

<div>Today for tomorrow</div> <div>A D</div> <div>A I</div>	<b>PHÒNG THÍ NGHIỆM THỰC HÀNH ỨNG DỤNG</b> <b>KHOA HỌC DỮ LIỆU VÀ TRÍ TUỆ NHÂN TẠO</b> Applied Data Science and Artificial Intelligence Lab
 <div data-bbox="791 651 1268 792"> <b>LẬP TRÌNH PYTHON CHO KHOA HỌC DỮ LIỆU</b>  <b>PYTHON FOR DATA SCIENCE</b> </div> <div data-bbox="930 869 1112 898">Hà Nội ★ 11.2020</div>	
Nothing is faster than Innovation together	

1

<div>Today for tomorrow</div> <div>A D</div> <div>A I</div>	<b>Thông tin giảng viên</b>	
	<b>TS. TẠ QUANG CHIẾU</b>	<div data-bbox="675 1294 888 1332"><b>Vị trí &amp; Đào tạo</b></div> <ul style="list-style-type: none"> <li>- Trưởng bộ môn Hệ thống thông tin và Tri thức, Khoa Công nghệ Thông tin</li> <li>- PhD &amp; Post-doc in Polytech Tours, France</li> </ul> <div data-bbox="675 1451 1091 1489"><b>Lĩnh vực giảng dạy, nghiên cứu</b></div> <ul style="list-style-type: none"> <li>- Big data, Data science, machine learning, AI</li> <li>- Các phương pháp cho bài toán lập kế hoạch</li> </ul> <div data-bbox="675 1585 799 1619"><b>Contact:</b></div> <div data-bbox="675 1630 1007 1662">(E): <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a></div> <div data-bbox="675 1671 882 1702">(M): 0913 522 275</div>
Dr. Tạ Quang Chiếu - (E): <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> - (M): 0913 522 275		<b>2/38</b>

2

<div style="display: flex; justify-content: space-between; padding: 2px;"> <span>A</span> <span>D</span> </div> <div style="display: flex; justify-content: space-between; padding: 2px;"> <span>A</span> <span>I</span> </div>	<h2 style="margin: 0;">Objectives</h2>
<ul style="list-style-type: none"> <li>• Hiểu được tầm quan trọng của Khoa học dữ liệu</li> <li>• Vận dụng được các bước trong quy trình thực hiện một dự án về khoa học dữ liệu.</li> <li>• Kiến thức, kỹ năng cần thiết để trở thành một nhà khoa học dữ liệu</li> <li>• Áp dụng được Python và các thư viện phổ biến trong giải quyết một số bài toán cơ bản của Khoa học dữ liệu.</li> </ul>	
<div style="display: flex; justify-content: space-between;"> <span>Dr. Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275</span> <span style="background-color: #000080; color: white; padding: 2px 5px;">3/38</span> </div>	

3

<div style="display: flex; justify-content: space-between; padding: 2px;"> <span>A</span> <span>D</span> </div> <div style="display: flex; justify-content: space-between; padding: 2px;"> <span>A</span> <span>I</span> </div>	<h2 style="margin: 0;">References</h2>
<ul style="list-style-type: none"> <li>[1]. Introducing Data Science</li> <li>[2]. Python Data Science Handbook</li> <li>[3]. Data Science from Scratch</li> <li>[4]. Python crash course</li> <li>[5]. Bài giảng Khoa học dữ liệu</li> </ul>	
<div style="display: flex; justify-content: space-between;"> <span>Dr. Tạ Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275</span> <span style="background-color: #000080; color: white; padding: 2px 5px;">4/38</span> </div>	

4

A A	D I	<b>PHẦN 2: NHẬP MÔN KHOA HỌC DỮ LIỆU</b>	
		FUNDAMENTALS OF DATA SCIENCE	
<p><b>CONTENT</b></p> <p><b>1</b> CHAPTER 1: INTRODUCTION</p> <p><b>2</b> CHAPTER 2: PYTHON LIBRARIES FOR DATA SCIENCE</p>			
Dr. Tà Quang Chiếu - [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> - [M]: 0913 522 275			5/38


5

A A	D I	<b>Content of chapter 2</b>	
		<p>2.1 Python Library for Data Science</p> <p>2.2 Numpy library</p> <p><b>2.3 Pandas library</b></p> <p>2.4 Matplotlib library</p> <p>2.5 Scikit-learn library</p>	
Dr. Tà Quang Chiếu - [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> - [M]: 0913 522 275			6/38

6

<b>A</b>	<b>D</b>	
<b>A</b>	<b>I</b>	
<h2>Phân tích và xử lý dữ liệu với Pandas (03)</h2>		
Dr. Tà Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275		7/38

7

<b>A</b>	<b>D</b>	
<b>A</b>	<b>I</b>	
<h2>4. Quan sát dữ liệu với Pandas</h2> 		
Dr. Tà Quang Chiếu – [E]: <a href="mailto:quangchieu.ta@gmail.com">quangchieu.ta@gmail.com</a> – [M]: 0913 522 275		8/38

8

A D  
A I

## 1. Viewing/Inspecting Data

- df.info():** Hiển thị thông tin tổng quan của dataframe df bao gồm: Số hàng, số cột, số lượng dữ liệu không null, kiểu dữ liệu của từng thuộc tính.

```

1 # sử dụng .info để quan sát dữ liệu Data frame
2 data_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 192 entries, 0 to 191
Data columns (total 7 columns):
time                192 non-null object
Ha Noi              192 non-null float64
Vinh                192 non-null float64
Da Nang             192 non-null float64
Nha Trang           192 non-null float64
Ho Chi Minh         192 non-null float64
Ca Mau              192 non-null float64
dtypes: float64(6), object(1)
memory usage: 10.6+ KB
```

Dr. Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
9/38

9

A D  
A I

## 1. Viewing/Inspecting Data

- df.shape:** Kích thước của dataframe
- df.count():** Đếm số dòng dữ liệu không null trong dataframe

```

1 #Xác định kích cỡ của Data Frame
2 print('Kích thước của Data:', data_df.shape)
3 #Đếm số Lượng hàng dữ liệu không null theo từng cột
4 print('Số liệu của từng cột:')
5 print(data_df.count())

Kích thước của Data: (192, 7)
Số liệu của từng cột:
time                192
Ha Noi              192
Vinh                192
Da Nang             192
Nha Trang           192
Ho Chi Minh         192
Ca Mau              192
dtype: int64
```

Dr. Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
10/38

10

## A D A I 1. Viewing/Inspecting Data

- df.describe ():** Thực hiện tính toán các đặc trưng thống kê của dataframe (các thuộc tính số) bao gồm: Tổng số giá trị, giá trị trung bình, độ lệch chuẩn, giá trị max, min...

```
1 #Thực hiện thống kê dữ liệu
2 data_df.describe()
```

	Ha Noi	Vinh	Da Nang	Nha Trang	Ho Chi Minh	Ca Mau
count	192.000000	192.000000	192.000000	192.000000	192.000000	192.000000
mean	27.712292	26.719896	25.522500	26.166875	26.159219	26.732552
std	2.749369	2.314602	1.932761	0.923510	1.719259	1.821799
min	21.680000	22.600000	20.930000	24.500000	23.220000	23.990000
25%	25.645000	24.875000	24.010000	25.485000	24.797500	25.315000
50%	27.685000	26.360000	25.310000	26.085000	25.930000	26.265000
75%	29.947500	28.022500	26.932500	26.795000	27.485000	28.057500
max	33.450000	32.570000	29.880000	28.680000	31.060000	31.370000

Những thông tin trên giúp cho chúng ta có cảm nhận tổng quan và sự phân tán về dữ liệu, từ đó ta tìm kiếm phương pháp phù hợp để xử lý tiếp theo.

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
11/38

11

## A D A I 1. Viewing/Inspecting Data

- df.describe (include=['O']):** Thực hiện tính toán các đặc trưng thống kê của dataframe (các thuộc tính có kiểu Object) bao gồm: Tổng số giá trị (count), số giá trị khác nhau xuất hiện trong thuộc tính (unique), Tên giá trị xuất hiện nhiều nhất (top), Số lần xuất hiện của thuộc tính đó (freq).

```
In [24]: 1 #Thông kê tập dữ liệu Train các thuộc tính có dtype: Object
          2 train_df.describe(include=['O'])
```

Out[24]:


	Name	Sex	Ticket	Cabin	Embarked	
count	891	891	891	204	889	
unique	891	2	681	147	3	
top	Mellinger, Miss. Madeleine Violet	male	347082	G6	S	
freq		1	577	7	4	644

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
12/38

12

A D
A I

## 2. Handling missing data with pandas



Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

13/38

13

A D
A I

## Missing data

**Các nguyên nhân dẫn đến missing data:**

- Khuyết ngẫu nhiên (Missing at Random – MAR):
- Khuyết hoàn toàn ngẫu nhiên (Missing Completely at Random – MCAR):
- Khuyết không ngẫu nhiên (Missing not at Random – MNAR):

	time	Ha Noi	Vinh	Da Nang	Nha Trang	Ho Chi Minh	Ca Mau
0	00 15-9-2019	25.65	24.79	24.01	25.06	25.48	24.97
1	01 15-9-2019		24.21	24.02	24.93	25.16	24.83
2	02 15-9-2019	25.05	23.73	23.89	24.79	24.8	24.55
3	03 15-9-2019	24.79	23.36	23.83		24.74	24.48
4	04 15-9-2019	24.59	23.05	23.69	24.82	24.8	24.38
5	05 15-9-2019	24.4		23.52	24.79	24.87	24.4
6	06 15-9-2019	24.38	22.79	23.68	25.1	24.71	24.41
7	07 15-9-2019	26.72	25.61	24.92	26.56	25.03	24.91
8	08 15-9-2019	28.84	26.93	26.51	26.53	25.75	25.85
9	09 15-9-2019	30.29	28.72	27.48	26.95	26.64	26.79
10	10 15-9-2019		29.97			27.68	27.53
11	11 15-9-2019	32.04	28.93	26.86	27.38	28.43	28.98
12	12 15-9-2019	31.31	28.94	26.65	27.47	28.29	29.24
13	13 15-9-2019	30.95		27.83	27.44	28	30.66
14	14 15-9-2019	30.56	30.62	26.49	27.16	27.67	30.97
15	15 15-9-2019	31.13	30.58	26.29	26.68	27.29	30.59
16	16 15-9-2019	30.8	30.2		26.45	27.29	29.13
17	17 15-9-2019	29.94	29.36	25.8	26.67	26.69	28.72
18	18 15-9-2019	28.53	27.48	24.82	25.92	25.81	27.46
19	19 15-9-2019	28.89	27.03	24.93	25.88	25.93	27.07
20	20 15-9-2019	28.06	26.41	24.7		25.97	26.75
21	21 15-9-2019	27.43	26.2	24.41	25.62	25.94	26.32
22	22 15-9-2019	26.98	25.79	24.17	25.6	25.9	26.29
23	23 15-9-2019	26.68	25.31	23.81	25.53	25.8	26.36

```

1 #Đọc file dữ liệu chứa missing
2 #Khởi tạo sử dụng thư viện Pandas
3 import pandas as pd
4 #Xác định đường dẫn tới file dữ liệu missing
5 path = 'Data_C4/Data_Temp_missing.csv'
6 #Đọc file dữ liệu csv với pandas
7 data_temp = pd.read_csv(path)
8 data_temp

```

	time	Ha Noi	Vinh	Da Nang	Nha Trang	Ho Chi Minh	Ca Mau
0	00 15-9-2019	25.65	24.79	24.01	25.06	25.48	24.97
1	01 15-9-2019	NaN	24.21	24.02	24.93	25.16	24.83
2	02 15-9-2019	25.05	23.73	23.89	24.79	24.8	24.55
3	03 15-9-2019	24.79	23.36	23.83	NaN	24.74	24.48

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

14/38

14

A D  
A I

## 2.1 Detecting Missing Values

- **Thống kê dữ liệu missing trong dataframe:**
  - `df.isnull().sum()`

```

1  #Thống kê số liệu missing trong Data frame
2  #Theo từng cột
3  print('Số lượng missing data trong file dữ liệu:')
4  print(data_temp.isnull().sum())

```

Số lượng missing data trong file dữ liệu:

time	0
Ha Noi	2
Vinh	2
Da Nang	2
Nha Trang	3
Ho Chi Minh	0
Ca Mau	0
dtype: int64	

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
15/38

15

A D  
A I

## 2.1 Detecting Missing Values (2)

- **Thống kê dữ liệu missing trong dataframe:**
  - Xây dựng hàm thống kê `missing_values()`

```

1  #Xây dựng hàm thống kê dữ liệu missing trong dataframe:
2  #-----
3  #Đầu vào của hàm là 1 biến Dataframe
4  #Đầu ra bao gồm các thông số:
5  #Tổng số cột của file dữ liệu
6  #Tổng số cột có chứa dữ liệu missing
7  #Danh sách các cột chứa dữ liệu missing với 2 thông số:
8  #Tổng số giá trị missing tương ứng với cột đó
9  #Tỷ lệ % dữ liệu missing trên tổng số dữ liệu của cột
10 def missing_values(df):
11     mis_val = df.isnull().sum()
12     mis_val_percent = 100 * df.isnull().sum() / len(df)
13     mis_val_table = pd.concat([mis_val, mis_val_percent], axis=1)
14     mis_val_table_ren_columns = mis_val_table.rename(
15         columns = {0 : 'Số giá trị Missing', 1 : 'Tỷ lệ % missing'})
16     mis_val_table_ren_columns = mis_val_table_ren_columns[
17         mis_val_table_ren_columns.iloc[:,1] != 0].sort_values(
18         'Tỷ lệ % missing', ascending=False).round(1)
19     print ("File dữ liệu bao gồm có: " + str(df.shape[1]) + " cột.\n"
20           "Có " + str(mis_val_table_ren_columns.shape[0]) +
21           " cột chứa missing values.")
22     return mis_val_table_ren_columns

```

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
16/38

16



## 2.1 Detecting Missing Values (3)

- Thống kê dữ liệu missing trong dataframe:**
  - Xây dựng hàm thống kê `missing_values()`

```
1 missing_values(data_temp)
```

File dữ liệu bao gồm có: 7 cột.  
Có 4 cột chứa missing values.

	Số giá trị Missing	Tỷ lệ % missing
Nha Trang	3	12.5
Ha Noi	2	8.3
Vinh	2	8.3
Da Nang	2	8.3

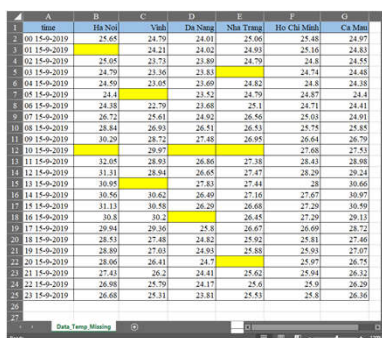
Dr. Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

17

## 2.1 Detecting Missing Values (4)

- `df.isnull().any(axis=1)`:** Kiểm tra trong từng hàng có thuộc tính nào chứa giá trị missing hay không? Nếu trong hàng chỉ cần có 1 thuộc tính missing – True

```
1 #Liệt kê danh sách các row bị missing data
2 #(Row có chứa thuộc tính bất kỳ bị missing - True )
3 #axis=1: Liệt kê các hàng / axis=0: Liệt kê các cột
4 data_temp.isnull().any(axis=1)
```



```
0 False
1 True
2 False
3 True
4 False
5 True
6 False
7 False
8 False
9 False
10 True
11 False
```

Dr. Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

18

## A D A I 2.1 Detecting Missing Values (5)

- df[df.isnull().any(axis=1)]:** Liệt kê chi tiết các hàng có chứa giá trị null trong một thuộc tính bất kỳ

```

1 #Liệt kê chi tiết các hàng có chứa giá trị null trong một thuộc tính bất kỳ
2 data_temp[data_temp.isnull().any(axis=1)]

```

	time	Ha Noi	Vinh	Da Nang	Nha Trang	Ho Chi Minh	Ca Mau
1	01 15-9-2019	NaN	24.21	24.02	24.93	25.16	24.83
3	03 15-9-2019	24.79	23.36	23.83	NaN	24.74	24.48
5	05 15-9-2019	24.40	NaN	23.52	24.79	24.87	24.40
10	10 15-9-2019	NaN	29.97	NaN	NaN	27.68	27.53
13	13 15-9-2019	30.95	NaN	27.83	27.44	28.00	30.66
16	16 15-9-2019	30.80	30.20	NaN	26.45	27.29	29.13
20	20 15-9-2019	28.06	26.41	24.70	NaN	25.97	26.75

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

19/38

19

## A D A I 2.1 Detecting Missing Values (6)

- pd.isnull(df["F1"]):** Liệt kê các hàng có chứa giá trị null trong một thuộc tính được chỉ định.

```

1 #Liệt kê các hàng có chứa giá trị null trong một cột được chỉ định.
2 x = pd.isnull(data_temp['Ha Noi'])
3 data_temp[x]

```

	time	Ha Noi	Vinh	Da Nang	Nha Trang	Ho Chi Minh	Ca Mau
1	01 15-9-2019	NaN	24.21	24.02	24.93	25.16	24.83
10	10 15-9-2019	NaN	29.97	NaN	NaN	27.68	27.53

```

1 #Liệt kê các hàng có chứa giá trị null trong một cột được chỉ định.
2 x = pd.isnull(data_temp['Nha Trang'])
3 data_temp[x]

```

	time	Ha Noi	Vinh	Da Nang	Nha Trang	Ho Chi Minh	Ca Mau
3	03 15-9-2019	24.79	23.36	23.83	NaN	24.74	24.48
10	10 15-9-2019	NaN	29.97	NaN	NaN	27.68	27.53
20	20 15-9-2019	28.06	26.41	24.70	NaN	25.97	26.75

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

20/38

20

A
D

A
I

## 2.2 Handling missing data

- Để xử lý dữ liệu missing cần phải hiểu sâu sắc tập dữ liệu, việc lựa chọn phương pháp nào phụ thuộc vào từng bài toán cụ thể, một số phương pháp xử lý dữ liệu missing cơ bản:

**1) Loại bỏ các missing (Deletion)**

**2) Thay thế các missing (Imputation)**

```

graph LR
    A[Handling Missing Data] --> B[Deletion]
    A --> C[Imputation]
    B --> B1[Deleting Rows (Listwise Deletion)]
    B --> B2[Pairwise Deletion]
    B --> B3[Deleting Columns]
    C --> D[Time-Series Problem]
    C --> E[General Problem]
    D --> D1[Data without Trend & without Seasonality]
    D --> D2[Data with Trend & without Seasonality]
    D --> D3[Data with Trend & with Seasonality]
    E --> E1[Categorical]
    E --> E2[Continuous]
    D1 --> D1a[Mean, Median, Mode, Random Sample Imputation]
    D2 --> D2a[Linear Interpolation]
    D3 --> D3a[Seasonal Adjustment + Interpolation]
    E1 --> E1a[Make NA as level, Multiple Imputation, Logistic Regression]
    E2 --> E2a[Mean, Median, Mode, Multiple Imputation, Linear Regression]
          
```

Dr. Tạ Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

21/38

21

A
D

A
I

## 1) Loại bỏ các missing (Deletion)

**df.dropna(axis=0) → loại bỏ hàng**

```

1 #1) Phương pháp 1: Loại bỏ các dữ liệu missing (Deletion)
2
3 #Xóa toàn bộ các hàng chứa missing data: axis=0 -> xóa hàng
4 data_new = data_temp.dropna(axis=0, how='any')
5 #Kết quả sau khi loại bỏ các row chứa missing
6 print(data_new)
          
```

	time	Ha Noi	Vinh	Da Nang	Nha Trang	Ho Chi Minh	Ca Mau
0	00 15-9-2019	25.65	24.79	24.01	25.06	25.48	24.97
2	02 15-9-2019	25.05	23.73	23.89	24.79	24.80	24.55
4	04 15-9-2019	24.59	23.05	23.69	24.82	24.80	24.38
6	06 15-9-2019	24.38	22.79	23.68	25.10	24.71	24.41
7	07 15-9-2019	26.72	25.61	24.92	26.56	25.03	24.91
8	08 15-9-2019	28.84	26.93	26.51	26.53	25.75	25.85
9	09 15-9-2019	30.29	28.72	27.48	26.95	26.64	26.79
11	11 15-9-2019	32.05	28.93	26.86	27.38	28.43	28.98
12	12 15-9-2019	31.31	28.94	26.65	27.47	28.29	29.24
14	14 15-9-2019	30.56	30.62	26.49	27.16	27.67	30.97
15	15 15-9-2019	31.13	30.58	26.29	26.68	27.29	30.59
17	17 15-9-2019	29.94	29.36	25.80	26.67	26.69	28.72
18	18 15-9-2019	28.53	27.48	24.82	25.92	25.81	27.46
19	19 15-9-2019	28.89	27.03	24.93	25.88	25.93	27.07
21	21 15-9-2019	27.43	26.20	24.41	25.62	25.94	26.32
22	22 15-9-2019	26.98	25.79	24.17	25.60	25.90	26.29
23	23 15-9-2019	26.68	25.31	23.81	25.53	25.80	26.36

Dr. Tạ Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

22/38

22

11

## 1) Loại bỏ các missing (Deletion)

**df.dropna(axis=1) → loại bỏ cột**

```

1 #1) Phương pháp 1: Loại bỏ các dữ liệu missing (Deletion)
2
3 #Xóa toàn bộ các cột chứa missing data: axis=1 -> xóa cột
4 data_new = data_temp.dropna(axis=1, how='any')
5 #Kết quả sau khi Loại bỏ các cột chứa missing
6 print(data_new)

```

	time	Ho Chi Minh	Ca Mau
0	00 15-9-2019	25.48	24.97
1	01 15-9-2019	25.16	24.83
2	02 15-9-2019	24.80	24.55
3	03 15-9-2019	24.74	24.48
4	04 15-9-2019	24.80	24.38
5	05 15-9-2019	24.87	24.40
6	06 15-9-2019	24.71	24.41
7	07 15-9-2019	25.03	24.91
8	08 15-9-2019	25.75	25.85
9	09 15-9-2019	26.64	26.79
10	10 15-9-2019	27.68	27.53
11	11 15-9-2019	28.43	28.98
12	12 15-9-2019	28.29	29.24
13	13 15-9-2019	28.00	30.66
14	14 15-9-2019	27.67	30.97
15	15 15-9-2019	27.29	30.59
16	16 15-9-2019	27.29	29.13
17	17 15-9-2019	26.69	28.72

Các cột Hà Nội, Vinh, Đà Nẵng, Nha Trang có chứa dữ liệu missing đã bị loại bỏ

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

23/38

23

## 2) Thay thế các missing (imputation)

**df.fillna(value) → thay thế bằng một giá trị cố định**

```

1 #PHƯƠNG PHÁP 2: Thay thế (Imputation)
2 #2.1) Thay thế các dữ liệu mất mát bằng một hằng số cố định
3 value = 25.0
4 #thay thế các giá trị missing bằng một giá trị cố định value
5 data_new = data_temp.fillna(value)
6 print(data_new)

```

	time	Hà Nội	Vinh	Đà Nẵng	Nha Trang	Ho Chi Minh	Ca Mau
0	00 15-9-2019	25.65	24.79	24.01	25.06	25.48	24.97
1	01 15-9-2019	25.00	24.21	24.02	24.93	25.16	24.83
2	02 15-9-2019	25.05	23.73	23.89	24.79	24.80	24.55
3	03 15-9-2019	24.79	23.36	23.83	25.00	24.74	24.48
4	04 15-9-2019	24.59	23.05	23.69	24.82	24.80	24.38
5	05 15-9-2019	24.40	25.00	23.52	24.79	24.87	24.40
6	06 15-9-2019	24.38	22.79	23.68	25.10	24.71	24.41
7	07 15-9-2019	26.72	25.61	24.92	26.56	25.03	24.91
8	08 15-9-2019	28.84	26.93	26.51	26.53	25.75	25.85
9	09 15-9-2019	30.29	28.72	27.48	26.95	26.64	26.79
10	10 15-9-2019	25.00	29.97	25.00	25.00	27.68	27.53
11	11 15-9-2019	32.05	28.93	26.86	27.38	28.43	28.98
12	12 15-9-2019	31.31	28.94	26.65	27.47	28.29	29.24
13	13 15-9-2019	30.95	25.00	27.83	27.44	28.00	30.66
14	14 15-9-2019	30.56	30.62	26.49	27.16	27.67	30.97
15	15 15-9-2019	31.13	30.58	26.29	26.68	27.29	30.59
16	16 15-9-2019	30.80	30.20	25.00	26.45	27.29	29.13

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

24/38

24

A D  
A I

## 2) Thay thế các missing (imputation)

df.fillna(method='pad') → thay thế bằng giá trị liền trước

```
1 #PHƯƠNG PHÁP 2: Thay thế (Imputation)
2 #2.2) Thay thế các dữ liệu mất mát bằng giá trị liền trước của nó
3 data_new2 = data_temp.fillna(method='pad')
4 print(data_new2)
```

	time	Ha Noi	Vinh	Da Nang	Nha Trang	Ho Chi Minh	Ca Mau
0	00 15-9-2019	25.65	24.79	24.01	25.06	25.48	24.97
1	01 15-9-2019	25.65	24.21	24.02	24.93	25.16	24.83
2	02 15-9-2019	25.05	23.73	23.89	24.79	24.80	24.55
3	03 15-9-2019	24.79	23.36	23.83	24.79	24.74	24.48
4	04 15-9-2019	24.59	23.05	23.69	24.82	24.80	24.38
5	05 15-9-2019	24.40	23.05	23.52	24.79	24.87	24.40
6	06 15-9-2019	24.38	22.79	23.68	25.10	24.71	24.41
7	07 15-9-2019	26.72	25.61	24.92	26.56	25.03	24.91
8	08 15-9-2019	28.84	26.93	26.51	26.53	25.75	25.85
9	09 15-9-2019	30.29	28.72	27.48	26.95	26.64	26.79
10	10 15-9-2019	30.29	29.97	27.48	26.95	27.68	27.53
11	11 15-9-2019	32.05	28.93	26.86	27.38	28.43	28.98
12	12 15-9-2019	31.31	28.94	26.65	27.47	28.29	29.24
13	13 15-9-2019	30.95	28.94	27.83	27.44	28.00	30.66
14	14 15-9-2019	30.56	30.62	26.49	27.16	27.67	30.97
15	15 15-9-2019	31.13	30.58	26.29	26.68	27.29	30.59
16	16 15-9-2019	30.80	30.20	26.29	26.45	27.29	29.13

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
25/38

25

A D  
A I

## 2) Thay thế các missing (imputation)

df.fillna(method='bfill') → thay thế bằng giá trị liền sau

```
1 #PHƯƠNG PHÁP 2: Thay thế (Imputation)
2 #2.3) Thay thế các dữ liệu mất mát bằng giá trị liền sau của nó
3 data_new3 = data_temp.fillna(method='bfill')
4 print(data_new3)
```

	time	Ha Noi	Vinh	Da Nang	Nha Trang	Ho Chi Minh	Ca Mau
0	00 15-9-2019	25.65	24.79	24.01	25.06	25.48	24.97
1	01 15-9-2019	25.05	24.21	24.02	24.93	25.16	24.83
2	02 15-9-2019	25.05	23.73	23.89	24.79	24.80	24.55
3	03 15-9-2019	24.79	23.36	23.83	24.82	24.74	24.48
4	04 15-9-2019	24.59	23.05	23.69	24.82	24.80	24.38
5	05 15-9-2019	24.40	22.79	23.52	24.79	24.87	24.40
6	06 15-9-2019	24.38	22.79	23.68	25.10	24.71	24.41
7	07 15-9-2019	26.72	25.61	24.92	26.56	25.03	24.91
8	08 15-9-2019	28.84	26.93	26.51	26.53	25.75	25.85
9	09 15-9-2019	30.29	28.72	27.48	26.95	26.64	26.79
10	10 15-9-2019	32.05	29.97	26.86	27.38	27.68	27.53
11	11 15-9-2019	32.05	28.93	26.86	27.38	28.43	28.98
12	12 15-9-2019	31.31	28.94	26.65	27.47	28.29	29.24
13	13 15-9-2019	30.95	30.62	27.83	27.44	28.00	30.66
14	14 15-9-2019	30.56	30.62	26.49	27.16	27.67	30.97
15	15 15-9-2019	31.13	30.58	26.29	26.68	27.29	30.59
16	16 15-9-2019	30.80	30.20	25.80	26.45	27.29	29.13

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
26/38

26



A D  
A I
2) Thay thế các missing (imputation)

**df.interpolate()** → thay thế giá trị bằng nội suy

```

1 #PHƯƠNG PHÁP 2: Thay thế (Imputation)
2 #2.4) Xử lý các giá trị missing theo phương pháp nội suy
3 #Sử dụng hàm interpolate để thay thế giá trị missing với tham số:
4 #Thuật toán nội suy: Tuyến tính (linear)
5 #Hướng nội suy: Tiến lên (forward)
6 data_new4 = data_temp.interpolate(method='linear', limit_direction='forward')
7 print(data_new4)

```

	time	Ha Noi	Vinh	Da Nang	Nha Trang	Ho Chi Minh	Ca Mau
0	00 15-9-2019	25.65	24.79	24.010	25.060	25.48	24.97
1	01 15-9-2019	25.35	24.21	24.020	24.930	25.16	24.83
2	02 15-9-2019	25.05	23.73	23.890	24.790	24.80	24.55
3	03 15-9-2019	24.79	23.36	23.830	24.805	24.74	24.48
4	04 15-9-2019	24.59	23.05	23.690	24.820	24.80	24.38
5	05 15-9-2019	24.40	22.92	23.520	24.790	24.87	24.40
6	06 15-9-2019	24.38	22.79	23.680	25.100	24.71	24.41
7	07 15-9-2019	26.72	25.61	24.920	26.560	25.03	24.91
8	08 15-9-2019	28.84	26.93	26.510	26.530	25.75	25.85
9	09 15-9-2019	30.29	28.72	27.480	26.950	26.64	26.79
10	10 15-9-2019	31.17	29.97	27.170	27.165	27.68	27.53
11	11 15-9-2019	32.05	28.93	26.860	27.380	28.43	28.98
12	12 15-9-2019	31.31	28.94	26.650	27.470	28.29	29.24
13	13 15-9-2019	30.95	29.78	27.830	27.440	28.00	30.66
14	14 15-9-2019	30.56	30.62	26.490	27.160	27.67	30.97
15	15 15-9-2019	31.13	30.58	26.290	26.680	27.29	30.59
16	16 15-9-2019	30.80	30.20	26.045	26.450	27.29	29.13

Dr. Tà Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275
27/38

27

A D  
A I
3. Chuyển đổi dữ liệu từ chuỗi sang số (labelEncoder) với pandas

X <sub>1</sub>	X <sub>2</sub>	y
5	8	calabar
9	3	uyo
8	6	owerrri
0	5	uyo
2	3	calabar
0	8	calabar
1	8	owerrri

LabelEncoder

→

```
{
  "calabar" ---> 0
  "owerrri" ---> 1
  "uyo" ---> 2
}
```

X <sub>1</sub>	X <sub>2</sub>	y
5	8	0
9	3	2
8	6	1
0	5	2
2	3	0
0	8	0
1	8	1

Dr. Tà Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275
28/38

28

### 3) LabelEncoder

Các model chỉ thực hiện trên dữ liệu dạng số. Do đó chúng ta cần phải chuyển đổi các nhãn sang số.

	Survived	Pclass	Sex	Age	Fare	Embarked	Title	IsAlone
0	0	3	male	22.0	7.2500	S	Mr	0
1	1	1	female	38.0	71.2833	C	Mrs	0
2	1	3	female	26.0	7.9250	S	Miss	1
3	1	1	female	35.0	53.1000	S	Mrs	0
4	0	3	male	35.0	8.0500	S	Mr	1
5	0	3	male	NaN	8.4583	Q	Mr	1
6	0	1	male	54.0	51.8625	S	Mr	1
7	0	3	male	2.0	21.0750	S	Master	0
8	1	3	female	27.0	11.1333	S	Mrs	0
9	1	2	female	14.0	30.0708	C	Mrs	0

Dr. Tạ Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

29/38

29

### 3) LabelEncoder

Sử dụng pandas.series.map()

```

1 #Chuyển đổi thuộc tính Sex về dạng số nguyên (int)
2 # trong đó: Female = 1; Male = 0
3 data_label['Sex'] = data_label['Sex'].map( {'female': 1, 'male': 0} ).astype(int)

1 #Hiển thị dữ liệu 10 mẫu đầu tiên sau khi đã chuyển đổi.
2 data_label.head(10)

```

	Survived	Pclass	Sex	Age	Fare	Embarked	Title	IsAlone
0	0	3	0	22	7.2500	S	Mr	0
1	1	1	1	38	71.2833	C	Mrs	0
2	1	3	1	26	7.9250	S	Miss	1
3	1	1	1	35	53.1000	S	Mrs	0
4	0	3	0	35	8.0500	S	Mr	1
5	0	1	0	54	51.8625	S	Mr	1
6	0	3	0	2	21.0750	S	Master	0
7	1	3	1	27	11.1333	S	Mrs	0
8	1	2	1	14	30.0708	C	Mrs	0
9	1	3	1	4	16.7000	S	Miss	0

Dr. Tạ Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275

30/38

30

A D  
A I

## 3) LabelEncoder

Sử dụng pandas.series.map()

```

1 #Chuyển đổi thuộc tính Embarked về dạng số nguyên (int)
2 # Trong đó: S = 0, C = 1, Q = 2
3 data_label['Embarked'] = data_label['Embarked'].map( {"S": 0, "C": 1, "Q": 2} ).astype(int)

1 #Hiển thị dữ liệu 10 mẫu cuối cùng sau khi đã chuyển đổi.
2 data_label.tail(10)

```

	Survived	Pclass	Sex	Age	Fare	Embarked	Title	IsAlone
15	0	3	0	2	29.1250	2	Master	0
16	0	3	1	31	18.0000	0	Mrs	0
17	0	2	0	35	26.0000	0	Mr	1
18	1	2	0	34	13.0000	0	Mr	1
19	1	3	1	15	8.0292	2	Miss	1
20	1	1	0	28	35.5000	0	Mr	1
21	0	3	1	8	21.0750	0	Miss	0
22	1	3	1	38	31.3875	0	Mrs	0
23	0	1	0	19	263.0000	0	Mr	0
24	1	3	1	60	7.8792	2	Miss	1

Dr. Tạ Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275
31/38

31

A D  
A I

## 3) LabelEncoder

Sử dụng pandas.series.map()

```

1 #Chuyển đổi dữ liệu thuộc tính Title:
2 #chuyển sang dạng số, với các giá trị tương ứng (Mr=1, Miss=2, Mrs=3, Master=4, Rare=5)
3 #Có thể sử dụng một biến kiểu Dictionary để chuyển đổi
4 title_mapping = {"Mr": 1, "Miss": 2, "Mrs": 3, "Master": 4, "Rare": 5}
5 data_label['Title'] = data_label['Title'].map(title_mapping).astype(int)

1 #Hiển thị dữ liệu 10 mẫu đầu tiên sau khi đã chuyển đổi.
2 data_label.head(10)

```

	Survived	Pclass	Sex	Age	Fare	Embarked	Title	IsAlone
0	0	3	0	22	7.2500	0	1	0
1	1	1	1	38	71.2833	1	3	0
2	1	3	1	26	7.9250	0	2	1
3	1	1	1	35	53.1000	0	3	0
4	0	3	0	35	8.0500	0	1	1
5	0	1	0	54	51.8625	0	1	1
6	0	3	0	2	21.0750	0	4	0
7	1	3	1	27	11.1333	0	3	0
8	1	2	1	14	30.0708	1	3	0
9	1	3	1	4	16.7000	0	2	0

Tham khảo thêm `sklearn.preprocessing.LabelEncoder()`

Dr. Tạ Quang Chiếu - [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) - [M]: 0913 522 275
32/38

32



A D
A I

# Thực hành

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
33/38

33

A D
A I

## Bài 19: Làm việc với Pandas

**Mô tả file dữ liệu: Bai19\_Personal.csv**

- File dữ liệu chứa thông tin của 300 bệnh nhân bị bệnh tim mạch

	A	B	C	D	E	F	G	H	I
	id	tuoi	gioitinh	loai	huyetap	cholesterol	nhieptim	thalassemia	Ketqua
2	1	63	Male	Typical angina	145	233	150	6	0
3	2	67	Male	Asymptomatic	160	286	108	3	1
4	3	67	Male	Asymptomatic	120	229	129	7	1
5	4	37	Male	Non-anginal pain	130	250	187	3	0
6	5	41	Female	Atypical angina	130	204	172		0
7	6	56	Male	Atypical angina	120	236	178	3	0
8	7	62	Female	Asymptomatic	140	268	160	3	1
9	8	57	Female	Asymptomatic	120	354	163	3	0
10	9	63	Male	Asymptomatic	130	254	147	7	1
11	10	53	Male	Asymptomatic	140	203	155	7	1
12	11	57	Male	Asymptomatic	140	192	148	6	0
13	12	56	Female	Atypical angina	140	294	153	3	0
14	13	56	Male	Non-anginal pain	130	256	142	6	1

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
34/38

34

A

D

A

I

## Bài 19: Làm việc với Pandas

**Chi tiết như sau:**

- Id:** Mã của bệnh nhân (số)
- Tuoi:** Tuổi của bệnh nhân (số)
- Gioitinh:** Giới tính của bệnh nhân (chuỗi: Male – Female)
- Loại:** Cho biết loại triệu chứng đau ngực mà bệnh nhân này mắc phải, với 4 giá trị: (Typical angina, Atypical angina, Non-anginal pain, Asymptomatic)
- Huyetap:** Huyết áp của bệnh nhân – đơn vị: mmhg (số)
- Cholesterol:** Chỉ số cholesterol của bệnh nhân – đơn vị: mg/dl (số)
- Nhiptim:** Thông số nhịp tim của bệnh nhân – đơn vị: lần/phút (số)
- Thalassemia:** Chỉ số Thalassemia của bệnh nhân chỉ gồm 3 giá trị (3: Bình thường | 4: Khiếm khuyết cố định | 7: Khiếm khuyết có thể đảo ngược)
- Ketqua:** Cho biết bệnh nhân có bị bệnh tim hay không? (0: Không bị bệnh tim mạch | 1: Bị bệnh tim mạch)

Dr. Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

35/38

35

A

D


A

I

## Bài 19: Làm việc với Pandas

**Yêu cầu 1:**

- Đọc dữ liệu từ file .csv vào biến kiểu dataframe**
- Hiển thị thông tin của 20 bệnh nhân đầu tiên và 30 bệnh nhân cuối cùng của tập dữ liệu.**
- Sử dụng phương thức .describe cho biết:**
  - Tuổi trung bình của các bệnh nhân trong tập dữ liệu
  - Tuổi của bệnh nhân trẻ nhất
  - Tuổi của bệnh nhân già nhất
  - Bao nhiêu bệnh nhân nam (Male)



Dr. Tạ Quang Chiêu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275

36/38

36

A
D
A
I
Bài 19: Làm việc với Pandas

**Yêu cầu 2:**

- Cho biết những cột nào trong dữ liệu có chứa missing data và số lượng missing là bao nhiêu.
- Liệt kê danh sách các bệnh nhân bị missing dữ liệu cột 'loai', cột 'thalassemia'

	id	tuoi	gioitinh	loai	huyetap	cholesterol	nhiptim	thalassemia	Ketqua
205	206	58	Male	NaN	128	259	130	7.0	1
218	219	59	Male	NaN	138	271	182	3.0	0
250	251	58	Male	NaN	146	218	105	7.0	1
270	271	66	Male	NaN	160	228	138	6.0	0
292	293	63	Male	NaN	140	187	144	7.0	1

	id	tuoi	gioitinh	loai	huyetap	cholesterol	nhiptim	thalassemia	Ketqua
4	5	41	Female	Atypical angina	130	204	172	NaN	0
20	21	64	Male	Typical angina	110	211	144	NaN	0
35	36	42	Male	Asymptomatic	140	226	178	NaN	0
240	241	41	Female	Atypical angina	126	306	163	NaN	0
265	266	52	Male	Asymptomatic	128	204	156	NaN	1
277	278	57	Male	Atypical angina	154	232	164	NaN	1
293	294	63	Female	Asymptomatic	124	197	136	NaN	1

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
37/38

37

A
D
A
I
Bài 19: Làm việc với Pandas

**Yêu cầu 3:**

- Xử lý dữ liệu missing ở cột 'loai' bằng cách thay thế các giá trị missing bằng giá trị cố định là một chuỗi: 'Asymptomatic'
- Xử lý dữ liệu missing ở cột 'thalassemia' bằng cách thay thế các giá trị missing bằng số 3

**Yêu cầu 4:**

- Chuyển đổi dữ liệu chuỗi (label) ở 2 cột **gioitinh** và **loai** sang dạng số

**Yêu cầu 5:**

- Lưu dataframe sau khi xử lý ở trên ra file: **Bai19\_personal\_finish.csv**

Dr. Tạ Quang Chiếu – [E]: [quangchieu.ta@gmail.com](mailto:quangchieu.ta@gmail.com) – [M]: 0913 522 275
38/38

38