Customer Behavior Analysis

Project Overview

This project explores and analyzes a retail customer shopping behavior dataset to help a company better understand how customers interact with its products, discounts, and subscription programs

Dataset exploration

Rows: 3,900

Columns: 18

Categorical Columns : gender, item_purchased, category, location , size, color, season , subscription_status , discount_applied, payment_method, frequency_of_purchases , age_group

Numerical Columns : customer_id, purchase_amount,review_rating,previous_purchases,purchase_frequency_days

Missing_values : 37 values in the Review Rating column

1.Data loading : a csv file using python : pandas

Exploration : using.info() and .describe() for summary statistics

```
 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	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3863.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.750065	25.351538
std	1125.977353	15.207589	23.685392	0.716983	14.447125
min	1.000000	18.000000	20.000000	2.500000	1.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000
50%	1950.500000	44.000000	60.000000	3.800000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

```
##Summary statistics of categorical columns
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	Prc C U
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3
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	2
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	I
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	I
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	I
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	I

Data Cleaning :

- Missing values : used median to replace the nulls for each product category in the column Review Rating
- Renamed columns by removing spaces between words and replaced it with"_" for a better handling
- Suppression of the column promo_code_used as it had the same values and the same purpose In this dataset as the column discount_applied
- Feature engineering : 2 columns were added : **age_group** (categorizing customer ages) , and **purchase_frequency_days** from purchase_data

	age	age_group
0	55	middle_aged
1	19	young_adult
2	50	middle_aged
3	21	young_adult
4	45	middle_aged
5	46	middle_aged
6	63	senior
7	27	young_adult
8	26	young_adult
9	57	middle_aged
10	53	middle_aged

	frequency_of_purchases	purchase_frequency_days
0	Fortnightly	14
1	Fortnightly	14
2	Weekly	7
3	Weekly	7
4	Annually	365
5	Weekly	7
6	Quarterly	90
7	Weekly	7
8	Annually	365
9	Quarterly	90
10	Bi-Weekly	14

Database Upload

Connecting Python to SQL Server (SSMS)

Data analysis using sql in ssms

Revenue by Gender : total revenue generated from both gender

	gender	revenue
1	Male	157890
2	Female	75191

High-Spending Discounted : customers that surpassed the average purchase amount despite using discounts

	customer_id	purchase_amount
1	2	64
2	3	73
3	4	90
4	7	85
5	9	97
6	12	68
7	13	72
8	16	81
9	20	90
10	22	62
11	24	88

Top 5 Rated Products : highest reviews

	item_purchased	average_product_rating
1	Gloves	3,86
2	Sandals	3,84
3	Boots	3,82
4	Hat	3,8
5	Skirt	3,78

Shipping Comparison : shipping type comparaison based on purchase amount

	shipping_type	purchase_amount
1	Express	60
2	Standard	58

Subscription Spending : total customers and revenues of subscribed vs non subscribed customers

	subscription_status	total_customers	avg_spend	total_revenue
1	Yes	1053	59	62645
2	No	2847	59	170436

Discount Percentage : top 5 products with the highest % of discounted purchases

	item_purchased	discount_rate
1	Hat	50.00
2	Sneakers	49.66
3	Coat	49.07
4	Sweater	48.17
5	Pants	47.37

Customer Segmentation : segmentation of customers based on purchase amount

Results Messages

	customer_segment	number_of_customers
1	New	83
2	Returning	701
3	loyal	3116

Top Products : top 3 products per each category based on total orders

	item_rank	category	item_purchased	total_orders
1	1	Accessories	Jewelry	171
2	2	Accessories	Belt	161
3	3	Accessories	Sunglasses	161
4	1	Clothing	Blouse	171
5	2	Clothing	Pants	171
6	3	Clothing	Shirt	169
7	1	Footwear	Sandals	160
8	2	Footwear	Shoes	150
9	3	Footwear	Sneakers	145
10	1	Outerwear	Jacket	163
11	2	Outerwear	Coat	161

Repeat Subscribers : most repeated buyers are not subscribed

Results Messages

	subscription_status	repeat_buyers
1	Yes	958
2	No	2518

Revenue by Age group : contribution of each group to revenue generation , young adults being the top contributors

	age_group	total_revenue
1	young_adult	62143
2	middle_aged	59197
3	adult	55978
4	senior	55763

Visualisations : dashboard power bi

Key Insights

- 🛍️ Clothing and Accessories are the most purchased categories.
- 💸 Young adults and middle-aged customers are the main contributors to revenue.
- 🔁 Repeat buyers are more likely to be subscribed customers.
- The majority of the customer base remain unsubscribed.