phenotypes

2023-10-10

Cleaning Phenotypes

K. Uckele October 10, 2023

```
##
## 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
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v readr     2.1.5
## v ggplot2   3.5.1     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
##
## Attaching package: 'EnvStats'
##
##
## The following objects are masked from 'package:stats':
##
##     predict, predict.lm
```

Red area (nectar guide data)

Load and filter data

```
redarea <- read_sheet(ss="19XHi3K57mDi2BpMPrlN0R2w16VSwV0YP6WYKYJHvuqs", sheet="red_area")
```

```
## ! Using an auto-discovered, cached token.
```

```
##   To suppress this message, modify your code or options to clearly consent to
##   the use of a cached token.
```

```
##   See gargle's "Non-interactive auth" vignette for more details:
```

```
##   <https://gargle.r-lib.org/articles/non-interactive-auth.html>
```

```
## i The googlesheets4 package is using a cached token for 'kuckele@ucsc.edu'.
## Auto-refreshing stale OAuth token.

## v Reading from "F2_phenotypes".

## v Range ''red_area''.
```

```
#make a new column that makes a unique ID for each plant
redarea <- mutate(redarea, unique_ID = paste0(plant_type, "_", ID))

#exclude parents and F1s
redarea <- filter(redarea, plant_type != "F1", plant_type != "P")

#exclude columns we don't need
redarea <- redarea %>% dplyr::select(-date, -photo_set, -plant_type, -ID, -labellum_photo, -stamen_phot

#rename columns
redarea <- redarea %>% dplyr::rename("RALA" = "red_labellum",
                                   "RAST" = "red_stamen")

#take a quick look at data frame
redarea
```

```
## # A tibble: 230 x 3
##        RALA  RAST unique_ID
##       <dbl> <dbl> <chr>
##  1   0      0    39_2
##  2   0.002  0    39_3
##  3  63.4    4.92 39_5
##  4  54.3   12.3  39_5
##  5  71.4    3.6  39_6
##  6 109.     9.3  39_6
##  7   4.7    1.01 39_9
##  8   1.4    1.1  39_9
##  9   0      0    39_10
## 10   0      0    39_10
## # i 220 more rows
```

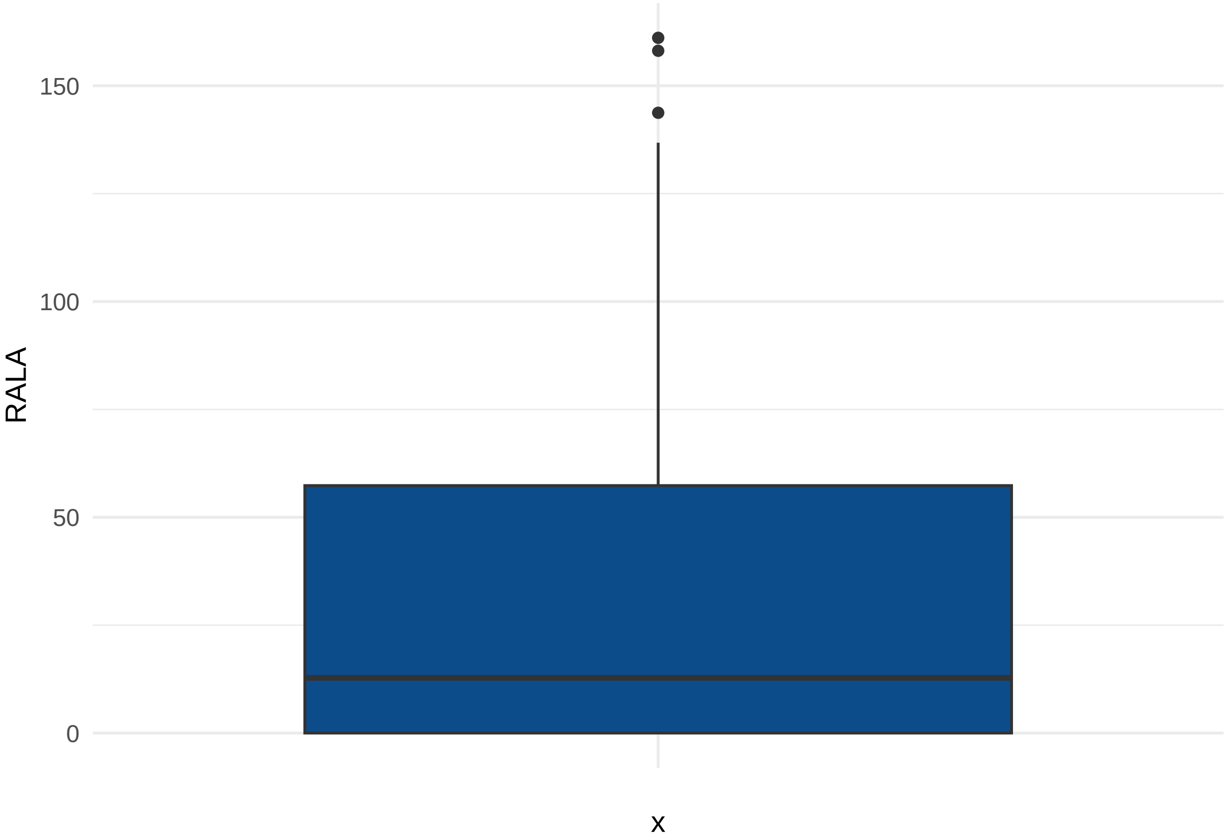
Collapse the replicate observations

```
#First, make a function to calculate the mode of categorical variables
#this function will output NA if there is a tie
Modes <- function(x) {
  ux <- unique(na.omit(x))
  tab <- tabulate(match(x, ux))
  if(sum(tab == max(tab)) == 2) {NA}
  else {ux[tab == max(tab)]}
}

#collapse the replicates by taking means
#first take the mean of continuous data
redarea_by_ID <- redarea %>% group_by(unique_ID) %>% reframe(
  tibble(
    across(where(is.double), \(x) mean(x, na.rm = TRUE)),
    across(where(is.character), Modes),
    across(where(is.factor), Modes)
  )
)
```

Identify potential outliers

```
## Boxplot of red labellum
ggplot(redarea_by_ID) +
  aes(x = "", y = RALA) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
## Identify potential outliers using the IQR criterion
# print outlier values
sort(boxplot.stats(redarea_by_ID$RALA)$out, decreasing = TRUE)
```

```
## [1] 161.1000 158.1000 143.7333
```

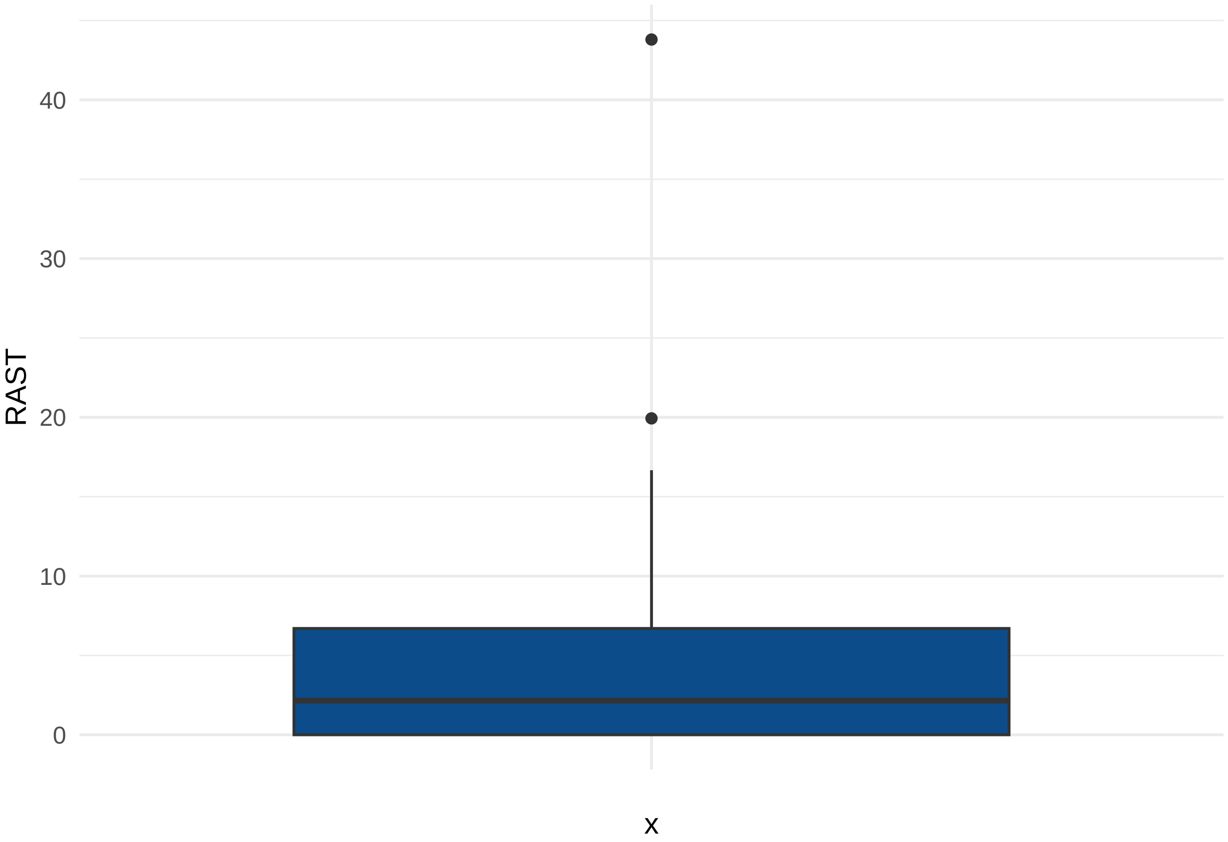
```
#Rosner test
rosnerTest(redarea_by_ID$RALA, k=length(boxplot.stats(redarea_by_ID$RALA)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
```

```
## Data:                            redarea_by_ID$RALA
##
## Sample Size:                     129
##
## Test Statistics:                 R.1 = 3.090004
##                                  R.2 = 3.150395
##                                  R.3 = 2.921505
##
## Test Statistic Parameter:        k = 3
##
## Alternative Hypothesis:          Up to 3 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
##
## Number of Outliers Detected:     0
##
##   i   Mean.i     SD.i    Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 33.37575 41.33465 161.1000      37 3.090004   3.468769   FALSE
## 2 1 32.37791 39.90677 158.1000      93 3.150395   3.466243   FALSE
## 3 2 31.38797 38.45461 143.7333      38 2.921505   3.463694   FALSE
```

```r
#no outliers detected

## Boxplot of red stamen
ggplot(redarea_by_ID) +
  aes(x = "", y = RAST) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values
boxplot.stats(redarea_by_ID$RAST)$out
```

```
## [1] 19.93333 43.80000
```

```
#Rosner test
rosnerTest(redarea_by_ID$RAST, k=length(boxplot.stats(redarea_by_ID$RAST)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            redarea_by_ID$RAST
##
## Sample Size:                     129
##
## Test Statistics:                 R.1 = 6.944386
##                                  R.2 = 3.566435
##
## Test Statistic Parameter:        k = 2
##
## Alternative Hypothesis:          Up to 2 observations are not
##                                  from the same Distribution.
##
```

```
## Type I Error:                    5%
##
## Number of Outliers Detected:     2
##
##   i   Mean.i     SD.i    Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 4.137136 5.711500 43.80000     106 6.944386   3.468769    TRUE
## 2 1 3.827270 4.516013 19.93333     105 3.566435   3.466243    TRUE
```

```
#two outliers detected
## Remove outliers
redarea_by_ID <- redarea_by_ID %>%
  mutate(RAST = na_if(RAST, 43.800)) %>%
  mutate(RAST = na_if(RAST, 19.933))
```

Inflorescence-level data

Load and filter data

```
#get access to google sheets
gs4_auth(email = "kuckele@ucsc.edu")
#read in the tab of interest and convert it to a commonly named df
inflor <- read_sheet(ss="19XHi3K57mDi2BpMPrlNOR2w16VSwVOYP6WYKYJHvuqs", sheet="inflorescences")
```

```
## v Reading from "F2_phenotypes".
```

```
## v Range ''inflorescences''.
```

```
#make a new column that makes a unique ID for each plant
inflor <- mutate(inflor, unique_ID = paste0(plant_type, "_", ID))

#exclude parents and F1s
inflor <- filter(inflor, plant_type != "F1", plant_type != "P")

#exclude columns we don't need
inflor <- inflor %>% dplyr::select(-date, -rep, -plant_type, -ID, -typeID)

#make sure categorical variables are factors
inflor$visible_EFnectar <- as.factor(inflor$visible_EFnectar)
inflor$visible_guides <- as.factor(inflor$visible_guides)

#rename columns
inflor <- inflor %>% dplyr::rename("INFA" = "infl_angle",
                                   "VEFN" = "visible_EFnectar",
                                   "CAL" = "callus_length",
                                   "VNG" = "visible_guides")

#take a quick look at data frame
inflor
```

```
## # A tibble: 588 x 5
##     INFA VEFN    CAL VNG   unique_ID
##    <dbl> <fct> <dbl> <fct> <chr>
##  1    25 1      4.89 0     39_2
##  2    20 1      3.9  0     39_2
##  3    70 1      4.1  0     39_2
##  4   100 1      5.27 0     39_2
##  5    18 1      4.4  0     39_3
##  6    15 1      5.19 0     39_3
##  7     0 1      3.48 0     39_3
##  8    37 1      4.09 1     39_4
##  9    55 0      4.7  0     39_5
## 10     5 1      4.81 1     39_5
## # i 578 more rows
```

Collapse the replicate observations

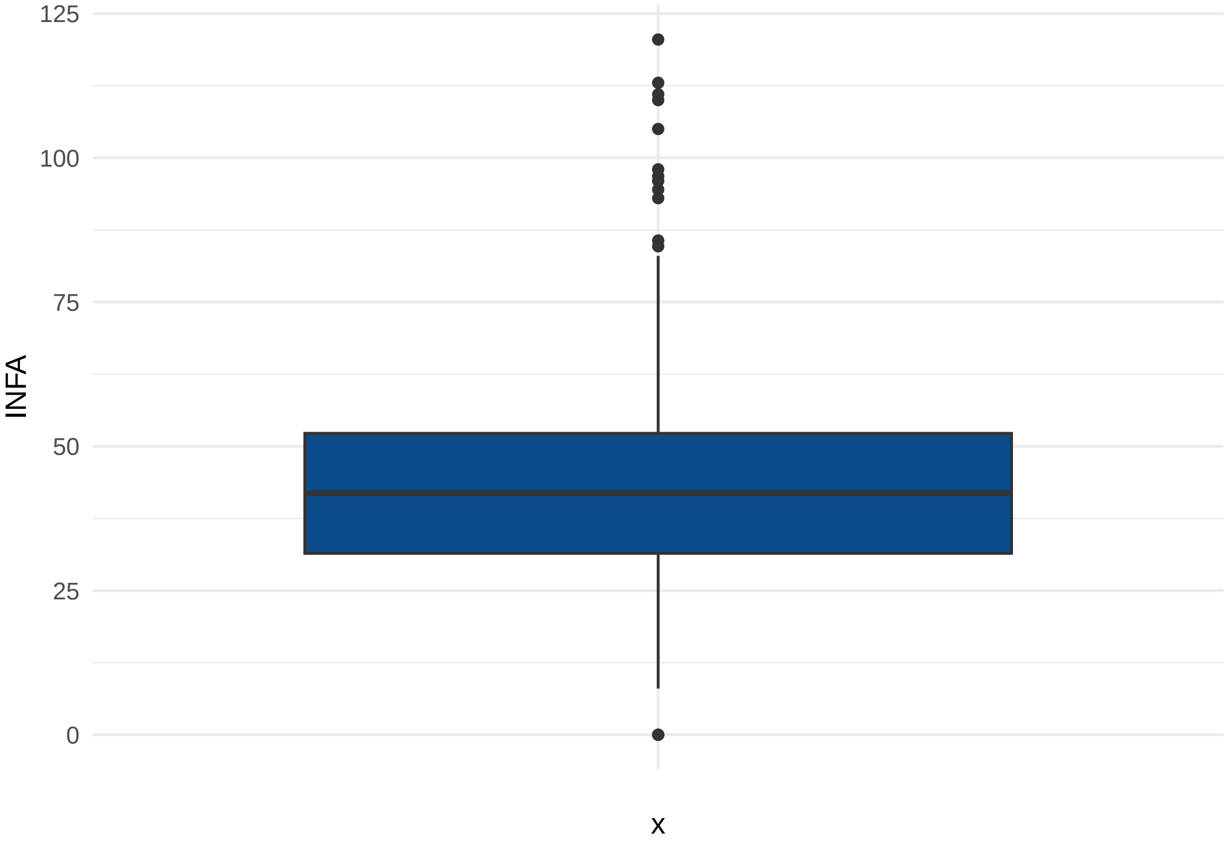
```
#collapse the replicates by taking means and modes
#take the mean of continuous data and the mode of the factors and characters
inflor_by_ID <- inflor %>% group_by(unique_ID) %>% reframe(
  tibble(
    across(where(is.double), \(x) mean(x, na.rm = TRUE)),
    across(where(is.character), Modes),
    across(where(is.factor), Modes)
  )
)
```

Fill in VNG based on red area data

```
# Join the two data frames by the unique_ID column
inflor_by_ID <- inflor_by_ID %>%
  left_join(redarea_by_ID, by = "unique_ID") %>%
  mutate(VNG = case_when(
    is.na(VNG) & (RAST > 0 | RALA > 0) ~ factor(1, levels = c(0, 1)),
    is.na(VNG) & RAST == 0 & RALA == 0 ~ factor(0, levels = c(0, 1)),
    TRUE ~ VNG  # Keep the original value if the condition is not met
  )) %>%
  # Remove the columns from redarea_by_ID if no longer needed
  select(-RAST, -RALA)
```

Identify potential outliers

```
## Boxplot of inflorescence angle
ggplot(inflor_by_ID) +
  aes(x = "", y = INFA) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values based on IQR criterion
sort(boxplot.stats(inflor_by_ID$INFA)$out, decreasing = TRUE)
```

```
##  [1] 120.50000 113.00000 111.00000 110.00000 105.00000  98.00000  96.80000
##  [8]  96.00000  94.50000  93.00000  85.66667  84.66667   0.00000   0.00000
## [15]   0.00000
```

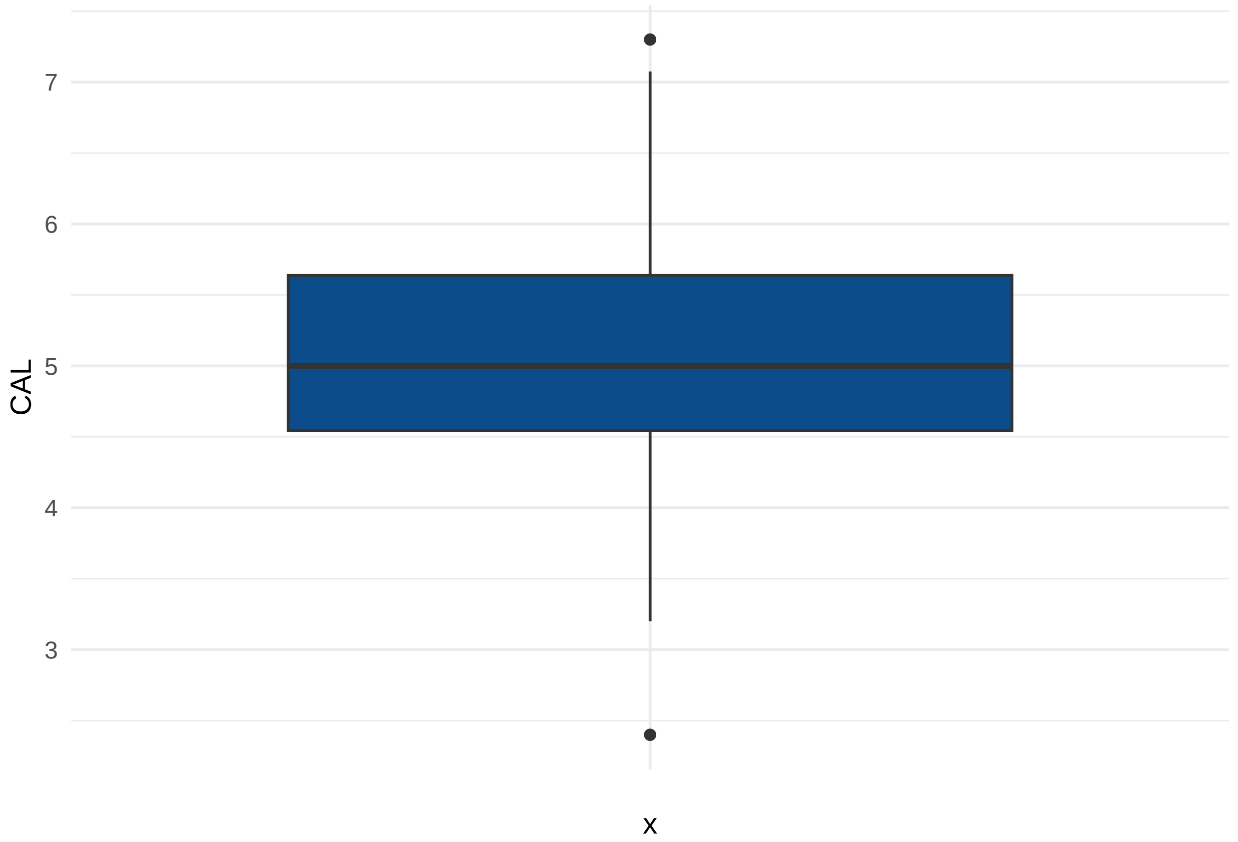
```
#Rosner test
rosnerTest(inflor_by_ID$INFA, k=length(boxplot.stats(inflor_by_ID$INFA)$out))
```

```
## Warning in rosnerTest(inflor_by_ID$INFA, k = length(boxplot.stats(inflor_by_ID$INFA)$out)): The true
## Although the help file for 'rosnerTest' has a table with information
## on the estimated Type I error level,
## simulations were not run for k > 10 or k > floor(n/2).
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            inflor_by_ID$INFA
##
## Sample Size:                     202
##
## Test Statistics:                 R.1  = 3.575298
##                                  R.2  = 3.342445
##                                  R.3  = 3.349837
##                                  R.4  = 3.406706
##                                  R.5  = 3.258849
##                                  R.6  = 2.983606
##                                  R.7  = 2.996105
##                                  R.8  = 3.030958
##                                  R.9  = 3.027914
##                                  R.10 = 3.023179
##                                  R.11 = 2.667927
##                                  R.12 = 2.665660
##                                  R.13 = 2.621376
##                                  R.14 = 2.646271
##                                  R.15 = 2.672478
##
## Test Statistic Parameter:        k = 15
##
## Alternative Hypothesis:          Up to 15 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
##
## Number of Outliers Detected:     0
##
##     i   Mean.i     SD.i     Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1   0 44.13531 21.35897 120.50000     201 3.575298   3.608511   FALSE
## 2   1 43.75539 20.71675 113.00000     135 3.342445   3.607023   FALSE
## 3   2 43.40917 20.17735 111.00000      98 3.349837   3.605525   FALSE
## 4   3 43.06951 19.64669 110.00000     129 3.406706   3.604019   FALSE
## 5   4 42.73148 19.10752 105.00000     128 3.258849   3.602505   FALSE
## 6   5 42.41540 18.63001  98.00000      62 2.983606   3.600981   FALSE
## 7   6 42.13180 18.24642  96.80000      72 2.996105   3.599448   FALSE
## 8   7 41.85145 17.86516  96.00000     163 3.030958   3.597906   FALSE
## 9   8 41.57234 17.47991  94.50000     107 3.027914   3.596355   FALSE
## 10  9 41.29810 17.10183  93.00000     125 3.023179   3.594795   FALSE
## 11 10 41.02882 16.73128  85.66667     167 2.667927   3.593225   FALSE
## 12 11 40.79511 16.45804  84.66667     138 2.665660   3.591646   FALSE
## 13 12 40.56421 16.18836  83.00000      90 2.621376   3.590057   FALSE
## 14 13 40.33968 15.93197  82.50000     121 2.646271   3.588458   FALSE
## 15 14 40.11543 15.67256  82.00000      15 2.672478   3.586849   FALSE
```

```r
#no outliers detected based on Rosner test

## Boxplot of callus length
ggplot(inflor_by_ID) +
  aes(x = "", y = CAL) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values based on IQR criterion
boxplot.stats(inflor_by_ID$CAL)$out
```

```
## [1] 7.3 2.4
```

```
#Rosner test
rosnerTest(inflor_by_ID$CAL, k=length(boxplot.stats(inflor_by_ID$CAL)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            inflor_by_ID$CAL
##
## Sample Size:                     202
##
## Test Statistics:                 R.1 = 3.118932
##                                  R.2 = 2.652569
##
## Test Statistic Parameter:        k = 2
##
## Alternative Hypothesis:          Up to 2 observations are not
##                                  from the same Distribution.
##
```

```
## Type I Error:                    5%
##
## Number of Outliers Detected:     0
##
##   i   Mean.i      SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 5.068003 0.8554221   2.4     140 3.118932   3.608511   FALSE
## 2 1 5.081277 0.8364433   7.3      65 2.652569   3.607023   FALSE
```

```
#no outliers detected based on Rosner test
```

Flower morphology data

Load and filter data

```
flo_morph <- read_sheet(ss="19XHi3K57mDi2BpMPrlNOR2w16VSwVOYP6WYKYJHvuqs", sheet="flower_morphology")
```

```
## v Reading from "F2_phenotypes".
```

```
## v Range ''flower_morphology''.
```

```
#make a new column that makes a unique ID for each plant
flo_morph <- mutate(flo_morph, unique_ID = paste0(plant_type, "_", ID))

#exclude parents and F1s
flo_morph <- filter(flo_morph, plant_type != "F1", plant_type != "P")

#exclude columns we don't need
flo_morph <- flo_morph %>% dplyr::select(-date, -plant_type, -ID, -rep)

#rename columns
flo_morph <- flo_morph %>% dplyr::rename("COL" = "Corolla_Length",
                                   "COLL" = "Corolla_Lobe_Length",
                                   "STAE" = "Stamen_exsertion",
                                   "TUA" = "tube_angle",
                                   "STATL" = "stamen_tip",
                                   "LABL" = "Labellum_Length",
                                   "LABW" = "Labellum_Width",
                                   "CLL" = "Labellum_lobe",
                                   "STAL" = "Stamen_Length",
                                   "STAW" = "Stamen_width",
                                   "ANL" = "Anther_Length",
                                   "ANW" = "Anther_width",
                                   "STIW" = "Stigma_Width",
                                   "STYL" = "Style_length")

#quick look at data
flo_morph
```

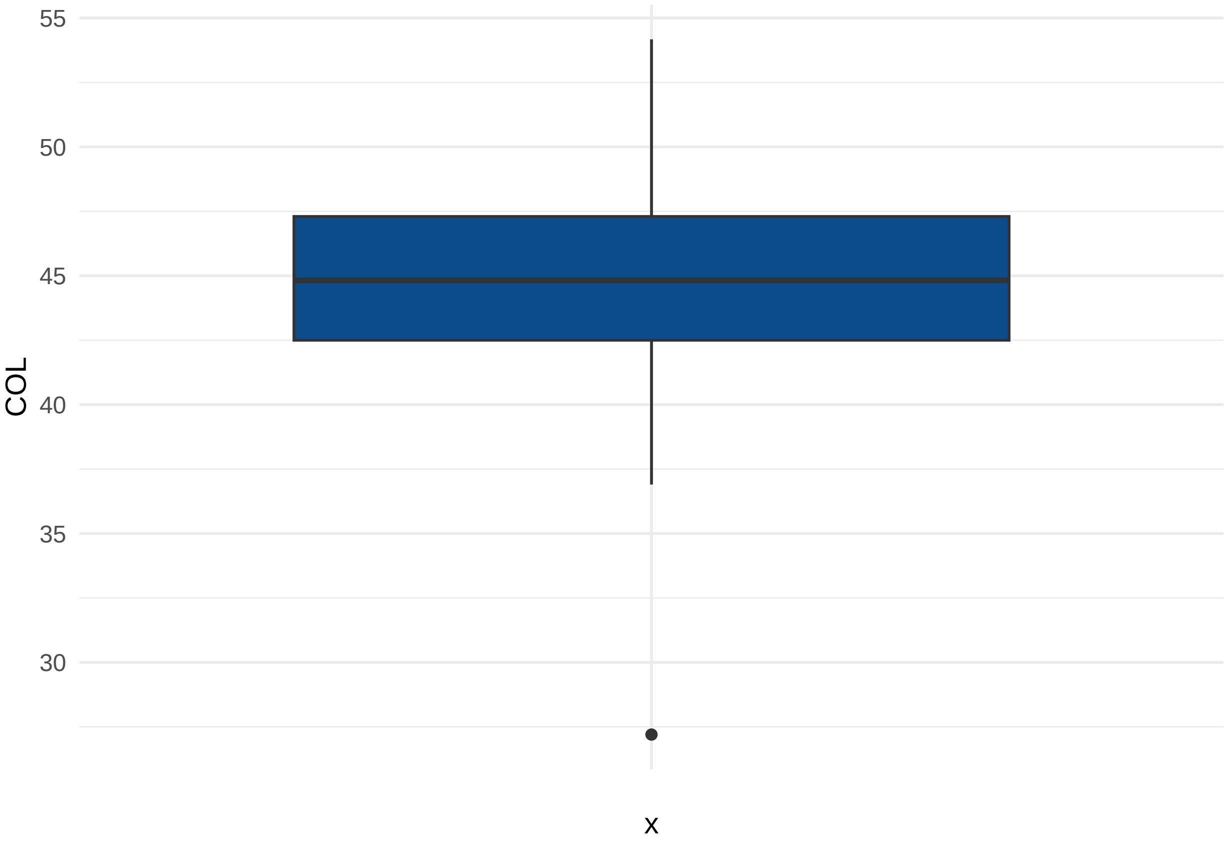
```
## # A tibble: 755 x 15
##      COL  COLL  STAE   TUA STATL  LABL  LABW   CLL  STAL  STAW   ANL   ANW  STIW
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  48.8  30.5  -2.3    24   8.2  54.4  29.6   7.5  52.4   9.4   6.4   2.7   2.6
##  2  52.1  35    -1.4    27   8.5  57.4  28.3   5.3  55.9   9.4   6.8   2.7   3
##  3  51.4  30.4  -2.4     7   8    58    25.5   6.6  55.6   9.4   6.6   2.8   2.7
##  4  45.7  27.4  -2.9    21   7.2  52.7  27.7   5.6  49.8   9.3   6.6   2.4   2.7
##  5  46.5  26.6  -2.1     5   6.5  53.3  26.9   6.7  51.2   8.6   6.7   2.4   2.8
##  6  47.3  28.3  -3.7    21   6.9  53.2  27.9   5.6  49.5   9.8   6.5   3.1   2.5
##  7  46.3  26.4  -1.9    13   7    50.8  22.3   4.2  48.9   8.6   6.4   2.4   2.3
##  8  50.6  29.6  -2.9    24  11.1  57.6  31.1   5    54.7   9.3   6.4   2.8   2.6
##  9  49.3  27.9  -3.1    11  11    58.1  31.2   6.9  55     9.5   6.8   2.8   3.2
## 10  51.9  32.7  -2.4    10  11.5  58.5  31.4   9    56.1   9.4   6.9   2.8   2.9
## # i 745 more rows
## # i 2 more variables: STYL <dbl>, unique_ID <chr>
```

Collapse the replicate observations

```
flo_morph_by_ID <- flo_morph %>% group_by(unique_ID) %>% reframe(
  tibble(
    across(where(is.double), \(x) mean(x, na.rm = TRUE)),
    across(where(is.character), Modes),
    across(where(is.factor), Modes)
  )
)
```

Identify potential outliers

```
## Corolla Length
ggplot(flo_morph_by_ID) +
  aes(x = "", y = COL) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values
sort(boxplot.stats(flo_morph_by_ID$COL)$out, decreasing = TRUE)
```

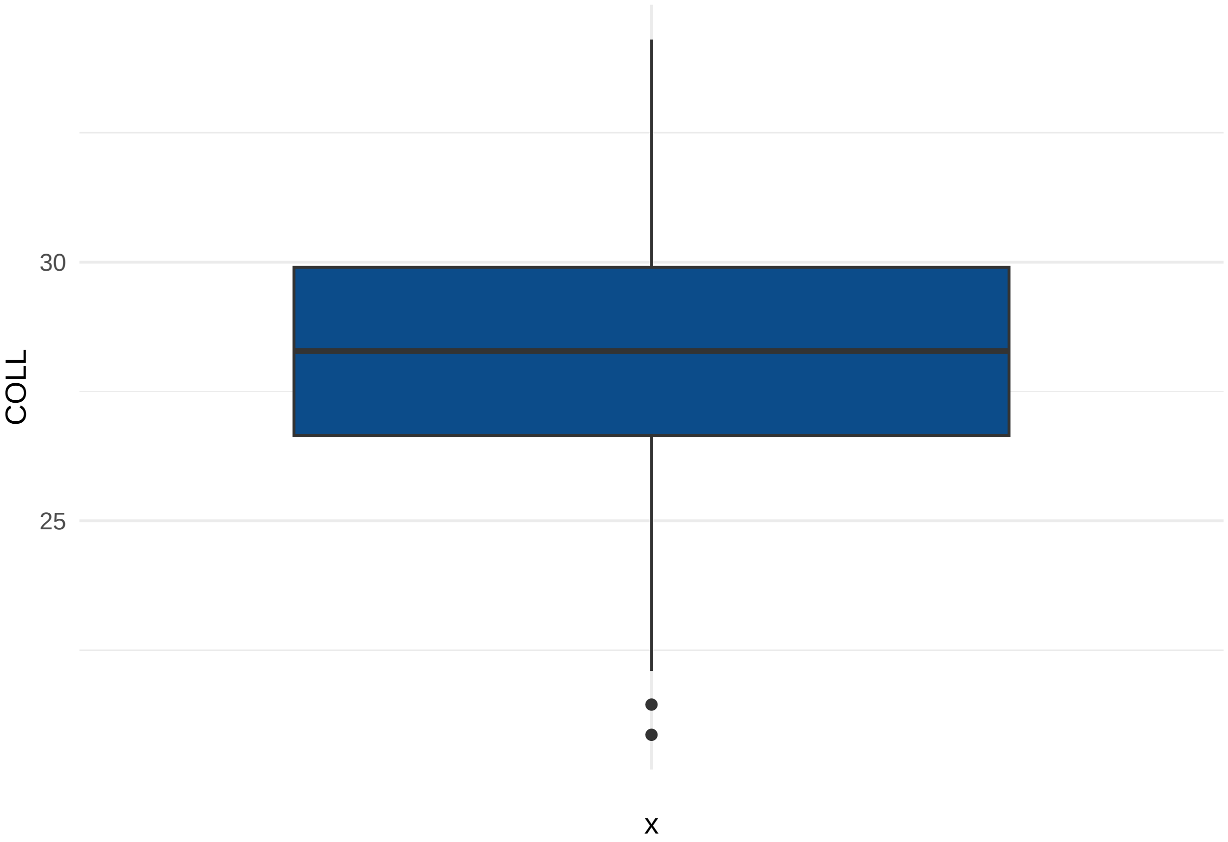
```
## [1] 27.2
```

```
#Rosner test
rosnerTest(flo_morph_by_ID$COL, k=length(boxplot.stats(flo_morph_by_ID$COL)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            flo_morph_by_ID$COL
##
## Sample Size:                     221
##
## Test Statistic:                  R.1 = 4.713177
##
## Test Statistic Parameter:        k = 1
##
## Alternative Hypothesis:          Up to 1 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
##
## Number of Outliers Detected:     1
##
##   i   Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 44.83215 3.741033  27.2     131 4.713177   3.635271    TRUE
```

```
#one outlier detected
## Remove outliers
flo_morph_by_ID <- flo_morph_by_ID %>%
  mutate(COL = na_if(COL, 27.2))

## Corolla_Lobe_Length
ggplot(flo_morph_by_ID) +
  aes(x = "", y = COLL) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values
sort(boxplot.stats(flo_morph_by_ID$COLL)$out, decreasing = TRUE)
```

```
## [1] 21.45000 20.86667
```

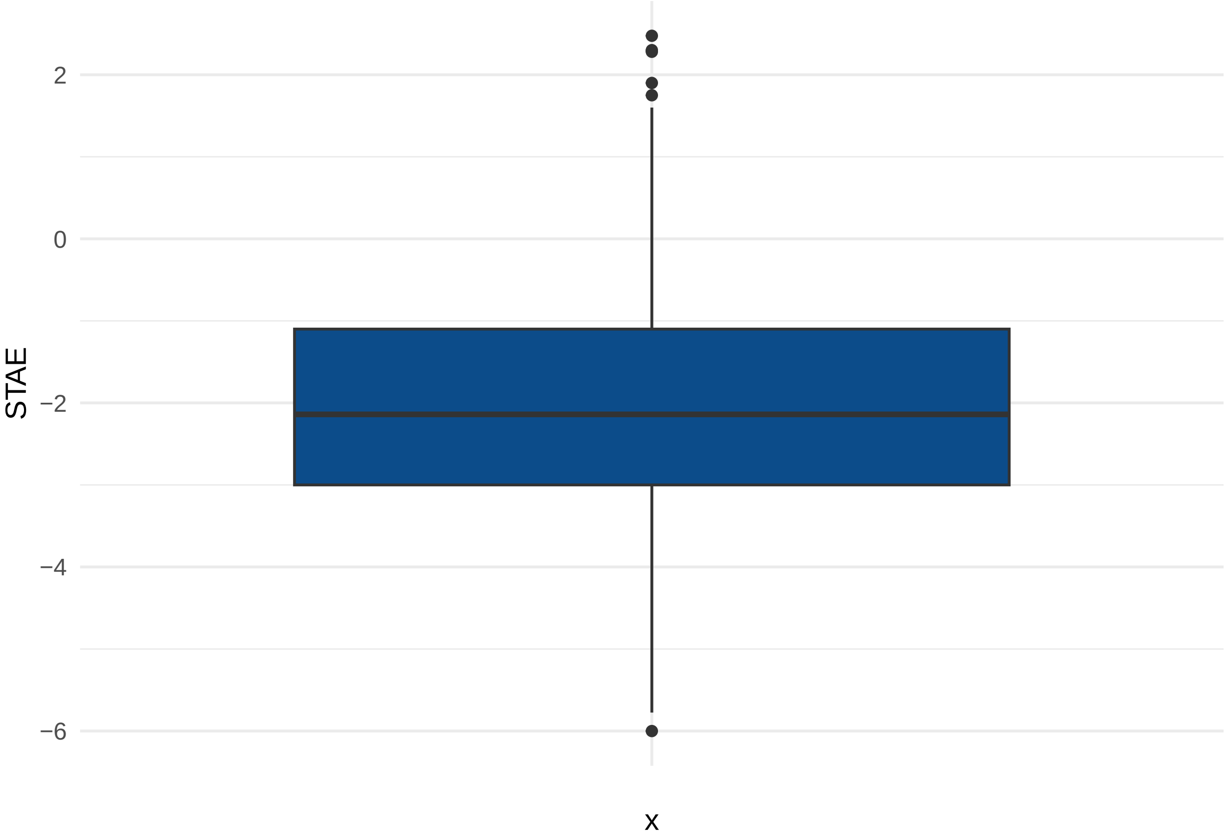
```
#Rosner test
rosnerTest(flo_morph_by_ID$COLL, k=length(boxplot.stats(flo_morph_by_ID$COLL)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            flo_morph_by_ID$COLL
##
## Sample Size:                     221
##
## Test Statistics:                 R.1 = 2.869232
##                                  R.2 = 2.699103
##
## Test Statistic Parameter:        k = 2
##
## Alternative Hypothesis:          Up to 2 observations are not
##                                  from the same Distribution.
##
```

```
## Type I Error:                     5%
##
## Number of Outliers Detected:      0
##
##   i   Mean.i     SD.i    Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 28.19631 2.554567 20.86667     189 2.869232   3.635271   FALSE
## 2 1 28.22963 2.511808 21.45000     145 2.699103   3.633930   FALSE
```

```
#no outliers detected

## Stamen_exsertion
ggplot(flo_morph_by_ID) +
  aes(x = "", y = STAE) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values
sort(boxplot.stats(flo_morph_by_ID$STAE)$out, decreasing = TRUE)
```

```
## [1]  2.475  2.300  2.280  1.900  1.750 -6.000
```

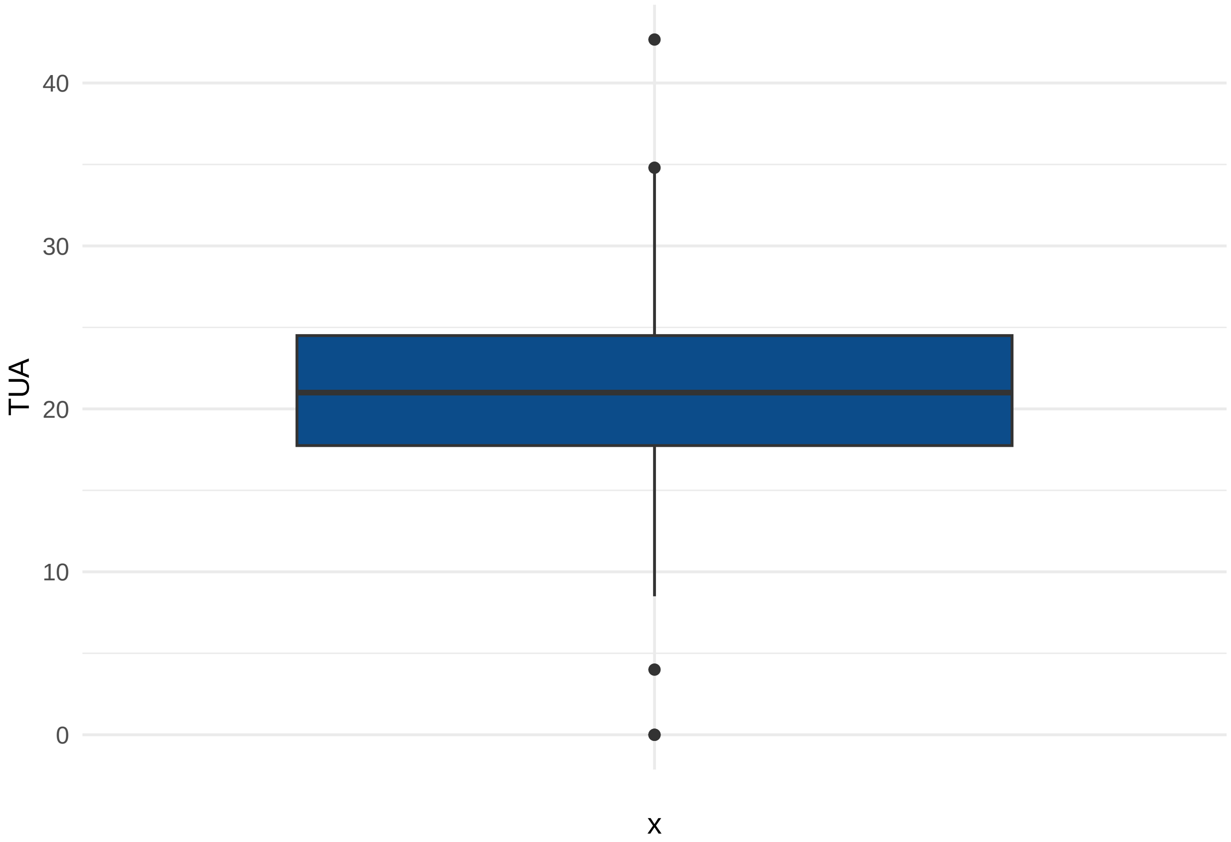
```
#Rosner test
rosnerTest(flo_morph_by_ID$STAE, k=length(boxplot.stats(flo_morph_by_ID$STAE)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
```

```
##
## Hypothesized Distribution:       Normal
##
## Data:                            flo_morph_by_ID$STAE
##
## Sample Size:                     221
##
## Test Statistics:                 R.1 = 2.893990
##                                  R.2 = 2.844152
##                                  R.3 = 2.891828
##                                  R.4 = 2.701007
##                                  R.5 = 2.651966
##                                  R.6 = 2.650546
##
## Test Statistic Parameter:        k = 6
##
## Alternative Hypothesis:          Up to 6 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
##
## Number of Outliers Detected:     0
##
##   i    Mean.i     SD.i  Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 -2.065339 1.568886  2.475     108 2.893990   3.635271   FALSE
## 2 1 -2.085977 1.542104  2.300     158 2.844152   3.633930   FALSE
## 3 2 -2.106005 1.516689  2.280     199 2.891828   3.632582   FALSE
## 4 3 -2.126124 1.490601  1.900      83 2.701007   3.631227   FALSE
## 5 4 -2.144677 1.468600  1.750     163 2.651966   3.629865   FALSE
## 6 5 -2.162708 1.447736 -6.000      68 2.650546   3.628495   FALSE
```

```r
#no outliers detected

## tube_angle
ggplot(flo_morph_by_ID) +
  aes(x = "", y = TUA) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```r
# print outlier values
sort(boxplot.stats(flo_morph_by_ID$TUA)$out, decreasing = TRUE)
```

```
## [1] 42.66667 34.80000  4.00000  0.00000  0.00000
```

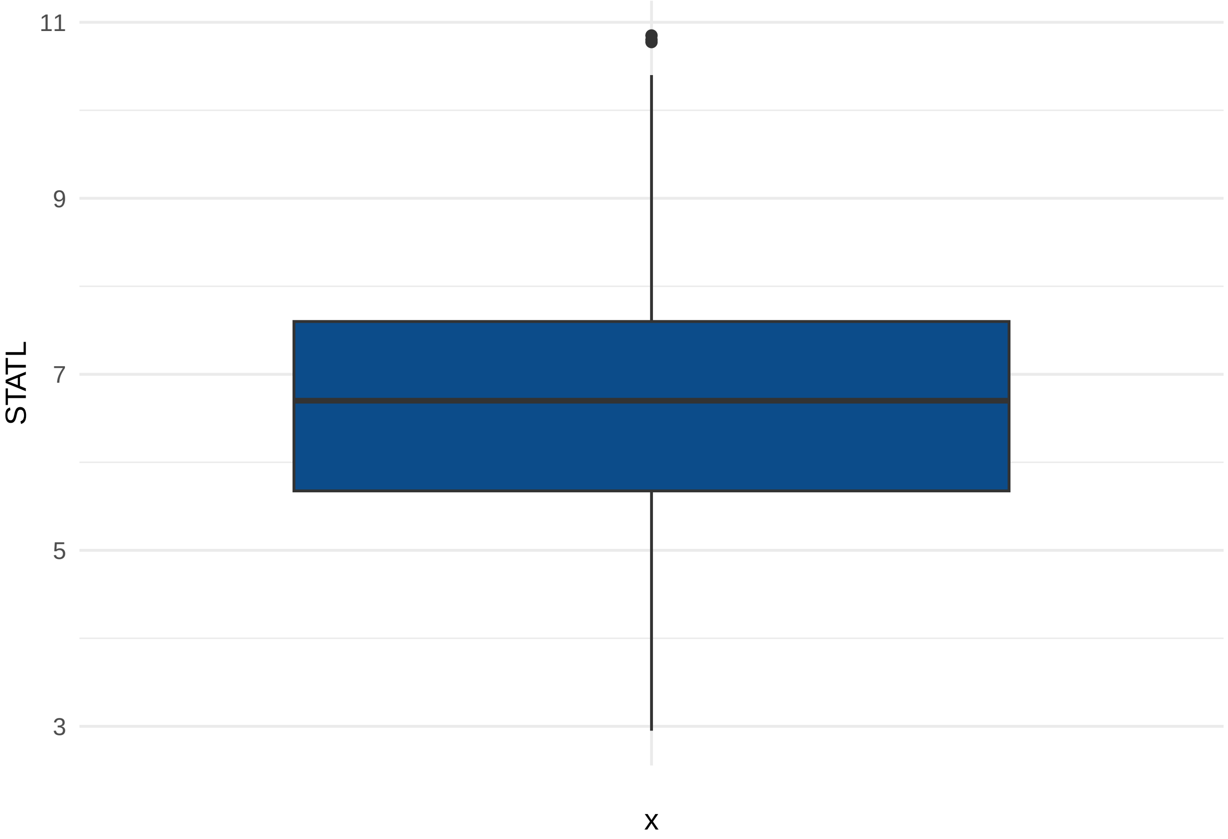
```r
#Rosner test
rosnerTest(flo_morph_by_ID$TUA, k=length(boxplot.stats(flo_morph_by_ID$TUA)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            flo_morph_by_ID$TUA
##
## Sample Size:                     221
##
## Test Statistics:                 R.1 = 3.571790
##                                  R.2 = 3.550304
##                                  R.3 = 3.665920
##                                  R.4 = 3.075028
##                                  R.5 = 2.493557
##
## Test Statistic Parameter:        k = 5
##
```

```
## Alternative Hypothesis:          Up to 5 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
##
## Number of Outliers Detected:     3
##
##   i   Mean.i     SD.i    Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 21.02323 6.059550 42.66667      54 3.571790   3.635271    TRUE
## 2 1 20.92485 5.893819  0.00000      30 3.550304   3.633930    TRUE
## 3 2 21.02040 5.734003  0.00000     131 3.665920   3.632582    TRUE
## 4 3 21.11682 5.566395  4.00000      51 3.075028   3.631227   FALSE
## 5 4 21.19570 5.455780 34.80000       1 2.493557   3.629865   FALSE
```

```
#three outliers detected
## Remove outliers
flo_morph_by_ID <- flo_morph_by_ID %>%
  mutate(TUA = round(TUA, digits=0)) %>%
  mutate(TUA = na_if(TUA, 43)) %>%
  mutate(TUA = na_if(TUA, 0))

## stamen_tip
ggplot(flo_morph_by_ID) +
  aes(x = "", y = STATL) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```r
# print outlier values
sort(boxplot.stats(flo_morph_by_ID$STATL)$out, decreasing = TRUE)
```

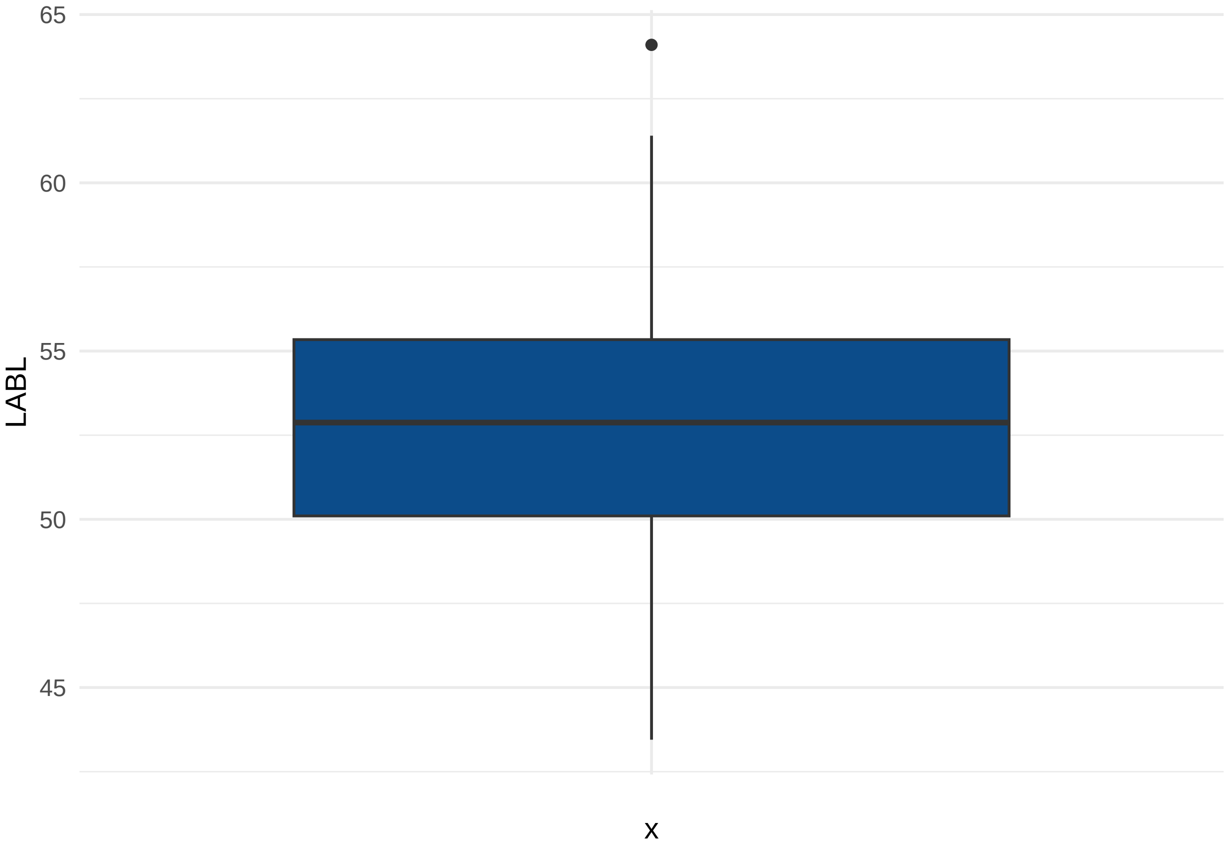
```
## [1] 10.850 10.800 10.775
```

```r
#Rosner test
rosnerTest(flo_morph_by_ID$STATL, k=length(boxplot.stats(flo_morph_by_ID$STATL)$out))
```

```
## 
## Results of Outlier Test
## -------------------------
## 
## Test Method:                     Rosner's Test for Outliers
## 
## Hypothesized Distribution:       Normal
## 
## Data:                            flo_morph_by_ID$STATL
## 
## Sample Size:                     221
## 
## Test Statistics:                 R.1 = 2.601881
##                                  R.2 = 2.616971
##                                  R.3 = 2.648860
## 
## Test Statistic Parameter:        k = 3
## 
## Alternative Hypothesis:          Up to 3 observations are not
##                                  from the same Distribution.
## 
## Type I Error:                    5%
## 
## Number of Outliers Detected:     0
## 
##   i   Mean.i     SD.i  Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 6.736554 1.580951 10.850      23 2.601881   3.635271   FALSE
## 2 1 6.717856 1.559874 10.800      10 2.616971   3.633930   FALSE
## 3 2 6.699216 1.538694 10.775      33 2.648860   3.632582   FALSE
```

```r
#no outliers detected

## Labellum_Length
ggplot(flo_morph_by_ID) +
  aes(x = "", y = LABL) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```r
# print outlier values
sort(boxplot.stats(flo_morph_by_ID$LABL)$out, decreasing = TRUE)
```

```
## [1] 64.1
```

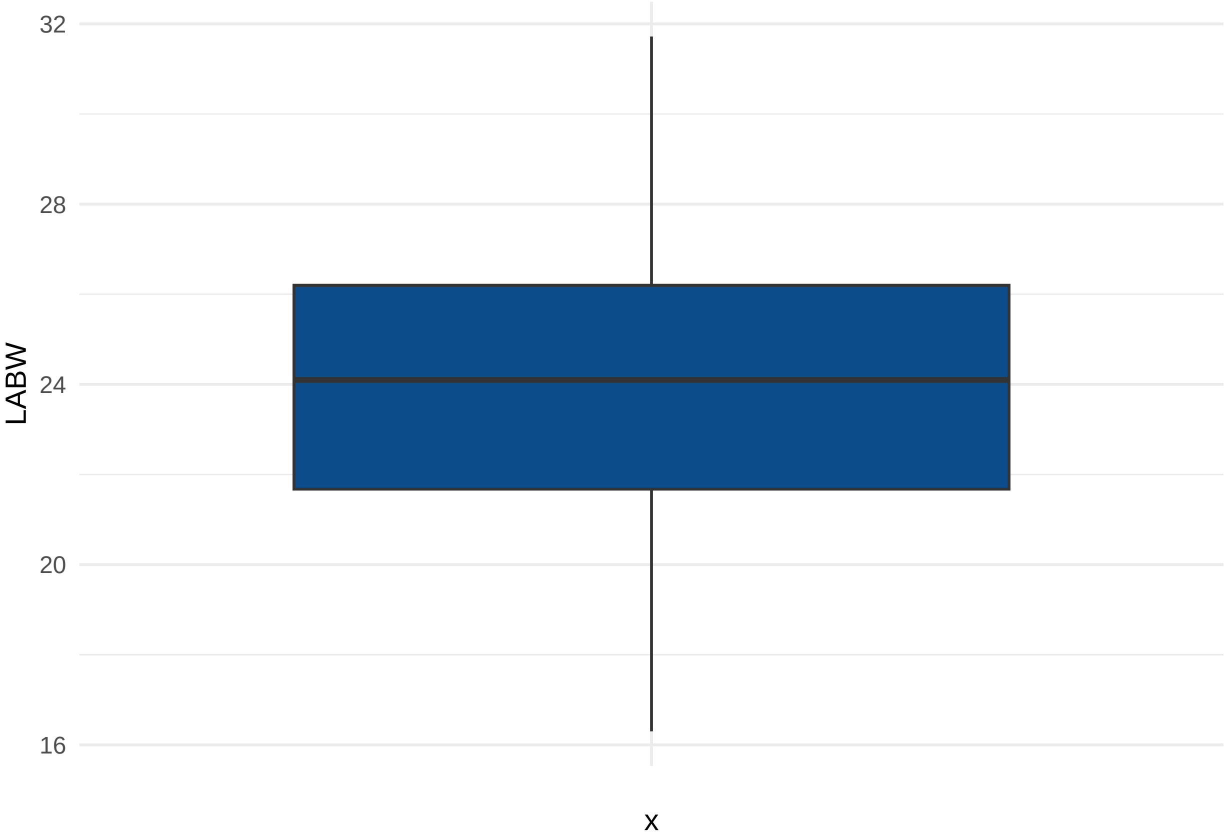
```r
#Rosner test
rosnerTest(flo_morph_by_ID$LABL, k=length(boxplot.stats(flo_morph_by_ID$LABL)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            flo_morph_by_ID$LABL
##
## Sample Size:                     221
##
## Test Statistic:                  R.1 = 3.095334
##
## Test Statistic Parameter:        k = 1
##
## Alternative Hypothesis:          Up to 1 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
```

```
##
## Number of Outliers Detected:     0
##
##   i   Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 52.88004 3.624798  64.1      23 3.095334   3.635271   FALSE
```

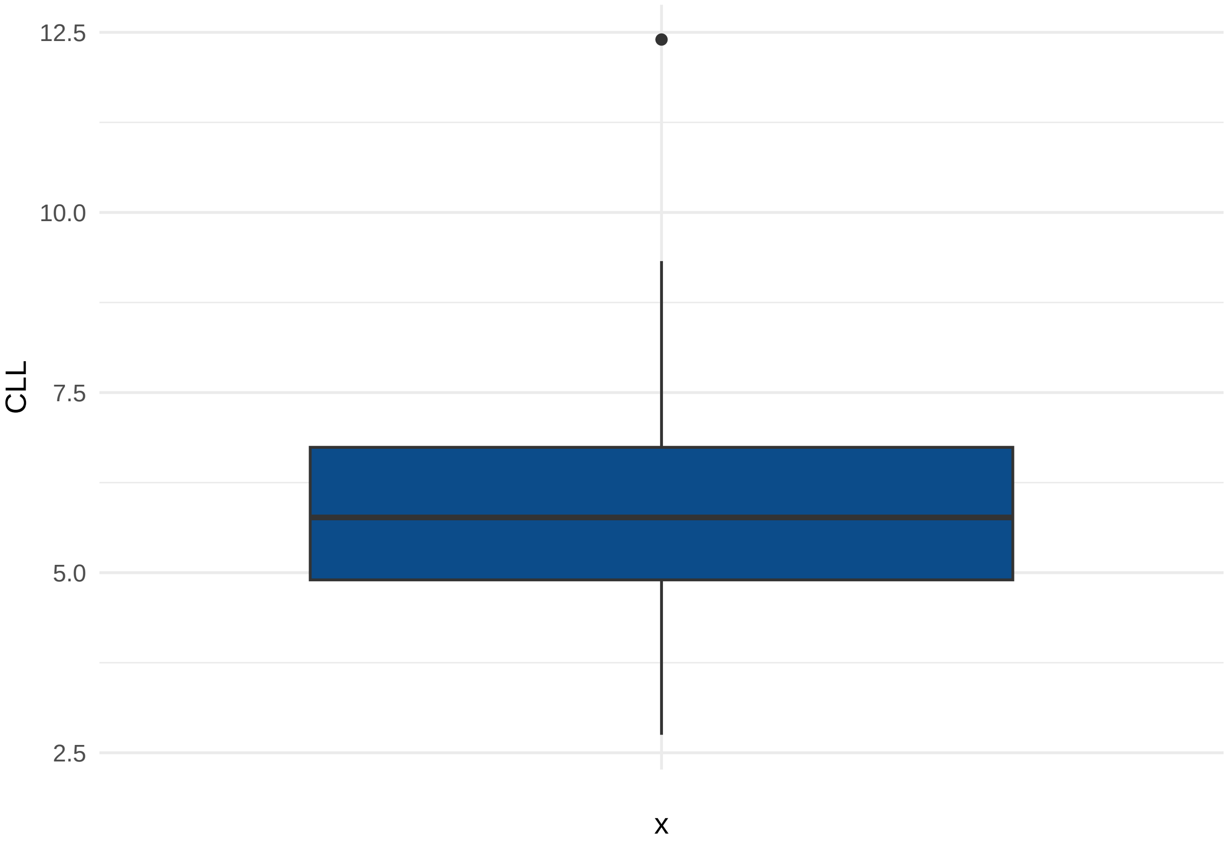
```
#no outliers detected

## Labellum_Width
ggplot(flo_morph_by_ID) +
  aes(x = "", y = LABW) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# no outliers detected

## Labellum_lobe
ggplot(flo_morph_by_ID) +
  aes(x = "", y = CLL) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values
sort(boxplot.stats(flo_morph_by_ID$CLL)$out, decreasing = TRUE)
```

```
## [1] 12.4
```

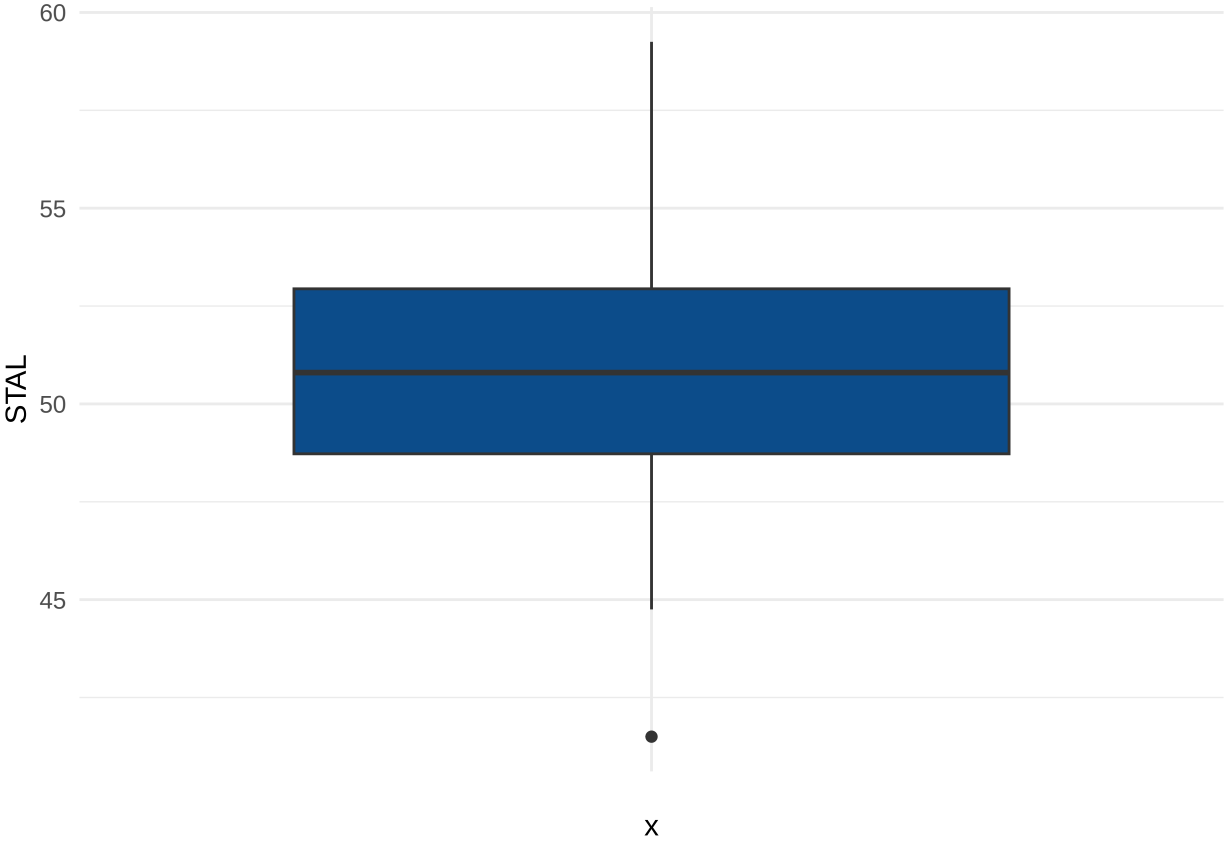
```
#Rosner test
rosnerTest(flo_morph_by_ID$CLL, k=length(boxplot.stats(flo_morph_by_ID$CLL)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            flo_morph_by_ID$CLL
##
## Sample Size:                     221
##
## Test Statistic:                  R.1 = 4.427964
##
## Test Statistic Parameter:        k = 1
##
## Alternative Hypothesis:          Up to 1 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
```

```
##
## Number of Outliers Detected:     1
##
##   i   Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 5.831342 1.483449  12.4     212 4.427964   3.635271    TRUE
```

```
#one outlier detected
## Remove outliers
flo_morph_by_ID <- flo_morph_by_ID %>%
  mutate(CLL = na_if(CLL, 12.4))

## Stamen_Length
ggplot(flo_morph_by_ID) +
  aes(x = "", y = STAL) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values
sort(boxplot.stats(flo_morph_by_ID$STAL)$out, decreasing = TRUE)
```

```
## [1] 41.5
```

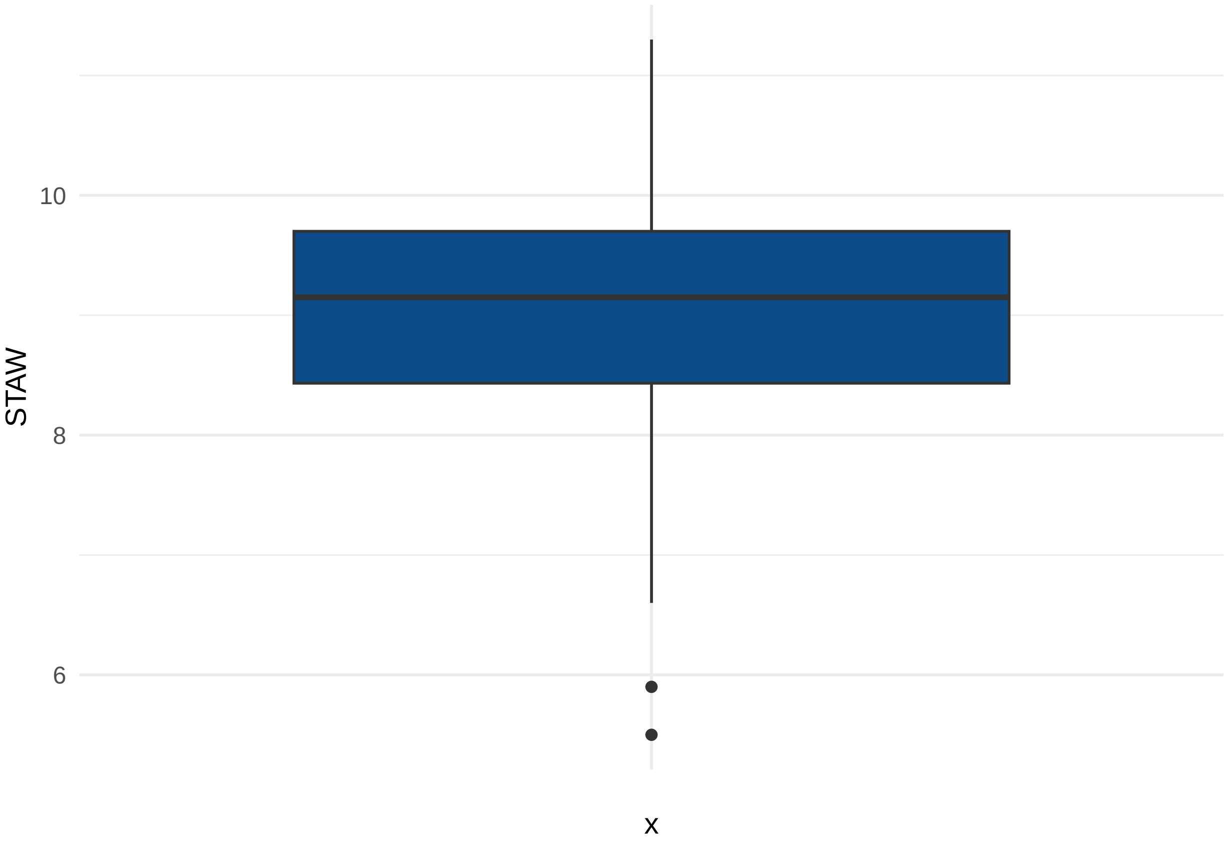
```
#Rosner test
rosnerTest(flo_morph_by_ID$STAL, k=length(boxplot.stats(flo_morph_by_ID$STAL)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
```

```
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            flo_morph_by_ID$STAL
##
## Sample Size:                     221
##
## Test Statistic:                  R.1 = 3.136281
##
## Test Statistic Parameter:        k = 1
##
## Alternative Hypothesis:          Up to 1 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
##
## Number of Outliers Detected:     0
##
##   i   Mean.i    SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 50.82884 2.97449  41.5     169 3.136281   3.635271   FALSE
```

```
#no outliers detected

## Stamen_width
ggplot(flo_morph_by_ID) +
  aes(x = "", y = STAW) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values
sort(boxplot.stats(flo_morph_by_ID$STAW)$out, decreasing = TRUE)
```

```
## [1] 5.9 5.5
```

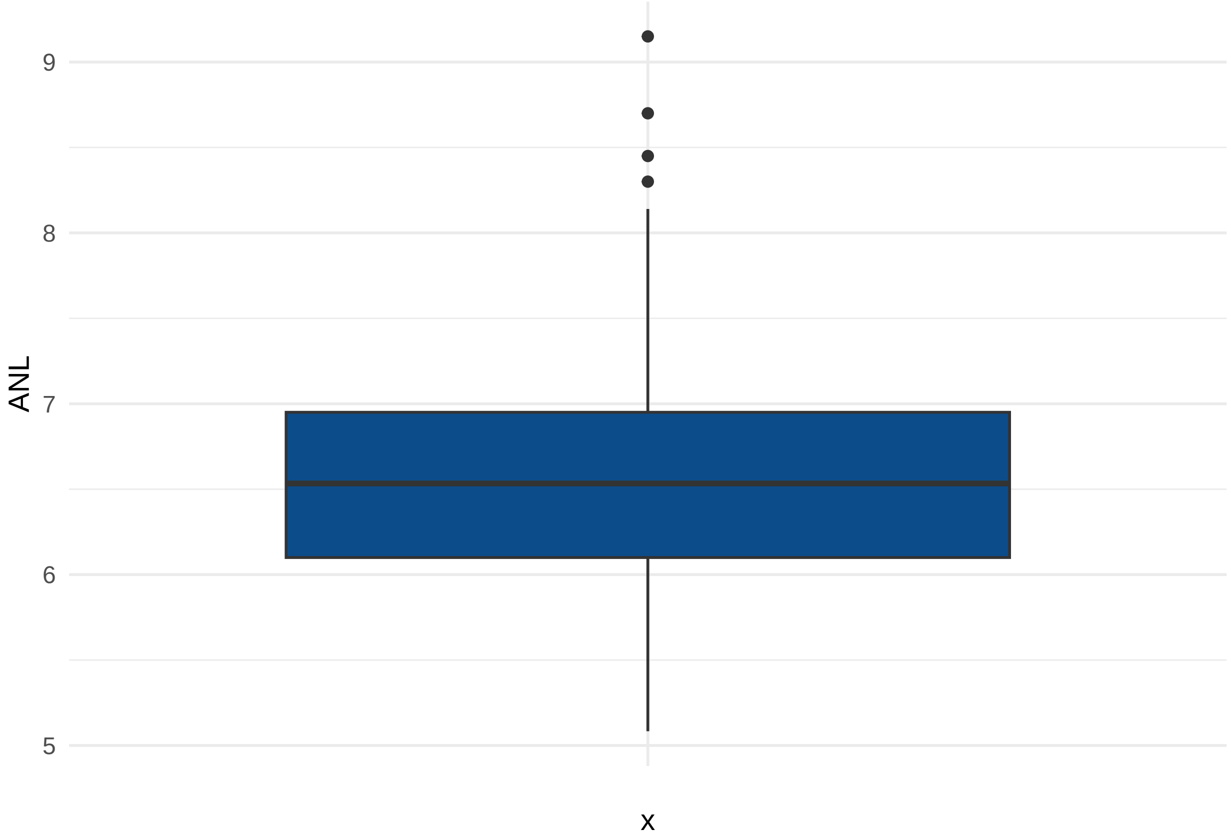
```
#Rosner test
rosnerTest(flo_morph_by_ID$STAW, k=length(boxplot.stats(flo_morph_by_ID$STAW)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            flo_morph_by_ID$STAW
##
## Sample Size:                     221
##
## Test Statistics:                 R.1 = 3.806009
##                                  R.2 = 3.504634
##
## Test Statistic Parameter:        k = 2
##
## Alternative Hypothesis:          Up to 2 observations are not
##                                  from the same Distribution.
##
```

```
## Type I Error:                      5%
##
## Number of Outliers Detected:      1
##
##   i   Mean.i      SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 9.050064 0.9327523   5.5      80 3.806009   3.635271    TRUE
## 2 1 9.066201 0.9034325   5.9     201 3.504634   3.633930   FALSE
```

```
#one outlier detected
## Remove outlier
flo_morph_by_ID <- flo_morph_by_ID %>%
  mutate(STAW = na_if(STAW, 5.5))

## Anther_Length
ggplot(flo_morph_by_ID) +
  aes(x = "", y = ANL) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values
sort(boxplot.stats(flo_morph_by_ID$ANL)$out, decreasing = TRUE)
```

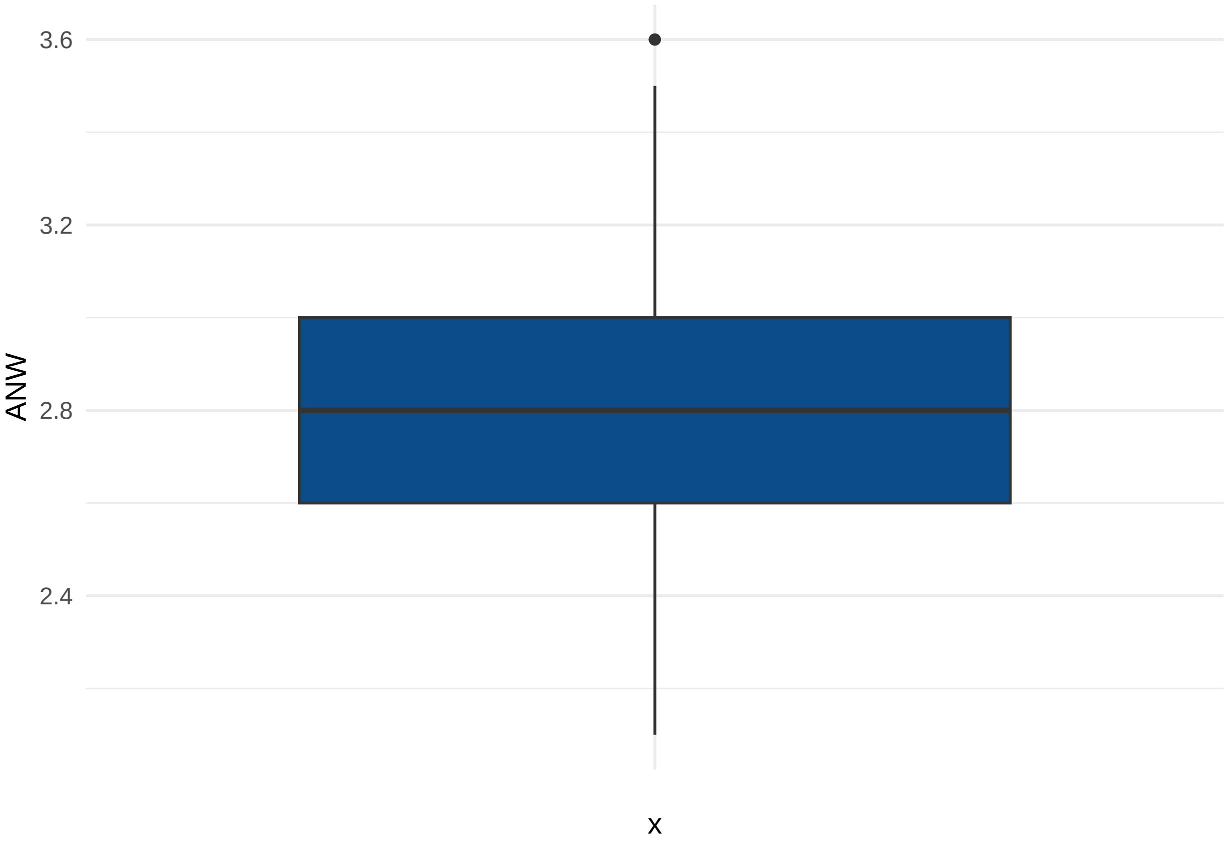
```
## [1] 9.15 8.70 8.45 8.30
```

```
#Rosner test
rosnerTest(flo_morph_by_ID$ANL, k=length(boxplot.stats(flo_morph_by_ID$ANL)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            flo_morph_by_ID$ANL
##
## Sample Size:                     221
##
## Test Statistics:                 R.1 = 3.963247
##                                  R.2 = 3.410173
##                                  R.3 = 3.106919
##                                  R.4 = 2.937045
##
## Test Statistic Parameter:        k = 4
##
## Alternative Hypothesis:          Up to 4 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
##
## Number of Outliers Detected:     1
##
##   i   Mean.i      SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 6.558371 0.6539155  9.15     113 3.963247   3.635271    TRUE
## 2 1 6.546591 0.6314662  8.70     221 3.410173   3.633930   FALSE
## 3 2 6.536758 0.6158004  8.45     142 3.106919   3.632582   FALSE
## 4 3 6.527982 0.6033337  8.30     197 2.937045   3.631227   FALSE
```

```r
#one outlier detected
## Remove outliers
flo_morph_by_ID <- flo_morph_by_ID %>%
  mutate(ANL = na_if(ANL, 9.15))

## Anther_width
ggplot(flo_morph_by_ID) +
  aes(x = "", y = ANW) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values
sort(boxplot.stats(flo_morph_by_ID$ANW)$out, decreasing = TRUE)
```

```
## [1] 3.6
```

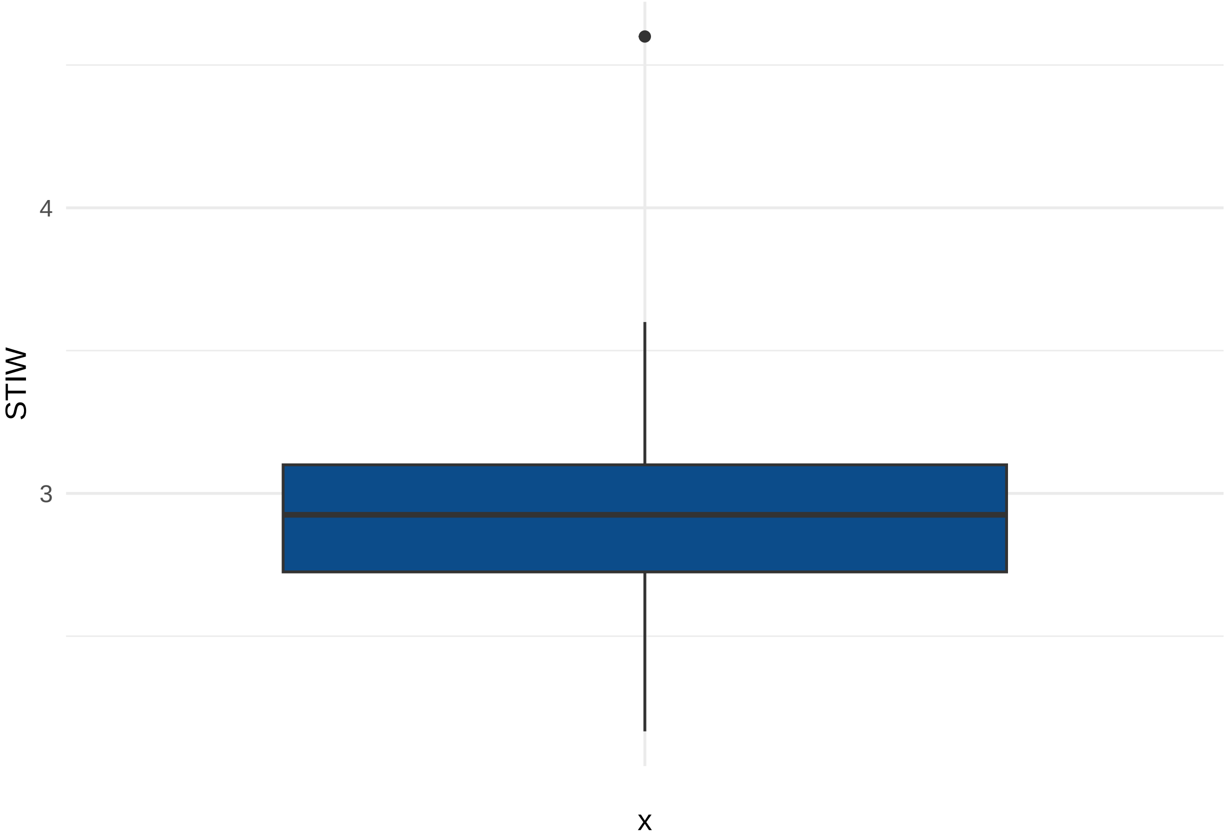
```
#Rosner test
rosnerTest(flo_morph_by_ID$ANW, k=length(boxplot.stats(flo_morph_by_ID$ANW)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            flo_morph_by_ID$ANW
##
## Sample Size:                     221
##
## Test Statistic:                  R.1 = 2.524466
##
## Test Statistic Parameter:        k = 1
##
## Alternative Hypothesis:          Up to 1 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
```

```
##
## Number of Outliers Detected:      0
##
##   i   Mean.i      SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 2.807504 0.3139263   3.6     102 2.524466   3.635271   FALSE
```

```
#no outliers detected

## Stigma_Width
ggplot(flo_morph_by_ID) +
  aes(x = "", y = STIW) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values
sort(boxplot.stats(flo_morph_by_ID$STIW)$out, decreasing = TRUE)
```

```
## [1] 4.6
```

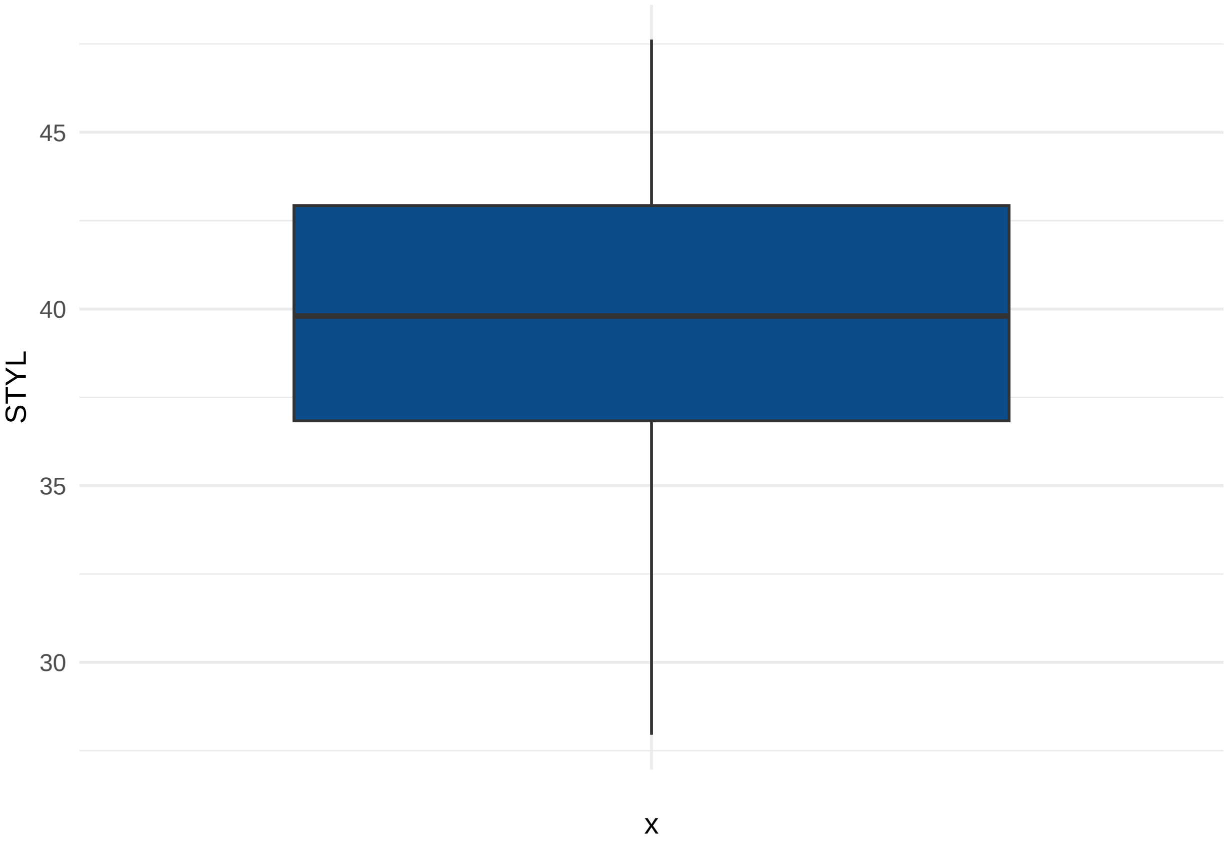
```
#Rosner test
rosnerTest(flo_morph_by_ID$STIW, k=length(boxplot.stats(flo_morph_by_ID$STIW)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
```

```
##
## Data:                            flo_morph_by_ID$STIW
##
## Sample Size:                     221
##
## Test Statistic:                  R.1 = 5.671614
##
## Test Statistic Parameter:        k = 1
##
## Alternative Hypothesis:          Up to 1 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
##
## Number of Outliers Detected:     1
##
##   i  Mean.i      SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 2.91727 0.2966933   4.6     115 5.671614   3.635271    TRUE
```

```
#one outlier detected
## Remove outliers
flo_morph_by_ID <- flo_morph_by_ID %>%
  mutate(STIW = na_if(STIW, 4.6))

## Style_length
ggplot(flo_morph_by_ID) +
  aes(x = "", y = STYL) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values
sort(boxplot.stats(flo_morph_by_ID$STYL)$out, decreasing = TRUE)
```

```
## numeric(0)
```

```
# no outliers detected
```

Nectar data

Load and filter data

```
nectar <- read_sheet(ss="19XHi3K57mDi2BpMPrlNOR2w16VSwVOYP6WYKYJHvuqs", sheet="nectar")
```

```
## v Reading from "F2_phenotypes".
```

```
## v Range ''nectar''.
```

```
#make a new column that makes a unique ID for each plant
nectar <- mutate(nectar, unique_ID = paste0(plant_type, "_", ID))

#exclude parents and F1s
nectar <- filter(nectar, plant_type != "F1", plant_type != "P")

#exclude columns we don't need
nectar <- nectar %>% dplyr::select(-year, -date, -plant_type, -ID, -rep, -fl_capsize)

# EFN_percent_sucrose is dependent on EFN_H2Ovolume_uL
# Split EFN_percent_sucrose into two different variables: 1) diluted in 40 uL or
```

```r
# 2) diluted in 50 uL (because these cannot be directly compared)

nectar <- nectar %>%
  mutate(
    EFN_percent_sucrose_H2Ovolume_40uL = if_else(
      EFN_H2Ovolume_uL == 40,        # Condition: Check if H2O volume is 40 µL
      EFN_percent_sucrose,           # If TRUE: Assign the value of EFN_percent_sucrose
      NA_real_                       # If FALSE: Assign NA (as a numeric NA)
    ),
    EFN_percent_sucrose_H2Ovolume_50uL = if_else(
      EFN_H2Ovolume_uL == 50,        # Condition: Check if H2O volume is 50 µL
      EFN_percent_sucrose,           # If TRUE: Assign the value of EFN_percent_sucrose
      NA_real_
    )
  )

#exclude columns we don't need
# I'm removing EFN_percent_sucrose_H2Ovolume_50uL because there are not enough observations to QTL map
nectar <- nectar %>% dplyr::select(-fl_nectarlength_mm, -fl_mg_sucrose, -EFN_H2Ovolume_uL, -EFN_percent

#rename columns
nectar <- nectar %>% dplyr::rename("VFN" = "fl_nectarvolume_uL", # volume floral nectar
                                   "FNSC" = "fl_percent_sucrose", # percent sucrose floral nectar
                                   "EFNSC40" = "EFN_percent_sucrose_H2Ovolume_40uL") # percent sucrose

#quick look at data
nectar
```

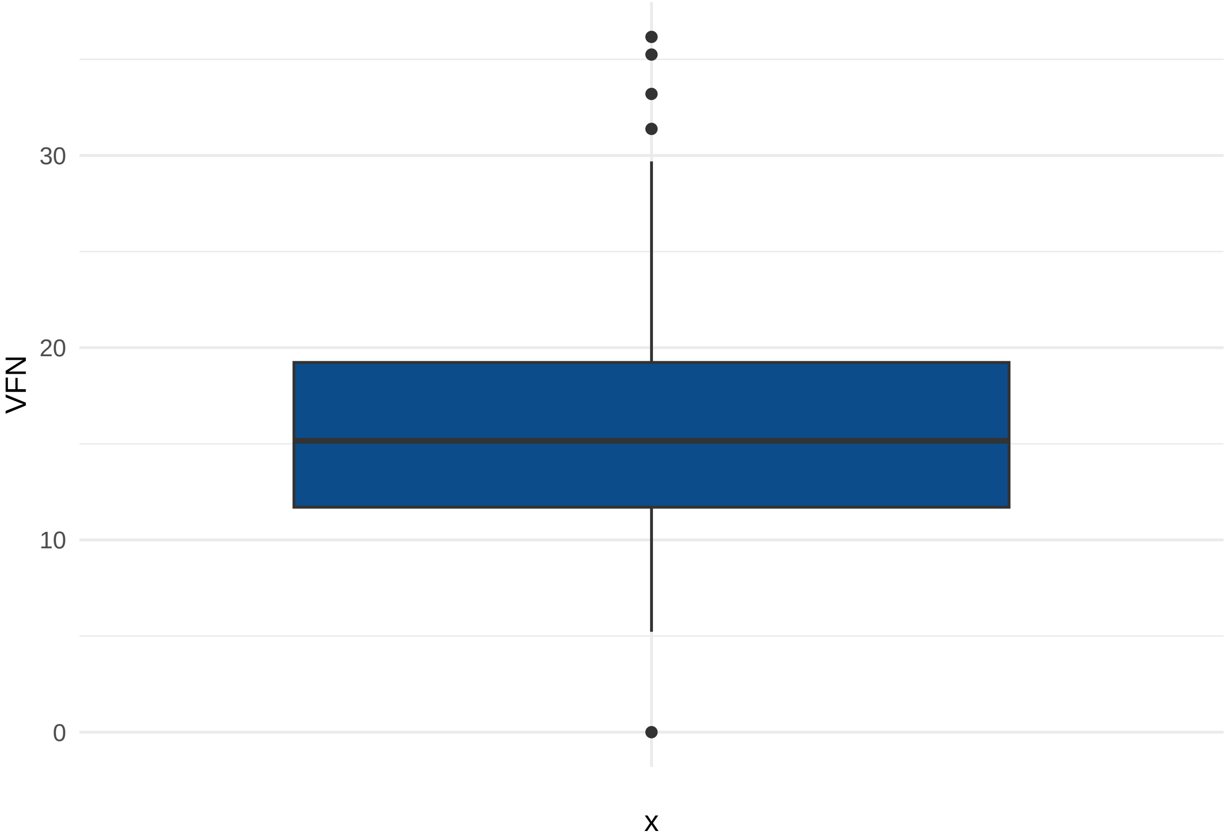
```
## # A tibble: 1,017 x 4
##      VFN  FNSC unique_ID EFNSC40
##    <dbl> <dbl> <chr>       <dbl>
##  1 13.8   32.5 39_2         13
##  2 14.9   27.5 39_2          5.5
##  3 20.8   28.5 39_2          5
##  4 21.2   24.5 39_2          3
##  5 18.1   32   39_2         18
##  6 13.3   33.5 39_3          8
##  7  5.9   31   39_3          9
##  8  1.78  16.5 39_3          0
##  9 16.2   36   39_3         15.5
## 10 11.1   36   39_3          6
## # i 1,007 more rows
```

Collapse the replicate observations

```r
nectar_by_ID <- nectar %>% group_by(unique_ID) %>% reframe(
  tibble(
    across(where(is.double), \(x) mean(x, na.rm = TRUE)),
    across(where(is.character), Modes),
    across(where(is.factor), Modes)
  )
)
```

Identify potential outliers

```
## fl_nectarlength_mm
ggplot(nectar_by_ID) +
  aes(x = "", y = VFN) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```
# print outlier values
sort(boxplot.stats(nectar_by_ID$VFN)$out, decreasing = TRUE)
```

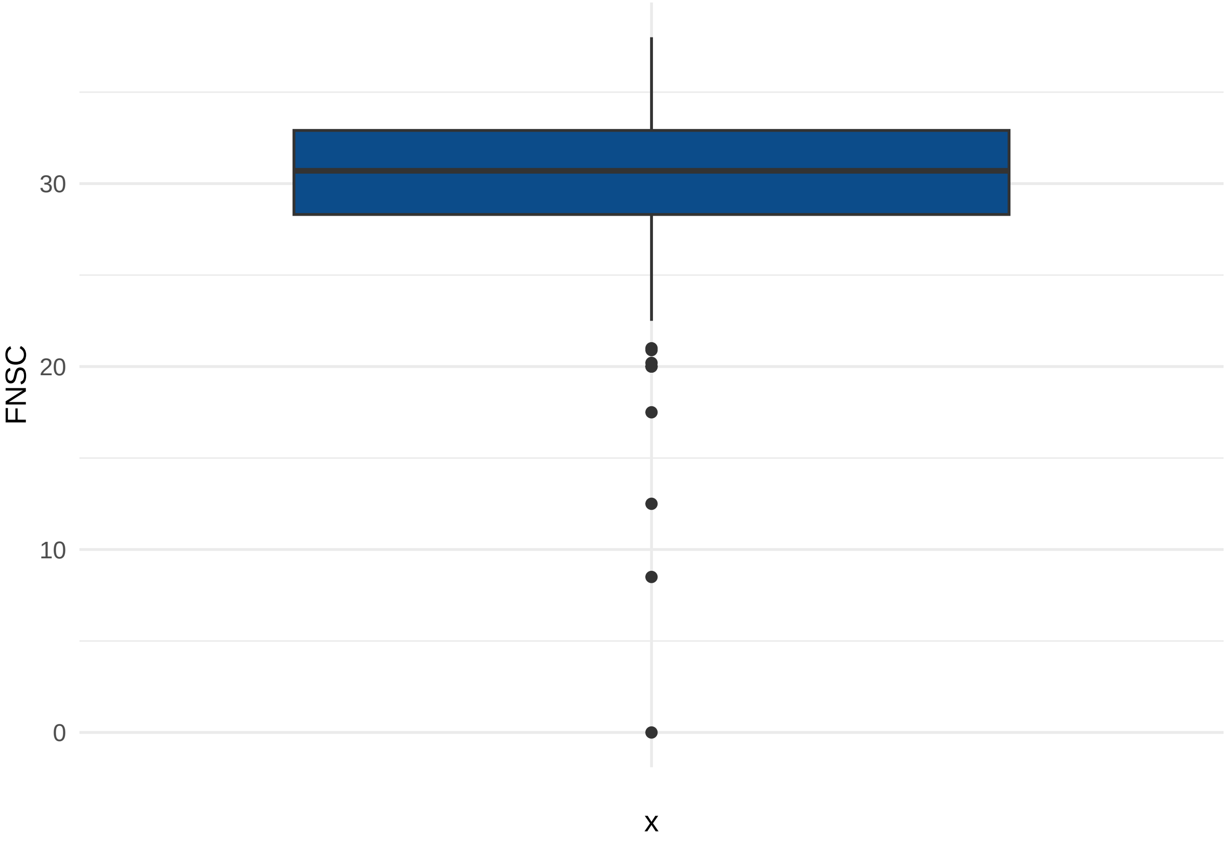
```
## [1] 36.16600 35.24900 33.20000 31.38125  0.00000
```

```
#Rosner test
rosnerTest(nectar_by_ID$VFN, k=length(boxplot.stats(nectar_by_ID$VFN)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            nectar_by_ID$VFN
##
## Sample Size:                     230
##
## Test Statistics:                 R.1 = 3.404418
##                                   R.2 = 3.343785
##                                   R.3 = 3.075063
##                                   R.4 = 2.826014
##                                   R.5 = 2.853453
##
## Test Statistic Parameter:         k = 5
##
## Alternative Hypothesis:           Up to 5 observations are not
##                                   from the same Distribution.
##
## Type I Error:                     5%
##
## Number of Outliers Detected:      0
##
##   i   Mean.i     SD.i    Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 15.95434 5.936891 36.16600     230 3.404418   3.647033   FALSE
## 2 1 15.86608 5.796699 35.24900     157 3.343785   3.645753   FALSE
## 3 2 15.78107 5.664577 33.20000      16 3.075063   3.644466   FALSE
## 4 3 15.70433 5.557062  0.00000     214 2.826014   3.643172   FALSE
## 5 4 15.77382 5.469664 31.38125     103 2.853453   3.641872   FALSE
```

```
#no outliers detected

## fl_percent_sucrose
ggplot(nectar_by_ID) +
  aes(x = "", y = FNSC) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```



```r
# print outlier values
sort(boxplot.stats(nectar_by_ID$FNSC)$out, decreasing = TRUE)
```

```
## [1] 21.0 20.9 20.2 20.0 20.0 17.5 12.5  8.5  0.0
```

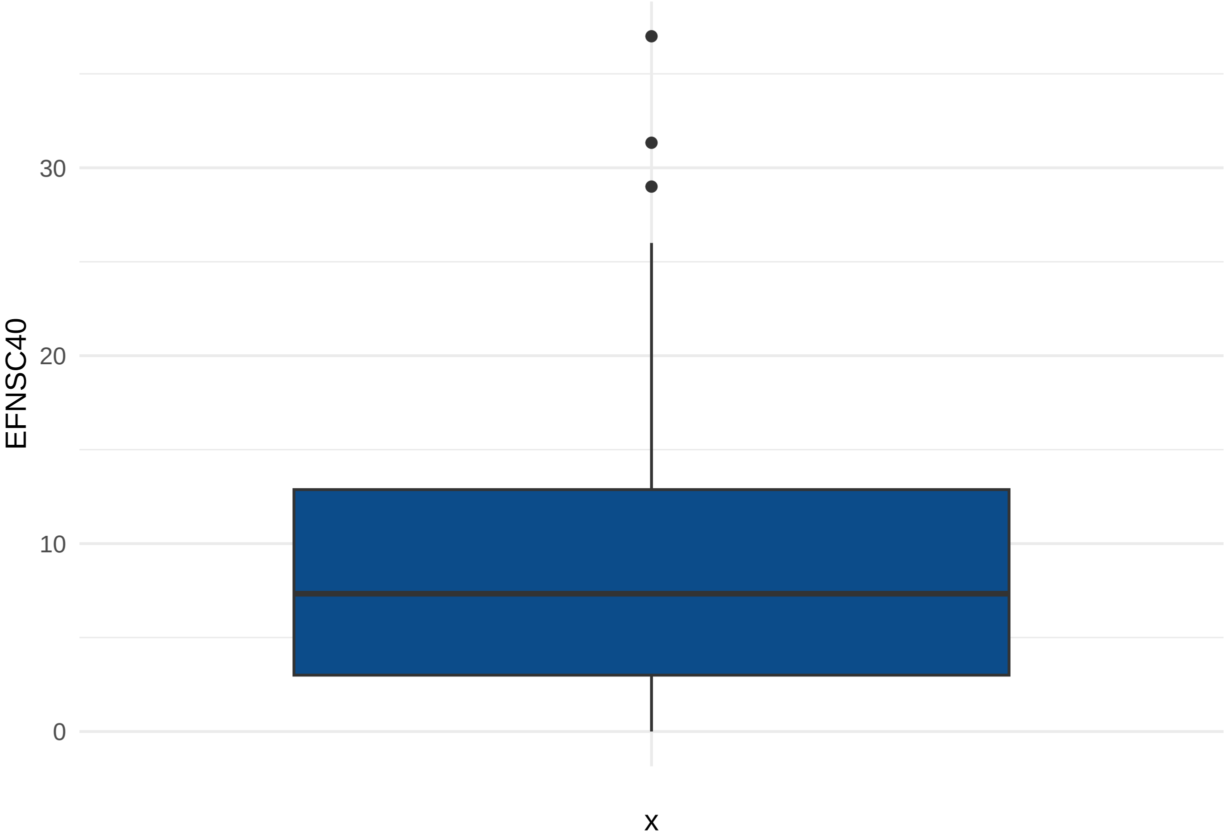
```r
#Rosner test
rosnerTest(nectar_by_ID$FNSC, k=length(boxplot.stats(nectar_by_ID$FNSC)$out))
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            nectar_by_ID$FNSC
##
## Sample Size:                     230
##
## Test Statistics:                 R.1 = 6.639724
##                                  R.2 = 5.322246
##                                  R.3 = 4.654127
##                                  R.4 = 3.531589
##                                  R.5 = 2.936567
##                                  R.6 = 3.001395
##                                  R.7 = 3.012250
##                                  R.8 = 2.874135
##                                  R.9 = 2.905676
##
## Test Statistic Parameter:        k = 9
##
## Alternative Hypothesis:          Up to 9 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
##
## Number of Outliers Detected:     3
##
##   i   Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 30.04947 4.525710   0.0     214 6.639724   3.647033    TRUE
## 2 1 30.18069 4.073597   8.5      22 5.322246   3.645753    TRUE
## 3 2 30.27578 3.819359  12.5     178 4.654127   3.644466    TRUE
## 4 3 30.35409 3.639746  17.5     218 3.531589   3.643172   FALSE
## 5 4 30.41096 3.545284  20.0      61 2.936567   3.641872   FALSE
## 6 5 30.45723 3.484124  20.0     147 3.001395   3.640566   FALSE
## 7 6 30.50392 3.420672  20.2      42 3.012250   3.639252   FALSE
## 8 7 30.55012 3.357575  20.9     155 2.874135   3.637932   FALSE
## 9 8 30.59359 3.301673  21.0      96 2.905676   3.636605   FALSE
```

```r
#three outliers detected
## Remove outliers
nectar_by_ID <- nectar_by_ID %>%
  mutate(FNSC = na_if(FNSC, 0.0)) %>%
  mutate(FNSC = na_if(FNSC, 8.5)) %>%
  mutate(FNSC = na_if(FNSC, 12.5))

## Extrafloral nectar percent sucrose (diluted in 40 uL)
ggplot(nectar_by_ID) +
  aes(x = "", y = EFNSC40) +
  geom_boxplot(fill = "#0c4c8a") +
  theme_minimal()
```

```
## Warning: Removed 64 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



```
# print outlier values
sort(boxplot.stats(nectar_by_ID$EFNSC40)$out, decreasing = TRUE)
```

```
## [1] 37.00000 31.33333 29.00000
```

```
#Rosner test
rosnerTest(nectar_by_ID$EFNSC40, k=length(boxplot.stats(nectar_by_ID$EFNSC40)$out))
```

```
## Warning in rosnerTest(nectar_by_ID$EFNSC40, k =
## length(boxplot.stats(nectar_by_ID$EFNSC40)$out)): 64 observations with
## NA/NaN/Inf in 'x' removed.
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
```

```
## Hypothesized Distribution:       Normal
##
## Data:                            nectar_by_ID$EFNSC40
##
## Number NA/NaN/Inf's Removed:     64
##
## Sample Size:                     166
##
## Test Statistics:                 R.1 = 3.993270
##                                  R.2 = 3.376433
##                                  R.3 = 3.154536
##
## Test Statistic Parameter:        k = 3
##
## Alternative Hypothesis:          Up to 3 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
##
## Number of Outliers Detected:     1
##
##   i   Mean.i     SD.i    Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 8.700638 7.086765 37.00000     110 3.993270   3.548694    TRUE
## 2 1 8.529126 6.753934 31.33333      77 3.376433   3.546821   FALSE
## 3 2 8.390076 6.533424 29.00000       2 3.154536   3.544935   FALSE
```

```
#two outliers detected
## Remove outliers
nectar_by_ID <- nectar_by_ID %>%
  mutate(EFNSC40 = round(EFNSC40, digits=2)) %>%
  mutate(EFNSC40 = na_if(EFNSC40, 37.00000))
## make sure outliers are removed
rosnerTest(nectar_by_ID$EFNSC40, k=2)
```

```
## Warning in rosnerTest(nectar_by_ID$EFNSC40, k = 2): 65 observations with
## NA/NaN/Inf in 'x' removed.
```

```
##
## Results of Outlier Test
## -------------------------
##
## Test Method:                     Rosner's Test for Outliers
##
## Hypothesized Distribution:       Normal
##
## Data:                            nectar_by_ID$EFNSC40
##
## Number NA/NaN/Inf's Removed:     65
##
## Sample Size:                     165
##
## Test Statistics:                 R.1 = 3.375910
##                                  R.2 = 3.154464
##
## Test Statistic Parameter:        k = 2
```

```
##
## Alternative Hypothesis:          Up to 2 observations are not
##                                  from the same Distribution.
##
## Type I Error:                    5%
##
## Number of Outliers Detected:     0
##
##   i   Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1 0 8.529394 6.753915 31.33      77 3.375910   3.546821   FALSE
## 2 1 8.390366 6.533483 29.00       2 3.154464   3.544935   FALSE
```

Combine all data

```
data <- inflor_by_ID %>%
  full_join(redarea_by_ID, by="unique_ID") %>%
  full_join(flo_morph_by_ID, by="unique_ID") %>%
  full_join(nectar_by_ID, by="unique_ID")


write.csv(data, "~/Dropbox/Costus/costus-genetic-mapping/phenotype/results/processed_data/phenotypic_da
```