uhuru dataset

Luna L Sanchez Reyes

2022-10-04

1. Describing the data that we are using

We are using the dataset from this study

Add a picture of an Acacia

2. Reading the data table into R

2.1 Getting the working directory of my R chunks The working directory of an Rmd file is not the same as the working directory of the Rstudio project you are in. We can check that by running the following:

```
r_proj_wd <- "/Users/lunasare/Desktop/data-science-course/fall-2022"
r_chunk_wd <- getwd()
r_proj_wd == r_chunk_wd
```

```
## [1] FALSE
```

As you can see, the working directory from my R project and the one used in my R chunks in an Rmd file are not the same.

Th working directory of an R chunk is equal to the path where the Rmd file is in, in my case, it is `r_chunk_wd`. This is nested withing my project directory, keep this in ming for when you have to provide the path of a file!

```
acacia <- read.csv(file = "../data-raw/ACACIA_DREPANOLOBIUM_SURVEY.txt", sep = "\t")
```

2.2 Reading a data table

3. Explore our data set

```
head(acacia)
```

```
##   SURVEY YEAR  SITE BLOCK TREATMENT    PLOT   ID HEIGHT AXIS1 AXIS2 CIRC
## 1      1 2012 SOUTH     1     TOTAL S1TOTAL  581   2.25  2.75  2.15   20
## 2      1 2012 SOUTH     1     TOTAL S1TOTAL  582   2.65  4.10  3.90   28
## 3      1 2012 SOUTH     1     TOTAL S1TOTAL 3111    1.5  1.70  0.85   17
## 4      1 2012 SOUTH     1     TOTAL S1TOTAL 3112   2.01  1.80  1.60   12
## 5      1 2012 SOUTH     1     TOTAL S1TOTAL 3113   1.75  1.84  1.42   13
## 6      1 2012 SOUTH     1     TOTAL S1TOTAL 3114   1.65  1.62  0.85   15
##   FLOWERS BUDS FRUITS ANT
## 1       0    0     10  CS
## 2       0    0    150  TP
## 3       2    1     50  TP
## 4       0    0     75  CS
## 5       0    0     20  CS
## 6       0    0      0   E
```

```
summary(acacia)
```

```
##      SURVEY       YEAR          SITE               BLOCK
##  Min.   :1   Min.   :2012   Length:157         Min.   :1.000
##  1st Qu.:1   1st Qu.:2012   Class :character   1st Qu.:2.000
##  Median :1   Median :2012   Mode  :character   Median :2.000
##  Mean   :1   Mean   :2012                      Mean   :2.089
##  3rd Qu.:1   3rd Qu.:2012                      3rd Qu.:2.000
##  Max.   :1   Max.   :2012                      Max.   :3.000
##
##   TREATMENT             PLOT                 ID          HEIGHT
##  Length:157         Length:157         Min.   : 101   Length:157
##  Class :character   Class :character   1st Qu.:1062   Class :character
##  Mode  :character   Mode  :character   Median :1301   Mode  :character
##                                        Mean   :1743
##                                        3rd Qu.:3118
##                                        Max.   :3199
##
##      AXIS1           AXIS2            CIRC          FLOWERS
##  Min.   :0.700   Min.   :0.550   Min.   : 4.00   Min.   : 0.0000
##  1st Qu.:1.400   1st Qu.:1.100   1st Qu.:10.00   1st Qu.: 0.0000
##  Median :1.800   Median :1.490   Median :13.00   Median : 0.0000
##  Mean   :1.972   Mean   :1.636   Mean   :13.76   Mean   : 0.4444
##  3rd Qu.:2.350   3rd Qu.:2.000   3rd Qu.:16.00   3rd Qu.: 0.0000
##  Max.   :5.550   Max.   :4.820   Max.   :35.20   Max.   :40.0000
##  NA's   :4       NA's   :4       NA's   :4       NA's   :4
##       BUDS             FRUITS           ANT
##  Min.   : 0.0000   Min.   :  0.00   Length:157
##  1st Qu.: 0.0000   1st Qu.:  0.00   Class :character
##  Median : 0.0000   Median :  0.00   Mode  :character
##  Mean   : 0.3595   Mean   : 20.03
##  3rd Qu.: 0.0000   3rd Qu.: 25.00
##  Max.   :50.0000   Max.   :300.00
##  NA's   :4         NA's   :4
```

```
colnames(acacia)
```

```
##  [1] "SURVEY"    "YEAR"      "SITE"      "BLOCK"     "TREATMENT" "PLOT"
##  [7] "ID"        "HEIGHT"    "AXIS1"     "AXIS2"     "CIRC"      "FLOWERS"
## [13] "BUDS"      "FRUITS"    "ANT"
```

Make sure that everuthing that is a number, is actually numeric.

One way to do this is withthe function `summary`, and checking at the type of data on each column visually.

Another way is using the type function

```
typeof(acacia[,"HEIGHT"])
```

```
## [1] "character"
```

```
acacia$HEIGHT
```

```
##   [1] "2.25" "2.65" "1.5"  "2.01" "1.75" "1.65" "1.2"  "1.45" "1.87" "2.38"
##  [11] "2.58" "2.65" "2.35" "1.88" "2.32" "2.39" "2.2"  "1.05" "2"    "1.28"
##  [21] "dead" "1.4"  "1.9"  "1.75" "1.8"  "2.7"  "2.02" "1.9"  "1.85" "1.65"
##  [31] "1.4"  "2.5"  "2.05" "2.26" "2.13" "1.8"  "1.85" "1.5"  "1.87" "1.58"
##  [41] "2.05" "1.75" "1.49" "1.28" "1.49" "1.07" "1.48" "1.25" "1.41" "1.6"
```

```
##  [51] "1.2"  "1.49" "1.5"  "1.65" "1.13" "1.25" "1.1"  "2.2"  "1.45" "1.6"
##  [61] "1.55" "1.5"  "1.03" "2.14" "1.2"  "1.05" "1.8"  "1.2"  "1.75" "1.45"
##  [71] "1.17" "2.15" "1.7"  "1.98" "1.26" "1.11" "1.14" "1.26" "1.3"  "1.29"
##  [81] "1.31" "1.15" "1.87" "1.47" "1.05" "2.1"  "1.99" "1.42" "1.5"  "1.06"
##  [91] "1.49" "1.8"  "1.93" "1.2"  "1.65" "1.52" "1.43" "1.25" "1.88" "1.03"
## [101] "1.1"  "1.4"  "1.05" "1.18" "1.4"  "1.37" "1.32" "1.55" "1.3"  "1.24"
## [111] "1.5"  "1.65" "2.17" "1.28" "1.07" "0.67" "0.68" "1.87" "1.35" "1.75"
## [121] "1.75" "1.64" "1.42" "dead" "0.9"  "dead" "1.8"  "2.47" "2.15" "1.7"
## [131] "1.9"  "1.95" "1.8"  "1.4"  "1"    "1.75" "1.28" "1"    "1.45" "1"
## [141] "1.03" "1.51" "1.17" "1.33" "1.3"  "1.13" "1.58" "1.06" "1.05" "1.45"
## [151] "1.15" "1.42" "1.02" "1.4"  "1.45" "1.95" "dead"
```

We identified a column that has problematic data. We need to fix it!

We are goinf to read the data table again, but we are gonna assign `NA` to the "dead" value that we do not want in our "HEIGHT" column.

```
acacia <- read.csv(file = "/Users/lunasare/Desktop/data-science-course/fall-2022/data-raw/ACACIA_DREPAN
                   sep = "\t",
                   na.strings = "dead")
```

Let's check if this worked!

```
acacia$HEIGHT
```

```
##   [1] 2.25 2.65 1.50 2.01 1.75 1.65 1.20 1.45 1.87 2.38 2.58 2.65 2.35 1.88 2.32
##  [16] 2.39 2.20 1.05 2.00 1.28   NA 1.40 1.90 1.75 1.80 2.70 2.02 1.90 1.85 1.65
##  [31] 1.40 2.50 2.05 2.26 2.13 1.80 1.85 1.50 1.87 1.58 2.05 1.75 1.49 1.28 1.49
##  [46] 1.07 1.48 1.25 1.41 1.60 1.20 1.49 1.50 1.65 1.13 1.25 1.10 2.20 1.45 1.60
##  [61] 1.55 1.50 1.03 2.14 1.20 1.05 1.80 1.20 1.75 1.45 1.17 2.15 1.70 1.98 1.26
##  [76] 1.11 1.14 1.26 1.30 1.29 1.31 1.15 1.87 1.47 1.05 2.10 1.99 1.42 1.50 1.06
##  [91] 1.49 1.80 1.93 1.20 1.65 1.52 1.43 1.25 1.88 1.03 1.10 1.40 1.05 1.18 1.40
## [106] 1.37 1.32 1.55 1.30 1.24 1.50 1.65 2.17 1.28 1.07 0.67 0.68 1.87 1.35 1.75
## [121] 1.75 1.64 1.42   NA 0.90   NA 1.80 2.47 2.15 1.70 1.90 1.95 1.80 1.40 1.00
## [136] 1.75 1.28 1.00 1.45 1.00 1.03 1.51 1.17 1.33 1.30 1.13 1.58 1.06 1.05 1.45
## [151] 1.15 1.42 1.02 1.40 1.45 1.95   NA
```

```
typeof(acacia$HEIGHT)
```

```
## [1] "double"
```

4. Visualize our data

For this, we are using the `ggplot` package. Let's install it and load it:

```
# install.packages("ggplot2")
library(ggplot2)
```

Now we are gonna create our first plotting layer with the function `ggplot`.

```
colnames(acacia)
```

```
##  [1] "SURVEY"    "YEAR"      "SITE"      "BLOCK"     "TREATMENT" "PLOT"
##  [7] "ID"        "HEIGHT"    "AXIS1"     "AXIS2"     "CIRC"      "FLOWERS"
## [13] "BUDS"      "FRUITS"    "ANT"
```

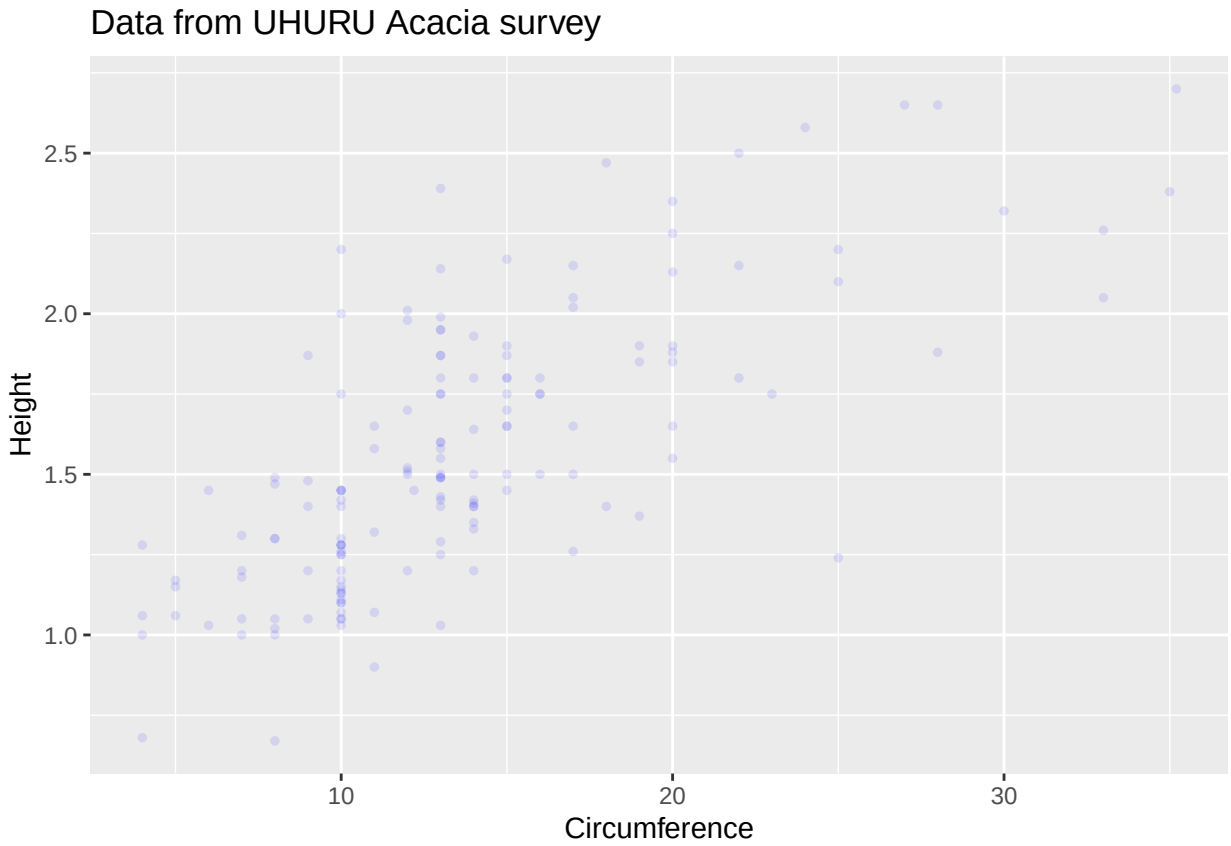
```
acacia$CIRC
```

```
##   [1] 20.0 28.0 17.0 12.0 13.0 15.0  9.0 12.2 13.0 35.0 24.0 27.0 20.0 28.0 30.0
##  [16] 13.0 10.0  8.0 10.0 10.0   NA 18.0 15.0 16.0 16.0 35.2 17.0 19.0 19.0 17.0
```

```
##  [31] 14.0 22.0 33.0 33.0 20.0 22.0 20.0 15.0 13.0 11.0 17.0 16.0 13.0 10.0 13.0
##  [46] 11.0  9.0 10.0 14.0 13.0 14.0  8.0 14.0 20.0 10.0 10.0 10.0 25.0 10.0 13.0
##  [61] 13.0 13.0 10.0 13.0 12.0  9.0 15.0  7.0 10.0 10.0  5.0 22.0 12.0 12.0 17.0
##  [76] 10.0 10.0 10.0 10.0 13.0  7.0 10.0 15.0  8.0 10.0 25.0 13.0 14.0 12.0  4.0
##  [91] 13.0 14.0 14.0 10.0 11.0 12.0 13.0 13.0 20.0 13.0 10.0 10.0 10.0  7.0 13.0
## [106] 19.0 11.0 20.0  8.0 25.0 16.0 15.0 15.0 10.0 10.0  8.0  4.0  9.0 14.0 15.0
## [121] 23.0 14.0 10.0   NA 11.0   NA 15.0 18.0 17.0 15.0 20.0 13.0 13.0 14.0  7.0
## [136] 13.0  4.0  4.0 10.0  8.0  6.0 12.0 10.0 14.0  8.0 10.0 13.0  5.0  7.0  6.0
## [151]  5.0 13.0  8.0  9.0 15.0 13.0   NA
```

```
ggplot(data = acacia, mapping = aes(x = CIRC, y = HEIGHT)) +
  geom_point(size = 1, color = "blue", alpha = 0.1) +
  labs(x = "Circumference", y = "Height", title = "Data from UHURU Acacia survey")
```

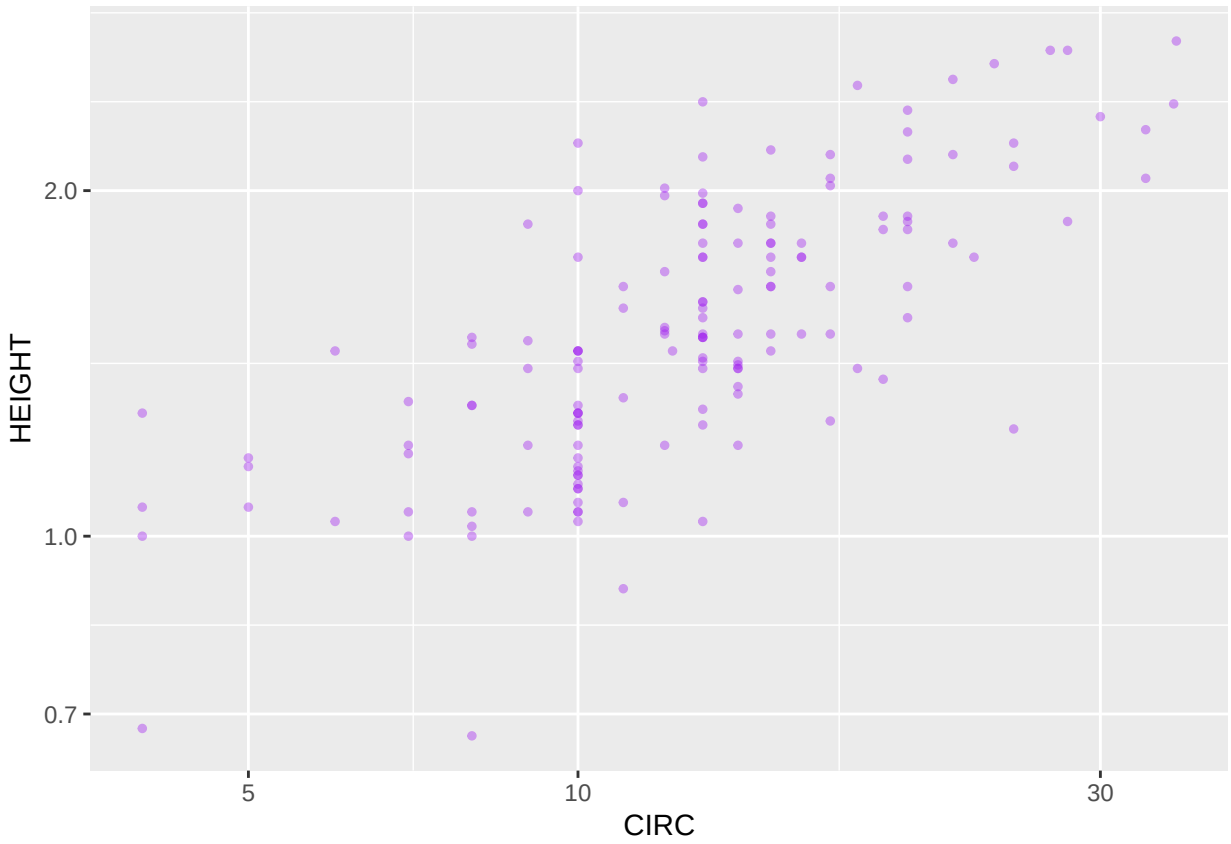
```
## Warning: Removed 4 rows containing missing values (geom_point).
```



To rescale the plotting of the axis to log scale we use the function `scale_y_log_10()`

```
ggplot( data = acacia, mapping = aes(x = CIRC, y = HEIGHT)) +
  geom_point(size =1, color = "purple", alpha = 0.4) +
  scale_x_log10() +
  scale_y_log10()
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



We have the information on experimental treatment in th etreatment column:

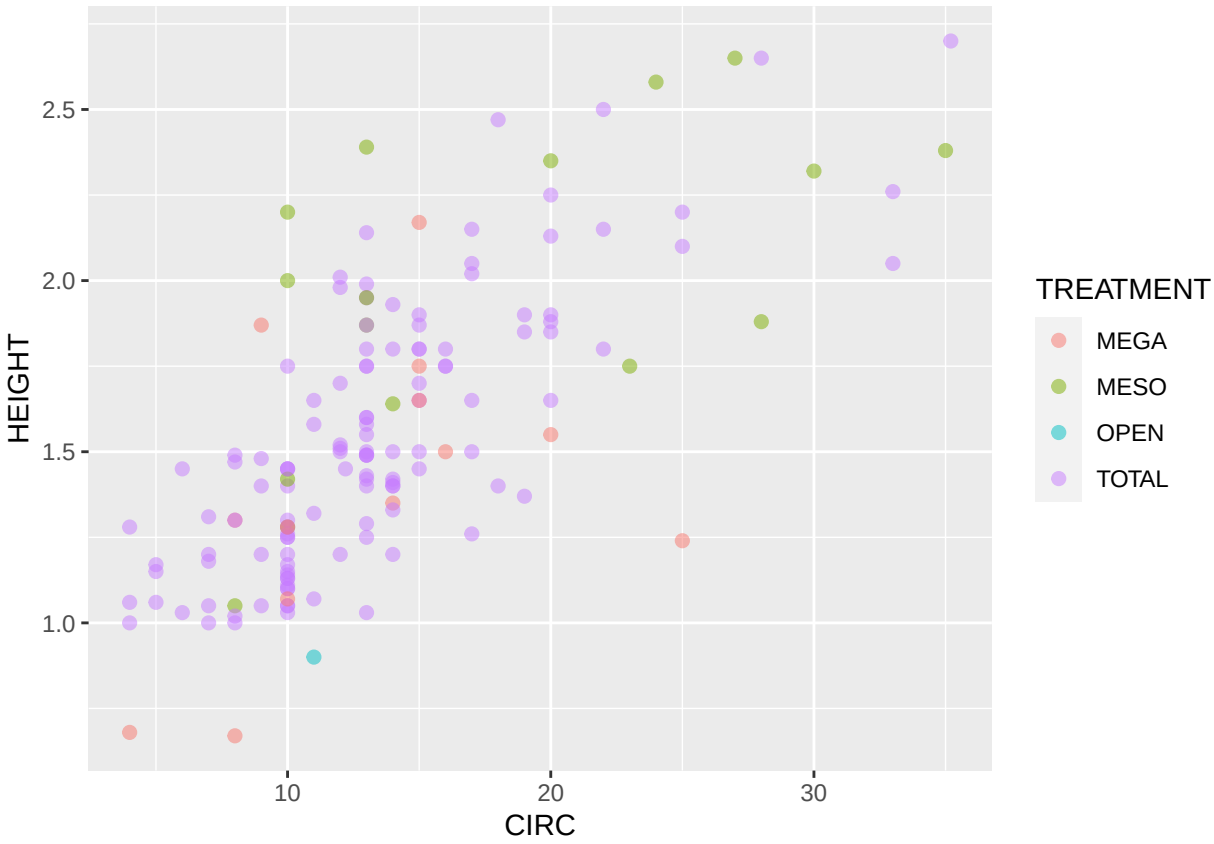
```
acacia$TREATMENT
```

```
##   [1] "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "MESO"
##  [10] "MESO"  "MESO"  "MESO"  "MESO"  "MESO"  "MESO"  "MESO"  "MESO"  "MESO"
##  [19] "MESO"  "MESO"  "OPEN"  "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL"
##  [28] "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL"
##  [37] "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL"
##  [46] "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL"
##  [55] "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL"
##  [64] "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL"
##  [73] "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL"
##  [82] "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL"
##  [91] "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL"
## [100] "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "MEGA"
## [109] "MEGA"  "MEGA"  "MEGA"  "MEGA"  "MEGA"  "MEGA"  "MEGA"  "MEGA"  "MEGA"
## [118] "MEGA"  "MEGA"  "MEGA"  "MESO"  "MESO"  "MESO"  "OPEN"  "OPEN"  "TOTAL"
## [127] "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL"
## [136] "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL"
## [145] "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL" "TOTAL"
## [154] "TOTAL" "TOTAL" "MESO"  "MESO"
```

Let's add information on treatment to our plot:

```
ggplot(data = acacia, mapping = aes (x = CIRC, y = HEIGHT, color = TREATMENT)) +
  geom_point(size = 2, alpha = 0.5)
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



4.2 Visualize a statistical analysis of correlation