week 1

me

8/6/2021

week 1 notes for Cleaning Data

dir

```
if(!file.exists("data")){
        dir.create("data")
}
```

Getting data from internet-download.file()

download.file() *url destfile method*

```
fileUrl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06hid.csv"
download.file(fileUrl, destfile = "./data/2006.csv", methods = "curl")
list.files("./data")
```

```
## [1] "2006.csv"          "community2006.csv" "condoms.csv"
## [4] "getdatagov.xlsx"   "Idaho.csv"         "xml.xml"
```

dont't forget to add the access date.

```
dateDownload <- date()
dateDownload
```

```
## [1] "Mon Aug 16 19:56:30 2021"
```

if the url start with **http** or **https** will be fine with using download.file()

loading flat files - read.table()

read.table(*file*,*header*,*sep*,*row.names*,*nrows*) *skip quote* you'll get an error

```
com <- read.table("./data/2006.csv")
head(com)
```

```
##
## 1 RT,SERIALNO,DIVISION,PUMA,REGION,ST,ADJUST,WGTP,NP,TYPE,ACR,AGS,BDS,BLD,BUS,CONP,ELEP,FS,FULP,GASP
## 2
## 3
## 4
## 5
## 6
```

```
com1 <- read.table("./data/2006.csv", sep = ",",header = TRUE, skip = 2, nrows = 8)
# com1
```

1 How many properties are worth $1,000,000 or more?

```
com <- read.table("./data/2006.csv", sep = ",",header = TRUE)

a <- subset(com, VAL>=24)

# is.na(com$VAL)
```

reading xlsx files

```
if(!file.exists("data")){dir.create("data")}
fileUrl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2FDATA.gov_NGAP.xlsx"
download.file(fileUrl, destfile = "./data/getdatagov.xlsx",method = "curl")
accessDate <- date()
```

xlsx package

```
#install.packages("xlsx")
library("xlsx")
gas <- read.xlsx("./data/getdatagov.xlsx",sheetIndex = 1, header = TRUE)
head(gas)
```

```
##   Table.Name..Contract          NA.               NA..1     NA..2
## 1       ContractNumber ContractorId          ExpiryDate CFileName
## 2   GS-00P-02-BSC-0201           23 2004-09-30 00:00:00      <NA>
## 3   GS-00P-02-BSC-0204            5 2003-10-31 00:00:00      NULL
## 4   GS-00P-02-BSC-0206            6 2004-10-31 00:00:00      <NA>
## 5   GS-00P-02-BSC-0207            4 2006-10-31 00:00:00      <NA>
## 6   GS-00P-02-BSC-0209            7 2004-10-31 00:00:00      <NA>
##                 NA..3 NA..4 NA..5 NA..6 NA..7 NA..8 NA..9 NA..10 NA..11 NA..12
```

```
## 1      ReactivationDt  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>   <NA>   <NA>   <NA>
## 2 2004-09-30 00:00:00  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>   <NA>   <NA>   <NA>
## 3                NULL  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>   <NA>   <NA>   <NA>
## 4 2004-11-02 00:00:00  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>   <NA>   <NA>   <NA>
## 5 2004-11-01 00:00:00  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>   <NA>   <NA>   <NA>
## 6 2004-11-01 00:00:00  <NA>  <NA>  <NA>  <NA>  <NA>  <NA>   <NA>   <NA>   <NA>
##   NA..13 NA..14 NA..15 NA..16 NA..17 NA..18 NA..19 NA..20 NA..21 NA..22 NA..23
## 1   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 2   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 3   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 4   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 5   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
## 6   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>   <NA>
##   NA..24
## 1   <NA>
## 2   <NA>
## 3   <NA>
## 4   <NA>
## 5   <NA>
## 6   <NA>
```

3

Read rows 18-23 and columns 7-15 into R and assign the result to a variable called:DAT what's the value of: sum(dat$Zip * dat$Ext,na.rm=T)?

```
rowIndex <- 18:23
colIndex <- 7:15
dat <- read.xlsx("./data/getdatagov.xlsx",sheetIndex = 1 ,rowIndex = rowIndex,colIndex = colIndex, head
sum(dat$Zip*dat$Ext,na.rm=T)
```

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

reading xml data

```
# install.packages("XML")
library(XML)
file <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Frestaurants.xml"
restaurant <- xmlTreeParse(sub("s","",file),useInternalNode = TRUE)

#class(restaurant)
rootNode <- xmlRoot(restaurant)
xmlName(rootNode)
```

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

```
names(rootNode)
```

```
##   row
## "row"
```

4

How many restaurants have zipcode 21231?

```
#rootNode[[1]][[3]][[2]]
zipcode <- xpathSApply(rootNode,"//zipcode",xmlValue)
# zipcode
d <- subset(zipcode,zipcode=="21231")

d <- subset(rootNode,xpathSApply(rootNode,"//zipcode",xmlValue)=="21231")
```

data.table package

```
#install.packages("data.able")
library("data.table")
fileUrl <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv"
c <- download.file(fileUrl, destfile = "./data/Idaho.csv")
idaho <- read.csv("./data/Idaho.csv",header = TRUE)

DT <- data.table(idaho)
# DT[1,]
# DT[c(3,5,10)]
DT[,list(mean(pwgtp1),sum(RAC1P))]
```

```
##          V1    V2
## 1: 98.21613 21745
```

```
DT[,table(VPS)]
```

```
## VPS
##   1   2   3   4   5   6   7   8   9  10  11  12
## 251  23 483  14   4 174  18 196 161 170   7   7
```

```
DT[,a:=RACBLK>0]
```

5

the ans is DT cuz it uses the subset from packages of data tables others uses bases packages it'a a tricky one...

```
# Url <- "https://d396qusza40orc.cloudfront.net/getdata%2Fdata%2Fss06pid.csv"
# download.file(Url,destfile = "./data/Idaho.csv")
library(data.table)

Idaho <- read.csv("./data/Idaho.csv",header = TRUE)

Idaho <- data.table(Idaho)
```

```
file <- tempfile()
write.table(Idaho, file = file,row.names=FALSE,col.names = TRUE, sep = ",", quote = FALSE)
DT <- fread(file)

system.time(fread(file))
```

```
##    user  system elapsed
##    0.06    0.02    0.04
```

time <- 1000 a <- replicate(time,system.time(mean(DT[DT$SEX==1,]$pwgtp15), mean(DT[DT$SEX==2,]$pwgtp15)))

sum(a[1,]) #9.76 sum(a[3,]) #12.16 plot(b)

b <- replicate(time, system.time(DT[,mean(pwgtp15),by=SEX])) sum(b[1,]) # 4.73 sum(b[3,]) #6.68

error

c <- replicate(time,system.time(rowMeans(DT)[DT$SEX==1],rowMeans(DT

#sum(c[1,])

incorrect ans

#d <- replicate(time,system.time(mean(DT$pwgtp15, by = DT$SEX))) #sum(d[1,]) #0.05

e <- replicate(time,system.time(sapply(split(DT$pwgtp15, DT$SEX),mean))) sum(e[1,]) #0.47 sum(e[3,]) #0.49

f <- replicate(time,system.time(tapply(DT$pwgtp15, DT$SEX,mean))) sum(f[1,]) #0.52 sum(f[3,]) #0.55