Online Retail Dataset - Exploratory Data Analysis Project

Author: Shaik Sohail
Tools Used: Python, Pandas, Matplotlib, Seaborn

Introduction

This project analyzes the Online Retail dataset, which contains transactional data from a UK-based e-commerce store. The goal is to explore sales trends, customer behavior, and business opportunities using Exploratory Data Analysis (EDA).

Objective

- Find top-selling products
- Identify top revenue-generating countries
- Understand customer purchasing patterns
- Provide actionable business insights

Dataset Description

The dataset contains 541,909 rows and 8 columns, including InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country.

Data Cleaning & Preparation

- Removed rows with missing CustomerID
- Removed negative or zero Quantity and UnitPrice
- Converted InvoiceDate to datetime format
- Added new column: TotalPrice = Quantity × UnitPrice
- Final cleaned dataset: 397,884 rows

Python Code Snippets (Data Cleaning)

```
# Drop rows with missing CustomerID
df = df.dropna(subset=["CustomerID"])

# Remove negative/zero quantities and prices
df = df[df["Quantity"] > 0]
df = df[df["UnitPrice"] > 0]

# Convert InvoiceDate to datetime
df["InvoiceDate"] = pd.to_datetime(df["InvoiceDate"])
```

```
# Create new column: TotalPrice
df["TotalPrice"] = df["Quantity"] * df["UnitPrice"]

# Reset index after cleaning
df = df.reset_index(drop=True)
```

Python Code Snippets (EDA)

```
# Top 10 products by quantity sold
top_products = df.groupby("Description")["Quantity"].sum().sort_values(ascending=False).head(10)

plt.figure(figsize=(10,6))
sns.barplot(x=top_products.values, y=top_products.index, palette="viridis")
plt.title("Top 10 Products by Quantity Sold")
plt.show()
```

Business Insights

- A small number of products contribute to the majority of sales.
- United Kingdom generates the highest revenue.
- Certain inexpensive products are sold in very large volumes.
- Outliers exist with unusually high transaction values.

Conclusion

This analysis highlights key products and markets driving revenue. Businesses can optimize inventory management and marketing strategies based on these findings.