Python for Beginners

Archer Infotech , PUNE

Python - Pandas

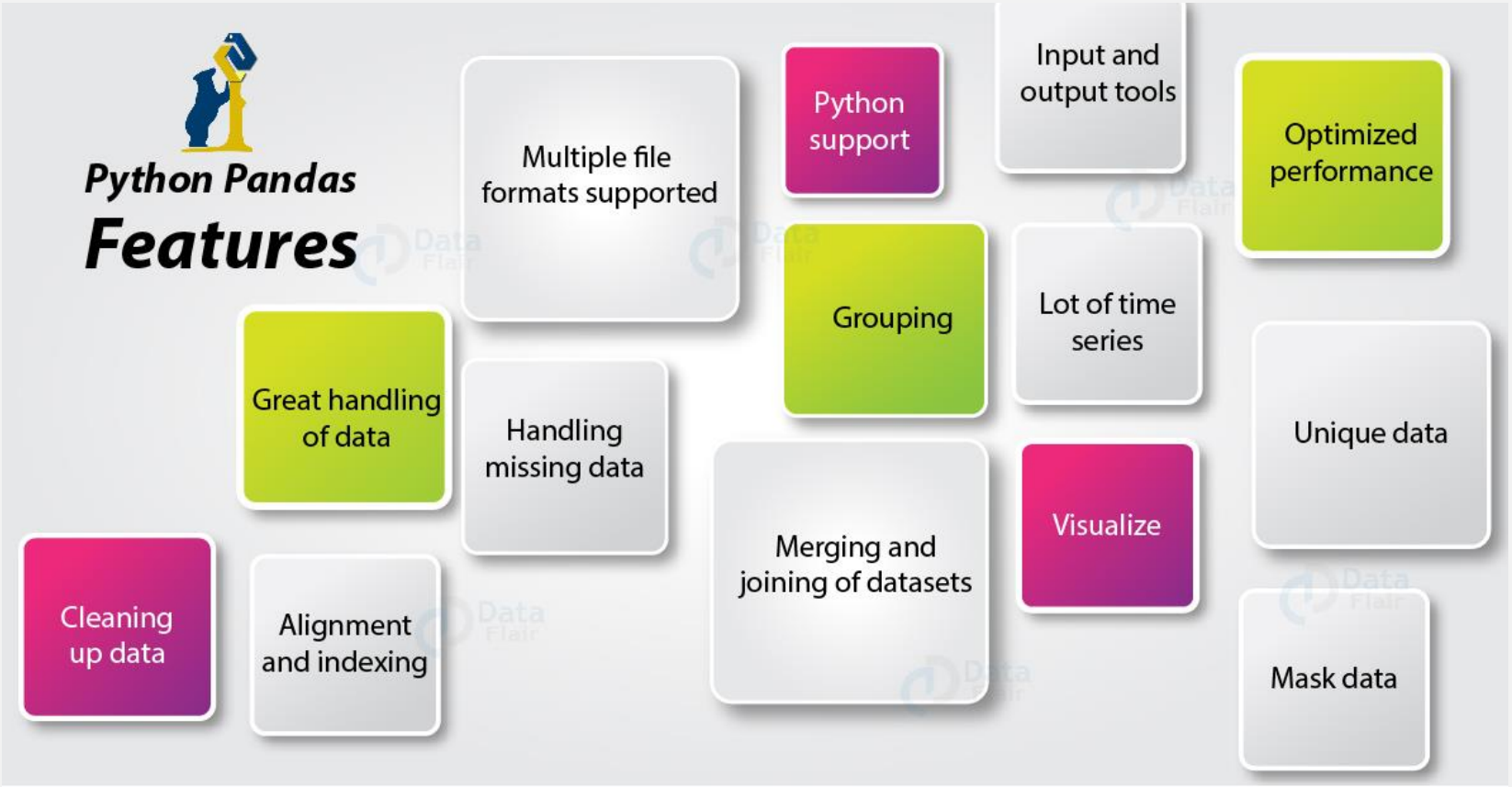
What is Pandas ?

- 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.

- This package comprises many data structures and tools for effective data manipulation and analysis. Python Pandas is used everywhere including commercial and academic sectors and in fields like economics, finance, analytics, statistics, etc.

- Pandas library is built on top of Numpy, meaning Pandas needs Numpy to operate.

Pandas Features



Installation & importing pandas

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

```
pip install pandas
```

- **The very first and the most important operation is to import Python Pandas library properly.**

```
import pandas as pd
```

Pandas data structures

- **Pandas deals with following three data structures**

  1. **Series**
  2. **DataFrame**
  3. **Panel**

These data structures are built on top of numpy array ,
which means they are faster

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.
e.g. DataFrame is a container of Series, Panel is a container of DataFrame

Pandas data structures

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.

Mutability

All Pandas data structures are value mutable (can be changed) and except Series all are size mutable. Series is size immutable.

Pandas data structures

Series :- **Is a 1-D array like structure with homogenous data**

Homogeneous data
Size immutable
Values are mutable

10	23	56	17	52	61	73	90	26	72

Pandas data structures

DataFrame :- **Is a 2-D array with heterogeneous data**

Heterogeneous data
Size & values are mutable

Name	Age	Gender	Rating
Steve	32	Male	3.45
Lia	28	Female	4.6
Vin	45	Male	3.9
Katie	38	Female	2.7

Pandas data structures

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.

Heterogeneous data
Size Mutable
Data Mutable

Pandas Series

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.

A pandas Series can be created using the following constructor −

```
pandas.Series( data, index, dtype, copy)
```

A series can be created using various inputs like −

Scalar Values
List
Dictionaries
Numpy Arrays

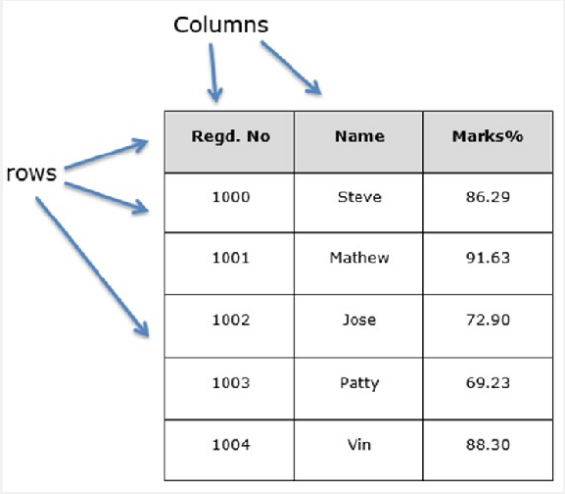
Pandas Series Basic Attributes

Attributes	Description
Series.index	Defines the index of the Series.
Series.shape	It returns a tuple of shape of the data.
Series.dtype	It returns the data type of the data.
Series.size	It returns the size of the data.
Series.empty	It returns True if Series object is empty, otherwise returns false.
Series.hasnans	It returns True if there are any NaN values, otherwise returns false.
Series.nbytes	It returns the number of bytes in the data.
Series.ndim	It returns the number of dimensions in the data.
Series.itemsize	It returns the size of the datatype of item.

Pandas Data Frame

A Data frame is a two-dimensional data structure, i.e., data is aligned in a tabular fashion in rows and columns.

Potentially columns are of different types
Size – Mutable
Labeled axes (rows and columns)
Can Perform Arithmetic operations on rows and columns



Regd. No	Name	Marks%
1000	Steve	86.29
1001	Mathew	91.63
1002	Jose	72.90
1003	Patty	69.23
1004	Vin	88.30

Pandas Data Frame

A pandas Data Frame can be created using the following constructor −

```
pandas.DataFrame( data, index, columns, dtype, copy)
```

A data frame can be created using various inputs like −

- Lists
- Dict
- Series
- Numpy ndarrays
- Another DataFrame

Pandas Data Frame Basic Functionality

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.
10	**tail()** Returns last n rows.

Pandas Descriptive Statistics

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

Pandas Iteration

To iterate over the rows of the DataFrame, we can use the following functions –

- **iteritems()** – to iterate over the (key,value) pairs
- **iterrows()** – iterate over the rows as (index,series) pairs
- **itertuples()** – iterate over the rows as namedtuples

Pandas Indexing & Selecting Data

Sr.No	Indexing & Description	
1	**.loc()**	Label based
2	**.iloc()**	Integer based
3	**.ix()**	Both Label and Integer based

THANK YOU !!!

Amol Patil - 9822291613