Python Libraries : PANDAS

Agenda

- Introduction of Pandas
- Pandas Dataframes
- Pandas Series
- Converting Dataframe & Series into arrays
- Functions to work view data with files

What is Pandas

- Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool

- The name Pandas is derived from word ‘Panel Data’, and was developed by Wes McKinney for high performance, flexible tool for data analysis.

- Using Pandas, we can accomplish five typical steps in the processing and analysis 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.

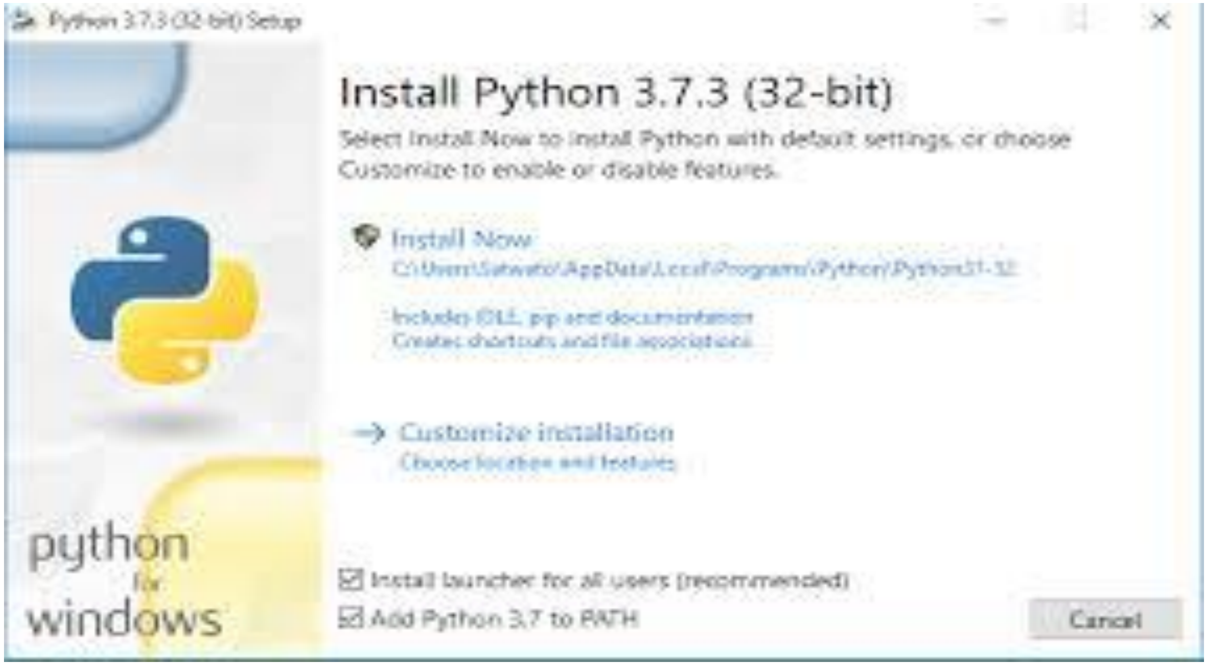
Key Features

Pandas

INTERNSHIPSTUDIO

- Fast and efficient Dataframe objects with customized indexing.
- Tools for loading data into in-memory from different file formats.
- Data alignment and integrated handling of missing data.
- Reshaping and pivoting of date sets.
- Label-based slicing, indexing and sub setting of large data sets.
- Group by data for aggregation and transformations.
- High performance merging and joining of data. Time Series functionality.

Environment Setup



Python 3.7.3 (32-bit) Setup

Install Python 3.7.3 (32-bit)

Select Install Now to install Python with default settings, or choose Customize to enable or disable features.

Install Now
C:\Users\Satwato\AppData\Local\Programs\Python\Python37-32
Includes IDLE, pip and documentation
Creates shortcuts and file associations
→ Customize installation
Choose location and features
python for windows
Cancel

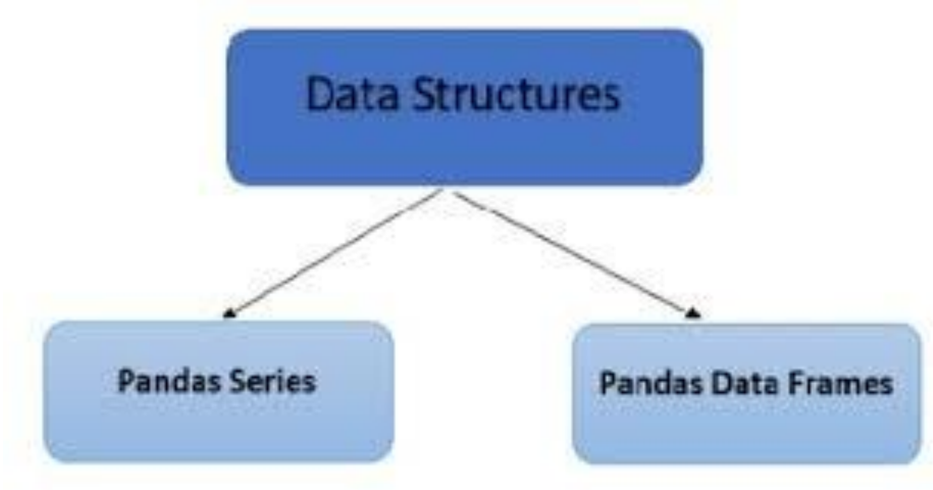
☑ Install launcher for all users (recommended)

☑ Add Python 3.7 to PATH

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

Introduction to Data Structures



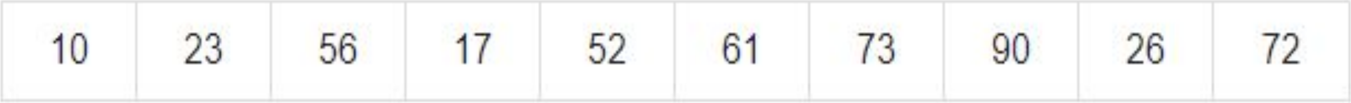
Pandas deals with the following three data structures –

1 Series
2 DataFrame
3 Panel

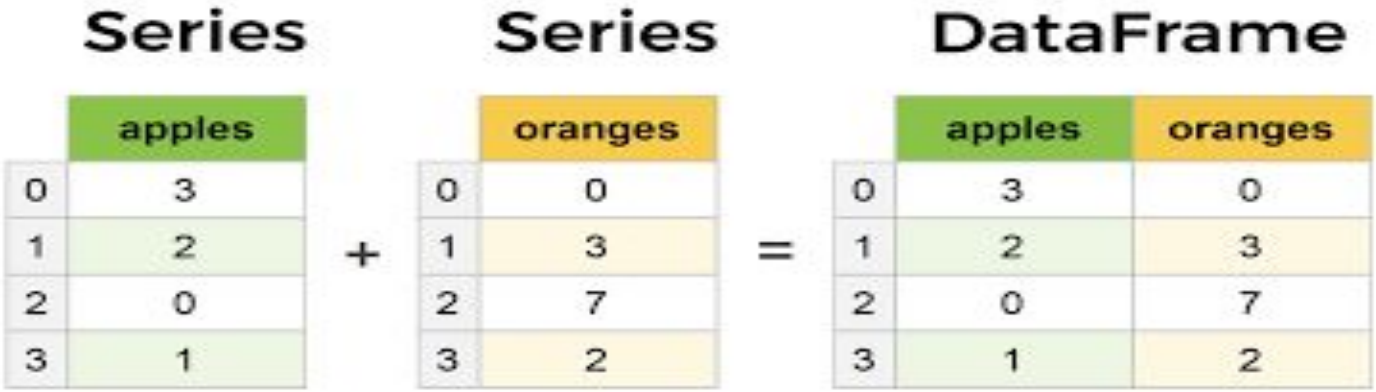
These data structures are built on top of Numpy array, which means they are fast.

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, …
- **<u>Key Points:</u>**
- Homogeneous data
- Size Immutable
- Values of Data Mutable



10	23	56	17	52	61	73	90	26	72



Series

	apples
0	3
1	2
2	0
3	1

+

Series

	oranges
0	0
1	3
2	7
3	2

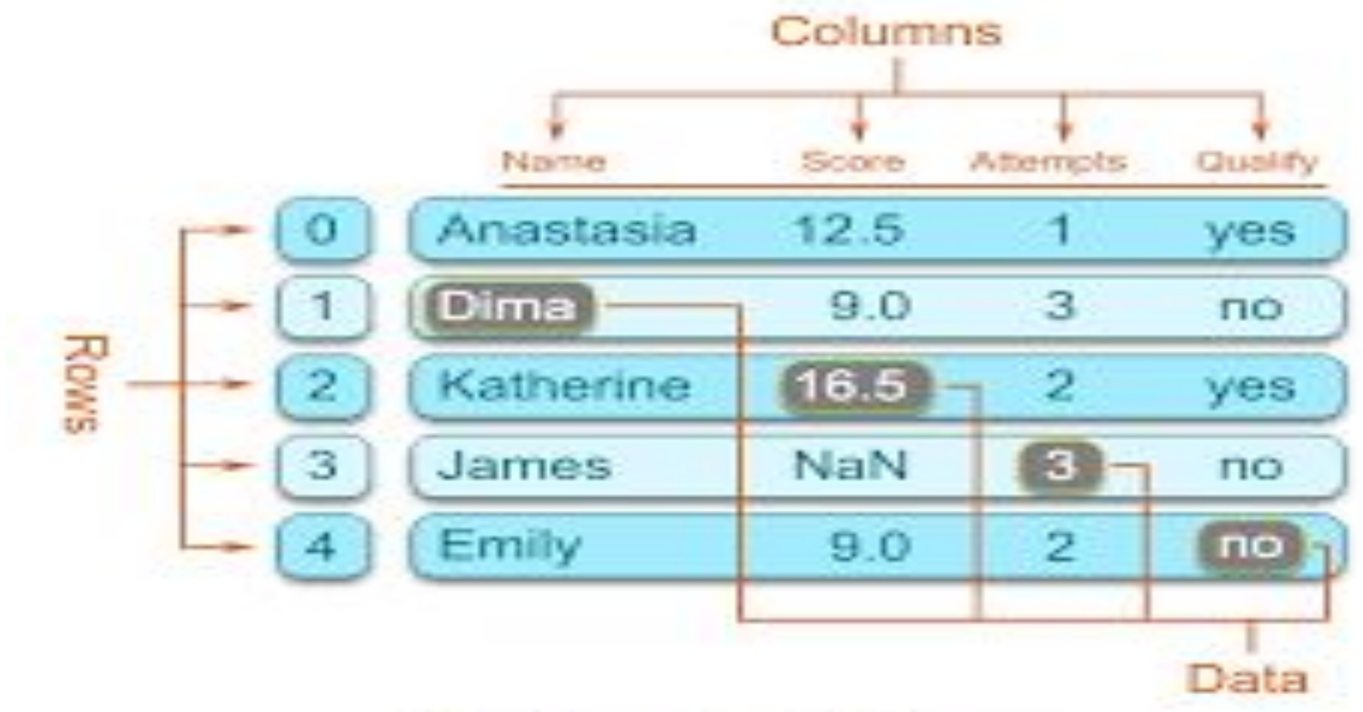
=

DataFrame

	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2

DataFrame

DataFrame is a two-dimensional array with heterogeneous data.



	Name	Score	Attempts	Qualify
0	Anastasia	12.5	1	yes
1	Dima	9.0	3	no
2	Katherine	16.5	2	yes
3	James	NaN	3	no
4	Emily	9.0	2	no

Pandas DataFrame

© w3resource.com

Data Type

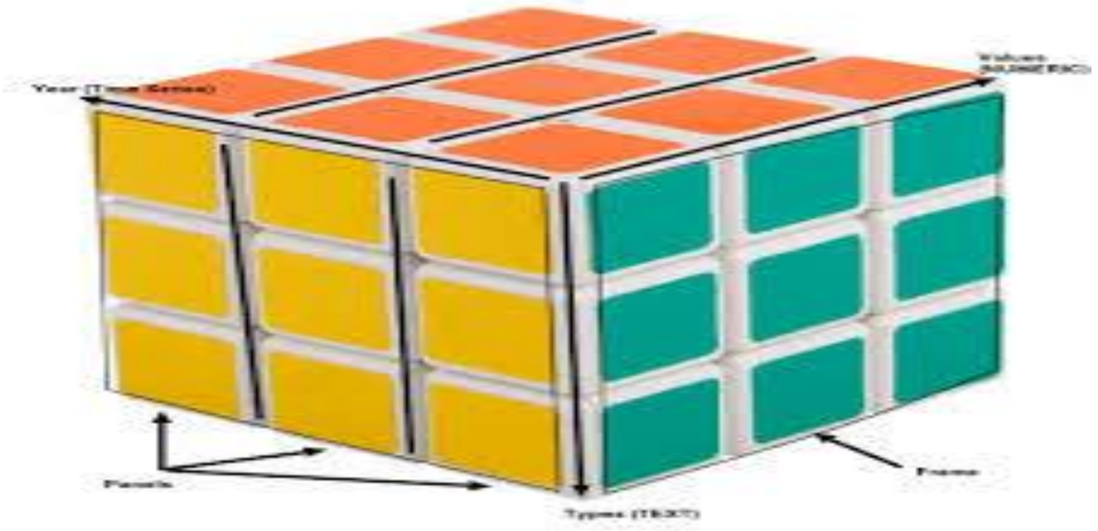
The data types of the four columns are as follows

Name	String
Score	Float
Attempt	Integer
Qualify	String

Key Points:-

- Heterogeneous data
- Size Mutable
- Data Mutable

Panel



Panel is a three-dimensional data structure with heterogeneous data.

A panel can be illustrated as a container of DataFrame.

Key Points:-

- Heterogeneous data
- Size Mutable
- Data Mutable

- Define the Pandas/Python pandas?
- Mention the different types of Data Structures in Pandas?
- Define DataFrame with example in Pandas?
- Define Series with example in Pandas?
- Define panel with example in Pandas?