DataFrame

❑ Phát hiện null (NaN) trong DataFrame

```
df = pd.DataFrame({'weight':[65,87,67,65,74,69], 'height':[1.67,1.82,1.68,np.nan,1.68,np.nan]})
df
```

	weight	height
0	65	1.67
1	87	1.82
2	67	1.68
3	65	NaN
4	74	1.68
5	69	NaN

```
# kiểm tra null
df.isnull()
```

	weight	height
0	False	False
1	False	False
2	False	False
3	False	True
4	False	False
5	False	True

```
df.isnull().any() # tính theo cột (axis=0)
```

```
weight    False
height     True
dtype: bool
```

DataFrame

Phát hiện null (NaN) trong DataFrame

	weight	height
0	65	1.67
1	87	1.82
2	67	1.68
3	65	NaN
4	74	1.68
5	69	NaN

```
df.isnull().sum() # đếm True theo mỗi cột
```

```
weight    0
height    2
dtype: int64
```

```
df[df['height'].isnull()] # cho biết các vận động viên có height là null (NaN)
```

	weight	height
3	65	NaN
5	69	NaN

DataFrame

❑ Trùng lắp (duplicate) trong DataFrame

	name	age
0	Mike	21
1	John	32
2	Bill	27
3	John	32
4	Mike	21
5	Bill	26

```
df.duplicated() # bị duplicate trên các cột name và age
```

```
0    False
1    False
2    False
3     True
4     True
5    False
```

```
df.duplicated(subset=['name']) # bị duplicate trên cột name
```

```
0    False
1    False
2    False
3     True
4     True
5     True
```

DataFrame

Trùng lắp (duplicate) trong DataFrame

	name	age
0	Mike	21
1	John	32
2	Bill	27
3	John	32
4	Mike	21
5	Bill	26

```
# in ra các dòng bị duplicate trên cột name
df[df.duplicated(subset=['name'])]
```

	name	age
3	John	32
4	Mike	21
5	Bill	26

```
# in ra các dòng không bị duplicate trên cột name
df[~df.duplicated(subset=['name'])]
```

	name	age
0	Mike	21
1	John	32
2	Bill	27

DataFrame

Xóa dữ liệu trùng lắp trong DataFrame

	name	age
0	Mike	21
1	John	32
2	Bill	27
3	John	32
4	Mike	21
5	Bill	26

```
df.drop_duplicates(keep='first', inplace=True) # xóa các dòng bị duplicate
df
```

	name	age
0	Mike	21
1	John	32
2	Bill	27
5	Bill	26

DataFrame

❑ Xử lý tạo cột mới trong DataFrame

- Tạo cột điểm trung bình

	toan	ly	hoa
0	3.5	4.0	5.0
1	8.0	9.0	9.0
2	7.0	8.5	9.5
3	4.0	5.5	8.0
4	8.5	6.0	7.0

```
# tính điểm trung bình
df['trungbinh'] = round((df['toan']+df['ly']+df['hoa'])/3,2)
df
```

	toan	ly	hoa	trungbinh
0	3.5	4.0	5.0	4.17
1	8.0	9.0	9.0	8.67
2	7.0	8.5	9.5	8.33
3	4.0	5.5	8.0	5.83
4	8.5	6.0	7.0	7.17

DataFrame

Xử lý tạo cột mới trong DataFrame

- Tạo cột điểm trung bình

	toan	ly	hoa
0	3.5	4.0	5.0
1	8.0	9.0	9.0
2	7.0	8.5	9.5
3	4.0	5.5	8.0
4	8.5	6.0	7.0

```
df['trungbinh'] = round(df.mean(axis=1),2)
df
```

	toan	ly	hoa	trungbinh
0	3.5	4.0	5.0	4.17
1	8.0	9.0	9.0	8.67
2	7.0	8.5	9.5	8.33
3	4.0	5.5	8.0	5.83
4	8.5	6.0	7.0	7.17

DataFrame

❑ Xử lý tạo cột mới trong DataFrame

- Tạo cột điểm trung bình

	toan	ly	hoa
0	3.5	4.0	5.0
1	8.0	9.0	9.0
2	7.0	8.5	9.5
3	4.0	5.5	8.0
4	8.5	6.0	7.0

```
df['trungbinh'] = df.apply(np.mean, axis=1)
```

```
# axis=1, tính theo dòng
df['trungbinh'] = df.apply(lambda x: round((x[0]+x[1]+x[2])/3,2), axis=1)
```

DataFrame

❑ Xử lý tạo cột mới trong DataFrame

- Tạo cột kết quả

	toan	ly	hoa	trungbinh
0	3.5	4.0	5.0	4.1675
1	8.0	9.0	9.0	8.6675
2	7.0	8.5	9.5	8.3325
3	4.0	5.5	8.0	5.8325
4	8.5	6.0	7.0	7.1675

```
# tính kết quả đạt/ không đạt
df['ketqua'] = df['trungbinh'].map(lambda x: 'dat' if x>=5 else 'khong dat')
df
```

	toan	ly	hoa	trungbinh	ketqua
0	3.5	4.0	5.0	4.1675	khong dat
1	8.0	9.0	9.0	8.6675	dat
2	7.0	8.5	9.5	8.3325	dat
3	4.0	5.5	8.0	5.8325	dat
4	8.5	6.0	7.0	7.1675	dat

DataFrame

❑ Xử lý tạo cột mới trong DataFrame

- Tạo cột xếp hạng

	toan	ly	hoa	trungbinh
0	3.5	4.0	5.0	4.1675
1	8.0	9.0	9.0	8.6675
2	7.0	8.5	9.5	8.3325
3	4.0	5.5	8.0	5.8325
4	8.5	6.0	7.0	7.1675

```
# xếp hạng theo điểm trung bình
df['hang'] = df['trungbinh'].rank(ascending=False)

df['hang'] = df['hang'].astype(int)
df.sort_values(by='hang')
```

	toan	ly	hoa	trungbinh	ketqua	hang
1	8.0	9.0	9.0	8.6675	dat	1
2	7.0	8.5	9.5	8.3325	dat	2
4	8.5	6.0	7.0	7.1675	dat	3
3	4.0	5.5	8.0	5.8325	dat	4
0	3.5	4.0	5.0	4.1675	khong dat	5