Budget Spending Data Analysis Project Report

1. Introduction

This report outlines the process and key findings from an exploratory data analysis (EDA) project focused on budget spending data. The primary goal was to load, clean, analyze, and visualize spending patterns from two provided datasets: `Budget_Spending_Data.csv` (raw data) and `Cleaned_Budget_Spending_Data.csv` (pre-processed data).

2. Project Objectives

The main objectives of this project were:

- To load and inspect both the raw and cleaned budget spending datasets.
- To perform initial data quality checks, including identifying missing values and data types.
- To conduct basic descriptive statistical analysis.
- To visualize spending distributions and trends.
- To identify top and bottom spending categories.
- To compare the characteristics of the raw versus the cleaned dataset.

3. Data Sources

Two primary CSV files were utilized for this analysis:

- `Budget_Spending_Data.csv`: This dataset represents the initial, raw budget spending information.
- `Cleaned_Budget_Spending_Data.csv`: This dataset is a pre-processed version of the budget spending data, presumably with some cleaning steps already applied.

4. Methodology

The analysis was conducted in a Jupyter Notebook environment using Python libraries such as `pandas` for data manipulation, `matplotlib.pyplot` for basic plotting, and `seaborn` for enhanced statistical visualizations. The systematic steps involved:

1. **Library Imports:** Importing essential libraries (`pandas`, `matplotlib.pyplot`, `seaborn`).
2. **Data Loading:** Both CSV files were loaded into pandas DataFrames (`df_raw` and `df_cleaned`). Error handling was implemented for file loading.
3. **Initial Data Inspection:** `df.head()`, `df.info()`, `df.describe()`, and `df.isnull().sum()` were used to understand the structure, data types, and presence of missing values in both datasets.
4. **Data Cleaning & Preparation:**
    - Columns identified as 'Amount' and 'Date' were converted to numeric and datetime types, respectively. Non-convertible values and associated rows were handled (dropped).
    - Duplicate rows were identified and removed from the cleaned dataset.
5. **Distribution Analysis:** Histograms and box plots were generated to understand the distribution of spending amounts and identify potential outliers. Box plots were also used to visualize spending distribution across different categories.
6. **Categorical Analysis:** Total spending was aggregated by 'Category' to identify top and bottom spending areas. Pie charts were used to illustrate the proportion of spending in the top categories.
7. **Time Series Analysis:** If a 'Date' column was available, monthly and quarterly spending trends were visualized using line plots and stacked bar charts to show category-wise spending over time.
8. **Data Comparison (Raw vs. Cleaned):** A comparative analysis was performed to highlight differences in total spending, number of records, and unique categories between the raw and cleaned datasets.
9. **Summary and Export:** A final summary of the cleaned dataset was provided, with an option to export the processed data to a new CSV file.

5. Key Findings and Insights

(***Please fill in your specific findings here after running all the code cells and reviewing their outputs.***)

- **Data Quality:**
    - [E.g., "The raw dataset had X% missing values in [column name] and required cleaning for [specific issue]."]
    - [E.g., "The cleaned dataset showed significantly fewer missing values and consistent data types for analysis."]
- **Spending Distribution:**
    - [E.g., "The distribution of spending amounts was right-skewed, indicating many small transactions and a few large ones."]
    - [E.g., "Outliers in spending were observed, potentially indicating unusually high expenses that might warrant further investigation."]
- **Top/Bottom Categories:**

    - [E.g., "The top 3 spending categories were [Category 1], [Category 2], and [Category 3], accounting for X% of total expenditure."]
    - [E.g., "Discretionary spending categories like 'Entertainment' showed high variability."]
- **Spending Trends Over Time:**
    - [E.g., "Monthly spending showed a clear upward trend from Q1 to Q3, possibly due to [reason]."]
    - [E.g., "There was a noticeable spike in spending during [month/quarter], predominantly driven by expenses in [Category Name]."]
- **Raw vs. Cleaned Data Comparison:**
    - [E.g., "The cleaning process resulted in a reduction of X rows, primarily due to duplicate removal and invalid data entries."]
    - [E.g., "Total spending in the cleaned data was slightly lower/higher than the raw data due to [reason, e.g., removal of erroneous entries]."]

6. Visualizations

The following types of visualizations were generated to support the analysis:

- **Histograms:** To show the frequency distribution of spending amounts.
- **Box Plots:** To illustrate the spread and outliers of spending amounts, both overall and by category.
- **Bar Charts:** To compare total spending across different categories.
- **Pie Charts:** To represent the proportion of spending in top categories.
- **Line Plots:** To depict monthly spending trends over time.
- **Stacked Area/Bar Charts:** To show how spending within different categories changed over time (monthly/quarterly).
- **Comparison Bar Chart:** To visually compare total spending between raw and cleaned datasets.

7. Conclusion and Next Steps

This project successfully provided a foundational analysis of the budget spending data. The systematic approach allowed for data inspection, cleaning, and the extraction of preliminary insights into spending patterns.

(***Please fill in your concluding remarks and potential next steps here.***)

- **Conclusion:** [E.g., "The cleaned dataset is now robust for further detailed financial planning and anomaly detection."]
- **Potential Next Steps:**
    - Deeper dive into specific categories with unusual spending patterns.
    - Integration with budget targets for variance analysis.

- Development of predictive models for future spending.
- Interactive dashboard creation for real-time monitoring.
- Further investigation of specific outliers or anomalies identified.