1. Data Analysis

- **Definition:** The process of inspecting, cleaning, transforming, and modeling data to discover useful information, draw conclusions, and support decision-making.
- **Importance:**
  - Extracts meaningful insights from raw data.
  - Informs business strategies and decisions.
  - Identifies patterns, trends, and anomalies.

2. Data Preprocessing

- **Definition:** The process of preparing raw data for analysis.
- **Importance:**
  - Real-world data is often incomplete, inconsistent, and noisy.
  - Improves data quality, accuracy, and suitability for analysis.
  - Crucial for reliable results and effective modeling.

3. Pandas

- **Definition:** A Python library providing data structures and functions for efficient data manipulation and analysis.
- **Key Features:**
  - DataFrames and Series for structured data.
  - Functions for reading, writing, cleaning, transforming, and analyzing data.
  - Integration with other Python libraries like NumPy.

4. NumPy

- **Definition:** A Python library for numerical computation, supporting arrays and mathematical operations.
- **Key Features:**
  - Multi-dimensional arrays (ndarrays).
  - Mathematical functions, linear algebra, random number generation.
  - Integration with Pandas for data manipulation.

5. Data Structures

- **DataFrame:**
  - A 2D labeled table-like structure with rows and columns.
  - Each column can have a different data type.
  - Fundamental data structure in Pandas.
- **Series:**
  - A 1D labeled array, a single column of a DataFrame.
-

6. Data Types

- **Definition:** The type of data stored in a variable (e.g., numeric, text, categorical).
- **Importance:**
  - Determines how data can be processed and analyzed.
  - Ensuring correct data types is crucial for accurate results.
-

7. Missing Values

- **Definition:** Data points where no value is recorded for a particular attribute.
- **Causes:**
  - Data entry errors.
  - Incomplete records.
  - System errors.
- **Handling:**
  - Imputation (replacing with estimated values).
  - Deletion (removing rows or columns).
-

8. Data Cleaning

- **Definition:** The process of identifying and correcting errors, inconsistencies, and inaccuracies in data.
- **Tasks:**
  - Handling missing values.
  - Removing duplicates.
  - Correcting data entry errors.
  - Handling outliers.
-

9. Data Transformation

- **Definition:** Converting data from one format or structure to another.
- **Tasks:**
  - Normalization/Scaling.
  - Converting data types.
  - Creating or modifying features.
  - Converting categorical variables.

10. Data Normalization/Scaling

- **Definition**: Adjusting numerical values to a standard range.
- **Reasons**:
  - Features on different scales can bias some machine learning algorithms.

    - Improves model performance and stability.
- **Methods**:
    - Min-Max scaling
    - Standardization (Z-score)

11. Categorical Variables

- **Definition:** Variables that represent categories or groups (e.g., colors, names, labels).
- **Handling:**
    - One-hot encoding.
    - Label encoding.

12. Feature Engineering

- **Definition**: The process of selecting, transforming, and creating new features from raw data.
- **Goal**: To improve the performance of machine learning models.

13. Data Visualization

- **Definition**: Representing data visually (e.g., charts, graphs).
- **Purpose**:
    - To understand data patterns and distributions.
    - To communicate findings effectively.
- **Libraries**:
    - Matplotlib
    - Seaborn