Dispersal Plasticity: Diagnostics

Jackson Anderson

```
####  Packages ####
#+ message = FALSE, warning = FALSE
library(tidyverse)
```

```
## -- Attaching packages ----------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.1
## v tidyr   1.1.1     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts -------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(nlme)
```

```
##
## Attaching package: 'nlme'
```

```
## The following object is masked from 'package:dplyr':
##
##     collapse
```

```
library(lattice)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(knitr)
library("grid")
library("gridExtra")
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

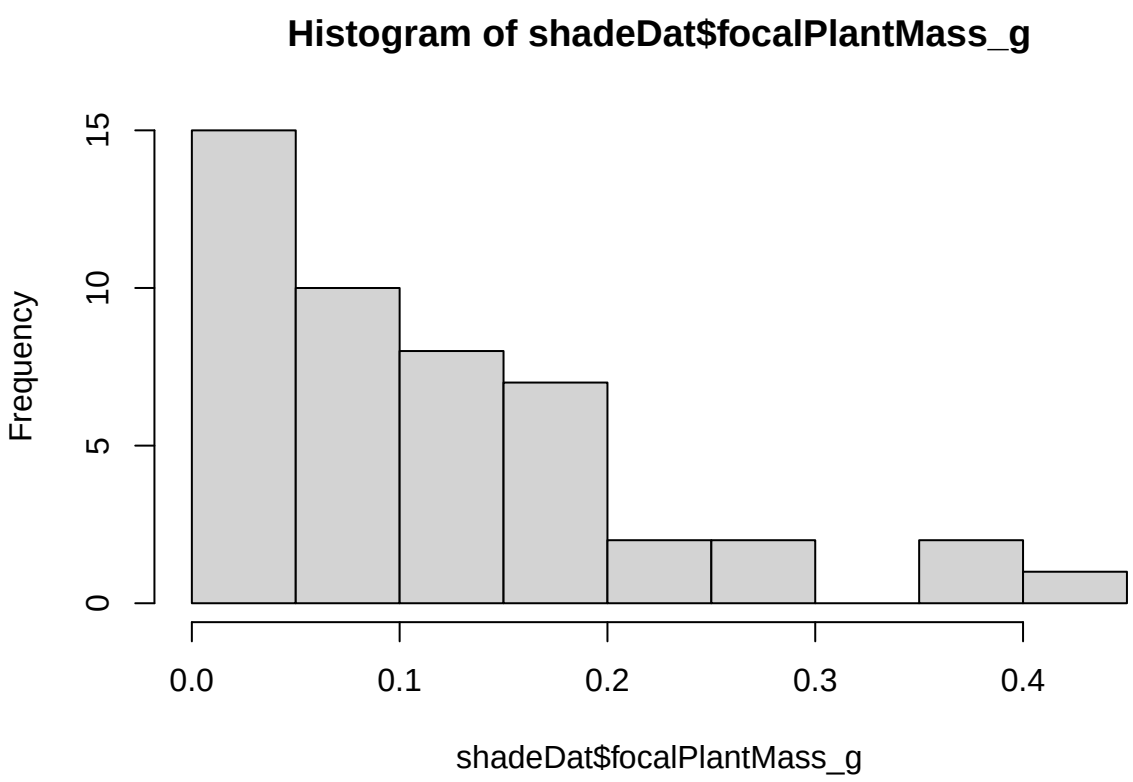
```
library(wesanderson)
```

```r
dat <- read_csv("~/Repos/EmeryLab/DispersalMasterData.csv")
```

```
## Parsed with column specification:
## cols(
##   gitNum = col_double(),
##   experiment = col_character(),
##   plantNum = col_character(),
##   survivorship = col_double(),
##   bin = col_double(),
##   treatmentActual = col_double(),
##   treatmentCat = col_character(),
##   phyllaryCount = col_double(),
##   raySeedCount = col_double(),
##   diskSeedCount1 = col_double(),
##   diskSeedCount2 = col_double(),
##   focalFlowerHeight_cm = col_double(),
##   focalPlantMass_g = col_double(),
##   comments = col_character()
## )
```

```r
#'Writing conditional to check for NA's in either disk seed count column and filtering accordingly
d2NA <- which(is.na(dat$diskSeedCount2) & !is.na(dat$diskSeedCount1))
d1NA <- which(!is.na(dat$diskSeedCount2) & is.na(dat$diskSeedCount1))
view(dat[d2NA,]) #no cases where d1 is NA and d2 is not, so subset the master data where disk seed coun



#'#### Data Cleaning
#'
#'Getting Mean disk seed count
dat$meanDiskCount <-((dat$diskSeedCount1) + (dat$diskSeedCount2))/2

#replacing observations where disk count 2 was an NA with the value of disk count 1 for meanDiskCount
dat[d2NA,"meanDiskCount"] <- dat$diskSeedCount1[d2NA]
#'Ratio of ray seeds to disk seeds
dat$phyllToTotal <- round(dat$phyllaryCount / (dat$phyllaryCount + dat$meanDiskCount), 3)
```

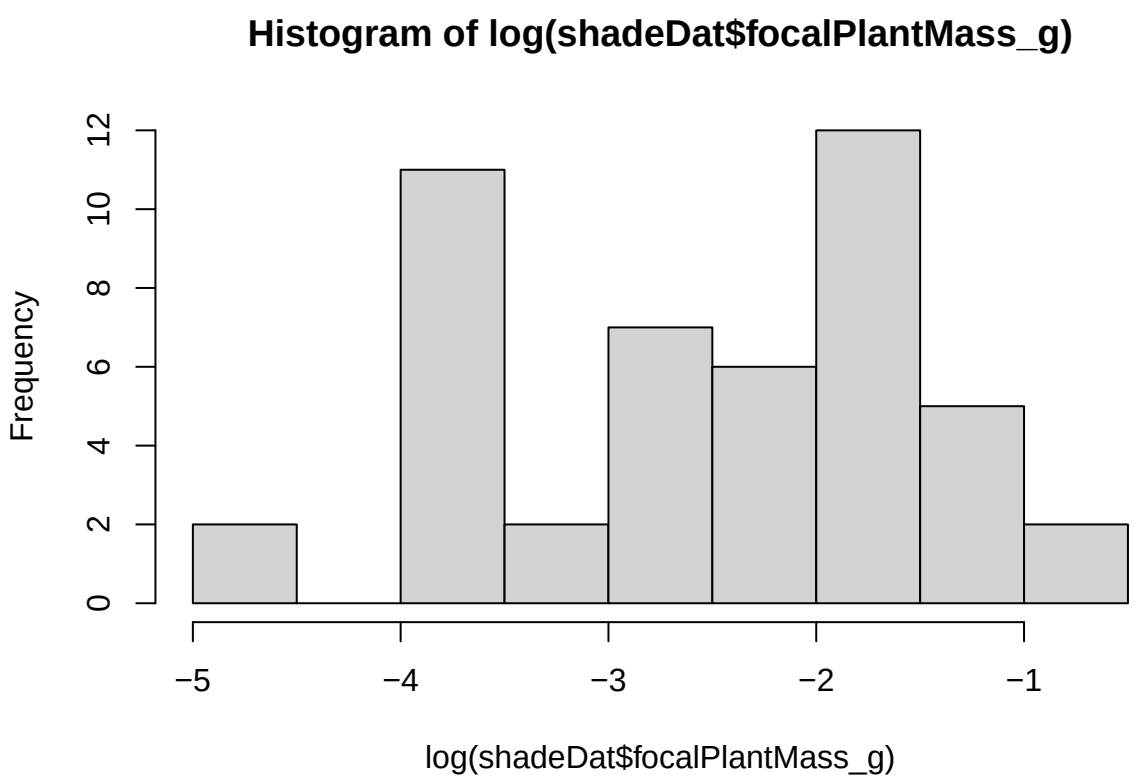
```r
#### Data Cleaning/Diagnostics: Shade Experiment ####
#'

#' #### Data
shadeDat <- dat %>%
  filter( experiment == "shade")

#' Conditional filtering
shadeDat <- shadeDat %>%
  filter(!is.na(phyllToTotal)) %>%
  mutate_at(vars(treatmentCat), factor) %>%
  filter(!is.na(treatmentCat))

#Response variable diagnostics: plant mass
hist(shadeDat$focalPlantMass_g) #highly skewed to the left
```




```
shapiro.test(shadeDat$focalPlantMass_g)# significantly different from normal distribution
```

```
##
##  Shapiro-Wilk normality test
##
## data:  shadeDat$focalPlantMass_g
## W = 0.86413, p-value = 6.209e-05
```

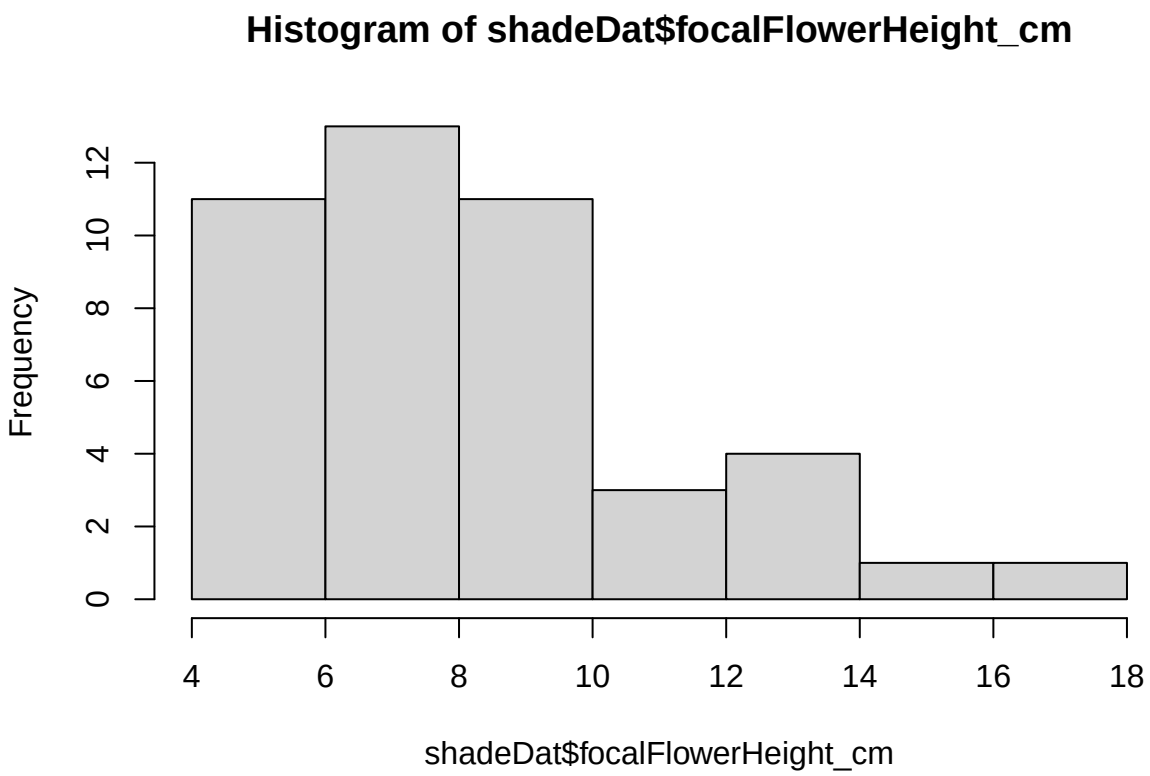
```
#log transformation
hist(log(shadeDat$focalPlantMass_g))#looks marginally better
```




```
shapiro.test(log(shadeDat$focalPlantMass_g))# similarity is marginal; dont feel entirely comfortable us
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(shadeDat$focalPlantMass_g)
## W = 0.95861, p-value = 0.09477
```

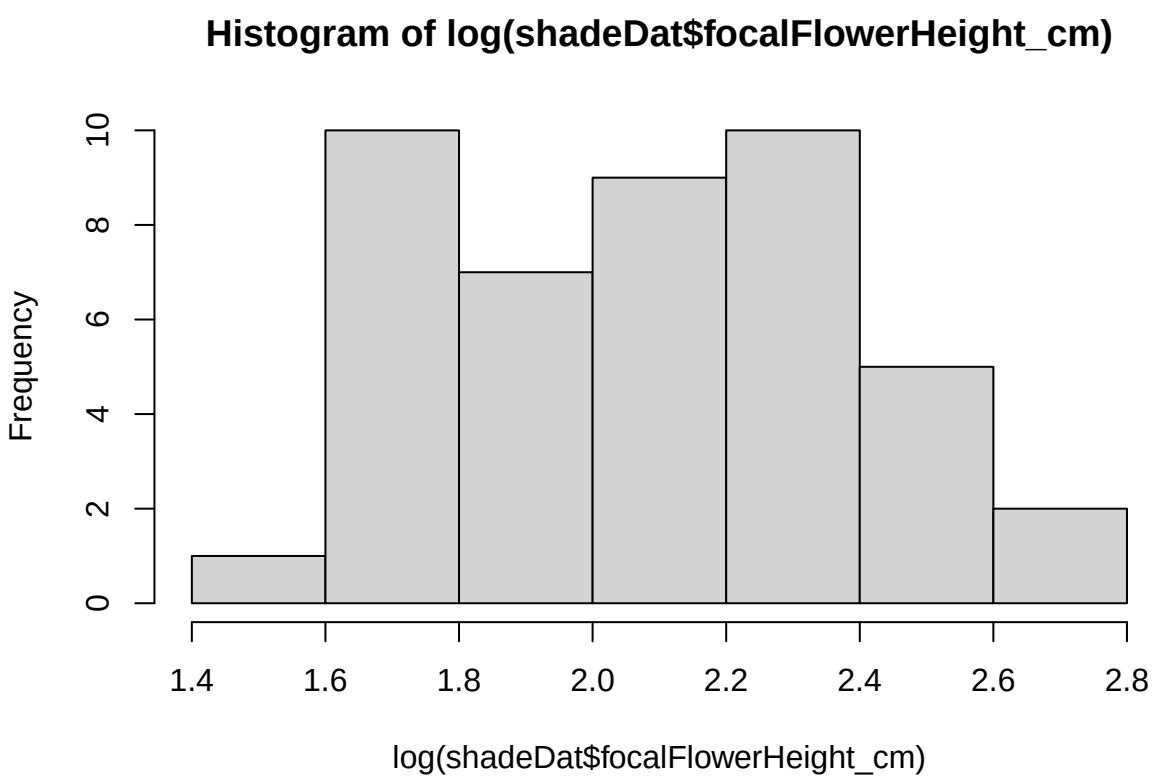
```
#Response variable diagnostics: flower height
hist(shadeDat$focalFlowerHeight_cm) #highly skewed to the left
```




```
shapiro.test(shadeDat$focalFlowerHeight_cm)# significantly different from normal dist
```

```
##
##  Shapiro-Wilk normality test
##
## data:  shadeDat$focalFlowerHeight_cm
## W = 0.93162, p-value = 0.01189
```

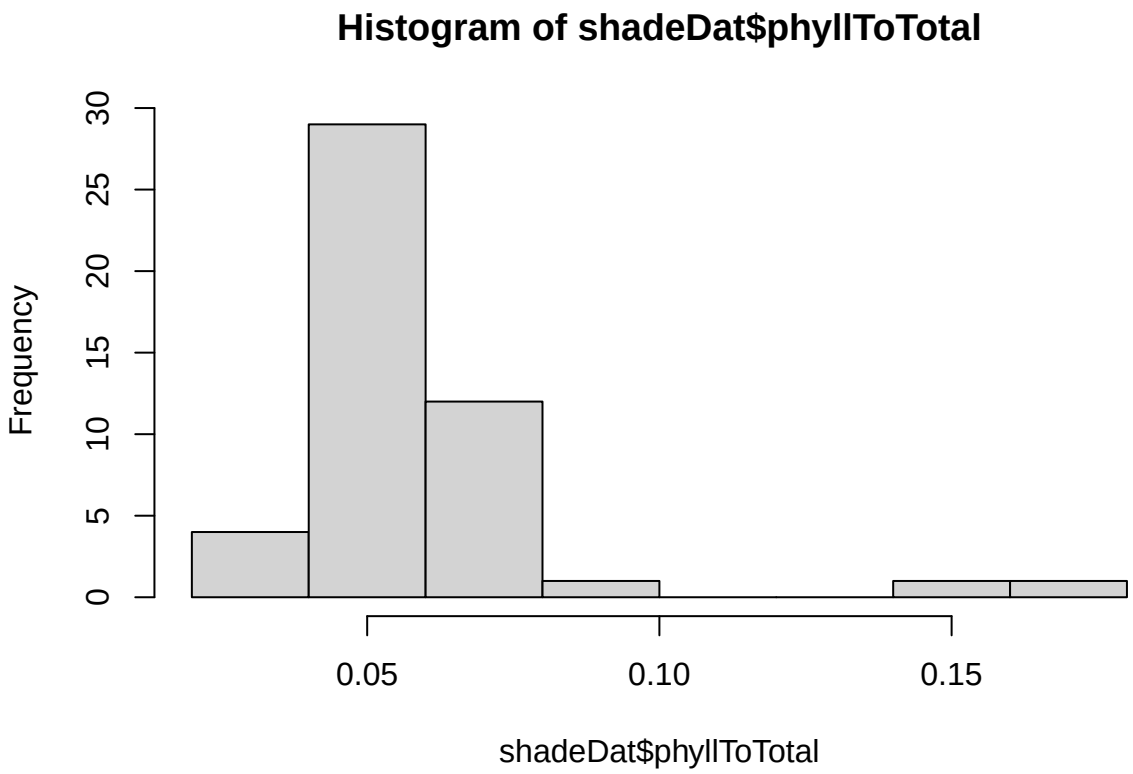
```
#log transformation
hist(log(shadeDat$focalFlowerHeight_cm))#looks a little better
```




```
shapiro.test(log(shadeDat$focalFlowerHeight_cm))#not significantly different from normal dist
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(shadeDat$focalFlowerHeight_cm)
## W = 0.9667, p-value = 0.2303
```

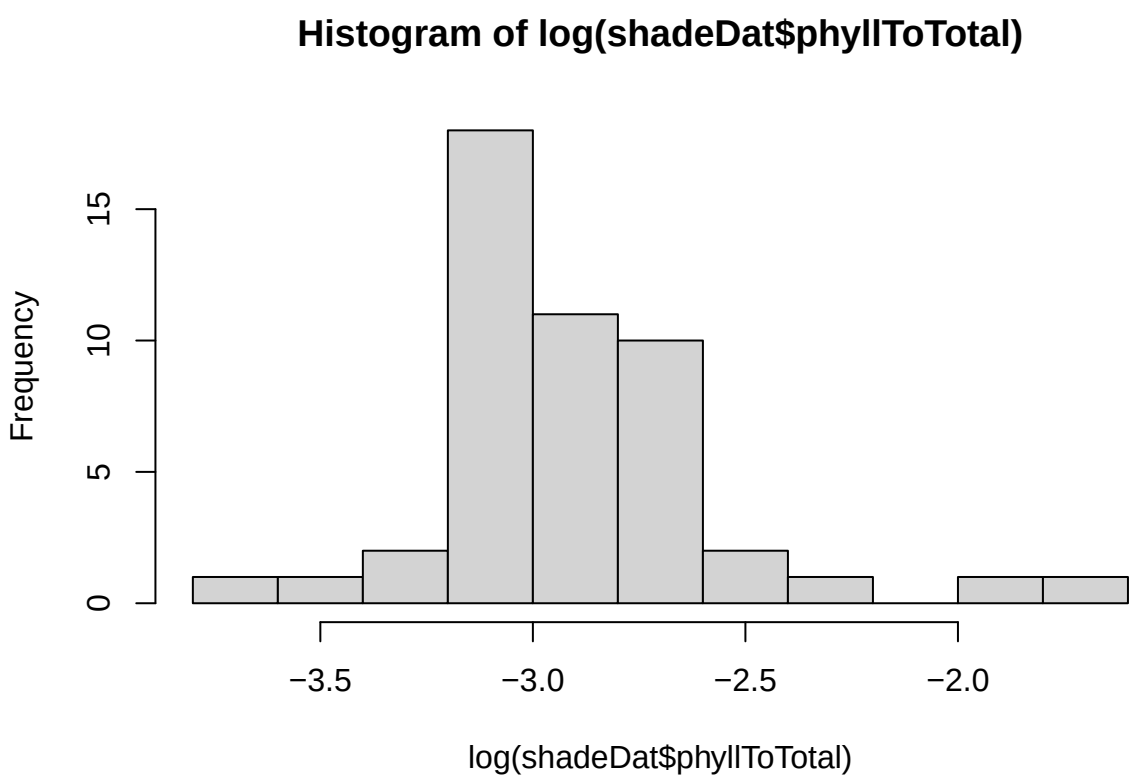
```
#Response variable diagnostics: seed proportion
hist(shadeDat$phyllToTotal) #highly skewed to the left
```




```
shapiro.test(shadeDat$phyllToTotal)# significantly different from normal dist
```

```
##
##  Shapiro-Wilk normality test
##
## data:  shadeDat$phyllToTotal
## W = 0.67632, p-value = 5.188e-09
```

```
#log transformation
hist(log(shadeDat$phyllToTotal))#looks a little better, but not great by any means
```

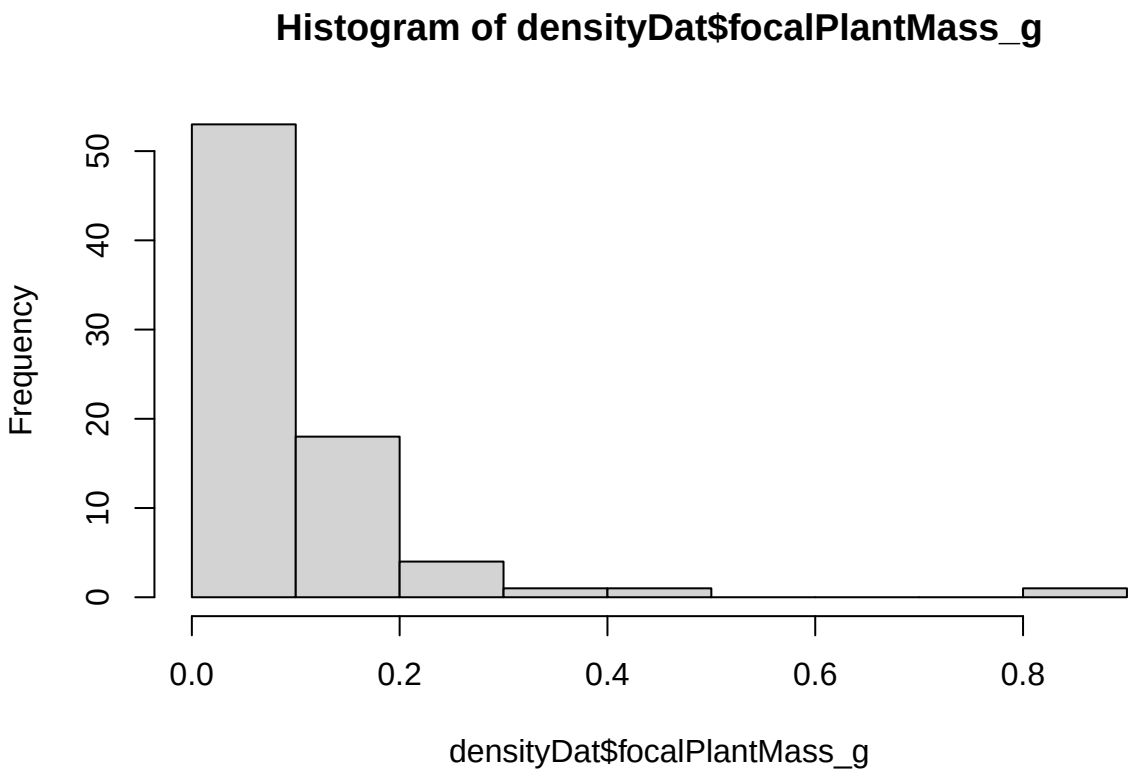



```
shapiro.test(log(shadeDat$phyllToTotal))#still significantly different from norm. dist
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(shadeDat$phyllToTotal)
## W = 0.89558, p-value = 0.0004542
```

Overall, the only response variable that withstood the requirements of normalcy was that of focal flower height, though it barely squeeked by

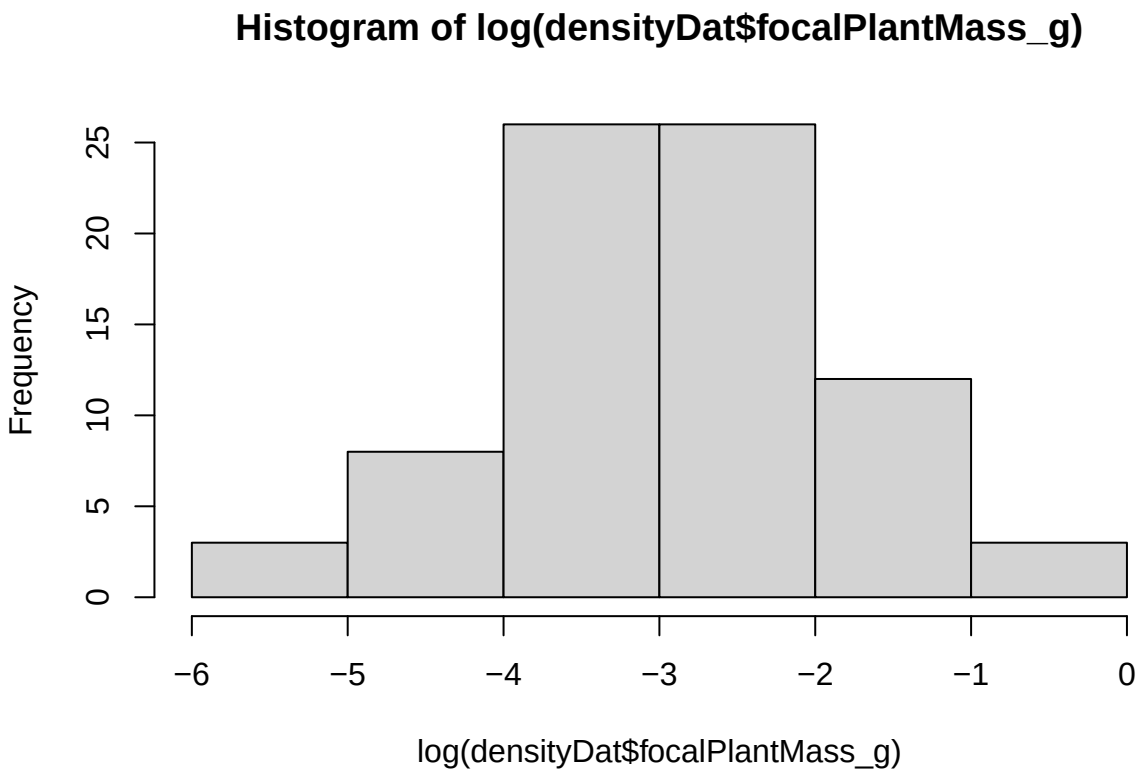
```
#### Data Cleaning/Diagnostics: Density Experiment ####
#'
#' #### Data
densityDat <- dat %>%
  filter( experiment == "density")

#' Conditional Filtering

densityDat <- densityDat %>%
  filter( !is.na(phyllToTotal)) %>%
  mutate_at(vars(treatmentCat), factor) %>%
  filter(!is.na(treatmentCat)) %>%
  filter(survivorship == 1)

#Response variable diagnostics: plant mass
hist(densityDat$focalPlantMass_g) #highly skewed to the left
```




```
shapiro.test(densityDat$focalPlantMass_g)# significantly different from normal distribution
```

```
##
##  Shapiro-Wilk normality test
##
## data:  densityDat$focalPlantMass_g
## W = 0.62751, p-value = 9.169e-13
```

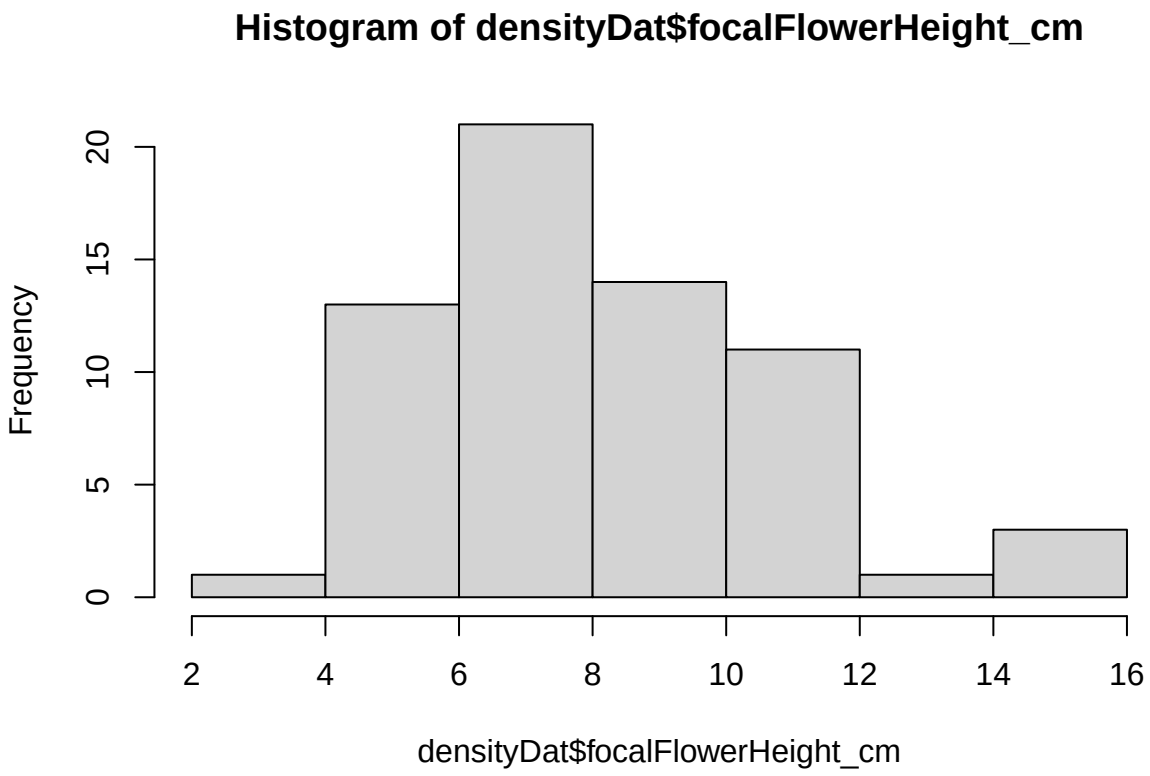
```
#log transformation
hist(log(densityDat$focalPlantMass_g))#looks much better
```




```
shapiro.test(log(densityDat$focalPlantMass_g))# fixed the issue
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(densityDat$focalPlantMass_g)
## W = 0.98842, p-value = 0.7079
```

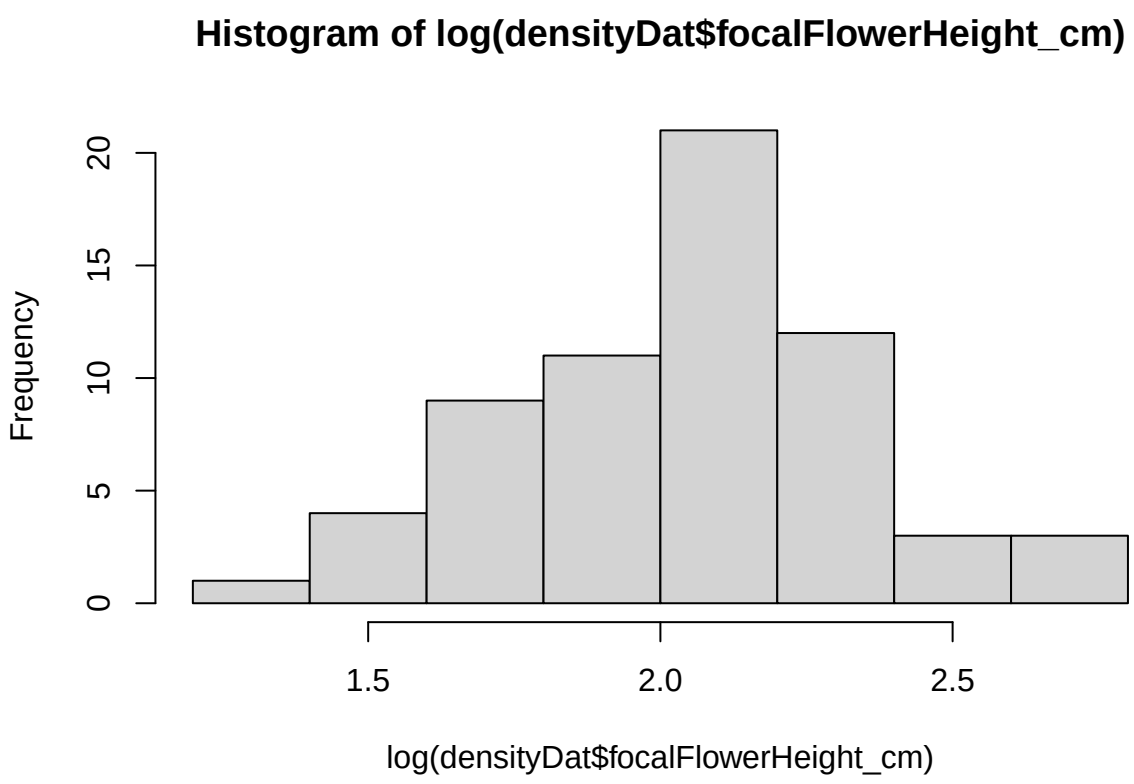
```
#Response variable diagnostics: flower height
hist(densityDat$focalFlowerHeight_cm) #highly skewed to the left
```




```
shapiro.test(densityDat$focalFlowerHeight_cm)# significantly different from normal dist
```

```
##
##  Shapiro-Wilk normality test
##
## data:  densityDat$focalFlowerHeight_cm
## W = 0.95389, p-value = 0.01785
```

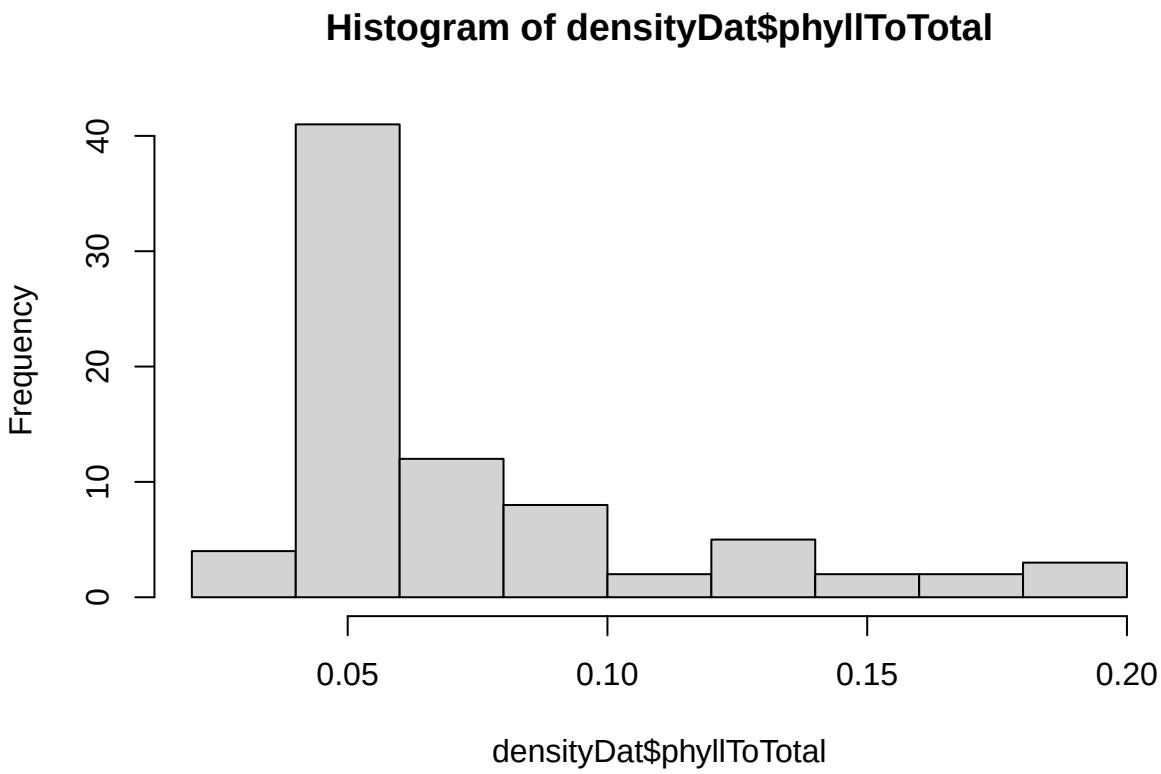
```
#log transformation
hist(log(densityDat$focalFlowerHeight_cm))#looks a little better
```




```
shapiro.test(log(densityDat$focalFlowerHeight_cm))#not significantly different from normal dist; fixed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(densityDat$focalFlowerHeight_cm)
## W = 0.99265, p-value = 0.9695
```

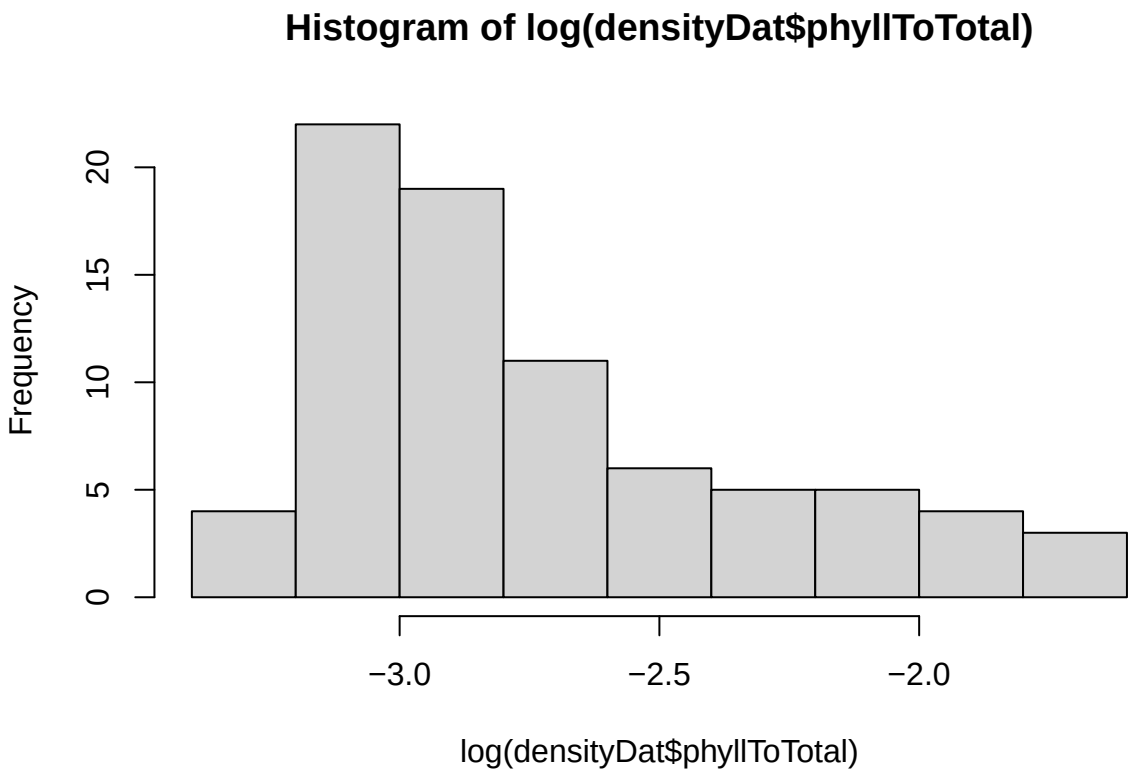
```
#Response variable diagnostics: seed proportion
hist(densityDat$phyllToTotal) #highly skewed to the left
```




```
shapiro.test(densityDat$phyllToTotal)# significantly different from normal dist
```

```
##
##  Shapiro-Wilk normality test
##
## data:  densityDat$phyllToTotal
## W = 0.77936, p-value = 1.534e-09
```

```
#log transformation
hist(log(densityDat$phyllToTotal))#does not look much better
```




```
shapiro.test(log(densityDat$phyllToTotal))#still significantly different from norm. dist
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(densityDat$phyllToTotal)
## W = 0.9022, p-value = 1.689e-05
```

For the density experiment, both the response variable distributions of plant mass/focal flower height became normal when log transformed. For the seed proportion, however, log transformation did not solve the issue

```
#### Data Cleaning/Diagnostics: Resource Experiment ####

#' #### Data
#'
#' Subsetting the data
resourceDat <- dat %>%
  filter(experiment == "resources")

#' Conditional Filtering
resourceDat <- resourceDat %>%
  filter(!is.na(phyllToTotal)) %>%
  filter(!is.na(treatmentCat)) %>%
  filter(survivorship == 1) %>%
  mutate_at(vars(treatmentCat), factor)

#' Replicates per treatment
```

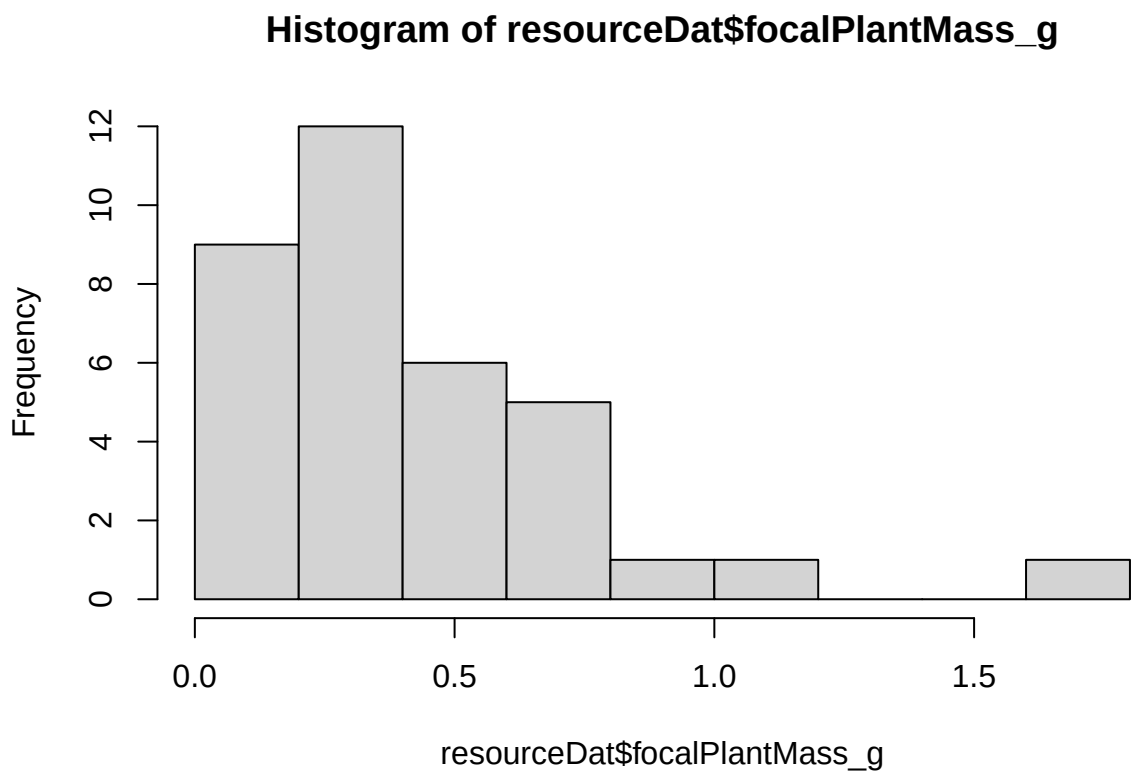
```
resource_reps <- summarize(group_by(resourceDat, treatmentCat), n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
resource_reps
```

```
## # A tibble: 4 x 2
##   treatmentCat `n()`
##   <fct>        <int>
## 1 Control         16
## 2 High             3
## 3 Low             13
## 4 Medium           7
```

```
#' The "high" treatment was removed due to low replication (3 data points)
resourceDat <- resourceDat %>%
  filter( treatmentCat != "High" ) %>%
  droplevels()

#Response variable diagnostics: plant mass
hist(resourceDat$focalPlantMass_g) #highly skewed to the left
```

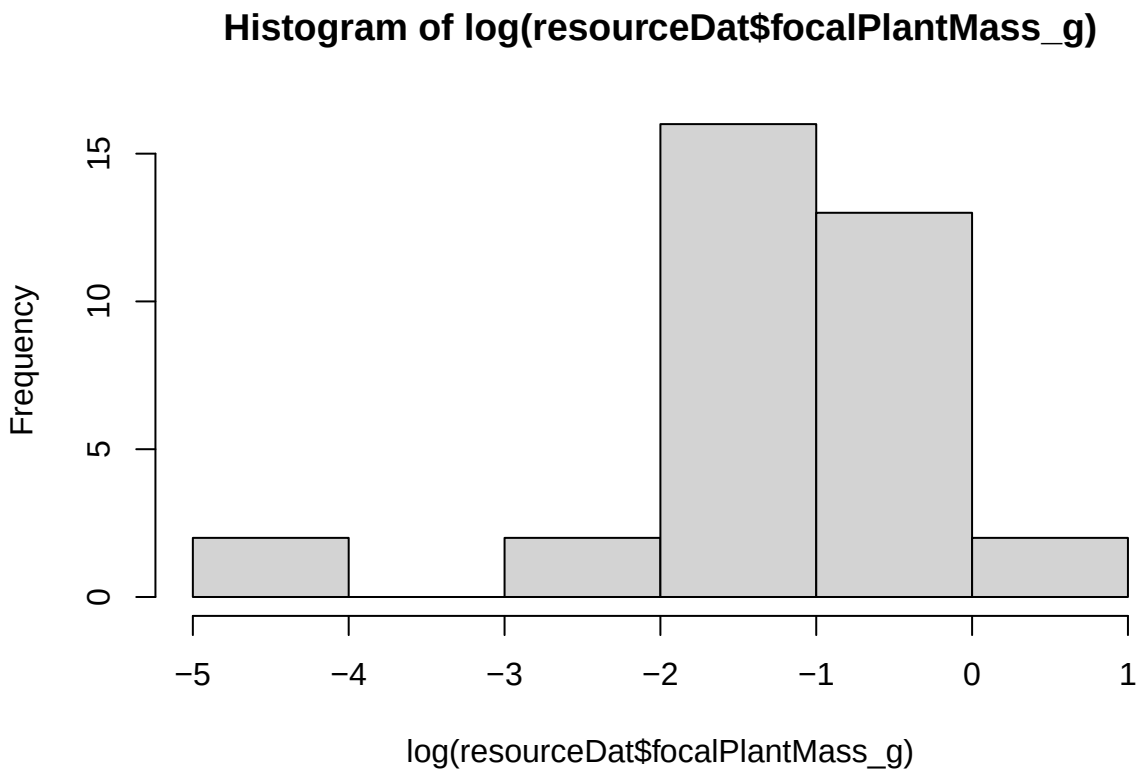



```
shapiro.test(resourceDat$focalPlantMass_g)# significantly different from normal distribution
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resourceDat$focalPlantMass_g
```

```
## W = 0.84846, p-value = 0.0002129
```

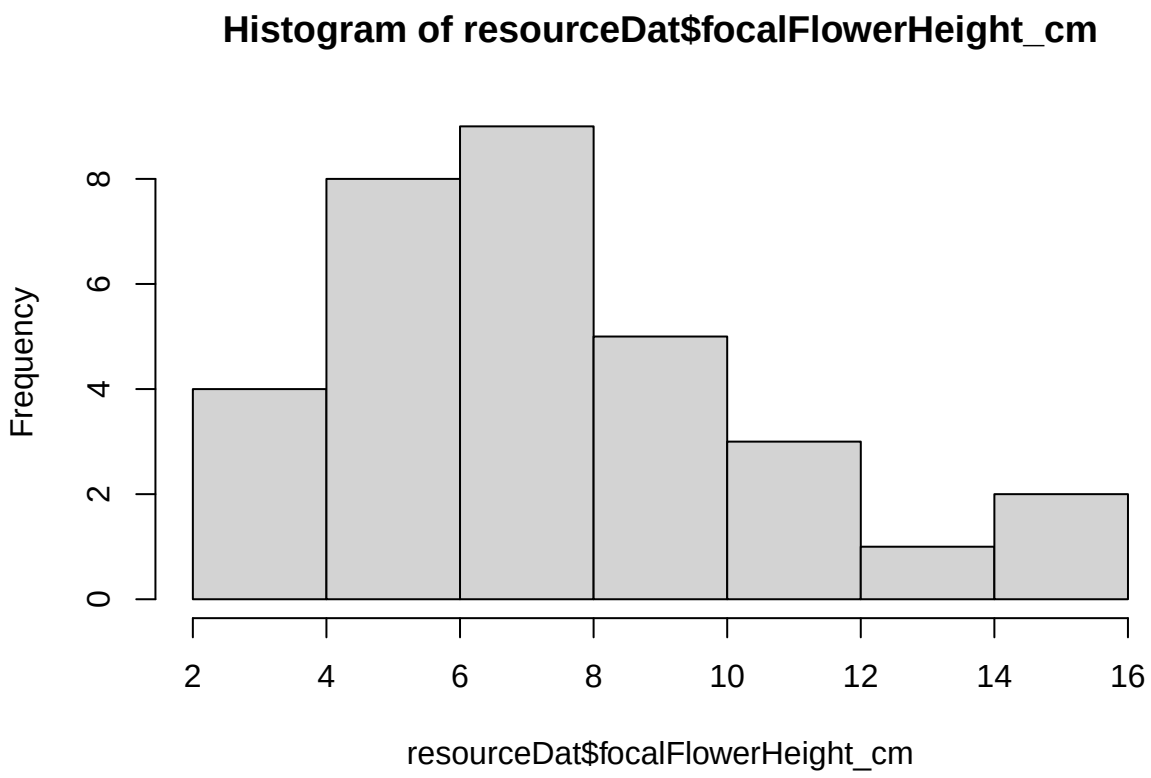
```
#log transformation
hist(log(resourceDat$focalPlantMass_g))#looks a little better; skewed to right now
```




```
shapiro.test(log(resourceDat$focalPlantMass_g))# did not fix the issue; significantly different from no
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(resourceDat$focalPlantMass_g)
## W = 0.84231, p-value = 0.0001551
```

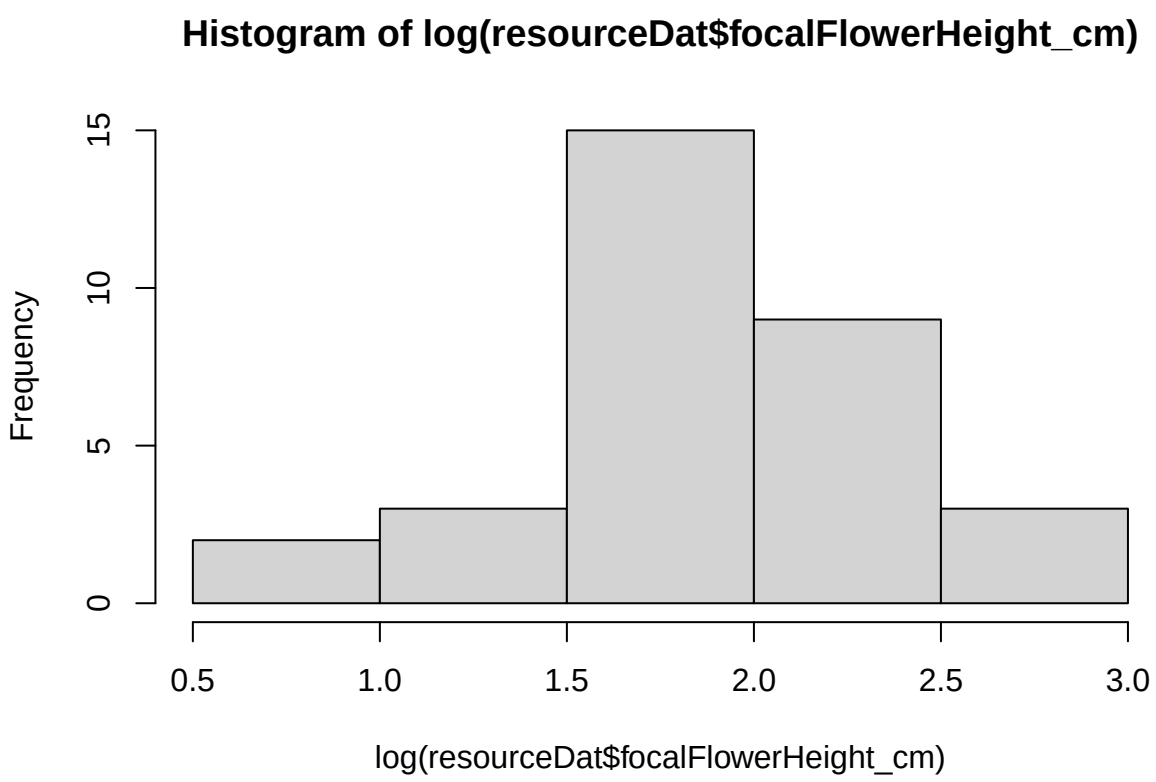
```
#Response variable diagnostics: flower height
hist(resourceDat$focalFlowerHeight_cm) #Slightly skewed to left, but overall seems OK
```




```
shapiro.test(resourceDat$focalFlowerHeight_cm)# marginally similar to normal distribution
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resourceDat$focalFlowerHeight_cm
## W = 0.93693, p-value = 0.0613
```

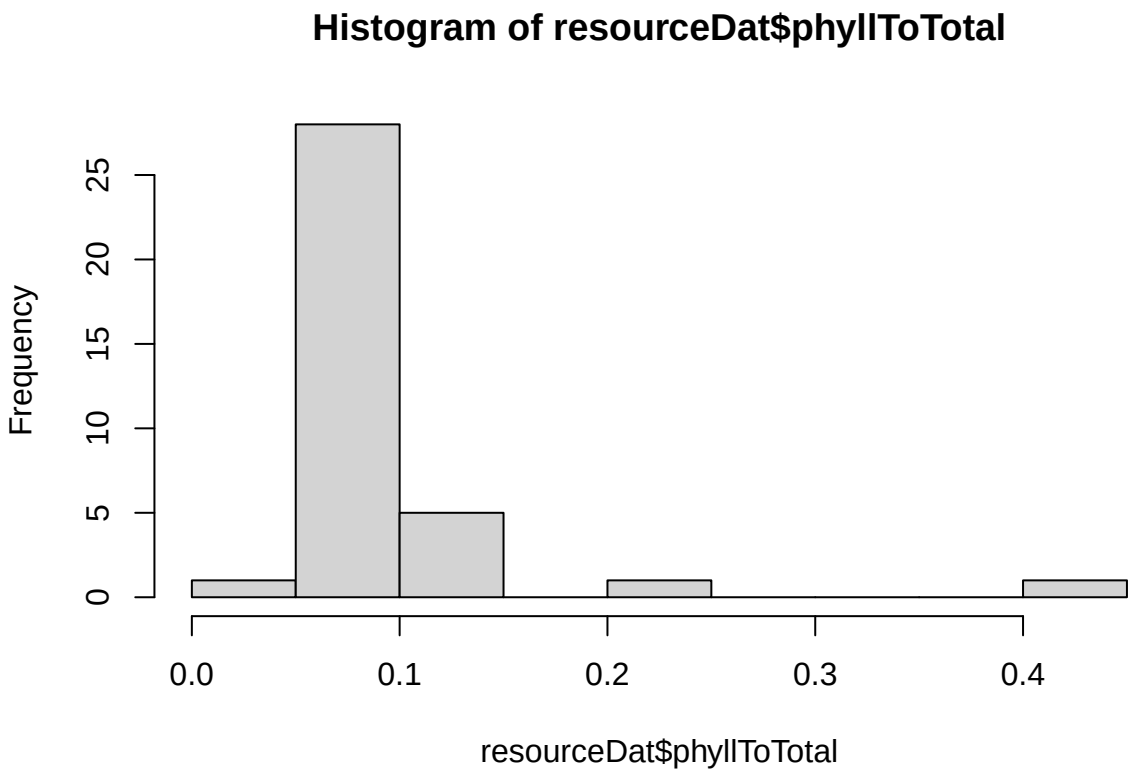
```
#log transformation
hist(log(resourceDat$focalFlowerHeight_cm))#looks a little better
```




```
shapiro.test(log(resourceDat$focalFlowerHeight_cm))#not significantly different from normal dist; fixed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(resourceDat$focalFlowerHeight_cm)
## W = 0.96361, p-value = 0.3437
```

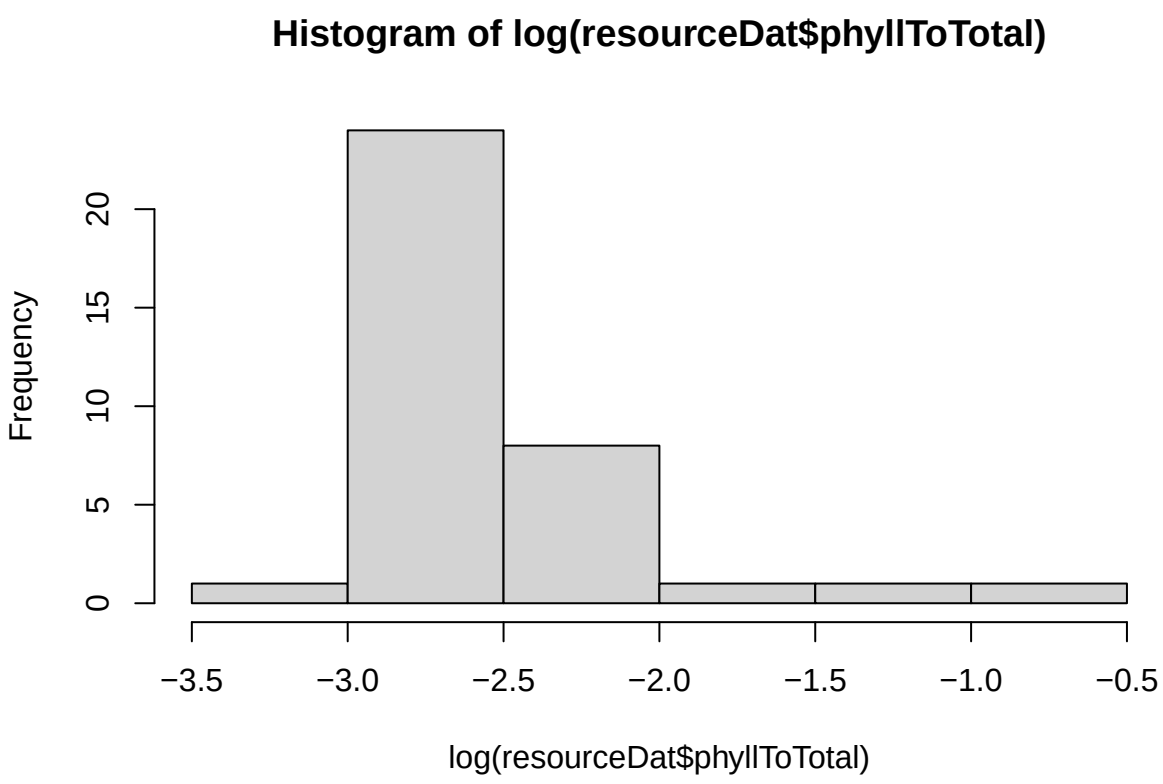
```
#Response variable diagnostics: seed proportion
hist(resourceDat$phyllToTotal) #highly skewed to the left
```




```
shapiro.test(resourceDat$phyllToTotal)# significantly different from normal dist
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resourceDat$phyllToTotal
## W = 0.48898, p-value = 4.79e-10
```

```
#log transformation
hist(log(resourceDat$phyllToTotal))#does not look much better
```




```
shapiro.test(log(resourceDat$phyllToTotal))#still significantly different from norm. dist
```

```
##
##  Shapiro-Wilk normality test
##
## data:  log(resourceDat$phyllToTotal)
## W = 0.78218, p-value = 7.348e-06
```

For the resource experiment, only the focal flower height was fixed when log transformed; the other two response variables remain unusable for standard models

Conclusion: the purpose of this diagnostics script is to demonstrate the fact that real-world data can be very problematic. Even though some of the response variable distributions were fixed by log transformation, many of them were not. Subsequent analyses on these log-transformed variables did not change the fact that there was high heteroskedasticity within these models. As a result, the use of log transformation ultimately failed. To address this issue, models that incorporated unequal residual variance structures were applied with much more success. Even still, those models provided problems of their own, such as the loss of degrees of freedom.

Some thoughts on the biology behind this: in every case, there was a skew to the left. There were platykurtic and leptokurtic distributions; normal too. The gamut of distributions was present! However, the consitency of leftward skew suggested that one of the following things may be true: 1) there was more variation in these variables than was sampled; i.e., even though outliers existed, realistically these data points may not have actually been outliers relative to the true mean of the population. It's a stretch, but it's possible, given that our replication numbers were somewhat low within each treatemt, and very low in a few. Point 3 below somewhat addresses this. 2) Our treatments themselves resulted in such low survivorship (certainly the case in the resource experiment) that our results were not a realistic guage of stress or trait response; maybe our treatments were too stressful. There is still survivorship data that needs to be collected and entered before a survivorship analysis can be completed, but the results of such an analysis will certainly be interesting. 3)

These models do not exclude outliers, as all of the data has yet to be collected. As a result, these outliers may be significantly affecting our response variable distribution.