Data exploration with Python: Basic Pandas

Option GIS-Python

hes·so business.

Jean-Paul Calbimonte

School of Management

Bachelor of Science HES-SO (BSc) in Business Information Technology

swissuniversities

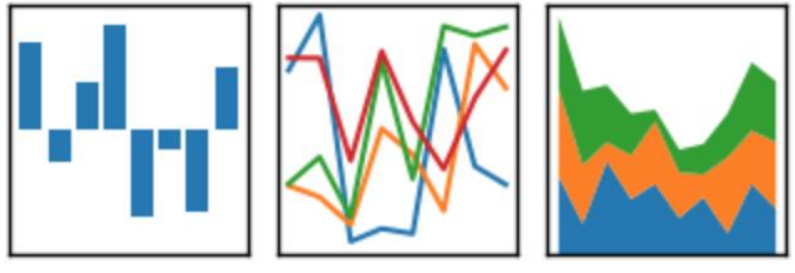
> Pandas

What is Pandas? http://pandas.pydata.org

- python package
- fast, flexible, and expressive **data structures**
- working with "**relational**" or "**labeled**" data
- high-level building block for doing practical, real world **data analysis**
- powerful and **flexible** open source data analysis / **manipulation** tool

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



> Pandas

When to use Pandas?

- Well suited for many **different kinds** of data
- Tabular data with **heterogeneously-typed** columns, as in an SQL table or Excel spreadsheet
- Ordered and unordered (not necessarily fixed-frequency) **time series** data.
- Arbitrary **matrix data** (homogeneously typed or heterogeneous) with row and column labels
- Any form of observational / **statistical data** sets.
- The data actually need not be labeled at all.

> Pandas

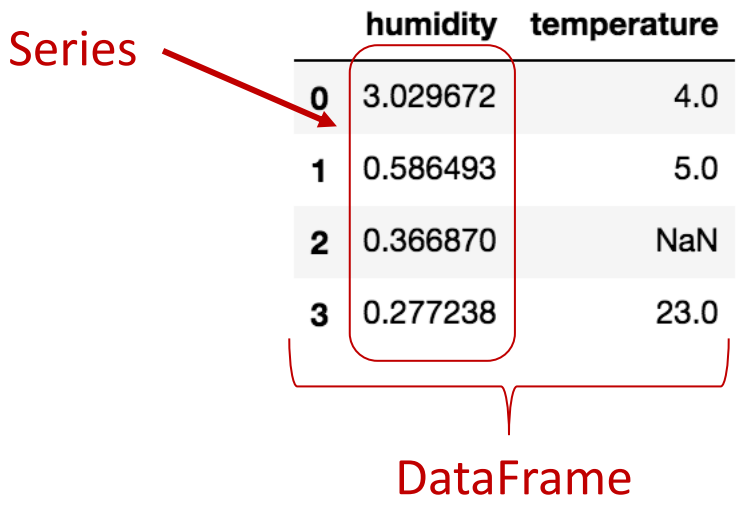
What can Pandas do?

- Handling of **missing data** (NaN) in floating point/non-floating point data
- Size **mutability**: columns can be inserted/deleted
- Powerful, flexible **group by functionality** to perform split-apply-combine operations on data sets, for both aggregating and transforming data
- Intelligent label-based **slicing**, fancy **indexing**, and subsetting of large data sets
- Intuitive **merging** and **joining** data sets
- Flexible reshaping and **pivoting** of data sets
- **Hierarchical labeling** of axes
- Robust **IO tools** for loading data from flat files CSV, Excel, HDFS5)
- **Time series-specific** functionality: date range generation and frequency conversion, moving window statistics, moving window linear regressions, etc.

> Pandas types

Series & DataFrame

Series	1D labeled homogeneously-typed array
DataFrame	General 2D labeled, size-mutable tabular structure with potentially heterogeneously-typed column



> Series and DataFrames

```
import pandas as pd
import numpy as np
s= pd.Series([4,5,np.nan,23,])
s
```

```
0   4.0
1   5.0
2   NaN
3   23.0
dtype: float64
```

Series

```
df=pd.DataFrame({'temperature':s,
                 'humidity':np.random.randn(4)})
df
```

	humidity	temperature
0	3.029672	4.0
1	0.586493	5.0
2	0.366870	NaN
3	0.277238	23.0

DataFrame

> Reading files

```
hospitalsFile = "../data/healthValais/hospitals.csv"
hospitals=pd.read_csv(hospitalsFile,encoding='latin1')
hospitals
```

	CLASSE	ETABLISSEMENT	Adresse	numero	npa	ville	telephone	site_internet	RuleID	RuleID_1	RuleID_2	beds
0	1	HVS - Hpital psychiatrique de Malvoz	Route de Morgins	10.0	1870	Monthey	0800 012 210	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	30
1	1	HVS - Clinique de Saint-Am	Vers Saint-Am	10.0	1890	St-Maurice	027/604.66.55	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	20
2	1	HVS - Hpital de Martigny	Avenue de la Fusion	27.0	1920	Martigny	027/603.90.00	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	10
3	1	HVS - Hpital de Sion	Avenue du Grand-Champsec	80.0	1951	Sion	027/603.40.00	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	11
4	1	HVS - Institut Central des Hpitaux ICH	Avenue du Grand-Champsec	86.0	1951	Sion	027/603.47.00	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	15
5	1	HVS - Hpital de Sierre	Rue St-Charles	14.0	3960	Sierre	027/603.70.00	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	30

> Exploring data

```
hospitals.head(3)
```

	CLASSE	ETABLISSEMENT	Adresse	numero	npa	ville	telephone	site_internet	RuleID	RuleID_1	RuleID_2	beds
0	1	HVS - Hpital psychiatrique de Malvoz	Route de Morgins	10.0	1870	Monthey	0800 012 210	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	30
1	1	HVS - Clinique de Saint-Am	Vers Saint-Am	10.0	1890	St-Maurice	027/604.66.55	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	20
2	1	HVS - Hpital de Martigny	Avenue de la Fusion	27.0	1920	Martigny	027/603.90.00	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	10

```
hospitals.columns
Index([u'CLASSE', u'ETABLISSEMENT', u'Adresse', u'numero', u'npa', u'ville',
u'telephone', u'site_internet', u'RuleID', u'RuleID_1', u'RuleID_2',  u'beds'],
dtype='object')
```

> Exploring metadata

```
hospitals.index
RangeIndex(start=0, stop=22, step=1)
```

```
len(hospitals)
22
```

```
hospitals.shape
(22, 12)
```

```
hospitals.dtypes
CLASSE                int64
ETABLISSEMENT        object
Adresse              object
numero              float64
npa                   int64
ville                object
telephone            object
site_internet        object
RuleID              float64
RuleID_1             object
RuleID_2             object
beds                  int64
dtype: object
```

> Describing data

```
hospitals['Adresse']
 0   Route de Morgins
 1        Vers Saint-Am
 2   Avenue de la Fusion
 3        Avenue du Grand-Champsec
 4        Avenue du Grand-Champsec
 5        Rue St-Charles
 6        Route de la Moubra
 7        erlandstrasse
 8        Pflanzettastrasse
 9        Route de Morgins
 10       Chemin du Grand-Ch e
 11       Av. de la Prairie
 12       Rue Pr Fleuri
 13       Route du L man
 14       Willy-Sp hlerstrasse
 15       Impasse Palace Bellevue
 16       Rte de L'Astoria
 17       Avenue Grand-Champsec
 18       Boulevard Paderewski
 19       Avenue de Belmont
 20       erlandstrasse
 21       Impasse Clairmont
Name: Adresse, dtype: object
```

```
hospitals['beds'].mean()
28.727272727272727
```

```
hospitals['beds'].max()
50
```

```
hospitals.describe()
```

	CLASSE	numero	npa	RuleID	beds
count	22.0	21.000000	22.000000	0.0	22.000000
mean	1.0	27.476190	2728.909091	NaN	28.727273
std	0.0	31.512885	1035.833178	NaN	12.585637
min	1.0	1.000000	1800.000000	NaN	10.000000
25%	1.0	3.000000	1875.000000	NaN	20.500000
50%	1.0	14.000000	1951.000000	NaN	30.000000
75%	1.0	29.000000	3948.000000	NaN	40.000000
max	1.0	90.000000	3963.000000	NaN	50.000000

> Adding Series

```
hospitals['newBeds']=0
hospitals
```

net	RuleID	RuleID_1	RuleID_2	beds	newBeds
.ch	NaN	Rule_1	Rule_1	30	0
.ch	NaN	Rule_1	Rule_1	20	0
.ch	NaN	Rule_1	Rule_1	10	0
.ch	NaN	Rule_1	Rule_1	11	0

```
hospitals['newBeds']=hospitals['beds']*2
hospitals.head()
```

net	RuleID	RuleID_1	RuleID_2	beds	newBeds
s.ch	NaN	Rule_1	Rule_1	30	60
s.ch	NaN	Rule_1	Rule_1	20	40
s.ch	NaN	Rule_1	Rule_1	10	20
s.ch	NaN	Rule_1	Rule_1	11	22

> Selecting data

```
hospitals[5:8]
```

	CLASSE	ETABLISSEMENT	Adresse	numero	npa	ville	telephone	site_internet	RuleID	RuleID_1	RuleID_2	beds	newBed
5	1	HVS - Hpital de Sierre	Rue St-Charles	14.0	3960	Sierre	027/603.70.00	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	30	6
6	1	HVS - Centre Valaisan de Pneumologie (CVP)	Route de la Moubra	87.0	3963	Crans-Montana	027/603.80.00	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	12	2
7	1	HVS - Hpital de Brigue	berlandstrasse	14.0	3900	Brig	027/604.33.33	http://www.spitalvs.ch/de/spital-wallis/stando...	NaN	Rule_1	Rule_1	45	9

```
hospitals.loc[11]
```

```
CLASSE                                                1
ETABLISSEMENT    HRC - Hpital de Vevey la Providence
Adresse                                 Av. de la Prairie
numero                                                3
npa                                                1800
ville                                             Vevey
telephone                                 021/977.55.55
site_internet    http://www.hopitalrivierachablais.ch
RuleID                                              NaN
RuleID_1                                         Rule_1
RuleID_2                                         Rule_1
beds                                                 34
newBeds                                              68
Name: 11, dtype: object
```

> Selecting data

```
hospitals.loc[9:12, ['ETABLISSEMENT','beds']]
```

	ETABLISSEMENT	beds
9	HRC - Hpital de Monthey	45
10	HRC - Hpital d'Aigle	23
11	HRC - Hpital de Vevey la Providence	34
12	Clinique de Valre	12

```
hospitals['ville'].unique()

array([u'Monthey', u'St-Maurice', u'Martigny', u'Sion', u'Sierre',
u'Crans-Montana', u'Brig', u'Visp', u'Aigle', u'Vevey', u'Saxon',
u'Leukerbad', u'Montreuy'], dtype=object)
```

> Filtering & Sorting data

```
hospitals.loc[(hospitals['Adresse'].str.startswith('Avenue')) &
              (hospitals['beds']<15)]
```

	CLASSE	ETABLISSEMENT	Adresse	numero	npa	ville	telephone	site_internet	RuleID	RuleID_1	RuleID_2	beds	newBe
2	1	HVS - Hpital de Martigny	Avenue de la Fusion	27.0	1920	Martigny	027/603.90.00	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	10	
3	1	HVS - Hpital de Sion	Avenue du Grand-Champsec	80.0	1951	Sion	027/603.40.00	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	11	

```
hospitals.sort_values(by='npa',ascending=False).head()
```

	CLASSE	ETABLISSEMENT	Adresse	numero	npa	ville	telephone	site_internet	RuleID	RuleID_1	RuleID_2	beds	newBe
21	1	HUG-Clinique de Crans-Montana	Impasse Clairmont	2.0	3963	Crans-Montana	027/485.61.11	http://www.hug-ge.ch/crans-montana	NaN	Rule_1	Rule_1	22	
16	1	Luzerner Hhenklinik Montana	Rte de L'Astoria	2.0	3963	Crans-Montana	027/485.81.81	http://www.lhm.ch	NaN	Rule_1	Rule_1	22	
6	1	HVS - Centre Valaisan de Pneumologie (CVP)	Route de la Moubra	87.0	3963	Crans-Montana	027/603.80.00	http://www.hopitalduvalais.ch	NaN	Rule_1	Rule_1	12	
15	1	Berner Klinik Montana	Impasse Palace Bellevue	1.0	3963	Crans-Montana	027/485.51.21	http://www.bernerklinik.ch	NaN	Rule_1	Rule_1	44	

> Grouping data

```
groups=hospitals.groupby('ville')

groups.get_group('Brig')
```

	CLASSE	ETABLISSEMENT	Adresse	numero	npa	ville	telephone	site_internet	RuleID	RuleID_1	RuleID_2	beds	newBeds
7	1	HVS - Hpital de Brigue	berlandstrasse	14.0	3900	Brig	027/604.33.33	http://www.spitalvs.ch/de/spital-wallis/stando...	NaN	Rule_1	Rule_1	45	90
20	1	HVS - Psychiatriezentrum Oberwallis	berlandstrasse	14.0	3900	Brig	027/604.33.33	http://www.hopitalvs.ch/de/unsere-fachbereiche...	NaN	Rule_1	Rule_1	42	84

> Iterating over data

```
for idx,row in hospitals.iterrows():
    print(row['ville'])
    print(idx)
```

```
Monthey
0
St-Maurice
1
Martigny
2
Sion
3
Sion
4
Sierre
5
Crans-Montana
6
Brig
7
Visp
8
Monthey
9
Aigle
10
```

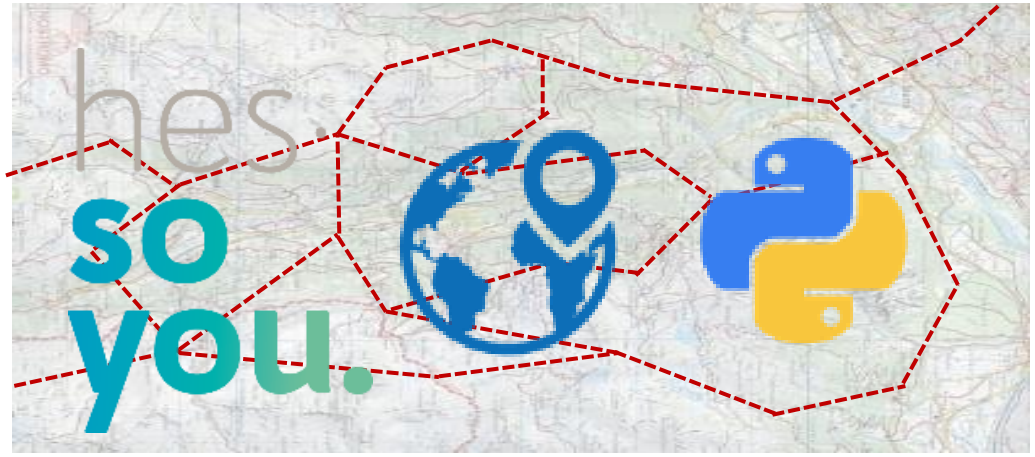
> Writing DataFrames into a file

```
hospitalsModif=hospitals.loc[:,['ETABLISSEMENT','newBeds','ville']]

hospitalsModif.to_csv('hospitals_modif.csv',sep=',',encoding='utf-8')
```

hospitals_modif.csv

```
,ETABLISSEMENT,newBeds,ville
0,HVS - Hpital psychiatrique de Malvoz,60,Monthey
1,HVS - Clinique de Saint-Am,40,St-Maurice
2,HVS - Hpital de Martigny,20,Martigny
3,HVS - Hpital de Sion,22,Sion
4,HVS - Institut Central des Hpitaux ICH,30,Sion
5,HVS - Hpital de Sierre,60,Sierre
6,HVS - Centre Valaisan de Pneumologie (CVP),24,Crans-Montana
7,HVS - Hpital de Brigue,90,Brig
8,HVS - Hpital de Vige,46,Visp
9,HRC - Hpital de Monthey,90,Monthey
10,HRC - Hpital d'Aigle,46,Aigle
11,HRC - Hpital de Vevey la Providence,68,Vevey
12,Clinique de Valre,24,Sion
13,Clinique CIC Valais,86,Saxon
14,Leukerbad Clinic,66,Leukerbad
15,Berner Klinik Montana,88,Crans-Montana
16,Luzerner Hhenklinik Montana,44,Crans-Montana
17,Clinique romande de radaptation (CRR SUVA),66,Sion
18,HRC - Hpital de Vevey le Samaritain,100,Vevey
19,HRC - Hpital de Montreux,66,Montreuy
20,HVS - Psychiatriezentrum Oberwallis,84,Brig
21,HUG-Clinique de Crans-Montana,44,Crans-Montana
```



Thank you for your attention.

School of Management
Route de la Plaine 2
3960 Sierre

hevs.ch/heg

swissuniversities