# Essential Research Toolkit for the Humanities

## Week 3: Looking at data

Anna Pryslopska

April 22, 2024

Psycholinguistics and Cognitive Modeling Lab

# Homework

# What I got

25/30 homeworks → do homework, learn things, get 3 points for free

**Expected:** `screenshot.{jpg, png, doc, pdf}` and `script.R`
**Got:** 1–4 files: `screenshot.{jpg, png, doc, pdf}` and `script.{R, doc, sec, png, pdf}`

Install & load the packages: `tidyverse, knitr, MASS, psych`

Print a long text & saves it to a variable. → saved, didn't print

Run your code:

```
a <- print("Why does this work?")  ✔
1 <- print("But this not?!")      ✖
```

Error message! Warning! Conflict! Solved your problems ✔

# How to fail this class

# Why this “pointless” exercise

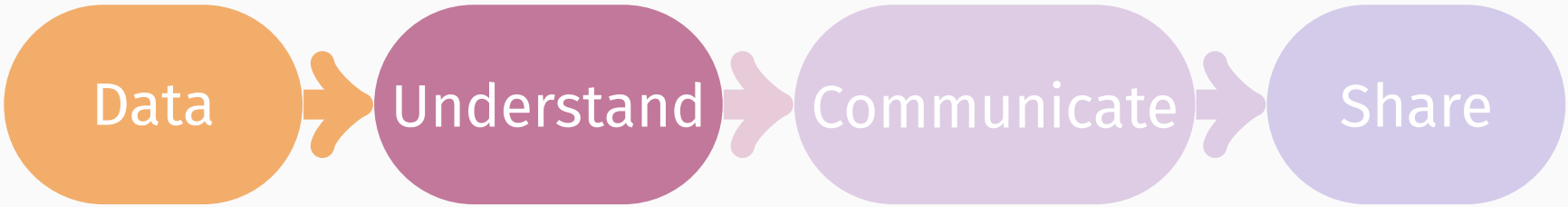
- ✔ Succesfully install R and RStudio
- ✔ Open and look at RStudio
- ✔ Name variables meaningfully
- ✔ Install and load packages
- ✔ Solved your problems
- ✔ Best honest effort
- ✖ Follow instructions
- ✖ Do what all programmers do with new IDE

Questions?

# Table of contents

# Where are we this week?

# Workflow




import, clean, transform, visualize, model

document, create clean and beautiful reports

connect, collaborate, backup

# Data types, formats, and encoding

# Data: What the statistitian sees

| | |
|---|---|
| nominal | marital status, religion |
| ordinal | grades, energy efficiency classes |
| interval | IQ, temperature in C and F |
| ratio | reaction times, population |

# Data: What the researcher sees

- **reading times**
- **acceptability judgments**
- **free response**
- completion
- recordings
- transcription, annotation
- brain waves
- texts
- video
- images
- ...

→ text files (`txt, csv, tsv`)

# Data: What R sees

| | |
|---|---|
| logical | `TRUE` |
| integer | `1` or `1L` |
| double | `1.0` |
| complex | `1+0i` |
| character | `"one"` |
| double "not a number" | `NaN` |
| double "infinity" | `Inf` |
| logical "missing" value | `NA` |
| special variable without a type | `NULL` |

...

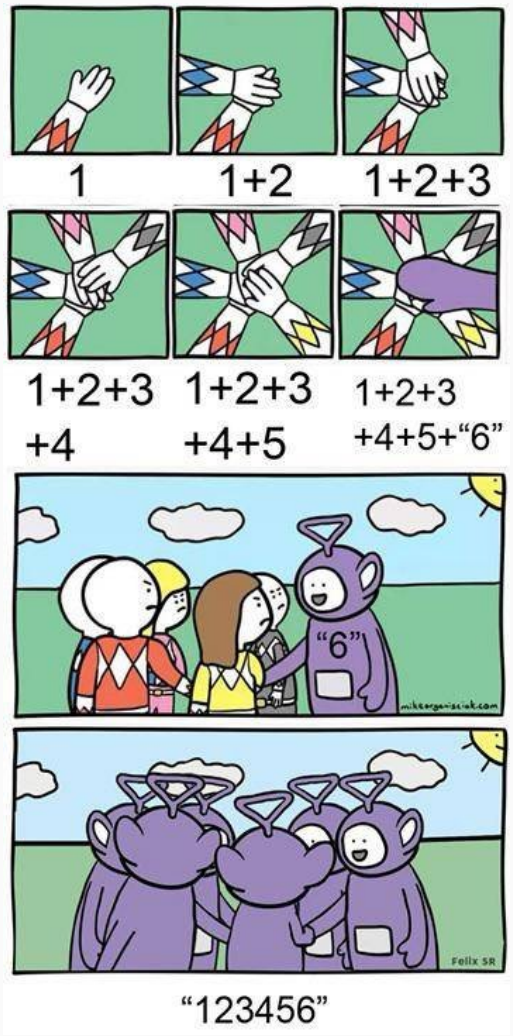
# Reap what you sow

Unsure? `typeof(1L)` or `is.numeric(1)`

Certain? `as.character(1)`

| | |
|---|---|
| `5L + 2 = 7` | integer |
| `3.7 * 3L = 11.1` | double |
| `99999.0e-1 - 3.3e+3 = 9999.9 - 3300 = 6699.9` | double |
| `10 / as.complex(2) = 5+0i` | complex |
| `as.character(5) / 5` = ⊘ | non-numeric argument! |
| `NaN == NaN` | NA |

# Lost in translation: How to mess up data




Make different types of entries in the same column.

```
慆獬档獩傒拔涋瘠湯堠泅灯澽浮獵步焠嵼沤⁴
数斀牧翳映牥涋娠敷杲ㄐ楠　睖⁸楦菶獩搠慦物
慬映　整傒ᶠ〈敨牵畤挠湩ⁱᵠ猸灥　窠ᵚ菡嫁ᵚ
菡β菡攜ᵬ♠銃菡嫁ᵚ菡攝邃♠ᵬ菡杸閃菡ₐ菡Ψ
鲃菡嫁隃菡ₔ菡攜邃菡ᶇ邃菡墜邃菡δ鯇菡ϒ靰
```

```
Falsches Äœben von Xylophonmusik
quÃ¤lt jeden grÃ¶ÃŸeren Zwerg.
Dis aux filles de faire la fÃªte
Ã<0xa0> l'heure du cinq Ã<0xa0>
sept.
áƒ¥áƒ<0x90>áƒ<0xa0>áƒ–áƒ£áƒšáƒ¡
áƒ'áƒ<0x90>áƒ<0xa0>áƒ“áƒ<0x90>
áƒ¡áƒ®áƒ•áƒ<0x90>
áƒ”áƒœáƒ<0x90>áƒ–áƒ” áƒšáƒ<0x90>áƒ
žáƒ<0x90>áƒ<0xa0>áƒ<0x90>áƒ™áƒ
<0x9d>áƒ‘áƒ–?
```

```
Falsches √úben von Xylophonmusik
qu√§lt jeden gr√∂√üeren Zwerg.
Dis aux filles de faire la f√™te
√† l'heure du cinq √† sept.
ŠÉ•ŠÉźŠÉ†ŠÉóŠÉ£ŠÉöŠÉ°
ŠÉíŠÉźŠÉ†ŠÉďŠÉź ŠÉ°ŠÉģŠÉēŠÉź
ŠÉĒŠÉúŠÉźŠÉĒŠÉĒ
ŠÉöŠÉźŠÉĕŠÉźŠÉ†ŠÉźŠÉôŠÉĔŠÉĎŠÉó?
```

```
Falsches Üben von Xylophonmusik
quält jeden größeren Zwerg.
Dis aux filles de faire la fête à
l'heure du cinq à sept.
ქართულს გარდა სხვა ენაზე
ღაპარაკობთ?
```

Make or change the encoding to something random.

# Lost in translation: How NOT to mess up data

→ Be careful and know what you put in

→ Be mindful of the character encoding (when in doubt, UTF-8)

# Data

DATA

# Experiment 1: Moses illusion

Q: Can a man marry his widow's sister? No 🐄💀🪦

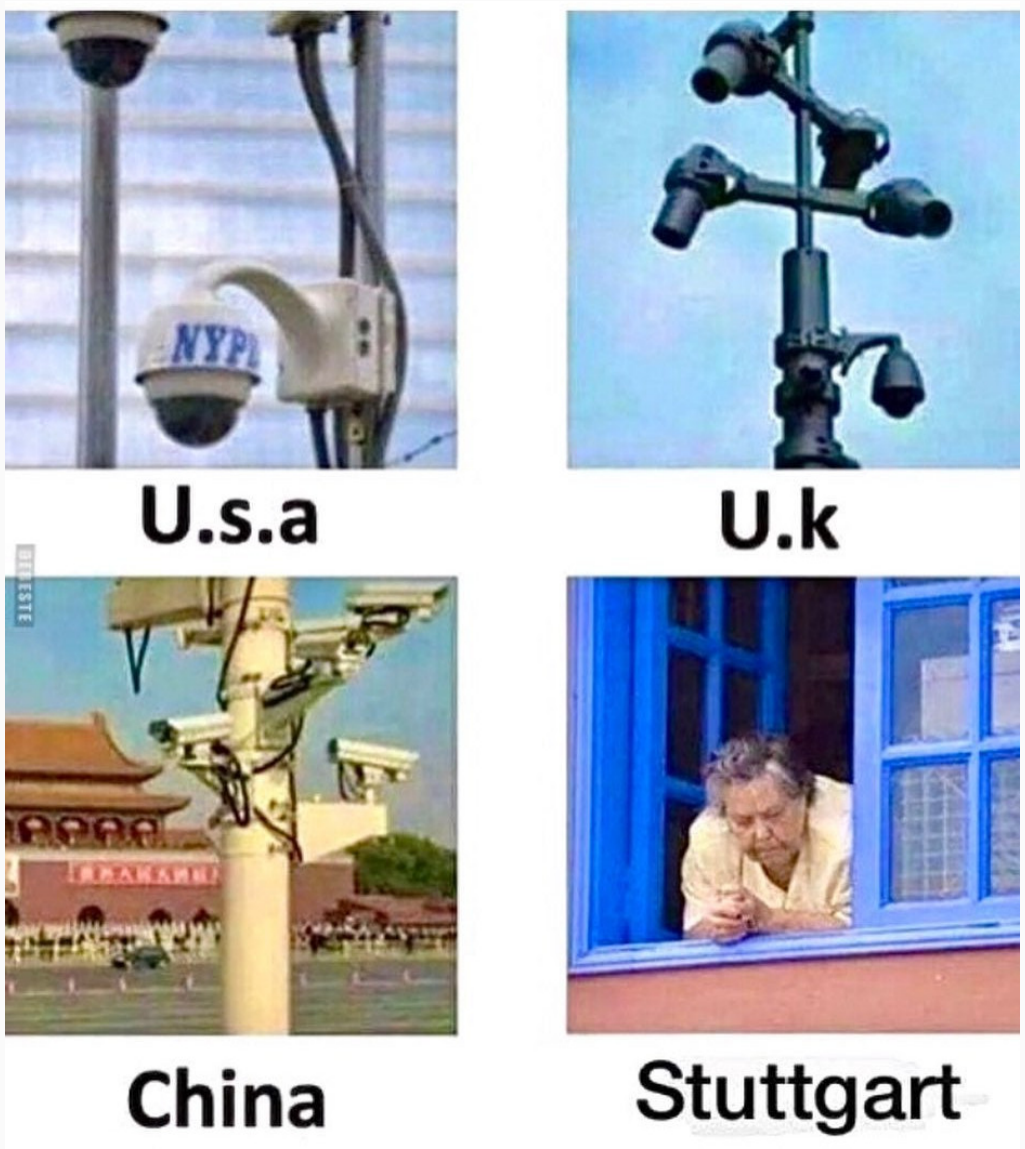

Data that we got:

- Answer to the Moses illusion and the control questions
- Answer to the distractor questions
- Answer time

...

# Inspecting data

# Look at what you did



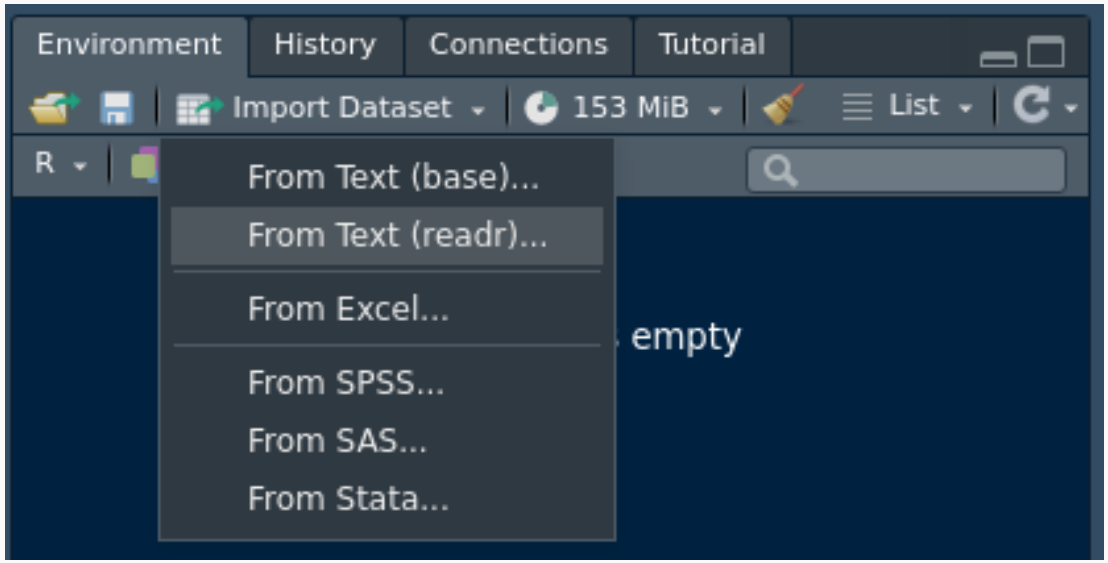
# Reading in data

Download `moses.csv` and `noisy.csv` from ILIAS and save it to your working directory

Make a (new) script and load the packages `tidyverse` and `psych`

Read in the file `moses.csv` and assign it to the name `moses`

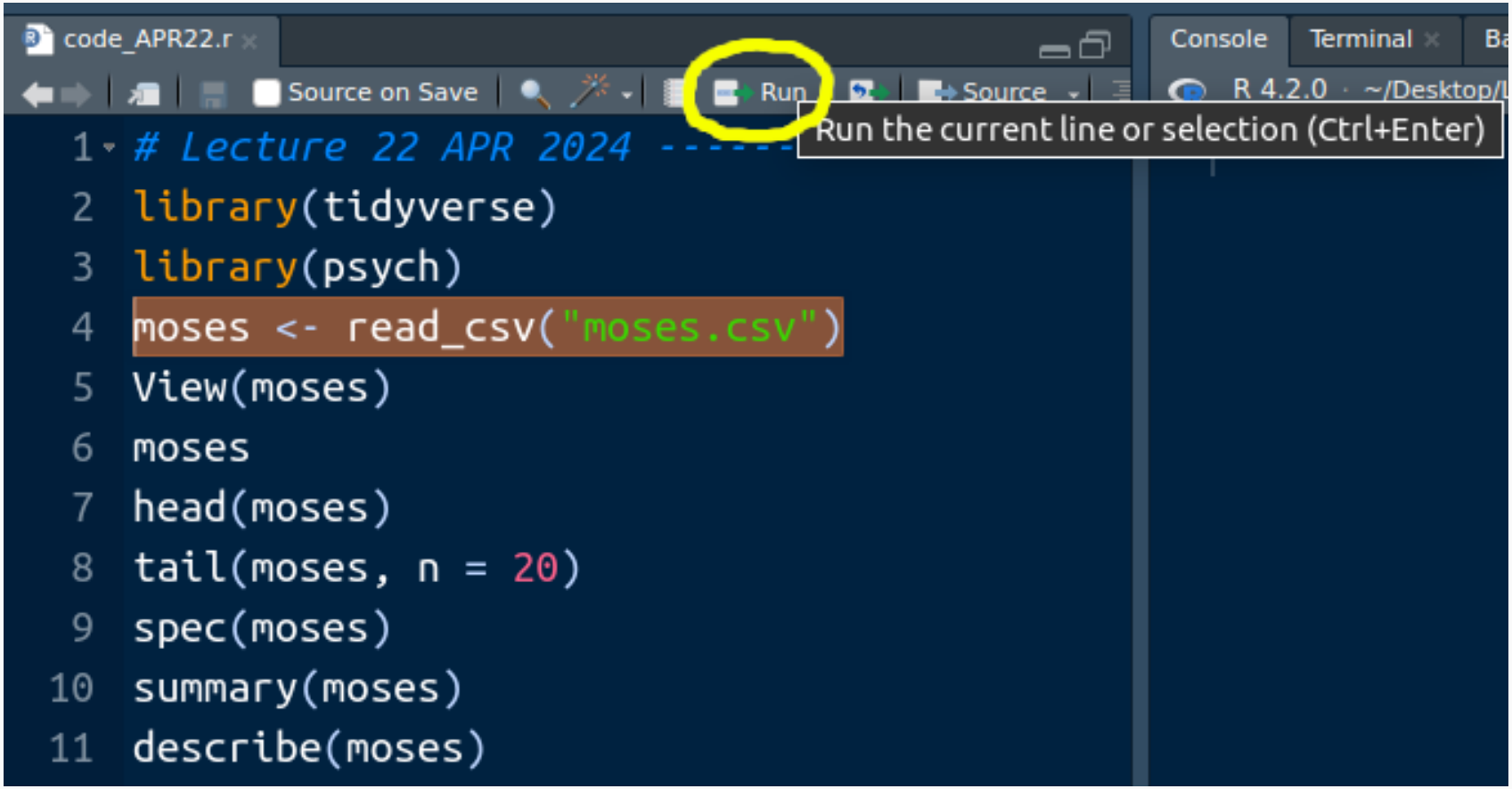
Select `Environment > Import Dataset > readr`



# Load the data into environment

```
moses <- read_csv("moses.csv")
```

CTRL+ENTER or CMD+RETURN or click on “Run”




You now have a data frame or tibble called `moses`.

# Look at what you did

| | |
|---|---|
| `View(moses)` | in the RStudio window |
| `moses` | in the console |
| `print(moses, n=Inf)` | in the console |
| `head(moses)` | first 6 rows |
| `tail(moses, n=20)` | last 20 rows |
| `spec(moses)` | column properties |
| `summary(moses)` | summary statistics |
| `describe(moses)` | summary statistics vol. 2 |
| `colnames(moses)` | column names |
| `summary(NAME)` | → calling function with one argument |
| `head(NAME, n=20)` | → calling function with two arguments |
| `dbinom(x=6, size=9, prob=0.5)` | 3 named arguments in order |

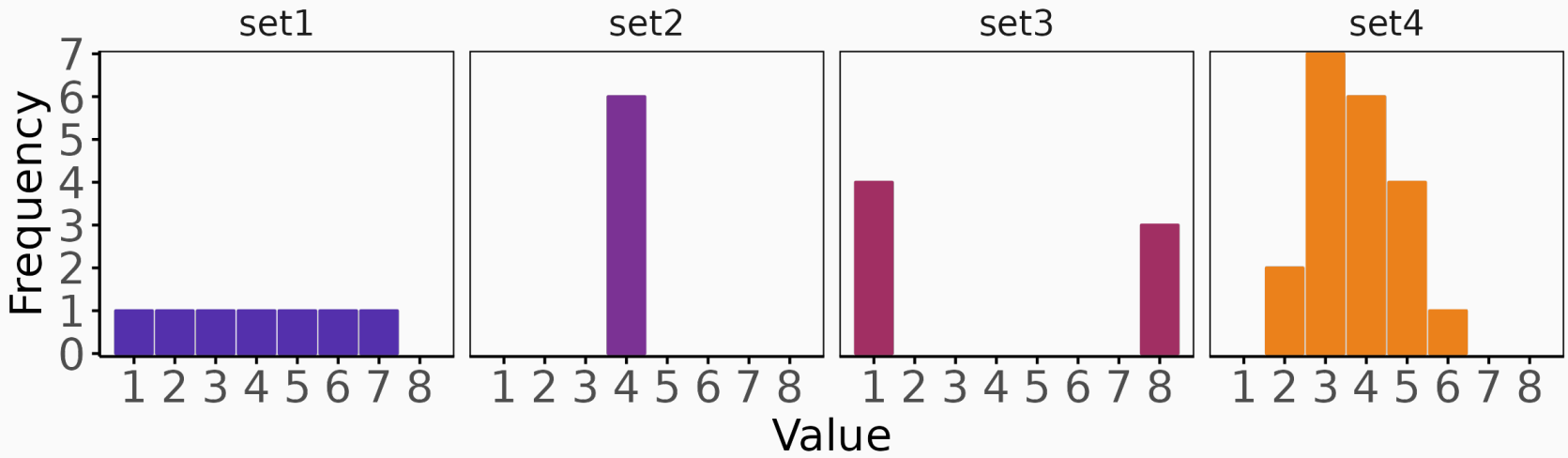
# Summarize

| | | |
|---|---|---|
| `Min.` | `min()` | minimal value |
| `Max.` | `max()` | maximal value |
| `Mean` | `mean()` | average |
| `1st Qu.` | `quantile()` | 25% |
| `Median` | `quantile()` | middle number == 2nd quantile == 50% |
| `3rd Qu.` | `quantile()` | 75% |
| `NA's` | TBA | nr missing data |

# Describe

| | | |
|---|---|---|
| | `colnames()` | item name |
| `vars` | `colnames()` | item number |
| `n` | TBA | number of valid cases |
| `mean` | `mean()` | mean |
| `median` | `median()` | median |
| `min` | `min()` | minimum |
| `max` | `max()` | maximum |
| `range` | `range()` | range |
| `sd` | `sd()` | standard deviation ($\sqrt{variance}$) |
| `trimmed` | | trimmed mean |
| `mad` | | median absolute deviation |
| `skew` | `skew()` | skew |
| `kurtosis` | `kurtosis()` | kurtosis |
| `se` | `mean_se()` | standard error |

# Central tendency




| Set | Values | Mean | Median | SD |
|---|---|---|---|---|
| 1 | 1,2,3,4,5,6,7 | 4 | 4 | 2 |
| 2 | 4,4,4,4,4,4 | 4 | 4 | 0 |
| 3 | 1,1,1,1,8,8,8 | 4 | 1 | 4 |
| 4 | 2,2,3,3,3,3,3,3,3,4,4,4,4,4,4,5,5,5,5,6 | 4 | 4 | 1 |

# What a mess

| | |
|---|---|
| Too much information | native, Instructions |
| Too little information | condition 1? 2?? 100??? |
| Missing information | NA |

Inconsistent information:

Q: **Margaret Thatcher was the former president/prime minister of which country?**

A: Don’t know, Great Britain, UK, United Kingdom, can’t answer, england, the UK, uk, Prime Minister of UK, can’t know, cant answer

Q: **According to the Bible, how many animals of each kind did Moses/Noah take on the ark?**

A: 2, Don’t know, Two, can’t answer, don’t know, don´t know, no idea, two, 2 of each kind, 2%2C 1 male 1 female, 42, 62,

# Clean up after yourselves

# Clean up after yourselves

| | |
|---|---|
| select meaningful columns | `select(WHERE, WHAT)` |
| remove missing values | `na.omit(WHERE)` |
| choose or remove data | `filter(WHERE, TRUE CONDITION)` |
| reorder values | `arrange(WHERE, HOW)` |
| rename columns | `rename(WHERE, NEW = OLD)` |
| create values | `mutate(WHERE, NEW = FUNCTION(OLD))` |

`= != ==` `=` is assignment, `==` is equality

Functions are executed, results are displayed, but nothing is saved.

# Wrap-up

# Summary

- ✔ scripts
- ✔ data types
- ✔ encoding
- ✔ reading in data
- ✔ inspecting data
- ⏩ Hands-on data cleanup, basic R operations, tidy code, data manipulation.

# Homework assignment due April 26th 15:30

- Read chapters 3, 4 and 5 of *R for Data Science* (Wickham et al. 2023)
- Complete assignment 2 ($\rightarrow$ ILIAS)

# References

 Wickham, Hadley, Mine Çetinkaya-Rundel, and Garrett Grolemund (2023). *R for data science: import, tidy, transform, visualize, and model data*. 2nd ed. O'Reilly Media, Inc. URL: https://r4ds.hadley.nz/.