

Class period 6

Pandas 101

Pandas

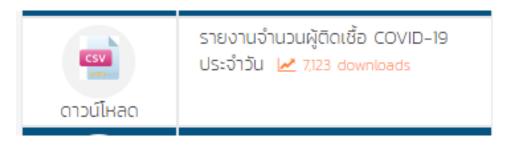


- Pandas เป็นหนึ่งใน package ที่สำคัญของ python ใช้สำหรับจัดการข้อมูลรูปแบบตาราง .CSV
- import pandas as pd

Download ข้อมูลรายงาน COVID-19 ประจำวัน ข้อมูลประจำประเทศไทย



https://data.go.th/dataset/covid-19-daily



covid-19-daily_data_dictionary

∠ 4,364 downloads

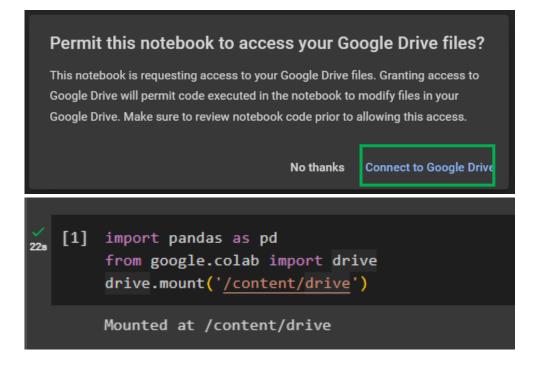
CSV = Comma Separated Values ในการจะใส่คาแต่ละค่า จะใช้ comma ในการแยก

Meta data = Data that description data ใช้อธิบายช้อมูล

การนำข้อมูลเข้า



- 1. สร้าง folder ใน google drive และนำข้อมูล .csv ที่ดาวน์โหลดเข้าไปเก็บไว้ใน folder ที่สร้าง
- 2. นำเข้า package pandas และ package ของ google.colab ที่ชื่อ drive เพื่อเชื่อมต่อ google drive กับ google colab
- import pandas as pd
- from google.colab import drive
- drive.mount('/content/drive')
- 3. กด Connect Google Drive และเลือก Account
- 4. กด select all และกด continue



Import os



- นำเข้า package os เพื่อใช้ในการทำงานต่างๆที่เกี่ยวกับไฟล์ เช่น การชี้ไฟล์ การลบไฟล์ การสร้างโฟลเดอร์ เป็นต้น
- โดยในกรณี google drive จะใช้ os เพื่อชี้ไฟล์ ว่าไฟล์ที่ต้องการใช้งานอยู่ path ไหนใน google drive ที่เชื่อม
- path คือเส้นทางที่อยู่ไฟล์ จะทำงานเหมือนกับ path ใน window เช่น
- E:\WORKSPACE\Basic Programming\confirmed-cases.csv
- หมายความว่า ไฟล์ confirmed-cases.csv อยู่ใน drive E โฟลเดอร์ WORKSPACE ในโฟลเดอร์ Basic Programming

การ set path



- path = '/content/drive/MyDrive/dataviz_2024_data'
- set 'path' ที่ชี้ไปยังโฟลเดอร์ที่เก็บไฟล์ .csv ไว้ใน google drive และเก็บ string ไว้ในตัวแปร path
- โดย path หรือเส้นทางที่ชี้ไปยังโฟลเดอร์และไฟล์ต่างๆ ใน os ของ window, mac หรือlinux จะใช้สัญลักษณ์ใน path แตกต่าง กัน
- Package os จะช่วยให้สามารถเชื่อม path โดยไม่ต้องคำนึงถึงสัญลักษณ์ เพราะ os จะใส่สัญลักษณ์เชื่อมให้เองตาม platform ที่ใช้งานอยู่ เช่น
- ถ้าใช้ os ของ window ก็จะเชื่อมด้วย \
- ถ้าใช้ os ของ mac หรือ linux จะเชื่อมด้วย /

คำสั่ง os.path.join



- เป็นคำสั่งที่ใช้สำหรับเชื่อม path เข้าด้วยกัน
- import os
- covid file path = os.path.join(path, confirmed-cases.csv')
- หมายความว่า เชื่อมตัวแปร path ที่ set ไว้ก่อนหน้านี้เข้ากับชื่อไฟล์ confirmed-cases.csvเก็บไว้ในตัวแปร covid_file_path
- print(covid_file_path)
- ผลลัพธ์จะได้เส้นทางไปยังไฟล์ที่ต้องการอยู่ในตัวแปร covid_file_path
- /content/drive/My Drive/dataviz_2024_data/confirmed-cases.csv

load data to memory (pd.read csv)



- pd.read_csv เป็นคำสั่งที่ใช้สำหรับโหลดข้อมูล
- data_covid = pd.read_csv(covid_file_path)
- โหลดข้อมูลไฟล์ confirmed-cases.csv ตามเส้นทาง covid file path
- data covid
- ผลลัพธ์จะได้หน้าไฟล์ csv

	No.	announce_date	notification_date	sex	age	nationality	province_of_isolation	province_of_onset	district_of_onset	
0	1	1/12/2020	NaN	หญิง	61.0	China	กรุงเทพมหานคร	กรุงเทพมหานคร	NaN	คนต่า เดินท จา ปร
1	2	1/17/2020	NaN	หญิง	74.0	China	กรุงเทพมหานคร	กรุงเทพมหานคร	NaN	คนต่า เดิน <i>ท</i> จา ปร
2	3	1/22/2020	NaN	หญิง	73.0	Thailand	นครปฐม	นครปฐม	เมือง	คนต่า เดินท จา ปร
3	4	1/22/2020	NaN	ชาย	68.0	China	กรุงเทพมหานคร	กรุงเทพมหานคร	NaN	คนต่า เดินท จา ปร
4	5	1/24/2020	NaN	หญิง	66.0	China	นนทบุรี	กรุงเทพมหานคร	NaN	คนต่า เดินท จา ปร
12648	12649	1/20/2021	1/19/2021	หญิง	44.0	Thailand	ชลบุรี	ชลบุรี	บางละมุง	Quara

Parameter: encoding



- ใช้สำหรับกำหนดภาษาของไฟล์ที่จะอ่าน เช่น
- data_covid = pd.read_csv(covid_file_path, encoding='utf-8')
- encoding='utf-8'อาจจะอ่านภาษาไทยได้แต่จะมี csv ภาษาไทยบางไฟล์ที่มันอ่านไม่ได้แล้วเกิด error
- การอ่านไฟล์ csv ที่มีข้อมูลที่เป็นภาษาไทยที่ครอบคลุมที่สุดจะใช้
- encoding='iso-8859-11'

คำสั่ง .head()



• data_covid.shape ชื่อตัวแปรที่เก็บข้อมูลตามด้วย .head() ใช้เพื่อให้แสดงชื่อคอลัมน์และข้อมูลในตาราง เฉพาะ 5 แถวแรก int, default=5

- สามารถกำหนดจำนวนคอลัมน์ที่ต้องการให้แสดงได้ เช่น
- data_covid.head(10) จะแสดงชื่อคอลัมน์และข้อมูลในตาราง 10 แถว

คำสั่ง .shape



- data covid.shape
- ชื่อตัวแปรที่เก็บข้อมูลตามด้วย .shape ใช้ตรวจสอบขนาดของข้อมูล ผลลัพธ์จะได้
- (839771, 11)
- หมายความว่า มีข้อมูลทั้งหมด 839,771 แถว มีคอลัมน์ 11 คอลัมน์

การชี้ค่าในข้อมูลตารางแบบ basic



- ใช้ชื่อคอลัมน์ในการดึงข้อมูลในคอลัมน์ที่ต้องการ
- data_covid['province_of_onset']

```
data_covid['province_of_onset']
          กรุงเทพมหานคร
          กรุงเทพมหานคร
                  นครปฐม
          กรุงเทพมหานคร
          กรุงเทพมหานคร
839766
              กาญจนบุรี
              กาญจนบุรี
839767
839768
              กาญจนบุรี
839769
              กาญจนบุรี
               กาญจนบุรี
839770
Name: province_of_onset, Length: 839771, dtype: object
```

การชี้ค่าในข้อมูลตารางแบบ basic



- การใช้ชื่อคอลัมน์และลำดับแถวในการดึงข้อมูลในแถวและคอลัมน์ที่ต้องการ
- data_covid['province_of_onset'][4]
- ผลลัพธ์จะได้ข้อมูลแถวที่ 4 นับจาก 0 ในคอลัมน์ province_of_onset
- 'กรุงเทพมหานคร'

การชี้ค่าในข้อมูลตารางแบบ .iloc



- โดยการมองมุมมองข้อมูลตารางในรูปแบบ numpy array หรือ matrix จะใช้ตำแหน่งเพื่อชี้ข้อมูลที่ต้องการ เช่น
- data covid.iloc[4,9]
- ผลลัพธ์จะได้ข้อมูลแถวที่ 4 คอลัมน์ที่ 9 (ในมุมมอง matrix คือหลักที่ 9) นับจาก 0 คือคอลัมน์ province_of_onset
- 'กรุงเทพมหานคร'

Table slicing การเลือกข้อมูลเฉพาะคอลัมน์ที่ต้องการ



- การเลือกข้อมูลเฉพาะคอลัมน์ที่ต้องการมาเก็บไว้ในตัวแปรเพื่อนำไปใช้งาน
- smaller_table = data_covid[['announce_date','province_of_onset','risk']]
- หมายความว่า เลือกข้อมูลคอลัมน์ชื่อ announce_date, province_of_onset, risk ในข้อมูลที่เก็บอยู่ในตัว แปร data_covid และเก็บข้อมูลเฉพาะคอลัมน์ที่เลือกไว้ในตัวแปร smaller_table
- ผลลัพธ์จะได้

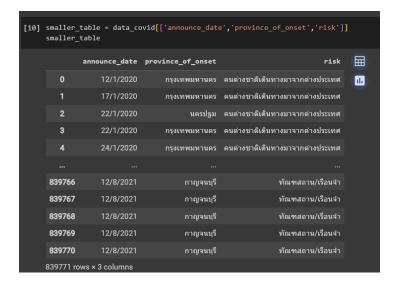


Table slicing การเลือกข้อมูลเฉพาะแถวที่ต้องการแบบง่าย



- การเลือกข้อมูลเฉพาะแถวที่ต้องการแบบง่าย คือการมองมุมมองข้อมูลตารางในรูปแบบ array แต่การนำไปใช้ งาน ใช้งานอะไรไม่ค่อยได้
- data covid.iloc[1:5,:]
- หมายความวา
- 1:5 คือเลือกข้อมูลที่อยู่ในแถวที่ 1 ไปจนถึงแถวที่ 4
- ,: คือเลือกทุกคอลัมน์ ดังนั้น
- data_covid.iloc[1:5,:] คือเลือกข้อมูลในตัวแปร data_covid ที่อยู่ในแถวที่ 1 ไปจนถึงแถวที่ 4 และ เลือกทุกคอลัมน์

Table slicing การเลือกข้อมูลเฉพาะแถวที่ต้องการแบบ advanced



- การเลือกข้อมูลเฉพาะแถวที่ต้องการแบบ advanced คือการใช logic query ในการเลือกข้อมูล
- data_covid[data_covid['province_of_onset'] == 'ขอนแก่น']
- หมายความว่า เลือกข้อมูลที่อยู่ในตัวแปร data_covid โดยกำหนดชื่อคอลัมน์ที่ต้องการคือ province_of_onset และต้องการข้อมูลทุกแถวที่มีข้อมูลในคอลัมน์ province_of_onset เป็นจังหวัดขอนแก่น

วิธีการเลือกข้อมูลเฉพาะแถวที่ต้องการแบบ advanced



- การทำงานของการเลือกข้อมูลเฉพาะแถวที่ต้องการแบบ advanced
- คือ การใส่แถวที่ต้องการ และใส่ list True/False ที่มีขนาดเท่ากับจำนวนแถว เพื่อเปรียบเทียบข้อมูลในแถว นั้นๆ ด้วยเงื่อนไข logical expression (True/False)) เช่น
- สร้างตารางใช้สำหรับยกตัวอย่าง
- eight_rows_covid = data_covid.iloc[:8,:]
- eight rows covid
- หมายความว่า เลือกข้อมูลในตัวแปร data_covid แถวที่ 0 ถึงแถวที่ 7 ทุกคอลัมน์เก็บไว้ในตัวแปร eight rows covid

การทำงานของการเลือกข้อมูลเฉพาะแถวที่ต้องการแบบ advanced



- ใส่แถวที่ต้องการด้วยการกำหนดค่า True(แถวที่ต้องการ)/False(แถวที่ไม่ต้องการ)
- eight_rows_covid[[True, True, False, True, True, True, True, False]]
- ผลลัพธ์จะได้ข้อมูลตารางตามค่ำ True/False ที่เลือกใน list คือแถวที่ 0,1,3,4,5,6
- เช่นเดียวกันกับการสร้าง list ของ logical expression แต่แทนที่จะเลือกเองโดยการใส่ list True/False ให้ กำหนดเงื่อนไขและข้อมูลที่ต้องการ เพื่อเปรียบเทียบและเลือกข้อมูลที่ตรงตามเงื่อนไข โดยถ้าตรงตามเงื่อนไข คือ True ไม่ตรงตามเงื่อนไขคือ False

การสร้าง list ของ logical expression

• Name: province of onset, dtype: bool



```
• eight rows covid['province of onset'] == 'กรุงเทพมหานคร'
• ผลลัพธ์จะได้
        True
        True
       False
• 3
        True
        True
• 5
        True
       True
       False
```

นำ list ของ logical expression ที่สร้างมาใช้งาน



- ซึ่งเมื่อนำมาใช้งานเลือกข้อมูลเฉพาะแถวที่ต้องการแบบ advanced คือการใช้ logic query ในการเลือกข้อมูล นั้นเอง
- eight rows covid[eight rows covid['province of onset'] == 'กรุงเทพมหานคร']
- ผลลัพธ์จะได้ข้อมูลทุกแถวที่มีข้อมูลในคอลัมน์ province_of_onset เป็น กรุงเทพมหานคร คือแถวที่ 0,1,3,4,5,6

Homework class period 6



- (ให้ใช้เฉพาะที่อาจารย์สอนไปแล้วในวิชานี้)
- คำนวณ อายุเฉลี่ย ของผู้หญิง และผู้ชาย ของข้อมูลทั้งหมด
- คำนวณ อายุเฉลี่ย ของผู้หญิง และผู้ชาย ของผู้ปวยในจังหวัดขอนแก่น
- หาจำนวนผู้ปวยที่เป็นคน "คนต่างชาติเดินทางมาจากต่างประเทศ"