PHASE 3 ASSIGNMENT

PROJECT TITLE:Preprocessing the Dataset

GITHUB LINK: https://github.com/swetha5611/market-basket-insightsss.git

https://github.com/swetha5611/innovation.git

PROBLEM DEFINITION: Market basket insights, or market basket analysis, is the process of discovering associations and patterns in customer transaction data. The primary goal is to uncover relationships between products or items that are frequently purchased together.

DOCUMENT:

Building the project by preprocessing the data

DATASET LINK ON: Market Basket Insights

https://www.kaggle.com/datasets/aslanahmedov/market-basket-analysis

➢ Pre-Requisites for Performing Market Basket Analysis :

Download the dataset before you start coding. Make sure you also have Jupyter Notebook installed on your device. If you are unfamiliar with the software, follow 365's beginner-friendly Jupyter Notebook tutorial or Introduction to Jupyter course to learn about its usage and installation.

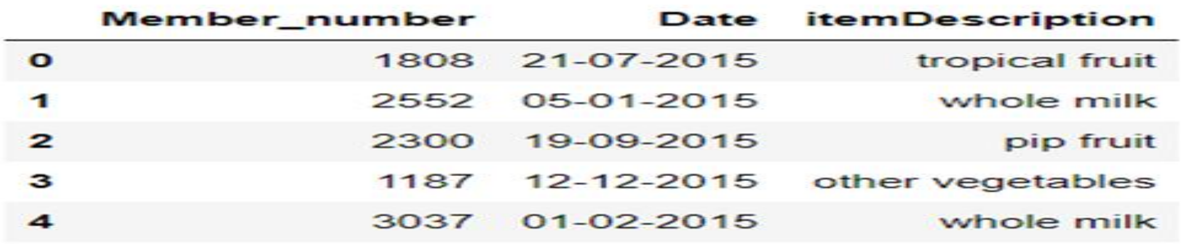
Finally, install the pandas and MLXtend libraries if you haven't already.

➢ Reading the Dataset.

Now, let's read the dataset as a pandas data frame and take a look at its head:

```
import pandas as pd

df = pd.read_csv('Groceries_dataset.csv')
df.head()
```



	Member_number	Date	itemDescription
0	1808	21-07-2015	tropical fruit
1	2552	05-01-2015	whole milk
2	2300	19-09-2015	pip fruit
3	1187	12-12-2015	other vegetables
4	3037	01-02-2015	whole milk

- **Data Preparation for Market Basket Analysis**

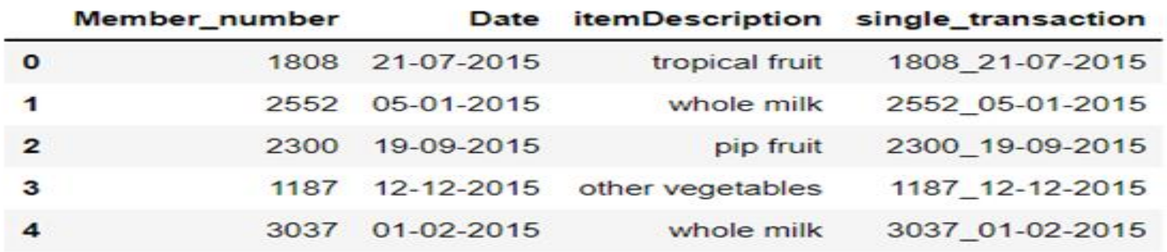
Before we perform market basket analysis, we need to convert this data into a format that can easily be ingested into the Apriori algorithm. In other words, we need to turn it into a tabular structure comprising ones and zeros, as displayed in the bread and milk example above.

To achieve this, the first group items that have the same member number and date:

```
df[‘single_transaction’] =
df[‘Member_number’].astype(str)+’_’+df[‘Date’].astype(str)
```

```
df.head()
```

This will provide us with a list of products purchased in the same transaction:

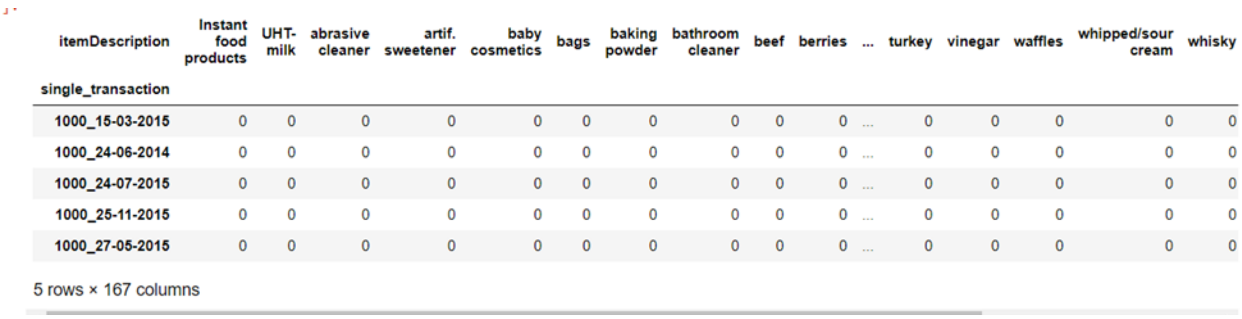


	Member_number	Date	itemDescription	single_transaction
0	1808	21-07-2015	tropical fruit	1808_21-07-2015
1	2552	05-01-2015	whole milk	2552_05-01-2015
2	2300	19-09-2015	pip fruit	2300_19-09-2015
3	1187	12-12-2015	other vegetables	1187_12-12-2015
4	3037	01-02-2015	whole milk	3037_01-02-2015

The “single_transaction” variable combines the member number, and date, and tells us the item purchased in one receipt.Now, let’s pivot this table to convert the items into columns and the transaction into rows:

```
df2 = pd.crosstab(df['single_transaction'], df['itemDescription'])
df2.head()
```

The resulting table tells us how many times each item has been purchased in one transaction:



itemDescription	Instant food products	UHT-milk	abrasive cleaner	artif. sweetener	baby cosmetics	bags	baking powder	bathroom cleaner	beef	berries	...	turkey	vinegar	waffles	whipped/sour cream	whisky
single_transaction																
1000_15-03-2015	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
1000_24-06-2014	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
1000_24-07-2015	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
1000_25-11-2015	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0
1000_27-05-2015	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0

5 rows × 167 columns

There are over a hundred columns while most people only shop for 2-3 items, which is why this table is sparse and mostly comprised of zeroes.The final data pre-processing step involves encoding all values in the above data frame to 0 and 1.This means that even if there are multiples of the same item in the same transaction, the value will be encoded to 1 since market basket analysis does not take purchase frequency into consideration.

```
def encode(item_freq):
    res = 0
    if item_freq > 0:
        res = 1
    return res


basket_input = df2.applymap(encode)
```

- **Build the Apriori Algorithm for Market Basket Analysis.**

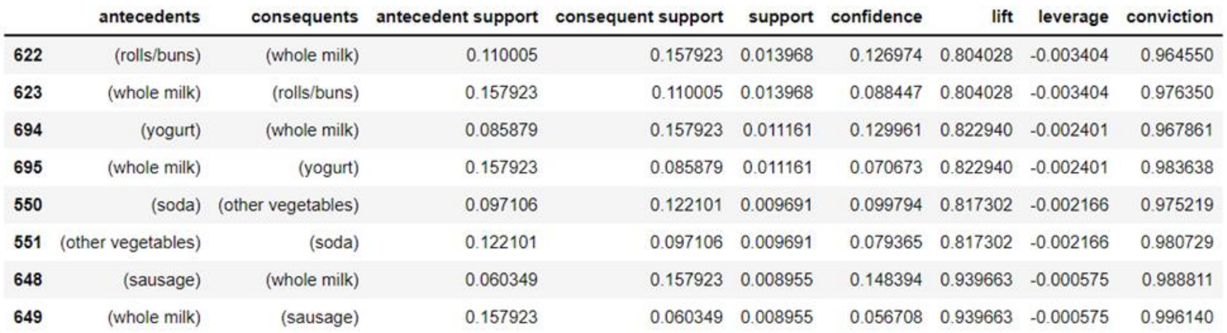
Now, let’s import the Apriori algorithm from the MLXtend Python package and use it to discover frequently-bought-together item combinations:

```
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules

frequent_itemsets = apriori(basket_input, min_support=0.001,
use_colnames=True)
```

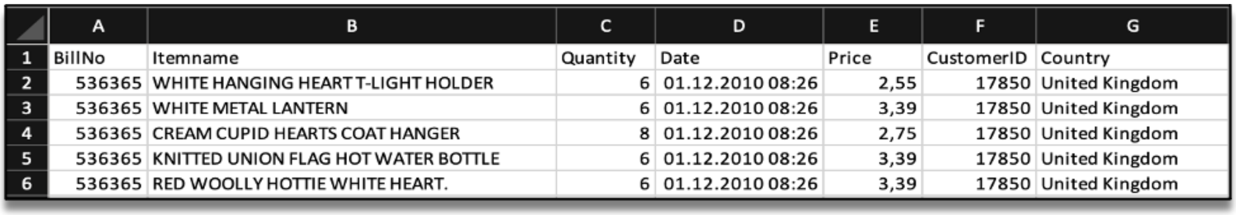
```
rules = association_rules(frequent_itemsets, metric="lift")

rules.head()
rules.sort_values(["support", "confidence","lift"],axis = 0, ascending =
False).head(8)
```



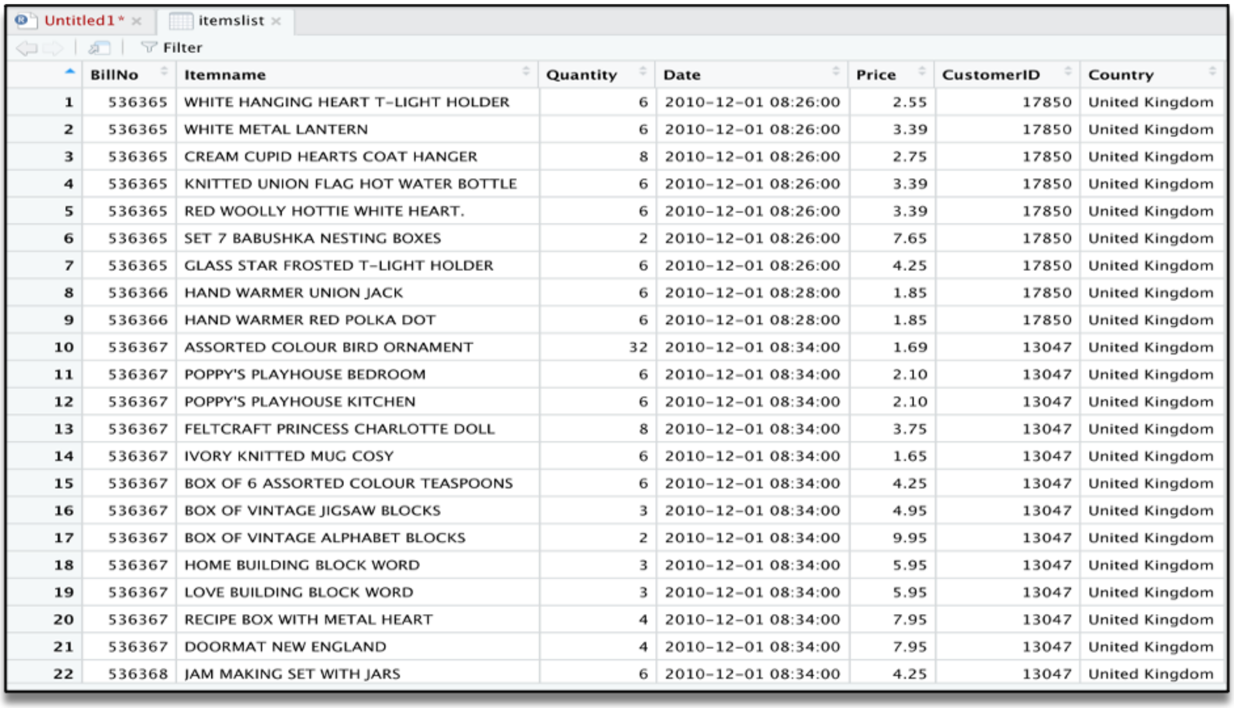
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
622	(rolls/buns)	(whole milk)	0.110005	0.157923	0.013968	0.126974	0.804028	-0.003404	0.964550
623	(whole milk)	(rolls/buns)	0.157923	0.110005	0.013968	0.088447	0.804028	-0.003404	0.976350
694	(yogurt)	(whole milk)	0.085879	0.157923	0.011161	0.129961	0.822940	-0.002401	0.967861
695	(whole milk)	(yogurt)	0.157923	0.085879	0.011161	0.070673	0.822940	-0.002401	0.983638
550	(soda)	(other vegetables)	0.097106	0.122101	0.009691	0.099794	0.817302	-0.002166	0.975219
551	(other vegetables)	(soda)	0.122101	0.097106	0.009691	0.079365	0.817302	-0.002166	0.980729
648	(sausage)	(whole milk)	0.060349	0.157923	0.008955	0.148394	0.939663	-0.000575	0.988811
649	(whole milk)	(sausage)	0.157923	0.060349	0.008955	0.056708	0.939663	-0.000575	0.996140

➢ Example Dataset:

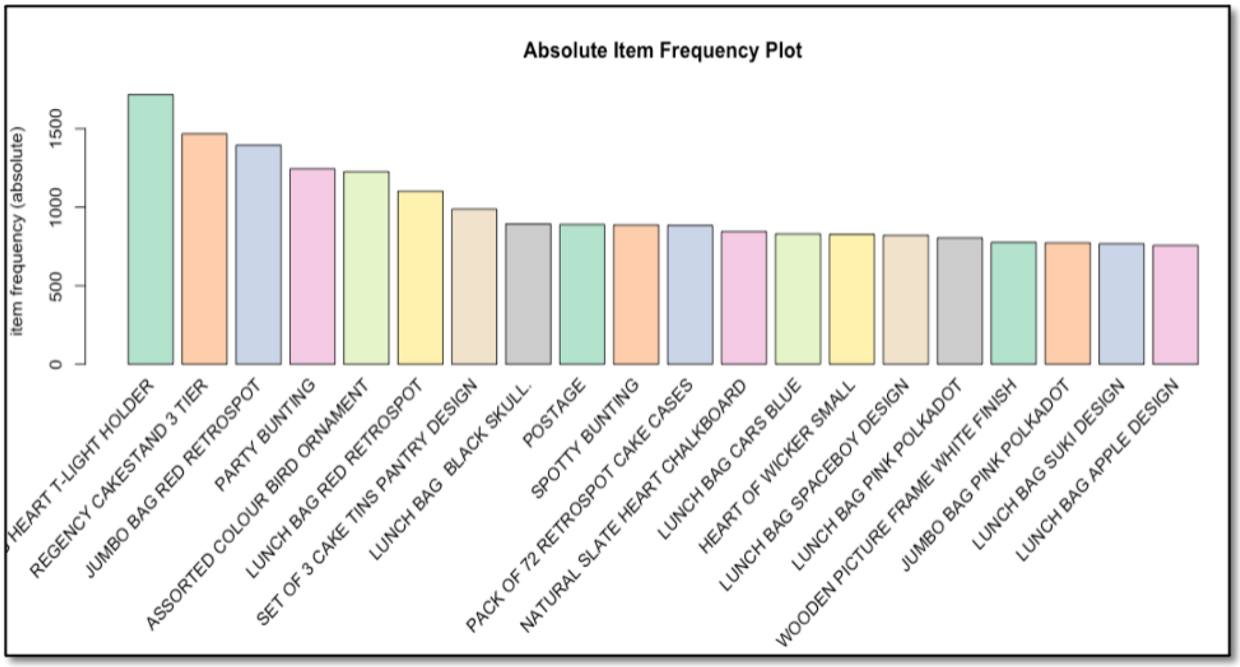


	A	B	C	D	E	F	G
1	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
2	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01.12.2010 08:26	2,55	17850	United Kingdom
3	536365	WHITE METAL LANTERN	6	01.12.2010 08:26	3,39	17850	United Kingdom
4	536365	CREAM CUPID HEARTS COAT HANGER	8	01.12.2010 08:26	2,75	17850	United Kingdom
5	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01.12.2010 08:26	3,39	17850	United Kingdom
6	536365	RED WOOLLY HOTTIE WHITE HEART.	6	01.12.2010 08:26	3,39	17850	United Kingdom

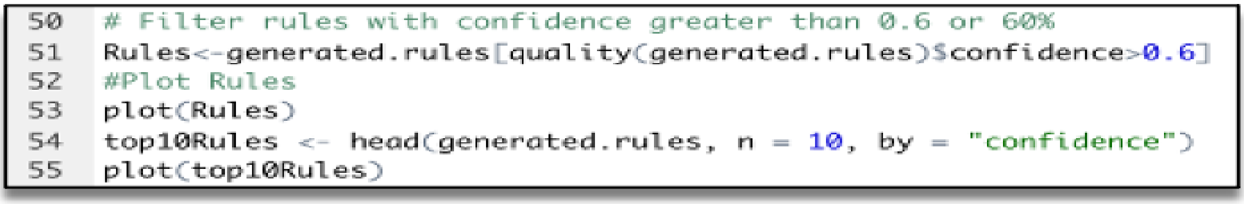
➢ Pre-processing data:



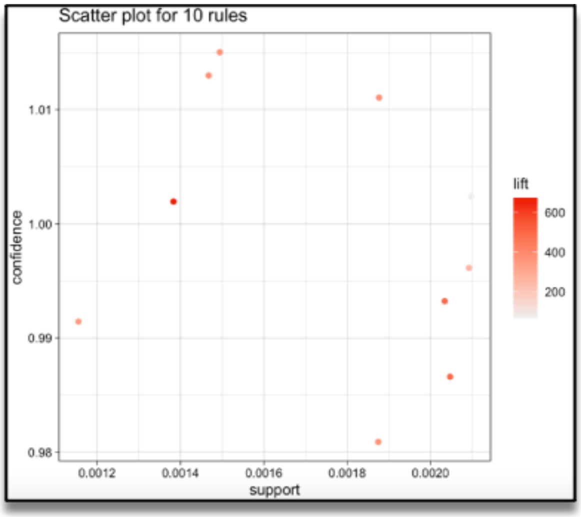
	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
1	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850	United Kingdom
2	536365	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
3	536365	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850	United Kingdom
4	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
5	536365	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850	United Kingdom
6	536365	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850	United Kingdom
7	536365	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850	United Kingdom
8	536366	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
9	536366	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850	United Kingdom
10	536367	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047	United Kingdom
11	536367	POPPY'S PLAYHOUSE BEDROOM	6	2010-12-01 08:34:00	2.10	13047	United Kingdom
12	536367	POPPY'S PLAYHOUSE KITCHEN	6	2010-12-01 08:34:00	2.10	13047	United Kingdom
13	536367	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	2010-12-01 08:34:00	3.75	13047	United Kingdom
14	536367	IVORY KNITTED MUG COSY	6	2010-12-01 08:34:00	1.65	13047	United Kingdom
15	536367	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	2010-12-01 08:34:00	4.25	13047	United Kingdom
16	536367	BOX OF VINTAGE JIGSAW BLOCKS	3	2010-12-01 08:34:00	4.95	13047	United Kingdom
17	536367	BOX OF VINTAGE ALPHABET BLOCKS	2	2010-12-01 08:34:00	9.95	13047	United Kingdom
18	536367	HOME BUILDING BLOCK WORD	3	2010-12-01 08:34:00	5.95	13047	United Kingdom
19	536367	LOVE BUILDING BLOCK WORD	3	2010-12-01 08:34:00	5.95	13047	United Kingdom
20	536367	RECIPE BOX WITH METAL HEART	4	2010-12-01 08:34:00	7.95	13047	United Kingdom
21	536367	DOORMAT NEW ENGLAND	4	2010-12-01 08:34:00	7.95	13047	United Kingdom
22	536368	JAM MAKING SET WITH JARS	6	2010-12-01 08:34:00	4.25	13047	United Kingdom



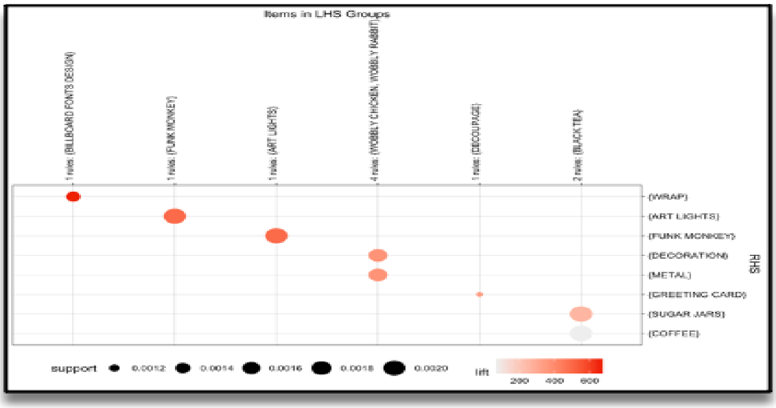
- **Scatter plot:**



```
# Filter rules with confidence greater than 0.6 or 60%
Rules<-generated.rules[quality(generated.rules)$confidence>0.6]
#Plot Rules
plot(Rules)
top10Rules <- head(generated.rules, n = 10, by = "confidence")
plot(top10Rules)
```



- ***Graph - Based Visualization and Group Method:***



- **Conclusion:**

Based on the results of these calculations can be used as a recommendation for retail owners to arrange the arrangement of product catalogs and take strategic steps to improve product marketing.. By utilizing the association rules which are discovered as a result of the analyses, the retailer can apply effective marketing and sales promotion strategies, he will be able increase customer engagement and improve customer experience and identify customer behavior.

SUBMITTED BY:

STUDENT REG NO:711221104056

NAAN MUDHALVAN: au711221104056