

CP3403 - Final Project
Analysis of Online Retail Sales Data

Kantapong Wongsanguan (14405427)
Pornnatcha Sarujikomjornwattana (14473015)
Yunsun Park (13980787)
Zambu Kyaw(13808179)

James Cook University Singapore

CP3403 Data Mining

Eric Tham

19, April, 2024

Table of Content

Abstract

1. Introduction
2. Business Scenario
 - 2.1 Business goals
 - 2.2 Expected outcome
3. Data Preparation
 - 3.1 Data used for the project
 - 3.2 Data description
 - 3.3 Data preprocessing
 - 3.3.1 Data Cleaning
 - 3.3.2 Data Transformation
 - 3.3.3 Categorization Process
 - 3.3.4 Data Reduction
 - 3.3.5 Data Refinement
4. Data Mining
 - 4.1 Association Rule Mining using Apriori algorithm
 - 4.1.1 ARM Model Selection
 - 4.1.2 Summary of output
 - 4.1.3 Limitation of selected model
 - 4.2 Data Visualization and Exploration
5. Recommendation for achieving business goals
6. Conclusion

References

Appendix

Abstract

This paper is part of the CP3403 Data Mining course at James Cook University Singapore. The main objective of this project is to apply data mining methods learned in class to real-world scenarios and make some suggestions to enhance business outcomes. The team chose to concentrate on online retail sales with the goal of leveraging data mining techniques to improve business outcomes. While the team has set specific goals for this project, it is purely for educational purposes and not tied to any particular organization. The primary focus is on analyzing customer buying habits using Association Rule Mining (ARM) to uncover interesting purchasing patterns. Additionally, we used clustering techniques to assist businesses in developing country-specific strategies and targeted marketing campaigns for annual events.

1. Introduction

The ability to analyze and interpret data effectively is becoming more than just a benefit in the retail industry. Data analytics is a key component of operational and strategic decisions in retail due to the continuous pace of market evolution and ongoing changes in consumer preferences.[2] Data mining, also known as knowledge discovery in data (KDD), is the process of extracting patterns, trends, relationships, and insightful information from massive amounts of data[4]. It is at the heart of this analytical revolution. Data mining facilitates the identification of significant relationships within large, complex data sets by utilizing a range of approaches and algorithms, including association rule mining, clustering, classification, and regression.

These methods are critical for gaining a deeper understanding of behaviors and preferences in a variety of industries, such as business, healthcare, finance, and marketing. In the realm of retail, specifically, data mining allows businesses to predict trends and market dynamics in addition to understanding customer behavior. Making informed suggestions and coming up with more effective planning strategies both depend on this predictive ability. As such, data mining is a powerful instrument that helps firms make better decisions while also guaranteeing that they can stay ahead of the competition by responding to both expected and unexpected shifts in the market.

2. Business Scenario

2.1 Business Goals

In today's digital world, online shopping is a big part of how we buy things. Every day, lots of people buy things online, which creates a huge amount of data. This data is both a great opportunity and a big challenge for online stores. Data mining is a way to make sense of this data. It helps businesses learn useful things from all the information available. [5]

As the owners of an online shop, we sell various products across different categories, to showcase and sell their products, while also offering customers a convenient way to make purchases. Over time, we have accumulated a wealth of data from our online retail sales, capturing valuable insights into customer behavior, purchasing patterns, and market trends. Recognizing the potential of this data, we are eager to leverage data mining techniques to extract meaningful insights that can guide our market planning strategies.

Currently, we are facing challenges with excessive stock levels and inconsistent monthly sales. Our sales tend to spike only during the last few months of the year, leading to inventory issues and missed opportunities for revenue generation. To address these issues and optimize our business operations, we are seeking the expertise of data science professionals to improve our stock management, boost sales performance, and achieve other beneficial business outcomes. By harnessing the power of data science, we aim to enhance our competitiveness, maximize profitability, and ensure sustainable growth in the dynamic online retail landscape.

Mass management of stocks challenges and missed revenue opportunities arise from the difficulties in controlling excessive stock levels and experiencing fluctuations in sales. In order to address these concerns and enhance overall business performance, we have established three key strategic objectives:

- **Customer Habit Analysis:**
Our goal is to identify common purchasing patterns among customers by applying Association Rule Mining (ARM). Through the suggestion of relevant product bundles, this analysis will enable us to more effectively tailor product recommendations and enhance customer engagement. This strategic initiative uses insights into frequently purchased item combinations to increase transaction value and customer satisfaction.

- **Country-Specific Recommendations:**

Through an analysis of product performance and customer preferences in various regions, we intend to create inventory strategies that are adapted to local necessities. With the help of top-selling products in important markets, this goal aims to increase sales and customer satisfaction.

- **Marketing Strategy for Yearly Events:**

Identifying key sales periods and analyzing historical sales trends will enable us to optimize our marketing efforts and inventory management strategically. By coordinating our marketing efforts and stock levels with projected consumer demand, especially during holidays and special occasions, we hope to optimize sales during periods of low consumer demand.

2.2 Expected Outcomes

The strategic application of data mining in our business operations is anticipated to lead to several beneficial outcomes:

- **Improved Stock Management:**

By gaining a deeper understanding of customer preferences, we will ensure that in-demand products are readily available while simultaneously reducing the inventory of less popular items. This approach will decrease storage costs and prevent stock shortages during critical sales periods, thereby optimizing our inventory turnover rate.

- **Increased Sales:**

Our targeted marketing and refined inventory management strategies, aligned with established customer demand patterns, are expected to significantly enhance sales volumes. Effective promotions and strategic stock placements during high-demand periods will exploit customer interest and buying potential.

- **Enhanced Customer Satisfaction:**

Aligning our product offerings with customer preferences and ensuring their availability during high-demand periods will markedly improve the shopping experience. This enhancement is likely to boost customer retention rates and promote brand loyalty, ultimately contributing to a stronger market presence and improved brand reputation.

3. Data Preparation

3.1 The dataset used for the report

The data used for this project is obtained from <https://archive.ics.uci.edu/dataset/502/online+retail+ii>

Chen, Daqing. (2019). Online Retail II. UCI Machine Learning Repository.

<https://doi.org/10.24432/C5CG6D>

3.2 Dataset description

Dataset name: Online Retail II

Instances: 1067371

Characteristic: Multivariate, Sequential, Time-Series, Text

Subject Area: Business

Dataset information

“This Online Retail II data set contains all the transactions occurring for a UK-based and registered, non-store online retail between 01/12/2009 and 09/12/2011. The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers.”

Attribute description

Attribute Name	Attribute Type	Description
InvoiceNO	Nominal	A 6-digit integral number uniquely assigned to each transaction. If this code starts with the letter 'c', it indicates a cancellation.
StockCode	Nominal	A 5-digit integral number uniquely assigned to each distinct product. Description: Product (item) name. Nominal.
Description	Numeric	Product (item) name.

Quantity	Numeric	The quantities of each product (item) per transaction.
InvoiceDate	Numeric	The day and time when a transaction was generated.
UnitPrice	Numeric	Product price per unit in sterling (£).
CustomerID	Nominal	A 5-digit integral number uniquely assigned to each customer.
Country	Nominal	The name of the country where a customer resides.

3.3 Data Preprocessing

This section outlines data preprocessing steps undertaken to prepare the source file, “Online Retail II” for further analysis, including clustering, classification, and association rule mining. Refer to the dataset description section above for the description of the source file.

The primary objective of this step is to clean and transform the dataset appropriately to facilitate effective data mining for pattern detection, classification for predictive analytics, and association rule mining for market basket analysis.

Tools Used

- Python: All stages of data manipulation, cleaning, and feature engineering.
- Pandas library: employed within the Python code for efficient data handling and operations.
- Weka: addition data manipulation specifically in data reduction (removal)
- Git: Version control and simple document sharing within team management.

The steps

***for full code for each snippet please refer to the appendix section below

3.3.1 Data Cleaning

Missing entries in ‘Description’ and CustomerID’ were identified and removed as specifically in these critical attributes, they represent incomplete records. As for error handling, transactions with negative quantities indicating returns and processing errors were identified and removed based on

analysis focus. Additionally, 'Description' attributes needed to be cleaned up since Weka had errors reading specific characters, i.e., '“', '“ “', etc.

```
13 def clean_description(desc):
14     if pd.isnull(desc):
15         return desc
16     desc = desc.replace('“', '')
17     desc = desc.replace('“ “', '')
18     desc = desc.replace('”', '')
19     desc = re.sub(pattern=r"['"]s", repl: "", desc, flags=re.IGNORECASE)
20     return desc
21
22 df['Description'] = df['Description'].apply(clean_description)
23
24 # Handle missing values
25 # numeric columns
26 numeric_columns = df.select_dtypes(include=[np.number]).columns.tolist()
27 numeric_columns.remove('Customer ID') # Exclude Customer ID from being filled with median
28 df[numeric_columns] = df[numeric_columns].fillna(df[numeric_columns].median())
```

Fig 1. Data cleaning 1

```
# Remove duplicates
df.drop_duplicates(inplace=True)

# Filter out negative quantities and prices
df = df[(df['Quantity'] > 0) & (df['Price'] > 0)]
```

Fig 2. Data_cleaning 2

3.3.2 Data Transformation

Data transformation of this dataset includes calculating a new attribute, 'TotalPrice' by multiplying 'Quantity' and 'Price' for each transaction to reflect the total transaction value of each item. Additionally, string attributes like 'InvoiceDate' needed to be converted from a string format into a DateTime format to facilitate time-based analysis. Additional transformation is needed to ensure all categorical data such as 'Country' is handled appropriately for analytical models.

Moreover, feature engineering was done as a part of data transformation to simplify the data. New time-related features from 'InvoiceDate' such as 'Year' and 'Month' were developed to analyze sales trends over time.

```
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])

# Feature engineering making 3 new attributes [Year, Month, TotalPrice]
df['Year'] = df['InvoiceDate'].dt.year
df['Month'] = df['InvoiceDate'].dt.month
df['TotalPrice'] = df['Quantity'] * df['Price']
```

Fig 3. Data transformation and feature engineering

3.3.3 Categorization process

This section specifically outlines the procedures used to preprocess and categorize item descriptions from the source dataset. The goal is to assign each item to a recognizable category based on keywords found in the item description.

Based on the most common words in the item descriptions, the categories analysis of the item descriptions are identified as several potential categories for the items.

```
5 df = pd.read_csv('online_retail/online_retail_II_clean_2011.csv')
6 df['Processed_Description'] = df['Description'].str.lower()
7 df['Processed_Description'] = df['Processed_Description'].apply(lambda x: re.sub(pattern=r'\W+', repl: ' ', x))
8 df['Words'] = df['Processed_Description'].str.split()
9
```



```
[('set', 58023), ('of', 49017), ('bag', 48315), ('red', 38703), ('heart', 35224), ('retrospot', 31368), ('vintage', 31115), ('design', 27247), ('pink', 26954), ('christmas', 21958), ('box', 21432), ('jumbo', 19899), ('cake', 19167), ('metal', 19082), ('white', 18646), ('blue', 17927), ('3', 17881), ('lunch', 17188), ('light', 16276), ('sign', 15614), ('hanging', 15301), ('t', 14829), ('6', 14588), ('pack', 14323), ('holder', 14310), ('card', 13245), ('paper', 13233), ('small', 12670), ('decoration', 11791), ('wooden', 11632), ('polkadot', 11330), ('tea', 11179), ('cases', 11063), ('glass', 10883), ('12', 10734), ('4', 10206), ('spaceboy', 9907), ('in', 9887), ('and', 9782), ('bottle', 9697), ('pantry', 9194), ('hot', 8787), ('home', 8784), ('with', 8698), ('large', 8607), ('paisley', 8586), ('tin', 8527), ('regency', 8301), ('rose', 8279), ('ceramic', 8195), ('green', 8163), ...]
```

Fig 4. Wordcounter & output

Thus, the categories were defined as the following code snippet.

```

7      if any(word in description for word in ['bag', 'lunch', 'jumbo']):
8          return 'Bags & Accessories'
9      elif any(word in description for word in ['vintage', 'retrospot', 'light', 'sign', 'hanging', 'decoration', 'wooden']):
10         return 'Home Decor'
11     elif any(word in description for word in ['card', 'paper', 'pack']):
12         return 'Stationery'
13     elif any(word in description for word in ['cake', 'box', 'holder']):
14         return 'Kitchenware'
15     elif 'christmas' in description:
16         return 'Christmas'

```

Fig 5. categorize.py

The ‘assign_category’ function is applied to each row in the DataFrame to assign a category based on the item’s description. This modifies the dataset by adding a new column attribute ‘Category’. Hence, this preprocessing and categorization process provides a structured way to understand and analyze the data based on the types of items it contains by utilizing simple text processing techniques and Pandas to assign item categories that are more meaningful for further analysis and reporting.

3.3.4 Data Reduction

The source dataset contains situational non-contributive attributes like ‘Invoice’ and ‘StockCode’ when their inclusion was not necessary for situational processes like clustering, classification, and associate rule mining. To reduce the amount of resources needed to process the data, in some cases data reduction is needed to focus the analysis on product-related characteristics and buying patterns rather than focusing on customer behavior for example. This process can be done within Weka for versatility in analysis.

The screenshot shows the 'Data Reduction' window in Weka. It features two panes. The left pane, titled 'No.' and 'Name', lists 12 attributes with checkboxes: Invoice (checked), StockCode (checked), Description (unchecked), Quantity (unchecked), InvoiceDate (checked and highlighted), Price (unchecked), Customer ID (checked), Country (checked), Year (unchecked), Month (unchecked), TotalPrice (unchecked), and Category (unchecked). A 'Remove' button is at the bottom. The right pane shows the selected attributes: Description, Quantity, Price, Year, Month (highlighted), TotalPrice, and Category.

Fig 6. Data reduction

3.3.5 Data Refinement

To refine a CSV file containing the dataset into an output suitable for use in the Weka rule associated with the Apriori algorithm for finding frequent itemsets and association rules. The file needs to be further preprocessed into a one-hot encoded matrix.

For each transaction, a one-hot encoded matrix is created, each row for each transaction and a column for each item purchased with binary '0' and '1' indicating the presence of each item in each transaction. A pandas DataFrame constructed as an 'Invoice' attribute is utilized as the first column to serve as a transaction ID (TID). This means that the output file includes binary attributes for each item category and the 'Invoice' ID as the first column.

Thus, the transactions are grouped by invoice ID and each entry corresponds to a transaction and the value is a list of all item categories bought within that transaction.
refer to the code snippet below.

```
8      # Create the one-hot encoded matrix
9      onehot = {}
10     invoices = transactions.index.tolist()
11
12     for idx, items in enumerate(transactions):
13         onehot[invoices[idx]] = {}
14         for item in items:
15             if pd.notnull(item):
16                 item = item.strip() # Clean the item
17                 onehot[invoices[idx]][item] = 1
18
```

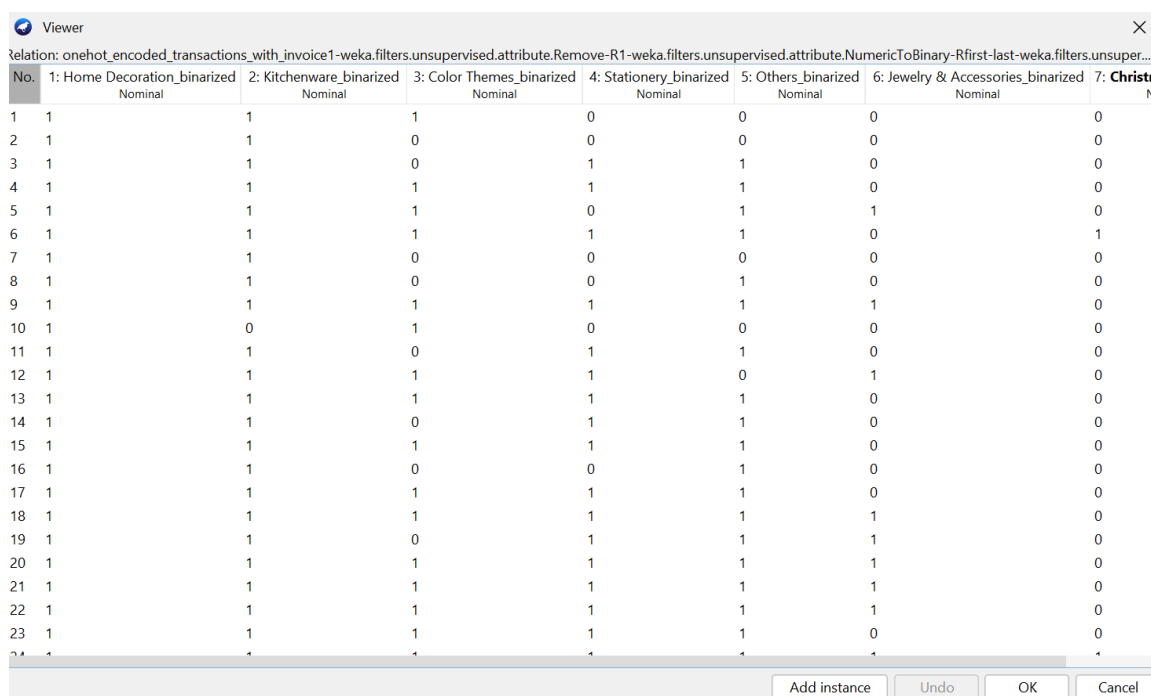
Fig 7. Binary matrix for apriori

4. Data Mining

4.1 Association Rule Mining

To satisfy business goals, our team decided to use ‘Association Rule Mining’ (ARM) or ‘Marketing Basket Analysis’. Association Rule Mining is one of the most popular data mining techniques that is widely used in organization. The primary goal of ARM is to identify frequent itemsets, which are combinations of items that appear together in transactions more often than would be expected by chance. Apriori is a preferred algorithm for this project. Here are the following steps:

Step 1: Check the data format and make sure that data is in binary.



The screenshot shows a 'Viewer' window with a table of data. The table has 7 columns: 'No.', '1: Home Decoration_binarized', '2: Kitchenware_binarized', '3: Color Themes_binarized', '4: Stationery_binarized', '5: Others_binarized', and '6: Jewelry & Accessories_binarized'. The 7th column is labeled '7: Christ' and has a 'N' in the first row. The data is binary, with values 0 or 1. The table is titled 'relation: onehot_encoded_transactions_with_invoice1-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.NumericToBinary-Rfirst-last-weka.filters.unsuper...'. At the bottom, there are buttons for 'Add instance', 'Undo', 'OK', and 'Cancel'.

No.	1: Home Decoration_binarized	2: Kitchenware_binarized	3: Color Themes_binarized	4: Stationery_binarized	5: Others_binarized	6: Jewelry & Accessories_binarized	7: Christ
1	1	1	1	0	0	0	0
2	1	1	0	0	0	0	0
3	1	1	0	1	1	0	0
4	1	1	1	1	1	0	0
5	1	1	1	0	1	1	0
6	1	1	1	1	1	0	1
7	1	1	0	0	0	0	0
8	1	1	0	0	1	0	0
9	1	1	1	1	1	1	0
10	1	0	1	0	0	0	0
11	1	1	0	1	1	0	0
12	1	1	1	1	0	1	0
13	1	1	1	1	1	0	0
14	1	1	0	1	1	0	0
15	1	1	1	1	1	0	0
16	1	1	0	0	1	0	0
17	1	1	1	1	1	0	0
18	1	1	1	1	1	1	0
19	1	1	0	1	1	1	0
20	1	1	1	1	1	1	0
21	1	1	1	1	1	1	0
22	1	1	1	1	1	1	0
23	1	1	1	1	1	0	0
24	1	1	1	1	1	1	1

Fig 8. Data table

Step 2: Set class to: No class

Step 3: Apply Apriori algorithm. In this stage, it is trial and error on adjusting minimum support and confidence. After a few attempts, we came up with the final rules.

Attempt 1:

Set support = 0.5 and lift = 1.1

```
=== Associator model (full training set) ===

Apriori
=====

Minimum support: 0.5 (20041 instances)
Minimum metric <lift>: 1.1
Number of cycles performed: 10

Generated sets of large itemsets:

Size of set of large itemsets L(1): 10
Size of set of large itemsets L(2): 9
Size of set of large itemsets L(3): 3

Best rules found:

1. Others_binarized=1 Home Decor_binarized=1 26813 ==> Kitchenware Items_binarized=1 22462    conf:(0.84) < lift:(1.19)> lev:(0.09) [3529] conv:(1.
2. Kitchenware Items_binarized=1 28301 ==> Others_binarized=1 Home Decor_binarized=1 22462    conf:(0.79) < lift:(1.19)> lev:(0.09) [3529] conv:(1.
3. Others_binarized=1 30012 ==> Home Decor_binarized=1 Kitchenware Items_binarized=1 22462    conf:(0.75) < lift:(1.16)> lev:(0.08) [3016] conv:(1.
4. Home Decor_binarized=1 Kitchenware Items_binarized=1 25969 ==> Others_binarized=1 22462    conf:(0.86) < lift:(1.16)> lev:(0.08) [3016] conv:(1.
5. Home Decor_binarized=1 33835 ==> Others_binarized=1 Kitchenware Items_binarized=1 22462    conf:(0.66) < lift:(1.13)> lev:(0.06) [2586] conv:(1.
6. Others_binarized=1 Kitchenware Items_binarized=1 23545 ==> Home Decor_binarized=1 22462    conf:(0.95) < lift:(1.13)> lev:(0.06) [2586] conv:(3.
7. Others_binarized=1 30012 ==> Kitchenware Items_binarized=1 23545    conf:(0.78) < lift:(1.11)> lev:(0.06) [2353] conv:(1.36)
8. Kitchenware Items_binarized=1 28301 ==> Others_binarized=1 23545    conf:(0.83) < lift:(1.11)> lev:(0.06) [2353] conv:(1.49)
```

Fig 9. Attempt 1

From the output, notice that every rule has an association with “Others”. Dealing with the "Others" category can be a bit tricky because it's a catch-all category that doesn't provide specific insights.

Attempt 2:

Now the “Others” category has been removed and tried to focus on confidence instead of lift.

Therefore we set support = 0.5, confidence = 0.9

```
1. Kitchenware_binarized=1 Color Themes_binarized=1 Stationery_binarized=1 19584 ==> Home Decoration_binarized=1 19045    <conf:(0.97)> lift:(1.15)
2. Color Themes_binarized=1 Jewelry & Accessories_binarized=1 19494 ==> Home Decoration_binarized=1 18794    <conf:(0.96)> lift:(1.14) lev:(0.05) [2
3. Color Themes_binarized=1 Stationery_binarized=1 22094 ==> Home Decoration_binarized=1 21231    <conf:(0.96)> lift:(1.13) lev:(0.05) [2519] conv:(
4. Kitchenware_binarized=1 Jewelry & Accessories_binarized=1 20479 ==> Home Decoration_binarized=1 19675    <conf:(0.96)> lift:(1.13) lev:(0.05) [23
5. Kitchenware_binarized=1 Color Themes_binarized=1 24857 ==> Home Decoration_binarized=1 23846    <conf:(0.96)> lift:(1.13) lev:(0.06) [2794] conv:
6. Stationery_binarized=1 Jewelry & Accessories_binarized=1 18966 ==> Home Decoration_binarized=1 18191    <conf:(0.96)> lift:(1.13) lev:(0.04) [212
7. Kitchenware_binarized=1 Color Themes_binarized=1 Christmas_binarized=0 19144 ==> Home Decoration_binarized=1 18209    <conf:(0.95)> lift:(1.12) l
8. Kitchenware_binarized=1 Stationery_binarized=1 25725 ==> Home Decoration_binarized=1 24467    <conf:(0.95)> lift:(1.12) lev:(0.05) [2680] conv:(3
9. Kitchenware_binarized=1 Stationery_binarized=1 Christmas_binarized=0 19808 ==> Home Decoration_binarized=1 18655    <conf:(0.94)> lift:(1.11) lev
10. Jewelry & Accessories_binarized=1 25330 ==> Home Decoration_binarized=1 23538    <conf:(0.93)> lift:(1.1) lev:(0.04) [2085] conv:(2.16)
```

Fig 10. Attempt 2

4.1.1 ARM Model Selection

The second model presents more useful information compared to the first model, therefore it will be the basis for developing our business goals. Moreover, this model produces a higher confidence level and the value of lift of every rule is greater than one. Key Matrix helps in making model selection:

- Confidence (conf): Measures the likelihood that the consequent occurs given the antecedent.

- Lift: Indicates the strength of a rule. A lift > 1 indicates that the antecedent and consequent appear together more often than expected by chance.

4.1.2 Summary of the output:

Scheme: Apriori algorithm

- Instances: 49,687
- Attributes = (Home Decoration, Kitchenware, Color Themes, Stationery, Jewelry & Accessories, Christmas)

Apriori Settings:

- Minimum Support: 0.35 (17,390 instances)
- Minimum Confidence: 0.9
- Number of Cycles: 13

Top 10 Association Rules:

{Kitchenware, Color Themes, Stationery} \rightarrow Home Decoration:

- Confidence: 97%
- Lift: 1.15

{Color Themes, Jewelry & Accessories} \rightarrow Home Decoration:

- Confidence: 96%
- Lift: 1.14

{Color Themes, Stationery} \rightarrow Home Decoration:

- Confidence: 96%
- Lift: 1.13

{Kitchenware, Jewelry & Accessories} \rightarrow Home Decoration:

- Confidence: 96%
- Lift: 1.13

{Kitchenware, Color Themes} \rightarrow Home Decoration:

- Confidence: 96%
- Lift: 1.13

{Stationery, Jewelry & Accessories} \rightarrow Home Decoration:

- Confidence: 96%
- Lift: 1.13

{Kitchenware, Stationery} \rightarrow Home Decoration:

- Confidence: 95%
- Lift: 1.12

{Jewelry & Accessories} → Home Decoration:

- Confidence: 93%
- Lift: 1.1

Above output clearly illustrates strong association rules to Home Decoration. When customers purchase two or more products from Kitchenware, Color Themes, Stationery, and Jewelry & Accessories, there's a likelihood over 93% that they will also purchase products in the Home Decoration category. Lift ranges between 1.1 to 1.15 show a strong relation between items like Kitchenware, Color Themes, Stationery, and Jewelry & Accessories with Home Decoration purchases. This means customers buying these items often also buy Home Decoration products. It's a useful insight for marketing and promoting complementary products to boost sales.

4.1.3 Limitation of the selected model

As the dataset grows larger, the algorithm's performance can decrease significantly due to its exhaustive search approach. In addition, Apriori doesn't consider the sequence in which items are purchased, which can be crucial in some retail scenarios. Lastly, the model only shows rules where all the relations point to Home Decoration which narrow useful aspects. The model might be biased or overly focused on one particular category.

4.2 Data Visualization and Data Exploration

To increase sales, simply offering bundles may not be efficient enough. At this data visualization and data exploration stage, we aim to identify additional insights which we can use further in market planning strategies. At this moment, we know that products from different categories should be bundled with products from Home Deco. However, it may not be efficient to sell the same bundles all year long. Now, we will want to visualize and explore the portion of each category in each month. To accomplish so, we used WEKA.

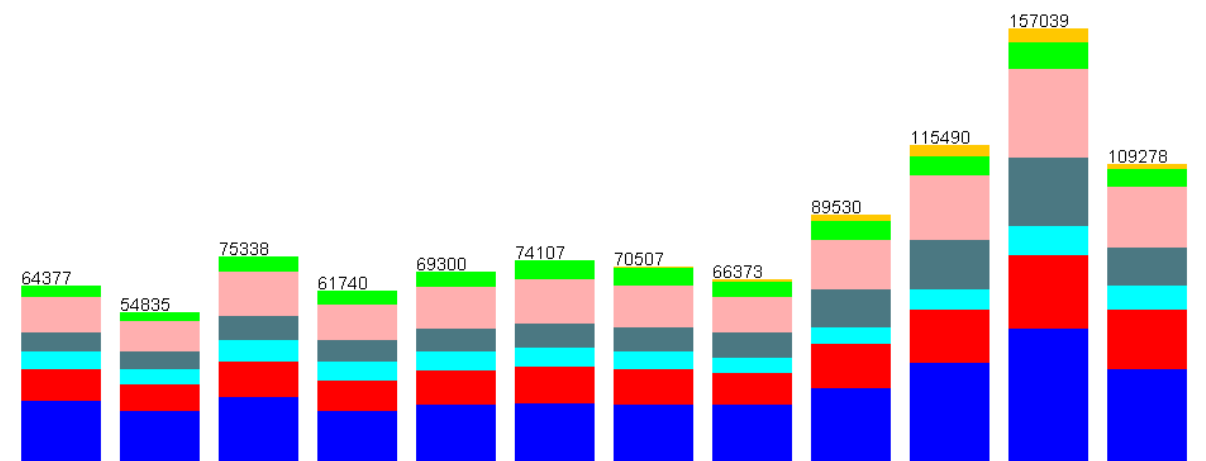


Fig 11. Bar chart from WEKA

These 12 bars illustrate how many products have been sold in each month from January to December accordingly and the color in each bar represents the category. . It is clear that November has the highest sales. Customers tend to purchase during the third quarter of the year. Moreover, to boost sales, the owner may want to offer sales or promotions during months with a low purchase rate along with recommending customers' preferred products determined by ARM.

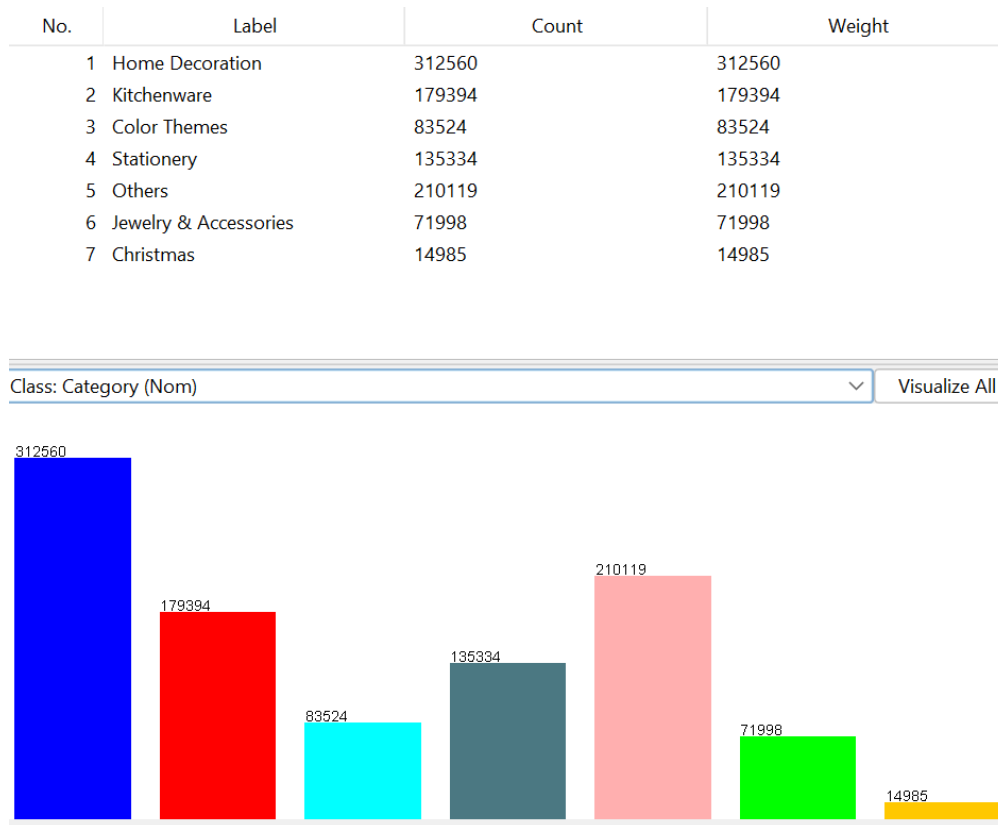


Fig 12. Category class from WEKA

Home decoration is in blue, kitchenware is in red, and so on. This chart simply illustrates the number of products purchased across different categories. We observe that Color Themes, Jewelry & Accessories, and Christmas items have lower purchase rates compared to Home Decoration and Kitchenware. Therefore, it's essential to promote these categories strategically. By referencing the first chart, we can determine the optimal timing for promotions. For instance, although Christmas products have the least sales, they are sold during the last four months of the year. Hence, stocking up on Christmas items during this period and promoting bundles that include Christmas products would increase the sales.

Data analyzed by top selling Countries

After conducting a comprehensive analysis of the data, particularly focusing on the geographic origin of our customers, we have identified the United Kingdom, Ireland, and Germany as our

top-performing markets in terms of sales volume. Our investigation revealed that these countries exhibit similar purchasing preferences, with a consistent interest in product categories such as Home Decoration, Kitchenware, and Stationery. Therefore, it is reasonable to infer that these top three countries share similar consumer buying habits, indicating potential opportunities for targeted marketing and product promotions across these regions.

Category (UK)	Average Price	Quantity	Total Sale
Home Decoration	3	3009826	9885906
Stationery	5	2064448	9607133
Kitchenware	4	1566144	6062565
Others	6	1213178	7103944
Jewelry & Accessories	3	774622	1966438
Color Themes	5	443011	2003662
Christmas	2	116525	222049

Fig 13. Data table of United Kingdom

Category (Ireland)	Average Price	Quantity	Total Sale
Home Decoration	3	104724	348636
Stationery	3	80608	214917
Kitchenware	4	66292	274211
Others	16	46978	737980
Color Themes	4	18934	78022
Jewelry & Accessories	2	14901	32902
Christmas	2	3892	6379

Fig 14. Data table of Ireland

Category (Germany)	Average Price	Quantity	Total Sale
Stationery	6	55203	303079
Kitchenware	3	51936	157213
Home Decoration	3	42327	119737
Others	4	36997	151804
Jewelry & Accessories	2	18792	34869
Color Themes	3	16528	55858
Christmas	2	3371	5236

Fig 15. Data table of Germany

Country	Count
United Kingdom	926039
IRELAND	17154
Germany	16432
France	13639

Fig 16. Quantity table sorted by Top 4 countries

In utilizing Weka for analysis, our findings from the J48 algorithm output indicate a predominant category, with the United Kingdom emerging as the primary focus.

```
J48 pruned tree
-----
: United Kingdom (1007914.0/81875.0)

Number of Leaves :      1

Size of the tree :      1

Time taken to build model: 115.46 seconds
```

Fig 17. J48 algorithm output

To expand our market and find new focus, with the help of research, found the focus for business.

Home Decoration should be promoted in India, Switzerland, Hong Kong. (Infographics: Which Countries Splurge the Most on Home Decor? (yahoo.com)

Stationery in China, Japan, Canada, Germany. &

(<https://finance.yahoo.com/news/global-stationery-products-market-poised-143200209.html>)

Kitchenware in U.S., China, and Germany.

(<https://www.statista.com/outlook/cmo/furniture/kitchen-dining-furniture/kitchenware/worldwide#global-comparison>)

To expand the market and achieve expected sales, India is a new market that needs to enter and other countries have our customer base only need to expand the market well.

China, India, and the U.S. have highest consumer market projections so it is also better to focus expanding marketing on these countries.




Rank ↕	Country	Consumer Market (2030 Projections)	% Change (from 2024)
1	 China	1,062,294,436	+15%
2	 India	772,929,623	+46%
3	 U.S.	348,393,863	+4%

Fig 18. Countries with top consumer markets [1]

5. Recommendation for achieving business goals

1. Strategic Product Bundling:

Themed Bundles: Create themed bundles that integrate Kitchenware, Color Themes, Stationery, and Home Decorations. For example, offer a "Home Office Refresh" bundle that includes stylish stationery items paired with home decor elements to appeal to customers interested in enhancing their workspace aesthetics.

Seasonal Bundles: Capitalize on seasonal trends by offering special bundles during holidays or seasonal changes. For instance, a "Winter Cozy" bundle could include warm color-themed decorations and kitchenware that resonate with the season's mood.

2. Targeted Promotions:

Email Campaigns: Use customer purchase history data to segment your audience and send targeted promotional emails. Customers who have shown an interest in Color Themes or Jewelry & Accessories could receive exclusive offers on these items when they are likely to complement a recent or frequent purchase of Home Decoration items.

Limited-Time Offers: Introduce time-limited promotions on bundles or individual products that encourage quick decision-making, such as "48-hour flash sale on kitchen and decor bundles."

3. Enhanced Cross-Selling Techniques:

Personalized Recommendations: Leverage machine learning algorithms to analyze customer purchase history and provide personalized product recommendations on the website and in the checkout process. For example, if a customer adds a home decoration item to their cart, they could immediately see a pop-up suggesting a matching color theme.

Point of Sale (POS) Suggestions: Train sales teams to suggest complementary products during the checkout process in physical stores or through online customer service interactions. For instance, when a customer buys a set of colorful vases, the system or salesperson can recommend matching table runners from the kitchenware range.

4. Geographic Expansion:

Market Analysis for Expansion: Identify and target international markets with similar buying habits. This approach leverages existing data to predict market behaviors in new regions, reducing the risk associated with entering new markets.

Localized Marketing Strategies: Once a target country is identified, tailor marketing efforts to resonate with local tastes and preferences while maintaining the brand's core identity. This could involve localized bundles or promotions that specifically cater to regional trends and cultural preferences.

6. Conclusion

The extensive analysis of the Online Retail II dataset has provided valuable insights into customer buying behaviors and market dynamics over the analyzed period. Our focused approach on data cleaning, transformation, and reduction facilitated a structured and insightful exploration into purchasing patterns, notably through the application of Association Rule Mining using the Apriori algorithm. The results have clearly outlined strong correlations between purchases in categories such as Kitchenware, Color Themes, Stationery, Jewelry & Accessories, and their association with Home Decoration purchases.

The data-driven findings have not only reinforced the importance of strategic category management but have also highlighted significant seasonal buying trends, which suggest opportunities for targeted marketing campaigns and promotional strategies. For instance, the increased transaction activity in the latter part of the year, particularly in November, presents a prime opportunity for maximizing sales through well-timed promotions and product bundling strategies.

Moreover, the geographical analysis has identified the United Kingdom, Ireland, and Germany as pivotal markets, providing a pathway to refine marketing efforts and product offerings specifically tailored to these regions. Considering the global consumer market trends, expanding into new markets like India, and further penetrating the U.S., China, and Japan, could strategically position the company to capture additional market share and enhance revenue streams.

As we move forward, the implementation of the recommended strategies such as product bundling, timed promotions, and market expansion should be pursued vigorously. These strategies are not only aligned with the current market analysis but also pave the way for sustainable growth and competitive advantage in the rapidly evolving online retail sector.

In conclusion, this project underscores the critical role of data analytics in understanding and leveraging consumer behavior to drive business decisions. With continued refinement and adaptation of our data analysis techniques, the potential for realizing enhanced operational efficiency and market responsiveness is substantial, ensuring that the company remains at the forefront of the online retail industry.

References

- [1]M. Lu, “The World’s Largest Consumer Markets in 2030,” *Visual Capitalist*, Feb. 08, 2024.
[Online]. Available: <https://www.visualcapitalist.com/the-worlds-largest-consumer-markets-in-2030/>
- [2]“The Advantages of Data-Driven Decision-Making | HBS Online,” *Business Insights Blog*, Aug. 26, 2019. [Online]. Available: <https://online.hbs.edu/blog/post/data-driven-decision-making>
- [3]K. Nikolopoulou, “What Is Data Mining? | Definition & Techniques,” *Scribbr*, Jul. 20, 2023.
[Online]. Available: <https://www.scribbr.com/ai-tools/data-mining/>
- [4]“UCI Machine Learning Repository.” [Online]. Available:
<https://archive.ics.uci.edu/dataset/352/online+retail>
- [5]A. Ghosh, “How Data Mining can help Organizations as well as Startups?,” Oct. 19, 2022.
[Online]. Available:
<https://www.linkedin.com/pulse/how-data-mining-can-help-organizations-well-startups-aritra-ghosh/>

Appendix

- cleanonlineretail.py

```
1 import pandas as pd
2 import re
3 import numpy as np
4
5 # Load CSV
6 csv_file_path = 'online_retail/online_retail_II.csv'
7 clean_csv_file_path = 'online_retail/online_retail_II_clean.csv'
8
9 # Read CSV into DataFrame
10 df = pd.read_csv(csv_file_path)
11
12 # Clean Description column
13 usage  KantapongWongJC14405427
14 def clean_description(desc):
15     if pd.isnull(desc):
16         return desc
17     desc = desc.replace('"""', '')
18     desc = desc.replace('"""', '')
19     desc = desc.replace('"""', '')
20     desc = re.sub(pattern=r"['"]s", repl: "", desc, flags=re.IGNORECASE)
21     return desc
22
23 df['Description'] = df['Description'].apply(clean_description)
24
25 # Handle missing values
26 # numeric columns
27 numeric_columns = df.select_dtypes(include=[np.number]).columns.tolist()
28 numeric_columns.remove('Customer ID') # Exclude Customer ID from being filled with median
29 df[numeric_columns] = df[numeric_columns].fillna(df[numeric_columns].median())
30
31 # categorical columns
32 categorical_columns = df.select_dtypes(include=['object', 'category']).columns.tolist()
33 df[categorical_columns] = df[categorical_columns].fillna(df[categorical_columns].mode().iloc[0])
```



```

34 # Remove duplicates
35 df.drop_duplicates(inplace=True)
36
37 # Filter out negative quantities and prices
38 df = df[(df['Quantity'] > 0) & (df['Price'] > 0)]
39
40 # Convert data types
41 df['Invoice'] = df['Invoice'].astype(str)
42 df['StockCode'] = df['StockCode'].astype(str)
43 df['Customer ID'] = df['Customer ID'].astype(str)
44 df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'])
45
46 # Feature engineering making 3 new attributes [Year, Month, TotalPrice]
47 df['Year'] = df['InvoiceDate'].dt.year
48 df['Month'] = df['InvoiceDate'].dt.month
49 df['TotalPrice'] = df['Quantity'] * df['Price']
50
51 # Save
52 df.to_csv(clean_csv_file_path, index=False)
53 print(f"Data cleaned and transformed, saved to {clean_csv_file_path}")
54
55 # Testing by printing specific lines for verification
56
57 with open(clean_csv_file_path, 'r', encoding='utf-8') as file:
58     for i, line in enumerate(file, 1):
59         if 151 <= i <= 155:
60             print(f'Line {i}: {line.rstrip()}')
61         elif i > 155:
62             break
63

```

- wordcount.py

```

1 import pandas as pd
2 from collections import Counter
3 import re
4
5 df = pd.read_csv('online_retail/online_retail_II_clean_2011.csv')
6 df['Processed_Description'] = df['Description'].str.lower()
7 df['Processed_Description'] = df['Processed_Description'].apply(lambda x: re.sub(pattern: r'\W+', repl: ' ', x))
8 df['Words'] = df['Processed_Description'].str.split()
9
10 all_words = [word for words_list in df['Words'] for word in words_list]
11 word_counts = Counter(all_words)
12
13
14 print(word_counts.most_common(100))
15

```

- categorize.py

```

1 import pandas as pd
2 df = pd.read_csv('online_retail/online_retail_II_clean_2011.csv')
3 usage new *
4 def assign_category(description): # categories and keywords
5     if pd.isna(description):
6         return 'Unknown'
7     description = description.lower()
8     if any(word in description for word in ['bag', 'lunch', 'jumbo']):
9         return 'Bags & Accessories'
10    elif any(word in description for word in ['vintage', 'retrospot', 'light', 'sign', 'hanging', 'decoration', 'wooden']):
11        return 'Home Decor'
12    elif any(word in description for word in ['card', 'paper', 'pack']):
13        return 'Stationery'
14    elif any(word in description for word in ['cake', 'box', 'holder']):
15        return 'Kitchenware'
16    elif 'christmas' in description:
17        return 'Christmas'
18    elif any(word in description for word in ['heart', 'metal', 'wooden']):
19        return 'Jewelry & Accessories'
20    elif any(word in description for word in ['red', 'pink', 'white', 'blue']):
21        return 'Color Themes'
22    else:
23        return 'Others'
24 df['Category'] = df['Description'].apply(assign_category)
25 df.to_csv(path_or_buf='online_retail/categorized_items.csv', index=False)
26 print("Categories assigned and file saved as 'categorized_items.csv'")
27

```

- ap.py

```

1
2 import pandas as pd
3 df = pd.read_csv('online_retail/categorized_items.csv')
4
5 # Group by 'Invoice' and aggregate the items in 'Category'
6 transactions = df.groupby('Invoice')['Category'].apply(list)
7
8 # Create the one-hot encoded matrix
9 onehot = {}
10 invoices = transactions.index.tolist()
11
12 for idx, items in enumerate(transactions):
13     onehot[invoices[idx]] = {}
14     for item in items:
15         if pd.notnull(item):
16             item = item.strip() # Clean the item
17             onehot[invoices[idx]][item] = 1
18
19 # dict into a DataFrame
20 onehot_df = pd.DataFrame.from_dict(onehot, orient='index').fillna(0)
21
22 # Reset index and push Invoice
23 onehot_df.reset_index(inplace=True)
24 onehot_df.rename(columns={'index': 'Invoice'}, inplace=True)
25
26 # Save to CSV
27 output_csv_path = 'onehot_encoded_transactions_with_invoice.csv'
28 onehot_df.to_csv(output_csv_path, index=False)
29

```