# Customer segmentation to define marketing strategy

Nagarjuna Yegavinti

**Problem Statement –**

The problem requires trainees 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. So, segmenting the customer profiles having similar key performance index can be done with unsupervised machine learning algorithm and a clustering problem.

**Data:**

Data contains eighteen variables each variable specifies as follows

- 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 instalment 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 instalment 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

Figure

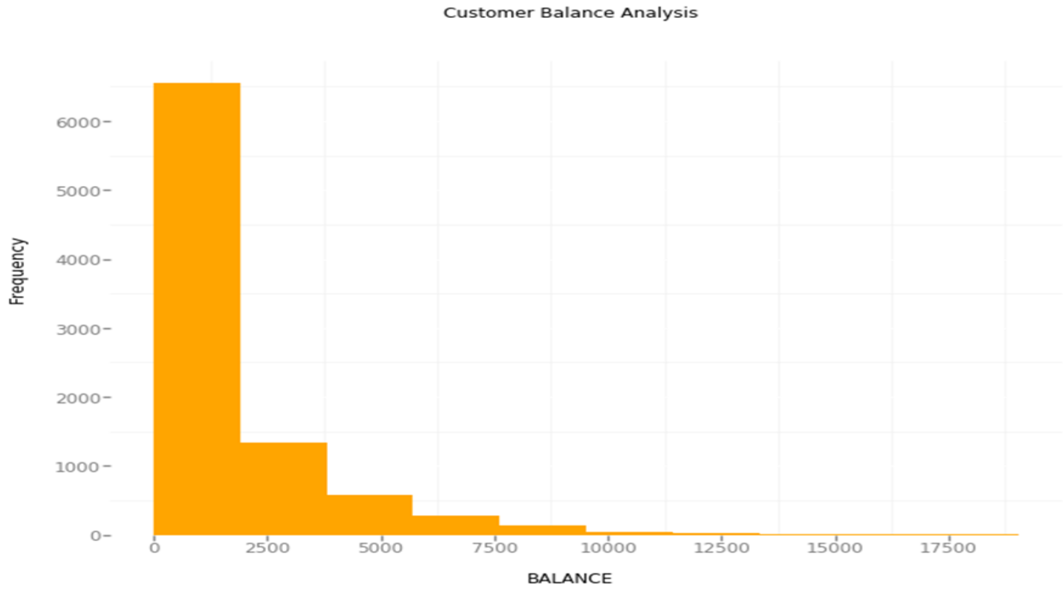


Figure.1.

Figure.1. Shows the number of customers and their balance amount with most of the customers having their balance below 2500.

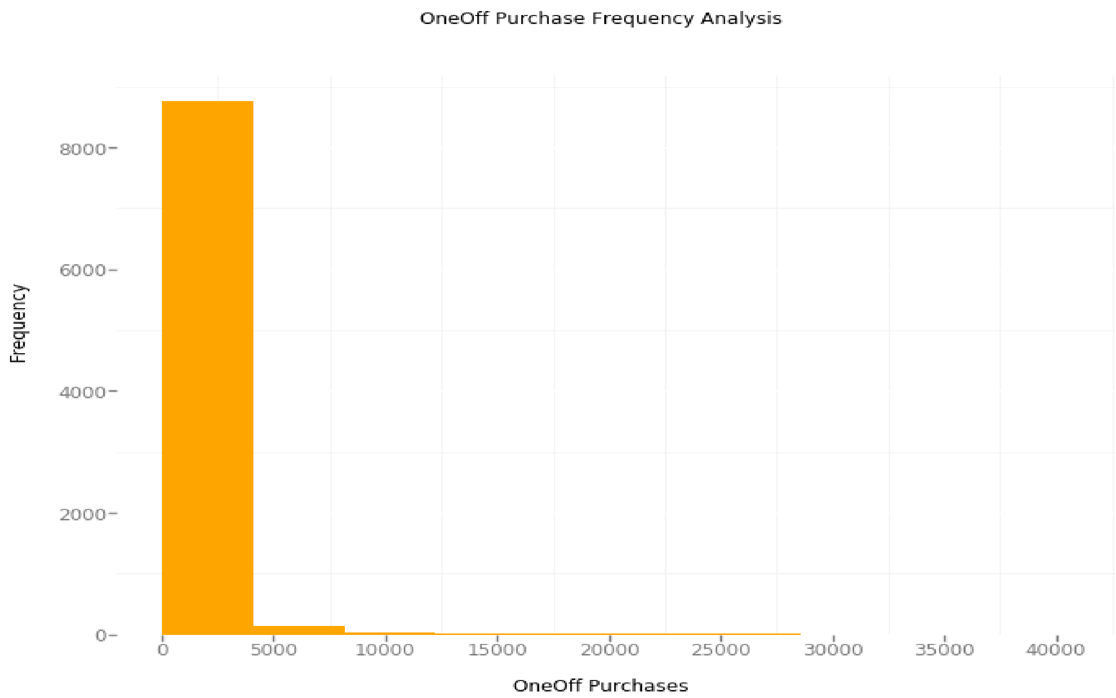


Figure.2.

Figure.2. shows the amount spent by number of customers on their one-off purchases. And in this most of the customers below 5000 on their one-off purchases.

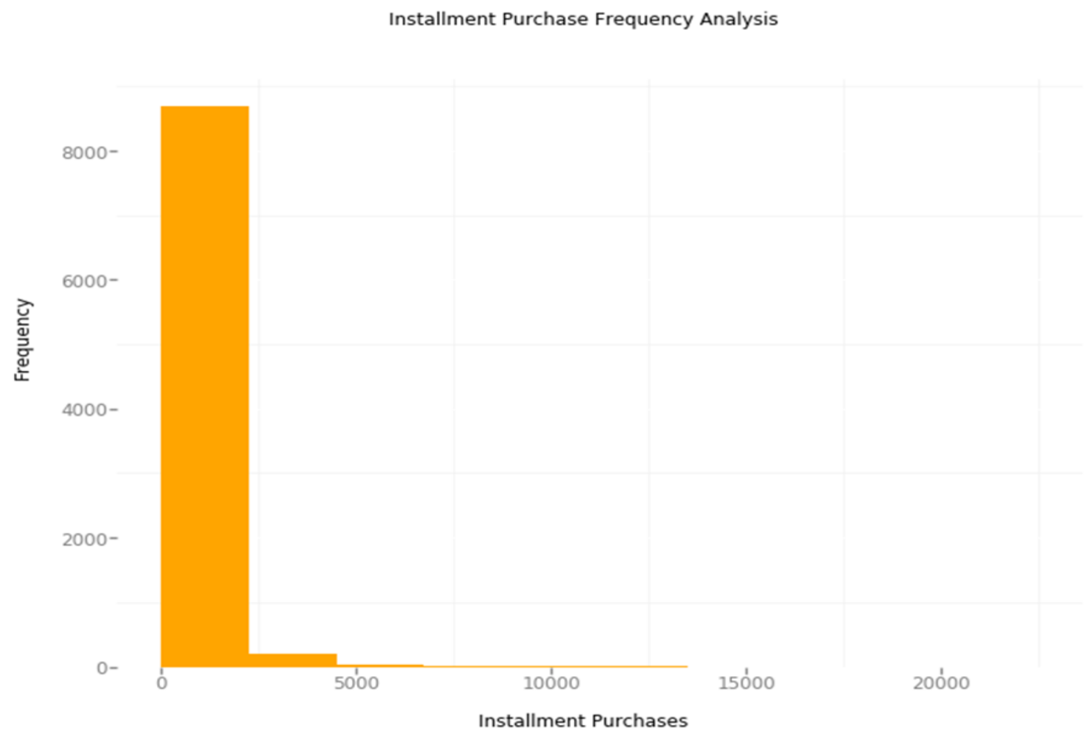


Figure.3.

Figure.3. shows the amount spent by number of customers on their instalment purchases. It shoes clearly that most of the customers having below 2500 as their instalment purchases.

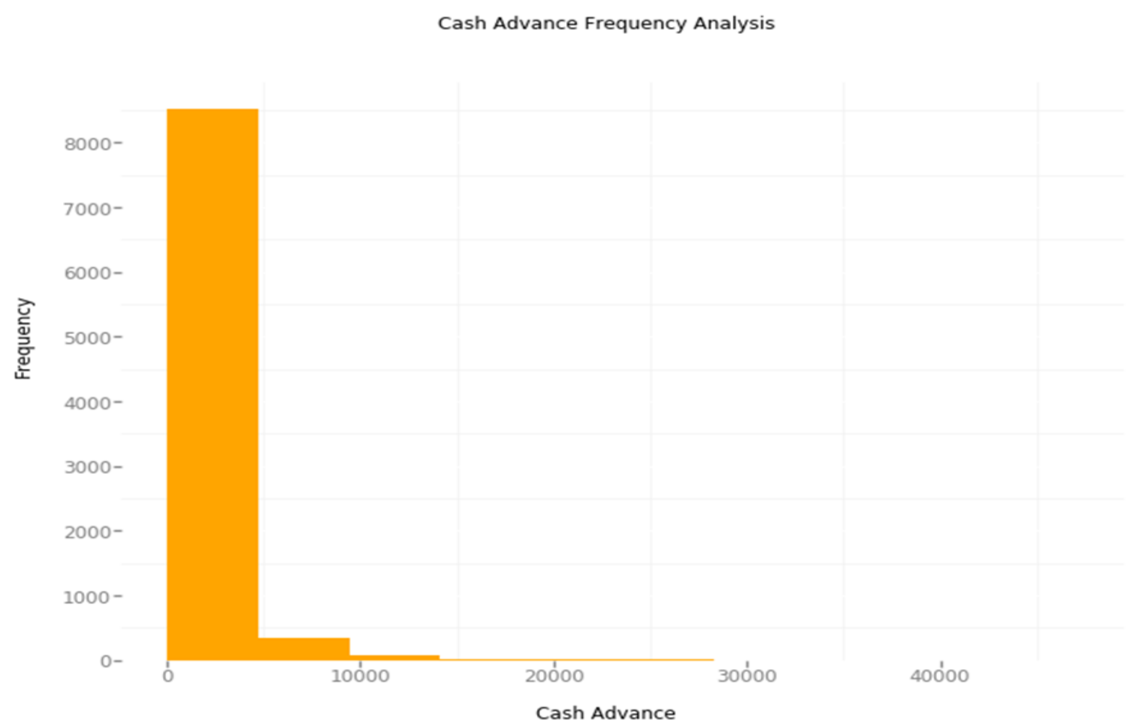


Figure.4.

Figure.4. shows the number of customers who took amount of cash advance.

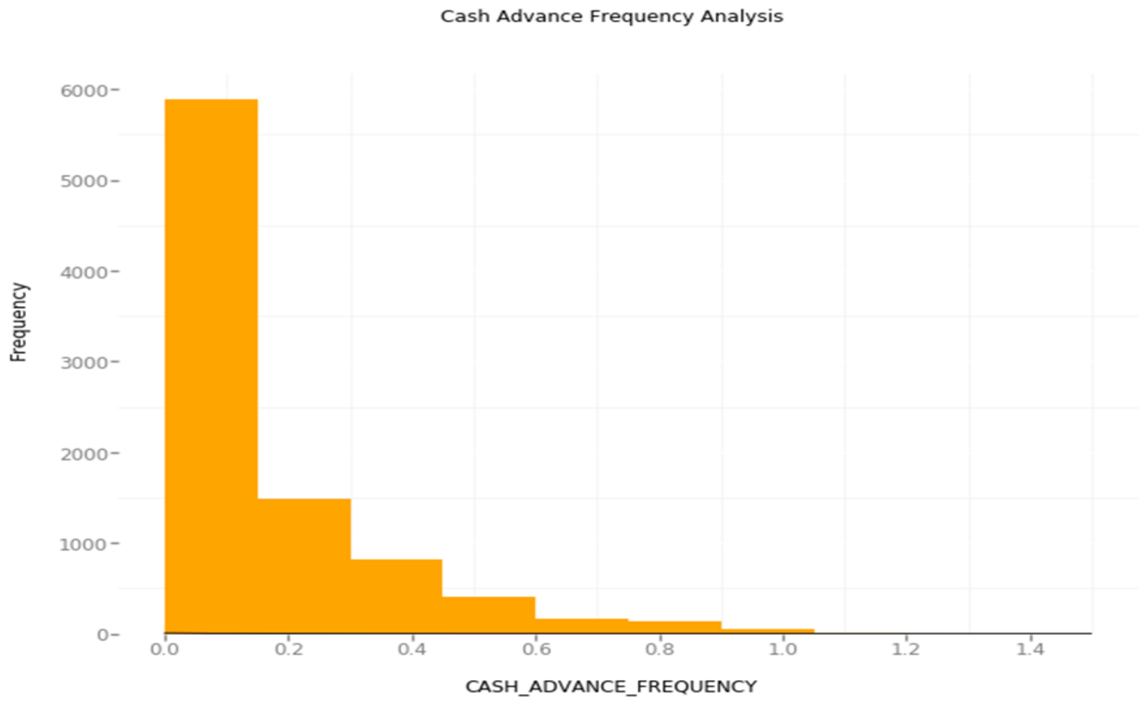


Figure.5.

Figure.5. shows the number of customers with their cash advance frequencies with most of customers having lowest cash advance transactions.

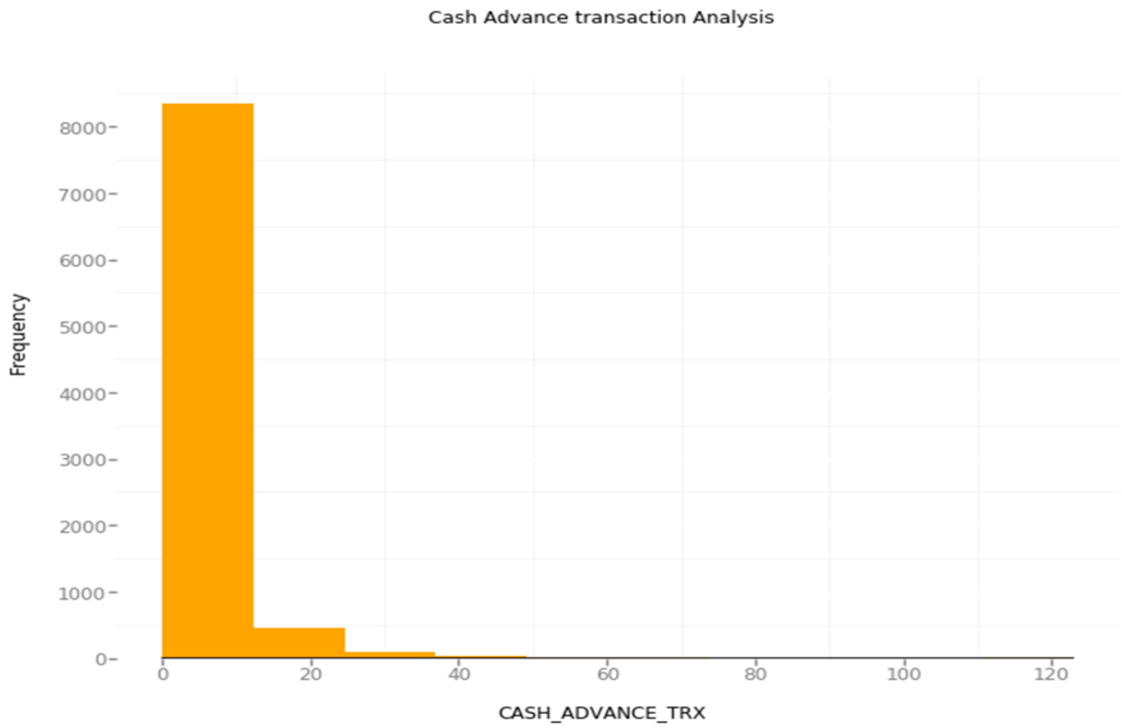


Figure.6.

Figure.6. shows the number of customers with number of cash advance transactions with most of the customers having 0 to 10 transactions and next 11 to 20 follows.

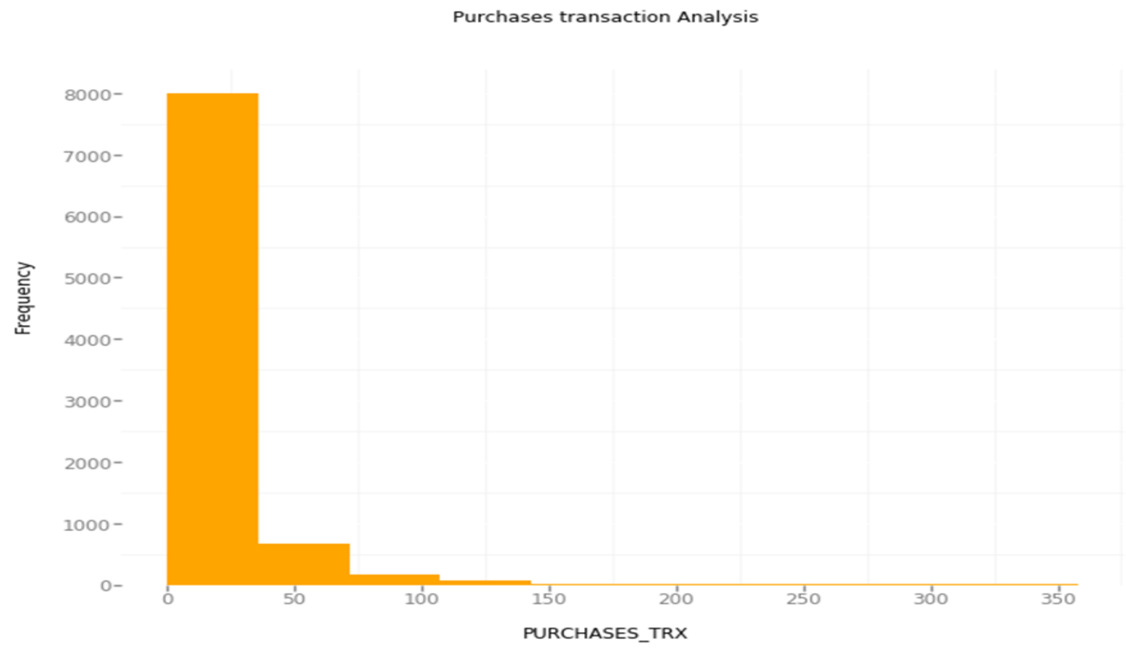


Figure.7.

Figure.7. shows number customers who's purchase transaction count. Here it showed that most of the customers having their purchase transactions below 50.

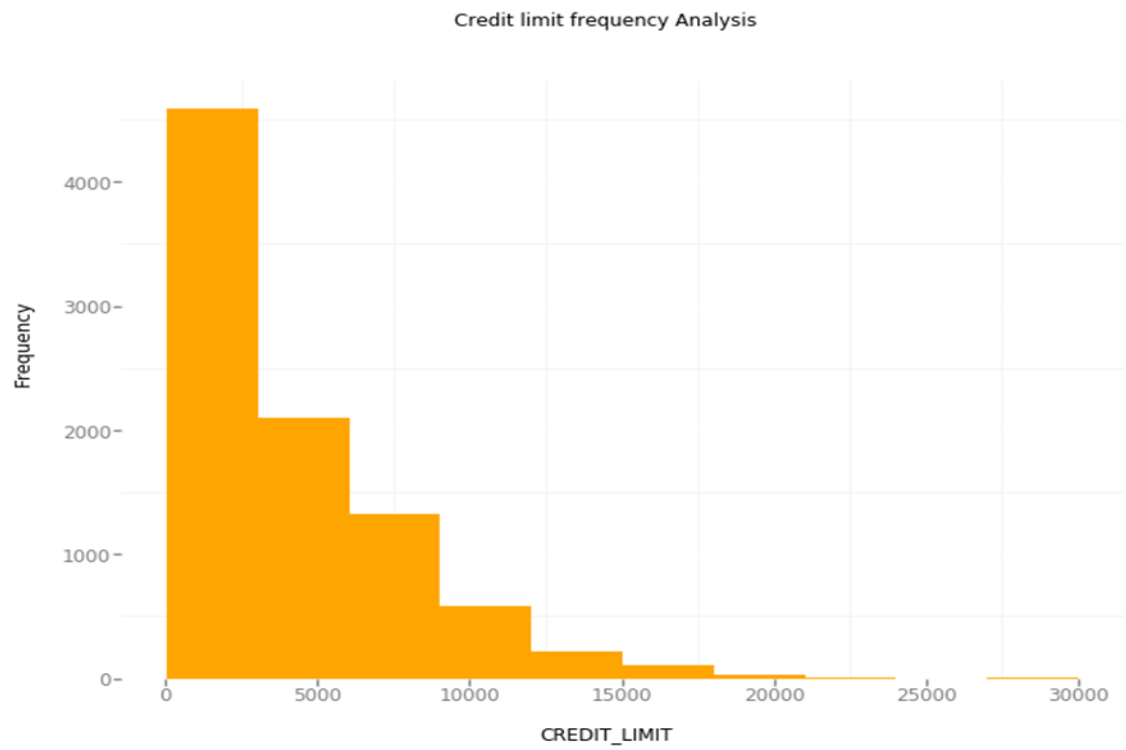


Figure.8. shows the number of customers with their credit limit. In this figure it is clear that maximum credit limit of an any customer is 30000 and most of the customers having their credit limit below 5000.

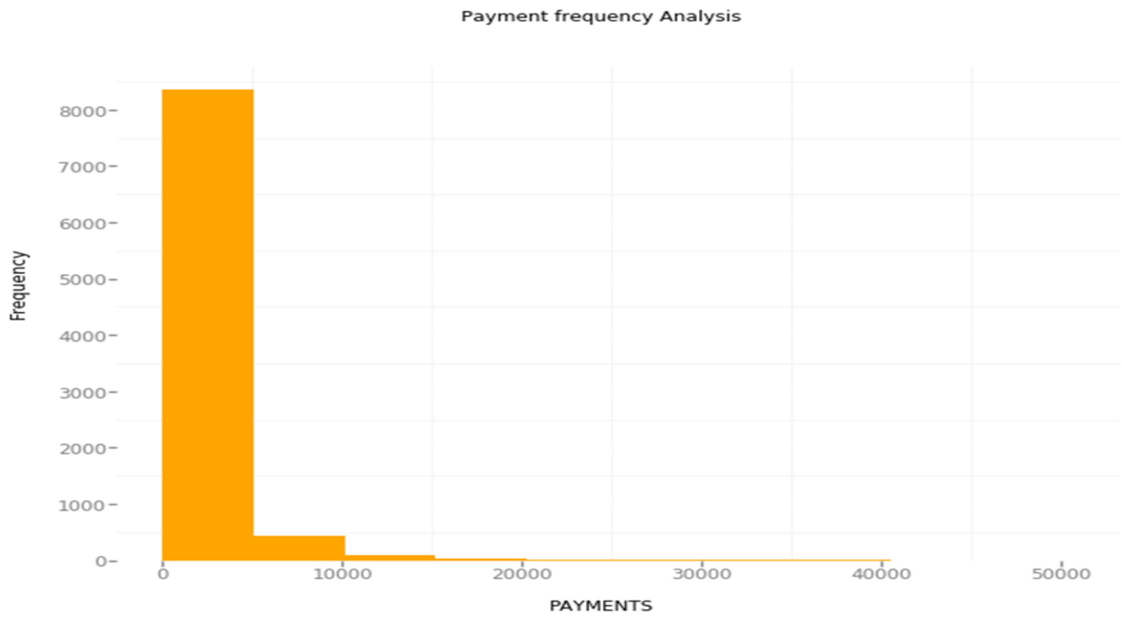


Figure.9. shows most of the customers having their payments below 5000 and with maximum of above 40000.

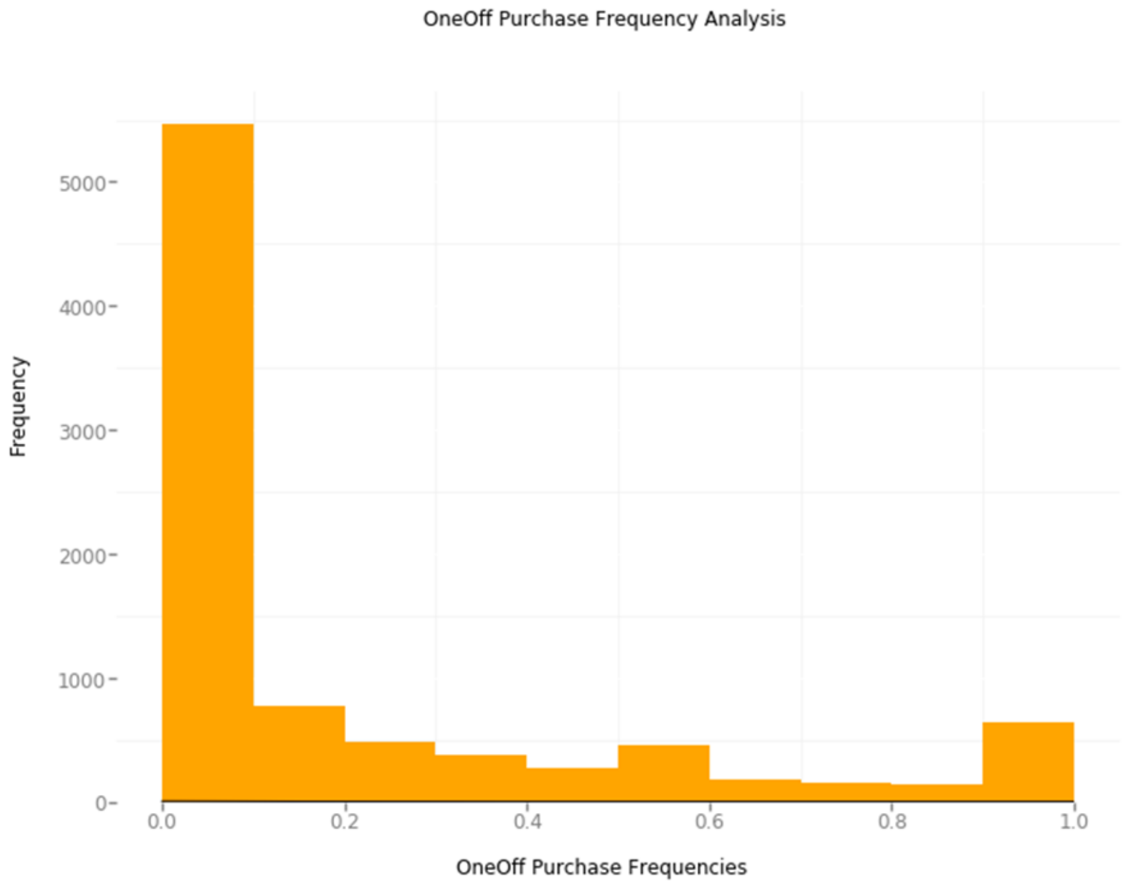


Figure.10. Showing one-off purchase frequencies

Figure.10. indicating that most of the customers are having a smaller number of one-off frequencies while there are significant number of customers who gone enough number of one-off purchase frequencies.

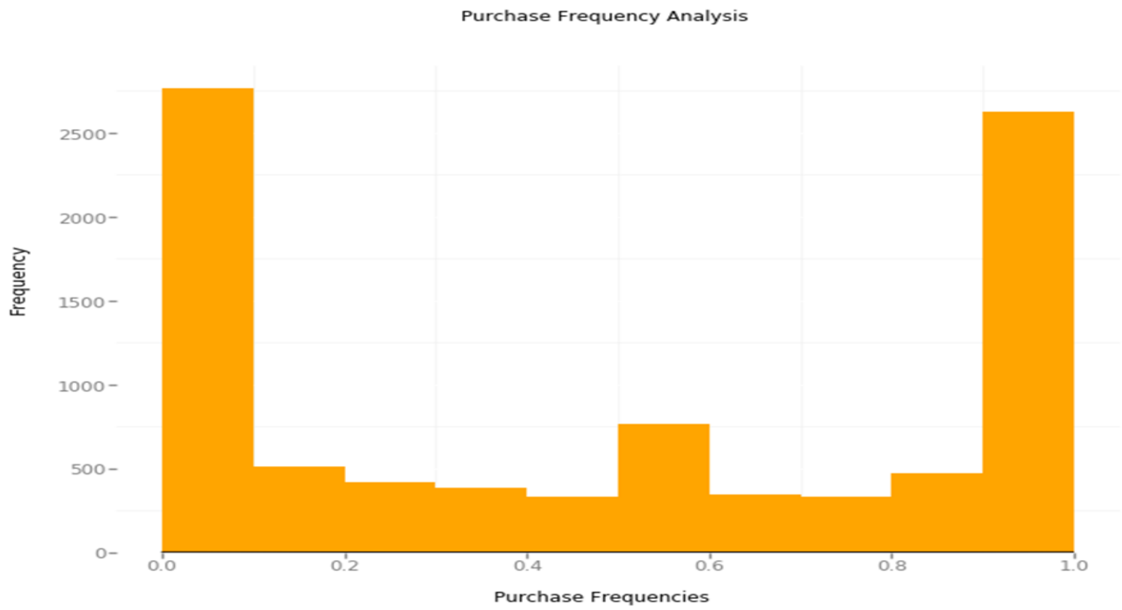


Figure.11. purchase frequencies of customers

Figure.11. indicates that purchase frequencies of customers who's transaction higher than 1.0 are more compared to lower levels while the customers who's purchase frequencies are less than 0.1 are higher compared to any other.

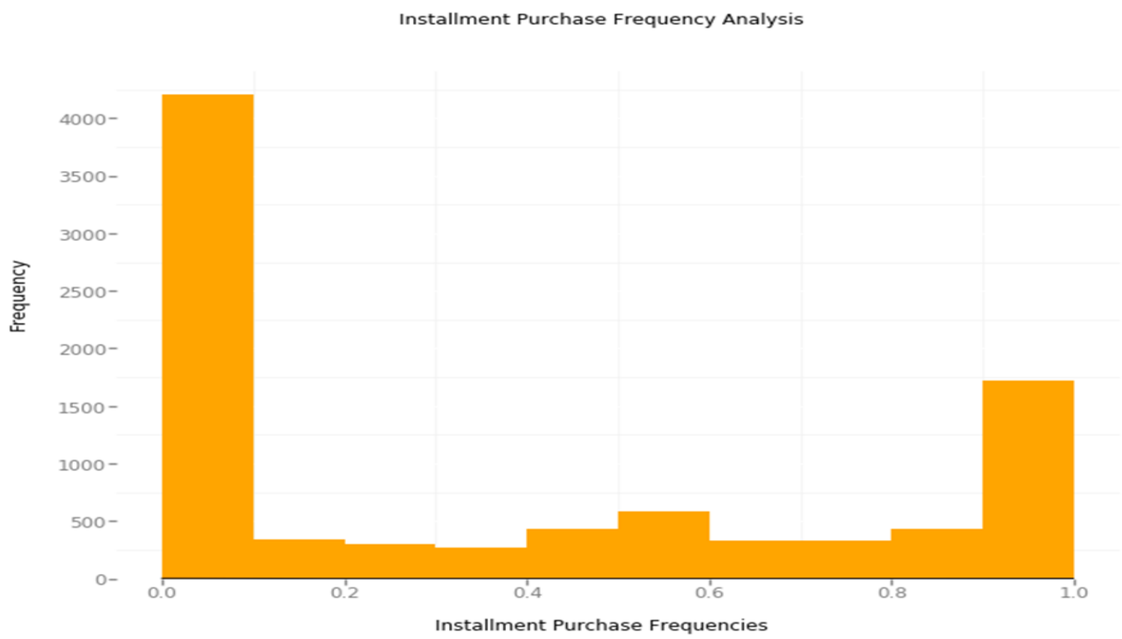


Figure.12. instalment purchase frequencies. Figure.12. indicates that there are significant number of instalment purchase transaction have been performed by customers.

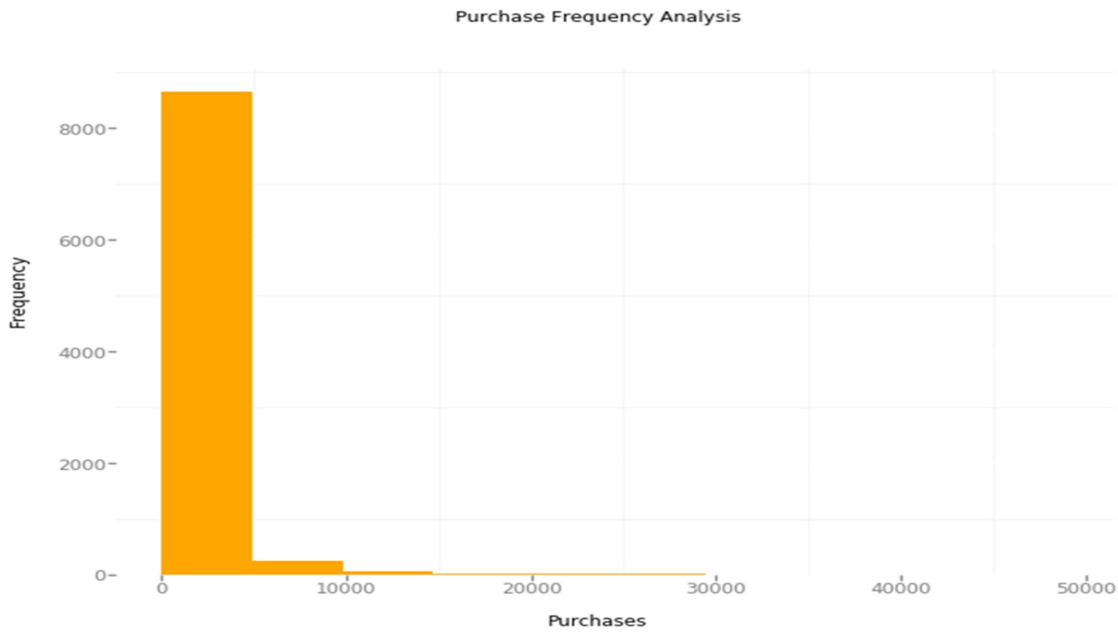


Figure.13. Purchases done by customers.

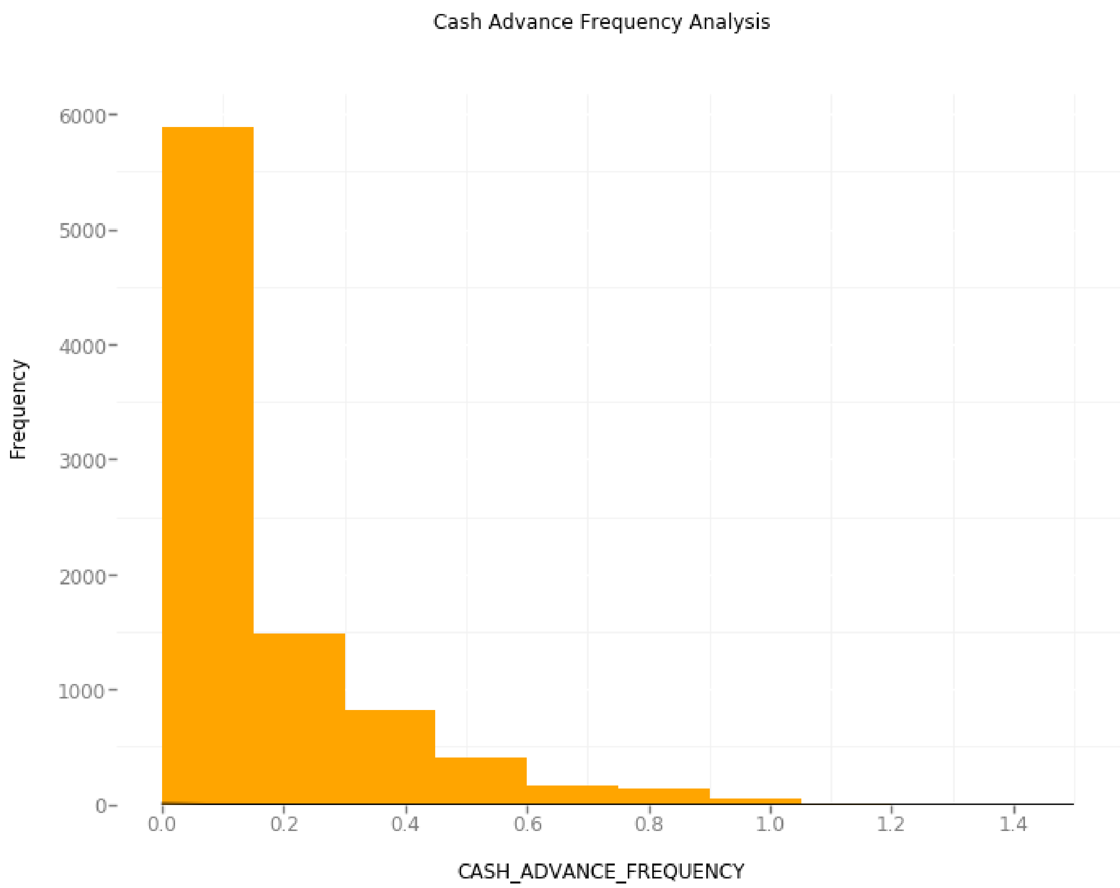


Figure.14. cash advance frequencies by customers

**Data pre-processing**

**Missing value treatment**

There are missing values present in MINIMUM_PAYMENTS and CREDIT_LIMIT variables, these missing values are treated with Knn imputation method as they are higher than 3%.

| Variables | 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 |

Table.1.

**Exploratory data Analysis /Advance Data Preparation**

Monthly average purchases were calculated from dividing the purchase with tenure. Monthly cash advances were calculated by dividing the cash advance with tenure.

#Monthly cash advance and monthly cash advance derivation

```
df['MONTH_AVG_PURCHASES'] = df['PURCHASES']/df['TENURE']

df['MONTHLY_CASH_ADVANCE'] = df['CASH_ADVANCE']/df['TENURE']
```

Extracting purchase types from one-off and instalment purchases

Four different purchase types have been defined from both one-off and instalment purchase types.

```
#Python code for creating a definition for new features
def purchase(credit):
    if (credit['ONEOFF_PURCHASES']==0) & (credit['INSTALLMENTS_PURCHASES']==0):
        return 'NONE'
    if (credit['ONEOFF_PURCHASES']>0) & (credit['INSTALLMENTS_PURCHASES']>0):
        return 'ONEOFF_INSTALLMENT'
    if (credit['ONEOFF_PURCHASES']>0) & (credit['INSTALLMENTS_PURCHASES']==0):
        return 'ONEOFF'
    if (credit['ONEOFF_PURCHASES']==0) & (credit['INSTALLMENTS_PURCHASES']>0):
        return 'INSTALLMENT'
#Applying the new features to data
df['PURCHASE_TYPE’] =df.apply(purchase,axis=1)
```

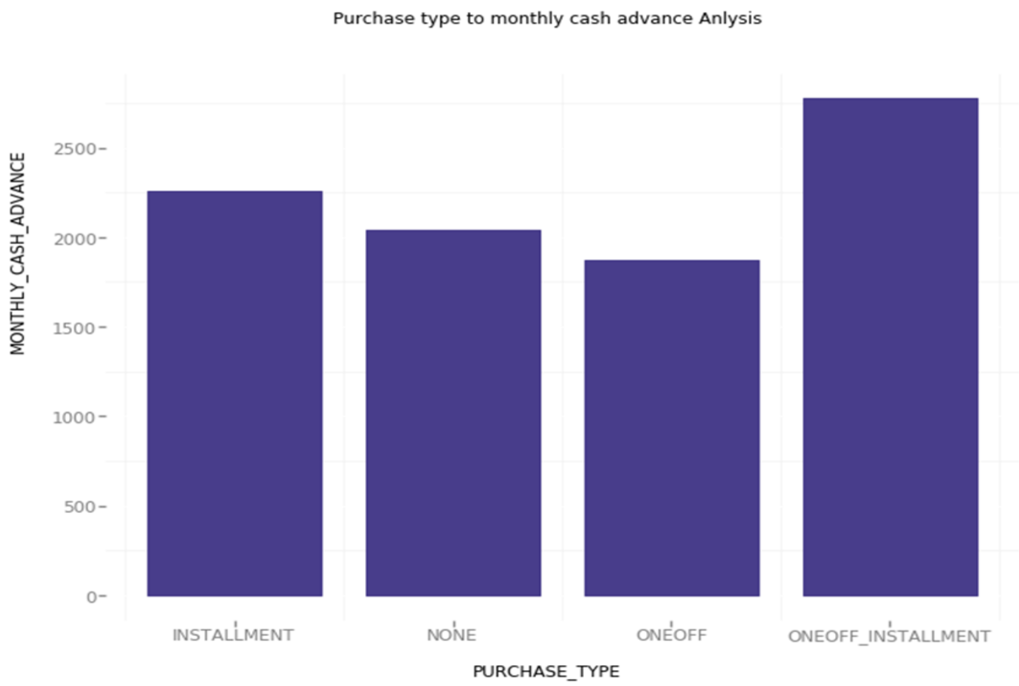


Figure.15.

From the figure.15. it indicates monthly cash advance for both one-off and instalment purchase types are high followed by instalment purchase types.

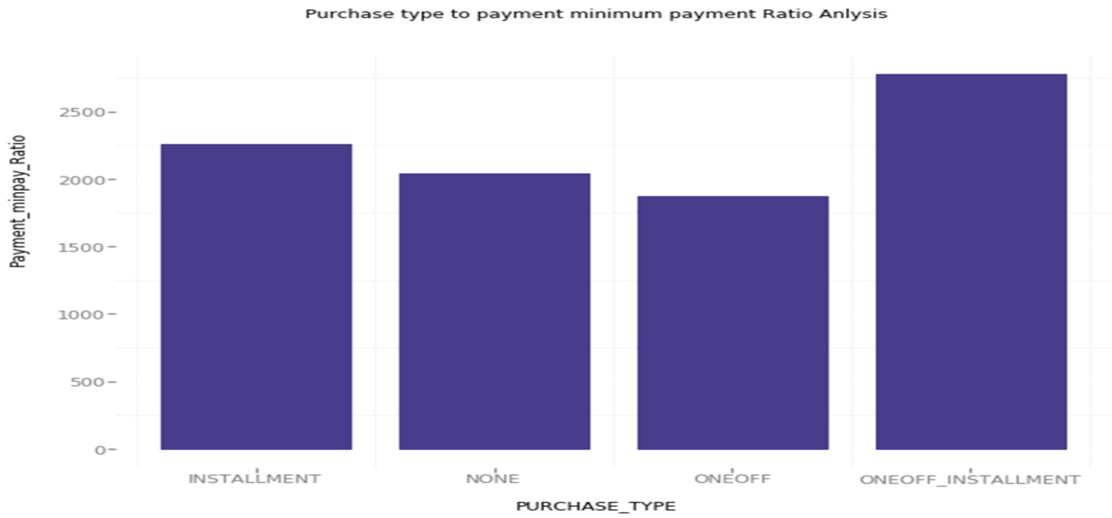


Figure.16. from the figure we can see that payment to minimum payment ratio were high for both one-off and instalment purchase types followed by instalment purchase types.

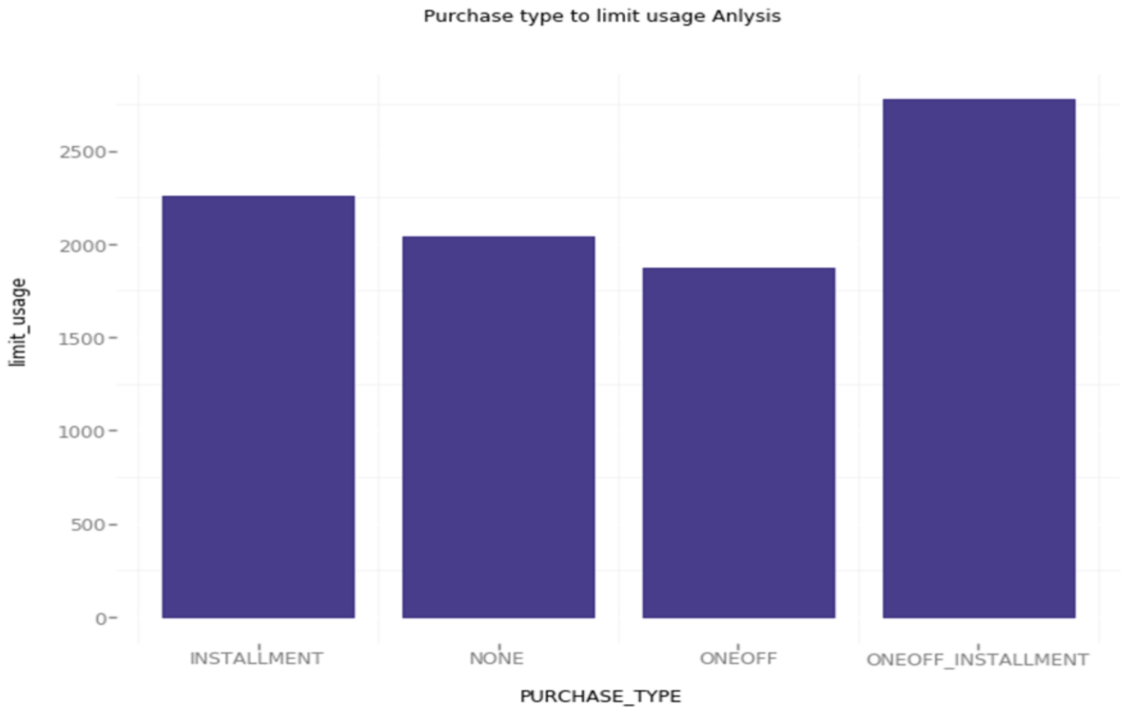


Figure.17. showing highest limit usage done by the customers who have done both one-off and instalment purchases followed by the instalment purchases.

Limit usage by each customer was calculated with the ratio of balance to credit limit ratio and payment to minimum payment ratio has been calculated.

```
# Limit usage calculation from balance to credit ratio

df['limit_usage']=df.apply(lambda x: x['BALANCE']/x['CREDIT_LIMIT'], axis=1)
```

#Payments to minimum payments ratio calculation

df['Payment_minpay_Ratio'] = df.apply(lambda x: x['PAYMENTS']/x['MINIMUM_PAYMENTS'], axis=1)

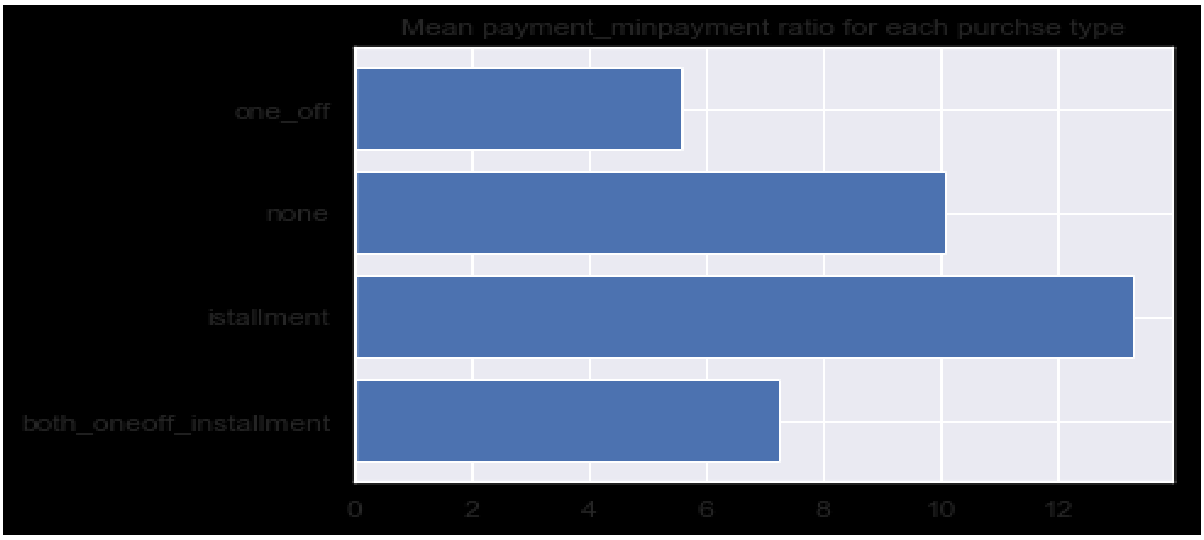


**Figure.18.**

**Outlier treatment**

There are different methods to treat outliers here all the values have been transformed into log values to overcome wide range of outliers

**Feature selection & Dimensionality reduction**

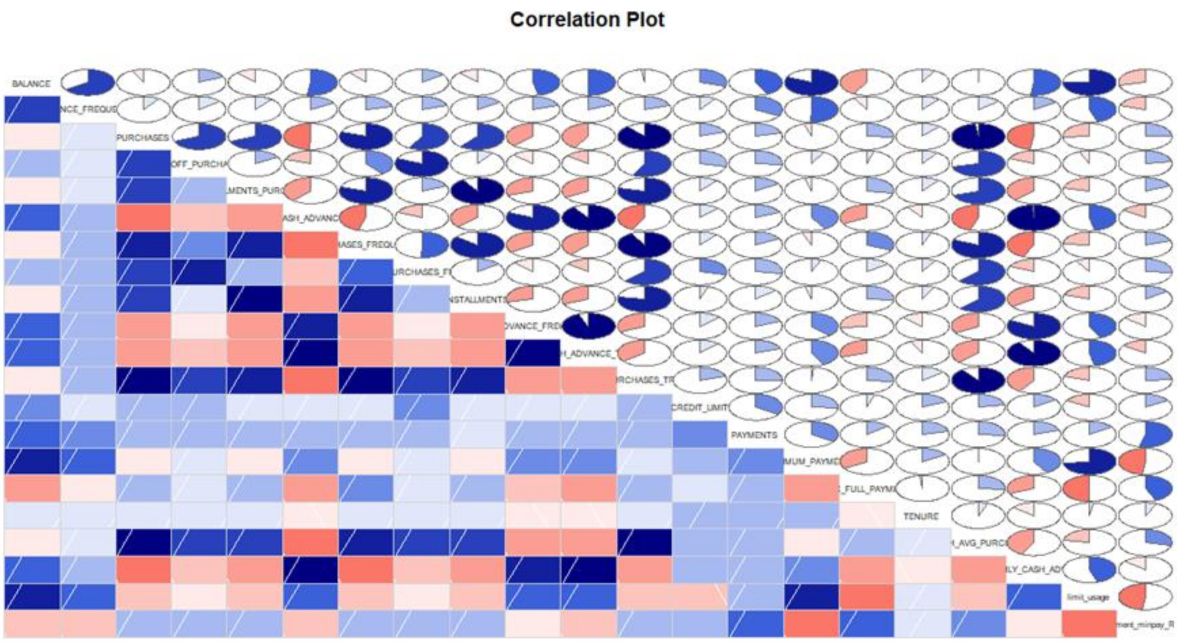


**Figure.18. correlation plot**

Correlation analysis has been checked to see if there are any variables which are correlated with each other. PURCHASES_FREQUENCY and CASH_ADVANCE_FREQUENCY variables had been removed for their high correlation with PURCHASE_TRX and CASH_ADVANCE_TRX variables. BALANCE and CASH_ADVANCE variables have been removed as they are used for feature extraction and their correlation with their extracted variables.

**Feature scaling**

Due to the variability in the range of variables all the variables have been brought down to normal range by applying normalization

```
#Normalisation

for i in cnames:

    print(i)

    df2[i] = (df2[i] - min(df2[i]))/(max(df2[i]) - min(df2[i]))
```

Dummy variables have been extracted for each category of newly derived variable PURCHASE_TYPE. Once dummy variable has been created, now each category includes instalment, one-off, none and both one-off and instalment purchase have been grouped into new categories.

**Application of clustering machine learning algorithm**

There are many machine learning algorithms for clustering the data, here it’s been applied Kmeans clustering model. Initially bar graph was plot to check to number of clusters to be built and it is determined to be the optimum no of clusters are four. Kmeans is the function which helps to build the kmeans algorithm on the data.

```
#Implement kmeans

kmeans_model = KMeans(n_clusters = 4).fit(df2.iloc[:,0:22])
```

**Conclusions**

```
K-means clustering with 4 clusters of sizes 1874, 2260, 2774, 2042

Cluster means:
```

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| BALANCE_FREQUENCY | 0.8586794 | 0.8447376 | 0.9556123 | 0.8989203 |
| PURCHASES | 0.5388247 | 0.5303136 | 0.6676056 | 0.0000000 |
| ONEOFF_PURCHASES | 0.5482454 | 0.0000000 | 0.6035077 | 0.0000000 |
| INSTALLMENTS_PURCHASES | 0.0000000 | 0.5720356 | 0.6132729 | 0.0000000 |
| ONEOFF_PURCHASES_FREQUENCY | 0.3704713 | 0.0000000 | 0.4884200 | 0.0000000 |
| PURCHASES_INSTALLMENTS_FREQUENCY | 0.0000000 | 0.7020842733 | 0.6849570164 | 0.0002032503 |
| CASH_ADVANCE_TRX | 0.15369796 | 0.06891247 | 0.13537861 | 0.34216493 |
| PURCHASES_TRX | 0.2742979395 | 0.3871019412 | 0.5382726636 | 0.0001491425 |

| | | | | |
|---|---|---|---|---|
| CREDIT_LIMIT | 0.6539738 | 0.6110515 | 0.6982519 | 0.6360695 |
| PAYMENTS | 0.5935374 | 0.5641628 | 0.6682951 | 0.6030079 |
| MINIMUM_PAYMENTS | 0.5151027 | 0.4874549 | 0.5362390 | 0.5483329 |
| PRC_FULL_PAYMENT | 0.11335603 | 0.27658904 | 0.2067088 1 | 0.05348513 |
| MONTH_AVG_PURCHASES | 0.4167865 | 0.4038627 | 0.5731052 | 0.0000000 |
| MONTHLY_CASH_ADVANCE | 0.2313755 | 0.1121314 | 0.1983623 | 0.5401009 |
| limit_usage | 0.10321402 | 0.07220634 | 0.0979956 6 | 0.15113644 |
| Payment_minpay_Ratio | 0.1501091 | 0.1449388 | 0.1851049 | 0.1293493 |
| data_INSTALLMENT | 0 | 1 | 0 | 0 |
| data_NONE | 0 | 0 | 0 | 1 |
| data_ONEOFF | 1 | 0 | 0 | 0 |
| data_ONEOFF_INSTALLMENT | 0 | 0 | 1 | 0 |

Table.2.

- We can see from the table.2. that out of the four clusters of customers, cluster 3 c ustomers are outperforming over other group customers and having good credit score with high both one-off and instalment purchases and on time payments.
- Group 2 customers are high PRC full payments and instalment purchase frequenc y.
- While the customers in group four are having maximum minimum payments and cash advance transactions.
- Customers belonging to cluster three can be given more credit limit and reward points.