



Python Pandas



tutorialspoint

SIMPLY EASY LEARNING



www.tutorialspoint.com



<https://www.facebook.com/tutorialspointindia>



<https://twitter.com/tutorialspoint>

About the Tutorial

Pandas is an open-source, BSD-licensed Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc.

In this tutorial, we will learn the various features of Python Pandas and how to use them in practice.

Audience

This tutorial has been prepared for those who seek to learn the basics and various functions of Pandas. It will be specifically useful for people working with data cleansing and analysis.

After completing this tutorial, you will find yourself at a moderate level of expertise from where you can take yourself to higher levels of expertise.

Prerequisites

You should have a basic understanding of Computer Programming terminologies. A basic understanding of any of the programming languages is a plus.

Pandas library uses most of the functionalities of NumPy. It is suggested that you go through our tutorial on NumPy before proceeding with this tutorial. You can access it from: [NumPy Tutorial](#).

Disclaimer & Copyright

© Copyright 2017 by Tutorials Point (I) Pvt. Ltd.

All the content and graphics published in this e-book are the property of Tutorials Point (I) Pvt. Ltd. The user of this e-book is prohibited to reuse, retain, copy, distribute or republish any contents or a part of contents of this e-book in any manner without written consent of the publisher.

We strive to update the contents of our website and tutorials as timely and as precisely as possible, however, the contents may contain inaccuracies or errors. Tutorials Point (I) Pvt. Ltd. provides no guarantee regarding the accuracy, timeliness or completeness of our website or its contents including this tutorial. If you discover any errors on our website or in this tutorial, please notify us at contact@tutorialspoint.com.

Table of Contents

About the Tutorial	i
Audience.....	i
Prerequisites.....	i
Disclaimer & Copyright.....	i
Table of Contents	ii
1. Pandas – Introduction.....	1
2. Pandas – Environment Setup	2
3. Pandas – Introduction to Data Structures	3
Dimension & Description.....	3
Series	4
DataFrame	4
Data Type of Columns	4
Panel.....	5
4. Pandas — Series.....	6
pandas.Series.....	6
Create an Empty Series.....	7
Create a Series f.....	7
rom ndarray.....	7
Create a Series f.....	8
rom dict	8
Create a Series f.....	9
rom Scalar.....	9
Accessing Data from Series with Position	10
Retrieve Data Using Label (Index)	11

5. Pandas – DataFrame	13
pandas.DataFrame	14
Create DataFrame	14
Create an Empty DataFrame	15
Create a DataFrame from Lists	15
Create a DataFrame from Dict of ndarrays / Lists	16
Create a DataFrame from List of Dicts.....	17
Create a DataFrame from Dict of Series	19
Column Selection.....	20
Column	20
Addition	20
Column Deletion.....	21
Row Selection, Addition, and Deletion	23
6. Pandas – Panel.....	26
pandas.Panel()	26
Create Panel	26
Selecting the Data from Panel	28
7. Pandas – Basic Functionality	30
DataFrame Basic Functionality	35
8. Pandas – Descriptive Statistics	45
Functions & Description	48
Summarizing Data	49
9. Pandas – Function Application	53
Table-wise Function Application	53
Row or Column Wise Function Application	54

Element Wise Function Application	55
10. Pandas – Reindexing	57
Reindex to Align with Other Objects	58
Filling while ReIndexing	58
Limits on Filling while Reindexing.....	60
Renaming.....	61
11. Pandas – Iteration	62
Iterating a DataFrame.....	62
iteritems()	63
iterrows()	64
itertuples()	64
12. Pandas – Sorting	66
By Label	66
Sorting Algorithm	70
13. Pandas – Working with Text Data	71
14. Pandas – Options and Customization	82
get_option(param)	82
set_option(param,value)	83
reset_option(param)	83
describe_option(param).....	84
option_context().....	84
15. Pandas – Indexing and Selecting Data	86
.loc().....	86
.iloc()	90
.ix().....	92

Use of Notations.....	93
16. Pandas – Statistical Functions	96
Percent_change.....	96
Covariance	97
Correlation.....	98
Data Ranking.....	98
17. Pandas – Window Functions	100
.rolling() Function	100
.expanding() Function.....	101
.ewm() Function	101
18. Pandas – Aggregations	103
Applying Aggregations on DataFrame	103
19. Pandas – Missing Data	108
Cleaning / Filling Missing Data.....	111
Replace NaN with a Scalar Value	111
Fill NA Forward and Backward	112
Drop Missing Values	113
Replace Missing (or) Generic Values	114
20. Pandas – GroupBy.....	116
Split Data into Groups	117
View Groups	117
Iterating through Groups.....	119
Select a Group	120
Aggregations.....	121
Transformations	123

Filtration	124
21. Pandas – Merging/Joining.....	125
Merge Using 'how' Argument.....	127
22. Pandas – Concatenation.....	131
Concatenating Objects	131
Time Series	136
23. Pandas – Date Functionality.....	139
24. Pandas – Timedelta.....	141
25. Pandas – Categorical Data.....	144
Object Creation	144
26. Pandas – Visualization	150
Bar Plot	151
Histograms.....	153
Box Plots	154
Area Plot	155
Scatter Plot	155
Pie Chart	156
27. Pandas – IO Tools.....	157
read.csv	157
28. Pandas – Sparse Data.....	161
29. Pandas – Caveats & Gotchas	164
30. Pandas – Comparison with SQL.....	169

1. Pandas – Introduction

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word Panel Data – an Econometrics from Multidimensional data.

In 2008, developer Wes McKinney started developing pandas when in need of high performance, flexible tool for analysis of data.

Prior to Pandas, Python was majorly used for data munging and preparation. It had very less contribution towards data analysis. Pandas solved this problem. 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.

2. Pandas – Environment Setup

Standard Python distribution doesn't come bundled with Pandas module. A lightweight alternative is to install NumPy using popular Python package installer, **pip**.

```
pip install pandas
```

If you install Anaconda Python package, Pandas will be installed by default with the following:

Windows

- **Anaconda** (from <https://www.continuum.io>) is a free Python distribution for SciPy stack. It is also available for Linux and Mac.
- **Canopy** (<https://www.enthought.com/products/canopy/>) is available as free as well as commercial distribution with full SciPy stack for Windows, Linux and Mac.
- **Python** (x,y) is a free Python distribution with SciPy stack and Spyder IDE for Windows OS. (Downloadable from <http://python-xy.github.io/>)

Linux

Package managers of respective Linux distributions are used to install one or more packages in SciPy stack.

For Ubuntu Users

```
sudo apt-get install python-numpy python-scipy python-matplotlibpythonipython-notebook python-pandas python-sympy python-nose
```

For Fedora Users

```
sudo yum install numpy scipy python-matplotlibpython python-pandas sympy python-nose atlas-devel
```

3. Pandas – Introduction to 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.

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, size-immutable.
Data Frames	2	General 2D labeled, size-mutable tabular structure with potentially heterogeneously-typed columns.
Panel	3	General 3D labeled, size-mutable array.

Building and handling two or more dimensional arrays is a tedious task, burden is placed on the user to consider the orientation of the data set when writing functions. But using Pandas data structures, the mental effort of the user is reduced.

For example, with tabular data (DataFrame) it is more semantically helpful to think of the **index** (the rows) and the **columns** rather than axis 0 and axis 1.

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 very less used.

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

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.

Key Points

- Heterogeneous data
- Size Mutable
- Data Mutable

End of ebook preview

If you liked what you saw...

Buy it from our store @ **<https://store.tutorialspoint.com>**