

Pandas 資料分析 (2)

張家瑋 副教授

國立臺中科技大學資訊工程系

選取多個 DataFrame 的欄位

```
In [10]: import pandas as pd
import numpy as np

movies = pd.read_csv('data/movie.csv')
movie_actor_director = movies[['actor_1_name', 'actor_2_name',
                               'actor_3_name', 'director_name']]
movie_actor_director.head()
```

```
Out[10]:
```

	actor_1_name	actor_2_name	actor_3_name	director_name
0	CCH Pounder	Joel David Moore	Wes Studi	James Cameron
1	Johnny Depp	Orlando Bloom	Jack Davenport	Gore Verbinski
2	Christoph Waltz	Rory Kinnear	Stephanie Sigman	Sam Mendes
3	Tom Hardy	Christian Bale	Joseph Gordon-Levitt	Christopher Nolan
4	Doug Walker	Rob Walker	NaN	Doug Walker

```
In [11]: type(movies[['director_name']])
```

```
Out[11]: pandas.core.frame.DataFrame
```

```
In [12]: type(movies['director_name'])
```

```
Out[12]: pandas.core.series.Series
```

```
In [13]: type(movies.loc[:, ['director_name']])
```

```
Out[13]: pandas.core.frame.DataFrame
```

```
In [14]: type(movies.loc[:, 'director_name'])
```

```
Out[14]: pandas.core.series.Series
```

```
In [15]: cols = ['actor_1_name', 'actor_2_name', 'actor_3_name', 'director_name']
movie_actor_director = movies[cols]
```

movie_actor_director

	actor_1_name	actor_2_name	actor_3_name	director_name
0	CCH Pounder	Joel David Moore	Wes Studi	James Cameron
1	Johnny Depp	Orlando Bloom	Jack Davenport	Gore Verbinski
2	Christoph Waltz	Rory Kinnear	Stephanie Sigman	Sam Mendes
3	Tom Hardy	Christian Bale	Joseph Gordon-Levitt	Christopher Nolan
4	Doug Walker	Rob Walker	NaN	Doug Walker
...
4911	Eric Mabius	Daphne Zuniga	Crystal Lowe	Scott Smith
4912	Natalie Zea	Valorie Curry	Sam Underwood	NaN
4913	Eva Boehnke	Maxwell Moody	David Chandler	Benjamin Roberds
4914	Alan Ruck	Daniel Henney	Eliza Coupe	Daniel Hsia
4915	John August	Brian Herzlinger	Jon Gunn	Jon Gunn

4916 rows × 4 columns

```
movies['actor_1_name', 'actor_2_name', 'actor_3_name', 'director_name']
```

```
-----
KeyError                                Traceback (most recent call last)
c:\users\jwchang\appdata\local\programs\python\python37\lib\site-packages\pandas\core\indexes\base.py in get_loc(self, key, method, tolerance)
    3360         try:
-> 3361             return self._engine.get_loc(casted_key)
    3362         except KeyError as err:
```

用方法 (Methods) 選取欄位 (1/3)

```
: movies = pd.read_csv('data/movie.csv')  
movies.dtypes.value_counts()
```

```
: float64    13  
   object     12  
   int64       3  
   dtype: int64
```

```
: movies.select_dtypes(include='object').head()
```

	color	director_name	actor_2_name	genres	actor_1_name	movie_title	actor_3_name	plot_keywords	
0	Color	James Cameron	Joel David Moore	Action Adventure Fantasy Sci-Fi	CCH Pounder	Avatar	Wes Studi	avatar future marine native paraplegic	http://www.imdb.com/title/tt0447746/
1	Color	Gore Verbinski	Orlando Bloom	Action Adventure Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	Jack Davenport	goddess marriage ceremony marriage proposal pirates	http://www.imdb.com/title/tt0440476/
2	Color	Sam Mendes	Rory Kinnear	Action Adventure Thriller	Christoph Waltz	Spectre	Stephanie Sigman	bomb espionage sequel spy terrorist	http://www.imdb.com/title/tt2381249/
3	Color	Christopher Nolan	Christian Bale	Action Thriller	Tom Hardy	The Dark Knight Rises	Joseph Gordon-Levitt	deception imprisonment lawlessness police officer	http://www.imdb.com/title/tt1345836/
4	NaN	Doug Walker	Rob Walker	Documentary	Doug Walker	Star Wars: Episode VII - The Force Awakens	NaN	NaN	http://www.imdb.com/title/tt2529004/

用方法 (Methods) 選取欄位 (2/3)

```
def shorten(col):  
    return (col.replace('facebook_likes', 'fb')  
            .replace('_for_reviews', ''))  
movies = movies.rename(columns=shorten)  
movies.select_dtypes(include='number').head()
```

	num_critic	duration	director_fb	actor_3_fb	actor_1_fb	gross	num_voted_users	cast_total_fb	facenumber_in_poster	num_user	budget	title_year
0	723.0	178.0	0.0	855.0	1000.0	760505847.0	886204	4834	0.0	3054.0	2370000000.0	2009
1	302.0	169.0	563.0	1000.0	40000.0	309404152.0	471220	48350	0.0	1238.0	300000000.0	2007
2	602.0	148.0	0.0	161.0	11000.0	200074175.0	275868	11700	1.0	994.0	245000000.0	2011
3	813.0	164.0	22000.0	23000.0	27000.0	448130642.0	1144337	106759	0.0	2701.0	250000000.0	2011
4	NaN	NaN	131.0	NaN	131.0	NaN	8	143	0.0	NaN	NaN	NaN

```
movies.select_dtypes(include=['int', 'object']).head()
```

	color	director_name	actor_2_name	genres	actor_1_name	movie_title	actor_3_name	plot_keywords	url
0	Color	James Cameron	Joel David Moore	Action Adventure Fantasy Sci-Fi	CCH Pounder	Avatar	Wes Studi	avatar future marine native paraplegic	http://www.imdb.com/title/tt0447746/
1	Color	Gore Verbinski	Orlando Bloom	Action Adventure Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	Jack Davenport	goddess marriage ceremony marriage proposal pirates	http://www.imdb.com/title/tt0440476/
2	Color	Sam Mendes	Rory Kinnear	Action Adventure Thriller	Christoph Waltz	Spectre	Stephanie Sigman	bomb espionage sequel spy terrorist	http://www.imdb.com/title/tt2381249/
3	Color	Christopher Nolan	Christian Bale	Action Thriller	Tom Hardy	The Dark Knight Rises	Joseph Gordon-Levitt	deception imprisonment lawlessness police officer	http://www.imdb.com/title/tt1345836/
4	NaN	Doug Walker	Rob Walker	Documentary	Doug Walker	Star Wars: Episode VII - The Force Awakens	NaN	NaN	http://www.imdb.com/title/tt2529004/

用方法 (Methods) 選取欄位 (3/3)

```
movies.filter(like='fb').head()
```

	director_fb	actor_3_fb	actor_1_fb	cast_total_fb	actor_2_fb	movie_fb
0	0.0	855.0	1000.0	4834	936.0	33000
1	563.0	1000.0	40000.0	48350	5000.0	0
2	0.0	161.0	11000.0	11700	393.0	85000
3	22000.0	23000.0	27000.0	106759	23000.0	164000
4	131.0	NaN	131.0	143	12.0	0

```
cols = ['actor_1_name', 'actor_2_name', 'actor_3_name', 'director_name']  
movies.filter(items=cols).head()
```

	actor_1_name	actor_2_name	actor_3_name	director_name
0	CCH Pounder	Joel David Moore	Wes Studi	James Cameron
1	Johnny Depp	Orlando Bloom	Jack Davenport	Gore Verbinski
2	Christoph Waltz	Rory Kinnear	Stephanie Sigman	Sam Mendes
3	Tom Hardy	Christian Bale	Joseph Gordon-Levitt	Christopher Nolan
4	Doug Walker	Rob Walker	NaN	Doug Walker

```
movies.filter(regex=r'\d').head()
```

	actor_3_fb	actor_2_name	actor_1_fb	actor_1_name	actor_3_name	actor_2_fb
0	855.0	Joel David Moore	1000.0	CCH Pounder	Wes Studi	936.0
1	1000.0	Orlando Bloom	40000.0	Johnny Depp	Jack Davenport	5000.0
2	161.0	Rory Kinnear	11000.0	Christoph Waltz	Stephanie Sigman	393.0
3	23000.0	Christian Bale	27000.0	Tom Hardy	Joseph Gordon-Levitt	23000.0
4	NaN	Rob Walker	131.0	Doug Walker	NaN	12.0

用方法 (Methods) 選取欄位 (3/3)

```
movies.filter(like='fb').head()
```

	director_fb	actor_3_fb	actor_1_fb	cast_total_fb	actor_2_fb	movie_fb
0	0.0	855.0	1000.0	4834	936.0	33000
1	563.0	1000.0	40000.0	48350	5000.0	0
2	0.0	161.0	11000.0	11700	393.0	85000
3	22000.0	23000.0	27000.0	106759	23000.0	164000
4	131.0	NaN	131.0	143	12.0	0

```
cols = ['actor_1_name', 'actor_2_name', 'actor_3_name', 'director_name']  
movies.filter(items=cols).head()
```

	actor_1_name	actor_2_name	actor_3_name	director_name
0	CCH Pounder	Joel David Moore	Wes Studi	James Cameron
1	Johnny Depp	Orlando Bloom	Jack Davenport	Gore Verbinski
2	Christoph Waltz	Rory Kinnear	Stephanie Sigman	Sam Mendes
3	Tom Hardy	Christian Bale	Joseph Gordon-Levitt	Christopher Nolan
4	Doug Walker	Rob Walker	NaN	Doug Walker

```
movies.filter(regex=r'\d').head()
```

	actor_3_fb	actor_2_name	actor_1_fb	actor_1_name	actor_3_name	actor_2_fb
0	855.0	Joel David Moore	1000.0	CCH Pounder	Wes Studi	936.0
1	1000.0	Orlando Bloom	40000.0	Johnny Depp	Jack Davenport	5000.0
2	161.0	Rory Kinnear	11000.0	Christoph Waltz	Stephanie Sigman	393.0
3	23000.0	Christian Bale	27000.0	Tom Hardy	Joseph Gordon-Levitt	23000.0
4	NaN	Rob Walker	131.0	Doug Walker	NaN	12.0

在三種參數是互斥關係，一次只能用一種。

在select_dtypes使用exclude排除欄位

```
movies.select_dtypes(exclude='float').head()
```

	color	director_name	actor_2_name	genres	actor_1_name	movie_title	num_voted_users	cast_total_fb	actor_3_name	
0	Color	James Cameron	Joel David Moore	Action Adventure Fantasy Sci-Fi	CCH Pounder	Avatar	886204	4834	Wes Studi	avatar future r
1	Color	Gore Verbinski	Orlando Bloom	Action Adventure Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	471220	48350	Jack Davenport	goddess marri
2	Color	Sam Mendes	Rory Kinnear	Action Adventure Thriller	Christoph Waltz	Spectre	275868	11700	Stephanie Sigman	bomb espior
3	Color	Christopher Nolan	Christian Bale	Action Thriller	Tom Hardy	The Dark Knight Rises	1144337	106759	Joseph Gordon-Levitt	deception imprison
4	NaN	Doug Walker	Rob Walker	Documentary	Doug Walker	Star Wars: Episode VII - The Force Awakens	8	143	NaN	

在select_dtypes使用exclude排除欄位

```
movies.select_dtypes(exclude='float').head()
```

	color	director_name	actor_2_name	genres	actor_1_name	movie_title	num_voted_users	cast_total_fb	actor_3_name	
0	Color	James Cameron	Joel David Moore	Action Adventure Fantasy Sci-Fi	CCH Pounder	Avatar	886204	4834	Wes Studi	avatar future re
1	Color	Gore Verbinski	Orlando Bloom	Action Adventure Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	471220	48350	Jack Davenport	goddess marri
2	Color	Sam Mendes	Rory Kinnear	Action Adventure Thriller	Christoph Waltz	Spectre	275868	11700	Stephanie Sigman	bomb espior
3	Color	Christopher Nolan	Christian Bale	Action Thriller	Tom Hardy	The Dark Knight Rises	1144337	106759	Joseph Gordon-Levitt	deception imprison
4	NaN	Doug Walker	Rob Walker	Documentary	Doug Walker	Star Wars: Episode VII - The Force Awakens	8	143	NaN	

對欄位名稱進行排序 (1/2)

```
movies = pd.read_csv('data/movie.csv')
def shorten(col):
    return (col.replace('facebook_likes', 'fb')
            .replace('_for_reviews', ''))
movies = movies.rename(columns=shorten)
```

```
movies.columns
```

```
Index(['color', 'director_name', 'num_critic', 'duration', 'director_fb',
       'actor_3_fb', 'actor_2_name', 'actor_1_fb', 'gross', 'genres',
       'actor_1_name', 'movie_title', 'num_voted_users', 'cast_total_fb',
       'actor_3_name', 'facenumber_in_poster', 'plot_keywords',
       'movie_imdb_link', 'num_user', 'language', 'country', 'content_rating',
       'budget', 'title_year', 'actor_2_fb', 'imdb_score', 'aspect_ratio',
       'movie_fb'],
      dtype='object')
```

```
cat_core = ['movie_title', 'title_year', 'content_rating', 'genres']
cat_people = ['director_name', 'actor_1_name', 'actor_2_name', 'actor_3_name']
cat_other = ['color', 'country', 'language', 'plot_keywords', 'movie_imdb_link']
cont_fb = ['director_fb', 'actor_1_fb', 'actor_2_fb', 'actor_3_fb',
           'cast_total_fb', 'movie_fb']
cont_finance = ['budget', 'gross']
cont_num_reviews = ['num_voted_users', 'num_user', 'num_critic']
cont_other = ['imdb_score', 'duration', 'aspect_ratio', 'facenumber_in_poster']
```

對欄位名稱進行排序 (2/2)

```
new_col_order = cat_core + cat_people + \
                 cat_other + cont_fb + \
                 cont_finance + cont_num_reviews + \
                 cont_other
set(movies.columns) == set(new_col_order)
```

True

```
movies[new_col_order].head()
```

	movie_title	title_year	content_rating	genres	director_name	actor_1_name	actor_2_name	actor_3_name	color	country	...	movie_fb
0	Avatar	2009.0	PG-13	Action Adventure Fantasy Sci-Fi	James Cameron	CCH Pounder	Joel David Moore	Wes Studi	Color	USA	...	33000
1	Pirates of the Caribbean: At World's End	2007.0	PG-13	Action Adventure Fantasy	Gore Verbinski	Johnny Depp	Orlando Bloom	Jack Davenport	Color	USA	...	0
2	Spectre	2015.0	PG-13	Action Adventure Thriller	Sam Mendes	Christoph Waltz	Rory Kinnear	Stephanie Sigman	Color	UK	...	85000
3	The Dark Knight Rises	2012.0	PG-13	Action Thriller	Christopher Nolan	Tom Hardy	Christian Bale	Joseph Gordon-Levitt	Color	USA	...	164000
4	Star Wars: Episode VII - The Force Awakens	NaN	NaN	Documentary	Doug Walker	Doug Walker	Rob Walker	NaN	NaN	NaN	...	0

5 rows × 28 columns



DataFrame的統計方法 (1/3)

```
movies = pd.read_csv('data/movie.csv')  
movies.shape
```

(4916, 28)

```
movies.size
```

137648 列數和行數的乘積

```
movies.ndim
```

2 ndim=2 為 Dataframe
ndim=1 為 Series

```
len(movies)
```

4916 列數為資料筆數

```
movies.count()
```

color	4897
director_name	4814
num_critic_for_reviews	4867
duration	4901
director_facebook_likes	4814
actor_3_facebook_likes	4893
actor_2_name	4903
actor_1_facebook_likes	4909
gross	4054
genres	4916
actor_1_name	4909
movie_title	4916
num_voted_users	4916
cast_total_facebook_likes	4916
actor_3_name	4893
facenumber_in_poster	4903
plot_keywords	4764
movie_imdb_link	4916
num_user_for_reviews	4895
language	4904
country	4911
content_rating	4616
budget	4432
title_year	4810
actor_2_facebook_likes	4903
imdb_score	4916
aspect_ratio	4590
movie_facebook_likes	4916
dtype:	int64

```
movies.min()
```

num_critic_for_reviews	1
duration	7
director_facebook_likes	0
actor_3_facebook_likes	0
actor_1_facebook_likes	0
gross	162
genres	Action
movie_title	#Horror
num_voted_users	5
cast_total_facebook_likes	0
facenumber_in_poster	0
movie_imdb_link	http://www.imdb.com/title/tt0006864/?ref_=fn_t...
num_user_for_reviews	1
budget	218
title_year	1916
actor_2_facebook_likes	0
imdb_score	1.6
aspect_ratio	1.18
movie_facebook_likes	0
dtype:	object

其他統計方法還有 max(), mean(), median(), std()

DataFrame的統計方法 (2/3)

```
movies.describe().T
```

	count	mean	std	min	25%	50%	75%	max
num_critic_for_reviews	4867.0	1.379889e+02	1.202394e+02	1.00	49.00	108.00	191.00	8.130000e+02
duration	4901.0	1.070908e+02	2.528602e+01	7.00	93.00	103.00	118.00	5.110000e+02
director_facebook_likes	4814.0	6.910145e+02	2.832954e+03	0.00	7.00	48.00	189.75	2.300000e+04
actor_3_facebook_likes	4893.0	6.312763e+02	1.625875e+03	0.00	132.00	366.00	633.00	2.300000e+04
actor_1_facebook_likes	4909.0	6.494488e+03	1.510699e+04	0.00	607.00	982.00	11000.00	6.400000e+05
gross	4054.0	4.764451e+07	6.737255e+07	162.00	5019656.25	25043962.00	61108412.75	7.605058e+08
num_voted_users	4916.0	8.264492e+04	1.383222e+05	5.00	8361.75	33132.50	93772.75	1.689764e+06
cast_total_facebook_likes	4916.0	9.579816e+03	1.816432e+04	0.00	1394.75	3049.00	13616.75	6.567300e+05
facenumber_in_poster	4903.0	1.377320e+00	2.023826e+00	0.00	0.00	1.00	2.00	4.300000e+01
num_user_for_reviews	4895.0	2.676688e+02	3.729348e+02	1.00	64.00	153.00	320.50	5.060000e+03
budget	4432.0	3.654749e+07	1.002427e+08	218.00	6000000.00	19850000.00	43000000.00	4.200000e+09
title_year	4810.0	2.002448e+03	1.245398e+01	1916.00	1999.00	2005.00	2011.00	2.016000e+03
actor_2_facebook_likes	4903.0	1.621924e+03	4.011300e+03	0.00	277.00	593.00	912.00	1.370000e+05
imdb_score	4916.0	6.437429e+00	1.127802e+00	1.60	5.80	6.60	7.20	9.500000e+00
aspect_ratio	4590.0	2.222349e+00	1.402940e+00	1.18	1.85	2.35	2.35	1.600000e+01
movie_facebook_likes	4916.0	7.348294e+03	1.920602e+04	0.00	0.00	159.00	2000.00	3.490000e+05

DataFrame的統計方法 (3/3)

```
movies.describe(percentiles=[.99]).T
```

	count	mean	std	min	50%	99%	max
num_critic_for_reviews	4867.0	1.379889e+02	1.202394e+02	1.00	108.00	5.466800e+02	3.130000e+02
duration	4901.0	1.070908e+02	2.528602e+01	7.00	103.00	1.890000e+02	5.110000e+02
director_facebook_likes	4814.0	6.910145e+02	2.832954e+03	0.00	48.00	1.600000e+04	2.300000e+04
actor_3_facebook_likes	4893.0	6.312763e+02	1.625875e+03	0.00	366.00	1.100000e+04	2.300000e+04
actor_1_facebook_likes	4909.0	6.494488e+03	1.510699e+04	0.00	982.00	4.492000e+04	6.400000e+05
gross	4054.0	4.764451e+07	6.737255e+07	162.00	25043962.00	3.264128e+08	7.605058e+08
num_voted_users	4916.0	8.264492e+04	1.383222e+05	5.00	33132.50	6.815846e+05	1.689764e+06
cast_total_facebook_likes	4916.0	9.579816e+03	1.816432e+04	0.00	3049.00	6.241390e+04	6.567300e+05
facenumber_in_poster	4903.0	1.377320e+00	2.023826e+00	0.00	1.00	8.000000e+00	4.300000e+01
num_user_for_reviews	4895.0	2.676688e+02	3.729348e+02	1.00	153.00	1.999240e+03	5.060000e+03
budget	4432.0	3.654749e+07	1.002427e+08	218.00	19850000.00	2.000000e+08	4.200000e+09
title_year	4810.0	2.002448e+03	1.245398e+01	1916.00	2005.00	2.016000e+03	2.016000e+03
actor_2_facebook_likes	4903.0	1.621924e+03	4.011300e+03	0.00	593.00	1.700000e+04	1.370000e+05
imdb_score	4916.0	6.437429e+00	1.127802e+00	1.60	6.60	8.500000e+00	9.500000e+00
aspect_ratio	4590.0	2.222349e+00	1.402940e+00	1.18	2.35	4.000000e+00	1.600000e+01
movie_facebook_likes	4916.0	7.348294e+03	1.920602e+04	0.00	159.00	9.385000e+04	3.490000e+05

DataFrame的統計方法 – skipna = False

```
movies.min()
```

num_critic_for_reviews	1
duration	7
director_facebook_likes	0
actor_3_facebook_likes	0
actor_1_facebook_likes	0
gross	162
genres	Action
movie_title	#Horror
num_voted_users	5
cast_total_facebook_likes	0
facenumber_in_poster	0
movie_imdb_link	http://www.imdb.com/title/tt0006864/?ref=fn_t...
num_user_for_reviews	1
budget	218
title_year	1916
actor_2_facebook_likes	0
imdb_score	1.6
aspect_ratio	1.18
movie_facebook_likes	0

dtype: object

```
movies.min(skipna=False)
```

num_critic_for_reviews	NaN
duration	NaN
director_facebook_likes	NaN
actor_3_facebook_likes	NaN
actor_1_facebook_likes	NaN
gross	NaN
genres	Action
movie_title	#Horror
num_voted_users	5
cast_total_facebook_likes	0
facenumber_in_poster	NaN
movie_imdb_link	http://www.imdb.com/title/tt0006864/?ref=fn_t...
num_user_for_reviews	NaN
budget	NaN
title_year	NaN
actor_2_facebook_likes	NaN
imdb_score	1.6
aspect_ratio	NaN
movie_facebook_likes	0

dtype: object

串連 DataFrame 的方法

```
movies = pd.read_csv('data/movie.csv')
def shorten(col):
    return (col.replace('facebook_likes', 'fb')
            .replace('_for_reviews', ''))
movies = movies.rename(columns=shorten)
movies.isna().head()
```

	color	director_name	num_critic	duration	director_fb	actor_3_fb	actor_2_name	actor_1_fb	gross	genres	...	num_user	language	country	content_ratin
0	False	False	False	False	False	False	False	False	False	False	...	False	False	False	Fals
1	False	False	False	False	False	False	False	False	False	False	...	False	False	False	Fals
2	False	False	False	False	False	False	False	False	False	False	...	False	False	False	Fals
3	False	False	False	False	False	False	False	False	False	False	...	False	False	False	Fals
4	True	False	True	True	False	True	False	False	True	False	...	True	True	True	Tru

5 rows × 28 columns



串連 DataFrame 的方法

```
(movies.isna().sum().head())
```

```
color            19
director_name    102
num_critic       49
duration         15
director_fb      102
dtype: int64
```

```
movies.isna().sum().sum()
```

2654

```
movies.isna().any()
```

```
color            True
director_name    True
num_critic       True
duration         True
director_fb      True
actor_3_fb       True
actor_2_name     True
actor_1_fb       True
gross            True
genres           False
actor_1_name     True
movie_title      False
num_voted_users  False
cast_total_fb    False
actor_3_name     True
facenumber_in_poster True
plot_keywords    True
movie_imdb_link  False
num_user        True
language         True
country          True
content_rating   True
budget           True
title_year       True
actor_2_fb       True
imdb_score       False
aspect_ratio     True
movie_fb         False
dtype: bool
```

```
movies.isna().any().any()
```

True

```
movies[['color']].max()
```

```
Series([], dtype: float64)
```

```
movies.select_dtypes(['object']).fillna('')
```

	color	director_name	actor_2_name	genres	actor_1_name	movie_title	actor_3_name	plot_keywords	
0	Color	James Cameron	Joel David Moore	Action Adventure Fantasy Sci-Fi	CCH Pounder	Avatar	Wes Studi	avatar future marine native paraplegic	http://www
1	Color	Gore Verbinski	Orlando Bloom	Action Adventure Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	Jack Davenport	goddess marriage ceremony marriage proposal pi...	http://www
2	Color	Sam Mendes	Rory Kinnear	Action Adventure Thriller	Christoph Waltz	Spectre	Stephanie Sigman	bomb espionage sequel spy terrorist	http://www
3	Color	Christopher Nolan	Christian Bale	Action Thriller	Tom Hardy	The Dark Knight Rises	Joseph Gordon-Levitt	deception imprisonment lawlessness police offi...	http://www
4		Doug Walker	Rob Walker	Documentary	Doug Walker	Star Wars: Episode VII - The Force Awakens			http://www
...
4911	Color	Scott Smith	Daphne Zuniga	Comedy Drama	Eric Mabius	Signed Sealed Delivered	Crystal Lowe	fraud postal worker prison theft trial	http://www
4912	Color		Valorie Curry	Crime Drama Mystery Thriller	Natalie Zea	The Following	Sam Underwood	cult fbi hideout prison escape serial killer	http://www
4913	Color	Benjamin Roberds	Maxwell Moody	Drama Horror Thriller	Eva Boehnke	A Plague So Pleasant	David Chandler		http://www
4914	Color	Daniel Hsia	Daniel Henney	Comedy Drama Romance	Alan Ruck	Shanghai Calling	Eliza Coupe		http://www
4915	Color	Jon Gunn	Brian Herzlinger	Documentary	John August	My Date with Drew	Jon Gunn	actress name in title crush date four word tit...	http://www

串連 DataFrame 的方法

```
(movies.isna().sum().head())
```

```
color          19
director_name  102
num_critic      49
duration        15
director_fb     102
dtype: int64
```

```
movies.isna().sum().sum()
```

```
2654
```

```
movies.isna().any()
```

```
color          True
director_name   True
num_critic      True
duration        True
director_fb     True
actor_3_fb      True
actor_2_name    True
actor_1_fb      True
gross           True
genres          False
actor_1_name    True
movie_title     False
num_voted_users False
cast_total_fb   False
actor_3_name    True
facenumber_in_poster True
plot_keywords   True
movie_imdb_link False
num_user        True
language        True
country         True
content_rating  True
budget          True
title_year      True
actor_2_fb      True
imdb_score      False
aspect_ratio    True
movie_fb        False
dtype: bool
```

```
movies.isna().any().any()
```

```
True
```

串連 DataFrame 的方法

```
movies[['color']].max()
```

```
Series([], dtype: float64)
```

```
movies.select_dtypes(['object']).fillna('')
```

	color	director_name	actor_2_name	genres	actor_1_name	movie_title	actor_3_name	plot_keywords	
0	Color	James Cameron	Joel David Moore	Action Adventure Fantasy Sci-Fi	CCH Pounder	Avatar	Wes Studi	avatar future marine native paraplegic	http://www.imdb.com/title/tt0447746/
1	Color	Gore Verbinski	Orlando Bloom	Action Adventure Fantasy	Johnny Depp	Pirates of the Caribbean: At World's End	Jack Davenport	goddess marriage ceremony marriage proposal pirates	http://www.imdb.com/title/tt0440476/
2	Color	Sam Mendes	Rory Kinnear	Action Adventure Thriller	Christoph Waltz	Spectre	Stephanie Sigman	bomb espionage sequel spy terrorist	http://www.imdb.com/title/tt2381249/
3	Color	Christopher Nolan	Christian Bale	Action Thriller	Tom Hardy	The Dark Knight Rises	Joseph Gordon-Levitt	deception imprisonment lawlessness police officer	http://www.imdb.com/title/tt1345836/
4		Doug Walker	Rob Walker	Documentary	Doug Walker	Star Wars: Episode VII - The Force Awakens			http://www.imdb.com/title/tt2529004/
...
4911	Color	Scott Smith	Daphne Zuniga	Comedy Drama	Eric Mabius	Signed Sealed Delivered	Crystal Lowe	fraud postal worker prison theft trial	http://www.imdb.com/title/tt0447746/
4912	Color		Valorie Curry	Crime Drama Mystery Thriller	Natalie Zea	The Following	Sam Underwood	cult fbi hideout prison escape serial killer	http://www.imdb.com/title/tt1596369/
4913	Color	Benjamin Roberds	Maxwell Moody	Drama Horror Thriller	Eva Boehnke	A Plague So Pleasant	David Chandler		http://www.imdb.com/title/tt0447746/
4914	Color	Daniel Hsia	Daniel Henney	Comedy Drama Romance	Alan Ruck	Shanghai Calling	Eliza Coupe		http://www.imdb.com/title/tt0447746/
4915	Color	Jon Gunn	Brian Herzlinger	Documentary	John August	My Date with Drew	Jon Gunn	actress name in title crush date four word title	http://www.imdb.com/title/tt0447746/

```
(movies.select_dtypes(['object'])
.fillna('')
.max())
```

```
color
director_name
actor_2_name
genres
actor_1_name
movie_title
actor_3_name
plot_keywords
movie_imdb_link
language
country
content_rating
dtype: object
```

Color
Étienne Faure
Zubaida Sahar
Western
Óscar Jaenada
Æon Flux
Óscar Jaenada
zombie|zombie spoof
http://www.imdb.com/title/tt5574490/?ref_=fn_t...
Zulu
West Germany
X

DataFrame 的算符運算

```
: # colleges = pd.read_csv('data/college.csv')  
# colleges + 5
```

```
: colleges = pd.read_csv('data/college.csv', index_col='INSTNM')  
college_ugds = colleges.filter(like='UGDS_')  
college_ugds.head()
```

	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2MOR	UGDS_NRA	UGDS_UNKN
INSTNM									
Alabama A & M University	0.0333	0.9353	0.0055	0.0019	0.0024	0.0019	0.0000	0.0059	0.0138
University of Alabama at Birmingham	0.5922	0.2600	0.0283	0.0518	0.0022	0.0007	0.0368	0.0179	0.0100
Amridge University	0.2990	0.4192	0.0069	0.0034	0.0000	0.0000	0.0000	0.0000	0.2715
University of Alabama in Huntsville	0.6988	0.1255	0.0382	0.0376	0.0143	0.0002	0.0172	0.0332	0.0350
Alabama State University	0.0158	0.9208	0.0121	0.0019	0.0010	0.0006	0.0098	0.0243	0.0137

DataFrame 的算符運算

```
name = 'Northwest-Shoals Community College'
college_ugds.loc[name]
```

```
UGDS_WHITE    0.7912
UGDS_BLACK    0.1250
UGDS_HISP     0.0339
UGDS_ASIAN    0.0036
UGDS_AIAN     0.0088
UGDS_NHPI     0.0006
UGDS_2MOR     0.0012
UGDS_NRA      0.0033
UGDS_UNKN     0.0324
Name: Northwest-Shoals Community College, dtype: float64
```

```
college_ugds.loc[name].round(2)
```

```
UGDS_WHITE    0.79
UGDS_BLACK    0.12
UGDS_HISP     0.03
UGDS_ASIAN    0.00
UGDS_AIAN     0.01
UGDS_NHPI     0.00
UGDS_2MOR     0.00
UGDS_NRA      0.00
UGDS_UNKN     0.03
Name: Northwest-Shoals Community College, dtype: float64
```

```
(college_ugds.loc[name] + .0001).round(2)
```

```
UGDS_WHITE    0.79
UGDS_BLACK    0.13
UGDS_HISP     0.03
UGDS_ASIAN    0.00
UGDS_AIAN     0.01
UGDS_NHPI     0.00
UGDS_2MOR     0.00
UGDS_NRA      0.00
UGDS_UNKN     0.03
Name: Northwest-Shoals Community College, dtype: float64
```

```
college_ugds + .00501
```

	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2MOR	UGDS_NRA	UGDS_UNKN
INSTNM									
Alabama A & M University	0.03831	0.94031	0.01051	0.00691	0.00741	0.00691	0.00501	0.01091	0.01881
University of Alabama at Birmingham	0.59721	0.26501	0.03331	0.05681	0.00721	0.00571	0.04181	0.02291	0.01501
Amridge University	0.30401	0.42421	0.01191	0.00841	0.00501	0.00501	0.00501	0.00501	0.27651
University of Alabama in Huntsville	0.70381	0.13051	0.04321	0.04261	0.01931	0.00521	0.02221	0.03821	0.04001
Alabama State University	0.02081	0.92581	0.01711	0.00691	0.00601	0.00561	0.01481	0.02931	0.01871
...
SAE Institute of Technology San Francisco	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Rasmussen College - Overland Park	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
National Personal Training Institute of Cleveland	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Bay Area Medical Academy - San Jose Satellite Location	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Excel Learning Center-San Antonio South	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

7535 rows × 9 columns

DataFrame 的算符運算

```
.045+.005
```

```
0.049999999999999996
```

```
(college_ugds + .00501) // .01
```

	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2MOR	UGDS_NRA	UGDS_UNKN
INSTNM									
Alabama A & M University	3.0	94.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0
University of Alabama at Birmingham	59.0	26.0	3.0	5.0	0.0	0.0	4.0	2.0	1.0
Amridge University	30.0	42.0	1.0	0.0	0.0	0.0	0.0	0.0	27.0
University of Alabama in Huntsville	70.0	13.0	4.0	4.0	1.0	0.0	2.0	3.0	4.0
Alabama State University	2.0	92.0	1.0	0.0	0.0	0.0	1.0	2.0	1.0
...
SAE Institute of Technology San Francisco	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Rasmussen College - Overland Park	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
National Personal Training Institute of Cleveland	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Bay Area Medical Academy - San Jose Satellite Location	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Excel Learning Center-San Antonio South	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

7535 rows × 9 columns

DataFrame 的算符運算

```
college_ugds_op_round = (college_ugds + .00501) // .01 / 100  
college_ugds_op_round.head()
```

	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2MOR	UGDS_NRA	UGDS_UNKN
INSTNM									
Alabama A & M University	0.03	0.94	0.01	0.00	0.00	0.0	0.00	0.01	0.01
University of Alabama at Birmingham	0.59	0.26	0.03	0.05	0.00	0.0	0.04	0.02	0.01
Amridge University	0.30	0.42	0.01	0.00	0.00	0.0	0.00	0.00	0.27
University of Alabama in Huntsville	0.70	0.13	0.04	0.04	0.01	0.0	0.02	0.03	0.04
Alabama State University	0.02	0.92	0.01	0.00	0.00	0.0	0.01	0.02	0.01

```
college_ugds_round = (college_ugds + .00001).round(2)  
college_ugds_round
```

	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2MOR	UGDS_NRA	UGDS_UNKN
INSTNM									
Alabama A & M University	0.03	0.94	0.01	0.00	0.00	0.0	0.00	0.01	0.01
University of Alabama at Birmingham	0.59	0.26	0.03	0.05	0.00	0.0	0.04	0.02	0.01
Amridge University	0.30	0.42	0.01	0.00	0.00	0.0	0.00	0.00	0.27
University of Alabama in Huntsville	0.70	0.13	0.04	0.04	0.01	0.0	0.02	0.03	0.04
Alabama State University	0.02	0.92	0.01	0.00	0.00	0.0	0.01	0.02	0.01
...
SAE Institute of Technology San Francisco	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Rasmussen College - Overland Park	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
National Personal Training Institute of Cleveland	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Bay Area Medical Academy - San Jose Satellite Location	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

DataFrame 的算符運算

```
college_ugds_op_round.equals(college_ugds_round)
```

True

```
college2 = (college_ugds
            .add(.00501)
            .floordiv(.01)
            .div(100)
            )
college2.equals(college_ugds_op_round)
```

True

比較缺失值

```
np.nan == np.nan
```

False

```
None == None
```

True

```
np.nan > 5
```

False

```
5 > np.nan
```

False

```
np.nan != 5
```

True

```
college = pd.read_csv('data/college.csv', index_col='INSTNM')
college_ugds = college.filter(like='UGDS_')
```

```
college_ugds == .0019
```

	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2MOR	UGDS_NRA	UGDS_UNKN
INSTNM									
Alabama A & M University	False	False	False	True	False	True	False	False	False
University of Alabama at Birmingham	False	False	False	False	False	False	False	False	False
Amridge University	False	False	False	False	False	False	False	False	False
University of Alabama in Huntsville	False	False	False	False	False	False	False	False	False
Alabama State University	False	False	False	True	False	False	False	False	False
...
SAE Institute of Technology San Francisco	False	False	False	False	False	False	False	False	False
Rasmussen College - Overland Park	False	False	False	False	False	False	False	False	False
National Personal Training Institute of Cleveland	False	False	False	False	False	False	False	False	False
Bay Area Medical Academy - San Jose Satellite Location	False	False	False	False	False	False	False	False	False
Excel Learning Center-San Antonio South	False	False	False	False	False	False	False	False	False

7535 rows × 9 columns

比較缺失值

```
college_self_compare = college_ugds == college_ugds  
college_self_compare.head()
```

	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2MOR	UGDS_NRA	UGDS_UNKN
INSTNM									
Alabama A & M University	True	True	True	True	True	True	True	True	True
University of Alabama at Birmingham	True	True	True	True	True	True	True	True	True
Amridge University	True	True	True	True	True	True	True	True	True
University of Alabama in Huntsville	True	True	True	True	True	True	True	True	True
Alabama State University	True	True	True	True	True	True	True	True	True

```
college_self_compare.all()
```

```
UGDS_WHITE    False  
UGDS_BLACK    False  
UGDS_HISP     False  
UGDS_ASIAN    False  
UGDS_AIAN     False  
UGDS_NHPI     False  
UGDS_2MOR     False  
UGDS_NRA      False  
UGDS_UNKN     False  
dtype: bool
```

乍看沒有遺失值，但透過all()可以發現其實每個欄位都有遺失值。

比較缺失值

```
(college_ugds == np.nan).sum()
```

```
UGDS_WHITE    0
UGDS_BLACK    0
UGDS_HISP     0
UGDS_ASIAN    0
UGDS_AIAN     0
UGDS_NHPI     0
UGDS_2MOR     0
UGDS_NRA      0
UGDS_UNKN     0
dtype: int64
```

```
college_ugds.isna().sum()
```

```
UGDS_WHITE    661
UGDS_BLACK    661
UGDS_HISP     661
UGDS_ASIAN    661
UGDS_AIAN     661
UGDS_NHPI     661
UGDS_2MOR     661
UGDS_NRA      661
UGDS_UNKN     661
dtype: int64
```

```
college_ugds.equals(college_ugds) 相同位置的 nan 會當成同樣的元素
```

```
True
```

```
college_ugds.eq(.0019) 與 college_ugds == .0019 相同
```

	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2MOR	UGDS_NRA	UGDS_UNKN
INSTNM									
Alabama A & M University	False	False	False	True	False	True	False	False	False
University of Alabama at Birmingham	False	False	False	False	False	False	False	False	False
Amridge University	False	False	False	False	False	False	False	False	False
University of Alabama in Huntsville	False	False	False	False	False	False	False	False	False
Alabama State University	False	False	False	True	False	False	False	False	False
...
SAE Institute of Technology San Francisco	False	False	False	False	False	False	False	False	False
Rasmussen College - Overland Park	False	False	False	False	False	False	False	False	False
National Personal Training Institute of Cleveland	False	False	False	False	False	False	False	False	False
Bay Area Medical Academy - San Jose Satellite Location	False	False	False	False	False	False	False	False	False
Excel Learning Center-San Antonio South	False	False	False	False	False	False	False	False	False

7535 rows × 9 columns

```
from pandas.testing import assert_frame_equal
assert_frame_equal(college_ugds, college_ugds) is None
```

```
True
```

單元測試，可正確比較 nan

轉置 DataFrame 運算的方向

```
college = pd.read_csv('data/college.csv', index_col='INSTNM')
college_ugds = college.filter(like='UGDS_')
college_ugds.head()
```

	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2MOR	UGDS_NRA	UGDS_UNKN
INSTNM									
Alabama A & M University	0.0333	0.9353	0.0055	0.0019	0.0024	0.0019	0.0000	0.0059	0.0138
University of Alabama at Birmingham	0.5922	0.2600	0.0283	0.0518	0.0022	0.0007	0.0368	0.0179	0.0100
Amridge University	0.2990	0.4192	0.0069	0.0034	0.0000	0.0000	0.0000	0.0000	0.2715
University of Alabama in Huntsville	0.6988	0.1255	0.0382	0.0376	0.0143	0.0002	0.0172	0.0332	0.0350
Alabama State University	0.0158	0.9208	0.0121	0.0019	0.0010	0.0006	0.0098	0.0243	0.0137

```
college_ugds.count()
```

```
UGDS_WHITE    6874
UGDS_BLACK    6874
UGDS_HISP     6874
UGDS_ASIAN    6874
UGDS_AIAN     6874
UGDS_NHPI     6874
UGDS_2MOR     6874
UGDS_NRA      6874
UGDS_UNKN     6874
dtype: int64
```

轉置 DataFrame 運算的方向

```
college_ugds.count(axis='columns').head()
```

```
INSTNM
Alabama A & M University      9
University of Alabama at Birmingham  9
Amridge University           9
University of Alabama in Huntsville  9
Alabama State University      9
dtype: int64
```

```
college_ugds.sum(axis='columns').head()
```

```
INSTNM
Alabama A & M University      1.0000
University of Alabama at Birmingham  0.9999
Amridge University           1.0000
University of Alabama in Huntsville  1.0000
Alabama State University      1.0000
dtype: float64
```

```
college_ugds.median(axis='index')
```

```
UGDS_WHITE    0.55570
UGDS_BLACK    0.10005
UGDS_HISP     0.07140
UGDS_ASIAN    0.01290
UGDS_AIAN     0.00260
UGDS_NHPI     0.00000
UGDS_2MOR     0.01750
UGDS_NRA      0.00000
UGDS_UNKN     0.01430
dtype: float64
```

axis 參數預設為 0，即為 index。參數若為 1，則代表 columns。

```
college_ugds_cumsum = college_ugds.cumsum(axis=1)
college_ugds_cumsum.head()
```

	UGDS_WHITE	UGDS_BLACK	UGDS_HISP	UGDS_ASIAN	UGDS_AIAN	UGDS_NHPI	UGDS_2MOR	UGDS_NRA	UGDS_UNKN
INSTNM									
Alabama A & M University	0.0333	0.9686	0.9741	0.9760	0.9784	0.9803	0.9803	0.9862	1.0000
University of Alabama at Birmingham	0.5922	0.8522	0.8805	0.9323	0.9345	0.9352	0.9720	0.9899	0.9999
Amridge University	0.2990	0.7182	0.7251	0.7285	0.7285	0.7285	0.7285	0.7285	1.0000
University of Alabama in Huntsville	0.6988	0.8243	0.8625	0.9001	0.9144	0.9146	0.9318	0.9650	1.0000
Alabama State University	0.0158	0.9366	0.9487	0.9506	0.9516	0.9522	0.9620	0.9863	1.0000

案例演練：確定大學校園的多樣性

練習 & 回家作業