How to CLEAN your data? Outliers and missing values

Last updated: November 13, 2017

Basics

- An outlier is an extreme value (much lower or higher than the rest)
- A missing value is a missing entry in your data-table (in R, it is denoted NA)

Why are outliers bad? Why are NAs bad?

What to do with NAs?

Here are some common approaches.

What are the advantages and disadvantages of each?

- Contact whoever provided you with the data and ask them to fill it out
- Interpolate (ie: replace by some guessed value)
- Exclude all observations with NAs in the entire dataset
- Exclude all observations with NAs in specific calculations

Examples: mammals sleep; emails; marioKart

How to detect outliers?

Visual guide:

- Boxplot
- Dotplot / Density / Histogram
- Scatterplot

Quantitative guide: more than 3 standard deviation away from the mean (in either direction).

What to do with outliers?

Here are some common approaches.

What are the advantages and disadvantages of each?

- Contact whoever provided you with the data and ask them why this value is so extreme
- Exclude this observation from analysis
- Leave it alone

Examples: teacher (fte and salary); emails; marioKart