**uestion 1**

The American Community Survey distributes downloadable data about United States communities. Download the 2006 microdata survey about housing for the state of Idaho using download.file() from here:

<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv>

and load the data into R. The code book, describing the variable names is here:

<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FPUMSDataDict06.pdf>

How many properties are worth $1,000,000 or more?

**R script:**

|  |
| --- |
| ## Question 1:  if (!file.exists("data")) {  dir.create("data") } fileUrl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv" download.file(fileUrl, destfile = "./data/06hid.csv", method = "curl") dateDownloaded <- date()  HD <- read.csv("./data/06hid.csv") sum(!is.na(HD[HD$VAL >= 24, 37])) |

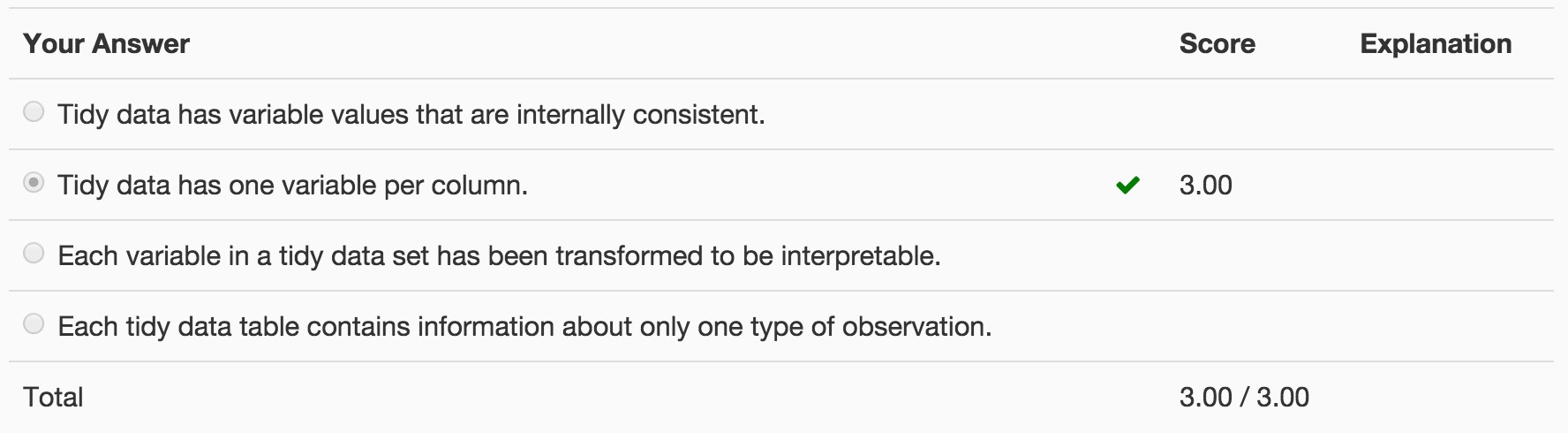
**Result:**

[](http://qiniupicbed.qiniudn.com/upload/8b81bcebf6bf74304d07c7e0317a5965.png)

**Question 2**

Use the data you loaded from Question 1. Consider the variable FES in the code book. Which of the “tidy data” principles does this variable violate?

**Result:**

[](http://qiniupicbed.qiniudn.com/upload/c3d80b846d3225e2dd07c5af9119ec37.png)

**Question 3**

Download the Excel spreadsheet on Natural Gas Aquisition Program here:

<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FDATA.gov_NGAP.xlsx>

Read rows 18-23 and columns 7-15 into R and assign the result to a variable called:

|  |
| --- |
| dat |

What is the value of:

|  |
| --- |
| sum(dat$Zip \* dat$Ext, na.rm = T) |

(original data source: <http://catalog.data.gov/dataset/natural-gas-acquisition-program>)

**R script:**

|  |
| --- |
| ## Question 3:  if (!file.exists("data")) {  dir.create("data") } fileUrl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FDATA.gov\_NGAP.xlsx" download.file(fileUrl, destfile = "./data/gas.xlsx", method = "curl") dateDownloaded <- date()  library(xlsx) rowIndex = 18 : 23 colIndex = 7 : 15 dat <- read.xlsx("./data/gas.xlsx", sheetIndex = 1, rowIndex = rowIndex,   colIndex = colIndex, header = TRUE) sum(dat$Zip \* dat$Ext, na.rm=T) |

**Result:**

[](http://qiniupicbed.qiniudn.com/upload/5339361b506afa8ce1379711259e263c.png)

**Question 4**

Read the XML data on Baltimore restaurants from here:

<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml>

How many restaurants have zipcode 21231?

**R script:**

|  |
| --- |
| ## Question 4:  # Method 1: remove the letter "s" from "https", and use the "http" instead library(XML) fileUrl <- "http://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml" doc <- xmlTreeParse(fileUrl, useInternal = TRUE) rootNode <- xmlRoot(doc)  sum(xpathSApply(rootNode, "//zipcode", xmlValue) == "21231")  # Method 2: download the XML file first, and then load it into R if (!file.exists("data")) {  dir.create("data") } fileUrl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml" download.file(fileUrl, destfile = "./data/restaurants.xml", method = "curl") dateDownloaded <- date()  library(XML) doc <- xmlTreeParse("./data/restaurants.xml", useInternal = TRUE) rootNode <- xmlRoot(doc)  sum(xpathSApply(rootNode, "//zipcode", xmlValue) == "21231") |

**Result:**

[](http://qiniupicbed.qiniudn.com/upload/a9b215adf99faef0fa41fc1ffe616dd6.png)

**Question 5**

The American Community Survey distributes downloadable data about United States communities. Download the 2006 microdata survey about housing for the state of Idaho using download.file() from here:

<https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv>

using the fread() command load the data into an R object

|  |
| --- |
| DT |

Which of the following is the fastest way to calculate the average value of the variable

|  |
| --- |
| pwgtp15 |

broken down by sex using the data.table package?

# download file from server

download.file(url = "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv",

destfile = "q5.csv",

method = "curl")

# load library

library(data.table)

# read csv file using fread

DT <- fread("q5.csv", header = TRUE)

# best way to calculate mean by sex

DT[,mean(pwgtp15),by=SEX]