E-Commerce Return Rate Reduction Analysis

Objective

The project aims to identify **why customers return products** and how return rates vary by **category, geography, and payment channel**. The insights are used to build a simple predictive model for return probability and an interactive dashboard for visualization.

Dataset

The analysis uses the **Olist Brazilian E-Commerce Public Dataset**, which contains real-world order data from multiple sellers across Brazil.
Key tables used:

- **Orders** → order-level details, timestamps, return flags
- **Order Items** → product, price, freight, seller info
- **Products** → product categories
- **Customers** → customer location (state)
- **Payments** → payment type and value
- **Reviews** → customer reviews

Data Preparation & Feature Engineering

1. Created a **binary flag** is_returned for canceled/unavailable orders.
2. Parsed timestamps to calculate **delivery delays and shipping times**.
3. Aggregated order items into order-level totals (price, freight, item count).
4. Mapped **product categories, seller states, and customer states**.
5. Added **main payment type** per order.
6. Created features like:
    - days_to_ship
    - days_to_deliver
    - promised_days
    - delivery_delay

Analysis & Insights

1. Overall Return Rate

- The dataset shows an overall return rate of ~ **3–5%** of total orders.

2. Return Rate by Product Category

- Certain categories (e.g., **computers, electronics, office supplies**) had the **highest return %**, suggesting product-specific issues.

3. Return Rate by State

- Geographic variation was observed. Some states had significantly higher return rates, indicating **logistics challenges**.

4. Return Rate by Payment Type

- Orders paid by **credit card** were more common, but boleto (bank slips) showed slightly higher return percentages.

5. Monthly Trends

- Seasonal spikes in returns were observed around **festive/shopping months**, possibly due to bulk orders and returns.

Predictive Modelling (Logistic Regression)

- Features: product category, payment type, seller state, customer state, price, freight, shipping times, delays.
- Model: **Logistic Regression with One-Hot Encoding** for categorical variables.
- Performance:
    - **ROC-AUC:** ~0.70 (moderate predictive power)
    - Insights:
        - Higher delivery delays increase probability of return.
        - Certain product categories are inherently more return-prone.

SQL Queries Used

1. Overall return rate
2. Return % by product category

3. Return % by month
4. Return % by state

Visualizations (Python + Seaborn/Matplotlib)

- **Bar Chart:** Top 10 product categories by return rate
- **Line Chart:** Monthly return rate trend
- **Bar Heatmap:** State-wise return rates

Deliverables

Python Notebook (data cleaning, analysis, logistic regression, SQL queries, visualization)
CSV dataset
Power BI Dashboard with drill-through filters (categories, states, payment type)

Key Recommendations

- **Improve logistics** in high-return states.
- **Tighter quality control** for categories with high return %.
- **Better expectation management** (e.g., delivery timelines, product descriptions).
- **Customer engagement** to reduce avoidable returns.

DASHBOARD