Customer Purchase Behavior Analysis

Abdullah Yassine

Spring 2025

1 Introduction

In the age of e-commerce, understanding customer purchasing behavior is critical for businesses looking to improve product recommendations, maximize revenue, and enhance customer loyalty. Retailers thrive when they can anticipate what customers want to buy next or group customers into segments for targeted marketing. This project focuses on uncovering patterns in customer transactions through frequent itemset mining and customer segmentation techniques. The insights generated could improve real-world business strategies such as personalized marketing, product bundling, and loyalty program design.

Personally, I found this project exciting because it combines two important areas of data science — pattern mining and clustering — both of which have immediate business value. Using real transactional data to model customer behavior mirrors problems faced by modern companies like Amazon, Target, and Etsy.

2 Data Description

The dataset used in this project is the **Online Retail Dataset** from the **UCI Machine Learning Repository**. It contains approximately 541,909 rows of transactional data for a UK-based online retailer during 2010–2011. Each row represents a product purchased as part of a single invoice.

Dataset Columns

- InvoiceNo: Unique invoice number

- StockCode: Unique product code
- Description: Product description
- Quantity: Number of units purchased
- InvoiceDate: Date and time of transaction
- UnitPrice: Price per unit
- CustomerID: Unique identifier for each customer
- Country: Country of customer

The raw data required significant cleaning, including removing missing CustomerIDs, excluding canceled transactions (invoices starting with 'C'), and filtering out negative quantities and prices.

3 Real-World Context

Customer segmentation and product recommendation are key techniques widely used by companies like Amazon and Target.

- **Amazon** uses collaborative filtering to recommend products based on customer behavior.
- **Target** leverages purchase patterns for targeted marketing.

Academic studies such as *Market Basket Analysis with Apriori Algorithm* highlight the effectiveness of association rule mining for boosting cross-sells and bundle sales.

4 Exploratory Analysis

Initial analysis showed certain products were extremely popular, such as *WHITE HANGING HEART T-LIGHT HOLDER* and *JUMBO BAG RED RETROSPOT.*

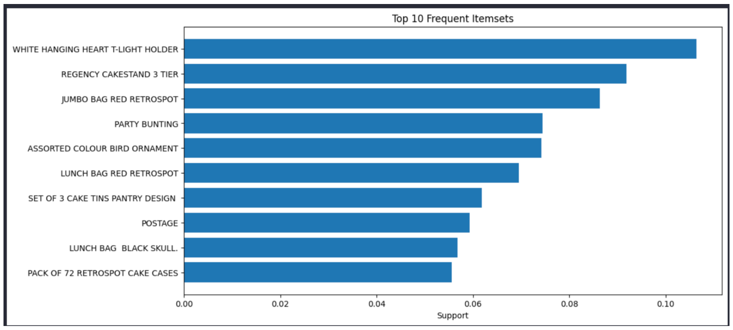

Figure 1: Top 10 Individual Products by Support

Description of Figure:

Figure 1 shows the **Top 10 Most Frequently Purchased Items** in the dataset, ranked by their *support* values. In this context, **support** measures the proportion of all transactions that contain a given product.

Key observations:

- **WHITE HANGING HEART T-LIGHT HOLDER** is the most purchased item, appearing in approximately 10% of all transactions.
- Other highly purchased items include:
  - *REGENCY CAKESTAND 3 TIER*
  - *JUMBO BAG RED RETROSPOT*
  - *PARTY BUNTING*
- All top 10 items have a support value greater than 5%, indicating strong popularity across customers.

Business implications:

- These products can be used as *anchor items* in promotions or bundled offers.
- Inventory management should prioritize keeping these high-demand items in stock.

- Marketing campaigns could highlight these products to attract broader customer interest.

Overall, this plot highlights a small set of products that contribute disproportionately to sales volume. Understanding which items are consistently purchased helps businesses optimize product placement, targeted promotions, and supply chain decisions.

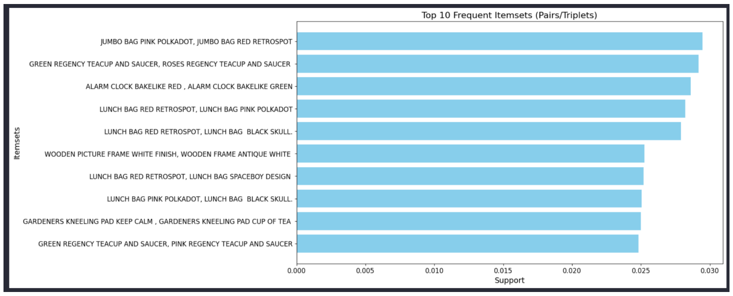

Figure 2: Top Frequent Pairs and Triplets by Support

Description of Figure:

Figure 2 shows the **Top 10 Most Frequent Itemsets** involving *pairs* and *triplets* of products. Here, **support** represents the proportion of all transactions that contain all items within the itemset together.

Key observations:

- The most common co-purchased items are:
  - *JUMBO BAG PINK POLKADOT* and *JUMBO BAG RED RETROSPOT*
  - *GREEN REGENCY TEACUP AND SAUCER* and *ROSES REGENCY TEACUP AND SAUCER*
  - *ALARM CLOCK BAKELIKE RED* and *ALARM CLOCK BAKELIKE GREEN*
- Many of the top itemsets consist of similar or complementary products (e.g., different styles of lunch bags, matching home decor items).

- The support values for these itemsets are lower than individual items (as expected) but still significant, exceeding 2%.

Business implications:

- Bundling these frequently bought-together items could increase average cart size and revenue.
- Marketing strategies could suggest complementary products during checkout.
- Product recommendations based on these combinations would likely have high success rates, improving customer satisfaction and loyalty.

Overall, this plot reveals natural groupings of products that customers tend to purchase together, offering valuable guidance for bundling, cross-selling, and promotion strategies.

5 Methods

5.1 Frequent Itemset Mining

The Apriori algorithm was used with a minimum support threshold of 2%. Association rules were generated where lift was greater than 1. Apriori is efficient because it expands itemsets level by level, pruning combinations that are not frequent early on.

5.2 Customer Segmentation

RFM features were engineered:

- Recency: Days since last purchase
- Frequency: Number of purchases
- Monetary: Total amount spent

Features were standardized, and clustering was performed using:

- **K-Means Clustering** (hard assignment)
- **Gaussian Mixture Models (GMM)** (soft probabilistic assignment)

5.3 Gaussian Mixture Model Formula

The probability model for GMM is:

$$p(x) = \sum_{k=1}^{K} \phi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$

where:

- ϕ_k are mixing coefficients
- μ_k are mean vectors
- Σ_k are covariance matrices

6 Results

6.1 Frequent Itemsets and Rules

Customers frequently bought:

- Multiple lunch bags (different styles)
- Matching tea cups and saucers
- Bundles of reusable shopping bags

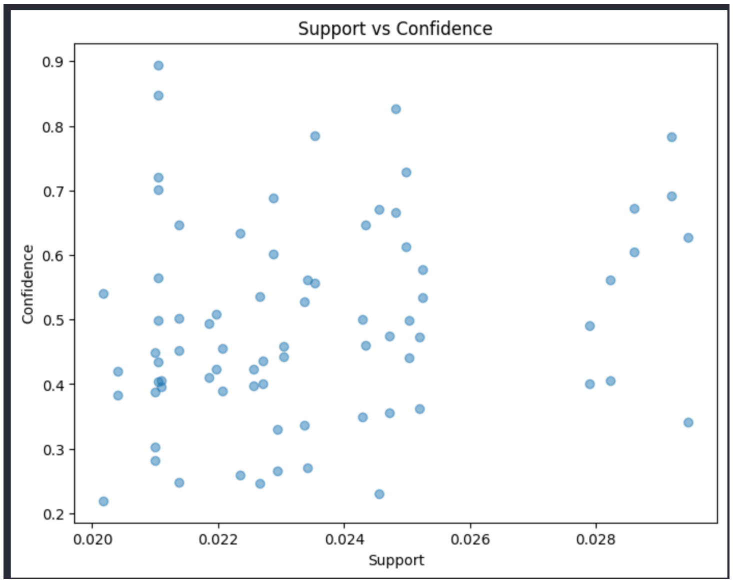

Figure 3: Association Rules: Support vs Confidence

Description of Figure:

Figure 3 presents a **scatter plot of Support versus Confidence** for the association rules generated from frequent itemsets. Each point on the graph represents a single association rule between items.

Understanding the axes:

- **Support (x-axis):** The proportion of all transactions that contain both the antecedent and consequent items. Higher support values indicate that the rule involves more commonly purchased items.
- **Confidence (y-axis):** The probability that a transaction containing the antecedent item(s) also contains the consequent item(s). Higher confidence means the rule is more reliable for making predictions.

Key observations:

- Most rules have support values ranging between 0.02 and 0.03, meaning they apply to 2%–3% of all transactions.

- Confidence values are more widely spread, ranging from 0.2 to over 0.9.
- There is no obvious linear relationship between support and confidence. Some rules have high confidence but relatively low support, indicating they are very reliable but occur less frequently.
- Densely packed points around confidence 0.4–0.5 suggest that many rules have moderate reliability.

Business implications:

- High-confidence rules, even with moderate support, are valuable for personalized recommendations because they represent strong purchasing patterns.
- Rules with both high support and high confidence are ideal candidates for automated marketing strategies such as ”Frequently Bought Together” promotions.
- Understanding the trade-off between support and confidence helps businesses balance between recommending highly reliable rules versus more popular item combinations.

Overall, this plot provides a comprehensive view of the reliability and prevalence of various product association rules, enabling better decision-making for recommendation systems and targeted offers.

6.2 Customer Segmentation: KMeans Clusters

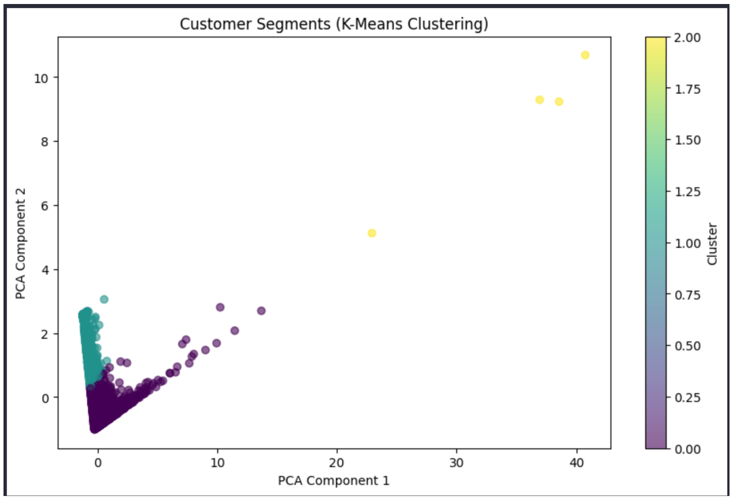

Figure 4: KMeans Customer Clusters (PCA Projection)

Description of Figure:

Figure 4 shows the **K-Means Clustering Results** visualized using a **2D Principal Component Analysis (PCA) projection**. Each point represents a customer, colored according to the cluster assignment determined by the K-Means algorithm.

Understanding the axes:

- **PCA Component 1** and **PCA Component 2** are synthetic features created by PCA to capture the maximum variance in the original three-dimensional RFM data (Recency, Frequency, Monetary).
- These components are linear combinations of the original features and allow for easier 2D visualization while preserving the general structure of the data.

Key observations:

- Three distinct clusters are visible, suggesting that customers naturally group into three main behavioral segments.

- The yellow cluster (Cluster 2) is noticeably separated, representing customers who are very different from the majority. These could be extremely high-value buyers (very frequent purchases, high spending).
- The dark purple and teal clusters (Clusters 0 and 1) represent customers who are closer together but still distinct in terms of their purchasing patterns.
- Some customers (points) are more spread out, indicating variability even within clusters, while most customers are densely packed near the center.

Business implications:

- The identification of clear clusters allows businesses to create tailored marketing strategies:
  - High-value customers (yellow cluster) can be targeted with loyalty rewards and premium offers.
  - Moderate customers can be encouraged to spend more through promotions.
  - Low-engagement customers could receive reactivation campaigns or targeted discounts.
- Understanding how customers group naturally helps prioritize marketing budget and resources effectively.

Overall, this plot demonstrates that customer behavior is not random but forms structured patterns that businesses can leverage to drive personalized engagement and improve lifetime value.

KMeans Clusters Summary:

- Cluster 2: Super buyers (extremely high frequency and monetary)
- Cluster 0: Loyal customers (good frequency and spending)
- Cluster 1: At-risk customers (long recency, low frequency)

6.3 Customer Segmentation: GMM Clusters

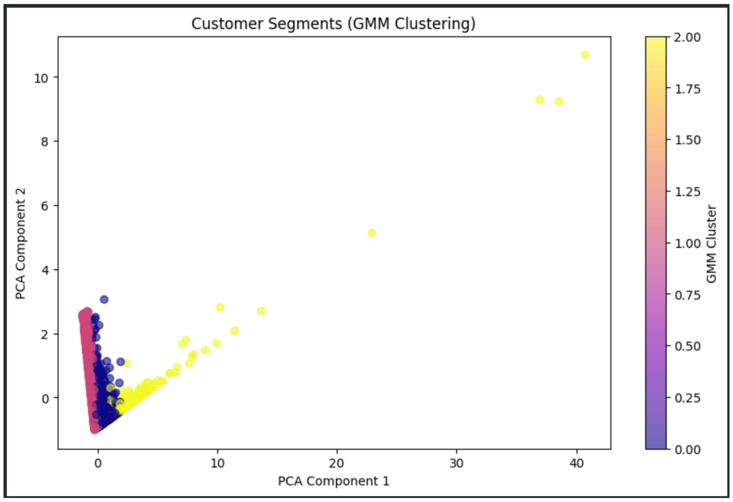

Figure 5: GMM Customer Clusters (PCA Projection)

Description of Figure:

Figure 5 displays the **Gaussian Mixture Model (GMM) Clustering Results** visualized using a **2D PCA projection**. Each point represents a customer, colored by their GMM-assigned cluster membership based on Recency, Frequency, and Monetary (RFM) features.

Understanding the axes:

- **PCA Component 1** and **PCA Component 2** are principal components derived from the original RFM variables.
- These components summarize the variation in customer behavior into two dimensions, making it possible to visualize customer groupings in 2D.

Understanding GMM clustering:

- Unlike K-Means, where each customer belongs to exactly one cluster, GMM assigns customers a probability of belonging to each cluster.

- Customers near the boundaries between clusters might have a high probability for multiple clusters, resulting in softer or overlapping cluster boundaries.

Key observations:

- Three main clusters are identified, matching the number found through K-Means, confirming the natural structure of the data.
- The yellow-colored cluster (Cluster 2) consists of customers that are extreme in purchasing behavior — possibly super-buyers or very high-frequency purchasers.
- The darker clusters (Clusters 0 and 1) represent the majority of customers, but some overlap is more visible compared to the K-Means result, showing how customer behavior is sometimes mixed between groups.
- There are a few highly dispersed points, indicating customers whose behavior deviates significantly from the general population.

Business implications:

- GMM clustering provides a probabilistic view of customer segmentation, allowing businesses to handle ”in-between” customers more flexibly.
- Customers with split probabilities between clusters could receive mixed marketing strategies tailored to their diverse behaviors.
- Identifying the distinct super-buyer cluster helps focus loyalty programs and premium offerings on the most valuable customers.

Overall, this GMM clustering visualization demonstrates that customer purchasing behavior follows underlying patterns but is not strictly divided. Soft clustering allows for more nuanced and flexible customer engagement strategies in real-world business applications.

GMM clustering produced similar findings to KMeans, confirming the customer segments’ robustness.

7 Conclusion

This project successfully applied frequent itemset mining and customer clustering techniques to extract valuable insights from e-commerce transactional data. Through the Apriori algorithm, we identified strong product associations, such as bundles of lunch bags and coordinated home decor items. Clustering customers based on Recency, Frequency, and Monetary (RFM) behavior revealed distinct customer segments: **super-buyers**, **loyal regulars**, and **at-risk customers**.

Key Takeaways:

- **Frequent itemset mining** highlights cross-sell and bundle opportunities that could immediately boost revenue and improve customer satisfaction.
- **Customer segmentation** provides a foundation for more effective marketing campaigns, allowing businesses to tailor promotions based on customer value and engagement level.
- The combination of market basket analysis and RFM clustering demonstrates how traditional data science methods can deliver actionable business strategies.

Potential Future Work

Looking ahead, several extensions could deepen the business and data science impact:

- **Seasonality Analysis:** Investigate how purchasing patterns vary across seasons, holidays, or sales events, enabling time-sensitive promotions.
- **Predictive Customer Churn Modeling:** Build classification models (e.g., Logistic Regression, Random Forests, XGBoost) to predict which customers are at risk of leaving, allowing proactive re-engagement strategies.

- **Dynamic Recommendation Systems:** Move beyond frequent itemsets by implementing more sophisticated recommendation models such as Collaborative Filtering, Matrix Factorization, or Neural Network-based recommendation engines.

- **Customer Lifetime Value (CLV) Modeling:** Estimate the future value of a customer using regression or survival analysis models to optimize long-term marketing investment.

- **Deep Learning for Pattern Discovery:** Use autoencoders or embedding techniques (e.g., Word2Vec for products) to uncover hidden structures in purchase behavior beyond simple item combinations.

Final Thoughts: By better understanding customer behavior and purchasing relationships, businesses can personalize the customer experience, increase customer retention, and drive sustainable revenue growth. Integrating more advanced machine learning models would allow for even more precise, dynamic, and scalable insights, ensuring businesses stay competitive in the rapidly evolving world of digital commerce.