Data Cleaning

UC Davis Statistics Club

Created by:
Christopher Wong
Academic Director

Introduction to Data Cleaning

- Data cleaning is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database.
- Data is not always clean. This is the case in today's world where massive amounts of data are being collected and they are not necessarily formatted nicely.
- In fact, in practice it is often the case that a data analyst will spend more time cleaning their data than performing statistical analysis.
- A great resource

Reading Data

- The first step to cleaning data is to read it into a program such as R
- We will use the following functions to read in our data:
  - read.table Tabular data stored in textual format
  - read.csv Comma separated values with period as decimal separator
  - read.csv2 Semicolon separated values with comma as decimal separator
  - read.delim Tab-delimited files with period as decimal separator
  - read.delim2 Tab-delimited files with comma as decimal separator
  - readLines Reads in the text line-by-line
  - read.xlsx or read_excel Reads in Excel files. Type library(xlsx) (requires packages "rJava" and "xlsxjars") or the "readxl" package.
- Arguments for each: header, col.names, na.strings, colClasses, stringsAsFactors

Inspecting the Data

- nrow
- ncol
- length
- head
- tail
- dim
- class
- which
- unique
- names
- range
- table
- summary
- typeof
- str

Type conversion

- Modes: numeric, complex, character, logical, list, function
- Class: How generic functions operate with it, usually same as mode if not assigned
- Typeof: The type of C structure that is used to store a basic type
- Coercion in R:
  - as.numeric
  - as.integer
  - as.character
  - as.logical
  - as.factor
- Recoding factors

String / character manipulation

- String normalization
- grep, grepl, gsub, strsplit, Regular expressions
- str_trim from the library “stringr”

Detection of errors

- Missing values
- Special values
- Outliers
- Inconsistencies

Before we get started

- Go to the Statistics Club Google Drive → Case Studies and Programming Workshops
- Download Data Cleaning Back-up Functions
- Data Cleaning Examples → Download them all
- Also install the packages "readxl" and "stringr"