Sprint 04

Tags	Data Transformation
Status	Done

🎱 Code

The Most Important Skill

Core Data Analyst Skills

การเล่นแร่แปรธาตุตัว data frame ให้อยู่ในรูปแบบ format แบบที่เราต้องการ ใน R เราใช้ library `dplyr` i.e. ในตระกูล `tidyverse` เพื่อปรับหน้าตา data frame แบบที่เราต้องการ

5 functions หลักของ dplyr ประกอบด้วย

- `select()` : เลือกคอลัมน์
- `filter()` : กรองข้อมูลด้วยเงื่อนไข
- `mutate()` : สร้างคอลัมน์ใหม่
- `arrange()` : เรียงข้อมูล
- `summarise()` อันนี้เขียนได้สองแบบ `summarize()` : สรุปผลสถิติ
- `group_by()` : จับกลุ่มข้อมูล

Typical workflow ของ data analyst ที่เขียน `R` คือ

- ดึงข้อมูลจาก SQL databases หรือ data format ต่างๆเข้าสู่ R
- เขียน `dplyr` เพื่อจัดการ data frame จะ merge, join, union, transform ทำได้หมดเลย
- ส่ง transformed data ให้ users ของเรา (e.g. `csv`, `excel`, `json`) หรือส่งไปให้ software อื่นๆใช้งานต่อ เช่น Power BI, Tableau, Google Sheets, Data Studio

▼ Data Transformation

What is Data Transformation?

เป็นการปรับเปลี่ยนหน้าตาข้อมูลจากรูปแบบหนึ่งไปเป็นอีกรูปแบบหนึ่งที่เหมาะสมกว่า

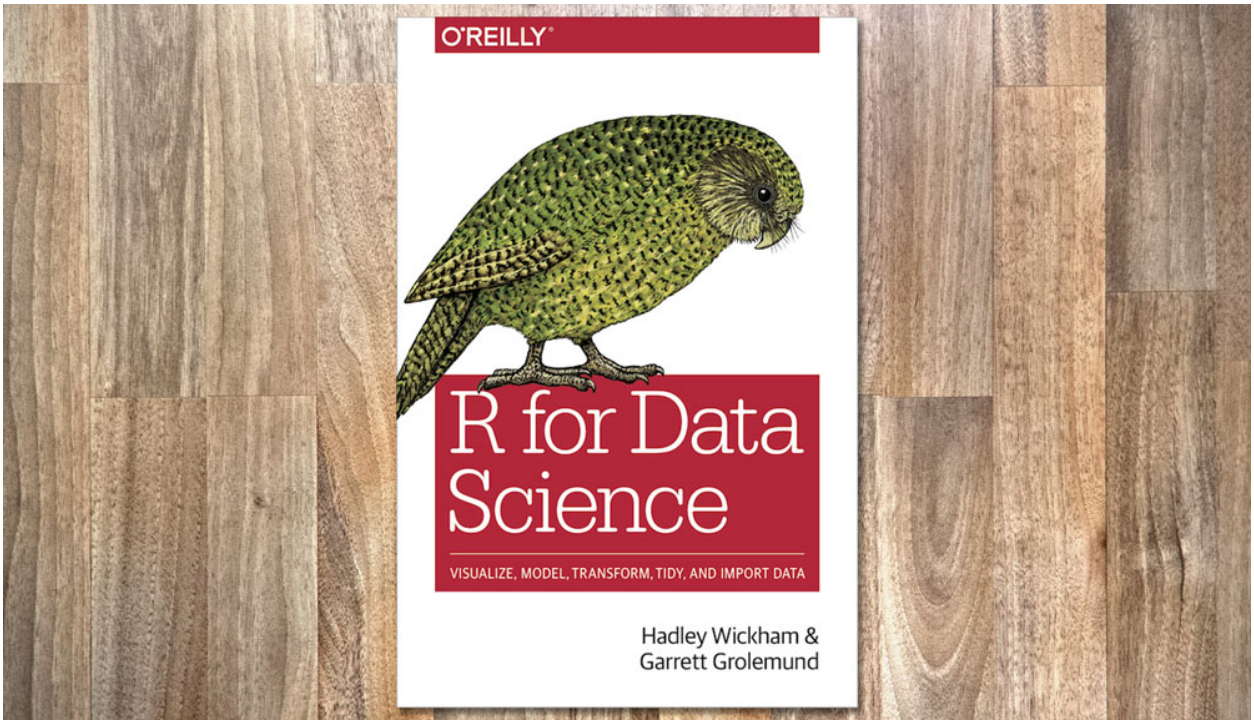
Export CSV File

🌵 ในวีดีโอแอดใช้ฟังก์ชัน `write.csv()` เป็น based R function นะครับ ไม่จำเป็นต้องโหลด library อะไรเพิ่ม แต่ถ้าใครอยากใช้ฟังก์ชันใหม่ๆอย่าง `write_csv()` ต้องโหลด library `readr` ก่อนนะครับ

จริงๆผลที่ได้ไม่ต่างกันเลย แต่ `write_csv()` หรือ `read_csv()` ที่อยู่ใน readr จะทำงานได้เร็วกว่านิดนึง และมีพฤติกรรมที่ไม่แปลกเหมือน based R functions เก่าๆ

Tip - วิธีสังเกตว่า functions ไหนเป็น **based R** เก่า หรือ **modern R** ให้ดูที่ตัว `.` ได้เลย ชื่อฟังก์ชันใหม่ๆจะใช้ `_` แทน `.` หมดแล้ว (เขียนแบบ snake case เช่น `hello_world()` เป็นต้น)

Data Transformation



🌵 R for Data Science เขียนโดย Hadley Wickham & Garrett Grolemund ส่วนตัวแอดยกเล่มนี้เป็นหนังสือ R ที่ดีที่สุดเลย และ Hadley ใจดีให้เราอ่านฟรีออนไลน์ด้วย สุดยอด!

นักเรียนที่เรียน Data Transformation 101 จบแล้ว อ่านหนังสือ **R for Data Science** บทที่ 5 ต่อได้ที่ https://r4ds.had.co.nz/transform.html

- อ่านหนังสือ
- เขียนสรุปใน Notion
- แชร์ในห้อง #r หรือ #notion ใน discord ได้เลยนะครับ

Note - ตอนอ่านบทที่ 5 แอดยังไม่ได้สอน `ggplot` ในการทำ data visualization นักเรียนสามารถข้ามเนื้อหาบางส่วนไปก่อนได้เลยนะครับ รอเรียน sprint ต่อไปได้เลย 😊

▼ Windows Terminal

command line คือคำสั่งที่เราสามารถใช้จัดการไฟล์และโฟลเดอร์ในคอมพิวเตอร์เราได้ (แบบไม่ต้องใช้เม้าส์เลย)

Basic commands ที่ data analyst ควรรู้จัก

- `echo`
- `pwd`
- `ls`
- `cd`
- `mkdir`
- `rmdir`
- `del`
- `move`
- `rename`

Echo (Print Text)

🌵 `echo` ใช้แสดงผลข้อความ text, string ใน terminal เหมือนเราเขียน `print()` ในภาษา R / Python เลย

```
input : echo Hello World
output : Hello World

input : set my_name=Folk
input : echo %my_name%
output : Folk

input : echo %time%
output : 10:51:01.77

input : echo %date%
output : 2023-11-21
```

Note : cls is clear data on window

Quick Start

`cd` ย่อยมาจาก change directory เป็นคำสั่งย้าย path หรือที่อยู่ไฟล์

`mkdir` ย่อมาจาก make directory เป็นคำสั่งสร้าง folder ใหม่

Note : directory = folder

`dir` เป็นการเรียกดูไฟล์ทั้งหมดที่อยู่ใน folder นั้น.

`type` m เป็นคำสั่งใช้ดูรายละเอียดภายในไฟล์ที่เลือก

```
input : cd Desktop

input : mkdir (name_folder)

input : dir

input : type (name_folder)
```

Change Directory

cd .. เป็นการย้าย folder ไปย้าย folder ก่อนหน้า 1 step

cd ..\.. เป็นการย้าย folder ไปย้าย folder ก่อนหน้า 2 step

cd name_folder\name_folder เป็นการย้าย folder ไปข้างหน้า

```
C:\Users\user\Desktop\Conicle>cd ..
output : C:\Users\user\Desktop>

C:\Users\user\Desktop\Conicle>cd ..\..
output : C:\Users\user>

C:\Users\user\Desktop>cd Conicle\Data
output : C:\Users\user\Desktop\Conicle\Data>
```

Create Text File

- การใช้ echo ในการสร้างไฟล์
- เขียนทับข้อมูลเดิม(เขียนเหมือนเดิม แค่เปลี่ยนข้อมูลที่จะใส่ใหม่)
- เพิ่มข้อมูลในบรรทัดต่อไปของไฟล์ (>>)

```
#Create File
C:\Users\user\Desktop\Conicle\Data>echo hello > hello.txt

#Create File ทับไฟล์เดิม
C:\Users\user\Desktop\Conicle\Data>echo I love a course > hel

#เพิ่มข้อมูลในไฟล์โดยไม่เขียนทับ
C:\Users\user\Desktop\Conicle\Data>echo I love myself >> hell
```

Delete File and Folder

- `dir /b` เป็นการเรียกดูเฉพาะแค่ชื่อไฟล์ใน Folder เท่านั้น
- `help dir` ดูว่าฟังก์ชัน dir ทำอะไรได้บ้าง
- `del (name_file)` เป็นการลบไฟล์
- `del *` เป็นการลบไฟล์ทั้งหมด

- `rmdir (name_folder)` เป็นการลบ folder # ลบได้แค่ folder ที่ว่างเท่านั้น
- `rmdir /s` สามารถลบ folder ที่มีข้อมูลอยู่ได้

Rename Files and Folder

- rename (old_name) (new_name) #rename or ren

```
C:\Users\user\Desktop\Conicle\Data>rename hello.txt this_is_c
C:\Users\user\Desktop\Conicle\Data>rename Test CoolFolder

C:\Users\user\Desktop\Conicle\Data>dir /b
CoolFolder
this_is_cool.txt
```

Find String

- findstr “(text)” (name_file) การหาข้อมูล
- findstr /N การหาข้อมูลโดยให้แสดงเลขแถวด้วย

```
C:\Users\user\Desktop\Conicle\Data>findstr /N "l" city.txt
3:Seoul
```

Download Data with Curl

🌵 Command `curl` จะขึ้นอยู่กับเวอร์ชันของ terminal (powershell) ที่เราใช้ด้วยนะครับ Windows บางเครื่องอาจจะหา `curl` ไม่เจอ

แอดแนะนำว่าลองโหลด PowerShell เวอร์ชัน 7.0+ จะมี command `Invoke-WebRequest` ใช้งานได้เหมือนกับ `curl` เลย

Reference: Invoke-WebRequest (Microsoft.PowerShell.Utility) - PowerShell | Microsoft Docs

- curl (URL) —output (new_name_file) การดึงข้อมูลจาก web_site

Ping

เช็คว่า web_bowser ที่เราต้องการเข้าถึงมันเข้าถึงได้จริงรึป่าว

- ping (name_website/URL)

Move Files

- move (name_file) (name_folder) ย้าย file ไปยัง folder ใน path เดียวกัน
- move (name_folder)\(name_file) . ย้าย file ไปยัว path ปัจจุบัน

Run File R

- Rscript (name_file) การรันไฟล์นามสกุล R