Customer Segmentation

Introduction

Marketing strategies are only effective if targeted on right consumers, Therefore Bank wants to group their costumers so that they can build offers and promotions based on types of customers they have. They want to offer products to customers which will be liked by them and will be willing to buy those offers.

So, we are going to make a clustering method which can group the costumers based on their credit card transaction, on base of that Banks can build their strategies.

Project Flow

Business understanding : to identify the business goals and to determine how to measure success.
Data understanding: to select relevant data and to understand this data. This means to understand the semantics of tables and columns and to know the data distributions.
Data preparation: to cleanse the selected data and to transform it, for example, by joining and by aggregation so that it is suitable for data mining analysis.
Modelling: To run the data mining algorithms.
Evaluation: To look at mining models, understand influencing factors and assess model accuracy.
Deployment: To score, this means to apply the data mining model to new data.

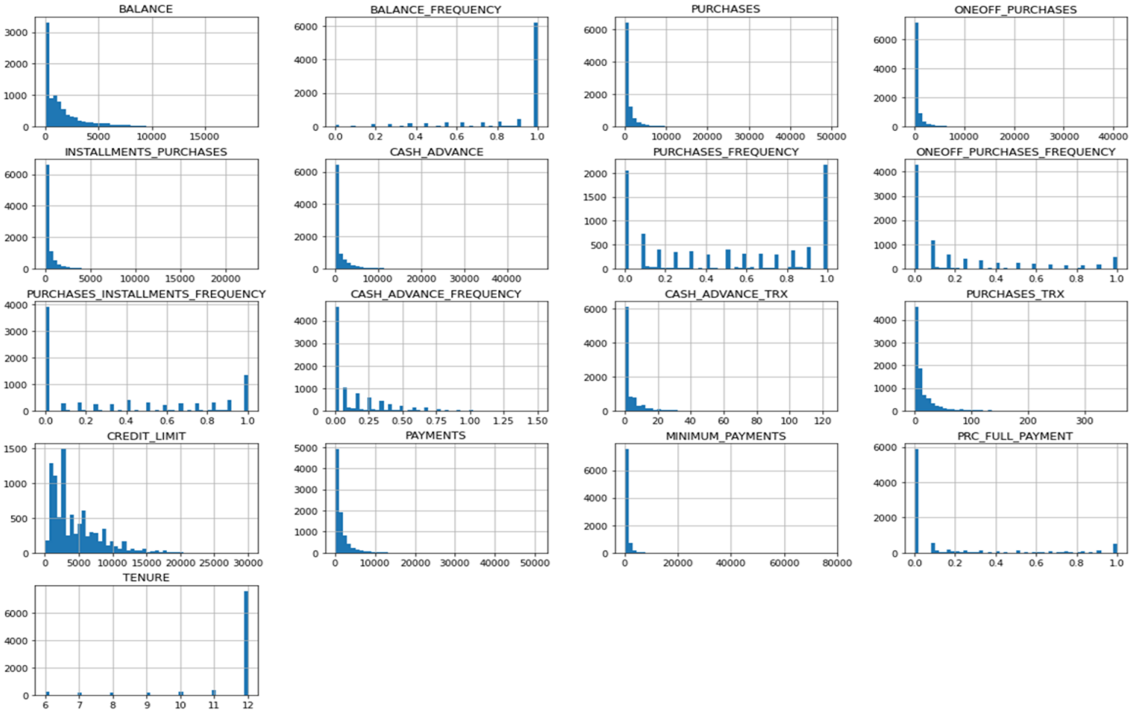
Data

The sample Dataset summarizes the use behaviour of about 9000 active credit card holders of 6 months duration. The file at a customer level with 18 behavioural variables.

Following is the Data Dictionary for Credit Card dataset:

- Cust_ID: - Identification of Credit Card holder (Categorical)
- Balance: - Balance amount left in their account to make purchases
- Balance Frequency: - How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
- Purchase: - Amount of purchases made from account
- One-off purchase: - Maximum purchase amount done in one-go
- Instalment Purchase: - Amount of purchase done in instalment
- Cash advance: - Cash in advance given by the user
- Purchase Frequency: - How frequently the Purchases are being made, score between 0 and 1
- One-off purchase frequency: - How frequently Purchases are happening in one-go
- Purchase Instalment frequency: - How frequently purchases in instalments are being done
- Cash advance frequency: - How frequently the cash in advance being paid
- Cash advanced TRX: - Number of Transactions made with "Cash in Advanced"
- Purchase TRX: - Number of purchase transactions made
- Credit limit: - Limit of Credit Card for user
- Payments: - Amount of Payment done by user
- Minimum payments: - Minimum amount of payments made by user
- PRCFullpayment: - Percent of full payment paid by user
- Tenure: -Tenure of credit card service for user

Data understanding and cleaning



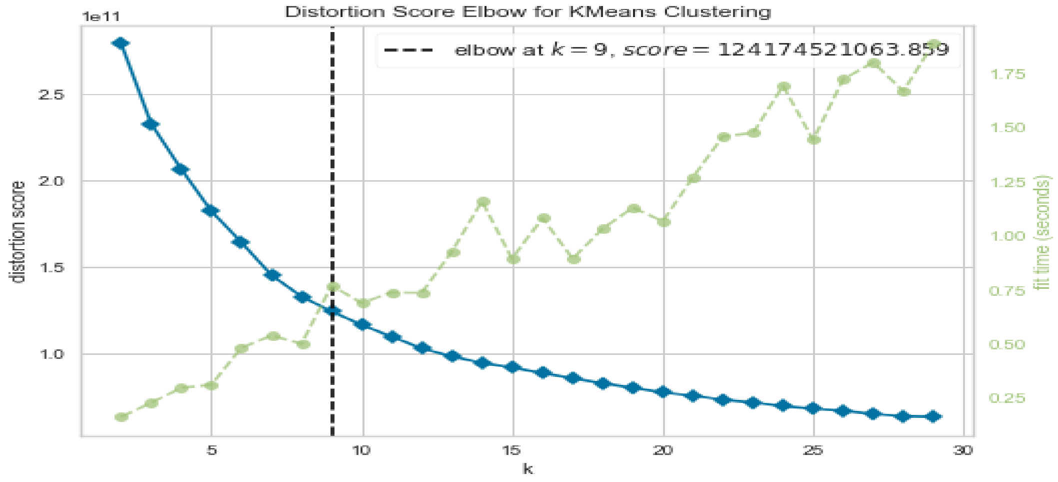
As we can see many histograms are tail heavy, they extend much farther to the right of the median than to the left. This may make it a bit harder for some Machine Learning algorithms to detect patterns.
These attributes have very different scales. So we will apply features scaling.
One of the most important transformations you need to apply to your data is feature scaling. With few exceptions, Machine Learning algorithms don't perform well when the input numerical attributes have very different scales
we also notice the presence of outlier for some variables for Balance, Purchase, one-off Purchase, Install Purchase, Cash Advanced, Credit Limit, Payment, Min Payment.
As you see percentages and outliers' numbers there are too many. If we drop outliers, we lose many rows. Because of that the sensible way is to make ranges to deal with extreme values. In this way, we will reduce the effects on our dataset without deleting outliers.
Also there are some null values which will be handled by replacing them with mean value of column.

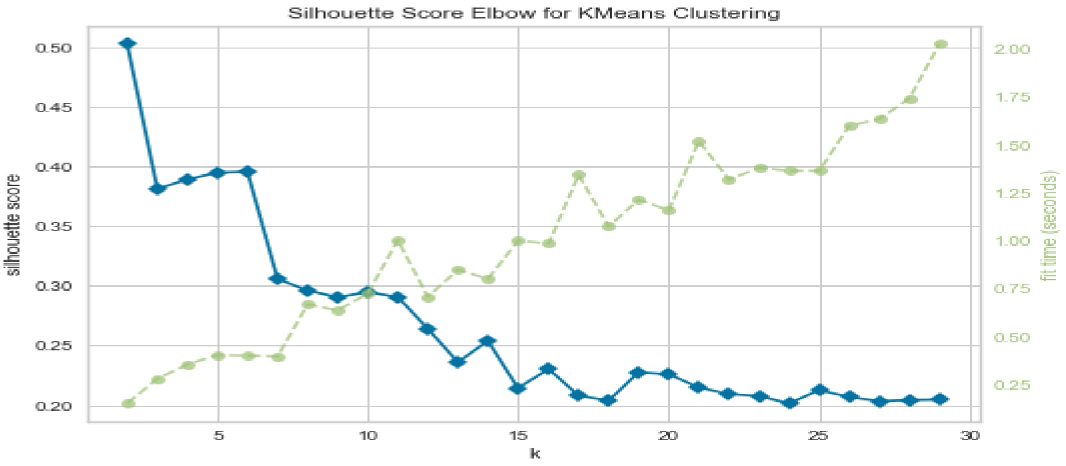
Experiments

To build a clustering method we are going to perform 2 different types of clustering methods: Kmeans and Agglomerative clustering, which we are going to compare based on number of clusters and silhouette scores and choose the one which best fits and use the cluster build by model to interpret the results.

What are Clusters ?

Clustering is basically defined as division of data into groups of similar objects. Each group called a cluster consists of objects that are similar between themselves and dissimilar compared of other groups.





As we can see from the graph, that there are not many breaking points after 9. distortion score- which computes the sum of squared distances from each point to its assigned centre.
From comparison of silhouette scores, we are going to get the optimal number of clusters n=2.

K-cluster

K means Clustering : clusters : 2 silhouette_score : 0.511761078211312

K means Clustering : clusters : 3 silhouette_score : 0.46711830473128096

K means Clustering : clusters : 9 silhouette_score : 0.35394159637130235

Agglomerative Cluster

Agglomerative Clustering : clusters : 2 linkage : complete silhouette_score : 0.868801771895

Agglomerative Clustering : clusters : 3 linkage : complete silhouette_score : 0.854623904544

Agglomerative Clustering : clusters : 9 linkage : complete silhouette_score : 0.708633390380

Agglomerative Clustering : clusters : 2 linkage : average silhouette_score : 0.9090802426640

Agglomerative Clustering : clusters : 3 linkage : average silhouette_score : 0.8855258191064

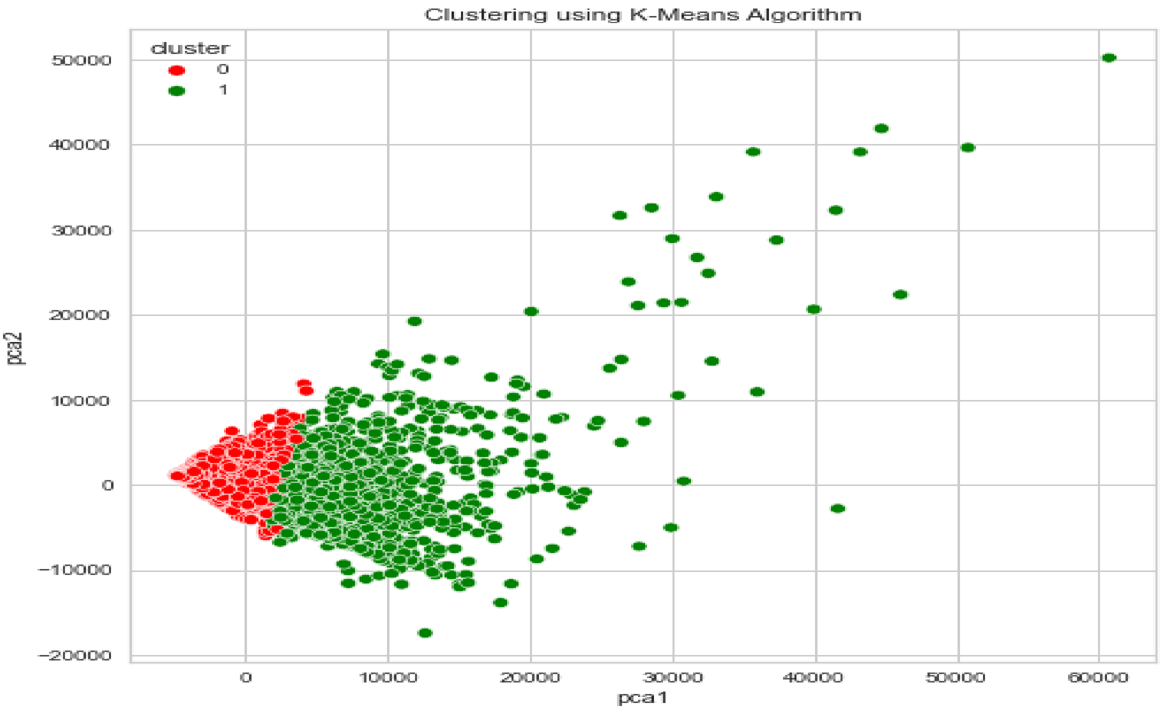
Agglomerative Clustering : clusters : 9 linkage : average silhouette_score : 0.7877547560651

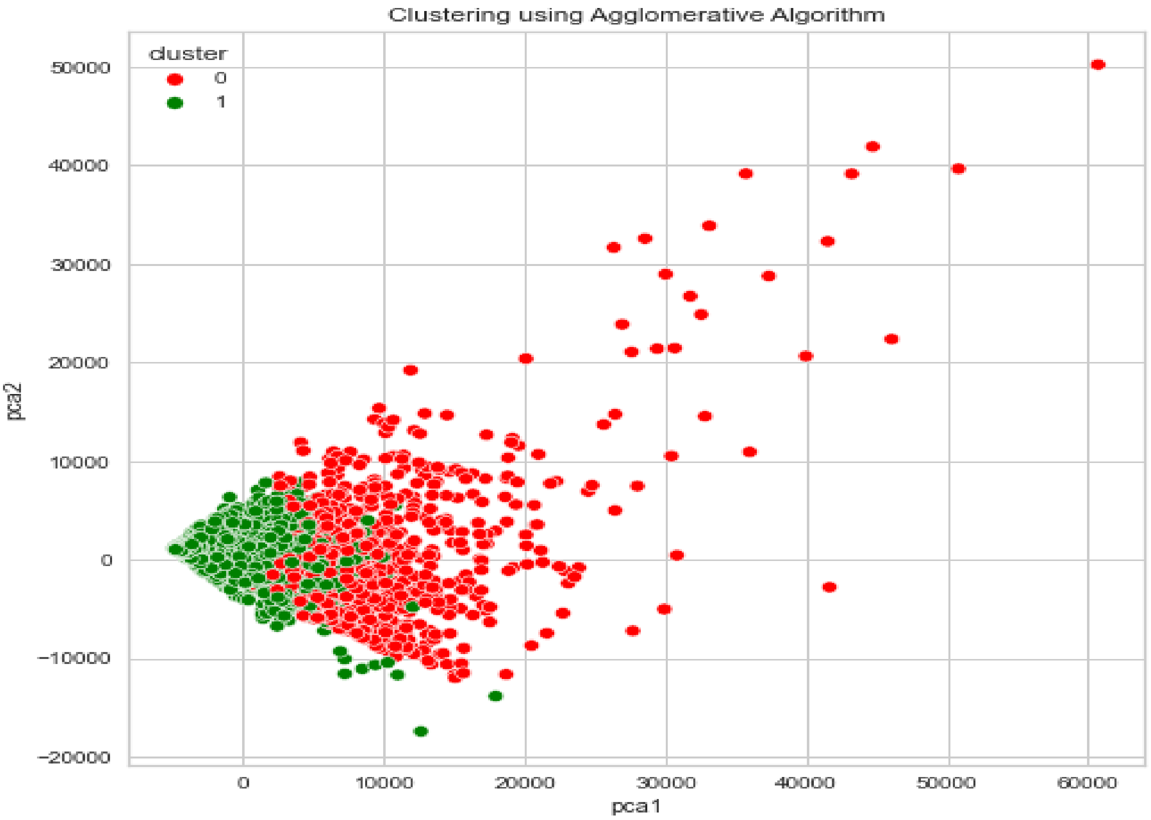
Agglomerative Clustering : clusters : 2 linkage : single silhouette_score : 0.87379145880434

Agglomerative Clustering : clusters : 3 linkage : single silhouette_score : 0.87328686073753

Agglomerative Clustering : clusters : 9 linkage : single silhouette_score : 0.81540345627150

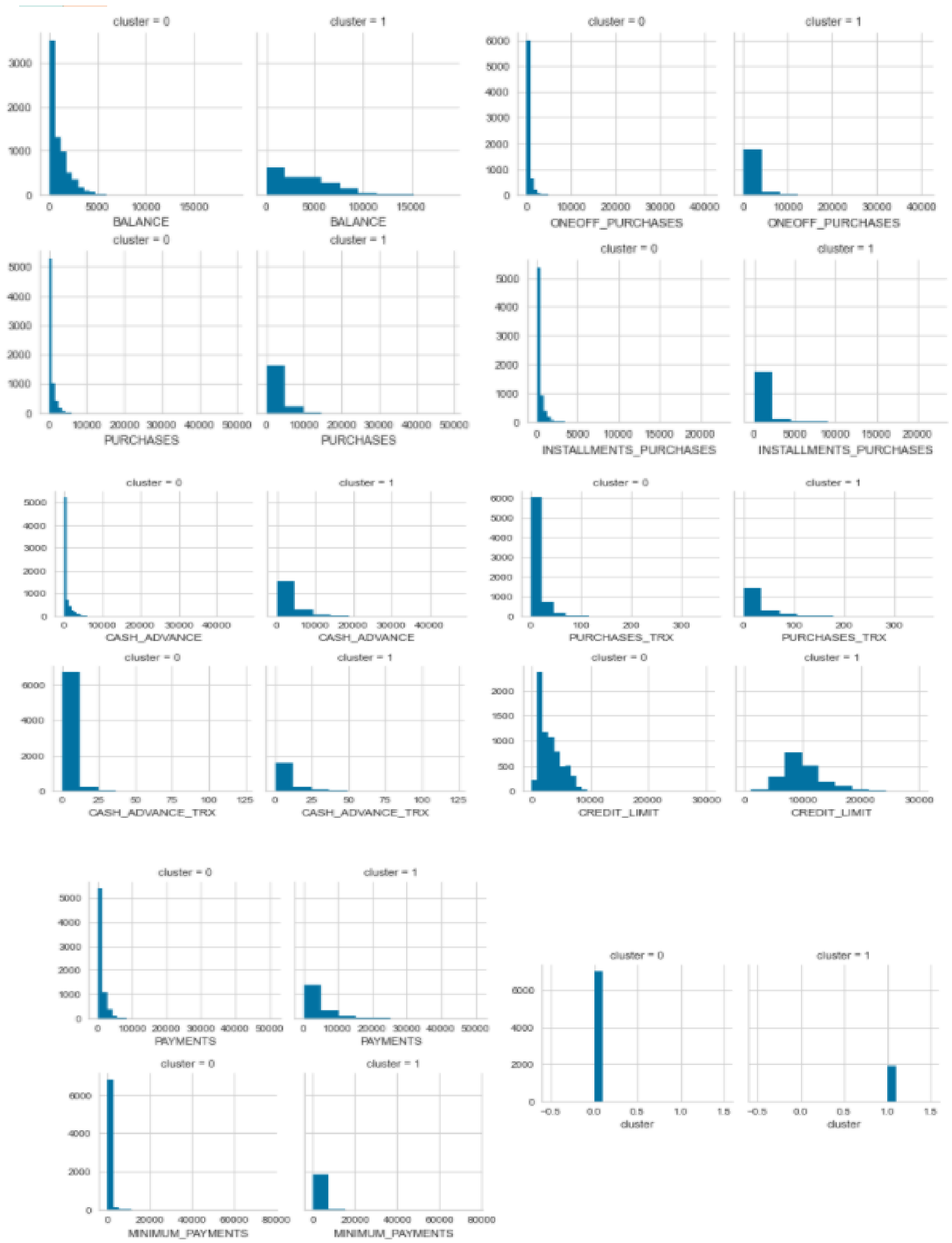
Analysing the scatter plot with n=2





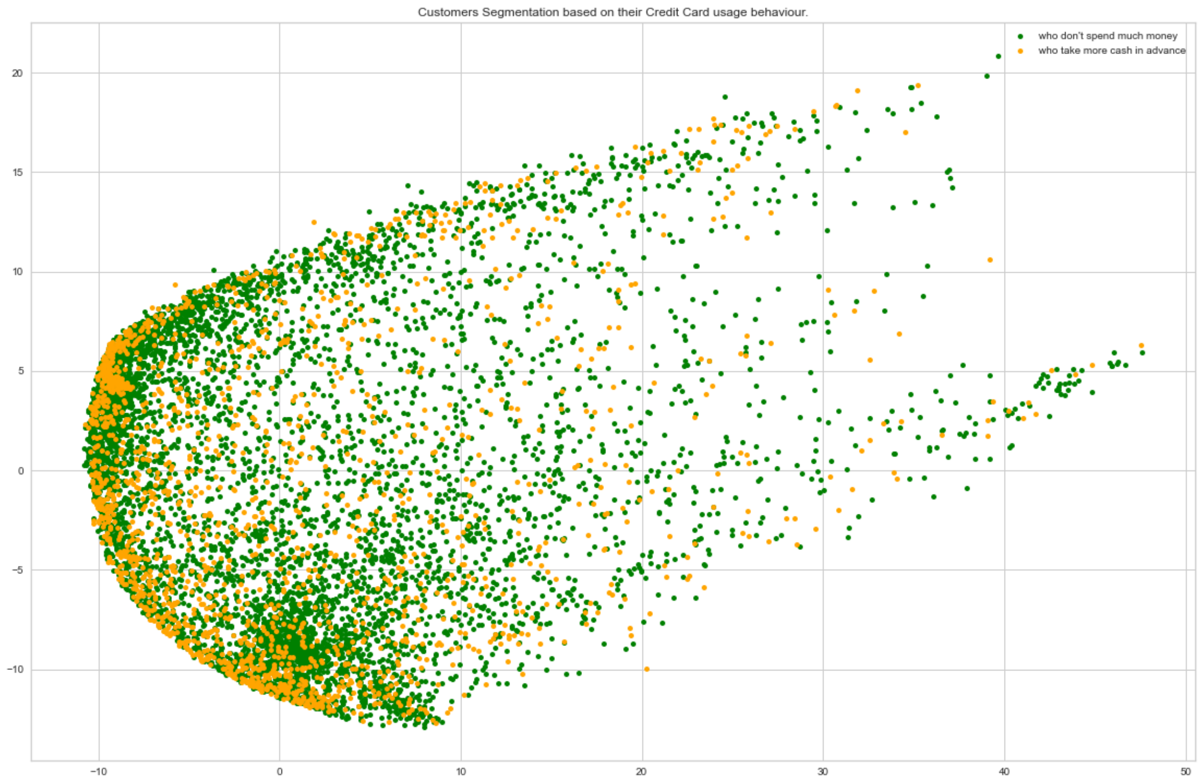
Here we can see the distribution of data points using k-means clustering as compared to Agglomerative Clustering. K means and HC give almost the same results in terms of number of clusters k=2. But the K means is more accurate (more precise scatter plot), and also simpler in implementation. In the end we can say that the K Means is the best method for the segmentation of this Dataset.

Interpretation of Clusters formed by K-means



These above Graph represent the distribution of datapoint of both clusters with respect to each column.

Visualization Of Cluster



Conclusion

- **Group 1**: lower in all fronts (Economical family), most of them have average credit limit with low to medium balance, less spending in the form of purchase and payments transaction and most of them do very less cash advanced transaction.
- **Group 2**: High spenders (Higher middle-class family), high credit limit who do more cash advance transaction, most of them have high balance for purchase who do often one-off purchase transaction.
- Different action can be taken according to the clusters we have separated. For example, bank should encourage the group 1 customer, who do payments in instalment by increasing their instalment options or offer 2* points for purchase transaction which they can redeem later.
- For high spenders, bank may increase their loyalty with special offer.