Organization: Schedule

Schedule of today

- Now - 14 (or 14.30 if you are enthusiastic still): Work on the data set
  - 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 by "experts"

```
## 'data.frame':    1599 obs. of  12 variables:
##  $ fixed.acidity       : num  12.7 9.8 6.5 8.6 7.5 7.6 10.1 6.4 6.1 6.7 ...
##  $ volatile.acidity    : num  0.6 0.66 0.88 0.52 0.58 0.5 0.935 0.4 0.58 0.46 ...
##  $ citric.acid         : num  0.49 0.39 0.03 0.38 0.14 0.29 0.22 NA 0.23 0.24 ...
##  $ residual.sugar      : num  2.8 3.2 NA 1.5 2.2 2.3 3.4 1.6 2.5 1.7 ...
##  $ chlorides           : num  0.075 0.083 0.079 0.096 0.077 NA 0.105 0.066 0.044 0.077 ...
##  $ free.sulfur.dioxide : num  5 21 23 5 27 5 11 5 16 18 ...
##  $ total.sulfur.dioxide: num  NA 59 47 18 60 NA 86 12 70 34 ...
##  $ density             : num  0.999 0.999 0.996 NA 0.996 ...
##  $ pH                  : num  3.14 3.37 NA 3.2 3.28 3.32 3.43 3.34 3.46 3.39 ...
##  $ sulphates           : num  0.57 0.71 0.5 0.52 0.59 NA 0.64 NA NA 0.6 ...
##  $ alcohol             : num  11.4 11.5 11.2 9.4 9.8 11.5 11.3 9.2 12.5 10.6 ...
##  $ quality             : int  5 7 4 5 5 6 4 5 6 6 ...
```

Reference: P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Data set 1: What makes a good wine?

Ideas:

- Which chemical properties are associated with a good wine rating?
- Plot chemical properties against each other
- Plot quality judgement depending on different properties
- Linear models for the relationship between properties
- Summary tables of the data using `dplyr`

Hint: It might be a good idea to transform the quality column to a factor (use `dplyr::mutate` and `as.factor()` if you want to do this).

Data set 2: Paralympic games from 1980-2016

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
grp_id	integer	Group ID as a count within team sports
type	character	Type of sport
year	double	year of games
guide	character	Guide (for vision impaired athletes)
pilot	character	Pilot (for vision impaired athletes)

Source: International Paralympic Committee (provided by tidytuesday)

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
08-03/athletes.csv')
```

Ideas

- Probably dplyr heavy task
- Create summaries of medal counts for different groups
- Explores questions such as:
  - 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?
- Plots: e.g. heat map showing the number of medals by medal type and sport/country/gender...

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 of the crab

Ideas for data

- Rather small and good to handle
- Explore Bergmann's rule (organisms are large in higher latitudes)
- Linear model to explore relationship between latitude and body size
- Anova or tests to compare size between locations
- Plot relationship between latitude and size
- ...

Source: Johnson, D. 2019. Fiddler crab body size in salt marshes from Florida to Massachusetts, USA at PIE and VCR LTER and NOAA NERR sites during summer 2016. ver 1. Environmental Data Initiative.

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 LTER and NOAA NERR sites during summer 2016. ver 1. Environmental Data Initiative.

Data set 4: Ice cover and temperature

Temperature and ice duration on lakes since 19th century

- 2 data sets that can be combined
- measurements of
  - ice start, end and duration on 2 lakes in Wisconsin
  - daily air temperature since 1870

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.

Source temperature data: Anderson, L. and D. Robertson. 2020. Madison Wisconsin Daily Meteorological Data 1869 - current ver 32. Environmental Data Initiative.

Data set 4: Ice cover and temperature

Ideas

- How did ice cover duration change over the years?
  - Plotting of the time series
  - Linear model to test effect
- How did air temperature change over the years?
  - Summarizing mean annual temperatur or mean temperature in winter
  - Linear model to test the effect
- Combine the two data sets together
  - How do ice duration on the lakes correlate with temperure (e.g. with mean winter temperature)

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
```

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.
- Google
- If you get stuck, ask in the chat or stop by in General
- Have a look at the cheat sheets (list of all cheat sheets)
  - ggplot2
  - dplyr
  - Data import cheat sheet: readr and tidyr cheat sheet

Read data from Excel (xlsx format)

Use the `read_xlsx()` function from the `readxl` package.

Read a single excel file:

```
# install.packages("readxl")
library(readxl)

# by default: reads the first sheet
my_data <- read_xlsx(path = "data/my_datafile1.xlsx")

# To read a specific sheet:
# refer to the sheet by its name
my_data <- read_xlsx(path = "data/my_datafile1.xlsx", sheet = "sheet1")
# or refer to it by its id
my_data <- read_xlsx(path = "data/my_datafile1.xlsx", sheet = 2)
```

- `read_xlsx` has similar options as `read_csv`, e.g. to skip rows etc.
- have a look at `?read_xlsx` to see all options

Read data from Excel (xlsx format)

If you data is **split into multiple excel sheets** (i.e. one excel file but multiple sheets) you can:

- read in all the excel sheets at once into a list
- use `bind_rows` to bind all the sheets together by row
- This requires all sheets to have the same format and columns

```
# create a variable with the path to the file you want to read
path_to_file <- "data/my_datafile1.xlsx"

# list all the sheets that exist in this file
sheets <- excel_sheets(path_to_file)

# use lapply to read all sheets into a list using read_xlsx
all_sheets <- lapply(sheets, function(x) read_xlsx(path_to_file, sheet = x))

# bind all rows from the list together
my_data <- do.call(bind_rows, all_sheets) # without id column
my_data <- do.call(bind_rows, list(all_sheets, .id = "sheet_id")) # with id column
```

Read data from Excel (xlsx format)

If you data is **split into multiple excel files** (i.e. multiple excel files but one sheet per file) you can:

- read in all the excel files at once into a list
- use `bind_rows` to bind all the files together by row
- This requires all files to have the same format and columns
- My code requires the files to all be in the same folder

```
# List all xlsx files at a given location
# this code will read in all xlsx files at the given location
# if you don't want this, you have to adjust the `pattern`
file_paths <- list.files(path = "./data", pattern = ".xlsx", full.names = TRUE)
# check if this actually list your desired files
file_paths

# use lapply to read all files into a list using read_xlsx
all_files <- lapply(file_paths, read_xlsx)

# bind all rows from the list together
my_data <- do.call(bind_rows, all_files) # without id column
my_data <- do.call(bind_rows, list(all_files, .id = "file_id")) # with id column
```

Now You: Working on your own research data

- Meet in your group (if you want)

- Start working on a data set

- Take breaks as you need and be back at 14:00

- Keep an eye on your group and the general chat

Now You: Working on a data set from me

- Think about the data set you would like to work on

- Build groups

- Meet in your group

- Start working on a data set

- Take breaks as you need and be back at 14:00

- Keep an eye on your group and the general chat

Presentation/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?
- If you have a plot or a result to share, please post a screenshot in the chat or share your screen.

Feedback

Please take 5 mins to complete the feedback survey for the Graduate center (don't use Internet Explorer)

https://votingo.cedis.fu-berlin.de/survey/PCNLP3

Please take another 5 mins to complete my more specific survey to further improve the course

https://forms.gle/xfwKLDDomwoY44i66

Feedback

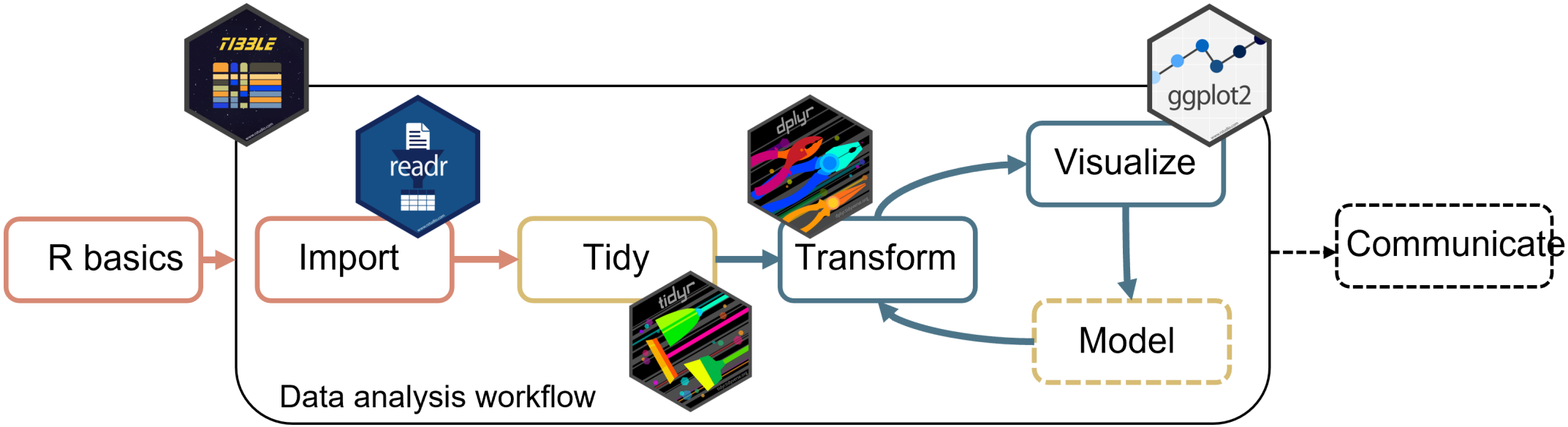
- Any other feedback or comments from your side?

Workshop: Reproducible documents with Rmarkdown

Tuesday 13.09. - Wednesday 14.09. (+ Thursday 15.09. optional)
1 p.m. - 5 p.m.
Write an email to me or to Simone if you want to join

Artwork by Allison Horst

Conclusion

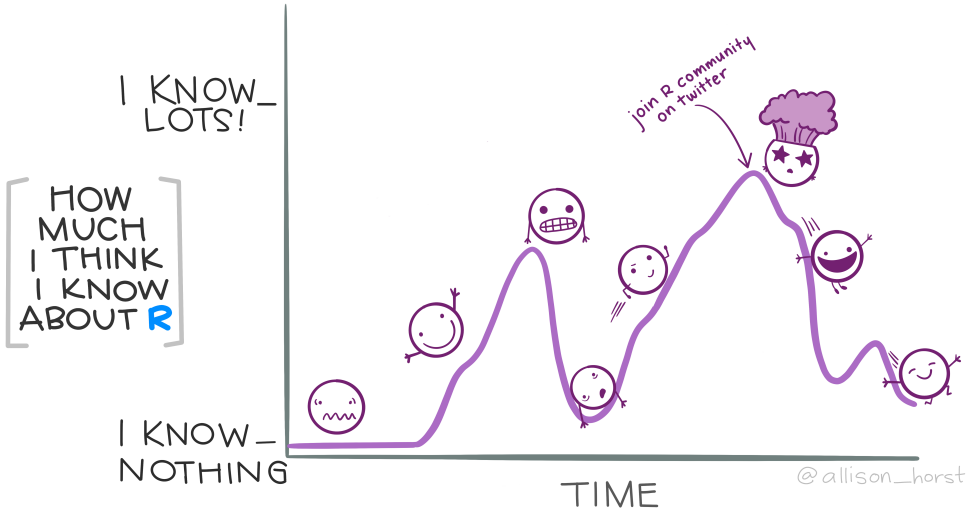


We learned a lot of stuff!

Conclusion

How to continue from here?

- Learning by doing!
- Have a look at some online ressoures, I recommend the R for Data Science book by Hadley Wickham
- If you use Twitter: Follow some people that post R content regarding your interest
- If you like plotting: Consider participating in the tidytuesday
- FU statistical consulting for questions regarding statistical methods



The End

Thanks a lot for participating!

Artwork by Allison Horst