A Capstone Project Report

On

Market Basket Analysis Using Python

Submitted by

D.AJAY KUMAR (192211089)
J.JOEL ANDREW (192211111)
M.RAM JEYANTH (192211167)
S.HARI HARAN(192211110)

Under the guidance of

Dr. C. ROHITH BHAT

In partial fulfillment for the completion of the course

CSA1358- THEORY OF COMPUTATION FOR POST CORRESPONDANCE PROBLEM

SIMATS ENGINEERING

THANDALAM

JULY 2024

DECLARATION

We, **D. Ajay Kumar, J. Joel Andrew, M. Ram Jeyanth and**, **S.Hari Haran** students of **Bachelor of Engineering in Computer Science & Engineering**, Department of Computer Science and Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, with this we declare that the work presented in this Capstone Project Work entitled **Market Basket Analysis Software** is the outcome of our bonafide work and is correct to the best of our knowledge. This work has been undertaken while taking care of engineering ethics.

D. Ajay Kumar
J. Joel Andrew
M. Ram Jeyanth
S. Hari Haran

Date: 25.07.2024

Place: Chennai

Table of Contents

ABSTRACT

This project focuses on market basket analysis using Python, implementing the Apriori algorithm to identify frequently purchased itemsets in transaction data. By preprocessing and analyzing the data, the algorithm generates association rules that reveal relationships between items. These patterns are visualized using various data visualization techniques, providing clear insights into purchasing behaviors. The project aims to optimize retail strategies by highlighting key findings and offering actionable recommendations. The results, including visualizations and detailed reports, are exported for further analysis and strategic planning, enhancing decision-making processes in retail management.

INTRODUCTION

Market basket analysis is a powerful data mining technique extensively used in the retail industry to uncover relationships between items that are frequently purchased together. By examining transaction data, businesses can identify patterns and correlations among products, which can be leveraged to optimize various aspects of retail operations. For example, understanding which items are commonly bought together can help in designing more effective store layouts, where related products are placed near encourage impulse purchases. Additionally, this analysis can inform marketing strategies, such as targeted promotions and discounts for complementary products, thereby increasing sales and customer engagement. Moreover, insights from market basket analysis can enhance inventory management by predicting demand for grouped items, reducing the likelihood of stockouts and overstock situations. The Apriori algorithm is particularly favored in market basket analysis due to its efficiency in identifying frequent item sets and generating meaningful association rules, making it a valuable tool for retail data analysis.

This project uses Python to implement the Apriori algorithm for market basket analysis. The process begins with data loading and preprocessing, ensuring the transaction data is clean and suitable for analysis. The Apriori algorithm is then applied to identify frequent item sets based on specified support and confidence thresholds. Visualizations are created to represent these itemsets and their relationships, providing a clear understanding of purchasing behaviors. In this project, we employ Python to execute a comprehensive market basket analysis using the Apriori algorithm. The first step involves loading and preprocessing the transaction data to ensure it is clean, structured, and ready for analysis. This includes handling any missing values, encoding categorical data, and converting the data into a format suitable for the Apriori algorithm. Once the data is prepared, the Apriori algorithm is applied to discover frequent item sets, which are sets of items that appear together in transactions more frequently than a specified threshold.

The insights derived from the market basket analysis are crucial for making informed business decisions and enhancing retail strategies. By analyzing the generated association rules and frequent itemsets, businesses can uncover significant patterns in customer purchasing behavior. For instance, identifying that certain products are often bought together can lead to strategic decisions such as bundling these items in promotional offers or placing them together in-store to boost sales. The recommendations based on these insights help in crafting targeted marketing strategies, such as personalized discounts or cross-selling opportunities, thereby increasing customer engagement and loyalty.

Furthermore, these insights are instrumental in improving inventory management. By predicting the demand for grouped items, businesses can adjust their stock levels to align with expected sales, minimizing the risk of stockouts and reducing excess inventory. The results and visualizations are exported in accessible formats, allowing for easy integration into existing business systems and reporting tools. This facilitates the application of the findings in real-world scenarios, helping businesses to make data-driven decisions that enhance overall operational efficiency and customer satisfaction.

PROJECT DESCRIPTION :

This project utilizes the Apriori algorithm to perform market basket analysis, applying Python to discover frequent item sets and generate association rules from transaction data. By visualizing these patterns, the project aims to provide actionable insights for optimizing retail strategies, improving product placement, and enhancing inventory management. The findings are exported in various formats for easy integration into business operations and decision-making processes.

PROBLEM DESCRIPTION :

Existing systems for market basket analysis often rely on traditional methods that may not efficiently handle large volumes of transaction data or uncover complex relationships between items. Many systems struggle with scalability issues, leading to performance bottlenecks when analyzing extensive datasets. Additionally, these systems may lack advanced visualization capabilities, making it challenging for users to interpret and act on the data insights effectively. As a result, businesses may miss out on valuable opportunities for optimizing product placement, crafting targeted promotions, and improving inventory management. There is a need for a more robust solution that leverages advanced algorithms and provides clear, actionable insights through intuitive visualizations to enhance decision-making and operational efficiency.

TOOL DESCRIPTION :

The tool leverages Python to implement the Apriori algorithm for market basket analysis, focusing on efficiently discovering frequent itemsets and generating association rules from transaction data. It handles data loading, preprocessing, and applies the Apriori algorithm to identify significant itemsets based on user-defined thresholds for support and confidence. The tool's core functionality includes identifying patterns and relationships among items, which are crucial for understanding purchasing behaviors.

In addition to its analytical capabilities, the tool incorporates advanced data visualization features designed to present the results in a user-friendly and intuitive format. These visualizations include bar charts, scatter plots, and network graphs, each chosen for their ability to convey different aspects of the data effectively. Bar charts provide a straightforward way to compare the frequency of itemsets, while scatter plots can highlight relationships and trends between items. Network graphs offer a visual representation of the associations between items, illustrating how they are connected through the generated rules.

The tool's visualization capabilities are not just for display; they play a crucial role in interpreting complex data patterns and deriving actionable insights. Users can interact with the visualizations to explore different facets of the data, making it easier to identify key trends and correlations. Furthermore, the tool supports exporting results in various formats such as CSV, Excel, and PDF. This functionality ensures that users can seamlessly integrate the findings into their business strategies and reporting systems, facilitating better decision-making. By providing a comprehensive and accessible view of the data, the tool enhances operational efficiency and helps businesses leverage insights to optimize their retail strategies and improve overall performance.

OPERATIONS :

- **Load and Prepare Data:** This feature enables users to import and preprocess transaction data from various sources such as CSV files or databases. It ensures the data is clean and formatted correctly for analysis using the Apriori algorithm, allowing for effective and accurate market basket analysis.
- **Apply Apriori Algorithm:** This functionality processes the cleaned transaction data to identify frequent item sets and generate association rules based on user-defined support and confidence thresholds. The algorithm helps in discovering meaningful patterns and relationships between items in the data.
- **Visualize Patterns:** This tool generates visual representations of the identified itemsets and association rules using bar charts, scatter plots, and network graphs. These visualizations facilitate the interpretation of complex data patterns and provide actionable insights for optimizing retail strategies.
- **Generate Insights:** Leveraging the results from the Apriori algorithm, this feature generates detailed reports and insights into purchasing behaviors. It highlights significant item relationships and trends, providing recommendations for improving product placement and marketing strategies.
- **Export Results:** This feature allows users to export the analysis results, visualizations, and reports in various formats such as CSV, Excel, and PDF. It also includes options to reset the tool and access information about the software, ensuring users can easily integrate findings into business strategies and reporting systems.

FUNCTIONALITIES :

The tool provides a range of functionalities designed to streamline market basket analysis and generate actionable insights from transaction data. Users can import and preprocess data from various sources, ensuring it is clean and properly formatted for analysis. The Apriori algorithm is then applied to identify frequent item sets and generate association rules, which are crucial for uncovering relationships between items. The tool's visualizations, including bar charts, scatter plots, and network graphs, help users interpret complex data patterns and understand purchasing behaviors.

The tool offers comprehensive functionalities aimed at simplifying market basket analysis and deriving actionable insights from transaction data. Users can easily import data from various sources, such as CSV files or databases, and preprocess it to ensure it is clean and well-structured for analysis. This preprocessing step is crucial for ensuring that the Apriori algorithm can accurately identify frequent itemsets and generate reliable association rules. The algorithm processes the data to uncover patterns and relationships between items based on user-defined support and confidence thresholds. The visual representation of these results through bar charts, scatter plots, and network graphs helps users understand complex data patterns and interpret purchasing behaviors more effectively. These visualizations make it easier to spot trends and correlations that might not be immediately apparent from raw data alone.

Additionally, the tool provides functionalities for generating detailed insights and reports based on the analyzed data. Users can access comprehensive summaries and recommendations derived from the association rules, which can be used to optimize retail strategies such as product placement, marketing campaigns, and inventory management.

IMPLEMENTATION: (PSEUDO CODE)

Initialize Data:

```
Try:
    Print "Loading data..."
    Set status_label to "Loading data..."
    Load data from FILE_PATH into dataframe 'df'
    Drop rows where 'Description' is NaN
    Group data by 'Invoice' and create a list of 'Description' per invoice
    Transform data using TransactionEncoder and create dataframe 'df_transformed'
    Print "Performing market basket analysis..."
    Set status_label to "Performing market basket analysis..."
    Apply Apriori algorithm on 'df_transformed' with min_support=0.01
    Generate association rules from frequent itemsets with metric="lift" and min_threshold=1
    Print "Displaying rules..."
    Set status_label to "Displaying rules..."
    Display rules in the Treeview
    Set status_label to "Data loaded successfully!"
Catch Exception:
    Print error message
    Show error dialog with the message
    Set status_label to "Error loading data"
```

Display Rules (with filtering):

```
Try:
    Clear all rows from Treeview
    Filter rules based on min_support, min_confidence, and min_lift
    For each filtered rule:
        Insert rule into Treeview with antecedents, consequents, support, confidence, and lift
Catch Exception:
    Print error message
    Show error dialog with the message
```

Plot Graph:

```
Try:
        Create a scatter plot of support vs confidence
        Set labels and title for the plot
        Embed plot into the Tkinter window
    Catch Exception:
        Print error message
        Show error dialog with the message
```

Save Rules to File (with file type):

```
    Try:
        Filter rules based on support, confidence, and lift values
        Depending on file_type ('csv', 'excel', 'json'):
            Save filtered rules to the corresponding file format
            Show success dialog with the file name
    Catch Exception:
        Print error message
        Show error dialog with the message
```

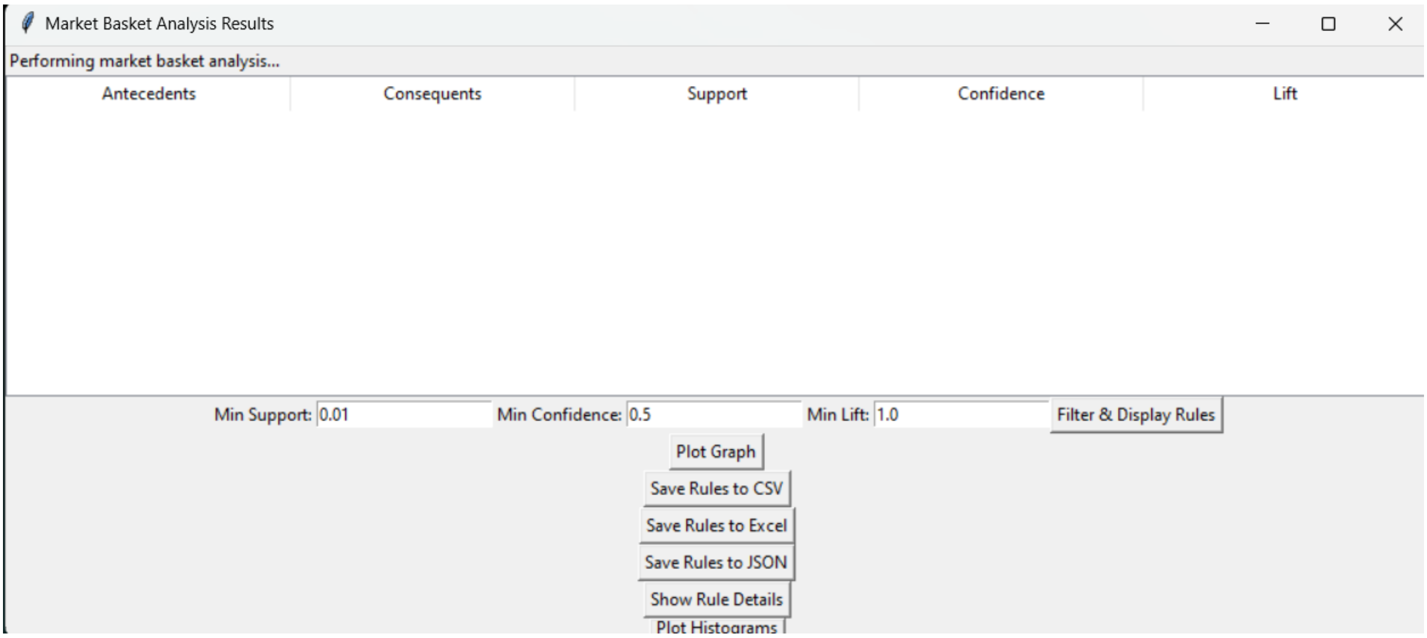
Show Rule Details:

```
    Try:
        Get selected item from Treeview
        Retrieve rule details (antecedents, consequents, support, confidence, lift)
        Show rule details in a message dialog
    Catch Exception:
        Print error message
        Show error dialog with the message
```

Plot Histograms:

```
    Try:
        Create histograms for support, confidence, and lift
        Set titles for each histogram
        Embed histograms into the Tkinter window
```

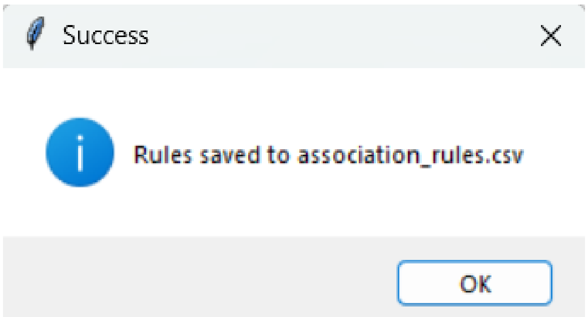
OUTPUT :



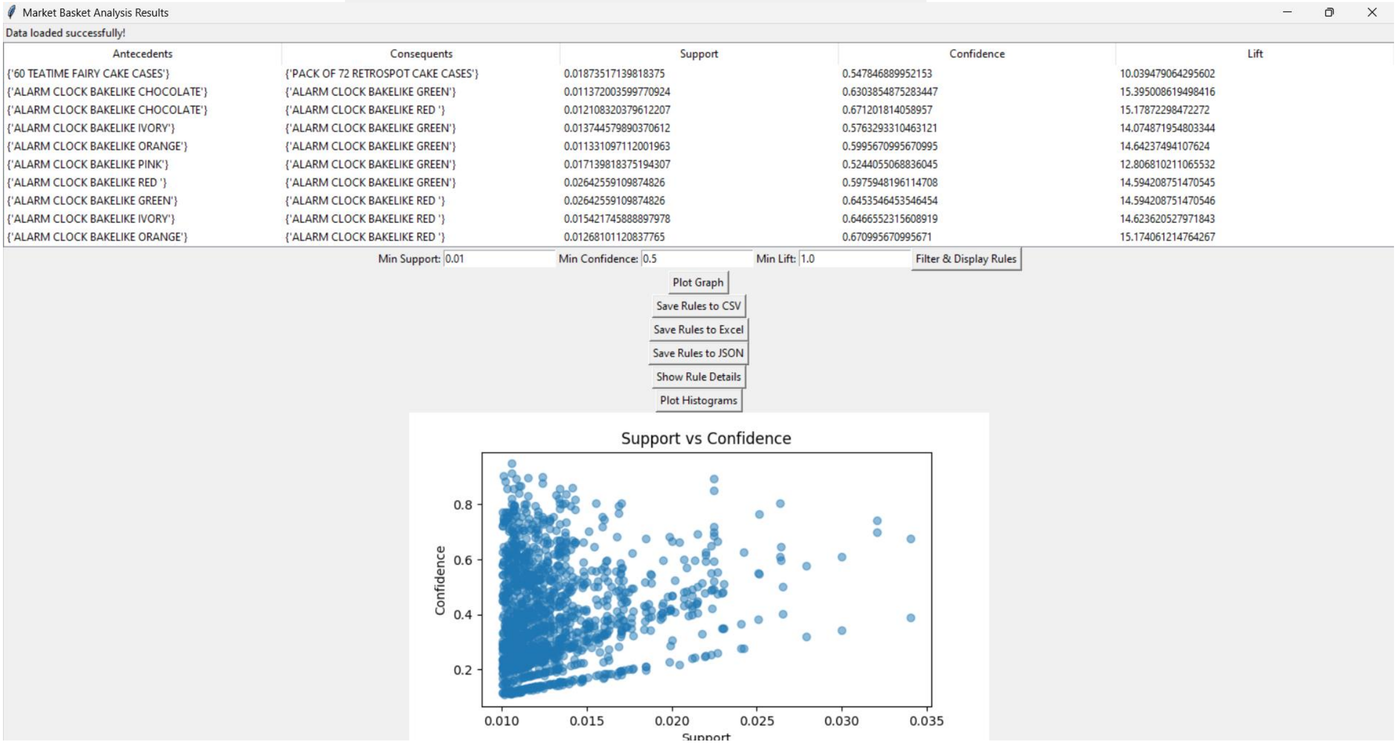
Performing market basket analysis...

Antecedents	Consequents	Support	Confidence	Lift

Min Support: 0.01 Min Confidence: 0.5 Min Lift: 1.0 Filter & Display Rules



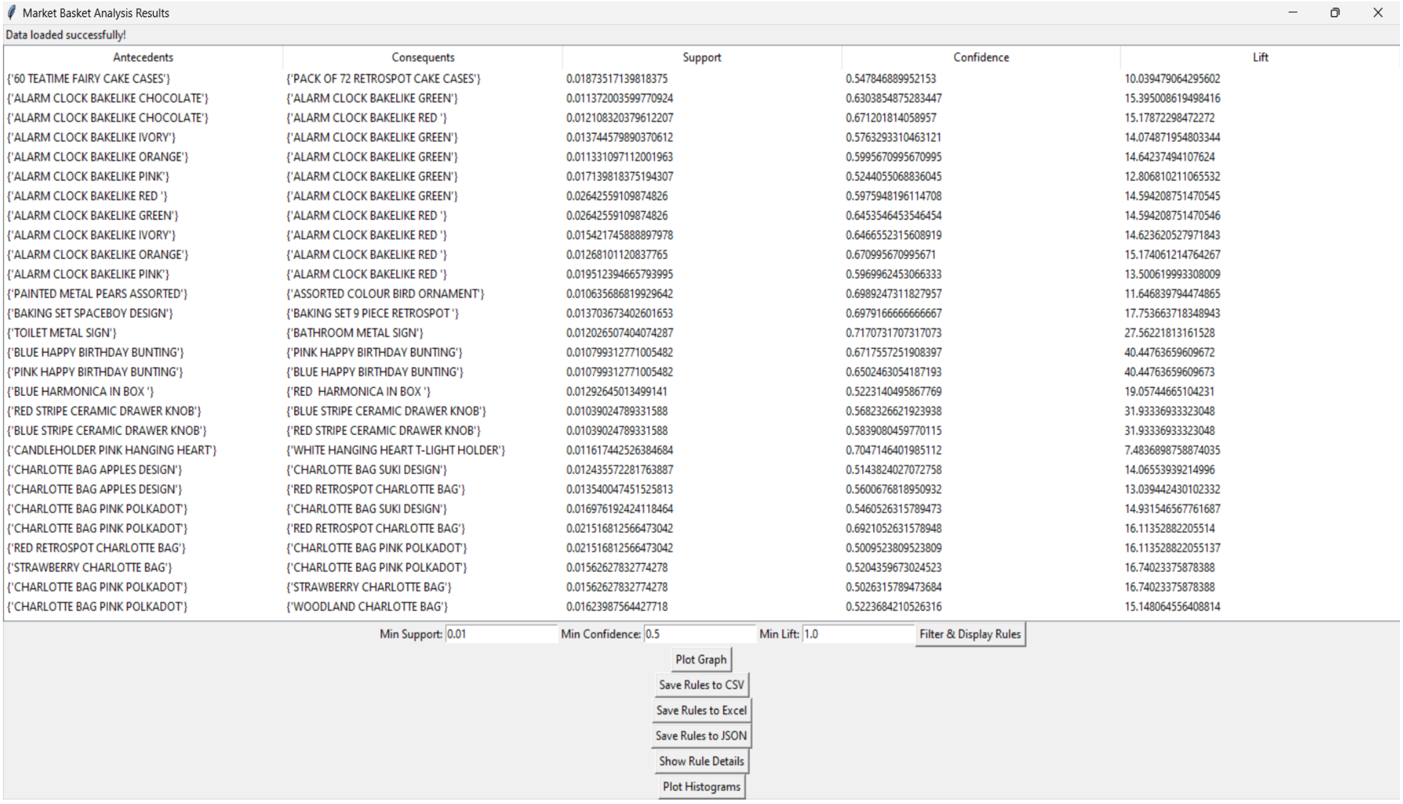
Rules saved to association_rules.csv



Data loaded successfully!

Antecedents	Consequents	Support	Confidence	Lift
{'60 TEATIME FAIRY CAKE CASES'}	{'PACK OF 72 RETROSPOT CAKE CASES'}	0.01873517139818375	0.547846889952153	10.039479064295602
{'ALARM CLOCK BAKELIKE CHOCOLATE'}	{'ALARM CLOCK BAKELIKE GREEN'}	0.011372003599770924	0.6303854875283447	15.395008619498416
{'ALARM CLOCK BAKELIKE CHOCOLATE'}	{'ALARM CLOCK BAKELIKE RED '}	0.012108320379612207	0.671201814058957	15.17872298472272
{'ALARM CLOCK BAKELIKE IVORY'}	{'ALARM CLOCK BAKELIKE GREEN'}	0.013744579890370612	0.5763293310463121	14.074871954803344
{'ALARM CLOCK BAKELIKE ORANGE'}	{'ALARM CLOCK BAKELIKE GREEN'}	0.011331097112001963	0.5995670995670995	14.64237494107624
{'ALARM CLOCK BAKELIKE PINK'}	{'ALARM CLOCK BAKELIKE GREEN'}	0.017139818375194307	0.5244055068836045	12.806810211065532
{'ALARM CLOCK BAKELIKE RED '}	{'ALARM CLOCK BAKELIKE GREEN'}	0.02642559109874826	0.5975948196114708	14.594208751470545
{'ALARM CLOCK BAKELIKE GREEN'}	{'ALARM CLOCK BAKELIKE RED '}	0.02642559109874826	0.6453546453546454	14.594208751470546
{'ALARM CLOCK BAKELIKE IVORY'}	{'ALARM CLOCK BAKELIKE RED '}	0.015421745888897978	0.6466552315608919	14.623620527971843
{'ALARM CLOCK BAKELIKE ORANGE'}	{'ALARM CLOCK BAKELIKE RED '}	0.012681011120837765	0.670995670995671	15.174061214764267

Min Support: 0.01 Min Confidence: 0.5 Min Lift: 1.0 Filter & Display Rules



Market Basket Analysis Results

Data loaded successfully!

Antecedents	Consequents	Support	Confidence	Lift
{'60 TEATIME FAIRY CAKE CASES'}	{'PACK OF 72 RETROSPOT CAKE CASES'}	0.01873517139818375	0.547846889952153	10.039479064295602
{'ALARM CLOCK BAKELIKE CHOCOLATE'}	{'ALARM CLOCK BAKELIKE GREEN'}	0.011372003599770924	0.6303854875283447	15.395008619498416
{'ALARM CLOCK BAKELIKE CHOCOLATE'}	{'ALARM CLOCK BAKELIKE RED '}	0.012108320379612207	0.671201814058957	15.17872298472272
{'ALARM CLOCK BAKELIKE IVORY'}	{'ALARM CLOCK BAKELIKE GREEN'}	0.013744579890370612	0.5763293310463121	14.074871954803344
{'ALARM CLOCK BAKELIKE ORANGE'}	{'ALARM CLOCK BAKELIKE GREEN'}	0.011331097112001963	0.5995670995670995	14.64237494107624
{'ALARM CLOCK BAKELIKE PINK'}	{'ALARM CLOCK BAKELIKE GREEN'}	0.017139818375194307	0.5244055068836045	12.806810211065532
{'ALARM CLOCK BAKELIKE RED '}	{'ALARM CLOCK BAKELIKE GREEN'}	0.02642559109874826	0.5975948196114708	14.594208751470545
{'ALARM CLOCK BAKELIKE GREEN'}	{'ALARM CLOCK BAKELIKE RED '}	0.02642559109874826	0.6453546453546454	14.594208751470546
{'ALARM CLOCK BAKELIKE IVORY'}	{'ALARM CLOCK BAKELIKE RED '}	0.015421745888897978	0.6466552315608919	14.623620527971843
{'ALARM CLOCK BAKELIKE ORANGE'}	{'ALARM CLOCK BAKELIKE RED '}	0.01268101120837765	0.670995670995671	15.174061214764267
{'ALARM CLOCK BAKELIKE PINK'}	{'ALARM CLOCK BAKELIKE RED '}	0.019512394665793995	0.5969962453066333	13.500619993308009
{'PAINTED METAL PEARS ASSORTED'}	{'ASSORTED COLOUR BIRD ORNAMENT'}	0.010635686819929642	0.6989247311827957	11.646839794474865
{'BAKING SET SPACEBOY DESIGN'}	{'BAKING SET 9 PIECE RETROSPOT '}	0.013703673402601653	0.6979166666666667	17.753663718348943
{'TOILET METAL SIGN'}	{'BATHROOM METAL SIGN'}	0.012026507404074287	0.7170731707317073	27.56221813161528
{'BLUE HAPPY BIRTHDAY BUNTING'}	{'PINK HAPPY BIRTHDAY BUNTING'}	0.010799312771005482	0.6717557251908397	40.44763659609672
{'PINK HAPPY BIRTHDAY BUNTING'}	{'BLUE HAPPY BIRTHDAY BUNTING'}	0.010799312771005482	0.6502463054187193	40.44763659609673
{'BLUE HARMONICA IN BOX '}	{'RED HARMONICA IN BOX '}	0.01292645013499141	0.5223140495867769	19.05744665104231
{'RED STRIPE CERAMIC DRAWER KNOB'}	{'BLUE STRIPE CERAMIC DRAWER KNOB'}	0.01039024789331588	0.5682326621923938	31.93336933323048
{'BLUE STRIPE CERAMIC DRAWER KNOB'}	{'RED STRIPE CERAMIC DRAWER KNOB'}	0.01039024789331588	0.5839080459770115	31.93336933323048
{'CANDLEHOLDER PINK HANGING HEART'}	{'WHITE HANGING HEART T-LIGHT HOLDER'}	0.011617442526384684	0.7047146401985112	7.4836898758874035
{'CHARLOTTE BAG APPLES DESIGN'}	{'CHARLOTTE BAG SUKI DESIGN'}	0.012435572281763887	0.5143824027072758	14.06553939214996
{'CHARLOTTE BAG APPLES DESIGN'}	{'RED RETROSPOT CHARLOTTE BAG'}	0.013540047451525813	0.5600676818950932	13.039442430102332
{'CHARLOTTE BAG PINK POLKADOT'}	{'CHARLOTTE BAG SUKI DESIGN'}	0.016976192424118464	0.5460526315789473	14.931546567761687
{'CHARLOTTE BAG PINK POLKADOT'}	{'RED RETROSPOT CHARLOTTE BAG'}	0.021516812566473042	0.6921052631578948	16.11352882205514
{'RED RETROSPOT CHARLOTTE BAG'}	{'CHARLOTTE BAG PINK POLKADOT'}	0.021516812566473042	0.5009523809523809	16.113528822055137
{'STRAWBERRY CHARLOTTE BAG'}	{'CHARLOTTE BAG PINK POLKADOT'}	0.01562627832774278	0.5204359673024523	16.74023375878388
{'CHARLOTTE BAG PINK POLKADOT'}	{'STRAWBERRY CHARLOTTE BAG'}	0.01562627832774278	0.5026315789473684	16.74023375878388
{'CHARLOTTE BAG PINK POLKADOT'}	{'WOODLAND CHARLOTTE BAG'}	0.01623987564427718	0.5223684210526316	15.148064556408814

Min Support: 0.01 Min Confidence: 0.5 Min Lift: 1.0

CONCLUSION:

In conclusion, the market basket analysis tool provides a comprehensive solution for understanding and optimizing retail strategies by leveraging the Apriori algorithm. This tool excels in preprocessing transaction data, identifying frequent item sets, and generating meaningful association rules, all of which contribute to uncovering valuable patterns in customer purchasing behavior. Its robust analytical capabilities are complemented by user-friendly visualizations, including scatter plots and histograms, which make complex data insights accessible and actionable.

The tool's extensive functionalities, including data filtering, graph plotting, and result exporting, provide users with the flexibility to customize their analysis to meet specific business needs. Data filtering allows users to refine their analysis by setting thresholds for support, confidence, and lift, ensuring that only the most relevant association rules are considered. This targeted approach helps in identifying actionable insights that are directly applicable to the business context. Graph plotting, on the other hand, transforms complex data patterns into intuitive visual representations such as scatter plots and histograms. These visual tools make it easier to interpret and communicate the findings, facilitating a clearer understanding of customer behavior and item relationships.

Result exporting enhances the tool's utility by allowing users to save and share their findings in various formats, including CSV, Excel, and JSON. This capability ensures that the analysis results can be easily integrated into existing business systems and reporting frameworks. By presenting insights in a clear and actionable manner, the tool empowers businesses to make data-driven decisions regarding product placement, inventory management, and targeted marketing strategies

REFERENCES :

[1] Raeder, Troy, and Nitesh V. Chawla. "Market basket analysis with networks." *Social network analysis and mining* 1 (2011): 97-113. http://dx.doi.org/10.1007/s13278-010-0003-7

[2] Kaur, Manpreet, and Shivani Kang. "Market Basket Analysis: Identify the changing trends of market data using association rule mining." Procedia computer science 85 (2016): 78-85.https://doi.org/10.1016/j.procs.2016.05.180

[3] Chen, Yen-Liang, Kwei Tang, Ren-Jie Shen, and Ya-Han Hu. "Market basket analysis in a multiple store environment." Decision support systems 40, no. 2 (2005): 339-354.http://dx.doi.org/10.1016/j.dss.2004.04.009

[4] Boztuğ, Yasemin, and Thomas Reutterer. "A combined approach for segment-specific market basket analysis." European Journal of Operational Research 187, no. 1 (2008): 294-312.https://doi.org/10.1016/j.ejor.2007.03.001

[5] Sjarif, Nilam Nur Amir, Nurulhuda Firdaus Mohd Azmi, Siti Sophiayati Yuhaniz, and Doris Hooi-Ten Wong. "A review of market basket analysis on business intelligence and data mining." International journal of business intelligence and data mining 18, no. 3 (2021): 383-394.https://doi.org/10.1504/IJBIDM.2021.114475

[6] Rao, Abishek B., and Jammula Surya Kiran. "Application of market–basket analysis on healthcare." International Journal of System Assurance Engineering and Management 14, no. Suppl 4 (2023): 924-929.https://doi.org/10.1007/s12652-019-01217-1

[7] Dhanabhakyam, M., and M. Punithavalli. "A survey on data mining algorithm for market basket analysis." Global Journal of Computer Science and Technology 11, no. 11 (2011): 23-28.. https://doi.org/10.1016/j.ipm.2023.103577

[8] Müller, Henning, Thierry Pun, and David Squire. "Learning from user behavior in image retrieval: Application of market basket analysis." *International Journal of Computer Vision* 56 (2004): 65-77. https://doi.org/10.1023/B:VISI.0000004832.02269.45

[9] Musalem, Andres, Luis Aburto, and Maximo Bosch. "Market basket analysis insights to support category management." European Journal of Marketing 52, no. 7/8 (2018): 1550-1573.DOI:10.1108/EJM-06-2017-0367

[10] Woo, Jongwook, and Yuhang Xu. "Market basket analysis algorithm with map/reduce of cloud computing." In Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA), p. 1. The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2011.

http://dx.doi.org/10.18687/LACCEI2016.1.1.307