D212 Data Mining II Task 3:
Market Basket Analysis

Part I: Research Question

Research Question:

What top related products/product groupings are items of interest for customers that we can incentivize by discounting to entice more customer purchases and reduce customer churn?

One Goal of Analysis:

1. Identify the top 3-5 products based on statistical supported evidence from high lift values.

Part II: Market Basket Justification

Explanation of Market Basket Analysis:

Market Basket Analysis is used to identify which products customers purchase together by using a variety of metrics such as support, confidence, and lift to scrutinize itemsets. To create itemsets, one has to consider antecedents (left hand side/LHS) and consequents (right hand side/RHS). The idea of antecedents and consequents is based on if a condition (antecedent) is met, then this result (consequent). For example, one of our dataset's final itemset pairings is Sandisk 128GB SD Card -> Sandisk 64GB SD Card or "If Sandisk 128gb SD Card is purchased, then Sandisk 64gb SD Card". The itemsets are scrutinized based on our three metrics of support, confidence, and lift. Support is the frequency of a item/itemset in our transaction data. Confidence is the probability of occurrence of the consequent given when the antecedent is present. Lift is the ratio of confidence to the consequent's probability of occurrence. A lift value above 1 indicates that the presence of the antecedent itemset increases the chances that the consequent itemset will occur in a specific transaction. By scrutinizing, or setting thresholds, for our three metrics, non-frequent or relevant itemsets are eliminated. These thresholds are dataset- and analysis-specific. The thresholds placed on our data were minimum support of 1%, minimum confidence of 15%, and minimum lift of 2.5.

Explanation of Apriori Algorithm:

Market Basket Analysis can be computationally expensive as the amount of transaction data and items increases. In our dataset, we have 7502 transactions and 119 unique items. The number of itemset pairings is exponentially even larger which is computationally expensive. To alleviate this, it is common to use the Apriori Algorithm. The Apriori Algorithm is based on the Apriori Principal: "All subsets of a frequent itemset must also be frequent". The algorithm takes a minimum support threshold, which for our analysis, was 1%. If an itemset is not above the minimum support threshold, then it is not included in our frequent itemsets list. The algorithm does not include any further variation that includes those itemsets. This reduces the computation expense as under-minimum, low support itemsets are automatically removed. From this itemset, I used the association_rules() function to add more thresholds, such as confidence. After, I scrutinized the association rules further by putting a minimum threshold of 2.5 for lift. This resulted in six final association rules which were actually only three unique pairings of itemsets.

Please see "Task 3 – Example of One Transaction.PNG" for a screenshot of one transaction and "D212 Performance Assessment Market Basket Analysis.ipynb" for my code.

Assumptions of Market Basket Analysis:

1. For a Market Basket Analysis to be effective, there must be a large number of real transactions to gleam any meaningful insights from it.
2. Items in the data without similar frequency can cause inaccuracies. Therefore, I chose to institute a minimum support level for items/itemsets of 1% (.01). This will remove any items without a decent frequency/support in our dataset.
3. To use the Apriori Algorithm for MBA, transaction data must be in an encoded format.

Part III: Data Prep/Analysis

Please see "encoded_dataset.csv" for the cleaned dataset suitable for market basket analysis.

Please see "Task 3 – Code to Generate Association Rules using Apriori Algorithm.PNG" for the screenshot of my error-free functional code using the Apriori Algorithm and association_rules.

Please see "Task 3 – Values for Support, Lift, and Confidence of the Association Rules Table.PNG" for a screenshot of the support, lift, and confidence values of the association rules table.

Please see below for my metric threshold values.

Metric	Threshold Amount
Support	0.01 (1%)
Confidence	0.15 (15%)
Lift	2.5

Please see "Task 3 – Top Three Rules.PNG" for a screenshot of the top three rules.

Please see "D212 Performance Assessment Market Basket Analysis.ipynb" for my code.

Part IV: Data Summary and Implications

Significance of Support, Confidence, and Lift:

Our three final association rules are:

1. SanDisk 128GB Ultra card -> SanDisk Ultra 64gb card
2. SanDisk Ultra 64GB card -> SanDisk 128GB Ultra card
3. Dust-Off Compressed Gas 2 pack, VIVO Dual LCD Mount -> FEIYOLD Blue Light Glasses

For rule to be included in our results, itemsets associations must have a minimum support of 1%, minimum confidence of 15%, and minimum lift of 2.5. The itemset must appear at least 1% of the time in our dataset. The probability of the consequent itemset's occurrence given when

the antecedent itemset is present must be at least 15%. A lift value above 1 indicates that the presence of the antecedent itemset increases the chances that the consequent itemset will occur in a specific transaction. This means that the antecedent and consequent association is significant. I wanted only rules that have itemsets which are significantly associated with each other, so I used the lift value of 2.5 as the minimum threshold. Finally, our final three rules contain itemsets with associations that are significantly related to each other, show up frequently in the dataset (1%), and are present together at least 15% of the time the antecedent itemset is present.

Practical Significance:

The practical significance of our findings is that our itemsets occurred in at least 1% of our transactions. This is still a small number of transactions (~317 out of 7502 transactions). This is not a large number of transactions. However, depending on current inventory, this may still be a potential area of opportunity for discounts/promotions. If the executives are satisfied with this result, then we would offer a discount/promotion. If they are not, then I would explore other support frequency values to capture itemsets that are present more in our dataset. From there, I can scrutinize further with confidence and lift thresholds.

Recommendations:

1. Offer a promotional discount when a customer is trying to purchase one of the three antecedent itemsets found in our three association rules for its corresponding consequent itemset. For example, a customer is trying to purchase a Dust-Off Compressed Gas 2-Pack and VIVO Dual LCD Monitor Mount together. We offer them a promotional discount for FEIYOLD Blue Light Blocking Glasses.
2. When expanding the final three rules to the entire rules table that satisfies our three thresholds, there are a total of six association rules. In these six association rules, there are only three unique itemset pairings between them as they are variations on their antecedent-consequent relationship. I would offer discounts similarly to recommendation number 1 but expand the discount regardless of antecedent-consequent relationship. If a customer is purchasing one of these four itemsets, then a promotional discount for another itemset will be offered to the customer, provided that the association rule exists in our table.