

Data Analysis and Visualization in R (IN2339)

Exercise Session 3 - Tidy Data & Combining Tables

Felix Brechtmann, Vicente Yépez, Žiga Avsec, Julien Gagneur

Quizzes (from the lecture)

The following quizzes will be solved orally by the students and the professor during the lecture.

1. Is the following dataset tidy?

Dataset: `mtcars` (Motor Trend Car Road Tests)

- mpg: Miles/(US) gallon
- cyl: Number of cylinders
- hp: Gross horsepower
- wt: Weight (1000 lbs)

```
##           model  mpg  cyl  hp   wt
## 1:      Mazda RX4 21.0   6 110 2.620
## 2:    Mazda RX4 Wag 21.0   6 110 2.875
## 3:      Datsun 710 22.8   4  93 2.320
## 4:   Hornet 4 Drive 21.4   6 110 3.215
## 5: Hornet Sportabout 18.7   8 175 3.440
## 6:      Valiant 18.1   6 105 3.460
```

1. Yes
2. No

2. Is the following dataset tidy?

```
##      country  1999  2000
## 1: Afghanistan   745  2666
## 2:      Brazil 37737 80488
## 3:      China 212258 213766
```

1. Yes
2. No

3. What transformations are required to tidy the following data?

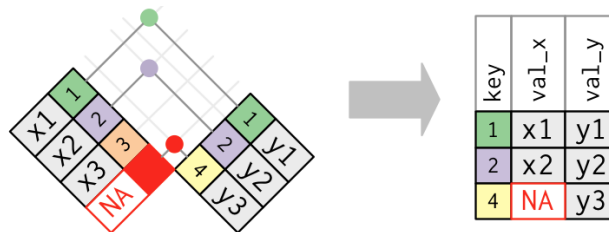
religion	<\$10k	\$10-20k	\$20-30k	\$30-40k	\$40-50k	\$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
Don't know/refused	15	14	15	11	10	35

1. cast
2. melt

3. cast and melt
4. unite
4. What transformations are required to tidy the following data?

id	year	month	element	d1	d2	d3	d4	d5	d6	d7	d8
MX17004	2010	1	tmax	—	—	—	—	—	—	—	—
MX17004	2010	1	tmin	—	—	—	—	—	—	—	—
MX17004	2010	2	tmax	—	27.3	24.1	—	—	—	—	—
MX17004	2010	2	tmin	—	14.4	14.4	—	—	—	—	—
MX17004	2010	3	tmax	—	—	—	—	32.1	—	—	—
MX17004	2010	3	tmin	—	—	—	—	14.2	—	—	—

1. separate
2. unite
3. melt and cast
4. melt, unite and cast
5. How do you perform the data table merge pictured here?



1. Inner, all = FALSE
2. Full, all = TRUE
3. Left, all.x = TRUE
4. Right, all.y = TRUE

Tutorial

The following exercises will be solved during the tutorial sessions.

Section 00 - Getting ready

```
library(data.table)
library(magrittr)
library(tidyr)
```

Section 01 - Tidy Data Warm Up

Hint: You can solve parts 1, 2 and 3 by looking at the printed output. You do not need to load the datasets.

1. Visually examine the dataset `AirPassengers`. Which of the following is true:

```
##      Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
## 1949 112 118 132 129 121 135 148 148 136 119 104 118
```

```
## 1950 115 126 141 135 125 149 170 170 158 133 114 140
## 1951 145 150 178 163 172 178 199 199 184 162 146 166
## 1952 171 180 193 181 183 218 230 242 209 191 172 194
## 1953 196 196 236 235 229 243 264 272 237 211 180 201
## 1954 204 188 235 227 234 264 302 293 259 229 203 229
## 1955 242 233 267 269 270 315 364 347 312 274 237 278
## 1956 284 277 317 313 318 374 413 405 355 306 271 306
## 1957 315 301 356 348 355 422 465 467 404 347 305 336
## 1958 340 318 362 348 363 435 491 505 404 359 310 337
## 1959 360 342 406 396 420 472 548 559 463 407 362 405
## 1960 417 391 419 461 472 535 622 606 508 461 390 432
```

- `AirPassengers` is tidy data: it has one year for each row.
- `AirPassengers` is not tidy: we need at least one column with a character vector.
- `AirPassengers` is not tidy: it is a matrix instead of a data frame.
- `AirPassengers` is not tidy: to be tidy we would have to wrangle it to have three columns (year, month and value), then each passenger count would have a row.

2. Visually examine the dataset `ChickWeight`. Which of the following is true:

```
##   weight Time Chick Diet
## 1     42    0     1     1
## 2     51    2     1     1
## 3     59    4     1     1
## 4     64    6     1     1
## 5     76    8     1     1
## 6     93   10     1     1
```

- `ChickWeight` is not tidy: each chick has more than one row.
- `ChickWeight` is tidy: each observation (a weight) is represented by one row. The chick from which this measurement came is one of the variables.
- `ChickWeight` is not tidy: we are missing the year column.
- `ChickWeight` is tidy: it is stored in a data frame.

3. Visually examine the dataset `spanish_vowels`. Is the data set tidy?

```
##           label rep frequency1 frequency2
## 1: p01-male-a    1  615.4477  1230.806
## 2: p01-male-a    2  644.6112  1281.965
## 3: p01-male-a    3  607.9174  1247.960
## 4: p01-male-e    1  476.9079  1612.076
## 5: p01-male-e    2  457.2205  1839.456
## ---
## 746: p50-female-o  2  577.1894  1310.138
## 747: p50-female-o  3  545.5014  1214.094
## 748: p50-female-u  1  405.7645  1491.935
## 749: p50-female-u  2  458.0345  1141.513
## 750: p50-female-u  3  457.4308  1181.657
```

4. The `example_product_data.csv` file describes the number of times a person bought product “a” and “b”. Load the file into a `data.table`.

5. Transform `product_dt` into a long format using `data.table` commands.

6. Transform the table from the long format back into a wide format. Check that it is equal to the original `data.table`.

Section 02 - Merge Warm Up

Prepare two tables by running the following code:

```
mtcars_dt <- as.data.table(mtcars)
mtcars_dt[, carname := rownames(mtcars)]

dt1 <- mtcars_dt[5:25,.(carname, mpg, cyl)]
dt2 <- mtcars_dt[1:10, .(carname, gear)]
```

1. How many rows does the inner merge of dt1 and dt2 have?
2. How many rows does the left merge of dt1 and dt2 have?
3. How many rows does the outer merge of dt1 and dt2 have?

Section 03 - Weather dataset

1. Read in the weather dataset `weather.txt` as a `data.table`. Please find this file on Moodle and save it to a folder called `extdata` in your working directory.
2. Why is this dataset messy?
3. How would a tidy version of it look like? Do not give the code, only describe how the tidy table would look like.
4. Create a tidy version of the weather dataset.

Homework

Please solve the exercises below at home. The solutions will be discussed in the central exercise.

Section 04 -Scattered data across many files

The `baby-names` folder contains 258 csv-files (`1999.girl.csv`, `1999.boy.csv`, ...) which store name frequencies for a particular year and sex.

1. Create a list containing all file paths in the folder, i.e. (`extdata/baby-names/1999.girl.csv`, `extdata/baby-names/1999.boys.csv`, ...). Refer to Appendix A of the script if you are having trouble with the file paths.
2. Name the list entries with the basename of the corresponding file path. The basename of a path is the actual name of the file, e.g. the basename of `extdata/baby-names/1999.girl.csv` is `1999.girl.csv`.
3. Read in the data from all files into one table. *Hint*: when you read many files and gather them into one table, be sure to add a column that identifies each file. `rbindlist()`
4. Is the data tidy? If not, tidy it up.

Section 05 - Small case-study: cleaning up a gene-expression dataset in yeast

In this section, we will read and clean up the data from the paper:

- Gagneur, Julien, et al. "Genotype-environment interactions reveal causal pathways that mediate genetic effects on phenotype." *PLoS Genet* 9.9 (2013): e1003803. <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003803>

You will not need to understand the data in depth for this exercise but if you are interested you can find a detailed description in section 7.1 of the script.

1. Read in the two files in the folder `eqt1`.

The first file contains the genotypes of yeast strains and a strain identifier. The second file contains information on how quickly each strain grows in different growth media.

```
gt <- fread('extdata/eqtl/genotype.txt')
dim(gt)
```

```
## [1] 158 1001
```

```
head(gt[,1:5])
```

```
##      strain      mrk_1      mrk_14      mrk_27      mrk_40
## 1: seg_01B Lab strain Lab strain Lab strain Lab strain
## 2: seg_01C Wild isolate Wild isolate Wild isolate Wild isolate
## 3: seg_01D Lab strain Lab strain Lab strain Lab strain
## 4: seg_02B Lab strain Lab strain Lab strain Lab strain
## 5: seg_02C Wild isolate Wild isolate Wild isolate Lab strain
## 6: seg_02D Lab strain Lab strain Lab strain Wild isolate
```

```
growth <- fread('extdata/eqtl/growth.txt')
head(growth)
```

```
##      strain      YPD      YPD_BPS YPD_Rapa      YPE      YPMalt
## 1: seg_01B 12.60399 10.460795 2.500311 5.265698 6.720447
## 2: seg_01C 10.79114 11.632019      NA 5.365259 7.429273
## 3: seg_01D 12.81727 10.423287 3.142154 5.577932 6.905589
## 4: seg_02B 10.29921 9.103611 4.314388 3.257843 4.924324
## 5: seg_02C 11.13278 9.263100 3.548543 3.815689 4.413402
## 6: seg_02D 13.91084 11.750178      NA 5.672890 7.926200
```

2. Come up with a strategy, how you can transform the two tables shown above into the single table shown below.

```
head(dt)
```

```
##      strain media growth_rate marker      gt
## 1: seg_01B YPD      12.60399 mrk_1 Lab strain
## 2: seg_01B YPD      12.60399 mrk_14 Lab strain
## 3: seg_01B YPD      12.60399 mrk_27 Lab strain
## 4: seg_01B YPD      12.60399 mrk_40 Lab strain
## 5: seg_01B YPD      12.60399 mrk_54 Lab strain
## 6: seg_01B YPD      12.60399 mrk_67 Lab strain
```

```
summary(dt)
```

```
##      strain      media      growth_rate      marker
## seg_01B: 5000 YPD      :158000 Min.    : 1.57 mrk_1 : 790
## seg_01C: 5000 YPD_BPS :158000 1st Qu.: 4.55 mrk_14 : 790
## seg_01D: 5000 YPD_Rapa:158000 Median : 6.93 mrk_27 : 790
## seg_02B: 5000 YPE      :158000 Mean    : 7.60 mrk_40 : 790
## seg_02C: 5000 YPMalt  :158000 3rd Qu.:10.70 mrk_54 : 790
## seg_02D: 5000      Max.    :16.27 mrk_67 : 790
## (Other):760000 NA's    :42000 (Other):785260
##      gt
## Lab strain :398145
## Wild isolate:391855
##
##
##
```

```
##  
##
```

3. Write code that implements your strategy to transform the two tables into the one shown above.

4. Spoiler: These are the things you can do with tidy data. When you are done run the following line of code and observe the result:

```
library(ggplot2)  
ggplot(dt[marker %in% c('mrk_5211', 'mrk_1653')], aes(marker, growth_rate, color=gt)) +  
  geom_boxplot() + facet_wrap(~media)
```