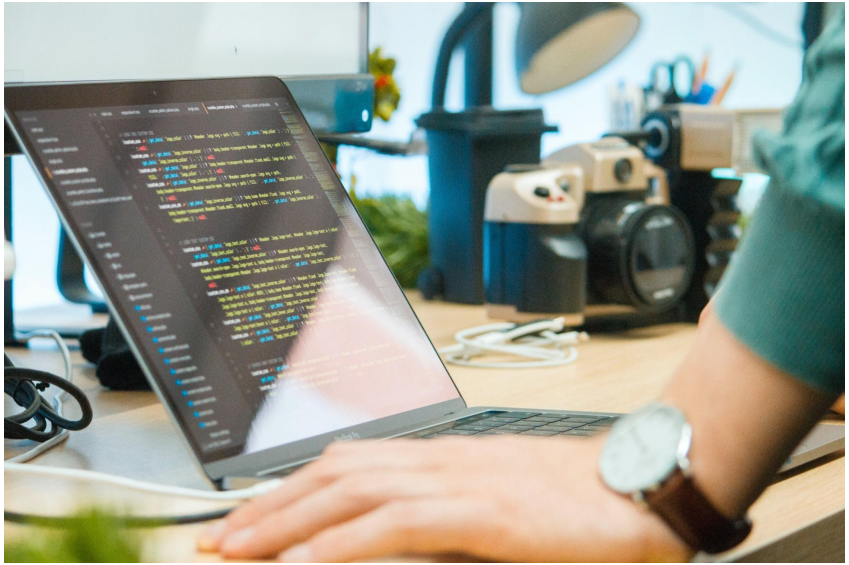
Indonesia AI
AI for Everyone, AI for Indonesia



Pandas

What is Pandas?

Pandas is a fast, powerful, flexible and easy to use open source **data analysis and manipulation tool**, built on top of the Python programming language.

Key Features

- **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

Essential Concept

Indonesia AI
AI for Everyone, AI for Indonesia

- **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, ...

  - Homogeneous data
  - Size Immutable
  - Values of Data Mutable

Essential Concept

Indonesia AI
AI for Everyone, AI for Indonesia

- **DataFrame**

  DataFrame is a two-dimensional array with heterogeneous data.

  - Heterogeneous data
  - Size Mutable
  - Data Mutable

What We Will Learn?

1. Data loading
2. Selecting/Indexing
3. Filtering
4. Sorting
5. Mutating/conditionally adding columns
6. Groupby/summarize

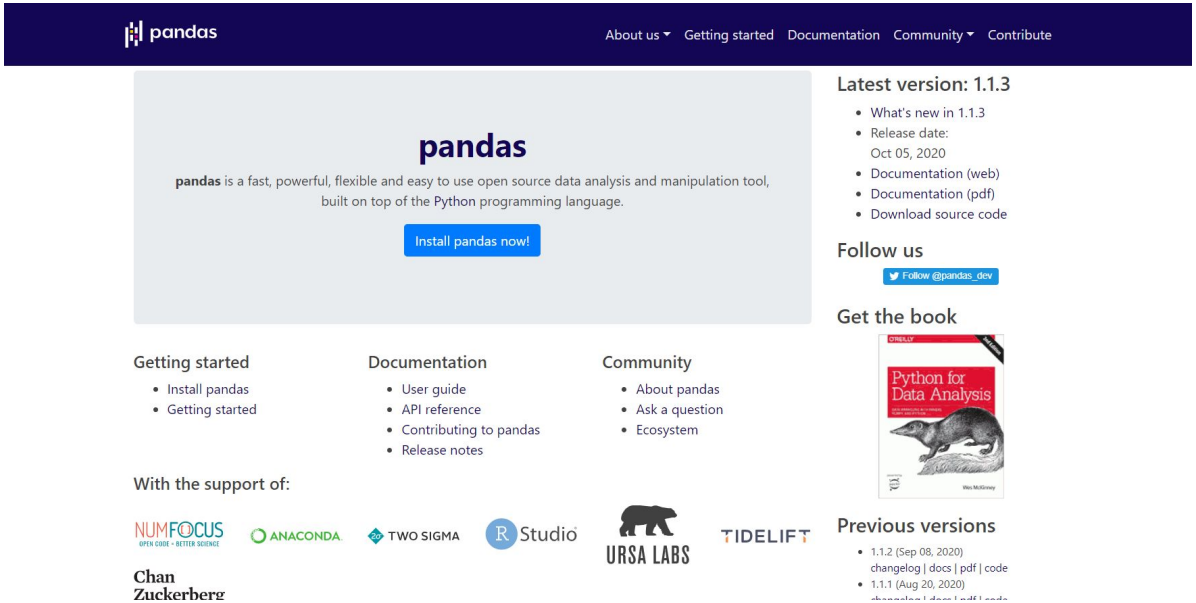
Quiz Session

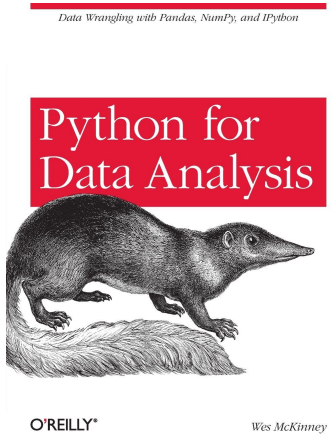
Quiz

What is a correct syntax to create a Pandas Series from a Python list?

a. pd.getSeries(myseries)
b. pd.series(myseries)
c. pd.createSeries(myseries)

Resources





Python for Data Analysis
(Wes McKinney)

DataFrame - Basic Functionality

Indonesia AI
AI for Everyone, AI for Indonesia

Let's get started!

Sr.No.	Attribute or Method & Description
1	**T** Transposes rows and columns.
2	**axes** Returns a list with the row axis labels and column axis labels as the only members.
3	**dtypes** Returns the dtypes in this object.
4	**empty** True if NDFrame is entirely empty [no items]; if any of the axes are of length 0.
5	**ndim** Number of axes / array dimensions.
6	**shape** Returns a tuple representing the dimensionality of the DataFrame.
7	**size** Number of elements in the NDFrame.
8	**values** Numpy representation of NDFrame.
9	**head()** Returns the first n rows.

DataFrame - T (Transpose)

Indonesia AI
AI for Everyone, AI for Indonesia

Returns the transpose of the DataFrame. The rows and columns will interchange.

```
# Create a Dictionary of series
d = {'Name':pd.Series(['Tom','James','Ricky','Vin','Steve','Smith','Jack']),
   'Age':pd.Series([25,26,25,23,30,29,23]),
   'Rating':pd.Series([4.23,3.24,3.98,2.56,3.20,4.6,3.8])}

# Create a DataFrame
df = pd.DataFrame(d)
print ("The transpose of the data series is:")
print df.T
```

Its **output** is as follows –

```
The transpose of the data series is:
         0     1       2      3      4      5      6
Age      25    26      25     23     30     29     23
Name     Tom   James   Ricky  Vin    Steve  Smith  Jack
Rating   4.23  3.24    3.98   2.56   3.2    4.6    3.8
```

DataFrame - dtypes

Indonesia AI
AI for Everyone, AI for Indonesia

Returns the data type of each column.

```
#Create a Dictionary of series
d = {'Name':pd.Series(['Tom','James','Ricky','Vin','Steve','Smith','Jack']),
   'Age':pd.Series([25,26,25,23,30,29,23]),
   'Rating':pd.Series([4.23,3.24,3.98,2.56,3.20,4.6,3.8])}

#Create a DataFrame
df = pd.DataFrame(d)
print ("The data types of each column are:")
print df.dtypes
```

Its **output** is as follows –

```
The data types of each column are:
Age     int64
Name    object
Rating  float64
dtype: object
```

DataFrame - shape

Indonesia AI
AI for Everyone, AI for Indonesia

Returns a tuple representing the dimensionality of the DataFrame. Tuple (a,b), where a represents the number of rows and b represents the number of columns.

```
#Create a Dictionary of series
d = {'Name':pd.Series(['Tom','James','Ricky','Vin','Steve','Smith','Jack']),
   'Age':pd.Series([25,26,25,23,30,29,23]),
   'Rating':pd.Series([4.23,3.24,3.98,2.56,3.20,4.6,3.8])}

#Create a DataFrame
df = pd.DataFrame(d)
print ("Our object is:")
print df
print ("The shape of the object is:")
print df.shape
```

Its **output** is as follows –

```
Our object is:
   Age   Name    Rating
0  25    Tom     4.23
1  26    James   3.24
2  25    Ricky   3.98
3  23    Vin     2.56
4  30    Steve   3.20
5  29    Smith   4.60
6  23    Jack    3.80

The shape of the object is:
(7, 3)
```

DataFrame - values

Indonesia AI
AI for Everyone, AI for Indonesia

Returns the actual data in the DataFrame as an NDarray.

```
#Create a Dictionary of series
d = {'Name':pd.Series(['Tom','James','Ricky','Vin','Steve','Smith','Jack']),
   'Age':pd.Series([25,26,25,23,30,29,23]),
   'Rating':pd.Series([4.23,3.24,3.98,2.56,3.20,4.6,3.8])}

#Create a DataFrame
df = pd.DataFrame(d)
print ("Our object is:")
print df
print ("The actual data in our data frame is:")
print df.values
```

Its **output** is as follows –

```
Our object is:
    Age   Name    Rating
0   25    Tom     4.23
1   26    James   3.24
2   25    Ricky   3.98
3   23    Vin     2.56
4   30    Steve   3.20
5   29    Smith   4.60
6   23    Jack    3.80
The actual data in our data frame is:
[[25 'Tom' 4.23]
[26 'James' 3.24]
[25 'Ricky' 3.98]
[23 'Vin' 2.56]
[30 'Steve' 3.2]
[29 'Smith' 4.6]
[23 'Jack' 3.8]]
```

DataFrame - Head & Tail

Indonesia AI
AI for Everyone, AI for Indonesia

To view a small sample of a DataFrame object, use the head() and tail() methods. head() returns the first n rows (observe the index values). The default number of elements to display is five, but you may pass a custom number.

```
#Create a Dictionary of series
d = {'Name':pd.Series(['Tom','James','Ricky','Vin','Steve','Smith','Jack']),
   'Age':pd.Series([25,26,25,23,30,29,23]),
   'Rating':pd.Series([4.23,3.24,3.98,2.56,3.20,4.6,3.8])}

#Create a DataFrame
df = pd.DataFrame(d)
print ("Our data frame is:")
print df
print ("The first two rows of the data frame is:")
print df.head(2)
```

Its **output** is as follows –

```
Our data frame is:
    Age   Name    Rating
0   25    Tom     4.23
1   26    James   3.24
2   25    Ricky   3.98
3   23    Vin     2.56
4   30    Steve   3.20
5   29    Smith   4.60
6   23    Jack    3.80

The first two rows of the data frame is:
   Age   Name   Rating
0  25    Tom    4.23
1  26    James  3.24
```

DataFrame - Functions

Sr.No.	Function	Description
1	count()	Number of non-null observations
2	sum()	Sum of values
3	mean()	Mean of Values
4	median()	Median of Values
5	mode()	Mode of values
6	std()	Standard Deviation of the Values
7	min()	Minimum Value
8	max()	Maximum Value
9	abs()	Absolute Value
10	prod()	Product of Values
11	cumsum()	Cumulative Sum
12	cumprod()	Cumulative Product

Quiz Session

Quiz

What is the output from this code ?

```
df = pd.DataFrame([[1, 2], [3, 4]], columns=list('AB'), index=['x', 'y'])
df2 = pd.DataFrame([[5, 6], [7, 8]], columns=list('AB'), index=['x', 'y'])

df.append(df2)
```

Indonesia AI
AI for Everyone, AI for Indonesia

Terima Kasih!

Indonesia AI | AI for Everyone, AI for Indonesia

contact@aiforindonesia.org