Credit Card Segmentation – Data Science Project

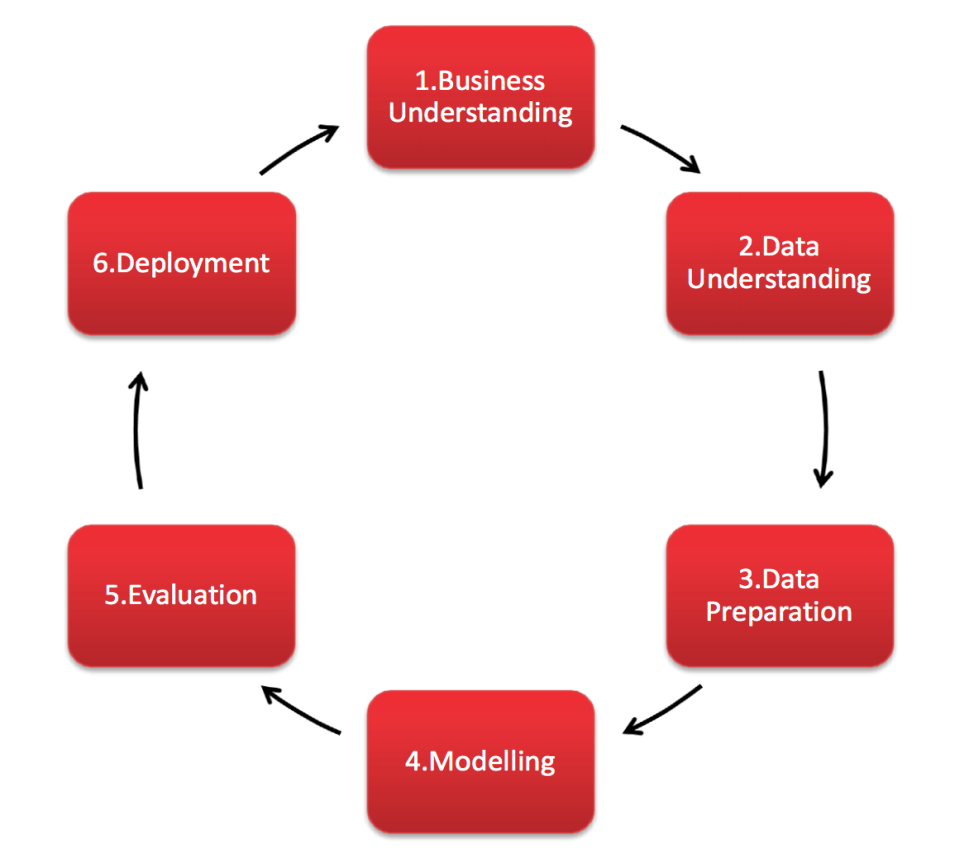
Authored by: Rishi Mukunthan T

Abstract:In this project, for the given dataset with credit card usage data of 8950 customers of a credit card company, customer segmentation was performed using unsupervised ML algorithm (K-Means Clustering). For the clusters identified, the features were analyzed to assign meaningful label to the segments of the customers and further marketing strategies were prepared for each group based on their features (customer behavior).

# Project Plan:

CRISP-DM (Cross Industry Standard Process for Data Mining) methodology was adapted to tackle this Customer Segmentation Project.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment



**Business Understanding:**

Domain knowledge is crucial for tackling any data science project. The objective here is to develop a customer segmentation model to define a unique marketing strategy for each segment identified.

The dataset consists of various features of credit card users without a target label. I have a credit line and am fairly familiar with credit card usage. Still, I browsed the internet to know the basics of the domain which will cover the Business Understanding process of the CRISP-DM process.

**Data Understanding:**

Data Understanding part will include understanding the meaning of each feature provided in the dataset which can be accomplished by the Data Dictionary already provided in the requirements document.

# EDA (Exploratory Data Analysis) will help understand the features and their inter-relations if any in extreme detail.

**Data Preparation:**

* Missing value removal/imputation
* Feature Selection, Feature Extraction and Dimensionality Reduction
* Feature Scaling

**Modelling**:

Since we are provided with unlabelled data, the problem statement comes under unsupervised learning or clustering. I will use K-Mean Clustering algorithm to accomplish this. Identified clusters will be labeled by human intervention (semi-supervised) based on the features exhibited by each cluster.

Additional Work done: Once we segment the customers, we can create our own label with the segmentation achieved and train supervised learning classification models on the labeled data generated from unsupervised learning process. Future customer data can be classified in to any of the segment using the supervised model.

**Evaluation:**

Evaluation of unsupervised learning model is fairly difficult since we don't have any labels. WCSS (Within Cluster Sum of Squares) is a good evaluating measure for relatively determining which no of cluster split is maximizing the variance between clusters and maximizing the similarity within cluster.

**Deployment:**

R scripts and .py scripts can be run from the user system to get the results of the project. Additionally, if time permits, we can develop and deploy a flask api using heroku to get future customer data as input and respond with the segment, the customers belong to, as output, so that business user can leverage this api so easily directly from web.

# Exploratory Data Analysis (Data Understanding):

As part of Business/Data understanding, I am creating an own data dictionary with my understanding. In real life scenario, it is better to reach out to business/client to gain better understanding of the features if there are less than 50 features. In some medical/genetic datasets, there could be millions of features, where data understanding is not completely feasible.

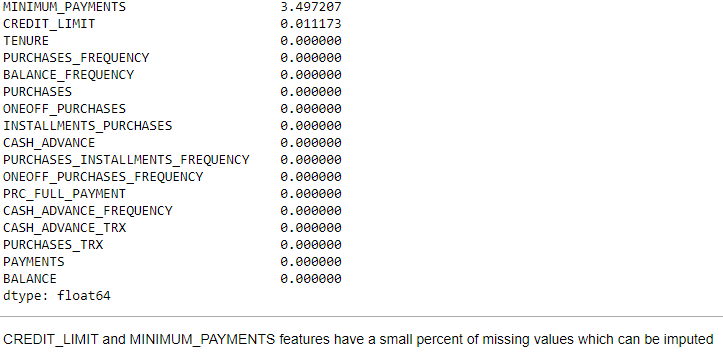
1. CUST\_ID Credit card holder ID - doesn't contain any valuable/meaningful insight so can be dropped.
2. BALANCE Monthly average balance (based on daily balance averages) - Balance is the amount you owe to your credit card company. Making a purchase increases balance while paying dues decreases balances.
3. BALANCE\_FREQUENCY Ratio of last 12 months with balance. Percentage of months with balance/due.
4. PURCHASES Total purchase amount spent during last 12 months
5. ONEOFF\_PURCHASES Total amount of one-off purchases
6. INSTALLMENTS\_PURCHASES Total amount of installment purchases
7. CASH\_ADVANCE Total cash-advance amount
8. PURCHASES\_FREQUENCY-Frequency of purchases (percentage of months with at least one purchase)
9. ONEOFF\_PURCHASES\_FREQUENCY Frequency of one-off-purchases
10. PURCHASES\_INSTALLMENTS\_FREQUENCY Frequency of installment purchases
11. CASH\_ADVANCE\_FREQUENCY Cash-Advance frequency
12. CASH\_ADVANCE\_TRX Average amount per cash-advance transaction
13. PURCHASES\_TRX Average amount per purchase transaction
14. CREDIT\_LIMIT Credit limit
15. PAYMENTS-Total payments (due amount paid by the customer to decrease their statement balance) in the period
16. MINIMUM\_PAYMENTS Total minimum payments due in the period.
17. PRC\_FULL\_PAYMENT- Percentage of months with full payment of the due statement balance
18. TENURE Number of months as a customer

The dataset contains 8950 rows, i.e., customer data, and 18 columns, i.e., features.

We will analyze the dataset to identify:

* Missing values
* Distribution of the numerical variables
* Outliers

**Missing Values:**

****

There are 2 ways to deal with missing values.

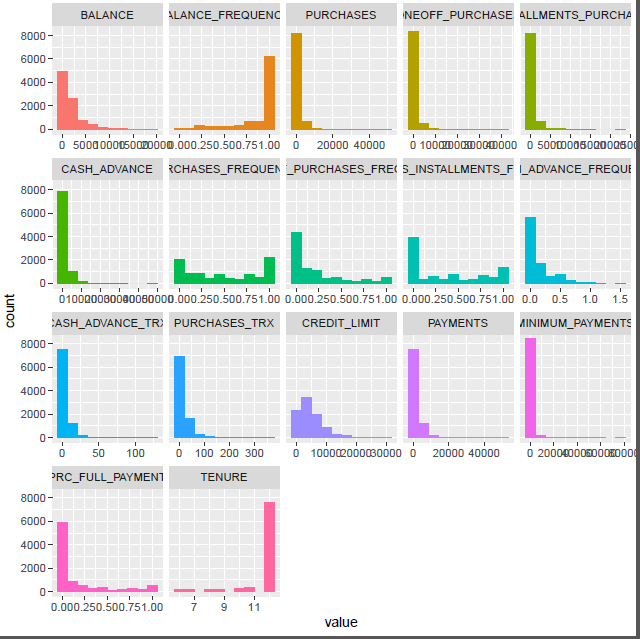
1. Imputing missing values with any central tendency measure or using predictive modelling algorithms like KNN to perform imputation.

2. Deleting either the observation or the feature which has missing values.

Option 2 won’t be suitable for this project because ideally we don’t want to drop customer observations just because they have a missing value in one column. Also since both columns which have missing values have less than 5%, dropping the entire feature is not advised. If in case, a particular features had more than 25 percent missing values then the feature should be dropped ideally.

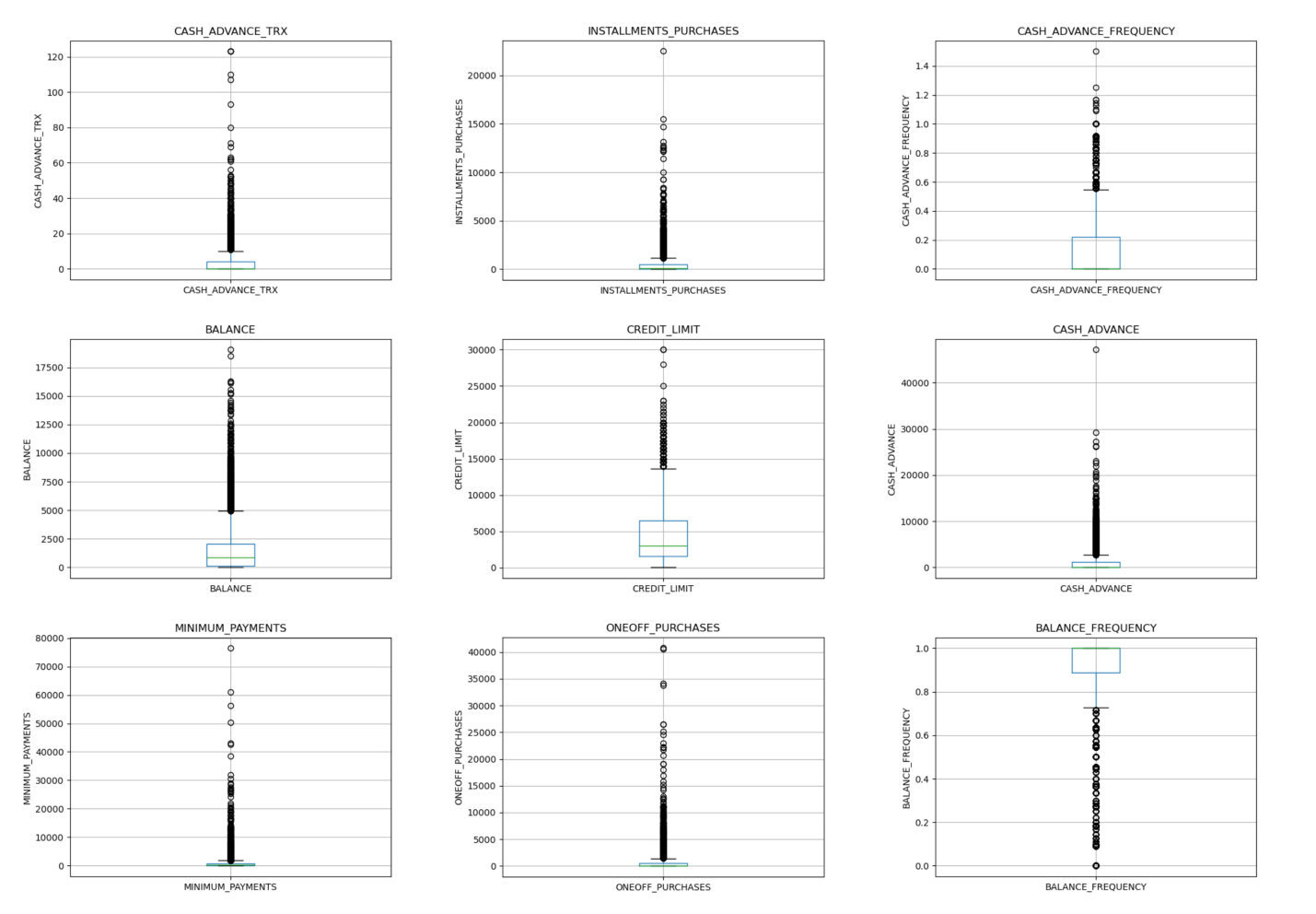
Since objective is to devise marketing strategy for each customer based on the cluster they fall in. So, we will choose option to impute without dropping any observation.

**Distribution of Numerical features:**



On analyzing the histograms of frequency distribution of all features, almost no variable seems to have a normal distribution so log transforming variables could improve performance and in reducing the skewness.

**Outlier Analysis:**

****

Each numerical variable were analyzed using box plots which revealed the presence of outliers in most variables. Removing the outlier observation data points would greatly reduce the number of observations. Also, outlier information could be meaningful and since we are not doing supervised learning and the main objective is to segment the given customers, it is not advisable to just drop customers. Log transformation can be performed to mitigate the effect of outlier (Treatment of outliers).

# Data Preprocessing/Preparation/Cleaning:

Data preprocessing steps include the following:

1. Missing value handling.
2. Outlier Treatment.
3. Feature Extraction and Selection
4. Feature Scaling

**Missing Value Handling:**

From EDA, we found that CREDIT\_LIMIT and MINIMUM\_PAYMENT features have missing values.

Imputing these missing values with mean or median will be problematic since both variables don't follow a normal distribution and are heavily skewed. For example, the imputed minimum\_payment with mean/median could in some cases produce meaningless values when compared with values of other features like balance, payments. Example, if you have high MAB (Monthly Average Balance) and no payments made then in span of 12 months the minimum payments due would have grown exponentially with the interest rate and could be high but mean/median wouldn't capture these. So, KNN imputation will be the best fit since it predicts the missing values with the approximation from the data points' nearest neighbors instead of generalizing a same value for all missing values.

**Feature Engineering - Deriving new KPIs:**

1. Monthly Average Purchase (MAP) = Total Purchases/Tenure --> Regular purchasers
2. Monthly Average Cash Advance (MA\_CA) = Total Cash Advance/Tenure --> Regular withdrawers
3. Average Amount per purchase (PURCHASES\_TRX) --> High value purchasers (Luxurious) or Low value purchasers (Economic)
4. Customer Purchase Behavior as Categorical Value by Purchase type. Categories - One-off, Installments, Both, None
5. Limit Usage (balance to credit limit ratio) --> High Credit Utilization lead to lower credit score.
6. Payments to minimum payments ratio --> Higher the value, more desirable/low-risk customer.
7. PRC\_FULL\_PAYMENT - Percentage of months with full payment of the due statement balance.

Below are the desirable KPI values expected in a customer profile in an ideal case scenario - This may not be completely applicable in some contexts but will prove good as an initial generalization.

* High MAP
* Low MA\_CA
* High average amount per purchase(not necessarily - depends)
* Both One\_Off and Installment purchase type
* Optimum(<30 percent) Limit\_Usage
* High Pay\_MinPay Ratio
* Reasonably high Full Payment - Too low might indicate financially unstable customer.

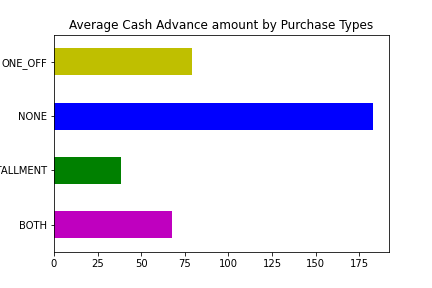
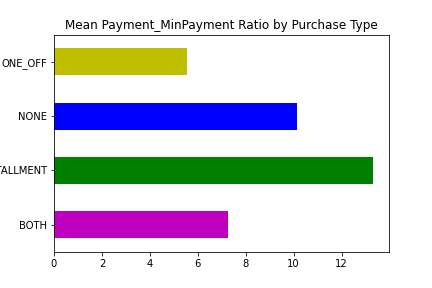
**Customer Profiling - Insights from KPIs**:

Building customer profiles based on the Purchase type. Basic analysis is done by creating customer profiles based on purchase types.



Customers who do Installment Purchases are highly likely to have greater payment to minimum payment ratio whereas who do one off purchases have the least payment to minimum payment ratio which is not a desirable quality in a customer because eventually these late payments or overdue payments maybe defaulted.

Installment payments usually reduce the payment burden on the customer by splitting huge payments to easy emi installment payments with reasonable interest margin. So, encouraging one-off purchasers to convert their purchases in to emi payments would be a win-win situation.

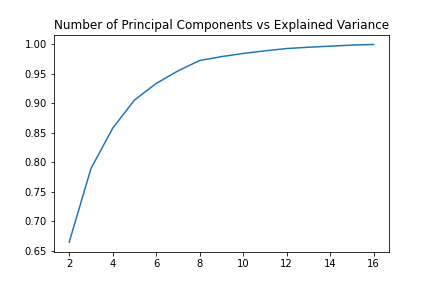
By the below graph, we could obviously see those who don’t make purchases, use their limits by taking cash advance. This behavior would differentiate them and further analysis on this group’s features should be analyzed.

**Principal Component Analysis:**

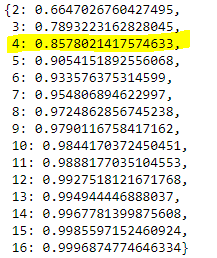
In high-dimensional space, finding the directions of maximum variance and we project it in to a smaller dimensional subspace preserving most variance in the data.

Example: Say Feature1(x axis) = (2, 2, 2, 2, 3, 2) and Feature2(y axis) = (2, 10, 5, 8, 16, 20). If you plot a trend line for this data points in 2d graph, you will get a vertical straight line parallel to y axis with x = 2. This vertical line covers almost most variance in the data, so we can say this as a new principal component axis (PC1). Now we have converted 2 features to 1 which is the dimensionality reduction.

We apply PCA for dimensionality reduction before doing clustering.

In the above graph, the x axis is the no of principal components and y axis percentage of variance in the data explained. Rule of thumb in dimensionality reduction is to preserve at least 80 percent variance.

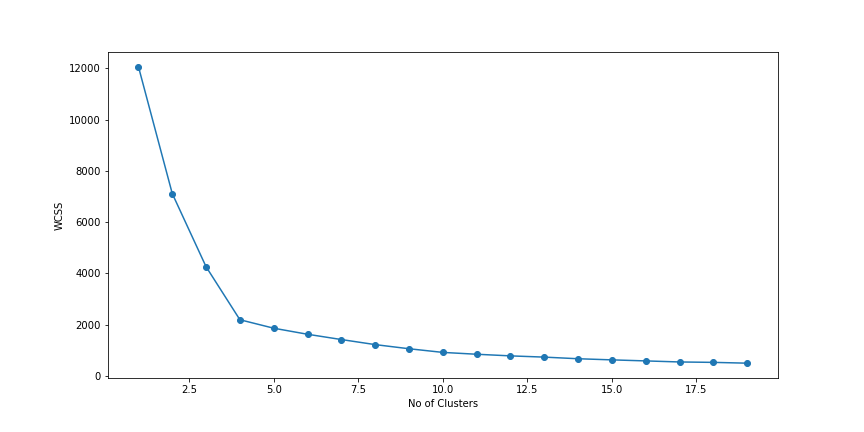
So, we have selected the no of principal components as 4 since it preserves around 85 percent variance.



# Modelling: Clustering

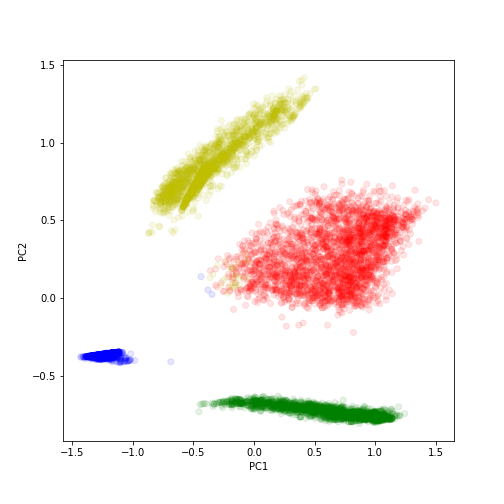
We will be applying the widely popular K-Means clustering of the prepared final data to identify clusters.

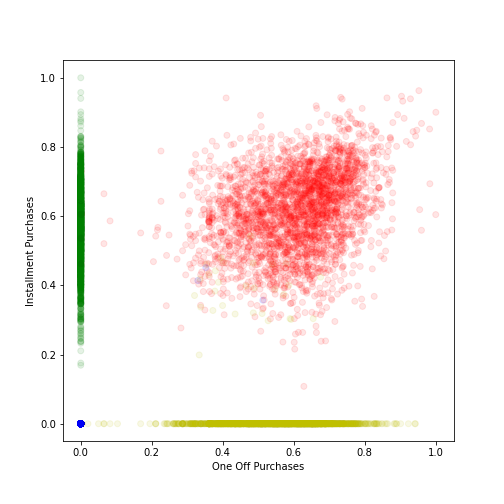
For performing K Means clustering algorithm, we must first choose the no of clusters we want. We decide on this by plotting the no of clusters vs WCSS (Within Cluster Sum of Squares or cluster errors). We then use the elbow method and our intuition from customer profiling to note that the ideal no of clusters would be 4. We can clearly visualize the elbow at 4 clusters.



**Results of K-Means Clustering:**

Plotting the observations taking PC1 and PC2, we can see the separation between clusters is so clear.

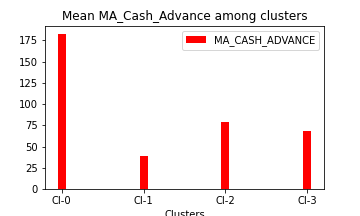




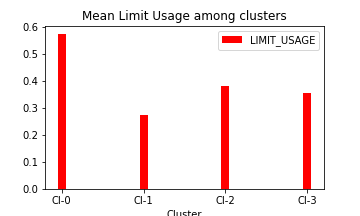


The above shows the count of observations tagged to each cluster and they are almost balanced. There are around 20 to 30 percent customers tagged to each cluster.

**Plotting the KPIs for each cluster:**



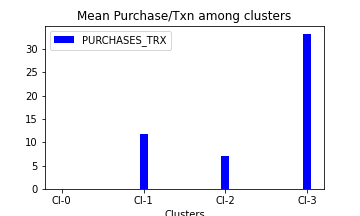
**Inference – C0 – Taking high monthly average cash advance.**



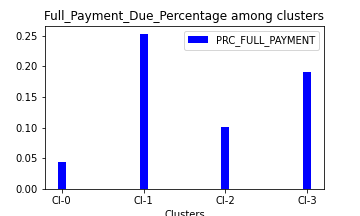
**Inference – C0 – Has high credit utilization/limit usage which leads to lower credit score.**



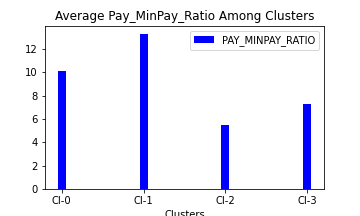
**Inference – C0 – Has almost no purchases whereas C3 is characterized by the highest monthly average purchase value.**

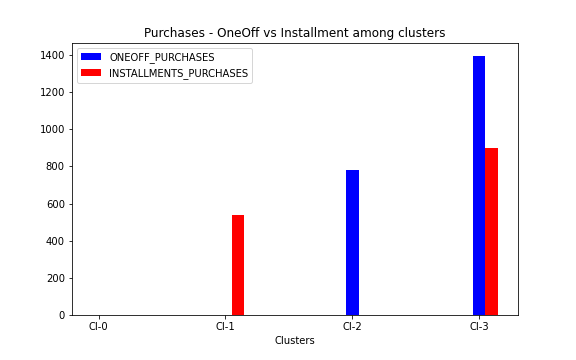


**Inference – C0, C3 has expected values but it is interesting to note that C2 has slightly higher mean monthly average purchase than C1 but the average purchase amount per transaction is higher for C1. So, C2 may be buying low cost/economic products with high volume whereas C1 is buying luxury/high cost purchases but the frequency is low.**

****

**Inference – C1 regularly pay full payment due on time whereas C0 may be characterized by paying only the minimum due and not paying due in full.**

**Inference –** **Even though C0 has low full payment due paying habit, in the tenure of 12 months, they have a reasonable payment to minimum payment ratio. C1 has the highest pay\_minpay\_ratio.**

* Here we can see clearly that C0 is characterized by no purchases whereas C1 identified with mostly Installment\_Purchases, C2 by mostly one\_off purchases and C3 by both type of purchases.

# Conclusion/Inferences:

Regular Withdrawers - (High Monthly Average Cash Advance)

Low Credit Scores - (High Credit Utilization)

* C0 - Regular Withdrawers, Non-Purchaser, Non-Full\_Due Payer, Low Credit Scorers
* C1 - Occasional Purchasers, Installment Purchasers, Prompt Full-Due Payers
* C2 - Occasional Purchasers, One-off Purchasers, Low Pay\_MinPay Ratio
* C3 - Regular Purchasers, Both Purchasers, Mostly Full Due Payers

To summarize we are labeling C0 – Withdrawers, C1 - Installment\_Purchasers, C2 - One\_Off\_Purchasers and C3 - Big\_Spenders

**Marketing Strategies:**

* C0 - This group of customers seem to be utilizing the card mostly for taking cash advance and they are not doing purchases with the card. So providing cash back offers for first purchase could convert them gradually to purchasers. Since they seem to already have high credit utilization, we should be cautious and refrain from offering any credit limit increase offers. Deals like providing low interest rate for installment purchases could also be beneficial.
* C1 - This group consists of moderate spenders who purchase on installments and pay their dues in full mostly. We should focus our marketing strategies on retaining these customers and maybe trying to increase their purchases by providing customized deals based on their spending pattern like movies, food deals etc.
* C2 - This group also consists of moderate spenders but with one major difference from the previous cluster in the aspect of purchase type. This group mostly does one-off purchases. Providing low interest rates for emi/installation purchases would encourage them to purchase on installments.
* C3 - This group consists of the "Big Spenders" who purchase in both one-off and installments. They also mostly pay their dues as full. We must be fully concerned to have the lowest churn rate in this group. So, making these customers loyal by providing reward points will be highly beneficial. Also, we can off credit limit increase offers to this group.