STA 445 HW3

Matthew Hashim

2/5/2024

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

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.

```
trees <- read_excel("Example_5.xls", sheet = "RawData", skip = 4)[1:3]
head(trees)
```

```
## # 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
```

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.

```
cars <- read_excel("Example_3.xls", sheet = "data", na = c(-9999, "NA"),
                   n_max = 34)[1:12]
tail(cars)
```

```
## # A tibble: 6 x 12
##   model          mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##   <chr>        <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 Lotus Europa  30.4     4  95.1   113  3.77  1.51  16.9     1     1     5     2
```

```
## 2 Ford Panter~  15.8     8 351     264  4.22  3.17  14.5     0     1     5     4
## 3 Ferrari Dino  19.7     6 145     175  3.62  2.77  15.5     0     1     5     6
## 4 Maserati Bo~  15       8 301     335  3.54  3.57  14.6     0     1     5     8
## 5 Volvo 142E    21.4     4 121     109  4.11  2.78  18.6     1     1     4     2
## 6 Tesla Model~  98      NA  NA     778 NA     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.

It is important to keep information consistant so that we can easily compare, merge, and append data sets together.

Change 1: Changed the order and names of the sheets so that each sheet goes Sites, Observations Change 2: Changed the date to be MM/DD/YYYY and removed "did not visit" Change 3: Added site names for every row Change 4: Reorder and Make column names consistent

```
Sites = NULL
Observations = NULL
Months <- c("May.xlsx", "June.xlsx", "July.xlsx", "August.xlsx", "September.xlsx", "October.xlsx")
for(i in Months){
  temp <- read_excel(i, sheet = "Sites", range = "A1:F10", na = "NA")
  Sites <- rbind(Sites, temp)
  temp <- read_excel(i, sheet = "Observations", range = "A1:C37")
  Observations = rbind(Observations, temp)
}
slice_sample(Sites, n = 10)
```

```
## # A tibble: 10 x 6
##    `Site Name`      `Pond Area` `Water Depth`    ph Date                Observer
##    <chr>                  <dbl>         <dbl> <dbl> <dttm>              <chr>
##  1 Calculus Vector          321            13   6.4  2020-06-17 00:00:00 Bob
##  2 Fennel Gardens            62             3.6 7    NA                  Charlie
##  3 Ephemeral Stream          28             2   7.1  2020-10-15 00:00:00 Charlie
##  4 Gigantic Pain            489             4   7.1  2020-10-17 00:00:00 Charlie
##  5 Happy Feet               398            10   6.8  2020-07-18 00:00:00 Charlie
##  6 Happy Feet               398            10   6.8  2020-08-18 00:00:00 Charlie
##  7 Indigo Flats             126             9   6.75 2020-10-19 00:00:00 Charlie
##  8 Deer Valley               74             4.4 6.9  2020-09-18 00:00:00 Bob
##  9 Bridger Valley           240             6   6.5  2020-05-16 00:00:00 Bob
## 10 Happy Feet               398            10   6.8  2020-09-18 00:00:00 Charlie
```

```
slice_sample(Observations, n = 10)
```

```
## # A tibble: 10 x 3
##    Site             Species    Count
##    <chr>            <chr>      <dbl>
##  1 Ephemeral Stream May Fly        4
```

```
##  2 Bridger Valley   Stone Fly      8
##  3 Gigantic Pain    Caddis Fly     2
##  4 Bridger Valley   May Fly        4
##  5 Gigantic Pain    Caddis Fly     2
##  6 Happy Feet       May Fly        4
##  7 Araphahoe Road   Stone Fly      8
##  8 Calculus Vector  Caddis Fly     2
##  9 Calculus Vector  Stone Fly      8
## 10 Indigo Flats     Caddis Fly     2
```