Week 5: Surve

Charlotte Hadley

Topics for today

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

This week we're going to be discussing surv

The goals for the lecture section of today

1. Identify what makes survey data messy
2. Correctly identify if data is "long" or "w
3. Understand how to use the {tidyr} pivot fu data
4. Understand how to use the tidyverse for co with survey data

We'll likely continue some of the lecture n

Surveys are an indis
healthcare

There are lots of abso
measures in healthcare

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

Bluutt et hqeusaen tmitehaastuierveesw
o f [tdeant amsecaineinncgel]e s s ...

- ... for understanding p
- ... for tracking patien
- ... for medical trials
- measurements

We need to understand
context of the patien

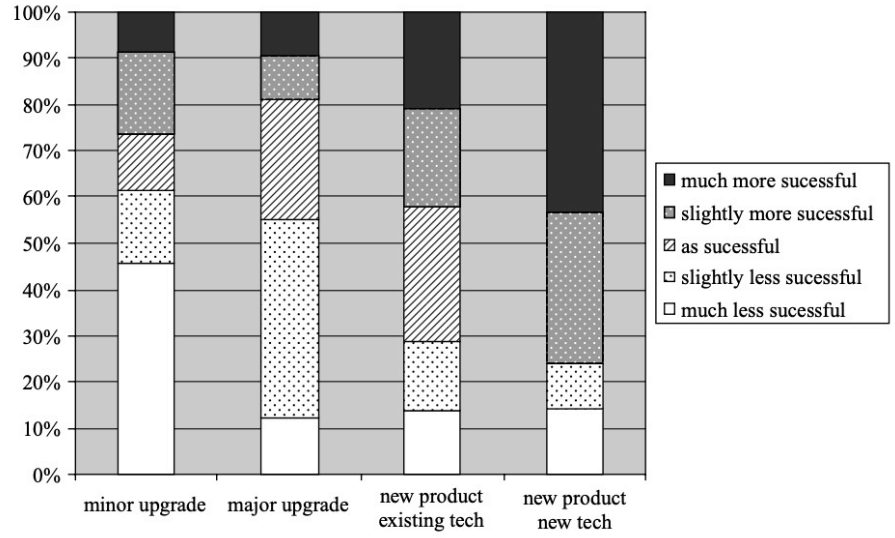
Surveys are an indis
healthcare

Surveys might be the s
we take in a study

NHS patient experience

"The importance of urb
and urban ecosystem se
COVID-19 pandemic"[1]

How successful do medi
companies rate their c



Source: Eatock, et al[2]

Surveys are an indis... healthcare

Surveys might be the study we take in a study

NHS patient experience surveys

"The importance of urban ... and urban ecosystem services during the COVID-19 pandemic"[1]

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

Surveys might instead ... context for other measures we take.

- Dietary studies might take samples but require ...
- Mental health studies ... physiological measures ... psychological surveys

Designing effective

We don't have enough time to go deep into everything in probably an entire undergraduate course in survey

> So why are we looking at surveys?

Designing effective

There are some tips in designing surveys.

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

However, these tips do not guarantee an effective survey.

In an effective survey:

- There is an overall g
- Each question is ask
  ist 'asking.
- Questions are unbiase
  leading.

The best way to test if your survey[3] is prete

Designing effective

Getting survey data in ... requires many different data wrangling tasks/skills.

The tidyverse ... developed by Hadley Wickham[4] in 2014.

We will use survey data as an introduction
wrangling:

- Pivoting data between wide and long forma
- Joining datasets [in the same way that SQL
- Wrangling survey questions with multiple c
- Wrangling survey questions that capture mu

Quantitative vs Qualitative measurements?

“Closed question” vs

Which is which?

Surveys can capture

Quantitative measurements collected with closed-ended questions.

Qualitative measurements collected with open-ended questions.

- Do you find it useful to wear a 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

If you were to recommend a tracking device to a friend what would you tell them?

Surveys can captu data

Ideally we would always use closed-form qu unfortunately this isn't always the case. T

Survey

Question "Where do you get most of your fi from?"

Question input

Example responses:

What do you think ab

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

- □ Yes
- □ No
- □ Don't know

This is a question sfr
bite-size gufidetivewr
questionnaires

Survey Mode

Methods of survey da

There are several methods for survey data collection:

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

The mode is considered a feature of the survey[7]:

Respondents answer the same question by mode[8]

Respondent demographics and survey topic

Inaccurate state-level 2016 US elections are have been strongly biased over-representation of graduates[9]

Survey Size

Survey Size

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

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

- These are reasonable
  studies" but for othe
  Serdar[10] et al
- For pre-test surveys
  to use sample at leas
  per Perneger et al[11]

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

t 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.

Sourced from[12]

LIKERT Scales (I)

In a Likert scale, responses are given a score.

response	response
Strongly	1
Disagree	2
Neutral	3
Agree	4
Strongly	5

question	response
Feel confident in	Strongly
Feel confident in	Neutral
Feel confident in	Disagree
Feel confident in	Strongly

It is absolutely meaningless to take the mean of Likert scale responses but people do it.

When you create your own survey, you're creating ordinal variables.

LIKERT Scales (II)

response	response
Strong (	1
Disagree	2
Neutral	3
Agree	4
Strong a	5

question	response
Feel confident in	Strong a
Feel confident in	Neutral
Feel confident in	Disagree
Feel confident in	Strong (

Our original question and ... variables an the intrinsic responses.

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

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

When we convert the q... LIKERT score we are the ... interval data

> But is this accurate...
>
> Is the distance between "Str... disagree" and "Disag... as that between "Neu... "Agree"?

LIKERT Scales (III)

response	response
Strong d	1
Disagree	2
Neutral	3
Agree	4
Strong a	5

question	response
Feel confident in	Strong a
Feel confident in	Neutral
Feel confident in	Disagree
Feel confident in	Strong d

If you want to compar
multiple Likert items then
meaning to the “media

You could perform fac
is well explained[14] by

When designing your s
also directly ask res
instead of trying to
+ “neutral” means.

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

Missing data in

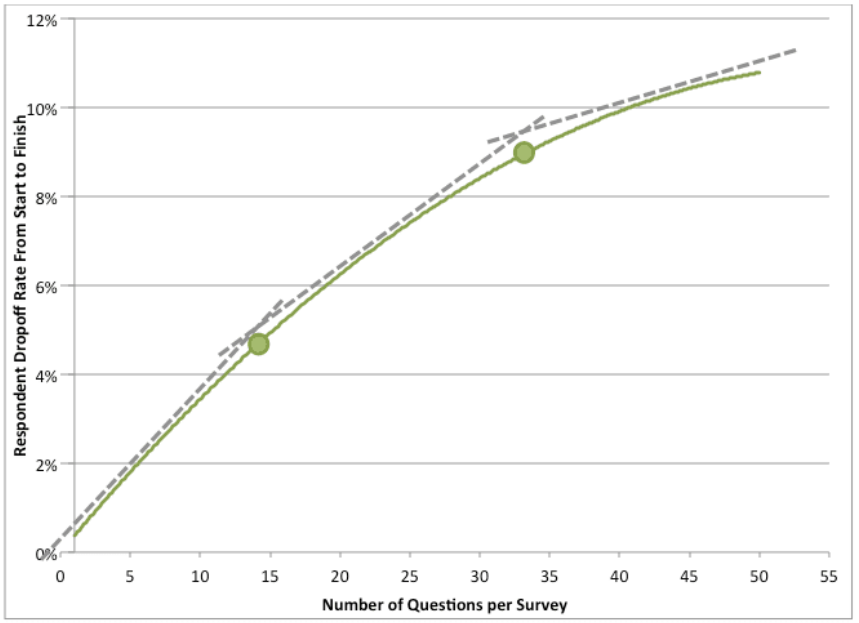
Missing data in surv

When designing a survey we ideally want res

However, research[15] has found that increa
rate (or survey abando
correlated with survey

Although, the effects may

- survey mode
- survey reward
- are questions skippab



Source: SurveyMonkey[15]

Missing data in surv

There are three kinds of missing data distr

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

Missingness is relate

- Missingess of respons related to question give up.
- Missingness might be respondent telling towar [eg extra-marital re subjects]

Data that is MNAR indicates study design.

However often it is difficult to det your data is indeed M

Missing data in surv

There are three kinds of missing data distr

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

Missingness is not ra
accounted for bye othe
could call this “cond

- Survey abandonment c
by including informa
order.
- Missingness might be
differences in demogra
males are less likely
depression surveys.

Data sthMAR’ provides us
methodology for imput

Missing data in surv

There are three kinds of missing data distr

- Missing Not at
- Missing at Ran
- Missing Completely at Random (MCAR)

Missingness is truly
in the dataset. There
variables.

In practice it is har
MNAR without actually seeing the
study for randomness.

- Randomly sampling a s
  questions for each pa

Planned missingness depends
on fairly advanced st
methodologies.

Survey tools

Survey tools

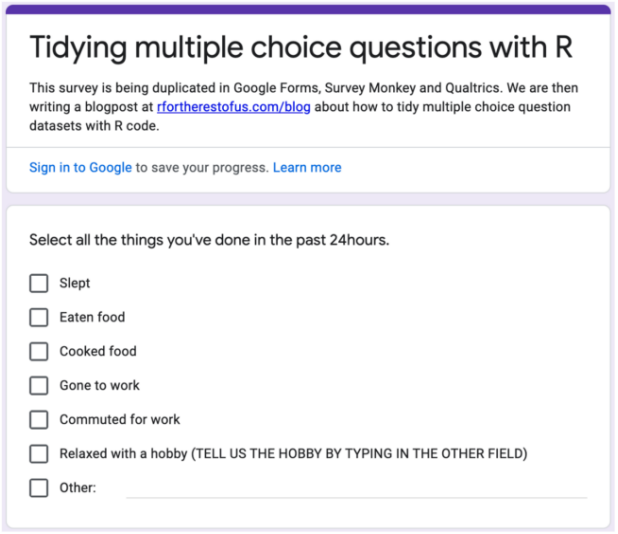
There's a plethora of survey tools available.

Many of these tools provide free tiers but [and for GDPR compliance].

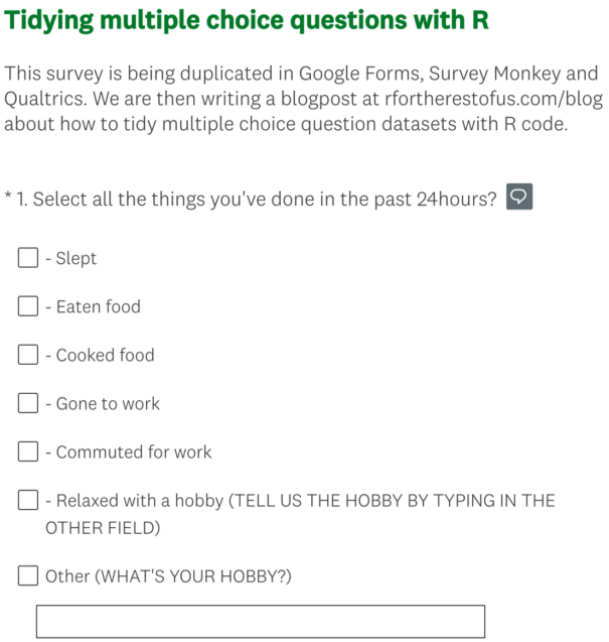
One thing that is useful in these tools is their unique format - interactive of

> Let's look first at this timeline format about

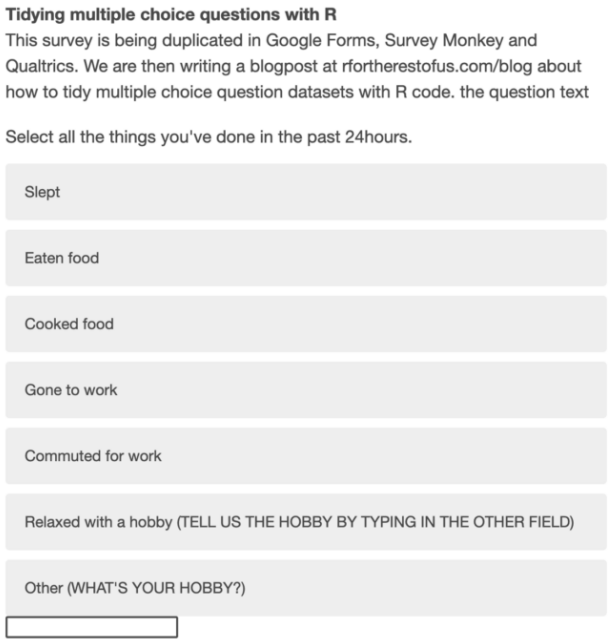
Google Forms



Survey Monkey



Qualtrics



Google Forms

Timestampyou've done in the past 24hours.
2022-04-21 Slept, Eaten food, Cooked food, Gone to work
2022-04-21 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), Played mobile games
2022-04-21 Slept, Eaten food, Cooked food
2022-04-21 Slept, Eaten food, Cooked food, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Pokemon
2022-04-21 Slept, Eaten food, Cooked food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Video Games
2022-04-21 Slept, Eaten food, Cooked food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), lifted weights
2022-04-21 Slept, Eaten food, Cooked food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Watched football
2022-04-21 Slept, Eaten food, Gone to work
2022-04-21 Slept, Eaten food, Gone to work, Commuted for work
2022-04-21 Slept, Eaten food, Gone to work, Commuted for work
2022-04-21 Slept, Eaten food, Gone to work
2022-04-21 Slept, Eaten food, Cooked food, Gone to work
2022-04-21 Slept, Eaten food, Cooked food, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), video games
2022-04-21 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?
2022-04-21 Slept, Eaten food, Cooked food, Gone to work
2022-04-21 Slept, Commuted for work, Other
2022-04-21 Slept, Eaten food, Cooked food, Gone to work
2022-04-21 Slept, Eaten food, Cooked food, Gone to work, Commuted for work
2022-04-21 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
2022-04-21 Slept, Eaten food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Diabetes
2022-04-21 Slept, Eaten food
2022-04-21 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
2022-04-21 Slept, Eaten food, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Stroll outside
2022-04-21 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)
2022-04-21 Slept, Eaten food, Cooked food
2022-04-21 Slept, Eaten food
2022-04-21 Slept, Eaten food, Gone to work, Commuted for work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), tennis
2022-04-21 Slept, Eaten food, Cooked food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Furniture restoration
2022-04-21 Slept, Eaten food, Cooked food
2022-04-21 Slept, Eaten food, Gone to work, Commuted for work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), TV
2022-04-21 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
2022-04-21 Slept, Eaten food, Gone to work
2022-04-21 Slept, Eaten food, Cooked food, Basketball
2022-04-21 Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), working
2022-04-21 Slept, Eaten food, Cooked food
2022-04-21 Slept, Eaten food, Cooked food, Gone to work
2022-04-21 Slept, Eaten food, Cooked food, Gone to work, Commuted for work, Nature walks
2022-04-21 Slept, Eaten food, Cooked food, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Warched TV
2022-04-21 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
2022-04-21 Slept, Eaten food, Cooked food, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Writing R code;-)
2022-04-21 Slept, Eaten food, Cooked food
2022-04-21 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
2022-04-21 Slept, Eaten food, Gone to work, Commuted for work
2022-04-21 Slept, Eaten food, Gone to work
2022-04-21 Slept, Eaten food, Gone to work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD)
2022-04-21 Slept, Eaten food, Gone to work, Commuted for work, Relaxed with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD), Gardening
2022-04-21 Slept, Eaten food, Cooked food, Writing, and visiting family
2022-04-21 Eaten food, Cooked food, Commuted for work, yoga
2022-04-21 Slept, Eaten food, Cooked food, Gone to work, Commuted for work
2022-04-21 Slept, Eaten food, Gone to work, Commuted for work
2022-04-21 Eaten food
2022-04-21 Slept, Eaten food
2022-04-21 Slept, Eaten food, Gone to work, Commuted for work
2022-04-21 Slept, Eaten food, Gone to work, Commuted for work

SurveyMonkey

Start Date	Select all the things you've done in the past 24hours?				
	- Slept	- Eaten foo	- Cooked f	- Gone to	- Commut	- Relaxed	Other (WHAT'S YOUR HOBBY?)
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to work			
2022-04-21	- Slept	- Eaten food		- Gone to w	- Commute	- Relaxed w	Played mobile games
2022-04-21	- Slept	- Eaten foo	- Cooked food				
2022-04-21	- Slept	- Eaten foo	- Cooked food			- Relaxed w	Pokemon
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to work		- Relaxed w	Video Games
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to work		- Relaxed w	lifted weights
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to work		- Relaxed w	Watched football
2022-04-21	- Slept	- Eaten food		- Gone to work			
2022-04-21	- Slept	- Eaten food		- Gone to w	- Commuted for work		
2022-04-21	- Slept	- Eaten food		- Gone to w	- Commuted for work		
2022-04-21	- Slept	- Eaten food		- Gone to work			
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to work			
2022-04-21	- Slept	- Eaten foo	- Cooked food			- Relaxed w	video games
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to w	- Commute	- Relaxed w	Coding/programming - but is it really relaxing?
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to work			
2022-04-21	- Slept				- Commuted for work		Other
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to work			
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to w	- Commuted for work		
2022-04-21	- Slept	- Eaten food		- Gone to w	- Commute	- Relaxed w	Mountain biking
2022-04-21	- Slept	- Eaten food		- Gone to work		- Relaxed w	Diabetes
2022-04-21	- Slept	- Eaten food					
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to w	- Commute	- Relaxed w	Reading
2022-04-21	- Slept	- Eaten food				- Relaxed w	Stroll outside
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to work		- Relaxed w	Playing the guitar and embrodering (not at the same time lol)
2022-04-21	- Slept	- Eaten foo	- Cooked food				
2022-04-21	- Slept	- Eaten food					
2022-04-21	- Slept	- Eaten food		- Gone to w	- Commute	- Relaxed w	tennis
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to work		- Relaxed w	Furniture restoration
2022-04-21	- Slept	- Eaten foo	- Cooked food				
2022-04-21	- Slept	- Eaten food		- Gone to w	- Commute	- Relaxed w	TV
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to w	- Commute	- Relaxed w	Amateur R
2022-04-21	- Slept	- Eaten food		- Gone to work			
2022-04-21	- Slept	- Eaten foo	- Cooked food				Basketball
2022-04-28	11:16:48					- Relaxed w	working
2022-04-21	- Slept	- Eaten foo	- Cooked food				
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to work			
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to w	- Commuted for work		Nature walks
2022-04-21	- Slept	- Eaten foo	- Cooked food			- Relaxed w	Warched TV
2022-04-21	- Slept	- Eaten food		- Gone to w	- Commute	- Relaxed w	Watching Series
2022-04-21	- Slept	- Eaten foo	- Cooked food			- Relaxed w	Writing R code;-)
2022-04-21	- Slept	- Eaten foo	- Cooked food				
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to work		- Relaxed w	Reading a book
2022-04-21	- Slept	- Eaten food		- Gone to w	- Commuted for work		
2022-04-21	- Slept	- Eaten food		- Gone to work			
2022-04-21	- Slept	- Eaten food		- Gone to work		- Relaxed	with a hobby (TELL US THE HOBBY BY TYPING IN THE OTHER FIELD)
2022-04-21	- Slept	- Eaten food		- Gone to w	- Commute	- Relaxed w	Gardening
2022-04-21	- Slept	- Eaten foo	- Cooked food				Writing, and visiting family
2022-04-28	13:07:53	- Eaten foo	- Cooked food		- Commuted for work		yoga
2022-04-21	- Slept	- Eaten foo	- Cooked fo	- Gone to w	- Commuted for work		
2022-04-21	- Slept	- Eaten food		- Gone to w	- Commuted for work		
2022-04-28	13:08:01	- Eaten food					
2022-04-21	- 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	ours. - 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/softwar

Because sthuecrhe’a huge variety in dastaimexpotrat
train your general purpose data wrangling s

In the lecture we’ll look at several real-w
wrangling.

In the workshop you’ll look at simpler data
learn. now

Survey file formats

- Excel files: Almost all tools read excel files. However, there are a wide range of options in how the data is arranged - inc
  - We use the {readxl} package for reading
  - If data is encoded with cell colour you'
    package.

- .csv files: Most tools provide a .csv export file, "o
  text file". They're likely to be well formatted
  - Files like .csv are read into R with the

Survey datasets f

Survey datasets (I)

We'll be looking at datasets during this week

- Emerging Adulthood Measured[17] at Multiple In
  - This is the 2nd instance of a large scal
  - Learn about the 1st instance (and how[18]. th
  - The actual survey questionnaire is available
  - The survey dataset comes as a coded CSV[19] fo
  - The actual survey data is available for
    downloaded via a URL.

osf.io/3zq5e

osf.io/download/c3 from

Survey datasets (II)

We'll be looking at data from these surveys during this week

- Emerging Adulthood Measured [17]

- Public Attitudes to Commercial Access to H[20]
  - The Wellcome Trust commissioned Ipsos MORI to s access to health data
  - The study can be read here [21]
  - The survey questions can be found at the[21].
  - The dataset is openly available but requires a UK account.[20]

Survey datasets (II)

We'll be looking at data from these surveys during this week

- Emerging Adulthood Measured [17]
- Public Attitudes to Commercia [20]

- British Election Study 2019
  - Since 1964 a post-election survey has be motivations and the impact of political
  - Data for all surveys is available for joined cross-sectional-data/
  - The 2019 questionnaire and information o britishelection.ac.uk/data-publications/2019-british-election-study/ [22]
  - The 2019 survey data is available from the UK Data Service [5] but you must first add to your UK Data Service account.

Survey datasets (IV)

We'll be looking at data from the datasets during this week

- Emerging Adulthood Measured at Multiple In
- Public Attitudes to Commercial Access to H
- British Election Study 2019

Task: Setup our pro

SLIDE 1 OF 3

1. Create a new empirical-weekly-surveys project.
2. Add a subfolder called data to store all of the datasets.
3. Create a separate .Rmd document for each of the studi

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

📝 Task: Obtain Emergi data

SLIDE 2 OF 3

1. Open the emerging-adulthood.Rmd file

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

- The codebook
- The dataset

2. Add a code chunk to load the {tidyverse} and {readxl} packages.
3. Add this code chunk to your .Rmd file to down

Task: Read in Emerging Adulthood data

SLIDE 3 OF 3

When we read in datasets we should always store the 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 (

Most fields you'll work with will have messy c
with:

```
1 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\":\"...
```

```
1 adulthood_raw_codebook %>%
2   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 clean names is with the clean_names() function in the janitor package.

Messy column names (

But still need to understand how these two datasets rela

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

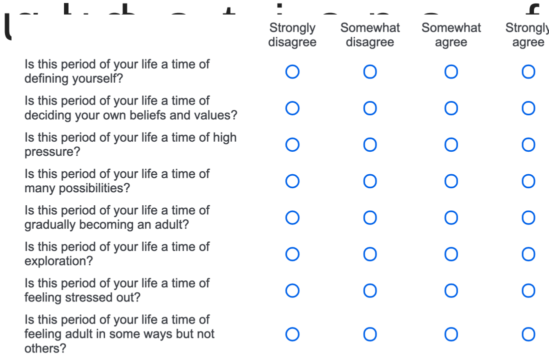
If we clean up the column names of adulthood_raw_codebook, then
this instantly will clean the codebook column nam

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

IDEA Questions (I)

Lest 't ake a look at the these … from

you suggest … da … questions … bo




```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_39pIdNoon8CePfP      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 dataset?

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

ResponseId	IDEA	IDEA	IDEA	IDEA	IDEA
R_BJN3bQqi1	3	4	4	3	4
R_2TGbiBXmA	4	4	4	4	3
R_12G7blqN2	4	4	4	4	4
R_39pldNoon	4	4	3	3	4
R_1QiKb2Lfd	4	4	3	4	3
R_pmwDTZyCy	3	4	3	3	4

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

IDEA Questions (III)

- Using `across()` to target multiple columns at onc

```
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 t

```
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()` hears the fitd doefd poreenpearing the d
> `{ggplot2}`

Wide vs long data (l

In wide datasets each variable is stored in a unique column.

Pers	Ag	Weig	Hei
Bob	32	16	18
Alic	24	15	17
Stev	64	14	16

In long datasets each row is an observation.

Pers	Vari	Val
Bob	Age	32
Bob	Weig	16
Bob	Heig	18
Alic	Age	24

However, datasets might be messy.
For instance, year is spread across multiple columns.

coun	variabl	200	200	200
UK	Superma	20	20	23
UK	Shoppin	40	42	46
US	Superma	30	36	38
US	Shoppin	80	90	98

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

coun	variabl	yea	val
UK	Superma	200	20
UK	Superma	200	20
UK	Superma	200	23
UK	Shoppin	200	40
UK	Shoppin	200	42
coun	variabl	yea	val

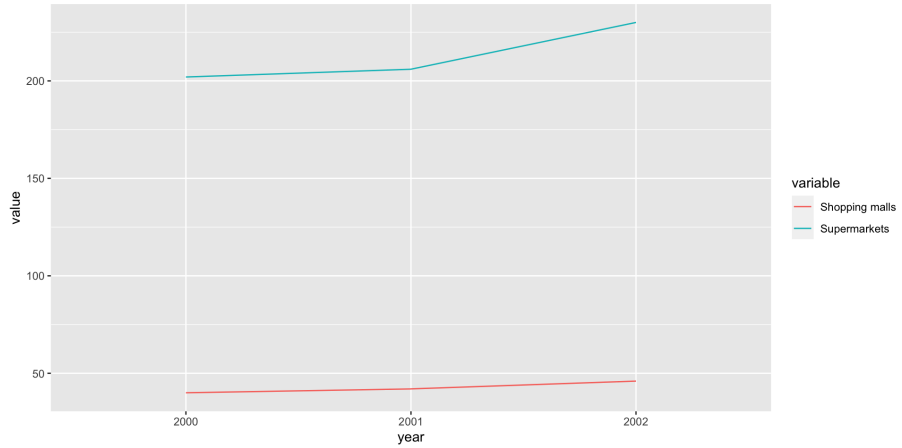
UK	Shoppin	200	46
US	Superma	200	30
US	Superma	200	36
US	Superma	200	38
US	Shoppin	200	80
US	Shoppin	200	90
US	Shoppin	200	98

Wide vs long data (l

The {ggplot2} package requi
data

coun	variabl	yea	val
UK	Superma	200	20
UK	Superma	200	20
UK	Superma	200	23
UK	Shoppin	200	4(
UK	Shoppin	200	42
UK	Shoppin	200	46
US	Superma	200	30
US	Superma	200	36
US	Superma	200	38
US	Shoppin	200	8(
US	Shoppin	200	9(
US	Shoppin	200	98

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



pivot_wider() and pi

The pivot_wider() and pivot_longer() functions are for tra
format and wide format, respectively

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

Pers	name	val
Bob	Age	3
Bob	Weig	16
Bob	Heig	18
Alic	Age	2
Alic	Weig	15
Alic	Heig	17
Ste	Age	6
Ste	Weig	14
Ste	Heig	16

Note that we only use tidyselect functions col

> Pre 2020 spread() and gather(). These {tidyr} func
> but are considered superseded functions.

IDEA Questions (IV)

We can now transform our actual dataset int

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

The remaining step is to use the case_when

```
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_BJN3bQqi	IDEA	Agree
R_BJN3bQqi	IDEA	Strong
R_BJN3bQqi	IDEA	Strong
R_BJN3bQqi	IDEA	Agree
R_BJN3bQqi	IDEA	Strong
R_BJN3bQqi	IDEA	Strong

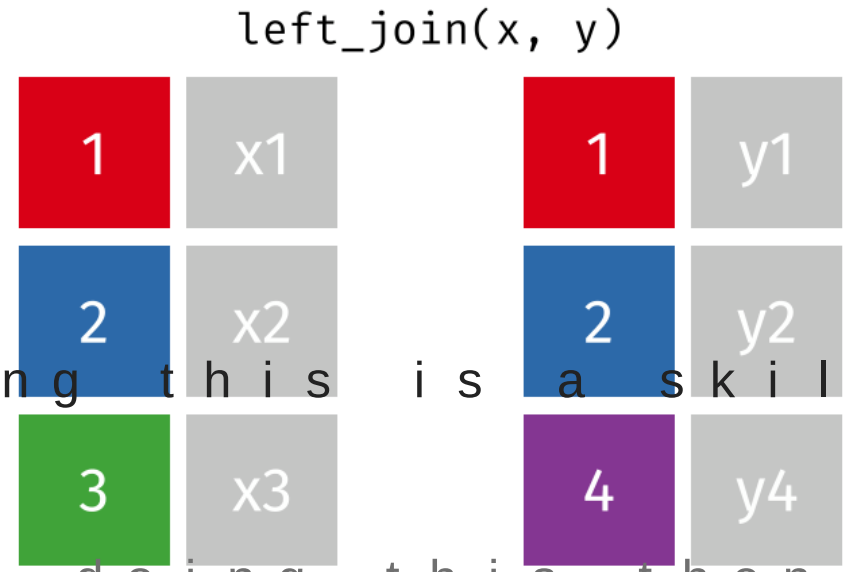
IDEA Questions (V)

We need to match up th[illegible]
with the actual questi[illegible]

To achieve this we're mutating join.

Is it 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.wickadenbuie[illegible]

IDEA Questions (VI)

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

ResponseId	name	value
R_BJN3bQqi	IDEA_Agree	
R_BJN3bQqi	IDEA_Strong	
R_BJN3bQqi	IDEA_Strong	
R_BJN3bQqi	IDEA_Agree	
R_BJN3bQqi	IDEA_Strong	
R_BJN3bQqi	IDEA_Strong	

Lest 'extract the variable labels from the codeb

Note that the column datased is for agreement

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

idea_question_labels %>%
  gt()
```

variabl	question_text
IDEA_1	Is this period of your possibilities?
IDEA_2	Is this period of your
IDEA_3	Is this period of your out?
IDEA_4	Is this period of your
IDEA_5	Is this period of finiymogu ry o
IDEA_6	Is this period of your own beliefs and values?
IDEA_7	Is this period of your some ways but not other
IDEA_8	Is this period of your becoming an adult?

IDEA Questions (VII)

Because the columfmenr ematmewse anreee dd itfo give left

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

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

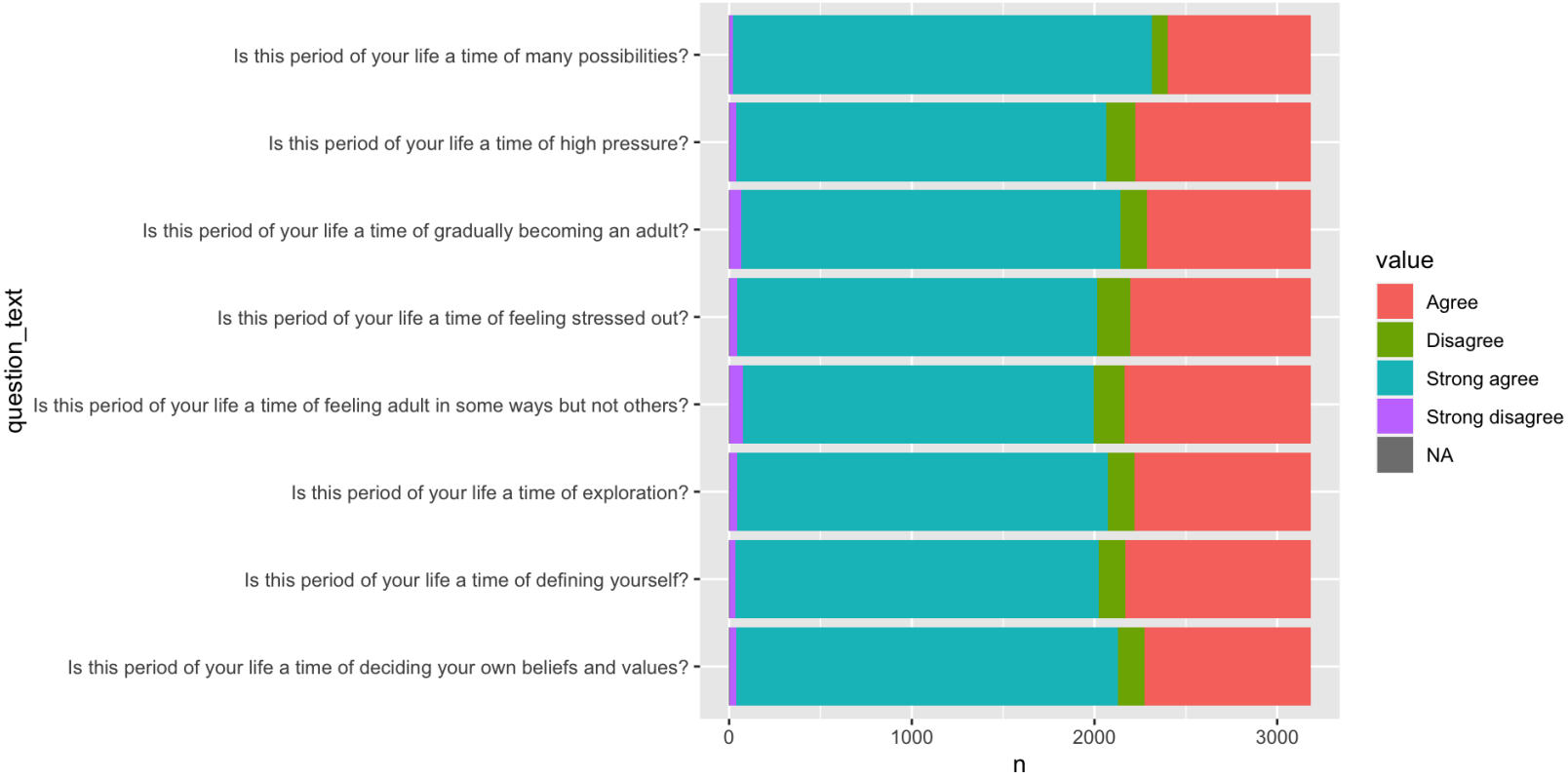
ResponseId	name	value	question_text
R_BJN3bQqi	IDEA	Agree	Is this period of your life a ti
R_BJN3bQqi	IDEA	Strong	Is this period of your life a ti
R_BJN3bQqi	IDEA	Strong	Is this period of your life a ti
R_BJN3bQqi	IDEA	Agree	Is this period of your life a ti
R_BJN3bQqi	IDEA	Strong	Is this period ofin iynogu ry oluirfsee laf ?t i
R_BJN3bQqi	IDEA	Strong	Is this period of your life a ti

> … so why did we cant how ask? count() to ta question_text and visualise the results.

IDEA Questions (VIII

Finally we can visualise the responses... r

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

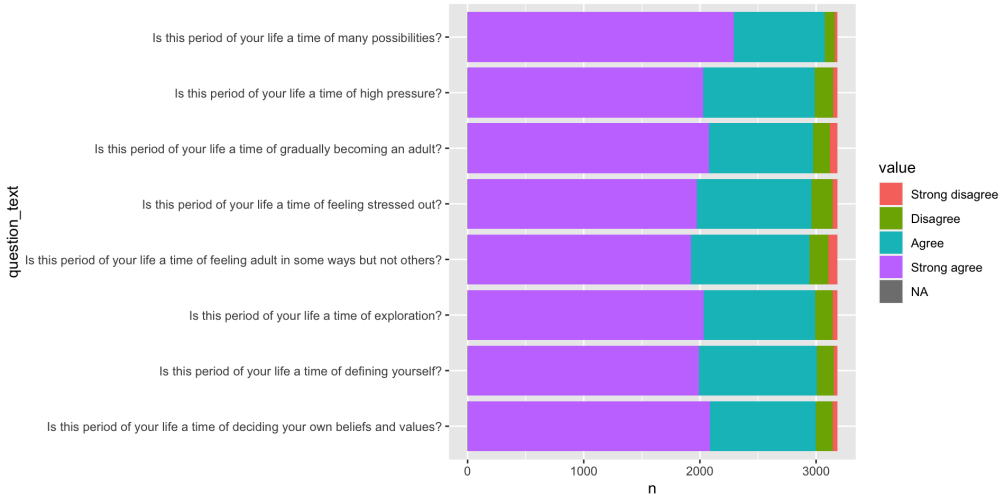


IDEA Questions (IX)

We need to use fct_relevel() to set the car

- Which part of the chart do we need to fill to color
- Which part of the chart do we need to targ

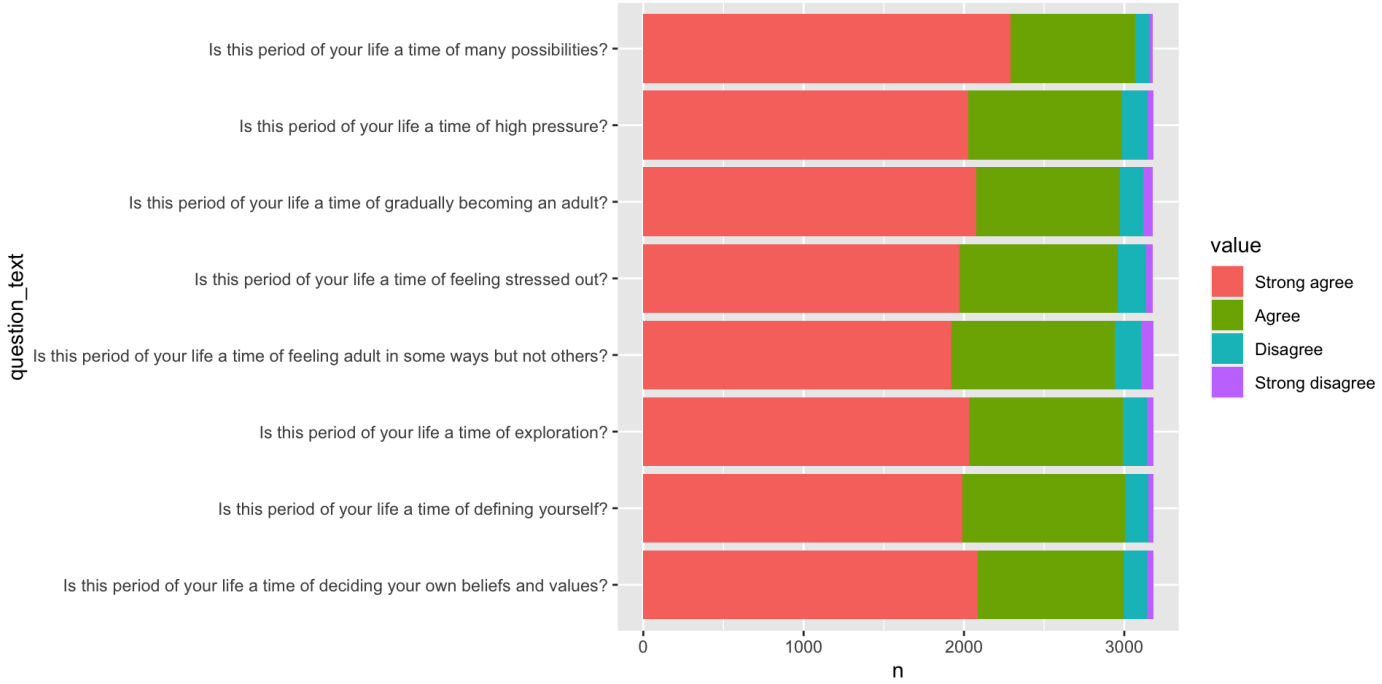
```
1  order_agree_responses <- c("Strong disagree", "Disagree", "Agree", "Strong agree")
2
3  idea_responses_clean %>%
4    count(question_text, value) %>%
5    mutate(value = fct_relevel(value, order_agree_responses)) %>%
6    ggplot(aes(x = n,
7               y = question_text,
8               fill = value)) +
9    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 m

Follow this same process to visualise the same s.urvey

> Lest’take 20mins for this

Please note that I’ve not created this cha during tshimsa tweereika’l.

Task: Obtain Commer Health Data

SLIDE 1 OF 2

1. Register for a FREE UK Data Service accoun
   beta.ukdataservice.ac.uk/myaccount/login
2. Navigate to the access data page for the c
   beta.ukdataservice.ac.uk/datacatalogue/stu
3. Download the SPSS dataset

> Ist’very useful to learn how to deal with S

1. Unzip the dataset and add the folder to th

📝 Task: Obtain Commercial Health Data

SLIDE 2 OF 2

1. Open up the `commercial-access-to-health.Rmd` RMarkdown document
2. Load the `haven` and `tidyverse` packages

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

```
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 augmente
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+… <dbl+… <dbl+… <dbl+… <dbl+… <dbl+… <dbl+> <dbl+… <dbl+… <dbl+… <dbl+… <dbl+… <dbl+… <dbl+… <dbl+… <dbl+…
 1 4 [Hea… 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 g… 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 [Jus… 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 g… 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 [Jus… 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 [Jus… 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 [Jus… 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 [Jus… 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 [Jus… 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 [Jus… 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 (I

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

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

The haitse si sa sslpiegchitally conf

- singular “label” mean (or a shortened vers
- plural “labels” means 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 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 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 quest {haven} output (I)

We have two different ways to extract labels from

- 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

I'd recommend using the {sjlabelled} package exclusively for labels from SPSS. It does provide more and also can list sc

Extracting the quest {haven} output (II)

The `sjlabelled::get_label()` function generates a named vector 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 column?

Extracting the quest

{haven} output (II)

This is one of many ways to tidy up this da

```
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

To convert all labelled columns to factors

```
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

Add respondent ID

The `row_number()` function gives us a neat way to

However, it's not necessarily that clever a solut

```
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 Da

I'd like you to extract the columns from th

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 Da

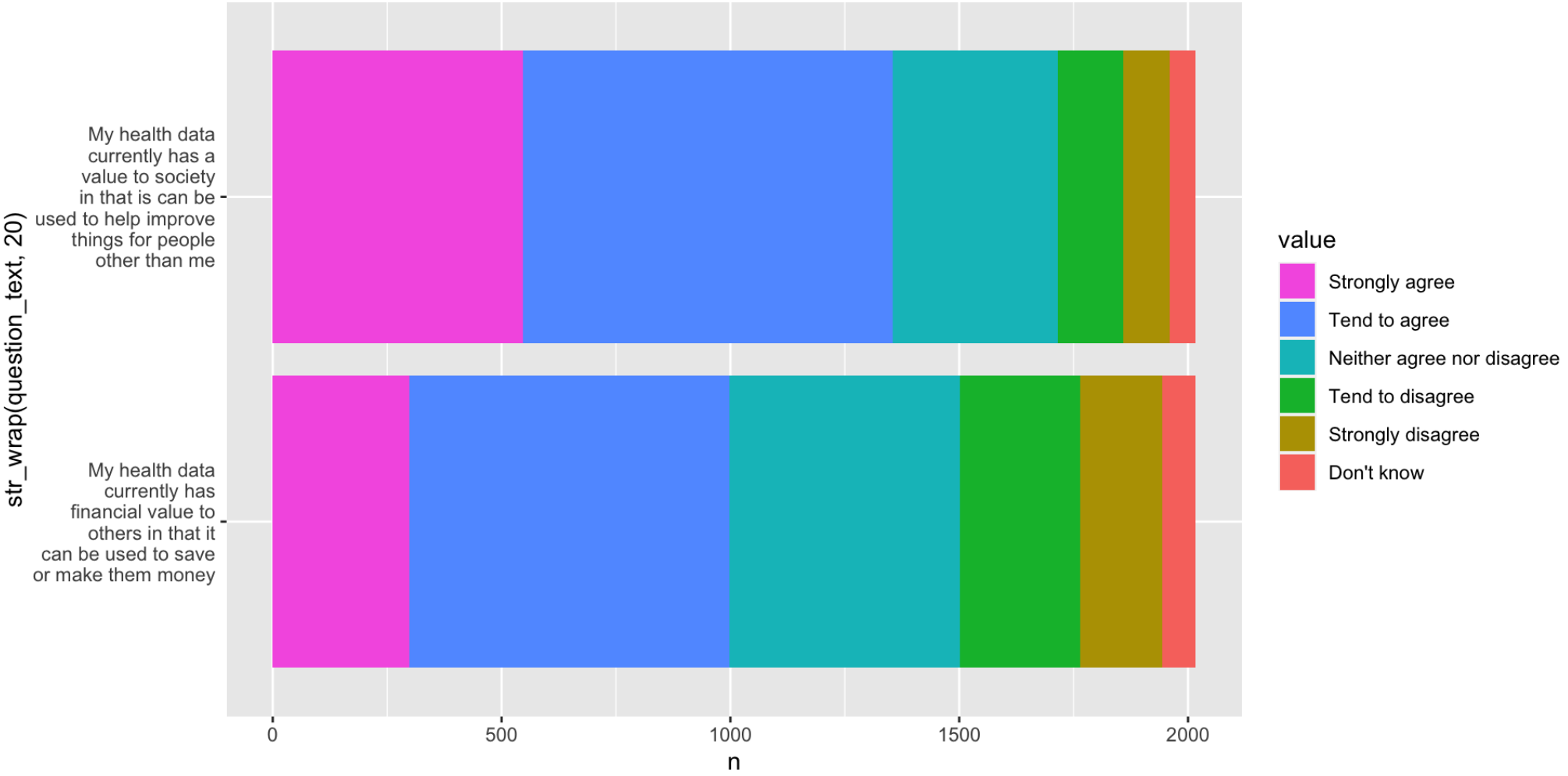
What do we need to do to this data so that {ggplot2}?

```
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 Da

Lest 'make this chart:



How wide is too

```
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 vox ep rank?!
- In principle we could combine:
  - votes
  - views
  - vox ep rank

```
1 buffy_raw <- read_csv(here::here("static", "datasets", "buffy", "buffy_data.csv"))
2 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

Tidy data is a useful concept for wrangling

But sometimes it's not something to conform to religiously

You might want to keep some width to your d

Wide data might also be more appropriate i

Other forms of un

Multiple pieces of d

Sometimes a single col
multiple variables.

You might also ask res
all that apply"

This is often the case
"where do you live?" d

```
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
```

```
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 Data

SLIDE 1 OF 2

1. Register for a FREE British Election Study user account: https://www.britishelectionstudy.com/wp-login.php?action=register
2. Navigate to the access data page: https://www.britishelectionstudy.com/data-object/2019-british-election-study-post-e
3. Download the SPSS dataset
4. Unzip the dataset and add the folder to th

Task: Obtain British Election Study Data

SLIDE 2 OF 2

1. Setup the british_election_study_data folder in your working directory
2. Read in the file SPSS

```
# 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 from?

Can you extract the co[illegible] dataset corresponding

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

K04: 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 from?

This is an open-ended question and is going to be really messy.

To properly analyse this data we would use the {tidytext} package and work with a tidyverse approach.

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

- `separate()`?
- `separate_rows()`?

```
1 british_election_data_raw %>%
2   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 well-being. (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. Tell, 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.