Exercise 1: Data import

Karsten Donnay

14.06.2021

Contents


```
library(knitr)

## Global options
options(max.print="75")
opts_chunk$set(echo=FALSE,
                 cache=FALSE,
               prompt=FALSE,
               tidy=TRUE,
               comment=NA,
               message=FALSE,
               warning=FALSE)
opts_knit$set(width=75)
rm(list = ls())
```

Guardian Data

In this exercise session we work with a dataset of Guardian news coverage gathered through their API and tweets sent by their official Twitter account last year. In later sessions, we will cover how to compile (and combine) equivalent data ourselves, until then you will work with these sample data also for the assignments, including today's first assignment.

Various ways to import data

For the first few illustrations today we work with our "sampledata", a simplified version of these data containing just a few key variables. The data can be found in the folder "data/" of this exercise.

Here we import the same dataset in 3 common fileformats: an R-data file, a comma separated file, and an

Microsoft excel sheet. This sample data has 3 columns, and about 2.500 rows containing only the links and sections of articles from the Guardian.

R Data

Load with **load()**.

```
rm(list = ls())
load("data/sampledata.Rda")
head(sampledata)
```

```
                                                                                                  id
1          artanddesign/2020/apr/06/andy-warhol-take-a-virtual-tour-around-the-tate-modern-exhibition
2                  artanddesign/2020/apr/06/bathtime-and-black-paint-tracey-emin-posts-lockdown-diary
3          artanddesign/2020/apr/06/how-i-became-the-duke-of-urbino-getty-museum-recreate-masterpiece
4                                       artanddesign/2020/apr/10/peter-saul-donald-trump-in-florida
5                                artanddesign/2020/apr/10/virtual-design-festival-coronavirus-lockdown
6 artanddesign/2020/apr/11/mick-rock-releases-unseen-photographs-of-1970s-rock-royalty-to-support-nhs

1          https://www.theguardian.com/artanddesign/2020/apr/06/andy-warhol-take-a-virtual-tour-around-
2                  https://www.theguardian.com/artanddesign/2020/apr/06/bathtime-and-black-paint-tracey
3          https://www.theguardian.com/artanddesign/2020/apr/06/how-i-became-the-duke-of-urbino-getty-mu
4                                       https://www.theguardian.com/artanddesign/2020/apr/10/peter-sau
5                                https://www.theguardian.com/artanddesign/2020/apr/10/virtual-design-fes
6 https://www.theguardian.com/artanddesign/2020/apr/11/mick-rock-releases-unseen-photographs-of-1970s-ro
     sectionId
1 artanddesign
2 artanddesign
3 artanddesign
4 artanddesign
5 artanddesign
6 artanddesign
```

```
dim(sampledata)
```

```
[1] 2498    3
```

Comma separated files

Use **read.csv()**.

```
sampledata_csv <- read.csv("data/sampledata.csv")
```

Attention: check the dimensions: only 1 column, but the dataset included 4 columns.

```
dim(sampledata_csv)
```

```
[1] 2498    1
```

```
head(sampledata_csv)
```

```
1                   artanddesign/2020/apr/06/andy-warhol-take-a-virtual-tour-around-the-tate-modern-exhi
2                                   artanddesign/2020/apr/06/bathtime-and-black-paint-tracey-emin-posts-
3                   artanddesign/2020/apr/06/how-i-became-the-duke-of-urbino-getty-museum-recreate-maste
4                                                                                 artanddesign/2020/apr/
5                                                                artanddesign/2020/apr/10/virtual-design-fe
6 artanddesign/2020/apr/11/mick-rock-releases-unseen-photographs-of-1970s-rock-royalty-to-support-nhs;ht
```

Inspect it with a text editor of your choice: you will see that values are not separated by commas, but by semicolons. We adjust the parsing from .csv accordingly:

```
sampledata_csv <- read.csv("data/sampledata.csv", sep = ";")
dim(sampledata_csv)
```

```
[1] 2498    3
```

Excel sheets

Install and use the **readxl** package and use the **read_xlsx()** command.

```
sampledata_xls <- readxl::read_xlsx("data/sampledata.xlsx")
```

Basic overiew

To get a basic overview of a dataset, we might use **str()**

```
str(sampledata)
```

```
'data.frame':   2498 obs. of  3 variables:
 $ id       : Factor w/ 3493 levels "australia-news/2020/apr/09/australian-government-experts-at-odds-wi
 $ link     : chr  "https://www.theguardian.com/artanddesign/2020/apr/06/andy-warhol-take-a-virtual-tou
 $ sectionId: Factor w/ 48 levels "australia-news",..: 29 29 29 29 29 29 29 29 29 29 ...
```

As mentioned above, **dim()** provides us with a basic overview of how many rows and columns are included in the dataset.

```
dim(sampledata)
```

```
[1] 2498    3
```

The **table()** command provides us with an easy overview of the distribution of a dichotomous or categorical variable.

```
table(sampledata$sectionId)
```

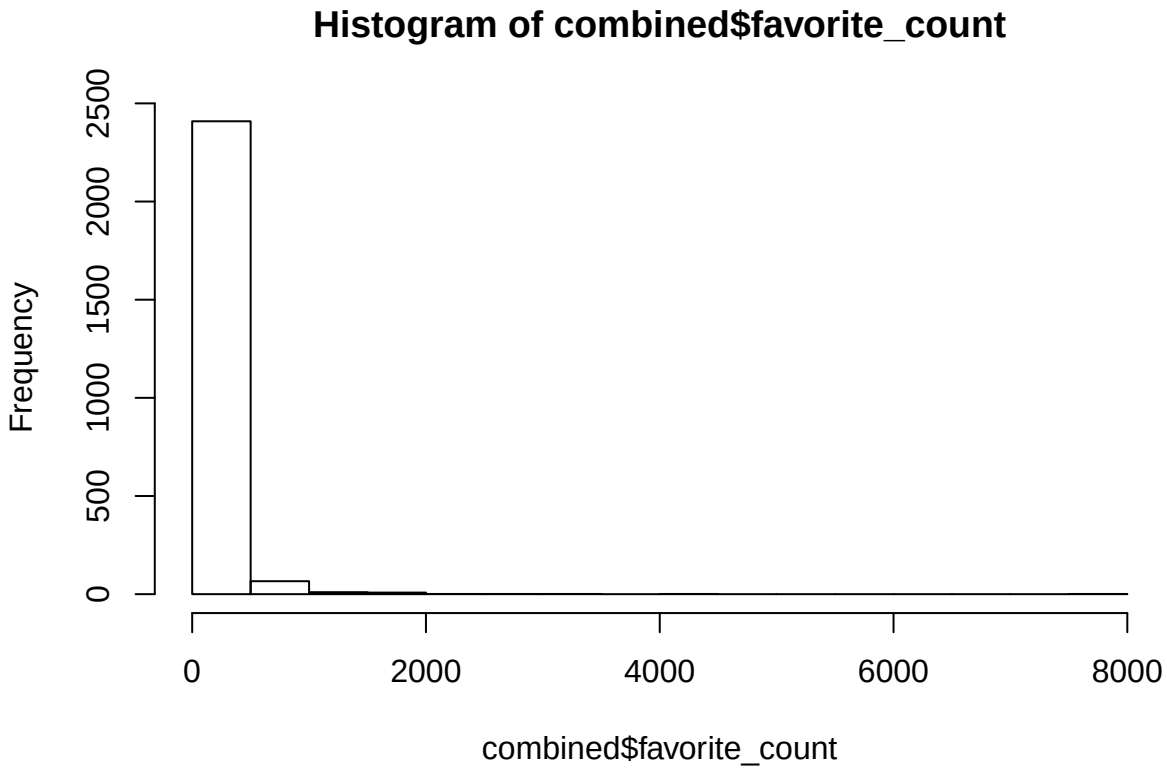
```
australia-news
            87
         world
           717
          film
            34
       society
            90
   environment
            46
global-development
            17
 lifeandstyle
            71
  crosswords
             0
         books
            52
      football
           155
 tv-and-radio
```

```
                                    41
                               politics
                                   127
                               science
                                    21
                                 sport
                                   104
                               business
                                   175
                               us-news
                                   139
                                 music
                                    53
                           commentisfree
                                   213
                              education
                                    37
                               uk-news
                                    84
                                 news
                                    18
                         the-last-taboo
                                     0
                                culture
                                    25
                                 stage
                                    13
                               fashion
                                    13
                            technology
                                    33
                                food
                                    16
                              travel
                                    19
                          artanddesign
                                    21
                               media
                                    25
                               games
                                     4
                           community
                                    19
                               money
                                    20
                 guardian-masterclasses
                                     0
                              global
                                     1
                         animals-farmed
                                     0
                        keep-connected
                                     0
shelter-supporting-those-who-are-struggling
```

```
                                0
                       membership
                                1
                              law
                                4
          global-health-progress
                                0
                          weather
                                0
                            focus
                                1
    guardian-us-press-office
                                0
                       inequality
                                1
                      theobserver
                                1
detectives-transforming-communities
                                0
   property-management-at-firstport
                                0
```

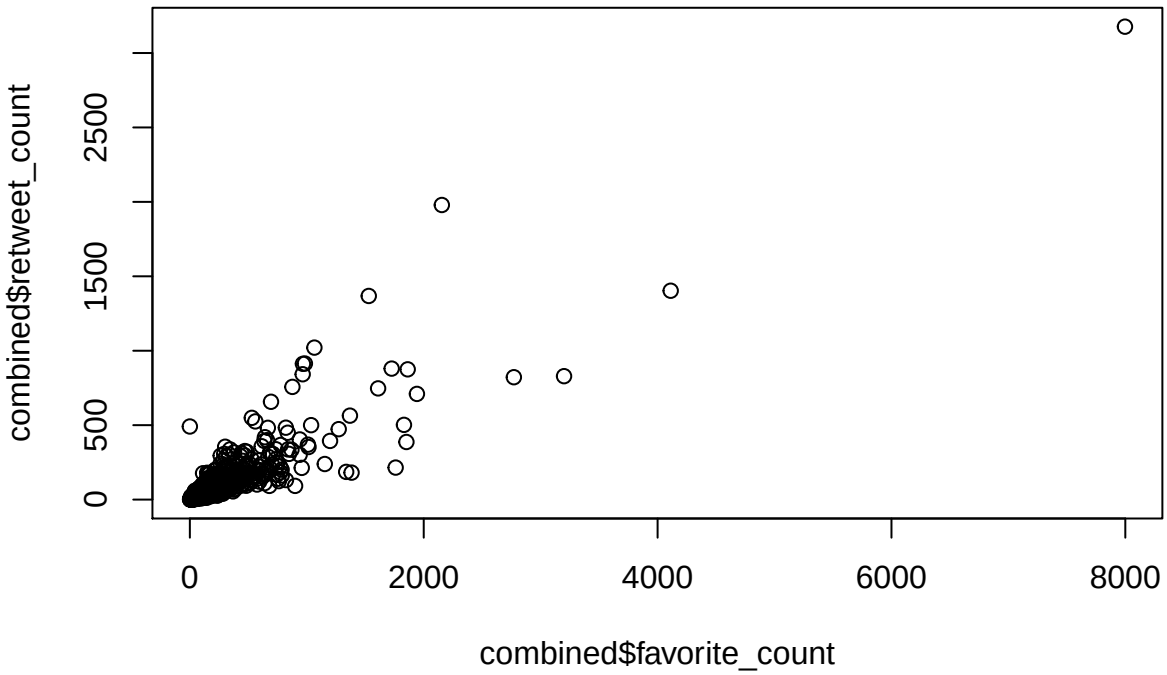
You can use **hist()** to plot a histogram of a numeric variable and get an overview.

```
load("data/combined.Rda")
hist(combined$favorite_count)
```



You can use **plot()** to plot two variables against each other.

```
plot(combined$favorite_count, combined$retweet_count)
```



Rmarkdown

This file is or is produced by a R Markdown file. You will find a detailed introduction to RMarkdown here: https://bookdown.org/yihui/rmarkdown/ and a summary sheet here: https://github.com/rstudio/cheatsheets/raw/master/rmarkdown-2.0.pdf.

In our case, RMarkdown files consist of sections that include text in combination with code "chunks''. Markdown files allow us to combine text with code. Markdown files keep the syntax simple and use comparable syntax to LaTeX and HTML. Often, the same commands that work in LaTeX work with Markdown as well. Basic text formatting is done with the following commands:

- you can make text italic by putting * or _ around it. E.g. *text* looks like *text*
- you can make text bold by putting ** it. E.g. **text** looks like **text**
- if you want to resemble code you need to put ‘ around it. E.g. ‘codetext‘ looks like `codetext`
- Sections are introduced by using #.
- Lists can be done with using a space and either *, -, or + and then again a space.

Additionally, you need the package `knitr` to compile or "knit" a Markdown File to an output format. In RStudio you can then chose if you want a PDF, HTML or Word Document.

Code chunks also allow for options. Code Chunks are introduced by "‘ and also closed like this. After the opening, we need to specify which type of programming language we want to insert (`r`) and then we e.g. can specify if we want to `echo` our code (`TRUE`) or whether we want to omit it (`FALSE`). Other options are e.g. `eval`, which indicates whether we want to evaluate the code in the chunk below or whether we just want to skip it (`eval=FALSE`).

Git and assignment

Next you may set up your own github account and download or clone the github repository accompanying the lecture and this exercise. You will find the assignment of the first exercise in the folder "assignments/assignment1" under the name "assignment1.Rmd" or its HTML and PDF version. You need to complete this assignment by adding the necessary code to the prepared RMarkdown file. Please change the

name to “firstname_lastname_assignment1.Rmd” and upload it in the Dropbox section for “Assignment 1” on **OLAT**. This will also be the submission format for the next assignments as well.