PRODUCT RETURN ANALYSIS USING PYTHON EDA FOR A U.S. ONLINE FASHION RETAILER

1. Background

A U.S.-based online fashion brand with growing national sales was facing an increase in product returns. This affected not just revenue, but also customer satisfaction and inventory planning. Their team had transaction-level data from their Shopify and warehouse systems but lacked the analytical capacity to extract actionable insights from it.

We were engaged to conduct a full exploratory data analysis (EDA) in Python to help the team understand **why customers were returning items**, **which SKUs or categories had abnormal return rates**, and **what operational or customer behaviors were contributing** to these issues.

2. Objective

- To identify return rate patterns based on product attributes, sizing, customer segments, and delivery timelines
- To visualize these patterns in an accessible format using Python
- To support return reduction strategy by providing a data-backed narrative for key stakeholder teams: merchandising, fulfillment, and customer success

3. Data Used

The client provided a **12-month dataset** with **145,000 order records** across their women's and unisex clothing lines. Key fields included:

- Order_ID
- Product_ID
- Category (Dresses, Tops, Bottoms, etc.)
- Size (XS–XXL)
- Returned (1/0)
- Days_To_Delivery
- ZIP_Code
- Customer_Age

- First_Time_Buyer (Yes/No)
- Reason_For_Return (Free text, semi-structured)
- Order_Date, Return_Date

4. Methodology

4.1 Data Cleaning and Preprocessing

- Handled missing and inconsistent entries in Reason_For_Return and Size columns
- Standardized Category and Size fields to resolve inconsistent naming conventions
- Extracted derived fields such as **Return Lag (days)** and **Order Month**
- Aggregated returns at the **SKU** and **ZIP code** level for geo-analysis

4.2 EDA and Visualization

- Univariate analysis of return rate across Category, Size, and Customer Age
- Bivariate analysis of Return Rate vs Days_To_Delivery and Size vs Category
- Used:
    - **Heatmaps** to identify high-return product-size combinations
    - **Boxplots** for delivery times across return status
    - **Bar charts** for return reasons and category-level return share
    - **Geographic return clustering** using ZIP codes (via plotly)

4.3 Tools Used

- pandas, numpy for manipulation
- matplotlib, seaborn, plotly for visualization
- nltk for light text preprocessing on return reason field
- Report compiled in APA format using Python-generated visuals

5. Key Results

- Overall return rate: **28.6%**, with Dresses and Bottoms contributing disproportionately
- **Top three findings**:

1. **Size M in Dresses had the highest absolute return volume**, driven by inconsistent fit remarks in return notes
2. Orders with **delivery time > 5 days** had a **12% higher return rate**, indicating potential mismatch in expectation vs receipt
3. **First-time buyers aged 18–25** had a 34% return rate, significantly above average

- Return notes analysis showed that **45% of free-text entries referenced fit or sizing issues**, especially in international shipments
- Certain ZIP code clusters had repeat returners with above-average behavior — shared with fraud prevention team

6. Report Output

- **PDF Report (15 pages)**:
  - Visual summaries (heatmaps, bar plots, customer return profiles)
  - Section-by-section interpretation of trends and what they mean
  - Category-wise return breakdown with callout boxes for high-return SKUs
  - Summary tables formatted in APA style
- **Jupyter Notebook (.ipynb)** with:
  - Full reproducible code
  - Annotated output cells
  - Reusable functions for return rate computation per field
- **CSV Output**:
  - Product_ID-level return stats
  - ZIP-level return risk score for mapping and segmentation

7. Business Impact

- **Return prediction logic** derived from EDA is being tested in the client's pre-order verification stage
- The merchandising team adjusted **size guide displays and description text** for three high-return categories

- Customer service team used findings to improve **FAQ and chatbot scripting** around fit and delivery expectations
- Estimated **10–12% reduction in preventable returns** in Q3 following the rollout of changes based on the analysis
- Fraud prevention flagged high-risk ZIP clusters for policy adjustment

8. Future Scope

- Train a **classification model using EDA results as input features** to predict return risk for new orders
- Extend analysis to include marketing source data to understand return behavior by campaign
- Automate EDA refresh every 30 days to update heatmaps and dashboards
- Deploy a real-time size recommendation widget powered by historical return rates and body-type feedback