W03-1: Cleaning a data set

This worksheet will foster your competence in manipulating strings and coercing data types using R. After completing this worksheet you should feel confident that you can clean columns of a data frame to get a technically correct data set.

Things you need for this worksheet

- R — the interpreter can be installed on any operation system. For Linux, you should use the r-cran packages supplied for your Linux distribution. If you use Ubuntu, this is one of many starting points. If you use windows, you could install R from the official CRAN web page.
- R Studio — we recommend to use R Studio for (interactive) programming with R. You can download R Studio from the official web page.
- your data sets from W01-1: Getting and organizing data

Learning log assignments

As always, please add these entries to your today's learning log at teachwiki:

- Favorite aspect of the session (if any)
- Superfluous aspect of the session (if any)
- Eureka effect (if any)
- Links to what I've learned so far (if any)
- Questions (if any)

For more information see this short howto.

As today's special, please complete the following assignment:

The Ebola chronology information from the CDC had to be copied directly from the html source in W01-1 and you likely get away with two files - one for the chronology from 1976 to 2014 and one covering just the cases in 2014. Today, we want to clean and combine the data sets. Since this is just the second worksheet dealing with R, we will again provide some guiding lines of the workflow.

We assume that you have two data sets, one covering the Ebola chronology between 1976 and 2014 and one covering only 2014.

Please write an R script for the following tasks and upload it to your learning log:

1. Read the Ebola chronology from 1976 to 2014 into a data frame
2. Rename the column headers to meaningful names
3. Remove the following rows:
   1. empty rows (tip: look for blank cells)
   2. the row with information about “March 2014-Present” since this will be provided by the second data set
4. Clear the column with information about the year. It is sufficient to define an outbreak just by the

starting year so there will be only one year date per cell left. Watch out for the event in 2008 (and just handle it separately).

5. Clean the column on total/relative death
    1. separate total from relative death and store the latter in a new column (tip: you can not do this in one step but you have to copy the relative death to the new column first before removing them in the original column)
    2. clean both columns so the cells will only contain digits
    3. convert both columns to numeric
6. Clean column with information on infected persons and convert it to numeric
7. Read the Ebola information on 2014 from the second file into a data frame
8. Remove empty lines and the ones containing “Total” as country name in this data frame
9. Create a new data frame holding the information of the second and having the same column number and names
10. Combine both data sets (chronology from 1976 - 2014 and details from 2014)
11. Save the combined data set as csv file with years in increasing order.

For most of the tasks above, the substr function is your way to go and you can use the regexpr to define a starting or end position for your character manipulations.

Applying an operation to selected cells

If you just want to apply something to selected columns/rows/cells, you can restrict your operations to these cells by subsetting the data frame on both sides of the “←” (e.g. to restrict something to cells containing a “*”, use:

```
var$col[grepl('\\*', var$col] ← substr(var$col[grepl('\\*', var$col], start,
end)
```

The “\\” is necessary since “*” is the multiplication operator. In order to use it as string, you have to use so called “escape characters” which is “\\” in R.

Getting into gear

As a means of assistance, the following code snippet shows you one possible solution for tasks 1 to 4.

"snippet 01"

```
ebola <- read.table("cdc_ebola_chronology_1976-2014.csv",
                    header = TRUE, sep = ",")

# new column header
colnames(ebola) <- c("Year", "Country", "Subtype", "Infected", "Death",
                     "Situation")

# Remove empty rows and first row since this will be replaced later
ebola <- ebola[(ebola$Year != "" & !grepl("March 2014-Present",
ebola$Year)),]

# Clear column Year and define outbreack by start year
# Special procedure for cell with 01.11.08
ebola$Year <- as.character(ebola$Year)
ebola$Year[ebola$Year == "01.11.08"] = "2008"
ebola$Year <- substring(
  ebola$Year,
  regexpr(pattern="[[:digit:]]", ebola$Year),
  regexpr(pattern="[[:digit:]]", ebola$Year) + 3)
```