

Data Mining Project Proposal

Yaw Sakyi

College of Computing and Software Engineering
Kennesaw State University
Marietta, USA
ysakyi@students.kennesaw.edu

Abstract—This document is a proposal for a data mining research project. The document covers the data set used in the project, the key discovery prospects which would be ascertained at the end of the project, and the techniques used to arrive at the said conclusions.

Index Terms—Data Mining, Knowledge Discovery in Databases.

I. INTRODUCTION

Data mining is the use of machine learning and statistical analysis to uncover patterns and other valuable information from large data sets[1]. As such, this project seeks to apply data mining techniques to discover meaningful patterns in a real world dataset of an online retailer in the United Kingdom. The project will focus on pattern discovery within a dataset, interpreting the discoveries, and performing critical evaluations of the discoveries made to identify whether these discoveries have substantial influence in a real world context. This project will use the Knowledge Discovery in Databases process, which involves selecting a dataset, initial processing of the data by cleaning and structuring, and transforming the data into a usable form which allows us to apply certain algorithms to perform informative evaluations and interpretation. This process ensures that any data analysis performed over the course of the project is systematic and principled.

II. DATASET DESCRIPTION

The dataset selected for this project is the Online Retail dataset on the UCI Machine learning repository. The dataset is a large real world transactional dataset of a UK based retailer. It contains every recorded line item sale for the retailer's online retail business over approximately one year, i.e; from December 1, 2010 to December 9, 2011. The online retailer primarily deals in unique, all-occasion giftware, and has a customer base which mainly consists of wholesalers as opposed to individual customers. This dataset is quite valuable for the Knowledge Discovery in Databases (KDD) process as it captures actual shopper behaviour. The dataset contains information on when an item was purchased, how much of said item was purchased, who purchased said item, and where the buyer is located. With this rich slew of information, the dataset is suitable for the findings the project seeks to undertake.

A. Data Scale and Structure.

- Total Instances: 541,910 transaction records
- Attribute Types:

Variables Table					
Variable Name	Role	Type	Description	Units	Missing Values
InvoiceNo	ID	Categorical	a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.		no
StockCode	ID	Categorical	a 5-digit integral number uniquely assigned to each distinct product		no
Description	Feature	Categorical	product name		no
Quantity	Feature	Integer	the quantities of each product [item] per transaction		no
InvoiceDate	Feature	Date	the day and time when each transaction was generated		no
UnitPrice	Feature	Continuous	product price per unit	sterling	no
CustomerID	Feature	Categorical	a 5-digit integral number uniquely assigned to each customer		no
Country	Feature	Categorical	the name of the country where each customer resides		no

Fig. 1. Figure showing the attributes present in the dataset.

- Categorical: These include the InvoiceNo, StockCode, Description, CustomerID, and Country.
- Integer: Quantity.
- Continuous/Floating Point: UnitPrice.
- Date/Time: InvoiceDate.
- Data Characteristics: Multivariate, Sequential/time-stamped,Integers.

B. Attribute Descriptions

Details of the key features of the dataset and their meaning:

- InvoiceNo: this is a unique identifier for each transaction. In the case where the code begins with the letter "c", that transaction is identified to be a cancellation or return transaction.
- StockCode: A unique alphanumeric code which identifies each product.
- Description: A text description of the product associated with the transaction.
- Quantity: Displays the number of units of a product purchased in a particular transaction.
- InvoiceDate: Shows a timestamp of when a transaction occurred.
- UnitPrice: This variable stores the price of each unit sold in the British Pound Sterling currency.
- CustomerID: A unique number which identifies a customer.
- Country: The country of residence of the customer associated with the transaction.

III. DISCOVERY QUESTIONS

The primary objective of this project is to unearth and interpret meaningful patterns with regards to customer purchasing behavior within the store's online retail dataset. Instead of predicting future outcomes, this project centers on answering the following questions to uncover relationships, and interesting occurrences present in the dataset.

A. Discovery Question 1: Product Association Patterns

- What products are frequently purchased together, and what product combinations happen between customer transactions?

This question seeks to uncover frequent item sets and association rules which reveal how certain products co-occur within customer invoices. As transactions are analyzed at the invoice level, we can identify complementary items and also common items which are purchased within a particular season.

B. Discovery Question 2: Customer Segmentation and Behavior

- Are there customer groups based on purchasing behavior, and what characteristics define these customers?

This question investigates if there are clusters of customers who have similar purchasing habits and if these habits result in the purchase of similar items. By asking this question we can also identify how often certain customers purchase items from the store as well as how much they spend.

C. Discovery Question 3: Anomalous Transactions and Customer Behavior.

- Are there seasonal purchasing behavior with respect to the four weather seasons?

Here we seek to identify through the purchase history of customers if within a particular climate season, customers purchase certain items in large quantities, and also how sales numbers look during the year. for example: during the winter seasons to sales see an uptick or vice versa.

D. Planned Data Mining Techniques

In order to answer the questions posed in the prior section, multiple data mining techniques will be implemented and applied to the dataset. Through the use of a combination of techniques, we can discover comprehensive patterns, validate these patterns, and better interpret the results of any findings at the end of the project. The following techniques emphasize an exploratory analysis of the dataset.

• Association Rule Mining.

- Techniques: Apriori and FP-growth.

The association rule mining techniques above will be applied to the dataset at the invoice level to discover the frequency of certain itemsets and other co-purchasing patterns which may exist between products.

The techniques used in the association rule mining aspect of the project directly answer the first discovery question: What products are frequently purchased together, and what

product combinations happen between customer transactions?. Metrics such as support, lift, and confidence will be used to evaluate the strength and usefulness of the discovered rules. By comparing the results of the Apriori and FP-Growth algorithms, the validity of any pattern discovered can critically tested.

• Clustering.

- Techniques: K-Means and DBSCAN

The clustering techniques above will be applied to the dataset to discover any natural groupings which may exist among customers. This will highlight any similarities between certain customers and also any differences which are present.

The algorithms used in this section provide answers to the second discovery question: Are there customer groups based on purchasing behavior, and what characteristics define these customers? The K-Means algorithm divides customers into a relative number of groups where each individual customer behavior closely relates to a particular mean value or centroid. Since the K-Means algorithm has certain limitations, DBSCAN is algorithm is used to perform clusterization based on arbitrary boundaries of density rather than spherical shapes. By comparing the results of both algorithms on the dataset, any customer clusters can be analyzed with more confidence and certainty.

• Temporal Pattern Analysis

- Techniques: Time-based aggregation, clustering by time frames.

Time based analysis will be performed on the dataset, particularly on the InvoiceDate attribute to explore the relation between specific periods during the year and the types of products that are purchased. By aggregating these dates by months and quarters, an inference can be made on the topic of recurring trends and seasonal shopping habits.

This analysis answers the question: Are there seasonal purchasing behavior with respect to the four weather seasons? The analysis would draw attention to any seasonal peaks of shopping which may exist through out the year and also unearth any pattern that may exist between any product and the season in which it had the most sales. Here the analysis focuses on showing a peak product at a specific time on the year within the boundaries of the dataset and not what would happen in the following year.

• Dimensionality Reduction

- Technique: Principal Component Analysis.

Principal Component Analysis helps to reduce the noise in the dataset thereby revealing the underlying structure in the data.

This helps to interpret the customer clusters which occur after clustering is performed. It further helps to provide a clearer answer to the question: Are there customer groups based on purchasing behavior, and what characteristics define these customers? PCA analysis often singles out characteristics in the dataset which contribute to the groupings found in the dataset.

IV. FLOWCHARTS SHOWING CONCEPTUAL ANALYSIS PIPELINES

A. The High-Level Overview Pipeline

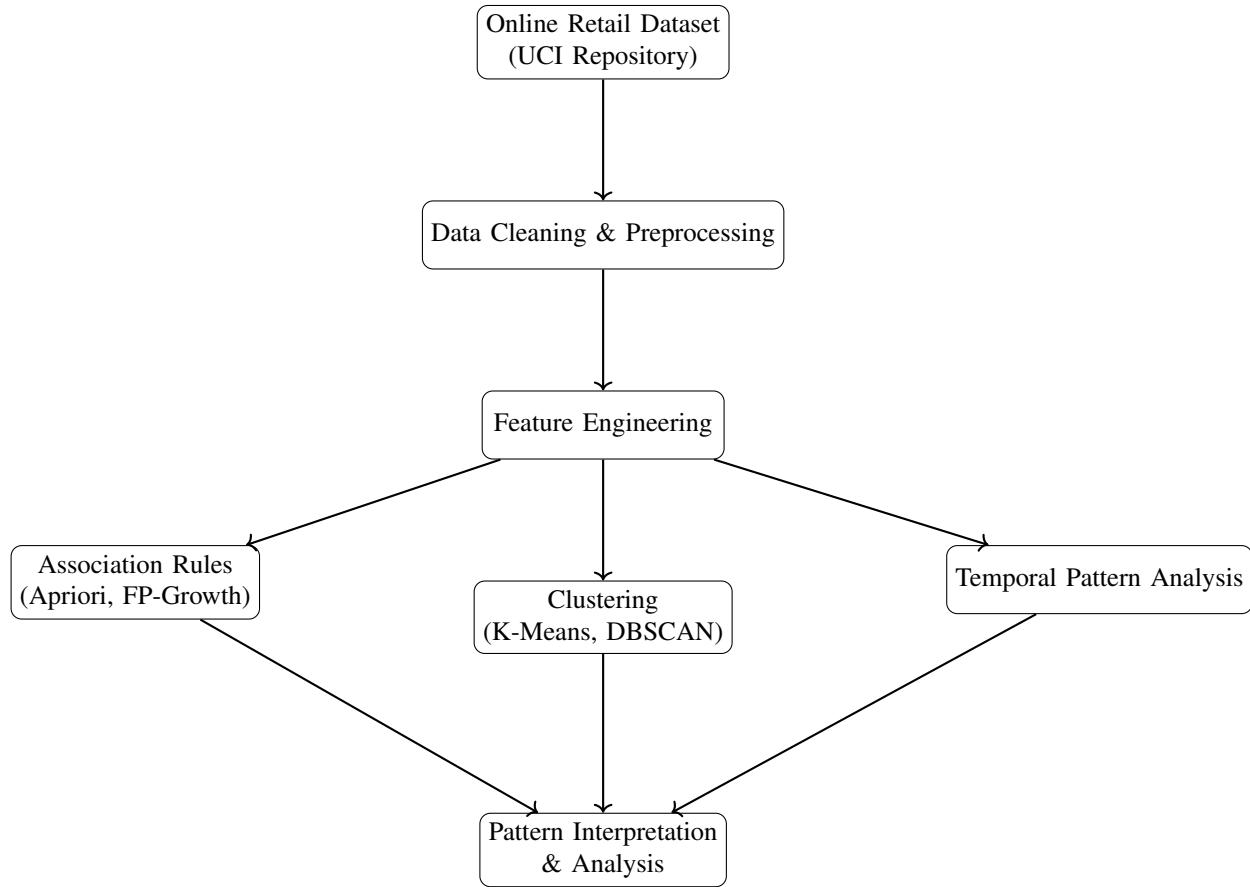


Fig. 2. Planned Data Mining Analysis Pipeline. *The figure illustrate a planned overview of the data mining pipeline for the project.*

B. Association Rule Mining Flowchart

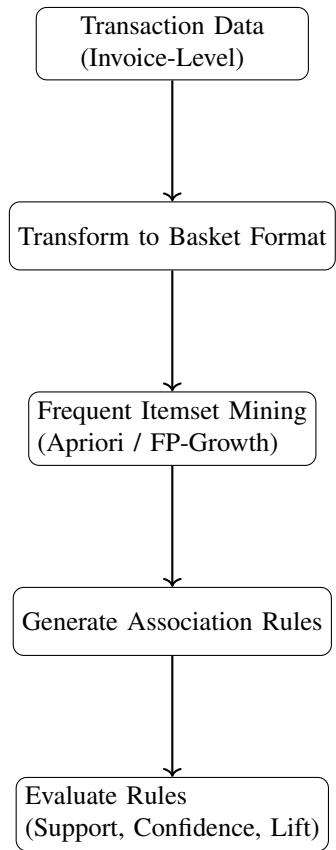


Fig. 3. Association Rule Mining Workflow

C. Clustering Flowchart

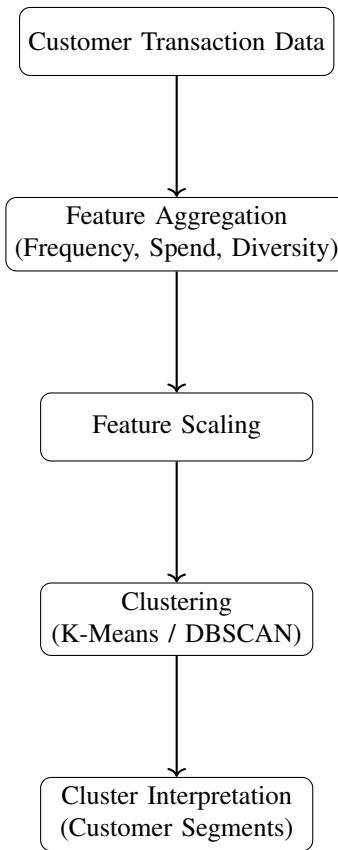


Fig. 4. Customer Clustering Workflow

D. Temporal Pattern Analysis Flowchart

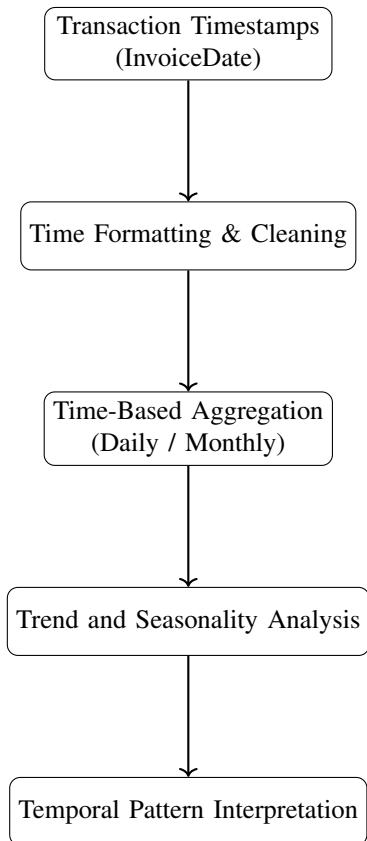


Fig. 5. Temporal Pattern Analysis Workflow

E. Dimensionality Reduction Flowchart (PCA)

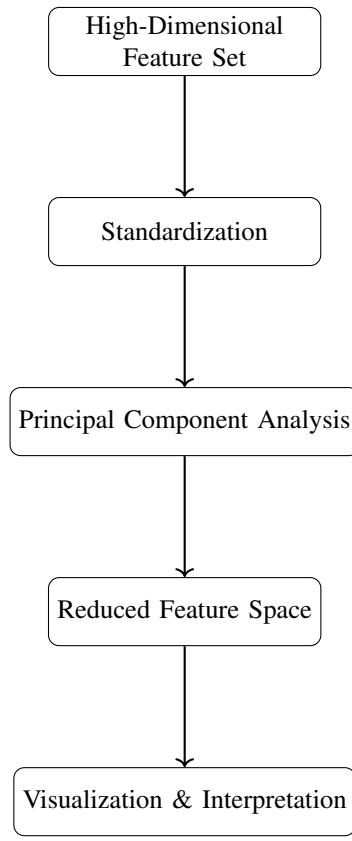


Fig. 6. Dimensionality Reduction Using PCA

V. TENTATIVE SCHEDULE

- **Module Two:** Completed by the 25th of February, 2026. Reserving the final week for minor tweaks and adjustments.
- **Module Three:** Completed within the week of April 1st.
- **Module Four:** Completed by the 23rd of April with the remaining days set for presentation planning.

REFERENCES

- [1] J. Holdsworth. *What is Data Mining?* June 2024. URL: <https://www.ibm.com/think/topics/data-mining> (visited on 02/05/2026).

A.I. was used in the drafting process of this proposal.