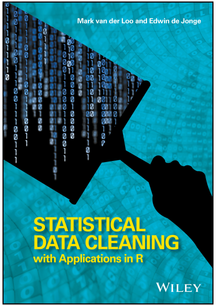
Reading Dirty Data


Mark van der Loo and Edwin de Jonge
STATISTICAL DATA CLEANING
with Applications in R
WILEY

Mark van der Loo and Edwin de Jonge

Statistics Netherlands Research & Development
Twitter: `@markvdloo` `@edwindjonge`

Tokyo | Institute of Statistical Mathemathics | 2020

大学共同利用機関法人 情報・システム研究機構
統計数理研究所
The Institute of Statistical Mathematics

Try the code together with your neighbour

`01raw/json_to_raw.R`

Reading Dirty Data

Reading raw data is about uniforming the technical layout of the data.

We saw the following issues in the example:

- Storage formats
- Data types
  - Locales. . .
  - Unit suffixes.
- Non-uniform formatting

Storage formats

Your raw data is often stored in external formats:

- A database (dump).
- Scraped data in html format.
- Data retrieved from an api (JSON/XML).
- csv / txt/ spss / stata / sas / etc.

Action:

Turn the data into a tabular format

Data types

Most important reading action is setting data types:

- numeric columns often contain units (“42%”, “42 $”, “42 km”, etc.)
- numeric columns often contain locale dependent formatting: “4.2” vs “4,2” or “4.200,42” vs “4,200.42”
- date columns (not in example): “4/2/1942” vs “42/2/4” vs “1942-4-2” etc.
- encoding of strings (for non english): “forty two” vs “tweeënveertig”
- footnotes/annotations[1] in values may corrupt data type: 42ˆ

Cleaning action

Fix the columns: remove non-relevant info and set data type

[1] Footnotes can be essential to data interpretation, but desastrous for technical parsing.

Uniforming values

Some text variables are structured, but contain variations:

- phone numbers: “+311234567890” vs “0031-12-34567890” vs “012-34567890”
- soc. sec. numbers: “1234 5678” vs “123456789”
- bank account numbers, etc.
- zip codes: e.g Dutch zip codes: “1234AB” vs “1234 AB” vs “1234 ab”.

Cleaning action

Reformat values into standard format

String manipulations

Replace text in columns

```r
library(stringr)
str_replace(c("3,4", "4,5"), ",", ".") # or use gsub
```

Change casing of characters

```r
str_to_lower("HUGE")
```

Tip: use readr::parse_* functions

```r
library(readr)
parse_number(c("3,4", "5,6"), locale = locale(decimal_mark = ","))
```

```
## [1] 3.4 5.6
```

```r
parse_date(c("3-4-25", "1-12-78"), format = "%d-%m-%y")
```

```
## [1] "2025-04-03" "1978-12-01"
```

Encoding

```r
bands <- c("Motörhead", "Iron Maiden")
Encoding(bands)
```

```
## [1] "unknown" "unknown"
```

```r
Encoding(bands) <- "latin1"
print(bands)
```

```
## [1] "MotÃ¶rhead"  "Iron Maiden"
```

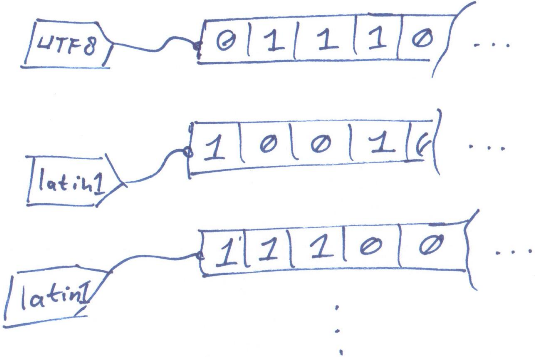
```r
Encoding(bands) <- "UTF-8"
print(bands)
```

```
## [1] "Motörhead"   "Iron Maiden"
```

Character vector x

Encoding(X) <- "UTF-8"

only changes the labels, not the data



Use iconv() to change the encoding.

Assigment

Read in the “01raw/backbone.xml” file and **reformat the zip code** into 4 digits, no space, two uppercase letters. “1234 ab” -> “1234AB”

```
library(XML)
backbone_xml <- XML::xmlParse("01raw/backbone.xml")
backbone <- xmlToDataFrame(backbone_xml)

# Fix zip code
...

# and save the result
write.csv( backbone, "01raw/my_backbone.csv"
         , row.names = FALSE, na="")
```