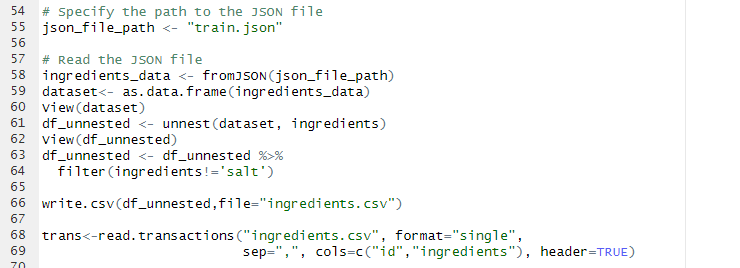
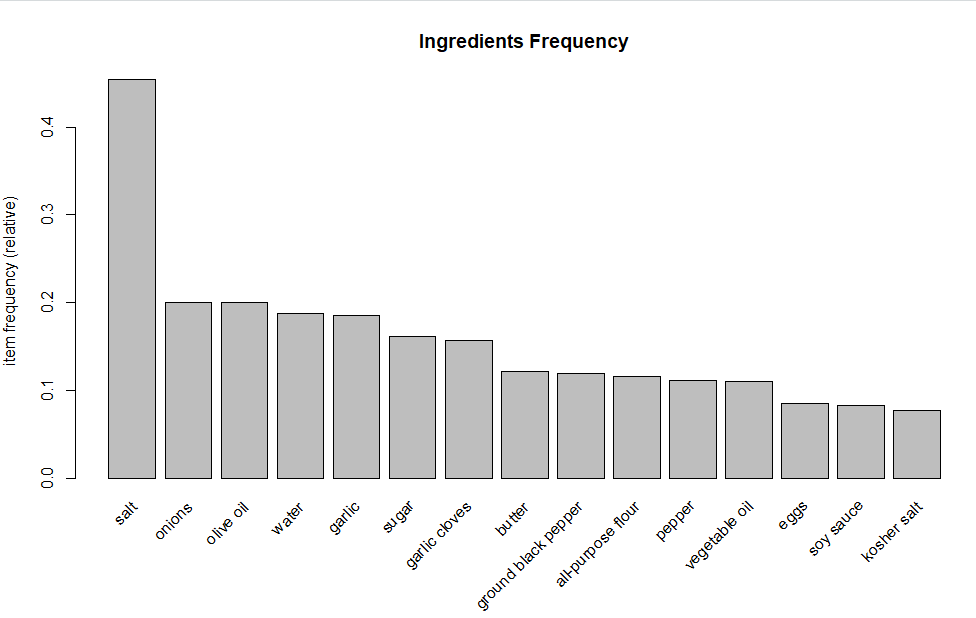
Association Rule Mining - Food Ingredients

In this report, we will delve into the food ingredients dataset and analyze and detail the findings of the relationships between ingredients. We start off by loading the relevant libraries using pacman and then read the json file into the transaction data type.



## Item Frequency Plot

Firstly, we plot the item frequency, this allows us to discern any ingredients that occur too often and don't provide any meaningful relations and remove them, in this case: salt.

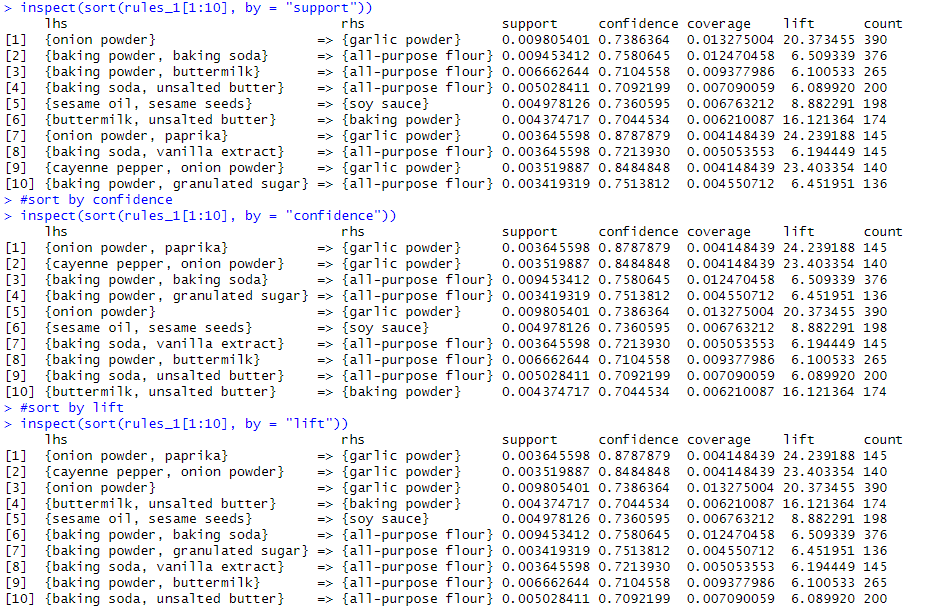


## Varying Support and Confidence Levels for the Apriori Model

We create 3 different apriori models and then compare the top 10 for support vs confidence vs lift for each model.

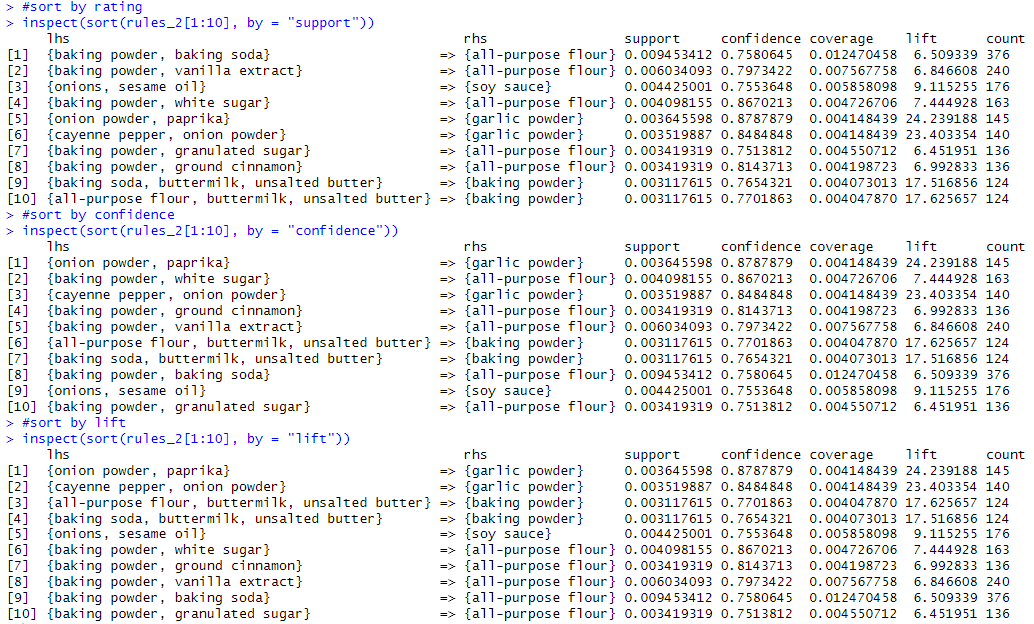
### Iteration 1

Support = 0.003, Confidence = 0.7



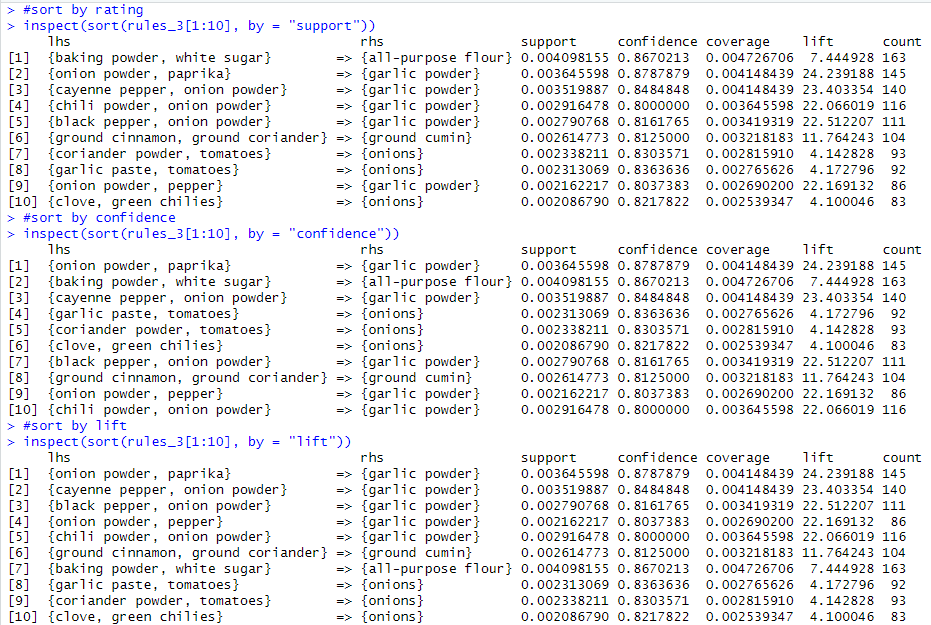
### Iteration 2

Support = 0.003, Confidence = 0.75



### Iteration 3

Support = 0.002, Confidence = 0.8



**Analysis:** As the support goes down and confidence increases, the rhs of the relationship focuses more on garlic powder and lhs on other spices, vice versa, the relationships focus more on bakery goods as the rhs focuses more on flour and the lhs on items for baked goods such as baking powder, sugar and eggs. Ordering by Support and confidence does not show much of a difference in the top 10, however ordering by lift shows other spices having a stronger relation with garlic powder than, the relationship between bakery ingredients showing this dataset will be helpful for identifying associations between spices.

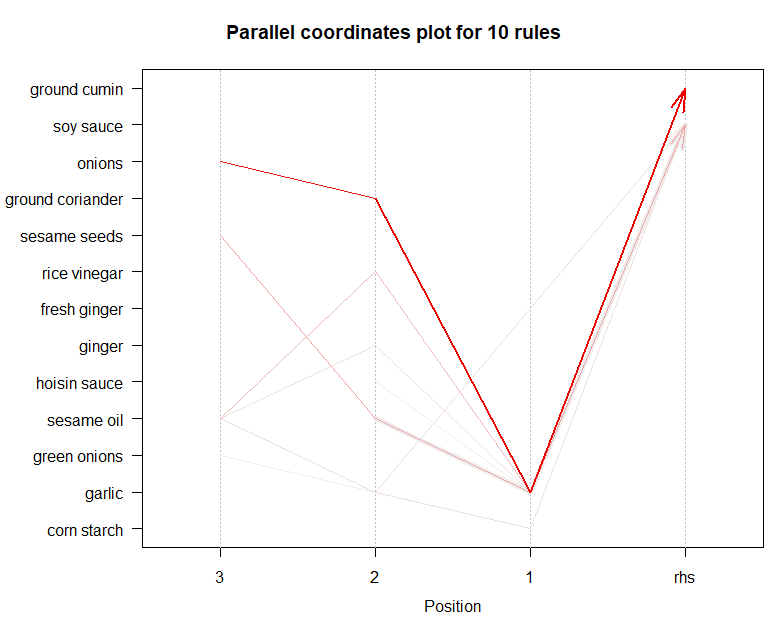
## Specified LHS for Apriori Model

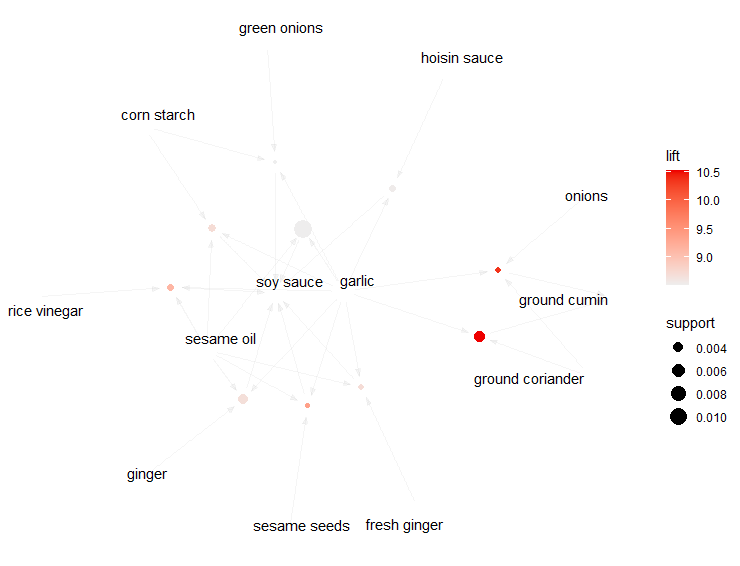
Analyzing this apriori model which has 146 rules for specific lhs ingredients. Since this specification results in a low number of rules, we will not plot the scatter point graph.



### Iteration 1





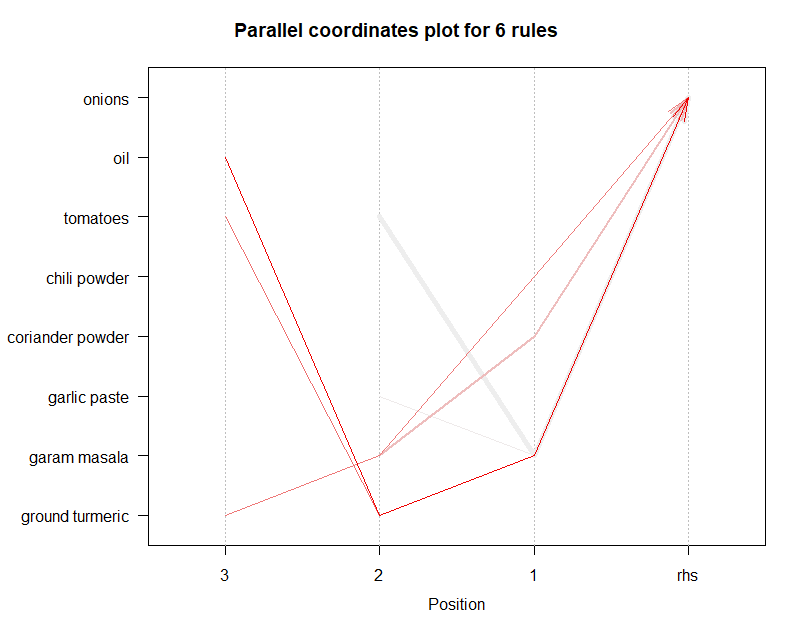


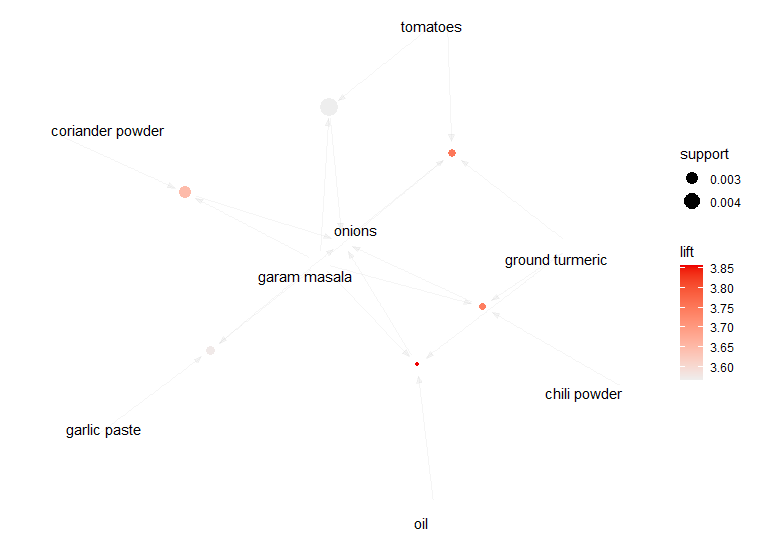
**Analysis:** For garlic, onions and ground coriander on the lhs and soy sauce on the rhs form the strongest relations, it also forms a less intense, but still prevalent relationship with sesame seeds and oils on the lhs and onions on the rhs. Typically, garlic appears in relations where there are 3 items on the lhs.

In the graph plot, we can see that soy sauce on the rhs and garlic on the lhs has the highest support, although its lift is almost none showing that their independent frequencies are high but their co-occurrence does not provide any additional information. Garlic seems to have the best relationship with onions, ground cumin, and ground coriander as the lift involving relationships with them is the highest.

### Iteration 2



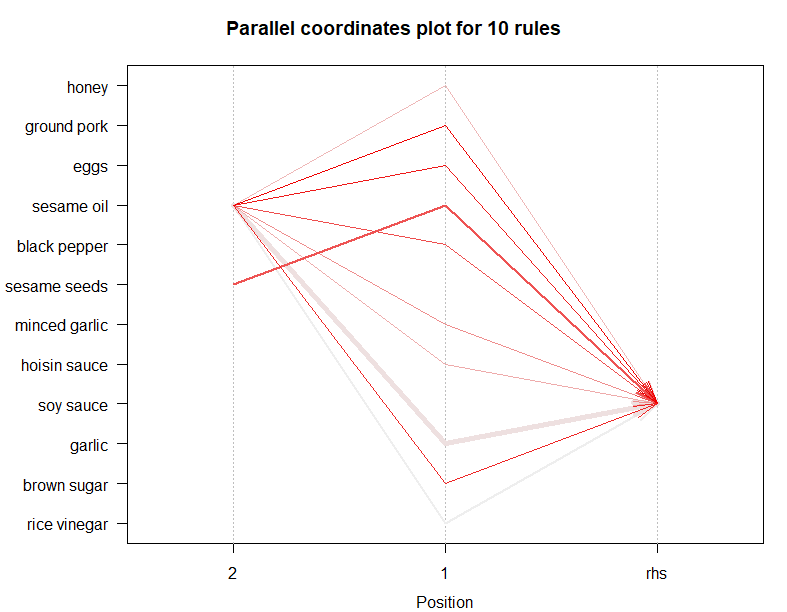




**Analysis:** Garam masala has strong relations with multiple ingredients such as oil, tomatoes and ground turmeric on the lhs. It occurs an equal number of times for both 3 ingredients on the lhs and 2 ingredients on the lhs. When we have garam masala on the lhs, we can see the resulting rule always has onions on the rhs.

The relationships with the highest support are {tomatoes, garam masala} => {onions}, however its lift is very low thus we don't consider it a significant relationship. The relationships that do however have a decent lift, usually include ground turmeric, oil and coriander powder.

### Iteration 3



### 

### 

**Analysis:** In this iteration, we specified we atleast want sesame oil once in each relationship. Almost all of the relationships involving this, results in soy sauce as the rhs. Typically sesame oil in this model has more than 3 ingredients on the left , however there are a few cases of only 2 ingredients, this indicates that there is a high interdependence of items.

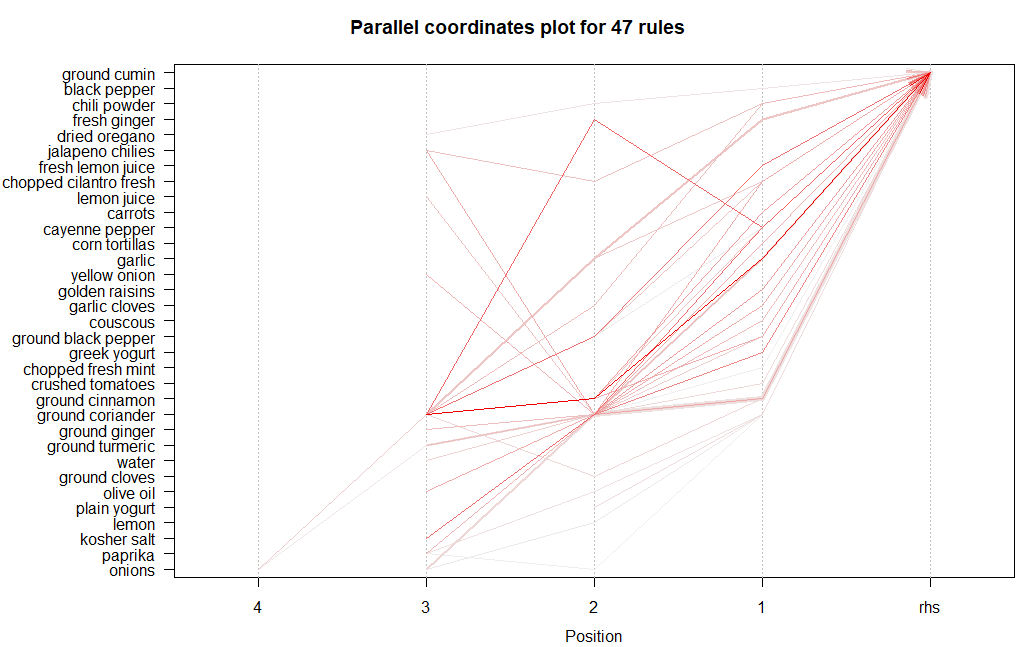
In this iteration, the lift and support for all the relationships is relatively higher showing stronger and more telling associations. Sesame seeds and garlic have a higher frequency however garlic does not form any strong relationships, thus it doesn't give more intel on the association rules. Brown sugar, eggs and sesame seeds, have a higher lift showing they occur more times together with sesame oil than they would independently

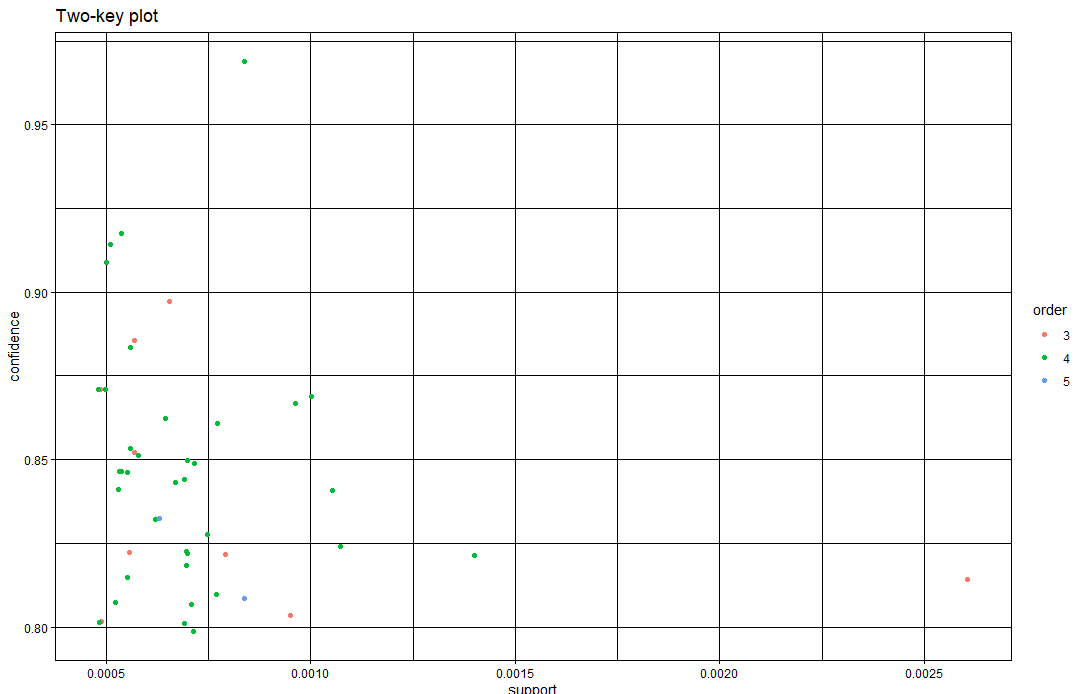
.

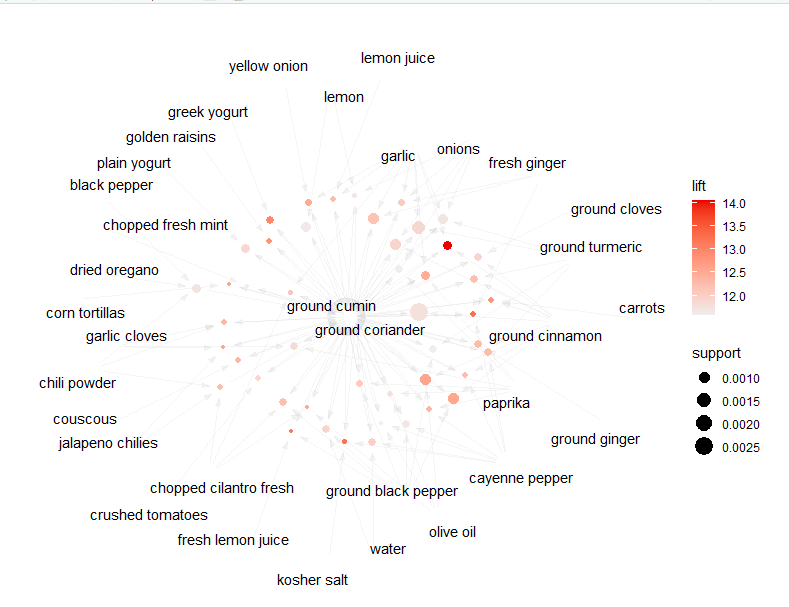
## Specified RHS for Apriori Model

### Iteration 1





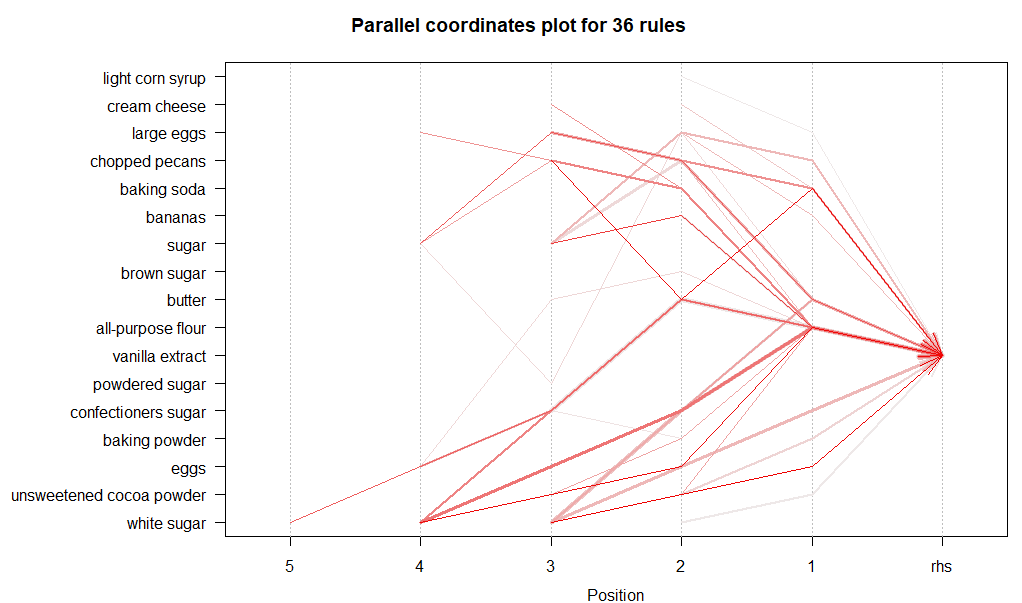


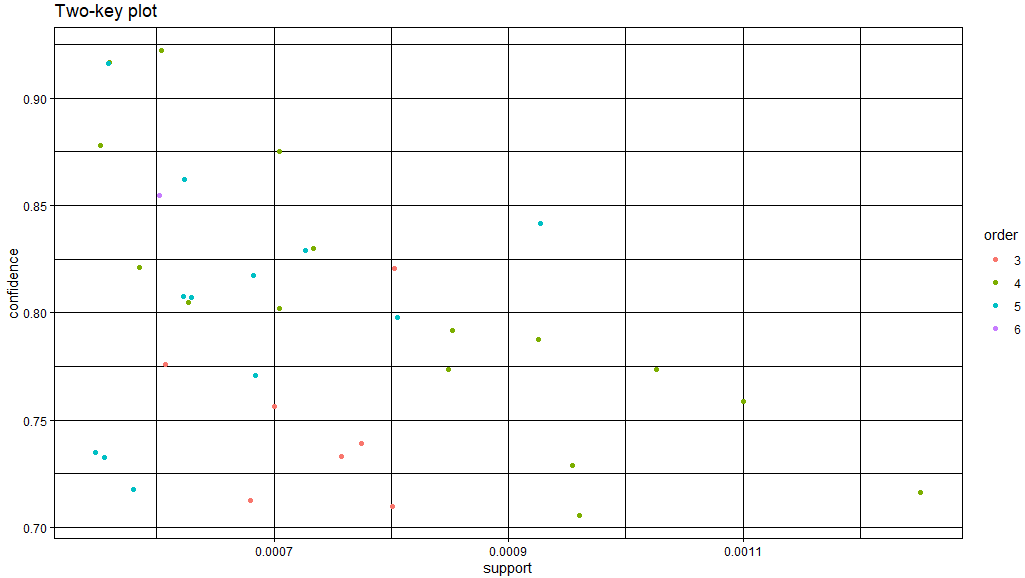


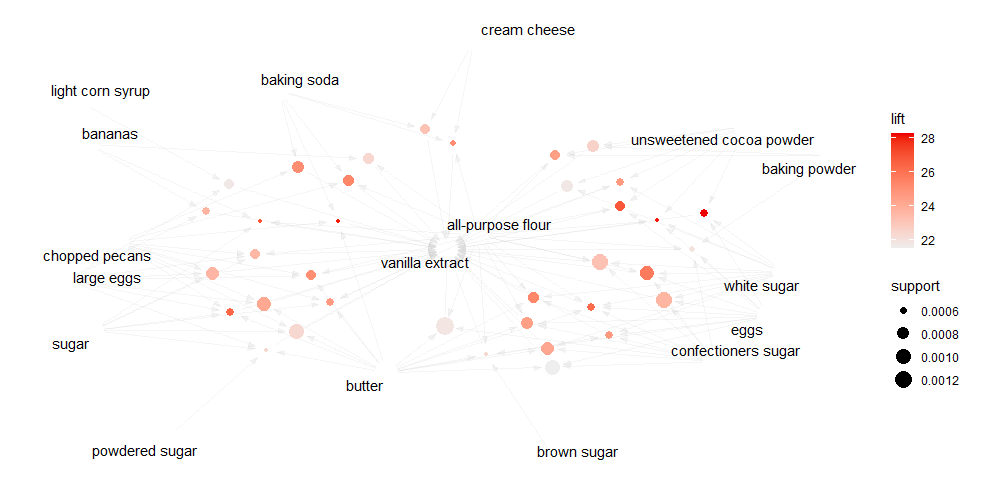
**Analysis:** In this model, when we specify ground cumin on the rhs, for the lhs, fresh chopped mint and ground coriander seem to have the highest support, while ground coriander, fresh ginger and ground cinnamon seem to form the strongest relationship with the ground cumin. Ground coriander usually appears when the lhs has 2 ingredients in it. Most rules have low support, indicating the involved itemsets are uncommon within the data. There are very few rules that have both high support and high confidence.

### Iteration 2





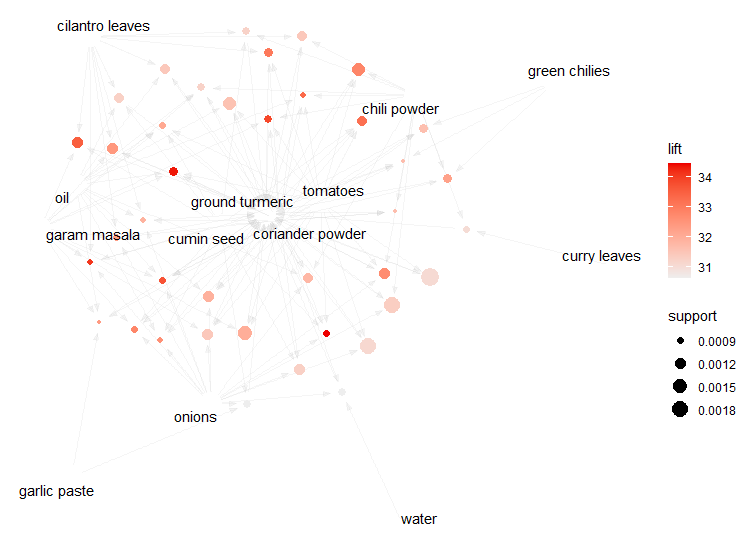
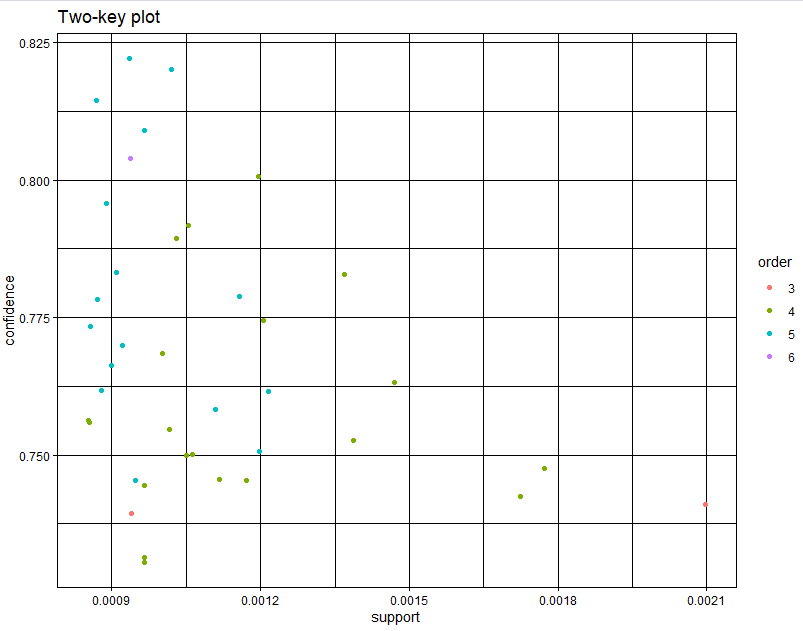
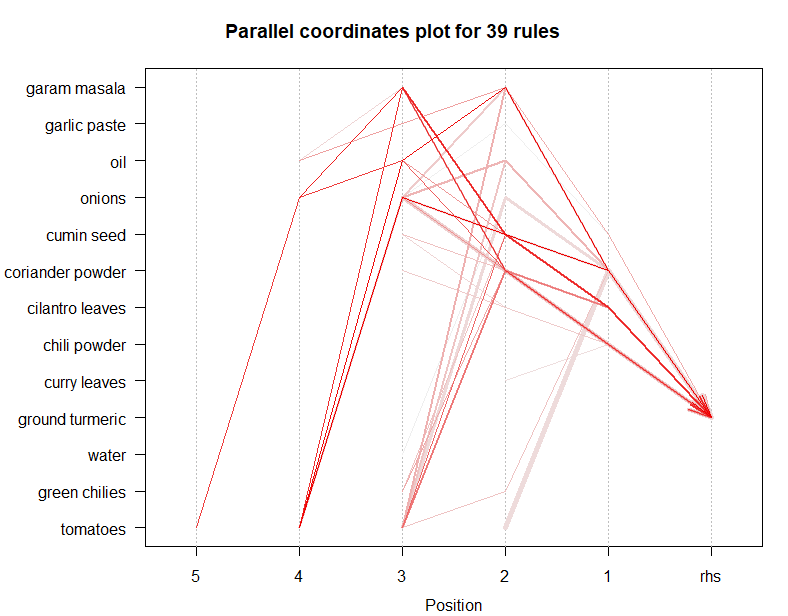




**Analysis:** In this model, when we specify vanilla on the rhs, all of the following itemsets include bakery focused goods as one would expect. Ingredients like sugar, butter, eggs and all purpose flour have the strongest relations as well as making up a majority of them. In the 2 key plot, the ingredients seem to belong to orders 4 and 5 the most, pointing to this model having more complex itemsets consisting of multiple ingredients. Several of the points also seem to be high in confidence (>85) but low in support pointing to them being highly reliable but infrequent in this dataset. In the graph plot, Items like sugar, baking powder and unsweetened cocoa powder have smaller nodes but are darker in color, indicating they have high lift despite lower support. All purpose flour is located in the middle of the graph, with multiple lines extending out from it indicating it is heavily involved in the rules formed for this apriori model.

### Iteration 3





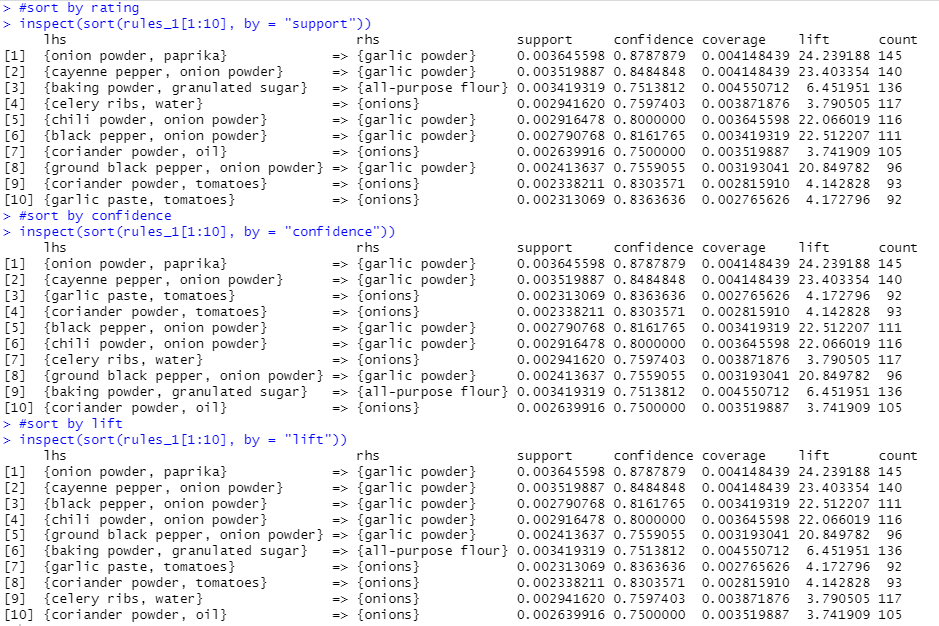
**Analysis:** In this model, since the rhs is ground turmeric, the following itemsets focus more on spices. Tomatoes, onions, garam masala and oil make up most of the relations as well as their co-occurrence being high. The itemsets in this model also have more than 3 ingredients almost every time. In the 2 key plot, the ingredients seem to belong to orders 4 and 5 the most, pointing to this model having more complex itemsets consisting of multiple ingredients. There are more low support and high confidence rules than in the other iterations pointing to this having a more diverse number of rules and them being more reliable. In the graph plot, you can see the items mentioned previously being positioned near the center showing their key involvement in the associations as well as their connection to the different lift and support size based clusters.

## Repeating 2 and 4 with at least 3 different items

Step 2 Repeated:

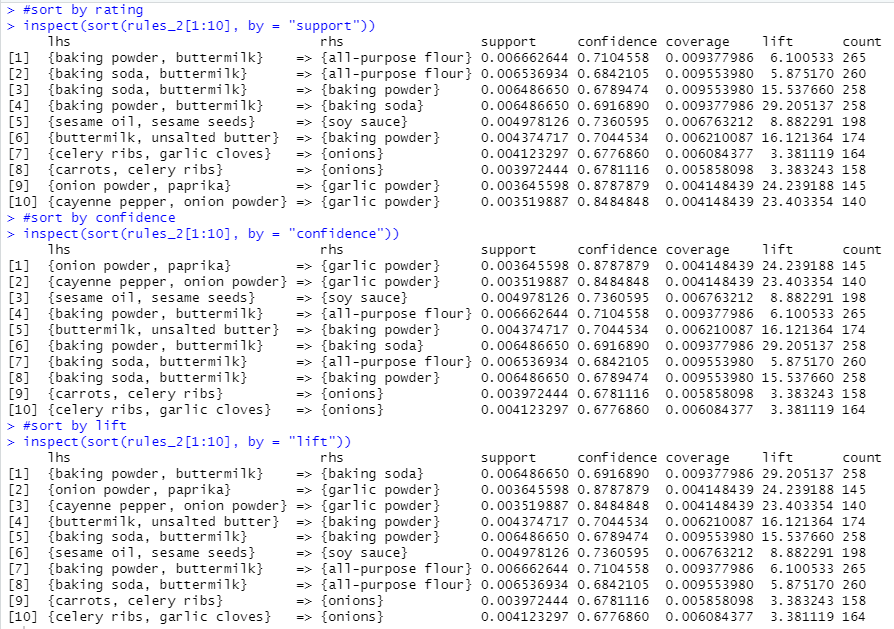
### Iteration 1

Support = 0.0023, Confidence = 0.75



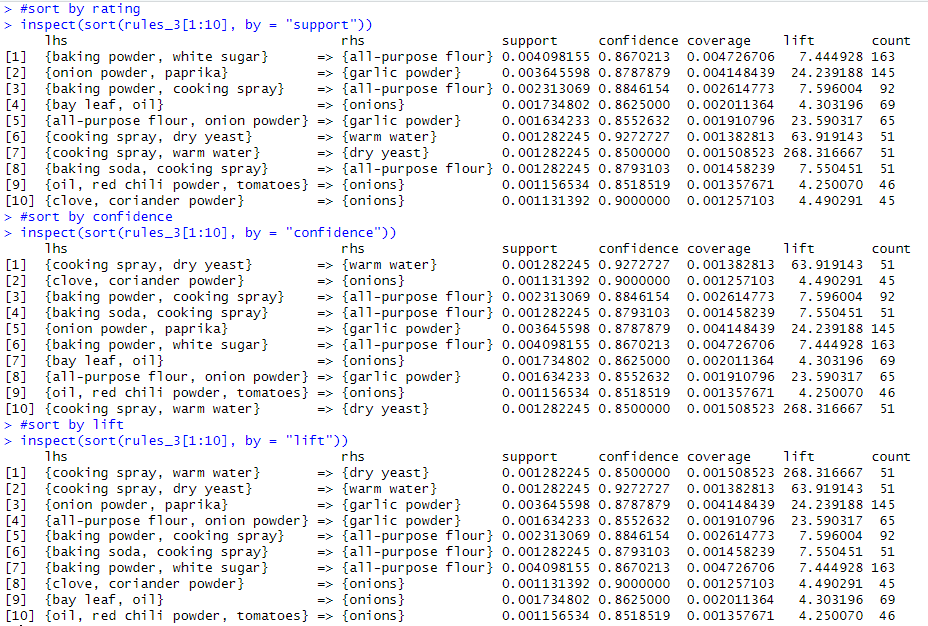
### Iteration 2

Support = 0.0033, Confidence = 0.67



### Iteration 3

Support = 0.0011, Confidence = 0.85

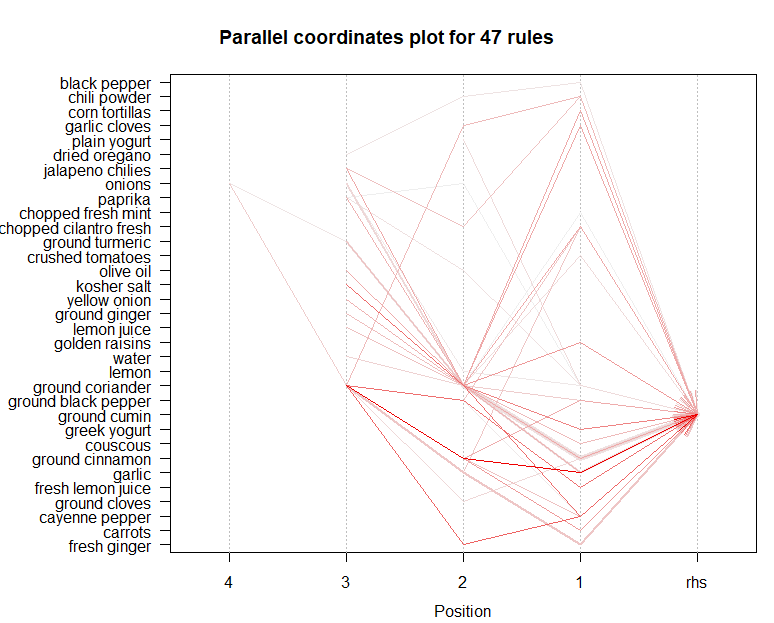


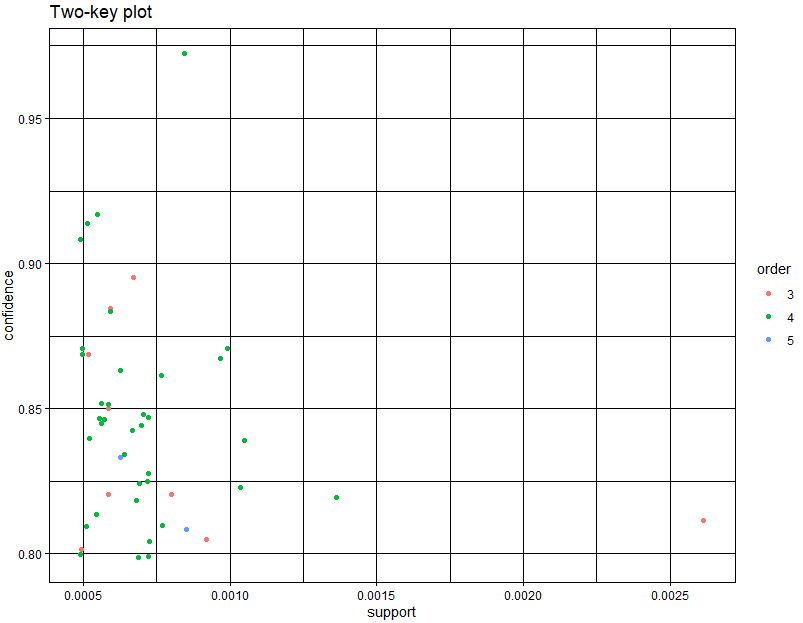
**Analysis:** As we increase the minlen of this apriori model, the stronger rules start to diversify, and their overall confidence, support and lift increases. New relations start to form and they have a higher lift and confidence than the previous ones such as cooking spray, dry yeast and dry water, which perhaps suggest cleaning items rather than just cooking ingredients that were prevalent in the dataset as well as different spices and vegetables. However the frequency of the previous items still appears to be higher than the new additions. Increasing the confidence and decreasing the support results in more of these non-direct cooking related ingredients while increasing the support and decreasing the confidence results in more vegetarian related items such as celery ribs, buttermilk, butter, garlic cloves and so on.

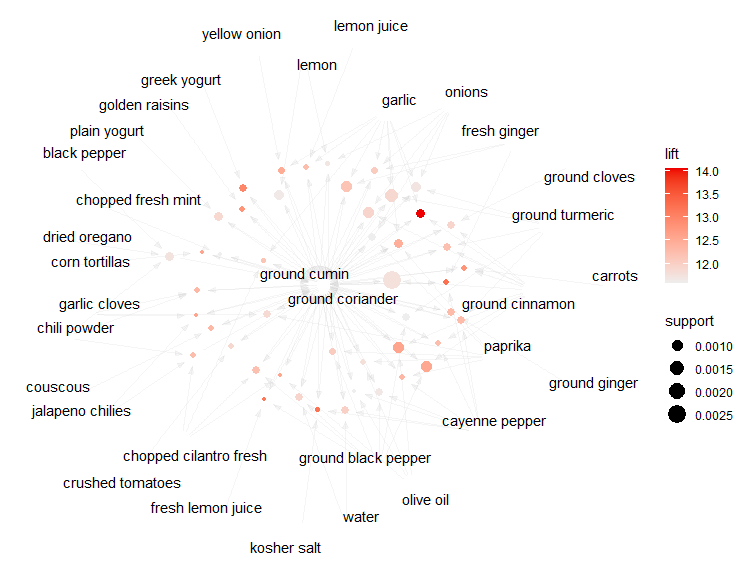
Step 5 Repeated:

### Iteration 1



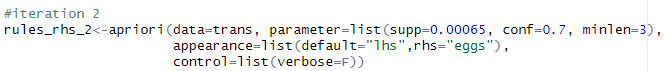


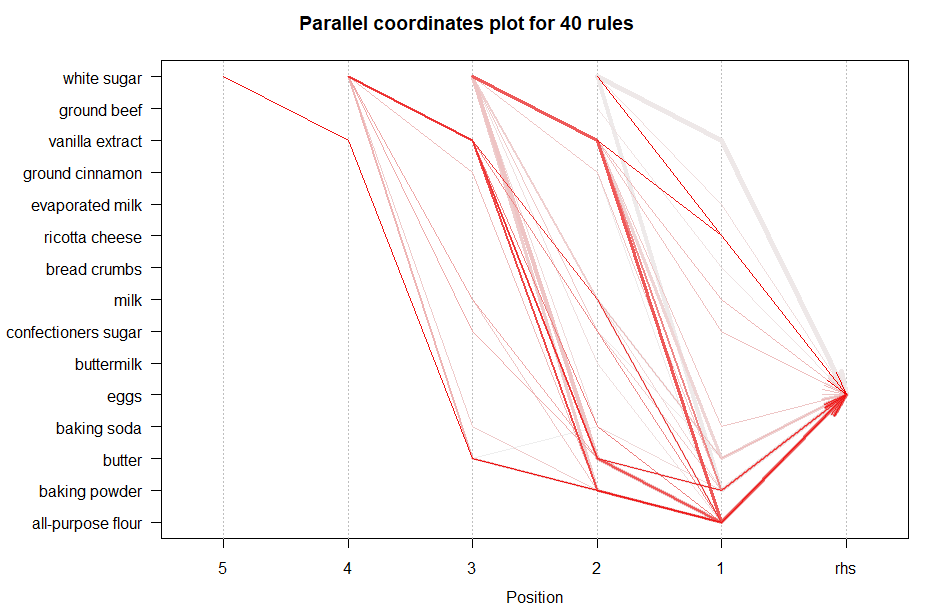
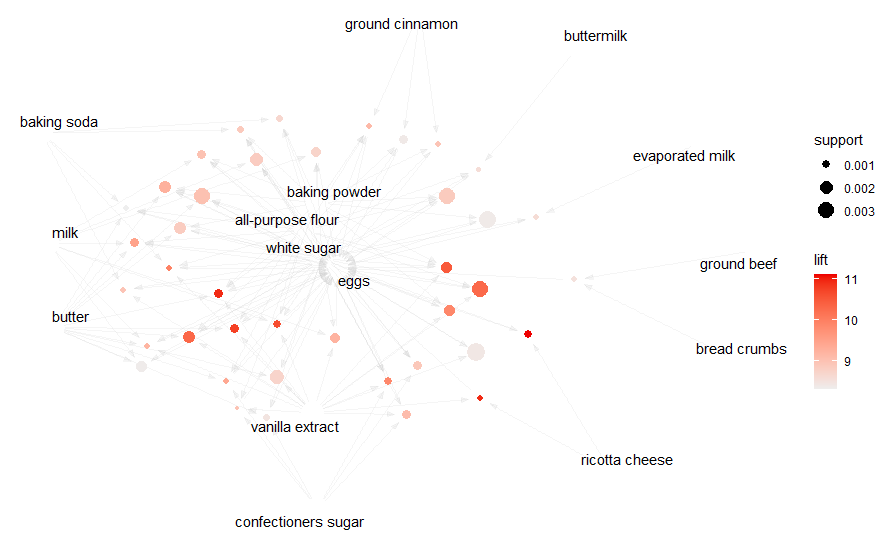
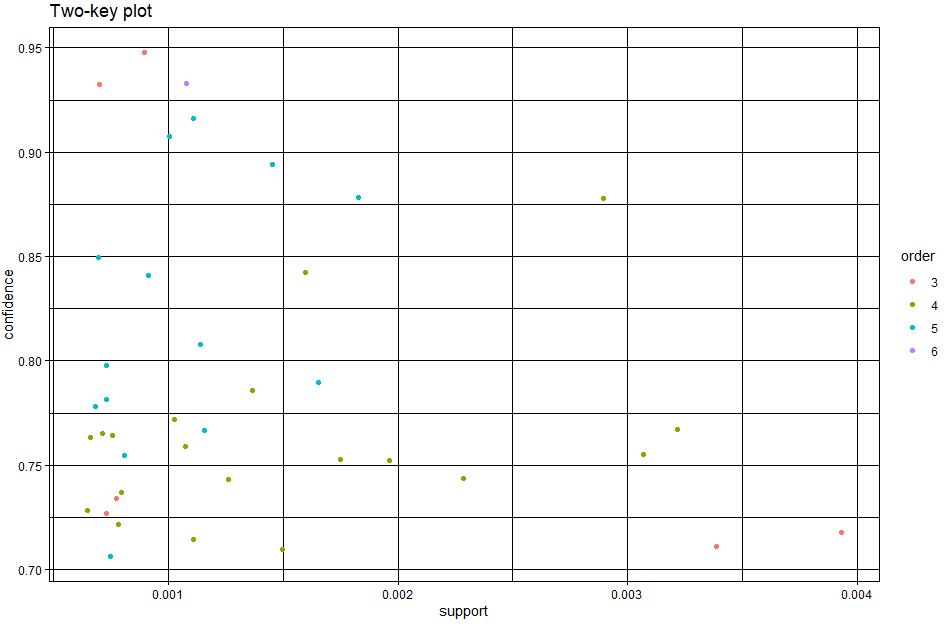




**Analysis:** In this iteration of the apriori model, we set ground cumin as our rhs, unlike the previous iterations, over here the rules are a bit more diverse and contain more ingredients. Typically the itemsets contain 2-3 ingredients on the lhs. In the 2-key plot, we can see that the most common order is 4 showing that it has a mediocre amount of complexity. It has low confidence and low support in general showing the rules are not too reliable. Ingraph plot, we can see that ground coriander is also at the center of the graph due to its high frequency of connections with other nodes. Garlic also has a strong association with ground cumin however its support is low suggesting while reliable, it is not frequent.

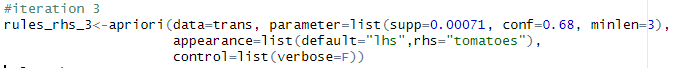
### Iteration 2

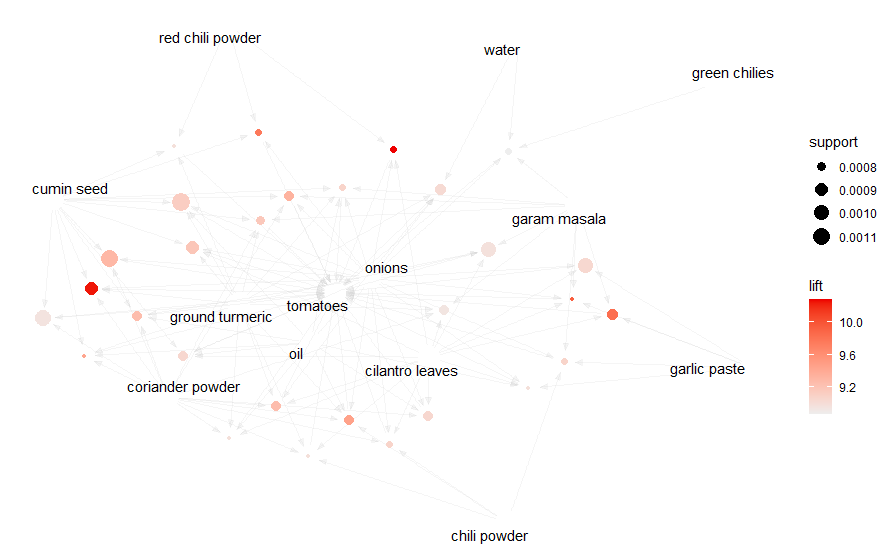
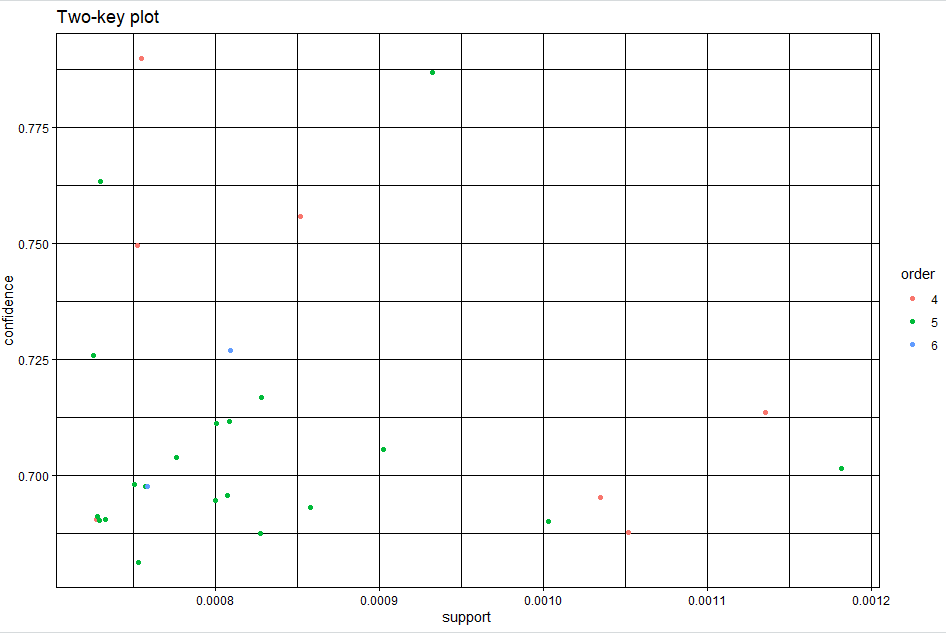
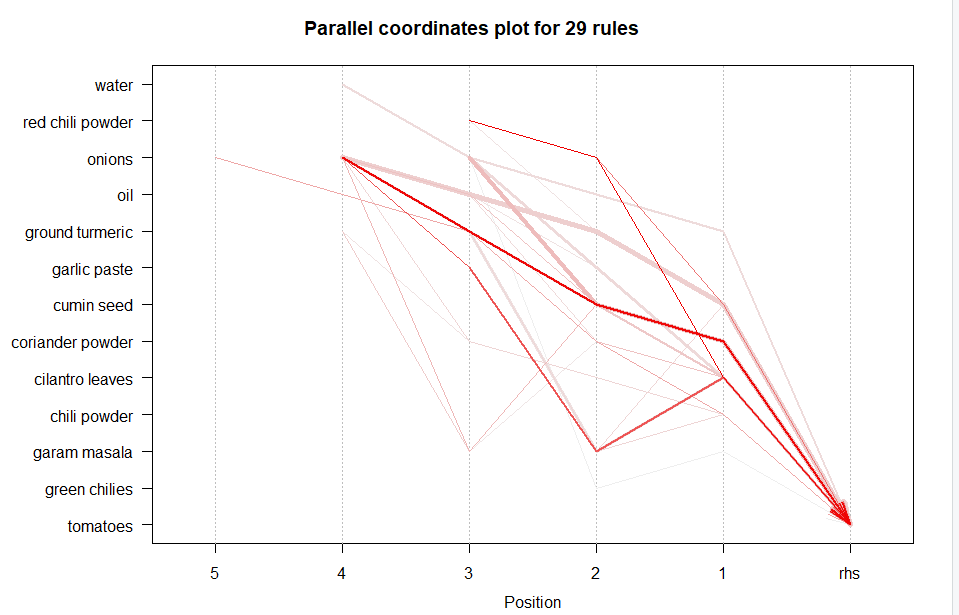




**Analysis:** In this model, since the rhs is eggs, the item sets seem to follow a simple pattern that starts off with white sugar and then different dairy products such as cheese, butter or milk, then baking powder or all purpose flour and then finally eggs with the exceptions and a few change sin ingredients of course, but these seem to have the strongest associations . In the 2 key plot, the ingredients seem to belong to orders 4 and 5 the most, pointing to this model having more complex itemsets consisting of a multitude of combinations of different ingredients. There are more high confidence rules than in the other iterations pointing to this having a far more reliable and robust set of rules . In the graph plot, you can see the items mentioned previously being positioned near the center showing their key involvement in the associations as well as their connection to the different lift and support size based clusters. While butter and ricotta cheese may ot have the highest support, it has the best lift indicating these specific rules are the more reliable available even though their frequency is low.

### Iteration 3





**Analysis:** In this model, when we specify tomatoes on the rhs, this leaves the associations more open to interesting associations that one might not have thought of unlike in previous iterations, where people can come to pre-conceived associations based on everyday knowledge. Ingredients like onions, oil, garam masala and cilantro leaves have the strongest relations as well as making up a majority of them. The 2 key plot indicates that while the itemsets are rare, some of them are highly predictive of the presence of tomatoes. In the graph plot, Items like cumin seed, water red chili powder and garam masala have smaller nodes but are darker in color, indicating they have high lift despite lower support therefore being specific and reliable rules . Spices like ground turmeric, coriander powder, red chili powder and cumin seed all form a cluster suggesting overlapping itemsets.