

**Faculty of Computing and Informatics**

**(FCI)**

TDS3301

DATA MINING

Assignment

Part 2 Association Rule Mining

Prepared by:

|  |  |
| --- | --- |
| **Name** | **ID** |
| Yap Kit Boon | 1141124580 |
| Thong Yong Lin | 1132702398 |
| Ooi Yi Jie | 1131122872 |
| Chew Siu Min | 1122703126 |

# PART 2: ASSOCIATION RULE MINING

Name: **Extended Bakery Dataset**

From: <https://wiki.csc.calpoly.edu/datasets/wiki/ExtendedBakery>

## 1. Objectives: What is the domain and what are the potential benefits to be derived from association rule mining. This is high level - not find patterns, but what would improve because of the use of the patterns.

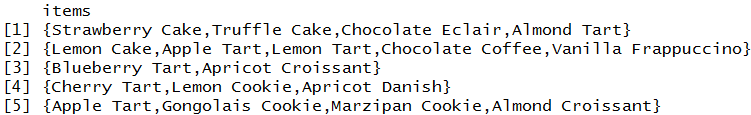
Association rule mining is a procedure of finding frequent patterns, associations, or correlations from data sets. It is frequently used in transactions data to discover interesting relations between items data and allows us to identify a set of rules that can be understood as “if this, then that”. As such, we will be able to gain insight on what items are always purchased together. Using the information obtained from the patterns, many improvements can be made such as related product recommendation, better shelf arrangement which can allow the store to gain more profits by increasing the sales.

## 2. Data set description: What is in the data, and what preprocessing was done to make it amenable for association rule mining. Where choices were made (e.g., parameter settings for discretization, or decisions to ignore an attribute), describe your reasoning behind the choices.

The dataset we choose is a full binary vector representation dataset which is showing 5000 transactions of 50 foods include 40 pastry items and 10 coffee drinks in a bakery chain. There is one “Receipt No” column representing receipt number and 50 columns representing 50 foods in bakery. Each receipt has its own receipt number followed by 0’s and 1’s indicating if that particular food item was purchased in the transaction, whereby 0 stands for no while 1 stands for yes.

## Preprocessing

Firstly, we removed the first column which is ‘Receipt No.’ column because it is not important. The dataset was also converted from data frame into matrix to be able to read in transaction format. As the items are still represented by numbers which is rather confusing, we converted it into item names. Our dataset are represented in the form below, each row represent a transaction and the contents are the items purchased in that particular transaction.



Decision to ignore an attribute

Other than the removed **‘Receipt No.’** column, we also decided to ignore **‘Quantity’** of purchased, and **‘Price’** of item because they are not important in applying association rules mining.

## 3. Rule mining process: Parameter settings, choice of algorithm, and the time required.

Parameter Setting

We set **support** to be **0.001**, **confidence** to be **0.8** and **minlen** to be **2**. This will give the result of 174 rules, which we think that this set of rules is appropriate to observe how the market is going. Our team have tried several different values of support, confidence and minlen to observe the number of rules provided.

Support

When we set support to **0.002** with the same confidence and minlen value, the set of rules decreases from 174 to 85, that’s mean that we get lesser rules to observe how food sell in bakery. However, when support value is 0.001, it results in 127 rules which we think that this number is sufficient for us to conduct our investigation in food sales in the bakery.

Confidence

We set confidence to be **0.08** is because we find the itemsets which carry 80% probability when customer buy a set of food, he or she will also going to buy the other considered food.

Minlen

We set it to **2**, it’s because we want to view itemsets with minimum 2 items.

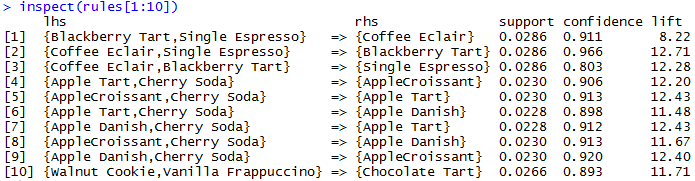
### Choice of algorithm

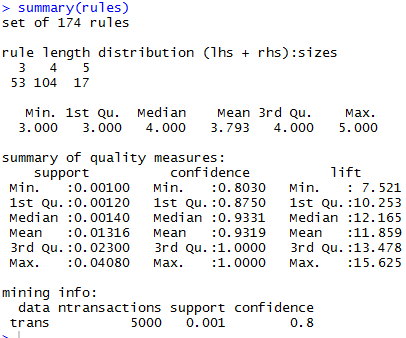
The rules were created using the ***apriori* function** on the dataset. Package arules is used to apply association rules while package arulesViz is used for additional features such as graphing and plotting the rules.

### Time required

Run time to finish all code for the experiment in Assignment Part 2.R: **within** **1 minute**

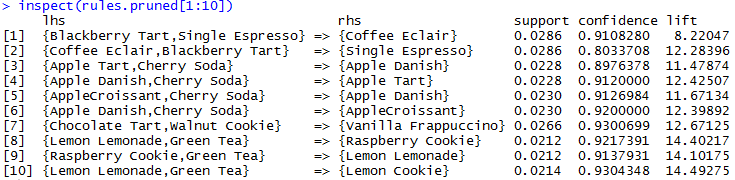
## 4. Resulting rules: Summary (number of rules, general description), and a selection of those you would show to a client.

These are the first 10 rules that generated by the *apriori* algorithm from the 174 set of rules. Before pruning out those redundant and infrequent rules, we cannot really get an accurate rules, so we decided to clean that out.

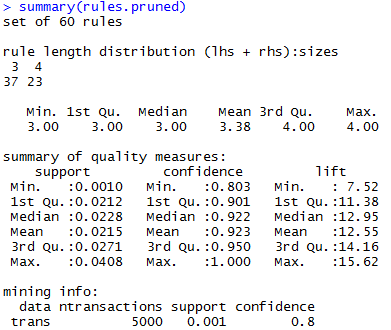


This is the summary result after we set the parameters that we mention is Question.3, we get 174 set of rules, from the summary result show up we know that there is 53 rules generated with 3 items in the set, 104 rules with 4 items in the set, and 17 rules with 5 items in the set. So after we clear out the infrequent and redundant rules, we get another amount of the rules.

After pruning:



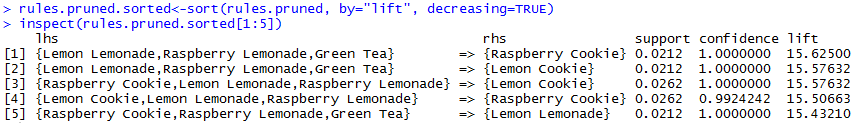
These are the 10 rules that generated after we did the pruning.



From the new result of the summary of the pruned rules, we get 60 set of rules. Which we can see that there is 37 rules with 3 items in the set, and 23 rules with 4 items in the set. Then, we tried to sort the pruned rules by several parameters.

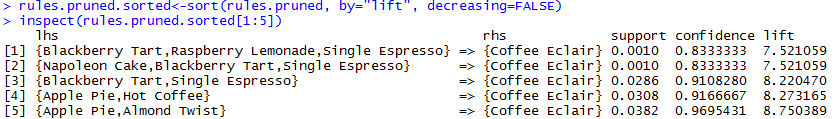
### The pruned rules sorted by the maximum lift

Sorting by maximum lift is to show which itemsets are getting more popular than others. It’s because the higher lift value, the better sales of the itemsets. Therefore, we can observe that good selling set of foods in bakery.



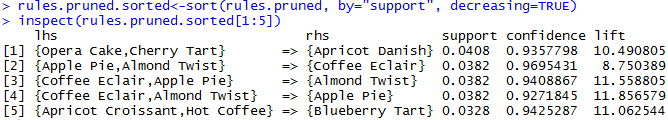
### The pruned rules sorted by the minimum lift

Sorting by minimum lift is to show which itemsets are getting less popular than others. In this case, we can observe that the combination of food with Coffee Eclair is getting less sales than other combinations.



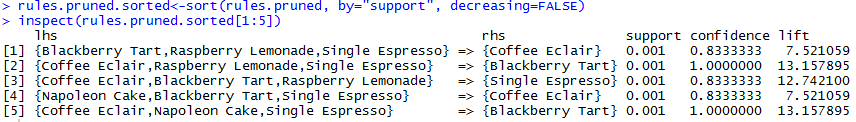
### The pruned rules sorted by the maximum support

Sorting by maximum support is to show the itemsets that more frequent appear that in this dataset than others. From the screenshot below, we understand that itemset that has highest occurrence is Opera Cake, Cherry Tart and Apricot Danish.



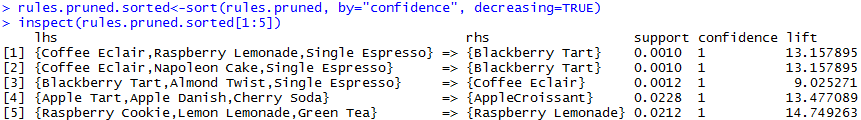
### The pruned rules sorted by the minimum support

Sorting by minimum support is to show the itemsets that less frequent appear that in this dataset than others. Therefore, from the result below, we found the combinations which carry lesser sales in bakery. That’s meaning that we should put more attention to those itemsets, discuss some alternative to improve the sales.



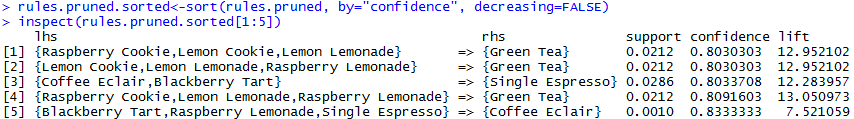
### The pruned rules sorted by the maximum confidence

Sorting by maximum confidence is to show those itemsets that carry higher probability to be purchased together. From one of the itemsets in this result, we can know that if customer buy Coffee Eclair, Raspberry Lemonade and Single Espresso, then he or she will definitely buy Blackberry Tart because it carry 100% confidence.

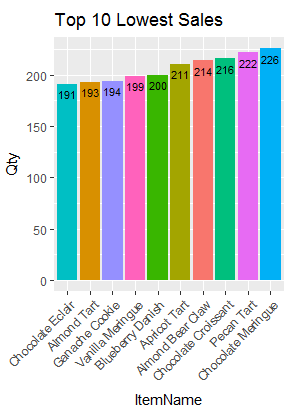
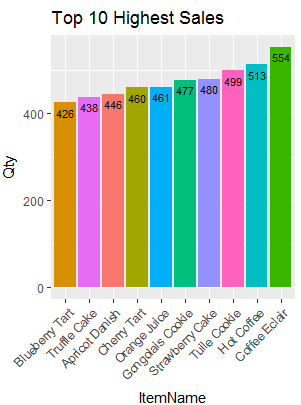


### The pruned rules sorted by the minimum confidence

Sorting by maximum confidence is to show the itemsets that carry lower probability to be purchased together. The 5 rules stated below are the rules that the client should consider not to group the items together when packaging the items.



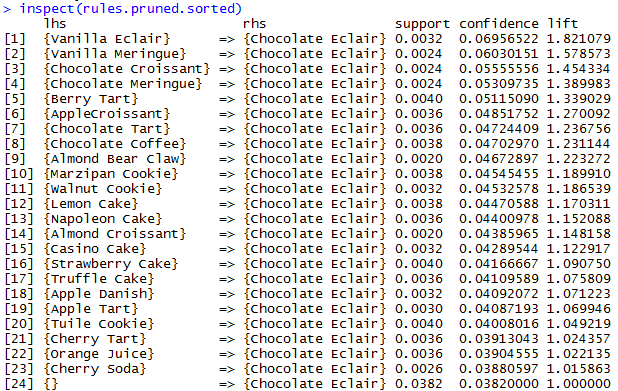
What we want to show to client is the itemsets that involves lower sales food in bakery, to find out a way of how to improve the food selling rate. Therefore, we found out the top ten highest sales and top 10 lowest sales foods in bakery.



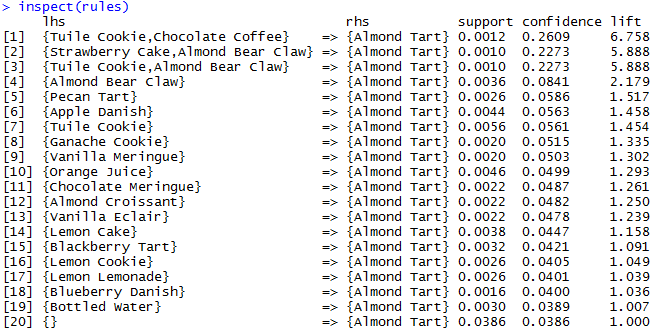
|  |  |
| --- | --- |
| Figure 1: A bar chart that display the top 10 highest sales in the bakery | Figure 2: A bar chart that display the top 10 lowest sales in the bakery |

Figure 1 shows the top 10 highest sales item, while Figure 2 shows the top 10 lowest sales based on the dataset. By using these two charts, we can now relate the number of sales with the rules we have discovered and shown at above. The charts also give an insight about the condition of item sales of the bakery shop to the client. With this, he can probably do some adjustment with the low sales items as recommended in Question 5.

The screenshot below is showing the set of rules with the lowest sales item, Chocolate Eclair at right hand side, to see which items would be purchased together with Chocolate Eclair. In this case, rule no.20 is considered whereby a customer who purchased Tuile Cookies (high sales item) would likely to purchase Chocolate Eclair as well.



The screenshot below is showing the set of rules with the lowest sales item, Almond Tart at right hand side, to see which items would be purchased together with Almond Tart. According to the screenshot below, rule no 7 and rule no. 10 are considered.



## 5. Recommendations: What should the client do because of the rules discovered?

* Reallocate the low sales items at the counter
  + In order to increase client’s bakery sales, we can recommend the client to reallocate the lower sales items or food near or at the counter to attract customers’ attention to the selected items. For example, by referring to Figure 2, Item No.6 (Chocolate Eclair) has the lowest sales among the 50 foods, so we can encourage them to put it near to the counter, and staff can suggest customer to purchase Chocolate Eclair when customer approaches to counter.
  + Below are the low sales items that can recommend to customer when they are making payment at the counter:
    - Chocolate Eclair
    - Almond Tart
    - Ganache Cookie
    - Vanilla Meringue
    - Blueberry Danish
    - Apricot Tart
    - Almond Bear Claw
    - Chocolate Croissant
    - Pecan Tart
    - Chocolate Meringue
* Reallocate the low sales items around high sales item
  + Reallocating related low sales items with high sales item can also attract customers’ attention to those low sales items, increasing the chance of the items to be purchased. For example, Item No.10 (one of lowest sales food) and Item No.28 (one of highest sales food), which are Almond Tart and Tuile Cookie, is one of the itemsets in association rules mining, we can suggest client to reallocate Almond Tart around Tuile Cookie, so that there are chances for the customers to buy Almond Tart along when they buy Tuile Cookie.
  + The items that can be relocate: **(format: low sales item -> around good sales item)**
    - Chocolate Eclair -> Blueberry Tart
    - Almond Tart -> Tuile Cookie
    - Ganache Cookie -> Truffle Cake
    - Vanilla Meringue -> Apricot Danish
    - Blueberry Danish -> Cherry Tart
* Held promotion for low sales items
  + The sales can be increased by offering discounts and promotions on low sales items. For instance, promoting Chocolate Eclair to customers by setting a discount price for it. We can also suggest client to set different low sales items as promotional items for different day. For example, customers can get discount on Chocolate Eclair on Mondays, Almond Tart on Tuesday, and so on. This allows low sales items to be introduced to customer in a more efficient way.
* Bundle low sales items with high sales item
  + Introducing the low sales items in a bundle of high sales items while offering discount on whole bundle is also a good way to increase the sales. As bundle is going to sell because by purchasing the bundle, the low sales items can be sold together. For example, one of the highest sales item no. 45, Hot Coffee can be sold together with one if the lowest sales item no.26, Vanilla Meringue with an offer price. By doing so, Vanilla Meringue could get a chance to gain popularity
  + For bundle, we recommend client to set up a tea time package whereby the package items included blackberry tart, raspberry lemonade, single espresso and coffee eclair. From the rules we can see that although these are the low sales item, but these item are always come together, so we recommend client to set up this package to push the sales of other related item to this package.