

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

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} *rr script != doc document*

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

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

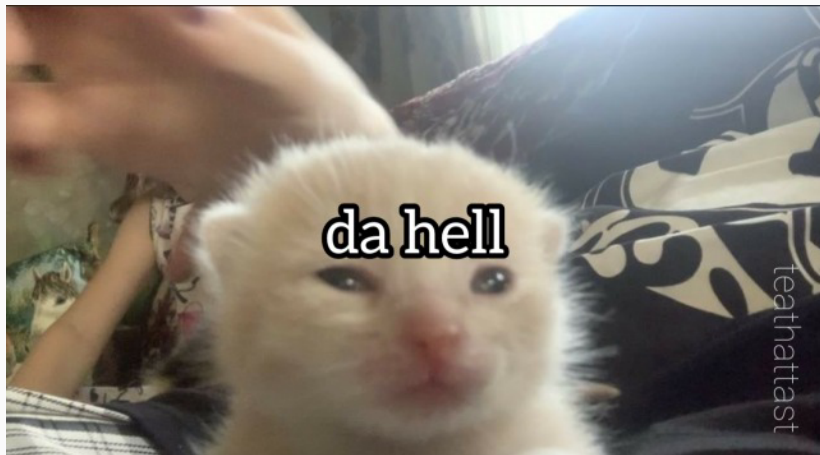
Run your code:

```
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

- ✓ Successfully 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

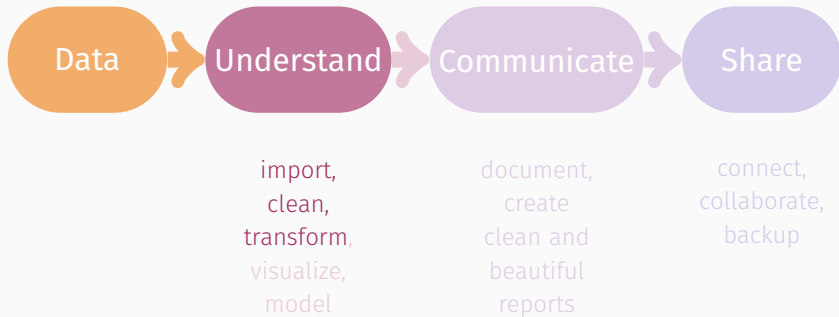
Questions?

Table of contents

1. Where are we this week?
2. Data types, formats, and encoding
3. Data
4. Inspecting data
5. Wrap-up

Where are we this week?

Workflow



Data types, formats, and encoding

Data: What the statistician 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

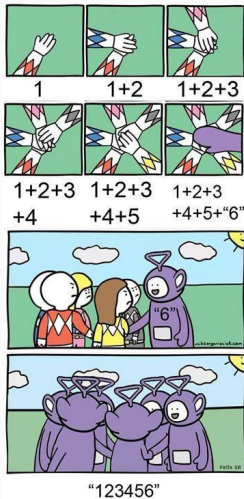
`99999.0e-1 - 3.3e+3 = 9999.9 - 3300 = 6699.9` double

`10 / as.complex(2) = 5+0i` complex

`as.character(5) / 5 = \emptyset` non-numeric argument!

`NaN == NaN` NA

Lost in translation: How to mess up data



Make different types of entries in the same column.

惺惺档徽徽拔淡瘠湯墩酒灯濛浮獵步燐峇派
 斂穀枚習啖妨淡娠敷果H編 咳。槍奏徽棚愁物
 僅映 整歇<散牵時挽漣·狹蠱 寐?箇媒
 箇β箇搥♠銃箇媒?箇搥邊♠箇枝閃箇。箇ψ
 鮑箇媒喻箇。箇搥邊箇箇箇整邊箇δ駭箇Υ駭

Falsches Aøben von Xylophonmusik
 quÄ=lt jeden grÄ¶ÄYeren Zwerg.
 Dis aux filles de faire la fÄete
 Ä<0xa0> l'heure du cinq Ä<0xa0>
 sept.
 äf'Yäf<0x90>äf<0xa0>äf-äffäfsäfi
 äf'äf<0x90>äf<0xa0>äf"äf<0x90>
 äfiäf@äf•äf<0x90>
 äf"äf@äf<0x90>äf-äf" äfsäf<0x90>äf
 zäf<0x90>äf<0xa0>äf<0x90>äf"äf
 <0x9d>äf'äf-?

Falsches üben von Xylophonmusik
 quv\$lt jeden grvðvüeren Zwerg.
 Dis aux filles de faire la fÿnte
 vÿt l'heure du cinq vÿt sept.
 SE•SEZSEtSEöSEfSEöSE"
 SEiSEZSEtSEdSEZ SE•SEöSESEZ
 SEöSEüSEZSEöSEE
 SEöSEZSEöSEZSEtSEZSEöSEöSEöSEö?

Falsches Üben von Xylophonmusik
 quält jeden größeren Zwerg.
 Dis aux filles de faire la fête à
 l'heure du cinq à sept.
 ღართუღს გართა სხვა ენაზე
 რაბარაკობ?

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



U.s.a



U.k



China



Stuttgart

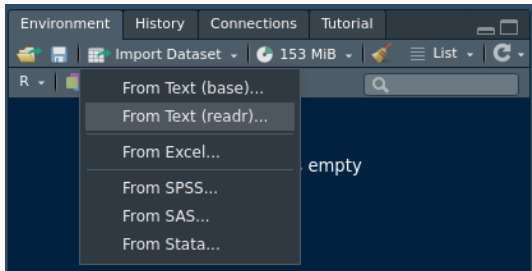
Reading in data

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`

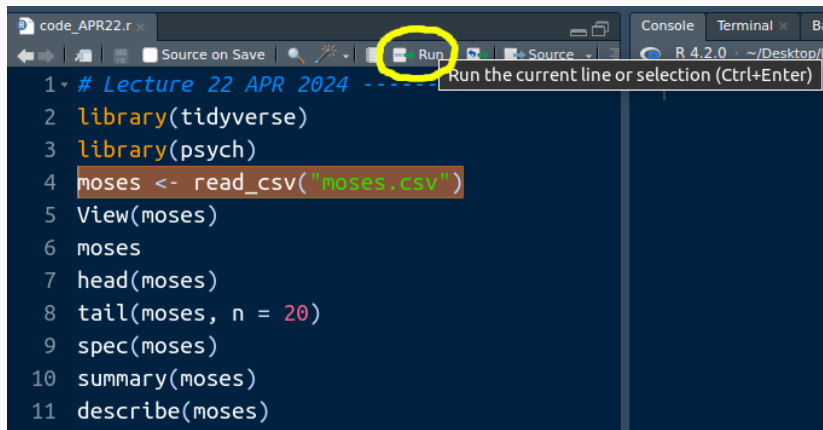
Select `Environment > Import Dataset > readr`



Load the data into environment

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

CTRL+ENTER or CMD+RETURN or click on "Run"



```
code_APR22.r x
Source on Save
Run
Run the current line or selection (Ctrl+Enter)
R 4.2.0 · ~/Desktop/

1 # Lecture 22 APR 2024 -----
2 library(tidyverse)
3 library(psych)
4 moses <- read_csv("moses.csv")
5 View(moses)
6 moses
7 head(moses)
8 tail(moses, n = 20)
9 spec(moses)
10 summary(moses)
11 describe(moses)
```

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

Look at what you did

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

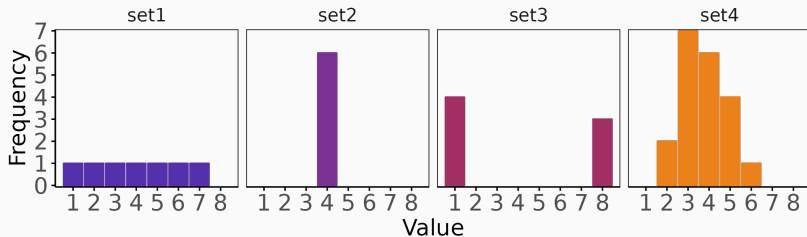
Summarize

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 ($\sqrt{\text{variance}}$)
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	<code>select(WHERE, WHAT)</code>
remove missing values	<code>na.omit(WHERE)</code>
choose or remove data	<code>filter(WHERE, TRUE CONDITION)</code>
reorder values	<code>arrange(WHERE, HOW)</code>
rename columns	<code>rename(WHERE, NEW = OLD)</code>
create values	<code>mutate(WHERE, NEW = FUNCTION(OLD))</code>

`=` `!=` `==`

`=` 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 (→ 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/>.