**Data Analytics Capstone Topic Approval Form**

**Student Name:** Ebenezer Essilfie Wilson

**Student ID:** 011064333

**Capstone Project Name:** Improving product bundles and promotions in grocery stores using Market Basket Analysis.

**Project Topic**: This project will capture the essence of leveraging association rule mining via market basket analysis to improve product bundling and promotional efforts in a grocery store setting. The project will use dataset from Kaggle: <https://www.kaggle.com/datasets/bhavikjikadara/grocery-store-dataset?resource=download>.

**This project does not involve human subjects research and is exempt from WGU IRB review.**

**Research Question:** How can association rule mining be applied to discover effective product combinations for cross-selling in a grocery store setting?

**Hypothesis**: **Null hypothesis**- There is no statistically significant relationship between products purchased together in a grocery store; association rule mining does not reveal effective product combinations for cross-selling. **Alternate Hypothesis**- There is a statistically significant relationship between products purchased together in a grocery store; association rule mining reveals effective product combinations for cross-selling.

**Context:** The research question— *“How can association rule mining be applied to discover effective product combinations for cross-selling in a grocery store setting?”*—is of significant scholarly and practical relevance in the domains of retail analytics and consumer behavior. Given the vast assortment of products offered in grocery stores, identifying meaningful associations among items frequently purchased together presents an opportunity to enhance sales strategies and operational efficiency. This is particularly pertinent in cross-selling, where understanding co-purchase patterns can directly inform product placement, promotional tactics, and inventory management. Association rule mining, a core technique within the field of data mining, is particularly suited to this analytical challenge. It enables the identification of statistically significant relationships between items in large transactional datasets. Utilizing metrics such as support, confidence, and lift, association rule mining facilitates the detection of product pairings that occur with greater frequency and reliability than would be expected by chance. These insights enable retailers to move beyond intuition-based decision-making, adopting a data-driven approach to designing product bundles and recommending related items. Incorporating association rule mining into retail operations offers several benefits. From a marketing perspective, it supports the development of targeted promotions and dynamic product suggestions tailored to observe customer behaviors. From a merchandising standpoint, it can inform optimal shelf layouts by placing commonly co-purchased items in proximity to one another, thereby enhancing convenience and encouraging unplanned purchases. Operationally, the ability to anticipate joint demand allows for more accurate inventory forecasting, reducing the risk of stockouts and overstock. Furthermore, the application of data analysis to this research question is essential for transforming raw transactional data into actionable business intelligence. Through systematic pattern discovery, businesses can uncover latent structures in consumer purchasing behavior that are not readily observable through manual inspection. This contributes not only to improved revenue performance but also to a more seamless and personalized customer shopping experience. In summary, the question under investigation is well-positioned to benefit from a rigorous data analysis approach. Association rule mining offers a robust methodological framework for uncovering actionable insights that support strategic cross-selling initiatives in the grocery retail sector.

**Data:** The dataset used is called Grocery Store Dataset from Kaggle and here is the link to it; <https://www.kaggle.com/datasets/bhavikjikadara/grocery-store-dataset?resource=download>.

These are the features of the dataset; Subcategory: This column categorizes the grocery items into subcategories, providing a detailed classification for easier analysis and organization.

Price: Represents the monetary value of the grocery item, indicating its cost or retail price in the specified currency.

Discount: Reflects any discounts or promotional offers applicable to the respective grocery items, providing insights into pricing strategies.

Rating: Indicates customer satisfaction or product quality based on user ratings, offering a measure of the overall perceived value of the grocery item.

Title: Describes the name or title of the grocery item, providing a concise identifier for easy reference and understanding.

Currency: Specifies the currency in which the prices are denominated, facilitating proper interpretation and comparison of monetary values.

Feature: Includes features or characteristics of the grocery item, offering additional information about its unique attributes or selling points.

Product Description: Provides a detailed textual description of the grocery item, offering comprehensive information about its specifications, uses, and other relevant details. This column is handy for understanding product details beyond what is captured in other columns.

The Grocery Store Dataset hosted on Kaggle and created by Bhavik Jikadara is publicly available under the licensing terms defined on the dataset’s Kaggle page. As the dataset is explicitly shared for public use; especially for learning, research, and project development and so it can be legally and ethically used for academic purposes such as a capstone project.

*Note: If you are using restricted information, please have the Third-Party Authorization Form signed by an authorized agent on behalf of the data owner. The data owner’s legal name is required on the form.*

**Data Gathering:**  This study utilizes secondary data obtained from a publicly available dataset titled “Grocery Store Dataset” curated by Bhavik Jikadara and hosted on the Kaggle open data platform. The dataset is designed to represent transactional-level purchase behavior in a grocery store setting and is particularly suitable for conducting market basket analysis due to its structure, which records multiple items associated with unique transaction identifiers.

Data acquisition will involve downloading the dataset directly from the Kaggle repository (https://www.kaggle.com/datasets/bhavikjikadara/grocery-store-dataset). Kaggle is a widely recognized platform for sharing datasets that support academic, research, and non-commercial applications. The dataset is publicly accessible and, at the time of use, carries no known licensing restrictions that would prohibit educational usage. Proper attribution to the original data publisher and platform will be maintained in accordance with academic standards.

Following the acquisition, the dataset will be imported into a secure, local computational environment for preprocessing and analysis using Python programming tools. The data will undergo a series of preparatory steps, including data parsing, formatting, and validation to ensure analytical readiness. Key tasks will involve transforming item lists into a binary-encoded matrix format suitable for the application of association rule mining algorithms such as Apriori or FP-Growth. Any missing or malformed entries will be addressed using standard data-cleaning procedures.

As the dataset does not contain personally identifiable information (PII) and is synthetic in nature, its use complies with ethical guidelines for secondary data research. No institutional review board (IRB) approval is required due to the anonymized and non-human subject status of the dataset.

In summary, the data-gathering methodology centers on the ethical use of an open-access, high-quality dataset that enables robust, replicable analysis of grocery purchase behavior. This approach supports the research objective of uncovering meaningful product associations for cross-selling within a retail context.

**Data Analytics Tools and Techniques**: The principal analytical method to be employed in this study is Association Rule Mining (ARM), a widely recognized technique in the field of data mining and unsupervised machine learning. Association rule mining is particularly well-suited for uncovering latent patterns and co-occurrence relationships within transactional datasets, such as those derived from grocery retail environments. This methodology facilitates the identification of frequent item combinations and the formulation of decision rules that can inform cross-selling strategies.

ARM operates by evaluating transactional records to extract item sets and generate if-then rules of the form “If Item A is purchased, then Item B is likely to be purchased.” These associations are quantitatively assessed using three key metrics:

Support, which measures the frequency with which an itemset appears in the dataset;

Confidence, which evaluates the likelihood of the consequent given the antecedent;

Lift, which indicates the strength of the association relative to random chance, thus providing insights into the practical relevance of the rule.

This study will apply the Apriori algorithm, a foundational ARM approach that leverages the anti-monotonic property of frequent item sets, ensuring computational efficiency by pruning infrequent subsets during rule generation. The rules extracted through this process will be analyzed to identify product pairings with high cross-promotional potential.

Data processing and analysis will be conducted using the Python programming environment. The following libraries will be utilized:

mlxtend, for implementing the Apriori algorithm and generating association rules;

pandas, for data structuring and manipulation.

matplotlib and seaborn, for visualizing rule distributions, item co-occurrence frequencies, and network relationships.

The application of association rule mining is methodologically appropriate for the selected dataset and research objectives. It enables the derivation of actionable insights into consumer purchasing behavior and supports data-driven decision-making related to merchandising, promotional design, and product placement. Moreover, its interpretability and scalability make it particularly suitable for both exploratory and applied retail analytics in academic and commercial contexts.

**Justification of Tools/Techniques:** Association Rule Mining (ARM) is an appropriate technique for analyzing the grocery store dataset because it is specifically designed to identify patterns and relationships among items in transactional data. This aligns with the study’s goal of uncovering product combinations for cross-selling. ARM uses key metrics—support, confidence, and lift—to evaluate how frequently items are purchased together and the strength of their association. The Apriori algorithm, known for its efficiency, will be applied to generate these rules. Since the data is unlabeled and exploratory in nature, ARM is particularly well-suited, offering meaningful, interpretable insights that can guide retail strategies such as product placement and targeted promotions.

**Project Outcomes**:

This capstone project aims to explore and apply association rule mining techniques to identifyeffective product combinations for cross-selling in a grocery store setting. The following are the key expected outcomes and deliverables:

1. Cleaned and Structured Dataset

A fully processed transactional dataset ready for analysis. This will include:

Parsed and structured transaction records

One-hot encoded item sets for algorithm compatibility

Any necessary data cleaning and preprocessing steps

2. Association Rules

A comprehensive set of association rules generated using the Apriori algorithm. These rules will include:

Support, confidence, and lift values

Identification of high-interest item combinations

Interpretation of frequent co-purchase patterns

3. Visualizations

Data visualizations to support analysis and communicate insights, including:

Bar charts of the top frequent items and item sets

Scatter plots of support vs. confidence/lift

Network graphs showing relationships between commonly co-purchased products

4. Business Insights and Recommendations

Actionable insights for grocery retailers based on the analysis, such as:

Suggested product bundles for cross-promotions

Recommendations for shelf arrangement based on item associations

Opportunities for targeted marketing based on purchase patterns

5. Final Report

A well-structured academic paper including:

Abstract, introduction, and literature review

Methodology (data collection, processing, and analysis)

Results with visualizations

Business implications and conclusion

Proper citations and references

6. Presentation Deck

A summary PowerPoint or Google Slides presentation highlighting:

The problem statement and objectives

Methods and analysis techniques

Key findings and visualizations

Strategic recommendations

**Projected Project End Date**: 8/31/2025

**Sources**:

Han, J., Pei, J., & Kamber, M. (2021).

Data Mining: Concepts and Techniques (4th ed.). Morgan Kaufmann.

This textbook provides the theoretical foundation for association rule mining, including the Apriori algorithm, and covers practical applications in retail.

Tan, P. N., Steinbach, M., Karpatne, A., & Kumar, V. (2018).

Introduction to Data Mining (2nd ed.). Pearson.

Discusses data preprocessing, association rule generation, and evaluation metrics like support, confidence, and lift in detail.

Agrawal, R., Imieliński, T., & Swami, A. (1993).

Mining association rules between sets of items in large databases. In Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (pp. 207–216). ACM.

Seminal paper introducing the concept of association rules and the Apriori algorithm.

Hipp, J., Güntzer, U., & Nakhaeizadeh, G. (2000).

Algorithms for association rule mining—a general survey. ACM SIGKDD Explorations Newsletter, 2(1), 58–64.

Provides an overview and comparison of different algorithms used in association rule mining, including scalability and efficiency.

Kaggle (n.d.).

Grocery Store Dataset by Bhavik Jikadara. Retrieved from: https://www.kaggle.com/datasets/bhavikjikadara/grocery-store-dataset

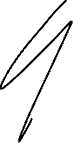
The dataset used for this study offers simulated retail transactions suitable for pattern mining and educational research.

**Course Instructor Signature/Date:**

The research is exempt from an IRB Review.

An IRB approval is in place (provide proof in appendix B).

Course Instructor’s Approval Status: Approved



Date: 6/17/2025

Reviewed by:

Comments: Click here to enter text.