STA 445 HW3

Sophia Kubisiak

02/25/2024

```
library(tidyverse)
library(readxl)
library(readr)
```

Problem 1

Download from GitHub the data file Example_5.xls. Open it in Excel and figure out which sheet of data we should import into R. At the same time figure out how many initial rows need to be skipped. Import the data set into a data frame and show the structure of the imported data using the `str()` command. Make sure that your data has $n = 31$ observations and the three columns are appropriately named. If you make any modifications to the data file, comment on those modifications.

```
url <- "https://github.com/dereksonderegger/444/raw/master/data-raw/Example_5.xls"
destfile <- "Example_5.xls"
Example_5 <- read_excel(destfile, range = "A5:C36" , sheet =2)
```

```
head(Example_5)
```

```
## # A tibble: 6 x 3
##   Girth Height Volume
##   <dbl>  <dbl>  <dbl>
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
## 6  10.8     83   19.7
```

```
str(Example_5)
```

```
## tibble [31 x 3] (S3: tbl_df/tbl/data.frame)
##  $ Girth : num [1:31] 8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
##  $ Height: num [1:31] 70 65 63 72 81 83 66 75 80 75 ...
##  $ Volume: num [1:31] 10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
```

Problem 2

Download from GitHub the data file Example_3.xls. Import the data set into a data frame and show the structure of the imported data using the `tail()` command which shows the last few rows of a data table.

Make sure the Tesla values are `NA` where appropriate and that both `-9999` and `NA` are imported as NA values. If you make any modifications to the data file, comment on those modifications.

```
url <- "https://github.com/dereksonderegger/444/raw/master/data-raw/Example_3.xls"
destfile <- "Example_3.xls"
Example_3 <- read_excel(destfile, range = "A1:L34" , sheet=2)
```

```
tail(Example_3)
```

```
## # A tibble: 6 x 12
##   model        mpg cyl   disp     hp    drat    wt  qsec vs       am  gear carb
##   <chr>      <dbl> <chr> <chr> <dbl>   <dbl> <dbl> <dbl> <chr> <dbl> <dbl> <chr>
## 1 Lotus Eur~  30.4 4     95.0~   113  3.77e0  1.51  16.9 1         1     5 2
## 2 Ford Pant~  15.8 8     351     264  4.22e0  3.17  14.5 0         1     5 4
## 3 Ferrari D~  19.7 6     145     175  3.62e0  2.77  15.5 0         1     5 6
## 4 Maserati ~  15   8     301     335  3.54e0  3.57  14.6 0         1     5 8
## 5 Volvo 142E  21.4 4     121     109  4.11e0  2.78  18.6 1         1     4 2
## 6 Tesla Mod~  98   NA    NA      778 -1.00e4  4.94  10.4 NA        0     1 NA
```

Problem 3

Download all of the files from GitHub `data-raw/InsectSurveys` directory here. Each month's file contains a sheet contains site level information about each of the sites that was surveyed. The second sheet contains information about the number of each species that was observed at each site. Import the data for each month and create a single `site` data frame with information from each month. Do the same for the `observations`. Document any modifications you make to the data files. Comment on the importance of consistency of your data input sheets.

```
data.1 <- read_excel("May.xlsx" , sheet = 1 , range = "A1:F10" )
data.2 <- read_excel("June.xlsx" , sheet = 1 , range = "A1:F10" )
data.3 <- read_excel("July.xlsx" , sheet = 1 , range = "A1:F10" )
data.4 <- read_excel("August.xlsx" , sheet = 1 , range = "A1:F10" )
data.5 <- read_excel("September.xlsx" , sheet = 1 , range = "A1:F10" )
data.6 <- read_excel("October.xlsx" , sheet = 1 , range = "A1:F10" )
```

```
sitesMayJune <- rbind(data.1 , data.2)
```

```
sitesJulyAug <- rbind(data.3 , data.4)
```

```
sitesSeptOct <- rbind(data.5 , data.6)
```

```
sitesMJJA <- rbind(sitesMayJune , sitesJulyAug)
```

```
sitesALL <- rbind(sitesSeptOct , sitesMJJA)
```

```
head(sitesALL)
```

```
## # A tibble: 6 x 6
##   `Site Name`      `Pond Area` `Water Depth`    ph Date                Observer
```

```
##   <chr>                <dbl>         <dbl> <dbl> <dttm>              <chr>
## 1 Araphahoe Road          34           3     6.2 2020-09-15 00:00:00 Bob
## 2 Bridger Valley         240           6     6.5 2020-09-16 00:00:00 Bob
## 3 Calculus Vector        321          13     6.4 2020-09-17 00:00:00 Bob
## 4 Deer Valley             74           4.4   6.9 2020-09-18 00:00:00 Bob
## 5 Ephemeral Stream        28           2     7.1 2020-09-15 00:00:00 Charlie
## 6 Fennel Gardens          62           3.6   7   2020-09-16 00:00:00 Charlie
```

For the sites, I first rearranged and renamed the sheet name to 'Sites' with a capital S for all of the data frames. I then changed the date column's format to the same format in every sheet. I then renamed the column names so that they all have a capital letter, and all are the same name.

```
data1 <- read_excel("May.xlsx" , sheet = 2 , range = "A1:C37" )
data2 <- read_excel("June.xlsx" , sheet = 2 , range = "A1:C37" )
data3 <- read_excel("July.xlsx" , sheet = 2 , range = "A1:C37" )
data4 <- read_excel("August.xlsx" , sheet = 2 , range = "A1:C37" )
data5 <- read_excel("September.xlsx" , sheet = 2 , range = "A1:C37" )
data6 <- read_excel("October.xlsx" , sheet = 2 , range = "A1:C37" )
```

```
obsMayJune <- rbind(data1 , data2)
```

```
obsJulyAug <- rbind(data3 , data4)
```

```
obsSeptOct <- rbind(data5 , data6)
```

```
obsMJJA <- rbind(obsMayJune , obsJulyAug)
```

```
obsALL <- rbind(obsSeptOct , obsMJJA)
```

```
head(obsALL)
```

```
## # A tibble: 6 x 3
##   Site           Species    Count
##   <chr>          <chr>      <dbl>
## 1 Araphahoe Road Caddis Fly     2
## 2 <NA>           May Fly        4
## 3 <NA>           Stone Fly      8
## 4 <NA>           Dragon Fly     7
## 5 Bridger Valley Caddis Fly     2
## 6 <NA>           May Fly        4
```

For the observations, I named all of the sheets 'Observations" with a capital O. I also had to change the column names so that all of them started with a capital letter, and were all the same.

It is really important to have consistency in your data. Otherwise, R is unable to correctly combing data together, and it makes your job more difficult trying to correct it in R when you can just correct the excel sheet themselves.