Marketing Analytics

Palash Jain

2/9/2019

Executive Summary

This report contains the business analytics performed on the sales data from 2013-01-02 to 2018-10-31. Three separate analysis were performed in order to understand various facets of the sales data. These were:

- Market Segmentation - This was performed in order to gain a better understanding of the customers. The segmentation is based on the infamous RFM (Recency, Frequency, Monetary Value) model and helps us to identify and target customer segments with appropriate products and marketing campaigns.
- Customer Churn analysis - Modelling customer churn in a non-contractual setting such as retail is a difficult problem as every transaction could a customer's last transaction. Here, an anomaly detection method is used to label customers spending pattern as normal or anamolous. This helps us identify customers who are most likely to churn and take actions to retain them.
- Market Basket Analysis - This analysis is based on a theory that if you buy a certain set of items, you are more or less likely to buy another set of items. Knowledge of such rules can really help drive the revenue up. Here the apriori algorithm is used to extract such rules from the sales data.

The Dataset

The dataset consists of 234,086 rows of sales data with 37 columns. Each row contains details about the purchase of an item by a customer. The starting date of this data is 2013-01-02 while the last date is 2018-10-31.

```
##  [1] "Invoice Code"                  "Continent"
##  [3] "State"                         "Country name EN"
##  [5] "Customer code"                 "Customer group code"
##  [7] "Customer salesman"             "Invoice Code__1"
##  [9] "DDMMYYYY"                      "YYYYMM"
## [11] "YYYY"                          "YYYYWW"
## [13] "Month (Long)"                  "Month (short)"
## [15] "Sales company code"            "Salesman code"
## [17] "Salesman name"                 "Market code"
## [19] "Market name"                   "Master item code"
## [21] "Master item name"              "Item code"
## [23] "Item name"                     "Format code"
## [25] "Format name"                   "Range code"
## [27] "Range name"                    "Packing code"
## [29] "Packing name EN"               "PCB"
## [31] "Qty"                           "Total Units (Qty*PCB)"
## [33] "Volume (Ltr)"                  "Total Volume (Ltr)"
## [35] "Total Sales Amount (€)"        "Currency"
## [37] "Total Sales Amount (Currency)"
```

Market Segmentation

Market segmentation allows us to group customers into segments based on purchasing behaviour, demographics amongst other things. This allows us to ensure that appropriate marketing campaigns are targeted to the relevant segment of customers. Market segmentation also helps us in identifying our best customers.

The first step in RFM based market segmentation is to generate an event log of transactions. This basically contains 3 columns of data: a customer ID, date of transaction and the total amount spent in that transaction (Table 1.). From the event log, a RFM matrix is generated (Table 2.) with the following definitions.

- Recency - The number of days passed between a customer's last transaction and the end of the period under observation (in this case '2018-11-01')
- Frequency - The number of transactions (invoices) the customer has had in the sales period under observation.
- Monetary Value - The average money (in euros) spent by a customer per transaction during the sales period under observation.

cust_id	invoice_code	date	amount	Days_since_Purchase
00000-A01	SIN-B10-1603-00005	2016-03-10	53.33	966
00000-A01	SIN-B10-1603-00057	2016-03-31	1032.20	945
00000-A01	SIN-B10-1603-00061	2016-03-31	110.90	945
00000-A01	SIN-B10-1603-00066	2016-03-31	1150.00	945
00000-A01	SIN-B10-1603-00067	2016-03-31	16.38	945
00000-A01	SIN-B10-1604-00006	2016-04-08	46.84	937

Table 1. The transactional event log created from the dataset.

cust_id	Recency	Frequency	First_Purchase	Monetary_Value
00000-A01	1	26	1568	647.30
00000-D02	366	14	945	769.39
00000-P07	1	305	763	1248.42
00001-D99	335	135	518	5421.79
00002-A01	9	208	2123	14971.22
00004-01	639	1	639	619.40

Table 2. The RFM matrix created from the transactional event log.

Once this matrix is calculated, two different types of market segmentations were used. First is the traditional RFM based market segmentation which scores every customers based on the RFM matrix and then categorizes them into segments based on that score. The second is a managerial segmentation performed on the basis of the absolute values in the RFM matrix. The rules used for each type of segmentation can be found in the appendix section at the end of this report.

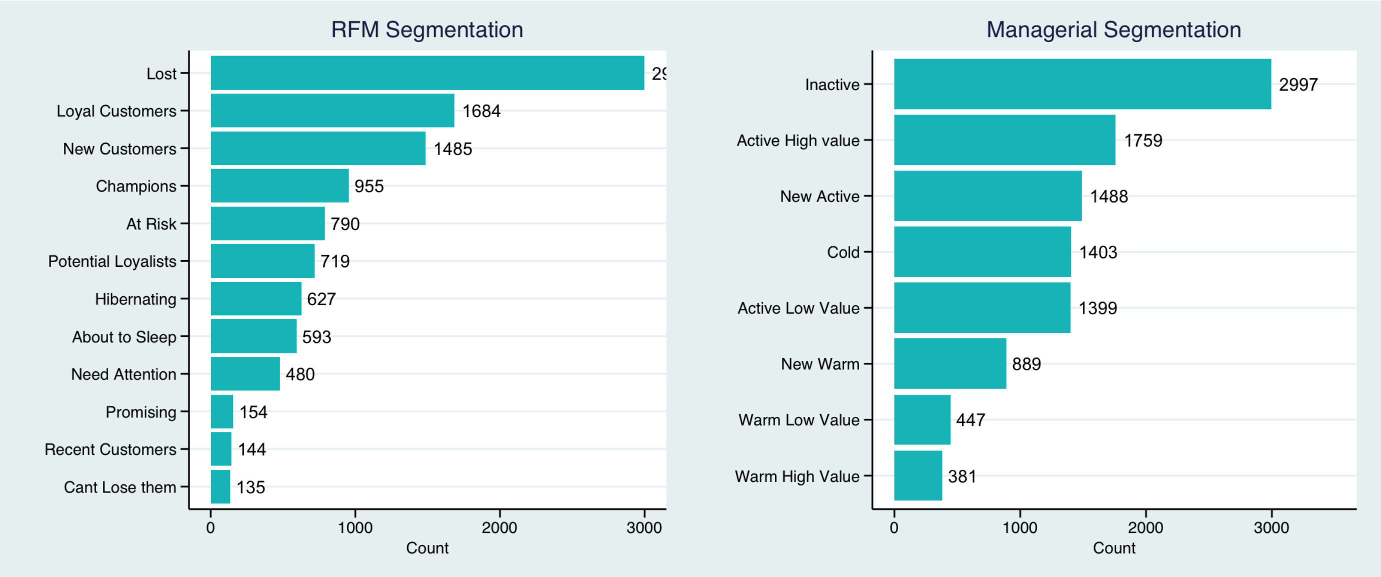
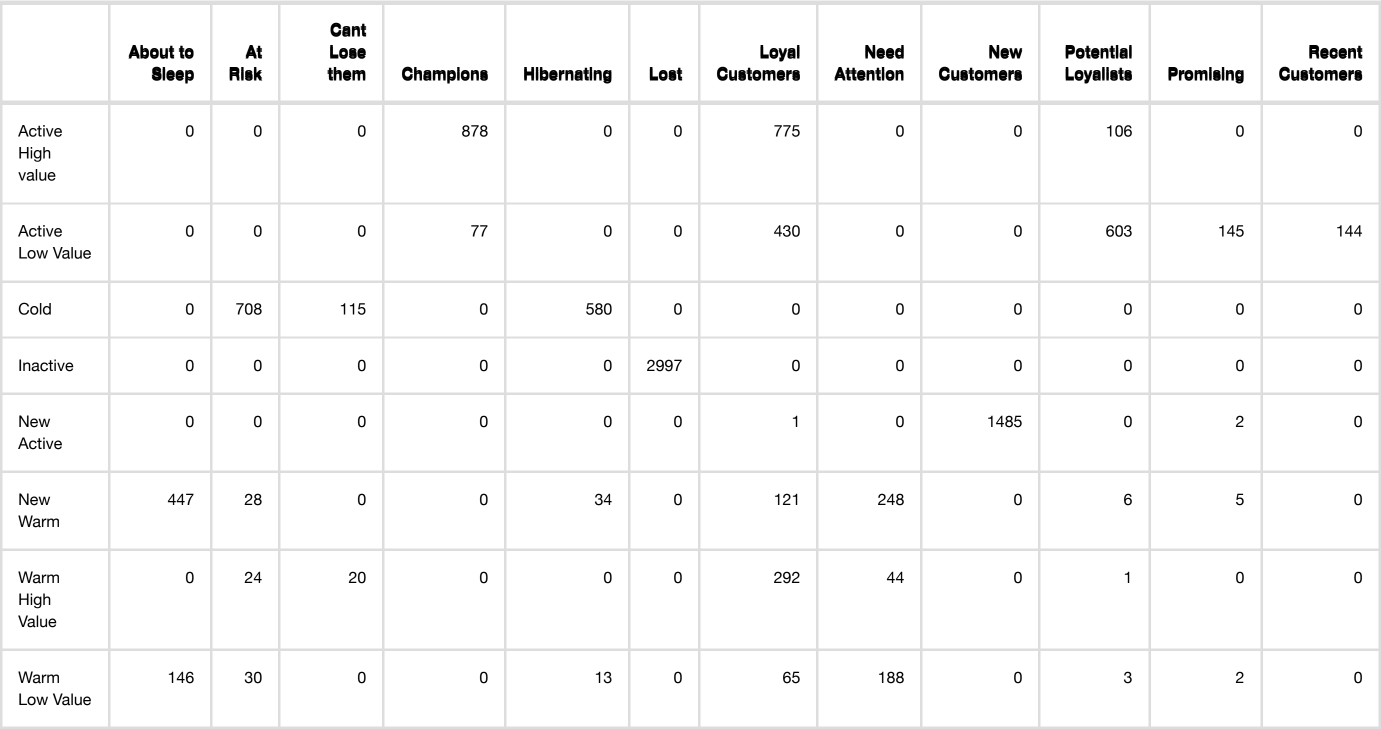


Figure 1. The market distribution according to the two segmentation strategies used.

As seen in Fig 1. A majority of our customers are lost or inactive. These are customers who have not interacted with us for over 3 years now. There is not much profit to be had in trying to appease these customers via campaigns. Having said that we do have a significant amount of loyal and active customers. These customers can be the most responsive to marketing campaigns and new products. The major cause for concern here are the “Can’t lose them” customers, as these were the customers who were both high value and frequent, but we have not heard from them for a while. We need to identify the reason for their lack of interaction with us and win them back. A full breakdown of the segments and relevant strategies is given below.

Customer Segment	**Activity**	**Actionable Tip**
Champions	Bought recently, buy often and spend the most!	Reward them. Can be early adopters for new products. Will promote your brand.
Loyal Customers	Spend good money with us often. Responsive to promotions.	Upsell higher value products. Ask for reviews. Engage them.
Potential Loyalist	Recent customers, but spent a good amount and bought more than once.	Offer membership / loyalty program, recommend other products.
Recent Customers	Bought most recently, but not often.	Provide on-boarding support, give them early success, start building relationship.
Promising	Recent shoppers, but haven’t spent much.	Create brand awareness, offer free trials
Customers Needing Attention	Above average recency, frequency and monetary values. May not have bought very recently though.	Make limited time offers, Recommend based on past purchases. Reactivate them.
About To Sleep	Below average recency, frequency and monetary values. Will lose them if not reactivated.	Share valuable resources, recommend popular products / renewals at discount, reconnect with them.
At Risk	Spent big money and purchased often. But long time ago. Need to bring them back!	Send personalized emails to reconnect, offer renewals, provide helpful resources.
Can’t Lose Them	Made biggest purchases, and often. But haven’t returned for a long time.	Win them back via renewals or newer products, don’t lose them to competition, talk to them.
Hibernating	Last purchase was long back, low spenders and low number of orders.	Offer other relevant products and special discounts. Recreate brand value.
Lost	Lowest recency, frequency and monetary scores.	Revive interest with reach out campaign, ignore otherwise.



	About to Sleep	At Risk	Cant Lose them	Champions	Hibernating	Lost	Loyal Customers	Need Attention	New Customers	Potential Loyalists	Promising	Recent Customers
Active High value	0	0	0	878	0	0	775	0	0	106	0	0
Active Low Value	0	0	0	77	0	0	430	0	0	603	145	144
Cold	0	708	115	0	580	0	0	0	0	0	0	0
Inactive	0	0	0	0	0	2997	0	0	0	0	0	0
New Active	0	0	0	0	0	0	1	0	1485	0	2	0
New Warm	447	28	0	0	34	0	121	248	0	6	5	0
Warm High Value	0	24	20	0	0	0	292	44	0	1	0	0
Warm Low Value	146	30	0	0	13	0	65	188	0	3	2	0

Table 3. Cross-tabulation of the two segmentation strategies (values represent counts).

The two types of segmentations mostly agree with each other(Table 3 & appendix).

Customer Churn Analysis

Churn modelling in non-contractual business is not a classification problem, it is an anomaly detection problem. Anomaly detection is a technique used to identify unusual patterns that do not conform to expected behaviour, called outliers. We want to be able to make claims like “9 times out of 10, Customer X will make his next purchase within Y days”. If Customer X does not make another purchase within Y days, we know that there is only a 1 in 10 chance of this happening, and that this behaviour is anomalous. Using the anomaly threshold obtained from the analysis, our customers’ purchasing behaviour can be classified as being normal or anomalous.

To be able to model this accurately the analysis is limited to only those businesses which have had at least 10 transactions with us.

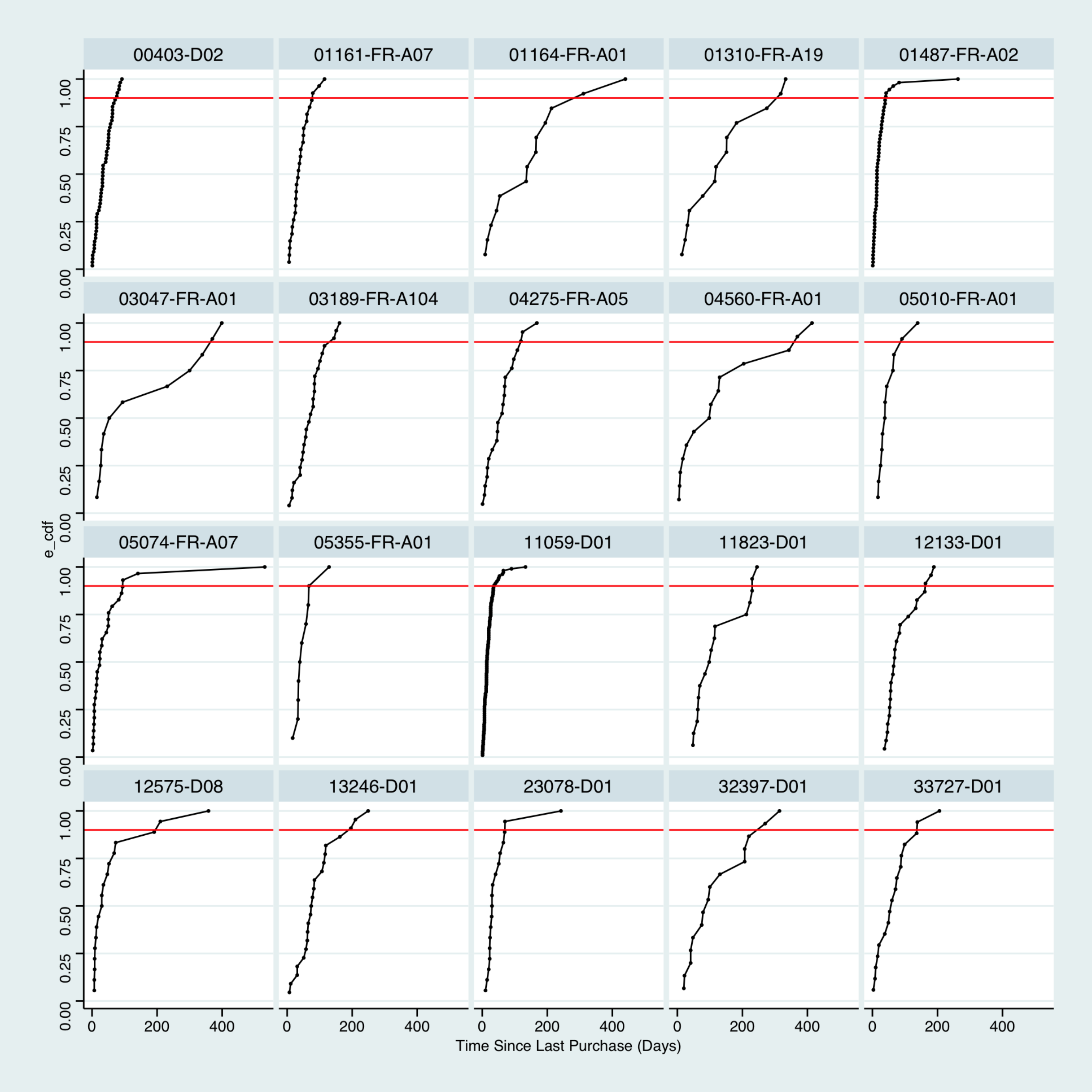


Figure 2. The probability of a business purchasing from us plotted agaisnt the number of days passed since their last purchase. The red line marks a 90% probability. The intersection point’s X-coordinate tell us the number of days in which that business is expected to make a transaction 9 out of 10 times.

cust_id	Recency	rfm_score	rfm_segment	managerial_segment	anomaly_threshold	status

cust_id	Recency	rfm_score	rfm_segment	managerial_segment	anomaly_threshold	status
00000-A01	1	54	Champions	Active High value	282.2	Normal
00000-D02	366	33	Loyal Customers	Warm High Value	63	Anomalous
00000-P07	1	54	Champions	Active High value	15.4	Normal
00001-D99	335	35	Loyal Customers	Active High value	Not Enough Transactions	Not Enough Transactions
00002-A01	9	55	Champions	Active High value	28	Normal
00004-01	639	22	Need Attention	New Warm	Not Enough Transactions	Not Enough Transactions

Table 4. Customers table with segmentation and anomaly detection incorporated.

As seen in Table 4, customer ‘00000-D02’ makes a purchase every 63 days (9 out of 10 times). This customer’s recency value of 366 means that this customer has not made a purchase in over a year. Our analysis suggests that their is only a 1 in 10 chance of this happening.

Market Basket Analysis

Market Basket Analysis is a modeling technique based upon the theory that if you buy a certain set of items, you are more or less likely to buy another set of items. It is an essential technique used to discover association rules that can help increase the revenue of a company.

transaction	items_bought
00000-A01 on 2014-07-17	EVASOL 20L P
00000-A01 on 2016-03-10	L043 - FLAT 24W P45 18x38
00000-A01 on 2016-03-31	SA/FORTE C86 24W*18x38 FP1, PROTOTYPE CORK, UNIVERSAL INNER BAG-FLAT PACK, OUTER BAG - PRINTED- ZIPPED, LABEL- SOCOMORE- 70X103MM- BLU, PROTOTYPE CORK, SOCOCLEAN AQUAFORTE, SA/FORTE 15233 300W 30X52 BK1, VARIOUS COSTS, SA/FORTE 15100 25W*21x21 FP1, PROTOTYPE CORK
00000-A01 on 2016-04-08	OUTER BAG - WHITE- SOCOMORE, UNIVERSAL INNER BAG-FLAT PACK

Table 5. The transaction data shows what item were bought together.

A transaction dataset is generated so that all items that are bought together in one invoice are in one row. Since in our dataset there are multiple invoices generated for the same customer on the same date, I group all items bought by one customer on a single day into one transaction.

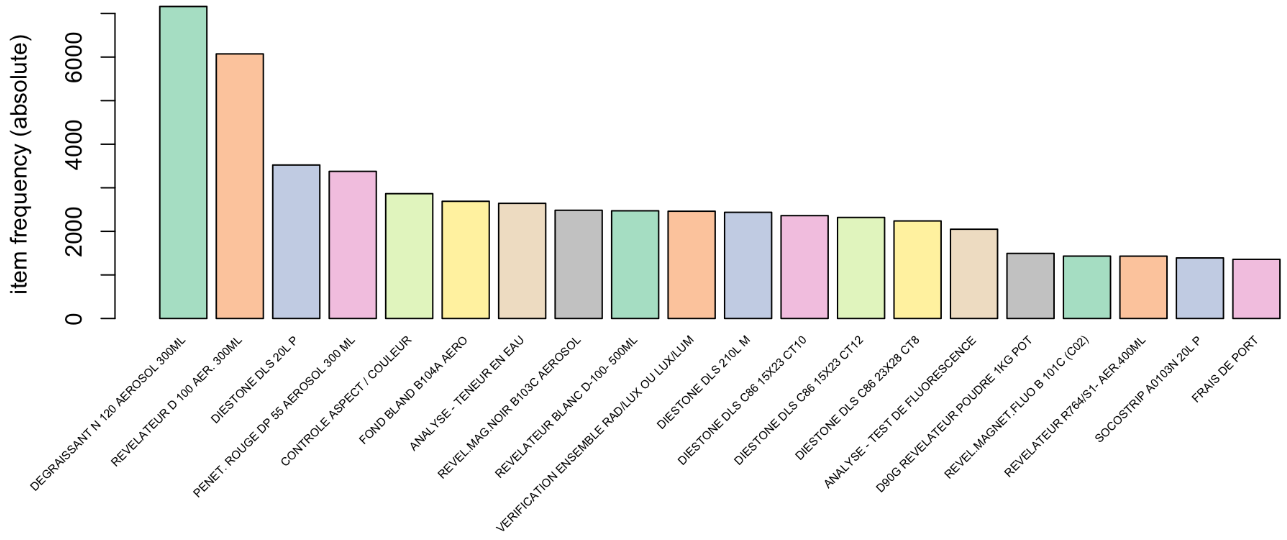


Figure 3. The top 20 most frequently bought items according to our sales data.

As seen in Fig 3. The most frequently bought item in our sales dataset is “Degraissant N 120 Aerosol 300 mL” with over 8000 purchases. Some other frequently bought items or services are “Diestone DLS 20L P”,“Analyse - Teneur EN EAU”,“Socostrip A0103N” and “Frais De Port”.

Following this, the apriori algorithm is used to form association rules as described above. The terminology of the results follows the statement :

“If item/service(s) X(s) is/are bought there is a Y % probability that item/service(s) Y(s) will be bought in the same invoice.”

- LHS - item/service(s) X(s)

- RHS - item/service(s) Y(s)
- Support - How many times does this association occur in the dataset (expressed as a proportion).
- Confidence - Probability of this rule holding (expressed as a proportion)
- Lift - How likely item Y is purchased when item X is purchased, while controlling for how popular item Y.
- Count - How many times does this association occur in the dataset (expressed as an absolute).

	LHS	RHS	support	confidence	lift	count
1	{DIMENSIONNEMENT DES INDICATIONS - PHOTO CLIENT}	{TIRAGE CLICHE CLIENT}	0.0013323	0.8270270	304.40426	153
4	{PENETRANT BASSE TEMP. LTP 82 AEROSOL}	{REVELAT.BASSE TEMP. D106-AEROSOL}	0.0011582	0.9109589	323.87832	133
11	{REPARATION METALLOSCOPE COMPRENANT :}	{VERIFICATION METALLOSCOPE/AIMANT PERMANENT}	0.0019506	0.8115942	112.83518	224
16	{PENET.ROUGE PR 25 AEROSOL 500ML}	{REVELATEUR R60 AEROSOL 500ML}	0.0013062	0.5434783	126.85357	150
17	{WADIS 24- AEROSOL 400ML}	{VERIFICATION METALLOSCOPE/AIMANT PERMANENT}	0.0024208	0.7020202	97.60121	278
18	{ANALYSE - CHLORE/FLUOR/SOUFRE ASME}	{ANALYSE - SENSIBILITE PENETRANT}	0.0010624	0.6559140	106.99410	122

Table 6. The association rules show which items are frequently bought together.

As seen in Table 6. If the item/service “Wadis 24- Aerosol 400 mL” is purchased by a customer, there is a 70 % (confidence column) that they will also purchase the item/service “Verification Metalloscope/Aimant Permanent”. This combination occurs in approximately 300 transactions in the dataset.

Appendix

Rules for RFM based segmentation

- Champions : Recency score and Frequency + Monetary Value Score in the range [4-5].
- Loyal Customers : Unsegmented customers with Recency score and Frequency + Monetary Value Score in the range [2-5] & [4-5] respectively.
- Potential Loyalists : Unsegmented customers with Recency score and Frequency + Monetary Value Score in the range [3-5] & [2-3] respectively.
- Recent Customers : Unsegmented customers with Recency score and Frequency + Monetary Value Score in the range [4-5] & [0-2] respectively.
- Promising : Unsegmented customers with Recency score and Frequency + Monetary Value Score in the range [3-4] & [0-2] respectively.
- Need Attention : Unsegmented customers with Recency score and Frequency + Monetary Value Score in the range [2-3] & [2-3] respectively.
- About to sleep : Unsegmented customers with Recency score and Frequency + Monetary Value Score in the range [2-3] & [0-2] respectively.
- At Risk : Unsegmented customers with Recency score and Frequency + Monetary Value Score in the range [0-2] & [2-3] respectively.
- Can’t Lose Them : Unsegmented customers with Recency score and Frequency + Monetary Value Score in the range [0-1] & [4-5] respectively.
- Hibernating : Unsegmented customers with Recency score and Frequency + Monetary Value Score in the range [0-2] & [0-2] respectively.

Rules for Managerial Segmentation

- Inactive : Customers whose last transaction with us was more than 3 years ago.
- Cold : Unsegmented customers whose last transaction with us was more than 2 years ago.
- Warm : Unsegmented customers whose last transaction with us was more than 1 year ago.
- Active : Customers whose last transaction with us was less than 1 year ago.
- Active High Value : Active customers whose average sales value per transaction is more than 500 Euros.
- Active Low Value : Active customers whose average sales value per transaction is less than 500 Euros.
- New Active : Active customers whose first purchase with us was less than 1 years ago.
- Warm High Value : Warm customers whose average sales value per transaction is more than 500 Euros.
- Warm Low Value : Warm customers whose average sales value per transaction is less than 500 Euros.
- New Warm : Warm customers whose first purchase with us was less than 2 years ago.