

Consumer Growth Analytics

An End-to-End Analysis of Purchase Activity to Drive Strategic Growth

1. Overview

This report details an end-to-end data analysis project initiated to uncover the causal factors driving customer loyalty and sales growth for a leading retail enterprise. The core business challenge was to strategically leverage consumer purchase data to optimize marketing, product, and engagement strategies.

By executing a comprehensive analytical roadmap, this project successfully transformed raw transactional data into actionable intelligence. The process involved:

- 1.1. **Data Engineering (Python):** Cleaning, transforming, and enriching the 3,900-record dataset using Pandas, which included creating new features like age group for segmentation.
- 1.2. **Database & SQL Analysis (PostgreSQL):** Loading the cleaned data into a relational database and executing 17 advanced SQL queries to analyze customer segments, loyalty drivers, and product performance.
- 1.3. **Visualization (Power BI):** Consolidating key findings into an interactive dashboard to monitor KPIs and identify trends.

2. Data Source and Preparation (ETL)

2.1. Dataset Profile

The foundation of this analysis is the `purchase_activity.csv` dataset, a flat file containing 3,900 records and 18 columns. Key attributes include:

- **Customer Demographics:**
`Customer ID, Age, Gender, Location`
- **Transaction Details:**
`Item Purchased, Category, Purchase Amount (USD), Payment Method`
- **Behavioral Data:**
`Season, Review Rating, Subscription Status, Shipping Type`
- **Loyalty Metrics:**
`Discount Applied, Previous Purchases, Frequency of Purchases`

2.2. Data Cleaning and Transformation (Python)

The raw data was loaded into a Pandas DataFrame for initial inspection and cleansing. The `CLV_Analysis.ipynb` notebook details the following key steps:

- **Initial Audit:** Used `.info()` to check structure and `.describe()` for summary statistics.

	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
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900	3900.000000	3900	3900
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2	NaN	6	7
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No	NaN	PayPal	Every 3 Months
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223	NaN	677	584
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN	25.351538	NaN	NaN
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN	14.447125	NaN	NaN
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN	1.000000	NaN	NaN
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN	13.000000	NaN	NaN
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN	25.000000	NaN	NaN
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN	38.000000	NaN	NaN
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN	50.000000	NaN	NaN

- **Missing Data Handling:** Checked for null values and imputed missing values in the `Review Rating` column using the median rating of each product category.
- **Column Standardization:** Renamed columns to snake case for better readability and documentation.

2.3. Feature Engineering:

- Created `age_group` column by binning customer ages.
- Created `purchase_frequency_days` column from purchase data.

2.4. Data Consistency Check: Verified if `discount_applied` and `promo_code_used` were redundant; dropped `promo_code_used`.



2.5. Database Integration: Upon successful cleaning and transformation, the finalized DataFrame was loaded into a PostgreSQL database named `clv_analysis` within a table called `consumer` using the `SQLAlchemy` and `psycopg2` libraries. This critical step provided a robust, relational environment for executing complex and high-performance SQL queries.

3. Strategic Analysis via Advanced SQL

Leveraging the structured `consumer` table in PostgreSQL, 17 distinct SQL queries were executed to answer specific business questions. The findings from `CLV_Analysis.sql` are summarized below.

3.1. Customer & Revenue Segmentation

- **Gender:** Male customers account for the majority of revenue (Q1).

	gender 	revenue 
1	Female	75191
2	Male	157890

- **Discount Users:** A specific segment of customers exists that applies discounts while still spending more than the average purchase amount, indicating an opportunity for targeted promotions (Q2).

	customer_id bigint	purchase_amount bigint
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
Total rows: 839		Query complete 00:00:00.127

- **Subscribers:** Non-subscribers generate significantly more total revenue and spend more on average per transaction (Q5).

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

- **Age Group:** Young Adults contribute the most to total revenue (Q10).

	age_group text	total_revenue numeric
1	Young Adult	62143
2	Middle-aged	59197
3	Adult	55978
4	Senior	55763

3.2. Product & Satisfaction Analysis

- **Top 5 Products by Rating:** 'Gloves', 'Sandals', 'Boots', 'Hat' and 'Skirt' are the top 5 most rated items (Q3).

	item_purchased text	Average Product Rating numeric
1	Gloves	3.86
2	Sandals	3.84
3	Boots	3.82
4	Hat	3.80
5	Skirt	3.78

- **Top 3 Products of Each Category:** (Q8).

	item_rank bigint	category text	item_purchased text	total_orders bigint
1	1	Accessories	Jewelry	171
2	2	Accessories	Sunglasses	161
3	3	Accessories	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

- **Top 5 Discounted Products:** (Q6).

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

- **Targeted Scaling:** The most successful (high-AOV, high-satisfaction) product categories in specific geographic markets (Q12).

	location text	category text	average_order_value numeric	average_rating numeric
1	Arizona	Outerwear	65.14	4.17
2	Georgia	Outerwear	93.00	4.60
3	New Jersey	Outerwear	61.00	4.02
4	Tennessee	Outerwear	68.80	4.18
5	Texas	Outerwear	72.17	4.00

3.3. Loyalty & Behavior Drivers

- **Payment Method:** Modern payment methods are correlated with a higher number of average lifetime purchases, suggesting these users are more digitally engaged and loyal (Q14).

	payment_method text	average_lifetime_purchases numeric	total_customers_using_method bigint
1	PayPal	26	677
2	Debit Card	26	636
3	Credit Card	26	671
4	Venmo	26	634
5	Bank Transfer	25	612
6	Cash	25	670

- **Shipping:** Customers using 'Express' shipping have a slightly higher average purchase amount (\$60.48) than those using 'Standard' (\$58.46). (Q4).

	shipping_type text	round numeric
1	Standard	58.46
2	Express	60.48

- **+Customer Loyalty:** Classified customers into New, Returning, and Loyal segments based on purchase history (Q7).

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

- **Repeat Buyers & Subscriptions:** Checked whether customers with more than 5 purchases are more likely to subscribe (Q9).

	subscription_status text	repeat_buyers bigint
1	No	2518
2	Yes	958

- **Faster Shipping vs. Loyalty:** Does the investment in faster shipping correlate with a measurable increase in loyalty (i.e., faster repeat purchases) for subscribed customers (Q11.)

	shipping_type text	avg_days_to_next_purchase numeric
1	2-Day Shipping	84
2	Store Pickup	86
3	Next Day Air	90
4	Standard	91
5	Free Shipping	93
6	Express	94

- **Most Loyal Customers:** Calculated using previous purchases as a proxy for Lifetime value.

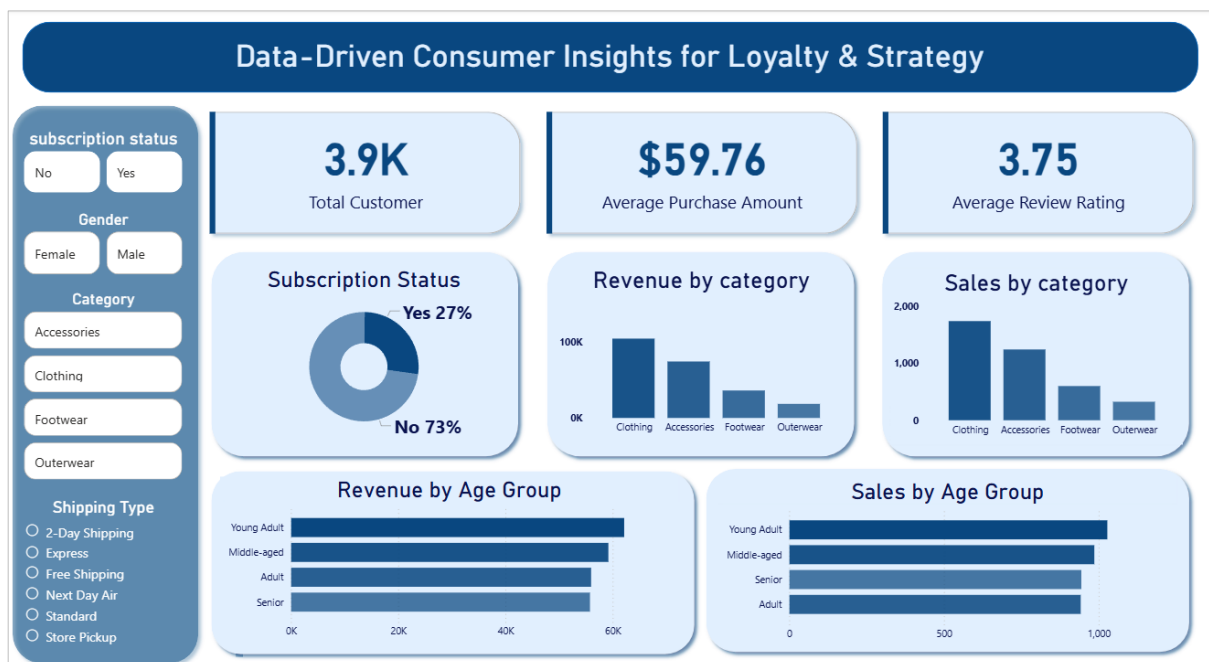
	gender text	payment_method text	median_age numeric	total_top_customers bigint
1	Male	Cash	44	78
2	Male	Credit Card	48	75
3	Male	PayPal	43	74
7	Female	Credit Card	45	36
8	Female	PayPal	40	31
9	Female	Bank Transfer	46	30

4. Interactive Visualization (Power BI Dashboard)

To consolidate these findings and provide stakeholders with a self-service analytics tool, an interactive Power BI dashboard was developed.

4.1. Dashboard Overview

The "Consumer Insights" dashboard provides a centralized, at-a-glance view of key performance indicators (KPIs) and trends, with slicers for filtering by **Subscription Status**, **Gender**, **Category**, and **Shipping Type**.



4.2. Key Performance Indicators (KPIs)

The main metrics highlighted on the dashboard are:

- **Total Revenue:** \$233.081K
- **Average Purchase Amount:** \$59.76
- **Total Customers:** 3.9K
- **Average Review Rating:** 3.75

4.3. Key Value Insights

The dashboard visually confirms and complements the SQL analysis:

- **Top Segments:** Bar charts for 'Revenue by Category' and 'Sales by Category' immediately draw attention to the most valuable category (Clothing).
- **Revenue by Age Group:** A bar chart clearly illustrates that Young Adults contribute the most to total revenue and sales.
- **Subscription by Gender:** A donut chart highlights that only male customers opt for subscription. There is no female subscribed customer.

5. Actionable Business Recommendations

- **Refine Marketing & Segmentation:** Develop targeted marketing campaigns for the highest-value segments.
- **Enhance Subscription & Loyalty Programs:** Re-evaluate the subscription value proposition.
- **Optimize Product & UX:** Investigate product quality and marketing for the lowest-rated product segment, indicating a clear quality or expectation mismatch that needs to be addressed to prevent churn.
- **Promote High-Loyalty Behaviors:** Reward repeat buyers to move them into the “Loyal” segment.
- **Review Discount Policy:** Balance sales boosts with margin control.
- **Product Positioning:** Highlight top-rated and best-selling products in campaigns.
- **Targeted Marketing:** Focus efforts on high-revenue age groups and express-shipping users.

6. Conclusion

This project successfully fulfilled its mandate, transitioning the company's analytics from basic reporting to actionable, strategic insight. By integrating Python for data engineering, SQL for deep analysis, and Power BI for visualization, a holistic view of customer behavior was created, and clear recommendations were delivered.

7. Project Artifacts

For a complete review of the technical analysis, all project files are available in the public GitHub repository. I also recommend connecting on LinkedIn.

- **GitHub Repository:**
www.github.com/arnavbhomia/consumer_growth_analytics
- **Author's LinkedIn:**
www.linkedin.com/in/arnavbhomia