Customer Shopping Behavior Analysis

Understanding purchasing patterns across 3,900 shopping records to drive strategic business decisions

Customer Shopping Behavior Analysis

1. Project Overview

This project focuses on understanding customer purchasing patterns by examining transactional data from approximately **3,900 shopping records** spanning multiple product categories. The objective of the analysis is to identify trends in spending habits, customer segmentation, product preferences, and subscription-related behaviors. These insights aim to support data-driven strategic and marketing decisions for business optimization.

2. Dataset Summary

- **Total Rows:** 3,900
- **Total Columns:** 18

Main Variables Included:

- **Customer Demographics:** Age, Gender, Location, Subscription Status

- **Purchase Information:** Purchased Item, Product Category, Amount Spent, Season, Size, Color
- **Behavioral Attributes:** Discount Usage, Promo Code Application, Past Purchases, Purchase Frequency, Product Review Rating, Shipping Method

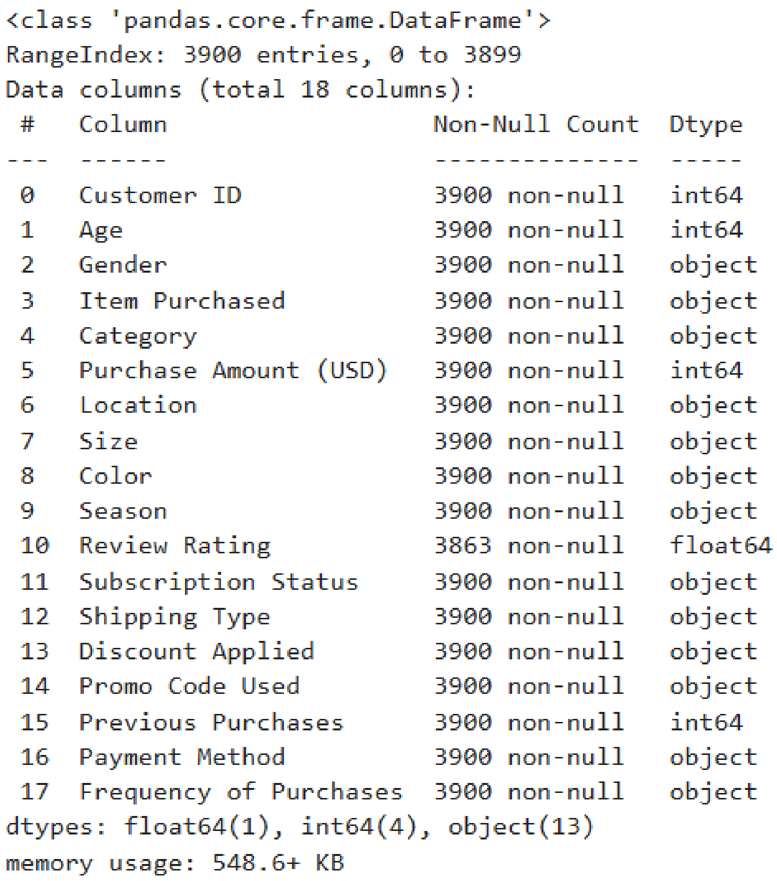
Data Quality Check:

- Identified **37 missing values** in the *Review Rating* field.

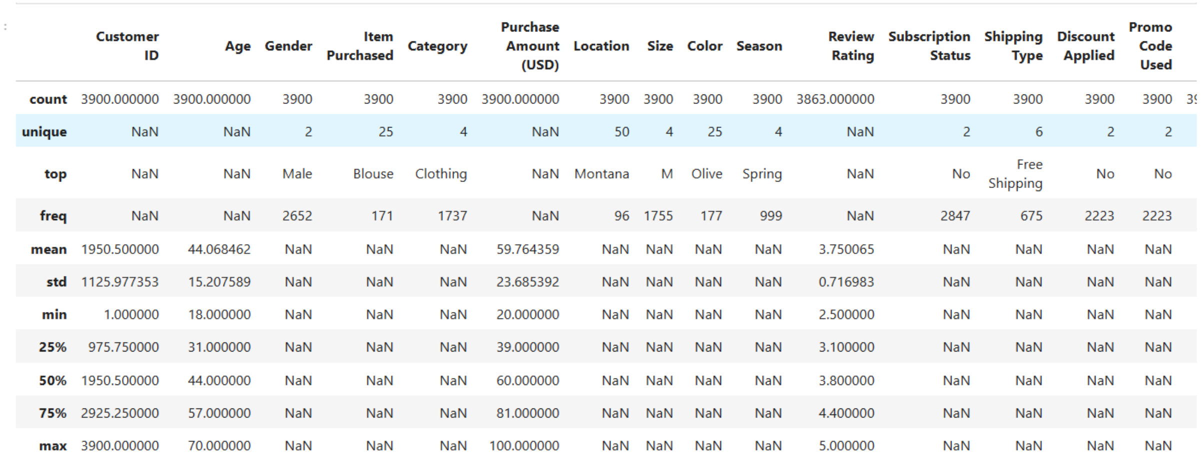
3. Exploratory Data Analysis (EDA) in Python

The initial analysis involved preparing and exploring the dataset using Python:

- **Data Import:** Loaded the dataset using *pandas* for manipulation and exploration.
- **Structural Review:** Utilized functions like df.info() to understand column types and dataset structure.
- **Statistical Summary:** Applied .describe() to generate descriptive statistics for key numerical attributes.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                  Non-Null Count  Dtype  
---  ------                  --------------  -----  
 0   Customer ID             3900 non-null   int64  
 1   Age                     3900 non-null   int64  
 2   Gender                  3900 non-null   object 
 3   Item Purchased          3900 non-null   object 
 4   Category                3900 non-null   object 
 5   Purchase Amount (USD)   3900 non-null   int64  
 6   Location                3900 non-null   object 
 7   Size                    3900 non-null   object 
 8   Color                   3900 non-null   object 
 9   Season                  3900 non-null   object 
 10  Review Rating           3863 non-null   float64
 11  Subscription Status     3900 non-null   object 
 12  Shipping Type           3900 non-null   object 
 13  Discount Applied        3900 non-null   object 
 14  Promo Code Used         3900 non-null   object 
 15  Previous Purchases      3900 non-null   int64  
 16  Payment Method          3900 non-null   object 
 17  Frequency of Purchases  3900 non-null   object 
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```



	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	
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900	3
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2	
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No	
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223	
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN	
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN	
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN	
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN	
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN	
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN	
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN	

4. Data Cleaning and Preparation in Python

- **Handling Missing Values**:
  Performed a missing-value check and filled the null entries in the *Review*

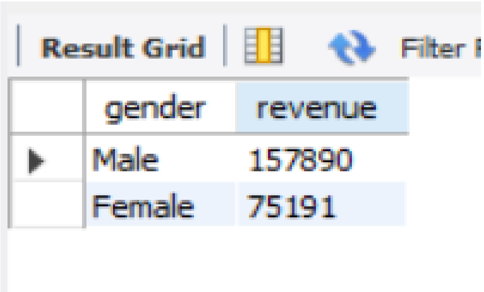
Rating column by assigning the median rating of the corresponding product category.

- **Column Standardization:**
  Converted all column names into snake_case format to ensure clean, consistent, and easily understandable documentation.
- **Feature Engineering:**
  - Created an age_group field by segmenting customers into defined age brackets.
  - Derived a purchase_frequency_days feature to calculate the time interval between customer purchases.
- **Data Consistency Verification**:
  Examined the relationship between *discount_applied* and *promo_code_used* to detect duplication. Since both variables conveyed overlapping information, the *promo_code_used* column was removed for clarity.
- **Database Integration**:
  Established a connection between Python and MySQL using a suitable connector, and uploaded the cleaned DataFrame into the MySQL database for further querying and analysis**.**

4. Data Analysis using SQL (Business Transactions)

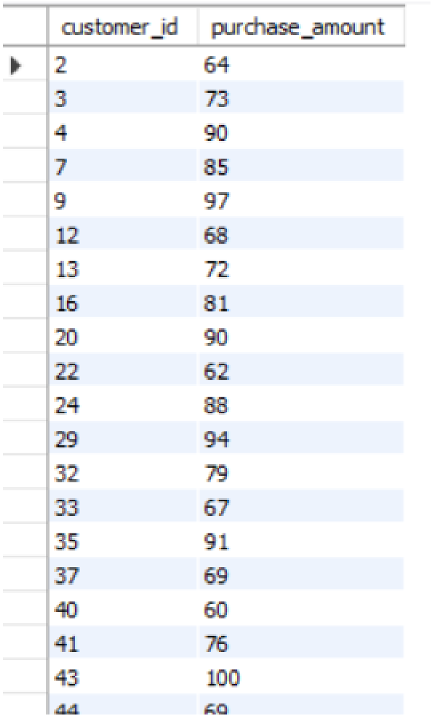
To address core business questions, structured analysis was carried out in MySQL using the cleaned transactional dataset. The SQL queries helped uncover patterns in customer behavior and overall business performance.

1. Revenue by Gender
Analyzed and compared the total spending contributed by male and female customers to understand which demographic segment drives higher revenue.



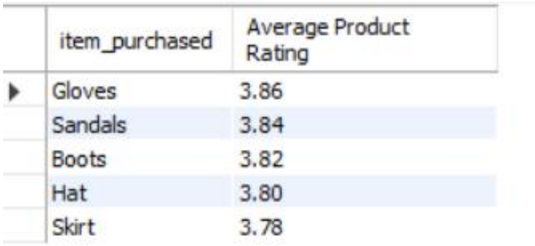
gender	revenue
Male	157890
Female	75191

2. High-Value Customers Using Discounts
Identified those customers who availed discounts yet maintained spending levels above the overall average purchase amount. This helped highlight valuable shoppers who respond to promotional offers without reducing their total spending.



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
32	79
33	67
35	91
37	69
40	60
41	76
43	100
44	69

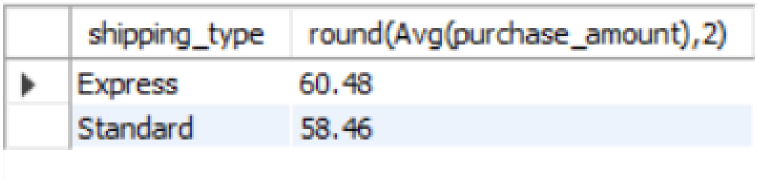
3. Top 5 Products by Rating – Found products with the highest average review ratings.



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

4. Comparison of Shipping Types

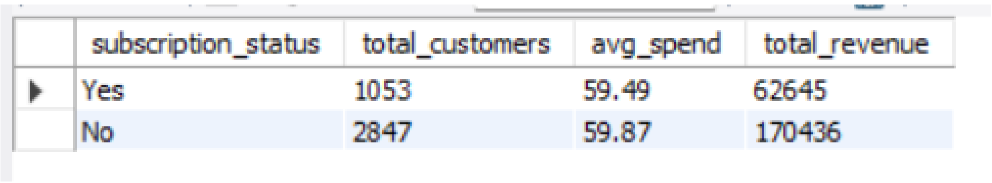
Evaluated how the choice of shipping method—Standard versus Express—impacts customer spending by comparing the average purchase amount across both shipping categories. This helped identify whether faster delivery options are associated with higher transaction values.



shipping_type	round(Avg(purchase_amount),2)
Express	60.48
Standard	58.46

5. Subscribers vs. Non-Subscribers

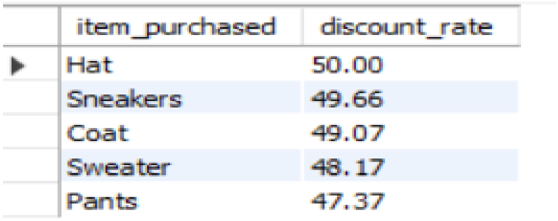
Compared average spend and total revenue across subscription status



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

6. Products Highly Driven by Discounts

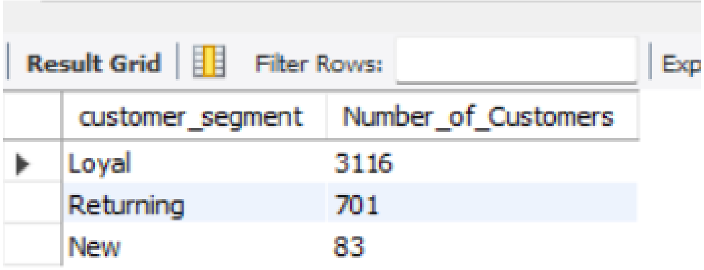
Determined the top five products with the largest share of purchases made using discounts. This analysis highlighted items whose sales performance is strongly influenced by promotional pricing.



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

7. Customer Segmentation Based on Purchase Behavior

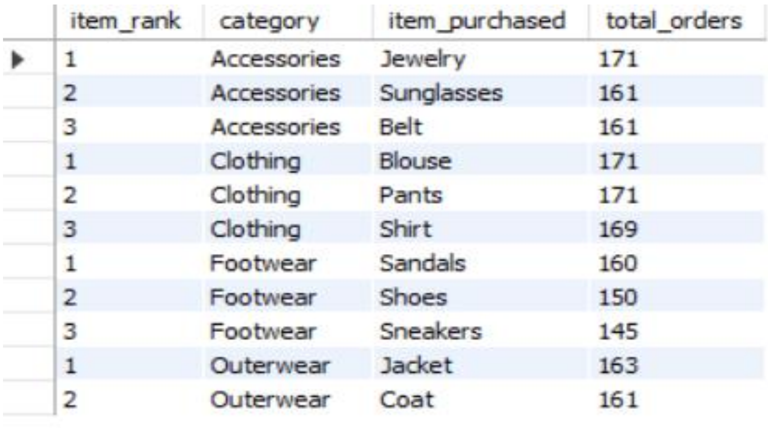
Grouped customers into three segments—**New**, **Returning**, and **Loyal**—by examining their purchase history and frequency. This segmentation provided a clearer understanding of customer lifecycle stages and overall engagement patterns.



customer_segment	Number_of_Customers
Loyal	3116
Returning	701
New	83

8. Top 3 Products Within Each Category

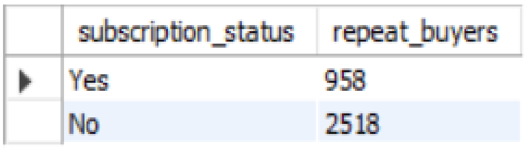
Identified the three most frequently purchased products under every product category to understand which items dominate customer demand in each segment.



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. Relationship Between Repeat Purchases and Subscriptions

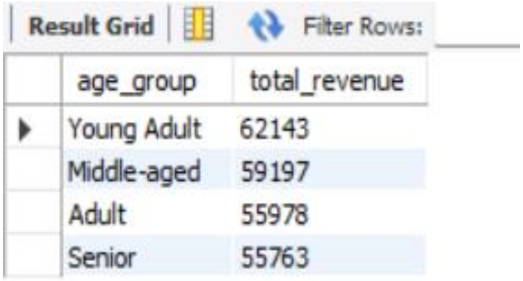
Analyzed whether customers with more than five purchases show a higher likelihood of opting for a subscription, helping assess the link between engagement level and subscription behavior.



subscription_status	repeat_buyers
Yes	958
No	2518

10. Revenue Contribution by Age Group

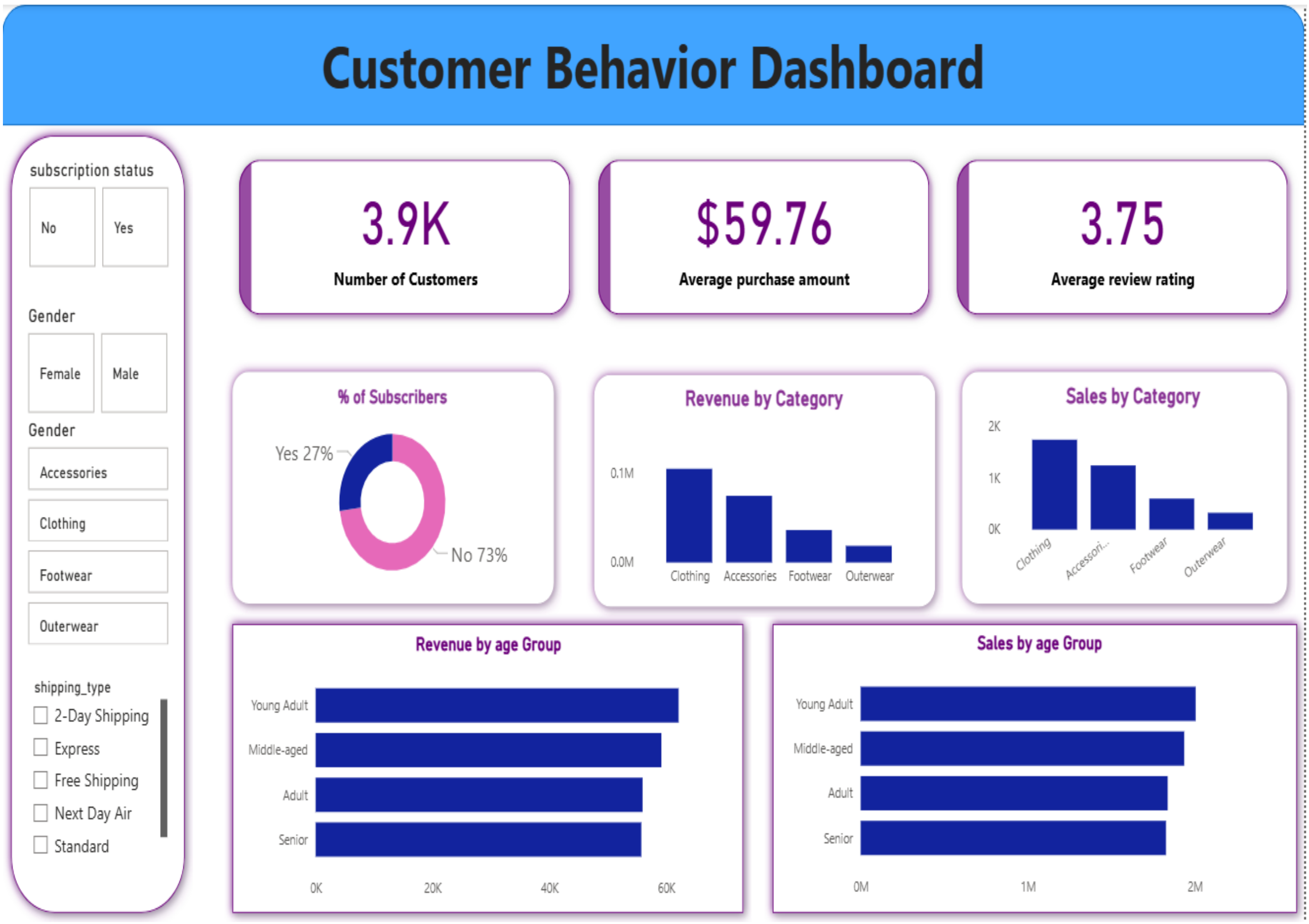
Calculated total revenue generated by each age bracket to determine which age groups contribute most to overall sales.



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

5. Power BI Dashboard

An interactive dashboard was developed in **Power BI** to visually showcase the key insights derived from the analysis. The dashboard highlights customer behavior patterns, spending trends, and product performance through intuitive charts and KPIs.



6. Business Recommendations

- **Increase Subscription Conversions:**
  Promote subscription-only perks and value-added benefits to encourage more customers to enroll.
- **Strengthen Customer Loyalty Initiatives:**
  Implement reward programs for frequent shoppers to help transition them into the “Loyal” customer segment.
- **Optimize Discount Strategy:**
  Review the current discount structure to ensure it drives sales effectively without negatively impacting overall profit margins.
- **Enhance Product Promotion:**
  Feature high-performing and top-rated products prominently in marketing campaigns to maximize visibility and sales.
- **Improve Targeted Marketing:**
  Direct personalized marketing efforts toward high-spending age groups and customers who prefer express shipping, as they show strong purchasing potential.