shrub-volume-dataset

Azul Carrillo

2022-10-17

Exercise 1: Data wrangling basics

#2. Describing the data that we are using

We are using the data set on this file

We are using data collected from Dr. Granger's study of the factors controlling the size and carbon storage of shrubs. Dr. Granger conducted an experiment looking at three different treatments and their effect on shrub volume at 4 different locations. The data set consists of five columns. Site, experiment, length, width and, height.

3. reading the data table into R

```
getwd()
```

```
## [1] "/Users/atziri/Bio 195-197/Data Science/documents"
```

```
shrubs <-read.csv(file = "../raw-data/shrub-volume-data.csv")
head(shrubs)
```

```
##   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
```

```
summary(shrubs)
```

```
##       site        experiment     length          width           height
##  Min.   :1.00   Min.   :1    Min.   :1.100   Min.   :0.500   Min.   :1.50
##  1st Qu.:1.75   1st Qu.:1    1st Qu.:2.050   1st Qu.:1.725   1st Qu.:2.60
##  Median :2.50   Median :2    Median :2.600   Median :2.100   Median :3.60
##  Mean   :2.50   Mean   :2    Mean   :2.558   Mean   :2.417   Mean   :4.55
##  3rd Qu.:3.25   3rd Qu.:3    3rd Qu.:3.025   3rd Qu.:2.875   3rd Qu.:6.75
##  Max.   :4.00   Max.   :3    Max.   :4.500   Max.   :4.800   Max.   :9.60
```

#4. Select the data from the "length" column and print it out. First use the library() function with the `dplyr` input to be able to use the select() function then use the select function to obtatin only the length column of the dataset

```
library(dplyr)
```

```
##
## 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
```

```
select(shrubs, length)
```

```
##    length
## 1     2.2
## 2     2.1
## 3     2.7
## 4     3.0
## 5     3.1
## 6     2.5
## 7     1.9
## 8     1.1
## 9     3.5
## 10    2.9
## 11    4.5
## 12    1.2
```

#5. Select the data from the site and experiment columns

```
select(shrubs, site, experiment)
```

```
##    site experiment
## 1     1          1
## 2     1          2
## 3     1          3
## 4     2          1
## 5     2          2
## 6     2          3
## 7     3          1
## 8     3          2
## 9     3          3
## 10    4          1
## 11    4          2
## 12    4          3
```

#6.Add a new column named "area" containing the area of the shrub, which is the length times the width Added a new column 'area' the area value is length multiplied by with i did this with the mutate() function

```
shrubs <- mutate(shrubs, area = length * width)
head(shrubs)
```

```
##   site experiment length width height  area
## 1    1          1    2.2   1.3    9.6  2.86
## 2    1          2    2.1   2.2    7.6  4.62
## 3    1          3    2.7   1.5    2.2  4.05
## 4    2          1    3.0   4.5    1.5 13.50
## 5    2          2    3.1   3.1    4.0  9.61
## 6    2          3    2.5   2.8    3.0  7.00
```

#7.Sort the data by length The function arrange will sort the data by length

```
arrange(shrubs, length)
```

```
##    site experiment length width height  area
## 1     3          2    1.1   0.5    2.3  0.55
## 2     4          3    1.2   1.8    2.7  2.16
## 3     3          1    1.9   1.8    4.5  3.42
## 4     1          2    2.1   2.2    7.6  4.62
## 5     1          1    2.2   1.3    9.6  2.86
## 6     2          3    2.5   2.8    3.0  7.00
## 7     1          3    2.7   1.5    2.2  4.05
## 8     4          1    2.9   2.7    3.2  7.83
## 9     2          1    3.0   4.5    1.5 13.50
## 10    2          2    3.1   3.1    4.0  9.61
## 11    3          3    3.5   2.0    7.5  7.00
## 12    4          2    4.5   4.8    6.5 21.60
```

#8.Filter the data to include only plants with heights greater than 5 The filter() function will allow me to add a limit to what i need to include in this case only plants with height greater than 5 >5

```
filter(shrubs, height > 5 )
```

```
##   site experiment length width height  area
## 1    1          1    2.2   1.3    9.6  2.86
## 2    1          2    2.1   2.2    7.6  4.62
## 3    3          3    3.5   2.0    7.5  7.00
## 4    4          2    4.5   4.8    6.5 21.60
```

#9. Filter the data to include only plants with heights greater than 4 and widths greater than 2 to filter more than one command using the filter() function you can use a comma `,` or the `&` symbol.

```
filter(shrubs, height > 4 & width > 2 )
```

```
##   site experiment length width height  area
## 1    1          2    2.1   2.2    7.6  4.62
## 2    4          2    4.5   4.8    6.5 21.60
```

#10. Filter the data to include only plants from Experiment 1 or Experiment 3

```
filter(shrubs, experiment == "1" | experiment == "3")
```

```
##   site experiment length width height  area
## 1    1          1    2.2   1.3    9.6  2.86
## 2    1          3    2.7   1.5    2.2  4.05
## 3    2          1    3.0   4.5    1.5 13.50
## 4    2          3    2.5   2.8    3.0  7.00
## 5    3          1    1.9   1.8    4.5  3.42
## 6    3          3    3.5   2.0    7.5  7.00
## 7    4          1    2.9   2.7    3.2  7.83
## 8    4          3    1.2   1.8    2.7  2.16
```

#11. Filter the data to remove rows with null values in the height column the !is.na()function will eliminate andy null values therfore here it will remuve null values from the heigth column

```
filter(shrubs, !is.na(height))
```

```
##    site experiment length width height  area
## 1     1          1    2.2   1.3    9.6  2.86
## 2     1          2    2.1   2.2    7.6  4.62
## 3     1          3    2.7   1.5    2.2  4.05
## 4     2          1    3.0   4.5    1.5 13.50
## 5     2          2    3.1   3.1    4.0  9.61
## 6     2          3    2.5   2.8    3.0  7.00
## 7     3          1    1.9   1.8    4.5  3.42
## 8     3          2    1.1   0.5    2.3  0.55
## 9     3          3    3.5   2.0    7.5  7.00
## 10    4          1    2.9   2.7    3.2  7.83
## 11    4          2    4.5   4.8    6.5 21.60
## 12    4          3    1.2   1.8    2.7  2.16
```

#12. Create a new data frame called shrub_volumes that includes all of the original data and a new column containing the volumes (length * width * height), and display it. by using the head function it shows only the first 6 rows and all columns so to display the full dataset i used the View() function

```
shrub_volumes <-  mutate(shrubs, volumes = length * width * height)
head(shrub_volumes)
```

```
##   site experiment length width height  area volumes
## 1    1          1    2.2   1.3    9.6  2.86  27.456
## 2    1          2    2.1   2.2    7.6  4.62  35.112
## 3    1          3    2.7   1.5    2.2  4.05   8.910
## 4    2          1    3.0   4.5    1.5 13.50  20.250
## 5    2          2    3.1   3.1    4.0  9.61  38.440
## 6    2          3    2.5   2.8    3.0  7.00  21.000
```

```
nrow(shrub_volumes)
```

```
## [1] 12
```

```
#View(shrub_volumes)
```

Exercise 2: Data agreggation

The following code calculates the average height of a plant at each site:

```
shrub_dims <- read.csv("../raw-data/shrub-volume-data.csv")
by_site <- group_by(shrub_dims, site)
avg_height <- summarize(by_site, avg_height = mean(height))
head(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
```

#1. Modify the code to calculate and print the average height of a plant in each experiment. i did this by creating a new subset by_experiment that will group the 3 experiments by number of experiment and then instead of using by_site subset in the new avg_height subset i used by_experiment. this gave me the average height of a plant in each experiment.

```
shrub_dims <- read.csv("../raw-data/shrub-volume-data.csv")
by_experiment <- group_by(shrub_dims, experiment)
avg_height <- summarize(by_experiment, avg_height = mean(height))
head(avg_height)
```

```
## # A tibble: 3 x 2
##   experiment avg_height
```

```
##        <int>      <dbl>
## 1          1       4.7
## 2          2       5.1
## 3          3       3.85
```

```
#View(avg_height)
```

#2. Use max() to determine the maximum height of a plant at each site.

```
shrub_dims <- read.csv("../raw-data/shrub-volume-data.csv")
by_site <- group_by(shrub_dims, site)
max_height <- summarize(by_site, max_height = max(height))
head(max_height)
```

```
## # A tibble: 4 x 2
##    site max_height
##   <int>      <dbl>
## 1     1        9.6
## 2     2        4
## 3     3        7.5
## 4     4        6.5
```

#3. Write the same code but as a pipeline (using the pipe |> or >) to get the same result. The pipe doesn't need subsets the higher function sends the next function a command when its connected by a pipe.

```
read.csv("../raw-data/shrub-volume-data.csv") %>%
  group_by(site) %>%
  summarize(max_height = max(height)) %>%
  head()
```

```
## # A tibble: 4 x 2
##    site max_height
##   <int>      <dbl>
## 1     1        9.6
## 2     2        4
## 3     3        7.5
## 4     4        6.5
```

3.

this is the original code

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))

What the pipe helps us do is send the command of the function above to the line that is attached to by the pipe.

#1. First I had to change the input of the rad.csv() function to give it direction and add the object shrubs_data doing this at the start of the code and not adding a pipe allowed the code to use that object

multiple times without having to read the csv every time the object needs to be used. #2. i had to modify all the pipe symbols because my system is not updated used the function colnames to see the name of each column

```
shrub_data <- read.csv("../raw-data/shrub-volume-data.csv") #1. added the object shrubs_data
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.
```

```
colnames(shrub_data)
```

```
## [1] "site"       "experiment" "length"     "width"      "height"
```

```
 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
```