Data Analysis 2

COURSE PRESENTER
(DR. Omima Fallatah)

SUBMITTED BY:

Name	ID
Jood Mohmmed Algarni	444005790
Noura Nawar Alhuthali	444001743

DEPARTMENT OF (DATA SCIENCE)
COLLEGE OF COMPUTERS
UMM AL-QURA UNIVERSITY

Market Basket Analysis

Data

we'll be working with the **E-Commerce Retail Dataset** this is a dataset containing transnational transactions made on a UK-based online retail store between 01/12/2010 and 09/12/2011 capturing customer purchasing behaviour over time. The data is available at Kaggle and can be downloaded from here: E-Commerce Data (kaggle.com) .

Defining Objectives

The primary objective her is to identify patterns in customer purchasing behaviour through Market Basket Analysis by understanding which products are frequently bought together, improve product placement, and gain insights into customer purchasing patterns, especially which items customers are likely to purchase together.

Data Exploration

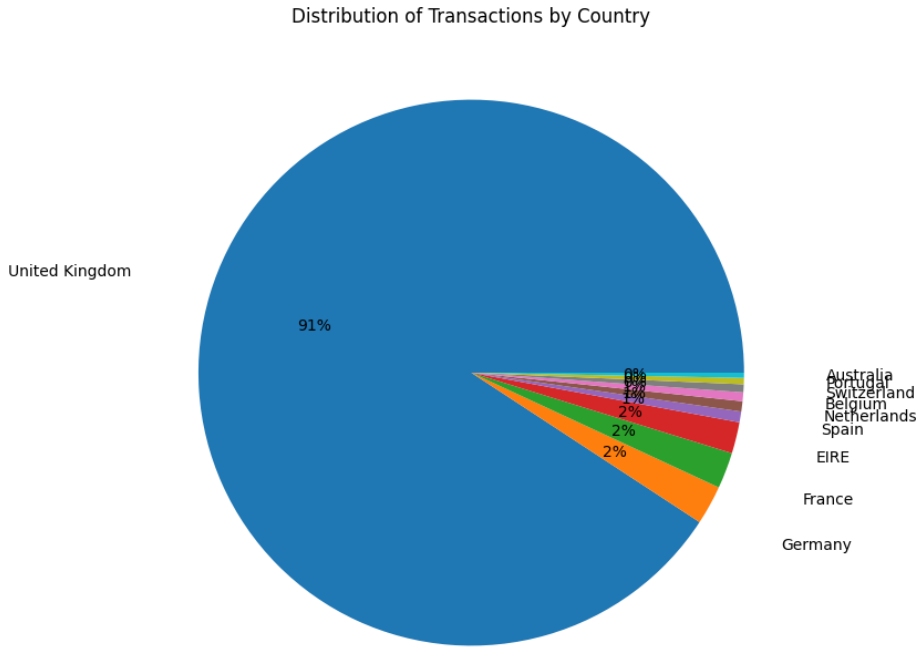
First we discovered number of rows and columns using (.shape) method its print that the data have 541909 rows, and 8 columns :
InvoiceNo: a unique 6-digit number assigned to each transaction. If this code starts with letter 'C', then the order was cancelled.
StockCode: a unique 5-digit number assigned to each distinct product.
Description: the product name.
Quantity: the number of each product (item) purchased per transaction.
InvoiceDate: the date and time each transaction was completed.
UnitPrice: the product price per unit in pounds sterling.
CustomerID: a unique 5-digit number assigned to each customer.
Country: the name of the country from where the purchase was made.

we check for null values, and we have null in CustomerID and Description coulmns this will be handled it Preprocessing section.

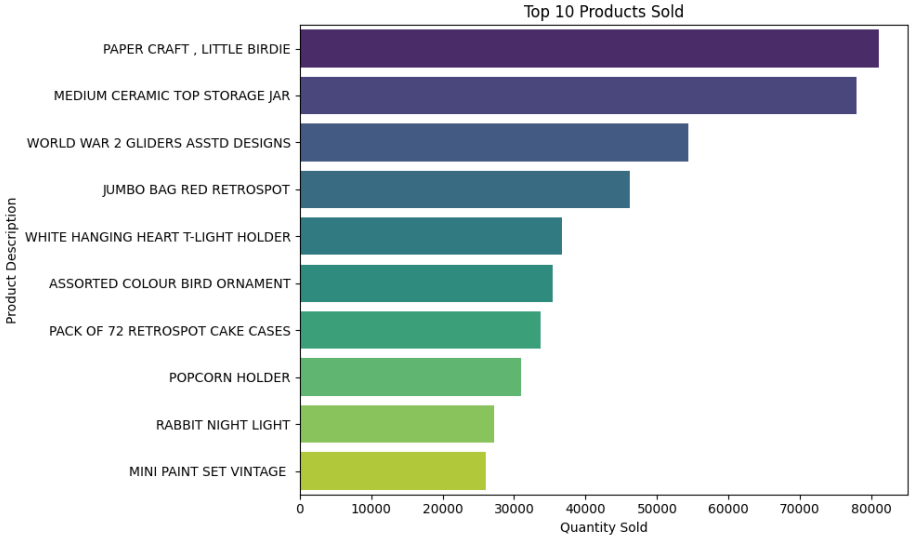
Data Preprocessing

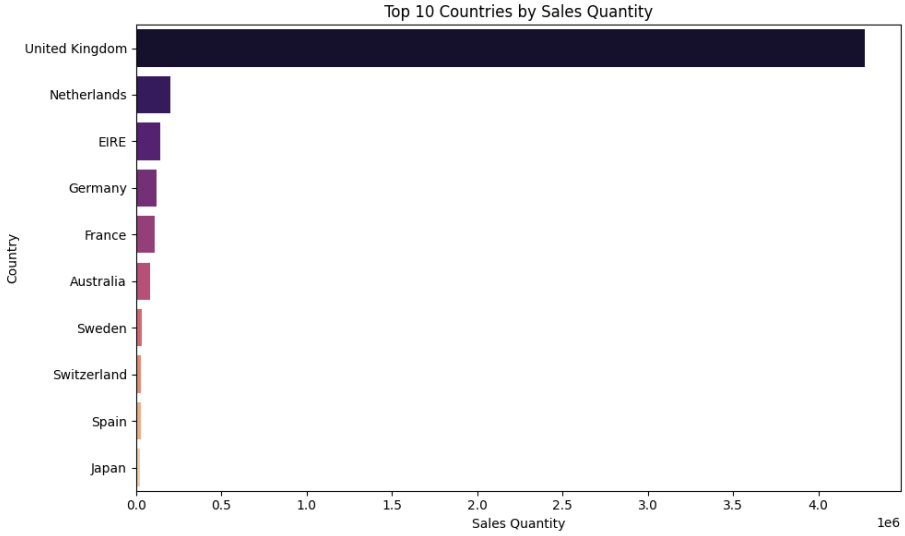
so first the data set contains a total of 136,534 null values, with the vast majority occurring in the CustomerID column and since we will not use it for analysis simply we will remove all rows with a null value using (drop) method , next step we converting the InvoiceNo column to a string data type, and then removing all rows that start with a “C” which mean canceled/returned transactions , we also Convert InvoiceDate to datetime objects so we can see sales per month, another thing we do that we choose transactions from only one country for our analysis.
first check how all the transactions are distributed by country with the(value_counts()) method and we limit it to the top 10 countries its show that UK is clearly number one in terms of the number of transactions.

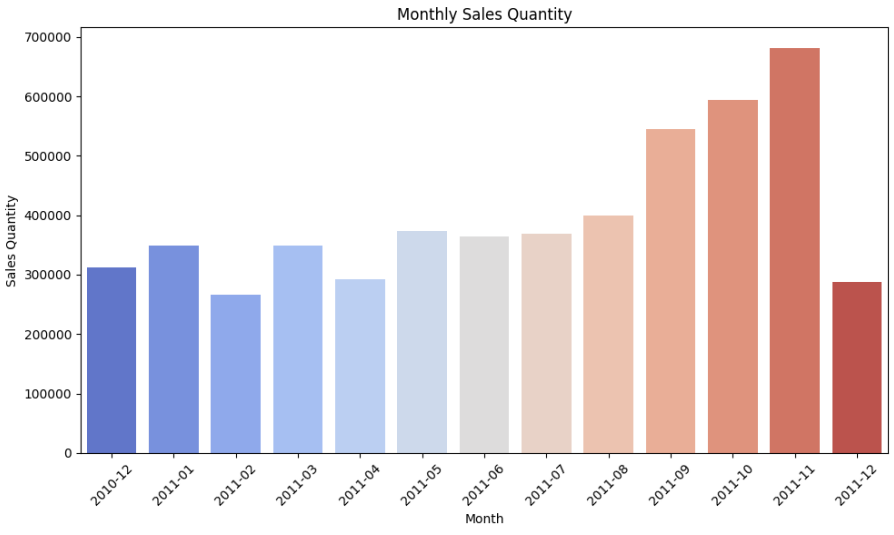
After all this be do some visualizations to get more insights about the data



It Obviously that 91% of all transactions originated from the UK.


Top 10 Products Sold
PAPER CRAFT , LITTLE BIRDIE
MEDIUM CERAMIC TOP STORAGE JAR
WORLD WAR 2 GLIDERS ASSTD DESIGNS
JUMBO BAG RED RETROSPOT
WHITE HANGING HEART T-LIGHT HOLDER
ASSORTED COLOUR BIRD ORNAMENT
PACK OF 72 RETROSPOT CAKE CASES
POPCORN HOLDER
RABBIT NIGHT LIGHT
MINI PAINT SET VINTAGE
Product Description
0
10000
20000
30000
40000
50000
60000
70000
80000
Quantity Sold


Top 10 Countries by Sales Quantity
United Kingdom
Netherlands
EIRE
Germany
France
Australia
Sweden
Switzerland
Spain
Japan
Country
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e6
Sales Quantity


Monthly Sales Quantity
700000
600000
500000
400000
300000
200000
100000
0
Sales Quantity
2010-12
2011-01
2011-02
2011-03
2011-04
2011-05
2011-06
2011-07
2011-08
2011-09
2011-10
2011-11
2011-12
Month

Data Processing

For the processing we did 3 main steps before applying Association Rule Mining

First **Creating a Transaction Basket**:
We group the data by InvoiceNo and Description to analyze item purchases this will show the quantity of each item purchased. So, we need to sum up these values and unstack them. Also change the index of the DataFrame to the InvoiceNo to display the quantity of each item purchased for every transaction.

Second **Encoding Values**:
We apply a function to hot encode the values, indicating whether a product was purchased (1) or not (0).

Third **Filtering Baskets**:
We filter for invoices with two or more items to focus on meaningful transactions because single-item invoice will be of no use.

Association Rule Mining

First to identify the most frequently purchased items in the dataset, we will apply the Apriori algorithm we set the minimum support value to 3%, meaning that only items appearing in at least 3% of the transactions will be considered.

	support	itemsets
99	0.121358	(WHITE HANGING HEART T-LIGHT HOLDER)
44	0.093197	(JUMBO BAG RED RETROSPOT)
80	0.090466	(REGENCY CAKESTAND 3 TIER)
6	0.084417	(ASSORTED COLOUR BIRD ORNAMENT)
71	0.082986	(PARTY BUNTING)
58	0.072841	(LUNCH BAG RED RETROSPOT)
86	0.064971	(SET OF 3 CAKE TINS PANTRY DESIGN)
52	0.064646	(LUNCH BAG BLACK SKULL.)
69	0.061004	(PAPER CHAIN KIT 50'S CHRISTMAS)
64	0.060939	(NATURAL SLATE HEART CHALKBOARD)

According to results the “White hanging Heart T-Light Holder” is the most frequently purchased item, with a support value of 0.121358. meaning it was purchased in 12% of all transactions.

final step in our analysis is to generate the rules along with their corresponding support, confidence, and lift values allowing us to extract useful insights about which items are more likely to be purchased together.

And this is the result:

	antecedents	consequents	support	confidence	lift
5	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.030957	0.777778	17.717202
4	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.030957	0.705185	17.717202
0	(JUMBO BAG PINK POLKADOT)	(JUMBO BAG RED RETROSPOT)	0.032908	0.624691	6.702899
6	(LUNCH BAG PINK POLKADOT)	(LUNCH BAG RED RETROSPOT)	0.030632	0.556080	7.634188
2	(LUNCH BAG BLACK SKULL.)	(LUNCH BAG RED RETROSPOT)	0.031478	0.486922	6.684737
3	(LUNCH BAG RED RETROSPOT)	(LUNCH BAG BLACK SKULL.)	0.031478	0.432143	6.684737
7	(LUNCH BAG RED RETROSPOT)	(LUNCH BAG PINK POLKADOT)	0.030632	0.420536	7.634188
1	(JUMBO BAG RED RETROSPOT)	(JUMBO BAG PINK POLKADOT)	0.032908	0.353105	6.702899

Conclusion:

After implementing association rules, we can see that the “Roses Regency Teacup and Saucer” and the “Green Regency Teacup and Saucer” have the highest lift value, indicating the strongest association between any two products with a combined support of 0.0309, these charming teacups were purchased together in 3.09% of all transactions, highlighting their popularity as a pair among customers this insight can be leveraged for targeted marketing strategies, such as bundling these products in promotions.