**Analysis Purpose and Target Audience**

The purpose of the analysis is to investigate two main questions for a coffee business that has provided us with a sample of one month of sales transaction data for three stores. The business provided this data in hopes that we could use data mining methods to develop a methodology for the business to expand across its entire enterprise. Ultimately, the company is interested in our analysis so that it can have a better understanding of its loyalty customers – since it views its loyalty customers as those most important to its business model. The business wants to understand the structure of purchases for those customers and any insight into the makeup of its sales transactions for those customers, as well.

**Discussion of the Data**

The dataset used is made up of transactions from the month of April for three stores. In the analysis we begin with considering the dataset (after it is combined from multiple tables) made-up of 41 variables and 49,894 transaction records. Of those, we homed in on 32 potentially useful variables (some, such as ‘product description’ or ‘customer email’ were considered noise for purposes of this analysis) across 24,582 loyalty reward card transactions.

This analysis is made meaningfully more complex than previous projects for two reasons. First, there were numerous transformations necessary to correctly align the data (date/time, unit of measures, multiple table joins). Second, and most important, there are multiple levels of depth within the analysis. The depth includes various degrees of analysis of four hierarchical product structure levels as well as an entirely separate analysis (using different elements of the dataset) of the makeup of transactions. By the end of the analysis, we will have utilized five unique variables in our product analysis and ten unique variables in our structural transaction makeup analysis.

**Questions**

The first question we attempt to answer is, “Can an analysis applied across the organization help us understand what product, type, category or group should be focused on if the company is looking to appeal to its loyalty card customers? That is, can this method offer overall suggestions of what we might decide to ‘upsell’ based on what our loyalty customers have previously purchased?”

The second question we attempt to answer is, “Based on the variables we are collecting as a business, today, can transactions from our loyalty card members be grouped into similar and dissimilar transactions, overall? Or more simply, based on that data are there some variables that account for more variation between our transactions than others?”

**Variable Relationships**

The product variables are a hierarchy that relate to each other in hierarchical form. For example, one such hierarchy in form {Group, Category, Type, Product} is {Beverages, Coffee, Organic brewed coffee, Brazilian Lg}. For this reason, we may find analysis at the Product level is to fine-grained and more understanding can be gained with analysis at the Category level. That is, it may make sense at the end of the analysis to suggest “Tea” to a customer as opposed to a “Spicy Eye Opener Chai Lg”. The data will help inform us what might make the most sense.

In the transaction data, we see correlations in ways we would expect. For example, unit price and wholesale price are highly correlated (0.95) and in birth date of the customer and how long they’ve been a customer (0.60). That would make sense because a younger customer might not have had opportunity to be a customer for as long as an older customer. Surprisingly, there are not any strong correlations with the Unit of Measure and other variables. We might have suspected younger customers buy smaller quantities because they may have less money. Overall, there are not a lot of correlations to help us understand the relationship between these variables, so a cluster analysis might shine a light where we otherwise are unable to observe relationship in two-dimensional space.

**Methods**

We will use two main methods in this analysis, described below.

Association Rule analysis technique is appropriate for this data because it contains a list of purchased items. These items are placed into a matrix for each loyalty customer for the month of April and examined based on which customer purchased which combination (or set) of items for the month. This method is preferred because it will provide the business with statistical values that show which product, category, or type might be more likely to be purchased by someone else in the dataset having also purchased a specific grouping of items. In demonstrating this on a small scale (with essentially sample data from April for only three stores), if the business likes the direction the entire analysis could be applied to the full set of transactions held by the business for more realistic insight.

Clustering methods are appropriate for the second question because the data will, essentially, present any similarities or dissimilarities to us within the transactions. Because the variables for the transactions are mostly able to be transformed to be continuous (including date/time which has been converted to a continuum from a specific point in time) we are also able to understand if, for example, years of being a customer has any influence on the variation of the dataset. The business would be able to identify if there are different groupings of transactions that might guide future advertising or even understand if typically, new employees, based on start date, are involved in one cluster versus another. This method is preferred because we do not have a response variable, but rather want to understand IF transactions can be grouped and distinguished from one another. To put this simply as to why we are analyzing this way: If, for example, two groups of transactions emerge and we look deeply at them, we might notice features within each group that can help us understand training opportunities. Without knowing the groups, however, we cannot yet ask those questions. The first step, performed with this analysis, is to identify the existence of clustering to inform further study. If identified, more analysis would be completed in the future.

Additionally, we performed a brief principal components analysis to help us understand the data structure better in preparation of our other analysis. Not surprisingly, we found that much of the variability in the dataset is driven by price variables. This makes sense given the set of data is essentially transaction records for various products. Additionally, 86% of the variation can be explained by the first 6 principal components which means variables within the data could be eliminated.

**Problems and Issues**

As mentioned elsewhere, having only a sample set of data proves difficult to provide actual outcomes to the business. However, if the direction we have gone with this analysis to show the capability of using data mining tools and techniques is desired by the business, we are very excited to continue working on these problems in the future.

**Analysis**

In analyzing the clusters, we performed many iterations to uncover any consistent patterns that might exist. We performed hierarchical, non-hierarchical, and k-means clustering to look for any sizes (up to 10 groups) of clusters that might be consistent in their groups. Unfortunately, all analysis of such clusters leads to the same conclusion – clustering methods, with these variables, do not produce a consistent outcome. Therefore, this data is not conducive to clustering methods. **Figure 1** shows a hierarchical cluster that does not display even groupings, no matter where it is cut. **Figure 2** shows that even though clusters appear, when moving from one iteration to the next the cluster within which the data is in the points switch groups, therefore resulting in an inconsistent result. Notice the blue band grow from run 1 to run 2 and then meld with all the yellow in run 3. This signifies data that is not grouping well. Again, clustering is found to be ineffective in helping us understand similarities or dissimilarities among this data, as the result of applying this method shows us.

**Evaluation**

Our evaluation of using association rule methods applied to the items purchased by loyalty members shows much promise for a larger analysis beyond the month of April for three stores. We identified that Product Category is a good level to work at because it makes the most sense if the business were upselling. **Figure 3** shows categories at that level in the hierarchy purchased by at least 1% of loyalty customers in April, as an example of the options this data provides us for further research. Through our analysis of the four hierarchy levels, we first identified that many flavorings and merchandise items were being associated to purchases. It is reasonable to assume that other food items, particularly ones that would ‘sell on their own’ without a companion product like “coffee/flavor” would be the target to identify. For this reason, we reduced the types of product groups (at a higher level in the hierarchy) to only look at Beverages and Food, excluding Add-ons, Merchandise, and Whole Bean Teas. In doing so, we discovered 32 item association rules that we used to further understand the purchasing data. Using this data, we determined Bakery items would be a good target for upselling because it continually rose to the top by its value of “lift” amongst other products. When examining bakery purchases in the entire dataset, we found, for example, a loyalty customer who bought Coffee, Drinking and Packaged Chocolate, and Tea was almost 1.2 times more likely to buy Bakery compared to a randomly chosen loyalty customer in the dataset. Further, there is a 95% probability that if they purchased those other items they will also purchase Bakery (again, only from those within this dataset, however). While this may not be relevant beyond this sample of data, the same method and concepts can be applied to a much larger roll-out, potentially being used to program POS interfaces to make product recommendations on-the-fly.

While we assessed many iterations of k-means clustering, hierarchical clustering, and non-hierarchical clustering, we failed to identify any particular groupings of data that would be helpful for future analysis. We did identify groupings that seemed to align to particular variables, such as price, when we split into nine groups. However, this is not helpful as if we wanted to split price into groups we could do so without such an analysis. I would conclude that this particular transaction data is not able to have unsupervised clustering methods applied to it in a way that is helpful for any future purpose or in applying to a larger population. We had high hopes, but this data told us a different story.

**Summary**

Overall, we found that using Association Rule Methods with this data could lead to some great outcomes for the business if applied to a larger dataset involving more customers and time periods. We did not, however, find much use for clustering with this data as it does not, to our surprise, yield much value from performing clustering methods. In the future, we might suggest more customer data be collected and an attempt be made at determining if specific sales could be predicted by certain customer demographics or patterns. For purposes of this analysis, however, we did not find meaningful clusters to further assess.

**Figures**

Figure 1: Clusters using complete linkage – Notice the unevenness of the groups connecting to the vertical lines as each horizontal colored line represents a “cut”. In equal groupings, the points at the bottom would be more evenly spread.

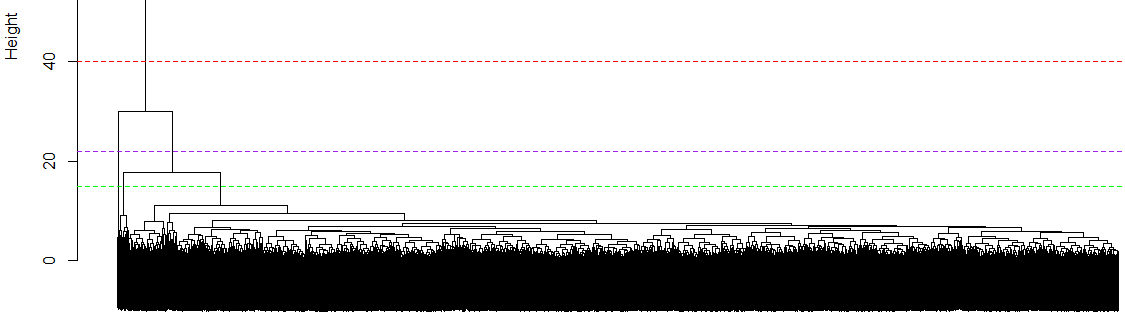


Figure 2: Five clusters showing groups switching between iterations using k-means

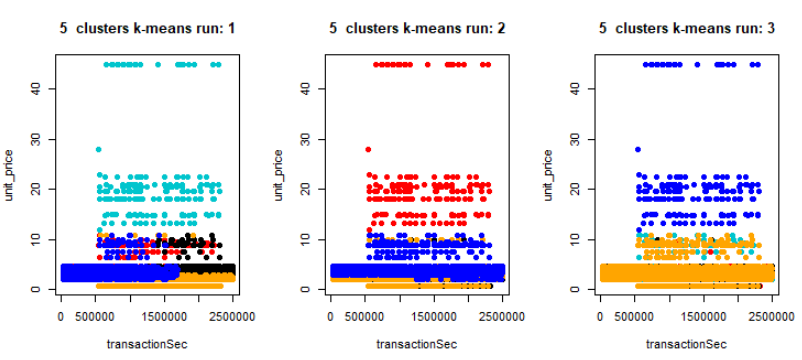
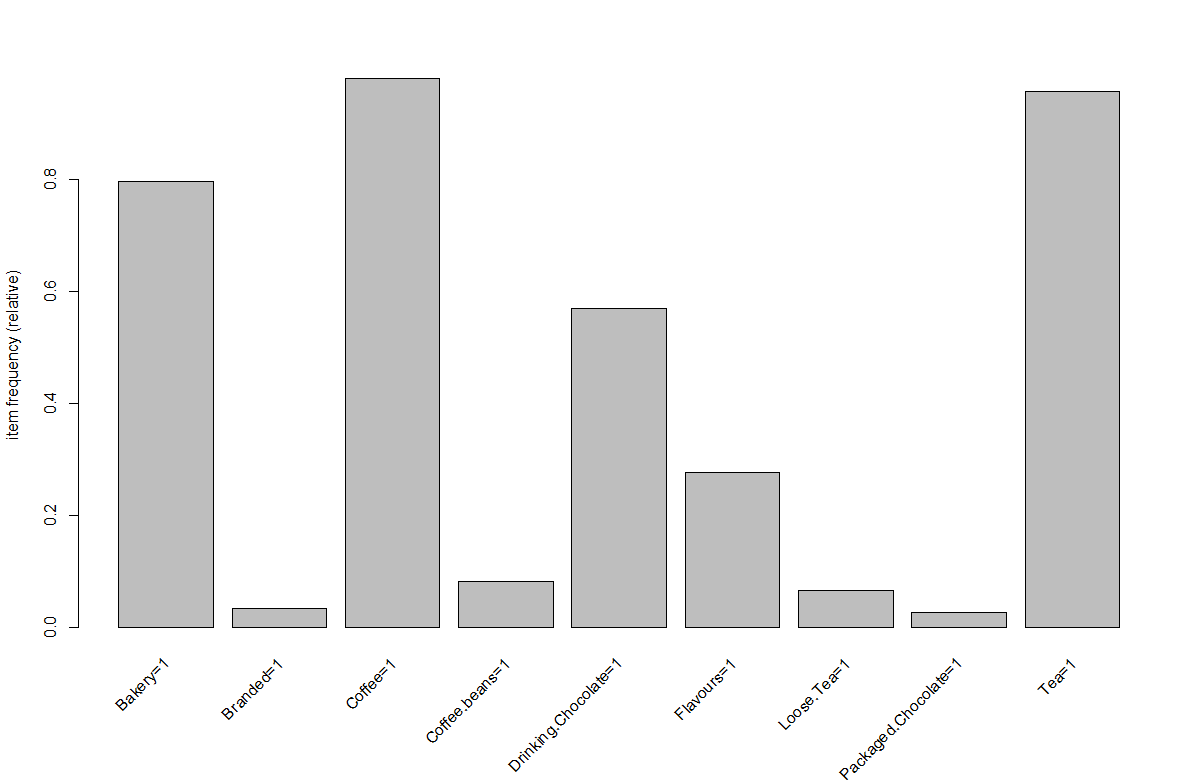


Figure 3: Plot of the various Categories of items purchased by 1% of Loyalty Card Customers



**References**

Dataset for Analysis Found at: <https://www.kaggle.com/ylchang/coffee-shop-sample-data-1113>

[Convert time interval to number in R](https://ro-che.info/articles/2017-05-18-r-time-interval-to-number)

[r - Combining date and time into a Date column for plotting - Stack Overflow](https://stackoverflow.com/questions/24105984/combining-date-and-time-into-a-date-column-for-plotting)

[clustering - DateTime Variable -- Kmeans - Cross Validated](https://stats.stackexchange.com/questions/221369/datetime-variable-kmeans)

[Data Cleanup: Remove NA rows in R - ProgrammingR](http://www.programmingr.com/examples/remove-na-rows-in-r/)

[r - How to drop columns by name in a data frame - Stack Overflow](https://stackoverflow.com/questions/5234117/how-to-drop-columns-by-name-in-a-data-frame)

[r - Drop data frame columns by name - Stack Overflow](https://stackoverflow.com/questions/4605206/drop-data-frame-columns-by-name)

[r - How to solve prcomp.default(): cannot rescale a constant/zero column to unit variance - Stack Overflow](https://stackoverflow.com/questions/40315227/how-to-solve-prcomp-default-cannot-rescale-a-constant-zero-column-to-unit-var/40317343)