**Online Sales Data Analysis Report**

**Project Overview**

This report presents an end-to-end analysis of an **Online Sales Dataset** containing 50,000 rows. The goal is to clean the data, analyze key trends, visualize insights, segment customers, and recommend strategies for business improvement.

**1. Data Cleaning and Preparation**

**1.1 Handled Missing Values**

* **CustomerID**: Missing values were replaced with 'Unknown' to retain orders without customer identification.
* **ShippingCost**: Null values filled with the **median** shipping cost of the dataset.

**1.2 Removed Duplicates**

* Duplicates based on InvoiceNo and StockCode were dropped (keeping the first occurrence).

**1.3 Standardized Categorical Data**

* Corrected typo in **PaymentMethod** ('paypall' → 'paypal').
* Normalized the **ReturnStatus** column to title case (e.g., Not Returned, Returned).

**1.4 Added Revenue Features**

* Revenue = Quantity × UnitPrice
* Net Revenue = Revenue × (1 - Discount)

**1.5 Filtered Invalid Transactions**

* Excluded rows where:
  + Quantity ≤ 0
  + UnitPrice ≤ 0

**2. Exploratory Data Analysis (EDA)**

**2.1 Top-Performing Products**

* **Headphones** and **Office Chairs** lead in **Net Revenue**.
* **White Mugs** show consistently high quantities sold.

**2.2 High Revenue Regions**

| **Country** | **Total Net Revenue** |
| --- | --- |
| **Australia** | Highest |
| **Spain** | High |
| **Germany** | High |

**2.3 Sales Channel Insights**

* **Online Sales** significantly outperform **In-Store Sales** in terms of revenue and volume.

**2.4 Seasonal Trends**

* Stronger sales during **Q4 months** (Oct - Dec).
* Sales dip in **February** and **August**, suggesting off-season months.

**2.5 Customer Segmentation (RFM Analysis)**

* **Segment 0** (High Recency, Frequency, and Monetary): Loyal, high-value customers.
* **Segments 1 & 2**: Infrequent or low-spending customers. Opportunities exist for re-engagement.

**3.Visualizations**

**3.1 Sales Trend Over Time**

*“Monthly Net Revenue shows an upward trend in Q4, with peaks in November and December.”*

*(Replace with your actual plot)*

**3.2 Top 10 Products by Net Revenue**

*“Headphones and Office Chairs dominate product sales.”*

**3.3 Customer Segmentation Scatter Plot (RFM Segments)**

*“Segment 0 customers are highly valuable based on their recency, frequency, and monetary scores.”*

**3.4 Sales by Country**

*“Australia leads total sales, followed by Spain and Germany.”*

**4. Insights and Recommendations**

**4.1 Increase Sales During Low-Performing Months**

* **Promotional Campaigns**: Run discounts or limited-time offers in **February** and **August**.
* **Product Bundling**: Encourage bundled purchases (e.g., Headphones + USB Cable).

**4.2 Retain High-Value Customers**

* **Loyalty Programs**: Offer reward points and exclusive discounts for **Segment 0** customers.
* **Personalized Engagement**: Use customer behavior data to personalize emails and product suggestions.

**4.3 Reduce Product Returns**

* **Detailed Product Descriptions**: Provide comprehensive specs, videos, and reviews, especially for electronics.
* **Enhanced Quality Control**: Focus on quality improvements for high-return products.

**4.4 Optimize Sales Channels**

* **Boost In-Store Promotions** in regions with low online penetration.
* **Leverage Online Channels** for wider product range and targeted marketing.

**5. Sales Forecasting**

**Forecast Model: ARIMA (1,1,1)**

| **Forecast Month** | **Predicted Net Revenue** |
| --- | --- |
| Next Month | $**XXXX.XX** |
| +2 Months | $**XXXX.XX** |
| +3 Months | $**XXXX.XX** |

*(Replace XXXX.XX with the actual forecast results from your ARIMA model)*