



End-to-End Retail Performance and Behavioural Analytics

ACKNOWLEDGMENT

I would like to sincerely thank the faculty, mentors, and peers who guided me during the development of this **Retail Analytics Project**. Their valuable feedback helped me understand how structured data analysis can transform raw business data into actionable insights. This project was not just an academic exercise but also a real-world simulation of how retail businesses operate across multiple channels, manage customers, handle store costs, and optimize returns.

Through this journey, I developed strong skills in SQL query building, data cleaning with Python, exploratory data analysis (EDA), and visualization using Power BI. Each stage of this project enhanced my ability to connect data with business decision-making. I am grateful for the opportunity to work on this project, as it gave me exposure to end-to-end data analytics workflows, including data integration, preprocessing, KPI analysis, and dashboard storytelling.

CHAPTER 1

INTRODUCTION

1.1 Project Objective:

The main objective of this project is to analyse retail sales, customer behaviour, store performance, and product returns to generate insights that can help in strategic decision-making. Modern retail businesses face enormous competition, and decisions backed by data are crucial for survival and growth. This project leverages SQL, Python (EDA), and Power BI to highlight important business metrics such as revenue trends, profit margins, customer loyalty, store costs, and product return patterns. The aim is not just to present numbers but to translate raw data into meaningful stories. By focusing on areas like sales performance by channel, store profitability, customer demographics, and return reasons, the project provides a 360-degree view of the retail ecosystem. This approach ensures that stakeholders can make data-driven strategies regarding pricing, marketing, product design, logistics, and customer engagement.

Overview of the Project (Retail Analytics):

This retail analytics project is designed to provide a comprehensive understanding of business performance across multiple dimensions, including sales, customers, stores, and product returns. The purpose is to move beyond raw transaction data and create a structured framework of insights that can support strategic decision-making at all levels of the organization.

The project is divided into several key stages:

1. Data Collection and Integration

The first stage focused on consolidating data from multiple sources — sales transactions, customer profiles, store details, product information, and return records. Together, these datasets capture the entire lifecycle of a retail transaction, from purchase to possible return. This integration created a unified dataset that

could support both granular (customer-level) and aggregate (region or store-level) analysis.

2. Data Cleaning and Preprocessing

Before analysis, the dataset was cleaned to remove duplicates, handle missing values, and standardize fields. SQL foreign keys and indexing ensured that all datasets (sales, customers, products, stores, returns) were linked seamlessly. This step was crucial because inconsistencies in customer IDs, product categories, or return records could otherwise distort results. Preprocessing ensured that the dataset was accurate, reliable, and analysis-ready.

3. SQL-Based Business Querying

SQL formed the backbone of the analysis. Business questions were framed around key metrics such as total revenue, customer segmentation, store profitability, product return rates, and sales channel efficiency. Queries were optimized using indexes and constraints to ensure efficiency. Each query was designed to address a specific business decision point, for example:

- Which stores generate the most profit relative to operating cost?
- Which customer age groups contribute the most revenue?
- Which sales channel yields higher profit per order?
- What is the trend of returns by product category and region?

4. Exploratory Data Analysis (EDA)

Python was used for EDA to complement SQL outputs with visual patterns and trends. EDA helped uncover seasonality in revenue, regional differences in costs, demographic trends among customers, and recurring reasons for product returns.

Visualization with Python libraries such as Pandas and Matplotlib enabled a deeper understanding of the raw data before dashboard creation.

5. Dashboard Development with Power BI

To translate technical outputs into business-friendly insights,

interactive dashboards were developed in Power BI. Four major dashboards were created:

- Sales Dashboard → Revenue, profit, product performance, and channel comparisons.
 - Store Dashboard → Store count, regional cost distribution, store-type mix, and profitability analysis.
 - Customer Dashboard → Age group trends, gender distribution, signup analysis, and top customers by profit.
 - Returns Dashboard → Return rate, regional variation, reasons for return, and channel-level breakdown.
- Each dashboard was designed with filters and drill-down features, allowing users to explore the data according to their needs.

6. Insight Generation and Business Implications

Finally, the results from SQL, EDA, and dashboards were combined into actionable insights. For instance, identifying that the online channel generates higher profit per order helps guide marketing strategy, while recognizing that Apparel has the highest return rate due to defects highlights the need for improved quality checks.

1.2 Problem Statement

Retail businesses often generate huge amounts of transactional and customer data, but most of it remains underutilized. Managers and executives may know their overall revenue but may not fully understand:

- Which sales channels (online vs in-store) generate higher profit per order?
- Which regions or stores are cost-efficient, and which are draining resources?
- Which customers contribute the most to long-term profit and how to retain them?
- Which product categories have the highest return rates and why?

Without such clarity, businesses may invest in unprofitable channels, fail to retain loyal customers, or miss opportunities to control costs. This lack of structured, centralized insight results in poor inventory management, suboptimal marketing campaigns, and reduced profitability.

This project addresses these challenges by developing an integrated analytical system where SQL queries generate KPIs, Python EDA uncovers patterns, and Power BI dashboards visualize insights in an interactive, business-friendly format.

1.3 Project Objectives

The project's specific objectives are:

1. **Sales Analysis** – Measure revenue, profit, and sales trends across time, channels, and products. Provide decision-makers with insights into which products and sales methods contribute the most value.
2. **Store Analysis** – Evaluate operating costs, profitability, and performance by store type and location. Identify cost hotspots (cities/regions) and compare profit vs. operating cost at the store level.
3. **Customer Insights** – Analyse customer demographics such as age, gender, and region. Segment customers by profit contribution, age group revenue, and signup trends. Highlight top-value customers and loyalty patterns.
4. **Return Analysis** – Study return rates over time, across regions, categories, and reasons. Help businesses control avoidable returns by identifying weak areas (e.g., defective products or delivery delays).
5. **Comprehensive KPI Monitoring** – Provide stakeholders with dashboards that combine all aspects of sales, store costs, customers, and returns into one framework for quick, data-driven decision-making.

CHAPTER 2

DATA COLLECTION AND SOURCES

2.1 Data Sources

The dataset used in this project is a consolidated retail dataset that combines information from multiple aspects of a business:

- **Sales Data:** Each row represents an order with details such as order date, product sold, customer ID, store ID, sales channel (online or in-store), quantity purchased, unit price, discount percentage, and the final total amount. This is the backbone of revenue and profitability analysis.
- **Customers Data:** Contains demographic and personal details about customers, such as their region, age, gender, and signup date. It allows segmentation by age groups, gender ratio, loyalty, and regional presence.
- **Products Data:** Provides details like product category (Apparel, Electronics, Personal Care, Home & Kitchen), product ID, and cost price. This data is crucial for analysing profit margins, best-sellers, and return rates by category.
- **Stores Data:** Captures details about store location (region, city), store type (flagship, franchise, or mall kiosk), and operating cost. It helps evaluate regional cost distribution, city-level cost hotspots, and profitability per store.
- **Returns Data:** Tracks products that were returned by customers, including the return ID, return reason (defective, late delivery, no longer needed, wrong item), and associated order ID. This dataset highlights customer dissatisfaction patterns and helps control avoidable costs.

Together, these datasets create a complete ecosystem of retail operations, enabling holistic analysis.

CHAPTER 3

DATA PREPROCESSING

Effective data preprocessing is the foundation of any reliable data analytics project. Raw retail data, although rich in transactional and demographic information, often contains inconsistencies, missing values, duplicate records, and irregular formats that can distort analysis. To ensure that the final insights are both accurate and actionable, the dataset was subjected to a series of systematic preprocessing steps. These steps not only improved the quality of the data but also ensured that it was optimized for SQL querying, exploratory data analysis (EDA), and dashboard visualization in Power BI.

3.1 Removal of Duplicates

One of the common challenges in transactional datasets is the presence of duplicate entries. In the retail dataset, duplicates could occur when:

- An order is recorded more than once due to system errors.
- Customer IDs or product IDs are repeated in multiple rows with identical attributes.
- Store transactions are mistakenly logged twice.

Duplicates can artificially inflate metrics such as total revenue, number of orders, or customer counts, leading to misleading results. To address this, duplicate entries were identified using a combination of attributes such as Order_ID, Customer_ID, Product_ID, Quantity, and Order_Date. SQL's DISTINCT and ROW_NUMBER() functions were used to flag duplicate rows, while Python's Pandas drop_duplicates() method was applied to the dataset exported for EDA. After cleaning, each transaction was uniquely represented, ensuring the integrity of downstream analysis.

3.2 Handling Missing Values

Missing values are another major issue in retail datasets, particularly in fields such as Customer Age, Region, Store Type, and Product

Category. Left untreated, missing values can disrupt calculations such as average revenue per customer, regional profitability comparisons, or product category return rates.

To address this challenge:

- **Customer Information:** If age group or gender was missing, values were inferred where possible based on similar customer profiles. If inference was not possible, the entry was marked as “Unknown” to avoid deletion of the record.
- **Product Details:** Missing product categories were mapped by checking Product_ID against the master product table.
- **Store Attributes:** For stores with missing region or type, details were cross-referenced using Store_ID.
- **Sales Transactions:** Records with missing transaction dates or invalid pricing data were carefully reviewed. Transactions without essential values (e.g., no Order_Date) were excluded, as they could not contribute meaningfully to time-based analysis.

By applying these treatments, the dataset maintained both completeness and reliability, ensuring that all key attributes were available for KPIs and dashboards.

3.3 Standardization of Column Naming and Data Formats

For smooth integration across SQL, Python, and Power BI, column headers and formats needed to be standardized. In the raw dataset, some columns had inconsistent or ambiguous naming (e.g., “TotalAmt” vs. “Total_Amount”, “Cust_ID” vs. “Customer_Id”). Such inconsistencies could lead to query errors or difficulties in joining tables.

The following steps were taken:

- **Consistent Naming:** All column names were converted to a standardized format using uppercase letters with underscores
- **Data Type Conversion:**

- Dates were converted into a standard datetime format (YYYY-MM-DD) to ensure accurate time-based grouping in SQL and Power BI.
- Price and cost columns were converted from string/object formats into numeric (FLOAT/DECIMAL) to support arithmetic calculations like profit, discount percentages, and revenue aggregation.
- Customer Age was converted to integer, and age groups were derived using conditional logic.
- **Currency Standardization:** All financial data (revenue, profit, cost, discount) was standardized in Indian Rupees (₹) for consistency.

This standardization step ensured compatibility across tools, minimized query errors, and made the dataset dashboard ready.

3.4 Creation of Indexes and Constraints

Since the dataset contained multiple related tables — Sales, Customers, Products, Stores, and Returns — proper indexing and relationship constraints were established in SQL to maintain integrity and improve performance.

- **Indexes Created:** Indexes were added to frequently queried columns such as Customer_Id, Product_Id, Store_Id, and Order_Date

Retails Project Queries. This drastically improved the speed of queries like “monthly profit by region” or “average revenue per customer by age group.”

- **Foreign Keys Added:**
 - Sales → Customers (via Customer_Id).
 - Sales → Products (via Product_Id).
 - Sales → Stores (via Store_Id).

- Returns → Sales (via Order_Id).

These relationships ensured that every sale linked to a valid customer, product, and store, and every return was tied back to a legitimate sales transaction.

By enforcing constraints, the database was protected against errors such as orphaned records (returns without sales) or invalid foreign keys.

3.5 Outlier Detection and Treatment

In retail data, extreme values can occur — such as unusually high discounts (e.g., 90%), negative profit margins, or customer ages that fall outside realistic ranges (e.g., 150 years). These outliers can distort visualizations and averages, making them unreliable.

- **Profit and Revenue Outliers:** Sales with negative or excessively high profit margins were investigated. Some represented genuine promotional sales, while others were due to incorrect unit price or cost entries. Incorrect records were corrected or removed.
- **Customer Age Outliers:** Unrealistic ages (e.g., <10 or >100 years) were flagged and replaced with “Unknown” or adjusted based on age group distribution.
- **Return Rate Anomalies:** Products with a 100% return rate were closely examined. Some cases were due to test transactions or incorrectly marked returns, which were excluded.

This step ensured that analysis results reflected realistic patterns rather than being driven by anomalies or data-entry errors.

3.6 Derivation of New Fields

To enhance analysis, several new calculated fields were derived:

- **Profit per Order:** $(\text{Unit_Price} - \text{Cost_Price}) \times \text{Quantity} \times (1 - \text{Discount\%})$
- **Age Group Buckets:** (e.g. adults, teenagers, seniors etc).

These derived attributes provided deeper insights that were not directly available in the raw data but were critical for business decisions.

3.7 Final Dataset Export

After cleaning, validation, and transformation, the dataset was saved in CSV format for Power BI ingestion and preserved in SQL tables for ongoing queries. The Python-cleaned version of the dataset was also maintained for EDA.

This structured and pre-processed dataset became the foundation for SQL analysis, EDA in Python, and dashboard development in Power BI, ensuring accuracy, consistency, and reliability across all stages of the project.

CHAPTER 4

DATA ANALYSIS AND VISUALIZATION

4.1 Introduction

This chapter presents the results of the analysis performed on the retail dataset using SQL queries, Python EDA, and Power BI dashboards. The focus is on understanding sales trends, customer behavior, store performance, and return patterns. Dashboards provide an interactive platform to visualize KPIs, trends, and comparisons, making insights more accessible to decision-makers.

4.2 Sales Dashboard

The Sales dashboard provides an overview of revenue, profit, orders, and product performance.

- Key Metrics: Revenue of ₹340K+, Profit of ₹206K, Total Orders = 3000+, Quantity Sold = 5000+.
- Channel Performance: Online sales contribute more than in-store, with higher profit margins per order.
- Top Products: A handful of products (e.g., Brand C Air, Brand B Specific) dominate sales volume.
- Revenue Trend: Peak sales observed mid-year (July) at ₹66.8K, followed by a decline in later months.
This dashboard highlights the importance of online sales and seasonality in revenue trends.

4.3 Store Dashboard

The Store dashboard evaluates operating costs, profitability, and performance by region, city, and store type.

- Key Metrics: Total store cost ~₹2.4M, Avg. store cost ~₹40K.
- Regional Distribution: The West region incurs the highest cost (~₹750K).
- Store Type Mix: Franchise (35%) is the largest type, followed by Flagship and Kiosk stores.

- City Insights: Chicago stores record the highest operational costs (~₹700K).
- Profitability: Robinson PLC Store emerges as the most profitable store.
This dashboard identifies high-cost regions/cities and pinpoints top-performing stores.

4.4 Customer Dashboard

The Customer dashboard profiles demographics, loyalty, and revenue contribution.

- Key Metrics: 800 customers, Avg. age = 44 years, balanced gender ratio (1.04).
- Age Group Insights: Mid-aged customers (31–50) contribute the highest revenue (~₹290K).
- Top Customers: A small group of customers (David, James, Jennifer, etc.) generate the highest profit.
- Regional Presence: West region is the most active, followed by South and East.
- Signup Trends: Customer signups are declining month-over-month.
This dashboard shows that mid-age groups drive most revenue, while retention of loyal customers is crucial.

4.5 Returns Dashboard

The Returns dashboard highlights product return patterns by region, reason, and sales channel.

- Key Metrics: 200 returns, ~7% return rate.
- Regional Trends: North region shows the highest return share (54%).
- Category Insights: Apparel and Personal Care products have the highest return percentages.

- Reasons for Return: Defective items (31.5%) are the main cause, followed by wrong item delivery (26.5%).
- Channel Comparison: Online accounts for 71% of returns, while in-store has fewer returns (29%).
This dashboard points to quality control issues in certain categories and highlights that online returns are a major cost driver.

4.6 Store Dashboard

The Stores Dashboard provides a comprehensive view of retail operations by analysing costs, profitability, and distribution across regions, cities, and store types. It reveals that the company operates 60 stores with a total operating cost of ₹2.41M and an average cost of ₹40.2K per store. Among them, Robinson PLC Store emerges as the most profitable, while the West region and Chicago city incur the highest costs, making them critical areas for cost management. The store network is balanced among Franchise (35%), Flagship (33%), and Kiosk (31%) formats, reflecting a mix of brand visibility and operational flexibility.

4.7 Conclusion of Analysis

From the dashboards, the following insights were derived:

- Online is the most profitable sales channel but also contributes most to returns.
- The West region has high sales and costs, requiring better cost management.
- Mid-aged customers are the core revenue group, while customer acquisition is slowing.
- Apparel and Personal Care categories need better quality assurance to reduce returns.

These findings create a data-driven foundation for improving sales, optimizing store operations, enhancing customer retention, and minimizing returns.

CHAPTER 5

TOOLS AND TECHNOLOGIES

Software Requirement: -

- Python (Jupyter notebook)

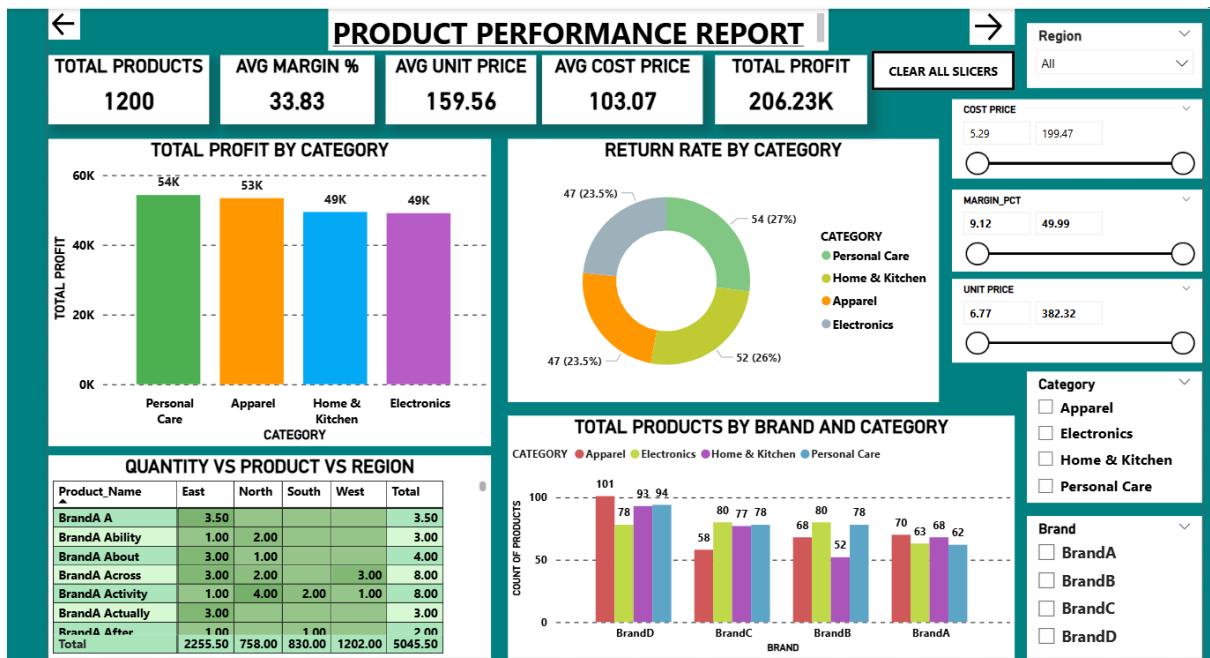
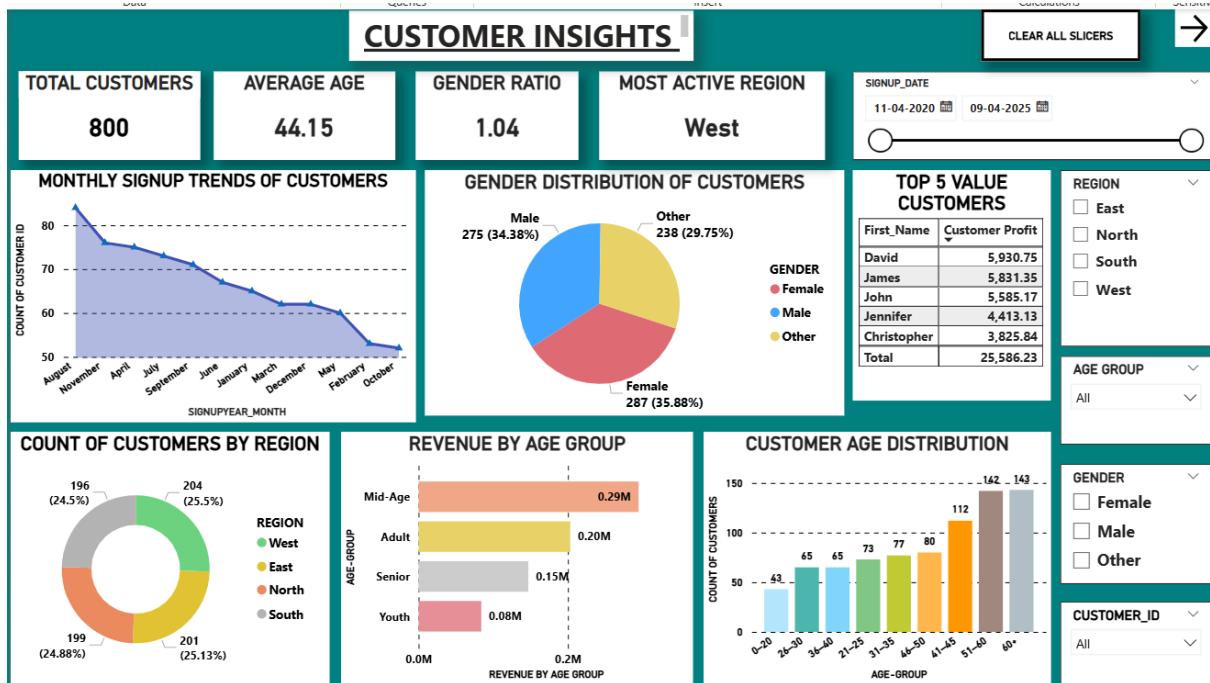
- Power bi
- Excel

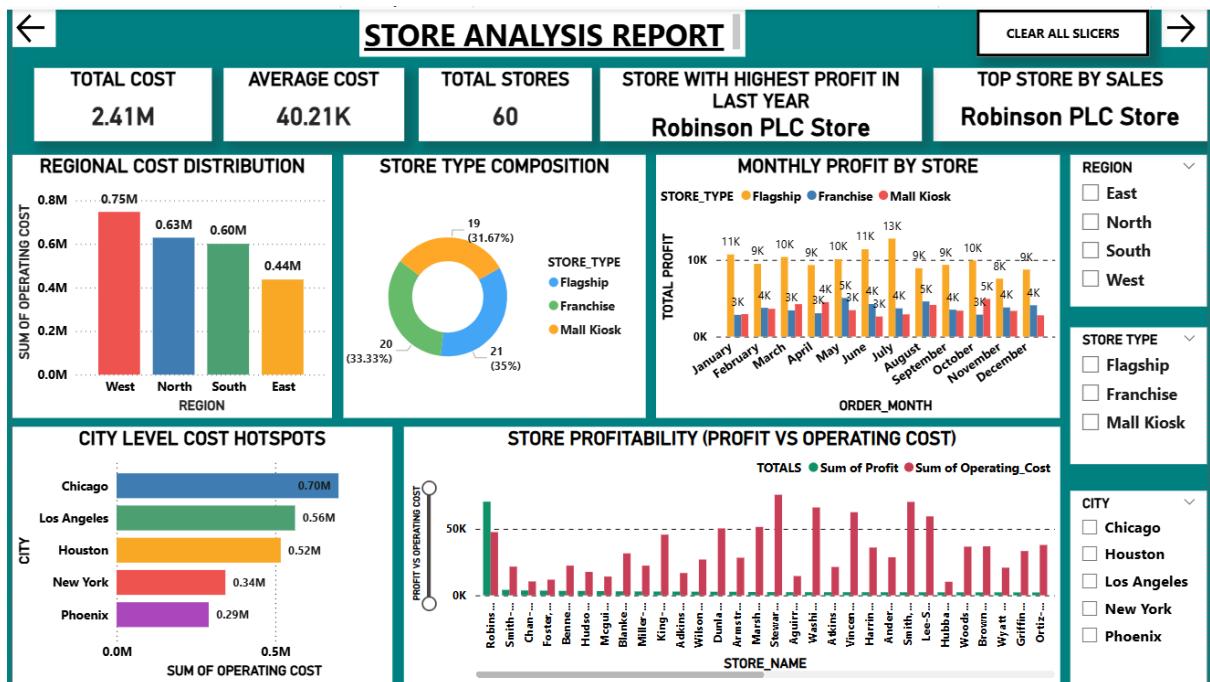
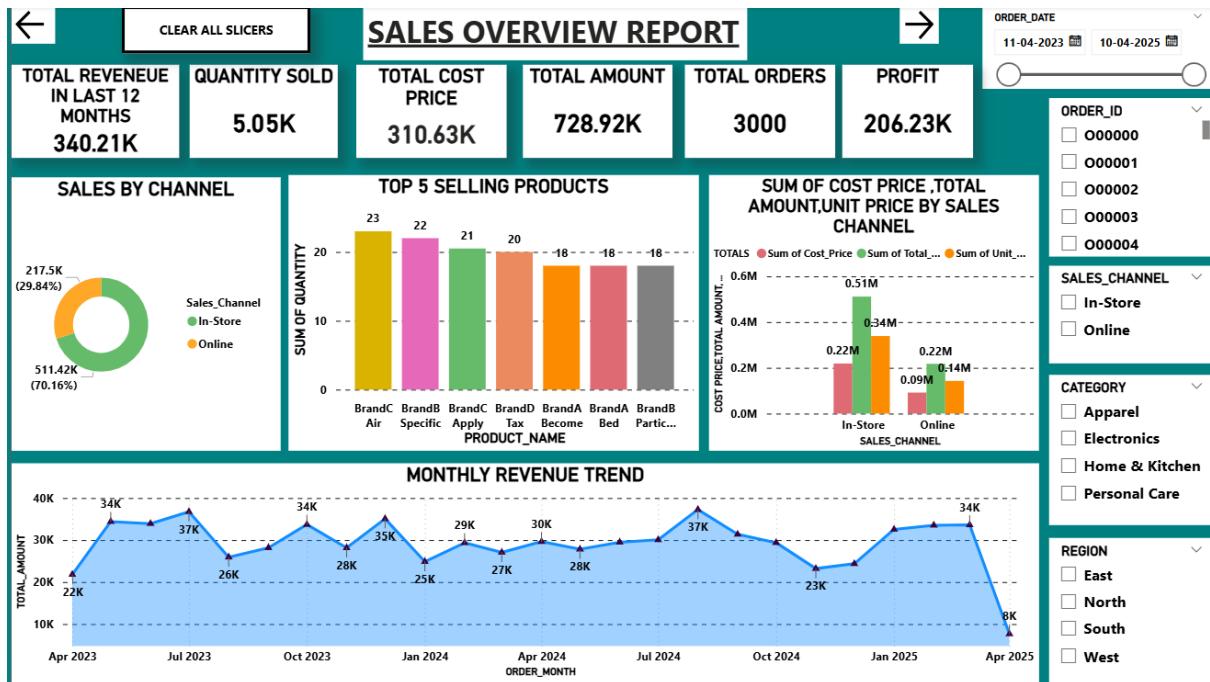
Main Software Libraries: -

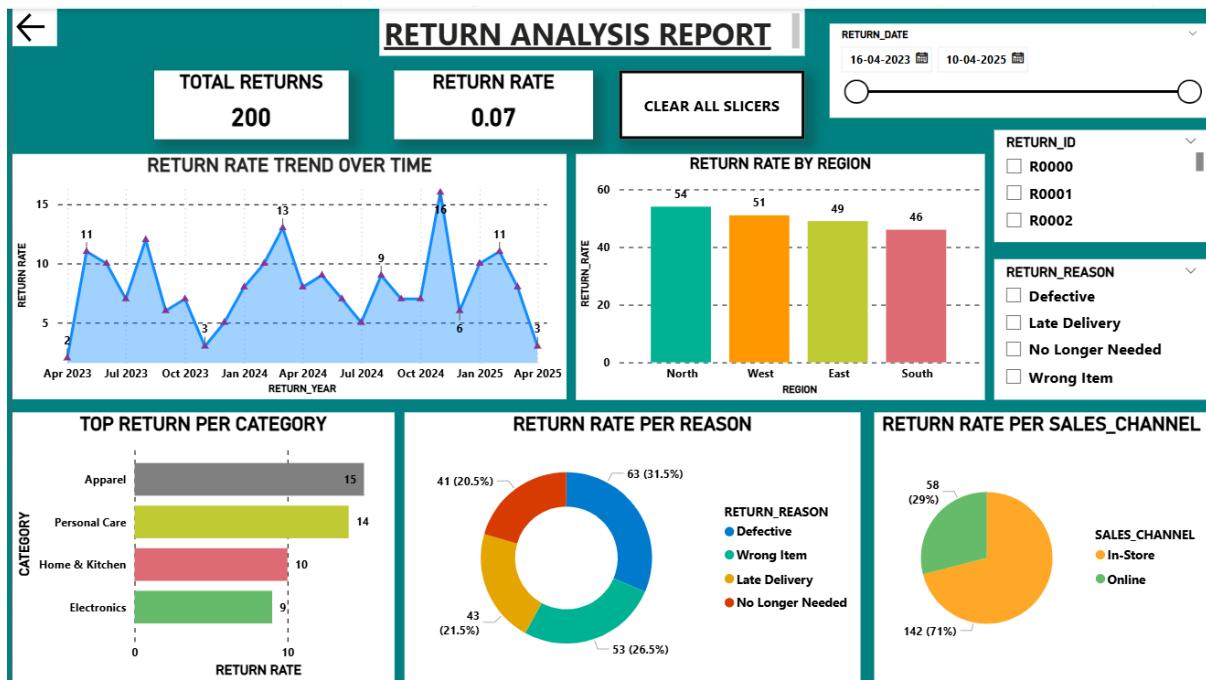
- Pandas
- Numpy
- CSV

CHAPTER 6

DASHBOARDS







CHAPTER 7

CONCLUSION

Primary Objective

The goal of this analysis was to apply data analytics techniques to extract actionable insights from the retail business ecosystem by integrating sales, customers, products, stores, and returns data. The dashboards explore key dimensions such as sales channel performance, store profitability, customer demographics, product return behaviour, and regional cost distribution. By visualizing these metrics interactively, the project delivers a holistic understanding of retail performance—empowering stakeholders to evaluate revenue growth, cost optimization, customer loyalty, and operational efficiency.

1. Sales Performance and Revenue Insights

- The company generated ₹340K+ revenue with a profit of ₹206K from 3000+ orders and 5000+ units sold.
- Online sales channels outperform in-store channels, contributing a higher profit margin per order.
- Top products like *Brand C Air* and *Brand B Specific* dominate in sales volume.
- Revenue trends peak in July (₹66.8K) before tapering off, highlighting seasonal fluctuations.

2. Store Profitability and Cost Distribution

- The company operates 60 stores, incurring a total operating cost of ₹2.41M (~₹40.2K average per store).
- The West region and Chicago city record the highest operating costs, making them priority areas for cost optimization.
- Robinson PLC Store is the most profitable store in the dataset.
- Store composition is balanced among Franchise (35%), Flagship (33%), and Kiosk (31%), offering both scale and flexibility.

3. Customer Demographics and Loyalty

- The customer base consists of 800 customers, with an average age of 44 years and a balanced gender ratio (1.04).
- Mid-aged groups (31–50 years) are the highest contributors to revenue (~₹290K).
- A small segment of loyal customers (David, James, Jennifer, etc.) contributes disproportionately to profit.
- Signup trends show a decline in new customer acquisition, signaling the need for targeted marketing to younger age groups.

4. Returns and Product Performance

- A total of 200 returns were recorded, with a 7% return rate.
- Apparel and Personal Care categories experience the highest return percentages.
- Defective products (31.5%) are the most common reason for returns, followed by wrong item delivery (26.5%).
- The North region contributes the highest return share (54%), while online sales channels account for 71% of returns, emphasizing a need for improved quality checks and logistics in online orders.

Implications for Stakeholders

Consumers (Buyers & Sellers):

- Buyers can identify high-value products and compare channels before purchasing.
- Sellers can position their listings competitively by analyzing return rates and channel profitability.
- Awareness of seasonal sales trends helps both buyers and sellers make better decisions.

Store Managers & Executives:

- High-cost regions (West, Chicago) require targeted cost control strategies.

- Profitable stores like Robinson PLC can be used as benchmarks for best practices.
- Store type analysis helps decide between expanding franchise networks or flagship outlets.

Marketing & Customer Engagement Teams:

- Focus on retaining high-value customers and incentivizing repeat purchases.
- Target underperforming customer segments (youth and new signups) with tailored promotions.
- Address declining signup trends with loyalty programs, referral schemes, and digital campaigns.

Operations & Product Managers:

- Reduce defective product rates in Apparel and Personal Care through stricter quality checks.
- Improve delivery accuracy to lower wrong item returns.
- Optimize logistics for online sales to reduce high return percentages.

Summary

This project demonstrates a data-driven approach to understanding the dynamics of the retail business. Insights from sales channels, customer behaviour, store costs, and product returns offer key takeaways for improving profitability and efficiency. By identifying high-value customers, highlighting cost-heavy regions, and uncovering reasons behind returns, the project provides a roadmap for better decision-making.

Through the combined power of SQL, Python EDA, and Power BI dashboards, the analysis promotes transparency in performance monitoring, highlights areas of improvement, and equips stakeholders with actionable insights to strengthen competitiveness in the retail market.

