E-commerce Data Analysis Project

Overview

This project involves analyzing an e-commerce dataset to extract insights and perform various data processing tasks. The dataset contains transaction records including invoice details, stock information, customer data, and country-specific sales data. The analysis includes data cleaning, exploration, visualization, and deriving key metrics to understand sales trends and customer behavior.

Project Structure

The project is organized in a Jupyter Notebook (E commerce.ipynb) and follows a structured approach:

1. **Data Loading and Initial Exploration**
    - Load the dataset from a CSV file.
    - Display the first few rows to understand the structure.
    - Check basic information about the dataset (data types, missing values, etc.).
2. **Data Cleaning**
    - Handle missing values in columns like StockCode,Description and CustomerID.
    - Remove duplicate records to ensure data integrity.
    - Convert data types where necessary (e.g., StockCode to numeric).
3. **Feature Engineering**
    - Create a new column Revenue by multiplying Quantity and UnitPrice.
    - Extract the month from the InvoiceDate for time-based analysis.
4. **Exploratory Data Analysis (EDA)**
    - Visualize the distribution of Quantity using a boxplot and histogram.
    - Identify and analyze outliers in the Quantity column.
    - Summarize sales by country to understand geographical trends.
    - Analyze monthly revenue trends using a bar chart.
5. **Key Insights**

    - The dataset contains transaction records from multiple countries, with the United Kingdom being the dominant market.
    - Outliers in the `Quantity` column were identified and analyzed.
    - Monthly revenue trends show fluctuations, which can be further investigated for seasonal patterns.

Tools and Libraries Used

- **Python**: Primary programming language.
- **Pandas**: For data manipulation and analysis.
- **Matplotlib and Seaborn**: For data visualization.
- **Scikit-learn**: For machine learning and data preprocessing (though not extensively used in this analysis).

Dataset Description

The dataset includes the following columns:

- `InvoiceNo`: Unique identifier for each invoice.
- `StockCode`: Code for the stock item.
- `Description`: Description of the stock item.
- `Quantity`: Quantity of the item purchased.
- `InvoiceDate`: Date and time of the invoice.
- `UnitPrice`: Price per unit of the item.
- `CustomerID`: Unique identifier for the customer.
- `Country`: Country where the transaction occurred.

How to Run the Project

1. Ensure you have Python installed (preferably Python 3.6 or higher).
2. Install the required libraries using `pip install pandas matplotlib seaborn scikit-learn`.
3. Open the Jupyter Notebook (`E commerce.ipynb`) and run each cell sequentially to reproduce the analysis.

Conclusion

This project provides a comprehensive analysis of an e-commerce dataset, highlighting key trends and metrics. The insights derived can be used to make data-driven decisions to optimize sales strategies and improve customer engagement.