MARKET BASKET ANALYSIS USING APRIORI AND FP-GROWTH ALGORITHMS TO ASSESS CONSUMER PURCHASE PATTERNS IN SELECTED AFRICAN MARKETS

Sandra Beldine Otieno

Introduction

Post the pandemic and global issues like the Russian Ukraine war Africa grapples with a mounting issue: the high cost of living.

As such consumers find themselves navigating a landscape characterized by economic uncertainties and increase in living expenses.

In response to these issues , consumers are proactively adopting cheaper alternatives, curtailing discretionary spending, and reassessing their purchasing priorities considering constrained financial resources.

Statement of the problem

Amidst this economic backdrop, retailers face a tough challenge, navigating uncharted waters where consumer purchase behavior is intricately woven with the complexities of their economic situations.

The uncertainty surrounding what products to stock on shelves becomes a paramount concern for retailers as they grapple with the ever-changing preferences and needs of a consumer base actively responding to the economic downturn.

The uncertainty surrounding what products to stock on shelves becomes a paramount concern for retailers as they grapple with the ever-changing preferences and needs of a consumer base actively responding to the economic downturn.

Objectives of the study

The main objective of the study was to assess consumer purchase behavior patterns in selected African markets using Apriori and FP-growth algorithms for market basket analysis, and to further segment these patterns based on country.

- To apply Apriori and FP-growth algorithms to survey data for identifying frequent item sets.
- To extract meaningful association rules from these item sets.
- To evaluate the significance of the association rules by assessing their strength and confidence.
- To compare the performance of Apriori and FP-Growth algorithms in market basket analysis.
- To develop detailed consumer behavior profiles for each country based on identified patterns.

Scope and Limitations

Urban Focus: The study primarily focused on urban-dwelling consumers in the 17 markets, potentially limiting the representation of rural or other regional populations.

Data Consistency: It assumed that the survey data collected by Kasi Insight maintained consistency in methodology, survey questions, and sampling procedures over time.

Response Accuracy: The study relied on the assumption that the responses provided by consumers were accurate and truthful, with the potential for response bias or inaccuracies in self-reported data.

Literature Reviewed

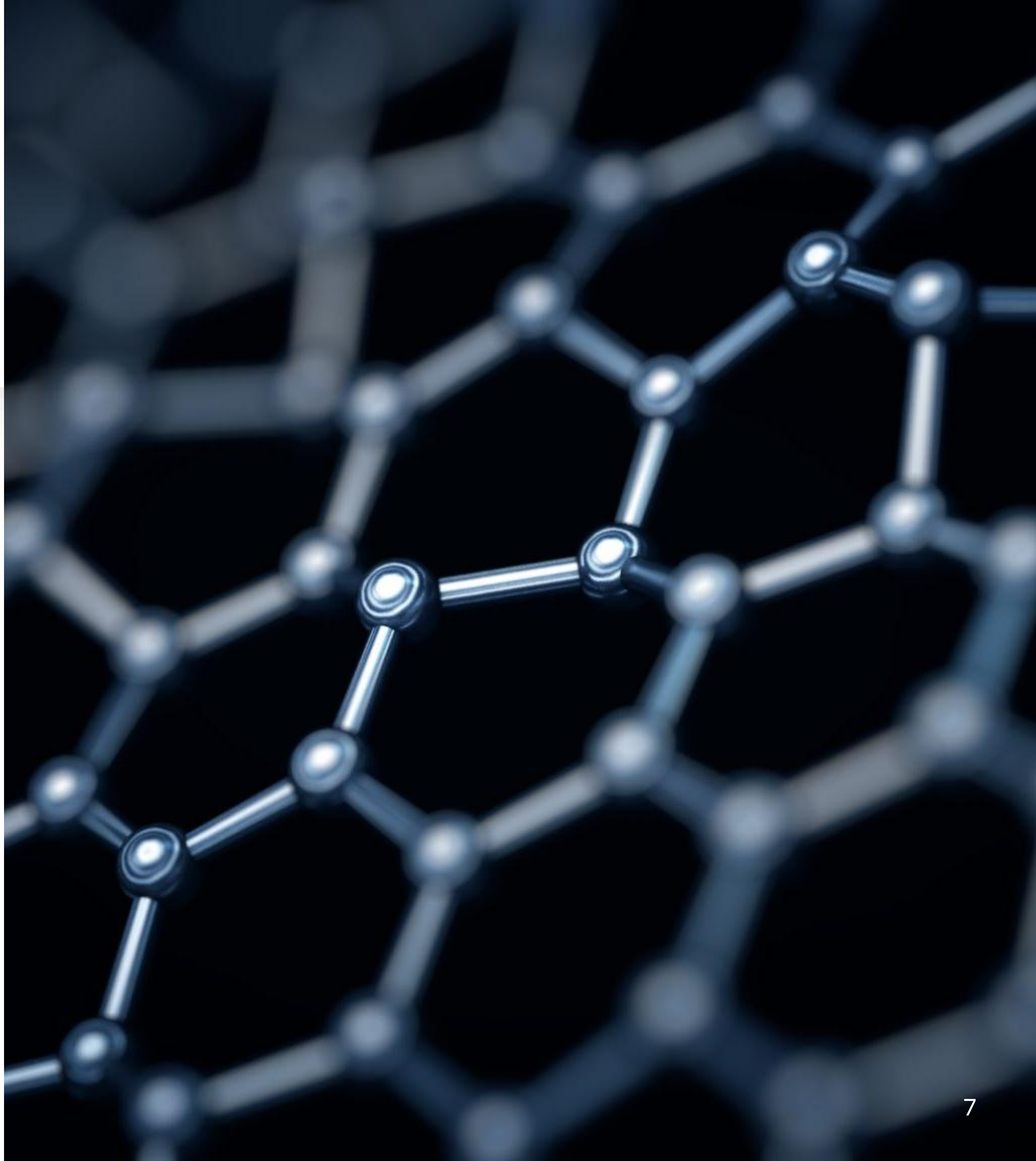
Study	Details
Izang et al. (2020)	Investigated four association rule mining algorithms: Apriori, FP-Growth, Association Outliers, and Supervised Association Rule. Emphasized the importance of algorithm selection based on factors like data volume, application domain, and customer preferences. Contributed to the practical considerations in implementing Market Basket Analysis systems.
Aldino et al. (2021)	Compared the efficiency and rule generation capabilities of FP-Growth and Apriori algorithms on transaction data. Provided practical insights into algorithm selection and performance.
Dio et al. (2022)	Undertook Market Basket Analysis in a beauty clinic context, identifying customer buying patterns for developing promotional menus. Relevant for practical applications of Market Basket Analysis in specific industries.
Yu-Unvan (2020)	Conducted Market Basket Analysis using Association Rules and provided practical insights by applying both Apriori and FP-Growth algorithms to supermarket sales data. Identified significant rules and associations in customer purchasing behavior.

Research Gap

Previous studies on Market Basket Analysis have primarily focused on individual retail entities like supermarkets, using Apriori and FP-Growth Algorithms. These studies typically rely on transactional data, which is readily available from such formal retail sources.

A significant portion of shopping in many African countries occurs in informal markets, making the collection of comprehensive transactional data challenging. This limits the breadth and applicability of traditional market basket analyses in these regions.

The study successfully conducted a Market Basket Analysis across 17 African markets, taking a novel approach to overcome the data collection challenges typical in informal settings. This method provides a broader and more inclusive perspective on consumer behavior patterns across the continent.



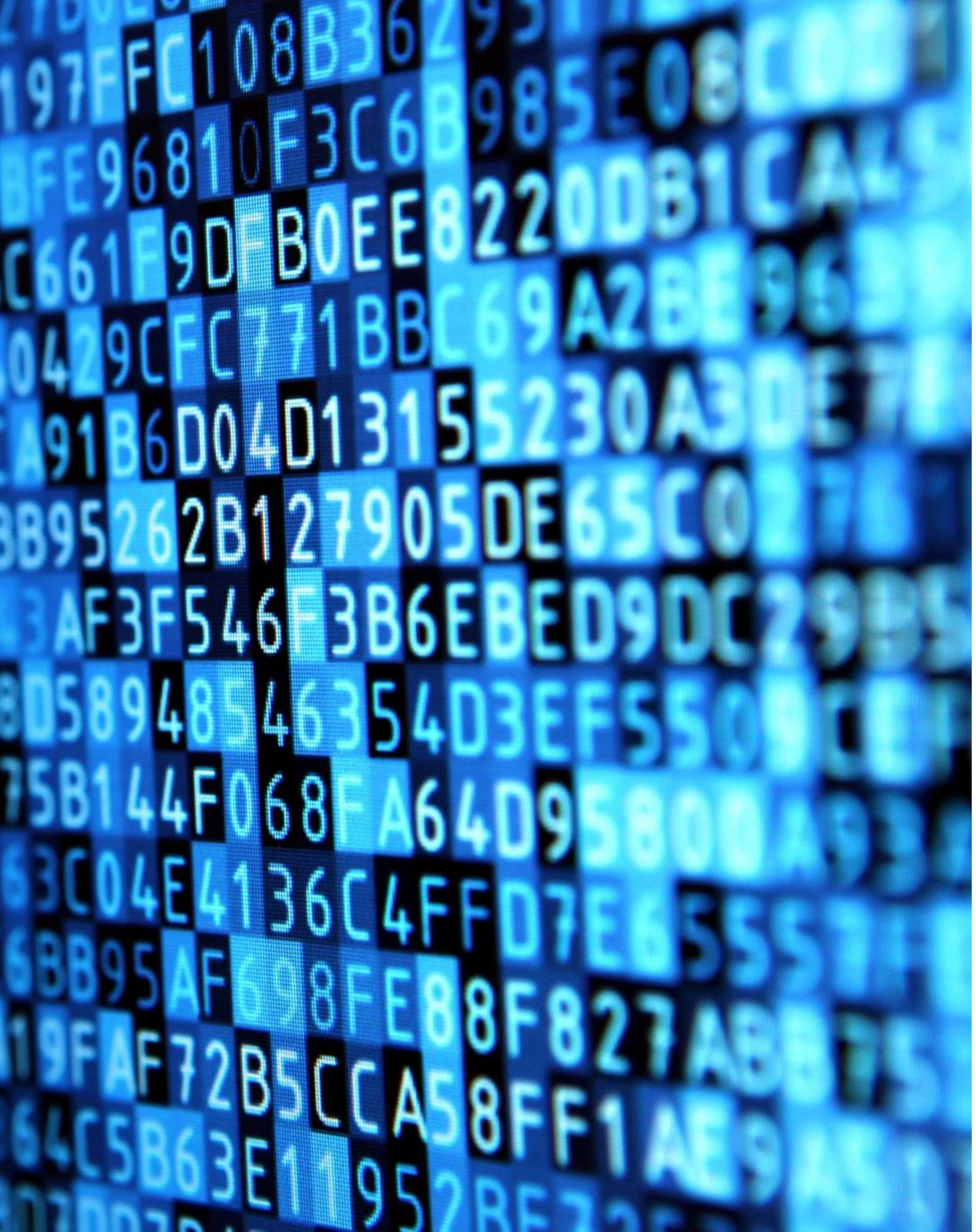
Research Methodology

The study utilized the CRISP-DM model.

The survey data was obtained from Kasi Insight, a market research firm, that has been consistently conducting monthly surveys across several African markets since 2016.

Upon acquiring the survey data from the Kasi Insight Portal, a series of essential preprocessing steps were applied to ensure the data's suitability for analysis.

Following the completion of data preprocessing, and exploratory data analysis, the analysis initiated the application of both the Apriori and FP-Growth algorithms.



Data Overview

- The survey spanned 17 African countries, including Algeria, Angola, Cameroon, Congo, Democratic Republic of Congo, Egypt, Ghana, Ivory Coast, Kenya, Mozambique, Namibia, Nigeria, Senegal, South Africa, Tanzania, Tunisia, and Zambia. The dataset included 62,187 rows and 105 columns, with the following variables:
- **Demographics:** Details such as gender, age group, marital status, highest educational level, occupation, household size, whether the household has children, and income level.
- **Purchases:** Data on monthly purchases divided into several categories:
    - Home and Personal Care Products:
    - Dry Foods:
    - Fruits and Vegetables
    - Fish and Meat Products
    - Beans and Legumes Products:
    - Protein Foods
    - Snacks
    - Non-alcoholic Beverages and Alcoholic Beverages

Data Preprocessing steps

- **Data Compilation**

Survey data from June, September, and December 2023 were combined using Python's pd.concat, ensuring consistency across the datasets.

- **Data Cleaning**

Missing Values: Applied listwise deletion for heavily missing data (e.g., tofu column) and mode imputation for lesser gaps.

Duplicates: Removed duplicate entries using unique respondent codes to maintain data integrity.

Outliers: Identified rare categories in the Yes/No responses, confirming no traditional outliers.

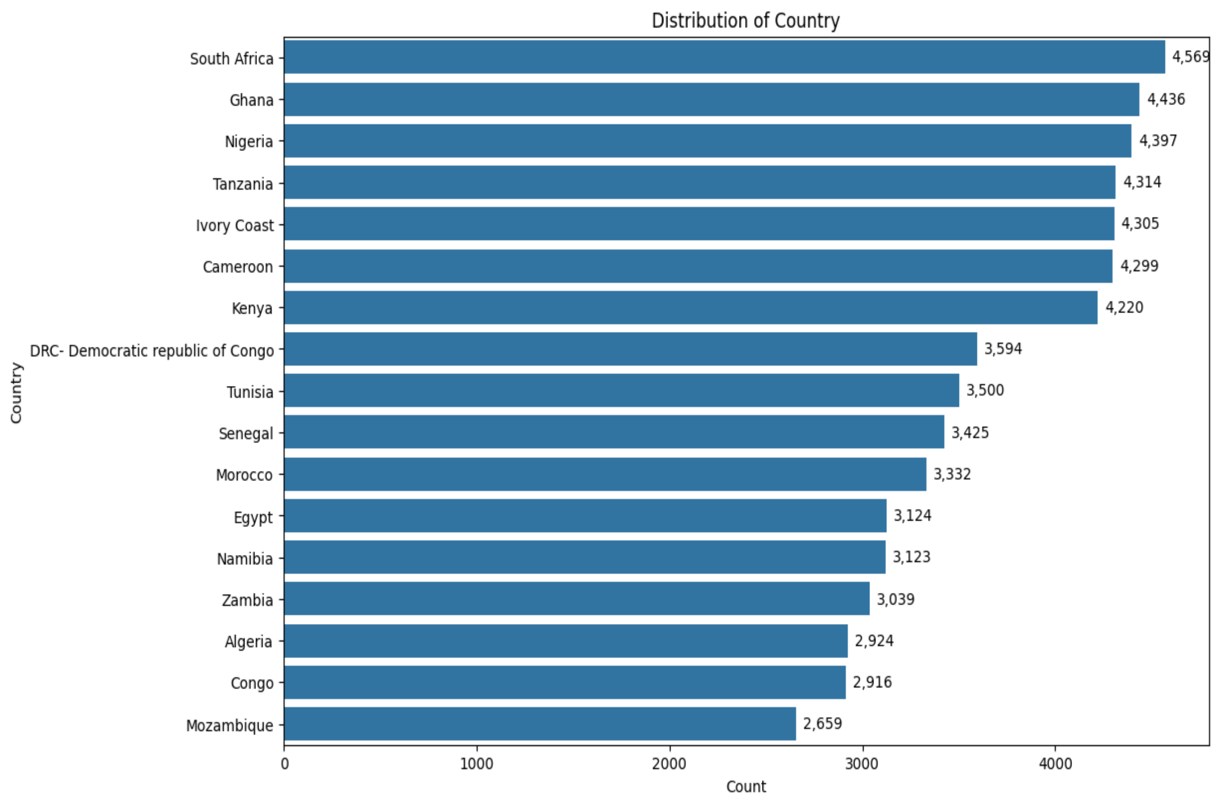
- **Data Transformation**

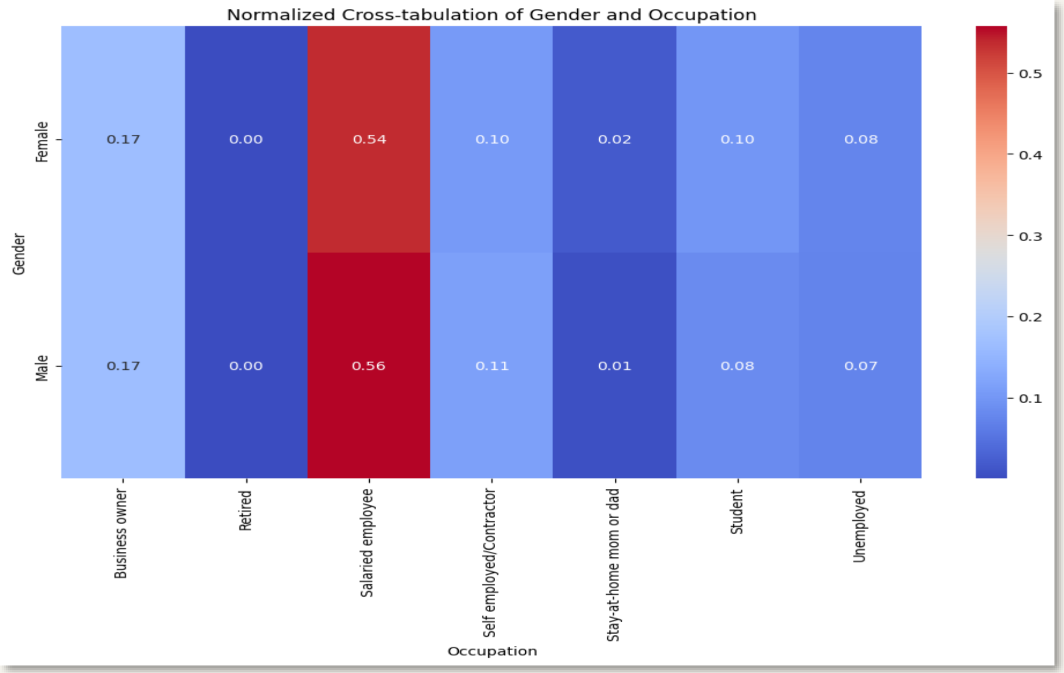
Employed one-hot encoding to convert categorical responses into binary format, making the data compatible for modeling.

- **Feature Engineering**

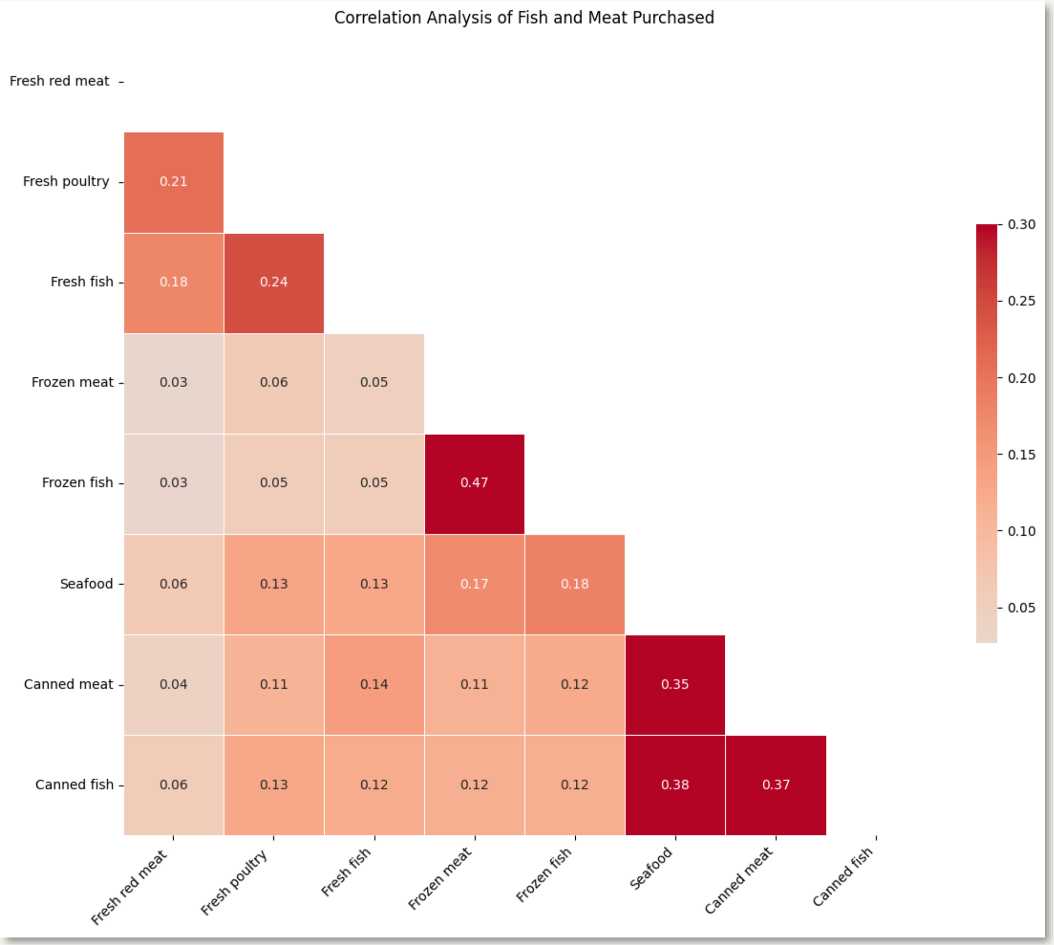
Refined product categories into three groups—food, non-food, and beverages—to optimize the dataset for market basket analysis and clustering algorithms.

Exploratory Data Analysis



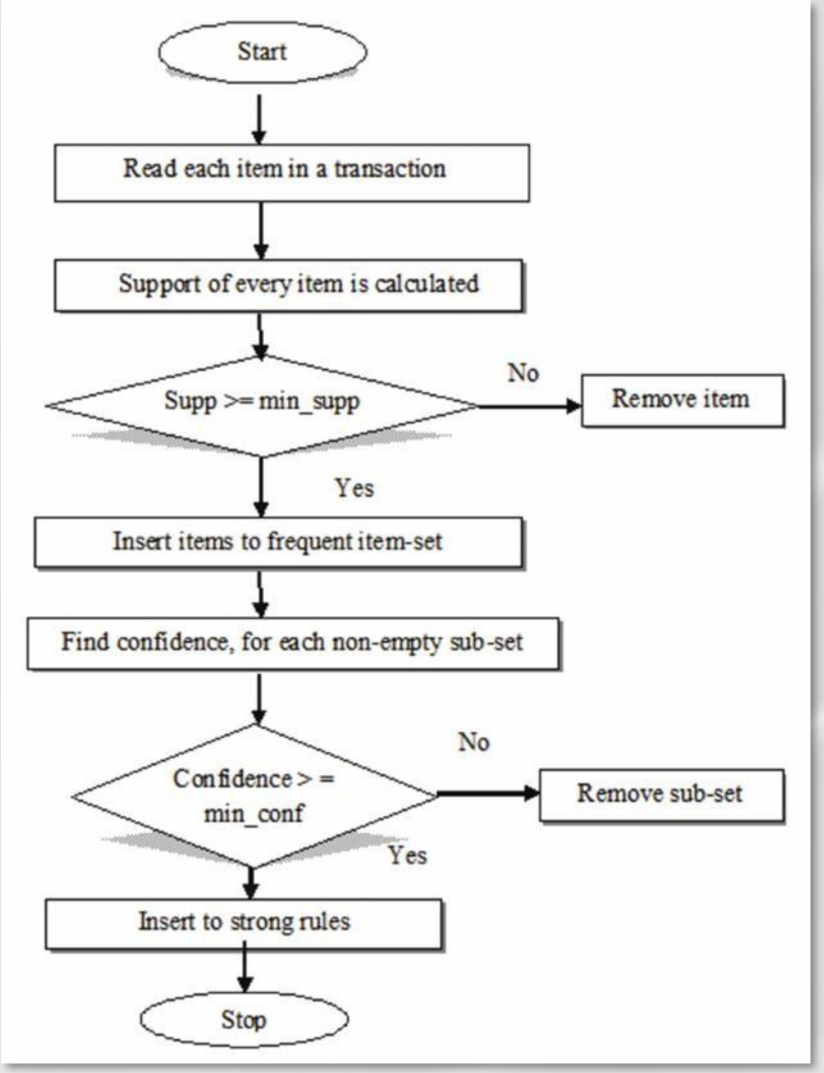


Exploratory Data Analysis

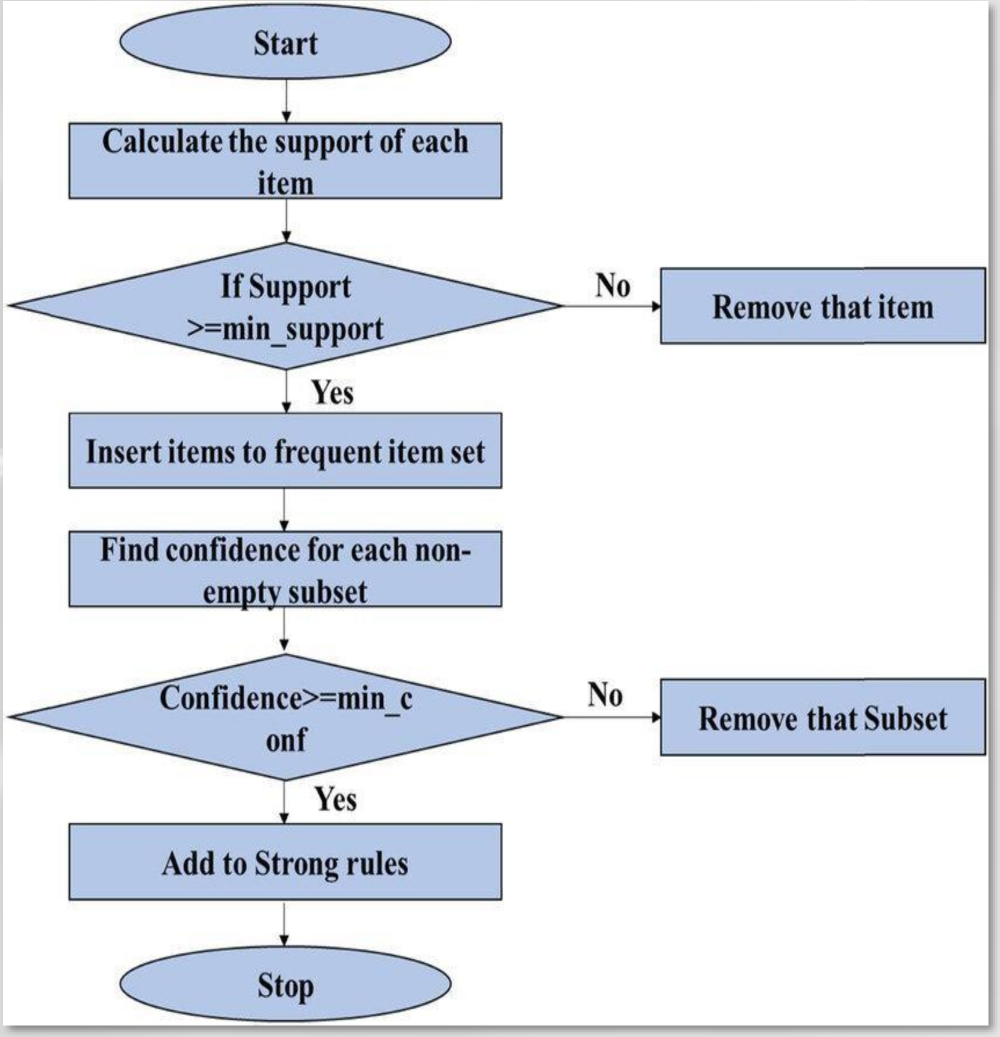


- Higher correlations indicate a strong relationship, meaning the products are often bought together.
- Lower correlations suggest a weaker relationship, indicating less likelihood of the products being purchased together.

Association Rule Mining Methodology



Apriori –Flow chart



FP Growth Pattern –Flow chart

Evaluation Metrics

- **Support** is a fundamental metric that helps identify the prevalence of an itemset, I within a dataset D.

$$Support\ (I) = \frac{Number\ of\ transactions\ containing\ I}{Total\ number\ of\ transactions}$$

- **Confidence** measures the strength or reliability of an inference made by an association rule, specifically the likelihood of observing the consequent of the rule given the presence of its antecedent:

$$Confidence(A \Rightarrow B) = \frac{Support(A \cup B)}{Support(A)}$$

- **Lift** offers a measure of the rule's effectiveness and strength by comparing the observed support of A and B appearing together against the expected support if A and B were statistically independent:

$$Lift(A \Rightarrow B) = \frac{Confidence(A \Rightarrow B)}{Support(B)}$$

- **Memory Usage:** Critical for processing large datasets, memory usage varies with the number of candidate item sets generated by Apriori and the size of the FP-tree in FP-Growth, which holds the dataset's frequency information.

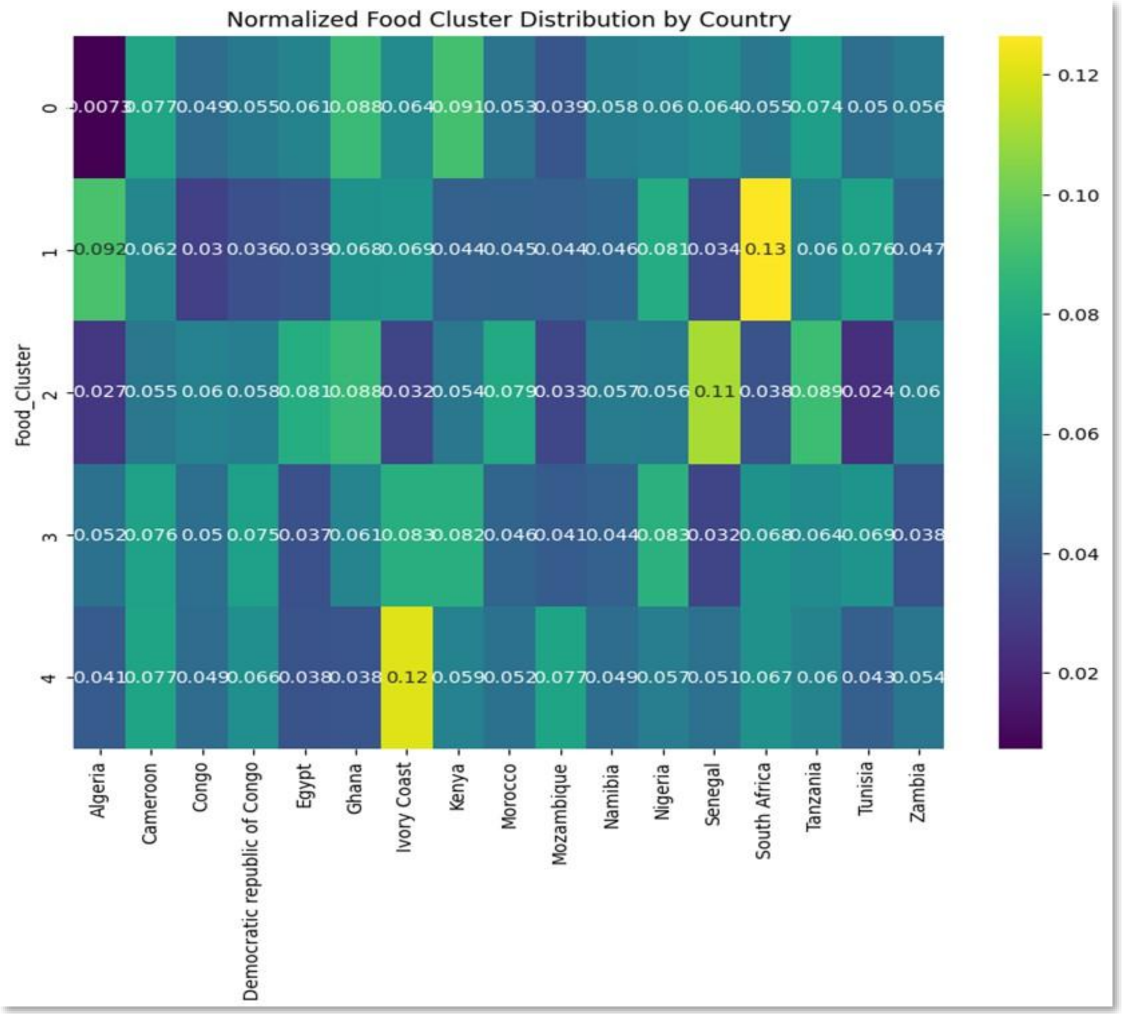
- **Efficiency and Effectiveness:** These metrics balance how quickly and resource-efficiently an algorithm processes data (efficiency) and the accuracy and relevance of the patterns identified (effectiveness). An algorithm might be fast but less thorough, or slower but more comprehensive in pattern detection.

Association Rule Mining Comparison Results

Metrics	Apriori	FP-Growth
Itemsets	31 itemsets	31 itemsets
Rules	180 rules	180 rules
Lift	1.064623	1.064623
Support	0.662500	0.662500
Conviction	3.935224	3.935224
Leverage	0.040481	0.040481
Time taken	0.0665 seconds	2.0778 seconds
Memory Usage	17.6283 MB	4.3542 MB

- The study compared the performance of the Apriori and FP-Growth algorithms across various categories. Both algorithms were equally effective in identifying the same number of item sets and generating an equivalent set of rules.
- Both algorithms were equally effective in identifying the same number of item sets and generating an equivalent set of rules.
- The Apriori algorithm excelled in execution speed across all categories.
- FP-Growth demonstrated a consistent advantage in memory efficiency, making it preferable for large datasets or memory-constrained environments.

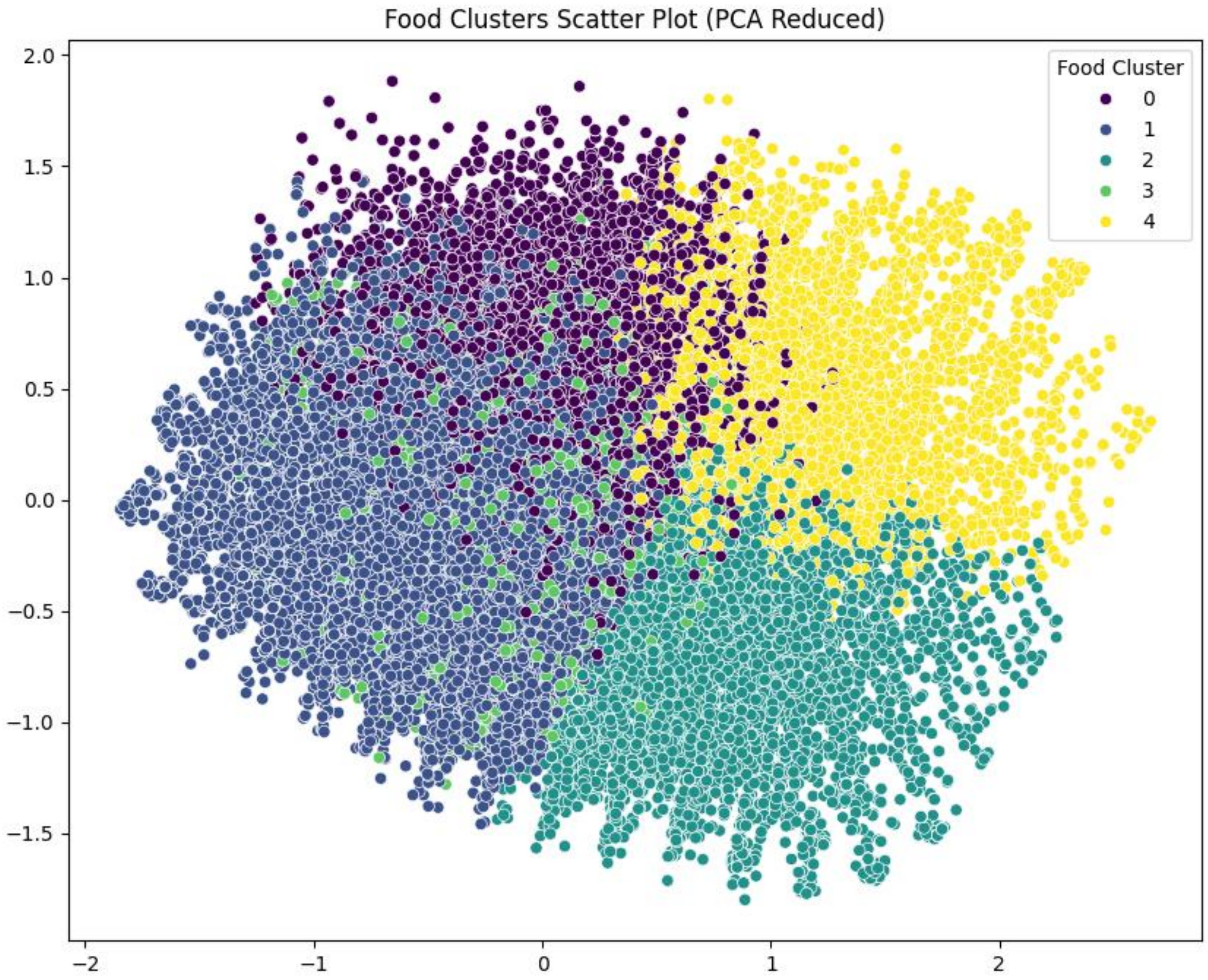
Clustering Results



The dietary clusters across several African countries reveal distinct consumption patterns: Cluster 0, prevalent in Kenya and Ghana, suggests a well-rounded diet integrating traditional and global foods, indicating economic stability and diverse food access.

Cluster 1, notable in South Africa, highlights a reliance on essential foods, possibly due to economic constraints or cultural significance. Cluster 2, characterized by Senegal, blends convenience with tradition, likely due to urbanization. Cluster 3, found in Ivory Coast, Kenya, and Nigeria, focuses on basic, cost-effective ingredients like flour and legumes, reflecting economic necessities or cultural preferences. Cluster 4, dominant in Ivory Coast, emphasizes a nutrient-dense diet rich in carbohydrates and proteins, possibly supported by specific agricultural policies.

Clustering Results



The plot shows distinct consumer purchasing patterns. Tight groupings (e.g., Cluster 0) indicate homogeneity in preferences, suitable for targeted promotions. Dispersed clusters (e.g., Cluster 4) suggest diverse preferences, requiring a broad range of product offerings.

Discussion

Survey Analysis and Algorithm Efficiency: The study involved a detailed examination of demographic trends and product purchase frequencies, with a particular focus on co-purchasing patterns among various food items. The data mining utilized both the Apriori and FP-Growth algorithms, revealing that Apriori was unexpectedly quicker, while FP-Growth was more memory efficient. This differentiation helps in choosing the suitable algorithm based on the complexity and scale of the data.

Consumer Clustering

- Cluster 0 in Kenya and Ghana reflects strong demand for a wide variety of both local and imported fresh and pantry foods. This suggests a market preference for stores that can serve as one-stop shops.
- Cluster 1 in South Africa points to a reliance on affordable and versatile staple foods, indicating a consumer preference for value and convenience in staple purchases.
- Cluster 2 in Senegal highlights a mix of traditional food offerings with convenience items like baked goods, suggesting a dual demand for time-saving products alongside fresh produce.
- Cluster 3 in Ivory Coast, Kenya, and Nigeria shows a focus on economical and non-perishable food items, implying a consumer trend towards cost-effective dietary solutions.
- Cluster 4 in Ivory Coast indicates a preference for energy-dense and protein-rich foods, reflecting a consumer interest in nutritious and hearty meal options.

Recommendations

- **Adapt Retail Strategies to Local Demands**

Kenya and Ghana (Cluster 0): Retailers should ensure a consistent supply of both local and imported fresh and pantry foods to meet the demand for diverse dietary preferences. Consider expanding the range of products offered to include both traditional local staples and international items.

- **Enhance Consumer Engagement and Education**

Engaging with local communities through events or partnerships can help retailers better understand and meet the local consumer needs and preferences, fostering loyalty and community ties.

- **Leverage Technological Insights for Inventory Management**

Utilize data analytics to track sales and inventory levels to keep popular items in stock and to anticipate shifts in consumer preferences based on seasonal and economic changes.

- **Collaborate with Local Producers**

Strengthen relationships with local farmers and producers to ensure a steady supply of fresh and staple foods, which can also support the local economy. This is particularly pertinent for markets focused on economical and staple food items.

Objective Recap

Frequent Item Sets Identified: Both Apriori and FP-Growth algorithms successfully identified frequent item sets, with each identifying 31 item sets for food products.

Strong Association Rules: Generated 180 association rules with metrics showing strong associations (e.g., lift, support, and confidence).

Algorithm Performance: Apriori was faster in execution time, while FP-Growth was more memory efficient.

Distinct Consumer Clusters: Clustering analysis revealed clear consumer segments with distinct purchasing behaviors, aiding in targeted marketing strategies.

THANK YOU