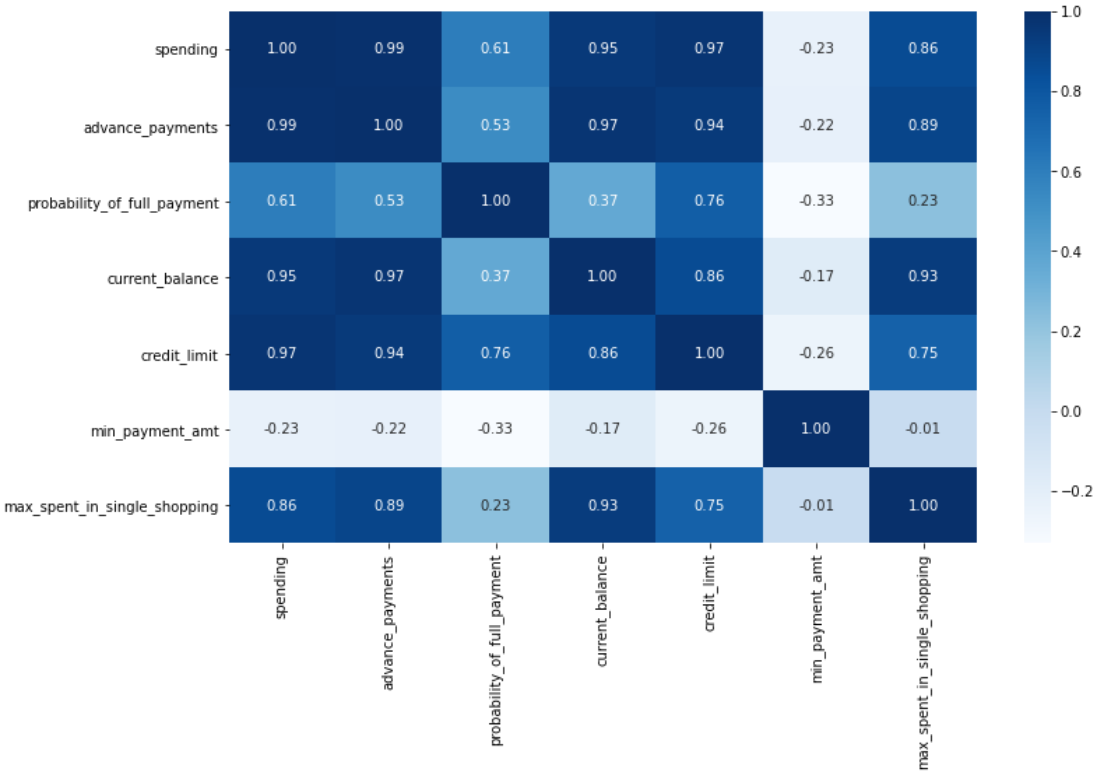



**Plot 2: Boxplot with outliers for seven variables**

## 1.1.3 Bivariate Analysis




**Plot 3: Pairplot of seven variables**



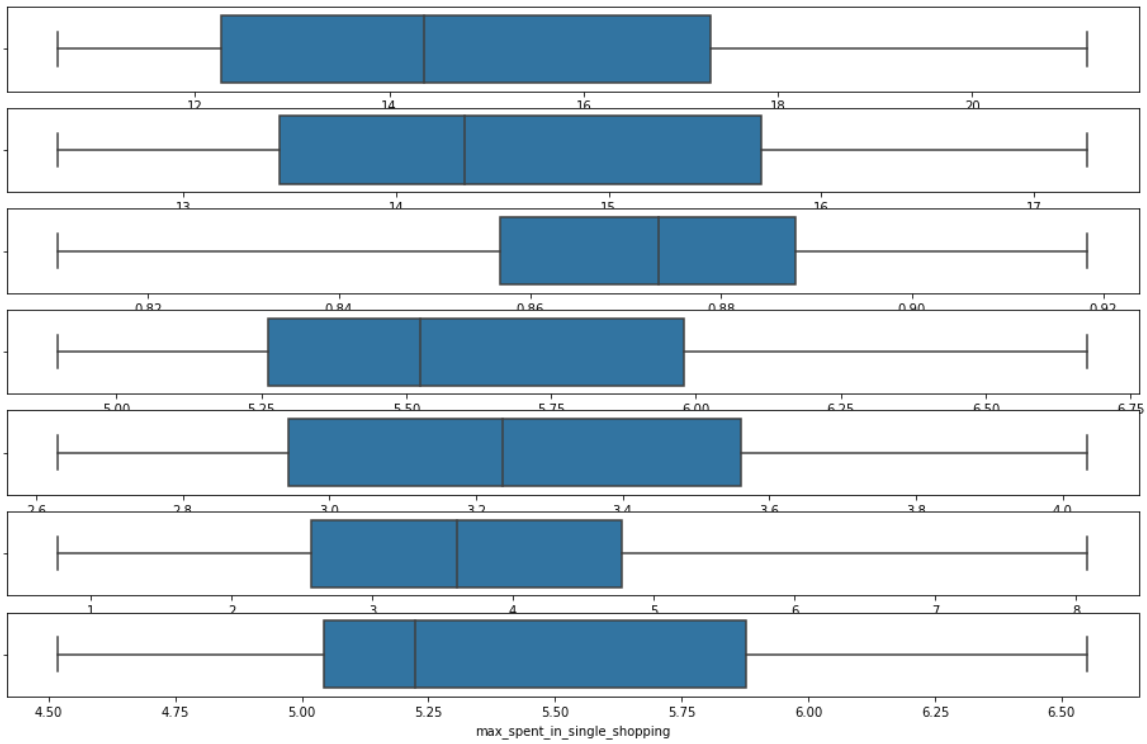

**Plot 4: Correlation heatmap of seven variables**

Following insights can be gathered from Univariate and bivariate analysis:

1. Most variables are moderately and positively skewed except 'probability_of_full_payment' which is negatively skewed.
2. probability_of_full_payment and min_payment_amt have outliers.
3. There is medium to strong correlation between most variables except for min_payment_amt, which shows negative correlation with other variables

### 1.1.4 Outlier Treatment

Since clustering will be performed in further analysis, outliers need to be treated as clustering models are sensitive to outliers that affect the clustering of datapoints.
The method of flooring and capping was used to treat the outliers. In the following boxplot, no outliers are present-




**Plot 5: Box-plot without outliers**

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

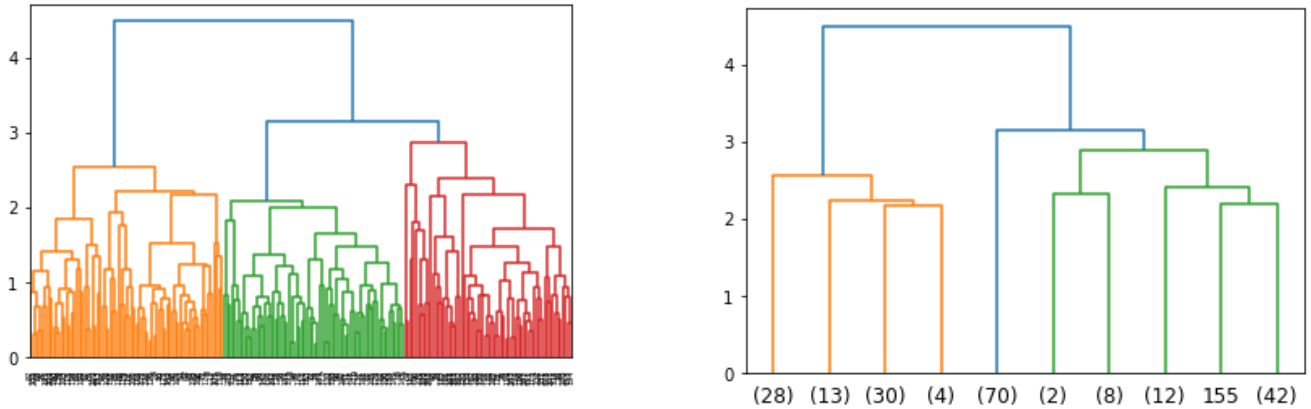
Clustering is a distance based algorithm. All distance based algorithms are affected by the scale of the variables. Since all the 7 variables have different ranges of values, it would be difficult for clustering algorithm to group and compare them. Hence, scaling is done to standardize all values and for easy clustering.

In this case, standard scaler was used to standardize all the variables.

Note: no columns were dropped since all the variables talks about credit spending and managing behaviour and profile of sample consumers. Hence, all the variables will help to cluster the data.

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

To obtain clusters, a dendrogram was constructed, which is shown as below:-




**Plot 6: Dendrogram for hierarchal clustering. The second one is truncated dendrogram**

Two hierarchal clustering methods were used, f-clustering and agglomerative clustering, to discern which model gives better clustering result. Following table shows the clustering results done by f-clustering and agglomerative clustering: -

| clusters | advance_payments | probability_of_full_payment | current_balance | credit_limit | r |
|---|---|---|---|---|---|
| 1 | 16.058000 | 0.881595 | 6.135747 | 3.648120 | |
| 2 | 13.291000 | 0.846845 | 5.258300 | 2.846000 | |
| 3 | 14.195846 | 0.884869 | 5.442000 | 3.253508 | |

**Table 3: Cluster groups formed after F-clustering**

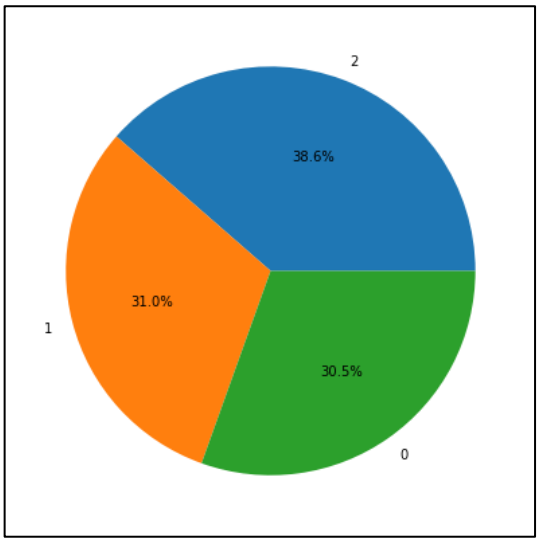
| Agglo_CLusters | spending | advance_payments | probability_of_full_payment | current_balance |
|---|---|---|---|---|
| 0 | 11.845781 | 13.234375 | 0.849104 | 5.226609 |
| 1 | 18.569231 | 16.235077 | 0.884386 | 6.183723 |
| 2 | 14.232716 | 14.261358 | 0.877623 | 5.500580 |

**Table 4: Cluster groups formed after agglomerative clustering**

For the analysis of clusters in this case, agglomerative clustering, with linkage='average' and distance='Euclidean' is being chosen because upon trying f-clustering method and other linkage methods (i.e., ward, complete etc), the agglomerative clustering gives much more distinctly profiled clusters.

### 1.3.1 Cluster Analysis: -

Three clustering profiles are captured, which approximately covers the max, min and median range of spending amounts. They are described as follows:

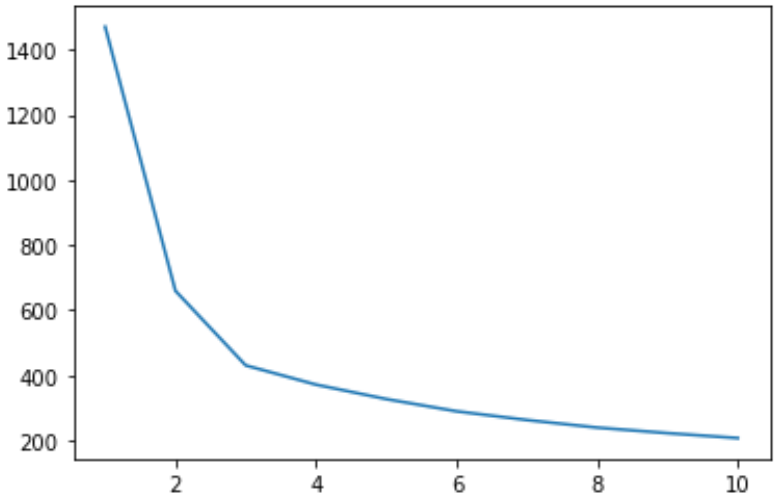



**Plot 7: Pie chart showing proportion of each cluster under agglomerative clustering**

- **Cluster 1-** this cluster spends the most, pays the most and has high probability of full payment as compared to other two clusters. 35.7% of customers fall in this cluster.
- **Cluster 2-** this cluster relatively spends less, pays less and has good probability of full payment but the probabilty is lower as compared to other two clusters. 31% customers fall in this cluster.
- **Cluster 0-** this is the cluster where 33.3% customers fall and stands in the middle of cluster 1 and 2 in terms of spending and payments. It's probabilty of payment to bank is higher than cluster 2 but lower than cluster 1

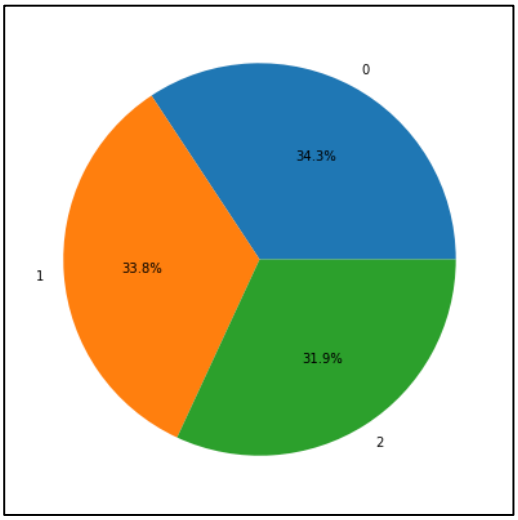
**1.4 Apply K-Means clustering on scaled data and determine optimum clusters. Apply elbow curve and silhouette score. Explain the results properly. Interpret and write inferences on the finalized clusters.**

In K-means clustering, the number of clusters have to predetermined to generate clusters. By making a plot of within sum of squares, number of 'k' clusters can be decided. The point where elbow is formed in the plot can be chosen as number of clusters.




**Plot 8: WSS plot for k-means clustering**

Plot 8 shows that elbow is formed at k=2 and 3 and can be chosen as optimal number of clusters, since wss does not decrease significantly after k=3 (the idea is to get number of clusters where wss decreases the most). For this business case study, k=3 is taken since two clusters don't make business sense and also the silhoutte score, which measures the goodness of clustering technique, comes out to be 0.40, which is positive and informs us that clusters, to a decent extent are apart from each other. If we choose n>3, the silhoutte score decreases, leading to overlapping of clusters.




**Plot 9: Pie chart showing proportion of each cluster using K-means clustering**

| Clus_kmeans | spending | advance_payments | probability_of_full_payment | current_balance |
|---|---|---|---|---|
| 0 | 11.856944 | 13.247778 | 0.848330 | 5.231750 |
| 1 | 14.437887 | 14.337746 | 0.881597 | 5.514577 |
| 2 | 18.495373 | 16.203433 | 0.884210 | 6.175687 |

**Table 5: Cluster groups formed after K-clustering**

Three clustering profiles are made, which approximately covers the max, min and median range of spending amounts. They are described as follows:

- **KCluster 1**- this cluster relatively spends the least, pays less and has good probability of full payment but the probabilty is lower as compared to other two clusters. 32.4% customers fall in this cluster
- **KCluster 2**- this is the cluster where 34.3% of customers fall and stands in the middle of cluster 1 and 2 in terms of spending and payments.
- **KCluster 0**- this cluster spends the most, pays the most and also has high probability of full payment as compared to other two clusters. 33.3% customers fall in this cluster It's probabilty of payment to bank by cans is higher than cluster 2 but lower than cluaster 1

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

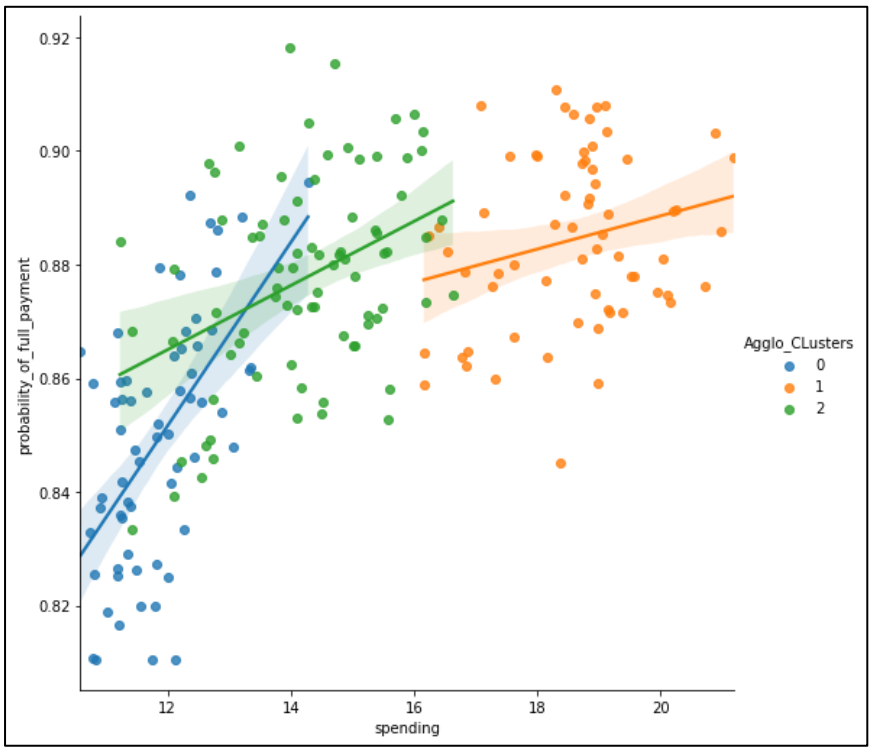
Based on both Hierarchal clustering and KMeans clustering we arrive at three clusters, which can be labelled as:

a) High Spender

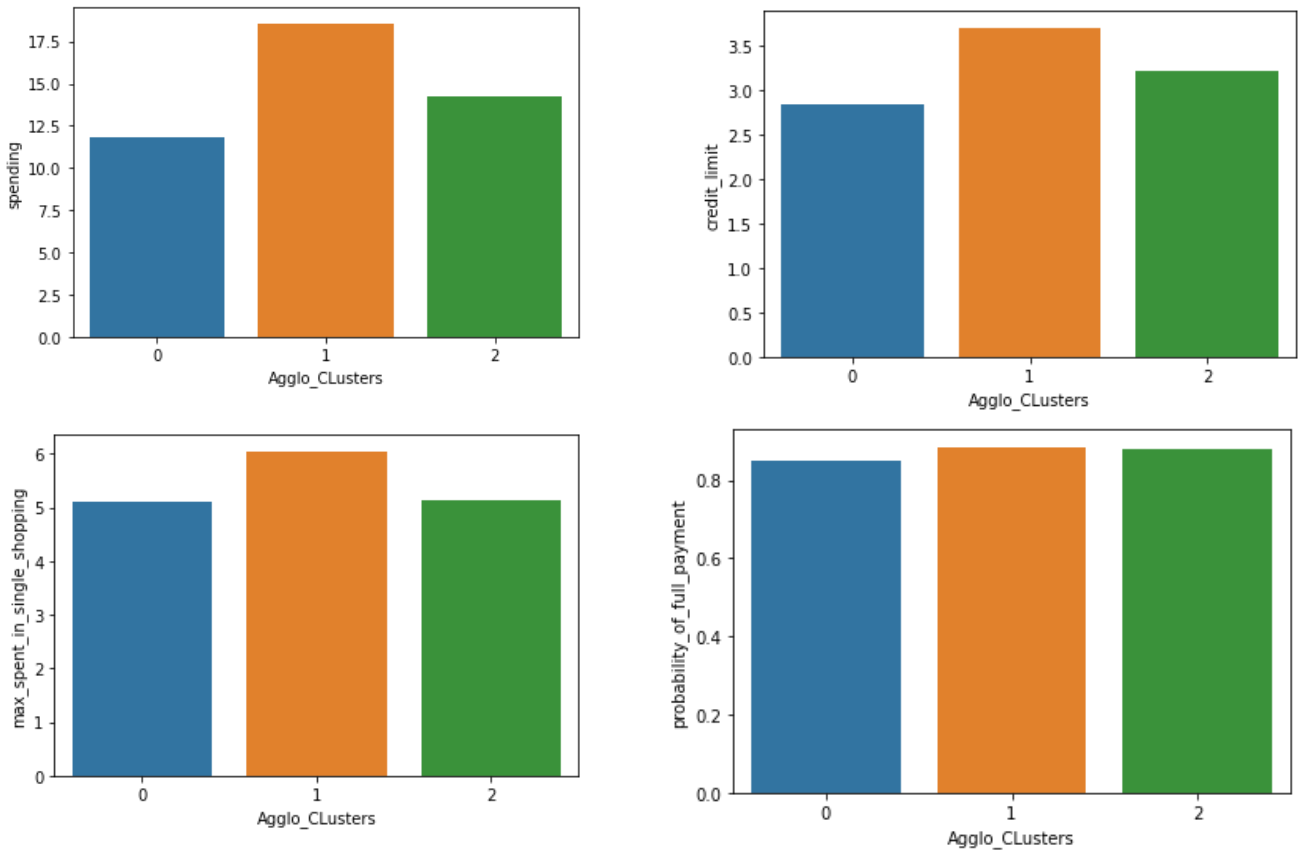
b) Medium Spenders

c) Low Spenders

For profiling theses clusters in detail, hierarchal agglomerative clustering information will be used since it gives better gap and distinction level between three clusters.




**Plot 10: Three Clusters shown in the scatter plot**




**Plot 11: Bar graphs showing level of spending, credit limit, max spending amount and probability of full payment for 3 clusters.**

On the basis the three clusters' formed and their visual analysis on their spending and payment behaviour, following strategies are suggested for each cluster: -

**Cluster 1: The High Spenders**

This is the high spending consumer segment which constitute 31% of the consumers. Since they have good probability of payment and spends a lot, their credit limit can be increased. Their spending range is in the range of affluent class, so discounts could be given in the premium labelled brands. This could be discerned from more detailed data on where they shop, which can help in knowing which brands to tie up with to give special discounts. give them more credit point benefit to increase their spending

**Cluster 2: The Medium Spenders**

This is the medium spending consumer segment which constitute 38.6% of the consumers Their credit limit should also be increased since they have good probability of repayment and spends the second highest and also constitute the largest consumer segment for the bank's credit card Since their spending range is in the middle class range, they are probably spending on everyday expenses. So, they should be given quick small credit benefits to incentivise spending through credit card. having more demographical data might help in discerning the consumer profile, which could further be clustered to generate specific credit based shopping incentive.

**Cluster 3: The Low Spenders**

This is the low spending consumer segment which constitute 30.4% of the consumers. Their probability of repayment is slightly less the other two segments, but mechanisms should be set up to improve their repament probability. Perhaps, incentive of lowering the interest on timely payment can help to achieve timely payments and more spending as well.