ML - Association Rule Mining

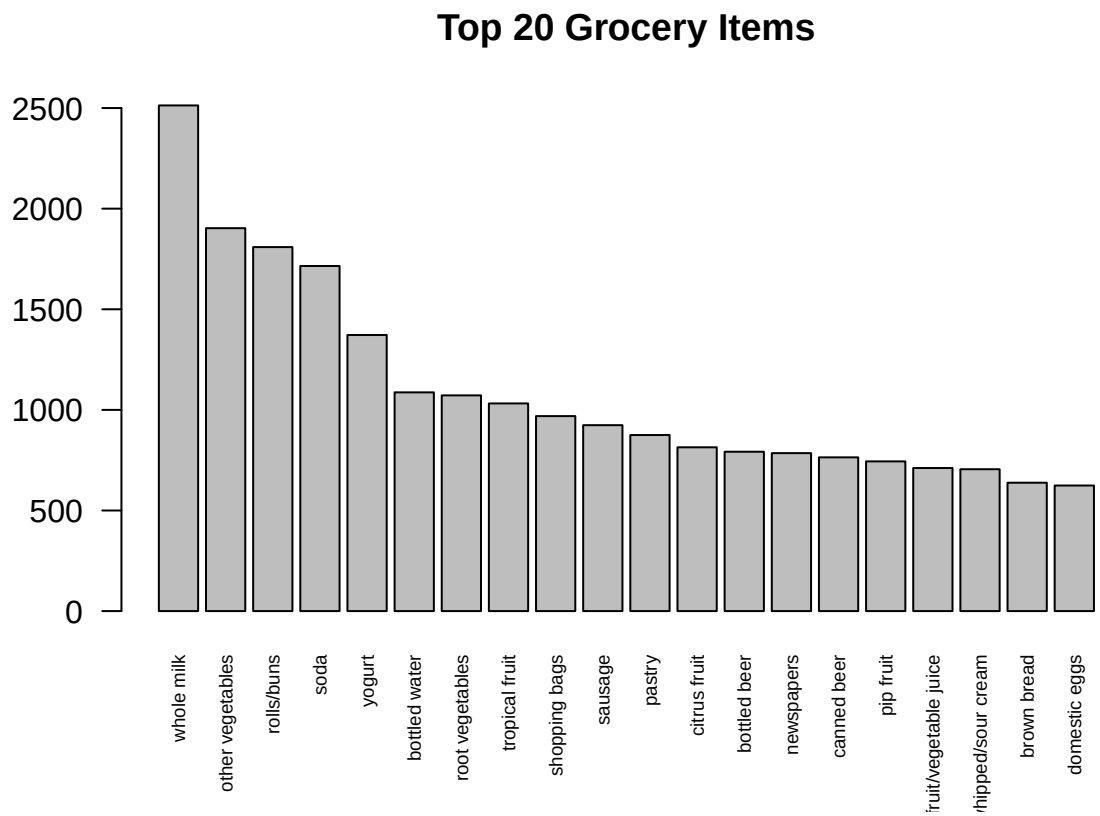
Mauro Cardona Reyes

2024-08-15

```
library(tidyverse)
library(igraph)
library(arules)  # has a big ecosystem of packages built around it
library(arulesViz)
library(igraph)
library(knitr)
```

Top 20 items across all baskets

```
##
##            whole milk      other vegetables            rolls/buns
##                  2513                  1903                  1809
##                  soda                yogurt         bottled water
##                  1715                  1372                  1087
##       root vegetables        tropical fruit         shopping bags
##                  1072                  1032                   969
##              sausage                pastry          citrus fruit
##                   924                   875                   814
##          bottled beer           newspapers           canned beer
##                   792                   785                   764
##             pip fruit fruit/vegetable juice    whipped/sour cream
##                   744                   711                   705
##           brown bread         domestic eggs
##                   638                   624
```



Understand data

```
summary(grocery_transactions)
print(grocery_transactions)
inspect(grocery_transactions[1:5])
```

First set of rules

support: 0.001 confidence: 0.01 Max length: 4 Lift: No lift yet

```
# Step 6: Association Rule Mining
min_support <- 0.001
min_confidence <- 0.01

grocery_rules <- apriori(grocery_transactions,
                parameter = list(supp = min_support, conf = min_confidence, maxlen= 4 ))
```

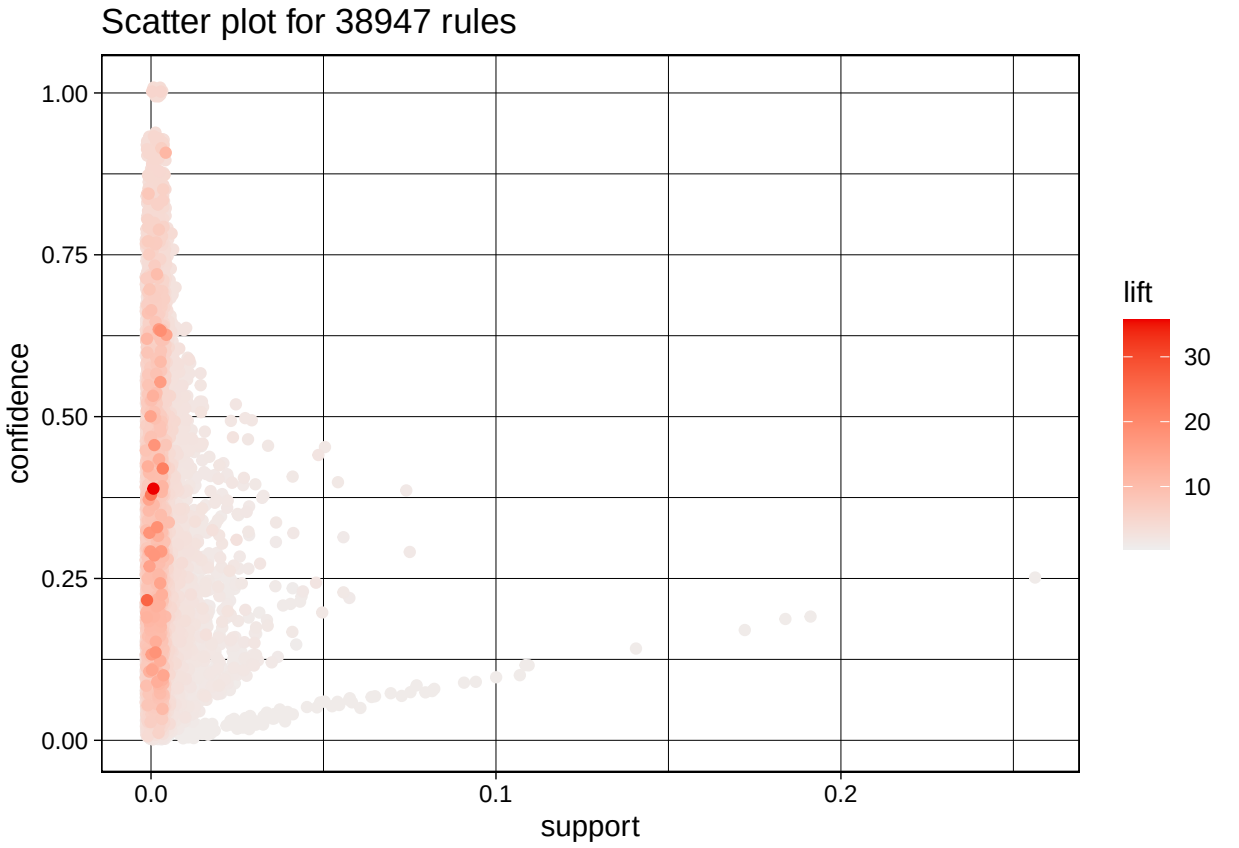
Inspect first set of rules

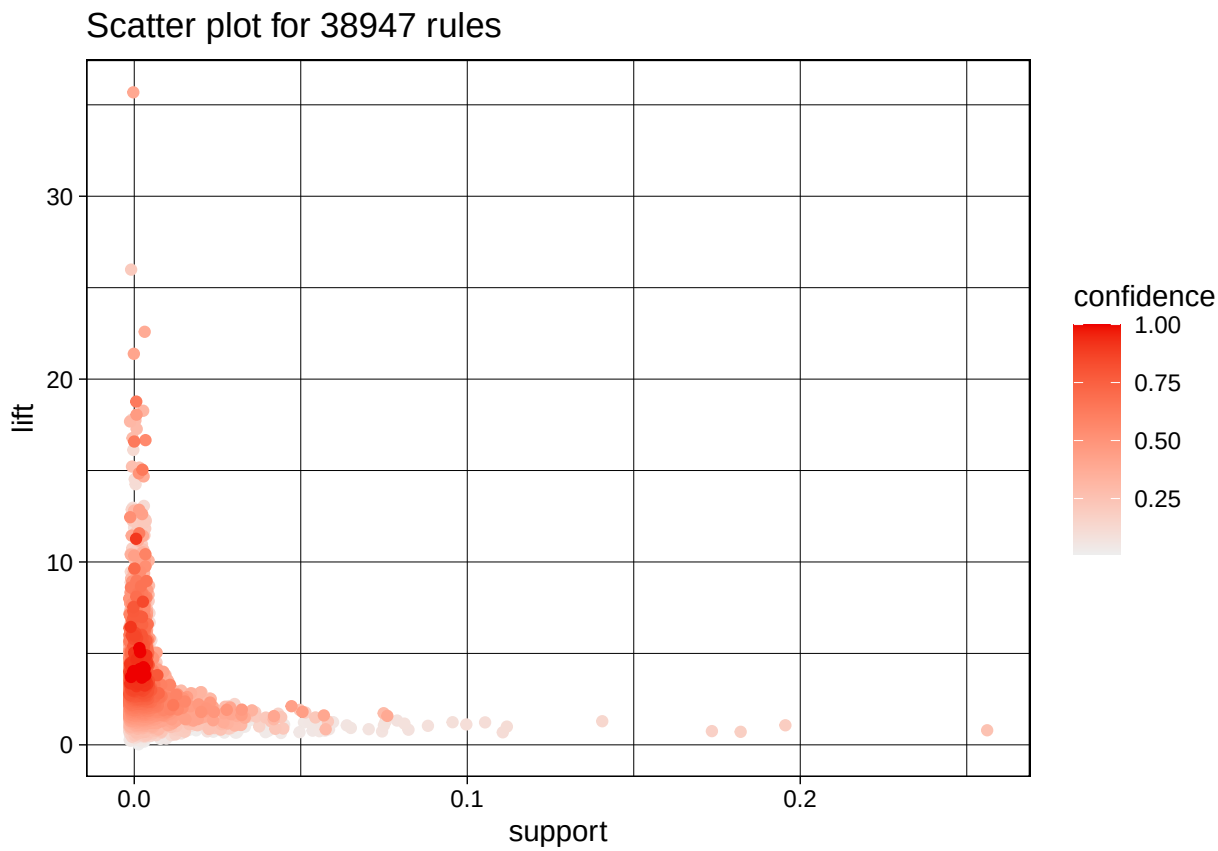
```
# The rules generated
inspect(grocery_rules) # 39,000 rules, with support .001, conf .01, too many to print
```

Too big, 39,000 rows, but good for graph below:

Graphs for first set of rules

These graphs are for visualizing as much of the data as possible, and then based of these graphs we pick a subset of the data to look at





Look at subsets based off graphs above

```
inspect(subset(grocery_rules, support > 0.05)) # 34 rules

inspect(subset(grocery_rules, support > 0.01)) #610 rules

inspect(subset(grocery_rules, confidence > 0.75)) #458
```

Most relevant subset:

There are 78, only 30 printed. Included to show evidence of inspection

```
relevant_subset = subset(grocery_rules, lift > 10) #78 rules

# Convert to a data frame
relevant_subset_df <- as(relevant_subset, "data.frame")

# Remove the 'coverage' column
relevant_subset_df <- relevant_subset_df[, !names(relevant_subset_df) %in% "coverage"]

# Display only the first 20 rules of this subset
kable(head(relevant_subset_df, 78), format = "markdown", col.names = c("Rule", "Support", "Confidence",
```

	Rule	Support	Confidence	Lift	Count
204	{softener} => {detergent}	0.0011185	0.2037037	10.60014	11
205	{detergent} => {softener}	0.0011185	0.0582011	10.60014	11
274	{liquor} => {red/blush wine}	0.0021352	0.1926606	10.02548	21
275	{red/blush wine} => {liquor}	0.0021352	0.1111111	10.02548	21
536	{Instant food products} => {hamburger meat}	0.0030503	0.3797468	11.42144	30
537	{hamburger meat} => {Instant food products}	0.0030503	0.0917431	11.42144	30
1131	{mayonnaise} => {mustard}	0.0014235	0.1555556	12.96516	14
1132	{mustard} => {mayonnaise}	0.0014235	0.1186441	12.96516	14
5922	{liquor,red/blush wine} => {bottled beer}	0.0019319	0.9047619	11.23527	19
5923	{bottled beer,liquor} => {red/blush wine}	0.0019319	0.4130435	21.49356	19
5924	{bottled beer,red/blush wine} => {liquor}	0.0019319	0.3958333	35.71579	19
5971	{popcorn,soda} => {salty snack}	0.0012201	0.6315789	16.69779	12
5972	{salty snack,soda} => {popcorn}	0.0012201	0.1304348	18.06797	12
6016	{Instant food products,soda} => {hamburger meat}	0.0012201	0.6315789	18.99565	12
6017	{hamburger meat,soda} => {Instant food products}	0.0012201	0.2105263	26.20919	12
6019	{Instant food products,rolls/buns} => {hamburger meat}	0.0010168	0.4347826	13.07672	10
6020	{hamburger meat,rolls/buns} => {Instant food products}	0.0010168	0.1176471	14.64631	10
6022	{Instant food products,whole milk} => {hamburger meat}	0.0015252	0.5000000	15.03823	15
6023	{hamburger meat,whole milk} => {Instant food products}	0.0015252	0.1034483	12.87866	15
6191	{sugar,whole milk} => {rice}	0.0012201	0.0810811	10.63243	12
6200	{butter,root vegetables} => {rice}	0.0010168	0.0787402	10.32546	10
6215	{fruit/vegetable juice,root vegetables} => {rice}	0.0011185	0.0932203	12.22429	11
7272	{ham,processed cheese} => {white bread}	0.0019319	0.6333333	15.04549	19
7273	{processed cheese,white bread} => {ham}	0.0019319	0.4634146	17.80345	19
7274	{ham,white bread} => {processed cheese}	0.0019319	0.3800000	22.92822	19
7276	{fruit/vegetable juice,processed cheese} => {ham}	0.0011185	0.3793103	14.57233	11
7277	{fruit/vegetable juice,ham} => {processed cheese}	0.0011185	0.2894737	17.46610	11
7280	{ham,soda} => {processed cheese}	0.0010168	0.2040816	12.31376	10
7291	{domestic eggs,processed cheese} => {white bread}	0.0011185	0.5238095	12.44364	11
7292	{domestic eggs,white bread} => {processed cheese}	0.0011185	0.1929825	11.64406	11
7297	{pip fruit,processed cheese} => {white bread}	0.0010168	0.4347826	10.32871	10
7301	{tropical fruit,white bread} => {processed cheese}	0.0015252	0.1744186	10.52397	15
7304	{soda,white bread} => {processed cheese}	0.0017285	0.1683168	10.15580	17
7307	{rolls/buns,white bread} => {processed cheese}	0.0013218	0.2031250	12.25604	13
8058	{baking powder,flour} => {sugar}	0.0010168	0.5555556	16.40807	10
8059	{baking powder,sugar} => {flour}	0.0010168	0.3125000	17.97332	10
8060	{flour,sugar} => {baking powder}	0.0010168	0.2040816	11.53530	10
8065	{baking powder,margarine} => {sugar}	0.0011185	0.3666667	10.82933	11
8066	{margarine,sugar} => {baking powder}	0.0011185	0.2037037	11.51394	11
8069	{domestic eggs,sugar} => {baking powder}	0.0010168	0.2040816	11.53530	10
8072	{sugar,whipped/sour cream} => {baking powder}	0.0013218	0.2708333	15.30831	13
8254	{curd,flour} => {sugar}	0.0011185	0.3548387	10.48000	11
8255	{curd,sugar} => {flour}	0.0011185	0.3235294	18.60767	11
8257	{flour,margarine} => {sugar}	0.0016268	0.4324324	12.77169	16
8258	{margarine,sugar} => {flour}	0.0016268	0.2962963	17.04137	16
8261	{sugar,whipped/sour cream} => {flour}	0.0010168	0.2083333	11.98221	10
8264	{citrus fruit,sugar} => {flour}	0.0010168	0.2127660	12.23715	10
8267	{root vegetables,sugar} => {flour}	0.0014235	0.2222222	12.78103	14

	Rule	Support	Confidence	Lift	Count
8269	{flour,soda} => {sugar}	0.0011185	0.3928571	11.60285	11
8273	{sugar,yogurt} => {flour}	0.0013218	0.1911765	10.99544	13
8279	{sugar,whole milk} => {flour}	0.0028470	0.1891892	10.88114	28
10508	{dessert,pip fruit} => {butter milk}	0.0014235	0.2857143	10.21818	14
11165	{sliced cheese,whipped/sour cream} => {ham}	0.0010168	0.2631579	10.10999	10
11167	{ham,pip fruit} => {sliced cheese}	0.0010168	0.2564103	10.46388	10
11203	{fruit/vegetable juice,ham} => {white bread}	0.0016268	0.4210526	10.00254	16
26411	{other vegetables,root vegetables,tropical fruit} => {turkey}	0.0012201	0.0991736	12.19215	12
26423	{butter,root vegetables,whole milk} => {rice}	0.0010168	0.1234568	16.18930	10
26431	{other vegetables,root vegetables,yogurt} => {rice}	0.0014235	0.1102362	14.45564	14
26435	{root vegetables,whole milk,yogurt} => {rice}	0.0014235	0.0979021	12.83823	14
26439	{other vegetables,root vegetables,whole milk} => {rice}	0.0018302	0.0789474	10.35263	18
26495	{curd,root vegetables,whole milk} => {herbs}	0.0012201	0.1967213	12.09221	12
26527	{bottled water,citrus fruit,whole milk} => {herbs}	0.0010168	0.1724138	10.59806	10
26551	{other vegetables,root vegetables,shopping bags} => {herbs}	0.0011185	0.1692308	10.40240	11
26611	{soda,white bread,whole milk} => {processed cheese}	0.0010168	0.2500000	15.08436	10
26698	{flour,root vegetables,whole milk} => {sugar}	0.0010168	0.3448276	10.18432	10
26699	{root vegetables,sugar,whole milk} => {flour}	0.0010168	0.2941176	16.91606	10
26703	{other vegetables,sugar,whole milk} => {flour}	0.0012201	0.1935484	11.13186	12
26707	{margarine,other vegetables,yogurt} => {flour}	0.0010168	0.1785714	10.27047	10
26719	{root vegetables,whipped/sour cream,whole milk} => {flour}	0.0017285	0.1827957	10.51343	17
26791	{domestic eggs,other vegetables,yogurt} => {soft cheese}	0.0011185	0.1929825	11.29751	11
26883	{citrus fruit,fruit/vegetable juice,tropical fruit} => {grapes}	0.0011185	0.2820513	12.60897	11
27102	{hard cheese,whipped/sour cream,yogurt} => {butter}	0.0010168	0.5882353	10.61522	10
27103	{butter,whipped/sour cream,yogurt} => {hard cheese}	0.0010168	0.2631579	10.73924	10
27359	{chocolate,rolls/buns,soda} => {candy}	0.0012201	0.3000000	10.03571	12
27603	{pip fruit,sausage,yogurt} => {sliced cheese}	0.0012201	0.3076923	12.55665	12
27707	{coffee,other vegetables,yogurt} => {oil}	0.0010168	0.2857143	10.18116	10
27735	{citrus fruit,fruit/vegetable juice,root vegetables} => {oil}	0.0010168	0.2941176	10.48061	10
28226	{hamburger meat,whipped/sour cream,yogurt} => {butter}	0.0010168	0.6250000	11.27867	10

Support 0.05% and above

Things to understand:

A big difference between this data set and the playlist data set is how frequently items appear. In this data set, the support at 0.05% only draws 34 rules (appears in 5% of baskets or more). Of these rules, 28 of them are just an item by itself! This gives us insight into the grocery shopping of individuals, and suggests people have very different grocery shopping lists. Besides these 28 items, there is a lot of diversity in what people buy.

These 28 items included: canned beef, coffee, beef, napkins, pork, newspapers, domestic eggs, butter, fruit/vegetable juice, pastry, bottled water, soda, yogurt, rolls/buns, other vegetables, milk, etc.

These are very common food items, so it makes sense that they appear constantly throughout the data. Milk has the highest support of 25%, which makes sense as our graph earlier had milk as the most common food

item.

The other rules included:

yogurt –> whole milk

other vegetables –> whole milk

rolls/buns –> whole milk

and the opposites of these as well! Notice that these are the top 3 items that appeared across all baskets, and their lifts stay between 1.2 to 1.5. This suggests that these associations are not at all powerful, and their appearence here is more than likely because they appear throughout all baskets, not because they have associations with each other.

The difference between the the grocery data and playlist data is why in the first graph above, I choose a support of 0.001 and confidence of 0.01. I wanted to capture many rules (graph captured 39,000 rules!), so I could find interesting patterns and then look at the subsets.

Confidence 75% and above

when looking at confidence of 75% and above, we find there are 458 rules. This is interesting, as it represents the probability of the right hand side showing up based on what is in the left hand side. However, we may run into an issue. If something is very common like milk, there is a chance that most products will have a high confidence when milk is on the right side! This doesn't necessarily indicate a strong association!

Therefore, when looking for rules it is better to just look at the lift (which tells us how much *more likely* an item is to be bought if it appears with another, and accounts for frequency of the right hand side)

However, some of the 458 rules with conf above 75% include:

Citrus fruit, fruit/vegetable juice, grapes –> tropical fruit (lift of 8!)

frozen meals, tropical fruit, yogurt –> whole milk (lift of 3)

When looking through the 75% confidence rules, my hypothesis came out true! As I scrolled down, the majority of the right hand side was milk! and the lift was around 3 on average for those which contained milk, suggesting that these rules weren't very meaningful. A similar pattern was discovered for other vegetables.

I included above one rule which had a lift of 8 (suggesting it is genuinely strong!) and one rule with milk on the right hand side and a lift of 3.

Notice that for the first rule, the foods make sense to be bought together. Clearly, the person is buying things to make something fruity! However, for the milk rule, notice that frozen meals, tropical fruit, yogurt, and then milk, don't really create an interesting meaningful rule. Rather, it seems to be a conglomerate of random things.

Lift of 10+

When looking at a lift above 10 (i.e., if item on left hand side appears, the item on the right hand side is 10 times as likely to appear), we find 78 rules. I decided on the number ten based on insights from the graph, and when we look at these 78 rules together, we find foods/items that tend to go together! For example:

softener –> detergent (lift: 10)

popcorn, soda –> salty snack (16)

ham, bread –> cheese (lift: 22)

bottled beer, red/blush wine –> liquor (35)

and much more!

These rules make *way* more sense than just looking at the confidence. It has found associations between cleaning products, alcohol, popcorn, and sandwich purchases that make perfect sense and by intuition are quite common!

Something else I noticed is that quite a few of the rules are 2-3 items together and the inclusion of a 4th. A lot of times this 4th item doesn't have as close of an association with the other items, BUT it is considered a necessity, such:

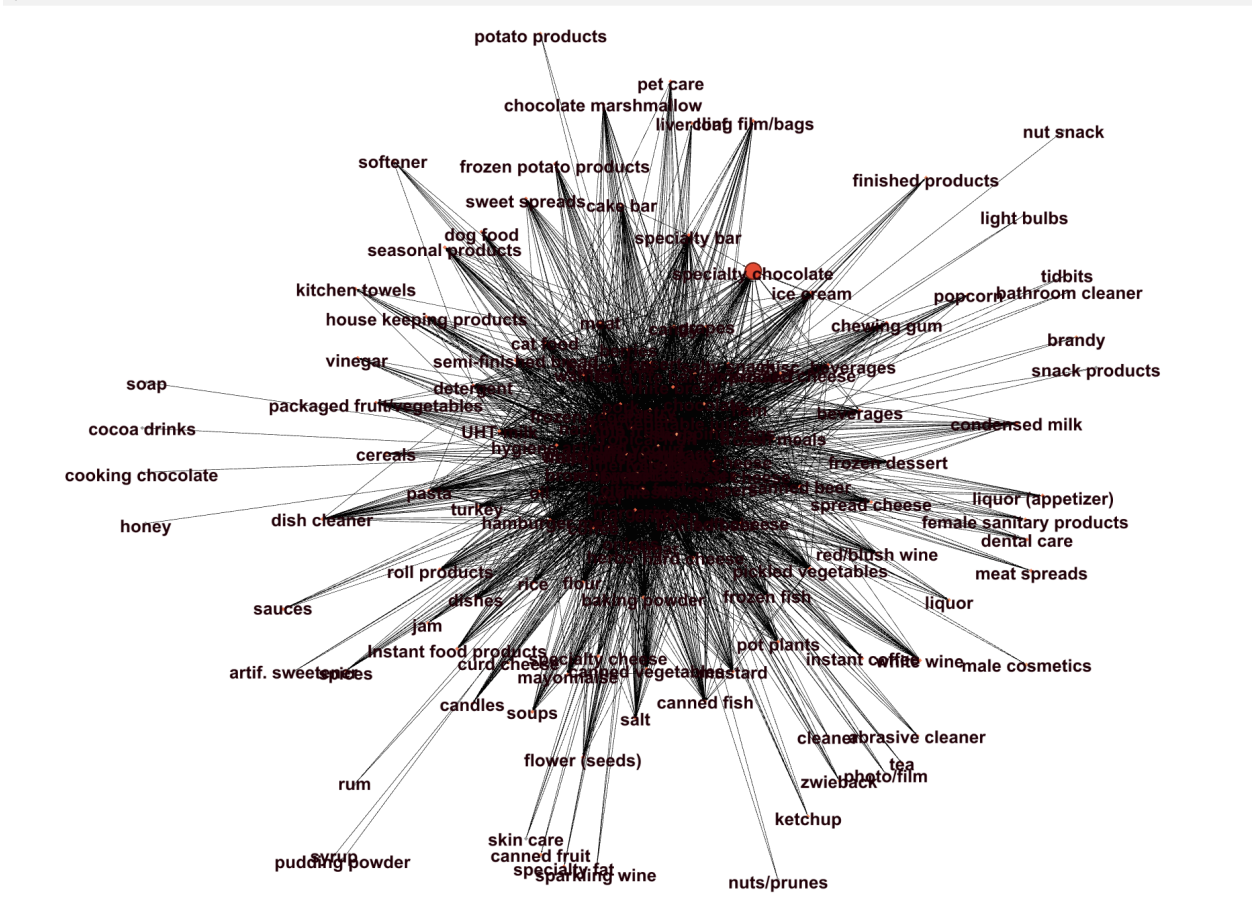
Butter, root vegetables, and whole milk –> rice

I personally do not necessarily see a connection with these foods, however they are considered essential household foods. Perhaps, when we get rules like these, it suggests that when people buy a lot of food, they tend to buy many necesities, etc.

First Igraph code

(Turned out pretty ugly)

```
# Convert grocery_rules to Graph and Export for Gephi
grocery_graph <- associations2igraph(grocery_rules, associationsAsNodes = FALSE)
igraph::write_graph(grocery_graph, file='grocery_rules.graphml', format = "graphml")
```



Above, we see an igraph based on the first settings, which were very broad. Turned out very ugly, decided to get a better subset before manipulating the igraph

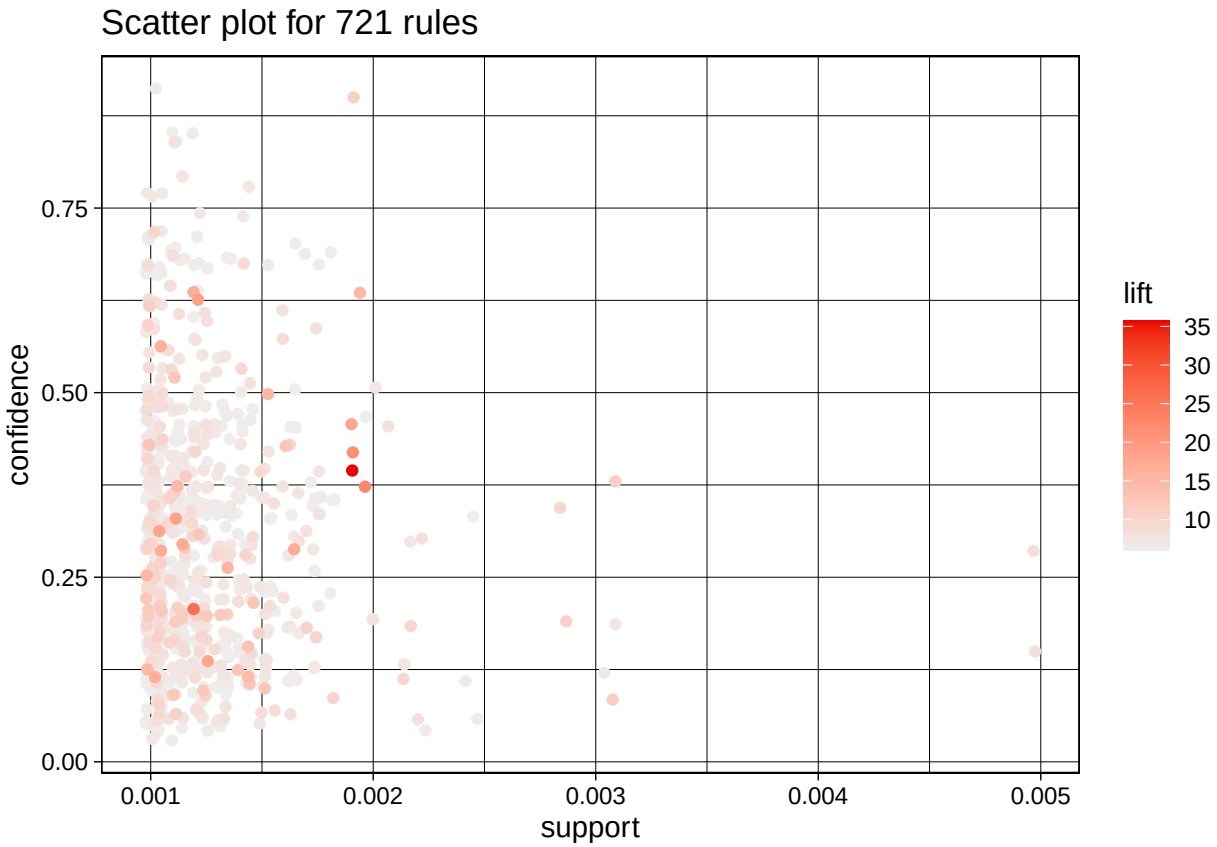
Second Rules Based on Previous Graphs

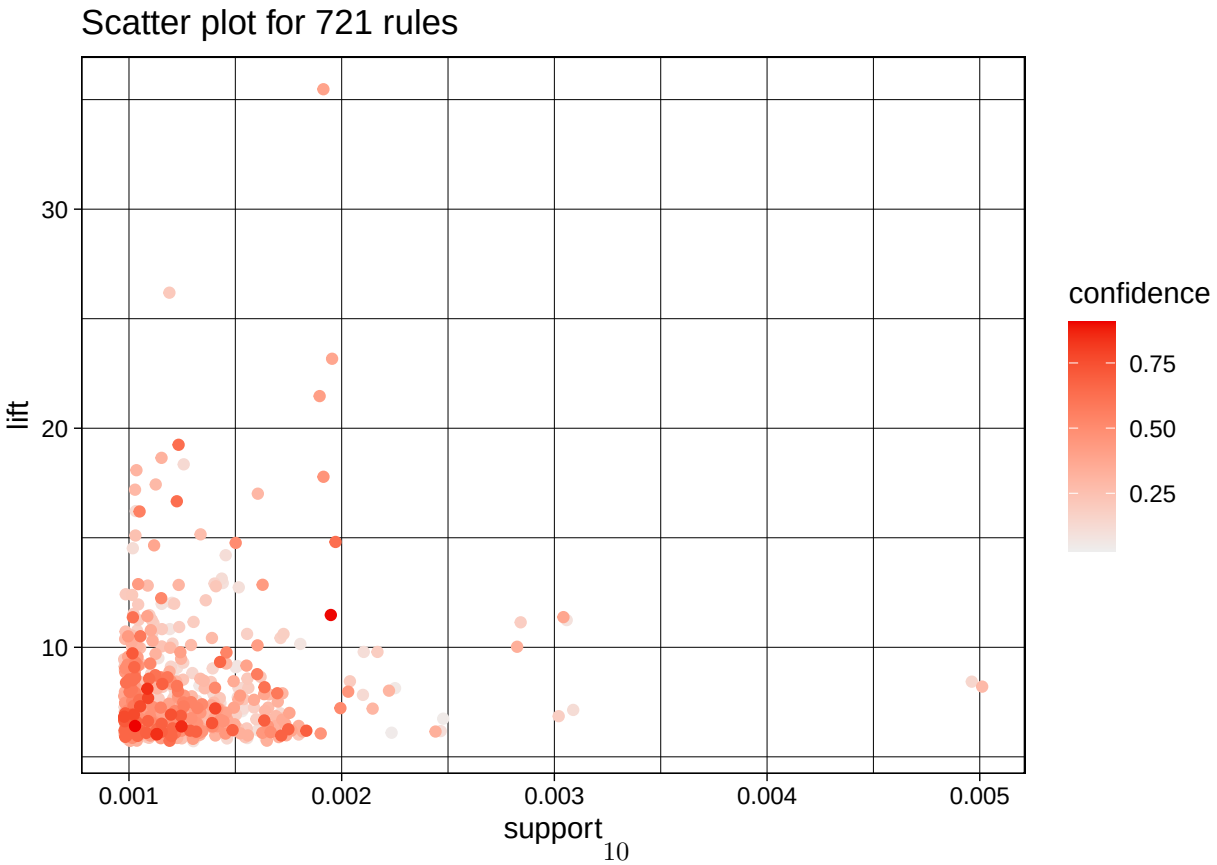
support: 0.001 confidence: 0.01 max length: 4 **lift:** 6

```
# Step 6: Association Rule Mining
f_grocery_rules <- apriori(grocery_transactions,
                 parameter = list(supp = 0.001, conf = .01, maxlen=4))

f_grocery_rules <- subset(f_grocery_rules, lift > 6)
```

Second Rule Graphs



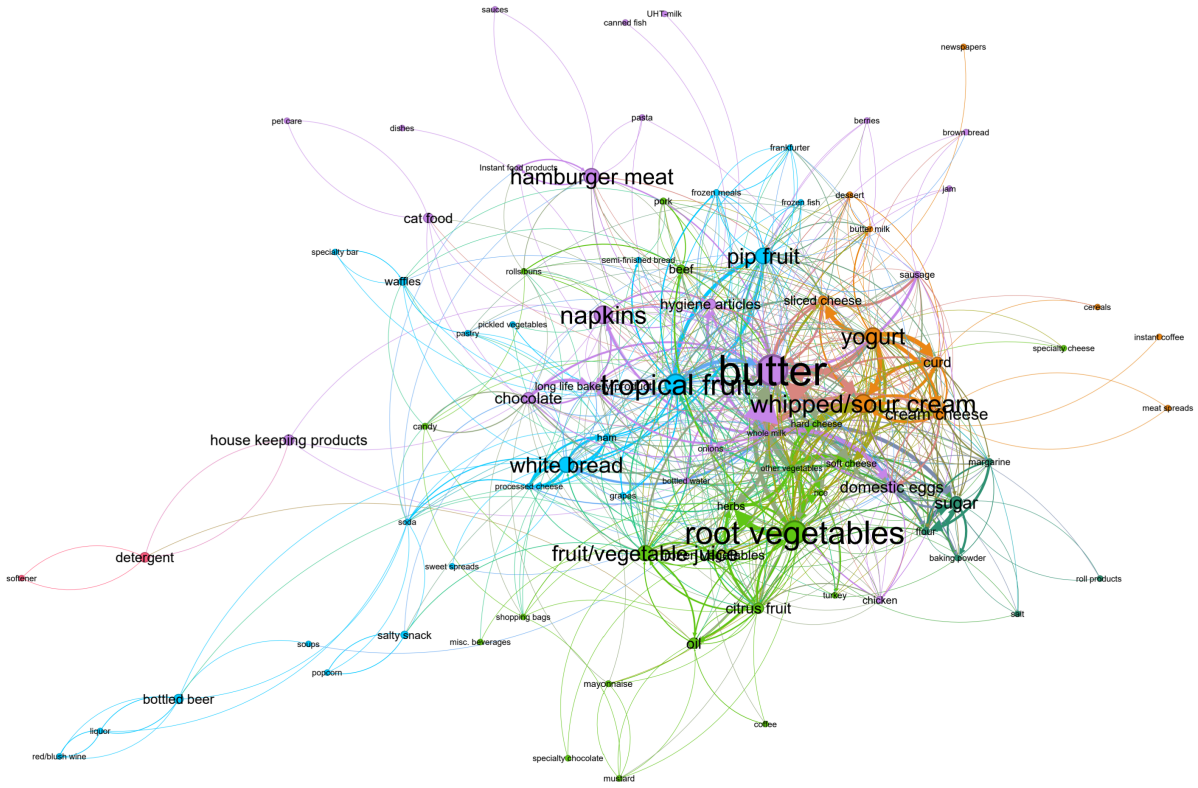


Based off of previous subsets and the analysis, these above are the graphs of our final rules for lift, support, and confidence.

Second and Final Igraph

```
# Step 7: Convert to Graph and Export for Gephi
f_grocery_graph <- associations2igraph(f_grocery_rules, associationsAsNodes = FALSE)
igraph::write_graph(f_grocery_graph, file='f_grocery_rules.graphml', format = "graphml")
```

Here is the network graph generated using Gephi:



By using the Gephi tutorial within the professor's github, we get this result. Gephi discovered 6 communities within the data! Gephi also got their share of all the data:

Purple: 27% Green: 25% Blue: 25% Orange: 12.64% Dark Green: 6.9% Pink: 2.3%

The graph is an association network, so in short terms it represents items that are bought together. They may indicate common ingredients, styles of food, necessities, etc. Something which worried me at first, was how small the whole milk node is. it is in the purple group, below the butter. I was concerned, as it was the most frequent item but was small in the results. However, I believe this is due to what was discussed earlier: frequency **does not** necessarily indicate strong association.

Individual Nodes

Look at the large nodes, like root vegetables, butter, white break, tropical fruit, etc. These are nodes that have lots of strong associations for many other products. As the nodes get smaller and further away, these have less associations and the associations are weaker.

Root vegetables for example, are bought often with oil, citrus fruit, fruit/vegetable juice, herbs, rice, other vegetables, etc. This can be determined by looking at the arrows and which directions they point in (although it is a little tough to do)

Another node or community we can look at, is sugar. Based off of the arrows, we can determine that flour, baking powder, margarine, domestic eggs, salt, and roll products have strong associations with sugar and with each other (whenever their arrows point at each other). However, something like roll products and salt, although they are in the same community, don't have an association.

Communities

We have already covered two communities, green (root vegetables) and dark green (sugar). The items within each group do seem to corespond to each other, and the identity of these communities makes sense.

The orange community also seems to make sense. Whipped/sour cream, yogurt, curd, cream cheese, sliced cheese do make a lot of sense together and are relatable! On the outskirts, we see smaller associations which do still make sense, just not as much and only with specific other items. For example, deserts, cereal, meat spread, coffee, etc. Notice how these only have a couple arrows pointing at them. Coffee only has one! What this has taught me is that the further away/smaller a point is from the middle of the group, the less relatable and less often it is bought with those things

The community which makes the most sense and is especially specific is pink, which only contains softner and detergent. The lift was 10, and the confidence was 90%, so them being together and their own category makes sense

Two communities I found to be broad were the purple and blue communities. Some things don't necessarily make sense to be a part of the same community. For example, in purple, napkins, butter, hamburgers, and chocolate are all in the same group! In order to understand it better, you must take a look at where the arrows are going and which arrow is touching which. If you look closely, napkins and hamburgers do **NOT** have a lie drawn towards each other, even though they are particularly large nodes in the same community. Rather, they connect with nodes within the same community, and have sort of their own subsets of associations. For example, hamburgers has sauces and instant food products nearby, with suaces only connected to hamburgers.

You can also find another sub community in purple, cat food and pet care products.

The blue community also has a lot of sub communities, for example:

Popcorn, salty snacks, soda Bread, processed cheese, ham Bottled beer, liquor, red/blush wine

My personal favorite, there are connections between soda and liquor! Could this be jack and coke?

In the future, I would love to dive further and understand why the blue and purple communities have strange associations. I would also like to perhaps include even more communities!