Peyton Bailey

March 23rd, 2025

D599 – Task 3

Part I: Research Question

A. Describe the purpose of your report by doing the following:

1. Propose one question relevant to a real-world organizational situation that you will answer using market basket analysis.

What product combinations are frequently purchased together at Allias Megastore?

2. Define one goal of the data analysis. Ensure your goal is reasonable within the scope of the provided scenario and is represented in the available data.

The analysis aims to identify top product combinations to optimize product placement within the store.

Part II: Market Basket Justification

B. Explain the reasons for using market basket analysis by doing the following:

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

"Market basket analysis is a data mining technique that analyzes patterns of co-occurrence and determines the strength of the link between products purchased together" (Turing, 2023). These patterns are discovered within the vast amount of customer sales data collected and stored. Market basket analysis utilizes the association rule {IF} → {THEN} to predict the probability of certain products being purchased together (Turing, 2023). The "IF" component is the antecedent, and the "THEN" component is the consequent. A set of items a customer purchases simultaneously is called an item set. Apriori is a standard algorithm that leverages these association rules between itemsets and provides different quantitative metrics to measure these rules' strengths. The Apriori algorithm uses a pre-defined probability to identify the itemsets that occur frequently in a dataset. This probability is a metric called support, which is the total number of transactions made for a particular product or itemset divided by the total number of transactions made. Then, the algorithm calculates the confidence of all possible rules. The confidence metric is the measure of the likelihood of purchasing an item "Y" given that "X" has been purchased (Datacamp, 2024). Another metric often used is lift, which compares the observed frequency of "X" and "Y" appearing together to the frequency expected if "X" and "Y" were independent. Lift values greater than one indicate a stronger association than chance. By inputting minimum values for these metrics, we tell the algorithm only to report rules above these cut-off points, thus eliminating rules that have no value in the decision-making process.

2. Provide one example of a transaction in the dataset.

OrderID: 536370

Products: ['INFLATABLE POLITICAL GLOBE ', 'SET2 RED RETROSPOT TEA TOWELS ', 'PANDA AND BUNNIES STICKER SHEET', 'RED TOADSTOOL LED NIGHT LIGHT', 'VINTAGE HEADS AND TAILS CARD GAME ', 'STARS GIFT TAPE ', 'VINTAGE SEASIDE JIGSAW PUZZLES', 'ROUND SNACK BOXES SET OF4 WOODLAND ', 'MINI PAINT SET VINTAGE ', 'MINI JIGSAW CIRCUS PARADE ', 'MINI JIGSAW SPACEBOY', 'SPACEBOY LUNCH BOX ', 'CIRCUS PARADE LUNCH BOX ', 'LUNCH BOX I LOVE LONDON', 'CHARLOTTE BAG DOLLY GIRL DESIGN', 'ALARM CLOCK BAKELIKE GREEN', 'ALARM CLOCK BAKELIKE RED ', 'ALARM CLOCK BAKELIKE PINK', 'SET 2 TEA TOWELS I LOVE LONDON ']

3. Summarize one assumption of market basket analysis.

One assumption of Market Basket Analysis is transaction independence. Each entry in a dataset represents an independent transaction, where a member makes purchases on a specific date. Medium article, “Boosting Sales with Data: The Power of Market Basket Analysis in Retail,” claims that “this assumption is foundational for market basket analysis, as the analysis hinges on understanding item associations within individual transactions” (Deniran, 2023). Thus, transactions that are not independent may impact the reliability of the association rules generated through the analysis.

Part III: Data Preparation and Analysis

C. Prepare the dataset for further analysis by doing the following:

1. Wrangle (i.e., transform) data by doing the following:

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

Ordinal – Order Priority, Customer Order Satisfaction

Nominal – Region, Payment Method

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

Order Priority – Ordinal Encoding

Customer Order Satisfaction – Ordinal Encoding

Region- Label Encoding

Payment Method – Label Encoding

c. Transactionalize the data for market basket analysis.

See attachment.

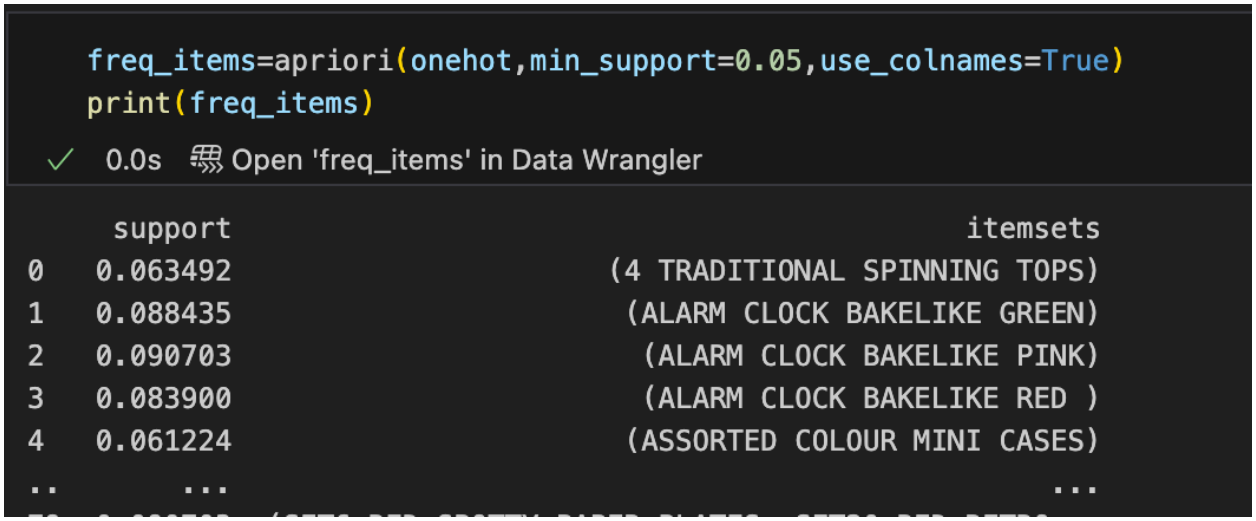
d. Explain and justify *each* step you took in parts C1a, C1b, and C1c.

The Order Priority variable has three categories: Low, Medium, and High. This classified the variable as an ordinal categorical variable since each category differed by name and order. The Customer Order Satisfaction variable is also ordinal, with each category representing a quality rating. The five categories are "Prefer Not to Answer," "Very Dissatisfied," "Dissatisfied," "Satisfied," and "Very Satisfied." Both variables were treated with ordinal encoding to preserve the order and quality rating. The Region and Payment Method are nominal categorical variables, meaning their categories differ by name only. Therefore, they were each treated with label encoding.

To transactionalize the data in preparation for market basket analysis, I grouped the data by "Order ID" and "Product Name" and converted the results into a list. This allowed each row to represent one transaction with the names of all products purchased during the transaction instead of multiple rows per transaction for each product. Afterward, this data was treated with one-hot encoding to make it compatible with the Apriori algorithm.

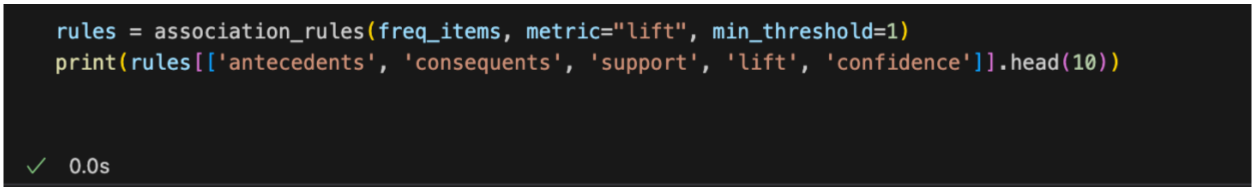
3. Execute the code used to generate association rules with the Apriori algorithm.

Provide a screenshot that demonstrates that the code is error-free.



```
freq_items=apriori(onehot,min_support=0.05,use_colnames=True)
print(freq_items)
```

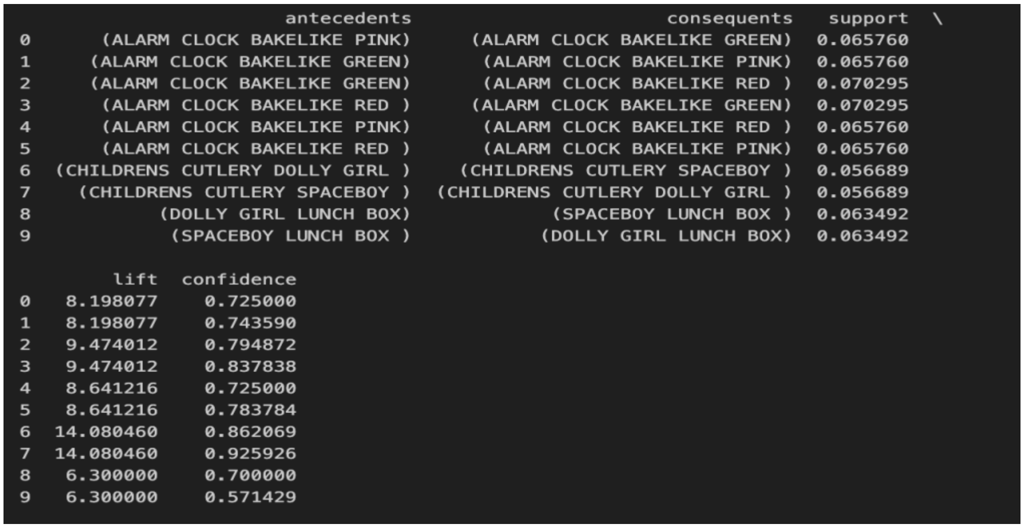
```
     support                          itemsets
0   0.063492       (4 TRADITIONAL SPINNING TOPS)
1   0.088435      (ALARM CLOCK BAKELIKE GREEN)
2   0.090703       (ALARM CLOCK BAKELIKE PINK)
3   0.083900       (ALARM CLOCK BAKELIKE RED )
4   0.061224      (ASSORTED COLOUR MINI CASES)
..       ...                               ...
```



```
rules = association_rules(freq_items, metric="lift", min_threshold=1)
print(rules[['antecedents', 'consequents', 'support', 'lift', 'confidence']].head(10))
```

✓ 0.0s

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



```
                       antecedents                      consequents   support  \
0      (ALARM CLOCK BAKELIKE PINK)     (ALARM CLOCK BAKELIKE GREEN)  0.065760
1     (ALARM CLOCK BAKELIKE GREEN)      (ALARM CLOCK BAKELIKE PINK)  0.065760
2     (ALARM CLOCK BAKELIKE GREEN)      (ALARM CLOCK BAKELIKE RED )  0.070295
3      (ALARM CLOCK BAKELIKE RED )     (ALARM CLOCK BAKELIKE GREEN)  0.070295
4      (ALARM CLOCK BAKELIKE PINK)      (ALARM CLOCK BAKELIKE RED )  0.065760
5      (ALARM CLOCK BAKELIKE RED )      (ALARM CLOCK BAKELIKE PINK)  0.065760
6  (CHILDRENS CUTLERY DOLLY GIRL )    (CHILDRENS CUTLERY SPACEBOY )  0.056689
7    (CHILDRENS CUTLERY SPACEBOY )  (CHILDRENS CUTLERY DOLLY GIRL )  0.056689
8           (DOLLY GIRL LUNCH BOX)            (SPACEBOY LUNCH BOX )  0.063492
9            (SPACEBOY LUNCH BOX )           (DOLLY GIRL LUNCH BOX)  0.063492

        lift  confidence
0   8.198077    0.725000
1   8.198077    0.743590
2   9.474012    0.794872
3   9.474012    0.837838
4   8.641216    0.725000
5   8.641216    0.783784
6  14.080460    0.862069
7  14.080460    0.925926
8   6.300000    0.700000
9   6.300000    0.571429
```

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

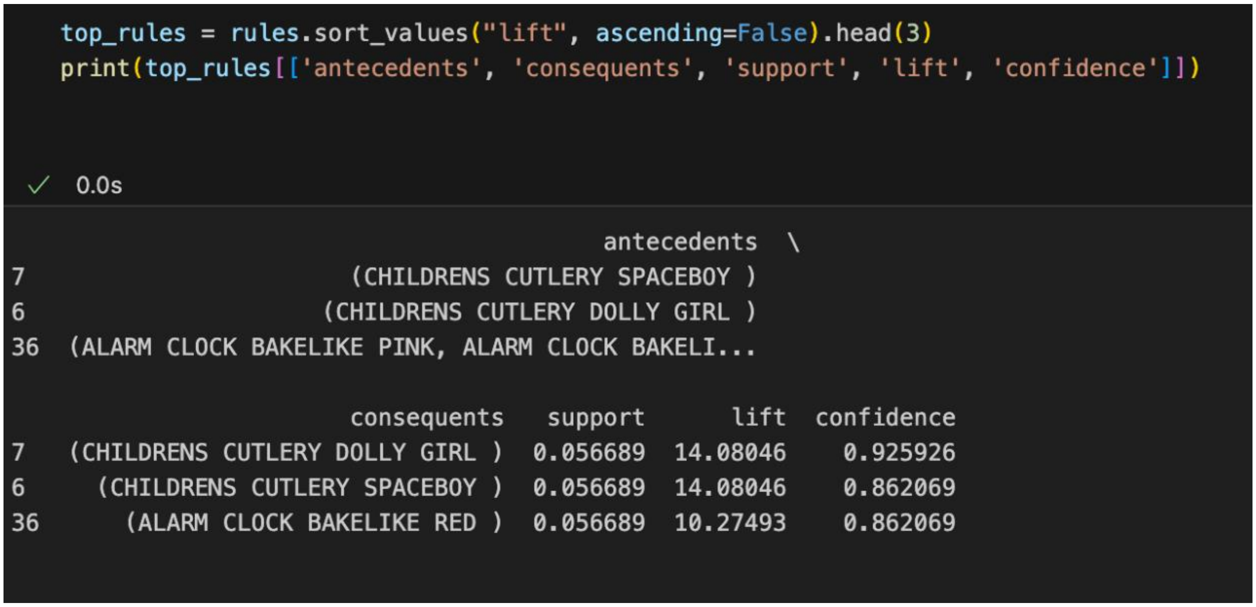
The top 3 rules have the following antecedent/consequent pairs:

Children’s Cutlery Spaceboy → Children’s Cutlery Dolly Girl

Children’s Cutlery Dolly Girl → Children’s Cutlery Spaceboy

Alarm Clock Bakelike Pink & Alarm Clock Bakelike Green → Alarm Clock Bakelike Red

Each of the rules has a support of 0.0567. The top two rules have a lift of 14.08, and the third, 10.27. The top rule has a confidence of 92.59%, and the bottom two both have a confidence of 86.21%.



```
top_rules = rules.sort_values("lift", ascending=False).head(3)
print(top_rules[['antecedents', 'consequents', 'support', 'lift', 'confidence']])
```

✓ 0.0s

```
                                      antecedents  \
7                        (CHILDRENS CUTLERY SPACEBOY )
6                      (CHILDRENS CUTLERY DOLLY GIRL )
36  (ALARM CLOCK BAKELIKE PINK, ALARM CLOCK BAKELI...

                       consequents   support      lift  confidence
7   (CHILDRENS CUTLERY DOLLY GIRL )  0.056689  14.08046    0.925926
6     (CHILDRENS CUTLERY SPACEBOY )  0.056689  14.08046    0.862069
36      (ALARM CLOCK BAKELIKE RED )  0.056689  10.27493    0.862069
```

Part IV: Data Summary and Implications

D. Summarize your data analysis by doing the following:

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

Support measures how frequently an item set appears in a dataset. These top three rules each have a support of 0.0567, which means that each of these antecedent/consequent pairs appears in 5.67% of the dataset. Lift evaluates how likely the consequent is to occur given the antecedent, compared to its baseline probability of occurring. A lift value of 1 indicates a positive association between the items, with higher values suggesting stronger relationships. The first of the top 3 rules (Rule #7) has a lift of 14.08, indicating that Children's Cutlery Dolly Girl is 14.08 times more likely to be purchased when Children's Cutlery Spaceboy is purchased. Interestingly, the 2nd rule has the same itemset but with a swapped antecedent and consequent. It also has the same lift value of 14.08, which adds even more confidence to this pair. This leads me to the last observed metric, confidence, a measure of how often the consequent is purchased when the antecedent is purchased. The first two rules share the same values for support and lift; however, the first rule has a confidence of 92.59%, and the second is 86.21%. Higher confidence levels suggest stronger dependencies between the antecedent and the consequent.

2. Explain the practical significance of your findings from the analysis.

The top two rules suggest a strong association between the Children's Cutlery Spaceboy and the Children's Cutlery Dolly Girl items. Allias Megastore should place these products together to provide easier customer access. The third rule pairs Alarm Clock Bakelike Pink and Alarm Clock Bakelike Green as the antecedent with Alarm Clock Bakelike Red as the consequent. These are the same products but in different colors, which suggests that customers may like to purchase a variety of colors. Products that have color variability should be placed in a way that customers are easily able to see those options.

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

Allias Megastore should place the associated item pairs within the top rules near each other. They should also include color variety when available. Beyond creating optimal product placement within the store, bundled discounts can be offered for products that appear within the top rules. This could lead to more sales, higher revenue, and more satisfied customers.

REFERENCES

Deniran, O. (2023). Boosting sales with data: The power of market basket analysis in retail. Medium **https://medium.com/@chemistry8526/boosting-sales-with-data-the-power-of-market-basket-analysis-in-retail-c79cc10a14df**

Srishti. (2023). Market basket analysis explained. Turing https://www.turing.com/kb/market-basket-analysis#market-basket-analysis-explained