

Data Analysis with Pandas

Tushar B. Kute,
<http://tusharkute.com>

Pandas

- 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 little 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.

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.

Installation

- 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-pandas python-sympy python-nose`**
- For Fedora Users
 - **`sudo yum install numpy scipy python-matplotlib python python-pandas sympy python-nose atlas-devel`**

Pandas data structures

- Pandas deals with the following two data structures –
 - Series
 - DataFrame
- These data structures are built on top of Numpy array, which means they are fast.

Pandas data structures

- 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
- Series
 - 1D labeled homogeneous array, size- immutable.
- Data Frames
 - General 2D labeled, size-mutable tabular structure with potentially heterogeneously typed columns.

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,

Roll	Name	Marks
1	Aneet	67.56
2	Asha	59.05
3	Anil	71.22

- The data is represented in rows and columns. Each column represents an attribute and each row represents a person.

Data types of columns

- The data types of the three columns are as follows –

Column	Type
Roll	Integer
Name	String
Marks	Float

- Key Points
 - Heterogeneous data
 - Size Mutable
 - Data Mutable

Series object

- Series is a one-dimensional labeled array capable of holding data of any type (integer, string, float, python objects, etc.).
- The axis labels are collectively called index.

pandas.Series

- A pandas Series can be created using the following constructor –
pandas.Series(data, index, dtype, copy)

Creating a Series

```
#import the pandas library and aliasing as pd
import pandas as pd
import numpy as np
sr = pd.Series()
print sr

data = np.array(['a', 'b', 'c', 'd'])
sr = pd.Series(data)
print sr
```

Creating a Series

Series using index

```
data = np.array(['a', 'b', 'c', 'd'])  
s = pd.Series(data, index=[100, 101, 102, 103])  
print s
```

Series from dict

```
data = {'a' : 0., 'b' : 1., 'c' : 2.}  
s = pd.Series(data)  
print s
```

Series from Scalar

```
s = pd.Series(5, index=[0, 1, 2, 3])  
print s
```

Accessing Series elements

```
s = pd.Series([1,2,3,4,5],index = ['a','b','c','d','e'])
#retrieve the first element
print s[0]
#retrieve the first three element
print s[:3]
#retrieve the last three element
print s[-3:]
#retrieve a single element
print s['a']
#retrieve multiple elements
print s[['a','c','d']]
#retrieve single elements [error]
print s['f']
```

DataFrames

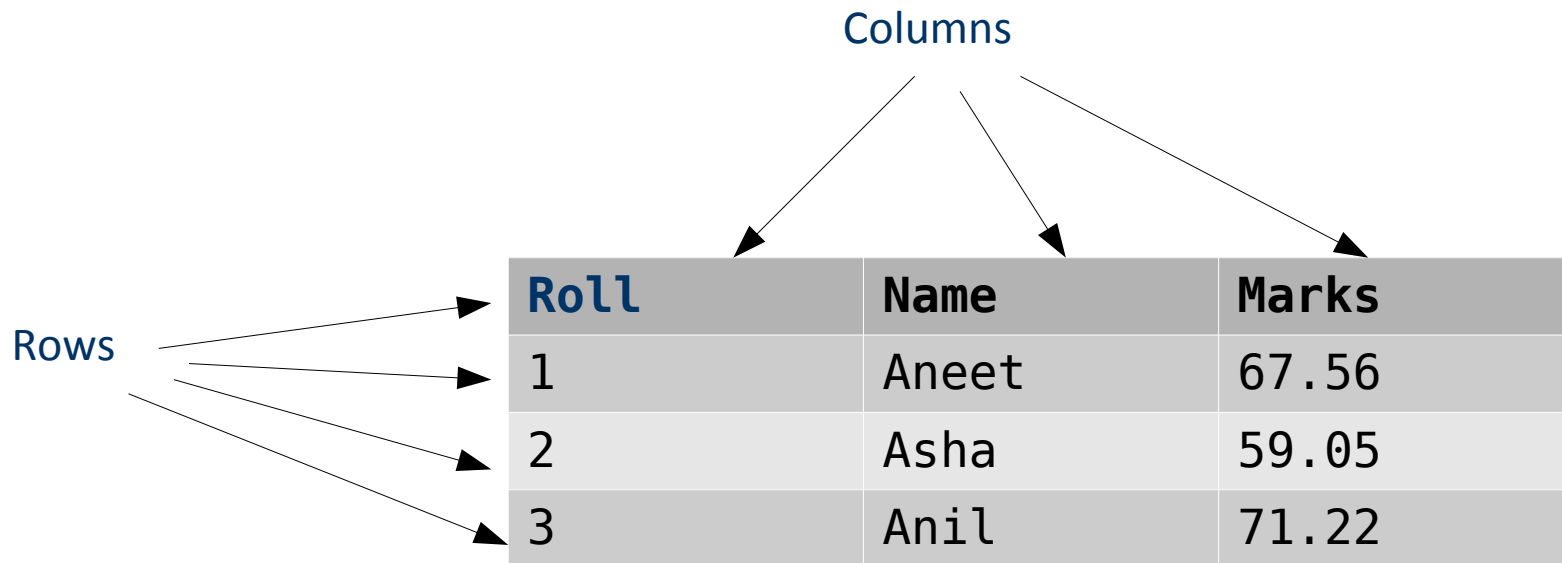
- A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns.
- Features of DataFrame
 - Potentially columns are of different types
 - Size – Mutable
 - Labeled axes (rows and columns)
 - Can Perform Arithmetic operations on rows and columns

DataFrame Structure

Columns

Rows

Roll	Name	Marks
1	Aneet	67.56
2	Asha	59.05
3	Anil	71.22

The diagram illustrates the structure of a DataFrame. It features a table with three columns: 'Roll', 'Name', and 'Marks'. The 'Roll' column is highlighted in blue. To the left of the table, the word 'Rows' is written in blue, with three arrows pointing to the first three rows of the table. Above the table, the word 'Columns' is written in blue, with three arrows pointing to the three columns of the table.

Creating DataFrames

- 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

Creating DataFrames

```
import pandas as pd
# Empty dataframe
df = pd.DataFrame()
print df
```

```
# Create df from list
data = [1,2,3,4,5]
df = pd.DataFrame(data)
print df
```

```
# Create df from list
data = [['Ashok',10],['Ana',12],['Asha',13]]
df = pd.DataFrame(data,columns=['Name','Age'])
print df
```

Thank you

This presentation is created using LibreOffice Impress 7.4.1.2, can be used freely as per GNU General Public License



@mitu_skillologies



@mITuSkillologies



@mitu_group



@mitu-skillologies



@MITUSkillologies

kaggle

@mituskillologies

Web Resources

<https://mitu.co.in>

<http://tusharkute.com>



@mituskillologies

contact@mitu.co.in
tushar@tusharkute.com