Assignment No 1

# Aim:

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

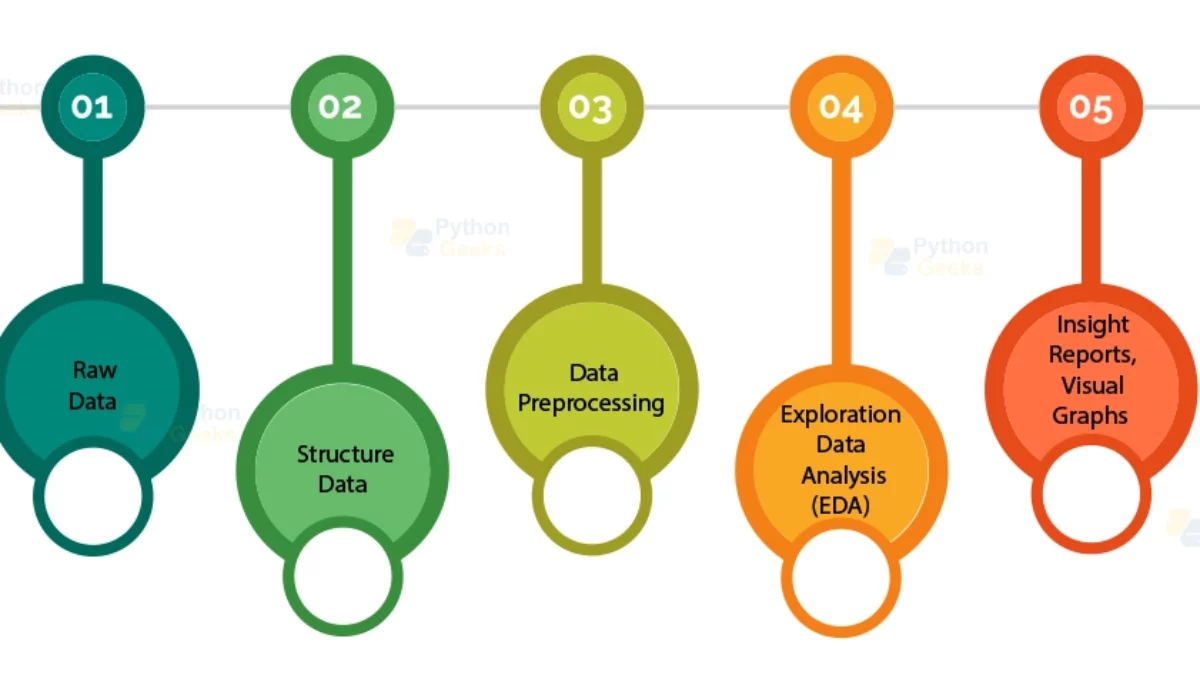
f) Indexing and selecting data, sort data,

g) Describe attributes of data, checking data types of each column,

h) counting unique values of data, format of each column, converting variable

# Theory:

What is Data Preprocessing?  
  
Data preprocessing is a crucial step in the data analysis pipeline. It includes cleaning, transforming, and organizing raw data into a usable format. Without proper preprocessing, data analysis or machine learning models may yield inaccurate results.  
  
Data is usually collected from multiple sources and stored in formats like CSV, XLS/XLSX, JSON, etc. Before conducting any statistical or machine learning task, we need to inspect, clean, and understand the data through exploratory techniques.



# Methods and Explanations of Operations

## a) Reading Data from Different Formats (CSV, XLS)

* - CSV (Comma-Separated Values): A plain text file format that stores tabular data.
* - XLS/XLSX: Excel spreadsheet files. Reading these formats requires additional libraries in Python like pandas with openpyxl or xlrd, and readxl in R.
* - Why Important: Helps bring external data into your working environment for processing and analysis.

## b) Find Shape of Data

* - Represents the number of rows and columns in a dataset.
* - Syntax: (rows, columns)
* - Why Important: Understanding the shape is the first step in knowing what you’re working with.

## c) Find Missing Values

* - Real-world datasets often have missing entries.
* - Missing values are typically represented as NaN, NA, or blanks.
* - Why Important: They can skew analysis, so you need to either fill, impute, or drop them.

## d) Find Data Type of Each Column

* - Helps in understanding how each column is interpreted: numerical, string, boolean, datetime, etc.
* - Why Important: Many operations are data-type sensitive. For instance, you can’t compute the mean of a text field.

## e) Finding Out Zeros

* - Zero values may or may not be significant.
* - Sometimes zeros indicate missing values or errors (especially in medical or survey data).
* - Why Important: Differentiating actual values from placeholders helps with data cleaning.

## f) Indexing and Selecting Data, Sorting Data

* - Indexing: Refers to accessing specific rows or columns.
* - Selection: Filter data based on condition.
* - Sorting: Reorder data based on column values (ascending/descending).
* - Why Important: Gives flexibility in analyzing data slices or ordering it for better visibility.

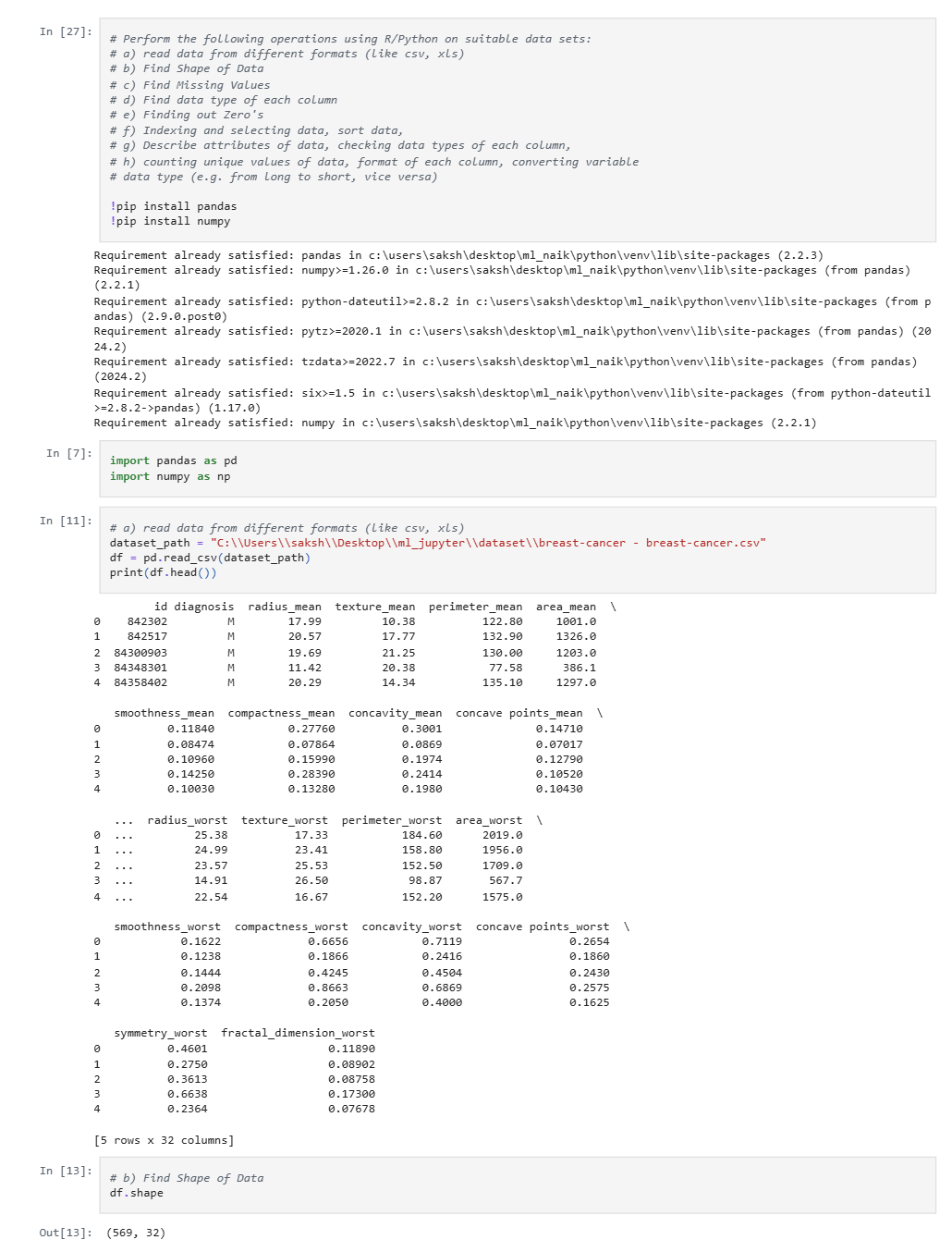
## g) Describe Attributes of Data, Checking Data Types of Each Column

* - Includes mean, median, min, max, standard deviation, quartiles.
* - Helps to get a statistical summary of each column.
* - Why Important: Quickly provides distribution and spread of numerical data, helping to identify outliers and trends.

## h) Counting Unique Values, Format of Each Column, Converting Variable Data Type

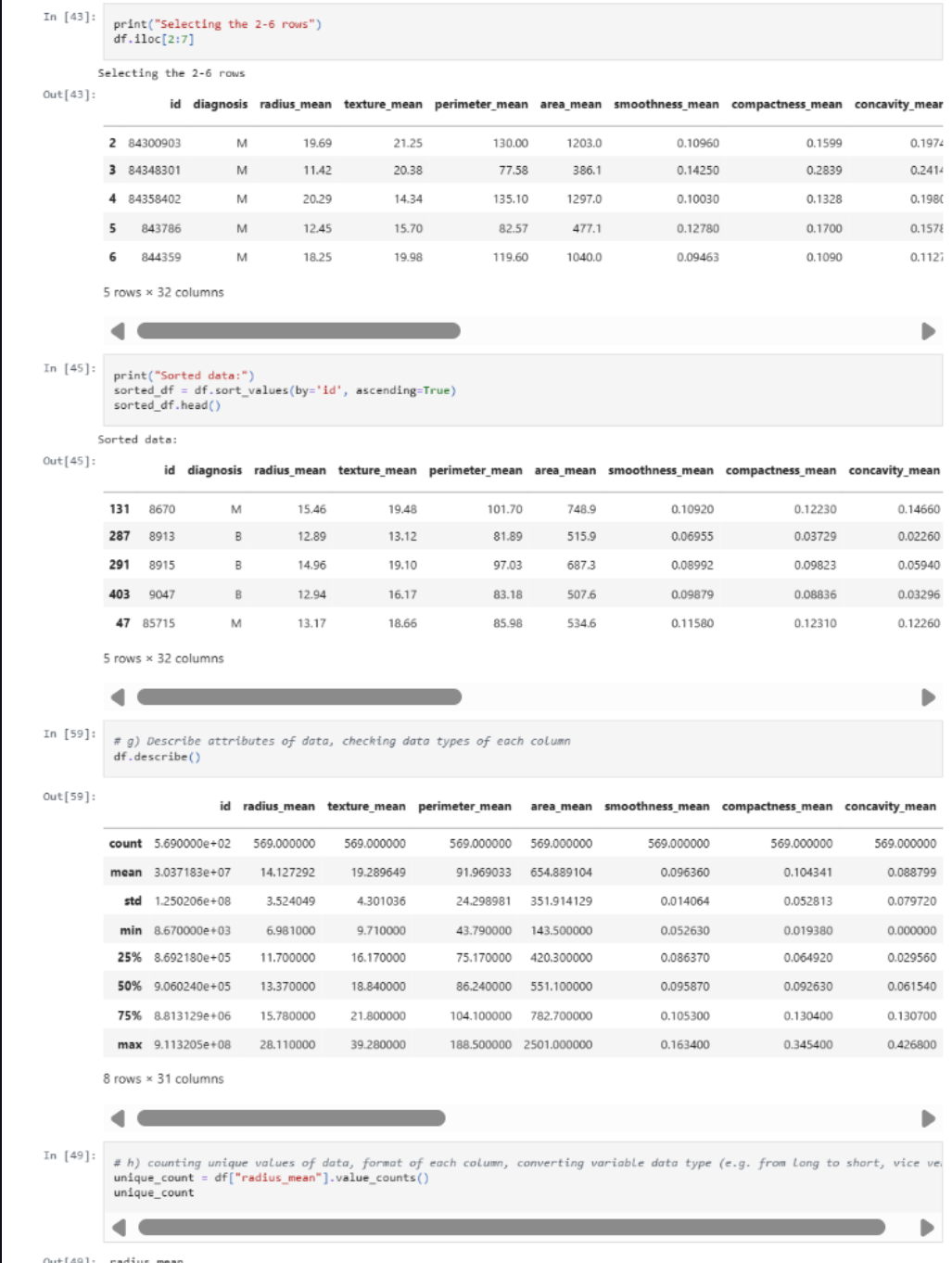
* - Unique Counts: Useful for categorical variables to know diversity.
* - Format: Consistency check (e.g., dates, strings).
* - Conversion: Sometimes needed for modeling (e.g., converting float to integer, or string to category).
* - Why Important: Improves model performance and ensures compatibility between tools and libraries.

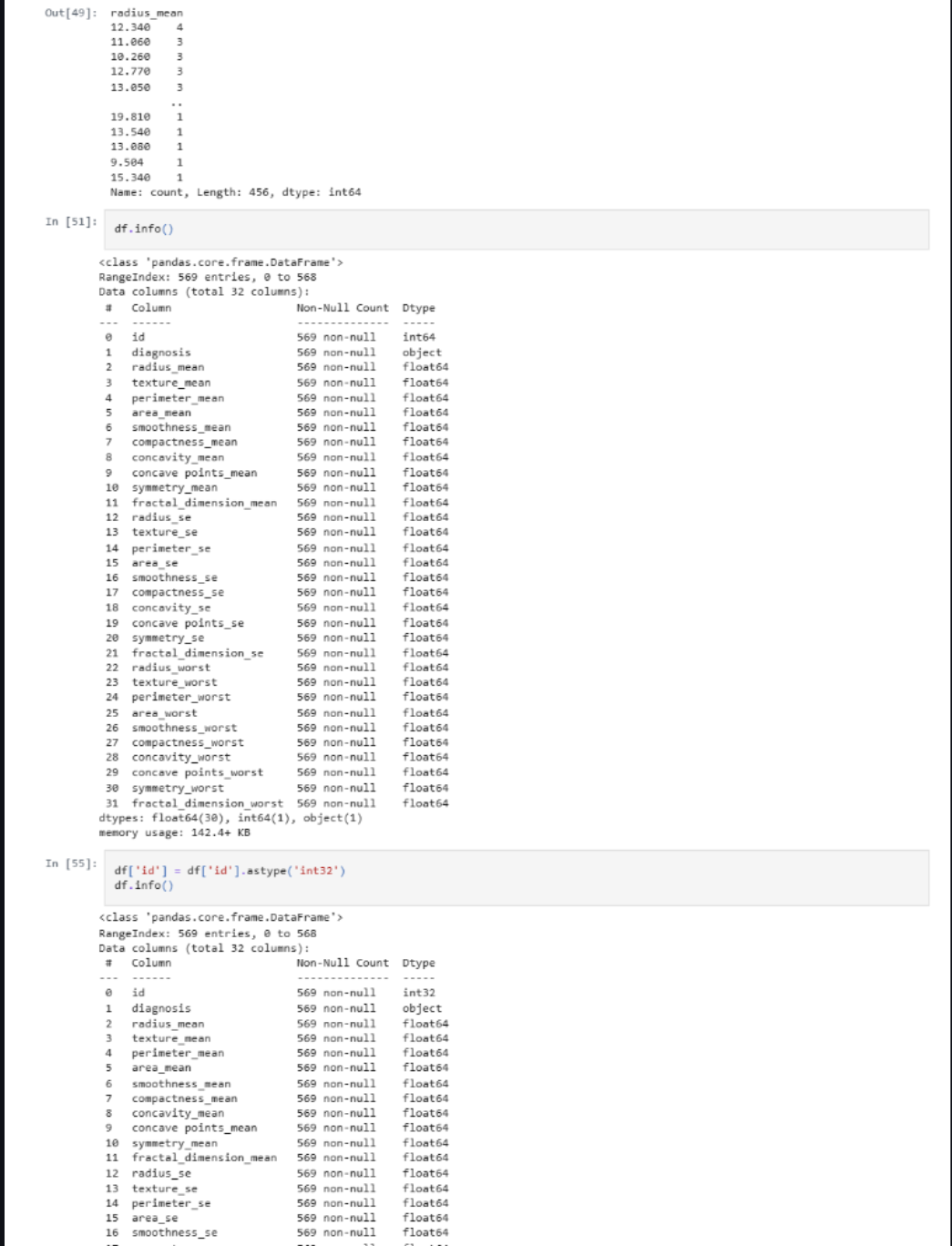
**Outputs:**

****

# 

# 





# Conclusion

We have successfully performed all essential data preprocessing tasks using R/Python as part of this assignment. We started with importing data from various formats and explored different techniques such as identifying missing values, checking and converting data types, indexing and sorting, detecting zeros, and summarizing attributes.Through this assignment, I have learned how crucial preprocessing is to prepare raw datasets for analysis or modeling. These skills are fundamental for ensuring the accuracy and reliability of results in any data-driven project.