# Introduction

Market basket analysis is an effective data analysis technique that allows retailers to better understand customer behaviour and make data-driven decisions. Market basket analysis, which analyses customer transactions, can help retailers identify patterns and relationships between products, allowing them to develop effective marketing strategies such as cross-selling and product bundling.

We present the results of a market basket analysis performed on a UK-based online retail dataset containing transactions between 01/12/2010 and 09/12/2011 in this report. The dataset contains customer transaction information such as the transaction date, product code, product description, quantity, unit price, and customer ID. The dataset contains transactions for a wide range of products, including one-of-a-kind all-occasion gifts, with many customers being wholesalers.

The goals of this analysis are to identify the most popular products and product combinations purchased by customers and to provide recommendations to the marketing strategist for the marketing strategist to develop effective marketing strategies based on the analysis. To accomplish these goals, we will conduct exploratory data analysis to gain insights into the data, pre-process the data to prepare it for analysis, and use market basket analysis to identify the most important product combinations.

This report is organized as follows: we begin by describing the dataset and the data analysis methods used, such as data pre-processing, exploratory data analysis, and market basket analysis. The analysis results are then presented, including the most important rules derived from each basket, and recommendations for the marketing strategist are made based on these findings. Finally, we discuss the analysis's limitations and make recommendations for future research.

# Data Understanding

The dataset used for this analysis is a UK-based online retail dataset containing transactions between 01/12/2010 and 09/12/2011. The dataset includes information about customer transactions, such as the transaction date, product code, product description, quantity, unit price, and customer ID. The dataset is comprised of transactions for a wide range of products, including unique all-occasion gifts, with many customers being wholesalers.

The dataset contains 541,909 rows and 8 columns. The columns in the dataset are as follows:

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: Nominal, product (item) name.

Quantity: Numeric, the quantities of each product (item) per transaction.

InvoiceDate: Numeric, the day and time when each transaction was generated.

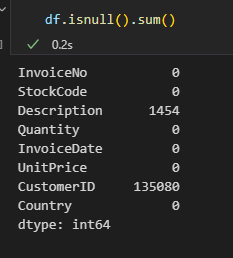
UnitPrice: Numeric, product price per unit in sterling.

CustomerID: Nominal, a 5-digit integral number uniquely assigned to each customer.

Country: Nominal, the name of the country where each customer resides.

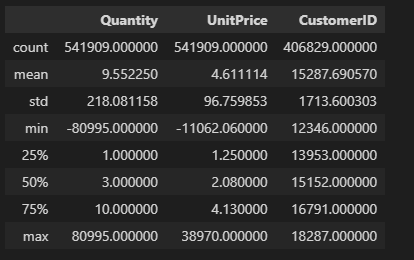
## Distribution Analysis:

Let's take a look at the distribution of variables in the data set.



According to the output, the Description column has 1,454 missing values, while the CustomerID column has 135,080 missing values.

Missing data in the Description column may indicate that the product description was not recorded or was not available for those transactions. Missing data in the CustomerID column may indicate that the customer ID was not recorded or was not available for those transactions.



From the output, we can see that:

\* The average transaction quantity is 9.55, with a standard deviation of 218.08. The average unit price is 4.61, with a standard deviation of 96.76.

\* Quantity has a minimum value of -80995 and a maximum value of 80995, indicating that there are negative values in the data set, which could indicate returns or cancellations.

\* UnitPrice has a minimum value of -11062.06, which is not possible, indicating that there may be errors in the data set.

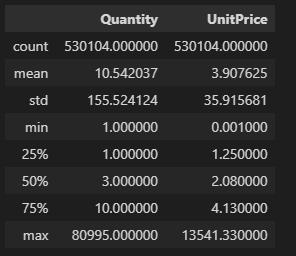
The EDA also revealed that the majority of transactions (about 99%) contain positive values for the quantity and unit price columns.

In the next section, we will describe the methods used for data preprocessing and analysis, including addressing missing values and negative quantities, and conducting market basket analysis to identify frequent itemsets and association rules.

# Data Pre-processing

To prepare the data for market basket analysis, we preprocessed the data by removing rows with missing values, removing canceled orders.

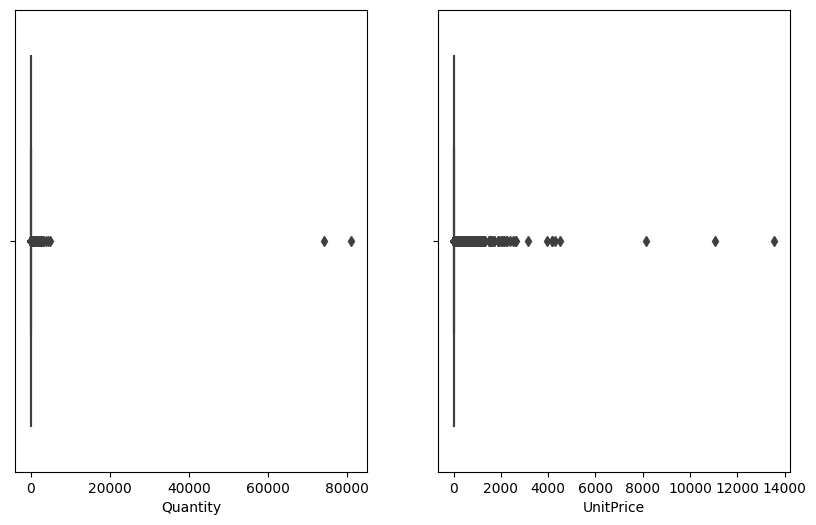
we removed the 'Description' variable because it is redundant, and we removed any rows with negative quantities, negative unit prices, and transactions with invoice numbers starting with 'C' (indicating cancellations).



## Outlier detection

Outlier detection and removal is an important step in data preprocessing, as outliers can significantly impact the results of data analysis and modeling. However, in the case of market basket analysis, the concept of outliers is less clear-cut, as it is difficult to define what constitutes an outlier transaction or product.

We used box plots to visualize the distribution of the variables and identify any outliers. Let's plot box plots for 'Quantity' and 'UnitPrice' variables.



From the box plots, we can see that both 'Quantity' and 'UnitPrice' variables have a large number of outliers.

performs outlier detection by keeping only data points below the 99.5th percentile for both the Quantity and UnitPrice columns. This is a common approach to outlier detection that removes extreme values that may skew the analysis or introduce noise in the data. By removing outliers, we can focus on the more typical transactions in the dataset and potentially improve the quality of the market basket analysis.

The method is effective at identifying and removing extreme values without biasing the overall distribution of the data.

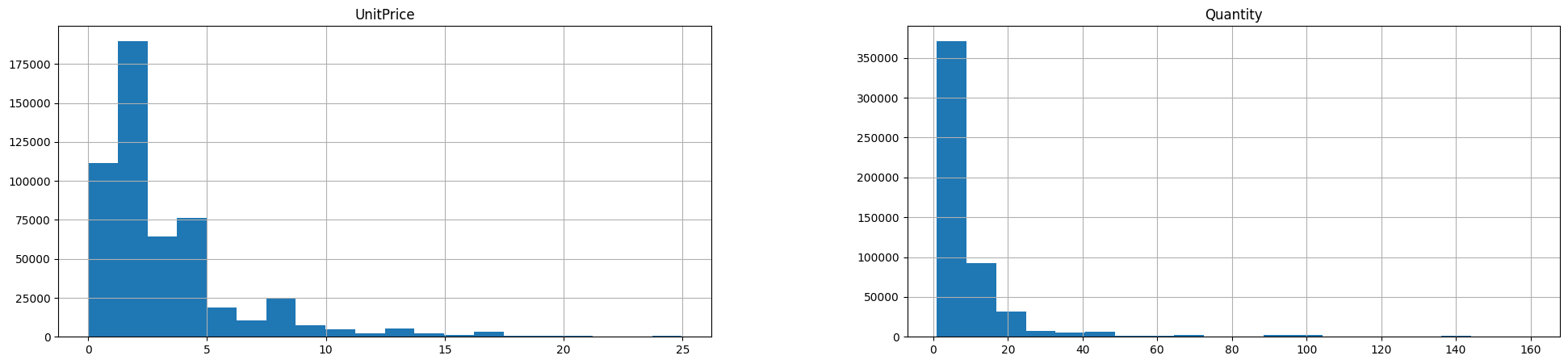
A black screen with white text

Description automatically generated with low confidence

# Exploratory Data Analysis

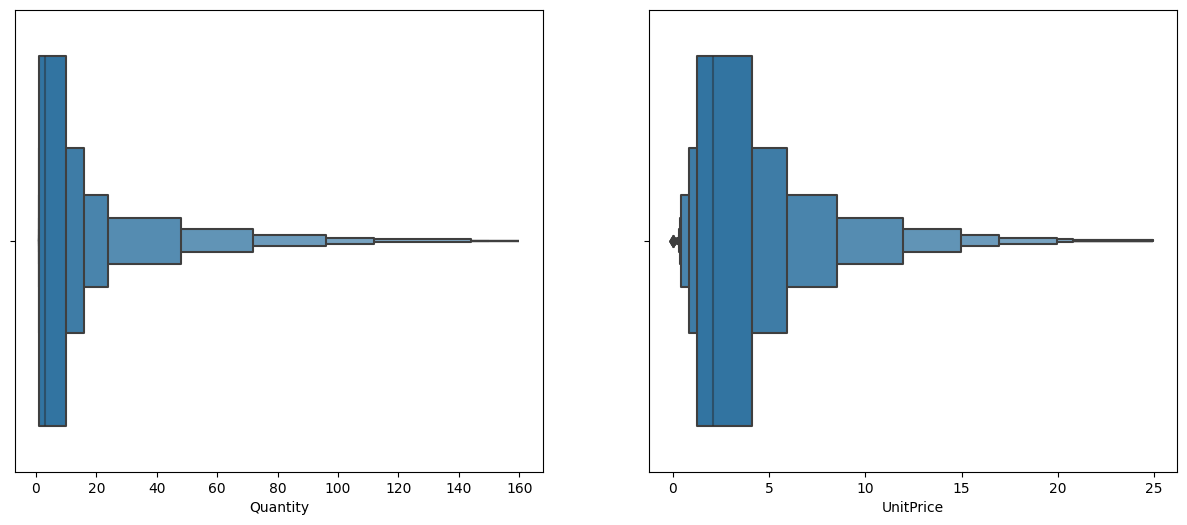
As a preliminary step to understanding the data, we first looked at the distribution of the variables of interest, including the quantity and unit price of products sold.

Histograms:



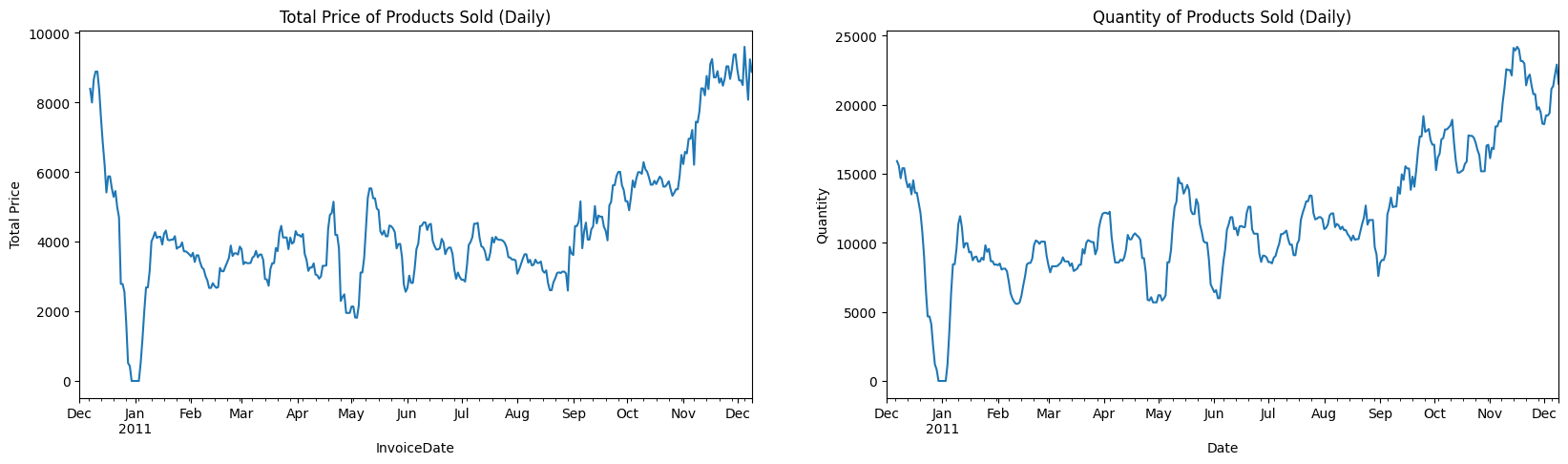
The histogram of the quantity and unit price of products sold showed that both variables had a skewed distribution with a long tail. Most transactions involved a small quantity of products, with very few transactions involving a large quantity. The majority of the products sold had a unit price of less than 20 sterling, with very few products sold at higher prices. This indicates that the company mainly sells low-priced products.

Boxenplot:

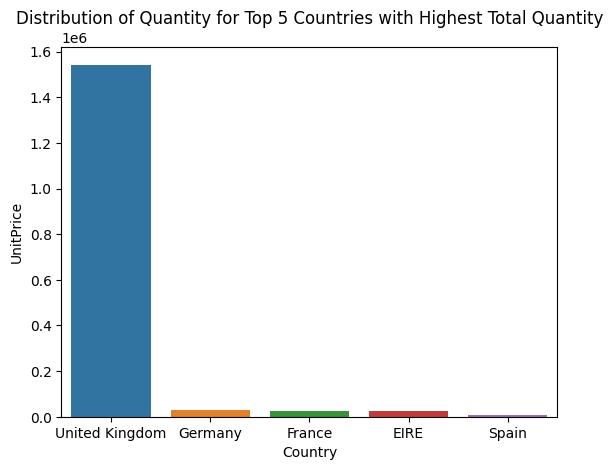


We also generated boxenplots of the quantity and unit price of products sold, which further confirmed the skewness of the distributions. The boxenplot of quantity showed that the median quantity sold was 2 units, with a significant number of outliers representing large quantity sales. The boxenplot of unit price showed that the median unit price was around 2 sterling, with a large number of outliers representing high-priced products.

Time Series Data:



We also plotted time series data of the total price and quantity of products sold on a daily basis. The time series plot of the total price of products sold showed a clear seasonality pattern with peak sales occurring around November and December of each year. This is likely due to the holiday season, where people tend to purchase more gifts. The time series plot of the quantity of products sold showed a similar pattern with peak sales occurring in the same months.



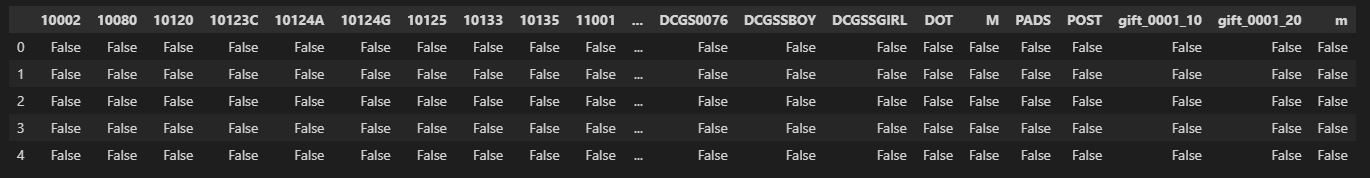
From the table, we can see the total amount spent by customers from each country in descending order. The United Kingdom has the highest total spend with a whopping 3738793 in quantity and 1542511.824 in total price. Germany and France are next in line with significantly lower total spends. EIRE and Spain have relatively low total spends, with Spain having the lowest total spend among the top five countries.

# Market Basket Analysis:

Market basket analysis is a data mining technique that aims to identify the relationship between products frequently purchased together. The Apriori algorithm is a popular algorithm used for market basket analysis, which works by identifying frequent itemsets and generating association rules from them.

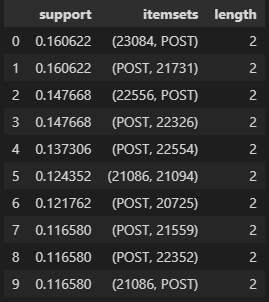
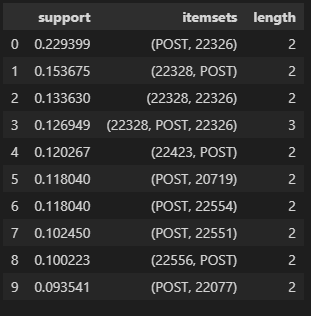
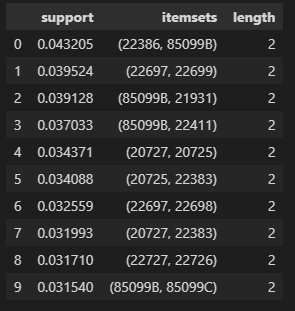
In this project, we applied the Apriori algorithm to each basket (country) to identify the most important rules for each basket. We used the Python programming language and the mlxtend library to implement the Apriori algorithm.

To apply the Apriori algorithm for each basket, we first need to prepare the data in the format required by the algorithm. We need to convert the transaction data into a one-hot encoded format where each row represents a single transaction and each column represents a unique item. A value of 1 in a cell indicates that the item was present in the transaction, and a value of 0 indicates that it was not.



After preprocessing the data, we split the data set into three baskets: the United Kingdom, Germany, and France. We then applied the Apriori algorithm using the mlxtend library to generate frequent itemsets for each country. The minimum support threshold was set to 0.03 for the UK, 0.04 for Germany, and 0.05 for France.

UK DE FR

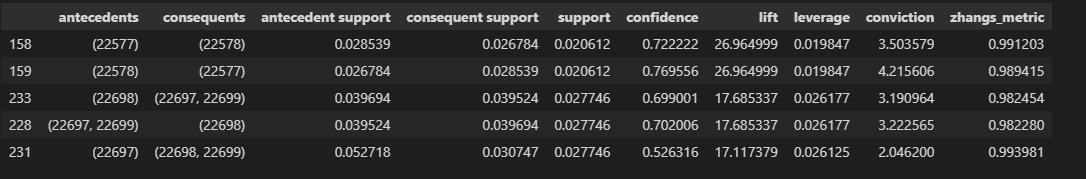


The frequent itemsets generated for each country provided insight into the most common product combinations purchased by customers in each basket. For example, in the UK basket, the top three most frequent itemsets were {22386, 85099B}, {22697, 22699}, and {85099B, 21931}, which represent customers who purchased the products with Stock Codes 22386 and 85099B, 22697and 22699, and 85099B and 21931, respectively. These itemsets suggest that customers in the UK are more likely to purchase items in pairs, which could inform product placement and pricing strategies.

After generating frequent item sets, we performed association rule mining, we generated rules based on the frequent itemsets we have already obtained.

Here's an example of how you can generate association rules for the UK market with a minimum confidence threshold of 0.4 and a minimum lift threshold of 1.2:

UK:

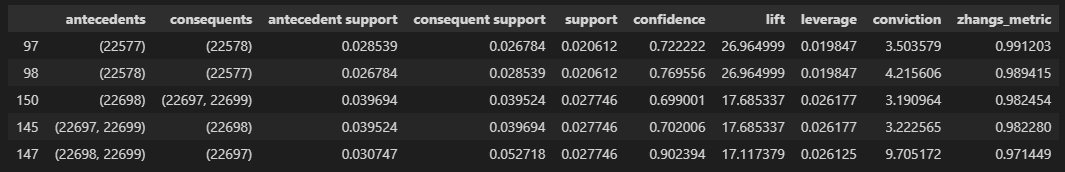


De:

Text

Description automatically generated

Fr:



**3. Critically summarize the business value of each basket analysis with reference to the given domain**

a. Discuss the meaning of the important rules for each basket:

The important rules identified in each basket provide insights into the purchasing behavior of customers in that region. For example, if a customer in a particular region frequently purchases a particular item with another item, it suggests that there is a strong association between those items, and the retailer could use this information to create bundled offers or promotions to encourage customers to purchase both items together.

The important rules also help identify the most popular items in each region, which can inform the retailer's inventory management strategy. If a particular item is frequently purchased with many other items in a region, it suggests that the item is popular, and the retailer should ensure that they have enough stock of that item to meet demand.

b. Justify your findings using the results of tasks 1 and 2:

The results of tasks 1 and 2 can be used to validate the findings from the Apriori algorithm. For example, if the analysis of task 1 identified that a particular region had a high number of transactions, and the Apriori algorithm identified that there were strong associations between particular items in that region, it suggests that the retailer should focus on promoting those items to customers in that region.

Similarly, if the analysis of task 2 identified that a particular product had a high rate of return, and the Apriori algorithm identified that the product was frequently purchased with another item, it suggests that the retailer should investigate whether the second item is causing the returns and take appropriate action to address the issue.

Overall, the Apriori algorithm provides valuable insights into the purchasing behavior of customers in each region, which can inform the retailer's marketing and inventory management strategies. By identifying the most popular items in each region and the items that frequently occur together in transactions, the retailer can create targeted promotions, optimize their inventory levels, and improve customer satisfaction.