Ravali Kandula
Shawn Deluz
Anqi Liu
Minhyeon Kwon
Amber Huang
Vinay Agarwal

Customer and Social Analytics

Market Basket Analysis

Introduction

Market basket analysis can be used to identify which items are often purchased together, which can help retailers deploy effective marketing and sales strategies to increase customer engagement and customer experience. By understanding the items that customers are likely to purchase together, the retailer can create targeted promotions and discounts to encourage customers to purchase more items. Additionally, the retailer can use the data to create more personalized experiences for customers, such as tailored product recommendations.
The question we are trying to answer based on this dataset is, to help the retailer deploy effective marketing and sales strategies to increase customer engagement and customer experience for its retail store.

Data Description

We have sourced the data on grocery store transactions from Kaggle. The data contains three columns;

1. Member_Id: This is a unique numeric value that signifies the members that shop at the retail outlet
2. Date of Transaction: The date of transaction in DD-MM-YYYY format
3. Products Purchased: The specific items that were purchased by the customer

	Member_number	Date	itemDescription
1	1808	21-07-2015	tropical fruit
2	2552	05-01-2015	whole milk
3	2300	19-09-2015	pip fruit
4	1187	12-12-2015	other vegetables
5	3037	01-02-2015	whole milk

Fig 1: Snapshot of the dataset

Data Cleaning

Before market basket analysis can begin, the data must be cleaned and manipulated into a usable format. First, the "member_number" variable in the dataset needed to be converted to a numeric data type. This allows us to then sort the data into transactions with the same member number and date, which represents an entire transaction. By doing this, every item that a customer bought in a single trip to the store is accounted for in the item description column of the same row. Now that transactions are in the same row, we can split each product item into its own column, which allows for connections between items to be made. In order to do this, we needed to find the maximum number of items in a single transaction, which was found to be 11 unique products. Finally, we removed any rows that did not have at least two products in that transaction, since these rows would not provide any links between products and would just add noise to our analysis. The data is now in a proper structure for network analysis to be conducted.

Data Analysis

Apriori rules are a type of association rule used in market basket analysis. They are used to identify relationships between items in a collection of data. Apriori rules are based on the concept of frequent item sets, which are sets of items that appear together frequently in the data. We used Apriori algorithm to identify rules/ relationships between sets of items. These rules can be better understood using three metrics:

1. Support: The percentage of transactions that contain all of the items in an itemset. The higher the support the more frequently the itemset occurs. Rules with a high support are preferred since they are likely to be applicable to a large number of future transactions.
2. Confidence: The probability that a transaction that contains the items on the left hand side of the rule also contains the item on the right hand side. The higher the confidence, the

greater the likelihood that the item on the right hand side will be purchased or the greater the return rate you can expect for a given rule.

3. Lift: The probability of all of the items in a rule occurring together divided by the product of the probabilities of the items on the left and right hand side occurring as if there was no association between them.

We identified **1250 rules** from the given dataset and further sorted it by the highest likelihood of items in the LHS leading to the purchase of the item on the RHS side. The table below shows the top ten closely related association rules.

Set of Rules: 1250

	lhs		rhs	support	confidence	coverage	lift	count
[1]	{whole milk, yogurt}	=>	{sausage}	0.001470195	0.13173653	0.011160118	2.183062	22
[2]	{sausage, whole milk}	=>	{yogurt}	0.001470195	0.16417910	0.008954825	1.911888	22
[3]	{citrus fruit}	=>	{specialty chocolate}	0.001403368	0.02641509	0.053127506	1.653872	21
[4]	{specialty chocolate}	=>	{citrus fruit}	0.001403368	0.08786611	0.015971665	1.653872	21
[5]	{sausage, yogurt}	=>	{whole milk}	0.001470195	0.25581395	0.005747126	1.619975	22
[6]	{flour}	=>	{tropical fruit}	0.001069233	0.10958904	0.009756750	1.617249	16
[7]	{tropical fruit}	=>	{flour}	0.001069233	0.01577909	0.067762630	1.617249	16
[8]	{beverages}	=>	{sausage}	0.001537022	0.09274194	0.016573109	1.536866	23
[9]	{sausage}	=>	{beverages}	0.001537022	0.02547065			
[10]	{soda, whole milk}	=>	{sausage}	0.001069233	0.09195402			

Fig 2: Snapshot of Top 10 Rules sorted by Lift

Now we understand the closely related association rules, we further try to visualize these rules in different ways to draw a better clarity about them.

Introductory Data Visualizations

- **Frequency Distribution**

  The frequency distribution is a statistical technique that is used to count the number of occurrences of each value in our dataset and display the results in the graph below. Frequency distributions can be used to identify patterns in the data, such as the most common values or the range of values. In our dataset the maximum number of

transactions involve whole milks as the most frequent, followed by vegetable purchases and more.

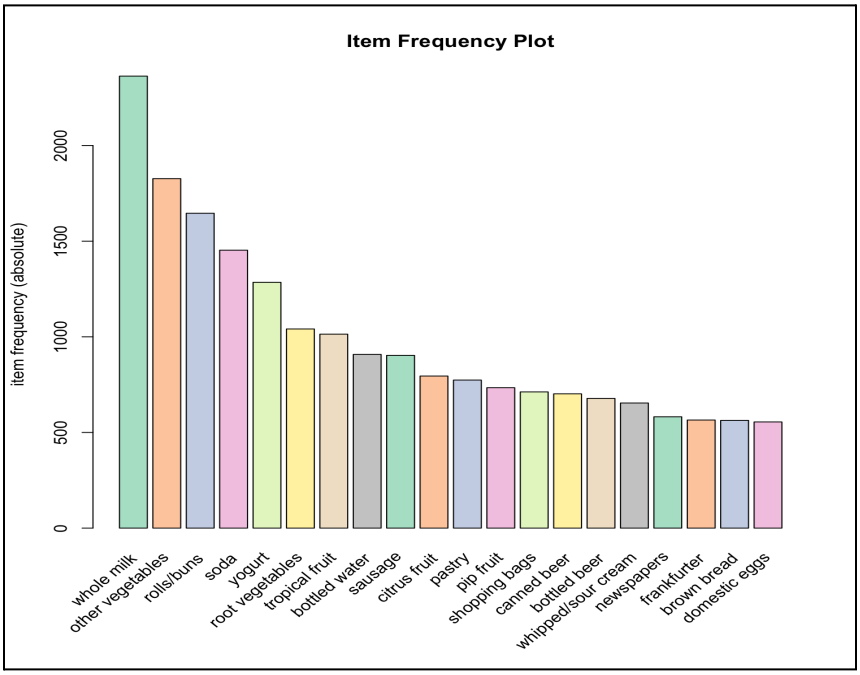

Fig 3: Frequency distribution

- **Scatter Plot**

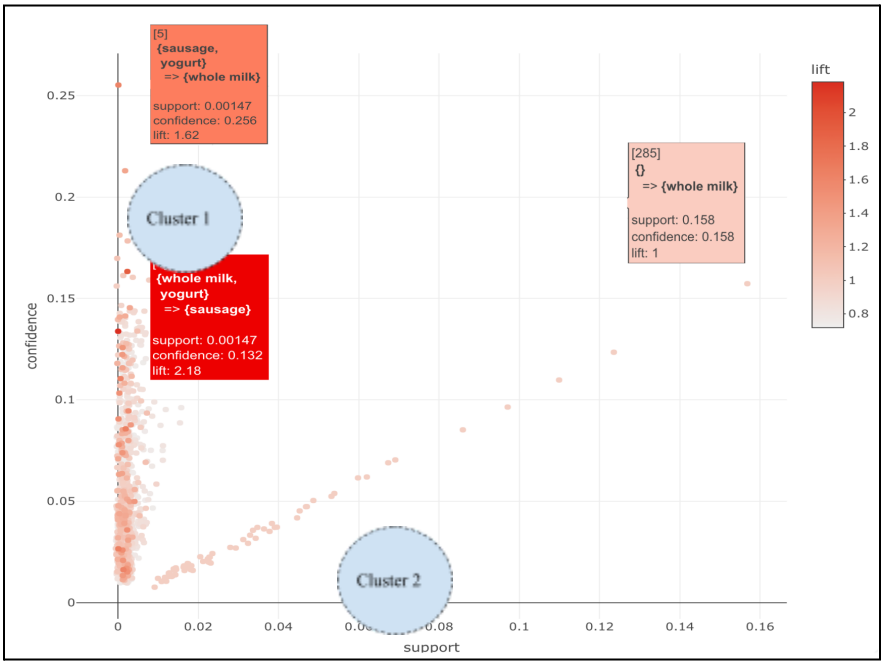

Fig 4: Scatter plot

A scatter plot is a type of graph used to show the relationship between two variables, support and confidence here. It is a type of data visualization that plots the values of two variables against each other on a two-dimensional graph.

From the scatter plot we can see that the rules having a strong association are dark red in color followed by a fading color for a relatively weaker rule. Thus, we infer that support and confidence do not significantly impact the strength of association rules, it is a direct function of lift.

We observe that the scatterplot is segmented into two clusters Cluster1 and Cluster2, from the plot above we can observe the clusters. Cluster 1 consists of rules that have items both on the LHS and RHS side of the rule, thus nearing the same levels of support ,confidence and lift. Cluster 2 has rules that have items only on the RHS of the rule thus having similar characteristics to the rules in that cluster. The items in Cluster 2 are the ones being purchased without any high direct association to any other items in the dataset.

- **Grouped Matrix**

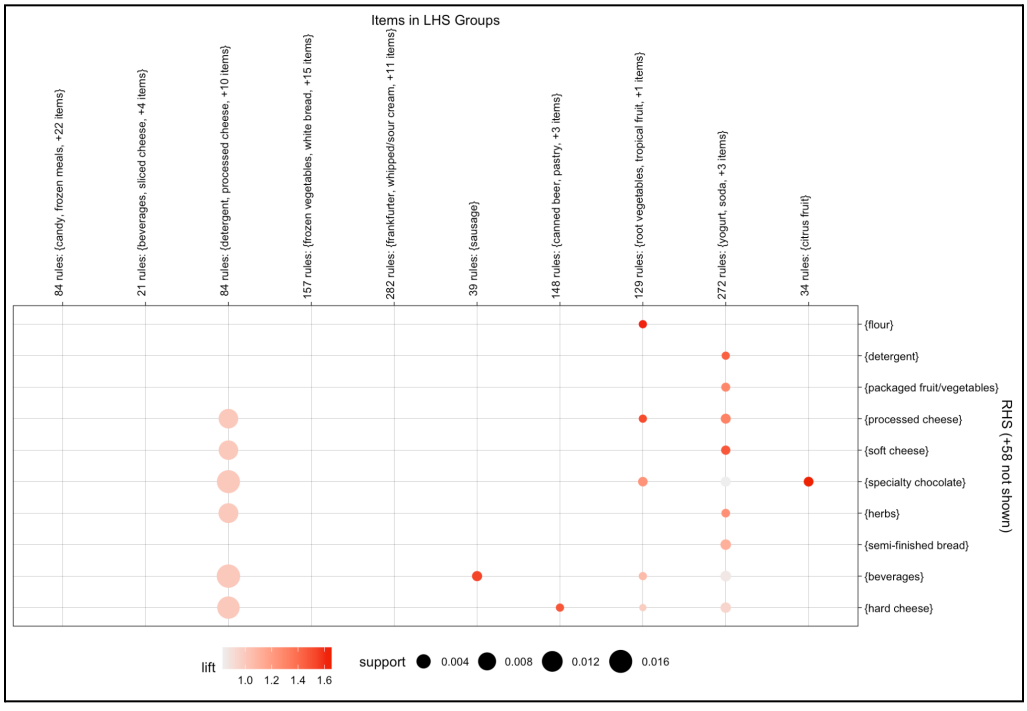

Fig 5: Grouped Matrix

A grouped matrix visualization for market basket analysis is a type of data visualization that displays the relationships between items in a market basket. The X-axis represents the items on the left hand side of the association rule and the Y-axis represents the items on the right side of the association rule.

On the matrix, a colored dot at the intersection of the items on LHS and RHS represents that an association exists between them. If there happens to be no association between the items, there would be no dot at the intersection.

The size of the dot represents the support and the color shade represents the lift value. The best case scenario would be to have as big of a dot as possible, indicating high confidence and dark red in color indicating a high lift value.

Graph

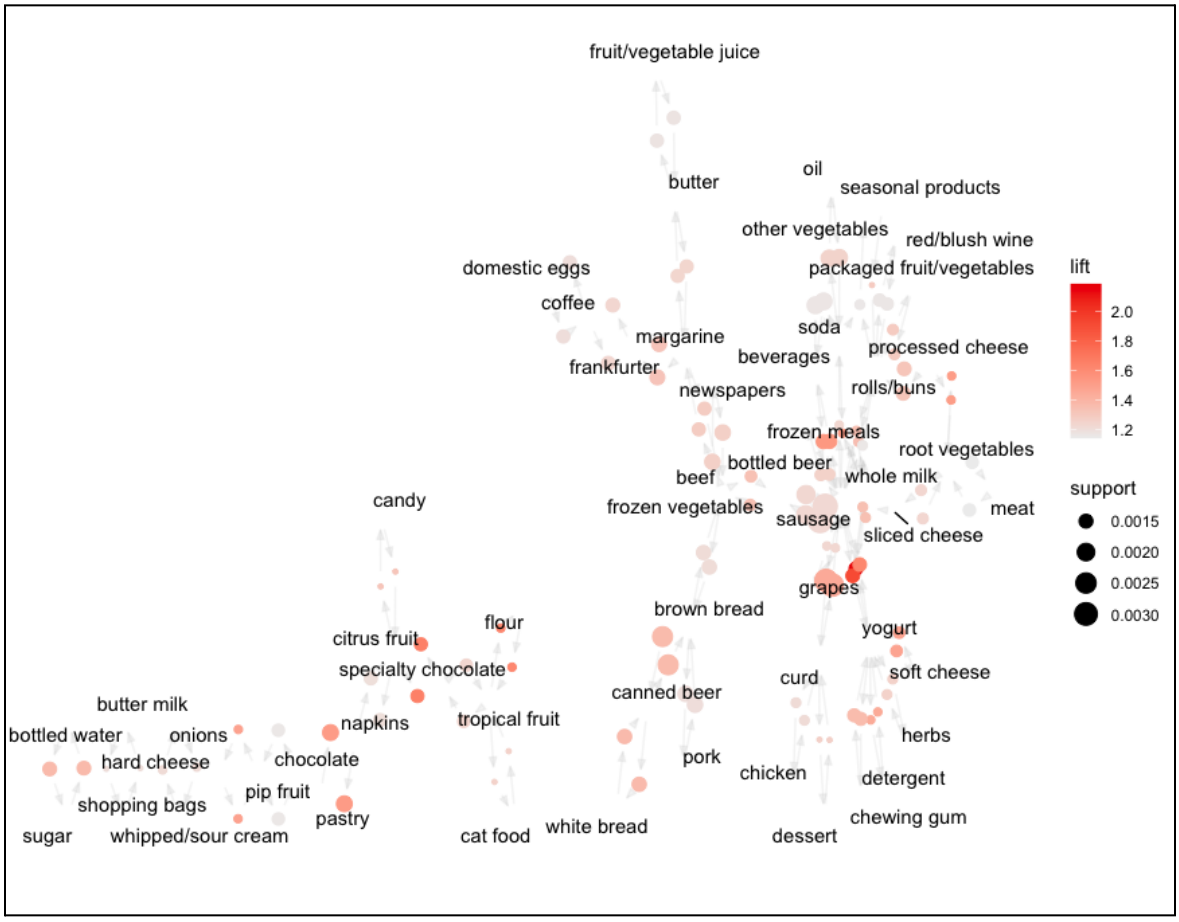

Fig 6: Directed Graph

A directed graph in graph theory is a graph in which the edges have a direction associated with them. The edges point from one vertex to another, and the direction of the edge indicates the relationship between the two vertices.

Here we are visualizing the association rules using vertices and edges where the vertices are annotated with item labels representing items, and rules are represented as a second set of vertices.The items are connected with itemsets/rules using arrows. The arrows pointing from items to rule vertices indicate LHS items and an arrow from a rule to an item indicates the RHS. The size of the nodes is driven by support levels, greater the support larger the size of node whereas the color of the nodes is driven by lift, greater the lift darker the color of node.

Thus, this graph-based visualization offers a very clear representation of rules but they tend to easily become cluttered and thus are only viable for very small sets of rules.

Entire Data Network Analysis

Now that the data is in a usable format, analysis can begin. In the entire dataset, there were 167 vertices (as there are 167 unique grocery products) and 14,963 edges. Because we do not know which direction any association goes at this point, we can disregard indegree and outdegree and focus on the all degree of all items instead, as all connections are undirected. A histogram of the degree of each vertex is shown in Figure 7 below.

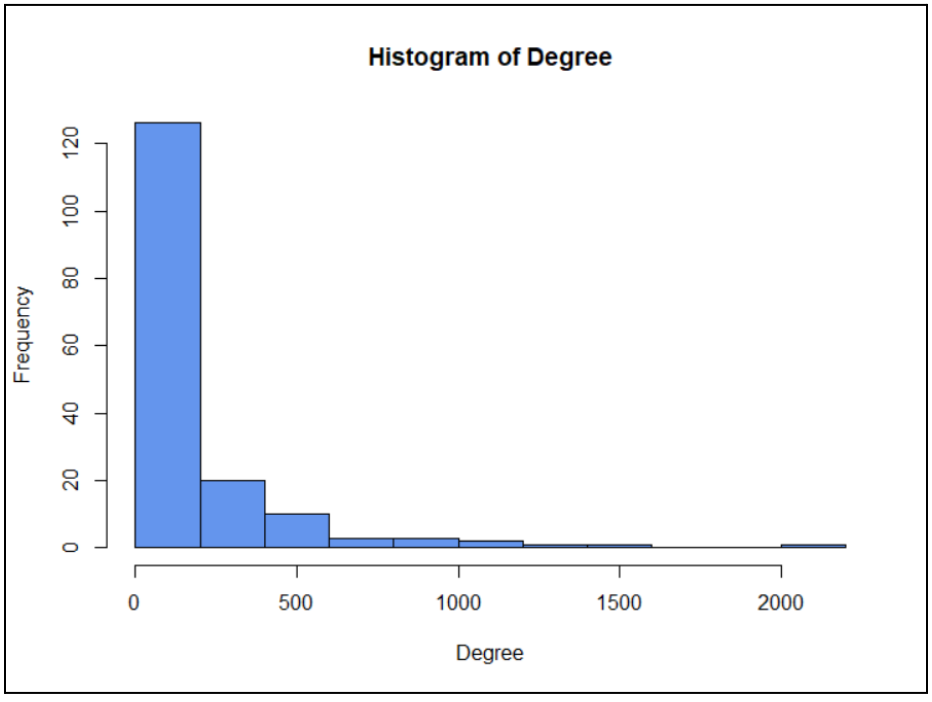

Fig 7: Histogram of Degree of Entire Dataset

This histogram shows the distribution of the amount of degrees each vertex has. As we would expect, the histogram is right skewed, as only a few products have a very large degree value, and most products have a degree value in the double digits range. The median degree value is the item vertices 62, and the mean degree value for the item vertices is 179.2 (pulled upward by the extreme values such as whole milk with a degree of 2066).

The graph below, Figure 8, shows the network analysis of the entire dataset. As we can see, there are many connections between every item vertex. The big green vertex represents whole milk in the data, as it has the largest degree of any product.

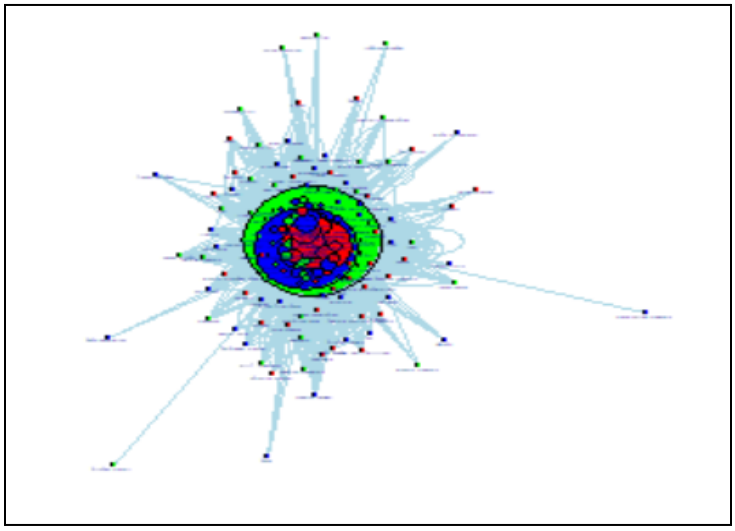

Fig 8: Entire Product Network

The item with the highest degree in the dataset is whole milk, with a degree of 2066. Because whole milk has the highest degree of any item, we performed subcomponent network analysis on the whole milk product. With so many unique products, it is difficult to see which products have the most centrality and highest degree from the above plot. However, when examining the perimeter of the plot, we can see which items are not central and have the fewest number of connections (preservation products, toilet cleaner, rubbing alcohol, etc.). These perimeter items make sense to have few connections, as they are more household items and do not seem to be connected with more frequently purchased items in any way. Since all of the data in undirected, hub and authority scores were not calculated, as there would be no difference between them. A

cluster edge betweenness graph was also produced for the entire dataset, but due to the amount of items in the dataset, it is difficult to extract meaning from it. We will revisit the cluster edge betweenness graph when we analyze a smaller subset of the data in the next section.
Centrality measures of the grocery item network were conducted, including calculating the edge density, reciprocity, closeness, and betweenness of each item. The mean distance between items is 1.75, undirected. The diameter of the network is 5, with the longest path being: bags - other vegetables - whole milk - soups - preservation products.

Subset Data Analysis

In order to better illustrate the links between grocery items, we conducted market basket analysis on a subset of items in the dataset. The items we analyzed are: sausage, soda, berries, waffles, candy, coffee, ice cream, chicken, yogurt, and ham. The same data cleaning process was applied to the subset as was done for the entire dataset. In the subsetted dataset, there are 10 vertices and 770 edges. **Figure 9** is a histogram of the degrees of each item in the subset. The item with the highest degree is soda, with 392 degrees. The mean degree value is 152 and the median is 94 degrees.

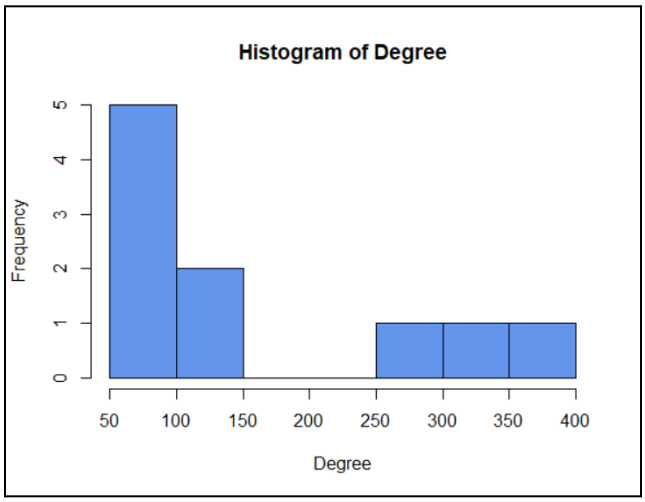

Fig 9: Histogram of Degrees for Subsetted Data

The subsetted network is displayed below in **Figure 10**. As we can see, the items that are most central in the network are soda, yogurt, and sausage. Items on the perimeter of the network are

purchased less often and thus have less connections throughout the network, such as waffles, ice cream, and candy.

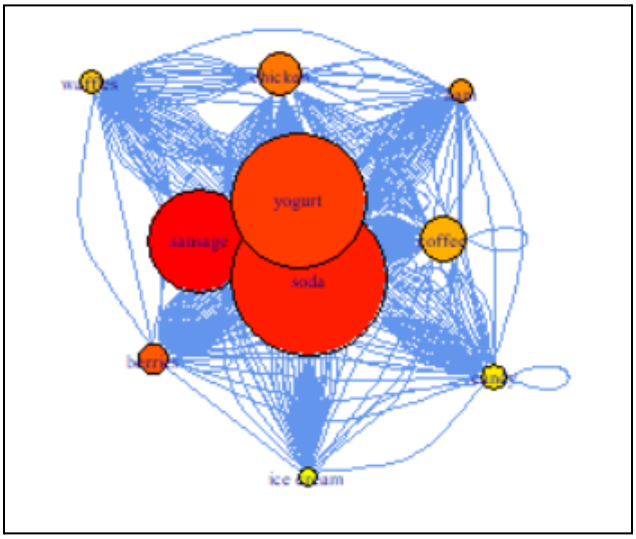

Fig 10: Subsetted Data Product Network

The diameter of this subset network is only 2, which means that every item is connected to each other in at least 1 transaction. This also means that the mean distance between items is 1. The edge density of the subset network is 17.11. Due to the limited amount of items in this subset, other centrality measures are not significant because every item is interconnected. The subset more so serves as a visualization to demonstrate how certain products are linked to each other.

Business Insights

The apriori analysis and market basket network analysis that was conducted can be used by managers and grocery store owners in order to maximize sales. Organizing stores by placing items that are frequently bought together in the same area, it can remind or entice customers to purchase both of those items. For example, it would be wise for grocery stores to have sausage and yogurt near each other, since those are linked in the apriori rules and commonly linked in the network analysis. These types of analysis can be done on a bunch of different products, and thus help grocery stores maximize their sales.

Works Cited

1. *Dedhia, Heeral. "Groceries Dataset." Kaggle, 17 Sept. 2020, https://www.kaggle.com/datasets/heeraldedhia/groceries-dataset.*