PANDAS

OVERVIEW

Pandas

- Panel Data Module
- objects: series and dataframes
- series similar to a table column
- dataframe similar to a table
- designed to manage indexed data (like SQL)

A Numerical Dataset

object x_i	Height (H)	Weight (W)	Foot (F)	Label (L)
x_1	5.00	100	6	green
x_2	5.50	150	8	green
x_3	5.33	130	7	green
x_4	5.75	150	9	green
x_5	6.00	180	13	red
x_6	5.92	190	11	red
x_7	5.58	170	12	red
x_8	5.92	165	10	red

- $N = 8$ items
- $M = 3$ (unscaled) attributes

Code for the Dataset

```
import pandas as pd
data = pd.DataFrame(
    {'id':[ 1,2,3,4,5,6,7,8],
    'Label':['green','green','green','green',
                  'red','red','red','red'],
    'Height':[5,5.5,5.33,5.75,6.00,5.92,5.58,5.92],
    'Weight':[100,150,130,150,180,190,170,165],
    'Foot':[6, 8, 7, 9, 13, 11, 12, 10]},
    columns=['id','Height','Weight','Foot','Label'])
```

```
ipdb> data
  id Height Weight Foot Label
0  1   5.00    100    6  green
1  2   5.50    150    8  green
2  3   5.33    130    7  green
3  4   5.75    150    9  green
4  5   6.00    180   13    red
5  6   5.92    190   11    red
6  7   5.58    170   12    red
7  8   5.92    165   10    red
```

Alternative Approach

```
data = pd.DataFrame(
  data = [ [1, 5, 100, 6, 'green'],
           [2,5.5, 150, 8, 'green'],
           [3, 5.33, 130, 7, 'green'],
           [4, 5.75, 150, 9, 'green'],
           [5, 6, 180, 13, 'red'],
           [6, 5.92, 190, 11, 'red'],
           [7, 5.58, 170, 12, 'red'],
           [8, 5.92, 165, 10, 'red']],
  columns = ['id','Height','Weight','Foot','Label'] )
```

```
ipdb> data
  id Height Weight Foot Label
0  1   5.00    100    6  green
1  2   5.50    150    8  green
2  3   5.33    130    7  green
3  4   5.75    150    9  green
4  5   6.00    180   13    red
5  6   5.92    190   11    red
6  7   5.58    170   12    red
7  8   5.92    165   10    red
```

Some Observations

- data in different shapes (dictionary, lists)
- different data types
- columns have custom names
- index can be in different formats

Typical Operations

- index values

```
> data.index
RangeIndex(start=0, stop=8, step=1)
```

- column names

```
> data.columns
Index(['id','Height','Weight','Foot',
       'Label'], dtype='object')
```

Data Selection

- selection via index:

1. .loc by label
2. .iloc by position

```
> data.iloc[5]
id            6
Height     5.92
Weight      190
Foot         11
Label       red
Name: 5, dtype: object
```

Data Selection

- selection of multiple indices

```
> data.iloc[[5,7]]
   id  Height  Weight  Foot Label
5   6    5.92     190    11   red
7   8    5.92     165    10   red
```

- se;ection via index object

```
> data.iloc[data.index[1:7:2]]
   id  Height  Weight  Foot  Label
1   2    5.50     150     8  green
3   4    5.75     150     9  green
5   6    5.92     190    11    red
```

Statistical Functions

- apply statistical functions

```
> data[['Height','Weight',
            'Foot']].mean()
Height      5.625
Weight    154.375
Foot        9.500
dtype: float64
```

Lambda Functions

- apply lambda functions

```
> data[['Height',
      'Weight']].apply(lambda x: x**2)
    Height  Weight
0  25.0000   10000
1  30.2500   22500
2  28.4089   16900
3  33.0625   22500
4  36.0000   32400
5  35.0464   36100
6  31.1364   28900
7  35.0464   27225
```

Adding Column(s)

```
> data['n_col']=['a','b','c','d',
                 'e','f','g','h']
```

```
> data
```

	id	Height	Weight	Foot	Label	n_col
0	1	5.00	100	6	green	a
1	2	5.50	150	8	green	b
2	3	5.33	130	7	green	c
3	4	5.75	150	9	green	d
4	5	6.00	180	13	red	e
5	6	5.92	190	11	red	f
6	7	5.58	170	12	red	g
7	8	5.92	165	10	red	h

Dropping Column(s)

```
> data.drop(['n_col'],axis=1,inplace=True
> data
   id  Height  Weight  Foot  Label
0   1    5.00     100     6  green
1   2    5.50     150     8  green
2   3    5.33     130     7  green
3   4    5.75     150     9  green
4   5    6.00     180    13    red
5   6    5.92     190    11    red
6   7    5.58     170    12    red
7   8    5.92     165    10    red
```

- axis: 1-columns, 0 - rows

Desribing the Dataset

```
import pandas as pd
data = pd.DataFrame(
    {'id':[ 1,2,3,4,5,6,7,8],
    'Label':['green','green','green','green',
                  'red','red','red','red'],
    'Height':[5,5.5,5.33,5.75,6.00,5.92,5.58,5.92],
    'Weight':[100,150,130,150,180,190,170,165],
    'Foot':[6, 8, 7, 9, 13, 11, 12, 10]},
    columns=['id','Height','Weight','Foot','Label'])
```

```
ipdb> data.describe()
             id    Height      Weight      Foot
count   8.00000  8.000000    8.000000   8.00000
mean    4.50000  5.625000  154.375000   9.50000
std     2.44949  0.343428   28.962722   2.44949
min     1.00000  5.000000  100.000000   6.00000
25%     2.75000  5.457500  145.000000   7.75000
50%     4.50000  5.665000  157.500000   9.50000
75%     6.25000  5.920000  172.500000  11.25000
max     8.00000  6.000000  190.000000  13.00000
```

A Dataset Illustration