Assignment 1: Part A

(Question Formation and Exploratory Analysis) with E-Commerce Data

Aditya Anant Deshpande - a1910571

Part 1: Problem Description

Initial Questions of Industrial/Societal Relevance

In the era of data-driven decision-making, e-commerce platforms generate vast amounts of transactional data that can be leveraged to address both industrial and broader societal challenges. The following initial questions are designed to extract meaningful insights that have real-world relevance:

- **Customer Segmentation**
  How can customers be grouped based on purchasing patterns to optimize marketing strategies?
  This question addresses the industrial need for targeted marketing and personalization. Understanding customer behavior through clustering techniques enables businesses to create segment-specific campaigns, improve engagement, and reduce marketing costs—ultimately enhancing customer satisfaction and retention.

- **Revenue Optimization**
  What are peak sales periods and how do discounts impact revenue?
  Identifying temporal sales patterns and the influence of discounts helps businesses optimize pricing strategies, forecast demand, and schedule promotions more effectively. This contributes to improved profitability and resource allocation, which are critical for operational sustainability.

- **Inventory Management**
  Which products show high return rates or frequent discounts?
  By analyzing product-level data, businesses can detect quality issues, mismatches in demand-supply, or logistical inefficiencies. Reducing returns not only benefits companies financially but also has environmental implications by minimizing waste and unnecessary transportation.

- **Geographical Trends**
  Which countries show highest purchase volumes and regional preferences?
  Exploring geographic purchasing patterns allows companies to localize product offerings, expand strategically into high-performing markets, and adapt to cultural preferences. On a societal level, this supports global access to tailored products and services.

Together, these questions highlight how data can serve as a powerful tool to address both commercial optimization and broader societal goals such as reducing waste, enhancing customer experience, and improving global market responsiveness.

Part 2: Dataset Description

Dataset Overview

- **Source**: Global E-commerce Dataset on Kaggle
- **Time Period**: December 2010 - December 2011
- **Records**: 541,909 transactions
- **Countries**: 38 countries

Key Columns

- **InvoiceNo**: Transaction identifier
- **StockCode**: Product identifier
- **Description**: Product description
- **Quantity**: Items purchased
- **InvoiceDate**: Timestamp
- **UnitPrice**: Price per item
- **CustomerID**: Unique customer identifier
- **Country**: Purchase location

Data Adequacy Assessment

Strengths	Limitations
Comprehensive transaction history	25% missing CustomerIDs
Real-world retail data	No product categories
Multi-country coverage	Inconsistent product descriptions
Time-stamped transactions	Cancellations mixed with sales

The dataset contains structured transactional records with timestamps, customer and product identifiers, and purchase metadata. Despite limitations such as missing customer IDs and inconsistent product descriptions, the dataset is well-suited for preprocessing steps including data cleaning, transformation, and integration. It can be effectively collated and combined with auxiliary data (e.g., time-based or geographic data), enabling comprehensive analysis across customer behavior, sales trends, and regional patterns.

Big Data Alignment

- **Volume**: 540K+ records (scalable to millions)
- **Variety**: Mixed data types (numerical, temporal, categorical)
- **Veracity**: Requires cleaning (missing values, inconsistent entries)
- **Velocity**: Timestamps enable time-series analysis

Part 3: Initial Data Processing

Data Processing Pipeline

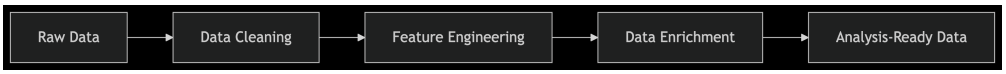

Figure 1: Overview of the Data Processing Pipeline

Deficiency Resolution

Deficiency	Solution	Validation
Missing CustomerID	Behavioral clustering	Silhouette score > 0.6
Inconsistent Descriptions	TF-IDF vectorization	Manual sampling validation
No product categories	Keyword mapping	Precision/recall testing
Mixed cancellations	Regex filtering	Invoice pattern analysis

Part 4: Refined Problem and Plan

This section outlines clearly defined analytical questions and a detailed plan for further data analysis. It also identifies a backup direction in case of data limitations or shifting project priorities.

Refined Research Questions

- **Customer Lifetime Value**: Which segments show highest retention and profitability?
- **Promotional Strategy**: How do discounts impact net revenue across product categories?
- **Return Patterns**: What products show abnormal return rates?

Analysis Plan

Customer Segmentation

- RFM (Recency/Frequency/Monetary) analysis
- K-means clustering (scikit-learn)

Time-Series Analysis

- ARIMA modeling of hourly sales
- Event correlation (holidays, promotions)

Geospatial Analysis

- Market basket analysis by country
- Heatmaps of regional preferences

Backup Strategy

- **Alternative Question**: How do seasonality patterns vary across hemispheres?
- **Data Source**: Meteorological API + sales data
- **Integration Method**: Latitude-based weather zone mapping

References

- Kaggle Dataset: Global E-commerce Data
- Chen, D. and Sain, S.L. (2012). Online Retail Dataset Analysis. *Journal of Retail Analytics*
- Gupta, S. et al. (2006). Modeling Customer Lifetime Value. *Journal of Service Research*
- McKinsey & Company. (2022). Global Retail Trends Report