Retail Sales Analysis: EDA + Regression + Time Series

📌 Project Overview

This project analyzes a real-world e-commerce retail dataset to uncover trends in sales, customers, and product performance. Using a combination of Python (Pandas, Matplotlib, Scikit-learn) and advanced data wrangling techniques, the study extracts insights such as revenue trends, top contributors, and seasonality.

🧰 Tools & Libraries

- **Pandas** – Data wrangling & aggregation
- **Matplotlib / Seaborn** – Visualization
- **Scikit-learn** – Regression modeling
- **Jupyter Notebook** – Data exploration environment

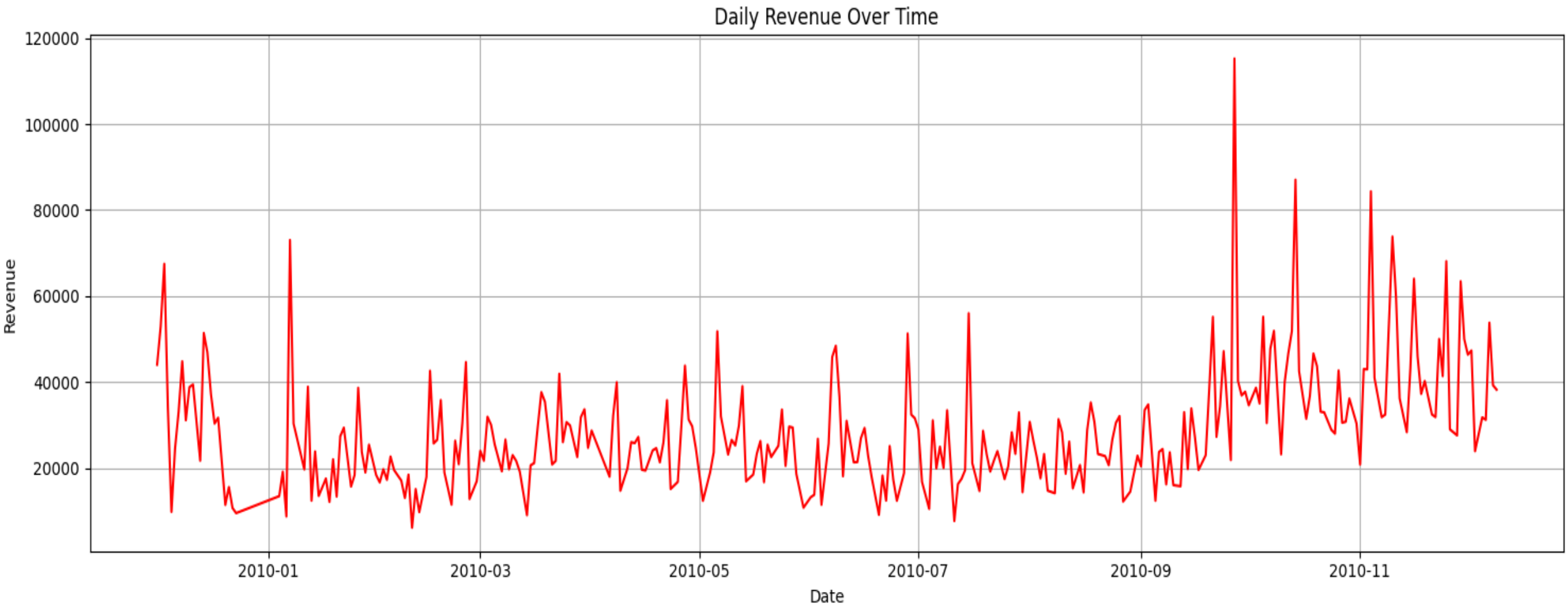
🔍 Analysis Breakdown

✅ 1. Data Cleaning & Preparation

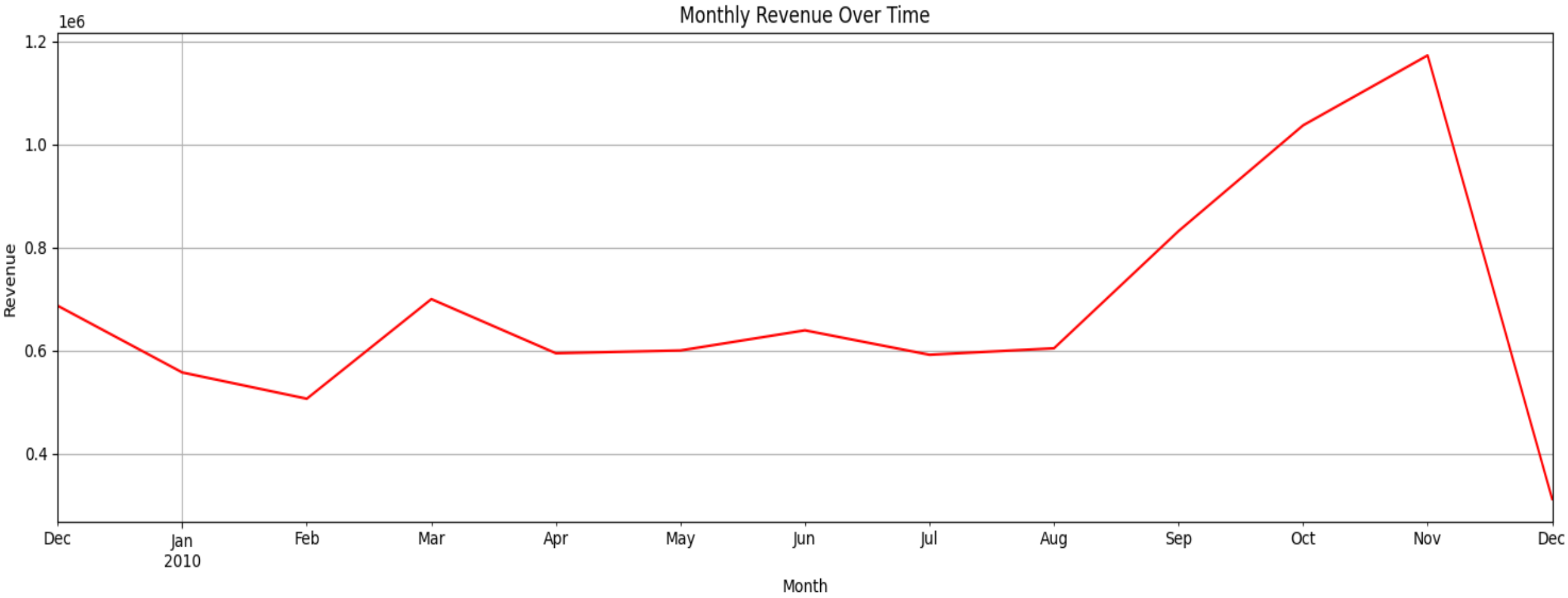
- Removed nulls from Description and Customer ID
- Filtered out invalid sales (zero/negative Quantity or Price)
- Created Revenue = Quantity × Price

📊 2. Revenue Trend Analysis

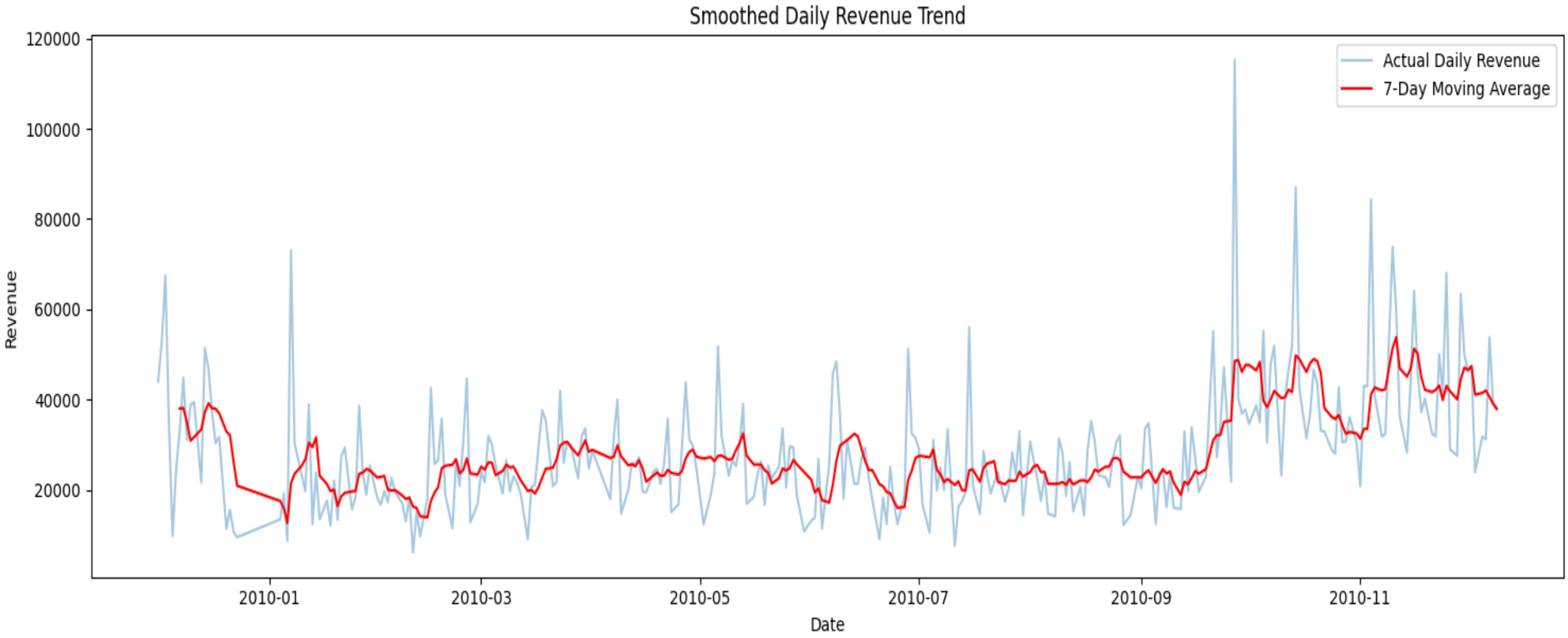
- **Daily Revenue**: Visualized fluctuations in day-to-day performance



- **Monthly Revenue**: Grouped by month to reveal high-performing periods

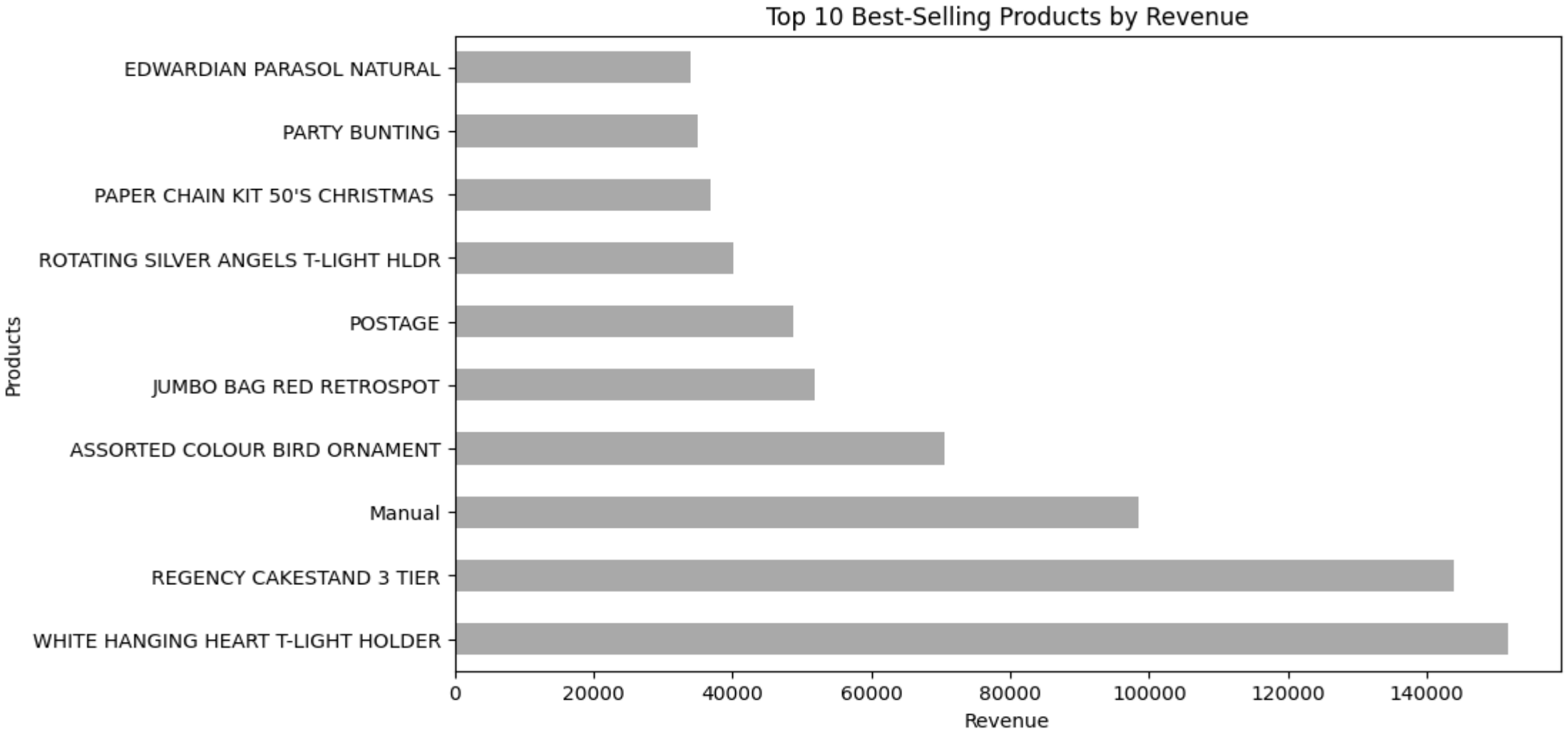


- **Smoothed Daily Revenue**: Applied 7-day rolling average to smooth noise

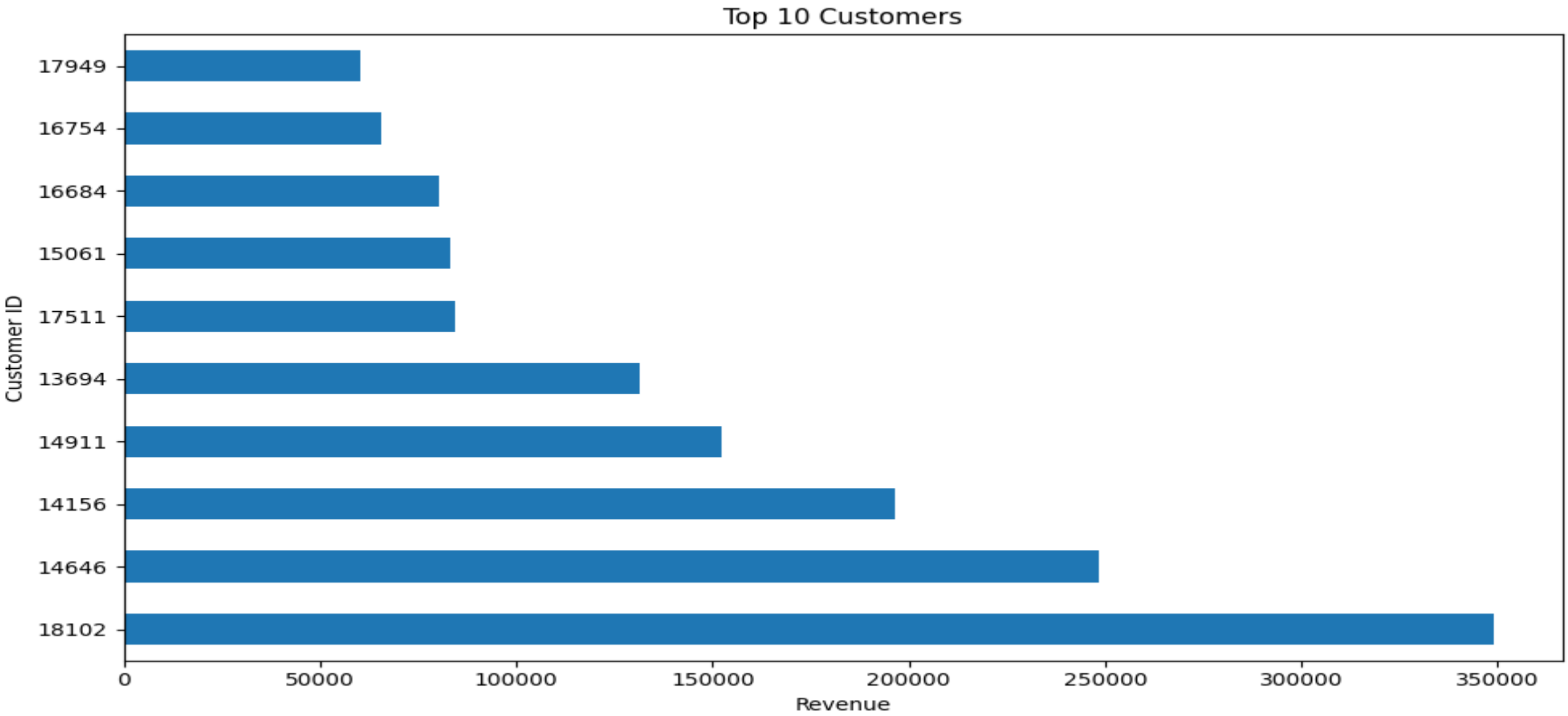


🛒 3. Product & Customer Insights

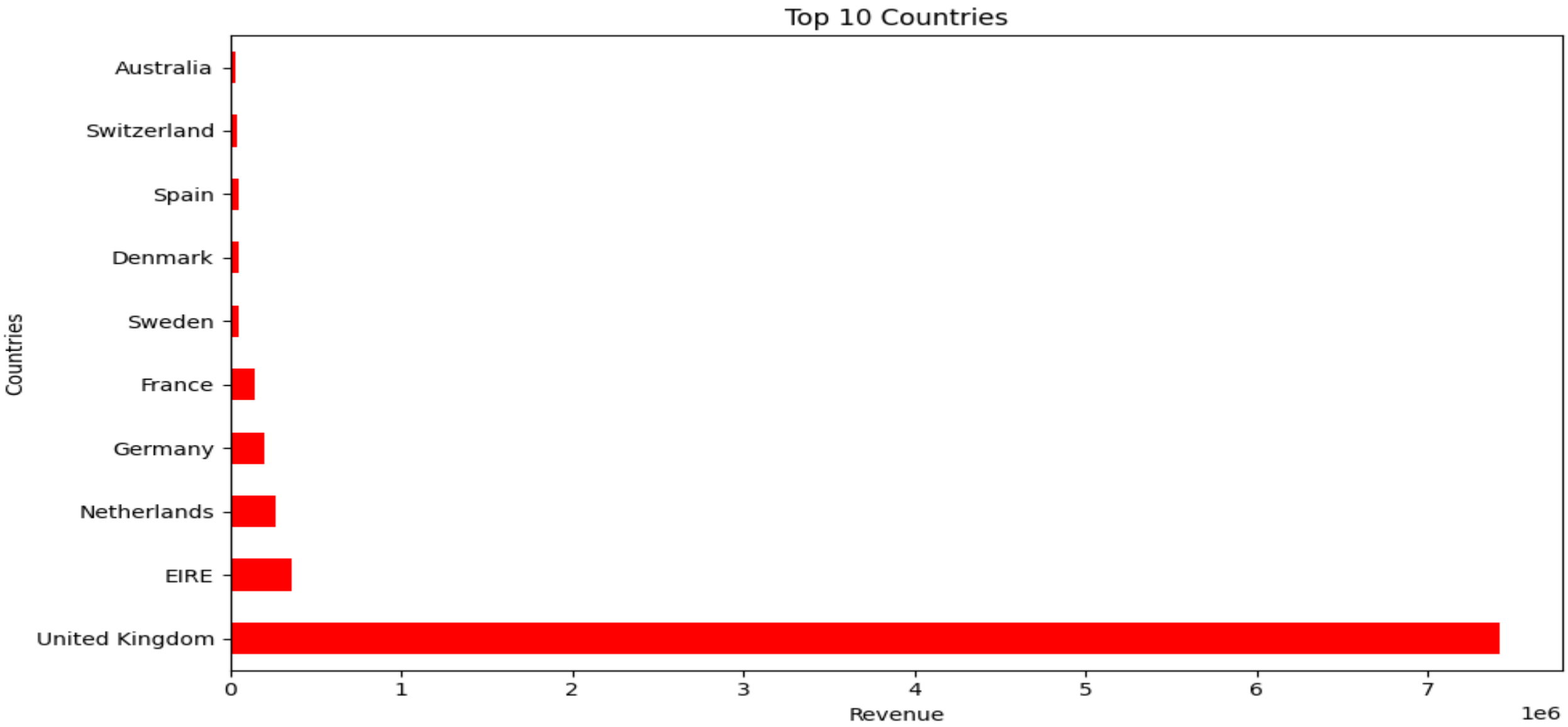
- **Top 10 Products**: Highest revenue-generating SKUs



- **Top Customers**: ID-wise ranking by total spend



- **Top Countries**: Global revenue distribution by region



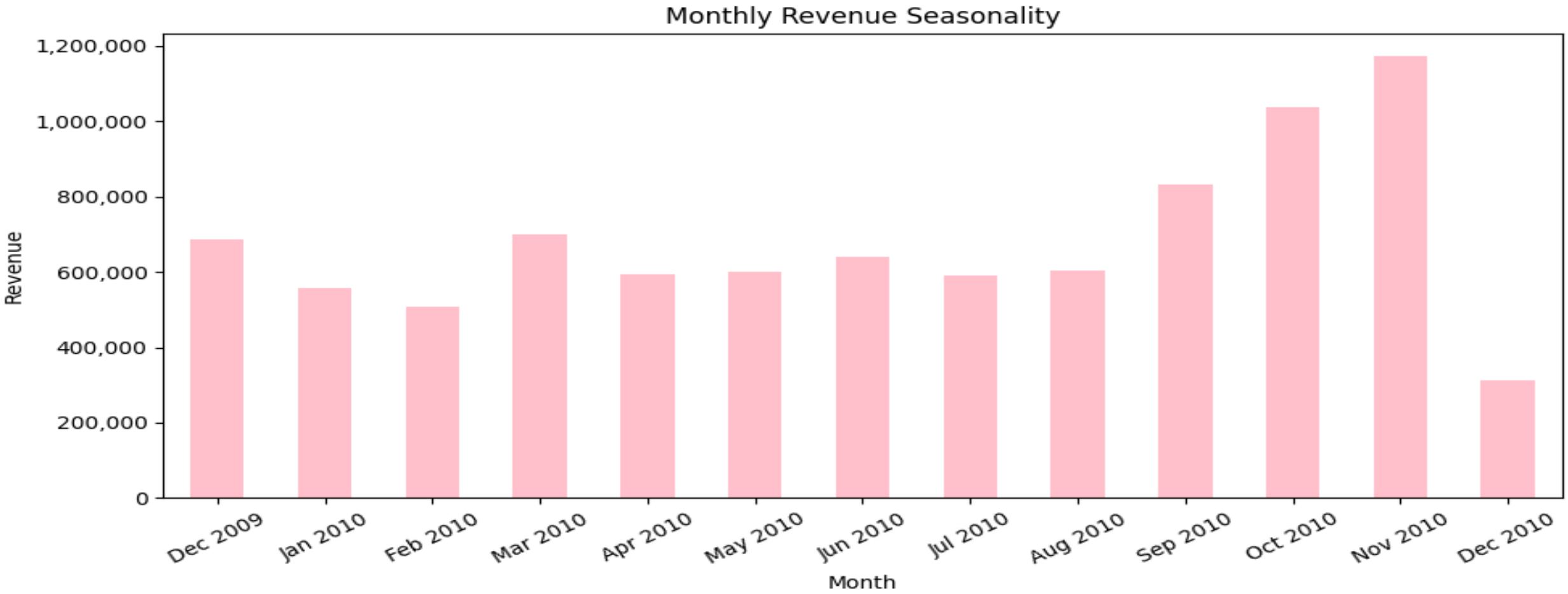
📈 4. Predictive Modeling (Regression)

- Encoded top 10 products and countries using One-Hot Encoding
- Features: Price, Month, Product, Country
- Split into training and test sets
- Achieved:
  - **R^2 Score**: 0.1853
  - **MSE**: 4091.54

🧪 Note: While R^2 was modest, the model helps understand feature contributions.

- 🗓️ **5. Seasonality Detection**

- Used monthly revenue bar chart with formatted currency and rotated labels
- Found spikes in sales during year-end months (likely holiday season)



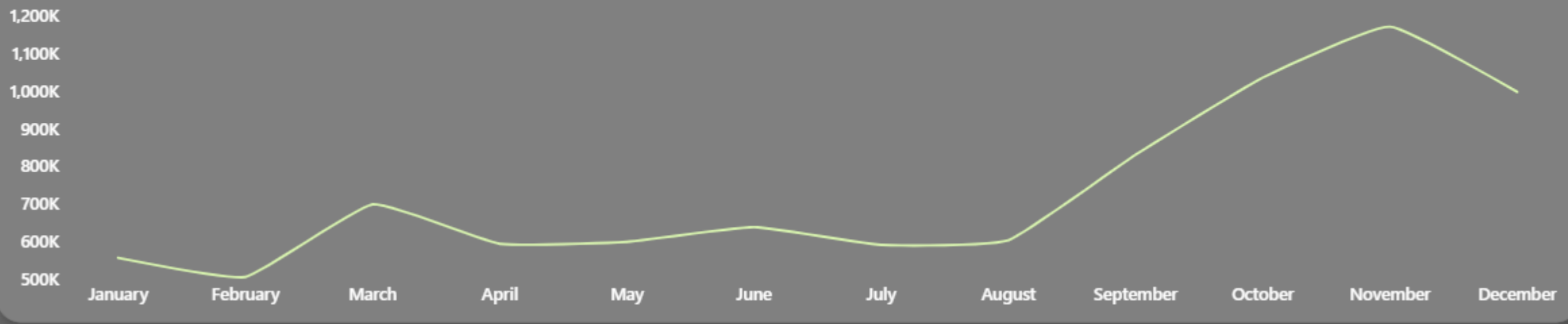
Power BI Dashboard

8.83M

Total Revenue

Daily Revenue Trend

Monthly Revenue Trend

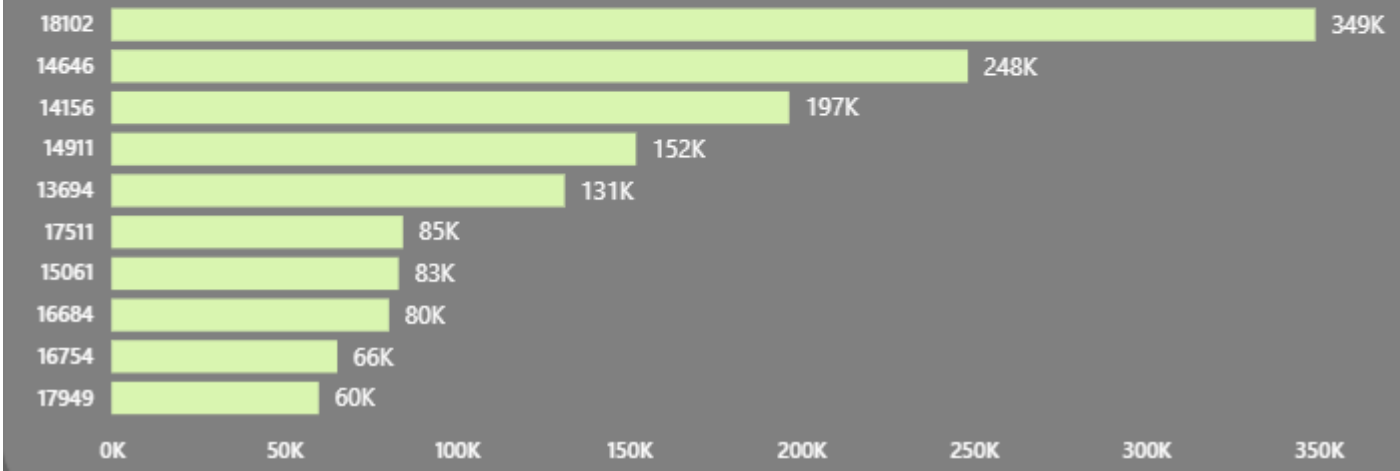


407.66K
Customers

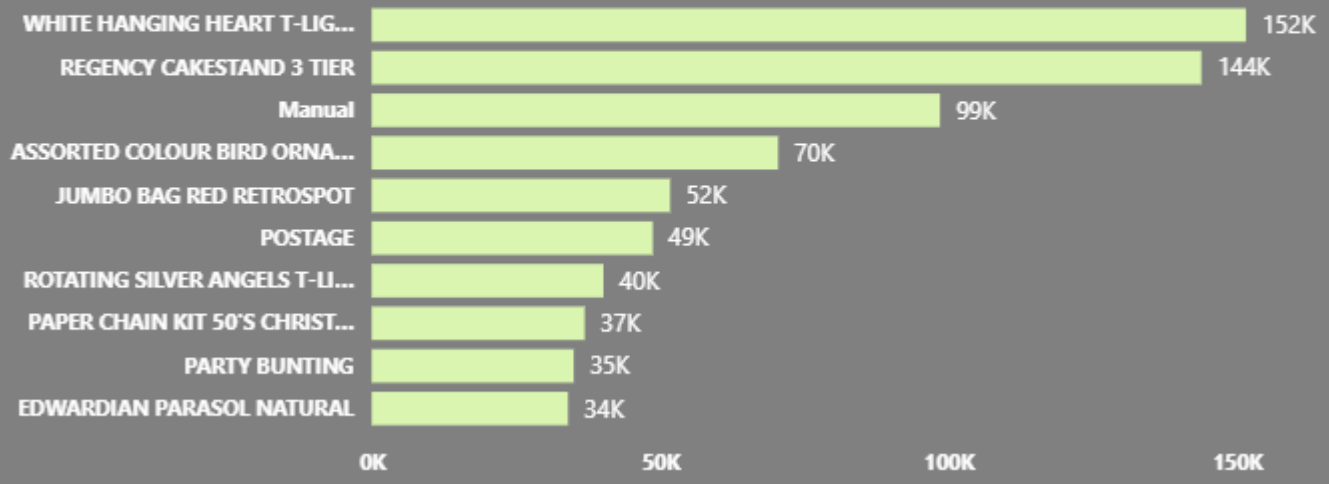
4430
Products

37
Countries

Top 10 Customers



Top 10 Products



Top 10 Countries

📁 Output Artifacts

🖼️ Power BI Dashboard (offline, screenshot exported)

📈 Matplotlib Charts (Python-generated insights)

🗒️ Cleaned Dataset: cleaned_online_sales.csv

📘 Full Python notebook with EDA + modeling

💡 Final Takeaways

- Data cleaning is crucial before any modeling
- Smoothed trends reveal business cycles more clearly than raw numbers
- Revenue is driven by a handful of top customers and products
- Seasonality is visible in monthly aggregates — critical for planning