Data Management: Introduction to Pandas

Michel Coppée

March

Overview

Introduction

pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

Code

```
import pandas as pd
```

Creating data - DataFrame

Code

```
data = pd.DataFrame({'Yes': [50, 21], 'No': [131, 2]})
print(data)
```

Result

	Yes	No
0	50	131
1	21	2

Code

```
data = pd.DataFrame({'Mike': ['Bald', 'Tall'], 'Anna': ['', 'Short']})
```

Result

	Mike	Anna
0	Bald	
1	Tall	Short

Row labels - index

Code

```
data = pd.DataFrame(
    {'Mike': ['Bald', 'Tall'], 'Anna': ['', 'Short']},
    index=['Hair', 'Size']
    )
print(data)
```

Result

	Mike	Anna
Hair	Bald	
Size	Tall	Short

Series

Code

```
data = pd.Series([30, 35, 40])
print(data)
```

Result

```
0    30
1    35
2    40
dtype: int64
```

Series

Code

```
data = pd.Series(
    [30, 35, 40],
    index=['2018 Sales', '2019 Sales', '2020 Sales'],
    name='Product A'
    )
print(data)
```

Result

```
2018 Sales      30
2019 Sales      35
2020 Sales      40
Name: Product A, dtype: int64
```

Reading data files

data.csv

```
Product A,Product B,Product C,
30,21,9,
35,34,1,
41,11,11
```

Code

```
wine_reviews = pd.read_csv("your_directory/winemag-data-130k-v2.csv")
print(wine_reviews.shape)
```

Result

```
(129971, 14)
```

Reading data files

Code

```
print(wine_reviews.head())
```

Result

	Unnamed: 0	country	description	designation	points	price	province	region_1
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN
2	2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley
			Pineapple					

Reading data files

Code

```
wine_reviews = pd.read_csv(
    "your_directory/winemag-data-130k-v2.csv",
    index_col=0
    )
print(wine_reviews.head())
```

Result

	country	description	designation	points	price	province	region_1	region_2
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN
2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley

Native accessors

Code

```
print(wine_reviews.country)
print(wine_reviews['country'])
```

Result

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

Code

```
print(wine_reviews['country'][0])
```

Result

'Italy'

Indexing in pandas - Index-based selection

Code

```
print(wine_reviews.iloc[0])
```

Result

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

Indexing in pandas - Index-based selection

Code

```
print(wine_reviews.iloc[:, 0])
print(wine_reviews.iloc[:3, 0])
print(wine_reviews.iloc[[0, 1, 2], 0])
print(wine_reviewsreviews.iloc[-5:])
```

Result

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

Result

```
0       Italy
1    Portugal
2          US
Name: country, dtype: object
```

Indexing in pandas - Label-based selection

Code

```
print(wine_reviews.loc[0, 'country'])
```

Result

	taster_name	taster_twitter_handle	points
0	Kerin O'Keefe	@kerinokeefe	87
1	Roger Voss	@vossroger	87
...	...	...	...
129969	Roger Voss	@vossroger	90
129970	Roger Voss	@vossroger	90

Remark

"iloc' includes the first element of the range and excludes the last one. So 0:10 will select entries 0,...,9. "loc", meanwhile, indexes inclusively. So 0:10 will select entries 0,...,10.

Conditional selection

Code

```
print(wine_reviews.country == 'Italy')
```

Result

```
0          True
1         False
          ...
129969    False
129970    False
Name: country, Length: 129971, dtype: bool
```

Code

```
print(wine_reviews.loc[wine_reviews.country == 'Italy'])
```

Result

	country	description	designation	points	price	province	region_1	region_2	taster_name
0	Italy	Aromas include tropical fruit, broom,	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe

Conditional selection

Code

```
print(wine_reviews.loc[
    (wine_reviews.country == 'Italy') & (wine_reviews.points >= 90)
    ])
```

Result

	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle
120	Italy	Slightly backward, particularly given the vint...	Bricco Rocche Prapó	92	70.0	Piedmont	Barolo	NaN	NaN	NaN
130	Italy	At the first it was quite muted and subdued, b...	Bricco Rocche Brunate	91	70.0	Piedmont	Barolo	NaN	NaN	NaN

Conditional selection

Code

```
print(wine_reviews.loc[
    (wine_reviews.country == 'Italy') | (wine_reviews.points >= 90)
    ])
```

Result

6	Italy	Here's a bright, informal red that opens with ...	Belsito	87	16.0	Sicily & Sardinia	Vittoria	NaN	Kerin O'Keefe	@kerinokeefe
...	...	...	...	...	...	...	...	...	...	...
129969	France	A dry style of Pinot Gris, this is crisp with ...	NaN	90	32.0	Alsace	Alsace	NaN	Roger Voss	@vossroger

Conditional selection

Code

```
print(wine_reviews.loc[
    wine_reviews.country.isin(['Italy', 'France'])
    ]) # Selects wines from Italy and France

print(wine_reviews.loc[
    wine_reviews.price.notnull()
    ]) # Selects wines for which price is not missing
```

Remark

The opposite of "notnull" is "isnull" and can be used in the same way.

Assigning data

Code

```
wine_reviews['critic'] = 'everyone'
print(wine_reviews['critic'])
```

Result

```
0         everyone
1         everyone
             ...
129969    everyone
129970    everyone
Name: critic, Length: 129971, dtype: object
```

Assigning data

Code

```
wine_reviews['index_backwards'] = range(len(reviews), 0, -1)
print(wine_reviews['index_backwards'])
```

Result

```
0         129971
1         129970
           ...
129969         2
129970         1
Name: index_backwards, Length: 129971, dtype: int64
```

Summary functions

Code

```
print(wine_reviews.points.describe())
print(wine_reviews.taster_name.describe())
```

Result

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

Result

```
count          103727
unique             19
top       Roger Voss
freq            25514
Name: taster_name, dtype: object
```

Summary functions

Code

```
print(wine_reviews.points.mean())
print(wine_reviews.taster_name.unique())
```

Result

88.44713820775404

Result

```
array(['Kerin O'Keefe', 'Roger Voss', 'Paul Gregutt',
       'Alexander Peartree', 'Michael Schachner', 'Anna Lee C. Iijima',
       'Virginie Boone', 'Matt Kettmann', nan, 'Sean P. Sullivan',
       'Jim Gordon', 'Joe Czerwinski', 'Anne Krebiehl\xa0MW',
       'Lauren Buzzeo', 'Mike DeSimone', 'Jeff Jenssen',
       'Susan Kostrzewa', 'Carrie Dykes', 'Fiona Adams',
       'Christina Pickard'], dtype=object)
```

Summary functions

Code

```
print(wine_reviews.taster_name.value_counts())
```

Result

```
Roger Voss              25514
Michael Schachner       15134
                          ...
Fiona Adams                27
Christina Pickard           6
Name: taster_name, Length: 19, dtype: int64
```

Maps

Code

```
review_points_mean = wine_reviews.points.mean()
points_remean = wine_reviews.points.map(lambda p: p - review_points_mean)
print(points_remean)
```

Result

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

Maps

Code

```
review_points_mean = wine_reviews.points.mean()

def remean_points(row):
    row.points = row.points - review_points_mean
    return row

wine_reviews.apply(remean_points, axis='columns')
```

Result

	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_har
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	-1.447138	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	-1.447138	15.0	Douro	NaN	NaN	Roger Voss	@vossroger

Maps - Remarks

If we had called wine_reviews.apply() with axis='index', then instead of passing a function to transform each row, we would need to give a function to transform each column.

Note that map() and apply() return new, transformed Series and DataFrames, respectively. They don't modify the original data they're called on.

Code

```
review_points_mean = wine_reviews.points.mean()
points_remean = wine_reviews.points - review_points_mean
print(points_remean)

print(wine_reviews.country + " - " + wine_reviews.region_1)
```

Groupwise analysis

Code

```
wine_reviews.groupby('points').points.count() # Same as value_counts()
```

Result

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

Groupwise analysis

Code

```
wine_reviews.groupby('points').price.min()
```

Result

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

Groupwise analysis

Code

```
wine_reviews.groupby('winery').apply(lambda df: df.title.iloc[0])
```

Result

```
winery
1+1=3                          1+1=3 NV Rosé Sparkling (Cava)
10 Knots                  10 Knots 2010 Viognier (Paso Robles)
                                        ...
àMaurice    àMaurice 2013 Fred Estate Syrah (Walla Walla V...
Štoka                           Štoka 2009 Izbrani Teran (Kras)
Length: 16757, dtype: object
```

Code - Meaning ?

```
wine_reviews.groupby(['country', 'province']).apply(
    lambda df: df.loc[df.points.idxmax()]
    )
```

Groupwise analysis

Code

```
wine_reviews.groupby(['country']).price.agg([len, min, max])
```

Result

	len	min	max
country			
Argentina	3800	4.0	230.0
Armenia	2	14.0	15.0
...	...	...	...
Ukraine	14	6.0	13.0
Uruguay	109	10.0	130.0

Multi-indexes

Code

```
countries = wine_reviews.groupby(
    ['country', 'province']).description.agg([len])
print(countries)
```

Result

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

Multi-indexes

Code

```
print(countries.reset_index())
```

Result

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

Sorting

Code

```
countries.sort_values(by='len')
countries.sort_values(by='len', ascending=False)
countries.sort_index()
countries.sort_values(by=['country', 'len'])
```

Data types

Code

```
print(wine_reviews.price.dtype)
```

Result

dtype('float64')

Code

```
print(wine_reviews.dtypes)
```

Result

```
country        object
description    object
                 ...
variety        object
winery         object
Length: 13, dtype: object
```

Data types

Code

```
wine_reviews.points.astype('float64')
```

Result

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

Missing Data

Remark

Entries missing values are given the value NaN, short for "Not a Number". For technical reasons these NaN values are always of the float64 dtype.

Code

```
wine_reviews[pd.isnull(reviews.country)]
```

Result

	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle
913	NaN	Amber in color, this wine has aromas of peach ...	Asureti Valley	87	30.0	NaN	NaN	NaN	Mike DeSimone	@worldwineguys
3131	NaN	Soft, fruity and juicy, this is a pleasant, si...	Partager	83	NaN	NaN	NaN	NaN	Roger Voss	@vossroger

Missing Data

Code

```
wine_reviews.region_2.fillna("Unknown")
```

Result

```
0          Unknown
1          Unknown
             ...
129969     Unknown
129970     Unknown
Name: region_2, Length: 129971, dtype: object
```

Missing Data

Code

```
wine_reviews.taster_twitter_handle.replace("@kerinokeefe", "@kerino")
```

Result

```
0              @kerino
1            @vossroger
                ...
129969       @vossroger
129970       @vossroger
Name: taster_twitter_handle, Length: 129971, dtype: object
```

Renaming

Code

```
wine_reviews.rename(columns={'points': 'score'})
```

Result

	country	description	designation	score	price	province	region_1	region_2	taster_name	taster_twitter_handle
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger

Renaming

Code

```
wine_reviews.rename_axis("wines", axis='rows').rename_axis(
    "fields", axis='columns')
```

Result

fields	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle
wines										
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	@kerinokeefe
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	@vossroger

Combining

Code

```
# Do not run these lines
french_wines = pd.read_csv("your_directory.csv")
british_wines = pd.read_csv("your_directory.csv")

pd.concat([french_wines, british_wines])
```

Code

```
# Do not run these lines
left = french_wines.set_index(['shared_index1', 'shared_index2'])
right = british_wines.set_index(['shared_index1', 'shared_index2'])

left.join(right, lsuffix='_FR', rsuffix='_UK')
```