

CLASS 2: Pandas

1. Data Cleaning in Pandas

Data cleaning in Pandas refers to the process of identifying and correcting errors, inconsistencies, and inaccuracies in datasets. Real-world data often contains missing values, duplicate records, incorrect formats, or irrelevant information, which can negatively affect analysis and model performance.

Key Data Cleaning Tasks:

- Handling missing or null values
- Removing duplicate entries
- Correcting data types
- Renaming columns for clarity
- Filtering irrelevant or noisy data

Effective data cleaning ensures that the dataset is reliable, consistent, and ready for analysis or machine learning.

2. Difference Between `loc` and `iloc`

Both `loc` and `iloc` are used for selecting data from Pandas DataFrames, but they differ in how indexing is performed.

- `loc` selects data using labels such as row names or column names. It is label-based indexing.
- `iloc` selects data using integer positions. It is position-based indexing.

Key Difference:

`loc` depends on index labels, while `iloc` depends purely on numerical positions.

3. Missing Values and Their Handling in Pandas

Missing values represent the absence of data in a dataset and are commonly indicated as NaN (Not a Number) in Pandas. Missing values may arise due to data collection errors, non-responses, or system issues.

Pandas Handles Missing Values By:

- Detecting missing data
- Removing rows or columns containing missing values
- Filling missing values using statistical methods (mean, median, mode)
- Forward or backward filling based on context

Proper handling of missing values is crucial to avoid biased or inaccurate model outcomes.

4. `groupby()` and Its Purpose

The `groupby()` function in Pandas is used to group data based on one or more categorical variables and perform aggregate operations on each group.

Common Uses:

- Summarizing data
- Computing statistics such as mean, sum, or count
- Comparing patterns across categories

`groupby()` is essential for exploratory data analysis and feature engineering.

5. Role of Pandas in Data Preprocessing for AI Models

Pandas plays a vital role in preparing raw data for artificial intelligence models. It provides tools for data manipulation, transformation, and analysis, making it easier to convert unstructured or messy data into a clean and structured format.

Importance in AI Preprocessing:

- Enables efficient handling of large datasets
- Supports feature selection and transformation
- Facilitates data normalization and encoding
- Integrates seamlessly with NumPy and Scikit-learn

Pandas acts as a bridge between raw data and AI model training.