📦 Project Summary: E-commerce Return Rate Reduction Analysis

🎯 Objective:

To identify why customers return products and how return behavior varies across product categories, customer segments, regions, and marketing channels. The goal is to reduce unnecessary returns and improve profitability.

🛠️ Tools Used:

- **Python (Google Colab):** Data cleaning, EDA, modeling
- **Power BI:** Dashboard creation, insights visualization

🧰 Technologies Used:

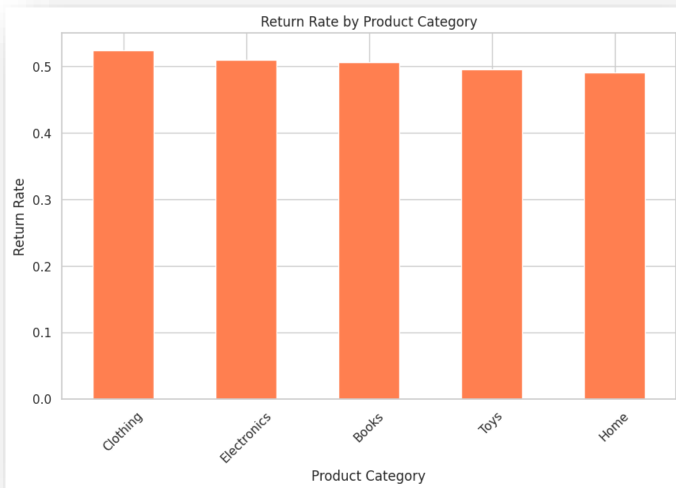
- **Jupyter Notebook**
- **Pandas** for data manipulation 🐼
- **Matplotlib & Seaborn** for visualization 📊
- **NumPy** for numerical operations ➗

🛠️ Step-by-Step Approach

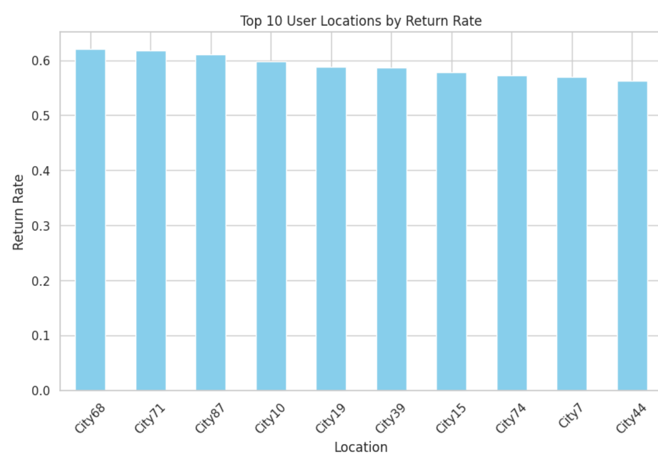
1. 📁 Data Preparation
   - Cleaned the raw order and return datasets
   - Handled missing values and standardized formats for seamless analysis
2. 📊 Exploratory Data Analysis (EDA)
   - Calculated overall return rate
   - Analyzed return patterns:
     - By product category
     - By user location (geography)

- By gender and age group
- By shipping method and applied discount

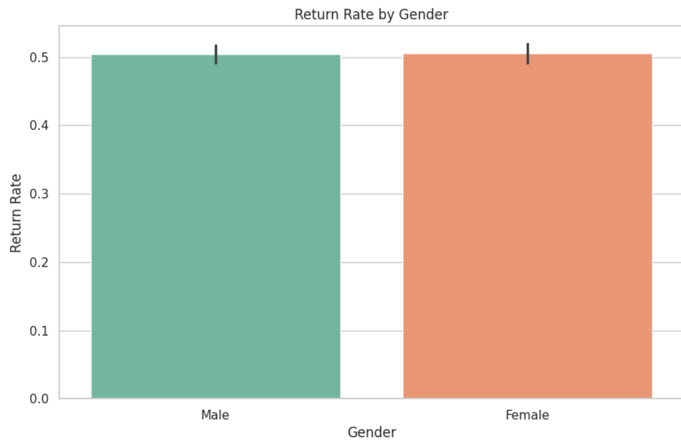
- Analyzed return rates by:
  - **Product Category** (Clothing = highest)



  - **User Location** (City43)



- **Gender** (Females marginally higher)

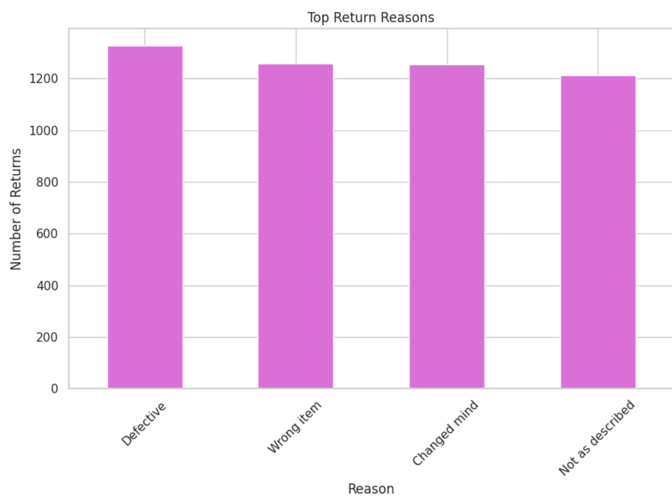


- **Age Group** (50–65)

- **Shipping Method** (Next Day)

- **Discounts** (Higher discounts → more returns)

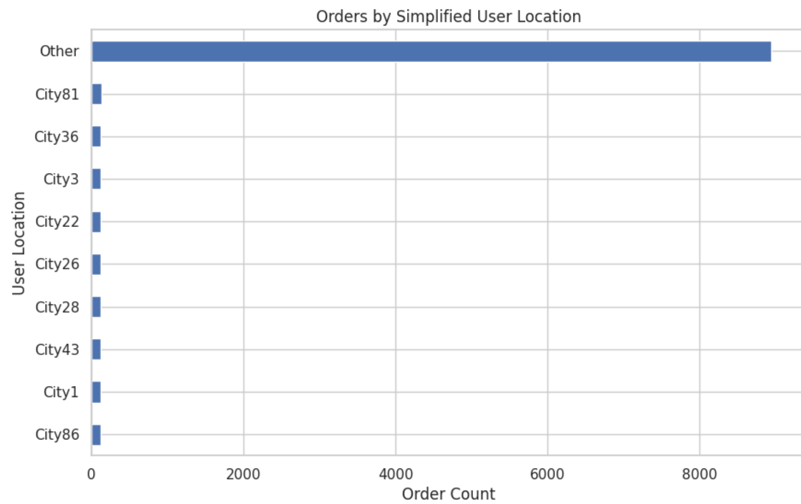
- **Return Reason** (Top = Defective)



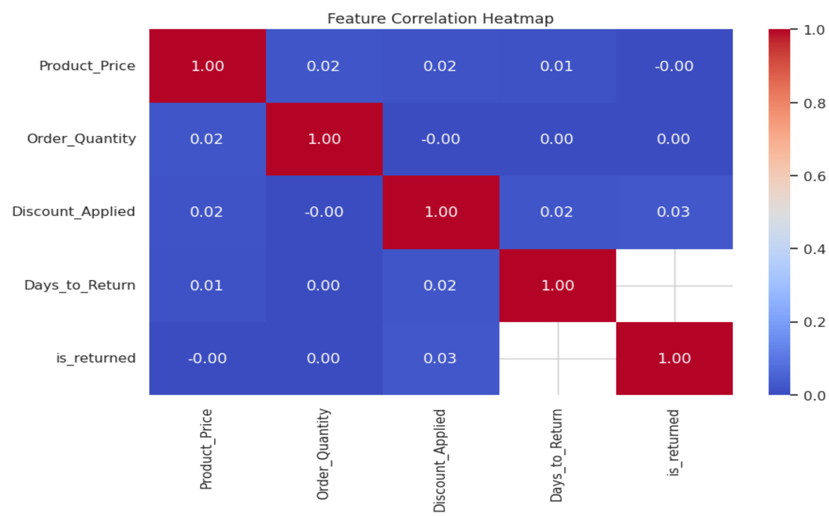
- Visualized return reason frequencies using bar plots and count charts

3. 🔢 Feature Engineering

- Encoded categorical variables (e.g., product category, user location)



- Selected relevant features for the return prediction model



4. 🤖 Predictive Modeling (Logistic Regression)
   - Split the data into training and testing sets
   - Built preprocessing pipelines for both categorical and numerical features
   - Trained a logistic regression model to predict the probability of return
   - Evaluated model performance using metrics like accuracy and AUC-ROC
5. 🧠 Risk Scoring & Interpretation
   - Calculated return risk scores for all records
   - Flagged products with a risk score > 0.8 as high return risk

- Set a custom threshold to better reflect business risk appetite
- Visualized risk segments using Matplotlib and grouped rare categories

count	10000.000000
mean	0.507649
std	0.059730
min	0.319937
25%	0.468770
50%	0.505467
75%	0.542714
max	0.690058

	Feature	Coefficient
73	User_Location_City71	0.621069
58	User_Location_City58	0.579577
90	User_Location_City87	0.554150
5	User_Location_City10	0.554125
22	User_Location_City25	-0.511963
37	User_Location_City39	0.504890
11	User_Location_City15	0.442745
45	User_Location_City46	-0.421284
69	User_Location_City68	0.412862
99	User_Location_City95	0.372000

6. 📊 Power BI Dashboard Development

- Imported cleaned dataset + high-risk product CSV
- Created multi-page dashboard:
  - **Page 1:** RETURNS & REFUND INTELLIGENCE DASHBOARD

- **Page 2:** BEHIND THE RETURNS: Trends, Risks & Actions

- Added advanced features:
  - Slicers for filters (Category, Location, Age, Gender)
  - Drill-through to product details
  - Report page tooltips for deeper context
  - Conditional formatting for risk scores
- Created dynamic text cards with insights (e.g., "Females contribute 60% of returns")

📤 Deliverables

- ✅ Cleaned dataset (cleaned_ecommerce_returns.csv)
- ✅ Python codebase (Colab notebook: data prep, EDA, modeling)
- ✅ Exported high-risk product CSV (return_risk_score > 0.6)
- ✅ Interactive Power BI dashboard (.pbix file) with:
  - Return insights

- Drill-through
- Risk score visualization
- Tooltip-enhanced KPIs

🧠 **Final Insight:**

Return rates are most influenced by product category (Clothing), city-specific behavior (City43), and marketing conditions like discounts and shipping speed. Prioritizing moderate-risk products (score > 0.6) can help reduce refunds.