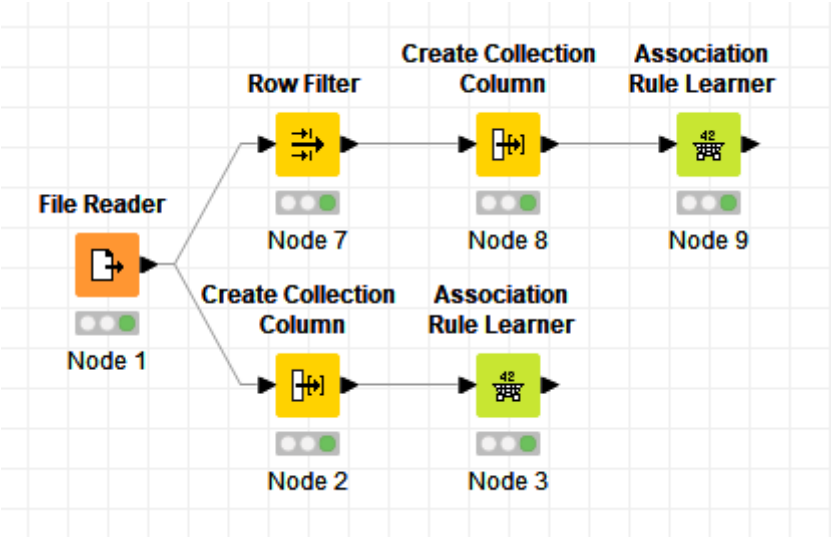
Association Rules

Introduction

For this project, the dataset from a supermarket containing products from transactions over one month was analyzed, it was the month of February, using market basket techniques.

Excel was used to analyze the dataset and knime to study the frequent item sets and association rules.



Analyzing the dataset in more detail, i could understand better the way this supermarket worked and the type and frequency of the clients and products sold.

The dataset has 168 different items and 50 % of sales are for about 12 % of the products, around 20 products.

I wanted to see how many transactions were made every day and, as can be seen below, this number was constant on the first 20 days with 325 transactions and after that increasing for 415 on the rest of the days excluding the last one that had 430.

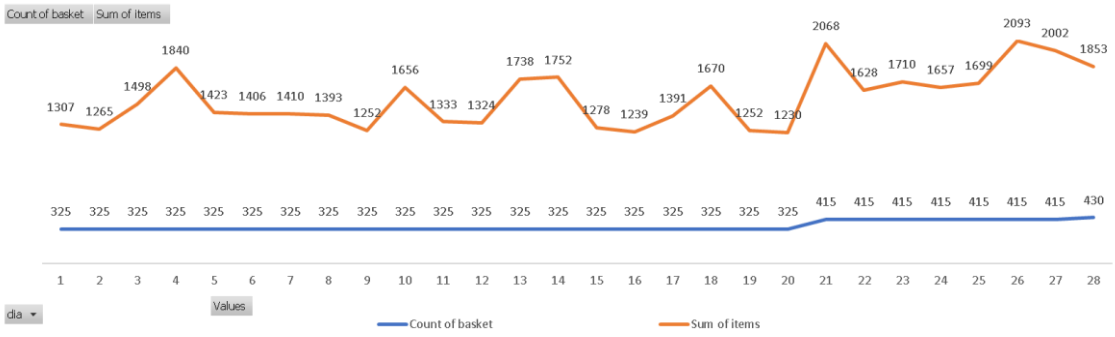


Fig. 1 Count of baskets and sum of items per day

The number of items sold per day, represented with the orange line on the upper image, had an interesting behavior because with this line we could identify what we think are the weekends, for example, the days 3 and 4 have an increase on the products sold and then we can see an identical increase on the days 10, 17 and 18 and on the days 25 and 26, but off course i could not know this for sure. I could see another two increases for one or two days on the days 13, 14 and 21, this might be due to some holidays. And finally, i could note that the end of the month had more products sold than the beginning and maybe that is because of the salary income that the clients might have.

Another interesting fact that noticed, is that the baskets that have at most seven items make more than 80% of the transactions on almost every day except on the days, that i see before, with more products sold. Those days have more baskets with more than seven items, but in those cases, the lowest value is 72% on the day 14, and this can lead us to think that this supermarket is a small place where people do small purchases.

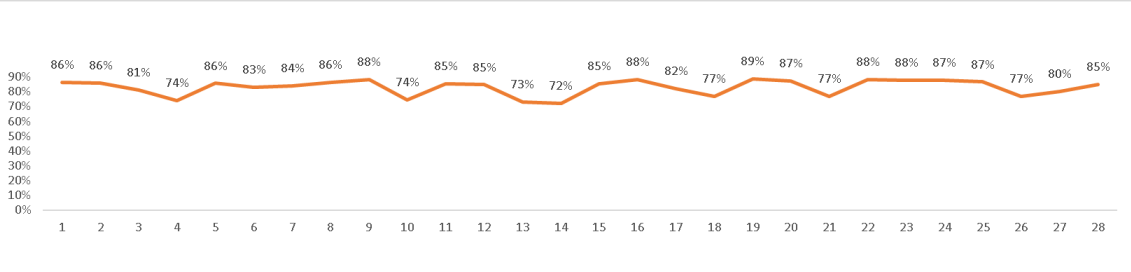


Fig. 2 Percentage of baskets per day with at most 7 items

Therefore, with this understanding of the data i tried to answer the questions below.

1. Study of the variation of the number of frequent item sets (free, closed and maximal) varying the support.

Since the maximum support is 25,6% for the frequent item set with Whole Milk, i decided to study the number of frequent item sets for each value of support from 1% to 25%.

I choose the minimum value of support of 1% because i think that with a value lower than that i stop having enough information on the relationship between the items and because of that, no good conclusions can be achieved.

Value of Support	Number of frequent item sets (per type)		
	Free	Closed	Maximal
0,01	333	333	243
0,02	122	122	91
0,03	63	63	50
0,04	41	41	34
0,05	31	31	27
0,06	21	21	19
0,07	19	19	17
0,08	13	13	13

0,09	10	10	10
0,1	8	8	8
0,11	6	6	6
0,12	5	5	5
0,13	5	5	5
0,14	4	4	4
0,15	4	4	4
0,16	4	4	4
0,17	4	4	4
0,18	3	3	3
0,19	2	2	2
0,2	1	1	1
0,21	1	1	1
0,22	1	1	1
0,23	1	1	1
0,24	1	1	1
0,25	1	1	1

I can see that, for the same confidence level, the number of closed item sets, which is always the same as the free item sets, are decreasing with the increase of the support. The rate of the decreasing lowers with the increase of the support because the number of frequent item sets is decreasing.

Being the maximal frequent item sets a subset of the closed item sets that have none of its super sets frequent, i can see on the table, that its number is lower than the one for the closed item sets for the lower levels of support, as expected.

When support reaches the value of 8%, i could see that the numbers are equal and this is explained by the fact that this frequent item sets doesn't have frequent super sets.

So maybe a support at 8% can be a good value to study the most important frequent item sets.

2. Study of the variation of the number of association rules fixing the support.

Fixing the minimum support at 0,01, beginning with the minimum confidence of 0,01 to the maximum of 0,55 and registering the number of association rules at every 0,05 increment of confidence i got the results on the table below.

Confidence	Number of Association Rules
0,01	522
0,05	513
0,1	427
0,15	311
0,2	231
0,25	170
0,3	125
0,35	89
0,4	62
0,45	31
0,5	15
0,55	7

Analyzing the numbers on the table, i see that the number of association rules are decreasing with the increase in the confidence, as expected, because i see a decrease of the most frequent associations between the products.

For instance, i can see that a client that already have yogurt and root vegetables will be more probable to buy whole milk (56,3%) than other vegetables (50%), as can be seen below.

Table "default" - Rows: 15 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	Support	Con...	Lift	Consequent	implies	Items
rule1	0.01	0.586	3.03	other vegetables	<---	[citrus fruit,root vegetables]
rule6	0.012	0.585	3.021	other vegetables	<---	[root vegetables,tropical fruit]
rule0	0.01	0.582	2.279	whole milk	<---	[yogurt,curd]
rule3	0.011	0.574	2.245	whole milk	<---	[butter,other vegetables]
rule4	0.012	0.57	2.231	whole milk	<---	[root vegetables,tropical fruit]
rule11	0.015	0.563	2.203	whole milk	<---	[yogurt,root vegetables]
rule7	0.012	0.553	2.162	whole milk	<---	[domestic eggs,other vegetables]
rule2	0.011	0.525	2.053	whole milk	<---	[yogurt,whipped/sour cream]
rule8	0.013	0.523	2.047	whole milk	<---	[rolls/buns,root vegetables]
rule10	0.014	0.518	2.025	whole milk	<---	[pip fruit,other vegetables]
rule13	0.015	0.517	2.025	whole milk	<---	[yogurt,tropical fruit]
rule14	0.022	0.513	2.007	whole milk	<---	[yogurt,other vegetables]
rule12	0.015	0.507	1.984	whole milk	<---	[other vegetables,whipped/sour cream]
rule5	0.012	0.502	2.595	other vegetables	<---	[rolls/buns,root vegetables]
rule9	0.013	0.5	2.584	other vegetables	<---	[yogurt,root vegetables]

Fig. 3 Association rules for the minimum confidence of 0,5

Table "default" - Rows: 7 | Spec - Columns: 6 | Properties | Flow Variables

Row ID	Support	Con...	Lift	Conseq...	implies	Items
rule1	0.01	0.586	3.03	other veget...	<---	[citrus fruit,root vegetables]
rule4	0.012	0.585	3.021	other veget...	<---	[root vegetables,tropical fruit]
rule0	0.01	0.582	2.279	whole milk	<---	[yogurt,curd]
rule2	0.011	0.574	2.245	whole milk	<---	[butter,other vegetables]
rule3	0.012	0.57	2.231	whole milk	<---	[root vegetables,tropical fruit]
rule6	0.015	0.563	2.203	whole milk	<---	[yogurt,root vegetables]
rule5	0.012	0.553	2.162	whole milk	<---	[domestic eggs,other vegetables]

Fig. 4 Association rules for the minimum confidence of 0,55

With this analysis, i could see that for low values of confidence i have association rules with a less important meaning and with a too high value for confidence i can leave out some important association rules that must be studied.

Therefore, the value of confidence used must have in attention to not only the confidence but also the lift and support values, as can be seen on the answers of the next questions.

3. Selection of one of the rules with largest lift. Justification of the value of the lift.

Considering the fig.5 i choose the lift (beef → root vegetables).

The lift value is a measure of importance of a rule. Lift is the ratio between confidence and the expected value for the confidence. In this case with the lift value greater than 1 (3.04) indicate a positive relation, it means that the rule beef and the rule root vegetables appear more often

together than expected, this means that the occurrence of the rule beef has a positive effect on the occurrence of the rule root vegetables.

Row ID	Support	Confidence	Lift	Consequent	implies	Items
rule21	0.01	0.359	3.295	root vegetables	<---	[citrus fruit,other vegetables]
rule7	0.01	0.234	3.267	whipped/sour cream	<---	[yogurt,other vegetables]
rule78	0.012	0.343	3.145	root vegetables	<---	[other vegetables,tropical fruit]
rule148	0.017	0.331	3.04	root vegetables	<---	[beef]
rule20	0.01	0.586	3.03	other vegetables	<---	[citrus fruit,root vegetables]
rule77	0.012	0.585	3.021	other vegetables	<---	[root vegetables,tropical fruit]
rule192	0.023	0.31	2.842	root vegetables	<---	[whole milk,other vegetables]
rule2	0.01	0.385	2.761	yogurt	<---	[whole milk,curd]

Fig. 5 subset with largest lift

4. What is the conviction of the rule used in 3)?

The conviction is calculated consider the function $\text{conv}(X \Rightarrow Y) = \frac{1 - \sup(Y)}{1 - \text{conf}(X \Rightarrow Y)}$.

Can be interpreted as the ratio of the expected frequency that beef occurs without root vegetables, it means the frequency that the rule makes an incorrect prediction. In this example the rule show that beef → root vegetables would be incorrect 33% more often with random association.

In this example the sup(root vegetable) is 0.1089 and conf(beef→root vegetable) is 0.331.

5. Are there differences among the frequent item sets in the beginning of the month and end of month?

I have created several scenarios assuming that purchasing behavior is impacted by the monthly or weekly pay cycle, or even by events like holidays and weekends:

Small basket – basket less than or equal to 7 items

Large basket – the others

Event - Assuming that the increase in “larger baskets” is related to some event, i analyze these days separately

End of month – the days when there is an increase in the number of baskets, subset day 21 to day 28.

Beginning of month - i have two branches;

1) to maintain consistency in relation to the weekly cycle at the end of the month, day 05 to day 12.

2) ignore the weekly cycle, subset day 01 to day 08.

Note: For this we used a support equal to 0,01.

Considering the fig.6 with a top 15 of most frequent item sets, the reflections are:

Most articles are common in all subsets. This is independent i) of the period of the month (beginning or end) and ii) of the type of purchase, be it a convenience purchase (small basket) or a larger purchase (larger basket).

This is expected to happen once small baskets represent more than 80% of sales.

[Pastry] has a strong presence in the small baskets, which is reflected in the subsets beginning and end of the month.

[Canned beer] has a strong presence in the small baskets, and more influence in beginning of month.

[newspaper] are present in the "small basket", however their importance is diluted in the total of transactions, being only relevant in the subsets: event and the end of the month.

only baskets at 7 items	
Support	Items
0.196	[whole milk]
0.156	[rolls/buns]
0.150	[soda]
0.131	[other vegetables]
0.093	[yogurt]
0.088	[bottled water]
0.080	[shopping bags]
0.080	[canned beer]
0.073	[bottled beer]
0.070	[tropical fruit]
0.070	[root vegetables]
0.068	[pastry]
0.068	[sausage]
0.064	[newspapers]
0.054	[citrus fruit]

only baskets <> 7items	
Support	Items
0.543	[whole milk]
0.498	[other vegetables]
0.363	[yogurt]
0.318	[rolls/buns]
0.299	[root vegetables]
0.293	[soda]
0.277	[whole milk, other vegetables]
0.275	[tropical fruit]
0.220	[citrus fruit]
0.220	[sausage]
0.218	[bottled water]
0.211	[whipped/sour cream]
0.210	[yogurt, whole milk]
0.199	[pip fruit]
0.195	[fruit/vegetable juice]

event	
Support	Items
0.287	[whole milk]
0.248	[other vegetables]
0.185	[soda]
0.173	[rolls/buns]
0.161	[yogurt]
0.143	[root vegetables]
0.128	[bottled water]
0.114	[tropical fruit]
0.105	[whole milk, other vegetables]
0.103	[newspapers]
0.101	[sausage]
0.101	[shopping bags]
0.100	[whipped/sour cream]
0.099	[citrus fruit]
0.094	[bottled beer]

beginning of month 01 to 08	
Support	Items
0.258	[whole milk]
0.202	[rolls/buns]
0.192	[other vegetables]
0.174	[soda]
0.146	[yogurt]
0.126	[bottled water]
0.108	[root vegetables]
0.100	[tropical fruit]
0.099	[shopping bags]
0.095	[citrus fruit]
0.093	[pastry]
0.089	[sausage]
0.078	[canned beer]
0.078	[whipped/sour cream]
0.077	[bottled beer]

beginning of month 05 to 12	
Support	Items
0.241	[whole milk]
0.186	[rolls/buns]
0.185	[other vegetables]
0.182	[soda]
0.137	[yogurt]
0.109	[bottled water]
0.103	[root vegetables]
0.102	[shopping bags]
0.097	[tropical fruit]
0.095	[pastry]
0.091	[sausage]
0.086	[pip fruit]
0.084	[canned beer]
0.080	[bottled beer]
0.079	[citrus fruit]

end of month 21 to 28	
Support	Items
0.258	[whole milk]
0.190	[other vegetables]
0.178	[rolls/buns]
0.162	[soda]
0.145	[yogurt]
0.108	[bottled water]
0.106	[tropical fruit]
0.104	[root vegetables]
0.103	[shopping bags]
0.099	[sausage]
0.090	[pastry]
0.083	[citrus fruit]
0.081	[newspapers]
0.079	[bottled beer]
0.075	[whole milk, other vegetables]

Fig. 6 top 15 of most frequent item sets

Conclusion

With this project i had the opportunity to understand with a real world example the way the sales of a company work and the support that the management of those companies can have, with frequent patterns recognition and association rules estimation to make the everyday decisions needed to keep the business running well.

With the first question, i could see how the study of the number of frequent item sets and the different types of item sets is important to understand the products that are sold more to the clients and how the variation of support can have an impact on that understanding because with a too low support i don´t have enough information about the relationships between the items and with a too high one it can ignore important relations.

Solving the second question i could see the impact that the confidence level can have on the number of association rules estimated. Once again, i could see that with a too low value, i have many rules computed but some of them not very meaningful and with a high value for the confidence i could lose some information about important association rules.

With the third question it's possible to understand the relation between rules and in this case of the supermarket, redefine the place of products consider the potential of joint sales or events to promote cross sales. However, there is an associated error when I am predicting the occurrence, this is where question four comes in which allows to measure the deviation or percentage of an incorrect prediction.

With the last question, i could see that i could use the market basket analysis techniques for parts of the dataset to focus my attention in important parts of the month for this business, to have a better understanding of the behaviors of the clients and products sold in those specific days.