School of Computer Sciences

Universiti Sains Malaysia

CDS513 – PREDICTIVE BUSINESS ANALYTICS

Academic Session 2020/2021 Semester 2

Assignment 1

C1 : Market Basket Analysis, Topic 2 : Retail dataset

Analysing Grocery Retail Dataset By Using Market Basket Analysis

Name	Matric No.
Nurul Afina binti Abd Ghani	P-COM0166/19

1.0 Introduction

Today's competition forces consumer goods manufacturers and retailers to differentiate from their competitors by offering goods or services that are tailored towards one or more subgroups or segments of the market. The retailers in the Fast-moving Consumer Goods (FMCG) sector is however highly limited in their ability to segment the market and to focus on the most promising segments, since the typical attraction of retail store's attraction area is too small to overlook a subgroup within the store's attraction area. Nevertheless, if different customer segments, in terms of their shopping behaviour, can be identified, these segments could then be treated differently in terms of marketing communication for example in terms of pricing and promotion, in order to achieve greater overall effect.

Market Basket Analysis is a data mining technique, and it has been widely used in marketing and e-commerce field to discover the association between products bought together by customers. It helps businesses in increasing their sales by analysing the purchasing behaviour of customers and pitching the right customer with the right product.

1.1 Background of the problem domain

Market basket analysis evaluates the many products that shoppers place in their baskets in order to better understand shopper behaviour and influence marketing tactics (Kamakura, 2012). While there are variety of ways to collect products on a shopping trip such as using shopping cart, basket, or hand-carry, a shopping basket is generally defined as a set of products purchased together on a single occasion. Market basket analysis is extremely useful for retailers because it provides a practical approach to better understanding how to manage and grow their brands. Retailers have three sources of growth to consider – either increasing the number of baskets, increasing the amount purchased in each basket or basket size, or increasing the frequency of specific baskets especially for baskets with high value categories.

While there are numerous ways to grow a brand or a business, some may be more feasible than others. Marketers will be able to develop more informed growth strategies if they have a clear understanding of the shopping basket patterns that should be expected in-store and how they interact with performance measures. In this assignment, we analyse a dataset of purchases from a grocery retail store. We want to discover the buying pattern and what products that are usually purchased together, and to discover a better way to visualise and analyse the results.

2.0 Market Basket Analysis

2.1 Methodology

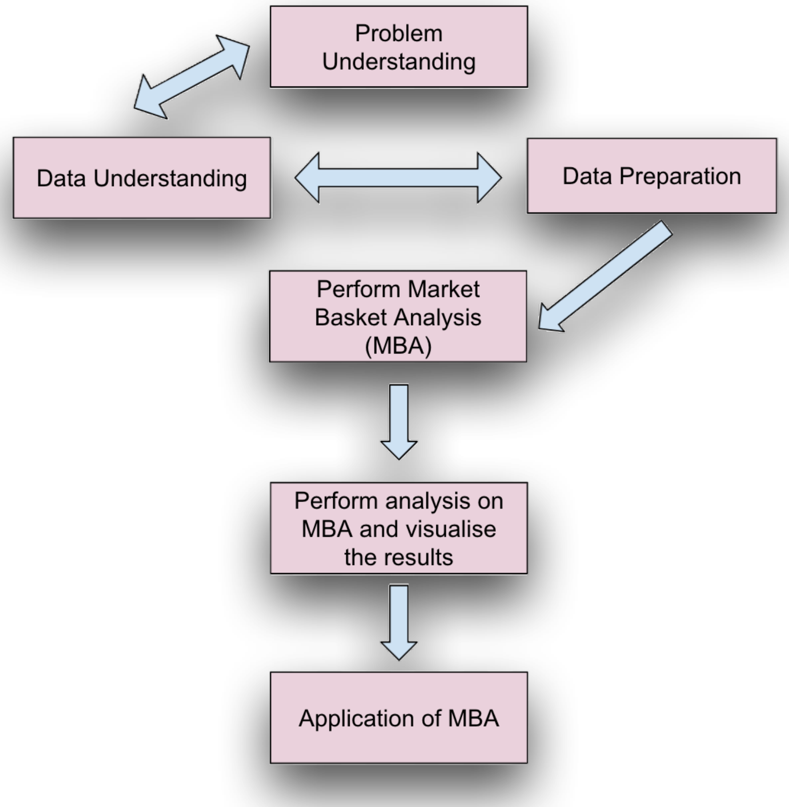

Problem Understanding
Data Understanding
Data Preparation
Perform Market Basket Analysis (MBA)
Perform analysis on MBA and visualise the results
Application of MBA

Figure 1: Framework of the assignment

Figure above describes the framework of the assignment. Jupyter Notebook (Python) will be used to clean and prepare the data. For the analysis portion, Rapid Miner and Python will be used to perform Market Basket Analysis.

Market basket analysis is a useful method for learning about customer buying patterns and preferences in the retail environment as it will uncover the relationship between two or more attributes. Specifically, the aim of the analysis is to figure out about certain patterns or groups of products that are likely to be purchased together and quantitatively report on these findings with the use of support, confidence and lift metrics. The end product of the Market Basket Analysis will be a set of association rules that describes the relationships between the products as a whole.

Association Rules Mining is a two step process, the first one is to find frequent item sets, and the second one is to generate strong association rules from the frequent item sets. Frequent item sets can be found using two methods, namely Apriori Algorithm and FP growth algorithm. Apriori algorithm generates all item sets by scanning the entire dataset. FP growth algorithm however only generates the frequent item sets according to the minimum support defined by the user. Since Apriori scans the whole dataset multiple times, it takes more time to generate the association rules, based on the size of the dataset. On the other hand, the FP growth algorithm doesn’t scan the whole dataset multiple times, hence, the FP growth algorithm is much faster than the Apriori algorithm.

To perform the Market Basket analysis (MBA), we will first create the Association Rules (AR). Then we will calculate Support, Confidence and Lift for each AR. Next, the minimum support will be determined and the number of AR will be selected to give the meaning of the associations. For Analysis of MBA, association graphs will be plotted with the best AR that reflects the optimal correlation, and other methods of analysis. After that, certain items will be selected for focus and analyse the pattern and impact. Finally, the results will be visualised.

2.2 Dataset and Preparation

The dataset that is used for this project is a Retail Dataset from Kaggle. This dataset is a table type dataset and there are 119 items and 7501 rows that represent transactions. In the below table, we can see the sum and support count of selected 8 products, top 5 and bottom 5 of list in descending order, in the frequency of how many they appear to be bought in the dataset. In this assignment, we assume that this dataset is a total of one week of transaction of a retail grocery store.

Table 1: Item Frequency and Support Count

Item	Quantity	Support Count (Quantity/Total Number of Transactions)
Mineral Water	1788	0.2384
Eggs	1348	0.1797
Spaghetti	1306	0.1741
French Fries	1282	0.1709
Chocolate	1229	0.1638
…		
Tea	29	0.0037
Bramble	14	0.0019
Cream	7	0.0009
Napkins	5	0.0007
Water Spray	3	0.0004

We need to prepare the data and make some changes on the dataset. Each row represents a transaction with items that were purchased. Each line is called a transaction and each column in a row represents an item. For each transaction, there can be only distinct items without repeating entries. This allows for us to create a binary (0,1) representation whether a particular item was purchased under a specific transaction.

In order to create a binary representation using Python, first, we convert the dataframe into a list of lists,

```
transactions = []
for i in range(0, len(trans_df)):
        transactions.append([str(trans_df.values[i,j])  for  j  in
range(0, 20) if str(trans_df.values[i,j])!='0'])
```

Then, by using Transaction Encoder, the dataset can be transformed into a logical data frame. Each column represents an item and each row represents a transaction for one purchase.

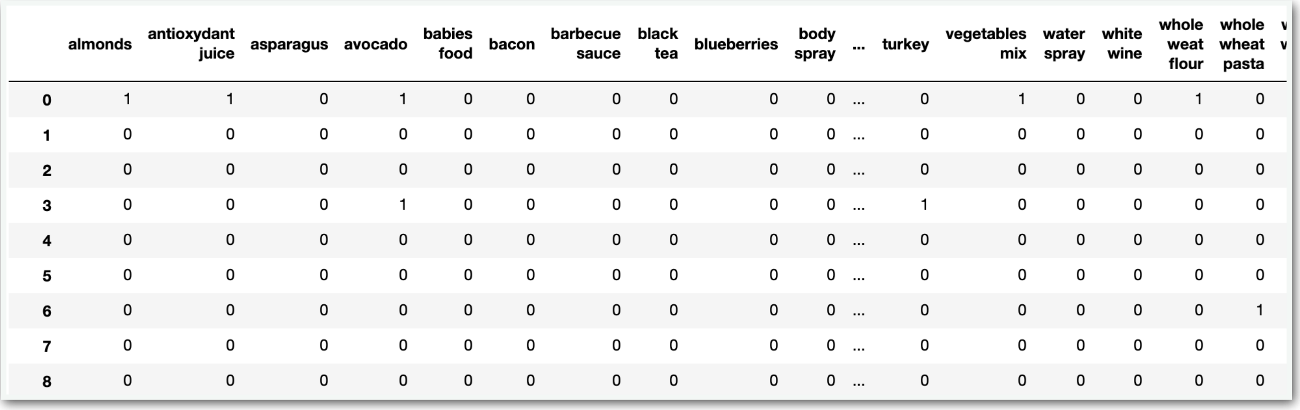
```
from mlxtend.preprocessing import TransactionEncoder

te = TransactionEncoder()
data = te.fit_transform(transactions)
data = pd.DataFrame(data, columns = te.columns_)
```

And the dataset will then be transformed as binary, with code :

```
data = data.astype("int"),
```

where 1 if the transaction occurs, 0 if not.

Now, we have the dataset looking like the figure below, and it’s now prepared.

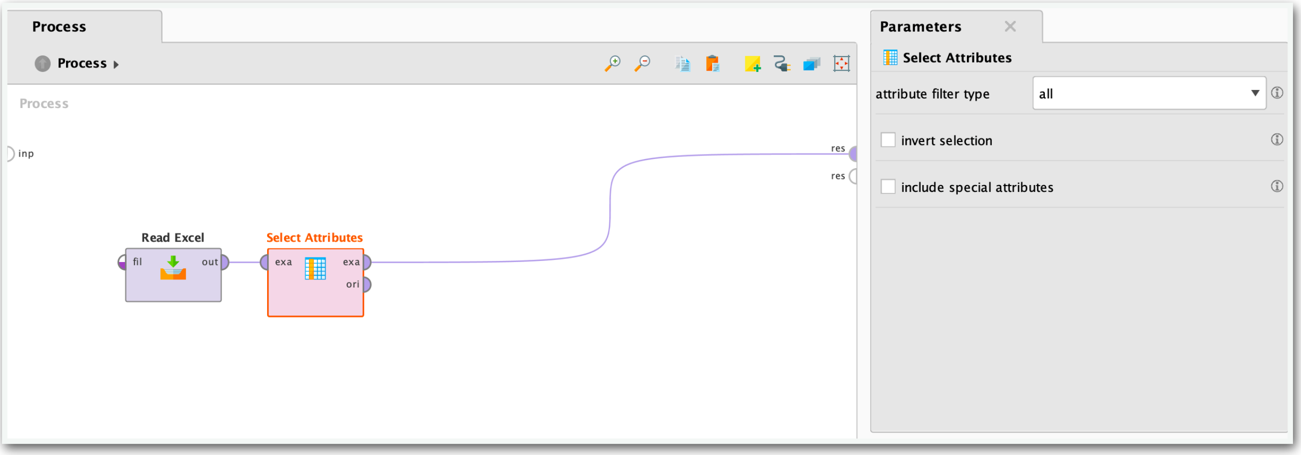


	almonds	antioxydant juice	asparagus	avocado	babies food	bacon	barbecue sauce	black tea	blueberries	body spray	...	turkey	vegetables mix	water spray	white wine	whole weat flour	whole wheat pasta
0	1	1	0	1	0	0	0	0	0	0	...	0	1	0	0	1	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
3	0	0	0	1	0	0	0	0	0	0	...	1	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	1
7	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0

Figure 2: Subset of Grocery Purchase Transactions in Binary Matrix Format

2.3 Market Basket Analysis by Using FP-Growth

For Market Basket Analysis by using FP-Growth algorithm, this analysis will be done by using Rapid Miner. First, all attributes will be chosen in order to find the association rules between the products.



Parameters

Select Attributes

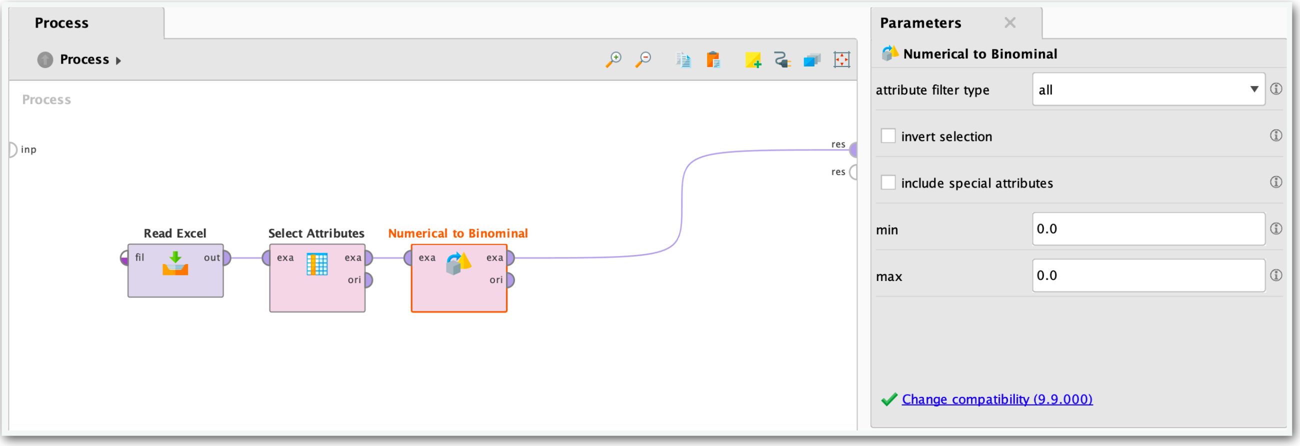
attribute filter type: all

□ invert selection

□ include special attributes

Figure 3: Parameters for Select Attributes operator

Next, the numerical to binomial operator will be chosen in order to make sure the attributes are in binomial as the next operator ‘FP-Growth’ only receives binomial input.



Parameters

Numerical to Binominal

attribute filter type: all

□ invert selection

□ include special attributes

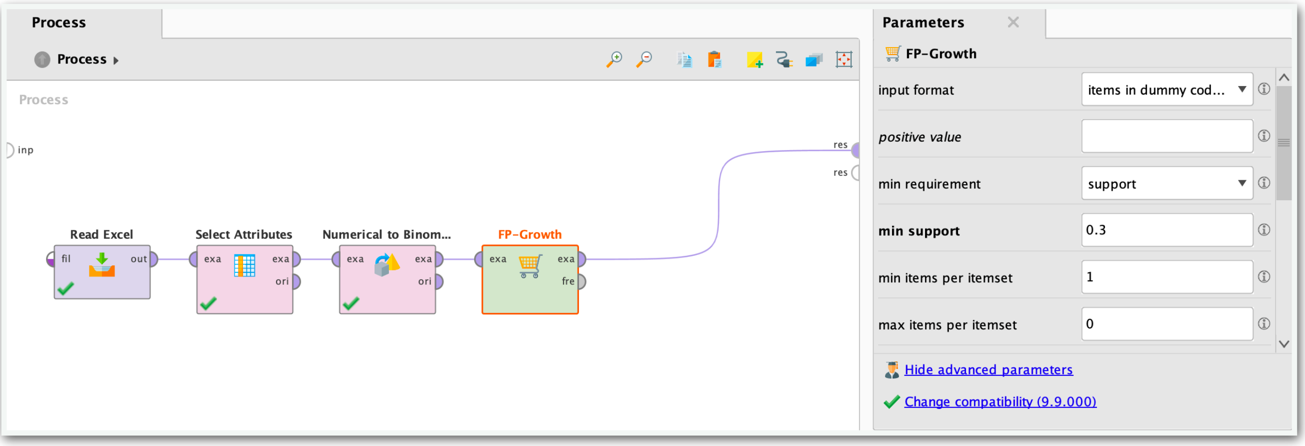
min: 0.0

max: 0.0

Change compatibility (9.9.000)

Figure 4: Parameters for Numerical to Binomial operator

After that, ‘FP-Growth’ operator will be selected. This operator will calculate the frequent item sets found in the data. Effectively, it goes through and identifies the frequency of all possible combinations of products that were purchased.



Parameters

FP-Growth

input format: items in dummy cod...

positive value:

min requirement: support

min support: 0.3

min items per itemset: 1

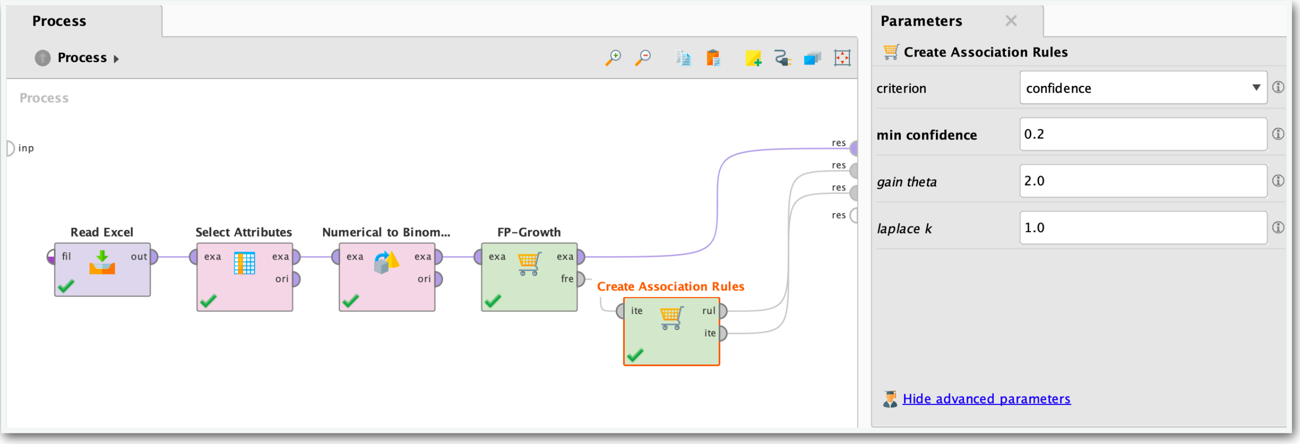
max items per itemset: 0

Hide advanced parameters

Change compatibility (9.9.000)

Figure 5: Parameters for FP-Growth operator

Finally, ‘Create Association Rules’ operator will be selected. This operator takes the product pairings that were frequently found by the FP-Growth operator and organises them into rules according to certain user-configurable parameters.



criterion	confidence
min confidence	0.2
gain theta	2.0
laplace k	1.0

Figure 6: Parameters for Create Association Rules operator

The association rules are shown in the table below. The support, confidence and lift for each of the association rules is also calculated.

Table 2: Association Rules sort by confidence

Premises	Conclusion	Support	Confidence	Lift
tomatoes	spaghetti	0.021	0.306	1.758
turkey	eggs	0.019	0.311	1.732
burgers	eggs	0.029	0.330	1.838
cake	mineral water	0.027	0.339	1.421
spaghetti	mineral water	0.060	0.343	1.439
whole wheat rice	mineral water	0.020	0.344	1.443
olive oil	spaghetti	0.023	0.348	2.000
pancakes	mineral water	0.034	0.355	1.489
tomatoes	mineral water	0.024	0.357	1.497
milk	mineral water	0.048	0.370	1.554
frozen vegetables	mineral water	0.036	0.375	1.572
chicken	mineral water	0.023	0.380	1.594
cooking oil	mineral water	0.020	0.394	1.654
ground beef	spaghetti	0.039	0.399	2.291
ground beef	mineral water	0.041	0.417	1.748
olive oil	mineral water	0.028	0.419	1.758
soup	mineral water	0.023	0.456	1.915

By definition, minimum support is the number of occurrences of an item set over the size of the example set. By decreasing the value of the minimum support, this will increase the number of item sets in the result. Suppose we want to find the association of items with a product that is sold at least 20 times per day. Thus, the minimum support would be 20 items per day multiplied by 7 days per week, and then divided by the total number of transactions. This left us with, (20*7)/7501 = 0.019. So the equivalent number 0.02 is then taken as the minimum support.

Confidence of an association rule indicates how likely Y is purchased when X is purchased (strength of a rule) – $P(Y|X)$. Confidence value should be close to 1 to say a rule holds enough confidence – dependency of an item on another item. In this assignment, we choose the minimum confidence as 0.3. The formula for confidence is shown as below :

$Confidence\ (X \rightarrow Y) = Support\ (X,Y)\ /\ Support\ (X)$

Lift is the ratio by which by the confidence of a rule exceeds the expected confidence. The basic rule of thumb is that the lift value is close to 1, it means that the rules were completely independent. If the lift values is greater than 1, it could be indicative of a useful rule pattern. The formula for lift is shown as below :

$Lift\ (X \rightarrow Y) = Support\ (X,Y)\ /\ (Support\ X \cdot Support\ Y)$

According to the results of the association rules, mineral water is the most popular consequent item. 12 out of 17 rules contain mineral water as the consequent item. From the above table, we can interpret the result of the first rule as the support is 0.021 calculated by dividing the number of transactions containing tomatoes and spaghetti by the total number of transactions. The confidence level shows that out of all the transactions that contain spaghetti 0.306 contain tomatoes too. The lift tells us that tomatoes is 1.758 times more likely to be bought by the customer who also buys spaghetti.

For the second rule, the support is 0.019 calculated by dividing the number of transactions containing turkey and eggs by the total number of transactions. The confidence level shows that out of all the transactions that contain eggs 0.311 contain turkey too. The lift tells us that turkey is 1.732 times more likely to be bought by the customer who also buys eggs.

A further discussion on the pattern of the result of this association rules will be discussed in Section 3.0.

2.4 Market Basket Analysis by Using Apriori

For Market Basket Analysis by using Apriori algorithm, this analysis will be done by using Jupyter Notebook. The minimum support chosen is 0.02. In total, there are 94 rules that are generated with 0.02 minimum support. The final step is to generate the rules with their corresponding support, confidence and lift. We can filter the data frame using standard pandas code. In this case, look for a lift greater than 1 and confidence greater than 0.3.

The code used is shown as below:

```
a_rules = apriori(basket_clean, min_support = 0.04,
use_colnames = True)

rules = association_rules(a_rules, metric =
'lift',  min_threshold = 1)

rules[(rules['lift'] >= 1) & (rules['confidence'] >= 0.3)]
```

The results of all association rules is shown in below table :

Table 3: Association Rules sort by confidence

antecedents	consequents	support	confidence	lift
(tomatoes)	(spaghetti)	0.021	0.306	1.758
(low fat yogurt)	(mineral water)	0.024	0.314	1.316
(frozen smoothie)	(mineral water)	0.020	0.320	1.342
(chocolate)	(mineral water)	0.053	0.321	1.348
(shrimp)	(mineral water)	0.024	0.330	1.385
(burgers)	(eggs)	0.029	0.330	1.838
(cake)	(mineral water)	0.027	0.339	1.421
(spaghetti)	(mineral water)	0.060	0.343	1.439
(whole wheat rice)	(mineral water)	0.020	0.344	1.443
(olive oil)	(spaghetti)	0.023	0.348	2.000
(pancakes)	(mineral water)	0.034	0.355	1.489
(tomatoes)	(mineral water)	0.024	0.357	1.497
(milk)	(mineral water)	0.048	0.370	1.554
(frozen vegetables)	(mineral water)	0.036	0.375	1.572
(chicken)	(mineral water)	0.023	0.380	1.594
(cooking oil)	(mineral water)	0.020	0.394	1.654
(ground beef)	(spaghetti)	0.039	0.399	2.291
(ground beef)	(mineral water)	0.041	0.417	1.748
(olive oil)	(mineral water)	0.028	0.419	1.758
(soup)	(mineral water)	0.023	0.456	1.915

From the above table, we see that mineral water is still the most popular consequent item. We can interpret the result of the first rule as the support is 0.021 calculated by dividing the number of transactions containing tomatoes and spaghetti by the total number of transactions. The confidence level shows that out of all the transactions that contain spaghetti 0.306 contain tomatoes too. The lift tells us that tomatoes is 1.758 times more likely to be bought by the customer who also buys spaghetti.

Next, we can interpret the result of the second rule as the support is 0.024 calculated by dividing the number of transactions containing low fat yogurt and mineral water by the total number of transactions. The confidence level shows that out of all the transactions that contain mineral water 0.314 contain low fat yogurt too. The lift tells us that low fat yogurt is 1.316 times more likely to be bought by the customer who also buys mineral water.

If we compare the result of association rules by using FP-Growth versus using Apriori, the association rules generated are similar. FP-Growth generated 17 association rules while Apriori generated 20 association rules.

3.0 Analysis of Market Basket Analysis

3.1 Association Graph

With the results that we get in section 2, the association graph with the best association rules that reflects the optimal correlation is shown as below:

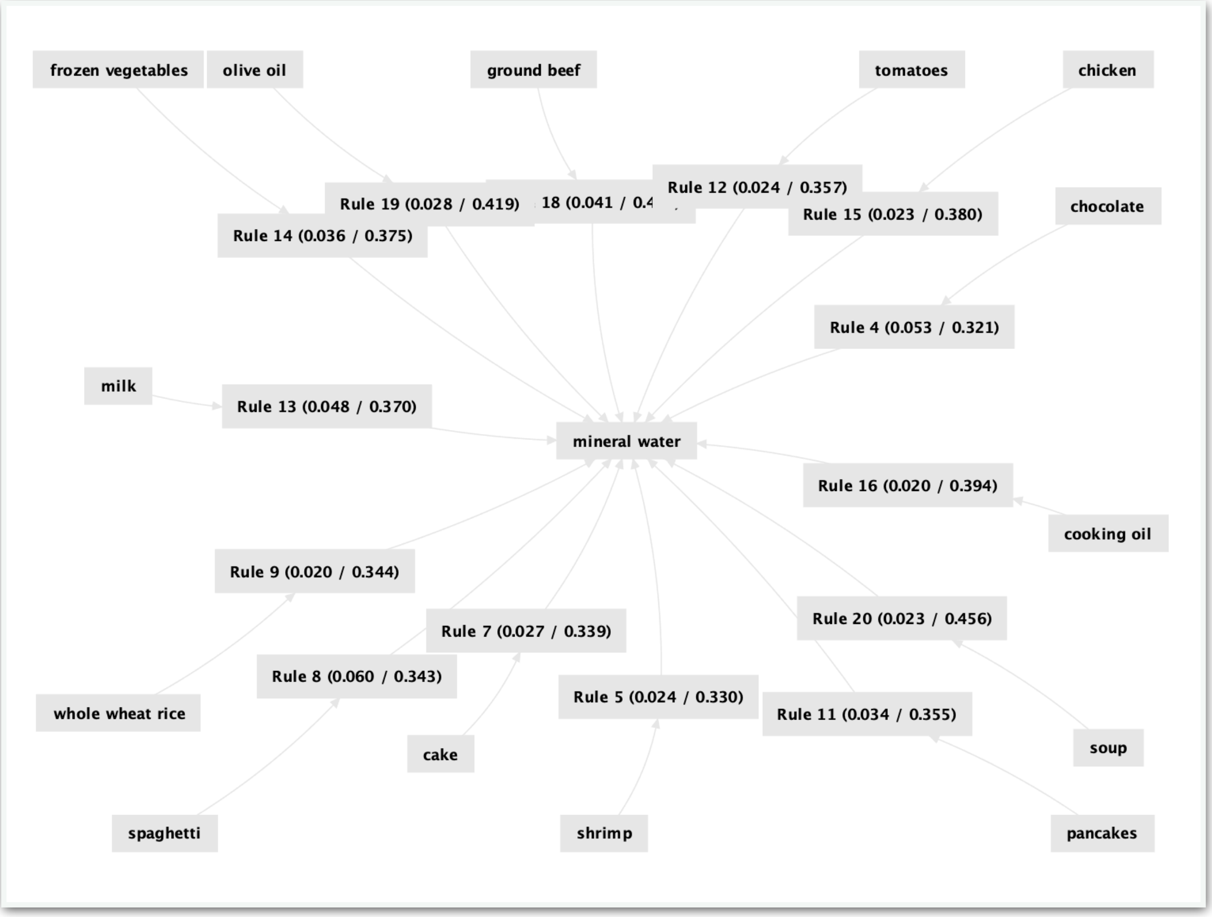


Figure 7: Association Graph

Referring to above figure, this graph is using FR Layout and Node Labels in RapidMiner, and it is filtered by showing only ‘mineral water’ as the conclusion. There are 12 antecedents that lead to mineral water as conclusion. We can see that mineral water is the most popular item, and the inventory and delivery of this product should always be put as priority as it generates sales in volume.

Other than that, there is also a classic pattern of ingredients of spaghetti can be found in the association rules. This can be shown in below figure.

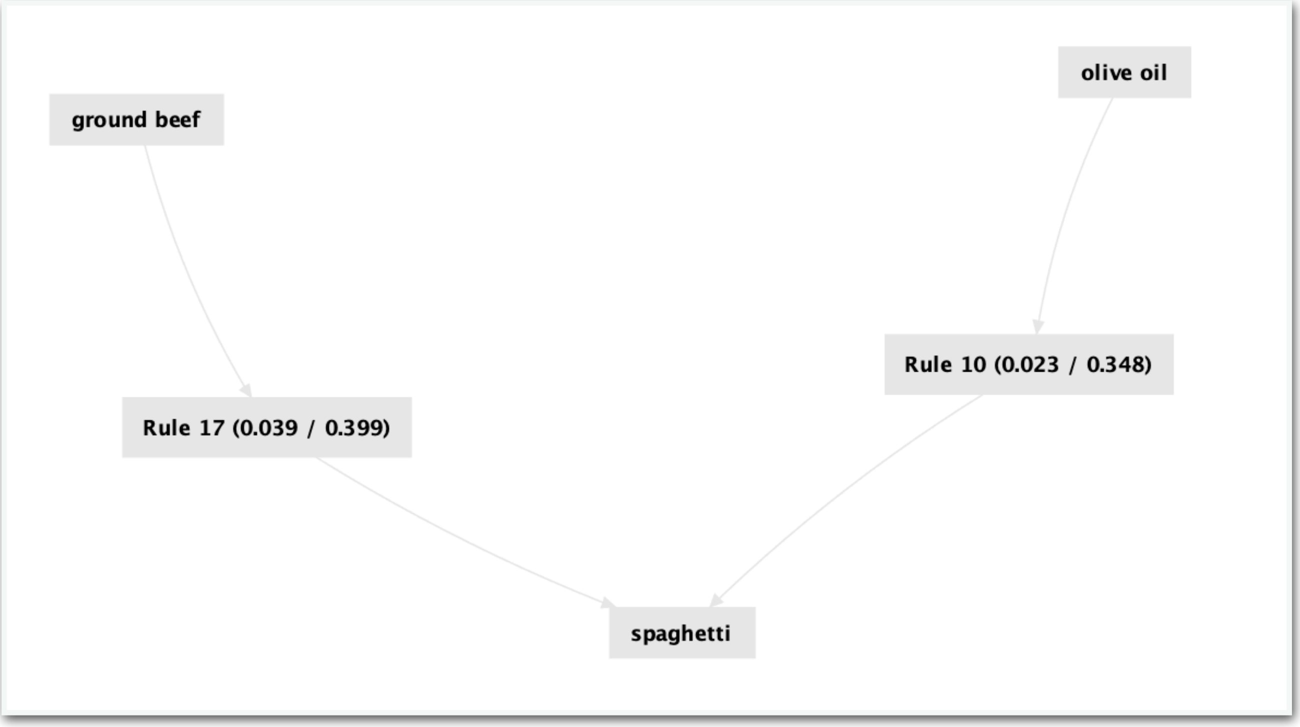


Figure 8 : Spaghetti as the conclusion

From the above figure we can see that customer that buy ground beef or olive oil will be likely to buy spaghetti. From this, retailer should put these three products together, and not to mention to make sure the mineral water is nearby too, as we know from Table 2, the lift tells us that spaghetti and ground beef is 1.439 and 1.749 times respectively more likely to be bought by the customer who also buys mineral water.

3.2 Analyse the pattern

If we refer to section 2, we can see the size of the basket for all the association rules is 2. If the retailer wanted to increase the size of the basket, the retailer can put some of the products together or do a promotion to bundle the products together as this will likely increase the possibility of customer buying the products together. For example, retailer can put spaghetti, ground beef, olive oil and mineral water together. Retailer could also do a promotion, for example by bundling olive oil and mineral water together, so that customer that come to buy ingredients for spaghetti will buy olive oil, mineral water and spaghetti, and this will increase the basket size.

The top 5 number of sales for this retailer is mineral water, eggs, spaghetti, french fries and chocolate. From this, it’s good if we can know the location of this grocery store, as the small size of basket and the products chosen look like the location of this grocery store could be in a neighbourhood that has a lot of singles. If there is a lot of single-person household in the neighbourhood, then it explains why the top 5 food are mineral water, eggs, spaghetti, french fries and chocolate. Due to the differences in food consumption patterns of live alone and multiple person households, single-person household grocery expenditures were found to be below multiple person households (Tariq, D'Souza & Allaway, 2016).

Another reason behind small-basket shoppers' shopping behaviour could be because of the lack of ability to plan ahead of time what is needed in a given time period. Or, shoppers might want to take advantages of price variations over time in a store, that is, buying more product categories when the prices are relatively low, and choose not to purchase when prices are high. However, for large-basket shopper, these shoppers usually are more likely to incur higher expenditure per trip and as a result, they will make fewer trips over a given time period compared to a typical small-basket shopper. In other words, the basket size, either small or large, will be highly correlated with two other consumer difference variables which is the amount of money spent on a grocery trip and also the frequency of grocery trips.

3.3 Visualisation

First, we will visualise the top 20 most frequent item in the dataset. This visualisation is done by using Python, with following code :

```
top_items = a_rules.sort_values('support', ascending = False)[:20]
for i in range(len(top_items.itemsets)):
    top_items.itemsets.iloc[i] = str(list(top_items.itemsets.iloc[i]))
fig = plt.figure(figsize = (10,10))
ax = fig.add_subplot(111)
ax.bar(top_items.itemsets, top_items.support)
for label in ax.xaxis.get_ticklabels():
    label.set_rotation(90)
plt.xlabel('Top 20 Most Frequent Item')
plt.ylabel(‘Support')
```

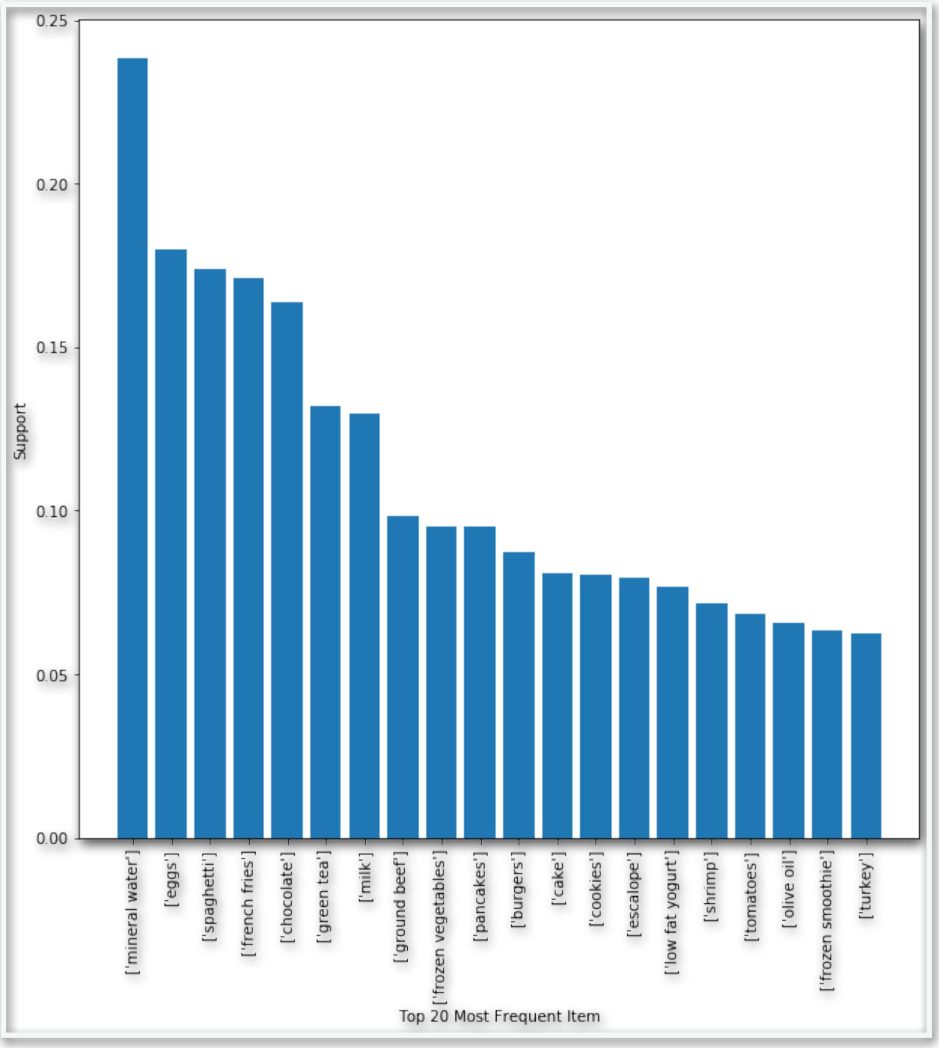


Figure 9 : Top 20 Most Frequent Item

Second visualisation that we will do is the visualisation of the association rules. By using the results from Apriori in Python, the results are visualise with following code :

```
rules['antecedents_'] = rules['antecedents'].apply(lambda a:
','.join(list(a)))
rules['consequents_'] = rules['consequents'].apply(lambda a:
','.join(list(a)))

pivot = rules[rules['lift']>1].pivot(index = 'antecedents_',
                    columns = 'consequents_', values= 'lift')

sns.heatmap(pivot, annot = True)
plt.yticks(rotation=0)
plt.xticks(rotation=90)
plt.show()
```

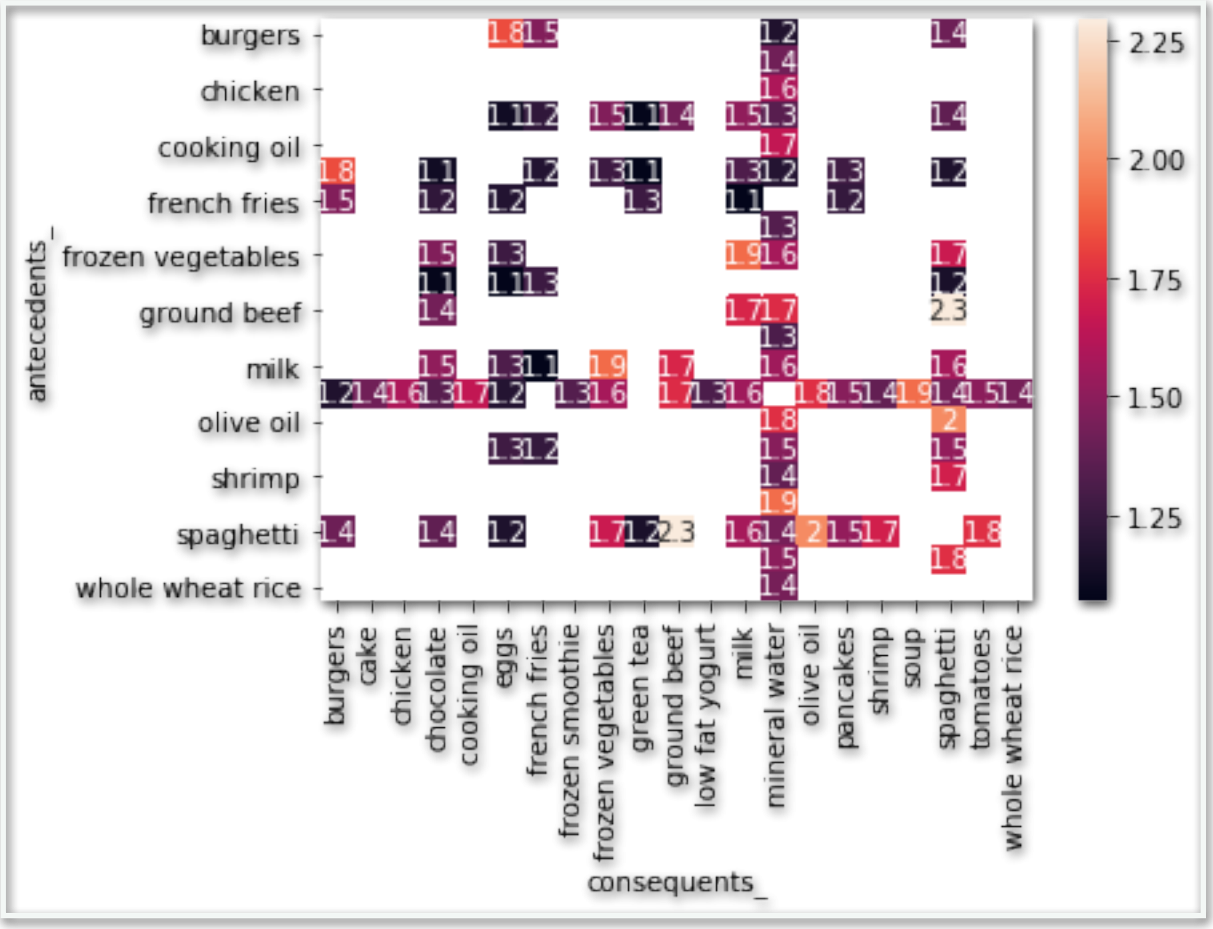


Figure 10: Market Basket Analysis Association Rules using Heatmap

In above figure, we show how we can visualise the Market Basket Analysis Association Rules by using Heat map. We will show all the rules, and the measurement is the lift.

4.0 Applications of Market Basket Analysis

4.1 Application of Market Basket Analysis in financial sector

Market Basket Analysis (MBA) can help financial sector for better targeting, and to increased sales and customer loyalty. Credit card history data is also a huge opportunity to use MBA. Bank frequently employs sales personnel at large malls to lure potential customers with attractive discounts on the go and they collaborate with retailers to show customers a multitude of offers they can get via purchasing through credit cards. All these customers' data can be used to discover different pairings of service offerings that are being used by customers. For instance, a customer that is using credit card has also turned a patron for the bank's home loan offer. There would be more of these basket-pairs in terms of service offerings being consumed by the similar customers. Bank can also used demographic data to know and to understand customers' banking behaviour which can help to narrow down the target market with the insights from MBA.

4.2 Application of Market Basket Analysis in healthcare sector

With the emergence of Electronic Medical Record (EMR), healthcare sector has a lot of data which could helped the doctors in order to no longer rely too much on their intuition while making critical clinical decisions. Market Basket Analysis applications in healthcare can be demonstrate such as analysing the workflow of diagnostic procedures, Up-selling and Cross-selling of Healthcare Systems, designing healthcare systems more user-friendly. MBA can be used to evaluate the effectiveness of medical treatments, by comparing the causes, symptoms, and courses of treatments. It can deliver an analysis of which courses of action prove effective. For example, the outcomes of patient groups treated with different drug regimens for the same disease or condition can be compared. The results then determine which treatments work best and are most cost-effective.

4.3 Application of Market Basket Analysis in education sector

One of the example of MBA in education sector is it can be used to do the analysis of an Undergraduate Curriculum. For instance, one can identify which courses that students fail simultaneously. The results then can be used to take at least two concrete actions, first is to suggest students to avoid problematic courses to be taken together at the enrolment process, and second is to invite degree coordinators to ponder about course distribution along the curriculum.

References

Kamakura, W. (2012). Sequential market basket analysis. Marketing Letters, 23(3), 505-516. doi: 10.1007/s11002-012-9181-6

Griva, A., Bardaki, C., Pramatari, K. and Papakiriakopoulos, D., 2018. Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems with Applications*, 100, pp.1-16.

Tariq, A., D’Souza, G., & Allaway, A. (2016). Grocery shopping, a one man job? Understanding the single shopper. Journal Of Consumer Marketing, 33(7), 574-584. doi: 10.1108/jcm-11-2015-1623

Market Basket Analysis for Banking: Better Targeting, Increased Sales and Customer Loyalty – Saksoft. (2020). Retrieved 26 May 2021, from https://www.saksoft.com/blog/market-basket-analysis-for-banking-better-targeting-increased-sales-and-customer-loyalty/