Pandas - Sélection, manipulation et analyse des données en Python (en français)

Pandas est une bibliothèque essentielle pour la manipulation et l’analyse des données en Python. Une des premières compétences à maîtriser est de savoir sélectionner et manipuler les données rapidement et efficacement.

In [26]:
```
import pandas as pd

reviews = pd.read_csv("./winedata.csv", index_col=0)
pd.set_option('display.max_rows', 5)
reviews.head()
```

Out[26]:

	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	v
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	N
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	(Avi
2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rair
3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree	NaN	St. Julian 2013 Reserve Late Harvest Riesling ...	Riesling	St.
4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	65.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	@paulgwine	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	Pinot Noir	C

1. Sélection des données

Accès natif

Pandas permet de sélectionner les colonnes d'un DataFrame de manière similaire à l'accès aux propriétés des objets en Python.

In [27]:
```
# Accéder à la colonne 'country'
reviews.country
```

Out[27]:
```
0            Italy
1         Portugal
             ...
129969      France
129970      France
Name: country, Length: 129971, dtype: object
```

In [28]:
```
# Accéder à la colonne 'country' avec l'opérateur []
reviews['country']
```

Out[28]:
```
0            Italy
1         Portugal
             ...
129969      France
129970      France
Name: country, Length: 129971, dtype: object
```

Sélection avec index

In [29]:
```
reviews.iloc[0]
```

Out[29]:
```
country                                                    Italy
description    Aromas include tropical fruit, broom, brimston...
                                        ...
variety                                             White Blend
winery                                                  Nicosia
Name: 0, Length: 13, dtype: object
```

In [30]:
```
reviews.iloc[1:3, 0]
```

Out[30]:
```
1    Portugal
2          US
Name: country, dtype: object
```

In [31]:
```
#avec une list
reviews.iloc[[0, 1, 2], 0]
```

Out[31]:
```
0       Italy
1    Portugal
2          US
Name: country, dtype: object
```

In [32]:
```
reviews.loc[0, 'country']
```

Out[32]:
```
'Italy'
```

2. Sélection conditionnelle

In [33]:
```
# Sélectionner les vins produits en Italie
reviews.loc[reviews.country == 'Italy']
```

Out[33]:

	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	
6	Italy	Here's a bright, informal red that opens with ...	Belsito	87	16.0	Sicily & Sardinia	Vittoria	NaN	Kerin O'Keefe	@kerinokeefe	Terre di Giurfo 2013 Belsito Frappato (Vittoria)	Frappato	
...	...	...	...	...	...	...	...	...	...	...	...	...	...
129961	Italy	Intense aromas of wild cherry, baking spice, t...	NaN	90	30.0	Sicily & Sardinia	Sicilia	NaN	Kerin O'Keefe	@kerinokeefe	COS 2013 Frappato (Sicilia)	Frappato	
129962	Italy	Blackberry, cassis, grilled herb and toasted a...	Sàgana Tenuta San Giacomo	90	40.0	Sicily & Sardinia	Sicilia	NaN	Kerin O'Keefe	@kerinokeefe	Cusumano 2012 Sàgana Tenuta San Giacomo Nero d...	Nero d'Avola	Cu

19540 rows × 13 columns

In [34]:
```
# Sélectionner les vins italiens avec un score supérieur ou égal à 90
reviews.loc[(reviews.country == 'Italy') & (reviews.points >= 90)]
```

Out[34]:

	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	
120	Italy	Slightly backward, particularly given the vint...	Bricco Rocche Prapó	92	70.0	Piedmont	Barolo	NaN	NaN	NaN	Ceretto 2003 Bricco Rocche Prapó (Barolo)	Nebbiolo	
130	Italy	At the first it was quite muted and subdued, b...	Bricco Rocche Brunate	91	70.0	Piedmont	Barolo	NaN	NaN	NaN	Ceretto 2003 Bricco Rocche Brunate (Barolo)	Nebbiolo	
...	...	...	...	...	...	...	...	...	...	...	...	...	
129961	Italy	Intense aromas of wild cherry, baking spice, t...	NaN	90	30.0	Sicily & Sardinia	Sicilia	NaN	Kerin O'Keefe	@kerinokeefe	COS 2013 Frappato (Sicilia)	Frappato	
129962	Italy	Blackberry, cassis, grilled herb and toasted a...	Sàgana Tenuta San Giacomo	90	40.0	Sicily & Sardinia	Sicilia	NaN	Kerin O'Keefe	@kerinokeefe	Cusumano 2012 Sàgana Tenuta San Giacomo Nero d...	Nero d'Avola	Cu

6648 rows × 13 columns

3. Fonctions de résumé

In [35]:
```
# Résumé statistique des points des vins
reviews.points.describe()
```

Out[35]:
```
count    129971.000000
mean         88.447138
                ...
75%          91.000000
max         100.000000
Name: points, Length: 8, dtype: float64
```

In [36]:
```
# Moyenne des points
reviews.points.mean()
```

Out[36]:
```
88.44713820775404
```

4. Mapping et transformation

In [37]:
```
# Recentrer les scores des vins autour de 0
review_points_mean = reviews.points.mean()
reviews.points.map(lambda p: p - review_points_mean)
```

Out[37]:
```
0        -1.447138
1        -1.447138
              ...
129969    1.552862
129970    1.552862
Name: points, Length: 129971, dtype: float64
```

5. Groupby et agrégations

In [38]:
```
# Compter le nombre de vins par points
reviews.groupby('points').points.count()
```

Out[38]:
```
points
80     397
81     692
      ...
99      33
100     19
Name: points, Length: 21, dtype: int64
```

In [39]:
```
# Le vin le moins cher par catégorie de points
reviews.groupby('points').price.min()
```

Out[39]:
```
points
80      5.0
81      5.0
       ...
99     44.0
100    80.0
Name: price, Length: 21, dtype: float64
```

6. Travailler avec les Multi-index

In [40]:
```
# Utilisation de Multi-index
countries_reviewed = reviews.groupby(['country', 'province']).description.agg([len])
countries_reviewed
```

Out[40]:

country	province	len
Argentina	**Mendoza Province**	3264
	Other	536
...	**...**	...
Uruguay	**San Jose**	3
	Uruguay	24

425 rows × 1 columns

In [41]:
```
# Réinitialiser l'index
countries_reviewed.reset_index()
```

Out[41]:

	country	province	len
0	Argentina	Mendoza Province	3264
1	Argentina	Other	536
...	...	...	...
423	Uruguay	San Jose	3
424	Uruguay	Uruguay	24

425 rows × 3 columns

7. Trier les données

In [42]:
```
# Trier par longueur (descendant)
countries_reviewed.sort_values(by='len', ascending=False)
```

Out[42]:

country	province	len
US	California	36247
	Washington	8639
...	...	...
Chile	Coelemu	1
Greece	Beotia	1

425 rows × 1 columns

In [43]:
```
# Trier par plusieurs colonnes
countries_reviewed.sort_values(by=['country', 'len'])
```

Out[43]:

country	province	len
Argentina	Other	536
	Mendoza Province	3264
...	...	...
Uruguay	Uruguay	24
	Canelones	43

425 rows × 1 columns

8. Gestion des types de données

In [44]:
```
# Connaître le type de chaque colonne
reviews.dtypes
```

Out[44]:
```
country        object
description    object
                ...
variety        object
winery         object
Length: 13, dtype: object
```

In [45]:
```
# Convertir les points en float
reviews.points.astype('float64')
```

Out[45]:
```
0         87.0
1         87.0
          ...
129969    90.0
129970    90.0
Name: points, Length: 129971, dtype: float64
```

9. Gestion des valeurs manquantes

In [46]:
```
# Sélectionner les entrées avec des valeurs manquantes
reviews[pd.isnull(reviews.country)]
```

Out[46]:

	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	w
913	NaN	Amber in color, this wine has aromas of peach ...	Asureti Valley	87	30.0	NaN	NaN	NaN	Mike DeSimone	@worldwineguys	Gotsa Family Wines 2014 Asureti Valley Chinuri	Chinuri	
3131	NaN	Soft, fruity and juicy, this is a pleasant, si...	Partager	83	NaN	NaN	NaN	NaN	Roger Voss	@vossroger	Barton & Guestier NV Partager Red	Red Blend	Ba Gu
...	...	...	...	...	...	...	...	...	...	...	...	...	
129590	NaN	A blend of 60% Syrah, 30% Cabernet Sauvignon a...	Shah	90	30.0	NaN	NaN	NaN	Mike DeSimone	@worldwineguys	Büyülübağ 2012 Shah Red	Red Blend	Büyü
129900	NaN	This wine offers a delightful bouquet of black...	NaN	91	32.0	NaN	NaN	NaN	Mike DeSimone	@worldwineguys	Psagot 2014 Merlot	Merlot	P

63 rows × 13 columns

In [47]:
```
# Remplacer les valeurs manquantes par "Unknown"
reviews.region_2.fillna("Unknown")
```

Out[47]:
```
0         Unknown
1         Unknown
           ...
129969    Unknown
129970    Unknown
Name: region_2, Length: 129971, dtype: object
```

10. Renommer et combiner des DataFrames

In [48]:
```
# Renommer la colonne 'points' en 'score'
reviews.rename(columns={'points': 'score'})
```

Out[48]:

	country	description	designation	score	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red
...	...	...	...	...	...	...	...	...	...	...	...	...
129969	France	A dry style of Pinot Gris, this is crisp with ...	NaN	90	32.0	Alsace	Alsace	NaN	Roger Voss	@vossroger	Domaine Marcel Deiss 2012 Pinot Gris (Alsace)	Pinot Gris
129970	France	Big, rich and off-dry, this is powered by inte...	Lieu-dit Harth Cuvée Caroline	90	21.0	Alsace	Alsace	NaN	Roger Voss	@vossroger	Domaine Schoffit 2012 Lieu-dit Harth Cuvée Car...	Gewürztraminer

129971 rows × 13 columns

In [49]:
```
# Concatenation de DataFrames
df1 = reviews.head()
df2 = reviews.tail()
pd.concat([df1, df2])
```

Out[49]:

	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red
...	...	...	...	...	...	...	...	...	...	...	...	...
129969	France	A dry style of Pinot Gris, this is crisp with ...	NaN	90	32.0	Alsace	Alsace	NaN	Roger Voss	@vossroger	Domaine Marcel Deiss 2012 Pinot Gris (Alsace)	Pinot Gris
129970	France	Big, rich and off-dry, this is powered by inte...	Lieu-dit Harth Cuvée Caroline	90	21.0	Alsace	Alsace	NaN	Roger Voss	@vossroger	Domaine Schoffit 2012 Lieu-dit Harth Cuvée Car...	Gewürztraminer

10 rows × 13 columns