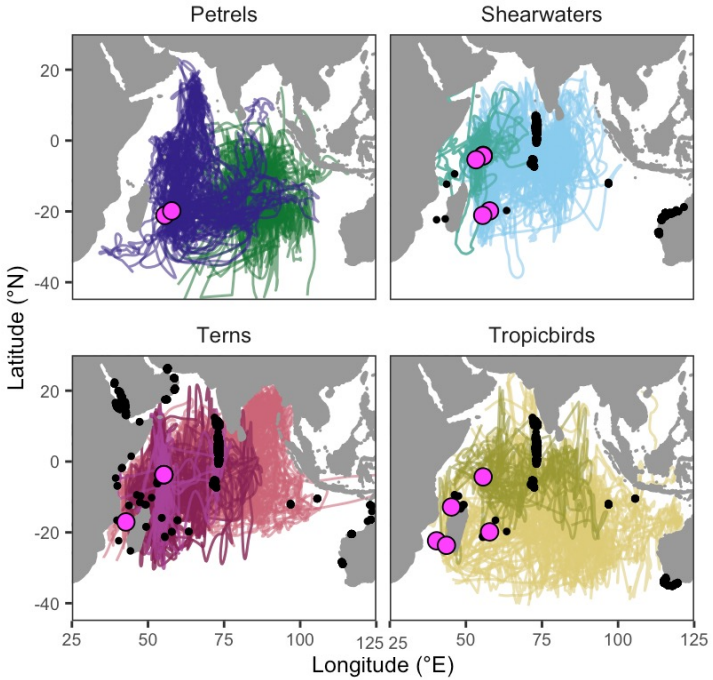
Reproducible data manipulation (in R)

Microteach, Alice Trevail



University *of* Exeter

Environment and Sustainability Institute

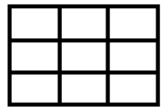
About today: Intended learning outcomes

1. Describe why we should manipulate our data in a reproducible way
2. Manipulate data = write simple code to change data from wide to long format

- No need to take part in activities if you would rather watch

Questions for you

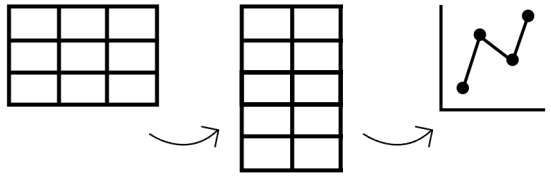
Yes/No – please raise your hands if ‘yes’



1: do you work with data?

2: do you manage data in excel?

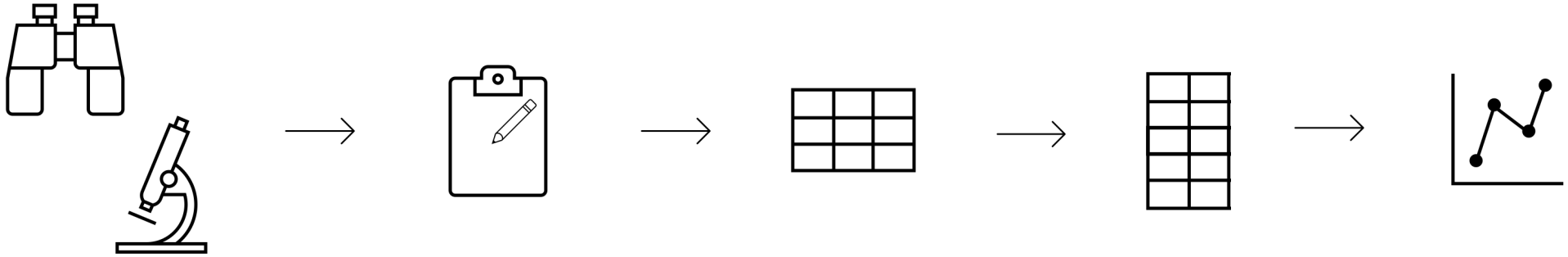
3: have you used the software R?



4: have you used R for data manipulation?

What is data manipulation?

- The way we collect data can be different to how we analyse it



- E.g., different rows/column structure

- To solve this problem, we need to re-organize our data

Why reproducible data manipulation?

- **Requirement:** journals & funding bodies mandate open access data & code
- **Avoid errors,** no more copy-paste
- Can be quick, easy, and fun!

Why reproducible data manipulation? (in R)


```
glimpse(penguins_raw)
```

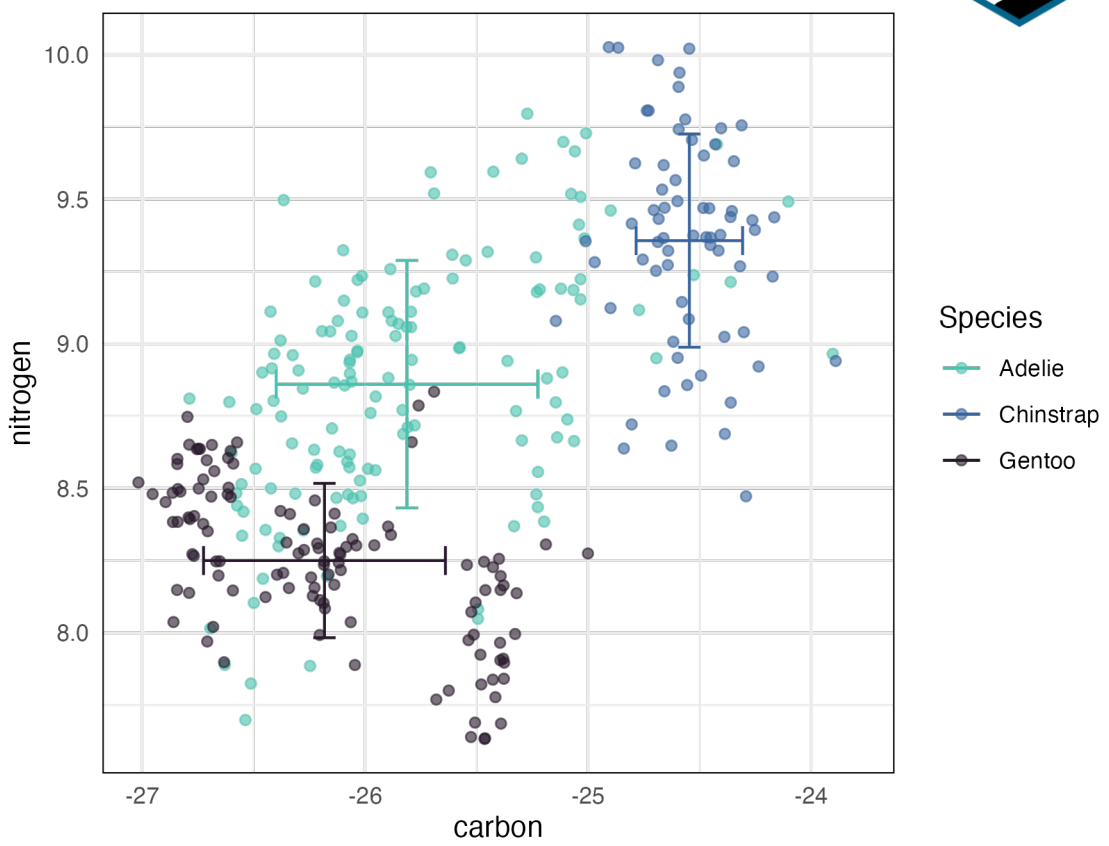
```
Rows: 344
Columns: 17
$ studyName             <chr> "PAL0708", "PAL0708", "PAL0708", "PAL0708", "PAL…
$ `Sample Number`       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 1…
$ Species               <chr> "Adelie Penguin (Pygoscelis adeliae)", "Adelie P…
$ Region                <chr> "Anvers", "Anvers", "Anvers", "Anvers", "Anvers"…
$ Island                <chr> "Torgersen", "Torgersen", "Torgersen", "Torgerse…
$ Stage                 <chr> "Adult, 1 Egg Stage", "Adult, 1 Egg Stage", "Adu…
$ `Individual ID`       <chr> "N1A1", "N1A2", "N2A1", "N2A2", "N3A1", "N3A2", …
$ `Clutch Completion`   <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "No", …
$ `Date Egg`            <date> 2007-11-11, 2007-11-11, 2007-11-16, 2007-11-16,…
$ `Culmen Length (mm)`  <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38.9, 39.2, 34…
$ `Culmen Depth (mm)`   <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17.8, 19.6, 18…
$ `Flipper Length (mm)` <dbl> 181, 186, 195, NA, 193, 190, 181, 195, 193, 190,…
$ `Body Mass (g)`       <dbl> 3750, 3800, 3250, NA, 3450, 3650, 3625, 4675, 34…
$ Sex                   <chr> "MALE", "FEMALE", "FEMALE", NA, "FEMALE", "MALE"…
$ `Delta 15 N (o/oo)`   <dbl> NA, 8.94956, 8.36821, NA, 8.76651, 8.66496, 9.18…
$ `Delta 13 C (o/oo)`   <dbl> NA, -24.69454, -25.33302, NA, -25.32426, -25.298…
$ Comments              <chr> "Not enough blood for isotopes.", NA, NA, "Adult…
```

Why reproducible data manipulation? (in R)

27 lines of code:

```
penguins_summary_isotopes <- penguins_example %>%
  pivot_longer(cols = carbon:nitrogen, names_to = "isotope", values_to = "value") %>%
  group_by(species, isotope) %>%
  summarize(mean = mean(value, na.rm = T),
            sd =  sd(value, na.rm = T)) %>%
  pivot_wider(id_cols = species, names_from = isotope, values_from=c(mean, sd))

ggplot(penguins_example, aes(x = carbon, y = nitrogen, col = species)) +
  geom_point(alpha = 0.6)+
  geom_errorbar(data = penguins_summary_isotopes,
                aes(x = mean_carbon, ymax = mean_nitrogen+sd_nitrogen, ymin = mean_nitrogen-sd_nitrogen, col = species),
                inherit.aes = F, width = 0.1)+
  geom_errorbar(data = penguins_summary_isotopes,
                aes(y = mean_nitrogen, xmax = mean_carbon+sd_carbon, xmin = mean_carbon-sd_carbon, col = species),
                inherit.aes = F, width = 0.1)+
  scale_colour_viridis_d(option = "mako", begin = 0.75, end = 0.1, name = "Species")+
  labs(title = "Isotope signatures among Palmer Penguins")+
  theme_minimal()+
  theme(panel.border = element_rect(fill = NA))
```



Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. doi: 10.5281/zenodo.3960218

`pivot_longer()`: An example

Q: do quadrats have different species counts?

Quadrat	SpeciesA	SpeciesB	SpeciesC
Q_1	10.00	12.00	15.00
Q_2	4.00	3.00	4.00

Wide data =
Observations across multiple columns and rows

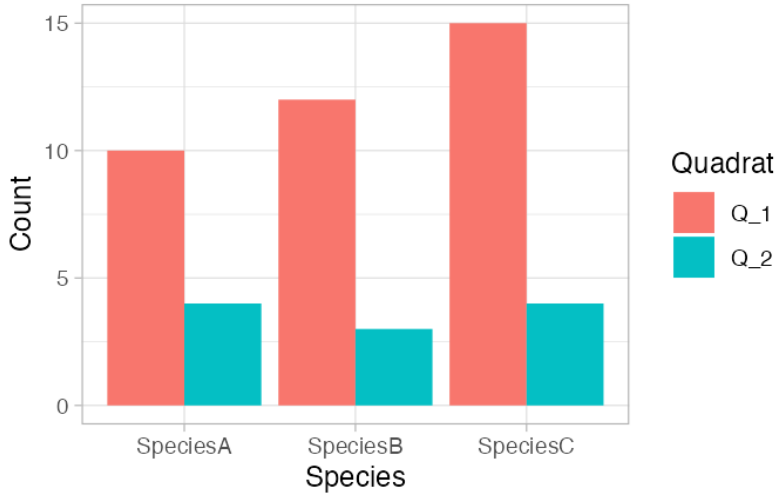
`pivot_longer()`: An example

Q: do quadrats have different species counts?

Quadrat	SpeciesA	SpeciesB	SpeciesC
Q_1	10.00	12.00	15.00
Q_2	4.00	3.00	4.00

```
pivot_longer(data,
        cols = SpeciesA:SpeciesC,
        names_to =  Species ,
        values_to =  Count  )
```

Quadrat	Species	Count
Q_1	SpeciesA	10.00
Q_1	SpeciesB	12.00
Q_1	SpeciesC	15.00
Q_2	SpeciesA	4.00
Q_2	SpeciesB	3.00
Q_2	SpeciesC	4.00



`pivot_longer()`: An example

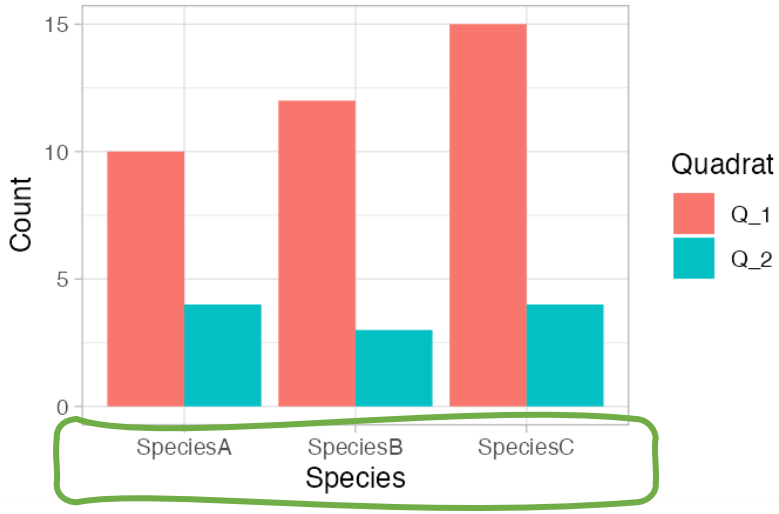
Q: do quadrats have different species counts?

Quadrat	SpeciesA	SpeciesB	SpeciesC
Q_1	10.00	12.00	15.00
Q_2	4.00	3.00	4.00

Columns containing data

```
pivot_longer(data,
        cols = SpeciesA:SpeciesC,
        names_to =  Species ,
        values_to =  Count  )
```

Quadrat	Species	Count
Q_1	SpeciesA	10.00
Q_1	SpeciesB	12.00
Q_1	SpeciesC	15.00
Q_2	SpeciesA	4.00
Q_2	SpeciesB	3.00
Q_2	SpeciesC	4.00



`pivot_longer()`: An example

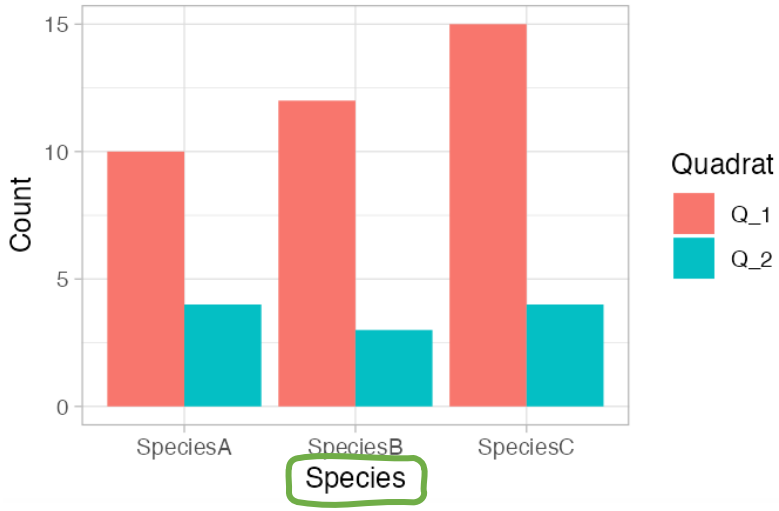
Q: do quadrats have different species counts?

Quadrat	SpeciesA	SpeciesB	SpeciesC
Q_1	10.00	12.00	15.00
Q_2	4.00	3.00	4.00

Columns containing data

```
pivot_longer(data,
        cols = SpeciesA:SpeciesC,
        names_to =  Species ,
        values_to =  Count  )
```

Quadrat	Species	Count
Q_1	SpeciesA	10.00
Q_1	SpeciesB	12.00
Q_1	SpeciesC	15.00
Q_2	SpeciesA	4.00
Q_2	SpeciesB	3.00
Q_2	SpeciesC	4.00



`pivot_longer()`: An example

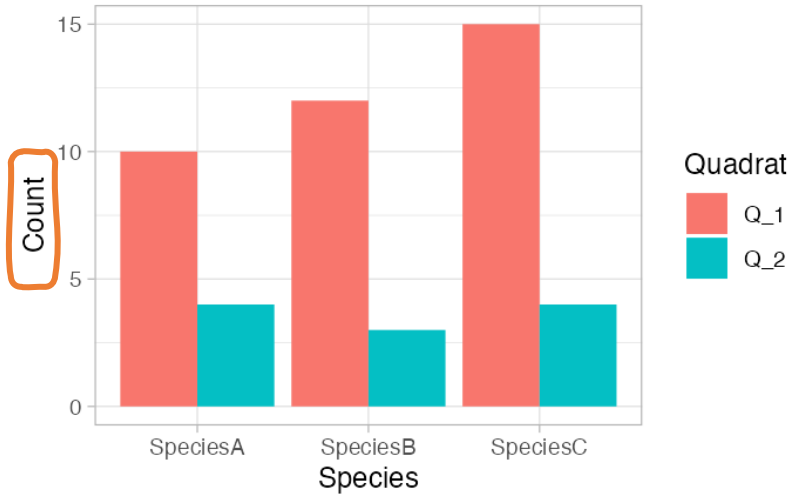
Q: do quadrats have different species counts?

Quadrat	SpeciesA	SpeciesB	SpeciesC
Q_1	10.00	12.00	15.00
Q_2	4.00	3.00	4.00

Values containing observations

```
pivot_longer(data,
        cols = SpeciesA:SpeciesC,
        names_to =  Species ,
        values_to =  Count  )
```

Quadrat	Species	Count
Q_1	SpeciesA	10.00
Q_1	SpeciesB	12.00
Q_1	SpeciesC	15.00
Q_2	SpeciesA	4.00
Q_2	SpeciesB	3.00
Q_2	SpeciesC	4.00



`pivot_longer()`: Your turn

Q: how do runners times change?

Athlete	run1	run2	run3
Alice	28:52	25:29	27:10
Olli	22:39	22:25	20:56

```
pivot_longer(data,
        cols = run1:run3,
        names_to =   ,
        values_to =    )
```

Your task =
Fill in the blanks

`pivot_longer()`: Your turn

Q: how do runners times change?

R Shiny app
alicetrevail.shinyapps.io/pivot_learn

Microteach: Learn some data manipulation! | Quadrats | Finish Times | Penguins

Find other examples here

Athlete	run1	run2	run3
Alice	28:52	25:29	27:10
Olli	22:39	22:25	20:56

Choose new column names

then Click here to pivot!

Enter name for new column that will contain old column names

names_to =

Enter name for new column that will contain values

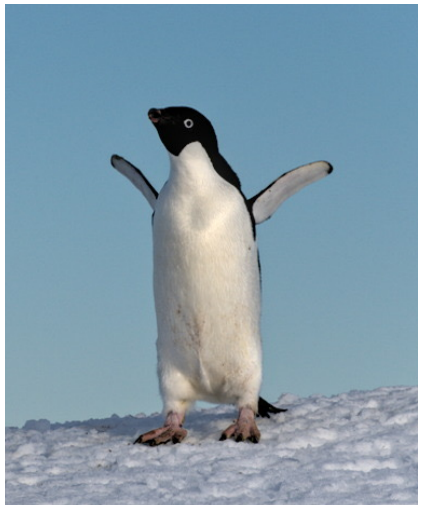
values_to =

Click here to `pivot_longer()`

```
pivot_longer(data,
        cols = run1:run3,
        names_to =   ,
        values_to =    )
```

Use these boxes to fill in the code

`pivot_longer()`: Test!

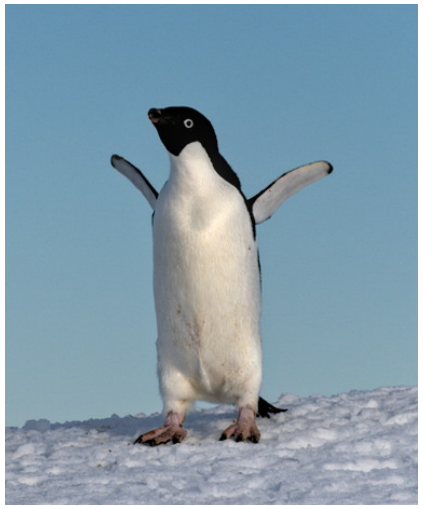


Q: how big are different penguin species?

species	bill_length_mm	flipper_length_mm	body_mass_g
Adelie	39.10	181	3750
Chinstrap	46.50	192	3500
Gentoo	46.10	211	4500

```
1       (data,
        cols =  bill_length_mm : body_mass_g ,
        2     =  Measurement ,
        values_to =  Size  )
```

`pivot_longer()`: Test!



Q: how big are different penguin species?

species	bill_length_mm	flipper_length_mm	body_mass_g
Adelie	39.10	181	3750
Chinstrap	46.50	192	3500
Gentoo	46.10	211	4500

```
pivot_longer( 1 ,
        2  =  bill_length_mm : body_mass_g ,
        names_to =  Measurement ,
        3       =  Size  )
```

Recap

We have learnt:

1. Why we should manipulate our data in R = *Reproducible & fun*
2. How to manipulate data from wide to long format = `pivot_longer()`

Find out more:

exeter-data-analytics.github.io

Workshop next Tuesday, 21st November

Exchange Lecture Theatre, 3-5pm