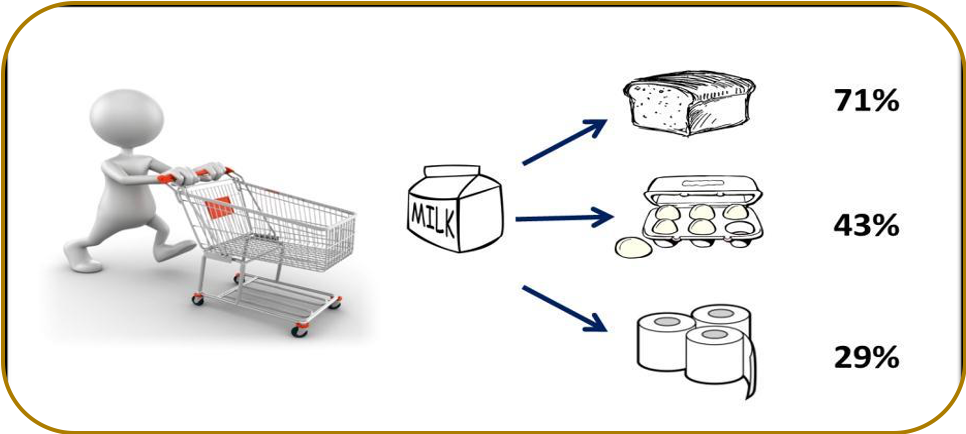
MARKET BASKET INSIGHTS

Phase- 4 Submission Document

**Project:** Market basket Insights

**Phase** 4:Development Part 2

**Topic**: continue building market basket insights. selecting a machine learning algorithm, training the model, and evaluating its performance. Perform different analysis as needed.



Market Basket Insights

**Introduction:**

Market basket analysis is a technique that helps retailers understand the buying patterns of their customers by analysing the items that are frequently purchased together. It can help them optimize their product placement, pricing, marketing, and cross-selling strategies.

**Purpose of Market Basket Insights:**

The purpose of market basket insights is to understand customers purchasing behaviour and identify cross-selling and up-selling opportunities.

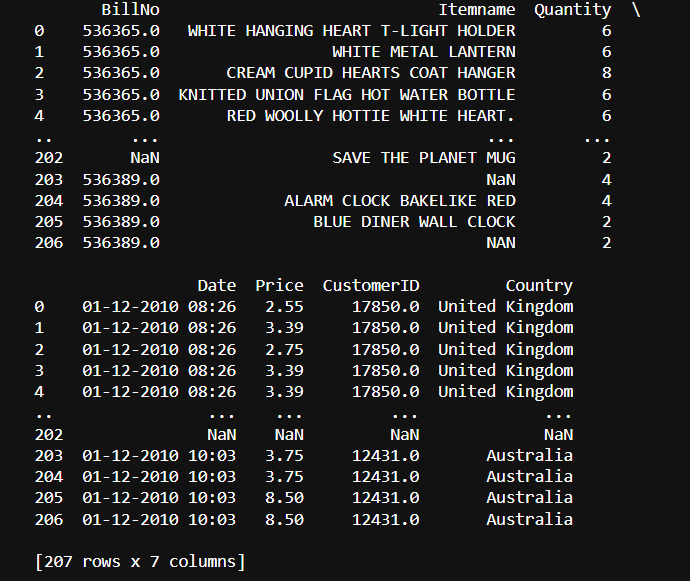
 Market basket analysis can also help retailers discover new product combinations, customer segments, and market trends.

**Algorithms of Market Basket Analysis:**

There are various algorithms that can be used to perform market basket insights

such as Apriori, FP-Growth, AIS, and SETM. These algorithms differ in their efficiency, scalability, and complexity. To perform market basket insights in Python, one can use libraries such as mlxtend or Orange3-Associate.

**Given data set:**



**Machine Learning algorithm:**

The Apriori algorithm is a classic data mining and machine learning technique used for discovering association rules in large datasets

It is commonly applied to market basket insights to find relationships between items that customers frequently purchase together optimize store layouts, and enhance sales strategies.

* Data Preparation:
* Start with a dataset where each row represents a transaction, and each column represents an item (product or service).
* Encode the data, typically using binary values (1 for item presence, 0 for absence) for each item in each transaction.
* **Apriori Principle**:
* The Apriori algorithm relies on the Apriori principle, which states that if an itemset is frequent (has high support), then all of its subsets must also be frequent.
* **Algorithm Steps**:
* Initialize with frequent itemsets of size 1 (single items).
* While there are frequent itemsets of size k, generate candidates of size k+1.
* Prune candidates that do not meet minimum support.
* Calculate confidence for the remaining candidates.
* Repeat the process until no more candidates can be generated.

**Training the model:**

Choose a machine learning algorithm: such as Apriori, FP-Growth, AIS, and SETM. These algorithms differ in their efficiency, scalability, and complexity. To perform market basket insights in Python, one can use libraries such as mlxtend or Orange3-Associate.

**Import apriori:**

import numpy as np

import pandas as pd

from mlxtend.frequent\_patterns import apriori

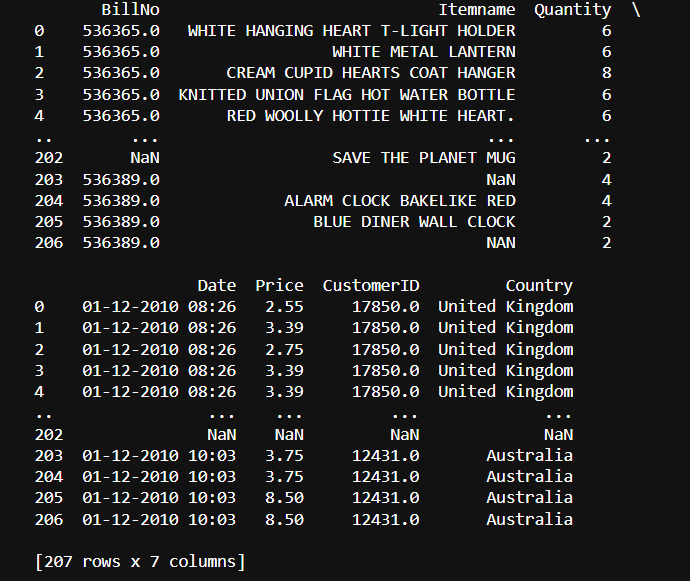
from mlxtend.frequent\_patterns import association\_rules

**PREPROCESSING:**

**Program:**

data.head(207)

**output:**



**Program:**

**Check the null values**

data.isnull().sum()

**output:**

BillNo 0

Itemname 1455

Quantity 0

Date 0

Price 0

CustomerID 134041

Country 0

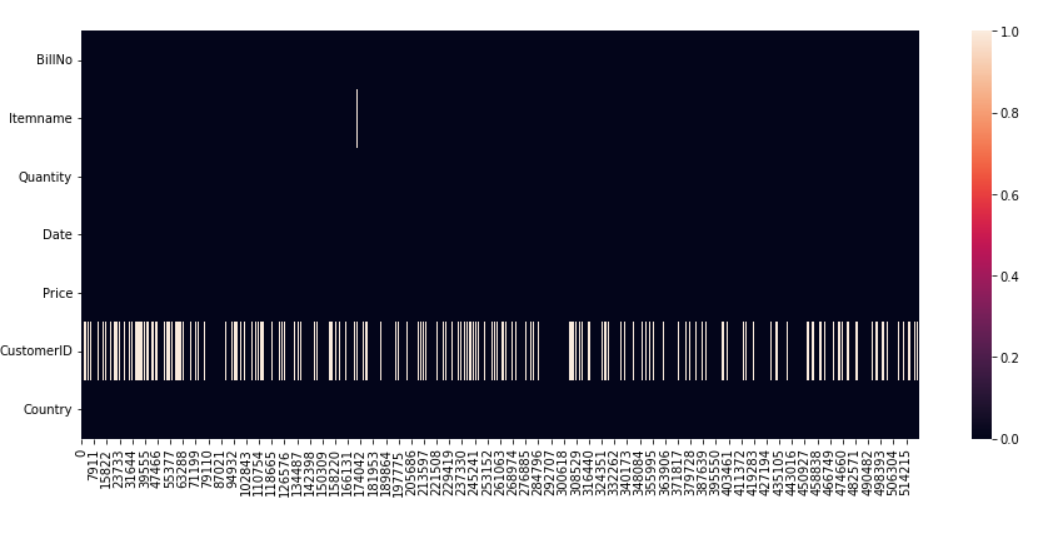
dtype: int64

**program:**

plt.figure(figsize=(15,6))

sns.heatmap(data.isna().transpose())

**output:**

****

Program:

**Remove null values:**

data = data.dropna(subset=["Itemname"])

**output:**

BillNo 0

Itemname 0

Quantity 0

Date 0

Price 0

CustomerID 132586

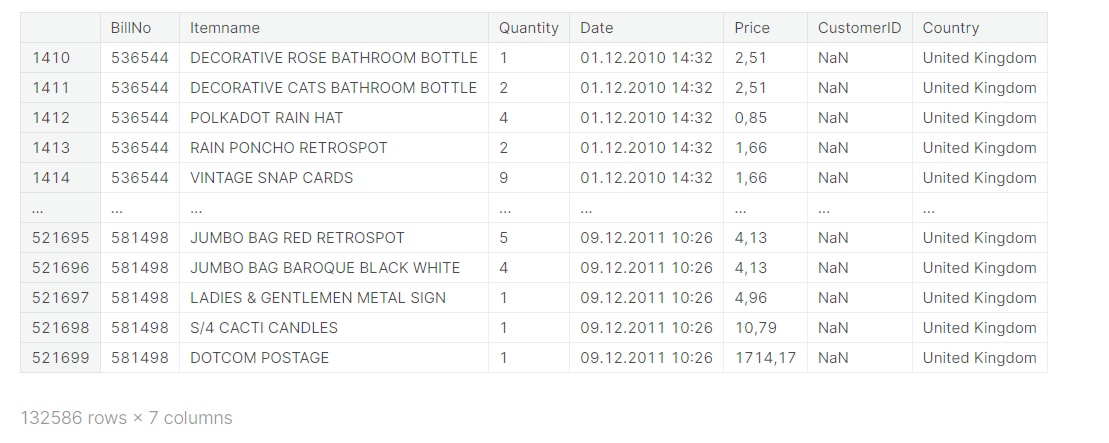
Country 0

dtype: int64

**program:**

data[data['CustomerID'].isnull()]

**output:**

****

**Program:**

data = data.fillna(0)

data[data["CustomerID"].isnull()]

**output:**

****

Program:

data.isnull().sum()

output:

BillNo 0

Itemname 0

Quantity 0

Date 0

Price 0

CustomerID 0

Country 0

dtype: int64

program:

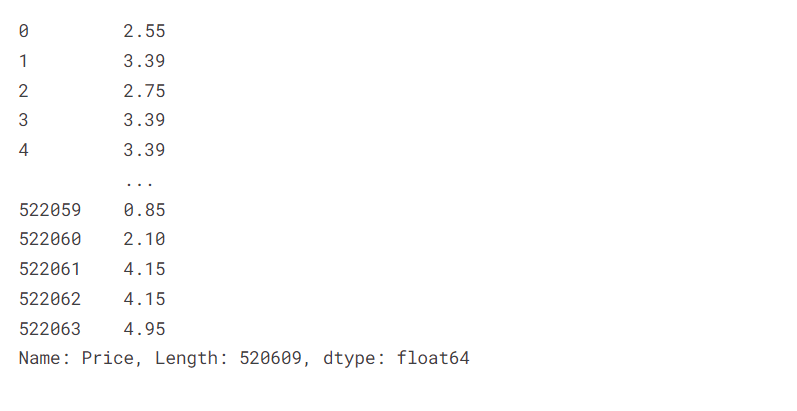
data["Date"] = pd.to\_datetime(data["Date"])

data["Price"] = data["Price"].str.replace(",",".")

data["Price"] = pd.to\_numeric(data["Price"])

data["Price"]

output:



Program:

Frequency:

top20items = pd.DataFrame(data["Itemname"].value\_counts().head(20))

top20items = top20items.reset\_index()

top20items.columns = ["Itemname","Frequency"]

labels = top20items["Itemname"]

sizes = top20items["Frequency"]

top20items

output:



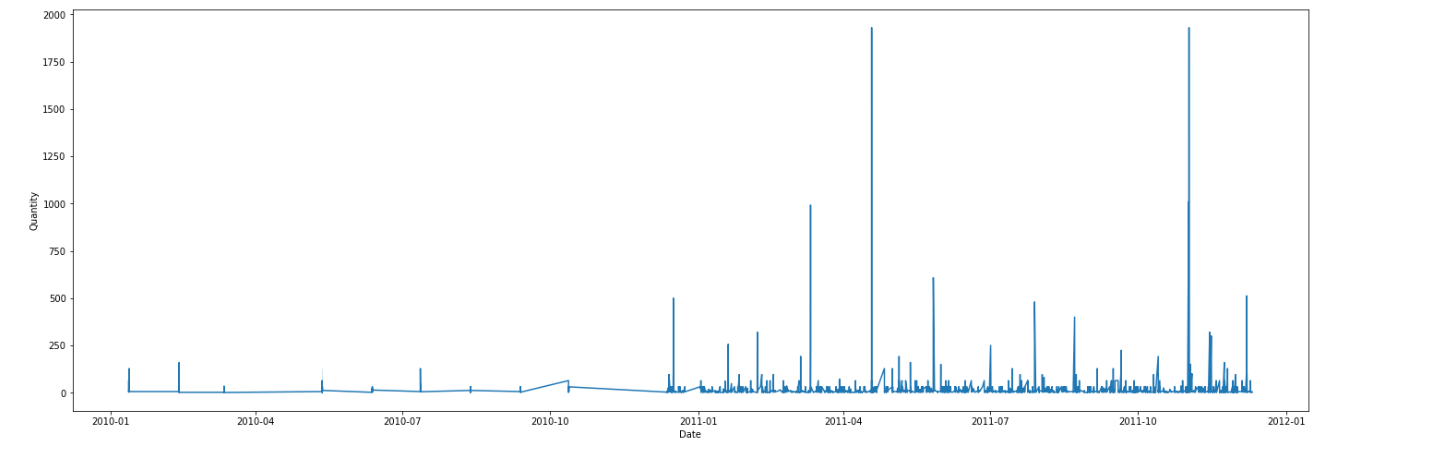
Program:

Quantity:

fig = plt.figure(figsize=(24,8))

sns.lineplot(x = t\_light["Date"], y = t\_light["Quantity"] )

output:



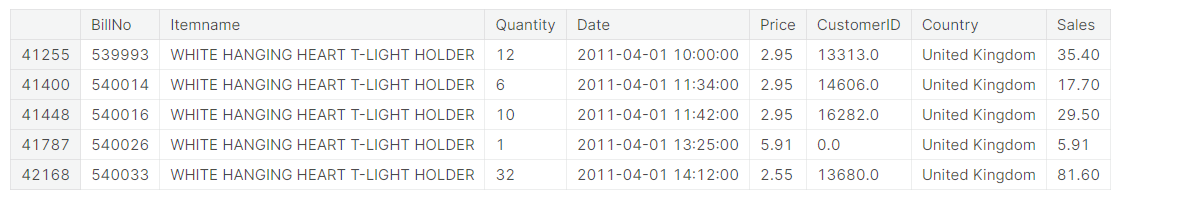
Program:

startdate = t\_light["Date"] >= "20110101"

t\_light\_2011 = t\_light.loc[startdate]

t\_light\_2011.head()

output:



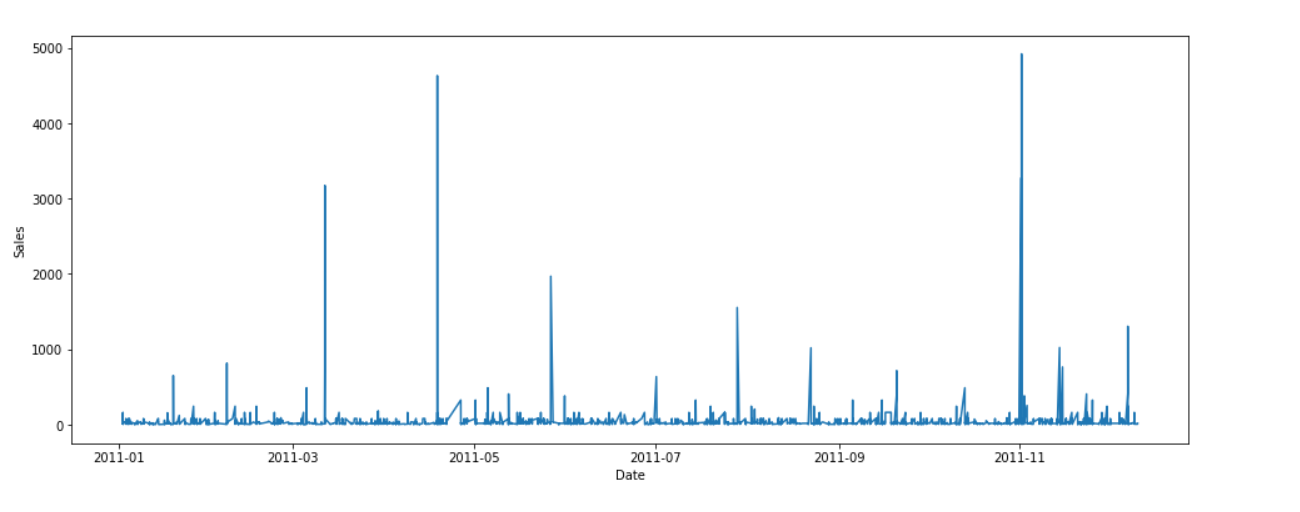
Program:

Sales frequency:

plt.figure(figsize=(16,6))

sns.lineplot(t\_light\_2011["Date"],t\_light\_2011["Sales"])

output:



Performing Association Analysis:

Association analysis, often referred to as market basket insights, is a data mining technique used to discover interesting relationships or associations between items in a transactional dataset.

* **Association Rules:**
* Association rules are generated by identifying itemsets with high support and confidence values.
* Rules typically take the form of "If {A}, then {B}," indicating that when customers buy item A, they are also likely to purchase item B.
* **Setting Thresholds:**

You can set minimum support and confidence thresholds to filter out rules. This allows you to focus on the most interesting and relevant associations.

* **Interpreting and Using Results**:
  + Analyze the generated rules to gain insights into customer behavior.
  + Use the insights to make business decisions, such as product placement, cross-selling, upselling, marketing campaigns, and inventory management.
* **A/B Testing:**

Implement A/B tests based on the insights from association analysis to validate the impact of changes in marketing or product placement on customer behavior.

**Program:**

mb\_dataset = pd.read\_csv('/kaggle/input/market-basket-analysis/Assignment-1\_Data.csv',delimiter=';')

mb\_dataset.head()

mb\_dataset.groupby('BillNo').BillNo.count().sort\_values()

mb\_dataset.groupby('Itemname').Itemname.count().sort\_values()

mb\_dataset.isnull().sum()

mbdatasetclean = mb\_dataset.iloc[:,0:2]

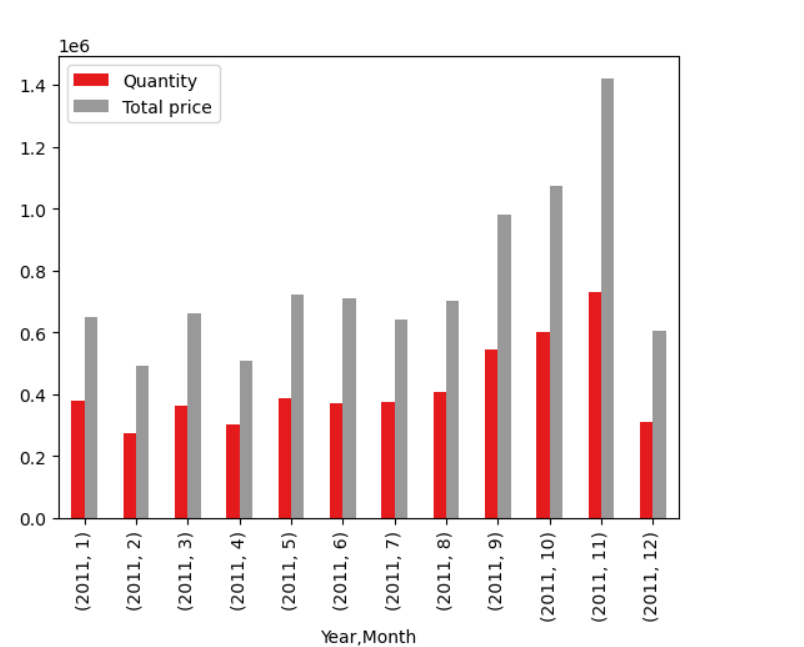
mbdatasetclean

mbdatasetclean[mbdatasetclean['Itemname'].isna()].index

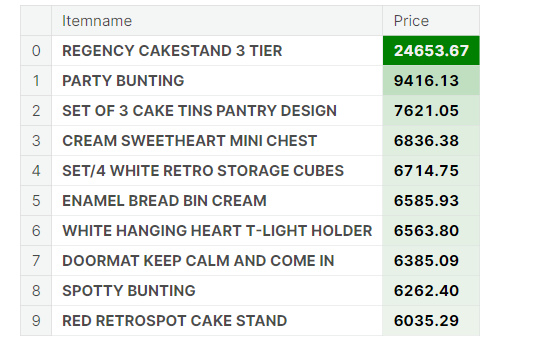
mbdatasetclean = mbdatasetclean.drop( mbdatasetclean[mbdatasetclean['Itemname'].isna()].index )

mbdatasetclean.isnull().sum()

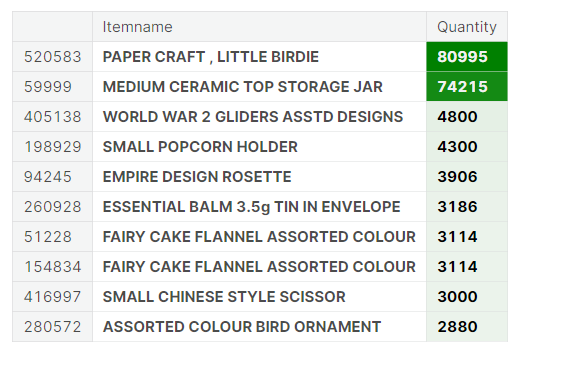
**Sales amount and quantity:**

****

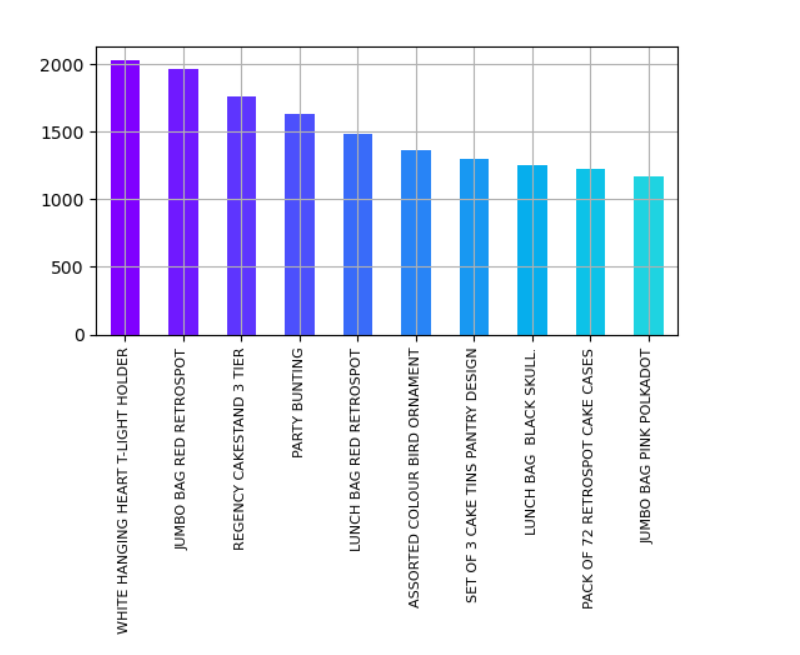
**Top 10 highest sales amount items:**

****

**Top 10 most purchased items:**

****

**Top 10 most frequently purchased items:**

****

**Data Analysis:**

Data analysis in the context of market basket insights, also known as market basket analysis or association analysis, involves examining transactional data to uncover patterns, relationships, and associations between items that customers purchase together

**Program:**

df\_market.createOrReplaceTempView("df")

itemname\_by\_country = sqlCtx.sql("""SELECT Country, Itemname, SUM(Quantity) as Quantity, SUM(TotPrice) as TotPrice FROM df GROUP BY Country, Itemname""")

itemname\_by\_country.createOrReplaceTempView("itemname\_by\_country")

top\_product\_country = sqlCtx.sql("""SELECT Country, Itemname, Quantity FROM

(SELECT Country, Itemname, Quantity, MAX(Quantity) OVER(PARTITION BY Country) AS Max\_Quant FROM itemname\_by\_country)

WHERE Quantity=MAX\_QUANT

""")

print("Best sellers by country")

top\_product\_country.orderBy('Quantity', ascending=False).show(40, truncate=False)

output:

Best sellers by country

[Stage 19:=============================> (2 + 2) / 4]

+--------------------+-----------------------------------+--------+

|Country |Itemname |Quantity|

+--------------------+-----------------------------------+--------+

|United Kingdom |PAPER CRAFT , LITTLE BIRDIE |80995 |

|Netherlands |RABBIT NIGHT LIGHT |4801 |

|France |RABBIT NIGHT LIGHT |4024 |

|Japan |RABBIT NIGHT LIGHT |3408 |

|Australia |MINI PAINT SET VINTAGE |2952 |

|Sweden |MINI PAINT SET VINTAGE |2916 |

|Germany |ROUND SNACK BOXES SET OF4 WOODLAND |1233 |

|Spain |CHILDRENS CUTLERY POLKADOT PINK |729 |

|Switzerland |PLASTERS IN TIN WOODLAND ANIMALS |639 |

|Norway |SMALL FOLDING SCISSOR(POINTED EDGE)|576 |

|Belgium |PACK OF 72 RETROSPOT CAKE CASES |480 |

|Singapore |CHRISTMAS TREE PAINTED ZINC |384 |

|Austria |SET 12 KIDS COLOUR CHALK STICKS |288 |

|Italy |FEATHER PEN,HOT PINK |240 |

|Iceland |ICE CREAM SUNDAE LIP GLOSS |240 |

|Portugal |POLKADOT PEN |240 |

|Hong Kong |ROUND SNACK BOXES SET OF4 WOODLAND |150 ||Greece |4 LAVENDER BOTANICAL DINNER CANDLES|48 |

|Greece |4 PEAR BOTANICAL DINNER CANDLES |48 |

|Lithuania |FELTCRAFT DOLL ROSIE |48 |

|Lithuania |RED HARMONICA IN BOX |48 |

|Brazil |ROSES REGENCY TEACUP AND SAUCER |24 |

|Brazil |SET OF 4 PANTRY JELLY MOULDS |24 |

|Brazil |DOLLY GIRL LUNCH BOX |24 |

|Brazil |SMALL HEART FLOWERS HOOK |24 |

|Brazil |GREEN REGENCY TEACUP AND SAUCER |24 |

|Brazil |PINK REGENCY TEACUP AND SAUCER |24 |

|Brazil |SET OF 6 SPICE TINS PANTRY DESIGN |24 |

|Brazil |SET/3 RED GINGHAM ROSE STORAGE BOX |24 |

|Lebanon |ASSTD FRUIT+FLOWERS FRIDGE MAGNETS |24 |

|RSA |PACK OF 6 BIRDY GIFT TAGS |12 |

+--------------------+-----------------------------------+--------+

only showing top 40 rows

program:

tot\_price\_by\_country = sqlCtx.sql("""SELECT Country, ROUND(SUM(TotPrice), 2) AS TotPrice FROM itemname\_by\_country GROUP BY Country""")

print("Total gain by country")

tot\_price\_by\_country\_pd = tot\_price\_by\_country.orderBy('TotPrice', ascending=False).toPandas()

tot\_price\_by\_country\_pd.head(40)

output:

|  |  |  |
| --- | --- | --- |
| 0 | United Kingdom | 9003097.96 |
| 1 | Netherlands | 285446.34 |
| 2 | Germany | 228867.14 |
| 3 | France | 209715.11 |
| 4 | Australia | 138521.31 |
| 5 | Spain | 61577.11 |
| 6 | Switzerland | 57089.90 |
| 7 | Belgium | 41196.34 |
| 8 | Sweden | 38378.33 |
| 9 | Japan | 37416.37 |
| 10 | Norway | 36165.44 |
| 11 | Portugal | 33747.10 |
| 12 | Singapore | 21279.29 |
| 13 | Italy | 17483.24 |
| 14 | Hong Kong | 15691.80 |
| 15 | Austria | 10198.68 |
| 16 | Israel | 8135.26 |
| 17 | Poland | 7334.65 |
| 18 | Greece | 4760.52 |
| 19 | Unspecified | 4749.79 |
| 20 | Iceland | 4310.00 |
| 21 | USA | 3580.39 |
| 22 | Malta | 2725.59 |
| 23 | United Arab Emirates | 1902.28 |
| 24 | Lebanon | 1693.88 |
| 25 | Lithuania | 1661.06 |
| 26 | Brazil | 1143.60 |
| 27 | RSA | 1002.31 |
| 28 | Bahrai | 754.789 |

**Program:**

x = tot\_price\_by\_country\_pd['Country']

y = tot\_price\_by\_country\_pd['TotPrice']

plt.bar(x, y)

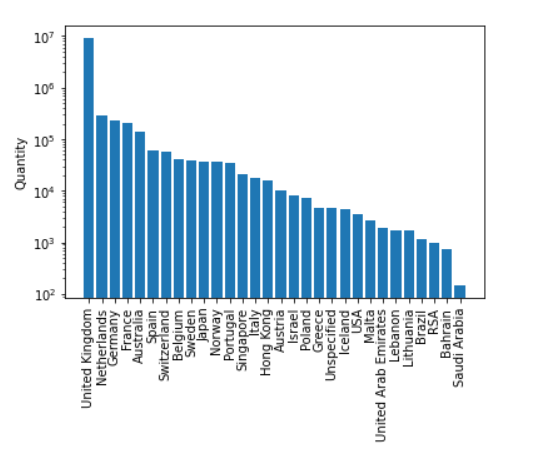
plt.yscale('log')

plt.ylabel('Quantity')

plt.xticks(rotation=90)

plt.show()

**output:**

****

**Generating insights:**

Generating insights in the context of market basket insights involves extracting valuable information and knowledge from transactional data to make informed decisions and improvements in various aspects of a business, particularly in retail and e-commerce.

* **Support, Confidence, and Lift Metrics:**
* Support, confidence, and lift are important metrics for assessing the strength and significance of association rules.
* Insights are derived by examining the values of these metrics. High confidence and lift suggest strong relationships, while high support indicates popularity
* **Identifying Product Associations**:
* By analyzing association rules and frequent itemsets, businesses can identify which products are often bought together. This information can be used for cross-selling and upselling strategies.
* For example, if customers frequently buy cameras with memory cards, insights can lead to bundling these items for promotions.
* **Marketing and Promotion Strategies**:
* Insights can be used to design marketing campaigns and promotions that are more likely to resonate with customers.
* For instance, businesses can create personalized recommendations based on the associations discovered in the data.
* **Customer Segmentation:**
* Insights can be used to segment customers based on their purchase behaviours. Different customer groups may exhibit distinct purchase patterns.
* Tailoring marketing and product recommendations to each segment can lead to more effective strategies.
* **Data-Driven Decision-Making:**
* Insights from market basket analysis provide a data-driven foundation for making decisions that aim to enhance customer satisfaction, increase revenue, and optimize business operations.

**Apriori Analysis:**

import numpy as np

import pandas as pd

from mlxtend.frequent\_patterns import apriori

from mlxtend.frequent\_patterns import association\_rules

from pyspark.ml.feature import OneHotEncoder

from pyspark.ml.feature import StringIndexer, CountVectorizer

from pyspark.sql.types import IntegerType

indexer = StringIndexer(inputCol="Itemname", outputCol="ItemnameIndex")

df\_group = indexer.fit(df\_market).transform(df\_market)

#df\_group = df\_group.withColumn("Itemname", df\_group["ItemnameIndex"].cast(IntegerType()))

df\_group = df\_group[['Country', 'BillNo', 'Itemname']].distinct()

df\_group = df\_group.groupBy('BillNo', 'Country').agg(collect\_list('Itemname').alias('Itemname\_vec'))

df\_group.show()

output:

+------+--------------+--------------------+

|BillNo| Country| Itemname\_vec|

+------+--------------+--------------------+

|536365|United Kingdom|[GLASS STAR FROST...|

|536369|United Kingdom|[BATH BUILDING BL...|

|536380|United Kingdom|[JAM MAKING SET P...|

|536388|United Kingdom|[HEART OF WICKER ...|

|536394|United Kingdom|[FANCY FONT BIRTH...|

|536395|United Kingdom|[BLACK HEART CARD...|

|536402|United Kingdom|[HOT WATER BOTTLE...|

|536404|United Kingdom|[PACK OF 72 RETRO...|

|536420|United Kingdom|[SCOTTIE DOG HOT ...|

|536425|United Kingdom|[GINGHAM HEART D...|

|536520|United Kingdom|[SET/10 PINK POLK...|

|536530|United Kingdom|[SILVER GLASS T-L...|

|536531|United Kingdom|[EDWARDIAN PARASO...|

|536532| Norway|[PLASTERS IN TIN ...|

|536534|United Kingdom|[HAND WARMER SCOT...|

|536535|United Kingdom|[LOVEBIRD HANGING...|

|536536|United Kingdom|[SET OF 2 CHRISTM...|

|536539|United Kingdom|[PLASTERS IN TIN ...|

|536565|United Kingdom|[MIRRORED WALL AR...|

|536575|United Kingdom|[EDWARDIAN PARASO...|

|536565|United Kingdom|[MIRRORED WALL AR...|

|536575|United Kingdom|[EDWARDIAN PARASO...|

+------+--------------+--------------------+

only showing top 20 rows

**Conclusion:**

In conclusion, market basket insights are a valuable asset for businesses, particularly those in retail and e-commerce, seeking to understand and leverage customer purchase patterns. By analyzing transactional data, companies can uncover associations and relationships between items that provide numerous benefits.

* **Enhanced Customer Experiences**: Market basket insights enable businesses to offer customers more relevant product recommendations. By understanding which items are frequently purchased together, businesses can improve the shopping experience and increase customer satisfaction.
* **Optimized Business Operations:** Insights from market basket analysis can lead to more effective inventory management. Businesses can adjust stock levels based on anticipated demand for associated items, reducing costs and minimizing stockouts or overstocking.
* **Strategic Marketing:** With knowledge of product associations, companies can design marketing campaigns and promotions that resonate with customers. Personalized recommendations and targeted marketing strategies are more likely to boost sales.