Credit Card Segmentation

Bharat Mohan Thakur 15-1-2020

Contents

Chapter 1: Introduction

1.1 Problem Statement

The objective of this project is to predict the count of bike rentals day wise.. Project will help accommodate in managing the number of bikes required on a daily basis, and being prepared for high demand of bikes during peak periods.

1.2 Data

The objective is to develop a customer segmentation to define marketing strategy. The sample dataset summarizes the usage behaviour of about 9000 active credit card holders during the last 6 months. The file is at a customer level with 18 behavioural variables. Table Description of data:

Number of attributes:

- CUST_ID Credit card holder ID
- BALANCE Monthly average balance (based on daily balance averages)
- BALANCE_FREQUENCY Ratio of last 12 months with balance

● PURCHASES Total purchase amount spent during last 12 months

● ONEOFF_PURCHASES Total amount of one-off purchases

● INSTALLMENTS_PURCHASES Total amount of installment purchases

● CASH_ADVANCE Total cash-advance amount

● PURCHASES_ FREQUENCY-Frequency of purchases (percentage of months with at least on purchase)

● ONEOFF_PURCHASES_FREQUENCY Frequency of one-off-purchases

● PURCHASES_INSTALLMENTS_FREQUENCY Frequency of installment purchases

● CASH_ADVANCE_ FREQUENCY Cash-Advance frequency

● AVERAGE_PURCHASE_TRX Average amount per purchase transaction

● CASH_ADVANCE_TRX Average amount per cash-advance transaction

● PURCHASES_TRX Average amount per purchase transaction

● CREDIT_LIMIT Credit limit

● PAYMENTS-Total payments (due amount paid by the customer to decrease their statement balance) in the period

● MINIMUM_PAYMENTS Total minimum payments due in the period.

● PRC_FULL_PAYMENT- Percentage of months with full payment of the due statement balance

● TENURE Number of months as a customer

	CUST_ID	BALANCE	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY
1	C10001	40.900749	0.818182	95.40	0.00	95.40	0.00000	0.166667	0.000000
2	C10002	3202.467416	0.909091	0.00	0.00	0.00	6442.94548	0.000000	0.000000
3	C10003	2495.148862	1.000000	773.17	773.17	0.00	0.00000	1.000000	1.000000
4	C10004	1666.670542	0.636364	1499.00	1499.00	0.00	205.78802	0.083333	0.083333
5	C10005	817.714335	1.000000	16.00	16.00	0.00	0.00000	0.083333	0.083333
6	C10006	1809.828751	1.000000	1333.28	0.00	1333.28	0.00000	0.666667	0.000000
7	C10007	627.260806	1.000000	7091.01	6402.63	688.38	0.00000	1.000000	1.000000
8	C10008	1823.652743	1.000000	436.20	0.00	436.20	0.00000	1.000000	0.000000
9	C10009	1014.926473	1.000000	861.49	661.49	200.00	0.00000	0.333333	0.083333
10	C10010	152.225975	0.545455	1281.60	1281.60	0.00	0.00000	0.166667	0.166667

PURCHASES_INSTALLMENTS_FREQUENCY	CASH_ADVANCE_FREQUENCY	CASH_ADVANCE_TRX	PURCHASES_TRX	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	PRC_FULL_PAYMENT	TENURE
0.083333	0.000000	0	2	1000	201.80208	139.50979	0.000000	12
0.000000	0.250000	4	0	7000	4103.03260	1072.34022	0.222222	12
0.000000	0.000000	0	12	7500	622.06674	627.28479	0.000000	12
0.000000	0.083333	1	1	7500	0.00000	312.34395	0.000000	12
0.000000	0.000000	0	1	1200	678.33476	244.79124	0.000000	12
0.583333	0.000000	0	8	1800	1400.05777	2407.24604	0.000000	12
1.000000	0.000000	0	64	13500	6354.31433	198.06589	1.000000	12
1.000000	0.000000	0	12	2300	679.06508	532.03399	0.000000	12
0.250000	0.000000	0	5	7000	688.27857	311.96341	0.000000	12
0.000000	0.000000	0	3	11000	1164.77059	100.30226	0.000000	12
1.000000	0.000000	0	12	1200	1083.30101	2172.69776	0.000000	12
0.000000	0.000000	0	6	2000	705.61863	155.54907	0.000000	12

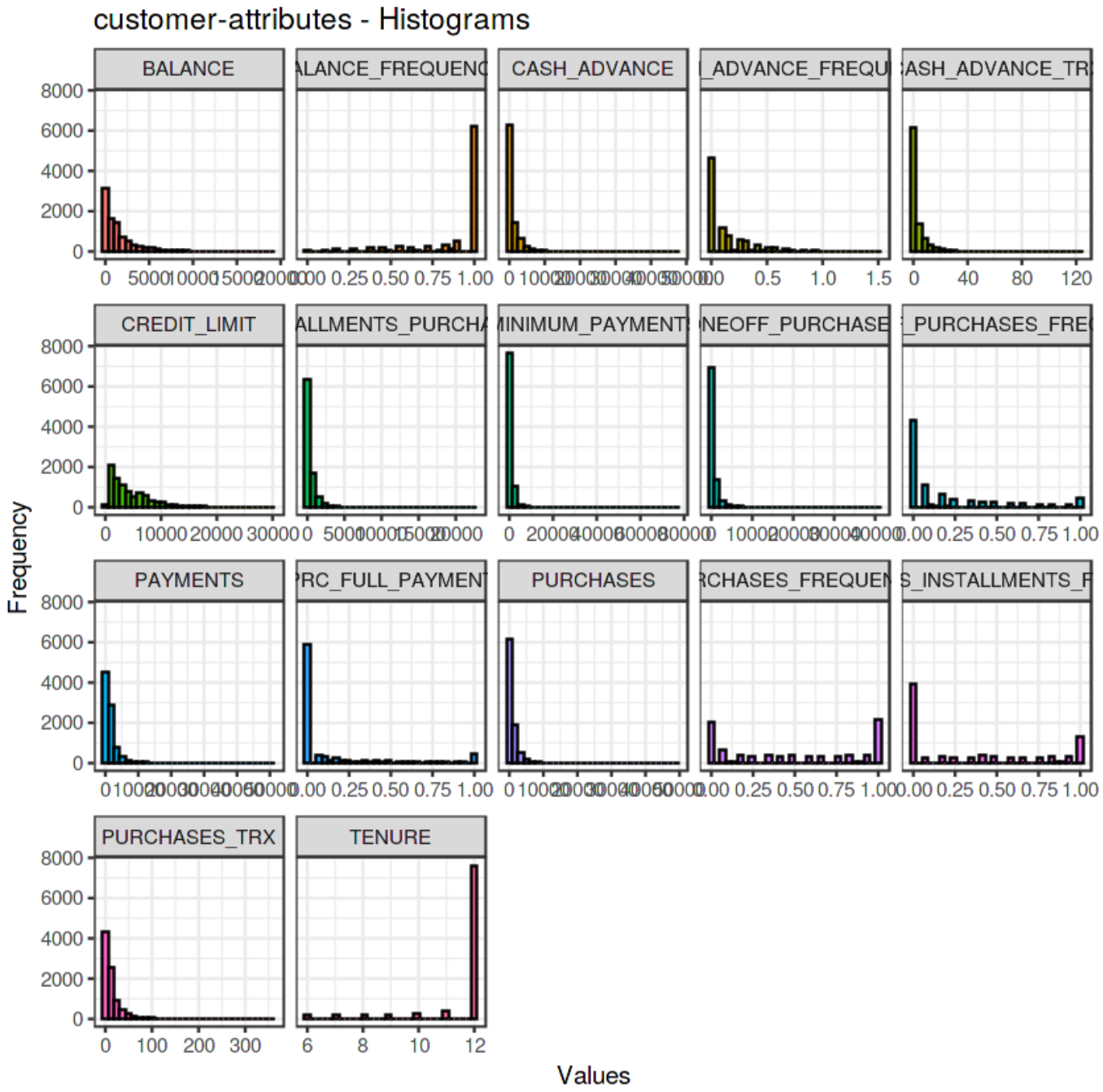
Chapter 2: Exploratory Data Analysis

2.1 Pre-Processing

We should look at the data before we start to create a model. In data mining, looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is known as Exploratory Data Analysis.

2.2 histogram of all variables

It can be observed from the below plot that some of the variables contains outlier as they are skewed


customer-attributes - Histograms
BALANCE
CASH_ADVANCE
CREDIT_LIMIT
PAYMENTS
PURCHASES
PURCHASES_TRX
TENURE
Frequency
Values

Creating Density Plot:

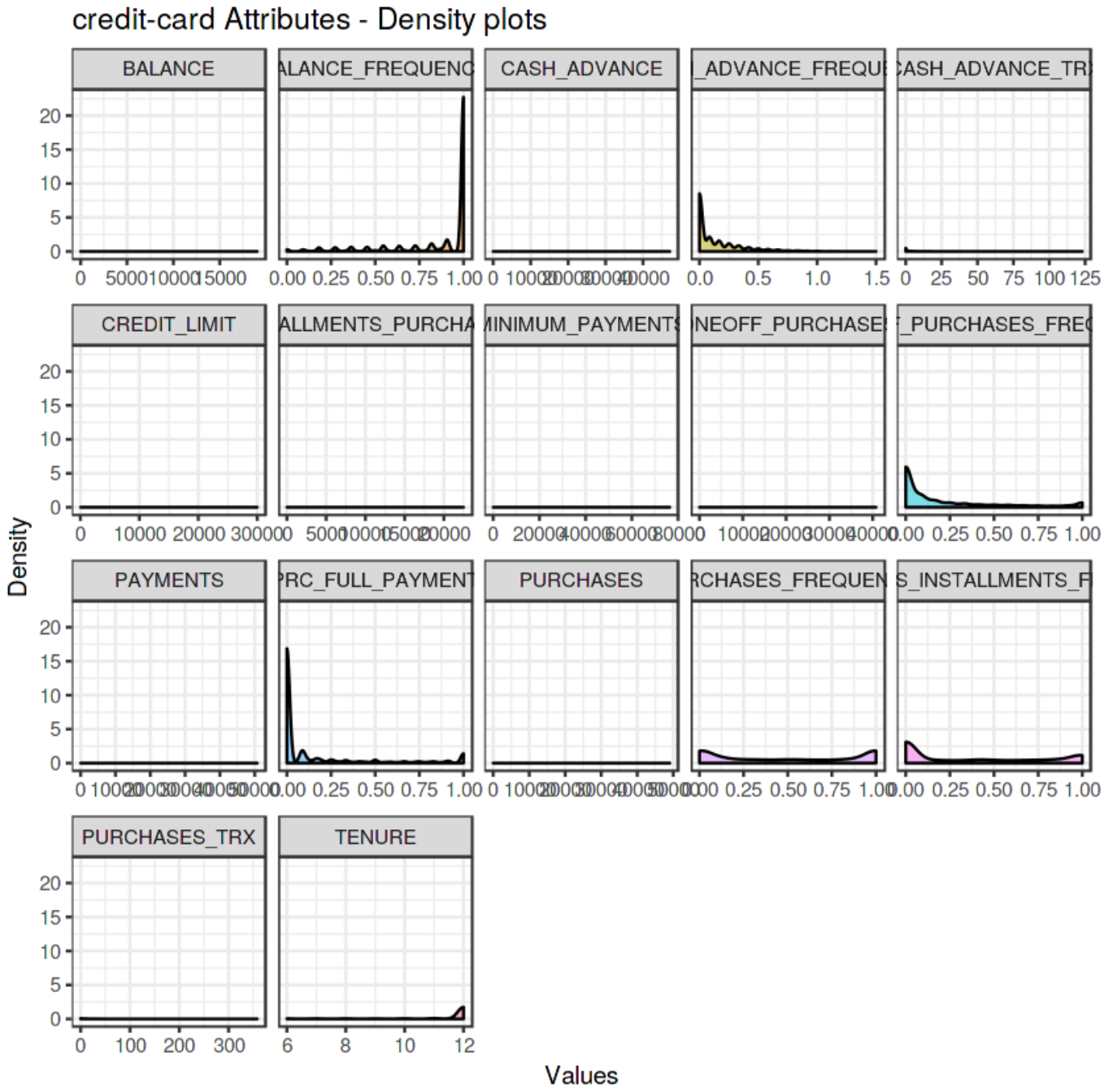
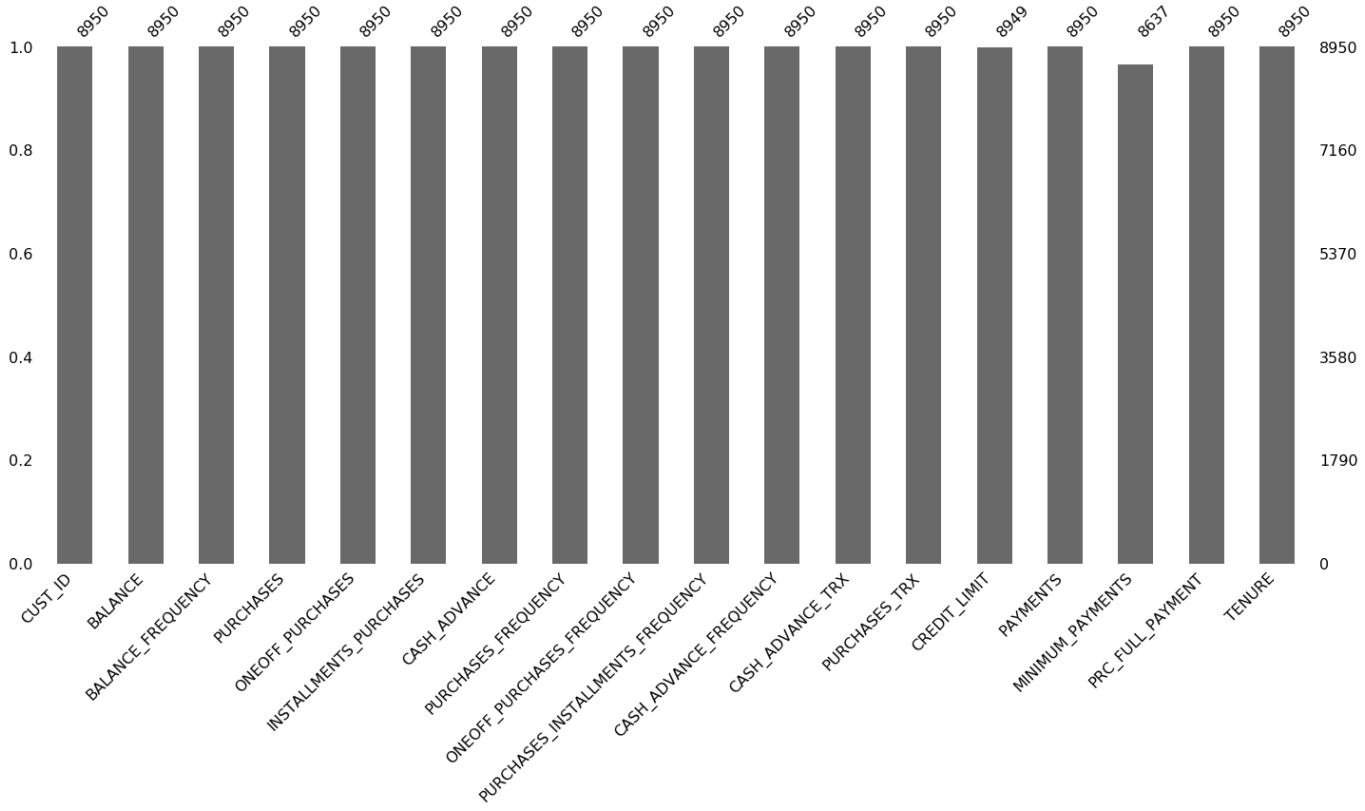


Fig 2.1: Scatter plot of all numerical variables

2.3 missing value analysis (using missigno library)

Barplot of missing value shows that there is no missing value in any row

Since missing values are more and data is not normally distributed so median will be best suited for imputation



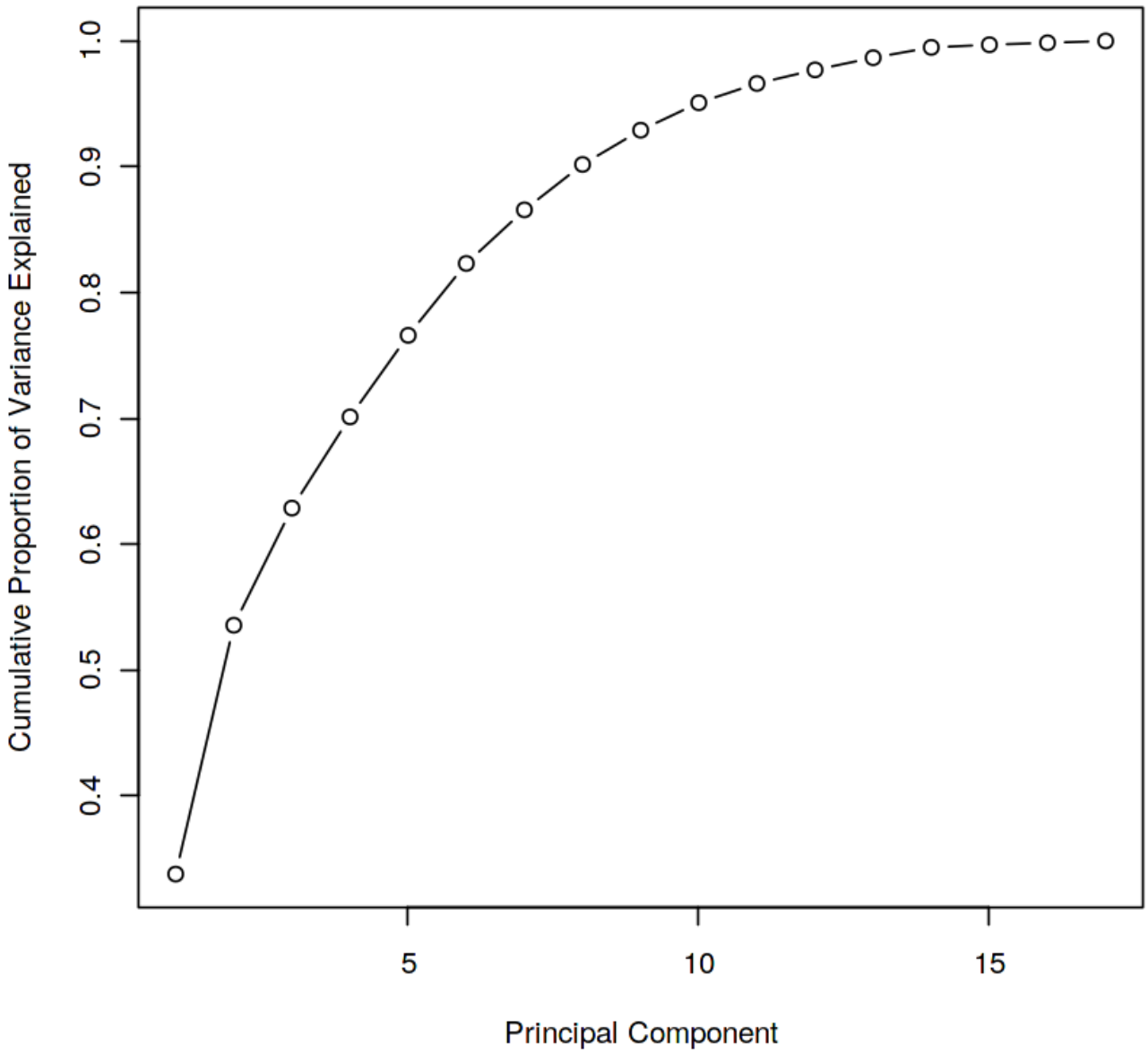
2.4 outlier removal detection by interquartile range :

By using python code it is seen that there are so many outliers so we cant remove it .since variables are skewed so log transformation is used to treat outlier.

Principal component analysis to select optimum number of features:

To select optimum no of principal component maximum variance plot is used.

It is seen from plot that 10 component can fairly explain 98.8% of variance .

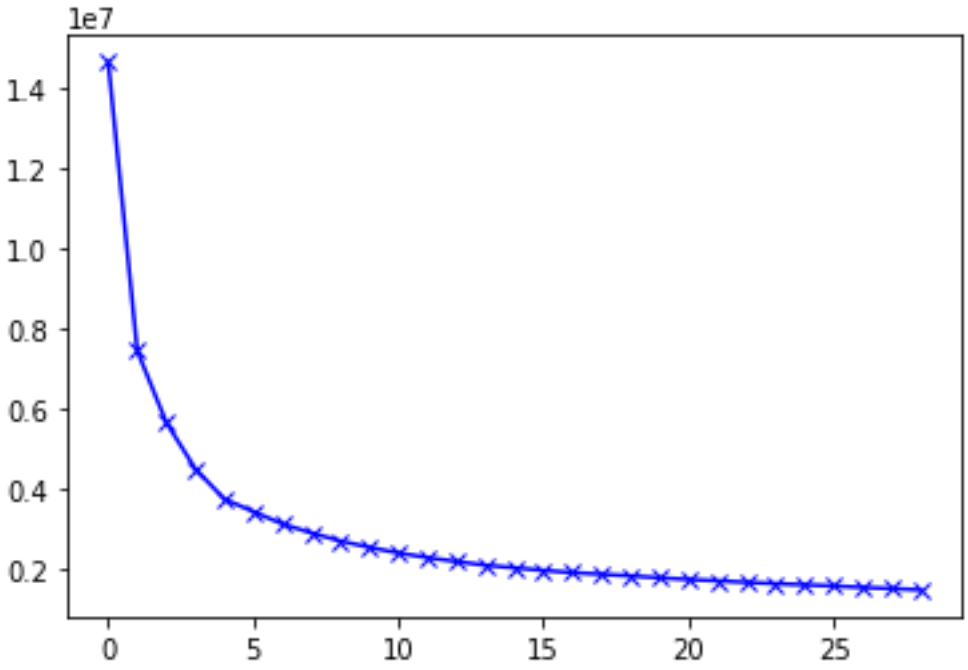


3. clustering model selection

- Hierarchical clustering can't handle big data well but K Means clustering can. This is because the time complexity of K Means is linear i.e. O(n) while that of hierarchical clustering is quadratic i.e. $O(n^2)$.
- In K Means clustering, since we start with random choice of clusters, the results produced by running the algorithm multiple times might differ. While results are reproducible in Hierarchical clustering.
- K Means is found to work well when the shape of the clusters is hyper spherical (like circle in 2D, sphere in 3D).
- K Means clustering requires prior knowledge of K i.e. no. of clusters you want to divide your data into. But, you can stop at whatever number of clusters you find appropriate in hierarchical clustering by interpreting the dendrogram

By seeing above point and comparing to data we have decided to have kmeans clustering

Selecting optimum number of cluster by using elbow method



To determine the optimal number of clusters, we have to select the value of k at the “elbow” ie the point after which the distortion/inertia start decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is **4**.
The clustered data points for different value of k:-

Percentage 100% 29% 14% 30% 16% 11%

Conclusion :

Percentage	cluster	BALANCE_FREQUENCY	PURCHASES	ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	CASH_ADVANCE	PURCHASES_FREQUENCY	ONEOFF_PURCHASES_FREQUENCY	PURCHASES_INSTALLMENTS_FREQUENCY
29	0	2277.724	0.920068	53.49933	49.1274	4.418937	2070.148	0.037801	0.028325
32	1	928.0083	0.956521	1970.408	1064.774	906.058	22.89546	0.885671	0.348066
22	2	330.7255	0.634516	369.0434	262.5154	107.037	53.0532	0.336333	0.111956
16	3	3236.864	0.978202	1645.872	1073.47	572.6193	2191.896	0.722705	0.347144

CASH_ADVANCE_FREQUENCY	CASH_ADVANCE_TRX	PURCHASES_TRX	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	PRC_FULL_PAYMENT	TENURE
0.008083	0.286468	6.770093	0.566461	4160.653	1674.75	1041.147	0.031351
0.727327	0.006905	0.098645	29.88156	4872.811	1886.362	613.6212	0.302894
0.206984	0.01338	0.203086	4.760577	3348.275	698.0788	273.8312	0.180065
0.495923	0.285891	7.370421	23.46947	5902.68	2947.158	1730.451	0.041267

Segmentation Analysis
<u>Segment 1 (Average customers) - 29%</u>
Insights:
High installment purchase spend and frequency. Primarily installment purchasers. Low one off usage and cash advance facility.

Strategy:

Push good one off offers along with attractive cash advance plans to increase purchases and monthly credit balance.

Segment 2 (Cash-advance customers) - 32%

Insights :

High cash-advances and monthly cash-advance, these customers perform more cash transactions rather than card usage at POS. High credit limit and good balance with low purchases and purchase freq.

Strategy:

Cross - sell one-off and installment offers in order to diversify the spend and increase profitability with offers like cashback,discounts on purchase ,lower fee charges for the cards. To boost installments charges we can keep interest rate low so they tend to buy more

Segment 3(Poor Customers) - 22%

Insights :

Low usage statistics on all factors. Low use of cash-advance facility.

Strategy:

Increase the credit limit to encourage spending behavior . Decrease interest rate on cash advances to gain profitability through volumes.

Segment 4 (Best customers) - 16%

Insights :

Highest Monthly avg purchase, one off and installment purchase with equal frequency on one-off and installment purchases. Highest credit limit and best repayment history. Average credit limit usage and low use of cash advance facility. Also maintains good amount of balance in their accounts

Strategy:

Incentivize cash advance facility usage with low interest rate offers accompanied by attractive deals on

one- off and installment purchases. These are customers with high repayment capacity and therefore capable of spending more. We can also introduce a scheme of loyalty points for these customers and provide better customer support so that they continue being a loyal customer and not churn in future