

Possible solution to the data wrangling task

Claudius Gräbner-Radkowitsch

Table of contents

1 Task 3: data wrangling	1
1.1 Import raw data	1
1.2 Make the data set tidy	2

```
here::i_am("quarto/Wrangel-Task.qmd")
library(here)
library(dplyr)
library(tidyr)
library(data.table)
```

1 Task 3: data wrangling

There are two main aspects of this task:

1. Import the raw data and make sure this works
2. Make the data tidy

1.1 Import raw data

First, I import the raw data:

```
data_path <- here("data/raw/wrangel_1.csv")
data_raw <- as_tibble(data.table::fread(
  file = data_path, header = TRUE))
head(data_raw)
```

```
# A tibble: 6 x 18
  country name    `2005` `2006` `2007` `2008` `2009` `2010` `2011` `2012` `2013`
  <chr>   <chr>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 Germany Growth    0.732   3.82   2.98  0.960 -5.69   4.18   3.93   0.418  0.438
2 Germany Educat~ NA       4.29   4.37  4.44   4.91   4.94   4.82   4.96   4.94
3 Germany Health~ 10.3    10.2   10.1  10.3   11.2   11.1   10.8   10.9   11.0
4 Italy   Growth    0.818   1.79   1.49 -0.962 -5.28   1.71   0.707 -2.98  -1.84
5 Italy   Educat~   4.24   4.53   4.11  4.39   4.52   4.33   4.12   4.06   4.14
6 Italy   Health~   8.34   8.44   8.14  8.53   8.95   8.92   8.77   8.78   8.78
# i 7 more variables: `2014` <dbl>, `2015` <dbl>, `2016` <dbl>, `2017` <dbl>,
#   `2018` <dbl>, `2019` <dbl>, `2020` <dbl>
```

Without `header = TRUE`, the file will not be imported correctly. You can see this from the columns V1, ...

1.2 Make the data set tidy

In a tidy data set,...

- every column corresponds to one variable
 - Here not satisfied, years should be placed in their proper column `year`
 - Also, under the column `name`, several variable names are mentioned
- every cell contains only one value
 - This is already satisfied
- every row corresponds to one observation
 - This is not true because the observations for Germany in 2008, for instance, are scattered across several rows

It is a good idea to describe how the tidy data set should look like. In the tidy data set, we should have the following columns:

- `country`, `year`, `Growth`, `EducationSpending`, `HealthSpending`

We then start with one step that we expect to be helpful and then see how to continue. A good first step to bring us closer to a tidy data set is to get rid of the year columns:

```
data_v1 <- tidyr::pivot_longer(
  data = data_raw,
  cols = -all_of(c("country", "name")), # also works without all_of
  names_to = "year",
```

```

    values_to = "value")
head(data_v1)

```

```

# A tibble: 6 x 4
  country name    year  value
  <chr>   <chr>  <chr> <dbl>
1 Germany Growth 2005    0.732
2 Germany Growth 2006    3.82
3 Germany Growth 2007    2.98
4 Germany Growth 2008    0.960
5 Germany Growth 2009   -5.69
6 Germany Growth 2010    4.18

```

A viable next step would be to transform the columns `name` and `value` into proper variable names with values. To this end, we make the data wider, taking the names for the new columns from the column `name` and the values from the column `value`:

```

data_v2 <- tidyr::pivot_wider(
  data = data_v1, names_from = "name", values_from = "value")
head(data_v2)

```

```

# A tibble: 6 x 5
  country year Growth EducationSpending HealthSpending
  <chr>   <chr>  <dbl>          <dbl>          <dbl>
1 Germany 2005    0.732          NA            10.3
2 Germany 2006    3.82           4.29           10.2
3 Germany 2007    2.98           4.37           10.1
4 Germany 2008    0.960           4.44           10.3
5 Germany 2009   -5.69           4.91           11.2
6 Germany 2010    4.18           4.94           11.1

```

Hurray, finished! All three requirements are met!

If there was also the task of saving the data:

```

data.table::fwrite(
  x = data_v2, file = here::here("data/tidy/wrangel_1_tidy.csv"))

```