Capstone Data Analysis Project Report

Introduction

The goal of this research is to look at various relationships among forest metrics including canopy density, regeneration, fern abundance, and tree cores. I collected this data in multiple forests in St. Lawrence County during summer 2022.

Figure 1: Here I am coring a tree!

Here, I conduct four statistical tests. Below I will provide a justification for each test I conduct to support my hypotheses.

Hypothesis 1: The amount of regeneration present in a site is impacted by Canopy Density. I expect to see a decrease in regeneration as canopy density increases. Justification: Madsen P. and Larson J.B. (1997). Natural regeneration of beech (*Fagus sylvatica L.*) with respect to canopy density, soil moisture, and soil carbon content. *Science Direct,* 97(2), 95-105.

This study found that "generally, the number of saplings was reduced by increased canopy density. The number of saplings varied greatly in plots with open canopies; whereas, the number of saplings was consistently low in plots with a closed canopy" (p. 99).

Hypothesis 2: As canopy density increases, the amount of bare ground on the forest floor will increase. Justification: Supporting literature: Goldblum D. (2009). The effects of treefall gaps on understory vegetation in New York State. *Journal of Vegetation Science,* 8(1), 125-132.

This study found greater vegetation abundance under more open canopies, which suggests there would be less vegetation abundance, and therefore more bare ground, under more closed canopies.

Hypothesis 3: The amount of ferns will be statistically different between the Kip and Donnerville forests. Justification: Rooney T.P. (2001). Deer impacts on forest ecosystems: a North American perspective. *Forestry: An International Journal of Forest Research,* 74(3), 201-208.

This study found that "excessive deer browsing can create forests dominated by ferns in the understory" (p. 205). As I have gone through game camera data from these sites, I have noticed differing amounts of deer at the forests so it seems like there could also be different amounts of ferns between forests.

Hypothesis 4: The amount of regeneration will differ between Peavine and Degrasse forests. Justification: Supporting literature: Madsen P. and Larson J.B. (1997). Natural regeneration of beech (*Fagus sylvatica L.*) with respect to canopy density, soil moisture, and soil carbon content. *Science Direct,* 97(2), 95-105.

This study looked at how light availability impacts regeneration. Because Peavine is a deciduous forest, and Degrasse is coniferous, there could be differing amounts of light reaching the forest floor as deciduous leaves are broader and tend to produce more shade. For this reason, it seems possible that there could be different amounts of regeneration between these two forests.

Analysis

First, I'll set up my script

Now, I will import my two (cleaned up) datasets. To see how the data were cleaned, view Data_Exploration.Rmd.

Statistical Test 1: Canopy Density and Regeneration Count

Before I begin the data analysis workflow, I need to manipulate my data because these two variables are found in different dataframes.

A. First I will summarize the regen data to get a dataframe that includes Forest, Plot, and the number of total regen (not distinguished by species) for that plot

```
regen_by_plot <- regen %>%
  group_by(Forest, Plot_num) %>%
  summarize(num_seedlings = mean(Regen_count))
```

B. Next I will summarize the density data to get a dataframe with the average canopy density (densiometer method) for each plot

```
avg_density <- density %>%
  group_by(Forest, Plot_num) %>%
  summarize(density = mean(Densiometer))
```

C. Now I will create a new column in each of the dataframes I created above that combines the forest and plot so that each site has a unique ID I can use for combining the dataframes

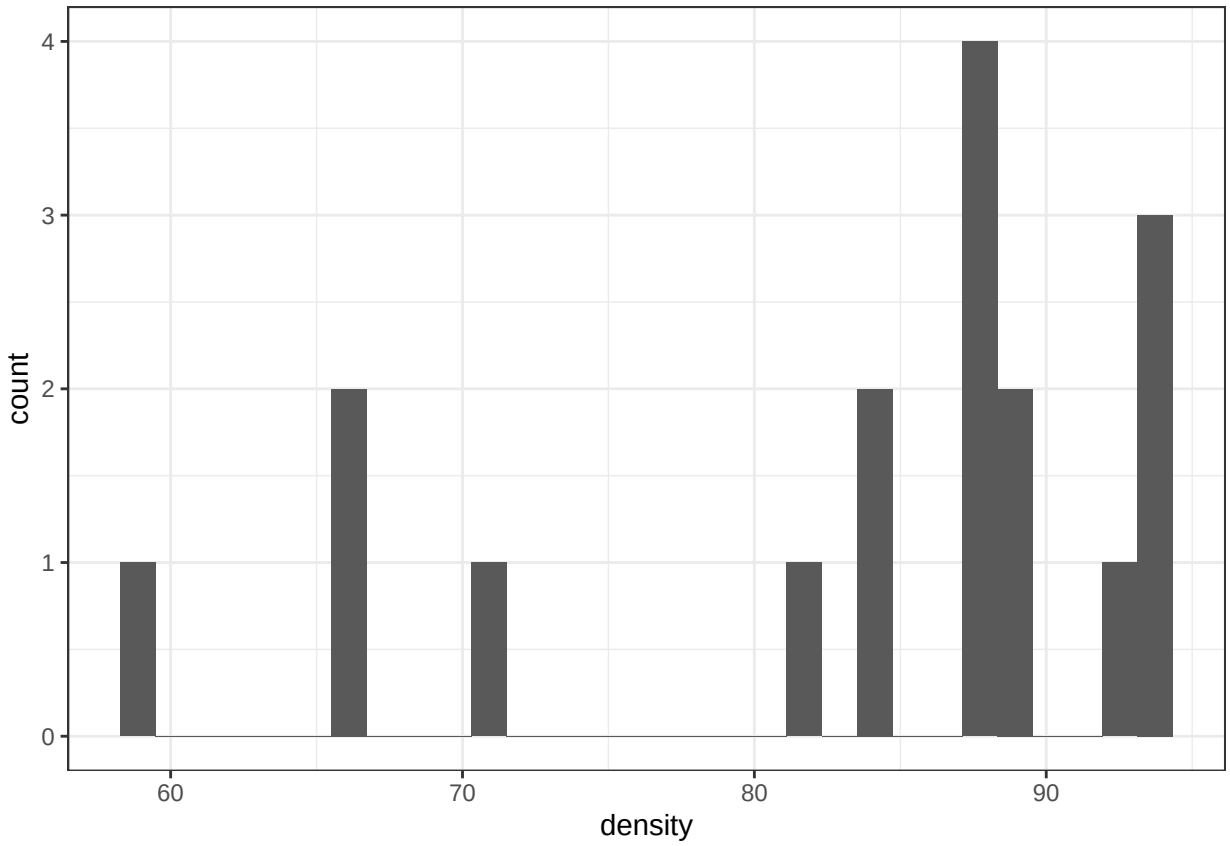
```
regen_by_plot$ID <- paste(regen_by_plot$Forest,regen_by_plot$Plot_num)

avg_density$ID <- paste(avg_density$Forest,avg_density$Plot_num)
```

D. Now I will join the dataframes

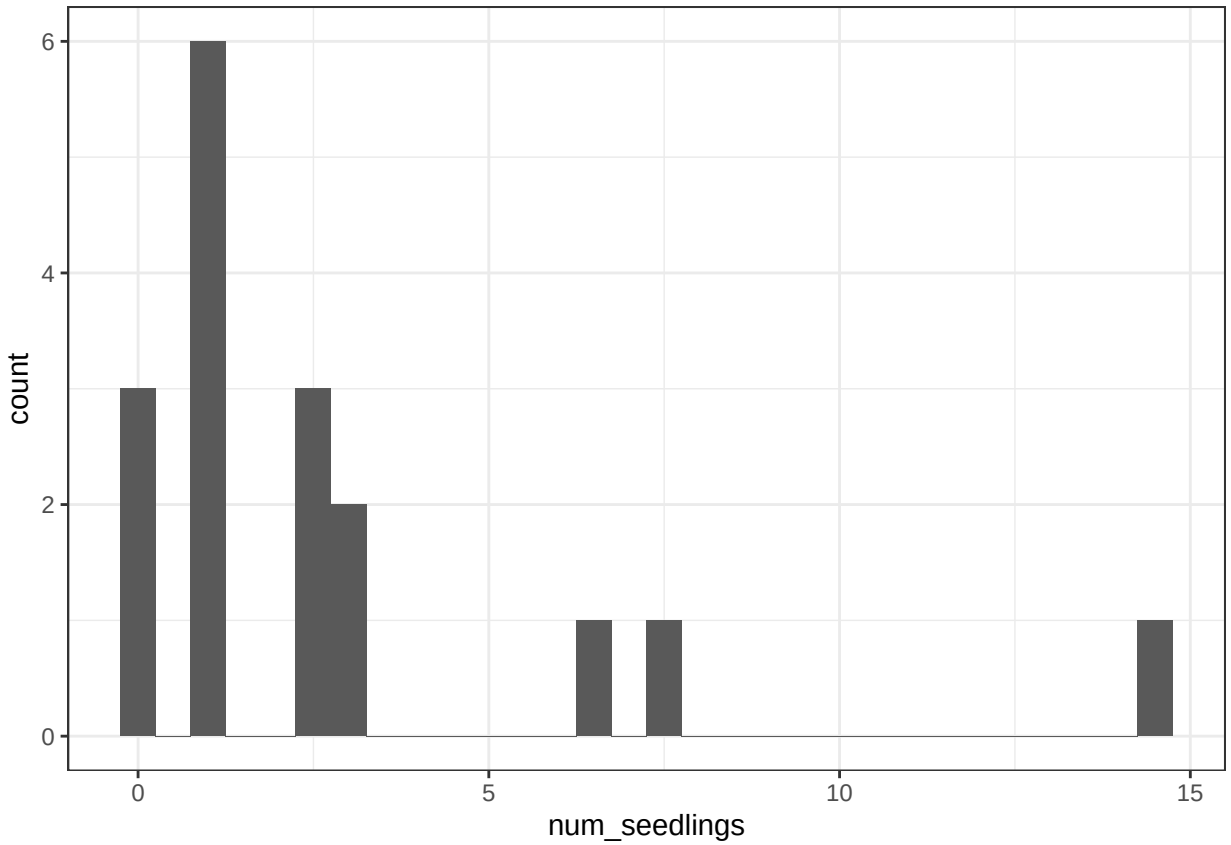
```
combined_data <- regen_by_plot %>% full_join(avg_density, by = c("ID")) %>%
  select(-c(Plot_num.x, Forest.x)) %>%
  rename(c("Forest"="Forest.y", "Plot_num" = "Plot_num.y"))
```

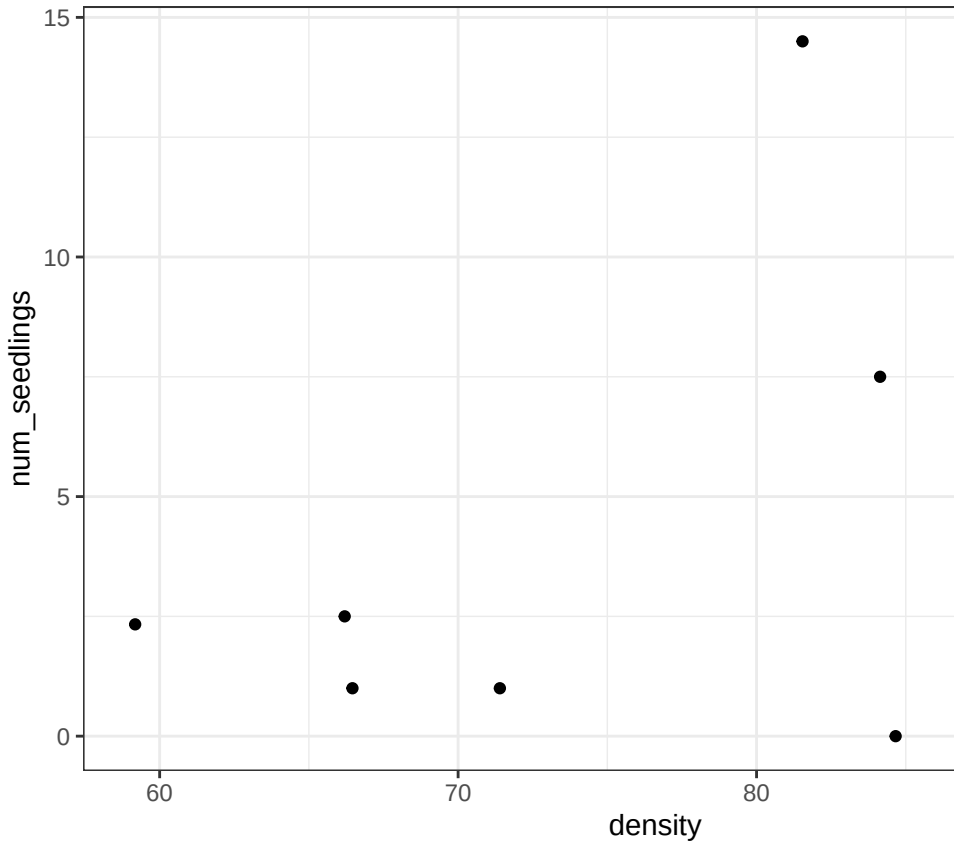
I will move into the analysis workflow

1. I will plot a histogram of each variable I am comparing Canopy density first.



Now I will look at the Regen_count variable


6
4
2
0
count
0
5
10
15
num_seedlings



2. I will plot the variables together

3. Guess relationship: Based on the data, it appears there could be a weak positive relationship between canopy density and the number of seedlings.

Total rise: 14

Total run: 58 to 94, so 44

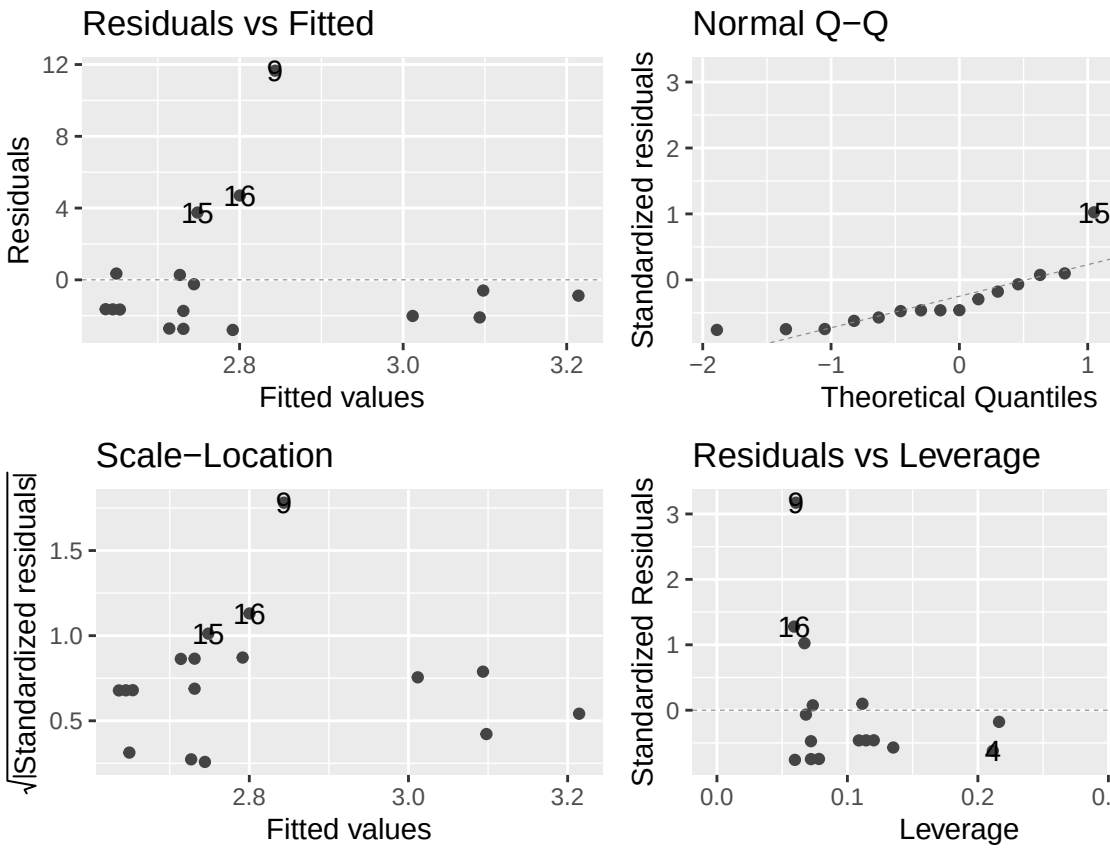
Slope: $15/44 = 0.34$

Y intercept: 0ish

Decide on statistical test: I will use a linear regression since the predictor variable (canopy density) is continuous, and the response variable (regeneration) is also continuous.

```
regen_dens_mod <- lm(num_seedlings ~ density, combined_data)
```

4. Create the Model



5. Check Assumptions

The normal Q-Q graph looks good, and indicates the data are fairly normally distributed, with some smaller and larger values deviating. The Residuals vs. fitted graph indicates relatively equal variance. I am comfortable moving forward.

6. Interpret Model:

```
## Analysis of Variance Table
##
## Response: num_seedlings
##           Df  Sum Sq Mean Sq F value Pr(>F)
## density    1   0.513  0.5132  0.0357 0.8527
## Residuals 15 215.591 14.3728
```

The P-value is 0.8527 which is large - accept null hypothesis (that there is no meaningful statistical relationship between regen and canopy density)

I will also run the summary

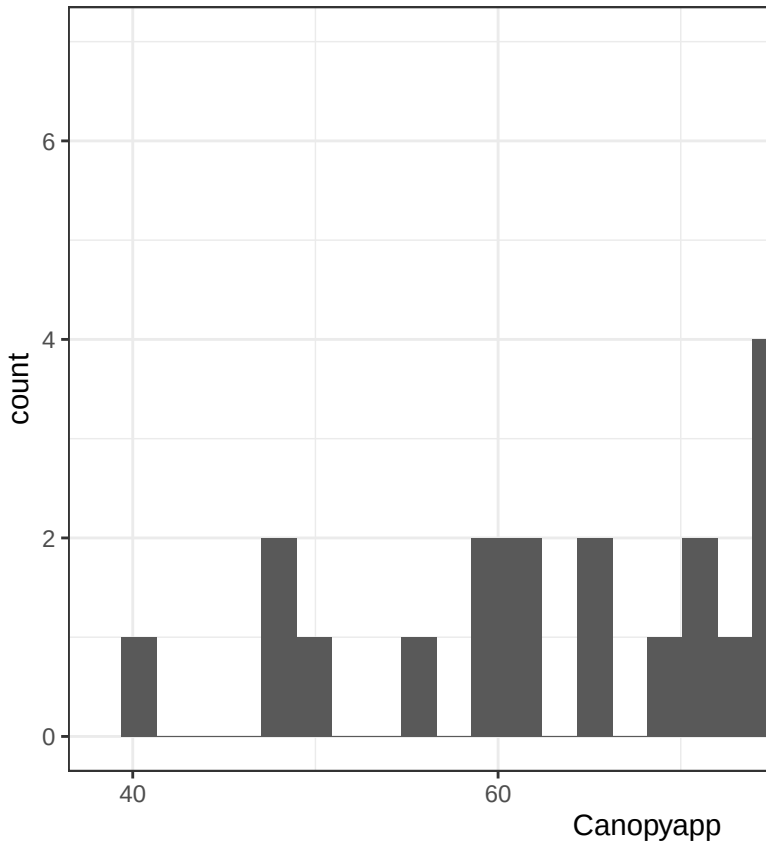
```
##
## Call:
## lm(formula = num_seedlings ~ density, data = combined_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.792 -2.012 -1.636  0.273 11.657
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.19621    7.37404   0.569    0.578
## density     -0.01659    0.08780  -0.189    0.853
##
## Residual standard error: 3.791 on 15 degrees of freedom
## Multiple R-squared:  0.002375,   Adjusted R-squared:  -0.06413
## F-statistic: 0.0357 on 1 and 15 DF,  p-value: 0.8527
```

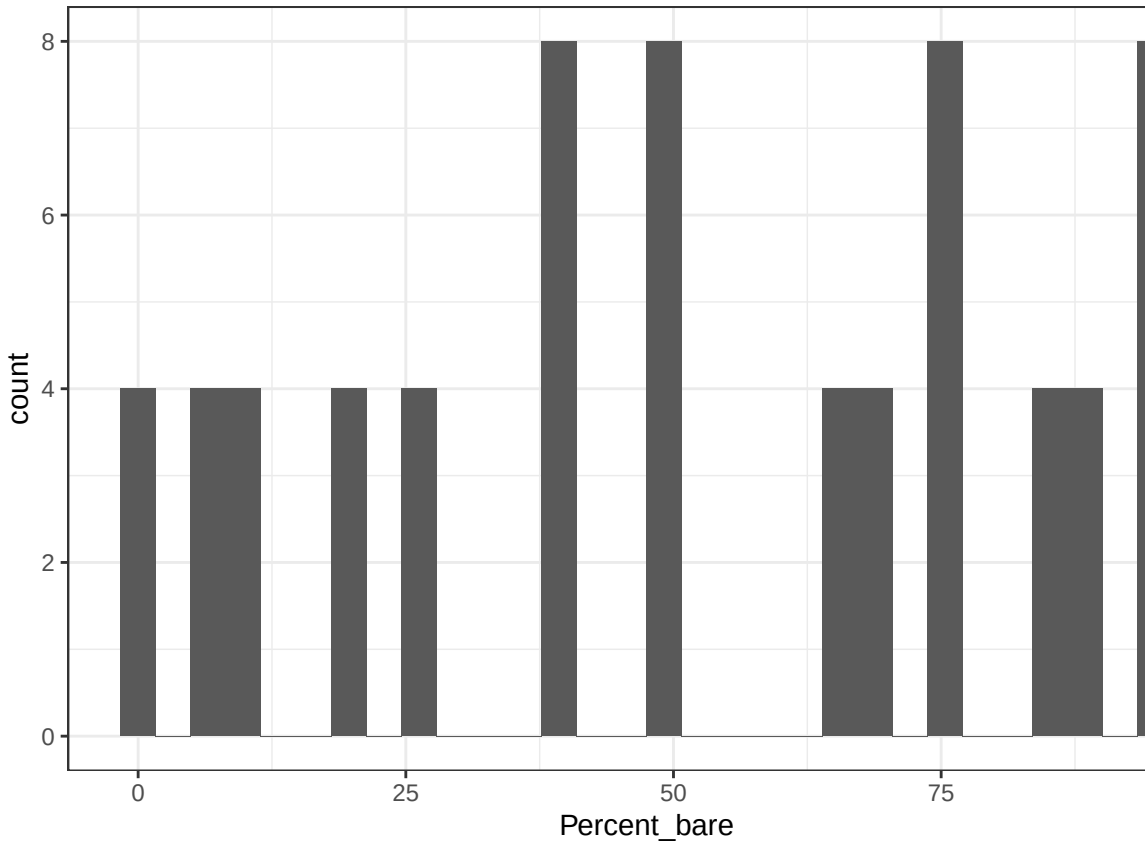
Estimates R-squared to be 0.002375, which means only 0.2% of variation in regen is explained by canopy density.

I will not make a publication plot since no statistically significant findings were found.

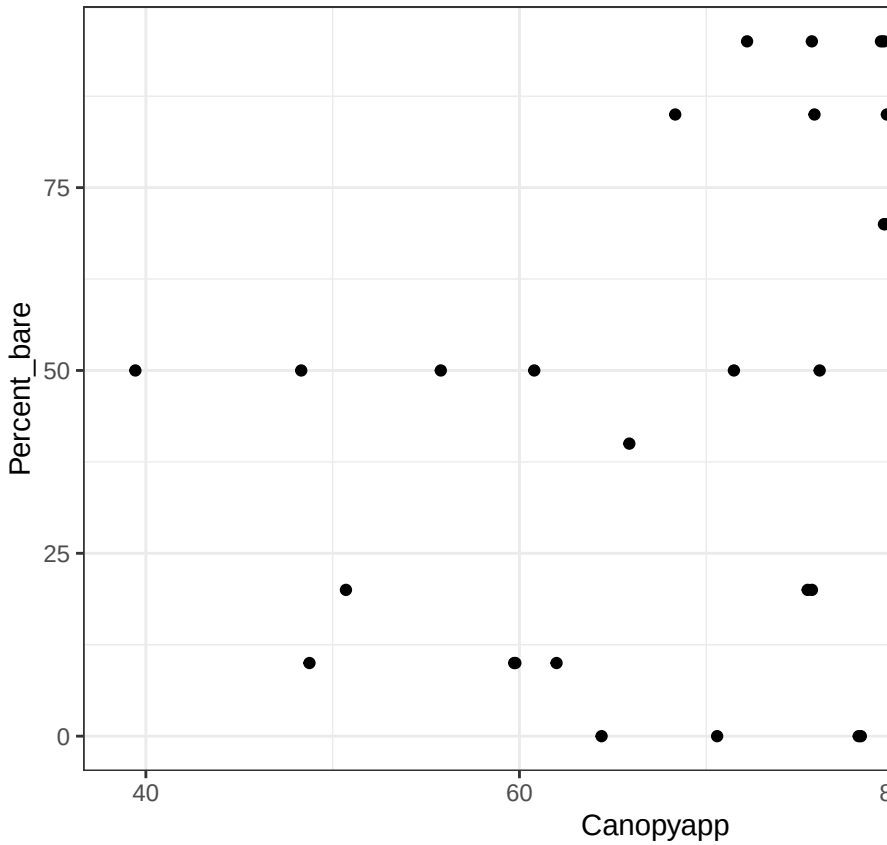
Statistical Test 2: Canopy Density and Percent Bare Ground



1. First I will plot each variable Canopy density first



Now percent bare ground



2. Now I will plot the variables together

3. Guess the relationship: From looking at the data, it appears there could be a weak positive relationship between canopy density and percent bare ground

Total rise: 100

Total run: 38 to 98 = 60

Slope: 100/60 = 1.67

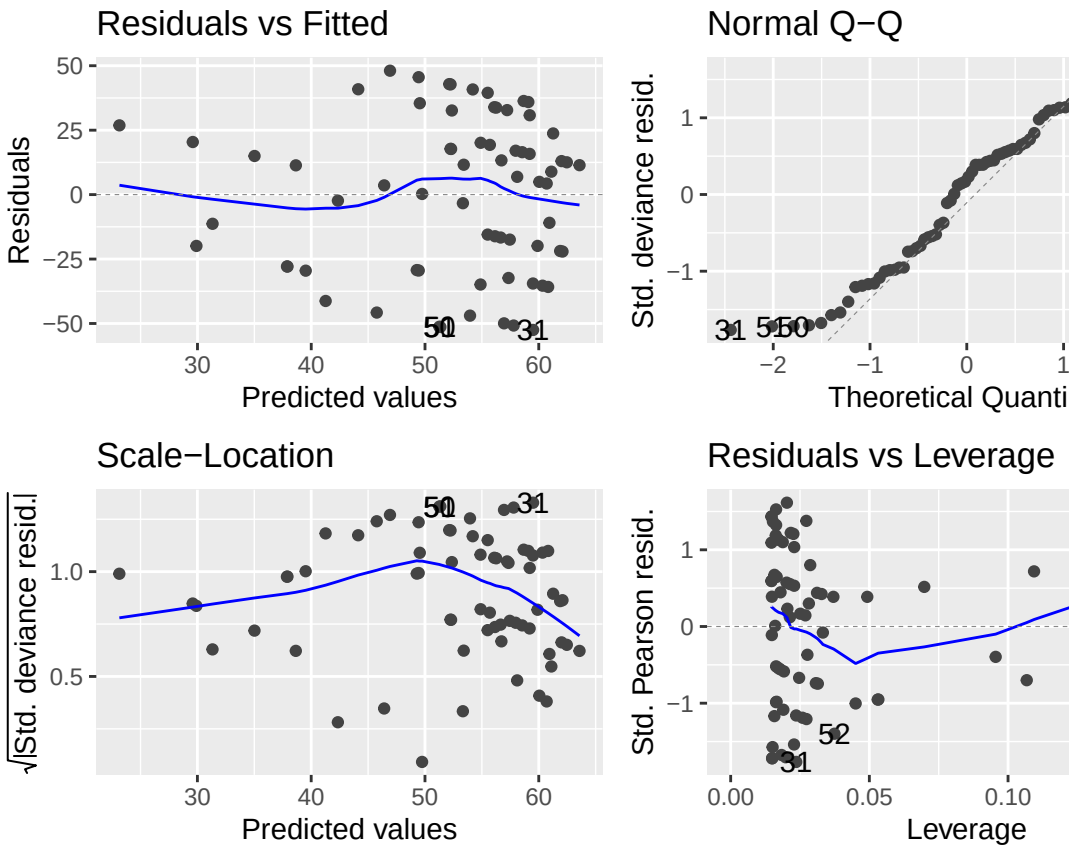
Y intercept: 0ish?

Null Hypothesis: there is no significant statistical relationship between canopy density and percent bare ground

4. Create a model Because the predictor (canopy density) is continuous but bounded between 0-100, and the response (percent bare ground) is continuous, I will use a generalized linear model.

I'll construct the model

```
density_bare_ground_glm <- glm(Percent_bare ~ Canopyapp, data = density, family = gaussian())
```



5. Check the assumptions

Normal Q-Q plot: shows that most of the data points fit a normal distribution, with some of the higher and lower points deviating, but nothing excessive

Residuals vs fitted plot: shows that variance of residuals is fairly equal (slight waiver in line but nothing too drastic)

Based on the autoplot() results, I feel moderately comfortable moving ahead and assuming these data meet the glm assumptions.

6. Interpret the model

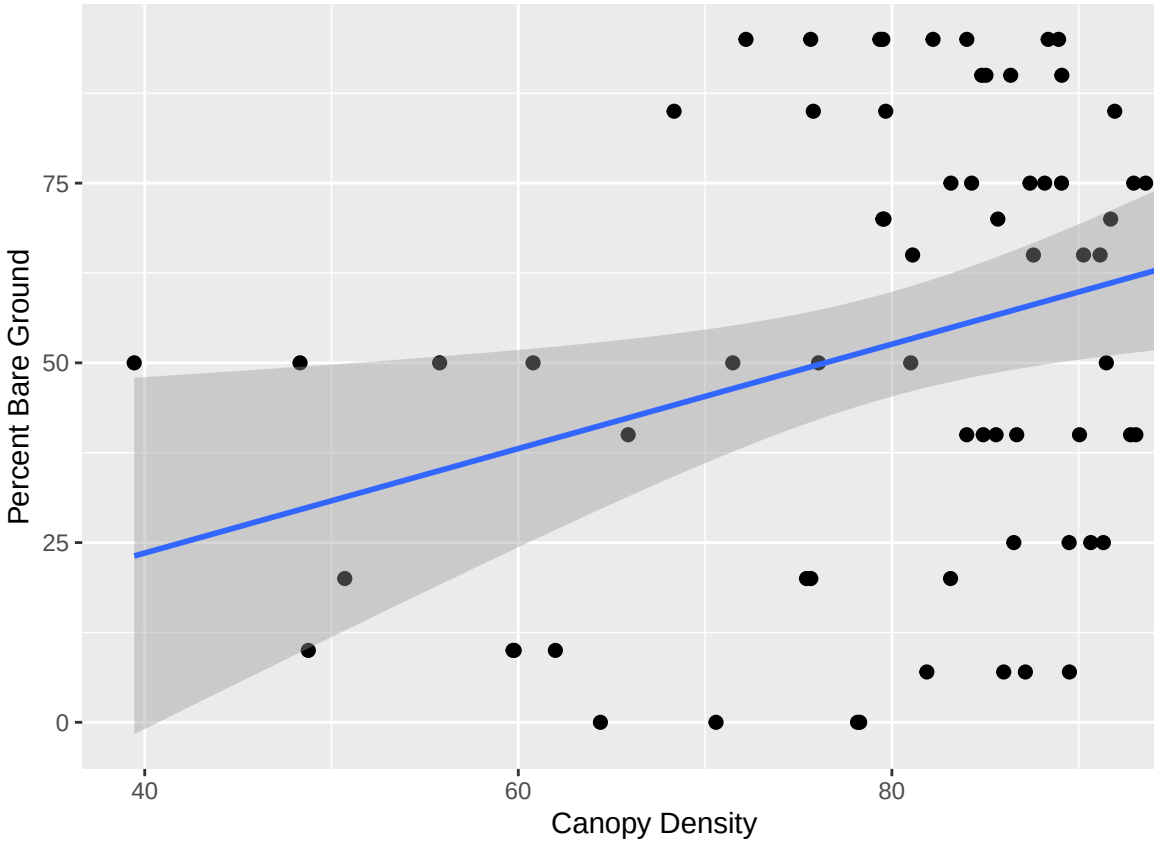
```
## Analysis of Deviance Table
##
## Model: gaussian, link: identity
##
## Response: Percent_bare
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev
## NULL                         67      65281
## Canopyapp  1   5530.9        66      59750
```

The P-value is 0.01603, which is fairly small. I reject the null hypothesis, and accept that canopy density has a significant effect on ground cover.

I will also run a summary of the test

```
##
## Call:
## glm(formula = Percent_bare ~ Canopyapp, family = gaussian(),
##     data = density)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -52.503  -27.881    5.924   21.244   48.075
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.5055    23.7379  -0.232    0.817
## Canopyapp     0.7263     0.2938   2.472    0.016 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 905.3044)
##
##     Null deviance: 65281  on 67  degrees of freedom
## Residual deviance: 59750  on 66  degrees of freedom
## AIC: 659.91
##
## Number of Fisher Scoring iterations: 2
```

Slope is estimated to be -5 and slope as 0.7, which is fairly similar to what I predicted.



7. Publication Plot

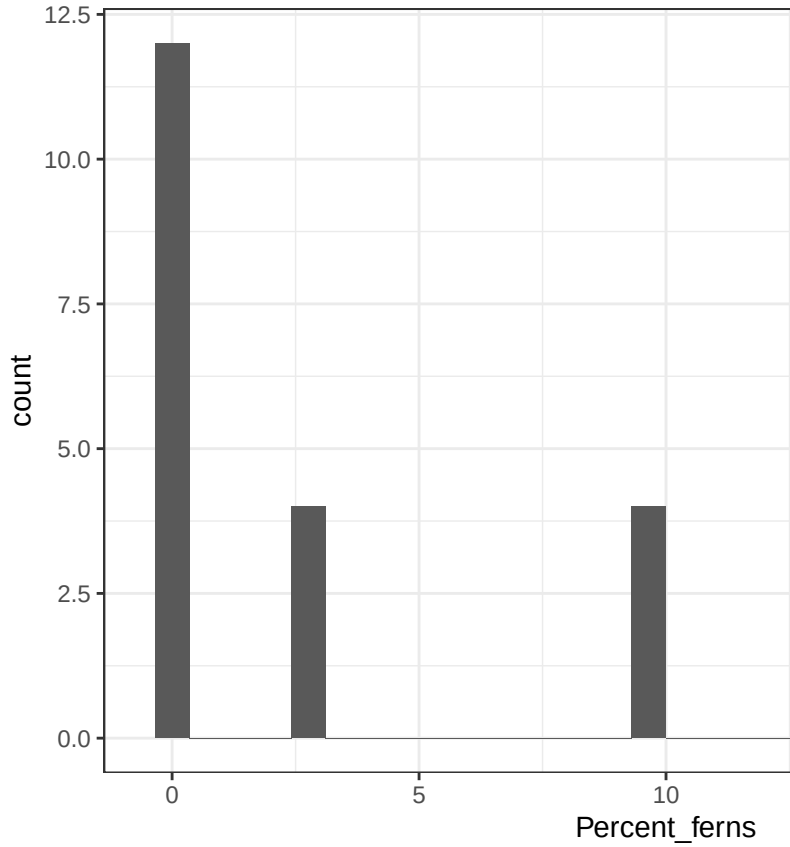
Swap to geom_smooth(method = "lm") as per notes in Data_Analysis.Rmd

Statistical Test 3: Percent Ferns Between Kip and Donnerville

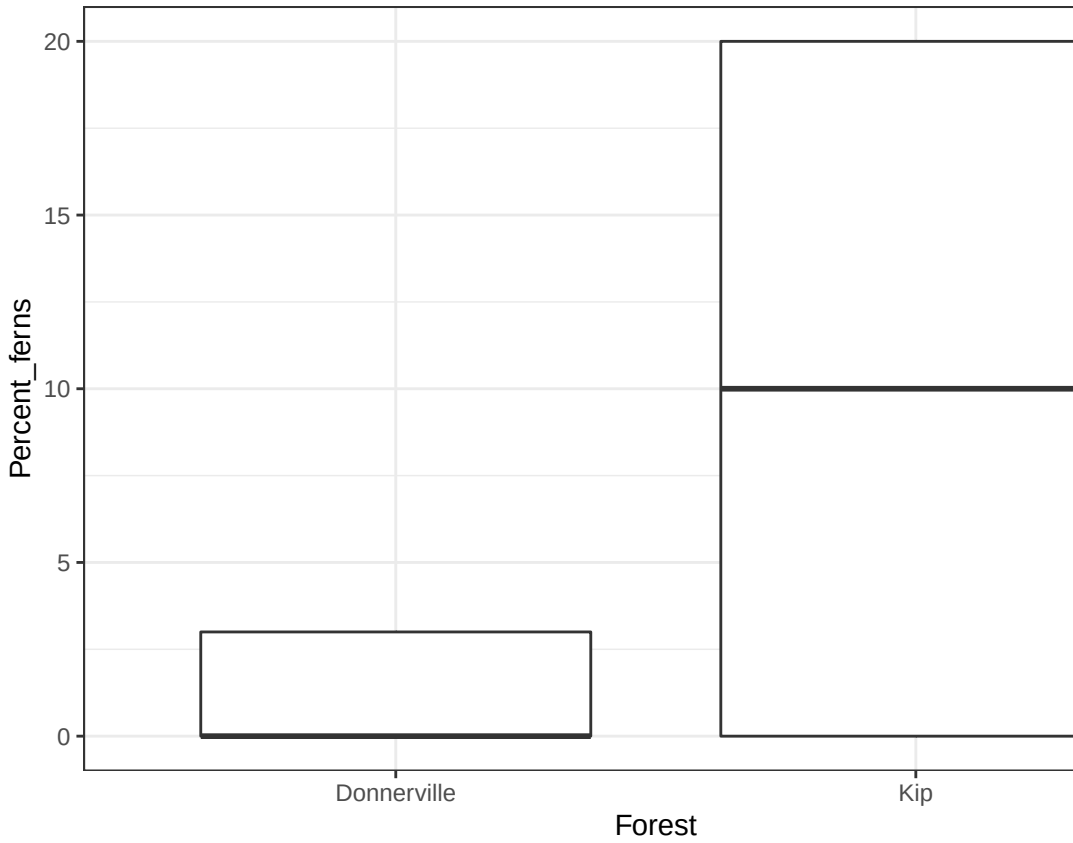
Before I start the data analysis workflow, I need to filter my data to only get fern data from Kip and Donnerville. The reason I choose to look at these two forests is because during my exploratory data phase I saw there could be a difference between the amount of ferns in these two forests.

```
ferns <- density %>%
  filter(Forest == "Kip" | Forest == "Donnerville")

ferns$Forest <- factor(ferns$Forest)
levels(ferns$Forest)
```

```
## [1] "Donnerville" "Kip"
```



1. **Plot each variable** First, I'll plot the percent ferns



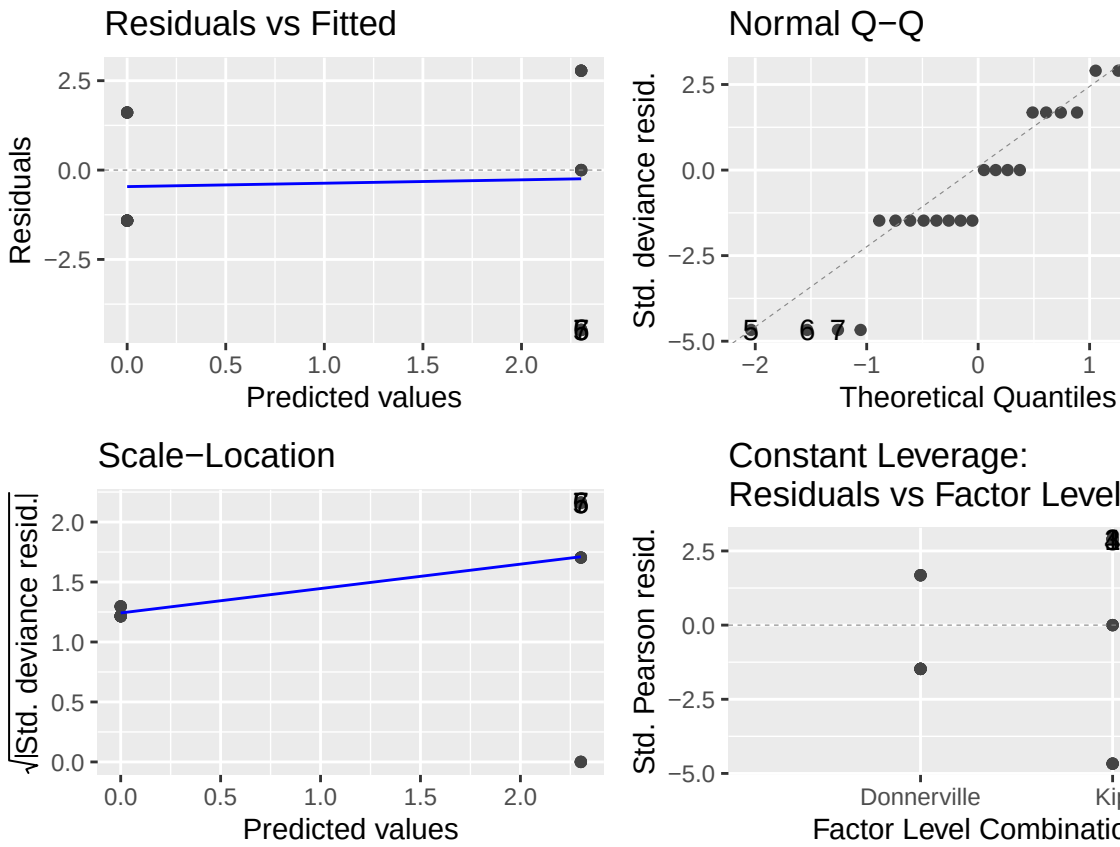
2. Plot variables together

3. Guess relationship: From looking at the data, I predict there to be a significant difference between the amount of ferns at each forest. It appears there are more ferns in the Kip forest.

Null hypothesis: there is no significant relationship between the forest and the amount of ferns present

4. Create model I will fit a generalized linear model because I am dealing with a categorical predictor variable with more than two groups (Forest) and a continuous, yet bounded, (0-1) response variable (percent ferns)

```
ferns_forest_glm <- glm(Percent_ferns ~ Forest, data = ferns, family = poisson())

#poisson family has a log link (because outcome must be positive and I'm dealing with percents) and ass
```



5. Check assumptions

Normal Q-Q plot: it appears the data fit to the normal distribution by sort of stepping up (which makes sense given the data) with some deviation for small and large values.

Residuals vs Fitted: shows that variance of residuals is relatively equal

Based on the autoplot() results, I will feel ok with proceeding.

6. Interpret the model

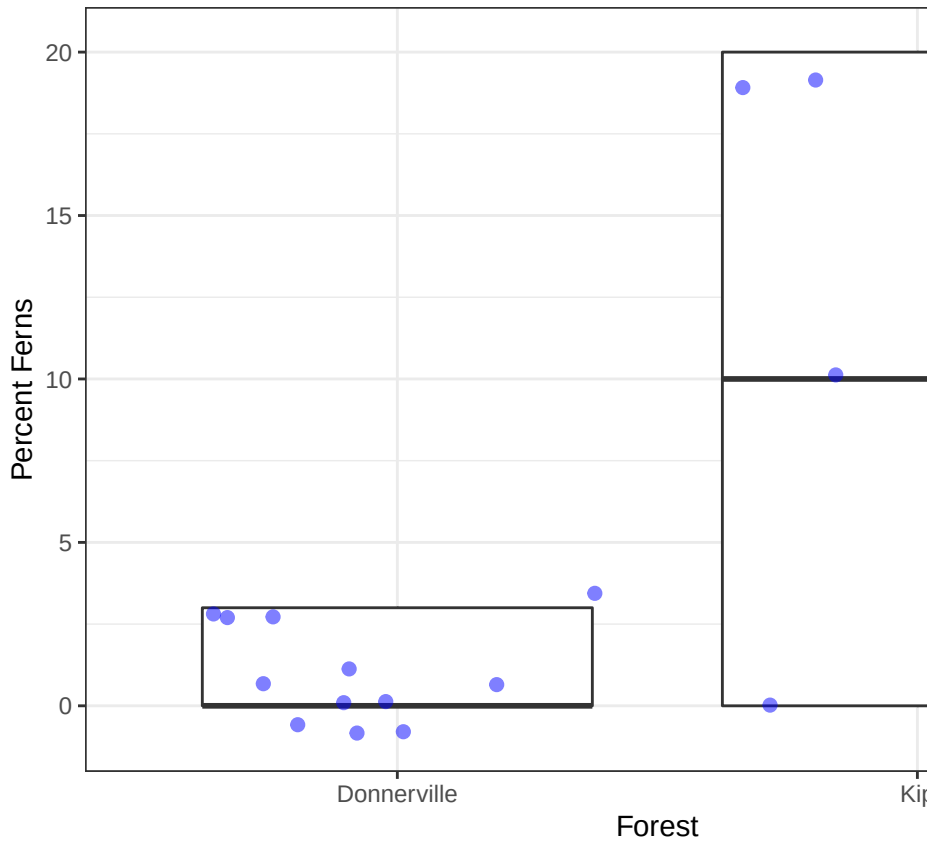
```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Percent_ferns
##
## Terms added sequentially (first to last)
##
##
##        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                     23     239.84
## Forest  1   102.57        22     137.27 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notics a very small p-value!

Also run the summary table

```
##
## Call:
## glm(formula = Percent_ferns ~ Forest, family = poisson(), data = ferns)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.4721  -1.4142  -0.7071   1.6099   2.7795
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 8.967e-15  2.887e-01   0.000        1
## ForestKip   2.303e+00  3.028e-01   7.605 2.84e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 239.84  on 23  degrees of freedom
## Residual deviance: 137.27  on 22  degrees of freedom
## AIC: 189.23
##
## Number of Fisher Scoring iterations: 6
```

The small p-value leads me to reject my null hypothesis. There is a significant difference in the amounts of ferns between Donnerville and Kip.



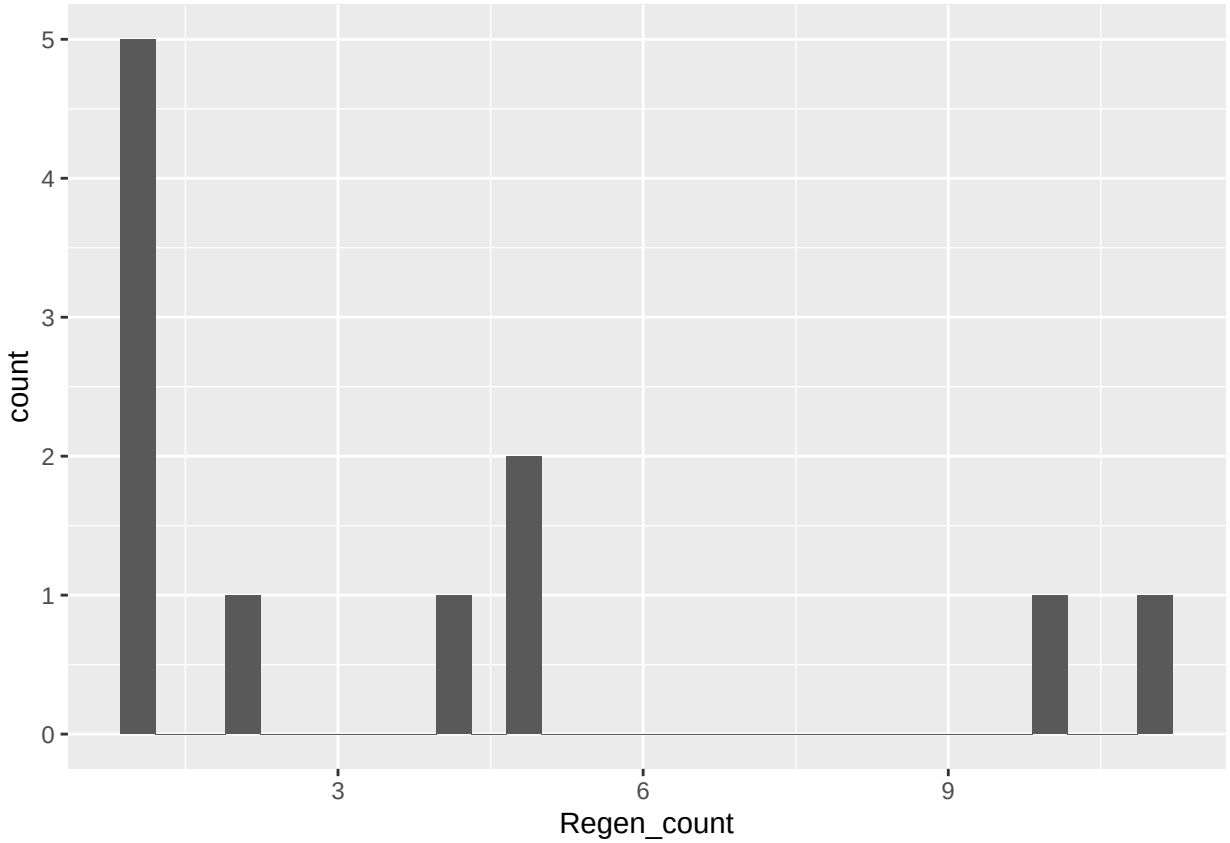
7. Now I will make a publication plot

Statistical Test 4: Regen count between Peavine and Degrasse

First I will filter my data to get regen count for just South Hammond and Degrasse forests. The reason I am choosing these forests is because I think it would be interesting to compare the amount of regen between a mostly coniferous forest (Degrasse) and a mostly deciduous forest (South Hammond).

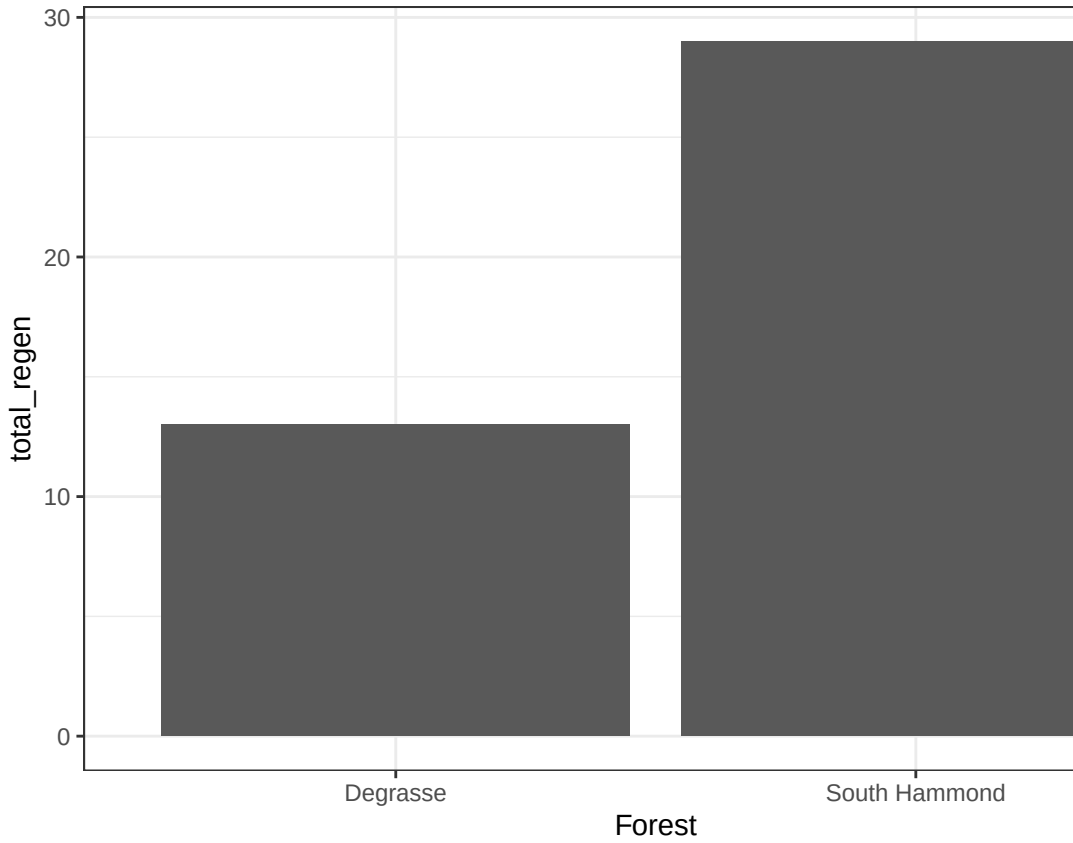
```
## [1] "Degrasse"      "South Hammond"
```

1. Plot the variables Now I will plot a histogram of regeneration count to look for any outliers



Now I will manipulate the data before I plot again to get a column with the total regen count regardless of species

```
total_regen <- regen_chi_squared %>% group_by(Forest) %>% summarise(total_regen = sum(Regen_count))
```



2. Plot varaibles together

3. Guess relationship: From looking at the data, it appears there is a significant difference between the amount of regeneration found between the degrasse and south hammond forests. It seems there is more regen in south hammond.

4. Create model Now, because I am dealing with a categorical variable with two levels (Forest) and count data (regen) I will run a chi squared test. first I will fit the data to a matrix.

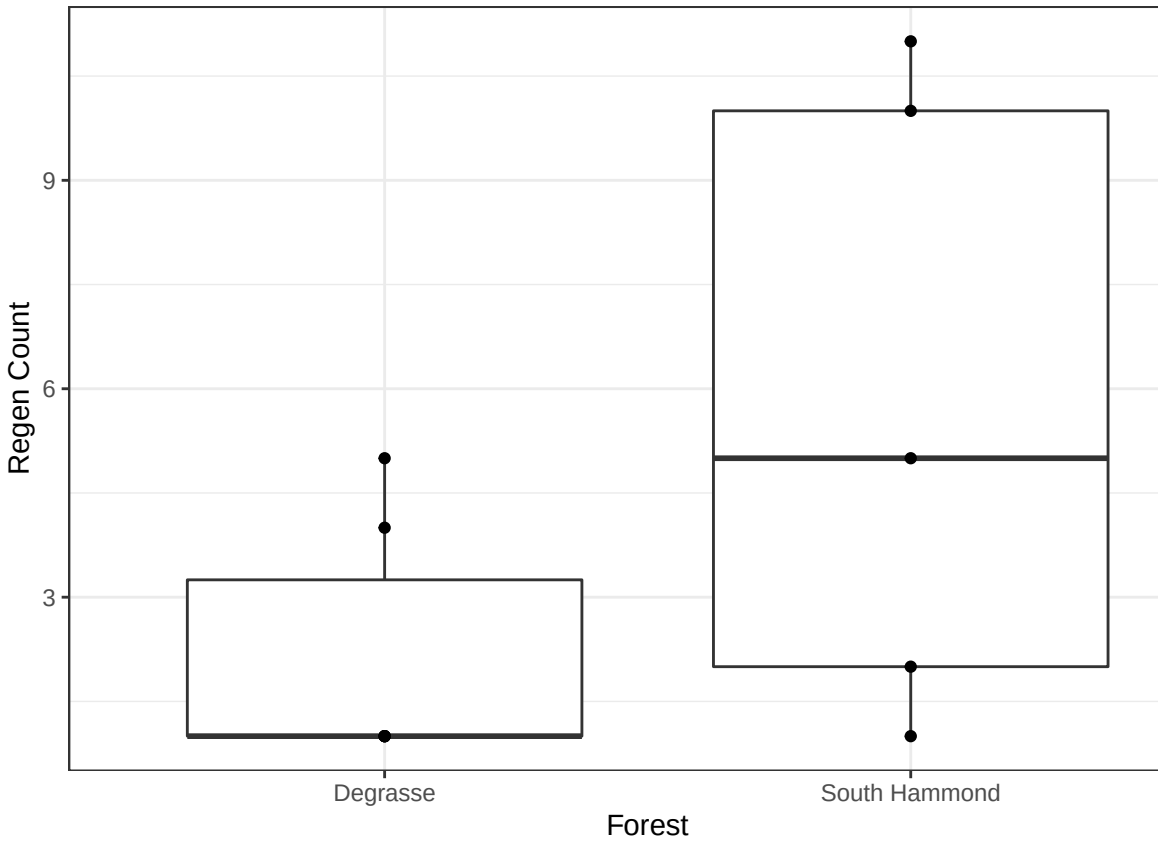
```
## Forest
##      Degrasse South Hammond
##            13            29
```

5. Check assumptions Count data are assumed to have a normal distribution, so I am not going to check any assumptions here.

6. Interpret the model Now I will run the chi squared test

```
##
##  Chi-squared test for given probabilities
##
## data:  regen.mat
## X-squared = 6.0952, df = 1, p-value = 0.01355
```

Because the p-value is relatively small (less than 0.05), I will reject the null hypothesis. It seems like there is a slightly significant difference between the amount of regen in South Hammond and Degrasse, meaning the forest has an impact on the amount of regen present.



7. Publication plot

Biological Summary

It was interesting to find that canopy density did not have a statistically significant impact on regeneration count in the forests. I would like to do more analysis with canopy density and regeneration in the future, possibly using the CanopyApp method rather then densiometer method in my analysis. From a biological standpoint, it seems like light (which is directly correlated with canopy density) would impact regeneration.

Since I found that as canopy density increased, percent bare ground also increased, it seems like there could also be a correlation between canopy density and regeneration. If there is more bare ground at a site, then it seems like there would be less regeneration.

I found that the amount of ferns was statistically different between Kip and Donnerville. Because ferns are an indicator of forest health, it would be interesting to compare the amount of ferns at other forests as well. Comparing ferns to other metrics (maybe canopy density or soil data) could also be interesting.

There was a statistically significant difference in the amount of regeneration between Degrasse and South Hammond. Considering possible biological reasons behind that finding, it could be interesting to do an analysis that looks at the tree species in both of these forests. I was initially expecting Degrasse to have more regen because it is coniferous, and I was thinking the absence of broad leaves would allow more light to reach the canopy floor. Instead, I found South Hammond had more regen. However, in another test I found that canopy density and regen were not statistically correlated, so perhaps there are other factors going on between Degrasse and South Hammond other than canopy density/light availability that could explain the difference in regen I found between these two forests.

Challenges

One of the challenges I encountered with this project was figuring out how to compare variables from two different data frames. I ended up using `full_join()` to merge the dataframes which I have done in the past but had forgotten how to do. I think `full_join()` was a really useful skill to refresh.

Another challenge I faced was figuring out which families to use for my generalized linear model statistical tests. I have not had much stats, so the concept of families was new to me. However, after some time on stack overflow and with help from GSWR and Erika, I figured out which families were the best fits.