DA Exercise Importing a Dataset - Advanced Implementation

Identifying and Importing a Dataset - Advanced Implementation

Task Overview

As a continuation of the previous example on data integration and summarization, this exercise focuses on finding and working with a real dataset.

Task Steps:

1. Find a Real Dataset

- Search online for a **real dataset** related to one of the following topics:
  - Business (e.g., sales, customer transactions, market trends)
  - Health (e.g., patient records, disease trends, hospital data)
  - Environment (e.g., climate change, air pollution, water quality)
  - Education (e.g., student performance, school rankings, university data)
  - Social Media (e.g., tweets, engagement metrics, trending topics)
  - Finance (e.g., stock prices, cryptocurrency, banking trends)
- The dataset should include **structured data** with at least one categorical variable and multiple numerical attributes.
- Justify why you chose the dataset and how it relates to real-world decision-making.

2. Load and Explore the Dataset

- Download the dataset and **load it into Python using Pandas**.
- Perform initial data exploration using:
  - `df.head()` – Display the first few rows.
  - `df.info()` – Check dataset structure, column types, and missing values.
  - `df.describe()` – Generate summary statistics.
- If multiple datasets are available, **merge them using a common key** to integrate different sources of information. Example:

```
merged_df = pd.merge(df1, df2, on="common_column", how="inner")
```

3. Perform Data Cleaning

- Identify missing values and handle them (drop or impute).
- Check for duplicate records and remove if necessary.
- Standardize column names and data formats.

4. Conduct Exploratory Data Analysis (EDA)

- Calculate:
  - Summary statistics for key numerical variables.
  - Distribution of categorical variables.
  - Relationships between different features using correlation analysis.
- Use Pandas `groupby()` and aggregation functions to summarize the dataset.

5. Visualize Insights

- Create at least **three visualizations**:
  1. **Histogram of a key numerical variable** – To observe distribution trends.
  2. **Bar chart of a categorical variable's aggregated value** – To compare different categories.
  3. **Scatter plot of two numerical variables** – To identify relationships.
- Additional optional visualizations:
  - Box plot to analyze variation in key variables.
  - Heatmap to check correlations between numerical variables.

6. Present Findings

- **Explain dataset characteristics:**
  - The source of the dataset.
  - Number of records and columns.
  - Key observations from the data.
- **Summarize insights from EDA and visualizations.**
  - What trends are visible?
  - Are there any outliers or patterns?
  - What decisions or insights can be drawn from this data?

Submission Requirements:

- Submit your cleaned dataset (CSV format).
- Include Python code and visualizations in a Jupyter Notebook or a Python script.
- Provide a short summary (one page) explaining the dataset, key findings, and challenges encountered.

Bonus Task (Optional):

- Apply **feature engineering** by creating a new variable such as:
  - A new metric relevant to the dataset (e.g., risk score, performance index, trend indicator)
  - Aggregated values over time or categories.