

Module - 2

Data Wrangling

Python Pandas

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures.

Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data — load, prepare, manipulate, model, and analyze.

Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

Key features of Pandas

- Fast and efficient DataFrame object with default and customized indexing.
- Tools for loading data into in-memory data objects from different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of date sets.
- Label-based slicing, indexing and subsetting of large data sets.
- Columns from a data structure can be deleted or inserted.
- Group by data for aggregation and transformations.
- High performance merging and joining of data.
- Time Series functionality.

Data structures

Pandas deals with the following three data structures –

- Series
- DataFrame
- Panel

These data structures are built on top of Numpy array, which means they are fast.

Data Structure	Dimensions	Description
Series	1	1D labeled homogeneous array, sizeimmutable.
Data Frames	2	General 2D labeled, size-mutable tabular structure with potentially heterogeneously typed columns.
Panel	3	General 3D labeled, size-mutable array.

Mutability

All Pandas data structures are value mutable (can be changed) and except Series all are size mutable. Series is size immutable.

Note – DataFrame is widely used and one of the most important data structures. Panel is used much less.

Series

Series is a one-dimensional array like structure with homogeneous data. For example, the following series is a collection of integers 10, 23, 56, ...

10 23 56 17 52 61 73 90 26 72

Key Points

- Homogeneous data

- Size Immutable

- Values of Data Mutable

DataFrame

DataFrame is a two-dimensional array with heterogeneous data. For example,

Name	Age	Gender	Rating
Steve	32	Male	3.45
Lia	28	Female	4.6
Vin	45	Male	3.9
Katie	38	Female	2.78

The table represents the data of a sales team of an organization with their overall performance rating. The data is represented in rows and columns. Each column represents an attribute and each row represents a person.

Data Type of Columns

The data types of the four columns are as follows –

Column	Type
Name	String
Age	Integer
Gender	String
Rating	Float

Key Points

- Heterogeneous data

- Size Mutable

- Data Mutable

A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns.

Features of DataFrame

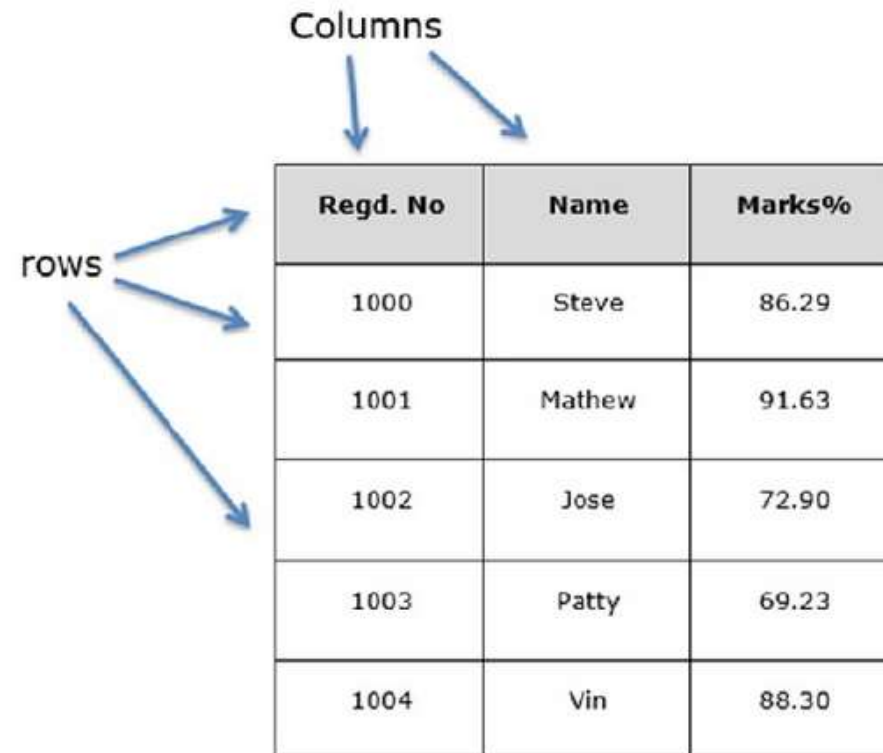
- Potentially columns are of different types

- Size – Mutable

- Labeled axes (rows and columns)

- Can Perform Arithmetic operations on rows and columns

Let us assume that we are creating a data frame with student's data.



The diagram illustrates a data frame structure. A table with three columns and six rows is shown. The columns are labeled 'Regd. No', 'Name', and 'Marks%'. The rows are labeled with student IDs: 1000, 1001, 1002, 1003, and 1004. Blue arrows point from the labels 'Columns' and 'rows' to their respective parts of the table.

Regd. No	Name	Marks%
1000	Steve	86.29
1001	Mathew	91.63
1002	Jose	72.90
1003	Patty	69.23
1004	Vin	88.30

pandas.DataFrame

A pandas DataFrame can be created using the following constructor –

```
pandas.DataFrame( data, index, columns, dtype, copy)
```

A pandas DataFrame can be created using various inputs like –

- Lists

- dict

- Series

- Numpy ndarrays

- Another DataFrame

```
#import the pandas library and aliasing as pd
import pandas as pd
df = pd.DataFrame()
print(df)
```

Create a DataFrame from Lists

The DataFrame can be created using a single list or a list of lists.

Example

```
import pandas as pd
data = [1,2,3,4,5]
df = pd.DataFrame(data)
Print(df)
```

```
import pandas as pd
data = [['Alex',10],['Bob',12],['Clarke',13]]
df = pd.DataFrame(data,columns=['Name','Age'])
print df
```

```
import pandas as pd
data = [['Alex',10],['Bob',12],['Clarke',13]]
df = pd.DataFrame(data,columns=['Name','Age'],dtype=float)
print df
```