Week 5: Surveys

Charlotte Hadley

Topics for today

```r
library(tidyverse)
library(gt)
library(readxl)
library(here)
library(janitor)
library(haven)
```

This week we’re going to be discussing surveys and wrangling survey data in R.

The goals for the lecture section of today is as follows:

1. Identify what makes effective surveys
2. Correctly identify if data is “long” or “wide”
3. Understand how to use the {tidyr} pivot functions for moving between “wide” and “long” data
4. Understand how to use the tidyverse for common data wrangling tasks when working with survey data

We’ll likely continue some of the lecture material into the workshop.

Surveys are an indispensible part of healthcare

There are lots of absolute quantitative measures in healthcare [datascience]

- Patient wait times
- Morbidity
- Biological samples
- Physiological health measurements
- Device-based measurements
- Anthropometric measurements
- Sensory measurements

But these measures on **there own** are often meaningless…

- … for understanding patient experiences
- … for tracking patient outcomes
- … for medical trials
- … for designing medical devices

We need to understand these measures in context of the patient/device/intervention.

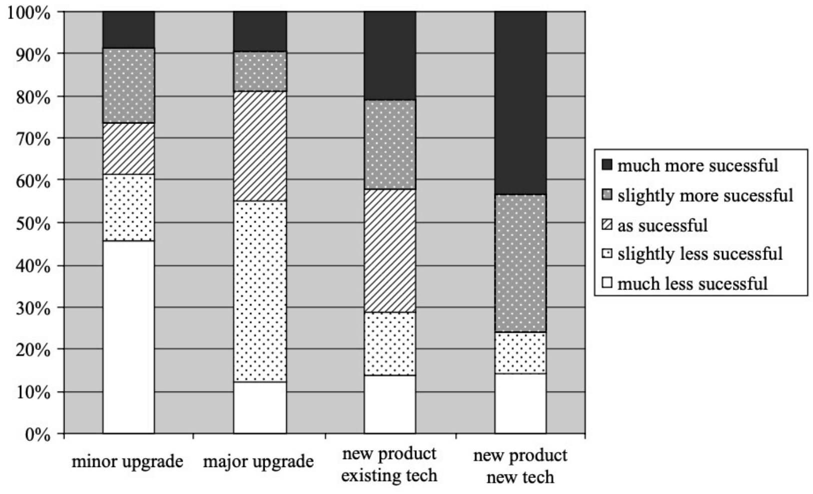
Surveys are an indispensible part of healthcare

Surveys might be the sole measurement we take in a study.

NHS patient experience surveys

“The importance of urban natural areas and urban ecosystem services during the COVID-19 pandemic”[1].

How successful do medical technology companies rate their devices[2]?



Source: Eatock, et al.@ 2

Surveys are an indispensible part of healthcare

Surveys might be the sole measurement we take in a study.

NHS patient experience surveys

“The importance of urban natural areas and urban ecosystem services during the COVID-19 pandemic”[1].

How successful do medical technology companies rate their devices[2]?

Surveys might instead provide additional context for other measurements that we take.

- Diet studies might take biological samples but require food surveys
- Mental health studies might track physiological measurements as well as psychological surveys

Designing effective surveys is hard

We don’t have enough time to go deep into how to design an effective survey - that’s probably an entire undergraduate course in its own right.

> So why are we looking at surveys?

Designing effective surveys is hard

There are some specific topics in designing surveys I want to cover.

These “tips” are geared towards designing surveys where you can easily analyse the data after running the survey.

However. These tips do not ensure an effective survey.

In an effective survey:

- There is an overall goal for the survey.
- Each question is asking what you think it’s asking.
- Questions are unbiased and are not leading.

The best way to test the effectiveness of your survey is pretesting[3]

Designing effective surveys is hard

Getting survey data into R for analysis requires many different data wrangling tasks/skills.

The `tidyverse` gets its name from the concept of “tidy data” defined by Hadley Wickham[4] back in 2014.

We will use survey data as an introduction to this concept and will cover 4 types of wrangling:

- Pivoting data between wide and long formats
- Joining datasets [in the same way that SQL databases are joined]
- Wrangling survey questions with multiple choices
- Wrangling survey questions that capture multiple pieces of information

Quantitative vs Qualitative measurements?

“Closed question” vs “Open question”?

Which is which?

Surveys can capture all sorts of data

Quantitative measurements can be collected with closed-ended questions.

- Do you own a fitness tracking device
  - Yes / No
- How often do you wear your tracking device
  - Every day / Some days / A few days / Rarely
- Since owning a tracking device do you feel like you know more about your activity?
  - Strong agree / Agree / Neither agree nor disagree / Disagree / Strong disagree

Qualitative measurements are collected using open-ended questions - or free-text fields.

> If you were to recommend device X what would you tell someone?

Surveys can capture all sorts of data

Ideally we would always use closed-form questions to capture quantitative information. But unfortunately this isn't always the case. Take this question from the 2019 British Election Survey[5]:

Question text: "Where do you get most of your information about politics or current affairs from?"

Question input type: free-form text

Example responses:

What do you think about this question?

Q. Do you support the idea that charities should not pay tax?

- □ Yes
- □ No
- □ Don’t know

This is a question from the NHS England’s bite-size guide to writing effective questionnaires[6].

Survey Mode

Methods of survey data collection (I)

There are lots of different methods/modes for survey data collection:

- Online (open/close)
- Telephone
- Mail
- Face-to-face
- Paper (observed)
- Mixed-mode
  - Same survey different modes
  - Multi-phase survey with different modes

There is considerable evidence for mode of data collection affecting survey results[7]:

Respondents answer questions differently by mode[8]

Respondent demographics vary by mode and survey topic

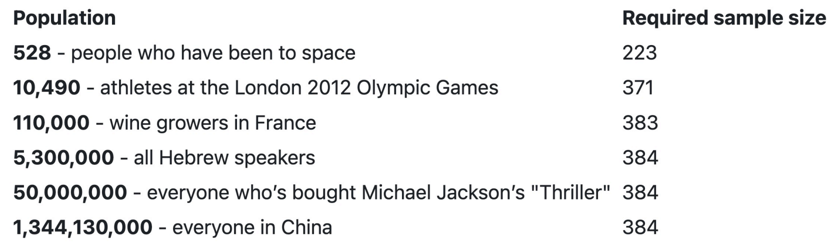
> Inaccurate state-level polls forn the 2016 US elections are considered to have been strongly biased by an over-representation of college graduates[9].

Survey Size

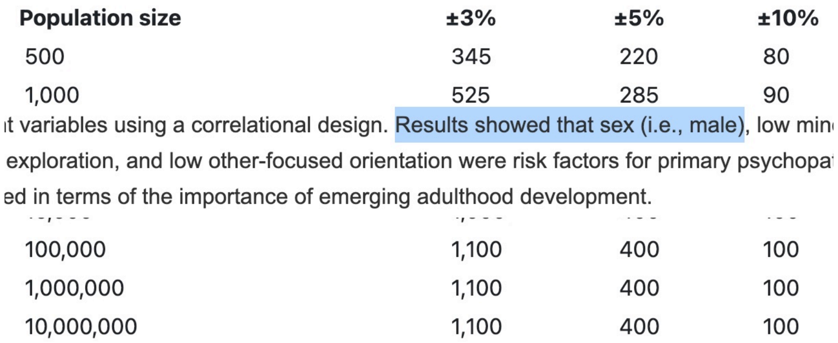
Survey Size

Many survey tools provide interactive calculators for estimating survey population requirements - surveymonkey.com/mp/sample-size-calculator.

- These are reasonable targets for “survey studies” but for other studies refer to Serdar et al[10].
- For pre-test surveys a useful heuristic is to use sample at least 30 participants as per Perneger et al[11]



Population	Required sample size
528 - people who have been to space	223
10,490 - athletes at the London 2012 Olympic Games	371
110,000 - wine growers in France	383
5,300,000 - all Hebrew speakers	384
50,000,000 - everyone who's bought Michael Jackson's "Thriller"	384
1,344,130,000 - everyone in China	384



Population size	±3%	±5%	±10%
500	345	220	80
1,000	525	285	90
[illegible]	[illegible]	[illegible]	[illegible]
100,000	1,100	400	100
1,000,000	1,100	400	100
10,000,000	1,100	400	100

ıt variables using a correlational design. Results showed that sex (i.e., male), low min
exploration, and low other-focused orientation were risk factors for primary psychopa
ed in terms of the importance of emerging adulthood development.

Source: Teller[12]

LIKERT Scales (I)

In a LIKERT scale responses are given a score.

response	response_score
Strong disagree	1
Disagree	2
Neutral	3
Agree	4
Strong agree	5

question	response_score
Feel confident in {ggplot2}?	Strong agree
Feel confident in {dplyr}?	Neutral
Feel confident in {tidyr}?	Disagree
Feel confident in {purrr}?	Strong disagree

It is absolutely meaningless to take the mean of a LIKERT scale[13] - but people still do it.

When you create your survey question you're creating an **ordinal variable**

LIKERT Scales (II)

response	response_score
Strong disagree	1
Disagree	2
Neutral	3
Agree	4
Strong agree	5

question	response_score
Feel confident in {ggplot2}?	Strong agree
Feel confident in {dplyr}?	Neutral
Feel confident in {tidyr}?	Disagree
Feel confident in {purrr}?	Strong disagree

Our original question defines an ordinal variable - there’s an intrinsic order to the responses.

```r
cat("Strong disagree < Disagree ... < Strong agree")
```

```
Strong disagree < Disagree ... < Strong agree
```

When we convert the question to a LIKERT score we are then working with a **interval data**.

> But is this accurate?
>
> Is the **distance** between “Strong disagree” and “Disagree” the same as that between “Neutral” and “Agree”?

LIKERT Scales (III)

response	response_score
Strong disagree	1
Disagree	2
Neutral	3
Agree	4
Strong agree	5

question	response_score
Feel confident in {ggplot2}?	Strong agree
Feel confident in {dplyr}?	Neutral
Feel confident in {tidyr}?	Disagree
Feel confident in {purrr}?	Strong disagree

If you want to compare answers across multiple LIKERT scales then there is some meaning to the “median” response.

You could perform factor analysis - which is well explained by Batterton and Hale[14].

When designing your survey you could also directly ask respondents a question instead of trying to guess what “disagree” + “neutral” means.

> Overall, do you feel confident with using the tidyverse?

Missing data in surveys

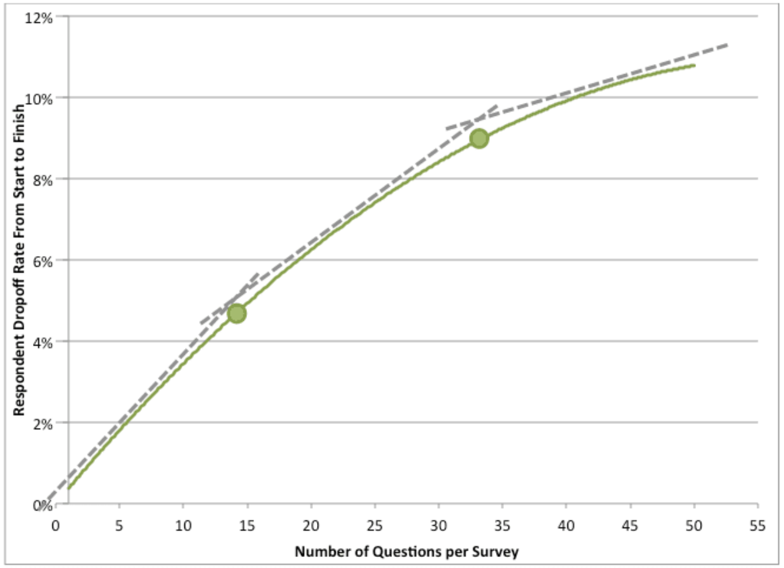
Missing data in surveys (I)

When designing a survey we ideally want respondents to answer all questions.

However, research is clear[15] that drop off rate (or survey abandonment) is correlated with survey length.

Although, this is strongly affected by:

- survey mode
- survey reward
- are questions skippable?



Source: SurveyMonkey[15]

Missing data in surveys: MNAR

There are three different kinds of missing data distribution:

- Missing Not at Random (MNAR)
- Missing at Random (MAR)
- Missing Completely at Random (MCAR)

Missingness is related to what is missing.

- Missingess of responses might be related to question order - respondents give up.
- Missingness might be due to respondent’s feeling towards questions [eg extra-marital relations, triggering subjects]

Data that’s MNAR indicates a bias in your study design.

However - it’s often difficult to determine if your data is indeed MNAR or not.

Missing data in surveys: MAR

There are three different kinds of missing data distribution:

- Missing Not at Random (MNAR)
- Missing at Random (MAR)
- Missing Completely at Random (MCAR)

Missingness is not random but can be accounted for by other variables/ We could call this “conditionally at random”.

- Survey abandonment could be modelled by including information about question order.
- Missingness might be due to known differences in demographics, for instance males are less likely to complete depression surveys.

Data that’s MAR provides us with a methodology for imputing missing values.

Missing data in surveys: MCAR

There are three different kinds of missing data distribution:

- Missing Not at Random (MNAR)
- Missing at Random (MAR)

- Missing Completely at Random (MCAR)

Missingness is truly randomly distributed in the dataset. There are no hidden variables.

In practice it is hard to verify MCAR over MNAR without specifically designing your study for randomness.

- Randomly sampling a subset of questions for each partipant.

Planned missingness designs[16] depend on fairly advanced statistical methodologies.

Survey tools

Survey tools

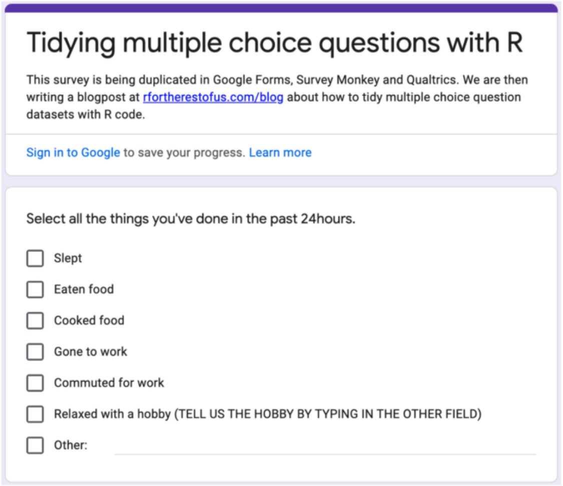
There’s a plethora of survey tools available.

Many of these tools provide free tiers but require subscriptions/licenses for wide scale use [and for GDPR compliance].

One thing that unifies all of these tools is that each one has it’s own unique data export format - irrespective of file format.

> Let’s look at this first and then talk about file formats

Google Forms



Tidying multiple choice questions with R

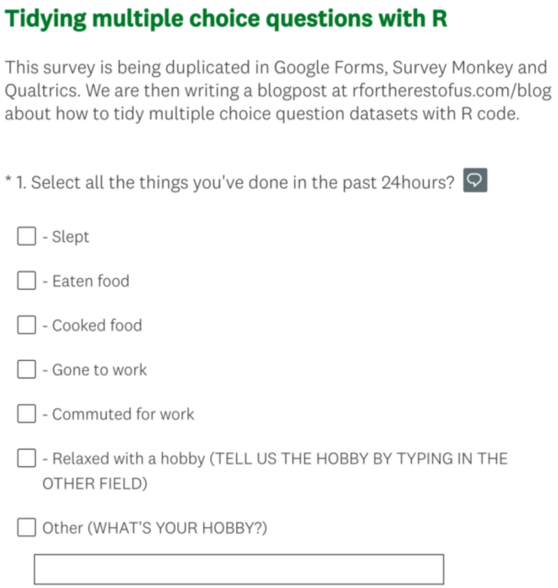
This survey is being duplicated in Google Forms, Survey Monkey and Qualtrics. We are then writing a blogpost at rfortherestofus.com/blog about how to tidy multiple choice question datasets with R code.

Sign in to Google to save your progress. Learn more

Select all the things you've done in the past 24hours.

- □ Slept
- □ Eaten food
- □ Cooked food
- □ Gone to work
- □ Commuted for work
- □ Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD)
- □ Other: ____

Survey Monkey



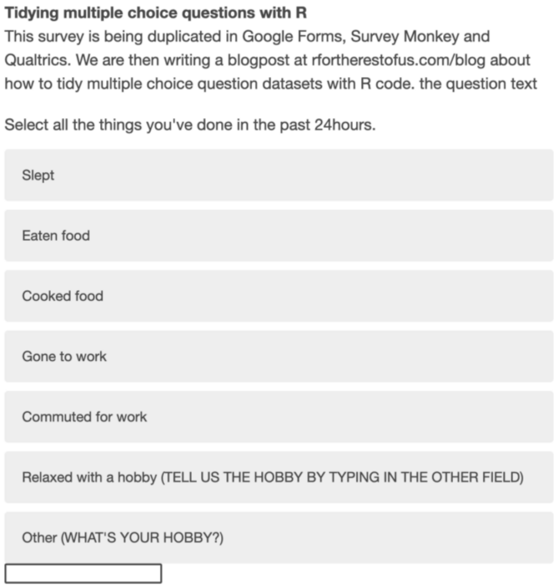
Tidying multiple choice questions with R

This survey is being duplicated in Google Forms, Survey Monkey and Qualtrics. We are then writing a blogpost at rfortherestofus.com/blog about how to tidy multiple choice question datasets with R code.

* 1. Select all the things you've done in the past 24hours?

- □ - Slept
- □ - Eaten food
- □ - Cooked food
- □ - Gone to work
- □ - Commuted for work
- □ - Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD)
- □ Other (WHAT'S YOUR HOBBY?)

Qualtrics



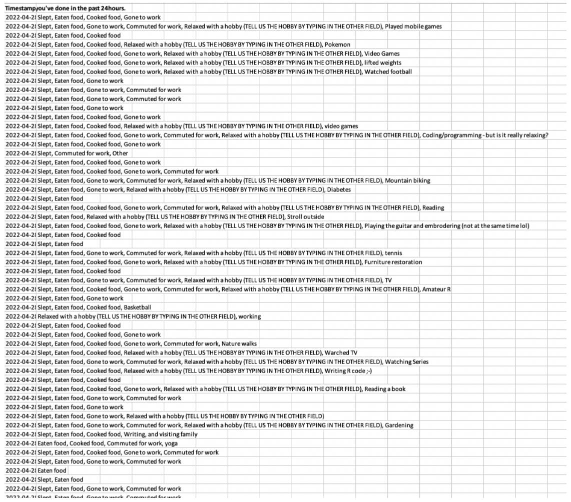
Tidying multiple choice questions with R

This survey is being duplicated in Google Forms, Survey Monkey and Qualtrics. We are then writing a blogpost at rfortherestofus.com/blog about how to tidy multiple choice question datasets with R code. the question text

Select all the things you've done in the past 24hours.

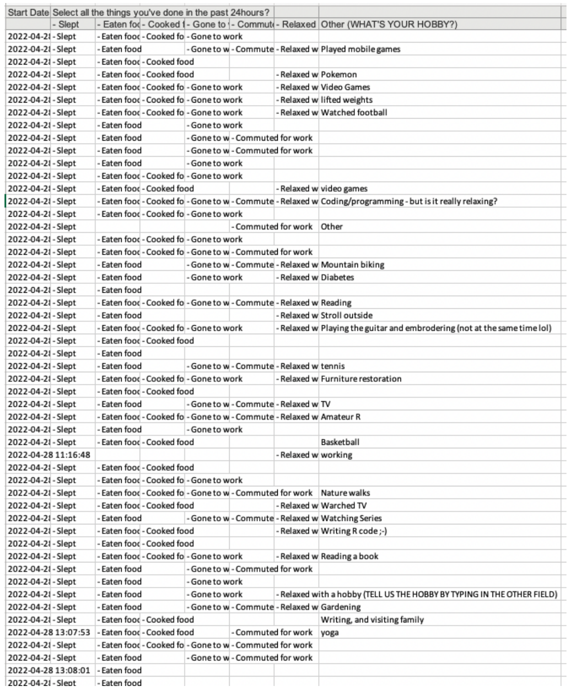
- Slept
- Eaten food
- Cooked food
- Gone to work
- Commuted for work
- Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD)
- Other (WHAT'S YOUR HOBBY?)

Google Forms



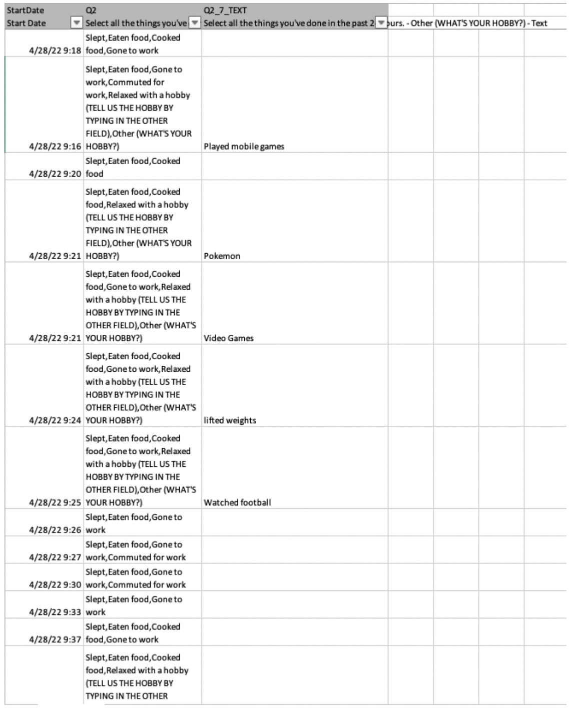
Timestamp	you've done in the past 24hours.
2022-04-2	Slept, Eaten food, Cooked food, Gone to work
2022-04-2	Slept, Eaten food, Gone to work, Commuted for work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Played mobile games
2022-04-2	Slept, Eaten food, Cooked food
2022-04-2	Slept, Eaten food, Cooked food, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Pokemon
2022-04-2	Slept, Eaten food, Cooked food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Video Games
2022-04-2	Slept, Eaten food, Cooked food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), lifted weights
2022-04-2	Slept, Eaten food, Cooked food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Watched football
2022-04-2	Slept, Eaten food, Gone to work
2022-04-2	Slept, Eaten food, Gone to work, Commuted for work
2022-04-2	Slept, Eaten food, Gone to work, Commuted for work
2022-04-2	Slept, Eaten food, Gone to work
2022-04-2	Slept, Eaten food, Cooked food, Gone to work
2022-04-2	Slept, Eaten food, Cooked food, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), video games
2022-04-2	Slept, Eaten food, Cooked food, Gone to work, Commuted for work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Coding/programming - but is it really relaxing?
2022-04-2	Slept, Eaten food, Cooked food, Gone to work
2022-04-2	Slept, Commuted for work, Other
2022-04-2	Slept, Eaten food, Cooked food, Gone to work
2022-04-2	Slept, Eaten food, Cooked food, Gone to work, Commuted for work
2022-04-2	Slept, Eaten food, Gone to work, Commuted for work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Mountain biking
2022-04-2	Slept, Eaten food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Diabetes
2022-04-2	Slept, Eaten food
2022-04-2	Slept, Eaten food, Cooked food, Gone to work, Commuted for work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Reading
2022-04-2	Slept, Eaten food, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Stroll outside
2022-04-2	Slept, Eaten food, Cooked food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Playing the guitar and embrodering (not at the same time lol)
2022-04-2	Slept, Eaten food, Cooked food
2022-04-2	Slept, Eaten food
2022-04-2	Slept, Eaten food, Gone to work, Commuted for work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), tennis
2022-04-2	Slept, Eaten food, Cooked food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Furniture restoration
2022-04-2	Slept, Eaten food, Cooked food
2022-04-2	Slept, Eaten food, Gone to work, Commuted for work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), TV
2022-04-2	Slept, Eaten food, Cooked food, Gone to work, Commuted for work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Amateur R
2022-04-2	Slept, Eaten food, Gone to work
2022-04-2	Slept, Eaten food, Cooked food, Basketball
2022-04-2	Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), working
2022-04-2	Slept, Eaten food, Cooked food
2022-04-2	Slept, Eaten food, Cooked food, Gone to work
2022-04-2	Slept, Eaten food, Cooked food, Gone to work, Commuted for work, Nature walks
2022-04-2	Slept, Eaten food, Cooked food, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Warched TV
2022-04-2	Slept, Eaten food, Gone to work, Commuted for work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Watching Series
2022-04-2	Slept, Eaten food, Cooked food, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Writing R code ;-)
2022-04-2	Slept, Eaten food, Cooked food
2022-04-2	Slept, Eaten food, Cooked food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Reading a book
2022-04-2	Slept, Eaten food, Gone to work, Commuted for work
2022-04-2	Slept, Eaten food, Gone to work
2022-04-2	Slept, Eaten food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD)
2022-04-2	Slept, Eaten food, Gone to work, Commuted for work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Gardening
2022-04-2	Slept, Eaten food, Cooked food, Writing, and visiting family
2022-04-2	Eaten food, Cooked food, Commuted for work, yoga
2022-04-2	Slept, Eaten food, Cooked food, Gone to work, Commuted for work
2022-04-2	Slept, Eaten food, Gone to work, Commuted for work
2022-04-2	Eaten food
2022-04-2	Slept, Eaten food
2022-04-2	Slept, Eaten food, Gone to work, Commuted for work
2022-04-2	Slept, Eaten food, Gone to work, Commuted for work

Survey Monkey



Start Date	Select all the things you've done in the past 24hours?					
	- Slept	- Eaten foo	- Cooked f	- Gone to	- Commute	- Relaxed	Other (WHAT'S YOUR HOBBY?)
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to work			
2022-04-2	- Slept	- Eaten food		- Gone to w	- Commute	- Relaxed w	Played mobile games
2022-04-2	- Slept	- Eaten foo	- Cooked food				
2022-04-2	- Slept	- Eaten foo	- Cooked food			- Relaxed w	Pokemon
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to work		- Relaxed w	Video Games
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to work		- Relaxed w	lifted weights
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to work		- Relaxed w	Watched football
2022-04-2	- Slept	- Eaten food		- Gone to work			
2022-04-2	- Slept	- Eaten food		- Gone to w	- Commuted for work		
2022-04-2	- Slept	- Eaten food		- Gone to w	- Commuted for work		
2022-04-2	- Slept	- Eaten food		- Gone to work			
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to work			
2022-04-2	- Slept	- Eaten foo	- Cooked food			- Relaxed w	video games
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to w	- Commute	- Relaxed w	Coding/programming - but is it really relaxing?
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to work			
2022-04-2	- Slept				- Commuted for work		Other
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to work			
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to w	- Commuted for work		
2022-04-2	- Slept	- Eaten food		- Gone to w	- Commute	- Relaxed w	Mountain biking
2022-04-2	- Slept	- Eaten food		- Gone to work		- Relaxed w	Diabetes
2022-04-2	- Slept	- Eaten food					
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to w	- Commute	- Relaxed w	Reading
2022-04-2	- Slept	- Eaten food				- Relaxed w	Stroll outside
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to work		- Relaxed w	Playing the guitar and embrodering (not at the same time lol)
2022-04-2	- Slept	- Eaten foo	- Cooked food				
2022-04-2	- Slept	- Eaten food					
2022-04-2	- Slept	- Eaten food		- Gone to w	- Commute	- Relaxed w	tennis
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to work		- Relaxed w	Furniture restoration
2022-04-2	- Slept	- Eaten foo	- Cooked food				
2022-04-2	- Slept	- Eaten food		- Gone to w	- Commute	- Relaxed w	TV
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to w	- Commute	- Relaxed w	Amateur R
2022-04-2	- Slept	- Eaten food		- Gone to work			
2022-04-2	- Slept	- Eaten foo	- Cooked food				Basketball
2022-04-28 11:16:48						- Relaxed w	working
2022-04-2	- Slept	- Eaten foo	- Cooked food				
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to work			
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to w	- Commuted for work		Nature walks
2022-04-2	- Slept	- Eaten foo	- Cooked food			- Relaxed w	Warched TV
2022-04-2	- Slept	- Eaten food		- Gone to w	- Commute	- Relaxed w	Watching Series
2022-04-2	- Slept	- Eaten foo	- Cooked food			- Relaxed w	Writing R code ;-)
2022-04-2	- Slept	- Eaten foo	- Cooked food				
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to work		- Relaxed w	Reading a book
2022-04-2	- Slept	- Eaten food		- Gone to w	- Commuted for work		
2022-04-2	- Slept	- Eaten food		- Gone to work			
2022-04-2	- Slept	- Eaten food		- Gone to work		- Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD)	
2022-04-2	- Slept	- Eaten food		- Gone to w	- Commute	- Relaxed w	Gardening
2022-04-2	- Slept	- Eaten foo	- Cooked food				Writing, and visiting family
2022-04-28 13:07:53		- Eaten foo	- Cooked food		- Commuted for work		yoga
2022-04-2	- Slept	- Eaten foo	- Cooked fo	- Gone to w	- Commuted for work		
2022-04-2	- Slept	- Eaten food		- Gone to w	- Commuted for work		
2022-04-28 13:08:01		- Eaten food					
2022-04-2	- Slept	- Eaten food					

Qualtrics



StartDate	Q2	Q2_7_TEXT	
Start Date	Select all the things you've	Select all the things you've done in the past 2	urs. - Other (WHAT'S YOUR HOBBY?) - Text
4/28/22 9:18	Slept,Eaten food,Cooked food,Gone to work		
4/28/22 9:16	Slept,Eaten food,Gone to work,Commuted for work,Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD),Other (WHAT'S YOUR HOBBY?)	Played mobile games	
4/28/22 9:20	Slept,Eaten food,Cooked food		
4/28/22 9:21	Slept,Eaten food,Cooked food,Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD),Other (WHAT'S YOUR HOBBY?)	Pokemon	
4/28/22 9:21	Slept,Eaten food,Cooked food,Gone to work,Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD),Other (WHAT'S YOUR HOBBY?)	Video Games	
4/28/22 9:24	Slept,Eaten food,Cooked food,Gone to work,Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD),Other (WHAT'S YOUR HOBBY?)	lifted weights	
4/28/22 9:25	Slept,Eaten food,Cooked food,Gone to work,Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD),Other (WHAT'S YOUR HOBBY?)	Watched football	
4/28/22 9:26	Slept,Eaten food,Gone to work		
4/28/22 9:27	Slept,Eaten food,Gone to work,Commuted for work		
4/28/22 9:30	Slept,Eaten food,Gone to work,Commuted for work		
4/28/22 9:33	Slept,Eaten food,Gone to work		
4/28/22 9:37	Slept,Eaten food,Cooked food,Gone to work		
	Slept,Eaten food,Cooked food,Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER		

Survey tools/software

Because there’s such a huge variety in data export format for survey data it’s important to train your general purpose data wrangling skills.

In the lecture we’ll look at several real-world datasets that require lots of complicated wrangling.

In the workshop you’ll look at simpler datasets and practice the same wrangling skills you’ll learn now.

Survey file formats

- Excel files: Almost all tools will provide .xlsx files. However, there’s a vast range of options in how the data is arranged - including a separate sheet for each question.
  - We use the {readxl} package for reading in these files
  - If data is encoded with cell colour you’ll need to use the more complex {tidyxl} package.

- .csv files: Most tools provide a .csv export, .csv files are an example of a “flat” or “plain text” file. They’re likely to be well formatted for reading into R.
  - Flat files like .csv are read into R with the {readr} package

Survey datasets for today

Survey datasets (I)

We'll be looking at 3 different datasets during this week:

- Emerging Adulthood Measured at Multiple Institutions data set[17]
    - This is the 2nd instance of a large scale survey across multiple institutions.
    - Learn about the 1st instance (and how the study works) from Reifman and Grahe[18].
    - The actual survey questionnaire is available here osf.io/3zq5e
    - The survey dataset is stored on OSF.com as a collection[19]
    - The actual survey data is available from osf.io/download/c3pf6 and can be downloaded via a URL.

Survey datasets (II)

We’ll be looking at 3 different datasets during this week:

- Emerging Adulthood Measured at Multiple Institutions data set[17]

- Public Attitudes to Commercial Access to Health Data[20]
  - The Wellcome Trust commissioned Ipsos MORI to survey opinions on commercial access to health data
  - The study can be read here[21]
  - The survey questions can be found at the end of the report[21].
  - The dataset is openly available on the UK Data Service[20] but requires you to have an account.

Survey datasets (II)

We’ll be looking at 3 different datasets during this week:

- Emerging Adulthood Measured at Multiple Institutions data set[17]
- Public Attitudes to Commercial Access to Health Data[20]

- British Election Study 2019
  - Since 1964 a post-election survey has been carried out to understand electoral motivations and the impact of political party campaigning.
  - Data for all surveys is available from britishelectionstudy.com/data-objects/cross-sectional-data/
  - The 2019 questionnaire and information on how it was rolled out is britishelectionstudy.com/data-object/2019-british-election-study-post-election-random-probability-survey/[22].
  - The 2019 survey data is available from the UK Data Service[5] but must be manually added to your UK Data Service account.

Survey datasets (IV)

We’ll be looking at 3 different datasets during this week:

- Emerging Adulthood Measured at Multiple Institutions data set17
- Public Attitudes to Commercial Access to Health Data20
- British Election Study 2019

📝 Task: Setup our project

SLIDE 1 OF 3

1. Create a new project called `eng7218-week-5_surveys`
2. Add a subfolder called `data` to store the datasets.
3. Create a separate `.Rmd` document for each of the studies:

- `emerging-adulthood.Rmd`
- `commercial-access-to-health-data.Rmd`
- `british-election-study-2019.Rmd`

📝 Task: Obtain Emerging Adulthood data

SLIDE 2 OF 3

1. Open the emerging-adulthood.Rmd file

There are two important files from https://osf.io/qtqpb/[19] that we need:

- The codebook
- The dataset

2. Add a code chunk to load the `{tidyverse}` and `{readxl}` packages.
3. Add this code chunk to your .Rmd to download these files

📝 Task: Read in Emerging Adulthood data

SLIDE 3 OF 3

When we read in datasets we should always assume they need cleaning, so let’s import these files with object names that indicate this.

```
adulthood_raw_data <- read_excel("data/emerging-adulthood_data.xlsx")
adulthood_raw_codebook <- read_excel("data/emerging-adulthood_codebook.xlsx")
```

Messy column names (I)

Most datafiles you’ll work with will have messy column names that are annoying to work with:

```r
glimpse(adulthood_raw_codebook)
```

```
Rows: 328
Columns: 6
$ `Variable Name`      <chr> "StartDate", "EndDate",
"Status", "Progress", "Du…
$ `Question text`      <chr> "Start Date", "End Date",
"Response Type", "Progr…
$ ...3                 <chr> "n/a", "n/a", "n/a", "n/a",
"n/a", "n/a", "n/a", …
$ responses            <chr> "qualtrics variable",
"qualtrics variable", "qual…
$ ...5                 <lgl> NA, NA, NA, NA, NA, NA, NA,
NA, NA, NA, NA, NA, N…
$ `Survey Question ID` <chr> "
{\"ImportId\":\"startDate\"}", "{\"ImportId\":\"…
```

```r
adulthood_raw_codebook %>%
  select(`Question text`)
```

```
# A tibble: 328 × 1
   `Question text`
   <chr>
 1 Start Date
 2 End Date
 3 Response Type
 4 Progress
 5 Duration (in seconds)
 6 Finished
 7 Recorded Date
 8 Response ID
 9 Recipient Last Name
10 Recipient First Name
# … with 318 more rows
```

The easiest way to solve this is with the `clean_names()` function from the `{janitor}` package.

Messy column names (II)

But let’s understand how these two datasets relate to one another.

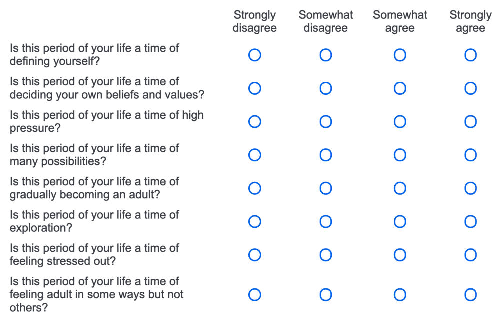
The `Variable Name` column of `adulthood_raw_codebook` contains the exact column names from `adulthood_raw_data`.

If we clean up the names of `adulthood_raw_data` these will no longer matchup. So in this instance let’s only clean the codebook column names:

```r
adulthood_raw_codebook <- read_excel("data/emerging-adulthood_codebook.xlsx") %>%
  clean_names()
```

IDEA Questions (I)

Let’s take a look at the these questions from the survey. Could you suggest a way to find these questions in the codebook?



	Strongly disagree	Somewhat disagree	Somewhat agree	Strongly agree
Is this period of your life a time of defining yourself?	○	○	○	○
Is this period of your life a time of deciding your own beliefs and values?	○	○	○	○
Is this period of your life a time of high pressure?	○	○	○	○
Is this period of your life a time of many possibilities?	○	○	○	○
Is this period of your life a time of gradually becoming an adult?	○	○	○	○
Is this period of your life a time of exploration?	○	○	○	○
Is this period of your life a time of feeling stressed out?	○	○	○	○
Is this period of your life a time of feeling adult in some ways but not others?	○	○	○	○

```r
adulthood_raw_codebook %>%
  filter(str_detect(question_text, "defining yourself"))
```

```
# A tibble: 1 × 6
  variable_name question_text                          x3    respo…¹ x5    surve…²
  <chr>         <chr>                                  <chr> <chr>   <lgl> <chr>
1 IDEA_5        Is this period of your life a time … Thin… 1-stro… NA    "{\"Im…
# … with abbreviated variable names ¹responses, ²survey_question_id
```

```r
idea_responses_raw <- adulthood_raw_data %>%
  select(ResponseId, starts_with("IDEA_"))
idea_responses_raw
```

```
# A tibble: 3,182 × 9
   ResponseId        IDEA_1 IDEA_2 IDEA_3 IDEA_4 IDEA_5 IDEA_6 IDEA_7 IDEA_8
   <chr>              <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
 1 R_BJN3bQqi1zUMid3      3      4      4      3      4      4      4      4
 2 R_2TGbiBXmAtxywsD      4      4      4      4      3      4      4      4
 3 R_12G7bIqN2wB2N65      4      4      4      4      4      4      3      3
 4 R_39pldNoon8CePfP      4      4      3      3      4      4      4      4
 5 R_1QiKb2LdJo1Bhvv      4      4      3      4      3      3      3      4
 6 R_pmwDTZyCyCycXwB       3       4       3       3       4       4       3       2
 7 R_2Quh0h3wxTjZjKP       4       3       4       4       3       4       3       3
 8 R_2CfdmFw1NTliv4e       4       3       3       4       3       2       2       3
 9 R_24kJPxVOxMshN3Q       4       4       3       2       4       4       3       3
10 R_3fv0VeHsW6AvJPk       4       4       2       2       4       4       3       4
# … with 3,172 more rows
```

IDEA Questions (II)

Can you tell me what “4” means in this dataset?

The adulthood_raw_codebook tells us that “4” encodes “Strong agree”.

We now need a way to transform all of these columns at once - can you suggest one?

There are two methods I can think of:

- One method we’ve already used in coding
- One method we’ll be introducing today.

```r
idea_responses_raw %>%
  head() %>%
  gt()
```

ResponseId	IDEA_1	IDEA_2	IDEA_3	IDEA_4	IDEA_5
R_BJN3bQqi1zUMid3	3	4	4	3	4
R_2TGbiBXmAtxywsD	4	4	4	4	3
R_12G7bIqN2wB2N65	4	4	4	4	4
R_39pldNoon8CePfP	4	4	3	3	4
R_1QiKb2LdJo1Bhvv	4	4	3	4	3
R_pmwDTZyCyCycXwB	3	4	3	3	4

IDEA Questions (III)

- Using `across()` to target multiple columns at once.

```r
idea_responses_raw %>%
  mutate(across(starts_with("IDEA_"),
                ~case_when(.x == 1 ~ "Strong disagree",
                           .x == 2 ~ "Disagree",
                           .x == 3 ~ "Agree",
                           .x == 4 ~ "Strong agree")))
```

- Using `pivot_longer()` to transform this from wide to long data.

```r
idea_responses_raw %>%
  pivot_longer(starts_with("IDEA_")) %>%
  mutate(value = case_when(value == 1 ~ "Strong disagree",
                           value == 2 ~ "Disagree",
                           value == 3 ~ "Agree",
                           value == 4 ~ "Strong agree"))
```

> Using `pivot_longer()` has the added benefit of preparing the data for `{ggplot2}`.

Wide vs long data (I)

In a **wide dataset** each variable is stored in a unique column.

Person	Age	Weight	Height
Bob	32	168	180
Alice	24	150	175
Steve	64	144	165

In a **long** dataset each row is a single observation.

Person	Variable	Value
Bob	Age	32
Bob	Weight	168
Bob	Height	180
Alice	Age	24

However, datasets might be *partially wide*. For instance, year is spread across multiple columns.

country	variable	2000	2001	2002
UK	Supermarkets	202	206	230
UK	Shopping malls	40	42	46
US	Supermarkets	305	360	380
US	Shopping malls	80	90	98

In the tidyverse **tidy data**[4] means long data.

country	variable	year	value
UK	Supermarkets	2000	202
UK	Supermarkets	2001	206
UK	Supermarkets	2002	230
UK	Shopping malls	2000	40
UK	Shopping malls	2001	42
country	variable	year	value

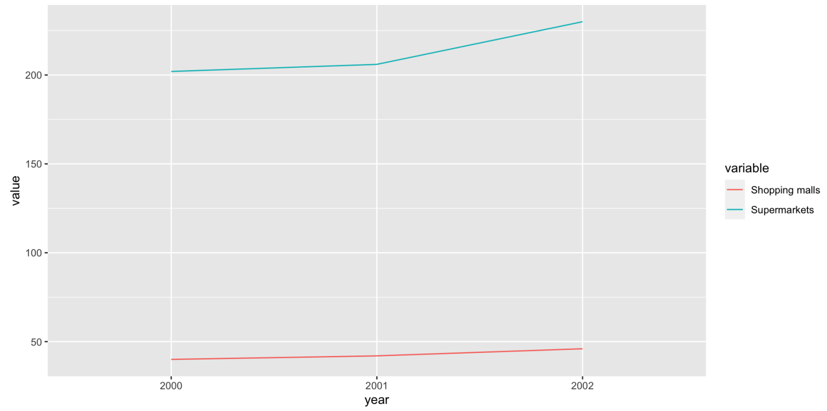
UK	Shopping malls	2002	46
US	Supermarkets	2000	305
US	Supermarkets	2001	360
US	Supermarkets	2002	380
US	Shopping malls	2000	80
US	Shopping malls	2001	90
US	Shopping malls	2002	98

Wide vs long data (II)

The {ggplot2} package requires long data

country	variable	year	value
UK	Supermarkets	2000	202
UK	Supermarkets	2001	206
UK	Supermarkets	2002	230
UK	Shopping malls	2000	40
UK	Shopping malls	2001	42
UK	Shopping malls	2002	46
US	Supermarkets	2000	305
US	Supermarkets	2001	360
US	Supermarkets	2002	380
US	Shopping malls	2000	80
US	Shopping malls	2001	90
US	Shopping malls	2002	98

```r
long_shops_data %>%
  filter(country == "UK") %>%
  ggplot(aes(x = year,
             y = value,
             group = variable,
             color = variable)) +
  geom_line()
```



pivot_wider() and pivot_longer()

The `pivot_wider()` and `pivot_longer()` functions are for transforming data to long format and wide format, respectively.

```r
tribble(
  ~Person, ~Age, ~Weight, ~Height,
    "Bob",  32L,    168L,    180L,
  "Alice",  24L,    150L,    175L,
  "Steve",  64L,    144L,    165L
  ) %>%
  pivot_longer(cols = Age:Height) %>%
  gt()
```

Person	name	value
Bob	Age	32
Bob	Weight	168
Bob	Height	180
Alice	Age	24
Alice	Weight	150
Alice	Height	175
Steve	Age	64
Steve	Weight	144
Steve	Height	165

Note that we can use **any** of the tidy selection functions to target our columns.

> Pre 2020 there were `spread()` and `gather()`. These functions are still in `{tidyr}` but are considered superceded by the `pivot_*()` functions.

IDEA Questions (IV)

We can now transform our actual dataset into long format as follows:

```r
idea_responses_raw %>%
  pivot_longer(starts_with("IDEA_"))
```

The remaining step is to use the case_when() function [which is newly introduced here]

```r
idea_responses_long <- idea_responses_raw %>%
  pivot_longer(starts_with("IDEA_")) %>%
  mutate(value = case_when(value == 1 ~ "Strong disagree",
                           value == 2 ~ "Disagree",
                           value == 3 ~ "Agree",
                           value == 4 ~ "Strong agree"))

idea_responses_long %>%
  head() %>%
  gt()
```

ResponseId	name	value
R_BJN3bQqi1zUMid3	IDEA_1	Agree
R_BJN3bQqi1zUMid3	IDEA_2	Strong agree
R_BJN3bQqi1zUMid3	IDEA_3	Strong agree
R_BJN3bQqi1zUMid3	IDEA_4	Agree
R_BJN3bQqi1zUMid3	IDEA_5	Strong agree
R_BJN3bQqi1zUMid3	IDEA_6	Strong agree

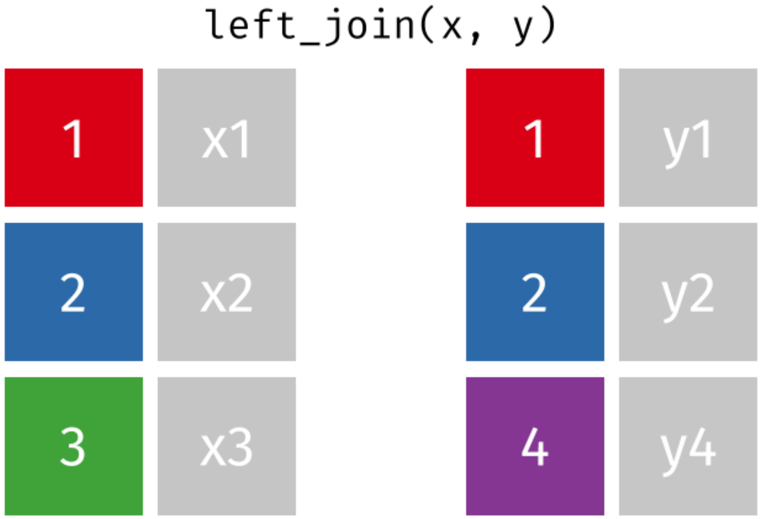
IDEA Questions (V)

We need to match up the question codes with the actual questions.

To achieve this we're going to use a **mutating join**.

It's worthwhile mentioning this is a skill you would use in SQL.

> If you're comfortable doing this then you'll be comfortable with basic SQL.



Source:
https://www.garrickadenbuie.com/project/tidyex

IDEA Questions (VI)

```r
idea_responses_long %>%
  head() %>%
  gt()
```

ResponseId	name	value
R_BJN3bQqi1zUMid3	IDEA_1	Agree
R_BJN3bQqi1zUMid3	IDEA_2	Strong agree
R_BJN3bQqi1zUMid3	IDEA_3	Strong agree
R_BJN3bQqi1zUMid3	IDEA_4	Agree
R_BJN3bQqi1zUMid3	IDEA_5	Strong agree
R_BJN3bQqi1zUMid3	IDEA_6	Strong agree

Let’s extract the variable names and labels from the codebook.

Note that the column names in these two datasets are **different**.

```r
idea_question_labels <- adulthood_raw_codebook %>%
  filter(str_detect(variable_name, "IDEA_")) %>%
  select(variable_name, question_text)

idea_question_labels %>%
  gt()
```

variable_name	question_text
IDEA_1	Is this period of your life a time of many possibilities?
IDEA_2	Is this period of your life a time of exploration?
IDEA_3	Is this period of your life a time of feeling stressed out?
IDEA_4	Is this period of your life a time of high pressure?
IDEA_5	Is this period of your life a time of defining yourself?
IDEA_6	Is this period of your life a time of deciding your own beliefs and values?
IDEA_7	Is this period of your life a time of feeling adult in some ways but not others?
IDEA_8	Is this period of your life a time of gradually becoming an adult?

IDEA Questions (VII)

Because the column names are different we need to give left_join a little help:

```r
idea_responses_clean <- idea_responses_long %>%
  left_join(idea_question_labels,
            by = c("name" = "variable_name"))

idea_responses_clean %>%
  head() %>%
  gt()
```

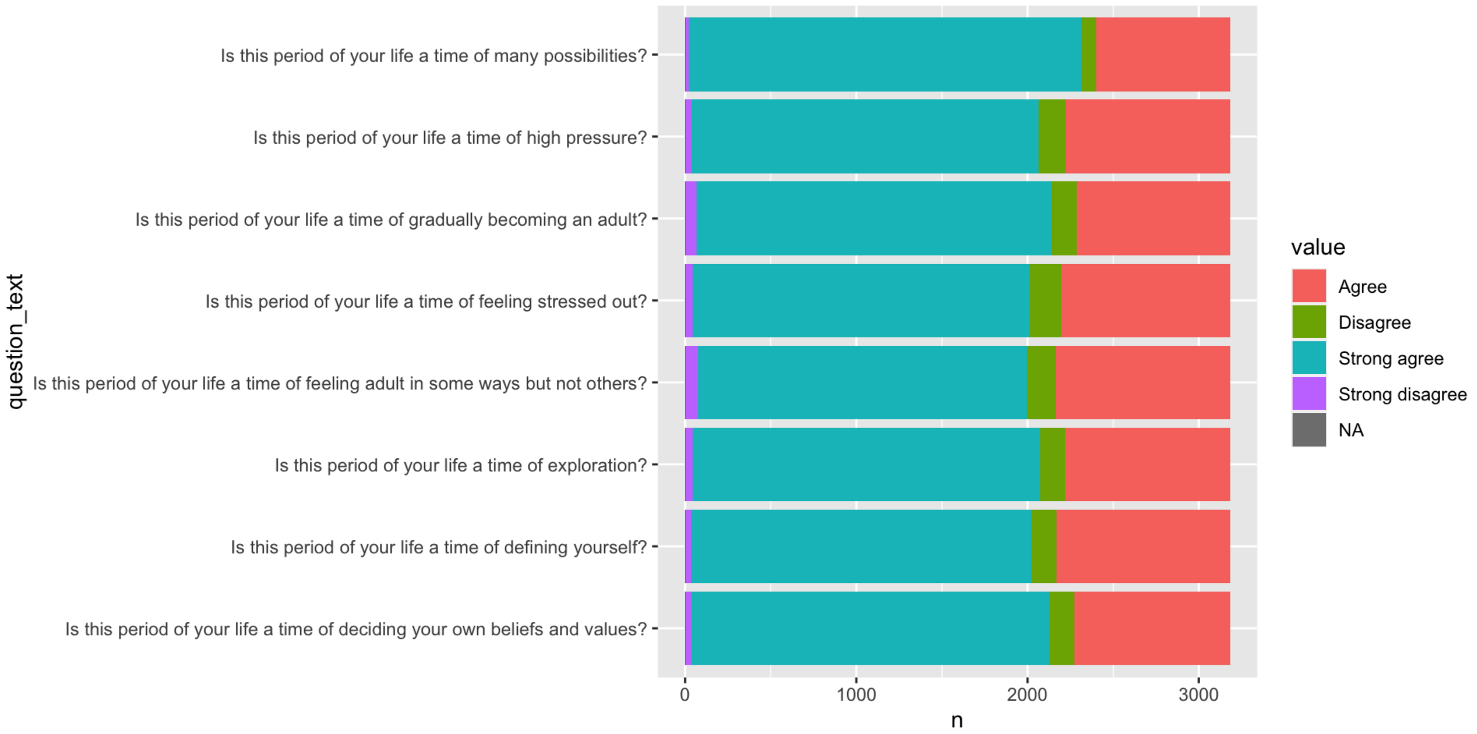
ResponseId	name	value	question_text
R_BJN3bQqi1zUMid3	IDEA_1	Agree	Is this period of your life a time of many possibilities?
R_BJN3bQqi1zUMid3	IDEA_2	Strong agree	Is this period of your life a time of exploration?
R_BJN3bQqi1zUMid3	IDEA_3	Strong agree	Is this period of your life a time of feeling stressed out?
R_BJN3bQqi1zUMid3	IDEA_4	Agree	Is this period of your life a time of high pressure?
R_BJN3bQqi1zUMid3	IDEA_5	Strong agree	Is this period of your life a time of defining yourself?
R_BJN3bQqi1zUMid3	IDEA_6	Strong agree	Is this period of your life a time of deciding your own beliefs and values?

> … so why did we do that all?! We can now use count() to tally responses per question_text and visualise the results.

IDEA Questions (VIII)

Finally we can visualise the responses…. right?!

```
idea_responses_clean %>%
  count(question_text, value) %>%
  ggplot(aes(x = n,
             y = question_text,
             fill = value)) +
  geom_col()
```

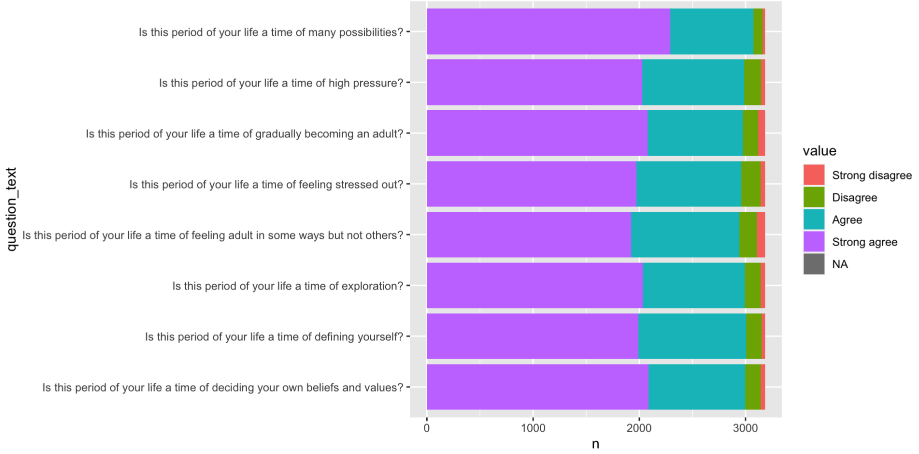


IDEA Questions (IX)

We need to use fct_relevel() to set the canonical order of the factor:

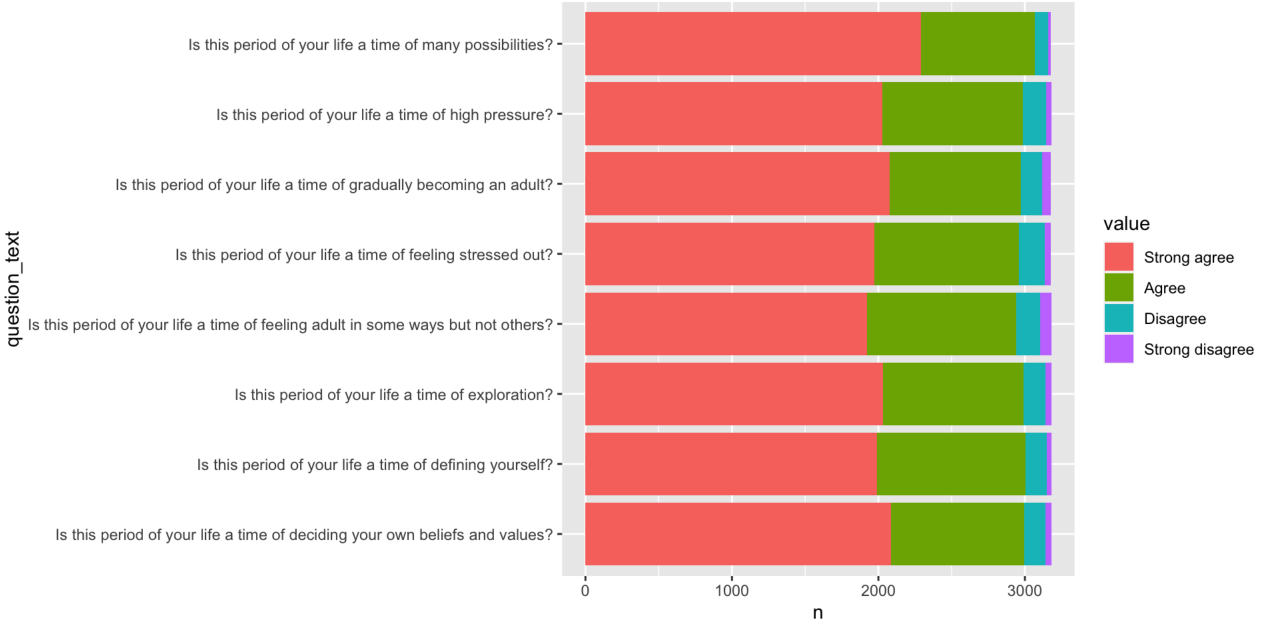
- Which part of the chart do we need to target to change the order of the fill colours?
- Which part of the chart do we need to target to change the order of the legend?

```
order_agree_responses <- c("Strong disagree", "Disagree", "Agree", "Strong agree")

idea_responses_clean %>%
  count(question_text, value) %>%
  mutate(value = fct_relevel(value, order_agree_responses)) %>%
  ggplot(aes(x = n,
             y = question_text,
             fill = value)) +
  geom_col()
```



IDEA Questions (X)

```r
order_agree_responses <- c("Strong disagree", "Disagree", "Agree", "Strong agree")

idea_responses_clean %>%
  drop_na(value) %>%
  count(question_text, value) %>%
  mutate(value = fct_relevel(value, order_agree_responses)) %>%
  ggplot(aes(x = n,
             y = question_text,
             fill = value)) +
  geom_col() +
  scale_fill_discrete(direction = -1) +
  guides(fill = guide_legend(reverse = TRUE))
```



📖 Exercise: Social media questions

SLIDE 1 OF 1

Follow this same process to visualise the responses to the “Social media” questions in the same survey.

> Let’s take 20mins for this

Please note that I’ve not created this chart myself - I’ll run through the same process later during this week’s material.

📝 Task: Obtain Commercial Access to Health Data

SLIDE 1 OF 2

1. Register for a FREE UK Data Service account - beta.ukdataservice.ac.uk/myaccount/login
2. Navigate to the access data page for the dataset - beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=8049
3. Download the SPSS dataset

> It’s very useful to learn how to deal with SPSS datasets now

1. Unzip the dataset and add the folder to the data folder in your RStudio project

📝 Task: Obtain Commercial Access to Health Data

SLIDE 2 OF 2

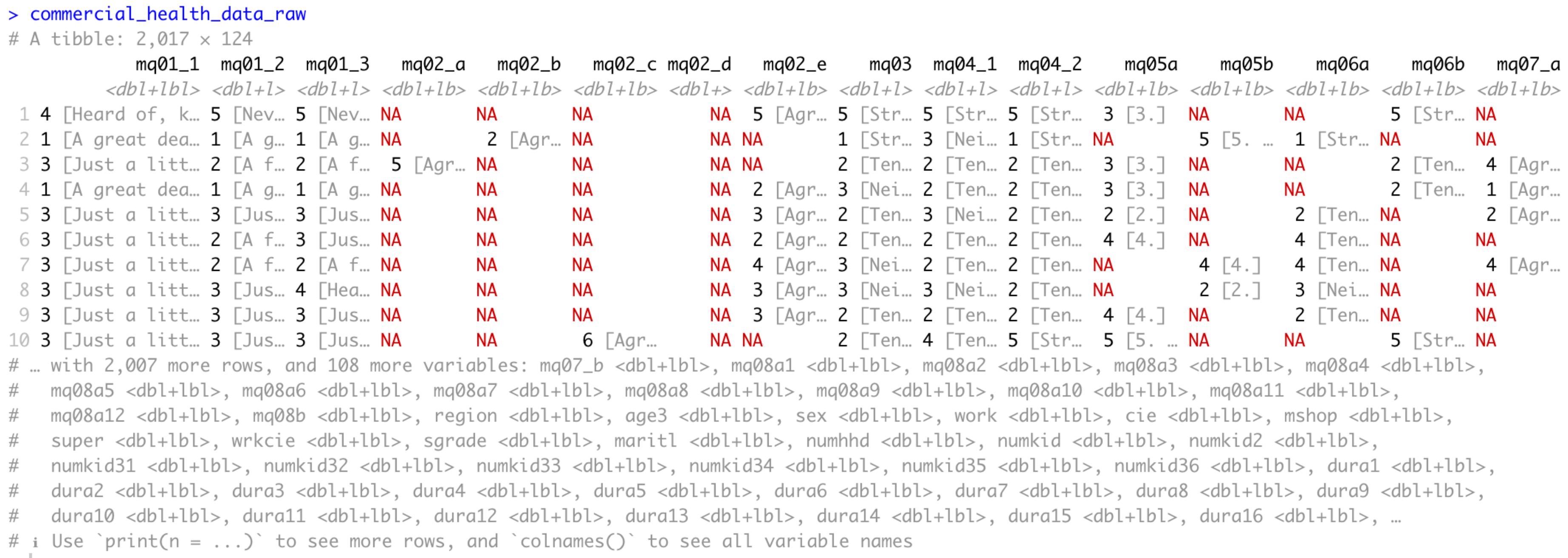
1. Open up the `commercial-access-to-health-data.Rmd` RMarkdown document
2. Load the `{haven}` and `{tidyverse}` package

Read in the dataset (it has a really long file path!)

```
commercial_health_data_raw <- read_spss("data/UKDA-8049-spss 2/spss/spss19/health_data_attitudes_spss_final.sav")
```

Tibbles are great (I)

We’ve seen before that tibbles are augmented data.frame - they can have additional attributes and print more prettily.



```
> commercial_health_data_raw
# A tibble: 2,017 × 124
            mq01_1  mq01_2  mq01_3   mq02_a   mq02_b   mq02_c mq02_d   mq02_e    mq03  mq04_1  mq04_2    mq05a    mq05b    mq06a    mq06b   mq07_a
         <dbl+lbl> <dbl+l> <dbl+l> <dbl+lb> <dbl+lb> <dbl+lb> <dbl+> <dbl+lb> <dbl+l> <dbl+l> <dbl+l> <dbl+lb> <dbl+lb> <dbl+lb> <dbl+lb> <dbl+lb>
 1 4 [Heard of, k… 5 [Nev… 5 [Nev… NA       NA       NA           NA  5 [Agr… 5 [Str… 5 [Str… 5 [Str…  3 [3.]  NA       NA        5 [Str… NA      
 2 1 [A great dea… 1 [A g… 1 [A g… NA        2 [Agr… NA           NA NA        1 [Str… 3 [Nei… 1 [Str… NA        5 [5. …  1 [Str… NA       NA      
 3 3 [Just a litt… 2 [A f… 2 [A f…  5 [Agr… NA       NA           NA NA        2 [Ten… 2 [Ten… 2 [Ten…  3 [3.]  NA       NA        2 [Ten…  4 [Agr…
 4 1 [A great dea… 1 [A g… 1 [A g… NA       NA       NA           NA  2 [Agr… 3 [Nei… 2 [Ten… 2 [Ten…  3 [3.]  NA       NA        2 [Ten…  1 [Agr…
 5 3 [Just a litt… 3 [Jus… 3 [Jus… NA       NA       NA           NA  3 [Agr… 2 [Ten… 3 [Nei… 2 [Ten…  2 [2.]  NA        2 [Ten… NA        2 [Agr…
 6 3 [Just a litt… 2 [A f… 3 [Jus… NA       NA       NA           NA  2 [Agr… 2 [Ten… 2 [Ten… 2 [Ten…  4 [4.]  NA        4 [Ten… NA       NA      
 7 3 [Just a litt… 2 [A f… 2 [A f… NA       NA       NA           NA  4 [Agr… 3 [Nei… 2 [Ten… 2 [Ten… NA        4 [4.]   4 [Ten… NA        4 [Agr…
 8 3 [Just a litt… 3 [Jus… 4 [Hea… NA       NA       NA           NA  3 [Agr… 3 [Nei… 3 [Nei… 2 [Ten… NA        2 [2.]   3 [Nei… NA       NA      
 9 3 [Just a litt… 3 [Jus… 3 [Jus… NA       NA       NA           NA  3 [Agr… 2 [Ten… 2 [Ten… 2 [Ten…  4 [4.]  NA        2 [Ten… NA       NA      
10 3 [Just a litt… 3 [Jus… 3 [Jus… NA       NA        6 [Agr…      NA NA        2 [Ten… 4 [Ten… 5 [Str…  5 [5. … NA       NA        5 [Str… NA      
# … with 2,007 more rows, and 108 more variables: mq07_b <dbl+lbl>, mq08a1 <dbl+lbl>, mq08a2 <dbl+lbl>, mq08a3 <dbl+lbl>, mq08a4 <dbl+lbl>,
#   mq08a5 <dbl+lbl>, mq08a6 <dbl+lbl>, mq08a7 <dbl+lbl>, mq08a8 <dbl+lbl>, mq08a9 <dbl+lbl>, mq08a10 <dbl+lbl>, mq08a11 <dbl+lbl>,
#   mq08a12 <dbl+lbl>, mq08b <dbl+lbl>, region <dbl+lbl>, age3 <dbl+lbl>, sex <dbl+lbl>, work <dbl+lbl>, cie <dbl+lbl>, mshop <dbl+lbl>,
#   super <dbl+lbl>, wrkcie <dbl+lbl>, sgrade <dbl+lbl>, maritl <dbl+lbl>, numhhd <dbl+lbl>, numkid <dbl+lbl>, numkid2 <dbl+lbl>,
#   numkid31 <dbl+lbl>, numkid32 <dbl+lbl>, numkid33 <dbl+lbl>, numkid34 <dbl+lbl>, numkid35 <dbl+lbl>, numkid36 <dbl+lbl>, dura1 <dbl+lbl>,
#   dura2 <dbl+lbl>, dura3 <dbl+lbl>, dura4 <dbl+lbl>, dura5 <dbl+lbl>, dura6 <dbl+lbl>, dura7 <dbl+lbl>, dura8 <dbl+lbl>, dura9 <dbl+lbl>,
#   dura10 <dbl+lbl>, dura11 <dbl+lbl>, dura12 <dbl+lbl>, dura13 <dbl+lbl>, dura14 <dbl+lbl>, dura15 <dbl+lbl>, dura16 <dbl+lbl>, …
# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

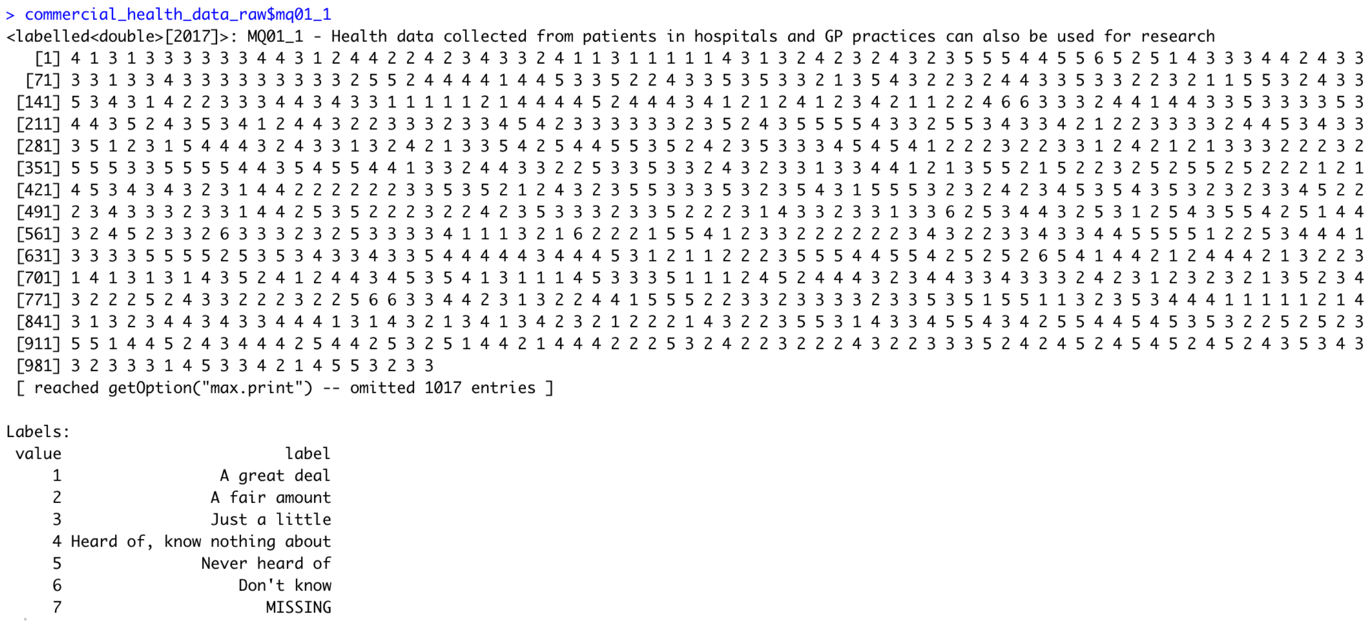
Tibbles are great (II)

The {haven} package creates a special “labelled” column class which contains:

- The question label
- The question values
- The question value labels

This is slightly confusing, but important:

- singular “label” means the question text (or a shortened version of it).
- plural “labels” means the question response label.



```
> commercial_health_data_raw$mq01_1
<labelled<double>[2017]>: MQ01_1 - Health data collected from patients in hospitals and GP practices can also be used for research
   [1] 4 1 3 1 3 3 3 3 3 3 4 4 3 1 2 4 4 2 2 4 2 3 4 3 3 2 4 1 1 3 1 1 1 1 1 4 3 1 3 2 4 2 3 2 4 3 2 3 5 5 5 4 4 5 5 6 5 2 5 1 4 3 3 3 4 4 2 4 3 3
  [71] 3 3 1 3 3 4 3 3 3 3 3 3 3 3 3 2 5 5 2 4 4 4 4 1 4 4 5 3 3 5 2 2 4 3 3 5 3 5 3 3 2 1 3 5 4 3 2 2 3 2 4 4 3 3 5 3 3 2 2 3 2 1 1 5 5 3 2 4 3 3
 [141] 5 3 4 3 1 4 2 2 3 3 3 4 4 3 4 3 3 1 1 1 1 1 2 1 4 4 4 4 5 2 4 4 4 3 4 1 2 1 2 4 1 2 3 4 2 1 1 2 2 4 6 6 3 3 3 2 4 4 1 4 4 3 3 5 3 3 3 3 5 3
 [211] 4 4 3 5 2 4 3 5 3 4 1 2 4 4 3 2 2 3 3 3 2 3 3 4 5 4 2 3 3 3 3 3 3 2 3 5 2 4 3 5 5 5 5 4 3 3 2 5 5 3 4 3 3 4 2 1 2 2 3 3 3 3 2 4 4 5 3 4 3 3
 [281] 3 5 1 2 3 1 5 4 4 4 3 2 4 3 3 1 3 2 4 2 1 3 3 5 4 2 5 4 4 5 5 3 5 2 4 2 3 5 3 3 3 4 5 4 5 4 1 2 2 2 3 2 2 3 3 1 2 4 2 1 2 1 3 3 3 2 2 2 3 2
 [351] 5 5 5 3 3 5 5 5 5 4 4 3 5 4 5 5 4 4 1 3 3 2 4 4 3 3 2 2 5 3 3 5 3 3 2 4 3 2 3 3 1 3 3 4 4 1 2 1 3 5 5 2 1 5 2 2 3 2 5 2 5 5 2 5 2 2 2 1 2 1
 [421] 4 5 3 4 3 4 3 2 3 1 4 4 2 2 2 2 2 2 3 3 5 3 5 2 1 2 4 3 2 3 5 5 3 3 3 5 3 2 3 5 4 3 1 5 5 5 3 2 3 2 4 2 3 4 5 3 5 4 3 5 3 2 3 2 3 3 4 5 2 2
 [491] 2 3 4 3 3 3 2 3 3 1 4 4 2 5 3 5 2 2 2 3 2 2 4 2 3 5 3 3 3 2 3 3 5 2 2 2 3 1 4 3 3 2 3 3 1 3 3 6 2 5 3 4 4 3 2 5 3 1 2 5 4 3 5 5 4 2 5 1 4 4
 [561] 3 2 4 5 2 3 3 2 6 3 3 3 2 3 2 5 3 3 3 3 4 1 1 1 3 2 1 6 2 2 2 1 5 5 4 1 2 3 3 2 2 2 2 2 2 3 4 3 2 2 3 3 4 3 3 4 4 5 5 5 5 1 2 2 5 3 4 4 4 1
 [631] 3 3 3 3 5 5 5 5 2 5 3 5 3 4 3 3 4 3 3 5 4 4 4 4 4 3 4 4 4 5 3 1 2 1 1 2 2 2 3 5 5 5 4 4 5 5 4 2 5 2 5 2 6 5 4 1 4 4 2 1 2 4 4 4 2 1 3 2 2 3
 [701] 1 4 1 3 1 3 1 4 3 5 2 4 1 2 4 4 3 4 5 3 5 4 1 3 1 1 1 4 5 3 3 3 5 1 1 1 2 4 5 2 4 4 4 3 2 3 4 4 3 3 4 3 3 3 2 4 2 3 1 2 3 2 3 2 1 3 5 2 3 4
 [771] 3 2 2 2 5 2 4 3 3 2 2 2 3 2 2 5 6 6 3 3 4 4 2 3 1 3 2 2 4 4 1 5 5 5 2 2 3 3 2 3 3 3 3 2 3 3 5 3 5 1 5 5 1 1 3 2 3 5 3 4 4 4 1 1 1 1 1 2 1 4
 [841] 3 1 3 2 3 4 4 3 4 3 3 4 4 4 1 3 1 4 3 2 1 3 4 1 3 4 2 3 2 1 2 2 2 1 4 3 2 2 3 5 5 3 1 4 3 3 4 5 5 4 3 4 2 5 5 4 4 5 4 5 3 5 3 2 2 5 2 5 2 3
 [911] 5 5 1 4 4 5 2 4 3 4 4 4 2 5 4 4 2 5 3 2 5 1 4 4 2 1 4 4 4 2 2 2 5 3 2 4 2 2 3 2 2 2 4 3 2 2 3 3 3 5 2 4 2 4 5 2 4 5 4 5 2 4 5 2 4 3 5 3 4 3
 [981] 3 2 3 3 3 1 4 5 3 3 4 2 1 4 5 5 3 2 3 3
 [ reached getOption("max.print") -- omitted 1017 entries ]

Labels:
 value                          label
     1                   A great deal
     2                  A fair amount
     3                  Just a little
     4 Heard of, know nothing about
     5                  Never heard of
     6                     Don't know
     7                        MISSING
```

Extracting the question label from {haven} output (I)

We have two different ways to extract labels from {haven} output

- Programming with purrr::attr_getter()

```
commercial_health_data_raw %>%
  map_chr(attr_getter("label"))
```

- Using sjlabelled::get_label()

```
sjlabelled::get_label(commercial_health_data_raw)
```

Unfortunately, the programming solution only works if all columns are labelled.

I'd recommend using the `{sjlabelled}` package exclusively for extracting question labels from SPSS. It does provide more tools but they're not compatible with `{haven}` and also cause conflicts.

Extracting the question label from {haven} output (II)

The `sjlabelled::get_label()` function generates a named list. We can convert this into a tibble with `enframe()`

```
commercial_health_data_qs_raw <- sjlabelled::get_label(commercial_health_data_raw) %>%
  enframe() %>%
  rename(question_text = value)
commercial_health_data_qs_raw
```

```
# A tibble: 124 × 2
   name    question_text
   <chr>   <chr>
 1 mq01_1 MQ01_1 - Health data collected from patients in hospitals and GP prac…
 2 mq01_2 MQ01_2 - Health data collected from patients in hospitals and GP prac…
 3 mq01_3 MQ01_3 - Health data collected from patients in hospitals and GP prac…
 4 mq02_a MQ02_A - As you may know, the NHS and other health services collect d…
 5 mq02_b MQ02_B - As you may know, the NHS and other health services collect d…
 6 mq02_c MQ02_C - As you may know, the NHS and other health services collect d…
 7 mq02_d MQ02_D - As you may know, the NHS and other health services collect d…
 8 mq02_e MQ02_E - As you may know, the NHS and other health services collect d…
 9 mq03   MQ03 - To what extent, if at all, would you support your health data …
10 mq04_1 MQ04_1 - To what extent do you agree or disagree with the following s…
# … with 114 more rows
```

> Can you help me write some code to remove the question label from the value column?

Extracting the question label from {haven} output (II)

This is one of many ways to tidy up this data:

```r
commercial_health_data_qs <- commercial_health_data_qs_raw %>%
  mutate(question_text = str_remove(question_text, toupper(name)),
         question_text = str_remove(question_text, " - "),
         question_text = str_remove(question_text, "MQ08A"))
commercial_health_data_qs
```

```
# A tibble: 124 × 2
   name    question_text
   <chr>   <chr>
 1 mq01_1 Health data collected from patients in hospitals and GP practices can…
 2 mq01_2 Health data collected from patients in hospitals and GP practices can…
 3 mq01_3 Health data collected from patients in hospitals and GP practices can…
 4 mq02_a As you may know, the NHS and other health services collect data about…
 5 mq02_b As you may know, the NHS and other health services collect data about…
 6 mq02_c As you may know, the NHS and other health services collect data about…
 7 mq02_d As you may know, the NHS and other health services collect data about…
 8 mq02_e As you may know, the NHS and other health services collect data about…
 9 mq03   To what extent, if at all, would you support your health data being a…
10 mq04_1 To what extent do you agree or disagree with the following statements?
# … with 114 more rows
```

Converting labelled columns to factors

To convert all labelled columns to factors we can use across()

```
commercial_health_data_factors <- commercial_health_data_raw %>%
  mutate(across(where(is.labelled), ~as_factor(.x)))
commercial_health_data_factors
```

```
# A tibble: 2,017 × 124
   mq01_1   mq01_2 mq01_3 mq02_a mq02_b mq02_c mq02_d mq02_e mq03  mq04_1 mq04_2
   <fct>    <fct>  <fct>  <fct>  <fct>  <fct>  <fct>  <fct>  <fct> <fct>  <fct> 
 1 Heard o… Never… Never… <NA>   <NA>   <NA>   <NA>   Agree… Stro… Stron… Stron…
 2 A great… A gre… A gre… <NA>   Agree… <NA>   <NA>   <NA>   Stro… Neith… Stron…
 3 Just a … A fai… A fai… Agree… <NA>   <NA>   <NA>   <NA>   Tend… Tend … Tend …
 4 A great… A gre… A gre… <NA>   <NA>   <NA>   <NA>   Agree… Neit… Tend … Tend …
 5 Just a … Just … Just … <NA>   <NA>   <NA>   <NA>   Agree… Tend… Neith… Tend …
 6 Just a … A fai… Just … <NA>   <NA>   <NA>   <NA>   Agree… Tend… Tend … Tend …
 7 Just a … A fai… A fai… <NA>   <NA>   <NA>   <NA>   Agree… Neit… Tend … Tend …
 8 Just a … Just … Heard… <NA>   <NA>   <NA>   <NA>   Agree… Neit… Neith… Tend …
 9 Just a … Just … Just … <NA>   <NA>   <NA>   <NA>   Agree… Tend… Tend … Tend …
10 Just a … Just … Just … <NA>   <NA>   Agree… <NA>   <NA>   Tend… Tend … Stron…
# … with 2,007 more rows, and 113 more variables: mq05a <fct>, mq05b <fct>,
#   mq06a <fct>, mq06b <fct>, mq07_a <fct>, mq07_b <fct>, mq08a1 <fct>,
#   mq08a2 <fct>, mq08a3 <fct>, mq08a4 <fct>, mq08a5 <fct>, mq08a6 <fct>,
#   mq08a7 <fct>, mq08a8 <fct>, mq08a9 <fct>, mq08a10 <fct>, mq08a11 <fct>,
#   mq08a12 <fct>, mq08b <fct>, region <fct>, age3 <fct>, sex <fct>,
#   work <fct>, cie <fct>, mshop <fct>, super <fct>, wrkcie <fct>,
#   sgrade <fct>, maritl <fct>, numhhd <fct>, numkid <fct>, numkid2 <fct>, …
```

> Notice how we don’t have a respondent ID column?

Add respondent ID

The `row_number()` function gives us a neat way to add a respondent ID.

However, it’s not necessarily that clever a solution in terms of anonymisation.

```r
commercial_health_data_clean <- commercial_health_data_factors %>%
  mutate(respondent_id = row_number()) %>%
  relocate(respondent_id)
commercial_health_data_clean
```

```
# A tibble: 2,017 × 125
   respon…¹ mq01_1 mq01_2 mq01_3 mq02_a mq02_b mq02_c mq02_d mq02_e mq03  mq04_1
      <int> <fct>  <fct>  <fct>  <fct>  <fct>  <fct>  <fct>  <fct>  <fct> <fct>
 1        1 Heard… Never… Never… <NA>   <NA>   <NA>   <NA>   Agree… Stro… Stron…
 2        2 A gre… A gre… A gre… <NA>   Agree… <NA>   <NA>   <NA>   Stro… Neith…
 3        3 Just … A fai… A fai… Agree… <NA>   <NA>   <NA>   <NA>   Tend… Tend …
 4        4 A gre… A gre… A gre… <NA>   <NA>   <NA>   <NA>   Agree… Neit… Tend …
 5        5 Just … Just … Just … <NA>   <NA>   <NA>   <NA>   Agree… Tend… Neith…
 6        6 Just … A fai… Just … <NA>   <NA>   <NA>   <NA>   Agree… Tend… Tend …
 7        7 Just … A fai… A fai… <NA>   <NA>   <NA>   <NA>   Agree… Neit… Tend …
 8        8 Just … Just … Heard… <NA>   <NA>   <NA>   <NA>   Agree… Neit… Neith…
 9        9 Just … Just … Just … <NA>   <NA>   <NA>   <NA>   Agree… Tend… Tend …
10       10 Just … Just … Just … <NA>   <NA>   Agree… <NA>   <NA>   Tend… Tend …
# … with 2,007 more rows, 114 more variables: mq04_2 <fct>, mq05a <fct>,
#   mq05b <fct>, mq06a <fct>, mq06b <fct>, mq07_a <fct>, mq07_b <fct>,
#   mq08a1 <fct>, mq08a2 <fct>, mq08a3 <fct>, mq08a4 <fct>, mq08a5 <fct>,
#   mq08a6 <fct>, mq08a7 <fct>, mq08a8 <fct>, mq08a9 <fct>, mq08a10 <fct>,
#   mq08a11 <fct>, mq08a12 <fct>, mq08b <fct>, region <fct>, age3 <fct>,
#   sex <fct>, work <fct>, cie <fct>, mshop <fct>, super <fct>, wrkcie <fct>,
#   sgrade <fct>, maritl <fct>, numhhd <fct>, numkid <fct>, numkid2 <fct>, …
```

Commercial Health Data Survey Q4 (I)

I’d like you to extract the columns from the survey data that correspond

Q4. To what extent do you agree or disagree with the following statements?

	“My health data currently has financial value to others in that it can be used to save or make them money.”	“My health data currently has a value to society in that it can be used to help improve things for people other than me.”
Base:	*All respondents (2,017)* %	*All respondents (2,017)* %
Strongly agree	15	28
Tend to agree	35	40
Neither agree nor disagree	25	18
Tend to disagree	12	7
Strongly disagree	9	5
Don’t know	3	3
Agree	50	67
Disagree	21	12

Commercial Health Data Survey Q4 (II)

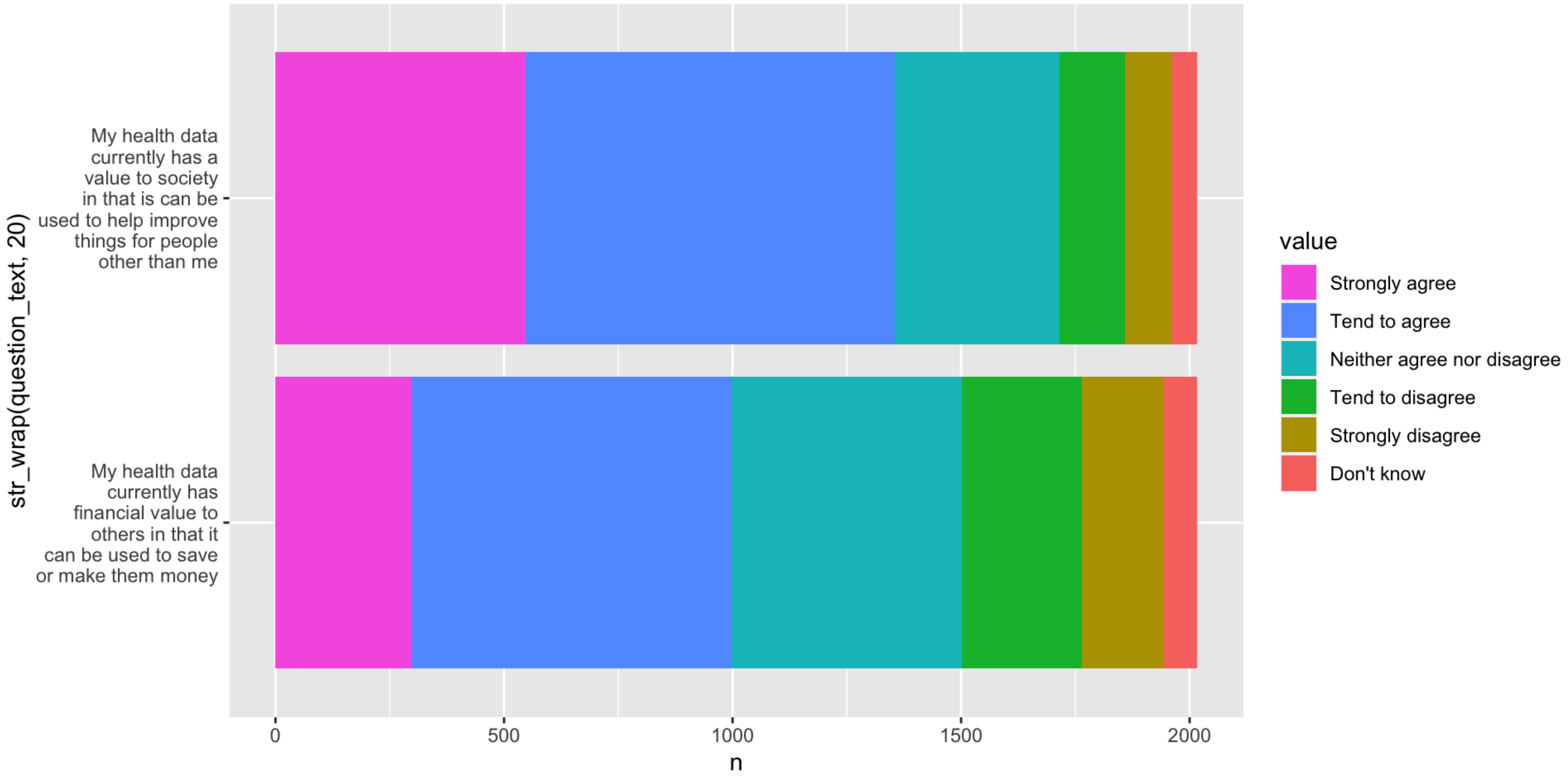
What do we need to do to this data so that we can tally responses and visualise it with {ggplot2}?

```r
commercial_health_data_clean %>%
  select(respondent_id, starts_with("mq04"))
```

```
# A tibble: 2,017 × 3
   respondent_id mq04_1                     mq04_2
           <int> <fct>                      <fct>
 1             1 Strongly disagree          Strongly disagree
 2             2 Neither agree nor disagree Strongly agree
 3             3 Tend to agree              Tend to agree
 4             4 Tend to agree              Tend to agree
 5             5 Neither agree nor disagree Tend to agree
 6             6 Tend to agree              Tend to agree
 7             7 Tend to agree              Tend to agree
 8             8 Neither agree nor disagree Tend to agree
 9             9 Tend to agree              Tend to agree
10            10 Tend to disagree           Strongly disagree
# … with 2,007 more rows
```

Commercial Health Data Survey Q4 (III)

Let’s make this chart:



How wide is too wide?

```r
tweetrmd::tweet_embed("https://twitter.com/charliejhadley/status/1522559488284413954?ref_src=twsrc%5Etfw")
```

Buffy ratings

In what ways is this dataset wide?

- Ratings are split across multiple columns
- Should we include `vox ep rank` in ratings?!
- In principle we could combine:
  - votes
  - views
  - vox ep rank

```r
buffy_raw <- read_csv(here::here("static", "datasets", "buffy", "buffy_data.csv"))
buffy_raw
```

```
# A tibble: 144 × 15
   no overal…¹ season no in…² title direc…³ writer air_d…⁴ views…⁵ imdb_…⁶ votes
         <dbl>  <dbl>   <dbl> <chr> <chr>   <chr>  <chr>     <dbl>   <dbl> <dbl>
 1           1      1       1 Welc… Charle… Joss … 3/10/1…    4.59     8    4548
 2           2      1       2 The … John T… Joss … 3/10/1…    4.59     7.8  3952
 3           3      1       3 Witch Stephe… Dana … 3/17/1…    4.63     7.7  3940
 4           4      1       4 Teac… Bruce … David… 3/24/1…    2.98     6.9  3800
 5           5      1       5 Neve… David … Rob D… 3/31/1…    4.09     7.4  3611
 6           6      1       6 The … Bruce … Matt … 4/7/19…    3.42     7.4  3770
 7           7      1       7 Angel Scott … David… 4/14/1…    3.39     8.5  3918
 8           8      1       8 I, R… Stephe… Ashle… 4/28/1…    2.47     6.7  3629
 9           9      1       9 The … Ellen … Rob D… 5/5/19…    2.56     7.7  3666
10          10      1      10 Nigh… Bruce … Joss … 5/12/1…    3.47     8.2  3614
# … with 134 more rows, 5 more variables: plot <chr>, runtime <dbl>,
#   deathcount <dbl>, `neilsen rating` <chr>, `vox ep rank` <dbl>, and
```

Long enough for what you need

Tidy data is a useful concept for wrangling, modelling and data visualisation[4].

But it’s not something to conform to religiously.

You might want to keep some width to your data to make it easy to quickly view.

Wide data might also be more appropriate if visualising your data with tables.

Other forms of untidy data

Multiple pieces of data in one cell

Sometimes a single column contains multiple variables.

This is often the case in poorly designed “where do you live?” questions:

```r
location_data <- tribble(
  ~id, ~address,
  1, "Las Vegas, USA",
  2, "Bristol, UK",
  3, "Kassala, Sudan"
)
location_data
```

```
# A tibble: 3 × 2
     id address
  <dbl> <chr>
1     1 Las Vegas, USA
2     2 Bristol, UK
3     3 Kassala, Sudan
```

You might also ask respondents to “select all that apply”

```r
device_ownership <- tribble(
  ~name, ~devices_owned,
  "Charlie", "Smart TV, Cell phone",
  "Mohammad", "Cell phone",
  "Christina", "Smart TV, Games Console, Cell phone"
)
device_ownership
```

```
# A tibble: 3 × 2
  name      devices_owned
  <chr>     <chr>
1 Charlie   Smart TV, Cell phone
2 Mohammad  Cell phone
3 Christina Smart TV, Games Console, Cell phone
```

📝 Task: Obtain British Election Survey Data

SLIDE 1 OF 2

1. Register for a FREE British Election Survey Data account - britishelectionstudy.com/wp-login.php?action=register
2. Navigate to the access data page for the dataset - britishelectionstudy.com/data-object/2019-british-election-study-post-election-random-probability-survey/
3. Download the SPSS dataset
4. Unzip the dataset and add the folder to the data folder in your RStudio project

📝 Task: Obtain British Election Survey Data

SLIDE 2 OF 2

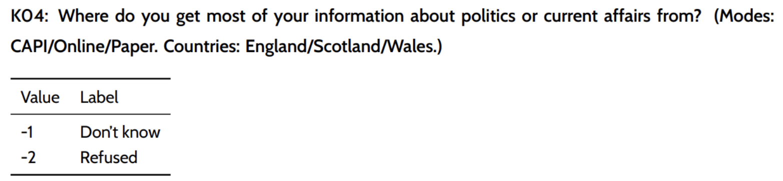
1. Setup the british-election-survey.Rmd for data wrangling
2. Read in the SPSS file

```
# A tibble: 3,946 × 415
   finalser…¹ agency    Y10A Y10B1    Y10B2   Y10B3   Y10B4   Y10B5   a01
        <dbl> <dbl+lb> <dbl> <dbl+lb> <dbl+l> <dbl+l> <dbl+l> <dbl+l> <chr+lbl>
 1      10102  1 [Ips…    NA  NA       NA      NA      NA      NA      "-2" [Ref…
 2      10103 NA           2  1 [Yes]  0 [No]  0 [No]  0 [No]  0 [No]  ""
 3      10105 NA           2  1 [Yes]  0 [No]  0 [No]  0 [No]  0 [No]  ""
 4      10110  1 [Ips…    NA  NA       NA      NA      NA      NA      "-1" [Don…
 5      10111  1 [Ips…    NA  NA       NA      NA      NA      NA      "-1" [Don…
 6      10202 NA           2  1 [Yes]  0 [No]  0 [No]  0 [No]  0 [No]  ""
 7      10206 NA           3  1 [Yes]  0 [No]  0 [No]  0 [No]  0 [No]  ""
 8      10208 NA           2  1 [Yes]  0 [No]  0 [No]  0 [No]  0 [No]  ""
 9      10210 NA           2  1 [Yes]  0 [No]  0 [No]  0 [No]  0 [No]  ""
10      10304 NA           2  1 [Yes]  0 [No]  0 [No]  0 [No]  0 [No]  ""
# … with 3,936 more rows, 406 more variables: a01_code <dbl+lbl>,
#   a02 <dbl+lbl>, a03 <dbl+lbl>, m02_1 <dbl+lbl>, m02_2 <dbl+lbl>,
#   m02_3 <dbl+lbl>, m02_4 <dbl+lbl>, m02_5 <dbl+lbl>, m02_6 <dbl+lbl>,
#   b01 <dbl+lbl>, b02 <dbl+lbl>, b04 <dbl+lbl>, b05 <dbl+lbl>,
#   b0601 <dbl+lbl>, b0602 <dbl+lbl>, b0603 <dbl+lbl>, b0604 <dbl+lbl>,
#   b0605 <dbl+lbl>, b0606 <dbl+lbl>, b0607 <dbl+lbl>, b0608 <dbl+lbl>,
#   b0609 <dbl+lbl>, b0610 <dbl+lbl>, b0611 <dbl+lbl>, b0612 <dbl+lbl>, …
```

Where do people get their information from?

Can you extract the column(s) from the dataset corresponding to this question?

> What can you tell me about this data and question?



KO4: Where do you get most of your information about politics or current affairs from? (Modes: CAPI/Online/Paper. Countries: England/Scotland/Wales.)

Value	Label
-1	Don't know
-2	Refused

Where do people get their information from?

This is an open-ended question that's going to be really messy to handle.

To properly analyse this we might need to use the {tidytext} package for text mining with a tidyverse approach.

But let's see what we can do by pretending it's multiple choice data and using

`separate()` … ?

`separate_rows()` … ?

```r
british_election_data_raw %>%
  select(finalserialno, k04)
```

```
# A tibble: 3,946 × 2
   finalserialno k04
           <dbl> <chr+lbl>
 1         10102 "family"
 2         10103 "-2" [Refused]
 3         10105 "Media- cross referencing and watching
parliamentary debates"
 4         10110 "parents"
 5         10111 "tv radio"
 6         10202 ""
 7         10206 "News, internet and conversation"
 8         10208 ""
 9         10210 "Mail on line \nNews on tv"
10         10304 "t v .     papers.       radio."
# … with 3,936 more rows
```

References

1. Grima, N. *et al.* The importance of urban natural areas and urban ecosystem services during the COVID-19 pandemic. *PLOS ONE* **15**, e0243344 (2020).
2. Eatock, J., Dixon, D. & Young, T. An exploratory survey of current practice in the medical device industry. *Journal of Manufacturing Technology Management* **20**, 218–234 (2009).
3. Presser, S. *et al.* Methods for Testing and Evaluating Survey Questions. *Public Opinion Quarterly* **68**, 109–130 (2004).
4. Wickham, H. Tidy Data. *Journal of Statistical Software* **59**, 1–23 (2014).
5. Fieldhouse, E. *et al.* British Election Study, 2019: Post-Election Random Probability Survey. (2019) doi:10.5255/UKDA-SN-8875-1.
6. NHS England. Writing an effective questionnaire. (2018).
7. Gallup. Why Phone and Web Survey Results Aren't the Same. *Gallup.com* (2018).
8. Sánchez Tomé, R. The impact of mode of data collection on measures of subjective wellbeing. (University of Lausanne, 2018).
9. AAPOR. An Evaluation of 2016 Election Polls in the U.S. *American Association for Public Opinion Research* (2016).
10. Serdar, C. C., Cihan, M., Yücel, D. & Serdar, M. A. Sample size, power and effect size revisited: Simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia Medica* **31**, 010502 (2021).
11. Perneger, T. V., Courvoisier, D. S., Hudelson, P. M. & Gayet-Ageron, A. Sample size for pre-tests of questionnaires. *Quality of Life Research* **24**, 147–151 (2015).
12. Teller, S. Sample size: How many people should take the survey? *on device research* (2014).
13. Jamieson, S. Likert scales: How to (ab)use them. *Medical Education* **38**, 1217–1218 (2004).
14. Batterton, K. A. & Hale, K. N. The Likert Scale What It Is and How To Use It. *Phalanx* **50**, 32–39 (2017).
15. SurveyMonkey. Does adding one more question impact survey completion rate? *SurveyMonkey*.
16. Rhemtulla, M., Savalei, V. & Little, T. D. On the Asymptotic Relative Efficiency of Planned Missingness Designs. *Psychometrika* **81**, 60–89 (2016).
17. Grahe, J. E. *et al.* Emerging Adulthood Measured at Multiple Institutions 2: The Data. *Journal of Open Psychology Data* **6**, 4 (2018).
18. Reifman, A. & Grahe, J. E. Introduction to the Special Issue of Emerging Adulthood. *Emerging Adulthood* **4**, 135–141 (2016).
19. Grahe, J. *et al.* EAMMi2 Public Data. (2022) doi:10.17605/OSF.IO/QTQPB.