**Pandas**

**Chức năng :**

* Dọn dẹp dữ liệu : xử lý các giá trị bị thiếu, trung lặp và định dạng dữ liệu chính xác
* Xử lý dữ liệu
* Phân tích
* Hình dung
* Phân tích chuỗi thời gian : xử lý và thao tác dữ liệu chuỗi thời gian

**Có bao nhiêu thành phần chính ?**

Có 2 thành phần chính là Series( 1 chiều ) và DataFrame ( dạng bảng, 2 chiều )

Series giống như mảng numpy nhưng có nhãn

Numpy các thành phần phải chung 1 kiểu dữ liệu còn series thì các thành phần khác kiểu dữ liệu cũng được.

Series hoàn toàn là 1 chiều có thể chứa bất kì loại dữ liệu nào.

Tạo series :

1. Tạo từ 1 danh sách

Pd.Series( data = [ -5, 2, 3, 21, 6 ], index = [‘a’, ‘b’, ‘c’, ‘d’, ‘e’ ] )

Index là không bắt buộc ( không ghi thì sẽ tự động gán )

1. Tạo từ 1 từ điển

dictionary = { 'a' : 1, 'b' : 3 , 'c' : 10 , 'd' : 2}

series = pd.Series( dictionary )

1. Tạo từ 1 mảng numpy

numpy\_arr = np.arange( 1, 100, 4)

series2 = pd.Series( numpy\_arr )

1. **Tạo từ 1 số ( một Scalar )**

series4 = pd.Series ( data = 3.14, index = [1, 2, 3, 4, 5, 6])

**Các thuộc tính của Series**

Thuộc tính name

* + - Ví dụ : series1 = pd.Series( data = np.random.randn(5), name = ‘random\_series’ )
    - Muốn xem tên : series1.name
    - Đổi tên : series1.rename( “tên mới muốn đổi” )

Thuộc tính index :

* + - Ví dụ : series2 = pd.Series( data = np.random.randn( 5 )
    - Xem chỉ số : series2.index

**Lấy thành phần trong series theo chỉ mục**

**Dùng iloc[]**

* + - Series.iloc[ [ 1 ,2, 3 ] ]
    - So sánh với loc : thành phần trong iloc phải là số, còn với loc thì không bắt buộc.

**Slicing series bằng dấu ngoặc vuông :**

* + - Series[ ‘a’ : ‘c ’] lấy các thành phần theo mảng từ chỉ mục a đến c
    - Series[ [ ‘a’, ‘c’, ‘d’ ] ] lấy các thành phần theo mảng chỉ mục

**Thoả điều kiện** :

* + - Series[ Series > Series.mean() ]

**Các phương thức trong Series**

**Cộng ( Cộng 2 series ) :**

* + - Nhãn giống nhau thì cộng vào chung
    - Khác nhãn thì cộng vào ra NaN ( NaN thuộc kiểu dữ liệu số )

**Tại sao NaN nhìn vào là chữ nhưng lại là kiểu dữ liệu float.**

**Exp :**

**Add**

**Pow**

**Mean**

**Astype**

**Ví dụ : s1.add( 3.14 ).pow(2).mean().astype(int)**

**DataFrame**

**DataFrame :** là 1 dictionary của các Series với key là mảng index còn value là dữ liệu trong các cột còn lại.

**Tạo DataFrames :**

pd.DataFrame( )

**Tạo từ 1 list :**

pd.DataFrame( [ ………]

[ ……… ] )

Index = [ ‘R1’, ‘R2’ , ‘R3’ ]

Columns = [ ‘C1’, ‘C2’, ‘C3’ ]

**Tạo từ 1 dictionary :**

ở DataFrame tạo từ dictionary thì key là chỉ số cột. ( do là 2 chiều )

ở Series tạo từ dictionary thì key là chỉ số dòng.

**Indexing and slicing DataFrames**

Có 1 số cách chính để chọn dữ liệu từ DataFrame

1. [ ]
2. .loc[ ]
3. .iloc[ ]
4. Boolean idexing
5. .query[ ]

**Indexing với []**

**Indexing với .loc và .iloc**

**Boolean indexing**

**Indexing với .query( )**

**Các phương thức phổ biến trên dataframe**

**Df.groupby**