Collage of computing

Data science

Data Analysis 2

DS3114

Project report

Task 2: Market Basket

Instructor: Omaima Fallatah

Student Name	**ID**
Fajr Faisal Al-Zahrani	44410657
Mayar Turki Al-Owaydhi	443003550

TABLE OF CONTACT

INTRODUCTION

In the age of big data, data analytics techniques have become vital tools for businesses across various sectors. The "Market Basket Analysis" project focuses on studying consumer behavior by analyzing data collected from purchasing transactions, enabling companies to understand patterns associated with buying behaviors. By employing algorithms such as FP-Growth to discover frequent itemset, analysts can identify the factors influencing purchasing decisions, which helps enhance marketing strategies and increase revenue.

DATASET

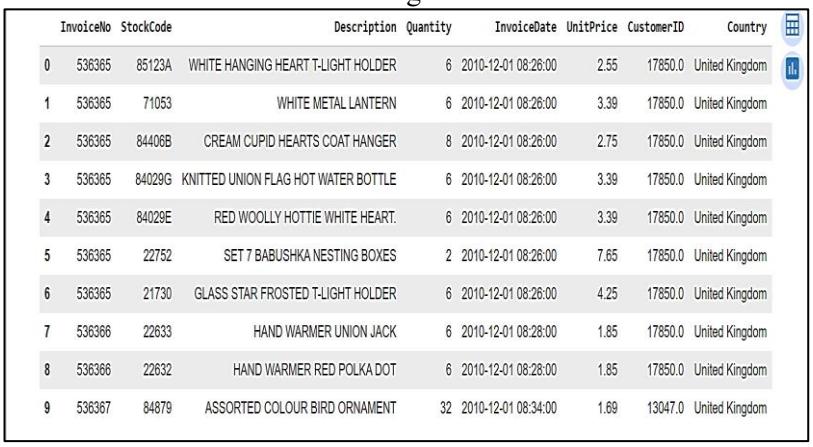
archive.ics.uci.edu. (n.d.). *UCI Machine Learning Repository*. [online] Available at: https://archive.ics.uci.edu/dataset/352/online+retail.

OBJECTIVES

1. **Data Analysis**: Study and analyze purchasing transaction data to understand prevalent patterns in consumer behavior.
2. **Frequent Itemset Discovery**: Apply the FP-Growth algorithm to identify commonly purchased product groups.
3. **Association Rule Evaluation**: Analyze the resulting rules to determine strong relationships between products, focusing on metrics such as support, confidence, and lift.
4. **Recommendation Development**: Provide recommendations based on the findings to improve marketing strategies and increase sales through effective promotion of related products.

EXPLORATORY DATA ANALYSIS (EDA)

Figure 1



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850.0	United Kingdom
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850.0	United Kingdom
7	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
8	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047.0	United Kingdom

- Description Figure1: shows the first few rows of the dataset loaded from the "Online Retail.xlsx" file. The displayed columns include **InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country.**

- Purpose: It provides a snapshot of the transaction data, allowing for initial validation of the data structure and contents, which is crucial for understanding the scope of the analysis.

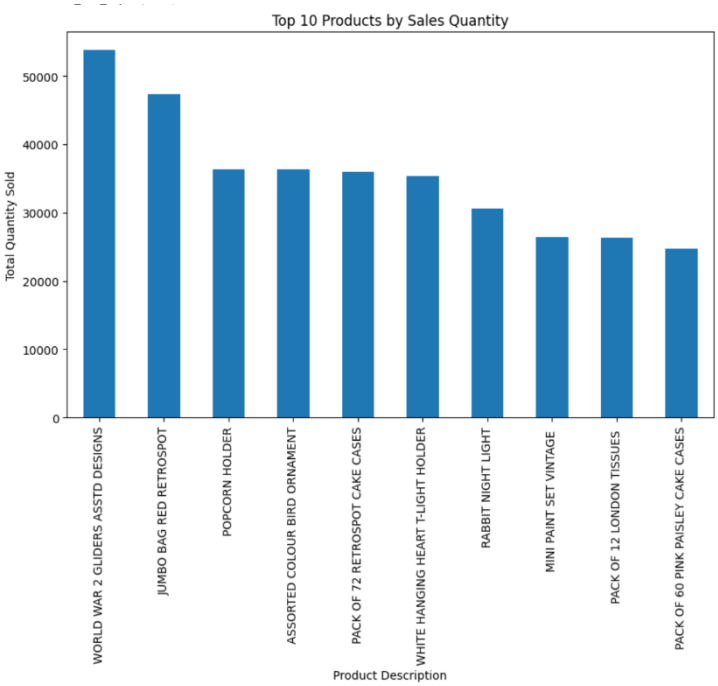
Figure 2

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Date	year	month	day	day_name
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom	2010-12-01 08:26:00	2010	12	1	Wednesday
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	2010-12-01 08:26:00	2010	12	1	Wednesday
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom	2010-12-01 08:26:00	2010	12	1	Wednesday
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	2010-12-01 08:26:00	2010	12	1	Wednesday
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom	2010-12-01 08:26:00	2010	12	1	Wednesday
5	536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	7.65	17850.0	United Kingdom	2010-12-01 08:26:00	2010	12	1	Wednesday
6	536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	4.25	17850.0	United Kingdom	2010-12-01 08:26:00	2010	12	1	Wednesday
7	536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom	2010-12-01 08:28:00	2010	12	1	Wednesday
8	536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	1.85	17850.0	United Kingdom	2010-12-01 08:28:00	2010	12	1	Wednesday
9	536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	1.69	13047.0	United Kingdom	2010-12-01 08:34:00	2010	12	1	Wednesday

- Description Figure2: captures the data processing steps, including data cleaning, date conversion, and the creation of additional columns such as year, month, day, and day name **from the InvoiceDate.**

- Purpose: These steps are crucial for preparing the data for analysis. Properly formatted dates allow for more nuanced insights when examining sales trends over time.

Figure3



- **Description Figure3**: features a bar chart displaying the top 10 products by total sales quantity. The chart includes product descriptions on the x-axis and the total quantity sold on the y-axis.

- **Purpose**: This visualization helps identify the best-selling products, providing insights into customer preferences and trends. Understanding which products are most popular can guide inventory management and marketing strategies.

TASK 2: MARKET BASKET

A Market Basket Analysis was conducted using sales data extracted from the "Online Retail.xlsx" file. The FP-Growth algorithm was applied to discover patterns and relationships between products purchased together by customers, with the aim of enhancing marketing strategies and increasing store sales.

FP-Growth:

1. **Data Aggregation:** Products associated with each invoice were grouped into a single list using groupby, allowing for the analysis of purchasing behavior for each customer.
2. **Data Transformation to Binary Format:** The list of products was converted into a binary-encoded DataFrame to facilitate the application of the FP-Growth algorithm.
3. **Application of FP-Growth Algorithm**: The FP-Growth algorithm was executed to identify frequent itemsets with a minimum support threshold of 1%.
4. **Results Evaluation**:

   - The frequent itemsets were sorted by support, displaying the top 10 itemsets.

   - Association rules were applied using lift as a metric to assess relationships, with a minimum lift threshold specified.

Results of FP-Growth:

Figure4

- Top 10 Frequent Itemsets:A set of the most frequently occurring products was obtained, indicating customer preferences, such as: (WHITE HANGING HEART T-LIGHT HOLDER),(REGENCY CAKESTAND 3 TIER),(JUMBO BAG RED RETROSPOT)

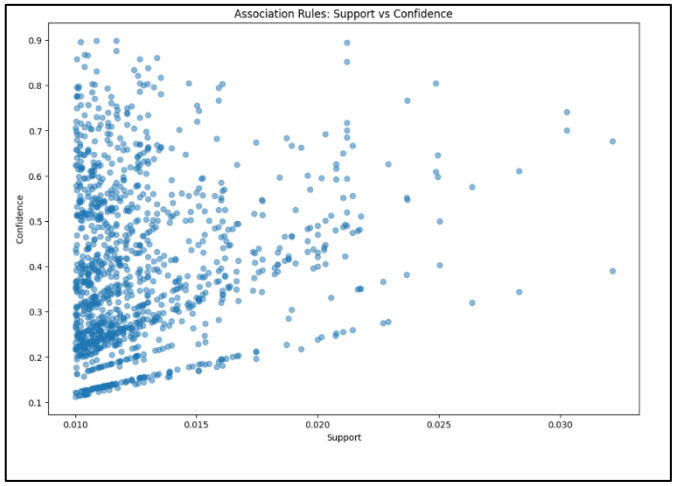
- Top 10 Association Rules:The results also included association rules that highlight relationships between products: (GREEN REGENCY TEACUP AND SAUCER) associated with (REGENCY CAKESTAND 3 TIER) with high support and confidence.

```
      support                            itemsets
0    0.088880  (WHITE HANGING HEART T-LIGHT HOLDER)
263  0.083745            (REGENCY CAKESTAND 3 TIER)
90   0.082432             (JUMBO BAG RED RETROSPOT)
446  0.065869                        (PARTY BUNTING)
40   0.062046             (LUNCH BAG RED RETROSPOT)
7    0.056641       (ASSORTED COLOUR BIRD ORNAMENT)
466  0.056293   (SET OF 3 CAKE TINS PANTRY DESIGN )
41   0.051506     (PACK OF 72 RETROSPOT CAKE CASES)
161  0.050000             (LUNCH BAG  BLACK SKULL.)
77   0.048880     (NATURAL SLATE HEART CHALKBOARD )
                                        antecedents  \
940   (ROSES REGENCY TEACUP AND SAUCER , REGENCY CAK...
1191                          (REGENCY TEA PLATE PINK)
342   (SET/6 RED SPOTTY PAPER CUPS, SET/20 RED RETRO...
953   (ROSES REGENCY TEACUP AND SAUCER , PINK REGENC...
941   (GREEN REGENCY TEACUP AND SAUCER, REGENCY CAKE...
268   (JUMBO STORAGE BAG SUKI, JUMBO BAG PINK POLKAD...
1193                          (REGENCY TEA PLATE PINK)
928   (REGENCY CAKESTAND 3 TIER, PINK REGENCY TEACUP...
179   (CHARLOTTE BAG PINK POLKADOT, STRAWBERRY CHARL...
210   (WOODLAND CHARLOTTE BAG, CHARLOTTE BAG SUKI DE...

                           consequents   support  confidence       lift
940    (GREEN REGENCY TEACUP AND SAUCER)  0.011699    0.899110  22.031167
1191          (REGENCY TEA PLATE GREEN )  0.010888    0.898089  60.260387
342      (SET/6 RED SPOTTY PAPER PLATES)  0.010232    0.895270  43.999051
953    (GREEN REGENCY TEACUP AND SAUCER)  0.021197    0.894137  21.909313
941   (ROSES REGENCY TEACUP AND SAUCER )  0.011699    0.875723  20.251084
268            (JUMBO BAG RED RETROSPOT)  0.010386    0.867742  10.526705
1193          (REGENCY TEA PLATE ROSES )  0.010502    0.866242  49.093367
928    (GREEN REGENCY TEACUP AND SAUCER)  0.013359    0.860697  21.089915
179        (RED RETROSPOT CHARLOTTE BAG)  0.012664    0.858639  21.179756
210        (RED RETROSPOT CHARLOTTE BAG)  0.010077    0.858553  21.177632
/usr/local/lib/python3.10/dist-packages/ipykernel/ipkernel.py:283: DeprecationWarn
```

Figure5

- Description: Displays the relationship between support and confidence for association rules, with support on the x-axis and confidence on the y-axis.

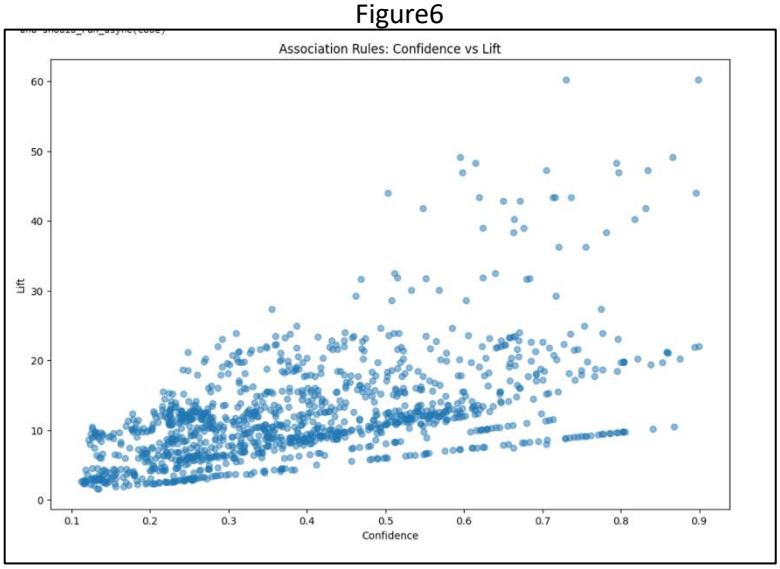


- Insights:

1. **General Trend**: Increased support does not guarantee high confidence, indicating variability in product relationships.
2. **Rule Quality**: Rules with high support and confidence are more valuable for marketing strategies.

- Description: Shows the relationship between confidence and lift, with confidence on the x-axis and lift on the y-axis.

Figure6



- Insights:

1. **Positive Correlation**: Higher confidence typically leads to higher lift, reflecting strong product associations.
2. **Marketing Implications**: Rules with high confidence and lift are ideal for marketing strategies.

CONCLUSION

In conclusion, this project provided a comprehensive analysis of customer purchasing behaviors through Market Basket Analysis using sales data extracted from the "Online Retail.xlsx" file. By applying the FP-Growth algorithm, we successfully identified patterns and relationships between products that are frequently purchased together.

The results revealed the top-selling products and the most common itemsets, offering valuable insights into consumer preferences. Additionally, the extracted association rules highlighted strong relationships among products, which can be leveraged for targeted marketing strategies. For instance, high support and confidence values indicated key products that could be effectively promoted together to boost sales.

The visualizations created, such as scatter plots comparing support versus confidence and confidence versus lift, helped clarify the quality of the association rules. These visual tools provided a clearer understanding of how product relationships function within the data, aiding in informed decision-making.

Overall, this project highlights the power of data analysis in understanding customer behavior and equips businesses with actionable insights to optimize marketing efforts and improve overall sales performance. By leveraging these findings, companies can better meet consumer needs and adapt to changing market dynamics.