Credit Card Segmentation

By, Sandeep A C

INTRODUCTION

Customer segmentation is the process of dividing customers into groups based on common characteristics so companies can market to each group effectively and appropriately.

Why Segment Customers?

Segmentation allows marketers to better tailor their marketing efforts to various audience subsets. Those efforts can relate to both communications and product development.

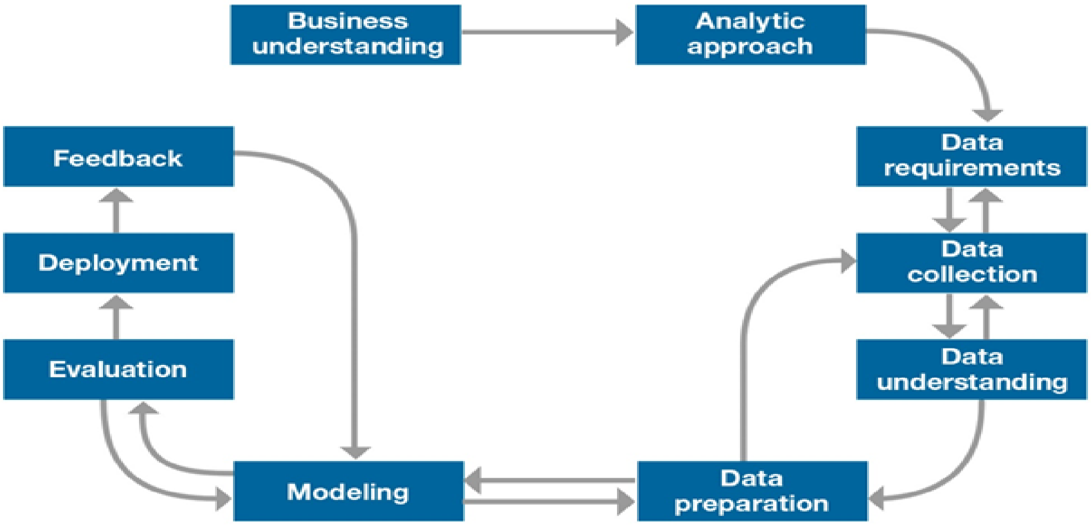
PROBLEM STATEMENT

This case 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.

SYSTEM REQUIREMENTS

Hardware Requirements: Micro Processor: Intel processor Main memory/RAM: 4GB Hard Disk: 40 GB HHD/SSD
Software Requirements: Operating system: windows, mac, Linux Framework: Anaconda, Jupyter Notebook Programming Language: Python 3

METHODOLOGY



Deriving KPI's and Extracting Insights from them.

As asked in the problem statement, we have to extract different KPI's(Key Performing Issues)

i. Monthly Average Purchases: We can obtain that by dividing Purchases with Tenure.

ii. Monthly Cash Advanced Amount: We can obtain that by dividing cash Advance with Tenure.

iii. Purchase Type: By observing the data, we can infer that there are two types of purchases. ONE_OFF_PURCHASES and INSTALLMENT_PURCHASES.

By Exploring more deeply, we can gain 4 meaningful insights,
There are FOUR types of Purchase behaviour in the data.
1. People who does not make any purchases.
2. People who make both types of purchases.
3. People who make only OneOff_Purchases
4. People who make only Installment_Purchases

We can also extract how many customers belong to each type of Purchase Behaviour.

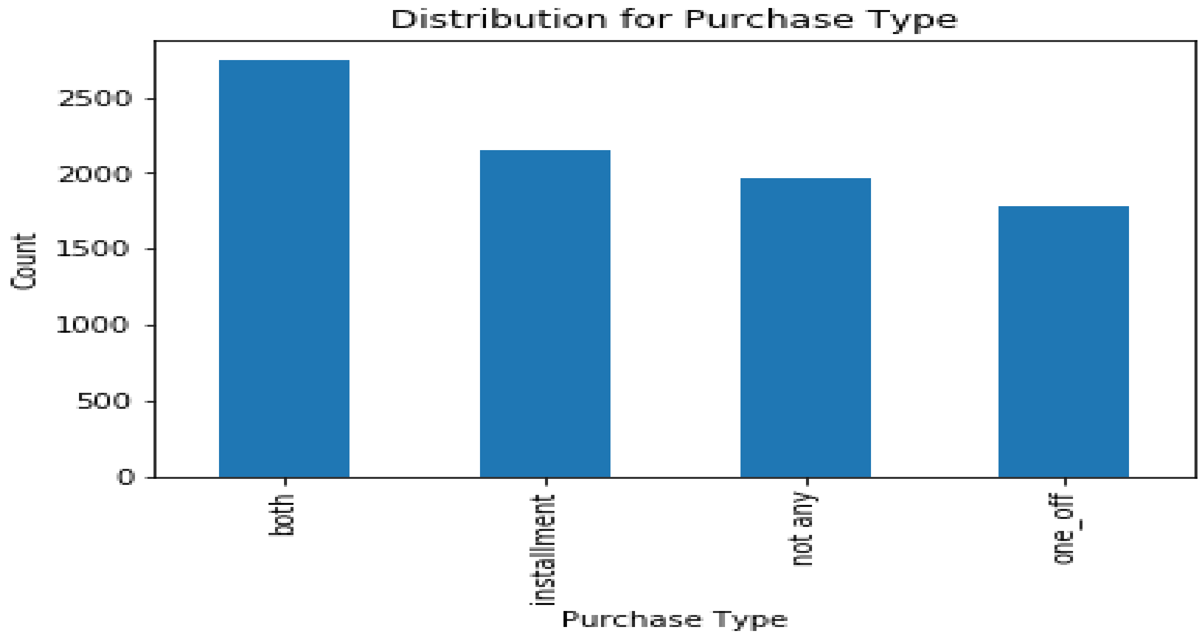


Fig: Distributions for Purchase Type

iv. Limit Usage: We can obtain it by dividing Balance with Credit limit.

v. Payments to Minimum payments Ratio: It is obtained by dividing Payments with minimum Payments.

Insights from derived KPI’s

i. Plotting Bar Graph for Average Payment to minimum Payment Ratio VS Purchase Type

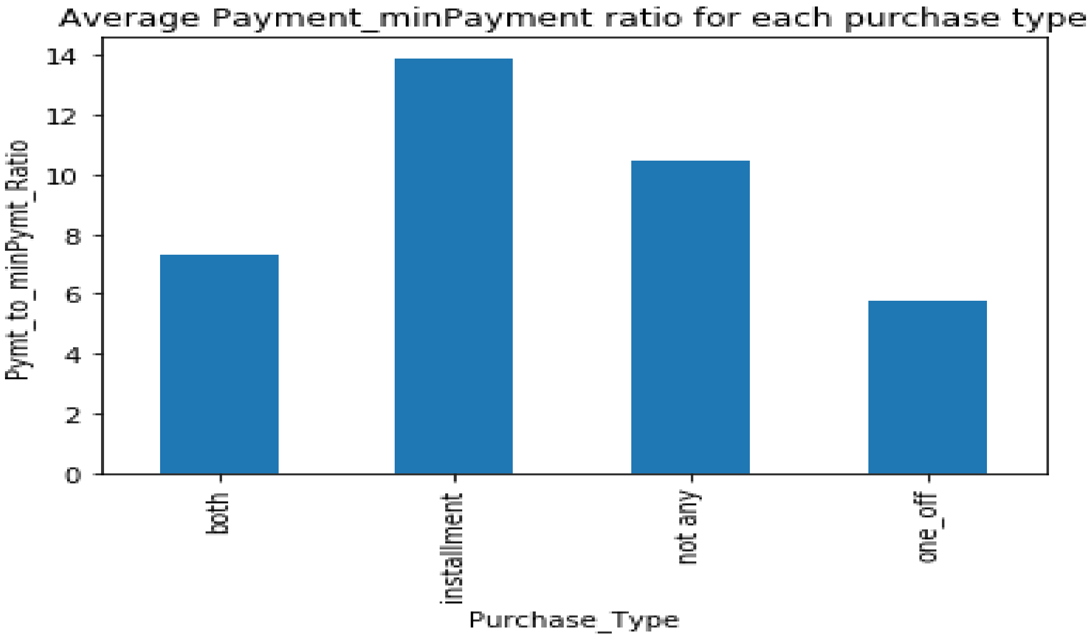


Fig: Average Payment to minimum Payment VS Purchase Type in Python

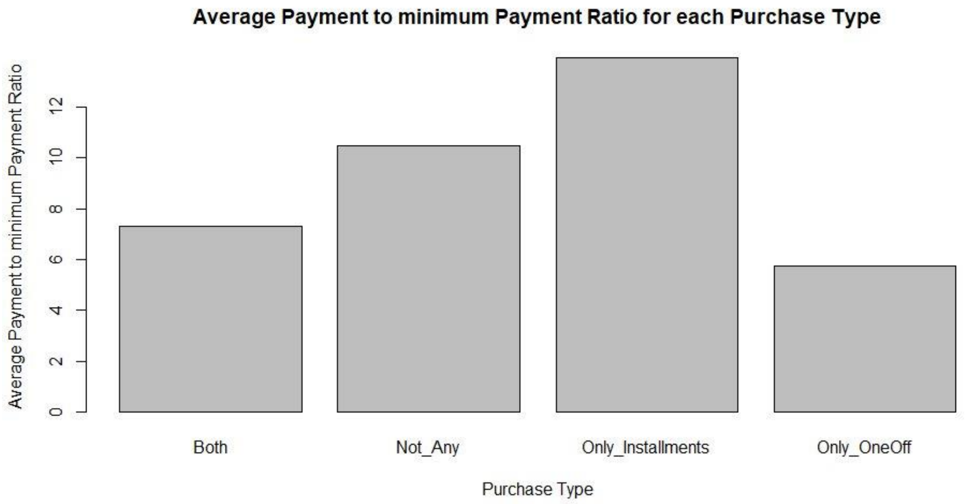


Fig: Average Payment to minimum Payment VS Purchase Type in R

Insight 1: Clients with Purchase Type Instalment are with more dues.

ii. Plotting Bar Graph for Average Limit Usage VS Purchase Type

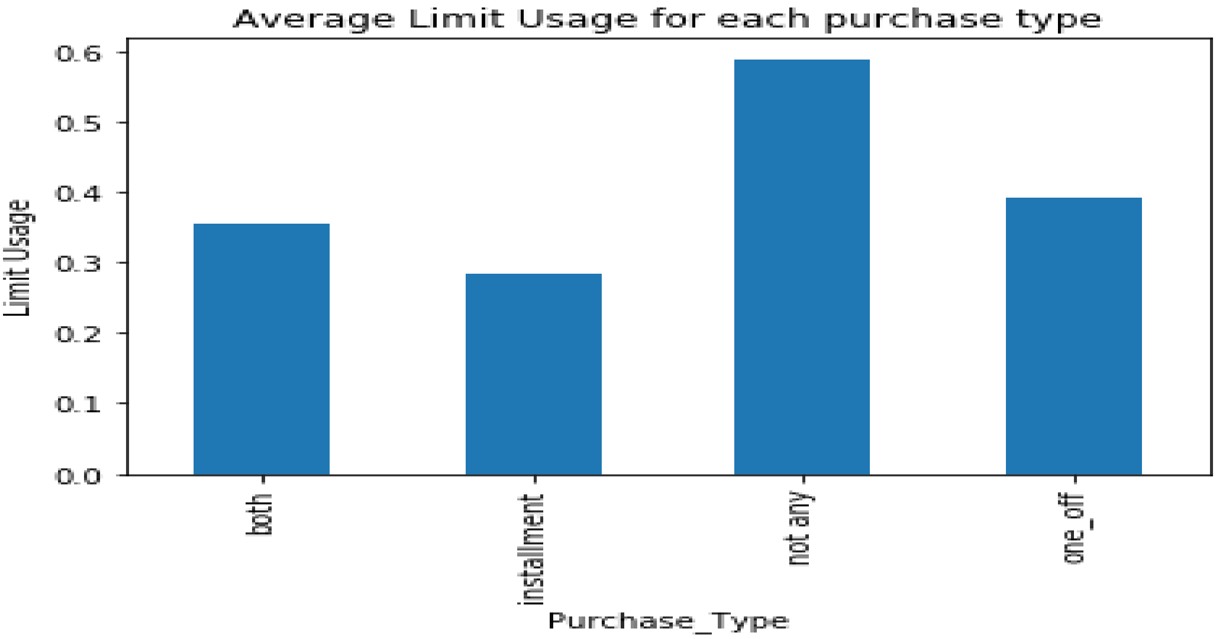


Fig: Average Limit Usage VS Purchase Type in Python

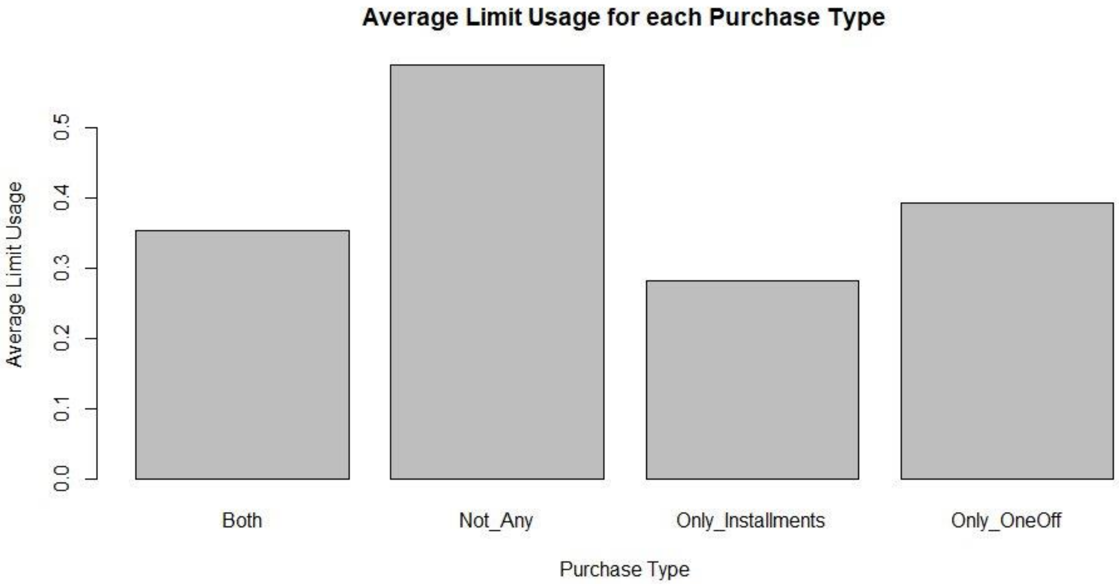


Fig: Average Limit Usage VS Purchase Type in R

Insight 2: Clients with Purchase type Instalment have good Credit Score

iii. Plotting Bar Graph for Average Monthly Cash Advance VS Purchase Type

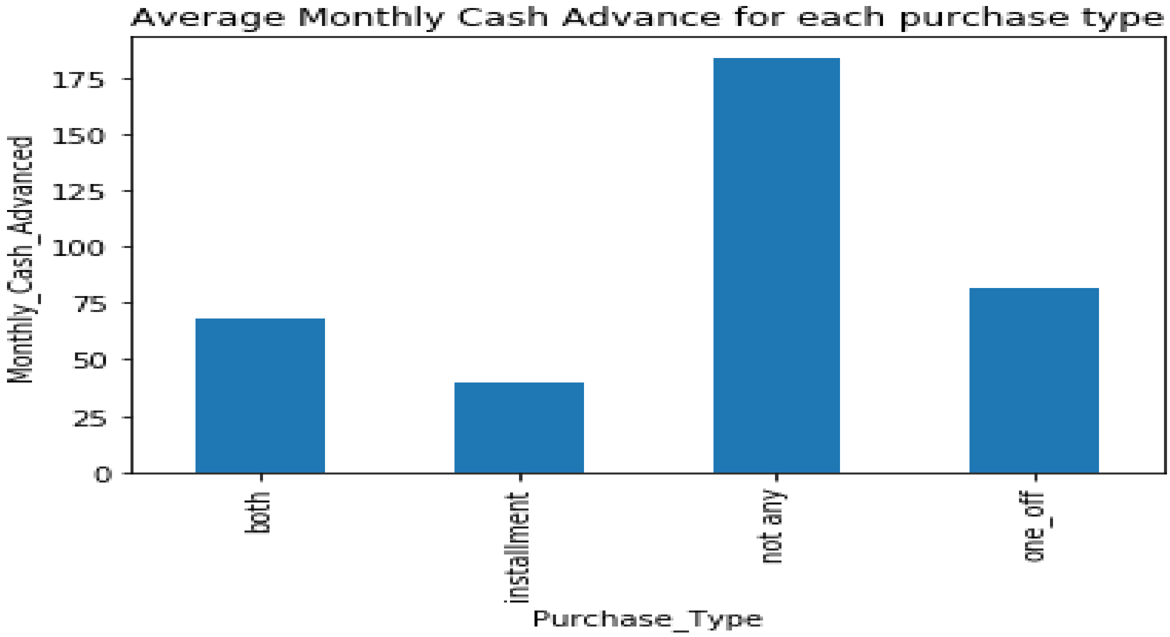


Fig: Average Monthly Cash Advance VS Purchase Type in Python

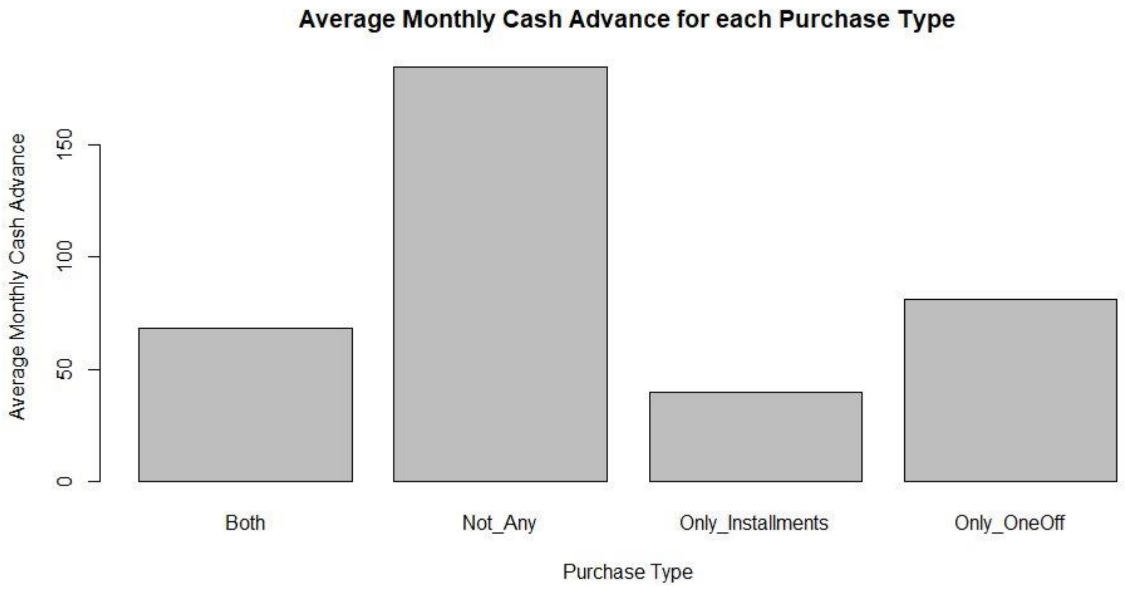


Fig: Average Monthly Cash Advance VS Purchase Type in R

Insight 3: Clients who don’t do any type of Purchases (One-off, Instalment), Take more Cash on Advance.

DESIGN AND IMPLEMENTATION

1. Collect the Data

Typical information-gathering methods include:

- Face-to-face or telephone interviews
- Surveys
- General research using published information about market categories
- Focus groups

In our case the dataset is already provided.

2. Data Pre-processing

It includes many phases.

i. Missing Value Analysis

Missing data can occur because of nonresponse, no information is provided for one or more items or for a whole unit. Sometimes missing values are caused by the researcher. For example, when the data collection is done improperly or mistakes are made in data entry, it is called Human Error.

There are different ways to deal with missing values in the data.

- **Imputation**

We can impute the missing values with its mean, median or mode.

There is also another method called KNN imputation, which is available in both R and Python.

- **Partial deletion**

When the missing values in the data are less, we can delete them since they do not affect the model accuracy as they are in less number.

In our case, we are doing partial deletion.

ii. Outlier Analysis

An **Outlier** is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining.

A simple way to way to detect outlier is to use boxplot.

For our data, the below images show the box plot, which are drawn in both R and Python.

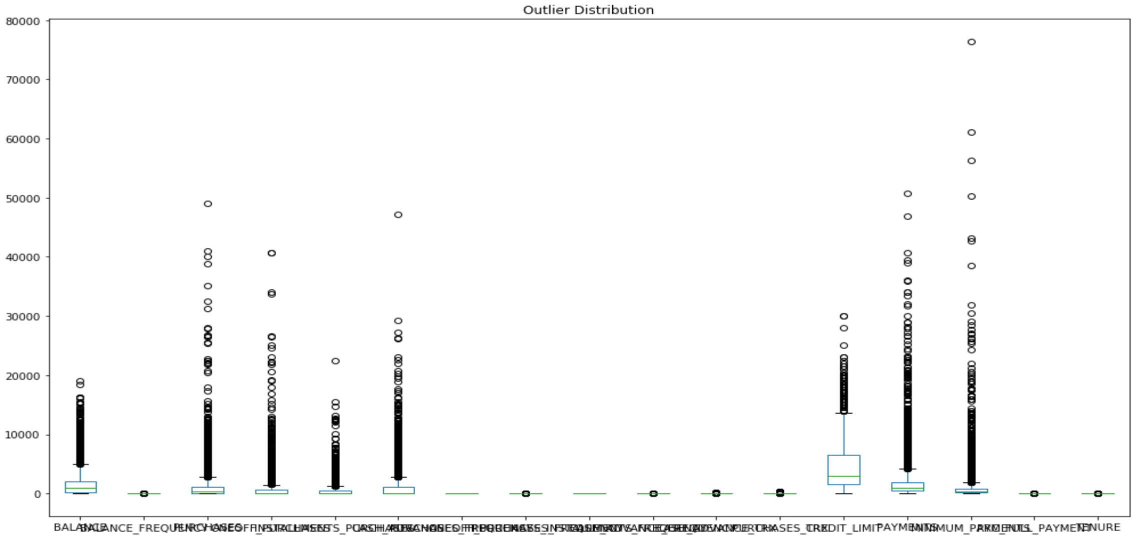


Fig: Box-Plot in Python

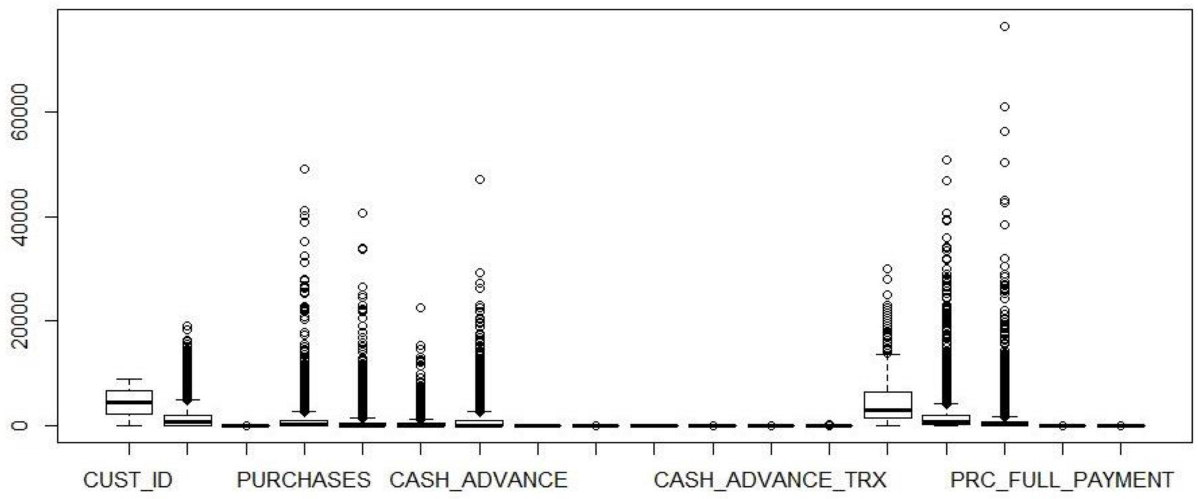


Fig: Box-Plot in R

If there is too much of data in outlier, then deleting them is not an efficient way, because we will lose lot of data.

iii. Feature Selection

For modelling, we have to select the columns or independent variables that will highly contribute to the dependent variable.

If there is no dependent variable, we have to find the collinearity between all the variables and if there is high collinearity between two variables, we can remove any one variable and keep the other. For Example, in our case, PURCHASES_FREQUENCY and PURCHASES_INSTALLMENTS_FREQUENCY have high correlation. Hence, we remove PURCHASES_INSTALLMENTS_FREQUENCY.

iv. Feature Scaling

Feature Scaling is a technique to standardize the independent **features** present in the data in a fixed range. There are different methods used in feature scaling,

- Rescaling (min-max normalization)
- Mean Normalization
- Standardization (Z-score Normalization)
- Scaling to Unit etc.

Standardization has to be applied when the data is normally/uniformly distributed. In our case, the data is not uniformly distributed. Therefore, we Normalize the data.

3. Clustering or Segmentation or Model Creation

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.

There are different clustering algorithms. Here we use **K-Means** clustering.

K-means algorithm is an iterative algorithm that tries to partition the dataset into **K** pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group**.** It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different as possible.

The way K-means algorithm works is as follows:

1. Specify number of clusters *K*.
2. Initialize centroids by first shuffling the dataset and then randomly selecting *K* data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.
4. Compute the sum of the squared distance between data points and all centroids.
5. Assign each data point to the closest cluster (centroid).
6. Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

- **Evaluation Method:**

We use **Elbow method** to Evaluate the number of clusters(K) to used.
Elbow method gives us an idea on what a good K number of clusters would be based on the sum of squared errors (SSE) between data points and their assigned clusters centroids.

The below images show the elbow graph for our data in Python.

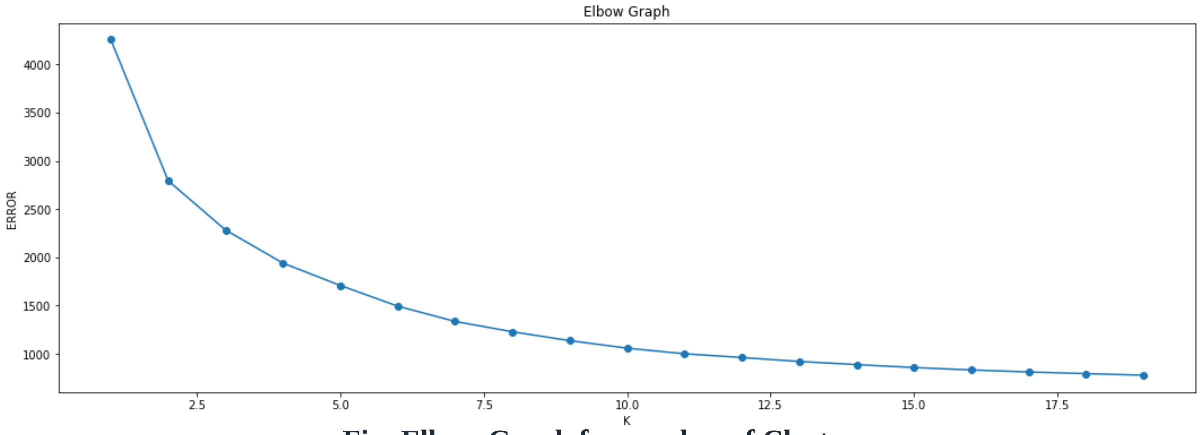


Fig: Elbow Graph for number of Clusters

From the graph, we can select 4 as the optimum number of clusters because, either 4 or 20 clusters, the error difference is very less. But when compared to 2 and 20 clusters, the error difference is more.

Finally, for the cluster of 4, our data is split as below.

Cluster Number	Number of Clients/Customers
0	3900
1	974
2	2487
3	1275

Table: Cluster VS Number of Clients

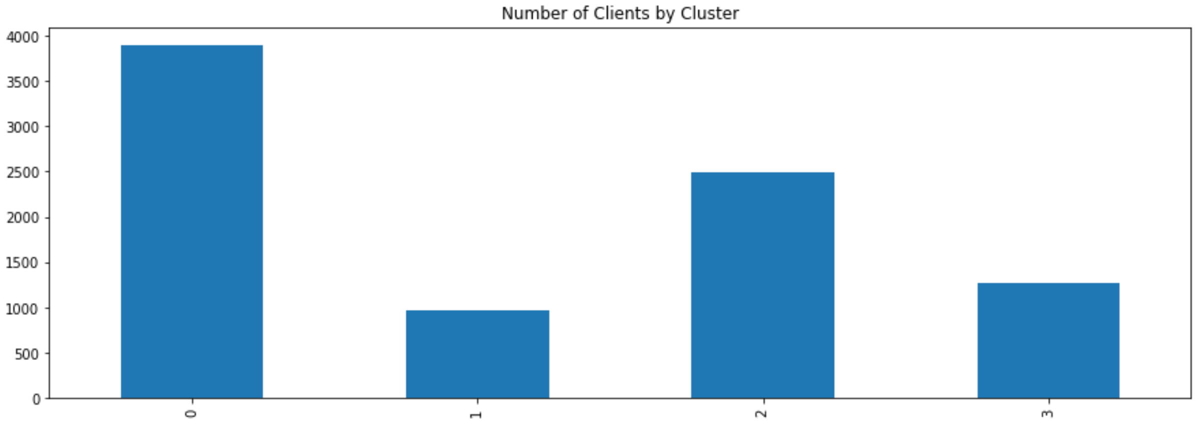


Fig: Clusters VS Number of Clients

In machine learning algorithms, the output will be different each time when we run the code and same cluster won’t have same number of count each time.
Finally, from the clusters, we can gain insights and derive suggested marketing strategies.

Insights and Suggested Marketing Strategies from Clusters

Cluster 0: (Count: 3900) **Insight:** These customers have maximum Average Purchase and good Monthly cash advance but this cluster doesn't do frequent instalment or one-off purchases. **Marketing Strategy:** They are potential target customers who are paying dues and doing purchases. We can increase credit limit or can lower down interest rate, can be given premium card or loyalty cards to increase transactions
Cluster 1: (Count: 974) **Insight:** They do very less One-off purchases and less cash advance and maintain very less Balance. **Marketing Strategy:** We can target them by giving them more offers.
Cluster 2: (Count: 2487) **Insight:** This group of customers who have highest Monthly cash advance and doing both instalment as well as one-off purchases, have comparatively good credit score but have poor average purchase score. **Marketing Strategy:** They take only cash on advance. We can target them by providing less interest rate on purchase transaction.
Cluster 3: (Count: 1275) **Insight:** Customers are doing maximum One-off transactions and has least payment ratio amongst all the cluster. **Marketing Strategy:** This group is having minimum paying ratio and using card for just one-off transactions (may be for utility bills only). This group seems to be risky group.