

Data analysis of Arabidopsis experiment B1

Nine Luijendijk

2023-05-26

Data can be found on GitHub

Loading the packages

```
library(readxl)
library(tidyverse)
library(gdata)
library(car)
library(cowplot)
library(magick)
library(here)
```

The first step is to load in the data from the Excel file and combine it with the metadata.

Loading file and generating complete dataframe

```
data_raw <- read_excel("data/arabidopsis_data.xlsx", col_types = c("text", rep("numeric", 8), "text"))

metadata <- tibble("Group" = c("L1", "L2", "S3", "S4", "L5", "L6", "S7", "S8"),
  "Population" = rep(c("Bovra", "Smadalen", "Helin", "Spiterstulen"), each = 2),
  "Size" = rep(c("Large", "Small", "Large", "Small"), each = 2),
  "Elevation" = rep("High", each = 8),
  "Condition" = rep(c("Stable", "Declining", "Declining", "Declining"), each = 2), #difference between
  "Soil" = rep(c("Riverbed", "Riverbed", "Scree", "Rocks"), each = 2),
  "Treatment" = c("Unsalted", "Salted", "Unsalted", "Salted", "Unsalted", "Salted", "Unsalted", "Salted"),
  "Conc_NaCl" = c(0, 50, 0, 50, 0, 50, 0, 50),
  "Conc_unit" = rep("mM", each = 8),
  "Length_unit" = rep("cm", each = 8),
  "Weight_unit" = rep("grams"), each = 8)

data <- mutate(data_raw,
  "Length_dif" = data_raw$Length_longest_leaf_a - data_raw$Length_longest_leaf_b,
  "Number_dif" = data_raw$Number_leaves_a - data_raw$Number_leaves_b) #calculating difference

data_complete <- left_join(data, metadata, by = "Group") #create one complete dataframe
```

Now to begin the data analysis, the Shapiro-Wilk values of each group are calculated too see whether the values are normally distributed.

```
names <- c("Wet_weight", "Length_dif", "Number_dif") #collect the names of the columns of interest

sw_results <- vector("list", 3) #prepare an empty list

for (i in 1:3) { #create a list of dataframes where every dataframe contains its Shapiro-Wilk p-values
  result <- data_complete %>% group_by(Size, Treatment) %>% summarize(sw = shapiro.test(!!sym(names[i]))$
```

```

result <- unite(result, "Treatment", Size, Treatment)
sw_results[[i]] <- result
}

merged <- sw_results %>% reduce(full_join, by= "Treatment") #merge all the dataframes from the list

tidy_sw_results <- pivot_longer(data = merged, cols = names,
                               names_to = "Info", values_to = "ShapiroWilk_p.value") #make the dataframe

tidy_sw_results %>% filter(ShapiroWilk_p.value > 0.05) #filter for which groups are normally distributed
tidy_sw_results %>% filter(ShapiroWilk_p.value < 0.05) #filter for which groups are NOT normally distributed

```

All groups of interest are normally distributed, except for the difference in number of leaves

To check whether there is a difference between the wet weights and growth (difference in length of the longest leaf) between the salted and unsalted groups, a t-test is used after checking for equal variance:

```

leveneTest(data_complete$Wet_weight ~ as.factor(Treatment), data = data_complete) #Pr(>F) = 0.4168, which means there is no significant difference

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 1    0.667 0.4168
##      71

leveneTest(data_complete$Length_dif ~ as.factor(Treatment), data = data_complete) #Pr(>F) = 0.4793, which means there is no significant difference

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 1    0.5058 0.4793
##      72

t.test(formula = data_complete$Length_dif ~ data_complete$Treatment,
       paired = FALSE, var.equal = TRUE)$p.value %>% round(.,3) #t-test, p = 0.005, which means there is a significant difference

## [1] 0.005

t.test(formula = data_complete$Wet_weight ~ data_complete$Treatment,
       paired = FALSE, var.equal = TRUE)$p.value %>% round(.,3) #t-test, p = 0.011, which means there is a significant difference

## [1] 0.011

wilcox.test(Number_dif ~ Treatment, data = data_complete) #Wilcoxon test, p = 0.5845, which means there is no significant difference

## Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
## compute exact p-value with ties

##
## Wilcoxon rank sum test with continuity correction
##
## data: Number_dif by Treatment
## W = 715, p-value = 0.5845
## alternative hypothesis: true location shift is not equal to 0

```

These differences are interesting to visualize as they are statistically significant:

```

##
## Wilcoxon rank sum test with continuity correction
##
## data: Number_dif by Treatment
## W = 715, p-value = 0.5845
## alternative hypothesis: true location shift is not equal to 0

```

These differences are interesting to visualize as they are statistically significant:

Generating the first plots

```

summary <- data_complete %>%
  group_by(Treatment) %>%
  summarize(mean_noleaves.increase = mean(Number_dif, na.rm = TRUE),

```

```

    mean_lengthleaf.increase = mean(Length_dif, na.rm = TRUE),
    mean_wetweight = mean(Wet_weight, na.rm = TRUE),
    stdevLength = sd(Length_dif, na.rm = TRUE),
    stdevWetweight = sd(Wet_weight, na.rm = TRUE),
    stdevNumber = sd(Number_dif, na.rm = TRUE))

plot_lengthdiffull <- summary %>% #plot the increase in longest leaf length grouped by population size
  ggplot(aes(x = Treatment, y = mean_lengthleaf.increase))+
  geom_col(aes(fill = Treatment))+
  geom_errorbar(aes(ymin = mean_lengthleaf.increase - stdevLength,
                    ymax = mean_lengthleaf.increase + stdevLength), width=.2)+
  theme_light()+
  labs(x="Treatment",
       y="Mean increase in the length\nof the longest leaf in cm")+
  theme(legend.position = "none", text = element_text(size=12), plot.title = element_text(size=14),
        plot.subtitle = element_text(size=12))

plot_wetweightfull <- summary %>% #plot the wet weight grouped by population size
  ggplot(aes(x = Treatment, y = mean_wetweight))+
  geom_col(aes(fill = Treatment))+
  geom_errorbar(aes(ymin = mean_wetweight - stdevWetweight,
                    ymax = mean_wetweight + stdevWetweight), width=.2)+
  theme_light()+
  labs(x="Treatment",
       y="Wet weight in grams")+
  theme(legend.position = "none", text = element_text(size=12), plot.title = element_text(size=14),
        plot.subtitle = element_text(size=12))

plot_numberdiffull <- summary %>% #plot the increase in number of leaves grouped by population size
  ggplot(aes(x = Treatment, y = mean_noleaves.increase))+
  geom_col(aes(fill = Treatment))+
  geom_errorbar(aes(ymin = mean_noleaves.increase - stdevNumber,
                    ymax = mean_noleaves.increase + stdevNumber), width=.2)+
  theme_light()+
  labs(x="Treatment",
       y="Mean increase in the number of leaves")+
  theme(legend.position = "none", text = element_text(size=12), plot.title = element_text(size=14),
        plot.subtitle = element_text(size=12))

plotgrid0 <- plot_grid(plot_lengthdiffull, plot_wetweightfull, plot_numberdiffull, #combine the 3 plots
  labels = c("A", "B", "C"),
  ncol = 3, nrow = 1)

```

Zooming in, the large population-salted and small population-salted groups will be compared using a t-test for the wet weight and length-increase and a Wilcoxon test for the increase in number of leaves (as this group's values were not normally distributed):

```

data_salted <- subset(data_complete, Treatment == "Salted")
leveneTest(data_salted$Wet_weight ~ as.factor(Size), data = data_salted) #Pr(>F) = 0.3754 which means e

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 1  0.8061 0.3754
##      35

```

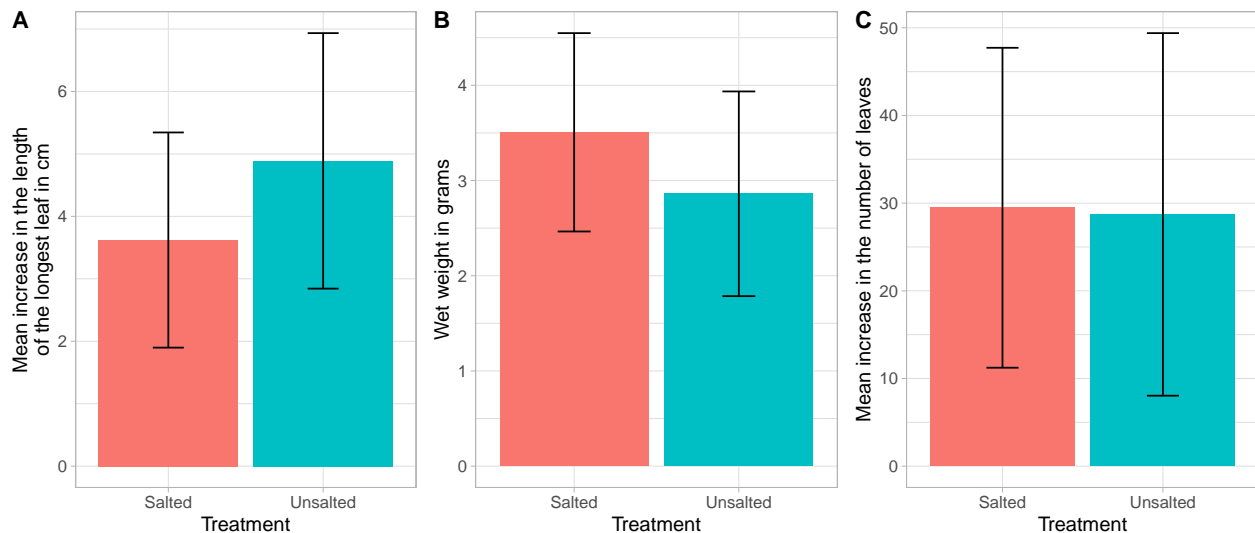


Figure 1: Figure 1: Bar graphs showing A) the increase in longest leaf length grouped by treatment, B) the increase in number of leaves grouped by treatment, and C) the increase in number of leaves grouped by treatment, with error bars depicting 1 standard deviation.

```
t.test(formula = data_salted$Wet_weight ~ data_salted$Size,
       paired = FALSE, var.equal = TRUE)$p.value %>% round(.,3) #t-test, p = 0.848, which means there is no significant difference

## [1] 0.848

leveneTest(data_salted$Length_dif ~ as.factor(Size), data = data_salted) #Pr(>F) = 0.7135 which means there is no significant difference

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 1  0.1369 0.7135
##      36

t.test(formula = data_salted$Length_dif ~ data_salted$Size,
       paired = FALSE, var.equal = TRUE)$p.value %>% round(.,3) #t-test, p = 0.563, which means there is no significant difference

## [1] 0.563

wilcox.test(Number_dif ~ Size, data = data_salted) #Wilcoxon test, p = 0.4735, which means there is no significant difference

##
## Wilcoxon rank sum test with continuity correction
##
## data: Number_dif by Size
## W = 205, p-value = 0.4735
## alternative hypothesis: true location shift is not equal to 0
```

The data can be plotted, where it's visualized how little difference there is:

Generating the figures

```
summary_salted <- data_salted %>%
  group_by(Size) %>%
  summarize(mean_noleaves.increase = mean(Number_dif, na.rm = TRUE),
            mean_lengthleaves.increase = mean(Length_dif, na.rm = TRUE),
            mean_wetweight = mean(Wet_weight, na.rm = TRUE),
```

```

    stdevlengthleave.increase = sd(Length_dif, na.rm = TRUE),
    stdevwetweight = sd(Wet_weight, na.rm = TRUE),
    stdevNumber = sd(Number_dif, na.rm = TRUE))

plot_length <- summary_salted %>% #plot the increase in longest leaf length grouped by population size
  ggplot(aes(x = Size, y = mean_lengthleaves.increase))+
  geom_col(aes(fill = Size))+
  geom_errorbar(aes(ymin = mean_lengthleaves.increase - stdevlengthleave.increase,
                    ymax = mean_lengthleaves.increase + stdevlengthleave.increase), width=.2)+
  theme_light()+
  labs(x="Population size",
       y="Mean increase in the length\nof the longest leaf in cm")+
  theme(legend.position = "none", text = element_text(size=12), plot.title = element_text(size=14),
        plot.subtitle = element_text(size=12))+
  scale_fill_manual(values=c("#e457b5", "#57e486"))

plot_noleaves <- summary_salted %>% #plot the increase in number of leaves grouped by population
  ggplot(aes(x = Size, y = mean_noleaves.increase))+
  geom_col(aes(fill = Size))+
  geom_errorbar(aes(ymin = mean_noleaves.increase - stdevNumber,
                    ymax = mean_noleaves.increase + stdevNumber), width=.2)+
  theme_light()+
  labs(x="Population size",
       y="Mean increase in the number of leaves")+
  theme(legend.position = "none", text = element_text(size=12), plot.title = element_text(size=15),
        plot.subtitle = element_text(size=12))+
  scale_fill_manual(values=c("#e457b5", "#57e486"))

plot_wetweight <- summary_salted %>% #plot the increase in number of leaves grouped by population
  ggplot(aes(x = Size, y = mean_wetweight))+
  geom_col(aes(fill = Size))+
  geom_errorbar(aes(ymin = mean_wetweight - stdevwetweight,
                    ymax = mean_wetweight + stdevwetweight), width=.2)+
  theme_light()+
  labs(x="Population size",
       y="Mean wet weight in grams")+
  theme(legend.position = "none", text = element_text(size=12), plot.title = element_text(size=15),
        plot.subtitle = element_text(size=12))+
  scale_fill_manual(values=c("#e457b5", "#57e486"))

plotgrid1 <- plot_grid(plot_length, plot_wetweight, plot_noleaves, #combine the 3 plots into 1 figure
  labels = c("A", "B", "C"),
  ncol = 3, nrow = 1)

```

Population size is only used as a proxy for inbreeding, but the genotypes of these plants are still unknown. This is why the populations need to be looked at individually as well:

```

data_salted %>%
  group_by(Population) %>%
  summarise(ShapiroWilk_p.value = shapiro.test(Length_dif)$p.value) #Shapiro-Wilk test, all p-values are

## # A tibble: 4 x 2
##   Population      ShapiroWilk_p.value
##   <chr>          <dbl>

```

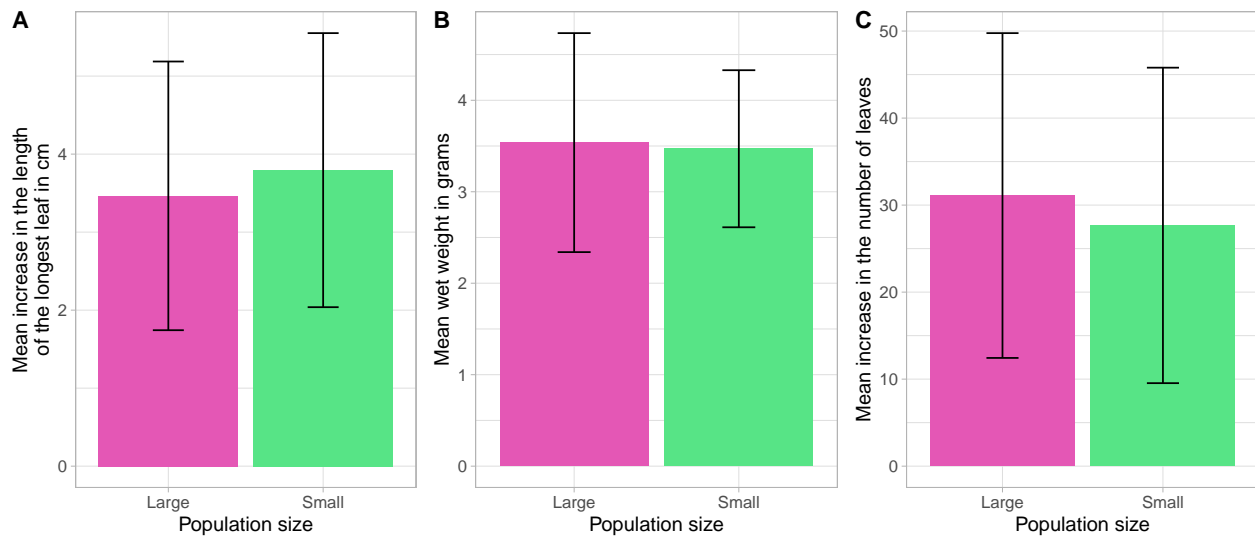


Figure 2: Figure 2: Bar graphs showing A) the increase in longest leaf length grouped by population size, B) the increase in number of leaves grouped by population size, and C) the increase in number of leaves grouped by population size, with error bars depicting 1 standard deviation, and the data filtered to only show the salted groups.

```
## 1 Bovra 0.489
## 2 Helin 0.112
## 3 Smadalen 0.273
## 4 Spiterstulen 0.725
```

```
data_salted %>%
  group_by(Population) %>%
  summarise(ShapiroWilk_p.value = shapiro.test(Wet_weight)$p.value) #Shapiro-Wilk test, all p-values are
```

```
## # A tibble: 4 x 2
##   Population ShapiroWilk_p.value
##   <chr> <dbl>
## 1 Bovra 0.0624
## 2 Helin 0.506
## 3 Smadalen 0.314
## 4 Spiterstulen 0.284
```

```
data_salted %>%
  group_by(Population) %>%
  summarise(ShapiroWilk_p.value = shapiro.test(Number_dif)$p.value) #Shapiro-Wilk test, one of the p-values
```

```
## # A tibble: 4 x 2
##   Population ShapiroWilk_p.value
##   <chr> <dbl>
## 1 Bovra 0.950
## 2 Helin 0.843
## 3 Smadalen 0.206
## 4 Spiterstulen 0.0215
```

```
leveneTest(data_salted$Length_dif ~ as.factor(Population), data = data_salted) #Pr(>F) = 0.8361, which is
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
```

```
## group 3 0.2847 0.8361
##      34
leveneTest(data_salted$Wet_weight ~ as.factor(Population), data = data_salted) #Pr(>F) = 0.7846, which means there is no significant difference in variance

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value Pr(>F)
## group 3 0.3567 0.7846
##      33
aov(Length_dif ~ Population, data_salted) %>% summary.aov() #Pr(>F) = 0.478, which means there is no significant difference in variance

##      Df Sum Sq Mean Sq F value Pr(>F)
## Population 3 7.64 2.546 0.847 0.478
## Residuals 34 102.19 3.006
aov(Wet_weight ~ Population, data_salted) %>% summary.aov() #Pr(>F) = 0.961, which means there is no significant difference in variance

##      Df Sum Sq Mean Sq F value Pr(>F)
## Population 3 0.34 0.1147 0.098 0.961
## Residuals 33 38.71 1.1732
## 1 observation deleted due to missingness
kruskal.test(Number_dif ~ Population, data = data_salted) #Kruskal-Wallis test, p = 0.03029, which means there is a significant difference in the distribution of the number of leaves

##
## Kruskal-Wallis rank sum test
##
## data: Number_dif by Population
## Kruskal-Wallis chi-squared = 8.9263, df = 3, p-value = 0.03029
summary_complete_pop <- data_salted %>%
  group_by(Population) %>%
  summarize(mean_noleaves.increase = mean(Number_dif, na.rm = TRUE),
            mean_lengthleaves.increase = mean(Length_dif, na.rm = TRUE),
            mean_wetweight = mean(Wet_weight, na.rm = TRUE),
            stdevlengthleave.increase = sd(Length_dif, na.rm = TRUE),
            stdevwetweight = sd(Wet_weight, na.rm = TRUE),
            stdevNumber = sd(Number_dif, na.rm = TRUE))
```

And again, these can be plotted:

Plotting more figures

```
plot_length2 <- summary_complete_pop %>% #plot the increase in longest leaf length grouped by population
  ggplot(aes(x = Population, y = mean_lengthleaves.increase))+
  geom_col(aes(fill = Population), color = c("#e457b5", "#e457b5", "#57e486", "#57e486"), linewidth = 2)+
  geom_errorbar(aes(ymin = mean_lengthleaves.increase - stdevlengthleave.increase,
                    ymax = mean_lengthleaves.increase + stdevlengthleave.increase), width=.2)+
  theme_light()+
  labs(x="Population",
       y="Mean increase in the length\nof the longest leaf in cm")+
  theme(legend.position = "none", text = element_text(size=12), plot.title = element_text(size=14),
        plot.subtitle = element_text(size=12))+
  scale_fill_manual(values=c("#8111ee", "#7eee11", "#f0860f", "#0f79f0"))

plot_noleaves2 <- summary_complete_pop %>% #plot the increase in number of leaves grouped by population
  ggplot(aes(x = Population, y = mean_noleaves.increase))+
  geom_col(aes(fill = Population), color = c("#e457b5", "#e457b5", "#57e486", "#57e486"), linewidth = 2)+
```

```

geom_errorbar(aes(ymin = mean_noleaves.increase - stdevNumber,
                  ymax = mean_noleaves.increase + stdevNumber), width=.2)+
theme_light()+
labs(x="Population",
     y="Mean increase in the number of leaves")+
theme(legend.position = "none", text = element_text(size=12), plot.title = element_text(size=15),
      plot.subtitle = element_text(size=12))+
scale_fill_manual(values=c("#8111ee", "#7eee11", "#f0860f", "#0f79f0"))

plot_wetweight2 <- summary_complete_pop %>% #plot the increase in number of leaves grouped by population
ggplot(aes(x = Population, y = mean_wetweight))+
geom_col(aes(fill = Population), color = c("#e457b5", "#e457b5", "#57e486", "#57e486"), linewidth = 2)+
geom_errorbar(aes(ymin = mean_wetweight - stdevwetweight,
                  ymax = mean_wetweight + stdevwetweight), width=.2)+
theme_light()+
labs(x="Population",
     y="Mean wet weight in grams")+
theme(legend.position = "none", text = element_text(size=12), plot.title = element_text(size=15),
      plot.subtitle = element_text(size=12))+
scale_fill_manual(values=c("#8111ee", "#7eee11", "#f0860f", "#0f79f0"))

plotgrid2 <- plot_grid(plot_length2, plot_wetweight2, plot_noleaves2, #combine the 3 plots into 1 figure
                      labels = c("A", "B", "C"),
                      ncol = 3, nrow = 1)

```

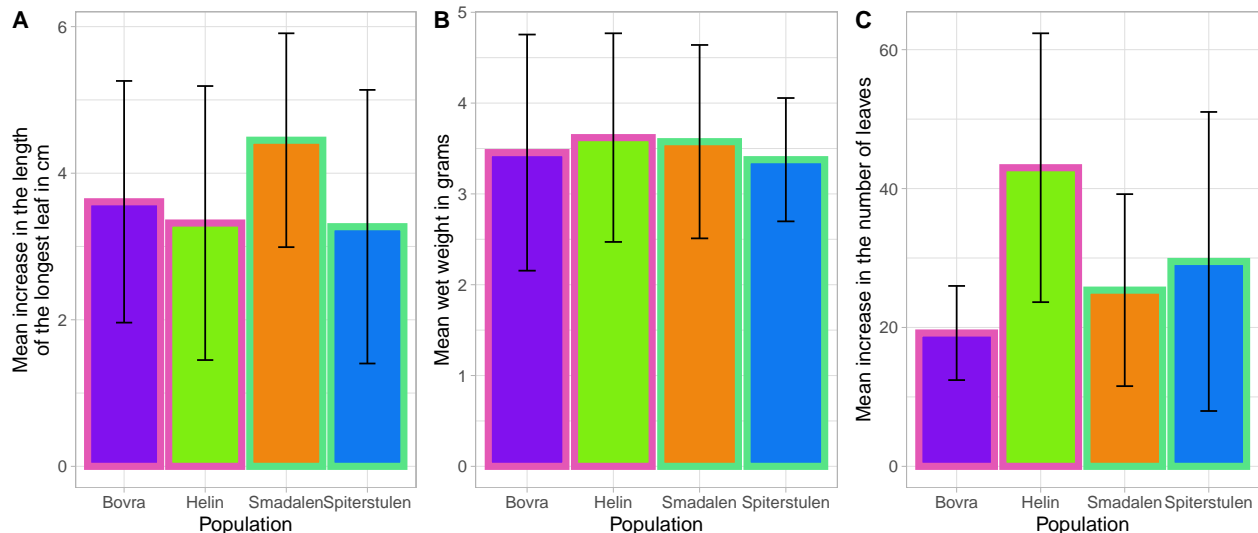


Figure 3: Figure 3: Bar graphs showing A) the increase in longest leaf length grouped by population, B) the increase in number of leaves grouped by population, and C) the increase in number of leaves grouped by population, with error bars depicting 1 standard deviation and the bar borders showing whether the population is large (pink) or small (green). The data is filtered to only show the salted groups.

Different types of phenotypes were scored, like the color and shape of the leaves. These scores can be visualized as well:

Generating even more figures


```

data_color <- data_complete %>% #prepare color data for plotting
  mutate(Color_graph = case_when(
    Color == 1 ~ "Green",
    Color == 2 ~ "Yellow",
    Color == 3 ~ "Red",
    Color == 4 ~ "Red and\yellow"))

plot_color <- data_color %>% #plot number of leaves of each color, grouped by treatment
  ggplot()+
  geom_bar(aes(x = Color_graph, fill = Treatment), position = position_dodge2(width = 0.9, preserve = "s
  theme_light()+
  labs(x="Color",
       y="Count")

plot_color2 <- data_color %>% #plot number of leaves of each color, grouped by population
  ggplot()+
  geom_bar(aes(x = Color_graph, fill = Population, color = Size), linewidth = 1,
           position = position_dodge2(width = 0.9, preserve = "single"))+
  theme_light()+
  labs(x="Color",
       y="Count")+
  scale_fill_manual(values=c("#8111ee", "#7eee11", "#f0860f", "#0f79f0"))+
  scale_color_manual(values = c("#e457b5", "#57e486"))+
  guides(color=guide_legend(title="Population size"))

plot_shape_salt <- data_complete %>% #plot number of leaves of each shape, grouped by treatment
  ggplot()+
  geom_bar(aes(x = Leaf_shape, fill = Treatment), position = position_dodge2(width = 0.9, preserve = "s
  theme_light() +
  labs(x="Leaf shape",
       y="Count")

plot_shape_popsiz <- data_salted %>% #plot number of leaves of each shape, grouped by population size
  ggplot()+
  geom_bar(aes(x = Leaf_shape, fill = Size), position = position_dodge2(width = 0.9, preserve = "single
  theme_light() +
  labs(x="Leaf shape",
       y="Count")+
  scale_fill_manual(values=c("#e457b5", "#57e486"))+
  guides(fill=guide_legend(title="Population size"))

plot_shape_pop <- data_salted %>% #plot number of leaves of each shape, grouped by population
  ggplot()+
  geom_bar(aes(x = Leaf_shape, fill = Population, color = Size), linewidth = 2,
           position = position_dodge2(width = 0.9, preserve = "single"))+
  theme_light() +
  labs(x="Leaf shape",
       y="Count")+
  scale_fill_manual(values=c("#8111ee", "#7eee11", "#f0860f", "#0f79f0"))+
  scale_color_manual(values = c("#e457b5", "#57e486"))+
  guides(color=guide_legend(title="Population size"))

img_leafshapes <- image_read(here("images/leaf_shapes.png")) %>% image_ggplot() #import the picture sho

```

```
plotgrid3 <- plot_grid(plot_color, plot_color2, plot_shape_salt, img_leafshapes, plot_shape_popsiz, pl
  labels = c("A", "B", "C", "D", "E", "F"),
  ncol = 2, nrow = 3)
```

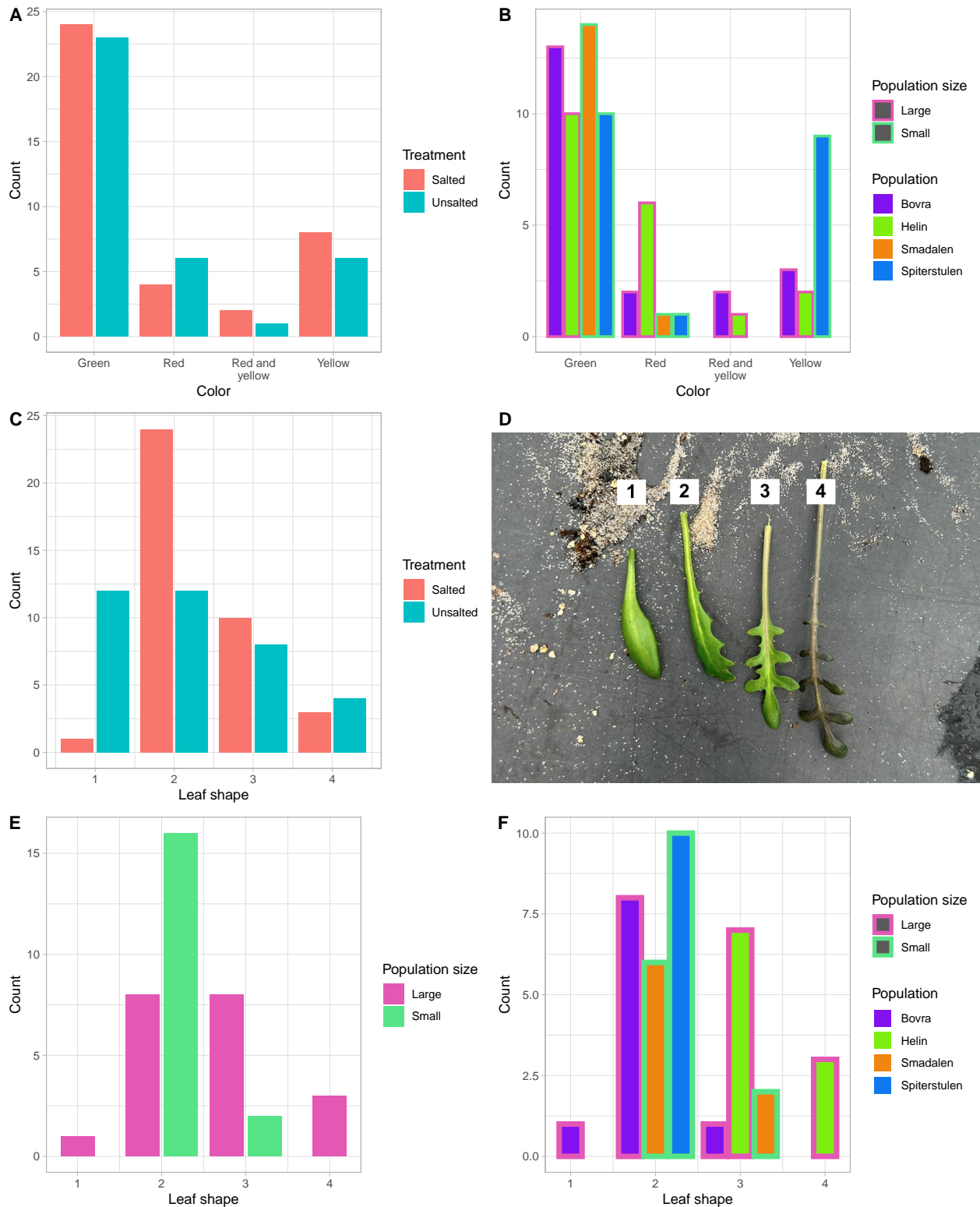


Figure 4: Figure 4: Bar graphs and an image showing A) the number of leaves of each color, grouped by treatment, B) the number of leaves of each color, grouped by population, C) the number of leaves of each shape, grouped by treatment D) the picture showing which shape is equal to which number, E) the number of leaves of each shape, grouped by population size, and F) the number of leaves of each shape, grouped by population. The data is filtered to only show the salted groups in B, E, and F.