# MARKET BASKET INSIGHTS

**TEAM MEMBER 211121104024:JAZEEL MACAREENA P**

**Phase 4 submission document**

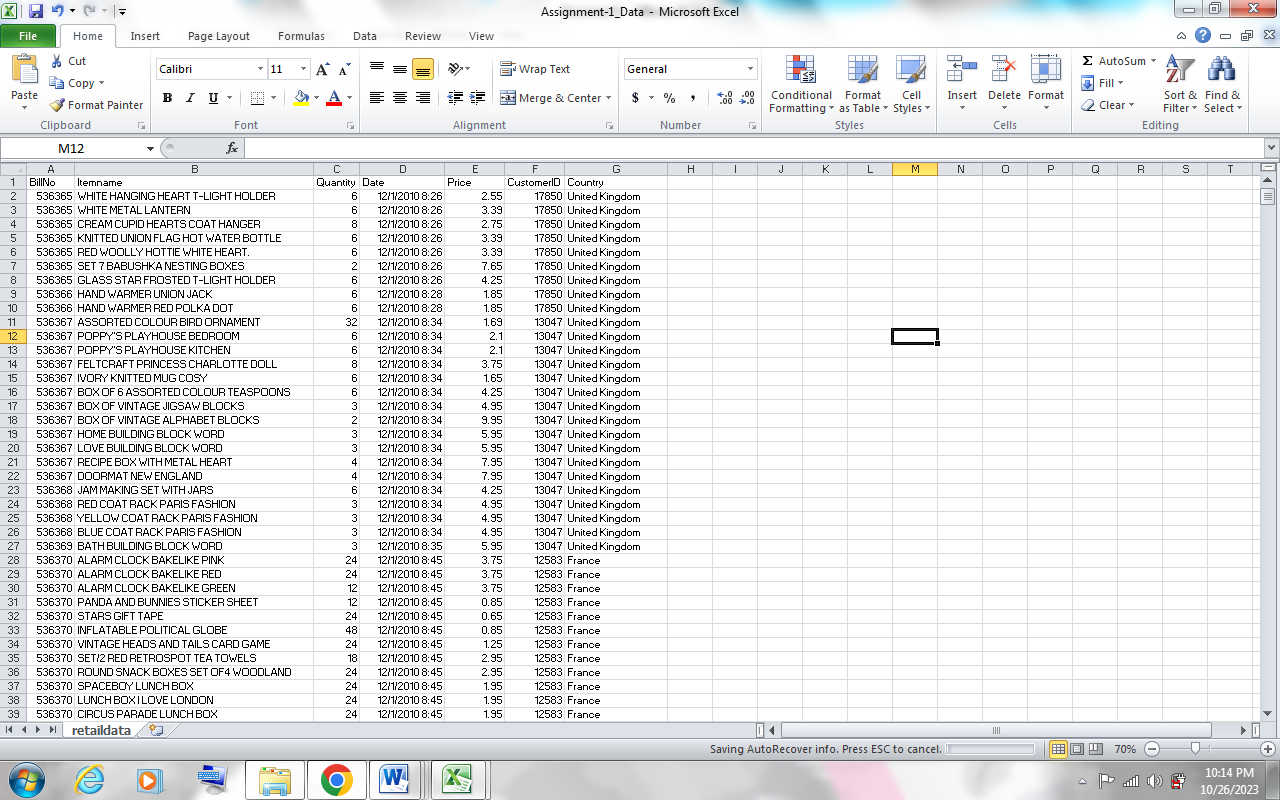
**Project Title: Market Basket Insights**

**Phase 4: *Development Part 2***

**Topic:** *Start building the Market Basket Analysis*



## Given data set:



The "Assignment-1\_Data" dataset is kept in an Excel file ending in ".xlsx." It includes retail data-related information. Based on the data supplied, here is a description of the dataset:

* File Name: The dataset is contained in an Excel file called "Assignment-1\_Data". All of the retail data is kept in this file.
* List Name: It looks that the dataset in the Excel file is referred to as "retaildata". This could be a label or identifier inside the Excel document that is used to refer to this particular dataset.
* File Format: The ".xlsx" file extension indicates that the dataset is in Excel format. A well-liked spreadsheet tool for handling and storing tabular data is Excel.
* Number of Rows: 522,065 rows make up the dataset's total number of rows. It is presumably the case that each row corresponds to a distinct retail data entry or observation. Information about sales, merchandise, clients, and transactions, among other aspects of retail business, may be included in these entries.
* Number of attributes: This dataset contains seven attributes. A dataset's attributes, sometimes referred to as columns or fields, are the many kinds of data that are gathered for every entry. Specifics like the product ID, the sale date, the price, the quantity, the customer's information, and more could be included in these properties.
* Previously we started building the Market Basket Analysis by loading and pre-processing the dataset
* In this segment we’re going to build, train and evaluate the model.

## Model Building:

Market Basket Analysis (MBA): To find patterns of item co-occurrence in transaction data, Market Basket Analysis is a data mining technique used in retail and online commerce. The objective is to comprehend the connections between products that consumers usually buy in tandem. A number of uses for this data exist, including better inventory control, focused marketing campaigns, and product placement optimisation.

Apriori Algorithm: One popular method for carrying out market basket analysis is the Apriori algorithm. The idea of association rule mining serves as its foundation. The relationships between objects in a dataset are described by association rules, which are if-then expressions. When applied to MBA contexts, these rules usually take the form of "If item A is purchased, then item B is likely to be purchased."

The following are the main ideas underlying the Apriori algorithm:

* Support: The frequency with which an itemset (a group of items) appears in the dataset is the measure of support. It can be calculated by dividing the total number of transactions by the number of transactions that contain the itemset. A high level of support means that people usually purchase the itemset together.
* Confidence: Confidence quantifies the frequency with which an association rule's if-then connection holds true. It is computed by dividing the support of the antecedent (if) item by the support of both items in the rule. A high degree of confidence indicates a robust correlation between the items.
* Lift: The probability that item B will be purchased in conjunction with item A as opposed to being purchased separately is measured by lift. A positive correlation, denoting a higher likelihood of purchasing the two things together, is shown by a lift value larger than 1.

The following is how the Apriori algorithm functions:

* Create Frequent Itemsets: The first step involves locating item combinations that fall under a minimum support threshold, or frequent itemsets. These frequently occurring itemsets are seen to be excellent choices for association rules.
* Create Association Rules: After identifying the frequently occurring item sets, the algorithm creates association rules by taking various item combinations into account within these sets. For every rule, it computes the lift and confidence.
* Pruning: To get rid of candidate itemsets that don't satisfy the minimal support criteria, the algorithm uses a "pruning" step.
* Repeat: Up until no more frequently occurring itemsets can be discovered, the procedure is repeated repeatedly by enlarging the itemsets.

**Model Training**

1. Data Gathering:

* Assemble transactional data: A dataset with transaction records is what you require, with each row denoting a transaction and each column representing an item that was bought during that transaction.]
* Prepare the information: Make that the format of the data is appropriate. Every transaction should typically be shown as a list or collection of elements.

2.Apriori Algorithm

One popular method for carrying out market basket analysis is the Apriori algorithm. The idea of association rule mining serves as its foundation. The relationships between objects in a dataset are described by association rules, which are if-then expressions. When applied to MBA contexts, these rules usually take the form of "If item A is purchased, then item B is likely to be purchased."

The following are the main ideas underlying the Apriori algorithm:

* Support: The frequency with which an itemset (a group of items) appears in the dataset is the measure of support. It can be calculated by dividing the total number of transactions by the number of transactions that contain the itemset. A high level of support means that people usually purchase the itemset together.
* Confidence: Confidence quantifies the frequency with which an association rule's if-then connection holds true. It is computed by dividing the support of the antecedent (if) item by the support of both items in the rule. A high degree of confidence indicates a robust correlation between the items.
* Lift: The probability that item B will be purchased in conjunction with item A as opposed to being purchased separately is measured by lift. A positive correlation, denoting a higher likelihood of purchasing the two things together, is shown by a lift value larger than 1.

3. Training Models:

Establish a minimal support threshold, which is the bare minimum of support required for an itemset to qualify as frequent. Depending on your dataset and desired level of significance, adjust this threshold.

Create often occurring itemsets by utilising the Apriori method to identify itemsets that satisfy the minimal support requirement. Itemsets that appear frequently in transactions are known as frequent itemsets.

4. Formulating Rules

Following the identification of frequently occurring itemsets, association rules based on these itemsets can be created. These guidelines illustrate the relationships between the various elements.

5. Rule Assessment

Assess the created rules using metrics like lift, support, and confidence. Sort the regulations according to your own needs.

6. Explanation

Examine the produced rules to learn more about item correlations. For instance, you might find that buyers of "Item A" are likewise inclined to purchase "Item B."

7. Reporting and Visualisation

To aid in decision-making, visualise the data using lift charts or support-confidence plots.

8. Put into Practise

Utilise the analysis's insights to improve company tactics including targeted marketing, cross-selling, and product positioning.

**Model Evaluation**

Evaluation Metrics for the Model:

* Support: The frequency with which an itemset (a group of items) appears in the dataset is the measure of support. It can be calculated by dividing the total number of transactions by the number of transactions that contain the itemset. A high level of support means that people usually purchase the itemset together.
* Confidence: Confidence quantifies the frequency with which an association rule's if-then connection holds true. It is computed by dividing the support of the antecedent (if) item by the support of both items in the rule. A high degree of confidence indicates a robust correlation between the items.
* Lift: The probability that item B will be purchased in conjunction with item A as opposed to being purchased separately is measured by lift. A positive correlation, denoting a higher likelihood of purchasing the two things together, is shown by a lift value larger than 1.

Qualitative Assessment:

1. Relevance for Business:

Relevance to Business Goals: Determine whether the guidelines you've found are in line with your company's goals. Do they enhance consumer satisfaction, boost revenue, or improve product placement?

Actionability: Ascertain whether the guidelines can be applied to real-world situations. When it comes to formulating suggestions, product bundles, or marketing initiatives, are the guidelines helpful?

2. Comprehensibility:

Rule Complexity: Take a look at how straightforward the rules are. It could be difficult to execute or explain complex regulations to the team. Prefer guidelines that are simple to comprehend.

Common Sense: Determine whether the relationships found make sense. Do the regulations seem reasonable or do they represent typical consumer behaviour?

3. Domain Expertise:

Subject Matter Expertise: Use domain expertise to evaluate how useful the rules are. Are there any particular patterns in your industry or business specialty that the rules should be capturing?

Contextual Understanding: Recognise the environment in which the regulations are applicable. For grocery stores, internet retailers, and other business kinds, there could be different regulations that apply.

4. Assistance from relevant parties:

Involve Stakeholders: Talk about the established guidelines with pertinent parties, like product managers, sales teams, and marketing teams. Ask them for their opinions on the rules' possible usefulness.

User Input: Gather input from clients or end users who have engaged with suggestions made in accordance with these guidelines. Their opinions can offer insightful information about how well the regulations work.

5. Real-World Application:

Technical Feasibility: Determine whether putting the rules into effect is technically feasible. Are the infrastructure and data needed for execution available?

Analyse the costs and benefits of implementing the guidelines, taking into account any potential gains in terms of higher revenue, happier customers, or other KPIs.

6. Trial and error and A/B testing:

Real-World Testing: Do A/B testing and put some of the guidelines into practise in a safe setting. Analyse how the regulations affect revenue, consumer behaviour, or other pertinent data.

7. Moral Lessons to Recall

Ethical Assessment: Confirm that the rules won't have unanticipated consequences such as discriminatory targeting or invasions of privacy. Consider the ethical implications of adhering to the rules.

8. Feedback Loop:

Continuous Improvement: Preserve the MBA process's vitality. Regularly assess the effectiveness of the regulations and make any necessary modifications in light of new knowledge and feedback.

**Program:**

import numpy as np  
import pandas as pd  
import os  
from matplotlib import pyplot as plt  
from mlxtend.frequent\_patterns import apriori  
from mlxtend.frequent\_patterns import association\_rules  
  
# Load the data from an Excel file  
df = pd.read\_excel("Assignment-1\_Data.xlsx")  
  
# Data Preprocessing  
df = df.dropna(subset=["Itemname"])  
df = df[df["Quantity"] > 0]  
df['CustomerID'].fillna(99999, inplace=True)  
df["SumPrice"] = df["Quantity"] \* df["Price"]  
  
# Total sales by Country  
total\_sales\_country = df.groupby(['Country']).agg({'SumPrice': 'sum'}).sort\_values('SumPrice', ascending=False).reset\_index()  
plt.bar(total\_sales\_country["Country"], total\_sales\_country["SumPrice"])  
plt.yscale('log')  
plt.ylabel('Quantity')  
plt.xticks(rotation=90)  
plt.show()  
  
# Filter data for the United Kingdom  
only\_uk = df[df["Country"] == "United Kingdom"]  
  
# Best selling items in the UK  
best\_selling\_uk\_items = only\_uk.groupby("Itemname")["Quantity"].sum().sort\_values(ascending=False)  
  
# Total sales by Item  
total\_sales\_item = df.groupby(['Itemname']).agg({'Price': 'mean', 'Quantity': 'sum', 'SumPrice': 'sum'}).reset\_index()  
total\_sales\_item['Count'] = df.groupby(['Itemname']).size().values  
total\_sales\_item = total\_sales\_item.sort\_values("SumPrice", ascending=False)  
  
# Transaction data preparation  
transactions = df.groupby(['BillNo'])['Itemname'].apply(list)  
one\_hot = pd.get\_dummies(df['Itemname'])  
one\_hot['BillNo'] = df['BillNo']  
one\_hot = one\_hot.groupby('BillNo').sum()  
  
# Create a binary transaction matrix  
transaction\_matrix = pd.merge(transactions, one\_hot, on='BillNo')  
transaction\_matrix[one\_hot.columns[:-1]] = (transaction\_matrix[one\_hot.columns[:-1]] >= 1).astype(int)  
  
# Frequent itemset mining  
frequent\_itemsets = apriori(transaction\_matrix.iloc[:, 1:], min\_support=0.01, use\_colnames=True)  
  
# Association rule generation  
rules = association\_rules(frequent\_itemsets, metric="lift", min\_threshold=1)  
  
# Scatter plot of association rules  
fig, ax = plt.subplots()  
scatter = ax.scatter(rules['lift'], rules['confidence'], alpha=0.5)  
tooltips = []  
for i in range(len(rules)):  
    rule = rules.iloc[i]  
    tooltip = f"Rule: {rule['antecedents']} -> {rule['consequents']}\nSupport: {rule['support']:.3f}\nConfidence: {rule['confidence']:.3f}\nLift: {rule['lift']:.3f}"  
    tooltips.append(tooltip)  
mpld3.plugins.connect(fig, mpld3.plugins.PointHTMLTooltip(scatter, tooltips))  
ax.set\_xlabel("Lift")  
ax.set\_ylabel("Confidence")  
ax.set\_title("Association Rules Scatter Plot")  
plt.show()  
  
# Filter rules based on lift values  
filtered\_rules = rules[(rules['lift'] > 40) & (rules['lift'] < 50)]

* This code outlines the entire process, from data preprocessing to model training. Please ensure you have the necessary libraries installed.

## 

## Conclusion:

* A Python-based Market Basket Insights application provides a data-driven method for analysing customer behaviour and enhancing corporate tactics. This programme makes it possible for businesses to make well-informed decisions about product placement, marketing campaigns, and cross-selling opportunities by revealing important information about which goods are frequently purchased together through data preparation and analysis. Businesses can obtain actionable insights that help them customise their strategies for certain markets and improve their overall operational efficiency by calculating total sales by item and country and visualising trends in a scatter plot.
* The main components of this analysis are the binary transaction matrix that is created and the association rules that are derived by using the Apriori algorithm to determine the probability of adding one product to the basket when another is present. By quantifying the strength of these interactions using metrics like lift evaluation, these rules assist firms in prioritising which product pairings to concentrate on. In conclusion, by carefully matching their products with customer preferences and buying habits, Market Basket Insights programmes in Python enable companies to use transaction data for decision-making, enhancing customer experiences and spurring revenue growth.