**Group #1 (Casey O’Malley, Brad Kim, Zoe Huang, Qi Ding)**

**Applied Analytics Frameworks and Methods II**

**Final Project Write-Up – Black Friday Dataset**

**Deliverable #2:**

**Problem Statement:**

The purpose of this report is to discuss the analytical work that our project group has completed during the Spring 2019 semester for our final project and presentation. Our group chose to work on the Black Friday dataset, which was originally released as part of a machine learning contest on Kaggle.com. The goal of the Kaggle competition was to predict the purchase amounts for transactions in the testing dataset, but our goals for this project expanded beyond those instructions. We looked at the dataset through the lens of a company trying to accomplish the following goal: use the analytical techniques and frameworks we learned this semester to develop insights about the company’s customers and products, which could then be leveraged to increase future sales.

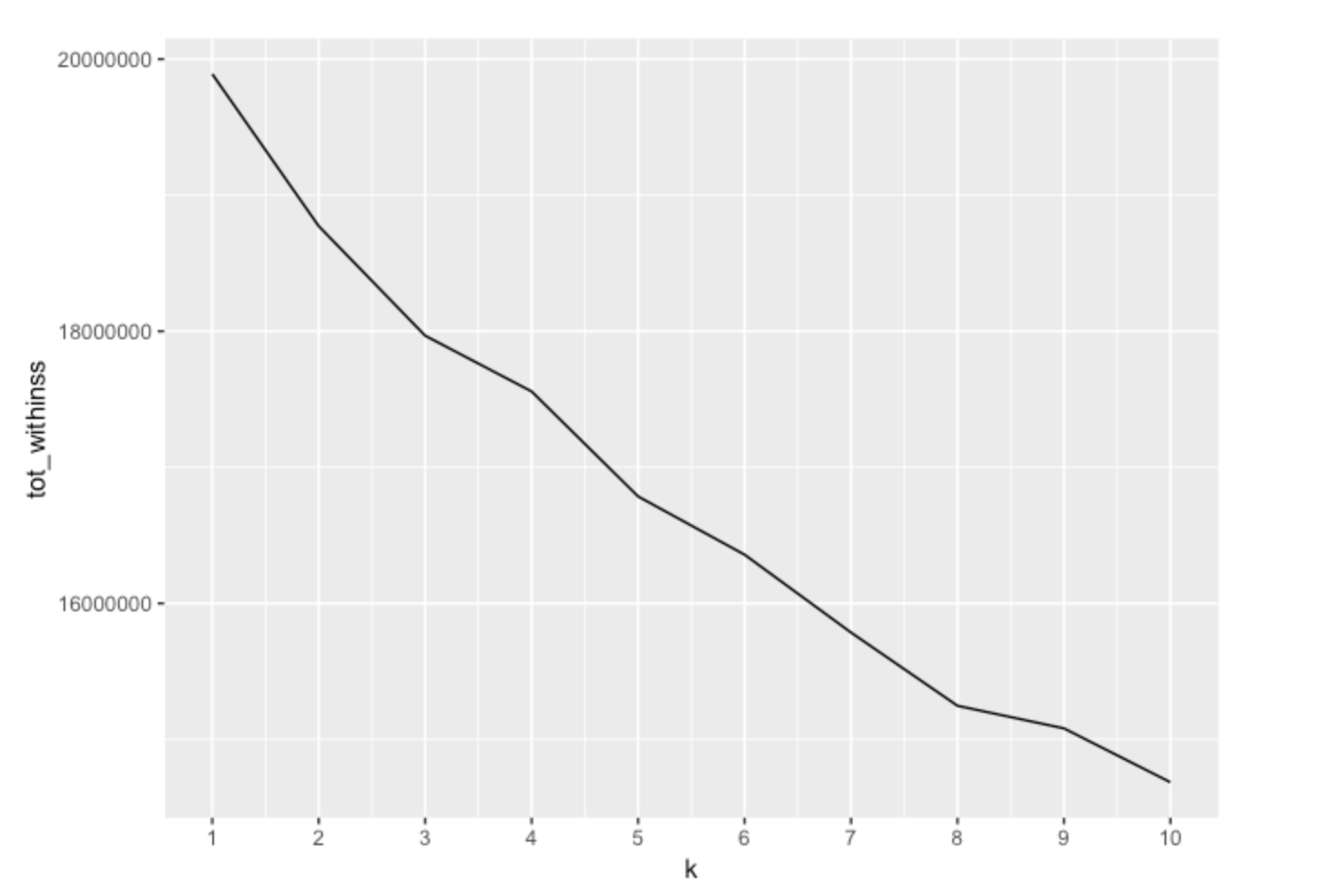
**Analytic Technique Reasoning:**

Our group chose three main techniques from this semester to conduct our analyses: K-means clustering, recommender system, and market basket analysis.

* K-means clustering was used to find different groups of customers within the dataset that have certain qualities in common. This was very relevant because k-means can process the variables that we were interested in analyzing and use those variables to develop unsupervised insights. In this case, k-means clustering could potentially help us find groups of customers within the data with similar demographic qualities and purchasing habits. If we are able to find interesting customer groupings in the k-means clusters, we may be able to use this information to create new market segmentation strategies based on the insights.
* Recommender systems are extremely relevant in this case and could help the company point its customers towards products that they may like based on their past purchase history and their perspectives on the products they already purchased. Since Black Friday sales are typically a characteristic of retail stores, it is imperative to understand how recommendations could be used to encourage more spending amongst customers. If developed properly, this system could eventually become a competitive advantage for the company that improves over time and keeps customers engaged.
* Similar to recommender systems, market basket analyses are highly relevant in the retail world. The insights from this analysis will help the company understand aggregate consumer behavior through the products that are commonly bought together. The company can then use these insights to create new promotional events that combine products that customers frequently buy together to increase the overall sales of these products.

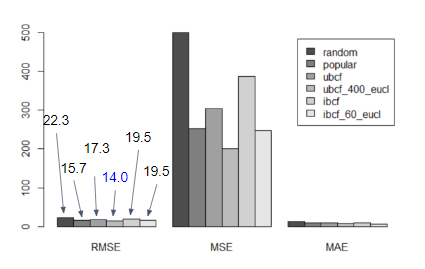
**Analytical Results:**

The results from the k-means clustering analysis were a bit disappointing. By creating dummy variables from categorical variables, we were able to process our data into a format that worked well with the kmeans function in R. Specifically, we combined the various demographic categories describing the customers to create an assortment of dummy variables with the Caret package; this was done to focus on clustering combinations of user qualities without using any product or transaction price information. We used a map function to find the total within sum of squares for the model, with the “centers” argument ranging from 1 to 10 clusters. Unfortunately, this work did not yield great results, as the elbow plot generated by our model did not show any “elbow” where the slope of the line flattens out. Instead, the graph presents a relatively straight shape for the first eight values and is difficult to interpret:



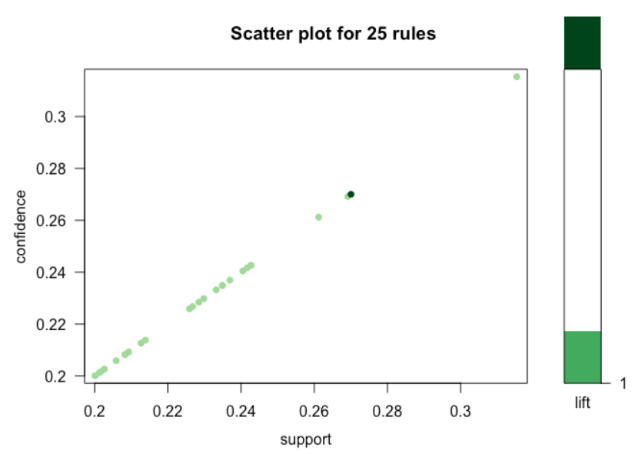
Nonetheless, we still wanted to see what the clustering results would yield and proceeded with a 4-cluster solution since that appeared to most closely resemble an elbow. The kmeans model assigned clusters in a somewhat lopsided fashion, with cluster sizes ranging from roughly 30,000 to 226,000. Precisely, the first cluster had 151,748 observations, the second 128,810, the third 30,423, and the fourth 226,596. Our disappointment with the elbow plot was compounded by the lack of insights we found about their purchasing habits. The average purchase amount for each cluster ranged from 8,798 to 9,497 – all hovering near the entire dataset’s average purchase amount of 9,333.86. Despite our efforts, none of these clusters had significantly different purchasing habits than the overall populations, which prohibited us from making any actionable insights based on the results that could increase sales.

The results from the recommender system were far more interesting than those from k-means clustering. For this method, each of the 18 product categories were analyzed instead of including thousands of individual products in the system; this greatly increased the simplicity of our analysis and the power of the system’s insights. To complete the analysis, we were required to make a few assumptions about the data. We assumed that 1 customer transaction was equal to 1 customer rating to keep a count of the product categories that were purchased by each user; this effectively let us model the recommendations without having explicit ratings for each product in the dataset. Once this data was processed into a workable data matrix format, four recommender algorithms were tested on the data: Random, Popular, User-Based Collaborative Filtering (UBCF), and Item-Based Collaborative filtering (IBCF). Ultimately, six models were built – two each for UBCF and IBCF – and we proceeded with the UBCF model that had the lowest RMSE (14.0) and MSE (200.78) of all the systems:

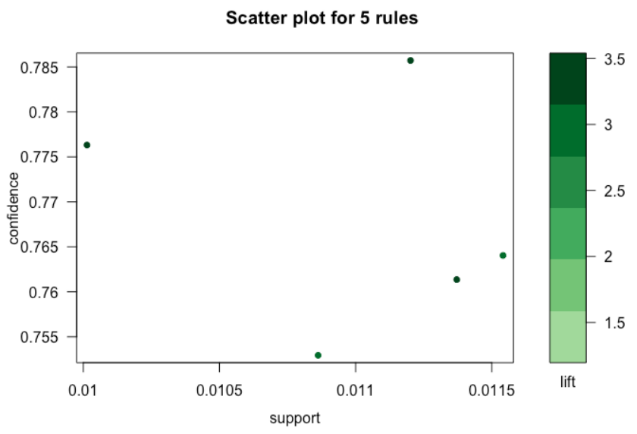


Once we saw these results, we proceeded to test the system on our data. Specifically, we tested the first five users in the dataset to see their top 5 unpurchased product categories. This returned the most highly recommended product categories for each user, effectively creating actionable insights for the company’s users that could expand their purchasing habits into new categories. We also predicted the expected number of products that would be purchased by each user in their unpurchased categories. For example, we found that the first user was most likely to purchase ~3.23 products of Product\_Category\_1 item #7 – a category from which they had never purchased before. Insights such as these could be used by the company to direct customers towards product categories they have never purchased before but are likely to purchase based on data from other users. Needless to say, a system that is successful in doing this would be hugely beneficial for our goal of increasing sales during future Black Friday events.

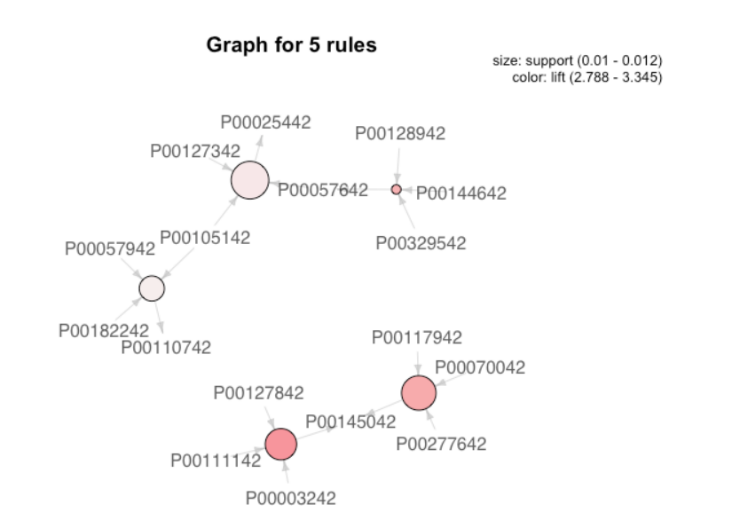
The results from the market basket analysis were also interesting and actionable. Similar to the previous analyses, the arules package requires data to be processed in a transactional format that can be analyzed by the apriori function. Once the data was manipulated and ready, we started out by creating a model using 0.05 for both the confidence and support parameters. Since these are relatively low thresholds, the model ended up generating 3,280 association rules with the data – way too many for us to process. We proceeded to increase the support parameter to 0.2, which reduced the rules to 25. Although this was much more reasonable to interpret, a red flag quickly emerged when we tried to plot this data:



The direct relationship shown in the plot indicated that only one product was included in each rule, which was validated when we saw that no lhs variables were being recorded in the summary table. We then tried building a third model that lowered the support parameter down to 0.01 and increased the confidence parameter to 0.75. This ended up generating 5 rules that looked much better (i.e. more random) when we plotted them:



Unlike the previous model, this set of rules contained products in both the lhs and rhs columns of the summary table. This immediately let us know that it was indeed generating the results we wanted to see. Despite the fact that we only had product IDs (not the actual product names or descriptions) to work with, we were able to visualize the rules in an interpretable way:



We cannot do much at this point without the underlying product information, but our analysis has proven that association rules can be identified for items that are frequently bought by the same users. Since the company does have this crucial information, they could analyze this model to determine if any of these associated products could work well together in a promotional event. If so, the company could almost certainly use the insights from this model to further encourage purchasing certain products in-tandem and thus increase their overall sales figures.

**Conclusions:**

The goal of our group was to think like this company and apply advanced analytical techniques to generate insights that could lead to higher sales figure during the next Black Friday event. We applied three techniques to accomplish this goal: k-means clustering, recommender system, and market basket analysis. While the k-means clusters did not generate any actionable results, both the recommender system and the market basket analysis were successful in this endeavor. We discussed those insights in previous sections, but will reiterate them here:

* A user-based collaborative filtering (UBCF) recommender system worked best with the data and allowed us to determine the expected ratings for users in product categories they had never purchased from before. It also allowed us to predict how many products a user would buy from a certain category that he or she has never purchased from before. The company can use this information to target customers with high unpurchased product scores and use this to achieve their goal of increasing sales.
* A set of association rules that use a relatively low support threshold and high confidence parameter yielded the best association rules in our model. While we cannot make any strategic moves without the underlying product information, the company would have enough information to apply these rules to their products. This could eventually lead to promotions that encourage purchasing certain products together, which would effectively help the company achieve its goal of increasing sales.

**Deliverable #1:**

1. Project name: A study of sales through consumer behaviors on a Black Friday

2. Data source: https://www.kaggle.com/mehdidag/black-friday

3. Description of dataset: The dataset is a sample of the transactions made in a retail store on a Black Friday. This dataset contains 550,000 observations and 12 different variables that are either numerical or categorical. It also contains missing values for some variables. This store can have a better understanding of customer purchasing behavior for its various products by performing different analytical techniques on this dataset. This includes analyses that predict the dependent variable (the amount of purchase) with the help of the information obtained from the other variables and clustering analyses to find different groups of consumers.

4. Description of each variable:

* User\_ID: Unique identifier of shopper. There are a total of 5891 unique users in the dataset;
* Product\_ID: Unique identifier of the product. There are a total of 3623 unique products in the dataset;
* Gender: the gender of the person making the transaction;
* Age: the age of the person making the transaction, already put into different groups;
* Occupation: Occupation of a shopper, already labeled with numbers 0 to 20;
* City\_Category: the category of user's living city. Cities are categorized into 3 different categories 'A', 'B' and 'C';
* Stay\_In\_Current\_City\_Years: how long the users have lived in this city;
* Marital\_Status: Marital status of shopper. 0 stands for ‘not married’ and 1 for ‘married’;
* Product\_Category\_1: Product category of purchase. There are 18 unique products listed in this category;
* Product\_Category\_2: Product category of purchase. There are 18 unique products listed in this category;
* Product\_Category\_3: Product category of purchase. There are 16 unique products listed in this category;
* Purchase: purchase amount in dollars in one specific transaction.

5. Issues and approaches to cope with them:

5.1 Issues:

* Some variables are labeled and we do not know what the labels stand for exactly. For example, the Occupation variable is labeled from 0 to 20 and the City\_Category variable has three different types “A, B, and C.” In the case of City\_Category, it is impossible to tell what the characters stand for but we presume that they are independent;
* Products in category\_1, category\_2, and category\_3 overlap too much. For instance, 17/18 unique products in category\_2 are also listed in category\_1;
* The Products\_categories variable is not listed as product\_ID, so it is hard to distinguish them from one to another;
* Read the CSV file, without having any column converted as factors. For example, currently Marital status is in either 0 or 1, which is not a factor. We should convert the Marital status column to either Married or Single, or at least convert it into a factor;
* Identify the columns in the dataset that have missing values in them. There are two features category\_2 and \_3 that do not have data for all the observations, and the proportion of NAs is very high;
* In the whole dataset, each observation represents one purchase instead of one user; however, for the future clustering analysis, we want to see more information about each customer.

5.2 Approaches:

* Missing Values: since the missing values are only in the secondary product category columns, we presume that these users didn’t buy any product in that category and replaced these NAs with zero. This dataset is already pretty clean in terms of missing value since other features all have data for all the observations. However, if there were any missing values in other features such as age, occupation, City\_Category, and Marital\_Status, then we would consider to remove these observations instead of using imputations;
* Data type: categorical variables are converted to factors for future analysis;
* Scaling for clustering analysis: for clustering analysis, we usually need to scale variables to make them unified. In this case, all categorical variables are converted to factors since these variables are already unified and we don’t have to rescale them;
* Group all the purchases made by one customer to one row. It can be expected that the categorical values stay the same in each entry, so the first entry is used. When the sum of all the purchases is used, the number of purchases is also included to deduce the mean later.

6. Questions to answer:

6.1 Clustering related: by performing K-means clustering and Hierarchical clustering, we could narrow down our target customer and make a recommendation based on the clustering.

* City: Among A, B, C which city generates the most sales?
* Occupation: which occupation generates the highest sales?
* Gender: between males and females who spend more?
* Stay\_In\_Current\_City\_Years: ranging from 0 years to 4 years, which group buy the most.
* Mixed combinations:
  + in a certain city – A, B, C – which occupation and gender devote the most in sales?
  + in a certain city, which resident group, staying in current city for ranging from 0 years to 4 years, generates the most sales?
  + for each occupation – 1,2,3,….18,19,20 – which gender and resident group plays a key role in the sale?

6.2 Which age groups have the greatest interest in the store? In this situation, losing interest means less purchases on Black Friday;

6.3 Top buyers: We will try to come up with new or edited categories by looking into the product\_ID variable and pair it with the cluster that we performed in 6.1;

6.4 Top sellers: Since we do not have a deep understanding or more details about what the categories are, we cannot diagnose if the categories are good or not. However, we can still look into which products are the best seller items;

6.5 Dimension reduction: We could also perform dimension reduction such as PCA and Factor Analysis on product variables that we create in 6.3 since there are too many items to select. This would serve as a step to simplify the data rather than a way to find analytical insights. Therefore, we put it as the last step.