**Task 3:** Market Basket Analysis

Krescens Kok

Department of Technology, Western Governors University

**D599:** Data Preparation and Exploration

Keiona Middleton

2025 September 25

**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.**

A question that is relevant to a real-world organizational situation, in this case, Allias Megastore, is what products are customers frequently purchasing together?

**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.**

A goal of this data analysis is to use a Market Basket analysis to identify which combinations of products are frequently purchased together by customers at Allias Megastore. This will help the company identify how it can structure its store to increase sales.

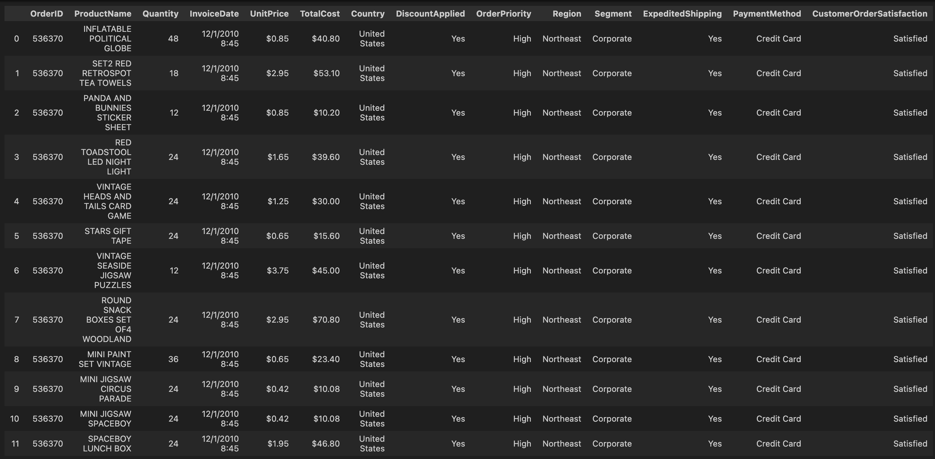
**Part II: Market Basket Justification**

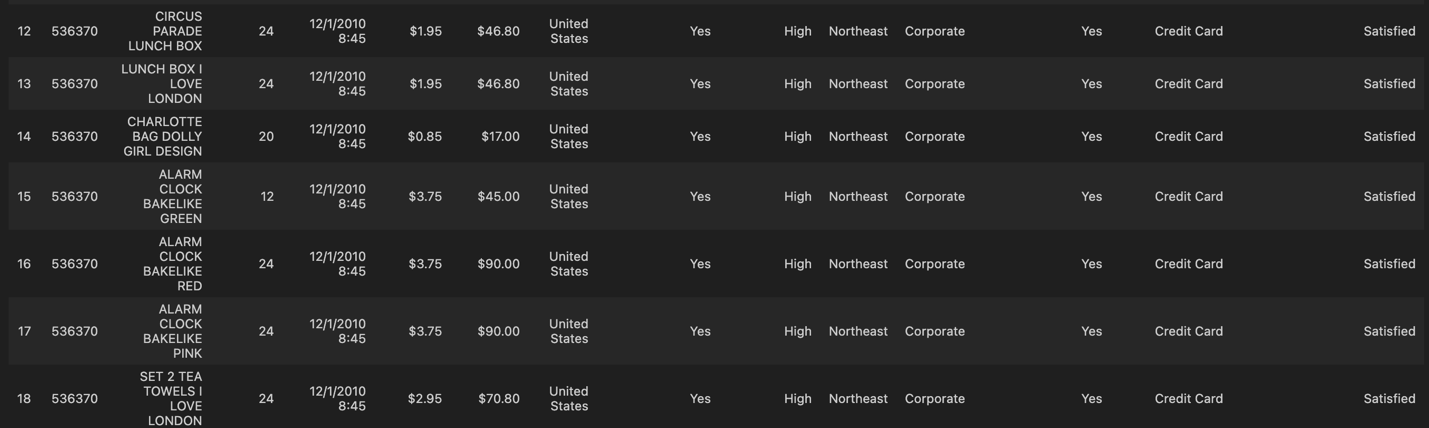
**B.  Discuss market basket analysis by doing the following:**

**1.  Explain how the Apriori algorithm, which is used for the market basket, analyzes the provided dataset, including expected outcomes.**

According to the course textbook, the Apriori algorithm starts by identifying individual items in the dataset that meet a minimum support threshold. It then iteratively generates itemsets, or combinations of products, of increasing length by joining frequent itemsets from the previous iteration. Afterwards, pruning occurs where itemsets that do not meet the minimum support threshold are dropped. After identifying the frequent itemsets, the algorithm generates association rules and calculates the confidence and lift to evaluate their strength (*Course |*, n.d.-d). The expected outcome of the algorithm is a set of frequent itemsets and the association rules that were generated.

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

**** The example provided of a one transaction is the OrderID = 536370. There were 19 items associated with this transaction.

****

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

An assumption of the market basket analysis is that the transactions are independent of one another. This means that it is assumed that each transaction does not directly affect another customer’s transaction. This is assumed because if this weren’t the case, the measures from the algorithm could give us biased results. For example, if multiple transactions came from the same customer who bought the same items over and over again, this would inflate the support/confidence, making an association rule seem stronger than it really is (Deniran, 2023).

**Part III: Data Preparation and Analysis**

**C.  Prepare the dataset for further analysis by doing the following using R or Python:**

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

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

The ordinal variables that I will be choosing are OrderPriority and CustomerOrderSatisfaction. OrderPriority is ordinal because the categories are High and Medium, where there can be a scale where High is greater than Medium. Similarly to CustomerOrderSatisfaction, there are 5 categories: Prefer not to answer, Very dissatisfied, dissatisfied, satisfied, and very satisfied, where this column represents the customer level of satisfaction with the order (range 1-5).

The nominal variables that I will be choosing are DiscountApplied, which has values yes or no, and Segment, which has values Consumer or Corporate.

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

**A screen shot of a computer

AI-generated content may be incorrect.Ordinal Encoder:**

**Nominal Encoder:**

**A screenshot of a computer

AI-generated content may be incorrect.**

**c.  Justify each step you took in part C1b.**

Ordinal encoding was used for the ordinal variables because the categories have a natural

order. The encoder assigns a unique integer to each category according to this order, preserving the ranking of information. This approach is advantageous because it allows the function to recognize the relative position of each category, making it suitable for ordinal data where the order of categories carries meaningful information. Lastly, the original ordinal columns were both dropped from the data frame to prevent redundancy.

One-hot encoding was applied to the nominal variables because there is no inherent order or relationship between their categories. This encoding method creates a separate binary variable for each category, making it well-suited for nominal data. In the code, the first category was dropped to reduce redundancy and prevent multicollinearity (GeeksforGeeks, 2025e). The value of the dropped category can still be inferred from the other dummy variables, since each observation can only belong to one category.

**d.  Export the dataset that includes all encoded variables.**

Attached to the submission as ‘Encoded\_All.csv’

**2.  Perform a market basket analysis by completing the following:**

**a.  Transactionalize the dataset with only the relevant variables for market basket analysis.**

**A screenshot of a computer

AI-generated content may be incorrect.** In the screenshot below, the data is being transactional by a group by function, grouping by the order ID.

**b.  Export the transactionalized dataset for market basket analysis with only the relevant variables.**

Attached to the submission as ‘Transactional.csv’

**c.  Execute the error-free code used to generate association rules with the Apriori algorithm. Provide a screenshot of the top three rules generated by the Apriori algorithm sorted by your chosen metric (i.e., confidence, support, or lift).**

The first screenshot reveals the code that was used for the Apriori algorithm and generating the rule sets. A min\_support = .01 means that only items that appear in at least 1% of the transactions will be included.

**A screen shot of a computer

AI-generated content may be incorrect.**

The rules were then generated with the association\_rules() function, where the lift was the metric used for a threshold, and the min\_threshold was set equal to 1. The lift was used because it tells us if the relationship between the antecedent and consequent is stronger than random chance. A lift score of greater than 1 reveals that there is a positive correlation between the antecedent and consequent and that buying a product/products increases the chance of another product/products being bought. A lift score of less than 1 reveals that there is a negative correlation between the antecedent and consequent, and buying a product/products decreases the chances of another A screenshot of a computer program

AI-generated content may be incorrect.product/products. When lift = 1, this means that there is no real association.

**Part IV: Data Summary and Implications**

**D.  Summarize your data analysis by doing the following:**

**1.  Justify the criteria used to generate the top three rules (e.g., "The association rules were sorted by lift in ascending order because...").**

The association rules were sorted by lift in descending order because the lift value reveals if the relationship between the antecedent and consequent is stronger than random chance (GeeksforGeeks, 2025f). I wanted to find the rules that had a strong association with each other, with non-random associations. However, a limitation of choosing to sort by the lift is that the results could reveal very rare itemsets that might not happen frequently.

With a lift score that is greater than 1, it shows a positive correlation, and the presence of the antecedent increases the presence of the consequent. Therefore, by sorting by the lift in descending order, we get the highest lift value at the top of the list, revealing which rules have the strongest relationship.

**2.  Explain support, lift, and confidence for the top three rules generated by the Apriori algorithm.**

The top 3 rules that were generated by the Apriori algorithm have the same scores for support, lift, and confidence, which are .011338, 88.2, and 1.

A support score of .011338 means that 1.13% of all transactions will have this itemset.

So, for the first rule, the RED SPOTTY PAPER PLATES, ROUND SNACK BOXES SET OF4 WOODLAND, ALARM CLOCK BAKELINE PINK, RED SPOTTY PAPER CUPS, ALARM CLOCK BAKELINE RED, ALARM CLOCK BAKELINE GREEN, AND PLASTERS IN TIN CIRCUS PARADE will show up in a transaction 1.13% of the time. For the 2nd rule, BLUE POLKADOT PARTY CANDLES, RED SPOTTY PAPER PLATES, SET10 PINK POLKADOT PARTY CANDLES, SET10 RED POLKADOT PARTY CANDLES, and SET6 RED SPOTTY PAPER CUPS will show up in a transaction 1.13% of the time. For the 3rd rule, SPACEBOY BIRTHDAY CARD, CHILDRENS CUTLERY DOLLY GIRL, ALARM CLOCK BAKELIKE RED, ALARM CLOCK BAKELIKE PINK, ROUND SNACK BOXES SET OF4 WOODLAND, CARD DOLLY GIRL, and CHILDRENS CUTLERY SPACEBOY will show up in a transaction 1.13% of the time.

The **lift score of 88.2** reveals that when the antecedent items are purchased, the consequent items are **88.2 times more likely to be purchased** compared to what would be expected if the two were independent (random chance). So, for the first rule, the RED SPOTTY PAPER PLATES, ROUND SNACK BOXES SET OF4 WOODLAND, and ALARM CLOCK BAKELINE PINK being bought means that RED SPOTTY PAPER CUPS, ALARM CLOCK BAKELINE RED, ALARM CLOCK BAKELINE GREEN, AND PLASTERS IN TIN CIRCUS PARADE are 88.2 times more likely to be purchased as well. For the 2nd rule, BLUE POLKADOT PARTY CANDLES and RED SPOTTY PAPER PLATES being bought means that SET10 PINK POLKADOT PARTY CANDLES, SET10 RED POLKADOT PARTY CANDLES, and SET6 RED SPOTTY PAPER CUPS are 88.2 times more likely to be purchased. Lastly, the SPACEBOY BIRTHDAY CARD, CHILDRENS CUTLERY DOLLY GIRL, ALARM CLOCK BAKELIKE RED and ALARM CLOCK BAKELIKE PINK being bought means that the ROUND SNACK BOXES SET OF4 WOODLAND, CARD DOLLY GIRL, and CHILDRENS CUTLERY SPACEBOY are 88.2 times more likely to be purchased.

The confidence score of 1 means that when the antecedent items are bought, the consequent items will always be bought. So for the first rule, if the RED SPOTTY PAPER PLATES, ROUND SNACK BOXES SET OF4 WOODLAND, and ALARM CLOCK BAKELINE PINK are bought, then the RED SPOTTY PAPER CUPS, ALARM CLOCK BAKELINE RED, ALARM CLOCK BAKELINE GREEN, AND PLASTERS IN TIN CIRCUS PARADE will always be bought. For the 2nd rule, if the BLUE POLKADOT PARTY CANDLES and RED SPOTTY PAPER PLATES are bought, then the SET10 PINK POLKADOT PARTY CANDLES, SET10 RED POLKADOT PARTY CANDLES, and SET6 RED SPOTTY PAPER CUPS will always be bought. Lastly, for the third rule, if SPACEBOY BIRTHDAY CARD, CHILDRENS CUTLERY DOLLY GIRL, ALARM CLOCK BAKELIKE RED and ALARM CLOCK BAKELIKE PINK are bought, then the ROUND SNACK BOXES SET OF4 WOODLAND, CARD DOLLY GIRL, and CHILDRENS CUTLERY SPACEBOY will always be bought as well.

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

For the top 3 rules, the lift score of 88.2 indicates that the antecedent and consequent will be purchased 88.2 times more frequently than what would be expected from random chance alone, or if they were independent. A confidence score of 1 means that if the antecedent is bought, then the consequent will also be bought 100% of the time. Lastly, the support of .011338 means that 1.13% of the transactions contain this itemset.

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

A recommended course of action for the Megastore is to take into consideration the top 3 rules to identify how the store can rearrange the antecedent and consequent items so that more customers will also buy the same items because they see the items next to each other. In addition, the store could promote packaged bundles, so that the items are marketed together and therefore, will be bought together as well. Lastly, the store could ensure that they have the antecedent and consequent item sets in stock, since if the antecedent is purchased, so is the consequent, so it is a good idea to have both in stock at all times.

**References**

*Course |*. (n.d.-d). <https://apps.cgp-oex.wgu.edu/learning/course/course-v1:WGUx+OEX0395+v01/block-v1:WGUx+OEX0395+v01+type@sequential+block@5415f5fe74ad4d6cb4f8c5a5689f490b/block-v1:WGUx+OEX0395+v01+type@vertical+block@1ce05965e7294dd3960aae7463ddde21>

Deniran, O. H. (2023, November 27). 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>

GeeksforGeeks. (2025e, August 5). *How to perform ordinal encoding using Sklearn*. GeeksforGeeks. <https://www.geeksforgeeks.org/machine-learning/how-to-perform-ordinal-encoding-using-sklearn/>

GeeksforGeeks. (2025f, September 18). *Apriori algorithm*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/apriori-algorithm/