# **Comprehensive Analysis of Online Retail Customer Purchase Behavior Using Data Mining Techniques**

# **Executive Summary**

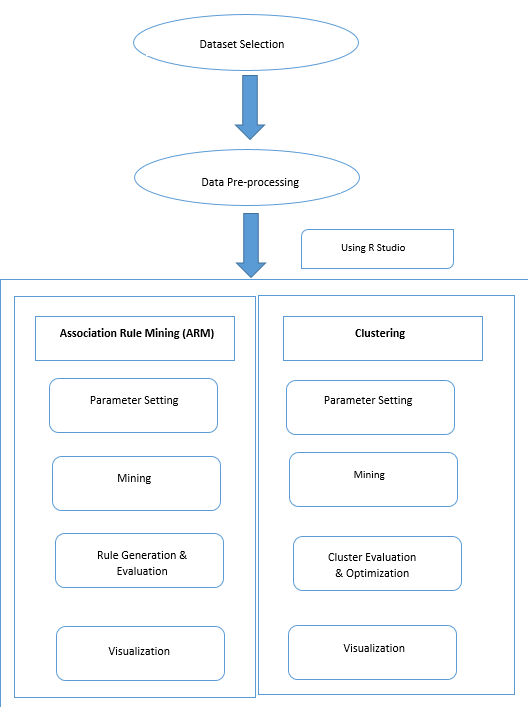
Through the analysis of the 'Online Retail' dataset from Kaggle, this report leveraged Association Rule Mining (ARM) and Clustering within R Studio to uncover customer purchase behaviors. ARM pinpointed patterns of co-purchased items, indicating opportunities for cross-selling and inventory strategy enhancements. Concurrently, Clustering classified customers into distinct segments based on purchase volume and frequency, guiding targeted marketing initiatives. The strategic interplay of these data mining techniques, depicted in detailed visualizations, provided a granular understanding of consumer trends and preferences. This integrative approach delivered actionable insights, shaping retail marketing, and operational strategies to foster customer engagement and optimize sales outcomes.

# **Customer Purchase Behaviour in Retail**

Retail businesses are inundated with vast amounts of transaction data, which, if harnessed effectively, can yield significant insights into customer purchase behavior. This data, drawn from various sources like in-store purchases, online sales, and loyalty programs, offers retailers a deep understanding of consumer preferences and habits (Gupta and Shukla, 2019). By analyzing this information, retailers can tailor their marketing strategies, enhance customer experiences, optimize inventory and product assortments, and predict future buying trends. Such data-driven approaches not only enable personalized customer engagement but also drive efficient business operations, helping retailers to remain competitive in a rapidly evolving market (Prior, 2023).

The objective of this analysis is to apply data mining techniques to the chosen dataset to uncover patterns in customer purchases. Goal is to identify products that are frequently bought together and segment customers based on their buying habits. These insights will enable retailers to offer more personalized shopping experiences and optimize their inventory management in line with consumer demand (Mandal et al., 2021).

# **Tool Selection**



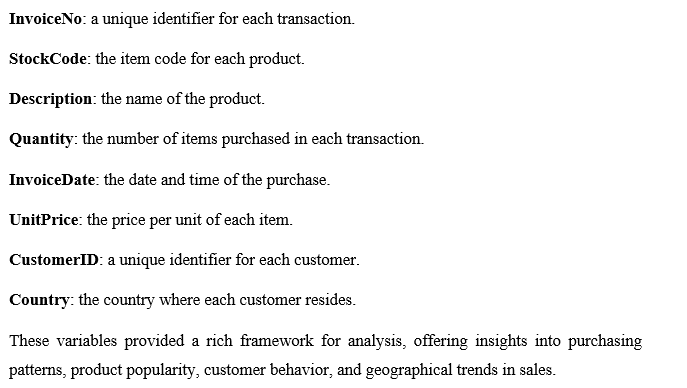
**Figure 1. Two-pronged data mining approach (self-created)**

This report employed a two-pronged data mining approach namely Association Rule Mining (ARM) and Clustering (As shown in Figure 1). Initially, an Apriori algorithm was utilized for ARM to discern patterns of items frequently purchased together, serving as the foundation for our cross-selling strategies (Chen et al., 2020). Subsequent clustering analysis segments customers into groups based on their purchasing behaviour. Data pre-processing involved normalization and handling of missing values to ensure robustness in the analysis. Parameters for the ARM, such as support and confidence thresholds, were carefully selected to balance insight granularity with statistical significance (Shahbazova et al., 2021).  
The rationale behind choosing both ARM and Clustering was that ARM excels in uncovering associations between products, while clustering is adept at segmenting customers based on purchasing profiles.

In the practical application of these data mining techniques, R Studio emerges as a powerful and versatile tool, offering a range of functionalities for data analysis and visualization. Utilizing R Studio's comprehensive suite of packages, such as '**arules**' for Association Rule Mining and '**cluster**' for clustering analysis, it was efficient to process, analyze, and extract meaningful insights from complex datasets like the one chosen for this report. R Studio's capabilities in handling large datasets, combined with its advanced data mining and visualization features, allowed for the exploration of patterns, correlations, and trends in the data in a more nuanced and sophisticated manner (Rhys, 2020). Further sections will apply Association Rule Mining and Clustering techniques to the dataset using R Studio and elucidate the insights obtained and their relevance to retail strategy.

# **Dataset Selection**

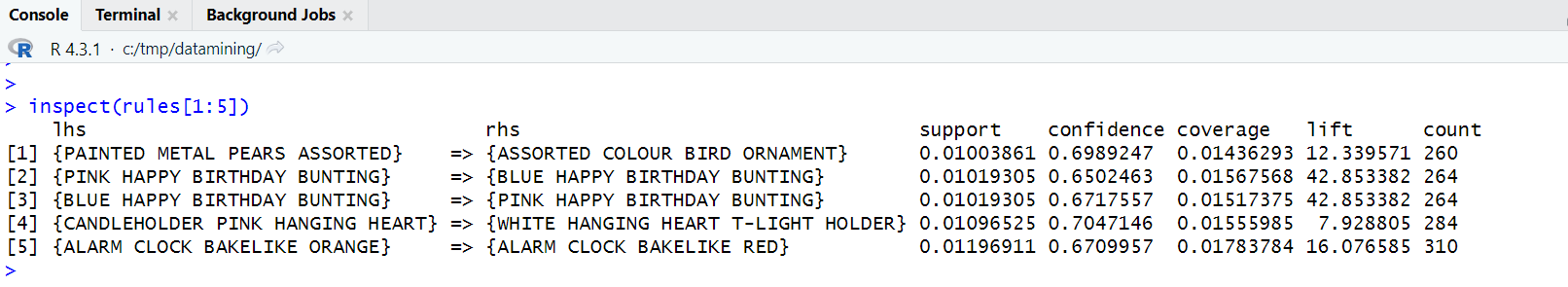
Dataset called 'Online Retail' taken from (Kaggle), is a comprehensive collection that comprised of transactional data from an e-commerce platform. It contained a variety of data types, including nominal, numeric, and temporal data points. The variables within this dataset include:



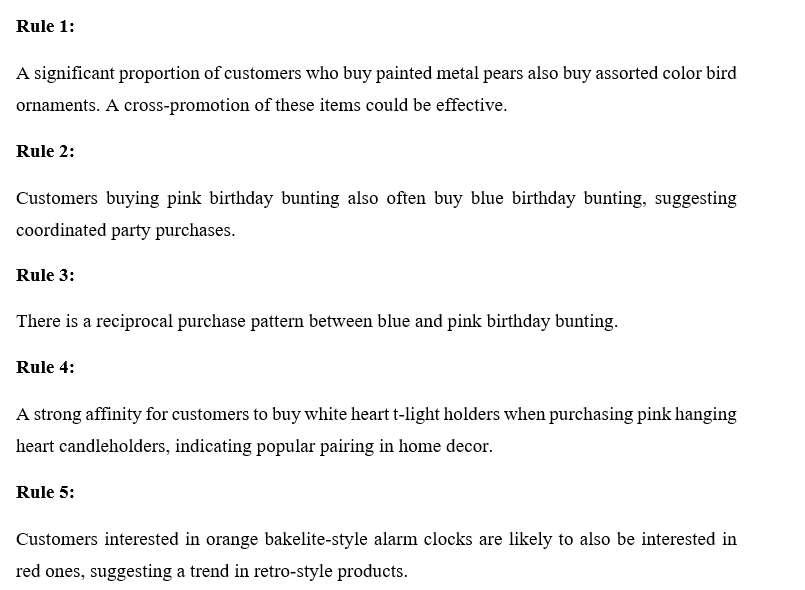
# **Potential insights**

Upon applying data mining tools to the 'Online Retail' dataset, the analysis yielded a variety of insights. It revealed purchasing patterns, highlighting products frequently bought together, which informed strategies for product placement, inventory bundling, and cross-selling. Customers were segmented based on purchasing behavior, aiding in the tailoring of marketing campaigns, personalization of discounts, and enhancement of customer retention strategies (Christy et al., 2021). The analysis identified seasonal trends, informing seasonal stock levels and promotional activities. Sales data analyzed by country provided geographical insights, which were utilized to customize marketing strategies for different regions and to optimize international supply chains. Additionally, a price sensitivity analysis was conducted, revealing how unit price changes affected the quantity sold, guiding the development of effective pricing strategies (Rajagopal, 2022).

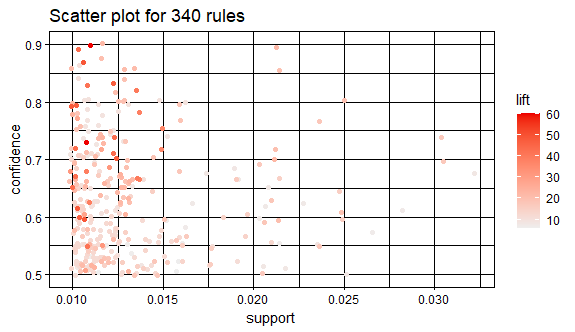
# **Association Rule Mining Interpretation**

Figure 2 presented below shows some of the rules and their insights obtained from R-Studio

**Figure 2. Rules [1] to [5] obtained in ARM (self-created using R-studio)**

****

In association rule mining, scatter plots visually represent key metrics such as support, confidence, and lift of mined rules (Chammas et al., 2023). They plot support on the x-axis and confidence on the y-axis, while color intensity or size indicates lift. This visualization aids in quickly identifying significant and reliable relationships between items in a dataset (Nwanganga and Chapple, 2020).



**Figure 3. Scatter plot of ARM (Self-created using R-Studio)**

Figure 3 illustrates a scatter plot visualizing the outcomes of association rule mining, highlighting significant connections between variables in a large dataset. The plot illustrated three critical metrics:

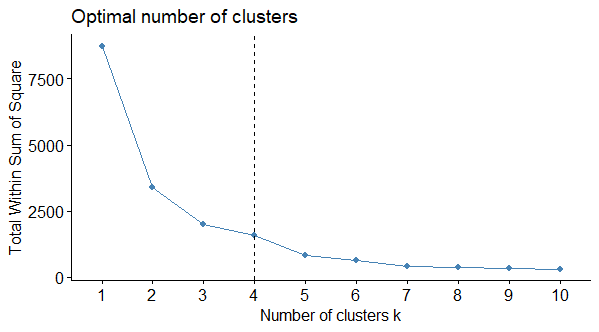
**Support**, shown on the x-axis, reflected the transaction proportion including the item set, ranging between 0.01 and 0.03. This indicated that these item sets featured in 1% to 3% of all transactions.

**Confidence**, displayed on the y-axis, measured the frequency of item Y being bought with item X (X → Y). The confidence levels in this plot varied from slightly above 0.5 to just below 0.9, signifying that in many cases, if X was bought, Y was also purchased 50% to 90% of the time.

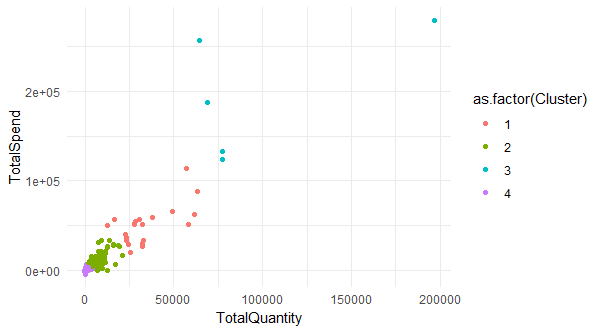
**Lift**, denoted by color intensity, assessed the strength of a rule. A lift value of 1 implied independent occurrence of items, above 1 indicated a positive association (items bought together more often than expected), and below 1 suggested a negative association. Darker dots signified higher lift values, ranging from 10 to 60, pointing to strong positive associations among item sets.

The scatter plot revealed that the association rule mining had successfully unearthed valuable patterns, with rules showing moderate to high confidence and diverse lift levels. Even rules with low support, thus less common, had high lift values, indicating potent but infrequent associations useful for cross-selling. This analysis underscored the effectiveness of association rule mining in retail for strategic product placement and promotions, enhancing sales by understanding customer buying behavior.

# **Clustering Interpretation**

The k-means clustering analysis presented in the above figures is a common technique used for this purpose. It divides the dataset into a pre-defined number of clusters, each represented by a centroid, which is essentially the average of all points in the cluster. The goal was to minimize the total variance within each cluster, thereby making the clusters as distinct as possible in terms of the variables considered, which in this case are **TotalSpend** and **TotalQuantity**.

**Figure 4. Elbow method (Self-created using R studio)**

Figure 4 is the elbow method which is used to identify the most appropriate number of clusters. The x-axis of the graph was labeled "Number of clusters k" and that ranges from 1 to 10. The y-axis was labeled "Total Within Sum of Square," which indicated the within-cluster sum of squares (a measure of variability within the clusters). Here, a vertical dashed line drawn at the point where k equals 3, suggests that the elbow method has determined that the optimal number of clusters for this particular dataset is 3, as there is a noticeable angle change in the plot around that point.

**Figure 5. Scatter plot of Clustering (Self-created using R studio)**

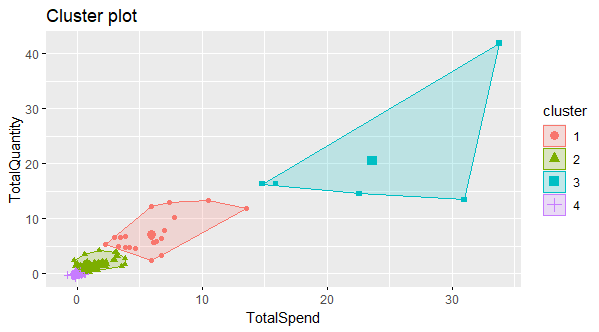
A shown in Figure 5, the scatter plot provided a past analysis of customer segments based on their total spending and the total quantity of purchases. Four distinct clusters, each indicated by a different color, were identified:

**Cluster 1 (Red):** This group consisted of customers who spent the least and purchased fewer items, likely representing occasional or low-volume shoppers.

**Cluster 2 (Green):** These customers showed moderately higher spending and quantity compared to Cluster 1, suggesting they were more regular shoppers or bought items with a slightly higher price point.

**Cluster 3 (Blue):** This cluster had customers with larger variations in spending and quantity, with some customers making significant purchases in terms of both spend and quantity, possibly indicating bulk buyers or those making substantial one-time purchases.

**Cluster 4 (Purple):** The most sparse group, these customers had exceptionally high spend and quantity levels, indicating they were likely the most valuable customers, potentially engaging in large-scale purchases or consistently choosing high-value items.

The plot highlighted a wide distribution of **TotalSpend** and **TotalQuantity**, with the majority of data points clustered towards the lower end of the scale and a few outliers indicating exceptionally high spending and purchasing behaviors. This analysis provided insights that could be critical for developing differentiated marketing strategies and personalized customer engagement based on the varying levels of spending and purchasing habits.

**Figure 6 Cluster plot (Self-created using R studio)**

As shown in Figure 6, the cluster plot visualized customer segments based on their total spending and total quantity of purchases, identifying four distinct clusters, each represented by a different color.

**Cluster 1 (Red):** This cluster contained customers with the lowest spending and quantity, indicating they were likely occasional shoppers or those who made small purchases.

**Cluster 2 (Green):** Customers in this cluster showed slightly higher levels of spending and quantity than those in Cluster 1, suggesting they might have purchased more frequently or opted for moderately priced items.

**Cluster 3 (Blue):** This cluster was the most sparsely populated, with customers demonstrating high levels of spending but not the highest quantity, possibly indicating purchases of high-value items but in smaller numbers.

**Cluster 4 (Purple):** Representing the smallest group, these customers had both high spending and high quantity, suggesting they were likely the most valuable customers, such as bulk buyers or those purchasing premium products.

Convex hulls were drawn around each cluster to indicate the general area containing all its members, thus providing a visual boundary. The plot suggested a progressive increase in both spending and quantity from Cluster 1 through Cluster 4. This analysis could help the business tailor marketing strategies and customer service to the different behaviors and needs of each customer segment.

# **Novelty and Significance of the Applications**

The combination of Association Rule Mining (ARM) and Clustering in the context of the 'Online Retail' dataset offers a novel and highly significant approach to retail data analysis.

The novelty lies in the sequential application of ARM followed by Clustering. By first applying ARM, it was identified that association rules reveal intricate purchase patterns and relationships between products, offering insights into what items are frequently bought together. This initial step is critical for understanding customer preferences and optimizing cross-selling strategies.

Subsequently, applying Clustering allows to segment customers into distinct groups based on their purchasing behavior and characteristics. This segmentation added a novel layer of personalization, as it tailored marketing and product recommendations to each customer group's preferences. This approach goes beyond traditional customer segmentation by incorporating the discovered association rules, making it unique.

The significance of this approach is profound. It will empower retailers to enhance customer experiences by offering personalized recommendations and promotions, ultimately increasing customer satisfaction and loyalty. Additionally, it optimized inventory management and product placement, leading to improved sales strategies and operational efficiencies. To sum up, the sequential application of ARM and Clustering represents a groundbreaking approach that can revolutionize how retailers leverage data to drive sales and improve customer engagement.

# **References**

CHAMMAS, M., BYERLY, S., LYNDE, J., MANTERO, A., SABERI, R., GILNA, G., PUST, G. D., RATTAN, R., NAMIAS, N., CRANDALL, M. & YEH, D. D. 2023. Association Between Child Access Prevention and State Firearm Laws With Pediatric Firearm-Related Deaths. *The Journal of surgical research,* 281**,** 223-227.

CHEN, C.-W., TSAI, C.-F., TSAI, Y.-H., WU, Y.-C. & CHANG, F.-R. 2020. Association rule mining for the ordered placement of traditional Chinese medicine containers: An experimental study. *Medicine (Baltimore),* 99**,** e20090-e20090.

CHRISTY, A. J., UMAMAKESWARI, A., PRIYATHARSINI, L. & NEYAA, A. 2021. RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University. Computer and information sciences,* 33**,** 1251-1257.

GUPTA, A. K. & SHUKLA, A. V. 2019. Online Retail Format Choice Behavior of Indian Customers for Reasoned Purchase: A Cultural Perspective. *Journal of international consumer marketing,* 31**,** 469-491.

KAGGLE. Available: <https://www.kaggle.com/datasets/puneetbhaya/online-retail> [Accessed 02-01-2024].

MANDAL, P., BASU, P. & SAHA, K. 2021. Forays into omnichannel: An online retailer’s strategies for managing product returns. *European journal of operational research,* 292**,** 633-651.

NWANGANGA, F. C. & CHAPPLE, M. 2020. *Practical machine learning in R / Fred Nwanganga, Mike Chapple,* Indianapolis, John Wiley and Sons.

PRIOR, D. D. 2023. *B2B customer engagement strategy : an introduction to managing customer experience / Daniel D. Prior,* Cham, Switzerland, Macmillan Palgrave.

RAJAGOPAL 2022. *Agile marketing strategies : new approaches to engaging consumer behavior / Rajagopal,* Cham, Switzerland, Springer.

RHYS, H. 2020. *Machine Learning with R, the tidyverse, and mlr / Hefin I. Rhys,* Shelter Island, New York, Manning.

SHAHBAZOVA, S. N., KACPRZYK, J., BALAS, V. E. & KREINOVICH, V. 2021. Customer and Product Clustering in Retail Business. Switzerland: Springer International Publishing AG.