**Project: Data Mining**

Pavan Kalyan B | PGP DSBA Oct\_C’21 | March 2022

**Table of contents**

**I CLUSTERING**

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

**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. Explain the results properly. Interpret and write inferences on the finalized clusters.

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

**List of Figures**

Fig 1.1.1 Box plot of data

Fig 1.1.2 Distribution of whole data

Fig 1.1.3 Scatterplot between two variables

Fig 1.1.4 Scatterplot between two variables

Fig 1.1.5 pairplot between all variables

Fig 1.1.6 Heat Map

Fig 1.4.3 scatterplot for kmeans data of 4 clusters

Fig 1.4.4 scatterplot for kmeans data of 4 clusters

**List of Tables**

Table 1.1.1 – First five rows in the data

Table 1.1.2 - Information of all the columns with memory

Table 1.1.3 - Description of the data

Table 1.2.1 - scaled data

Table 1.3.1 – data segregated according to cluster

Table 1.3.2 - Aggregate data with clusters

Table 1.4.1 - cluster values for each number

Table 1.5.1 - cluster values with frequency

**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).

**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. Explain the results properly. Interpret and write inferences on the finalized clusters.

**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](https://olympus.mygreatlearning.com/courses/63575/files/4626497/download?verifier=9Esd3uL7DfdwYx0FGHAvjif4sVYZsqWovubgulyi&wrap=1)

**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 and do exploratory data analysis (3 pts). Describe the data briefly. Interpret the inferences for each (3 pts). Initial steps like head() .info(), Data Types, etc . Null value check. Distribution plots(histogram) or similar plots for the continuous columns. Box plots, Correlation plots. Appropriate plots for categorical variables. Inferences on each plot. Summary stats, Skewness, Outliers proportion should be discussed, and inferences from above used plots should be there. There is no restriction on how the learner wishes to implement this but the code should be able to represent the correct output and inferences should be logical and correct.**

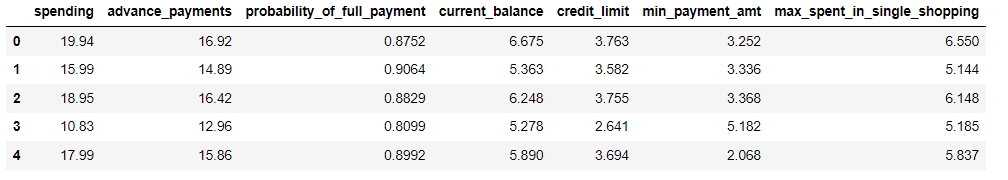


Table 1.1.1 - First five rows in the data

* Data set has 7 variables with with different aspects of credit usage. Through this information we will analyze the data.

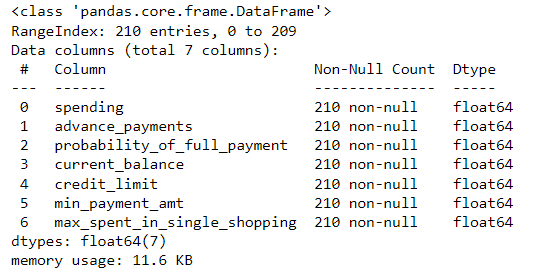


Table 1.1.2 - Information of all the columns with memory

* We can observe that all the columns in the bank data are continuous and numeric and there are no null values in the data.

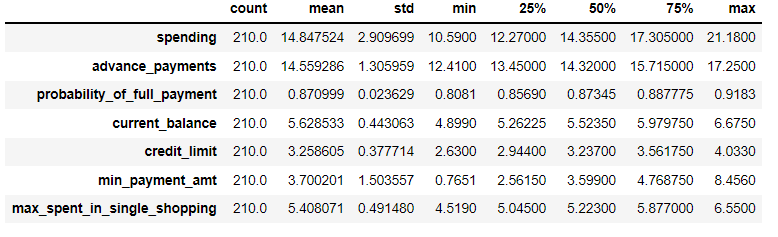


Table - 1.1.3 Description of the data

* We can observe that data has not been scaled.
* Spending amount ranges between 10.5900 to 21.1800.
* Advance\_payments has the minimum value of 12.4100 and highest of 17.2500.
* Probability\_of\_full\_payment ranges between 0.8081 to 0.9183.
* Credit\_limit is between 2.6300 to 4.0330.

**Univariate Analysis:**

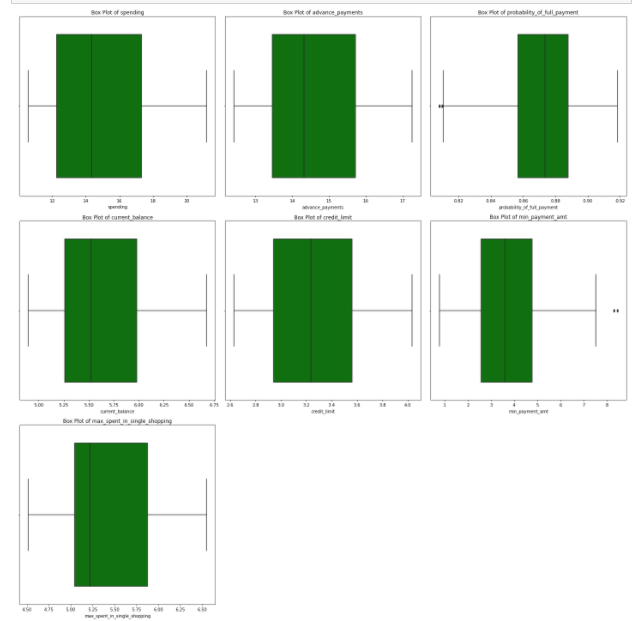


Fig 1.1.1 Box plot of data

* We can observe that there are outliers in **’min\_payment\_amt’** and **‘probability\_of\_full\_payment’.**

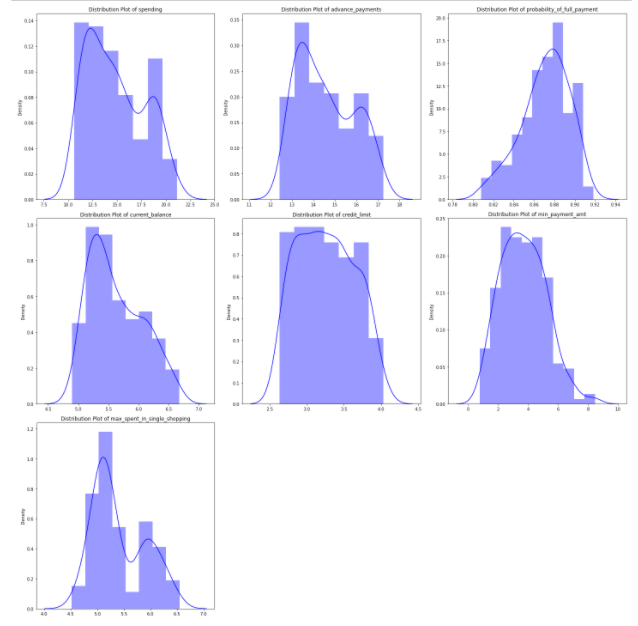


Fig 1.1.2 Distribution of whole data

* We can observe that ‘min\_payment\_amt’ is symmetric and normally distributed.
* ‘spending’, ‘advance\_payments’, ‘max\_spent\_in\_single\_shopping’, ‘current\_balance’ are right skewed.
* ‘probability\_of\_full\_payment’ is left skewed.

**Bivariate Analysis:**

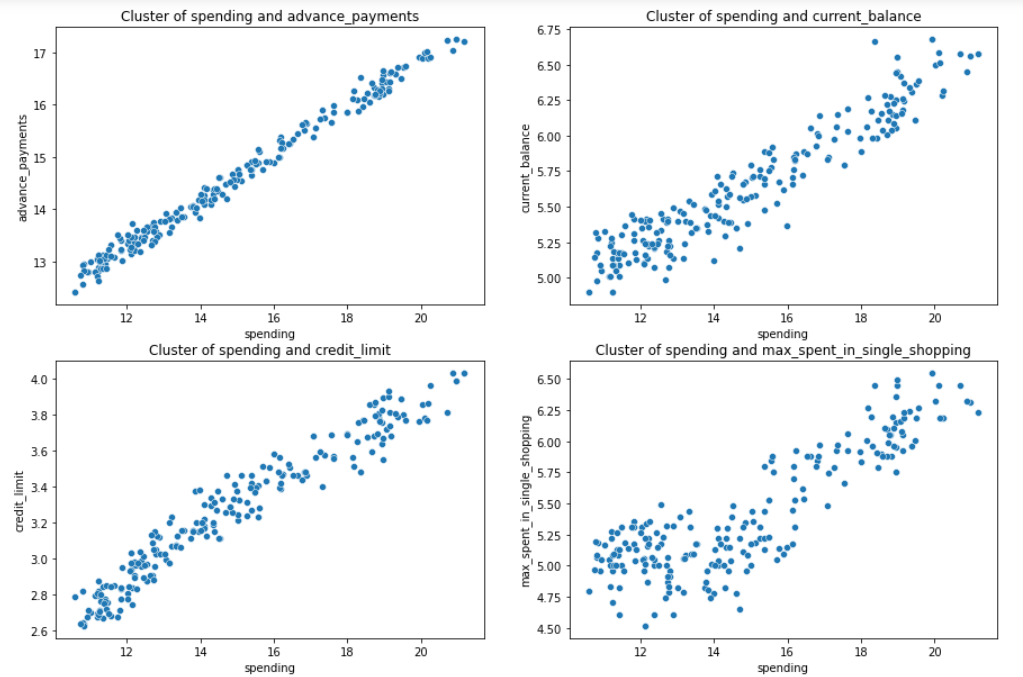


Fig 1.1.3 Scatterplot between two variables

* We can observe there is a **strong positive correlation** between spending and credit limit, spending and advance payments.
* There is **moderately positive correlation** between spending and current balance, spending and max spent in single shopping.

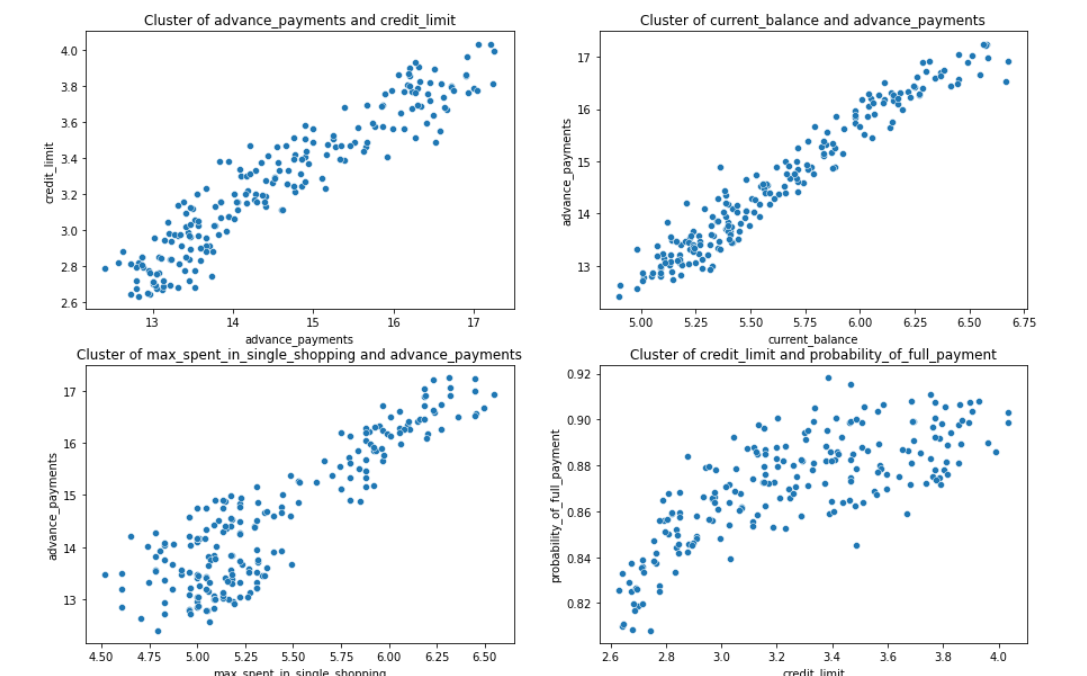


Fig 1.1.4 Scatterplot between two variables

* We can observe there is a **strong positive correlation** between credit limit and advance payments.
* There is **moderately positive correlation** between current balance and advance payments , advance payments and max spent in single shopping.

**Multivariate Analysis:**

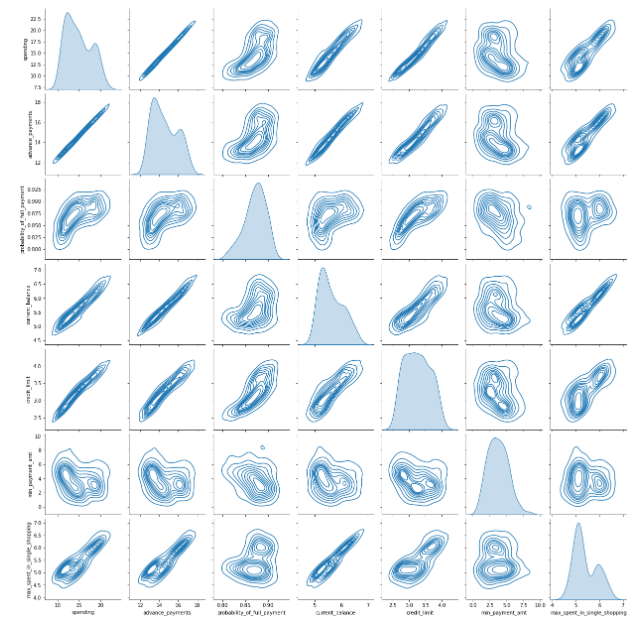


Fig 1.1.5 pairplot between all variables

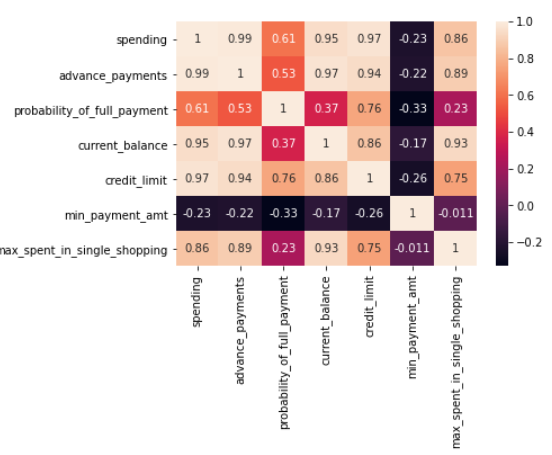


Fig 1.1.6 Heat Map

* We can observe that spending advance payments, current balance, credit limit, are highly correlated.
  1. **Do you think scaling is necessary for clustering in this case? Justify The learner is expected to check and comment about the difference in scale of different features on the bases of appropriate measure for example std dev, variance, etc. Should justify whether there is a necessity for scaling and which method is he/she using to do the scaling. Can also comment on how that method works.**

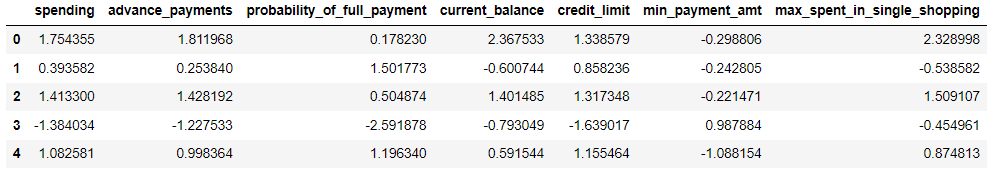
****

Table 1.2.1 - scaled data

* As the data is not aligned properly, we scaled the data.

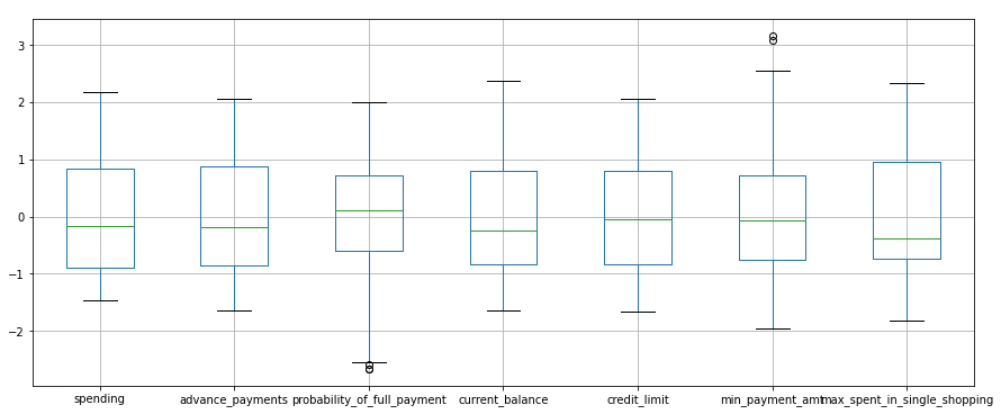


Table 1.2.2 - Box plots of scaled data

* Probability of full payment and min payment has outliers in the data.
* Other columns doesn’t have any outliers.
  1. **Apply hierarchical clustering to scaled data (3 pts). Identify the number of optimum clusters using Dendrogram and briefly describe them (4). Students are expected to apply hierarchical clustering. It can be obtained via Fclusters or Agglomerative Clustering. Report should talk about the used criterion, affinity and linkage. Report must contain a Dendrogram and a logical reason behind choosing the optimum number of clusters and Inferences on the dendrogram. Customer segmentation can be visualized using limited features or whole data but it should be clear, correct and logical. Use appropriate plots to visualize the clusters.**

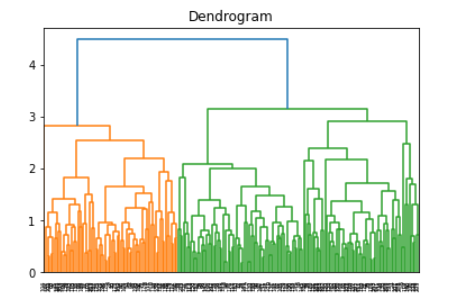


Fig 1.3.1 Dendrogram

* We choosed a dendrogram with average method and Euclidean distance
* As we have more amount of data, we will segregate the data into three cluster

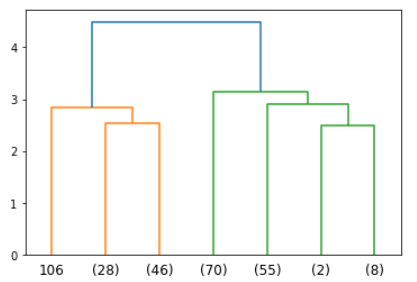


Fig 1.3.2 - Truncated dendrogram

* We truncated the dendrogram into 7 last formations merged.

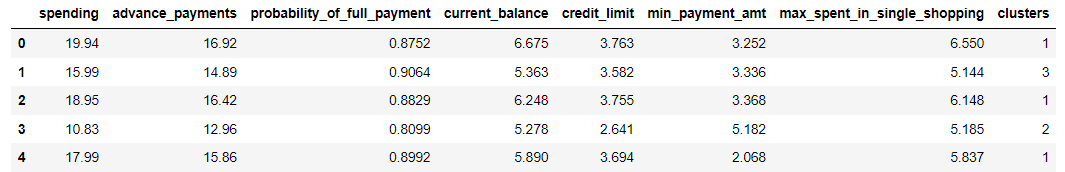


Table 1.3.1 - data segregated according to cluster

* The data has been segregated in to three clusters using the centroid method.
* Using this clusters we will segregate the data into 3 ways.

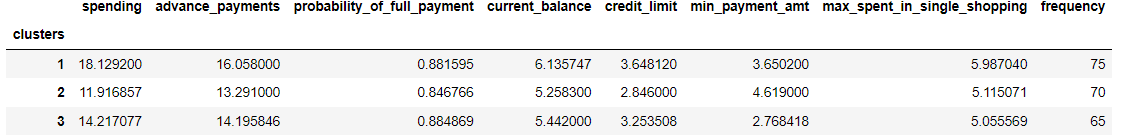


Table 1.3.2 - Aggregate data with clusters

* 1 cluster has a frequency of 75 and which is the highest weightage.
* 2 cluster has a frequency of 70 weightage.
* 3 cluster has a frequency of 65 weightage which is low
  1. **Apply K-Means clustering on scaled data and determine optimum clusters (2 pts). Apply elbow curve and silhouette score (3 pts). Interpret the inferences from the model (2.5 pts). K-means clustering code application with different number of clusters. Calculation of WSS(inertia for each value of k) Elbow Method must be applied and visualized with different values of K. Reasoning behind the selection of the optimal value of K must be explained properly. Silhouette Score must be calculated for the same values of K taken above and commented on. Report must contain logical and correct explanations for choosing the optimum clusters using both elbow method and silhouette scores. Append cluster labels obtained from K-means clustering into the original data frame. Customer Segmentation can be visualized using appropriate graphs.**

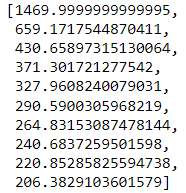


Table 1.4.1 - cluster values for each number

* Each value in the cluster is taken and highly dropped value is taken as the kmeans cluster.

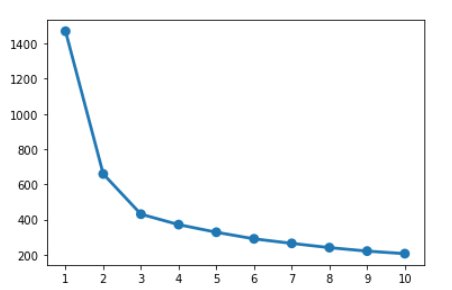


Fig 1.4.1 wss point plot

* We can observe that after three there is less drop in values, so we take 4 as kmeans cluster and start the model.

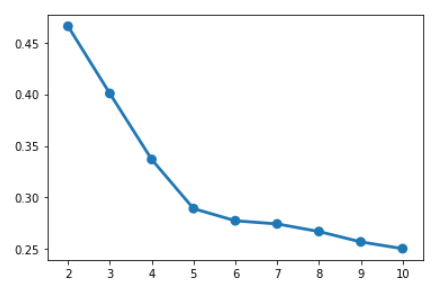


Fig 1.4.2 - silhouette score point plot

* Here we can observe that 5 is the perfect value while looking into the graph.

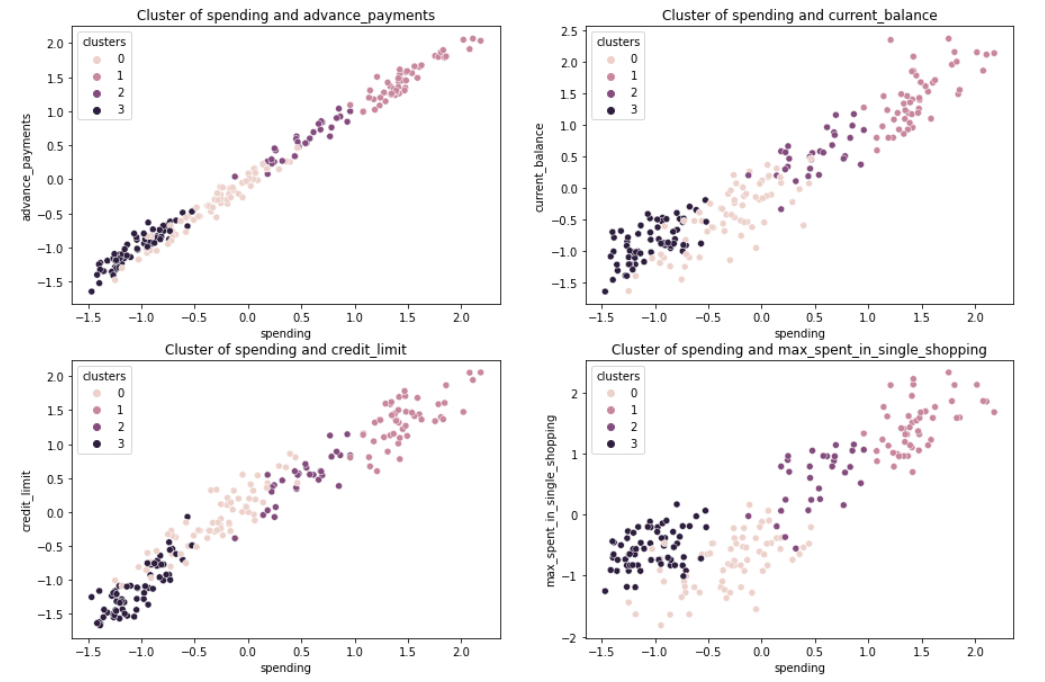


Fig 1.4.3 – scatterplot for kmeans data of 4 clusters

* Cluster of spending and advance payments, and spending and credit limit are strongly positively correlated.
* Other two plots are moderately positive in correlation.

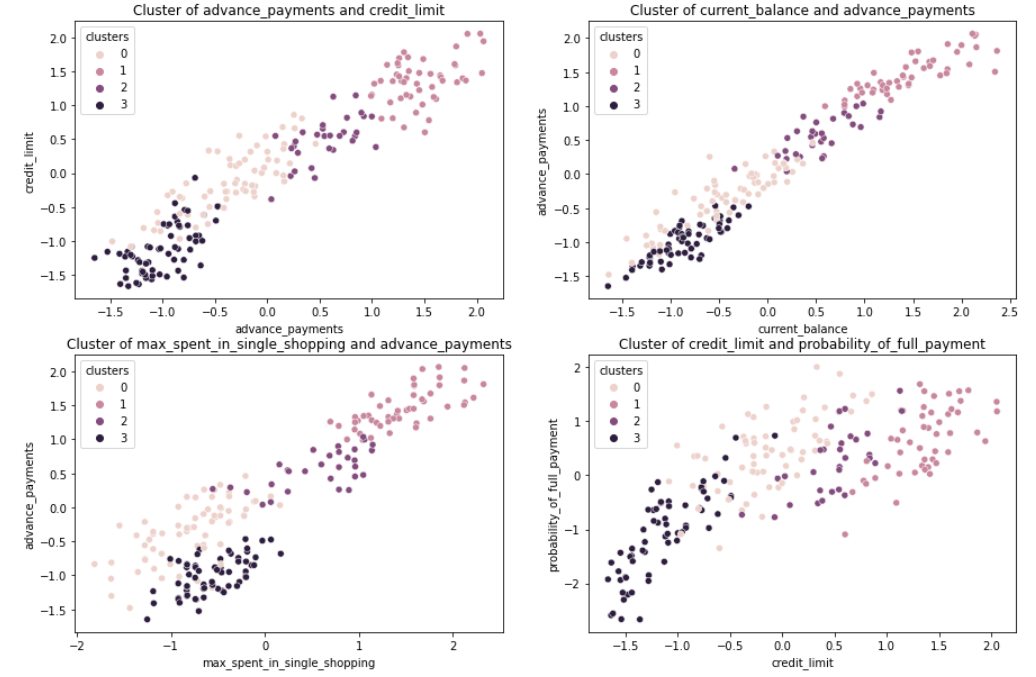


Fig 1.4.4 – scatterplot for kmeans data of 4 clusters

* Cluster of current balance and advance payments are strongly positively correlated.
* Other three plots are moderately positive correlated.
  1. **Describe cluster profiles for the clusters defined (2.5 pts). Recommend different promotional strategies for different clusters in context to the business problem in-hand (2.5 pts ). After adding the final clusters to the original dataframe, do the cluster profiling. Divide the data in the finalyzed groups and check their means. Explain each of the group briefly. There should be at least 3-4 Recommendations. Recommendations should be easily understandable and business specific, students should not give any technical suggestions. Full marks will only be allotted if the recommendations are correct and business specific. variable means. Students to explain the profiles and suggest a mechanism to approach each cluster. Any logical explanation is acceptable.**

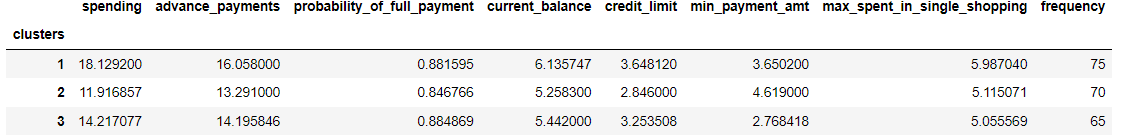
****

Table 1.5.1 - cluster values with frequency

When we look into final cluster merged with original data set and by taking average values,

These are the recommendations.

Cluster 1 – Platinum customers

Cluster 2 – Gold customers

Cluster 3 – Silver customers

Cluster 1 has high spending and we can give them high limit where as gold customers are medium privileged members. They use credit cards but not frequently. Silver customers are less usage of credit cards.

**--------------------------------------------------------------------------------------------------------------------------**