Analyzing Product Recommendations in E-Commerce for Delivered Korea

Anthony Fang

San Francisco State University

BUS 895: Master's Thesis

Fall 2024

Abstract

This study focuses on optimizing product recommendations and analyzing customer purchase behavior using transactional data from Delivered Korea, an international logistics company specializing in shipping Korean products worldwide. Utilizing Shopify datasets with over 114,000 transactions, the research explores sales trends, customer preferences, and the effectiveness of recommendation algorithms. The analysis segments data geographically, categorizing shipping destinations into continents and consolidating product categories for simplified yet meaningful insights.

Results highlight the importance of high-lift category pairs, revealing actionable insights for targeted marketing and customer engagement. While the predictive model achieved an accuracy of 79% for overall purchase predictions, category-specific performance varied, suggesting room for improvement through integrated machine learning approaches. Future efforts will aim to refine thresholds for confidence and support in association rules and experiment with hybrid recommendation systems to enhance accuracy and customer satisfaction.

Table of Contents

Introduction

This study examines the business challenges faced by Delivered Korea, an international logistics company specializing in shipping Korean products to customers worldwide. Using a comprehensive dataset sourced from Shopify, which includes transactional details such as product categories, shipping destinations, and customer purchase patterns, this research aims to uncover actionable insights.

The primary objective is to analyze sales trends, identify popular product categories, and explore geographic variations in purchasing behavior. By leveraging customer purchase history and applying advanced analytical techniques, the study seeks to recommend products effectively, enhance customer satisfaction, and support the company in optimizing its market strategies. This research also evaluates the relationships between product categories and identifies patterns that can inform personalized marketing and operational decision-making.

Literature Review and Research Questions

Effective Product Recommendations in E-Commerce and Logistics

In the domain of e-commerce and logistics, the effective recommendation of products has become a critical business strategy, enhancing customer satisfaction and operational efficiency. Prior research in recommendation systems often focuses on leveraging historical customer purchase data to identify patterns and predict future preferences.

Association Rule Mining and the Apriori Algorithm

Methods such as association rule mining, commonly operationalized through the Apriori algorithm, have demonstrated strong capabilities in uncovering relationships between product categories (Agrawal & Srikant, 1994). Metrics such as support, confidence, and lift are widely used to evaluate the strength of these associations and their potential for actionable recommendations.

Applications of Market Basket Analysis

Studies in market basket analysis have highlighted its application in identifying frequently co-purchased items, which can guide inventory management and targeted marketing campaigns (Hahsler et al., 2007). Furthermore, research emphasizes the importance of category consolidation for simplifying complex datasets and enhancing interpretability, particularly when analyzing large-scale transactional data (Lin et al., 2020).

Geographic Segmentation in Sales Analysis

In cross-sectional analyses, grouping geographic regions into continents has been shown to reveal broader sales trends and regional preferences, aiding in strategic decision-making for global businesses.

Advancements in Predictive Modeling

Predictive modeling in e-commerce has evolved, with hybrid approaches combining association rules and machine learning algorithms to improve recommendation accuracy (Zhang et al., 2018). While high overall prediction accuracy (e.g., 79% in this study) indicates the effectiveness of existing methodologies, category-specific performance variations underscore the need for further refinement and adaptation to unique customer behaviors.

Research Questions

This thesis aims to explore the challenges and opportunities in product recommendations for Delivered Korea, leveraging customer transactional data to enhance marketing strategies and operational efficiency. The study focuses on the following key research questions:

Primary Research Question

How can Delivered Korea recommend products effectively to customers based on their purchase history and preferences?

This question addresses the overarching goal of the study: optimizing product recommendations to improve customer satisfaction and increase sales by leveraging transactional data and advanced analytical techniques.

Secondary Research Questions

1. **Identifying Popular Product Categories.**

What are the most popular product categories across different customer segments?

This question explores customer preferences and category-level trends by analyzing sales volumes and identifying top-performing categories (e.g., "K-POP," "Fashion," "Beauty"). Understanding these preferences can help Delivered Korea prioritize inventory and promotional strategies.

2. **Geographic Variations in Purchasing Behaviors.**

How do purchasing behaviors vary geographically, and what insights can be drawn from region-specific trends?

By segmenting the shipping data into continents and analyzing category-level demand across regions, this question investigates cultural and logistical influences on purchasing behavior. For example, identifying the dominant product categories in North America compared to Asia.

3. **Strongest Product Category Associations.**

Which product category pairs exhibit the strongest associations, and how can these insights guide targeted recommendations?

Using metrics such as lift, confidence, and leverage, this question focuses on uncovering relationships between product categories. These insights inform cross-selling strategies, such as pairing "K-POP" items with "Entertainment" for customers likely to purchase both.

4. **Effectiveness of Geographic Segmentation.**

What is the effectiveness of geographic segmentation in improving sales recommendations?

This question evaluates the impact of consolidating shipping destinations into six continents. It examines whether geographic segmentation enhances the interpretability and accuracy of recommendations compared to analyzing data at the country level.

5. **Accuracy of Predictions and Model Improvement.**

How accurate are predictions generated through association rule mining and related algorithms, and how can their performance be improved?

This question delves into the predictive power of the Apriori algorithm and evaluates the model's overall and category-specific accuracy. The analysis includes the identification of factors leading to false positives or negatives and suggestions for refining model parameters, such as lowering confidence and support thresholds.

6. **Influence of Customer Purchase Patterns.**

How do customer purchase patterns (e.g., single-category vs. multi-category buyers) influence product recommendations?

By analyzing customer segmentation (e.g., 11,363 customers buying a single category vs. 6,552 buying multiple categories), this question explores how purchase patterns affect the design and success of recommendation systems.

7. **Implications of Removing Low-Frequency Categories.**

What are the implications of removing low-frequency categories (e.g., "Other") for simplifying analysis and improving recommendations?

This question examines the trade-off between simplifying the dataset and retaining potentially meaningful insights. It evaluates whether excluding less frequent categories improves model performance or limits recommendation diversity.

8. **Enhancing Business Relevance with Lift and Purchase Frequency.**

How can the lift metric combined with purchase frequency enhance business relevance for category pair analysis?

This question investigates the dual role of lift and counts in identifying category pairs with both strong associations and high purchase volumes. For example, high-lift pairs with substantial counts are prioritized for marketing campaigns.

9. **Optimizing Thresholds in Association Rule Mining.**

What are the optimal thresholds for confidence and support in association rule mining to balance prediction accuracy and interpretability?

This question seeks to refine the association rule model by determining the most effective parameters for identifying actionable rules, particularly for categories with fewer transactions.

10. **Performance Variations Across Market Sizes.**

How does the performance of recommendation models vary between high-demand regions like North America and Europe compared to smaller markets?

This question evaluates the scalability and applicability of recommendations across different market sizes, identifying opportunities to adapt strategies for underrepresented regions.

Methodology and Analysis

Methodological Framework

This chapter outlines the modeling approach used to analyze transactional data from Delivered Korea and provide actionable product recommendations. The study employs both descriptive and predictive methodologies to uncover patterns in customer purchase behavior and build a recommendation system.

Data Preparation

Before modeling, the dataset underwent significant preprocessing steps:

Cleaning and Aggregation

Unnamed and irrelevant columns were removed.

Missing values in key variables such as "category" and "shipping country" were handled to preserve data integrity.

Cross-Sectional Transformation

The panel dataset, which recorded multiple transactions per customer, was transformed into a cross-sectional format.

Categories were consolidated into eight main groups (e.g., combining "Fashion" and "Clothing" into "Fashion"), and shipping countries were aggregated into six continents to simplify geographic analysis.

Feature Engineering

Model Description

The dataset was prepared for analysis by encoding categorical variables (e.g., product categories) and standardizing numerical features (e.g., shipping fees).

The study integrates association rule mining and predictive modeling to achieve its objectives:

Association Rule Mining

The Apriori algorithm is used to uncover relationships between product categories, identifying frequently co-purchased items. Key metrics include: **Support:** Measures the frequency of an item or itemset in the dataset. **Confidence:** Indicates the likelihood of purchasing one product given another is purchased. **Lift:** Evaluates the strength of an association relative to random chance, with lift values greater than 1 suggesting a meaningful relationship.

The association rules form the foundation for cross-selling strategies by highlighting category pairs with high lift and substantial purchase counts.

Predictive Modeling

A binary classification model is developed to predict customer purchase behavior for specific product categories. The modeling process includes: Converting purchases into binary outcomes (e.g., 1 if purchased, 0 otherwise).Training the model using 70% of the data and validating it on the remaining 30%.Evaluating model performance using metrics such as overall accuracy, category-wise accuracy, true positives, false positives, and false negatives.

The model identifies potential product recommendations for individual customers based on past purchase behavior and learned patterns.

The models were implemented in the following steps:

Association Rule Mining with the Apriori Algorithm

Parameters:

Minimum support: Adjusted to retain meaningful associations.

Confidence threshold: Set to 30%, ensuring rules had a reasonable predictive value.

Lift threshold: Values above 1.0 were prioritized for identifying strong relationships.

Example: If customers who purchase "K-POP" frequently also buy "Entertainment," a rule like "If K-POP, then Entertainment" is generated.

Predictive Model Training

The training phase focused on identifying patterns in customer purchase behavior.

Example: A rule may predict that if a customer buys "Fashion," they are likely to purchase "Beauty" products as well.

Evaluation Metrics

- **Holistic Accuracy:** The overall accuracy of predictions across all categories (79% in this study).
- **Category-Specific Accuracy:** Analyzed the model's performance in predicting purchases for individual categories, with high demand categories (e.g., "K-POP") showing stronger results compared to less frequent categories.
- **Confusion Matrix Analysis:**
    - **True Positives (TP):** Correctly predicted purchases.
    - **False Positives (FP):** Incorrectly predicted purchases.
    - **True Negatives (TN):** Correctly identified non-purchases.
    - **False Negatives (FN):** Missed purchase predictions.

Limitations and Future Enhancements

While the model achieved promising results, several limitations were identified:

- **Category Imbalance:** Categories with fewer transactions (e.g., "Food & Health") were more challenging to predict accurately.
- **Simplified Geography:** Aggregating shipping countries into continents reduced granularity, potentially overlooking regional variations.
- **Static Rules:** The association rules are static and do not adapt to real-time changes in customer behavior.

Future enhancements could include:

- **Lowering thresholds:** Adjust thresholds for confidence and support to capture more subtle patterns.

- **Incorporating machine learning** : Leverage collaborative filtering or hybrid recommendation approaches to improve prediction accuracy and personalization.

Data and Analysis Procedures

Data Understanding

The dataset used in this study originates from Shopify and consists of transactional data for Delivered Korea, capturing essential details about customer purchases, shipping destinations, and product categories. Below is an in-depth understanding of the data structure and characteristics:

Dataset Overview

Source: The dataset is a CSV file titled "DK Product Recommendation Data_2024_1107.csv" (Figure 1)and includes transaction-level data collected from various stores on the Shopify platform.

```
DK Product Recommendation Data
File  Edit  View

suiteNO,packageNO,date,shippingFee,shippingcounty,storename,category,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
D1234,146659,1/1/2023 0:58,200.14,United States,Weverse Global Shop,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
D3654,141004,1/1/2023 1:46,50.86,Brunei Darussalam,Weverse Shop,Game/Toy/Book,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
D3654,144729,1/1/2023 1:46,38.19,Brunei Darussalam,Weverse Shop,Stationery/Photocard/Photo album,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
D3654,149472,1/1/2023 1:46,50.86,Brunei Darussalam,Weverse Shop,Stationery/Photocard/Photo album,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B4702,146875,1/1/2023 2:31,330.21,United States,Weverse Shop,Albums/CD/DVD,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B4702,146875,1/1/2023 2:31,330.21,United States,Weverse Shop,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B4702,146875,1/1/2023 2:31,330.21,United States,Weverse Shop,Game/Toy/Book,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B4702,146875,1/1/2023 2:31,330.21,United States,Weverse Shop,Stationery/Photocard/Photo album,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B4702,146875,1/1/2023 2:31,330.21,United States,Weverse Shop,Clothing/Shoes/Jewelry/Bag,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B4702,146875,1/1/2023 2:31,330.21,United States,Weverse Shop,Computer/PC Accessories,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B2026,145363,1/1/2023 2:47,79.98,Singapore,Weverse,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
E2818,150506,1/1/2023 3:07,62.63,United States,WEVERSE,Stationery/Photocard/Photo album,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
D4664,149432,1/1/2023 4:36,62.63,United States,Namoo Shop,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A2641,133525,1/1/2023 7:51,298.76,Mexico,Weverse ,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A2641,133525,1/1/2023 7:51,298.76,Mexico,Weverse ,Clothing/Shoes/Jewelry/Bag,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A2641,133525,1/1/2023 7:51,298.76,Mexico,Weverse ,Stationery/Photocard/Photo album,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A2641,134803,1/1/2023 7:51,309.83,Mexico,Weverse ,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A2641,134803,1/1/2023 7:51,309.83,Mexico,Weverse ,Stationery/Photocard/Photo album,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A2641,134803,1/1/2023 7:51,309.83,Mexico,Weverse ,Clothing/Shoes/Jewelry/Bag,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A2641,147469,1/1/2023 7:51,298.76,Mexico,Weverse ,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A2641,147469,1/1/2023 7:51,298.76,Mexico,Weverse ,Albums/CD/DVD,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
E2144,149048,1/1/2023 8:13,171.16,United States,Weverse shop,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A5034,147903,1/1/2023 8:16,258.47,France,Weverse Shop,Albums/CD/DVD,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A5034,147903,1/1/2023 8:16,258.47,France,Weverse Shop,Game/Toy/Book,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
C0934,121324,1/1/2023 8:39,134.92,United States,seller,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B9837,146918,1/1/2023 8:47,40.01,United States,bunjang,Albums/CD/DVD,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B9837,121454,1/1/2023 8:49,51.35,United States,bunjang,Albums/CD/DVD,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
C1019,148392,1/1/2023 9:47,163.2,Bahrain,Weverse shop,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
C1019,148392,1/1/2023 9:47,163.2,Bahrain,Weverse shop,Albums/CD/DVD,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B0118,148488,1/1/2023 10:16,119.12,India,SB be keyringx4,Home/Kitchen/Supplies,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B0118,148488,1/1/2023 10:16,119.12,India,SB be keyringx4,Game/Toy/Book,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B0118,148488,1/1/2023 10:16,119.12,India,SB be keyringx4,Stationery/Photocard/Photo album,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B0118,148488,1/1/2023 10:16,119.12,India,SB be keyringx4,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B0546,139792,1/1/2023 12:31,28.41,United Kingdom,Weverse shop,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
C1455,146876,1/1/2023 14:29,119.24,United Kingdom,Weverse Shop,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
C1455,146876,1/1/2023 14:29,119.24,United Kingdom,Weverse Shop,Stationery/Photocard/Photo album,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
E4183,147872,1/1/2023 15:11,13.8,Japan,Bunjang,Electronics,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A2641,147470,1/1/2023 17:11,298.76,Mexico,Weverse ,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A2641,147470,1/1/2023 17:11,298.76,Mexico,Weverse ,Albums/CD/DVD,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A2641,147471,1/1/2023 17:11,309.83,Mexico,Weverse shop,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A2641,147471,1/1/2023 17:11,309.83,Mexico,Weverse shop,Albums/CD/DVD,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
A2641,147472,1/1/2023 17:11,287.7,Mexico,Weverse shop ,Albums/CD/DVD,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
E2161,146587,1/1/2023 17:33,107.78,United States,WEVERSE,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
E1839,147466,1/1/2023 18:14,158.27,United States,weverse,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
E1839,147466,1/1/2023 18:14,158.27,United States,weverse,Electronics,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B7545,147927,1/1/2023 18:21,46.55,United Kingdom,twt m_z7w,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B7545,147927,1/1/2023 18:21,46.55,United Kingdom,twt m_z7w,Stationery/Photocard/Photo album,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
D8378,139197,1/1/2023 18:25,21.1,Germany,Kpoptown.com,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B2162,143218,1/1/2023 19:31,251.66,Russian Federation,gift,1. K-POP,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
B2162,143218,1/1/2023 19:31,251.66,Russian Federation,gift,Game/Toy/Book,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,
```

Figure 1. Dataset Preview from DK Product Recommendation Data_2024_1107.csv

Size and Structure:

Number of Rows: 114,317 (representing individual transactions).

Number of Columns: 48.

Several columns are unnamed or sparsely populated, requiring careful preprocessing to extract meaningful insights.

- **Key Variables (Figure 2):**
  - **suiteno:** Unique customer or suite identifier used to track individual transactions.
  - **packageno:** A unique identifier for each package associated with the transaction.
  - **date:** The date and time of the transaction, essential for time-series analysis.
  - **shippingFee:** The cost associated with the shipment, reflecting logistics expenses.
  - **shippingcounty:** The country to which the shipment was delivered, enabling geographic segmentation.
  - **storename:** The store from which the product was purchased, indicating store-specific preferences.
  - **category:** Product category (e.g., "K-POP," "Albums/CD/DVD"), which is critical for recommendation analysis.

Summary Statistics:

The dataset contains 18,364 unique customers (suiteno).

"K-POP" is the most common category, appearing in 53,451 transactions, making up approximately 58.78% of the total records.

Key Variables and Domains

Variable	Description	Domain
`suiteNO`	Unique customer identifier (alphanumeric)	String
`packageNO`	Package number for the transaction	Float
`date`	Date and time of the transaction	Datetime (YYYY-MM-DD HH:mm)
`shippingFee`	Cost of shipping for the transaction	Float (with missing values)
`shippingcounty`	Country where the product is shipped	String
`storename`	Name of the store where the transaction occurred	String
`category`	Product category for the transaction	String (e.g., K-POP, Game/Toy/Book)
`Unnamed: 7` to `Unnamed: 47`	Mostly empty or irrelevant columns	Various

Figure 2

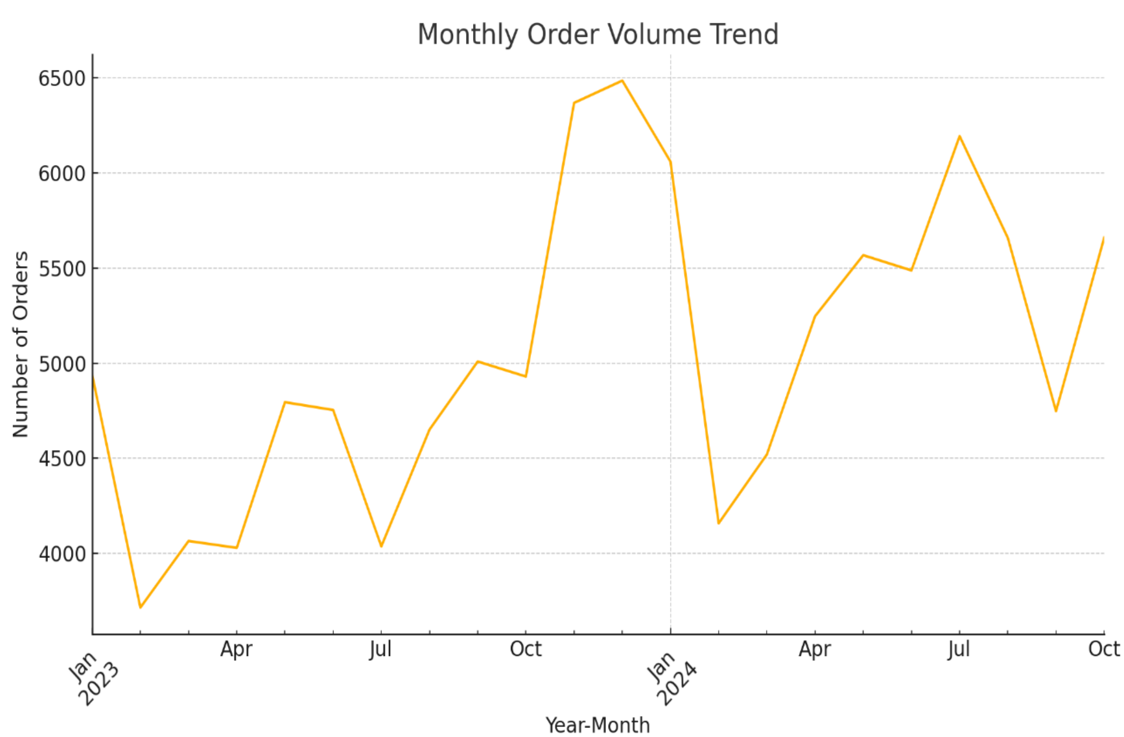


Figure 3. Monthly Order Volume Trend: Figure 3 illustrates the monthly order volume trend, showcasing seasonal patterns or growth trends.

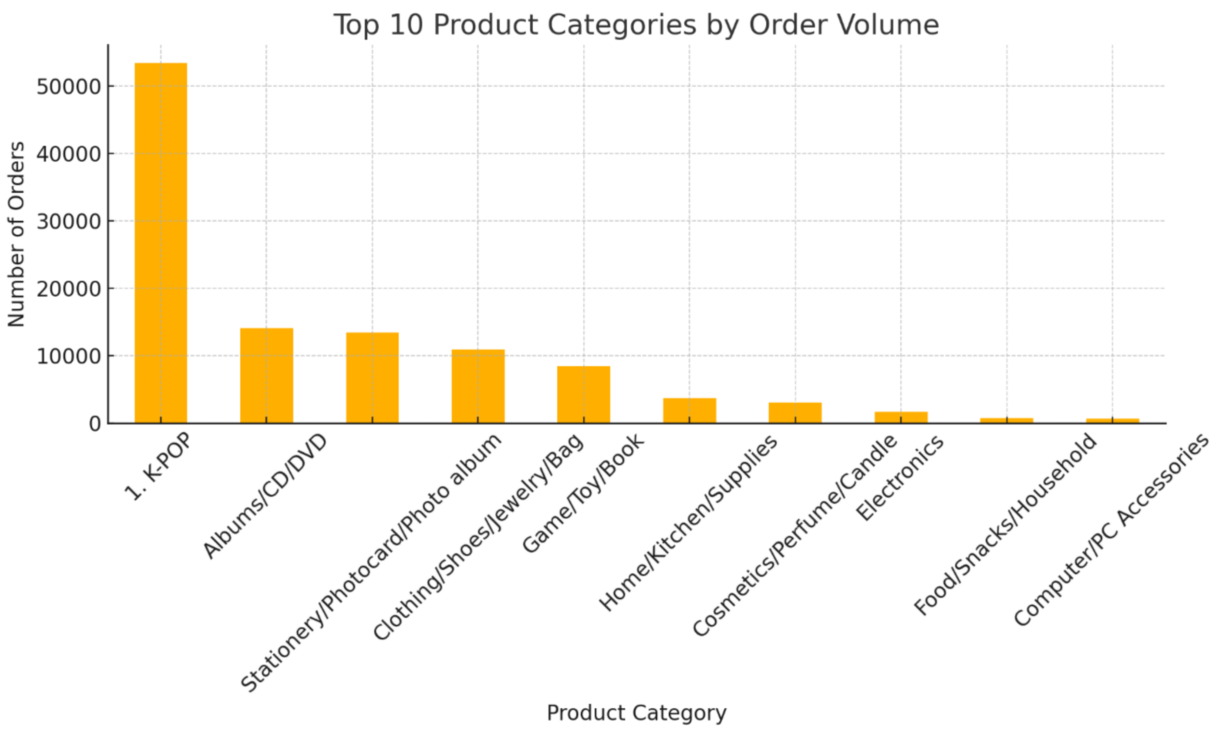


Figure 4 .Top 10 Product Categories by Order Volume:

This bar chart displays the most popular product categories, showing customer preferences and demand for different types of items.

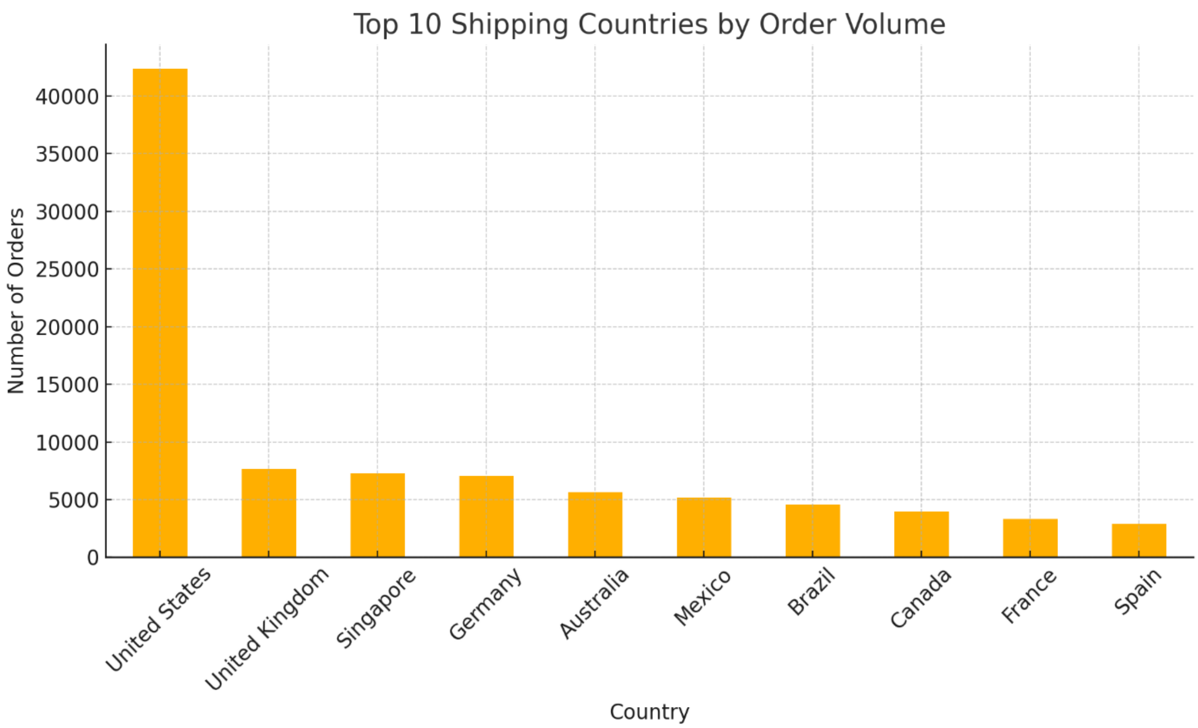


Figure 5 .Top 10 Shipping Countries by Order Volume

This bar chart reveals the countries with the highest number of orders, indicating where the products are most popular geographically

Data Quality Assessment

- **Missing Data:**
  - **shippingcounty:** Missing in 9 records.
  - **category:** Missing in 3,229 records (~2.82% of the dataset). These missing values were either imputed or excluded to ensure analytical integrity.
- **Unnamed Columns:**

The dataset includes numerous unnamed columns (e.g., Unnamed: 7 to Unnamed: 47), most of which have minimal or no data and were excluded during preprocessing.

- **Data Types:**

Several columns, such as shippingFee, contain mixed types (e.g., numeric and text), necessitating type conversions for proper analysis.

- **Outliers:**

Extreme values in shippingFee were identified and addressed to prevent skewing the analysis.

Table 1. Summary of Dataset Characteristics

Metric	Details
Basic Information	48 columns, 114317 rows. Mixed data types (39 object, 9 float). Memory usage: 41.9 MB.
Summary Statistics	See full summary in the detailed report. Includes mean, standard deviation, min, and max for numeric columns.
Missing Values (Top Columns)	Missing values detected in key columns: 'suiteNO' (1), 'date' (1), 'shippingFee' (8), 'shippingcounty' (9), 'storename' (1140), 'category' (3229). Most 'Unnamed' columns are nearly empty.
Duplicate Rows	830 duplicate rows detected.
Outlier Detection (Numeric Columns)	No outliers detected in 'packageNO', 'Unnamed: 19', 'Unnamed: 20', 'Unnamed: 27', 'Unnamed: 28', 'Unnamed: 35', 'Unnamed: 36', 'Unnamed: 43', 'Unnamed: 44'.

Data Cleaning

The data cleaning process was crucial to ensure the accuracy and reliability of the analysis. Based on the identified data quality issues, two cleaning approaches were considered:

Approach 1: Excluding all missing and irrelevant data, (which resulted in a 5.07% reduction of the dataset.)

Step 1: Handling Missing Values

shippingcounty:

Missing values in the shippingcounty column (9 records) were either imputed based on the customer's previous transactions or excluded if imputation was not possible.

category:

Missing values in the category column (~2.82% of records) were excluded from the dataset, as this variable is critical for analysis and imputation was deemed unreliable.

storename:

Rows with missing values in the storename column were retained, as this variable was not essential for the main analysis.

Step 2: Removing Irrelevant Columns

Unnamed Columns:

Columns labeled Unnamed: 7 through Unnamed: 47 were dropped due to minimal or no data, reducing unnecessary noise in the dataset.

Non-Informative Columns:

Uniform or irrelevant columns with no analytical significance were excluded to streamline the dataset.

Step 3: Addressing Outliers

shippingFee:

Extreme values in the shippingFee column were identified using interquartile ranges (IQR).

Valid outliers (e.g., large shipping costs for bulk orders) were retained to preserve realistic data patterns, while erroneous entries were corrected or removed.

Final Cleaning Results

Remaining Dataset Dimensions:

Number of Rows: 110,952.

Number of Columns: 7 key variables retained for analysis (suiteno, packageno, date, shippingFee, shippingcounty, storename, category).

Final Cleaning Results (Figure 6)

suiteNO	category	storename	shippingcounty	packageNO	date	shippingFee
D1234	1. K-POP	Weverse Global Shop	United States	146659	1/1/2023 0:58	200.14
D3654	Game/Toy/Book	Weverse Shop	Brunei Darussalam	141004	1/1/2023 1:46	50.86
D3654	Stationery/Photocard/Photo album	Weverse Shop	Brunei Darussalam	144729	1/1/2023 1:46	38.19
D3654	Stationery/Photocard/Photo album	Weverse Shop	Brunei Darussalam	149472	1/1/2023 1:46	50.86
B4702	Albums/CD/DVD	Weverse Shop	United States	146875	1/1/2023 2:31	330.21
B4702	1. K-POP	Weverse Shop	United States	146875	1/1/2023 2:31	330.21
B4702	Game/Toy/Book	Weverse Shop	United States	146875	1/1/2023 2:31	330.21
B4702	Stationery/Photocard/Photo album	Weverse Shop	United States	146875	1/1/2023 2:31	330.21
B4702	Clothing/Shoes/Jewelry/Bag	Weverse Shop	United States	146875	1/1/2023 2:31	330.21

Figure 6.Final Cleaning Results of Approach 1

cleaning approaches, showcasing the impact on dataset dimensions.

Approach 2: Focused Cleaning on Key Variables

This approach minimized data loss by focusing on cleaning the category and shippingcounty columns, resulting in a 2.94% reduction. Given the study's focus on product categories, the second approach was chosen to retain more data while ensuring the integrity of key variables. (Figure 6-2)

suiteNO	shippingcountry	category
D1234	United States	1. K-POP
D3654	Brunei Darussalam	Game/Toy/Book
D3654	Brunei Darussalam	Stationery/Photocard/Photo album
D3654	Brunei Darussalam	Stationery/Photocard/Photo album
B4702	United States	Albums/CD/DVD
B4702	United States	1. K-POP
B4702	United States	Game/Toy/Book
B4702	United States	Stationery/Photocard/Photo album
B4702	United States	Clothing/Shoes/Jewelry/Bag
B4702	United States	Computer/PC Accessories
B2026	Singapore	1. K-POP
E2818	United States	Stationery/Photocard/Photo album
D4664	United States	1. K-POP

Figure 6-2. Final Cleaning Results of Approach 2

Data Transformation

To prepare the dataset for analysis, several transformations were performed to simplify the structure and enhance interpretability while maintaining critical information for modeling:

Category Consolidation:

Product categories were grouped into eight main groups to reduce complexity and improve analytical focus. Examples include:

"Game/Toy/Book" and "Stationery/Photocard/Photo album" were merged into a broader "Entertainment" category.

"Fashion" and related categories were unified under a single "Fashion" category.

	K-POP	Entertainment	Fashion	Home & Living	Beauty	Electronics	Food & Health	Automotive
Included Items	K-POP merchandise	Music, stationery, games, toys, books	Clothing, shoes, jewelry, bags	Home and kitchen supplies, furniture	Cosmetics, perfumes, candles, facial masks	Electronics, computer accessories	Food, snacks, pet supplies, fitness equipment	Car parts and accessories
Reason	Kept separate due to its distinct merchandise and fanbase.	Commonly associated with leisure activities.	Standard fashion items, often grouped together in retail.	Logical grouping for home-related items, making it easy to locate essentials.	Related to personal care, providing a consistent theme for beauty products.	Combines main electronics with accessories as tech products.	Related to health, wellness, and nutrition, forming a cohesive category.	Includes vehicle-related products for car maintenance and customization.

Other
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0
0

Figure 7. Categorization of Product Groups

Rare categories accounting for 1.66% of transactions were grouped as "Other", and later excluded to simplify the analysis. (Figure 7)

Geographic Aggregation:

Shipping destinations (shippingcounty) were aggregated into six major continents: Asia, Europe, North America, South America, Africa, and Oceania.

This aggregation enabled a high-level analysis of geographic trends while minimizing noise from country-level details.

Cross-Sectional Transformation:

The transactional dataset, originally in panel format, was transformed into a cross-sectional format. (Figure 8)

Each row represents a unique customer (suiteno), with columns summarizing key metrics such as purchase counts per product category, total spending, and geographic region.

suiteNO	K-POP	Entertainment	Fashion	Home & Living	Beauty	Electronics	Food & Health	Automotive	Total	continent_encoded
A0266	3	12	4	3	0	3	1	0	26	Asia
A0519	2	0	0	0	0	0	0	0	2	Europe
A0523	6	2	1	1	0	0	0	0	10	North America
A0604	2	5	0	0	0	0	0	0	7	Asia
A0610	0	29	21	12	3	11	0	0	76	North America
A0613	1	0	0	0	0	0	0	0	1	Asia
A0619	3	0	0	0	0	0	0	0	3	South America
A0627	16	5	0	0	0	0	0	0	21	North America
A0673	7	1	1	0	0	0	0	0	9	Europe
A0680	13	1	3	0	5	0	0	0	22	North America
A0681	0	3	1	0	0	6	0	0	10	Asia
A0686	14	5	0	0	0	0	0	0	19	North America
A0694	1	2	0	0	0	0	0	0	3	South America
A0738	12	6	5	3	1	1	0	0	28	Europe
A0741	5	3	5	3	0	1	0	0	17	North America
A0744	3	2	1	0	0	0	0	0	6	Europe
A0756	0	0	2	0	1	0	0	0	3	North America
A0774	3	1	0	0	0	0	0	0	4	North America
A0794	2	2	0	0	0	0	0	0	4	North America
A0810	0	1	2	0	0	1	0	0	4	North America
A0818	1	0	0	0	0	0	0	0	1	North America
A0819	24	14	9	0	1	1	0	0	49	South America
A0824	14	1	1	1	0	0	0	0	17	Oceania
A0836	1	1	0	0	0	0	0	0	2	North America
A0838	3	0	0	0	0	0	0	0	3	Asia
A0843	19	17	4	2	0	0	0	0	42	North America

Figure 8.Cross-Sectional Dataset

Understanding the Cross-Sectional Dataset

After transforming the original transactional data into a cross-sectional format, the dataset provides a customer-centric view that summarizes key attributes and purchase behaviors for each customer (suiteNO). This transformation simplifies the data structure, making it suitable for pattern recognition and predictive modeling while retaining critical details for analysis.

Key Features :

Customer Identifier (suiteNO):

Each row in the dataset represents a unique customer, identified by their suiteNO.

This structure ensures that the focus is on customer-specific behavior rather than individual transactions, facilitating personalized insights.

Category-Level Purchase Counts:

Columns represent the total number of purchases made by a customer in each of the eight consolidated product categories: (Figure 9)

"K-POP"

"Entertainment"

"Fashion"

"Home & Living"

"Beauty"

"Electronics"

"Food & Health"

"Automotive"

These counts highlight customer preferences across categories, allowing for personalized product recommendations.

Geographic Region (continent_encoded):

Customers are grouped into six geographic regions based on their shipping destination: Asia, Europe, North America, South America, Africa, and Oceania.

Benefits :

Simplifies analysis and enhances interpretability.

Suitable for machine learning applications, particularly for classification tasks.

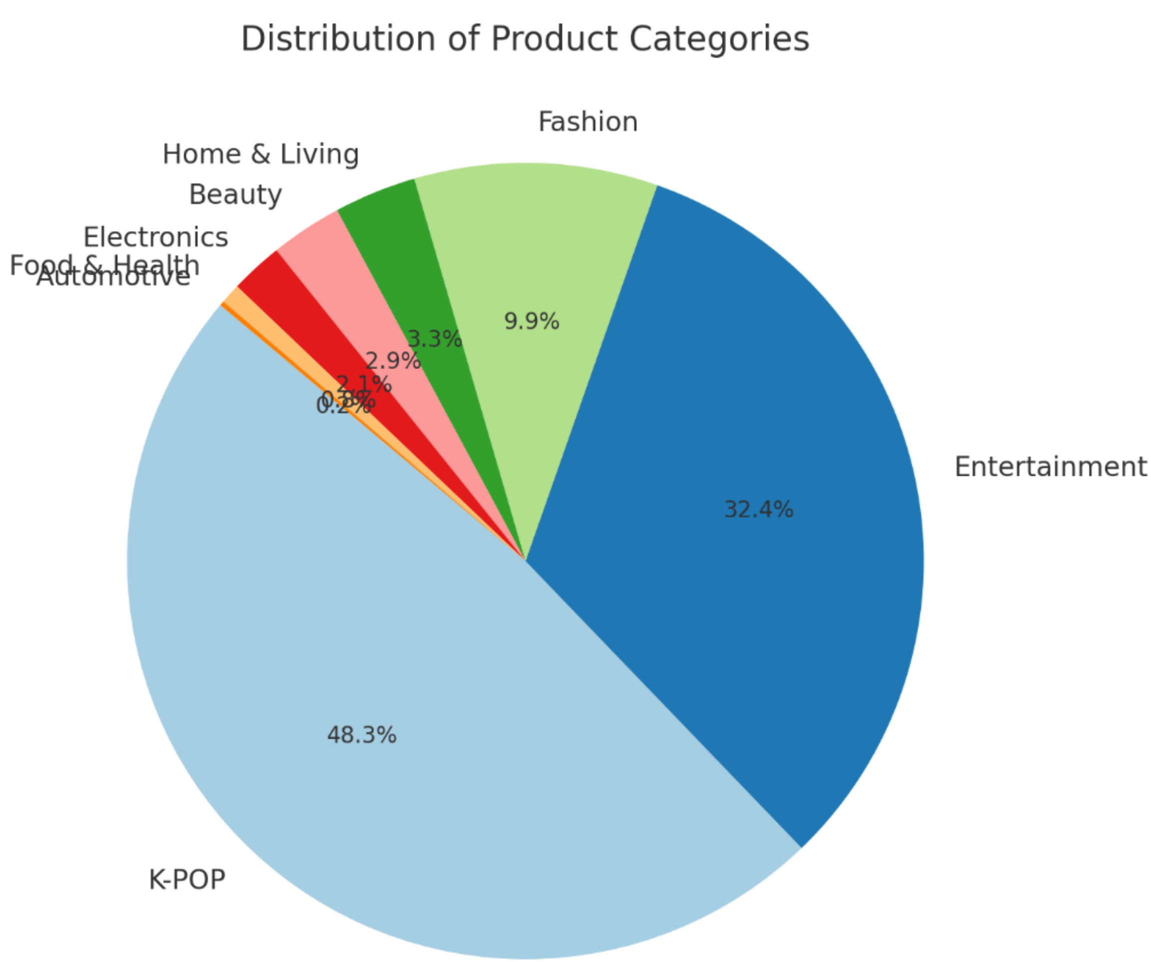


Figure 9.Distribution of Product Categories

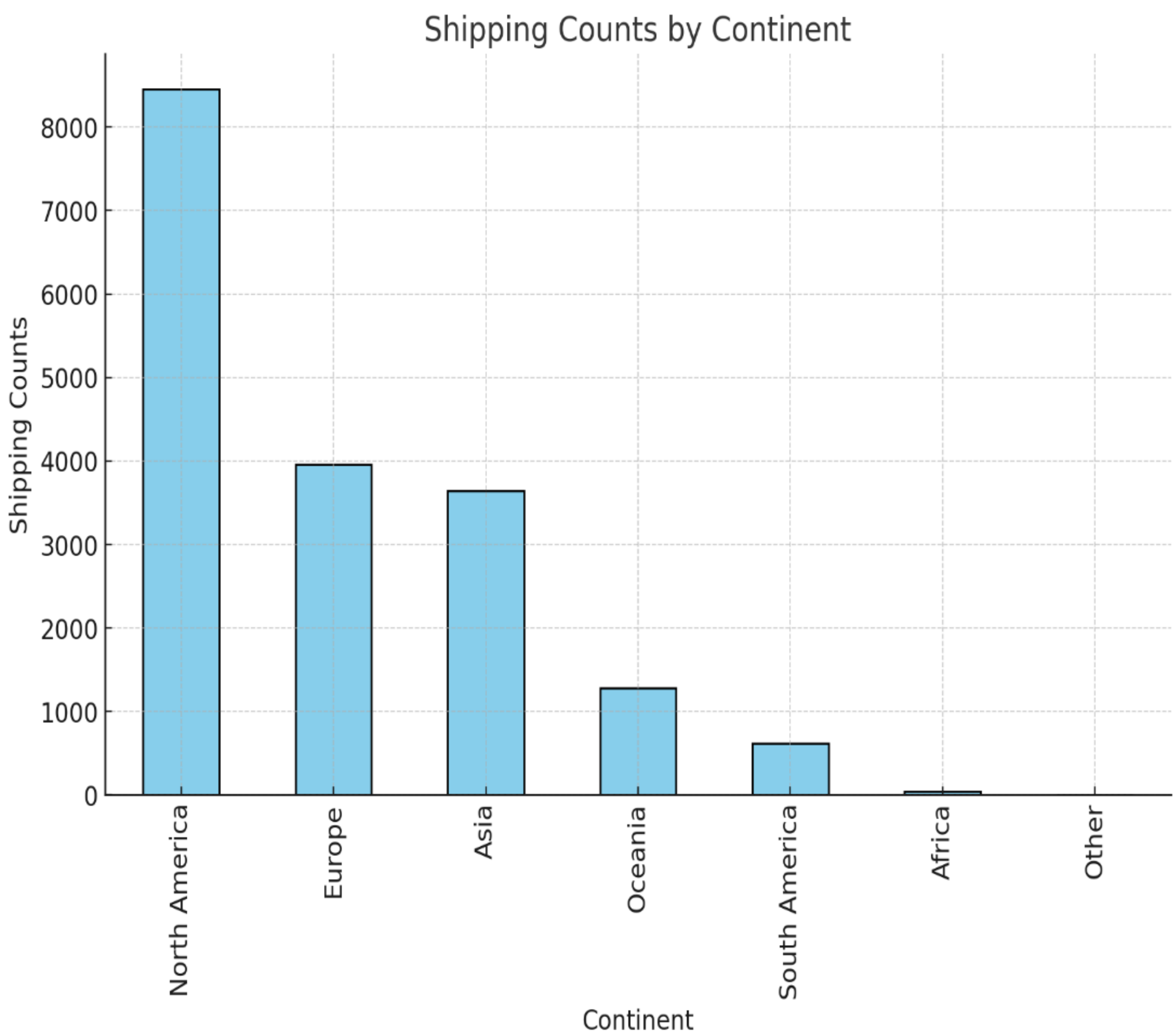


Figure 10.Shipping Counts by Continent

Analysis of Product Category Relationships

Correlation and Shipping Counts

Understanding the relationships between product categories is crucial for uncovering meaningful patterns in customer purchases. This analysis leverages correlation and shipping counts to identify category pairs frequently purchased together, providing actionable insights for product recommendations and marketing strategies.

Methodology

To analyze product category relationships, the study employs two key metrics: correlation and shipping counts.

Correlation Analysis: Correlation measures the strength and direction of the relationship between product categories. High positive correlation values indicate that customers often purchase specific categories together. For example, a strong correlation between "K-POP" and "Albums/CD/DVD" suggests these categories appeal to a similar customer base and are frequently co-purchased.

Shipping Counts: Shipping counts quantify how often two categories are ordered together. While correlation identifies relative relationships, shipping counts add a business context by providing the absolute frequency of co-purchases. This dual perspective helps prioritize category pairs that are not only related but also popular among customers.

Findings

Heatmap Analysis: A correlation heatmap (Figure 11) visualizes the strength of relationships between product categories.Key observations include high correlations between "K-POP" and "Entertainment”.

A high correlation was observed between "K-POP" and "Entertainment" categories. This suggests that customers purchasing K-POP merchandise often complement their orders with related items like albums or DVDs.

Moderate correlations between "Fashion" and "Beauty" indicate that these categories attract overlapping customer segments, likely due to shared demographic appeal.

Shipping Counts: The shipping counts for each category pair were overlaid onto the correlation analysis to identify high-frequency co-purchases. (Figure 12)

Insights:

"Electronics" and "Fashion" emerged as a high-shipping-count pair, reflecting their broad customer appeal and potential for bundling.

Certain category pairs, such as "Food & Health" and "Home & Living," exhibited high shipping counts but low correlation, suggesting general demand rather than specific associations.

Bar Chart Representation: To rank category pairs, a bar chart was created, showcasing the strongest relationships (based on correlation) alongside their shipping counts.

Examples:

"K-POP" and "Albums/CD/DVD" ranked highest, with both strong correlation and significant shipping counts.

Other notable pairs included "Fashion" and "Beauty," as well as "Home & Living" with "Electronics."

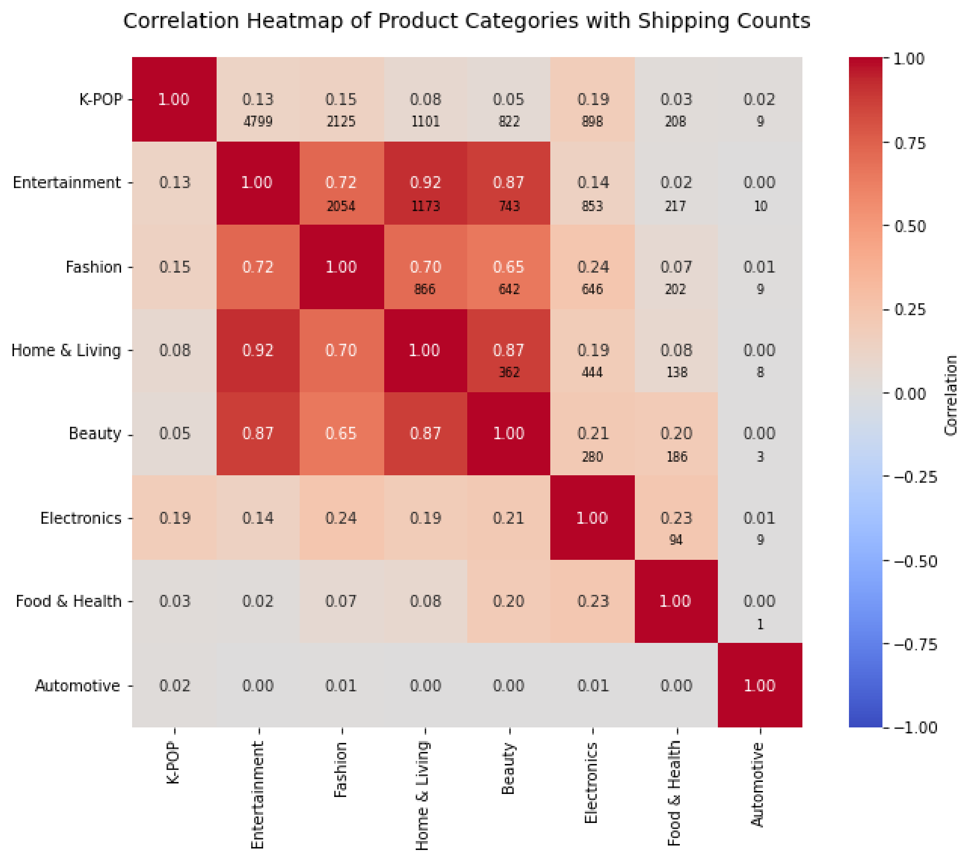

Correlation Heatmap of Product Categories with Shipping Counts
K-POP
Entertainment
Fashion
Home & Living
Beauty
Electronics
Food & Health
Automotive
K-POP 1.00 0.13 4799 0.15 2125 0.08 1101 0.05 822 0.19 898 0.03 208 0.02 9
Entertainment 0.13 1.00 0.72 2054 0.92 1173 0.87 743 0.14 853 0.02 217 0.00 10
Fashion 0.15 0.72 1.00 0.70 866 0.65 642 0.24 646 0.07 202 0.01 9
Home & Living 0.08 0.92 0.70 1.00 0.87 362 0.19 444 0.08 138 0.00 8
Beauty 0.05 0.87 0.65 0.87 1.00 0.21 280 0.20 186 0.00 3
Electronics 0.19 0.14 0.24 0.19 0.21 1.00 0.23 94 0.01 9
Food & Health 0.03 0.02 0.07 0.08 0.20 0.23 1.00 0.00 1
Automotive 0.02 0.00 0.01 0.00 0.00 0.01 0.00 1.00
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Correlation

Figure 11.Correlation Heatmap of Product Categories with shipping Counts

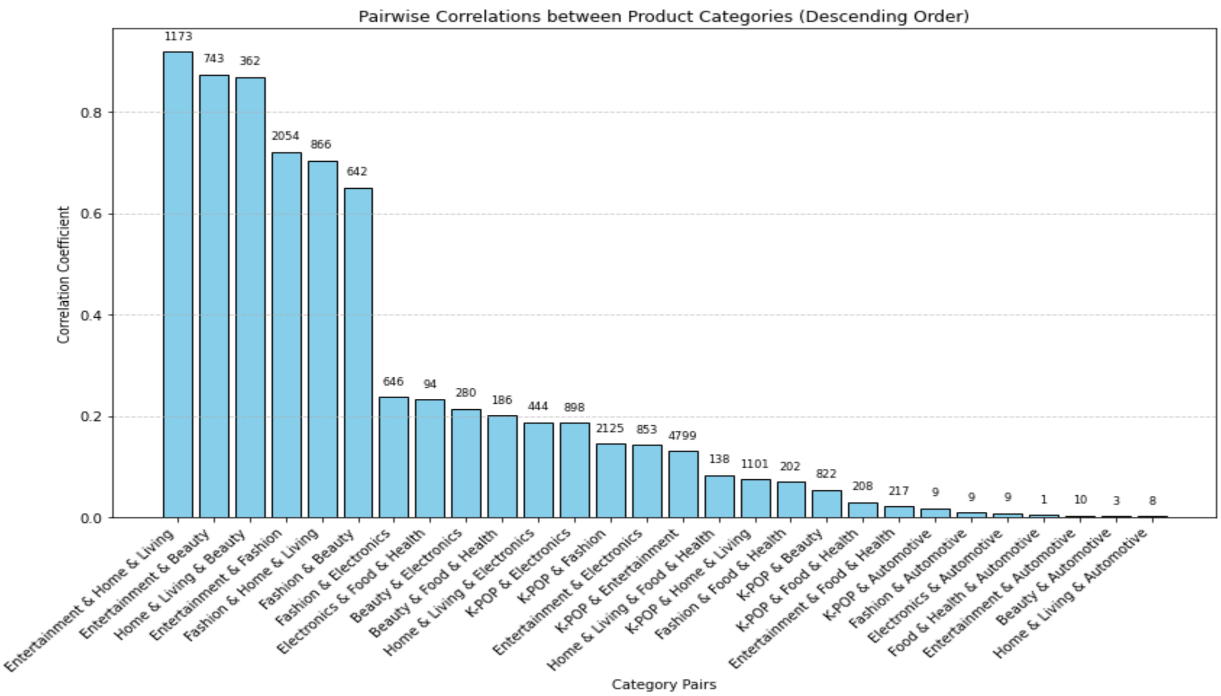


Figure 12.Pairwise Correlation Between Product Categories

Practical Implications

Marketing Strategies:

High-correlation and high-shipping-count pairs can guide cross-selling and bundling strategies.

Inventory Management:

Understanding co-purchase trends allows for better inventory allocation, ensuring frequently paired items are stocked together to reduce logistical inefficiencies.

Regional Customization:

Segmenting analysis by geographic regions supports the development of region-specific marketing strategies.

Data Preprocessing and Customer Purchase Patterns

Following the exploration of product category relationships, this section delves into data preprocessing and the analysis of customer purchase patterns by category. This step is essential for structuring the dataset for predictive modeling and understanding customer behavior in terms of category preferences.

Data Preprocessing and Binary Conversion (Figure 13, Figure 14)

The initial dataset consisted of purchase counts for various product categories. To prepare it for analysis, the data was transformed into a binary format:

Binary Conversion Rule:

Purchase counts >0 converted to 1 (indicating a purchase).Purchase counts of 0 remained 0

Significance:

Simplifies analysis and ensures consistency when identifying patterns across categories.

suiteNO	K-POP	Entertainment	Fashion	Home & Living	Beauty	Electronics	Food & Health	Automotive
A0266	3	12	4	3	0	3	1	0
A0519	2	0	0	0	0	0	0	0
A0523	6	2	1	1	0	0	0	0
A0604	2	5	0	0	0	0	0	0
A0610	0	29	21	12	3	11	0	0

suiteNO	K-POP	Entertainment	Fashion	Home & Living	Beauty	Electronics	Food & Health	Automotive
A0266	1	1	1	1	0	1	1	0
A0519	1	0	0	0	0	0	0	0
A0523	1	1	1	1	0	0	0	0
A0604	1	1	0	0	0	0	0	0
A0610	0	1	1	1	1	1	0	0

Figure 13. Transformation of Category Counts to Binary Indicators for Product Analysis

suiteNO	K-POP	Entertainment	Fashion	Home & Living	Beauty	Electronics	Food & Health	Automotive
A0266	1	1	1	1	0	1	1	0
A0519	1	0	0	0	0	0	0	0
A0523	1	1	1	1	0	0	0	0
A0604	1	1	0	0	0	0	0	0
A0610	0	1	1	1	1	1	0	0
A0613	1	0	0	0	0	0	0	0
A0619	1	0	0	0	0	0	0	0
A0627	1	1	0	0	0	0	0	0
A0673	1	1	1	0	0	0	0	0
A0680	1	1	1	0	1	0	0	0
A0681	0	1	1	0	0	1	0	0
A0686	1	1	0	0	0	0	0	0
A0694	1	1	0	0	0	0	0	0
A0738	1	1	1	1	1	1	0	0
A0741	1	1	1	1	0	1	0	0
A0744	1	1	1	0	0	0	0	0
A0756	0	0	1	0	1	0	0	0
A0774	1	1	0	0	0	0	0	0
A0794	1	1	0	0	0	0	0	0
A0810	0	1	1	0	0	1	0	0
A0818	1	0	0	0	0	0	0	0
A0819	1	1	1	0	1	1	0	0
A0824	1	1	1	1	0	0	0	0
A0836	1	1	0	0	0	0	0	0
A0838	1	0	0	0	0	0	0	0
A0843	1	1	1	1	0	0	0	0
A0858	1	1	1	0	1	1	0	0

Figure 14. Binary Representation of Product Categories Across Transactions

Customer Purchase Patterns by Category

After binary conversion, customer purchase patterns were analyzed to understand the diversity of their shopping behavior. The findings revealed two distinct groups:

Single-Category Buyers:

11,363 customers (63.4% of the dataset) purchased from only one category.

This group likely represents specialized buyers with focused interests, such as fans of K-POP merchandise. (Figure 15)

suiteNO	K-POP	Entertainment	Fashion	Home & Living	Beauty	Electronics	Food & Health	Automotive
A0519	1	0	0	0	0	0	0	0
A0613	1	0	0	0	0	0	0	0
A0619	1	0	0	0	0	0	0	0
A0818	1	0	0	0	0	0	0	0
A0838	1	0	0	0	0	0	0	0
A0938	0	0	1	0	0	0	0	0
A0966	1	0	0	0	0	0	0	0
A0988	0	1	0	0	0	0	0	0
A1001	0	0	1	0	0	0	0	0
A1017	0	0	1	0	0	0	0	0
A1085	0	0	1	0	0	0	0	0
A1124	1	0	0	0	0	0	0	0
A1125	0	0	0	0	1	0	0	0
A1263	1	0	0	0	0	0	0	0
A1344	1	0	0	0	0	0	0	0
A1353	0	1	0	0	0	0	0	0
A1372	1	0	0	0	0	0	0	0
A1374	1	0	0	0	0	0	0	0
A1422	1	0	0	0	0	0	0	0
A1511	1	0	0	0	0	0	0	0
A1588	0	0	1	0	0	0	0	0
A1589	0	0	1	0	0	0	0	0
A1595	0	0	1	0	0	0	0	0
A1627	0	1	0	0	0	0	0	0

Figure 15. Binary Representation of Product Categories for only one Transactions

Multi-Category Buyers:

6,552 customers (36.6%) purchased from more than one category.(Figure 16)

Multi-category buyers are valuable for cross-selling opportunities, as they exhibit broader interests.

suiteNO	K-POP	Entertainment	Fashion	Home & Living	Beauty	Electronics	Food & Health	Automotive
A0266	1	1	1	1	0	1	1	0
A0523	1	1	1	1	0	0	0	0
A0604	1	1	0	0	0	0	0	0
A0610	0	1	1	1	1	1	0	0
A0627	1	1	0	0	0	0	0	0
A0673	1	1	1	0	0	0	0	0
A0680	1	1	1	0	1	0	0	0
A0681	0	1	1	0	0	1	0	0
A0686	1	1	0	0	0	0	0	0
A0694	1	1	0	0	0	0	0	0
A0738	1	1	1	1	1	1	0	0
A0741	1	1	1	1	0	1	0	0
A0744	1	1	1	0	0	0	0	0
A0756	0	0	1	0	1	0	0	0
A0774	1	1	0	0	0	0	0	0
A0794	1	1	0	0	0	0	0	0
A0810	0	1	1	0	0	1	0	0
A0819	1	1	1	0	1	1	0	0
A0824	1	1	1	1	0	0	0	0
A0836	1	1	0	0	0	0	0	0
A0843	1	1	1	1	0	0	0	0
A0858	1	1	1	0	1	1	0	0
A0861	0	1	1	0	0	0	0	0
A0870	1	1	1	0	0	0	0	0
A0913	1	1	1	0	0	1	0	0

Figure 16. Binary Representation of Product Categories for more than one Transactions

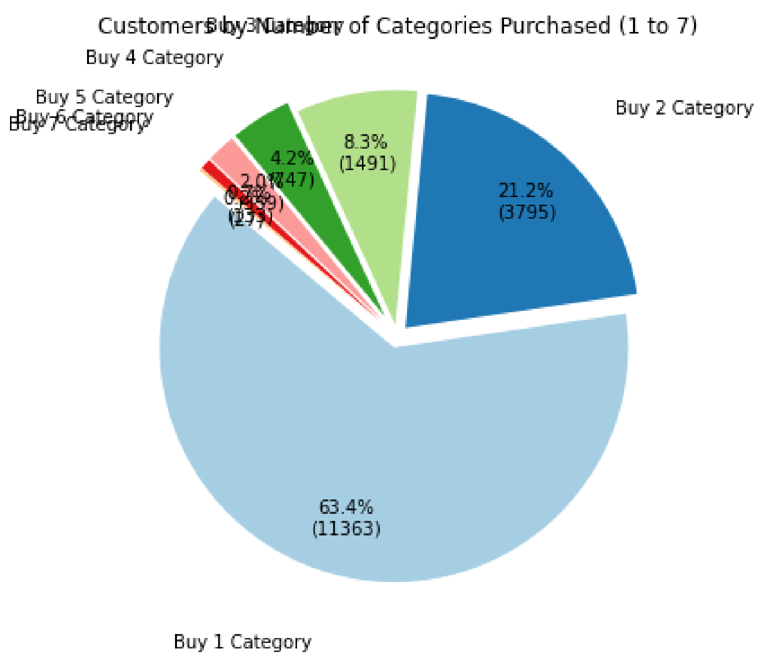


Figure 17.Pie chart of Transactions of one, two, three, four, five, six, seven Categories

Visualization:

A pie chart and grouped binary data illustrate these customer segments, highlighting the predominance of single-category buyers but also the potential of multi-category buyers for targeted marketing campaigns. (Figure 17)

Client Purchase Counts by Number of Categories

To deepen the analysis, the number of categories purchased by each customer was further broken down:

Major Observations:

The majority of customers (63.4%) purchased only one category.

A smaller but significant portion purchased from 2 (21.2%), 3 (8.3%), or more categories, tapering off to just 27 customers buying from all 7 categories.

Bar Chart Representation (Figure 18):

The chart demonstrates the distribution of customers across different category counts, emphasizing the long tail of specialized buyers versus the broader interests of a smaller group.

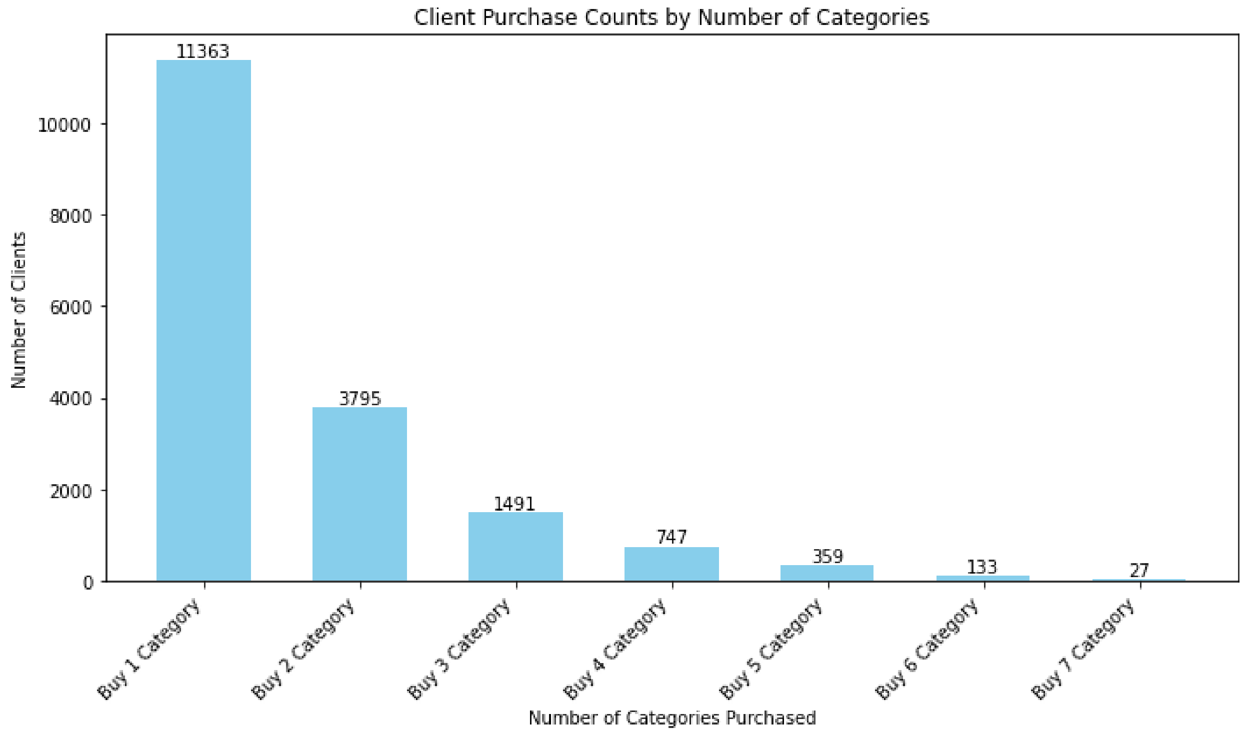


Figure 18.Client Purchase Counts by Number of Categories

Implications:

Multi-category buyers offer cross-selling opportunities.Single-category buyers can be targeted with related products in the same category to deepen their engagement.

Leveraging the Apriori Algorithm for Product Recommendation

The Apriori algorithm is a key component in this study's recommendation framework, enabling the identification of associations between product categories. By systematically evaluating customer purchase data, the algorithm uncovers frequently co-purchased items, making it an essential tool for building actionable product recommendations.

Key Metrics Used in the Apriori Algorithm

The Apriori algorithm identifies relationships between items based on their co-occurrence in transaction data.

1.Support:

Indicates how frequently a product category appears in transactions.The algorithm scans all transactions to identify individual items or itemsets that meet a predefined support threshold.

Support is calculated as:

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

Analysis of Support Ratios and Counts (Figure 19)

The support metric plays a foundational role in identifying frequently purchased product categories in the dataset. Support measures the proportion of transactions in which a specific category appears, helping prioritize items for association rule mining and recommendation systems.

	Support Ratio	Support Count
K-POP	0.587770984	10574
Entertainment	0.435908838	7842
Fashion	0.30233463	5439
Home & Living	0.096275709	1732
Electronics	0.089327404	1607
Beauty	0.077543079	1395
Food & Health	0.026459144	476
Automotive	0.00717065	129

Figure 19. Support Ratios and Counts for Product Categories

Key Observations from Support Analysis

High-Support Categories:

K-POP is the most prominent category, appearing in 58.78% of transactions with a total count of 10,574 purchases.

Entertainment follows with 43.59% support and 7,842 transactions.

Fashion ranks third at 30.23% support with 5,439 transactions.

Low-Support Categories:

Categories such as Food & Health and Automotive exhibit much lower support ratios (2.65% and 0.71%, respectively), indicating niche demand.

Visual Representation (Figure 20)

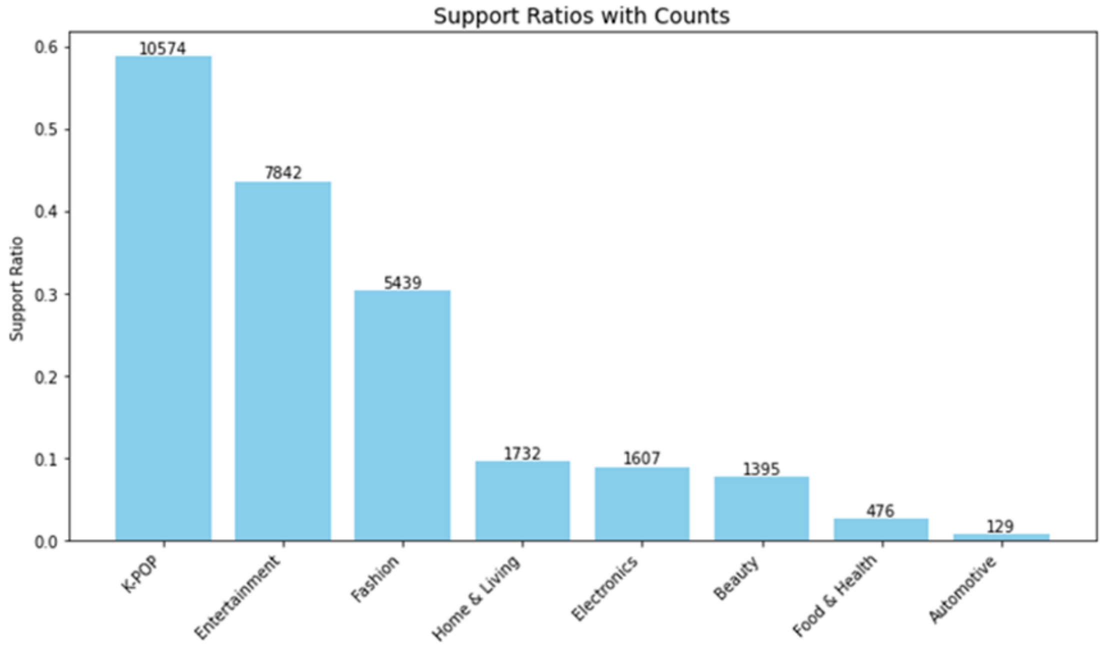


Figure 20.Support Ratios with Counts

The support distribution is visualized using a bar chart, emphasizing the dominance of categories like K-POP, Entertainment, and Fashion in the dataset. These categories are key targets for building association rules and cross-selling strategies due to their broad appeal.

Implications for Product Recommendations

High-support categories ensure broader customer engagement and relevance in recommendations.

Low-support categories may still hold strategic value for targeted campaigns, particularly in niche markets or specific regions.

By leveraging support ratios, this analysis identifies priority categories for developing effective and data-driven product recommendations.

2.Confidence:

Confidence is a critical metric in association rule mining, measuring the likelihood that a customer will purchase one category (item B) given that they have already purchased another (item A). Mathematically, confidence is calculated as:

$$\text{Confidence} = \frac{\text{Number of times A and B are purchased together}}{\text{Number of times A is purchased}}$$

This metric directly evaluates the predictive strength of associations, making it a valuable tool for designing effective recommendation strategies.

Confidence Threshold: 20% (Figure 21)

In this analysis, pairs with a confidence greater than 0.2 (20%) are considered strong candidates for recommendations. The 20% threshold was chosen because it strikes a balance between relevance and practical application, ensuring that recommendations are both meaningful and actionable. For instance:

A confidence of 0.6 (60%) indicates that if a customer buys "Home & Living," there is a 60% likelihood that they will also buy "Entertainment."

A confidence of 0.25 (25%) for "Food & Health → K-POP" suggests that 25% of customers who buy "Food & Health" also purchase "K-POP."

Highlights from the Graph

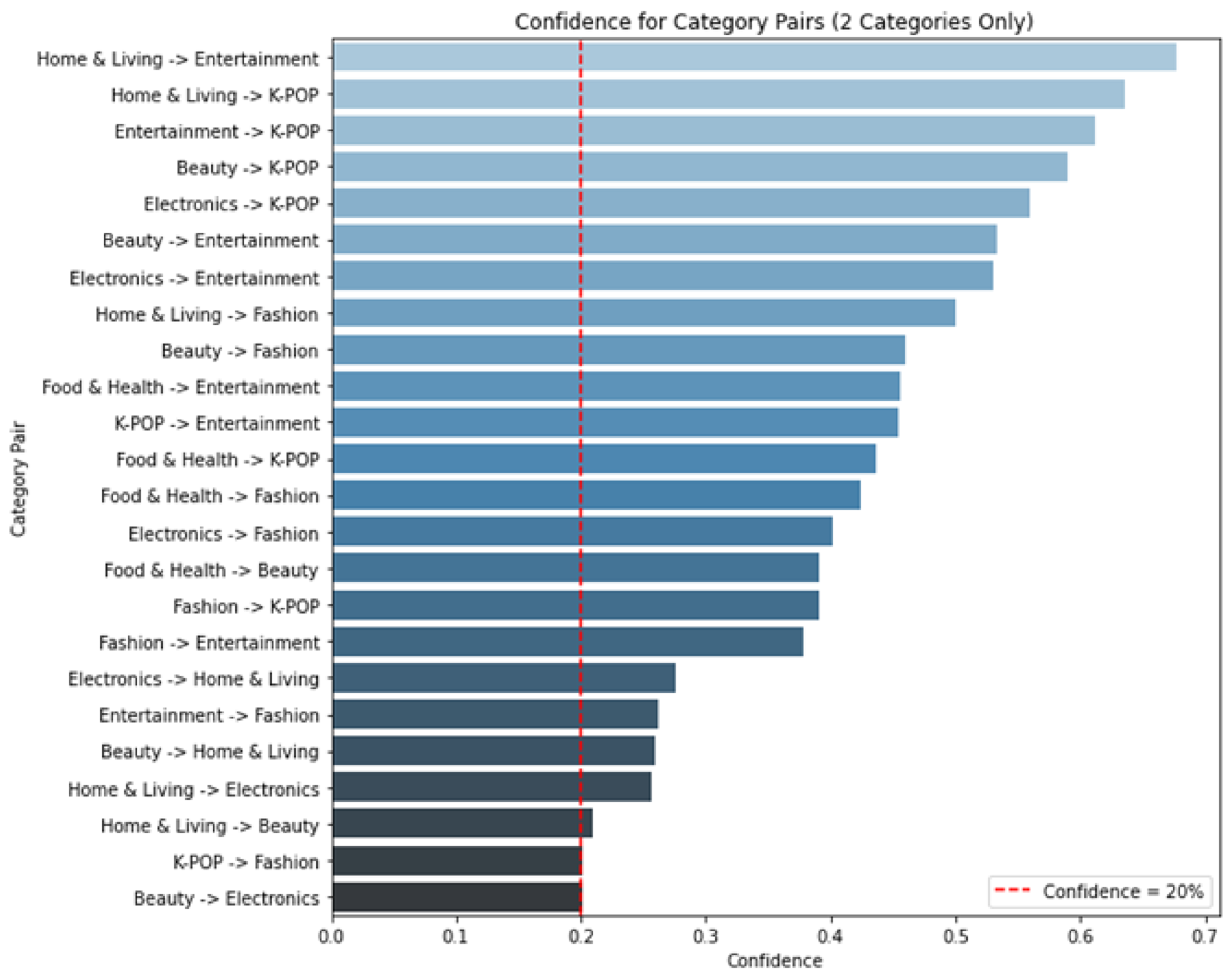


Figure 21. Confidence Levels for Product Category Pairs with a Threshold of 20%

The bar chart presents a ranked view of confidence values for category pairs, with notable examples including:

High Confidence Pairings:

"Home & Living → Entertainment" (confidence ≈ 0.6) indicates a strong relationship, suggesting opportunities for cross-selling or bundling products from these categories.

"Beauty → K-POP" and "Electronics → K-POP" also show significant confidence levels, hinting at customer tendencies to combine these purchases.

Moderate Confidence Pairs:

Categories such as "Food & Health → Entertainment" and "Electronics → Fashion" demonstrate moderate confidence, which may still be leveraged for targeted marketing.

Lower Confidence Pairs:

Pairs like "Beauty → Electronics" fall below the threshold, indicating weaker associations that may not be as effective for recommendations.

Applications of Confidence

Targeted Marketing:

High-confidence pairs can be used to craft tailored promotional campaigns. For example, customers buying "Home & Living" items can be shown advertisements for "Entertainment" products.

Product Bundling:

Pairs with high confidence are ideal for creating bundles or discount packages, encouraging customers to purchase related products.

Customer Insights:

Understanding which category combinations have high confidence levels provides insights into customer preferences and shopping patterns.

The association rule model based on the Apriori algorithm is presented as a robust framework for predictive analytics in product recommendation systems. This section integrates the methodology and role of metrics—such as confidence and lift—into a clear and logical workflow, demonstrating how these elements enable the identification of meaningful patterns in customer purchasing behavior.

Advanced Analyses

Building on the Apriori algorithm, lift emerges as a pivotal metric for understanding the strength of associations between product categories. This section delves deeper into the concept of lift, its calculation, and the significant regional insights derived from the dataset.

3.**Lift:**

Lift measures the strength of an association between two product categories by comparing their co-occurrence frequency to what would be expected if they were independent. Mathematically, it $\text{Lift}(A, B) = \frac{P(A \cap B)}{P(A) \times P(B)}$ is expressed as:

Interpretation:

Lift > 1: A positive association; the items are purchased together more often than expected.

Lift = 1: No association; the items are independent.

Lift < 1: A negative association; the items are less likely to be purchased together.

Example: If the lift between "Beauty" and "Food & Health" is 5.03, customers who purchase "Beauty" are five times more likely to buy "Food & Health" compared to random chance. This high lift value highlights a strong cross-category relationship.

Calculating Lift for Association Rules (Figure 22,Figure 23)

Using the Apriori algorithm, lift values were calculated for frequently purchased category pairs. The top category pairs with high lift values and their corresponding counts are summarized

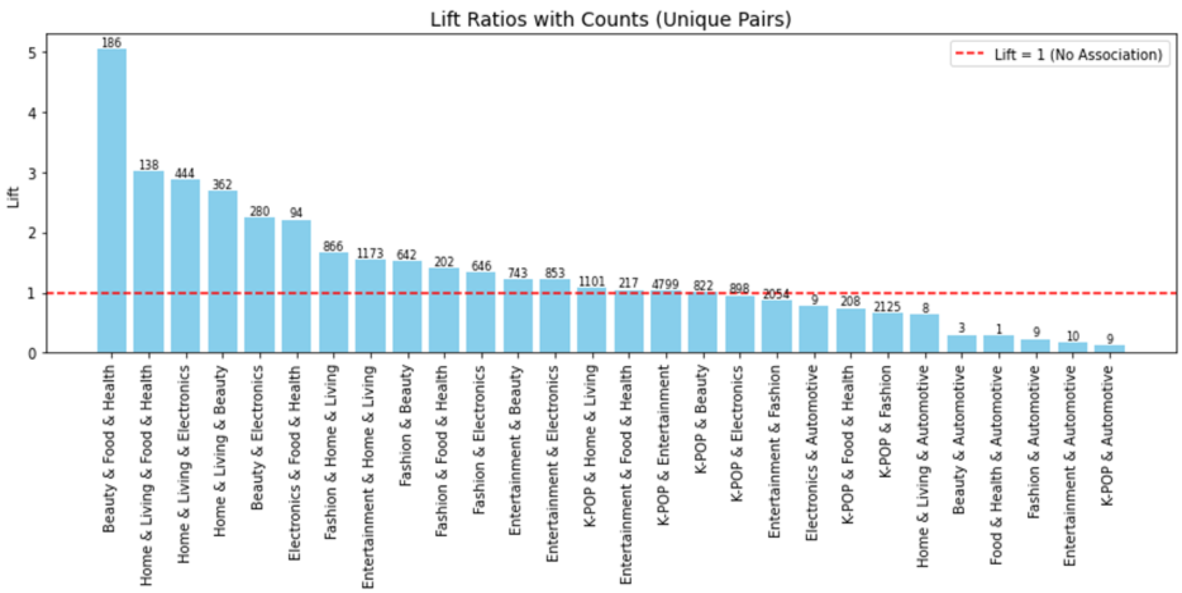


Figure 22.Lift Ratios with Counts

in the following table:

Category Pair	Lift	Count
Beauty & Food & Health	5.039	186
Home & Living & Food & Health	3.011	138
Food & Health & Electronics	2.869	94
Home & Living & Electronics	2.695	444
Beauty & Electronics	2.247	280
Beauty & Home & Living	2.211	362
Entertainment & Home & Living	1.553	1173
Fashion & Home & Living	1.522	642
Fashion & Electronics	1.329	646
Entertainment & Electronics	1.218	853

Figure 23. Lift and Count Metrics for Top Product Category Pairs

The associated bar chart visually depicts these relationships, emphasizing category pairs with both high lift and substantial support counts.

Regional Analysis of Lift and Category Pairing

The dataset was segmented by region to analyze how lift varies geographically. Two primary observations were drawn:

1.North America and Europe: (Figure 24 ,Figure 25)

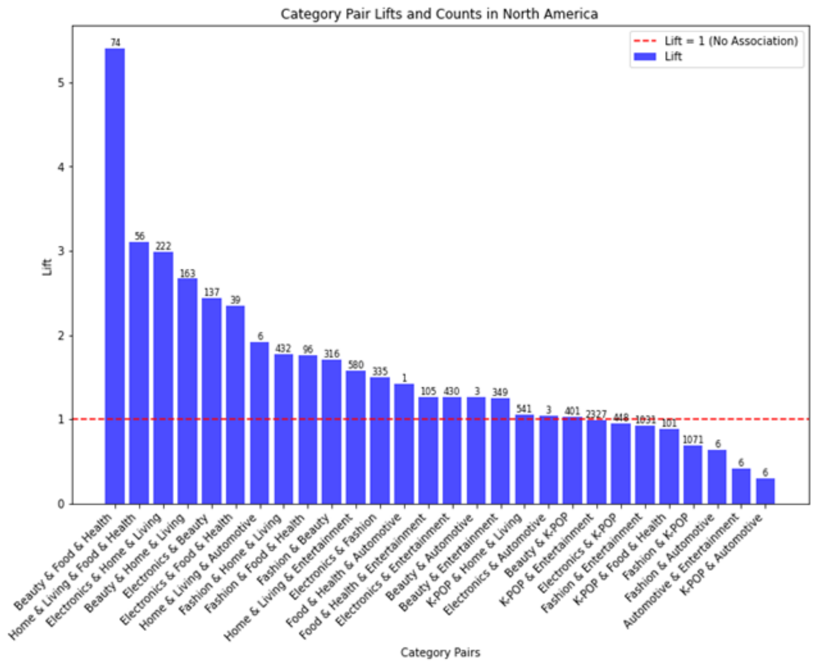


Figure 24.Category Pair Lifts and Counts in North America

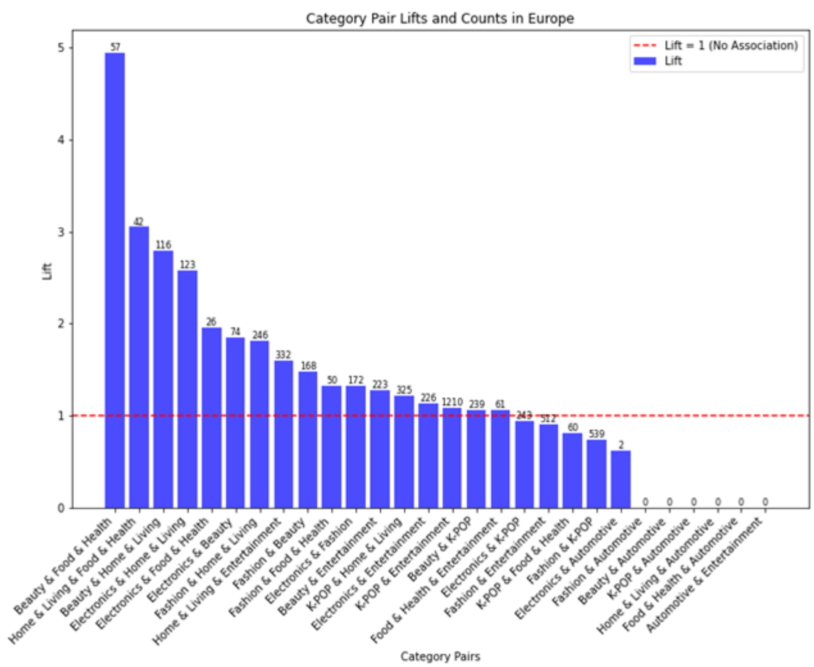


Figure 25.Category Pair Lifts and Counts in Europe

These regions demonstrated higher transaction counts, providing more meaningful lift values.

For instance, in North America, the "Home & Living → Entertainment" pair had a lift of 3.0, reflecting a significant co-purchase trend.

2.Smaller Markets (Asia, Africa, Oceania, South America): (Figure 26, Figure 27,Figure 28 ,Figure 29)

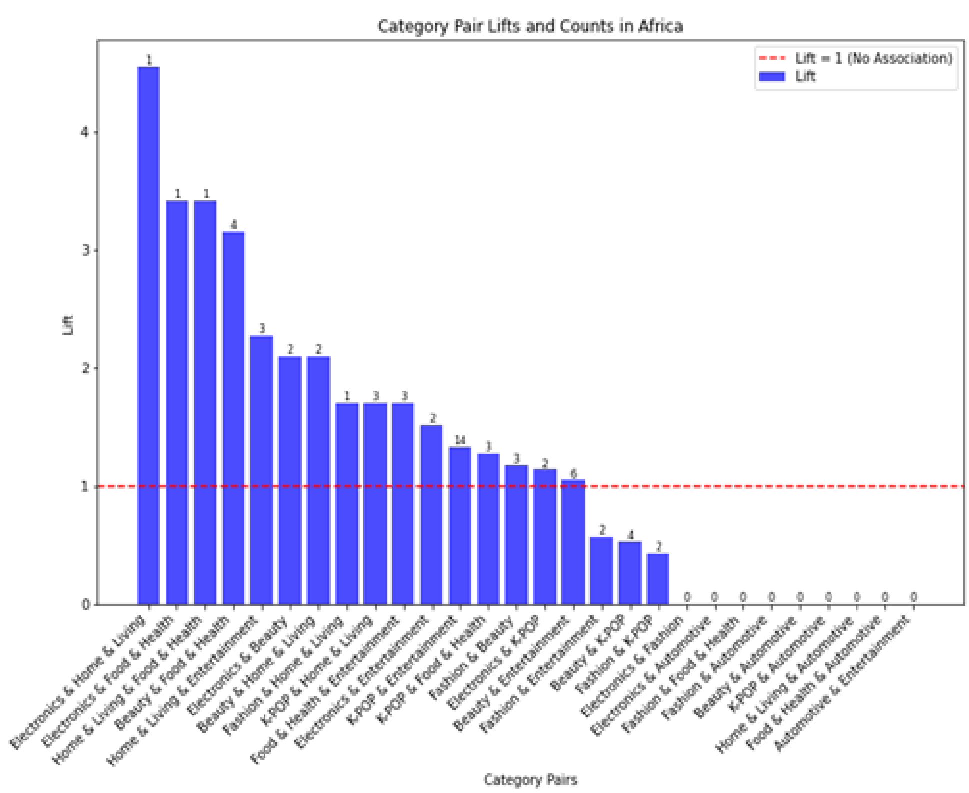


Figure 26.Category Pair Lifts and Counts in Africa

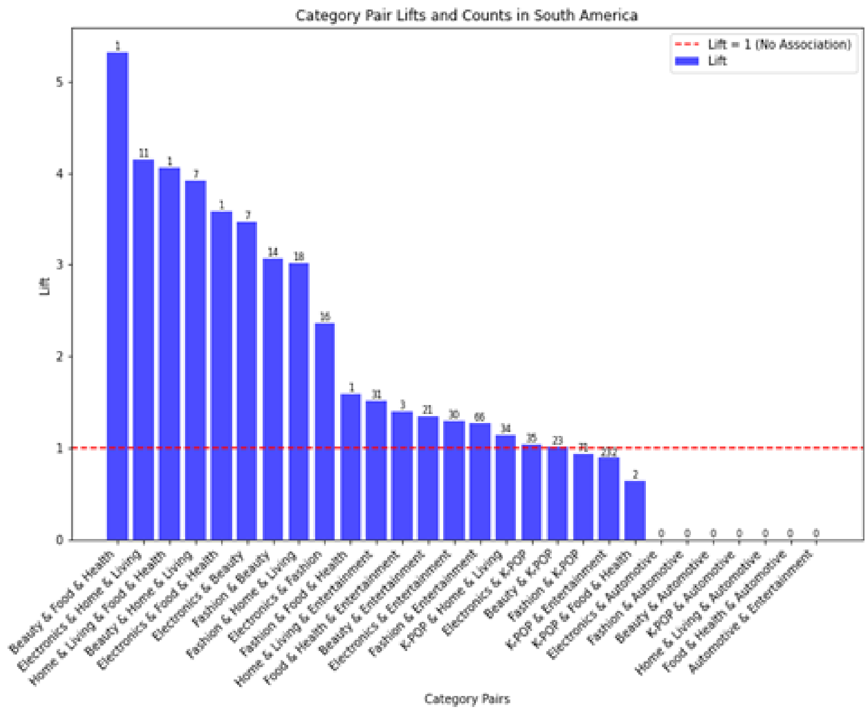


Figure 27.Category Pair Lifts and Counts in South America

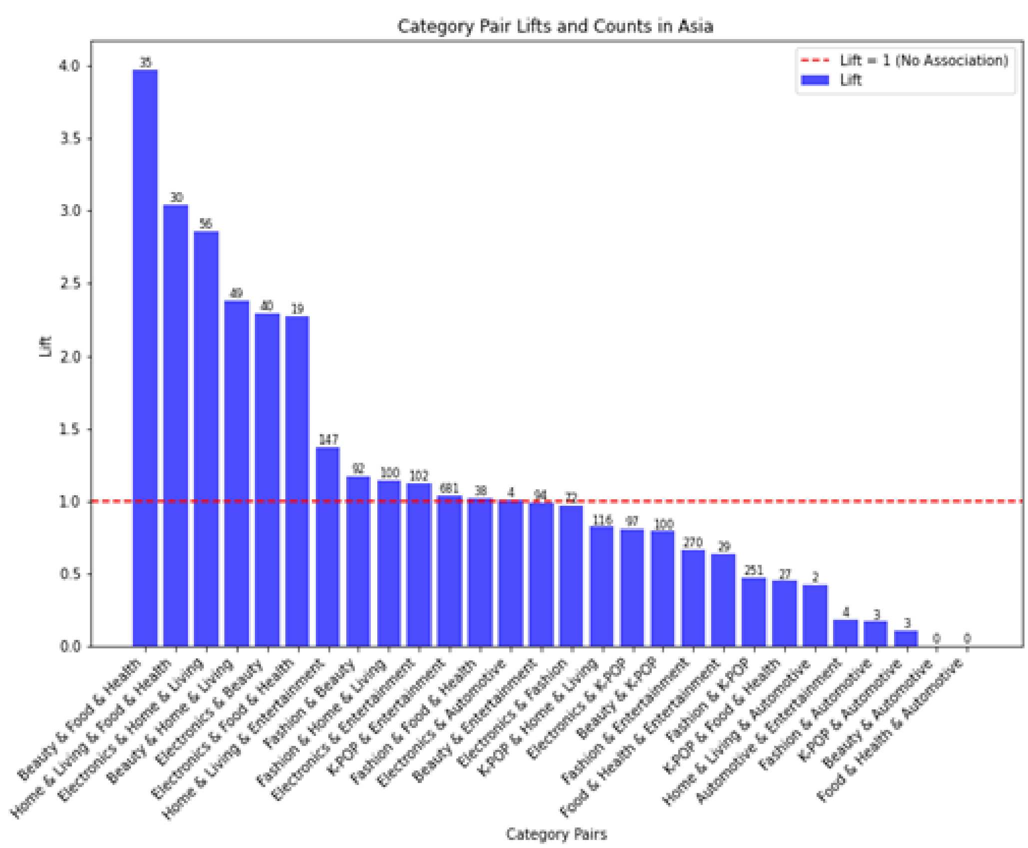


Figure 28.Category Pair Lifts and Counts in Asia

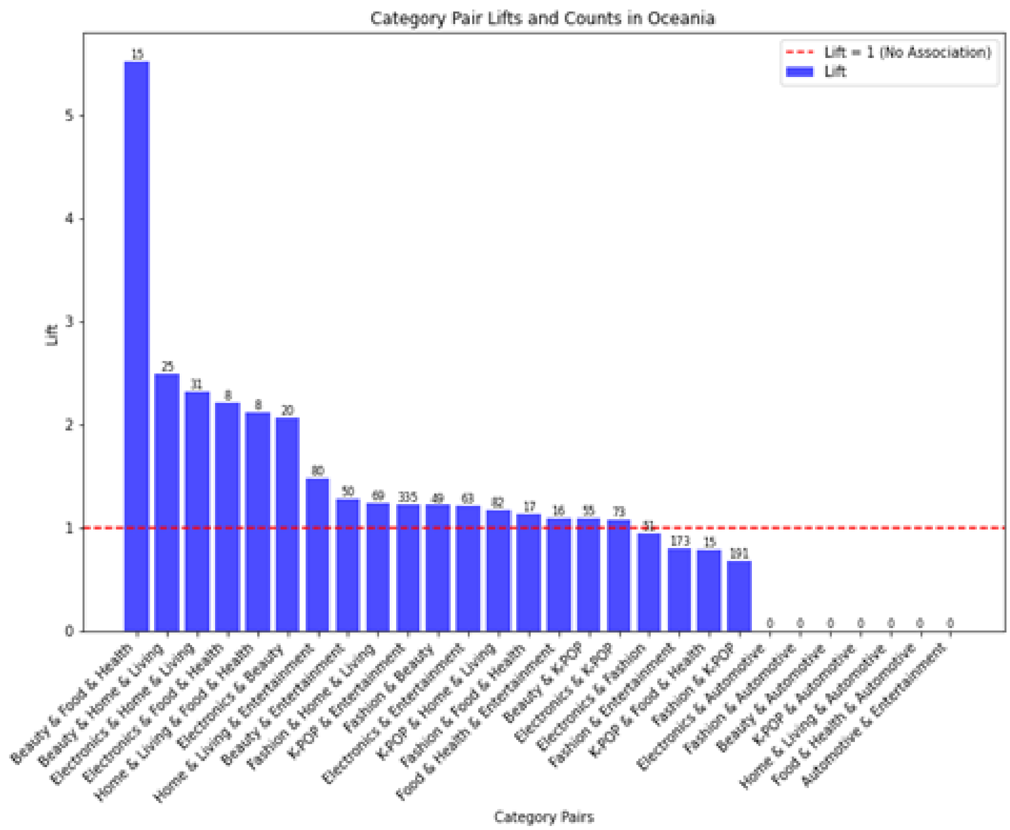


Figure 29.Category Pair Lifts and Counts in Oceania

The limited transaction counts in these regions resulted in less reliable lift values.

However, niche pairings, such as "Beauty → Food & Health" in Asia, still provided actionable insights.

Heatmap Analysis: Lift Values with Pair Counts (Figure 30)

The heatmap presented above provides a comprehensive visualization of lift values alongside pair counts, offering critical insights into category associations within the dataset. This representation is particularly powerful for identifying both the strength of relationships and the frequency of co-occurrence, making it an essential tool for strategic decision-making in recommendation systems.

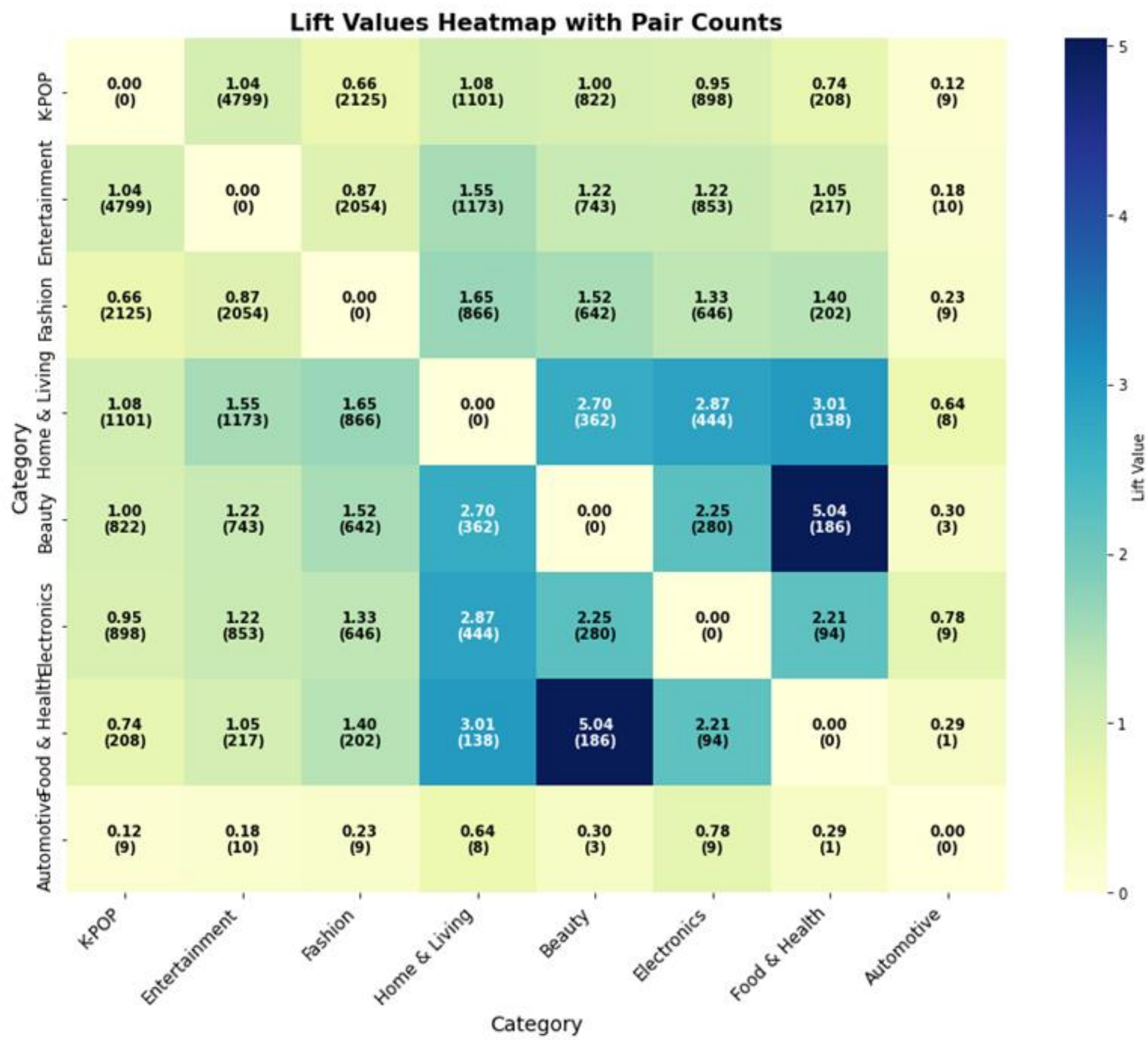


Figure 30.Lift Values Heatmap with Pair Counts

Key Features of the Heatmap

Lift Value Representation:

The heatmap shades range from light to dark, with darker shades indicating higher lift values. This visual distinction helps to quickly identify category pairs with strong associations.

For example, "Beauty → Food & Health" stands out with the highest lift value of 5.04, demonstrating a strong and significant co-purchase relationship.

Pair Counts Overlaid:

Pair counts, displayed in parentheses, complement the lift values by adding context to the significance of each relationship.

High lift values with substantial pair counts (e.g., "Home & Living → Electronics" with a lift of 2.87 and 444 occurrences) highlight meaningful and actionable insights for cross-selling or bundling strategies.

Diagonal Values:

The diagonal elements represent self-pairings (e.g., "K-POP → K-POP") and are logically assigned a lift value of 0 since they do not reflect cross-category associations. This helps maintain focus on inter-category relationships.

Significance and Unique Contributions of the Heatmap

Dual Context: Strength and Frequency:

By combining lift values with pair counts, the heatmap provides a dual-layered analysis. High lift values alone may indicate a strong relationship, but without corresponding counts, their business relevance diminishes. The overlay of counts ensures that only significant and practical pairings are considered.

Strategic Decision-Making:

The heatmap aids in prioritizing categories for recommendation strategies. For example, "Home & Living → Beauty" with a lift of 3.01 and 138 occurrences is a notable pair worth exploring in promotional campaigns targeting diverse product interests.

Regional Customization:

The heatmap’s format can be adapted to regional datasets, offering a granular view of category relationships in different markets. This enhances the precision of recommendations for specific customer segments.

Quick Identification of Weak Associations:

Low lift values and minimal pair counts (e.g., "Automotive → K-POP" with a lift of 0.12 and 9 occurrences) help filter out irrelevant or insignificant relationships, ensuring resources are allocated efficiently.

Clarity of Insights:

The heatmap visually simplifies a complex dataset, making it easier to interpret for stakeholders and readers.

Comprehensive Analysis:

It demonstrates the application of lift values and pair counts in a practical and business-oriented context, aligning with the objectives of the study.

Support for Recommendations:

By providing actionable insights into cross-category relationships, the heatmap directly supports the thesis's goal of optimizing product recommendations for Delivered Korea.

Conclusion

The heatmap of lift values with pair counts serves as an invaluable analytical tool, bridging the gap between theoretical metrics and practical business applications.

Association Rule Model Framework

The association rule model employs a four-step process, illustrated in the first diagram. It begins with a training phase, utilizing 70% of the dataset to calculate lift values between categories, identifying the most frequently associated items. The prediction phase applies these rules to forecast the likely purchases of customers, based on past behavior. The model evaluates its effectiveness by comparing predicted purchases against actual ones in the test set, ensuring the recommendations align with reality.

Training Phase: The lift values between categories are computed to reveal which products are commonly purchased together.

Prediction Phase: (Figure 30)

Step 1: Patterns are identified, such as "customers buying K-POP also buy Entertainment."

Step 2: These patterns are used to predict future purchases for individual customers.

Prediction Performance

The predicted results are validated by comparing them with actual customer behavior.

Confidence and Lift in Rule Generation

The second diagram explains the role of confidence and lift in generating product recommendation rules. Confidence measures the likelihood of a consequent (e.g., buying a product) given the antecedent, with a threshold of 30% ensuring relevance. Lift values greater than 1 indicate a meaningful relationship beyond random association, strengthening the predictive capability of the model. For instance:

If "K-POP" has a high confidence of association with "Entertainment," the model recommends these categories together.

Metrics in the Model Process

The third diagram presents a detailed breakdown of the model process and metrics in tabular form:

Find Frequent Itemsets: The model identifies combinations of products that occur frequently, based on a minimum support threshold.

Generate Association Rules: Rules are filtered using a minimum confidence threshold (e.g., 0.4) to ensure reliability.

Evaluate Rules: The lift value validates the strength of these associations. A lift > 1 confirms a significant relationship.

Predict Purchases: Using filtered rules, the model predicts additional purchases customers might make.

The role of each metric is critical:

Support ensures statistical significance.

Confidence validates reliability.

Lift measures the strength of associations, enabling actionable insights.

	Step	What Happens	Role of Support	Role of Confidence	Role of Lift
1	Find Frequent Itemsets	The model identifies combinations of products (itemsets) that occur frequently in the training data. This is based on a minimum support threshold (e.g., 0.01).	Ensures the itemsets are frequent enough to be meaningful. For example, if a combination occurs in 1% or more of transactions, it is kept.		
2	Generate Association Rules	The model generates rules from frequent itemsets. Each rule has an antecedent (e.g., K-POP) and a consequent (e.g., Entertainment). Rules are filtered based on a minimum confidence threshold (e.g., 0.4).		Filters the rules to ensure reliability. For example, if at least 40% of K-POP buyers also buy Entertainment, the rule is kept.	
3	Evaluate Rules	Each rule is evaluated using lift to check how strong the association is. Lift > 1 indicates a meaningful relationship, lift = 1 indicates no relationship, and lift < 1 indicates a negative relationship.			Measures the strength of the relationship between products. Higher lift values (>1) indicate a strong association beyond random chance.
4	Predict Purchases	The model uses the filtered rules to predict which products customers might buy, based on what they have already purchased.			

Figure 30. Steps in Generating and Evaluating Association Rules with Support, Confidence, and Lift"

Code Implementation for Association Rule Mining and Predictive Modeling (Figure 31)

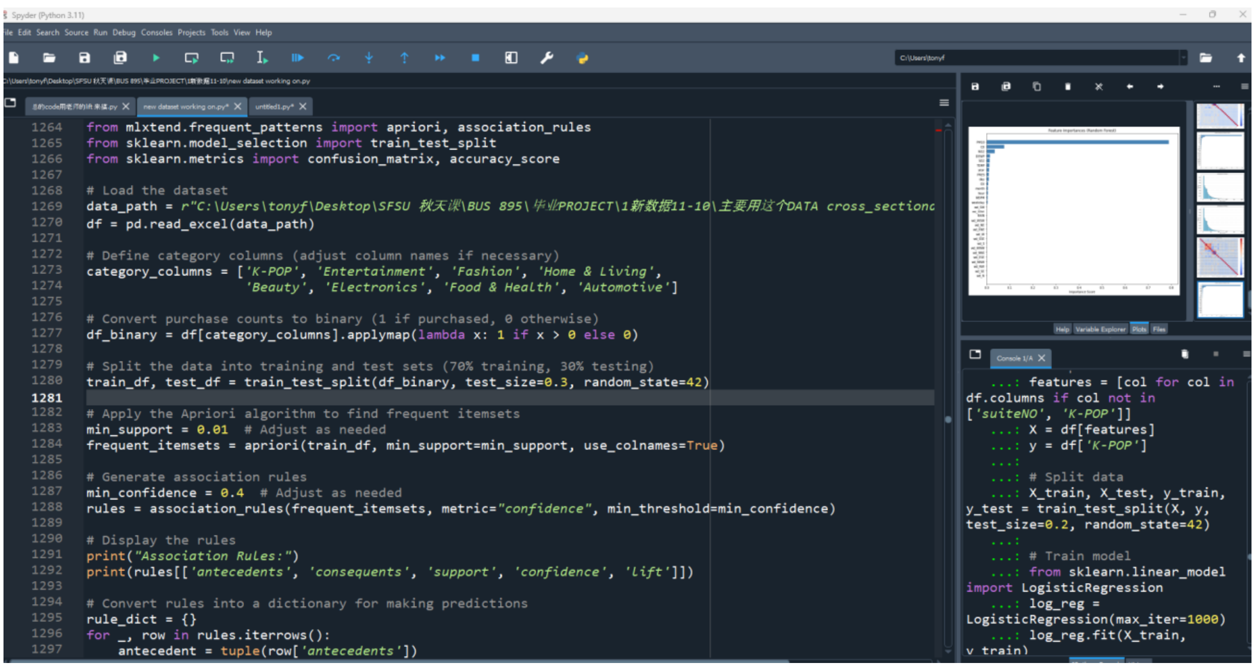


Figure 31. Code Implementation for Association Rule Mining and Predictive Modeling

This section presents the technical implementation of the Apriori algorithm and predictive modeling processes, outlining their roles in generating actionable product recommendations for Delivered Korea. By adhering to a systematic approach, this study bridges theoretical methodologies with practical applications, providing a robust framework for analyzing customer behavior and optimizing recommendations.

Principles and Methodology

The implementation is guided by the following principles:

1. **Data Preparation:** Ensuring the dataset is clean and ready for analysis through transformations like binary encoding of purchase data.
2. **Frequent Itemset Mining**: Identifying combinations of frequently purchased items using the Apriori algorithm.
3. **Association Rule Generation**: Creating rules based on predefined thresholds for support, confidence, and lift.
4. **Predictive Modeling**: Applying the rules to a test dataset to make predictions and evaluate their effectiveness.

Coding Steps and Workflow

1. Data Preparation

The dataset was first preprocessed to focus on customer purchase patterns. Purchase counts were transformed into binary format, where 1 indicates a purchase and 0 indicates no purchase. This transformation simplifies the input data for association rule mining.

```
df_binary = df[category_columns].applymap(lambda x: 1 if x > 0
else 0)
```

This step ensures consistency in data representation and is critical for the effective application of the Apriori algorithm.

2. Splitting Data into Training and Testing Sets

The dataset was divided into 70% training and 30% testing subsets to evaluate the model's predictive capabilities. This split ensures that the rules generated are tested on unseen data, providing a realistic assessment of their accuracy.

3. Mining Frequent Itemsets

The Apriori algorithm was applied to the training dataset to identify frequent itemsets that meet the minimum support threshold of 1% (min_support=0.01). This parameter specifies the minimum proportion of transactions in which an itemset must appear to be considered frequent. For instance, a min_support of 0.01 means that only itemsets appearing in at least 1% of the transactions will be retained.

This threshold is carefully chosen to balance the inclusion of meaningful patterns while excluding noise from infrequent combinations, ensuring that the generated itemsets are both statistically significant and practically relevant.

```
frequent_itemsets = apriori(train_df, min_support=0.01,
use_colnames=True)
```

This step uncovers foundational relationships between product categories, such as frequent co-purchases, which are essential for generating association rules.

4. Generating Association Rules

Rules were generated from the frequent itemsets using a confidence threshold of 40% (min_threshold=0.4), ensuring the reliability of the associations. The confidence metric measures the likelihood that the consequent (predicted item) is purchased, given that the antecedent (already purchased item) has occurred.

For example, a confidence threshold of 40% means that at least 40% of the transactions containing the antecedent must also contain the consequent for the rule to be considered valid. This threshold helps strike a balance between including meaningful associations and avoiding rules that might occur due to chance.

Each generated rule consists of:

- **Antecedent:** The item(s) already purchased by a customer.
- **Consequent:** The item(s) predicted to be purchased based on the antecedent.
- **Metrics:** Support, confidence, and lift, which evaluate the statistical significance and strength of the association.

```
rules = association_rules(frequent_itemsets,
metric="confidence", min_threshold=0.4)
```

This step is critical for identifying actionable insights from the data, forming the foundation for recommendation strategies based on customer purchasing patterns.

5. Storing Rules in a Dictionary

The rules were stored in a dictionary to facilitate quick lookup during the prediction phase. Each antecedent maps to its corresponding consequent, confidence score, and lift value.

6. Predictive Modeling

To generate predictions based on the association rules, a function was developed to apply the rules to the test dataset. This function compares each customer's purchased categories against the antecedents in the rule dictionary. If a rule's antecedent matches the customer's purchases, the corresponding consequent category is predicted as a likely purchase.

The logic works as follows:

1. **Input:** The test dataset (test_data), rule dictionary (rule_dict), and the list of category columns (category_columns).
2. **Purchased Categories Identification:** For each customer, their purchased categories are identified by filtering the test data for categories with a value of 1.
3. **Rule Matching:** The antecedents in the rule dictionary are checked against the customer's purchased categories. If all items in an antecedent are present in the customer's purchases, the consequent category associated with that rule is added to the predicted categories.
4. **Output:** A dictionary containing:
    - **Purchased:** The actual categories the customer purchased.
    - **Predicted:** The categories predicted based on the rules.

The function implementation is as follows:

```
def make_predictions_using_rules(test_data, rule_dict,
category_columns):

predictions = {}
     for idx, row in test_data.iterrows():
        purchased_categories = [cat for cat in
category_columns if row[cat] == 1]
        predicted_categories = set()
        for antecedent, (consequent, confidence) in
rule_dict.items():
            if all(item in purchased_categories for item in
antecedent):
                predicted_categories.add(consequent)
        predictions[idx] = {
```

```
            'Purchased': purchased_categories,
            'Predicted': list(predicted_categories)
        }
    return predictions
```

This function is critical for applying the association rules to new data, enabling the system to generate personalized product recommendations for customers based on their past purchasing behavior. It bridges the gap between rule generation and practical application in predictive modeling.

7. Evaluating Predictions

The predictions were evaluated against actual purchases in the test data. Key performance metrics included:

- **Overall Accuracy**: The proportion of correct predictions across all categories.
- Category-Specific Accuracy: The accuracy for each individual category.
- Confusion Matrix: A detailed breakdown of true positives, false positives, true negatives, and false negatives.

8. Visualizing Results

Results were visualized to highlight the model's performance. Accuracy scores for each category were presented as bar plots, and confusion matrices provided a granular view of prediction accuracy.

```
plt.figure(figsize=(10, 6))
sns.barplot(x=list(category_accuracies.keys()),
y=list(category_accuracies.values()))
plt.title('Accuracy for Each Category')
plt.xlabel('Category')
plt.ylabel('Accuracy')
```

```
plt.xticks(rotation=45)
plt.show()
```

Additionally, the overall accuracy and the distribution of correct vs. incorrect predictions were illustrated to provide stakeholders with an accessible understanding of the model's capabilities.

Key Insights from the Implementation

Efficiency and Interpretability: The structured use of the Apriori algorithm ensures that rules are both computationally efficient and easy to interpret, making them suitable for business applications.

Scalability: The modular design of the code allows for adjustments in support, confidence, and lift thresholds, making it adaptable to datasets of varying sizes and characteristics.

Actionable Results: By focusing on high-confidence and high-lift associations, the model prioritizes rules that offer the greatest potential for cross-selling and customer engagement.

Significance

This coding framework is integral , demonstrating the practical application of advanced analytical techniques to real-world e-commerce challenges. By transforming theoretical principles into actionable insights, this implementation highlights the effectiveness of data-driven approaches in optimizing product recommendations. The methodology, combined with its visual and statistical outputs, provides a comprehensive foundation for Delivered Korea's strategic decision-making processes.

Results and Discussion

Holistic Assessment of Prediction Accuracy Across Categories

Objectives

The analysis of prediction accuracy across categories aims to:

Holistically Evaluate Accuracy: Assess how well the model predicts customer purchases across all categories. By considering all products collectively, this metric provides an overarching view of the model's effectiveness.

Implementation in Python:

```
rule_dict = {}
for _, row in rules.iterrows():
    antecedent = tuple(row['antecedents'])
    consequent = list(row['consequents'])[0]
    confidence = row['confidence']
    rule_dict[antecedent] = (consequent, confidence)

def make_predictions_using_rules(test_data, rule_dict,
category_columns):
    predictions = {}
    for idx, row in test_data.iterrows():
        purchased_categories = [cat for cat in category_columns
if row[cat] == 1]
        predicted_categories = set()
        for antecedent, (consequent, confidence) in
rule_dict.items():
            if all(item in purchased_categories for item in
antecedent):
                predicted_categories.add(consequent)
        predictions[idx] = {'Purchased': purchased_categories,
'Predicted': list(predicted_categories)}
    return predictions
```

Model Performance Evaluation:

Confusion Matrix: Captures True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) to assess prediction performance.

Overall Accuracy Calculation:
Accuracy=Number of Correct PredictionsTotal Number of Predictions\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}Accuracy=Total Number of PredictionsNumber of Correct Predictions

Results

The predictions were evaluated using a confusion matrix and visual plots: (Figure 32)

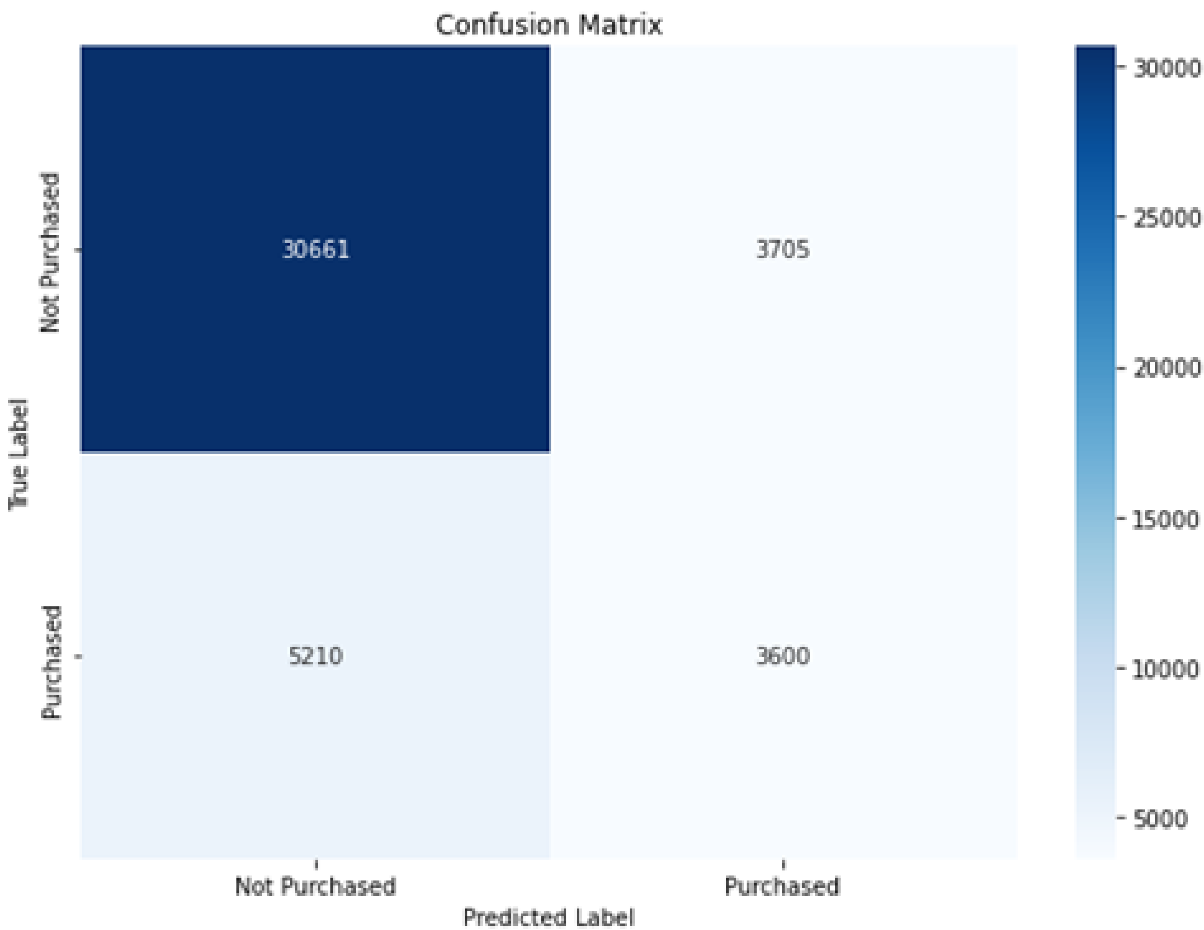


Figure 32.Confusion Matrix

Confusion Matrix Details:

- True Negatives (TN = 30,661): Non-buyers correctly identified.
- False Positives (FP = 3,705): Predicted as buyers but did not purchase.
- False Negatives (FN = 5,210): Missed potential buyers.
- True Positives (TP = 3,600): Correctly identified buyers.

- Accuracy: The model achieved a 79% accuracy (Figure 33)

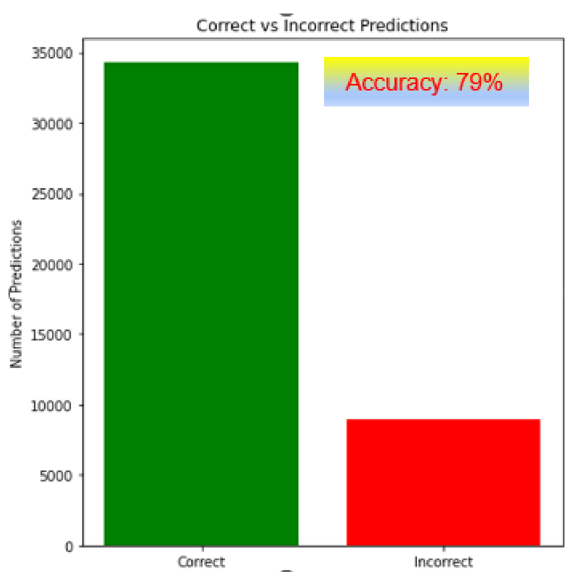


Figure 33.Correct vs Incorrect Predictions

Visual Analysis

1. Confusion Matrix Heatmap: Displays classification results with distinct sections for TN, FP, FN, and TP.
2. Correct vs. Incorrect Predictions:

- Correct predictions (green) dominate, highlighting the model's effectiveness.
- Incorrect predictions (red) provide room for refinement.

3. Accuracy Explanation:

- For five customers and eight categories, there are 5×8=405 \times 8 = 405×8=40 predictions. If 30 predictions are correct, accuracy is 30/40=75%30/40 = 75\%30/40=75%. This simple breakdown communicates the prediction process to a general audience.

Implications

The holistic assessment underscores the model's utility in capturing general trends while revealing areas for improvement:

1. Strengths:

High accuracy across multiple categories indicates reliable recommendations for broad customer segments.

The model effectively identifies true non-buyers, conserving marketing resources.

2. Challenges:

False positives and negatives highlight the need for enhanced precision in targeting actual buyers.

Category-specific insights reveal variability in prediction reliability, informing where additional focus is required.

Summary

This analysis demonstrates how association rules can effectively predict purchase patterns in e-commerce. By leveraging both holistic and detailed accuracy metrics, Delivered Korea can refine its product recommendation strategies to enhance customer engagement and business outcomes.

Per-Consequent Prediction Accuracy

This section presents an analysis of Per-Consequent Prediction Accuracy, focusing on evaluating how accurately the system predicts purchases for each specific product category (consequent) using association rules. This metric is derived from evaluation methods in associative classification, where association rules are used for predictive modeling (Liu et al., 1998; Thabtah, 2007).

Definition and Formula

The prediction accuracy for each consequent category is calculated as:

$$\text{Per-Consequent Prediction Accuracy} = \frac{\text{Number of Correct Predictions for the Consequent}}{\text{Total Number of Predictions Made for the Consequent}}$$

This formula measures the reliability of the model in predicting each specific category, assessing the quality of category-specific recommendations. It is crucial for:

Tailoring marketing strategies to align with customer preferences.

Identifying areas for improvement in the recommendation algorithms.

Example Calculation

For instance:

If the system predicts "K-POP" for 10 customers and 7 of those customers actually purchase "K-POP," the prediction accuracy for the "K-POP" consequent is:

Per-Consequent Prediction Accuracy for K-POP= 7/10 =70%

Dataset Breakdown and Analysis Workflow

Each row in the test dataset is structured to include the following key components:

Antecedents: Categories the customer has purchased.

Consequent: Categories predicted by the model using association rules.

Actual: Categories the customer actually purchased.

Correct Prediction Indicator:: A binary indicator (1 for correct, 0 for incorrect) reflecting whether the prediction matches the actual purchase.

Code and Implementation Steps

To implement the predictive framework, the following key steps were followed:

1. Frequent Itemset Mining:

The Apriori algorithm identifies frequent itemsets based on a minimum support threshold.

This ensures that only significant patterns are used for generating recommendations.

2. Rule Generation:

Association rules are created from frequent itemsets using confidence thresholds to filter reliable rules.

Rules include metrics like support, confidence, and lift to evaluate their strength and reliability.

3. Prediction Generation:

For each transaction in the test dataset, the system evaluates whether a customer's purchase matches the antecedents of any generated rules.

Predictions are made for corresponding consequents based on these matches.

Below is a code snippet highlighting the key steps for rule generation and prediction:

Generate association rules

```
min_confidence = 0.4  # Confidence threshold for filtering
rules
rules = association_rules(frequent_itemsets,
metric="confidence", min_threshold=min_confidence)

# Prediction logic
for idx, row in test_df.iterrows():
    suite_no = df.iloc[idx].get('suiteNO', f'Index-{idx}')
    purchased_categories = [cat for cat in category_columns
if row[cat] == 1]
    predicted_categories = set()

    # Apply rules to generate predictions
    for _, rule_row in rules.iterrows():
        antecedent = tuple(rule_row['antecedents'])
        consequent = list(rule_row['consequents'])[0]
        if all(item in purchased_categories for item in
antecedent):
```

```
predicted_categories.add(consequent)
```

SuiteNo	Antecedent	Consequent	Support	Confidence	Lift Value	Prediction	Actual	Correct
F6503	Entertainment, Electronics	Home & Living	0.022631621	0.478991597	4.980958858	1	0	0
F6503	Electronics, Fashion	Home & Living	0.018819979	0.520879121	5.416540685	1	0	0
F6503	Entertainment, K-POP, Electronics	Fashion	0.027237354	0.681908549	2.265771597	1	1	1
F6503	Entertainment, Electronics, Fashior	K-POP	0.027237354	0.900262467	1.534931661	1	1	1
F6503	Electronics, K-POP, Fashion	Entertainment	0.027237354	0.897905759	2.059622445	1	1	1
F6503	Entertainment, Electronics	K-POP	0.027237354	0.576470588	4.885258491	1	1	1
F6503	Electronics, K-POP	Entertainment	0.027237354	0.541009464	4.75431415	1	1	1
F6503	Electronics, Fashion	Entertainment	0.027237354	0.753846154	2.843973821	1	1	1
F6503	Entertainment, K-POP, Electronics	Home & Living	0.020169936	0.504970179	5.25110608	1	0	0
F6503	Entertainment, Electronics	Home & Living	0.020169936	0.426890756	6.830794529	1	0	0
F6503	Electronics, K-POP	Entertainment	0.020169936	0.400630915	6.115327407	1	1	1
F6503	Electronics, K-POP, Fashion	Home & Living	0.01675534	0.552356021	5.743864056	1	0	0
F6503	Electronics, Fashion	Home & Living	0.01675534	0.463736264	7.420369465	1	0	0
F6503	Entertainment, Electronics, Fashior	Home & Living	0.017549432	0.580052493	6.031875351	1	0	0
F6503	Electronics, Fashion	Entertainment	0.017549432	0.485714286	7.414060606	1	1	1
F6503	Electronics, K-POP, Entertainment,	Home & Living	0.015961248	0.586005831	6.093783178	1	0	0
F6503	Electronics, Entertainment, Fashior	Home & Living	0.015961248	0.527559055	8.441615224	1	0	0
F6503	Electronics, K-POP, Fashion	Entertainment	0.015961248	0.52617801	8.03170871	1	1	1

Figure 34 . Key Metrics for Predictions Generated from Association Rules

Results and Observations

The table (Figure 34) generated from this analysis highlights the following key metrics for each prediction:

1. Support: The frequency of the antecedent in the dataset, which indicates its significance.
2. Confidence: The likelihood of the consequent being purchased when the antecedent is purchased.
3. Lift Value: The strength of association between antecedent and consequent categories.
4. Prediction vs. Actual: Evaluates the system's reliability in predicting customer behavior.

Per-Consequent Prediction Accuracy Analysis

To quantify the reliability of specific category recommendations, Per-Consequent

Breakdown of Predictions by Antecedent and Consequent Categories (Figure 35,36,37,38,39)

SuiteNo	Antecedent	Consequen	Support	Confidence	Lift Value	Prediction	Actual	Correct
E7973	Electronics	K-POP	0.05034543	0.572202166	0.975594622	1	0	0
E7973	Electronics, Fashion	K-POP	0.030334313	0.83956044	1.431435772	1	0	0
A6145	Entertainment	K-POP	0.265067895	0.608014572	1.036654144	1	0	0
H1467	Entertainment	K-POP	0.265067895	0.608014572	1.036654144	1	0	0
B7433	Entertainment	K-POP	0.265067895	0.608014572	1.036654144	1	1	1
B7433	Electronics	K-POP	0.05034543	0.572202166	0.975594622	1	1	1
B7433	Entertainment, Electronics	K-POP	0.039942825	0.845378151	1.441354869	1	1	1
B7433	Entertainment, Electronics	K-POP	0.027237354	0.576470588	4.885258491	1	1	1
H6258	Entertainment	K-POP	0.265067895	0.608014572	1.036654144	1	1	1
B9148	Entertainment	K-POP	0.265067895	0.608014572	1.036654144	1	0	0
B9148	Home & Living	K-POP	0.062495037	0.649876135	1.108027373	1	0	0

SuiteNo	Antecedent	Consequent	Support	Confidence	Lift Value	Prediction	Actual	Correct
K3710	Entertainment, Home & Living, K-POP, Fashion	Electronics	0.015961248	0.442731278	5.031872724	1	0	0
K3710	Home & Living, K-POP, Fashion	Electronics	0.015961248	0.406060606	8.594153298	1	0	0
A7586	Home & Living, K-POP, Fashion	Electronics	0.01675534	0.426262626	4.844697881	1	1	1
A7586	Entertainment, Home & Living, Fashion	Electronics	0.017549432	0.413857678	4.70370915	1	1	1
A7586	Entertainment, Home & Living, K-POP, Fashion	Electronics	0.015961248	0.442731278	5.031872724	1	1	1
A7586	Home & Living, K-POP, Fashion	Electronics	0.015961248	0.406060606	8.594153298	1	1	1
B0492	Home & Living, K-POP, Fashion	Electronics	0.01675534	0.426262626	4.844697881	1	0	0
B0492	Entertainment, Home & Living, Fashion	Electronics	0.017549432	0.413857678	4.70370915	1	0	0
B0492	Entertainment, Home & Living, K-POP, Fashion	Electronics	0.015961248	0.442731278	5.031872724	1	0	0
B0492	Home & Living, K-POP, Fashion	Electronics	0.015961248	0.406060606	8.594153298	1	0	0
B4655	Home & Living, K-POP, Fashion	Electronics	0.01675534	0.426262626	4.844697881	1	0	0

SuiteNo	Antecedent	Consequent	Support	Confidence	Lift Value	Prediction	Actual	Correct
E7973	Electronics	Entertainment	0.047248471	0.53700361	1.231782598	1	0	0
E7973	Electronics	Entertainment	0.039942825	0.453971119	1.712659767	1	0	0
E7973	Electronics, Fashion	Entertainment	0.030254904	0.837362637	1.920748214	1	0	0
E7973	Electronics, Fashion	Entertainment	0.027237354	0.753846154	2.843973821	1	0	0
E7973	Electronics, Fashion	Entertainment	0.017549432	0.485714286	7.414060606	1	0	0
E7973	Electronics, Fashion	Entertainment	0.015961248	0.441758242	8.217225315	1	0	0
F6114	K-POP	Entertainment	0.265067895	0.451936095	1.036654144	1	0	0
L9137	K-POP	Entertainment	0.265067895	0.451936095	1.036654144	1	0	0
I9136	K-POP	Entertainment	0.265067895	0.451936095	1.036654144	1	0	0
H4736	K-POP	Entertainment	0.265067895	0.451936095	1.036654144	1	0	0
F3962	K-POP	Entertainment	0.265067895	0.451936095	1.036654144	1	0	0
B7433	K-POP	Entertainment	0.265067895	0.451936095	1.036654144	1	1	1
B7433	Electronics	Entertainment	0.047248471	0.53700361	1.231782598	1	1	1
B7433	Electronics, K-POP	Entertainment	0.039942825	0.793375394	1.819849971	1	1	1

SuiteNo	Antecedent	Consequent	Support	Confidence	Lift Value	Prediction	Actual	Correct
E7973	Electronics	Fashion	0.036131184	0.410649819	1.36446258	1	1	1
B7433	Electronics	Fashion	0.036131184	0.410649819	1.36446258	1	0	0
B7433	Electronics, K-POP	Fashion	0.030334313	0.602523659	2.002000117	1	0	0
B7433	Entertainment, Electronics	Fashion	0.030254904	0.640336134	2.127639298	1	0	0
B7433	Entertainment, K-POP, Electronics	Fashion	0.027237354	0.681908549	2.265771597	1	0	0
B9148	Home & Living	Fashion	0.049630747	0.516102395	1.714848933	1	0	0
B9148	Entertainment, Home & Living	Fashion	0.04240451	0.647272727	2.150687455	1	0	0
F6503	Electronics	Fashion	0.036131184	0.410649819	1.36446258	1	1	1
F6503	Electronics, K-POP	Fashion	0.030334313	0.602523659	2.002000117	1	1	1
F6503	Entertainment, Electronics	Fashion	0.030254904	0.640336134	2.127639298	1	1	1

SuiteNo	Antecedent	Consequent	Support	Confidence	Lift Value	Prediction	Actual	Correct
D1743	Entertainment, Beauty, Fashion	Home & Living	0.013578972	0.518181818	5.388491855	1	0	0
D1743	Entertainment, Electronics, Fashion	Home & Living	0.017549432	0.580052493	6.031875351	1	0	0
D1743	Entertainment, Beauty, K-POP, Fashion	Home & Living	0.012149607	0.523972603	5.44870932	1	0	0
D1743	Entertainment, Beauty, Fashion	Home & Living	0.012149607	0.463636364	7.418770937	1	0	0
D1743	Electronics, K-POP, Entertainment, Fashion	Home & Living	0.015961248	0.586005831	6.093783178	1	0	0
D1743	Electronics, Entertainment, Fashion	Home & Living	0.015961248	0.527559055	8.441615224	1	0	0
A7586	Electronics, K-POP	Home & Living	0.021281664	0.422712934	4.395725826	1	1	1
A7586	Entertainment, Electronics	Home & Living	0.022631621	0.478991597	4.980958858	1	1	1
A7586	Beauty, Fashion	Home & Living	0.01484952	0.426940639	4.439689075	1	1	1
A7586	Electronics, Fashion	Home & Living	0.018819979	0.520879121	5.416540685	1	1	1
A7586	Entertainment, Beauty, K-POP	Home & Living	0.014531883	0.413092551	4.295684965	1	1	1
A7586	Entertainment, K-POP, Electronics	Home & Living	0.020169936	0.504970179	5.25110608	1	1	1
A7586	Entertainment, Electronics	Home & Living	0.020169936	0.426890756	6.830794529	1	1	1

Figure 35,36,37,38,39. Breakdown of Predictions by Antecedent and Consequent Categories

A detailed view of the predictions shows:

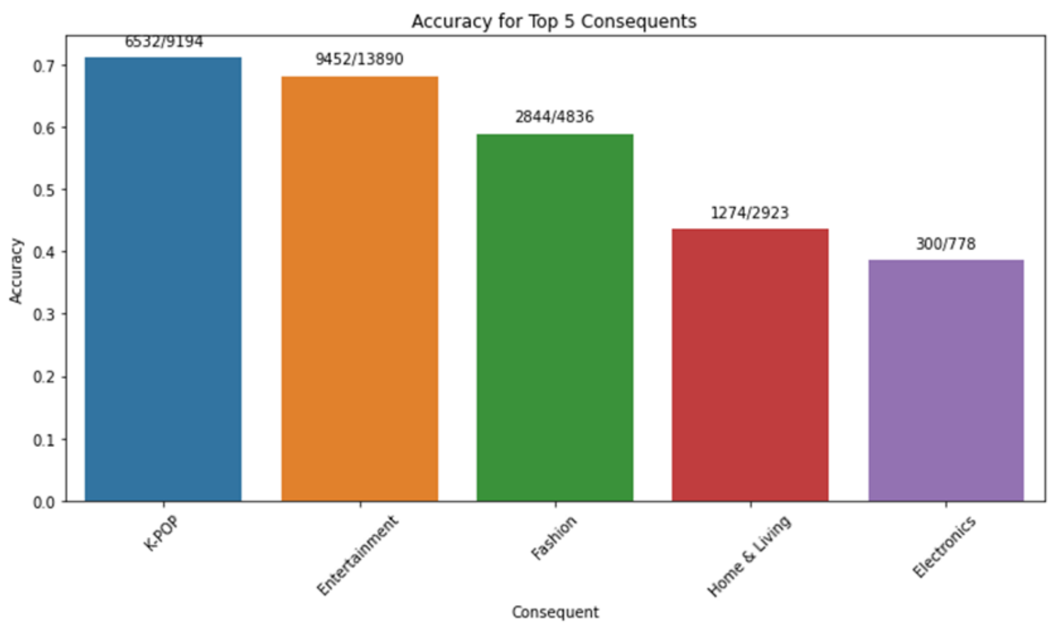
1. High-Lift Pairs:

Categories such as "K-POP" and "Entertainment" exhibit high lift values, indicating strong associations that are reliable for cross-selling strategies.

2. Low-Lift Pairs:

Pairs with lower lift values and prediction accuracy suggest potential areas for improvement in data preprocessing or model refinement.

Accuracy Analysis for Top 5 Consequents (Figure 40)



Consequent	Accuracy
K-POP	0.7104633
Entertainment	0.6804895
Fashion	0.5880893
Home & Living	0.4358535
Electronics	0.3856041

Figure 40. Accuracy Analysis for Top 5 Consequents

The analysis focused on the top 5 recommended categories, including:

1. K-POP:

Achieved the highest accuracy (71.05%) with substantial support (6,532 correct predictions out of 9,194 total).

2. Entertainment:

Recorded an accuracy of 68.05%, showing its strength as a secondary recommendation category.

3. Fashion:

Noted for moderate accuracy (58.80%), reflecting customer diversity in purchasing patterns.

4. Home & Living and Electronics:

Both exhibited lower accuracy, suggesting these categories might require refined rules or additional data to enhance predictive performance.

The bar chart provided a visual summary of these accuracies, helping stakeholders identify strengths and weaknesses across different product categories.

Implications for Business Strategies

1. Marketing Insights:

High-Accuracy Categories: Categories like "K-POP" and "Entertainment" can be prioritized for promotional campaigns to leverage their strong predictive associations.

2. Operational Efficiency:

Inventory Management: Understanding category-specific prediction quality supports better inventory management and supply chain decisions, reducing overstock and stockouts.

3.Recommendation Optimization:

Model Refinement: Insights from lower-accuracy categories inform improvements in model parameters, such as fine-tuning confidence thresholds or leveraging hybrid recommendation approaches.

By incorporating these results and coding steps, this section demonstrates the effectiveness of association rule mining for targeted marketing while acknowledging opportunities for further refinement in model performance. This analysis strengthens the predictive framework and provides actionable insights for Delivered Korea to enhance customer satisfaction and operational strategies.

Relation to Associative Classification

The approach aligns with evaluation methods used in associative classification, where association rules are employed for predictive tasks, and performance is measured using accuracy metrics (Liu et al., 1998; Thabtah, 2007). By assessing the per-consequent prediction accuracy, we can determine how effectively each rule predicts its consequent, similar to how class-wise accuracy is evaluated in classification models.

Conclusion

By evaluating the Per-Consequent Prediction Accuracy, we gain valuable insights into the performance of the association rules for each product category. This analysis helps in:

Identifying which categories the model predicts well, enabling targeted marketing efforts.

Recognizing categories with lower accuracy, prompting further investigation and model refinement.

Enhancing overall business strategies by aligning recommendations with customer purchasing behaviors.

Category-wise Prediction Accuracy (Figure 41)

This section presents an analysis of category-wise prediction accuracy, which focuses on evaluating how well the system predicts purchases for each individual product category. Unlike the holistic and consequent-specific accuracies discussed earlier, this metric provides insights into the model's ability to capture customer behavior at a granular, category-specific level.

Category-wise Prediction Accuracy Formula

The accuracy for each category is calculated as:

$$\text{Category-wise Prediction Accuracy} = \frac{\text{Number of Correct Predictions for the Category}}{\text{Total Actual Purchases of the Category}}$$

SuiteNo	Prediction	Predicted Category	Actual	Actual Category	Correct
A0266	1	Electronics, Entertainment, K-POP, Fashion, Home & Living	1	K-POP, Entertainment, Fashion, Home & Living, Electronics, Food & Health	0
A0519	1	Entertainment	1	K-POP	0
A0523	1	Entertainment, Electronics, K-POP, Fashion	1	K-POP, Entertainment, Fashion, Home & Living	0
A0604	1	Entertainment, K-POP	1	K-POP, Entertainment	1
A0610	1	Electronics, Entertainment, K-POP, Fashion, Home & Living	1	Entertainment, Fashion, Home & Living, Beauty, Electronics	0
A0613	1	Entertainment	1	K-POP	0
A0619	1	Entertainment	1	K-POP	0
A0627	1	Entertainment, K-POP	1	K-POP, Entertainment	1
A0673	1	Entertainment, K-POP	1	K-POP, Entertainment, Fashion	0
A0680	1	Entertainment, Home & Living, K-POP, Fashion	1	K-POP, Entertainment, Fashion, Beauty	0
A0681	1	Entertainment, Home & Living, K-POP, Fashion	1	Entertainment, Fashion, Electronics	0
A0686	1	Entertainment, K-POP	1	K-POP, Entertainment	1
A0694	1	Entertainment, K-POP	1	K-POP, Entertainment	1
A0738	1	Electronics, Entertainment, K-POP, Fashion, Home & Living	1	K-POP, Entertainment, Fashion, Home & Living, Beauty, Electronics	0
A0741	1	Electronics, Entertainment, K-POP, Fashion, Home & Living	1	K-POP, Entertainment, Fashion, Home & Living, Electronics	1
A0744	1	Entertainment, K-POP	1	K-POP, Entertainment, Fashion	0

Figure 41.Category-Wise Prediction Accuracy

This formula measures how effectively the model predicts customer purchases for specific categories based on historical patterns.

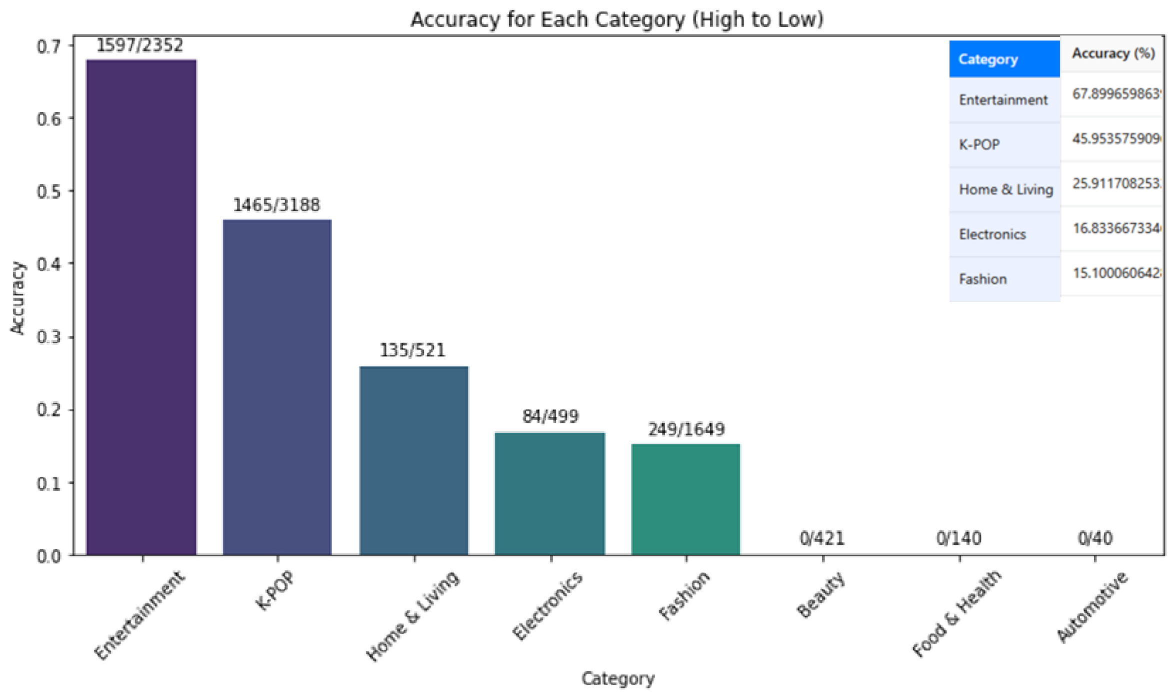


Figure 42.Accuracy for Each Category

Observations and Results

The bar chart illustrates the prediction accuracy across categories. Key observations include:

1. Entertainment (67.9%):

Achieved the highest category-wise accuracy, reflecting strong predictive performance for this widely purchased category. The high volume of transactions for "Entertainment" likely provided the model with sufficient data to identify reliable patterns.

2. K-POP (45.9%):

The second most accurate category. Although widely purchased, its moderate accuracy suggests variability in customer preferences within the category.

3. Home & Living (25.9%):

Demonstrated lower accuracy, possibly due to more diverse purchasing behaviors or fewer transactions compared to "Entertainment" or "K-POP."

4. Electronics (16.8%) and Fashion (15.1%):

These categories showed the lowest accuracies, potentially due to sparse data or overlapping patterns with other categories, making predictions more challenging.

5. Other Categories (Beauty, Food & Health, Automotive):

These categories were excluded due to insufficient occurrences in the dataset, which made reliable prediction impossible.

Key Insights

Impact of Category Popularity: The accuracy differences stem largely from the frequency of category transactions in the dataset. Categories with more occurrences (e.g., "Entertainment") have higher accuracy, as the model learns more robust patterns from larger data volumes.

Model Limitations: Categories with fewer transactions or more diverse behaviors pose challenges, as the limited data impacts the model's ability to generalize.

Why This Accuracy Differs: Unlike holistic accuracy, which evaluates all predictions collectively, and consequent-specific accuracy, which focuses on rule-based predictions, this metric isolates individual categories. It highlights disparities in model performance due to varying transaction frequencies and customer behaviors.

Supporting Coding

The following Python code demonstrates how the system calculates category-wise prediction accuracy:

```
# Calculate category-wise accuracy
category_accuracies = {}
for category in category_columns:
    total_actual = test_df[category].sum()  # Total
occurrences of the category
    correct_predictions = sum(
        (test_df[category] == 1) & (predictions[category]
== 1)
    )  # Correct predictions
    category_accuracies[category] = correct_predictions /
total_actual if total_actual > 0 else 0

# Display category-wise accuracies
for category, accuracy in category_accuracies.items():
    print(f"{category}: {accuracy:.2%}")
```

This snippet computes the accuracy for each category by comparing actual and predicted values, ensuring a detailed breakdown of the model's performance.

Implications

Tailored Recommendations: High-performing categories like "Entertainment" and "K-POP" can be leveraged for targeted marketing, while efforts to improve predictions for low-accuracy categories can expand opportunities.

Dataset Enhancement: Increasing the volume of transactions for low-accuracy categories (e.g., through promotions or data augmentation) can improve the model's learning capacity and overall accuracy.

Model Refinement: Adding contextual features (e.g., time of purchase or regional preferences) or using hybrid recommendation systems could boost accuracy for diverse categories like "Home & Living."

Why It Matters

Accurate and personalized recommendations build customer trust and loyalty. For example, suggesting "K-POP" albums to customers who frequently buy "K-POP" products aligns with their preferences, increasing the likelihood of repeat purchases. Conversely, low accuracy in categories like "Fashion" signals a need for more nuanced modeling approaches, ensuring that recommendations resonate with customers.

By incorporating these findings, this underscores the importance of refining prediction models to enhance both customer satisfaction and business performance.

Results and Discussion

Key Findings

Category Insights: High-lift category pairs like "Beauty" and "Food & Health" demonstrated significant potential for targeted marketing and bundling strategies, supported by robust confidence and lift values.   High-Lift, High-Count Pairs These pairs have both a significant Lift and a sufficiently high Pair Count, making them highly actionable for marketing and business strategies:K-POP & Entertainment:Entertainment & Home & Living:Pair Count:

Model Performance: The predictive model achieved 79% overall accuracy, with notable variations in category-specific performance, such as 68% accuracy for "Entertainment" and lower results for niche categories like "Electronics."

Customer Behavior: Single-category buyers formed the majority, but multi-category buyers offered higher cross-selling opportunities.

Recommendations for Delivered Korea

Based on the findings of this thesis, the following recommendations are proposed to enhance Delivered Korea’s operations and customer engagement strategies:

Optimize High-Lift Category Pairs:

Focus marketing efforts on frequently co-purchased categories, such as "K-POP" and " Entertainment." These categories can be bundled in promotional campaigns to encourage higher average transaction values.

Introduce dynamic discounts or loyalty rewards for customers purchasing complementary categories to drive cross-selling opportunities.

Enhance Geographic Customization

Develop region-specific marketing campaigns tailored to the preferences of high-demand regions like North America and Europe.

Consider reintroducing country-level granularity for regions where trends may differ significantly between neighboring markets, such as in Asia.

Improve Customer Segmentation

Utilize customer purchase patterns to segment buyers into single-category and multi-category groups, tailoring promotional strategies to their unique preferences.

For single-category buyers, recommend products within their favored category to deepen their engagement, while encouraging multi-category buyers to explore complementary products.

Refine Recommendations for Niche Categories

Increase visibility and appeal for low-frequency categories, such as "Food & Health" and "Automotive," through targeted promotions or partnerships with niche influencers.

Adopt Real-Time Recommendation Systems

Implement real-time recommendation systems that adapt dynamically to customer interactions and transactional data. This approach would allow Delivered Korea to provide more personalized suggestions and improve customer satisfaction.

Leverage Seasonal and Cultural Trends

Incorporate seasonal data and cultural preferences into the recommendation models to align offerings with customer needs during peak seasons or holidays.

For example, promoting "Home & Living" and "Electronics" categories during the holiday season could resonate with North American customers.

Conclusion and Future Work

Conclusion

This study analyzed customer purchase behaviors and product recommendations for Delivered Korea, employing association rule mining through the Apriori algorithm and predictive modeling. These approaches provided actionable insights into sales trends, customer preferences, and the effectiveness of recommendation algorithms. The results highlight the potential for using data-driven strategies to improve customer satisfaction and operational efficiency.

Future Work

To address the limitations identified and build on the current findings, future efforts should consider:

Hybrid Recommendation Systems

Integrate collaborative filtering and machine learning models with association rule mining to enhance personalization and adapt to changing customer behaviors.

Experiment with deep learning models like neural collaborative filtering for improved accuracy in multi-category predictions.

Dynamic Rule Updates

Develop a framework for periodically updating association rules to capture shifts in purchase patterns due to seasonal trends or marketing campaigns.

Data Augmentation

Enrich the dataset with external features like customer demographics, time of purchase, and product reviews to improve the robustness of predictive models.

Real-Time Systems

Deploy real-time recommendation systems that dynamically adjust based on customer interactions and transactional data.

Category-Level Enhancements

Focus on underperforming categories by increasing their representation through targeted promotions or synthetic data generation techniques.

Final Remarks

The recommendations presented in this thesis offer Delivered Korea practical steps to enhance its recommendation strategies, operational efficiency, and customer engagement. By leveraging advanced analytics, the company can maintain a competitive edge in the global e-commerce market. Future enhancements, including the adoption of hybrid recommendation systems and real-time personalization, will further position Delivered Korea as a leader in providing tailored shopping experiences.

References

1. Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB)*, 487–499.
2. Hahsler, M., Chelluboina, S., Hornik, K., & Buchta, C. (2007). The arules R-Package Ecosystem: Analyzing Interesting Patterns from Large Transaction Data Sets. *Journal of Machine Learning Research*, 12, 1977–1981.
3. Lin, C. Y., Hu, Y. H., & Tsai, C. F. (2020). A hybrid recommendation approach for improving decision accuracy in e-commerce environments. *Decision Support Systems*, 129, 113199. https://doi.org/10.1016/j.dss.2020.113199
4. Shopify. (n.d.). Shopify data documentation. Retrieved from https://www.shopify.com
5. Zhang, S., Yao, L., Sun, A., & Tay, Y. (2018). Deep learning-based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, 52(1), 1–38. https://doi.org/10.1145/3285029
6. San Francisco State University. (n.d.). Thesis formatting guidelines. Retrieved from https://grad.sfsu.edu

7. Han, J., Pei, J., & Kamber, M. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann. (This textbook provides a comprehensive overview of data mining techniques, including association rule mining and classification evaluation metrics.)
8. Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2(1), 37-63.
9. Liu, B., Hsu, W., & Ma, Y. (1998). Integrating classification and association rule mining. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98) (pp. 80-86). AAAI Press.(This paper introduces associative classification and discusses evaluating prediction accuracy using association rules.)
10. Thabtah, F. (2007). A review of associative classification mining. The Knowledge Engineering Review, 22(1), 37-65.(This review paper covers various associative classification methods and evaluation metrics, emphasizing accuracy assessment.)
11. Tan, P.-N., Steinbach, M., & Kumar, V. (2005). Introduction to Data Mining. Addison-Wesley.(Summary: This textbook provides an introduction to data mining concepts, including classification, association analysis, and evaluation methods like per-class accuracy.Reference to Per-Class Evaluation:"In multiclass classification, per-class accuracy provides detailed insights by computing the accuracy for each class separately." (Tan et al., 2005, Chapter 4)

Appendices

Appendix A

Detailed Code