Class 3 Week 1 Quiz

-----------------------------------------------------------------------------

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

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

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

THe Val colum has 24 codes (1-24) representing the house values. 24th is $1M+

> getwd()

[1] "F:/Skydrive/Certification/Data Science Specialization/R\_Working\_Dir"

> dir.create("3Week1")

> setwd("F:/Skydrive/Certification/Data Science Specialization/R\_Working\_Dir/3Week1")

> getwd()

[1] "F:/Skydrive/Certification/Data Science Specialization/R\_Working\_Dir/3Week1

fileURL <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv"

download.file(url=fileURL,destfile="idaho\_housing.csv",method="auto")

> list.files()

[1] "idaho\_housing.csv"

> dateDownLoaded <- date()

> housing <- read.table("idaho\_housing.csv",sep = ",", header=TRUE, quote="")

> length(housing$VAL[!is.na(housing$VAL) & housing$VAL==24])

[1] 53

try this:

length(which((housing[,housing$VAL]==24)))

table(housing$VAL)[[24]]

-------------------------------------------------------------------

What is the value of: sum(datZip∗dat Ext,na.rm=T)

Helpful Info:  
<https://class.coursera.org/getdata-015/forum/thread?thread_id=6>

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

Had to get the package xlsx. Search RSeek.org. Found the zipped file here:

http://cran.r-project.org/web/packages/xlsx/index.html

In RStudio, clicked INstall on the Packages tab, pointed to the zipped file   
THen found a problem with package rJava. Installed rJava from Packages tab  
Then told I needed xlsxjars. stalled xlsxjars from Packages tab.  
Finally package(xlsx) installed

install.packages("F:/Skydrive/Certification/Data Science Specialization/Class 3 - Getting and Cleaning Data/xlsx\_0.5.7.zip", repos = NULL)

install.packages("rJava")

install.packages("xlsxjars")

library(xlsx)

> download.file(url=fileURL, destfile="NGAP.xlsx", mode="wb", method="auto")  
(Had to use the wb mode. Useful values are "w", "wb" (binary), "a" (append) and "ab".)

> dat <- read.xlsx(file="NGAP.xlsx",sheetIndex=1,colIndex=colIndex,startRow=18, endRow=23, header=TRUE)

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

--------------------------------------------------------------------

How many restaurants have zipcode 21231?

The course notes for this week are very good!

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

Load XML Package - Library(XML)

Load the data (its in RAM)

> doc <- xmlTreeParse(file=fileURL,useInternal=TRUE)

> rootNode <- xmlRoot(doc) # Wrapper for the entire document

> names(rootNode) #return names of elements

> rootNode[[1]][[1]] #ﬁrst sub component in the ﬁrst element

> xmlSApply(rootNode, xmlValue) #returns every single tagged element in the doc

> zipcode <- xpathSApply(rootNode,"//zipcode",xmlValue)

> length(zipcode[zipcode==21231])

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?

# write the file url and file destination to an object   
fileurl <- 'https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv'

# download from the URL   
download.file(fileurl, "3WeekQ1.csv", method="auto" )

# load the data.table package   
library(data.table)

# read the data   
# Similar to read.table but faster and more convenient. All controls such as sep, colClasses and nrows are automatically detected

DT <- fread("3WeekQ1.csv")

# time the processes

system.time(replicate(5000,(mean(DT$pwgtp15,by=DT$SEX)))) # fastest by far but does not return data by sex

system.time(replicate(5000,(tapply(DT$pwgtp15,DT$SEX,mean)))) # Returns the right data

system.time(replicate(5000,(sapply(split(DT$pwgtp15,DT$SEX),mean)))) # Returns the right data FASTEST

system.time(replicate(5000,(mean(DT[DT$SEX==1,]$pwgtp15); mean(DT[DT$SEX==2,]$pwgtp15)))) # 2 seperate calcs

system.time(replicate(5000,(DT[,mean(pwgtp15),by=SEX]))) # Returns the right data

system.time(replicate(5000,(rowMeans(DT)[DT$SEX==1]; rowMeans(DT)[DT$SEX==2]))) # 2 seperate calcs

**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 properties are worth $1,000,000 or more?

|  |  |  |  |
| --- | --- | --- | --- |
| **Your Answer** |  | **Score** | **Explanation** |
| 31 |  |  |  |
| 2076 |  |  |  |
| 25 |  |  |  |
| 53 | Correct | 3.00 |  |
| Total |  | 3.00 / 3.00 |  |

**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?

|  |  |  |  |
| --- | --- | --- | --- |
| **Your Answer** |  | **Score** | **Explanation** |
| Each tidy data table contains information about only one type of observation. |  |  |  |
| Tidy data has variable values that are internally consistent. |  |  |  |
| Numeric values in tidy data can not represent categories. |  |  |  |
| Tidy data has one variable per column. | Correct | 3.00 |  |
| Total |  | 3.00 / 3.00 |  |

**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>)

|  |  |  |  |
| --- | --- | --- | --- |
| **Your Answer** |  | **Score** | **Explanation** |
| 36534720 | Correct | 3.00 |  |
| 33544718 |  |  |  |
| NA |  |  |  |
| 154339 |  |  |  |
| Total |  | 3.00 / 3.00 |  |

**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?

|  |  |  |  |
| --- | --- | --- | --- |
| **Your Answer** |  | **Score** | **Explanation** |
| 100 |  |  |  |
| 28 |  |  |  |
| 130 |  |  |  |
| 127 | Correct | 3.00 |  |
| Total |  | 3.00 / 3.00 |  |

**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?

|  |  |  |  |
| --- | --- | --- | --- |
| **Your Answer** |  | **Score** | **Explanation** |
| rowMeans(DT)[DT$SEX==1]; rowMeans(DT)[DT$SEX==2] |  |  |  |
| sapply(split(DT$pwgtp15,DT$SEX),mean) | Correct | 3.00 |  |
| tapply(DT$pwgtp15,DT$SEX,mean) |  |  |  |
| DT[,mean(pwgtp15),by=SEX] |  |  |  |
| mean(DT[DT$SEX==1,]$pwgtp15); mean(DT[DT$SEX==2,]$pwgtp15) |  |  |  |
| mean(DT$pwgtp15,by=DT$SEX) |  |  |  |
| Total |  | 3.00 / 3.00 |  |