Supermarket Sales Analysis

Overview

This project focuses on analyzing a supermarket sales dataset to uncover insights into customer purchasing behavior, product performance, and sales trends. The analysis is conducted using Python and its data science libraries, with a structured approach to data cleaning, exploration, and visualization.

Dataset Details

- **Source**: Supermarket Sales Dataset
- **Size**: 1000entries, 16 columns
- **Key Attributes**:
  - Branch & City: Three branches located in Yangon, Naypyitaw, and Mandalay.
  - Customer Type: Normal or Member.
  - Product Line: Categories like Health & Beauty, Food & Beverages, etc.
  - Sales Details: Unit Price, Quantity, Tax (5%), Total.
  - Transaction Info: Date, Time, Payment Method, and Customer Ratings.

Data Processing Steps

1. **Loading the Data**:
   - Used `pandas` to load the dataset and previewed it with `df.head()`.

2. **Handling Missing Values**:
   - Checked for missing values using `df.isnull().sum()`.
   - Filled missing values using forward fill (`df.fillna(method='ffill')`) and linear interpolation.

3. **Data Cleaning**:
   - Removed duplicate entries with `df.drop_duplicates(inplace=True)`.
   - Standardized column names (e.g., lowercase, underscores instead of spaces).
   - Encoded categorical variables using one-hot encoding for analysis.

Exploratory Data Analysis (**EDA**)

1. **Summary Statistics**:
   - Used `df.describe()` to generate numeric summaries.
   - Verified data types with `df.info()`.

2. Visualizations:
   - **Gender Distribution**: Horizontal bar chart to show the proportion of male and female customers.
   - **Product Line Analysis**: Bar chart to compare the popularity of different product categories.
   - **Sales vs**. **Tax**: Scatter plot to visualize the relationship between sales and tax.
   - **Total Sales by Gender**: Boxplot to compare sales distribution across genders.

Key Insights
- The dataset includes sales data from three branches in Yangon, Naypyitaw, and Mandalay.
- Customers are categorized as either "Normal" or "Member."
- Popular Product Lines: Health & Beauty and Food & Beverages are the most purchased categories.
- Payment Methods: Most transactions are completed using E-wallets and credit cards.
- Missing values in the "Tax 5%" and "Total" columns were successfully handled during data cleaning.

Technologies Used
- Python: For data manipulation and analysis.
- Pandas: For data cleaning and processing.
- NumPy: For numerical computations.
- Seaborn & Matplotlib: For creating visualizations.

How to Run the Project
1. **Clone the repository**:
   git clone https://github.com/nourhanfarag1610 /supermarket-sales-analysis.git

2. **Install the required dependencies**:
    pip install pandas numpy seaborn matplotlib

3. **Open and run the Jupyter Notebook or Colab link provided in the repository**.
https://colab.research.google.com/

Next Steps

- Predictive Analysis: Use machine learning models to predict future sales trends.
- Dashboard Development: Create an interactive dashboard for real-time sales tracking.
- Time-Based Analysis: Compare sales performance across different time periods to identify seasonal trends.

Contact

For inquiries, feel free to reach out to [Nourhan Farag] at [norahassan4420@gmail.com]