

Class period 6

บทที่ 4 การจัดการข้อมูลในรูปแบบตาราง
Pandas 101

1

Pandas

- Pandas เป็นหนึ่งใน package ที่สำคัญของ python ใช้สำหรับการจัดการข้อมูลในรูปแบบตาราง CSV
- `import pandas as pd`

2

Download ข้อมูลรายงาน COVID-19 ประจำวัน ข้อมูลประจำประเทศไทย

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



งานจากเว็บไซต์ covid-19 ประจำวัน [ดูไฟล์ download](#)

CSV = Comma Separated Values
ในการแยกค่าข้อมูลด้วยเครื่องหมาย comma ในการแทน

Meta data = Data that description data
ข้อมูลอธิบายข้อมูล

3

การนำข้อมูลเข้า

- สร้าง folder ใน google drive และนำไฟล์ `csv` ที่เรามีอยู่มาอัปโหลดไว้ใน folder ที่สร้าง
- นำ package pandas มาใช้ google.colab เพื่อ drive เพื่อเชื่อมกับ google drive ที่ google colab

```

import pandas as pd
from google.colab import drive
drive.mount('/content/drive')

3. nk Connect Google Drive และเลือก Account
4. nk select all when continue
  
```



4

Import os

- นำ package os เพื่อใช้ในการทำงานต่างๆที่เกี่ยวข้องกับไฟล์ เช่น การเขียนไฟล์ การลบไฟล์ การสร้างไฟล์ลอคเตอร์ เป็นต้น
- โดยในการใช้ google drive จะมี os เพื่อใช้เพื่อว่าไฟล์ที่ต้องการใช้จาก path ใน google drive ที่เขียน
- path คือเส้นทางที่ชี้ไปยังตำแหน่งของไฟล์ path ใน window เช่น
- E:\WORKSPACE\Basic Programming\confirmed-cases.csv
- หมายความว่า ไฟล์ `confirmed-cases.csv` อยู่ใน drive E โฟลเดอร์ `WORKSPACE` ในโฟลเดอร์ `Basic Programming`

5

การ set path

```

path = '/content/drive/My Drive/dataviz_2024_data'

• data 'path' ที่ใช้กับโฟลเดอร์ที่เก็บไฟล์ csv ใน google drive และเก็บ string ไว้ในตัวแปร path
• โดย path หรือชื่อของโฟลเดอร์จะแตกต่างกันไปใน os ของ window, mac หรือ linux ทำให้มีสัญลักษณ์ path แต่ละตัว
• Package os จะช่วยให้สามารถเขียน path โดยไม่ต้องกังวลว่าสัญลักษณ์ของแต่ละ platform จะต่างกันหรือไม่
• ถ้าใช้ os ของ window ก็เขียนแบบนี้
• ถ้าใช้ os ของ mac หรือ linux ก็เขียนแบบนี้
  
```

6

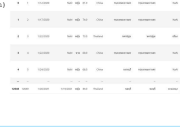
คำสั่ง os.path.join()

- เป็นคำสั่งที่ใช้สำหรับเชื่อม path เข้าด้วยกัน
- `import os`
- `covid_file_path = os.path.join(path, 'confirmed-cases.csv')`
- หมายความว่า เชื่อมตัวแปร path ที่เราได้จากขั้นตอนที่ 4 เข้ากับชื่อไฟล์ `confirmed-cases.csv` ที่เราได้มาอัปโหลดไว้ใน drive
- `print(covid_file_path)`
- ผลลัพธ์จะได้ออกมาในรูปแบบที่พร้อมสำหรับการนำไปใช้กับ `covid_file_path`
- `/content/drive/My Drive/dataviz_2024_data/confirmed-cases.csv`

7

load data to memory คำสั่ง pd.read_csv()

- `pd.read_csv` เป็นคำสั่งที่ใช้สำหรับโหลดข้อมูล
- `data_covid = pd.read_csv(covid_file_path)`
- โดยที่ข้อมูลที่ได้คือ `confirmed-cases.csv` ตามเส้นทาง `covid_file_path`
- `data_covid`
- ผลลัพธ์ที่ได้คือ csv



8

Parameter: encoding

- ใช้สำหรับกำหนดการเข้ารหัสไฟล์ที่จะอ่าน เช่น
- `data_covid = pd.read_csv(covid_file_path, encoding='utf-8')`
- `encoding='utf-8'` ถ้าจะอ่านภาษาไทยในไฟล์ csv ภาษาไทยบางไฟล์ที่มันไม่ได้เข้ารหัส error
- การอ่านไฟล์ csv ที่มีข้อมูลที่เป็นภาษาไทยที่ระบบคอมพิวเตอร์จะรู้
- `encoding='iso-8859-11'`

9

การทำงานของเลือกข้อมูลเฉพาะแถวที่ต้องการแบบ advanced

- โด่แถวที่ต้องการด้วยการกำหนดค่า True(แถวที่ต้องการ)/False(แถวที่ไม่ต้องการ)
- `eight_rows_covid[(True, True, False, True, True, True, False)]`
- ผลลัพธ์คือข้อมูลตารางตามค่า True/False ที่เลือกใน list คือแถวที่ 0, 1, 3, 4, 5, 6
- เทคนิคเกี่ยวกับการสร้าง list ของ logical expression แต่ละตัวที่จะเลือกโดยใช้การใส่ list True/False ไม่กำหนดเงื่อนไขและข้อมูลที่ต้องการ เพื่อเปรียบเทียบและเลือกข้อมูลที่ต้องการตามเงื่อนไข โดยถ้าตรงตามเงื่อนไขคือ True ไม่ตรงตามเงื่อนไขคือ False

19

การสร้าง list ของ logical expression

```

eight_rows_covid['province_of_onset'] == 'กรุงเทพมหานคร'
• ผลลัพธ์คือ
• 0 True
• 1 True
• 2 False
• 3 True
• 4 True
• 5 True
• 6 True
• 7 False
• Name: province_of_onset, dtype: bool

```

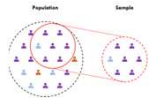
20

นำ list ของ logical expression ที่สร้างมาใช้งาน

- ใช้เพื่อมาใช้งานเลือกข้อมูลเฉพาะแถวที่ต้องการแบบ advanced คือการใช้ logic query ในการเลือกข้อมูลนั่นเอง
- `eight_rows_covid[eight_rows_covid['province_of_onset'] == 'กรุงเทพมหานคร']`
- ผลลัพธ์คือข้อมูลเฉพาะแถวที่มีข้อมูลในคอลัมน์ province_of_onset เป็น กรุงเทพมหานคร คือแถวที่ 0, 1, 3, 4, 5, 6

21


Bias ในข้อมูล



- Bias คืออคติหรือความลำเอียงที่อาจแฝงอยู่ในข้อมูล ซึ่งอาจเกิดจากการเก็บข้อมูล การออกแบบแบบสำรวจ หรืออคติในการเลือกข้อมูล

22

สำรวจ Bias ในชุดข้อมูล



- ชุดข้อมูลไม่มีความเป็นธรรม เช่น หรือข้อจำกัดที่จัดเป็นตัวอย่างเป็นเช่นนั้น
- ให้น้ำหนักกับระดัข้อมูลเพื่อระบุ bias ที่อาจเกิดขึ้น เช่น การมีตัวแทนของบางกลุ่มน้อยเกินไปหรือมากเกินไป

23

เทคนิคการประมวลผลข้อมูล

- ตรวจสอบเทคนิคการประมวลผลข้อมูลที่สามารถตรวจสอบ bias เช่น
- การจัดการกับข้อมูลที่หายไป
- การจัดการกับค่าผิดปกติ (outliers)
- การสร้างความสมดุลของการกระจายตัวของข้อมูล

24

Algorithmic Bias



- Bias สามารถขยายผลหรือบิดเบือนได้ทั้งจาก ฐานข้อมูลที่มี และโมเดลการเรียนรู้ของเครื่อง
- ตัวอย่างของ algorithmic bias เช่น ระบบการแนะนำภาพยนตร์ที่เลือกเฉพาะหนังที่ผู้ชายดู
- อคติด้านจริยธรรมการนำข้อมูลเชิงลึกไปใช้และ ความเสี่ยงที่ข้อมูลจะเป็นการเลือกปฏิบัติในการตัดสินใจเลือกปฏิบัติ

25

กรณีศึกษาและตัวอย่างจากโลกแห่งความเป็นจริง

- นำมาสอนกรณีศึกษาและตัวอย่างจากโลกแห่งความเป็นจริงที่ขึ้นกับอคติด้วยข้อมูลที่มีเฉพาะทางจริง
- ในระบบการวินิจฉัยทางการแพทย์ที่ใช้ข้อมูลเชิงลึกของ bias นี้อาจส่งผลต่อความแม่นยำและการวินิจฉัย
- กรณีศึกษาเกี่ยวกับอคติของระบบที่ได้อินพุตจากการวินิจฉัยทางการแพทย์ และนำมาประมวลผลให้มีความถูกต้องในการวินิจฉัย
- <https://www.profit.com/resources/shocking-bias>

26

Homework class period 6

- (ให้วิเคราะห์การอ่านต่อไปนี้ในวิชา)
- ด้านผล อายุเฉลี่ย ของผู้หญิง และผู้ชาย ของข้อมูลทั้งหมด
- ด้านผล อายุเฉลี่ย ของผู้หญิง และผู้ชาย ของกลุ่มในจังหวัดขอนแก่น
- หากจำนวนผู้หญิงที่เกินกว่า "ค่าเฉลี่ยที่คิดหาจากค่าเฉลี่ย"

27