ETC5510: Introduction to Data Analysis

Week of Tidy Data

Stuart Lee and Nick Tierney

16th Mar 2020

About your instructors

Stuart

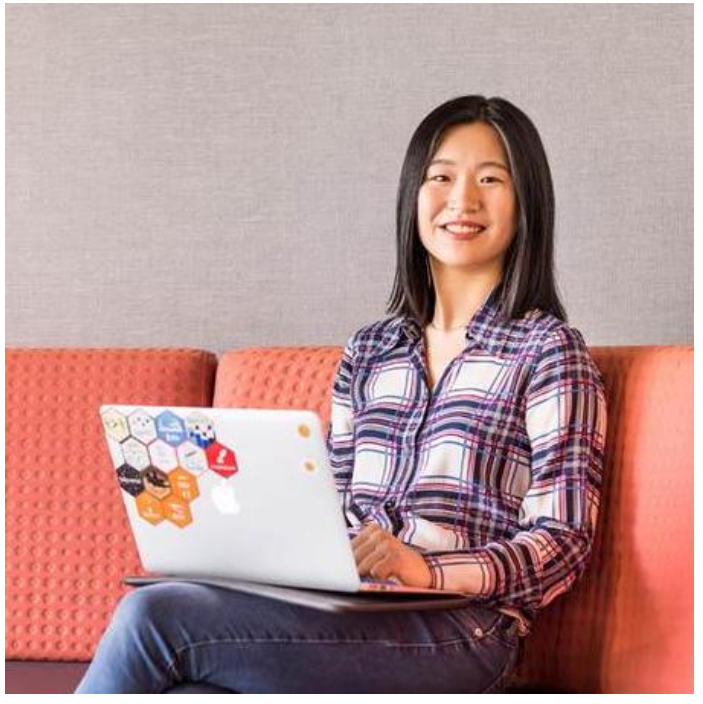
- 🎓 Bachelor of Mathematical Sciences at University of Adelaide
- 🎓 PhD Candidate in Statistics at Monash EBS.
- Research: genomics, data visualisation, statistical computing
- ❤️: board games, cooking, music, reading and video games

Steph

- 🎓 Bachelor of Economics and Bachelor of Commerce from Monash
- Studying a Masters of Statistics at QUT, based at Monash.
- Loves to read 📖, any and all recommendations are welcome.
- Has an R package called taipan, and another called sugarbag.

Sherry

- 🎓 Bachelor of Commerce 2018
- Honours in Econometrics 2019 with Di Cook
- Commenced PhD programme 2020
- Created her first ever R package, `quickdraw`
- Loves puzzles games like jigsaws 🧩.



Nick

- 🎓 Bachelor of Psychological Sciences UQ
- 🎓 PhD in Statistics at QUT.
- Research: missing data, data visualisation, statistical computing
- R 📦: `naniar`, `visdat`,
- `#rstats` 🎤: Credibly Curious w Saskia Freytag
- ❤️ outdoors, especially: 🥾, 🏃, and 🧗.

Di

- Professor at Monash University in Melbourne Australia, doing research in statistics, data science, visualisation, and statistical computing.
- Created the current version of the course
- Likes to play all sorts of sports, tennis, soccer, hockey, cricket, and go boogie boarding.

Your Turn: Making the groups

We are going to set up the groups for doing assignment work.

1. Find your name from the list at this link
2. Find the other people in the class with the same quote as you (feel free to wander around the class!)
3. Grab your gear and claim a table to work together at.

Your Turn: Ask your team mates these questions:

1. What is one food you'd never want to taste again?
2. If you were a comic strip character, who would you be and why?

LASTLY, come up with a name for your team (we have provided a suggested name, but you are free to change it!) and tell this to a tutor, along with the names of members of the team.

05:00

Traffic Light System

Traffic Light System

Red Post-it

- I need a hand
- Slow down

Green Post-it

- I am up to speed
- I have completed the thing

Recap

- packages are installed with _ ?
- packages are loaded with _ ?
- Why do we care about Reproducibility?
- Output + input of rmarkdown
- I have an assignment group
- If I have an assignment group, have recorded my assignment group in the ED survey

Today: Outline

- An aside on learning
- Tidy Data
- Terminology of data
- Different examples of data
- Steps in making data tidy
- Lots of examples

A note on difficulty

- This is not a programming course - it is a course about **data, modelling, and computing**.
- At the moment, you might be sitting there, feeling a bit confused about where we are, what are are doing, what R is, and how it even works.
- That is OK!
- The theory of this class will only get you so far
- The real learning happens from doing the data analysis - the **pressure of a deadline can also help.**

Tidy Data

You're ready to sit down with a newly-obtained dataset, excited about how it will open a world of insight and understanding, and then find you can't use it. You'll first have to spend a significant amount of time to restructure the data to even begin to produce a set of basic descriptive statistics or link it to other data you've been using.

--John Spencer (Measure Evaluation)

Tidy Data

"Tidy data" is a term meant to provide a framework for producing data that conform to standards that make data easier to use. Tidy data may still require some cleaning for analysis, but the job will be much easier.

--John Spencer (Measure Evaluation)

Example: US graduate programs

- Data from a study on US grad programs.
- Originally came in an excel file containing rankings of many different programs.
- Contains information on four programs:
  1. Astronomy
  2. Economics
  3. Entomology, and
  4. Psychology

Example: US graduate programs

```
library(tidyverse)
grad <- read_csv(here::here("slides/data/graduate
grad
## # A tibble: 412 x 16
##    subject Inst  AvNumPubs AvNumCits PctFacGran
##    <chr>   <chr>     <dbl>     <dbl>       <db
##  1 econom… ARIZ…       0.9       1.57         31
##  2 econom… AUBU…       0.79      0.64         77
##  3 econom… BOST…       0.51      1.03         43
##  4 econom… BOST…       0.49      2.66         36
##  5 econom… BRAN…       0.3       3.03         36
##  6 econom… BROW…       0.84      2.31         27
```

Example: US graduate programs

Good things about the format:

```
## # A tibble: 6 x 16
##   subject Inst  AvNu
##   <chr>   <chr>
## 1 econom… ARIZ…
## 2 econom… AUBU…
## 3 econom… BOST…
## 4 econom… BOST…
## 5 econom… BRAN…
## 6 econom… BROW…
## # … with 9 more vari
## #   PctFemaleStud <d
```

Rows contain information about the institution

Columns contain types of information, like average number of publications, average number of citations, % completion,

Example: US graduate programs

Easy to make summaries:

```
grad %>% count(subject)
## # A tibble: 4 x 2
##   subject          n
##   <chr>        <int>
## 1 astronomy       32
## 2 economics      117
## 3 entomology      27
## 4 psychology     236
```

Example: US graduate programs

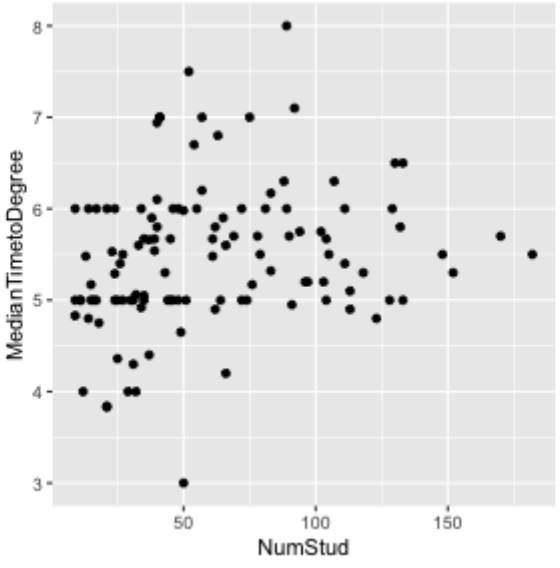
Easy to make summaries:

```
grad %>%
  filter(subject == "economics") %>%
  summarise(
    mean = mean(NumStud),
    s = sd(NumStud)
  )
## # A tibble: 1 x 2
##    mean     s
##   <dbl> <dbl>
## 1  60.7  39.4
```

Example: US graduate programs

Easy to make a plot

```
grad %>%
  filter(subject == "ec
  ggplot(aes(x = NumStu
  geom_point() +
  theme(aspect.ratio =
```



Your Turn: download exercises for today's lecture!

- Notice the `data/` directory with many datasets!
- Open `graduate-programs.Rmd`
- Answer these questions:
  - "What is the average number of graduate students per economics program?"
  - "What is the best description of the relationship between number of students and median time to degree?"
- Use the traffic light system if you need a hand.

03:00

What could this image say about R?

03:00

Terminology of data: Variable

- A quantity, quality, or property that you can measure.
- For the grad programs, these would be all the column headers.

```
## # A tibble: 412 x 16
##    subject Inst  AvNumPubs AvNumCits PctFacGran
##    <chr>   <chr>     <dbl>     <dbl>        <db
##  1 econom… ARIZ…      0.9       1.57         31
##  2 econom… AUBU…      0.79      0.64         77
##  3 econom… BOST…      0.51      1.03         43
##  4 econom… BOST…      0.49      2.66         36
##  5 econom… BRAN…      0.3       3.03         36
##  6 econom… BROW…      0.84      2.31         27
##  7 econom… CALI…      0.99      2.31         56
```

Terminology of data: Observation

- A set of measurements made under similar conditions
- Contains several values, each associated with a different variable.
- For the grad programs, this is institution, and program, uniquley define the observation.

```
## # A tibble: 412 x 16
##    subject Inst  AvNumPubs AvNumCits PctFacGran
##    <chr>   <chr>     <dbl>     <dbl>        <db
##  1 econom… ARIZ…       0.9      1.57          31
##  2 econom… AUBU…       0.79     0.64          77
##  3 econom… BOST…       0.51     1.03          43
##  4 econom… BOST…       0.49     2.66          36
```

Terminology of data: Value

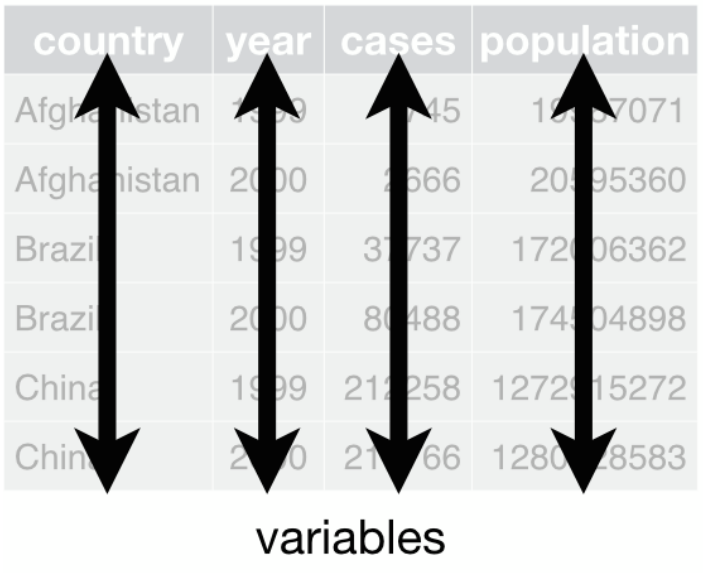
- Is the state of a variable when you measure it.
- The value of a variable typically changes from observation to observation.
- For the grad programs, this is the value in each cell

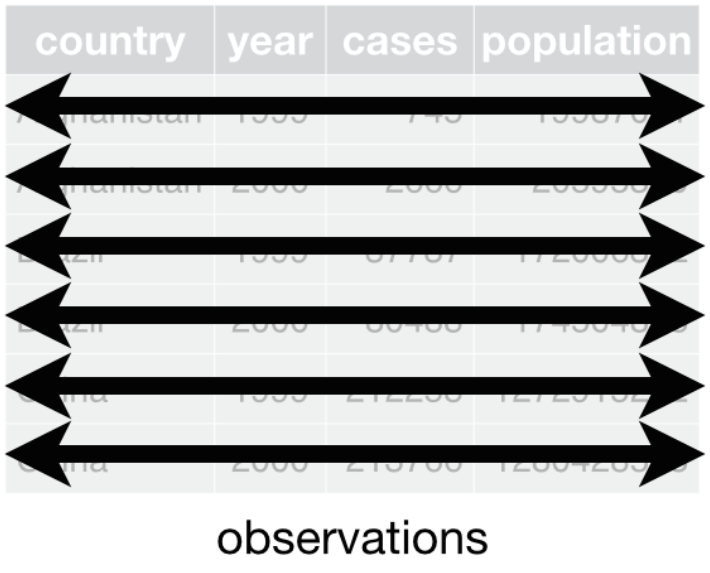
```
## # A tibble: 412 x 16
##    subject Inst  AvNumPubs AvNumCits PctFacGran
##    <chr>   <chr>     <dbl>     <dbl>        <db
##  1 econom… ARIZ…      0.9      1.57          31
##  2 econom… AUBU…      0.79     0.64          77
##  3 econom… BOST…      0.51     1.03          43
##  4 econom… BOST…      0.49     2.66          36
##  5 econom… BRAN…      0.3      3.03          36
```

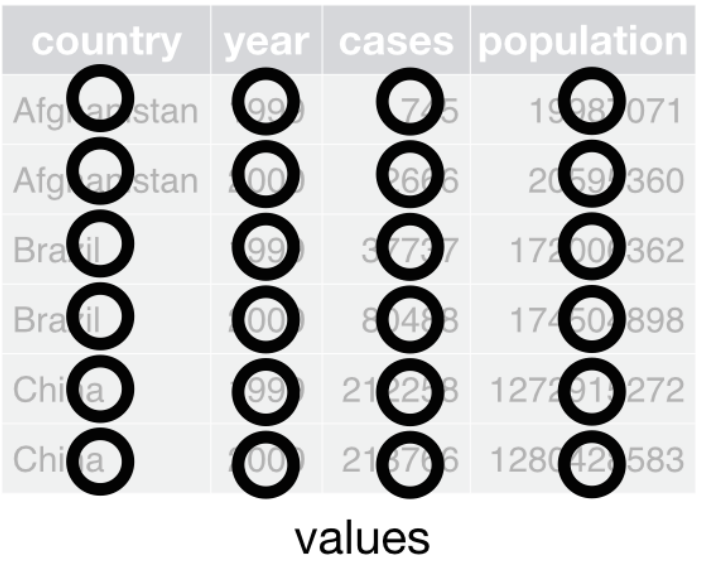
Tidy tabular form

Tabular data is a set of values, each associated with a variable and an observation. Tabular data is **tidy** iff (if and only if):

- Each variable in its own column,
- Each observation in its own row,
- Each value is placed in its own `cell`.


country
year
cases
population
variables


country
year
cases
population
observations


country
year
cases
population
values

The grad program

Is in **tidy** tabular form.

```
## # A tibble: 412 x 16
##    subject Inst  AvNumPubs AvNumCits PctFacGran
##    <chr>   <chr>     <dbl>     <dbl>       <db
##  1 econom… ARIZ…       0.9       1.57        31
##  2 econom… AUBU…       0.79      0.64        77
##  3 econom… BOST…       0.51      1.03        43
##  4 econom… BOST…       0.49      2.66        36
##  5 econom… BRAN…       0.3       3.03        36
##  6 econom… BROW…       0.84      2.31        27
##  7 econom… CALI…       0.99      2.31        56
##  8 econom… CARN…       0.43      1.67        35
```

Different examples of data

For each of these data examples, **let's try together to identify the variables and the observations** - some are HARD!

Your Turn: Genes experiment 🤔

```
## # A tibble: 3 x 12
##   id    `WI-6.R1` `WI-6.R2` `WI-6.R4` `WM-6.R1`
##   <chr>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 Gene…      2.18      2.20      4.20      2.63
## 2 Gene…      1.46     0.585      1.86     0.515
## 3 Gene…      2.03     0.870      3.28     0.533
## # … with 4 more variables: `WI-12.R4` <dbl>, `W
## #   `WM-12.R2` <dbl>, `WM-12.R4` <dbl>
```

02:00

Melbourne weather 😨

```
## # A tibble: 1,593 x 12
##    X1                 X2 X3    X4        X5    X9    
##    <chr>           <dbl> <chr> <chr> <dbl> <dbl> <d
##  1 ASN00086282  1970 07    TMAX    141   124   
##  2 ASN00086282  1970 07    TMIN     80    63   
##  3 ASN00086282  1970 07    PRCP      3    30   
##  4 ASN00086282  1970 08    TMAX    145   128   
##  5 ASN00086282  1970 08    TMIN     50    61   
##  6 ASN00086282  1970 08    PRCP      0    66   
##  7 ASN00086282  1970 09    TMAX    168   168   
##  8 ASN00086282  1970 09    TMIN     19        
##  9 ASN00086282  1970 09    PRCP      0
```

02:00

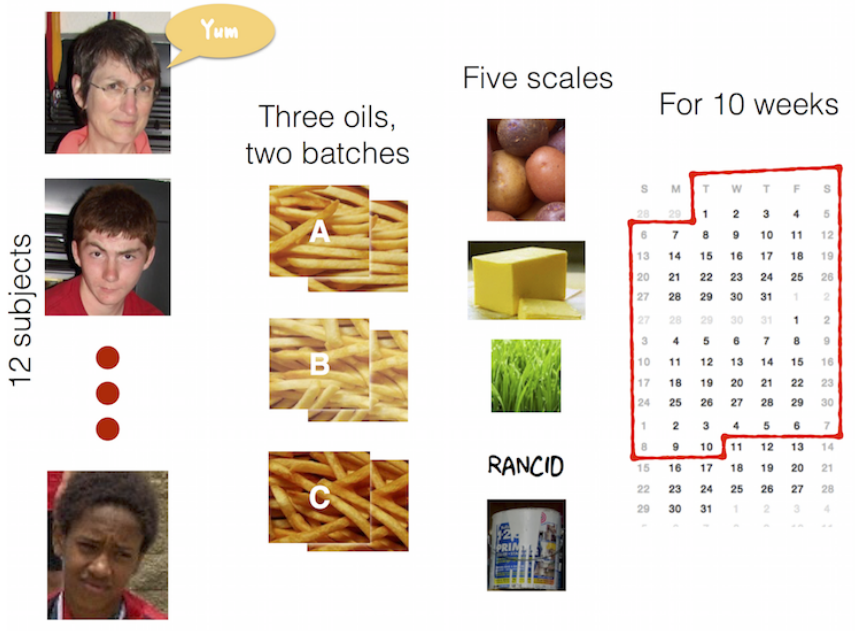
Tuberculosis notifications data taken from WHO 🤧

```
## # A tibble: 3,202 x 22
##    country  year new_sp_m04 new_sp_m514 new_sp_
##    <chr>   <dbl>      <dbl>       <dbl>      <
##  1 Afghan… 1997         NA          NA
##  2 Afghan… 1998         NA          NA
##  3 Afghan… 1999         NA          NA
##  4 Afghan… 2000         NA          NA
##  5 Afghan… 2001         NA          NA
##  6 Afghan… 2002         NA          NA
##  7 Afghan… 2003         NA          NA
##  8 Afghan… 2004         NA          NA
##  9 Afghan… 2005         NA          NA
```

02:00

French fries

- 10 week sensory experiment
- 12 individuals assessed taste of french fries on several scales (how potato-y, buttery, grassy, rancid, paint-y do they taste?)
- fried in one of 3 different oils, replicated twice.



French fries: Variables? Observations?

```
## # A tibble: 696 x 9
##      time treatment subject   rep potato buttery
##     <dbl>     <dbl>   <dbl> <dbl>  <dbl>   <dbl>
##  1     1         1       3     1    2.9       0
##  2     1         1       3     2   14         0
##  3     1         1      10     1   11         6.4
##  4     1         1      10     2    9.9       5.9
##  5     1         1      15     1    1.2       0.1
##  6     1         1      15     2    8.8       3
##  7     1         1      16     1    9         2.6
##  8     1         1      16     2    8.2       4.4
##  9     1         1      19     1    7         3.2
```

Rude Recliners data

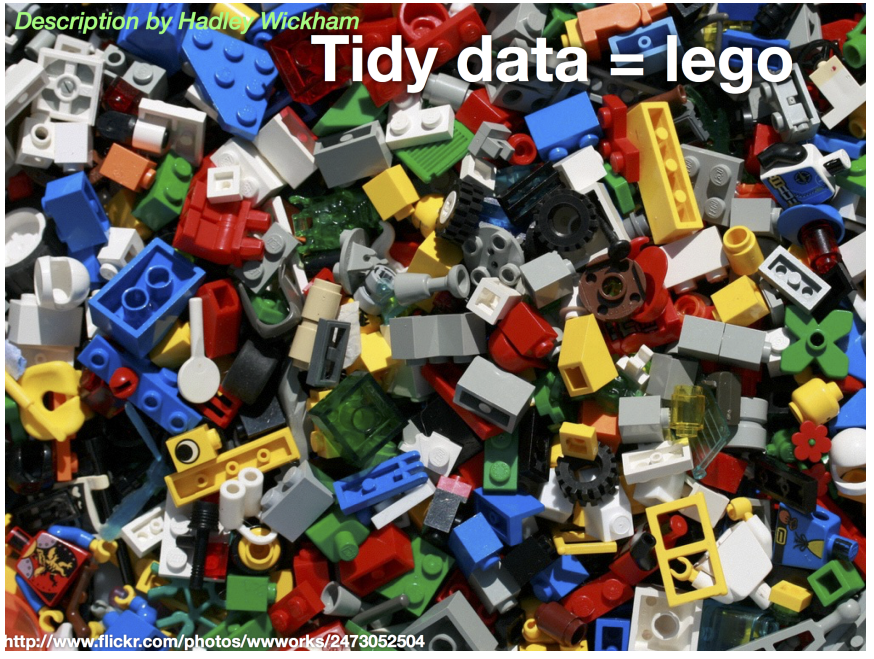
- data is collated from this story: 41% Of Fliers Think You're Rude If You Recline Your Seat
- What are the variables?

```
## # A tibble: 3 x 6
##   V1      `V2:Always` `V2:Usually` `V2:About ha
##   <chr>         <dbl>        <dbl>
## 1 No, no…         124          145
## 2 Yes, s…           9           27
## 3 Yes, v…           3            3
```

Messy vs tidy

Messy data is messy in its own way. You can make unique solutions, but then another data set comes along, and you have to again make a unique solution.

Tidy data can be though of as legos. Once you have this form, you can put it together in so many different ways, to make different analyses.



Data Tidying verbs

- `pivot_longer`: Specify the **names_to** (identifiers) and the **values_to** (measures) to make longer form data.
- `pivot_wider`: Variables split out in columns
- `separate`: Split one column into many

one more time: `pivot_longer`

```
pivot_longer(<DATA>,
             <COLS>,
             <NAMES_TO>
             <VALUES_TO>)
```

- **cols** to select are those that represent values, not variables.
- **names_to** is the name of the variable whose values for the column names.
- **values_to** is the name of the variable whose values are spread over the cells.

pivot_longer: example

```
## # A tibble: 3 x 3
##   country     `1999`
## * <chr>        <int>
## 1 Afghanistan    745
## 2 Brazil       37737
## 3 China       212258
```

```
table4a %>%
  pivot_longer(cols = c
                       names_to
                       values_t
## # A tibble: 6 x 3
##   country     year
##   <chr>       <chr>
## 1 Afghanistan 1999
## 2 Afghanistan 2000
## 3 Brazil      1999
## 4 Brazil      2000
## 5 China       1999
```

Tidying genes data

Tell me what to put in the following?

- **cols** are the columns that represent values, not variables.
- **names_to** is the name of new variable whose values for the column names.
- **values_to** is the name of the new variable whose values are spread over the cells.

```
## # A tibble: 3 x 12
##   id    `WI-6.R1` `WI-6.R2` `WI-6.R4` `WM-6.R1`
##   <chr>     <dbl>     <dbl>     <dbl>     <dbl>
## 1 Gene…      2.18     2.20       4.20      2.63
## 2 Gene…      1.46     0.585      1.86      0.515
## 3 Gene…      2.03     0.870      3.28      0.533
```

Tidy genes data

```
## # A tibble: 3 x 12
##   id    `WI-6.R1` `W
##   <chr>     <dbl>
## 1 Gene…      2.18
## 2 Gene…      1.46
## 3 Gene…      2.03
## # … with 4 more vari
## #   `WM-12.R2` <dbl>
```

```
genes_long <- genes %>%
  pivot_longer(cols = -id,
               names_to =
               values_to =
```

```
genes_long
## # A tibble: 33 x 3
##    id     variable  exp
##    <chr>  <chr>     <dbl
##  1 Gene 1 WI-6.R1    2.1
##  2 Gene 1 WI-6.R2    2.2
##  3 Gene 1 WI-6.R4    4.2
```

Separate columns

```
## # A tibble: 33 x 3
##    id      variable    <
##    <chr>   <chr>       <
##  1 Gene 1  WI-6.R1
##  2 Gene 1  WI-6.R2
##  3 Gene 1  WI-6.R4
##  4 Gene 1  WM-6.R1
##  5 Gene 1  WM-6.R2
##  6 Gene 1  WI-12.R1
##  7 Gene 1  WI-12.R2
##  8 Gene 1  WI-12.R4
##  9 Gene 1  WM-12.R1
```

```
genes_long %>%
  separate(col = variab
           into = c("tr
## # A tibble: 33 x 4
##    id      trt   left
##    <chr>   <chr> <chr
##  1 Gene 1  WI    6.R1
##  2 Gene 1  WI    6.R2
##  3 Gene 1  WI    6.R4
##  4 Gene 1  WM    6.R1
##  5 Gene 1  WM    6.R2
##  6 Gene 1  WI    12.R
```

Separate columns

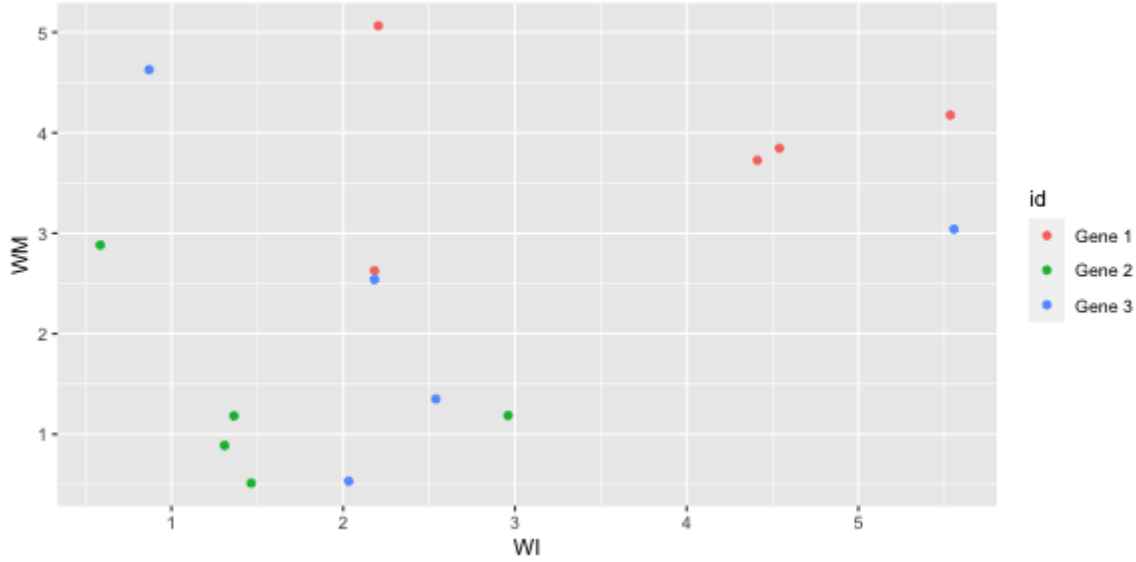
```
genes_long_tidy <- genes_long %>%
  separate(variable, c("trt", "leftover"), "-") %
  separate(leftover, c("time", "rep"), "\\.")

genes_long_tidy
## # A tibble: 33 x 5
##    id      trt   time  rep    expr
##    <chr>   <chr> <chr> <chr> <dbl>
##  1 Gene 1 WI    6     R1     2.18
##  2 Gene 1 WI    6     R2     2.20
##  3 Gene 1 WI    6     R4     4.20
##  4 Gene 1 WM    6     R1     2.63
```

Now let's use `pivot_wider` to examine different aspects

Examine treatments against each other

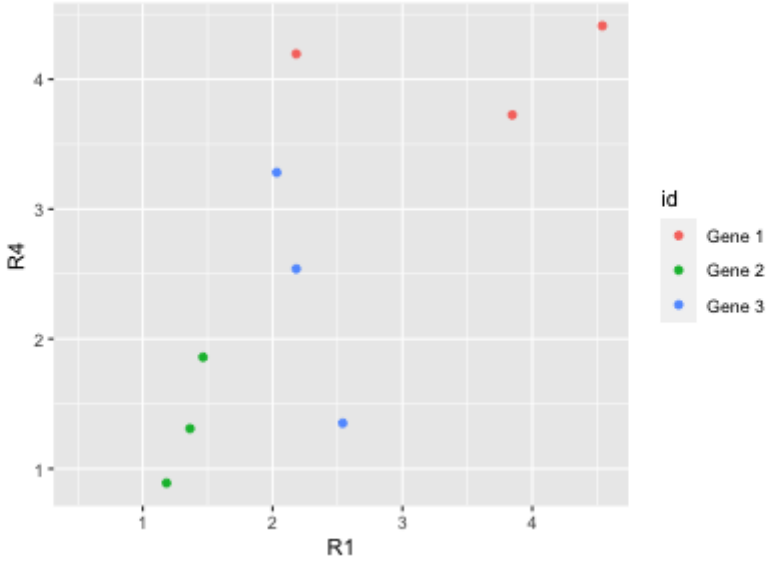
```
genes_long_tidy %>%
  pivot_wider(id_cols =
                names_fro
                values_fr
  ggplot(aes(x=WI, y=WM
```



Generally, some negative association within each gene, WM is low if WI is high.

Examine replicates against each other

```
genes_long_tidy %>%
  pivot_wider(id_cols =
              names_fro
              values_fr
  ggplot(aes(x=R1, y=R4
  geom_point() + coord_
```



Roughly, replicate 4 is like replicate 1, eg if one is low, the other is low.

That's a good thing, that the replicates are fairly similar.

Your turn: Demonstrate with koala bilby data (live code)

Here is a little data to practice `pivot_longer`, `pivot_wider` and `separate` on.

- Read over `koala-bilby.Rmd`
- pivot_longer the data into long form, naming the two new variables, `label` and `count`
- Separate the labels into two new variables, `animal`, `state`
- pivot_wider the long form data into wide form, where the columns are the states.
- pivot_wider the long form data into wide form, where the columns are the animals.

Exercise 1: Rude Recliners

- Open `rude-recliners.Rmd`
- This contains data from the article 41% Of Fliers Think You're Rude If You Recline Your Seat.
- V1 is the response to question: "Is it rude to recline your seat on a plane?"
- V2 is the response to question: "Do you ever recline your seat when you fly?".

```
## # A tibble: 3 x 6
##   V1      `V2:Always` `V2:Usually` `V2:About ha
##   <chr>         <dbl>        <dbl>
## 1 No, no…         124          145
## 2 Yes, s…           9           27
```

Exercise 1: Rude Recliners (15 minutes)

Answer the following questions in the rmarkdown document.

- A) What are the variables and observations in this data?
- 1B) Put the data in tidy long form (using the names V2 as the key variable, and `count` as the value).
- 1C) Use the `rename` function to make the variable names a little shorter.

Exercise 1: Answers

Your Turn: Turn to the people next to you and ask 2 questions:

- Are you more of a dog or a cat person?
- What languages do you know how to speak?

03:00

Exercise 2: Tuberculosis Incidence data (15 minutes)

Open: `tb-incidence.Rmd`

Tidy the TB incidence data, using the Rmd to prompt questions.

Exercise 3: Currency rates (15 minutes)

- open `currency-rates.Rmd`
- read in `rates.csv`
- Answer the following questions:

1. What are the variables and observations?
2. pivot_longer the five currencies, AUD, GBP, JPY, CNY, CAD, make it into tidy long form.
3. Make line plots of the currencies, describe the similarities and differences between the currencies.

Exercise 4: Australian Airport Passengers (optional!)

- Open `oz-airport.Rmd`
- Contains data from the web site Department of Infrastructure, Regional Development and Cities, containing data on Airport Traffic Data 1985–86 to 2017–18.
- Read the dataset, into R, naming it `passengers`
- Tidy the data, to produce a data set with these columns
  - airport: all of the airports.
  - year
  - type_of_flight: DOMESTIC, INTERNATIONAL
  - bound: IN or OUT

Lab quiz

Time to take the lab quiz.

Learning is where you:

1. Receive information accurately
2. Remember the information (long term memory)
3. In such a way that you can reapply the information when appropriate

Your Turn:

Go to the data source at this link: bit.ly/dmac-noaa-data

- "Which is the best description of the temperature units?"
- "What is the best description of the precipitation units"
- "What does -9999 mean?"

Recap

- Traffic Light System: Green = "good!" ; Red = "Help!"
- R + Rstudio
- Functions are _
- columns in data frames are accessed with _ ? If you have questions, place a red sticky note on your laptop.

If you are done, place a green sticky on your laptop

Traffic Light System

Red Post it

- I need a hand
- Slow down

Green Post it

- I am up to speed
- I have completed the thing

That's it!