Pandas Python Library

Pandas is a software library written for the Python programming language for data manipulation and analysis.
In particular, it offers data structures and operations for manipulating numerical tables and time series.
It is free software released under the three-clause BSD license.
The name is derived from the term "panel data", an econometrics term for data sets that include observations over multiple time periods for the same individuals.

This is the most important library for Machine Learning because we can deal with our data set by using Pandas.

Install Pandas library using pip as

sudo pip install pandas

Features of Pandas

- DataFrame object for data manipulation with integrated indexing.
- Tools for reading and writing data between in-memory data structures and different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of data sets.
- Label-based slicing, fancy indexing, and subsetting of large data sets.
- Data structure column insertion and deletion.
- Group by engine allowing split-apply-combine operations on data sets.
- Data set merging and joining.
- Hierarchical axis indexing to work with high-dimensional data in a lower-dimensional data structure.
- Time series-functionality: Date range generation[4] and frequency conversion, moving window statistics, moving window linear regressions, date shifting and lagging.
- Provides data filtration.

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.

Dimension & Description

The best way to think of these data structures is that the higher dimensional data structure is a container of its lower dimensional data structure.
For example, DataFrame is a container of Series, Panel is a container of DataFrame.

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.

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, …

Key Points

- Homogeneous data
- Size Immutable
- Values of Data Mutable

10	23	56	17	52	61	73	90	26	72

DataFrame

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

Marvellous Infosystems Training Data set

Weight	Pattern	Label
35	Rough	Tennis
47	Rough	Tennis
90	Smooth	Cricket
48	Rough	Tennis
90	Smooth	Cricket
35	Rough	Tennis
92	Smooth	Cricket
35	Rough	Tennis
35	Rough	Tennis
35	Rough	Tennis
96	Smooth	Cricket
43	Rough	Tennis
110	Smooth	Cricket
35	Rough	Tennis
95	Smooth	Cricket

The table represents the data of a balls.
We are going to user to use this data set in machine learning application.
The data is represented in rows and columns. Each column represents an attribute and each row represents a person.
We consider data frame as our excel sheet.

Key Points

- Heterogeneous data
- Size Mutable
- Data Mutable

Panel

Panel is a three-dimensional data structure with heterogeneous data. It is hard to represent the panel in graphical representation. But a panel can be illustrated as a container of DataFrame.
We consider panel as excel file which contains multiple excel sheets (Data frame).

Key Points

- Heterogeneous data
- Size Mutable
- Data Mutable