Executive Summary: E-Commerce Analytics & Data Mining

Project Title: Economic analysis of Brazilian Dataset 2011

Date: December 2, 2025

---

1. Project Overview & Objectives

This project aimed to apply a full data mining pipeline—from ETL (Extract, Transform, Load) to storytelling—on a transactional e-commerce dataset. The primary objective was to uncover hidden patterns in customer behavior and sales trends to inform decision-making regarding inventory planning, customer retention, and cross-selling strategies.

Key Business Questions:

- How can we segment customers to improve marketing targeting?
- Which products are frequently bought together (bundling opportunities)?
- What are the seasonal trends affecting revenue?

---

2. Methodology: ETL & Data Preparation

Before analysis, the raw dataset underwent rigorous cleaning and transformation:

- **Data Cleaning:** Removed duplicate records and handled missing values in critical fields like `CustomerID` and `Description`.
- **Data Integrity:** Filtered out invalid transactions, specifically negative quantities (returns) and zero unit prices to ensure analytical accuracy.
- **Feature Engineering:** Created new features including `TotalSpend` (Quantity × UnitPrice) and extracted date components (Month, Year) for time-series analysis.
- **Transformation:** Applied scaling (MinMax) for clustering algorithms and One-Hot Encoding for Association Rule Mining.

Tools Used: Python, Pandas, Scikit-learn, Mlxtend, Matplotlib/Seaborn.

---

3. Data Mining Techniques Applied

We utilized two primary unsupervised learning techniques to extract insights:

1. **K-Means Clustering (Customer Segmentation):**

    - We applied the **RFM (Recency, Frequency, Monetary)** model to quantify customer value.
    - Using the Silhouette Score method, we determined the optimal number of clusters to be **4**.
    - This grouped customers based on their purchasing habits rather than static demographics.
2. **Association Rule Mining (Market Basket Analysis):**
    - We utilized the **FP-Growth Algorithm** (an optimized alternative to Apriori) to identify frequent itemsets.
    - We generated rules based on a minimum support of 2% and a confidence threshold of 30% to find strong product relationships.

---

4. Key Findings & Insights

A. Customer Segmentation (The 4 Personas) Our clustering analysis revealed four distinct customer groups:

- **Cluster 0 (At-Risk/Low Value):** Customers who purchased long ago and spent little. They represent the largest segment but the lowest revenue.
- **Cluster 1 (VIPs):** High frequency and high monetary value. These are the most loyal and profitable customers.
- **Cluster 2 (New/Developing):** Recent buyers with low frequency. They have high potential to be converted into loyalists.
- **Cluster 3 (Big Spenders):** Customers who buy infrequently but make large purchases when they do.

B. Market Basket Trends

- Strong associations were found between homeware items. For example, customers buying **"Green Regency Teacup and Saucer"** are highly likely to buy **"Roses Regency Teacup and Saucer"**.
- We identified several rules with a **Lift > 1**, indicating that these items are statistically more likely to be bought together than separately.

C. Seasonal Trends

- The data indicates a clear seasonal spike in **November**, likely driven by holiday shopping, followed by a drop in subsequent months.

---

5. Strategic Recommendations

Based on the data mining results, we propose the following actions:

1. **Launch a Tiered Loyalty Program:** Target **Cluster 1 (VIPs)** with exclusive early access to products to maintain retention. For **Cluster 0 (At-Risk)**, implement automated "We Miss You" email campaigns with discount codes to re-engage them.

2. **Implement Product Bundling:** Create physical bundles for high-lift item pairs (e.g., selling Teacup sets together) at a 5-10% discount. This will increase the Average Order Value (AOV).
3. **Inventory Optimization:** Increase stock levels by **20% in October** to prepare for the November sales peak identified in the time-series analysis, preventing stockouts during critical revenue periods.