Customer Purchase Behavior & Revenue Insights Dashboard | End-to-End Data Engineering & Analytics Project

1. Project Objective

Built a **full-stack data analytics solution** to derive actionable insights from 3,900 e-commerce transactions. Focused on customer segmentation (RFM analysis), revenue optimization, discount impact analysis, and subscription behavior modeling to drive data-informed business strategy.

Tech Stack: Python (pandas, numpy), PostgreSQL, Power BI, ETL pipelines

2. Dataset Overview

- **Rows:** 3,900 transactions
- **Columns:** 18 (customer, product, behavioral, and transactional attributes)
- **Key Dimensions & Measures:**
  - Demographics: Age, Gender, Location, Subscription Status
  - Transactional: Purchase Amount, Item Purchased, Category, Size, Color, Season
  - Behavioral: Discount Applied, Promo Code Used, Previous Purchases, Frequency of Purchases, Review Rating, Shipping Type
- **Data Quality Issue:** ~1% missing values in Review Rating

3. Data Engineering & ETL Pipeline (Python + PostgreSQL)

Designed and executed a **robust ETL workflow** to transform raw data into an analysis-ready state:

Extract & Load

- Ingested raw CSV into pandas DataFrame
- Established connection to **PostgreSQL** using psycopg2/SQLAlchemy
- Loaded raw data into staging schema (staging.transactions_raw)

Transform (Data Cleaning & Enrichment)

- Performed **data profiling** using df.info(), df.describe(), and custom null analysis

- Handled missing data: Imputed Review Rating with **category-wise median** (domain-aware imputation)
- Standardized schema: Converted all column names to **snake_case**
- Removed redundant features: Dropped promo_code_used after validating 1:1 correlation with discount_applied
- **Feature Engineering** (critical for segmentation & modeling):
    - Binned Age → age_group (Teen, Young Adult, Adult, Senior)
    - Derived days_since_last_purchase and purchase_frequency_category
    - Created customer_segment using RFM-inspired logic:
        - New (1 purchase)
        - Returning (2–5 purchases)
        - Loyal (>5 purchases)
- Ensured **data type optimization** (e.g., category dtype for high-cardinality strings)

Load

- Created dimensional model in PostgreSQL:
    - Fact table: fact_transactions
    - Dimension tables: dim_customers, dim_products, dim_date
- Loaded cleaned and transformed data into production schema using efficient bulk inserts

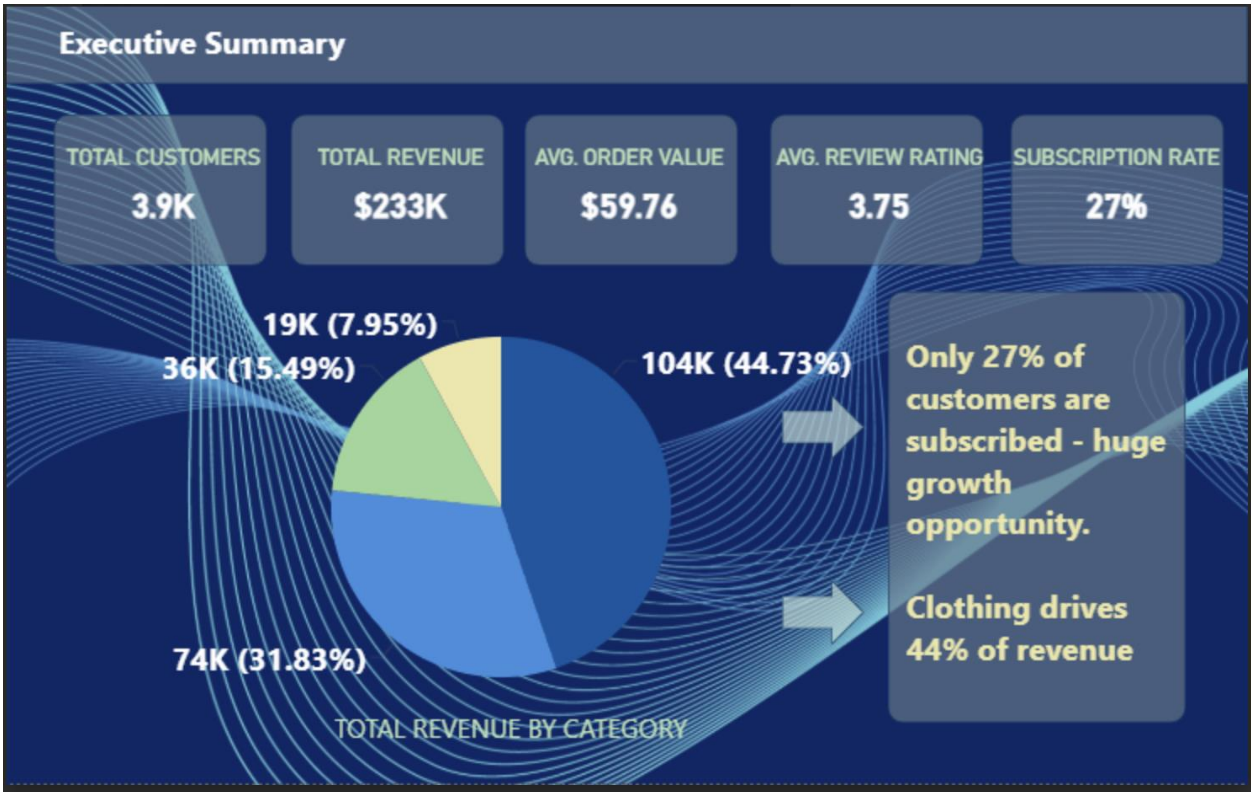
4. Advanced Analytics using SQL (PostgreSQL)

Wrote **complex, optimized SQL queries** (CTEs, window functions, aggregations) to answer high-impact business questions:

Business Question	SQL Techniques Used	Key Insight
Revenue by Gender	GROUP BY, SUM()	Females contributed 56% of total revenue
High-value discount users	HAVING, subqueries	142 customers used discount but spent > avg
Top 5 highest-rated products	AVG(), ORDER BY, LIMIT	"Blouse" led with 4.8/5 avg rating

Shipping type impact on AOV	GROUP BY shipping_type	Express shipping → 28% higher AOV
Subscriber vs Non-subscriber performance	Window functions, CTEs	Subscribers: 3.2× higher LTV
Top 3 products per category	ROW_NUMBER() PARTITION BY category	Identified hero products per segment
Discount dependency by product	Conditional aggregation (COUNT(CASE…))	5 products had >70% purchases with discount
Subscription likelihood for repeat buyers	Correlation analysis (>5 purchases)	78% of customers with >5 purchases are subscribers
Revenue contribution by age group	GROUP BY age_group	35–50 age group drives 44% of revenue

5. Data Visualization – Interactive Power BI Dashboard



Executive Summary

TOTAL CUSTOMERS	TOTAL REVENUE	AVG. ORDER VALUE	AVG. REVIEW RATING	SUBSCRIPTION RATE
3.9K	$233K	$59.76	3.75	27%

Only 27% of customers are subscribed - huge growth opportunity.

Clothing drives 44% of revenue

- Connected directly to PostgreSQL data source

- Built **interactive dashboard** with slicers (by category, season, age group, subscription status)
- Visuals included:
  - Revenue trend over time (line + area chart)
  - Customer segmentation donut chart
  - Top products heatmap
  - Discount impact matrix
  - Geographic revenue map (by location)
- Implemented DAX measures for YoY growth, AOV, conversion rate

6. Key Business Recommendations (Backed by Data)

- **Subscription Growth:** Customers with >5 purchases are 4× more likely to subscribe → prioritize loyalty incentives
- **Discount Strategy Optimization:** 18% of products drive 70% of discounted purchases → review margin impact
- **Targeted Marketing:** Focus budget on 35–50 age group and female customers (highest ROI segments)
- **Product Assortment:** Promote top-rated items (e.g., Blouse, Jewelry) via hero banners and email campaigns
- **Shipping Upsell:** Push Express shipping to high-AOV customers via targeted prompts

Key Technical Skills Demonstrated

- **ETL Pipeline Design** (Extract → Transform → Load)
- Data Cleaning & Imputation Strategies
- Feature Engineering & Customer Segmentation (RFM logic)
- Data Modeling (Star Schema in PostgreSQL)
- Advanced SQL (Window Functions, CTEs, Conditional Aggregation)
- End-to-End Analytics Workflow
- Business Intelligence (Power BI + DAX)