Phase-5

Documentation

Date	1 November 2023
Team ID	Proj-212168-Team-2
Project Name	Market Basket Insights
Maximum marks	

INDEX

List of figures and tables:

Abstract:

Market basket analysis is a critical module within the field of retail analytics that enables businesses to gain valuable insights into consumer purchasing behaviors. This module employs advanced data mining techniques to identify patterns, associations, and correlations among products frequently purchased together by customers. This abstract outlines the key components and objectives of the market basket analysis module, highlighting its significance in optimizing inventory management, personalized marketing strategies, and overall business profitability.

Introduction:

Market basket insights, also known as market basket analysis or association analysis, is a powerful data mining technique used by businesses and retailers to uncover valuable patterns and relationships within their transactional data. This method involves examining the purchasing habits of customers to identify which products or items are frequently bought together. By doing so, businesses can gain a deeper understanding of customer behavior, improve sales and marketing strategies, and enhance the overall customer experience.

Market basket insights are essential for various industries, including retail, e-commerce, and hospitality, as they enable organizations to make data-driven decisions and optimize product placements, pricing, and promotions. This analysis helps uncover cross-selling opportunities, allowing businesses to suggest complementary products to customers, increasing revenue and customer satisfaction.

Literature Survey:

Real-time Market Basket Insights:

In the context of real-time data analysis, the paper "Stream-Apriori: A Real-Time Sequential Pattern Mining Algorithm" by X. Zhang et al. (2017) presents a method for continuous market basket analysis.

Market Basket Insights for Cross-Selling and Upselling:

R. W. Palmatier et al.'s paper, "Linking the Service Profit Chain to Customer Loyalty" (2007), demonstrates how market basket insights can be used to improve customer loyalty and profitability through cross-selling and upselling.

Frequent Itemset Mining:

J. Han, J. Pei, Y. Yin, and R. Mao's paper, "Mining Frequent Patterns without Candidate Generation" (2000), presented the FP-growth algorithm, which improved upon the Apriori algorithm's efficiency in finding frequent itemsets.

Market Basket Analysis in Retail:

Michael J. A. Berry and Gordon S. Linoff's book, "Data Mining Techniques: For Marketing, Sales, and Customer Support" (1997), provides a comprehensive overview of how market basket analysis is used in retail, including case studies and practical applications.

Problem Definition:

The ultimate objective of this project is to gain profound insights into customer purchasing behavior and unveil potential cross-selling opportunities for a retail establishment. The project necessitates the application of association analysis techniques, notably the Apriori algorithm, to identify products that frequently co-occur in customer transactions.

The insights generated through this analysis will be instrumental in optimizing various facets of the business, from inventory management to marketing strategies, ultimately driving revenue growth and customer satisfaction.

Design Thinking:

Empathy:

Empathy in the context of market basket insights refers to the ability to understand and connect with the needs, preferences, and behaviors of customers as they make purchasing decisions.

Developing empathy for market basket insights is crucial for businesses to effectively analyze and respond to consumer behavior, tailor their marketing strategies, and optimize their product offerings.

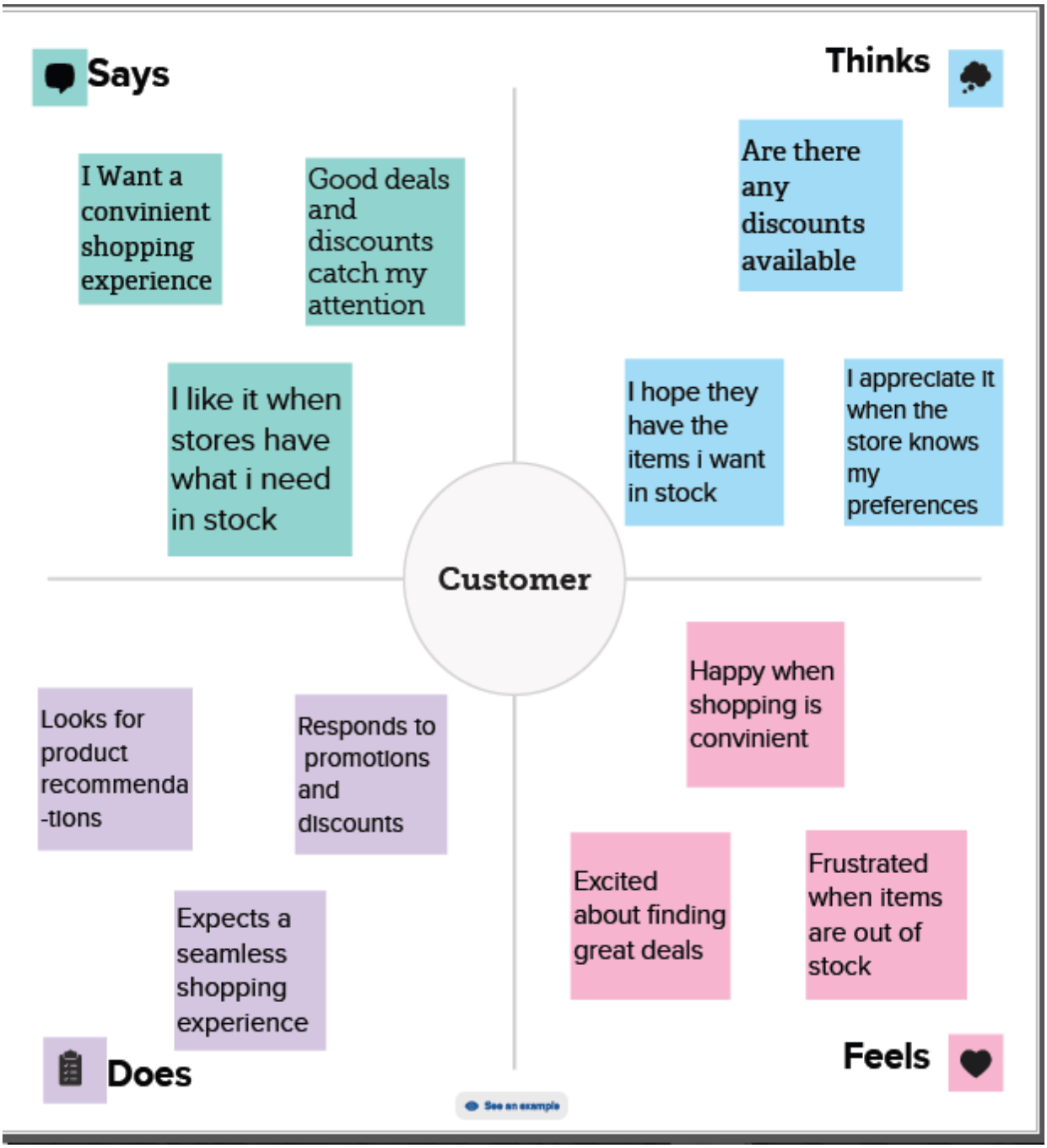

Figure 1

Brainstorm, Idea listing and grouping:

Market basket insights involve analyzing customer purchasing behavior to uncover patterns, relationships, and opportunities that can inform business decisions.

Grouping ideas is a helpful way to organize and categorize your thoughts, concepts, or strategies.

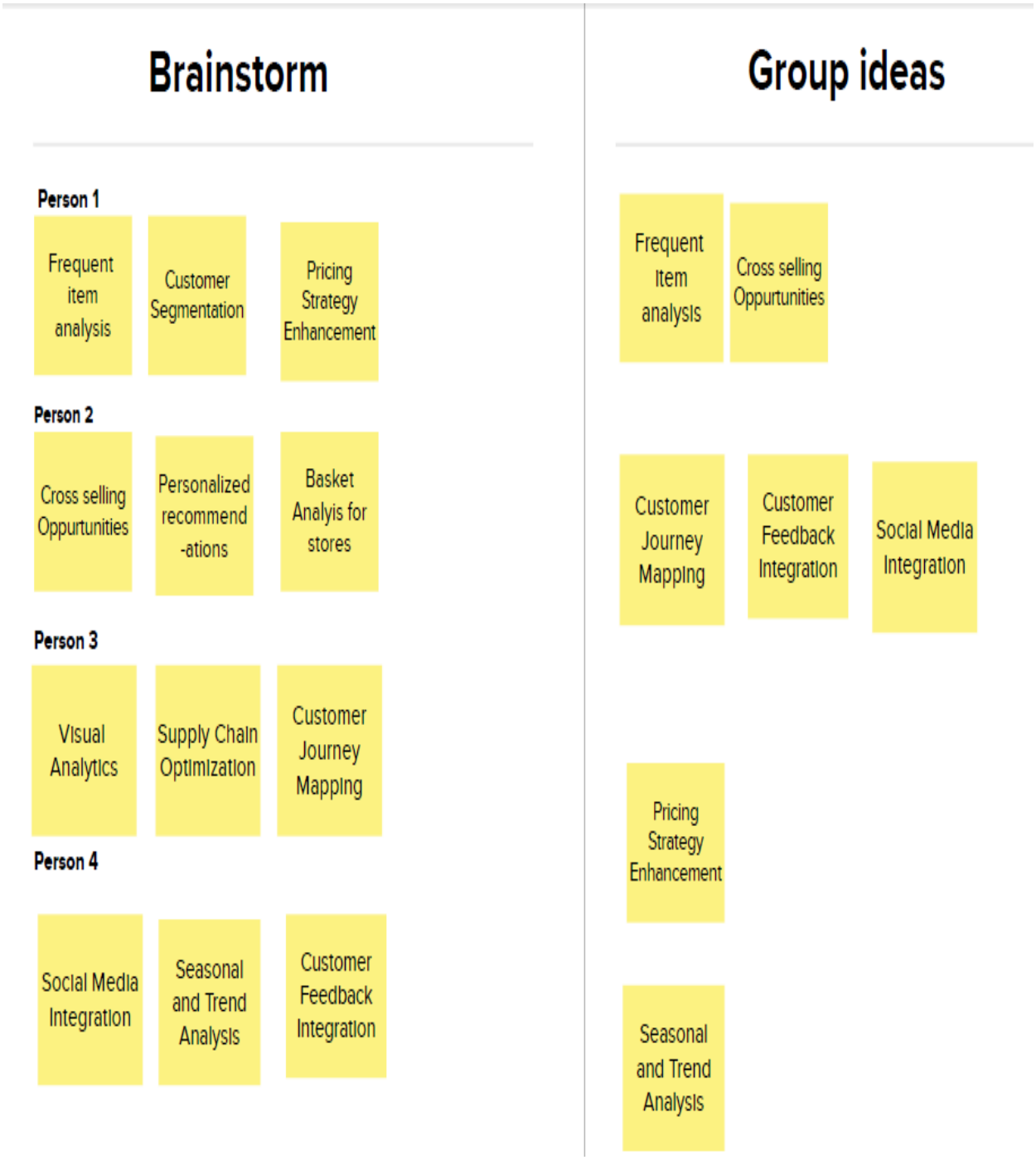

Figure 2

Idea Prioritization:

Prioritization is a critical process for individuals and organizations to allocate their time, resources, and efforts effectively. It involves determining which tasks, projects, or goals are most important and should be tackled first.

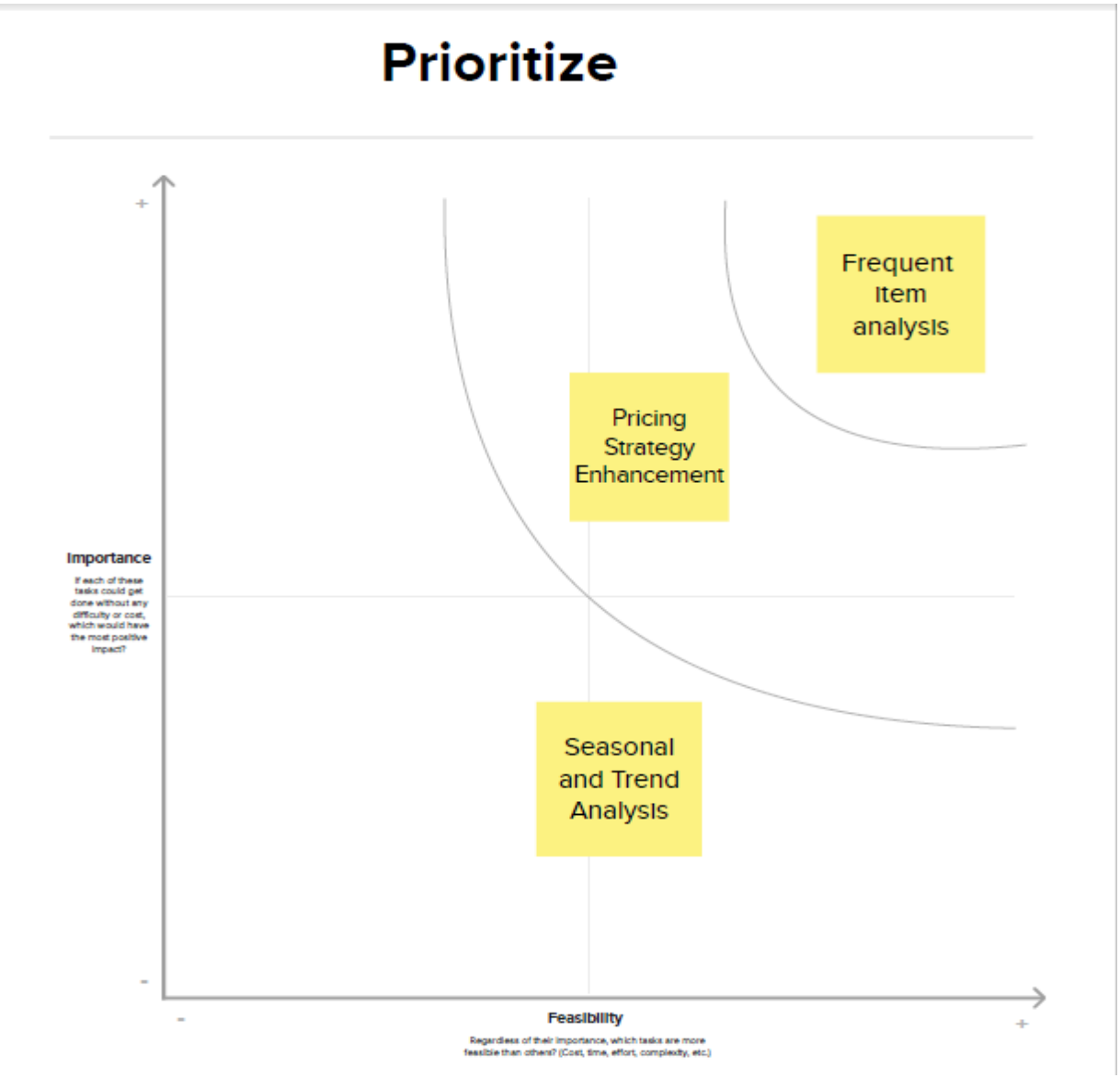

Figure 3

Innovation and Problem Solving:
Problem Statement:

1.Understanding Purchase Patterns	This understanding is crucial for optimizing product placements, enhancing marketing strategies, and increasing overall sales.
2. Customer Segmentation	Segmenting customers based on their shopping habits, preferences, and demographics.
3. Personalized Recommendations	Personalized product recommendations requires a deep understanding of individual preferences, purchase histories, and browsing behaviors.
4. Pricing Strategy Enhancement	Determine optimal pricing strategies, including bundling options, discounts, and promotions, that attract and retain customers while ensuring profitability.
5. Data Management and Analysis	Effectively handling and analyzing extensive transactional data is a resource-intensive task.
6. Real-Time Insights	In an era of rapid e-commerce and instantaneous decision-making, retailers often require real-time or near-real-time market basket insights to remain agile and responsive.

Table 1

Importing Dataset:

To import a dataset for market basket insights, you can use various programming languages and libraries, depending on your preferences and the format of your dataset. Common choices include Python with libraries like Pandas or R.

1. Prepare Your Dataset:

Ensure your dataset is in a suitable format for analysis, such as CSV, Excel, or a database. If it's not in the right format, you may need to perform data preprocessing to clean and format it appropriately.

2.Import the Library:

```
import pandas as pd
```

3.Load the Dataset:

Depending on the format of your dataset, you can use functions like pd.read_csv(), pd.read_excel()

```
datasets=pd.read_csv('dataset.csv')
```

Data cleaning and Analysis:

Data cleaning and analysis are crucial steps in any data science or analytics project. Data cleaning involves preparing the dataset for analysis by addressing missing values, handling outliers, and ensuring data consistency, while data analysis involves exploring and extracting insights from the cleaned data.

Data Cleaning:

1.Load the Data
2.Handle Missing Values
3.Dealing with Duplicates
4.Outlier Detection and Treatment
5.Data Transformation:

Data Analysis:

1.Exploratory Data Analysis (EDA)
2.Hypothesis Formulation
3.Statistical Analysis
4.Machine Learning and Predictive Analysis (Optional)
5.Visualization:

Data Visualization:

Data visualization is a powerful tool for gaining insights from market basket analysis, which is often used in retail to understand the relationships between products that customers purchase together. You can use libraries like Python's Matplotlib, Seaborn for creating data visualizations.

Scatter plot:

A scatter plot is used to visualize the relationship between two numerical variables.

Use plt.scatter() to create a scatter plot. Set a title, labels for the x and y axes, and a legend using plt.title(), plt.xlabel(), plt.ylabel(), and plt.legend().

Histogram:

Histogram in Python is a common data visualization technique, and use libraries like Matplotlib and Seaborn to easily generate histograms. Use plt.hist() to create the histogram. Specify the number of bins using the bins parameter.

Heatmap:

A heatmap is a graphical representation of data where individual values are represented as colors. It is a way to visualize data in a matrix or grid format. Heatmaps are commonly used to depict data in various fields, including data analysis, statistics, and data visualization.

Pairplot:

A pairplot is a type of data visualization that shows pairwise relationships between variables in a dataset. Create a pairplot using sns.pairplot().

Model Development and Evaluation:

Developing and evaluating a market basket analysis model typically involves the use of association rule mining algorithms, such as the Apriori algorithm, and the evaluation of these rules using relevant metrics. You can use programming languages like Python to accomplish this task.

Steps:

- First, load your transaction data into a pandas DataFrame. Each row represents a transaction, and each column represents an item, with binary values (1 for purchased, 0 for not purchased).
- Then use the Apriori algorithm to find frequent itemsets based on a minimum support threshold.
- Association rules are generated using a minimum confidence threshold.

- The code then displays the frequent itemsets and association rules.
- You can evaluate the rules based on other metrics like lift, conviction, etc. In the example, we filtered the rules with a minimum lift threshold of 0.5.

Code Sample:

```
#Import Packages
    import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    import seaborn as sns
    from mlxtend.frequent_patterns import apriori
    from mlxtend.frequent_patterns import association_rules

#import dataset
    datasets=pd.read_csv('dataset.csv')

#Data cleaning
    datasets.head()

    datasets.isnull().sum()

    datasets.info()
    df=datasets.fillna({'Itemname':'xyz'})
    df

    df1=datasets.fillna(value=datasets['CustomerID'].mean())
    df1

    df1.isnull().sum()
```

```
  #Finding Outliers
       Q1=df1['Quantity'].quantile(0.25)
       Q3=df1['Price'].quantile(0.75)
       IQR=Q3-Q1
       lowerbound=Q1-1.5*IQR
       upperbound=Q3+1.5*IQR
       outliers=df1[(df1['Quantity']<lowerbound)|
                          (df1['Price']>upperbound)]
        print(outliers)



#Scatter plot
     df1= pd.DataFrame(datasets)
     x=df1['Quantity']
     y=df1['Price']
     plt.scatter(x,y)
     plt.xlabel('Quantity')
     plt.ylabel('Price')
     plt.title('Market Basket Analysis')
     plt.show()

#Histogram
     sns.histplot(datasets,x='Price',bins=10,color='b')
#Heat map

    correlation_matrix = data.corr()
    sns.heatmap(correlation_matrix, annot=True)
    plt.title('Correlation Heatmap')
    plt.show()

#Formatting the transaction data in a suitable format for analysis
       df=pd.DataFrame(datasets)
       items_df=df[‘Itemname’].str.split(‘,’,expand=True)
```

```
    transaction_data=pd.concat([df,items_df],axis=1)

    transaction_data=transaction_data.drop('Itemname',axis=1)
    print(transaction_data.head( ))

# Converting items to Boolean columns
      df_encoded=pd.get_dummies(transaction_data,prefix='',
prefix_sep='') .groupby(level=0, axis=1).max( )
     df_encoded.to_csv('transaction_data_encoded.csv', index=False)

# Association Rule mining
    frequent_itemsets=apriori(df_encoded,min_support=0.007,
use_colnames=True)
    rules=association_rules(frequent_itemsets,metric="confidence",
min_threshold=0.5)

# Display information of the rules
    print("Association Rules:")
     print(rules.head())
```

The plot depicts the relationship between support, confidence, and lift for the generated association rules.

```
Plt.figure(figsize=(12,8))
Sns.scatterplot(x="support",y="confidence",size="lift",data=rules,
                        Hue="shift",palatte="viridis",sizes=(20,200))
plt.xlabel('Support')
plt.ylabel('Confidence')
plt.title('Market Basket Analysis-Support vs Confidence (size=lift)')
plt.legend(title='lift',loc='upper right',bbox_to_anchor=(1.2,1))
plt.show()
```

Output Screenshot:

```
In [4]: datasets.head()
```

0.027 seconds Explain... Format Copy 4

Out[4]:

	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom

Figure 4

```
In [5]: datasets.isnull().sum()
```

0.183 seconds Explain... Format Copy 5

```
Out[5]: BillNo          0
        Itemname       1455
        Quantity          0
        Date              0
        Price             0
        CustomerID   134041
        Country           0
        dtype: int64
```

Figure 5

```
In [6]: datasets.info()
```

0.198 seconds Explain... Format Copy 6

```
Out[6]: <class 'pandas.core.frame.DataFrame'>
RangeIndex: 522064 entries, 0 to 522063
Data columns (total 7 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   BillNo      522064 non-null  object
 1   Itemname    520609 non-null  object
 2   Quantity    522064 non-null  int64
 3   Date        522064 non-null  object
 4   Price       522064 non-null  float64
 5   CustomerID  388023 non-null  float64
 6   Country     522064 non-null  object
dtypes: float64(2), int64(1), object(4)
memory usage: 27.9+ MB
```

Figure 6

```
In [7]: df=datasets.fillna({'Itemname':'xyz'})
        df
```

0.072 seconds Explain... Format Copy 7

Out[7]:

	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...
522059	581587	PACK OF 20 SPACEBOY NAPKINS	12	09-12-2011 12:50	0.85	12680.0	France
522060	581587	CHILDREN'S APRON DOLLY GIRL	6	09-12-2011 12:50	2.10	12680.0	France
522061	581587	CHILDRENS CUTLERY DOLLY GIRL	4	09-12-2011 12:50	4.15	12680.0	France
522062	581587	CHILDRENS CUTLERY CIRCUS PARADE	4	09-12-2011 12:50	4.15	12680.0	France
522063	581587	BAKING SET 9 PIECE RETROSPOT	3	09-12-2011 12:50	4.95	12680.0	France

522064 rows × 7 columns

Figure 7

```
In [9]: df1=datasets.fillna(value=datasets['CustomerID'].mean())
        df1
```

0.267 seconds Explain... Format Copy

Out[9]:

	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
0	536365	WHITE HANGING HEART T-LIGHT HOLDER	6	01-12-2010 08:26	2.55	17850.0	United Kingdom
1	536365	WHITE METAL LANTERN	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
2	536365	CREAM CUPID HEARTS COAT HANGER	8	01-12-2010 08:26	2.75	17850.0	United Kingdom
3	536365	KNITTED UNION FLAG HOT WATER BOTTLE	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
4	536365	RED WOOLLY HOTTIE WHITE HEART.	6	01-12-2010 08:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...
522059	581587	PACK OF 20 SPACEBOY NAPKINS	12	09-12-2011 12:50	0.85	12680.0	France
522060	581587	CHILDREN'S APRON DOLLY GIRL	6	09-12-2011 12:50	2.10	12680.0	France
522061	581587	CHILDRENS CUTLERY DOLLY GIRL	4	09-12-2011 12:50	4.15	12680.0	France
522062	581587	CHILDRENS CUTLERY CIRCUS PARADE	4	09-12-2011 12:50	4.15	12680.0	France
522063	581587	BAKING SET 9 PIECE RETROSPOT	3	09-12-2011 12:50	4.95	12680.0	France

522064 rows × 7 columns

Figure 8

```
In [10]: df1.isnull().sum()

Out[10]: BillNo        0
         Itemname      0
         Quantity      0
         Date          0
         Price         0
         CustomerID    0
         Country       0
         dtype: int64
```

Figure 9

In [17]:
```
Q1=df1['Quantity'].quantile(0.25)
Q3=df1['Price'].quantile(0.75)
IQR=Q3-Q1
lowerbound=Q1-1.5*IQR
upperbound=Q3+1.5*IQR
outliers=df1[(df1['Quantity']<lowerbound)|(df1['Price']>upperbound)]
print(outliers)
```

Out[17]:

	BillNo	Itemname	Quantity	Date	Price	CustomerID	Country
16	536367	BOX OF VINTAGE ALPHABET BLOCKS	2	01-12-2010 08:34	9.95	13047.0	United Kingdom
45	536370	POSTAGE	3	01-12-2010 08:45	18.00	12583.0	France
65	536374	VICTORIAN SEWING BOX LARGE	32	01-12-2010 09:09	10.95	15100.0	United Kingdom
150	536382	3 TIER CAKE TIN GREEN AND CREAM	2	01-12-2010 09:45	14.95	16098.0	United Kingdom
151	536382	3 TIER CAKE TIN RED AND CREAM	2	01-12-2010 09:45	14.95	16098.0	United Kingdom
...	...	...	...	...	...	...	...
521922	581574	POSTAGE	2	09-12-2011 12:09	18.00	12526.0	Germany
521923	581578	POSTAGE	3	09-12-2011 12:16	18.00	12713.0	Germany
521941	581578	BOX OF VINTAGE ALPHABET BLOCKS	6	09-12-2011 12:16	11.95	12713.0	Germany
522004	581580	TABLECLOTH RED APPLES DESIGN	2	09-12-2011 12:20	9.95	12748.0	United Kingdom
522047	581586	RED RETROSPOT ROUND CAKE TINS	24	09-12-2011 12:49	8.95	13113.0	United Kingdom

[31717 rows x 7 columns]

Figure 10

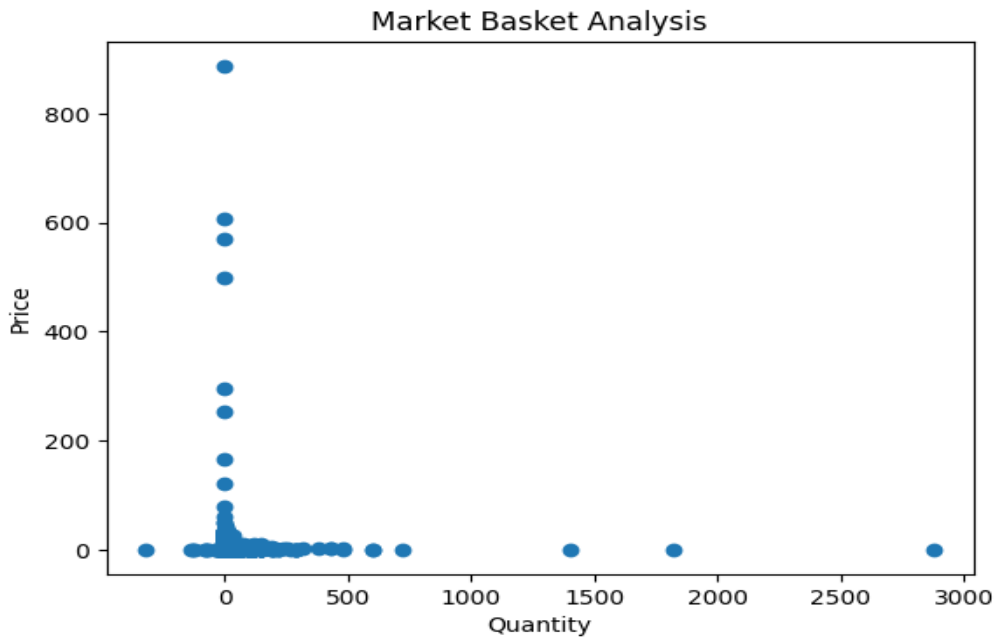

Figure 11

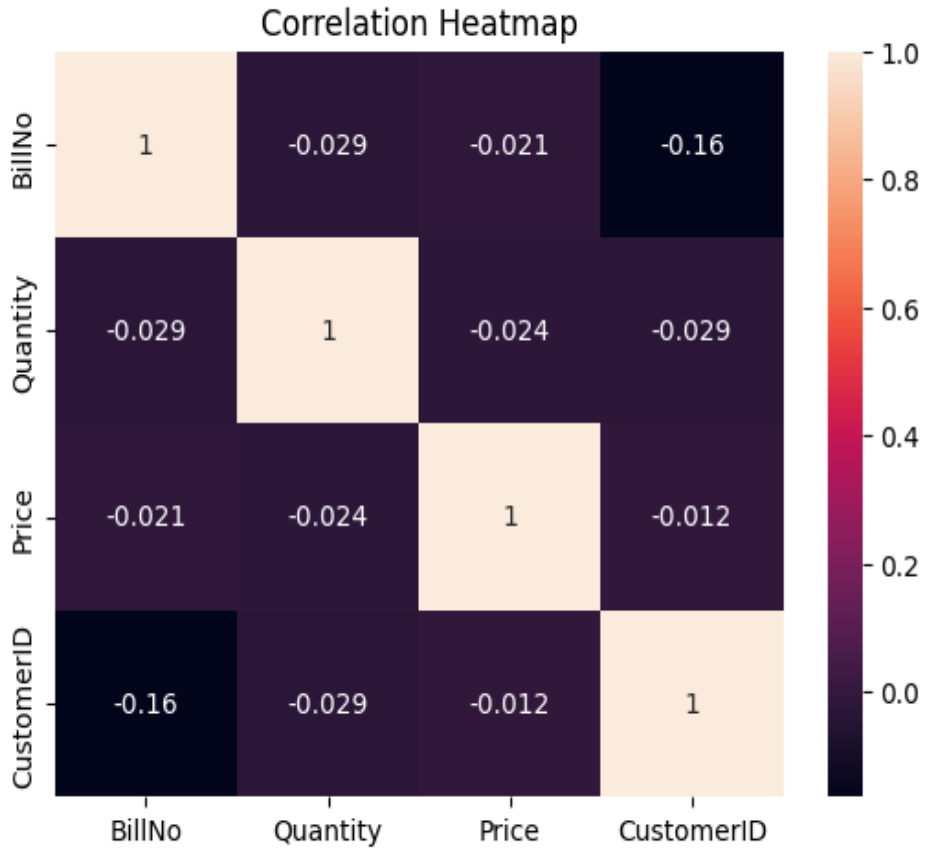

Figure 12

	0	1	\
0	WHITE HANGING HEART T-LIGHT HOLDER	WHITE METAL LANTERN	
1	HAND WARMER UNION JACK	HAND WARMER RED POLKA DOT	
2	ASSORTED COLOUR BIRD ORNAMENT	POPPY'S PLAYHOUSE BEDROOM	
3	JAM MAKING SET WITH JARS	RED COAT RACK PARIS FASHION	
4	BATH BUILDING BLOCK WORD	None	

	2	3	\
0	CREAM CUPID HEARTS COAT HANGER	KNITTED UNION FLAG HOT WATER BOTTLE	
1	None	None	
2	POPPY'S PLAYHOUSE KITCHEN	FELTCRAFT PRINCESS CHARLOTTE DOLL	
3	YELLOW COAT RACK PARIS FASHION	BLUE COAT RACK PARIS FASHION	
4	None	None	

	4	5	\
0	RED WOOLLY HOTTIE WHITE HEART.	SET 7 BABUSHKA NESTING BOXES	
1	None	None	

Figure 13

```
                                0                                   1  \
0  RED WOOLLY HOTTIE WHITE HEART.        SET 7 BABUSHKA NESTING BOXES
1                            None                                None
2         IVORY KNITTED MUG COSY  BOX OF 6 ASSORTED COLOUR TEASPOONS
3                            None                                None
4                            None                                None

                                6                                   7    \
0  GLASS STAR FROSTED T-LIGHT HOLDER                                None
1                            None                                None
2       BOX OF VINTAGE JIGSAW BLOCKS  BOX OF VINTAGE ALPHABET BLOCKS
3                            None                                None
4                            None                                None

                        8                       9    ...   534   535
536  \
0                       None                    None  ...  None  None
None
1                       None                    None  ...  None  None
None
1                       None                    None  ...  None  None
None
2  HOME BUILDING BLOCK WORD  LOVE BUILDING BLOCK WORD  ...  None  None
None
3                       None                    None  ...  None  None
None
4                       None                    None  ...  None  None
None

    537   538   539   540   541   542   543
0  None  None  None  None  None  None  None
1  None  None  None  None  None  None  None
2  None  None  None  None  None  None  None
3  None  None  None  None  None  None  None
4  None  None  None  None  None  None  None

[5 rows x 544 columns]
```

Figure 14

```
Association Rules:
                        antecedents                        consequents
\
0            (CHOCOLATE BOX RIBBONS)           (6 RIBBONS RUSTIC CHARM)
1  (60 CAKE CASES DOLLY GIRL DESIGN)  (PACK OF 72 RETROSPOT CAKE CASES)
2      (60 TEATIME FAIRY CAKE CASES)  (PACK OF 72 RETROSPOT CAKE CASES)
3   (ALARM CLOCK BAKELIKE CHOCOLATE)       (ALARM CLOCK BAKELIKE GREEN)
4   (ALARM CLOCK BAKELIKE CHOCOLATE)        (ALARM CLOCK BAKELIKE PINK)

   antecedent support  consequent support   support  confidence       li
ft  \
0            0.012368            0.039193  0.007036    0.568889  14.5150
44
1            0.018525            0.054529  0.010059    0.543027   9.9584
09
2            0.034631            0.054529  0.017315    0.500000   9.1693
55
3            0.017150            0.042931  0.011379    0.663462  15.4541
51
4            0.017150            0.032652  0.009125    0.532051  16.2947
42

   leverage  conviction  zhangs_metric
0  0.006551    2.228676       0.942766
1  0.009049    2.068984       0.916561
2  0.015427    1.890941       0.922902
3  0.010642    2.843862       0.951613
4  0.008565    2.067210       0.955009
```

Figure 15

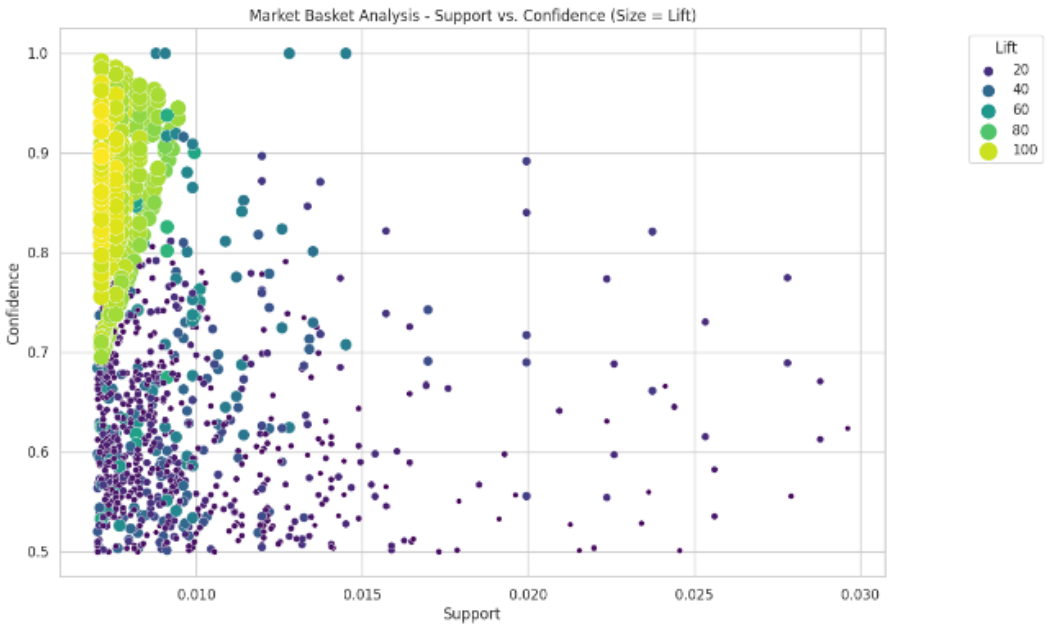

Figure 16

Conclusion:

Market basket insights provide a data-driven foundation for retailers to make informed decisions about product offerings, pricing, marketing, and customer experience. By harnessing the power of these insights, businesses can enhance their competitiveness, boost sales, and ultimately, better serve their customers. As technology and data analytics continue to evolve, market basket analysis will remain a vital tool for optimizing retail operations in an ever-changing marketplace.

References:

Apriori Algorithm and Association Rule Mining:

Rakesh Agrawal and Ramakrishnan Srikant's paper titled "Fast Algorithms for Mining Association Rules" (1994) introduced the Apriori algorithm, a fundamental method for discovering associations between items in transaction data.