Data Analysis Task for Students

Objective: The objective of this task is to explore a given dataset, apply preprocessing steps, analyze correlations, create data visualizations, and select important features. This exercise will help you better understand how to process and derive meaningful insights from data.

Instructions:

1. Dataset Overview

- Select a dataset of your choice. You can use any public dataset from platforms such as Kaggle, UCI Machine Learning Repository, or use the provided dataset.
- Briefly explore the dataset and summarize the basic information, such as the number of rows and columns, types of features (numerical, categorical), and any interesting observations.

2. Data Preprocessing

- **Handle Missing Values**: Identify any missing values in the dataset. Use techniques like mean/median imputation for numerical features or mode imputation for categorical features. Alternatively, decide whether rows or columns with many missing values should be dropped.
- **Data Normalization/Standardization**: Normalize or standardize the numerical features so that they have a similar scale, especially if they vary widely in magnitude.
- **Categorical Encoding**: Convert categorical features into numerical ones using techniques such as One-Hot Encoding or Label Encoding.

3. Correlation Analysis

- **Correlation Matrix**: Compute a correlation matrix for all the numerical features in the dataset. Visualize the correlation matrix using a heatmap to better understand relationships between variables.
- **Identify Strong Relationships**: Analyze the correlation matrix to identify pairs of variables that have a strong positive or negative correlation (e.g., |correlation| > 0.7). Discuss the potential implications of these relationships.

4. Data Visualization

- Create at least **3 different visualizations** that provide insights into the dataset. Examples include:
    - **Histogram**: For visualizing the distribution of numerical features.
    - **Box Plot**: To identify outliers in key features.
    - **Scatter Plot**: To visualize relationships between pairs of features with high correlation.
    - **Bar Chart**: To show the frequency distribution of categorical variables.
- Use appropriate labels, titles, and legends to make your visualizations clear and informative.

5. Feature Selection

- **Feature Importance using Correlation**: Based on the correlation analysis, identify features that may be redundant (e.g., features that have very high correlation with each other).
- **Feature Selection Technique**: Apply a feature selection method such as Recursive Feature Elimination (RFE), feature importance from a decision tree, or Principal Component Analysis (PCA) to determine the most important features.
- **Explain Your Choices**: Provide a short explanation for why you chose certain features over others for building a predictive model.

6. Summary and Report

- Summarize your findings in a short report. Discuss:
    - Any interesting relationships or insights discovered in the data.
    - The steps you took to preprocess the data and the rationale behind each step.
    - The features you selected and why you believe they are important.

Deliverables:

- Jupyter Notebook (or similar) with all the code for the data preprocessing, analysis, and visualizations.
- A short report (1-2 pages) summarizing your findings and explaining your data analysis workflow.

Note: Make sure to include comments in your code so that it is easy to follow, and ensure that all plots are well-labeled and easy to interpret.

Additional Tips:

- Feel free to use libraries such as Pandas, NumPy, Matplotlib, and Scikit-Learn to complete your analysis.
- Be creative in your visualizations and aim to present your insights in a clear and meaningful way.