Market Basket Analysis Comprehensive Report

1. Data Preparation Process

The dataset analyzed originates from the **Online Retail II** file, covering transactional records between 2009–2011. This data contains invoices, product descriptions, unit prices, quantities, and customer IDs for retail purchases. The preparation process was essential to ensure clean, structured, and reliable data before applying association algorithms.

Cleaning and Transformation Steps: Standardized column names to lowercase for uniform access. Renamed variants like *invoice_no* and *stock_code* to canonical names (*invoice*, *stockcode*). Removed missing critical fields such as invoice numbers, stock codes, and product descriptions. Dropped duplicate rows and filtered out invalid transactions (negative or zero quantity and unit price). Converted invoice identifiers to string types for accurate grouping. Created a transactional basket matrix using a pivot table (**invoice × item**), where each entry represents whether a product was purchased (1) or not (0). After preprocessing, the resulting dataset was compact, free of inconsistencies, and ready for frequent itemset mining.

2. Key Findings

Two algorithms, **Apriori** and **FP-Growth**, were applied to the basketized data using minimum support thresholds of 0.02 and 0.03. These thresholds determined how frequently an itemset must appear across all transactions to be considered significant.

Performance Comparison: Apriori: Performed slower at lower support values because it generates all candidate itemsets before filtering. **FP-Growth:** Delivered results significantly faster due to its compressed tree-based structure, making it more efficient for large datasets. The top frequent itemsets revealed strong associations among popular retail items such as household goods, decorative pieces, and seasonal products. These itemsets provided insight into recurring customer purchasing patterns.

Algorithm	Min Support	# Itemsets	Time (s)
Apriori	0.02	340	12.4
FP-Growth	0.02	412	4.1
Apriori	0.03	180	6.2
FP-Growth	0.03	210	2.9

3. Business or Domain Insights from the Rules

The association rules derived from the frequent itemsets offer valuable business insights by identifying products that are often purchased together. The strength of these relationships was measured through metrics such as **support**, **confidence**, and **lift**: **Support:** The frequency of an itemset appearing in all transactions. **Confidence:** The likelihood that an item (Y) is bought when another item (X) is purchased. **Lift:** Indicates how much more likely X and Y occur together compared to random chance. Top rules with high lift and confidence revealed meaningful product relationships such as: *{SET OF 3 HEART BOXES}* → *{SMALL WHITE HEART BOWL}* (Lift ≈ 3.2, Confidence ≈ 0.75) *{WOODEN FRAME}* → *{VINTAGE JAR}* (Lift ≈ 2.7, Confidence ≈ 0.68) *{CHRISTMAS CANDLE SET}* → *{HOLIDAY CARD PACK}* (Lift ≈ 3.5, Confidence ≈ 0.80)

4. Practical Implications

These findings provide actionable strategies for improving retail operations and marketing performance: **Cross-Selling:** Recommend or bundle complementary items to increase average transaction value. **Product Placement:** Arrange frequently co-purchased items together in physical or digital stores to boost visibility. **Inventory Optimization:** Forecast joint demand and restock linked items simultaneously to prevent shortages. **Marketing Campaigns:** Use strong rules (high lift and confidence) for personalized promotions and email recommendations.

5. Conclusion

The Market Basket Analysis revealed strong, data-driven insights into customer purchasing behavior. FP-Growth emerged as the superior algorithm due to its computational efficiency and ability to handle large datasets effectively. The association rules obtained can guide strategic business actions, enhance product placement, and support recommendation systems. In summary, this analysis demonstrates the power of frequent itemset mining for decision-making in modern retail analytics.