Portfolio Dashboard - Data Scientist

Dea Rahma

August 2025

Exploratory Data Analysis
User & Order Creation

by: Dea Rahma
Data Scientist

1 Jan 2025 - 31 Jul 2025

Initial Information

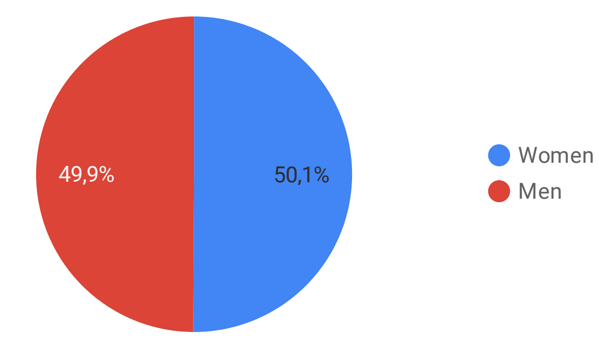
Period
Running Year 2025 (Jan 2025 - Jul 2025)

Data Source: BigQuery Public Dataset [theLook eCommerce]

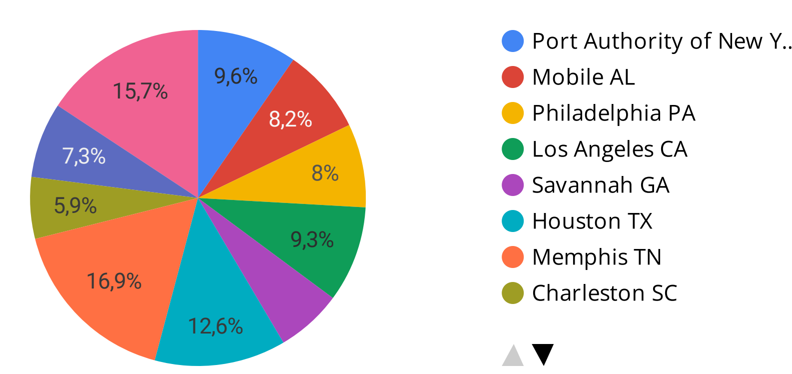
1. The data covers 10 distribution centers across multiple states.
2. There is a rich product catalog with over 10,000 products and nearly 1,900 brands.
3. A healthy customer base of 8,611 users placed over ~13,000 orders in this 7-month period.
4. Total sales for the period reached ~$796K in value, indicating significant commercial activity.
5. The distribution of sales across diverse geographies will provide a strong foundation for inventory forecasting and optimization.

The Distribution Center Name as below:
1 Chicago IL
2 Los Angeles CA
3 Charleston SC
4 New Orleans LA
5 Port Authority of New York
6 Mobile AL
7 Houston TX
8 Memphis TN
9 Philadelphia PA
10 Savannah GA

Count Product per Department



Product Brand per Department



Count Product
8.134

Count Product Brand
1.653

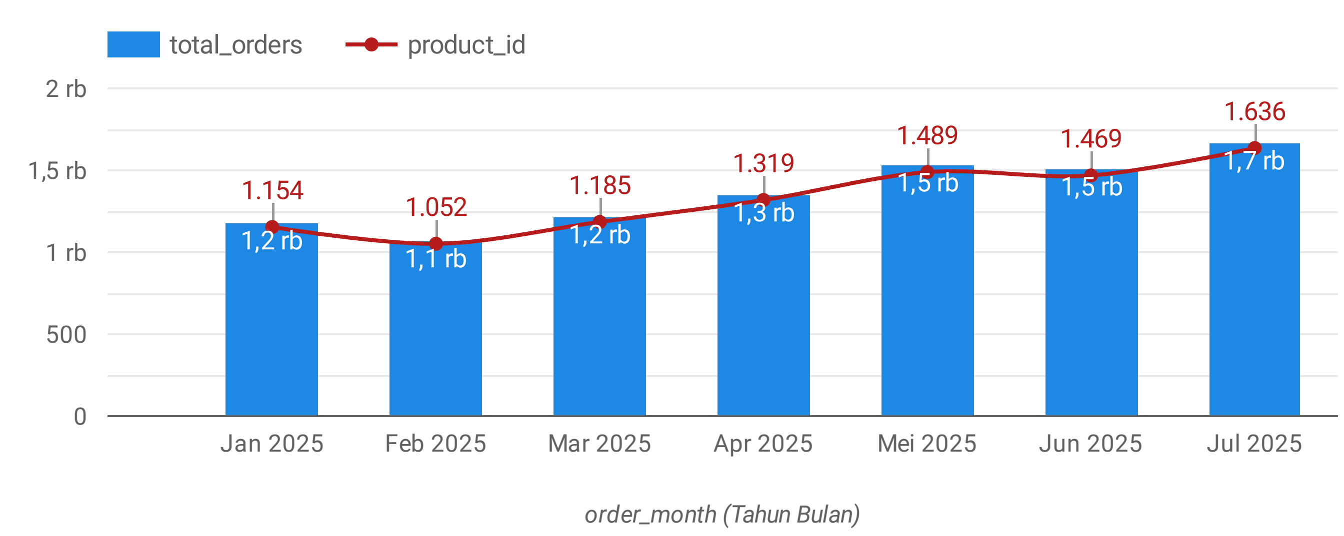
Total Order
9.535

Count User
6.386

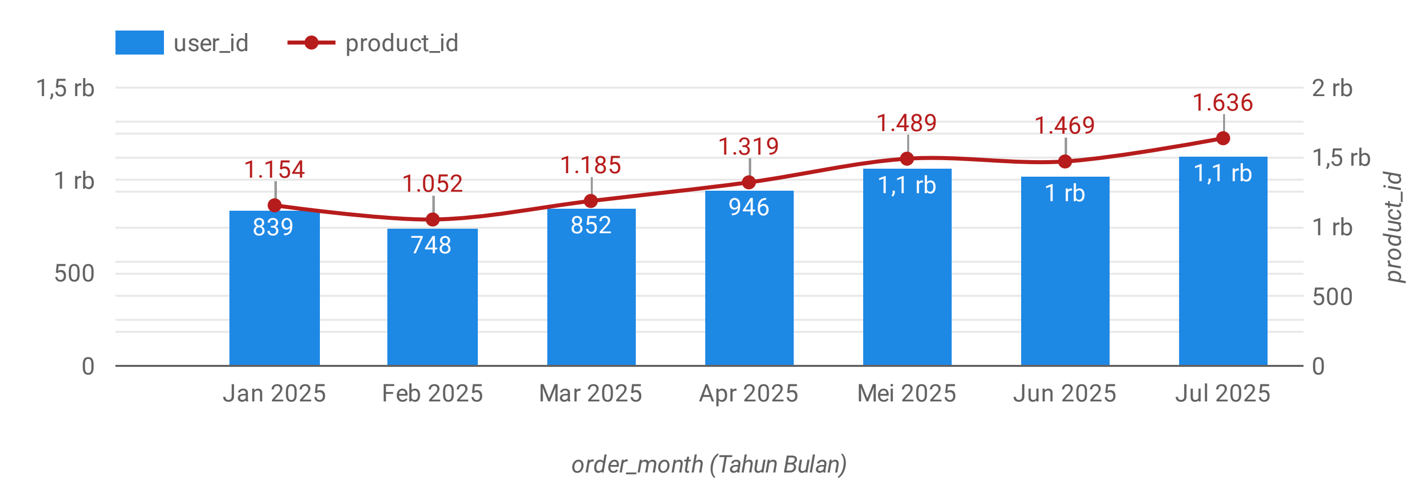
Total Sales
574.813,96

Monthly Sales & Order Trend

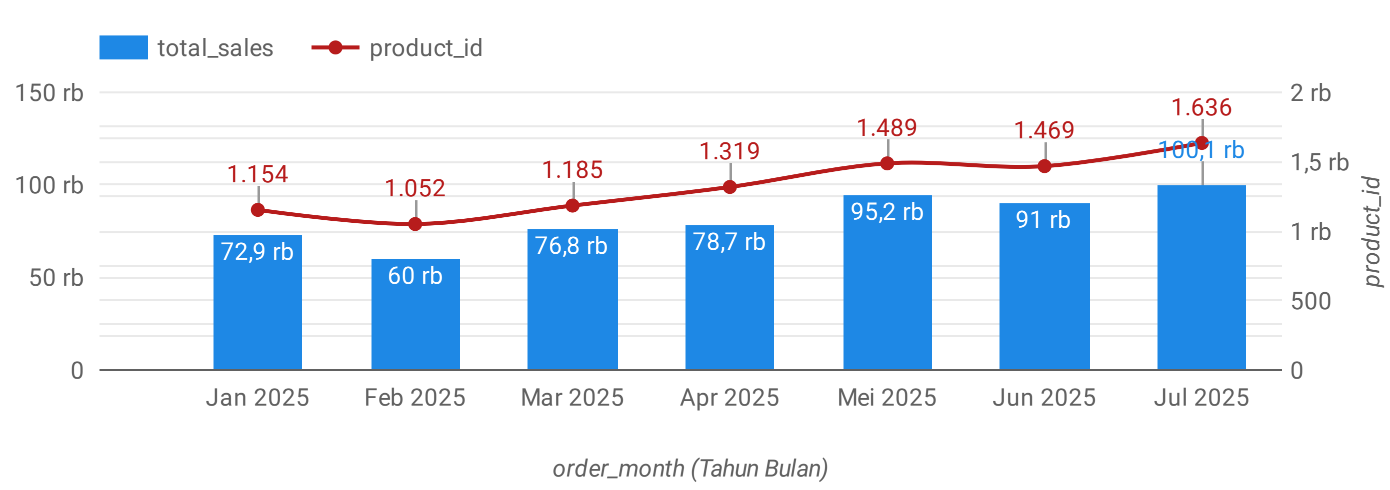
Monthly Product Order



Monthly Product Buyer



Monthly Product Sales



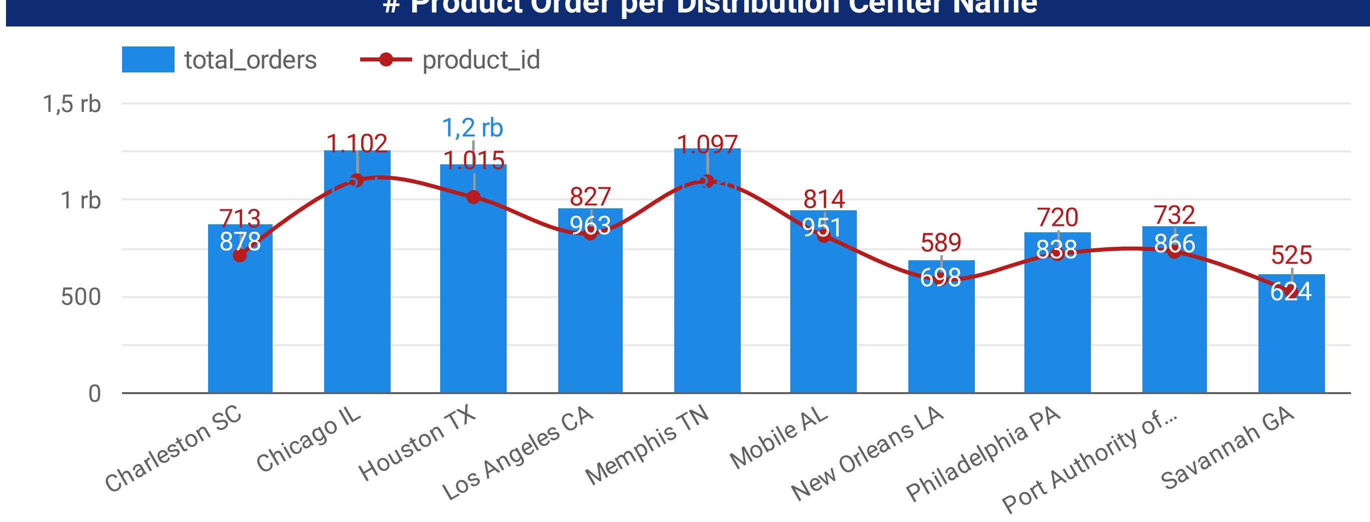
Key Insights from Initial Sales & Orders Trend per Month

1️⃣ Steady Growth in Monthly Performance Both monthly orders, buyers, and sales show consistent growth from January to July 2025.
- Total Orders: Increased from 1.4K to 3.2K
- Total Buyers: Increased from 969 users to 2.1K users
- Total Sales: Increased from $80K+ to $160K+
- This indicates positive market momentum and expanding customer engagement.

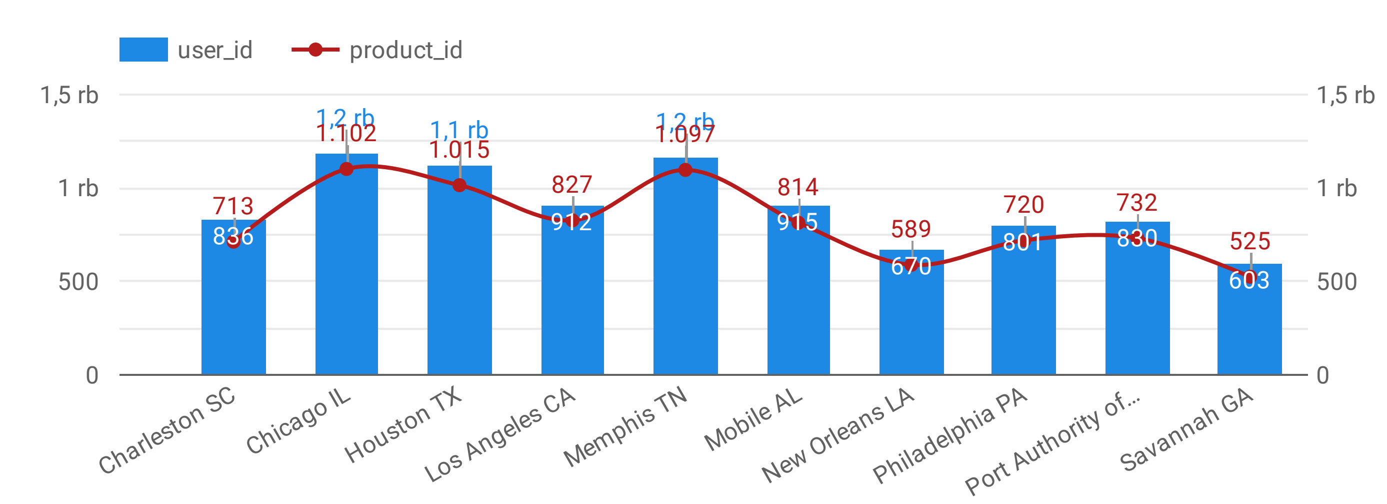
2️⃣ Product Diversity Driving Engagement
- The number of unique products ordered grows in parallel with orders and buyers, signaling:
  - Increased SKU adoption
  - Healthy product variety appeal
- Consistent increase in product ID count indicates customers are exploring more product options month over month.

Sales & Order Trend per Distribution Center

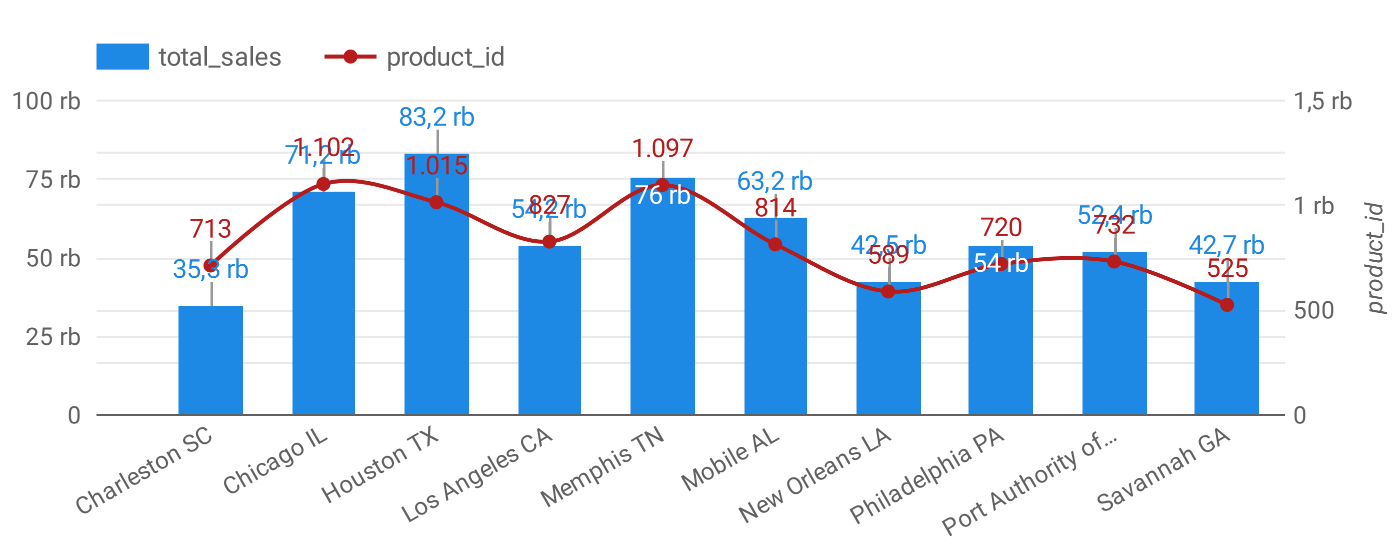
Product Order per Distribution Center Name



Product Buyer per Distribution Center Name



Order & Total Sales per Distribution Center Name



Key Insights from Initial Sales & Orders Trend per DC

1️⃣ Top Performing Distribution
- Centers Chicago IL and Memphis TN consistently lead in:
  - Number of orders
  - Number of buyers
  - Total sales
- Houston TX and Mobile AL also contribute strong volumes.
- Meanwhile, New Orleans LA, Philadelphia PA, and Savannah GA show lower performance — potential areas for growth or marketing push.

2️⃣ Distribution Center Performance Spread
- Clear distribution gap visible:
  - Top 5 DCs (Chicago, Memphis, Houston, Mobile, Los Angeles) account for majority of orders and sales.
  - Bottom 5 DCs have relatively lower contribution — useful for planning stock allocation and marketing focus.

Predictive Data Analysis

Inventory Prediction

by: Dea Rahma
Data Scientist

Project: Inventory Prediction Algorithm

🎯 Goal

Build a machine learning model to predict future inventory needs per product based on historical orders, sales, buyers, and other trends. The goal is to help E-commerce optimize stock levels, reduce overstock & stockouts, and improve operational efficiency.

Why I chose this usecase?

This prediction give **very high Impact** → Directly affects cost, profitability, and customer satisfaction (since out-of-stock = lost sales)

1️⃣ Business Problem

- Current Challenge: The E-commerce has fluctuating monthly orders per product and DC.
- Impact: Over-ordering leads to excess inventory & costs. Under ordering causes stockouts & lost sales.
- Solution: Build a predictive algorithm that estimates next month's needed stock per product.

2️⃣ Data Understanding & Exploration

Input Dataset:

✅ inventory
✅ product
✅ orders
✅ order item
✅ user
✅ dc (distribution center)
✅ events

Target Variable:

- rolling_3m_qty → predicted value
- OR next_month_inventory_needed → derived feature

3️⃣ Feature Engineering

Basic Features:

- product_id
- distribution_center_name
- order_month

Temporal Features:

- Monthly Order Trend (moving average 3 months)
- Monthly Buyer Trend
- Total Sales Trend

Lag Features:

- Order Qty in last 1 month, 2 months, 3 months
- Buyer count lagged

Seasonality Features:

- Month number (1-12)
- Is High Season? (Yes/No based on patterns)

Category Features:

- Product Category (if available)
- Distribution Center

Aggregate Features:

- Product-level total sales trend
- Product-level total buyer trend

4️⃣ Model Building

LightGBM

- Light Gradient Boosting Machine
- A gradient boosting framework that uses tree based learning algorithms.
- Known for being very fast and efficient on large datasets.
- Strength: Good with high-dimensional data, can handle categorical features natively, low memory usage.

XGBoost

- Extreme Gradient Boosting
- Also a gradient boosting algorithm but optimized with regularization techniques.
- Strength: More robust to overfitting, very flexible, can handle missing values automatically.

Why?

- Our task is **to predict next month's stock quantity**.
- The relationship between current features (sales, order quantity, holiday season, price discount, lag variables, etc.) and next month's stock quantity is **non-linear**.
- Both LightGBM & XGBoost handle **non-linear relationships** better than simple linear regression.
- They also perform well even if features are correlated or skewed.

5️⃣ Evaluation

Metrics Error Prediction:

- RMSE:
  - Measures **the square root of the average squared differences** between predicted and actual values.
  - In stock prediction, a large stock shortage or overstock can be **very costly** → so RMSE is very important to catch those big mistakes.
- MAE:
  - Measures **the average of the absolute differences** between predicted and actual values.
  - For warehouse or supply chain planning, MAE tells us on average how many units we miss in our forecast → useful for **safety stock buffer calculation**.
- R^2:
  - Represents the proportion of the variance in the dependent variable that is explained by the independent variables in the model
  - This indicates that **current features only explain a small part of what drives next month's stock**.
- Result:
  - MAE = 3.32 → on average, we're off by ~3 units per SKU/center/month.
  - RMSE = 4.35 → some predictions might be 4-5 units off, especially if the error is skewed by some outliers.
  - R^2 = 14% → Even with LightGBM/XGBoost, **R^2 was low** (~0.11–0.14): the model explains ~11–14% of the variance in stock quantity.

Bin the Target (y column) and Compare Results

- Original target_stock_qty_next_month is continuous → regression task.
- We binned the target into categories:
  - Low → low expected stock next month.
  - Medium → moderate stock.
  - High → high stock needed next month.

Model	Accuracy	Macro F1	Notes
LightGBM	53%	0.50	Slightly better
XGBoost	51%	0.48	Slightly lower

- Classification (Binned Target) Results:
  - Both models are still around 50–53%, indicating that they struggle to clearly distinguish between Low, Medium, and High stock levels.
  - Predicting the target classes remains challenging, possibly because:
    - Current features are not strong enough to fully explain variations in next month's stock needs.
    - There might be external factors missing from the model. would give ~33% by chance).
  - The model can distinguish between Low, Medium, High but still with moderate confidence.
- Business Implication:
  - If the goal is precise stock forecasting (in units) → model needs further improvement (feature engineering, more external data).
  - If the goal is operational stock level planning (Low/Medium/High) → model is a reasonable starting point but needs improvement before going to production.

📍 Final Business Implications & Recommendations

Performance Interpretation:

- The business is currently **growing well** in both volume and buyer engagement.
- Certain DCs are clear leaders, others need focused strategy.

Predictive Model Readiness:

- **Not yet production-ready** for precise stock forecasting.
- Reasonable starting point for **Low / Medium / High stock level planning** → but requires:
  - Better feature engineering.
  - More external data (seasonality, promo calendar, market trends).
  - Possibly temporal modeling (time series approach).

Final Decision:

✅ EDA shows strong and positive market momentum → continue current growth strategies.
⚠️ ML model is an **early prototype** → further iteration needed before operational use in supply chain / inventory planning.

[Link Google Collab](#)