Journal of Statistical Software

August 2014, Volume 59, Issue 10. *http://www.jstatsoft.org/*

Tidy Data

Hadley Wickham
RStudio

Abstract

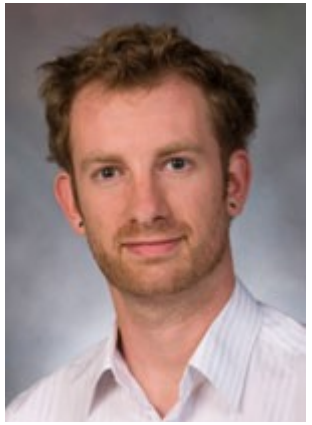
A huge amount of effort is spent cleaning data to get it ready for analysis, but there has been little research on how to make data cleaning as easy and effective as possible. This paper tackles a small, but important, component of data cleaning: data tidying. Tidy datasets are easy to manipulate, model and visualize, and have a specific structure: each variable is a column, each observation is a row, and each type of observational unit is a table. This framework makes it easy to tidy messy datasets because only a small set of tools are needed to deal with a wide range of un-tidy datasets. This structure also makes it easier to develop tidy tools for data analysis, tools that both input and output tidy datasets. The advantages of a consistent data structure and matching tools are demonstrated with a case study free from mundane data manipulation chores.

Keywords: data cleaning, data tidying, relational databases, R.

1. Introduction

It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data (Dasu and Johnson 2003). Data preparation is not just a first step, but must be repeated many times over the course of analysis as new problems come to light or new data is collected. Despite the amount of time it takes, there has been surprisingly little research on how to clean data well. Part of the challenge is the breadth of activities it encompasses: from outlier checking, to date parsing, to missing value imputation. To get a handle on the problem, this paper focuses on a small, but important, aspect of data cleaning that I call data *tidying*: structuring datasets to facilitate analysis.

The principles of tidy data provide a standard way to organize data values within a dataset. A standard makes initial data cleaning easier because you do not need to start from scratch and reinvent the wheel every time. The tidy data standard has been designed to facilitate initial exploration and analysis of the data, and to simplify the development of data analysis tools that work well together. Current tools often require translation. You have to spend time



`http://had.co.nz/`

*… **tidy** datasets are all alike but every **messy** dataset is messy in its own way.*

« Les familles heureuses se ressemblent toutes. Les familles malheureuses sont malheureuses chacune à leur manière. »

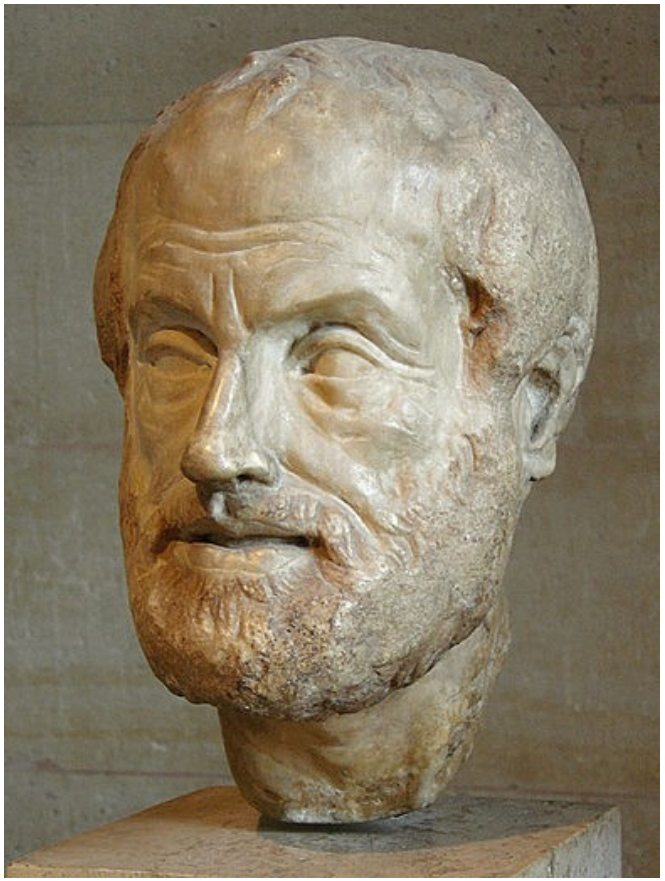
Le principe d'Anna Karenine

En d’autres termes, le succès demande que plusieurs conditions soient réunies. Une seule condition manquée est suffisante pour conduire à l’échec.

https://deselection.wordpress.com/2010/11/12/le-principe-danna-karenine/

Version Aristote

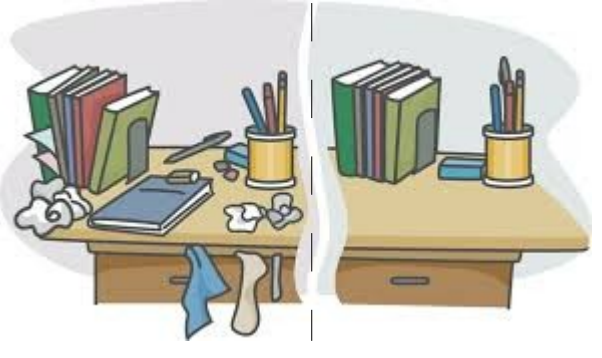
https://en.wikipedia.org/wiki/Anna_Karenina_principle



Much earlier, Aristotle states the same principle in the Nichomachean Ethics (Book 2):

Again, it is possible to fail in many ways (for evil belongs to the class of the unlimited, as the Pythagoreans conjectured, and good to that of the limited), while to succeed is possible only in one way (for which reason also one is easy and the other difficult – to miss the mark easy, to hit it difficult); for these reasons also, then, excess and defect are characteristic of vice, and the mean of virtue; For men are good in but one way, but bad in many.

messy



tidy

Love where you Live

LITTLE MISS TIDY

by Roger Hargreaves

Morceaux choisis

It is often said that 80% of data analysis is spent on the process of cleaning and preparing the data.

data tidying: structuring datasets to facilitate analysis.

This paper [...] provides a comprehensive ``philosophy of data"

Since most real world datasets are not tidy...

Tidy datasets provide a standardized way to link the structure of a dataset (its physical layout) with its semantics (its meaning).

Tidy data

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

Messy data is any other arrangement of the data.

Messy data

Real datasets can, and often do, violate the three precepts of tidy data in almost every way imaginable. While occasionally you do get a dataset that you can start analyzing immediately, this is the exception, not the rule. This section describes the five most common problems with messy datasets, along with their remedies:

- *Column headers are values, not variable names.*
- *Multiple variables are stored in one column.*
- *Variables are stored in both rows and columns.*
- *Multiple types of observational units are stored in the same table.*
- *A single observational unit is stored in multiple tables.*

Surprisingly, most messy datasets, including types of messiness not explicitly described above, can be tidied with a small set of tools: melting, string splitting, and casting.

Exemple

	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

Column headers are values, not variable names

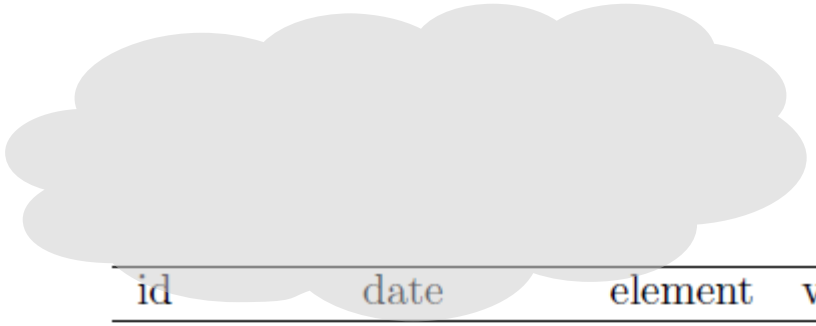
religion	<$10k	$10–20k	$20–30k	$30–40k	$40–50k	$50–75k	...
Agnostic	27	34	60	81	76	137	
Atheist	12	27	37	52	35	70	
Buddhist	27	21	30	34	33	58	
Catholic	418	617	732	670	638	1116	
Don't know/refused	15	14	15	11	10	35	
Evangelical Prot	575	869	1064	982	881	1486	
Hindu	1	9	7	9	11	34	
Historically Black Prot	228	244	236	238	197	223	
Jehovah's Witness	20	27	24	24	21	30	
Jewish	19	19	25	25	30	95	

3 variables :
- religion
- revenu
- effectif

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10–20k	34
Agnostic	$20–30k	60
Agnostic	$30–40k	81
Agnostic	$40–50k	76
Agnostic	$50–75k	137
Agnostic	$75–100k	122
Agnostic	$100–150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

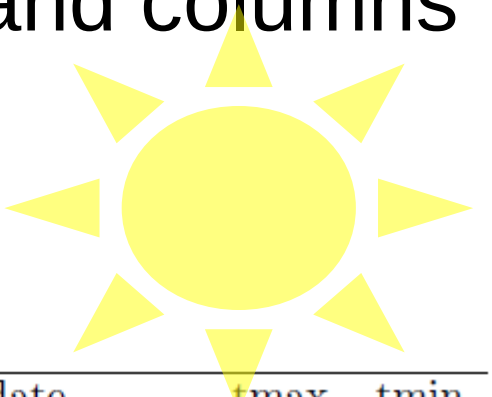
Chaque colonne représente une variable ; chaque ligne, une observation

Variables are stored in both rows and columns



id	date	element	value
MX17004	2010-01-30	tmax	27.8
MX17004	2010-01-30	tmin	14.5
MX17004	2010-02-02	tmax	27.3
MX17004	2010-02-02	tmin	14.4
MX17004	2010-02-03	tmax	24.1
MX17004	2010-02-03	tmin	14.4
MX17004	2010-02-11	tmax	29.7
MX17004	2010-02-11	tmin	13.4
MX17004	2010-02-23	tmax	29.9
MX17004	2010-02-23	tmin	10.7

Cette colonne contient un nom de variable !



id	date	tmax	tmin
MX17004	2010-01-30	27.8	14.5
MX17004	2010-02-02	27.3	14.4
MX17004	2010-02-03	24.1	14.4
MX17004	2010-02-11	29.7	13.4
MX17004	2010-02-23	29.9	10.7
MX17004	2010-03-05	32.1	14.2
MX17004	2010-03-10	34.5	16.8
MX17004	2010-03-16	31.1	17.6
MX17004	2010-04-27	36.3	16.7
MX17004	2010-05-27	33.2	18.2

Une variable par colonne, une observation par ligne

Tidy tools

1) Manipulation

2) Visualisation

3) Modélisation

Manipulation

- **Filter**: subsetting or removing observations based on some condition.
- **Transform**: adding or modifying variables. These modications can involve either a single variable (e.g., log-transformation), or multiple variables (e.g., computing density from weight and volume).
- **Aggregate**: collapsing multiple values into a single value (e.g., by summing or taking means).
- **Sort**: changing the order of observations.

All these operations are made easier when there is a consistent way to refer to variables. Tidy data provides this because each variable resides in its own column.

Ensure input and output-tidiness

plyr,dplyr packages

Visualisation

Tidy visualization tools only need to be input-tidy as their output is visual.

It provides a comprehensive "philosophy of data": one that underlies my work in the plyr (Wickham 2011) and ggplot2 (Wickham 2009) packages.

Modelisation

Modeling is the driving inspiration of this work because most modeling tools work best with tidy datasets.

Every statistical language has a way of describing a model as a connection among different variables, a domain specic language that connects responses to predictors

Conclusion

Apart from tidying, there are many other tasks involved in cleaning data: ***parsing dates and numbers***, ***identifying missing values***, ***correcting character encodings*** *(for international data),* ***matching similar but not identical values*** *(created by typos), verifying experimental design, and filling in structural missing values, not to mention model-based data cleaning that identifies suspicious values. Can we develop other frameworks to make these tasks easier?*