Nicole Gallo
D599 – Data Preparation and Exploration
July 14, 2025
TCN1 Task 3: Market Basket Analysis

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 most frequently purchased together by corporate customers in the Northeast 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.

   The goal of the data analysis is **to identify high-confidence product combinations** that can inform cross-selling strategies **for corporate customers in the Northeast region**, enabling targeted product bundles and promotional campaigns.

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 (MBA) is a data mining technique that identifies associations between items that are frequently purchased together.

   In this analysis, each sales transaction is represented as a "basket" of products, and the Apriori algorithm is used to identify patterns of co-purchased items. MBA helps identify which products tend to be bought together in the same orders placed by corporate customers in the Northeast region.

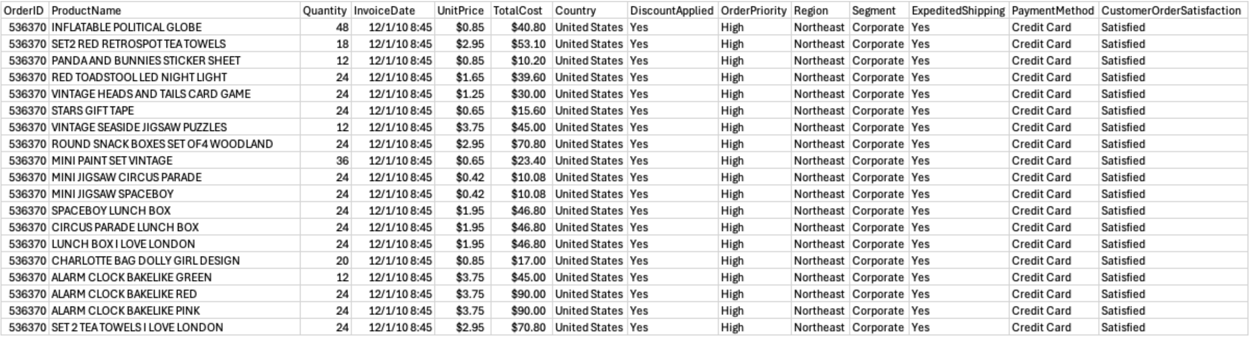
   The expected outcome is a set of association rules that show product combinations with high frequency, strong confidence in the likelihood of buying item B when item A is purchased, and the strength of the rule used when compared to random chance. This set of association rules can be used to create bundled offers, inform targeted marketing campaigns, and optimize product placements.

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

   An example of a transaction in the dataset is **OrderID 536370**, placed by a corporate customer in the Northeast region that includes the following 19 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



OrderID	ProductName	Quantity	InvoiceDate	UnitPrice	TotalCost	Country	DiscountApplied	OrderPriority	Region	Segment	ExpeditedShipping	PaymentMethod	CustomerOrderSatisfaction
536370	INFLATABLE POLITICAL GLOBE	48	12/1/10 8:45	$0.85	$40.80	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	SET2 RED RETROSPOT TEA TOWELS	18	12/1/10 8:45	$2.95	$53.10	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	PANDA AND BUNNIES STICKER SHEET	12	12/1/10 8:45	$0.85	$10.20	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	RED TOADSTOOL LED NIGHT LIGHT	24	12/1/10 8:45	$1.65	$39.60	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	VINTAGE HEADS AND TAILS CARD GAME	24	12/1/10 8:45	$1.25	$30.00	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	STARS GIFT TAPE	24	12/1/10 8:45	$0.65	$15.60	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	VINTAGE SEASIDE JIGSAW PUZZLES	12	12/1/10 8:45	$3.75	$45.00	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	ROUND SNACK BOXES SET OF4 WOODLAND	24	12/1/10 8:45	$2.95	$70.80	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	MINI PAINT SET VINTAGE	36	12/1/10 8:45	$0.65	$23.40	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	MINI JIGSAW CIRCUS PARADE	24	12/1/10 8:45	$0.42	$10.08	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	MINI JIGSAW SPACEBOY	24	12/1/10 8:45	$0.42	$10.08	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	SPACEBOY LUNCH BOX	24	12/1/10 8:45	$1.95	$46.80	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	CIRCUS PARADE LUNCH BOX	24	12/1/10 8:45	$1.95	$46.80	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	LUNCH BOX I LOVE LONDON	24	12/1/10 8:45	$1.95	$46.80	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	CHARLOTTE BAG DOLLY GIRL DESIGN	20	12/1/10 8:45	$0.85	$17.00	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	ALARM CLOCK BAKELIKE GREEN	12	12/1/10 8:45	$3.75	$45.00	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	ALARM CLOCK BAKELIKE RED	24	12/1/10 8:45	$3.75	$90.00	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	ALARM CLOCK BAKELIKE PINK	24	12/1/10 8:45	$3.75	$90.00	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied
536370	SET 2 TEA TOWELS I LOVE LONDON	24	12/1/10 8:45	$2.95	$70.80	United States	Yes	High	Northeast	Corporate	Yes	Credit Card	Satisfied

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

   An assumption we can make of the MBA is that it assumes products frequently purchased together in the past will likely be purchased together again in the future. This assumption supports the idea that historical purchasing behavior can be used to guide future promotions, bundling offers, or product stocking decisions.

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 variables:
   1. `CustomerOrderSatisfaction` – This variable is ranked on a scale from 0 to 4, 0 = “Prefer not to answer” and 4 = “Very Satisfied”

2. `OrderPriority` – The order priority is labeled as “Low”, “Medium”, or “High”, representing urgency

Nominal variables:
1. `Region` – This variable indicates the geographic location of the customer (e.g., Northeast, Southeast, etc.)
2. `Segment` – The market segment (corporate vs consumer) to which the customer belongs

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

For ordinal variables, I performed ordinal encoding:

- `CustomerOrderSatisfaction` values were converted to numbers 0 through 4
- `OrderPriority` values were mapped to numerical ranks: Low = 1, Medium = 2, High = 3

For nominal variables, I performed one-hot encoding:

- `Region` and `Segment` were each expanded into binary columns
    - *Example* – `Region_Northeast`, `Segment_Corporate`; this allows for easy filtering and analysis without imposing an order.

c. Transactionalize the data for market basket analysis.

I started by filtering the dataset to include only transactions made by corporate customers in the Northeast region, based on our research question: **What product combinations are most frequently purchased together by corporate customers in the Northeast region?**

Next, I grouped the data by `OrderID` and `ProductName`, summing the quantities of each product per order. I created a basket matrix in Juypter Lab using Python so that each row can represent a unique order, and each column can represent a product.

Each cell is then converted into binary values: 1 if the product was purchased in the order, 0 if not.

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

- **Encoding**: This step demonstrates the proper use of ordinal and one-hot encoding on relevant categorical variables, as part of data wrangling best practices.
- **Saving Encoded Dataset**: This step ensures that the full, clean version of the dataset is preserved and available for inspection or future analysis. This includes the encodings from the previous step.

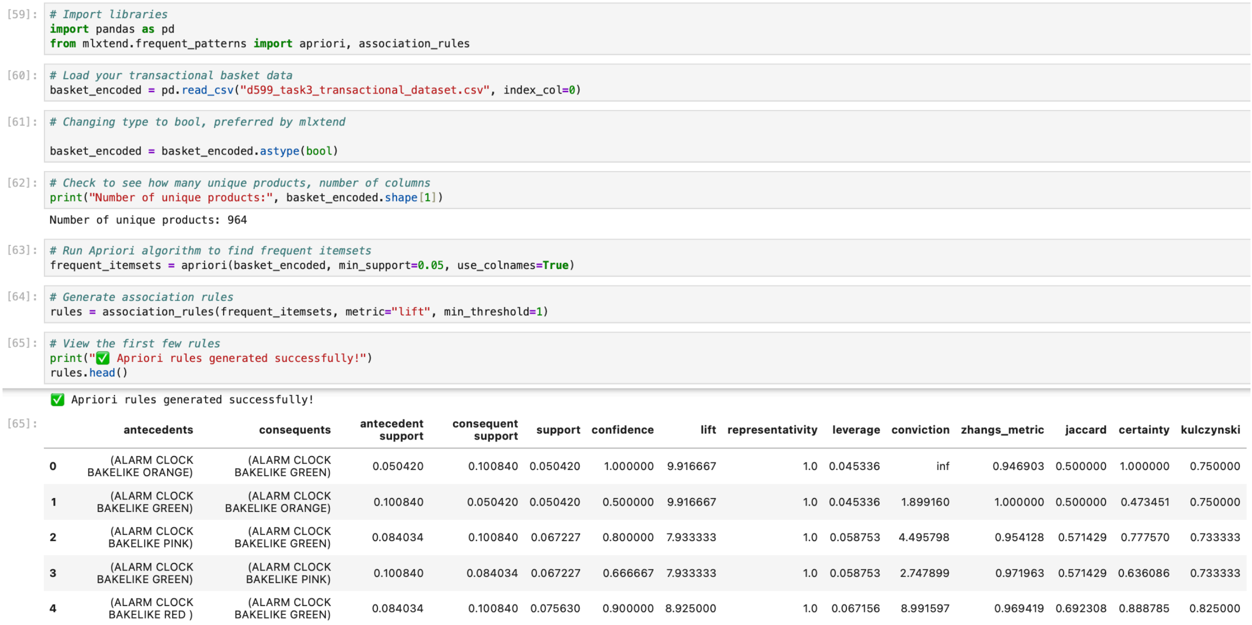
- **Filtering**: This step uses filtering for the target groups, narrowing down to corporate customers in the Northeast region. This aligns directly with the research question and improves the relevance of the findings.
- **Transactionalization**: This step transforms the filtered data into a binary matrix. This matrix is then saved to ensure any future use of the dataset can be used and repurposed.

2. Include a copy of the cleaned dataset.

   You can see the encoded dataset in the attached file **d599_task3_encoded_dataset.csv**.

   You can see the transactional matrix in the attached file **d599_task3_transactional_dataset.csv**.

3. Execute the code used to generate association rules with the Apriori algorithm. Provide a screenshot that demonstrates that the code is error-free.



```
[59]: # Import libraries
      import pandas as pd
      from mlxtend.frequent_patterns import apriori, association_rules
```

```
[60]: # Load your transactional basket data
      basket_encoded = pd.read_csv("d599_task3_transactional_dataset.csv", index_col=0)
```

```
[61]: # Changing type to bool, preferred by mlxtend

      basket_encoded = basket_encoded.astype(bool)
```

```
[62]: # Check to see how many unique products, number of columns
      print("Number of unique products:", basket_encoded.shape[1])

      Number of unique products: 964
```

```
[63]: # Run Apriori algorithm to find frequent itemsets
      frequent_itemsets = apriori(basket_encoded, min_support=0.05, use_colnames=True)
```

```
[64]: # Generate association rules
      rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
```

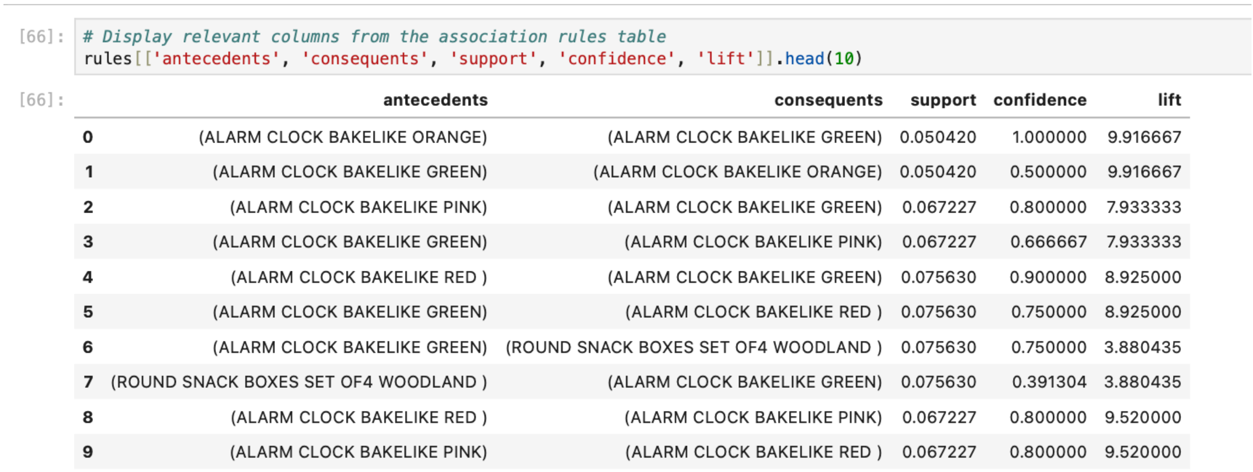
```
[65]: # View the first few rules
      print("✅ Apriori rules generated successfully!")
      rules.head()

      ✅ Apriori rules generated successfully!
```

[65]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	representativity	leverage	conviction	zhangs_metric	jaccard	certainty	kulczynski
0	(ALARM CLOCK BAKELIKE ORANGE)	(ALARM CLOCK BAKELIKE GREEN)	0.050420	0.100840	0.050420	1.000000	9.916667	1.0	0.045336	inf	0.946903	0.500000	1.000000	0.750000
1	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE ORANGE)	0.100840	0.050420	0.050420	0.500000	9.916667	1.0	0.045336	1.899160	1.000000	0.500000	0.473451	0.750000
2	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE GREEN)	0.084034	0.100840	0.067227	0.800000	7.933333	1.0	0.058753	4.495798	0.954128	0.571429	0.777570	0.733333
3	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE PINK)	0.100840	0.084034	0.067227	0.666667	7.933333	1.0	0.058753	2.747899	0.971963	0.571429	0.636086	0.733333
4	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.084034	0.100840	0.075630	0.900000	8.925000	1.0	0.067156	8.991597	0.969419	0.692308	0.888785	0.825000

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



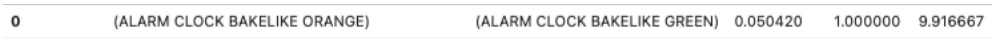
```
[66]: # Display relevant columns from the association rules table
      rules[['antecedents', 'consequents', 'support', 'confidence', 'lift']].head(10)
```

[66]:

	antecedents	consequents	support	confidence	lift
0	(ALARM CLOCK BAKELIKE ORANGE)	(ALARM CLOCK BAKELIKE GREEN)	0.050420	1.000000	9.916667
1	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE ORANGE)	0.050420	0.500000	9.916667
2	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE GREEN)	0.067227	0.800000	7.933333
3	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE PINK)	0.067227	0.666667	7.933333
4	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.075630	0.900000	8.925000
5	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.075630	0.750000	8.925000
6	(ALARM CLOCK BAKELIKE GREEN)	(ROUND SNACK BOXES SET OF4 WOODLAND)	0.075630	0.750000	3.880435
7	(ROUND SNACK BOXES SET OF4 WOODLAND)	(ALARM CLOCK BAKELIKE GREEN)	0.075630	0.391304	3.880435
8	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE PINK)	0.067227	0.800000	9.520000
9	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE RED)	0.067227	0.800000	9.520000

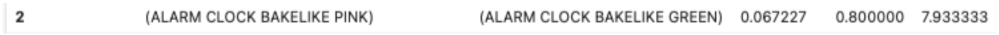
5. Explain the top **three** relevant rules generated by the Apriori algorithm. Include a screenshot of the top **three** relevant rules.

   Example #1 –
   If a customer buys an ALARM CLOCK BAKELIKE ORANGE, they also buy an ALARM CLOCK BAKELINE GREEN
   - Support: 5.04%
   - Confidence: 100%
   - Lift: 9.92



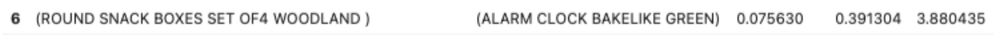
0	(ALARM CLOCK BAKELIKE ORANGE)	(ALARM CLOCK BAKELIKE GREEN)	0.050420	1.000000	9.916667

   Example #3 –
   If a customer buys an ALARM CLOCK BAKELIKE PINK, they also buy an ALARM CLOCK BAKELIKE GREEN
   - Support: 6.72%
   - Confidence: 80%
   - Lift: 7.93



2	(ALARM CLOCK BAKELIKE PINK)	(ALARM CLOCK BAKELIKE GREEN)	0.067227	0.800000	7.933333

   Example #2 –
   If a customer buys ROUND SNACK BOXES SET OF4 WOODLAND, they also buy ALARM CLOCK BAKELIKE GREEN
   - Support: 7.56%
   - Confidence: 39.1%
   - Lift: 3.88



6	(ROUND SNACK BOXES SET OF4 WOODLAND)	(ALARM CLOCK BAKELIKE GREEN)	0.075630	0.391304	3.880435

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.

   Based on the results from the association rules table, we can draw conclusions when evaluating Support, Confidence, and Lift:

   Support represents how frequently the item combination appears in the dataset. As an example, *ALARM CLOCK BAKELIKE ORANGE* and *ALARM CLOCK BAKELINE GREEN* had a support of 5.04%, meaning that this combination appeared in about 5 out of 100 transactions. The higher the support, the more relevant this association rule is to other customers.

   Confidence represents the likelihood that the consequent is purchased when the antecedent is purchased. In this same example, the confidence was 100%, meaning every customer who bought *ALARM CLOCK BAKELIKE ORANGE* also bought *ALARM CLOCK BAKELINE GREEN*. This can help us conclude that there's a strong relationship between these two products, and there might be value in bundling or cross-promoting this combination.

   Lift represents the strength of the rule compared to random chance. A lift greater than 1 indicates there's a positive correlation. The examples we've chosen had lift values well above 7, one of them being 9.92, meaning the likelihood of the consequent being purchased is nearly 10 times higher than it would be by random chance. A high lift value suggests strong actionable patterns in a customer's purchasing behavior.

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

   The practical significance of these Market Basket Analysis findings is directly related to the marketing, merchandising, and sales strategy for Allias Megastore. The rules we've created and analyzed provide insights that suggest Allias Megastore to do the following:

   - Create product bundles that package related items (e.g., clock sets or color assortments), increasing average transaction value.
   - Use "Frequently Bought Together" suggestions (via an e-commerce platform, email marketing campaigns, or physical and digital advertisements) to encourage multi-item purchases
   - Optimize in-store or online product placements (equip new merchandising strategies) by placing frequently co-purchased items near each other for convenience and accessibility.

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

   Based on our research question, *"What product combinations are most frequently purchased together by corporate customers in the Northeast region?"*, we recommend Allias Megastore to implement targeted bundling strategies for corporate customers in the Northeast region.

1. Bundle high-confidence product pairs and offer them at a slight discount to encourage multi-item purchases.
2. Highlight these product pairings using a “frequently bought together” feature in various digital and physical marketing campaigns.
3. Explore in-store or online product placements using merchandising techniques for convenient and optimal shopping.

Implementing these recommendations can improve average transaction value (order size), support overall customer experience, and increase marketing performance for corporate clients in a high-value region like the Northeast.

E. **Panopto Demonstration** - https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=7e6425e5-84fc-444f-8f23-b31b00d6f5d4

Sources

F. Beyond WGU resources, I’ve used the following website to further expand my knowledge on the *pandas* library with respect to `.map()`:
   a. https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.applymap.html