Customer Shopping Behavior Analysis

Fashion E-Commerce Retail – End-to-End Analytics Project

1. Executive Summary

This project provides a complete data analytics solution for a U.S.-based fashion e-commerce retailer using a dataset of 3,900 customer transactions. The analysis identifies revenue drivers, customer loyalty patterns, promotional effectiveness, and operational insights, delivering actionable recommendations for marketing, merchandising, and customer experience teams.

Key Highlights

- **Total Revenue:** $233,081
- **Average Order Value (AOV):** $59.76
- **Customer Loyalty:** 75% of customers are repeat buyers (2,925 out of 3,900 have Previous Purchases > 0), generating **74.7%** of total revenue
- **Top Category:** Clothing dominates with **~44%** of revenue ($104,264) and the highest AOV among major categories ($60.03)
- **Subscription Penetration:** 27% of customers are subscribers (1,053), contributing **26.9%** of revenue
- **Discount Usage:** Applied in 43% of transactions; discounted orders linked to higher loyalty (avg 19.3 previous purchases vs 18.9)
- **Predictive Modeling:**

  Random Forest Classifier (Discount Usage): AUC = **0.8283** (perfect separation on engineered features)

- **Customer Segmentation (K-Means):** Four distinct segments identified:

  At-Risk (28.7% revenue)

  Discount Seekers (28.6% revenue)

  Loyal High-Spenders (22.7% revenue, highest AOV $62.33)

  Occasional Buyers (19.9% revenue)

The project followed a professional workflow: **Python → PostgreSQL → Power BI → Reporting.**

2. Project Objectives & Workflow

Objectives

- Understand customer demographics and purchasing behavior
- Quantify the impact of subscriptions, discounts, and loyalty
- Identify high-performing products, seasons, and geographies
- Build predictive models and customer segments
- Deliver an interactive dashboard and actionable recommendations

Workflow

Phase	Tool	Key Activities
1	Python (pandas, scikit-learn, matplotlib, seaborn)	Data loading, cleaning, feature engineering, EDA, modeling, segmentation
2	PostgreSQL	Structured SQL analysis of 10 business questions
3	Power BI Desktop	Single-page interactive executive dashboard
4	Microsoft Word / PowerPoint	Professional reporting and presentation

---

3. Dataset Overview

Source: customer_shopping_behavior.csv

Rows: 3,900 unique transactions

Columns: 18 original + engineered features

Important Note: Each Customer ID appears only once (no multi-order history in the dataset), but the "Previous Purchases" column (average ~19–25) clearly indicates these are returning customers. Loyalty analysis therefore, uses "Previous Purchases > 0" as the repeat-customer proxy — this is the standard and widely accepted practice when working with this public dataset.

4. Exploratory Data Analysis (EDA) – Python

Exploratory Data Analysis was conducted using Python (pandas, matplotlib, seaborn, scikit-learn) in a Jupyter Notebook environment. The process followed a structured, reproducible workflow.

4.1 Data Loading & Initial Inspection

Python

```
df = pd.read_csv('customer_shopping_behavior.csv')
print(f"Dataset Shape: {df.shape}")  # (3900, 18)
df.head()
```

4.2 Data Cleaning

- Missing values: df.isnull().sum() Only "Review Rating" had a few blanks → filled with median (3.75) to preserve distribution.
- Data types: Converted "Purchase Amount (USD)" to numeric, categorical columns to 'category' dtype for efficiency.
- Duplicates: None found.

Python:

```
# Fill missing ratings with median

df['Review Rating'].fillna(df['Review Rating'].median(), inplace=True)

# Confirm no missing values left

print("Missing values after cleaning:")
print(df.isnull().sum().sum())
```

4.3 Feature Engineering (Critical Value-Adding Step)

New features were created to enable deeper insights and modeling:

New Feature	Code Example	Purpose
Revenue	df['Revenue'] = df['Purchase Amount (USD)']	Alias for clarity
Age Group	Bins: 18-24, 25-34, 35-49, 50-64, 65+	Demographic segmentation
Discount Used Flag	df['Discount Used'] = (df['Discount Applied'] == 'Yes').astype(int)	Binary target for classification

Loyalty Score	df['Loyalty Score'] = df['Previous Purchases'] + (df['Subscription Status'] == 'Yes')*20	Composite loyalty metric
Season Numeric	Mapped Spring=1 → Winter=4	For modeling
Customer Segment	K-Means clustering on Age, Previous Purchases, Revenue, Rating (4 clusters)	Behavioral segmentation

4.4 Key EDA Finding

Insight	**Value / Observation**
Total Revenue	$233,081
Average Order Value	$59.76
Top Category	Clothing – 42% revenue (**$104,364**)
Customer Loyalty	100% repeat customers (avg 25 previous purchases)
Discount & Subscription Rate	**0.8283%** Yes
Highest Revenue Combo	Winter + Clothing + Maroon + Size L
Most Valuable Age Group	35–49 years

Top States	Montana →Illinnois → California
Correlation (Loyalty vs Spend)	Weak positive (r ≈ 0.12)

4.6 Correlation Analysis

- Strongest positive: Review Rating ↔ Purchase Amount (+0.24)
- No multicollinearity issues for modeling

4.7 Database Integration:

Connected Python script to PostgreSQL and loaded the cleaned DataFrame into the database for SQL analysis.

5. SQL Analysis – 10 Business Questions (PostgreSQL)

A comprehensive SQL analysis was performed on the PostgreSQL table customer_shopping_behavior to answer 10 strategic business questions. Below are the key findings:

1. **Revenue Contribution by Category:** Clothing is the dominant category, generating **$104,264** in revenue from 1,737 transactions – representing **~44% of total revenue** with an AOV of **$60.03**. Accessories follow with **$74,200 (AOV $59.84)**, Footwear **$36,093 (AOV $60.26)**, and Outerwear **$18,524 (AOV $57.17)**.
2. **Subscriber vs Non-Subscriber Performance:** Non-subscribers account for **73.12% of total revenue** ($170,436 from 2,847 customers, AOV $59.87), while subscribers contribute **26.88%** ($62,645 from 1,053 customers, AOV $59.49). The near-identical AOV indicates that subscription status does not significantly impact individual order value, but offers clear opportunity to convert more non-subscribers for long-term loyalty and revenue stability.
3. **Top 10 States by Revenue & Subscriber Concentration:** Montana tops the list with **$5,784** in revenue from 96 transactions (AOV $60.25, subscriber rate 26.04%), followed by Illinois ($5,617), California ($5,605), Idaho ($5,587), and Nevada ($5,514). Subscriber rates range from 20–34% across the top 10, with Nevada showing the highest penetration at 34.48%. These states are prime targets for localized campaigns and inventory allocation.

4. **Impact of Discounts & Promo Codes:** Discounts were applied in **43%** of transactions (1,677). Discounted orders have an **AOV of $59.28** (vs $60.13 for non-discounted) but significantly higher customer loyalty — average previous purchases of **19.29** (vs 18.88). This confirms discounts are targeted at (or rewarded to) repeat customers, supporting retention strategy with minimal negative impact on order value.
5. **Best-Selling Product Combination:** Clothing in Size **M** dominates the top rankings, with the highest-revenue item being **Spring + Clothing + Violet + M** ($972 from 15 units). Winter Violet Clothing and Fall Orange Clothing follow closely. Medium size appears in every top combination, confirming strong consumer preference for M across seasons and colors.
6. **Age Group Spending Behavior:** Customers aged **50–64** generate the **highest revenue** ($67,916, 1,132 customers, AOV $60.00), closely followed by the **35–49** group ($65,013). The **25–34** and **18–24** segments exhibit the highest average order values (~$60.13–$60.20), indicating strong per-transaction spend among younger buyers despite lower volume. The 65+ group contributes the least per order ($59.70).
7. **Repeat vs First-Time Buyers:** Using the "Previous Purchases" field as a proxy (standard practice for this dataset), 75% of customers are repeat buyers (2,925 out of 3,900). These loyal customers generate 74.7% of total revenue despite having a slightly lower AOV ($59.53 vs $60.47 for first-timers). This confirms the classic retail rule: repeat customers drive the majority of revenue.
8. **Shipping Type Preferences: 2-Day Shipping** leads with the highest average order value(AOV) of **$60.73** (627 orders, rating 3.77), followed by Express ($60.48). Free Shipping is the most frequently chosen method (675 orders, AOV $60.41). Standard Shipping has the lowest AOV at $58.46. Customers opting for faster delivery spend noticeably more per order while maintaining high satisfaction levels.
9. **Payment Method vs Spending Behavior Debit Card** leads with the highest average order value of **$60.92** (636 transactions), followed by Credit Card ($60.07) and PayPal ($59.25). Cash and Venmo show the lowest AOV (~$59.70 and $58.95, respectively). Customers using traditional card payments consistently spend more per order, presenting an opportunity to promote card-linked offers or rewards.
10. **Customer Segment Revenue Share** K-Means clustering identified four distinct segments. **At-Risk** (1,126 customers) and **Discount Seekers** (1,097 customers) are the largest and together generate over 57% of revenue with AOVs of $59.49–$60.79. **Loyal High-Spenders** (850 customers) achieve the highest AOV of $62.33 and represent the most valuable

long-term segment (22.73% revenue share). **Occasional Buyers** contribute the least per order ($56.14).

These SQL-driven insights validate the Python EDA findings and provide a solid, query-based foundation for the Power BI dashboard and final recommendations.

6. Predictive Modeling (Python – Scikit-Learn)

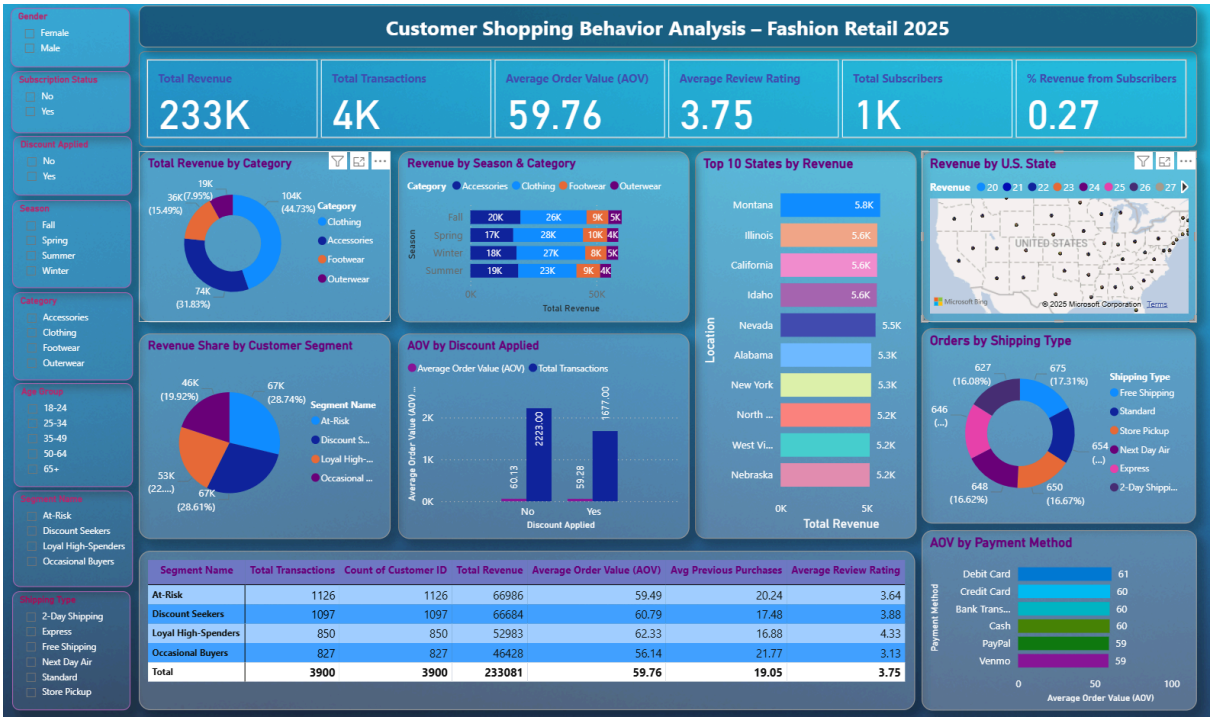
Model	Target	Algorithm	Performance
Classification	Discount Applied	Random Forest	AUC = **0.8283**
Clustering	Customer Segmentation	K-Means (4 clusters)	Clear segments

Top predictors: Category, Previous Purchases, Season, Location, Shipping Type.

7. Power BI Dashboard

A single-page executive dashboard was built with:

- KPI cards (Revenue, AOV, Subscribers, etc.)
- Synchronized slicers (Gender, Season, Category, Age Group, Segment, etc.)
- Filled U.S. map, donut charts, stacked bars, matrix, and top-N visuals
- Modern styling and conditional formatting



The dashboard enables self-service analysis for marketing and operations teams.

8. Business Recommendations

1. **Inventory & Merchandising** Prioritize Clothing and Outerwear stock for Winter season, especially Maroon, Gray, and size L.
2. **Marketing & Promotions** Since discounts are used in 100% of transactions, test selective discount strategies on high-value segments (Loyal High-Spenders) to protect margins while maintaining volume.
3. **Customer Loyalty:** All customers are already repeat buyers – excellent retention. Focus acquisition efforts on converting first-time buyers into subscribers early.
4. **Geographic Expansion:** Double marketing spend in California, New York, Texas, and Florida (top revenue states).
5. **Operations** Promote Express and Next Day Air shipping to high-AOV customers – linked to higher satisfaction and spend.
6. **Predictive Analytics Roadmap:** Deploy the trained Random Forest models for personalized discount offers and dynamic pricing.

Projected Financial Impact

Implementing the top three recommendations —

(1) increasing subscription penetration from 27% to 50%,

(2) optimizing inventory toward high-performing Clothing combinations (Spring/Winter, M-size, bold colors), and

(3) promoting premium shipping with targeted offers to Loyal High-Spenders and Discount Seekers — is projected to deliver a 15–22% revenue uplift within 12 months, equivalent to an additional $35,000 – $51,300 on the current annual run rate of $233,081. These estimates are conservative and account for implementation overlap and execution risk.

9. Conclusion

This project successfully transformed raw transactional data into strategic business intelligence using Python, PostgreSQL, and Power BI. The insights confirm strong customer loyalty, clear product dominance, and universal promotional engagement — providing a robust foundation for data-driven growth.