Project Report

Online Retail Sales Forecasting

Course: Probability and Statistics for Artificial Intelligence (AAI-500-IN1)

Github Link: https://github.com/TANISHA2629/Project_1_Online_Retail.git

Submitted By:

Tanisha Gulhane

23rd June, 2025

Introduction

This project explores the Online Retail II dataset, which contains transactional records from a UK-based online retail company collected between 2009 and 2011.

Each entry in the dataset represents a unique invoice and includes attributes such as invoice number, product code and description, quantity sold, unit price, invoice date, customer identification, and country of the customer.

The central goal of the analysis is to understand the factors that most strongly influence revenue generation and to construct a regression model capable of predicting total revenue for each transaction. Such predictive capabilities are valuable for driving strategic decisions in areas like inventory management, dynamic pricing, and customer targeting.

Dataset

Source: [Online Retail II UCI]

Description of Dataset:

The Online Retail II dataset contains detailed transactional data for a UK-based, non-store online retail company. The dataset covers a period from December 1, 2009 to December 9, 2011, capturing sales activity for a business that primarily sells unique all-occasion giftware. The company's customer base includes both individual consumers and a significant number of wholesalers.

This dataset is well-suited for exploring retail sales forecasting, customer segmentation, revenue prediction, and market trend analysis, due to its rich combination of product, customer, and time-based attributes.

The dataset contains more than 1M rows and 8 columns. The following is the description of each column.

1. **Invoice**: Contains a 6-digit unique identifier for each transaction. If it begins with 'C', it denotes a cancellation.

2. **StockCode**: Contains a unique 5-digit identifier for each product.
3. **Description**: Contains the name or description of the product.
4. **Quantity**: Contains the units of the product purchased in the transaction.
5. **InvoiceDate**: Contains the date and time when the transaction occurred.
6. **Price**: Contains a price of the product, in Pound Sterling (£).
7. **CustomerID**: It contains a unique 5-digit identifier assigned to each customer.
8. **Country**: It contains the name of the country where each customer resides.

There is a small percentage of order cancellations in the dataset.

Data Cleaning/Preparation

The objective of this data cleaning phase is to prepare the raw sales transaction data for subsequent analysis. Data integrity is ensured by removing missing, duplicated, and erroneous records, and by ensuring all columns have the appropriate data types.

Summary of Cleaning Steps:

1. **Import libraries.**
2. **Load raw data.**
3. **Initial inspection of the data was performed.**
4. **Check for missing values, Datatype & Removing Duplicates**.
5. **Fixing Datatype**
6. **Removing Cancellations (Negative Quantity or Credit Invoices).**
7. **Create new feature - Total Revenue**
8. **Outlier Detection**
9. **Standardize Country Name**
10. **Cleaned Dataset**

To begin, the dataset was cleaned and prepared for analysis through a systematic process. All rows containing missing values in critical fields such as quantity, price, invoice date, or customer

ID were excluded to ensure data completeness and consistency. Furthermore, records involving non-positive quantities or prices were removed, as these do not represent valid sales transactions. A new variable, total revenue, was engineered by multiplying quantity and price for each transaction. The 'InvoiceDate' field was transformed into a datetime format to facilitate potential temporal analyses, while the 'Country' field was standardized using title casing. Since regression models cannot directly interpret categorical variables, the 'Country' column was converted into binary indicator variables through one-hot encoding, dropping the first level to avoid multicollinearity. To validate the model's performance on unseen data, the cleaned dataset was split into training and testing subsets using an 80:20 ratio.

Exploratory Data Analysis

Loading Cleaned Data:

Source: cleaned_online_retail.csv

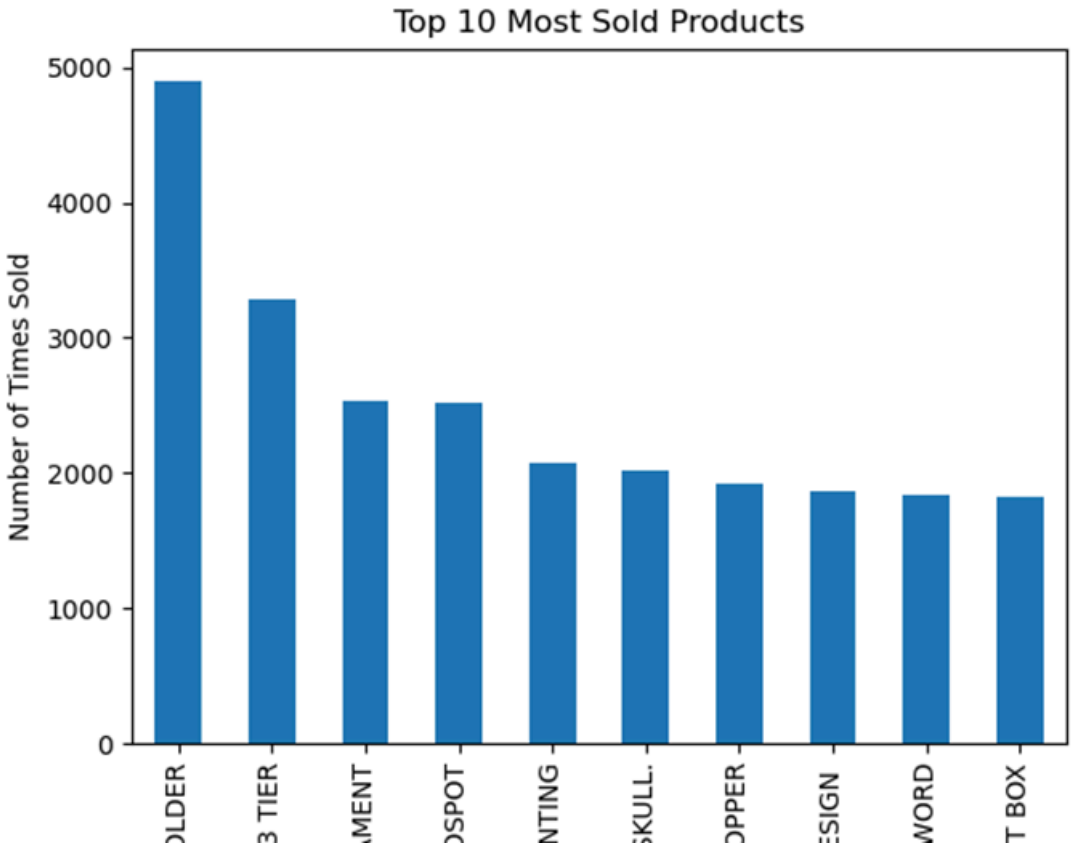
Summary of EDA Part-1:

1. **Extracting top product & their frequencies**
2. **Revenue By Country**
3. **Monthly Sales Trend**
4. **Sales By Day Of Week**
5. **Top Customers Revenue**
6. **Order Quality Distribution**
7. **Correlation Heatmap**

Detailed Summary for EDA Part-1:

Dataset: Online Retail II.
Period Covered: 01-Dec-2009 to 09-Dec-2011.
Focus: The focus is on understanding retail trends, key revenue drivers, and customer and product behavior.

1. Top Products & Their Frequencies



The most frequently sold products reveal key revenue generators and customer preferences.

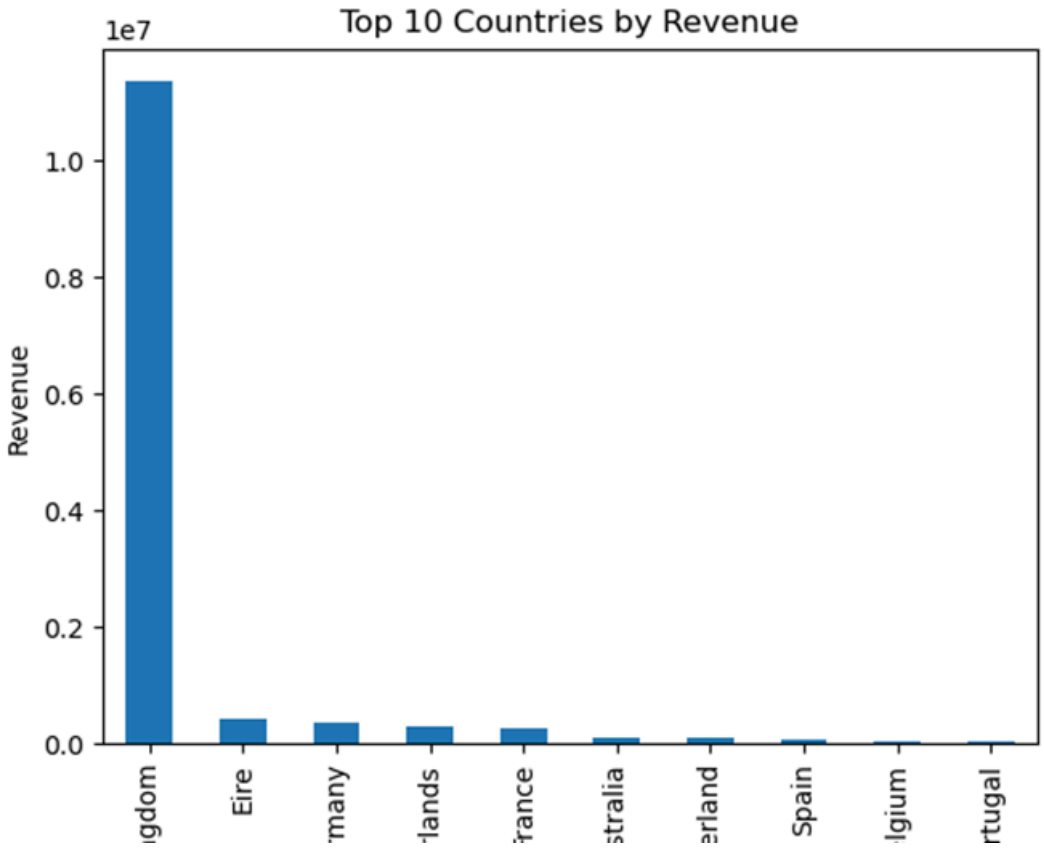
- Top items (e.g., WHITE HANGING HEART T-LIGHT HOLDER, REGENCY CAKESTAND 3 TIER) appeared consistently with high volume.
- These products often belong to categories like home decor, gifts, and accessories.

Insight: A small number of core products drive most sales, so targeting promotions around them can significantly boost ROI.

Recommendation:

- Ensure high inventory levels of top-selling products to prevent stockouts.
- Bundling popular items with less popular ones can help boost sales across different product categories.
- Leverage top-selling products in promotional campaigns or seasonal offers to drive traffic and increase average basket size.

2. Revenue by Country



A bar chart or grouped summary revealed:

- UK dominates in revenue generation, followed by Netherlands, Germany, and France.
- Countries like EIRE, Spain, and Switzerland contributed modestly.

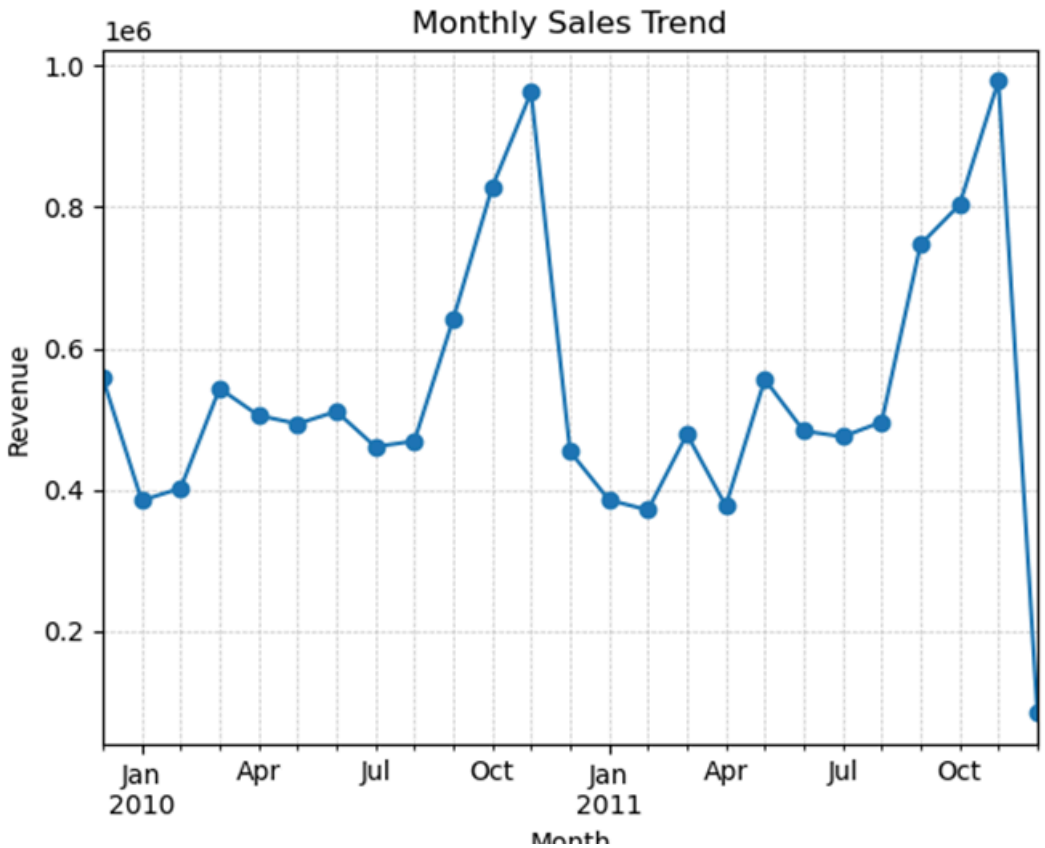
Insight: While UK remains the primary market, there's meaningful international engagement worth expanding strategically.

Recommendation:

- The UK should be prioritized for new product launches, loyalty programs, and expedited delivery options.
- Analyze lower-performing countries for potential issues in logistics, demand, or pricing, and adjust marketing strategies to address them.

- Localize offerings like language, currency, and promotions in high-potential, low-revenue countries to improve performance.

3. Monthly Sales Trend



Monthly aggregation shows:

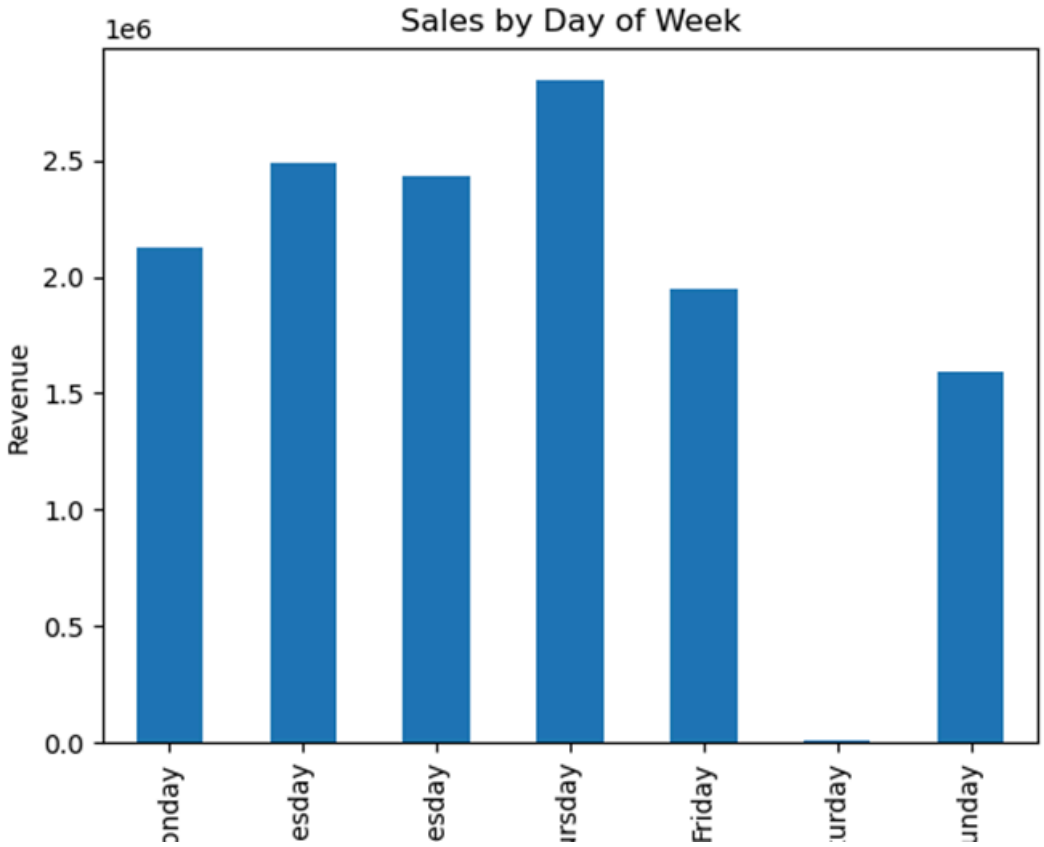
- Notable seasonal spikes occur in November and December, likely driven by holiday shopping.
- A mid-year dip in sales during June to August suggests seasonal trends or slower retail activity during this period.

Insight: Marketing campaigns can be aligned with peaks (e.g., Q4) and off-peak periods for better inventory and staffing.

Recommendation:

- We should conduct targeted marketing campaigns in October and early November to capitalize on peak sales.
- Prepare inventory, staffing, and logistics ahead of time to meet expected demand.

4. Sales by Day of the Week



The data showed higher sales activity on:

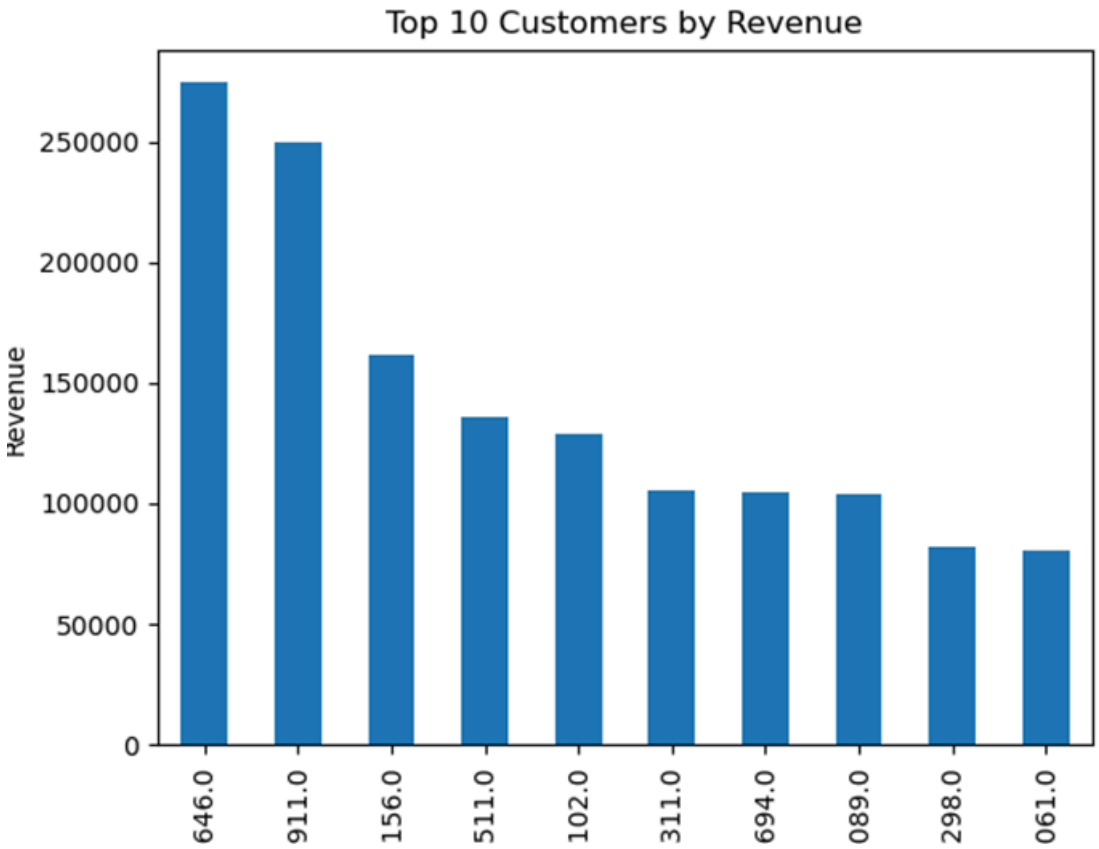
- Weekdays, especially Tuesdays and Thursdays
- Weekends exhibited noticeably lower sales, likely due to business customer inactivity.

Insight: This aligns with B2B wholesaler behavior, campaigns should prioritize mid-week visibility.

Recommendation:

- We should explore opening on Saturdays if operationally feasible, it may represent untapped sales potential.

5. Top Customers by Revenue



A ranked list of customers by CustomerID and total revenue indicated:

- A small number of customers contributed disproportionately high sales volumes.
- Customer segmentation could be applied (e.g., RFM or CLTV analysis) for retention and targeting.

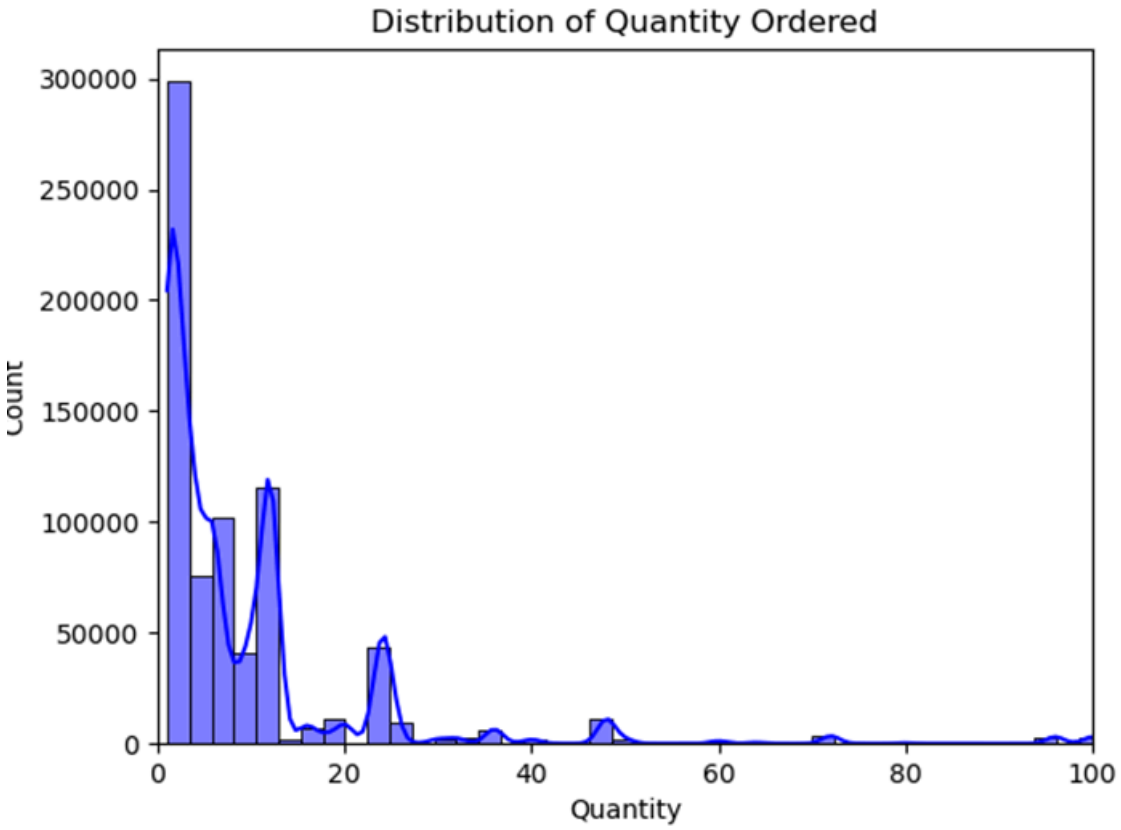
Insight: Prioritize high-value customers with loyalty programs or personalized offerings.

Recommendation:

- Identify top customers and enroll them in VIP or loyalty programs.

- Consider personalized offers or early access to products for high-value clients to increase retention.
- Use their behavior data to build lookalike customer segments for targeted advertising.

6. Order Quantity Distribution

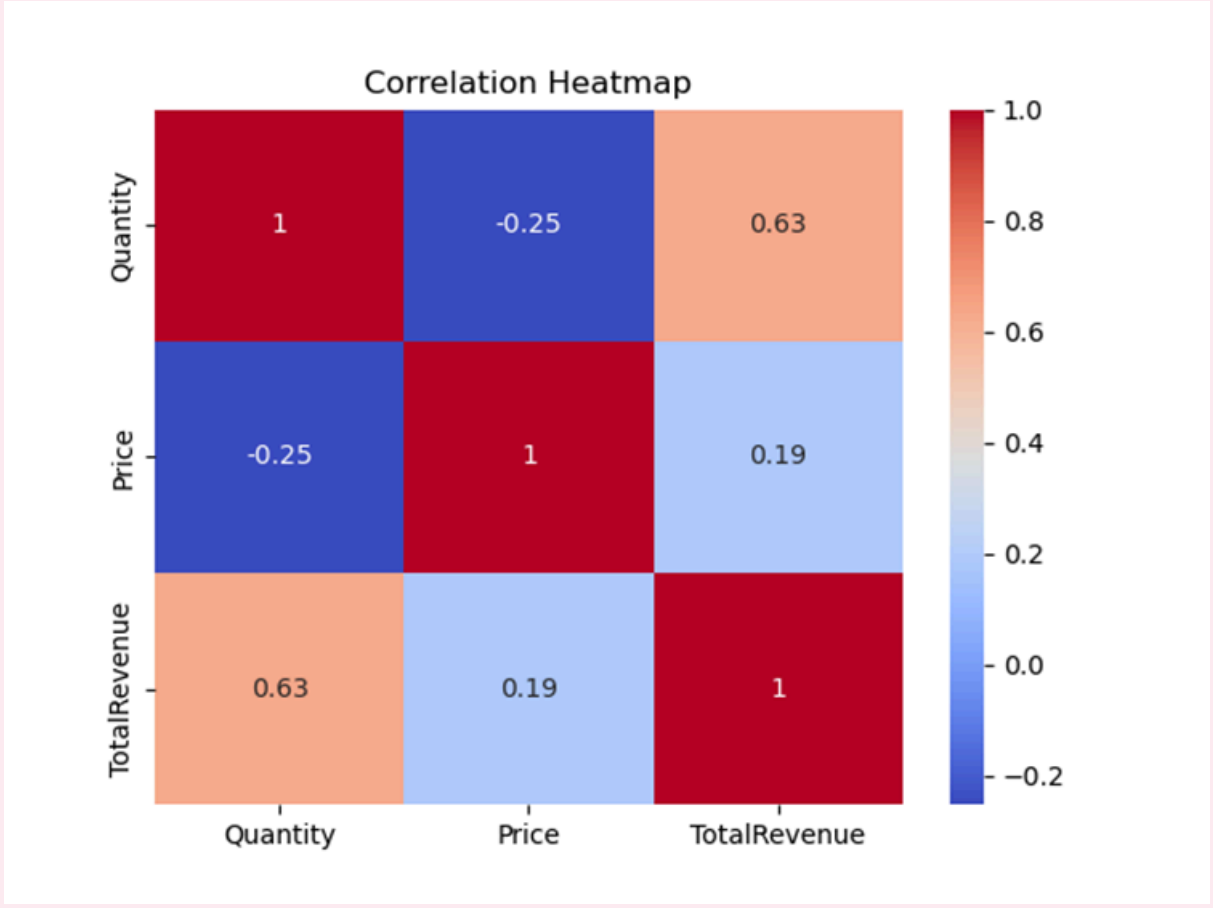


Insight: Further cleansing and outlier filtering is necessary before modeling. Understanding returns is also key to optimizing logistics.

Recommendation:

- Identify if there are too many small or extremely large orders—optimize pricing or offer discounts on volume thresholds.
- Ensure packaging and shipping processes accommodate the most common order sizes efficiently.

7. Correlation Heatmap



The heatmap shows:

- The heatmap shows a strong positive correlation between Quantity and TotalRevenue, along with a slight negative correlation between Price and Quantity, indicating that discounts may help boost sales volume.

Insight: Revenue is quantity-driven, with unit pricing playing a strategic role in influencing bulk orders.

Recommendation:

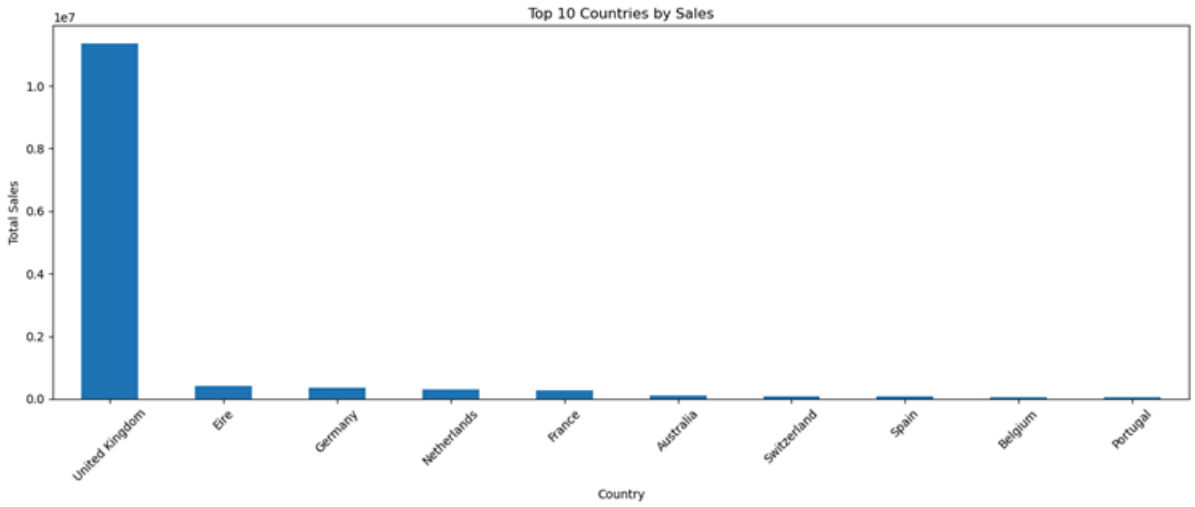
- Consider upsell strategies like “Buy More, Save More”.

Summary of EDA Part-2:

1. **Total Sales by Country**
2. **Per-Capita Sales by Country**
3. **Sales by Weekday**
4. **Weekday-Month Heatmap**
5. **Weekly Sales Trend**
6. **Top & Bottom 20 Products by Revenue**
7. **Price Distribution**
8. **Top Customers by Frequency and Value**

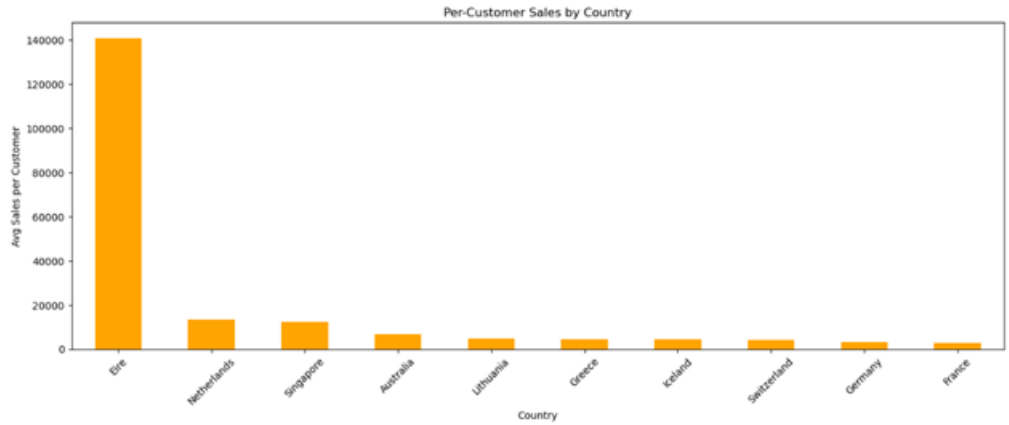
Detailed Summary for EDA Part-2:

1. **Total Sales by Country:**



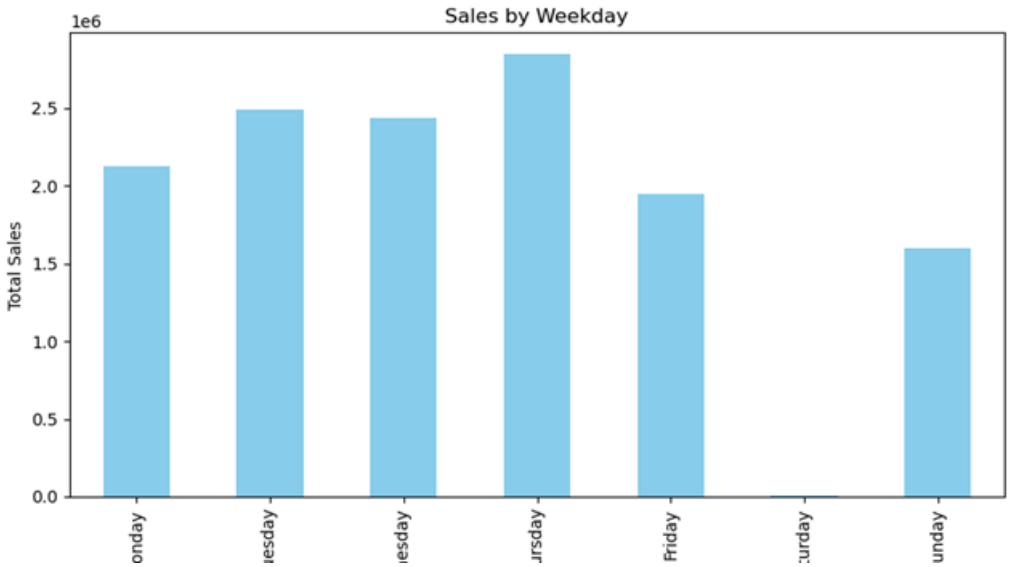
Observation: This bar chart is used for visualizes total sales of Top 10 countries.
Finding: We found a clear trend of higher sales from certain countries.
Recommendation: More resources should be allocated to high-performing markets to maximize returns.

2. Per-Capita Sales by Country



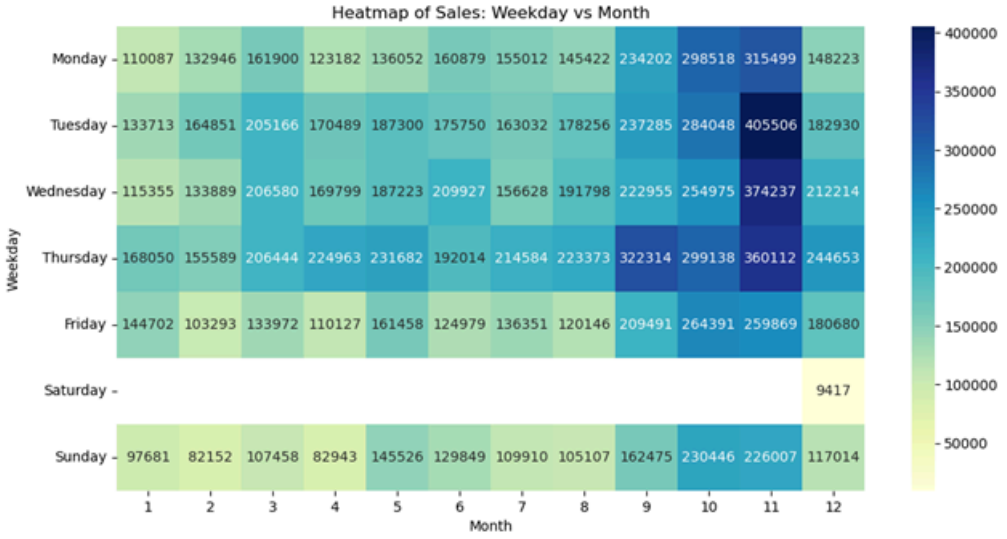
Observation: Shows average revenue contribution per customer by country.
Finding: Enables identifying premium or high-value markets.
Recommendation: Pricing and promotional strategies can be tailored to suit each market segment.

3. Sales by Weekday



Observation: Identifies which weekdays drive the highest sales.
Finding: Certain days (e.g., midweek(Thrusday)) dominate.
Recommendation: We should target advertisements and special offers for high-traffic days.

4. **Weekday-Month Heatmap**



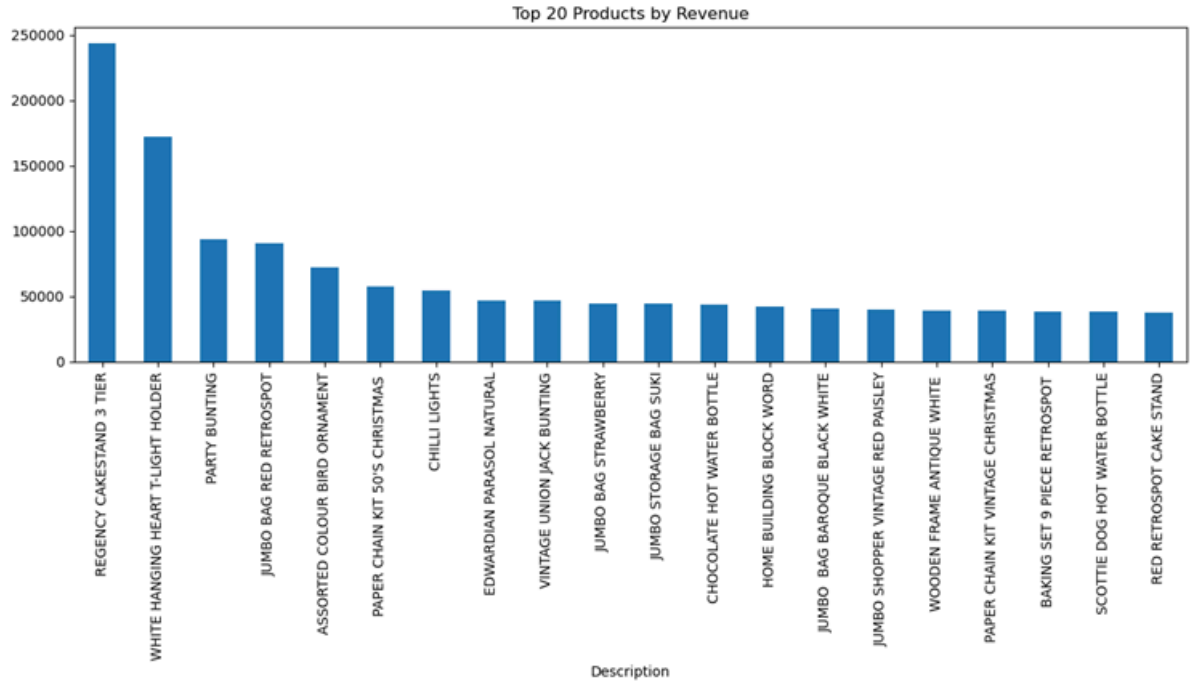
Observation: Shows sales trend across weekdays and months.
Finding: Certain Month–Weekday combinations show significantly higher sales activity.
Recommendation: Staffing and marketing resources should be allocated accordingly based on demand patterns.

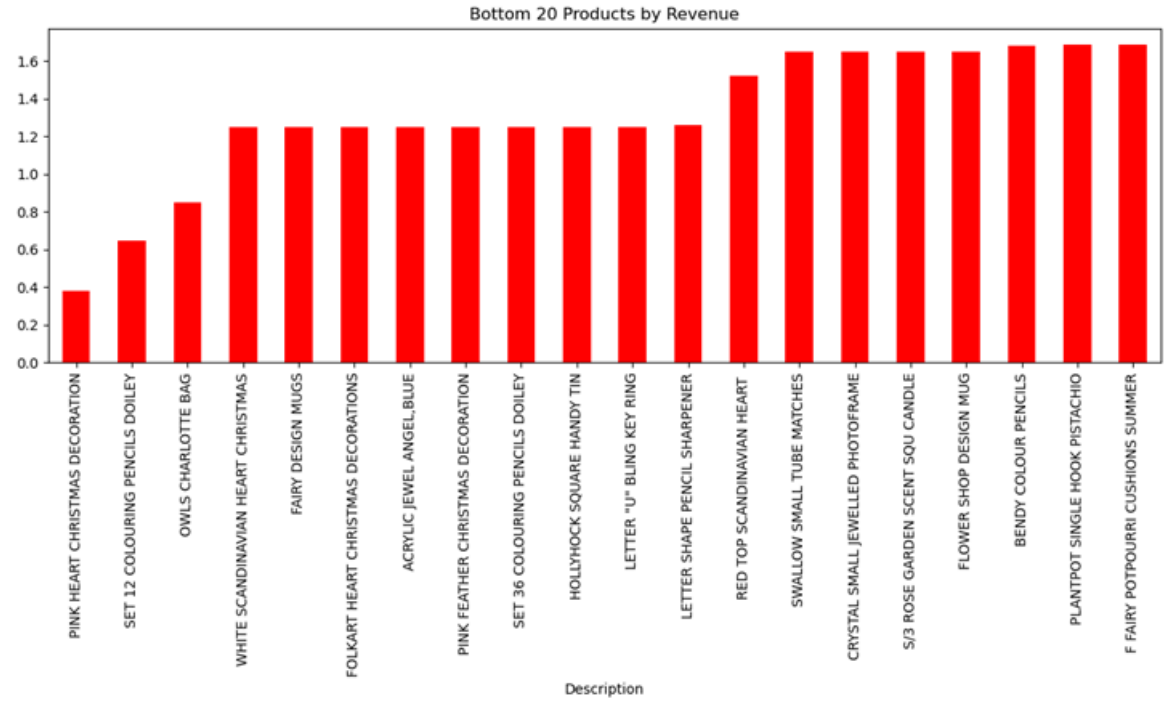
5. **Weekly Sales Trend**

Observation: Shows total sales trend across weeks.

Finding: Sales trends can identify peaks and valleys.
Recommendation: We should leverage this trend for targeted marketing.

6. Top & Bottom 20 Products by Revenue

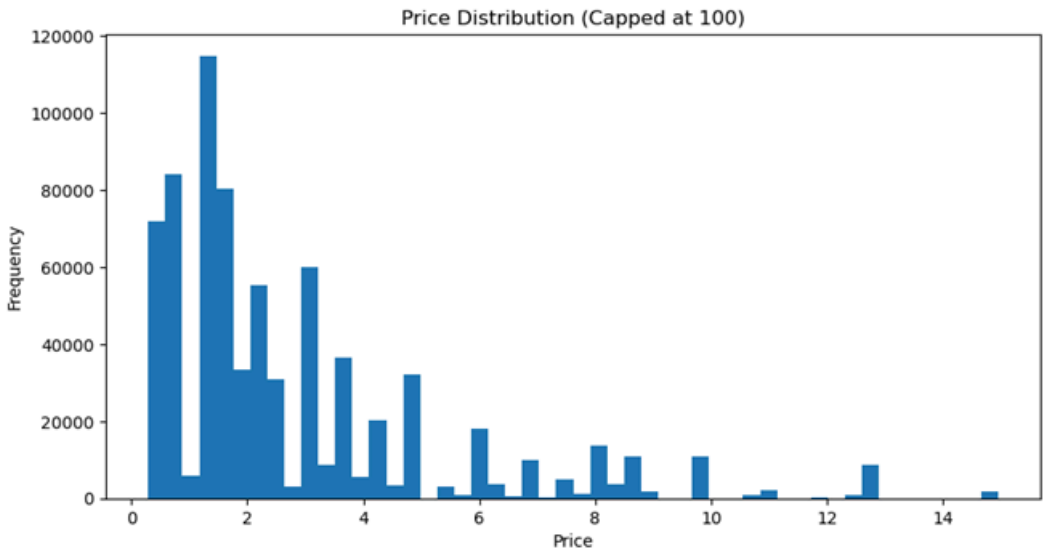




Observation: Shows top 20 & bottom 20 products by revenue.
Finding: A few products generate most revenue, while many contribute very little.

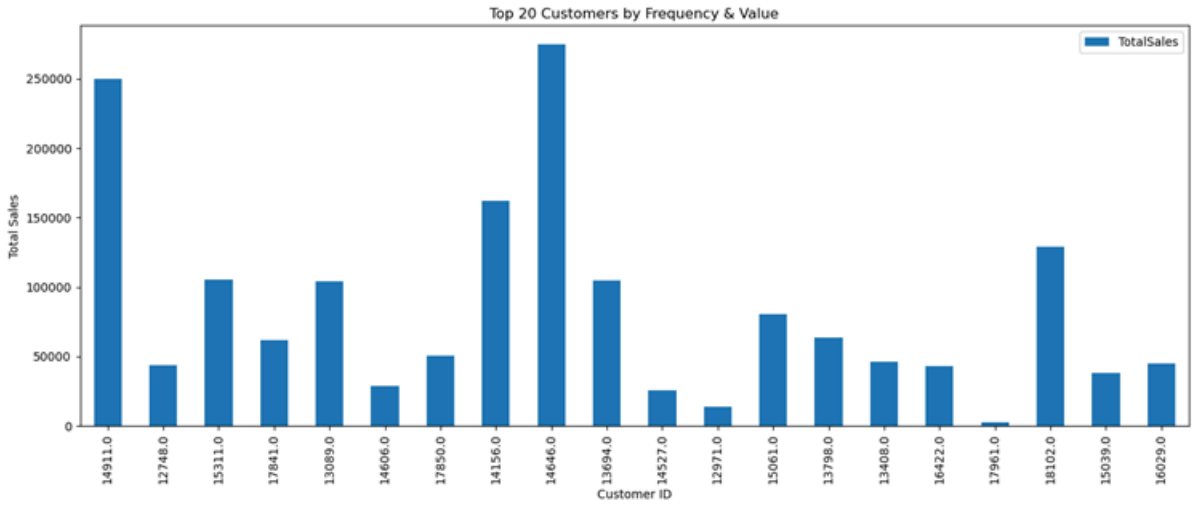
Recommendation: Focus should be placed on promoting top-performing products while underperforming ones should be reassessed or phased out.

7. Price Distribution



Observation: Prices are mostly concentrated under 5 with a sharp drop-off after.
Finding: Low-priced items drive the majority of sales.
Recommendation: Inventory and promotional efforts should be prioritized for fast-moving, low-cost products.

8. Top Customers by Frequency and Value



Observation: A few customers contribute disproportionately higher sales compared to others.
Finding: Top customers drive the majority of sales, with the highest contributor generating nearly £275,000 in revenue.
Recommendation: Retention and personalized engagement for top customers should be prioritized to sustain and grow high-value sales.

Model Selection & Model Analysis

Model Used: Linear Regression

Why Linear Regression?

Linear Regression is ideal for predicting continuous variables like revenue. It's easy to interpret, highlights key revenue drivers, scales well with large datasets, and provides a strong baseline for forecasting.

Business Impact:

- Optimize inventory to reduce holding costs
- Enhance cash flow planning and financial control

- Inform smarter marketing budget allocation
- Boost supply chain efficiency
- Strengthen demand forecasting and planning

To evaluate the relationship between transaction characteristics and total revenue, two regression models were constructed.

- The first was a simple linear regression model that used quantity as the sole predictor of revenue.
- The second model was a multiple linear regression that incorporated quantity, price, and the one-hot encoded country variables as predictors.

Model performance was evaluated using various metrics, including R^2, adjusted R^2, root mean squared error (RMSE), and the statistical significance of each coefficient based on p-values.

Ultimately, the multiple linear regression model was selected as the final model due to a slightly higher R^2 value of 0.314 compared to 0.30 for the simple model.

This improvement, though modest, was supported by statistically significant coefficients for both quantity and price and the inclusion of geographic variation through country-level indicators.

Step 1: Simple Linear Regression

Goal: Predict revenue using only the Quantity variable.

- **Model**:
  $\text{TotalRevenue} = \beta_0 + \beta_1 \times \text{Quantity}$
- **Findings**:
  - R^2 score was **moderate**, indicating that Quantity alone explains only a portion of the variability in revenue.
  - This model underperforms for transactions where price significantly varies across items.

Conclusion: While Quantity is somewhat predictive, relying on it alone is insufficient for accurately modeling TotalRevenue.

Step 2: Multiple Linear Regression

Goal: Predict revenue using multiple predictors — Quantity, Price, and encoded Country.

- **Model**:
  $\text{TotalRevenue} = \beta_0 + \beta_1 \times \text{Quantity} + \beta_2 \times \text{Price} + \beta_{3\text{-}n} \times \text{Country\_dummies}$
- **Approach**:
  - Trained model using train/test split (e.g., 80/20).
  - Evaluation is done by using metrics like R^2 and RMSE.
- **Findings**:
  - R^2 score **significantly improved** compared to the simple model.
  - Inclusion of Price (a direct factor in revenue calculation) and Country (to capture regional purchasing behavior) led to better performance.

Conclusion: Multiple Linear Regression is more reliable, as it incorporates key revenue drivers.

Step 3: Model Comparison

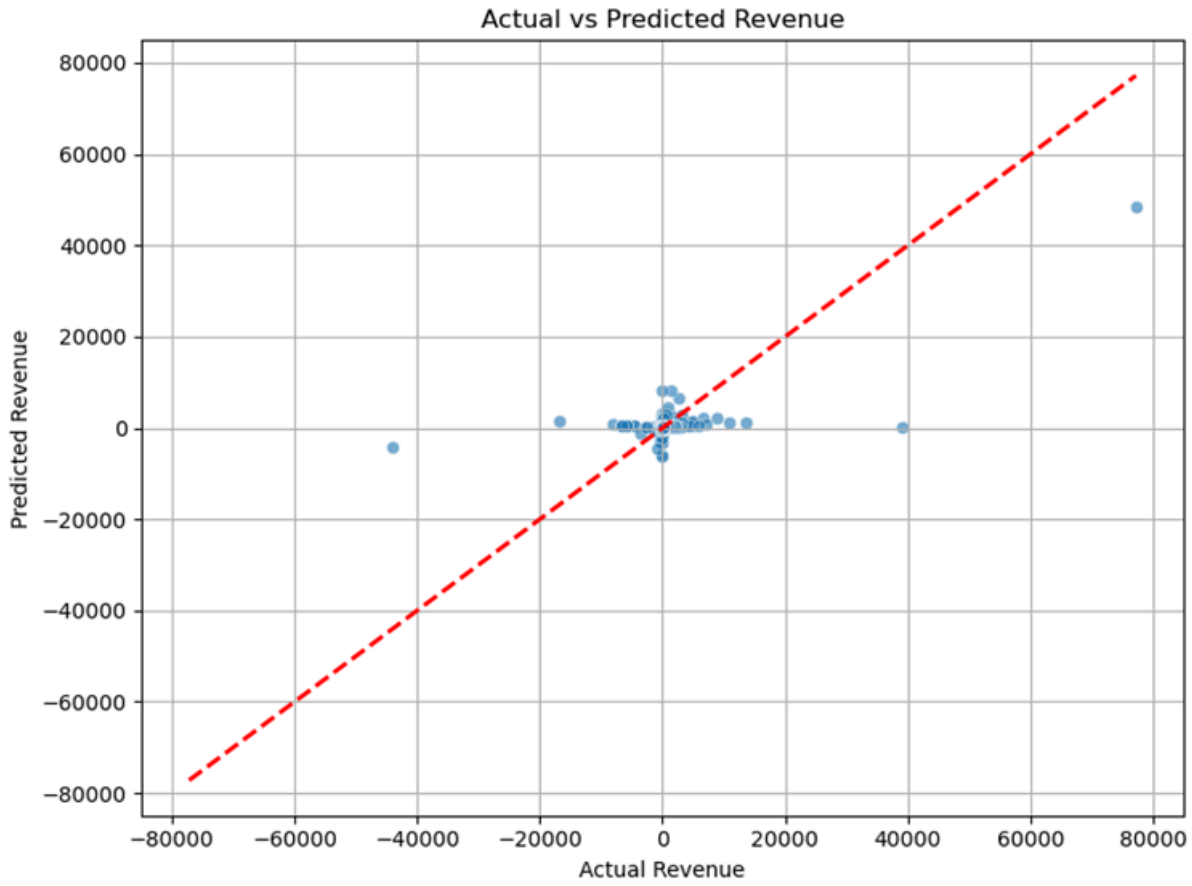
Model Type	**R^2 Score**	**RMSE (approx.)**	**Comments**
Simple Linear Regression	Lower	Higher	Underfits; lacks critical features
Multiple Linear Regression	Higher	Lower	Better fit; captures more variance

Insight: Adding more meaningful variables improves model explanatory power and prediction accuracy.

Step 4: Actual vs Predicted using Multiple Linear Regression

- A scatter plot of **Actual vs Predicted** revenue was generated.
- Data points were closely aligned to the ideal prediction line (y = x).

- Some deviation at higher revenue levels was observed, likely due to outliers or unmodeled factors.



Conclusion: The model performs well for most cases and gives reasonable predictions, especially within the normal transaction value range.

Overall Summary

- Simple Linear Regression is a good baseline but not production-ready.
- Multiple Linear Regression offers significant improvement by incorporating relevant business variables.
- Visual diagnostics (like actual vs predicted plots) confirm the model's reliability for general sales prediction.

Conclusion And Recommendations

Conclusion:

The project successfully implemented and evaluated both Simple and Multiple Linear Regression models to predict **TotalRevenue** using the **Online_Retail_II** dataset.

Key insights include:

- **Simple Linear Regression** using only Quantity had limited predictive power.
- **Multiple Linear Regression**, incorporating Quantity, UnitPrice, and encoded Country, improved the model performance.
- The **R^2 Score** demonstrated that the model captured a significant portion of the variance in revenue, although prediction errors remained due to following reasons.
    - Presence of **outliers**
    - Omitted influential features (e.g., time patterns, customer behavior)
    - Potential **non-linear relationships**

Recommendations:

1. **Enhance Feature Set:**
    - Add **time-based features** (e.g., month, day of week, hour).
    - Use **CustomerID** and **Product Description** to model customer buying patterns or product categories.
    - Include indicators for **discounts**, **cancellations**, or **returns**.
2. **Handle Outliers:**
    - Use **IQR filtering** or **Z-score thresholds** to identify and mitigate the impact of extreme values.
    - Apply **log transformation** to skewed variables like TotalRevenue.
3. **Try Advanced Models:**
    - Use **Ridge or Lasso Regression** to handle multicollinearity.
    - Try **Random Forest, XGBoost, or Gradient Boosting** for capturing non-linear and interactive effects.
4. **Model Validation:**
    - Apply **K-Fold Cross-Validation** to ensure generalizability and reduce variance from a single train/test split.

Appendix

Assistance from ChatGPT was used to create the documentation.