

Thanks a lot for participating!



Artwork by [Allison Horst](#)

Bring your own data

Day 4 - Introduction to Data Analysis with R

Selina Baldauf

Freie Universität Berlin - Theoretical Ecology

March 3, 2025

Organization

Schedule of today

- **Now - 14** (or 14.30 if you are enthusiastic still): Work on the data set(s)
 - Take break(s) as best fits your needs
- **14 (14.30) - 15**: Short feedback round
 - What did you find out about your data set? Plots, summaries, ...
 - Which methods did you use?
 - Did you learn something new?
 - Was there something you struggled with?
 - ...
- **15-16**: Feedback, conclusion

Data set 1: What makes a good wine?

Physicochemical properties of wine and quality judgements

```
Error in loadNamespace(x): there is no package called 'EcoData'
```

Data set 1: What makes a good wine?

Ideas - know methods

- Plot of wine quality against chemical properties
- Plot of distribution of chemical properties
- Summary tables using `dplyr`

Ideas - new methods

- **Correlation plots:** How are the different wine properties correlated with each other?
- **PCA:** How are the wine properties related to each other?



Frederik Vandaele - originally posted to Flickr as Château Pétrus, CC BY 2.0,
<https://commons.wikimedia.org/w/index.php?curid=5145286>

Data set 1: What makes a good wine?

Hints

- Transform the quality column to a factor before plotting: use `dplyr::mutate` and `as.factor()` to transform the column
- Try the `janitor::clean_names()` function



Frederik Vandaele - originally posted to Flickr as Château Pétus, CC BY 2.0,
<https://commons.wikimedia.org/w/index.php?curid=5145286>

Data set 2: Paralympic games from 1980-2016

Most important variables:

| variable | class | description |
|----------|-----------|------------------------------------|
| gender | character | Binary gender |
| event | character | Event name |
| medal | character | Medal type |
| athlete | character | Athlete name (LAST NAME first name |
| abb | character | Country abbreviation |
| country | character | Country name |
| type | character | Type of sport |
| year | double | year of games |

Data set 2: Paralympic games from 1980-2016

Get the data:

```
athletes <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-08
```

Ideas - know methods

- Create **summaries** of medal counts for different groups with **dplyr**
- Did the ratio of men/women winning medals change over time?
- Which countries were the most successful ones? Does this differ between sports type?
 - Which types of sports accumulated the most medals?
- **Make plots** such as:
 - Age distribution of athletes winning gold, silver and bronze
 - Compare the total number of medals over the years between winter and summer Olympics

Data set 2: Paralympic games from 1980-2016

Hints

- To reduce complexity of the data, first filter only the athletes that won a medal
(`!is.na(medal)`)

Data set 3: Crab data set

Atlantic marsh fiddler crab (*Minuca pugnax*)

- Crab from Florida is expanding northward due to ocean warming
- Data on 13 marshes across a range of latitude in the USA
- Recording of the size of the crab
- Rather small and good to handle



Image by LTER under CC BY-SA 4.0

Data set 3: Crab data set

| | date | latitude | site | size | air_temp | air_temp_sd | water_temp | water_temp_sd |
|---|------------|----------|------|-------|----------|-------------|------------|---------------|
| 1 | 2016-07-24 | 30 | GTM | 12.43 | 21.792 | 6.391 | 24.502 | 6.121 |
| 2 | 2016-07-24 | 30 | GTM | 14.18 | 21.792 | 6.391 | 24.502 | 6.121 |
| 3 | 2016-07-24 | 30 | GTM | 14.52 | 21.792 | 6.391 | 24.502 | 6.121 |
| 4 | 2016-07-24 | 30 | GTM | 12.94 | 21.792 | 6.391 | 24.502 | 6.121 |
| 5 | 2016-07-24 | 30 | GTM | 12.45 | 21.792 | 6.391 | 24.502 | 6.121 |
| 6 | 2016-07-24 | 30 | GTM | 12.99 | 21.792 | 6.391 | 24.502 | 6.121 |

| | name |
|---|------------------------------|
| 1 | Guana Tolomoto Matanzas NERR |
| 2 | Guana Tolomoto Matanzas NERR |
| 3 | Guana Tolomoto Matanzas NERR |
| 4 | Guana Tolomoto Matanzas NERR |
| 5 | Guana Tolomoto Matanzas NERR |
| 6 | Guana Tolomoto Matanzas NERR |

Source: [Johnson, D. 2019](#). Fiddler crab body size in salt marshes from Florida to Massachusetts, USA at PIE and VCR

Selina Baldauf // Bring your own data

Data set 3: Crab data set

Ideas - known methods

- Explore Bergmann's rule (organisms are large in higher latitudes)
- t-tests to compare size between locations
- Plot relationship between latitude and size
- Plot distributions of variables



Image by LTER under CC BY-SA 4.0

Data set 4: Ice cover and temperature

Temperature and ice duration on lakes since 19th century

- 2 data sets with measurements of
 - ice start, end and duration on 2 lakes in Wisconsin
 - daily air temperature since 1870
- Explore the effect of climate change on ice cover



Image by LTER under CC BY-SA 4.0

Source ice data: [Magnuson, J.J., S.R. Carpenter, and E.H. Stanley. 2021.](#) North Temperate Lakes LTER: Ice Duration - Madison Lakes Area 1853 - current ver 35. Environmental Data Initiative.

Data set 4: Ice cover and temperature

Ice data:

| | lakeid | ice_on | ice_off | ice_duration | year |
|---|--------------|------------|------------|--------------|------|
| 1 | Lake Mendota | <NA> | 1853-04-05 | NA | 1852 |
| 2 | Lake Mendota | 1853-12-27 | <NA> | NA | 1853 |
| 3 | Lake Mendota | 1855-12-18 | 1856-04-14 | 118 | 1855 |
| 4 | Lake Mendota | 1856-12-06 | 1857-05-06 | 151 | 1856 |
| 5 | Lake Mendota | 1857-11-25 | 1858-03-26 | 121 | 1857 |
| 6 | Lake Mendota | 1858-12-08 | 1859-03-14 | 96 | 1858 |

Temperature data:

| | sampledate | year | ave_air_temp_adjusted |
|---|------------|------|-----------------------|
| 1 | 1870-06-05 | 1870 | 20.0 |
| 2 | 1870-06-06 | 1870 | 18.3 |
| 3 | 1870-06-07 | 1870 | 17.5 |
| 4 | 1870-06-09 | 1870 | 13.3 |
| 5 | 1870-06-10 | 1870 | 13.9 |
| 6 | 1870-06-11 | 1870 | 15.0 |

Data set 4: Ice cover and temperature

Ideas - known methods

- How did ice cover duration change over the years?
- How did air temperature change over the years?
 - Summarize mean annual temperature or mean temperature in winter
- How do ice duration on the lakes correlate with temperature (e.g. with mean winter temperature)



Image by LTER under CC BY-SA 4.0

Data set 4: Ice cover and temperature

Hints

- For some plots it might make sense to summarize annual means first
- Use `dplyr::left_join` to combined the tables with annual mean temperature and ice duration
 - Have a look at the last slides of the `dplyr` session or look at the help
 - Ask me if you need help with this

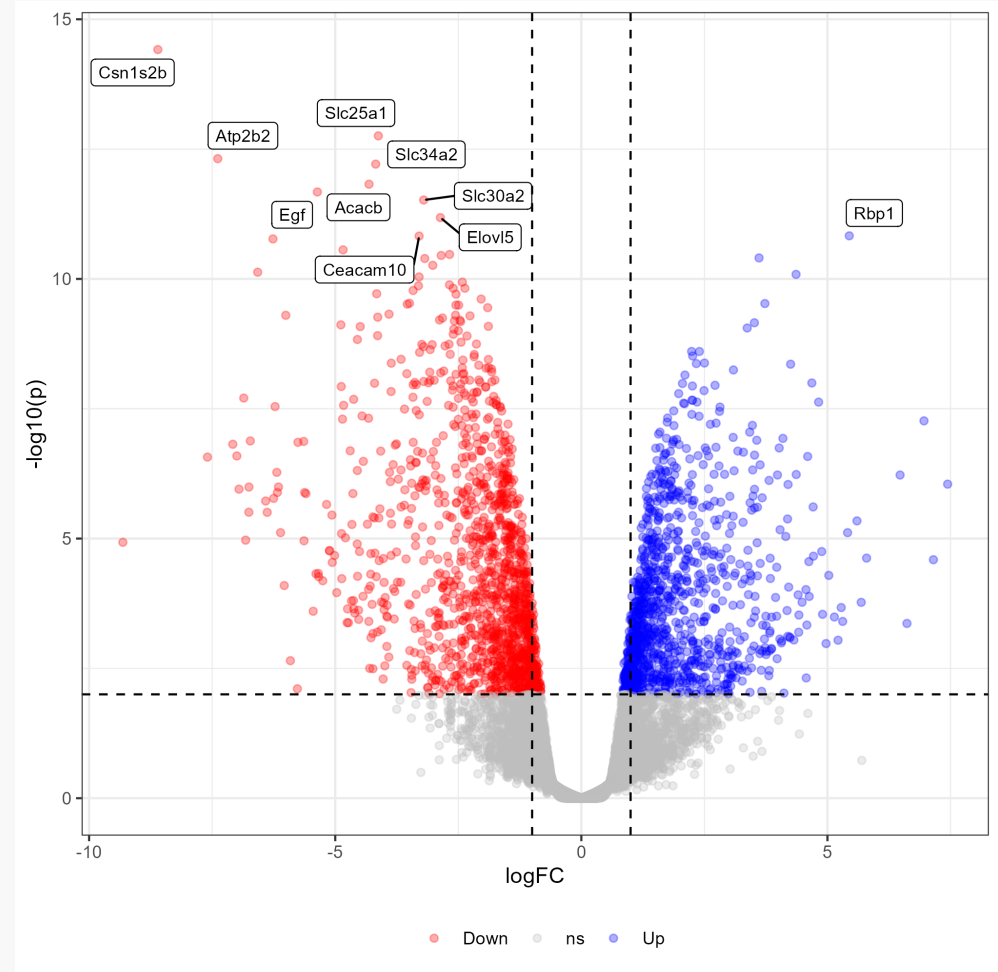
Data set 5: RNAseq data

- Data from [FU et al. 2015, Nature Cell Biology](#)
- Data found via [Tutorial on heat maps](#) using this data
- 3 csv files:
 - **heatmap_genes.csv**: A list of the names of interesting genes to look at (Genes used in Figure 6b in paper)
 - **DE_results.csv**: Gene expression in luminal cells in pregnant versus lactating mice
 - logFC, AveExpr, t, p-value
 - Also contains non-significantly expressed genes
 - **normalized_counts**: Normalized counts for genes for the different samples

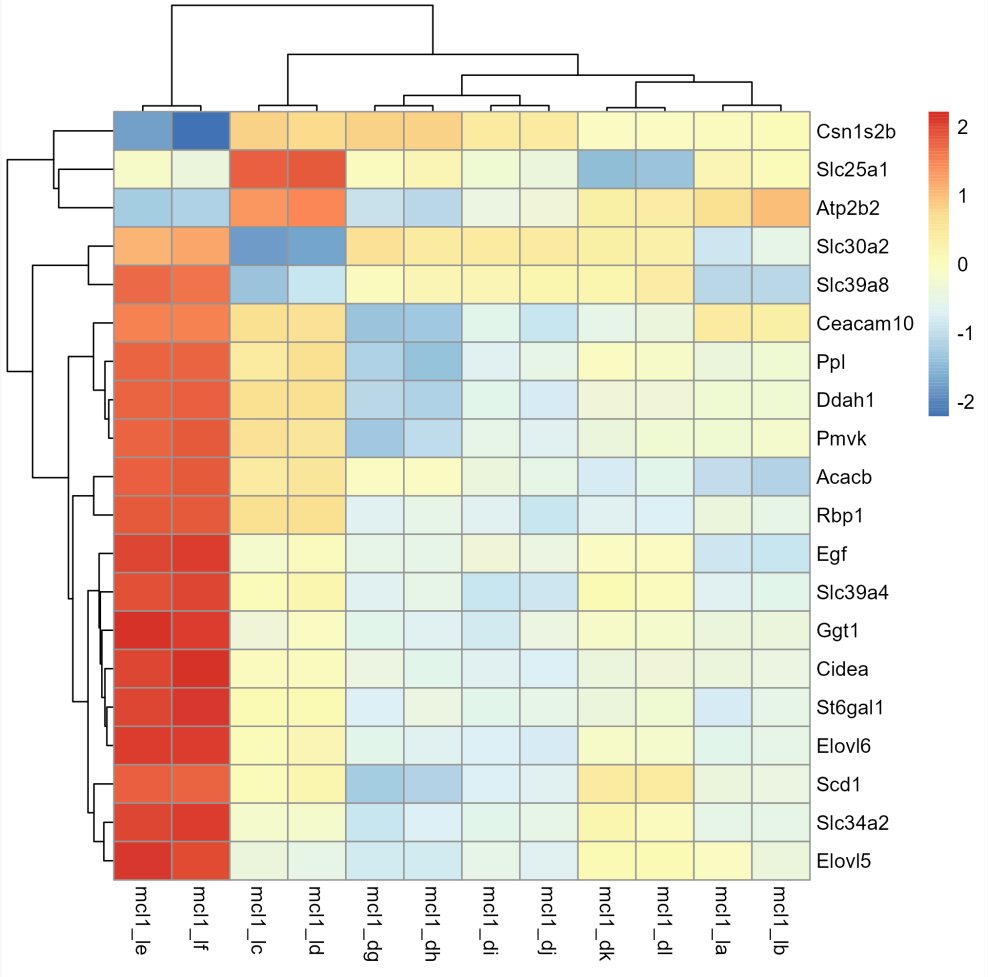
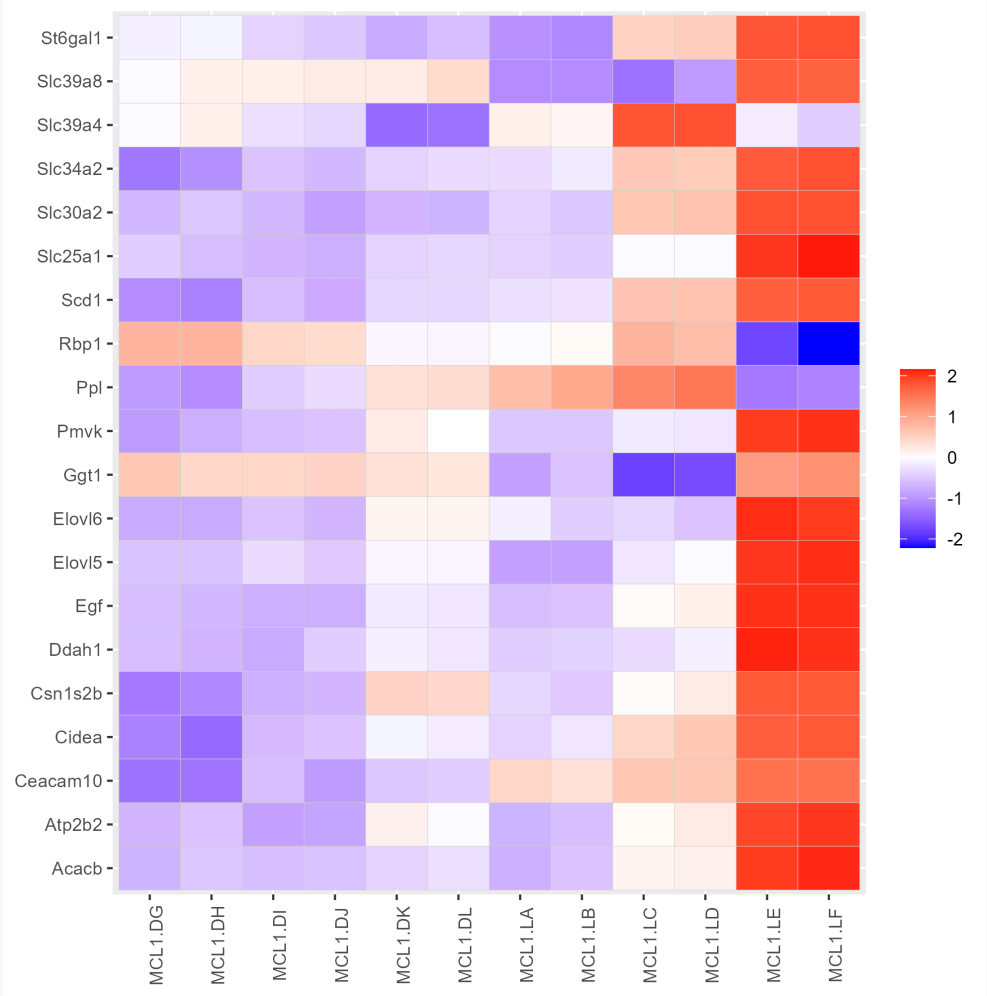
Data set 5: RNAseq data

Ideas:

- Create a heatmap of the top 20 most significant genes (see plot in the tutorial)
- Create a heatmap of the interesting genes (see Fig. 6 in the paper)
- Create a volcano plot of the data similar to the one [here](#)



Data set 5: RNAseq data



Data set 5: RNAseq data

Some tips:

Data cleaning:

- Read in the data and then use the `janitor::clean_names` function to make the column headers nicer
- Join `DE_results` and `normalized_counts` by their shared columns
- Use `select` to remove columns you don't need for analysis to get a better overview
- Filter only significant genes ([tutorial](#)) defines them as `p_value < 0.01 & abs(logFC) > 0.58`

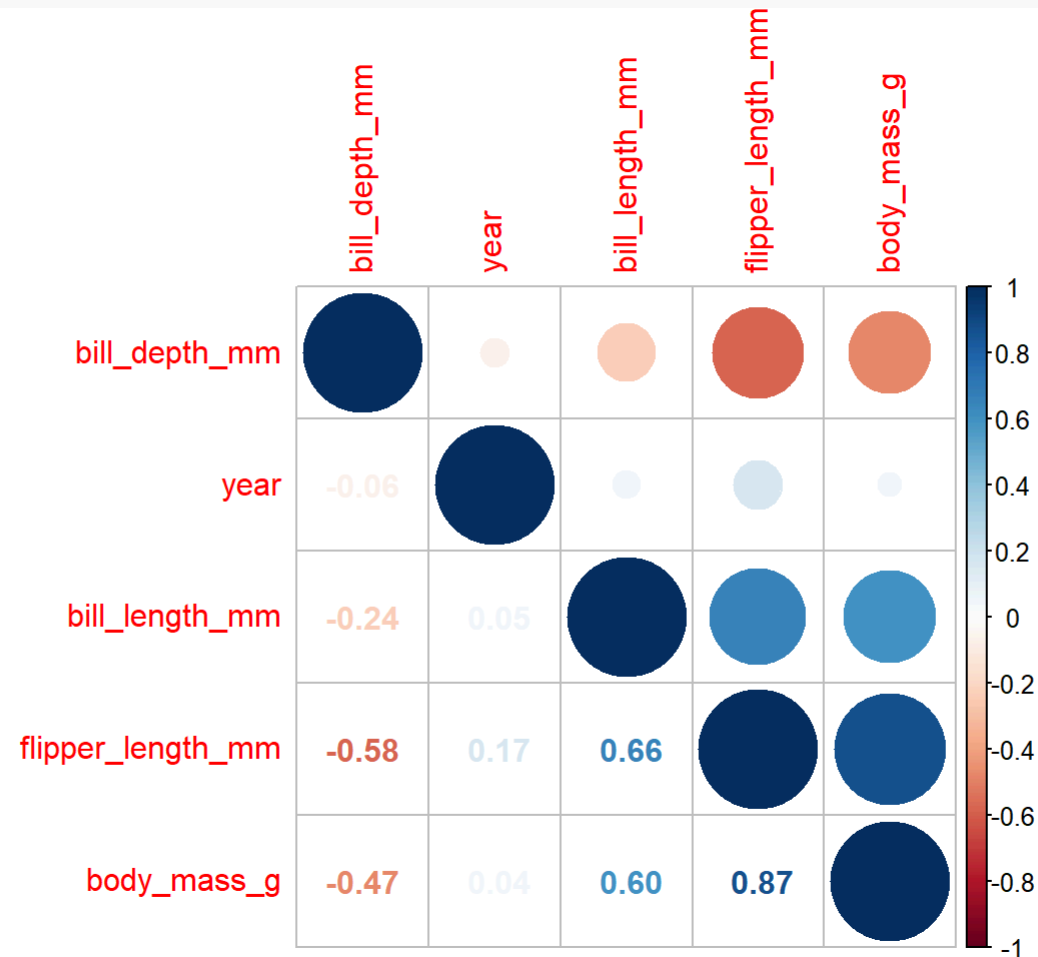
Data set 5: RNAseq data

Some tips:

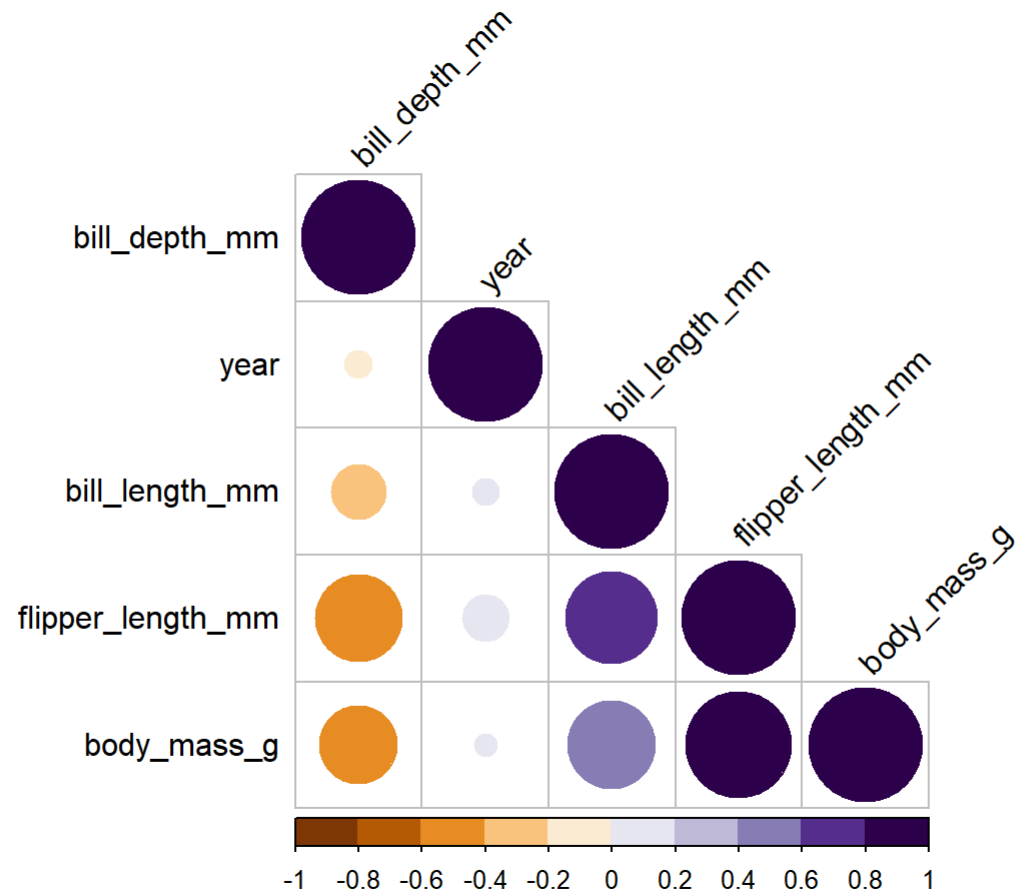
Data analysis:

- Heatmap with ggplot or with `pheatmap::pheatmap()`
 - `pheatmap` takes a matrix as input (use `as_matrix` on tibble to transform)
- scale the counts -> have a look at the `scale` function
 - `pheatmap` can scale but with ggplot you have to scale before plotting

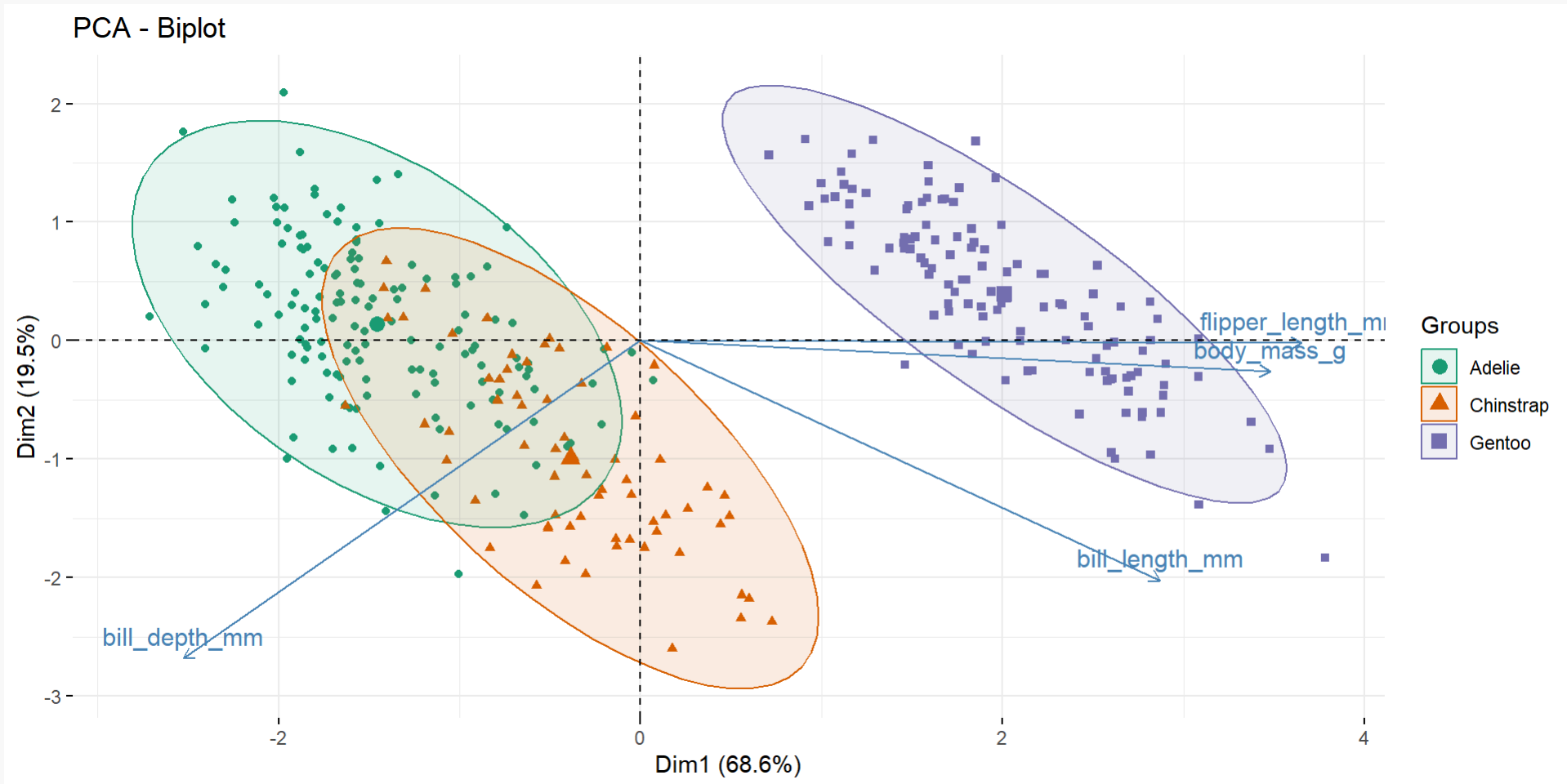
New methods: Correlation plots



New methods: Correlation plots



New methods: PCAS



New methods: Correlation plots and PCAS

- `corrplot` package for correlation plots
- `factoextra` package for PCA visualization
- Tutorial for PCAs in R, PCA tutorial for penguins
- Correlations and PCAs do not work with `NA` values: use `tidyr::drop_na()` to remove all `NA` values from the data first
- These plots work for: penguins, wine, piecrab dataset

Some general tips

- First make a plan:
 - What do you want to achieve and what are the steps?
 - Try to think in technical terms
 - Start with something small, e.g. reading in the data and bringing it into the right format.
 - If you want, stop by in general to discuss your plan or write me in the chat
- Google
- If you get stuck, ask in the chat or stop by in General
- Have a look at the [package cheat sheets](#)

Now you

Working with real research data

Meet in your group (if you want)

Work on your data set

Take breaks as you need and **be back** at 2 p.m.

Keep an eye on your **group** and the **general chat**

Sharing

In 1-2 mins:

- What was the highlight of your analysis?
 - Your favorite plot
 - Some cool code
 - A problem that you finally solved
 - Something new you learned
- What was difficult?
- If you want: Share a screenshot in the chat or share your screen

Feedback

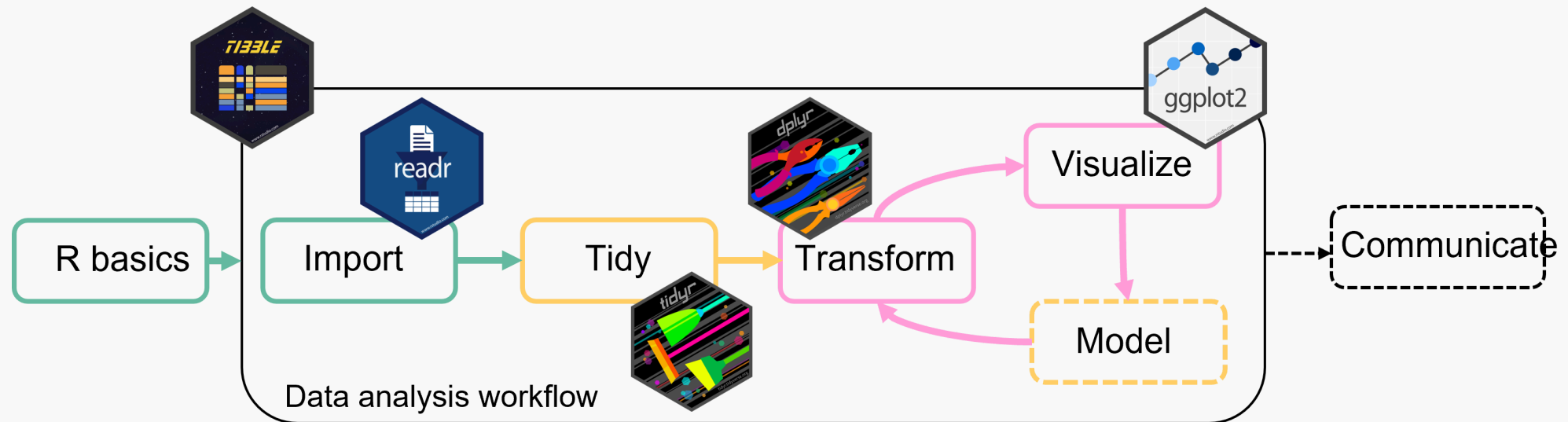
Please take 10 mins to complete the feedback survey for the Graduate center (don't use Internet Explorer)

<https://votingo.cedis.fu-berlin.de/PCNLP3>

Feedback

- Any other feedback or comments from your side?

Conclusion

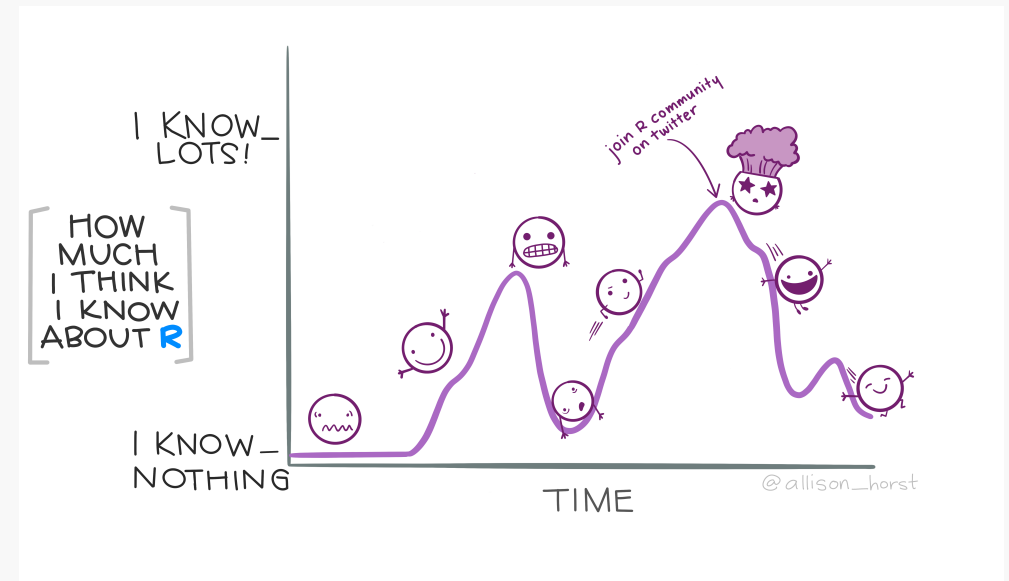


We learned a lot of stuff!

Conclusion

How to continue from here?

- Learning by doing!
- Have a look at some [online resources](#), I recommend the R for Data Science book by Hadley Wickham
- If you like plotting: Consider participating in the [tidytuesday](#)
- [FU statistical consulting](#) for questions regarding statistical methods
- [R Consulting by me](#)
- [Tools and Tips lecture](#)



Artwork by [Allison Horst](#)

The End