Data Science: Wrangling - Assessments

Hector Antigua

10/27/2021

## Assessment Part 1: Data Import

In this part of the assessment, you will answer several multiple choice questions that review the concepts of data import. You can answer these questions without using R, although you may find it helpful to experiment with commands in your console.

In the second part of the assessment on the next page, you will import real datasets and learn more about useful arguments to readr functions. The second part of the assessment will require you to program in R.

### **Question 1**

Which of the following is NOT part of the data wrangling process?

**Respuesta:** *Checking correlations between your variables*

### **Question 2**

Which files could be opened in a basic text editor?

**Respuesta:** *data.txt* *data.csv* *data.tsv*

### **Question 3**

You want to analyze a file containing race finish times for a recent marathon. You open the file in a basic text editor and see lines that look like the following:

initials,state,age,time  
vib,MA,61,6:01  
adc,TX,45,5:45  
kme,CT,50,4:19

**Respuesta:** *A comma-delimited file with a header*

### **Question 4**

Assume the following is the full path to the directory that a student wants to use as their working directory in R: “/Users/student/Documents/projects/”

Which of the following lines of code CANNOT set the working directory to the desired “projects” directory?

**Respuesta:** *setwd(/Users/student/Documents/projects/)*

### **Question 5**

We want to copy the “murders.csv” file from the dslabs package into an existing folder “data”, which is located in our HarvardX-Wrangling projects folder. We first enter the code below into our RStudio console.

> getwd()  
[1] "C:/Users/UNIVERSITY/Documents/Analyses/HarvardX-Wrangling"  
> filename <- "murders.csv"  
> path <- system.file("extdata", package = "dslabs")

Which of the following commands would NOT successfully copy “murders.csv” into the folder “data”?

**Respuesta:** *file.copy(file.path(path, “murders.csv”), getwd())*

### **Question 6**

You are not sure whether the murders.csv file has a header row. How could you check this?

**Respuesta:** *Open the file in a basic text editor.* *In the RStudio “Files” pane, click on your file, then select “View File”.* *Use the command read\_lines (remembering to specify the number of rows with the n\_max argument).*

### **Question 7**

What is one difference between read\_excel() and read\_xlsx()?

**Respuesta:** *read\_excel() reads both .xls and .xlsx files by detecting the file format from its extension, while read\_xlsx() only reads .xlsx files.*

### **Question 8**

You have a file called “times.txt” that contains race finish times for a marathon. The first four lines of the file look like this:

initials,state,age,time  
vib,MA,61,6:01  
adc,TX,45,5:45  
kme,CT,50,4:19

Which line of code will NOT produce a tibble with column names “initials”, “state”, “age”, and “time”?

**Respuesta:** *race\_times <- read.csv(“times.txt”)*

### **Question 9**

You also have access to marathon finish times in the form of an Excel document named “times.xlsx”. In the Excel document, different sheets contain race information for different years. The first sheet is named “2015”, the second is named “2016”, and the third is named “2017”.

Which line of code will NOT import the data contained in the “2016” tab of this Excel sheet?

**Respuesta:** *times\_2016 <- read\_xlsx(“times.xlsx”, sheet = “2”)*

### **Question 10**

You have a comma-separated values file that contains the initials, home states, ages, and race finish times for marathon runners. The runners’ initials contain three characters for the runners’ first, middle, and last names (for example, “KME”).

You read in the file using the following code.

race\_times <- read.csv(“times.csv”)

What is the data type of the initials in the object race\_times?

**Respuesta:** *factors* Nota:*In previous versions of R, this was true, but is not any more. read.csv() no longer automatically converts characters to factors. If you want to read in character columns as factors, you can supply the argument stringsAsFactors = T.*

### **Question 11**

Which of the following is NOT a real difference between the readr import functions and the base R import functions?

**Respuesta:** *The base R import functions can read .csv files, but cannot read files with other delimiters, such as .tsv files, or fixed-width files.*

### **Question 12**

You read in a file containing runner information and marathon finish times using the following code.

race\_times <- read.csv(“times.csv”, stringsAsFactors = F)

What is the class of the object race\_times?

**Respuesta:** *data frame*

### **Question 13**

Select the answer choice that summarizes all of the actions that the following lines of code can perform. Please note that the url below is an example and does not lead to data.

url <- "https://raw.githubusercontent.com/MyUserName/MyProject/master/MyData.csv "  
dat <- read\_csv(url)  
download.file(url, "MyData.csv")

**Respuesta:** *Create a tibble in R called dat that contains the information contained in the csv file stored on Github. Download the csv file to the working directory and name the downloaded file “MyData.csv”.*

#Assessment Part 2: Data Import

In this part of the assessment, you will import real datasets and learn more about useful arguments to readr functions. You will encounter common issues that arise when importing raw data. This part of the assessment will require you to program in R.

Use the readr package in the tidyverse library:

library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.4 v purrr 0.3.4  
## v tibble 3.1.2 v dplyr 1.0.6  
## v tidyr 1.1.3 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

### **Question 14**

Inspect the file at the following URL:

<https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data>

Which readr function should be used to import this file?