**Task 3: Market Basket Analysis**

Joseph Emmanuel Cayetano

Western Governors University

D599: Data Preparation and Exploration Task 3

October 20, 2024

**A1. Propose one question you will answer using market basket analysis**

One question that is relevant to a real-world organizational situation that I will answer using market basket analysis is "Which products are frequently purchased together?" By identifying the products that are usually purchased together, it can help the organization create targeted discounts and recommend bundled offers to customers, resulting in higher sales.

**A2. Define one goal of the data analysis**

One goal of the market basket analysis is to enhance complementary sales for the organization. By identifying which products are often purchased together, the organization can develop targeted marketing campaigns to promote those product combinations, which leads to increased customer engagement and sales.

**B1. Explain how the market basket technique analyzes the provided dataset, including expected outcomes**

By definition, market basket analysis is "a data mining technique retailers use to increase sales by better understanding customer purchasing patterns" (Amruta, 2024, par. 1). The market basket technique analyzes the dataset by converting the original dataset into a transactions data. It groups the dataset by using the column OrderID, so that every unique OrderID represents a separate transaction in the dataset. Each transaction has a group of products that were purchased together. The market basket analysis focuses on identifying frequent itemsets. If there is a group of products that is purchased together frequently, it will be considered as a frequent itemset. Once frequent itemsets are identified, association rules are created. Association rules show the relationship between different items purchased together. There are a couple of expected outcomes when performing the market basket technique in the dataset. First, by discovering the products that are often purchased together, the organization can sell those products as a bundle at a discounted price, resulting in increased sales by encouraging customers to buy more items. Second, the organization can improve its product recommendations to customers.

**B2. Provide one example of a transaction in the dataset**

A screenshot of a computer

Description automatically generated

One example of a transaction in the dataset is the OrderID 581316. In this order, the customer purchased three products in a single transaction. The products purchased are: RED RETROSPOT SUGAR JAM BOWL, GLASS SONGBIRD STORAGE JAR, and REGENCY SUGAR BOWL GREEN. Looking at the dataset, while there are three separate rows for three different products, those three products still came from a single transaction because they have the same OrderID value and the same InvoiceDate value. This is an example of a single transaction that would be in the transaction data to be used for market basket analysis.

**B3. Summarize one assumption of market basket analysis**

One assumption of market basket analysis is that the items frequently purchased together have an underlying relationship. These associations are not random, and they reflect customer preferences and purchasing behavior, which can be used by organizations to create personalized recommendations or special promotions to customers. This results in higher sales for organizations and improves customer satisfaction.

**C1a. Select x number of categorical variables, choosing at least two ordinal variables and at least two nominal variables**

The two ordinal variables in the dataset that I chose are OrderPriority and CustomerOrderSatisfaction. The two nominal variables in the dataset that I chose are PaymentMethod and Region.

A close-up of a computer code

Description automatically generated

The OrderPriority variable has two unique values: 'High' and 'Medium'. The OrderPriority variable is an ordinal variable because its values have a ranking, with 'High' being greater or more important than 'Medium'. The OrderPriority variable has ranked values but the differences between those values are not measurable in numerical terms.

A close up of words

Description automatically generated

The CustomerOrderSatisfaction variable has five unique values: 'Very Satisfied', 'Satisfied', 'Prefer not to answer', 'Dissatisfied', and 'Very Dissatisfied'. The CustomerOrderSatisfaction variable is an ordinal variable because its values have a clear and meaningful order where 'Very Satisfied' is better than 'Satisfied, and 'Satisfied' is better than 'Dissatisfied'. Also, since the differences between those values are not numerically measurable, it makes it ordinal.

A close up of a box

Description automatically generated

The PaymentMethod variable has two unique values: 'Credit Card' and 'PayPal'. The PaymentMethod variable is a nominal variable because its values have no meaningful order, and they are simply different labels. There is no value in the variable that is lower or higher than the other. They are simply different methods of payment.

A close up of a text

Description automatically generated

The Region variable has two unique values: 'Northeast' and Southeast'. The Region variable is a nominal variable because its values have no ranking, and they are simply labels for different areas. No region is greater than or less than the other. The values are just labels without any quantitative relationships between them.

**C1b. Perform the appropriate encoding method (ordinal, label encoding, one-hot encoding) for each variable selected in part C1a**

A close up of a text

Description automatically generated

I performed ordinal encoding on the OrderPriority variable because ordinal encoding is used for variables that have a hierarchy, which OrderPriority has in its values where 'High' is greater or more important than 'Medium'. Ordinal encoding converts the string values into integer values that reflect their natural order. In this code, I used the replace() function to encode 'High' as 2 and 'Medium' as 1 in the OrderPriority column. 'High' has a higher integer value since it has a higher priority than 'Medium'.



I performed ordinal encoding on the CustomerOrderSatisfaction variable because ordinal encoding is used for variables that have a meaningful order, which CustomerOrderSatisfaction has in its values where 'Very Satisfied' is better than 'Satisfied', and 'Satisfied' is better than 'Dissatisfied'. Ordinal encoding converts the values into integer values that reflect the order of the satisfaction levels. In this code, I used the replace() function to encode 'Very Satisfied' as 5, 'Satisfied' as 4, 'Prefer not to answer' as 3, 'Dissatisfied' as 2, and 'Very Dissatisfied' as 1 in the CustomerOrderSatisfaction column.



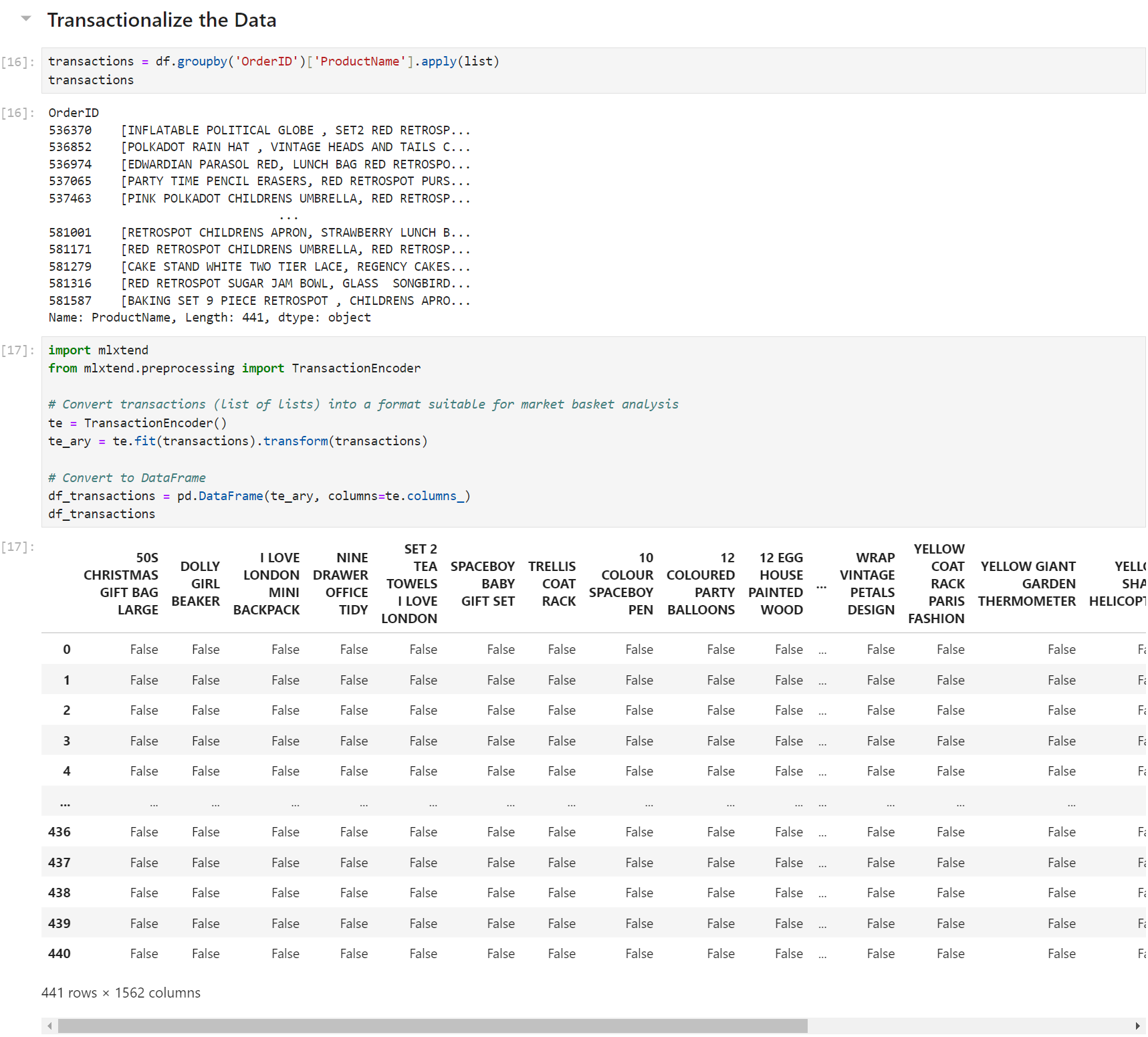
I performed label encoding on the PaymentMethod variable because label encoding is used for variables that have no hierarchy and have very few values, which describes PaymentMethod. Label encoding converts the string values into integer values, and it will treat the values simply as different categories, not as ranked or ordered data. In this code, I used the replace() function to encode 'Credit Card' as 0 and 'PayPal' as 1 in the PaymentMethod column.

A screenshot of a computer code

Description automatically generated

I performed one-hot encoding for the Region variable because one-hot encoding is appropriate for variables with values that have no meaningful order between them, which describes Region. One-hot encoding treats each value in the Region variable independently, which makes sense because there is no region that is less than or greater than the other. I used the get\_dummies() function, which creates two new columns: Region\_Northeast and Region\_Southeast. Rows that had the value 'Northeast' in the Region column will have a value of 1 in the Region\_Northeast column, 0 otherwise. Rows that had the value 'Southeast' in the Region column will have a value of 1 in the Region\_Southeast column, 0 otherwise. I converted the two new columns into an int data type to have numeric values instead of Boolean values.

**C1c. Transactionalize the data for market basket analysis**



Here, I converted the original dataset into a transaction data where market basket analysis can be performed. I started the process by grouping the products within each order using the groupby() function and storing them into a variable called transactions. Then, I imported the TransactionEncoder() function from mlxtend because it is the function that converts the data into a format that is appropriate for market basket analysis. Using the transform() function, it converted all the data in the transactions variable into a binary matrix, where columns represent products and rows represent transactions. The binary matrix is now stored in the variable te\_ary. The variable te\_ary is then converted into a pandas dataframe where each column represents a product, and it can either have a value of True or False, depending on whether the product was included in the transaction. Looking at the final dataframe, it has 441 transactions and 1,562 unique products.

**C1d. Explain and justify each step you took in parts C1a, C1b, and C1c**

For C1a, the two ordinal variables that I chose are OrderPriority and CustomerOrderSatisfaction, and the two nominal variables that I chose are PaymentMethod and Region. The variable OrderPriority is an ordinal variable because its values 'High' and 'Medium' can be ranked where 'High' is greater or more important than 'Medium'. The variable CustomerOrderSatisfaction is also an ordinal variable because its values are ranked and have a meaningful order where 'Very Satisfied' is better than 'Satisfied', and 'Satisfied' is better than 'Dissatisfied'. The variable PaymentMethod is a nominal variable because its values 'Credit Card' and 'Paypal' have no meaningful order, and they are simply different modes of payment. There is no payment that is better or worse than the other. The variable Region is also a nominal variable because its values 'Northeast' and 'Southeast' are not ranked and they are simply labels for different areas.

For C1b, I performed the appropriate encoding method for each variable. For the variable OrderPriority, I performed ordinal encoding because ordinal encoding is used for variables that have a hierarchy, which OrderPriority has in its values where 'High' is greater or more important than 'Medium'. Ordinal encoding converts the string values into integer values that reflect their natural order. I used the replace() function to encode 'High' as 2 and 'Medium' as 1 in the OrderPriority column. For the variable CustomerOrderSatisfaction, I performed ordinal encoding because ordinal encoding is used for variables that have a meaningful order, which CustomerOrderSatisfaction has in its values where 'Very Satisfied' is better than 'Satisfied', and 'Satisfied' is better than 'Dissatisfied'. Ordinal encoding converts the values into integer values that reflect the order of the satisfaction levels. I used the replace() function to encode 'Very Satisfied' as 5, 'Satisfied' as 4, 'Prefer not to answer' as 3, 'Dissatisfied' as 2, and 'Very Dissatisfied' as 1 in the CustomerOrderSatisfaction column. For the variable PaymentMethod, I performed label encoding because label encoding is used for variables that have no hierarchy and have very few values, which describes PaymentMethod. Label encoding converts the string values into integer values, and it will treat the values simply as different categories, not as ranked or ordered data. I used the replace() function to encode 'Credit Card' as 0 and 'PayPal' as 1 in the PaymentMethod column. For the variable Region, I used one-hot encoding because one-hot encoding is appropriate for variables with values that have no meaningful order between them, which describes Region. One-hot encoding treats each value in the Region variable independently, which makes sense because there is no region that is less than or greater than the other. I used the get\_dummies() function, which creates two new columns: Region\_Northeast and Region\_Southeast. Rows that had the value 'Northeast' in the Region column will have a value of 1 in the Region\_Northeast column, 0 otherwise. Rows that had the value 'Southeast' in the Region column will have a value of 1 in the Region\_Southeast column, 0 otherwise.

For C1c, I converted the original dataset into a transaction data where market basket analysis can be performed. I started the process by grouping the products within each order using the groupby() function and storing them into a variable called transactions. Then, I imported the TransactionEncoder() function from mlxtend because it is the function that converts the data into a format that is appropriate for market basket analysis. Using the transform() function, it converted all the data in the transactions variable into a binary matrix, where columns represent products and rows represent transactions. The binary matrix is now stored in the variable te\_ary. The variable te\_ary is then converted into a pandas dataframe where each column represents a product, and it can either have a value of True or False, depending on whether the product was included in the transaction. Looking at the final dataframe, it has 441 transactions and 1,562 unique products. The dataframe is now prepared and ready to be used by the Apriori algorithm to create association rules that show frequently purchased products.

**C2. Include a copy of the cleaned dataset**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

This is the dataset that has been cleaned and undergone various encoding processes. Both OrderPriority and CustomerOrderSatisfaction variables were ordinally encoded, the PaymentMethod was label encoded, and the Region column was one-hot encoded. This cleaned data is now prepared for further data analysis. This dataset is named “cleaned\_dataset.csv” and I have included it in my submission.

**A screenshot of a computer

Description automatically generated**

This is the dataset that has been converted into a binary matrix format to be used for market basket analysis. Each row represents a single transaction, and each column represents a unique product. The values in the table are either True or False. True means the product was purchased within the transaction, and False otherwise. This data is now ready to be used by the Apriori algorithm to create association rules that show frequently purchased products. This dataset is named “cleaned\_transactions\_dataset.csv” and I have included it in my submission.

**C3. Execute the code used to generate association rules with the Apriori algorithm and provide a screenshot**

A screenshot of a computer

Description automatically generated

In this code, I used the Apriori Algorithm to generate frequent itemsets and association rules from the binary matrix data. I set the minimum support threshold to 1%, which means only itemsets with at least 1% support will be considered frequent. The frequent itemsets are stored in the variable frequent\_itemsets. Then, I used the association\_rules() function to generate association rules from the frequent itemsets. Lift is the metric that was used to generate the rules and this metric measures how likely multiple items are to be purchased together compared to being purchased independently. Rules that have a lift value of at least 1 are the only rules to be included in the association rules table because we are only interested in rules that have a positive association. The association rules will include support, lift, and confidence values.

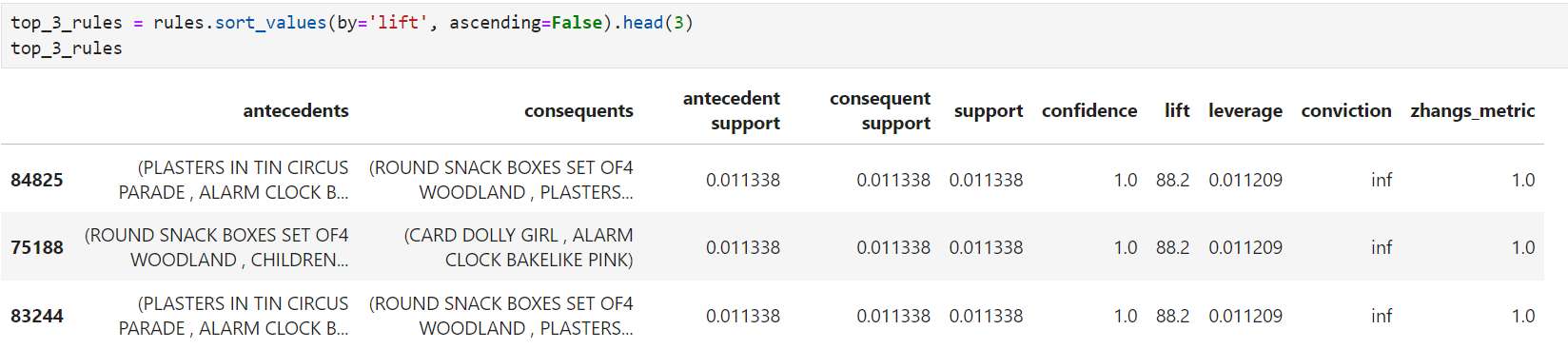
**C4. Provide values for the support, lift, and confidence of the association rules table. Include a screenshot of the values**

A screenshot of a computer

Description automatically generated

Here is the association rules table. This table describes relationships between different items that are often purchased together. Each association rule has support, confidence, and lift values. Support measures how often an item or an itemset appears in all of the transactions. Confidence measures how likely the consequent (right) will get purchased if the antecedent (left) is purchased. A higher confidence value means stronger association between the consequent and antecedent. Lift measures how likely multiple items are to be purchased together compared to being purchased independently. A rule that has a lift value of at least 1 indicates that the products are more likely to be purchased together than if they were purchased randomly.

**C5. Explain the top three relevant rules generated by the Apriori algorithm. Include a screenshot of the top three relevant rules**



Here are the top three rules that have the highest lift values. The first rule has four products as antecedents and they are: 'ALARM CLOCK BAKELIKE PINK', 'ALARM CLOCK BAKELIKE RED' 'PLASTERS IN TIN CIRCUS PARADE' and 'SET6 RED SPOTTY PAPER PLATES'. As for the consequents, the rule has four of them and they are: 'ALARM CLOCK BAKELIKE GREEN', 'PLASTERS IN TIN SPACEBOY', 'ROUND SNACK BOXES SET OF4 WOODLAND' and 'SET6 RED SPOTTY PAPER CUPS'. The support value is 1.1%, which means the combination of antecedents and consequents in this rule appear together in 1.1% of the total transactions. The confidence value is 1.0, which means every time the antecedents are purchased, the consequents also get purchased 100% of the time. The lift value is 88.2, which means both antecedents and consequents are 88.2 times more likely to be purchased together than if their purchases were independent.

The second rule has four products as antecedents and they are: 'ALARM CLOCK BAKELIKE RED', 'CHILDRENS CUTLERY SPACEBOY', 'ROUND SNACK BOXES SET OF4 WOODLAND' and 'SPACEBOY BIRTHDAY CARD'. As for the consequents, the rule has two of them and they are: 'ALARM CLOCK BAKELIKE PINK' and 'CARD DOLLY GIRL'. The support value is 1.1%, which means the combination of antecedents and consequents in this rule appear together in 1.1% of the total transactions. The confidence value is 1.0, which means every time the antecedents are purchased, the consequents also get purchased 100% of the time. The lift value is 88.2, which means both antecedents and consequents are 88.2 times more likely to be purchased together than if their purchases were independent.

The third rule has three products as antecedents and they are: 'ALARM CLOCK BAKELIKE RED', 'PLASTERS IN TIN CIRCUS PARADE', and 'SET6 RED SPOTTY PAPER PLATES'. As for the consequents, the rule has four of them and they are: 'ALARM CLOCK BAKELIKE PINK', 'PASTERS IN TIN SPACEBOY', 'ROUND SNACK BOXES SET OF4 WOODLAND', and 'SET6 RED SPOTTY PAPER CUPS'. The support value is 1.1%, which means the combination of antecedents and consequents in this rule appear together in 1.1% of the total transactions. The confidence value is 1.0, which means every time the antecedents are purchased, the consequents also get purchased 100% of the time. The lift value is 88.2, which means both antecedents and consequents are 88.2 times more likely to be purchased together than if their purchases were independent.

Overall, these three rules represent a strong and reliable association between their antecedents and consequents, despite each rule only appearing in 1.1% of all transaction in the dataset. These rules can be used to create targeted discounts and recommend bundled offers to customers.

**D1. Discuss the significance of support, lift, and confidence from the results of the analysis**

Support, confidence, and lift values are important metrics in market basket analysis because these metrics help identify which combinations of products have strong relationships. Support measures how frequently an itemset appears in the dataset. A higher support value means an itemset appears frequently in transactions. Confidence measures the probability that a transaction containing the antecedent also contains the consequent. A higher confidence value means that when the antecedent is purchased, the consequent is also often purchased. Lift measures the strength of association between the antecedent and consequent compared to random co-occurrence. A higher lift value means the antecedent and consequent are purchased together more frequently than expected by chance.

Looking at the general association rules table, each rule has support, confidence, and lift values. These three metrics are essential in assessing the strength and relevance of the association rules. In the top 3 association rules table we created from earlier, we can see that all three rules are highly reliable because they have high confidence and lift values, which indicate that there is a high probability that the items get purchased together than independently.

**D2. Explain the practical significance of your findings from the analysis**

Conducting market basket analysis on the dataset holds practical significance. The association rules can be used to enhance product recommendations by recommending complementary products to customers while their shopping, which increases the likelihood of them buying more items. Using one of our association rules for example, knowing that 'DOLLY GIRL BEAKER' and 'DOLLY GIRL CHILDRENS BOWL' frequently get purchased together, if the customer buys 'DOLLY GIRL BEAKER' they should also get recommended to buy 'DOLLY GIRL CHILDRENS BOWL' because those two items frequently get purchased together. This benefits businesses because it could significantly increase their sales, while also increasing customer engagement by offering relevant suggestions.

Another way businesses can take advantage of the results from market basket analysis is to create targeted discounts and bundling. They can offer discounts on products that are frequently purchased together or bundle the products to entice customer to purchase multiple items, which increases overall sales.

**D3. Recommend a course of action for the real-world organizational situation from part A1 that is based on the results from part D1**

To answer the question from part A1 about "Which products are frequently purchased together?", we have identified the products that were frequently purchased together based on the market basket analysis results. There are two recommended courses of actions for frequently purchased products. The first recommended action is to display frequently purchased together products during the online checkout process. For example, if a customer selects the item 'PLASTER IN TIN SPACEBOY', they should get recommended to buy 'ALARM CLOCK BAKELIKE PINK' as a complementary purchase because both items are often purchased together based on the market basket analysis results. This increases the likelihood of customers buying more items, and thus increases sales. The second recommended action is to bundle commonly purchased together products with a slight discount to encourage customers to buy multiple items together, increasing the overall order size. For example, bundling the frequently purchased together items like 'ALARM CLOCK BAKELIKE RED', 'CHILDRENS CUTLERY SPACEBOY' and 'SPACEBOY BIRTHDAY CARD' together with a slight discount, it encourages customers to purchase them all together. This drives more sales to the business.

**E. Panopto Presentation**

Here is the link to my Panopto Presentation: https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=c2e66138-22a0-4c85-ab51-b210012ffd57

**References**

Amruta. (2024, October 14). *Market basket analysis: A comprehensive guide for businesses*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/