E-Commerce Data

COURSE PRESENTER:
(DR.Omaima A. Fallatah)

SUBMITTED BY:

Name	ID
Layan Adel Babkour	444002368
Reham Faisal Alsubhi	444003014

DEPARTMENT OF(INFORMATION SCIENCE-DATA SCIENCE)

COLLEGE OF COMPUTER AND INFORMATION SYSTEMS

Table of content

1.Introduction

This dataset focuses on analyzing a dataset containing real e-commerce transactions, sourced from the UCI Machine Learning Repository. The data includes all transactions from December 2010 to December 2011 for a UK-based online retail company that specializes in selling unique gifts for various occasions, with wholesalers making up a significant portion of its customer base. The dataset includes details such as invoice number, product code, description, quantity sold, invoice date, unit price, customer ID, and the country of the customer. This data will be utilized to analyze purchasing behaviors, sales trends, and transaction patterns to uncover insights that can help optimize business strategies.

1.1. The our goals

- Use association rule learning techniques to perform Market Basket Analysis.
- Customer behavior analysis and understanding their preferences

1.2.Link dataset

E-Commerce Data

2.Exploratory Data Analysis

Attribute about dataset

-InvoiceNo : Invoice number corresponding to the product purchase.
-StockCode : Identifier of the purchased product. Each identifier is different.
-Description : Description of the purchased product.
-Quantity : Quantity of product purchased
-InvoiceDate : Date of invoice, from 01/12/2010 to 09/12/2011 .
-UnitPrice : Price of one product.
-CustomerID : Identifier of the customer. Each identifier is different.
-Country : Country where the customer places the order.

Datasets

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

1. *Figure*: *Displays the Dataset.*

Libraries used

```
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
import networkx as nx
import seaborn as sns
```

2. *Figure*: *Shows the dataset libraries.*

```
Number of rows: 541909, number of columns: 8
```

3. *Figure: Shows the number of features.*

This code displays the number of columns and rows in the dataset.

- Number of Columns:8
- Number of rows:541909

	Name	dtypes	Missing	Missing_%	Uniques	First Row	Last Row
0	InvoiceNo	object	0	0.000000	25900	536365	581587
1	StockCode	object	0	0.000000	4070	85123A	22138
2	Description	object	1454	0.268311	4223	WHITE HANGING HEART T-LIGHT HOLDER	BAKING SET 9 PIECE RETROSPOT
3	Quantity	int64	0	0.000000	722	6	3
4	InvoiceDate	object	0	0.000000	23260	12/1/2010 8:26	12/9/2011 12:50
5	UnitPrice	float64	0	0.000000	1630	2.55	4.95
6	CustomerID	float64	135080	24.926694	4372	17850.0	12680.0
7	Country	object	0	0.000000	38	United Kingdom	France

4. *Figure: Representing a wide range of transactions with both textual and numerical data.*

The DataFrame contains , representing a wide range of transactions with both textual and numerical data. Some columns provide identifiers or descriptions, while others include quantities or prices. This variety captures product and transaction diversity across different countries, helping us understand the data distribution and address missing values for more accurate analysis.

	Quantity	UnitPrice	CustomerID
count	541909.00	541909.00	406829.00
mean	9.55	4.61	15287.69
std	218.08	96.76	1713.60
min	-80995.00	-11062.06	12346.00
25%	1.00	1.25	13953.00
50%	3.00	2.08	15152.00
75%	10.00	4.13	16791.00
max	80995.00	38970.00	18287.00

5. *Figure: Shows describe function in our data*

The df.describe() function in the Pandas library is used to obtain a Basic Statistics

Before cleaning

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
622	536414	22139	NaN	56	12/1/2010 11:52	0.00	NaN	United Kingdom
1443	536544	21773	DECORATIVE ROSE BATHROOM BOTTLE	1	12/1/2010 14:32	2.51	NaN	United Kingdom
1444	536544	21774	DECORATIVE CATS BATHROOM BOTTLE	2	12/1/2010 14:32	2.51	NaN	United Kingdom
1445	536544	21786	POLKADOT RAIN HAT	4	12/1/2010 14:32	0.85	NaN	United Kingdom
1446	536544	21787	RAIN PONCHO RETROSPOT	2	12/1/2010 14:32	1.66	NaN	United Kingdom

6. *Figure*: *Representing data before cleaning*

The code cleans the data by identifying rows with missing values, then filters the dataset to remove rows with negative quantities or missing values in the "CustomerID" column, and finally displays a sample of the cleaned data

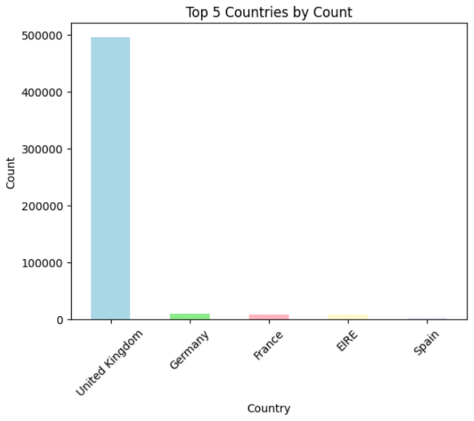
After cleaning:

```
  InvoiceNo StockCode                          Description  Quantity  \
0    536365    85123A   WHITE HANGING HEART T-LIGHT HOLDER         6
1    536365     71053                  WHITE METAL LANTERN         6
2    536365    84406B       CREAM CUPID HEARTS COAT HANGER         8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE         6
4    536365    84029E       RED WOOLLY HOTTIE WHITE HEART.         6

      InvoiceDate  UnitPrice  CustomerID         Country
0  12/1/2010 8:26       2.55     17850.0  United Kingdom
1  12/1/2010 8:26       3.39     17850.0  United Kingdom
2  12/1/2010 8:26       2.75     17850.0  United Kingdom
3  12/1/2010 8:26       3.39     17850.0  United Kingdom
4  12/1/2010 8:26       3.39     17850.0  United Kingdom
```

7. *Figure*: *Representing data after cleaning*

Country	
United Kingdom	495478
Germany	9495
France	8557
EIRE	8196
Spain	2533



8. *Figure*: *Representing the top five countries with the highest occurrence*

This chart displays the top five countries with the highest occurrence, helping us understand which country is making the most purchases from us.

```
basket = basket.applymap(encode_units)
StockCode  10002  10080  10120  10123C  10124A  10124G  10125  10133  10135  \
InvoiceNo
536365         0      0      0       0       0       0      0      0      0
536366         0      0      0       0       0       0      0      0      0
536367         0      0      0       0       0       0      0      0      0
536368         0      0      0       0       0       0      0      0      0
536369         0      0      0       0       0       0      0      0      0

StockCode  11001  ...  90214Y  90214Z  BANK CHARGES  C2  CRUK  D  DOT  M  \
InvoiceNo         ...
536365         0  ...       0       0             0   0     0  0    0  0
536366         0  ...       0       0             0   0     0  0    0  0
536367         0  ...       0       0             0   0     0  0    0  0
536368         0  ...       0       0             0   0     0  0    0  0
536369         0  ...       0       0             0   0     0  0    0  0

StockCode  PADS  POST
InvoiceNo
536365        0     0
536366        0     0
536367        0     0
536368        0     0
536369        0     0
```

9. *Figure*: *Display Transaction data in a suitable format for basket analysis*

This result shows me that the purpose of this code is to structure transaction data in a suitable format for basket analysis, where each row represents an invoice, each column represents a product, and the value (1 or 0) indicates whether the product was purchased in that invoice. This data can then be used in techniques such as Association Rules to extract common purchasing patterns between products.

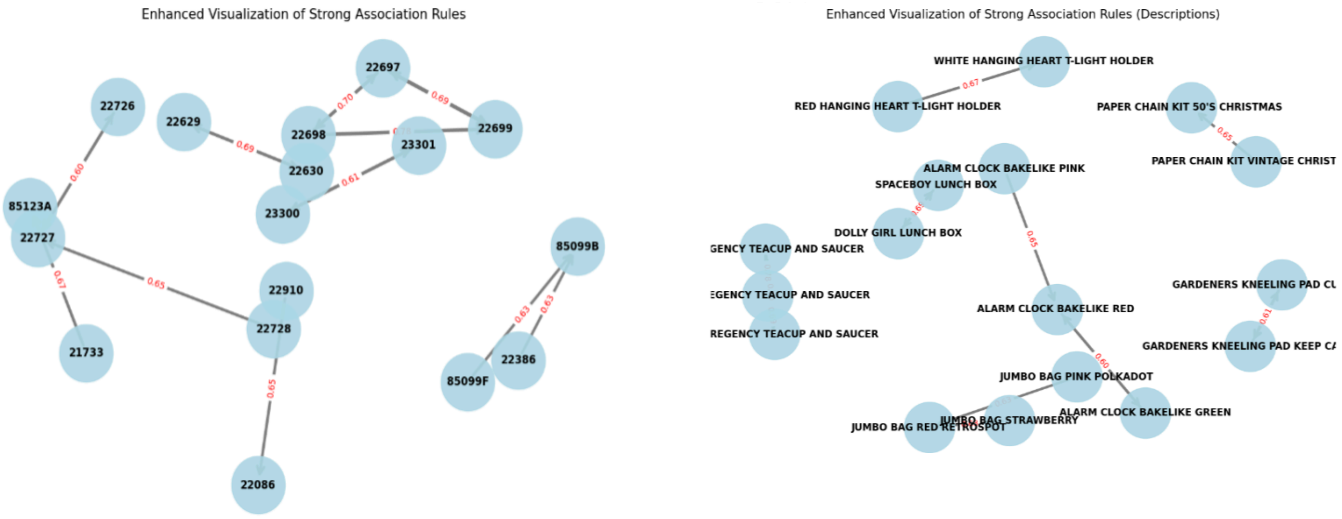
3.Implement Basket Analysis Algorithm:

FP-Growth algorithm

Using the FP-Growth algorithm, we calculated the percentage of transactions that contain the product or product group and represented it in association rules. The results showed the values of support, confidence, and lift.

4.Calculate Key Metrics:

Antecedent Description	ConsequentDescription	Confidence	Lift
ALARM CLOCK BAKELITE RED	ALARM CLOCK BAKELITE GREEN	0.604333	16.996386
ALARM CLOCK BAKELITE GREEN	ALARM CLOCK BAKELITE RED	0.671736	16.996386
RED HANGING HEART T-LIGHT HOLDER	WHITE HANGING HEART T-LIGHT HOLDER	0.673049	7.550530
JUMBO BAG PINK POLKA DOT	JUMBO BAG RED RETROSPOT	0.626866	8.693843
PAPER CHAIN KIT VINTAGE CHRISTMAS	PAPER CHAIN KIT 50'S CHRISTMAS	0.647059	14.651261
ROSES REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.690932	22.187826
GREEN REGENCY TEACUP AND SAUCER	ROSES REGENCY TEACUP AND SAUCER	0.782923	22.187826
PINK REGENCY TEACUP AND SAUCER	GREEN REGENCY TEACUP AND SAUCER	0.827338	26.568210
GREEN REGENCY TEACUP AND SAUCER	PINK REGENCY TEACUP AND SAUCER	0.665702	26.568210
GARDENERS KNEELING PAD CUP OF TEA	GARDENERS KNEELING PAD KEEP CALM	0.729134	21.401429
GARDENERS KNEELING PAD KEEP CALM	GARDENERS KNEELING PAD CUP OF TEA	0.612434	21.401429



10. *Figure*: *Display graph that helps visualize the strong associations between products.*

The table shows the correlation analysis results between products using Lift and Confidence to understand the strength of the relationship between products that frequently appear together in transactions.

Lift measures how the likelihood of purchasing the Consequent item increases when the Antecedent item is purchased. A Lift value greater than 1 indicates a stronger relationship. For instance, the Lift value of 26.57 between the GREEN REGENCY TEACUP AND SAUCER and the PINK REGENCY TEACUP AND SAUCER signifies a very strong relationship.

Confidence indicates the probability that the Consequent item will be purchased after the Antecedent item. High Confidence values suggest a strong connection between the two products. For example, the confidence ratio of 0.782923 between the ROSES REGENCY TEACUP AND SAUCER and the GREEN REGENCY TEACUP AND SAUCER shows a 78.29% chance of purchasing the first product after the second.

There are two graphs: the first illustrates the relationship between customer IDs and the stock codes they purchase, while the second focuses on product descriptions, highlighting which products frequently appear together in transactions. Together, these graphs offer a comprehensive view of purchasing patterns and product relationships, clearly showing strong connections between certain products.

5.Summary and Insight:

Products with higher support are of greater marketing significance, as they reflect a higher frequency of being bought together by customers.

This can be leveraged for several purposes, such as:

-**Improving marketing strategies**: By offering promotions or discounts on products that tend to be purchased together

-**Optimizing product placement in stores** Strategically placing frequently bought-together items next to each other to enhance customer convenience and encourage additional purchase

-**Enhancing recommendation systems**: In e-commerce, this information can be used in recommendation systems; when a customer buys one product, other frequently bought-together items can be suggested based on past data.

6.Conclusion:

This analysis of e-commerce transaction data highlights the importance of understanding customer purchasing behaviors through Market Basket Analysis. By using the FP-Growth algorithm, we were able to identify strong associations between products, as evidenced by the high values of lift and confidence between certain items. These insights provide valuable opportunities for businesses to optimize marketing strategies, product placement, and recommendation systems. By leveraging the frequent purchasing patterns of customers, businesses can enhance customer satisfaction and increase sales. The results emphasize the potential of data-driven decisions in improving both operational efficiency and the overall customer experience in e-commerce platforms.