uhuru-dataset-visualization.Rmd

Marcos Padilla-Ruiz

2022-10-04

1. Describing the data we are using

add a picture of an acacia ### 2. reading the data table into R

```
getwd()
```

```
## [1] "/Users/marcos/Desktop/BIO 197/Data_Science_Project/scripts"
```

```
acacia <- read.csv(file = "/Users/marcos/Desktop/BIO 197/Data_Science_Project/raw_data/ACACIA_DREPANOLO
```

3. explore our data

```
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
```

```
tail(acacia)
```

```
##     SURVEY YEAR  SITE BLOCK TREATMENT    PLOT   ID HEIGHT AXIS1 AXIS2 CIRC
## 152      1 2012 SOUTH     3     TOTAL S3TOTAL 2175   1.42  1.45  1.30   13
## 153      1 2012 SOUTH     3     TOTAL S3TOTAL 2176   1.02  1.20  1.00    8
## 154      1 2012 SOUTH     3     TOTAL S3TOTAL 2177    1.4  1.20  1.00    9
## 155      1 2012 SOUTH     3     TOTAL S3TOTAL 2178   1.45  2.10  2.05   15
## 156      1 2012 SOUTH     3      MESO  S3MESO 1421   1.95  2.20  1.60   13
## 157      1 2012 SOUTH     3      MESO  S3MESO 1422   dead    NA    NA   NA
##     FLOWERS BUDS FRUITS ANT
## 152       0    0      0  TP
## 153       0    0      0  TP
## 154       0    0      0  TP
## 155       0    0     20  TP
## 156       0    0      2  CS
## 157      NA   NA     NA
```

```
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"
```

```
nrow(acacia)
```

```
## [1] 157
```

make sure that everything that is a number is actually numeric use functon summary to do this and check that the type of data corresponds to this nanother wya is to use th 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"
```

idenitifed a colummn thta has problmeatic data so wee need to fix it so were gonna read the data table and assign “NA” to “dead” value in the height column

```
acacia <- read.csv(file = "/Users/marcos/Desktop/BIO 197/Data_Science_Project/raw_data/ACACIA_DREPANOLO
                   sep = "\t", na.strings = "dead")
getwd()
```

```
## [1] "/Users/marcos/Desktop/BIO 197/Data_Science_Project/scripts"
```

```
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
```

4. visualize our data

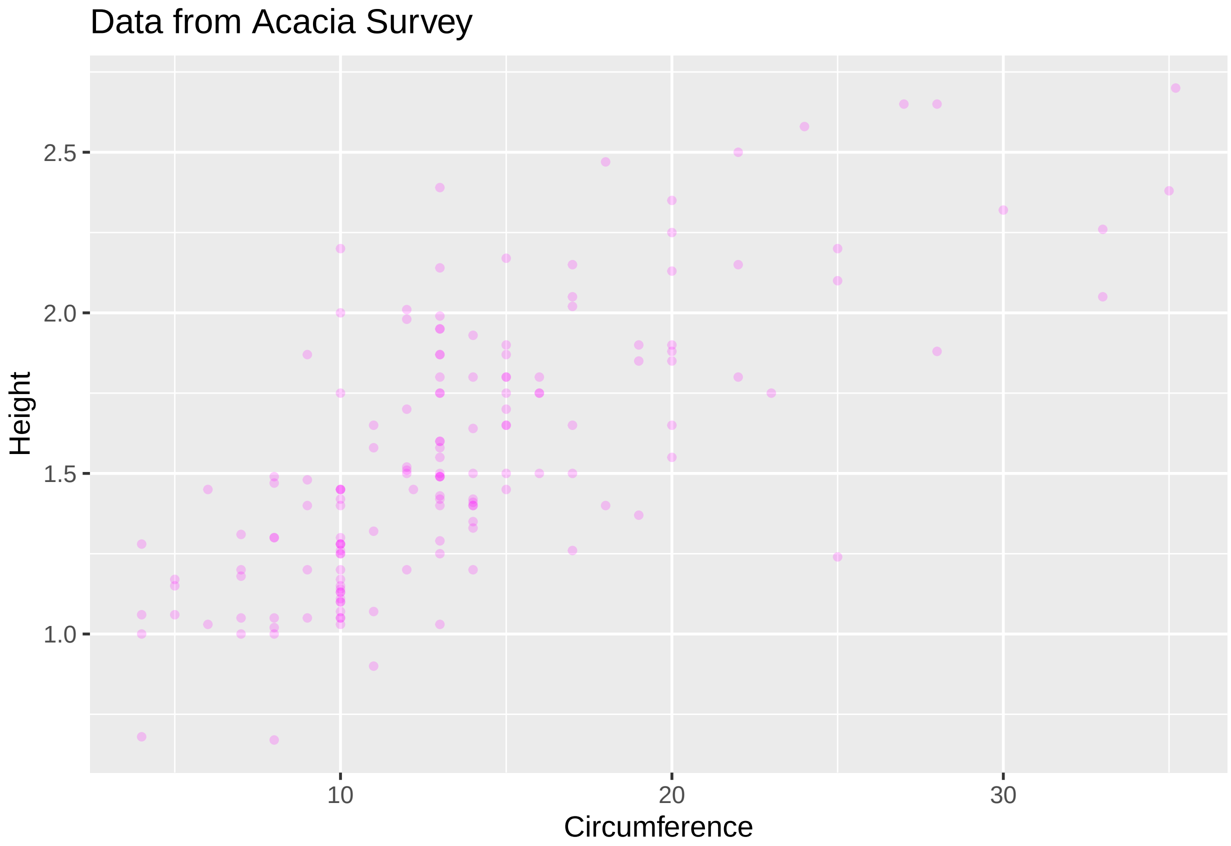
for this we use ggplot package let install it first

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

now we are gonna creat oiuur first plotting layer with ggplot function

```
ggplot(data = acacia, mapping = aes(x = CIRC, y = HEIGHT)) +
  geom_point(size = 1, col= "magenta", alpha = 0.2) +
  labs(x = "Circumference", y = "Height", title = "Data from Acacia Survey")
```

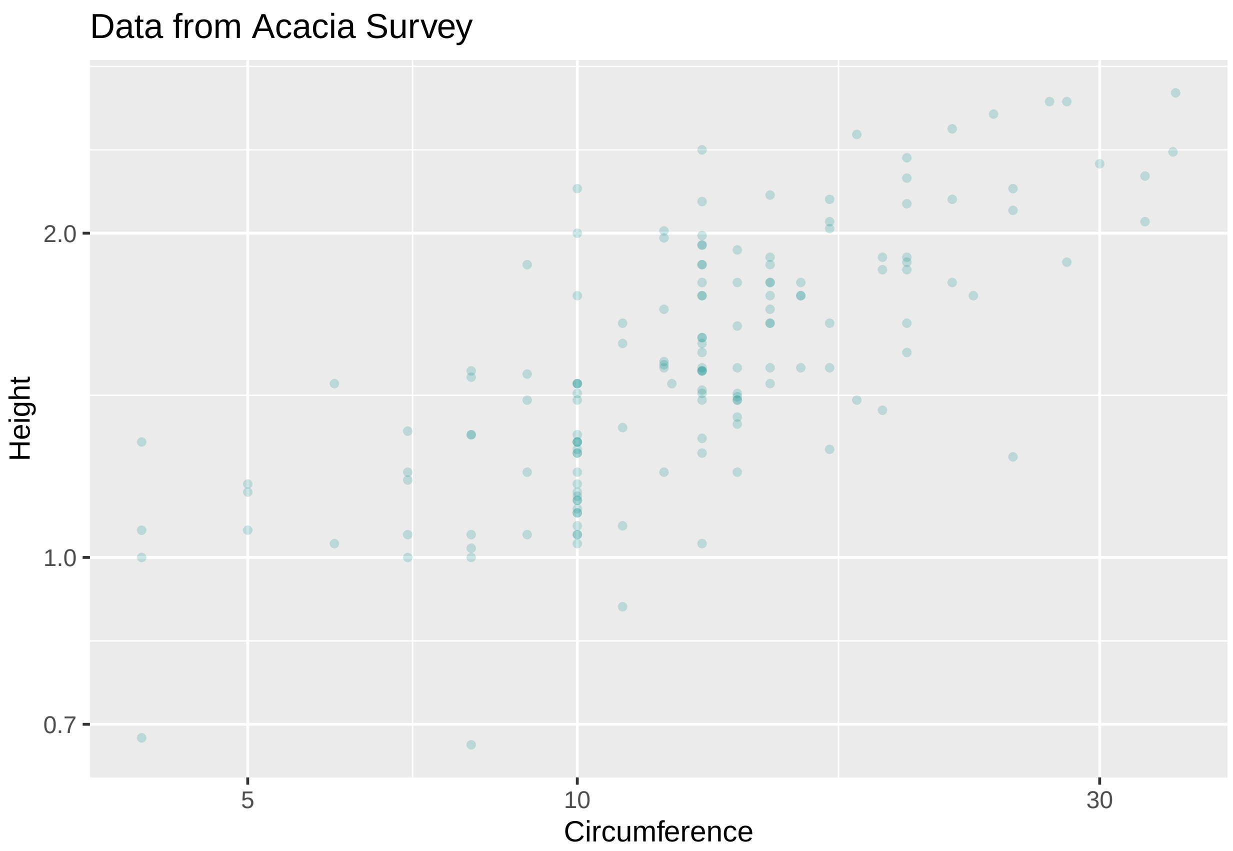
```
## Warning: Removed 4 rows containing missing values (geom_point).
```



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, col= "cyan4", alpha = 0.2) +
  scale_x_log10() +
  scale_y_log10() +
  labs(x = "Circumference", y = "Height", title = "Data from Acacia Survey")
```

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



we have the informaiton pon experimental treatmenet in treatmnt column

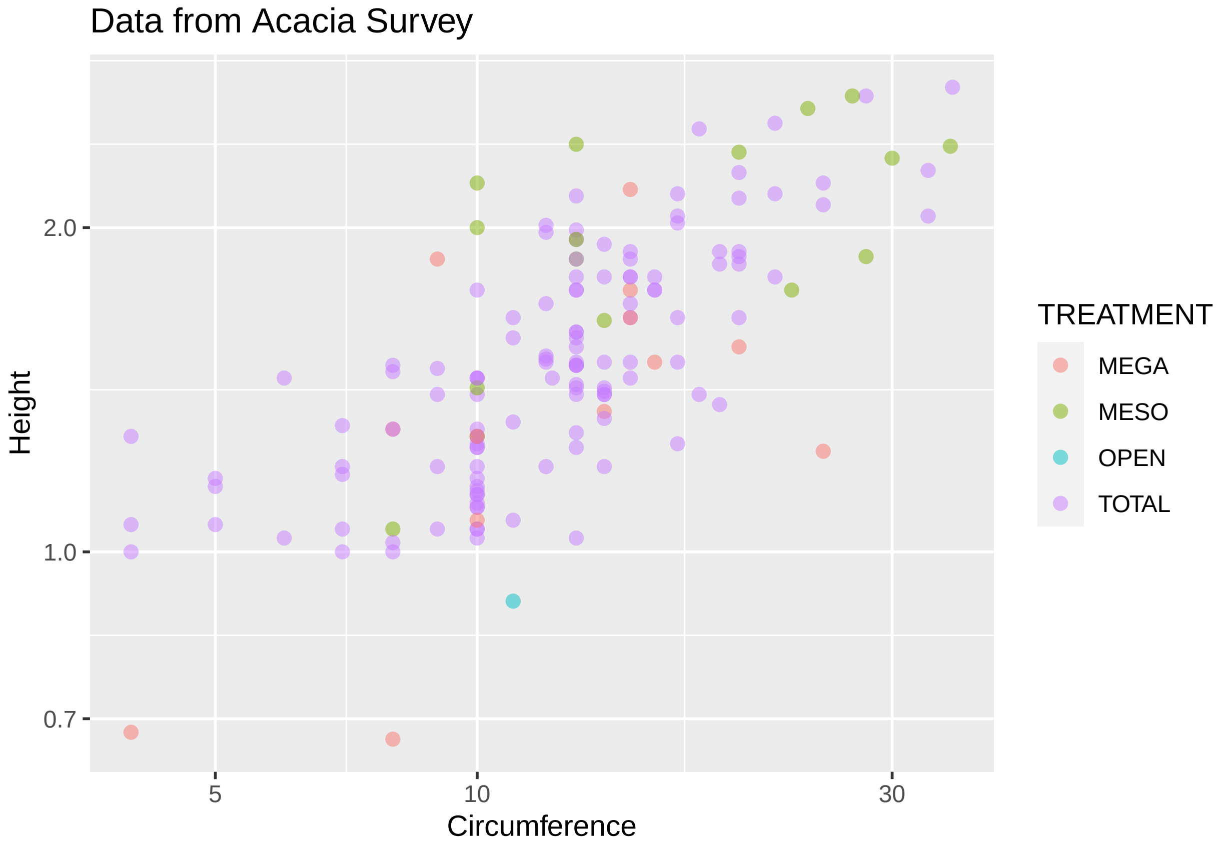
```
acacia$TREATMENT
```

```
##   [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"
```

lets add this information to our plot:

```
ggplot(acacia, mapping = aes (x = CIRC, y = HEIGHT, color = TREATMENT)) +
  geom_point(size = 2, alpha = 0.5) +
  labs(x = "Circumference", y = "Height", title = "Data from Acacia Survey") +
  scale_x_log10() +
  scale_y_log10()
```

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



4.2 visualize a statistical anlaysis of correlation