Credit Card Segmentation

By :- Rahul Prasad Sah

Contents :-

Introduction

Problem Statement:-

- The objective of this project 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.

Data:-

The details of data attribute in the dataset 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 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 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

Methodology

- Data Pre-Preprocessing :-
- Data pre-processing is the first stage of any type of project. We do this by looking at plots of independent variables vs target variables. If the data is messy, we try to improve it by sorting deleting extra rows and columns. This stage is called as Exploratory Data Analysis. This stage generally involves data cleaning, merging, sorting, looking for outlier analysis, looking for missing values in the data, Imputing missing values if found by various methods such as mean, median, mode, KNN imputation, etc.

Missing value Analysis

- In this step we look for missing values in the dataset like empty row column cell which was left after removing special characters and punctuation marks.

- Some missing values are in form of NA. missing values left behind after outlier analysis; missing values can be in any form. Unfortunately, in this dataset we found missing values.

```
CUST_ID                                   False
BALANCE                                   False
BALANCE_FREQUENCY                         False
PURCHASES                                 False
ONEOFF_PURCHASES                          False
INSTALLMENTS_PURCHASES                    False
CASH_ADVANCE                              False
PURCHASES_FREQUENCY                       False
ONEOFF_PURCHASES_FREQUENCY                False
PURCHASES_INSTALLMENTS_FREQUENCY          False
CASH_ADVANCE_FREQUENCY                    False
CASH_ADVANCE_TRX                          False
PURCHASES_TRX                             False
CREDIT_LIMIT                               True
PAYMENTS                                  False
MINIMUM_PAYMENTS                           True
PRC_FULL_PAYMENT                          False
TENURE                                    False
```

1. **Monthly average purchase and cash advance amount for credit card :-**

```
0      7.950000
1      0.000000
2     64.430833
3    124.916667
4      1.333333

                       Name: Monthly_avg_purchase

                                 0    12
                                 1    12
                                 2    12
                                 3    12
                                 4    12

                                                           Name: TENURE
                                                           0       95.40
                                                           1        0.00
                                                           2      773.17
                                                           3     1499.00
                                                           4       16.00
```

2. Monthly Cash Advanced Amount

Monthly cash advanced amount is : 4302

3. Purchases by type (one-off, instalments)

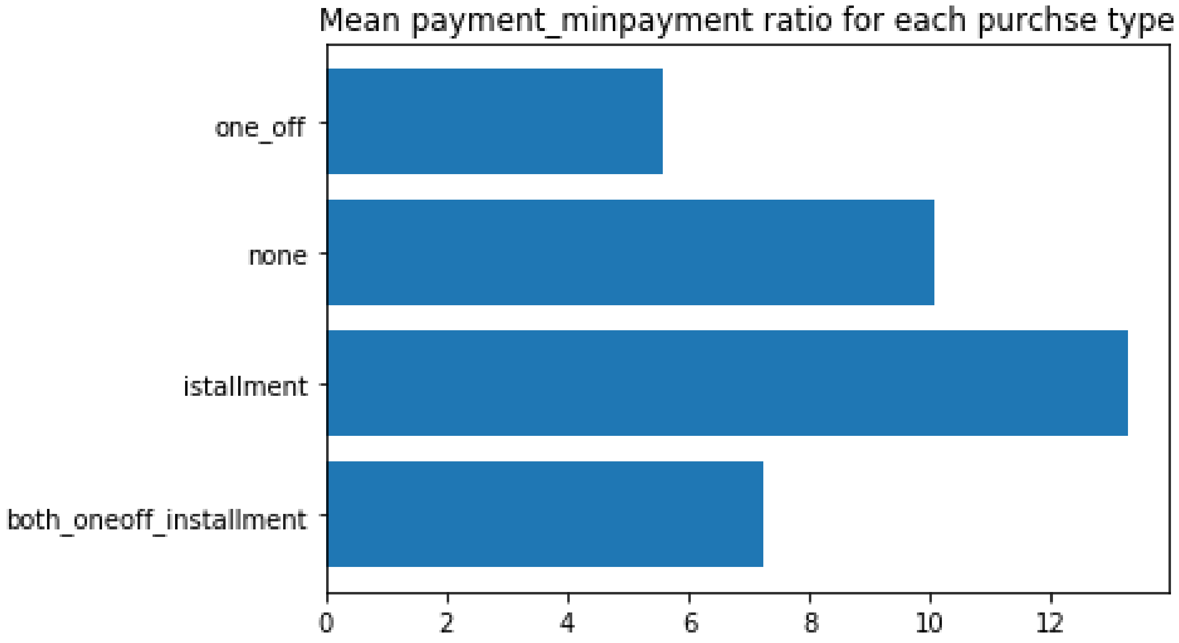
ONEOFF_PURCHASES	INSTALLMENTS_PURCHASES	
0	0.00	95.40
1	0.00	0.00
2	773.17	0.00
3	1499.00	0.00
4	16.00	0.00
5	0.00	1333.28
6	6402.63	688.38
7	0.00	436.20
8	661.49	200.00
9	1281.60	0.00
10	0.00	920.12

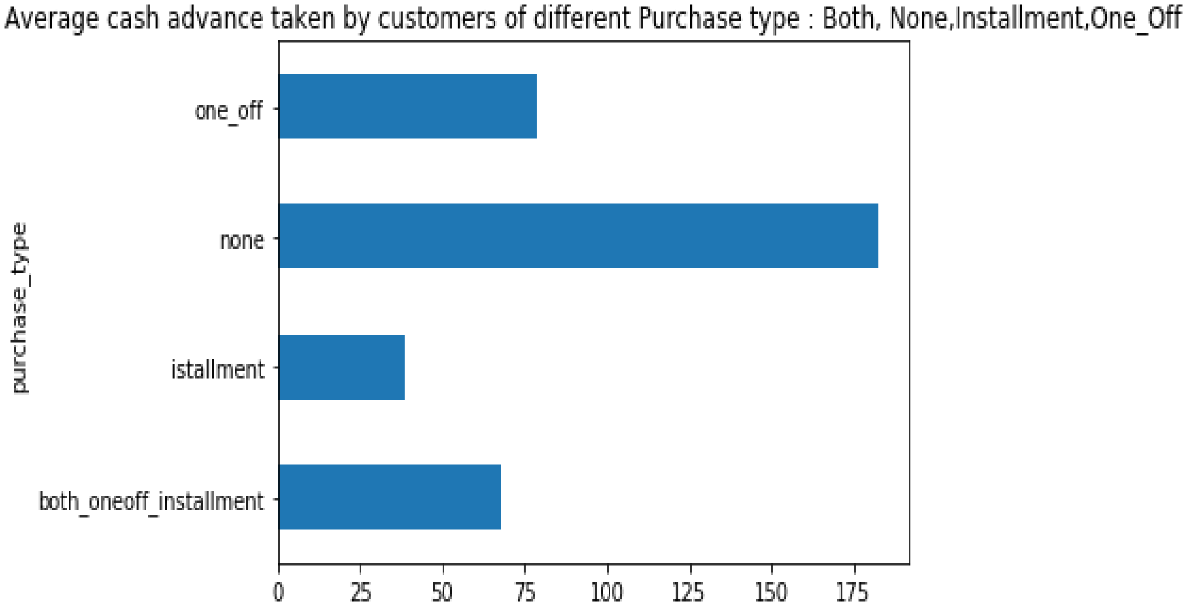
4. Finding the customers ONEOFF_PURCHASES and INSTALLMENTS PURCHASES details :-

```
both_oneoff_installment    2774
istallment                 2260
none                       2042
one_off                    1874
```

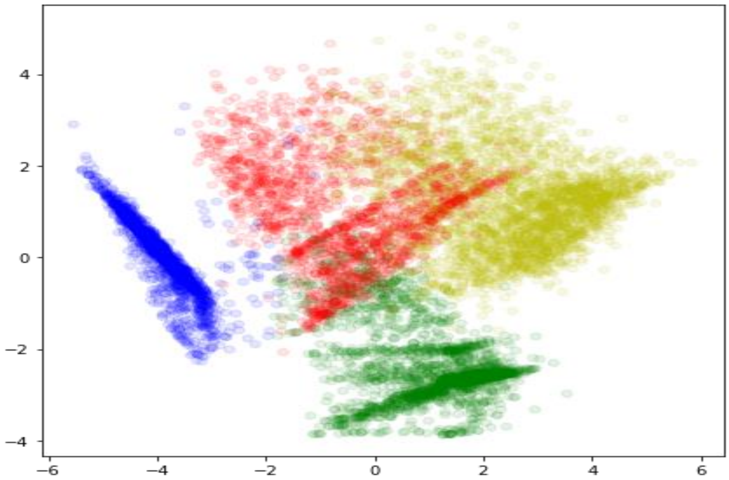
1. Analysis of data via visualization:-

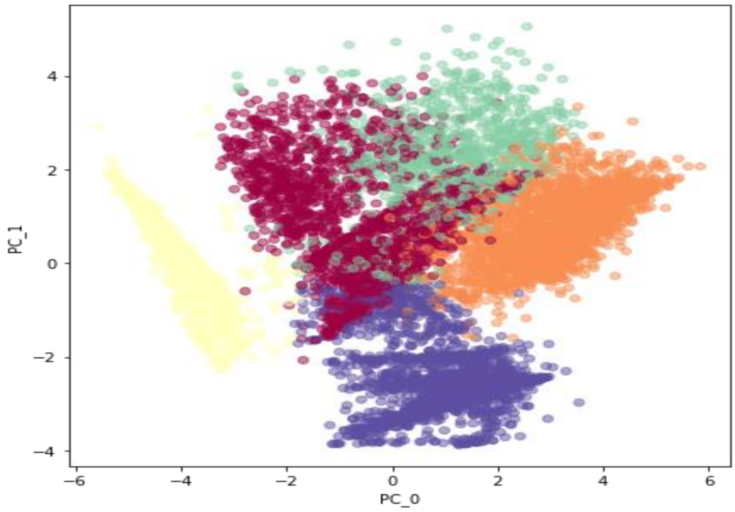
- By Bar Plots :-

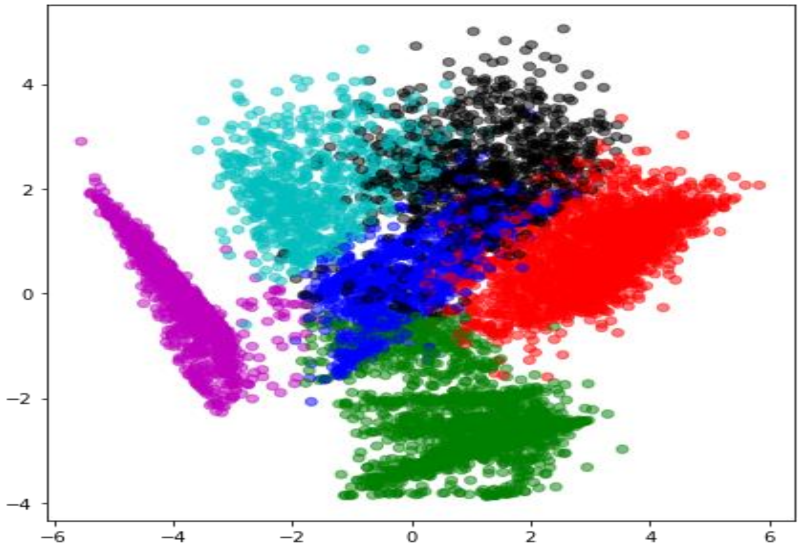




By Scatter Plot :-

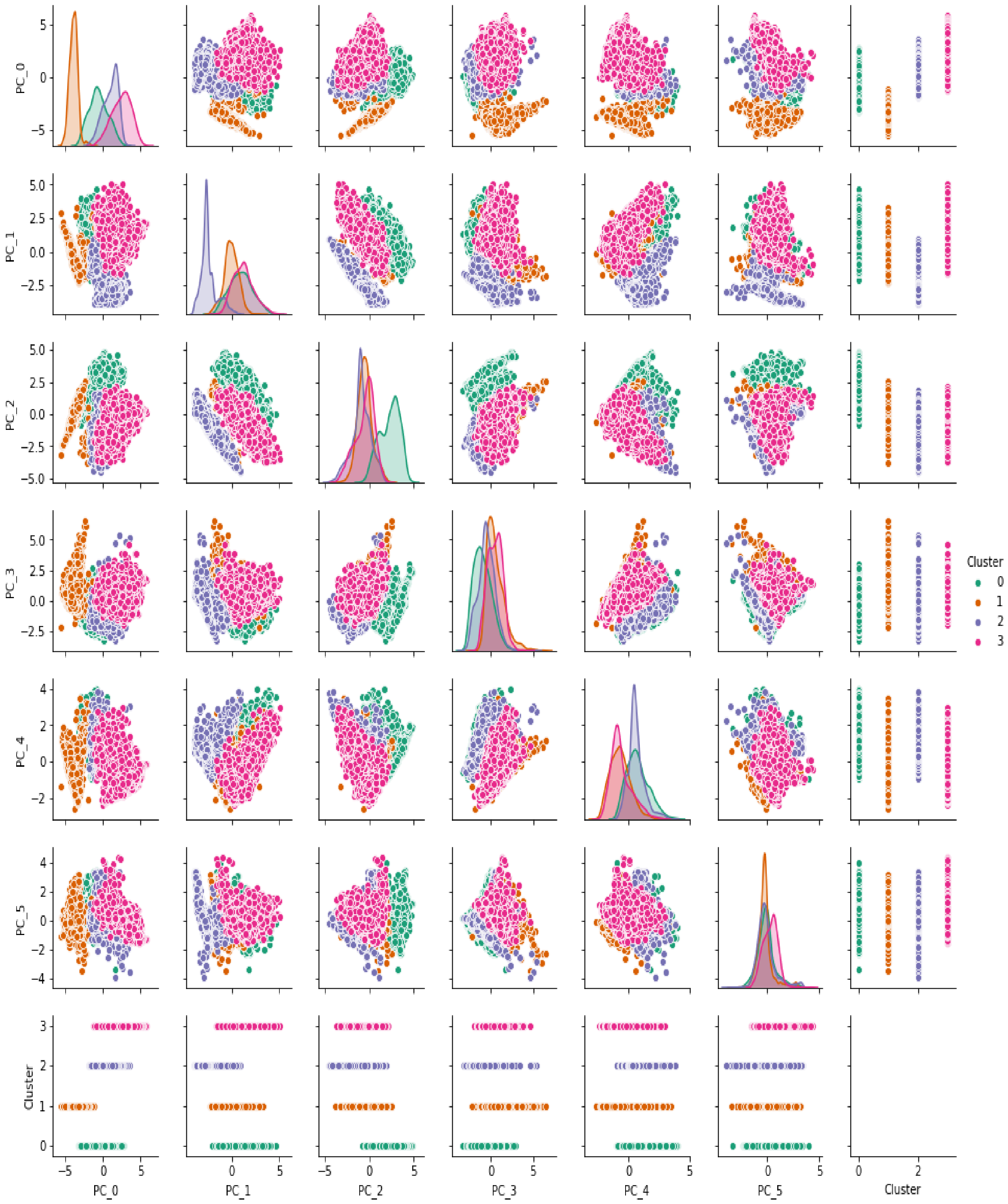




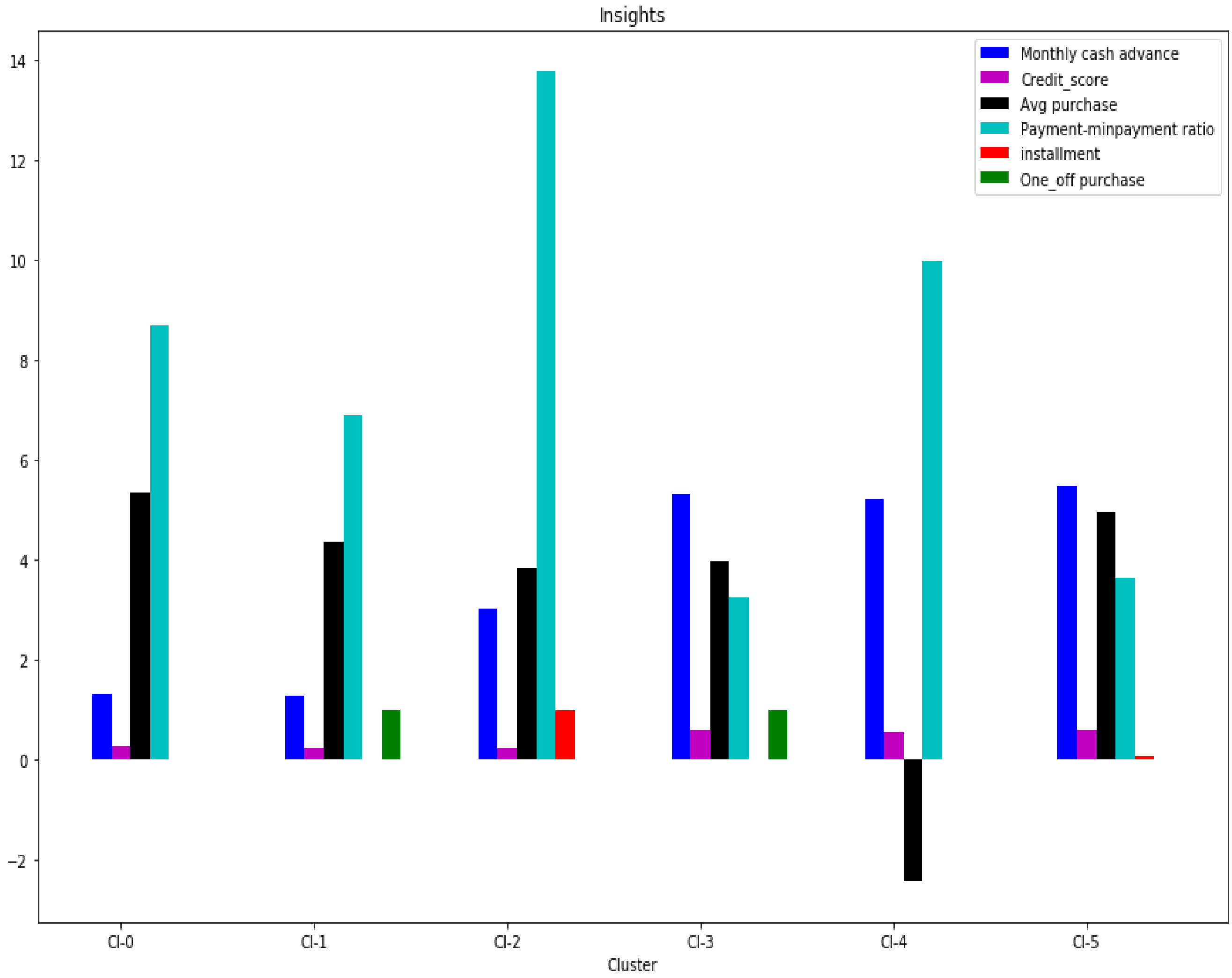


By Pair plots :-

Pairwise Plots for all Numerical variables:



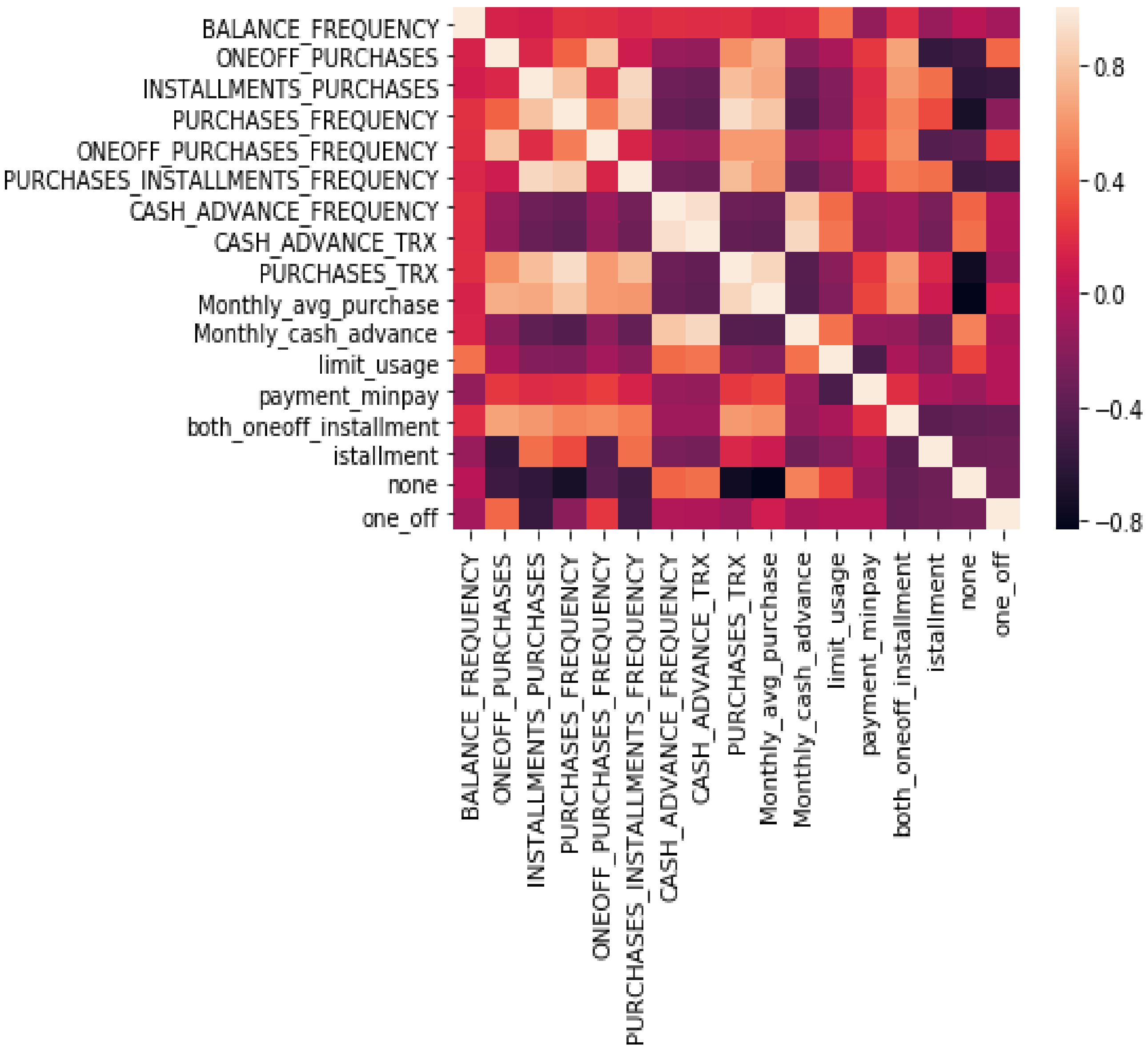
By Subplots Bar Graph :-



- Feature Selection:-

In this dataset we have to develop a customer segmentation to define marketing strategy : -

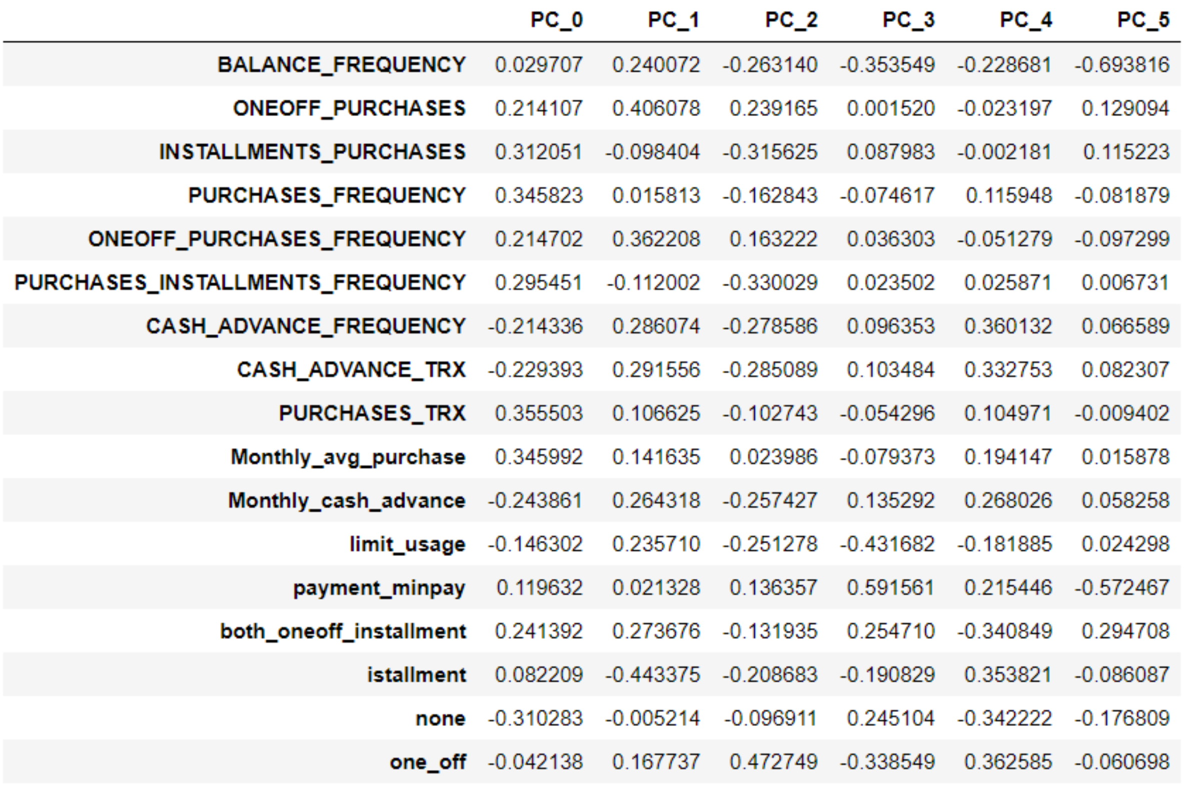
- **Correlation analysis** – This requires only numerical variables. Therefore, we will filter out only numerical variables and feed it to correlation analysis. We do this by plotting correlation plot for all numerical variables. There should be no correlation between independent variables but there should be high correlation between independent variable and dependent variable. So, we plot the correlation plot. we can see that in correlation plot faded colour like skin colour indicates that two variables are highly correlated with each other.



Applying Principal Component Analysis (PCA):-

PCA is essentially a method that reduces the dimension of the feature space in such a way that new variables are orthogonal to each other (i.e. they are independent or not correlated)

After applying PCA to every feature, we come up with following results :-



	PC_0	PC_1	PC_2	PC_3	PC_4	PC_5
BALANCE_FREQUENCY	0.029707	0.240072	-0.263140	-0.353549	-0.228681	-0.693816
ONEOFF_PURCHASES	0.214107	0.406078	0.239165	0.001520	-0.023197	0.129094
INSTALLMENTS_PURCHASES	0.312051	-0.098404	-0.315625	0.087983	-0.002181	0.115223
PURCHASES_FREQUENCY	0.345823	0.015813	-0.162843	-0.074617	0.115948	-0.081879
ONEOFF_PURCHASES_FREQUENCY	0.214702	0.362208	0.163222	0.036303	-0.051279	-0.097299
PURCHASES_INSTALLMENTS_FREQUENCY	0.295451	-0.112002	-0.330029	0.023502	0.025871	0.006731
CASH_ADVANCE_FREQUENCY	-0.214336	0.286074	-0.278586	0.096353	0.360132	0.066589
CASH_ADVANCE_TRX	-0.229393	0.291556	-0.285089	0.103484	0.332753	0.082307
PURCHASES_TRX	0.355503	0.106625	-0.102743	-0.054296	0.104971	-0.009402
Monthly_avg_purchase	0.345992	0.141635	0.023986	-0.079373	0.194147	0.015878
Monthly_cash_advance	-0.243861	0.264318	-0.257427	0.135292	0.268026	0.058258
limit_usage	-0.146302	0.235710	-0.251278	-0.431682	-0.181885	0.024298
payment_minpay	0.119632	0.021328	0.136357	0.591561	0.215446	-0.572467
both_oneoff_installment	0.241392	0.273676	-0.131935	0.254710	-0.340849	0.294708
istallment	0.082209	-0.443375	-0.208683	-0.190829	0.353821	-0.086087
none	-0.310283	-0.005214	-0.096911	0.245104	-0.342222	-0.176809
one_off	-0.042138	0.167737	0.472749	-0.338549	0.362585	-0.060698

- FACTOR ANALYSIS :-

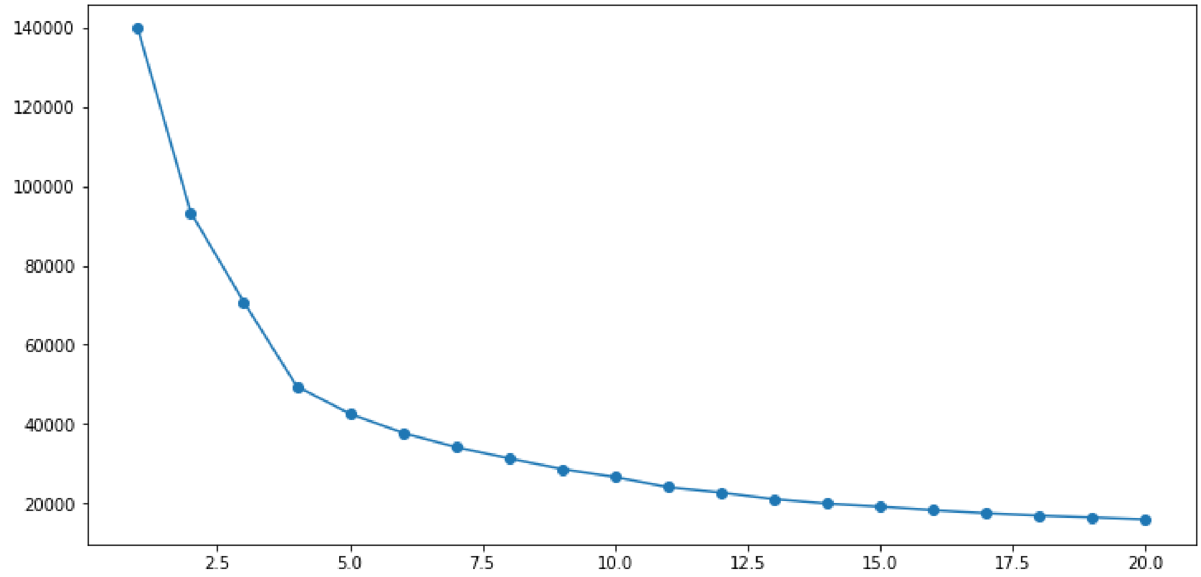
```
PC_0    0.402058
PC_1    0.180586
PC_2    0.147294
PC_3    0.081606
PC_4    0.065511
PC_5    0.041594
```

Cluster Analysis :-

- Based on the intuition on type of purchases made by customers and their distinctive behavior exhibited based on the purchase_type (as visualized above in Insights from KPI) , I am starting with 4 clusters and I came up with result :-

3	2769
2	2224
1	2088
0	1869

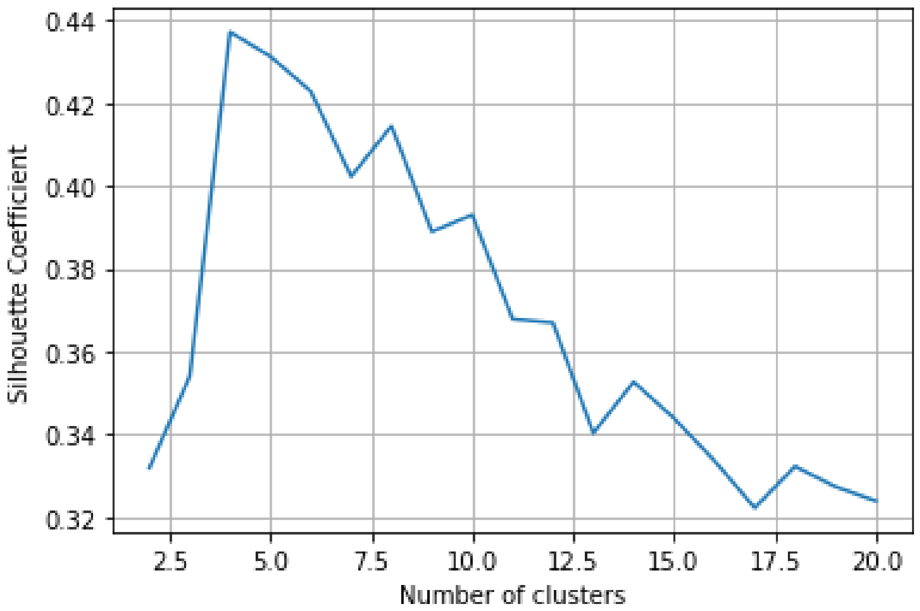
- **Identifying cluster Error :-**

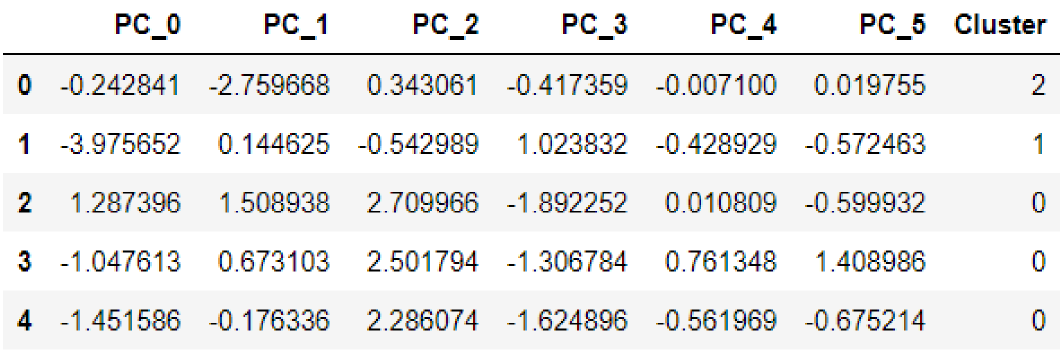


- **Silhouette Coefficient :-**

- It is a method of interpretation and validation of consistency within clusters of data. The technique provides a succinct graphical representation of how well each object has been classified. The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters.

- In my data set I used Silhouette Coefficient to classified the data and to see how how similar an object is to its own cluster compared to other clusters. And I came up with the following result in graphical representation.





	PC_0	PC_1	PC_2	PC_3	PC_4	PC_5	Cluster
0	-0.242841	-2.759668	0.343061	-0.417359	-0.007100	0.019755	2
1	-3.975652	0.144625	-0.542989	1.023832	-0.428929	-0.572463	1
2	1.287396	1.508938	2.709966	-1.892252	0.010809	-0.599932	0
3	-1.047613	0.673103	2.501794	-1.306784	0.761348	1.408986	0
4	-1.451586	-0.176336	2.286074	-1.624896	-0.561969	-0.675214	0

Conclusion & Suggestion After lots of Analysis :-

- I divided the data set into four groups and accordingly I suggest each group uses from the dataset.

Group 2 :-

- They are potential target customers who are paying dues and doing purchases and maintaining comparatively good credit score
- We can increase credit limit or can lower down interest rate
- Can be given premium card /loyality cards to increase transactions

Group 1 :-

- They have poor credit score and taking only cash on advance. We can target them by providing less interest rate on purchase transaction.

Group 0 :-

- This group is has minimum paying ratio and using card for just oneoff transactions (may be for utility bills only). This group seems to be risky group.

Group 3 :-

- This group is performing best among all as customers are maintaining good credit score and paying dues on time.
- Giving rewards point will make them perform more purchases.

>>>>>>>>>>>>>>>>>>>> End <<<<<<<<<<<<<<<<<<<<<<<