2025

Retail Customer Analytics and Data Visualization Report

Contents

1. Executive Summary

This project analyzes customer behavior to identify purchasing patterns, high-value customer segments, and overall sales trends. The insights help the client understand how customer requirements vary across the global market. All findings are presented through an interactive Power BI dashboard for improved decision-making and visual clarity.

2. Data Summary

The dataset contains 3,900 transactions with 18 columns, covering four major product categories: Clothing, Accessories, Footwear, and Outerwear.
Key features include:

- **Customer demographics:** Age, Gender, Location, Subscription Status
- **Purchase details:** Item Purchased, Category, Purchase Amount, Season, Size, Color
- **Shopping behavior:** Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type

The dataset contains 37 missing values in the *Review Rating* column, which were handled during preprocessing.

3. Tools and Technologies

The project was developed using **Python, SQL, and Power BI**.

- **Python (Google Colab):** Used for data exploration, cleaning, and preprocessing.
- **PostgreSQL:** The cleaned dataset was stored and queried for deeper analysis using SQL.
- **Power BI:** Used to build interactive dashboards and convert analytical findings into visual representations.

4.Data cleaning and Preparation

- **Imported the dataset using Pandas** to begin the analysis process.
- Used df.head() to view the first few rows and get an initial understanding of the dataset structure

	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	Previous Purchases	Payment Method	Frequency of Purchases
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	Venmo	Fortnightly
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	Cash	Fortnightly
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	Credit Card	Weekly
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	PayPal	Weekly
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	PayPal	Annually

- Checked for missing values using **df.isna()** to identify columns that require cleaning.

	0
Customer ID	0
Age	0
Gender	0
Item Purchased	0
Category	0
Purchase Amount (USD)	0
Location	0
Size	0
Color	0
Season	0
Review Rating	37
Subscription Status	0
Shipping Type	0
Discount Applied	0
Promo Code Used	0
Previous Purchases	0
Payment Method	0
Frequency of Purchases	0

- **Handled missing values:**
  Grouped the dataset by category, calculated the median review rating for each group, and filled the null values using the respective median.

- **Improved readability:**
  Renamed all column names to snake_case for better consistency, readability, and documentation.
- **Feature Engineering:**
  Created an **age_group** column using the qcut method.
  Created a **frequency_of_purchases** column based on the purchase history.

```
#create new column frequency_purchases
freequncy_purchases={
    'Fortnightly':14,
    'Weekly':7,
    'Annually':365,
    'Quarterly':90,
    'Bi-Weekly':14,
    'Monthly':30,
    'Every 3 Months':90
}
df['frequency_purchases']= df['frequency_of_purchases'].map(freequncy_purchases)
```

- **Ensured data consistency:**
  Verified whether promo_code_used and discount_applied contained overlapping or

redundant information. Since both represented similar behavior, the promo_code_used column was dropped.

- **Prepared dataset for SQL analysis:**
  Converted the cleaned dataset into a CSV file, downloaded it, and imported it into PostgreSQL for further querying and analysis.

```
from google.colab import files
df.to_csv("customer.csv",index=False)
files.download("customer.csv")
```

5.Data Analyzing using SQL

- Total Revenue According to Gender (Q1: What is the total revenue generated by male vs. female customers?)

	gender text	revenue numeric
1	Female	75191.00
2	Male	157890.00

- Identify customers who spend more than the average purchase amount (Q2: Which customers used a discount but still spent more than the average purchase amount?)

	customer_id integer	purchase_amount numeric (10,2)
1	2	64.00
2	3	73.00
3	4	90.00
4	7	85.00
5	9	97.00
6	12	68.00
7	13	72.00
8	16	81.00
9	20	90.00
10	22	62.00
11	24	88.00

Total rows: 839 Query complete 00:00:00.115

- Identify the top 5 average review ratings (Q3: Which are the top 5 products with the highest average review rating?)

	item_purchased text	avg numeric
1	Gloves	3.8614285714285714
2	Sandals	3.8443750000000000
3	Boots	3.8187500000000000
4	Hat	3.8012987012987013
5	Skirt	3.7848101265822785

- Understand purchase amount by shipping type (Q4: Compare the average Purchase Amounts between Standard and Express Shipping.)

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

- Compare average spend and total revenue across subscription status (Q5: Do subscribed customers spend more? Compare average spend and total revenue between subscribers and non-subscribers.)

	subscription_status text	total_customer bigint	avg_spend numeric	total_revenue numeric
1	Yes	1053	59.49	62645.00
2	No	2847	59.87	170436.00

- Highest percentage of discount purchases(Q6: Which 5 products have the highest percentage of purchases with discounts applied?)

	item_purchased text	discount_rate numeric
1	Hat	50.00
2	Sneakers	49.00
3	Coat	49.00
4	Sweater	48.00
5	Pants	47.00

- Classified customers into New,Returning, and loyal segments based on the purchase history(Q7: Segment customers into New, Returning, and Loyal based on their total number of previous purchases, and show the count of each segment.)

	customer_segment text	Number of Customers bigint
1	Loyal	3116
2	New	83
3	Returning	701

- List most purchased top 3 products in each category (Q8: What are the top 3 most purchased products within each category?)

	item_rank bigint	category text	item_purchased text	total_orders bigint
1	1	Accessori...	Jewelry	171
2	2	Accessori...	Sunglasses	161
3	3	Accessori...	Belt	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

- Check whether customer with >5 purchases are more likely to subscribe(Q9: Are customers who are repeat buyers (more than 5 previous purchases) also likely to subscribe?)

	subscription_status text	count bigint
1	No	2518
2	Yes	958

- Understand revenue contribution from each age group(Q10: What is the revenue contribution of each age group?)

	age_group text	revenue numeric
1	Young_Adult	62143.00
2	Middle Age	59197.00
3	Adult	55978.00
4	Senior	55763.00

6.Data visualization

Power BI was used to design interactive dashboards that present key insights from the cleaned and processed dataset. The dashboard includes cards, bar charts, donut charts, and slicers for dynamic filtering. These visualizations help the client easily explore customer behavior, purchase patterns, and revenue trends.

7.Business Recommendation

- **Focus marketing and product strategies on the Young Adult segment.**
  This age group shows the highest purchase amount, meaning they respond well to promotions and contribute significantly to revenue.

- **Promote the subscription program more aggressively.**
  Subscribers tend to spend more on average, so increasing subscription adoption can directly boost total revenue.

- **Introduce exclusive rewards for Loyal and Returning customers**
  Examples: early access to sales, loyalty points, birthday discounts.
  This helps maintain engagement and strengthens long-term customer retention.

- **Offer targeted discounts for the Clothing category.**
  Clothing generates the highest revenue, so strategic discounts can attract more new customers while increasing repeat purchases.