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.

Research question: What product combinations are most commonly purchased together by corporate customers in the Southeast region?

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.

One goal of the data analysis is to identify trends in product combinations, in order to influence the company's marketing and sales strategies for corporate customers in the Southeast region.

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.

First, the Apriori algorithm identifies all products that meet a minimum number of transactions (support threshold). Then, it removes any products that did not meet the support threshold from future algorithm iterations. Then, the algorithm iteratively pairs up the remaining items and checks if the number of transactions in which those two items are purchased together also exceeds the support threshold. Then, the algorithm creates itemsets of three items based on what 2-itemsets surpassed the support threshold. Then, the algorithm checks each 3-itemset to see if they include any 2-itemsets that were removed in the previous iteration. If they do include a removed 2-item set, they are pruned from the algorithm. This process continues until no frequent itemsets are found. At this point, the algorithm examines the confidence and lift of each frequent itemset and creates association rules based on those metrics.

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

Transaction with OrderID 539407: This was a transaction that occurred on 12/17/2010 at 12:56. The customer was corporate and located in the Southeast region of the United States. They purchased 12 Blue Harmonicas in Box, 48 Felt Toadstools Small, 192 Felt Farm Animal White Bunnies, 8 Retrospot Party Bag Sticker Sets, and 72 Mini Paint Sets Vintage. They paid with Credit Card and were dissatisfied with their order.

3. Summarize one assumption of market basket analysis.

One assumption Market Basket Analysis makes is that customer behavior will remain consistent as time goes on. In order for our findings to be relevant, we have to assume that tomorrow's customers will behave similarly to today's customers. If this assumption is true, our analysis of past customer behavior can be used to influence future marketing and sales strategies.

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.

<u>Ordinal variables:</u>

1. CustomerOrderSatisfaction - This variable is ranked on a 5-point scale from "Prefer Not To Answer" to "Very Satisfied"
2. OrderPriority - This variable is ranked on a 3-point scale, either low, medium, or high

<u>Nominal variables:</u>

1. Region - This variable indicates the geographic location of the customer
2. Segment - This variable indicates the market segment to which the customer belongs. (corporate v. consumer)

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

For the ordinal variables, I utilized ordinal encoding:

- CustomerOrderSatisfaction values were converted to numbers 0 to 4
- OrderPriority values were converted to numbers 1 to 3

For the nominal variables, I utilized one-hot encoding:

-Region and Segment were each expanded into binary columns, such as Region_Southeast or Segment_Corporate.

c. Justify each step you took in part C1b.

Ordinal encoding of CustomerOrderSatisfaction and OrderPriority values allows for easier analysis while also preserving the order

One-hot encoding of Region and Segment allows for easier filtering for the Market Basket Analysis, and avoids assigning an arbitrary or misleading order to the data

d. Export the dataset that includes all encoded variables.

You will find this dataset attached as EncodedMegastoreDataset.csv

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

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

First, I filtered out all orders not made in the Southeast region

Then, I grouped all orders by OrderID and ProductName, summing the quantities of each product per order.

Next, I created a basket matrix in Python where each row represented a unique order and each column represents a product name. If an order contained a given product name, the corresponding cell contains a boolean True. If not, a boolean False.

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

You will find this data attached as TransactionalizedMegastoreDataset.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).

These are the top three rules generated by the Apriori algorithm sorted by lift:

	antecedents	consequents	antecedent supp	consequent supp	support	confidence	lift
34	frozenset({'PACK OF 6 SKULL PAPER PLATES'})	frozenset({'PACK OF 6 SKULL PAPER CUPS'})	0.05555555556	0.05555555556	0.05555555556	1	18
35	frozenset({'PACK OF 6 SKULL PAPER CUPS'})	frozenset({'PACK OF 6 SKULL PAPER PLATES'})	0.05555555556	0.05555555556	0.05555555556	1	18
8	frozenset({'JUMBO BAG PEARS'})	frozenset({'JUMBO BAG APPLES'})	0.05555555556	0.06666666667	0.05555555556	1	15
9	frozenset({'JUMBO BAG APPLES'})	frozenset({'JUMBO BAG PEARS'})	0.06666666667	0.05555555556	0.05555555556	0.8333333333	15
3	frozenset({'ALARM CLOCK BAKELIKE GREEN'})	frozenset({'ALARM CLOCK BAKELIKE RED '})	0.08888888889	0.06666666667	0.06666666667	0.75	11.25
2	frozenset({'ALARM CLOCK BAKELIKE RED '})	frozenset({'ALARM CLOCK BAKELIKE GREEN'})	0.06666666667	0.08888888889	0.06666666667	1	11.25

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 rules with the highest lift are much more likely to be purchased together than to be purchased independently. Because our research question specifically asks for product combinations that are purchased together, lift is the most important criterion to answer our research question

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

Support: Support tells us what portion of our total transactions include a particular item or item combination. For example, our top three association rules(Pack of 6 Paper Skull Plates + Pack of 6 Paper Skull cups, Jumbo Bag Pears + Jumbo Bag Apples, Alarm Clock Bakelike Green + Alarm Clock Bakelike Red) have supports (rounded to the nearest tenth of a percent) of 5.6%, 5.6%, and 6.7% respectively. This means that 5.6% of transactions include the combination in our first rule, 5.6% of transactions include the combination in our second rule, and 6.7% of transactions include the combination in our third rule.

Confidence: Confidence tells us what portion of our transactions that contain one item also contain the other item. For example, for our top three association rules, the confidences (to the nearest percent) are 100%, 100%, and 75% respectively. This means that 100% of orders that contain Pack of 6 Paper Skull Plates also contain Pack of 6 Paper Skull cups, 100% of orders that contain Jumbo Bag Pears also contain Jumbo Bag Apples, and 95% of transactions that contain Alarm Clock Bakelike Green also contain Alarm Clock Bakelike Red.

Lift: Lift tells us how much more likely a customer is to purchase one item if they also purchase the other compared to how likely a customer is to purchase that item independently of the other. For example, for our top three association rules, the lifts (to the nearest tenth) are 18, 15, and 11.25 respectively. This means that the products in our top three association rules are very closely linked, since they are 18 times more likely, 15 times more likely, and 11.25 times more likely, respectively, to be purchased together than to be purchased independently.

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

The practical significance of these findings are related to the marketing and sales strategies of Allias Megastore. These association rules provide insights that can guide the Allias team to implement new ideas, such as offering bundle discounts on items frequently purchased together or optimizing the placement of products near other products they are frequently purchased with, in order to increase multi-item sales and increase customer satisfaction.

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

I recommend the following course of action for Allias Megastore based on my findings:

- Place products that are frequently purchased together near one another both in physical stores and on online platforms
- Implement a “Frequently purchased together” feature on online platforms
- Offer bundle discounts on items that are purchased together

I believe these actions will both increase multi-item sales and offer a convenience that will improve customer satisfaction scores.

F. Acknowledge reference sources used to support the Python or R code application. All references listed should also include an in-text citation in the code annotation. Be sure the sources are reliable. If no sources were used for coding, state, "No sources used."

No sources used.