

## Getting and Cleaning Data - Week 2 Quiz

```
library(httr)
library(jsonlite)

# Question #1

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

## <oauth_endpoint>
##   authorize: https://github.com/login/oauth/authorize
##   access:   https://github.com/login/oauth/access_token

# B. To make your own application, register at
#   https://github.com/settings/developers. Use any URL for the homepage URL
#   (http://github.com is fine) and http://localhost:1410 as the callback url
#
# Authenticate with key and secret below.
myapp <- oauth_app("github",
  key = "ca3a9589ef2819e1566f",
  secret = "e492571c91a14f0e4dc9373c48b102cf22fadcfb"
)

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

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

#E. Get names of Prof. Leek's repositories
jsondata$name

## [1] "2018" "ads2020"
## [3] "advdatasci" "advdatasci-project"
## [5] "advdatasci-swirl" "advdatasci15"
## [7] "advdatasci16" "advdatasci_swirl"
## [9] "ballgown" "big_course"
## [11] "bookdown-start" "capitalIn21stCenturyinR"
## [13] "careerplanning" "COVID-19"
## [15] "crsra" "cshlcbg-labs"
## [17] "dataanalysis" "datascientist"
## [19] "datasharing" "datawomenontwitter"
## [21] "day1" "derfinder"
## [23] "derfinder-1" "DSM"
```

```
## [25] "EDA-Project"          "escalatr"
## [27] "euclideo"             "firstpaper"
## [29] "futureofstats"        "gcd"

jsondata[19,47]

## [1] "2013-11-07T13:25:07Z"

# Question #2

# Download the .csv file and assign to the variable 'acs'

download.file("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv", "acs.csv")
acs <- read.csv("acs.csv")

library(sqldf)

## Loading required package: gsubfn
## Loading required package: proto
## Warning in doTryCatch(return(expr), name, parentenv, handler): unable to load shared object '/Library
## dlopen(/Library/Frameworks/R.framework/Resources/modules/R_X11.so, 6): Library not loaded: /opt/X
## Referenced from: /Library/Frameworks/R.framework/Versions/3.6/Resources/modules/R_X11.so
## Reason: image not found
## Could not load tcltk. Will use slower R code instead.
## Loading required package: RSQLite

head(sqldf("select pwgtp1 from acs where AGE<50"))

##   pwgtp1
## 1      87
## 2      88
## 3      94
## 4      91
## 5     539
## 6     192

# Question #3

head(sqldf("select distinct AGE from acs"))

##   AGE
## 1  43
## 2  42
## 3  16
## 4  14
## 5  29
## 6  40

# Question #4

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

num_chars <- nchar(htmlCode)
num_chars[c(10,20,30,100)]
```

```
## [1] 45 31 7 25
```

```
# Question #5
```

```
download.file("https://d396qusza40orc.cloudfront.net/getdata%2Fwksst8110.for","data.csv")
data <- read.fwf("data.csv",c(10,9,4,9),skip=4)
head(data)
```

```
##           V1    V2    V3    V4
## 1 03JAN1990 23.4 -0.4 25.1
## 2 10JAN1990 23.4 -0.8 25.2
## 3 17JAN1990 24.2 -0.3 25.3
## 4 24JAN1990 24.4 -0.5 25.5
## 5 31JAN1990 25.1 -0.2 25.8
## 6 07FEB1990 25.8  0.2 26.1
```

```
sum(data$V4)
```

```
## [1] 32426.7
```