Cleaning Datasets for Analysis

Contents

A brief explanation of messy data

"Messy data" are data that do not conform to the expected formats required to be read by a machine, or have sensible but inconsistent formats that complicate machine reading. For example, you might recognize "2022-02-03" and "2/3/2022" as two different date formats for February 3, 2022. But a computer program would need explicit instruction on date formats to understand that these two character strings share the same interpretation.

Reproducible data cleaning workflows

Scripted data cleaning workflows allow you to document how you approach data cleaning in a reproducible and easy-to-share manner. Data cleaning scripts are self-documenting; if you save your R scripts for data cleaning then you have a record of your approach, the decisons you made, and a means of reproducing your cleaning steps. While it may seem easier to clean messy data manually in a spreadsheet application, the reproducibility and transparency benefit of scripted workflows makes the scripted approach the widely preferred method for data analytics.

Load data from file

Name	Birthday
Harry Potter	31st July, 1980
Ronald Weasley	1st March, 1980
Hermione Granger	19 September, 1979
Neville Longbottom	30th July, 1980
Ginny Weasley	11th August, 1981
Luna Lovegood	13th February, 1981
Fred and George Weasley	1st April, 1978
James Potter	27th March, 1960
Lily Potter	30th January, 1960
Sirius Black	3rd November, 1959
Remus Lupin	10th March, 1960
Minerva McGonagall	4th October
Rubeus Hagrid	6th December
Severus Snape	9th January, 1960
Voldermort	31st Decmeber

Comma separated values (`.csv`)

Excel spreadsheet (`.xlsx`)

Single tab

Multiple tabs

Common messes and their fixes

Missing values

Interpolate

Treat as zero

Ignore

Inconsistent types

Detecting type

Coercing type

Inconsistent formats

Detecting type

Correcting type

Typos

Off-by-one errors

Extra white space