PBI

HOW TO WORK WITH PISA IN R

WE WILL SEE

PUBLISHED BY WE WILL SEE

TUFTE-LATEX.GOOGLECODE.COM

First printing, July 2013

Contents

Introduction to PISA

Here there should be an introduction to the PISA data. Who is doing this study and why.

Few words about data structure and availability [1] [2].

[1] The programme for international student assessment (pisa). http://www.oecd.org/pisa/

[2] The organisation for economic co-operation and development (oecd). http://www.oecd.org/

Get your data

In order to work with PISA data in R you need to load the data first. There are at least two way how to do this.

You can download raw data from PISA website [3]. The raw data is available as compressed text files and you can read these files with the `read.fwf()` function. [4]

The second, much easier, approach is to install R package that already consists required data. There are two sets of packages that you may be interested in. Packages with PISA data and packages with supplementary functions that makes it easier to analyse this data set.

[3] The programme for international student assessment (pisa). http://www.oecd.org/pisa/

[4] The `read.fwf()` function is a standard way to read text files in the fixed width format.

R packages with PISA data

Right now there are There are five packages with PISA data. Each package contains data from single PISA study. These packages have following names: `PISA2000lite`, `PISA2003lite`, `PISA2006lite`, `PISA2009lite`, `PISA2012lite`.

Installation of R package requires the download first. Since the datasets are large be prepared to download about 200MB from Internet. But you need to do this only once. per dataset.

In order to install any of these data packages you will need the `devtools` package. In the chapter you will find more details how to install that one.

Suppose that you have the `devtools` package. Than to get data from study PISA 2009 you need to run following commands.

```
library(devtools)
install_github("PISA2009lite", "pbiecek")
```

As a result you shall see an output like that:Depending on your Internet bandwidth it may take a while.

```
Installing github repo(s) PISA2009lite/master from pbiecek
Downloading PISA2009lite.zip from https://github.com/pbiecek/PISA2009lite/archive/master.zip
Installing package from /var/folders/g3/j8pnss9j3130g4nhj31wxxm0000103/T//RtmptdZ54R/PISA2009lite.zip
Installing PISA2009lite
'/Library/Frameworks/R.framework/Resources/bin/R' --vanilla CMD INSTALL  \
  '/private/var/folders/g3/j8pnss9j3130g4nhj31wxxm0000103/T/RtmptdZ54R/PISA2009lite-master'  \
  --library='/Library/Frameworks/R.framework/Versions/3.0/Resources/library'  \
  --with-keep.source --install-tests

* installing *source* package 'PISA2009lite' ...
** data
*** moving datasets to lazyload DB
** demo
** help
*** installing help indices
** building package indices
** testing if installed package can be loaded
* DONE (PISA2009lite)
```

If there is not `ERROR` in your output it looks like everything went smoothly. The package is installed. In order to work with it you need to load it. Use the `library()` function for that.

```
library(PISA2009lite)
```

You will find five data sets in this package [actually ten, I will explain this later]. These are: data from student questionnaire, school questionnaire, parent questionnaire, cognitive items and scored cognitive items.

```
dim(student2009)
## [1] 515958    437
dim(parent2009)
## [1] 106287     90
dim(school2009)
## [1] 18641   247
dim(item2009)
## [1] 515958    273
```

```
dim(scoredItem2009)
## [1] 515958    227
```

For most of variables in each data set there is a dictionary which decode answers for particular question. Dictionaries for all questions for a given data set are stored as a list of named vectors, these lists are named after corresponding data sets [just add suffix 'dict'].

For example fist six entries in a dictionary for variable CNT in the data set `student2009`.

```
head(student2009dict\$CNT)
##          ALB          ARG          AUS          AUT          AZE
##    "Albania"  "Argentina"  "Australia"    "Austria" "Azerbaijan"
##          BEL
##    "Belgium"
```

Selecting a subset of countries

In some cases you would not work on whole datasets, but only on some subset of countries. You can do this by subseting the dataset. For example, let's take only three countries out of the dataset

```
student2009selected <- subset(student2009, CNT %in% c("ITA", "FRA", "POL"))
dim(student2009selected)
## [1] 40120   437
```

R packages with supplementary functions

To make it easier to work with PISA data you may use the package `PISAtools`. The installation is similar to the installation of dataset.

```
library(devtools)
install_github("PISAtools", "pbiecek")
library(PISAtools)
```

And the package is ready to use. In next chapters we will show some useful functions that are available there.

Additional datasets

In the `PISAtools` package you will find some additional dataset that might be helpful when working with PISA data. Let's introduce them one by one.

dataset: countryOntology

This ontology was derived from FAO website [5]. It contains information about 211 countries. It may be usefull to verify to which group a given country belongs. Let's see columns in this dataset.

[5] Food and agriculture organization of the united nations (fao). http://www.fao.org/home/en/

```
head(countryOntology,2)
```

Last column `IS_IN_GROUP` describes to which groups given country belongs. Other columns are just different classifications of particular country.

```
  ISO3 ISO2 UN_CODE UNDP_CODE FAOSTAT_CODE GAUL_CODE FAOTERM_CODE AGROVOC_CODE NAME_EN
1  GRD   GD     308       GRN           86        99        15417         3384 Grenada
2  LBY   LY     434       LIB          124       145        15442         4312   Libya
  LISTNAME_EN
1     Grenada
2       Libya
                                                                          IS_IN_GROUP
1 "FAO_2006,CARICOM_1985,World,CARICOM,CARIFORUM,NFIDC,Americas,Caribbean,FAO,SIDS"
2                       "CEN_SAD,CAEU,World,AMU,northern_Africa,Africa,FAO,COMESA"
```

Data manipulation

Here there should be an information how to reshape data.

Subselect variables, combine student and school datasets and similar things.

Statistical procedures

Here there should be an information how to do some simple statistics with the data.

Like rankings, weighted averages, weighted regression.

Maybe mixed effect model or generalized mixed effect model.

Data Visualisation

Some examples how to create charts with the use of this data.

Introduction to R

Very short introduction to R. With references to other materials.

Bibliography

Food and agriculture organization of the united nations (fao). http://www.fao.org/home/en/.

The organisation for economic co-operation and development (oecd). http://www.oecd.org/.

The programme for international student assessment (pisa). http://www.oecd.org/pisa/.