Basic Exploratory Data Analysis (EDA) Assignment - ChatGPT Mini

The goal of this assignment is to perform a comprehensive Exploratory Data Analysis (EDA) on the Breast Cancer dataset using Python. You will use libraries such as NumPy, pandas, Seaborn, and Matplotlib to analyze the data, visualize key insights, and draw meaningful conclusions.

Dataset: Breast Cancer Dataset

The Breast Cancer dataset contains data on breast cancer cases, including features such as the size of the tumor, texture, perimeter, and number of mitoses. The target variable indicates whether the cancer is benign or malignant. This dataset is commonly used for classification problems and is excellent for practicing EDA.

To load the dataset, use the following code:

```python
from sklearn.datasets import load_breast_cancer
import pandas as pd

# Load the dataset
breast_cancer = load_breast_cancer()

# Create a DataFrame
df_cancer = pd.DataFrame(data=breast_cancer.data, columns=breast_cancer.feature_names)

# Add the target variable
df_cancer['target'] = breast_cancer.target

# Display the first few rows
df_cancer.head()
```

1. Data Loading and Exploration

1. Load the Breast Cancer dataset using pandas. Display the first five rows of the dataset. What are the features (columns) available in the dataset?

2. What is the shape of the dataset? How many rows and columns does it have?

3. Use the `info()` function to display the data types of each column. Are all columns of the appropriate data type for analysis?

4. Use the `describe()` function to display summary statistics for the numerical columns. What are the mean, median, and standard deviation of `mean radius`?

5. Are there any missing values in the dataset? If so, which columns contain them, and how many missing values are there?

2. Data Cleaning

1. If there are any missing values, decide whether to remove them or fill them with appropriate values. What approach did you choose, and why?

2. Ensure that the target variable (`target`) is correctly labeled (0 for malignant, 1 for benign). Are there any inconsistencies in the target variable?

3. Descriptive Statistics

1. Calculate and interpret the mean, median, and standard deviation for the `mean area` and `mean compactness`. What do these statistics tell you about the data?

2. Create a histogram for `mean radius`. What does the distribution look like? Is it skewed? If so, in which direction?

3. Use a boxplot to visualize the distribution of `mean texture`. Are there any outliers in the data? What is the interquartile range (IQR)?

4. Group the data by the `target` variable and calculate the mean for `mean perimeter`. Is there a noticeable difference between the mean perimeter of benign and malignant cases?

4. Data Visualization

1. Create a bar chart to show the number of benign vs. malignant cases. What does the chart tell you about the distribution of cases?

2. Visualize the relationship between `mean radius` and `mean texture` using a scatter plot. Do you notice any patterns or correlations between these two features?

3. Use a pairplot to visualize the relationship between all features and the `target` variable. What relationships or clusters can you identify from the pairplot?

4. Create a heatmap to visualize the correlation matrix of the dataset. Which features are most strongly correlated with each other? How might this affect your analysis?

5. Feature Engineering

1. Create a new feature called `area_to_perimeter_ratio` by dividing `mean area` by `mean perimeter`. How does this new feature differ between benign and malignant cases?

2. Analyze the impact of `area_to_perimeter_ratio` on the classification of cancer. Does this new feature provide any additional insights or separability between benign and malignant cases?

3. Are there other combinations of features that could be useful? Create at least one more derived feature and analyze its impact on the classification.

6. Conclusion

1. Summarize the key findings from your EDA. What were the most significant patterns or trends you observed?

2. Based on your analysis, which features seem to be the most important for distinguishing between benign and malignant cases?

3. What recommendations would you make for further analysis or model building using this dataset? Are there any areas where more data or different features would be beneficial?