

Modern Data Science Methods for Educational Research

R for Data Analysis in Educational Research

อ.ดร.ประภาศิริ รัชประภาพรกุล

ภาควิชาวิจัยและจิตวิทยาการศึกษา
คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

February 11, 2023

1. มโนทัศน์พื้นฐาน
○○○○○○○○○○

2. เครื่องมือสำรวจข้อมูลเบื้องต้น
○○○

3. Tidying Data
○○○

– Reshaping data
○○○○○○○○○○

– Separate()/Unite()
○○○○○○○○

4. My First Tidy Data
○○



KRURU
EDU @ CHULALONGKORN

1. มโนทัศน์พื้นฐาน

กิจกรรม : Messydata

1. ดาวน์โหลดไฟล์ข้อมูล messydata.xlsx
2. นำไฟล์ข้อมูล messydata.xlsx เข้าโปรแกรม R

```
1 dat <- read_excel("messydata.xlsx", na="-")
```

3. ลองสำรวจข้อมูลข้างต้นแล้วตอบคำถาม

วัตถุประสงค์ของการวิจัยคือ เพื่อเปรียบเทียบทักษะการแก้ปัญหาของนักเรียน
ภายหลังจากได้รับการจัดการเรียนรู้ด้วย วิธีการสอนแบบบรรยาย (Lecture) กับ
วิธีการสอนแบบใช้ปัญหาเป็นฐาน (PBL)

ท่านคิดว่า ข้อมูล messydata.xlsx

มีความพร้อมที่จะนำไปวิเคราะห์เพื่อตอบวัตถุประสงค์ดังกล่าวหรือไม่?

กิจกรรม : Messydata

```
# A tibble: 6 x 5
  ...1 Lecture.pre PBL.pre Lecture.post PBL.post
  <chr>          <dbl>   <dbl>         <dbl>   <dbl>
1 Ancient One      20      NA           45      NA
2 Adam Warlock     16      NA           34      NA
3 Captain America  NA      18           NA      67
4 Colossus         NA      25           NA      93
5 Captain Marvel   13      NA           50      NA
6 Diablo          NA      17           NA      71
```

กิจกรรม : การเปลี่ยนชื่อคอลัมน์

นอกจากการเปลี่ยนชื่อคอลัมน์ผ่านอาร์กิวเมนต์ `col_names` แล้ว
ผู้วิเคราะห์ยังสามารถเลือกเปลี่ยนชื่อคอลัมน์เป็นรายตัวได้ โดยใช้ฟังก์ชัน `names()`
ดังนี้

```
1 names(dat)
```

```
[1] "...1"          "Lecture.pre"    "PBL.pre"       "Lecture.p
```

```
1 names(dat)[1]<-"id"
```

```
2 names(dat)
```

```
[1] "id"            "Lecture.pre"   "PBL.pre"       "Lecture.p
```

กิจกรรม : การเปลี่ยนชื่อคอลัมน์

```
1 head(dat)
```

```
# A tibble: 6 x 5
```

| | id <chr> | Lecture.pre <dbl> | PBL.pre <dbl> | Lecture.post <dbl> | PBL.post <dbl> |
|---|-----------------|----------------------|------------------|-----------------------|-------------------|
| 1 | Ancient One | 20 | NA | 45 | NA |
| 2 | Adam Warlock | 16 | NA | 34 | NA |
| 3 | Captain America | NA | 18 | NA | 67 |
| 4 | Colossus | NA | 25 | NA | 93 |
| 5 | Captain Marvel | 13 | NA | 50 | NA |
| 6 | Diablo | NA | 17 | NA | 71 |

ภาพรวมของการเตรียมข้อมูล

- ▶ Tidying Data
- ▶ Manipulating Data
- ▶ Missing Values Analysis and Imputation
- ▶ Outlier Detection and Handling
- ▶ Data Reduction
- ▶ Feature Selection
- ▶ ...

การเตรียมข้อมูลภายใต้กระบวนการวิเคราะห์ข้อมูล

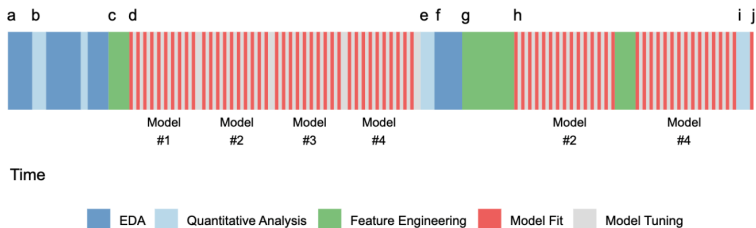


Figure 1: ที่มา : Max Khun, & Kjell Johnson (2019)

ตัวอย่าง : โมเดลทำนายการได้ขึ้นเงินเดือนของพนักงาน

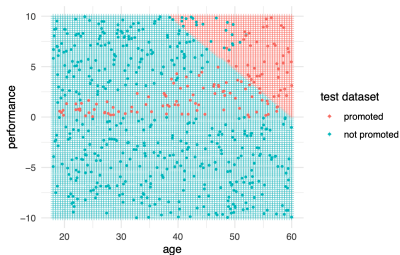


Figure 2: logistic regression ที่ไม่ได้ทำ feature engineering

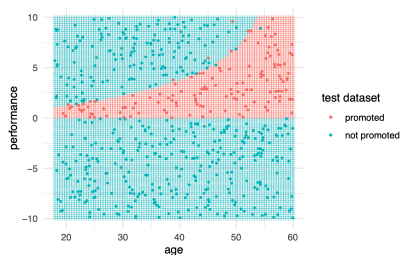


Figure 3: logistic regression ที่มีการทำ feature engineering

Tidy Data



Tidy Data

หน่วยข้อมูล

ตัวแปร

| id | Student | Teaching Method | Problem Solving (PreTest) | Problem Solving (PostTest) | Gain Score |
|----|---------|-----------------|---------------------------|----------------------------|------------|
| 1 | บุญมี | Lecture | 20 | 45 | 25 |
| 2 | บุญมาก | Lecture | 16 | 34 | 18 |
| 3 | บุญกัน | PBL | 18 | 67 | 49 |
| 4 | บุญถึง | PBL | 25 | 93 | 68 |
| 5 | บุญธรรม | Lecture | 13 | 50 | 37 |
| 6 | บุญใหญ่ | PBL | 17 | 71 | 54 |
| 7 | บุญนิก | Lecture | 14 | 48 | 34 |
| 8 | บุญนา | PBL | 20 | 83 | 63 |
| 9 | บุญยะระ | PBL | 28 | 75 | 47 |
| 10 | บุญเม่ง | PBL | 211 | 73 | -138 |
| 11 | บุญสี | Lecture | 22 | 32 | 10 |
| 12 | บุญระง | PBL | 17 | 68 | 51 |
| 13 | บุญถึง | Lecture | 14 | 70 | 56 |
| 14 | บุญจึง | Lecture | 15 | 640 | 625 |
| 15 | บุญเดิม | Lecture | 19 | 55 | 36 |
| 16 | บุญยศ | PBL | 15 | 82 | 67 |
| 17 | บุญชน | Lecture | 18 | 59 | 41 |
| 18 | บุญเม่ง | PBL | 12 | 68 | 56 |
| 19 | บุญสืบ | Lecture | 8 | 62 | 54 |
| 20 | บุญหลือ | PBL | 25 | 87 | 62 |

Figure 4: ที่มา : สิวะโชติ ศรีสุทธียากร (2564)

2. เครื่องมือสำรวจข้อมูลเบื้องต้น

ฟังก์ชันพื้นฐานใน R สำหรับสำรวจข้อมูล

- ▶ `str()` - ใช้สำรวจโครงสร้างโดยรวมของชุดข้อมูล
- ▶ `head()` และ `tail()` - ใช้เรียกดูตารางข้อมูลส่วนหัว และส่วนท้าย
- ▶ `names()` - ใช้เรียกดูชื่อคอลัมน์ในชุดข้อมูล
และยังสามารถใช้เปลี่ยนชื่อคอลัมน์ได้ด้วย
- ▶ `summary()` - เรียกดูค่าสถิติเบื้องต้นของตัวแปรแต่ละตัวภายในชุดข้อมูล

กิจกรรม : การสำรวจข้อมูล

จากชุดข้อมูล messydata.xlsx ขอให้ผู้เรียน

- ▶ ทดลองใช้ฟังก์ชันพื้นฐานข้างต้น สำรวจชุดข้อมูลผลลัพธ์ที่ได้เป็นอย่างไร
- ▶ ชุดข้อมูลที่นำเข้ามาจาก messydata.xlsx มีคุณสมบัติ tidy data หรือไม่ อย่างไร
- ▶ ผู้เรียนคิดว่า tidy data ของชุดข้อมูล messydata.xlsx ควรมีหน้าตาเป็นอย่างไร

โปรด upload รูป tidy data ของท่านที่นี่ —> upload เลย หรือ scan QR code

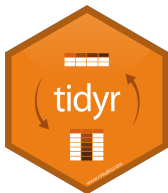


3. Tidying Data

Tidying Data

- ▶ Reshaping data
- ▶ Splitting and Combining column
- ▶ Removing duplicate data

Tidyr package



```
1 install.packages("tidyr")
2 library(tidyr)
```

– Reshaping data

Long and Wide Format data

▶ ชุดข้อมูล messydata.xlsx เป็นแบบ long หรือ wide format ?

| ระดับการศึกษา | เพศ | จำนวนนักศึกษา |
|----------------------|------|---------------|
| ต่ำกว่าปริญญาตรี | ชาย | 198,086 |
| ต่ำกว่าปริญญาตรี | หญิง | 151,587 |
| ปริญญาตรี | ชาย | 572,497 |
| ปริญญาตรี | หญิง | 889,112 |
| ประกาศนียบัตรบัณฑิต | ชาย | 3,410 |
| ประกาศนียบัตรบัณฑิต | หญิง | 7,133 |
| ปริญญาโท | ชาย | 39,849 |
| ปริญญาโท | หญิง | 54,215 |
| ประกาศนียบัตรชั้นสูง | ชาย | 680 |
| ประกาศนียบัตรชั้นสูง | หญิง | 1,081 |
| ปริญญาเอก | ชาย | 11,375 |
| ปริญญาเอก | หญิง | 12,027 |

| ระดับการศึกษา | ชาย | หญิง |
|----------------------|---------|---------|
| ต่ำกว่าปริญญาตรี | 198,086 | 151,587 |
| ปริญญาตรี | 572,497 | 889,112 |
| ประกาศนียบัตรบัณฑิต | 3,410 | 7,133 |
| ปริญญาโท | 39,849 | 54,215 |
| ประกาศนียบัตรชั้นสูง | 680 | 1,081 |
| ปริญญาเอก | 11,375 | 12,027 |

Figure 6: wide format data ที่มา : สิวะโชติ ศรีสุทธียากร (2564)

Figure 5: long format data ที่มา : สิวะโชติ ศรีสุทธียากร (2564)

Reshaping data: wide → long format



```
1 gather(data, ..., key, value)
```

- ▶ data ชุดข้อมูลประเภท wide format
- ▶ ... คอลัมน์ทั้งหมดใน dat ต้องการยุบมาไว้ภายใต้คอลัมน์ใหม่
- ▶ key ชื่อคอลัมน์ใหม่สำหรับเก็บ header หรือชื่อคอลัมน์ที่อยู่ใน ...
- ▶ value ชื่อคอลัมน์ใหม่สำหรับเก็บข้อมูลที่อยู่ภายใต้ ...

Reshaping data: wide → long format using gather()

```
1 # messy data
2 head(dat)
```

```
# A tibble: 6 x 5
```

| | id | Lecture.pre | PBL.pre | Lecture.post | PBL.post |
|---|-----------------|-------------|---------|--------------|----------|
| | <chr> | <dbl> | <dbl> | <dbl> | <dbl> |
| 1 | Ancient One | 20 | NA | 45 | NA |
| 2 | Adam Warlock | 16 | NA | 34 | NA |
| 3 | Captain America | NA | 18 | NA | 67 |
| 4 | Colossus | NA | 25 | NA | 93 |
| 5 | Captain Marvel | 13 | NA | 50 | NA |
| 6 | Diablo | NA | 17 | NA | 71 |

Reshaping data: wide → long format using gather()

```
1 long_dat <- gather(dat,  
2                     Lecture.pre,  
3                     PBL.pre,  
4                     Lecture.post,  
5                     PBL.post,  
6     key = "method_time",  
7     value = "score")
```

Reshaping data: wide → long format using gather()

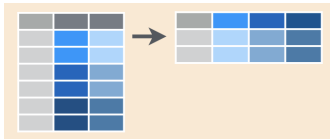
```
1 head(long_dat)
```

```
# A tibble: 6 x 3
```

| | id <chr> | method_time <chr> | score <dbl> |
|---|-----------------|----------------------|----------------|
| 1 | Ancient One | Lecture.pre | 20 |
| 2 | Adam Warlock | Lecture.pre | 16 |
| 3 | Captain America | Lecture.pre | NA |
| 4 | Colossus | Lecture.pre | NA |
| 5 | Captain Marvel | Lecture.pre | 13 |
| 6 | Diablo | Lecture.pre | NA |

► ข้อมูลข้างต้นเรียกว่า Tidy Data ได้แล้วหรือไม่ ? เพราะเหตุใด ?

Reshaping data: long → wide format using spread()



1 `spread(data, key, value)`

- ▶ `data` คือ dataframe ที่ต้องการแปลงจาก long เป็น wide format
- ▶ `key` คือ คอลัมน์ใน data ที่ต้องการ expand ไปอยู่บน header ของตาราง
- ▶ `value` คือ คอลัมน์ใน data ที่ต้องการย้ายไปอยู่ภายใต้ header ใหม่

Reshaping data: long → wide format using spread()

```
1 wide_dat <- spread(long_dat,  
2   key = "method_time",  
3   value = "score")
```

Reshaping data: long → wide format using spread()

```
1 head(wide_dat)
```

```
# A tibble: 6 x 5
```

| | id <chr> | Lecture.post <dbl> | Lecture.pre <dbl> | PBL.post <dbl> | PBL.pre <dbl> |
|---|-----------------|-----------------------|----------------------|-------------------|------------------|
| 1 | Adam Warlock | 34 | 16 | NA | NA |
| 2 | Ancient One | 45 | 20 | NA | NA |
| 3 | Captain America | NA | NA | 67 | 18 |
| 4 | Captain Marvel | 50 | 13 | NA | NA |
| 5 | Colossus | NA | NA | 93 | 25 |
| 6 | Deadpool | NA | NA | 83 | 20 |

- Separate()/Unite()

Separate Column using separate()

ชุดข้อมูล `long_dat` ยังไม่ใช่ tidy data ปัญหาหนึ่งที่เราพบคือคอลัมน์ `method_time` มีข้อมูลทั้งของวิธีการสอน และเวลาที่วัดค่าสังเกต รวมกันอยู่

```
1 separate(data, col, into, sep)
```

- ▶ `data` คือชุดข้อมูลที่ต้องการแยกคอลัมน์
- ▶ `col` คือคอลัมน์ที่ต้องการแยกข้อมูลออกจากกัน
- ▶ `into` ชื่อคอลัมน์ใหม่สำหรับเก็บข้อมูลที่แยกออกจากกัน
- ▶ `sep` คือตัวคั่นหรือเงื่อนไขที่ใช้สำหรับแยกข้อมูลใน `col`

Separate Column using separate()

ทดลองแยกคอลัมน์ method.time ในชุดข้อมูล long_dat

```
1 separated_dat <- separate(long_dat,  
2                             col = "method_time",  
3                             into = c("method","time"),  
4                             sep=".[.]")
```

Separate Column using separate()

```
1 head(separated_dat)
```

```
# A tibble: 6 x 4
```

| | id <chr> | method <chr> | time <chr> | score <dbl> |
|---|-----------------|-----------------|---------------|----------------|
| 1 | Ancient One | Lecture | pre | 20 |
| 2 | Adam Warlock | Lecture | pre | 16 |
| 3 | Captain America | Lecture | pre | NA |
| 4 | Colossus | Lecture | pre | NA |
| 5 | Captain Marvel | Lecture | pre | 13 |
| 6 | Diablo | Lecture | pre | NA |

► ชุดข้อมูล separated_dat เป็น tidy data แล้วหรือไม่ ? เพราะเหตุใด?

Combing Column using unite()

นอกจากแยกคอลัมน์แล้วยังสามารถยุบรวมคอลัมน์เข้าด้วยกันได้ด้วย

```
1 unite(data, ..., col, sep)
```

- ▶ data คือชุดข้อมูลที่ต้องการยุบรวมคอลัมน์เข้าด้วยกัน
- ▶ ... คือคอลัมน์ใน dat ที่ต้องการยุบรวมคอลัมน์เข้าด้วยกัน
- ▶ col คือชื่อคอลัมน์ใหม่ภายหลังยุบรวมคอลัมน์
- ▶ sep คือตัวคั่นระหว่างข้อมูลใหม่ที่ยุบรวมกัน

Combining Column using unite()

```
1 combine_dat <- unite(separated_dat,  
2   method, time,  
3   col = "method.time",  
4   sep = "-")
```


Combing Column using unite()

```
1 head(combine_dat)
```

```
# A tibble: 6 x 3
```

| | id <chr> | method.time <chr> | score <dbl> |
|---|-----------------|----------------------|----------------|
| 1 | Ancient One | Lecture-pre | 20 |
| 2 | Adam Warlock | Lecture-pre | 16 |
| 3 | Captain America | Lecture-pre | NA |
| 4 | Colossus | Lecture-pre | NA |
| 5 | Captain Marvel | Lecture-pre | 13 |
| 6 | Diablo | Lecture-pre | NA |

4. My First Tidy Data

กิจกรรม : My First Tidy data

ขอให้ผู้เรียนดำเนินการจัดระเบียบข้อมูล `separated_dat` ให้เป็น Tidy data

```
# A tibble: 6 x 4
```

| | id <chr> | method <chr> | post <dbl> | pre <dbl> |
|---|-----------------|-----------------|---------------|--------------|
| 1 | Adam Warlock | Lecture | 34 | 16 |
| 2 | Adam Warlock | PBL | NA | NA |
| 3 | Ancient One | Lecture | 45 | 20 |
| 4 | Ancient One | PBL | NA | NA |
| 5 | Captain America | Lecture | NA | NA |
| 6 | Captain America | PBL | 67 | 18 |