# pandas_start

July 2, 2023

## 1 DataFrames Pandas

Un Data Frame est un objet qui est présent dans la plupart des logiciels de traitements de données, c'est une matrice à 2 dimensions, chaque colonne a un type et toutes les cellules de cette colonne sont de ce type (nombre, dates, texte). Une cellule peut contenir une valeur manquante. On peut considérer chaque colonne comme les variables d'une table (pandas.Dataframe - cette page contient toutes les méthodes de la classe).

[1]:
```
from jyquickhelper import add_notebook_menu
add_notebook_menu()
```

[1]:
```
<IPython.core.display.HTML object>
```

Quelques liens : An Introduction to Pandas
Tous les exemples utilisent un jeu de données de l'ONU contenant par pays ("country"), par année ("year") et par secteur("code" ou "sub_item"), la valeur ajoutée monétaire du secteur dans ce pays cette année là ("VA1" ou "VA2"), la monnaie ("currency"), et la masse salariale du secteur cette année là ("WAGE1"). Les données uilisées pour l'exemple (accessible sur github) peuvent être remplacées par n'importe quelle table disponible sur le site : data.un.org.

### 1.1 Lecture et écriture sur disque

[2]:
```
import pandas
df = pandas.read_csv("UN_Data.csv", sep=",")
```

[3]:
```
df.head()
```

[3]:
```
     country                                sub_item  year        currency  \
0  Argentina  Agriculture, hunting, forestry fishing  1993  Argentine peso
1  Argentina                    Mining and quarrying  1993  Argentine peso
2  Argentina                           Manufacturing  1993  Argentine peso
3  Argentina       Electricity, gas and water supply  1993  Argentine peso
4  Argentina                            Construction  1993  Argentine peso

            VA1 code           VA2         WAGE1
0  1.214900e+10   AB  1.214900e+10  2.123000e+09
1  3.525000e+09    C  3.525000e+09  8.007000e+08
2  3.890700e+10    D  3.890700e+10  1.766600e+10
3  4.461000e+09    E  4.461000e+09  2.213000e+09
4  1.339300e+10    F  1.339300e+10  4.355000e+09
```

[4]:
```
df.to_excel("exemple.xlsx", index=False)
```

### 1.2 Manipulation basique

[5]:
```
df["VA1"]
```

[5]:
```
0       1.214900e+10
1       3.525000e+09
2       3.890700e+10
3       4.461000e+09
4       1.339300e+10
5       3.929400e+10
6       1.613400e+10
7       4.320200e+10
8       2.166090e+11
9       1.308500e+10
10      3.818000e+09
11      4.159600e+10
12      4.730000e+09
13      1.431100e+10
14      4.279800e+10
15      1.825100e+10
16      4.859900e+10
17      2.358460e+11
18      1.380850e+10
19      4.838400e+09
20      4.450210e+10
21      5.111000e+09
22      1.341400e+10
23      4.119850e+10
24      1.905990e+10
25      5.133940e+10
26      2.423343e+11
27      1.527000e+10
28      5.888900e+09
29      4.772340e+10
            …
4209    5.057137e+13
4210    2.880633e+12
4211    8.321318e+12
4212    8.204542e+12
4213    9.316080e+11
4214    3.400098e+12
4215    9.794983e+12
4216    5.716383e+12
4217    9.124555e+12
4218    6.033646e+13
4219    3.258018e+12
4220    1.591774e+13
4221    1.062152e+13
4222    1.137058e+12
4223    3.842038e+12
4224    1.168852e+13
4225    6.920770e+12
4226    1.100281e+13
4227    7.974029e+13
4228    3.872766e+12
```

```
4229    1.321210e+13
4230    1.095207e+13
4231    1.355214e+12
4232    5.032111e+12
4233    1.347735e+13
4234    8.063945e+12
4235    1.334013e+13
4236    8.804319e+13
4237    4.549229e+12
4238    1.074291e+14
Name: VA1, Length: 4239, dtype: float64
```

[6]:
```
df[1:3]
```

[6]:
```
     country              sub_item  year        currency           VA1 code  \
1  Argentina  Mining and quarrying  1993  Argentine peso  3.525000e+09    C
2  Argentina         Manufacturing  1993  Argentine peso  3.890700e+10    D

             VA2         WAGE1
1  3.525000e+09  8.007000e+08
2  3.890700e+10  1.766600e+10
```

La première ligne a pour indice 0 :

[7]:
```
df[0:3]
```

[7]:
```
     country                                sub_item  year        currency  \
0  Argentina  Agriculture, hunting, forestry fishing  1993  Argentine peso
1  Argentina                    Mining and quarrying  1993  Argentine peso
2  Argentina                           Manufacturing  1993  Argentine peso

             VA1 code           VA2         WAGE1
0  1.214900e+10   AB  1.214900e+10  2.123000e+09
1  3.525000e+09    C  3.525000e+09  8.007000e+08
2  3.890700e+10    D  3.890700e+10  1.766600e+10
```

[8]:
```
df[["country","year"]]
```

[8]:
```
        country  year
0     Argentina  1993
1     Argentina  1993
2     Argentina  1993
3     Argentina  1993
4     Argentina  1993
5     Argentina  1993
6     Argentina  1993
7     Argentina  1993
8     Argentina  1993
9     Argentina  1994
10    Argentina  1994
11    Argentina  1994
12    Argentina  1994
13    Argentina  1994
14    Argentina  1994
15    Argentina  1994
```

```
16    Argentina  1994
17    Argentina  1994
18    Argentina  1995
19    Argentina  1995
20    Argentina  1995
21    Argentina  1995
22    Argentina  1995
23    Argentina  1995
24    Argentina  1995
25    Argentina  1995
26    Argentina  1995
27    Argentina  1996
28    Argentina  1996
29    Argentina  1996
…           …    …
4209  Venezuela  1998
4210  Venezuela  1999
4211  Venezuela  1999
4212  Venezuela  1999
4213  Venezuela  1999
4214  Venezuela  1999
4215  Venezuela  1999
4216  Venezuela  1999
4217  Venezuela  1999
4218  Venezuela  1999
4219  Venezuela  2000
4220  Venezuela  2000
4221  Venezuela  2000
4222  Venezuela  2000
4223  Venezuela  2000
4224  Venezuela  2000
4225  Venezuela  2000
4226  Venezuela  2000
4227  Venezuela  2000
4228  Venezuela  2001
4229  Venezuela  2001
4230  Venezuela  2001
4231  Venezuela  2001
4232  Venezuela  2001
4233  Venezuela  2001
4234  Venezuela  2001
4235  Venezuela  2001
4236  Venezuela  2001
4237  Venezuela  2002
4238  Venezuela  2002

[4239 rows x 2 columns]
```

Documentation describe :

[9]:
```
df.describe()
```

[9]:
```
              year           VA1           VA2         WAGE1
count  4239.000000  4.239000e+03  4.233000e+03  4.239000e+03
```

```
mean   1989.912715  1.802346e+12  1.801822e+12  6.371376e+11
std      11.140271  1.086676e+13  1.087452e+13  3.957470e+12
min    1966.000000  2.000000e+00  0.000000e+00  1.000000e+00
25%    1981.000000  8.538500e+09  8.513000e+09  2.773500e+09
50%    1991.000000  4.565900e+10  4.565900e+10  1.555100e+10
75%    1999.000000  2.597610e+11  2.601340e+11  9.198150e+10
max    2010.000000  2.711389e+14  2.711389e+14  9.220360e+13
```

[10]:
```
df.loc[0]
```

[10]:
```
country                                   Argentina
sub_item    Agriculture, hunting, forestry fishing
year                                           1993
currency                             Argentine peso
VA1                                      1.2149e+10
code                                             AB
VA2                                      1.2149e+10
WAGE1                                     2.123e+09
Name: 0, dtype: object
```

[11]:
```
dfy = df.set_index("year")
dfy.loc[1993]
```

[11]:
```
        country                                          sub_item  \
year
1993  Argentina             Agriculture, hunting, forestry fishing
1993  Argentina                                Mining and quarrying
1993  Argentina                                        Manufacturing
1993  Argentina                  Electricity, gas and water supply
1993  Argentina                                         Construction
1993  Argentina  Wholesale retail trade, repair of motor vehicl…
1993  Argentina              Transport, storage and communications
1993  Argentina  Financial intermediation real estate, renting …
1993  Argentina                                        Total Economy
1993    Bolivia                                        Total Economy
1993     Brazil             Agriculture, hunting, forestry fishing
1993     Brazil                                Mining and quarrying
1993     Brazil                                        Manufacturing
1993     Brazil                  Electricity, gas and water supply
1993     Brazil                                         Construction
1993     Brazil  Wholesale retail trade, repair of motor vehicl…
1993     Brazil              Transport, storage and communications
1993     Brazil  Financial intermediation real estate, renting …
1993     Brazil                                        Total Economy
1993      Chile                                        Total Economy
1993   Colombia             Agriculture, hunting, forestry fishing
1993   Colombia                                Mining and quarrying
1993   Colombia                                        Manufacturing
1993   Colombia                  Electricity, gas and water supply
1993   Colombia                                         Construction
1993   Colombia  Wholesale retail trade, repair of motor vehicl…
1993   Colombia              Transport, storage and communications
1993   Colombia  Financial intermediation real estate, renting …
1993   Colombia                                        Total Economy
```

| | | |
|---|---|---|
| 1993 | Denmark | Agriculture, hunting, forestry fishing |
| … | … | … |
| 1993 | Spain | Financial intermediation real estate, renting … |
| 1993 | Spain | Total Economy |
| 1993 | Sweden | Agriculture, hunting, forestry fishing |
| 1993 | Sweden | Mining and quarrying |
| 1993 | Sweden | Manufacturing |
| 1993 | Sweden | Electricity, gas and water supply |
| 1993 | Sweden | Construction |
| 1993 | Sweden | Wholesale retail trade, repair of motor vehicl… |
| 1993 | Sweden | Transport, storage and communications |
| 1993 | Sweden | Financial intermediation real estate, renting … |
| 1993 | Sweden | Total Economy |
| 1993 | Thailand | Agriculture, hunting, forestry fishing |
| 1993 | Thailand | Mining and quarrying |
| 1993 | Thailand | Manufacturing |
| 1993 | Thailand | Electricity, gas and water supply |
| 1993 | Thailand | Construction |
| 1993 | Thailand | Wholesale retail trade, repair of motor vehicl… |
| 1993 | Thailand | Transport, storage and communications |
| 1993 | Thailand | Financial intermediation real estate, renting … |
| 1993 | Thailand | Total Economy |
| 1993 | Turkey | Total Economy |
| 1993 | Venezuela | Agriculture, hunting, forestry fishing |
| 1993 | Venezuela | Mining and quarrying |
| 1993 | Venezuela | Manufacturing |
| 1993 | Venezuela | Electricity, gas and water supply |
| 1993 | Venezuela | Construction |
| 1993 | Venezuela | Wholesale retail trade, repair of motor vehicl… |
| 1993 | Venezuela | Transport, storage and communications |
| 1993 | Venezuela | Financial intermediation real estate, renting … |
| 1993 | Venezuela | Total Economy |

| year | currency | VA1 | code | VA2 | WAGE1 |
|---|---|---|---|---|---|
| 1993 | Argentine peso | 1.214900e+10 | AB | 1.214900e+10 | 2.123000e+09 |
| 1993 | Argentine peso | 3.525000e+09 | C | 3.525000e+09 | 8.007000e+08 |
| 1993 | Argentine peso | 3.890700e+10 | D | 3.890700e+10 | 1.766600e+10 |
| 1993 | Argentine peso | 4.461000e+09 | E | 4.461000e+09 | 2.213000e+09 |
| 1993 | Argentine peso | 1.339300e+10 | F | 1.339300e+10 | 4.355000e+09 |
| 1993 | Argentine peso | 3.929400e+10 | GH | 3.929400e+10 | 1.092000e+10 |
| 1993 | Argentine peso | 1.613400e+10 | I | 1.613400e+10 | 6.213000e+09 |
| 1993 | Argentine peso | 4.320200e+10 | JK | 4.320200e+10 | 8.039000e+09 |
| 1993 | Argentine peso | 2.166090e+11 | TOT | 2.117210e+11 | 8.955300e+10 |
| 1993 | boliviano | 2.255600e+10 | TOT | 2.255600e+10 | 8.821000e+09 |
| 1993 | real | 9.560000e+08 | AB | 9.560000e+08 | 1.960000e+08 |
| 1993 | real | 1.480000e+08 | C | 1.480000e+08 | 3.600000e+07 |
| 1993 | real | 3.672000e+09 | D | 3.672000e+09 | 1.006000e+09 |
| 1993 | real | 3.930000e+08 | E | 3.930000e+08 | 2.480000e+08 |
| 1993 | real | 1.044000e+09 | F | 1.044000e+09 | 2.050000e+08 |
| 1993 | real | 1.172000e+09 | GH | 1.172000e+09 | 4.960000e+08 |
| 1993 | real | 6.820000e+08 | I | 6.820000e+08 | 3.250000e+08 |
| 1993 | real | 5.559000e+09 | JK | 5.559000e+09 | 1.460000e+09 |

```
1993              real  1.655200e+10  TOT  1.655200e+10  6.363000e+09
1993      Chilean peso  1.660178e+13  TOT  1.660178e+13  6.582086e+12
1993    Colombian peso  7.823940e+12   AB  7.823937e+12  1.544580e+12
1993    Colombian peso  2.237211e+12    C  2.237211e+12  6.405010e+11
1993    Colombian peso  8.275730e+12    D  8.275730e+12  3.349123e+12
1993    Colombian peso  1.607909e+12    E  1.607909e+12  4.512400e+11
1993    Colombian peso  3.648933e+12    F  3.648933e+12  1.326030e+12
1993    Colombian peso  6.521734e+12   GH  6.521734e+12  2.652296e+12
1993    Colombian peso  3.750715e+12    I  3.750715e+12  1.697697e+12
1993    Colombian peso  8.494946e+12   JK  8.494946e+12  1.571843e+12
1993    Colombian peso  5.059828e+13  TOT  5.059828e+13  1.882861e+13
1993      Danish krone  2.584700e+10   AB  2.584700e+10  6.645000e+09
…                  …           …   …            …             …
1993            peseta  1.106390e+13   JK  1.106390e+13  3.062300e+12
1993            peseta  6.162560e+13  TOT  6.162560e+13  3.006070e+13
1993     Swedish krona  2.749800e+10   AB  2.749800e+10  8.543000e+09
1993     Swedish krona  3.467000e+09    C  3.467000e+09  2.414000e+09
1993     Swedish krona  2.601340e+11    D  2.601340e+11  1.779260e+11
1993     Swedish krona  4.278100e+10    E  4.278100e+10  8.195000e+09
1993     Swedish krona  7.949000e+10    F  7.949000e+10  6.192000e+10
1993     Swedish krona  1.396150e+11   GH  1.396150e+11  1.110160e+11
1993     Swedish krona  8.499200e+10    I  8.499200e+10  5.889300e+10
1993     Swedish krona  3.169450e+11   JK  3.169450e+11  9.304900e+10
1993     Swedish krona  1.328816e+12  TOT  1.328816e+12  8.516030e+11
1993              baht  2.739010e+11   AB  2.739010e+11  3.091700e+10
1993              baht  3.778900e+10    C  3.778900e+10  5.041000e+09
1993              baht  7.990670e+11    D  7.990670e+11  2.956290e+11
1993              baht  7.315900e+10    E  7.315900e+10  1.835000e+10
1993              baht  2.109040e+11    F  2.109040e+11  8.201700e+10
1993              baht  3.711790e+11   GH  3.711790e+11  5.910000e+10
1993              baht  2.366030e+11    I  2.366030e+11  5.369300e+10
1993              baht  2.778150e+11   JK  2.778150e+11  6.928200e+10
1993              baht  2.784561e+12  TOT  2.784561e+12  8.666570e+11
1993  New Turkish lira  1.976292e+09  TOT  1.976292e+09  6.119037e+08
1993           bolivar  2.916830e+11   AB  2.916830e+11  7.117200e+10
1993           bolivar  8.411670e+11    C  8.411670e+11  9.630300e+10
1993           bolivar  9.605900e+11    D  9.605900e+11  2.825770e+11
1993           bolivar  1.420870e+11    E  1.420870e+11  3.532400e+10
1993           bolivar  3.322990e+11    F  3.322990e+11  1.072110e+11
1993           bolivar  1.045332e+12   GH  1.045332e+12  4.531180e+11
1993           bolivar  4.044280e+11    I  4.044280e+11  1.226970e+11
1993           bolivar  6.991300e+11   JK  6.991300e+11  1.585720e+11
1993           bolivar  5.449065e+12  TOT  5.449065e+12  1.863825e+12

[120 rows x 7 columns]
```

[12]:
```python
dfycc = df.set_index(["year", "country", "code"])
dfycc.head()
```

[12]:
```
                                                  sub_item        currency  \
year country   code
1993 Argentina AB    Agriculture, hunting, forestry fishing  Argentine peso
               C                        Mining and quarrying  Argentine peso
```

```
                D                                Manufacturing  Argentine peso
                E           Electricity, gas and water supply  Argentine peso
                F                                Construction  Argentine peso

                                VA1           VA2         WAGE1
year country   code
1993 Argentina AB    1.214900e+10  1.214900e+10  2.123000e+09
               C     3.525000e+09  3.525000e+09  8.007000e+08
               D     3.890700e+10  3.890700e+10  1.766600e+10
               E     4.461000e+09  4.461000e+09  2.213000e+09
               F     1.339300e+10  1.339300e+10  4.355000e+09
```

Documentation : sortlevel

[13]:
```
dfycc.sort_index(inplace=True)
dfycc.head()
```

[13]:
```
                                                sub_item      currency  \
year country code
1966 Denmark AB    Agriculture, hunting, forestry fishing  Danish krone
             C                        Mining and quarrying  Danish krone
             D                                Manufacturing  Danish krone
             E           Electricity, gas and water supply  Danish krone
             F                                Construction  Danish krone

                            VA1           VA2         WAGE1
year country code
1966 Denmark AB    5.694000e+09  5.694000e+09  1.191000e+09
             C     2.530000e+08  2.530000e+08  8.000000e+07
             D     1.543800e+10  1.543800e+10  1.095700e+10
             E     1.320000e+09  1.320000e+09  3.430000e+08
             F     6.928000e+09  6.928000e+09  4.857000e+09
```

[14]:
```
dfycc.loc[1993, "Brazil", "TOT"]
```

[14]:
```
sub_item    Total Economy
currency             real
VA1            1.6552e+10
VA2            1.6552e+10
WAGE1           6.363e+09
Name: (1993, Brazil, TOT), dtype: object
```

[15]:
```
dfycc.sort_index(level=2, inplace=True)
dfycc.head()
```

[15]:
```
                                                sub_item      currency  \
year country code
1966 Denmark AB    Agriculture, hunting, forestry fishing  Danish krone
1967 Denmark AB    Agriculture, hunting, forestry fishing  Danish krone
1968 Denmark AB    Agriculture, hunting, forestry fishing  Danish krone
1969 Denmark AB    Agriculture, hunting, forestry fishing  Danish krone
1970 Bolivia AB    Agriculture, hunting, forestry fishing     boliviano

                            VA1           VA2         WAGE1
year country code
```

```
1966 Denmark AB    5.694000e+09  5.694000e+09  1.191000e+09
1967 Denmark AB    5.419000e+09  5.419000e+09  1.213000e+09
1968 Denmark AB    5.686000e+09  5.686000e+09  1.221000e+09
1969 Denmark AB    6.707000e+09  6.707000e+09  1.245000e+09
1970 Bolivia AB    2.240000e+03  2.000000e+03  4.030000e+02
```

Documentation : reset_index

[16]:
```
df.reset_index(drop=False, inplace=True)
        # le mot-clé drop pour garder ou non les colonnes servant d'index
        # inplace signifie qu'on modifie l'instance et non qu'une copie est modifiée
        # donc on peut aussi écrire dfi2 = dfi.reset_index(drop=False)
```

[17]:
```
df.columns
```

[17]:
```
Index(['index', 'country', 'sub_item', 'year', 'currency', 'VA1', 'code',
       'VA2', 'WAGE1'],
      dtype='object')
```

[18]:
```
df.index
```

[18]:
```
RangeIndex(start=0, stop=4239, step=1)
```

[19]:
```
df.loc[1993]
```

[19]:
```
index                      1993
country                   Mexico
sub_item    Mining and quarrying
year                       2002
currency        Mexican new peso
VA1                  7.72065e+10
code                           C
VA2                  7.72065e+10
WAGE1                1.84915e+10
Name: 1993, dtype: object
```

## 2 Manipulation avancée

[20]:
```
df.dtypes
```

[20]:
```
index         int64
country      object
sub_item     object
year          int64
currency     object
VA1         float64
code         object
VA2         float64
WAGE1       float64
dtype: object
```

### 2.0.1 filter

filter : on sélectionne un sous-ensemble de lignes qui vérifie une condition
Filter consiste à sélectionner un sous-ensemble de lignes du dataframe. Pour filter sur plusieurs conditions, il faut utiliser les opérateurs logique & (et), | (ou), ~ (non)

- filter
- mask
- where
- pandas: filter rows of DataFrame with operator chaining
- Indexing and Selecting Data

[21]:
```
subset = df [ (df.year == 1993) & (df.code == "AB") ]
subset.head()
df.filter(items=["country", "year", "VA1"])
```

[21]:

| | country | year | VA1 |
|---|---|---|---|
| 0 | Argentina | 1993 | 1.214900e+10 |
| 1 | Argentina | 1993 | 3.525000e+09 |
| 2 | Argentina | 1993 | 3.890700e+10 |
| 3 | Argentina | 1993 | 4.461000e+09 |
| 4 | Argentina | 1993 | 1.339300e+10 |
| 5 | Argentina | 1993 | 3.929400e+10 |
| 6 | Argentina | 1993 | 1.613400e+10 |
| 7 | Argentina | 1993 | 4.320200e+10 |
| 8 | Argentina | 1993 | 2.166090e+11 |
| 9 | Argentina | 1994 | 1.308500e+10 |
| 10 | Argentina | 1994 | 3.818000e+09 |
| 11 | Argentina | 1994 | 4.159600e+10 |
| 12 | Argentina | 1994 | 4.730000e+09 |
| 13 | Argentina | 1994 | 1.431100e+10 |
| 14 | Argentina | 1994 | 4.279800e+10 |
| 15 | Argentina | 1994 | 1.825100e+10 |
| 16 | Argentina | 1994 | 4.859900e+10 |
| 17 | Argentina | 1994 | 2.358460e+11 |
| 18 | Argentina | 1995 | 1.380850e+10 |
| 19 | Argentina | 1995 | 4.838400e+09 |
| 20 | Argentina | 1995 | 4.450210e+10 |
| 21 | Argentina | 1995 | 5.111000e+09 |
| 22 | Argentina | 1995 | 1.341400e+10 |
| 23 | Argentina | 1995 | 4.119850e+10 |
| 24 | Argentina | 1995 | 1.905990e+10 |
| 25 | Argentina | 1995 | 5.133940e+10 |
| 26 | Argentina | 1995 | 2.423343e+11 |
| 27 | Argentina | 1996 | 1.527000e+10 |
| 28 | Argentina | 1996 | 5.888900e+09 |
| 29 | Argentina | 1996 | 4.772340e+10 |
| … | … | … | … |
| 4209 | Venezuela | 1998 | 5.057137e+13 |
| 4210 | Venezuela | 1999 | 2.880633e+12 |
| 4211 | Venezuela | 1999 | 8.321318e+12 |
| 4212 | Venezuela | 1999 | 8.204542e+12 |
| 4213 | Venezuela | 1999 | 9.316080e+11 |
| 4214 | Venezuela | 1999 | 3.400098e+12 |
| 4215 | Venezuela | 1999 | 9.794983e+12 |

```
4216  Venezuela  1999  5.716383e+12
4217  Venezuela  1999  9.124555e+12
4218  Venezuela  1999  6.033646e+13
4219  Venezuela  2000  3.258018e+12
4220  Venezuela  2000  1.591774e+13
4221  Venezuela  2000  1.062152e+13
4222  Venezuela  2000  1.137058e+12
4223  Venezuela  2000  3.842038e+12
4224  Venezuela  2000  1.168852e+13
4225  Venezuela  2000  6.920770e+12
4226  Venezuela  2000  1.100281e+13
4227  Venezuela  2000  7.974029e+13
4228  Venezuela  2001  3.872766e+12
4229  Venezuela  2001  1.321210e+13
4230  Venezuela  2001  1.095207e+13
4231  Venezuela  2001  1.355214e+12
4232  Venezuela  2001  5.032111e+12
4233  Venezuela  2001  1.347735e+13
4234  Venezuela  2001  8.063945e+12
4235  Venezuela  2001  1.334013e+13
4236  Venezuela  2001  8.804319e+13
4237  Venezuela  2002  4.549229e+12
4238  Venezuela  2002  1.074291e+14

[4239 rows x 3 columns]
```

### 2.0.2 union : concaténation de deux Data Frames

union = concaténation de deux DataFrame (qui n'ont pas nécessaire les mêmes colonnes). On peut concaténer les lignes ou les colonnes

- concat
- Merge, join, and concatenate

[22]:
```
concat_ligne = pandas.concat((df,df))
concat_ligne[ (concat_ligne.year == 1993) & (concat_ligne.code == "AB") &␣
↪(concat_ligne.country == "Argentina")]
```

[22]:
```
   index    country                                sub_item  year  \
0      0  Argentina  Agriculture, hunting, forestry fishing  1993
0      0  Argentina  Agriculture, hunting, forestry fishing  1993

         currency           VA1 code           VA2         WAGE1
0  Argentine peso  1.214900e+10   AB  1.214900e+10  2.123000e+09
0  Argentine peso  1.214900e+10   AB  1.214900e+10  2.123000e+09
```

### 2.0.3 sort : tri des lignes

sort

[23]:
```
tri = df.sort_values( by=["year", "country"], ascending=[1,0])
tri.tail(10)
```

[23]:

```
     index  country                                          sub_item  year  \
559    559    Chile                                Mining and quarrying  2009
560    560    Chile                                       Manufacturing  2009
561    561    Chile                   Electricity, gas and water supply  2009
562    562    Chile                                        Construction  2009
563    563    Chile  Wholesale retail trade, repair of motor vehicl…  2009
564    564    Chile               Transport, storage and communications  2009
565    565    Chile  Financial intermediation real estate, renting …  2009
566    566    Chile                                       Total Economy  2009
269    269  Bolivia                                       Total Economy  2009
270    270  Bolivia                                       Total Economy  2010

         currency           VA1 code           VA2         WAGE1
559  Chilean peso  1.404654e+13    C  1.404654e+13  1.588307e+12
560  Chilean peso  1.126610e+13    D  1.126610e+13  3.666442e+12
561  Chilean peso  3.633492e+12    E  3.633492e+12  3.030204e+11
562  Chilean peso  6.804767e+12    F  6.804767e+12  4.335883e+12
563  Chilean peso  8.163060e+12   GH  8.163060e+12  5.358225e+12
564  Chilean peso  6.600354e+12    I  6.600354e+12  2.894256e+12
565  Chilean peso  1.819692e+13   JK  1.819692e+13  5.991782e+12
566  Chilean peso  8.650220e+13  TOT  8.650220e+13  3.658516e+13
269     boliviano  1.021160e+11  TOT  1.021160e+11  3.381017e+10
270     boliviano  1.159344e+11  TOT  1.159344e+11  3.647705e+10
```

### 2.0.4 group by : grouper des lignes qui partagent une valeur commune

Cette opération consiste à grouper les lignes qui partagent une caractéristique commune (une ou ou plusieurs valeurs par exemple). Sur chaque groupe, on peut calculer une somme, une moyenne…

- groupby
- sum
- cumsum
- mean
- count
- SQL GROUP BY
- Group By: split-apply-combine
- group by customisé

[24]:

```
df[["country", "code", "year"]].cumsum(0).head()
```

[24]:

```
                                        country    code  year
0                                     Argentina      AB  1993
1                             ArgentinaArgentina     ABC  3986
2                     ArgentinaArgentinaArgentina    ABCD  5979
3            ArgentinaArgentinaArgentinaArgentina   ABCDE  7972
4  ArgentinaArgentinaArgentinaArgentinaArgentina  ABCDEF  9965
```

### 2.0.5 pivot : utiliser des valeurs présentes dans colonne comme noms de colonnes

pivot (tableau croisé dynamique)
Cette opération consiste à créer une seconde table en utilisant utiliser les valeurs d'une colonne comme nom de colonnes.

- pivot

- Reshaping and Pivot Tables
- Tableau croisé dynamique - wikipédia

[25]:
```
df.columns
```

[25]:
```
Index(['index', 'country', 'sub_item', 'year', 'currency', 'VA1', 'code',
       'VA2', 'WAGE1'],
      dtype='object')
```

[26]:
```
df.head()
```

[26]:
```
   index    country                                sub_item  year  \
0      0  Argentina  Agriculture, hunting, forestry fishing  1993
1      1  Argentina                     Mining and quarrying  1993
2      2  Argentina                            Manufacturing  1993
3      3  Argentina       Electricity, gas and water supply  1993
4      4  Argentina                             Construction  1993

         currency           VA1 code           VA2         WAGE1
0  Argentine peso  1.214900e+10   AB  1.214900e+10  2.123000e+09
1  Argentine peso  3.525000e+09    C  3.525000e+09  8.007000e+08
2  Argentine peso  3.890700e+10    D  3.890700e+10  1.766600e+10
3  Argentine peso  4.461000e+09    E  4.461000e+09  2.213000e+09
4  Argentine peso  1.339300e+10    F  1.339300e+10  4.355000e+09
```

[27]:
```
dfcopy = df.copy()
```

[28]:
```
dfcopy["index"] = df.apply(lambda x: "{0}-{1}".format(x["country"], x["year"]), axis=1)
```

[29]:
```
gr = dfcopy[["index", "code", "VA1"]].groupby(["index", "code"]).sum().reset_index()
gr.head()
```

[29]:
```
             index code           VA1
0  Argentina-1993   AB  1.214900e+10
1  Argentina-1993    C  3.525000e+09
2  Argentina-1993    D  3.890700e+10
3  Argentina-1993    E  4.461000e+09
4  Argentina-1993    F  1.339300e+10
```

[30]:
```
piv = gr.pivot(index="index", columns="code", values="VA1")
piv.head()
```

[30]:
```
code                      AB             C             D             E  \
index
Argentina-1993  1.214900e+10  3.525000e+09  3.890700e+10  4.461000e+09
Argentina-1994  1.308500e+10  3.818000e+09  4.159600e+10  4.730000e+09
Argentina-1995  1.380850e+10  4.838400e+09  4.450210e+10  5.111000e+09
Argentina-1996  1.527000e+10  5.888900e+09  4.772340e+10  5.232400e+09
Argentina-1997  1.529300e+10  5.632500e+09  5.338210e+10  5.501700e+09

code                       F            GH             I            JK  \
index
Argentina-1993  1.339300e+10  3.929400e+10  1.613400e+10  4.320200e+10
```

```
Argentina-1994  1.431100e+10  4.279800e+10  1.825100e+10  4.859900e+10
Argentina-1995  1.341400e+10  4.119850e+10  1.905990e+10  5.133940e+10
Argentina-1996  1.352680e+10  4.454100e+10  2.050140e+10  5.237490e+10
Argentina-1997  1.508030e+10  4.912050e+10  2.295190e+10  5.468300e+10

code                     TOT
index
Argentina-1993  2.166090e+11
Argentina-1994  2.358460e+11
Argentina-1995  2.423343e+11
Argentina-1996  2.546081e+11
Argentina-1997  2.730922e+11
```

[31]:
```
piv.tail()
```

[31]:
```
code                     AB             C             D             E  \
index
Venezuela-1998  2.461132e+12  5.412143e+12  7.463681e+12  8.678680e+11
Venezuela-1999  2.880633e+12  8.321318e+12  8.204542e+12  9.316080e+11
Venezuela-2000  3.258018e+12  1.591774e+13  1.062152e+13  1.137058e+12
Venezuela-2001  3.872766e+12  1.321210e+13  1.095207e+13  1.355214e+12
Venezuela-2002  4.549229e+12           NaN           NaN           NaN

code                      F            GH             I            JK  \
index
Venezuela-1998  3.443028e+12  8.775614e+12  4.893346e+12  7.898823e+12
Venezuela-1999  3.400098e+12  9.794983e+12  5.716383e+12  9.124555e+12
Venezuela-2000  3.842038e+12  1.168852e+13  6.920770e+12  1.100281e+13
Venezuela-2001  5.032111e+12  1.347735e+13  8.063945e+12  1.334013e+13
Venezuela-2002           NaN           NaN           NaN           NaN

code                     TOT
index
Venezuela-1998  5.057137e+13
Venezuela-1999  6.033646e+13
Venezuela-2000  7.974029e+13
Venezuela-2001  8.804319e+13
Venezuela-2002  1.074291e+14
```

### 2.0.6 join : fusionner deux Data Frames en associant les lignes qui partagent une valeur commune

Fusionner deux tables consiste à apparier les lignes de la première table avec celle de la seconde si certaines colonnes de ces lignes partagent les mêmes valeurs. On distingue quatre cas :
INNER JOIN - inner : on garde tous les appariements réussis
LEFT OUTER JOIN - left : on garde tous les appariements réussis et les lignes non appariées de la table de gauche
RIGHT OUTER JOIN - right : on garde tous les appariements réussis et les lignes non appariées de la table de droite
FULL OUTER JOIN - outer : on garde tous les appariements réussis et les lignes non appariées des deux tables
Exemples et documentation : * merging, joining * join * merge ou DataFrame.merge * jointures SQL - illustrations avec graphiques en patates

[32] :

```

```

## 3 Exercice: moyennes par groupes

Calculer par exemple pour chaque pays, la moyenne des salaires au cours des années.

[33] :

```

```