Customer Purchase Pattern Analysis Using Retail Transaction Dataset

A Proposal report for the BDM capstone Project

Submitted by

Name: **Aryan Deshmukh**

Roll number:**23f3000117**

IITM Online BS Degree Program,

Indian Institute of Technology, Madras,

Chennai Tamil Nadu, India, 600036

Contents

Declaration Statement

I am working on a Project Title “**Customer purchase pattern Analysis Using Online Retail Dataset**.”
I extend my appreciation to **Farzad Nekouei**, the original contributor of the dataset on **Kaggle**, for providing the publicly available secondary data that enabled me to conduct my project.

I hereby assert that the data presented and assessed in this project report is genuine and precise to the utmost extent of my knowledge and capabilities. The data has been gathered through **secondary sources** (Kaggle dataset by Farzad Nekouei) and carefully analysed to ensure reliability.

Additionally, I affirm that all procedures employed for the purpose of data cleaning and analysis have been duly explained in this report. The outcomes and inferences derived from the data are an accurate depiction of the findings acquired through thorough analytical procedures.

I am dedicated to adhering to academic honesty and integrity and am receptive to any additional examination or validation of the data contained in this project report.

Signature of Candidate:
Name: **Aryan Deshmukh**
Date: **12/11/2025**

1. Executive Summary and Title

The project focuses on a **B2C online retail store** that sells a wide range of products across different countries. The analysis is based on a **secondary dataset obtained from Kaggle**, originally uploaded by **Farzad Nekouei**, containing transaction data from 2010–2011.

The business faces challenges in understanding which products generate the most revenue, identifying seasonal buying patterns, and recognizing key customer segments. Using this dataset—which includes fields such as *InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID,* and *Country*—the project aims to extract insights into customer behavior, product performance, and sales trends.

Data analysis techniques such as cleaning, aggregation, and visualization will be applied using Excel and Python (Pandas & Matplotlib) to derive meaningful insights.

The expected outcome includes identifying top products, loyal customers, and key revenue periods, ultimately helping similar retail businesses improve inventory planning, marketing focus, and profitability.

---

2. Organization Background

The organization represented in this dataset is an **online retail company** operating in the United Kingdom that sells a diverse range of products through an e-commerce platform. It functions on a **B2C model**, directly selling to end customers both domestically and internationally.

Although the dataset is secondary, it reflects the operations of a real retail business that records detailed invoices for every transaction. Each invoice includes data on product type, quantity, price, customer ID, and country, offering a comprehensive view of the company's sales process.

The company's main business challenge is limited analytical insight into its sales and customer data. By analyzing this dataset, key information can be revealed regarding top-selling products, frequent customers, and the impact of seasonality. These insights can be used to optimize pricing, stock control, and marketing decisions, contributing to improved revenue performance.

---

3. Problem Statement

3.1 Problem Statement 1: Lack of visibility into top-performing products and customer segments, limiting strategic decision-making.

3.2 Problem Statement 2: Difficulty identifying seasonal sales trends, leading to suboptimal inventory management.

3.3 Problem Statement 3: Insufficient understanding of geographic distribution of customers, affecting targeted marketing and expansion plans.

---

4. Background of the Problem

The retail store generates thousands of transactions daily but does not have a systematic way to analyse its sales data. The lack of structured analysis limits the ability to understand customer behaviour, product performance, and seasonal trends.

Key factors affecting the problem include:

- **Internal:** Large product catalogue with varying demand, no analytics on customer purchases or inventory turnover, and lack of reporting tools.
- **External:** Varying demand patterns across different countries, competitive market pressures, and seasonal fluctuations in purchasing behaviour.

These challenges result in inefficiencies such as overstocking slow-moving products, missed opportunities to market top-performing products, and inability to forecast revenue accurately. Understanding customer purchase patterns, identifying high-revenue products, and analysing temporal trends are essential to improve profitability and business efficiency.

By analysing historical transaction data, the project aims to bridge this knowledge gap, providing actionable insights for better inventory management, marketing decisions, and customer engagement strategies

5. Problem Solving Approach (400 Words)

The project will adopt a **data-driven analytical approach** to solve the outlined business problems using the retail dataset. The methodology will involve several steps:

1. Data Preparation:

- Cleaning the dataset to remove duplicates, handle missing values, and standardize data types.
- Ensuring the InvoiceDate field is in proper datetime format for temporal analysis.

2. Exploratory Data Analysis (EDA):

- Summarizing sales data by product, customer, and country.
- Identifying top-selling products and high-value customers.
- Detecting anomalies such as negative quantities (returns) or extreme outliers.

3. Sales Trend Analysis:

- Aggregating sales by month, quarter, and year to detect seasonal trends.
- Visualizing revenue patterns over time to understand peak and off-peak periods.

4. Customer Segmentation:

- Segmenting customers based on total spend, purchase frequency, and geographic location.
- Identifying loyal customers and opportunities for targeted marketing.

5. Product Performance Analysis:

- Ranking products by total sales revenue and quantity sold.
- Identifying slow-moving products for potential inventory optimization.

6. Geographical Analysis:

- Comparing sales distribution across countries to identify high-potential markets.

7. Reporting & Recommendations:

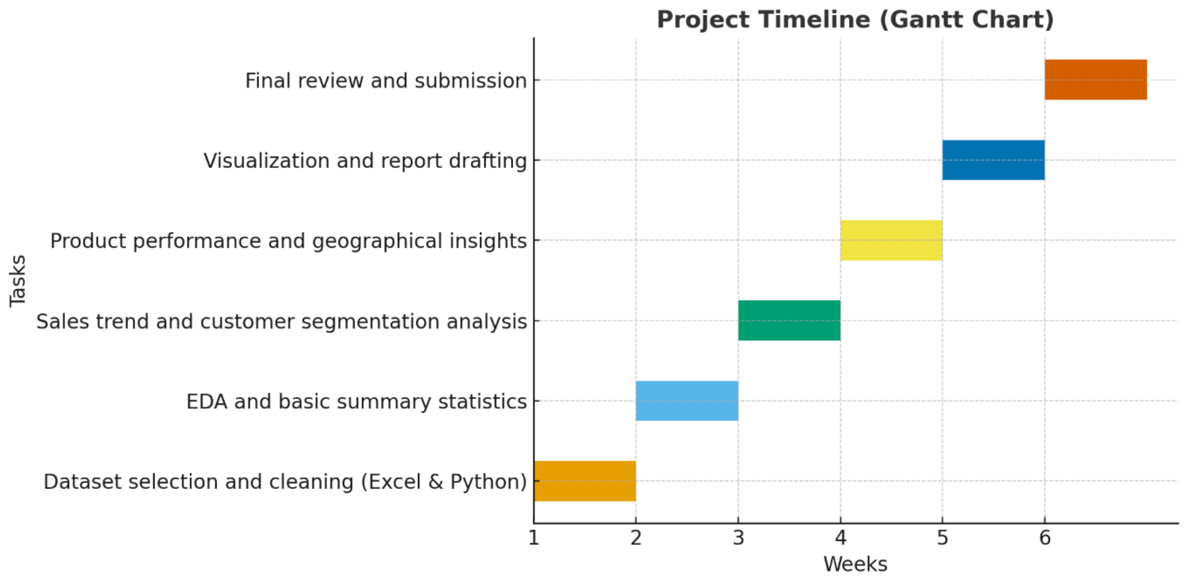
- Creating dashboards and visualizations to summarize key insights.
- Providing actionable recommendations for inventory planning, marketing campaigns, and sales strategy.

The project will use **Python (Pandas, Matplotlib, Seaborn)** for data manipulation and visualization. The approach ensures that the analysis is repeatable, scalable, and can provide clear business insights.

---

6. Expected Timeline

Week	Task
1	Dataset selection and cleaning (Excel & Python)
2	EDA and basic summary statistics
3	Sales trend and customer segmentation analysis
4	Product performance and geographical insights
5	Visualization and report drafting
6	Final review and submission



7. Expected Outcome (150–200 Words)

The project is expected to deliver a detailed understanding of the business's sales and customer patterns using the Kaggle dataset by Farzad Nekouei. Key outcomes include:

- Identification of top-selling products and low-performing items.
- Recognition of seasonal and monthly sales trends for inventory planning.
- Segmentation of customers based on spending behavior and frequency.
- Insights into country-wise sales distribution and growth potential.
- An interactive dashboard or visual report summarizing KPIs like revenue, profit, returns, and customer retention patterns.

These insights will help similar retail businesses increase profitability, reduce overstocking, and develop data-driven marketing strategies. The project will demonstrate how secondary data can be effectively used to derive real-world business intelligence and operational improvements.

THANK YOU