CIDM 6355 Week 10 RapidMiner Lab - Association Rule Mining

Purposes:

- Understand what association rules are, how they are found, and the benefits of using them;
- Recognize the necessary data format for association rule mining;
- Illustrate rule support and rule confidence;
- Develop association rule models in RapidMiner;
- Interpret the rules generated by RapidMiner and explain their significance.

Dataset: Please download the dataset titled "marketbasket.csv" from WT Class to use it with this lab session. After completing the lab, please answer all the questions highlighted in yellow in LA5 Submission on WTClass.

Objectives (Business Understanding)

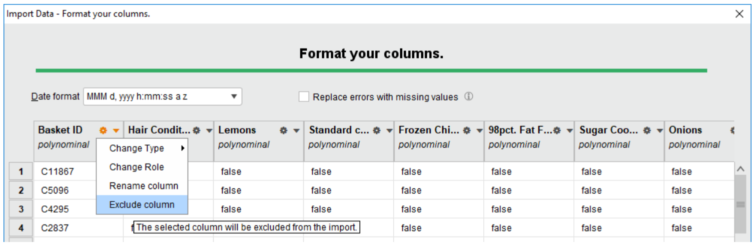
Customers usually purchase multiple items each time when they shop in a grocery store. The manager in a grocery store wants to know what products are usually sold together and then improve their advertising and product placement strategies to increase their sales and profit. Your goal is to help this manager identify those associations from existing transactions and then take advantages of them.

Data Understanding

In this lab, you will do market basket analysis for a grocery store. You will use the dataset ''marketbasket.csv". Open this file in Excel and take a look at it. You will find it has 303 columns (Basket ID and 302 items sold in this store) and 1362 rows (header row and 1361 transactions made at a grocery store). In the association rule mining, we do not need the first column, so it must be unselected or excluded before running the data. In each transaction row, a "true" value indicates the existence of that particular item in the customer's basket and a "false" indicates the absence of that particular item in the customer's basket.

Data Preparation

1. Import the ''marketbasket.csv" dataset to a new blank RM process using the Read CSV operator.
    1.1. Select the dataset in Step 1
    1.2. Use the default setting in Step 2-Specify your data format
    1.3. In step 3 – Format your columns, please notice two issues: 1) Basket ID cannot be included in your association rule model (you either exclude it here or unselect it later using the Select Attributes operator); 2) all other attributes are polynominal, which is not correct because each attribute has true or false. In addition, polynominal attributes cannot be analyzed by the rule mining operator in RM. We will deal with the first issue here (see the screenshot below) while tackle the second one later (you might want to manually change each attribute to binomial. However, it will be very time-consuming to do this for 302 attributes).



Import Data - Format your columns.

Format your columns.

Date format MMM d, yyyy h:mm:ss a z

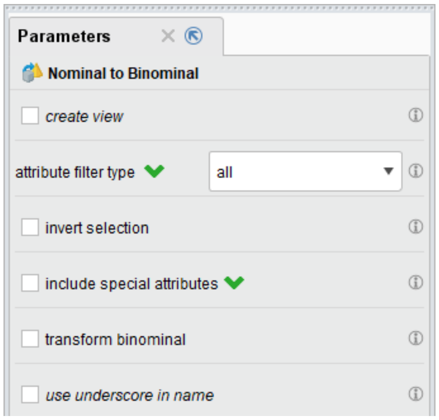
□ Replace errors with missing values

	Basket ID polynominal	Hair Condit... polynominal	Lemons polynominal	Standard c... polynominal	Frozen Chi... polynominal	98pct. Fat F... polynominal	Sugar Coo... polynominal	Onions polynominal
1	C11867		false	false	false	false	false	false
2	C5096		false	false	false	false	false	false
3	C4295		false	false	false	false	false	false
4	C2837	f			false	false	false	false

The selected column will be excluded from the import.

1.4. Once you select “Exclude column”, this first column is shaded. Then, click Finish.

2. Change all the 302 attributes to binomial using the operator Nominal to Binominal with the default parameters (see the screenshot in the right). Run this process and then take a look at the Statistics view to see if all the attributes are binominal. Note: If your RapidMiner shows that all the 302 attributes are binominal, you do not need to add this operator.



Parameters

Nominal to Binominal

□ create view	
attribute filter type	all
□ invert selection	
□ include special attributes	
□ transform binominal	
□ use underscore in name	

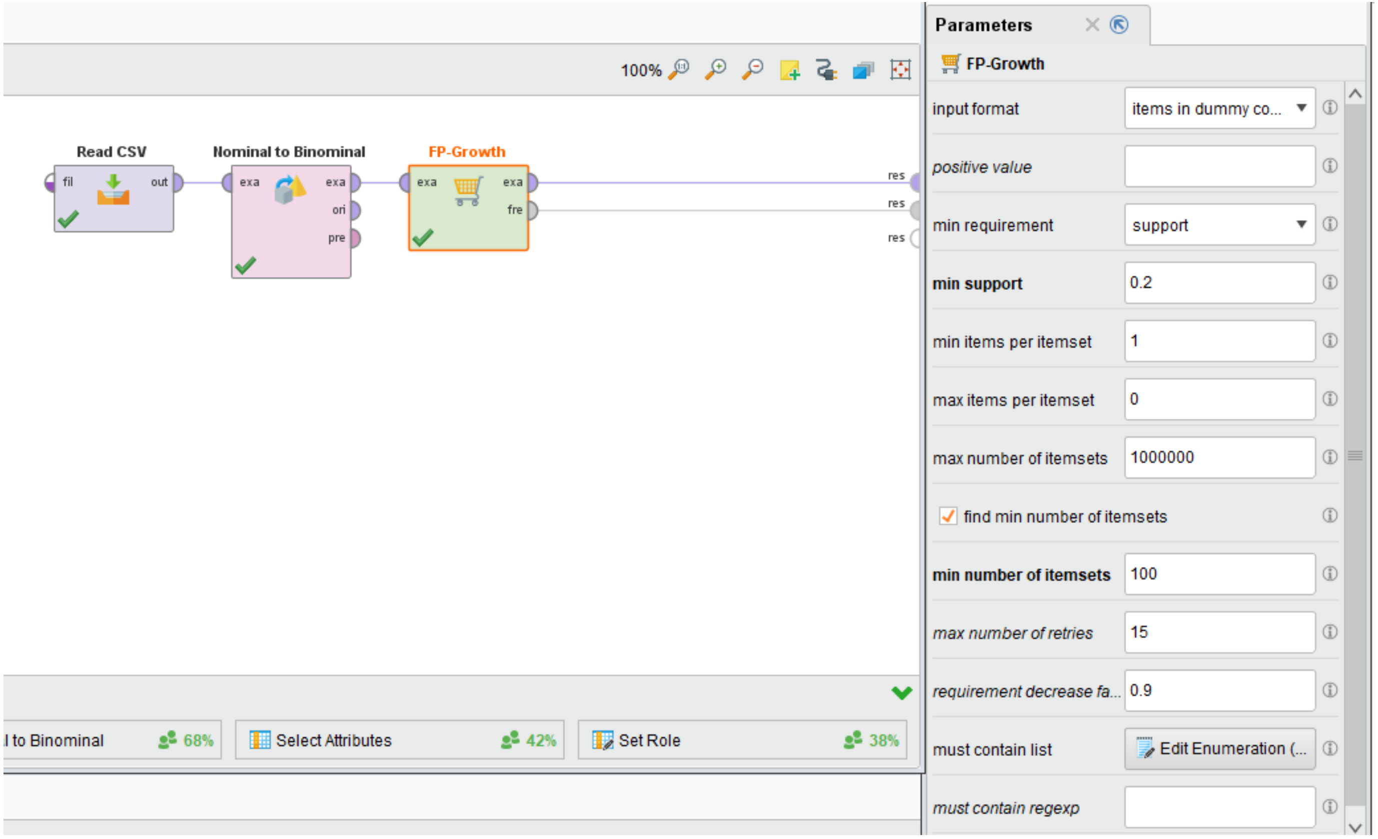
Note from RM: The Nominal to Binominal operator is used for changing the type of nominal attributes to a binominal type. This operator not only changes the type of selected attributes but it also maps all values of these attributes to binominal values i.e. true and false. For example, if a nominal attribute with name 'costs' and possible nominal values 'low', 'moderate', and 'high' is transformed, the result is a set of three binominal attributes 'costs = low', 'costs = moderate', and 'costs = high'. Only the value of one of these attributes is true for a specific example, the value of the other attributes is false. Examples of the original ExampleSet where the 'costs' attribute had value 'low', in the new ExampleSet these examples will have attribute 'costs=low' value set to 'true', value of 'cost=moderate' and ' cost=high' attributes will be 'false'. Numeric attributes of the input ExampleSet remain unchanged.

Modeling

As discussed in our lecture video, association rule mining is performed in two steps (operators): 1) frequent itemsets are generated using a search algorithm and then 2) find association rules from frequent itemsets. We are going to describe each step as below.

3. Generate Frequent Itemsets

3.1. In RapidMiner, the native algorithm to use is FP-Growth (Apriori is also supported, but only as a Weka module, under Extensions). Add the FP-Growth operator into your process and set its parameters as below.



Parameters

FP-Growth

input format	items in dummy co...
positive value	
min requirement	support
min support	0.2
min items per itemset	1
max items per itemset	0
max number of itemsets	1000000
☑ find min number of itemsets	
min number of itemsets	100
max number of retries	15
requirement decrease fa...	0.9
must contain list	Edit Enumeration (...
must contain regexp	

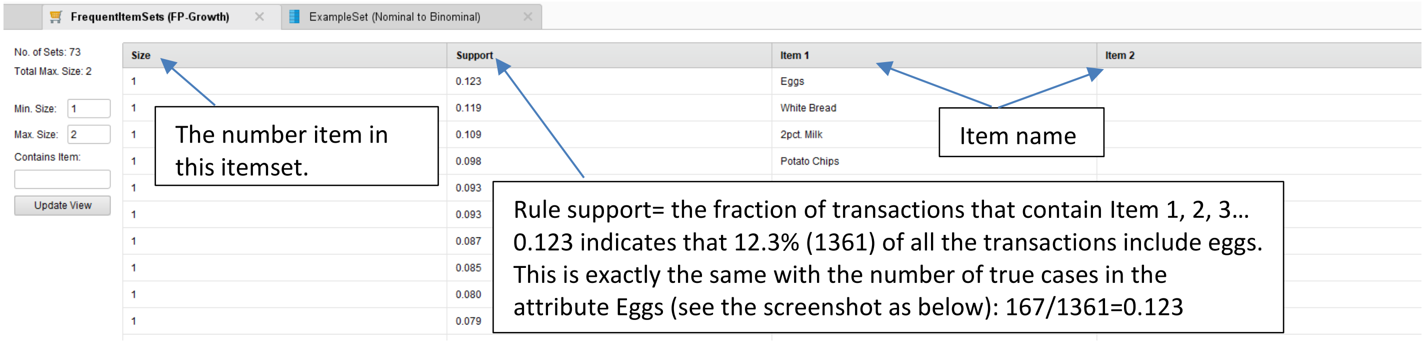
Note: that the default support threshold is .95, which is too high. Change this to 0.2. Notice that as long as "find min number of itemsets" is checked, RM will reduce the min support value to even lower till (close to) 100 frequent itemsets are identified. Save this process as Lab10_Step 3.1.

Note: RM 9.0 or later versions update this operator to make it support several different formats for the input data. For the columns, the three available input formats are illustrated in the second tutorial, together with necessary pre-processing. Here's the summary:

- item list in a column: All the items belonging to a transaction appear in a single column, separated by item separators, in a CSV-like format. As with CSV files, the items can be quoted, and escape characters are available. You can trim item names.
- items in separate columns: All the items belonging to a transaction appear in separate columns. For each transaction, the first item name appears in the first column, the second item name in the second column, etc. The number of columns corresponds to the basket with the maximum number of items. Missing values indicate no item. You can trim item names.
- items in dummy coded columns: Every item in the set of all items has its own column, and the item name is the column name. For each transaction, the binominal values (true/false) indicate whether the item can be found in the basket. If your data is binominal but does not identify the values as true/false, you may have to set the positive value parameter.

Our dataset is organized and structured as the third type. However, in a real case, a majority of original datasets belong to the first or second type.

3.2. Run the process and you will find the FrequentItemSets (similar to the one below). You can sort each column at the ascending or descending order.


FrequentItemSets (FP-Growth)
ExampleSet (Nominal to Binominal)
Update View

No. of Sets: 73
Total Max. Size: 2

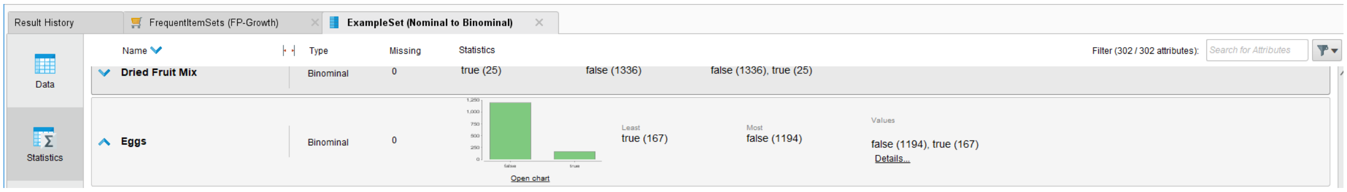
Min. Size: 1
Max. Size: 2
Contains Item:

Size	Support	Item 1	Item 2
1	0.123	Eggs	
1	0.119	White Bread	
1	0.109	2pct. Milk	
1	0.098	Potato Chips	
1	0.093		
1	0.093		
1	0.087		
1	0.085		
1	0.080		
1	0.079		

The number item in this itemset.

Item name

Rule support= the fraction of transactions that contain Item 1, 2, 3… 0.123 indicates that 12.3% (1361) of all the transactions include eggs. This is exactly the same with the number of true cases in the attribute Eggs (see the screenshot as below): 167/1361=0.123



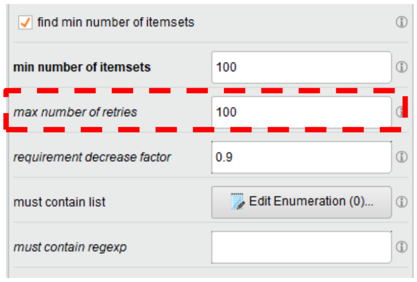
Name	Type	Missing	Statistics			
Dried Fruit Mix	Binominal	0	true (25)	false (1336)	false (1336), true (25)	
Eggs	Binominal	0		Least true (167)	Most false (1194)	Values false (1194), true (167)

Pause and Think: Please find out which **two-item** itemset is the most common one in the data (i.e., with the highest support and size =2). How many orders or transactions contain this most common **two-item** itemset? Note that you need to convert the *support* value to *the number of orders* to get the answer. Please show how you calculate it.

3.3. As you can see above, RM only gives us 73 itemsets, which is smaller than the number we set for *min number of itemsets*. You may wonder why. The reason is that we use the default number of *max number of retries* (i.e., 15).

Note: This parameter *max number of retries* is only available when find min number of itemsets is checked.

When automatically decreasing the value for minimum support / minimum frequency, this parameter determines how many times the Operator may decrease the value before giving up. Increase this number to get more results. The operator reduces minimum support value 15 times and then stops. That is why we only got 73 itemsets.



☑ find min number of itemsets

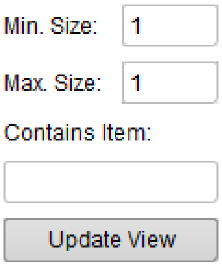
min number of itemsets	100
max number of retries	100
requirement decrease factor	0.9
must contain list	Edit Enumeration (0)...
must contain regexp	

3.4. Change the value of *max number of retries* (i.e., 15) to 100 and save your process as Lab10_Step3.4. Run this process and then you will have more than 100 itemsets.

Pause and Think: Take a look at your FrequentItemSets. What is the minimal support in your FrequentItemSets this time? How many one-item itemsets and two-item itemsets do you have, respectively?

Tip: you can set Min.Size and Max Size as 1 and then click Update View to quickly find the number of one-item itemsets. Similarly, you will find the number of two-item itemsets (see the seceenshots in the right).

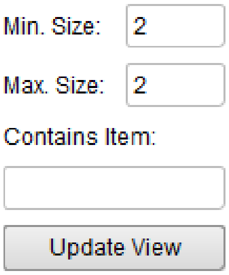
Think about why it is hard to see three-item or four-item itemsets in your results? Typically, as the size of an itemset increases, it will get harder to reach the threshold of the rule support.



Min. Size: 1

Max. Size: 1

Contains Item:



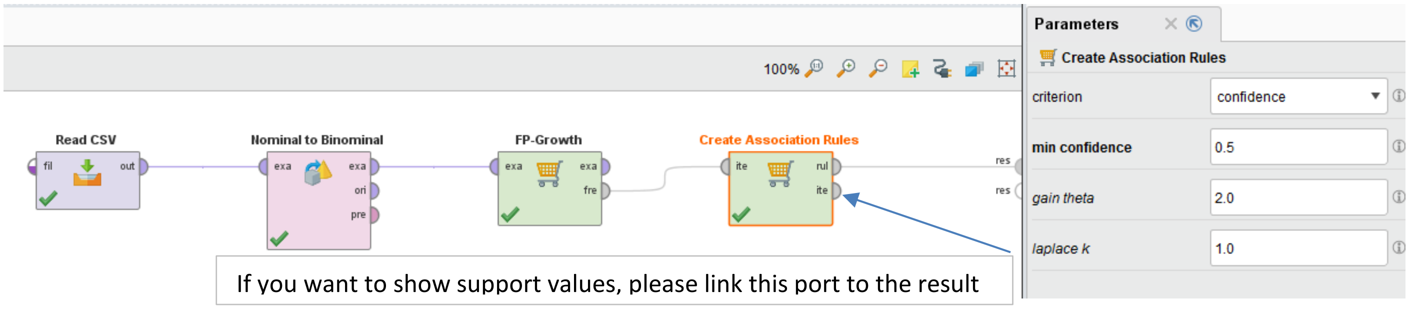
Min. Size: 2

Max. Size: 2

Contains Item:

4. Find Association Rules

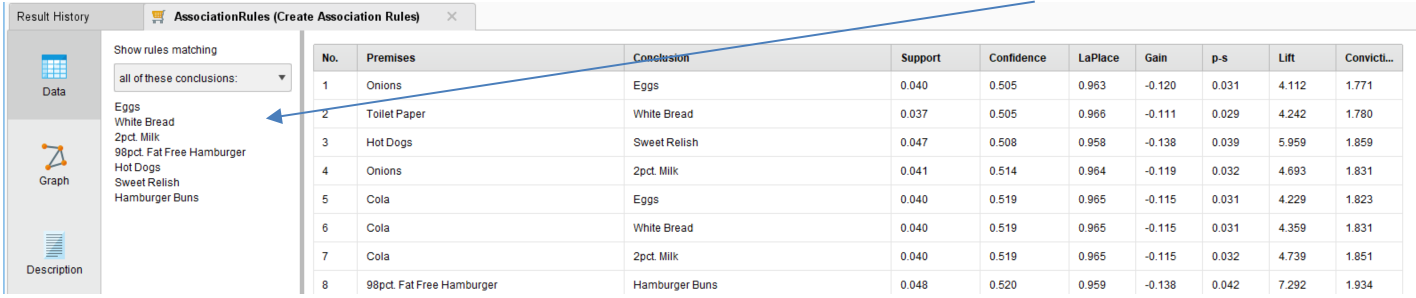
4.1. The second step in association rule mining is to find association rules from frequent itemsets. To do this, add the Create Association Rules operator into your model, as shown below. Note that min confidence is changed to 0.5. Save this process as Lab10_Step4.1.



criterion	confidence
min confidence	0.5
gain theta	2.0
laplace k	1.0

If you want to show support values, please link this port to the result

4.2. Run this process and you will find those rules are displayed in three different views: Data (or Table), Graph, and Description. First of all, in the Data view (see below), you will find Premises, Conclusion, Support, Confidence, and other criteria (see the last page for the description of each measure). In a rule, {X} → {Y}, X is the Premises and Y is Conclusion. You can also show particular rules using the filter in the left.



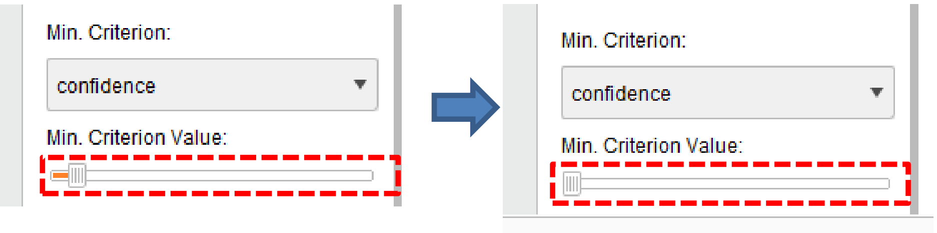
Show rules matching

all of these conclusions:

Eggs
White Bread
2pct. Milk
98pct. Fat Free Hamburger
Hot Dogs
Sweet Relish
Hamburger Buns

No.	Premises	Conclusion	Support	Confidence	LaPlace	Gain	p-s	Lift	Convicti...
1	Onions	Eggs	0.040	0.505	0.963	-0.120	0.031	4.112	1.771
2	Toilet Paper	White Bread	0.037	0.505	0.966	-0.111	0.029	4.242	1.780
3	Hot Dogs	Sweet Relish	0.047	0.508	0.958	-0.138	0.039	5.959	1.859
4	Onions	2pct. Milk	0.041	0.514	0.964	-0.119	0.032	4.693	1.831
5	Cola	Eggs	0.040	0.519	0.965	-0.115	0.031	4.229	1.823
6	Cola	White Bread	0.040	0.519	0.965	-0.115	0.031	4.359	1.831
7	Cola	2pct. Milk	0.040	0.519	0.965	-0.115	0.032	4.739	1.851
8	98pct. Fat Free Hamburger	Hamburger Buns	0.048	0.520	0.959	-0.138	0.042	7.292	1.934

4.3. In the bottom of the Data view, you find the Min. Criterion (see the screenshot in the right). In order to view all the rules generated, please move the bar to the very left.

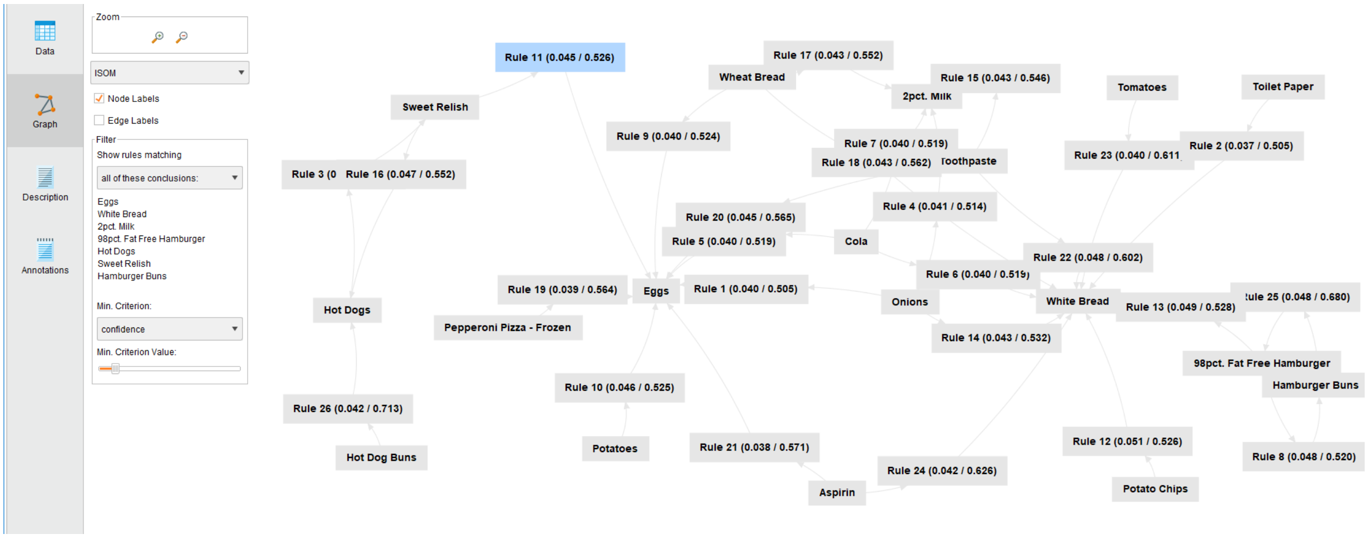


Min. Criterion:

confidence

Min. Criterion Value:

4.4. In the Graph view, you can view those associations in a graph (like a social network). As you can see one item could be conclusion or premises, depending on the arrow incoming or outgoing.



ISOM

☑ Node Labels

□ Edge Labels

Filter

Show rules matching

all of these conclusions:

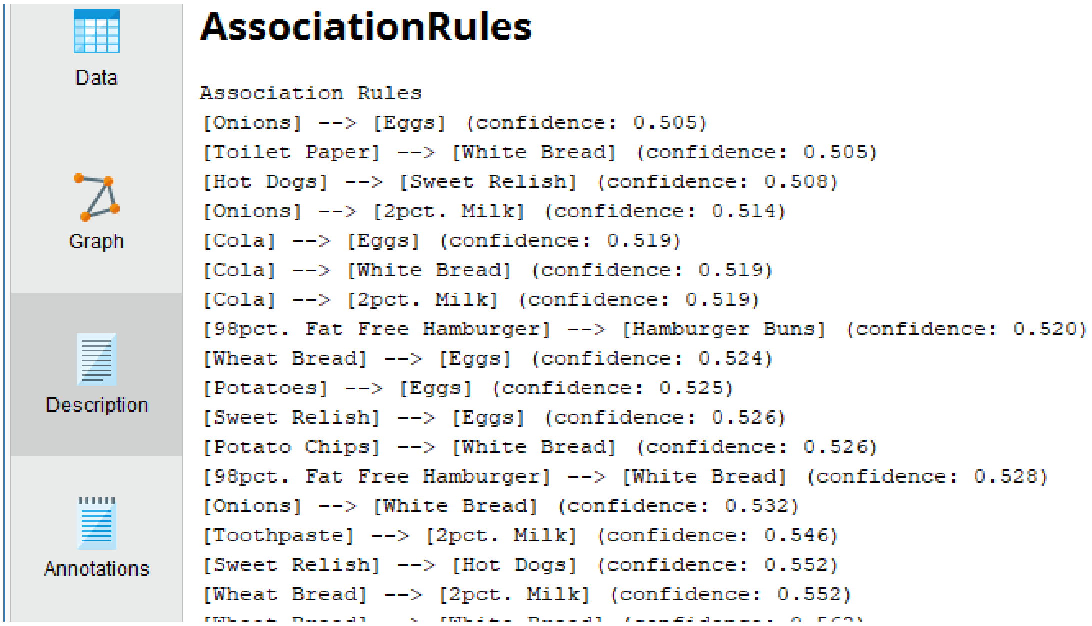
Eggs
White Bread
2pct. Milk
98pct. Fat Free Hamburger
Hot Dogs
Sweet Relish
Hamburger Buns

Min. Criterion:

confidence

Min. Criterion Value:

4.5. Finally, take a look at the Description view. These rules are displayed in the same order as the Data view (sorted by the confidence values at the ascending order).



AssociationRules

```
Association Rules
[Onions] --> [Eggs] (confidence: 0.505)
[Toilet Paper] --> [White Bread] (confidence: 0.505)
[Hot Dogs] --> [Sweet Relish] (confidence: 0.508)
[Onions] --> [2pct. Milk] (confidence: 0.514)
[Cola] --> [Eggs] (confidence: 0.519)
[Cola] --> [White Bread] (confidence: 0.519)
[Cola] --> [2pct. Milk] (confidence: 0.519)
[98pct. Fat Free Hamburger] --> [Hamburger Buns] (confidence: 0.520)
[Wheat Bread] --> [Eggs] (confidence: 0.524)
[Potatoes] --> [Eggs] (confidence: 0.525)
[Sweet Relish] --> [Eggs] (confidence: 0.526)
[Potato Chips] --> [White Bread] (confidence: 0.526)
[98pct. Fat Free Hamburger] --> [White Bread] (confidence: 0.528)
[Onions] --> [White Bread] (confidence: 0.532)
[Toothpaste] --> [2pct. Milk] (confidence: 0.546)
[Sweet Relish] --> [Hot Dogs] (confidence: 0.552)
[Wheat Bread] --> [2pct. Milk] (confidence: 0.552)
[Wheat Bread] --> [White Bread] (confidence: 0.562)
```

5. Understand Association Rules**:**
   5.1. Take a look at your Data view. How many rules are identified?

   5.2. Among them, which rule has the highest confidence? What does this rule mean?

   5.3. In the descriptive view, please find the confidence for the rules [Hot Dogs] --> [Sweet Relish] and [Sweet Relish] --> [Hot Dogs]. What does each rule mean? Are the two confidence values the same? Why?

Deployment

6. Use Association Rules to Solve Real-World Problems
   The following questions are not required in LA5. You can share your thought about the following questions in our Data Miner Community.
   6.1. If you are going to provide some suggestions about shelf management for this store, which item(s) would you consider storing on the same shelf with “Hamburger Buns” for the convenience of your customers? Why? Hint: in addition to those rules you find, you still need to think about the reality (e.g., some products require refrigerator).

   6.2. Which of the following is a better idea? Why?
   - Keep "2pt Milk" in only one aisle in the store.
   - Split "2pt Milk" to smaller batches and distribute it to several different spots in the store. If the second one is a better idea, near which items would you put "2pt Milk" to?

   6.3. Assume that reducing the price of an item (i.e. putting it on sale) will increase the demand for this item, which may also increase the demand for associated items. Suppose that White Bread products are overstocked. In addition to White Bread, which other item(s) would you consider running promotions on to reduce the stock of white bread? Why?

Appendix

The operator Create Association Rules have multiple criteria (Source: RM Help file).

- confidence: The confidence of a rule is defined conf(X implies Y) = supp(X ∩Y)/supp(X) . Be careful when reading the expression: here supp(X∩Y) means "support for occurrences of transactions where X and Y both appear", not "support for occurrences of transactions where either X or Y appears". Confidence ranges from 0 to 1. Confidence is an estimate of Pr(Y | X), the probability of observing Y given X. The support supp(X) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset.
- lift: The lift of a rule is defined as lift(X implies Y) = supp(X ∩ Y)/((supp(Y) x supp(X)) or the ratio of the observed support to that expected if X and Y were independent. Lift can also be defined as lift(X implies Y) =conf(X implies Y)/supp(Y). Lift measures how far from independence are X and Y. It ranges within 0 to positive infinity. Values close to 1 imply that X and Y are independent and the rule is not interesting.
- conviction: conviction is sensitive to rule direction i.e. conv(X implies Y) is not same as conv(Y implies X). Conviction is somewhat inspired in the logical definition of implication and attempts to measure the degree of implication of a rule. Conviction is defined as conv(X implies Y) =(1 - supp(Y))/(1 - conf(X implies Y))
- gain: When this option is selected, the gain is calculated using the gain theta parameter.
- laplace: When this option is selected, the Laplace is calculated using the laplace k parameter.
- ps: When this option is selected, the ps criteria is used for rule selection.