



DataFrame

❑ Tạo DataFrame

- Tạo dataframe từ tập tin

Data Source Type	Read Function
CSV	read_csv
JSON	read_json
HTML	read_html
XML	read_xml
Excel	read_excel



DataFrame

- Các tham số thông dụng :
 - `filepath_or_buffer`: đường dẫn tập tin
 - `skiprows = k`: bỏ qua k dòng đầu tiên
 - `nrows = k`: chỉ lấy k dòng đầu tiên
 - `names = [col1, col2...]`: danh sách các cột sẽ được sử dụng cho DataFrame.
 - `header = k`: tên cột tương ứng với dòng thứ k sẽ làm header cho DataFrame.
 - `index_col = col`: đặt col làm index của DataFrame. col có thể là None nếu dataset không có cột index





DataFrame

❑ Tạo DataFrame

- Tạo dataframe từ tập tin .csv

```
df = pd.read_csv('data/insurance.csv')
```

```
df.head()
```

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520



DataFrame

❑ Tạo DataFrame

- Tạo dataframe từ tập tin .csv

```
df = pd.read_csv('data/drinks.csv', index_col=0) # index_col=n chọn cột n làm cột index (n>=0)
```

```
df.head()
```

	country	beer_servings	spirit_servings	wine_servings	total_litres_of_pure_alcohol	continent
0	Afghanistan	0	0	0	0.0	AS
1	Albania	89	132	54	4.9	EU
2	Algeria	25	0	14	0.7	AF
3	Andorra	245	138	312	12.4	EU
4	Angola	217	57	45	5.9	AF



DataFrame

❑ Tạo DataFrame

- Tạo dataframe từ tập tin .xlsx

```
# pip install xlrd  
movies = pd.read_excel('data/movies.xlsx', index_col=0)
```

```
movies.head()
```

	title	genres
movieId		
1	Toy Story (1995)	Adventure Animation Children Comedy Fantasy
2	Jumanji (1995)	Adventure Children Fantasy
3	Grumpier Old Men (1995)	Comedy Romance
4	Waiting to Exhale (1995)	Comedy Drama Romance
5	Father of the Bride Part II (1995)	Comedy



DataFrame

❑ Tạo DataFrame

- Tạo dataframe từ tập tin .xlsx

```
ratings = pd.read_excel('data/movies.xlsx', 'ratings') # sheet ratings
```

```
ratings.head()
```

	userId	movieId	rating	timestamp
0	1	31	2.5	1260759144
1	1	1029	3.0	1260759179
2	1	1061	3.0	1260759182
3	1	1129	2.0	1260759185
4	1	1172	4.0	1260759205