Name - Manaswi Brijesh Shekokar
PRN No – 22310637
Roll No – 282024
Batch – B2
Year – SY B

Assignment No – 1

Problem Statement :
1. Perform the following operations using R/Python on suitable data sets:
a) read data from different formats (like csv, xls)
b) Find Shape of Data
c) Find Missing Values
d) Find data type of each column
e) Finding out Zero's
f) Indexing and selecting data, sort data,
g) Describe attributes of data, checking data types of each column,
h) counting unique values of data, format of each column, converting variable data type (e.g. from long to short, vice versa)

Objective
This code's main goal is to conduct preliminary preparation and exploration of the breast cancer dataset in order to comprehend its contents, structure, and statistical characteristics. This EDA is probably a prelude to additional analysis, including developing a machine learning model to categorize diagnoses of breast cancer as either benign or malignant.

Main Functions
The code utilizes functions from the pandas and numpy libraries to explore and manipulate the dataset. Below are the key functions used:

1. import pandas as pd and import numpy as np: Import the required libraries for data manipulation and numerical operations.
2. pd.read_csv('breast-cancer.csv'): Loads the dataset from a CSV file into a pandas DataFrame.
3. dataset.shape: Returns the dimensions of the dataset (rows, columns).
4. dataset.info(): Provides a summary of the dataset, including column names, data types, and non-null counts.
5. dataset.describe(): Generates descriptive statistics (count, mean, std, min, max, quartiles) for numerical columns.
6. dataset.isnull().sum(): Checks for missing values in each column.
7. dataset.iloc[]: Indexes and selects specific rows or columns (e.g., dataset.iloc[2:5] for rows 2 to 4, or dataset.iloc[:, :-1] for all columns except the last).
8. dataset.sort_values(by='column_name', ascending=True): Sorts the dataset based on a specified column.
9. dataset['column_name'].value_counts(): Counts the frequency of unique values in a specified column.
10. dataset['column_name'].astype(type): Converts the data type of a column (e.g., float or int).

Methodology
The code follows a structured approach to EDA:

1. **Library Import**: Import pandas and numpy to enable data manipulation and analysis.
2. **Data Loading**: Load the breast cancer dataset from a CSV file into a DataFrame.
3. **Dataset Overview**:
    - Check the shape to understand the dataset's size (569 rows, 33 columns).
    - Use info() to examine data types and missing values.

    - Use describe() to summarize numerical features statistically.
4. **Missing Value Check**: Identify null values in each column (e.g., Unnamed: 32 has 569 missing values).
5. **Data Selection and Indexing**: Extract specific rows (e.g., rows 2–4) or columns (e.g., all except the last) using iloc.
6. **Sorting**: Sort the dataset by columns like radius_mean, perimeter_mean, and concavity_mean to observe trends or outliers.
7. **Value Counting**: Analyze the distribution of values in columns like symmetry_mean and concave points_worst.
8. **Data Type Conversion**: Convert columns (area_worst to float, compactness_worst to int) to ensure compatibility with downstream tasks.

Advantages

1. **Comprehensive Exploration**: The code provides a thorough initial understanding of the dataset's structure, statistics, and missing values, which is crucial for any data-driven project.
2. **Ease of Use**: Utilizes pandas' intuitive functions, making it accessible for beginners and efficient for experienced users.
3. **Data Cleaning Insight**: Identifies missing values (e.g., all 569 entries in Unnamed: 32 are null), guiding decisions like dropping irrelevant columns.
4. **Flexibility**: Sorting and value counting allow for customizable analysis based on specific features of interest.
5. **Preprocessing Support**: Data type conversion prepares the dataset for machine learning algorithms that require specific input types.

Disadvantages

1. **Limited Visualization**: The code lacks graphical representations (e.g., histograms, scatter plots), which could enhance understanding of distributions and relationships between features.
2. **No Handling of Missing Values**: While it identifies null values (e.g., Unnamed: 32), it doesn't address them (e.g., by dropping or imputing), leaving this step incomplete.
3. **Basic Analysis**: The EDA is rudimentary and doesn't include advanced techniques like correlation analysis, outlier detection, or feature engineering.
4. **Data Type Conversion Issues**: Converting compactness_worst to int results in loss of precision (e.g., all values become 0 due to truncation), which could distort analysis or modeling.
5. **No Statistical Testing**: It doesn't perform hypothesis testing or statistical comparisons (e.g., between malignant and benign cases), limiting deeper insights.

Conclusion

This code serves as a foundational step in analyzing the breast cancer dataset, focusing on loading, inspecting, and basic manipulation. While it excels in providing a quick overview and identifying issues like missing data, it falls short in visualization, advanced analysis, and complete preprocessing. For a more robust analysis, additional steps like handling missing values, visualizing data, and computing feature correlations would be beneficial.