Introduction

The problem addressed in this project involves analyzing an online retail dataset from Kaggle. The goal is to identify the top 10 countries with the highest sales and the top 10 products with the highest sales. The challenge in this project was to preprocess the dataset to identify relevant information, clean the data, and perform visual analysis. The dataset contains transactional data, including product details, country, price, and quantity sold. Various technologies such as Pandas, Matplotlib, and Apache Spark were used to manipulate and visualize the data.

Data Collection

The dataset used in this project is publicly available on Kaggle and contains detailed sales data for an online retail store. The dataset includes columns like Order ID, Product Name, Quantity, Price, Country, and Customer ID. The data was cleaned by handling missing values and duplicates, and only relevant columns for the analysis (Country, Product Name, Quantity, and Price) were kept. The dataset was processed using Pandas to prepare it for visualization.

Data Processing

The data was cleaned and filtered to remove rows with missing or invalid values. The Quantity column was used to calculate the total sales for each product in every country. Sales were calculated as the product of quantity and price for each transaction. To identify the top 10 countries and products, we aggregated the sales data based on the country and product, then sorted the results in descending order.

Data Analysis

The data analysis focused on finding the top-selling countries and the most popular products.

- **Top 10 Selling Countries**: The total sales for each country were calculated and sorted to identify the 10 countries with the highest sales.
- **Top 10 Selling Products**: Similarly, the top 10 products were identified by calculating the total sales for each product.

Bar charts were used to visualize the sales distribution across countries, and a separate bar chart was created for the top-selling products. The following figures demonstrate the visual representations of the findings.

Figure I: Bar chart of top 10 countries by sales

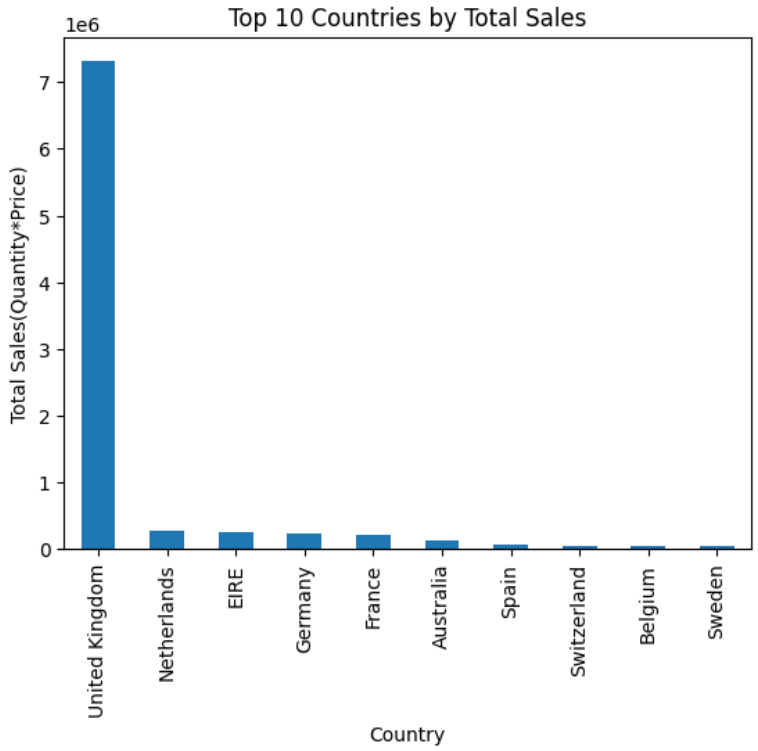
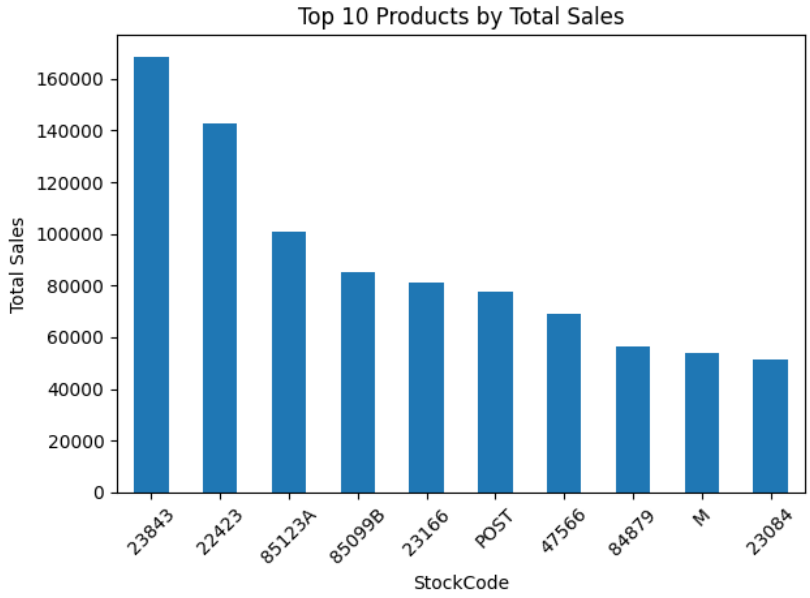


Figure II: Bar chart of top 10 products by sales



Conclusion

This project successfully analyzed the online retail dataset to identify the top-selling countries and products. The visual analysis revealed valuable insights into sales distribution by country and product. In the future, further analysis could be done to understand customer preferences, seasonal trends,

and other factors affecting sales. Additionally, predictive models could be developed to forecast future sales trends.