Project Proposal

Hernan Munoz
Kennesaw State University
Email: hmunoz2@students.kennesaw.edu

I. Dataset Description

A. Online Retail II

The dataset being used for this research is the **Online Retail II** dataset. This dataset contains all transactions between December 2010 and December 2011 for an online retail company. This company is a giftshop for all times throughout the year.

https://www.kaggle.com/datasets/jillwang87/online-retail-ii

B. Table Details

This dataset is comprised of 525,461 rows and 8 columns. Each row is made of the Invoice Number, StockCode, Description of the product, Quantity of items per transaction, InvoiceDate, Price per Unit, Customer ID, and the Country of the transaction. These will be used to find purchasing patterns

TABLE I
Online Retail II Attributes

Variable	Description
InvoiceNo	Nominal, a 6-digit integral number per transaction.
StockCode	Nominal, a 5-digit integral number per product.
Description	Nominal, product name.
Quantity	Numeric, quantity per transaction.
InvoiceDate	Numeric, date and time of transaction.
UnitPrice	Numeric, product price per unit.
CustomerID	Nominal, unique identifier per customer.
Country	Nominal, name of the customer's country.

C. Data Quality

Early analysis shows 2 main data quality issues that will effect the mining process. First, this dataset shows many missing values in the Customer ID attribute. This could impact any customer-specific research. Second, the Invoice attribute include a ”C” prefix, which explains cancelled transactions. These transactions need to be handled as they result in negative values.

D. Sample Data Table

Invoice	StockCode	Description	Qty	Price	ID
536365	85123A	WHITE HEART LIGHT	6	2.55	17850
536367	84879	BIRD ORNAMENT	32	1.69	13047
536370	22728	ALARM CLOCK BAKERY	24	3.75	12583

TABLE II
Sample of the Online Retail II transaction data.

II. Discovery Questions

In order to ensure a successful research into this dataset, these discovery questions will be used to find relationships in the data and not predict any specific outcome.

A. Item Assocations

”Which products are most often bought together in a single transaction?” When looking at retail transactions, item association is an important relationship to find. Retailers are constantly finding which items pair best together, so they can improve sales. This is interesting to research as it shows what customers will naturally pair together, which could later be used to improve store layouts and promotions.

B. Regional Product Preferences

”How does a customer's country impact their product preferences?” This company sells to multiple countries. This is valuable as businesses consider the geographic location very important for business. Focusing on this question can lead to interesting discoveries, such as differences between some countries, while on the other hand finding similarities one might not have expected.

III. Planned Techniques

In order to properly investigate the discovery questions, 2 different data mining techniques will be used. Association Rules and Clustering will be the 2 techniques used.

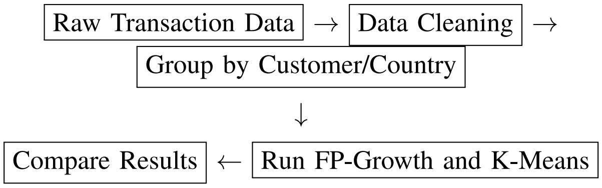
A. Assocation Rules

Since the first discovery question focuses on Item associations, it is best to use Association Rules. I plan to use the FP_Growth Algorithm because this dataset is very large with 500k+ rows. This algorithm will be much more efficient. I will be looking for patterns to see which items have a strong relationship. This will show natural associations between items, and show which products are bought together most often.

B. Clustering

For the second question, Region Product Preferences, the Clustering technique will be used. This technique relates heavily to this question, as it involves grouping/clustering customer product preferences. This question will require grouping transaction data by customer and country to find data like Average Quantity. I will use the ”Elbow Method” to find the best number of cluster for this dataset. Similarly, I will be using K-Means to see if customers from different countries naturally fall into different groups or have similar buying patterns.

C. Project Analysis Pipeline



IV. Project Schedule and Challenges

To make sure that this project is completed in time with each milestone deadline, this timeline will be used to ensure a steady workload.

A. Preliminary Timeline

The following milestones outline the phase-by-phase approach to the discovery process:

- **Milestone 1: Project Proposal (Due Feb 5)**
  This initial phase outlines the plans for this project. This includes making the Proposal pdf and initialize the GitHub Repository. This step is vital to ensuring a managable project timeline. This phase determines all discovery questions and data mining techniques that will be used in this project. A 500k+ row dataset was chosen as it will benefit the project far more than a smaller dataset would. A dataset of this scale makes relationships far more noticeable and clear. This accomplishes the primary objective of finding deeper connections rather than basic statistics.
- **Milestone 2: Data Preprocessing and EDA (Due Mar 5)**
  This milestone is by far one of the most important of the process. As mentioned before, this dataset has many null values. This will require cleaning, one of the main parts of this milestone. This milestone involves working on data management and data quality. Once managing any null values and filtering out the ”C” prefixes, the next step of this process will be the Exploratory Data Analysis. This will be done using Pandas. This is what makes this milestone by far the most important for research. This proposal will ensure a steady timeline and plan. By having this structured plan, this will help make sure this data is analyzed accurately.
- **Milestone 3: Model Implementation (Due Apr 2)**
  This milestone will be the implementation phase. This phase will transition from data preparation to active mining. The first step of this phase will be focusing on the first discovery question, Item Associations. By utilizing the FP-Growth algorithm, any product pairings will be found. From here, it is important to find which are statistically significant and which are coincidental. From here, the second discovery question will be the next focus. This will require transforming the data into a customer-focused format. Here we will implement the K-Means algorithm and use the Elbow method to find the optimal number of clusters. This milestone will be important to later evaluating the previously-made discovery questions.
- **Milestone 4: Evaluation and Final Report (Due Apr 30)**
  The final month of the semester will be dedicated to this milestone. Milestone 4 involves the evaluation and interpretation from Milestone 3. The goal of this Milestone will be to determine if the association rules and clusters give credible business insights. This phase will determine the conclusions made in the final report and presentation. The presentation will visualize and explain the most interesting discoveries and explain how these discoveries impact the discovery questions made in Milestone 1. By ensuring an accurate understanding of the data researched in this project, this Milestone will provide credible findings to be used later in the presentation slides.

B. Anticipated Technical Challenges

There are three primary challenges expected to be encountered during this project.

1) Computational Complexity and Scalability: With over 500,000 records, the first major hurdle is the size of this dataset. Using the FP-Growth algorithm works much better with large datasets. The Apriori method would be much less efficient with a dataset of this size. While FP-Growth is better, it could still be slow. This comes down to the efficiency of the hardware being used in this project. A personal laptop is being used so there is a possibility the laptop's specifications and age could slow down testing.

2) Data Transformation and Similarity Metrics: In order to effectively use the clustering algorithms chosen in Milestone 1, the data has to be transformed before doing any work on it. This requires normalization and aggregation. Data is currently a collection of transactions, which means it needs to be made into a collection of customers in order to cluster customers. This is necessary to answer the second discovery question. This will require alot of transformation to accomplish.

3) Missing Values and Data Integrity: The second discovery question requires a collection of customer that will be clustered. One main issue that will be encountered is the null values in the Customer ID column. This column is null for almost 25% of the dataset. This will have to be handled before moving forward. This leaves a very difficult decision to be made. K-Means requires a unique ID, so these missing values need to be either deleted or grouped. Deleting these values could remove very important data and give an inaccurate result. On the other hand, grouping these values would make one artificial cluster that wouldnt be grouped together otherwise. This artificial cluster can very drastically skew data since 25% of the dataset is missing the Customer ID.

References

[1] J. Wang, ”Online Retail II Dataset,” Kaggle. [Online]. Available: https://www.kaggle.com/datasets/jillwang87/online-retail-ii.

[2] J. Han, M. Kamber, and J. Pei, ”Data Mining: Concepts and Techniques,” 3rd ed. Morgan Kaufmann, 2011.