



Class period 6

บทที่ 4 การจัดการข้อมูลในรูปแบบตาราง

Pandas 101





Pandas



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

Download ข้อมูลรายงาน COVID-19 ประจำวัน ข้อมูลประจำประเทศไทย

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

 ดาวน์โหลด	รายงานจำนวนผู้ติดเชื้อ COVID-19 ประจำวัน  7,123 downloads
 ดาวน์โหลด	covid-19-daily_data_dictionary  4,364 downloads

CSV = Comma Separated Values

ในการจะใส่ค่าแต่ละค่า จะใช้ comma ในการแยก

Meta data = Data that description data

ใช้อธิบายข้อมูล

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

Permit this notebook to access your Google Drive files?

This notebook is requesting access to your Google Drive files. Granting access to Google Drive will permit code executed in the notebook to modify files in your Google Drive. Make sure to review notebook code prior to allowing this access.

No thanks

Connect to Google Drive

✓
22s

```
[1] import pandas as pd
    from google.colab import drive
    drive.mount('/content/drive')
```

Mounted at /content/drive



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/My Drive/dataviz_2024_data'`
- Set '`path`' ที่ชี้ไปยังโฟลเดอร์ที่เก็บไฟล์ .csv ไว้ใน google drive และเก็บ string ไว้ในตัวแปร `path`
- โดย `path` หรือเส้นทางที่ชี้ไปยังโฟลเดอร์และไฟล์ต่างๆ ใน os ของ window, mac หรือ linux จะใช้สัญลักษณ์ใน `path` แตกต่างกัน
- Package `os` จะช่วยให้สามารถเชื่อม `path` โดยไม่ต้องคำนึงถึงสัญลักษณ์ เพราะ `os` จะใส่สัญลักษณ์เชื่อมให้เองตาม platform ที่ใช้งานอยู่ เช่น
- ถ้าใช้ `os` ของ window ก็เชื่อมด้วย `\`
- 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.csv เก็บไว้ในตัวแปร 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 คำสั่ง 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	คนตำ เดินV จา ปร
1	2	1/17/2020		NaN	หญิง	74.0	China	กรุงเทพมหานคร	กรุงเทพมหานคร	NaN	คนตำ เดินV จา ปร
2	3	1/22/2020		NaN	หญิง	73.0	Thailand	นครปฐม	นครปฐม	เมือง	คนตำ เดินV จา ปร
3	4	1/22/2020		NaN	ชาย	68.0	China	กรุงเทพมหานคร	กรุงเทพมหานคร	NaN	คนตำ เดินV จา ปร
4	5	1/24/2020		NaN	หญิง	66.0	China	นนทบุรี	กรุงเทพมหานคร	NaN	คนตำ เดินV 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']  
  
0      กรุงเทพมหานคร  
1      กรุงเทพมหานคร  
2      นครปฐม  
3      กรุงเทพมหานคร  
4      กรุงเทพมหานคร  
...  
839766   กาญจนบุรี  
839767   กาญจนบุรี  
839768   กาญจนบุรี  
839769   กาญจนบุรี  
839770   กาญจนบุรี  
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` ในข้อมูลที่เก็บอยู่ในตัวแปร `data_covid` และเก็บข้อมูลเฉพาะคอลัมน์ที่เลือกไว้ในตัวแปร `smaller_table`
- ผลลัพธ์จะได้

```
[10] smaller_table = data_covid[['announce_date', 'province_of_onset', 'risk']]
      smaller_table
```

	announce_date	province_of_onset	risk
0	12/1/2020	กรุงเทพมหานคร	คนต่างชาติเดินทางมาจากต่างประเทศ
1	17/1/2020	กรุงเทพมหานคร	คนต่างชาติเดินทางมาจากต่างประเทศ
2	22/1/2020	นครปฐม	คนต่างชาติเดินทางมาจากต่างประเทศ
3	22/1/2020	กรุงเทพมหานคร	คนต่างชาติเดินทางมาจากต่างประเทศ
4	24/1/2020	กรุงเทพมหานคร	คนต่างชาติเดินทางมาจากต่างประเทศ
...
839766	12/8/2021	กาญจนบุรี	ท้องถิ่นสถาน/เรือนจำ
839767	12/8/2021	กาญจนบุรี	ท้องถิ่นสถาน/เรือนจำ
839768	12/8/2021	กาญจนบุรี	ท้องถิ่นสถาน/เรือนจำ
839769	12/8/2021	กาญจนบุรี	ท้องถิ่นสถาน/เรือนจำ
839770	12/8/2021	กาญจนบุรี	ท้องถิ่นสถาน/เรือนจำ

839771 rows x 3 columns



Table slicing การเลือกข้อมูลเฉพาะแถวที่ต้องการแบบง่าย

- การเลือกข้อมูลเฉพาะแถวที่ต้องการแบบง่าย คือการมองมุมมองข้อมูลตารางในรูปแบบ array แต่การนำไปใช้งาน ใช้งานอะไรไม่ค่อยได้
- `data_covid.iloc[1:5, :]`
- หมายความว่า
- `1:5` คือเลือกข้อมูลที่อยู่ในแถวที่ 1 ไปจนถึงแถวที่ 4
- `, :` คือเลือกทุกคอลัมน์ ดังนั้น
- `data_covid.iloc[1:5, :]` คือเลือกข้อมูลในตัวแปร `data_covid` ที่อยู่ในแถวที่ 1 ไปจนถึงแถวที่ 4 และเลือกทุกคอลัมน์

Table slicing การเลือกข้อมูลเฉพาะแถวที่ต้องการแบบ advanced



- การเลือกข้อมูลเฉพาะแถวที่ต้องการแบบ advanced คือการใช้ 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

การทำงานของการทำงานการเลือกข้อมูลเฉพาะแถวที่ต้องการแบบ advanced



- ใส่แถวที่ต้องการด้วยการกำหนดค่า True(แถวที่ต้องการ)/False(แถวที่ไม่ต้องการ)
- `eight_rows_covid[[True, True, False, True, True, True, True, False]]`
- ผลลัพธ์จะได้ข้อมูลตารางตามค่า True/False ที่เลือกใน list คือแถวที่ 0, 1, 3, 4, 5, 6
- เช่นเดียวกันกับการสร้าง list ของ logical expression แต่แทนที่จะเลือกเองโดยการใส่ list True/False ให้กำหนดเงื่อนไขและข้อมูลที่ต้องการ เพื่อเปรียบเทียบและเลือกข้อมูลตรงตามเงื่อนไข โดยถ้าตรงตามเงื่อนไขคือ True ไม่ตรงตามเงื่อนไขคือ False



การสร้าง list ของ logical expression

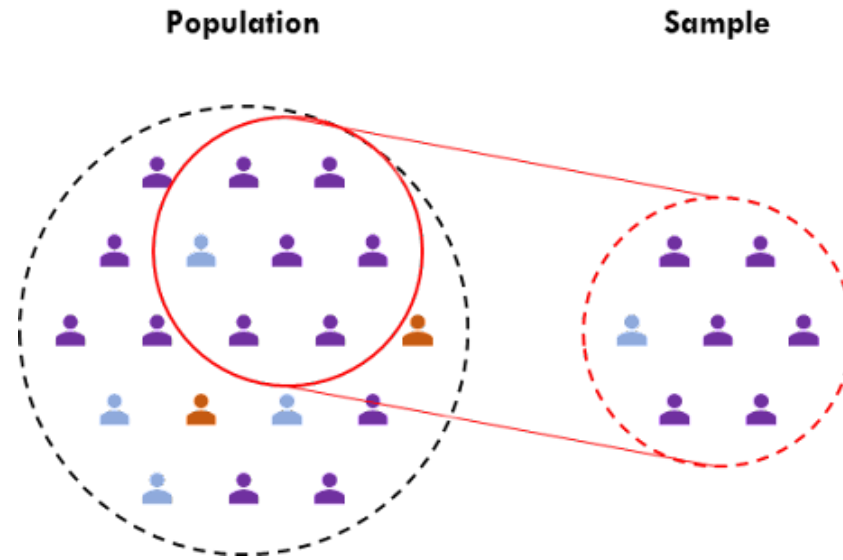
- `eight_rows_covid['province_of_onset'] == 'กรุงเทพมหานคร'`
- ผลลัพธ์จะได้
- 0 True
- 1 True
- 2 False
- 3 True
- 4 True
- 5 True
- 6 True
- 7 False
- Name: province_of_onset, dtype: bool



นำ list ของ logical expression ที่สร้างมาใช้งาน

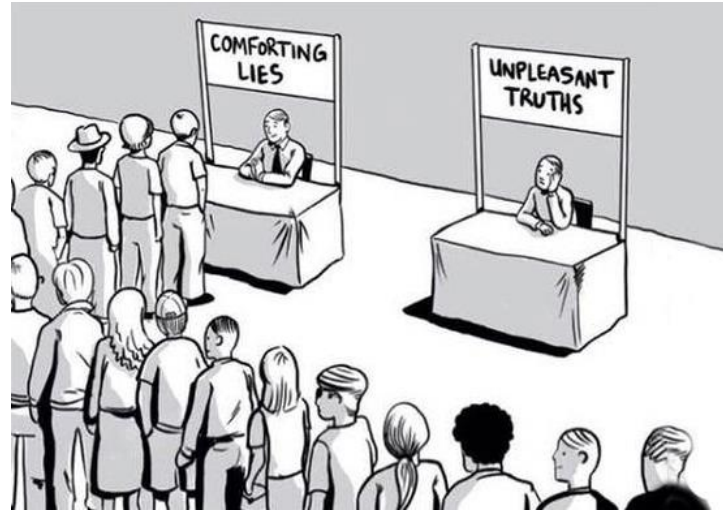
- ซึ่งเมื่อนำมาใช้งานเลือกข้อมูลเฉพาะแถวที่ต้องการแบบ advanced คือการใช้ logic query ในการเลือกข้อมูลนั่นเอง
- `eight_rows_covid[eight_rows_covid['province_of_onset'] == 'กรุงเทพมหานคร']`
- ผลลัพธ์จะได้ข้อมูลทุกแถวที่มีข้อมูลในคอลัมน์ province_of_onset เป็น กรุงเทพมหานคร คือแถวที่ 0, 1, 3, 4, 5, 6

Bias ในข้อมูล



- bias คืออคติหรือความลำเอียงที่อาจแฝงอยู่ในข้อมูล ซึ่งอาจเกิดจากวิธีการเก็บข้อมูล การออกแบบแบบสำรวจ หรือความผิดพลาดในการป้อนข้อมูล

สำรวจ Bias ในชุดข้อมูล



- ใช้ชุดข้อมูลในโลกแห่งความเป็นจริงที่มี bias หรือข้อจำกัดที่รู้จักเป็นตัวอย่างในชั้นเรียน
- ให้นักเรียนวิเคราะห์ชุดข้อมูลเพื่อระบุ bias ที่อาจเกิดขึ้น เช่น การมีตัวแทนของกลุ่มบางกลุ่มน้อยเกินไปหรือมากเกินไป

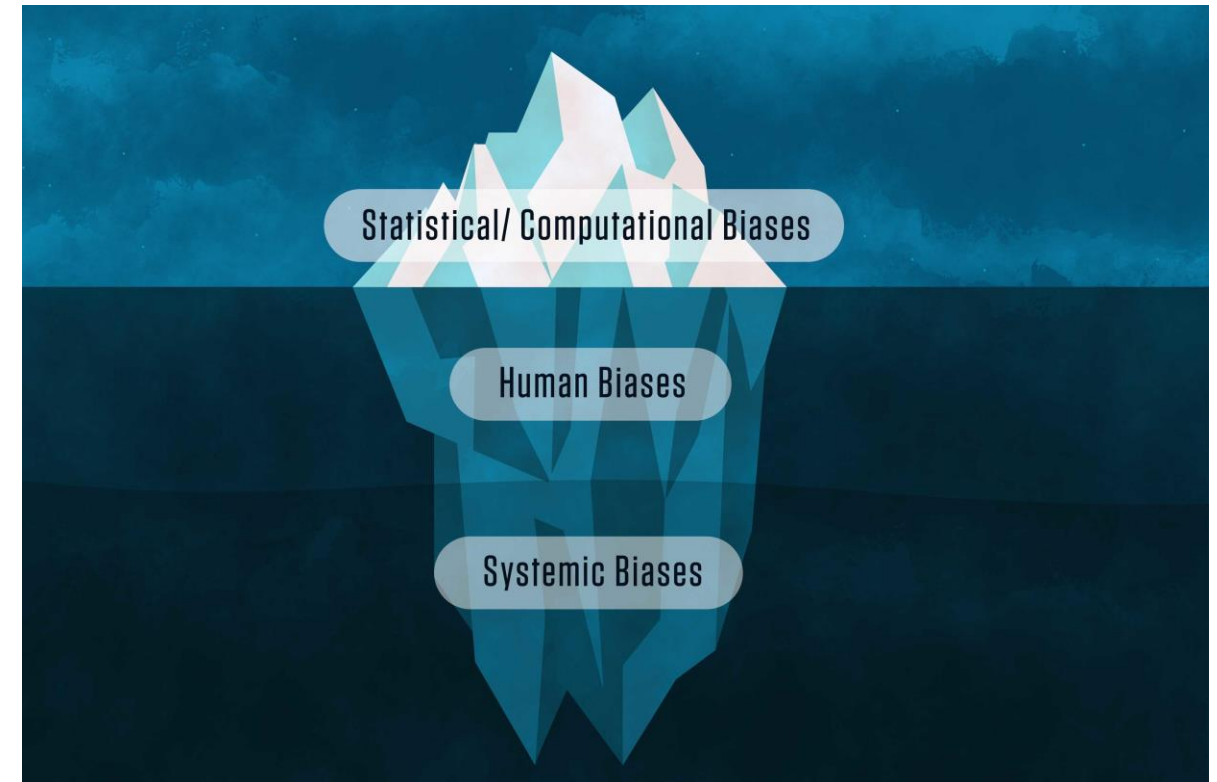
เทคนิคการประมวลผลข้อมูล



- ครอบคลุมเทคนิคการประมวลผลข้อมูลที่สามารถช่วยลด bias เช่น
- การจัดการกับข้อมูลที่ขาดหายไป
- การจัดการกับค่าผิดปกติ (outliers)
- การสร้างความสมดุลของการกระจายตัวของคลาส

Algorithmic Bias

- bias สามารถถูกขยายผลหรือเกิดขึ้นซ้ำๆ ผ่านอัลกอริทึม และโมเดลการเรียนรู้ของเครื่อง
- ตัวอย่างของ algorithmic bias เช่น ระบบการทำนายอาชญากรรมที่ลำเอียงหรืออัลกอริทึมการรับสมัครงานที่เลือกปฏิบัติตามเพศหรือเชื้อชาติ
- นัยยะด้านจริยธรรมของการใช้อัลกอริทึมที่ลำเอียง และความสำคัญของความเป็นธรรมและการไม่เลือกปฏิบัติในการตัดสินใจเชิงอัลกอริทึม





กรณีศึกษาและตัวอย่างจากโลกแห่งความเป็นจริง

- นำเสนอกรณีศึกษาและตัวอย่างจริงของอัลกอริทึมที่ลำเอียงหรือระบบที่ขับเคลื่อนด้วยข้อมูลที่มีผลกระทบเชิงลบ
- วิเคราะห์กรณีเหล่านี้ในการอภิปรายในชั้นเรียนเพื่อระบุที่มาของ bias นัยยะด้านจริยธรรม และวิธีแก้ปัญหาที่อาจเกิดขึ้น
- กระตุ้นให้นักเรียนคิดไตร่ตรองถึงบทเรียนที่ได้เรียนรู้จากกรณีศึกษาเหล่านี้ และนำมาประยุกต์ใช้กับแนวปฏิบัติในการเขียนโปรแกรมของตนเอง
- <https://www.prolific.com/resources/shocking-ai-bias>



Homework class period 6

- (ให้ใช้เฉพาะที่อาจารย์สอนไปแล้วในวิชานี้)
- คำนวณ อายุเฉลี่ย ของผู้หญิง และผู้ชาย ของข้อมูลทั้งหมด
- คำนวณ อายุเฉลี่ย ของผู้หญิง และผู้ชาย ของผู้ป่วยในจังหวัดขอนแก่น
- หาจำนวนผู้ป่วยที่เป็นคน "คนต่างชาติเดินทางมาจากต่างประเทศ"