IBM Naan Mudhalavan Project

Group-1 Artificial Intelligence

Team Member -4

311521106082

SAI KISHORE A

Project name

Market Basket Analysis

Phase-3

Preprocessing Data

Market Basket Analysis

**Preprocessing Data:**

**Data Collection:**

Collect transaction data, which typically includes a list of items bought in each transaction. This data should be in a structured format, such as a CSV or Excel file.

**Data Import:**

Import your data into a Pandas DataFrame. You can use libraries like pandas and numpy for data manipulation.

**Data Exploration:**

Explore your data to understand its structure. Use functions like head(), info(), and describe() to get a sense of what your data looks like.

**Data Cleaning:**

Remove any duplicate transactions or items, as they may distort the analysis.

**Data Transformation:**

You need to transform your data into a format suitable for market basket analysis. This often means converting it into a binary format, where each row represents a transaction, and each column represents an item. You can use one-hot encoding to do this.

**Consolidation:**

Group the data by transaction ID to consolidate it. This will allow you to analyze transaction-level data.

**Frequent Itemset Generation:**

Use algorithms like Apriori or FP-Growth to find frequent itemsets in your data. Libraries like mlxtend provide implementations of these algorithms.

**Association Rule Mining:**

Generate association rules from the frequent itemsets to find interesting patterns.

**Rule Filtering and Interpretation:**

Filter the rules based on your criteria (e.g., lift, confidence) and interpret the results to find meaningful associations.

**Visualization and Reporting:**

Use data visualization libraries like matplotlib or seaborn to present your findings in a clear and understandable way.

This is a high-level overview of the preprocessing steps for market basket analysis in Python. Depending on your specific dataset and analysis goals, you may need to perform additional data preprocessing and fine-tuning.

**Python code:**

import pandas as pd

from mlxtend.frequent\_patterns import apriori, association\_rules

pd.set\_option('display.max\_columns', None)

pd.set\_option('display.max\_rows', None)

pd.set\_option('display.width', 500)

df = pd.read\_excel("/kaggle/input/market-basket-analysis/Assignment-1\_Data.xlsx")

def outlier\_thresholds(dataframe, variable):

quartile1 = dataframe[variable].quantile(0.01)

quartile3 = dataframe[variable].quantile(0.99)

interquantile\_range = quartile3 - quartile1

up\_limit = quartile3 + 1.5 \* interquantile\_range

low\_limit = quartile1 - 1.5 \* interquantile\_range

return low\_limit, up\_limit

def replace\_with\_thresholds(dataframe, variable):

low\_limit, up\_limit = outlier\_thresholds(dataframe, variable)

dataframe.loc[(dataframe[variable] < low\_limit), variable] = low\_limit

dataframe.loc[(dataframe[variable] > up\_limit), variable] = up\_limit

def retail\_data\_prep(dataframe):

dataframe = dataframe[dataframe["Quantity"] > 0]

dataframe = dataframe[dataframe["Price"] > 0]

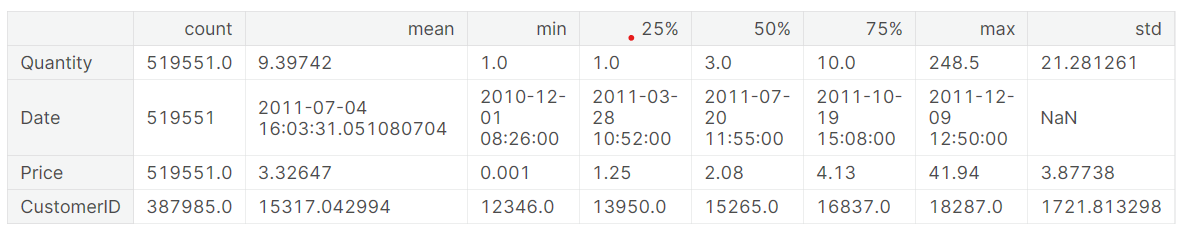
replace\_with\_thresholds(dataframe, "Quantity")

replace\_with\_thresholds(dataframe, "Price")

return dataframe

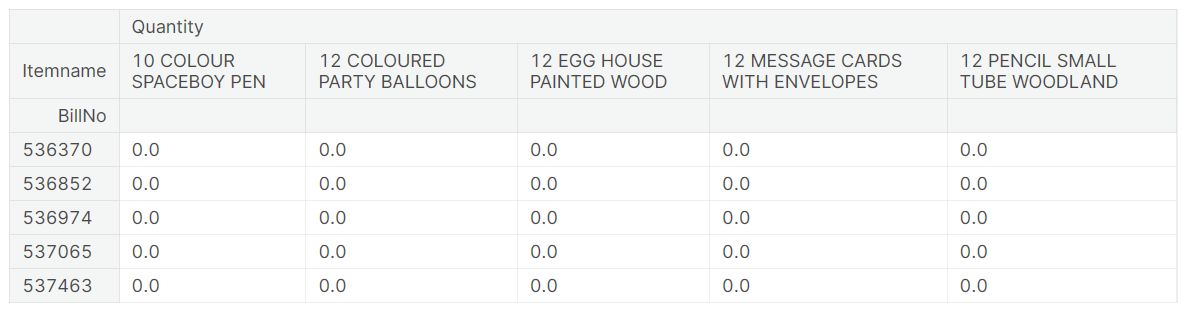
df = retail\_data\_prep(df)

df.describe()



df\_fr = df[df['Country'] == "France"]

df\_fr.groupby(['BillNo', 'Itemname']).agg({"Quantity": "sum"}).unstack().fillna(0).iloc[0:5, 0:5]



fr\_inv\_pro\_df=df\_fr.groupby(['BillNo', 'Itemname']). \

agg({"Quantity": "sum"}). \

unstack(). \

fillna(0). \

applymap(lambda x: 1 if x > 0 else 0)

frequent\_itemsets = apriori(fr\_inv\_pro\_df.astype("bool"),

min\_support=0.01,

use\_colnames=True)

frequent\_itemsets.sort\_values("support", ascending=False).head()

