

Introduction to Data Wrangling with `dplyr`

2021-07-07

Tidy Data

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	127291272
China	2000	213766	128042583

variables

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	127291272
China	2000	213766	128042583

observations

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	127291272
China	2000	213766	128042583

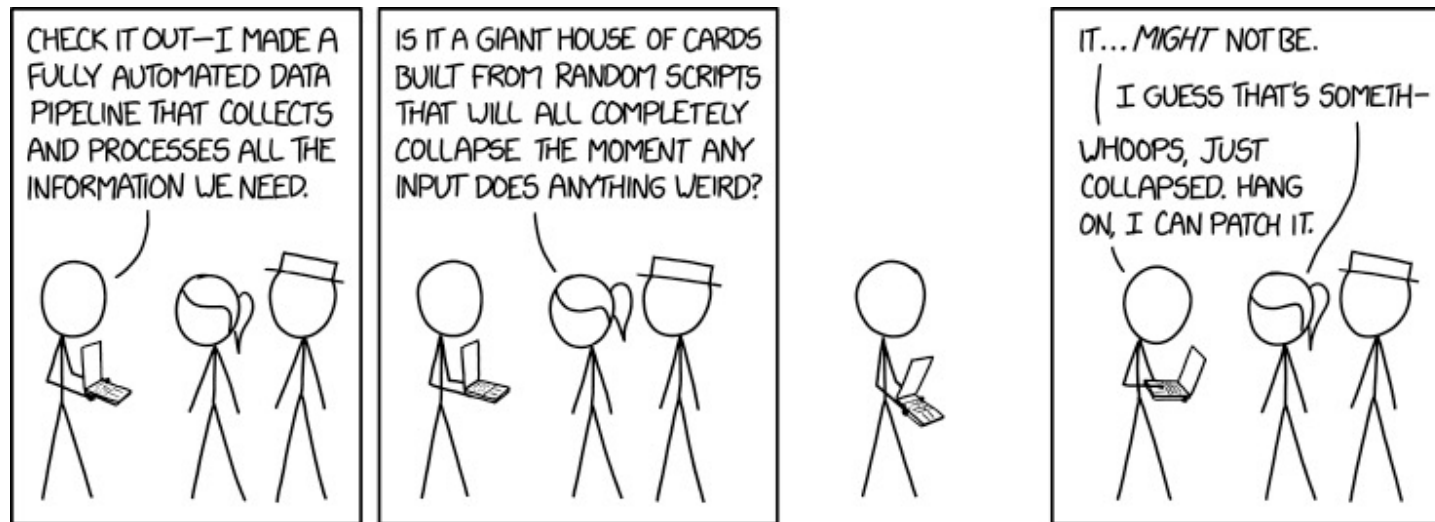
values

1. Each variable is in a column.
2. Each observation is a row.
3. Each value is a cell.

Data Wrangling

data wrangling – organizing your data into the form you want

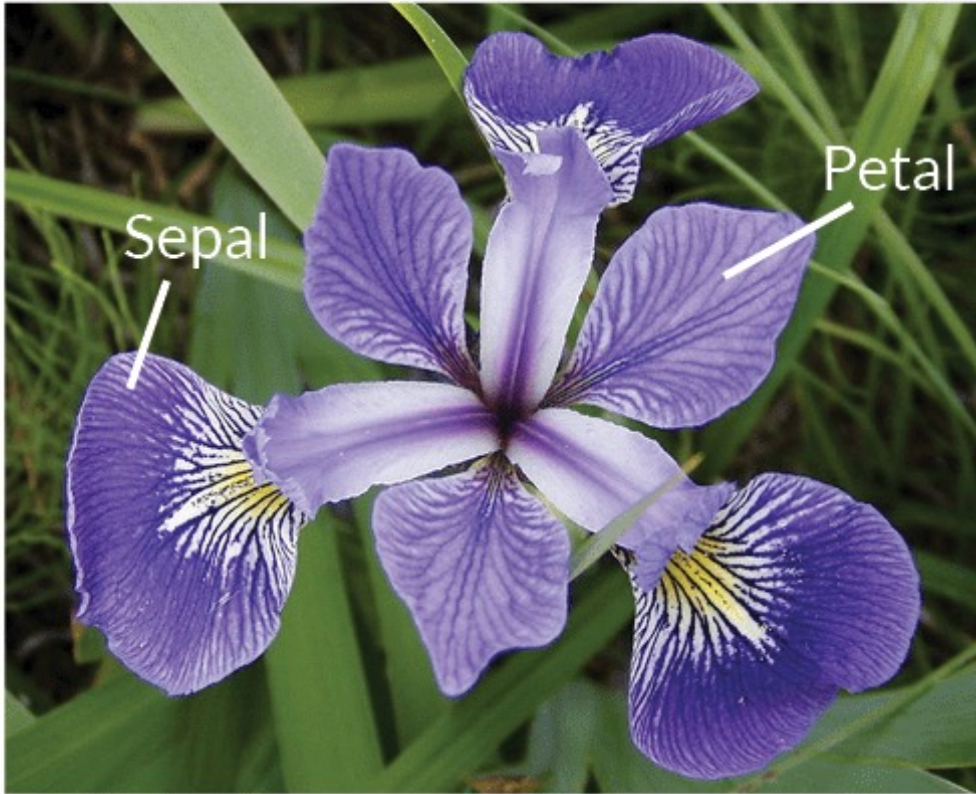
- Everyone spends most of their time on wrangling ! It's hard!
- The tidyverse makes it much easier though; that's its primary purpose



dplyr

- [mutate\(\)](#) adds new variables that are functions of existing variables
- [select\(\)](#) picks variables based on their names.
- [filter\(\)](#) picks cases based on their values.
- [summarise\(\)](#) reduces multiple values down to a single summary.
- [arrange\(\)](#) changes the ordering of the rows.
- You also need to know [group by\(\)](#) which allows you do any function by group.

The Data



Iris Versicolor



Iris Setosa



Iris Virginica

DEMO WITH
RStudio

Resources

- R <https://www.r-project.org/>
- Bioconductor <https://www.bioconductor.org/>
- Tidyverse <https://www.tidyverse.org/packages/>

Where to Get Help

- RStudio Help Panel
- Google the function or error message
- Ask me or fellow students on Slack