Organization: Schedule

Schedule of today

- 9-13 (or 14 if you are enthusiastic still): Work on the data set
  - take break(s) as best fits your needs and your group
- 13 (14)-15: Short 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?
  - ...
- 15-16: Feedback, conclusion

Organization: Groups

- Group 1 (own data):
- other groups depending on data set preferences (~ 4 people by group)

In the groups

- have a look at the data set first
- discuss potential questions, methods, ...
- decide how you want to work together
    - divide tasks
    - all do same task
    - everybody on their own but available for questions
- If you can, answer questions from the other group members
- You can also work on multiple data sets if wish

Organization: Groups

What I basically want to say: Just have fun with the data, try new things and organize yourselves!

Data set suggestions

The `EcoData` R package includes many different data sets.

Get the package:

```
#install.packages("devtools") # run this only if you don't have devtools installed
devtools::install_github(repo = "florianhartig/EcoData", subdir = "EcoData",
dependencies = T, build_vignettes = T)
```

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?

Get the data:

```
library(tidyverse)
wine <- as_tibble(EcoData::wine)
```

Ideas:

- plotting chemical properties against each other
- plotting quality judgement depending on different properties
- linear models for the relationship between properties
- summary tables of the data

Data set 2: Birds in Australia

Effect of forest fragmentation on bird abundance

- bird abundance in the fragment
- area (size of the fragment)
- distance to the next forest fragment
- distance to the next large forest fragment
- years since site was isolated
- grazing intensity of the surrounding
- altitude

```
## # A tibble: 56 x 8
##    Site ABUND  AREA  DIST LDIST YR.ISOL GRAZE   ALT
##   <int> <dbl> <dbl> <int> <int>   <int> <int> <int>
## 1     1   5.3   0.1    39    39    1968     2   160
## 2     2   2     0.5   234   234    1920     5    60
## 3     3   1.5   0.5   104   311    1900     5   140
## # ... with 53 more rows
```

Image by Marcel Holyoak

Data set 2: Birds in Australia

Get the data

```
library(tidyverse)
birds <- as_tibble(EcoData::birdabundance)
```

Ideas

- good for plotting
- rather clear and small data set
- explore the effects of environment on bird abundance (try models)

Image by Marcel Holyoak

Data set 3: 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 3: 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 men or women win more medals overall? 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...

Some general tips

For everyone

- google
- Have a look at the cheat sheets (list of all cheat sheets)
  - ggplot2
  - dplyr
  - lubridate: working with dates and times (convert character columns to date time, calculate differences in time, ...)
  - stringr: working with strings (e.g. creating substrings, extracting parts from strings, ...)

Some general tips

For those working with their own data

- data import cheat sheet: readr and tidyr cheat sheet
- janitor package): some simple functions for examining and cleaning dirty data
  - `clean_names()`: clean your column names with one function (e.g. in case you have white space)
  - `excel_numeric_to_date()`: convert strange excel serial numbers to actual dates
- readxl package: read directly from Excel files instead of using `readr` with csv files
  - check out `?read_xlsx` to see how to read different sheets

Now You

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

- Build groups using the chat

- Meet in your group

- Start working on a data set

- Take breaks as you need and be back at 13:30

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?

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/PCNLP3

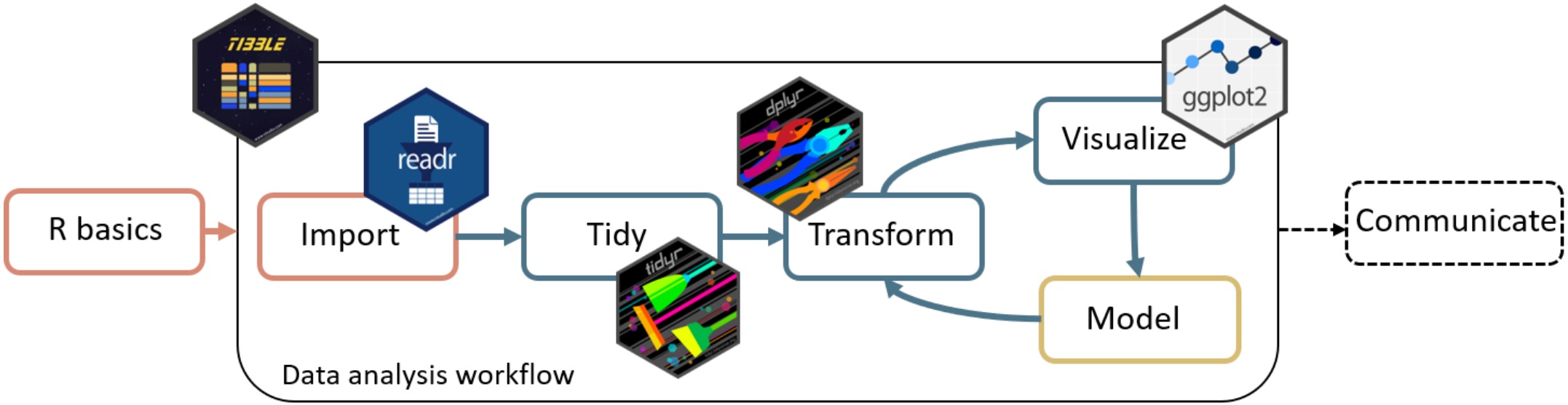
Please take another 5 mins to complete my more specific survey to further improve the course

https://forms.gle/xfwKLDDomwoY44i66

Feedback

- Any other feedback or comments from your side?

Conclusion

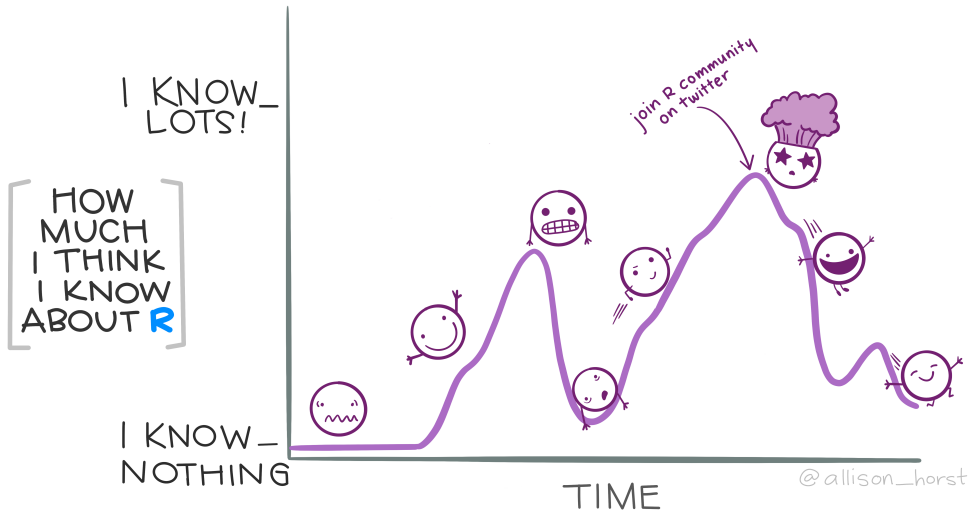


We learned a lot of stuff!

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



Artwork by Allison Horst

The End

Thanks a lot for participating!

Artwork by Allison Horst