Project 3: Market Basket analysis

ELVESSA TATUM

Background info

- Apriori is an algorithm used to find frequent itemsets and generate association rules in a dataset, often used for Market Basket Analysis.
- The core idea is that frequent patterns (item combinations) can be expanded one item at a time, and all subsets of a frequent itemset must also be frequent.
- The 2 basic Operations of Apriori are the Join step and the Prune step
- Join Step — Join sets of items together to create larger candidate sets.
- Prune Step — Eliminate candidate sets that contain infrequent subsets.
- The Apriori principle states the following:
- If an itemset is frequent, then all of its subsets must also be frequent
- If an itemset is not frequent, then all of its supersets cannot be frequent
- The support of an itemset never exceeds the support of its subsets
- Overfitting can be reduced by setting appropriate minimum support and confidence thresholds, and by filtering rules for relevance (e.g., high lift values), rather than keeping all possible rules.
- The benefits are that Apriori is an easy to understand algorithm, and join and prune steps are easy to implement on large itemsets in large databases
- Some issues with Apriori is it requires high computation if the itemsets are very large and the minimum support is kept very low. Also, the entire database needs to be scanned.

Our Purpose and Dataset information

- Today, I plan on using the UCI Online Retail Dataset to discover patterns in customer purchasing behavior through Market Basket Analysis. This dataset contains transactional data from an online retail store registered in the United Kingdom, covering purchases made between December 1, 2010 and December 9, 2011. Each observation represents an item purchased in a specific invoice, allowing for detailed tracking of buying habits, returns, and overall spending behavior.
- This dataset was chosen because it is relatively simple, with only six main features (these can be seen to the right), yet large enough to provide rich insights due to its thousands of recorded transactions. Its structure makes it ideal for applying association rule mining without overwhelming complexity.
- The primary goal of this project is to find frequent item combinations that customers often purchase together and generate meaningful association rules. To achieve this, I will use the Apriori algorithm, which is well-suited for this type of analysis because it systematically finds frequent patterns while minimizing computational cost through smart pruning strategies. I chose Apriori over other models like FP-Growth because it is easier to interpret and provides a clear view of the frequency relationships between items. Ultimately, these findings could help improve product placement strategies, bundling, and targeted marketing.

InvoiceNo	ID	Categorical	a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation		no
StockCode	ID	Categorical	a 5-digit integral number uniquely assigned to each distinct product		no
Description	Feature	Categorical	product name		no
Quantity	Feature	Integer	the quantities of each product (item) per transaction		no
InvoiceDate	Feature	Date	the day and time when each transaction was generated		no
UnitPrice	Feature	Continuous	product price per unit	sterling	no
CustomerID	Feature	Categorical	a 5-digit integral number uniquely assigned to each customer		no
Country	Feature	Categorical	the name of the country where each customer resides		no

Importing Data

LET'S TAKE A LOOK AT OUR DATA FROM A GLANCE

[7]:
```
import pandas as pd

retail_data = pd.read_csv('Online Retail.csv')
retail_data.head()
```

[7]:

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

Importing data

- Let's use describe and info, and check for missing values

```
# checking missing values
missing = retail_data.isnull().sum()
missing = missing[missing > 0].sort_values(ascending=False)
print(missing)
```

```
CustomerID     135080
Description      1454
dtype: int64
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype  
---  ------       --------------   -----  
 0   InvoiceNo    541909 non-null  object 
 1   StockCode    541909 non-null  object 
 2   Description  540455 non-null  object 
 3   Quantity     541909 non-null  int64  
 4   InvoiceDate  541909 non-null  object 
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object 
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

	Quantity	UnitPrice	CustomerID
count	541909.000000	541909.000000	406829.000000
mean	9.552250	4.611114	15287.690570
std	218.081158	96.759853	1713.600303
min	-80995.000000	-11062.060000	12346.000000
25%	1.000000	1.250000	13953.000000
50%	3.000000	2.080000	15152.000000
75%	10.000000	4.130000	16791.000000
max	80995.000000	38970.000000	18287.000000

Data pre-processing

- Let's do some pre-processing by getting rid of missing values and canceled transactions

```
# drop missing InvoiceNo or Description
retail_data.dropna(subset=['InvoiceNo', 'Description'], inplace=True)

# remove canceled transactions (InvoiceNo starting with 'C')
retail_data = retail_data[~retail_data['InvoiceNo'].astype(str).str.startswith('C')]
```

Data pre-processing continued

- Let's continue pre-processing by creating itemsets.

```
# create the basket
basket = retail_data.groupby(['InvoiceNo', 'Description'])['Quantity'].sum().unstack().fillna(0)

# convert to 1/0
basket = (basket > 0).astype(bool)  # use bool type for better performance
```

Data exploration and analysis

- Let's make some association rules and filter for high confidence and lift

```python
from mlxtend.frequent_patterns import apriori, association_rules
# apply Apriori algorithm
frequent_itemsets = apriori(basket, min_support=0.01, use_colnames=True)

# generate association rules
rules = association_rules(frequent_itemsets, metric="lift", min_threshold=1.0)

# filter for high-confidence rules
filtered_rules = rules[(rules['confidence'] >= 0.9) & (rules['lift'] >= 1.2)].copy()

filtered_rules['antecedents'] = filtered_rules['antecedents'].apply(lambda x: ', '.join(list(x)))
filtered_rules['consequents'] = filtered_rules['consequents'].apply(lambda x: ', '.join(list(x)))

# print clean results
print(filtered_rules[['antecedents', 'consequents', 'support', 'confidence', 'lift']])
```

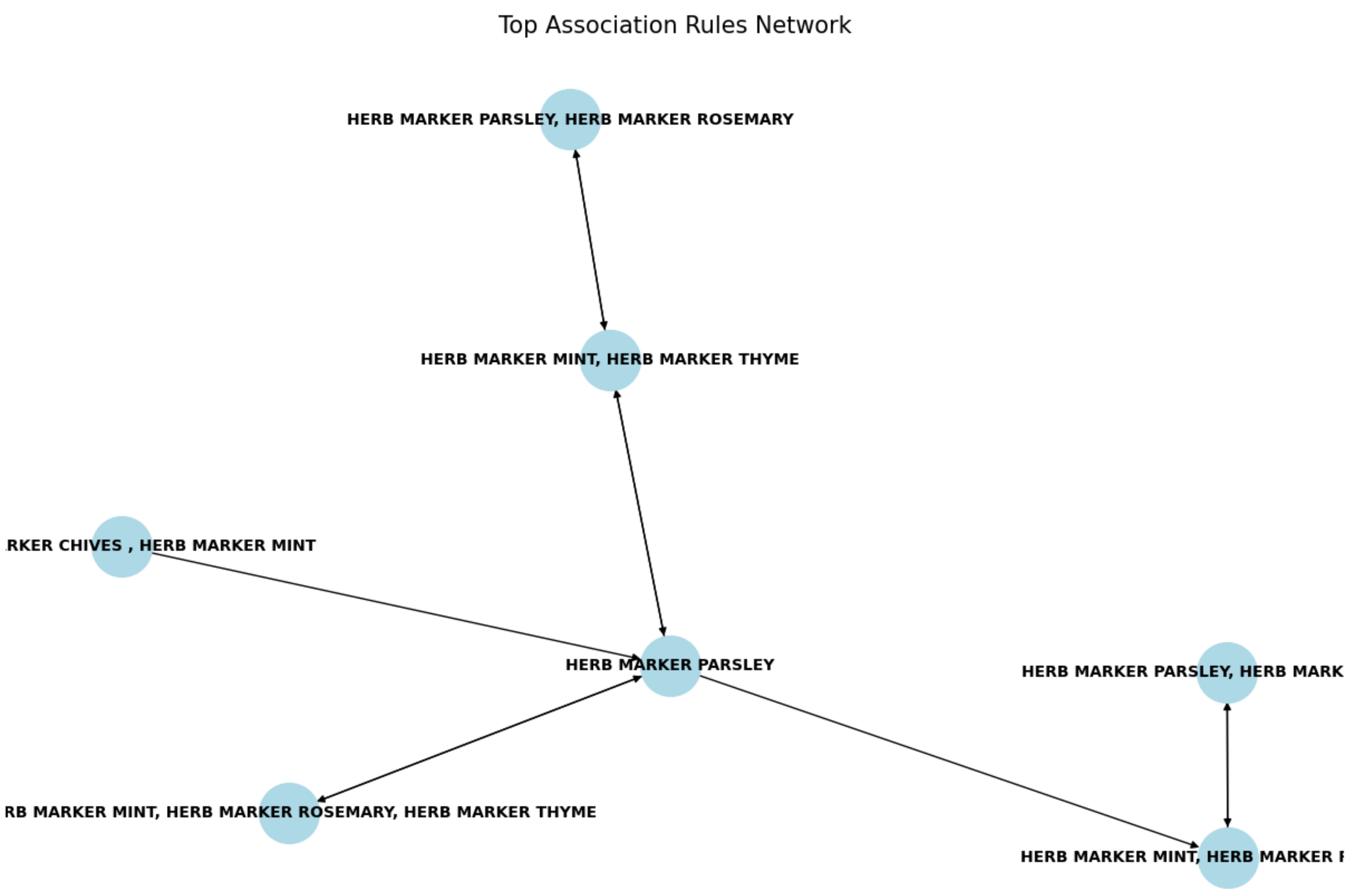
Data exploration and analysis continued

- Let's take a look at the itemsets created.

	antecedents	consequents	support	confidence	lift
296	COFFEE MUG PEARS DESIGN	COFFEE MUG APPLES DESIGN	0.0100	0.952381	63.492063
583	HERB MARKER BASIL	HERB MARKER MINT	0.0105	0.913043	76.086957
584	HERB MARKER PARSLEY	HERB MARKER BASIL	0.0100	0.909091	79.051383
588	HERB MARKER BASIL	HERB MARKER THYME	0.0105	0.913043	76.086957
590	HERB MARKER CHIVES	HERB MARKER MINT	0.0100	0.909091	75.757576
...	...	...	...	...	...
3567	HERB MARKER PARSLEY, HERB MARKER ROSEMARY	HERB MARKER MINT, HERB MARKER THYME	0.0100	1.000000	95.238095
3568	HERB MARKER PARSLEY, HERB MARKER THYME	HERB MARKER MINT, HERB MARKER ROSEMARY	0.0100	0.952381	95.238095
3571	HERB MARKER PARSLEY	HERB MARKER MINT, HERB MARKER ROSEMARY, HERB M...	0.0100	0.909091	90.909091
3603	JUMBO BAG WOODLAND ANIMALS, JUMBO BAG APPLES, ...	JUMBO BAG RED RETROSPOT	0.0100	0.909091	9.620010
3673	JUMBO BAG WOODLAND ANIMALS, JUMBO SHOPPER VINT...	JUMBO BAG RED RETROSPOT	0.0100	0.909091	9.620010

Data exploration and analysis continued

LET'S TAKE A LOOK AT THE BEST ASSOCIATION RULES NETWORK



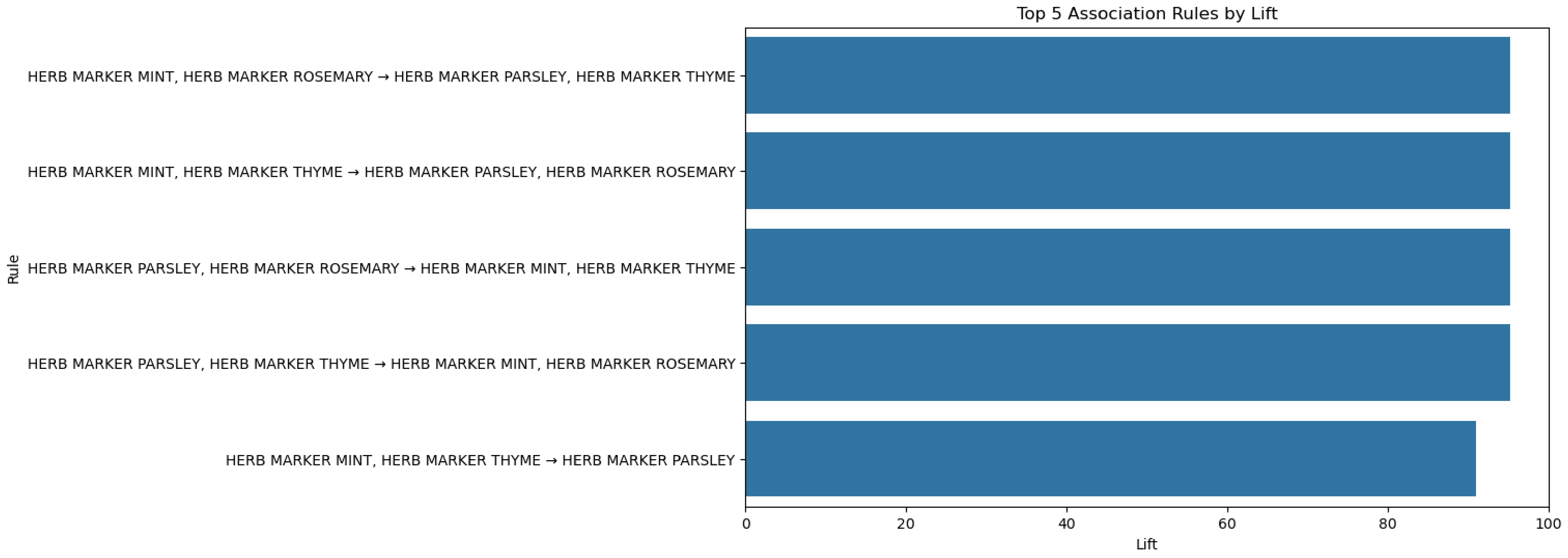
Data exploration and analysis continued

FINALLY, LET'S LOOK AT THE TOP 10 AND TOP 5 RULES BY LIFT

Top 10 Rules by Lift:

	antecedents	consequents	support	confidence	lift
3565	HERB MARKER MINT, HERB MARKER ROSEMARY	HERB MARKER PARSLEY, HERB MARKER THYME	0.0100	1.000000	95.238095
3566	HERB MARKER MINT, HERB MARKER THYME	HERB MARKER PARSLEY, HERB MARKER ROSEMARY	0.0100	0.952381	95.238095
3567	HERB MARKER PARSLEY, HERB MARKER ROSEMARY	HERB MARKER MINT, HERB MARKER THYME	0.0100	1.000000	95.238095
3568	HERB MARKER PARSLEY, HERB MARKER THYME	HERB MARKER MINT, HERB MARKER ROSEMARY	0.0100	0.952381	95.238095
2361	HERB MARKER MINT, HERB MARKER THYME	HERB MARKER PARSLEY	0.0105	1.000000	90.909091
3562	HERB MARKER MINT, HERB MARKER ROSEMARY, HERB M...	HERB MARKER PARSLEY	0.0100	1.000000	90.909091
3571	HERB MARKER PARSLEY	HERB MARKER MINT, HERB MARKER ROSEMARY, HERB M...	0.0100	0.909091	90.909091
2364	HERB MARKER PARSLEY	HERB MARKER MINT, HERB MARKER THYME	0.0105	0.954545	90.909091
2358	HERB MARKER PARSLEY	HERB MARKER MINT, HERB MARKER ROSEMARY	0.0100	0.909091	90.909091
2348	HERB MARKER CHIVES , HERB MARKER MINT	HERB MARKER PARSLEY	0.0100	1.000000	90.909091

Data exploration and analysis continued



Conclusion

- The analysis revealed that a significant number of the strongest association rules centered around herb marker products such as "HERB MARKER MINT," "HERB MARKER PARSLEY," "HERB MARKER ROSEMARY," and "HERB MARKER THYME." These items frequently appeared together in transactions with exceptionally high confidence and lift values, suggesting that customers who purchase one herb marker are extremely likely to purchase others as well. For example, some item combinations exhibited a confidence of 100% and a lift of over 90, indicating very strong and non-random associations.
- These findings are valuable for business applications. Retailers could strategically bundle these herb markers together or offer targeted promotions to encourage bulk purchases, maximizing revenue per customer. Additionally, these insights could guide inventory management by highlighting which items are often purchased together, allowing for better stock planning.
- Overall, the Apriori algorithm provided clear, actionable insights into customer buying patterns, demonstrating its strength as a tool for market basket analysis despite challenges such as memory demands and the need for careful parameter tuning to avoid overfitting.