MITx: 14.310x Data Analysis for Social Scientists

Hel

Bookmarks

Module 3: Gathering and Collecting Data, Ethics, and Kernel Density Estimates > Module 3: Homework > Questions 1 - 3

Bookmark

In this problem set, we will guide you through different ways of accessing real data sets and how to summarize and describe them properly. First, we will go through some of the data that is collected by the World Bank. We will do some cleaning on the data before we start analyzing it. Then, we will try to do a simple web scrapping exercise where we will analyze the data as well.

Let’s start with the data sets of the World Bank. Please complete the following steps:

1. Go to the World Bank Datasets website: http://data.worldbank.org/.
2. Once you are there, use the data catalogue to find the Gender Stats data and download the file in csv format.
3. Save the file in your computer in a folder where you can get it easily. In my case, I have saved the files in the following directory: *"/Users/raz/Dropbox/14.31 edX Building the Course/Problem Sets/PSET 3/Gender_Stats_csv"*.

NOTE: It is important to work in the same directory that the files are or to use the whole path when you specify opening a data set. To know in which directory you are currently working, you can use the command getwd(). Similarly, in order to set a different directory, you can use the command setwd().

Analyzing the Data:

Finger Exercises due Oct 17, 2016 at 05:00 IST

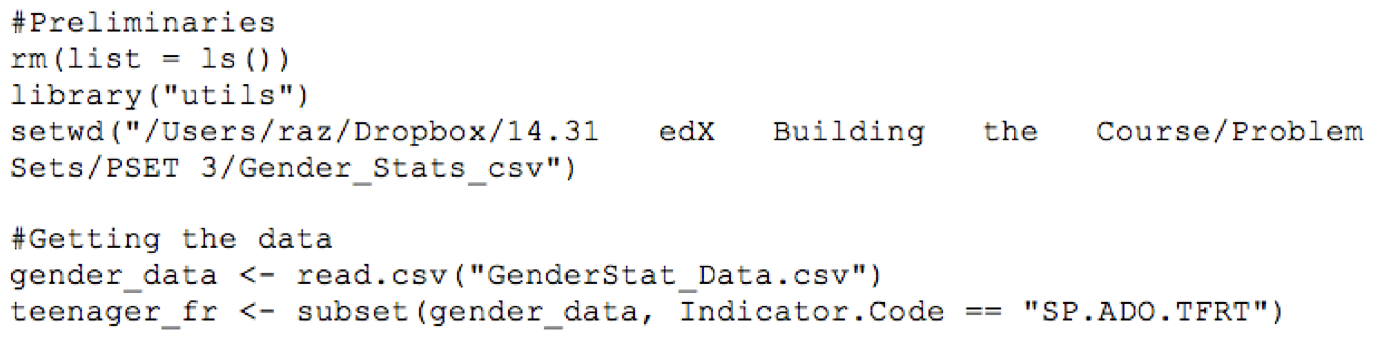
Module 3: Homework
Homework due Oct 10, 2016 at 05:00 IST

Exit Survey

For the purpose of analyzing the data, we are going to use the package “utils”. Once you have uploaded the data to R, you are going to see that there are multiple indicators of gender, countries and years in the data. In this case, we are just interested in analyzing the data for one indicator that is the *Adolescent Fertility Rate.*

The *Adolescent Fertility Rate* measures the annual number of births to women 15 to 19 years of age per 1,000 women in that age group. It represents the risk of childbearing among adolescent women 15 to 19 years of age. It is also referred to as the age-specific fertility rate for women aged 15-19. Once you have completed this problem set, you’ll have more information on how this rate has evolved over time and how it varies across different groups of countries.

Take a look at the following lines of code, whose main purpose is to upload the data in a data frame and to choose the proper indicator. Please, try to understand the code and then run it in your computer. Remember to set the directory accordingly to the folder where you saved the files.



```
#Preliminaries
rm(list = ls())
library("utils")
setwd("/Users/raz/Dropbox/14.31  edX  Building  the  Course/Problem
Sets/PSET 3/Gender_Stats_csv")

#Getting the data
gender_data <- read.csv("GenderStat_Data.csv")
teenager_fr <- subset(gender_data, Indicator.Code == "SP.ADO.TFRT")
```

Using the code, answer the following questions:

Question 1

(1/1 point)

What is the purpose of the line **rm(list = ls())**?

● a. To remove all the current existing objects in R ✓

○ b. To change the current directory path

○ c. To list all the files in the current directory3

○ d. To look in the web for the World Bank dataset.

EXPLANATION

You should be able to look for the help file of the command rm(). In particular you should run help("rm") in the command window of R-studio. The main argument for rm() correspond to the objects that you want to remove from your current session in R. In this case we can also specify a list, by using inside the command rm(), the command ls() –which lists all the files-, we are able to remove all the current existing objects in our R session.

You have used 1 of 2 submissions

Question 2

(1/1 point)

What part of the code specifies the creation of a data frame that contains only the relevant information for the analysis that we are interested?

○ a. setwd("/Users/raz/Dropbox/14.31 edX Building the Course/Problem Sets/PSET 3/Gender_Stats_csv")

○ b. gender_data <- read.csv("GenderStat_Data.csv")

● c. teenager_fr <- subset(gender_data, Indicator.Code == "SP.ADO.TFRT") ✓

○ d. library("utils")

EXPLANATION

The command subset is used to select part of the data frame. The first argument corresponds to the original data frame we are interested extracting information from. The second argument corresponds to the condition that the new data should satisfy. In our case it is that the variable **Indicator.Code** must be equal to **"SP.ADO.TFRT"**. This is the indicator code for the Adolescent fertility rate. to

You have used 1 of 2 submissions

Question 3

(1/1 point)

If you ran the provided code in R, you should have found that now you have two different data frames in R. One named gender_data with 180,601 observations and 62 variables. The other one named teenager_fr with 263 observations and 62 variables. This seems to be kind of inefficient since we won't use the first data frame and it is rather large.

If you were interested in removing this object, what would be the command that would allow you to do this?

rm(gender_data) ✓ **Answer:** rm(gender_data) **or** remove(gender_data)

EXPLANATION

As it was previously specified the command in R that allows you to remove objects is called rm(). Here we just specify that we want to remove the data frame that we called gender_data.

You have used 2 of 2 submissions