# Wrangling: Comprehensive Tests

## Assessment Part 1: Data Import

### **Question 1**

1/1 point (graded)

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

Importing data into R

Formatting dates/times

Checking correlations between your variables

Tidying data

correct

**Answer**

Correct:

Data analyses, such as checking correlations or creating visualizations, are done AFTER the data has been processed into a tidy format.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 2**

0.75/1 point (graded)

Which files could be opened in a basic text editor?

Select ALL that apply.

data.txt

data.csv

data.xlsx

data.tsv

partially correct

**Answer**

Incorrect:

Try again. Microsoft Excel files use a proprietary format that can’t be viewed with a text editor.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Partially correct (0.75/1 point)

Review

### **Question 3**

1/1 point (graded)

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

What type of file is this?

A comma-delimited file without a header

A tab-delimited file with a header

A white space-delimited file without a header

A comma-delimited file with a header

correct

**Answer**

Correct:

This file has a header that describes the contents of each column. Values are separated by commas.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 4**

1/1 point (graded)

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?

setwd("~/Documents/projects/")

setwd("/Users/student/Documents/projects/")

setwd(/Users/student/Documents/projects/)

dir <- "/Users/student/Documents/projects"   
setwd(dir)

correct

**Answer**

Correct:

Correct: this code will not work. You need to use quotation marks when providing a full path to the working directory.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 5**

1/1 point (graded)

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”?



file.copy(file.path(path, "murders.csv"), getwd())



setwd("data")

file.copy(file.path(path, filename), getwd())



file.copy(file.path(path, "murders.csv"), file.path(getwd(), "data"))



file.location <- file.path(system.file("extdata", package = "dslabs"), "murders.csv")  
file.destination <- file.path(getwd(), "data")  
file.copy(file.location, file.destination)

correct

**Answer**

Correct:

This command does NOT copy the “murders.csv” file into your “data” folder; instead it copies it into the parent directory, “HarvardX-Wrangling”. You need to specify the “data” folder, either by changing the working directory or by adding it to the file path.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 6**

1/1 point (graded)

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

Select ALL that apply.

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).

correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 7**

1/1 point (graded)

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

read\_excel() also reads meta-data from the excel file, such as sheet names, while read\_xlsx() only reads the first sheet in a file.

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

read\_excel() is part of the **readr** package, while read\_xlsx() is part of the **readxl** package and has more options.

read\_xlsx() has been replaced by read\_excel() in a recent **readxl** package update.

correct

**Answer**

Correct:

You can use read\_excel for both .xls and .xlsx files, while read\_xlsx only reads .xlsx files.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 8**

1/1 point (graded)

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”?

race\_times <- read\_csv("times.txt")

race\_times <- read.csv("times.txt")

race\_times <- read\_csv("times.txt", col\_names = TRUE)

race\_times <- read\_delim("times.txt", delim = “,”)

correct

**Answer**

Correct:

This code will import the comma-separated values table called “times.txt”, but the base R function read.csv does not produce a tibble. It creates a data frame.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 9**

1/1 point (graded)

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?

times\_2016 <- read\_excel("times.xlsx", sheet = 2)

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

times\_2016 <- read\_excel("times.xlsx", sheet = "2016")

times\_2016 <- read\_xlsx("times.xlsx", sheet = 2)

correct

**Answer**

Correct:

When the sheet argument is contained within quotes, the function expects a sheet name. There is no sheet named “2” in this spreadsheet, so the code will not work.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 10**

1/1 point (graded)

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?

integers

characters

factors

logical

correct

**Answer**

Correct:

If you don’t supply the argument stringsAsFactors = F, the read.csv function will automatically convert characters to factors.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 11**

1/1 point (graded)

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

The import functions in the readr package all start as read\_, while the import functions for base R all start with read.

Base R import functions automatically convert character columns to factors.

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

Base R functions import data as a data frame, while readr functions import data as a tibble.

correct

**Answer**

Correct:

This statement is false. The base R import functions can read files with other delimiters like .tsv using read.delim and can read fixed-width files using read.fwf.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 12**

1/1 point (graded)

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?

data frame

tibble

matrix

vector

correct

**Answer**

Correct:

The read.csv function generates a data frame containing these data.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 13**

1/1 point (graded)

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")

Create a tibble in R called dat that contains the information contained in the csv file stored on Github and save that tibble to the working directory.

Create a matrix 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”.

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 randomly assign it a temporary name that is very likely to be unique.

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”.

correct

**Answer**

Correct:

The read\_csv command creates the tibble in R and the download.file command downloads the csv file from the internet to the working directory with the specified name.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Use the **readr** package in the **tidyverse**library:

library(tidyverse)

### **Question 14**

1.0/1.0 point (graded)

Inspect the file at the following URL:

<http://mlr.cs.umass.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data>

Which **readr** function should be used to import this file?

read\_table()

read\_csv()

read\_csv2()

read\_tsv()

None of the above

correct

**Answer**

Correct:

Correct - this is a comma-separated value file.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 15**

0.0/1.0 point (graded)

Check the documentation for the **readr** function you chose in the previous question to learn about its arguments. Determine which arguments you need to the file from the previous question:

url <- "http://mlr.cs.umass.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data"

Does this file have a header row? Does the **readr** function you chose need any additional arguments to import the data correctly?

Yes, there is a header. No arguments are needed.

Yes, there is a header. The header=TRUE argument is necessary.

Yes, there is a header. The col\_names=TRUE argument is necessary.

No, there is no header. No arguments are needed.

No, there is no header. The header=FALSE argument is necessary.

No, there is no header. The col\_names=FALSE argument is necessary.

incorrect

**Answer**

Incorrect:

Try again. Read the documentation for the **readr** function you chose. header is not an argument - what argument is used instead?

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

### **Question 16**

2.0/2.0 points (graded)

Inspect the imported data from the previous question.

How many rows are in the dataset?

  correct

569 Loading

How many columns are in the dataset?

  correct

## 32 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Section 2: Tidy Data

## 2.1 Reshaping Data

## Assessment Part 1: Reshaping Data

 Bookmark this page

Part 1 consists of 8 questions are conceptual questions about tidy data and reshaping data. They do not necessarily require R, but you may benefit from checking your work on the console.

Part 2 consists of 7 questions which require you to write code in R to apply the new concepts about tidy data and reshaping data.

### **Question 1**

1/1 point (graded)

A collaborator sends you a file containing data for three years of average race finish times.

age\_group,2015,2016,2017

20,3:46,3:22,3:50

30,3:50,3:43,4:43

40,4:39,3:49,4:51

50,4:48,4:59,5:01

Are these data considered “tidy” in R? Why or why not?

Yes. These data are considered “tidy” because each row contains unique observations.

Yes. These data are considered “tidy” because there are no missing data in the data frame.

No. These data are not considered “tidy” because the variable “year” is stored in the header.

No. These data are not considered “tidy” because there are not an equal number of columns and rows.

correct

**Answer**

Correct:

The year is a variable and should be stored as a column instead of across multiple columns in the header.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 2**

1/1 point (graded)

Below are four versions of the same dataset. Which one is in a tidy format?



state abb region population total

Alabama AL South 4779736 135

Alaska AK West 710231 19

Arizona AZ West 6392017 232

Arkansas AR South 2915918 93

California CA West 37253956 1257

Colorado CO West 5029196 65



state abb region var people

Alabama AL South population 4779736

Alabama AL South total 135

Alaska AK West population 710231

Alaska AK West total 19

Arizona AZ West population 6392017

Arizona AZ West total 232



state abb Northeast South North Central West

Alabama AL NA 4779736 NA NA

Alaska AK NA NA NA 710231

Arizona AZ NA NA NA 6392017

Arkansas AR NA 2915918 NA NA

California CA NA NA NA 37253956

Colorado CO NA NA NA 5029196



state abb region rate

Alabama AL South 2.82e-05

Alaska AK West 2.68e-05

Arizona AZ West 3.63e-05

Arkansas AR South 3.19e-05

California CA West 3.37e-05

Colorado CO West 1.29e-05

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 3**

1/1 point (graded)

Your file called “times.csv” has age groups and average race finish times for three years of marathons.

age\_group,2015,2016,2017

20,3:46,3:22,3:50

30,3:50,3:43,4:43

40,4:39,3:49,4:51

50,4:48,4:59,5:01

You read in the data file using the following command.

d <- read\_csv("times.csv")

Which commands will help you “tidy” the data?



tidy\_data <- d %>%

gather(year, time, `2015`:`2017`)



tidy\_data <- d %>%

spread(year, time, `2015`:`2017`)



tidy\_data <- d %>%

gather(age\_group, year, time, `2015`:`2017`)



tidy\_data <- d %>%

gather(time, `2015`:`2017`)

correct

**Answer**

Correct:

This code will gather the years from 2015 to 2017 into a single column and create a single column called “time” that contains the time for each age group and each year.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 4**

1/1 point (graded)

You have a dataset on U.S. contagious diseases, but it is in the following wide format:

> head(dat\_wide)

state year population HepatitisA Mumps Polio Rubella

Alabama 1990 4040587 86 19 76 1

Alabama 1991 4066003 39 14 65 0

Alabama 1992 4097169 35 12 24 0

Alabama 1993 4133242 40 22 67 0

Alabama 1994 4173361 72 12 39 0

Alabama 1995 4216645 75 2 38 0

You want to transform this into a tidy dataset, with each row representing an observation of the incidence of each specific disease (as shown below):

> head(dat\_tidy)

state year population disease count

Alabama 1990 4040587 HepatitisA 86

Alabama 1991 4066003 HepatitisA 39

Alabama 1992 4097169 HepatitisA 35

Alabama 1993 4133242 HepatitisA 40

Alabama 1994 4173361 HepatitisA 72

Alabama 1995 4216645 HepatitisA 75

Which of the following commands would achieve this transformation to tidy the data?

Pay attention to the column names.



dat\_tidy <- dat\_wide %>%

gather (key = count, value = disease, HepatitisA, Rubella)



dat\_tidy <- dat\_wide %>%

gather(key = count, value = disease, -state, -year, -population)



dat\_tidy <- dat\_wide %>%

gather(key = disease, value = count, -state)



dat\_tidy <- dat\_wide %>%

gather(key = disease, value = count, HepatitisA:Rubella)

correct

**Answer**

Correct:

In this command, you properly specified that the “key” column will be called “disease”, the value of each entry will be called “count”, and that the columns HepatitisA through Rubella will all be included in the gather command.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 5**

1/1 point (graded)

You have successfully formatted marathon finish times into a tidy object called tidy\_data. The first few lines are shown below.

age\_group year time

20 2015 03:46

30 2015 03:50

40 2015 04:39

50 2015 04:48

20 2016 03:22

Select the code that converts these data back to the wide format, where each year has a separate column.



tidy\_data %>% spread(time, year)



tidy\_data %>% spread(year, time)



tidy\_data %>% spread(year, age\_group)



tidy\_data %>% spread(time, year, `2015`:`2017`)

correct

**Answer**

Correct:

This code tells the function to create new columns for each year and spread the time values over those cells.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 6**

1/1 point (graded)

You have the following dataset:

> head(dat)

state abb region var people

Alabama AL South population 4779736

Alabama AL South total 135

Alaska AK West population 710231

Alaska AK West total 19

Arizona AZ West population 6392017

Arizona AZ West total 232

You would like to transform it into a dataset where population and total are each their own column (shown below):

state abb region population total

Alabama AL South 4779736 135

Alaska AK West 710231 19

Arizona AZ West 6392017 232

Arkansas AR South 2915918 93

California CA West 37253956 1257

Colorado CO West 5029196 65

Which code would best accomplish this?



dat\_tidy <- dat %>% spread(key = var, value = people)



dat\_tidy <- dat %>% spread(key = state:region, value = people)



dat\_tidy <- dat %>% spread(key = people, value = var)



dat\_tidy <- dat %>% spread(key = region, value = people)

correct

**Answer**

Correct:

In this command, you properly specify that the column “var” will be used as the new column names, and that the column “people” should be spread into these two columns.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 7**

1/1 point (graded)

A collaborator sends you a file containing data for two years of average race finish times, "times.csv":

age\_group,2015\_time,2015\_participants,2016\_time,2016\_participants

20,3:46,54,3:22,62

30,3:50,60,3:43,58

40,4:39,29,3:49,33

50,4:48,10,4:59,14

You read in the data file:

d <- read\_csv("times.csv")

Which of the answers below best makes the data tidy?



tidy\_data <- d %>%

gather(key = “key”, value = “value”, -age\_group) %>%

separate(col = key, into = c(“year”, “variable\_name”), sep = “.”) %>%

spread(key = variable\_name, value = value)



tidy\_data <- d %>%

gather(key = “key”, value = “value”, -age\_group) %>%

separate(col = key, into = c(“year”, “variable\_name”), sep = “\_”) %>%

spread(key = variable\_name, value = value)



tidy\_data <- d %>%

gather(key = “key”, value = “value”) %>%

separate(col = key, into = c(“year”, “variable\_name”), sep = “\_”) %>%

spread(key = variable\_name, value = value)



tidy\_data <- d %>%

gather(key = “key”, value = “value”, -age\_group) %>%

separate(col = key, into = “year”, sep = “\_”) %>%

spread(key = year, value = value)

correct

**Answer**

Correct:

This column gathers the column names 2015\_time, 2015\_participants, 2016\_time, and 2016\_participants into one column called “key”, with the values for each stored in the column “value.” The key column is then separated into two columns, “year” and “variable\_name”. The two entries for “variable\_name”, time and participants, are then spread into their own columns.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 8**

0.33/1 point (graded)

You are in the process of tidying some data on heights, hand length, and wingspan for basketball players in the draft. Currently, you have the following:

> head(stats)

key value

allen\_height 75

allen\_hand\_length 8.25

allen\_wingspan 79.25

bamba\_height 83.25

bamba\_hand\_length 9.75

bamba\_wingspan 94

Select all of the correct commands below that would turn this data into a “tidy” format with columns "height", "hand\_length" and "wingspan".



tidy\_data <- stats %>%

separate(col = key, into = c("player", "variable\_name"), sep = "\_", extra = "merge") %>%

spread(key = variable\_name, value = value)



tidy\_data <- stats %>%

separate(col = key, into = c("player", "variable\_name1", "variable\_name2"), sep = "\_", fill = "right") %>%

unite(col = variable\_name, variable\_name1, variable\_name2, sep = "\_") %>%

spread(key = variable\_name, value = value)



tidy\_data <- stats %>%

separate(col = key, into = c("player", "variable\_name"), sep = "\_") %>%

spread(key = variable\_name, value = value)

partially correct

**Answer**

Incorrect:

This is an efficient way to separate the key column into two new columns, “player” and “variable\_name”, while keeping the full variable names using the extra command.

Try again. This wrangling code does not generate the correct variable names. Pay close attention to your separate command.

Try again. If you use this separate command, the second word of the variable hand\_length will be lost, because you do not account for the extra underscore.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Assessment Part 2: Reshaping Data

 Bookmark this page

Use the following libraries for these questions:

library(tidyverse)  
library(dslabs)

### **Question 9**

1/1 point (graded)

Examine the built-in dataset co2. This dataset comes with base R, not **dslabs** - just type co2 to access the dataset.

Is co2 tidy? Why or why not?

co2 is tidy data: it has one year for each row.

co2 is tidy data: each column is a different month.

co2 is not tidy: there are multiple observations per column.

co2 is not tidy: to be tidy we would have to wrangle it to have three columns (year, month and value), and then each co2 observation would have a row.

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 10**

1/1 point (graded)

Run the following command to define the co2\_wide object:

co2\_wide <- data.frame(matrix(co2, ncol = 12, byrow = TRUE)) %>%

setNames(1:12) %>%

mutate(year = as.character(1959:1997))

Use the gather() function to make this dataset tidy. Call the column with the CO2 measurements co2 and call the month column month. Name the resulting object co2\_tidy.

Which code would return the correct tidy format?

co2\_tidy <- gather(co2\_wide,month,co2,year)

co2\_tidy <- gather(co2\_wide,co2,month,-year)

co2\_tidy <- gather(co2\_wide,co2,month,year)

co2\_tidy <- gather(co2\_wide,month,co2,-year)

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 11**

1/1 point (graded)

Use co2\_tidy to plot CO2 versus month with a different curve for each year:

co2\_tidy %>% ggplot(aes(as.numeric(month), co2, color = year)) + geom\_line()

What can be concluded from this plot?

CO2 concentrations increased monotonically (never decreased) from 1959 to 1997.

CO2 concentrations are highest around May and the yearly average increased from 1959 to 1997.

CO2 concentrations are highest around October and the yearly average increased from 1959 to 1997.

Yearly average CO2 concentrations have remained constant over time.

CO2 concentrations do not have a seasonal trend.

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 12**

1/1 point (graded)

Load the admissions dataset from **dslabs**, which contains college admission information for men and women across six majors, and remove the applicants percentage column:

library(dslabs)

data(admissions)

dat <- admissions %>% select(-applicants)

Your goal is to get the data in the shape that has one row for each major, like this:

major men women

A 62 82

B 63 68

C 37 34

D 33 35

E 28 24

F 6 7

Which command could help you to wrangle the data into the desired format?

dat\_tidy <- spread(dat, major, admitted)

dat\_tidy <- spread(dat, gender, major)

dat\_tidy <- spread(dat, gender, admitted)

dat\_tidy <- spread(dat, admitted, gender)

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 13**

1/1 point (graded)

Now use the admissions dataset to create the object tmp, which has columns major, gender, key and value:

tmp <- gather(admissions, key, value, admitted:applicants)

tmp

Combine the key and gender and create a new column called column\_name to get a variable with the following values: admitted\_men, admitted\_women, applicants\_men and applicants\_women. Save the new data as tmp2.

Which command could help you to wrangle the data into the desired format?



tmp2 <- spread(tmp, column\_name, key, gender)



tmp2 <- gather(tmp, column\_name, c(gender,key))



tmp2 <- unite(tmp, column\_name, c(gender, key))



tmp2 <- spread(tmp, column\_name, c(key,gender))



tmp2 <- unite(tmp, column\_name, c(key, gender))

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 14**

1/1 point (graded)

Which function can reshape tmp2 to a table with six rows and five columns named major, admitted\_men, admitted\_women, applicants\_menand applicants\_women?

gather()

spread()

separate()

unite()

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

## 2.2 Combining Tables

## Assessment: Combining Tables

 Bookmark this page

### **Question 1**

1/1 point (graded)

You have created data frames tab1 and tab2 of state population and election data, similar to our module videos:

> tab1

state population

Alabama 4779736

Alaska 710231

Arizona 6392017

Delaware 897934

District of Columbia 601723

> tab2

state electoral\_votes

Alabama 9

Alaska 3

Arizona 11

California 55

Colorado 9

Connecticut 7

> dim(tab1)

[1] 5 2

> dim(tab2)

[1] 6 2

What are the dimensions of the table dat, created by the following command?

dat <- left\_join(tab1, tab2, by = “state”)

3 rows by 3 columns

5 rows by 2 columns

5 rows by 3 columns

6 rows by 3 columns

correct

**Answer**

Correct:

When we use a left\_join command, all rows in the left-hand table (in this case, tab1) are retained in the final table, so we expect to have five rows. In addition, columns from both tables will be included in the final “dat” table so we expect to have three columns.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 2**

1/1 point (graded)

We are still using the tab1 and tab2 tables shown in question 1. What join command would create a new table “dat” with three rows and two columns?



dat <- right\_join(tab1, tab2, by = “state”)



dat <- full\_join(tab1, tab2, by = “state”)



dat <- inner\_join(tab1, tab2, by = “state”)



dat <- semi\_join(tab1, tab2, by = “state”)

correct

**Answer**

Correct:

The semi\_join command takes tab1 and limits it to states that are also in tab2, without adding the additional columns in tab2. This gives us three rows (states in both tables) and two columns (state and population, the two columns in tab1).

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 3**

0.75/1 point (graded)

Which of the following are real differences between the join and bind functions?

Please select all correct answers.

Binding functions combine by position, while join functions match by variables.

Joining functions can join datasets of different dimensions, but the bind functions must match on the appropriate dimension (either same row or column numbers).

Bind functions can combine both vectors and dataframes, while join functions work for only for dataframes.

The join functions are a part of the dplyr package and have been optimized for speed, while the bind functions are inefficient base functions.

partially correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Partially correct (0.75/1 point)

Review

### **Question 4**

1/1 point (graded)

We have two simple tables, shown below, with columns x and y:

> df1

x y

a a

b a

> df2

x y

a a

a b

Which command would result in the following table?

> final

x y

b a



final <- union(df1, df2)



final <- setdiff(df1, df2)



final <- setdiff(df2, df1)



final <- intersect(df1, df2)

correct

**Answer**

Correct:

The setdiff() command returns rows in df1 but not df2, which matches our table final.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### Introduction to Questions 5-7

Install and load the **Lahman** library. This library contains a variety of datasets related to US professional baseball. We will use this library for the next few questions and will discuss it more extensively in the Regression course. For now, focus on wrangling the data rather than understanding the statistics.

The Batting data frame contains the offensive statistics for all baseball players over several seasons.  Filter this data frame to define top as the top 10 home run (HR) hitters in 2016:

library(Lahman)

top <- Batting %>%

filter(yearID == 2016) %>%

arrange(desc(HR)) %>% # arrange by descending HR count

slice(1:10) # take entries 1-10

top %>% as\_tibble()

Also Inspect the Master data frame, which has demographic information for all players:

Master %>% as\_tibble()

### **Question 5**

1/1 point (graded)

Use the correct join or bind function to create a combined table of the names and statistics of the top 10 home run (HR) hitters for 2016. This table should have the player ID, first name, last name, and number of HR for the top 10 players. Name this data frame top\_names.

Identify the join or bind that fills the blank in this code to create the correct table:

top\_names <- top %>% \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ %>%

select(playerID, nameFirst, nameLast, HR)

Which bind or join function fills the blank to generate the correct table?

rbind(Master)

cbind(Master)

left\_join(Master)

right\_join(Master)

full\_join(Master)

anti\_join(Master)

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 6**

1/1 point (graded)

Inspect the Salaries data frame. Filter this data frame to the 2016 salaries, then use the correct bind join function to add a salary column to the top\_names data frame from the previous question. Name the new data frame top\_salary. Use this code framework:

top\_salary <- Salaries %>% filter(yearID == 2016) %>%

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ %>%

select(nameFirst, nameLast, teamID, HR, salary)

Which bind or join function fills the blank to generate the correct table?

rbind(top\_names)

cbind(top\_names)

left\_join(top\_names)

right\_join(top\_names)

full\_join(top\_names)

anti\_join(top\_names)

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 7**

2/2 points (graded)

Inspect the AwardsPlayers table. Filter awards to include only the year 2016.

How many players from the top 10 home run hitters won at least one award in 2016?

Use a set operator.

  correct

3 Loading

How many players won an award in 2016 but were not one of the top 10 home run hitters in 2016?

Use a set operator.

  correct

44 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

## 2.3 Web Scraping

## Assessment: Web Scraping

 Bookmark this page

### Introduction: Questions 1-3

Load the following web page, which contains information about Major League Baseball payrolls, into R: [https://web.archive.org/web/20181024132313/http://www.stevetheump.com/Payrolls.htm](https://web.archive.org/web/20181024132313/http:/www.stevetheump.com/Payrolls.htm)

library(rvest)

url <- "https://web.archive.org/web/20181024132313/http://www.stevetheump.com/Payrolls.htm"

h <- read\_html(url)

We learned that tables in html are associated with the table node.  Use the html\_nodes() function and the table node type to extract the first table. Store it in an object nodes:

nodes <- html\_nodes(h, "table")

The html\_nodes() function returns a list of objects of class xml\_node. We can see the content of each one using, for example, the html\_text() function. You can see the content for an arbitrarily picked component like this:

html\_text(nodes[[8]])

If the content of this object is an html table, we can use the html\_table() function to convert it to a data frame:

html\_table(nodes[[8]])

You will analyze the tables from this HTML page over questions 1-3.

### **Question 1**

2.5/2.5 points (graded)

Many tables on this page are team payroll tables, with columns for rank, team, and one or more money values.

Convert the first four tables in nodes to data frames and inspect them.

Which of the first four nodes are tables of team payroll?

Check all correct answers. Look at table content, not column names.

None of the above

Table 1

Table 2

Table 3

Table 4

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 2**

2.0/2.0 points (graded)

For the last 3 components of nodes, which of the following are true? (Check all correct answers.)

Check all correct answers.

All three entries are tables.

All three entries are tables of payroll per team.

The last entry shows the average across all teams through time, not payroll per team.

None of the three entries are tables of payroll per team.

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3**

1/1 point (graded)

Create a table called tab\_1 using entry 10 of nodes. Create a table called tab\_2 using entry 19 of nodes.

Note that the column names should be c("Team", "Payroll", "Average"). You can see that these column names are actually in the first data row of each table, and that tab\_1 has an extra first column No. that should be removed so that the column names for both tables match.

Remove the extra column in tab\_1, remove the first row of each dataset, and change the column names for each table to c("Team", "Payroll", "Average"). Use a full\_join() by the Team to combine these two tables.

Note that some students, presumably because of system differences, have noticed that entry 18 instead of entry 19 of nodes gives them the tab\_2 correctly; be sure to check entry 18 if entry 19 is giving you problems.

How many rows are in the joined data table?  correct

\(\) Loading

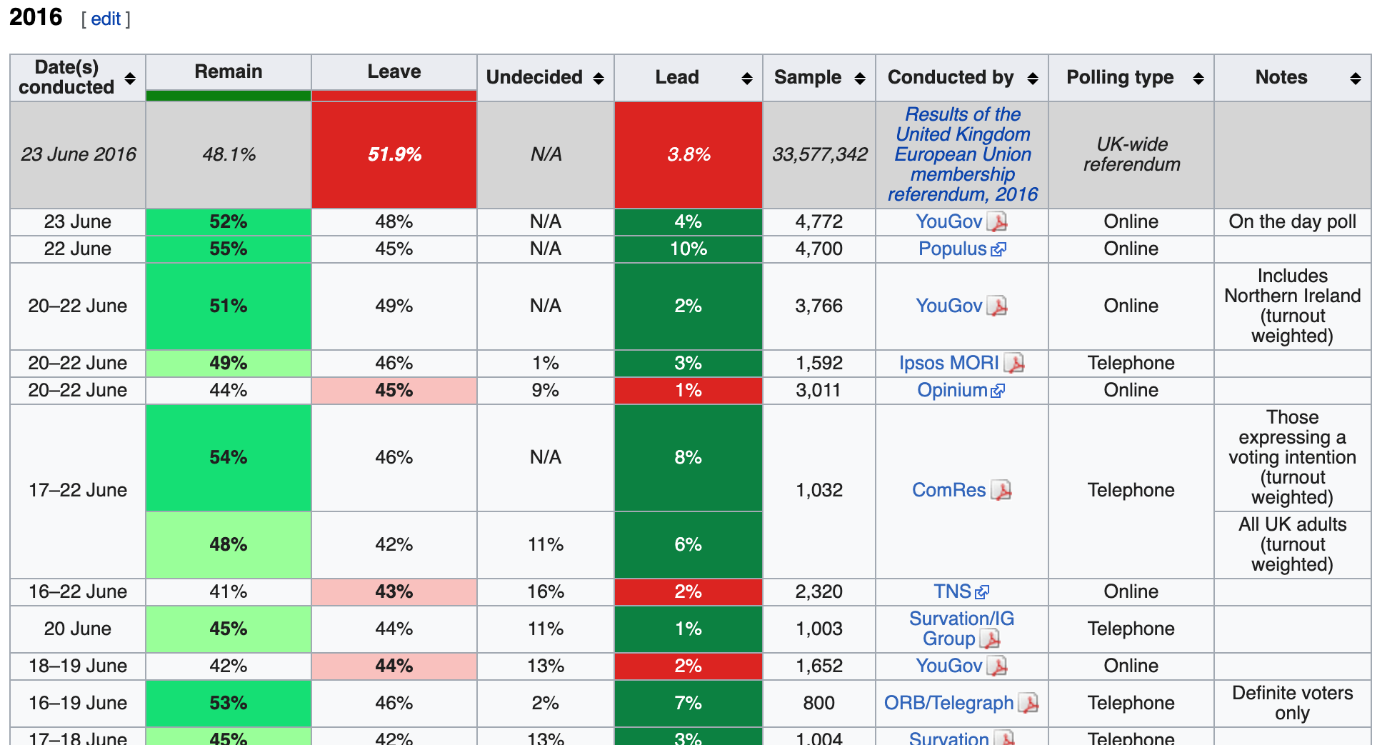
Submit

You have used 4 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

#### Introduction: Questions 4 and 5

The Wikipedia page on [opinion polling for the Brexit referendum](https://en.wikipedia.org/w/index.php?title=Opinion_polling_for_the_United_Kingdom_European_Union_membership_referendum&oldid=896735054), in which the United Kingdom voted to leave the European Union in June 2016, contains several tables. One table contains the results of all polls regarding the referendum over 2016:



Use the **rvest** library to read the HTML from this Wikipedia page (make sure to copy both lines of the URL):

library(rvest)  
library(tidyverse)  
url <- "https://en.wikipedia.org/w/index.php?title=Opinion\_polling\_for\_the\_United\_Kingdom\_European\_Union\_membership\_referendum&oldid=896735054"

### **Question 4**

1/1 point (graded)

Assign tab to be the html nodes of the "table" class.

How many tables are in this Wikipedia page?  correct

\(\) Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 5**

1/1 point (graded)

Inspect the first several html tables using html\_table() with the argument fill=TRUE (you can read about this argument in the documentation). Find the first table that has 9 columns with the first column named "Date(s) conducted".

What is the first table number to have 9 columns where the first column is named "Date(s) conducted"?  correct

\(\) Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Section 3: String Processing

## 3.1 String Processing Part 1

## Assessment: String Processing Part 1

 Bookmark this page

### **Question 1**

1/1 point (graded)

Which of the following is NOT an application of string parsing?

Removing unwanted characters from text.

Extracting numeric values from text.

Formatting numbers and characters so they can easily be displayed in deliverables like papers and presentations.

Splitting strings into multiple values.

correct

**Answer**

Correct:

Formatting text and numbers for deliverables is not an application of string parsing. String parsing is used as part of the data wrangling process.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 2**

1/1 point (graded)

Which of the following commands would not give you an error in R?



cat(" LeBron James is 6’8\" ")



cat(' LeBron James is 6'8" ')



cat(` LeBron James is 6'8" `)



cat(" LeBron James is 6\’8" ")

correct

**Answer**

Correct:

This would correctly print out your string. Because the string is enclosed in double quotes, (“”), you must use an escape character before the inches symbol (“).

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 3**

1/1 point (graded)

Which of the following are advantages of the **stringr** package over string processing functions in base R? Select all that apply.

Base R functions are rarely used for string processing by data scientists so it’s not worth learning them.

Functions in stringr all start with “str\_”, which makes them easy to look up using autocomplete.

Stringr functions work better with pipes.

The order of arguments is more consistent in stringr functions than in base R.

correct

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 4**

1/1 point (graded)

You have a data frame of monthly sales and profits in R:

> head(dat)

# A tibble: 5 x 3

Month Sales Profit

<chr> <chr> <chr>

January $128,568 $16,234

February $109,523 $12,876

March $115,468 $17,920

April $122,274 $15,825

May $117,921 $15,437

Which of the following commands could convert the sales and profits columns to numeric? Select all that apply.



dat %>% mutate\_at(2:3, parse\_number)



dat %>% mutate\_at(2:3, as.numeric)



dat %>% mutate\_all(parse\_number)



dat %>% mutate\_at(2:3, funs(str\_replace\_all(., c("\\$|,"), ""))) %>%

mutate\_at(2:3, as.numeric)

correct

**Answer**

Correct:

You can use the parse\_number command to remove all non-numeric characters. Combining this with the mutate\_at command allows you to reformat column two and three (Sales and Profit).

You can use the str\_replace\_all command to replace both the “$” and “,” characters, by specifying these in the “pattern” argument of the command. Combining this function with the mutate\_at command allows you to reformat both column two and three (Sales and Profit). You then need to use the “as.numeric” command to convert these columns from character strings to numbers.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

## 3.2 String Processing part 2

## Assessment: String Processing Part 2

 Bookmark this page

### **Question 1**

1/1 point (graded)

In the video, we use the function not\_inches to identify heights that were incorrectly entered

not\_inches <- function(x, smallest = 50, tallest = 84) {

inches <- suppressWarnings(as.numeric(x))

ind <- is.na(inches) | inches < smallest | inches > tallest

ind

}

In this function, what TWO types of values are identified as not being correctly formatted in inches?

Values that specifically contain apostrophes (‘), periods (.) or quotations (“).

Values that result in NA’s when converted to numeric

Values less than 50 inches or greater than 84 inches

Values that are stored as a character class, because most are already classed as numeric.

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 2**

1/1 point (graded)

Which of the following arguments, when passed to the function not\_inches(), would return the vector c(FALSE)?

c(175)

c(“5’8\””)

c(70)

c(85) (the height of Shaquille O'Neal in inches)

correct

**Answer**

Correct:

The entry 70 can be converted to a numeric entry by as.numeric and is within the range that we set. Therefore, the result of this function would be FALSE (i.e., our entry is correctly formatted in inches).

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 3**

1/1 point (graded)

Our function not\_inches() returns the object ind. Which answer correctly describes ind?

ind is a logical vector of TRUE and FALSE, equal in length to the vector x (in the arguments list). TRUE indicates that a height entry is incorrectly formatted.

ind is a logical vector of TRUE and FALSE, equal in length to the vector x (in the arguments list). TRUE indicates that a height entry is correctly formatted.

ind is a data frame like our reported\_heights table but with an extra column of TRUE or FALSE. TRUE indicates that a height entry is incorrectly formatted.

ind is a numeric vector equal to reported\_heights$heights but with incorrectly formatted heights replaced with NAs.

correct

**Answer**

Correct:

Our function returns a logical vector, with TRUE indicating that a height entry is incorrectly formatted and FALSE indicating that a height entry is formatted properly in inches. We then use this logical vector to filter our raw\_heights data to only show incorrectly formatted entries.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 4**

1/1 point (graded)

Given the following code

> s

[1] "70" "5 ft" "4'11" "" "." "Six feet"

What pattern vector yields the following result?

str\_view\_all(s, pattern)

70

5 ft

4’11

.

Six feet

pattern <- "\\d|ft"

pattern <- "\d|ft"

pattern <- "\\d\\d|ft"

pattern <- "\\d|feet"

correct

**Answer**

Correct:

This regex identifies any numeric characters or the text “ft”.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

### **Question 5**

1/1 point (graded)

You enter the following set of commands into your R console. What is your printed result?

animals <- c("cat", "puppy", "Moose", "MONKEY")

pattern <- "[a-z]"

str\_detect(animals, pattern)

TRUE

TRUE TRUE TRUE TRUE

TRUE TRUE TRUE FALSE

TRUE TRUE FALSE FALSE

correct

**Answer**

Correct:

While your first three strings have at least one lowercase letter [a-z], the string MONKEY does not have any lowercase letters and will return a FALSE.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

### **Question 6**

1/1 point (graded)

You enter the following set of commands into your R console. What is your printed result?

animals <- c("cat", "puppy", "Moose", "MONKEY")

pattern <- "[A-Z]$"

str\_detect(animals, pattern)

FALSE FALSE FALSE FALSE

FALSE FALSE TRUE TRUE

FALSE FALSE FALSE TRUE

TRUE TRUE TRUE FALSE

correct

**Answer**

Correct:

Your regex pattern tells str\_detect to look for an uppercase ([A-Z]) letter at the end of the string ($): this is only true for the string “MONKEY”.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 7**

1/1 point (graded)

You enter the following set of commands into your R console. What is your printed result?

animals <- c("cat", "puppy", "Moose", "MONKEY")

pattern <- "[a-z]{4,5}"

str\_detect(animals, pattern)

FALSE TRUE TRUE FALSE

TRUE TRUE FALSE FALSE

FALSE FALSE FALSE TRUE

TRUE TRUE TRUE FALSE

correct

**Answer**

Correct:

Your regex command tells R to look for either 4 or 5 lowercase letters in a row anywhere in the string. This is true for the animals “puppy” and “Moose”.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 8**

1/1 point (graded)

Given the following code:

animals <- c("moose", "monkey", "meerkat", "mountain lion")

Which TWO “pattern” vectors would yield the following result?

> str\_detect(animals, pattern)

[1] TRUE TRUE TRUE TRUE



pattern <- “mo\*”



pattern <- “mo?”



pattern <- “mo+”



pattern <- “moo\*”

correct

**Answer**

Correct:

This regex pattern looks for an “m” followed by zero or more “o” characters. This is true for all strings in the animal vector.

This regex pattern looks for an “m” followed by zero or one “o” characters. This is true for all strings in the animal vector. Even though “moose” has two “o”s after the “m”, it still matches the pattern.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 9**

1/1 point (graded)

You are working on some data from different universities. You have the following vector:

> schools

[1] "U. Kentucky" "Univ New Hampshire" "Univ. of Massachusetts" "University Georgia"

[5] "U California" "California State University"

You want to clean this data to match the full names of each university:

> final

[1] "University of Kentucky" "University of New Hampshire" "University of Massachusetts" "University of Georgia"

[5] "University of California" "California State University"

What of the following commands could accomplish this?



schools %>%

str\_replace("Univ\\.?|U\\.?", "University ") %>%

str\_replace("^University of |^University ", "University of ")



schools %>%

str\_replace("^Univ\\.?\\s|^U\\.?\\s", "University ") %>%

str\_replace("^University of |^University ", "University of ")



schools %>%

str\_replace("^Univ\\.\\s|^U\\.\\s", "University") %>%

str\_replace("^University of |^University ", "University of ")



schools %>%

str\_replace("^Univ\\.?\\s|^U\\.?\\s", "University") %>%

str\_replace("University ", "University of ")

correct

**Answer**

Correct:

This code properly replaces all versions of “U”, “U.”, “Univ” and “Univ.” with “University” and then adds the word “of”.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 10**

1/1 point (graded)

Rather than using the pattern\_with\_groups vector from the video, you accidentally write in the following code:

problems <- c("5.3", "5,5", "6 1", "5 .11", "5, 12")

pattern\_with\_groups <- "^([4-7])[,\\.](\\d\*)$"

str\_replace(problems, pattern\_with\_groups, "\\1'\\2")

What is your result?

[1] "5'3" "5'5" "6 1" "5 .11" "5, 12"

[1] "5.3" "5,5" "6 1" "5 .11" "5, 12"

[1] "5'3" "5'5" "6'1" "5 .11" "5, 12"

[1] "5'3" "5'5" "6'1" "5’11" "5’12"

correct

**Answer**

Correct:

Yes. You forgot to check for any spaces in your regex pattern. While the first two entries of “problems” have commas and periods correctly replaced, the last three entries are not identified as part of the pattern and are not replaced.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

### **Question 11**

1/1 point (graded)

You notice your mistake and correct your pattern regex to the following

problems <- c("5.3", "5,5", "6 1", "5 .11", "5, 12")

pattern\_with\_groups <- "^([4-7])[,\\.\\s](\\d\*)$"

str\_replace(problems, pattern\_with\_groups, "\\1'\\2")

What is your result?

[1] "5'3" "5'5" "6 1" "5 .11" "5, 12"

[1] "5.3" "5,5" "6 1" "5 .11" "5, 12"

[1] "5'3" "5'5" "6'1" "5 .11" "5, 12"

[1] "5'3" "5'5" "6'1" "5’11" "5’12"

correct

**Answer**

Correct:

The new regex pattern now checks for one character, either a comma, period or space, between the first digit and the last one or two digits, and replaces it with an apostrophe (‘). However, because your last two problem strings have additional space between the digits, they are not corrected.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 12**

1/1 point (graded)

In our example, we use the following code to detect height entries that do not match our pattern of x’y”:

converted <- problems %>%

str\_replace("feet|foot|ft", "'") %>%

str\_replace("inches|in|''|\"", "") %>%

str\_replace("^([4-7])\\s\*[,\\.\\s+]\\s\*(\\d\*)$", "\\1'\\2")

pattern <- "^[4-7]\\s\*'\\s\*\\d{1,2}$"

index <- str\_detect(converted, pattern)

converted[!index]

Which answer best describes the differences between the regex string we use as an argument in str\_replace("^([4-7])\\s\*[,\\.\\s+]\\s\*(\\d\*)$", "\\1'\\2") and the regex string in pattern <- "^[4-7]\\s\*'\\s\*\\d{1,2}$"?

The regex used in str\_replace() looks for either a comma, period or space between the feet and inches digits, while the pattern regex just looks for an apostrophe; the regex in str\_replace allows for one or more digits to be entered as inches, while the pattern regex only allows for one or two digits.

The regex used in str\_replace() allows for additional spaces between the feet and inches digits, but the pattern regex does not.

The regex used in str\_replace() looks for either a comma, period or space between the feet and inches digits, while the pattern regex just looks for an apostrophe; the regex in str\_replace allows none or more digits to be entered as inches, while the pattern regex only allows for the number 1 or 2 to be used.

The regex used in str\_replace() looks for either a comma, period or space between the feet and inches digits, while the pattern regex just looks for an apostrophe; the regex in str\_replace allows for none or more digits to be entered as inches, while the pattern regex only allows for one or two digits.

correct

**Answer**

Correct:

This answer describes two important differences in the str\_replace() regex and the pattern regex.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

### **Question 13**

1/1 point (graded)

You notice a few entries that are not being properly converted using your str\_replace() and str\_detect() code:

yes <- c("5 feet 7inches", “5 7”)

no <- c("5ft 9 inches", "5 ft 9 inches")

s <- c(yes, no)

converted <- s %>%

str\_replace("feet|foot|ft", "'") %>%

str\_replace("inches|in|''|\"", "") %>%

str\_replace("^([4-7])\\s\*[,\\.\\s+]\\s\*(\\d\*)$", "\\1'\\2")

pattern <- "^[4-7]\\s\*'\\s\*\\d{1,2}$"

str\_detect(converted, pattern)

[1] TRUE TRUE FALSE FALSE

It seems like the problem may be due to spaces around the words feet|foot|ft and inches|in. What is another way you could fix this problem?



converted <- s %>%

str\_replace("\\s\*(feet|foot|ft)\\s\*", "'") %>%

str\_replace("\\s\*(inches|in|''|\")\\s\*", "") %>%

str\_replace("^([4-7])\\s\*[,\\.\\s+]\\s\*(\\d\*)$", "\\1'\\2")



converted <- s %>%

str\_replace("\\s+feet|foot|ft\\s+”, "'") %>%

str\_replace("\\s+inches|in|''|\"\\s+", "") %>%

str\_replace("^([4-7])\\s\*[,\\.\\s+]\\s\*(\\d\*)$", "\\1'\\2")



converted <- s %>%

str\_replace("\\s\*|feet|foot|ft", "'") %>%

str\_replace("\\s\*|inches|in|''|\"", "") %>%

str\_replace("^([4-7])\\s\*[,\\.\\s+]\\s\*(\\d\*)$", "\\1'\\2")



converted <- s %>%

str\_replace\_all(“\\s”, “”) %>%

str\_replace("\\s|feet|foot|ft", "'") %>%

str\_replace("\\s|inches|in|''|\"", "") %>%

str\_replace("^([4-7])\\s\*[,\\.\\s+]\\s\*(\\d\*)$", "\\1'\\2")

correct

**Answer**

Correct:

You can add a none or more space character (\\s\*) before and after each word to properly replace the word and any additional spaces with an apostrophe.

## 3.3 String Parsing part 3

## Assessment Part 1: String Processing Part 3

 Bookmark this page

In this part of the assessment, you will answer several multiple choice questions that review the concepts of string processing. 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 a real dataset and use string processing to clean it for analysis. This will require you to write code in R.

Want even more practice with regular expressions? Complete the lessons and exercises in the [RegexOne](https://regexone.com/) online interactive tutorial!

### **Question 1**

1/1 point (graded)

s <- c("5'10", "6'1\"", "5'8inches", "5'7.5")

tab <- data.frame(x = s)

If you use the extract code from our video, the decimal point is dropped. What modification of the code would allow you to put the decimals in a third column called “decimal”?



extract(data = tab, col = x, into = c(“feet”, “inches”, “decimal”),

regex = "(\\d)'(\\d{1,2})(\\.)?"



extract(data = tab, col = x, into = c("feet", "inches", "decimal"),

regex = "(\\d)'(\\d{1,2})(\\.\\d+)"



extract(data = tab, col = x, into = c("feet", "inches", "decimal"),

regex = "(\\d)'(\\d{1,2})\\.\\d+?"



extract(data = tab, col = x, into = c("feet", "inches", "decimal"),

regex = "(\\d)'(\\d{1,2})(\\.\\d+)?")

correct

**Answer**

Correct:

In this code, you extract three groups: one digit for “feet”, one or two digits for “inches”, and an optional decimal point followed by at least one digit for “decimal”.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 2**

0.75/1 point (graded)

You have the following table, schedule:

>schedule

day staff

Monday Mandy, Chris and Laura

Tuesday Steve, Ruth and Frank

You want to turn this into a more useful data frame.

Which two commands would properly split the text in the “staff” column into each individual name? Select ALL that apply.

str\_split(schedule$staff, ",|and")

str\_split(schedule$staff, ", | and ")

str\_split(schedule$staff, ",\\s|\\sand\\s")

str\_split(schedule$staff, "\\s?(,|and)\\s?")

partially correct

**Answer**

Incorrect:

This regex will correctly split each “staff” string into three names by properly accounting for the space after the comma as well as the spaces before and after the “and”, but it’s not the only one.

This regex command is the same as the one above, except that the spaces are written as \\s, but it’s not the only one.

Try again. While this regex seems more flexible by allowing for the possibility of spaces before and after either the comma or the “and”, it will split Mandy into “M” and “y” because spaces are allowed but not required.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer

Partially correct (0.75/1 point)

Review

### **Question 3**

1/1 point (graded)

You have the following table, schedule:

> schedule

day staff

Monday Mandy, Chris and Laura

Tuesday Steve, Ruth and Frank

What code would successfully turn your “Schedule” table into the following tidy table?

> tidy

day staff

<chr> <chr>

Monday Mandy

Monday Chris

Monday Laura

Tuesday Steve

Tuesday Ruth

Tuesday Frank



tidy <- schedule %>%

mutate(staff = str\_split(staff, ", | and ")) %>%

unnest()



tidy <- separate(schedule, staff, into = c("s1","s2","s3"), sep = “,”) %>%

gather(key = s, value = staff, s1:s3)



tidy <- schedule %>%

mutate(staff = str\_split(staff, ", | and ", simplify = TRUE)) %>%

unnest()

correct

**Answer**

Correct:

The mutate command creates a column “staff”. Each row in the “staff” column is a character vector of length three, with the names of each staff member. We unnest this character vector using the unnest() function from tidyr.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

Show Answer

Correct (1/1 point)

Review

### **Question 4**

1/1 point (graded)

Using the gapminder data, you want to recode countries longer than 12 letters in the region “Middle Africa” to their abbreviations in a new column, “country\_short”. Which code would accomplish this?



dat <- gapminder %>% filter(region == "Middle Africa") %>%

mutate(recode(country,

"Central African Republic" = "CAR",

"Congo, Dem. Rep." = "DRC",

"Equatorial Guinea" = "Eq. Guinea"))



dat <- gapminder %>% filter(region == "Middle Africa") %>%

mutate(country\_short = recode(country,

c("Central African Republic", "Congo, Dem. Rep.", "Equatorial Guinea"),

c("CAR", "DRC", "Eq. Guinea")))



dat <- gapminder %>% filter(region == "Middle Africa") %>%

mutate(country = recode(country,

"Central African Republic" = "CAR",

"Congo, Dem. Rep." = "DRC",

"Equatorial Guinea" = "Eq. Guinea"))



dat <- gapminder %>% filter(region == "Middle Africa") %>%

mutate(country\_short = recode(country,

"Central African Republic" = "CAR",

"Congo, Dem. Rep." = "DRC",

"Equatorial Guinea" = "Eq. Guinea"))

correct

**Answer**

Correct:

This code properly recodes each country in a new column “country\_short”.

Submit

You have used 2 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

## Assessment Part 2: String Processing Part 3

 Bookmark this page

Import raw Brexit referendum polling data from Wikipedia:

library(rvest)

library(tidyverse)

library(stringr)

url <- "https://en.wikipedia.org/w/index.php?title=Opinion\_polling\_for\_the\_United\_Kingdom\_European\_Union\_membership\_referendum&oldid=896735054"

tab <- read\_html(url) %>% html\_nodes("table")

polls <- tab[[5]] %>% html\_table(fill = TRUE)

You will use a variety of string processing techniques learned in this section to reformat these data.

### **Question 5**

1/1 point (graded)

Some rows in this table do not contain polls. You can identify these by the lack of the percent sign (%) in the Remain column.

Update polls by changing the column names to c("dates", "remain", "leave", "undecided", "lead", "samplesize", "pollster", "poll\_type", "notes") and only keeping rows that have a percent sign (%) in the remain column.

How many rows remain in the polls data frame?  correct

129 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 6**

3/3 points (graded)

The remain and leave columns are both given in the format "48.1%": percentages out of 100% with a percent symbol.

Which of these commands converts the remain vector to a proportion between 0 and 1?

Check all correct answers.

as.numeric(str\_remove(polls$remain, "%"))

as.numeric(polls$remain)/100

parse\_number(polls$remain)

str\_remove(polls$remain, "%")/100

as.numeric(str\_replace(polls$remain, "%", ""))/100

parse\_number(polls$remain)/100

correct

Submit

You have used 1 of 3 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (3/3 points)

Review

### **Question 7**

3/3 points (graded)

The undecided column has some "N/A" values. These "N/A"s are only present when the remain and leave columns total 100%, so they should actually be zeros.

Use a function from **stringr** to convert "N/A" in the undecided column to 0. The format of your command should be function\_name(polls$undecided, "arg1", "arg2").

What function replaces function\_name?  correct

What argument replaces arg1?

Omit the quotation marks.

  correct

What argument replaces arg2?

Omit the quotation marks.

  correct

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (3/3 points)

Review

### **Question 8**

3.5/3.5 points (graded)

The dates column contains the range of dates over which the poll was conducted. The format is "8-10 Jan" where the poll had a start date of 2016-01-08 and end date of 2016-01-10. Some polls go across month boundaries (16 May-12 June).

The end date of the poll will always be one or two digits, followed by a space, followed by the month as one or more letters (either capital or lowercase). In these data, all month abbreviations or names have 3, 4 or 5 letters.

Write a regular expression to extract the end day and month from dates. Insert it into the skeleton code below:

temp <- str\_extract\_all(polls$dates, \_\_\_\_\_)

end\_date <- sapply(temp, function(x) x[length(x)]) # take last element (handles polls that cross month boundaries)

Which of the following regular expressions correctly extracts the end day and month when inserted into the blank in the code above?

Check all correct answers.

"\\d?\\s[a-zA-Z]?"

"\\d+\\s[a-zA-Z]+"

"\\d+\\s[A-Z]+"

"[0-9]+\\s[a-zA-Z]+"

"\\d{1,2}\\s[a-zA-Z]+"

"\\d{1,2}[a-zA-Z]+"

"\\d+\\s[a-zA-Z]{3,5}"

correct

Submit

You have used 1 of 3 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (3.5/3.5 points)

## Section 4: Dates, Time and Text Mining

## Assessment Part 1: Dates, Times, and Text Mining

 Bookmark this page

This assessment reviews several concepts about dates, times, and text mining. In part 1 on this page, you will practice extracting and manipulating dates in real datasets. In part 2 on the next page, you will walk through a sentiment analysis of a novel using steps covered in the previous section.

Use the following libraries and options for coding questions:

library(dslabs)  
library(lubridate)  
options(digits = 3) # 3 significant digits

IMPORTANT: Some of these exercises use **dslabs** datasets that were added in a July 2019 update. Make sure your package is up to date with the command install.packages("dslabs").

### **Question 1**

1/1 point (graded)

Which of the following is the standard ISO 8601 format for dates?

MM-DD-YY

YYYY-MM-DD

YYYYMMDD

YY-MM-DD

correct

**Answer**

Correct:

This is proper ISO 8601 formatting for dates.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 2**

1/1 point (graded)

Which of the following commands could convert this string into the correct date format?

dates <- c("09-01-02", "01-12-07", "02-03-04")



ymd(dates)



mdy(dates)



dmy(dates)

It is impossible to know which format is correct without additional information.

correct

**Answer**

Correct:

The formatting of these dates is ambiguous. They could be formatted as ymd, mdy or dmy. We need more information about our data to be able to select the correct command.

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 3**

2/2 points (graded)

Load the brexit\_polls data frame from **dslabs**:

data(brexit\_polls)

How many polls had a start date (startdate) in April (month number 4)?  correct

25 Loading

Use the round\_date() function on the enddate column with the argument unit="week". How many polls ended the week of 2016-06-12?

Read the documentation to learn more about round\_date().

  correct

13 Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

Review

### **Question 4**

1/1 point (graded)

Use the weekdays() function from **lubridate** to determine the weekday on which each poll ended (enddate).

On which weekday did the greatest number of polls end?

Monday

Tuesday

Wednesday

Thursday

Friday

Saturday

Sunday

correct

Submit

You have used 1 of 2 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 5**

2/2 points (graded)

Load the movielens data frame from **dslabs**.

data(movielens)

This data frame contains a set of about 100,000 movie reviews. The timestamp column contains the review date as the number of seconds since 1970-01-01 (epoch time).

Convert the timestamp column to dates using the **lubridate** as\_datetime() function.

Which year had the most movie reviews?  correct

2000 Loading

Which hour of the day had the most movie reviews?  correct

20 Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (2/2 points)

## Assessment Part 2: Dates, Times, and Text Mining

 Bookmark this page

In this part of the assessment, you will walk through a basic text mining and sentiment analysis task.

Project Gutenberg is a digital archive of public domain books. The R package **gutenbergr** facilitates the importation of these texts into R. We will combine this with the **tidyverse** and **tidytext** libraries to practice text mining.

Use these libraries and options:

library(tidyverse)  
library(gutenbergr)  
library(tidytext)  
options(digits = 3)

You can see the books and documents available in **gutenbergr** like this:

gutenberg\_metadata

### **Question 6**

1/1 point (graded)

Use str\_detect() to find the ID of the novel Pride and Prejudice.

How many different ID numbers are returned?  correct

\(\) Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 7**

1/1 point (graded)

Notice that there are several versions of the book. The gutenberg\_works() function filters this table to remove replicates and include only English language works. Use this function to find the ID for Pride and Prejudice.

What is the correct ID number?

Read the gutenberg\_works() documentation to learn how to use the function.

  correct

\(\) Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 8**

1/1 point (graded)

Use the gutenberg\_download() function to download the text for Pride and Prejudice. Use the **tidytext** package to create a tidy table with all the words in the text. Save this object as words.

How many words are present in the book?  correct

\(\) Loading

Submit

You have used 2 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 9**

1/1 point (graded)

Remove stop words from the words object. Recall that stop words are defined in the stop\_words data frame from the **tidytext** package.

How many words remain?  correct

\(\) Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 10**

1/1 point (graded)

After removing stop words, detect and then filter out any token that contains a digit from words.

How many words remain?  correct

\(\) Loading

Submit

You have used 4 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (1/1 point)

Review

### **Question 11**

3/3 points (graded)

Analyze the most frequent words in the novel after removing stop words and tokens with digits.

How many words appear more than 100 times in the book?  correct

\(\) Loading

What is the most common word in the book?  correct

How many times does that most common word appear?  correct

\(\) Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (3/3 points)

Review

### **Question 12**

3/3 points (graded)

Define the afinn lexicon:

afinn <- get\_sentiments("afinn")

Note that this command will trigger a question in the R Console asking if you want to download the AFINN lexicon. Press 1 to select "Yes" (if using RStudio, enter this in the Console tab).

Use this afinn lexicon to assign sentiment values to words. Keep only words that are present in both words and the afinn lexicon. Save this data frame as afinn\_sentiments.

How many elements of words have sentiments in the afinn lexicon?  correct

\(\) Loading

What proportion of words in afinn\_sentiments have a positive value?  correct

\(\) Loading

How many elements of afinn\_sentiments have a value of 4?  correct

\(\) Loading

Submit

You have used 1 of 10 attemptsSome problems have options such as save, reset, hints, or show answer. These options follow the Submit button.

SaveSave Your Answer Show Answer

Correct (3/3 points)

## 