

Pandas (Panel data)

1. Pandas is a python package, first introduced in Jan 2008.
2. it represents multi-dim structure data with labels.
3. it is used in data analysis and data preprocessing(manipulation).
4. it is not the part of default python and we need to download it
`pip install pandas`

Variables:

```
df.shape
df.columns
df.size
df.dtypes
```

Methods for Analysis:

```
df.head()
df.tail()
df.describe()
df.info()
df.col_label.unique()
df.col_label.value_counts()
df.col_label.isnull()
df.col_label.isnull().sum()
df.isnull()
df.isnull().sum()
df.sort_values()
df.fillna()
df.ffill()
df.bfill()
df.replace()
df.dropna()
df.drop_duplicates()
df.groupby()
```

Commonly used 2 types(objects) in pandas:

1. Series
2. DataFrame

```
DataFrame
-----
```

1. Generally represents 2 dim array in tabular form.
2. we may provide label row & column indexing
3. it is mutable in size

1. Creation of DataFrame
 - A. from list of lists
 - B. from list of dicts
 - C. from dict of lists
 - D. from tuple of list (loop with append)
 - E. from other DataFrame

```
In [1]: import numpy as np
import pandas as pd
```

Panel Data : data of an org or individual based on time

structured data

```
In [2]: df = pd.DataFrame()
print(df)
```

```
Empty DataFrame
Columns: []
Index: []
```

```
In [3]: df = pd.DataFrame([['Sonu', 20, 2000], ['monu', 21, 19000]])
df
```

```
Out[3]:
```

	0	1	2
0	Sonu	20	2000
1	monu	21	19000

labels ----> column name

index ----> row name

```
In [4]: df = pd.DataFrame([['Sonu', 20, 2000], ['monu', 21, 19000]], index=["r1", "r2"])
df
```

```
Out[4]:
```

	0	1	2
r1	Sonu	20	2000
r2	monu	21	19000

```
In [5]: df = pd.DataFrame([['Sonu', 20, 2000], ['monu', 21, 19000]], index=["r1", "r2"], columns=["a1", 'a2', 'a3'])
df
```

```
Out[5]:
```

	a1	a2	a3
r1	Sonu	20	2000
r2	monu	21	19000

```
In [6]: df = pd.DataFrame({'empid':[10, 20, 30], 'salary':[10000, 19000, 20000]})
df
```

```
Out[6]:
```

	empid	salary
0	10	10000
1	20	19000
2	30	20000

dataframe each column is series

series is represent in 1d array

```
In [7]: empid = df['empid']  
empid
```

```
Out[7]: 0    10  
        1    20  
        2    30  
        Name: empid, dtype: int64
```

```
In [8]: type(df)
```

```
Out[8]: pandas.core.frame.DataFrame
```

```
In [9]: type(empid)
```

```
Out[9]: pandas.core.series.Series
```

```
In [10]: df
```

```
Out[10]:
```

	empid	salary
0	10	10000
1	20	19000
2	30	20000

```
In [21]: df['name'] = ['Sonu', 'Monu', 'Tonu']  
df
```

```
Out[21]:
```

	empid	org	salary	name	area
0	10	10	10000	Sonu	up
1	20	20	19000	Monu	delhi
2	30	30	20000	Tonu	HP

```
In [12]: df['area']=pd.Series(['up', 'delhi', 'HP'])  
df
```

```
Out[12]:
```

	empid	salary	name	area
0	10	10000	Sonu	up
1	20	19000	Monu	delhi
2	30	20000	Tonu	HP

```
In [13]: df[0:1]
```

```
Out[13]:
```

	empid	salary	name	area
0	10	10000	Sonu	up

```
In [14]: df.insert(1, 'poc', [10, 20, 30])
df
```

```
Out[14]:
```

	empid	poc	salary	name	area
0	10	10	10000	Sonu	up
1	20	20	19000	Monu	delhi
2	30	30	20000	Tonu	HP

Operation:

inplace: change inn existing

copy: return a new of

```
In [15]: df1 = df.rename({'poc':'org'}, axis=1)
df1
```

```
Out[15]:
```

	empid	org	salary	name	area
0	10	10	10000	Sonu	up
1	20	20	19000	Monu	delhi
2	30	30	20000	Tonu	HP

```
In [16]: df
```

```
Out[16]:
```

	empid	poc	salary	name	area
0	10	10	10000	Sonu	up
1	20	20	19000	Monu	delhi
2	30	30	20000	Tonu	HP

```
In [17]: df.rename({'poc':'org'}, axis=1, inplace=True)
df
```

```
Out[17]:
```

	empid	org	salary	name	area
0	10	10	10000	Sonu	up
1	20	20	19000	Monu	delhi
2	30	30	20000	Tonu	HP

```
In [18]: df.rename({0: 11, 1:12, 2:'ducat'}, axis=0)
```

```
Out[18]:
```

	empid	org	salary	name	area
11	10	10	10000	Sonu	up
12	20	20	19000	Monu	delhi
ducat	30	30	20000	Tonu	HP

Change postion of column

```
In [19]: df.reindex(["empid", 'area', 'name', 'salary'], axis=1)
```

```
Out[19]:
```

	empid	area	name	salary
0	10	up	Sonu	10000
1	20	delhi	Monu	19000
2	30	HP	Tonu	20000

```
In [20]: df.reindex([2,0,1], axis=0)
```

```
Out[20]:
```

	empid	org	salary	name	area
2	30	30	20000	Tonu	HP
0	10	10	10000	Sonu	up
1	20	20	19000	Monu	delhi

axis 0 means column

axis 1 means row

Date Frame Operation:

```
df.at[rowlabel,collabel]  
df.iat[rowpos,colpos]
```

How to update an element:

```
df.at[rowlabel,collabel]=value  
df.iat[rowpos,colpos]=value
```

How to select a column:

```
df['col_label']  
df.col_label
```

How to select multiple columns:

```
df[['col_label1','col_label2']]
```

DataFrame Slicing:

```
label slicing: df.loc[row_label_slicing,col_label_slicing]  
pos slicing: df.iloc[row_pos_slicing,col_pos_slicing]  
boolean slicing: df[condition] df[(condition1) & (condition2)] df[(condition1) | (condition2)]
```

How to append a dataframe into existing dataframe:

```
df3=df1.append(df2)  
df3=df1.append(df2,ignoreIndex=True)
```

How to convert type of a column:

```
df['col']=pd.to_type(df.col) ex: df['col']=pd.to_numeric(df.col)
```