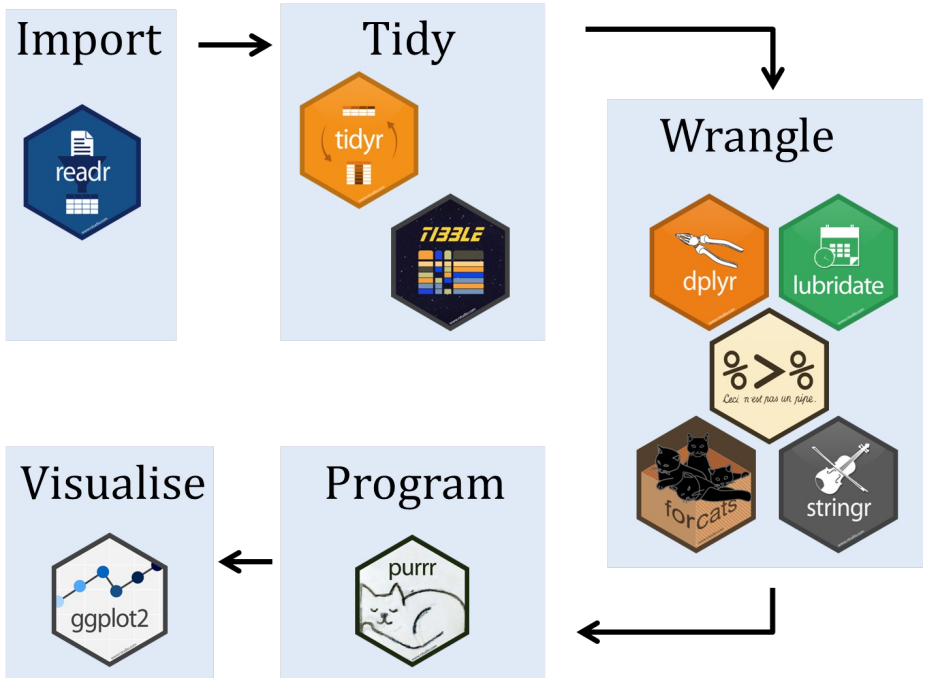
Starting with data

Peter Verhaar

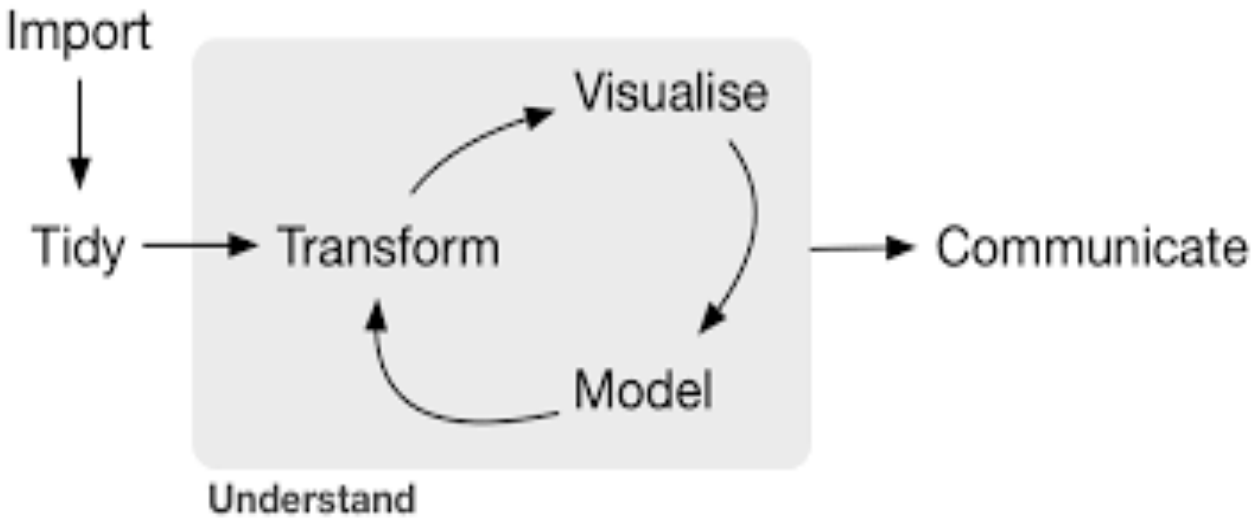
Universiteit
Leiden
The Netherlands



"an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures" (https://www.tidyverse.org/)

Data acquisition

Data visualisation

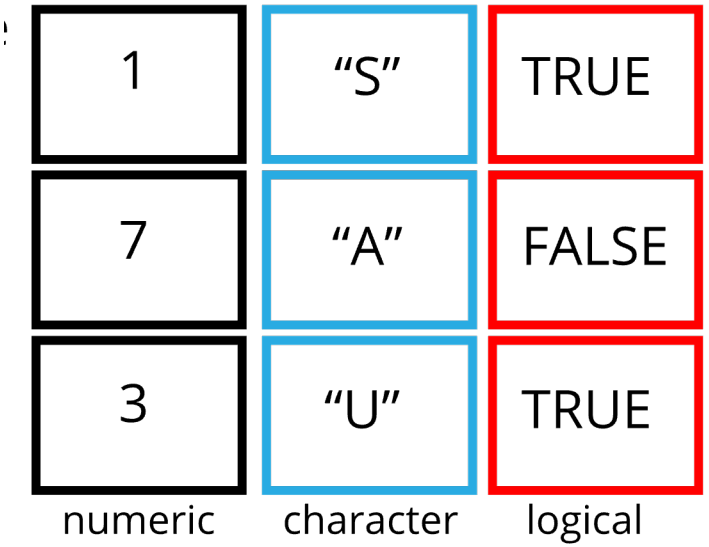


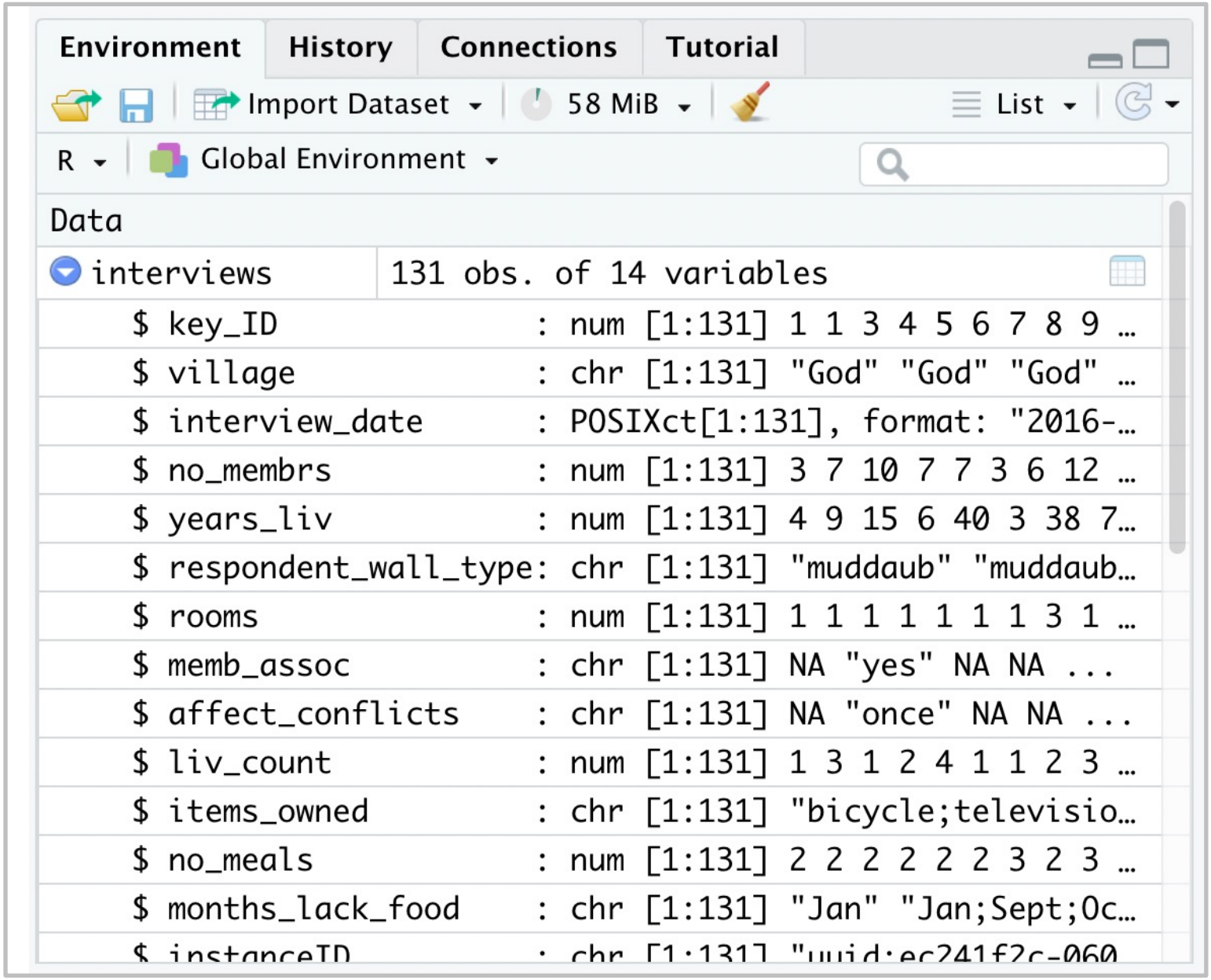
Data analysis

Hadley Wickham, *R for Data Science* (O'Reilly, 2016)

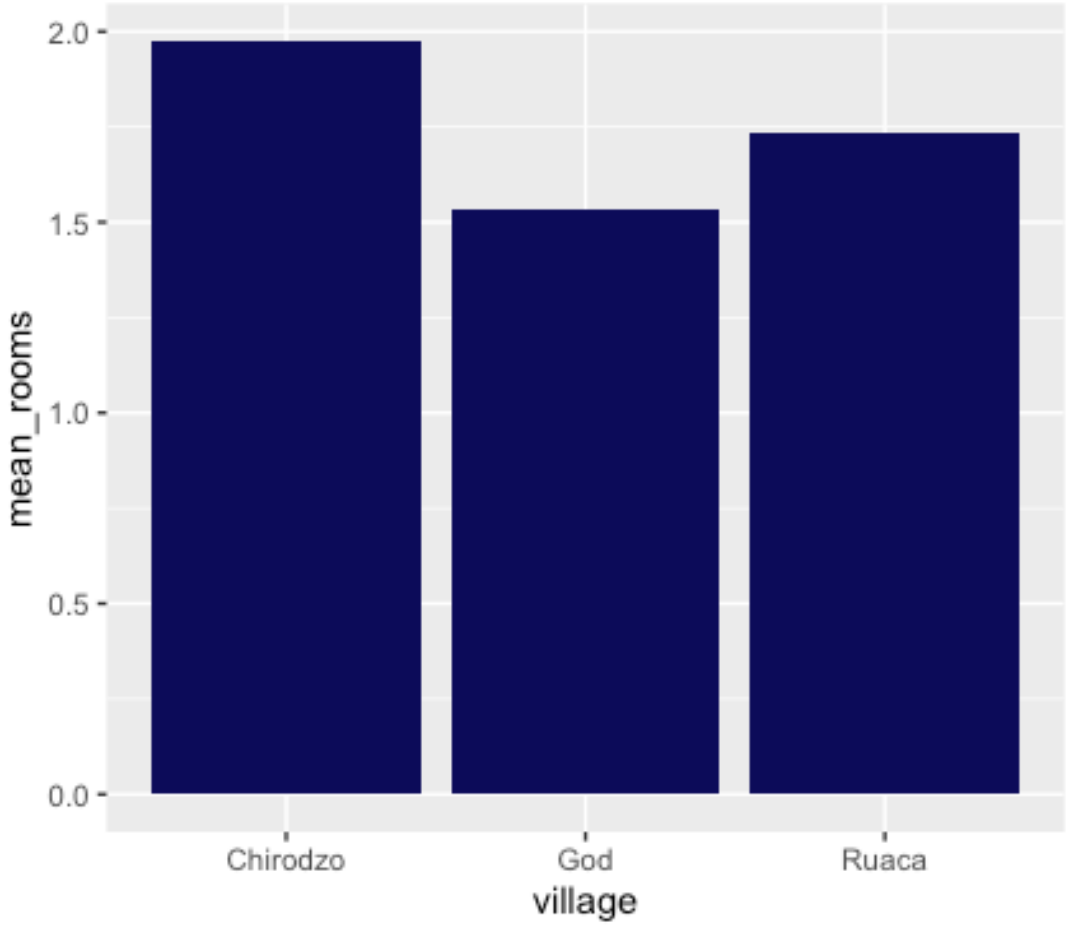
Data frame

- Data structure used to store tabular data
- Each column (variable) is also a vector
- Tidyverse uses the term ‘Tibble’

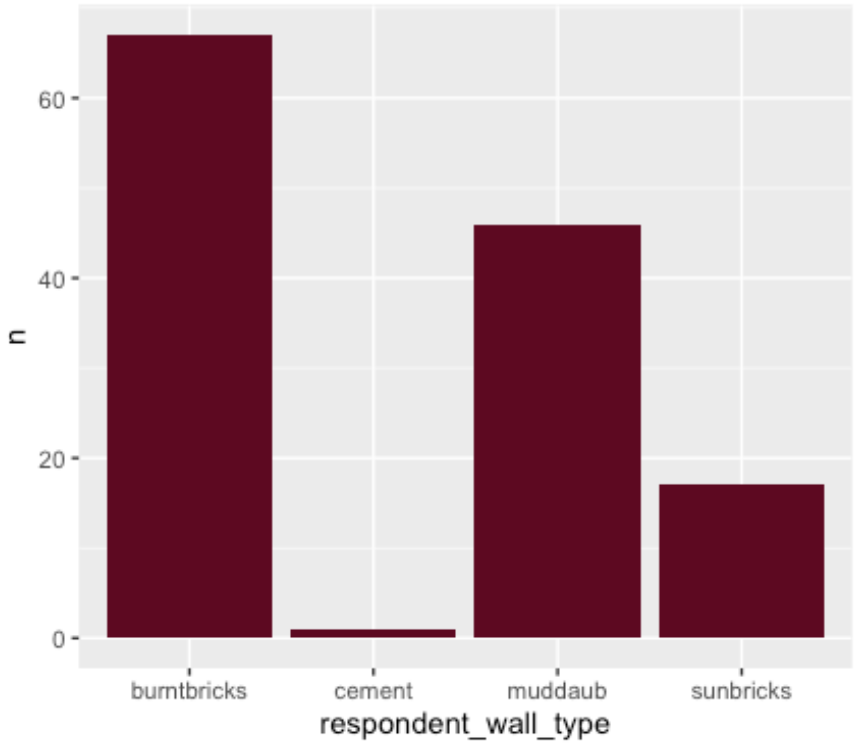



Environment
History
Connections
Tutorial
Import Dataset
58 MiB
List
R
Global Environment
Data
interviews
131 obs. of 14 variables
$ key_ID : num [1:131] 1 1 3 4 5 6 7 8 9 …
$ village : chr [1:131] "God" "God" "God" …
$ interview_date : POSIXct[1:131], format: "2016-…
$ no_membrs : num [1:131] 3 7 10 7 7 3 6 12 …
$ years_liv : num [1:131] 4 9 15 6 40 3 38 7…
$ respondent_wall_type: chr [1:131] "muddaub" "muddaub…
$ rooms : num [1:131] 1 1 1 1 1 1 1 3 1 …
$ memb_assoc : chr [1:131] NA "yes" NA NA ...
$ affect_conflicts : chr [1:131] NA "once" NA NA ...
$ liv_count : num [1:131] 1 3 1 2 4 1 1 2 3 …
$ items_owned : chr [1:131] "bicycle;televisio…
$ no_meals : num [1:131] 2 2 2 2 2 2 3 2 3 …
$ months_lack_food : chr [1:131] "Jan" "Jan;Sept;Oc…

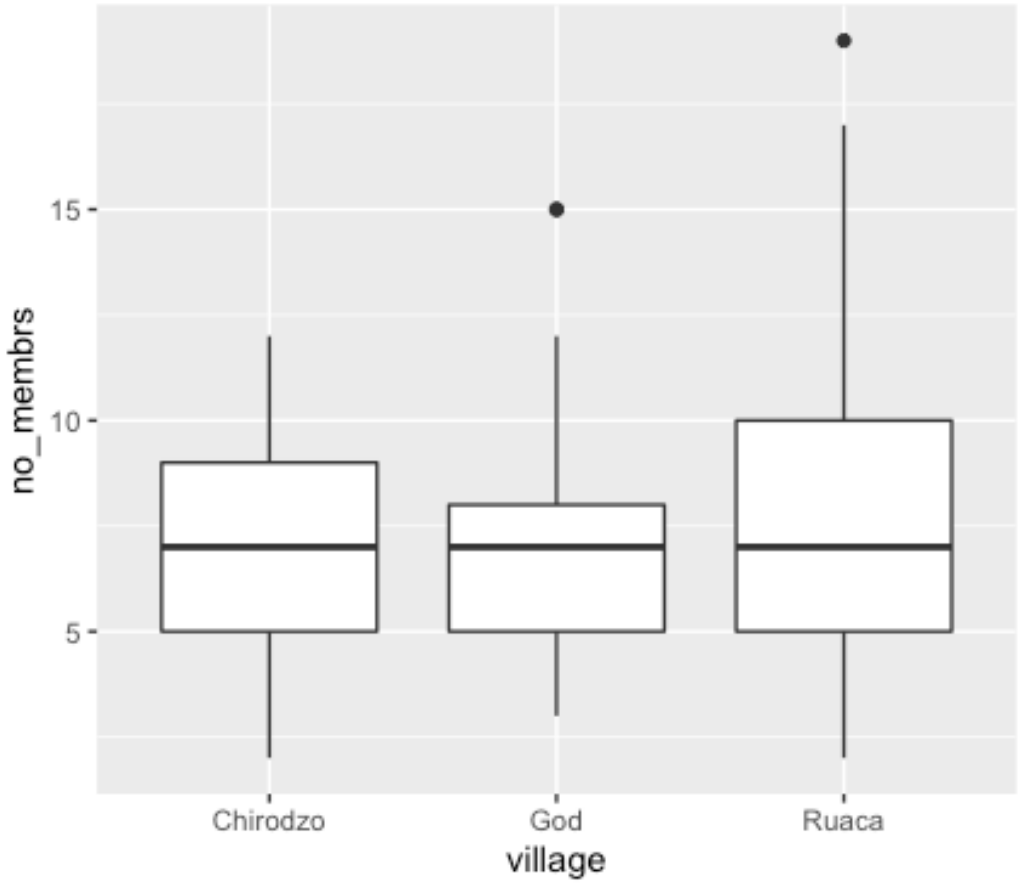
Average number of rooms in each village



Wall types


60
40
20
0
n
burntbricks
cement
muddaub
sunbricks
respondent_wall_type

Number of people living in houses



Exercise 1

- Create a new tibble (named ***interview_100***) containing only the data in row 100 of the interviews dataset.
- Create a new tibble (named ***interview_last***) containing only the data from the last row. Use nrow() to find the index of the last row. You can use the tail() function to check your results.

Exercise 2

- Extract the row that is in the middle of the dataset. Store the data on this middle row in a tibble named ***interview_middle***.
  To find the index of the middle row, you can work with the *median()* function, which takes a vector as a parameter. This vector needs to start at index 1, and needs to end at the index of the last row. Tip: use the nrow() function.
- Combine nrow() with the minus ('-') notation to reproduce the behavior of head(interviews), keeping the first 6 rows of the interviews dataset only. Assign the result to ***interview_head***

Exercise 3

How many respondents are members of an irrigation association?

Make a bar chart, with two values on the X-axis: “No” and “Yes”. The dataset contains the words ‘yes’ and ‘no’ in lower case, but these words need to be capitalised in the plot. Missing values (NA values) can be ignored.