Customer Behavior Analysis Report

1. Introduction

Understanding customer behavior is critical for improving business performance, customer retention, and revenue growth. This project presents an end-to-end customer behavior analysis using a dataset containing **3,900 customer records and 18 variables**.

The analysis was carried out in three main stages:

1. Data exploration and cleaning using Python
2. Structured analysis using SQL in a MySQL database
3. Interactive visualization using Power BI

The objective of this report is to provide actionable insights into customer spending patterns, product performance, subscription behavior, discount effectiveness, and demographic trends.

2. Dataset Overview

The dataset includes transactional and demographic information related to customer purchases.

Dataset Summary:

- Total Records: 3,900
- Total Columns: 18

Key Variables:

- Customer ID
- Gender
- Age Group
- Product Category
- Item Purchased
- Purchase Amount
- Discount Applied
- Shipping Type
- Subscription Status
- Previous Purchases

- Review Rating

This dataset allows a comprehensive analysis of both **customer demographics and purchasing behavior**.

3. Data Preparation and Exploratory Data Analysis (Python)

The dataset was initially analyzed in **Python using Jupyter Notebook**.

3.1 Exploratory Data Analysis

EDA was performed to:

- Understand data distributions
- Identify missing and inconsistent values
- Analyze purchase amount trends
- Explore relationships between customer attributes

3.2 Data Cleaning and Feature Engineering

The following steps were applied:

- Missing values were replaced using appropriate statistical methods
- New columns were created to enhance analytical depth
- Data types and formats were standardized

After cleaning, the dataset was ready for database integration and advanced analysis.

4. Database Integration and SQL Analysis (MySQL)

The cleaned dataset was connected from Jupyter Notebook to a **MySQL database**, where structured queries were executed to extract insights.

4.1 Revenue Contribution by Gender

Purpose:
To identify which gender contributes more to total revenue.

gender	revenue
Male	157890
Female	75191

Interpretation:
The output shows variation in revenue contribution across genders. This insight can help businesses design gender-focused marketing campaigns and product strategies.

4.2 High-Value Purchases with Discounts

Purpose:
To identify customers who made purchases above the average purchase amount while discounts were applied.

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

Interpretation:
The results indicate that discounts can still generate high-value purchases, suggesting that strategic discounting can increase revenue without reducing profitability.

4.3 Top 5 Products by Average Customer Rating

Purpose:
To identify the highest-rated products based on customer reviews.

Top 5 Products by Average Rating

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

Interpretation:
These products demonstrate strong customer satisfaction and should be prioritized in marketing campaigns and inventory planning.

4.4 Shipping Type vs Average Purchase Amount

Purpose:
To compare average spending between Standard and Express shipping customers.

Average Purchase Amount by Shipping Type

shipping_type	average_amount
Express	60.48
Standard	58.46

Interpretation:
Customers using Express shipping tend to have a higher average purchase amount, indicating a correlation between faster delivery options and higher spending.

4.5 Subscription Status and Customer Spending

Purpose:
To compare average spend and total revenue between subscribed and non-subscribed customers.

Subscription Status vs Revenue

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

Interpretation:
Subscribed customers generally spend more and generate higher total revenue, highlighting the importance of subscription-based business models.

4.6 Products with the Highest Discount Usage

Purpose:
To identify products with the highest percentage of discounted purchases.

Top Discounted Products

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

Interpretation:
These products are highly price-sensitive, and discount strategies should be optimized carefully to maintain profitability.

4.7 Customer Segmentation Based on Purchase History

Purpose:
To segment customers into New, Returning, and Loyal groups based on previous purchases.

Customer Segmentation Output

customer_segment	Number of Customers
Loyal	3116
Returning	701
New	83

Interpretation:
Loyal customers form a valuable segment and should be targeted with retention-focused strategies such as loyalty rewards and personalized offers.

4.8 Top Products Within Each Category

Purpose:
To identify the top-performing products within each product category.

Top Products per 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

Interpretation:
These products represent the strongest contributors within their categories and should receive priority in inventory and promotions.

4.9 Repeat Buyer Analysis

Purpose:
To evaluate whether subscribed customers are more likely to be repeat buyers.

Repeat Buyers by Subscription Status

subscription_status	repeat_buyers
Yes	958
No	2518

Interpretation:
The results show that subscribed customers are more likely to be repeat buyers, reinforcing the value of customer subscription programs.

4.10 Revenue Contribution by Age Group

Purpose:
To analyze revenue distribution across different age groups.

Revenue by Age Group

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

Interpretation:
Young Adults contribute the highest share of revenue, followed by middle-aged customers. Marketing efforts should prioritize these segments.

5. Power BI Dashboard Overview

The final dataset was connected to **Power BI**, where an interactive dashboard was developed.

Dashboard Features:

- Key performance indicators (Total Customers, Average Purchase Amount, Average Rating)
- Category-wise revenue and purchase analysis

- Subscription and age-group segmentation
- Interactive filters for deeper exploration

6. Key Business Insights

- Clothing is the highest revenue-generating category
- Subscribed customers show higher spending and repeat purchases
- Young Adults are the most valuable customer segment
- Express shipping is associated with higher purchase values
- Discounts influence purchase behavior but require optimization

7. Business Recommendations

1. Increase subscription adoption through exclusive benefits
2. Focus marketing campaigns on high-revenue age groups
3. Strengthen inventory for top-performing categories
4. Use targeted discounts instead of broad discounting
5. Promote faster shipping options to increase order value

8. Conclusion

This project demonstrates a complete data analytics pipeline using Python, SQL, and Power BI. The insights derived from this analysis can help business owners improve decision-making, customer retention, and overall business performance.

Appendix A: SQL Queries

```
Select * from mytable;
select gender,sum(purchase_amount) as revenue
from mytable
group by gender;

select customer_id, purchase_amount
from mytable
where discount_applied = 'Yes' and purchase_amount >= (select avg(purchase_amount) from mytable);

-- Q3: Which are the top 5 products with the highest average rating?

select
    item_purchased,
    round(avg(cast(review_rating as decimal(10,2))), 2) as "Average Product Rating"
from mytable
group by item_purchased
order by avg(review_rating) desc
limit 5;

-- Compare the average purchase amounts betweenstandard express shipping
select
    shipping_type,
    round(avg(purchase_amount), 2) as average_amount
from mytable
where shipping_type in ('Standard', 'Express')
group by shipping_type;
-- Q4. Do subscribed customers spend more? compare average spen and total revenue between sibcscribers and non subcribers.
select subscription_status,
count(customer_id) as total_customers,
round(avg(purchase_amount), 2) as avg_spend,
round(sum(purchase_amount),2) as total_revenue
from mytable
group by subscription_status
order by total_revenue, avg_spend desc;

-- Q5 which 5 products have the highest percentage  of purchase with  discounts applied

select
    item_purchased,
    round(100.0 * sum(case when discount_applied = 'Yes' then 1 else 0 end) / count(*), 2) as discount_rate
from mytable
group by item_purchased
order by discount_rate desc
limit 5;

-- Q6 segment customers into returning and loyal based on the pervious purchses and show the count pf each segment.

with customer_type as (
select customer_id, previous_purchases,
case when previous_purchases = 1 then 'New'
when previous_purchases between 2 and 10 then 'Returning'
else 'Loyal'
end as customer_segment
from mytable
)
select customer_segment, count(*) as "Number of Customers"
from customer_type
group by customer_segment;
```

```
-- 5
with item_counts as (
select category,
item_purchased,
count(customer_id) as total_orders,
ROW_NUMBER() over (partition by category order by count(customer_id) DESC) as item_rank
from mytable
group by category, item_purchased
)

select item_rank, category, item_purchased, total_orders
from item_counts
where item_rank <=3;

-- 7 Are customers are repeat buyyers?

select subscription_status,
count(customer_id) as repeat_buyers
from mytable
where previous_purchases >5
group by subscription_status;

-- 8 What is the revenue contribution of each age group?

select age_group,
sum(purchase_amount) as total_revenue
from mytable
group by age_group
order by total_revenue desc;
```