



Class period 8

Pandas 101 part3

Quiz สุ่มชื่อ (5 นาที)



- ให้นำรูปแบบข้อมูลในตารางโดยใช้ .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)`

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,:])`



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 ตัวเลขตามจำนวนแถวทั้งหมดเพื่อใช้วนลูปเข้าทีละแถว
- `[0, 1, 2, 3, ..., 839771]`



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 คือลำดับแถวจากกลุ่มและลำดับคอลัมน์(หลัก)ของข้อมูลที่ต้องการ ดังนั้น
- `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

print แสดงข้อมูล



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

Quiz ในห้อง (15 นาที)



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

- `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()`



เตรียม list วาง วนลูปและสร้างเงื่อนไข

- `TF=list()`
- สร้าง list วางเก็บไว้ในตัวแปร TF เพื่อเตรียมรับผลลัพธ์ True False ที่ได้จากการวนลูป
- `for each_row in data_covid.iterrows():`
- วนลูปอ่านค่าในข้อมูลตารางตัวแปร `data_covid` ทีละแถวและเก็บในตัวแปร `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

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 ถ้าข้อมูลปีในคอลัมน์ announce_date เท่ากับ 2021 ให้เขียน True เข้าไปใน list ถ้าไม่ ให้เขียน False
- ผลลัพธ์จะได้ list True False ตามจำนวนแถวในข้อมูลตาราง data_covid ที่เลือก True เฉพาะปี 2021

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

```
data_covid[TF].head()
```

	No.	announce_date	Notified date	sex	age	Unit	nationality	province_of_isolation	risk
124	125	6/3/2021	5/3/2021	หญิง	55.0	ปี	Thailand	ปทุมธานี	Cluster คลาด
6885	6886	1/1/2021	31/12/2020	หญิง	40.0	ปี	Thailand	กรุงเทพมหานคร	Cluster สมุทรสา
6886	6887	1/1/2021	31/12/2020	หญิง	21.0	ปี	Thailand	ปทุมธานี	Cluster ระย
6887	6888	1/1/2021	31/12/2020	หญิง	20.0	ปี	Thailand	นครปฐม	Cluster บ้าน
6888	6889	1/1/2021	31/12/2020	หญิง	47.0	ปี	Thailand	สมุทรสาคร	Cluster สมุทรสา



Function ตัวช่วยใน pandas

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

.describe()

- ใช้สำหรับคำนวณค่าทางสถิติของข้อมูลที่เป็นตัวเลข
- ไม่นำข้อมูลที่เป็น 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



.mean()

- ใช้สำหรับช่วยคำนวณค่าเฉลี่ยของข้อมูลโดยไม่สนใจ missing (ลบข้อมูลแถวที่มีค่าเป็น none ให้อัตโนมัติ)
- `data_covid[data_covid['sex']=='ชาย']['age'].mean()`
- ผลลัพธ์จะได้
- 34.96292702130938



.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	False	True	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 11 columns