Customer Shopping Behaviour Analysis:

1. Project Overview

This project focuses on analyzing customer shopping behavior using transactional data from **3,900 purchases** across multiple product categories. The objective is to identify key insights related to **spending habits, customer segmentation, product preferences, and subscription trends**, enabling data-driven strategic decision-making for businesses.

2. Dataset Summary

- **Total Records:** 3,900
- **Total Features:** 18

Key Attributes Include:

- **Customer Demographics:** Age, Gender, Location, Subscription Status
- **Purchase Information:** Item Purchased, Product Category, Purchase Amount, Season, Size, Colour
- **Shopping Behaviour Indicators:** Discount Applied, Promo Code Usage, Previous Purchases, Purchase Frequency, Review Ratings, Shipping Type

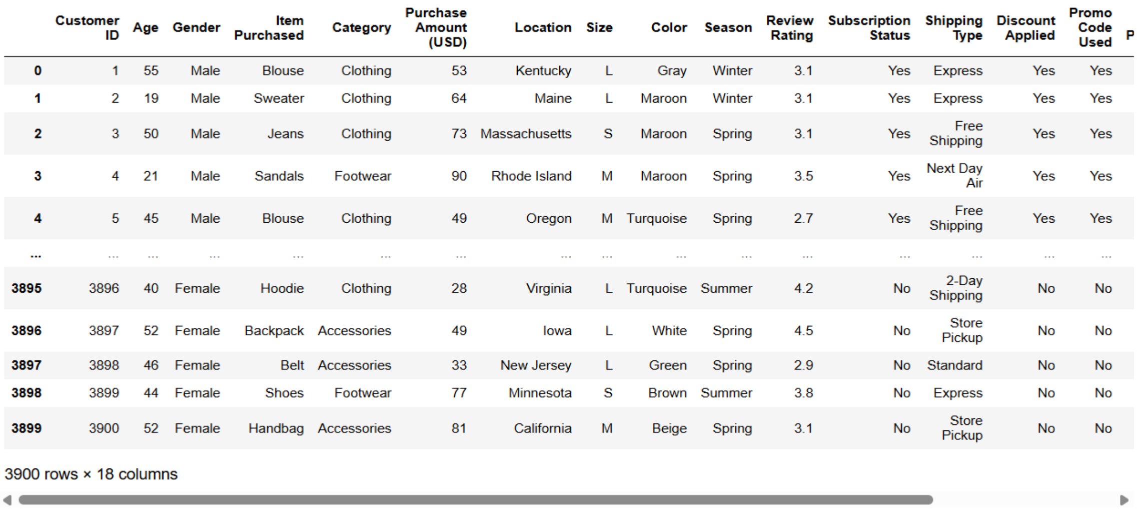
Data Quality Note:

- The Review Rating column contains **37 missing values**, which require appropriate handling during data preprocessing.

3. Exploratory Data Analysis (EDA) Using Python

The exploratory data analysis phase focused on preparing, cleaning, and structuring the dataset using Python to ensure reliable insights:

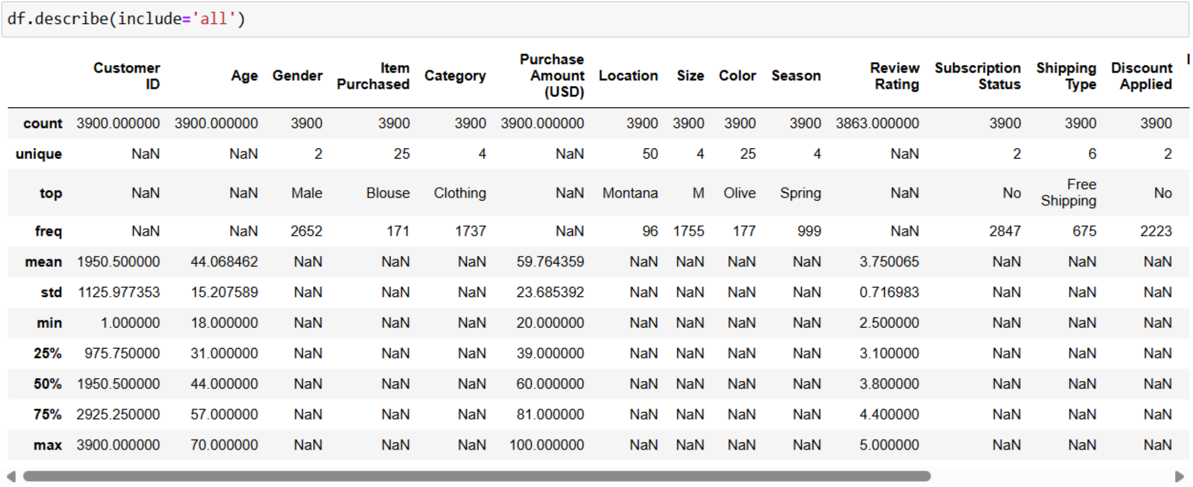
- **Data Loading:**
  The dataset was imported into Python using the pandas library.



	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	P
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3895	3896	40	Female	Hoodie	Clothing	28	Virginia	L	Turquoise	Summer	4.2	No	2-Day Shipping	No	No	
3896	3897	52	Female	Backpack	Accessories	49	Iowa	L	White	Spring	4.5	No	Store Pickup	No	No	
3897	3898	46	Female	Belt	Accessories	33	New Jersey	L	Green	Spring	2.9	No	Standard	No	No	
3898	3899	44	Female	Shoes	Footwear	77	Minnesota	S	Brown	Summer	3.8	No	Express	No	No	
3899	3900	52	Female	Handbag	Accessories	81	California	M	Beige	Spring	3.1	No	Store Pickup	No	No	

3900 rows × 18 columns

- **Initial Exploration:**
  Basic data inspection was performed using df.info() to understand the data structure and df.describe() to obtain summary statistics.



```
df.describe(include='all')
```

	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

- **Handling Missing Values:**
  Null values were identified, and missing entries in the Review Rating column were imputed using the median rating within each product category, preserving category-specific trends.

- **Column Standardization:**
  Column names were converted to snake case to improve readability, consistency, and documentation clarity.

- **Feature Engineering:**
  - An age_group feature was created by binning customer ages into meaningful ranges.
  - A purchase_frequency_days feature was derived from purchase-related data to better capture buying patterns.

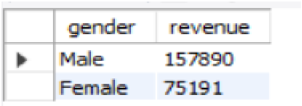
- **Data Consistency Check:**
  Redundancy between discount_applied and promo_code_used was evaluated. Since both conveyed similar information, promo_code_used was removed to reduce duplication.

- **Database Integration:**
  The cleaned and transformed DataFrame was then connected to a **MYSQL Workbench** database and stored for further SQL-based analysis.

4. Data Analysis Using SQL (Business Transactions)

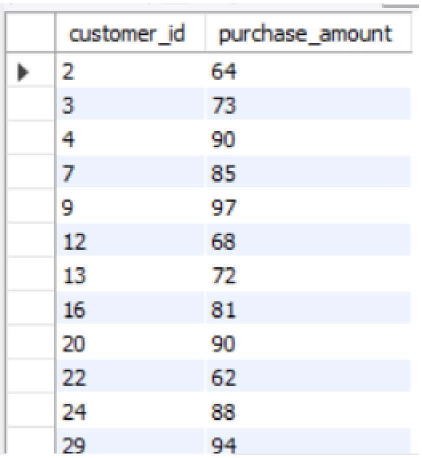
Structured queries were executed in **MYSQL Workbench** to extract actionable business insights from the transactional data. The analysis addressed the following key questions:

1. **Revenue by Gender:**
   Compared total revenue contributions from male and female customers.



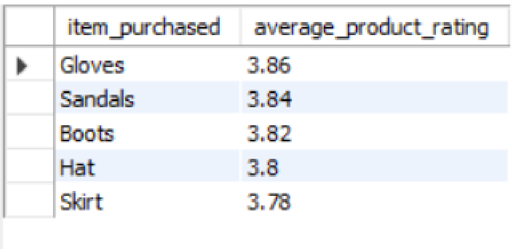
gender	revenue
Male	157890
Female	75191

2. **High-Spending Discount Users:**
   Identified customers who applied discounts yet still spent above the overall average purchase value.



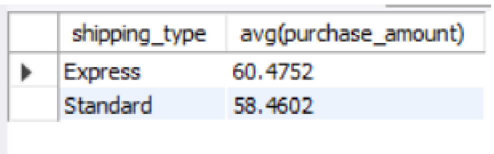
customer_id	purchase_amount
2	64
3	73
4	90
7	85
9	97
12	68
13	72
16	81
20	90
22	62
24	88
29	94

3. **Top 5 Products by Rating:**
   Determined the five products with the highest average customer review ratings.



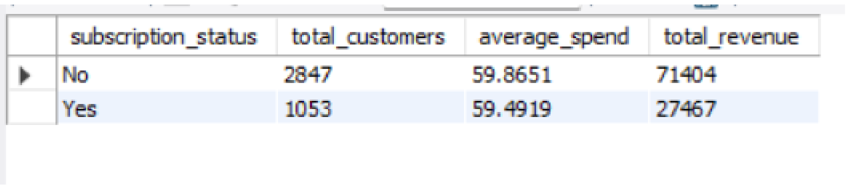
item_purchased	average_product_rating
Gloves	3.86
Sandals	3.84
Boots	3.82
Hat	3.8
Skirt	3.78

4. **Shipping Type Comparison:**
   Analyzed differences in average purchase amounts between **Standard** and **Express** shipping options.



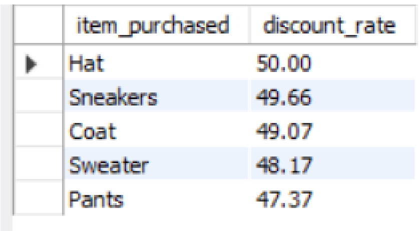
shipping_type	avg(purchase_amount)
Express	60.4752
Standard	58.4602

5. **Subscribers vs. Non-Subscribers:**
   Compared both average spending and total revenue between subscribed and non-subscribed customers.



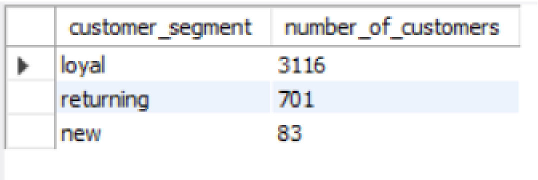
subscription_status	total_customers	average_spend	total_revenue
No	2847	59.8651	71404
Yes	1053	59.4919	27467

6. **Discount-Dependent Products:**
   Identified the top five products with the highest proportion of purchases made using discounts.



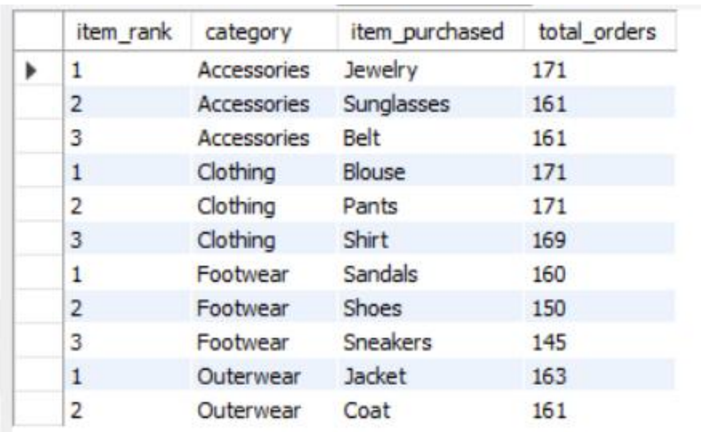
item_purchased	discount_rate
Hat	50.00
Sneakers	49.66
Coat	49.07
Sweater	48.17
Pants	47.37

7. **Customer Segmentation:**
   Segmented customers into **New**, **Returning**, and **Loyal** groups based on their purchase history.



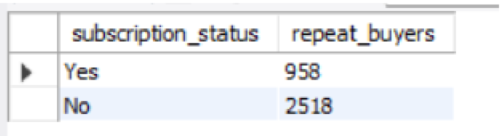
customer_segment	number_of_customers
loyal	3116
returning	701
new	83

8. **Top 3 Products per Category:**
   Ranked and listed the three most frequently purchased products within each category.



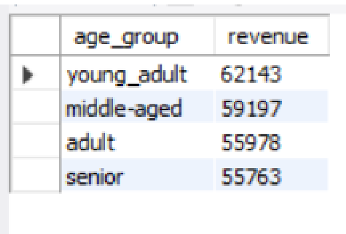
item_rank	category	item_purchased	total_orders
1	Accessories	Jewelry	171
2	Accessories	Sunglasses	161
3	Accessories	Belt	161
1	Clothing	Blouse	171
2	Clothing	Pants	171
3	Clothing	Shirt	169
1	Footwear	Sandals	160
2	Footwear	Shoes	150
3	Footwear	Sneakers	145
1	Outerwear	Jacket	163
2	Outerwear	Coat	161

9. **Repeat Buyers & Subscriptions:**
   Evaluated whether customers with more than five purchases showed a higher likelihood of having a subscription.



subscription_status	repeat_buyers
Yes	958
No	2518

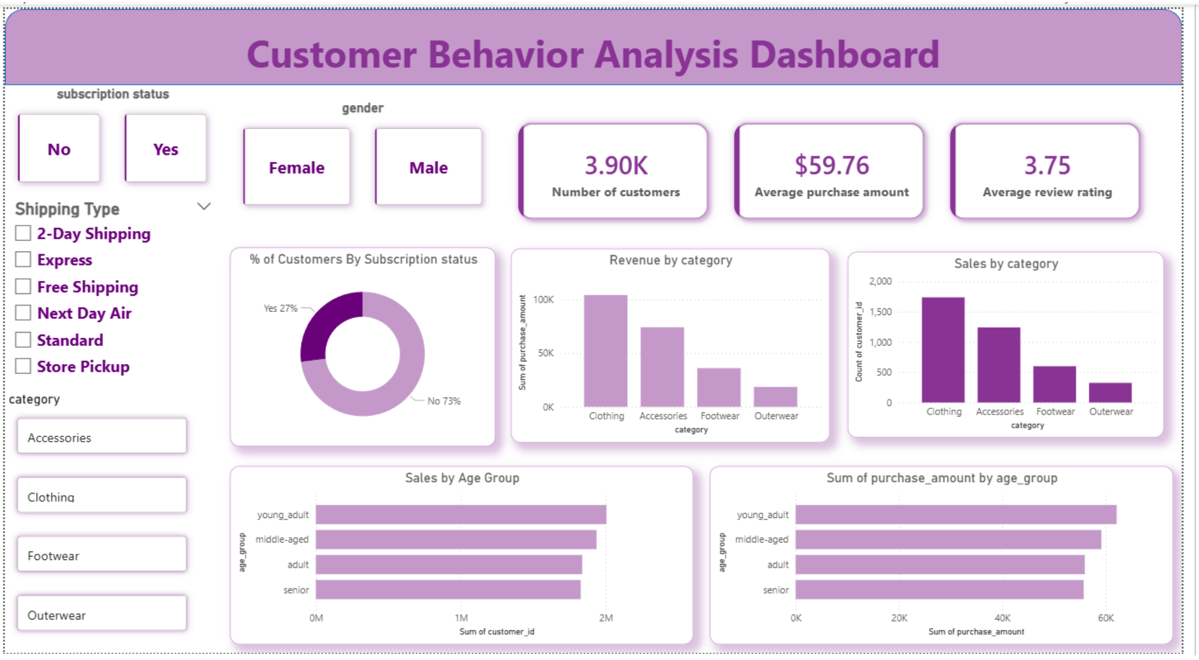
10. **Revenue by Age Group:**
    Calculated and compared total revenue contributions across different age groups.



age_group	revenue
young_adult	62143
middle-aged	59197
adult	55978
senior	55763

5. Interactive Dashboard in Power BI

An interactive **Power BI dashboard** was developed to visually communicate key insights from the analysis. The dashboard enables stakeholders to explore trends related to **customer demographics, purchasing behavior, product performance, discounts, and subscription status** through dynamic filters and intuitive visualizations, supporting faster and more informed decision-making.



Customer Behavior Analysis Dashboard

3.90K Number of customers

$59.76 Average purchase amount

3.75 Average review rating

Business Recommendations

- **Strengthen Subscription Adoption:**
  Promote exclusive subscriber-only benefits such as early access to sales, free express shipping, and loyalty points to encourage repeat purchases and build a stable stream of recurring revenue.

- **Enhance Customer Loyalty Programs:**
  Implement tier-based loyalty rewards (Returning → Loyal) to incentivize frequent purchases, increase customer lifetime value, and reduce churn.

- **Optimize Discount Strategy:**
  Reassess discount-heavy products and introduce data-driven, targeted promotions to boost sales while protecting profit margins and avoiding over-reliance on discounts.

- **Targeted & Lifecycle-Based Marketing:**
  Deploy personalized marketing campaigns tailored to high-revenue age groups and customer lifecycle stages (New, Returning, Loyal), ensuring more relevant messaging, higher engagement, and improved conversion rates.