Supermarket Sales Analysis

Overview

This project analyzes a supermarket sales dataset to extract insights about customer purchasing behavior, product performance, and overall sales trends. The analysis is conducted using Python and various data science libraries.

Dataset

- **Source:** `Python Project Data - Supermarket Sales.csv`
- **Size:** 1006 entries, 16 columns
- **Main Attributes:**
  - Invoice ID, Branch, City (Yangon, Naypyitaw, Mandalay)
  - Customer Type, Gender, Product Line
  - Unit Price, Quantity, Tax (5%), Total
  - Date, Time, Payment Method, Rating

Data Processing Steps

1. **Loading the Data:**
   - Used `pandas` to read the dataset.
   - Displayed sample data with `df.head()`.
2. **Handling Missing Values:**
   - Identified missing values using `df.isnull().sum()`.
   - Used forward fill (`df.fillna(method='ffill')`) and linear interpolation to fill missing values.
3. **Data Cleaning:**
   - Removed duplicates using `df.drop_duplicates(inplace=True)`.
   - Standardized column names (lowercase, underscores instead of spaces).
   - Encoded categorical variables using one-hot encoding.

Exploratory Data Analysis (EDA)

- **Summary Statistics:**
  - Used `df.describe()` for numeric summary.
  - Checked data types with `df.info()`.
- **Visualizations:**
  - **Gender Distribution:** Horizontal bar chart.
  - **Product Line Analysis:** Bar chart to compare product popularity.
  - **Sales vs. Tax:** Scatter plot.

    - **Total Sales by Gender:** Boxplot.

Insights

- The dataset contains three supermarket branches located in Yangon, Naypyitaw, and Mandalay.
- Most customers are either "Normal" or "Member" types.
- Products related to "Health & Beauty" and "Food & Beverages" are among the most popular.
- The majority of transactions are paid using E-wallets and credit cards.
- Certain records had missing values in the "Tax 5%" and "Total" columns, which were handled during data cleaning.

Technologies Used

- Python
- Pandas
- NumPy
- Seaborn
- Matplotlib

How to Run

Clone this repository:

git clone https://github.com/yourusername/supermarket-sales-analysis.git

Install dependencies:
pip install pandas numpy seaborn matplotlib

Open and run the Jupyter Notebook or Colab link provided:

Colab Notebook

Next Steps

- Perform predictive analysis using machine learning.
- Develop a dashboard for real-time sales tracking.
- Expand analysis to compare sales performance across different time periods.

Contact

For any inquiries, feel free to reach out at Mohrael Avraim