Branch: master ▾ Find file Copy path

cleaningdataR / week2_quiz_Baker.R

aurquhart Adam B answers

8da54df on Oct 20, 2014

1 contributor

Raw Blame History

111 lines (80 sloc) 3.89 KB

```
# Remove everything from the workspace
rm(list = ls())

# Windows
setwd('C://Users//ABaker//Documents//GitHub//Coursera//Getting and Cleaning Data')
# OS X
setwd('/Users/adam_baker_1/GitHub/Coursera/Getting and Cleaning Data')

if (!file.exists("data")) {
  dir.create("data")
}

# Question 1 - What time was the data sharing repo created?

library(httr)
library(httpuv)
library(jsonlite)
library(dplyr)

# 1. Find OAuth settings for github:
#    http://developer.github.com/v3/oauth/
github <- oauth_endpoints("github")


# 2. Register an application at https://github.com/settings/applications;
#    Use any URL you would like for the homepage URL (http://github.com is fine)
#    and http://localhost:1410 as the callback url
#
#    Insert your client ID and secret below - if secret is omitted, it will
#    look it up in the GITHUB_CONSUMER_SECRET environmental variable.
myapp <- oauth_app("github", "9e975720d681777b66d4", "b2a99834bf1543521472d02fd031ed022c63e983")

# 3. Get OAuth credentials
github_token <- oauth2.0_token(oauth_endpoints("github"), myapp)

# 4. Use API
gtoken <- config(token = github_token)
req <- GET("https://api.github.com/users/jtleek/repos", gtoken)
stop_for_status(req)
content(req)

json1 = content(req)
```

```r
json2 = jsonlite::fromJSON(toJSON(json1))
head(json2)

json2[json2$full_name == "jtleek/datasharing",] # 2013-11-07T13:25:07Z

# Question 2 - Which of the following commands will select only the data for the probability weights pwgtp1 wit
# sqldf("select * from acs where AGEP < 50")
# sqldf("select pwgtp1 from acs where AGEP < 50")
# sqldf("select pwgtp1 from acs")
# sqldf("select * from acs")

library(sqldf)

# Using method = "curl" doesn't seem to work on Windows
download.file("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv", destfile = "./data/acs_wk2

# Using method = "curl" on OS X works
download.file("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv", destfile = "./data/acs_wk2

acs <- read.csv("./data/acs_wk2.csv")
head(acs)

sqldf("select * from acs where AGEP < 50")          # NO - selects all data for ages less than 50
sqldf("select pwgtp1 from acs where AGEP < 50")     # YES - selects the correct column and for ages less than 5
sqldf("select pwgtp1 from acs")                     # NO - only selects pwgtp1 column
sqldf("select * from acs")                          # NO - selects everything

# Question 3 - Using the same data frame you created in the previous problem, what is the equivalent function t

unique(acs$AGEP) # this is the desired result

sqldf("select unique AGEP from acs")                # NO - syntax error
sqldf("select distinct AGEP from acs")              # YES? - generates the same list but in a different format
sqldf("select distinct pwgtp1 from acs")            # NO - selects the distinct values of the wrong column
sqldf("select unique * from acs")                   # NO - syntax error

# Question 4 - How many characters are in the 10th, 20th, 30th and 1000th lines of HTML from this page

con = url("http://biostat.jhsph.edu/~jleek/contact.html ")
htmlCode = readLines(con)
close(con)
htmlCode

nchar(htmlCode[10]) # 45
nchar(htmlCode[20]) # 31
nchar(htmlCode[30]) # 7
nchar(htmlCode[100]) # 25

# Answer = 45 31 7 25

# Question 5 - Read the data set into R and report the sum of the numbers in the fourth of the nine columns

# Downloading data from the Web
fileUrl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fwksst8110.for"

# Using method = "curl" doesn't seem to work on Windows
download.file(fileUrl, destfile = "./data/ac_survey.csv")
```

```r

# Using method = "curl" on OS X works
download.file(fileUrl, destfile = "./data/q5_data.for", method = "curl")
list.files('./data/')

?read.fwf

q5_df <- read.fwf(file = "./data/q5_data.for", widths = c(15, 4, 1, 3, 5, 4), header = FALSE, sep = "\t", skip
head(q5_df)

# Need to sum up the V6 column
sum(q5_df$V6) # 32426.7
```