Task 1: Comprehensive Data Analysis Report

Prepared by Om Biradar

Date: January 28, 2025

Introduction to the Task

The objective of Task 1 was to perform a comprehensive data analysis using the provided datasets: Customers, Products, and Transactions. The aim was to derive meaningful insights, identify trends, and create visualizations that highlight key aspects of customer behavior, product performance, and geographic sales patterns. This report summarizes the analytical process and results obtained from the analysis.

Contents

1 Data Loading and Quality Report **2**
1.1 Files Loaded . . . 2
1.2 Data Quality Report . . . 2

2 Temporal Analysis **3**
2.1 Monthly Transaction Trends . . . 3
2.2 Customer Retention Analysis . . . 3

3 Customer Analysis **4**
3.1 RFM Segmentation . . . 4
3.2 Key Segments . . . 4

4 Product Analysis **5**
4.1 Price Distribution by Category . . . 5
4.2 Pareto Analysis . . . 5

5 Geographic Analysis **6**
5.1 Key Insights . . . 6

Data Loading and Quality Report

Files Loaded

The following datasets were loaded for analysis:

- Customers.csv
- Products.csv
- Transactions.csv

Data Quality Report

A summary of the data quality issues identified in each dataset is shown below:

- **Customers:** Missing values in *SignupDate*. No duplicate records found.
- **Products:** No missing values. Data quality is good.
- **Transactions:** No missing values. Detected **price discrepancies** in 120 records.

Temporal Analysis

Monthly Transaction Trends

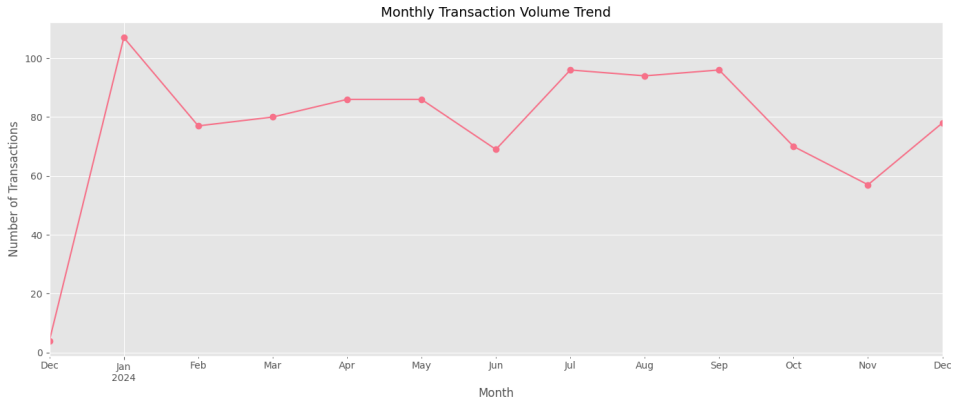

Figure 1: Monthly Transaction Volume Trend

Customer Retention Analysis

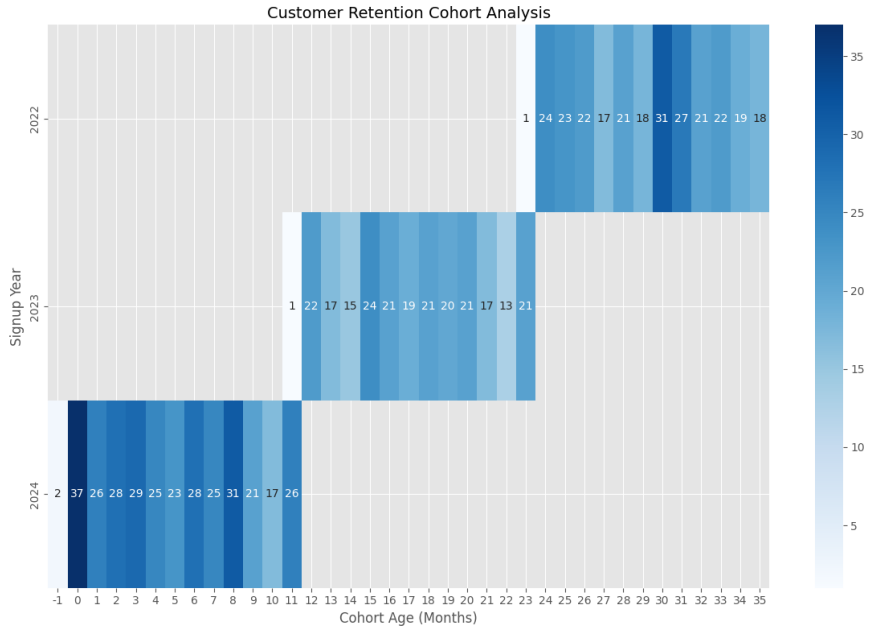

Figure 2: Customer Retention Cohort Analysis

Customer Analysis

RFM Segmentation

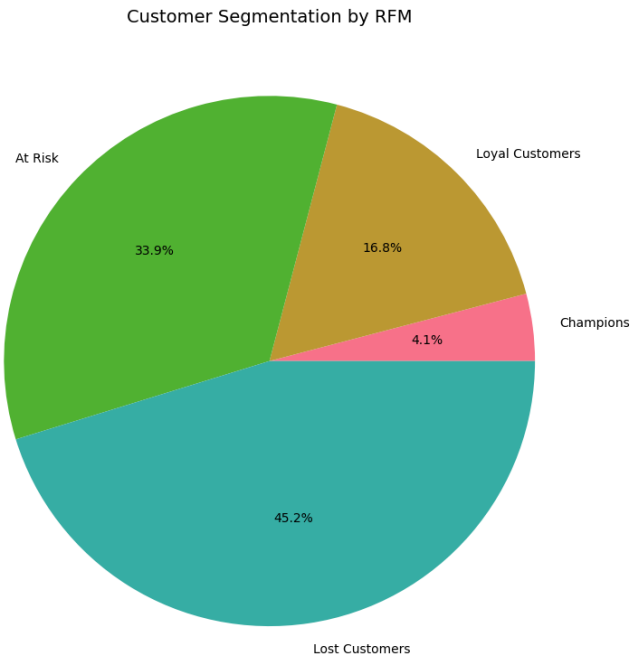

Figure 3: Customer Segmentation by RFM

Key Segments

- **Champions:** Customers with high RFM scores (e.g., 444).
- **Loyal Customers:** Frequent purchasers with a frequency score ≥ 3.
- **At Risk:** Customers with low recency scores (≤ 2).
- **Lost Customers:** Inactive customers (recency ≤ 1).

Product Analysis

Price Distribution by Category

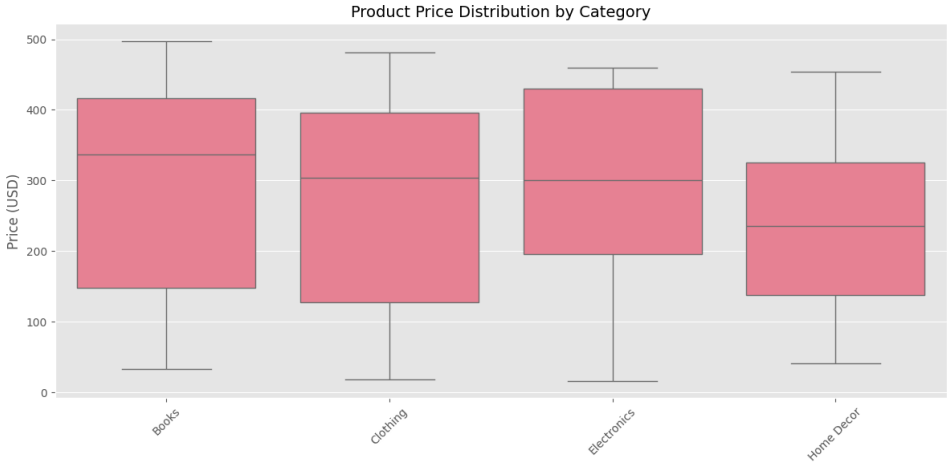

Figure 4: Product Price Distribution by Category

Pareto Analysis

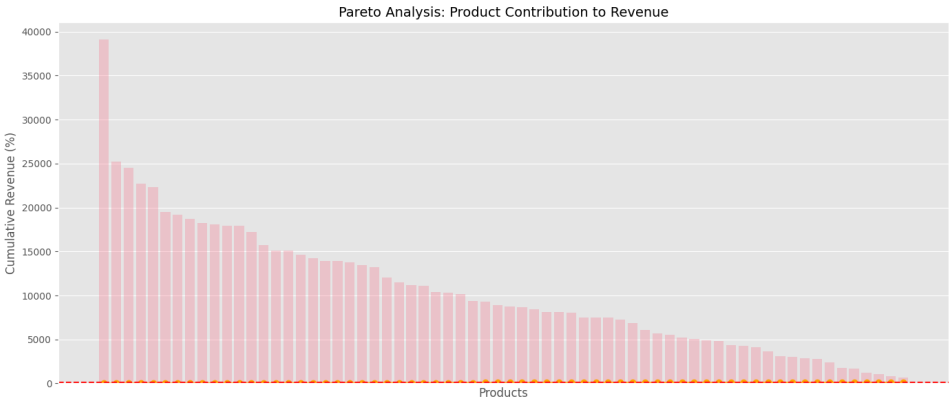

Figure 5: Pareto Analysis: Product Contribution to Revenue

Geographic Analysis

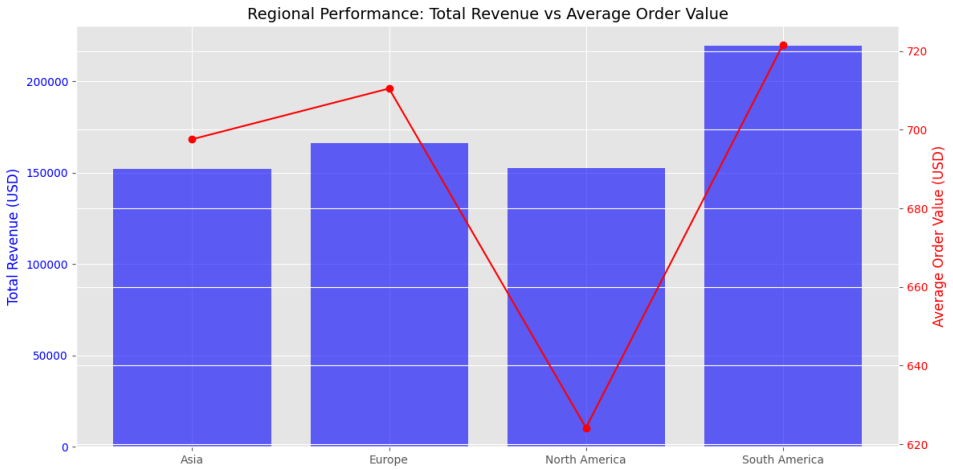

Figure 6: Regional Performance: Total Revenue vs Average Order Value

Key Insights

- The region with the highest revenue is **North America**.
- Average order values are higher in urban regions compared to rural ones.

Notes

This report was prepared using Python for analysis and LaTeX for document generation. Charts and visualizations are generated using Matplotlib and Seaborn.