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

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

```
## [1] "Sat Jun 27 10:07:28 2020"
```

```
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?

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

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

```
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

pwgtp15

```
DT <- data.table::fread("https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv")
system.time(DT[,mean(pwgtp15),by=SEX])
```

```
##    user  system elapsed
##    0.02    0.00    0.02
```