Week 2 - Class Worksheet

Get: Importing, Scraping and Exporting Data with R

Dr. Anil Dolgun

Last updated: 18 June, 2020

Working in small groups or pairs, complete the following exercises.

Required Packages

The following packages will be required or may come in handy.

```
library(readr)
library(xlsx)
library(readxl)
library(foreign)
library(gdata)
library(rvest)
```

Exercises

1 Display your working directory using `getwd()` command.

2 Create a folder and name it “**Week2**” on your desktop and set this as your working directory. (Hint: You can set your working directory using `setwd()` command.)

Population Data

The following exercises (exercise 3-5) are based on Population density and regional dataset (population.csv) (../data/population.csv) which contains 2245 rows of time series data between 2009 to 2016. This data is taken from OECD located at http://stats.oecd.org/Index.aspx?DataSetCode=REGION_DEMOGR (http://stats.oecd.org/Index.aspx?DataSetCode=REGION_DEMOGR) where you can find more about the data structure. The variables for population.csv dataset consists of:

`Region` [Character]: Region name

`Territory Level and Typology` [Factor]: Description of the region (Country, Large regions, Small regions,…)

`2009, 2010, ...,2016` [Numeric]: Population density growth index per referenced time.

3 Import csv version of Population dataset (population.csv (../data/population.csv)) using Base R functions.

4 Repeat exercise 3, this time use `readr` functions.

5 Import spss version of Population data (population.sav (../data/population.sav)) using `foreign` package.

Population - Migration Data

Exercise 6 is based on population-migration dataset (population-migration.xls) (../data/population-migration.xls) which contains two sheets named “Population Density and Regional” and “Inter-regional Migration” in .xls format. “Inter-regional Migration” sheet contains 1564 rows of timeseries data between 2009 to 2016. This data set is also taken from OECD located at http://stats.oecd.org/Index.aspx?DataSetCode=REGION_DEMOGR (http://stats.oecd.org/Index.aspx?DataSetCode=REGION_DEMOGR). The variables for population-migration data (“Inter-regional Migration” sheet) contains:

`Region` [Character]: Region name

`Territory Level and Typology` [Factor]: Description of the region (Country, Large regions, Small regions,…)

`2009, 2010, ...,2016` [Numeric]: Migration (All persons inflows minus outflows)

6 Import population-migration dataset (population-migration.xls) (../data/population-migration.xls) using `xlsx` or `readxl` functions. (Hint: Use `sheet` argument.)

Most Popular Baby Names by Sex and Mother’s Ethnic Group, New York City Data

The following exercises (exercise 7-9) is based on Popular Baby Names, NYC dataset which contains 22,035 rows of data recorded in 2011. This data set is taken from NYC Open Data which is located at https://data.cityofnewyork.us/Health/Most-Popular-Baby-Names-by-Sex-and-Mother-s-Ethnic/25th-nujf (https://data.cityofnewyork.us/Health/Most-Popular-Baby-Names-by-Sex-and-Mother-s-Ethnic/25th-nujf) containing the following variables:

`BIRTHYEAR` [Integer]: Year of birth

`GENDER` [Character]: Gender

`ETHNICITY` [Character]: Mother’s ethnicity (categories: HISPANIC, ASIAN AND PACIFIC ISLANDER, WHITE NON HISPANIC, BLACK NON HISPANIC)

`NAME` [Character]: Child first name

`COUNT` [Integer]: Count

`RANK` [Integer]: Rank

The url for the csv file is located at https://data.cityofnewyork.us/api/views/25th-nujf/rows.csv?accessType=DOWNLOAD (https://data.cityofnewyork.us/api/views/25th-nujf/rows.csv?accessType=DOWNLOAD)

7 Use the url for Popular Baby Names, NYC dataset to import in R.

8 After you import the dataset save it as a .csv file.

9 Repeat exercise 8, this time save it as .Rdata.

List of Cities in Australia by Population Data Table

The following exercise is based on the html data table taken from Wikipedia located at https://en.wikipedia.org/wiki/List_of_cities_in_Australia_by_population (https://en.wikipedia.org/wiki/List_of_cities_in_Australia_by_population) containing following variables:

`RANK` [Integer]: Rank

`GCCSA/SUA` [Character]: Greater Capital City Statistical Areas/Significant Urban Areas

`State/Territory` [Character]: State/Territory names

`June 2016[2] population` [Numeric]: Population

`Percentage of national population` [Character]: Ratio of the population as a percentage

10 Using the html link for Cities in Australia by Population, import the data table and save it to your working directory which you set up previously.

The following exercise is based on txt version of Cities in Australia dataset (aus.txt) (../data/aus.txt).

11 Import txt version of Cities in Australia by Population data set aus.txt (../data/aus.txt) using `readr` functions.

Finished?

If you have finished the above tasks, work through the weekly list of tasks posted on the Canvas announcement page.

Return to Course Website (../index.html)