DATA ANALYSIS

COURSE PRESENTER
(DR. Omaimah)

Online Retail project

STUDENT	ID
SHAHAD AMER	444005703
AREEJ TALEB	444002403

DEPARTMENT OF (INFORMATION SCIENCE-DATA SCIENCE)
COLLEGE OF COMPUTER AND INFORMATION SYSTEMS
UMM AL-QURA UNIVERSITY

2024

INTRODUCTION

Online retail datasets are collections of transactional data generated by e-commerce platforms. This dataset captures a wide range of information related to customer purchases, product details, and sales trends. They serve as valuable resources for analyzing consumer behaviour and informing business strategies.

Key columns in the data:

- **InvoiceNo(InvoiceNumber):**
  A unique number for each sales invoice. If the number starts with "C," it means it's a return or cancellation.
- **StockCode(ProductCode):**
  The code that uniquely identifies each product.
- **Description:**
  A simple text description of the product.
- **Quantity:**
  The number of units sold for each product in the invoice. This can be negative if the product was returned.
- **InvoiceDate(InvoiceDate):**
  The date and time the invoice was generated.
- **UnitPrice(UnitPrice):**
  The price per unit of each product.
- **CustomerID:**
  A unique identifier for each customer.
- **Country:**
  The country where the customer is located.

OBJECTIVE

The goal of using this dataset is to apply **Market Basket Analysis** to identify which products are frequently chosen and purchased together. Through this analysis, we can extract **rules** that link these products and uncover patterns in customer behaviour. This information helps improve marketing strategies by recommending related products, enhancing the customer experience

DATA PREPARATION

There are important steps to clean and prepare a retail dataset for analysis. Here's a clear explanation about what we did:

Dataset info: first examined the dataset to understand its structure, including the number of entries and the types of information available.

Column Renaming: rename the columns to make them clearer. For instance, changed 'StockCod' to 'product_code' and 'UnitPrice' to 'price_per_unit'.

Whitespace Removal: cleaned up the 'product_description' column by removing any extra spaces at the beginning or end of the text. This ensures that all descriptions are formatted consistently.

Duplicate Removal: eliminated duplicate entries from the dataset, ensuring that each transaction is represented only once.

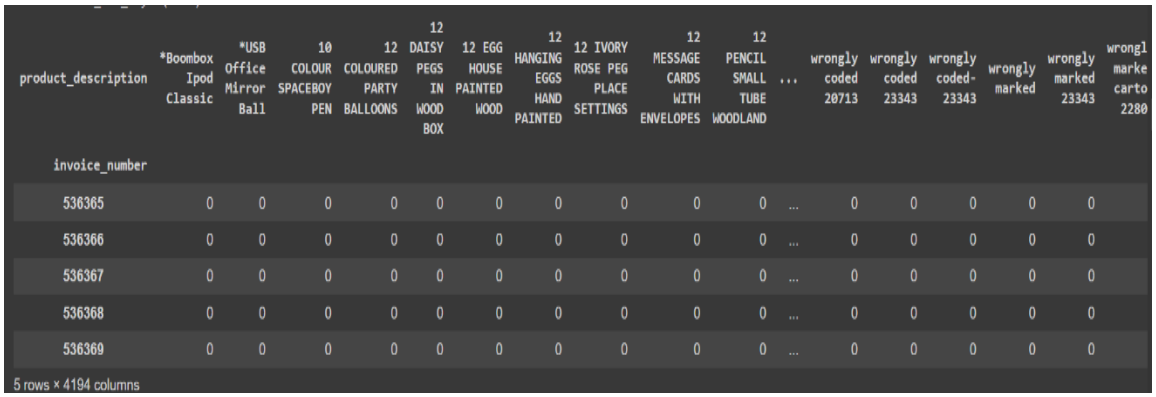
Missing Value Handling: removed any rows that had missing values in the 'invoice_number' column. Since this field is essential for identifying transactions, it's important to have complete data here.

Data Type Conversion: changed the 'invoice_number' column to a string format. This is typically done to treat these identifiers correctly, as they are not used for mathematical operations.

Filtering Records: filtered out any transactions that had an 'invoice_number' containing the letter 'C', which likely indicates canceled transactions. This helps in focusing on valid transactions only.

MARKET BASKET ANALYSIS

In this step, we transformed the data into a binary matrix to facilitate the application of **Market Basket Analysis**. The goal of this process is to reorganize the data so that each invoice is represented as a row, each product as a column, and the values are 1 if the product was purchased in the invoice and 0 if it was not.
The code performs the following steps:
Group the data by **invoice number** and **product description** to determine the total **quantities sold** for each product in each invoice.
Reshape the data so that each product becomes a separate column.
Fill missing values with 0, indicating that the product was not purchased in that invoice.
Convert the values to integers (int) and apply a function to assign 1 if the quantity is greater than zero, and 0 if the product was not purchased.
The result is a binary matrix that can be used to extract **rules** that help identify products frequently bought together. This supports more effective marketing decisions, such as product recommendations



product_description	*Boombox Ipod Classic	*USB Office Mirror Ball	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	...	wrongly coded 20713	wrongly coded 23343	wrongly coded- 23343	wrongly marked	wrongly marked 23343	wrongl marke carto 2280
invoice_number																	
536365	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
536366	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
536367	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
536368	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	
536369	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	

5 rows × 4194 columns

Frequent Itemsets

In this step, we utilized the **FPGrowth** algorithm to identify frequent itemsets from our prepared basket data,The aim of this analysis is to discover products that are often purchased together, which forms the foundation of Market Basket Analysis.
Note:
The results were sorted based on the highest support values, highlighting the products that occur most frequently in purchase transactions.

```
warnings.warn(
```

	support	itemsets
0	0.109661	(WHITE HANGING HEART T-LIGHT HOLDER)
92	0.101509	(JUMBO BAG RED RETROSPOT)
294	0.096511	(REGENCY CAKESTAND 3 TIER)
574	0.081809	(PARTY BUNTING)
42	0.075889	(LUNCH BAG RED RETROSPOT)
7	0.070600	(ASSORTED COLOUR BIRD ORNAMENT)
603	0.067204	(SET OF 3 CAKE TINS PANTRY DESIGN)
43	0.064050	(PACK OF 72 RETROSPOT CAKE CASES)
172	0.061769	(LUNCH BAG BLACK SKULL.)
79	0.060605	(NATURAL SLATE HEART CHALKBOARD)

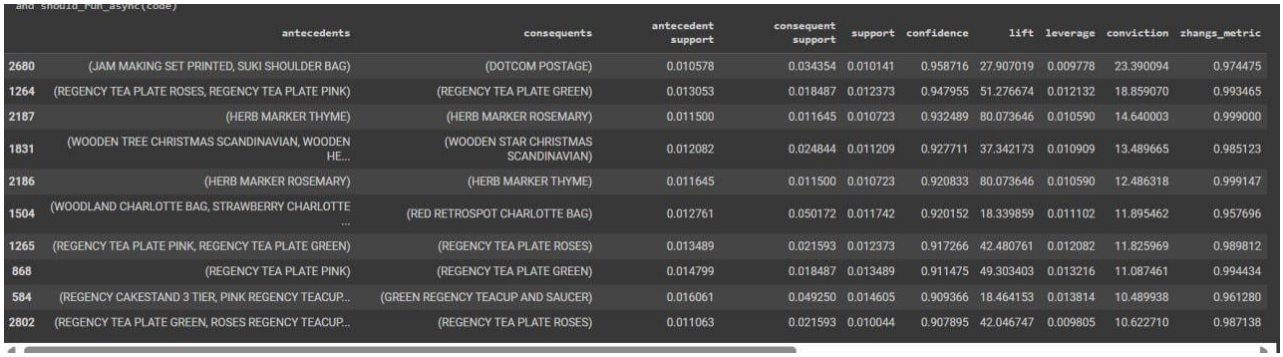
Results and Insights:
The output revealed that certain products are frequently bought together, with notable examples including:

- **WHITE HANGING HEART T-LIGHT HOLDER** with a support of 10.96%.
- **JUMBO BAG RED RETROSPOT** with a support of 10.15%.
- **REGENCY CAKESTAND 3 TIER** with a support of 9.65%.

These products are frequently observed in purchase transactions, indicating their high popularity among customers, Based on these findings, this data can be leveraged to enhance marketing strategies by recommending these items to customers or by offering promotional deals on related products

Association rules

In this step, we used the association rules function to extract rules from the frequent item sets, aiming to uncover relationships between products. Specifically, we wanted to determine how likely the purchase of one product (the antecedent) leads to the purchase of another (the consequent). By setting the metric to "lift" with a minimum threshold of 1, we identified rules indicating strong associations. The output includes important metrics such as support, confidence, and lift, which help evaluate the strength and significance of these relationships.



	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
2680	(JAM MAKING SET PRINTED, SUKI SHOULDER BAG)	(DOTCOM POSTAGE)	0.010578	0.034354	0.010141	0.958716	27.907019	0.009778	23.390094	0.974475
1264	(REGENCY TEA PLATE ROSES, REGENCY TEA PLATE PINK)	(REGENCY TEA PLATE GREEN)	0.013053	0.018487	0.012373	0.947955	51.276674	0.012132	18.859070	0.993465
2187	(HERB MARKER THYME)	(HERB MARKER ROSEMARY)	0.011500	0.011645	0.010723	0.932489	80.073646	0.010590	14.640003	0.999000
1831	(WOODEN TREE CHRISTMAS SCANDINAVIAN, WOODEN HE...	(WOODEN STAR CHRISTMAS SCANDINAVIAN)	0.012082	0.024844	0.011209	0.927711	37.342173	0.010909	13.489665	0.985123
2186	(HERB MARKER ROSEMARY)	(HERB MARKER THYME)	0.011645	0.011500	0.010723	0.920833	80.073646	0.010590	12.486318	0.999147
1504	(WOODLAND CHARLOTTE BAG, STRAWBERRY CHARLOTTE ...	(RED RETROSPOT CHARLOTTE BAG)	0.012761	0.050172	0.011742	0.920152	18.339859	0.011102	11.895462	0.957696
1265	(REGENCY TEA PLATE PINK, REGENCY TEA PLATE GREEN)	(REGENCY TEA PLATE ROSES)	0.013489	0.021593	0.012373	0.917266	42.480761	0.012082	11.825969	0.989812
868	(REGENCY TEA PLATE PINK)	(REGENCY TEA PLATE GREEN)	0.014799	0.018487	0.013489	0.911475	49.303403	0.013216	11.087461	0.994434
584	(REGENCY CAKESTAND 3 TIER, PINK REGENCY TEACUP...	(GREEN REGENCY TEACUP AND SAUCER)	0.016061	0.049250	0.014605	0.909366	18.464153	0.013814	10.489938	0.961280
2802	(REGENCY TEA PLATE GREEN, ROSES REGENCY TEACUP...	(REGENCY TEA PLATE ROSES)	0.011063	0.021593	0.010044	0.907895	42.046747	0.009805	10.622710	0.987138

Additional Note

The results are sorted by confidence, enabling us to prioritize the most reliable rules based on the likelihood of one product leading to another.

Results and Insights: The output refers to a set of association rules between products, Let's discuss higher3

1. invoice number(2680):

- **Antecedents: (JAM MAKING SET PRINTED, SUKI SHOULDER BAG)**
- **Consequent: (DOTCOM POSTAGE)**
- **Confidence: 95.87%**
- **Lift: 27.91**

Interpretation:

This result indicates a strong correlation between purchasing the jam making set and the SUKI shoulder bags, leading to the purchase of DOTCOM postage. The high confidence suggests that most customers who bought these items also bought the stamps, presenting an opportunity to boost sales through joint promotional offers.

2. invoice number(1264):

- **Antecedents:** (REGENCY TEA PLATE ROSES, REGENCY TEA PLATE PINK)
- **Consequent:** (REGENCY TEA PLATE GREEN)
- **Confidence:** 94.80%
- **Lift:** 51.28

Interpretation:

This result shows that customers who buy the rose or pink tea plates are likely to purchase the green tea plates as well. The high lift value indicates a strong association, suggesting the potential to enhance sales by promoting a variety of tea plates in different colors.

---

3. invoice number(2187):

- **Antecedents:** (HERB MARKER THYME)
- **Consequent:** (HERB MARKER ROSEMARY)
- **Confidence:** 93.25%
- **Lift:** 80.07

Interpretation:

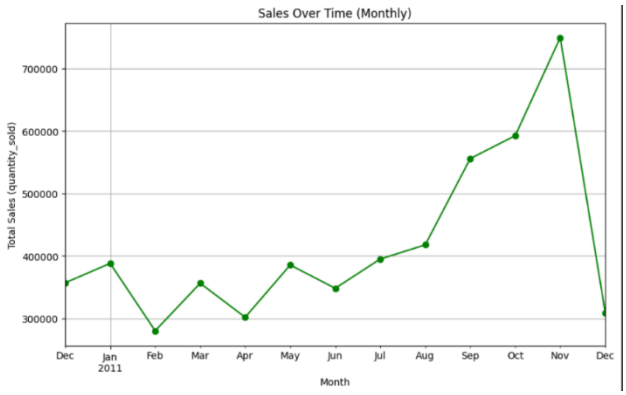
This rule indicates that customers purchasing the thyme herb marker are also likely to buy the rosemary herb marker. The high lift value reflects a strong opportunity for cross-promotional strategies, which can increase sales by bundling related products together.

Importance of the Results:

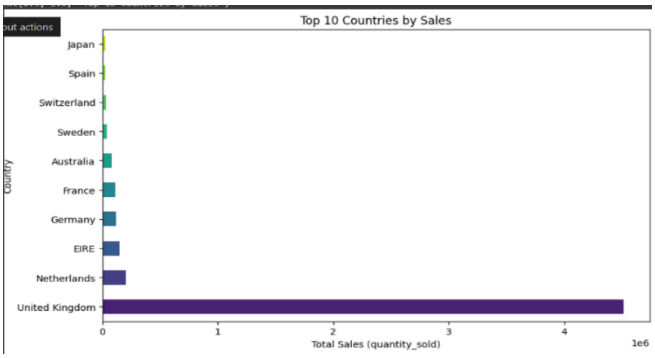
These results allow us to identify customer behavioral patterns, enabling us to enhance our marketing strategies through:

- Providing personalized recommendations to customers based on their previous purchases.
- Developing targeted promotional campaigns that include related products.
- Optimizing inventory planning to ensure the availability of popular products that are frequently purchased together.

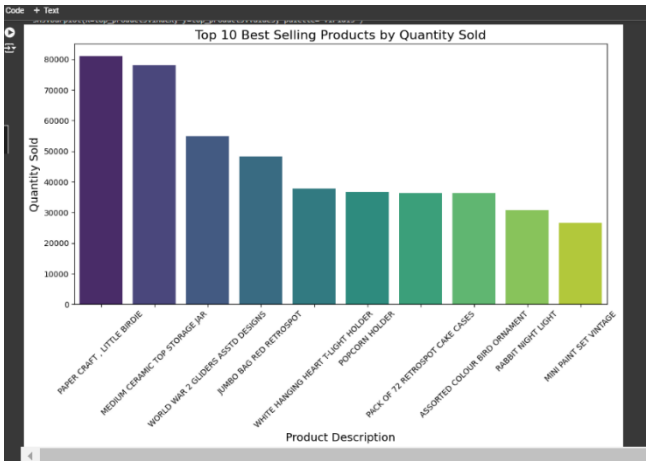
VISUALISATION



It shows the sales over time (month), the highest sales were in November.



It shows the top 10 country by sales, and the best-selling country was United Kingdom.



It shows the top 10 selling products by quantity sold, the most selling product was paper craft, little bride.