BUSINESS CASES WITH DATA SCIENCE

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS – MAJOR IN BUSINESS ANALYTICS

Business Case 3 – Gift-a-Lot Recommender System

Group R

Andreia Bastos, number: 20210604

João Silva, number: 20211014

Pauline Richard, number: 20211019

Tiago Quaresma, number: 20210766

April, 2022

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

Index of Figures

1. INTRODUCTION

The aim of this project is to improve consumers' shopping experience on the company's website. The company plans to develop: Market Basket Analysis, which determines which items are substitutes and complementary, as well as the most common types of consumer behavior in the industry; a recommender system that would assist users in making better decisions by suggesting items that they might be interested in; and finally, a cold start that introduces new buyers to relevant products.
The analysis will be based on a dataset with 541 909 observations that occurred between December 1, 2010 and December 9, 2011. Hereunder is an overview of the phases involved in achieving the output:

- Business and Data Understanding: Developing intuition and understanding the data considering the business context.
- Data Preparation and Transformation: Investigate the need for cleaning and unveiling potential consumer behavior patterns; prepare the data to enable modeling.
- Evaluation of the Results: Modelling alternative techniques to maximize the capability to predict and select the best model.
- Deployment and Application to the Business: provide input on the deployment strategy to implement the model and quick overview of limitations and future steps for improvement.

2. BUSINESS AND DATA UNDERSTANDING

2.1. Business Context and Objectives

Gift-a-Lot is a UK-based non-store online retailer with about 80 members of staff. The company was established in 1981 mainly selling unique all-occasion gifts. For years in the past, the merchant relied heavily on direct mailing catalogs, and orders were taken by phone calls. It was only 2 years ago that the company launched its own website and shifted completely to the web. Since then, the company has maintained a steady and healthy number of customers from all parts of the United Kingdom and the world. The company also uses Amazon.co.uk to market and sell its products.

With the exponential expansion in the number of alternative choices accessible, particularly for online buying, recommender systems have become an indispensable tool for retail businesses, assisting customers in their purchasing and facilitating their decisions. These technologies can assist consumers in discovering products they might not have discovered otherwise and in making purchasing decisions. Customers tend to buy more when given easy alternatives, according to certain studies.

With this in consideration, the company hired our team to develop a recommender system to enhance the user experience while placing orders on the website and to recommend things that the customer enjoys in order to simplify user selections and increase sales. This system will be based on information gathered about customers by the company.

2.2. Data Description and Assessing Data Quality

As we previously indicated, we have been handed a dataset composed of 541 909 observations that occurred between December 1st, 2010 and December 9th, 2011, with the following variables:

- InvoiceNo: Invoice number. Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.

- StockCode: Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
- Description: Product (item) name. Nominal.
- Quantity: The quantities of each product (item) per transaction. Numeric.
- InvoiceDate: Invoice Date and time. Numeric, the day and time when each transaction was generated.
- UnitPrice: Unit price. Numeric, Product price per unit in pounds.
- CustomerID: Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
- Country: Country name. Nominal, the name of the country where each customer resides.

There were 25900 valid transactions in total during that time period, with 4070 distinct items and 4372 clients from 37 different countries. It's also important to note that a large number of the consumers are wholesalers.

2.3. Using Positive Transactions Only

Our dataset is showing the Quantity of the items that have been bought for each transaction. However, since we are working with data from an online retailer, we could observe that sometimes the transaction gets canceled and appears as a negative value in the 'Quantity' column. We want to proceed with items that the customer actually bought, so we are deciding to not use those values, and only focus on the positive ones. We applied the same logic for the 'UnitPrice'.

3. DATA PREPARATION AND TRANSFORMATION

3.1. Null Values and Duplicates

To perform any Machine Learning Model, we should first handle the Null Values. When checking the dataset, we could see there are only NaN values in 'Description' and 'CustomerID'. The latter one will not impact our analysis; however, we need to drop all rows with absence of values in Description, because without the name of the Product we cannot do any association. It represents a total of 1454 entries dropped.

We also identified 5268 duplicates, that we decided to aggregate (add the quantities) since we considered it to be different purchases from the same customer.

3.2. Coherence Check

We ran a coherence check to make sure that the data is logically consistent and can be reliably used for our analysis. First, we checked if all rows had an invoice number, which was the case and then, we dropped all of them containing a "c", since it indicates the transactions have been canceled. Also, there were several observations indicating gift transactions, which were also deleted.

Finally, we looked at the description and we drove to the conclusion that the product names were written in UPPER letter, while all descriptions in lower case were only comments about errors or a lack of information. We therefore decided to drop also all the rows that contain that kind of 'Description', to avoid any misleading or useless association analysis. We are also dropping administrative charges such as 'POSTAGE', 'CARRIAGES' and 'BANK CHARGES'.

3.3. Feature Engineering

In our Data Exploration, we saw that the whole data set contains 37 different countries, and that the UK represented 92.2% of the transactions. So, to make the dataset smaller and more accurate to the customer's characteristics, without having to do 37 different market basket analysis, we kept the 'United Kingdom' and change all the rest of the countries as 'Rest of the World'.

Moreover, we also split 'InvoiceDate' into three new variables: date, month and hour, so we can analyze better the seasonality and purchasing pattern.

Finally, we created a new variable 'TotalPrice', corresponding to the quantity multiplied by the price, so we can get the total amount paid per clients' basket.

At the end of our Data Preparation, in total 19439 rows have been removed, which means 96.419% of data has been kept.

3.4. Data Visualization

To get a better understanding of the data, we chose to manage it in different ways
We grouped the number invoice by country, we found that Australia and Brazil are the only ones in the southern hemisphere and the remainder are all European countries in the northern hemisphere. The United Kingdom accounts for 92.1% of all transactions (Figure 1), it's a significant volume that requires special attention.

Then we divided the items by invoice number to discover the Top 10 gifts that were the most popular in terms of more frequently bought. The one with the highest percentage was WHITE HANGING HEART T-LIGHT HOLDER representing 13.9% of all gifts. (Figure 2)

Also, using the new feature "Total Price", we allocated the total value of each order into classes to get an idea of what the regular price range is, each time a customer makes an order. The price range with the most invoices is between 200£ and 500£, representing 44.3% of the total number of invoices. (Figure 3)

Finally, we sorted the occurrences by date to see whether any patterns or seasonality existed. To apply this distribution, we split 'InvoiceDate' into date and hour. We could first conclude that more gifts are purchased during wintertime, most probably due to the Christmas period, with an increase of the InvoiceNo in October, November and December (Figure 4).

Looking at the hourly purchasing pattern, we could conclude that the purchases are made between 6 am and 8 pm and are more frequent in the middle of the day (between 12 and 3pm) (Figure 5). Finally, taking a look at the weekly purchases, we could see that most of them are made during weekdays, with Thursday at the top, and fewer during the end of the weekend, including none on Saturday. (Figure 6).

4. CUSTOMER BEHAVIOUR ANALYSIS

Before proceeding to the modeling, we performed an RFM analysis (Figure 7) to understand our customers' behavior. RFM analysis is a marketing technique used to quantitatively rank and group customers based on the recency, frequency and monetary total of their recent transactions to identify the best customers and perform targeted marketing campaigns, given this, we can see why it would be favorable to apply.

This method resulted in seven categories, from the best customers, **Can't Loose Them** (2531 customers), to the worst, **Require Activation** (35 customers). For the best segmentation, for instance, we see that the customers' purchasing expenses are 3118£ (on average) whereas in the worst it is around 143£. These segmentations also allow us to personalize the marketing approach: For the best customers treatment is unnecessary because they are highly engaged - often make purchases, do them a lot and spend a lot of money. However, tools for retaining them should be considered. The marketing strategy here could be to give them personalized offers (for instance, money coupons) based on their previous activity (purchases). Additional discounts may not be needed for this group due to their high involvement already. For the worst, however, a cost-effective treatment can be a company newsletter (with new in-stock products, new services, etc.)

5. RESULTS OF THE MODEL

5.1. Market Basket Analysis

As seen in part 3, we are dividing the basket analysis in two, one for the UK products and one for the rest of the world. The Basket Data will contain the quantity for each item bought per InvoiceNo.

Construct the Basket data
We are grouping the data by the transaction and the items ('InvoiceNo' and 'Description') and showing the quantity bought. To reshape the dataframe, we sum up the values and unstacked them and we are setting the InvoiceNo as the index. Therefore, we can see the number of items bought but per transaction and customer. We are basically picturing the basket that each customer paid at the end. After, we need to display the items into one transaction per row with each product bought (1) or not bought (0), using 1 hot encoded.

Filter the Transactions
After we structured the data, we want to filter the transaction with at least 2 or more items bought, so we can work on their association. We generated frequent items, using the Apriori Algorithm with a support of 3%, to have enough output to work on. We added the column 'length', that calculates the number of items bought.

Comparing the results from UK and the Rest of the World in , we could see that 177 and 61 transactions, respectively, are considered as frequently bought items. For the UK, it seems WHITE HANGING HEART T-LIGHT HOLDER is the most frequently bought items with the support value of 0.121445$, meaning it has been bought 2 325 times out of the whole transaction.

Apply the Association rule
Finally, we can now apply the association rules to extract the information of which items are better to be sold together. In order to determine this we have 3 metrics to consider:

- **Support:** It shows the relative frequency at which the rule occurs, meaning the higher the better because it represents the most useful relationship. Low support can also help to find hidden relationships.
- **Lift**: it is good to have a high lift, because it means the item occurs more often than expected (looking at its ratio between the observed support if the two rules were independent). It helps us to know how accurate our predictions are.
- **Confidence**: it is good to have a high confidence value, it shows the reliability of the rule. However, for this last one it is important to also understand the business environment.

Results

From the result of the UK association rules, we could see that PINK REGENCY TEACUP AND SAUCER and GREEN REGENCY TEACUP AND SAUCER have the highest lift (14,609), meaning they have the highest association, and they are very good to be sold together. Knowing that a lift with a value over 1 is considered enough to determine there is an association between 2 items. For the Rest of the world, the best association is different, it is between the SPACEBOY LUNCH BOX and DOLLY GIRL BOX with a lift of 6,597.

Then, looking at the support value, we can see the value for the best association in UK is 0,0349 and for the Rest of the World, 0,0632. It means that 3,49% and 6,32%, respectively, of those 2 items have been sold together, out of all the transactions.

Finally, for the confidence value, it is impacted by the antecedents and consequents. For the UK scenario, the consequent is higher than the antecedent (0,618 < 0,825), so we have to apply the rule number 2, which is PINK REGENCY TEACUP AND SAUCER --> GREEN REGENCY TEACUP AND SAUCER. It means that clients tend to buy the pink regency teacup and saucer before the green, so in terms of discount, it would be more efficient for example to give a discount on the Green if they are buying the Rose (to encourage them to buy the green). For the Rest of the World, it is the other around, the antecedent is higher than the consequent (0,742 > 0,562), so we apply rule number 1, which is DOLLY GIRL LUNCH BOX --> SPACECBOY LUNCH BOX. Presented on the Figure 8 and 9, are the rules with confidence greater or equal to 0.7 for the UK and Rest of the World, respectively.

5.2. Recommender System

For the regular customers we chose to implement the Alternating Least Squares, a matrix factorization algorithm, since it’s simple and scales well with large dataset.

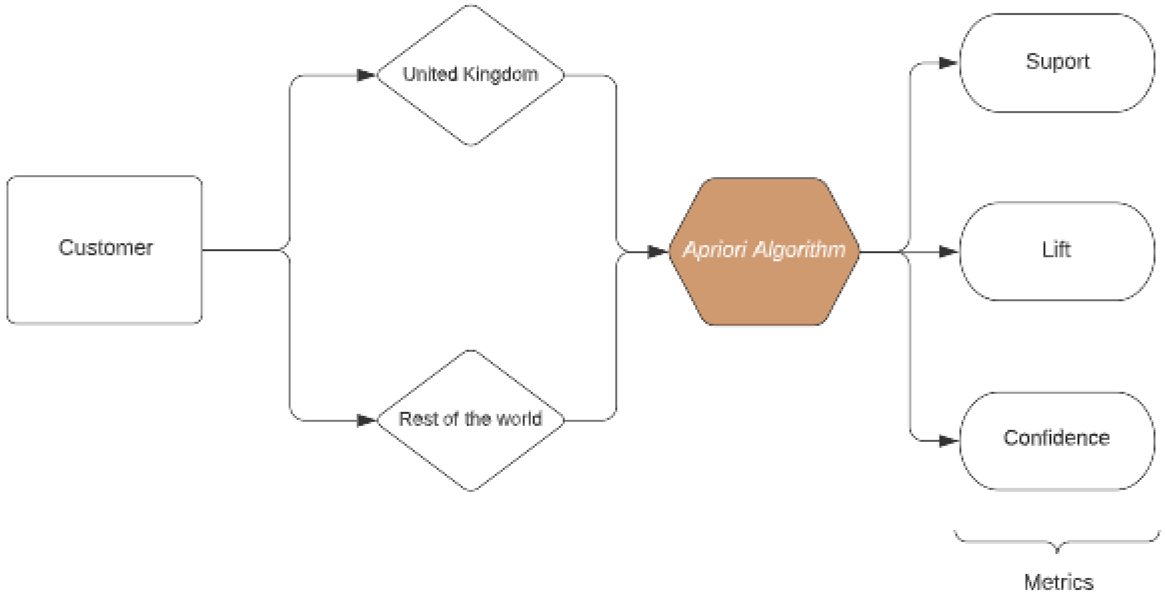
Before creating the recommender system, we must deal with another issue: Data Sparsity. In order to apply the ALS method, we must start by creating a user-item matrix. This matrix will represent the interactions between the CustomerID and the StockCode, meaning the number of times a user purchased a certain item. But since most of the matrix is represented by zeros, we are facing data sparsity. To solve this, we set up a threshold for the minimum user-item interactions: customers who have less than five purchases or products purchased less than five times would be considered. To implement the Alternating Least Squares algorithm we created two functions: **implicit_als** and **recommend**. The first function computes the algorithm with implicit data (with the) whereas the second recommends items for a given user given a trained model. By joining these two functions we created **rec_sys**, that receives a Customer ID as input and returns product recommendations (using the Stock Code) to the client.

5.2. Cold Start

For the regular customers we chose to implement the Alternating Least Squares, a matrix factorization algorithm, since it's simple and scales well with large dataset.

Before creating the recommender system, we must deal with another issue: Data Sparsity. In order to apply the ALS method, we must start by creating a user-item matrix. This matrix will represent the interactions between the CustomerID and the StockCode, meaning the number of times a user purchased a certain item. But since most of the matrix is represented by zeros, we are facing data sparsity. To solve this, we set up a threshold for the minimum user-item interactions: customers who have less than five purchases or products purchased less than five times would be considered. To implement the Alternating Least Squares algorithm we created two functions: implicit_als and recommend. The first function computes the algorithm with implicit data (with the alpha value set to 30) whereas the second recommends items for a given user given a trained model. By joining these two functions we created rec_sys, that receives a Customer ID as input and returns product recommendations (using the Stock Code) to the client.

The following organigram shows the visually our approach:



6. DEPLOYMENT AND APPLICATION TO THE BUSINESS

From the Market Basket Analysis and association Rules we can perform some data-driven marketing strategy and insights for the business. Those recommendations are focused on the UK market since it's the biggest, but the same can be applied for the Rest of the World.

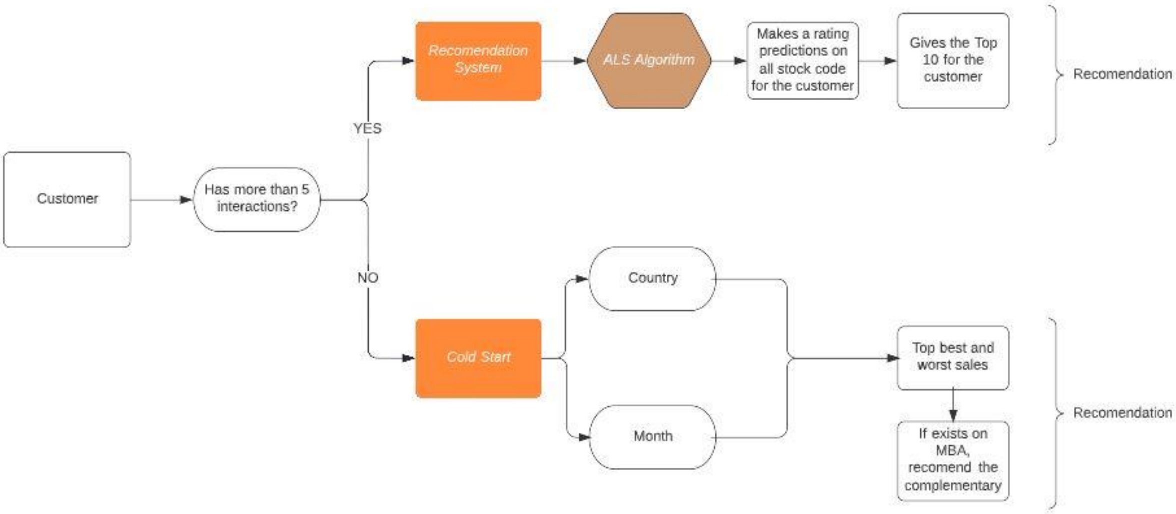
- First, we could focus on doing **Discount and Recommendations** on the on the consequent items. Therefore, we could use the success of one product, and increase the purchase of other items less bought but highly associated. For example, we saw in the visualization of the data that the JUMBO BAG RED RETROSPOT is one of the Top 10 gifts sold (representing 12.9% of the sales). In the Association Rules results we can see it is associated with the

JUMBO BAG PINK RETROSPOT. Knowing this, we could add a pop-up recommendation or discounts on the PINK BAG when the JUMBO BAG RED is selected.

- Secondly, we could also regroup Items together into a **Grouping Gifts,** with a lower price than if the two items were bought individually at the same time. It will be a good incentive for clients that were hesitating to buy both, it will generate more sales and therefore more profit.
- Lastly, we could use the Association Rules to **display items** that are highly associated next to each other on the website.

Those three recommendations all have the same aim: to attract the attention of the customer on the consequent items, increase the sales and therefore the profit.

Regarding the deployment of the recommendation system, we proceeded as follows:
Combining the two functions, **cold_start** and **rec_sys**, we created our recommendation system, **recommender_system**. The function receives as input the customer ID, the customers' country and the current month. Then it proceeds as follows: if it is a new user or an uncommon one (less than five purchases) the function will return the cold start approach. If, however, we are dealing with a regular customer (five or more purchases) then it will return the ALS method. So, as we can see, it is a very simple system, easy to implement and maintain.


Customer
Has more than 5 interactions?
YES
Recomendation System
ALS Algorithm
Makes a rating predictions on all stock code for the customer
Gives the Top 10 for the customer
Recomendation
NO
Cold Start
Country
Month
Top best and worst sales
If exists on MBA, recomend the complementary
Recomendation

7. CONCLUSION

It's difficult to create an effective recommendation system. The best recommender systems are created when a large amount of high-quality data, including both explicit and implicit information, is accessible.

When looking at some other types of recommender systems, we stumbled upon good alternatives to our final model that, however, required explicit data such as customer rating, and this was a big drawback in our implementation.

While the study provides an overview of how simple data based almost entirely on transactions can be used to develop recommender systems, the findings reveal that the collaborative-filtering strategy offered as a final output has limited capabilities. As a result, we suggest the company to use explicit data to create a more reliable recommendation system in the future.

8. REFRENCES

- Macmanus, R., & Web. (2009, January 28). 5 problems of Recommender Systems. ReadWrite. Retrieved April 18, 2022, from https://readwrite.com/5_problems_of_recommender_systems/
- Milankovich, M. (2017, July 5). The cold start problem for Recommender Systems. Medium. Retrieved April 18, 2022, from https://medium.com/@markmilankovich/the-cold-start-problem-for-recommender-systems-89a76505a7
- Rosenthal, E. (n.d.). Intro to implicit matrix factorization: Classic ALS with Sketchfab Models: Ethan Rosenthal. Intro to Implicit Matrix Factorization: Classic ALS with Sketchfab Models | Ethan Rosenthal. Retrieved April 18, 2022, from https://www.ethanrosenthal.com/2016/10/19/implicit-mf-part-1/
- Victor. (2018, July 10). Als implicit collaborative filtering. Medium. Retrieved April 18, 2022, from https://medium.com/radon-dev/als-implicit-collaborative-filtering-5ed653ba39fe

9. APPENDIX

Figure 1: Country Performance by Number of Invoice

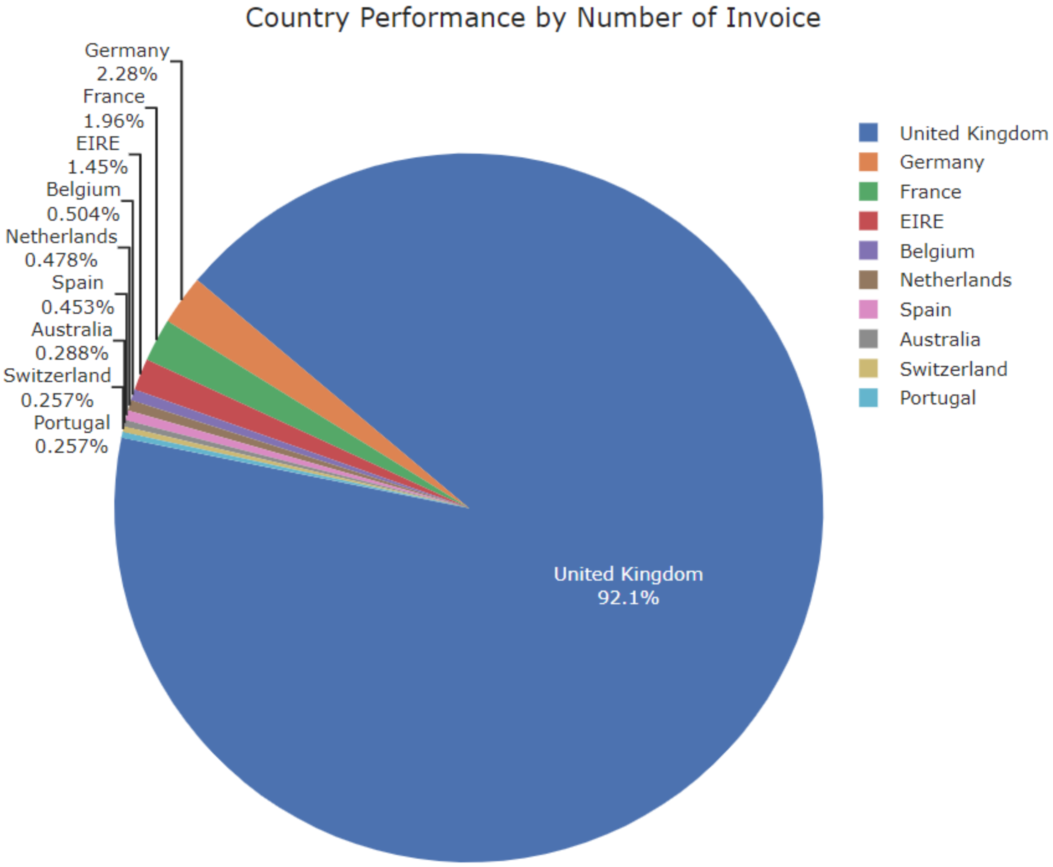


Figure 2: Distribution of Products by Number of Invoice

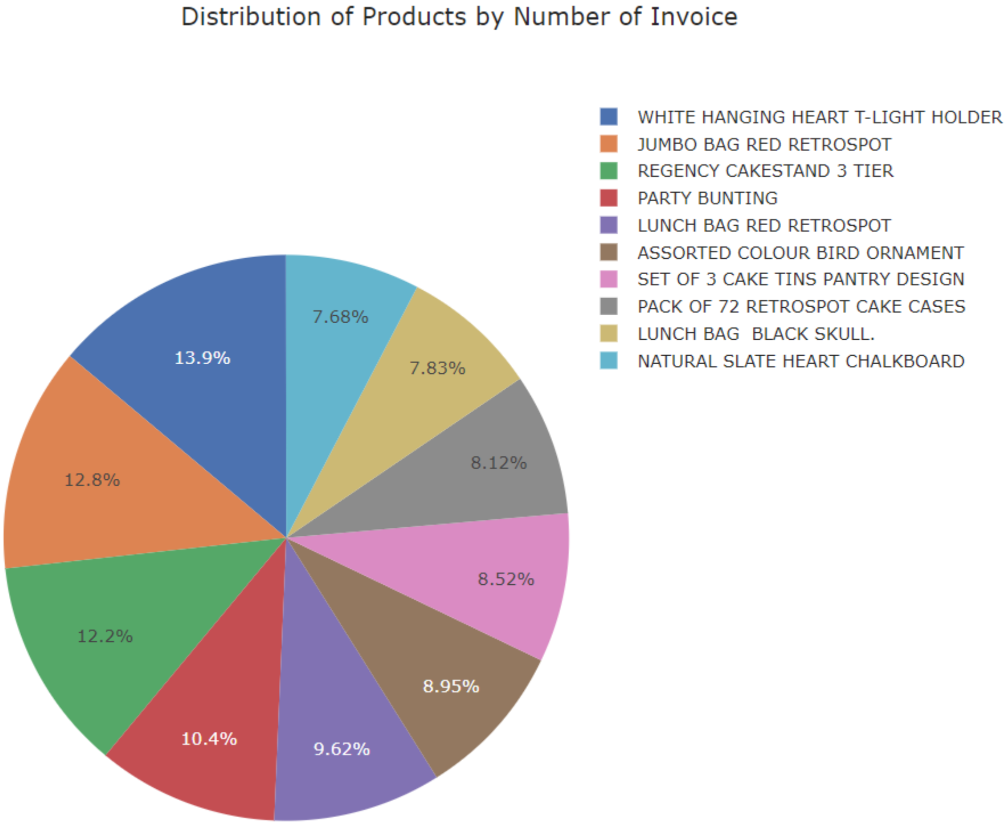


FIGURE 3: Distribution of orders total price

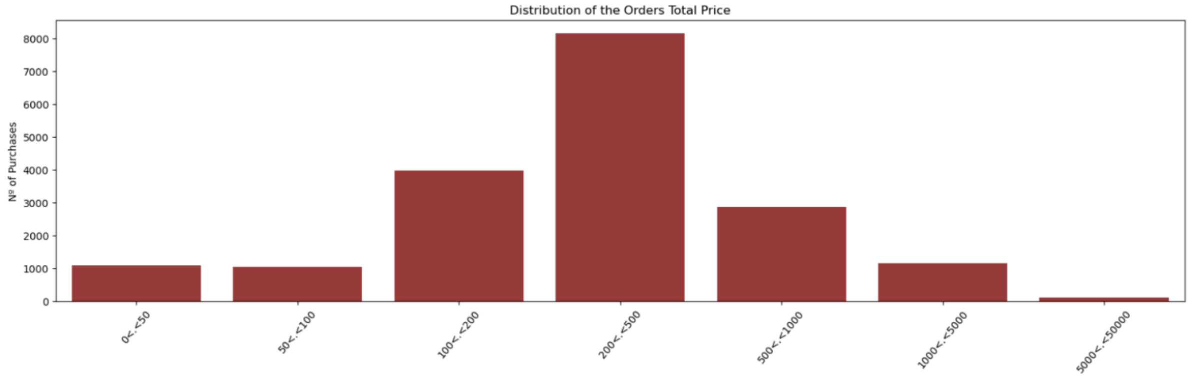


FIGURE 4: Distribution of events per date

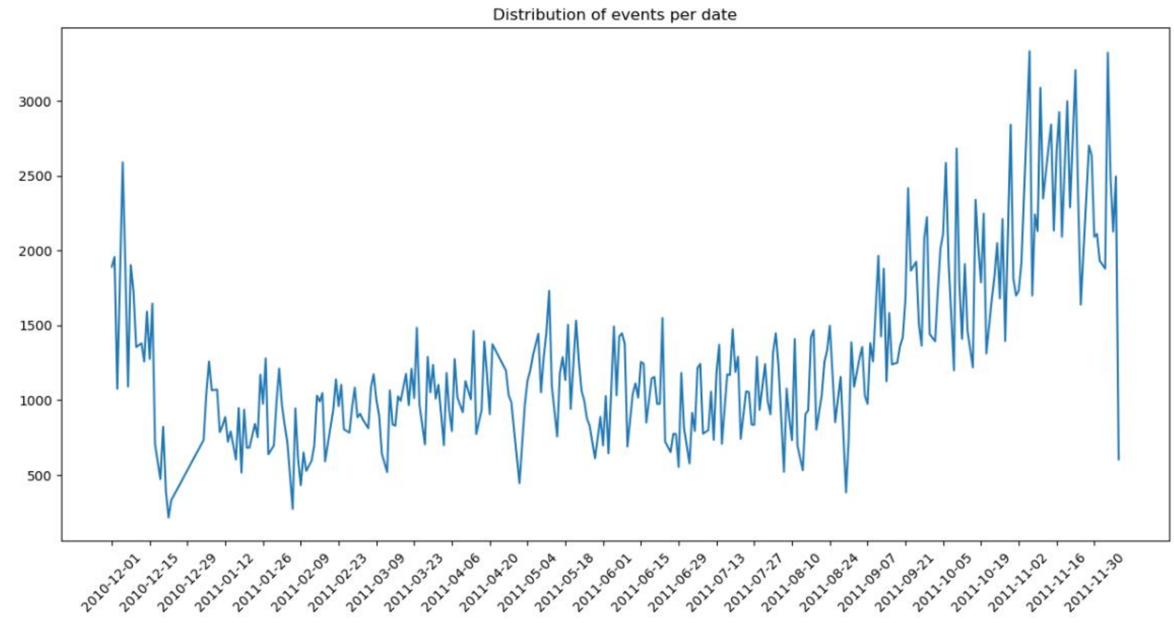


FIGURE 5: Most Frequent hour of the day

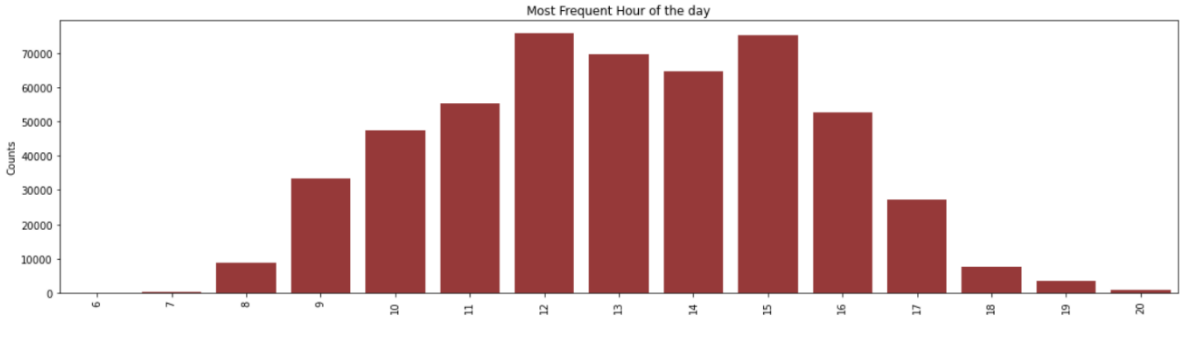


FIGURE 6: Most Frequent day of the week

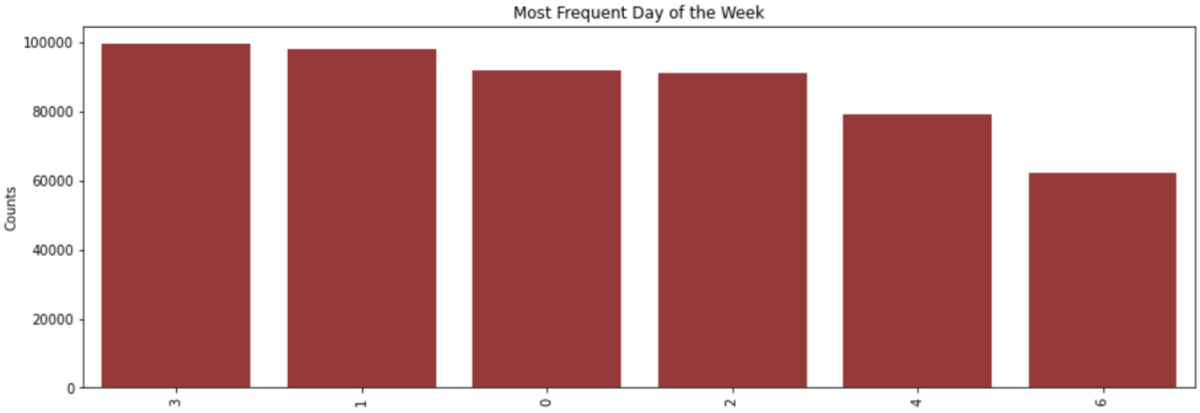


FIGURE 7: RFM analysis

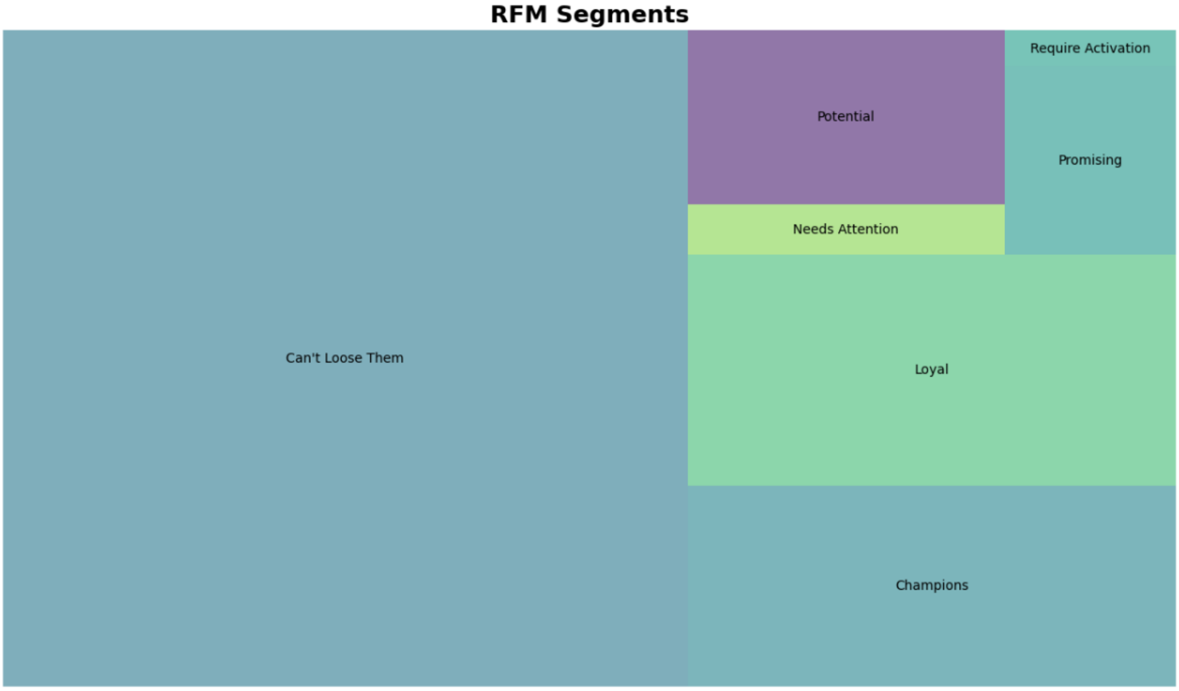
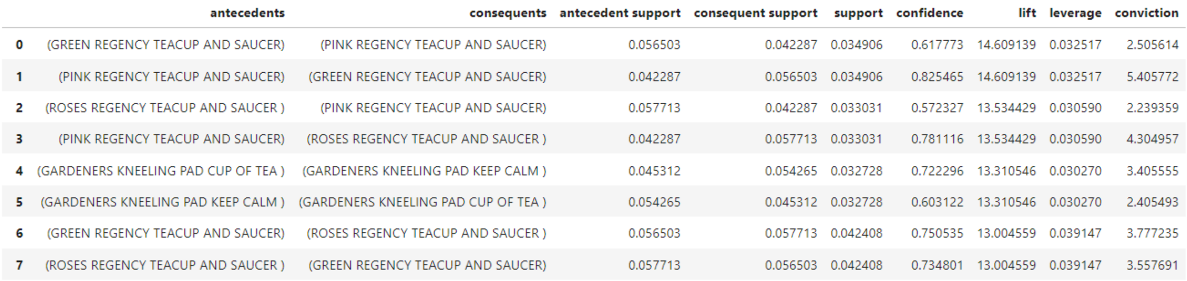
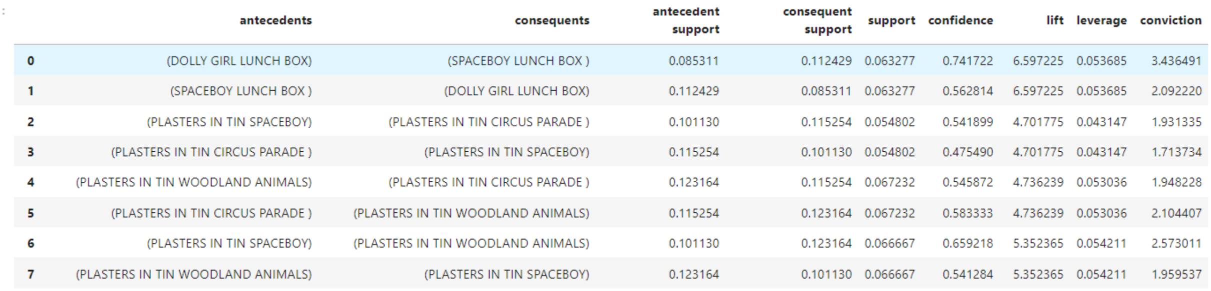


FIGURE 8: UK Market Basket rules



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.056503	0.042287	0.034906	0.617773	14.609139	0.032517	2.505614
1	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.042287	0.056503	0.034906	0.825465	14.609139	0.032517	5.405772
2	(ROSES REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.057713	0.042287	0.033031	0.572327	13.534429	0.030590	2.239359
3	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.042287	0.057713	0.033031	0.781116	13.534429	0.030590	4.304957
4	(GARDENERS KNEELING PAD CUP OF TEA)	(GARDENERS KNEELING PAD KEEP CALM)	0.045312	0.054265	0.032728	0.722296	13.310546	0.030270	3.405555
5	(GARDENERS KNEELING PAD KEEP CALM)	(GARDENERS KNEELING PAD CUP OF TEA)	0.054265	0.045312	0.032728	0.603122	13.310546	0.030270	2.405493
6	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.056503	0.057713	0.042408	0.750535	13.004559	0.039147	3.777235
7	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.057713	0.056503	0.042408	0.734801	13.004559	0.039147	3.557691

FIGURE 9: Rest of the World Market Basket rules



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(DOLLY GIRL LUNCH BOX)	(SPACEBOY LUNCH BOX)	0.085311	0.112429	0.063277	0.741722	6.597225	0.053685	3.436491
1	(SPACEBOY LUNCH BOX)	(DOLLY GIRL LUNCH BOX)	0.112429	0.085311	0.063277	0.562814	6.597225	0.053685	2.092220
2	(PLASTERS IN TIN SPACEBOY)	(PLASTERS IN TIN CIRCUS PARADE)	0.101130	0.115254	0.054802	0.541899	4.701775	0.043147	1.931335
3	(PLASTERS IN TIN CIRCUS PARADE)	(PLASTERS IN TIN SPACEBOY)	0.115254	0.101130	0.054802	0.475490	4.701775	0.043147	1.713734
4	(PLASTERS IN TIN WOODLAND ANIMALS)	(PLASTERS IN TIN CIRCUS PARADE)	0.123164	0.115254	0.067232	0.545872	4.736239	0.053036	1.948228
5	(PLASTERS IN TIN CIRCUS PARADE)	(PLASTERS IN TIN WOODLAND ANIMALS)	0.115254	0.123164	0.067232	0.583333	4.736239	0.053036	2.104407
6	(PLASTERS IN TIN SPACEBOY)	(PLASTERS IN TIN WOODLAND ANIMALS)	0.101130	0.123164	0.066667	0.659218	5.352365	0.054211	2.573011
7	(PLASTERS IN TIN WOODLAND ANIMALS)	(PLASTERS IN TIN SPACEBOY)	0.123164	0.101130	0.066667	0.541284	5.352365	0.054211	1.959537