Getting and Cleaning Data: Quiz 1

Author: Sherri Verdugo Date: July 12, 2014
========== This is the Quiz 1 Attempt for the Getting and Cleaning Data. John Hopkins University and part of the data science course series.

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 (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 (https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FPUMSDataDict06.pdf)

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

A. 47
B. 53
C. 164
D. 31

```
idaho_h <- read.csv("getdata-data-ss06hid.csv")
head(idaho_h, 2)
```

```
##   RT SERIALNO DIVISION PUMA REGION ST  ADJUST WGTP NP TYPE ACR AGS BDS BLD
## 1  H      186        8  700      4 16 1015675   89  4    1   1  NA   4   2
## 2  H      306        8  700      4 16 1015675  310  1    1  NA  NA   1   7
##   BUS CONP ELEP FS FULP GASP HFL INSP KIT MHP MRGI MRGP MRGT MRGX PLM RMS
## 1   2   NA  180  0    2    3   3  600   1  NA    1 1300    1    1   1   9
## 2  NA   NA   60  0    2    3   3   NA   1  NA   NA   NA   NA   NA   1   2
##   RNTM RNTP SMP TEL TEN VACS VAL VEH WATP YBL FES  FINCP FPARC GRNTP GRPIP
## 1   NA   NA  NA   1   1   NA  17   3  840   5   2 105600     2    NA    NA
## 2    2  600  NA   1   3   NA  NA   1    1   3  NA     NA    NA   660    23
##   HHL HHT  HINCP HUGCL HUPAC HUPAOC HUPARC LNGI MV NOC NPF NPP NR NRC
## 1   1   1 105600     0     2      2      2    1  4   2   4   0  0   2
## 2   1   4  34000     0     4      4      4    1  3   0  NA   0  0   0
##   OCPIP PARTNER PSF R18 R60 R65 RESMODE SMOCP SMX SRNT SVAL TAXP WIF
## 1    18       0   0   1   0   0       1  1550   3    0    1   24   3
## 2    NA       0   0   0   0   0       2    NA  NA    1    0   NA  NA
##   WKEXREL WORKSTAT FACRP FAGSP FBDSP FBLDP FBUSP FCONP FELEP FFSP FFULP
## 1       2        3     0     0     0     0     0     0     0    0     0
## 2      NA       NA     0     0     0     0     0     0     0    0     0
##   FGASP FHFLP FINSP FKITP FMHP FMRGIP FMRGP FMRGTP FMRGXP FMVYP FPLMP
## 1     0     0     0     0    0      0     0      0      0     0     0
## 2     0     0     0     0    0      0     0      0      0     0     0
##   FRMSP FRNTMP FRNTP FSMP FSMXHP FSMXSP FTAXP FTELP FTENP FVACSP FVALP
## 1     0      0     0    0      0      0     0     0     0      0     0
## 2     0      0     0    0      0      0     0     0     0      0     0
##   FVEHP FWATP FYBLP wgtp1 wgtp2 wgtp3 wgtp4 wgtp5 wgtp6 wgtp7 wgtp8 wgtp9
## 1     0     0     0    87    28   156    95    26    25    95    93    93
## 2     0     0     1   539   363   293   422   566   289    87   242   453
##   wgtp10 wgtp11 wgtp12 wgtp13 wgtp14 wgtp15 wgtp16 wgtp17 wgtp18 wgtp19
## 1     91     87    166     90     25    153     89    148     82     25
## 2    453    334    358    414    102    281     99    108    278    131
##   wgtp20 wgtp21 wgtp22 wgtp23 wgtp24 wgtp25 wgtp26 wgtp27 wgtp28 wgtp29
## 1    180     90     24    140     92     25     27     86     84     87
## 2    407    447    264    352    238    390    336    122    374    482
##   wgtp30 wgtp31 wgtp32 wgtp33 wgtp34 wgtp35 wgtp36 wgtp37 wgtp38 wgtp39
## 1     93     90    149     91     28    143     81    144     95     27
## 2    468    335    251    613    104    284    116     91    326    102
##   wgtp40 wgtp41 wgtp42 wgtp43 wgtp44 wgtp45 wgtp46 wgtp47 wgtp48 wgtp49
## 1     22     90    171     27     83    153    148     92     91     91
## 2    361    107    253    321    289     96    343    564    274    118
##   wgtp50 wgtp51 wgtp52 wgtp53 wgtp54 wgtp55 wgtp56 wgtp57 wgtp58 wgtp59
## 1     93     90     26     94    142     24     91     29     84    148
## 2    118    321    261    130    463    294    479    391    307    476
##   wgtp60 wgtp61 wgtp62 wgtp63 wgtp64 wgtp65 wgtp66 wgtp67 wgtp68 wgtp69
## 1     30     93    143     24     88    147    145     91     83     83
## 2    283    116    353    323    374    106    236    380    313     90
##   wgtp70 wgtp71 wgtp72 wgtp73 wgtp74 wgtp75 wgtp76 wgtp77 wgtp78 wgtp79
## 1     86     81     27     93    151     28     79     25    101    157
## 2     94    292    401     81    494    346    496    615    286    454
##   wgtp80
## 1    129
## 2    260
```

```
length(idaho_h$VAL[!is.na(idaho_h$VAL) & idaho_h$VAL==24])
```

```
## [1] 53
```

The answer is: 53 housing units in this survey were worth more than $1,000,000.

Question 2

Using the data from question 1. Consider the var FES in the codebook. Which of the “tidy data” principles does this variable violate?

```
idaho_h <- read.csv("getdata-data-ss06hid.csv")
table(idaho_h$FES)
```

```
##
##    1    2    3    4    5    6    7    8
## 1730  826  236  638  151   40  305  125
```

```
summary(idaho_h$FES)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     1.0     1.0     2.0     2.7     4.0     8.0    2445
```

```
idaho_h$FES[1:5]
```

```
## [1]  2 NA  7  1  1
```

The answer is: tidy data has one variable per column… FES has: gender, marital status and empoloyement status.

Question 3

Download the Excel spreadsheet on Natural Gas Aquisition Program here:

https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FDATA.gov_NGAP.xlsx (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 (http://catalog.data.gov/dataset/natural-gas-acquisition-program))

A. 154339
B. 0
C. NA
D. 36534720

```
library(xlsx)
```

```
## Loading required package: rJava
## Loading required package: xlsxjars
```

```
# Start and End row: 18 23
rowIndex <- 18:23
colIndex <- 7:15
dat <- read.xlsx(file="gov_NGAP.xlsx", sheetIndex=1, colIndex=colIndex, rowIndex=rowIndex, header=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
```

```
sum(dat$Zip*dat$Ext, na.rm=T)
```

```
## [1] 36534720
```

The answer is D) 36534720

Question 4

Read the XML data on Baltimore restaurants from here:

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

How many restaurants have zipcode 21231?

A. 127
B. 100
C. 17
D. 130

```
library(XML)
file <- "http://d396qusza40orc.cloudfront.net/getdata/data/restaurants.xml"
my.doc <- xmlTreeParse(file=file,useInternal=TRUE)
root.Node <- xmlRoot(my.doc)
xmlName(root.Node)
```

```
## [1] "response"
```

```
zipcode <- xpathSApply(root.Node, "//zipcode", xmlValue)
length(zipcode[zipcode==21231])
```

```
## [1] 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 (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?

A. mean(DT\(pwgtp15,by=DT\)SEX)
B. tapply(DT\(pwgtp15,DT\)SEX,mean)
C. mean(DT[DT$SEX==1,]\(pwgtp15); mean(DT[DT\)SEX==2,]$pwgtp15)
D. rowMeans(DT)[DT$SEX==1]; rowMeans(DT)[DT$SEX==2]
E. DT[,mean(pwgtp15),by=SEX]
F. sapply(split(DT\(pwgtp15,DT\)SEX),mean)

```
library(data.table)
DT <- fread(input="getdata-data-ss06pid.csv", sep=",")
system.time(mean(DT$pwgtp15,by=DT$SEX))
```

```
##    user  system elapsed
##       0       0       0
```

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

```
##    user  system elapsed
##   0.002   0.001   0.003
```

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

```
##    user  system elapsed
##   0.043   0.005   0.050
```

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

```
##    user  system elapsed
##   0.001   0.000   0.001
```

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

```
##    user  system elapsed
##   0.003   0.001   0.005
```

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

```
##    user  system elapsed
##   0.002   0.000   0.001
```