# ML_Assignment_STA380_jgscott

**Aditya Kumar, Barnana Ganguly ,Zihao Zhu, Nawen Deng**

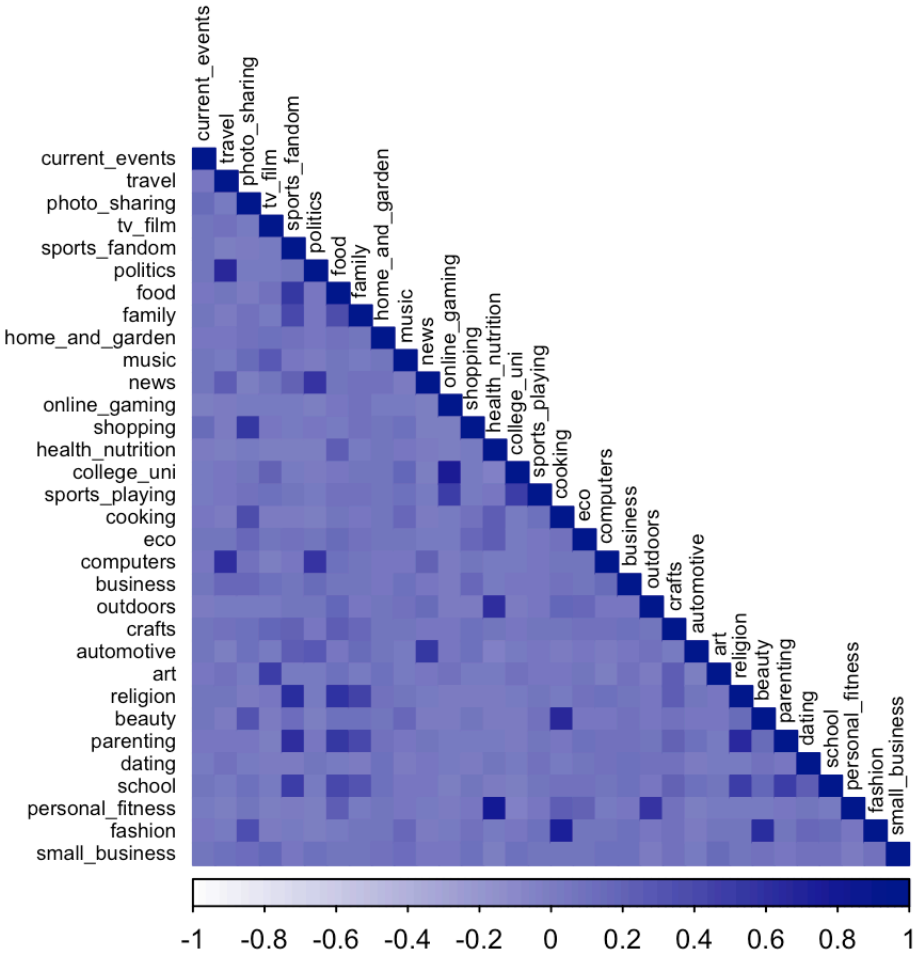
**2023-08-14**

# Market Segmentation

## (a) Data Pre-Proccessing

For the purpose of improving the significance of our market segmentation, I removed the following four categories: chatter, adult, spam, and uncategorized. Given that “adult” and “spam” categories are intended for content filtering due to their inappropriate nature, and “uncategorized” and “chatter” lack specific significance, I have excluded these four columns from the dataset.

```
## corrplot 0.92 loaded
```



There are several intutiative significant correlation between: college_uni and online gaming/ sports playing shopping and photosharing outdoor and health nutrition beauty and fashion personal fitness and health nutrition/outdoors religion and parenting/school

There are also some unexpected significant correlation between: politics and travel religion and food/sports_fandom parenting and sports_fandom fashion and cooking

Due to the limited distinguishing capacity of certain frequently occurring terms such as photo-sharing, I employed TF-IDF to reassess the significance of each term for each follower. TF, denoting term-frequency, quantifies the occurrence frequency of a term in a follower's tweets, giving greater importance to terms that appear more frequently. IDF, or inverse-document-frequency, gauges a term's occurrence across the entire dataset, reducing its significance for individual followers when it is more commonly present across the dataset

I utilize the 'cosine' metric to assess similarity, which computes the cosine value of the angle between two vectors. This approach evaluates divergence in direction rather than magnitude. For instance, with followers A, B, and C having attributes like A={'travelling':10,'cooking':5}, B={'travelling':20,'cooking':10}, C= {'travelling':10,'cooking':12}, I would perceive A as more similar to B than to C, despite A and C being 'nearer' in terms of values.

# (b) Define Market Segment

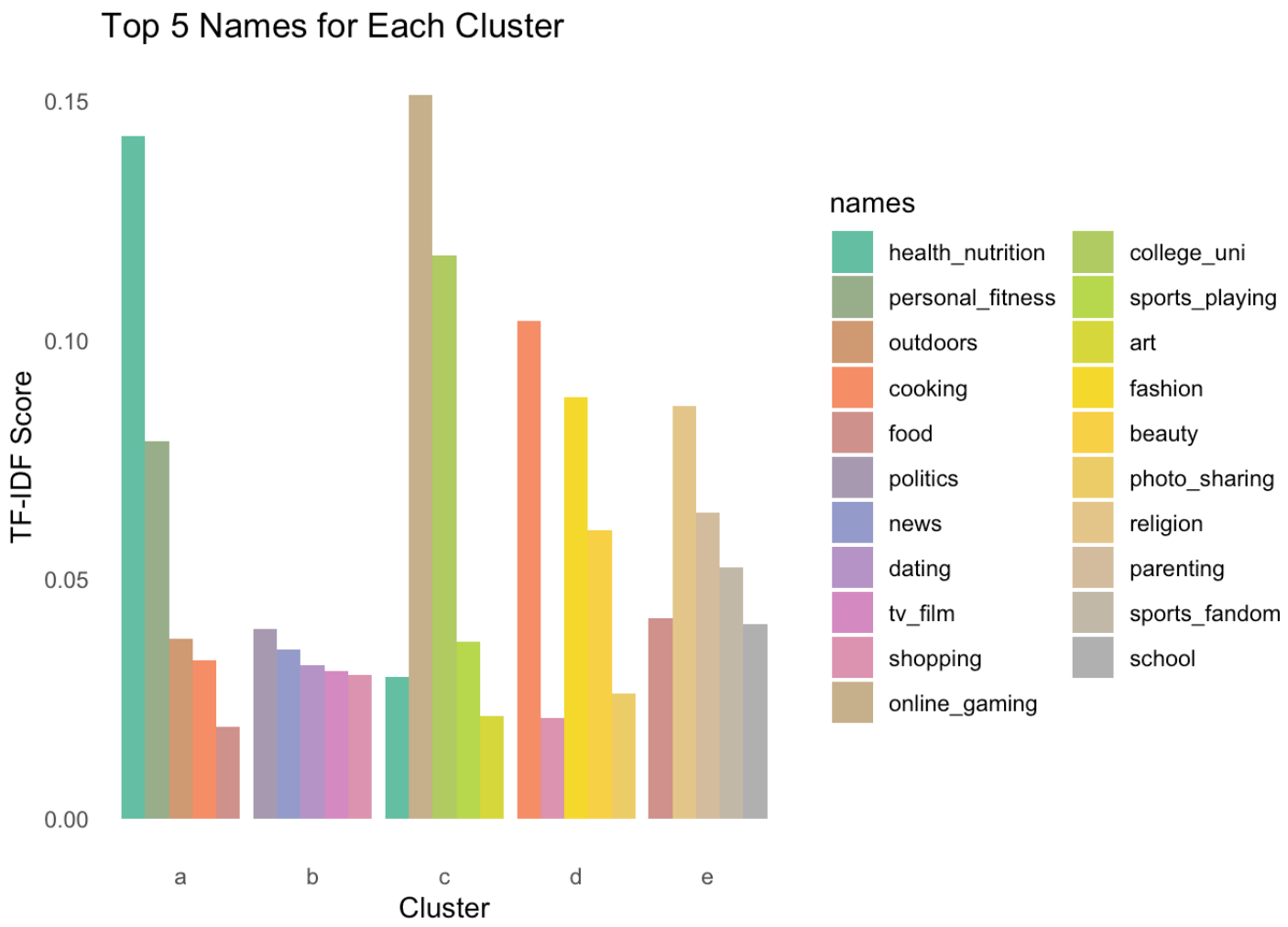
I will define a "market segment" as a cluster of correlated interests. I chose this because of the way data was collected i.e., categorizing by themes which entails commonality and would be efficient to identify their clusters

```
##
## Attaching package: 'proxy'
```

```
## The following objects are masked from 'package:stats':
##
##     as.dist, dist
```

```
## The following object is masked from 'package:base':
##
##     as.matrix
```

After evaluating various results across different values of K (using elbow graph), I opted for k=5 as our final parameter, as its outcome aligns more logically with our interpretation.




**Top 5 Names per Cluster:**

| A | B | C | D | E |
|---|---|---|---|---|
| health_nutrition | politics | online_gaming | cooking | religion |
| personal_fitness | news | college_uni | fashion | parenting |
| outdoors | dating | sports_playing | beauty | sports_fandom |
| cooking | tv_film | health_nutrition | photo_sharing | food |
| food | shopping | art | shopping | school |

Based on the topics with notable TFIDF scores within the clusters, I can deduce that the initial cluster signifies individuals with a strong emphasis on health and nutrition, likely well-educated individuals and homemakers; the second cluster encapsulates adults with religion specific inclination who are also interested in shopping and other modern practises; the third cluster embodies college/high school students; and the fourth cluster reflects those concerned with contemporary affairs, predominantly working individuals and fifth cluster captures female population who are interested in cooking, fashion, and beauty.

## (c) Marketing Strategy for Each Group:

*Cluster 1*: With an emphasis on health and nutrition, this group is potentially the fitness enthusiasts and could benefit from the company sharing nutritious cooking recipes that highlight their products. Companies can collaborate with the renowned chefs to promote their products more organically.

*Cluster 2*: With specially crafted social media efforts, this demographic, which is distinguished by its religious propensity and curiosity in contemporary activities, could be attracted. Their religious beliefs and current fashion trends might be combined to produce a novel marketing strategy.

*Cluster 3*: This segment comprises of young audience and getting college/high school students interested demands a novel strategy. To grab and hold their interest, the business should establish engrossing social media campaigns with interactive gaming components and marketing offers.

*Cluster 4*: Working class people who make up the majority of the cluster with contemporary affairs interests could be attracted to them by strategically sponsoring social activities and perhaps making pertinent political contributions. This would result in more coverage in media outlets including newspapers, television, and news websites, successfully grabbing the interest of this audience.

*Cluster 5*: This audience, which is mostly made up of women (maybe housewives) interested in food, clothing, and beauty, offers an opportunity for interesting content. The business may concentrate on giving them useful material that relates to their interests, possibly working with influencers from these fields to increase interaction.

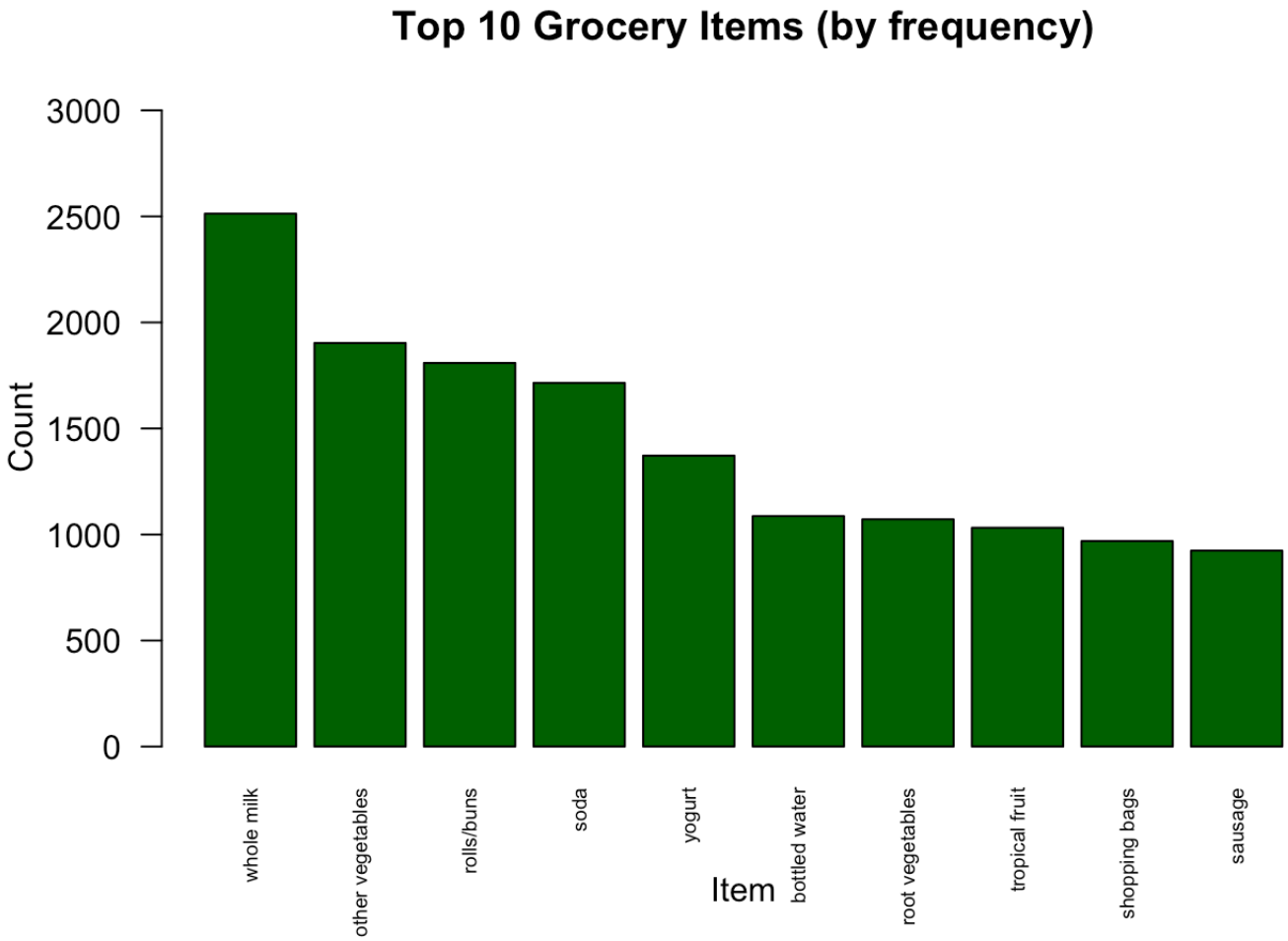
# Association Rule Mining

## (a) Data Pre-Proccessing

```
##   user              value
## 1    1        citrus fruit
## 2    1 semi-finished bread
## 3    1           margarine
## 4    1         ready soups
## 5    2      tropical fruit
## 6    2              yogurt
```

The table above shows the head of transaction dataframe before splitting it by transactions.

```
##
##       whole milk other vegetables       rolls/buns             soda
##             2513             1903             1809             1715
##           yogurt    bottled water  root vegetables   tropical fruit
##             1372             1087             1072             1032
##    shopping bags          sausage
##              969              924
```




### (b) Apriori Algorithm implementation Specifically *formatted lists of baskets* are required by the Apriori Algorithm. In this case, one “transaction” of items per person

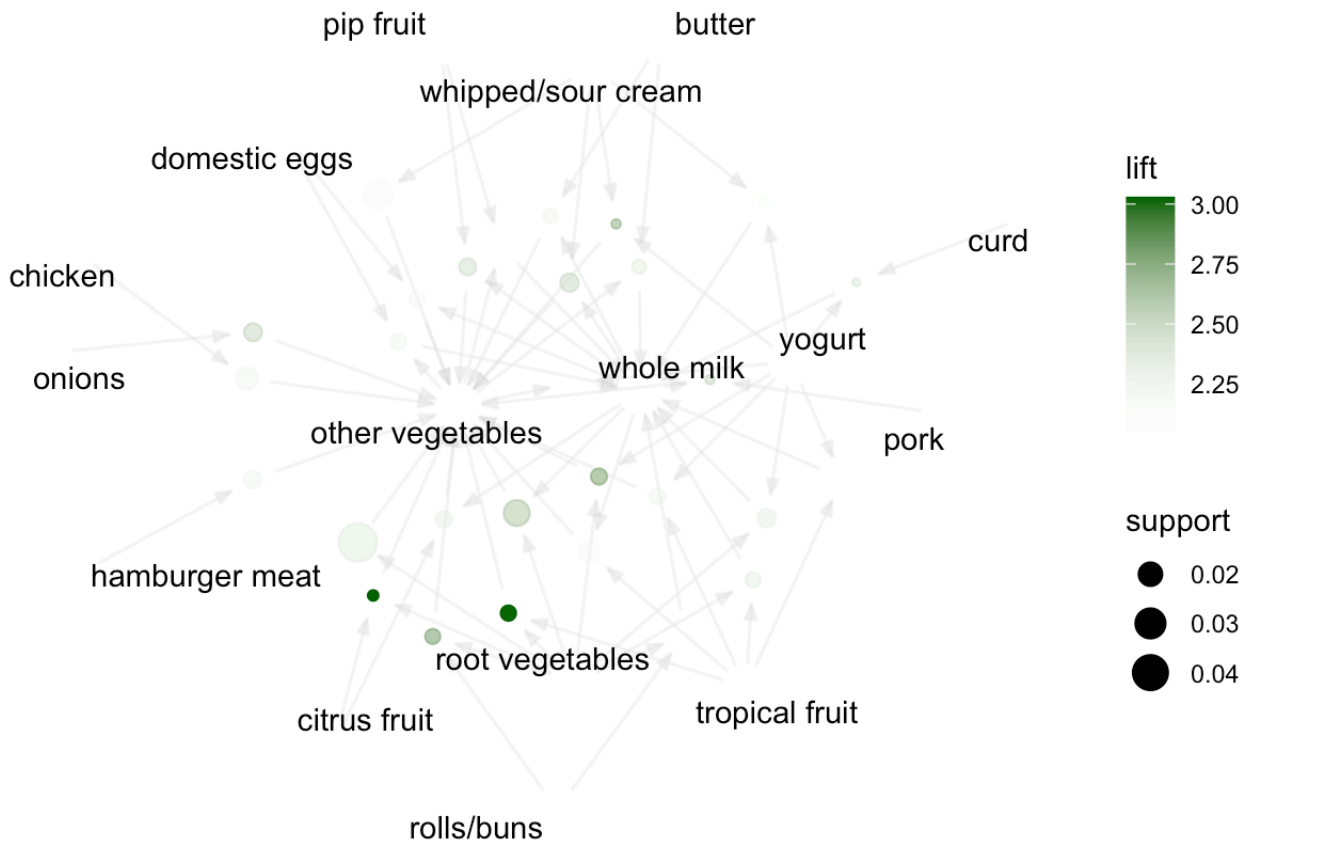
```
## [1] "Total no. of transactions are:  9835"
```

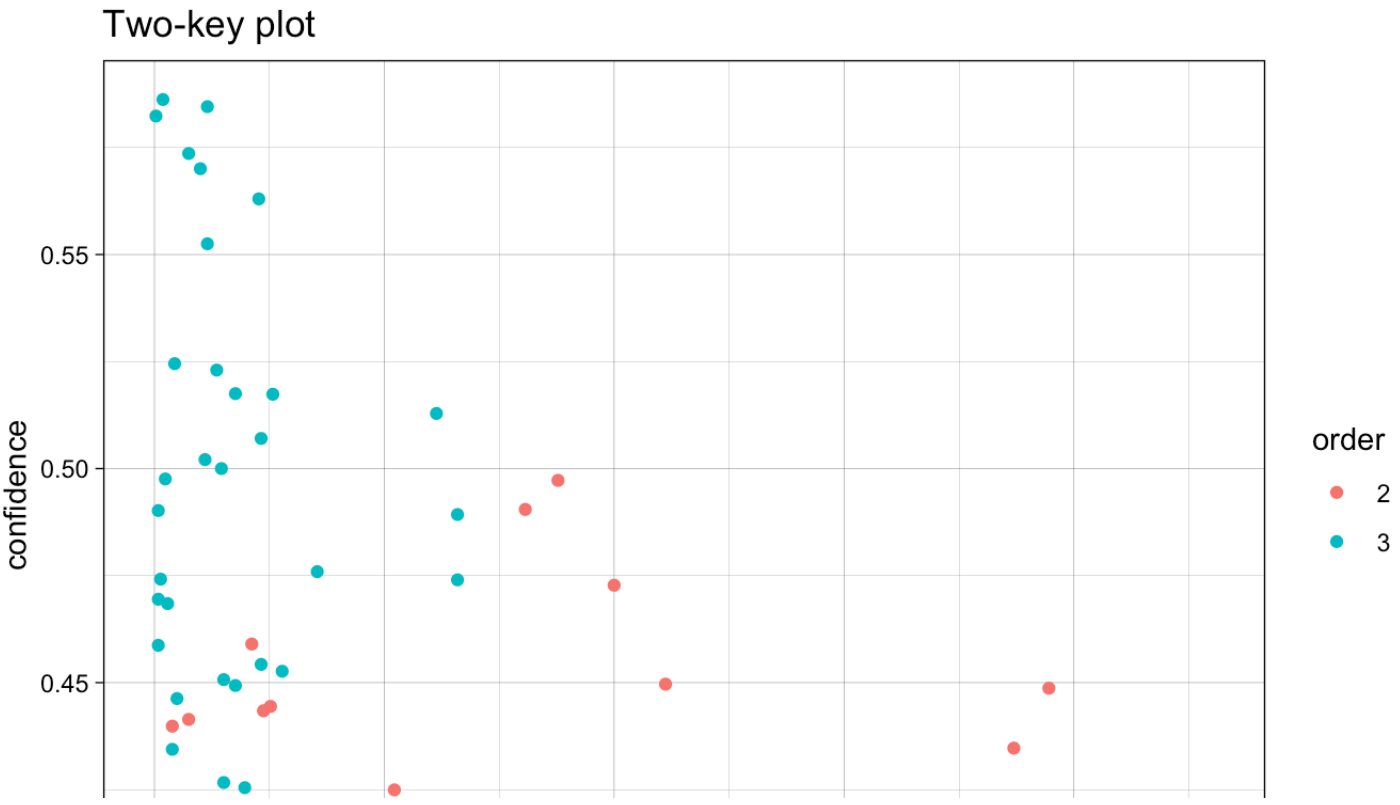
I wanted to find interesting patterns in the grocery purchase data. I tested different combinations of “support” and “confidence” values. I think of support as a measure of how often a certain combination of items appears together in all the shopping transactions. Confidence measures how likely one item is to be bought if another item is already in the basket. My goal was to find a combination of support and confidence values that gave the highest “average lift.” Lift means how much more likely items are to be bought together than if they were bought independently. In simple terms, high lift values suggest a strong connection between items.

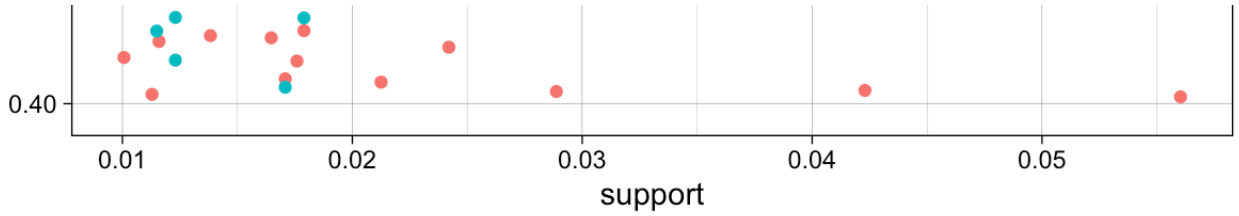
I tested different support values from 0.009 to 0.05 and confidence values from 0.2 to 0.5. I wanted to find the values that provided the highest lift, indicating a strong connection between items in the shopping baskets. The results showed that the best combination was a support of 0.009 and a confidence of 0.5, with an average lift of 2.2255. However, there's a trade-off: if you increase support, you capture more transactions, but the lift might decrease, showing a weaker connection between items. I decided to balance these factors by choosing a slightly higher support of 0.01 and a slightly lower confidence of 0.4. This balance allowed for more transactions while still maintaining a reasonable lift value.

Now, I ran the association rule model using specific support and confidence values. Then, from the results, I selected rules with a lift greater than 2, as the average lift was close to 2. This resulted in 29 strong rules of association. Among these rules, the most common item purchased was "whole milk," followed by "other vegetables." This suggests that many people across different shopping baskets are consistently interested in buying whole milk and/or other vegetables.

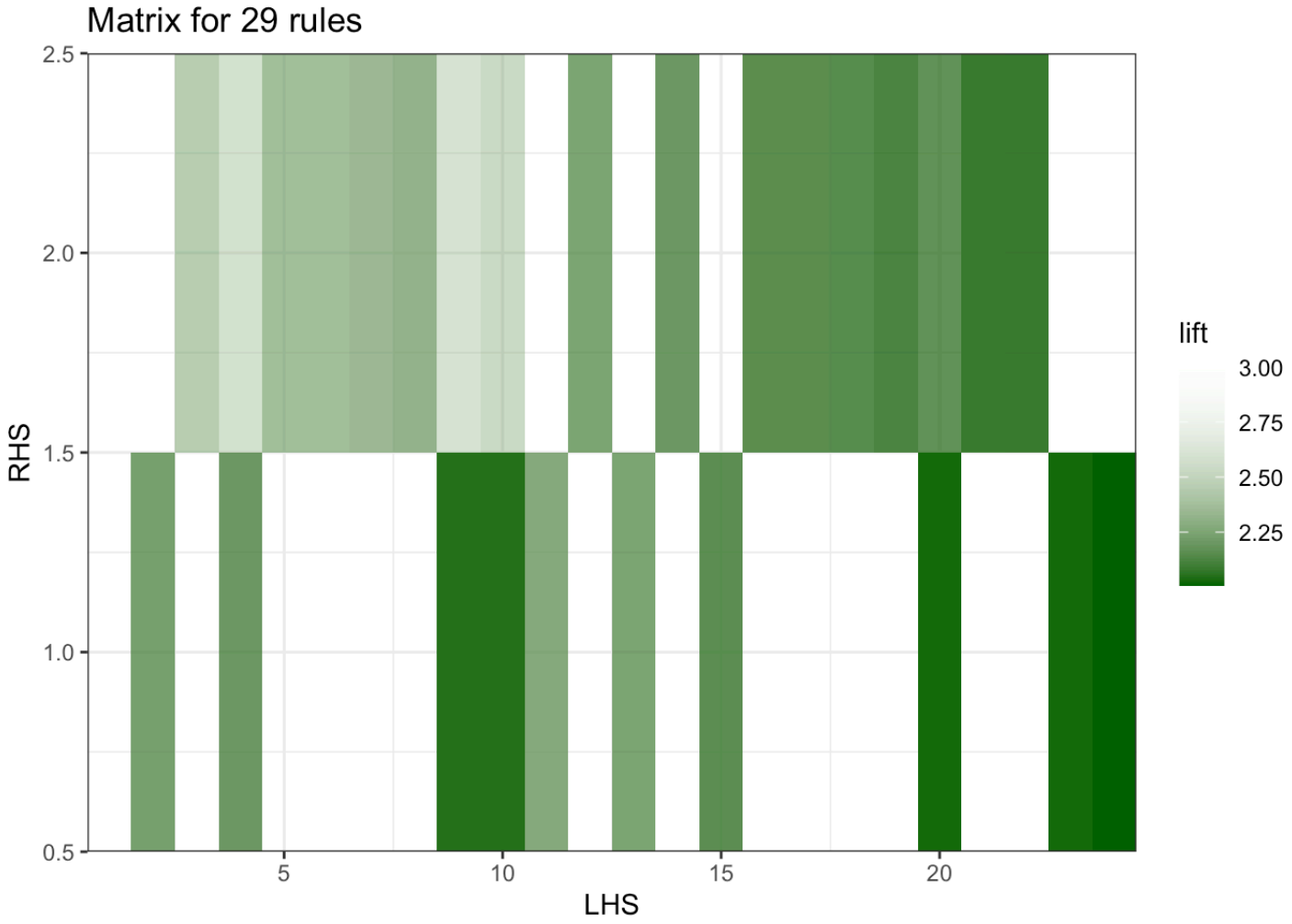
```
## Available control parameters (with default values):
## layout    =  stress
## circular  =  FALSE
## ggraphdots    =  NULL
## edges     =  <environment>
## nodes     =  <environment>
## nodetext  =  <environment>
## colors    =  c("#EE0000FF", "#EEEEEEFF")
## engine    =  ggplot2
## max   =  100
## verbose   =  FALSE
```

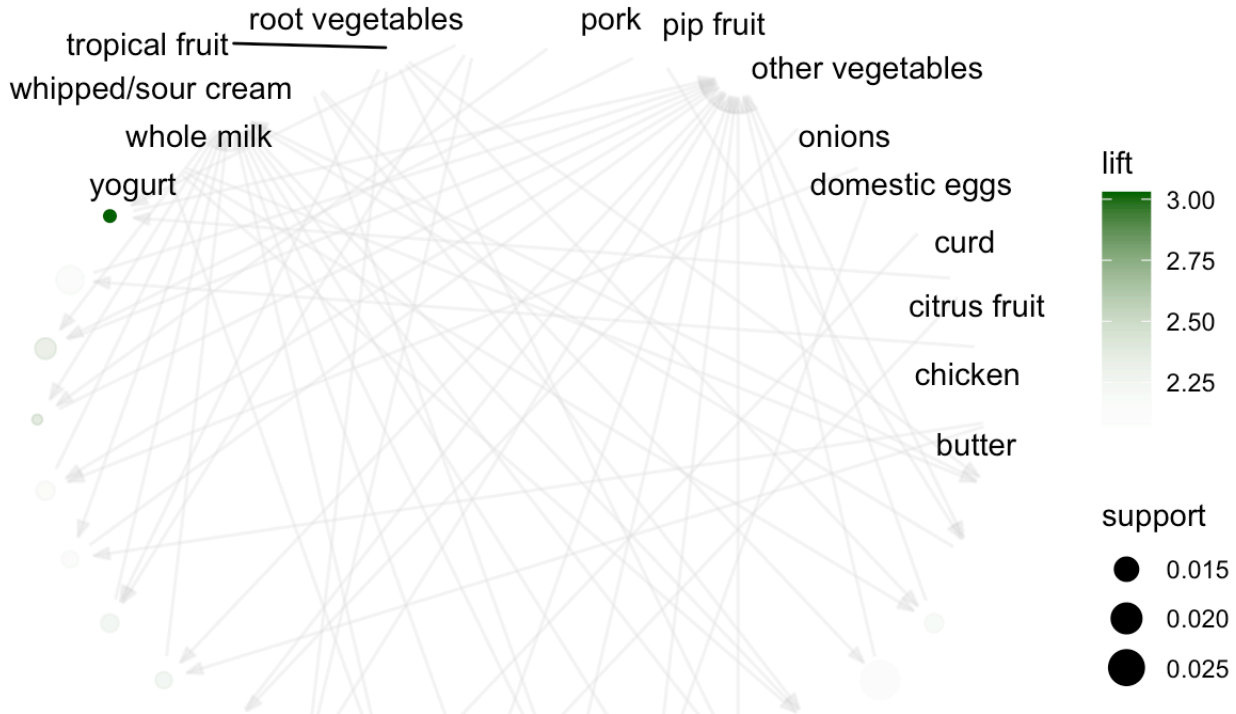

pip fruit
butter
whipped/sour cream
domestic eggs
curd
chicken
yogurt
onions
whole milk
other vegetables
pork
hamburger meat
root vegetables
citrus fruit
tropical fruit
rolls/buns
lift
3.00
2.75
2.50
2.25
support
0.02
0.03
0.04


Two-key plot
confidence
0.55
0.50
0.45
order
2
3




```
## Itemsets in Antecedent (LHS)
##  [1] "{citrus fruit,root vegetables}"   "{root vegetables,tropical fruit}"
##  [3] "{root vegetables,whole milk}"     "{root vegetables,yogurt}"
##  [5] "{onions}"                         "{pork,whole milk}"
##  [7] "{whipped/sour cream,whole milk}"  "{pip fruit,whole milk}"
##  [9] "{rolls/buns,root vegetables}"     "{whipped/sour cream,yogurt}"
## [11] "{curd,yogurt}"                    "{root vegetables}"
## [13] "{butter,other vegetables}"        "{citrus fruit,whole milk}"
## [15] "{domestic eggs,other vegetables}" "{chicken}"
## [17] "{butter,whole milk}"              "{hamburger meat}"
## [19] "{domestic eggs,whole milk}"       "{tropical fruit,yogurt}"
## [21] "{tropical fruit,whole milk}"      "{whipped/sour cream}"
## [23] "{other vegetables,pip fruit}"     "{other vegetables,yogurt}"
## Itemsets in Consequent (RHS)
## [1] "{whole milk}"       "{other vegetables}"
```


Matrix for 29 rules
RHS
LHS
lift
3.00
2.75
2.50
2.25
2.5
2.0
1.5
1.0
0.5
5
10
15
20


root vegetables
tropical fruit
whipped/sour cream
whole milk
yogurt
pork
pip fruit
other vegetables
onions
domestic eggs
curd
citrus fruit
chicken
butter
lift
3.00
2.75
2.50
2.25
support
0.015
0.020
0.025

The visualizations above gives us the strength f the associations. The first graph illustrates the relative value of the different basket elements. With branches stretching forth to other products, the center section contains whole milk and other veggies that used to be the most popular.

The next one provides a two-key plot for the entire set of values as a function of support and confidence, not only for the subset.

The 3rd graph is a matrix representation of the rule matrix, with the lift indicated by a color scale. These can be matched to the lift values mentioned above, giving me the precise basket contents.

# (c) Choice of parameters

Higher degrees of support provided too few rules for us to examine, so I went with support= 0.009 instead. Because I want to ensure that item on LHS will also appear if item on RHS appears, I set confidence = 0.5. This solely takes into account how well-liked goods on RHS are, not those on LHS. There is a higher likelihood that things on the RHS will contain items on the LHS if items on the rhs appear frequently overall.

I choose our final itemlists based on lift since lift gauges the likelihood that an item on LHS will be bought when an item on RHS is bought in order to account for this bias. The maximum average lift for these selected values of support and lift is 2.2255. Therefore, I sorted the items by lift and rank the top 20 rules, which is the result generated by the algorithm.

# (d) Recommendation

This data-driven analysis provides invaluable insights for store managers responsible for perishable inventory management. It also holds significance for optimizing product placement tactics. By grouping frequently co-purchased items on the same shelf or aisle, retailers can enhance customer convenience and drive supplementary purchases, leading to heightened sales and improved customer satisfaction.