What is pandas in python?

Pandas is an open-source Python library that provides easy-to-use data structures and data analysis tools for working with structured data. It is one of the most popular libraries used in data science and data analysis workflows. Pandas is designed to efficiently manipulate and analyze data, making it an essential tool for data manipulation, cleaning, and transformation tasks.

The primary data structures in Pandas are:

Series: A one-dimensional labeled array capable of holding any data type. It is similar to a NumPy array but with additional features like labeled indexing.

DataFrame: A two-dimensional labeled data structure, which can hold data in rows and columns, similar to a table or an Excel spreadsheet. It is the most commonly used data structure in Pandas and is highly versatile for data analysis tasks.

Pandas provides a wide range of functions and methods for tasks such as:

Loading and saving data in various formats (CSV, Excel, SQL databases, etc.)

Data cleaning and preprocessing (removing missing values, handling duplicates, etc.)

Data manipulation (filtering, sorting, grouping, merging, etc.)

Data aggregation and summarization (mean, sum, count, etc.)

Data visualization (built on top of Matplotlib and Seaborn)

Time series data analysis and manipulation

Handling missing data

Reshaping and pivoting data

Applying functions to data elements

And much more!

To use Pandas in your Python code, you'll need to import it at the beginning of your script or notebook:

python

import pandas as pd

**The os library in Python**

1. File and Directory Operations:

os.listdir(path): Returns a list of filenames in the specified directory.

os.getcwd(): Returns the current working directory.

os.chdir(path): Changes the current working directory to the specified path.

os.mkdir(path): Creates a new directory at the specified path.

os.makedirs(path): Creates multiple directories at the specified path, including any necessary parent directories.

os.remove(path): Removes a file at the specified path.

os.rmdir(path): Removes an empty directory at the specified path.

os.removedirs(path): Removes multiple directories at the specified path, including any parent directories if they become empty.

4. Environment Variables:

os.getenv(varname): Returns the value of the environment variable varname.

os.putenv(varname, value): Sets the value of the environment variable varname to value.

os.environ: A dictionary-like object containing the current environment variables.

2. Process Management:

os.system(command): Executes the system command in a subshell.

os.spawnl(mode, path, ...args): Spawns a new process using the given path and arguments.

os.getpid(): Returns the current process ID.

os.getppid(): Returns the parent process ID.

3. Path Manipulation:

os.path.join(path1, path2, ...): Joins one or more path components intelligently, using the appropriate separator for the current operating system.

os.path.exists(path): Checks if the specified path exists.

os.path.isfile(path): Checks if the specified path points to a regular file.

os.path.isdir(path): Checks if the specified path points to a directory.

NumPy (Numerical Python) is a fundamental package in Python for scientific computing. It provides support for working with arrays and matrices, along with a wide range of mathematical functions to operate on these arrays efficiently. NumPy is a crucial library for various data science, machine learning, and scientific applications.

Here are some key features and concepts related to NumPy:

1. **Arrays**: NumPy provides the **ndarray** (n-dimensional array) data structure that allows you to create and manipulate arrays of various dimensions. Arrays in NumPy are more efficient and versatile compared to Python's built-in lists.
2. **Array Creation**: You can create arrays using various methods, such as from existing data, using functions like **numpy.array()**, **numpy.zeros()**, **numpy.ones()**, **numpy.random.rand()**, and more.
3. **Array Operations**: NumPy supports element-wise mathematical operations (addition, subtraction, multiplication, division, etc.) on arrays, which are often faster and more concise than equivalent operations on Python lists.
4. **Universal Functions (ufuncs)**: NumPy provides a wide range of built-in mathematical functions (ufuncs) that operate element-wise on arrays. Examples include **numpy.sin()**, **numpy.cos()**, **numpy.exp()**, **numpy.sqrt()**, and more.
5. **Indexing and Slicing**: You can access and manipulate specific elements of arrays using indexing and slicing, similar to Python lists. NumPy also supports advanced indexing techniques like boolean indexing and fancy indexing.
6. **Array Broadcasting**: Broadcasting is a powerful feature that allows NumPy to perform operations between arrays of different shapes without explicitly replicating the smaller array.
7. **Array Manipulation**: NumPy offers functions to reshape, concatenate, split, stack, and transpose arrays, enabling you to efficiently transform and manipulate your data.
8. **Linear Algebra**: NumPy provides functions for performing various linear algebra operations, including matrix multiplication (**numpy.dot()**, **numpy.matmul()**), eigenvalue decomposition, singular value decomposition, and more.
9. **Statistical Functions**: NumPy includes a variety of statistical functions for calculating mean, median, standard deviation, variance, and more.
10. **File I/O**: NumPy can read and write data to/from disk in various formats, such as text files and binary files.

# Arithmetic operations

result = arr\_1d + 10

# result: [11 12 13 14 15]

# Element-wise operations

squared = np.square(arr\_1d)

# squared: [ 1 4 9 16 25]

# Matrix multiplication

mat\_product = np.dot(arr\_2d, arr\_2d)

# mat\_product: [[ 30 36 42]

# [ 66 81 96]

# [102 126 150]]

Array Indexing and Slicing:

You can access elements and slices of arrays using indexing and slicing:

python

Copy code

# Accessing elements

element = arr\_1d[2]

# element: 3

# Slicing

slice = arr\_1d[1:4]

# slice: [2 3 4]

Array Functions:

NumPy provides various functions for array manipulation and computation

# Sum of array elements

total\_sum = np.sum(arr\_1d)

# total\_sum: 15

# Mean of array elements

mean\_value = np.mean(arr\_1d)

# mean\_value: 3.0

|  |  |
| --- | --- |
| Function/Method | Description |
| numpy.array() | Create an array from a Python list or iterable. |
| numpy.zeros() | Create an array of all zeros. |
| numpy.ones() | Create an array of all ones. |
| numpy.empty() | Create an uninitialized array. |
| numpy.arange() | Create an array with values within a specified range. |
| numpy.linspace() | Create an array with evenly spaced values over a range. |
| numpy.eye() | Create an identity matrix or array. |
| numpy.random.rand() | Create an array with random values in the [0, 1) range. |
| numpy.random.randint() | Create an array with random integers in a specified range. |
| numpy.shape() | Get the dimensions of an array. |
| numpy.ndim() | Get the number of dimensions (axes) of an array. |
| numpy.size() | Get the total number of elements in an array. |
| numpy.reshape() | Reshape an array to a specified shape. |
| numpy.flatten() | Return a 1D copy of the array. |
| numpy.transpose() | Transpose an array (swap dimensions). |
| numpy.concatenate() | Join arrays along an existing axis. |
| numpy.vstack() | Stack arrays vertically (along rows). |
| numpy.hstack() | Stack arrays horizontally (along columns). |
| numpy.split() | Split an array into multiple sub-arrays. |
| numpy.max() / .min() | Find the maximum / minimum value in an array. |
| numpy.argmax() / .argmin() | Find the index of the maximum / minimum value. |
| numpy.mean() | Calculate the mean (average) of array elements. |
| numpy.median() | Calculate the median of array elements. |
| numpy.std() | Calculate the standard deviation of array elements. |
| numpy.sum() | Calculate the sum of array elements. |
| numpy.dot() | Perform matrix multiplication or dot product. |
| numpy.matmul() | Perform matrix multiplication. |
| numpy.diagonal() | Get the diagonal of a 2D array. |
| numpy.unique() | Find unique elements in an array. |
| numpy.sort() | Sort the elements of an array. |
| numpy.argsort() | Return indices that would sort an array. |
| numpy.clip() | Clip array values to a specified range. |
| numpy.copy() | Create a copy of an array. |
| numpy.save() / .load() | Save and load array data to/from a binary file. |
| numpy.savetxt() / .loadtxt() | Save and load array data to/from a text file. |

|  |  |
| --- | --- |
| Method | Description |
| Dataframe Creation |  |
| pd.DataFrame(data) | Create a DataFrame from various data sources. |
| df = pd.read\_csv(file) | Read a CSV file and create a DataFrame. |
| df = pd.read\_excel(file) | Read an Excel file and create a DataFrame. |
| df = pd.DataFrame(dict) | Create a DataFrame from a dictionary. |
| Basic Operations |  |
| df.head(n) | Display the first n rows of the DataFrame. |
| df.tail(n) | Display the last n rows of the DataFrame. |
| df.shape | Get the number of rows and columns. |
| df.info() | Get information about DataFrame data types. |
| Data Selection |  |
| df[col] | Select a single column by name. |
| df[[col1, col2]] | Select multiple columns by names. |
| df.loc[row, col] | Select a cell by label-based index. |
| df.iloc[row, col] | Select a cell by integer-based index. |
| Filtering Data |  |
| df[df['col'] > value] | Filter rows based on a condition. |
| df.query('condition') | Filter rows using a query string. |
| Sorting |  |
| df.sort\_values(col) | Sort DataFrame by values in a column. |
| df.sort\_index() | Sort DataFrame by index. |
| Aggregation |  |
| df.groupby(col).agg(func) | Apply an aggregation function. |
| df.mean() | Calculate mean value of numeric columns. |
| df.sum() | Calculate sum of numeric columns. |
| Data Cleaning |  |
| df.dropna() | Drop rows with missing values. |
| df.fillna(value) | Fill missing values with a specific value. |
| Data Modification |  |
| df['col'] = value | Modify values in a column. |
| df.rename(columns) | Rename columns. |
| df.drop(columns) | Drop specified columns. |
| Joining Data |  |
| pd.concat([df1, df2]) | Concatenate DataFrames vertically. |
| df.merge(other\_df) | Merge DataFrames based on specified columns. |
| Pivot and Reshape |  |
| df.pivot\_table() | Create a pivot table. |
| df.melt() | Unpivot/melt a DataFrame. |
| Statistical Summary |  |
| df.describe() | Generate summary statistics for numeric data. |
| Saving Data |  |
| df.to\_csv(file) | Save DataFrame to a CSV file. |
| df.to\_excel(file) | Save DataFrame to an Excel file. |