R: A Hitchhikers Guide to Reproducible Research

- (Practice) Over and over

Brendan Palmer,
Clinical Research Facility - Cork &
School of Public Health
@B_A_Palmer

CRF-C
HRB Clinical Research Facility Cork

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh
School of Public Health

Tibbles

- The tidyverse equivalent of data.frames

4 main points of difference:

1. Printing in the console
2. Tibbles don’t change the input
3. Interacting with older code
4. Subsetting - the use of a placeholder (“.”)

- Open the script 03_tibbles.R

readr and more

- fast way to read rectangular data (like csv, tsv)
- read_csv(): comma separated (CSV) files
- read_tsv(): tab separated files
- read_delim(): general delimited files

- readxl supports both the legacy .xls format and the modern xml-based .xlsx format
- Need to load explicitly

- read_sas(): SAS files
- read_sav(): SPSS files
- read_dta(): Stata files
- Also need to load explicitly

- Open the script 04_readr.R

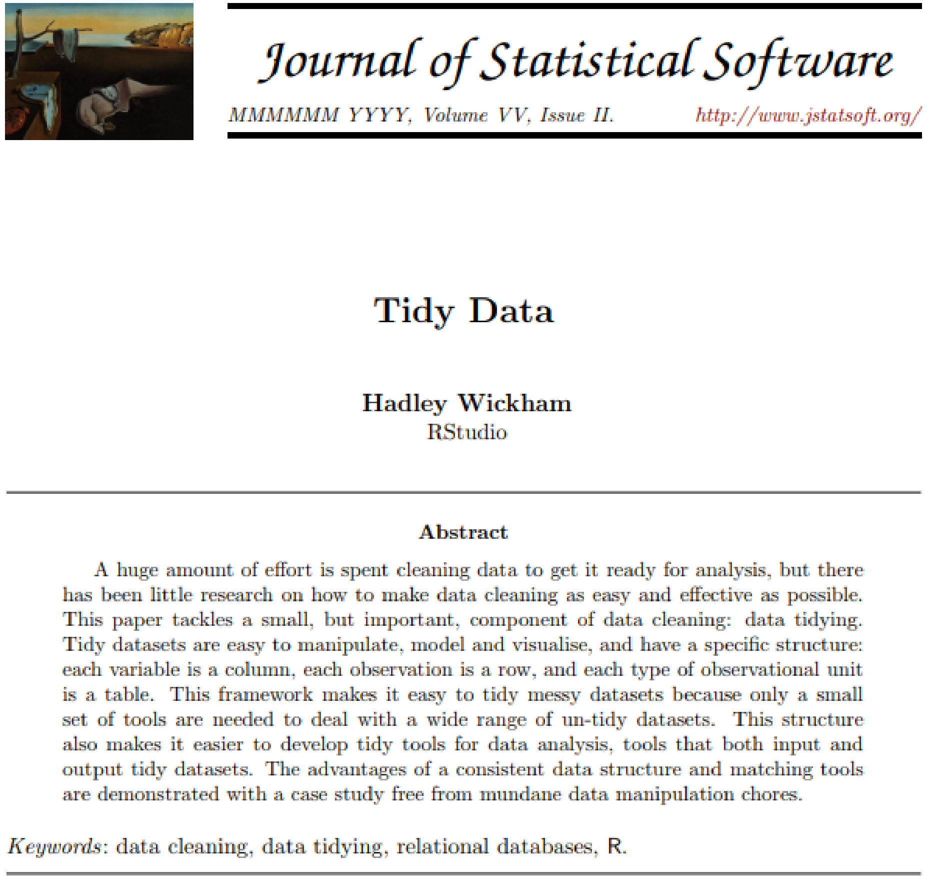
tidyr for…, erm…, tidying

- The tidyverse works with tidy data

Main functions:

- gather()
- separate()
- spread()
- unite()

- Open the script 05_tidyr.R



Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Tidy Data

Hadley Wickham
RStudio

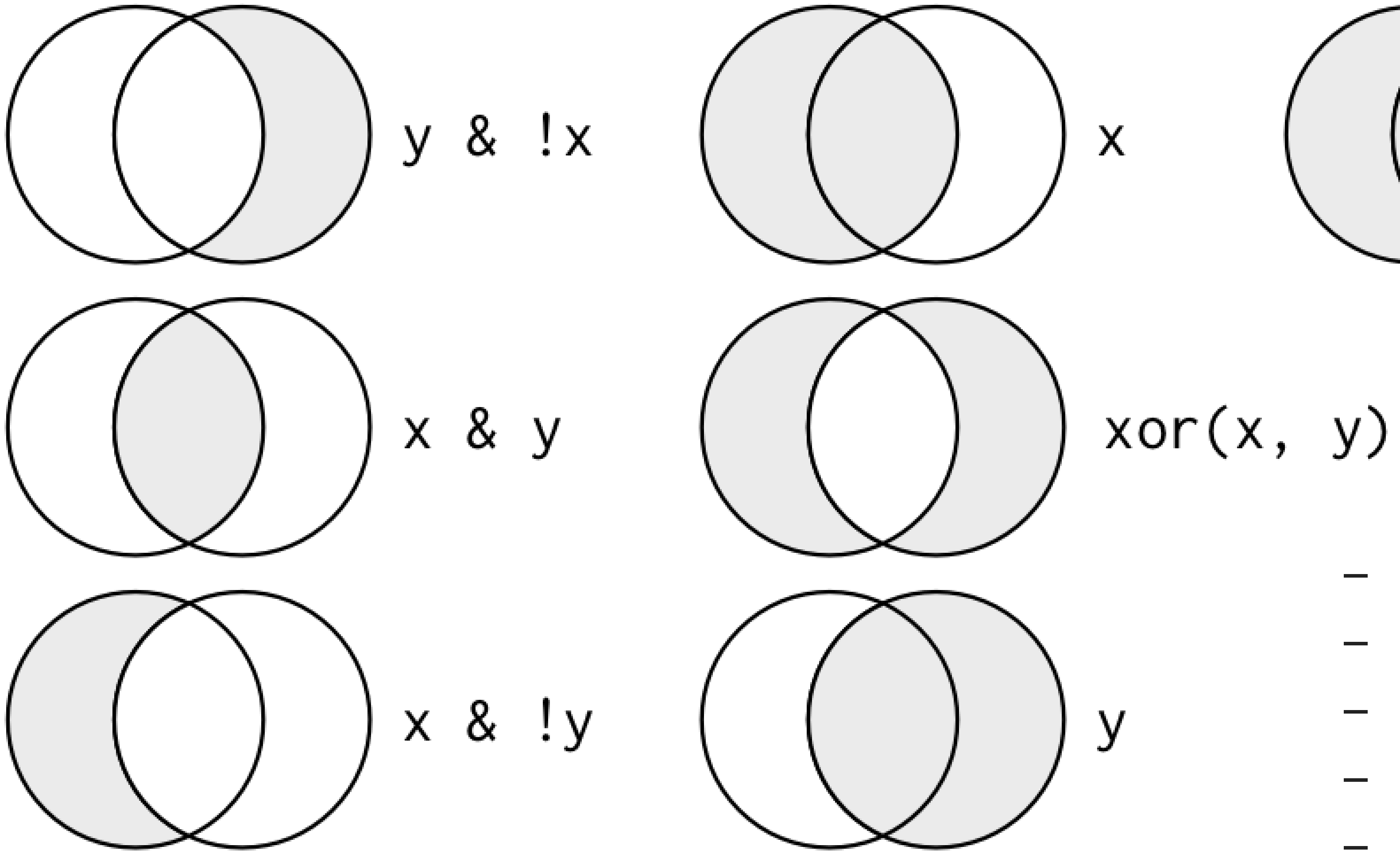
Abstract

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 visualise, 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.

DOI: 10.18637/jss.v059.i10

Logical operators and conditional subsetting



- & -> AND
- | -> OR (inclusive)
- ! -> NOT
- == -> EQUAL
- != -> NOT EQUAL

Image: Hadley Wickham R for Data Science Chapter 3

dplyr for data transformation

- Solves the most common data manipulation challenges

Main functions:

- select()
- filter()
- mutate()
- group_by()
- summarise()
- ………… and many many more

- Open the script 06_dplyr.R

Joining data frames

- Important to understand the chain of relations between the tables

- Variables used to connect each pair of tables are called **keys**

  - primary keys
  - foreign keys

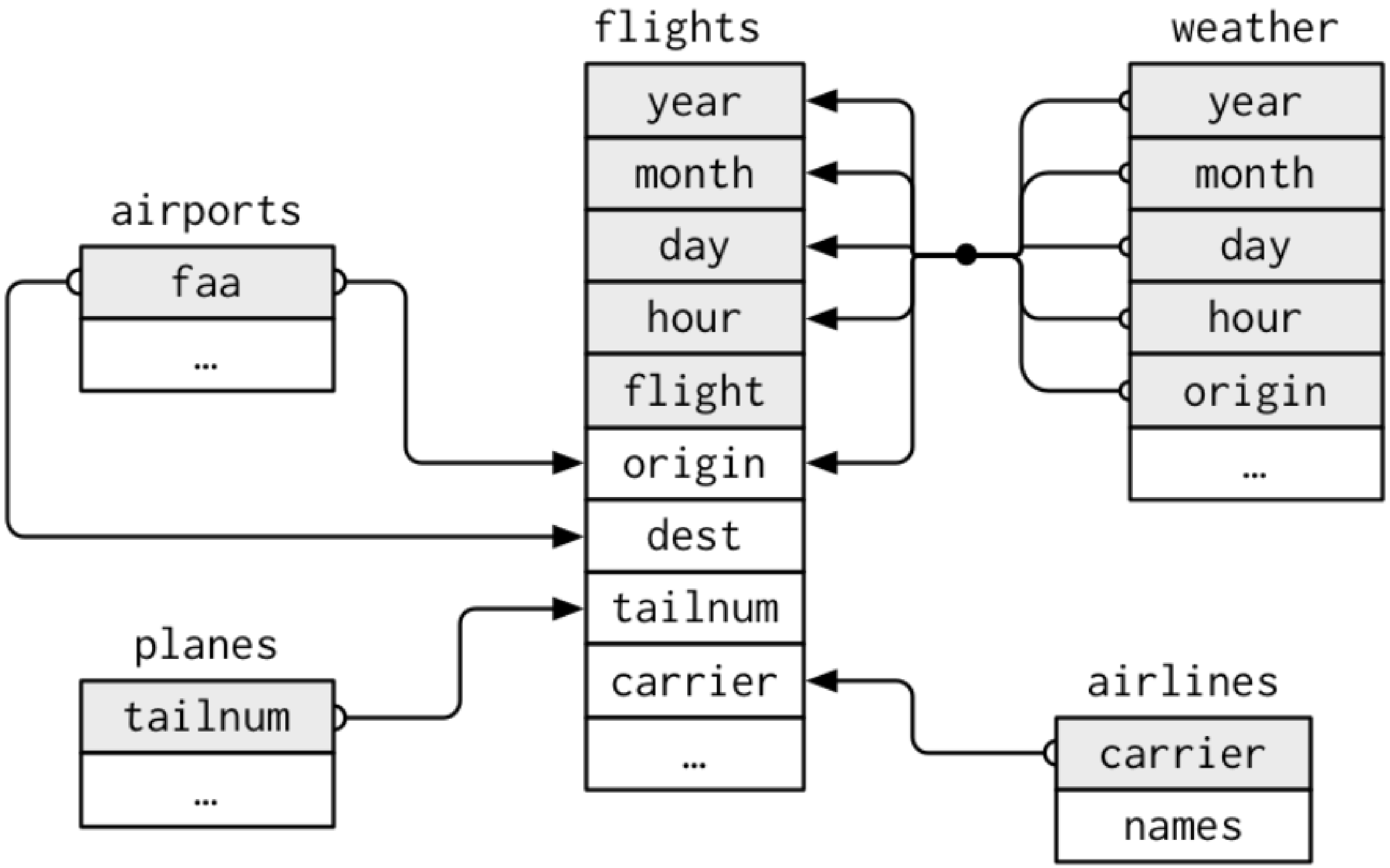
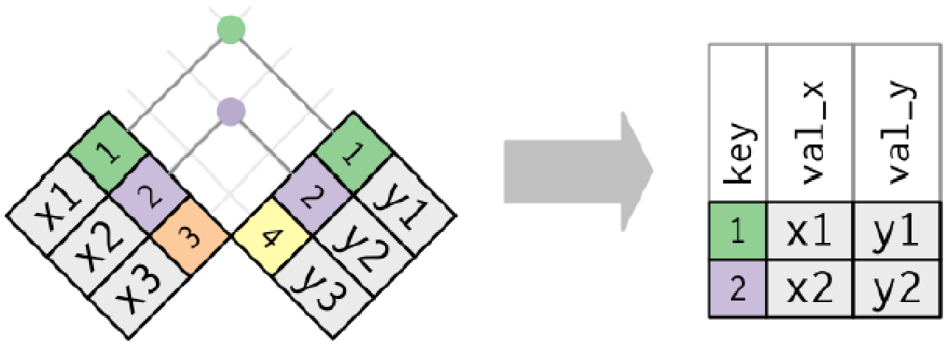


Image: Hadley Wickham R for Data Science Chapter 10

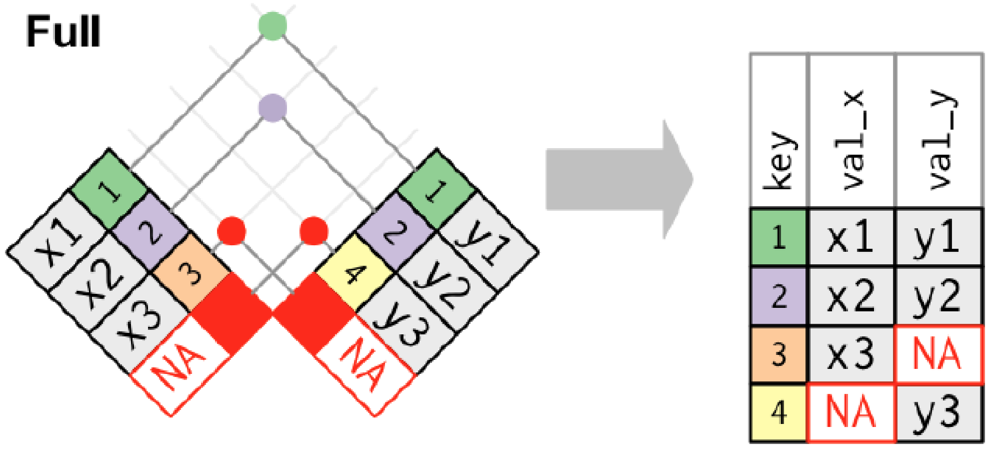
Types of join

- Inner join: Matches pairs of observations whenever their keys are equal
- Unmatched rows are not included



key	val_x	val_y
1	x1	y1
2	x2	y2

- Outer joins:
- left join keeps all the observations in x
- right join keeps all the observations in y
- full join keeps all the observations in a and y


Full

key	val_x	val_y
1	x1	y1
2	x2	y2
3	x3	NA
4	NA	y3

Image: Hadley Wickham R for Data Science Chapter 10

Worksheet

- 07_practice_worksheet.R