DPLYR

WHAT IS DPLYR?

- A grammar of data manipulation that helps to solve the most common data manipulation challenges

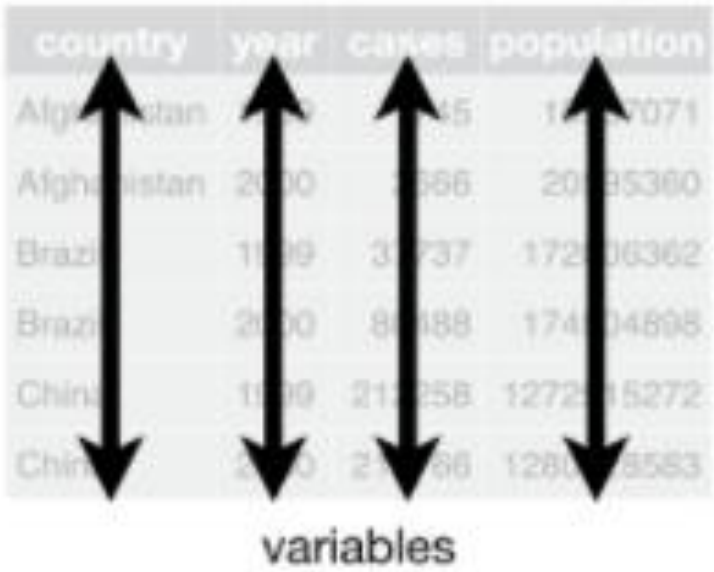
WHAT CHALLENGES?

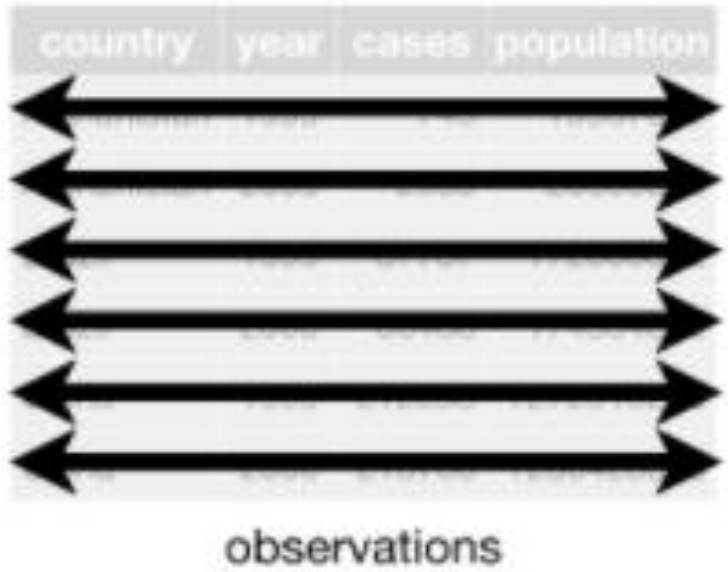
- Add new variables that are functions of existing variables
- Pick variable based on their names
- Pick cases based on their values
- Reduce mutiple values down to a single summary
- Change the ordering of the rows

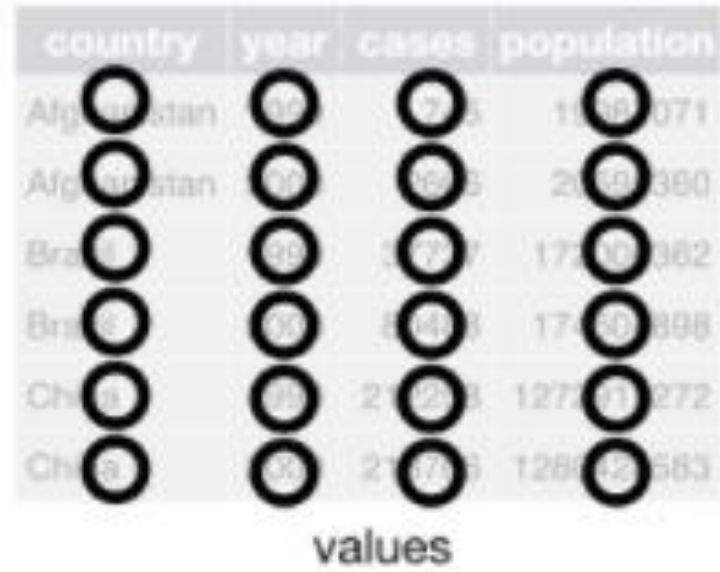
- mutate()
- select()
- filter()
- summarise()
- arrange()

WHAT IS TIDY DATA?

1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.


country
year
cases
population
variables


country
year
cases
population
observations


country
year
cases
population
values

	treatmenta	treatmentb
John Smith	—	2
Jane Doe	16	11
Mary Johnson	3	1

	John Smith	Jane Doe	Mary Johnson
treatmenta	—	16	3
treatmentb	2	11	1

person	treatment	result
John Smith	a	—
Jane Doe	a	16
Mary Johnson	a	3
John Smith	b	2
Jane Doe	b	11
Mary Johnson	b	1

row	a	b	c
A	1	4	7
B	2	5	8
C	3	6	9

(a) Raw data

row	column	value
A	a	1
B	a	2
C	a	3
A	b	4
B	b	5
C	b	6
A	c	7
B	c	8
C	c	9

(b) Molten data

Pipe operator %>%

- Shortcut: ctrl/cmd + shift + m
- Links together a sequence of functions by taking the output of preceding functions as input for subsequent functions
- Increased readability

Single table „verbs“

- Rows:
  - filter() chooses rows based on column values.
  - slice() chooses rows based on location.
  - arrange() changes the order of the rows.
- Columns:
  - select() changes whether or not a column is included.
  - rename() changes the name of columns.
  - mutate() changes the values of columns and creates new columns.
  - relocate() changes the order of the columns.
- Groups of rows:
  - summarise() collapses a group into a single row.

EXERCISE: SPLIT, APPLY, COMBINE

1. Split the dataset into subsets or groups
2. Apply functions to the groups or subsets
3. Combine the results

Task: What is the average, minimum and maximum weight of a human in that dataset?

Reshaping data: „wide“ and „long“

- „Wide“ (horizonal) formats have grouped categorical data.
- In ´“long“ (vertical) formats, every row represents an observation belonging to a particular category.

city	size	amount
New York	large	23
New York	small	14
London	large	22
London	small	16
Beijing	large	121
Beijing	small	56

spread()

gather()

city	large	small
New York	23	14
London	22	16
Beijing	121	56

spread vs. gather by RStudio

```
gather(key = "spice", value = "correct", cinnamon_1:nutmeg_3)
```

baker	cinnamon_1	cardamom_2	nutmeg_3
Emma	1	0	1
Harry	1	1	1
Ruby	1	0	1
Zainab	0	NA	0

baker	spice	correct
Emma	cinnamon_1	1
Harry	cinnamon_1	1
Ruby	cinnamon_1	1
Zainab	cinnamon_1	0
Emma	cardamom_2	0
Harry	cardamom_2	1
Ruby	cardamom_2	0
Zainab	cardamom_2	NA
Emma	nutmeg_3	1
Harry	nutmeg_3	1
Ruby	nutmeg_3	1
Zainab	nutmeg_3	0

Dataset: „dataset_1.xlsx“ under „econdata“

- Variables: country name, region, GPD per capita, capital stock, total-factor-productivity, size of population
- Task 1: Plot the development of the logged average GDP per capita by region between 1950 and 2017.
- Task 2: Create a table of the top 10 countries with the highest growth rates during the period 1950 – 2017 using the following formula:

$$(x_{2017} - x_{1970}) / x_{1970}$$

Sources

- Introduction to dplyr: https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html
- Data transformation: https://r4ds.had.co.nz/transform.html
- Data Wrangling with R: https://cengel.github.io/R-data-wrangling/
- Data management with R Tidyr: https://www.gis-blog.com/data-management-with-r-tidyr-part-1/