Data Transformation: Part III

Data Structure Conversion: Wide -> Long data

Cultivate Learning Innovation Lab Workshop
August 31, 2020 | Monday | 2:00 - 3:30 p.m.

[황보 민] Min Hwangbo

PhD Candidate in Learning Sciences & Human Development
Graduate Certificate Recipient in Demographic Methods

Learning Agenda

- Why do we convert wide data to long data?
- Data transformation: `gather`
- Next steps: Challenge!

[Intro] Inspiration

- Land
  - *“The University of Washington & Cultivate Learning acknowledges that it sits on Indigenous Land, which touches the shared waters of all tribes and bands within the Duwamish, Suquamish, Tulalip, and Muckleshoot Tribes.”*
- People
  - Aimée Dechter | Affiliate Assistant Professor & Former Research Coordinator at Center for Studies in Demography & Ecology (CSDE)
  - [Chuck] Charles C. Lanfear | PhD Candidate & R Guru | 2020 Distinguished Teaching Award Recipient
  - [Jose] Jose Hernandez | Data Science Fellow & Research Staff @ eScience Institute
  - [Liz] Elizabeth Sanders | Associate Professor & Quantitative Researcher @ College of Education
  - [RStudio] Hadley Wickham & Garrett Grolemund | Authors of R for Data Science | RStudio Chief Scientist, Data Scientist & Statistician | Creators of RStudio
  - [이근열] Keun Yeol, Lee. | Professor in Busan National University & Qualitative/Dialect/Linguistics Researcher
  - [本橋智光] Motohashi, Tomomitsu. (2018). Maeshoritaizen data bunseki no tame no SQL/R/Python jissen technique (데이터 전처리 대전. 2019).
  - [Nicolas] Nicolas Pröllochs | Tenure-track Professor of Data Science in University of Giessen & Social Network Analysis / Text Mining Expert

Learning Agenda

- Why do we convert wide data to long data?
- Data transformation: `gather`
- Next steps: Challenge!

[Intro] Data Transformation?

- *“Process of **converting** data from one format or structure into another format or structure.... **fundamental** aspect of most **data integration** and **data management.**”*[1]
    - Data framing
    - Data manipulation
    - Data wrangling
    - Data management
- Data Transformation Steps
    - Data discovery: Where's my data?
    - Data mapping: Explore your data - What's in it? Using data transformation technique
    - Code generation: Cheatsheet & Stackflow & Pre-Built codes + Ctrl + C / Ctrl + V
    - Code execution: Ctrl + Enter
    - Data review: head(data), tail(data), View(data), missing data, quality of your data, etc...

[1] Source: Wikipedia

[Intro] Packages: dplyr & tidyr

- dplyr: “The grammar of data manipulation.”
  - **mutate()**: adds new variables that are functions of existing variables
  - **select()**: picks variables based on their names.
  - **filter()**: picks cases based on their values.
  - **summarise()**: reduces multiple values down to a single summary.
  - **arrange()**: changes the ordering of the rows.
- tidyr: “The goal of tidyr is to help you create tidy data.”
  - “Every **column** is **variable**.”
  - “Every **row** is an **observation**.”
  - “Every **cell** is a single **value**.”
- Cheatsheet!: Data Transformation Cheatsheet

[Intro] tidyverse Workflow

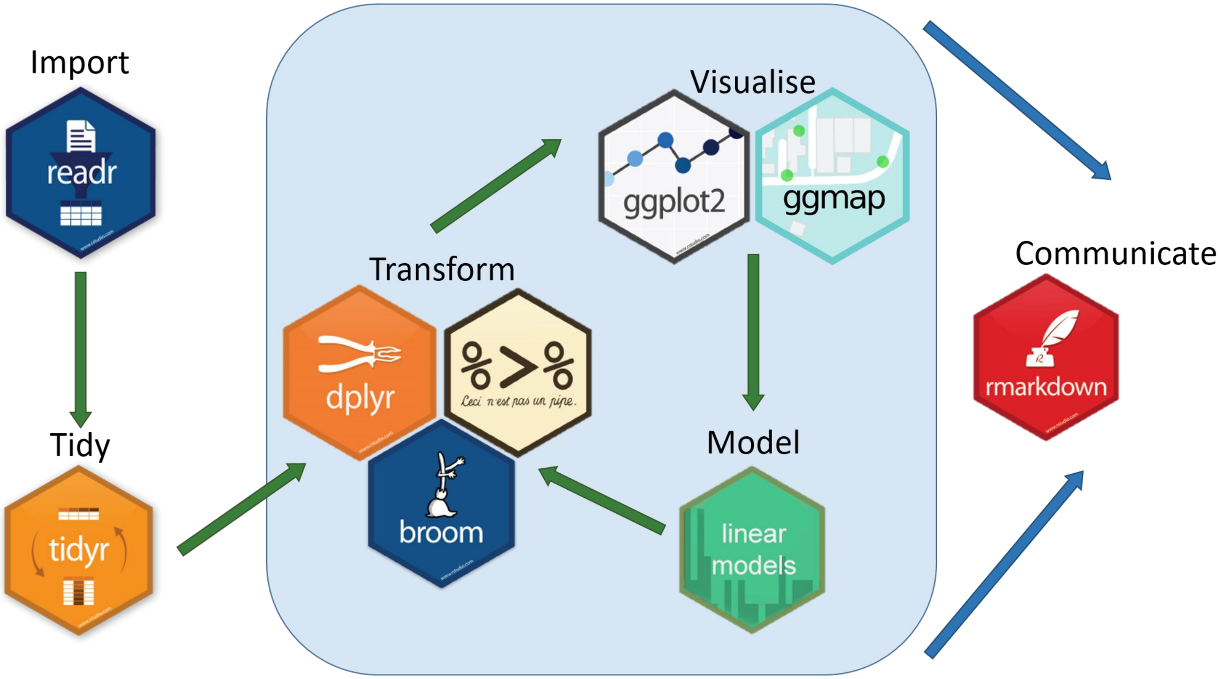


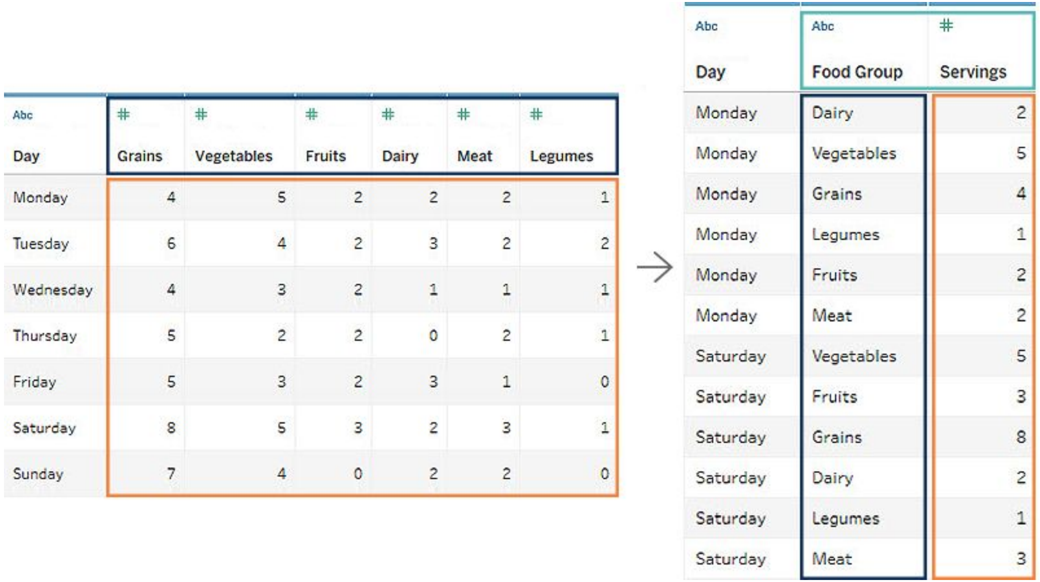
Image retrieved from: Glanz, Hunter. (2019). What is the Tidyverse? *Teach Data Science blog @* https://teachdatascience.com/tidyverse/

[Why] Wide Data -> Long data?

- *“**Wide data** has a column for each variable. Whereas **long format data** has a column for possible variable types & a column for the values of those variables.”* [1]
    - Wide format: Required for Multivariate analysis of variance (MANOVA) or repeated measures in SPSS.
    - Long format:
        - Mixed models (containing fixed & random effects) or most of the survival analysis
        - Also required for most data visualization softwares for survey data analysis (i.e. Qualtrics -> Tableau) - Common @ Cultivate Learning

[1] The analysis factor. (n.d.). *The Wide and Long Data Format for Repeated Measures Data.* https://www.theanalysisfactor.com/wide-and-long-data/

[Why] Wide Data -> Long data?[1]



Day	Grains	Vegetables	Fruits	Dairy	Meat	Legumes
Monday	4	5	2	2	2	1
Tuesday	6	4	2	3	2	2
Wednesday	4	3	2	1	1	1
Thursday	5	2	2	0	2	1
Friday	5	3	2	3	1	0
Saturday	8	5	3	2	3	1
Sunday	7	4	0	2	2	0

→

Day	Food Group	Servings
Monday	Dairy	2
Monday	Vegetables	5
Monday	Grains	4
Monday	Legumes	1
Monday	Fruits	2
Monday	Meat	2
Saturday	Vegetables	5
Saturday	Fruits	3
Saturday	Grains	8
Saturday	Dairy	2
Saturday	Legumes	1
Saturday	Meat	3

[1] Salesforce. (n.d.). *Get your data Tableau-teady.* https://www.tableau.com/learn/get-started/data-structure

Questions?

[Prep] Preparation

- Will not cover in this session (*See Session 2: Data Transformation*)
  - Package preparation
    - Installing packages
    - Loading packages
  - Data preparation
    - Loading data set
    - Framing data set
- **Will cover in this session**
  - Checking **`Null`** values
  - **`gather`** function to gather multiple columns into one column.
- Will not cover in this session (*See Session 2: Data Transformation*)
  - Saving it as a csv file

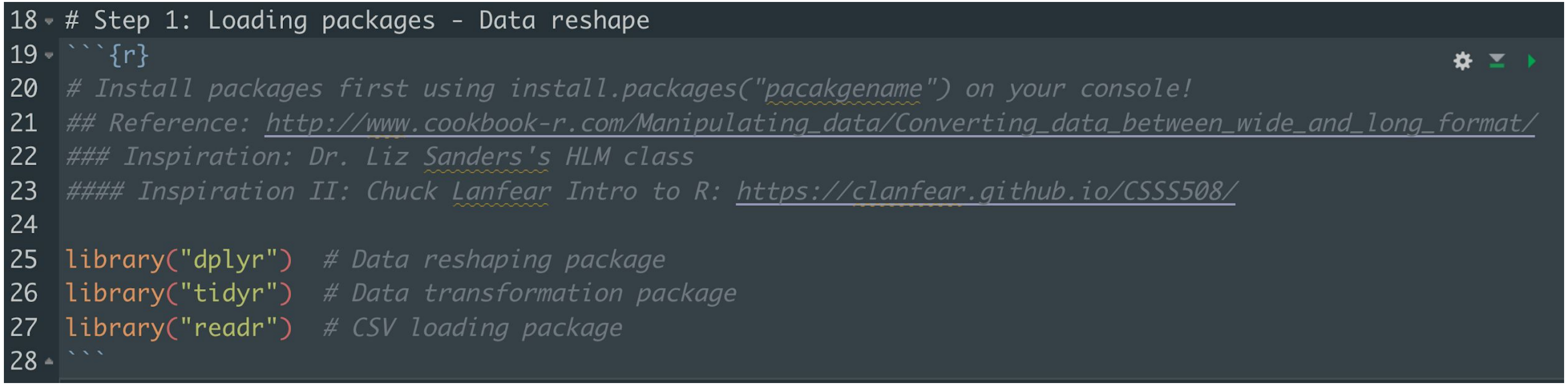
10 min Break

Coffee, Tea & Snack Time

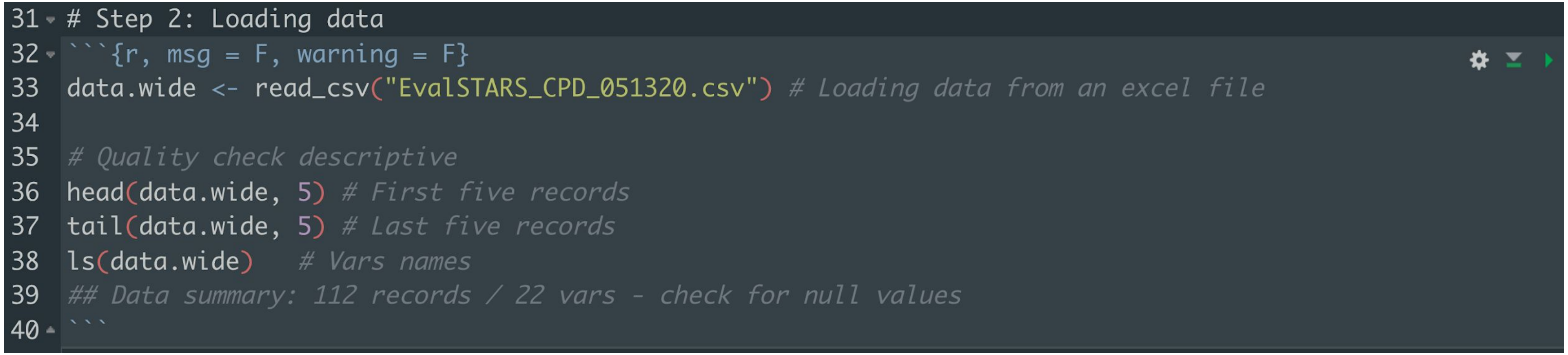
Learning Agenda

- Why do we convert wide data to long data?
- Data transformation: `gather`
- Next steps: Challenge!

[Prep] Step 1 & 2: Loading Packages & Data Set

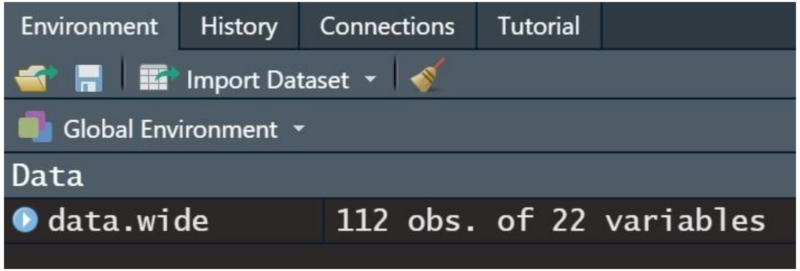


~~~
# Step 1: Loading packages - Data reshape
```{r}
# Install packages first using install.packages("pacakgename") on your console!
## Reference: http://www.cookbook-r.com/Manipulating_data/Converting_data_between_wide_and_long_format/
### Inspiration: Dr. Liz Sanders's HLM class
#### Inspiration II: Chuck Lanfear Intro to R: https://clanfear.github.io/CSSS508/

library("dplyr")  # Data reshaping package
library("tidyr")  # Data transformation package
library("readr")  # CSV loading package
```
~~~



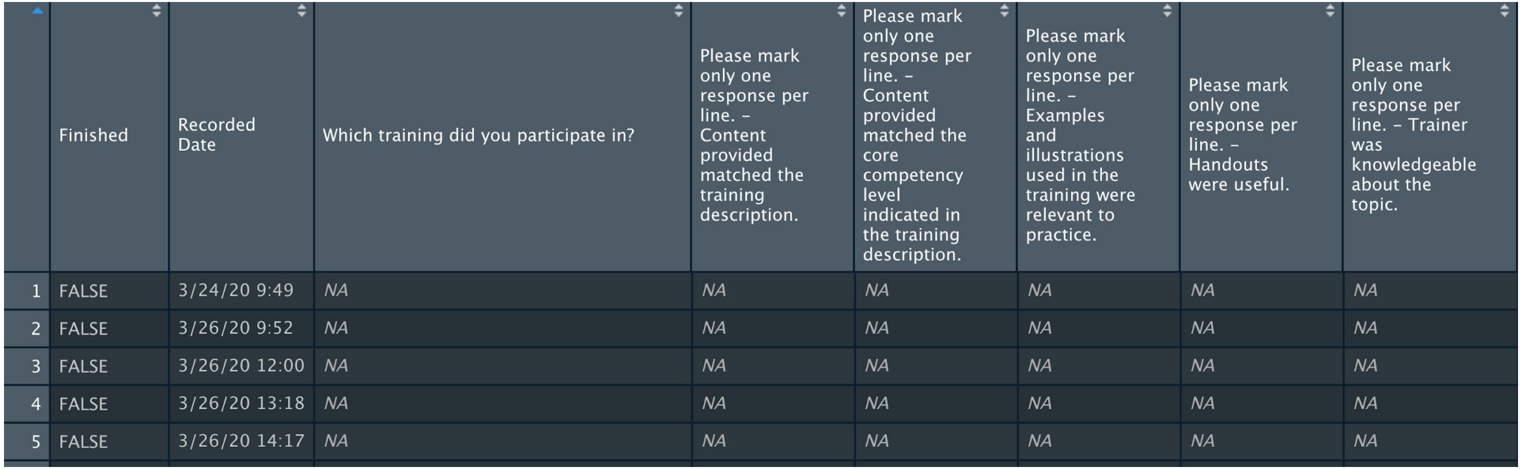
~~~
# Step 2: Loading data
```{r, msg = F, warning = F}
data.wide <- read_csv("EvalSTARS_CPD_051320.csv") # Loading data from an excel file

# Quality check descriptive
head(data.wide, 5) # First five records
tail(data.wide, 5) # Last five records
ls(data.wide)    # Vars names
## Data summary: 112 records / 22 vars - check for null values
```
~~~

[Prep] Data exploration: There are NA records!



Data

```
data.wide    112 obs. of 22 variables
```



	Finished	Recorded Date	Which training did you participate in?	Please mark only one response per line. – Content provided matched the training description.	Please mark only one response per line. – Content provided matched the core competency level indicated in the training description.	Please mark only one response per line. – Examples and illustrations used in the training were relevant to practice.	Please mark only one response per line. – Handouts were useful.	Please mark only one response per line. – Trainer was knowledgeable about the topic.
1	FALSE	3/24/20 9:49	NA	NA	NA	NA	NA	NA
2	FALSE	3/26/20 9:52	NA	NA	NA	NA	NA	NA
3	FALSE	3/26/20 12:00	NA	NA	NA	NA	NA	NA
4	FALSE	3/26/20 13:18	NA	NA	NA	NA	NA	NA
5	FALSE	3/26/20 14:17	NA	NA	NA	NA	NA	NA

[Prep] Options for NA records

- Option 1: Probably the best option :p
  - Do nothing, who cares...
- Option 2: Delete all NA records in excel... which will take couple hours to figure it out if your data set a ton of records....
- Option 3:
  - Identify number of NA records in R first then
  - **`Filter`** your original data set with no NA records.

[Prep] Step 3: Checking Null values

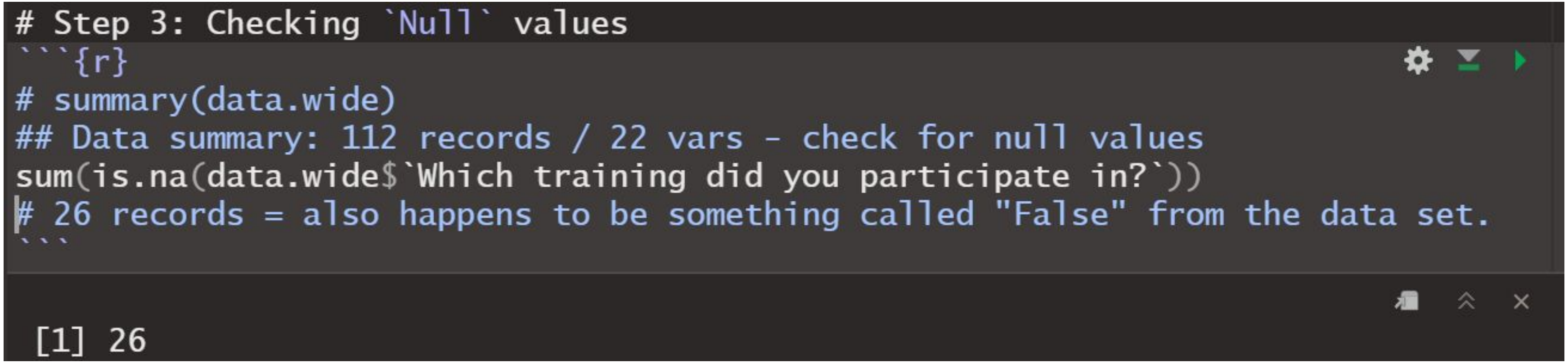
```
sum()            # Summing number of counts within ()

is.na()          # Will generate a list of whether (yourdataframe) has a record of NA

                 or null values
```

- We’re going to figure out how many records are available based whether participant(s)’s indication of which training that they participated in?

```
sum(is.na(data.wide$`Which training did you participate in?`))
```

[Prep] Step 3: Checking Null values

```
sum(is.na(data.wide$`Which training did you participate in?`))
```



~~~
# Step 3: Checking `Null` values
```{r}
# summary(data.wide)
## Data summary: 112 records / 22 vars - check for null values
sum(is.na(data.wide$`Which training did you participate in?`))
# 26 records = also happens to be something called "False" from the data set.
```
~~~

```
[1] 26
```

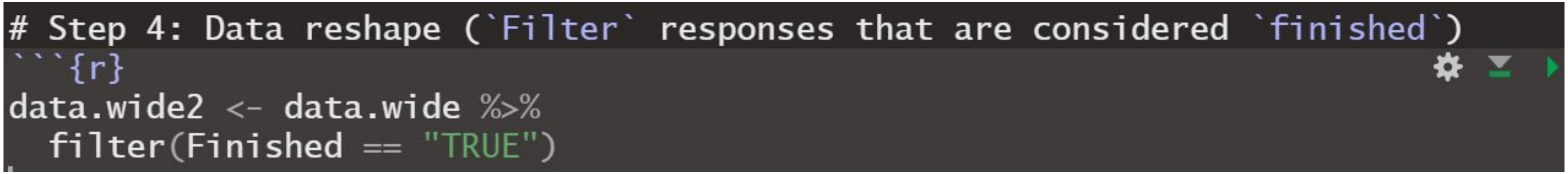
[Prep] Step 4: Filter “Finished” responses

```
filter (ColumnName == “Value”)          # Filter it based on value of the column name
```

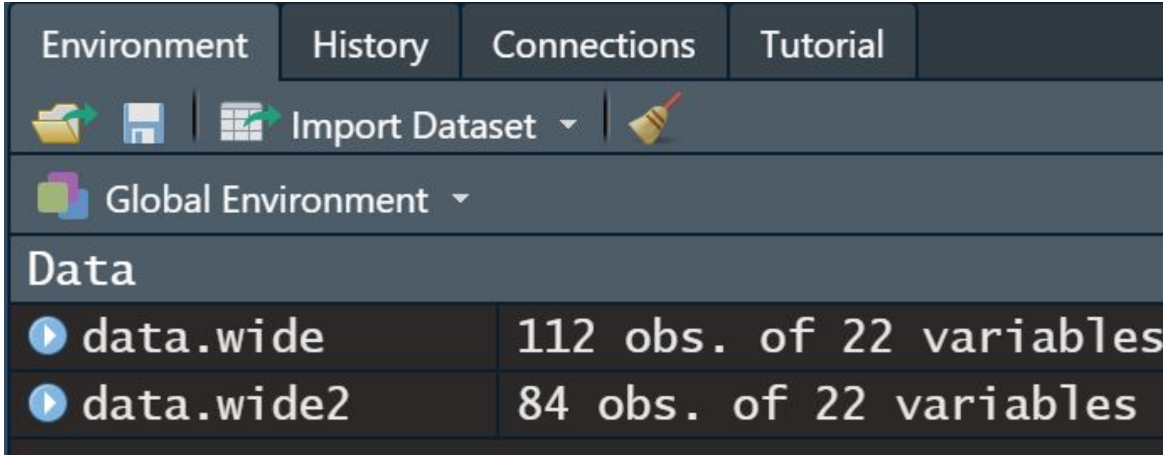
- We’re going to only filter responses that are considered “finished” and create the revised data set into data.wide2

```
data.wide2 <- data.wide %>%

 filter(Finished == "TRUE")
```

[Prep] Step 4: Filter “Finished” responses



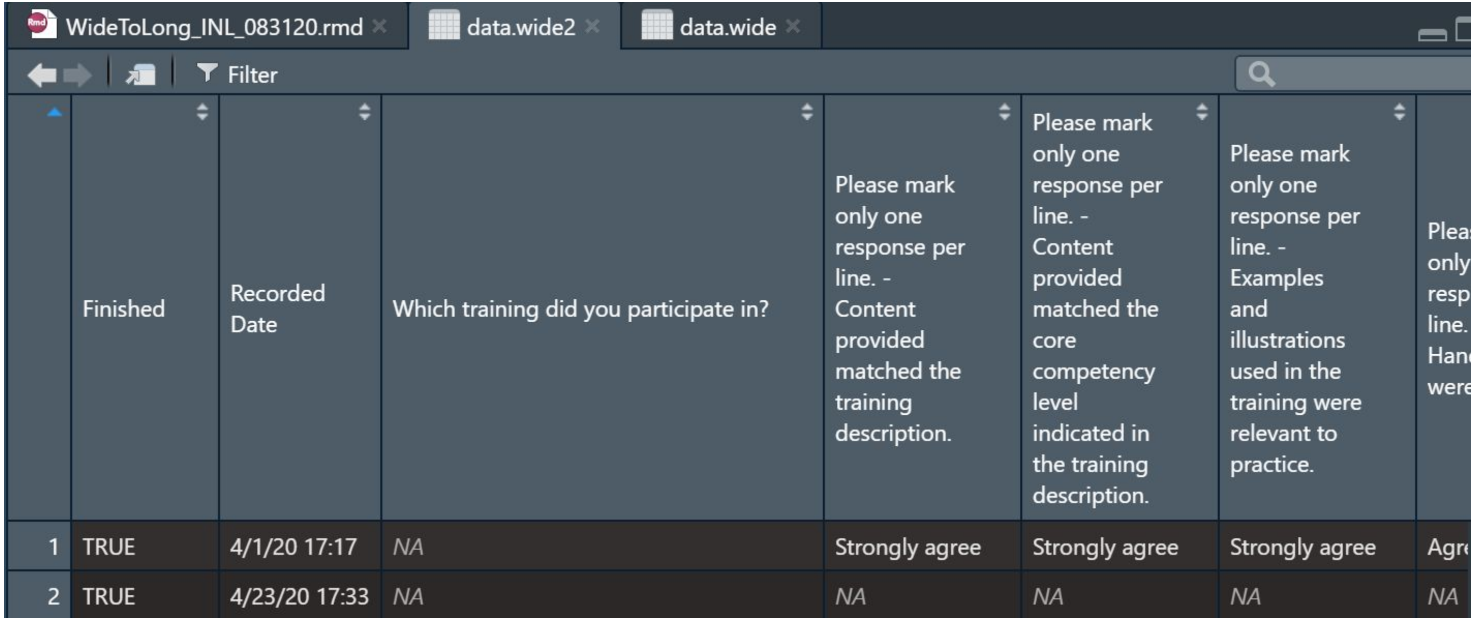
~~~
# Step 4: Data reshape (`Filter` responses that are considered `finished`)
```{r}
data.wide2 <- data.wide %>%
  filter(Finished == "TRUE")
~~~



```
Data
data.wide     112 obs. of 22 variables
data.wide2    84 obs. of 22 variables
```

[Prep] Step 4: Filter “Finished” responses

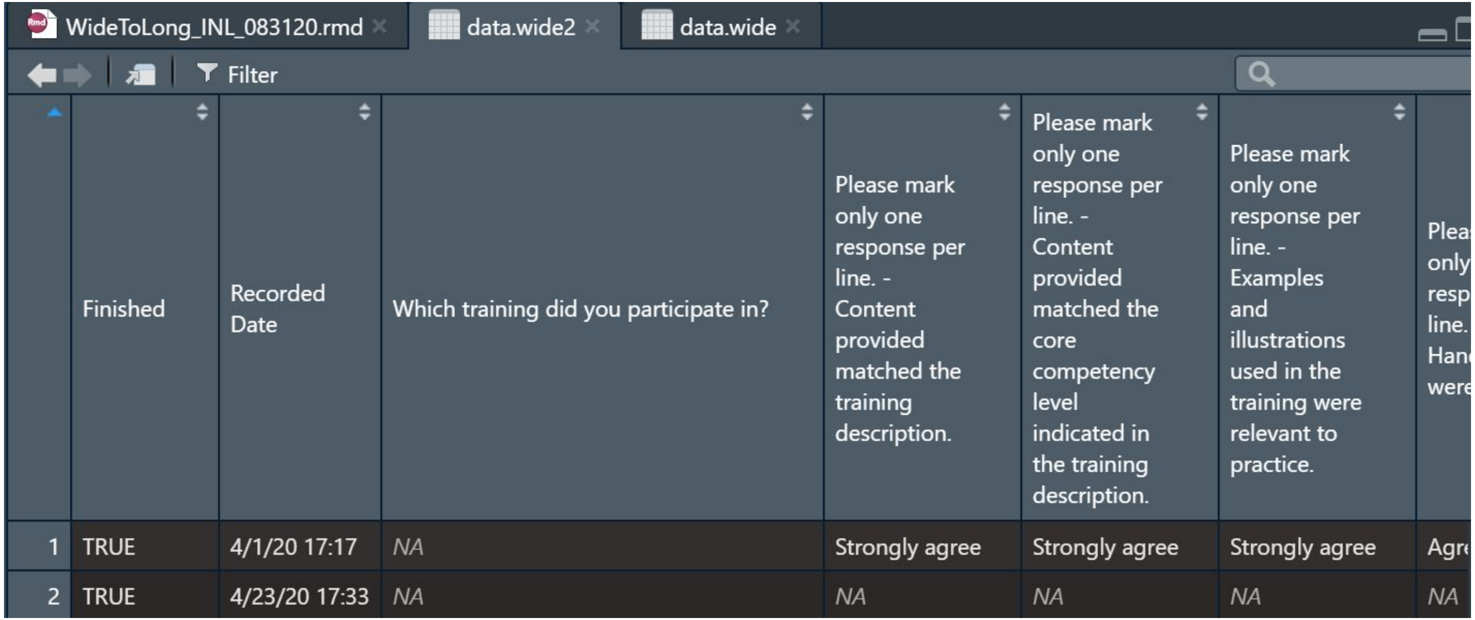
- 112 records - 26 records = 84 records?
- Something is going on with 2 responses…



	Finished	Recorded Date	Which training did you participate in?	Please mark only one response per line. - Content provided matched the training description.	Please mark only one response per line. - Content provided matched the core competency level indicated in the training description.	Please mark only one response per line. - Examples and illustrations used in the training were relevant to practice.	Plea only resp line. Han were
1	TRUE	4/1/20 17:17	NA	Strongly agree	Strongly agree	Strongly agree	Agre
2	TRUE	4/23/20 17:33	NA	NA	NA	NA	NA

[Prep] Step 4: Filter “Finished” responses

- 112 records - 26 records = 84 records?
- Something is going on with 2 responses... (don’t forget to do your quality check!)



	Finished	Recorded Date	Which training did you participate in?	Please mark only one response per line. - Content provided matched the training description.	Please mark only one response per line. - Content provided matched the core competency level indicated in the training description.	Please mark only one response per line. - Examples and illustrations used in the training were relevant to practice.	Plea only resp line. Han were
1	TRUE	4/1/20 17:17	NA	Strongly agree	Strongly agree	Strongly agree	Agre
2	TRUE	4/23/20 17:33	NA	NA	NA	NA	NA

[Transformation] Step 5: `gather` Wide -> Long

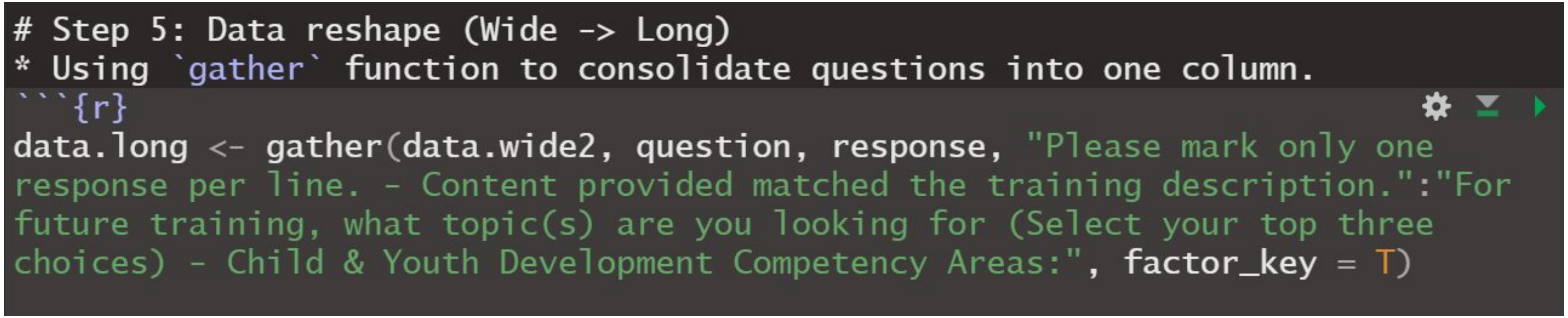
- Using `**gather**` function to gather questions from multiple columns into one column.

gather(dataset, column1, column2, “start column from data set: end column from data set”, factor_key = True)

- Column 1: usually **higher hierarchy** of a column that you’d like to gather it as (i.e. question)
- Column 2: **Value** of the Column 1

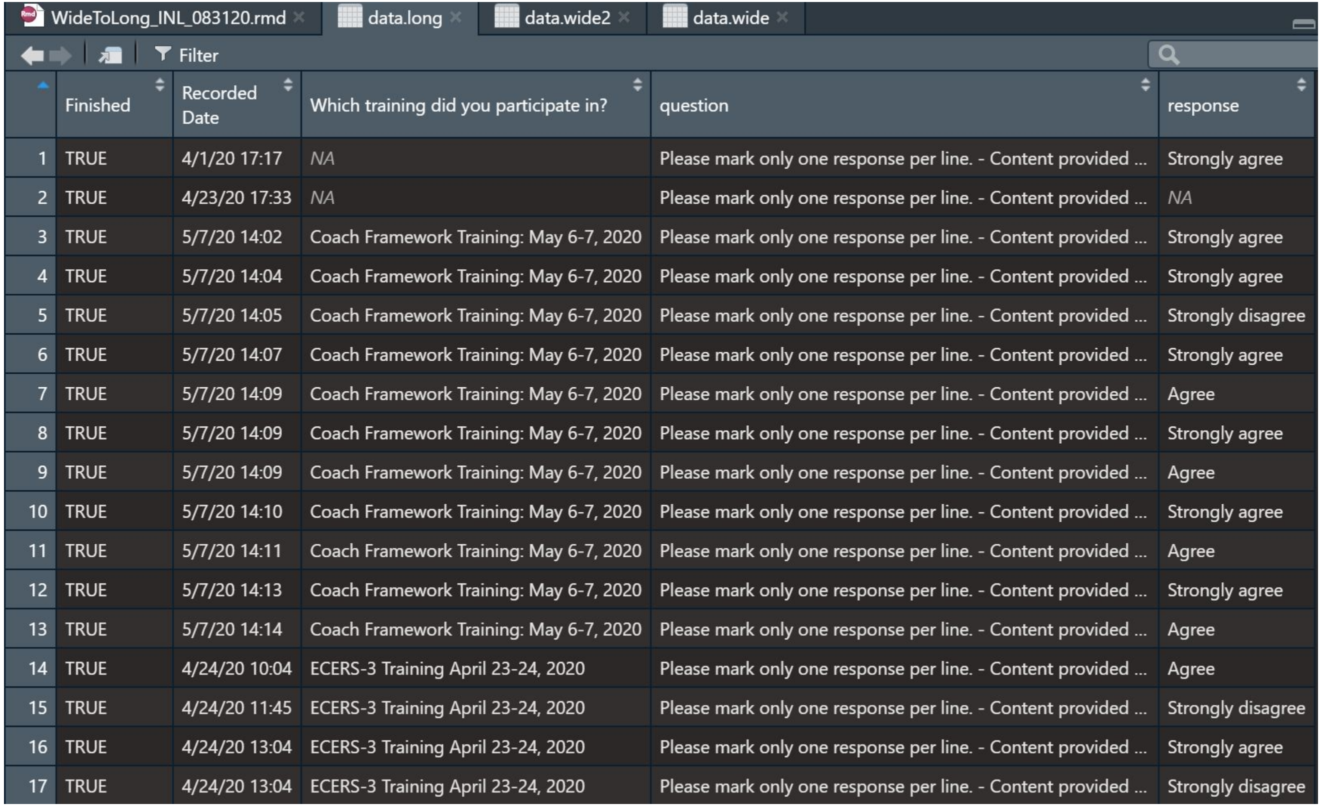
[Transformation] Step 5: `gather` Wide -> Long

data.long <- **gather**(data.wide2, question, response, "Please mark only one response per line. - Content provided matched the training description.":"For future training, what topic(s) are you looking for (Select your top three choices) - Child & Youth Development Competency Areas:", factor_key = T)



~~~
# Step 5: Data reshape (Wide -> Long)
* Using `gather` function to consolidate questions into one column.
```{r}
data.long <- gather(data.wide2, question, response, "Please mark only one
response per line. - Content provided matched the training description.":"For
future training, what topic(s) are you looking for (Select your top three
choices) - Child & Youth Development Competency Areas:", factor_key = T)
~~~

[Transformation] Step 5: `gather` Wide -> Long



	Finished	Recorded Date	Which training did you participate in?	question	response
1	TRUE	4/1/20 17:17	NA	Please mark only one response per line. - Content provided ...	Strongly agree
2	TRUE	4/23/20 17:33	NA	Please mark only one response per line. - Content provided ...	NA
3	TRUE	5/7/20 14:02	Coach Framework Training: May 6-7, 2020	Please mark only one response per line. - Content provided ...	Strongly agree
4	TRUE	5/7/20 14:04	Coach Framework Training: May 6-7, 2020	Please mark only one response per line. - Content provided ...	Strongly agree
5	TRUE	5/7/20 14:05	Coach Framework Training: May 6-7, 2020	Please mark only one response per line. - Content provided ...	Strongly disagree
6	TRUE	5/7/20 14:07	Coach Framework Training: May 6-7, 2020	Please mark only one response per line. - Content provided ...	Strongly agree
7	TRUE	5/7/20 14:09	Coach Framework Training: May 6-7, 2020	Please mark only one response per line. - Content provided ...	Agree
8	TRUE	5/7/20 14:09	Coach Framework Training: May 6-7, 2020	Please mark only one response per line. - Content provided ...	Strongly agree
9	TRUE	5/7/20 14:09	Coach Framework Training: May 6-7, 2020	Please mark only one response per line. - Content provided ...	Agree
10	TRUE	5/7/20 14:10	Coach Framework Training: May 6-7, 2020	Please mark only one response per line. - Content provided ...	Strongly agree
11	TRUE	5/7/20 14:11	Coach Framework Training: May 6-7, 2020	Please mark only one response per line. - Content provided ...	Agree
12	TRUE	5/7/20 14:13	Coach Framework Training: May 6-7, 2020	Please mark only one response per line. - Content provided ...	Strongly agree
13	TRUE	5/7/20 14:14	Coach Framework Training: May 6-7, 2020	Please mark only one response per line. - Content provided ...	Agree
14	TRUE	4/24/20 10:04	ECERS-3 Training April 23-24, 2020	Please mark only one response per line. - Content provided ...	Agree
15	TRUE	4/24/20 11:45	ECERS-3 Training April 23-24, 2020	Please mark only one response per line. - Content provided ...	Strongly disagree
16	TRUE	4/24/20 13:04	ECERS-3 Training April 23-24, 2020	Please mark only one response per line. - Content provided ...	Strongly agree
17	TRUE	4/24/20 13:04	ECERS-3 Training April 23-24, 2020	Please mark only one response per line. - Content provided ...	Strongly disagree

Summary

- Understanding how data can be reshaped can help you in a long run to feed your data to any software.
- It takes more time to plan these transformation structure than actual execution.
- Try to recycle what you have tried out last time :)

Next Steps

- Challenge! Institute scenario!

Institute Scenario

- Cultivate Learning institute manager has collected a data set of “session evaluation” from the last 2020 virtual institute.
- The current data set seems it has each session on each tab, and it’s not well organized as we hoped for.
- We communicated to the client that this would take a while to clean up before we attempt to visualize this data on Tableau - the platform where the client wants to visualize the data set.
- **You have 72 hours to think thoroughly and document your plan on how to organize this data set to feed into Tableau.**

Thank you!

감사합니다!