**MARKET BASKET INSIGHTS**

**Machine learning algorithms :**

Algorithms Used in Market Basket Analysis

1. Apriori Algorithm

2. AIS Algorithm

3. SETM Algorithm

4. FP Growth

**1.Apriori Algorithm**

Apriori Algorithm is a widely-used and well-known Association Rule algorithm and

Is a popular algorithm used in market basket analysis. It is also considered accurate

And overtop AIS and SETM algorithms. It helps to find frequent itemsets in

Transactions and identifies association rules between these items. The limitation of

The Apriori Algorithm is frequent itemset generation. It needs to scan the database

Many times, leading to increased time and reduced performance as a

Computationally costly step because of a large dataset. It uses the concepts of

Confidence and Support.

**2.AIS Algorithm**

The AIS algorithm creates multiple passes on the entire database or transactional

Data. During every pass, it scans all transactions. As you can see, in the first pass, it

Counts the support of separate items and determines then which of them are

Frequent in the database. Huge itemsets of every pass are enlarged to generate

Candidate itemsets. After each scanning of a transaction, the common itemsets

Between the itemsets of the previous pass and the items of this transaction are

Determined. This algorithm was the first published algorithm which is developed to

Generate all large itemsets in a transactional database. It focused on the

Enhancement of databases with the necessary performance to process decision

Support. This technique is bounded to only one item in the consequent.

**3.SETM Algorithm**

This Algorithm is quite similar to the AIS algorithm. The SETM algorithm creates

Collective passes over the database. As you can see, in the first pass, it counts the

Support of single items and then determines which of them are frequent in the

Database. Then, it also generates the candidate itemsets by enlarging large itemsets

Of the previous pass. In addition to this, the SETM algorithm recalls the

TIDs(transaction ids) of the generating transactions with the candidate itemsets.

**4.FP Growth**

FP Growth is known as Frequent Pattern Growth Algorithm. FP growth algorithm is

A concept of representing the data in the form of an FP tree or Frequent Pattern.

Hence FP Growth is a method of Mining Frequent Itemsets. This algorithm is an

Advancement to the Apriori Algorithm. There is no need for candidate generation

To generate a frequent pattern. This frequent pattern tree structure maintains the

Association between the itemsets.

A Frequent Pattern Tree is a tree structure that is made with the earlier itemsets of

The data. The main purpose of the FP tree is to mine the most frequent patterns.

Every node of the FP tree represents an item of that itemset. The root node

Represents the null value, whereas the lower nodes represent the itemsets of the

Data. The association of these nodes with the lower nodes, that is, between

Itemsets, is maintained while creating the tree.

Market Basket analysis using Apriori Algorithm

* [1. | Loading and Cleaning data](https://www.kaggle.com/code/earije/market-basket-analysis-with-apriori#section-one)
* [2. | Exploratoty Data Analysis](https://www.kaggle.com/code/earije/market-basket-analysis-with-apriori#section-two)
* [3. | Market Basket Analysis](https://www.kaggle.com/code/earije/market-basket-analysis-with-apriori#section-three)
* [4. | Conclusion](https://www.kaggle.com/code/earije/market-basket-analysis-with-apriori#section-four)

## 1. | **Loading and Cleaning data**

### 1-1. | Loading data

|  | BillNo | Itemname | Quantity | Date | Price | CustomerID | Country |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 536365 | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 01.12.2010 08:26 | 2,55 | 17850.0 | United Kingdom |
| 1 | 536365 | WHITE METAL LANTERN | 6 | 01.12.2010 08:26 | 3,39 | 17850.0 | United Kingdom |
| 2 | 536365 | CREAM CUPID HEARTS COAT HANGER | 8 | 01.12.2010 08:26 | 2,75 | 17850.0 | United Kingdom |
| 3 | 536365 | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 01.12.2010 08:26 | 3,39 | 17850.0 | United Kingdom |
| 4 | 536365 | RED WOOLLY HOTTIE WHITE HEART. | 6 | 01.12.2010 08:26 | 3,39 | 17850.0 | United Kingdom |

**<class 'pandas.core.frame.DataFrame'>**

**RangeIndex: 522064 entries, 0 to 522063**

**Data columns (total 7 columns):**

**# Column Non-Null Count Dtype**

**--- ------ -------------- -----**

**0 BillNo 522064 non-null object**

**1 Itemname 520609 non-null object**

**2 Quantity 522064 non-null int64**

**3 Date 522064 non-null object**

**4 Price 522064 non-null object**

**5 CustomerID 388023 non-null float64**

**6 Country 522064 non-null object**

**dtypes: float64(1), int64(1), object(5)**

**memory usage: 27.9+ MB**

Out[4]:

**BillNo 0**

**Itemname 1455**

**Quantity 0**

**Date 0**

**Price 0**

**CustomerID 134041**

**Country 0**

**dtype: int64**

### 1-2. | Dropping data with negative or zero quantity

In [6]:

**df**=**df**.**loc[df['Quantity']**>**0]**

### 1-3. | Dropping data with zero price

In [8]:

**df**=**df**.**loc[df['Price']**>**'0']**

### 1-4. | Dropping Non-product data.

In [10]:

**df**=**df**.**loc[(df['Itemname']**!=**'POSTAGE')**&**(df['Itemname']**!=**'DOTCOM POSTAGE')**&**(df['Itemname']**!=**'Adjust bad debt')**&**(df['Itemname']**!=**'Manual')]**

### 1-5. | Filling null data

In [12]:

**df**=**df**.**fillna('-')**

**df**.**isnull()**.**sum()**

Out[12]:

**BillNo 0**

**Itemname 0**

**Quantity 0**

**Date 0**

**Price 0**

**CustomerID 0**

**Country 0**

**dtype: int64**

### 1-6. | Splitting data into year and month

In [13]:

**df['Year']**=**df['Date']**.**apply(**lambda **x:x**.**split('.')[2])**

**df['Year']**=**df['Year']**.**apply(**lambda **x:x**.**split(' ')[0])**

**df['Month']**=**df['Date']**.**apply(**lambda **x:x**.**split('.')[1])**

**df**.**head()**

Out[13]:

|  | BillNo | Itemname | Quantity | Date | Price | CustomerID | Country | Year | Month |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 536365 | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 01.12.2010 08:26 | 2,55 | 17850.0 | United Kingdom | 2010 | 12 |
| 1 | 536365 | WHITE METAL LANTERN | 6 | 01.12.2010 08:26 | 3,39 | 17850.0 | United Kingdom | 2010 | 12 |
| 2 | 536365 | CREAM CUPID HEARTS COAT HANGER | 8 | 01.12.2010 08:26 | 2,75 | 17850.0 | United Kingdom | 2010 | 12 |
| 3 | 536365 | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 01.12.2010 08:26 | 3,39 | 17850.0 | United Kingdom | 2010 | 12 |
| 4 | 536365 | RED WOOLLY HOTTIE WHITE HEART. | 6 | 01.12.2010 08:26 | 3,39 | 17850.0 | United Kingdom | 2010 | 12 |

### 1-7. | Creating a Total price column

In [14]:

**df['Price']**=**df['Price']**.**str**.**replace(',','.')**.**astype('float64')**

**df['Total price']**=**df**.**Quantity**\***df**.**Price**

**df**.**head()**

Out[14]:

|  | BillNo | Itemname | Quantity | Date | Price | CustomerID | Country | Year | Month | Total price |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 536365 | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 01.12.2010 08:26 | 2.55 | 17850.0 | United Kingdom | 2010 | 12 | 15.30 |
| 1 | 536365 | WHITE METAL LANTERN | 6 | 01.12.2010 08:26 | 3.39 | 17850.0 | United Kingdom | 2010 | 12 | 20.34 |
| 2 | 536365 | CREAM CUPID HEARTS COAT HANGER | 8 | 01.12.2010 08:26 | 2.75 | 17850.0 | United Kingdom | 2010 | 12 | 22.00 |
| 3 | 536365 | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 01.12.2010 08:26 | 3.39 | 17850.0 | United Kingdom | 2010 | 12 | 20.34 |
| 4 | 536365 | RED WOOLLY HOTTIE WHITE HEART. | 6 | 01.12.2010 08:26 | 3.39 | 17850.0 | United Kingdom | 2010 | 12 | 20.34 |

### 1-8. | Checking the Total price in each month.

In [15]:

**df**.**groupby(['Year','Month'])['Total price']**.**sum()**

Out[15]:

**Year Month**

**2010 12 778386.780**

**2011 01 648311.120**

**02 490058.230**

**03 659979.660**

**04 507366.971**

**05 721789.800**

**06 710158.020**

**07 642528.481**

**08 701411.420**

**09 981408.102**

**10 1072317.070**

**11 1421055.630**

**12 606953.650**

**Name: Total price, dtype: float64**

**It is appropriate to look at 12-month increments to implement data analytics properly, so I'll drop the data for 2020 Dec.**

In [16]:

**df**=**df**.**loc[df['Year']**!=**'2010']**

## 2. | **Exploratoty Data Analysis**

### 2-1. | Sales amount and quantity

**2468101200.2M0.4M0.6M0.8M1M1.2M1.4M**

**CountryAustraliaBelgiumFranceGermanyGreeceHong KongIcelandIsraelItalyLebanonNetherlandsPolandPortugalSingaporeSpainSwedenSwitzerlandUnited KingdomAustriaJapanNorwaySaudi ArabiaUnited Arab EmiratesBrazilUSAUnspecifiedBahrainMaltaRSAMonthly sales amount in each country in 2021MonthSales amount**

**Most of the sales amounts are occupied by the UK.**

**United KingdomNetherlandsGermanyFranceAustraliaSpainSwitzerlandBelgiumSwedenJapanNorwayPortugalItalyHong KongSingaporeAustriaIsraelPolandUnspecifiedGreeceIcelandUSAMaltaUnited Arab EmiratesLebanonBrazilRSABahrainSaudi Arabia5M6M5.5M6.5M**

**Sales amount in each country in 2021CountrySales amount**

### 2-2. | Category

### **Top 10 highest sales amount items**

Out[20]:

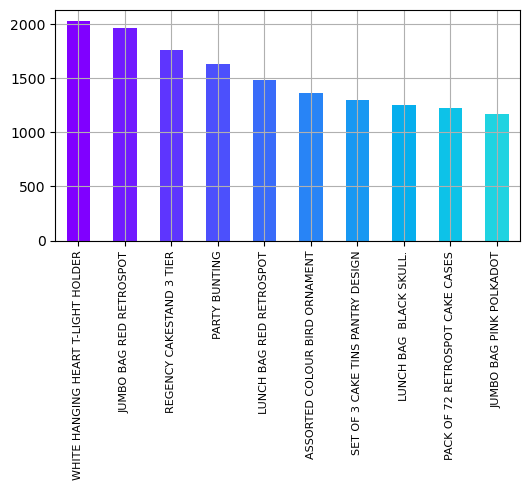
|  | Itemname | Price |
| --- | --- | --- |
| 0 | REGENCY CAKESTAND 3 TIER | 24653.67 |
| 1 | PARTY BUNTING | 9416.13 |
| 2 | SET OF 3 CAKE TINS PANTRY DESIGN | 7621.05 |
| 3 | CREAM SWEETHEART MINI CHEST | 6836.38 |
| 4 | SET/4 WHITE RETRO STORAGE CUBES | 6714.75 |
| 5 | ENAMEL BREAD BIN CREAM | 6585.93 |
| 6 | WHITE HANGING HEART T-LIGHT HOLDER | 6563.80 |
| 7 | DOORMAT KEEP CALM AND COME IN | 6385.09 |
| 8 | SPOTTY BUNTING | 6262.40 |
| 9 | RED RETROSPOT CAKE STAND | 6035.29 |

### **Top 10 most purchased items**

Out[21]:

|  | Itemname | Quantity |
| --- | --- | --- |
| 520583 | PAPER CRAFT , LITTLE BIRDIE | 80995 |
| 59999 | MEDIUM CERAMIC TOP STORAGE JAR | 74215 |
| 405138 | WORLD WAR 2 GLIDERS ASSTD DESIGNS | 4800 |
| 198929 | SMALL POPCORN HOLDER | 4300 |
| 94245 | EMPIRE DESIGN ROSETTE | 3906 |
| 260928 | ESSENTIAL BALM 3.5g TIN IN ENVELOPE | 3186 |
| 51228 | FAIRY CAKE FLANNEL ASSORTED COLOUR | 3114 |
| 154834 | FAIRY CAKE FLANNEL ASSORTED COLOUR | 3114 |
| 416997 | SMALL CHINESE STYLE SCISSOR | 3000 |
| 280572 | ASSORTED COLOUR BIRD ORNAMENT | 2880 |

### **Top 10 most frequently purchased items**

****

## 3. | **Market Basket Analysis**

**Since the UK is the most purchased country, let insight into the item combination purchased in the UK.**

### 3-1. | Implementing Apriori

Out[26]:

| Itemname | 10 COLOUR SPACEBOY PEN | 12 COLOURED PARTY BALLOONS | 12 DAISY PEGS IN WOOD BOX | 12 EGG HOUSE PAINTED WOOD | 12 HANGING EGGS HAND PAINTED | 12 IVORY ROSE PEG PLACE SETTINGS | 12 MESSAGE CARDS WITH ENVELOPES | 12 PENCIL SMALL TUBE WOODLAND | 12 PENCILS SMALL TUBE RED RETROSPOT | 12 PENCILS SMALL TUBE SKULL | ... | ZINC STAR T-LIGHT HOLDER | ZINC SWEETHEART SOAP DISH | ZINC SWEETHEART WIRE LETTER RACK | ZINC T-LIGHT HOLDER STAR LARGE | ZINC T-LIGHT HOLDER STARS LARGE | ZINC T-LIGHT HOLDER STARS SMALL | ZINC TOP 2 DOOR WOODEN SHELF | ZINC WILLIE WINKIE CANDLE STICK | ZINC WIRE KITCHEN ORGANISER | ZINC WIRE SWEETHEART LETTER TRAY |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| BillNo |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 539993 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 540001 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 540002 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | ... | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

**3 rows × 3893 columns**

Out[30]:

|  | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction | zhangs\_metric |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | (ALARM CLOCK BAKELIKE GREEN) | (ALARM CLOCK BAKELIKE RED) | 0.05 | 0.05 | 0.03 | 0.64 | 12.41 | 0.03 | 2.64 | 0.97 |
| 1 | (ALARM CLOCK BAKELIKE RED) | (ALARM CLOCK BAKELIKE GREEN) | 0.05 | 0.05 | 0.03 | 0.59 | 12.41 | 0.03 | 2.32 | 0.97 |
| 2 | (GARDENERS KNEELING PAD KEEP CALM) | (GARDENERS KNEELING PAD CUP OF TEA) | 0.05 | 0.05 | 0.03 | 0.60 | 13.23 | 0.03 | 2.40 | 0.98 |
| 3 | (GARDENERS KNEELING PAD CUP OF TEA) | (GARDENERS KNEELING PAD KEEP CALM) | 0.05 | 0.05 | 0.03 | 0.72 | 13.23 | 0.03 | 3.39 | 0.97 |
| 4 | (PINK REGENCY TEACUP AND SAUCER) | (GREEN REGENCY TEACUP AND SAUCER) | 0.04 | 0.05 | 0.03 | 0.82 | 15.50 | 0.03 | 5.25 | 0.98 |

### 3-2. | The top 5 of the highest support value of items(antecedents)

#### **Support(item) = Transactions comprising the item / Total transactions**

Out[32]:

|  | antecedents | consequents | support |
| --- | --- | --- | --- |
| 13 | frozenset({'JUMBO BAG RED RETROSPOT'}) | frozenset({'JUMBO BAG PINK POLKADOT'}) | 0.05 |
| 12 | frozenset({'JUMBO BAG PINK POLKADOT'}) | frozenset({'JUMBO BAG RED RETROSPOT'}) | 0.05 |
| 16 | frozenset({'JUMBO STORAGE BAG SUKI'}) | frozenset({'JUMBO BAG RED RETROSPOT'}) | 0.04 |
| 17 | frozenset({'JUMBO BAG RED RETROSPOT'}) | frozenset({'JUMBO STORAGE BAG SUKI'}) | 0.04 |
| 15 | frozenset({'JUMBO SHOPPER VINTAGE RED PAISLEY'}) | frozenset({'JUMBO BAG RED RETROSPOT'}) | 0.04 |

**In the top support value of purchase, it means that "JUMBO BAG PINK RETROSPOT" is present in 5% of all purchases.**

### 3-3. | The top 5 of the highest confidence value of items

#### **Confidence = Transactions comprising antecedent and consequent / Transactions comprising antecedent**

Out[33]:

|  | antecedents | consequents | confidence |
| --- | --- | --- | --- |
| 4 | frozenset({'PINK REGENCY TEACUP AND SAUCER'}) | frozenset({'GREEN REGENCY TEACUP AND SAUCER'}) | 0.82 |
| 30 | frozenset({'PINK REGENCY TEACUP AND SAUCER'}) | frozenset({'ROSES REGENCY TEACUP AND SAUCER'}) | 0.78 |
| 6 | frozenset({'GREEN REGENCY TEACUP AND SAUCER'}) | frozenset({'ROSES REGENCY TEACUP AND SAUCER'}) | 0.75 |
| 7 | frozenset({'ROSES REGENCY TEACUP AND SAUCER'}) | frozenset({'GREEN REGENCY TEACUP AND SAUCER'}) | 0.73 |
| 3 | frozenset({'GARDENERS KNEELING PAD CUP OF TEA'}) | frozenset({'GARDENERS KNEELING PAD KEEP CALM'}) | 0.72 |

**In the top confidence value of the purchase, it means that 82% of the customers who bought "PINK REGENCY TEACUP AND SAUCER" also bought "GREEN REGENCY TEACUP AND SAUCER".**

### 3-4. | The top 5 of the highest lift value of items

#### **Lift = Confidence (antecedent -> consequent) / Support(antecedent)**

In [34]:

**rules[['antecedents','consequents','lift']]**.**sort\_values('lift',ascending**=False**)[:5]**.**style**.**background\_gradient(cmap**=**cm)**.**set\_precision(2)**

Out[34]:

|  | antecedents | consequents | lift |
| --- | --- | --- | --- |
| 4 | frozenset({'PINK REGENCY TEACUP AND SAUCER'}) | frozenset({'GREEN REGENCY TEACUP AND SAUCER'}) | 15.50 |
| 5 | frozenset({'GREEN REGENCY TEACUP AND SAUCER'}) | frozenset({'PINK REGENCY TEACUP AND SAUCER'}) | 15.50 |
| 31 | frozenset({'ROSES REGENCY TEACUP AND SAUCER'}) | frozenset({'PINK REGENCY TEACUP AND SAUCER'}) | 14.36 |
| 30 | frozenset({'PINK REGENCY TEACUP AND SAUCER'}) | frozenset({'ROSES REGENCY TEACUP AND SAUCER'}) | 14.36 |
| 6 | frozenset({'GREEN REGENCY TEACUP AND SAUCER'}) | frozenset({'ROSES REGENCY TEACUP AND SAUCER'}) | 13.86 |

**In the top list value of the purchase, it means that customers are 15.5 times more likely to buy "GREEN REGENCY TEACUP AND SAUCER" if you sell "PINK REGENCY TEACUP AND SAUCER".**

### 3-5. | The best combination of the items

Out[35]:

|  | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction | zhangs\_metric |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 4 | frozenset({'PINK REGENCY TEACUP AND SAUCER'}) | frozenset({'GREEN REGENCY TEACUP AND SAUCER'}) | 0.04 | 0.05 | 0.03 | 0.82 | 15.50 | 0.03 | 5.25 | 0.98 |
| 30 | frozenset({'PINK REGENCY TEACUP AND SAUCER'}) | frozenset({'ROSES REGENCY TEACUP AND SAUCER'}) | 0.04 | 0.05 | 0.03 | 0.78 | 14.36 | 0.03 | 4.24 | 0.97 |
| 6 | frozenset({'GREEN REGENCY TEACUP AND SAUCER'}) | frozenset({'ROSES REGENCY TEACUP AND SAUCER'}) | 0.05 | 0.05 | 0.04 | 0.75 | 13.86 | 0.04 | 3.78 | 0.98 |
| 7 | frozenset({'ROSES REGENCY TEACUP AND SAUCER'}) | frozenset({'GREEN REGENCY TEACUP AND SAUCER'}) | 0.05 | 0.05 | 0.04 | 0.73 | 13.86 | 0.04 | 3.55 | 0.98 |
| 3 | frozenset({'GARDENERS KNEELING PAD CUP OF TEA'}) | frozenset({'GARDENERS KNEELING PAD KEEP CALM'}) | 0.05 | 0.05 | 0.03 | 0.72 | 13.23 | 0.03 | 3.39 | 0.97 |

## 4. | **Conclusion**

**• The most purchased item is PAPER CRAFT, LITTLE BIRDIE.**

* **The most frequently purchased item is WHITE HANGING HEART T-LIGHT HOLDER.**
* **The best combination items are PINK REGENCY TEACUP AND SAUCER and GREEN REGENCY TEACUP AND SAUCER.**