Reading Files

Vignesh

6/26/2020

Getting and Cleaning Data - Week 1 Quiz- Solutions

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?

Solution: 53

```r
fileURL <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv"
download.file(fileURL, destfile = "community-survery.csv")
dateDownloaded <- date()
dateDownloaded
```

```
## [1] "Sat Jun 27 10:01:23 2020"
```

```r
data <- read.csv("community-survery.csv")
sum(data$VAL == 24, na.rm = TRUE)
```

```
## [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?

Solution: Tidy data has one variable per column

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(datZip * dat$Ext,na.rm=T)

Solution: I tried working on the solution but every time I ran library(xlsx) function I got the following error message. Please suggest a way around this error or a resolution, will be grateful of you!!

> library(xlsx) Error: package or namespace load failed for ‘xlsx’: .onLoad failed in loadNamespace() for ‘rJava’, details: call: fun(libname, pkgname) error: JAVA_HOME cannot be determined from the Registry In addition: Warning message: package ‘xlsx’ was built under R version 4.0.2

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
library(XML)
fileURLBalti <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml"
fileURLBalti
```

```
## [1] "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml"
```

```r
BaltiResto <- xmlTreeParse(sub("s", "", fileURLBalti), useInternal=TRUE)
rootNode <- xmlRoot(BaltiResto)
zip <- xpathSApply(rootNode, "//zipcode", xmlValue)
sum(zip == 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

pwgtp15dat <- openxlsx::read.xlsx(file = fileNGAP)

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

```
##         RT SERIALNO SPORDER PUMA ST  ADJUST PWGTP AGEP CIT COW DDRS DEYE DOUT
##     1:  P      186       1  700 16 1015675    89   43   1   7    2    2    2
##     2:  P      186       2  700 16 1015675    92   42   1   4    2    2    2
##     3:  P      186       3  700 16 1015675   107   16   1   1    2    2    2
##     4:  P      186       4  700 16 1015675    91   14   1  NA    2    2   NA
##     5:  P      306       1  700 16 1015675   309   29   1   5    2    2    2
##    ---                                                                       
## 14927:  P  1357874       2  900 16 1015675    28   74   1   2    2    2    2
## 14928:  P  1357880       1  500 16 1015675   121   22   1   1    2    2    2
## 14929:  P  1357880       2  500 16 1015675   112   22   1   1    2    2    2
## 14930:  P  1358490       1  700 16 1015675   353   28   1   1    2    2    2
## 14931:  P  1358490       2  700 16 1015675   386   23   1   1    2    2    2
##        DPHY DREM DWRK ENG FER GCL GCM GCR INTP JWMNP JWRIP JWTR LANX MAR MIG
##     1:    2    2    2  NA  NA   2  NA  NA    0    15     1    1    2   1   1
##     2:    2    2    2  NA   2   2  NA  NA    0    NA    NA   NA    2   1   1
##      3:    2    2    2  NA  NA  NA  NA  NA    0     5     1    1    2   5   1
##      4:    2    2   NA  NA  NA  NA  NA  NA   NA    NA    NA   NA    2   5   1
##      5:    2    2    2  NA  NA  NA  NA  NA    0    50     8    1    2   5   1
##    ---
## 14927:    1    2    2  NA  NA   2  NA  NA    0     5     1    1    2   1   1
## 14928:    2    2    2  NA   2  NA  NA  NA    0     2     1    1    2   1   1
## 14929:    2    2    2  NA  NA  NA  NA  NA    0    30     1    1    2   1   1
## 14930:    2    2    2  NA  NA  NA  NA  NA    0     2     1    1    2   5   1
## 14931:    2    2    2  NA  NA  NA  NA  NA    0    20     1    1    2   5   1
##         MIL MILY MLPA MLPB MLPC MLPD MLPE MLPF MLPG MLPH MLPI MLPJ MLPK NWAB
##      1:   3    2    0    0    1    0    0    0    0    0    0    0    0    3
##      2:   5   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA    2
##      3:  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA    3
##      4:  NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA
##      5:   2    2    1    0    0    0    0    0    0    0    0    0    0    3
##    ---
## 14927:   5   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA    2
## 14928:   5   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA    2
## 14929:   5   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA    2
## 14930:   5   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA    3
## 14931:   5   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA   NA    3
##         NWAV NWLA NWLK NWRE OIP PAP REL RETP SCH SCHG SCHL  SEMP SEX SSIP SSP
##      1:    5    3    3    3   0   0   0    0   1   NA   10 50000   1    0   0
##      2:    5    2    2    3   0   0   1    0   1   NA    9     0   2    0   0
##      3:    5    3    3    3   0   0   2    0   3    5    7     0   1    0   0
##      4:   NA   NA   NA   NA  NA  NA   2   NA   3    4    4    NA   2   NA  NA
##      5:    5    3    3    3   0   0   0    0   1   NA   12     0   1    0   0
##    ---
## 14927:    5    2    2    3   0   0   1    0   1   NA    7     0   2    0   0
## 14928:    5    2    2    3   0   0   0    0   3    6   11     0   2    0   0
## 14929:    5    2    2    3   0   0   1    0   1   NA    9     0   1    0   0
## 14930:    5    3    3    3   0   0   0    0   1   NA   11     0   1    0   0
## 14931:    5    3    3    3   0   0   3    0   1   NA   10     0   1    0   0
##          WAGP WKHP WKL WKW YOEP UWRK ANC ANC1P ANC2P DECADE DRIVESP DS ESP ESR
##      1: 50000   50   1  52   NA    1   2   920   148     NA       1  2  NA   1
##      2:   800    4   1  20   NA    2   1   920   999     NA      NA  2  NA   6
##      3:  4800   20   1  52   NA    1   2   920   148     NA       1  2   2   1
##      4:    NA   NA  NA  NA   NA   NA   1   920   999     NA      NA  2   2  NA
##      5: 34000   50   1  52   NA    1   2   902   920     NA       6  2  NA   1
##    ---
## 14927: 12000   25   1  52   NA    1   1   939   999     NA       1  1  NA   1
## 14928:  5000   30   1  52   NA    1   2    50    32     NA       1  2  NA   1
## 14929: 28000   40   1  52   NA    1   2    50    22     NA       1  2  NA   1
## 14930: 13600   40   1  50   NA    1   4   999   999     NA       1  2  NA   1
## 14931: 19000   40   1  52   NA    1   4   999   999     NA       1  2  NA   1
##         HISP INDP JWAP JWDP LANP MIGPUMA MIGSP MSP NAICSP NATIVITY NOP OC OCCP
##      1:    1 7690   88   46   NA      NA    NA   1  5617Z        1  NA  0 4200
##      2:    1 7870   NA   NA   NA      NA    NA   1  611M1        1  NA  0 2340
##      3:    1 8680  200  119   NA      NA    NA   6   722Z        1   1  1 4020
##      4:    1   NA   NA   NA   NA      NA    NA  NA               1   1  1   NA
##      5:    1 9590   72   23   NA      NA    NA   6   928P        1  NA  0 7140
##    ---
## 14927:    1 8680   81   41   NA      NA    NA   1   722Z        1  NA  0 4020
## 14928:    1 8270   82   43   NA      NA    NA   1   6231        1  NA  0 3600
## 14929:    1 4470  130   82   NA       NA    NA   1   4244        1  NA  0 9610
## 14930:    1 8680   45   12   NA       NA    NA   6   722Z        1  NA  0 4060
## 14931:    1 4870   92   49   NA       NA    NA   6  4441Z        1  NA  0 4700
##         PAOC  PERNP  PINCP POBP POVPIP POWPUMA POWSP QTRBIR RAC1P RAC2P RAC3P
##      1:   NA 100000 100000   53    501     600    16      3     1     1    69
##      2:    2    800    800   41    501      NA    NA      3     1     1    69
##      3:   NA   4800   4800   16    501     600    16      2     1     1    69
##      4:   NA     NA     NA   41    501      NA    NA      4     1     1    69
##      5:   NA  34000  34000   36    333     400    16      1     9    67    43
##    ---
## 14927:    4  12000  12000   16    152     900    16      2     1     1    69
## 14928:    4   5000   5000   16    245     500    16      1     1     1    69
## 14929:   NA  28000  28000   41    245     700    16      1     1     1    69
## 14930:   NA  13600  13600   16    246     600    16      4     1     1    69
## 14931:   NA  19000  19000   41    246     700    16      3     1     1    69
##         RACAIAN RACASN RACBLK RACNHPI RACNUM RACSOR RACWHT RC SFN SFR   SOCP VPS
##      1:       0      0      0       0      1      0      1  0  NA  NA 371011   9
##      2:       0      0      0       0      1      0      1  0  NA  NA 253000  NA
##      3:       0      0      0       0      1      0      1  1  NA  NA 352010  NA
##      4:       0      0      0       0      1      0      1  1  NA  NA         NA
##      5:       1      0      1       0      2      0      0  0  NA  NA 493011   1
##    ---
## 14927:       0      0      0       0      1      0      1  0  NA  NA 352010  NA
## 14928:       0      0      0       0      1      0      1  0  NA  NA 311010  NA
## 14929:       0      0      0       0      1      0      1  0  NA  NA 537061  NA
## 14930:       0      0      0       0      1      0      1  0  NA  NA 353022  NA
## 14931:       0      0      0       0      1      0      1  0  NA  NA 411011  NA
##         WAOB FAGEP FANCP FCITP FCOWP FDDRSP FDEYEP FDOUTP FDPHYP FDREMP FDWRKP
##      1:    1     0     0     0     0      0      0      0      0      0      0
##      2:    1     0     0     0     0      0      0      0      0      0      0
##      3:    1     0     0     0     0      0      0      0      0      0      0
##      4:    1     0     0     0     0      0      0      0      0      0      0
##      5:    1     0     0     0     0      0      0      0      0      0      0
##    ---
## 14927:    1     0     0     0     0      0      0      0      1      0      1
## 14928:    1     0     0     0     0      0      0      0      0      0      0
## 14929:    1     0     0     0     0      0      0      0      0      0      0
## 14930:    1     0     0     0     0      0      0      0      0      0      0
## 14931:    1     0     0     0     0      0      0      0      0      0      0
##         FENGP FESRP FFERP FGCLP FGCMP FGCRP FHISP FINDP FINTP FJWDP FJWMNP
##      1:     0     0     0     0     0     0     0     0     0     0      0
##      2:     0     0     0     0     0     0     0     0     0     0      0
##      3:     0     0     0     0     0     0     0     0     0     0      0
##      4:     0     0     0     0     0     0     0     0     0     0      0
##      5:     0     0     0     0     0     0     0     0     0     0      0
##    ---
## 14927:     0     0     0     0     0     0     1     0     1     0      0
## 14928:     0     0     0     0     0     0     0     0     0     0      0
## 14929:     0     0     0     0     0     0     0     0     0     0      0
## 14930:     0     0     0     0     0     0     0     0     0     0      0
## 14931:     0     0     0     0     0     0     0     0     0     0      0
##         FJWRIP FJWTRP FLANP FLANXP FMARP FMIGP FMIGSP FMILPP FMILSP FMILYP FOCCP
##      1:      0      0     0      0     0     0      0      0      0      0     0
##      2:      0      0     0      0     0     0      0      0      0      0     0
##     3:      0      0     0      0     0     0      0      0      0      0     0
##     4:      0      0     0      0     0     0      0      0      0      0     0
##     5:      0      0     0      0     0     0      0      0      0      0     0
##    ---
## 14927:      0      0     0      0     0     0      0      0      0      0     0
## 14928:      0      0     0      0     0     0      0      0      0      0     0
## 14929:      0      0     0      0     0     0      0      0      0      0     0
## 14930:      0      0     0      0     0     0      0      0      0      0     0
## 14931:      0      0     0      0     0     0      0      0      0      0     0
##         FOIP FPAP FPOBP FPOWSP FRACP FRELP FRETP FSCHGP FSCHLP FSCHP FSEMP FSEXP
##     1:    0    0     0      0     0     0     0      0      0     0     0     0
##     2:    0    0     0      0     0     0     0      0      0     0     0     0
##     3:    0    0     0      0     0     0     0      0      0     0     0     0
##     4:    0    0     0      0     0     0     0      0      0     0     0     0
##     5:    0    0     0      0     0     0     0      0      0     0     0     0
##    ---
## 14927:    1    1     0      0     1     0     1      0      0     0     0     0
## 14928:    0    0     0      0     0     0     0      0      0     0     0     0
## 14929:    0    0     0      0     0     0     0      0      0     0     0     0
## 14930:    0    0     0      0     0     0     0      0      0     0     0     0
## 14931:    0    0     0      0     0     0     0      0      0     0     0     0
##         FSSIP FSSP FWAGP FWKHP FWKLP FWKWP FYOEP pwgtp1 pwgtp2 pwgtp3 pwgtp4
##     1:     0    0     0     0     0     0     0     87     28    153     93
##     2:     0    0     1     0     0     1     0     88     30    167     96
##     3:     0    0     0     0     0     0     0     94     33    163    110
##     4:     0    0     0     0     0     0     0     91     28    161    100
##     5:     0    0     0     0     0     0     0    539    365    288    414
##    ---
## 14927:     1    1     0     0     0     0     0     12     12     32     50
## 14928:     0    0     0     0     0     0     0     39    105     34     36
## 14929:     0    0     0     0     0     0     0     28     98     27     33
## 14930:     0    0     0     0     0     0     0    397    625    576    388
## 14931:     0    0     1     0     0     0     0    481    694    554    357
##         pwgtp5 pwgtp6 pwgtp7 pwgtp8 pwgtp9 pwgtp10 pwgtp11 pwgtp12 pwgtp13
##     1:     26     26     95     93     93      92      87     163      91
##     2:     27     25     95    100     99      90      91     164      92
##     3:     33     29    119    112    109     110     101     184     103
##     4:     28     26     98    106    106      98      88     162      90
##     5:    573    293     86    245    450     456     334     352     417
##    ---
## 14927:     31      8     28     28      5      25      40      50      39
## 14928:    125     33    227    139    120     125     183     121      36
## 14929:    134     34    170    124    117     104     168     143      34
## 14930:    588    107    333    384    399     108     328     550     341
## 14931:    643    108    314    424    426     124     374     666     385
##         pwgtp14 pwgtp15 pwgtp16 pwgtp17 pwgtp18 pwgtp19 pwgtp20 pwgtp21 pwgtp22
##     1:      25     153      89     149      83      25     180      89      23
##     2:      26     155      95     154      86      24     190      89      25
##     3:      32     190     116     178      95      29     219     104      29
##     4:      26     164     103     149      90      24     190      93      25
##     5:     103     283     100     108     282     129     408     442     261
##    ---
## 14927:      28      28      22      24      30      10       5      51      47
## 14928:     103     240     214     147     107     178     108      37     103
## 14929:     113     209     177     135      96     162      83      35     125
## 14930:     102     117     389     311      98     543     349     331      96
## 14931:     107     147     431     399     108     526     404     351     106
##        pwgtp23 pwgtp24 pwgtp25 pwgtp26 pwgtp27 pwgtp28 pwgtp29 pwgtp30 pwgtp31
##     1:     139      91      24      26      87      82      86      90      90
##     2:     142      96      26      30      88      85      92     100      94
##     3:     182     118      32      35     106      98      97     104     105
##     4:     145      95      25      27      88      82      90      90      93
##     5:     349     237     383     333     124     367     481     458     336
##    ---
## 14927:      22       6      29      51      32      25      41      20       6
## 14928:      35      37     123      46     207     109     115     111     190
## 14929:      39      37     100      42     201      97      99     111     210
## 14930:     109     354     105     600     356     353     353     648     335
## 14931:     124     403     115     665     423     368     418     647     330
##        pwgtp32 pwgtp33 pwgtp34 pwgtp35 pwgtp36 pwgtp37 pwgtp38 pwgtp39 pwgtp40
##     1:     151      91      28     144      81     146      95      27      22
##     2:     154      91      29     154      88     151      96      27      25
##     3:     191     107      35     176     101     168     112      34      28
##     4:     157      86      26     146      80     144      93      27      23
##     5:     255     614     102     284     117      93     327     102     356
##    ---
## 14927:       8       8      26      26      38      26      22      48      63
## 14928:     116      34     142     204     237     126     125     222     127
## 14929:     106      32     115     212     256     106     133     202     110
## 14930:     118     330     581     676     353     318     514     115     345
## 14931:     120     405     652     672     363     347     639     131     386
##        pwgtp41 pwgtp42 pwgtp43 pwgtp44 pwgtp45 pwgtp46 pwgtp47 pwgtp48 pwgtp49
##     1:      89     173      27      84     153     149      93      89      91
##     2:      93     163      29      83     156     153      96      90      91
##     3:     100     184      38     109     192     174     125     113      95
##     4:      92     160      28      85     156     148      96      99      90
##     5:     106     256     326     290      96     346     571     268     117
##    ---
## 14927:       9       8      34      63      27       8      25      27       7
## 14928:      34     115      42      37     119      43     225     111     110
## 14929:      29     101      34      37     115      41     176      94     107
## 14930:     356      99     122     365     108     578     367     327     319
## 14931:     418     105     127     370     118     595     352     422     373
##        pwgtp50 pwgtp51 pwgtp52 pwgtp53 pwgtp54 pwgtp55 pwgtp56 pwgtp57 pwgtp58
##     1:      92      90      27      91     139      25      91      29      84
##     2:      98      93      27      98     150      28      96      30      85
##     3:     107     103      35     104     164      33     110      34      97
##     4:      97      93      27     100     151      26      93      30      87
##     5:     118     320     263     128     453     298     480     393     307
##    ---
## 14927:      28      58      51      42      21      20      21      24      25
## 14928:     110     190     116      35     142     206     223     121     117
## 14929:      93     193     123      39     128     206     248     112     115
## 14930:     616     337     103     358     521     627     339     363     635
## 14931:     662     415     110     389     602     677     371     393     806
##        pwgtp59 pwgtp60 pwgtp61 pwgtp62 pwgtp63 pwgtp64 pwgtp65 pwgtp66 pwgtp67
##     1:     149      30      94     140      24      90     147     148      93
##     2:     152      30      98     144      24      92     161     163      97
##     3:     188     40    110    176     34    113    190    186    107
##     4:     144     33    101    152     23     95    148    149     92
##     5:     480    282    117    347    323    377    106    239    386
##   ---
## 14927:       6     10     33     46     36      9     23     39     37
## 14928:     237    125     31    133     33     34    125     38    201
## 14929:     183    106     32    120     33     35    113     37    198
## 14930:     107    334    380    650    555    361    638    112    378
## 14931:     124    371    349    635    681    405    614    132    486
##        pwgtp68 pwgtp69 pwgtp70 pwgtp71 pwgtp72 pwgtp73 pwgtp74 pwgtp75 pwgtp76
##     1:      82      84      87      82      27      92     150      28      78
##     2:      88      89      93      84      26      90     159      30      87
##     3:      99     101     109      90      28      92     177      33     105
##     4:      86      84      87      81      28      94     164      29      88
##     5:     309      90      96     294     400      80     489     340     491
##   ---
## 14927:      27      36      23       7      10      11      25      30      27
## 14928:     123     112     108     200     122      38     103     185     202
## 14929:     129     115     118     167     117      36      89     167     170
## 14930:     341     355     109     315     559     368     101     106     378
## 14931:     360     377     111     387     610     402     112     109     404
##        pwgtp77 pwgtp78 pwgtp79 pwgtp80
##     1:      25      99     159     129
##     2:      27      98     167     131
##     3:      30     104     206     156
##     4:      27     104     156     138
##     5:     612     282     462     259
##   ---
## 14927:      22      29      39      45
## 14928:     132     114     199     122
## 14929:     121     104     186     117
## 14930:     339     111     572     359
## 14931:     364     109     600     372
```

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

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