CUSTOMER BEHAVIOUR ANALYSIS

Business Problem Statement & Project Documentation

End-to-End Portfolio Project

Python • SQL • Tableau

Dataset: Shopping Trends (3,900 Customers)

Total Revenue: ₹2,33,081 | Avg Purchase: $59.76 | Avg Rating: 3.75

1. Executive Summary

A mid-sized retail business operating across the United States is experiencing stagnating revenue growth despite a large and diverse customer base of over 3,900 active shoppers. While the company has accumulated rich transactional data spanning customer demographics, purchasing patterns, seasonal trends, and marketing interactions, this data has never been systematically analysed to drive strategic decisions.

The business currently operates without a clear understanding of which customer segments drive the most value, what causes customers to stop purchasing (churn), how effective its discounts and promotions actually are, and which products and categories should be prioritised by season. As a result, marketing budgets are being spent without measurable ROI, and high-value customers are not being identified or retained proactively.

This project delivers a complete, data-driven intelligence system — built end-to-end using Python, SQL, and Tableau — to answer these critical business questions and provide actionable recommendations.

3,900	$233K	$59.76	3.75 / 5
Total Customers	**Total Revenue**	**Avg Purchase**	**Avg Rating**
Unique customer records	Aggregate purchase value	Per transaction	Customer satisfaction

2. Business Context & Problem Statement

2.1 Background

The company sells across four product categories — Clothing, Accessories, Footwear, and Outerwear — through multiple shipping methods and payment channels. Customers span multiple age groups (Young Adults, Adults, Seniors), purchase across all four seasons, and vary widely in engagement levels from weekly buyers to those who purchase only once a year. Despite this rich diversity, the company currently treats all customers uniformly, applying blanket discounts and promotions without segmentation.

2.2 Core Business Problem

The Central Challenge

The business cannot identify which customers are worth retaining, which are at risk of churning, which marketing strategies are working, and which product-season combinations are most profitable — because there is no unified analytical framework to connect customer behaviour to business outcomes.

2.3 Specific Pain Points

- **Revenue Leakage:** No customer segmentation exists — all 3,900 customers receive the same marketing and discount treatment regardless of their lifetime value or purchase frequency.
- **Discount Inefficiency:** 32.14% of customers use promo codes, but the business does not know if these drive incremental revenue or merely discount existing intent.
- **Silent Churn:** Churn signals are present in the data (low-frequency, low-engagement customers) but are never acted upon proactively.
- **Missed Seasonal Opportunities:** Seasonal demand patterns vary significantly across product categories, but inventory and promotions are not aligned to these patterns.
- **Operational Blindspots:** No operational benchmark exists to evaluate which shipping method or payment channel leads to higher customer satisfaction.

3. Project Objectives

This project is designed to solve the above business problems through three structured analytical layers, each implemented using a dedicated tool in the data analytics stack:

Objective	Description
Objective 1	Segment customers by value, frequency, and engagement to enable personalised marketing and retention strategies.
Objective 2	Identify high-value loyal customers and flag at-risk (churn-prone) customers before they are lost.
Objective 3	Quantify the real revenue impact of discounts, promo codes, and subscription plans.

Objective 4	Analyse seasonal and category-level product demand to optimise inventory and promotional timing.
Objective 5	Benchmark operational performance across shipping types, payment methods, and locations.
Objective 6	Build interactive Tableau dashboards so business stakeholders can explore these insights without technical skills.

4. Tech Stack & Tool Responsibilities

Python	SQL	Tableau
EDA, Statistical Testing, Churn Scoring	Data Cleaning, Segmentation, BI Views	Dashboards, Visual Storytelling

Python — Exploratory Data Analysis & Statistical Intelligence

Python serves as the analytical engine for this project. The dataset is loaded, cleaned, and explored using pandas and NumPy. Seaborn and Matplotlib are used for static visualisations, including distribution of purchase amounts, category-level spending boxplots, and age-vs-spend scatter plots. Plotly is used for creating interactive charts on gender distribution and location-based patterns.

Statistical techniques applied include: IQR-based outlier detection to identify anomalous high-spend customers, Welch's independent t-test to determine whether gender statistically influences spending behaviour (p-value analysis), Pearson correlation analysis between age and purchase amount, and 95% confidence interval estimation to bound the true population mean of purchase amounts.

The most business-critical Python contribution is the churn risk scoring model: a rule-based function assigns a churn score to each customer based on their purchase frequency and subscription status, flagging customers with scores ≥ 3 as high churn risk. This produces a predicted churn rate across the customer base — a metric the business had no visibility into previously.

SQL — Data Structuring, Segmentation & Business Intelligence Views

SQL is used on the MySQL database (customer_behaviour schema) to perform 30+ structured queries across three difficulty tiers. The foundational layer handles data cleaning: renaming columns, adding a derived age_group column (Minor, Young Adult, Adult, Senior) using CASE logic, and enabling safe updates.

Intermediate queries perform multi-dimensional revenue intelligence — breaking down revenue by location, season, gender, age group, and category simultaneously. The marketing effectiveness queries compare average transaction values and order counts for customers with and without discounts, promo codes, and subscriptions, answering the question of whether these tools genuinely lift revenue.

The advanced SQL section implements a full RFM-style (Recency-Frequency-Monetary) segmentation system. Each customer is assigned a monetary segment (High/Medium/Low Value), an engagement segment (Highly/Moderately/Low Engaged), and a frequency segment (High/Medium/Low Frequency). High-value loyal customers are identified through a compound HAVING clause, and churn-risk customers are isolated using opposite criteria. Finally, a permanent CREATE VIEW statement (customer_behavior_intelligence) consolidates all key dimensions into a reusable BI layer that Tableau connects to for live reporting.

Tableau — Interactive Dashboard & Visual Storytelling

Tableau consumes the SQL view and raw data to build two primary dashboards. The Customer Segmentation Dashboard presents the full segmentation breakdown by purchase frequency (Weekly to Every 3 Months), a treemap of age-group purchase rates showing Adults (37.97%) and Young Adults (31.82%) as dominant segments, a loyal repeat-buyer table, the Top 10 High-Value Customers by score, and a high-value customer horizontal bar chart comparing previous purchases against purchase amounts across 26 filtered customers.

The Product & Category Performance Dashboard provides a shipping-type vs payment-method operations heatmap, a Category-Season cross-analysis revenue matrix (Clothing leads at $104,264 followed by Accessories at $74,200), size trend analysis (M is the dominant size at 44,410 units), colour preference rankings, seasonal product demand by item, and item-wise previous purchase performance. Both dashboards include Location and Category filter controls for slice-and-dice exploration.

5. Research Questions This Project Answers

Each question below is a real business question a retail analytics team would face, mapped to the specific tool used to answer it:

Business Question	Analytical Approach
Which customers are high-value and loyal?	SQL: RFM segmentation + HAVING filter on spend ≥ $300, previous purchases ≥ 10, weekly/fortnightly frequency
Who is at risk of churning?	Python: Rule-based churn score; SQL: Low-frequency, low-spend, low-engagement filter
Does gender affect how much customers spend?	Python: Welch's t-test on male vs. female purchase amounts
Which age group drives the most revenue?	SQL: GROUP BY age_group; Tableau: Age-wise purchase rate treemap
Do discounts actually increase revenue?	SQL: Discount/promo cross-analysis comparing total revenue, avg purchase, and order count
Which product categories perform best by season?	SQL: GROUP BY season, category; Tableau: Category-Season matrix
Which shipping method leads to highest satisfaction?	SQL: Shipping type vs avg review rating
What is the true range of average purchase amount?	Python: 95% confidence interval estimation
What are the dominant colour and size preferences?	Tableau: Colour preference bar chart; size trend bar chart
How do subscribed vs non-subscribed customers differ?	SQL: Subscription vs frequency cross-tab with avg spend

6. Dataset Overview

The dataset used is the Shopping Trends dataset, a synthetic but realistic retail transaction dataset containing 3,900 customer records across 19 original features. Each row represents one customer's shopping profile.

Feature	Description
customer_id	Unique identifier for each customer (1–3,900)
age / age_group	Customer age and derived group (Young Adult, Adult, Senior)
gender	Male / Female
item_purchased / category	Specific item and product category (Clothing, Accessories, Footwear, Outerwear)
purchase_amount_usd	Transaction value in USD (target metric for revenue analysis)
location	US state (50 states represented)
size / color / season	Product attributes used for preference and demand analysis
review_rating	Customer satisfaction rating (1–5 scale)
subscription_status	Whether customer has an active subscription (Yes/No)
payment_method / shipping_type	Operational attributes for cross-analysis
discount_applied / promo_code_used	Binary marketing indicators
previous_purchases	Count of prior transactions — key loyalty indicator
frequency_of_purchases	Purchase cadence: Weekly, Fortnightly, Monthly, Quarterly, Every 3 Months, Annually

7. Key Insights & Business Findings

Customer Segmentation

- Purchase frequency is evenly distributed: Every 3 Months (584), Quarterly (563), Annually (572), Monthly (553), Weekly (539), Fortnightly (542) — suggesting no dominant buying pattern, which indicates an opportunity to shift customers toward higher-frequency tiers.
- Adults (35–54) represent 37.97% of the purchase rate, the largest age segment, followed by Young Adults (31.82%). Senior customers at 30.21% represent a significant and potentially under-served segment.

Revenue & Product Intelligence

- Clothing is the highest-revenue category at $104,264, more than 40% above Accessories ($74,200) and nearly 3x Outerwear ($18,524).
- Clothing revenue is stable across all four seasons (Fall $26,220 / Spring $27,692 / Summer $23,078 / Winter $27,274), suggesting it is a year-round staple that should anchor inventory planning.
- Medium (M) is the dominant size at 44,410 units, followed by Large (27,071), Small (16,429), and XL (10,961) — size assortment should reflect this distribution.

Marketing Effectiveness

- 32.14% of customers use promo codes. The SQL cross-analysis reveals that discount and promo usage does not significantly increase average purchase value, suggesting these tools are attracting existing buyers rather than driving incremental spend.
- Subscribed customers tend to purchase more frequently (Weekly/Fortnightly distribution is higher in the subscribed cohort), making subscription conversion a priority retention strategy.

Churn Risk

- A meaningful percentage of the customer base scores 3+ on the churn risk model (annually purchasing, no subscription), representing an immediate outreach and re-engagement opportunity for the marketing team.

8. Recommendations

Recommendation	Rationale & Action
Launch Tiered Loyalty Programme	Use SQL RFM segmentation to create three customer tiers (Gold, Silver, Bronze). Offer exclusive benefits to High Value + Highly Engaged customers to reduce churn and increase share of wallet.
Convert Discount Users to Subscribers	Since 32% of customers use promo codes without significant lift in spend, redirect that budget toward subscription conversion campaigns — subscribed customers spend more consistently.
Proactive Churn Intervention	Deploy the Python churn score model as a monthly batch process. Customers with score ≥ 3 should receive personalised re-engagement emails with time-limited offers within 30 days of identification.
Seasonal Inventory Optimisation	Use the Category-Season SQL cross-analysis to build a seasonal stock plan. Prioritise Clothing year-round, Outerwear in

	Fall/Winter, and align promotional spend to Spring peaks in Accessories.
Senior Customer Strategy	Seniors (30.21% of purchases) are underweighted in marketing. Design dedicated Senior-friendly product bundles and communication strategies — this is a high-potential growth segment.
Operational Review	Review shipping type vs satisfaction ratings from SQL Query 17. Redirect customers to higher-rated shipping options and communicate this as a service quality improvement.

9. Project Deliverables Summary

Deliverable	Contents
Python Notebook (.ipynb)	Full EDA, hypothesis testing, confidence interval, churn scoring model with visualisations
SQL Script (.sql)	30+ queries across 3 tiers + RFM segmentation + Churn identification + BI View
Tableau Workbook (.twbx)	2 interactive dashboards: Customer Segmentation + Product & Category Performance
Business Document (.docx)	This document — problem statement, objectives, findings, and recommendations

Customer Behaviour Analysis — Portfolio Project

Tools: Python (pandas, seaborn, scipy, plotly) • MySQL • Tableau • Dataset: Shopping Trends (3,900 records)

Author – Soumi Mukherjee
LinkedIn - www.linkedin.com/in/soumimukherjeeofficial
Mail – soumi.mukherjee2003@gmail.com