Getting and Cleaning Data - Data Science - Quiz 1 - Coursera

Getting and Cleaning Data Quiz 1

This is Quiz 1 from the Getting and Cleaning Data course within the Data Science Specialization on Coursera. Topics include reading XML, excel files, and extracting data.

Questions

---

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?

- **53**

```
download.file("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv", destf
ile = "quiz1data.csv")

data <- read.csv("quiz1data.csv")

nrow(data[which(data$VAL == 24),])

## [1] 53
```

---

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?

- **Tidy data has one variable per column.**

Explanation:

FES 1 Family type and employment status b .N/A (GQ/vacant/not a family) 1 .Married-couple family: Husband and wife in LF 2 .Married-couple family: Husband in labor force, wife .not in LF 3 .Married-couple family: Husband not in LF, .wife in LF 4 .Married-couple family: Neither husband nor wife in .LF 5 .Other family: Male householder, no wife present, in .LF 6 .Other family: Male householder, no wife present, .not in LF 7 .Other family: Female householder, no husband .present, in LF 8 .Other family: Female householder, no husband .present, not in LF

---

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)

- **36534720**

```
require(xlsx)

## Loading required package: xlsx

## Loading required package: rJava

## Loading required package: xlsxjars

download.file("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FDATA.gov_NGAP.xlsx"
, destfile = "quiz1data2.xlsx")


row <- 18:23

col <- 7:15


dat <- read.xlsx("quiz1data2.xlsx", sheetIndex = 1, colIndex = col, rowIndex = row, heade
r = TRUE)


head(dat)

##     Zip CuCurrent PaCurrent PoCurrent       Contact Ext          Fax email

## 1 74136         0         1         0 918-491-6998   0 918-491-6659    NA

## 2 30329         1         0         0 404-321-5711  NA         <NA>    NA

## 3 74136         1         0         0 918-523-2516   0 918-523-2522    NA
## 4 80203         0         1         0 303-864-1919   0         <NA>    NA
## 5 80120         1         0         0 345-098-8890 456         <NA>    NA
##   Status
## 1      1
## 2      1
## 3      1
## 4      1
## 5      1
```

---

4. Read the XML data on Baltimore restaurants from here:

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

How many restaurants have zipcode 21231?

- **127**

```
library(XML)
URL<-"http://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml"
doc <- xmlTreeParse(URL, useInternal = TRUE)
rootNode <- xmlRoot(doc)
xmlName(rootNode)
## [1] "response"
zips <- xpathSApply(rootNode, "//zipcode", xmlValue)

length(zips[which(zips=="21231")])
## [1] 127
```

---

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
##The following are ways to calculate the average value of the variable
```

The following are ways to calculate the average value of the variable

```
##pwgtp15

##broken down by sex. Using the data.table package, which will deliver the fastest user time?
```

broken down by sex. Using the data.table package, which will deliver the fastest user time?

```
download.file("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv", destfile="quiz1data4.csv")


library(data.table)

DT <- fread(input="quiz1data4.csv", sep=",")


system.time(mean(DT$pwgtp15,by=DT$SEX))
```

```
##    user  system elapsed

##   0.001   0.000   0.000
```

```
system.time(tapply(DT$pwgtp15,DT$SEX,mean))
```

```
##    user  system elapsed

##   0.001   0.000   0.001
```

```
system.time(sapply(split(DT$pwgtp15,DT$SEX),mean))
```

```
##    user  system elapsed

##   0.001   0.000   0.000
```

```
system.time(DT[,mean(pwgtp15),by=SEX])
```

```
##    user  system elapsed

##   0.004   0.001   0.005
```

```
system.time(mean(DT[DT$SEX==1,]$pwgtp15)) + system.time(mean(DT[DT$SEX==2,]$pwgtp15))
```

```
##    user  system elapsed

##   0.025   0.001   0.027
```
