```
อายุเฉลี่ย ของ ผู้ป่วยหญิง ในจังหวัดขอนแก่น 41.57142857142857
```

การจัดการ Missing Value

- ลบ record ที่เป็น missing
- แทนที่ ค่า missing ด้วยค่าที่เหมาะสม mean, default, category-unknown
- ใช้ ค่าจาก columns อื่นๆ ช่วยประมาณค่า ค่าใน column ที่หายไป (regression, deep learning, etc.)

ลบ record (dropna)

missing = `None` `NA` (not autorized) `NaN` (not a number)

In [ ]:
```
this_data.shape #(12423, 3)
```

Out[ ]:
```
(12653, 3)
```

In [ ]:
```
print(data_covid.shape)
print(data_covid.dropna().shape)
```

```
(12653, 10)
(6655, 10)
```

In [ ]:
```
this_data.dropna().shape
```

Out[ ]:
```
(8478, 3)
```

In [ ]:
```
this_data.dropna()
```

Out[ ]:

	sex	age	province_of_onset
0	หญิง	61.0	กรุงเทพมหานคร
1	หญิง	74.0	กรุงเทพมหานคร
2	หญิง	73.0	นครปฐม
3	ชาย	68.0	กรุงเทพมหานคร
4	หญิง	66.0	กรุงเทพมหานคร
...	...	...	...
12648	หญิง	44.0	ชลบุรี
12649	หญิง	52.0	ระยอง
12650	หญิง	23.0	ระยอง
12651	หญิง	29.0	ระยอง
12652	หญิง	22.0	ตาก

8478 rows × 3 columns

In [ ]:
```
this_data.dropna()
print(this_data.shape)
```

```
(12653, 3)
```

In [ ]:
```
this_data.dropna(inplace=True)
```

```
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:1: SettingWithCopyWar
ning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stabl
e/user_guide/indexing.html#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.
```

In [ ]:
```
this_data.dropna(inplace=True)
print(this_data.shape)
```

```
(8478, 3)
/usr/local/lib/python3.6/dist-packages/ipykernel_launcher.py:1: SettingWithCopyWar
ning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stabl
e/user_guide/indexing.html#returning-a-view-versus-a-copy
  """Entry point for launching an IPython kernel.
```

In [ ]:
```
this_data = data_covid[['sex','age','province_of_onset']]
this_data.shape
```

Out[ ]:
```
(12653, 3)
```

In [ ]:
```
this_data.dropna().shape
```

Out[ ]:
```
(8478, 3)
```

In [ ]:
```
this_data.dropna(subset=['age']).shape
```

Out[ ]:
```
(8881, 3)
```

แทน missing ด้วยค่าที่เหมาะสม (`fillna`)

In [ ]:
```
this_data = data_covid[['sex','age','province_of_onset']]
```

In [ ]:
```
this_data.head()
```

Out[ ]:

	sex	age	province_of_onset
0	หญิง	61.0	กรุงเทพมหานคร
1	หญิง	74.0	กรุงเทพมหานคร
2	หญิง	73.0	นครปฐม
3	ชาย	68.0	กรุงเทพมหานคร
4	หญิง	66.0	กรุงเทพมหานคร

In [ ]:
```
print(f'จำนวน record ก่อน drop missing ใน province {this_data.shape[0]}')
print(f'จำนวน record หลัง drop missing ใน province {this_data.dropna(subset=["provin
```

```
จำนวน record ก่อน drop missing ใน province 12653
จำนวน record หลัง drop missing ใน province 10662
```

In [ ]:
```
this_data_updated = this_data.fillna(value={'sex':'ไม่รู้', 'age':-1, 'province_of_ons
this_data_updated.head()
```

Out[ ]:

	sex	age	province_of_onset
0	หญิง	61.0	กรุงเทพมหานคร
1	หญิง	74.0	กรุงเทพมหานคร
2	หญิง	73.0	นครปฐม
3	ชาย	68.0	กรุงเทพมหานคร
4	หญิง	66.0	กรุงเทพมหานคร

In [ ]:
```
this_data_updated[this_data_updated['sex']=='ไม่รู้']
```

Out[ ]:

	sex	age	province_of_onset
4391	ไม่รู้	-1.0	สมุทรสาคร
4392	ไม่รู้	-1.0	สมุทรสาคร
4393	ไม่รู้	-1.0	สมุทรสาคร
4394	ไม่รู้	-1.0	สมุทรสาคร
4395	ไม่รู้	-1.0	สมุทรสาคร
...	...	...	...
12635	ไม่รู้	-1.0	สมุทรสาคร
12636	ไม่รู้	-1.0	สมุทรสาคร
12637	ไม่รู้	-1.0	สมุทรสาคร
12638	ไม่รู้	-1.0	สมุทรสาคร
12639	ไม่รู้	-1.0	สมุทรสาคร

2502 rows × 3 columns

In [ ]:
```
this_data_updated['province_of_onset']=='ไม่รู้'
```

Out[ ]:
```
0        False
1        False
2        False
3        False
4        False
         ...
12648    False
12649    False
12650    False
12651    False
12652    False
Name: province_of_onset, Length: 12653, dtype: bool
```

In [ ]:
```
data_covid[this_data_updated['province_of_onset']=='ไม่รู้']
```

Out[ ]:

	No.	announce_date	notification_date	sex	age	nationality	province_of_isolation	pro
414	415	3/22/2020	3/22/2020	ชาย	29.0	Thailand	กรุงเทพมหานคร	
529	530	3/22/2020	3/20/2020	ชาย	37.0	Thailand	กรุงเทพมหานคร	
555	556	3/22/2020	3/21/2020	หญิง	6.0	Thailand	กรุงเทพมหานคร	
576	577	3/22/2020	3/20/2020	ชาย	NaN	Thailand	กรุงเทพมหานคร	
588	589	3/22/2020	3/21/2020	ชาย	NaN	Thailand	กรุงเทพมหานคร	
...	...	...	...	...	...	...	...	
12405	12406	1/18/2021	1/17/2021	NaN	NaN	NaN	สมุทรสาคร	
12406	12407	1/18/2021	1/17/2021	NaN	NaN	NaN	สมุทรสาคร	
12407	12408	1/18/2021	1/17/2021	NaN	NaN	NaN	สมุทรสาคร	
12408	12409	1/18/2021	1/17/2021	NaN	NaN	NaN	สมุทรสาคร	
12409	12410	1/18/2021	1/17/2021	NaN	NaN	NaN	สมุทรสาคร	

1991 rows × 10 columns

In [ ]:
```
this_data_updated_2 = this_data.fillna(value={'province_of_onset':'ไม่รู้'})
```

In [ ]:
```
this_data_updated_2[this_data_updated['sex']=='ไม่รู้']
```

Out[ ]:

	sex	age	province_of_onset
4391	NaN	NaN	สมุทรสาคร
4392	NaN	NaN	สมุทรสาคร
4393	NaN	NaN	สมุทรสาคร
4394	NaN	NaN	สมุทรสาคร
4395	NaN	NaN	สมุทรสาคร
...	...	...	...
12635	NaN	NaN	สมุทรสาคร
12636	NaN	NaN	สมุทรสาคร
12637	NaN	NaN	สมุทรสาคร
12638	NaN	NaN	สมุทรสาคร
12639	NaN	NaN	สมุทรสาคร

2502 rows × 3 columns

การวนลูป record ในตาราง (`.iterrows`)

In [ ]:
```
this_data = data_covid[['sex','age','province_of_onset']]
```

In [ ]:
```
for each_row in this_data.iterrows():
    # print(each_row)
    # print(each_row[1])
    if (each_row[1]['age'] == 20) and (each_row[1]['province_of_onset'] == 'ขอนแก่น
        print(each_row)
```

```
(9334, sex                    ชาย
age                    20
province_of_onset    ขอนแก่น
Name: 9334, dtype: object)
```

การวนลูป แบบมองตารางแพนด้าส์(pandas dataframe) เป็น numpy array (`.iloc`)

In [ ]:
```
for each_row in range(this_data.shape[0]):
    if (this_data.iloc[each_row,1] == 20) and (this_data.iloc[each_row,2] == 'ขอนแ
        print(each_row)
        print(this_data.iloc[each_row,:])
```

```
9334
sex                    ชาย
age                     20
province_of_onset    ขอนแก่น
Name: 9334, dtype: object
```

Quiz ตัดตารางออกมาเฉพาะปี 2021 `announce_date` ในปี 2021

Hint

- วนลูปหา index ของปี 2021
- ตัดตารางมาเฉพาะ ปี 2021

In [ ]:
```
TF=list()
for each_row in data_covid.iterrows():
    if each_row[1]['announce_date'].split('/')[2] == '2021':
        TF.append(True)
    else:
        TF.append(False)

data_covid[TF].head()
```

Out[ ]:

	No.	announce_date	notification_date	sex	age	nationality	province_of_isolation	provin
6884	6885	1/1/2021	12/31/2020	หญิง	40.0	Thailand	กรุงเทพมหานคร	
6885	6886	1/1/2021	12/31/2020	หญิง	21.0	Thailand	ปทุมธานี	
6886	6887	1/1/2021	12/31/2020	หญิง	20.0	Thailand	นครปฐม	
6887	6888	1/1/2021	12/31/2020	หญิง	47.0	Thailand	สมุทรสาคร	
6888	6889	1/1/2021	12/31/2020	หญิง	36.0	Cambodia	ปทุมธานี	กรุ

Function ตัวช่วยใน pandas

`.describe()` คำนวณค่าทางสถิติของข้อมูลที่เป็นตัวเลข

In [ ]:
```
data_covid.describe()
```

Out[ ]:

	No.	age
count	12653.000000	8881.000000
mean	6327.000000	38.054748
std	3652.750813	15.125178
min	1.000000	0.100000
25%	3164.000000	27.000000
50%	6327.000000	36.000000
75%	9490.000000	48.000000
max	12653.000000	97.000000

`.mean()` คำนวณค่าเฉลี่ยของข้อมูลโดยไม่สนใจ missing

In [ ]:
```
data_covid[data_covid['sex']=='ชาย']['age'].mean()
```

Out[ ]:
```
39.46076940331987
```

.isnull()

In [ ]:
```
data_covid.isnull()
```

Out[ ]:

	No.	announce_date	notification_date	sex	age	nationality	province_of_isolation	prov
0	False	False	True	False	False	False	False	
1	False	False	True	False	False	False	False	
2	False	False	True	False	False	False	False	
3	False	False	True	False	False	False	False	
4	False	False	True	False	False	False	False	
...	...	...	...	...	...	...	...	
12648	False	False	False	False	False	False	False	
12649	False	False	False	False	False	False	False	
12650	False	False	False	False	False	False	False	
12651	False	False	False	False	False	False	False	
12652	False	False	False	False	False	False	False	

12653 rows × 10 columns

next

- .describe
- .mean
- HW วนลูป missing .isnull
- ต่อตารางแกน X แกน y
- .groupby
- save table

In [ ]:
```
```

In [ ]:
```
import pandas as pd
```

In [ ]:
```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

In [ ]:
```
import os

path = '/content/drive/My Drive/dataviz_2021_data'

covid_file_path = os.path.join(path,'pm-20-jan-2021.csv')
print(covid_file_path)
```

```
/content/drive/My Drive/dataviz_2021_data/pm-20-jan-2021.csv
```

In [ ]:
```
data_covid = pd.read_csv(covid_file_path)
data_covid.head()
```

Out[ ]:

	No.	announce_date	notification_date	sex	age	nationality	province_of_isolation	province_o
0	1	1/12/2020	NaN	หญิง	61.0	China	กรุงเทพมหานคร	กรุงเทพ
1	2	1/17/2020	NaN	หญิง	74.0	China	กรุงเทพมหานคร	กรุงเทพ
2	3	1/22/2020	NaN	หญิง	73.0	Thailand	นครปฐม	
3	4	1/22/2020	NaN	ชาย	68.0	China	กรุงเทพมหานคร	กรุงเทพ
4	5	1/24/2020	NaN	หญิง	66.0	China	นนทบุรี	กรุงเทพ

.isnull()?? None??

In [ ]:
```
data_covid.isnull()
```

Out[ ]:

	No.	announce_date	notification_date	sex	age	nationality	province_of_isolation	prov
0	False	False	True	False	False	False	False	
1	False	False	True	False	False	False	False	
2	False	False	True	False	False	False	False	
3	False	False	True	False	False	False	False	
4	False	False	True	False	False	False	False	
...	...	...	...	...	...	...	...	
12648	False	False	False	False	False	False	False	
12649	False	False	False	False	False	False	False	
12650	False	False	False	False	False	False	False	
12651	False	False	False	False	False	False	False	
12652	False	False	False	False	False	False	False	

12653 rows × 10 columns

.any() เอาค่าความจริงภายในแต่ละ column มา `OR` กัน

In [ ]:
```
data_covid.isnull().any()
```

Out[ ]:
```
No.                      False
announce_date            False
notification_date         True
sex                       True
age                       True
nationality               True
province_of_isolation     True
province_of_onset         True
district_of_onset         True
risk                      True
dtype: bool
```

.all() เอาค่าความจริงภายในแต่ละ column มา `AND` กัน

In [ ]:
```
data_covid.isnull().all()
```

Out[ ]:
```
No.                      False
announce_date            False
notification_date        False
sex                      False
age                      False
nationality              False
province_of_isolation    False
province_of_onset        False
district_of_onset        False
risk                     False
dtype: bool
```

In [ ]:
```
data_covid.iloc[0,0].isnull()
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-8-655f7c695cf3> in <module>()
----> 1 data_covid.iloc[0,0].isnull()

AttributeError: 'numpy.int64' object has no attribute 'isnull'
```

In [ ]:
```
data_covid['No.'][0].isnull()
```

```
---------------------------------------------------------------------------
AttributeError                            Traceback (most recent call last)
<ipython-input-9-85eef827ab2a> in <module>()
----> 1 data_covid['No.'][0].isnull()

AttributeError: 'numpy.int64' object has no attribute 'isnull'
```

In [ ]:
```
data_covid.iloc[0,0]
```

Out[ ]:
```
1
```

In [ ]:
```
data_covid.iloc[:1,0].isnull()
```

Out[ ]:
```
0    False
Name: No., dtype: bool
```

ต่อตารางแกน X แกน y

- ต่อแกน y คือ เพิ่ม records (เพิ่มจำนวนข้อมูล)
- ต่อแกน x คือ เพิ่ม columns (เพิ่มรายละเอียดของข้อมูล)

ต่อแกน Y pd.concat()

In [ ]:
```
data_covid['province_of_onset']=='ขอนแก่น'
```

Out[ ]:
```
0        False
1        False
2        False
3        False
4        False
         ...
12648    False
12649    False
12650    False
12651    False
12652    False
Name: province_of_onset, Length: 12653, dtype: bool
```

In [ ]:
```
dataKK = data_covid[data_covid['province_of_onset']=='ขอนแก่น']
dataKK.head()
```

Out[ ]:

	No.	announce_date	notification_date	sex	age	nationality	province_of_isolation	provin
180	181	3/18/2020	3/15/2020	ชาย	33.0	Thailand	ขอนแก่น	
462	463	3/22/2020	3/21/2020	หญิง	36.0	Thailand	ขอนแก่น	
1466	1467	3/30/2020	3/26/2020	ชาย	19.0	Thailand	ขอนแก่น	
1970	1971	4/3/2020	3/31/2020	หญิง	70.0	Thailand	ขอนแก่น	
2637	2638	4/15/2020	4/14/2020	หญิง	63.0	Thailand	ขอนแก่น	

In [ ]:

```
dataUD = data_covid[data_covid['province_of_onset']=='อุดรธานี']
dataUD.head()
```

Out[ ]:

	No.	announce_date	notification_date	sex	age	nationality	province_of_isolation	province
424	425	3/22/2020	3/21/2020	ชาย	33.0	Thailand	อุดรธานี	
434	435	3/22/2020	3/20/2020	หญิง	47.0	Thailand	อุดรธานี	
471	472	3/22/2020	3/22/2020	หญิง	26.0	Thailand	อุดรธานี	
883	884	3/25/2020	3/24/2020	ชาย	25.0	Thailand	อุดรธานี	
885	886	3/25/2020	3/24/2020	หญิง	20.0	Thailand	อุดรธานี	

In [ ]:
```
dataMS = data_covid[data_covid['province_of_onset']=='มหาสารคาม']
dataMS.head()
```

Out[ ]:

	No.	announce_date	notification_date	sex	age	nationality	province_of_isolation	prov
346	347	3/21/2020	3/20/2020	ชาย	34.0	Thailand	นนทบุรี	
789	790	3/24/2020	3/28/2020	ชาย	48.0	Thailand	มหาสารคาม	
6690	6691	12/31/2020	12/30/2020	หญิง	42.0	Thailand	มหาสารคาม	
10802	10803	1/12/2021	1/11/2021	หญิง	25.0	Thailand	สมุทรสาคร	

In [ ]:
```
dataMYisan = pd.concat([dataKK,dataUD,dataMS])
dataMYisan
```

Out[ ]:

	No.	announce_date	notification_date	sex	age	nationality	province_of_isolation	prov
180	181	3/18/2020	3/15/2020	ชาย	33.0	Thailand	ขอนแก่น	
462	463	3/22/2020	3/21/2020	หญิง	36.0	Thailand	ขอนแก่น	
1466	1467	3/30/2020	3/26/2020	ชาย	19.0	Thailand	ขอนแก่น	
1970	1971	4/3/2020	3/31/2020	หญิง	70.0	Thailand	ขอนแก่น	
2637	2638	4/15/2020	4/14/2020	หญิง	63.0	Thailand	ขอนแก่น	
2673	2674	4/17/2020	4/16/2020	ชาย	68.0	Thailand	ขอนแก่น	
5948	5949	12/26/2020	12/25/2020	หญิง	32.0	Thailand	ขอนแก่น	
6082	6083	12/27/2020	12/26/2020	หญิง	36.0	Thailand	ขอนแก่น	
9333	9334	1/7/2021	1/6/2021	ชาย	17.0	Thailand	ขอนแก่น	
9334	9335	1/7/2021	1/6/2021	ชาย	20.0	Thailand	ขอนแก่น	
10610	10611	1/12/2021	1/11/2021	ชาย	17.0	Thailand	ขอนแก่น	
11517	11518	1/16/2021	1/15/2021	ชาย	12.0	Thailand	ขอนแก่น	
11697	11698	1/17/2021	1/16/2021	หญิง	17.0	Thailand	ขอนแก่น	
11698	11699	1/17/2021	1/16/2021	หญิง	37.0	Thailand	ขอนแก่น	

	No.	announce_date	notification_date	sex	age	nationality	province_of_isolation	prov
424	425	3/22/2020	3/21/2020	ชาย	33.0	Thailand	อุดรธานี	
434	435	3/22/2020	3/20/2020	หญิง	47.0	Thailand	อุดรธานี	
471	472	3/22/2020	3/22/2020	หญิง	26.0	Thailand	อุดรธานี	
883	884	3/25/2020	3/24/2020	ชาย	25.0	Thailand	อุดรธานี	
885	886	3/25/2020	3/24/2020	หญิง	20.0	Thailand	อุดรธานี	
941	942	3/26/2020	3/25/2020	ชาย	31.0	Thailand	อุดรธานี	
1059	1060	3/27/2020	3/26/2020	ชาย	54.0	United States of America	อุดรธานี	
2013	2014	4/4/2020	4/3/2020	หญิง	49.0	Thailand	อุดรธานี	
5810	5811	12/24/2020	12/23/2020	ชาย	30.0	Thailand	อุดรธานี	
7433	7434	1/3/2021	NaN	หญิง	66.0	Thailand	ระยอง	
346	347	3/21/2020	3/20/2020	ชาย	34.0	Thailand	นนทบุรี	
789	790	3/24/2020	3/28/2020	ชาย	48.0	Thailand	มหาสารคาม	
6690	6691	12/31/2020	12/30/2020	หญิง	42.0	Thailand	มหาสารคาม	
10802	10803	1/12/2021	1/11/2021	หญิง	25.0	Thailand	สมุทรสาคร	

ต่อแกน X

- จับ 2 ตารางมาต่อกันเลย (merge)
- เลือกมาเพิ่มเฉพาะบาง column (map)

In [ ]:
```
data_province = data_covid[['No.','announce_date','province_of_onset']]
data_province
```

Out[ ]:

	No.	announce_date	province_of_onset
0	1	1/12/2020	กรุงเทพมหานคร
1	2	1/17/2020	กรุงเทพมหานคร
2	3	1/22/2020	นครปฐม
3	4	1/22/2020	กรุงเทพมหานคร
4	5	1/24/2020	กรุงเทพมหานคร
...	...	...	...
12648	12649	1/20/2021	ชลบุรี
12649	12650	1/20/2021	ระยอง
12650	12651	1/20/2021	ระยอง
12651	12652	1/20/2021	ระยอง
12652	12653	1/20/2021	ตาก

12653 rows × 3 columns

In [ ]:
```
data_human = data_covid[['No.','age','sex','nationality']]
data_human
```

Out[ ]:

	No.	age	sex	nationality
0	1	61.0	หญิง	China
1	2	74.0	หญิง	China
2	3	73.0	หญิง	Thailand
3	4	68.0	ชาย	China
4	5	66.0	หญิง	China
...	...	...	...	...
12648	12649	44.0	หญิง	Thailand
12649	12650	52.0	หญิง	Thailand
12650	12651	23.0	หญิง	Thailand
12651	12652	29.0	หญิง	Thailand
12652	12653	22.0	หญิง	Burma

12653 rows × 4 columns

แบบง่าย รู้ว่าสองตาราง record ตรงกัน

In [ ]:
```
full_table1 = data_human.merge(data_province)
full_table1.head()
```

Out[ ]:

	No.	age	sex	nationality	announce_date	province_of_onset
0	1	61.0	หญิง	China	1/12/2020	กรุงเทพมหานคร
1	2	74.0	หญิง	China	1/17/2020	กรุงเทพมหานคร
2	3	73.0	หญิง	Thailand	1/22/2020	นครปฐม
3	4	68.0	ชาย	China	1/22/2020	กรุงเทพมหานคร
4	5	66.0	หญิง	China	1/24/2020	กรุงเทพมหานคร

sort

In [ ]:
```
data_human2 = data_human.sort_values('age')
data_human2
```

Out[ ]:

	No.	age	sex	nationality
1987	1988	0.10	ชาย	Thailand
11497	11498	0.11	ชาย	Burma
6477	6478	0.11	หญิง	Thailand
1675	1676	0.30	ชาย	Japan
1075	1076	0.40	ชาย	Thailand
...	...	...	...	...
12635	12636	NaN	NaN	Thailand
12636	12637	NaN	NaN	Thailand
12637	12638	NaN	NaN	Thailand
12638	12639	NaN	NaN	Thailand
12639	12640	NaN	NaN	Thailand

12653 rows × 4 columns

In [ ]:
```
full_table2 = data_human2.merge(data_province)
full_table2.head()
```

Out[ ]:

	No.	age	sex	nationality	announce_date	province_of_onset
0	1988	0.10	ชาย	Thailand	4/4/2020	ระยอง
1	11498	0.11	ชาย	Burma	1/16/2021	สมุทรสาคร
2	6478	0.11	หญิง	Thailand	12/30/2020	เพชรบุรี
3	1676	0.30	ชาย	Japan	4/1/2020	กรุงเทพมหานคร
4	1076	0.40	ชาย	Thailand	3/27/2020	สุราษฎร์ธานี

In [ ]:
```
data_human2_renamed = data_human2.rename(columns={'No.': 'patientNumber'})
data_human2_renamed.head()
```

Out[ ]:

	patientNumber	age	sex	nationality
1987	1988	0.10	ชาย	Thailand
11497	11498	0.11	ชาย	Burma
6477	6478	0.11	หญิง	Thailand
1675	1676	0.30	ชาย	Japan
1075	1076	0.40	ชาย	Thailand

In [ ]:
```
data_human2_renamed.merge(data_province)
```

```
---------------------------------------------------------------------------
MergeError                                Traceback (most recent call last)
<ipython-input-25-ee0d027e7ef8> in <module>()
----> 1 data_human2_renamed.merge(data_province)

/usr/local/lib/python3.6/dist-packages/pandas/core/frame.py in merge(self, right,
how, on, left_on, right_on, left_index, right_index, sort, suffixes, copy, indicat
or, validate)
   7961             copy=copy,
   7962             indicator=indicator,
-> 7963             validate=validate,
   7964         )
   7965

/usr/local/lib/python3.6/dist-packages/pandas/core/reshape/merge.py in merge(left,
right, how, on, left_on, right_on, left_index, right_index, sort, suffixes, copy,
indicator, validate)
     85         copy=copy,
     86         indicator=indicator,
---> 87         validate=validate,
     88     )
     89     return op.get_result()

/usr/local/lib/python3.6/dist-packages/pandas/core/reshape/merge.py in __init__(se
lf, left, right, how, on, left_on, right_on, axis, left_index, right_index, sort,
suffixes, copy, indicator, validate)
    643             warnings.warn(msg, UserWarning)
    644
--> 645         self._validate_specification()
    646
    647         # note this function has side effects

/usr/local/lib/python3.6/dist-packages/pandas/core/reshape/merge.py in _validate_s
pecification(self)
   1215                 if len(common_cols) == 0:
   1216                     raise MergeError(
-> 1217                         "No common columns to perform merge on. "
   1218                         f"Merge options: left_on={self.left_on}, "
   1219                         f"right_on={self.right_on}, "

MergeError: No common columns to perform merge on. Merge options: left_on=None, ri
ght_on=None, left_index=False, right_index=False
```

In [ ]:
```
full_table3 = data_human2_renamed.merge(data_province,left_on='patientNumber',right
full_table3.head()
```

Out[ ]:

	patientNumber	age	sex	nationality	No.	announce_date	province_of_onset
0	1988	0.10	ชาย	Thailand	1988	4/4/2020	ระยอง
1	11498	0.11	ชาย	Burma	11498	1/16/2021	สมุทรสาคร
2	6478	0.11	หญิง	Thailand	6478	12/30/2020	เพชรบุรี
3	1676	0.30	ชาย	Japan	1676	4/1/2020	กรุงเทพมหานคร
4	1076	0.40	ชาย	Thailand	1076	3/27/2020	สุราษฎร์ธานี

map() เลือกมาเฉพาะบาง column มาแปะเพิ่มเข้าไป

In [ ]:
```
data_human2_renamed.head()
```

Out[ ]:

	patientNumber	age	sex	nationality
1987	1988	0.10	ชาย	Thailand
11497	11498	0.11	ชาย	Burma
6477	6478	0.11	หญิง	Thailand
1675	1676	0.30	ชาย	Japan
1075	1076	0.40	ชาย	Thailand

In [ ]:
```
data_covid.head()
```

Out[ ]:

	No.	announce_date	notification_date	sex	age	nationality	province_of_isolation	province_o
0	1	1/12/2020	NaN	หญิง	61.0	China	กรุงเทพมหานคร	กรุงเทพ
1	2	1/17/2020	NaN	หญิง	74.0	China	กรุงเทพมหานคร	กรุงเทพ
2	3	1/22/2020	NaN	หญิง	73.0	Thailand	นครปฐม	
3	4	1/22/2020	NaN	ชาย	68.0	China	กรุงเทพมหานคร	กรุงเทพ
4	5	1/24/2020	NaN	หญิง	66.0	China	นนทบุรี	กรุงเทพ

คุณสมบัติของ pandas ก็คือเราสามารถสร้าง column ใหม่ให้ตาราง df ได้ โดย

`df['ชื่อ column ใหม่'] = (list ที่มีจำนวนสมาชิกเท่ากับจำนวน record ของ df)`

In [ ]:
```
data_human2_renamed.head()
```

Out[ ]:

	patientNumber	age	sex	nationality
1987	1988	0.10	ชาย	Thailand
11497	11498	0.11	ชาย	Burma
6477	6478	0.11	หญิง	Thailand
1675	1676	0.30	ชาย	Japan
1075	1076	0.40	ชาย	Thailand

In [ ]:
```
data_human2_renamed['num'] = range(data_human2_renamed.shape[0])
data_human2_renamed
```

Out[ ]:

	patientNumber	age	sex	nationality	num
1987	1988	0.10	ชาย	Thailand	0
11497	11498	0.11	ชาย	Burma	1
6477	6478	0.11	หญิง	Thailand	2
1675	1676	0.30	ชาย	Japan	3
1075	1076	0.40	ชาย	Thailand	4
...	...	...	...	...	...
12635	12636	NaN	NaN	Thailand	12648
12636	12637	NaN	NaN	Thailand	12649
12637	12638	NaN	NaN	Thailand	12650
12638	12639	NaN	NaN	Thailand	12651
12639	12640	NaN	NaN	Thailand	12652

12653 rows × 5 columns

In [ ]:
```
data_human2_renamed['patientNumber'].map(data_covid.set_index('No.')['risk']) #.map
```

Out[ ]:
```
1987     สัมผัสใกล้ชิดกับผู้ป่วยยืนยันรายก่อนหน้านี้
11497                                 Cluster สมุทรสาคร
6477     สัมผัสใกล้ชิดกับผู้ป่วยยืนยันรายก่อนหน้านี้
1675     สัมผัสใกล้ชิดกับผู้ป่วยยืนยันรายก่อนหน้านี้
1075     สัมผัสใกล้ชิดกับผู้ป่วยยืนยันรายก่อนหน้านี้
                              ...
12635                                 Cluster สมุทรสาคร
12636                                 Cluster สมุทรสาคร
12637                                 Cluster สมุทรสาคร
12638                                 Cluster สมุทรสาคร
12639                                 Cluster สมุทรสาคร
Name: patientNumber, Length: 12653, dtype: object
```

In [ ]:
```
data_human2_renamed['detail'] = data_human2_renamed['patientNumber'].map(data_covid
data_human2_renamed
```

Out[ ]:

	patientNumber	age	sex	nationality	num	detail
1987	1988	0.10	ชาย	Thailand	0	สัมผัสใกล้ชิดกับผู้ป่วยยืนยันรายก่อนหน้านี้
11497	11498	0.11	ชาย	Burma	1	Cluster สมุทรสาคร
6477	6478	0.11	หญิง	Thailand	2	สัมผัสใกล้ชิดกับผู้ป่วยยืนยันรายก่อนหน้านี้
1675	1676	0.30	ชาย	Japan	3	สัมผัสใกล้ชิดกับผู้ป่วยยืนยันรายก่อนหน้านี้
1075	1076	0.40	ชาย	Thailand	4	สัมผัสใกล้ชิดกับผู้ป่วยยืนยันรายก่อนหน้านี้
...	...	...	...	...	...	...
12635	12636	NaN	NaN	Thailand	12648	Cluster สมุทรสาคร
12636	12637	NaN	NaN	Thailand	12649	Cluster สมุทรสาคร
12637	12638	NaN	NaN	Thailand	12650	Cluster สมุทรสาคร
12638	12639	NaN	NaN	Thailand	12651	Cluster สมุทรสาคร
12639	12640	NaN	NaN	Thailand	12652	Cluster สมุทรสาคร

12653 rows × 6 columns

HW7 สร้างตารางใหม่ ที่ค่าใน `sex` เป็น missing ทั้งหมด

- สรุปว่าทำไม record นั้นๆถึงเป็น missing

- .groupby
- create pandas table
- simple visualization
- save table

.groupby()

https://www.kaggle.com/crawford/python-groupby-tutorial#

In [ ]:
```
data_covid.groupby('nationality') ## จัดกลุ่มค่าที่เหมือนกันไว้ด้วยกัน
```

Out[ ]:
```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x7f0e67c0d9b0>
```

In [ ]:
```
data_covid.groupby('nationality').count()
```

Out[ ]:

nationality	No.	announce_date	notification_date	sex	age	province_of_isolation	province_of_o
Afghanistan	1	1	1	1	1	1	
Albania	3	3	3	3	3	3	
American	1	1	1	1	1	1	
American	1	1	1	1	1	1	
American Samoa	1	1	1	1	1	1	
...	...	...	...	...	...	...	
deutsch	1	1	1	1	1	1	
thailand	2	2	2	2	2	2	
ต่างด้าว	21	21	21	21	21	21	
ไทยใหญ่	1	1	1	1	1	1	
ไทใหญ่	1	1	1	1	1	1	

100 rows × 9 columns

In [ ]:
```
data_covid.groupby('nationality').mean()
```

Out[ ]:

nationality	No.	age
Afghanistan	4346.000000	31.000000
Albania	2377.333333	40.666667
American	3862.000000	39.000000
American	3905.000000	46.000000
American Samoa	11687.000000	21.000000
...	...	...
deutsch	3898.000000	55.000000
thailand	6823.000000	53.500000
ต่างด้าว	2944.952381	20.000000
ไทยใหญ่	10379.000000	30.000000
ไทใหญ่	5983.000000	33.000000

100 rows × 2 columns

In [ ]:
```
data_covid.groupby('nationality').max(' ') ### ทำไมใส่ ' ' แล้วรันได้!!
```

Out[ ]:

nationality	No.	age
Afghanistan	4346	31.0
Albania	3522	51.0
American	3862	39.0
American	3905	46.0
American Samoa	11687	21.0
...	...	...
deutsch	3898	55.0
thailand	6824	57.0
ต่างด้าว	2998	43.0
ไทยใหญ่	10379	30.0
ไทใหญ่	5983	33.0

100 rows × 2 columns

ทำ HW7 ด้วย groupby()

In [ ]:
```
data_covid['sex'].isnull()
```

Out[ ]:
```
0        False
1        False
2        False
3        False
4        False
         ...
12648    False
12649    False
12650    False
12651    False
12652    False
Name: sex, Length: 12653, dtype: bool
```

In [ ]:
```
missing_sex = data_covid[data_covid['sex'].isnull()]
missing_sex
```

Out[ ]:

	No.	announce_date	notification_date	sex	age	nationality	province_of_isolation	pro
4391	4392	12/20/2020	NaN	NaN	NaN	Burma	สมุทรสาคร	
4392	4393	12/20/2020	NaN	NaN	NaN	Burma	สมุทรสาคร	
4393	4394	12/20/2020	NaN	NaN	NaN	Burma	สมุทรสาคร	
4394	4395	12/20/2020	NaN	NaN	NaN	Burma	สมุทรสาคร	
4395	4396	12/20/2020	NaN	NaN	NaN	Burma	สมุทรสาคร	
...	...	...	...	...	...	...	...	
12635	12636	1/20/2021	1/19/2021	NaN	NaN	Thailand	สมุทรสาคร	
12636	12637	1/20/2021	1/19/2021	NaN	NaN	Thailand	สมุทรสาคร	
12637	12638	1/20/2021	1/19/2021	NaN	NaN	Thailand	สมุทรสาคร	
12638	12639	1/20/2021	1/19/2021	NaN	NaN	Thailand	สมุทรสาคร	
12639	12640	1/20/2021	1/19/2021	NaN	NaN	Thailand	สมุทรสาคร	

2502 rows × 10 columns

In [ ]:
```
missing_sex.groupby('nationality').describe()
```

Out[ ]:

	No.									
	count	mean	std	min	25%	50%	75%	max	count	me
nationality										
Burma	1366.0	5124.517570	472.244857	4392.0	4733.25	5096.5	5467.75	7075.0	0.0	N
Cambodia	61.0	6861.278689	275.220828	6543.0	6567.00	7035.0	7130.00	7160.0	0.0	N
Thailand	51.0	7819.274510	1649.397527	6272.0	7321.50	7509.0	7540.50	12640.0	1.0	2

In [ ]:
```
missing_sex.groupby('province_of_onset').describe()
```

Out[ ]:

									No.
	count	mean	std	min	25%	50%	75%	max	cour
province_of_onset									
ชลบุรี	18.0	7506.055556	6.637436	7497.0	7501.25	7505.5	7509.75	7524.0	1.
ระยอง	6.0	6289.000000	49.010203	6267.0	6268.25	6269.5	6270.75	6389.0	0.
สมุทรสาคร	1358.0	5136.421944	639.463677	4392.0	4731.25	5092.5	5461.75	12640.0	0.

In [ ]:
```
missing_sex.groupby(['province_of_onset','nationality']).describe() ## groupby หลา
```

Out[ ]:

		count	mean	std	min	25%	50%	75'
province_of_onset	nationality							
ชลบุรี	Thailand	18.0	7506.055556	6.637436	7497.0	7501.25	7505.5	7509.7
ระยอง	Thailand	1.0	6389.000000	NaN	6389.0	6389.00	6389.0	6389.0
สมุทรสาคร	Thailand	8.0	10251.125000	3294.200724	6272.0	6273.75	12636.5	12638.2
	Burma	1349.0	5103.564122	436.284721	4392.0	4729.00	5088.0	5455.0

In [ ]:
```
misssing_sex_no_burma = missing_sex[missing_sex['nationality']!='Burma']
misssing_sex_no_burma.groupby('risk').describe()
```

Out[ ]:

									No.
	count	mean	std	min	25%	50%	75%	max	count
risk									
Cluster ชลบุรี	18.0	7506.055556	6.637436	7497.0	7501.25	7505.5	7509.75	7524.0	1.0
Cluster ระยอง	7.0	6363.285714	201.569437	6267.0	6268.50	6270.0	6330.00	6809.0	0.0
Cluster สมุทรสาคร	895.0	11423.836872	1231.787200	6124.0	11176.50	11881.0	12191.50	12640.0	0.0
อยู่ระหว่างการสอบสวน	72.0	6854.111111	269.187068	6543.0	6564.75	7027.0	7128.25	7160.0	0.0

In [ ]:
```
missing_sex.groupby('risk').describe()
```

Out[ ]:

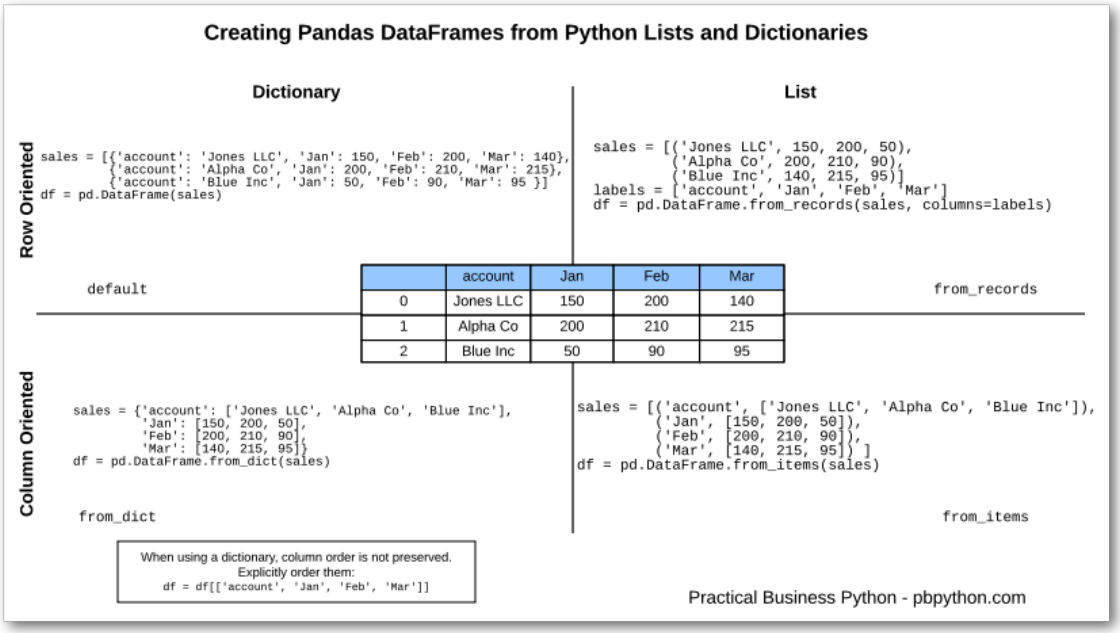
risk	count	mean	std	min	25%	50%	75%	max	No. count	m
Cluster ชลบุรี	18.0	7506.055556	6.637436	7497.0	7501.25	7505.5	7509.75	7524.0	1.0	
Cluster ระยอง	8.0	6452.250000	313.277946	6267.0	6268.75	6270.5	6494.00	7075.0	0.0	I
Cluster สมุทรสาคร	2258.0	7619.011515	3200.279792	4392.0	4978.25	5572.5	11763.75	12640.0	0.0	I
อยู่ระหว่างการสอบสวน	74.0	6852.945946	269.058169	6543.0	6564.25	7027.0	7127.75	7160.0	0.0	I

In [ ]:
```
missing_sex[missing_sex['risk']=='Cluster สมุทรสาคร'].groupby('announce_date').count
```

Out[ ]:

announce_date	No.	notification_date	sex	age	nationality	province_of_isolation	province_of_ons
1/14/2021	172	172	0	0	0	172	
1/17/2021	311	311	0	0	0	311	
1/18/2021	269	269	0	0	0	269	
1/20/2021	5	5	0	0	5	5	
1/5/2021	4	4	0	0	0	4	
1/7/2021	109	109	0	0	0	109	
12/20/2020	516	0	0	0	516	516	5
12/21/2020	360	0	0	0	360	360	3
12/22/2020	397	0	0	0	397	397	3
12/25/2020	35	0	0	0	35	35	
12/26/2020	12	0	0	0	12	12	
12/27/2020	36	18	0	0	18	36	
12/28/2020	14	0	0	0	14	14	
12/30/2020	3	0	0	0	1	3	
12/31/2020	15	0	0	0	15	15	

create pandas table



In [ ]:
```
records = [{'account':'Jones LLC','Jan':150,'Feb':200,'Mar':140},
           {'account':'Alpha Co','Jan':200,'Feb':210,'Mar':215},
           {'account':'Blue Inc','Jan':50,'Feb':90,'Mar':95}]
records_df = pd.DataFrame(records)
records_df
```

Out[ ]:

	account	Jan	Feb	Mar
0	Jones LLC	150	200	140
1	Alpha Co	200	210	215
2	Blue Inc	50	90	95

Simple Visualization

In [ ]:
```
df = pd.read_csv('https://raw.github.com/pandas-dev/pandas/master/pandas/tests/io/d
df
```

Out[ ]:

	SepalLength	SepalWidth	PetalLength	PetalWidth	Name
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

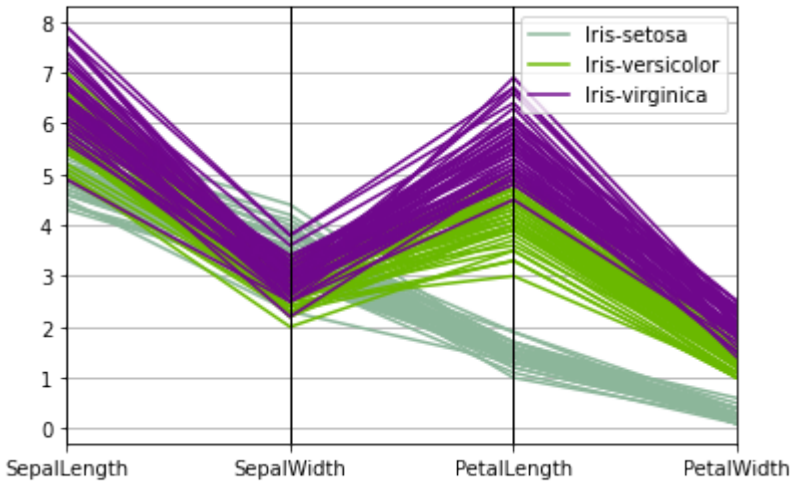
150 rows × 5 columns

In [ ]:
```
df.groupby('Name').count()
```

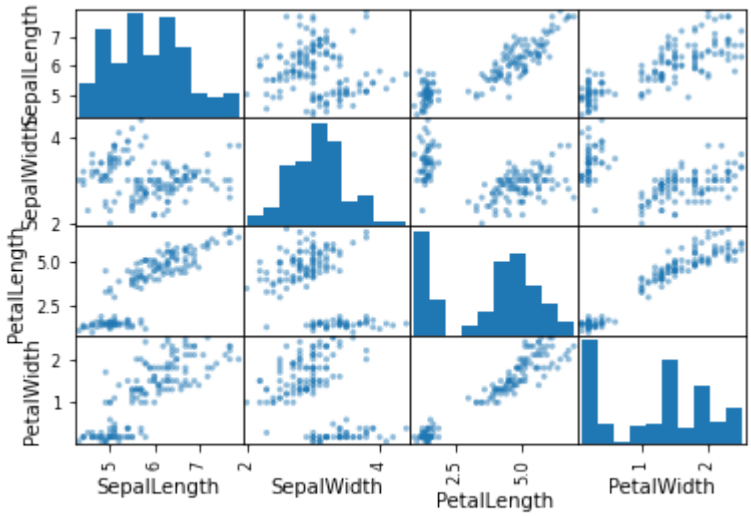
Out[ ]:

Name	SepalLength	SepalWidth	PetalLength	PetalWidth
Iris-setosa	50	50	50	50
Iris-versicolor	50	50	50	50
Iris-virginica	50	50	50	50

In [ ]:
```
pd.plotting.parallel_coordinates(df, 'Name');
```



In [ ]:
```
pd.plotting.scatter_matrix(df);
```



save table

In [ ]:
```
announce_date_count = missing_sex[missing_sex['risk']=='Cluster สมุทรสาคร'].groupby(
announce_date_count
```

Out[ ]:

announce_date	No.	notification_date	sex	age	nationality	province_of_isolation	province_of_ons
1/14/2021	172	172	0	0	0	172	
1/17/2021	311	311	0	0	0	311	
1/18/2021	269	269	0	0	0	269	
1/20/2021	5	5	0	0	5	5	
1/5/2021	4	4	0	0	0	4	
1/7/2021	109	109	0	0	0	109	
12/20/2020	516	0	0	0	516	516	5
12/21/2020	360	0	0	0	360	360	3
12/22/2020	397	0	0	0	397	397	3
12/25/2020	35	0	0	0	35	35	
12/26/2020	12	0	0	0	12	12	
12/27/2020	36	18	0	0	18	36	
12/28/2020	14	0	0	0	14	14	
12/30/2020	3	0	0	0	1	3	
12/31/2020	15	0	0	0	15	15	

In [ ]:
```
announce_date_count.to_csv('announce_date_count.csv')
```