D599: Data Exploration and Preparation Task 3
By Eric Williams

## Part I: Research Question

### A1:QUESTION FOR ANALYSIS

The underlying question for this market basket analysis is: which combination of products are most often bought together?

### A2:DATA ANALYSIS GOAL

The goal of this analysis is to identify which combination of products are often bought together in order to make practical changes to increase sales. If we can identify which products are often bought together, we can provide recommendations to customers for things they might already be looking for.

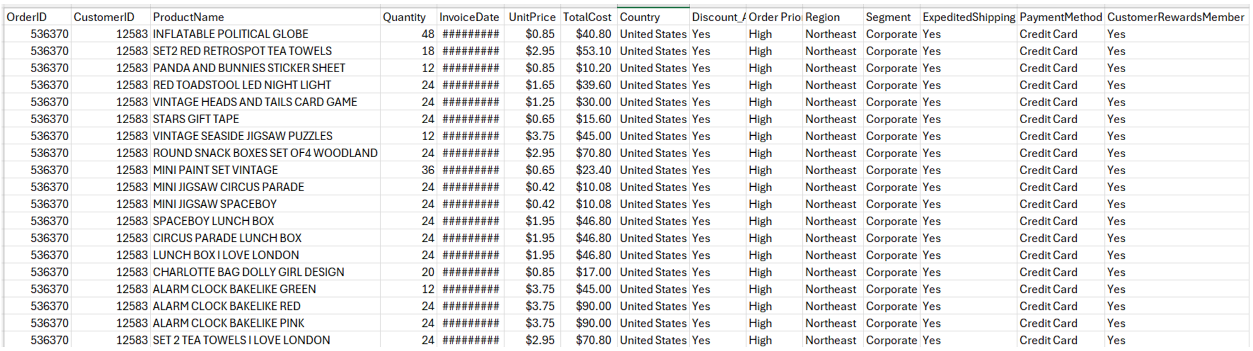
## Part II: Market Basket Justification

### B1:MARKET BASKET ANALYSIS

According to the WGU course materials, “Market Basket Analysis is a data mining technique used to identify relationships between products frequently purchased together by customers.” Essentially, by analyzing the data on sales, this analysis can discover patterns in customer behavior. With the findings of a market basket analysis, businesses can optimize where to place products. Essentials like eggs, milk, and break are rarely placed next to each other in a supermarket to ensure the customers spend longer in the store and see more products they may be interested in buying.

### B2:TRANSACTION EXAMPLE

For this example, I will use the following transaction in the dataset:



| OrderID | CustomerID | ProductName | Quantity | InvoiceDate | UnitPrice | TotalCost | Country | Discount_A | Order Prio | Region | Segment | ExpeditedShipping | PaymentMethod | CustomerRewardsMember |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 536370 | 12583 | INFLATABLE POLITICAL GLOBE | 48 | ######### | $0.85 | $40.80 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | SET2 RED RETROSPOT TEA TOWELS | 18 | ######### | $2.95 | $53.10 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | PANDA AND BUNNIES STICKER SHEET | 12 | ######### | $0.85 | $10.20 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | RED TOADSTOOL LED NIGHT LIGHT | 24 | ######### | $1.65 | $39.60 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | VINTAGE HEADS AND TAILS CARD GAME | 24 | ######### | $1.25 | $30.00 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | STARS GIFT TAPE | 24 | ######### | $0.65 | $15.60 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | VINTAGE SEASIDE JIGSAW PUZZLES | 12 | ######### | $3.75 | $45.00 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | ROUND SNACK BOXES SET OF4 WOODLAND | 24 | ######### | $2.95 | $70.80 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | MINI PAINT SET VINTAGE | 36 | ######### | $0.65 | $23.40 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | MINI JIGSAW CIRCUS PARADE | 24 | ######### | $0.42 | $10.08 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | MINI JIGSAW SPACEBOY | 24 | ######### | $0.42 | $10.08 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | SPACEBOY LUNCH BOX | 24 | ######### | $1.95 | $46.80 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | CIRCUS PARADE LUNCH BOX | 24 | ######### | $1.95 | $46.80 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | LUNCH BOX I LOVE LONDON | 24 | ######### | $1.95 | $46.80 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | CHARLOTTE BAG DOLLY GIRL DESIGN | 20 | ######### | $0.85 | $17.00 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | ALARM CLOCK BAKELIKE GREEN | 12 | ######### | $3.75 | $45.00 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | ALARM CLOCK BAKELIKE RED | 24 | ######### | $3.75 | $90.00 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | ALARM CLOCK BAKELIKE PINK | 24 | ######### | $3.75 | $90.00 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |
| 536370 | 12583 | SET 2 TEA TOWELS I LOVE LONDON | 24 | ######### | $2.95 | $70.80 | United States | Yes | High | Northeast | Corporate | Yes | Credit Card | Yes |

This table gives us a snapshot of one single purchase with a unique Order ID. Every order in the dataset tracks which customers made the purchase, how many products they bought, what the cost and price was, when and where the purchase was made, whether the products were discounted or high priority, if the customer was a rewards member, and what the payment method was.

## B3:ASSUMPTION SUMMARY

The most important assumption in market basket analysis is that the findings will be useful--meaning that by analyzing purchase patterns in the past can help us predict purchase patterns in the future. This is not necessarily always true, as people, society, and trends change. However, it is assumed that finding a connection between products in the past can give us insights into how customers will behave in the future.

## Part III: Data Preparation and Analysis

## C1a:CATEGORICAL VARIABLES

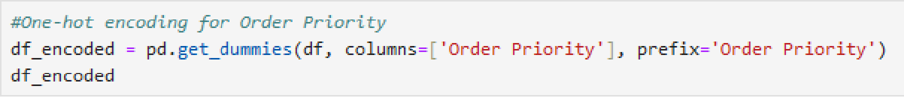
An ordinal variable is a variable that is categorical that can be ordered. The first ordinal variable I will use is order priority, because this can be ordered as “high” or “medium,” meaning one data point can be ordered above the other. The other ordinal variable I will use is Expedited Shipping. Although the column has a simple “yes” or “no” entry for the datapoints, another way to interpret this is “fast” shipping and “slow” shipping. As a matter of convenience one can clearly be ordered over the other.

The nominal variables I will use are segment (which can either be corporate or consumer) and Payment Method (which can be either PayPal or credit card).

Note for **C1d:EXPLANATION AND JUSTIFICATION OF STEPS**: Below I will outline the code for encoding and transactionalizing the data. I will include justification of each step as it is performed.

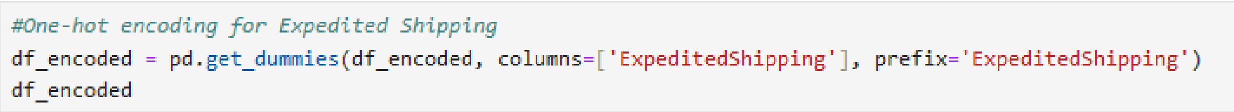
## C1b:ENCODING

The data needs to be encoded before it is ready for the Apriori algorithm. It cannot interpret raw data, so we need to convert our data into numerical dataset. Because there are only two variables for both our ordinal and nominal variables, we can do one hot encoding to prepare the data for the Apriori algorithm.



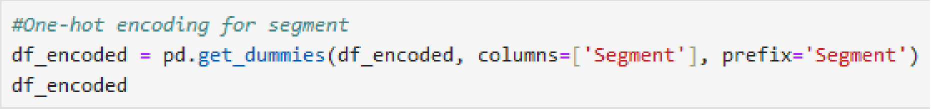
```
#One-hot encoding for Order Priority
df_encoded = pd.get_dummies(df, columns=['Order Priority'], prefix='Order Priority')
df_encoded
```

The next step was to encode Expedited Shipping. I decided to treat no expedited shipping as slow shipping, whereas expedited shipping is fast.

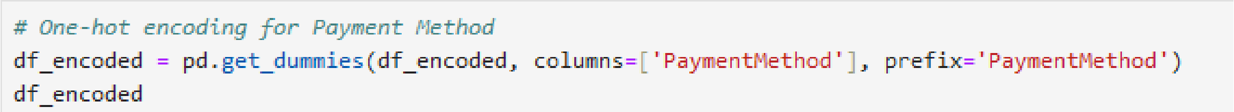


```
#One-hot encoding for Expedited Shipping
df_encoded = pd.get_dummies(df_encoded, columns=['ExpeditedShipping'], prefix='ExpeditedShipping')
df_encoded
```

Both the Segment and Payment Method columns need to be one hot encoded. This is because the data is nominal, and is not inherently a tiered structure such as “high” or “low” or a number ranking.



```
#One-hot encoding for segment
df_encoded = pd.get_dummies(df_encoded, columns=['Segment'], prefix='Segment')
df_encoded
```

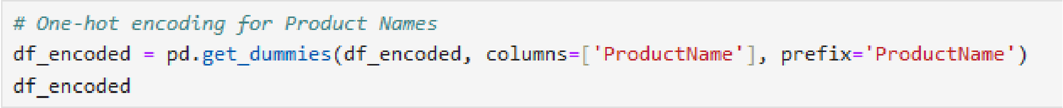


```
# One-hot encoding for Payment Method
df_encoded = pd.get_dummies(df_encoded, columns=['PaymentMethod'], prefix='PaymentMethod')
df_encoded
```

## C1c:TRANSACTIONALIZE DATA

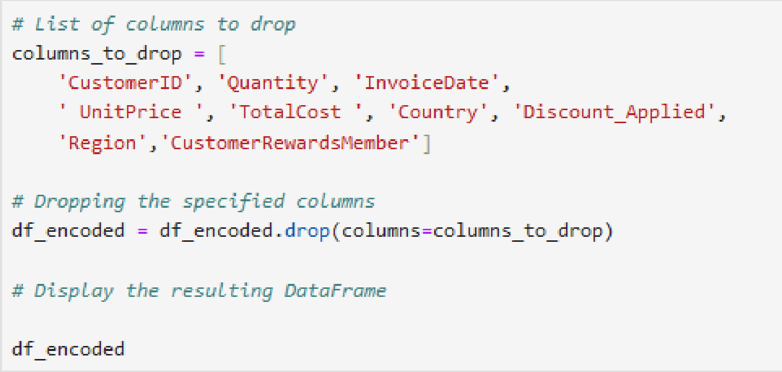
Here is the code and result I used to transactionalize the data. This process essentially turns the products into True and False values representing whether or not they are present in a particular order. This is necessary for the Apriori analysis to be able to read the data.

First, I one hot encoded the products:



```
# One-hot encoding for Product Names
df_encoded = pd.get_dummies(df_encoded, columns=['ProductName'], prefix='ProductName')
df_encoded
```

Then I dropped the columns we will not need in the algorithm



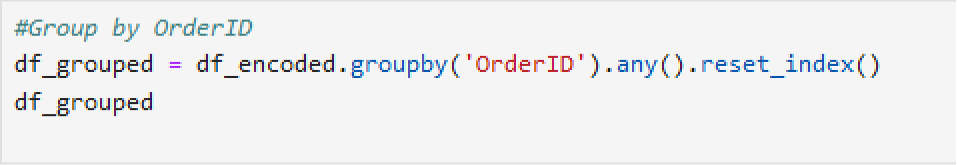
```
# List of columns to drop
columns_to_drop = [
    'CustomerID', 'Quantity', 'InvoiceDate',
    ' UnitPrice ', 'TotalCost ', 'Country', 'Discount_Applied',
    'Region','CustomerRewardsMember']

# Dropping the specified columns
df_encoded = df_encoded.drop(columns=columns_to_drop)

# Display the resulting DataFrame

df_encoded
```

Then I grouped the order by order ID:

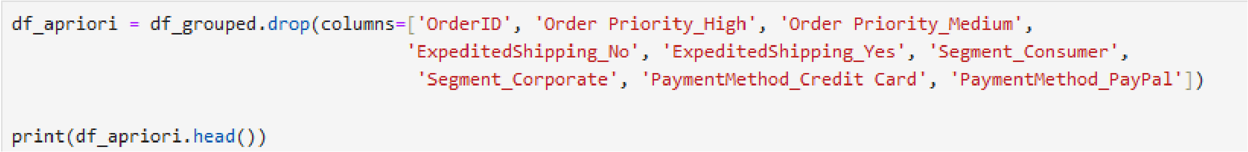


```
#Group by OrderID
df_grouped = df_encoded.groupby('OrderID').any().reset_index()
df_grouped
```

## C2:CLEAN DATASET COPY

At this point, what the professor tells me to do and what the evaluators want diverges. The professor told me that the cleaned data for submission should only include the product data, while the last evaluation I submitted stated “The transactional data and the selected nominal and ordinal variables should be combined into the same dataset. The combined dataset should be transformed so that it is suitable for market basket analysis.” I decided to include both copies of the cleaned dataset.

The first dataset was exported here. Then I dropped all the non-product columns and exported another dataset:



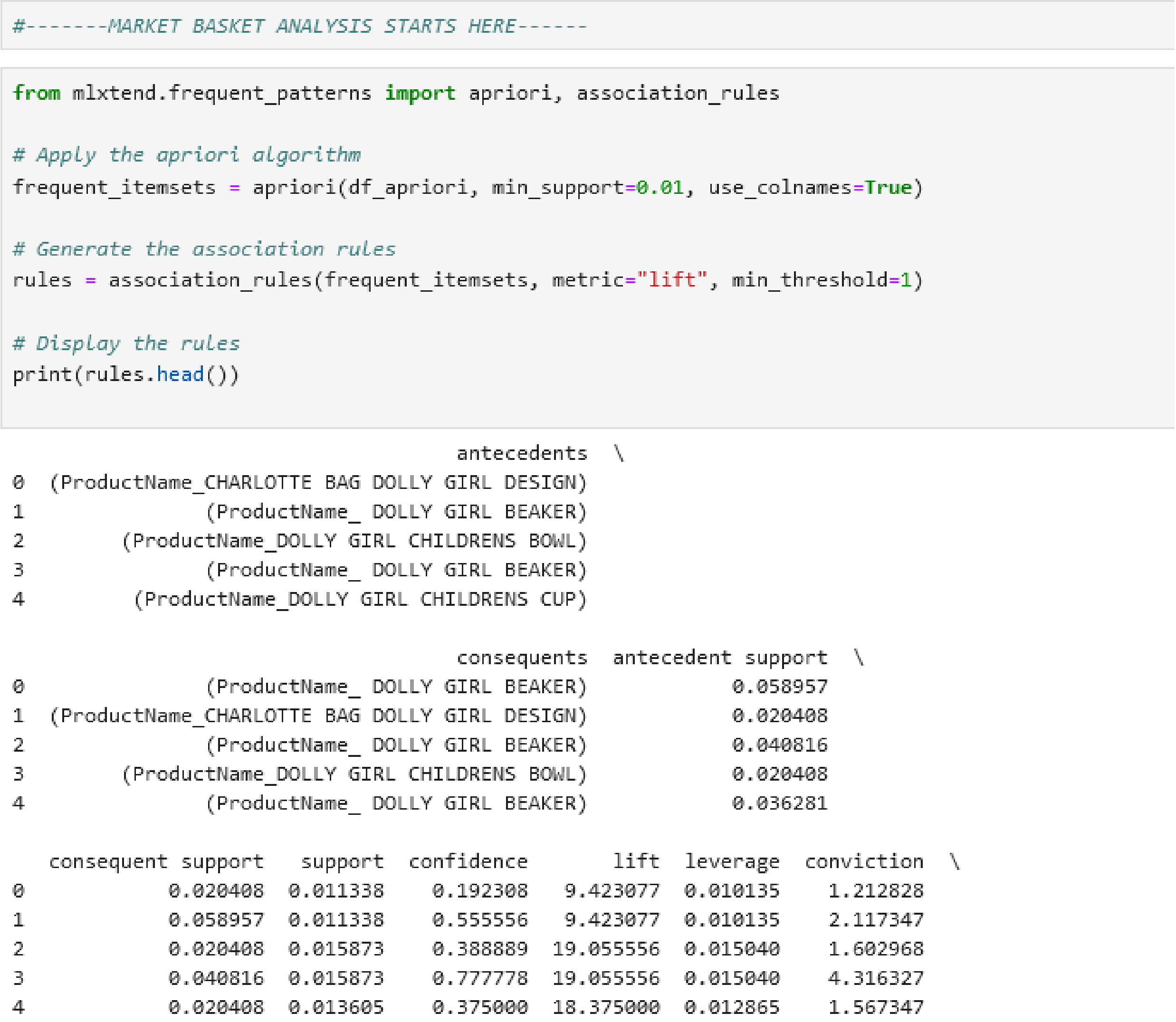
```
df_apriori = df_grouped.drop(columns=['OrderID', 'Order Priority_High', 'Order Priority_Medium',
                                      'ExpeditedShipping_No', 'ExpeditedShipping_Yes', 'Segment_Consumer',
                                       'Segment_Corporate', 'PaymentMethod_Credit Card', 'PaymentMethod_PayPal'])

print(df_apriori.head())
```

The professor told me I am to only perform the market basket analysis on the products, NOT including the ordinal and nominal variables. So this is the final dataset I need to perform market basket analysis.

## C3:EXECUTE CODE

Here is the code I used to run the Apriori algorithm, as well as a sample of the output:



```
#-------MARKET BASKET ANALYSIS STARTS HERE------
```

```
from mlxtend.frequent_patterns import apriori, association_rules

# Apply the apriori algorithm
frequent_itemsets = apriori(df_apriori, min_support=0.01, use_colnames=True)

# Generate the association rules
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)

# Display the rules
print(rules.head())
```

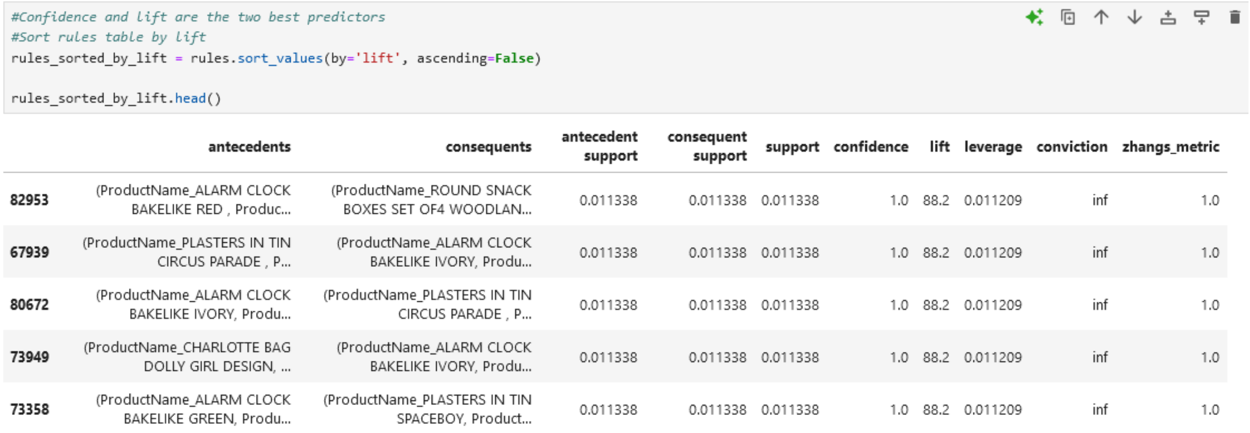
```
                                     antecedents  \
0  (ProductName_CHARLOTTE BAG DOLLY GIRL DESIGN)   
1               (ProductName_ DOLLY GIRL BEAKER)   
2        (ProductName_DOLLY GIRL CHILDRENS BOWL)   
3               (ProductName_ DOLLY GIRL BEAKER)   
4         (ProductName_DOLLY GIRL CHILDRENS CUP)   

                                     consequents  antecedent support  \
0               (ProductName_ DOLLY GIRL BEAKER)            0.058957   
1  (ProductName_CHARLOTTE BAG DOLLY GIRL DESIGN)            0.020408   
2               (ProductName_ DOLLY GIRL BEAKER)            0.040816   
3        (ProductName_DOLLY GIRL CHILDRENS BOWL)            0.020408   
4               (ProductName_ DOLLY GIRL BEAKER)            0.036281   

   consequent support   support  confidence       lift  leverage  conviction  \
0            0.020408  0.011338    0.192308   9.423077  0.010135    1.212828   
1            0.058957  0.011338    0.555556   9.423077  0.010135    2.117347   
2            0.020408  0.015873    0.388889  19.055556  0.015040    1.602968   
3            0.040816  0.015873    0.777778  19.055556  0.015040    4.316327   
4            0.020408  0.013605    0.375000  18.375000  0.012865    1.567347   
```

## C4:SUPPORT, LIFT, AND CONFIDENCE VALUES and C5:RELEVANT RULES

Next, I sorted the data by significance to find the best, most predictive rules. Here are the support, lift, and confidence values of the three most important rules from the algorithm. I will break them down in the next section:



```
#Confidence and lift are the two best predictors
#Sort rules table by lift
rules_sorted_by_lift = rules.sort_values(by='lift', ascending=False)

rules_sorted_by_lift.head()
```

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction | zhangs_metric |
|---|---|---|---|---|---|---|---|---|---|---|
| 82953 | (ProductName_ALARM CLOCK BAKELIKE RED , Produc... | (ProductName_ROUND SNACK BOXES SET OF4 WOODLAN... | 0.011338 | 0.011338 | 0.011338 | 1.0 | 88.2 | 0.011209 | inf | 1.0 |
| 67939 | (ProductName_PLASTERS IN TIN CIRCUS PARADE , P... | (ProductName_ALARM CLOCK BAKELIKE IVORY, Produ... | 0.011338 | 0.011338 | 0.011338 | 1.0 | 88.2 | 0.011209 | inf | 1.0 |
| 80672 | (ProductName_ALARM CLOCK BAKELIKE IVORY, Produ... | (ProductName_PLASTERS IN TIN CIRCUS PARADE , P... | 0.011338 | 0.011338 | 0.011338 | 1.0 | 88.2 | 0.011209 | inf | 1.0 |
| 73949 | (ProductName_CHARLOTTE BAG DOLLY GIRL DESIGN, ... | (ProductName_ALARM CLOCK BAKELIKE IVORY, Produ... | 0.011338 | 0.011338 | 0.011338 | 1.0 | 88.2 | 0.011209 | inf | 1.0 |
| 73358 | (ProductName_ALARM CLOCK BAKELIKE GREEN, Produ... | (ProductName_PLASTERS IN TIN SPACEBOY, Product... | 0.011338 | 0.011338 | 0.011338 | 1.0 | 88.2 | 0.011209 | inf | 1.0 |

Rule #1

The rule regarding row 82953  uses the antecedents ALARM CLOCK BAKELIKE RED, CHILDRENS CUTLERY SPACEBOY, SPACEBOY BIRTHDAY CARD, and ALARM CLOCK BAKELIKE PINK to predict the purchase of the consequents ROUND SNACK BOXES SET OF4 WOODLAND, CARD DOLLY GIRL, CHILDRENS CUTLERY DOLLY GIRL. A support rating of 0.0113 means that 1.13% of all transactions include all of those toys in the same basket. A confidence rating of 1.0 means that when the antecedents are bought, the consequents are bought 100% of the time. A lift rating of 88.2 means that the likelihood of buying the consequent products increases 88 times when the antecedent products are bought.

Rule #2

The rule regarding row 67939 is similar in predictive strength as the first rule. Rule 2 says that if a customer buys PLASTERS IN TIN CIRCUS PARADE, ALARM CLOCK BAKELIKE GREEN, CHARLOTTE BAG DOLLY GIRL DESIGN, and ALARM CLOCK BAKELIKE PINK, then we can predict the purchase of the consequents ALARM CLOCK BAKELIKE IVORY, PLASTERS IN TIN SPACEBOY. A support rating of 0.0113 means that 1.13% of all transactions include all of those toys in the same basket. A confidence rating of 1.0 means that when the antecedents are bought, the consequents are bought 100% of the time. A lift rating of 88.2 means that the likelihood of buying the consequent products increases 88 times when the antecedent products are bought.

Rule #3

The rule regarding row 80672 is similar in predictive strength as the first two rules. Rule 3 says that if a customer buys ALARM CLOCK BAKELIKE IVORY, ALARM CLOCK BAKELIKE GREEN, ALARM CLOCK BAKELIKE RED, and PLASTERS IN TIN SPACEBOY, then we can predict the purchase of the consequents PLASTERS IN TIN CIRCUS PARADE, CHARLOTTE BAG DOLLY GIRL DESIGN, and ALARM CLOCK BAKELIKE PINK. A support rating of 0.0113 means that 1.13% of all transactions include all of those toys in the same basket. A confidence rating of 1.0 means that when the antecedents are bought, the consequents are bought 100% of the time. A lift rating of 88.2 means that the likelihood of buying the consequent products increases 88 times when the antecedent products are bought.

## Part IV: Data Summary and Implications

### D1:SIGNIFICANCE OF RESULTS DISCUSSION

The interpretation and significance of these rules are given above. In each of the three rules above, I clearly highlight the precise rules on which purchases predict other purchases.

### D2:SIGNIFICANCE OF FINDINGS DISCUSSION and D3:RECOMMENDED COURSE OF ACTION

Finding a perfect predictive correlation between two groups of products should impact the way they are marketed, sold, and placed in the store. If you can be relatively certain a consumer will buy a second product after buying the first, the store should ensure that every consumer buying the first product is aware that the second product exists. Also, in a physical store, putting products with their predictive products next to each other in the store will ensure that consumers see both products and are able to buy both easily. If the products are purchased online, then the company should recommend the predicted products whenever the predictor products are bought. Other potential ideas for commonly purchased products would include cross marketing or possibly being sold together.

# Sources

No sources were used except for WGU official materials.