Marketing Analysis Project

This guide explains gives an overview of and instructions for how to run the code for my Data Mining project that analyzes online retail data.

Link to report: https://drive.google.com/file/d/1YYeTljqAXa6NXFZ4zyB7iuecDFh4_0ei/view?usp=share_link

Link to video presentation: https://youtu.be/29LF0Ggw4vM

Project Description

This project explores the application of data mining techniques to extract actionable business insights from retail transaction data. The project addresses the common challenge retailers face when dealing with massive amounts of transaction data but lacking the analytical tools to transform this information into strategic business decisions. Using the Online Retail II dataset from the UCI Machine Learning Repository, which contains over one million transactions from a UK-based online giftware retailer spanning December 2009 to September 2011, I implemented three complementary analytical approaches to understand customer behavior, product relationships, and temporal sales patterns.

The project employs market basket analysis to identify products frequently purchased together, K-Means clustering with RFM analysis for customer segmentation, and time series analysis to uncover seasonal trends. Through comprehensive data preprocessing, association rule mining, and machine learning techniques, the project demonstrates how retailers can leverage existing transaction data to optimize cross-selling strategies, develop targeted marketing campaigns, and improve inventory management decisions. The analysis provides concrete, quantifiable insights that directly translate into business applications for increasing revenue, improving customer satisfaction, and optimizing operational efficiency.

Research Questions and Answers

Question 1: Are there specific products that frequently appear together in the same invoice?

Answer: Yes, market basket analysis identified 76 strong association rules with an average lift of 22.18 and confidence of 62%. The strongest associations were found between complementary products:

- Poppy's Playhouse bedroom and kitchen items (lift: 52.7, confidence: 79.2%)
- Matching Christmas decorations like wooden star and heart Scandinavian items (lift: 42.6)
- Related gardening products like "Keep Calm" and "Cup of Tea" kneeling pads (lift: 38.9)

These findings reveal clear cross-selling opportunities and suggest that customers purchase thematically related items together.

Question 2: Can I identify distinct groups of customers based on purchase frequency, recency, and monetary value?

Answer: Yes, RFM clustering analysis successfully identified four distinct customer segments:

- **High-Value Active Customers** (3,610 customers): Recent purchases (52 days), high frequency (197 purchases), high monetary value ($4,107 average)

- **Semi-Active Medium-Value Customers** (1,513 customers): Moderate recency (347 days), medium frequency (60 purchases), medium value ($996)
- **At-Risk Low-Value Customers** (819 customers): Long recency (601 days), low frequency (27 purchases), low value ($384)
- **Anomalous Cluster** (1 customer): Likely data artifact requiring refined preprocessing

This segmentation enables targeted marketing strategies tailored to each customer group's behavior patterns.

Question 3: Are there geographical patterns in customer segments?

Answer: This question was not fully addressed in the current analysis due to limited geographical data in the dataset. The Online Retail II dataset contains primarily UK-based customers with limited geographical granularity, preventing meaningful spatial analysis of customer segments. Future analysis would require more detailed location data to explore geographical patterns effectively.

Question 4: Are there seasonal patterns in sales?

Answer: Yes, time series analysis revealed moderate seasonal patterns in sales data. Linear regression feature importance rankings showed:

- **Year** (25.99% importance): Most significant factor affecting sales
- **Month** (12.03% importance): Strong seasonal influence on purchasing patterns
- **Day of Week** (8.73% importance): Moderate weekly patterns
- **Day of Month** (1.28% importance): Minimal daily variation within months

The analysis confirmed seasonal trends exist but are less dramatic than initially expected, with monthly and yearly patterns being the primary drivers of sales variation rather than daily fluctuations.

System Overview

The project consists of five main components:

1. Data Preparation (`data/excel2csv.py`)
2. Data Cleaning (`data/data_cleaner.py`)
3. Market Basket Analysis (`market_basket/market_basket_analysis.py`)
4. Customer Segmentation (`clustering/cluster_analysis.py`)
5. Time Series Analysis (`time_series/time_series_analysis.py`)

Dependencies

Install the required dependencies with pip:

```
pip install pandas numpy scikit-learn mlxtend seaborn matplotlib
```

Running the Analysis

1. Data Preparation

First, convert the Excel data to CSV and set up Git LFS:

```
python data/excel2csv.py
```

This script:

- Combines multiple Excel sheets into one CSV file
- Configures Git LFS for handling large files
- Creates `online_retail_II_combined.csv`

2. Data Cleaning

Clean the combined dataset:

```
python data/data_cleaner.py
```

This script:

- Normalizes prices and quantities
- Handles missing Customer IDs
- Parses dates into programming-friendly format
- Creates `online_retail_II_cleaned.csv`

3. Market Basket Analysis

Run the market basket analysis to discover product associations:

```
python market_basket/market_basket_analysis.py
```

4. Customer Segmentation

Perform RFM analysis and customer clustering:

```
python clustering/cluster_analysis.py
```

5. Time Series Analysis

Analyze sales trends and seasonality:

```
python time_series/time_series_analysis.py
```

Project Structure

```
marketing-analysis-project/
├── data/
│   ├── excel2csv.py
│   ├── data_cleaner.py
│   ├── online_retail_II.xlsx
│   ├── online_retail_II_combined.csv
│   └── online_retail_II_cleaned.csv
├── market_basket/
│   └── market_basket_analysis.py
├── clustering/
│   └── cluster_analysis.py
├── time_series/
│   └── time_series_analysis.py
└── README.md
```

Results

The analysis produces:

- Clean, normalized dataset with proper date formatting
- Product association rules with support, confidence, and lift metrics
- Customer segments based on RFM (Recency, Frequency, Monetary) analysis
- Sales forecasting with RMSE and MAPE accuracy metrics
- Seasonal pattern analysis