W04-1: Summarizing a data set

This worksheet will train you in applying some quick summarizing functions to get a first overview of a data set. After completing this worksheet you should be able to compute simple aggregation statistics and handle date values.

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 W02-1: CSV I/O and some simple modifications and W03-1: Cleaning a data set

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:

As part of W02-1 and W03-1 you have already modified two data sets on the chronology of natural disasters and Ebola. Today, we will combine both data sets an try to quickly retrieve some information on how serious the consequences of Ebola are compared to other natural disasters.

Since this is already your third contact with R, the following tasks will be less detailed than in the previous worksheets.

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

1. Prepare a combined data frame from the natural disaster and the Ebola data set (see code snippet below for help on changing countries to common names). The combined data set should have the columns listed below. The disaster.* type categories should be set to “Ebola” if the respective row is taken from the Ebola data set.
    - year
    - disaster.group
    - disaster.subgroup
    - disaster.type

    - disaster.subtype
    - country
    - deaths
    - total_affected (i.e. sum of deaths and non-deaths but affected/infected persons; this might be something different then the column named total_affected in the natural disaster data set)
2. In the combined data set, delete all events of category disaster.subtype with ID “Viral Infectious Diseases” if they fall into an Ebola year and Ebola affected country (see tip below)
3. Print the following to the command line and replace <N>, <NN> etc. by the actual numbers (i.e. compute them first):
    1. “On average, each Ebola outbreak leads to the deaths of <N> persons as compared to <NN> for other natural disaster events in the countries affected by Ebola deaths.”
    2. “On average, each Ebola outbreak directly affects <N> persons as compared to <NN> for other natural disaster events in the countries affected by Ebola ebents.”
    3. “Ebola deaths rank on <N> of <NN> worldwide and on <NNN> of <NNNN> within countries directly affected by Ebola (level: disaster.subgroup).”
    4. “The mean time span between two Ebola years is <N> years”

In addition to uploading your code, please copy the resulting print statements of your code to your learning log.

Matching countries

Not all country names do match between the both data sets which makes it difficult to remove double entries. Code snippet 01 will help you handle it by replacing non-matching names of the Ebola chronology with the ones used in the natural disaster data set. The last command should return nothing, otherwise you have to include another line like the ones already printed.

"Snippet 01"

```
# Modify country names not compatible between the data sets
# The content of the Ebola data set is stored in data frame ebola.
ebola$country <- as.character(ebola$country)

ebola$country[grepl(
  paste("Zaire (Democratic Republic of the Congo - DRC)",
        "Zaire",
        "Democratic Republic of the Congo \\(formerly Zaire\\)",
        "Democratic Republic of Congo",
        "Democratic Republic of the Congo", sep = "|"),
  ebola$country)] <- "Zaire/Congo Dem Rep"

ebola$country[grepl("Republic of Congo",ebola$country)] <- "Congo"

ebola$country[grepl("England",ebola$country)] <- "United Kingdom"

ebola$country[grepl("USA",ebola$country)] <- "United States"

ebola$country[grepl("Sudan \\(South Sudan\\)",ebola$country)] <- "South Sudan"
```

```
ebola$country[grepl("Côte d'Ivoire \\(Ivory Coast\\)",ebola$country)]
<-
  "Cote d'Ivoire"

unique(ebola$country[!(ebola$country %in% natcat$country)])
```

Remove double events
To remove double events, one has to check the “Viral Infectious Diseases” entries in the disaster subtype category of the natural disaster data set and remove the ones listed in the Ebola data. The crux of the matter is that your selection must not include any Ebola year and any Ebola country but the specific combinations of the outbreak year and the affected country.

In general, if you want to select all entries in one vector which are also part of another one, you can use the `%in%` function. However, since you need fixed combinations of years and countries the %in% function would fail for something like

```
natcat$year %in% ebola$year & natcat$country %in% ebola$country
```

which would return all matching years and all matching countries. Therefore you have to define a key first which combines the year and country information and then use this key for the selection. Since this is worksheet 4 not 14, code snippet 02 provides you with the line you need.

"Snippet 02"

```
# The content of the combined data set is stored in data frame comb.
# Define a unique year/country key for the ebola events and use it for
the
# selection of all viral infectious diseases co-occuring to Ebola. By
inverting
# the selection, the new data frame has no double entries.
comb <- comb[!(comb$disaster.subtype == "Viral Infectious Diseases" &
                 paste0(comb$year, comb$country) %in%
                 paste0(ebola$year, ebola$country)),]
```

From:
http://moc.environmentalinformatics-marburg.de/ - **Marburg Open Courseware**

Permanent link:
http://moc.environmentalinformatics-marburg.de/doku.php?id=courses:msc:data-management:worksheets:dm-ws-04-1

Last update: **2014/11/13 08:52**