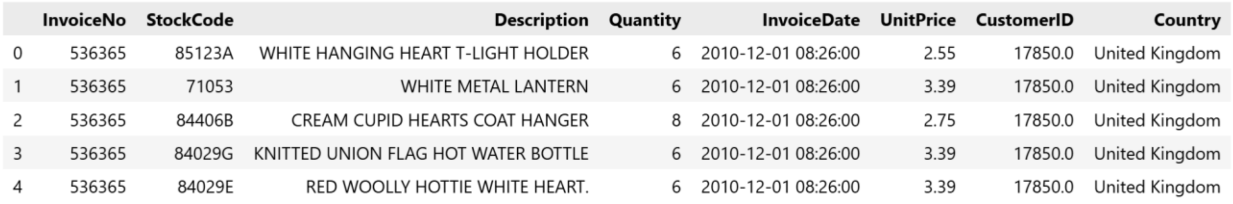
Lab 6: Association Rule Mining with Apriori and FP-Growth

Name: Sudhansu Sekhar Dash

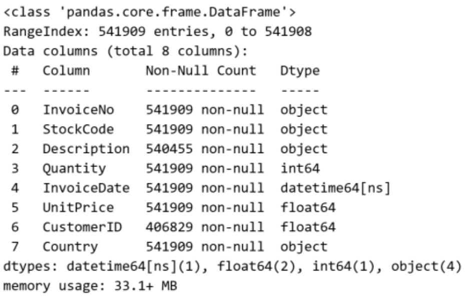
UC ID: 005033968

This lab focused on understanding how association rule mining can be used to discover meaningful relationships between items in transactional data. Using the Online Retail dataset, the lab demonstrated how two popular algorithms Apriori and FP-Growth identify frequent itemsets and generate rules that explain which items are commonly purchased together. The overall goal was to translate raw transaction records into actionable insights that could support retail decision-making such as product placement, bundling, and targeted marketing.

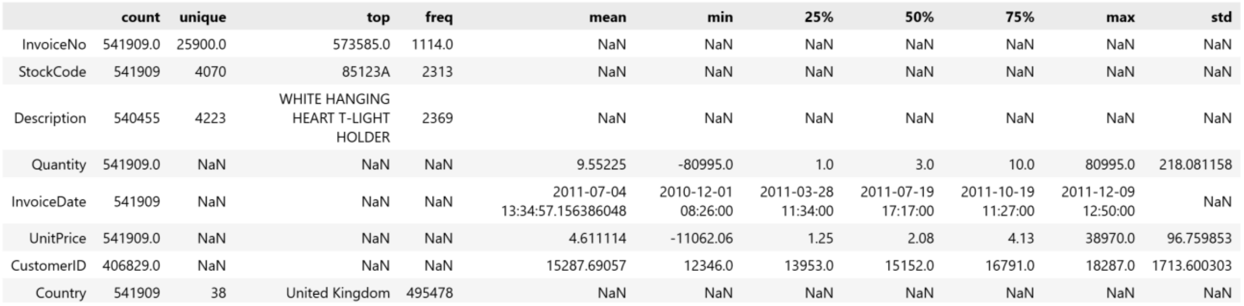
Data preparation and Exploration: The Online Retail dataset was first loaded and examined to understand its structure and content. Missing values were removed, negative quantities were filtered out, and item descriptions were standardized to ensure the data accurately represented real purchases. Basic visualizations such as bar charts were used to display the most frequently purchased products, giving an initial sense of customer buying behavior. A co-occurrence heatmap further illustrated item relationships by showing how often products appeared together in transactions.



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

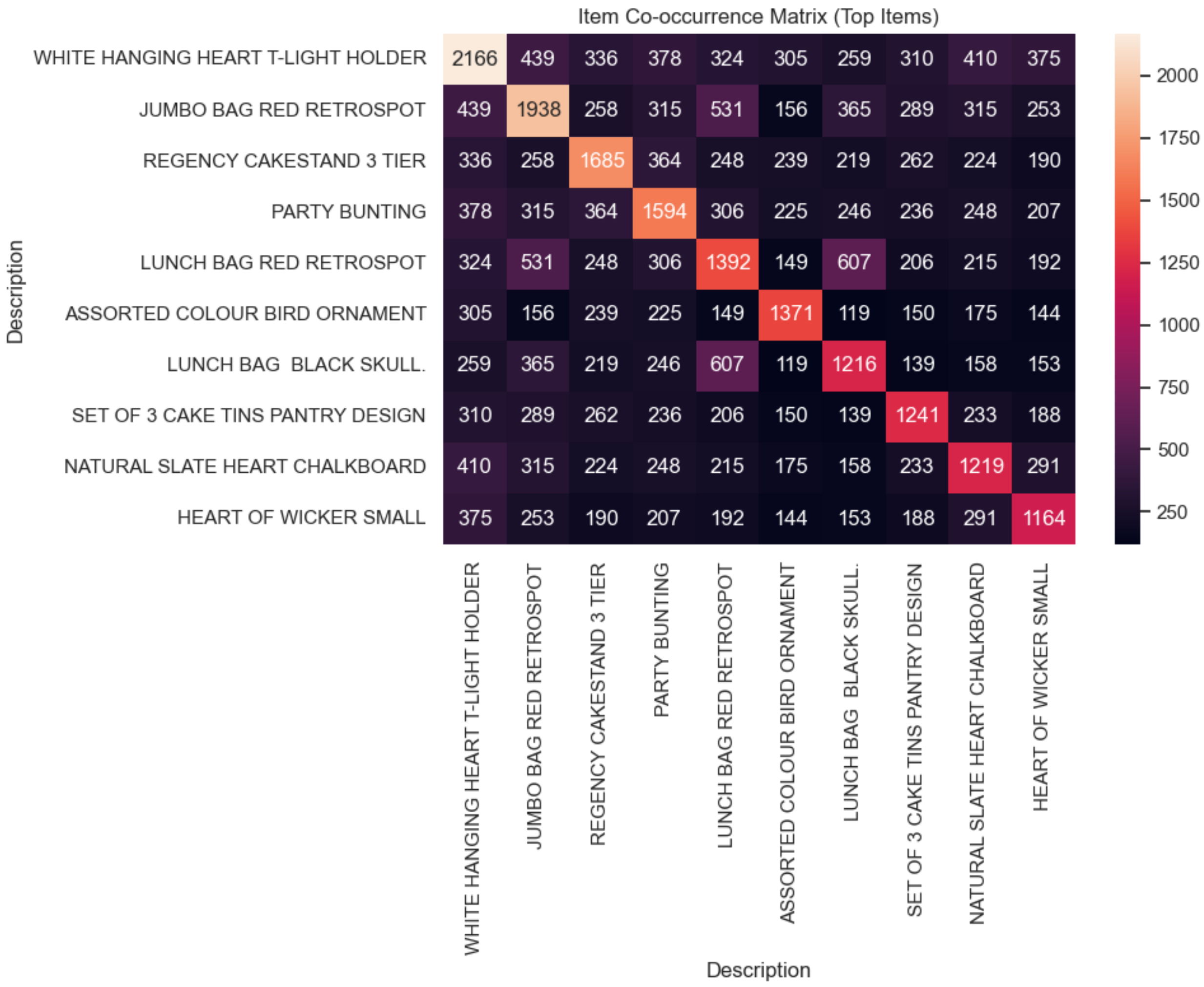


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```



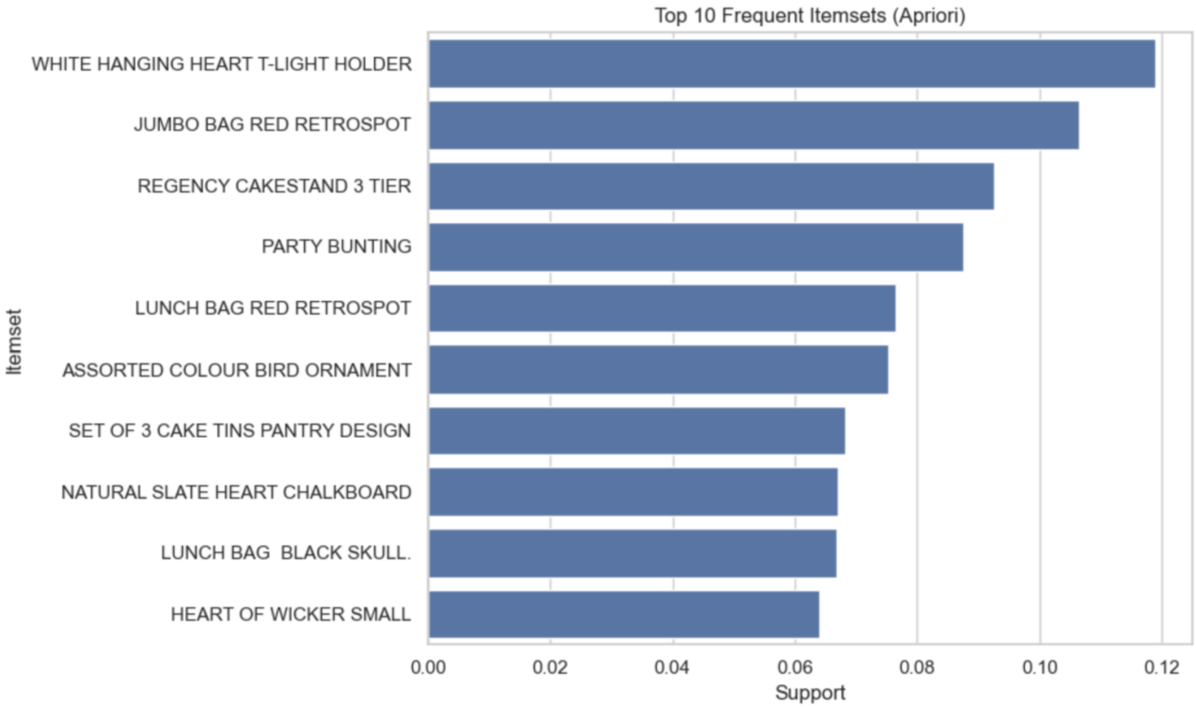
	count	unique	top	freq	mean	min	25%	50%	75%	max	std
InvoiceNo	541909.0	25900.0	573585.0	1114.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
StockCode	541909	4070	85123A	2313	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Description	540455	4223	WHITE HANGING HEART T-LIGHT HOLDER	2369	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Quantity	541909.0	NaN	NaN	NaN	9.55225	-80995.0	1.0	3.0	10.0	80995.0	218.081158
InvoiceDate	541909	NaN	NaN	NaN	2011-07-04 13:34:57.156386048	2010-12-01 08:26:00	2011-03-28 11:34:00	2011-07-19 17:17:00	2011-10-19 11:27:00	2011-12-09 12:50:00	NaN
UnitPrice	541909.0	NaN	NaN	NaN	4.611114	-11062.06	1.25	2.08	4.13	38970.0	96.759853
CustomerID	406829.0	NaN	NaN	NaN	15287.69057	12346.0	13953.0	15152.0	16791.0	18287.0	1713.600303
Country	541909	38	United Kingdom	495478	NaN	NaN	NaN	NaN	NaN	NaN	NaN

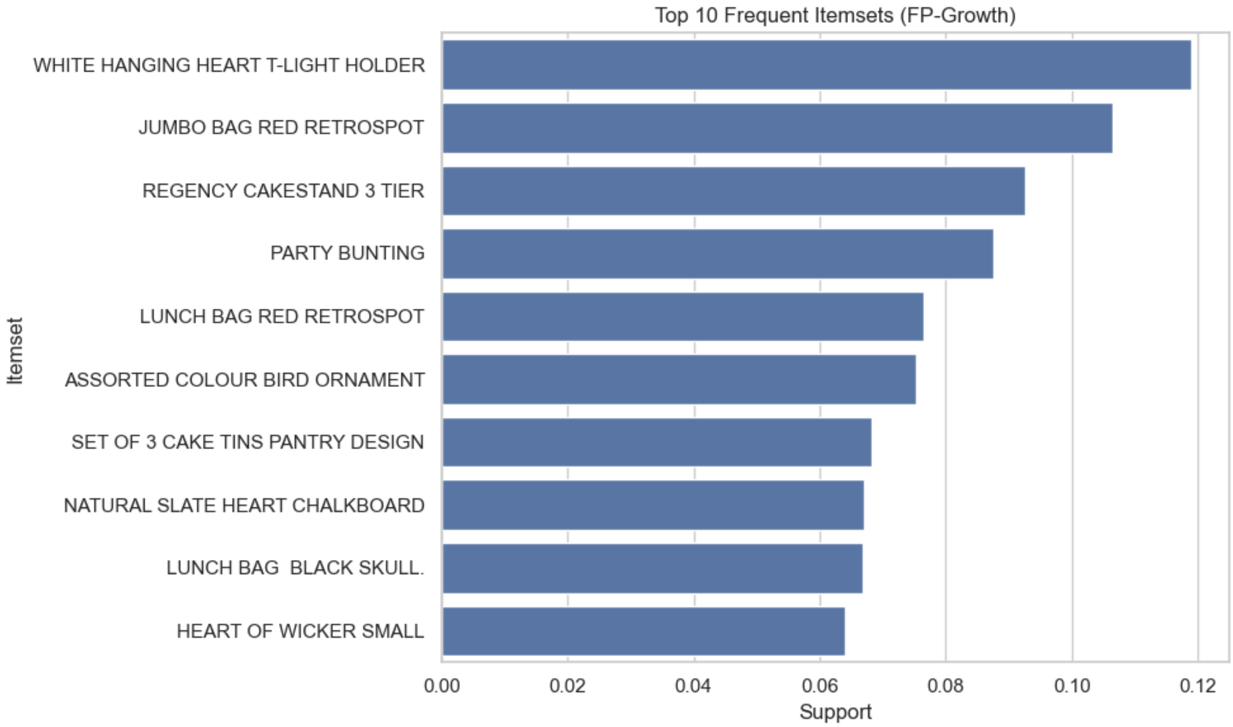
Next, the dataset was transformed into a basket matrix format where each row represented a transaction, and each column represented a product. Binary values indicated whether an item was present in a transaction, enabling the application of frequent itemset mining techniques.



Description	WHITE HANGING HEART T-LIGHT HOLDER	JUMBO BAG RED RETROSPOT	REGENCY CAKESTAND 3 TIER	PARTY BUNTING	LUNCH BAG RED RETROSPOT	ASSORTED COLOUR BIRD ORNAMENT	LUNCH BAG BLACK SKULL.	SET OF 3 CAKE TINS PANTRY DESIGN	NATURAL SLATE HEART CHALKBOARD	HEART OF WICKER SMALL
WHITE HANGING HEART T-LIGHT HOLDER	2166	439	336	378	324	305	259	310	410	375
JUMBO BAG RED RETROSPOT	439	1938	258	315	531	156	365	289	315	253
REGENCY CAKESTAND 3 TIER	336	258	1685	364	248	239	219	262	224	190
PARTY BUNTING	378	315	364	1594	306	225	246	236	248	207
LUNCH BAG RED RETROSPOT	324	531	248	306	1392	149	607	206	215	192
ASSORTED COLOUR BIRD ORNAMENT	305	156	239	225	149	1371	119	150	175	144
LUNCH BAG BLACK SKULL.	259	365	219	246	607	119	1216	139	158	153
SET OF 3 CAKE TINS PANTRY DESIGN	310	289	262	236	206	150	139	1241	233	188
NATURAL SLATE HEART CHALKBOARD	410	315	224	248	215	175	158	233	1219	291
HEART OF WICKER SMALL	375	253	190	207	192	144	153	188	291	1164

The Apriori algorithm was applied to identify frequent itemsets based on minimum support thresholds. These itemsets and their support values were visualized using Seaborn bar charts to highlight the most common product combinations. The FP-Growth algorithm was then implemented using the same support threshold for fair comparison. FP-Growth demonstrated better efficiency due to its tree-based structure that reduced repeated scans of the dataset.





Association rules were generated from the frequent itemsets using confidence thresholds, and metrics such as support, confidence, and lift were analyzed to interpret the strength and reliability of item relationships. Scatter plots of confidence versus lift helped visually identify strong rules.

Key Insights Gained

Strong purchasing patterns highlighted frequent product pairings that supported cross-selling and promotional strategies. FP-Growth consistently outperformed Apriori in terms of speed while producing similar results. Visualizations made complex patterns easier to understand and supported strategic decision-making applications such as targeted marketing and recommendation systems.

Challenges Encountered

The main challenges included file path errors, dataset formatting issues, parameter tuning for optimal support and confidence levels, and performance constraints with Apriori. These were resolved through careful data cleaning, threshold adjustments, and algorithm selection optimization.

Conclusion

This Lab provided valuable hands-on experience with association rule mining techniques, reinforcing the importance of data preparation, algorithm efficiency, and visualization in transforming raw transaction data into actionable business insights.

Github:github.com/SudhanshuDash/DataMining-Lab-6-Association-Rule-Mining-with-Apriori-and-FP-Growth