

Exercise: Customer Behavior Analysis

As a newly hired Data Scientist, your task is to analyze the customer purchasing behavior of an online retail company. This exercise involves cleaning the data, creating basic features, and performing some exploratory data analysis (EDA) to help the company better understand its customers and their purchasing patterns.

Dataset

Use the [Online Retail Dataset](#), a transactional dataset from a UK-based online retailer from December 2010 to December 2011.

Instructions

1. Data Cleaning

- **Inspect the dataset:**
 - Load the dataset and check for missing values in critical columns (e.g., `CustomerID`, `Quantity`, `UnitPrice`).
 - Identify and remove any duplicate transactions.
 - Handle missing or invalid values in `Quantity` or `UnitPrice` (e.g., remove rows where `Quantity` is 0 or negative).
 - **Handle missing `CustomerID`:**
 - For transactions where `CustomerID` is missing, label them as "guest".
 - **Save the cleaned dataset** as `cleaned_online_retail.csv`.
-

2. Feature Engineering

- **Calculate a `TotalPrice` column:**
 - Multiply `Quantity` by `UnitPrice` for each transaction.
 - **Create additional time-based features:**
 - Extract the `Month` and `Day of Week` from `InvoiceDate` to understand sales trends.
 - Create a binary feature for transactions that occur during **peak hours** (e.g., between 9 AM and 6 PM).
-

3. Exploratory Data Analysis (EDA)

- **Sales Trends:**

- Plot the total sales (**TotalPrice**) over time, aggregated by **Month** and **Day of Week**.
 - Identify trends or seasonal patterns. Are there any noticeable peaks?
 - **Customer Spending:**
 - Find the top 10 customers by total spending (**TotalPrice**).
 - Create a bar chart to visualize the distribution of total spending across customers.
 - **Average Transaction Value:**
 - Calculate the average value of a transaction and visualize its distribution.
 - Are there any extremely high or low values?
-

4. Insights and Recommendations

Write a brief report (1-2 paragraphs) summarizing your findings from the EDA:

- What insights did you gain about sales patterns (e.g., seasonality, peak hours)?
 - Which customers are the most valuable?
 - Based on your analysis, propose one marketing or business strategy to improve sales or customer retention.
-

Deliverables

1. **Cleaned Dataset:** `cleaned_online_retail.csv`.
2. **Feature-Engineered Dataset:** `feature_engineered_online_retail.csv` (including **TotalPrice**, **Month**, **Day of Week**, and peak hour feature).
3. **EDA Visualizations:** Include in the notebook (graphs for sales trends, spending distribution, etc.).
4. **Report:** A brief summary with insights and recommendations.