Getting and Cleaning Data Quiz 1 (JHU) Coursera

Question 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 housing units in this survey were worth more than $1,000,000?

```
# fread url requires curl package on mac
# install.packages("curl")

library(data.table)
housing <- data.table::fread("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid

# VAL attribute says how much property is worth, .N is the number of rows
# VAL == 24 means more than $1,000,000
housing[VAL == 24, .N]

# Answer:
# 53
```

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?

Answer

Tidy data one variable per column

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)

```
fileUrl <- "http://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FDATA.gov_NGAP.xlsx"
download.file(fileUrl, destfile = paste0(getwd(), '/getdata%2Fdata%2FDATA.gov_NGAP.xlsx'), m

dat <- xlsx::read.xlsx(file = "getdata%2Fdata%2FDATA.gov_NGAP.xlsx", sheetIndex = 1, rowInde
sum(dat$Zip*dat$Ext,na.rm=T)

# Answer:
# 36534720
```

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?

Use http instead of https, which caused the message Error: XML content does not seem to be XML: 'https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml'.

```
# install.packages("XML")
library("XML")
fileURL<-"https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml"
doc <- XML::xmlTreeParse(sub("s", "", fileURL), useInternal = TRUE)
rootNode <- XML::xmlRoot(doc)

zipcodes <- XML::xpathSApply(rootNode, "//zipcode", XML::xmlValue)
xmlZipcodeDT <- data.table::data.table(zipcode = zipcodes)
xmlZipcodeDT[zipcode == "21231", .N]

# Answer:
# 127
```

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?

```
DT <- data.table::fread("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv"

# Answer (fastest):
system.time(DT[,mean(pwgtp15),by=SEX])
```