Data Analysis and Visualization Exercise 5

Vicente Yépez, Žiga Avsec

3 January 2019

Setup

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

Questions

Q1 Product dataset

The example_product_data file describes the number of times a person bought product "a" and "b"

```r
messy_file <- file.path('extdata', 'example_product_data.csv')
messy_dt <- fread(messy_file)
messy_dt
##             name producta productb
## 1:      John Doe       NA       12
## 2:     Marry Doe        3        1
## 3: John Johnson        5        1
```

Why is this data-set messy? Which columns should a tidy version of this table have?

A1

Q2 Product dataset

Tranform `messy_dt` into a tidy from.

A2

Q3 Weather dataset

Read in the weather dataset `weather.txt`. Why is this dataset messy? How would a tidy version of it look like?

A3

Q4 Weather dataset

Create a tidy version of the weather dataset.

A4

Q5 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. 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()`

A5

Q6

Is the data tidy? If not, tidy it up.

A6

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

Here, we will read and clean up the data from the paper:

- *Bauer et.al., 2007, Coordination of Growth Rate, Cell Cycle, Stress Response, and Metabolic Activity in Yeast, MBoC*, http://www.molbiolcell.org/content/19/1/352.abstract

Read in the data:

```
original_dt <- fread("extdata/gene_expression.tds")
dim(original_dt)
## [1] 5537   40
head(original_dt, n = 2)
##           GID       YORF
## 1: GENE1331X A_06_P5820
## 2: GENE4924X A_06_P5866
##                                                                                     NAME
## 1: SFB2       || ER to Golgi transport || molecular function unknown || YNL049C || 1082129
## 2:       || biological process unknown || molecular function unknown || YNL095C || 1086222
##    GWEIGHT G0.05  G0.1 G0.15  G0.2 G0.25  G0.3 N0.05 N0.1 N0.15  N0.2 N0.25
## 1:       1 -0.24 -0.13 -0.21 -0.15 -0.05 -0.05  0.20 0.24 -0.20 -0.42 -0.14
## 2:       1  0.28  0.13 -0.40 -0.48 -0.11  0.17  0.31 0.00 -0.63 -0.44 -0.26
##    N0.3 P0.05  P0.1 P0.15  P0.2 P0.25 P0.3 S0.05  S0.1 S0.15 S0.2 S0.25 S0.3
## 1: 0.09 -0.26 -0.20 -0.22 -0.31  0.04 0.34 -0.51 -0.12  0.09 0.09  0.20 0.08
## 2: 0.21 -0.09 -0.04 -0.10  0.15  0.20 0.63  0.53  0.15 -0.01 0.12 -0.15 0.32
##    L0.05 L0.1 L0.15 L0.2 L0.25 L0.3 U0.05  U0.1 U0.15  U0.2 U0.25 U0.3
## 1:  0.18 0.18  0.13 0.20  0.17 0.11 -0.06 -0.26 -0.05 -0.28 -0.19 0.09
## 2:  0.16 0.09  0.02 0.04  0.03 0.01 -1.02 -0.91 -0.59 -0.61 -0.17 0.18
```

Column description:

- GID - gene ID
- YORF - Some other ID
- NAME - gene description composed of:
  - Gene name
  - Biological process
  - Molecular function
  - Systematic ID
  - Some other ID
- GWEIGHT - some type of weight
- G0.05, .., P0.03 - gene expression values for measured at different nutritient and growth rates:
  - Nutritients (G, N, P, . . .):
    - G = Glucose
    - L = Leucine
    - P = Phosphate
    - S = Sulphate
    - N = Ammonia
    - U = Uracil
  - Growth rate (0.05, 0.3, . . .)

Q6

Why is this dataset not tidy?

A6

Q7 - Transorm it into a tidy form

Provide a tidy dataset in the folowing form:

```
##    name           biological_process            molecular_function
## 1: SFB2        ER to Golgi transport    molecular function unknown
## 2:        biological process unknown    molecular function unknown
## 3: QRI7 proteolysis and peptidolysis metalloendopeptidase activity
## 4: CFT2        mRNA polyadenylation*                   RNA binding
## 5: SSO2              vesicle fusion*              t-SNARE activity
## 6: PSP2   biological process unknown    molecular function unknown
##    systematic_name nutrient rate expression
## 1:         YNL049C  Glucose 0.05      -0.24
## 2:         YNL095C  Glucose 0.05       0.28
## 3:         YDL104C  Glucose 0.05      -0.02
## 4:         YLR115W  Glucose 0.05      -0.33
```

```
        ## 5:         YMR183C  Glucose 0.05       0.05
        ## 6:         YML017W  Glucose 0.05      -0.69
```

A7