TDS 3301

DATA MINING

Semester 2, 2016/2017

**Assignment Part II**

**Group members:**

Amir Ridhwan Ahmad Azman (1132701767)

Liya Saffura Shamsun Anwar (1132702377)

Ili Fadhilah Ahmad Hizzad (1151303720)

Suraya Ibrahim Shah (1151303737)

**CONTENTS**

Objectives 3

Data Set Description 4

Rule Mining Process 5

Resulting Rules 6

Recommendations 8

**OBJECTIVES**

The objectives of this project is to find insights by using association rule mining. The dataset used is in Bakery business domain. The dataset contains a year’s worth of information regarding sales made by several small bakery shops. Some potentially beneficial patterns and findings could be derived after performing association rule mining. Such as, finding items that has high support and sell other items that are somewhat similar to that item, put items that has high confidence together in the bakery and increase sales of items that has low support & confidence by making promotions, or placing them in an obvious place or next to items that has high support.

**DATA SET DESCRIPTIONS**

The dataset that are used is called The Extended Bakery dataset, and for this project, we worked on the 5000 receipts dataset. Using R, we have created two data frames, named, ‘goods’ and ‘df’. The ‘goods’ data frame contains information of goods sold in the bakeries such as ‘item\_id’, ‘flavour’, ‘food’, ‘price’, ‘type\_food’. The data are obtained from ‘EB-Build\_goods.sql’, which was modified so that it can be read as a csv file. Data frame named ‘df’ is containing data regarding transactions made at the bakeries. The ‘df’ data frame is from csv file ‘5000i.csv’ and it has a few attributes such as, ‘transaction\_id’, ‘item\_count’ and ‘item\_id’.

As the datasets are dirty and messy, some preprocessing task is done in order to clean it. Firstly, ‘Flavor’ attribute and ‘food’ attribute is merged into ‘item’ column, for example, ‘Flavor’= ‘Chocolate’ and ‘food’= ‘cake’ therefore ‘item’= ‘Chocolate’Cake’. All row in the ‘item’ column is then modified to delete “ ‘ “ symbols and replace it with a space, therefore, ‘item’= ‘Chocolate Cake’. Other preprocessing step used is, merging of data frames. The goods and df data frames are merged based on the item\_id. This results in a new data frame called ‘transaction\_bakery’, that has both data frame, ‘goods’ and ‘df’ together, and grouped based on receipts.

Lastly, other than the ‘item\_id’ and ‘items’ attributes in data frame ‘goods’ are ignored. This is done due to because they deemed unnecessary towards the association rules mining objective.

**RULE MINING PROCESS**

Technique used to uncover insights and association rules for this project is apriori algorithm from ‘arules’ library. The apriori algorithm was performed on ‘transaction\_bakery’, which is a list of items grouped according to transaction\_id. Parameter was set to a support of 0.005 and confidence of 0.5. After that, the rules are sorted by lift. Then, we removed some redundant rules after finding them.

**RESULTING RULES**

A total number of 41 rules were obtained from our rule mining process. It is a collection of rules named ‘cleaned\_rules’ which contains the information of associations between a certain item or itemset with another item, along with the support, confidence and lift values for each association. We have found certain itemsets that has a relationship with other items.

There are a few rules that has a confidence of 1. The first rule that we have obtained is that if a transaction contains Green Tea,Lemon Lemonade,Raspberry Lemonade, then it will also contain Raspberry Cookie. Secondly, a transaction that contains Lemon Lemonade,Raspberry Cookie,Raspberry Lemonade will also contain Lemon Cookie. There are a total of 5 rules with confidence of 1 and lift above 15 that contains the combination of Raspberry flavoured items and Lemon flavoured items. From these five rules, four of them contains Green Tea. There are a total of 15 rules that contains various combinations of Green Tea, Raspberry flavoured items and Lemon flavoured items. From this, we can say that a person will buy Green Tea with Raspberry and Lemon flavoured items. These three items has a high association with one another. Figure 1 shows the screenshot of the results obtained for this rule.

There are also a few rules that show Apple Croissant, Apple Danish, Cherry Soda and Apple Tart are highly associated with one another. With a confidence of more than 0.9 and lift of 13.505509, the rules show that a person who buys Apple Croissant, Apple Danish and Cherry Soda will buy Apple Tart. Other than that, a person who buys Apple Danish and Apple Tart will buy Apple Croissant. There are a few more rules to prove this statement.

Another rule that we think is useful is where a person will buy Apple Pie if he buys Almond Twist and Coffee Eclair. There are two rules in which one of them with a confidence of 1 and lift of 12.787724, shows that Apple Pie is the consequent of Almond Twist, Coffee Eclair and Hot Coffee. Even so, there are rules with a low lift value for association between Apple Pie with Coffee Eclair, and also Almond Twist with Coffee Eclair independently.

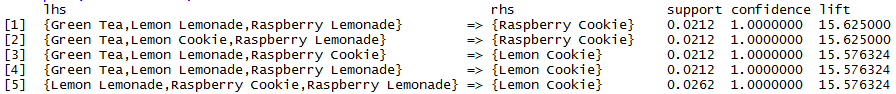


Figure 1

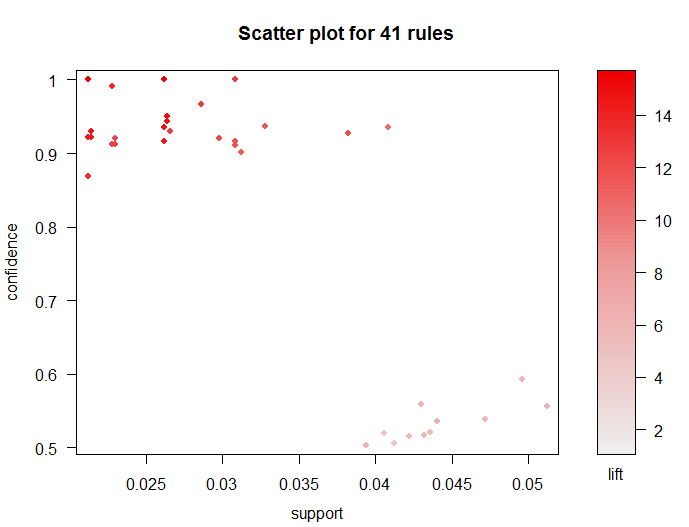


Figure 2: Scatter plot for final rules (after removing redundant rules)

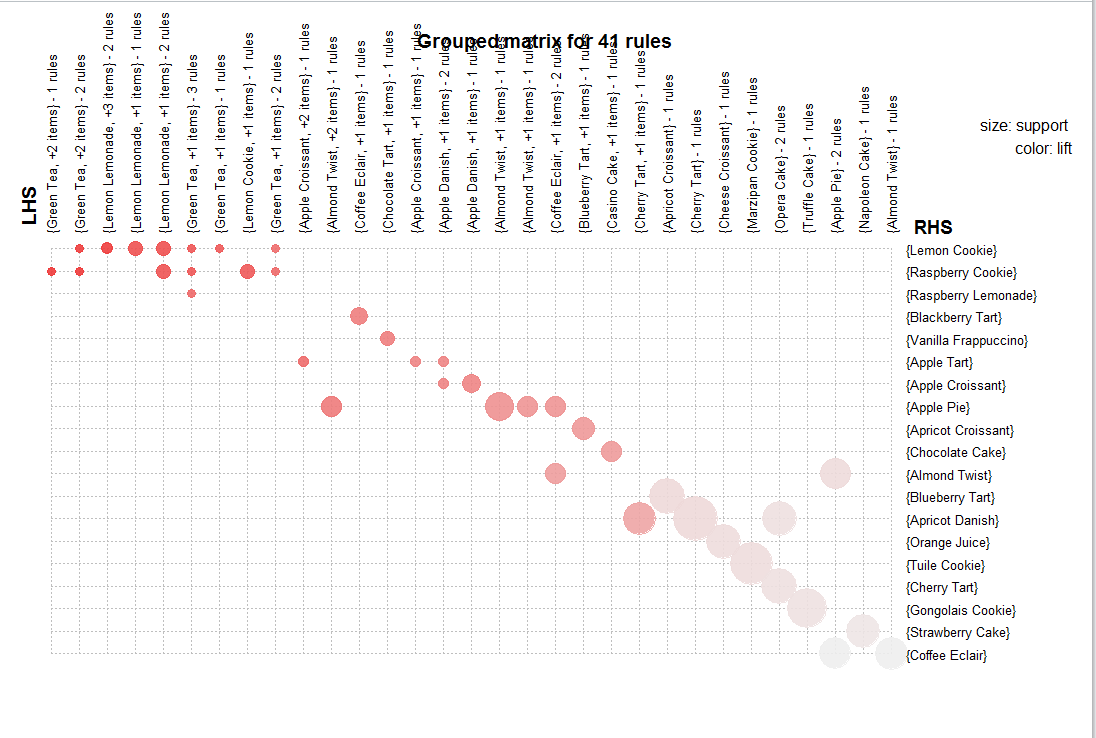


Figure 2: Grouped matrix with k = 41.

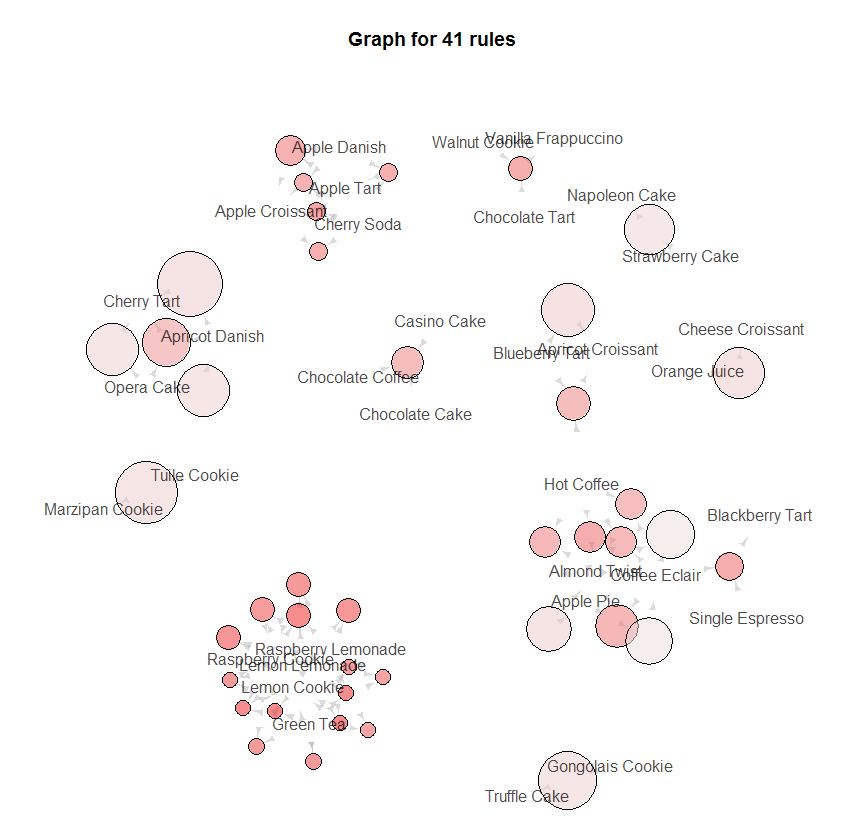


Figure 3: Graph-based visualization with items and rules as vertices.

**RECOMMENDATIONS**

Item combinations that have higher lifts should be put together to increase their sales. The items that did not have high lifts should be on promotion to boost sales.

From the rules obtained, the client should place Green Tea, Raspberry and Lemon flavoured items together. Other than that, the client should place Apple Danish, Apple Croissant, Apple Tart and Cherry Soda together. The client can also take action by placing Almond Twist, Apple Pie and Coffee Eclair together and also make promotion based on them since they have a low lift value.