Data Wrangling

Erwing

2023-05-09

Dr. Granger investigates factors which control carbon storage units and size of shrubs. The experiment involves analysis of three treatments which affect shrub volume at four areas.

This code is to load in the csv, was not previously able to load in with simple read.csv(file, 'shrub-volume-data.csv') needed to specify file location

Exercise 6

##Load in package dplyr

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Load in the data sets for shrub data, species, and plots.

```
shrub_data <- read.csv('/Users/sanch/Documents/Schooldownloads/BIO197/Rstudio/Homeworks/Homeworks/data-
```

```
species <- read.csv('/Users/sanch/Documents/Schooldownloads/BIO197/Rstudio/Homeworks/Homeworks/data-raw/
```

```
plots <- read.csv('/Users/sanch/Documents/Schooldownloads/BIO197/Rstudio/Homeworks/Homeworks/data-raw/pl
```

##Prints out the data from column "length"

```
print(shrub_data$length)
```

```
##  [1] 2.2 2.1 2.7 3.0 3.1 2.5 1.9 1.1 3.5 2.9 4.5 1.2
```

##Prints out the data from column "length"

```
print(shrub_data$site)
```

```
##  [1] 1 1 1 2 2 2 3 3 3 4 4 4
```

##Prints out the data from column "length"

```
print(shrub_data$experiment)
```

```
##  [1] 1 2 3 1 2 3 1 2 3 1 2 3
```

##This code makes the area column from the length times the width

```
area <- (shrub_data$length*shrub_data$width)
```

##This code adds the area column previously made to the shrub_data dataframe

```
shrub_data$area <- area
```

##This code checks that the column was actually added

```
View(shrub_data)
```

##Was able to load in dplyr package for the arrange function, but was not able to use it because library() fx would present ##Error: package or namespace load failed for 'dplyr' in loadNamespace(i, c(lib.loc, .libPaths()), versionCheck = vI[[i]]): namespace 'rlang' 1.0.6 is already loaded, but >= 1.1.0 is required In addition: Warning message:package 'dplyr' was built under R version 4.2.3

##Used this code to arrange by length instead

```
shrub_data_column <- shrub_data[order(shrub_data$length),]
```

##Ensure data frame is arranged by length by view code

```
View(shrub_data_column)
```

##Was not able to get the filter function to work properly, ended up getting only a single row produced which had values not even related to the original data frame

##This code creates the shrub_volumes column(vector) by multiplying the existing length, width, height of the shrubd_data_column dataframe

```
shrub_volumes <- (shrub_data_column$length*shrub_data_column$width*shrub_data_column$height)
```

##This code adds shrub_volumes to the shrub_data_column dataframe which can also be viewed with the view function

```
shrub_data_column$shrub_volumes <- shrub_volumes
View(shrub_data_column)
```

Exercise 7 Data aggregation

It is desired to create a summary for Dr. Granger's plants for each site and for each experiment. In the following section we will be coding for obtaining various values based on factors such as max, or from calculated values such as volume.

```
shrub_dims <- read.csv('/Users/sanch/Documents/Schooldownloads/BIO197/Rstudio/Homeworks/Homeworks/data-
```

##Grouping by site code is found below, but could not use had to resort to another code

```
by_site <- group_by(shrub_dims, site)
```

##Taking the average height of each plant in each experiment code is found below

```
avg_height <- summarize(by_site, avg_height = mean(height))
```

##Printing the average height of each plant

```
print(avg_height)
```

```
## # A tibble: 4 x 2
##    site avg_height
##   <int>      <dbl>
## 1     1       6.47
## 2     2       2.83
## 3     3       4.77
## 4     4       4.13
```

##Code to take the max height for each plant at each site of shrub_dims. Was unsure how to go about using the max() so I used tapply() and involved max within

```
tapply(shrub_dims$height,shrub_dims$site,max)
```

```
##   1   2   3   4
## 9.6 4.0 7.5 6.5
```

##Was not able to complete 3 as was unsure how to go about applying previous code used but in the form of a pipeline as it is all in one step.

##Exercise 8

##Broken code below read.csv("shrub-volume-data.csv") shrub_data |> mutate(volume = length * width * height) |> group_by(site) |> summarize(mean_volume = max(volume)) shrub_data |> mutate(volume = length * width * height) group_by(experiment) |> summarize(mean_volume = mean(volume))

##Fixed code below, all it required was a simple piping from the mutate() code to the group_by() code for the experiment section, with a quick skim I determined there was a |> missing. Here is the working code.

```
read.csv('/Users/sanch/Documents/Schooldownloads/BIO197/Rstudio/Homeworks/Homeworks/data-raw/shrub-volu
```

```
##    site experiment length width height
## 1     1          1    2.2   1.3    9.6
## 2     1          2    2.1   2.2    7.6
## 3     1          3    2.7   1.5    2.2
## 4     2          1    3.0   4.5    1.5
## 5     2          2    3.1   3.1    4.0
## 6     2          3    2.5   2.8    3.0
## 7     3          1    1.9   1.8    4.5
## 8     3          2    1.1   0.5    2.3
## 9     3          3    3.5   2.0    7.5
## 10    4          1    2.9   2.7    3.2
## 11    4          2    4.5   4.8    6.5
## 12    4          3    1.2   1.8    2.7
```

```
shrub_data |>
    mutate(volume = length * width * height) |>
    group_by(site) |>
    summarize(mean_volume = max(volume))
```

```
## # A tibble: 4 x 2
##    site mean_volume
##   <int>       <dbl>
## 1     1        35.1
## 2     2        38.4
## 3     3        52.5
## 4     4       140.
```

```
shrub_data |>
    mutate(volume = length * width * height) |>
group_by(experiment) |>
    summarize(mean_volume = mean(volume))
```

```
## # A tibble: 3 x 2
##   experiment mean_volume
##        <int>       <dbl>
## 1          1        22.0
## 2          2        53.8
## 3          3        22.1
```