MARKET BASKET ANALYSIS:

Introduction to Market Basket Analysis

Introduction to Market Basket Analysis

Market basket analysis is one of the key applications of machine learning in retail. It analyses the past buying patterns of customers to determine which products they frequently purchase together.

large retailers like Amazon, Flipkart, etc use Market basket analysis to analyse customer buying behaviors to find associations between the different items that customers place in their "shopping baskets". By acquiring insight into which items are frequently purchased together by customers, retailers can design marketing strategies based on the finding of these associations/rules.

2.2 Understanding our Dataset

Our dataset contains all transnational data of transactions that occurred between 01/12/2010 and 09/12/2011 for a UK-based non-store online retailer that sells unique all-occasion gifts and many of the company's customers are wholesalers.

Our dataset contains 24391 transactions (This includes transactions from all the countries including cancelled transactions and might require pre-processing).

Data Cleaning:

Missing Values:
It is evident from the below output that our dataset has some missing values. We are missing 1521 values for Description which is 0.28% of the total data and about 24% in the column CustomerID.
Since we don’t require CustomerID information for market basket analysis, it is better to drop the entire column and as the missing values in the Description are not huge, we'll drop them as well.

Duplicates:
There are 5192 duplicate rows in our dataset. We’ll drop them using drop_duplicates() function.

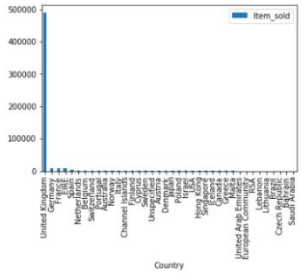
Dropping the Cancelled Transactions:

As this is a non-store online retail dataset, we have some cancelled transactions. For cancelled transactions the value in the quantity column is negative. For our analysis, we are only interested in the items bought.

We have 20130 out of 24391 that is not cancelled.

Top Three Countries – The UK, Germany and France:

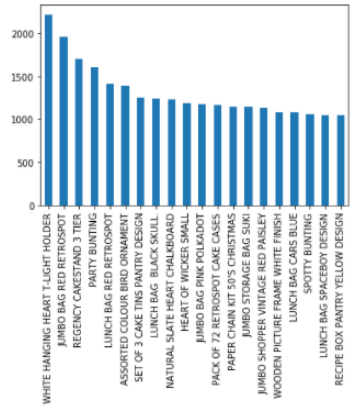
It is evident from the below output that The UK, Germany and France are our Top three countries. The UK has total 18188 Transactions out of 24391. Germany and France have 457 and 392 transactions respectively.



UK BASKET

Creating A Basket the UK:

Top selling items in the UK:



From The above output, it is evident that WHITE HANGING HEART T-LIGHT HOLDER , JUMBO BAG RED RETROSPOT, and EGENCY CAKESTAND 3 TIER are the top 3 sold items in the UK. However, we are interested in the top items sold together – top association rules.

Applying Apriori Algorithm

Steps Followed:

Step1: Creating a basket for the UK

Selecting all the transactions where country is United Kingdom.

UK_basket = (data[data['Country']== 'United Kingdom'])

Step2: Encoding

Transform the data into a sparse matrix format i.e., product description will be column and InoiceNo as index and then encode our basket into a binary dataframe – if an item is bought in a transaction replace it with 1 else with 0.

Step3: Selecting only the transactions which have at least two items

#Considering only the transactions having more or two item

UK_basket_atleast_2items = UK_basket[(UK_basket>0).sum(axis =1)>=2]

Step4: **Applying Apriori**

The algorithm finds the most frequent itemsets and relevant association rules.

As the UK basket have a large no. of transactions, let us consider minimum support as 3%, which means a product should appear in at least 3% of the total transactions.

Association Rules in the UK:

Total number of Rules in the UK basket with minimum support: 0.03 and minimum lift: 1:

```
#with minimum support: 0.03 and minimum lift: 1, there are 48 rules
rules = association_rules(UK_frequent_itemsets, metric="lift", min_threshold=1).sort_values('lift', ascending = False).reset_ind
print('there are ',len(rules),'rules for the UK basket')
```

```
there are  48 rules for the UK basket
```

```
rules
```

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.042264	0.056473	0.034887	0.825465	14.617093	0.032500	5.405948
1	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.056473	0.042264	0.034887	0.617773	14.617093	0.032500	2.505674
2	(PINK REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.042264	0.057682	0.033013	0.781116	13.541798	0.030575	4.305101
3	(ROSES REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.057682	0.042264	0.033013	0.572327	13.541798	0.030575	2.239413
4	(GARDENERS KNEELING PAD KEEP CALM)	(GARDENERS KNEELING PAD CUP OF TEA)	0.054235	0.045287	0.032711	0.603122	13.317793	0.030254	2.405555
5	(GARDENERS KNEELING PAD CUP OF TEA)	(GARDENERS KNEELING PAD KEEP CALM)	0.045287	0.054235	0.032711	0.722296	13.317793	0.030254	3.405662
6	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.057682	0.056473	0.042385	0.734801	13.011639	0.039127	3.557807
7	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.056473	0.057682	0.042385	0.750535	13.011639	0.039127	3.777361
8	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.052663	0.056170	0.034041	0.646383	11.507574	0.031083	2.669077

Reading the first rule:

(PINK REGENCY TEACUP AND SAUCER) -------> (GREEN REGENCY TEACUP AND SAUCER)

Lift: As we can see from the above rules, (PINK REGENCY TEACUP AND SAUCER) and (GREEN REGENCY TEACUP AND SAUCER) has the highest lift value, hence these two items are highly associated with each other. The higher the lift value, the higher the associations between the items. Any lift value greater than 1 is worth considering, in this case, it is 14.6 which is very high, It means these two items are very good to be sold together.

Support:
Antecedents((PINK REGENCY TEACUP AND SAUCER) is present in 4.22% or 929 times of the total UK transactions.
Consequents(GREEN REGENCY TEACUP AND SAUCER) – present in 5.64% or 1239 times the total UK transactions.
And both the items are present in 3.48% or 766 times the total UK transactions.

confidence: This measure of the likelihood of occurrence of consequent on a cart given that the cart already has the antecedents. So, in this case, there are 83% chances of finding GREEN REGENCY TEACUP AND SAUCER given PINK REGENCY TEACUP AND SAUCER is already in the basket.

Selecting the rules:

There are total 48 rules for the UK basket. However, if we see the above output there are two rules for the same itemsets.
For example, let us take Rule 0 and Rule 1. Both the rules have the same items that are, PINK REGENCY TEACUP AND SAUCER and GREEN REGENCY TEACUP AND SAUCER.

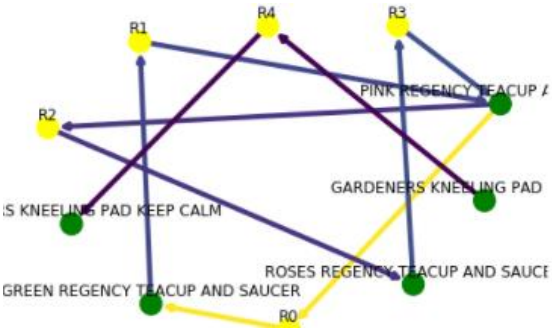
GREEN REGENCY TEACUP AND SAUCER is there in 1018 times (5.6%) of the total UK transactions and PINK REGENCY TEACUP is there 763 times (4.2%) and both the items are there in 618 (3.4%) of the total transitions.

This is a very valuable piece of information, as now we know which item to put on discount and how to increase the sale of PINK REGENCY TEACUP. In this case, we could give a discount on PINK REGENCY TEACUP AND SAUCER after the customer has bought GREEN REGENCY TEACUP AND SAUCER.

It is very important for marketing and business strategies to choose the Right Rule.
Hence, we will select the rules where **Antecedent Support > Consequent Support**
So, The Top 5 Final Association Rules for the UK:

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.056473	0.042264	0.034887	0.617773	14.617093	0.032500	2.505674
(ROSES REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.057682	0.042264	0.033013	0.572327	13.541798	0.030575	2.239413
(GARDENERS KNEELING PAD KEEP CALM)	(GARDENERS KNEELING PAD CUP OF TEA)	0.054235	0.045287	0.032711	0.603122	13.317793	0.030254	2.405555
(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.057682	0.056473	0.042385	0.734801	13.011639	0.039127	3.557807
(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.056170	0.052663	0.034041	0.606028	11.507574	0.031083	2.404578

Visual representation the top 5 rules in the UK:



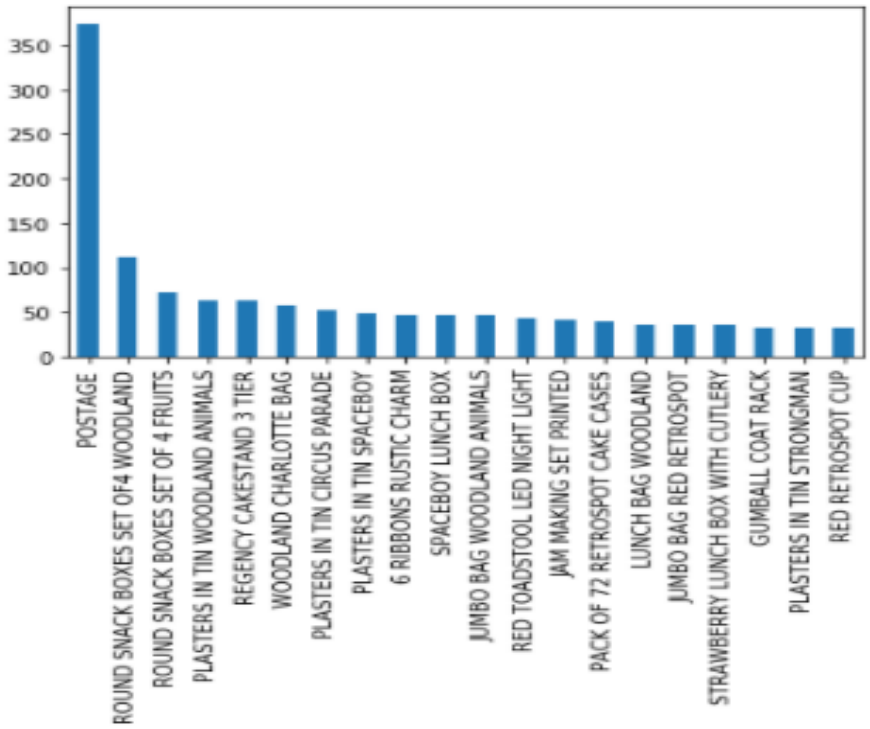
R0,R1,R2,R3,R4 – are the rules(yellow dots) and the arrow coming to the rules is from antecedents and the arrows going from the rules circle are towards consequents.

Germany Basket:

Association Rules for Germany:

No. of transactions in Germany: 457
And 427 out of 457 transactions has at least 2 items.

Top Items sold in Germany:
From the below output we see can that 'Postage' is the most sold item. However, 'Postage' is not any item but is a postage or delivery charge. So, we'll delete it from all the transactions as it might affect our association rules.

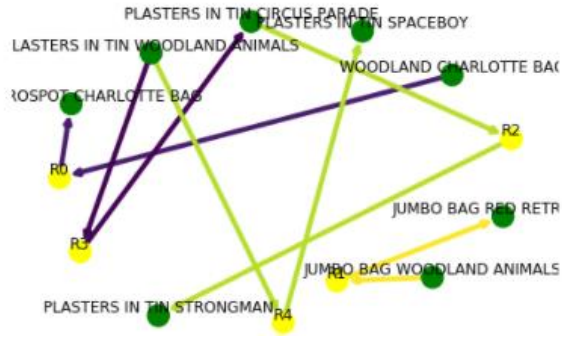


Minimum Support: 0.05 or 5%
Total No. of Rules: 30 (support > 0.05 and lift>1)
Total No. of Final Rules: 15 (antecedent support > consequent support)

Top 5 Final Rules in Germany:

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(WOODLAND CHARLOTTE BAG)	(RED RETROSPOT CHARLOTTE BAG)	0.135831	0.074941	0.063232	0.465517	6.211746	0.053052	1.730755
(RED RETROSPOT CHARLOTTE BAG)	(WOODLAND CHARLOTTE BAG)	0.074941	0.135831	0.063232	0.843750	6.211746	0.053052	5.530679
(JUMBO BAG WOODLAND ANIMALS)	(JUMBO BAG RED RETROSPOT)	0.105386	0.081967	0.051522	0.488889	5.964444	0.042884	1.796151
(JUMBO BAG RED RETROSPOT)	(JUMBO BAG WOODLAND ANIMALS)	0.081967	0.105386	0.051522	0.628571	5.964444	0.042884	2.408575
(PLASTERS IN TIN STRONGMAN)	(PLASTERS IN TIN CIRCUS PARADE)	0.074941	0.124122	0.051522	0.687500	5.538915	0.042220	2.802810

Visual representation of the top 5 rules in Germany:



From the above output, we can see that WOODLAND CHARLOTTE BAG and RED RETROSPOT CHARLOTTE BAG have a high association with a lift value of 6.21 and are good to be sold together.

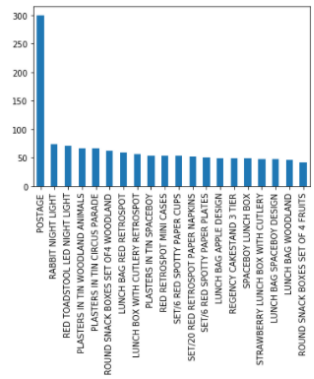
France Basket:

Association Rules for France:

Top items sold in France:

```
France_basket['Description'].value_counts()
France_basket['Description'].value_counts()[:20].plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x17af9ffeda0>
```

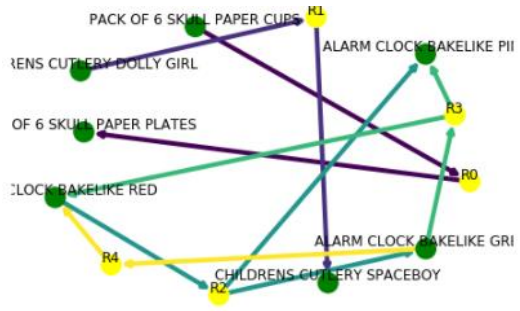


Similarly for France, we will delete 'Postage' from all the transactions as it is the postage charge.

No. of transactions in France: 392
No. of Transactions in France having two or more items: 366
Minimum Support: 0.07 or 7%
Total No. of Rules: 36 (support > 0.07 and lift>1)
Final No. of Rules to be considered: 18 (where antecedent support > consequent support)

antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(PACK OF 6 SKULL PAPER CUPS)	(PACK OF 6 SKULL PAPER PLATES)	0.068306	0.060109	0.054645	0.800000	13.309091	0.050539	4.699454
(CHILDRENS CUTLERY DOLLY GIRL)	(CHILDRENS CUTLERY SPACEBOY)	0.076503	0.073770	0.068306	0.892857	12.103175	0.062662	8.644809
(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELI...	0.101093	0.079235	0.068306	0.675676	8.527493	0.060296	2.839026
(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED, ALARM CLOCK BAKELIK...	0.103825	0.079235	0.068306	0.657895	8.303085	0.060079	2.691467
(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.103825	0.101093	0.084699	0.815789	8.069701	0.074203	4.879781
(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE RED, ALARM CLOCK BAKELIK...	0.109290	0.084699	0.068306	0.625000	7.379032	0.059049	2.440801
(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE RED)	0.109290	0.101093	0.079235	0.725000	7.171622	0.068187	3.268753
(SET/6 RED SPOTTY PAPER PLATES)	(SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETRO...	0.136612	0.109290	0.106557	0.780000	7.137000	0.091627	4.048684
(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE GREEN)	0.109290	0.103825	0.079235	0.725000	6.982895	0.067888	3.258818
(SET/6 RED SPOTTY PAPER CUPS)	(SET/20 RED RETROSPOT PAPER NAPKINS, SET/6 RED...	0.147541	0.109290	0.106557	0.722222	6.608333	0.090433	3.206557

Visualization of the top 5 rules:



From the above output, we can say that the PACK OF 6 (SKULL PAPER CUPS and PACK OF 6 SKULL PAPER PLATES) and CHILDRENS CUTLERY DOLLY GIRL and CHILDRENS CUTLERY SPACEBOY are the most highly

associated items with lift value of 13 and 12 and confidence of 80% and 89% respectively and are most frequently bought together.

It means if a customer buys CHILDRENS CUTLERY DOLLY GIRL there are 89% chances that he will buy CHILDRENS CUTLERY SPACEBOY.

Conclusion:

Upon performing Market Basket Analysis for Top 3 countries – the UK, Germany and France it can be seen that the top items sold together in the UK are TEACUP AND SAUCER, and in France, they are PAPER CUPS and PAPER PLATES and ALARM CLOCKS whereas in Germany kid's items/toys are the top-selling items together.

The result of the above analysis is a piece of valuable information, and it can be used for data-driven campaigning, building marketing strategies and decision making.

The strategies for non-store online retail may include:

i) **Discounts**- Discount on Item B on purchase of item A
ii) **Product bundling:** Discounted/lower combo price of items A and B compared to individual price
iii) **Recommendations**: Recommend associated items with a lower price when a customer adds item A to the basket
iv) **Trending Items:** Suggestion of trending items customers buy etc