CHAPTER 6: PRACTICAL IMPLEMENTATION OF MARKET BASKET ANALYSIS ON THE DATASET

6.1. DATA COLLECTION

MBA was originally designed for the use with large datasets that are usually collected by others (i.e., not the scholars or the research team). Thus, the researchers can seek out partnerships with organizations that will provide data in exchange for conducting analysis and presenting results to those who provided the data. It is also possible for the research team to collect the data but given the time involved, effort and issues to access the data sources, most MBA studies are done on the data collected by other parties. In the chapter of literature review we were able to see that researchers have been able to create data-sharing partnerships in many different fields.

In addition to using data collected by organizations, researchers can implement a third-party data collection strategy consisting of reliance on publicly available information. For example, O'Boyle and Aguinis (2012) investigated issues regarding individual performance by using data on academics, entertainers, politicians, and amateur and professional athletes.

The data for this research was collected from a Belgian known super market chain and are related to customer transactions for the period from April 2018 to June 2018. Data is extracted from the information system of the company and come exclusively from customer transactions through loyalty cards. Every transaction is a record of the purchase carried out by the customer whoever used the loyalty card at the time of the purchase.

6.2. DATASET DESCRIPTION

Number of Attributes: 11

Time period of the data: 3 months (April, May, June)

Number of transactions: 78839

Number of customers: 42468

6.2.1. ATTRIBUTES

S. No	Attribute	Description

1	random_cust_no	This is a customer ID which was replaced by random number by the company itself.
2	orig_invoice_id	This is the most important attribute of the data. It can be called invoice ID or Transaction ID. The whole analysis is based on transactions. It is explained in detail under define the transactions.
3	date_of_day	This is the day when the transaction took place.
4	minute_desc	For more accuracy we do not just keep track of day of the transactions but also time of transaction. Since in case of a super market there are hundreds of transactions happening every day.
5	mikg_art_no	It is the code of the article / item.
6	art_name	Name of the article / item
7	catman_buy_domain_desc	The categories are in hierarchy level. This is the top level of the category under which the article falls. Category_1
8	pcg_main_cat_desc	This is the second level of category under which the article comes. Category_2
9	pcg_cat_desc	This is the third level of category under which the article comes. This is the important level of category since we used this category in our analysis. The association rules are based on this level of category. Category_3
10	pcg_sub_cat_desc	This is the last and final level of category under which the article comes. Category_4
11	quantity	This is the quantity of product bought by the customer. We used this attribute to change the transactional data to binary representation.

TABLE 2: ATTRIBUTES DESCRIPTION

6.3. TOOLS USED FOR ANALYSIS

- Jupyter notebook

- Python Libraries
    - Pandas
    - Datetime
    - Matplotlib
    - Mlxtend
    - Random
    - xlsxwriter
- MS Excel

6.4. DATA EXPLORING & TRANSFORMATION

Before doing anything with the data as soon as we read in the data in jupyter notebook, we start exploring it. Data exploring is the most important step for any analysis. While, exploring the data the issues that needed fixing were as follows:

1. The attribute *date_of_day* was read as string variable but in this research, time is a very crucial factor, so its data type was changed to date type.
2. For comparison of data of each month we needed to separate the data based on each month. Since date_of_day attribute was changed to date type. It was very easy to get the month of each transaction.
3. As explained earlier in the long tail effect the data sparsity problem tackled up to some extent by considering only the transactions in which total items purchased were more than 8 items. While exploring the data it's important to look at the data distribution. For market basket analysis it has always been observed to have the long tail effect. The long tail effect has assumed its present connotation and dynamics by Chris Anderson [21]. This term often refers to data products purchase in supermarkets describing their distribution as a long tail in which a small number of products is purchased more frequently whereas a large one is purchased less frequently. This phenomenon creates data sparsity problem and worsens even more their elaboration. For this research from the total number of transactions there were selected, those that included at least the purchase of 8 products. That is how, to some extent, it was tried to resolve the data sparsity problem and prevent the removal of any market basket product, which may present a risk of information loss. Figure below shows the curve of long tail effect.

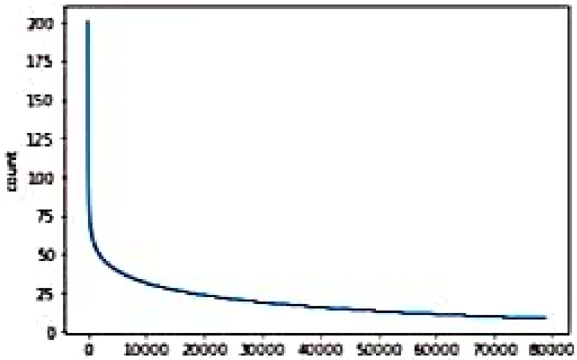


figure 10: long tail effect in our dataset

Some general exploring of data was done after all the fixes. In the table we can see the number of unique transactions, number of unique customers, number of unique articles sold, and number of articles sold in the period of 3 months as well as in each month (i.e. April, May, June).

Time	3 months	April	May	June
Number of Transactions	78839	24803	27710	26326
Number of Customers	42468	18743	20436	19347
Number of Unique Articles Sold	42694	31286	31035	31462
Number of Total Articles Sold	1541130	482776	544023	514331

TABLE 3: DATA EXPLORING

Since there are several categories and sub categories in the super market. We can easily identify the hierarchy of the categories based on the number of unique categories in the table below.

Category/Sub category	Total
catman_buy_domain_desc (Category_1)	87
pcg_main_cat_desc (Category_2)	348
pcg_cat_desc (Category_3)	1334
pcg_sub_cat_desc (Category_4)	3783

TABLE 4: CATEGORY AND SUBCATEGORIES DISTRIBUTION

6.5. BINARY REPRESENTATION

Binary data are used in information systems because data are categorized for attributes value by 0 and 1. It represents either the attribute category is present or absent in the data.

For super markets, the databases are very large and usually represented in binary format. The binary format is a matrix in which Transactions forms the rows and the items/ attributes forms the columns. For a specific transaction, if an item is purchased then the matrix position is made as 1. If the item is not purchased in the transaction, then the matrix position will be made 0. Binary data sets are interesting and useful due to its computational efficiency and minimal storage capacity.

6.6. ANALYSIS & DISCUSSION

6.6.1. Association rules from dataset

Since we have the data for three months, we divide the data into subset of each month to compare the analysis output of each month.

Firstly, we look at the comparison of number of rules when the minimum support count is less than 0.01 and minimum lift is 1.

Secondly, for all the data and for each of the subsets we have applied market basket analysis and recorded the results of the three indicators, i.e. confidence, support and lift. For every application of a market basket analysis process, the minimum level of confidence, support and lift established in the program, is 0.5, 0.01 and 1 respectively. The results of the number of rules that came out, are registered in table below.

Time	**Minimum support = 0.01** **Lift >= 1**	**Minimum support = 0.01** **Confidence >= 0.5** **Lift >= 1**
3 months	17282	1534
April	16564	1371
May	19380	1760
June	19272	1900

TABLE 5: ASSOCIATION RULES WITH DIFFERENT THRESHOLDS

In the first section we looked at the number of association rules with minimum support 0.01 and lift greater than 1. This criterion was applied to overall data as well as to monthly data. Our results were:

May > June > 3 months > April

We also calculated the number of association rules by adding degree of confidence greater than 0.5, the result changed to:

June > May > 3 months > April

6.6.2. Time development of association rules

The 128 rules, at the level of 1334 subcategories of products, resulting from the extraction process of overall 3 months data were used as a base for additional rules analysis at monthly level. For each of the 128 rules, there have been recorded the values for the three indicators, lift, support and confidence during each month (i.e. April, May, June). Thus, the table was created with the monthly development of association rules measurement indicators. The table can be found at the end of the document in ANNEX section in the end of the document.

In the end of the ANNEX we can notice that there are 2 association rules **{'COLA KZH'} => {'GROENTEN'}** and **{'GROENTEN'} => {'COLA KZH'}** which do not exist in the month of May, because they didn't pass the threshold of minimum support of 0.01 and lift greater than equal to 1. But these two rules only do not exist in the month of May They exist in the month of April and June.

Similarly, there are 2 association rules **{'GROENTEN'} => {'KOEKJES'}** and **{'KOEKJES'} => {'GROENTEN'}** which do not exist in the month of April and June. But they do exist in the month of May.

Whereas, the 128 association rules where chosen based on the threshold of minimum support as 0.05 and lift greater than and is equal to 1 from overall dataset. Whereas, to match with these 128 association rules the threshold set for monthly data was minimum support 0.01 and lift >= 1

Through these examples we can also say that there is a difference in the strength of the association rules in different time periods Its not just minor fluctuations in their strengths but sometimes it can be greater.

6.6.3. Top 30 association rules comparison

To show the fluctuation of degree of confidence of top 30 rule in each month. We have selected the rules with highest degree of confidence. In the table below, we can see the top 30 rules and their degree of confidence in 3 months all together as well as individually.

We can notice some major differences in rule number 4, 9, 18 and 19 in which we can see that those rules had lower degree of confidence in the month of April compared to the month of May and June. Also, in rule number 30 we can see that the degree of confidence for this rule is increasing with time.

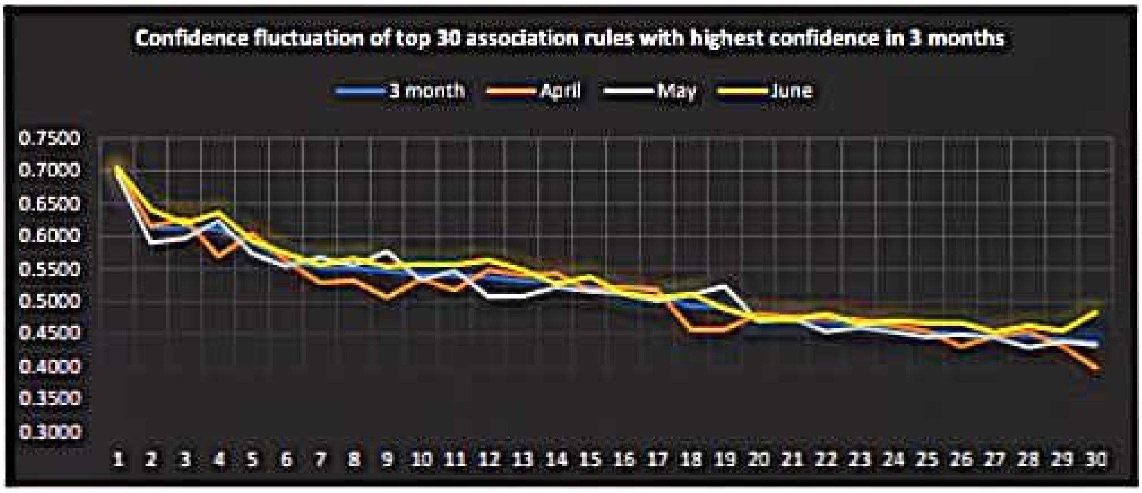


figure 11: Degree of confidence of top 30 rules

1	{'GROENTEN', 'CUCURBITACEAE'} => {'TOMATEN'}	16	{'GEBAK'} => {'BROOD'}
2	{'CUCURBITACEAE'} => {'TOMATEN'}	17	{'TOMATEN', 'BROOD'} => {'GROENTEN'}
3	{'CUCURBITACEAE', 'TOMATEN'} => {'GROENTEN'}	18	{'KRUIDEN + BLOEMEN'} => {'TOMATEN'}
4	{'PAPRIKA'} => {'TOMATEN'}	19	{'GROENTEN', 'TOMATEN'} => {'CUCURBITACEAE'}
5	{'SALADES', 'TOMATEN'} => {'GROENTEN'}	20	{'SCHARRELKIP'} => {'GROENTEN'}
6	{'SALADES', 'GROENTEN'} => {'TOMATEN'}	21	{'BANANEN'} => {'BROOD'}
7	{'BOLLEN'} => {'TOMATEN'}	22	{'STENGELGROENTEN'} => {'GROENTEN'}
8	{'KRUIDEN + BLOEMEN'} => {'GROENTEN'}	23	{'KIP'} => {'GROENTEN'}
9	{'PAPRIKA'} => {'CUCURBITACEAE'}	24	{'SALADES'} => {'GROENTEN'}
10	{'PAPRIKA'} => {'GROENTEN'}	25	{'EXOTISCH/TROPISCH'} => {'GROENTEN'}
11	{'WORTELGROENTE'} => {'TOMATEN'}	26	{'EXOTISCH/TROPISCH'} => {'TOMATEN'}
12	{'CUCURBITACEAE'} => {'GROENTEN'}	27	{'SALADES'} => {'BROOD'}
13	{'BOLLEN'} => {'GROENTEN'}	28	{'EUROPA'} => {'GROENTEN'}
14	{'TOMATEN'} => {'GROENTEN'}	29	{'SALADES'} => {'TOMATEN'}
15	{'WORTELGROENTE'} => {'GROENTEN'}	30	{'EXOTISCH/TROPISCH'} => {'BESSEN'}

TABLE 6: 30 ASSOCIATION RULES WITH HIGHEST DEGREE OF CONFIDENCE

Degree of support of top 30 rule in each month. In figure 12 and table 7 we can see the top 30 associations rules with highest support in 3 months all together as well as individually.

When we look at the figure, we can notice that there is a difference in the support level of most of the association rules. But for most of them the support is increasing with the increase in time. Most of the association rules had lower support in the month of April but it eventually increased in the month of June.

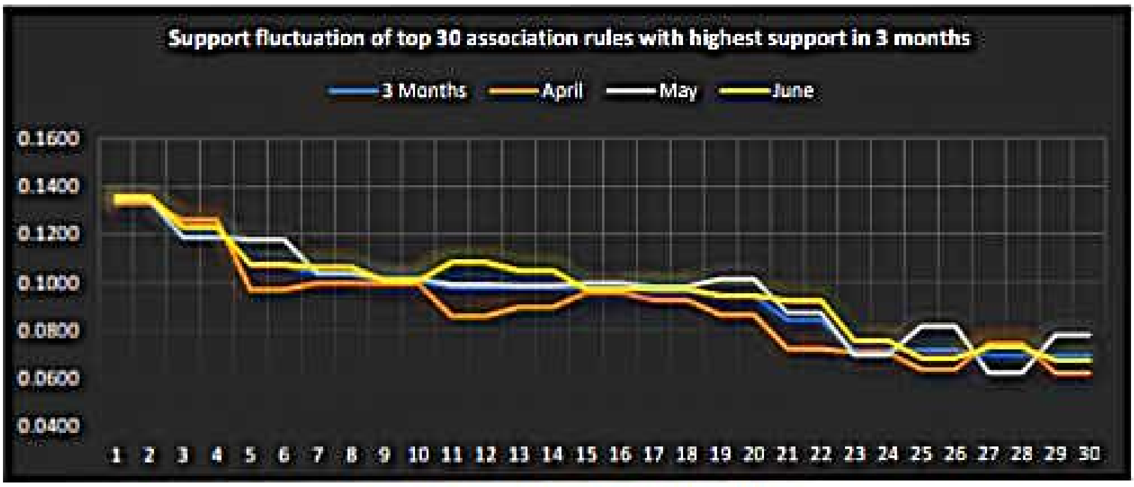


figure 12: Degree of support of top 30 rules

1	{'GROENTEN'} => {'TOMATEN'}	16	{'SALADES'} => {'BROOD'}
2	{'TOMATEN'} => {'GROENTEN'}	17	{'SALADES'} => {'TOMATEN'}
3	{'GROENTEN'} => {'BROOD'}	18	{'TOMATEN'} => {'SALADES'}
4	{'BROOD'} => {'GROENTEN'}	19	{'GROENTEN'} => {'CUCURBITACEAE'}
5	{'CUCURBITACEAE'} => {'TOMATEN'}	20	{'CUCURBITACEAE'} => {'GROENTEN'}
6	{'TOMATEN'} => {'CUCURBITACEAE'}	21	{'TOMATEN'} => {'BESSEN'}
7	{'TOMATEN'} => {'BROOD'}	22	{'BESSEN'} => {'TOMATEN'}
8	{'BROOD'} => {'TOMATEN'}	23	{'KIP'} => {'GROENTEN'}
9	{'SALADES'} => {'GROENTEN'}	24	{'GROENTEN'} => {'KIP'}
10	{'GROENTEN'} => {'SALADES'}	25	{'CUCURBITACEAE'} => {'BROOD'}
11	{'BESSEN'} => {'BROOD'}	26	{'BROOD'} => {'CUCURBITACEAE'}
12	{'BROOD'} => {'BESSEN'}	27	{'GEBAK'} => {'BROOD'}
13	{'BESSEN'} => {'GROENTEN'}	28	{'BROOD'} => {'GEBAK'}
14	{'GROENTEN'} => {'BESSEN'}	29	{'CUCURBITACEAE'} => {'SALADES'}
15	{'BROOD'} => {'SALADES'}	30	{'SALADES'} => {'CUCURBITACEAE'}

TABLE 7: 30 ASSOCIATION RULES WITH HIGHEST DEGREE OF SUPPORT

In the table 8 we can see what the top 30 association rules with the highest Lift are and in figure 13 we can see what the lift values of those 30 association rules in 3 months are all together and each month periodically.

In the comparison of lift we can notice that there is a clear difference in the lift of association rules number 1 and 2 in all the 3 months. But as the lift gets smaller and smaller the difference in different time periods decreases. Rule 5 and 6 also had higher lift in the month of April but in the month of May and June it dropped.

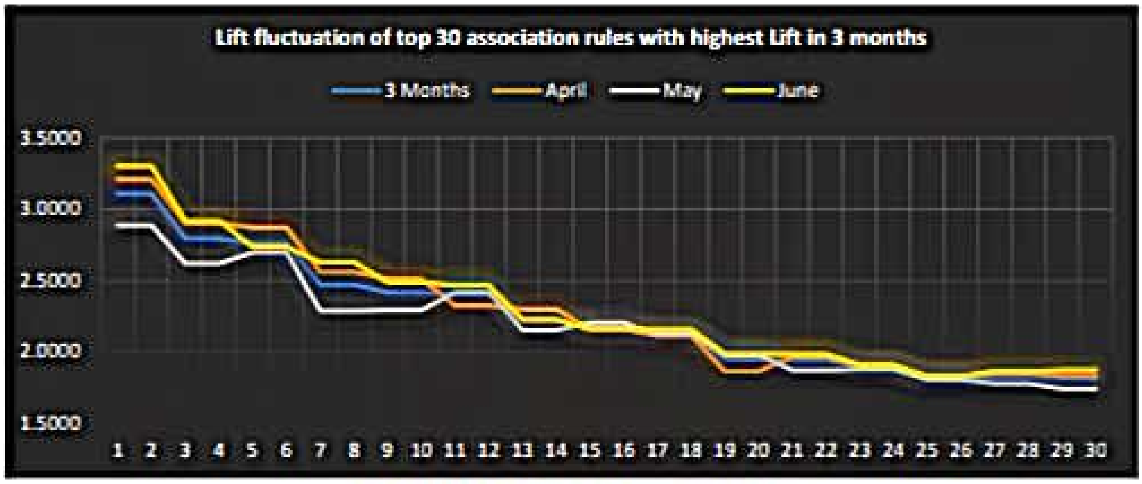
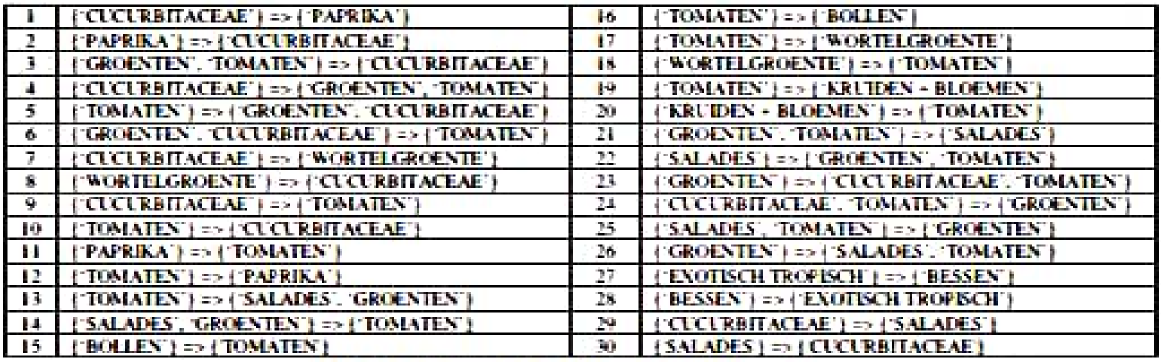


figure 13: Lift of top 30 association rules



1	{'CUCURBITACEAE'} => {'PAPRIKA'}	16	{'TOMATEN'} => {'BOLLEN'}
2	{'PAPRIKA'} => {'CUCURBITACEAE'}	17	{'TOMATEN'} => {'WORTELGROENTE'}
3	{'GROENTEN', 'TOMATEN'} => {'CUCURBITACEAE'}	18	{'WORTELGROENTE'} => {'TOMATEN'}
4	{'CUCURBITACEAE'} => {'GROENTEN', 'TOMATEN'}	19	{'TOMATEN'} => {'KRUIDEN + BLOEMEN'}
5	{'TOMATEN'} => {'GROENTEN', 'CUCURBITACEAE'}	20	{'KRUIDEN + BLOEMEN'} => {'TOMATEN'}
6	{'GROENTEN', 'CUCURBITACEAE'} => {'TOMATEN'}	21	{'GROENTEN', 'TOMATEN'} => {'SALADES'}
7	{'CUCURBITACEAE'} => {'WORTELGROENTE'}	22	{'SALADES'} => {'GROENTEN', 'TOMATEN'}
8	{'WORTELGROENTE'} => {'CUCURBITACEAE'}	23	{'GROENTEN'} => {'CUCURBITACEAE', 'TOMATEN'}
9	{'CUCURBITACEAE'} => {'TOMATEN'}	24	{'CUCURBITACEAE', 'TOMATEN'} => {'GROENTEN'}
10	{'TOMATEN'} => {'CUCURBITACEAE'}	25	{'SALADES', 'TOMATEN'} => {'GROENTEN'}
11	{'PAPRIKA'} => {'TOMATEN'}	26	{'GROENTEN'} => {'SALADES', 'TOMATEN'}
12	{'TOMATEN'} => {'PAPRIKA'}	27	{'EXOTISCH TROPISCH'} => {'BESSEN'}
13	{'TOMATEN'} => {'SALADES', 'GROENTEN'}	28	{'BESSEN'} => {'EXOTISCH TROPISCH'}
14	{'SALADES', 'GROENTEN'} => {'TOMATEN'}	29	{'CUCURBITACEAE'} => {'SALADES'}
15	{'BOLLEN'} => {'TOMATEN'}	30	{SALADES} => {CUCURBITACEAE}

TABLE 8: 30 ASSOCIATION RULES WITH HIGHEST LIFT