

Class period 8

บทที่ 5 การเตรียมข้อมูลสำหรับการแสดงผล 1 (ต่อ)
Pandas 101 part3

1

Quiz สุ่มชื่อ (5 นาที)

- ให้อ่านค่าข้อมูลในตารางโดยใช้ .iloc ซึ่งแบบ numpy array หรือ matrix ให้ผลลัพธ์ออกมาเหมือนใช้ .iterrows()
- `this_data = data_covid[['sex', 'age', 'province_of_onset']]`
- `for each_row in this_data.iterrows():`
- `if (each_row[1]['age'] == 20) and (each_row[1]['province_of_onset'] == 'ขอนแก่น):`
- `print(each_row)`

2

การวนลูปอ่านข้อมูลแบบมองตาราง (pandas dataframe) เป็น numpy array หรือ matrix (.iloc[])

```
• this_data = data_covid[['sex', 'age', 'province_of_onset']]

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

3

การวนลูป

- `for each_row in range(this_data.shape[0]):`
- วนลูปอ่านลำดับแถวในตารางทีละแถวเก็บไว้ในตัวแปร each_row โดยจำนวนแถวทั้งหมดในตารางสามารถหาได้จาก
- `this_data.shape`
- `(839771, 3)`
- `this_data.shape[0]`
- `839771`
- `range(this_data.shape[0])` คือ สร้าง list ตัวเลขตามจำนวนแถวทั้งหมดเพื่อใช้วนลูปเข้าถึงแถว
- `[0, 1, 2, 3, ..., 839771]`

4

การใช้ .iloc[] ในการชี้ข้อมูลในตาราง

- ```
• if (this_data.iloc[each_row,1] == 20) and (this_data.iloc[each_row,2] ==
 'ขอนแก่น):
```
- สร้างเงื่อนไขสำหรับเลือกเฉพาะข้อมูลที่ต้องการคือ age=20 และ province\_of\_onset=ขอนแก่น โดยใช้ด้วย .iloc ตามด้วย each\_row คือลำดับแถวจากลูปและลำดับคอลัมน์(หลัก)ของข้อมูลที่ต้องการ ดังนี้
  - `this_data.iloc[each_row,1] == 20` คือ ชี้ข้อมูลแต่ละแถวในคอลัมน์ age(หลักที่ 1 นับจาก 0) ตรวจสอบว่าเท่ากับ 20 ในตัวแปร this\_data ที่เก็บตารางข้อมูล
  - `this_data.iloc[each_row,2] == 'ขอนแก่น):` คือ ชี้ข้อมูลแต่ละแถวในคอลัมน์ province\_of\_onset(หลักที่ 2) ตรวจสอบว่าเท่ากับ 'ขอนแก่น' ในตัวแปร this\_data ที่เก็บตารางข้อมูล

5

## print() แสดงข้อมูล

- ```
• print(each_row)
• print(this_data.iloc[each_row,:])
```
- เมื่อผ่านเงื่อนไขให้
 - `print each_row` คือ ตัวเลขลำดับแถว และ
 - `print this_data.iloc[each_row,:])` คือ ข้อมูลในแถวนั้นๆ ทุกคอลัมน์

6

Quiz ในห้อง (15 นาที)

- ตัดตารางออกมาเฉพาะปี 2021 announce_date ในปี 2021
- Hint
- วนลูปหา index ของปี 2021
- ตัดตารางมาเฉพาะ ปี 2021

6

7

เฉลย

```
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()
```

7

8

เตรียม list() วาง วนลูปและสร้างเงื่อนไข

- TF=list()
- สร้าง list วางเก็บไว้ในตัวแปร TF เพื่อเตรียมรับผลลัพธ์ True False ที่ได้จากการวนลูป
- for each_row in data_covid.iterrows():
- วนลูปอ่านค่าในข้อมูลตารางตัวแปร data_covid ที่จะแถวและเก็บในตัวแปร each_row
- if each_row[1]['announce_date'].split('/')[2] == '2021':
- สร้างเงื่อนไขเพื่อหาข้อมูลที่มีค่าในคอลัมน์ announce_date เท่ากับ 2021 โดยการใช้ .split('/')[2] เพื่อแยกข้อความให้เหลือเฉพาะปีสำหรับการเปรียบเทียบ
- ข้อมูลวันเดือนปีในคอลัมน์ announce_date รูปแบบเป็น 3/11/2021 คือ วัน/เดือน/ปี
- ดังนั้นแยกด้วยสัญลักษณ์ / ปีจะอยู่ตำแหน่งที่ 2 นับจาก 0

8

9

เขียน True False เข้าไปใน List

- if each_row[1]['announce_date'].split('/')[2] == '2021':
- TF.append(True)
- else:
- TF.append(False)
- ถ้าผ่านเงื่อนไขให้เขียน True เข้าไปใน list ที่เตรียมไว้ ถ้าไม่ผ่านให้เขียน False
- หมายความว่า ถ้าข้อมูลเป็นคอลัมน์ announce_date เท่ากับ 2021 ให้เขียน True เข้าไปใน list ถ้าไม่ ให้เขียน False
- ผลลัพธ์จะได้ list True False ตามจำนวนแถวในข้อมูลตาราง data_covid ที่เลือก True เฉพาะปี 2021

9

10

ผลลัพธ์

- data_covid[TF].head()
- ให้เลือกข้อมูลตาม list true false ที่ได้จากการวนลูปเลือกเฉพาะปี 2021
- ผลลัพธ์จะได้ตารางข้อมูลที่มีแค่ข้อมูลในคอลัมน์ announce_date เท่ากับ 2021

```
data_covid[TF].head()
  No.  announce_date  notified_date  sex  age  last  nationality  province_province
154  155    6/3/2021    5/5/2021  หญิง  55.0  0  Thailand  กรุงเทพมหานคร
4858  4859    1/1/2021    31/12/2020  หญิง  41.0  0  Thailand  กรุงเทพมหานคร
4859  4860    1/1/2021    31/12/2020  หญิง  21.0  0  Thailand  กรุงเทพมหานคร
4867  4868    1/1/2021    31/12/2020  หญิง  20.0  0  Thailand  กรุงเทพมหานคร
4868  4869    1/1/2021    31/12/2020  หญิง  41.0  0  Thailand  กรุงเทพมหานคร
```

10

11

Function ตัวช่วยใน pandas

- .describe() คำนวณค่าทางสถิติของข้อมูลที่เป็นตัวเลข
- .mean() คำนวณค่าเฉลี่ยของข้อมูลโดยไม่นับ missing
- .isnull() ตรวจสอบข้อมูลที่ missing (none)

11

12

.describe()

- ใช้สำหรับคำนวณค่าทางสถิติของข้อมูลที่เป็นตัวเลข
- ไม่นำข้อมูลที่ป็น missing หรือ none มาคำนวณ
- (ลบข้อมูลแถวที่มีค่าเป็น none ให้อัตโนมัติ)

```
data_covid.describe()
```

	No.	age
count	839771.000000	763605.000000
mean	419886.000000	35.700109
std	242421.150791	16.597799
min	1.000000	0.750000
25%	209943.500000	24.000000
50%	419886.000000	34.000000
75%	629828.500000	46.000000
max	839771.000000	440.000000

12

13

.mean()

- ใช้สำหรับหาค่าเฉลี่ยของข้อมูลโดยไม่สนใจ missing (ลบข้อมูลแถวที่มีค่าเป็น none ให้อัตโนมัติ)

```
data_covid[data_covid['sex']=='male']['age'].mean()
```

ผลลัพธ์จะได้

```
34.96292702130938
```

13

14

.isnull()

- ใช้ตรวจสอบค่า missing ในข้อมูลตาราง
- True คือ missing (ค่าว่าง)
- False คือไม่ใช่ค่าว่าง

```
data_covid.isnull()
```

	No.	announce_date	Notified date	sex	age	Unit	nationality	province_of_isolation
0	False	False	True	False	False	False	False	False
1	False	False	True	False	False	False	False	False
2	False	False	True	False	False	False	False	False
3	False	False	True	False	False	False	False	False
4	False	False	True	False	False	False	False	False
...
839766	False	False	False	False	False	True	False	False
839767	False	False	False	False	False	False	False	False
839768	False	False	False	False	False	False	False	False
839769	False	False	False	False	False	False	False	False
839770	False	False	False	False	False	False	False	False

839771 rows x 9 columns

14

15