Final Cap Dap draft

Sierra Hess

11/29/2022

Introduction:

I chose a data set collected by Mario Zuliani, Laura Brussa, Jessica Cunsolo, Angela Zuliani, and Christopher Lortie in 2021(“The effects of varying temperature on the germination of California natives and invasive plant species”). This dataset contains data about how temperature affects the germination and plant biomass of four different plant species found in a California forest. The data includes information about the time spent observing the plants, the number of repetitions, the light strength, temperature intensity, soil moisture, number of plants, and masses. From this data, I would like to model the masses based on temperature and how the given plants grow over time depending on temperature. I can also look into how light intensity and soil moisture had an affect on plant growth and mass and how different species react differently to different light, moisture and temperature conditions.

What are my expected results?

Test #1: Light intensity relation to dry mass.

I hypothesize that there will a positive relationship between light and dry mass for all species, meaning that there will be the highest masses when the light intensity is the highest. (Light intensity is measured in lightbulb strength of 40, 60, and 100 percent)The article “Timing and Duration of Supplemental Lighting during the Seedling Stage Influence Quality and Flowering in Petunia and Pansy,” by Wook Oh from Jeju National University, Erik Runkle from Michigan State University and Ryan M Warner who is also from Michigan State University looks into how light affects the dry mass of seedlings. They found that in both growing seasons, there was a positive slope in the data meaning that as the light increased, the dry mass increased. Also, the article gave r squared values which tell us how much of the variability in the data can be related back to the model that considered dry mass and light. In the first season, only 54% of the variability was related to the model, but in the second season, 78% of the variability was related to the model. Since different species require different amounts of light, I predict that some plants will survive better in high light intensity while others may be killed off by too high of a light intensity and may grow better in lower light areas. In general, I think that there will be a consistently positive relationship with a decreasing slope as light intensity increases and eventually, there will probably be an asymptote where the light is too intense to support life within a species. However, since the data only looks at three light intensities (high, low, medium), I don’t think that we will see that in this data. In order to visualize the model, I will model the data by looking at histograms, scatter plots, and a final box plot. I can use histograms by modeling just one species at a time and how mass changes based on light intensity and I can look at how the mass changes when we model both independent variables together. I also plan to use an anova test since we are looking at continuous data for the predictor and two types of categorical data data for the response.

Test #2

I predict that there will be a positive relationship between light intensity and germination. When there is a high light intensity, I expect that there will be increased levels of germination. According to the journal, "Effect of light on seed germination and seedling shape of succulent species from Mexico," by Joel Flores, Claudia González-Salvatierra, and Enrique Jurado, "The influence of light on germination has also been associated with plant growth form (seeds from columnar cacti being neutral photoblastic, and the barrel-shaped and globose being positive photoblastic); perenniality (light promotes the germination of annual species); plant size (seeds from shorter plants have a stronger light requirement for germination than those from taller plants); and seed size (seeds requiring light are small),"(Flores, González-Salvatierra, Jurado). From this, I believe that the more light that is present, generally, the more the plant is able to grow and the higher the number of germinated plants there will be. Without sufficient light, plants can't germinate, so I assume that the more light that is present, the more likely a plant is to grow and the more plants that will germinate with high light than low light.I plan to model this data by looking at bar graphs since a bar graph will show exactly how the different light intensities affect the number of plants germinated and this will allow us to visually see the differences in plant growth depending on light intensity. I also plan to use a multiple regression test since we have two discrete variables.

Test #3

I predict that there will be a positive relationship between soil moisture and germination. In the article, "Soil Moisture & Corn Seed Depth by R.L." by Bob Nielsen, Nielsen wrote about how, "adequate soil moisture at seed depth (not too wet, not too dry) during those first 48 hours helps ensure rapid germination of the seed. If the soil at seed depth is excessively dry, the seed will remain inert until moisture is replenished. If soil moisture is excessive at the seed depth (e.g., saturated), the seeds may die and rot." (Nielsen). This means that the seed requires a very specific amount of moisture to be germinated. Since the data shows different moisture levels in different pots, so it is likely that the moisture had an effect on the growth of the plants and light intensity wasn't the only factor affecting the data. I plan to model this data with a scatter plot that will show the relationship between the moisture and germination. Here, I will also use a multiple regression test since we are looking at a continuous independent variable and a discrete dependent variable.

Analysis

Before we begin any analysis, we need to check our data for possible errors/outliers.

```
summary(dry_mass)
```

```
##          date                          species       factor     table_number
##  02/06/2021:210   Bromus rubens         :210   High  :280   Min.   :1
##  09/09/2021:210   Layia platyglossa     :210   Low   :280   1st Qu.:1
##  22/07/2021:210   Phacelia tanacetifolia:210   Medium:280   Median :2
##  28/10/2021:210   Salvia columbariae    :210                Mean   :2
##                                                             3rd Qu.:3
##                                                             Max.   :3
##    pot_number        mass
##  Min.   : 1.0   Min.   :0.0000
##  1st Qu.:18.0   1st Qu.:0.0080
##  Median :35.5   Median :0.1065
##  Mean   :35.5   Mean   :0.2257
##  3rd Qu.:53.0   3rd Qu.:0.3127
##  Max.   :70.0   Max.   :2.4650
```

```
summary(germination)
```

```
##    calendar_date   julian_date        pot_ID                        species
##  01/08/2021: 210   Min.   :105.0   Min.   : 1.0   Bromus rubens          :2520
##  02/09/2021: 210   1st Qu.:144.0   1st Qu.:18.0   Layia platyglossa      :2940
##  03/05/2021: 210   Median :196.0   Median :35.5   Phacelia tanacetifolia:2520
##  04/06/2021: 210   Mean   :197.7   Mean   :35.5   Salvia columbarie      :2730
##  04/07/2021: 210   3rd Qu.:249.0   3rd Qu.:53.0
##  04/10/2021: 210   Max.   :297.0   Max.   :70.0
##  (Other)   :9450
##                               census        pendant_ID       table_ID
##  Greenhouse Temperature Experiment:10710   Min.   :15566   Min.   :1
##                                            1st Qu.:15573   1st Qu.:1
##                                            Median :15575   Median :2
##                                            Mean   :15580   Mean   :2
##                                            3rd Qu.:15585   3rd Qu.:3
##                                            Max.   :15598   Max.   :3
##
##  lightbulb_strength    factor     number_planted number_germinated
##  Min.   : 40.00     High  :3570   Min.   :40     Min.   : 0.00
##  1st Qu.: 40.00     Low   :3570   1st Qu.:40     1st Qu.: 0.00
##  Median : 60.00     Medium:3570   Median :40     Median : 2.00
##  Mean   : 66.67                   Mean   :40     Mean   : 8.68
##  3rd Qu.:100.00                   3rd Qu.:40     3rd Qu.:16.00
##  Max.   :100.00                   Max.   :40     Max.   :37.00
##
##  soil_moisture
##  Min.   : 1.0
##  1st Qu.: 8.0
##  Median :14.0
##  Mean   :14.7
##  3rd Qu.:20.0
##  Max.   :47.0
##
```

Based on the output, everything seems correctly formatted and the values seem to be reasonable with no clear outliers.

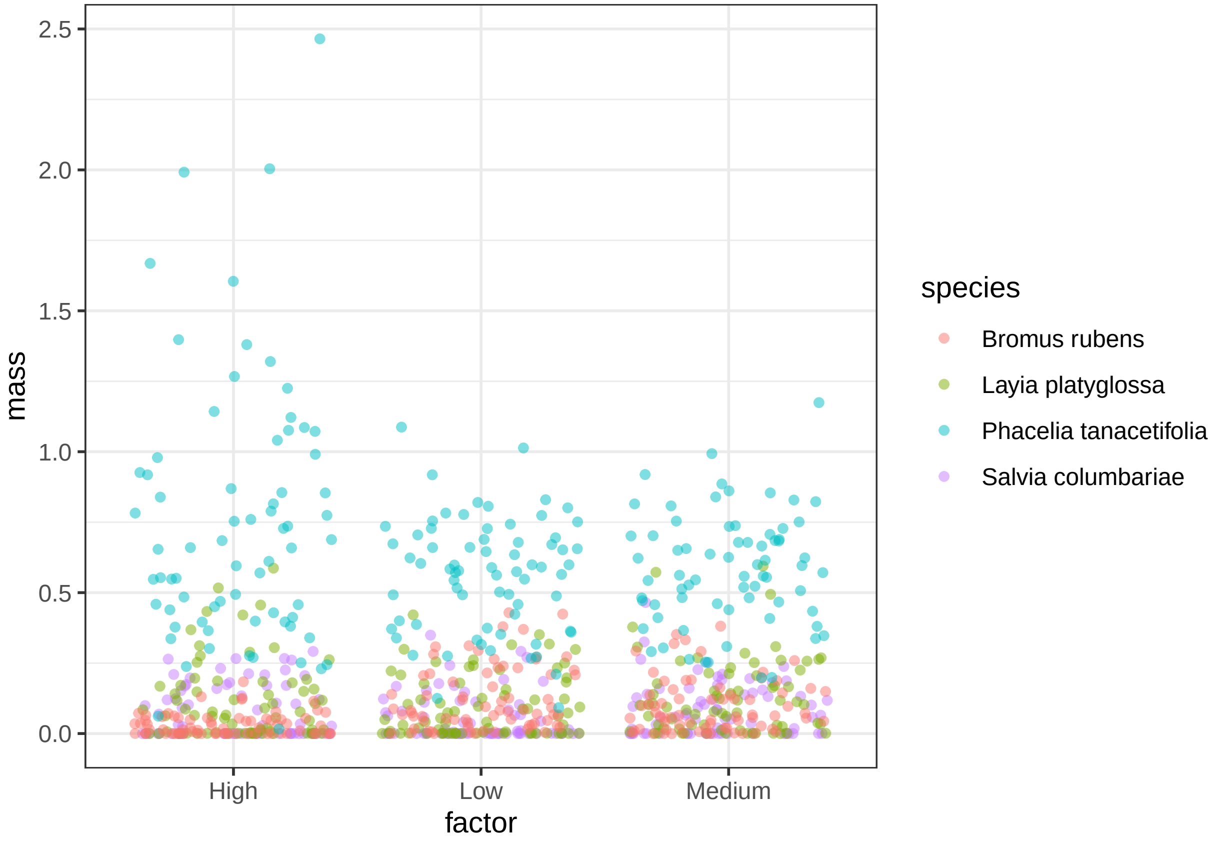
Also, by looking at the values on the chart, we can see which data appears to be continuous and which data is categorical which we can use to fit the correct models when running our tests.

Model #1: Light intensity relation to dry mass.

We can first look at a plot of our data

Lets first look at a scatterplot to see how the light intensity's relation to mass differs between species.

```
ggplot(dry_mass, aes(factor, mass, color = species, group = species))+
     geom_point(position = "jitter", size = 1.25, alpha = 0.5)+
     theme_bw()
```



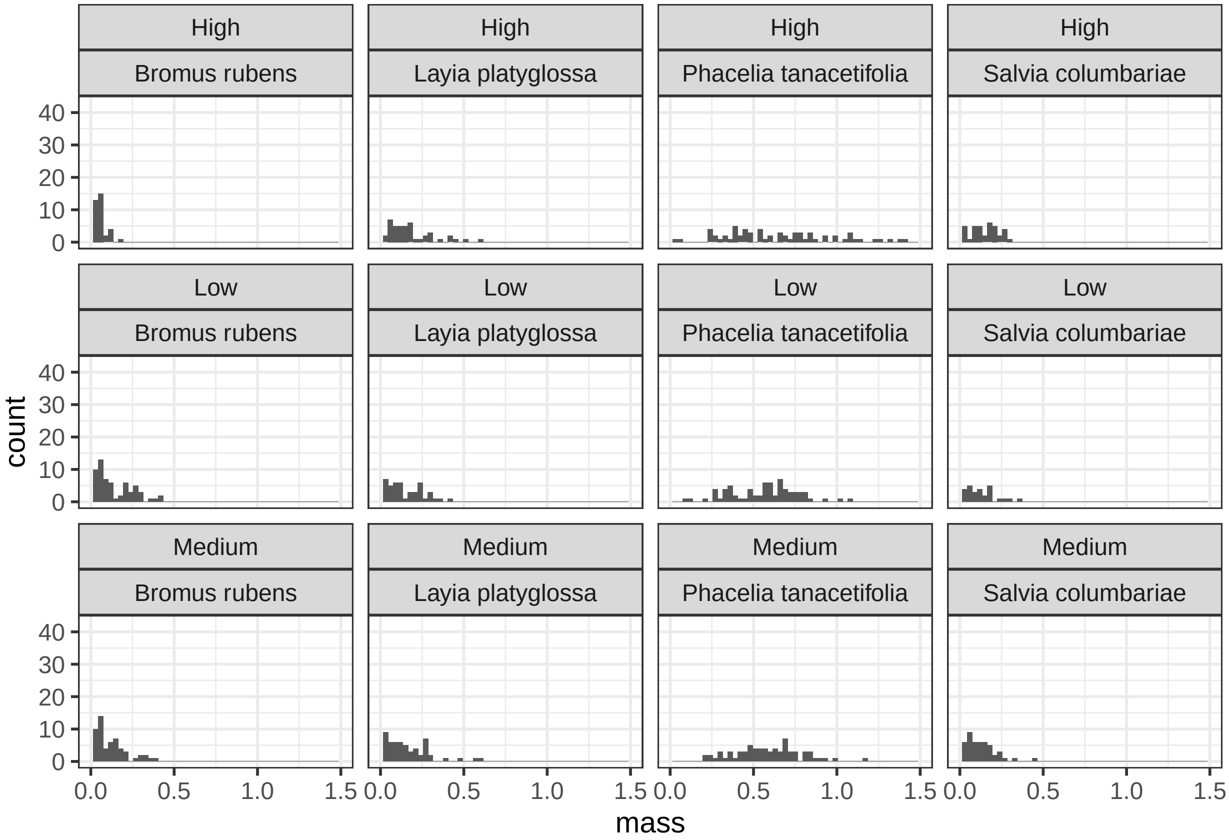
This shows us that there appear to be somewhat higher masses for high lighting in the Phaclia tanaccetifolia species but it isn't clear if there is much corellation between factors and mass besides that for all species.

Lets see if a histogram gives us a clearer picture.

```
ggplot(dry_mass, aes(x = mass))+
      geom_histogram(binwidth=.03)+
       facet_wrap(~ factor + species, ncol = 4)+
       theme_bw()+
       xlim(-.0002, 1.5)
```

```
## Warning: Removed 5 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 24 rows containing missing values (`geom_bar()`).
```



This shows us that there seems to be varying reactions to light intensity across the different species. Some species have higher masses with higher light while other species do better in conditons with lower light, and we can't see a definate correlation when looking at mass and light factors.

Now lets create a model and test our hypothesis statistically

Again, we hypothesized that there would be higher masses for higher light factors when we group by species.

For this model, we are comparing 2 types of categorical data with continuous data, so we should compare the two using a two way anova test.

```
mod = lm(mass ~ factor + species, data = dry_mass)
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: mass
##            Df Sum Sq Mean Sq  F value  Pr(>F)
## factor      2  0.165  0.0824   2.3567 0.09536 .
## species     3 46.269 15.4231 441.0214 < 2e-16 ***
## Residuals 834 29.166  0.0350
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod)
```

```
##
```

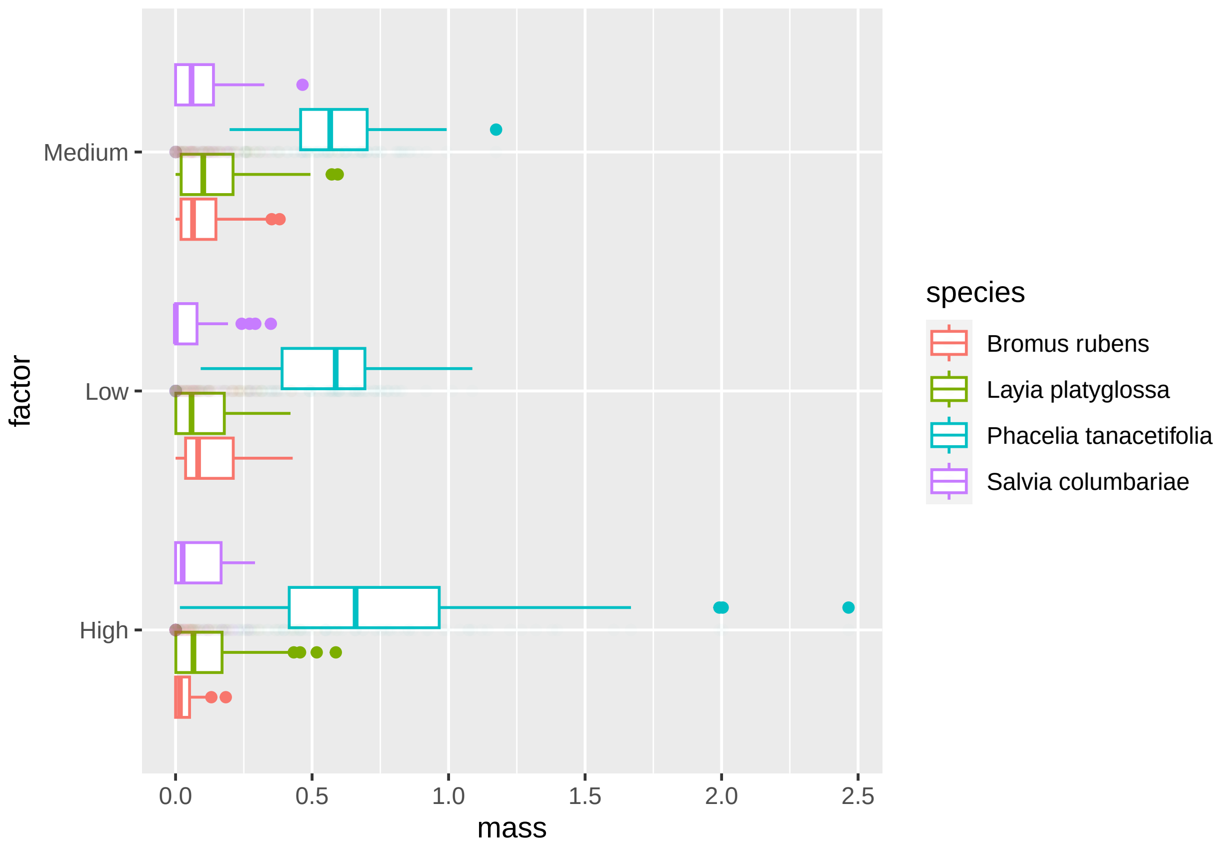
```
## Call:
## lm(formula = mass ~ factor + species, data = dry_mass)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.63332 -0.08969 -0.03840  0.06835  1.81568
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     0.10347    0.01580   6.546 1.03e-10 ***
## factorLow                      -0.03417    0.01580  -2.162   0.0309 *
## factorMedium                   -0.01979    0.01580  -1.252   0.2110
## speciesLayia platyglossa        0.02881    0.01825   1.579   0.1148
## speciesPhacelia tanacetifolia  0.54586    0.01825  29.910  < 2e-16 ***
## speciesSalvia columbariae     -0.01378    0.01825  -0.755   0.4505
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.187 on 834 degrees of freedom
## Multiple R-squared:  0.6142, Adjusted R-squared:  0.6119
## F-statistic: 265.6 on 5 and 834 DF,  p-value: < 2.2e-16
```

In this model, the factor has a p value $> .05$, so is not significantly useful in the model, but may be somewhat useful. The species has a p value $< .05$, so is useful in the model.

This means that there is a clear relationship between species and mass, but the light factor doesn't always accurately predict mass, so we can reject our hypothesis given that there isn't a super clear relationship between light factor and mass.

Lets summarize with a final plot

```
ggplot(dry_mass, aes(factor, mass, color = species))+
  geom_boxplot() +
  geom_point(alpha = 0.01) +
  coord_flip()
```



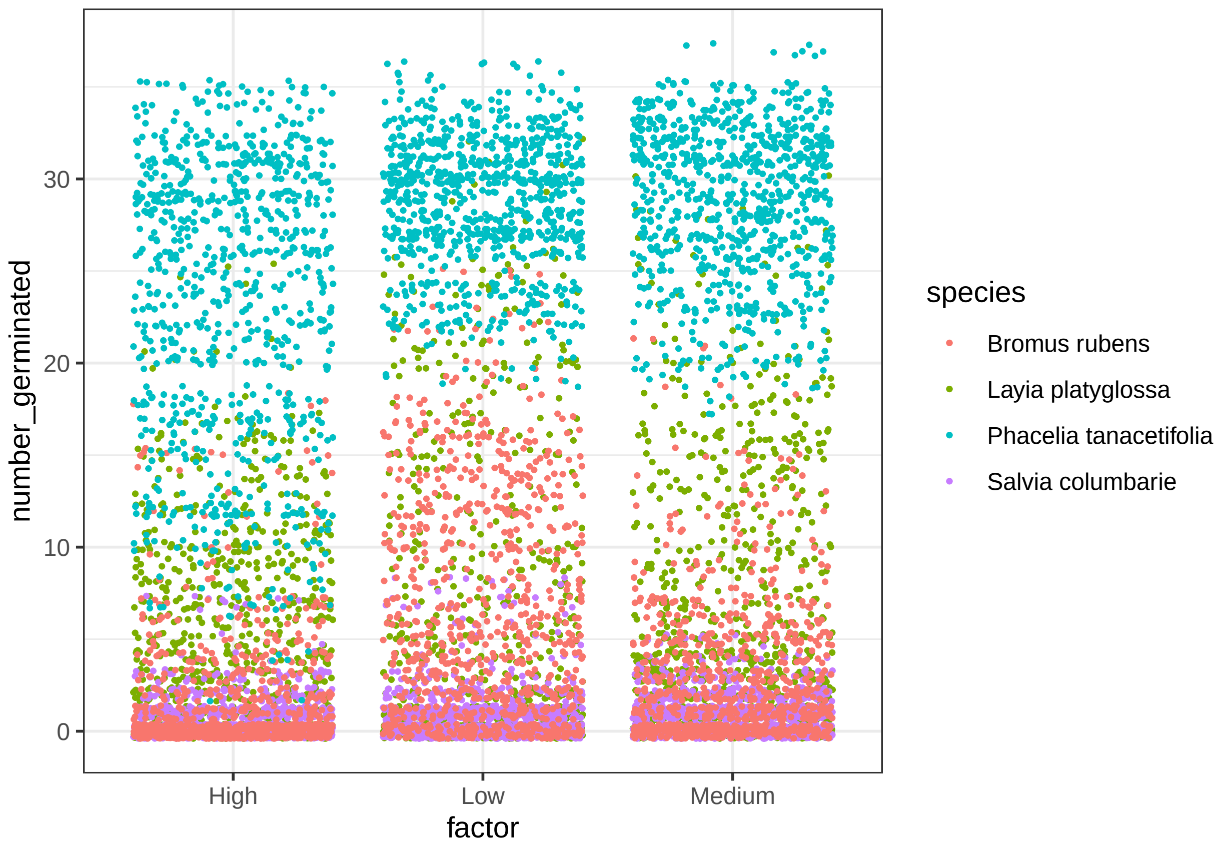
This plot shows us that there seems to be fairly similar average masses despite light factors for each of the species. There is some variation of masses, but not a significant ammount.

Model #2: Light intensity relation to germination

We can now look at a plot of our data

Lets first look at a scatterplot to see how the light intensity's relation to mass differs between species.

```
ggplot(germination, aes(factor, number_germinated, color = species, group = species))+
     geom_point(position = "jitter", size = .5, alpha = 2)+
     theme_bw()
```



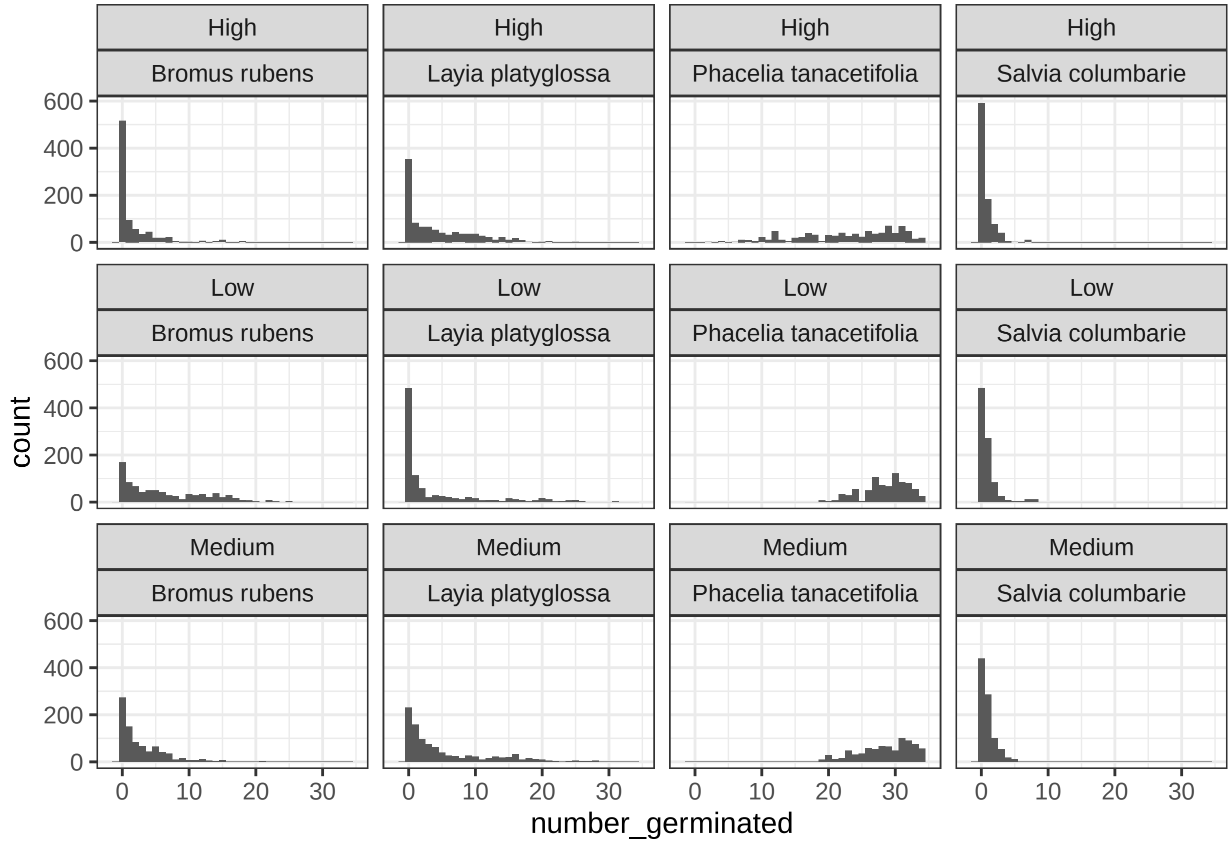
There seems to be somewhat of a correlation between factor and number germinated based on the differences in color gradient on the plot.

Lets see if a histogram gives us a clearer picture.

```
ggplot(germination, aes(x = number_germinated))+
      geom_histogram(binwidth= 1)+
       facet_wrap(~ factor + species, ncol = 4)+
       theme_bw()+
       xlim(-2,35)
```

```
## Warning: Removed 20 rows containing non-finite values (`stat_bin()`).

## Warning: Removed 24 rows containing missing values (`geom_bar()`).
```



Here, we can see that there again appears to be somewhat of a difference between how the many of the species germinate based on light intensity.

Now we can look at our data with an anova test

We can use an anova test since we are comparing our continuous count data to two types of categorical data.

```
mod = lm(number_germinated ~ factor + species, data = germination)
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: number_germinated
##                Df  Sum Sq Mean Sq  F value    Pr(>F)
## factor          2   12115    6058   247.68 < 2.2e-16 ***
## species         3 1122159  374053 15293.90 < 2.2e-16 ***
## Residuals 10704  261795      24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod)
```

```
##
## Call:
## lm(formula = number_germinated ~ factor + species, data = germination)
##
## Residuals:
```

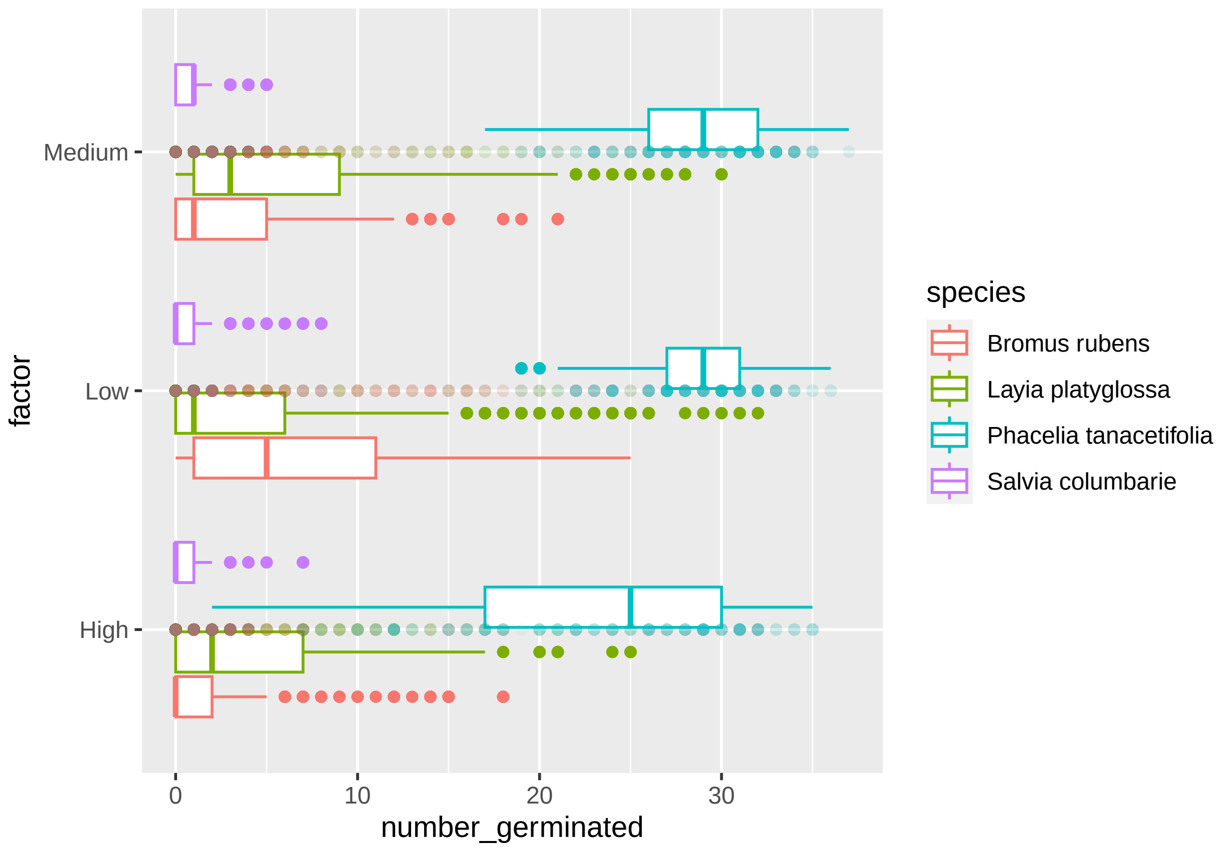
```
##      Min       1Q   Median       3Q      Max
## -23.4696  -3.1252  -0.7701   1.8748  26.3291
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     2.1664     0.1195  18.134  < 2e-16 ***
## factorLow                       2.4669     0.1171  21.075  < 2e-16 ***
## factorMedium                    1.9588     0.1171  16.734  < 2e-16 ***
## speciesLayia platyglossa        1.0376     0.1343   7.728 1.19e-14 ***
## speciesPhacelia tanacetifolia  23.3032     0.1393 167.260  < 2e-16 ***
## speciesSalvia columbarie       -2.8633     0.1366 -20.958  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.945 on 10704 degrees of freedom
## Multiple R-squared:  0.8125, Adjusted R-squared:  0.8124
## F-statistic:  9275 on 5 and 10704 DF,  p-value: < 2.2e-16
```

P values are both $< .05$, so there is statistical significance to a model that uses factor and species to predict the number germinated.

Therefore there is a clear relation between factor and number germinated.

Lets summarize with a final plot

```
ggplot(germination, aes(factor, number_germinated, color = species))+
  geom_boxplot() +
  geom_point(alpha = 0.01) +
  coord_flip()
```



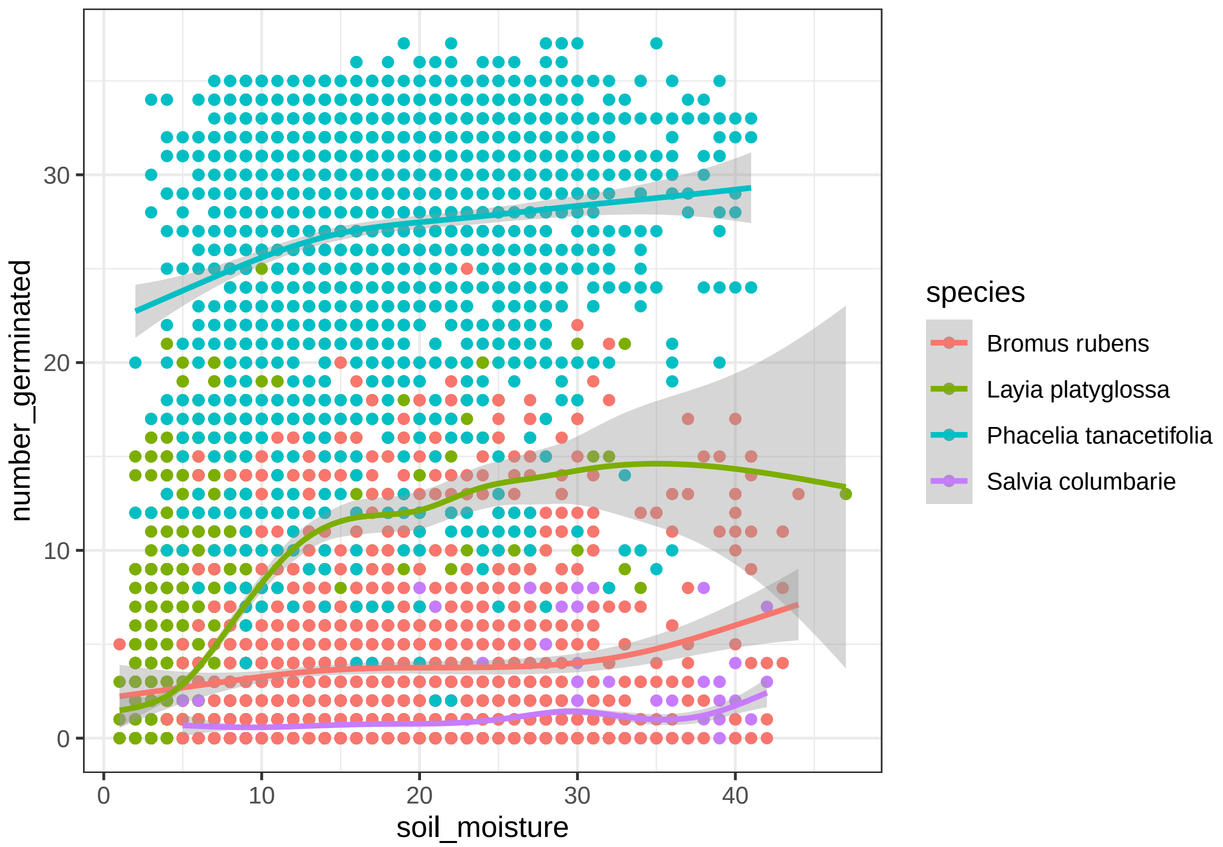
This plot shows us how there is a clear difference between the mean number germinated based on the light factor for the different species.

Model #3: Soil moisture relation to Germination

We can first look at a plot of our data

```
ggplot(germination, aes(x = soil_moisture, y = number_germinated, color = species, group = species))+
  geom_point() +
  geom_smooth() +
  theme_bw()
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



This plot shows us that plants appear to germinate the best when the moisture levels are not too high and not too low.

Now lets do a statistical test

Since we have continuous data (soil moisture, integer) and a discrete data(number germinated) seperated by species, we can fit a multiple linear regression and run an anova test.

```
mod = lm(number_germinated ~ soil_moisture + species, data = germination)
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: number_germinated
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## soil_moisture     1   79043   79043  3240.9 < 2.2e-16 ***
## species           3 1055943  351981 14432.0 < 2.2e-16 ***
## Residuals     10705  261084      24
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(mod)
```

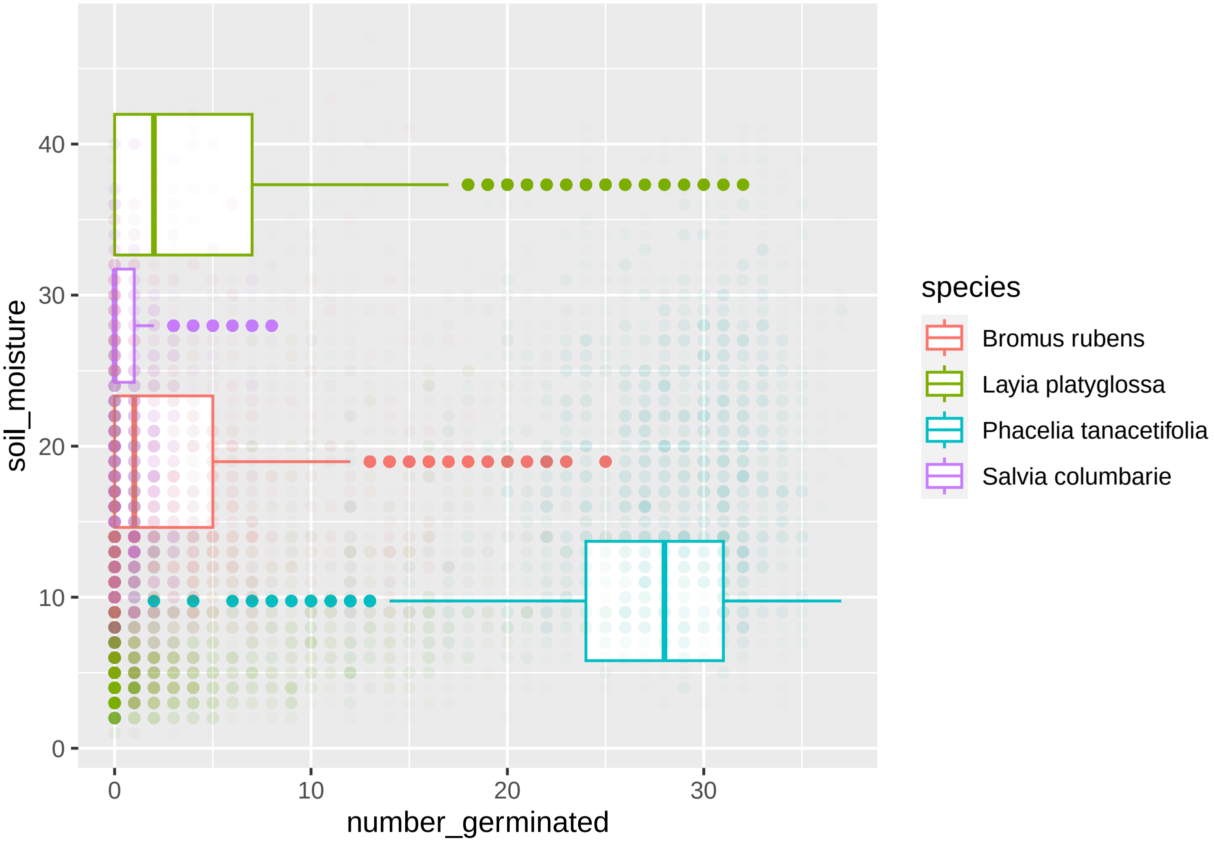
```
##
## Call:
## lm(formula = number_germinated ~ soil_moisture + species, data = germination)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.5740  -3.1766  -0.4423   1.9528  27.4892
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    0.806953   0.157980   5.108 3.31e-07 ***
## soil_moisture                  0.164706   0.007182  22.933  < 2e-16 ***
## speciesLayia platyglossa       2.715657   0.152736  17.780  < 2e-16 ***
## speciesPhacelia tanacetifolia 23.143501   0.139301 166.140  < 2e-16 ***
## speciesSalvia columbarie      -2.900910   0.136435 -21.262  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.939 on 10705 degrees of freedom
## Multiple R-squared:  0.813,  Adjusted R-squared:  0.8129
## F-statistic: 1.163e+04 on 4 and 10705 DF,  p-value: < 2.2e-16
```

This tells us that the model is useful and there is a clear relationship between the soil moisture and the number of plants germinated.

Lets summarize with a final plot

```
ggplot(germination, aes(soil_moisture, number_germinated, color = species))+
  geom_boxplot() +
  geom_point(alpha = 0.01) +
  coord_flip()
```



This plot shows us how there is a clear difference between the mean number germinated based on the soil moisture for the different species.

Biological Summary

We first found that there was no clear relationship between light intensity and dry mass. We saw that there was a large p value which meant that there wasn't a clear correlation. This was somewhat surprising since I expected the light factor to have a strong relationship with plant growth, but there was only a weak correlation (if any) between the light intensity and plant mass. Next, we looked at how light affected germination. We saw that there was a significant correlation between light factor and germination since our p value was low. This makes sense because light is required for plant growth, so you would expect too little light to result in no germination, but too high of a light factor would kill off the plant as well. Finally, we can see that there is a clear difference in the number of germinated plants with different soil moistures given the low p value for model 3. This makes sense since soil is required for plant growth, but some species of plants also can't grow in conditions where they are over saturated with water.

Challenges

One challenge that I faced was figuring out which statistical tests to do and how to get the data into the correct forms to do the test. For example, originally did several tests for each species separately before I realized that I could put all of my data into one singular test. In the future, I will make sure I find the best method for analyzing my data before jumping in and trying to make several repetitive plots.

Also, I had a lot of trial and error when trying to remember code. I had to look through old assignments in order to remember how to set up different graphs and save data in the correct formats and was able to use stack overflow to help me understand how I could make my plots more efficient.