Project title –
Market Segmentation of bank customer using Unsupervised Learning

By-
Lakshya Tekwani

Abstract

- The sample Dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months.
- Segmenting Bank Customers to launch a targeted ad and marketing campaign by dividing the customers into distinct groups according to their demand and provide them services accordingly.
- Using K-mean Clustering algorithm (Unsupervised Learning Algorithm)
- After cleaning , evaluating and analyzing data we can segment the customers based on their behavior in the banking system and help the bank to provide customer specific solutions.

Motivation

- I choose this particular task because everybody has a different banking need , some need high credit limit, some needs high cash advances, some prefers installment transactions on the other hand some users prefer one-off purchases therefore to serve them what they need , banks needs to segment their customers according to their needs and serve them better .
- It will help bank to increase the frequency of transactions and also the capital that customers are depositing in the bank or they are lending from the bank and overall increasing the profits for the bank.

Dataset

- Data Source: https://www.kaggle.com/arjunbhasin2013/ccdata
- The sample Dataset summarizes the usage behavior of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioral variables.
- This data is a csv file containing 8950 rows indicating 8950 customers and 18 columns indicating different transactional behaviour.

DATA DIRECTORY

Following is the Data Dictionary for Credit Card dataset :-

1. CUSTID : Identification of Credit Card holder (Categorical)
2. BALANCE : Balance amount left in their account to make purchases (
3. BALANCEFREQUENCY : How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated)
4. PURCHASES : Amount of purchases made from account
5. ONEOFFPURCHASES : Maximum purchase amount done in one-go
6. INSTALLMENTSPURCHASES : Amount of purchase done in installment
7. CASHADVANCE : Cash in advance given by the user
8. PURCHASESFREQUENCY : How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased)
9. ONEOFFPURCHASESFREQUENCY : How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased)
10. PURCHASESINSTALLMENTSFREQUENCY : How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done)
11. CASHADVANCEFREQUENCY : How frequently the cash in advance being paid
12. CASHADVANCETRX : Number of Transactions made with "Cash in Advanced"
13. PURCHASESTRX : Numbe of purchase transactions made
14. CREDITLIMIT : Limit of Credit Card for user
15. PAYMENTS : Amount of Payment done by user
16. MINIMUM_PAYMENTS : Minimum amount of payments made by user
17. PRCFULLPAYMENT : Percent of full payment paid by user
18. TENURE : Tenure of credit card service for user of Credit Card holder (Categorical)

Data Preparation and Cleaning

- The data set is almost clear and has no Nan-values ,therefore it wasn't much of a task to clean the data set.
- But there were a few null values which were replaced by the average of the column values.
- Sum of null values->

```
CUST_ID                             0
BALANCE                             0
BALANCE_FREQUENCY                   0
PURCHASES                           0
ONEOFF_PURCHASES                    0
INSTALLMENTS_PURCHASES              0
CASH_ADVANCE                        0
PURCHASES_FREQUENCY                 0
ONEOFF_PURCHASES_FREQUENCY          0
PURCHASES_INSTALLMENTS_FREQUENCY    0
CASH_ADVANCE_FREQUENCY              0
CASH_ADVANCE_TRX                    0
PURCHASES_TRX                       0
CREDIT_LIMIT                        1
PAYMENTS                            0
MINIMUM_PAYMENTS                  313
PRC_FULL_PAYMENT                    0
TENURE                              0
dtype: int64
```

Research Question

- The main purpose of this research is to increase the revenue generating activity of the bank by providing what customer needs the most. By studying the behavior of customer we are predicting what all services do a customer prefer the most and then we will try to maximize it in order to increase the frequency of transactions .
- We will divide customers into segments based upon their common traits and it will help bank to provide most beneficial offers and services according to customer’s need.

Methods

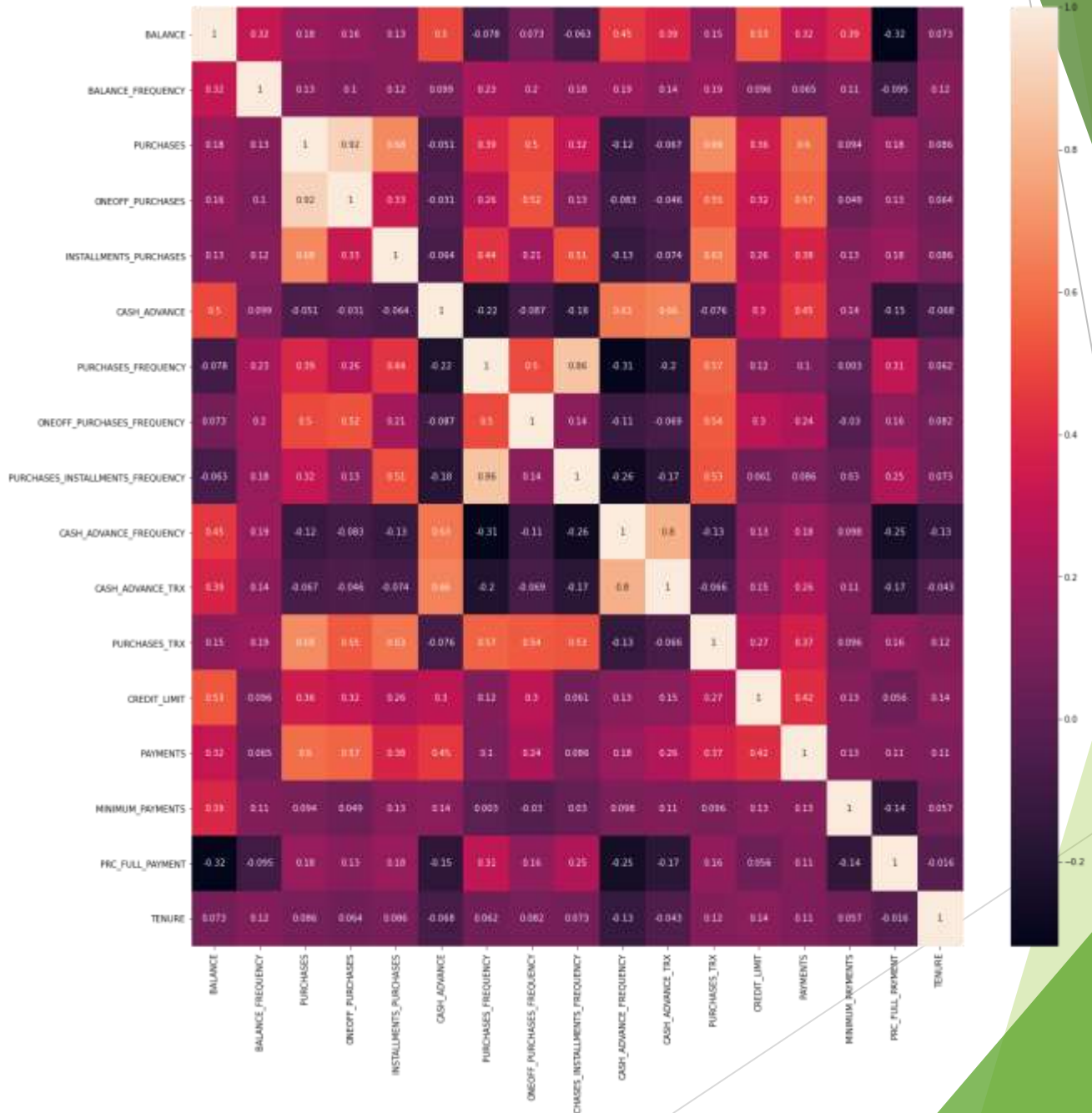
- In my project I used Unsupervised learning to cluster the customers using the K-mean clustering algorithms.
- **Kmeans** algorithm is an iterative algorithm that tries to partition the dataset into *K* pre-defined distinct non-overlapping clusters where each data point belongs to **only one group**. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.

Methods

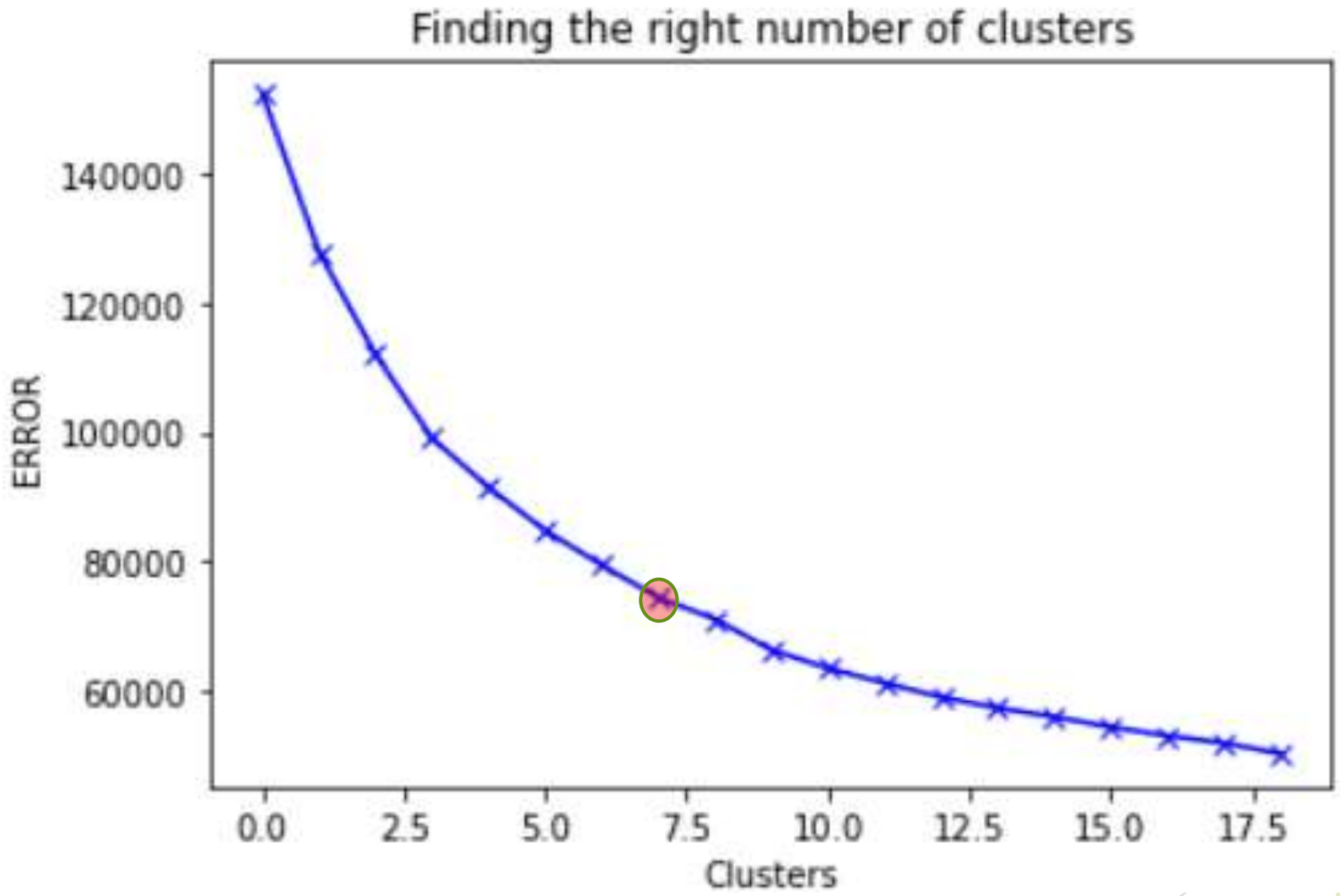
1. I analyzed all the features available in the data .
2. I cleaned the data so that none of its cell is empty .
3. I scaled the data to make it fit for the K-Mean Algortihm.
4. I computed the error for k=1 to k=20.
5. I analyzed the variation of error and selected the most suitable value of K is 7
6. I made a K-mean model with K=7 .
7. I analyzed the groups and their needs to make a clear insight for the banks

Findings

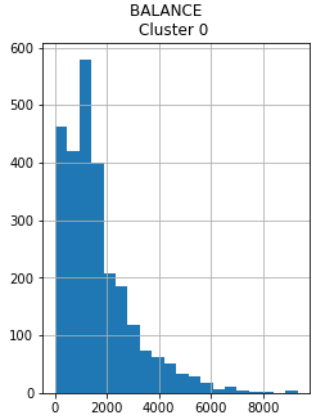
CORRELATION heatmap

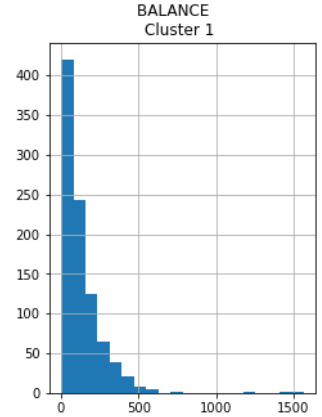


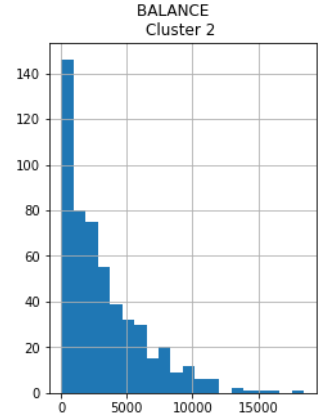
SELECTING THE KNEE POINT

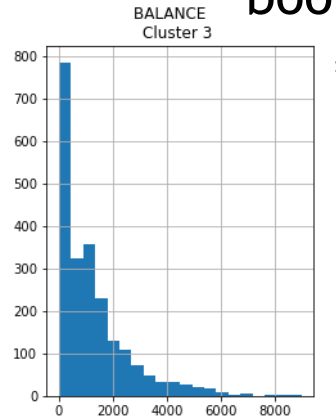


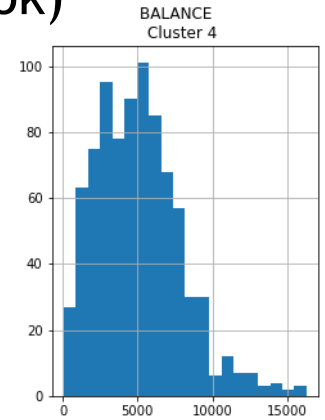
CLUSTER FEATURE COMPARISION-(rest in Jupiter note book)

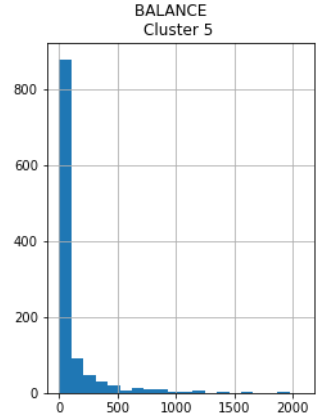


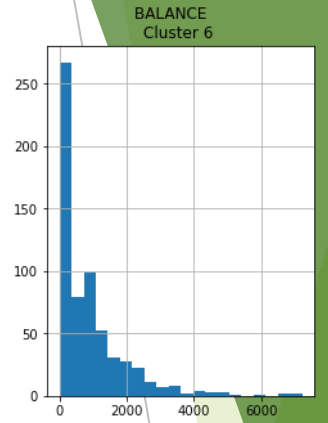


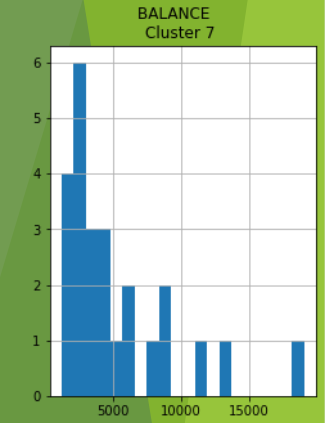


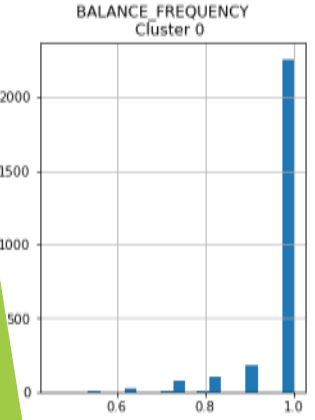


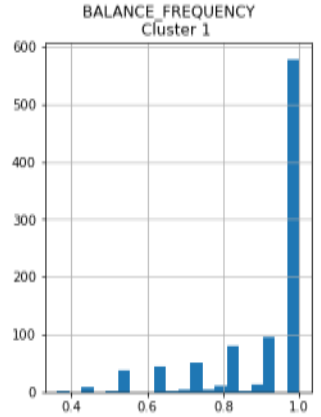


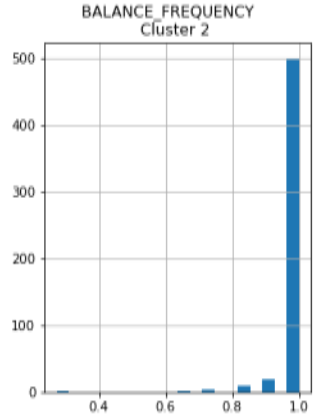


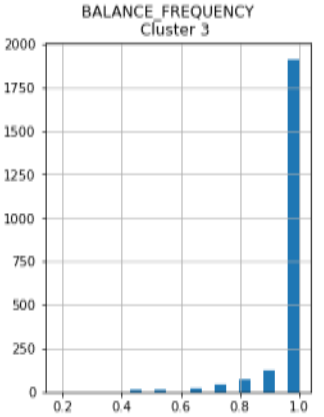


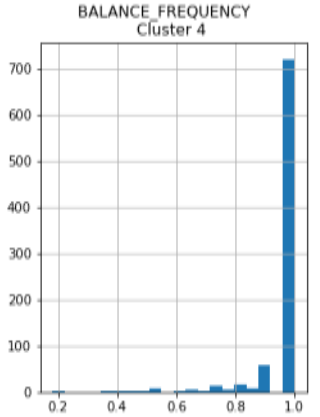


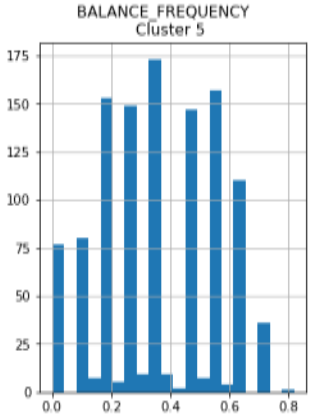


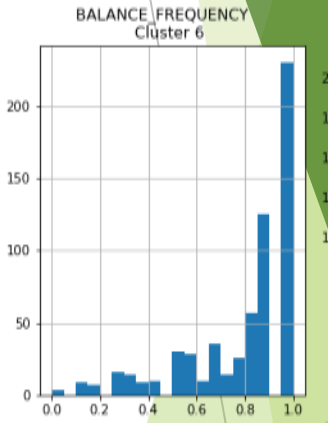


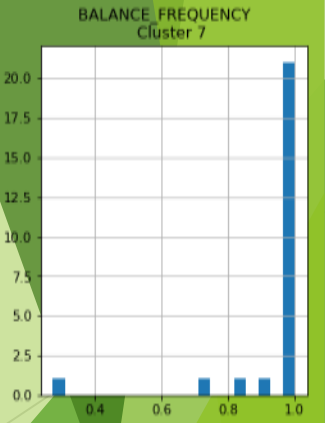


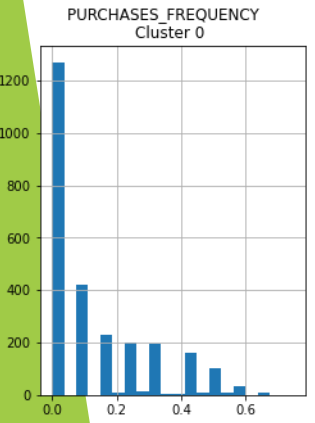


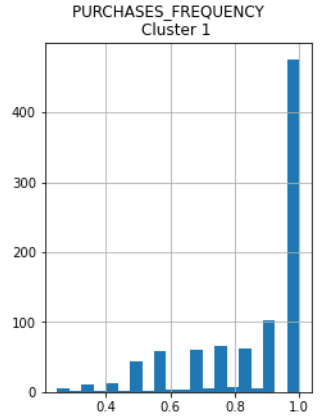


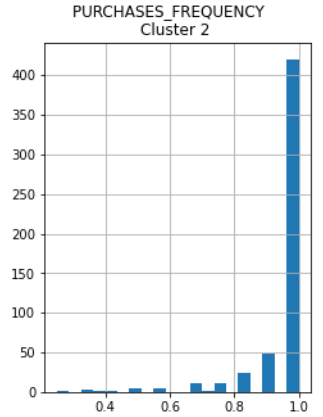


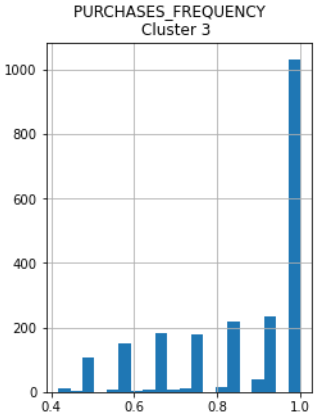


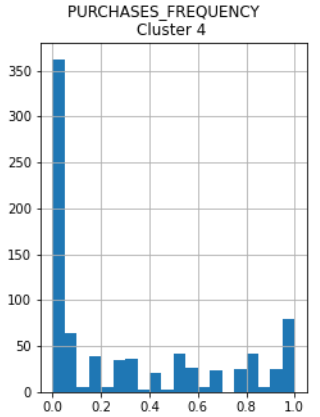


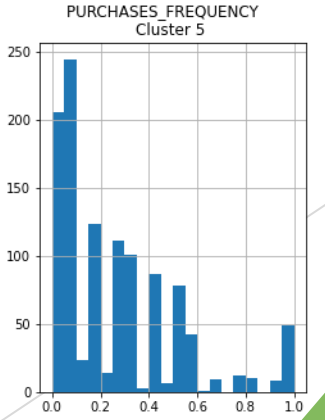


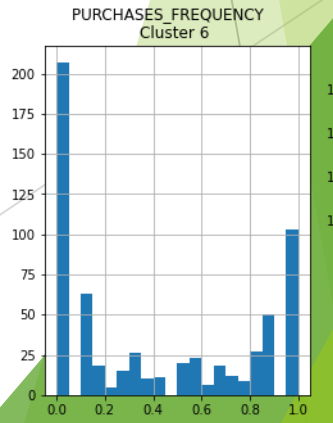


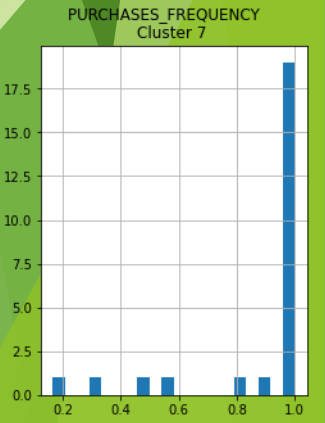












Limitations

- Banking-User preferences may change with time , so we need to update the data frequently in order to get customers engagement.
- Dimensionality reduction and Principal Component analysis is not done therefore the trained model may not be sufficient enough.
- Maybe the K value I chose wasn’t good enough but it was the knee point.

Conclusion – cluster centres

BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY	CASH_ADVANCE_FREQUENCY	CASH_ADVANCE_TRX	PURCHASES_TRX	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	PRC_FULL_PAYMENT	TENURE
0	0.027822	0.402697	-0.363266	-0.242628	-0.413061	-0.073046	-0.896141	-0.406575	-0.793813	0.141919	-0.003886	-0.494519	-0.293057	-0.245100	-0.003782	-0.453919
1	-0.689187	0.146308	0.150583	0.017057	0.324242	-0.448821	0.923043	0.384551	0.739958	-0.642453	-0.458227	0.249203	0.143074	-0.101981	-0.291042	2.172631
2	0.804265	0.474697	2.138432	1.756938	1.828474	-0.193371	1.158942	1.730366	1.099548	-0.296641	-0.210594	2.557005	1.191073	1.191044	0.358630	0.225787
3	-0.152500	0.399820	0.046345	-0.041449	0.185645	-0.323471	0.923471	0.262092	0.824741	-0.392080	-0.314430	0.250734	-0.167551	-0.154508	0.008713	-0.334421
4	1.694087	0.393012	-0.209727	-0.143219	-0.232836	2.061699	-0.456252	-0.167493	-0.418092	1.945963	1.976011	-0.255929	1.040024	0.850283	0.546782	-0.389971
5	-0.699450	-2.190441	-0.322267	-0.243102	-0.315239	-0.315320	-0.575005	-0.458367	-0.447304	-0.513097	-0.370101	-0.426230	-0.194923	-0.208223	-0.254012	0.118685
6	-0.330968	-0.343592	-0.282512	-0.198956	-0.302098	0.071633	-0.195183	-0.266590	-0.235749	0.312320	0.002112	-0.385861	-0.562938	-0.390576	-0.205772	-0.035587
7	1.819469	0.301643	12.207777	12.296446	6.272587	0.242957	1.002411	2.032081	0.866148	-0.392172	-0.124384	4.369530	3.261192	8.782991	1.003073	1.152351

Insights And Actions

- **First Customer Cluster** -> the most economical customer , doesn't spends to much and doesn’t tends to pay money in full. So not much to offer him
- **Second Customer Cluster** -> the smart spender , maintains low balance and pays pretty well therefore using the credit card most effectively , so we need to offer more credit limit
- **Third Customer Cluster** -> this customer with highest purchase frequency and does pretty good purchases in form of one-off purchases and installments as well and pays the money pretty well. so we can provide him with transactional offers or points so he can use more and more of our services.
- **Fourth Customer Cluster** -> the average customer , one who is most concerned to buy things on installments. so we can provide him with installments offers
- **Fifth Customer Cluster** -> the one who doesn't spends much , and keeps his balance full and uses cash advances the most and pays well. we can offer him some discounts related to cash_advances in order to increase his transactions
- **Sixth Customer Cluster** -> one of the defaulter customer , he doesn't maintains his balance and he is offered very low credit limit . we should encourage him to pay installments in time and maybe remove him from our base as he may not be able to pay his debt on time.
- **Seventh Customer Cluster** -> another economical customer who uses most of his money in form of cash_advances and doesn't repay well.we should encourage him to use installments and give him offers related to that.

Acknowledgements

I would like to express my special thanks of gratitude to our instructor **Ilkay Altintas** and **Leo Porter** who helped me a lot in course of this project, and also helped me in doing a lot of Research and to know about so many new techniques. I would also like to thanks **Arjun Bhasin** for his publicly available dataset. I would also like to thank my friends and family to gather all the possible conclusions for my project.

References

- Majority of the work done in this project is my own.
- I have received help from the Jupiter notebooks and resources available in course.

Link to notebook-
https://drive.google.com/file/d/1VgS7FiwDt8VTzWR8l2JcH7kw7Fd1jEEf/view?usp=sharing

Link to pdf-

https://drive.google.com/file/d/1lr0kn22nVoWzc6DQpW090x67VS22nFwW/view?usp=sharing