PURCHASING BEHAVIOR IN A B2B ONLINE RETAILER

Capstone Project

Abstract

The main purpose of this project is applying unsupervised machine learning algorithms such as clustering and association rules in order to provide great insight on a business transactional database from a product-centric and customer-centric approaches. This can result in better planning and more actionable strategies that could be reflected in higher revenue.

Diana Moyano

Dec 2018

Table of Contents

INTRODUCTION

The prediction of customer behaviour, operational analytics and supply chain analysis are some of the methods used in this particular industry, and it is usually referred to customers as individuals or end users. However, how do all this vary when **dealing with wholesalers**? B2B transactions are also a big part of the retail sector, hence understanding what **patterns** they follow and how they can be **predicted** has become pivotal in the creation of revenue.

Are there any purchasing patterns in these online retailer's UK customers based on their transactions?

The question above can be solved from different perspectives. This database will be analyzed taking a **product-centric approach** through the use of **association rules and the Apriori algorithm** in order to understand what products are usually bought together, allowing the business to offer special discounts and promotions that can increase sales in the future. Another way to analyze it is through a **customer-centric approach**, as this can provide information on patterns or special needs these customers may have; **clustering** may be the strategy to consider in order to get those insights relevant for the business.

DATASET

https://archive.ics.uci.edu/ml/datasets/Online+Retail#

The **Online Retail** dataset (available since 2015) provides information about all the transactions an online UK company has had between 2010 and 2011. It sells unique all-occasion gifts to mostly wholesalers.

Its attributes are:

- Invoice number
- Stock Code
- Description of the product
- Quantity
- Invoice Date
- Unit Price
- Customer ID
- Country

AT A GLANCE...

There are 541,909 observations, representing a transaction of a particular stock code. Some relevant aspects noticed during a brief data exploration will be explained below, and further analysis will be required in order to determine how relevant this information is for the scope of this project:

ADDITIONAL CONSIDERATIONS BASED ON THE BUSINESS PROBLEM

- A Customer-driven marketing approach that aims to understand customer behaviors in order to generate effective offers and promotions that are relevant to their needs. This can also lead to establishing loyalty programs that can be reflected and more steady future revenue.

- Product recommendation is an analytical process that are based on correlations with what other customers who bought the same product are also buying another one (this is called collaborative filtering). It is important to keep in mind that this method is being widely used in the industry (especially by Amazon), so it's relevant to determine will be differentiated from the rest.

LITERATURE REVIEW

Discovering Association Rules in Transaction Databases[1]

- Association rules has 2 parts
  - Antecedent→ item found in the data
  - Consequent→ item found in combination with the consequent
- An association rule has 2 numbers that provide information about how uncertain the rule is
  - Support→ number of transactions that have both the antecedent and the consequent. In other words, it gives an idea of the probability of finding this combination in the whole dataset
  - Confidence→ Ratio of the number of transactions of the consequent and the antecedent to the number of transactions only including the antecedent.
- The main purpose of the Apriori algorithm is to generate frequent itemsets starting with 1 item, then with 2, 3 and so on until it has generated itemsets for all sizes
- A good way to measure the strength of an association rule is through its benchmark confidence or lift. If this one is greater than 1, the association rule can be worth considering.

Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining[2]

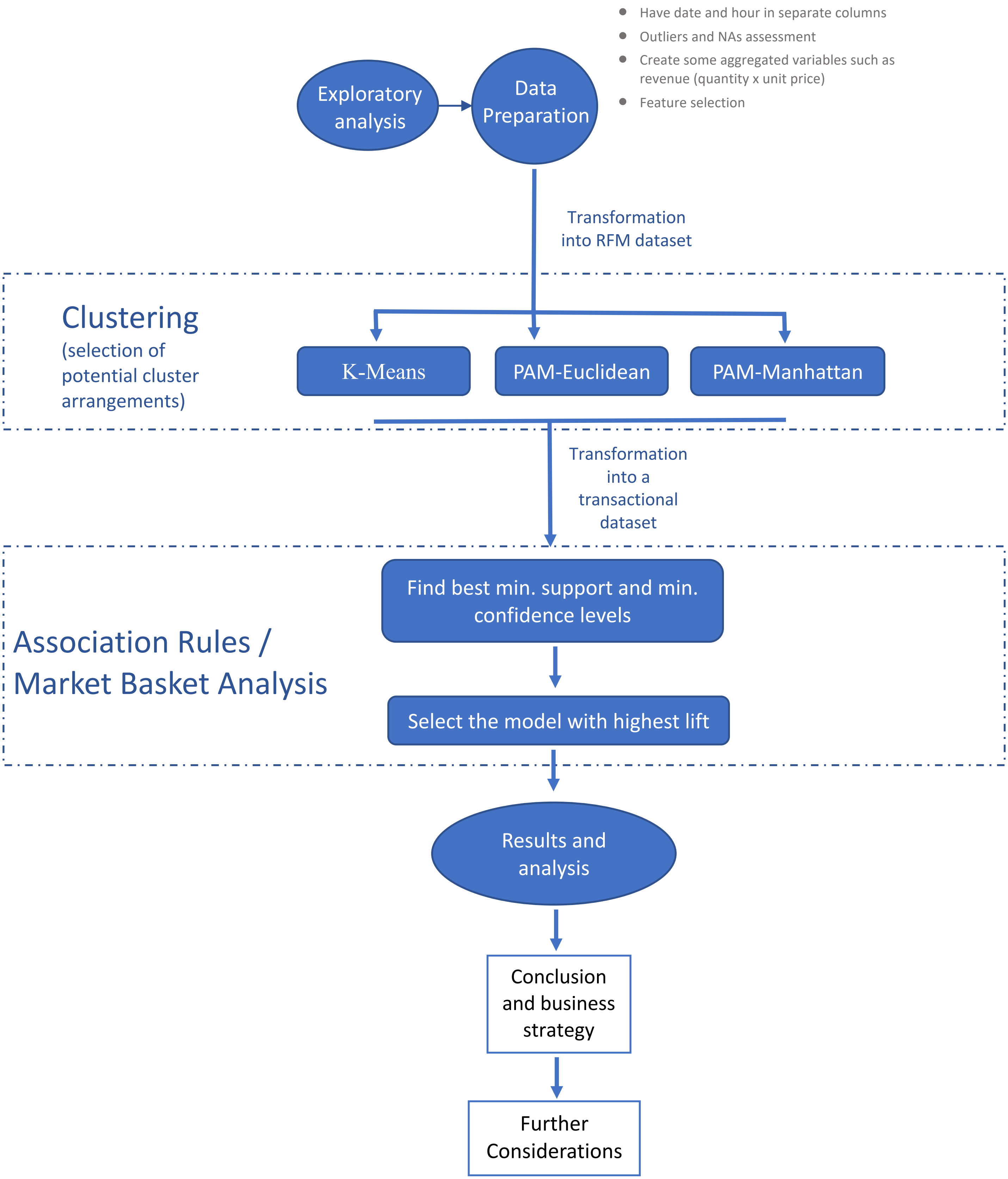
- In this article, researches look into an online retail dataset and performed clustering analysis by using the RFM model by using SAP. This one provides information about the recency, frequency and monetary per customer.
  - Recency – How recently did the customer purchase? (dataset's latest transaction date - customer's latest transaction date)
  - Frequency – How often do they purchase? (count of all unique invoices per customer)
  - Monetary Value – How much do they spend (each time on average)? (sum or revenue divide by the customer frequency)
- Based on this information, they were able to identify what groups of customers result more profitable for the company
- This dataset has information about the location where this transaction occurred (Zip code) which allowed the researchers find relevant insights on this regard.
- They have also used a decision tree algorithm in order to enhance their clustering analysis, as one of the clusters was very diverse. Nested segments were designed, and this group was segmented in sub-categories.

---

[1] Source: https://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/lecture-notes/Lecture_16.pdf

[2] Chen, Daqing. Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. 18th July, 2012

METHODOLOGY



APPROACH IN DETAIL

1. Data wrangling
2. Transformation into a RFM dataset
3. Clustering (selection of potential cluster arrangements)
4. Each cluster group is transformed into a transactional dataset
5. The Apriori algorithm (association rules) is applied to each group
6. Selection of the best cluster arrangement based on lift
7. Results and Analysis
8. Conclusion
9. Further Considerations

DATA WRANGLING

FILE: 01_DATAWRANGLING.RMD

Special focus was put in this section due to the type of machine learning method used for this project. When using unsupervised learning, there are no labels assigned to the observations (unlike supervised learning such as classification), which suggest a more exploratory standpoint in order to find potential patterns that are not entirely evident to the analyst. In order to find these insights, the dataset must be as clean as possible in order to avoid all the noise that could distort the information needed solve the business problem.

This dataset consists of approximately 542,000 transactions with 8 variables. Below is a small sample of how it looks like

InvoiceNo <fctr>	**Stock Code** <fctr>	**Description** <fctr>	**Quantity** <int>	**InvoiceDate** <fctr>	**UnitPrice** <dbl>	**Customer ID** <int>	**Country** <fctr>
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850	United Kingdom

The transactions coming from a country different than the UK were deleted (out of project's scope), resulting in 495,478 observations.

A sapply function was used to find NAs, all of them (133,600 obs.) related to the Customer ID variable. Since the business problem has a customer-centric focus, this data would not provide any relevant insight when proceeding with clustering. In addition to this, some transactions were not related to the sale of products (e.g. bank charges, payment to Amazon, etc.) and some others referred to damaged or lost stock. All these observations were deleted.

Two new variables were created:
- Revenue: the result of the unit price x quantity
- Date_Order: it was extracted from the InvoiceDate, as we will not focus on the hour of the transaction.

Two columns were deleted:
- Country
- InvoiceDate

InvoiceNo <fctr>	StockCode <fctr>	Description <fctr>	Quantity <int>	UnitPrice <dbl>	CustomerID <int>	Date_Order <date>	Revenue <dbl>
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2.55	17850	2010-12-01	15.30
536365	71053	WHITE METAL LANTERN	6	3.39	17850	2010-12-01	20.34
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2.75	17850	2010-12-01	22.00
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	3.39	17850	2010-12-01	20.34
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	3.39	17850	2010-12-01	20.34
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	7.65	17850	2010-12-01	15.30

Negative Values and Cancelled Transactions

The distribution of the quantity, unit price and revenue were analyzed the summary function

```
    Quantity            UnitPrice             Revenue
 Min.   :-80995.00   Min.   :    0.00   Min.   :-168469.60
 1st Qu.:     2.00   1st Qu.:    1.25   1st Qu.:      3.75
 Median :     4.00   Median :    1.95   Median :     10.20
 Mean   :    11.08   Mean   :    3.26   Mean   :     18.70
 3rd Qu.:    12.00   3rd Qu.:    3.75   3rd Qu.:     17.70
 Max.   : 80995.00   Max.   :38970.00   Max.   : 168469.60
```

We can observe that the unit price only has positive values, whereas the quantity and sales contain negative values. Since the unit price doesn't have negative values, we can infer that all those negative values come from the quantity (given that revenue is a calculated value from these two variables).

2% of the dataset is related to negative quantities. After doing some tests on this particular group, it was possible to conclude that all transactions with negative values were assigned with a C.

InvoiceNo <chr>	StockCode <fctr>	Description <fctr>	Quantity <int>	UnitPrice <dbl>	CustomerID <int>	Date_Order <date>	Revenue <dbl>	Cancelled <chr>
536379	D	Discount	-1	27.50	14527	2010-12-01	-27.50	C
536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	4.65	15311	2010-12-01	-4.65	C
536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	1.65	17548	2010-12-01	-19.80	C
536391	21984	PACK OF 12 PINK PAISLEY TISSUES	-24	0.29	17548	2010-12-01	-6.96	C
536391	21983	PACK OF 12 BLUE PAISLEY TISSUES	-24	0.29	17548	2010-12-01	-6.96	C
536391	21980	PACK OF 12 RED RETROSPOT TISSUES	-24	0.29	17548	2010-12-01	-6.96	C

At first, it was assumed that the C stands for a cancellation. However, some other type of transactions assigned with this letter were also related to discounts, postage, a manual entry or a commission (representing 0.08% of the whole dataset). These ones were deleted, as they were not are able to provide sufficient insight based on the project's scope.

The remaining observations assigned with a C were referred as cancellations of previous transactions. As they may not provide any relevant insights when applying a clustering algorithm (they were cancelled out, hence the customer did not want them at the end), these were deleted along with the initial transaction in order not to affect the whole distribution.

A column with the absolute value of the sales was created, and then duplicates could be found based on 3 variables:

- Absolute value in sales

- Invoice number
- Description

Final arrangements were made (removal of some outliers), resulting in a dataset of 344,094 observations that will be used for clustering analysis and association rules.

InvoiceNo <chr>	Description <fctr>	Quantity <int>	UnitPrice <dbl>	CustomerID <int>	Date_Order <date>	Revenue <dbl>
536365	WHITE HANGING HEART T-LIGHT HOLDER	6	2.55	17850	2010-12-01	15.30
536365	WHITE METAL LANTERN	6	3.39	17850	2010-12-01	20.34
536365	CREAM CUPID HEARTS COAT HANGER	8	2.75	17850	2010-12-01	22.00
536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	3.39	17850	2010-12-01	20.34
536365	RED WOOLLY HOTTIE WHITE HEART.	6	3.39	17850	2010-12-01	20.34
536365	SET 7 BABUSHKA NESTING BOXES	2	7.65	17850	2010-12-01	15.30

CLUSTERING

FILE: 02_CLUSTERING.RMD

Data Preparation

In order to have a customer-centric approach via clustering, the dataset will require an arrangement for RFM analysis. As previously explained in the literature review, this method is usually used for clustering analysis when we information about the customerID, date and monetary value are available in every transaction.

Recency – How recently did the customer purchase? (dataset's latest transaction date - customer's latest transaction date)

Frequency – How often do they purchase? (count of all unique invoices per customer)

Monetary Value – How much do they spend (each time on average)? (sum or revenue divide by the customer's frequency)

The data was also scaled as some of the clustering methods use Euclidean distance. Monetary, for instance, has way higher values that would affect the recency and frequency.

Finally, each row name was assigned with its corresponding customer ID.

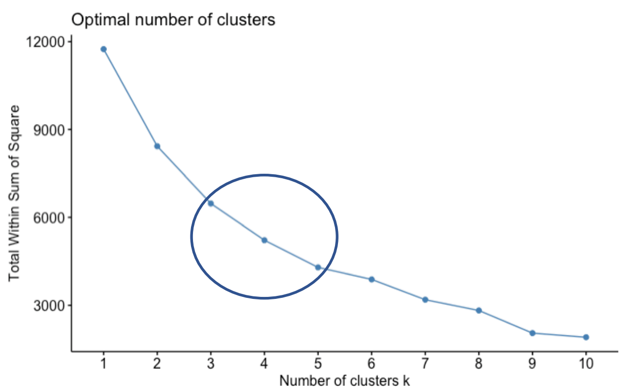
Below is a sample of the final dataset.

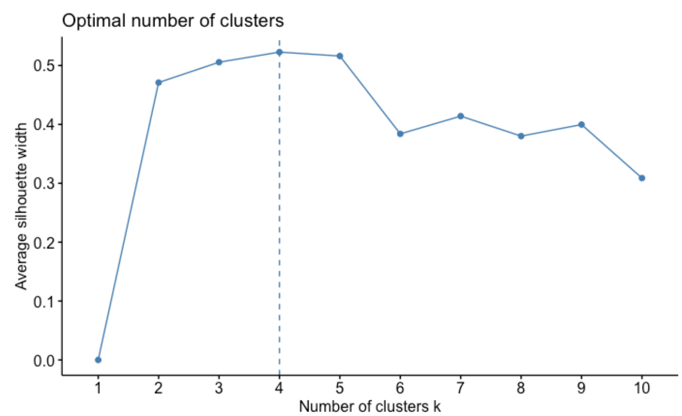
```
        Recency Frequency_1    Monetary
12747 -0.9016500  0.94572457  0.09201299
12748 -0.9217179 28.19860946 -0.49755032
12749 -0.8916161  0.10717427  1.18232442
12820 -0.8916161 -0.03258412 -0.28069915
12821  1.2255503 -0.45185927 -0.64574600
12822 -0.2193405 -0.31210089  0.32958601
```

Clustering Methods

In order to determine the best clustering arrangement and the optimal number of clusters, different methods were applied and results were compared.

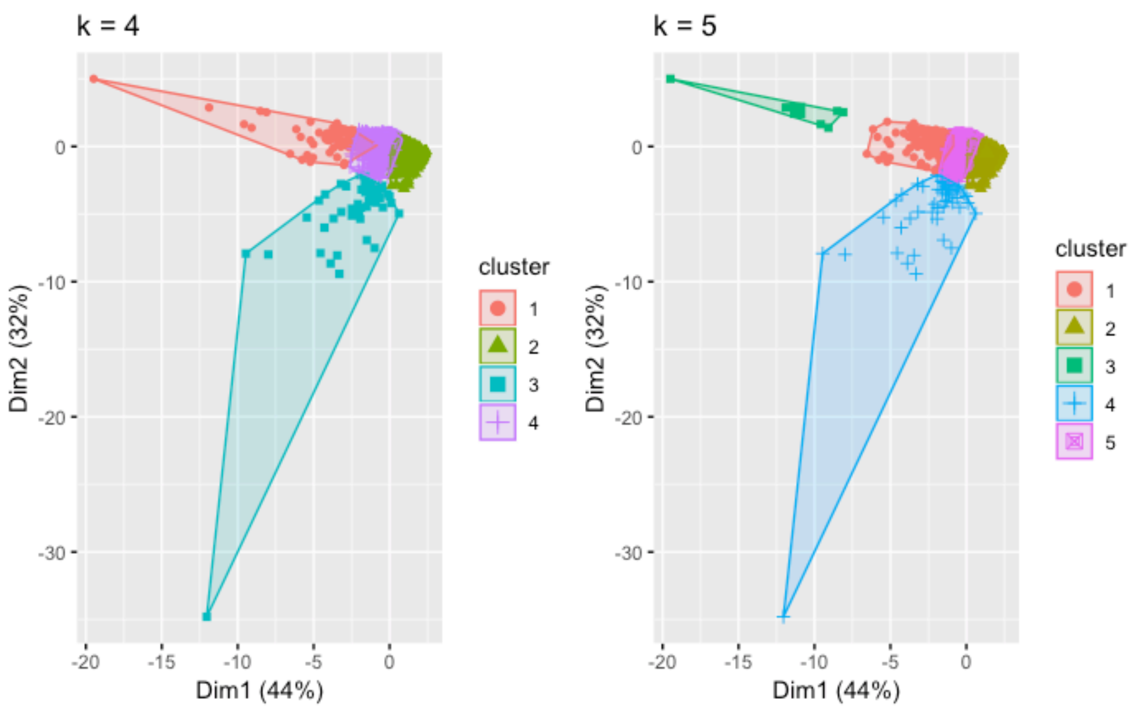
An elbow and silhouette methods with a k-means approach were first used for basic exploration. An arrangement with 4 and 5 clusters were considered.



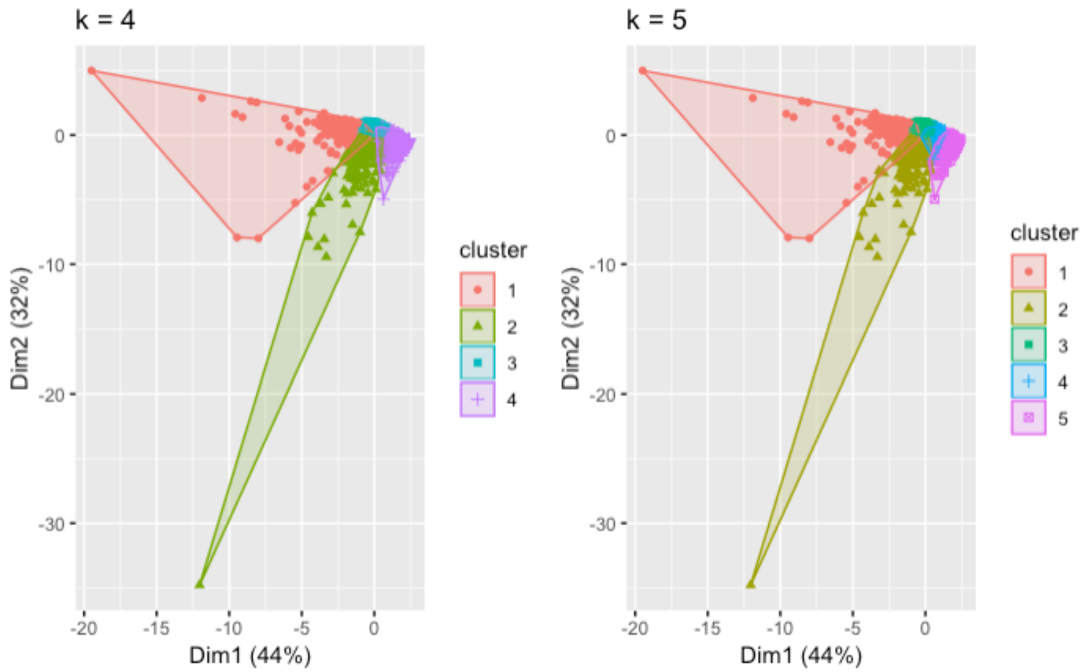


3 different clustering methods were used:

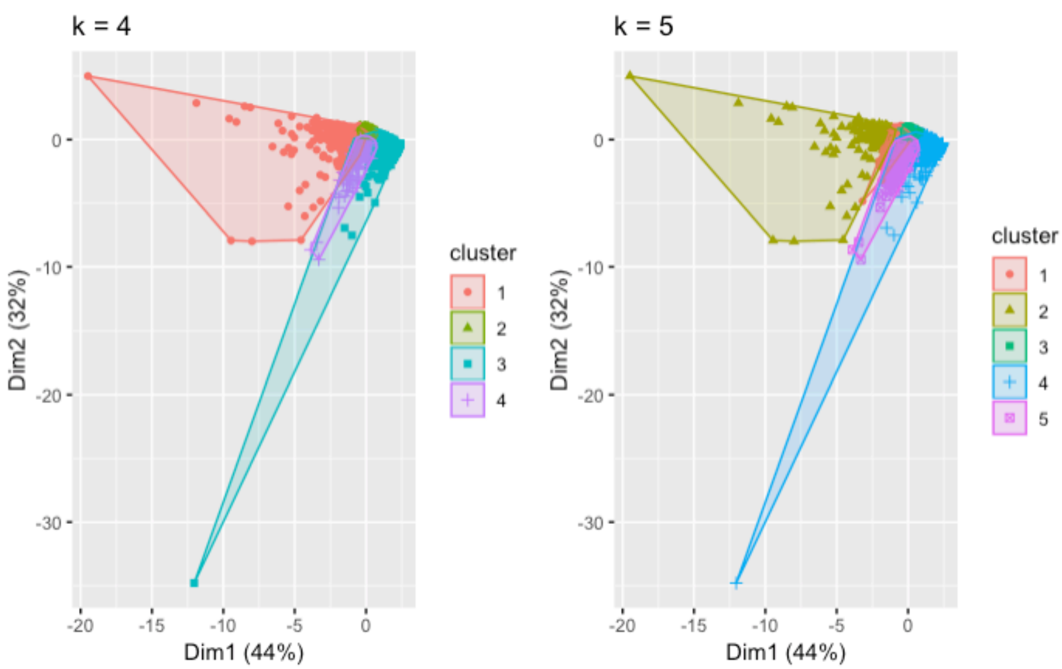
- **K-means**



- **PAM with Euclidean metric**



- **PAM with Manhattan metric**



All these clustering methods and arrangements were fitted into the original dataset.

				K-Means		PAM-Euclidean		PAM-Manhattan	
CustomerID	Recency	Frequency_1	Monetary	k4	k5	pe4	pe5	pm4	pm5
6 rows	<int>	<int>	<dbl>	<int>	<int>	<int>	<int>	<int>	<int>
12747	2	11	381.46	4	5	1	1	1	1
12748	0	206	150.72	1	3	1	1	1	2
12749	3	5	808.18	4	5	2	2	1	1
12820	3	4	235.59	4	5	3	3	2	3
12821	214	1	92.72	2	2	4	4	3	4
12822	70	2	474.44	4	5	2	2	4	5

For our next stage (Association Rules), we will need to join both the Retail and the RFM datasets, having in common the CustomerID.Quantity, unit price and revenue will be no longer needed.

InvoiceNo	Description	CustomerID	Date_Order	k4	k5	pe4	pe5	pm4	pm5
<fctr>	<fctr>	<fctr>	<date>	<int>	<int>	<int>	<int>	<int>	<int>
536365	WHITE HANGING HEART T-LIGHT HOLDER	17850	2010-12-01	1	2	1	1	3	4
536365	WHITE METAL LANTERN	17850	2010-12-01	1	2	1	1	3	4
536365	CREAM CUPID HEARTS COAT HANGER	17850	2010-12-01	1	2	1	1	3	4
536365	KNITTED UNION FLAG HOT WATER BOTTLE	17850	2010-12-01	1	2	1	1	3	4
536365	RED WOOLLY HOTTIE WHITE HEART.	17850	2010-12-01	1	2	1	1	3	4
536365	SET 7 BABUSHKA NESTING BOXES	17850	2010-12-01	1	2	1	1	3	4
536365	GLASS STAR FROSTED T-LIGHT HOLDER	17850	2010-12-01	1	2	1	1	3	4
536366	HAND WARMER UNION JACK	17850	2010-12-01	1	2	1	1	3	4
536366	HAND WARMER RED POLKA DOT	17850	2010-12-01	1	2	1	1	3	4
536367	ASSORTED COLOUR BIRD ORNAMENT	13047	2010-12-01	4	4	1	1	1	1

Association Rules

PREPROCESSING

FILE: 03_AR_PREPROCESSING.RMD

In order to proceed with the association rules, we first needed to prepare the different datasets for this process. These are all listed below (total=28 datasets):

- The whole dataset
- K=4 cluster arrangement
  - K-means
    - K=1 dataset
    - K=2 dataset
    - K=3 dataset
    - K=4 dataset
  - PAM-Euclidean
    - K=1 dataset
    - K=2 dataset
    - K=3 dataset
    - K=4 dataset
  - PAM-Manhattan
    - K=1 dataset
    - K=2 dataset
    - K=3 dataset
    - K=4 dataset
- K=5 cluster arrangement
  - K-means
    - K=1 dataset
    - K=2 dataset
    - K=3 dataset
    - K=4 dataset
    - K=5 dataset
  - PAM-Euclidean
    - K=1 dataset
    - K=2 dataset
    - K=3 dataset
    - K=4 dataset
    - K=5 dataset
  - PAM-Manhattan
    - K=1 dataset
    - K=2 dataset
    - K=3 dataset
    - K=4 dataset
    - K=5 dataset

Below is an example of the data preparation for the whole dataset. The same process will be performed on each cluster.

- The dataset is rearranged in a way that all transactions with the same Invoice number and date will be grouped. All the products will be also grouped in one column, separated by a comma

InvoiceNo	Date_Order	V1
536365	2010-12-01	WHITE HANGING HEART T-LIGHT HOLDER,WHITE METAL L...
536366	2010-12-01	HAND WARMER UNION JACK,HAND WARMER RED POLK...
536367	2010-12-01	ASSORTED COLOUR BIRD ORNAMENT,POPPY'S PLAYHOU...
536368	2010-12-01	JAM MAKING SET WITH JARS,RED COAT RACK PARIS FASHI...
536369	2010-12-01	BATH BUILDING BLOCK WORD
536371	2010-12-01	PAPER CHAIN KIT 50'S CHRISTMAS
536372	2010-12-01	HAND WARMER RED POLKA DOT,HAND WARMER UNION...
536373	2010-12-01	WHITE HANGING HEART T-LIGHT HOLDER,WHITE METAL L...
536374	2010-12-01	VICTORIAN SEWING BOX LARGE
536375	2010-12-01	WHITE HANGING HEART T-LIGHT HOLDER,WHITE METAL L...

- The InvoiceNo and the Date_Order are removed
- The remaining column is renamed "Products"

Products <chr>
WHITE HANGING HEART T-LIGHT HOLDER,WHITE METAL LANTERN,CREAM CUPID HEARTS COAT H...
HAND WARMER UNION JACK,HAND WARMER RED POLKA DOT
ASSORTED COLOUR BIRD ORNAMENT,POPPY'S PLAYHOUSE BEDROOM ,POPPY'S PLAYHOUSE KITCHE...
JAM MAKING SET WITH JARS,RED COAT RACK PARIS FASHION,YELLOW COAT RACK PARIS FASHION,...
BATH BUILDING BLOCK WORD
PAPER CHAIN KIT 50'S CHRISTMAS

- The new dataset is saved as a csv file in "D:\Users\dmoyano\Desktop\Github\Association_Rules"

This is done with every subset of clusters until we have all the cvs files ready for analysis with association rules

DETERMINING THE MIN. SUPPORT AND CONFIDENCE

FILE: 04_AR_SELECTION.RMD

Now we need to determine the minimum support and confidence that will be used as a comparison method among the clustering groups created from the step above.

We also required an arrangement of the dataset, this time with a product-based approach. In order to perform the Apriori algorithm, the monetary value is no longer required, but the product names become pivotal, as we need to understand how the purchase of one item may result in the purchase of another one.

To do so, the dataset was rearranged, so each row represents an invoice number that consists of one or more items purchased on a particular date (a basket of products).

Products <chr>
WHITE HANGING HEART T-LIGHT HOLDER,WHITE METAL LANTERN,CREAM CUPID HEARTS COAT HANGER,KNITTED U...
HAND WARMER UNION JACK,HAND WARMER RED POLKA DOT
ASSORTED COLOUR BIRD ORNAMENT,POPPY'S PLAYHOUSE BEDROOM ,POPPY'S PLAYHOUSE KITCHEN,FELTCRAFT PRIN...
JAM MAKING SET WITH JARS,RED COAT RACK PARIS FASHION,YELLOW COAT RACK PARIS FASHION,BLUE COAT RACK ...
BATH BUILDING BLOCK WORD
PAPER CHAIN KIT 50'S CHRISTMAS

In order to select the min. support and confidence levels that will be applied, we first selected one of the cluster groups (in this case, the observations assigned to the cluster #1 from the K4 column).

Lift will be assessed by having different combinations of the min. support and confidence levels:

We started by using a support level of 10% and a conf. level of 80%. However, there was no set of rules with that combination. Let's try supp.=10% and conf.=70%. That also resulted in 0 set of rules.

At this point, it is important to understand that this matrix has a massive number of products (3837 unique items), which means at least thousands of combinations. A support of 10% may be too high for the nature of this dataset.

A min support of 1% was applied, resulting in 90 rules. In order to find a combination with a goo lift, we will do some tests with the following parameters:

Min. support	Min. confidence
-1%	-70%
-1.5%	-75%
	-80%

1st group of the 4-cluster arrangement under K-means method

	MIN. CONFIDENCE	SET OF RULES	MEDIAN LIFT	MEAN LIFT	
1% SUPPORT	**70%**	**34**	23.03	25.22	
	75%	**14**	26.87	29.72	
	80%	**9**	28.22	35.19	**Selected**
1.5% SUPPORT	**70%**	**4**	27.1	26.76	
	75%	**4**	27.1	26.76	
	80%	set of 0 rules			

Combinations with 1.5% support do not provide much information: 2 out of 3 produce only 10 set of rules. Discounts and promotions only based on 10 set of rules may not be enough, especially when the data collected represent 2 years of transactions.

A 1% support gives more set of rules, and the highest lift is presented when the confidence is 80%

<u>A min. support of 1% and a min. confidence of 80% were chosen to evaluate the performance of each clustering method</u>.

The apriori algorithm was applied to every cluster and the results were compared based on the overall median and mean of the lift. At this point, it is important to understand what is considered a good lift under this business problem. This is a product-based approach and the main goal is to find potential relationships among some products that are purchased together.

The lift is a ratio between the probability of purchasing both X and Y products and the product of the probability of purchasing product X times the probability of purchasing product Y.

$$Lift(X \rightarrow Y) = \frac{support(X \cup Y)}{support(X).support(Y)}$$

If we are looking for complimentary products, we should be looking for a higher value in the numerator compared to the denominator. In other words, we are looking for products whose probability of being purchased together is higher than the product of their probabilities when purchased separately.

- When the lift is below 1, the product sets are substitutes (e.g. milk vs soy milk)
- When the lift is above 1, the product sets are complementary (e.g. a printer and ink cartridges)
- The closer the lift is to 1, it means that both the occurrence of the antecedent has almost no effect on the occurrence of the consequent

We are ideally looking for values above 1. The higher the better, as it indicates a stronger antecedent's influence over the consequent.

RESULTS

FILE: 04_AR_SELECTION.RMD

Results were compared in order to select the most appropriate arrangement for further analysis. Among the 4-cluster methods, PAM Manhattan seemed to perform better than the rest in terms of weighed lift average, while PAM Euclidean worked best for the 5-cluster arrangement. The following chart gives approximations to the results obtained from the Rmd file. The results in detail can be found in the APPENDIX A

		DATASET	# OF BASKETS	SUPPORT PORTION	MEAN LIFT	WEIGHED LIFT AVERAGE
		Whole DS	16577			
4 CLUSTERS	K-MEANS	Cluster 1	11892	118.92	35.19	25.24458467
		Cluster 2	1519	15.19	41.83	3.833007782
		Cluster 3	383	3.83	43.313	1.000716595
		Cluster 4	2783	27.83	64.76	10.87211679
						40.95
	PAM EUCLIDEAN	Cluster 1	7383	73.83	41.564	18.5116132
		Cluster 2	2520	25.2	39.923	6.069008868
		Cluster 3	5204	52.04	46.93	14.73268505
		Cluster 4	1470	14.7	42.844	3.799280931
						43.11
	PAM MANHATTAN	Cluster 1	9416	94.16	47.163	26.7893351
		Cluster 2	3296	32.96	38.44	7.643013814
		Cluster 3	1461	14.61	42.622	3.756454244
		Cluster 4	2404	24.04	43.828	6.355945708
						44.54

		DATASET	# OF BASKETS	SUPPORT PORTION	MEAN LIFT	WEIGHED LIFT AVERAGE
		Whole DS	16577			
5 CLUSTERS	K-MEANS	Cluster 1	3	0.03	2	0.000361947
		Cluster 2	11150	111.5	40.91	27.51683055
		Cluster 3	570	5.7	38.9	1.33757616
		Cluster 4	3352	33.52	53.6	10.83834228
		Cluster 5	1502	15.02	44.12	3.997601496
						43.69
	PAM EUCLIDEAN	Cluster 1	7332	73.32	41.14	18.19620438
		Cluster 2	2118	21.18	30.43	3.887961634
		Cluster 3	5089	50.89	44.29	13.59665862
		Cluster 4	1319	13.19	39.43	3.137369247
		Cluster 5	719	7.19	45.71	1.982595765
						40.80
	PAM MANHATTAN	Cluster 1	5127	51.27	17.06	5.276384147
		Cluster 2	5277	52.77	37.475	11.92951529
		Cluster 3	2811	28.11	34.42	5.83667853
		Cluster 4	1435	14.35	42.05	3.640088677
		Cluster 5	1927	19.27	38.636	4.491257284
						31.17

LIFT PER CLUSTER ARRANGEMENT

Both arrangements provide relevant information about potential groups in the online retail's customer base. However, The PAM Euclidean metric for 5 clusters seems to show a clearer delimitation of the groups, as this arrangement provides information about some of the most profitable customers, some of the most loyal ones or some customers that the company might lose if there is no action.

<u>The 5-cluster PAM-Euclidean arrangement was selected</u>

CLUSTER FEATURES

FILE: 05_RESULTS.RMD

PAMEuclidean5 <int>	NoCustomers <int>	Percentage <dbl>	AvgRecency <dbl>	MaxRecency <int>	MinRecency <int>	AvgFrequency <dbl>	MaxFrequency <int>	MinFrequency <int>	AvgMonetary <dbl>	MaxMonetary <dbl>	MinMonetary <dbl>
1	426	10.88	15.08685	372	0	17.211268	206	8	401.5493	4327.62	33.24
2	647	16.52	51.56414	290	0	3.273570	19	1	759.0119	14844.77	404.80
3	1659	42.36	34.44002	103	0	3.067511	9	1	231.4952	428.22	0.00
4	631	16.11	159.29477	222	93	2.090333	12	1	246.1310	931.50	2.90
5	553	14.12	293.45931	373	225	1.300181	8	1	273.5552	2002.40	3.75

- #1: 10.88% of the dataset
    - The most frequent customers
    - The second highest in avg. monetary
    - Customers who have purchased recently the most
- #2: 16.52% of the dataset
    - The second most frequent customers
    - The most profitable ones in avg. monetary
    - 3rd in recency
- #3: 42.36% of the dataset
    - 3rd most frequent customers
    - Their average monetary is the lowest of all
    - They represent the biggest cluster in the dataset
    - 2nd customers who have purchased recently
- #4: 16.11% of the dataset
    - Its recency is the second highest. They might not be customers anymore
    - Its frequency is the second lowest
    - The average monetary is similar to the 3^{rd} group
- #5: 14.12% of the dataset
    - Its recency is the highest. They might not be customers anymore
    - Its frequency is the lowest
    - The average monetary is slightly higher to the 3^{rd} group

ASSOCIATION RULES FOR THE CLUSTER SELECTED

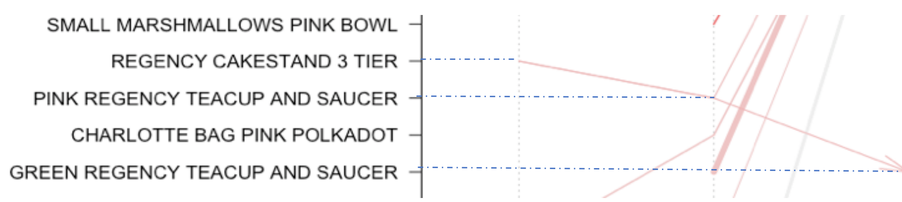
FILE: 05_RESULTS.RMD

Each group was analyzed based on the plots obtained in the results that also be found in the APPENDIX B These plots are:

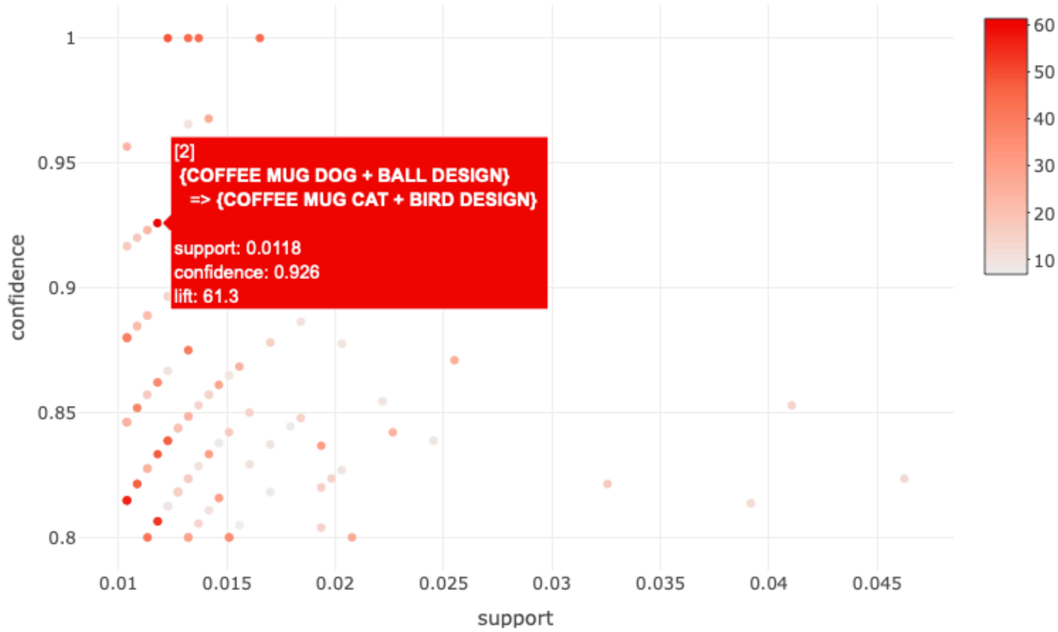
- Most relevant rules (non-redundant)
- Interactive plot with x=support, y=confidence, color=lift (deep red means a higher lift)
- Top-10 rule network: better visualization of the rules present in the group
- Parallel analysis: another way to show the relationship among the items.

Some of the relevant insights obtained are listed below

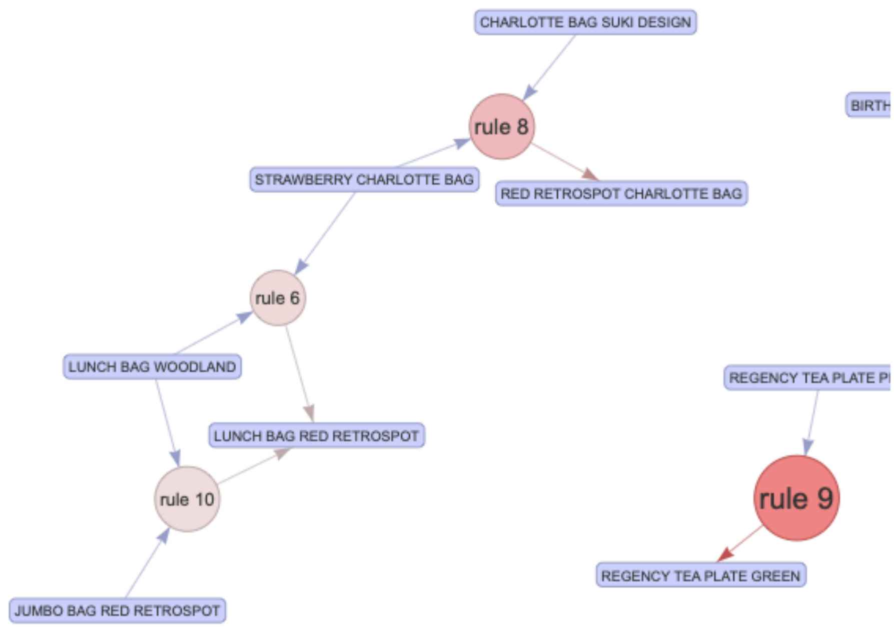
- Group 1
  - Some of the most relevant rules are
    - Back door -> Key Fob (one of the highest lifts)
    - Shed -> Key fob (one of the highest lifts)
    - Set 3 Retrospot tea -> Coffee
    - Sugar->Coffee (highest lift)
    - Regency Tea Plate Green -> Regency Tea Plate Roses
  - The parallel plot provides additional information for a set of 3 rules. For instance, if a customer purchases the Regency Cakestand 3 tier and the Pink Regency Teacup and Saucer, he/she is more likely to buy the Green Recency Teacup and Saucer



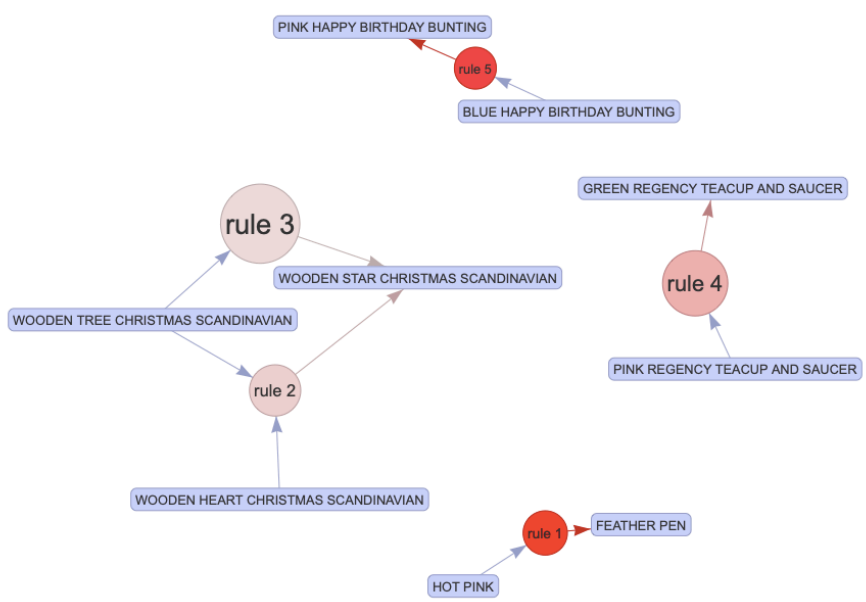
- Group 2
  - This group has more combinations to consider. Some of the ones with high lift are shown in the interactive plot below:



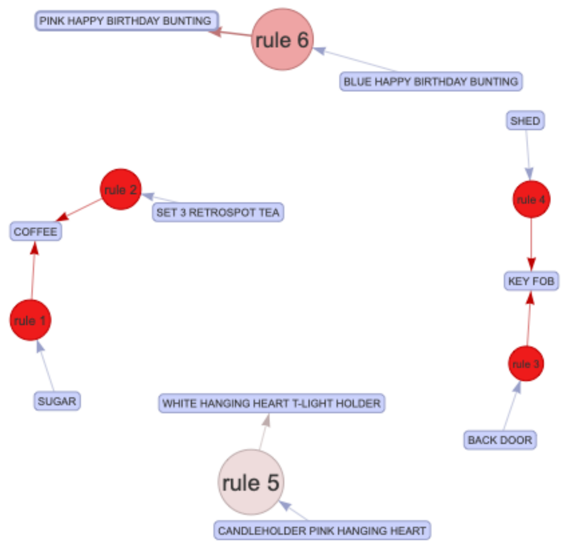
- In addition to this, this group seems to have a rule network with one of the longest relation among some items (rules 8, 6 and 10)



- Group 3
  - This group seems to be a bit more seasonal than the rest, as their most relevant rules are related to Christmas or to special occasions.
  - The parallel plot also shows interdependency among the rule 2 and 3 (related to the Christmas items below displayed)



- Group 4
    - Not much information can be obtained for this particular group. However, we can observe that some of the rules are very similar to the ones of the 1st group (related to the Retrospot tea, sugar and coffee, as well as the shed, the back door and the key fob)



- Group 5
    - This group did not provide as much info as the rest. However, some of their most relevant rules are alphabet stencil craft->happy stencil craft (lift=50) or kitchen metal sign->bathroom metal sign

CONCLUSIONS

Unsupervised learning allowed us to find some insights that would allow us to build some business strategies with a customer-based approach (through clustering) and a product-based approach (through association rules).

Through the evaluation and selection of the appropriate clustering method, we were able to obtain a set of clusters that present a good degree of lift. The products associated can be used in promotions and discounts that can increase the frequency these groups purchase or the amount purchased in every transaction.

Based on the information above, we can conclude the following

- The most profitable customers are the 1st and the 2nd ones, representing approximately 28% of the whole dataset and roughly 60% of the company's revenue
- The first group can be considered as the most loyal customer base given the frequency of purchase. Potential promotions and discounts offered based on products can be used to keep that loyalty. The association rules obtained in the prior section can provide insights on what products to promote or give discounts
- The second group may not be the most frequent customers, but it is certainly one of the most profitable ones, given the avg. monetary they present. Give the average monetary levels this particular group has, discounts per volume might be one of the best ways to approach and retain these customers. They also seem to be focused on gift items, hence promotions involving these products may also result in higher revenue.
- The 3rd group has the potential to become either group 1 or 2, as their recency is similar to these groups. We might look into either increasing their frequency or the amount purchased through a product-based approach given by the association rules. Potential actions toward this group is finding out if their purchase habits are based on

seasonality. It's important to remember that the kind of products sold by this retailer is for unique occasion, hence moments like Christmas, Valentine's day, etc. might have an impact

- The 4th and 5th groups' purchase habits might be seasonal as well. However, they also present the highest recency values. The 4th however, could be dormant customers that can be reach out through promotions and some customer service.
- Product placement is pivotal when displaying products in the company's website. The association rules found in this project should be easy to select when the customer is about to make a transaction. The online retailer can also track how these products were purchased together, allowing us to access more information that can improve the algorithms used.

The assessment of the different cluster arrangements through association rules provided us with the tools to compare the effectiveness these unsupervised methods have. In addition to this, the results obtained are relevant from a business perspective, as it provides insights that can be translated into actionable items.

FUTURE DISCUSSION

These ones include

- Understanding **seasonality** and do some analysis based on time series. It is important to know that we may require more years in the dataset in order to make a more accurate prediction
- Understanding in what way the **number of products purchased** can affect the way the association rules behave. The Apriori algorithm used in this project only considers if the product was purchased or not, and not how much of that product was purchased.
- More **in-depth analysis on the most profitable customers**: kind of products, habits, days of the week they regularly do transactions. This way, the online retailer can provide a more customized treatment that can result in higher revenue.
- Extend this model to the **non-UK customer base** in order to understand insights that can provide opportunities to expand their business overseas.

SOURCES

- Collin, James; Gates, Lee. *Retail Analytics Report*. http://scet.berkeley.edu/wp-content/uploads/UCBSCETRetailAnalyticsReport.pdf
- Davenport, Thomas. *Realizing the Potential of Retail Analytics*. http://analytics.typepad.com/files/retailanalytics.pdf
- Deloitte. *Analytics in Retail: Going to Market with a Smarter Approach*. https://www2.deloitte.com/content/dam/Deloitte/ch/Documents/consumer-business/ch-cb-en-Deloitte-Analytics-in-retail-0514.pdf
- EKN. *The Future of Retail Analytics*. https://www.sas.com/content/dam/SAS/en_us/doc/research2/ekn-report-future-retail-analytics-106717.pdf
- Fuloria, Sanjay. *How Advanced Analytics Will Inform and Transform U.S. Retail*. https://www.cognizant.com/InsightsWhitepapers/How-Advanced-Analytics-Will-Inform-and-Transform-US-Retail.pdf
- MIT Open Courseware. *Discovering Association Rules in Transaction Databases*. https://ocw.mit.edu/courses/sloan-school-of-management/15-062-data-mining-spring-2003/lecture-notes/Lecture_16.pdf
- Peng, Roger. *Exploratory Data Analysis with R*. Lulu.com, 2012
- Pollack, Joshua. *Retail Clustering Methods*. http://www.parkeravery.com/pov_Retail_Clustering_Methods.html
- Chen, Daqing. *Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining.* 18th July, 2012

APPENDIX A: DETAILED APRIORI RESULTS PER CLUSTER ARRANGEMENT

##4 CLUSTERS

KMEANS

C1

```
set of 57 rules

rule length distribution (lhs + rhs):sizes
 2  3 
37 20 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   2.000   2.000   2.351   3.000   3.000 

summary of quality measures:
    support           confidence          lift            count      
 Min.   :0.01006   Min.   :0.8000   Min.   : 8.838   Min.   :28.00  
 1st Qu.:0.01006   1st Qu.:0.8750   1st Qu.:50.618   1st Qu.:28.00  
 Median :0.01042   Median :0.9355   Median :73.958   Median :29.00  
 Mean   :0.01202   Mean   :0.9223   Mean   :64.762   Mean   :33.47  
 3rd Qu.:0.01401   3rd Qu.:0.9667   3rd Qu.:81.882   3rd Qu.:39.00  
 Max.   :0.02550   Max.   :1.0000   Max.   :92.690   Max.   :71.00  
```

C2

```
set of 19 rules

rule length distribution (lhs + rhs):sizes
 2  3 
 8 11 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   2.000   3.000   2.579   3.000   3.000 

summary of quality measures:
    support           confidence          lift          count      
 Min.   :0.01118   Min.   :0.8000   Min.   : 7.60   Min.   :17.00  
 1st Qu.:0.01250   1st Qu.:0.8718   1st Qu.:18.73   1st Qu.:19.00  
 Median :0.01250   Median :0.9444   Median :36.81   Median :19.00  
 Mean   :0.01537   Mean   :0.9263   Mean   :44.20   Mean   :23.37  
 3rd Qu.:0.01875   3rd Qu.:1.0000   3rd Qu.:63.33   3rd Qu.:28.50  
 Max.   :0.02368   Max.   :1.0000   Max.   :80.00   Max.   :36.00  
```

C3

```
set of 372348 rules

rule length distribution (lhs + rhs):sizes
     2      3 
  3918 368430 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   3.000   3.000   2.989   3.000   3.000 

summary of quality measures:
    support           confidence          lift             count       
 Min.   :0.01042   Min.   :0.8000   Min.   : 4.585   Min.   : 4.000  
 1st Qu.:0.01042   1st Qu.:0.8000   1st Qu.:25.600   1st Qu.: 4.000  
 Median :0.01042   Median :1.0000   Median :38.400   Median : 4.000  
 Mean   :0.01100   Mean   :0.9335   Mean   :40.776   Mean   : 4.222  
 3rd Qu.:0.01042   3rd Qu.:1.0000   3rd Qu.:54.857   3rd Qu.: 4.000  
 Max.   :0.07292   Max.   :1.0000   Max.   :96.000   Max.   :28.000  
```

C4

```
set of 9 rules

rule length distribution (lhs + rhs):sizes
2 3 
5 4 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   2.000   2.000   2.444   3.000   3.000 

summary of quality measures:
    support           confidence          lift           count      
 Min.   :0.01034   Min.   :0.8037   Min.   :10.62   Min.   :123.0  
 1st Qu.:0.01127   1st Qu.:0.8040   1st Qu.:25.98   1st Qu.:134.0  
 Median :0.01177   Median :0.8439   Median :28.22   Median :140.0  
 Mean   :0.01299   Mean   :0.8658   Mean   :35.19   Mean   :154.4  
 3rd Qu.:0.01228   3rd Qu.:0.8733   3rd Qu.:50.64   3rd Qu.:146.0  
 Max.   :0.02035   Max.   :1.0000   Max.   :56.90   Max.   :242.0  
```

PAM - EUCLIDEAN

C1

```
set of 24 rules

rule length distribution (lhs + rhs):sizes
 2  3
10 14

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.583   3.000   3.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.01016   Min.   :0.8000   Min.   : 8.331   Min.   : 75.00
 1st Qu.:0.01148   1st Qu.:0.8234   1st Qu.:21.410   1st Qu.: 84.75
 Median :0.01151   Median :0.8635   Median :34.197   Median : 85.00
 Mean   :0.01314   Mean   :0.8996   Mean   :41.564   Mean   : 97.00
 3rd Qu.:0.01331   3rd Qu.:1.0000   3rd Qu.:61.533   3rd Qu.: 98.25
 Max.   :0.02600   Max.   :1.0000   Max.   :86.871   Max.   :192.00
```

C3

```
set of 5 rules

rule length distribution (lhs + rhs):sizes
2 3
4 1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     2.0     2.0     2.2     2.0     3.0

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.01018   Min.   :0.8372   Min.   :29.49   Min.   :53
 1st Qu.:0.01037   1st Qu.:0.8438   1st Qu.:32.28   1st Qu.:54
 Median :0.01134   Median :0.8556   Median :40.73   Median :59
 Mean   :0.01210   Mean   :0.8946   Mean   :46.93   Mean   :63
 3rd Qu.:0.01383   3rd Qu.:0.9365   3rd Qu.:62.74   3rd Qu.:72
 Max.   :0.01479   Max.   :1.0000   Max.   :69.40   Max.   :77
```

C2

```
set of 249 rules

rule length distribution (lhs + rhs):sizes
  2   3
 63 186

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.747   3.000   3.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.01031   Min.   :0.8000   Min.   : 7.456   Min.   :26.00
 1st Qu.:0.01111   1st Qu.:0.8407   1st Qu.:17.898   1st Qu.:28.00
 Median :0.01190   Median :0.8913   Median :42.729   Median :30.00
 Mean   :0.01306   Mean   :0.8946   Mean   :43.320   Mean   :32.92
 3rd Qu.:0.01309   3rd Qu.:0.9394   3rd Qu.:66.026   3rd Qu.:33.00
 Max.   :0.03768   Max.   :1.0000   Max.   :90.036   Max.   :95.00
```

C4

```
set of 19 rules

rule length distribution (lhs + rhs):sizes
 2  3
 7 12

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.632   3.000   3.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.01020   Min.   :0.8095   Min.   : 7.742   Min.   :15.00
 1st Qu.:0.01224   1st Qu.:0.8536   1st Qu.:19.014   1st Qu.:18.00
 Median :0.01224   Median :0.9444   Median :38.591   Median :18.00
 Mean   :0.01499   Mean   :0.9211   Mean   :45.290   Mean   :22.05
 3rd Qu.:0.01801   3rd Qu.:1.0000   3rd Qu.:67.002   3rd Qu.:26.50
 Max.   :0.02311   Max.   :1.0000   Max.   :81.722   Max.   :34.00
```

PAM - MANHATTAN

C1

```
set of 20 rules

rule length distribution (lhs + rhs):sizes
 2  3
 8 12

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     2.0     3.0     2.6     3.0     3.0

summary of quality measures:
    support          confidence          lift            count
 Min.   :0.01030   Min.   :0.8042   Min.   : 8.755   Min.   : 97.0
 1st Qu.:0.01104   1st Qu.:0.8292   1st Qu.:22.534   1st Qu.:104.0
 Median :0.01104   Median :0.8624   Median :49.877   Median :104.0
 Mean   :0.01195   Mean   :0.9084   Mean   :47.163   Mean   :112.5
 3rd Qu.:0.01189   3rd Qu.:1.0000   3rd Qu.:64.500   3rd Qu.:112.0
 Max.   :0.01742   Max.   :1.0000   Max.   :90.548   Max.   :164.0
```

C2

```
set of 2 rules

rule length distribution (lhs + rhs):sizes
2
2

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2       2       2       2       2       2

summary of quality measures:
    support          confidence          lift           count
 Min.   :0.01304   Min.   :0.8462   Min.   :30.00   Min.   :43.00
 1st Qu.:0.01312   1st Qu.:0.8586   1st Qu.:34.22   1st Qu.:43.25
 Median :0.01319   Median :0.8710   Median :38.44   Median :43.50
 Mean   :0.01319   Mean   :0.8710   Mean   :38.44   Mean   :43.50
 3rd Qu.:0.01327   3rd Qu.:0.8834   3rd Qu.:42.66   3rd Qu.:43.75
 Max.   :0.01335   Max.   :0.8958   Max.   :46.88   Max.   :44.00
```

C3

```
set of 391 rules

rule length distribution (lhs + rhs):sizes
  2   3
 21 370

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.000   2.946   3.000   3.000

summary of quality measures:
    support          confidence          lift            count
 Min.   :0.01026   Min.   :0.8000   Min.   : 6.255   Min.   :15.00
 1st Qu.:0.01026   1st Qu.:0.8824   1st Qu.:30.458   1st Qu.:15.00
 Median :0.01026   Median :0.9375   Median :41.534   Median :15.00
 Mean   :0.01081   Mean   :0.9412   Mean   :41.210   Mean   :15.81
 3rd Qu.:0.01094   3rd Qu.:1.0000   3rd Qu.:54.148   3rd Qu.:16.00
 Max.   :0.02599   Max.   :1.0000   Max.   :73.100   Max.   :38.00
```

C4

```
set of 184 rules

rule length distribution (lhs + rhs):sizes
  2   3
 60 124

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.674   3.000   3.000

summary of quality measures:
    support          confidence          lift            count
 Min.   :0.01040   Min.   :0.8000   Min.   : 7.185   Min.   :25.00
 1st Qu.:0.01164   1st Qu.:0.8571   1st Qu.:30.462   1st Qu.:28.00
 Median :0.01414   Median :0.8993   Median :52.980   Median :34.00
 Mean   :0.01431   Mean   :0.9040   Mean   :44.028   Mean   :34.41
 3rd Qu.:0.01538   3rd Qu.:0.9515   3rd Qu.:55.726   3rd Qu.:37.00
 Max.   :0.03825   Max.   :1.0000   Max.   :85.893   Max.   :92.00
```

##5 CLUSTERS

KMEANS

C1

```
set of 28 rules

rule length distribution (lhs + rhs):sizes
 2  3
12 16

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.571   3.000   3.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.01019   Min.   :0.8015   Min.   : 7.643   Min.   : 49.00
 1st Qu.:0.01164   1st Qu.:0.8339   1st Qu.:20.227   1st Qu.: 56.00
 Median :0.01289   Median :0.8603   Median :24.319   Median : 62.00
 Mean   :0.01401   Mean   :0.8897   Mean   :36.919   Mean   : 67.36
 3rd Qu.:0.01404   3rd Qu.:1.0000   3rd Qu.:56.576   3rd Qu.: 67.50
 Max.   :0.02682   Max.   :1.0000   Max.   :77.565   Max.   :129.00
```

C2

```
set of 20 rules

rule length distribution (lhs + rhs):sizes
 2  3
 8 12

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     2.0     3.0     2.6     3.0     3.0

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.01134   Min.   :0.8077   Min.   : 7.656   Min.   :17.00
 1st Qu.:0.01268   1st Qu.:0.8629   1st Qu.:18.973   1st Qu.:19.00
 Median :0.01268   Median :0.9196   Median :41.964   Median :19.00
 Mean   :0.01514   Mean   :0.9199   Mean   :44.153   Mean   :22.70
 3rd Qu.:0.01818   3rd Qu.:1.0000   3rd Qu.:63.137   3rd Qu.:27.25
 Max.   :0.02335   Max.   :1.0000   Max.   :78.895   Max.   :35.00
```

C3

```
set of 443 rules

rule length distribution (lhs + rhs):sizes
  2   3
 56 387

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.000   2.874   3.000   3.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.01007   Min.   :0.8000   Min.   : 9.115   Min.   : 7.00
 1st Qu.:0.01007   1st Qu.:0.8750   1st Qu.:15.444   1st Qu.: 7.00
 Median :0.01151   Median :0.8889   Median :23.167   Median : 8.00
 Mean   :0.01184   Mean   :0.9144   Mean   :34.974   Mean   : 8.23
 3rd Qu.:0.01295   3rd Qu.:1.0000   3rd Qu.:43.438   3rd Qu.: 9.00
 Max.   :0.02590   Max.   :1.0000   Max.   :99.286   Max.   :18.00
```

C4

```
set of 372348 rules

rule length distribution (lhs + rhs):sizes
     2      3
  3918 368430

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.000   2.989   3.000   3.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.01042   Min.   :0.8000   Min.   : 4.585   Min.   : 4.000
 1st Qu.:0.01042   1st Qu.:0.8000   1st Qu.:25.600   1st Qu.: 4.000
 Median :0.01042   Median :1.0000   Median :38.400   Median : 4.000
 Mean   :0.01100   Mean   :0.9335   Mean   :40.776   Mean   : 4.222
 3rd Qu.:0.01042   3rd Qu.:1.0000   3rd Qu.:54.857   3rd Qu.: 4.000
 Max.   :0.07292   Max.   :1.0000   Max.   :96.000   Max.   :28.000
```

C5

```
set of 9 rules

rule length distribution (lhs + rhs):sizes
2 3
6 3

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   2.000   2.333   3.000   3.000

summary of quality measures:
    support           confidence          lift           count
 Min.   :0.01022   Min.   :0.8103   Min.   :26.91   Min.   : 94.0
 1st Qu.:0.01153   1st Qu.:0.8283   1st Qu.:29.01   1st Qu.:106.0
 Median :0.01207   Median :0.8740   Median :36.53   Median :111.0
 Mean   :0.01324   Mean   :0.8873   Mean   :41.58   Mean   :121.8
 3rd Qu.:0.01338   3rd Qu.:0.9216   3rd Qu.:55.73   3rd Qu.:123.0
 Max.   :0.02099   Max.   :1.0000   Max.   :57.32   Max.   :193.0
```

PAM - EUCLIDEAN

C1

```
set of 23 rules

rule length distribution (lhs + rhs):sizes
 2  3
 9 14

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.609   3.000   3.000

summary of quality measures:
    support           confidence          lift             count
 Min.   :0.01023   Min.   :0.8000   Min.   : 8.258   Min.   : 75.00
 1st Qu.:0.01146   1st Qu.:0.8240   1st Qu.:21.486   1st Qu.: 84.00
 Median :0.01146   Median :0.8590   Median :26.690   Median : 84.00
 Mean   :0.01317   Mean   :0.8991   Mean   :41.143   Mean   : 96.61
 3rd Qu.:0.01323   3rd Qu.:1.0000   3rd Qu.:61.622   3rd Qu.: 97.00
 Max.   :0.02577   Max.   :1.0000   Max.   :87.298   Max.   :189.00
```

C2

```
set of 371 rules

rule length distribution (lhs + rhs):sizes
  2   3
 76 295

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.000   2.795   3.000   3.000

summary of quality measures:
    support           confidence          lift             count
 Min.   :0.01038   Min.   :0.8000   Min.   : 6.976   Min.   :22.00
 1st Qu.:0.01133   1st Qu.:0.8387   1st Qu.:16.090   1st Qu.:24.00
 Median :0.01274   Median :0.8780   Median :31.430   Median :27.00
 Mean   :0.01366   Mean   :0.8855   Mean   :34.367   Mean   :28.94
 3rd Qu.:0.01416   3rd Qu.:0.9310   3rd Qu.:57.390   3rd Qu.:30.00
 Max.   :0.04625   Max.   :1.0000   Max.   :78.481   Max.   :98.00
```

C3

```
set of 5 rules

rule length distribution (lhs + rhs):sizes
2 3
4 1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     2.0     2.0     2.2     2.0     3.0

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.01041   Min.   :0.8030   Min.   :28.20   Min.   :53.0
 1st Qu.:0.01061   1st Qu.:0.8333   1st Qu.:30.79   1st Qu.:54.0
 Median :0.01179   Median :0.8587   Median :39.64   Median :60.0
 Mean   :0.01242   Mean   :0.8865   Mean   :44.29   Mean   :63.2
 3rd Qu.:0.01375   3rd Qu.:0.9375   3rd Qu.:58.39   3rd Qu.:70.0
 Max.   :0.01552   Max.   :1.0000   Max.   :64.43   Max.   :79.0
```

C4

```
set of 19 rules

rule length distribution (lhs + rhs):sizes
 2  3
10  9

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   2.000   2.474   3.000   3.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.01061   Min.   :0.8065   Min.   : 7.40   Min.   :14.00
 1st Qu.:0.01288   1st Qu.:0.8450   1st Qu.:20.45   1st Qu.:17.00
 Median :0.01288   Median :0.9375   Median :28.87   Median :17.00
 Mean   :0.01651   Mean   :0.9205   Mean   :42.02   Mean   :21.79
 3rd Qu.:0.02045   3rd Qu.:1.0000   3rd Qu.:60.00   3rd Qu.:27.00
 Max.   :0.02500   Max.   :1.0000   Max.   :77.65   Max.   :33.00
```

C5

```
set of 34 rules

rule length distribution (lhs + rhs):sizes
 2  3
15 19

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.559   3.000   3.000

summary of quality measures:
    support           confidence          lift            count
 Min.   :0.01111   Min.   :0.8125   Min.   : 7.50   Min.   : 8.00
 1st Qu.:0.01250   1st Qu.:0.8750   1st Qu.:21.84   1st Qu.: 9.00
 Median :0.01250   Median :1.0000   Median :46.86   Median : 9.00
 Mean   :0.01536   Mean   :0.9385   Mean   :45.32   Mean   :11.06
 3rd Qu.:0.01597   3rd Qu.:1.0000   3rd Qu.:65.45   3rd Qu.:11.50
 Max.   :0.03611   Max.   :1.0000   Max.   :80.00   Max.   :26.00
```

PAM - MANHATTAN

C1

```
set of 20 rules

rule length distribution (lhs + rhs):sizes
 2  3
 6 14

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.0     2.0     3.0     2.7     3.0     3.0

summary of quality measures:
    support           confidence          lift            count      
 Min.   :0.01014   Min.   :0.8000   Min.   :10.71   Min.   :52.00  
 1st Qu.:0.01048   1st Qu.:0.8106   1st Qu.:11.15   1st Qu.:53.75  
 Median :0.01112   Median :0.8279   Median :17.64   Median :57.00  
 Mean   :0.01145   Mean   :0.8444   Mean   :23.61   Mean   :58.70  
 3rd Qu.:0.01229   3rd Qu.:0.8492   3rd Qu.:29.95   3rd Qu.:63.00  
 Max.   :0.01385   Max.   :1.0000   Max.   :56.35   Max.   :71.00  
```

C2

```
set of 26 rules

rule length distribution (lhs + rhs):sizes
 2  3
11 15

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.577   3.000   3.000

summary of quality measures:
    support           confidence          lift             count       
 Min.   :0.01004   Min.   :0.8000   Min.   : 7.685   Min.   : 53.00  
 1st Qu.:0.01175   1st Qu.:0.8288   1st Qu.:20.207   1st Qu.: 62.00  
 Median :0.01222   Median :0.8792   Median :26.961   Median : 64.50  
 Mean   :0.01413   Mean   :0.8998   Mean   :38.364   Mean   : 74.58  
 3rd Qu.:0.01454   3rd Qu.:1.0000   3rd Qu.:58.103   3rd Qu.: 76.75  
 Max.   :0.02709   Max.   :1.0000   Max.   :85.129   Max.   :143.00  
```

C3

```
set of 3 rules

rule length distribution (lhs + rhs):sizes
2 3
2 1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   2.000   2.333   2.500   3.000

summary of quality measures:
    support           confidence          lift           count     
 Min.   :0.01031   Min.   :0.8511   Min.   :27.83   Min.   :29.0  
 1st Qu.:0.01209   1st Qu.:0.8787   1st Qu.:28.73   1st Qu.:34.0  
 Median :0.01387   Median :0.9062   Median :29.63   Median :39.0  
 Mean   :0.01280   Mean   :0.8953   Mean   :34.42   Mean   :36.0  
 3rd Qu.:0.01405   3rd Qu.:0.9174   3rd Qu.:37.72   3rd Qu.:39.5  
 Max.   :0.01422   Max.   :0.9286   Max.   :45.81   Max.   :40.0  
```

C4

```
set of 390 rules

rule length distribution (lhs + rhs):sizes
  2   3
 20 370

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.000   2.949   3.000   3.000

summary of quality measures:
    support           confidence          lift             count      
 Min.   :0.01045   Min.   :0.8000   Min.   : 6.243   Min.   :15.00  
 1st Qu.:0.01045   1st Qu.:0.8824   1st Qu.:29.917   1st Qu.:15.00  
 Median :0.01045   Median :0.9375   Median :41.090   Median :15.00  
 Mean   :0.01100   Mean   :0.9409   Mean   :40.622   Mean   :15.79  
 3rd Qu.:0.01114   3rd Qu.:1.0000   3rd Qu.:53.185   3rd Qu.:16.00  
 Max.   :0.02577   Max.   :1.0000   Max.   :71.800   Max.   :37.00  
```

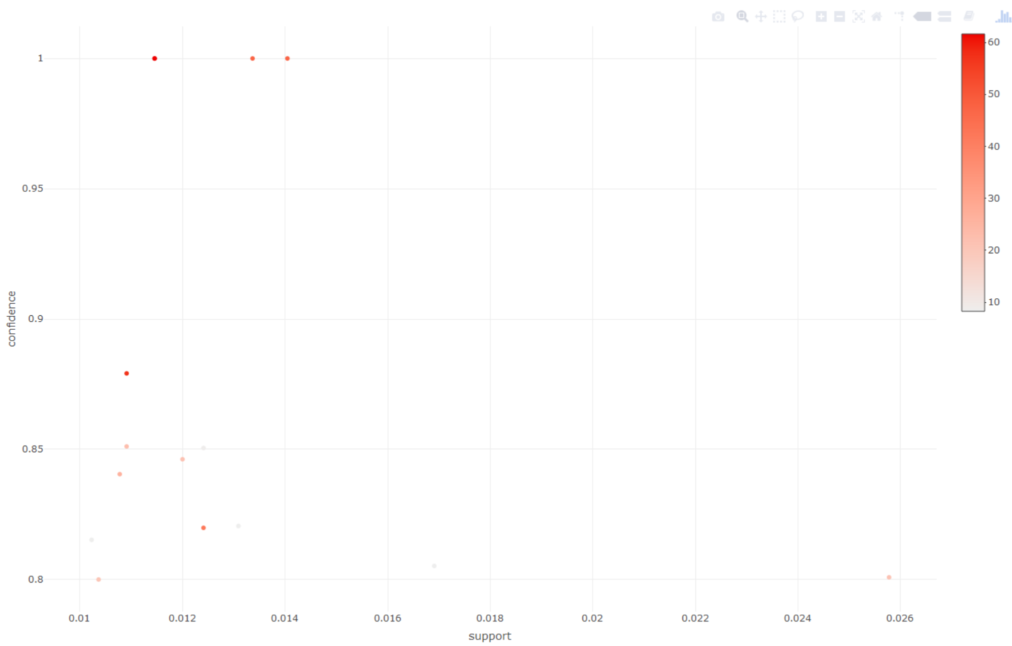
C5

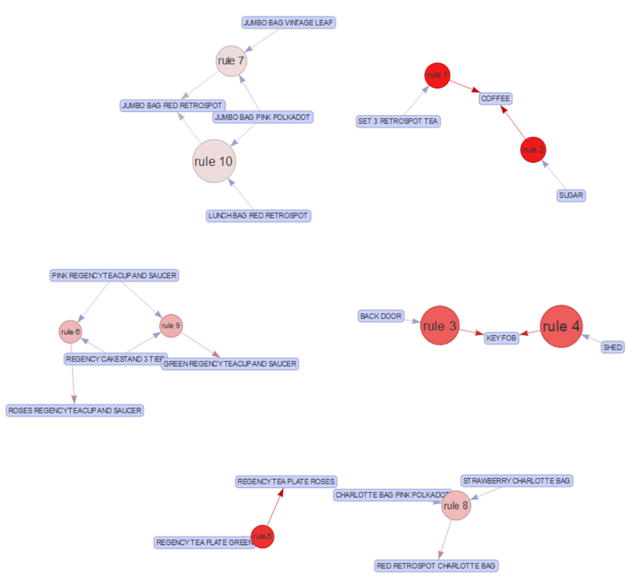
```
set of 231 rules

rule length distribution (lhs + rhs):sizes
  2   3
 63 168

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   2.000   3.000   2.727   3.000   3.000

summary of quality measures:
    support           confidence          lift             count      
 Min.   :0.01037   Min.   :0.8000   Min.   : 6.724   Min.   :20.00  
 1st Qu.:0.01141   1st Qu.:0.8550   1st Qu.:26.431   1st Qu.:22.00  
 Median :0.01349   Median :0.9032   Median :37.804   Median :26.00  
 Mean   :0.01456   Mean   :0.9022   Mean   :39.553   Mean   :28.06  
 3rd Qu.:0.01556   3rd Qu.:0.9600   3rd Qu.:53.186   3rd Qu.:30.00  
 Max.   :0.04201   Max.   :1.0000   Max.   :91.810   Max.   :81.00  
```

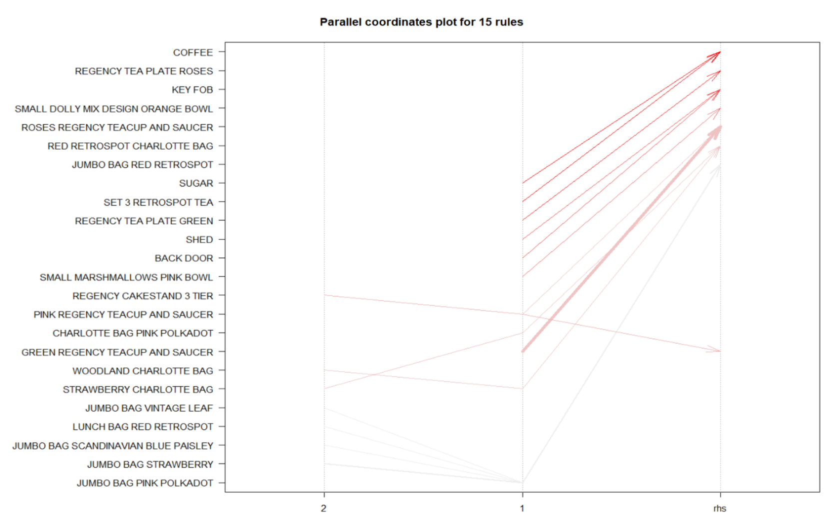
APPENDIX B: ASSOCIATION RULES FOR THE 5-CLUSTER ARRANGEMENT WITH PAM EUCLIDEAN METRIC

GROUP 1

	lhs		rhs	support	confidence	lift	count
	<fctr>	<fctr>	<fctr>	<dbl>	<dbl>	<dbl>	<dbl>
[1]	{REGENCY TEA PLATE GREEN}	=>	{REGENCY TEA PLATE ROSES}	0.01090959	0.8791209	57.049499	80
[2]	{SET 3 RETROSPOT TEA}	=>	{COFFEE}	0.01145507	1.0000000	61.621849	84
[3]	{SUGAR}	=>	{COFFEE}	0.01145507	1.0000000	61.621849	84
[4]	{BACK DOOR}	=>	{KEY FOB}	0.01336424	1.0000000	47.616883	98
[5]	{SHED}	=>	{KEY FOB}	0.01404609	1.0000000	47.616883	103
[6]	{SMALL MARSHMALLOWS PINK BOWL}	=>	{SMALL DOLLY MIX DESIGN ORANGE BOWL}	0.01240965	0.8198198	42.336188	91
[7]	{GREEN REGENCY TEACUP AND SAUCER}	=>	{ROSES REGENCY TEACUP AND SAUCER}	0.02577390	0.8008475	21.277588	189
[8]	{PINK REGENCY TEACUP AND SAUCER,REGENCY CAKESTAND 3 TIER}	=>	{GREEN REGENCY TEACUP AND SAUCER}	0.01077322	0.8404255	26.113731	79
[9]	{PINK REGENCY TEACUP AND SAUCER,REGENCY CAKESTAND 3 TIER}	=>	{ROSES REGENCY TEACUP AND SAUCER}	0.01090959	0.8510638	22.611779	80
[10]	{JUMBO BAG PINK POLKADOT,JUMBO BAG SCANDINAVIAN BLUE PAISLEY}	=>	{JUMBO BAG RED RETROSPOT}	0.01022774	0.8152174	8.360824	75

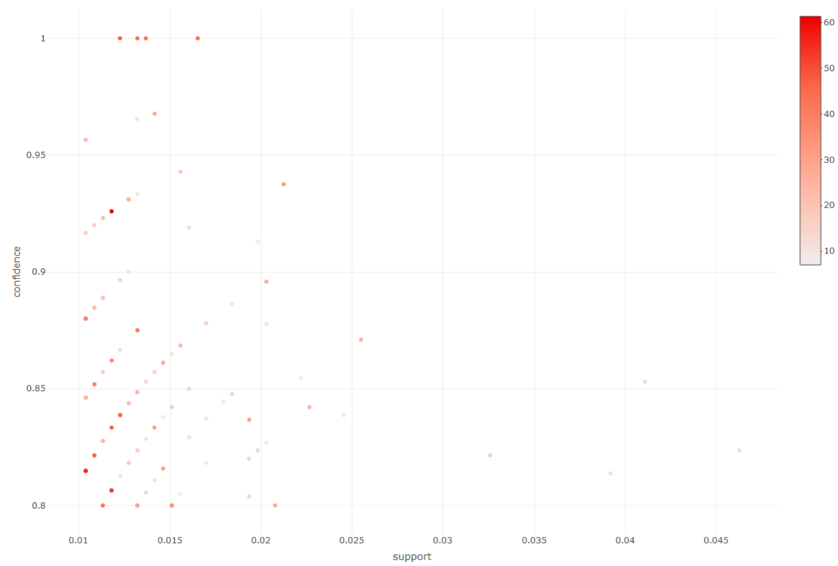


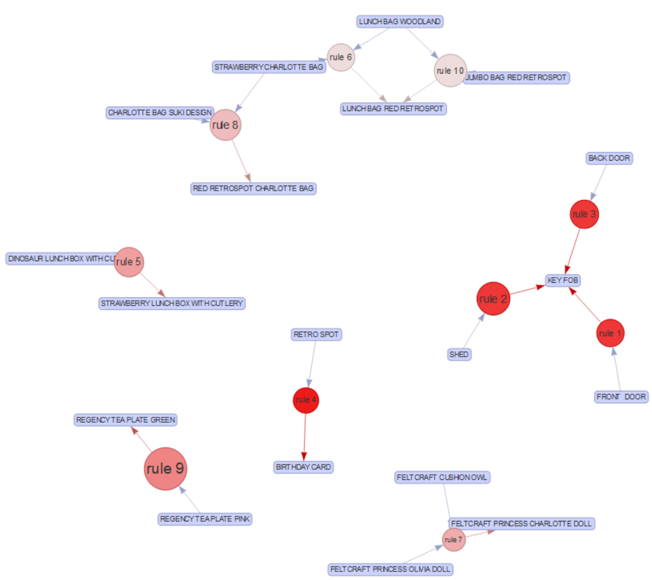


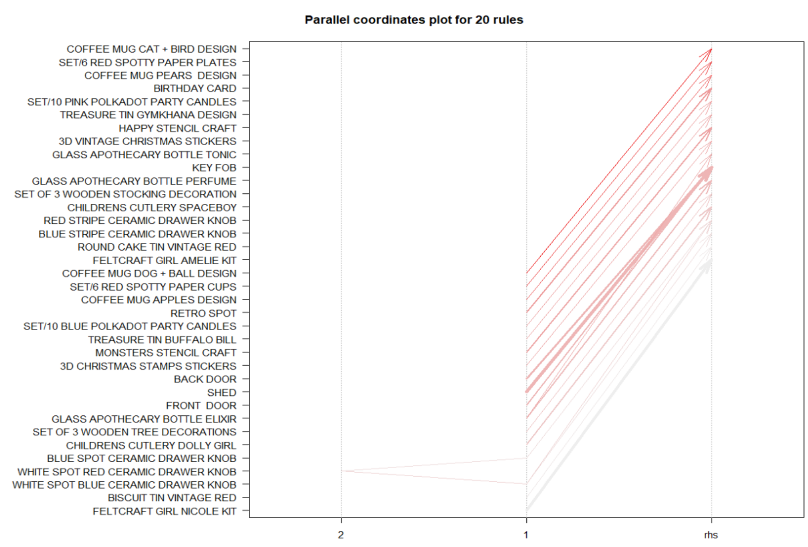


GROUP 2

	lhs		rhs	support	confidence	lift	count
	<fctr>	<fctr>	<fctr>	<dbl>	<dbl>	<dbl>	<dbl>
[1]	{BISCUIT TIN VINTAGE RED}	=>	{ROUND CAKE TIN VINTAGE RED}	0.01179802	0.8620690	36.53448	25
[2]	{COFFEE MUG DOG + BALL DESIGN}	=>	{COFFEE MUG CAT + BIRD DESIGN}	0.01179802	0.9259259	61.31366	25
[3]	{SET OF 6 SNACK LOAF BAKING CASES}	=>	{SET OF 12 MINI LOAF BAKING CASES}	0.01321378	0.8000000	30.27143	28
[4]	{COFFEE MUG APPLES DESIGN}	=>	{COFFEE MUG PEARS DESIGN}	0.01179802	0.8064516	53.40222	25
[5]	{CHILDRENS CUTLERY POLKADOT BLUE}	=>	{CHILDRENS CUTLERY POLKADOT PINK}	0.01510146	0.8000000	30.82182	32
[6]	{CHILDRENS CUTLERY DOLLY GIRL}	=>	{CHILDRENS CUTLERY SPACEBOY}	0.01321378	0.8750000	41.20278	28
[7]	{FRONT DOOR}	=>	{KEY FOB}	0.01321378	1.0000000	44.14583	28
[8]	{SET/10 BLUE POLKADOT PARTY CANDLES}	=>	{SET/10 PINK POLKADOT PARTY CANDLES}	0.01179802	0.8333333	47.72523	25
[9]	{TREASURE TIN BUFFALO BILL}	=>	{TREASURE TIN GYMKHANA DESIGN}	0.01085418	0.8214286	47.04344	23
[10]	{DINOSAUR LUNCH BOX WITH CUTLERY}	=>	{STRAWBERRY LUNCH BOX WITH CUTLERY}	0.01415762	0.9677419	26.98217	30

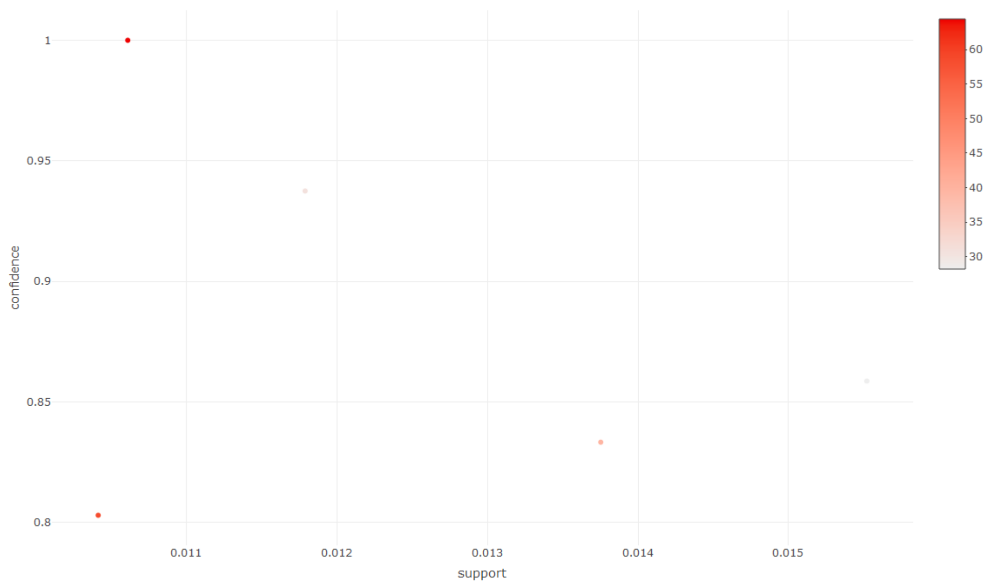


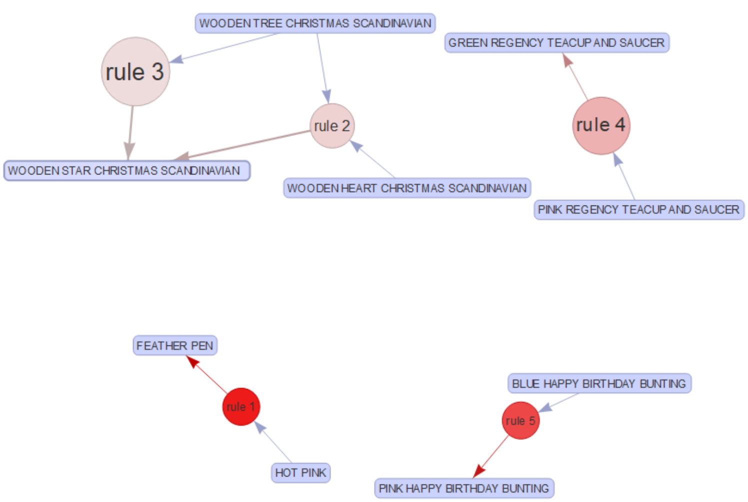


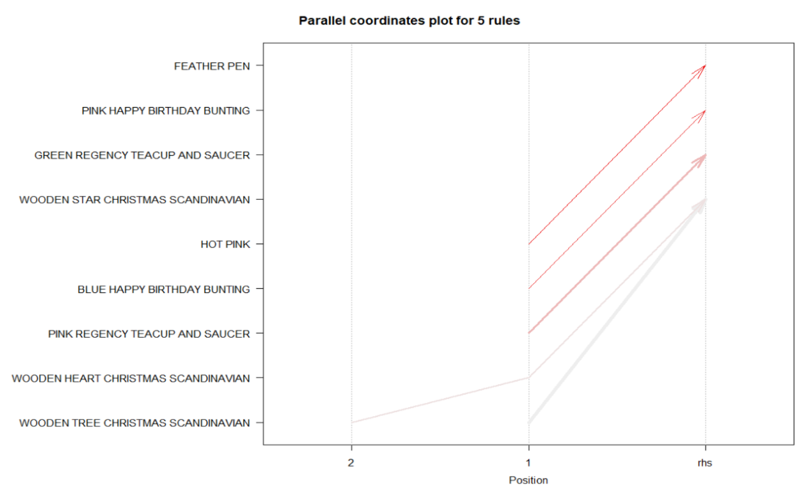


GROUP 3

```
    lhs                                    rhs                                 support confidence     lift count
[1] {BLUE HAPPY BIRTHDAY BUNTING}        => {PINK HAPPY BIRTHDAY BUNTING}       0.01041257  0.8030303 58.39177    53
[2] {HOT PINK}                           => {FEATHER PEN}                       0.01060904  1.0000000 64.43038    54
[3] {PINK REGENCY TEACUP AND SAUCER}     => {GREEN REGENCY TEACUP AND SAUCER}   0.01375246  0.8333333 39.64174    70
[4] {WOODEN TREE CHRISTMAS SCANDINAVIAN} => {WOODEN STAR CHRISTMAS SCANDINAVIAN} 0.01552063  0.8586957 28.19846    79
[5] {WOODEN HEART CHRISTMAS SCANDINAVIAN,
     WOODEN TREE CHRISTMAS SCANDINAVIAN} => {WOODEN STAR CHRISTMAS SCANDINAVIAN} 0.01178782  0.9375000 30.78629    60
```

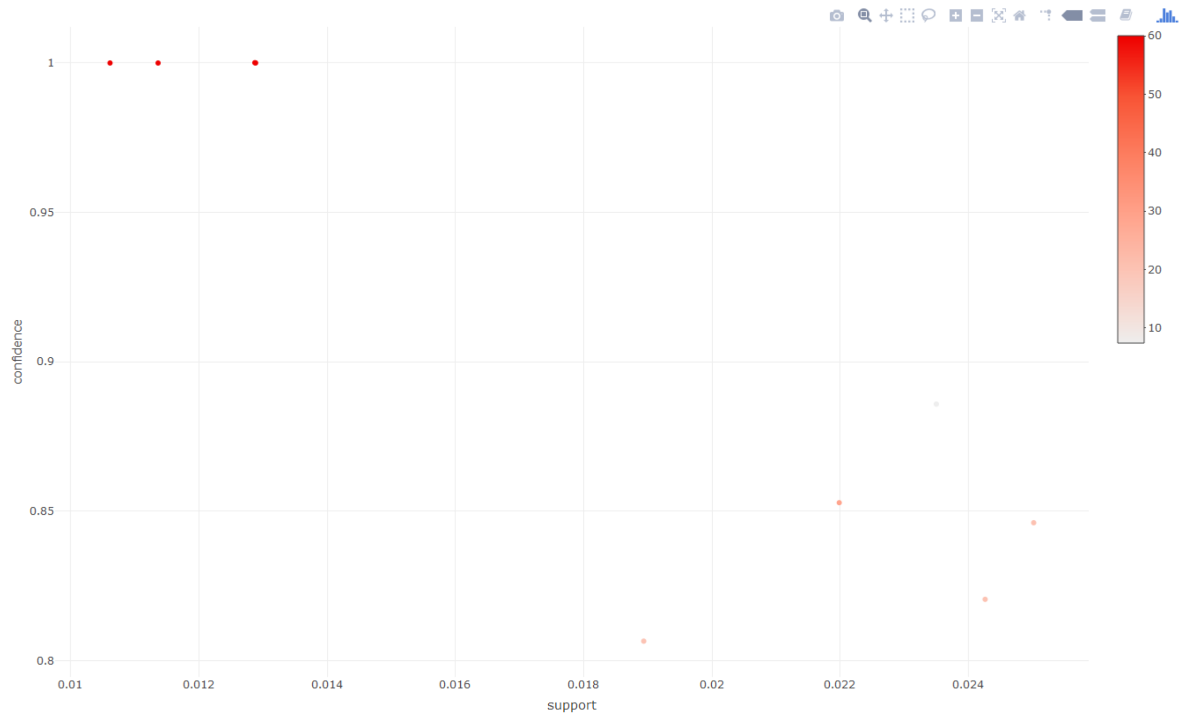


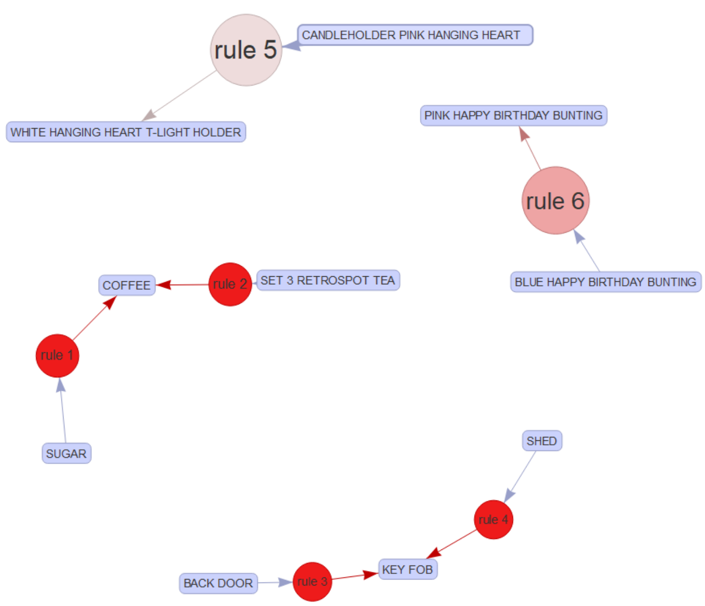


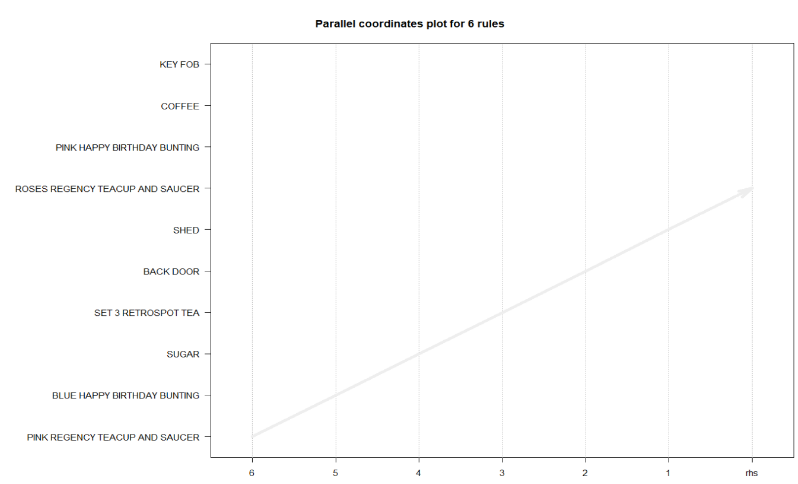


GROUP 4

	lhs <fctr>	<fctr>	rhs <fctr>	support <dbl>	confidence <dbl>	lift <dbl>	count <dbl>
[1]	{SUGAR}	=>	{COFFEE}	0.01287879	1.0000000	60.000000	17
[2]	{SET 3 RETROSPOT TEA}	=>	{COFFEE}	0.01287879	1.0000000	60.000000	17
[3]	{BACK DOOR}	=>	{KEY FOB}	0.01060606	1.0000000	60.000000	14
[4]	{SHED}	=>	{KEY FOB}	0.01136364	1.0000000	60.000000	15
[5]	{BLUE HAPPY BIRTHDAY BUNTING}	=>	{PINK HAPPY BIRTHDAY BUNTING}	0.02196970	0.8529412	28.868778	29
[6]	{PINK REGENCY TEACUP AND SAUCER}	=>	{ROSES REGENCY TEACUP AND SAUCER}	0.02500000	0.8461538	20.683761	33
[7]	{PINK REGENCY TEACUP AND SAUCER}	=>	{GREEN REGENCY TEACUP AND SAUCER}	0.02424242	0.8205128	20.435414	32
[8]	{CANDLEHOLDER PINK HANGING HEART}	=>	{WHITE HANGING HEART T-LIGHT HOLDER}	0.02348485	0.8857143	7.399638	31
[9]	{REGENCY CAKESTAND 3 TIER,ROSES REGENCY TEACUP AND SAUCER}	=>	{GREEN REGENCY TEACUP AND SAUCER}	0.01893939	0.8064516	20.085210	25







GROUP 5

	lhs		rhs	support	confidence	lift	count
	<fctr>	<fctr>	<fctr>	<dbl>	<dbl>	<dbl>	<dbl>
[1]	{WOODEN PICTURE FRAME WHITE FINISH}	=>	{WOODEN FRAME ANTIQUE WHITE}	0.01250000	0.8181818	26.77686	9
[2]	{COFFEE MUG PEARS DESIGN}	=>	{COFFEE MUG APPLES DESIGN}	0.01111111	0.8888889	45.71429	8
[3]	{LUNCH BAG DOLLY GIRL DESIGN}	=>	{LUNCH BAG SPACEBOY DESIGN}	0.01388889	0.8333333	25.00000	10
[4]	{KITCHEN METAL SIGN}	=>	{BATHROOM METAL SIGN}	0.01388889	1.0000000	48.00000	10
[5]	{SET 3 RETROSPOT TEA}	=>	{COFFEE}	0.01250000	1.0000000	55.38462	9
[6]	{SET 3 RETROSPOT TEA}	=>	{SET/5 RED RETROSPOT LID GLASS BOWLS}	0.01250000	1.0000000	30.00000	9
[7]	{SUGAR}	=>	{COFFEE}	0.01250000	1.0000000	55.38462	9
[8]	{SUGAR}	=>	{SET/5 RED RETROSPOT LID GLASS BOWLS}	0.01250000	1.0000000	30.00000	9
[9]	{ALPHABET STENCIL CRAFT}	=>	{HAPPY STENCIL CRAFT}	0.01111111	0.8888889	49.23077	8
[10]	{PINK REGENCY TEACUP AND SAUCER,REGENCY CAKESTAND 3 TIER}	=>	{GREEN REGENCY TEACUP AND SAUCER}	0.01666667	0.8571429	19.90783	12