**Credit Card Cluster Analysis**

Andrew Thomas

Joshua Gaze

Phong Ong

Sophea Hummel

IST 707 Data Analytics

3/28/2022

**About the Dataset**

The dataset chosen for project work was the “Credit Card Dataset for Clustering” which was made publicly available on Kaggle in 2018. This dataset draws from a bank’s population of active credit card holders consisting of 8,950 records across 18 attributes. The variables listed below describe the data attributes as they originally appeared in the dataset:

|  |  |
| --- | --- |
| **Variable** | **Description** |
| CUST*ID* | Identification of Credit Card holder (Categorical) |
| BALANCE | Balance amount left in their account to make purchases |
| BALANCEFREQUENCY | How frequently the Balance is updated, score between 0 and 1 (1 = frequently updated, 0 = not frequently updated) |
| PURCHASES | Amount of purchases made from account |
| ONEOFF*PURCHASES* | Maximum purchase amount done in one-go |
| INSTALLMENTSPURCHASES | Amount of purchase done in installment |
| CASH*ADVANCE* | Cash in advance given by the user |
| PURCHASESFREQUENCY | How frequently the Purchases are being made, score between 0 and 1 (1 = frequently purchased, 0 = not frequently purchased) |
| ONEOFFPURCHASESFREQUENCY | How frequently Purchases are happening in one-go (1 = frequently purchased, 0 = not frequently purchased) |
| PURCHASESINSTALLMENTSFREQUENCY | How frequently purchases in installments are being done (1 = frequently done, 0 = not frequently done) |
| CASHADVANCEFREQUENCY | How frequently the cash in advance being paid |
| CASHADVANCETRX | Number of Transactions made with "Cash in Advanced" |
| PURCHASES*TRX* | Number of purchase transactions made |
| CREDITLIMIT | Limit of Credit Card for user |
| PAYMENTS | Amount of Payment done by user |
| MINIMUM\_PAYMENTS | Minimum amount of payments made by user |
| PRCFULLPAYMENT | Percent of full payment paid by user |
| TENURE | Tenure of credit card service for user |

**Data Mining Problem**

Credit card usage behaviors are the intended analysis findings for this dataset. With that in mind, the following questions have been identified based on our anticipated modeling and clustering analysis plans:

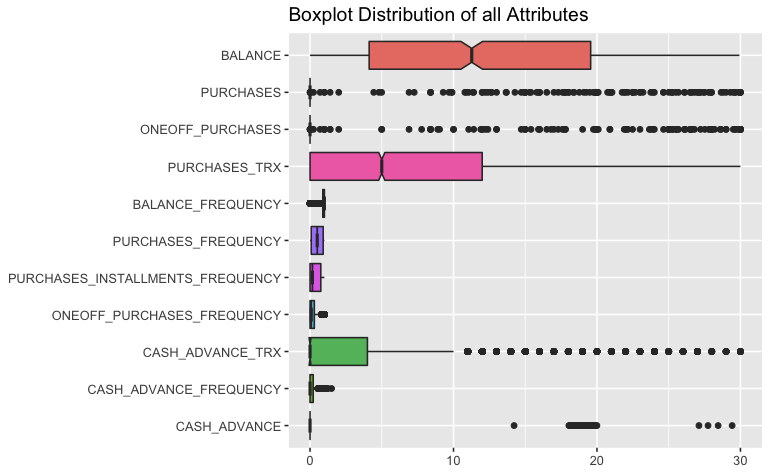
Using the attributes provided in the dataset, segment the credit card holders of the bank so that we can provide a better business insight to improve their overall satisfaction with the financial services the bank provides?

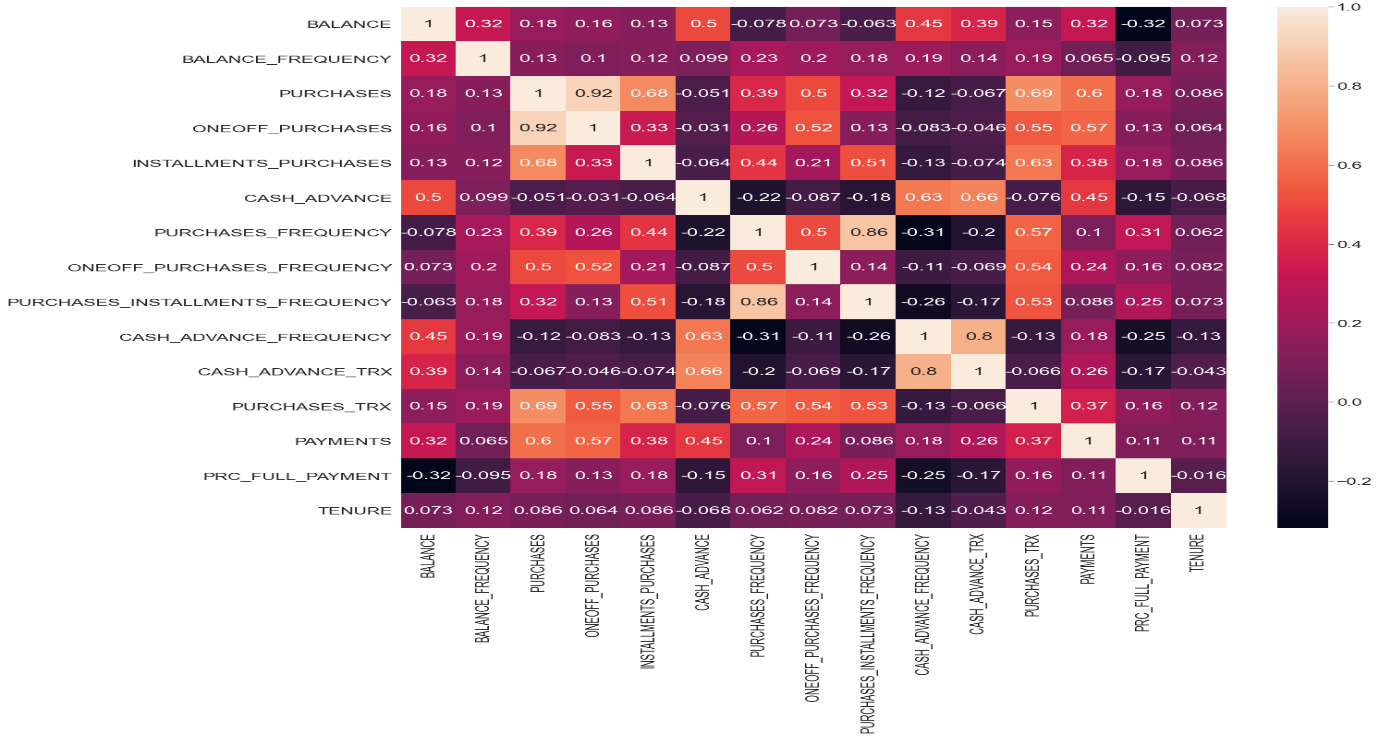
***What amount of the balance are you paying each cycle?***

***How many of the card owners are paying over the minimum monthly payment?***

**Exploratory Data Analysis**

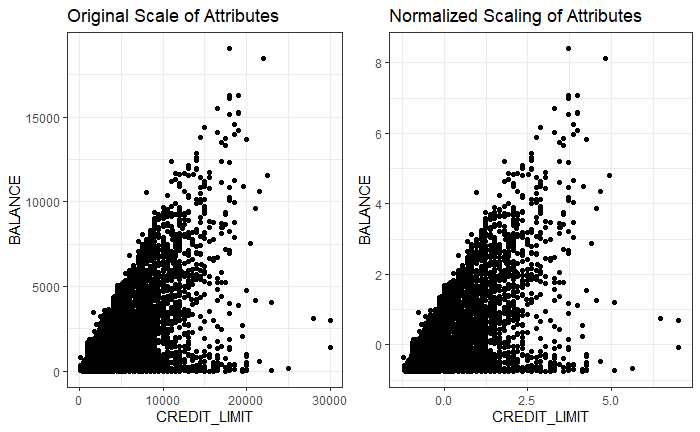
A quick dimensional glance of our dataset shows that it contains 8950 observations across 18 variables. Of those variables, there is a single identifier variable, CUST\_ID, serving as a means of uniquely identifying each record in the dataset, where the remaining variables are all quantitative in nature covering a wide array of aspects regarding credit card usage.

The figure shown above displays a grouped boxplot visualization that allows us to discern multitude of variables having an underlying distribution that is positively skewed. The one variable of slight exception to this observation was in the underlying distribution of *BALANCE*, where it was roughly a normal distribution with only a slight degree of positive skewness. Knowing this now may give rise to how we deal with outliers or missing values. With a skewed distribution such as what is shown above, it’d be prudent to utilize median over mean as the measure of central tendency when dealing with NA’s. Due to the number of records in the dataset, we decided to keep outliers in the dataset as we didn’t want to diminish the sizes of our training and testing datasets when generating models.



**Data Preprocessing**

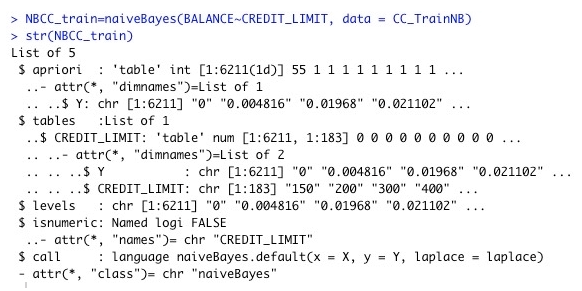
Our first task of preprocessing the dataset was the removal of the *CUST\_ID* column as it provided zero contextual knowledge towards our data questions and merely served as an identifying variable. With our remaining 17 attributes, we know that in its current state it would be fairly difficult to compare variable impact when *PAYMENTS* ranges in values from 0 to 50721.5 and *TENURE* ranges in values from 6 to 12. Therefore, before we can run most machine learning algorithms for useful insights, it’d be prudent to normalize the scale of all our quantitative variables to have a mean of 0 and a standard deviation of 1. Normalizing the dataset attributes is what follows below

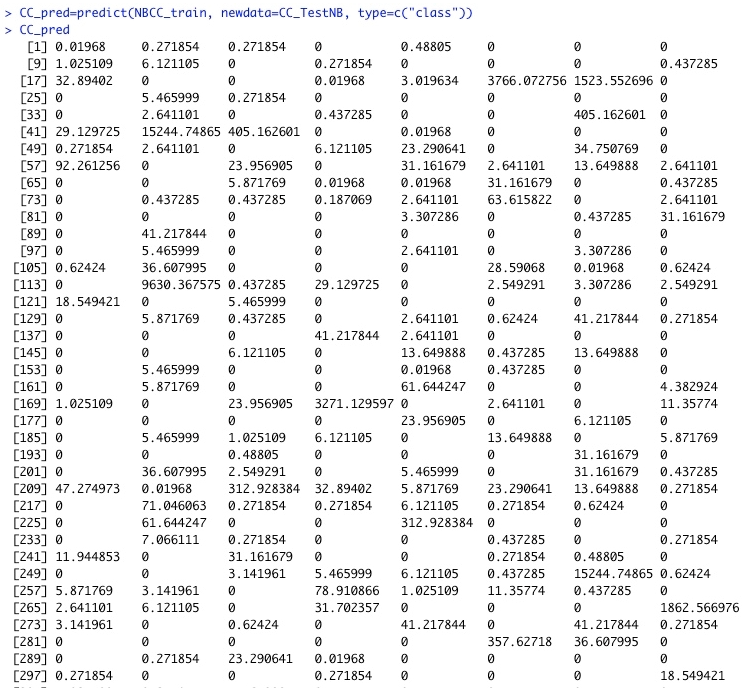


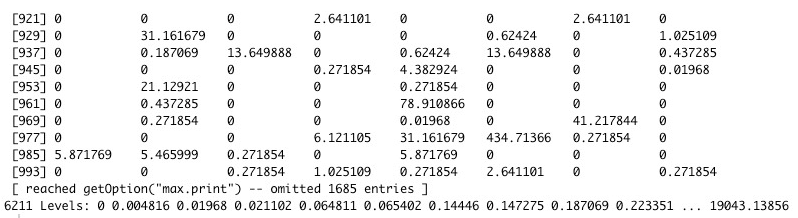
**Naïve-Bayes**

With the focus on credit limit within the dataset, predictions on card user’s balance was attempted through the Naïve Bayes model. Using the discretized dataset, the model applied rendered poor results that were not reflective of approximations or existing balances. Some predicted balances were in amounts of thousands of dollars off from the actual amount. Further review of the dataset along with the model results, concluded that the dataset was not appropriate for this model application.

(Below is a preview of the Naïve-Bayes output, not all records were included to conserve space)



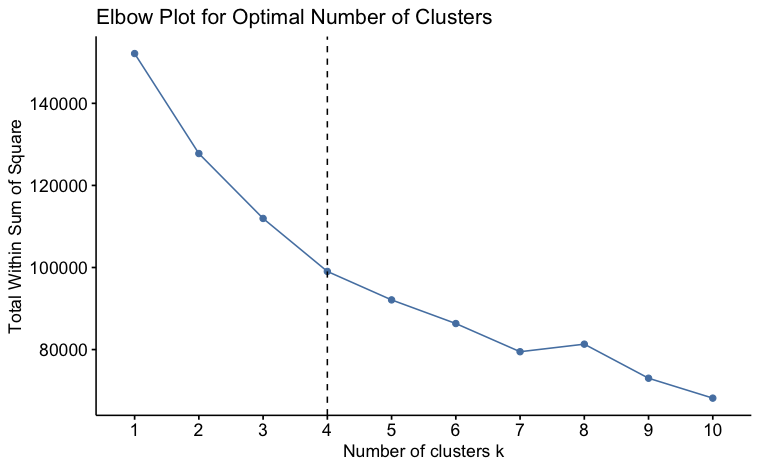




**K-Means Clustering**

K-means clustering is an unsupervised machine learning approach for clustering. When clustering, our aim is to construct groups of records from the dataset where these groups are disjoint in nature having points that a similar with one another. Where we can use these clusters of points and discover the underlying patterns that comprise cardholders within each grouping.

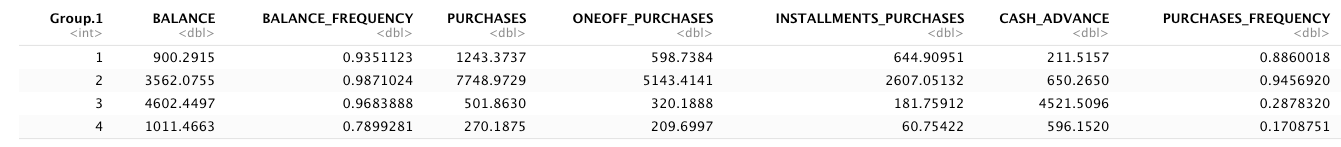
Our first intuitions from previous exploratory data analysis led us to hypothesize a value of *k* between 4 and 7 approach for the number of clusters we should proceed with. Running the algorithm resulted



The above figure displays an “elbow-plot” that can aid us in assessing the most proper number of clusters we should utilize using the K-Means approach when segmenting the bank’s cardholders.

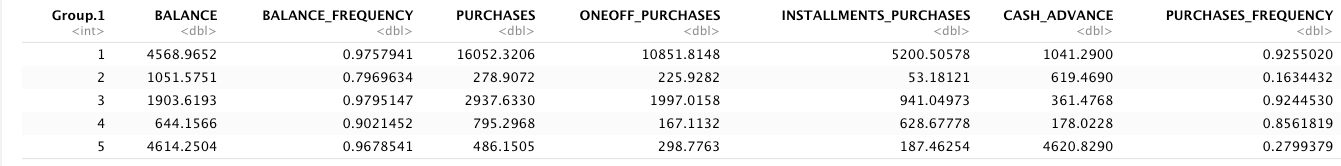
We ran the k-means classification algorithm with values of k between 4 and 7. The elbow-plot suggested this be a good range to evaluate with *k=4* possibly being the best. The four tables following below, displaying cluster size and composition, shall give us more information to confirm this thought before moving forward . We will also include the “quality” of the partition, which is calculated by dividing the *Between Sum of Squares* by the *Total Sum of Squares*.

k = 4



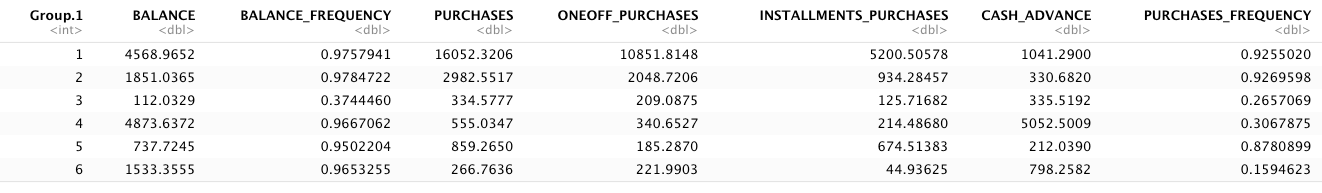
Where cluster 1 has 3367 records, cluster 2 has 402 records, cluster 3 has 1197 records, and cluster 4 has 3984 records. The quality of this partition scheme is 34.89%.

k = 5



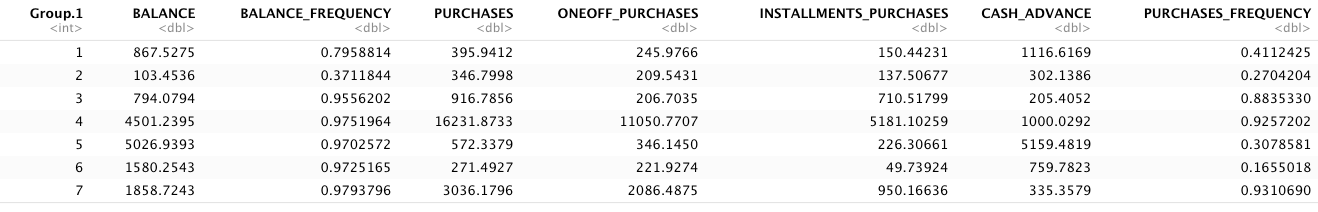
Where cluster 1 has 83 records, cluster 2 has 3914 records, cluster 3 has 1387 records, cluster 4 has 2415 records, and cluster 5 has 1151 records. The quality of this partition scheme is 39.87%.

k = 6



Where cluster 1 has 83 records, cluster 2 has 1323 records, cluster 3 has 1310 records, cluster 4 has 970 records, cluster 5 has 2227 records, and cluster 6 has 3037 records. The quality of this partition scheme is 44.26%.

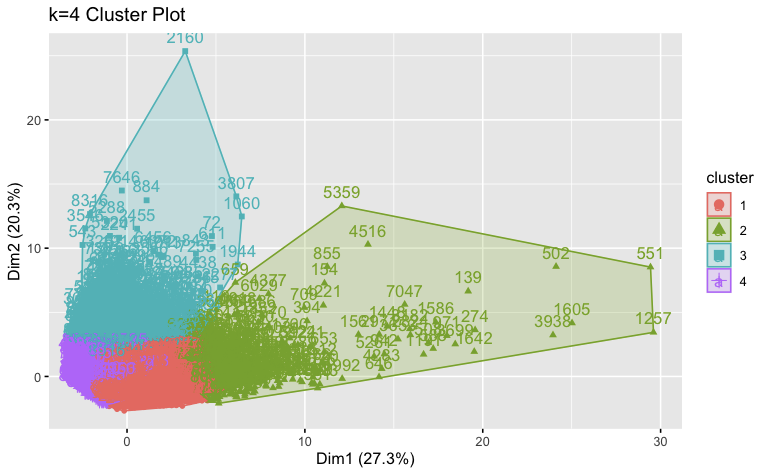
k = 7



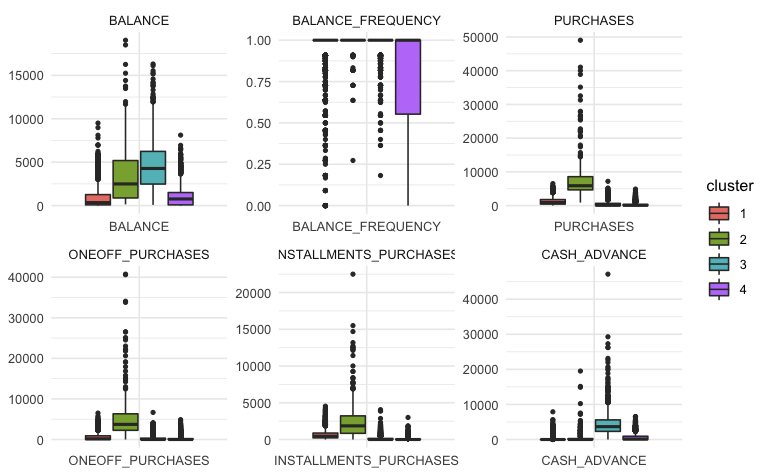
Where cluster 1 has 628 records, cluster 2 has 1187 records, cluster 3 has 2049 records, cluster 4 has 81 records, cluster 5 has 894 records, cluster 6 has 2846 records, and cluster 7 has 1265 records. The quality of this partition scheme is 47.76%.

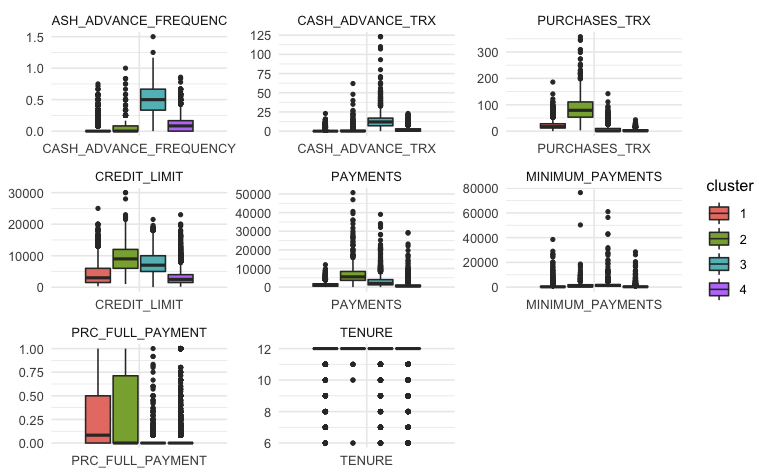
When increasing the value of *k*, we indeed see an increased quality of the partition scheme. But we must also consider the downside of an increase in quality percentage, that the number of records in each cluster goes down and thus makes it more difficult in a business setting to devote resources to addressing each of the clusters. Which leads us to continue that with a sufficient quality of partitioning and applicable settings in the business context that utilizing *k=4* clusters is how we should proceed.

The figure below displays a visualization of the partition scheme selected where k was equal to 4. A quick glance of the plot showcases that there is minimal to no overlap across the clusters, providing further validation that k=4 is a sufficient choice for the proper number of clusters to segment the bank’s cardholders.



Now that we have tuned the parameters of the k-means clustering algorithm with k=4, we must now interpret the types of cardholders belonging to each of the four clusters. Following below are grouped boxplots of the clusters across the attributes of the dataset.





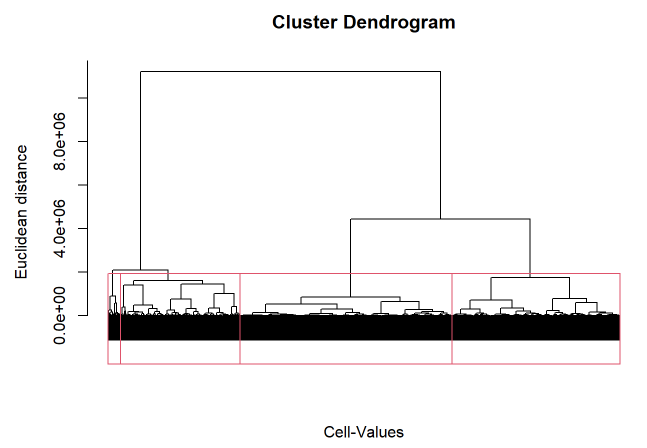
In regard to *cluster 1*, we see that it has a relatively lower *BALANCE*, but is also second in relation to the number of purchases made during a cycle period which indicates a large amount of transaction activity. The *PRC\_FULL\_PAYMENT* shows that despite their balance may be, they are likely to pay off a large majority of it at the end of the cycle period. This leads us to conjecture that the cardholders within *cluster 1* are high frequency users of the credit card with a moderate tier of income that is spending money on lower priced goods. While also confidently paying off a good majority of their balance at the end of their cycle period.

In regard to *cluster 2*, their *ONEOFF\_PURCHASES* is far above the rest meaning that its cardholders have the level of financial assets to be able to make such large payments on one individual item. *CREDIT\_LIMIT* also supports this assertion as banks are more likely to allow higher limits to those who bring in enough money to support this level of activity. They also possess the highest PURCHASES across the clusters in relation to the number of transactions in their cycle period*.* This leads us to conjecture that the cardholders within *cluster 2* are high frequency users of the credit card with a higher tier of income spending their money on higher priced consumer products. While also confidently paying off a good majority of their balance at the end of their cycle period.

In regard to *cluster 3,* we see that their *BALANCE* and *PURCHASES* have a lower amount of engagement with utilizing their credit card. This cluster was the highest in terms of *CASH\_ADVANCE,* meaning that they were likely to borrow some amount of money against their credit card’s line of credit. This leads us to conjecture that the cardholders of *cluster 3* are infrequent users of the credit card that take out a high number of items against their line of credit. Meaning that cardholders falling into this category should closely be examined to ensure they are not falling behind on their balance payoffs.

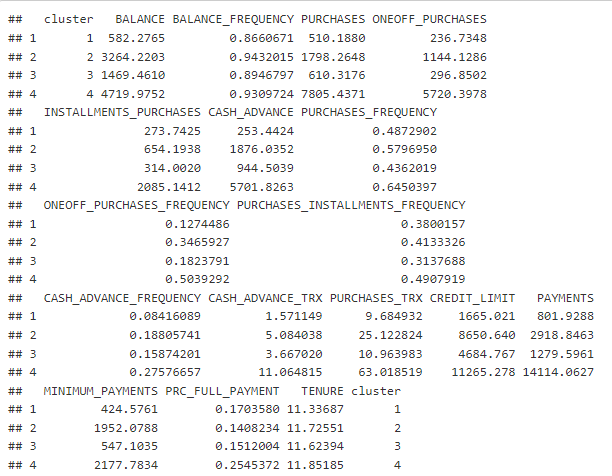
In regard to *cluster 4*, their *PURCHASES* indicates that they are the group of cardholders that seldomly uses their credit card. *CREDIT\_LIMIT* insinuates a lower tier of funds available and *BALANCE\_FREQUENCY* shows that they’re quite sporadic in paying off the balance they do carry. This leads us to conjecture that the cardholders within *cluster 4* are infrequent users of the credit card, with a relatively lower income tier spending their money on lower priced consumer products.

**Hierarchical Clustering**



The dendrogram above shows the clusters found from hierarchical clustering using the ward method.

All other hierarchical clustering methods did not perform well and placed almost all users into the same cluster.



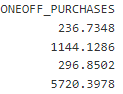
The table above shows the average value for each attribute in each cluster.

**Interpretations**

We can look at a couple of variables to compare the clusters and determine the types of people in each cluster.



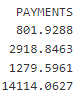
When we look at PURCHASES, this column is the average amount of purchases made from the account. Cluster 1 and 3 seem to be normal consumers spending less than a thousand a month on their credit line. Cluster 2 spends a little more at almost two thousand. This cluster could be more wealthy people. Cluster 4 spends substantially more than any other cluster at almost eight thousand a month. This cluster could be very wealthy people or business accounts.



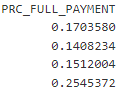
We can also look at ONEOFF\_PURCHASES, which is the average maximum amount purchased in one transaction. Similar to PURCHASES, cluster 1 and 3 are very similar as they have an average max of just under three hundred. Cluster 2 has an average max purchase of 1144 and cluster 4 has an average max purchase of 5720 dollars.



To tie these two columns together, we can look at CREDIT\_LIMIT, which is the average limit of credit each person can spend on their credit line. This is where we start to see the clusters separate from each other. Cluster 1 has the smallest credit limit at 1665, cluster 3 comes next at 4684, cluster 2 is next at 8650, and cluster 4 has the highest credit limit at 11265. Even though cluster 1 has a much smaller credit limit than cluster 3, they have similar spending behaviors.



PAYMENTS column shows the amount of payment made by the user. In this column the first 3 clusters are closer together but still follow the same patterns with cluster 1 at 800, cluster 3 at 1279, and cluster 2 at 2918. Cluster 4 is substantially more at 14114. Looking at PAYMENTS and CREDIT\_LIMIT, we can infer that the users in cluster 4 with higher CREDIT\_LIMIT are spending much more than the others in cluster 4 since the average in PAYMENTS is much higher than the average CREDIT\_LIMIT.



The last column we will look at is PRC\_FULL\_PAYMENT which is the average percentage of full payment by the user. The first 3 clusters are similar with cluster 1 at 17%, cluster 2 at 14%, and cluster 3 at 15%. Cluster 4 likes to pay more of the payment at 25%.

Overall, hierarchical clustering did a decent job at separating users into their clusters. Cluster 1 are users that spend the least and may have the smallest capital. Cluster 2 are users that spend the 2nd most and may have the 2nd highest capital. Cluster 3 are users that spend the 2nd least and may have the 2nd least capital. Cluster 4 are users that spend the most and may have the most capital.

In conclusion, we were pleased with what we discovered and find it very interesting that the outputs from both K-means and Hierarchical were similar and told the same story. The accuracies of their predictability are complimentary, which provided a good peace of mind to all of us. This crucial factor reinforces the conclusion for each method and demonstrates that the approach and output was solid. Our questions, ***What amount of the balance are you paying each cycle?*** and ***How many of the card owners are paying over the minimum monthly payment?*** were answered with accuracy from both methods. It would be interesting to explore this further and tune these methods more or even try other modeling techniques that might fit this type of data even better.