Sales Data Analysis and Visualization of Online Retail Dataset Using R

(Major Project Presentation)

PRESENTED BY:
ASHWINJEET SANDHU
O23MCA110142

CU
CHANDIGARH UNIVERSITY

CHANDIGARH UNIVERSITY
Discover. Learn. Empower.

Project Overview

- Objective: Analyse and visualize e-commerce sales data using R.
- Dataset: Online Retail dataset containing transactions of a UK-based online store.
- Goal: Extract insights such as top customers, revenue trends, and product demand.
- Tools Used: R, RStudio, and popular data science libraries (dplyr, ggplot2, etc.)

Problem Statement

- E-commerce businesses generate large volumes of sales data daily.
- Making sense of this data is crucial for understanding customer behavior and improving business decisions.
- Raw data often contains missing, inconsistent, or redundant entries.
- There is a need to clean, analyze, and visualize this data to extract meaningful insights.
- Businesses often lack clear visibility into top-performing products and high-value customers.
- Visual representation of data helps in faster and more informed decision-making.

Objectives of the Project

- To clean and preprocess the e-commerce sales data for accurate analysis.
- To perform exploratory data analysis (EDA) using R.
- To identify top-performing products and top customers by revenue.
- To analyse sales trends over time (e.g., monthly revenue).
- To generate clear and interactive visualizations using R packages like ggplot2.

Tools & Technologies Used in the Project

- Programming Language: R
- Development Environment: RStudio
- Data Manipulation: dplyr, readr
- Data Visualization: ggplot2, scales
- Date Handling: lubridate
- File Format: CSV (Comma-Separated Values)
- Version Control & Collaboration: Git, GitHub

Software Requirements

- R Programming Language – version 4 or higher
- RStudio IDE – preferred for script execution and visualization
- Required R packages:
  - dplyr – For data manipulation
  - ggplot – For data visualization
  - lubridate – For date-time handling
  - readr – For importing data
  - scales – For formatting values in plot
- Operating System – Compatible with Windows, Linux and macOS

Hardware Requirements

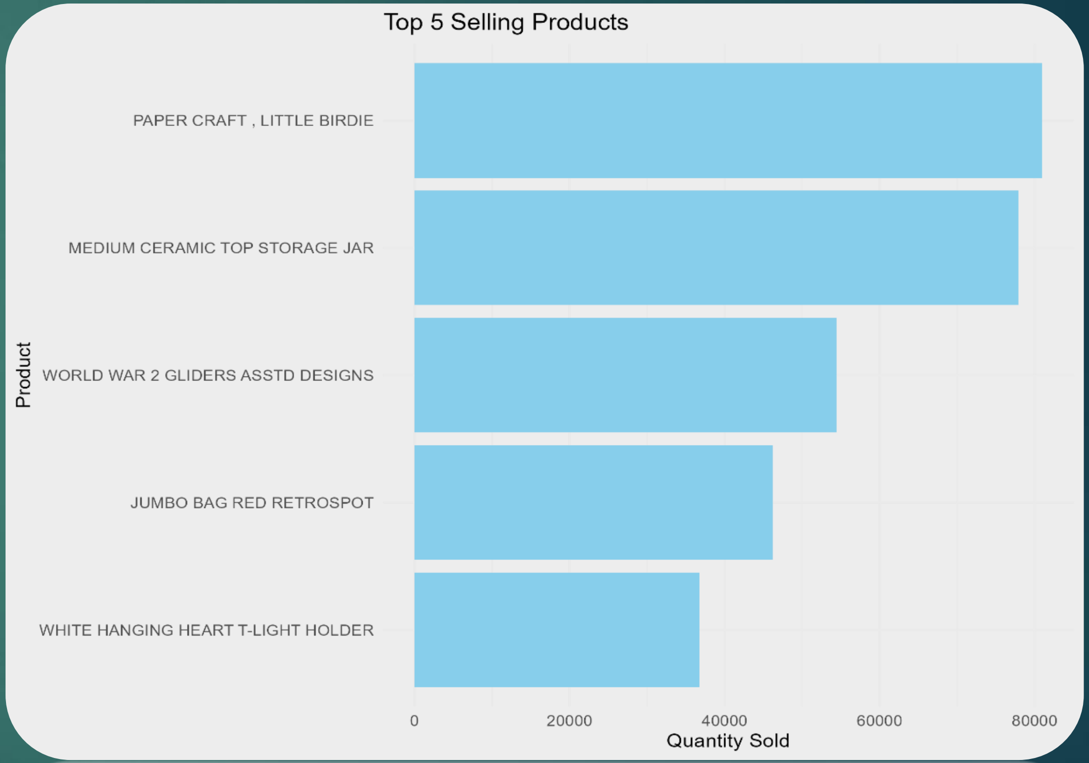
- Processor: Dual-core or higher (Intel i3/i5 or AMD equivalent
- RAM: Minimum 4 GB (8 GB recommended for smooth performance
- Storage:
  - At least 2 GB free space for R, RStudio, and project files
  - Additional space for data storage and output images
- Display: 1366x768 resolution or higher (for clear visualization)
- Internet Connection: Required for downloading packages and GitHub access

Data Description

- Dataset Source: Online Retail Dataset
- Total Records: 541909
- Columns:
  - InvoiceNo: Unique transaction Identifier
  - StockCode: Product code
  - Description: Product name
  - Quantity: Number of items sold
  - InvoiceDate: Date and Time of transaction
  - UnitPrice: Price per item
  - CustomerID: Unique customer Identifier
  - Country: Customer’s country of residence

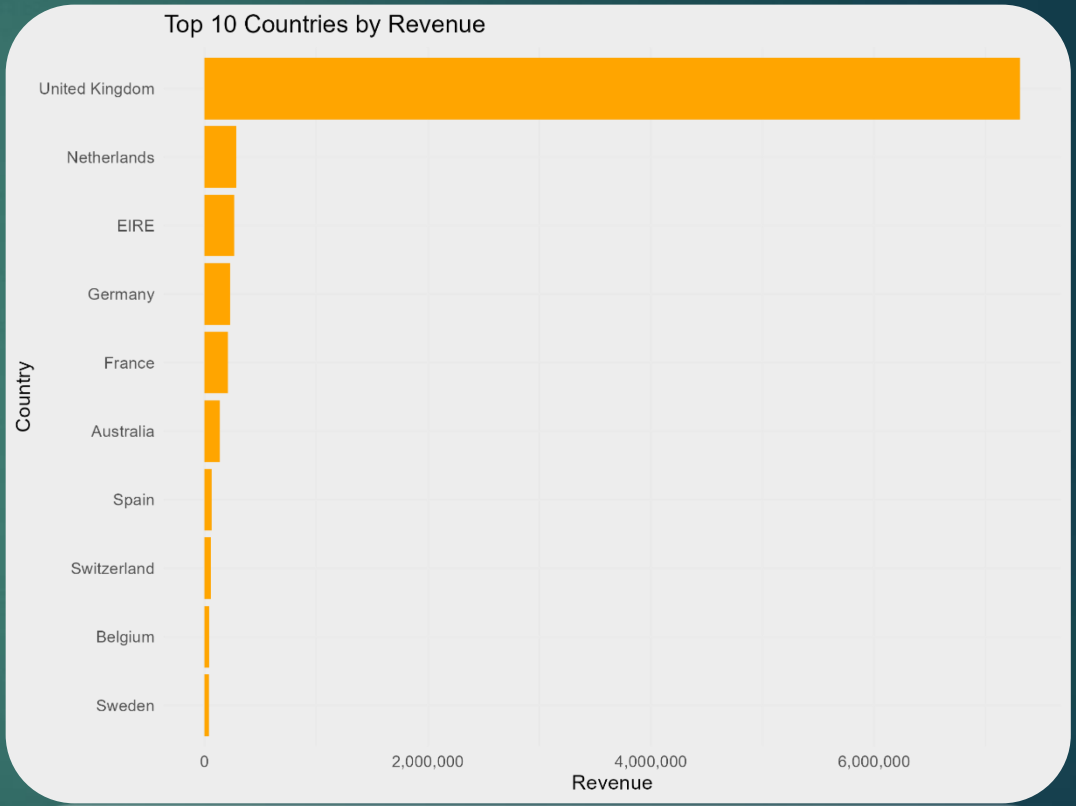
Top 5 Products by Revenue

- This bar chart displays the top 5 products that generated the highest revenue
- Helps identify bestselling items contributing significantly to total income
- Useful for inventory planning, marketing focus and customer targeting



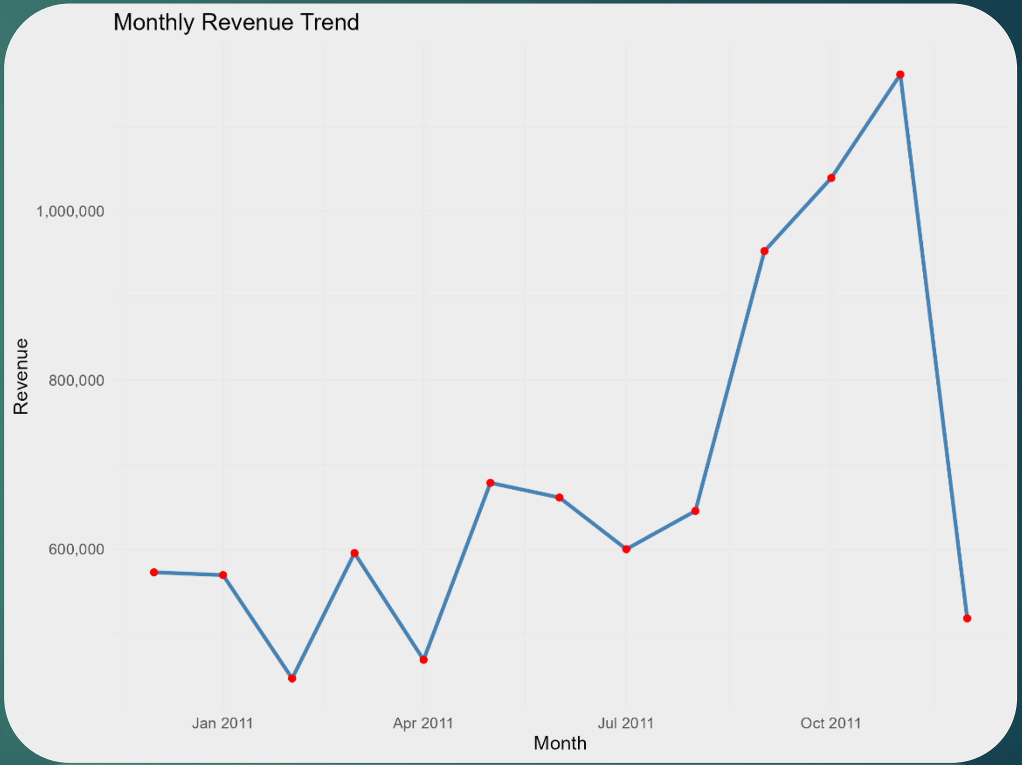
Top 10 Countries by Revenue

- Highlights international market contributions to total revenue
- Helps identify geographical performance trends
- Useful for strategic decisions like market expansion or localized marketing
- Reveals customer bases beyond the home country, aiding global planning



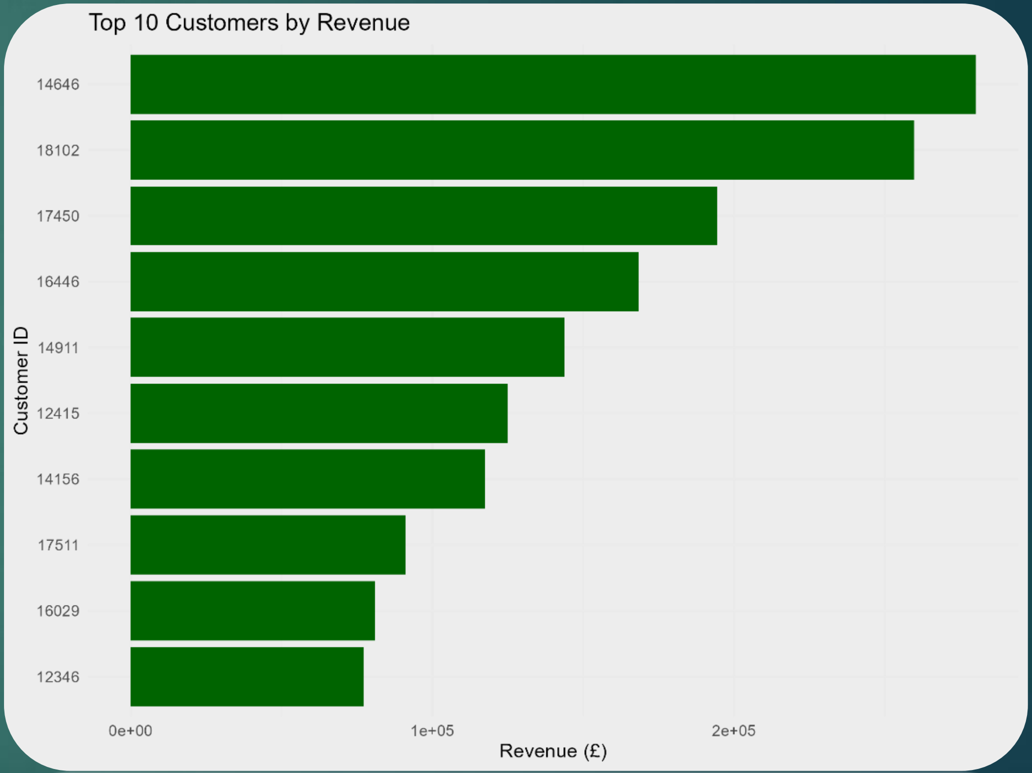
Monthly Revenue Trend

- Shows sales performance over time
- Helps identify seasonal patterns and sales spikes/drops
- Useful for forecasting future revenue and understanding growth trends
- Insights derived using lubridate for date parsing and ggplot2 for visualization



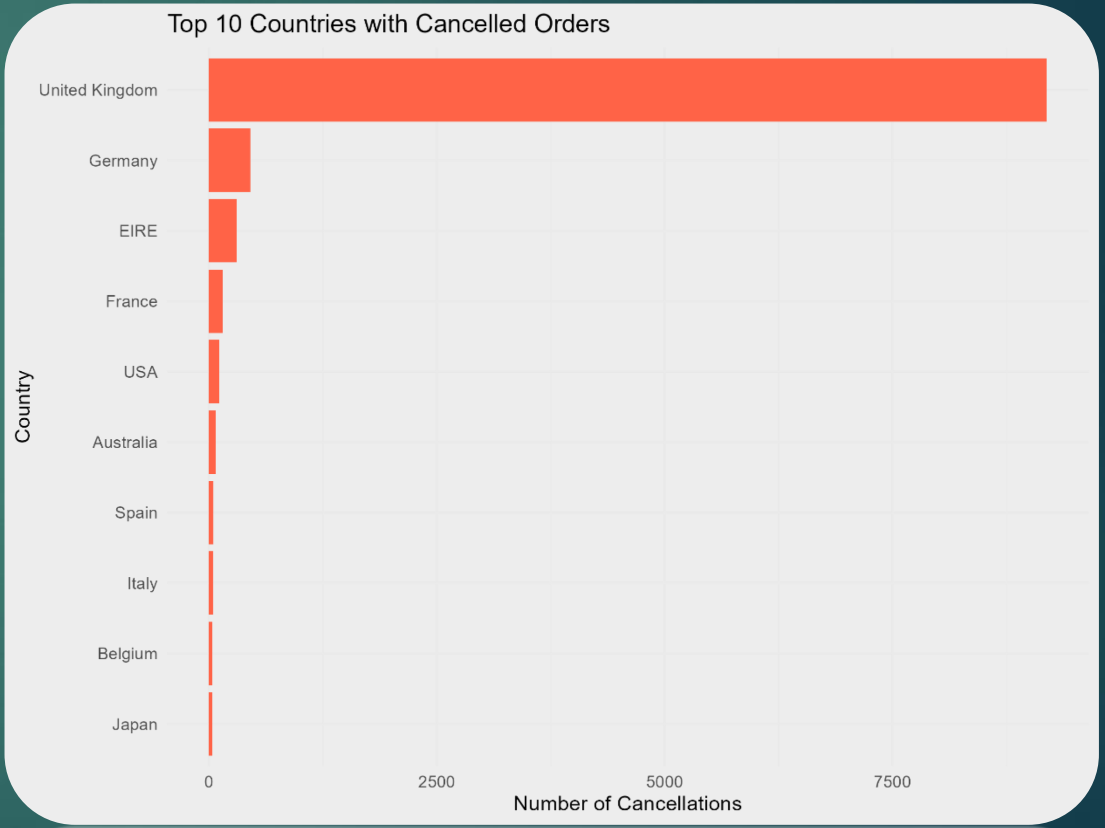
Top 10 Customers by Revenue

- Displays highest paying customers by total revenue
- Helps identify high-value customers for targeted marketing or loyalty programs
- Revenue calculated by summing transactions for each unique CustomerID
- Visualized using ggplot2 using coord_flip() for better readability



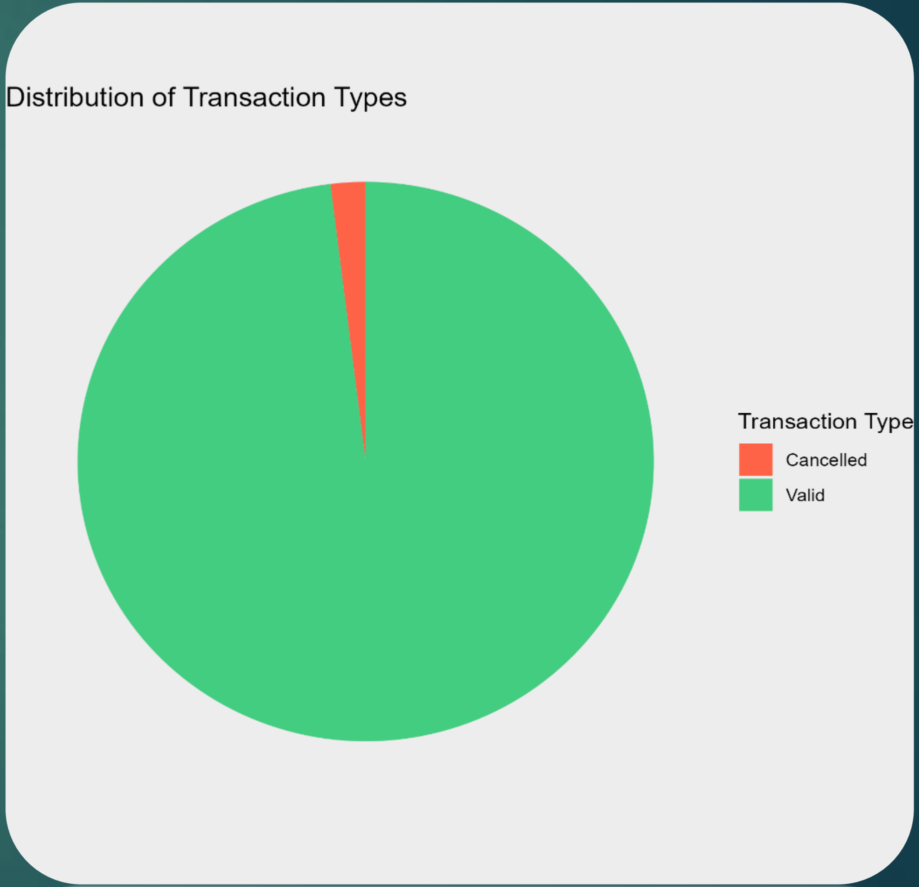
Top 10 Countries with Cancelled Orders

- Highlights countries with the highest frequency of cancelled transactions
- Useful for identifying potential logistics or customer satisfaction issues
- Data filtered where Quantity < 0 to represent cancelled items
- Enables the business to take targeted actions in regions with high return/cancellation rates



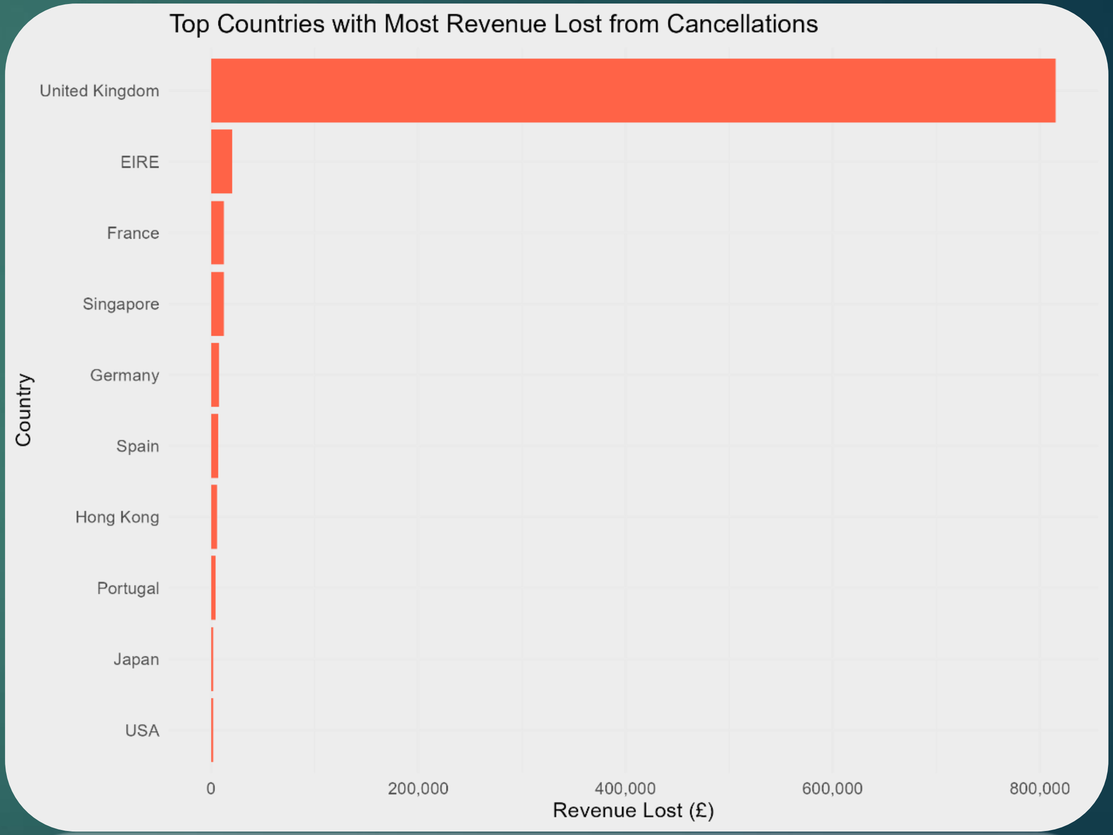
Distribution of Transaction Types

- Transactions were categorized as:
  - Valid (Quantity > 0)
  - Invalid (Quantity < 0)
- This pie chart visualizes the proportion of valid vs cancelled transactions
- Help identify the frequency of cancellations, which is crucial for understanding customer behavior and possible operational issues
- Majority of the transactions are usually valid, but cancelled transactions also form a noticeable portion of total orders



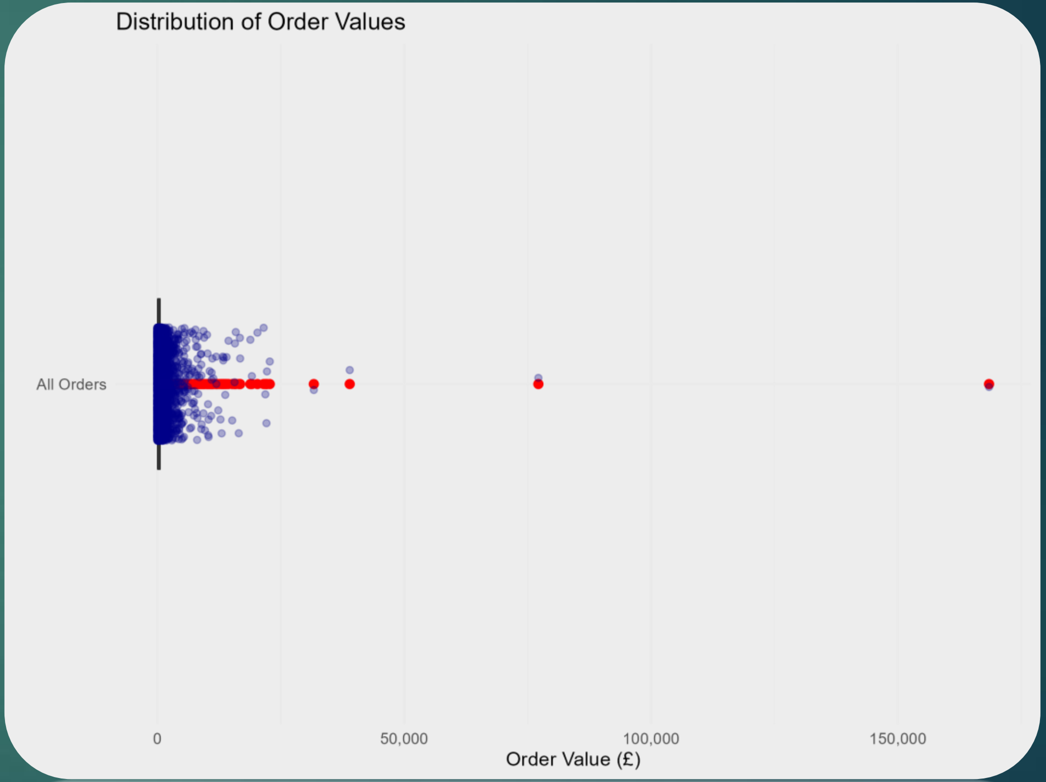
Top Countries with Most Revenue Lost from Cancellations

- This bar chart highlights the top countries where the most revenue was lost due to cancelled orders
- The chart helps pinpoint regions causing significant financial impact, aiding in
  - Investing potential causes of high cancellations
  - Improving operational efficiency
  - Enhancing customer service in high-loss regions
- Cancellations were identified using transactions with negative quantities, and revenue loss was calculated accordingly


Top Countries with Most Revenue Lost from Cancellations

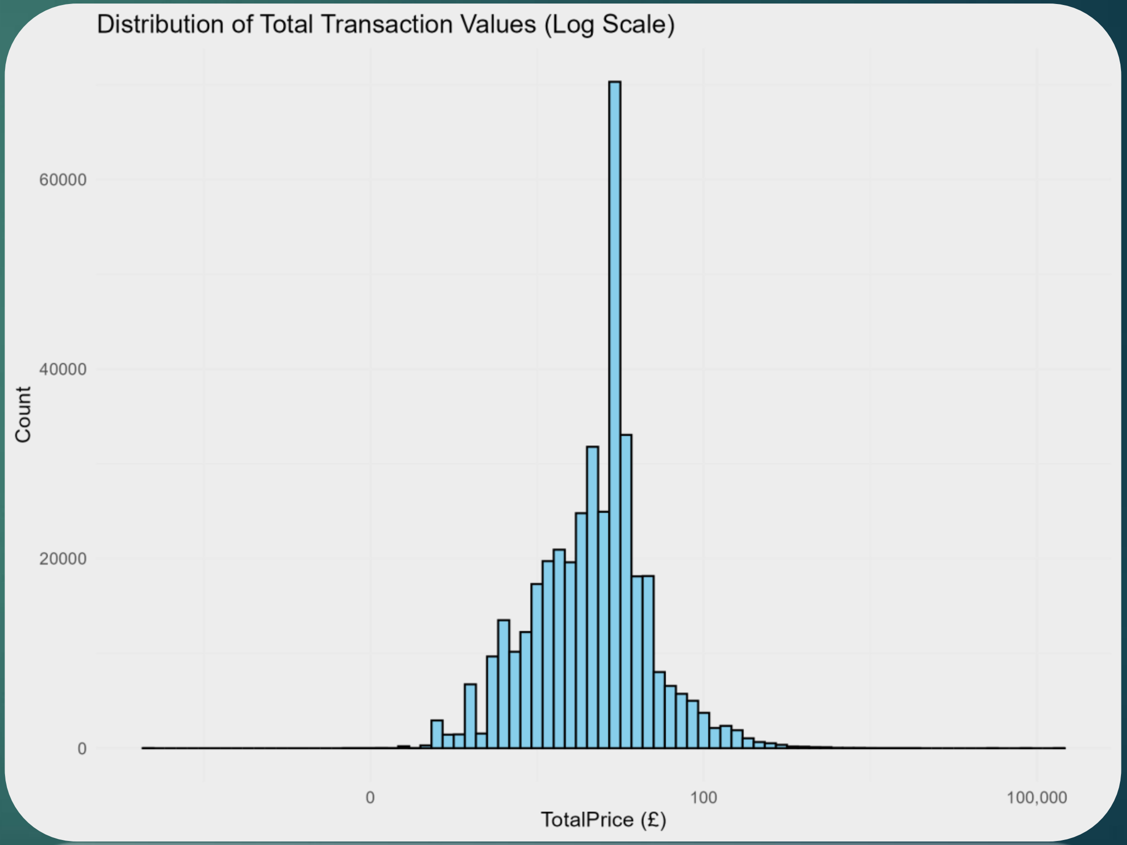
Distribution of Order Values

- This visualization shows the distribution of total order values across all invoices
- Each order’s value was calculated by summing the prices of items in that invoice
- A boxplot is used to capture:
  - Median order value (central line in the box)
  - Interquartile Range (IQR)
  - Outliers
- Jittered points provides a clearer picture of individual order distributions



Distribution of Total Transaction Values

- Show how transaction values are distributed across all orders
- Log scale on x-axis improves visibility of small and large transactions
- Most transactions falls within the lower price range
- Highlights presence of a few high-value orders
- Helps in analyzing customer spending behaviour



Key Insights

- The UK is the largest market by revenue and number of transactions
- A small number of customers contribute to a significant portion of the revenue
- High cancellation rates observed in certain countries lead to major revenue losses
- Most products sold are low in value, but there are occasional high-value orders
- Sales trends shows seasonal spikes, likely around holidays

Challenges Faced

- Data Cleaning: Dealing with missing values, duplicates, and invalid entries took considerable time
- Data Understanding: Interpreting business context from raw transactional data was challenging
- Visualizing Effectively: Choosing the right type to communicate insights clearly requires several iterations
- R Programming: Faced initial difficulties in syntax and debugging in R and ggplot2
- Time Management: Balancing project work with academic schedule and finalizing all components on time

Thank You

PRESENTED BY:

ASHWINJEET SANDHU

023MCA110142

CU
CHANDIGARH UNIVERSITY

CHANDIGARH UNIVERSITY

Discover. Learn. Empower.