HowTo Get data from the EUROSTAT

A. Blejec

October 23, 2014

Contents

1 Introduction

To make reproducible reports from the EUROSTAT data, I will explore the possibility to download data from their repository. They provide data in different formats, with footnotes and labels included or not included. The most promising for automatic retreival is the so called **TSV** format which is in fact tab separated file. This file is included in the .zip or .gz file. See the readme file.

2 Finding the file name

We will explore it later.

3 Get the data

Asumming that we have the desired file code, we can get it by composing the URL:

```
> fcode <- "tps00001"
> # fcode <- 'ei_bsco_m' fcode <- 'nama_gdp_k'
> lfn <- paste(fcode, ".tsv", sep = "")
> upre <- "http://epp.eurostat.ec.europa.eu/NavTree_prod/everybody/BulkDownloa
> upost <- ".tsv.gz"
> furl <- paste(upre, fcode, upost, sep = "")
> furl
[1] "http://epp.eurostat.ec.europa.eu/NavTree_prod/everybody/BulkDownloadListi


> temp <- tempfile()
> download.file(furl, temp)
> first <- readLines(temp, 1)
> data <- read.table(gzfile(temp), sep = "\t", header = TRUE, row.names = 1,
+     na.strings = ": ")
> unlink(temp)
```

Description

```
> first
[1] "indic_de,geo\\time\t2003 \t2004 \t2005 \t2006 \t2007 \t2008 \t2009 \t2010
> first <- strsplit(first, "\t")[[1]]
> colNames <- first[-1]
> units <- strsplit(first[1], "\\\\")[[1]][1]
> colVar <- strsplit(first[1], "\\\\")[[1]][2]
> units <- strsplit(units, ",")[[1]]
> units
[1] "indic_de" "geo"
> colVar
[1] "time"
> colNames
 [1] "2003 " "2004 " "2005 " "2006 " "2007 " "2008 " "2009 " "2010 "
 [9] "2011 " "2012 " "2013 " "2014 "
> attr(data, "units") <- units
> attr(data, "colVar") <- colVar
> attr(data, "colNames") <- colNames
```

Show the data structure

```
> dim(data)
[1] 43 12
> dimnames(data)
```

```
[[1]]
 [1] "JAN,AL"   "JAN,AT"   "JAN,BA"   "JAN,BE"   "JAN,BG"   "JAN,CH"
 [7] "JAN,CY"   "JAN,CZ"   "JAN,DE"   "JAN,DK"   "JAN,EA17" "JAN,EA18"
[13] "JAN,EE"   "JAN,EL"   "JAN,ES"   "JAN,EU27" "JAN,EU28" "JAN,FI"
[19] "JAN,FR"   "JAN,HR"   "JAN,HU"   "JAN,IE"   "JAN,IS"   "JAN,IT"
[25] "JAN,LI"   "JAN,LT"   "JAN,LU"   "JAN,LV"   "JAN,ME"   "JAN,MK"
[31] "JAN,MT"   "JAN,NL"   "JAN,NO"   "JAN,PL"   "JAN,PT"   "JAN,RO"
[37] "JAN,RS"   "JAN,SE"   "JAN,SI"   "JAN,SK"   "JAN,TR"   "JAN,UK"
[43] "JAN,XK"

[[2]]
 [1] "X2003" "X2004" "X2005" "X2006" "X2007" "X2008" "X2009" "X2010"
 [9] "X2011" "X2012" "X2013" "X2014"
```

Number of rows and columns for later use

```
> nRows <- dim(data)[1]
> nCols <- dim(data)[2]
```

Structure of the first few

```
> str(data[, 1:min(5, nCols)])
'data.frame':        43 obs. of  5 variables:
 $ X2003: Factor w/ 43 levels "10142362 ","10192649 ",..: 14 41 20 3 40 38 37
 $ X2004: Factor w/ 43 levels "10116742 ","10195347 ",..: 14 41 20 3 40 38 37
 $ X2005: Factor w/ 43 levels "10097549 ","10198855 ",..: 14 41 20 3 40 38 37
 $ X2006: num  3149143 8254298 3842650 10511382 7718750 ...
 $ X2007: Factor w/ 43 levels "10066158 ","10254233 ",..: 14 42 20 4 39 38 40
```

First few lines

```
> head(data[, 1:min(5, nCols)])
           X2003     X2004     X2005    X2006     X2007
JAN,AL  3102781   3119548   3134975   3149143  3152625
JAN,AT  8100273   8142573   8201359   8254298  8282984
JAN,BA 3830349 e 3837414 e 3842532 e  3842650  3844017
JAN,BE 10355844  10396421  10445852  10511382 10584534
JAN,BG  7845841   7801273   7761049   7718750 7572673 b
JAN,CH  7313853   7364148   7415102   7459128  7508739
```

4 Clean the data

Some data contain labels and can be suffixed with a space. So we have to clean the data.

```
> X <- data
> X <- as.data.frame(apply(X, 2, function(x) as.numeric(sapply(x, FUN = functi
+     "", u)))))
> dimnames(X)[[1]] <- dimnames(data)[[1]]
```

Row names are structured, Composed as `unit,indicator,geo`. The structure is

```
> getUnit <- function(x = data, id = 1) {
+     attr(x, "units")[id]
+ }
> getUnit(data, 1)
[1] "indic_de"
```

Show the structure

```
> str(X[, 1:min(5, nCols)])
'data.frame':        43 obs. of  5 variables:
 $ X2003: num  3102781 8100273 3830349 10355844 7845841 ...
 $ X2004: num  3119548 8142573 3837414 10396421 7801273 ...
 $ X2005: num  3134975 8201359 3842532 10445852 7761049 ...
 $ X2006: num  3149143 8254298 3842650 10511382 7718750 ...
 $ X2007: num  3152625 8282984 3844017 10584534 7572673 ...
```

and first few lines

```
> head(X[, 1:min(5, nCols)])
          X2003    X2004    X2005    X2006    X2007
JAN,AL  3102781  3119548  3134975  3149143  3152625
JAN,AT  8100273  8142573  8201359  8254298  8282984
JAN,BA  3830349  3837414  3842532  3842650  3844017
JAN,BE 10355844 10396421 10445852 10511382 10584534
JAN,BG  7845841  7801273  7761049  7718750  7572673
JAN,CH  7313853  7364148  7415102  7459128  7508739
```

SessionInfo

Windows 7 x64 (build 7601) Service Pack 1

- R version 3.0.2 (2013-09-25), `x86_64-w64-mingw32`
- Locale: `LC_COLLATE=Slovenian_Slovenia.1250`, `LC_CTYPE=Slovenian_Slovenia.1250`, `LC_MONETARY=Slovenian_Slovenia.1250`, `LC_NUMERIC=C`, `LC_TIME=Slovenian_Slovenia.1250`
- Base packages: base, datasets, graphics, grDevices, stats, utils
- Other packages: knitr 1.6
- Loaded via a namespace (and not attached): evaluate 0.5.5, formatR 0.10, stringr 0.6.2, tools 3.0.2

Project path: `D:/_Y/R/Rome`
Main file : `../doc/getEurostat-knitr.Rnw`

View as vignette

Project files can be viewed by pasting this code to R console:

```
> projectName <-"Rome";  mainFile <-"getEurostat-knitr"


> commandArgs()
> library(tkWidgets)
> openPDF(file.path(dirname(getwd()),"doc",
> paste(mainFile,"PDF",sep=".")))
> viewVignette("viewVignette", projectName, #
> file.path("../doc",paste(mainFile,"Rnw",sep=".")))
> #
```