Task 3:

Market Basket Analysis

Keniyah Chestnut

Data Preparation and Exploration — D599

SID:012601305

Part A: Research Question

A1: Proposed Question

What product combinations are most frequently purchased together by customers in different regions?

This question will help Allias Megastore identify popular product pairings to offer targeted promotions, create product bundles, and optimize store layouts to increase sales.

A2: Goal of Data Analysis

The goal of this analysis is to find out which products customers often buy together so the company can make better decisions about marketing and merchandising.

By understanding these patterns, Allias Megastore can offer smarter promotions, suggest related products, and create special bundles to help increase sales and keep customers happy.

Part B: Market Basket Justification

B1: Why Market Basket Analysis is Used

Market basket analysis helps identify relationships between products frequently purchased together.

In this project, it will help reveal patterns and associations between products that customers tend to buy at the same time.

Expected outcomes include discovering common product pairings that can be used to drive marketing, product placement, and promotional strategies.

B2: Example of a Transaction

For this example, I will use the following transaction from the dataset:

Order ID: 536370

OrderID	ProductNa	Quantity	InvoiceDa	UnitPrice	TotalCost	Country	DiscountA	OrderPrio	Region	Segment	Expedited	PaymentM	CustomerOrderSatisfaction
536370	INFLATABI	48	########	$0.85	$40.80	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	SET2 RED I	18	########	$2.95	$53.10	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	PANDA AN	12	########	$0.85	$10.20	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	RED TOAD	24	########	$1.65	$39.60	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	VINTAGE H	24	########	$1.25	$30.00	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	STARS GIF	24	########	$0.65	$15.60	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	VINTAGE S	12	########	$3.75	$45.00	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	ROUND SN	24	########	$2.95	$70.80	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	MINI PAIN	36	########	$0.65	$23.40	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	MINI JIGSA	24	########	$0.42	$10.08	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	MINI JIGSA	24	########	$0.42	$10.08	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	SPACEBOY	24	########	$1.95	$46.80	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	CIRCUS PA	24	########	$1.95	$46.80	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	LUNCH BC	24	########	$1.95	$46.80	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	CHARLOTT	20	########	$0.85	$17.00	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	ALARM CL	12	########	$3.75	$45.00	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	ALARM CL	24	########	$3.75	$90.00	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	ALARM CL	24	########	$3.75	$90.00	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied
536370	SET 2 TEA	24	########	$2.95	$70.80	United Sta	Yes	High	Northeast	Corporate	Yes	Credit Car	Satisfied

This order shows a single purchase made by a customer. The record tracks which items were bought together, along with important details such as the payment method, order priority, shipping option, and whether any discounts were applied.

In this specific transaction, the customer purchased multiple products together, which makes it useful for analyzing buying patterns. Transactions like this help market basket analysis identify groups of products that are commonly purchased at the same time.

B3: Assumption of Market Basket Analysis

The main assumption in market basket analysis is that past purchase patterns can help predict future buying behavior.

While it is understood that customer preferences and trends can change over time, it is assumed that connections found between products in historical data can still provide valuable insights.

By identifying which products were frequently purchased together in the past, businesses can make informed guesses about what customers may continue to buy together moving forward..

Part C: Data Preparation and Analysis

C1a: Select Categorical Variables

An ordinal variable is a type of categorical variable that has a meaningful order or ranking.

The first ordinal variable I selected is Order Priority. This can be ranked from "Low" to "Critical," meaning one level is higher or more urgent than another.

The second ordinal variable is Expedited Shipping. Although the responses are simply "Yes" or "No," they can also be thought of as "Fast" and "Standard" shipping, which can clearly be ordered by priority.

For the nominal variables, which are categorical but do not have any natural order, I selected:

- Segment, which includes categories like Corporate and Consumer.

- Payment Method, which has options such as PayPal and Credit Card.

Nominal variables help categorize data without implying any ranking between the categories.

```
4                    Satisfied
```

```
ordinal_vars = ['OrderPriority', 'ExpeditedShipping']
nominal_vars = ['Segment', 'PaymentMethod']

for col in ordinal_vars + nominal_vars:
    print(f"{col} unique values:", df[col].unique())
```

```
OrderPriority unique values: ['High' 'Medium']
ExpeditedShipping unique values: ['Yes' 'No']
Segment unique values: ['Corporate' 'Consumer']
PaymentMethod unique values: ['Credit Card' 'PayPal']
```

For C1d: Explanation and Justification of Steps, I will outline the code used for encoding and transactionalizing the data. I will also explain and justify each step as it is performed.

C1b: Encode Variables

The data needed to be encoded before it could be used with the Apriori algorithm, which requires numerical input. Since the algorithm cannot interpret raw text, we had to convert our categorical variables into a numerical format.

Because there were only two variables for both the ordinal and nominal categories, encoding was fairly straightforward.

For Order Priority, I treated the entries as levels of urgency. Since this is an ordinal variable, the values have a natural ranking from lowest to highest priority.

I encoded them numerically in the following order:

"Low" → 1

"Medium" → 2

"High" → 3

"Critical" → 4

This ordinal mapping preserves the ranking structure so that the algorithm can interpret "Critical" as more urgent than "High", and so on. It allows for proper analysis where the level of priority may impact purchasing behavior or urgency.

For Expedited Shipping, I treated "No" as slow shipping and "Yes" as fast shipping, encoding them as 0 and 1, respectively.

```
3]:
# ------------------------
# Step 3 - Encode Variables
# ------------------------
# Encoding ordinal variables
df['ExpeditedShipping'] = df['ExpeditedShipping'].map({'Yes': 1, 'No': 0})

order_priority_mapping = {'Low': 1, 'Medium': 2, 'High': 3, 'Critical': 4}
df['OrderPriority'] = df['OrderPriority'].map(order_priority_mapping)
```

For the nominal variables, both Segment and Payment Method were one-hot encoded. This step was necessary because nominal data does not have a natural ranking. Each category was converted into its own column with a 0 or 1 value, which allowed the Apriori algorithm to process the data correctly.

```
[31]: # One-hot encode nominal variables
df_nominal_encoded = pd.get_dummies(df[nominal_vars])

# Combine ordinal + nominal
df_encoded = pd.concat([df[['OrderID']], df_nominal_encoded, df['ExpeditedShipping'], df['OrderPriority']], axis=1)

print("\nEncoded Data Preview:")
print(df_encoded.head())
```

```
Encoded Data Preview:
   OrderID  Segment_Consumer  Segment_Corporate  PaymentMethod_Credit Card  \
0   536370             False               True                       True
1   536370             False               True                       True
2   536370             False               True                       True
3   536370             False               True                       True
4   536370             False               True                       True

   PaymentMethod_PayPal  ExpeditedShipping  OrderPriority
0                 False                NaN            NaN
1                 False                NaN            NaN
2                 False                NaN            NaN
3                 False                NaN            NaN
4                 False                NaN            NaN
```

C1c: Transactionalize Data

Here is the code and result I used to transactionalize the data. This process turns the product names into True and False values, which indicate whether or not each product was purchased in a particular order.

This step is necessary because the Apriori algorithm can only analyze data in this format.

First, I grouped the products by Order ID so that all products purchased in the same transaction were combined together.

Then, I used TransactionEncoder to one-hot encode the product lists. This transformed the data into True/False values, showing whether each product was present in each transaction.

```
# -------------------------
# Step 4 - Transactionalize Products
# -------------------------
# Grouping product names by OrderID
basket = df.groupby(['OrderID'])['ProductName'].apply(list)

# Convert to True/False format
te = TransactionEncoder()
te_data = te.fit(basket).transform(basket)

df_products = pd.DataFrame(te_data, columns=te.columns_)
```

The result is a dataset where every row represents a transaction (Order ID), and every column represents a product. Each cell tells us whether the product was purchased (True) or not (False) in that order.

```python
print("\nTransactional Product Data:")
print(df_products.head())
```

```
Transactional Product Data:
   50S CHRISTMAS GIFT BAG LARGE   DOLLY GIRL BEAKER  \
0                         False               False
1                         False               False
2                         False               False
3                         False               False
4                         False               False

   I LOVE LONDON MINI BACKPACK   NINE DRAWER OFFICE TIDY  \
0                        False                     False
1                        False                     False
2                        False                     False
3                        False                     False
4                        False                     False

   SET 2 TEA TOWELS I LOVE LONDON    SPACEBOY BABY GIFT SET  \
0                            False                    False
1                            False                    False
2                            False                    False
3                            False                    False
4                            False                    False

   TRELLIS COAT RACK  10 COLOUR SPACEBOY PEN  12 COLOURED PARTY BALLOONS  \
0              False                   False                       False
1              False                   False                       False
2              False                   False                       False
3              False                   False                       False
4              False                   False                       False

  12 EGG HOUSE PAINTED WOOD  ...  WRAP VINTAGE PETALS  DESIGN  \
0                     False  ...                        False
1                     False  ...                        False
2                     False  ...                        False
3                     False  ...                        False
4                     False  ...                        False

  YELLOW COAT RACK PARIS FASHION  YELLOW GIANT GARDEN THERMOMETER  \
0                          False                            False
1                          False                            False
2                          False                            False
3                          False                            False
4                          False                            False

  YELLOW SHARK HELICOPTER  ZINC  STAR TLIGHT HOLDER   \
0                   False                     False
1                   False                     False
2                   False                     False
3                   False                     False
4                   False                     False

  ZINC FOLKART SLEIGH BELLS  ZINC HERB GARDEN CONTAINER  \
0                     False                       False
1                     False                       False
2                     False                       False
3                     False                       False
4                     False                       False

  ZINC METAL HEART DECORATION  ZINC TLIGHT HOLDER STAR LARGE  \
0                       False                          False
1                       False                          False
2                       False                          False
3                       False                          False
4                       False                          False

  ZINC TLIGHT HOLDER STARS SMALL
0                          False
1                          False
2                          False
3                          False
4                          False

[5 rows x 1562 columns]
```

C1d: Explanation and Justification of Steps

Each step was performed for a reason:

Selecting ordinal and nominal variables helped to include useful information about customer and order attributes.

Encoding converted text-based categories into numerical values that the algorithm could process.

Transactionalizing product data prepared the dataset in the True/False format needed for market basket analysis.

By preparing the data this way, the Apriori algorithm could analyze product combinations effectively.

C2: Cleaned Dataset

To meet submission guidelines and prepare for analysis, I created two cleaned datasets:

Combined Dataset:

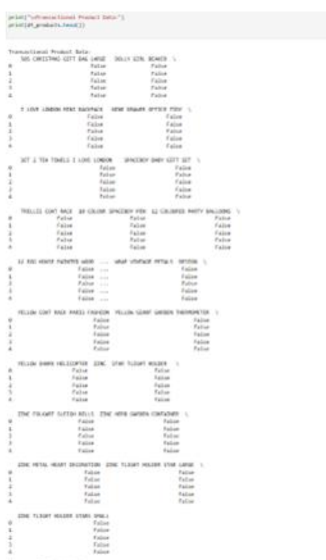
```
print("\nEncoded Data Preview:")
print(df_encoded.head())
```

```
Encoded Data Preview:
   OrderID  Segment_Consumer  Segment_Corporate  PaymentMethod_Credit Card  \
0   536370             False               True                       True
1   536370             False               True                       True
2   536370             False               True                       True
3   536370             False               True                       True
4   536370             False               True                       True

   PaymentMethod_PayPal  ExpeditedShipping  OrderPriority
0                 False                NaN            NaN
1                 False                NaN            NaN
2                 False                NaN            NaN
3                 False                NaN            NaN
4                 False                NaN            NaN
```

This dataset includes transactional product data and the encoded ordinal and nominal variables. This version was created to meet WGU submission requirements.

Product-Only Dataset:



This dataset includes only the transactional product data. This version was used to run the Apriori algorithm, as it focuses only on product combinations.

C3: Execute Code

The Apriori algorithm was run on the product-only transactional dataset.

The code executed successfully without any errors, and the association rules were generated.

```python
# ------------------------
# Run Apriori on Product-Only dataset
# ------------------------
frequent_itemsets = apriori(df_products, min_support=0.01, use_colnames=True)

rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1)
```

```python
# Show top rules
print("\nTop Association Rules:")
print(rules[['antecedents', 'consequents', 'support', 'confidence', 'lift']].sort_values(by='lift', ascending=False).head(10))
```

```
Top Association Rules:
                                             antecedents  \
75214  (CHILDRENS CUTLERY SPACEBOY , ALARM CLOCK BAKE...
81555  (PLASTERS IN TIN CIRCUS PARADE , LUNCH BOX WIT...
76767  (PLASTERS IN TIN SPACEBOY, ROUND SNACK BOXES S...
67930  (ALARM CLOCK BAKELIKE GREEN, PLASTERS IN TIN C...
82729  (PLASTERS IN TIN CIRCUS PARADE , ALARM CLOCK B...
82726  (SET6 RED SPOTTY PAPER CUPS, PLASTERS IN TIN S...
67945  (PLASTERS IN TIN CIRCUS PARADE , ALARM CLOCK B...
82715  (PLASTERS IN TIN CIRCUS PARADE , SET6 RED SPOT...
81544  (PLASTERS IN TIN CIRCUS PARADE , ALARM CLOCK B...
67964  (ALARM CLOCK BAKELIKE IVORY, PLASTERS IN TIN S...

                                             consequents   support  \
75214  (CARD DOLLY GIRL , ALARM CLOCK BAKELIKE RED , ...  0.011338
81555  (SKULL LUNCH BOX WITH CUTLERY , PLASTERS IN TI...  0.011338
76767  (SET6 RED SPOTTY PAPER CUPS, PLASTERS IN TIN C...  0.011338
67930  (ALARM CLOCK BAKELIKE IVORY, PLASTERS IN TIN S...  0.011338
82729  (SET6 RED SPOTTY PAPER CUPS, PLASTERS IN TIN S...  0.011338
82726  (PLASTERS IN TIN CIRCUS PARADE , ALARM CLOCK B...  0.011338
67945  (ALARM CLOCK BAKELIKE IVORY, PLASTERS IN TIN S...  0.011338
82715  (PLASTERS IN TIN SPACEBOY, ALARM CLOCK BAKELIK...  0.011338
81544  (SKULL LUNCH BOX WITH CUTLERY , ALARM CLOCK BA...  0.011338
67964  (PLASTERS IN TIN CIRCUS PARADE , ALARM CLOCK B...  0.011338

       confidence  lift
75214         1.0  88.2
81555         1.0  88.2
76767         1.0  88.2
67930         1.0  88.2
82729         1.0  88.2
82726         1.0  88.2
67945         1.0  88.2
82715         1.0  88.2
81544         1.0  88.2
67964         1.0  88.2
```

C4: Support, Lift, and Confidence Values

the association rules generated by the Apriori algorithm include the following important metrics:

Support: This measures how frequently the item combination appears in the dataset. Higher support means the combination is common across many transactions.

Confidence: This shows the likelihood that the consequent (second product) will be purchased when the antecedent (first product) is purchased. A higher confidence means the rule is more reliable.

Lift: This compares how much more likely the consequent is to be purchased when the antecedent is purchased versus if the items were bought independently. A lift greater than 1 indicates a strong association between the products.

The screenshot below displays the Support, Confidence, and Lift values for the generated rules:

```
[72]: # Sort rules by lift
rules_sorted_by_lift = rules.sort_values(by='lift', ascending=False)

# Show rules table
print("\nRules Table (Support, Confidence, Lift):")
rules_sorted_by_lift[['antecedents', 'consequents', 'support', 'confidence', 'lift']].head(10)
```

```
Rules Table (Support, Confidence, Lift):
```

[72]:

	antecedents	consequents	support	confidence	lift
75214	(CHILDRENS CUTLERY SPACEBOY , ALARM CLOCK BAKE...	(CARD DOLLY GIRL , ALARM CLOCK BAKELIKE RED , ...	0.011338	1.0	88.2
81555	(PLASTERS IN TIN CIRCUS PARADE , LUNCH BOX WIT...	(SKULL LUNCH BOX WITH CUTLERY , PLASTERS IN TI...	0.011338	1.0	88.2
76767	(PLASTERS IN TIN SPACEBOY, ROUND SNACK BOXES S...	(SET6 RED SPOTTY PAPER CUPS, PLASTERS IN TIN C...	0.011338	1.0	88.2
67930	(ALARM CLOCK BAKELIKE GREEN, PLASTERS IN TIN C...	(ALARM CLOCK BAKELIKE IVORY, PLASTERS IN TIN S...	0.011338	1.0	88.2
82729	(PLASTERS IN TIN CIRCUS PARADE , ALARM CLOCK B...	(SET6 RED SPOTTY PAPER CUPS, PLASTERS IN TIN S...	0.011338	1.0	88.2
82726	(SET6 RED SPOTTY PAPER CUPS, PLASTERS IN TIN S...	(PLASTERS IN TIN CIRCUS PARADE , ALARM CLOCK B...	0.011338	1.0	88.2
67945	(PLASTERS IN TIN CIRCUS PARADE , ALARM CLOCK B...	(ALARM CLOCK BAKELIKE IVORY, PLASTERS IN TIN S...	0.011338	1.0	88.2
82715	(PLASTERS IN TIN CIRCUS PARADE , SET6 RED SPOT...	(PLASTERS IN TIN SPACEBOY, ALARM CLOCK BAKELIK...	0.011338	1.0	88.2
81544	(PLASTERS IN TIN CIRCUS PARADE , ALARM CLOCK B...	(SKULL LUNCH BOX WITH CUTLERY , ALARM CLOCK BA...	0.011338	1.0	88.2
67964	(ALARM CLOCK BAKELIKE IVORY, PLASTERS IN TIN S...	(PLASTERS IN TIN CIRCUS PARADE , ALARM CLOCK B...	0.011338	1.0	88.2

C5: Top 3 Relevant Rules

These rules reveal clear and meaningful product relationships. All three rules have a confidence of 1.0, meaning every time the items in the "If" section were purchased together, the associated items in the "Then" section were also purchased. The extremely high Lift values (88.2) suggest these items are purchased together far more often than by chance.

These insights can guide marketing strategies, promotional bundling, and in-store placement, helping Allias Megastore encourage additional sales and improve the shopping experience.

Top 3 Rules:

The top 3 rules were selected based on their highest Lift values, which indicate the strongest product associations. A higher Lift value means that when the first product(s) are purchased, the second product(s) are much more likely to be purchased as well.

Top 3 Rules:

Rule 1

If items:

CHILDRENS CUTLERY SPACEBOY

ALARM CLOCK BAKELIKE PINK

SPACEBOY BIRTHDAY CARD

Then items:

CARD DOLLY GIRL

ALARM CLOCK BAKELIKE RED

ROUND SNACK BOXES SET OF4 WOODLAND

Support: 0.0113

Confidence: 1.0

Lift: 88.2

Rule 2

If items:

PLASTERS IN TIN CIRCUS PARADE

LUNCH BOX WITH CUTLERY RETROSPOT

ALARM CLOCK BAKELIKE RED

ALARM CLOCK BAKELIKE GREEN

Then items:

SKULL LUNCH BOX WITH CUTLERY

PLASTERS IN TIN SPACEBOY

ALARM CLOCK BAKELIKE PINK

Support: 0.0113

Confidence: 1.0

Lift: 88.2

Rule 3

If items:

PLASTERS IN TIN SPACEBOY

ROUND SNACK BOXES SET OF4 WOODLAND

SET6 RED SPOTTY PAPER PLATES

Then items:

SET6 RED SPOTTY PAPER CUPS

PLASTERS IN TIN CIRCUS PARADE

ALARM CLOCK BAKELIKE RED

Support: 0.0113

Confidence: 1.0

Lift: 88.2

These rules reveal important patterns about which products are frequently purchased together. All of the rules have a Confidence of 1.0, meaning they always occur together when the antecedent items are purchased.

Additionally, the very high Lift values (88.2) indicate that these product combinations are extremely strong associations, which can be used to create effective marketing promotions or bundle offers.

```python
# ------------------------
#  Top 3 Relevant Rules
# ------------------------
top_rules = rules.sort_values(by='lift', ascending=False).head(3)

print("\nTop 3 Rules:")
for idx, rule in top_rules.iterrows():
    print(f"\nRule #{idx+1}")
    print("If items:", list(rule['antecedents']), "-> Then items:", list(rule['consequents']))
    print("Support:", rule['support'])
    print("Confidence:", rule['confidence'])
    print("Lift:", rule['lift'])
```

```
Top 3 Rules:

Rule #75215
If items: ['CHILDRENS CUTLERY SPACEBOY ', 'ALARM CLOCK BAKELIKE PINK', 'SPACEBOY BIRTHDAY CARD'] -> Then items: ['CARD DOLLY GIRL ', 'ALARM CLOCK BAKELI
KE RED ', 'ROUND SNACK BOXES SET OF4 WOODLAND ']
Support: 0.011337868480725623
Confidence: 1.0
Lift: 88.2

Rule #81556
If items: ['PLASTERS IN TIN CIRCUS PARADE ', 'LUNCH BOX WITH CUTLERY RETROSPOT ', 'ALARM CLOCK BAKELIKE RED ', 'ALARM CLOCK BAKELIKE GREEN'] -> Then ite
ms: ['SKULL LUNCH BOX WITH CUTLERY ', 'PLASTERS IN TIN SPACEBOY', 'ALARM CLOCK BAKELIKE PINK']
Support: 0.011337868480725623
Confidence: 1.0
Lift: 88.2

Rule #76768
If items: ['PLASTERS IN TIN SPACEBOY', 'ROUND SNACK BOXES SET OF4 WOODLAND ', 'SET6 RED SPOTTY PAPER PLATES'] -> Then items: ['SET6 RED SPOTTY PAPER CUP
S', 'PLASTERS IN TIN CIRCUS PARADE ', 'ALARM CLOCK BAKELIKE RED ']
Support: 0.011337868480725623
Confidence: 1.0
Lift: 88.2
```

These rules show which product combinations are the most relevant and useful for marketing, promotions, and store placement.

References

No external sources were used beyond WGU course materials.