جامعة أم القرى
UMM AL-QURA UNIVERSITY

REPORT

DATA ANALYSIS (2)

DS3114

Dr. Omaima Fallatah

Team Members:

- Faten Matouq Almowallad 444000266
- Ayah Bakur Alhawsawi 444006678

https://colab.research.google.com/drive/193dc74CvY2OJ3leKVXrPT-69u9AsCa3t#scrollTo=8r575pAGdR3Y

Tasks: 2

Market Basket Analysis

INTRODUCTION

Market Basket Analysis (MBA) is a data analysis technique commonly used in retail and e-commerce. It identifies associations between items purchased together, helping businesses understand customer behavior and improve product placements, promotional strategies, and recommendation systems. The core method used here is the *apriori algorithm*, which helps identify frequent itemsets and generate association rules that reveal interesting relationships between products.

ABOUT DATASET

The dataset contains transaction records with attributes such as BillNo, Itemname, Quantity, and Country. Each row represents a unique item bought in a particular transaction. The task focuses on analyzing these transactions to discover meaningful item associations.

Key features of the dataset include:

- **Transaction ID**: Identifies each transaction uniquely.
- **Items**: List of items purchased in a given transaction.
- **Format**: CSV file, loaded with different delimiter options to ensure data integrity during the import process.

The dataset required some adjustments during import to handle mixed data types and parsing errors.

Data Preprocessing

Preprocessing steps were crucial to ensure that the data was suitable for the analysis. The following steps were taken:

- **Data Cleaning**: The dataset had some inconsistencies, including mixed types in certain columns and missing values. These were addressed by skipping problematic lines during import.
- **Data Structuring**: To apply the apriori algorithm, the data was transformed into a format where each row represents a transaction and each column represents an item, with a binary value indicating whether an item was purchased in that transaction.
- **Transaction Matrix**: Created a matrix where each transaction was represented as a row, with items as columns (1 if the item was present in a transaction, 0 otherwise).

ANALYSIS & RESULTS

The analysis focused on generating frequent itemsets using the apriori algorithm and deriving association rules:

- **Apriori Algorithm**: The Apriori algorithm was used to find frequent itemsets in the transaction data. Itemsets with a minimum support of 1% were considered, meaning that these itemsets appear in at least 1% of all transactions.
- **Frequent Itemsets**: The apriori algorithm was used to find combinations of items that are frequently bought together. The minimum support threshold was set to ensure that only combinations with significant occurrences were considered.
- **Association Rules**: Rules were derived from the frequent itemsets using confidence and lift as measures. These rules indicate how likely it is for a customer to buy a certain item when they have already purchased another.
  - **Support**: Indicates how frequently an itemset appears in the dataset.
  - **Confidence**: Measures the likelihood of purchasing an item given that another item is already purchased.
  - **Lift**: Evaluates how much more likely the purchase of an item is, given the presence of another, compared to a random purchase.

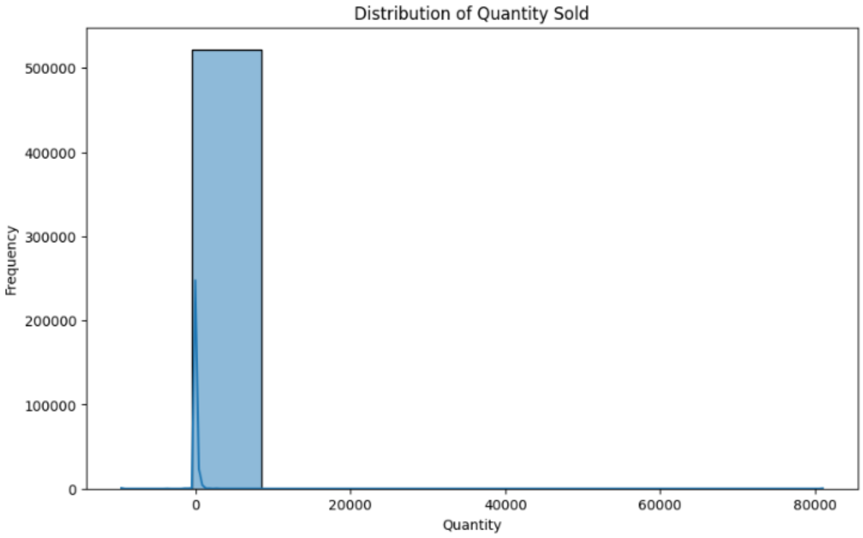
Key findings included identification of item pairs with high confidence and lift values, indicating strong associations between certain products.

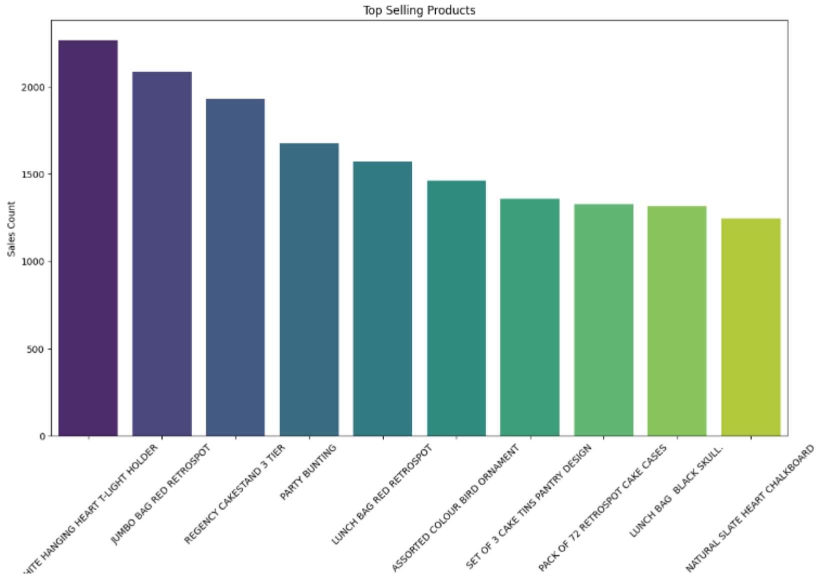
Visual Analysis

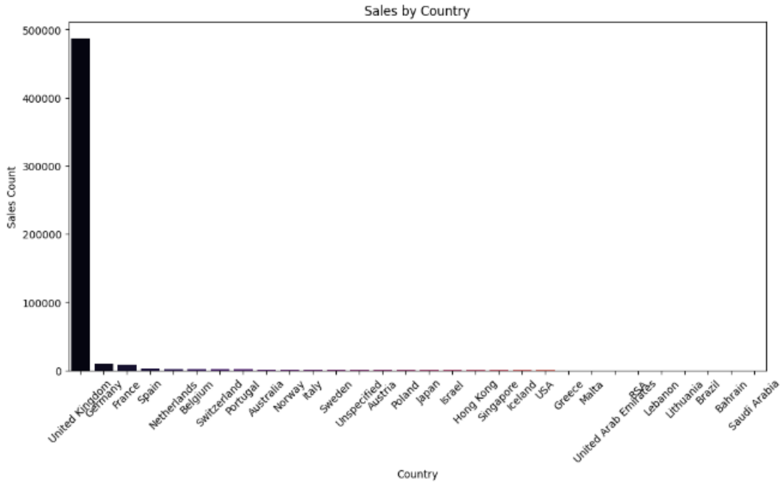
Several visualizations were used to illustrate the results:

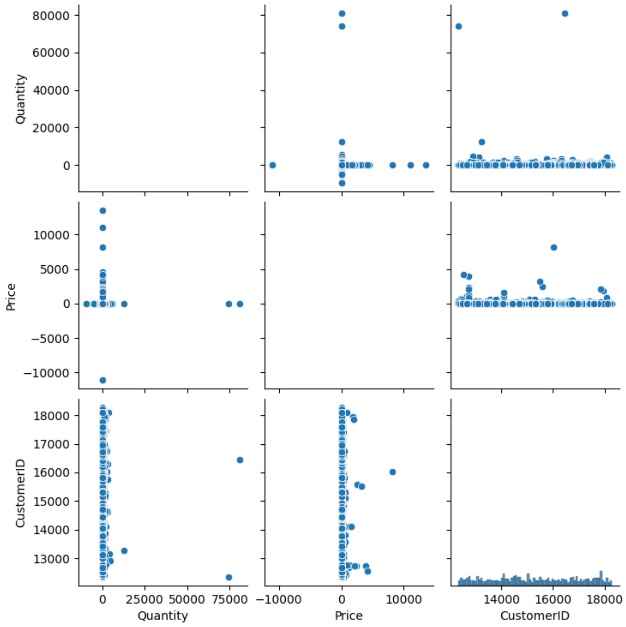
- **Heatmaps**: Showed the correlation between items, helping to identify which items are commonly purchased together.
- **Bar Charts**: Displayed the support values of the most frequent itemsets, giving a clear picture of the top combinations.
- **Scatter Plots**: Used to visualize the distribution of support vs. confidence for the derived association rules.
- These visuals help to quickly interpret the strength and reliability of various associations, providing actionable insights for decision-making.

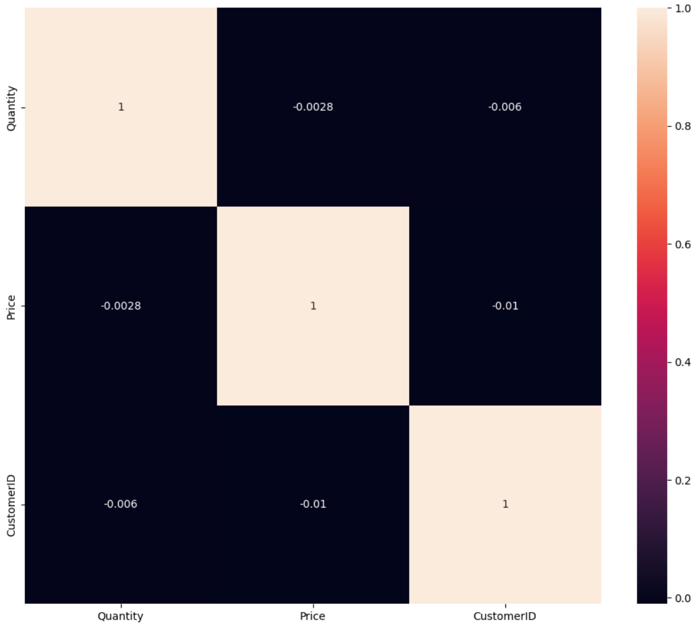
- **Top Selling Products**: A bar plot of the top 10 most frequently purchased items highlighted key products driving sales.
- **Quantity Sold**:
  - A distribution of the quantity sold per item revealed variations in purchase volumes.
  - A histogram of Quantity sold per item was created to visualize how often different quantities of items are sold.
- **Sales by Country**: Sales were also analyzed by country, showing how sales volume differs across regions.

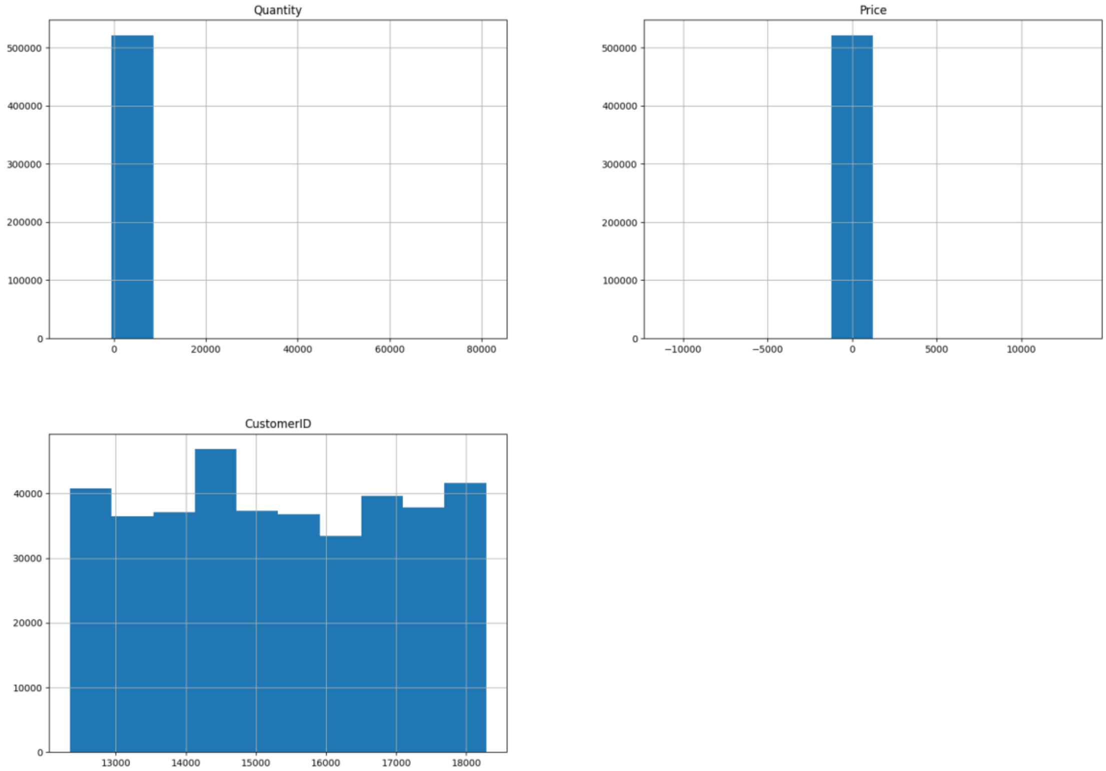

Distribution of Quantity Sold
Frequency
500000
400000
300000
200000
100000
0
0
20000
40000
60000
80000
Quantity


Top Selling Products
Sales Count
2000
1500
1000
500
0
HITE HANGING HEART T-LIGHT HOLDER
JUMBO BAG RED RETROSPOT
REGENCY CAKESTAND 3 TIER
PARTY BUNTING
LUNCH BAG RED RETROSPOT
ASSORTED COLOUR BIRD ORNAMENT
SET OF 3 CAKE TINS PANTRY DESIGN
PACK OF 72 RETROSPOT CAKE CASES
LUNCH BAG BLACK SKULL.
NATURAL SLATE HEART CHALKBOARD


Sales by Country
Sales Count
500000
400000
300000
200000
100000
0
United Kingdom
Germany
France
Spain
Netherlands
Belgium
Switzerland
Portugal
Australia
Norway
Italy
Sweden
Unspecified
Austria
Poland
Japan
Israel
Hong Kong
Singapore
Iceland
USA
Greece
Malta
United Arab Emirates
RSA
Lebanon
Lithuania
Brazil
Bahrain
Saudi Arabia
Country


Quantity
80000
60000
40000
20000
0
Price
10000
5000
0
−5000
−10000
CustomerID
18000
17000
16000
15000
14000
13000
0
25000
50000
75000
Quantity
−10000
0
10000
Price
14000
16000
18000
CustomerID


Quantity
Price
CustomerID
1
-0.0028
-0.006
-0.0028
1
-0.01
-0.006
-0.01
1
Quantity
Price
CustomerID
1.0
0.8
0.6
0.4
0.2
0.0


Quantity
500000
400000
300000
200000
100000
0
0
20000
40000
60000
80000
Price
500000
400000
300000
200000
100000
0
−10000
−5000
0
5000
10000
CustomerID
40000
30000
20000
10000
0
13000
14000
15000
16000
17000
18000

CONCLUSION

Market Basket Analysis helped uncover valuable insights into customer purchasing behavior. The Apriori algorithm identified frequent itemsets and association rules, which can be used to optimize store layout, recommend products, or design promotional offers. Visual analyses of top-selling items and country-wise sales further helped in understanding the broader trends in the dataset.

REFERENCES

Ahmedov, A. (2021) *Market Basket Analysis*. Kaggle. Available at: https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis (Accessed: 25 October 2024).