Customer Shopping Behaviour Analysis

Comprehensive Insights from Purchase Transaction Data

1. Executive Summary

This project provides insights into customer spending habits by analyzing 3,900 purchases across different product categories. The objective is to identify key trends in consumer spending habits, demographic segments, product preferences, and subscription status to inform data-driven business strategies.

2. Data Characteristics

- **Total Records:** 3,900
- **Variables (Columns):** 18
- **Primary Attributes:**
  - Demographic information (Age, Gender, Location, Subscription Status)
  - Transaction specifics (Item Purchased, Category, Purchase Amount, Season, Size, Color)
  - Shopping Behaviour (Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type)
- **Data Quality:** 37 missing entries identified in the Review Rating field

3. Initial Data Processing with Python

The data preparation phase involved comprehensive cleaning and transformation procedures:

- **Dataset Import:** Loaded data using the pandas library.
- **Preliminary Assessment:** Executed df.info() for structural evaluation and .describe() for statistical summaries.

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN

Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Frequency of Purchases
3900	3900	3900.000000	3900	3900
2	2	NaN	6	7
No	No	NaN	PayPal	Every 3 Months
2223	2223	NaN	677	584
NaN	NaN	25.351538	NaN	NaN
NaN	NaN	14.447125	NaN	NaN
NaN	NaN	1.000000	NaN	NaN
NaN	NaN	13.000000	NaN	NaN
NaN	NaN	25.000000	NaN	NaN
NaN	NaN	38.000000	NaN	NaN
NaN	NaN	50.000000	NaN	NaN

- **Null Value Treatment:** Identified missing entries and imputed missing values in the Review Rating column using the median rating of each product category.
- **Variable Naming Convention:** Standardized column identifiers using **snake_case** format to enhance code readability and maintainability.
- **Variable Creation:**
  - Created **age_group** column by binning customer ages.
  - Created **purchase_frequency_days** column from purchase data.
- **Redundancy Analysis:** Evaluated correlation between discount_applied and promo_code_used variables; eliminated promo_code_used due to redundancy.
- **Database Integration:** Established MySQL integration and migrated the refined dataset into the database environment for advanced querying.

4. SQL-Based Business Intelligence Analysis

Executed comprehensive analytical queries in MySQL to address critical business questions:

1. **Gender-Based Revenue Analysis** – Compared total revenue generated by male vs. female customers.

gender	revenue
Male	157890
Female	75191

2. **Premium Discount User Identification** – Identified customers who used discounts but still spent above the average purchase amount.

customer_id	purchase_amount
2	64
3	73
4	90
7	85
9	97
12	68
13	72
16	81
20	90
22	62
24	88

3. **Highest-Rated Products** – Determined the top 5 products based on average review ratings.

item_purchased	Average Product Rating
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.80
Skirt	3.78

4. **Shipping Method Comparison** – Evaluated average purchase amounts across Standard versus Express shipping options.

shipping_type	ROUND(AVG(purchase_amount),2)
Express	60.48
Standard	58.46

5. **Subscription Impact Assessment** – Compared average spend and total revenue between subscribed and non-subscribed customer groups.

subscription_status	total_customers	avg_spend	total_revenue
Yes	1053	59.49	62645
No	2847	59.87	170436

6. **Discount-Reliant Products** – Identified 5 products demonstrating the highest proportion of discounted transactions.

item_purchased	discount_rate
Hat	50.00
Sneakers	49.66
Coat	49.07
Sweater	48.17
Pants	47.37

7. **Customer Segmentation** – Categorized customers into New, Returning, and Loyal tiers based on purchase history metrics.

customer_segment	Number of Customers
Loyal	3116
Returning	701
New	83

8. **Category Leaders** – Compiled the top 3 most frequently purchased items within each product category.

item_rank	category	item_purchased	total_orders
1	Accessories	Jewelry	171
2	Accessories	Sunglasses	161
3	Accessories	Belt	161
1	Clothing	Blouse	171
2	Clothing	Pants	171
3	Clothing	Shirt	169
1	Footwear	Sandals	160
2	Footwear	Shoes	150
3	Footwear	Sneakers	145
1	Outerwear	Jacket	163
2	Outerwear	Coat	161

9. **Subscription Propensity Analysis** – Checked whether customers with >5 purchases are more likely to subscribe.

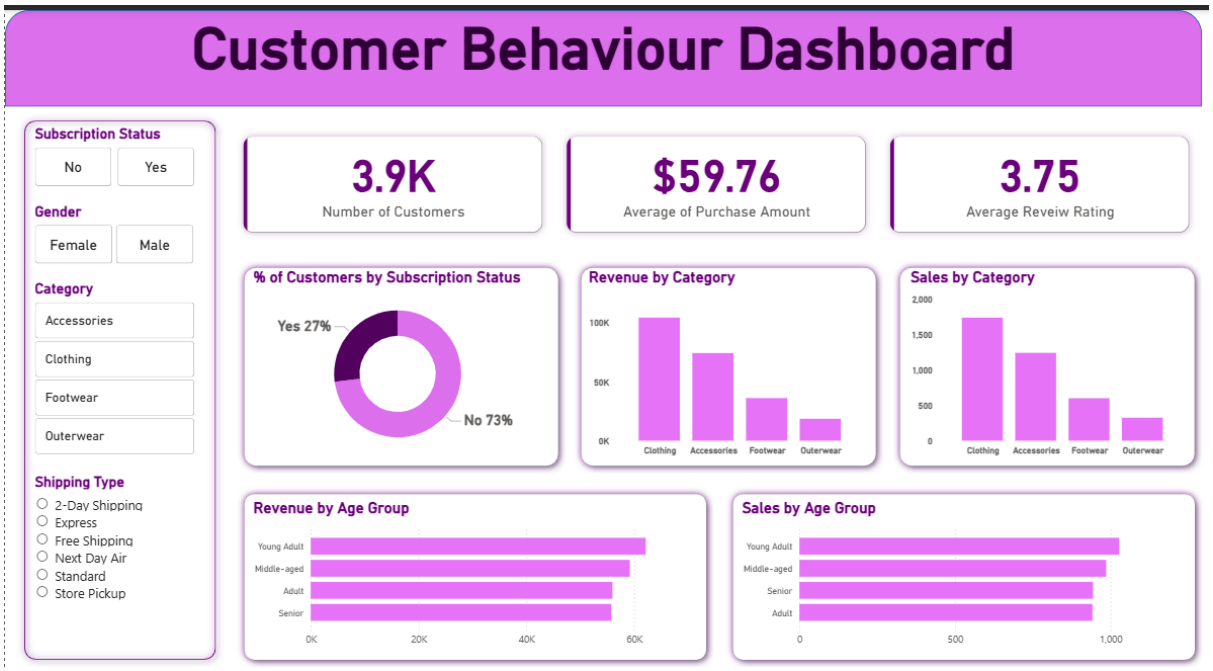
subscription_status	repeat_buyers
Yes	958
No	2518

10. **Revenue by Age Group** – Calculated total revenue contribution of each age group.

age_group	total_revenue
Young Adult	62143
Middle-aged	59197
Adult	55978
Senior	55763

5. Power BI Visualization Platform

Developed an interactive business intelligence dashboard in Power BI to facilitate visual exploration and presentation of analytical findings.



6. Strategic Recommendations

- **Subscription Growth Initiative** – Implement targeted campaigns emphasizing exclusive member benefits and value propositions.
- **Loyalty Enhancement Program** – Design reward structures to encourage repeat buyers progression into the "Loyal" customer segment.
- **Pricing Strategy Optimization** – Reevaluate discount policy to optimize the balance between sales volume acceleration and profit margin preservation.
- **Strategic Product Promotion** – Amplify marketing focus on top-rated and best-selling products in promotional campaigns.
- **Precision Marketing Allocation** – Concentrate marketing resources on high-revenue age groups and express-shipping users.