Homework 2

Jasmine Williamson

2024-10-09

Question 1) Which variables are your response variables? Which are your predictor variables (if relevant)? Are they same-scale or mixed scale? Categorical, continuous, or ordinal?

Response Variable: Salamander total count. My current matrix defines the plot ID as the objects. I have 127 sites with 7 plots each, and 889 rows in the matrix. I am working on a site-level matrix to use for this class, which will define sites as the objects with 127 rows, and will include salamander density per 9^2m plot as response variable.

Predictor Variables: temperature, humidity, soil moisture, elevation, downed wood cover, canopy cover, veg cover, fine woody debris cover. These variables are mixed scale. The first four are continuous, and the last four are percent cover categories numbered 1-4.

```
## 'data.frame':    127 obs. of  2 variables:
##  $ oss : int  0 0 0 0 0 8 3 0 0 6 ...
##  $ enes: int  0 0 0 0 1 0 1 0 0 3 ...
```

```
## 'data.frame':    127 obs. of  16 variables:
##  $ site_id   : chr  "10024 _ 1 _ 2023" "10024 _ 1 _ 2024" "10078 _ 1 _ 2023" "10078 _ 1 _ 2024" ...
##  $ landowner : Factor w/ 4 levels "BLM","ODF","PB",..: 4 4 4 4 4 4 4 4 4 1 ...
##  $ tree_farm : Factor w/ 3 levels "CL","NC","SP": 1 1 1 1 1 1 1 1 1 1 ...
##  $ stand     : int  10024 10024 10078 10078 10153 10153 10185 10185 10258 10302 ...
##  $ trt       : Factor w/ 5 levels "BS","BU","HB",..: 1 1 3 3 1 1 1 1 1 5 ...
##  $ year      : int  2023 2024 2023 2024 2023 2024 2023 2024 2023 2023 ...
##  $ jul_date  : num  142 86 74 94 124 129 100 142 79 121 ...
##  $ weather   : Factor w/ 6 levels "C","PC","R","S",..: 1 2 2 1 3 1 3 1 1 1 ...
##  $ elev      : num  850 834 1852 1923 2717 ...
##  $ temp      : num  56.9 53.8 37.4 43.4 45.8 46.8 45.3 56.9 44 42.7 ...
##  $ hum       : num  63.4 72.3 89.8 83.2 92.7 96 91.1 69.2 91.9 85.7 ...
##  $ canopy_cov: num  0 0 0 0 0 0 0 0 0 3.9 ...
##  $ veg_cov   : num  3.1 3.1 3.4 3.4 2.4 3.7 2.7 3.4 1.6 1.3 ...
##  $ dwd_cov   : num  1.7 1.6 1 1.4 1.4 2.3 1.4 2.3 1.3 1.3 ...
##  $ fwd_cov   : num  2.9 2.4 1 1.1 1.7 3.3 2 2 2.6 2.9 ...
##  $ soil_moist: num  13.8 28.2 30.1 26.1 25.7 ...
```

Question 2) Do you have missing values in your data? If so, how will you account for them? Will you need to use different methods for different variables?

I have missing values for observer, but that is a data entry error that I need to fix. Otherwise I do not have any missing data.

```r
na_count <- colSums(is.na(dat))
print(na_count)
```

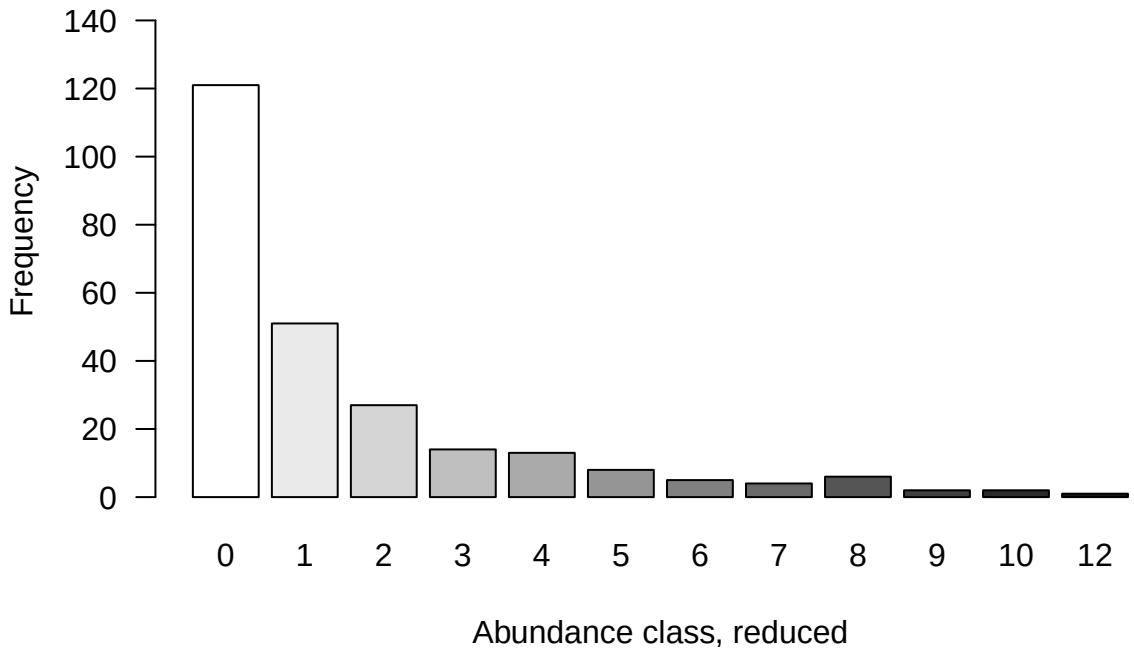
```
##    site_id  landowner  tree_farm      stand        trt       year   jul_date
##          0          0          0          0          0          0          0
##        lat       long    weather       elev       temp        hum canopy_cov
##          0          0          0          0          0          0          0
##    veg_cov    dwd_cov    fwd_cov soil_moist        oss       enes
##          0          0          0          0          0          0
```

Question 3) Is there a need for data transformation? If so, what transformations are you considering and why? Is your decision based on statistical or ecological criteria, or both?

Looking at the salamander data, four of my six species are present in very low numbers,so I am going to remove them from the dataset. OSS and ENES were the target species, so it is unsurprising that the other species were rarely found based on the types of habitat searched.

The salamander data is zero-skewed, so I should transform it. If I drop the species that include non-zero values in less that 5% of the surveys, the distribution looks a little better, but we still are very right-skewed and zero-skewed. The best option for highly right-skewed and zero-skewed data is to use a log +1 (or plus some other relevant value).

```
## ANFE TAGR PLDU AMGR ENES  OSS
##    1    3    4    5  107  163
```



Question 4) Is there a need for data standardization? What standardizations will you use? Is your decision based on statistical or ecological criteria, or both?

Salamander data: I surveyed identical numbers of plots for the same amount of time for each site, so my salamander data does not need to be standardized by survey effort.

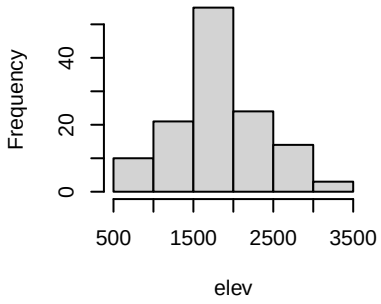
Env data: Most of my data is on similar scales, for example percent humidity and temperature in F are similar in scale. Elevation consists of larger values than the rest, so that one might benefit from being transformed. Most of this data is normally distributed. humidity is a little weird, but I'll probably drop it later anyways.

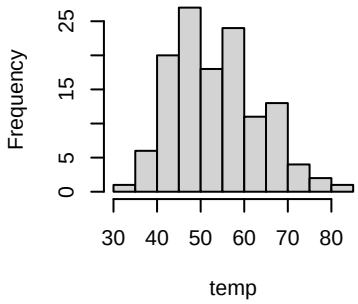
The coefficient of variation value (cv) is < 50, so apparently standardization won't make a difference? Am i interpreting that correctly? Since cv values are very low for both the species and environmental data sets, I wont standardize any of it for now.

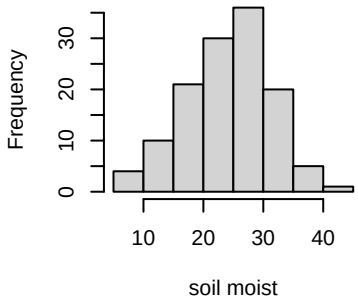
```
##                     oss        enes
## nbr.val      127.0000000 127.0000000
## nbr.null      50.0000000  71.0000000
## nbr.na         0.0000000   0.0000000
## min            0.0000000   0.0000000
## max           10.0000000  12.0000000
## range         10.0000000  12.0000000
## sum          257.0000000 138.0000000
## median         1.0000000   0.0000000
## mean           2.0236220   1.0866142
## SE.mean        0.2311415   0.1646954
## CI.mean.0.95   0.4574222   0.3259273
## var            6.7851519   3.4448194
## std.dev        2.6048324   1.8560225
## coef.var       1.2872129   1.7080786
```

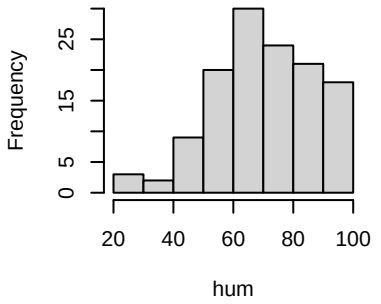
```
##          site_id landowner tree_farm        stand trt         year     jul_date
## nbr.val       NA        NA        NA 1.270000e+02  NA 1.270000e+02 1.270000e+02
## nbr.null      NA        NA        NA 0.000000e+00  NA 0.000000e+00 0.000000e+00
## nbr.na        NA        NA        NA 0.000000e+00  NA 0.000000e+00 0.000000e+00
## min           NA        NA        NA 9.757000e+03  NA 2.023000e+03 7.200000e+01
## max           NA        NA        NA 4.081800e+05  NA 2.024000e+03 1.530000e+02
## range         NA        NA        NA 3.984230e+05  NA 1.000000e+00 8.100000e+01
## sum           NA        NA        NA 6.437737e+06  NA 2.569810e+05 1.431000e+04
## median        NA        NA        NA 1.242100e+04  NA 2.023000e+03 1.140000e+02
## mean          NA        NA        NA 5.069084e+04  NA 2.023472e+03 1.126772e+02
## SE.mean       NA        NA        NA 9.833234e+03  NA 4.447583e-02 1.895728e+00
## CI.mean       NA        NA        NA 1.945968e+04  NA 8.801635e-02 3.751591e+00
## var           NA        NA        NA 1.227995e+10  NA 2.512186e-01 4.564108e+02
## std.dev       NA        NA        NA 1.108149e+05  NA 5.012171e-01 2.136377e+01
## coef.var      NA        NA        NA 2.186094e+00  NA 2.477015e-04 1.896016e-01
##          weather         elev         temp          hum  canopy_cov
## nbr.val       NA 1.270000e+02  127.0000000  127.0000000 127.0000000
## nbr.null      NA 0.000000e+00    0.0000000    0.0000000  75.0000000
## nbr.na        NA 0.000000e+00    0.0000000    0.0000000   0.0000000
## min           NA 6.310000e+02   33.1000000   23.6000000   0.0000000
## max           NA 3.261140e+03   81.4000000  100.0000000   3.9000000
## range         NA 2.630140e+03   48.3000000   76.4000000   3.9000000
## sum           NA 2.319362e+05 6805.2000000 8977.9000000 117.4000000
## median        NA 1.801000e+03   52.5000000   69.9000000   0.0000000
```

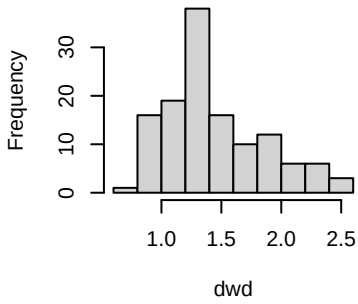
```
## mean          NA 1.826269e+03   53.5842520   70.6921260   0.9244094
## SE.mean       NA 4.956824e+01    0.8743262    1.5298005   0.1126191
## CI.mean       NA 9.809408e+01    1.7302658    3.0274302   0.2228699
## var           NA 3.120403e+05   97.0846707  297.2167629   1.6107487
## std.dev       NA 5.586057e+02    9.8531554   17.2399757   1.2691527
## coef.var      NA 3.058725e-01    0.1838816    0.2438741   1.3729335
##                veg_cov      dwd_cov      fwd_cov   soil_moist
## nbr.val  127.00000000 127.00000000 127.00000000  127.0000000
## nbr.null   0.00000000   0.00000000   0.00000000    0.0000000
## nbr.na     0.00000000   0.00000000   0.00000000    0.0000000
## min        0.90000000   0.70000000   0.90000000    6.2000000
## max        4.00000000   2.60000000   4.00000000   41.4000000
## range      3.10000000   1.90000000   3.10000000   35.2000000
## sum      374.10000000 189.10000000 264.60000000 3075.9100000
## median     3.40000000   1.40000000   2.00000000   24.8400000
## mean       2.94566929   1.48897638   2.08346457   24.2197638
## SE.mean    0.08856475   0.03669288   0.07205245    0.6284900
## CI.mean    0.17526703   0.07261412   0.14258968    1.2437632
## var        0.99615173   0.17098863   0.65932758   50.1649531
## std.dev    0.99807401   0.41350771   0.81198989    7.0827222
## coef.var   0.33882758   0.27771274   0.38973060    0.2924356
```





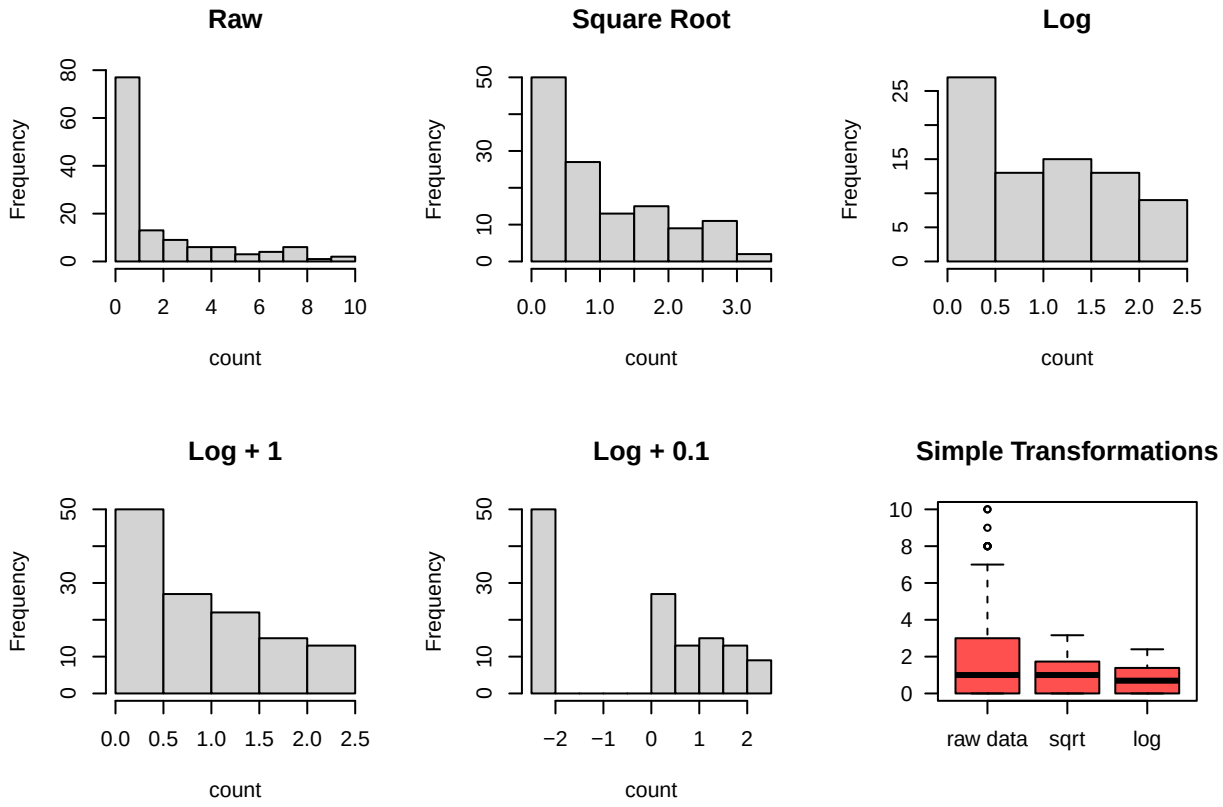




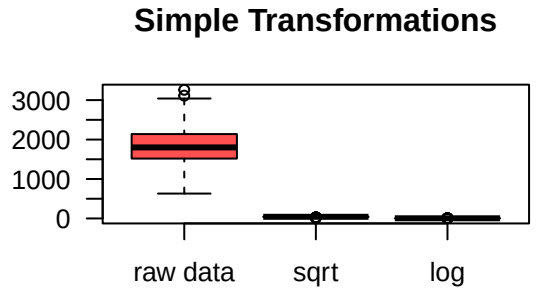


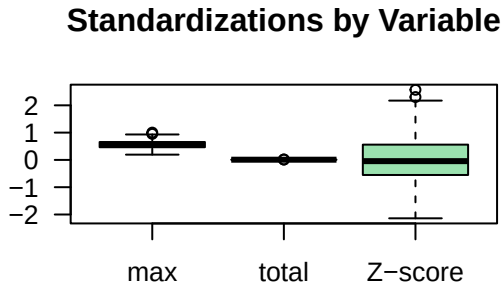
Question 5) Considering the histograms of the data, how effective do you think your transformation/standardization is?

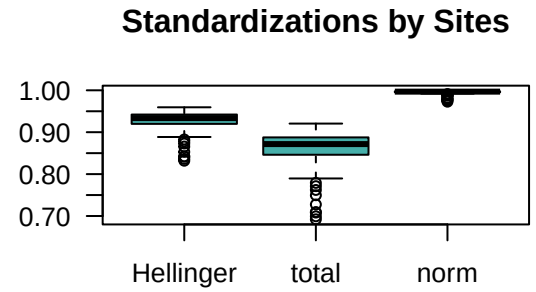
I think the log+1 histogram looks the best and is the transformation I should move forward with. The square root also does not look bad.

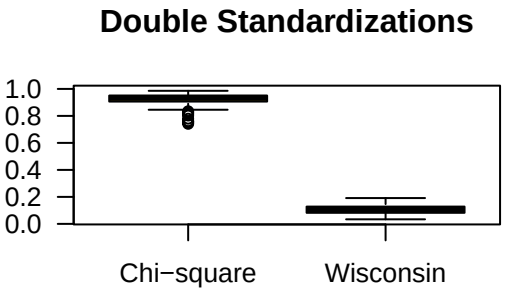


I ran some options with elevation since the scale is the farthest from the rest of my variables. I think z-scoring it is a good option, but might not be necessary since the cv value is so low? Unsure about this.



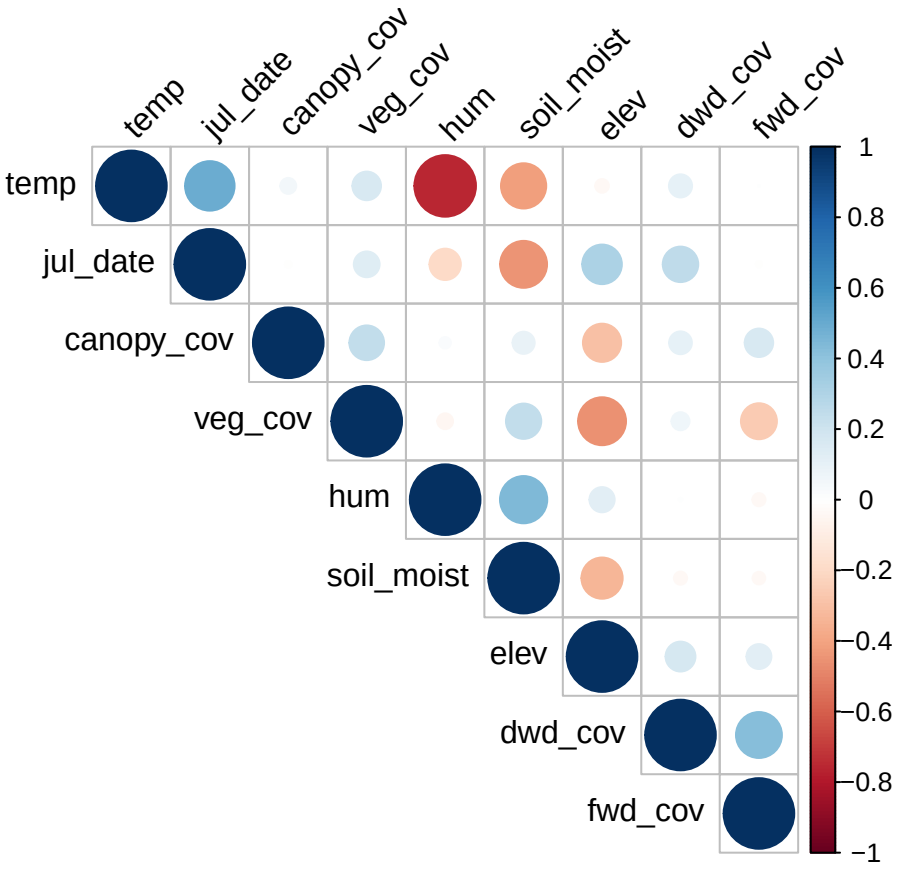






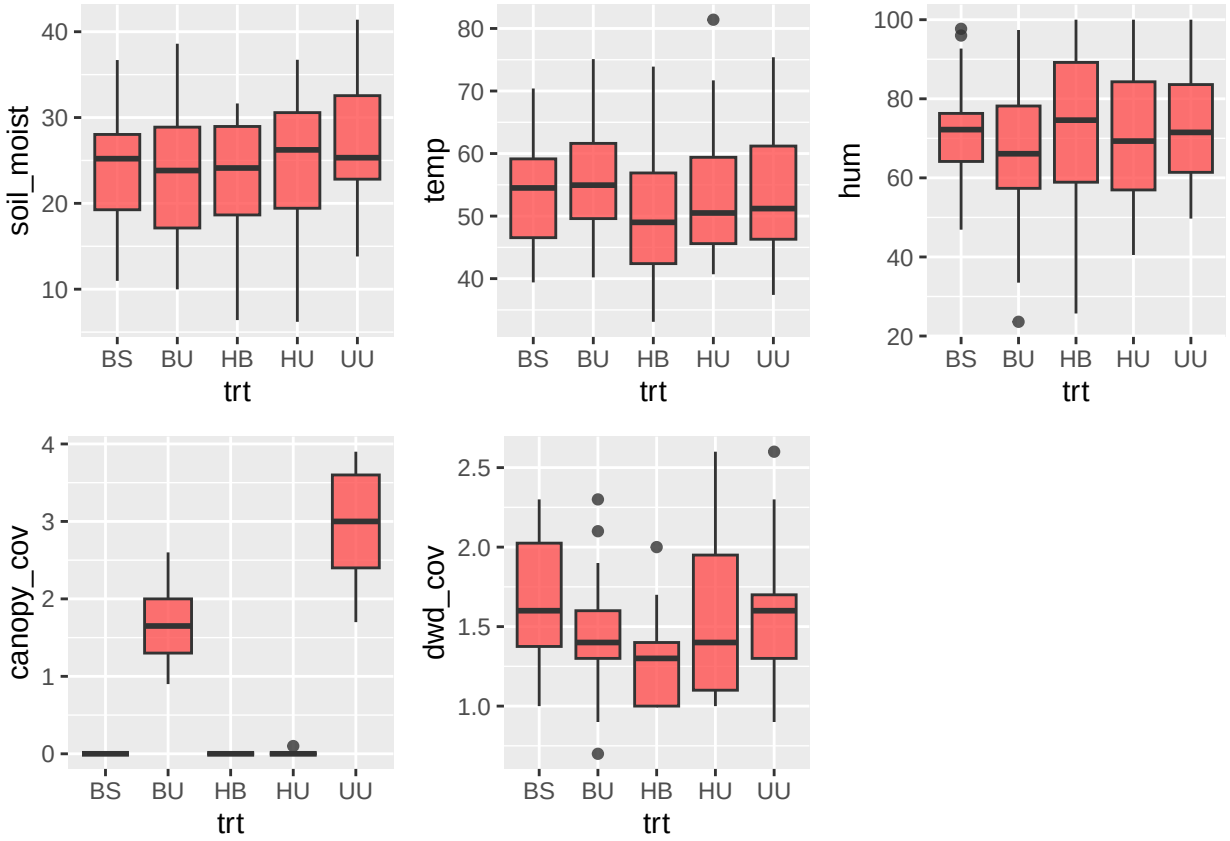
Question 6) If you are working with environmental predictors in your data, do any of them covary? Which ones will you remove?

Out of the continuous variables, it looks like temperature and humidity covary. I may remove humidity because I think temperature is a more reliable predictor of salamander behavior.



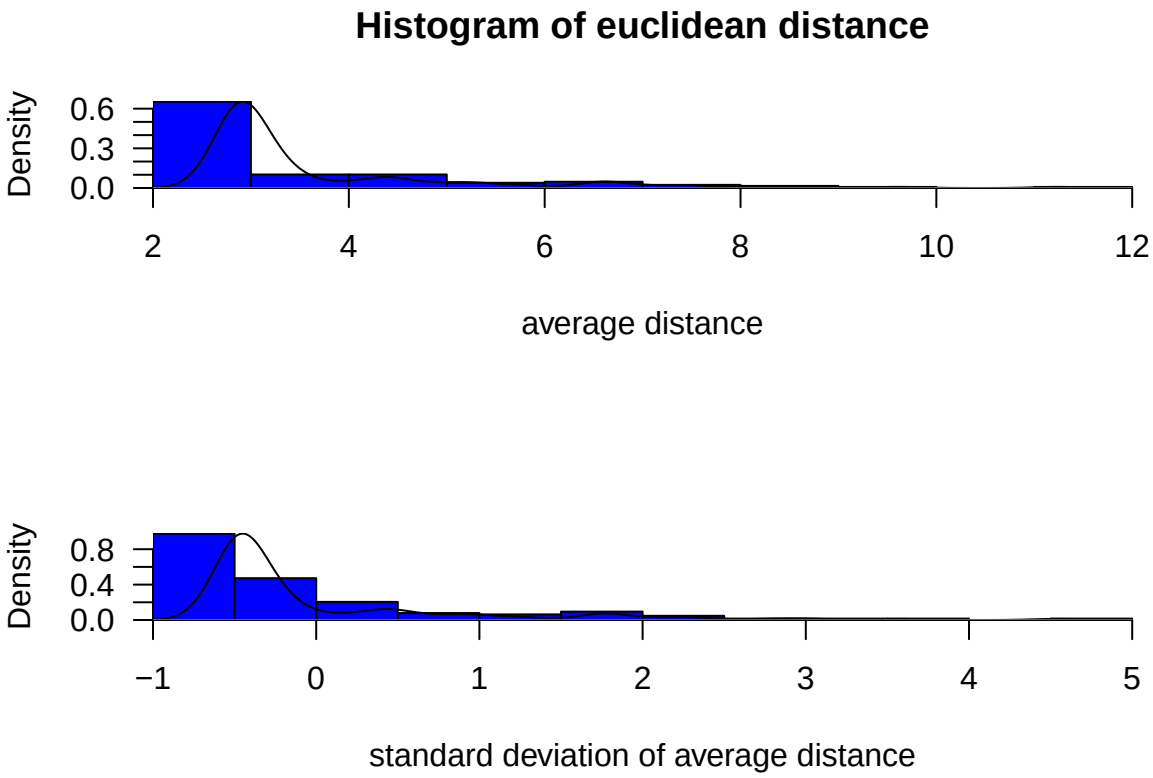
The categorical comparisons show a few potential trends. For the most part, soil moisture, temp, and humidity are all fluctuating around a similar level across treatments. Downed wood looks the most interesting, with more in the salvage logged and control plots than in the harvested and burned plots, which makes sense given my time on the ground. It's clear that canopy cover is highly related to treatment type. Which is not groundbreaking, seeing as my treatment types include logging.

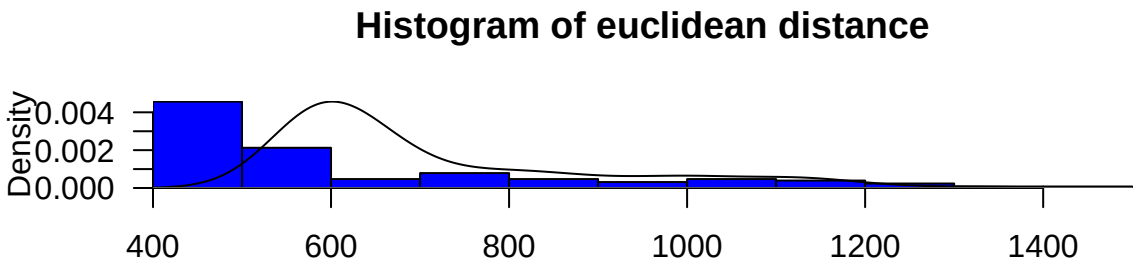
Treatments:
- BS = burned, salvage logged
- BU = burned, unharvested
- HB = harvested, burned
- HU = harvested, unburned
- UU = unharvested, unburned

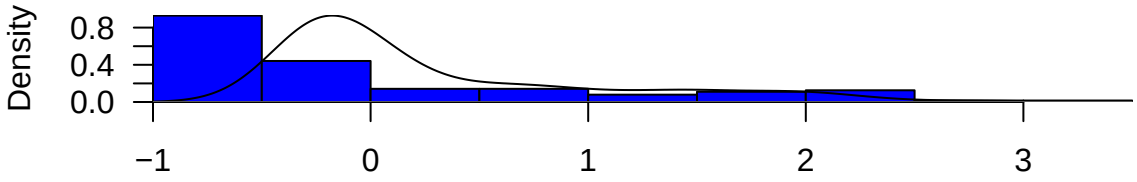


Question 7) Do you have outliers in your data? How will you handle them? Do you think there are ecological reasons for keeping any outliers in your analysis?

I have a few outliers, and some of them appear to be sites that are high in elevation. The outliers for the environmental and salamander only overlap at one site. Does that mean I dont need to be worried about removing them?


Histogram of euclidean distance
Density
0.6
0.3
0.0
2
4
6
8
10
12
average distance
Density
0.8
0.4
0.0
−1
0
1
2
3
4
5
standard deviation of average distance






```
saloutlier
```

```
##                   avedist    sd
## 10439 _ 1 _ 2024    7.101 2.166
## 12033 _ 1 _ 2023   11.480 4.782
## 12213 _ 1 _ 2023    7.491 2.399
## 12349 _ 1 _ 2024    9.737 3.741
## 12804 _ 1 _ 2024    7.484 2.395
## 29122 _ 1 _ 2023    8.797 3.179
## 408162 _ 1 _ 2024   8.349 2.911
```

```
envoutlier
```

```
##                   avedist    sd
## 10498 _ 1 _ 2023 1124.215 2.041
## 12213 _ 1 _ 2023 1174.919 2.250
## 12213 _ 1 _ 2024 1202.817 2.365
## 12239 _ 1 _ 2023 1132.261 2.074
## 12239 _ 1 _ 2024 1166.865 2.217
## 20039 _ 1 _ 2023 1229.986 2.477
## 20039 _ 1 _ 2024 1300.372 2.767
## 21026 _ 1 _ 2023 1447.255 3.372
## 34017 _ 1 _ 2023 1205.790 2.377
## 34017 _ 1 _ 2024 1177.641 2.261
```

```r
intersect(rownames(saloutlier),rownames(envoutlier))
```

```
## [1] "12213 _ 1 _ 2023"
```