Retail Sales and Customer Behavior Analytics

MSDS-632-M51: Big Data
University of the Cumberlands
June 15, 2025

- Sushila Sipai
- Satyanarayana Reddy Muttana
- Yanjie Liu

Introduction

Retailers today collect vast amounts of data from POS systems, online platforms, and customer interactions. This project focuses on using Apache Spark, Pandas, and Matplotlib to analyze a real-world retail dataset for insights into product performance and customer behavior.

Preview

Analyzed 500,000+ retail transactions (UCI Online Retail dataset)

Objectives and Goals: : Clean, transform, and analyze retail transaction data, understand customer behavior, sales trends, seasonal patterns

Technologies and Tools: Apache Spark, PySpark, Pandas, Matplotlib, SQLite, Google Colab

Big Data Characteristics – 5Vs

Volume: 500k records – simulates mid-size retail platform

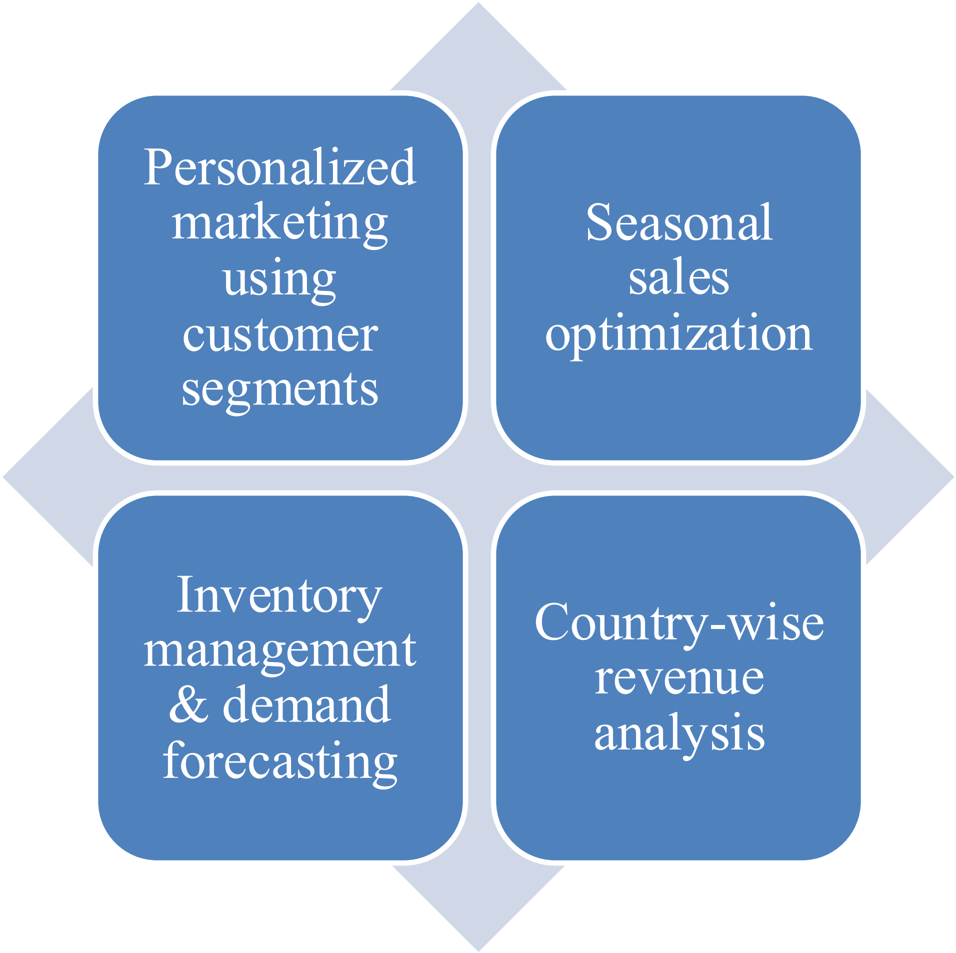
Velocity: Mimics real-time transaction flow

Variety: Structured fields (e.g., InvoiceNo, Date, Price)

Veracity: Data cleaning needed (nulls, returns)

Value: Insight into customer segments and sales drivers

Business Goals & Drivers



Personalized marketing using customer segments

Seasonal sales optimization

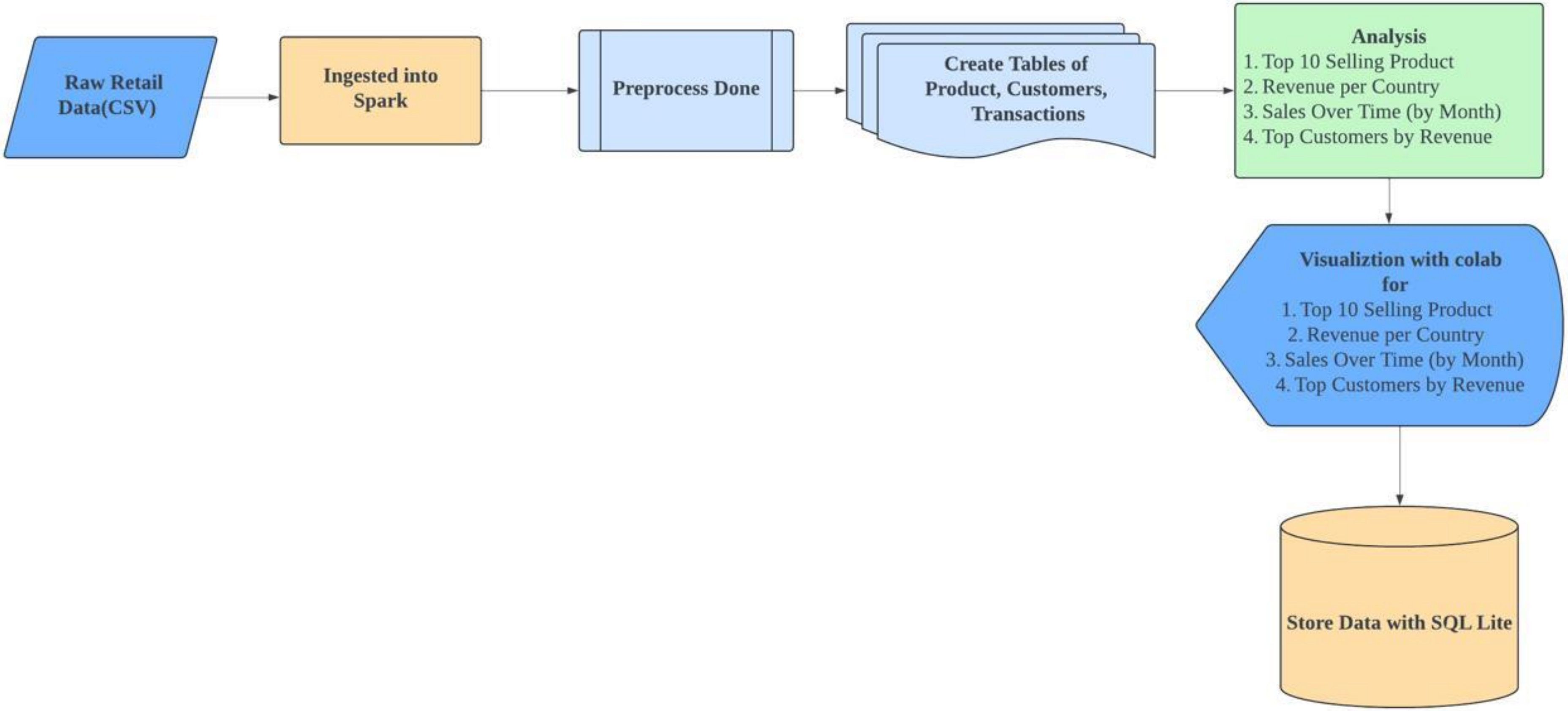
Inventory management & demand forecasting

Country-wise revenue analysis

Technical Architecture

- Foundation: Java 11, Spark 3.5.6, Winutils
- Processing: Apache Spark (local mode), PySpark
- Extension: Pandas, Matplotlib, Google Colab, SQLite
- Seamless Spark–Python interoperability

Architecture Diagram



Raw Retail Data(CSV)

Ingested into Spark

Preprocess Done

Create Tables of Product, Customers, Transactions

Analysis
1. Top 10 Selling Product
2. Revenue per Country
3. Sales Over Time (by Month)
4. Top Customers by Revenue

Visualiztion with colab for
1. Top 10 Selling Product
2. Revenue per Country
3. Sales Over Time (by Month)
4. Top Customers by Revenue

Store Data with SQL Lite

Big Data Lifecycle

1. Business Case Evaluation

2. Data Identification (UCI Retail CSV)

3. Acquisition & Filtering (cleaning nulls, malformed entries)

4. Data Transformation (revenue metrics)

5. Analysis (top products/customers, time series)

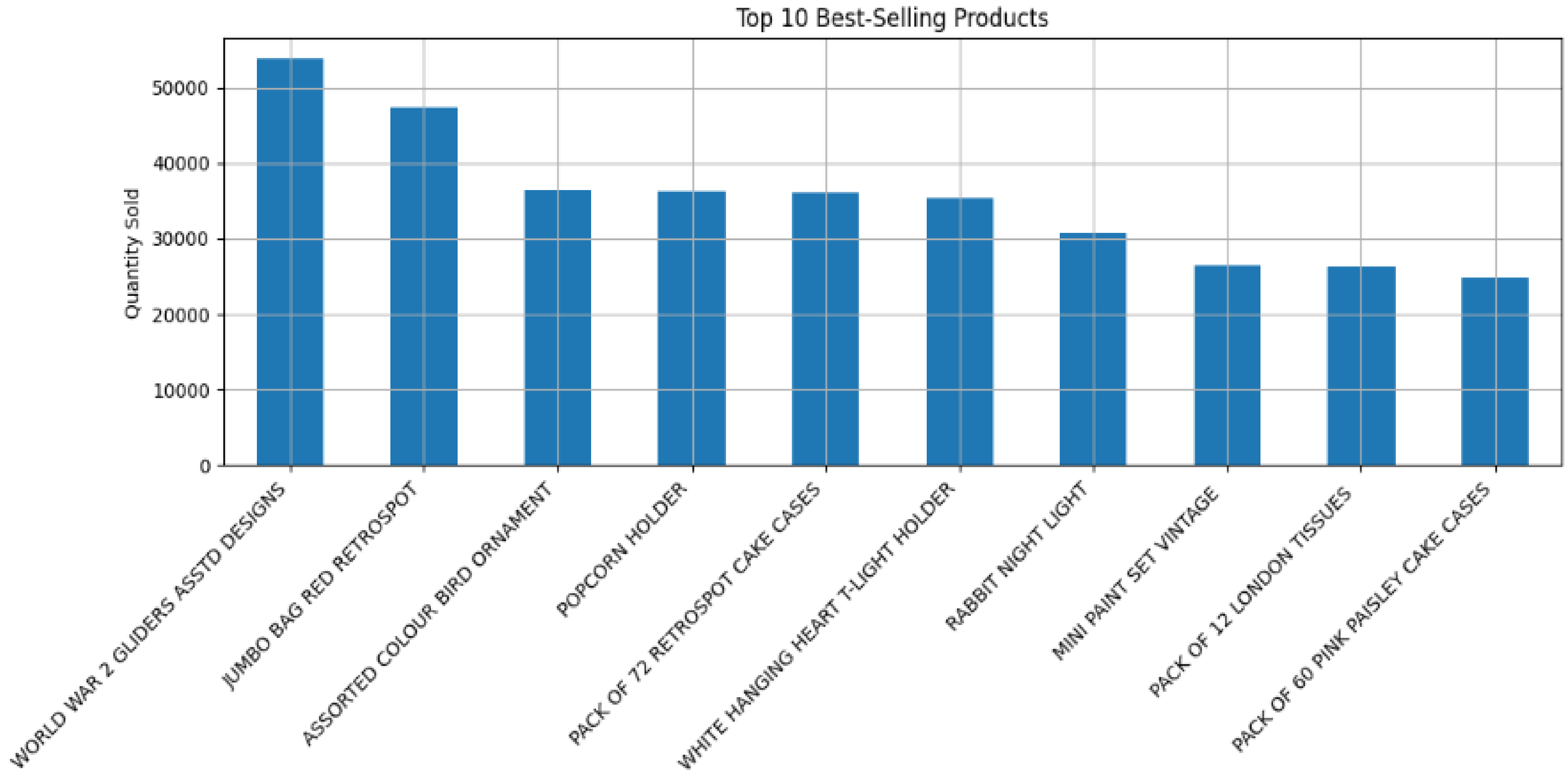
6. Visualization (line, bar, histograms)

7. Interpretation & Deployment (SQLite storage)

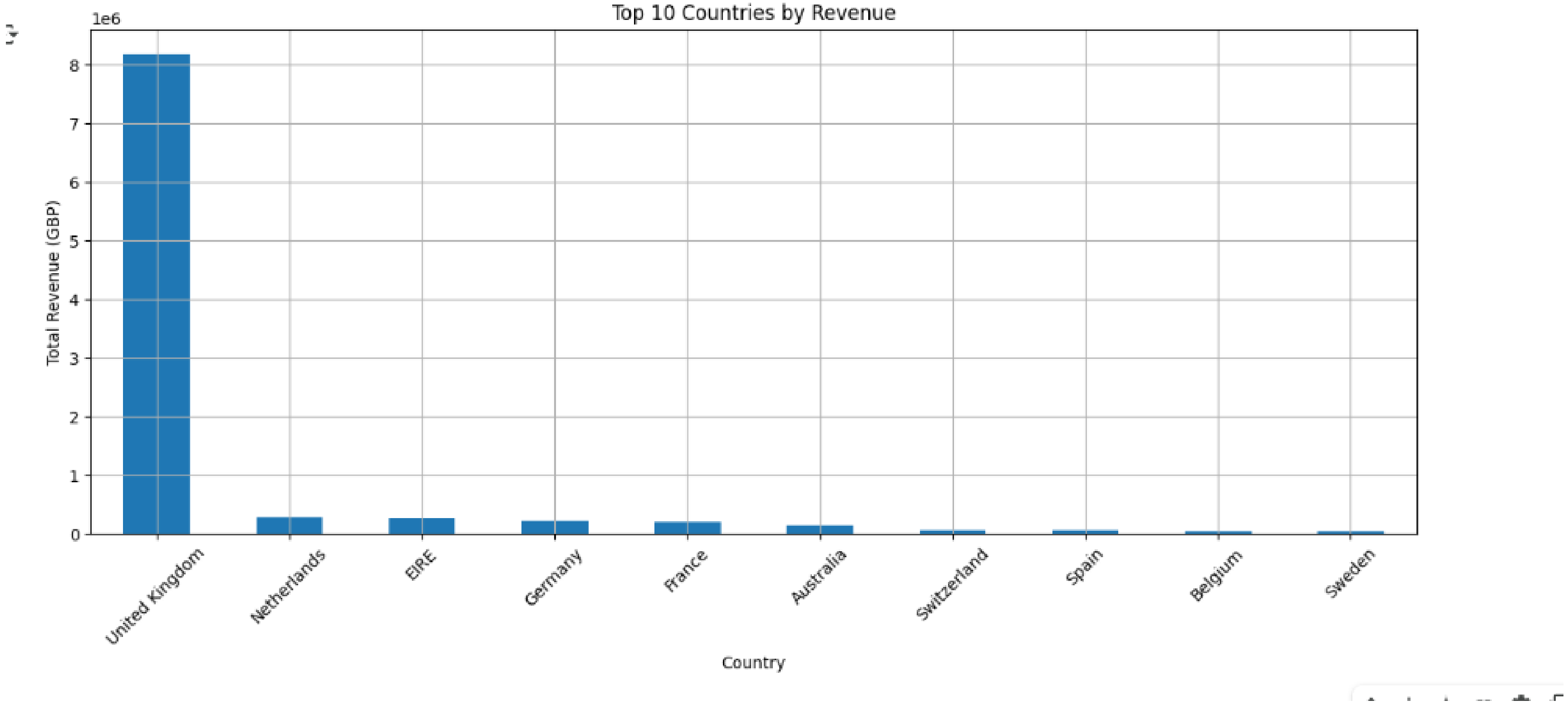
Key Analysis & Results

- Top 10 Best-Selling Products
- Top 10 Countries by Total Revenue
- Sales Trend over Time
- Customer Distribution by Revenue

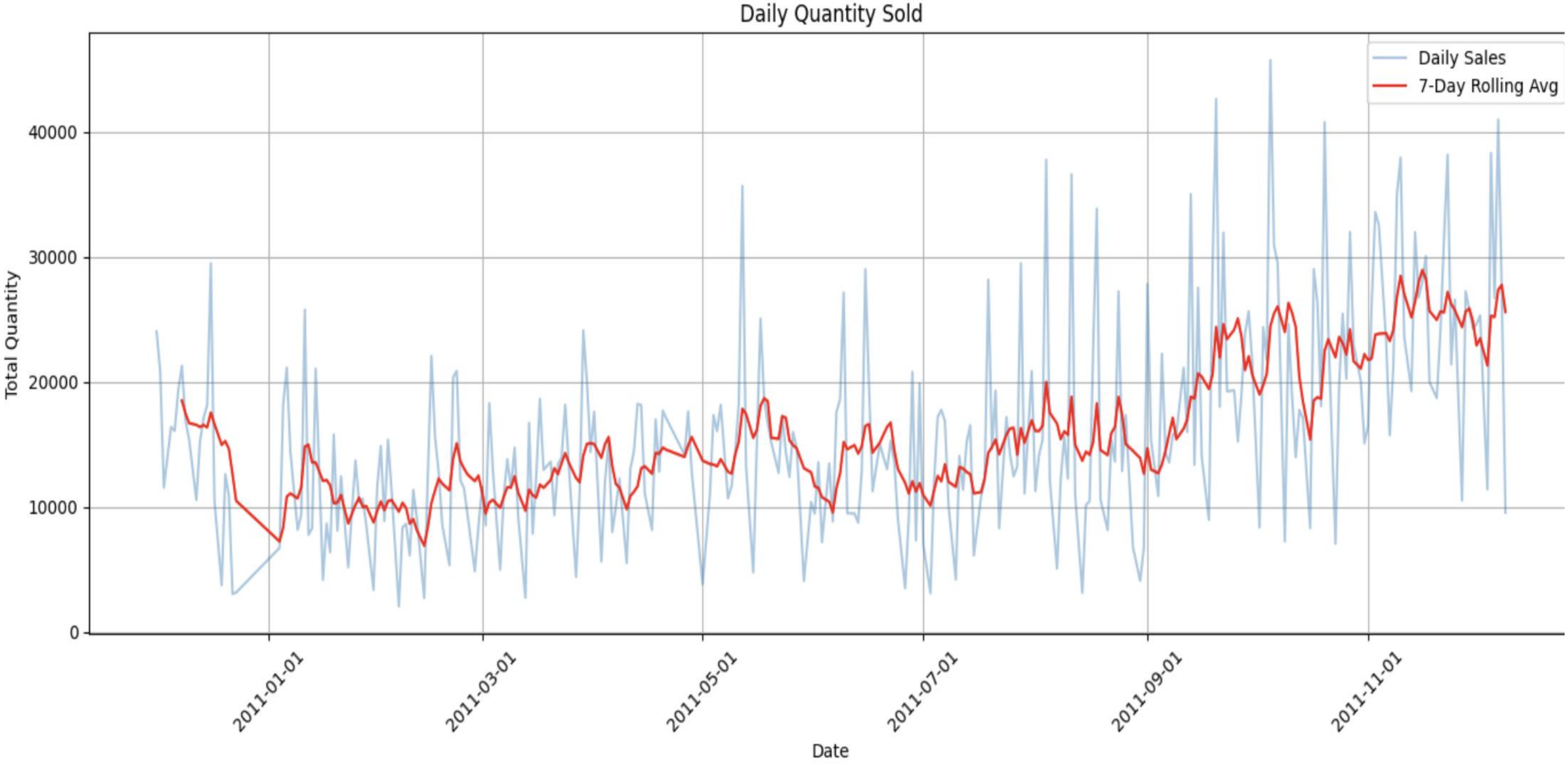
Top 10 Best-Selling Products



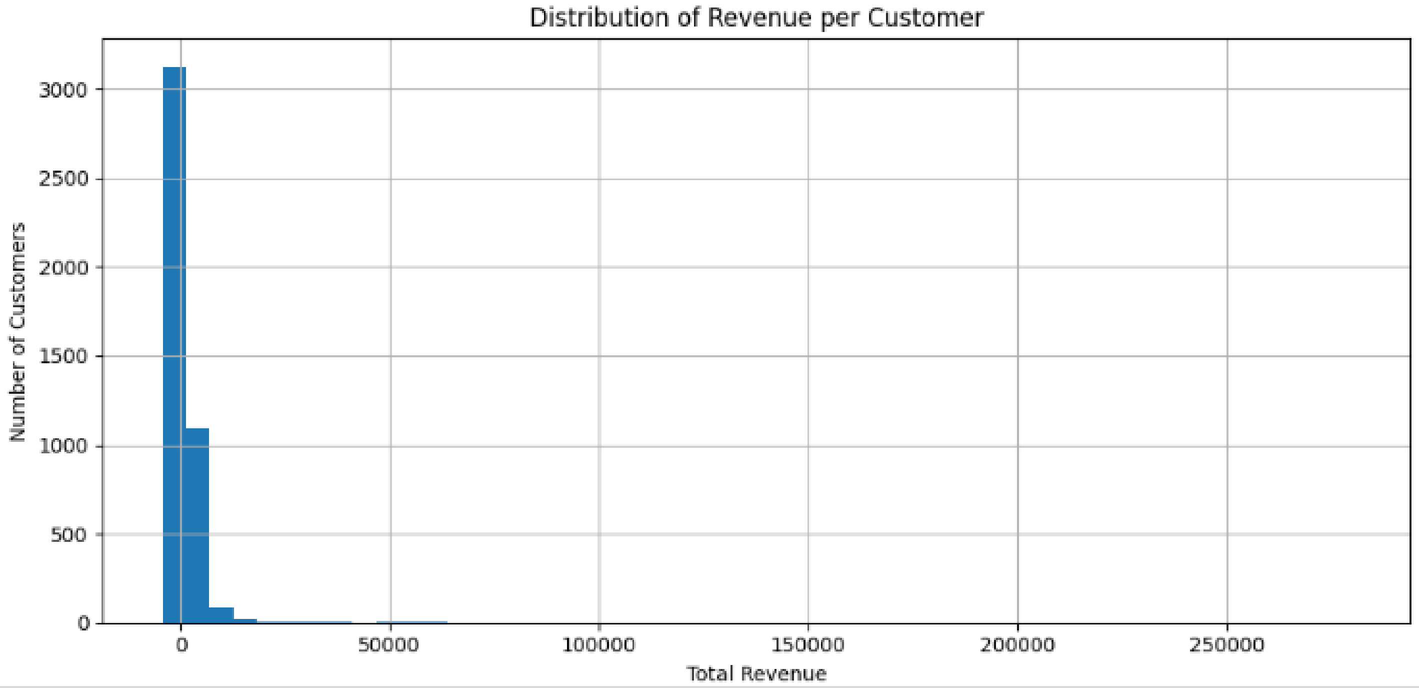
Top 10 Countries by Revenue



Daily Quantity Sold



Distribution of Revenue per Customer



Challenges Faced in the Project

- **Data Quality**: Missing values and product returns required cleaning.
- **Infrastructure Setup**: Spark installation and JVM/winutils configuration were time-consuming.
- **Visualization Accuracy**: Ensuring proper time intervals and scale of the graphs to avoid clutter in Matplotlib.
- **Learning Curve**: Required skills in Spark, SQL, and Python.

Technologies Used

Apache Spark – Distributed data processing

PySpark – Python API for Spark

Pandas – Exploratory data analysis

SQLite – Lightweight relational storage

Matplotlib – Data visualization

Google Colab – Cloud collaboration

Key Takeaways

SPARK ENABLES SCALABLE DATA PROCESSING

REVENUE AND SALES INSIGHTS EASILY DERIVED

VISUALIZATIONS HELP COMMUNICATE DATA TRENDS

RETAIL ANALYTICS IS EFFECTIVE FOR CUSTOMER BEHAVIOR INSIGHTS

References

- Chen, D. (2012). UCI Machine Learning Repository.
- EMC Education Services. (2015). Data Science & Big Data Analytics.
- Hashem et al. (2015). Information Systems, 47, 98–115.
- Russom, P. (2011). TDWI Big Data Analytics Report.
- ChatGPT