Recap data wrangling

Some worked examples and exercises

Applied Data Science using R, Session 12

Prof. Dr. Claudius Gräbner-Radkowitsch
Europa-University Flensburg, Department of Pluralist Economics
www.claudius-graebner.com | @ClaudiusGraebner | claudius@claudius-graebner.com

Europa-Universität Flensburg

Europa-Universität Flensburg
International Institute of Management and Economic Education
Department of Pluralist Economics

A worked example

Vantage point

```
# A tibble: 8 × 7
  country     Variable          `2008` `2009` `2010` `2011` `2012`
  <chr>       <chr>              <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 Germany     HealthSpending     10.3   11.2   11.1   10.8   10.9
2 Germany     EducationSpending   4.44   4.91   4.94   4.82   4.96
3 Italy       HealthSpending      8.53   8.95   8.92   8.77   8.78
4 Italy       EducationSpending   4.39   4.52   4.33   4.12   4.06
5 Netherlands HealthSpending      9.28   9.99  10.2   10.2   10.5
6 Netherlands EducationSpending   5.03   5.43   5.49   5.46   5.41
7 Spain       HealthSpending      8.38   9.11   9.12   9.17   9.16
8 Spain       EducationSpending   4.53   4.91   4.86   4.90   4.47
```

Intermediate step

```
# A tibble: 40 × 4
   country Variable          year  Value
   <chr>   <chr>             <chr> <dbl>
 1 Germany HealthSpending    2008  10.3
 2 Germany HealthSpending    2009  11.2
 3 Germany HealthSpending    2010  11.1
 4 Germany HealthSpending    2011  10.8
 5 Germany HealthSpending    2012  10.9
 6 Germany EducationSpending 2008   4.44
 7 Germany EducationSpending 2009   4.91
 8 Germany EducationSpending 2010   4.94
 9 Germany EducationSpending 2011   4.82
10 Germany EducationSpending 2012   4.96
# … with 30 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Goal: tidy data

```
# A tibble: 20 × 4
   country     year  HealthSpending EducationSpending
   <chr>       <chr>          <dbl>             <dbl>
 1 Germany     2008            10.3              4.44
 2 Germany     2009            11.2              4.91
 3 Germany     2010            11.1              4.94
 4 Germany     2011            10.8              4.82
 5 Germany     2012            10.9              4.96
 6 Italy       2008             8.53             4.39
 7 Italy       2009             8.95             4.52
 8 Italy       2010             8.92             4.33
 9 Italy       2011             8.77             4.12
10 Italy       2012             8.78             4.06
11 Netherlands 2008             9.28             5.03
12 Netherlands 2009             9.99             5.43
13 Netherlands 2010            10.2              5.49
14 Netherlands 2011            10.2              5.46
15 Netherlands 2012            10.5              5.41
16 Spain       2008             8.38             4.53
17 Spain       2009             9.11             4.91
18 Spain       2010             9.12             4.86
19 Spain       2011             9.17             4.90
20 Spain       2012             9.16             4.47
```

A worked example

```
# A tibble: 8 × 7
  country     Variable          `2008` `2009` `2010` `2011` `2012`
  <chr>       <chr>              <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
1 Germany     HealthSpending      10.3   11.2   11.1   10.8   10.9
2 Germany     EducationSpending    4.44   4.91   4.94   4.82   4.96
3 Italy       HealthSpending       8.53   8.95   8.92   8.77   8.78
4 Italy       EducationSpending    4.39   4.52   4.33   4.12   4.06
5 Netherlands HealthSpending       9.28   9.99  10.2   10.2   10.5
6 Netherlands EducationSpending    5.03   5.43   5.49   5.46   5.41
7 Spain       HealthSpending       8.38   9.11   9.12   9.17   9.16
8 Spain       EducationSpending    4.53   4.91   4.86   4.90   4.47
```

```
# A tibble: 40 × 4
   country Variable          year  Value
   <chr>   <chr>             <chr> <dbl>
 1 Germany HealthSpending    2008  10.3
 2 Germany HealthSpending    2009  11.2
 3 Germany HealthSpending    2010  11.1
 4 Germany HealthSpending    2011  10.8
 5 Germany HealthSpending    2012  10.9
 6 Germany EducationSpending 2008   4.44
 7 Germany EducationSpending 2009   4.91
 8 Germany EducationSpending 2010   4.94
 9 Germany EducationSpending 2011   4.82
10 Germany EducationSpending 2012   4.96
# … with 30 more rows
# ℹ Use `print(n = ...)` to see more rows
```

```
# A tibble: 20 × 4
   country     year  HealthSpending EducationSpending
   <chr>       <chr>          <dbl>             <dbl>
 1 Germany     2008            10.3              4.44
 2 Germany     2009            11.2              4.91
 3 Germany     2010            11.1              4.94
 4 Germany     2011            10.8              4.82
 5 Germany     2012            10.9              4.96
 6 Italy       2008             8.53             4.39
 7 Italy       2009             8.95             4.52
 8 Italy       2010             8.92             4.33
 9 Italy       2011             8.77             4.12
10 Italy       2012             8.78             4.06
11 Netherlands 2008             9.28             5.03
12 Netherlands 2009             9.99             5.43
13 Netherlands 2010            10.2              5.49
14 Netherlands 2011            10.2              5.46
15 Netherlands 2012            10.5              5.41
16 Spain       2008             8.38             4.53
17 Spain       2009             9.11             4.91
18 Spain       2010             9.12             4.86
19 Spain       2011             9.17             4.90
20 Spain       2012             9.16             4.47
```

```r
intermediate_step <- vantage_point %>%
  pivot_longer(
    cols = -c("country", "Variable"),
    names_to = "year",
    values_to = "Value")
```

```r
final_result <- intermediate_step %>%
  pivot_wider(
    names_from = "Variable",
    values_from = "Value")
```

```r
final_result <- vantage_point %>%
  pivot_longer(
    cols = -c("country", "Variable"),
    names_to = "year",
    values_to = "Value") %>%
  pivot_wider(
    names_from = "Variable",
    values_from = "Value")
```

Take-Aways: the general wrangling workflow

1. After reading in the raw data print it using `head()`
2. Then write down how the desired version of the data set looks like
3. Think step-by-step how you can reach the final version
   - Either backwards from the goal, or forwards from the start
4. After thinking about the steps, write down the functions you need
5. Start coding

Your turn

- Get together in groups of two
- Each of you works on the following task, after 5 minutes you explain to each other **what** you did, **how** you did it, and **why** you did it
  - Important: before start coding, explicate your final goal and make yourself a plan of how to proceed!

Download the data set from `ex1.csv`, import it and make it tidy!

A more complex worked example

Vantage point

```
# A tibble: 8 × 7
  country     Variable    `2008`  `2009`  `2010`  `2011`  `2012`
  <chr>       <chr>        <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 Germany     GDP_total  3.14e12 2.96e12 3.09e12 3.21e12 3.22e12
2 Germany     Population 8.21e 7 8.19e 7 8.18e 7 8.03e 7 8.04e 7
3 Italy       GDP_total  1.97e12 1.87e12 1.90e12 1.91e12 1.86e12
4 Italy       Population 5.88e 7 5.91e 7 5.93e 7 5.94e 7 5.95e 7
5 Netherlands GDP_total  7.55e11 7.28e11 7.38e11 7.49e11 7.41e11
6 Netherlands Population 1.64e 7 1.65e 7 1.66e 7 1.67e 7 1.68e 7
7 Spain       GDP_total  1.24e12 1.20e12 1.20e12 1.19e12 1.15e12
8 Spain       Population 4.60e 7 4.64e 7 4.66e 7 4.67e 7 4.68e 7
```

Goal: Average GDP per capita

```
# A tibble: 4 × 2
  country     GDP_pc_avg
  <chr>            <dbl>
1 Germany          38455.
2 Italy            32124.
3 Netherlands      44694.
4 Spain            25709.
```

```
# A tibble: 40 × 4
   country Variable   year    value
   <chr>   <chr>      <chr>   <dbl>
 1 Germany GDP_total  2008  3.14e12
 2 Germany GDP_total  2009  2.96e12
 3 Germany GDP_total  2010  3.09e12
 4 Germany GDP_total  2011  3.21e12
 5 Germany GDP_total  2012  3.22e12
 6 Germany Population 2008  8.21e 7
 7 Germany Population 2009  8.19e 7
 8 Germany Population 2010  8.18e 7
 9 Germany Population 2011  8.03e 7
10 Germany Population 2012  8.04e 7
# … with 30 more rows
```

```
# A tibble: 20 × 4
   country year  GDP_total Population
   <chr>   <chr>     <dbl>      <dbl>
 1 Germany 2008    3.14e12   82110097
 2 Germany 2009    2.96e12   81902307
 3 Germany 2010    3.09e12   81776930
 4 Germany 2011    3.21e12   80274983
 5 Germany 2012    3.22e12   80425823
 6 Italy   2008    1.97e12   58826731
 7 Italy   2009    1.87e12   59095365
 8 Italy   2010    1.90e12   59277417
 9 Italy   2011    1.91e12   59379449
10 Italy   2012    1.86e12   59539717
# … with 10 more rows
```

```
# A tibble: 20 × 5
   country year  GDP_total Population GDP_pc
   <chr>   <chr>     <dbl>      <dbl>  <dbl>
 1 Germany 2008    3.14e12   82110097 38278.
 2 Germany 2009    2.96e12   81902307 36190.
 3 Germany 2010    3.09e12   81776930 37761.
 4 Germany 2011    3.21e12   80274983 39977.
 5 Germany 2012    3.22e12   80425823 40069.
 6 Italy   2008    1.97e12   58826731 33530.
 7 Italy   2009    1.87e12   59095365 31615.
 8 Italy   2010    1.90e12   59277417 32058.
 9 Italy   2011    1.91e12   59379449 32229.
10 Italy   2012    1.86e12   59539717 31185.
# … with 10 more rows
```

Your turn

- Get together in groups of two and download the data set **`ex2.csv`**
- Implement the pathway of transformations just discussed:

```
# A tibble: 8 × 7
  country     Variable    `2008`  `2009`  `2010`  `2011`  `2012`
  <chr>       <chr>        <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 Germany     GDP_total  3.14e12 2.96e12 3.09e12 3.21e12 3.22e12
2 Germany     Population 8.21e 7 8.19e 7 8.18e 7 8.03e 7 8.04e 7
3 Italy       GDP_total  1.97e12 1.87e12 1.90e12 1.91e12 1.86e12
4 Italy       Population 5.88e 7 5.91e 7 5.93e 7 5.94e 7 5.95e 7
5 Netherlands GDP_total  7.55e11 7.28e11 7.38e11 7.49e11 7.41e11
6 Netherlands Population 1.64e 7 1.65e 7 1.66e 7 1.67e 7 1.68e 7
7 Spain       GDP_total  1.24e12 1.20e12 1.20e12 1.19e12 1.15e12
8 Spain       Population 4.60e 7 4.64e 7 4.66e 7 4.67e 7 4.68e 7
```

```
# A tibble: 4 × 2
  country     GDP_pc_avg
  <chr>            <dbl>
1 Germany          38455.
2 Italy            32124.
3 Netherlands      44694.
4 Spain            25709.
```

```
# A tibble: 40 × 4
   country Variable   year    value
   <chr>   <chr>      <chr>   <dbl>
 1 Germany GDP_total  2008  3.14e12
 2 Germany GDP_total  2009  2.96e12
 3 Germany GDP_total  2010  3.09e12
 4 Germany GDP_total  2011  3.21e12
 5 Germany GDP_total  2012  3.22e12
 6 Germany Population 2008  8.21e 7
 7 Germany Population 2009  8.19e 7
 8 Germany Population 2010  8.18e 7
 9 Germany Population 2011  8.03e 7
10 Germany Population 2012  8.04e 7
# … with 30 more rows
```

```
# A tibble: 20 × 4
   country year  GDP_total Population
   <chr>   <chr>     <dbl>      <dbl>
 1 Germany 2008    3.14e12   82110097
 2 Germany 2009    2.96e12   81902307
 3 Germany 2010    3.09e12   81776930
 4 Germany 2011    3.21e12   80274983
 5 Germany 2012    3.22e12   80425823
 6 Italy   2008    1.97e12   58826731
 7 Italy   2009    1.87e12   59095365
 8 Italy   2010    1.90e12   59277417
 9 Italy   2011    1.91e12   59379449
10 Italy   2012    1.86e12   59539717
# … with 10 more rows
```

```
# A tibble: 20 × 5
   country year  GDP_total Population GDP_pc
   <chr>   <chr>     <dbl>      <dbl>  <dbl>
 1 Germany 2008    3.14e12   82110097 38278.
 2 Germany 2009    2.96e12   81902307 36190.
 3 Germany 2010    3.09e12   81776930 37761.
 4 Germany 2011    3.21e12   80274983 39977.
 5 Germany 2012    3.22e12   80425823 40069.
 6 Italy   2008    1.97e12   58826731 33530.
 7 Italy   2009    1.87e12   59095365 31615.
 8 Italy   2010    1.90e12   59277417 32058.
 9 Italy   2011    1.91e12   59379449 32229.
10 Italy   2012    1.86e12   59539717 31185.
# … with 10 more rows
```

Final task

- Consider the data set `ex3.csv`:

```
# A tibble: 13 × 7
   country       income              `2012` `2011` `2010` `2009` `2008`
   <chr>         <chr>                <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
 1 Chile         High income          4.51   4.44   4.09   3.88   4.07
 2 China         Upper middle income  7.05   6.90   6.34   5.80   5.44
 3 Germany       High income          9.45   9.30   9.45   8.97   9.62
 4 India         Lower middle income  1.51   1.41   1.34   1.29   1.19
 5 Italy         High income          6.33   6.68   6.84   6.72   7.56
 6 Namibia       Upper middle income  1.61   1.54   1.48   1.45   1.44
 7 Netherlands   High income          9.40   9.51  10.3    9.71  10.0
 8 Nicaragua     Lower middle income  0.784  0.808  0.774  0.764  0.794
 9 Peru          Upper middle income  1.63   1.65   1.55   1.43   1.34
10 Saudi Arabia  High income         16.9   16.4   16.3   15.3   15.1
11 South Africa  Upper middle income  8.08   7.87   8.30   8.01   8.57
12 Spain         High income          5.76   5.87   5.87   6.20   7.06
13 United States High income         15.8   16.6   17.4   16.8   18.3
```

- Compute the deviation from the mean CO2 emissions for each country in each year, then compute the average deviation per income group!

- Finally, take this result and average the deviations per group over time!