Final project

Daniel Gonzalez-Suarez

4/3/2020

Contents

Final Assignment

This is the second assingment for the Ecological Data Analysis in R course. In order to start, packages and functions will be loaded first. We'll use *tidiverse*, *dplyr*, *vegan*, *viridis* and a customized function that will be available in the repository.

Loading data and tidying it up.

Here, steps for tidying the dataset will be presented. Raw data was taken during completion of underwater transects, then input in Excel. Data will be imported as a .csv file.

```
#df <- read.csv("./data/df_daniel.csv")
df <- read.csv("D:/UABC/II_Semestre/Datos_R/eda_finalproject/data/df_daniel.csv") #this one's for knitt

# We'll only need provinces, localities, sites, transects, species and abundances
df <- df %>% select(province, locality, siteref, transect, sp, length, abundance)

head(df, n=3L)
```

```
##     province     locality siteref transect                      sp length
## 1 Puntarenas ISLA_TORTUGA       A       A1   Gnathanodon speciosus     20
## 2 Puntarenas ISLA_TORTUGA       A       A1    Haemulon maculicauda     15
## 3 Puntarenas ISLA_TORTUGA       A       A1 Microspathodon dorsalis     25
##   abundance
## 1         5
## 2         4
## 3         2
```

```
tail(df, n=3L)
```

```
##        province    locality siteref transect                    sp length
## 1502 Guanacaste CUAJINIQUIL       G      G35         Sabellidae sp      5
## 1503 Guanacaste CUAJINIQUIL       G      G35 Stenorhynchus debilis     10
## 1504 Guanacaste CUAJINIQUIL       G      G35   Toxopneustes roseus      5
##      abundance
```

```
## 1502         4
## 1503         2
## 1504         4
```

```
str(df)
```

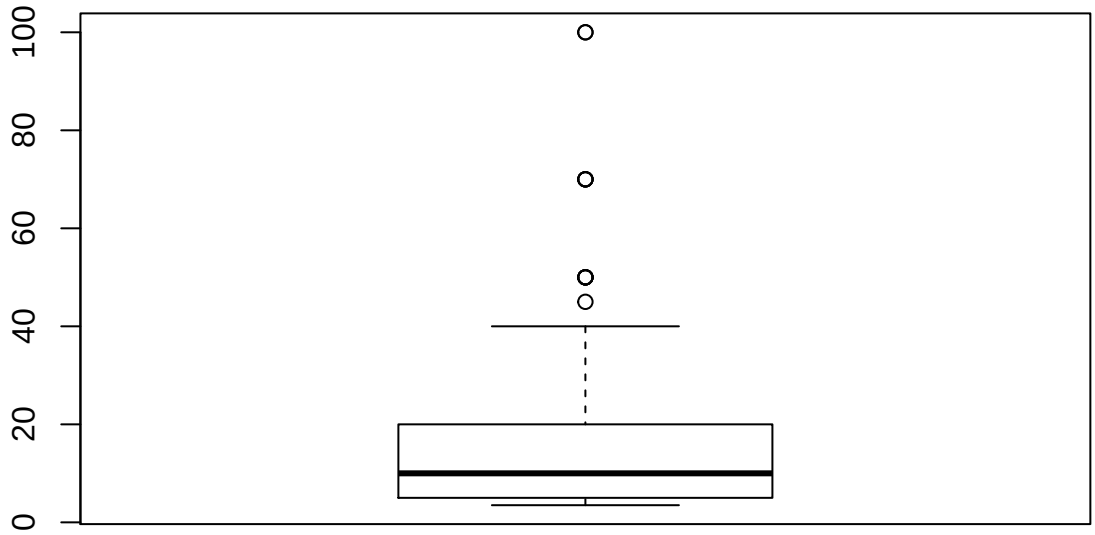
```
## 'data.frame':    1504 obs. of  7 variables:
##  $ province : Factor w/ 2 levels "Guanacaste","Puntarenas": 2 2 2 2 2 2 2 2 2 2 ...
##  $ locality : Factor w/ 3 levels "CUAJINIQUIL",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ siteref  : Factor w/ 7 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ transect : Factor w/ 35 levels "A1","A2","A3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ sp       : Factor w/ 117 levels "Abudefduf concolor",..: 46 50 67 67 2 1 107 107 23 39 ...
##  $ length   : num  20 15 25 30 10 5 10 15 15 15 ...
##  $ abundance: int  5 4 2 3 3 5 6 8 1 1 ...
```

Before continuing, let's remember our question: will there be a difference between communities of two provinces with different fishing pressures?

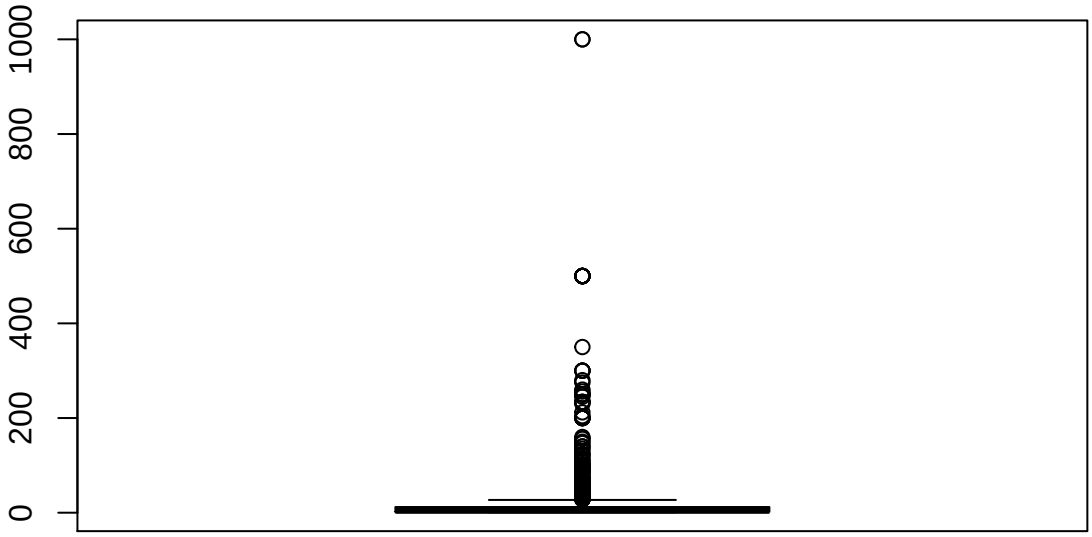
Data Exploration

1. Outliers

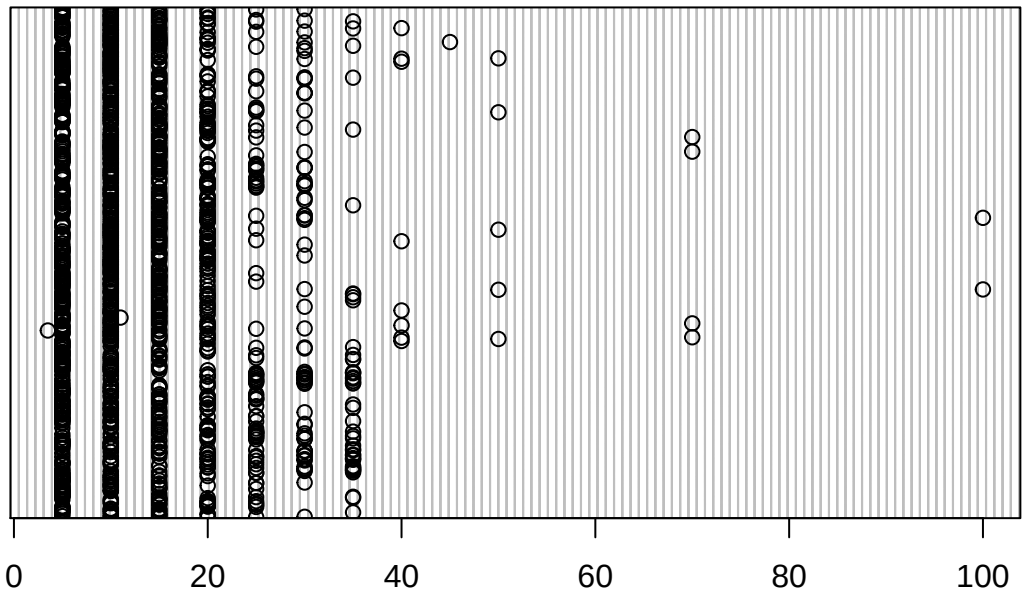
```
boxplot(df$length)
```



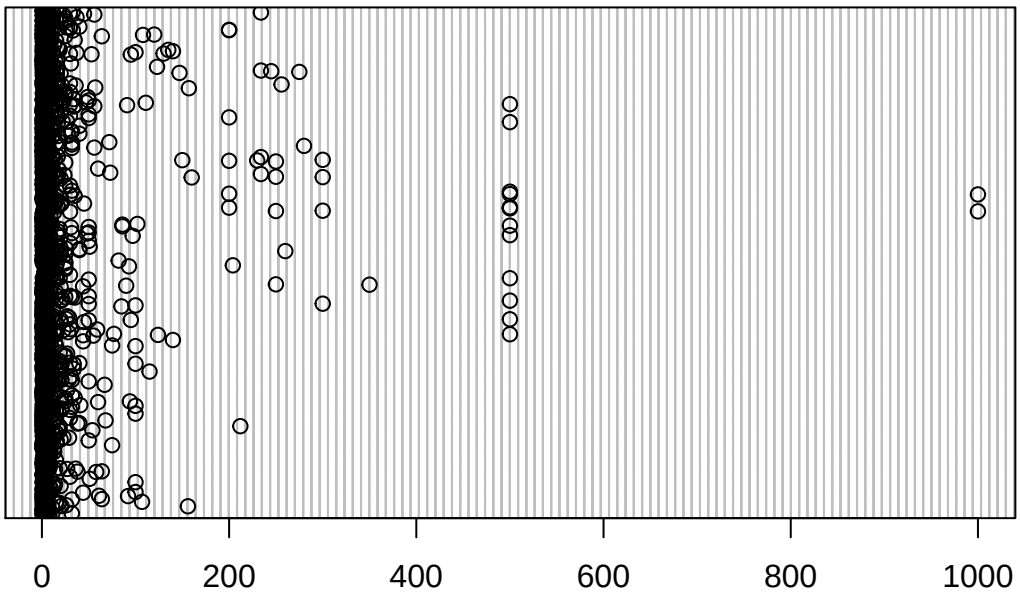

```
boxplot(df$abundance)
```




```
dotchart(df$length)
```



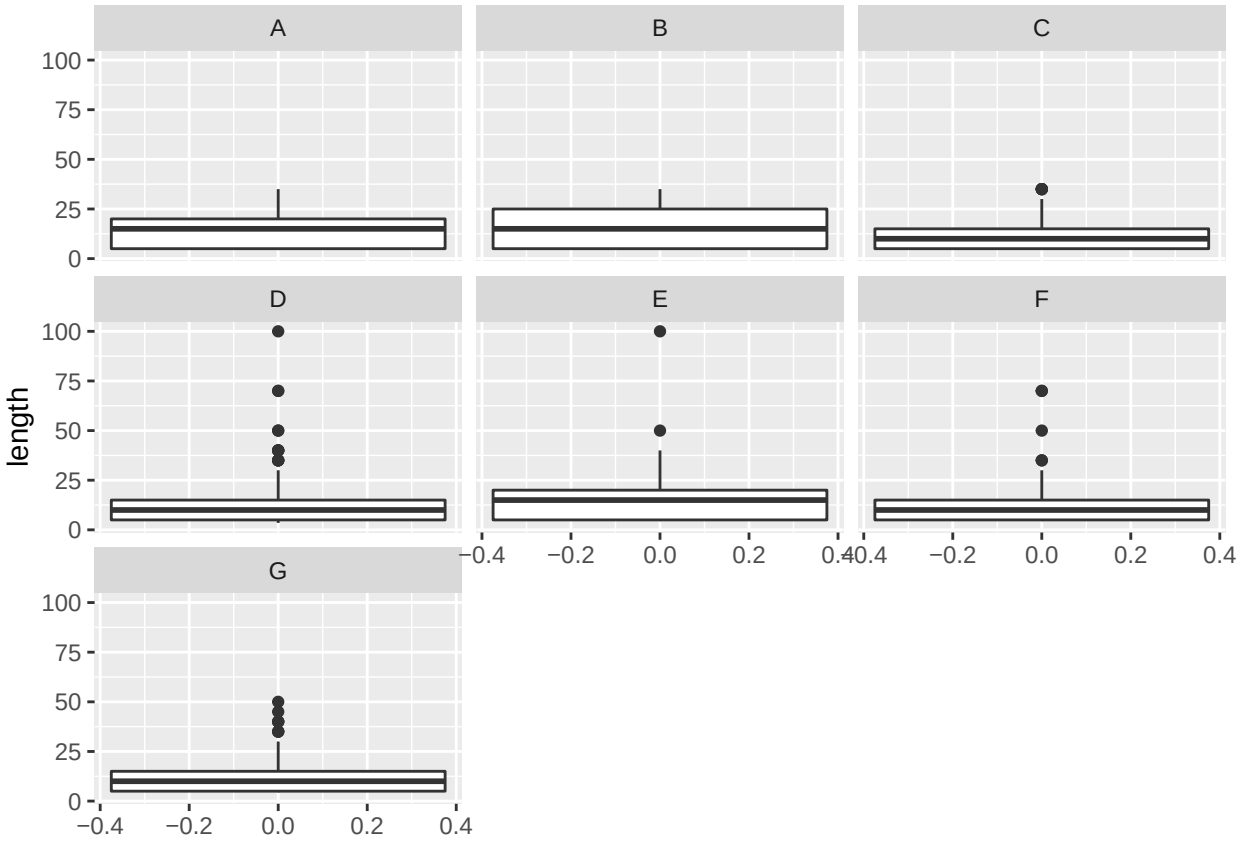

```
dotchart(df$abundance)
```



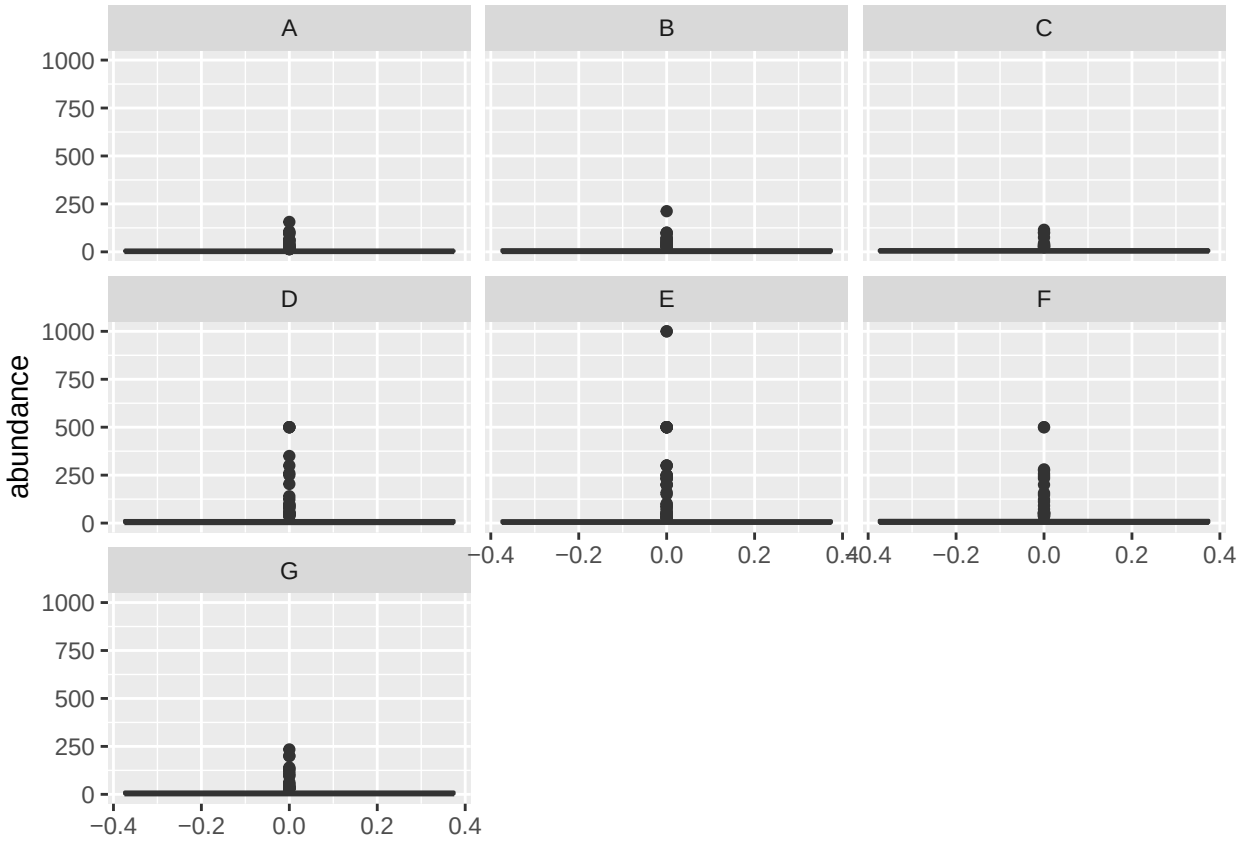
Not that uncommon to see abundances represented like that, I think. Lot's of zeroes and some outliers for abundance and no errors regarding length.

2. Homogeneity of variance

```
p <- ggplot(data=df, aes(X=province, y=length)) +
  geom_boxplot() + facet_wrap(~siteref)
p
```




```
p<- ggplot(data=df, aes(X=province, y=abundance)) +
  geom_boxplot() + facet_wrap(~siteref)
p
```



3. Normality

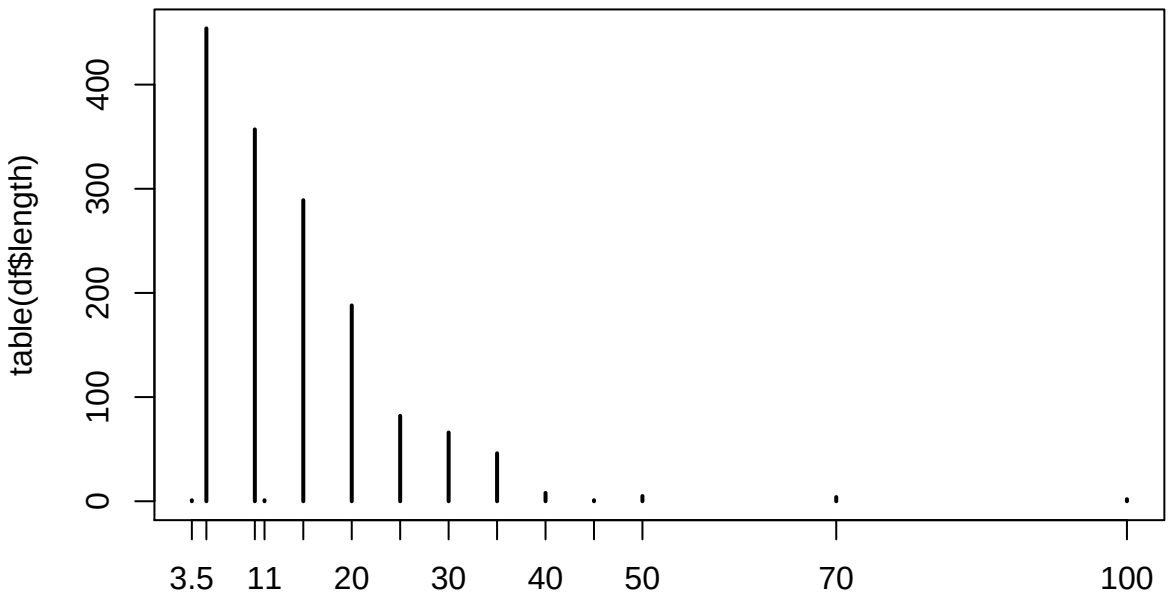
Data is not normal, as seen in histograms and confirmed by Q-Q plots.

4. Zeroes

```
range(df$length)
```

```
## [1]   3.5 100.0
```

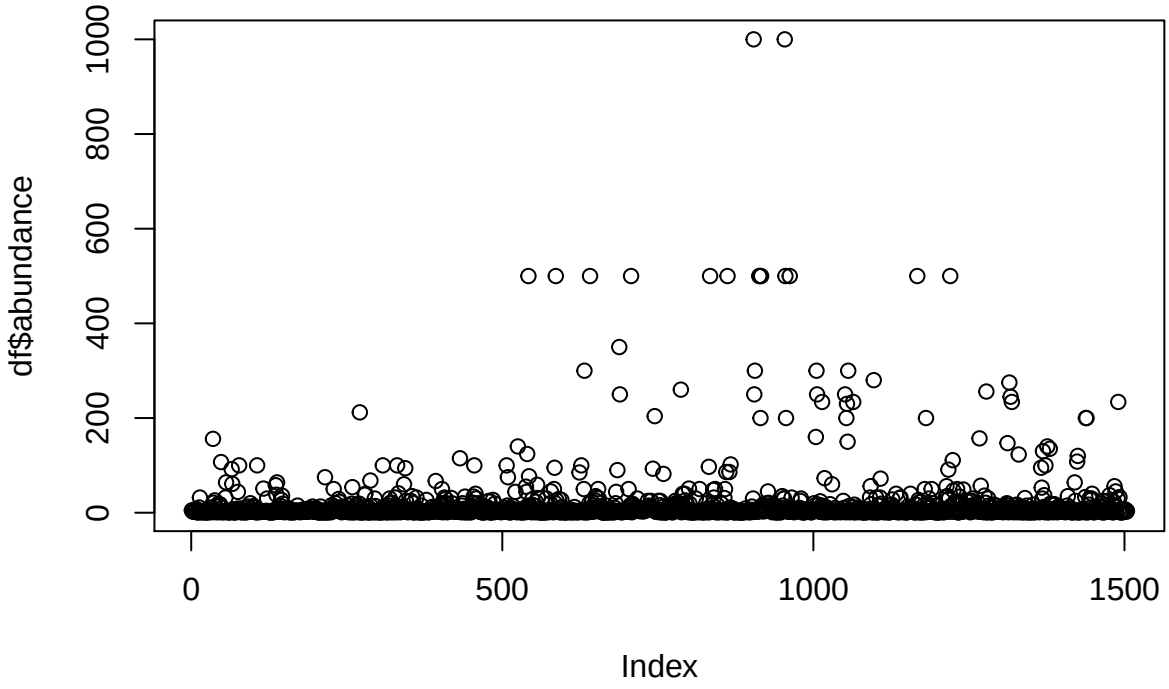
```
plot(table(df$length))
```




```
range(df$abundance)
```

```
## [1]    1 1000
```

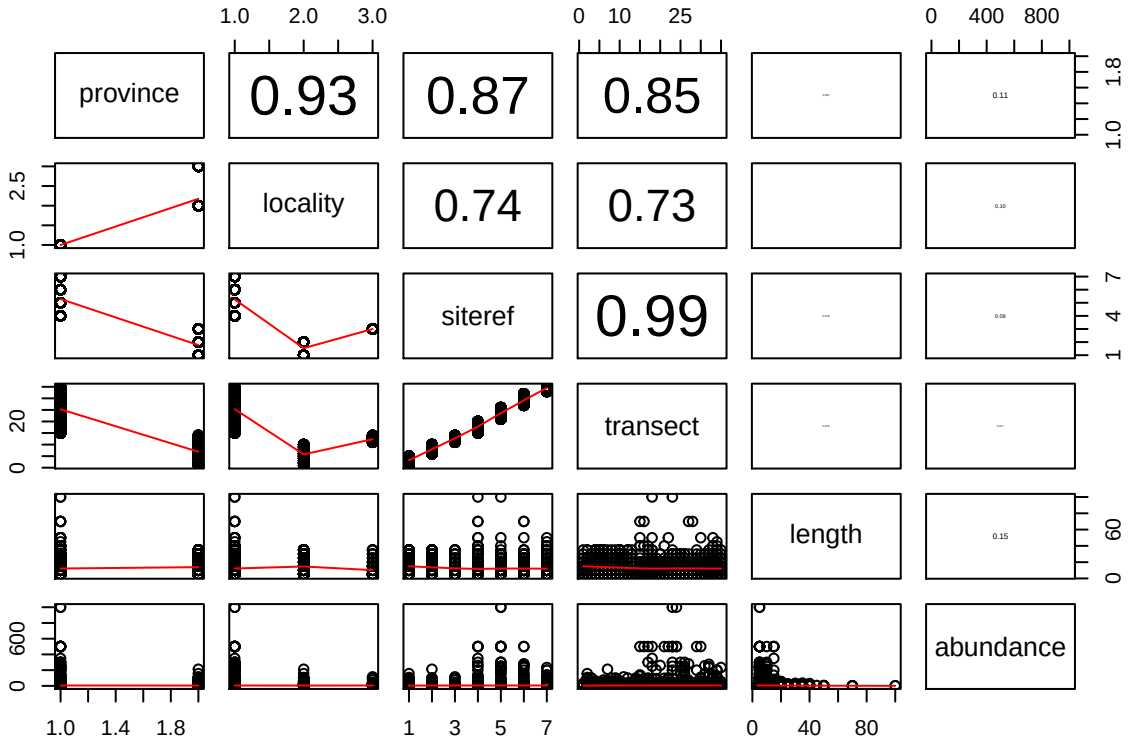
```
plot(df$abundance)
```



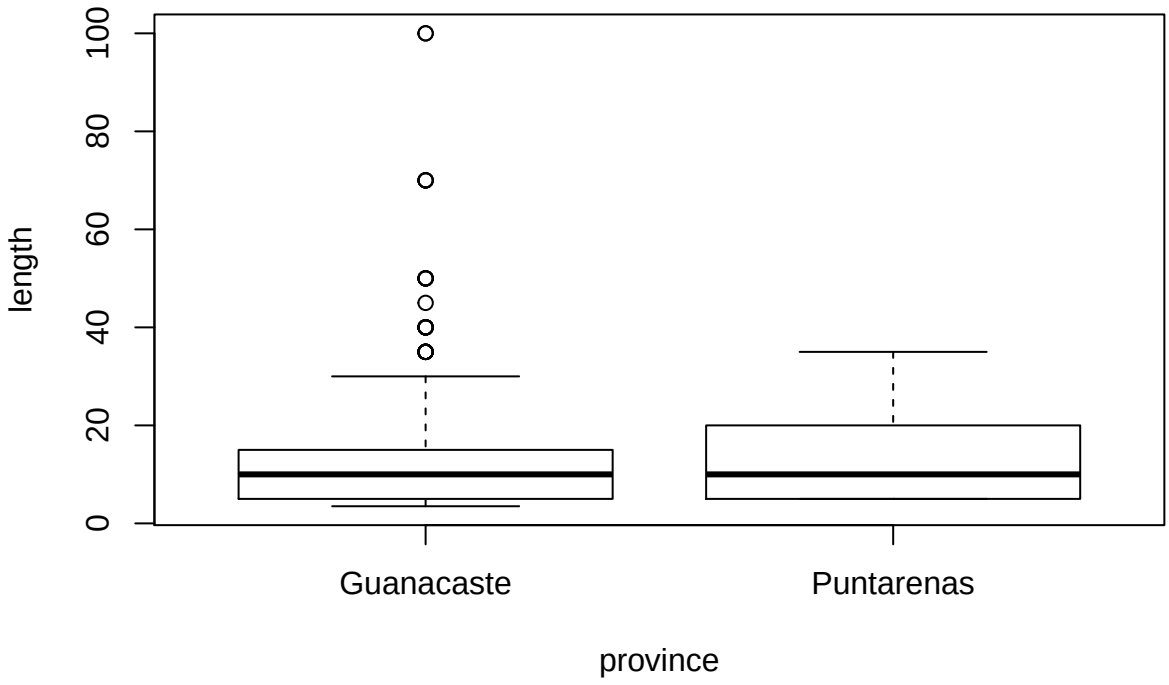
There are lots of zeroes for abundance, which we'll have to take into account when analyzing data.

5. Collinearity X

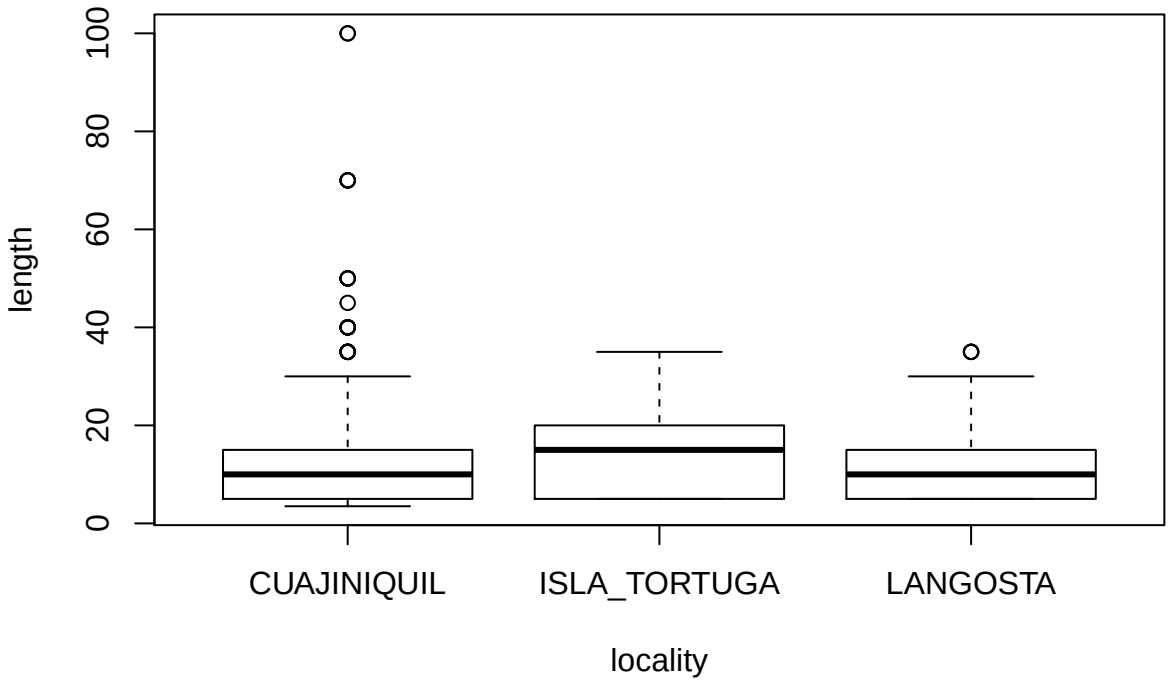
```
pairs(~ province + locality + siteref + transect + length + abundance,
      lower.panel=panel.smooth, upper.panel=panel.cor,
      data=df)
```



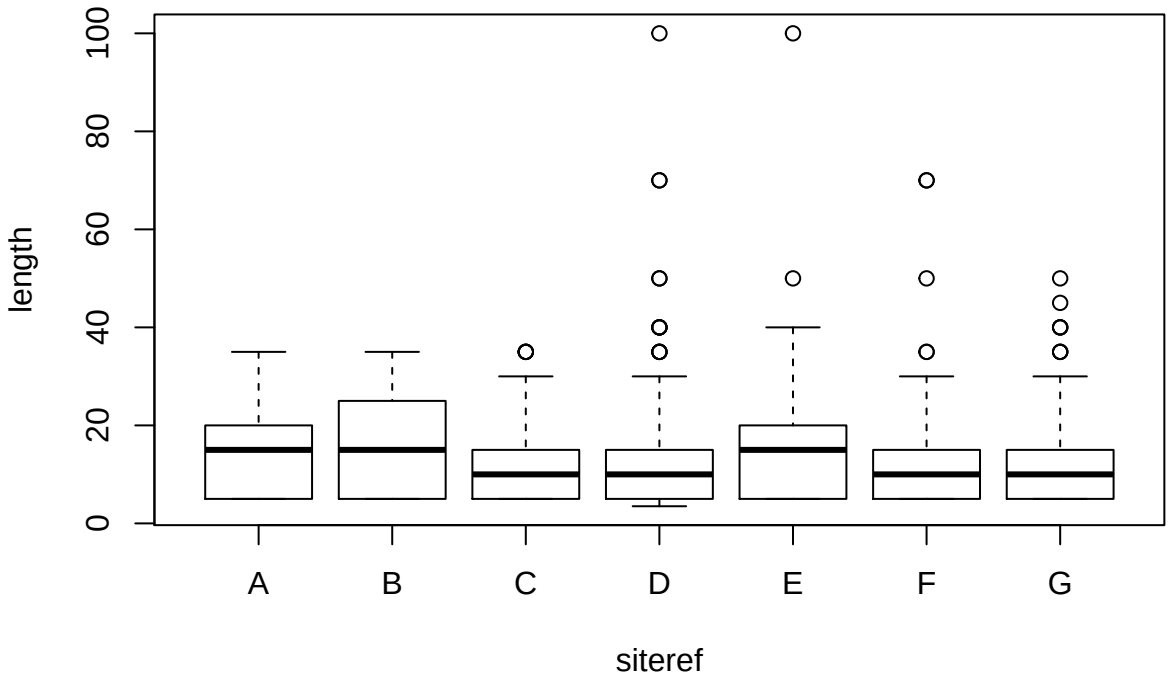

```
plot(length ~ province, data=df)
```



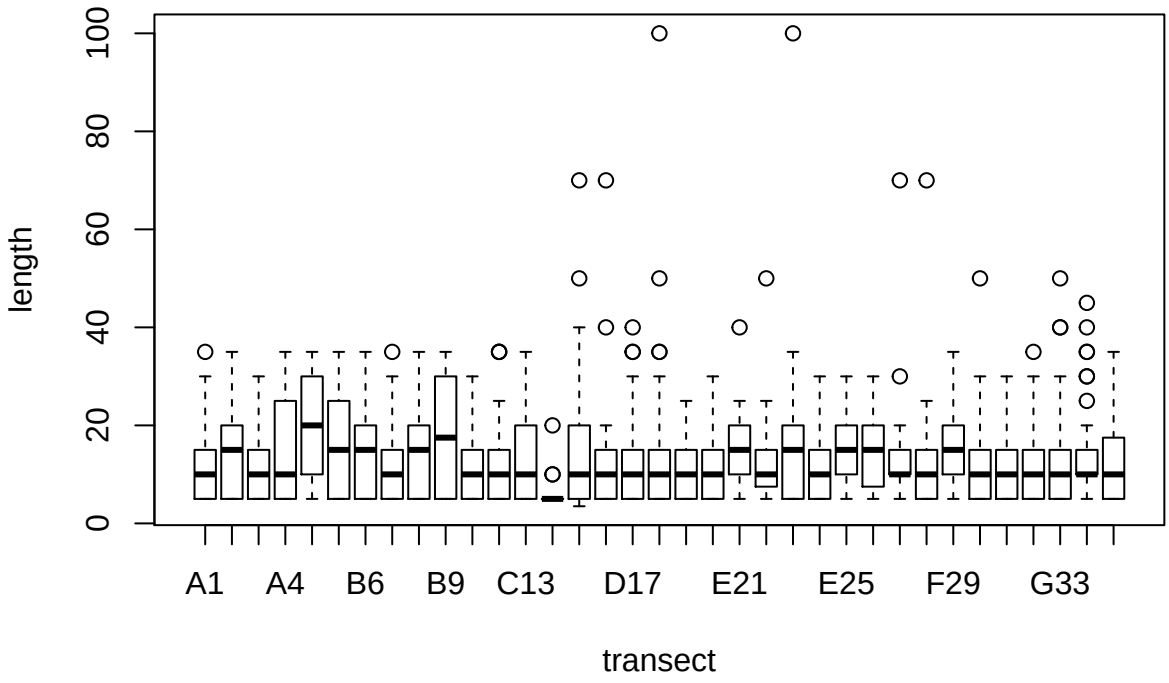

```
plot(length ~ locality, data=df)
```



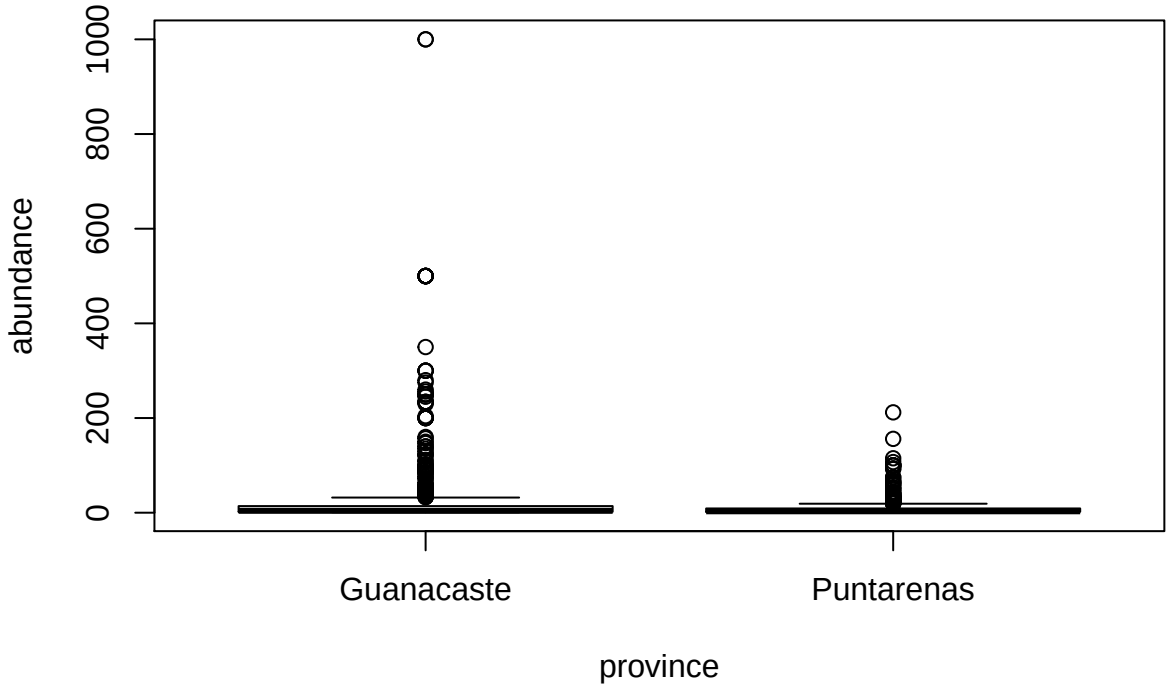

```
plot(length ~ siteref, data=df)
```



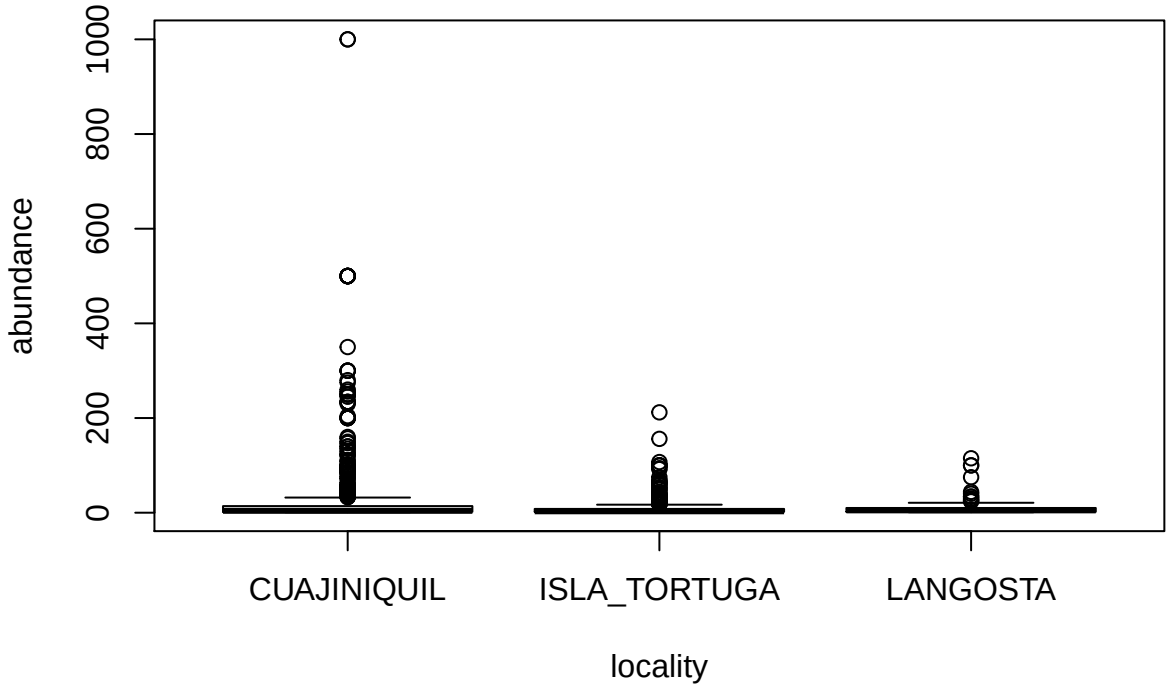

```
plot(length ~ transect, data=df)
```



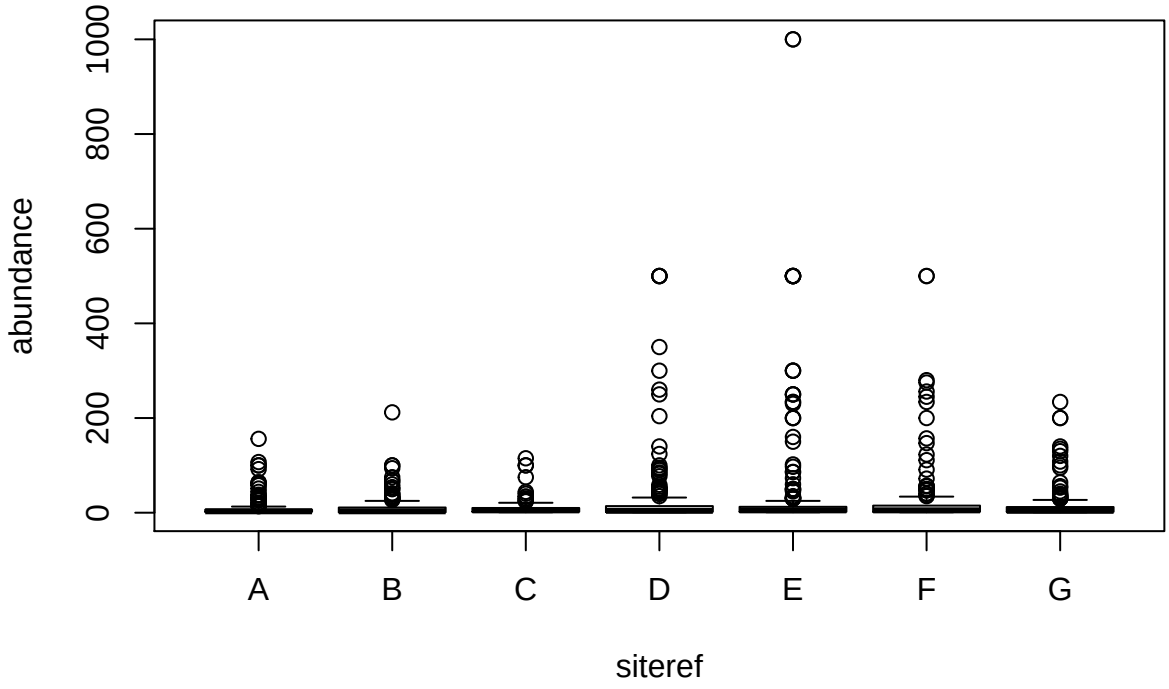

```
plot(abundance ~ province, data=df)
```



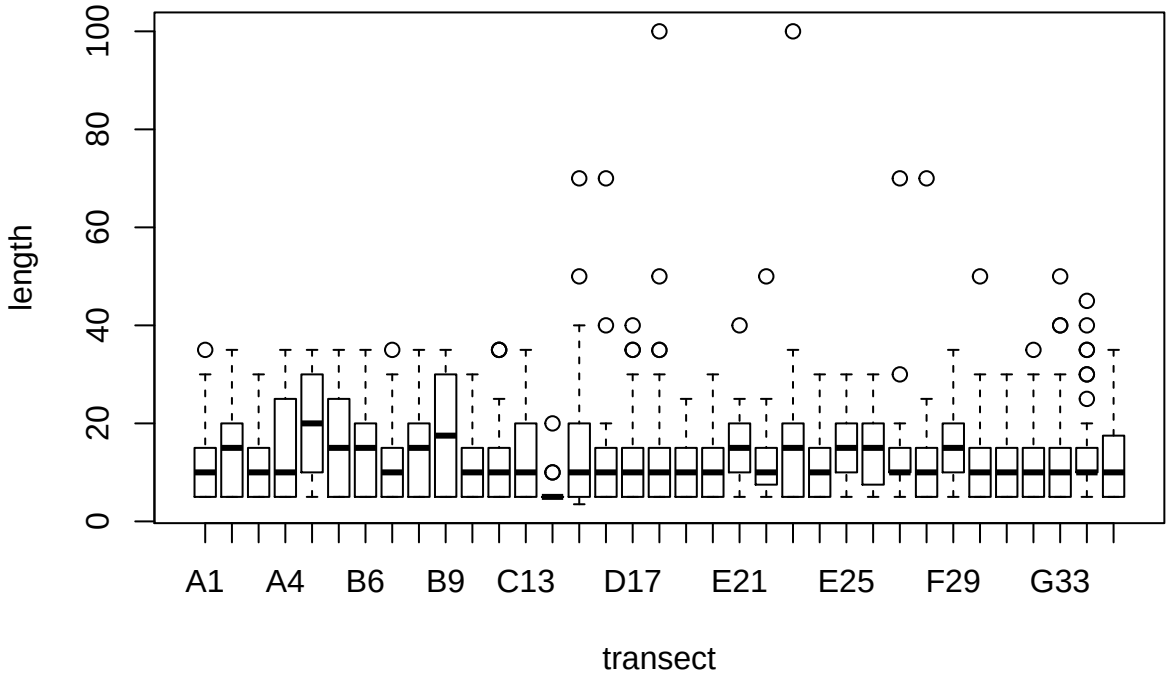

```
plot(abundance ~ locality, data=df)
```



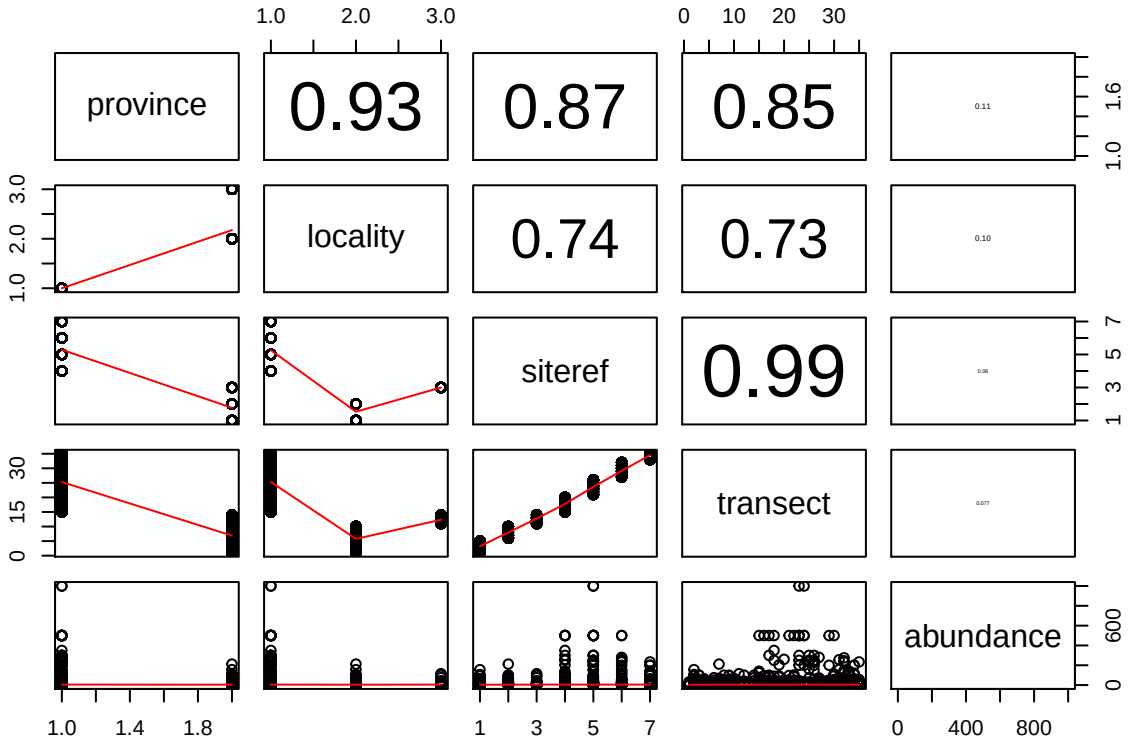

```
plot(abundance ~ siteref, data=df)
```




```
plot(length ~ transect, data=df)
```



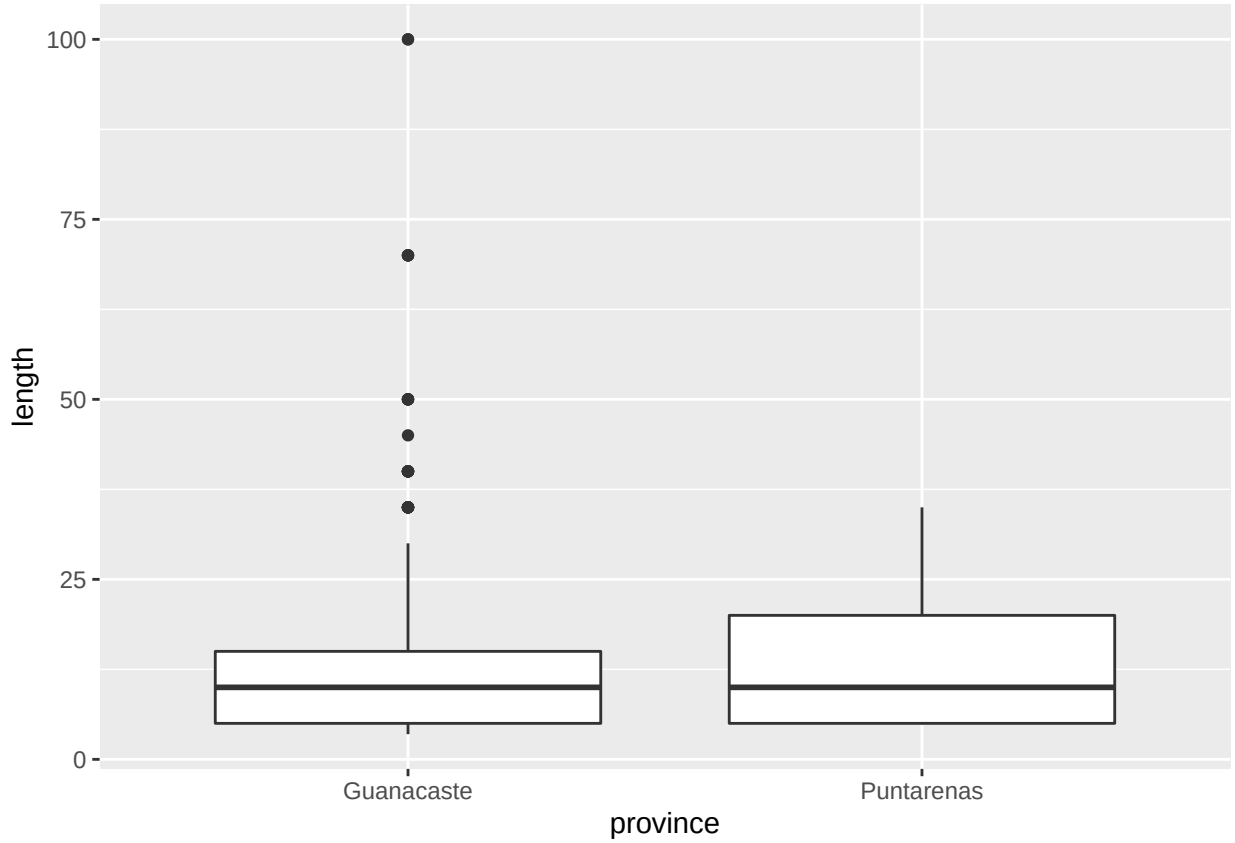

```
pairs(~ province+locality+siteref+transect+abundance,
      lower.panel=panel.smooth, upper.panel=panel.cor,
      data=df)
```



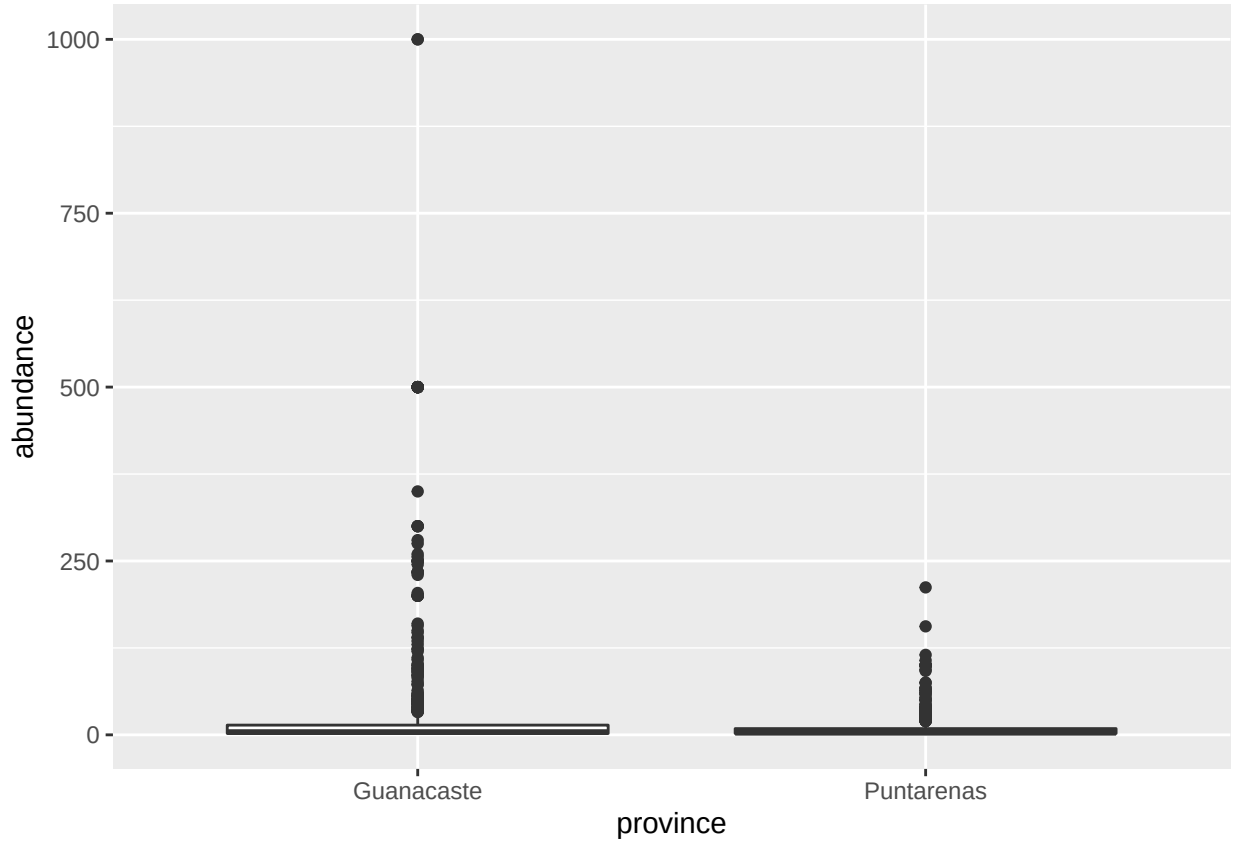
We'll take into account province, locality, site and transects as x variables and abundance as y variables; our question does not requires us to study length, so it's something we could have dropped before. We'll still explore it for the sake of it.

6. Relationships X and Y

```
p <- ggplot(df, aes(x=province, y=length)) +
  geom_boxplot()
p
```

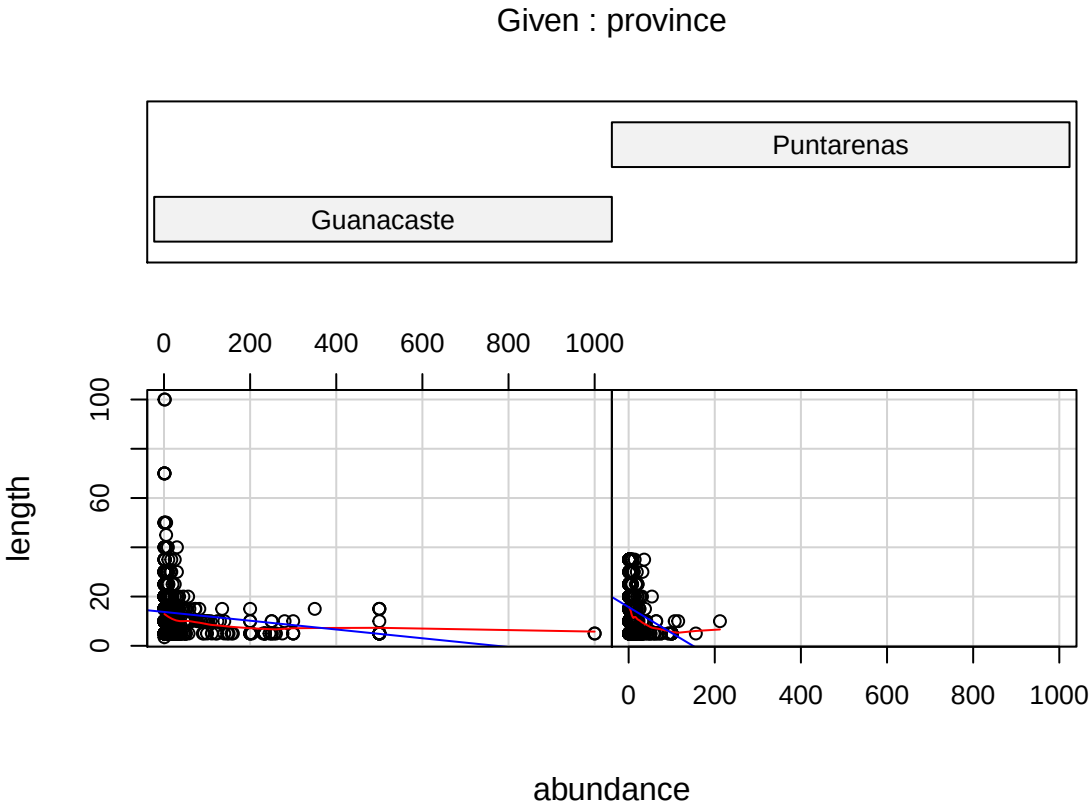



```
p <- ggplot(df, aes(x=province, y=abundance)) +
  geom_boxplot()
p
```

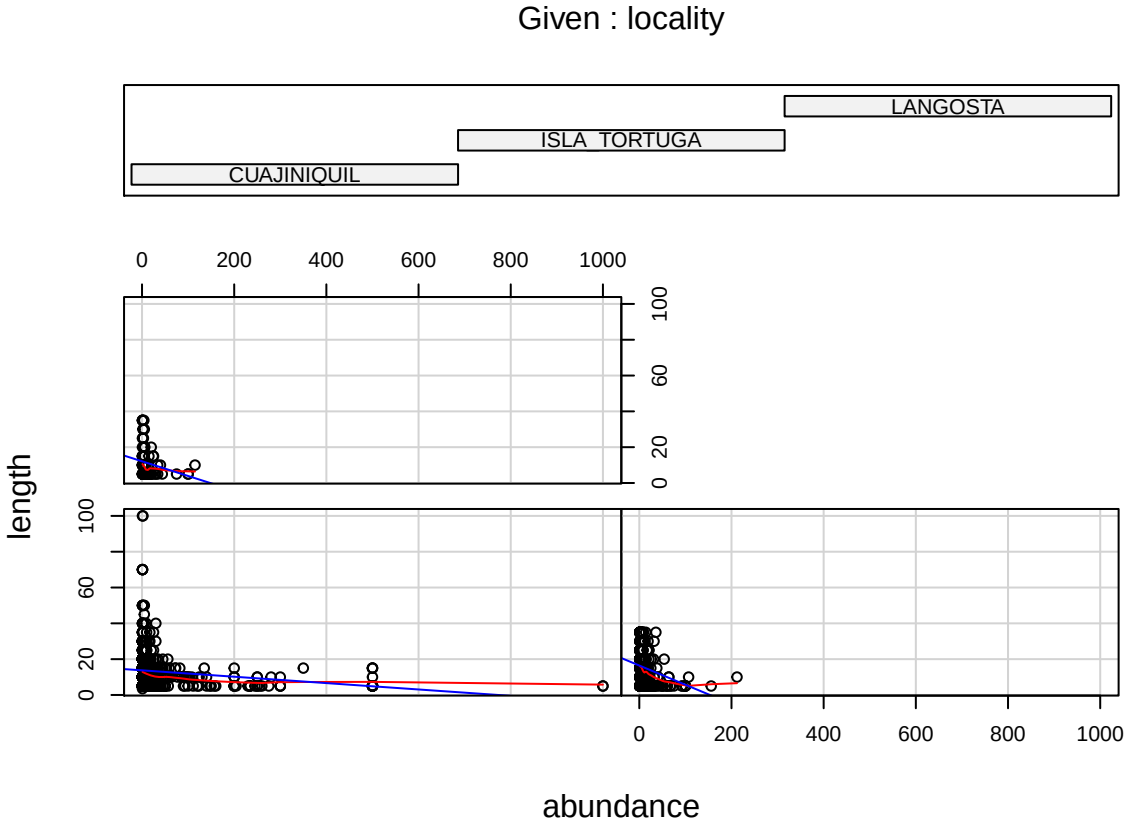


7. Interactions

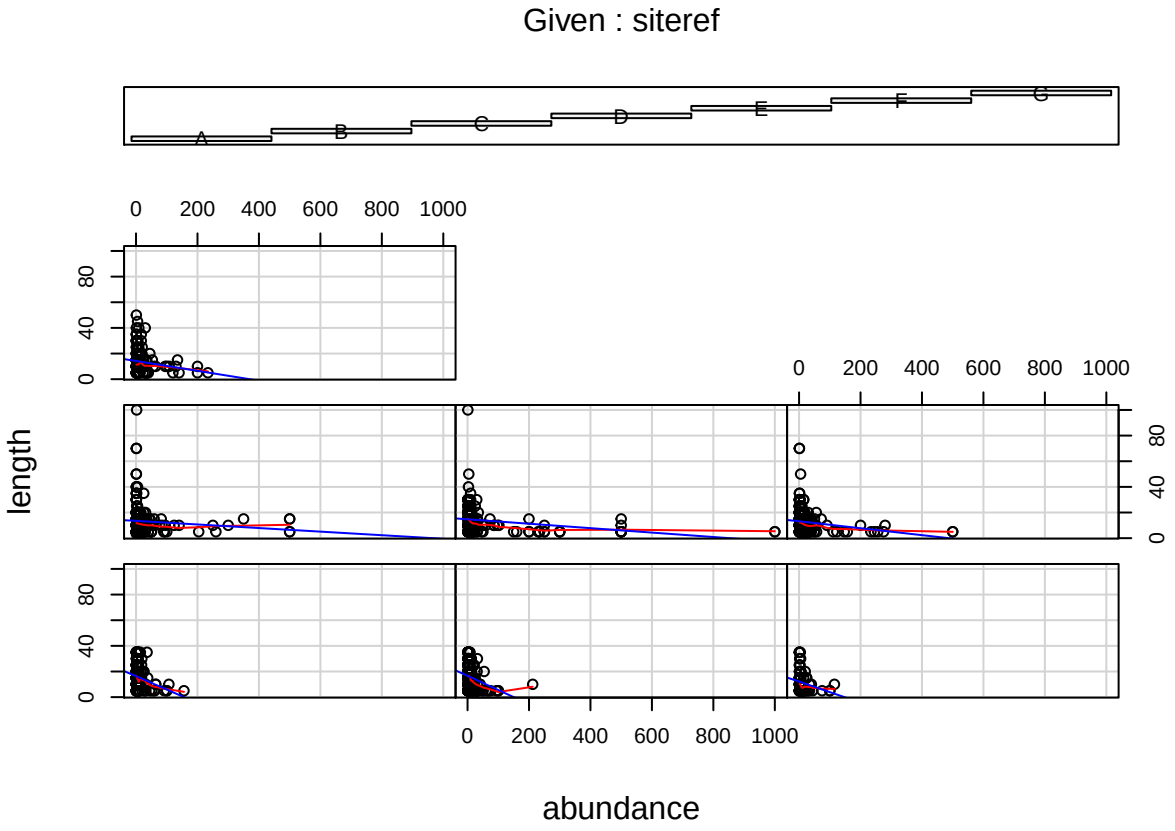
```
coplot(length ~ abundance | province ,
       data=df,
       panel=function(x,y,...) {
         panel.smooth(x,y,span=0.8,iter=5,...)
         abline(lm(y ~ x), col="blue") } )
```




```
coplot(length ~ abundance | locality ,
       data=df,
       panel=function(x,y,...) {
         panel.smooth(x,y,span=0.8,iter=5,...)
         abline(lm(y ~ x), col="blue") } )
```




```
coplot(length ~ abundance | siteref ,
       data=df,
       panel=function(x,y,...) {
         panel.smooth(x,y,span=0.8,iter=5,...)
         abline(lm(y ~ x), col="blue") } )
```




```
#coplot(length ~ abundance | transect ,
#       data=df,
#       panel=function(x,y,...) {
#         panel.smooth(x,y,span=0.8,iter=5,...)
#         abline(lm(y ~ x), col="blue") } )
```

There seems to be no interaction.

8. Independence of Y

Both length and density could be dependent on province, locality, site because of fishing pressure effects. In transects, I'd say there could be independence.

Indices

Data setup

We'll need to create an abundance matrix in order to calculate indices. Let's create a dataframe for our indices.

```
#Let's create a dataframe for our indices.
#We only need abundance

# first, we'll summarise species disregarding length
df_sp <- df %>% select(province, locality, siteref, transect, sp, abundance)
head(df_sp)
```

```
##     province     locality siteref transect                    sp abundance
```

```
## 1 Puntarenas ISLA_TORTUGA       A       A1   Gnathanodon speciosus         5
## 2 Puntarenas ISLA_TORTUGA       A       A1    Haemulon maculicauda         4
## 3 Puntarenas ISLA_TORTUGA       A       A1 Microspathodon dorsalis         2
## 4 Puntarenas ISLA_TORTUGA       A       A1 Microspathodon dorsalis         3
## 5 Puntarenas ISLA_TORTUGA       A       A1    Abudefduf troschelii         3
## 6 Puntarenas ISLA_TORTUGA       A       A1      Abudefduf concolor         5
```

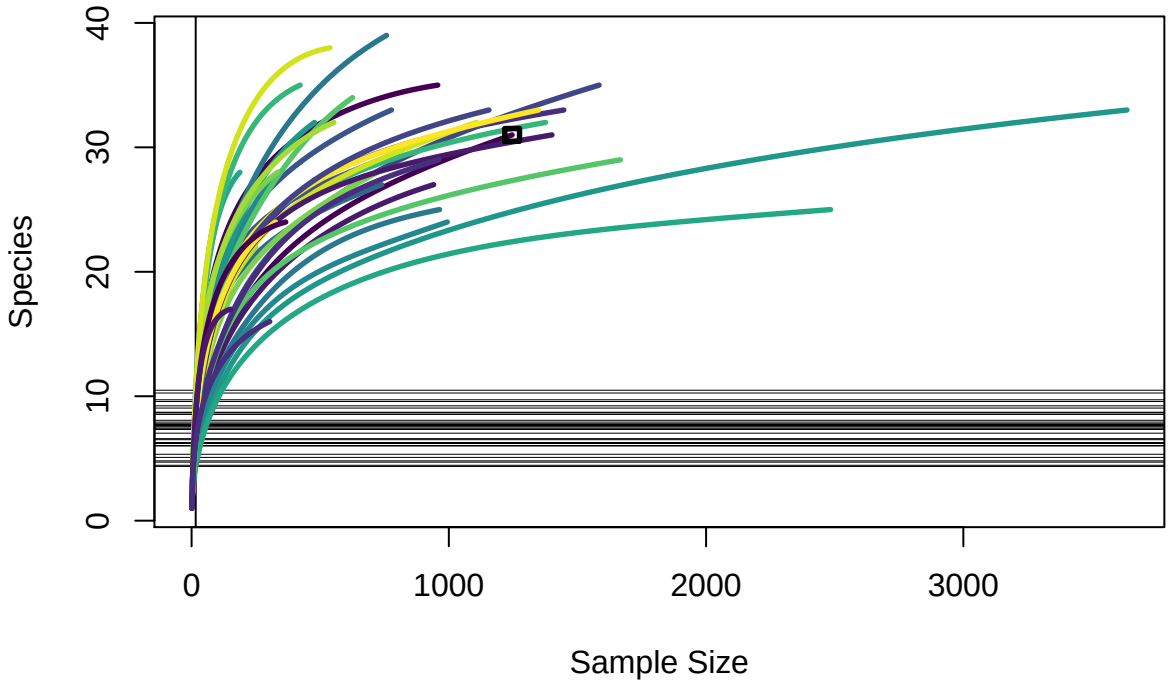
```
str(df_sp)
```

```
## 'data.frame':    1504 obs. of  6 variables:
##  $ province : Factor w/ 2 levels "Guanacaste","Puntarenas": 2 2 2 2 2 2 2 2 2 2 ...
##  $ locality : Factor w/ 3 levels "CUAJINIQUIL",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ siteref  : Factor w/ 7 levels "A","B","C","D",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ transect : Factor w/ 35 levels "A1","A2","A3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ sp       : Factor w/ 117 levels "Abudefduf concolor",..: 46 50 67 67 2 1 107 107 23 39 ...
##  $ abundance: int  5 4 2 3 3 5 6 8 1 1 ...
```

```
df_sp <- summarise(group_by(df_sp, province, locality, siteref, transect, sp),
                        sum_abundance = sum(abundance, na.rm = TRUE))

# ab_sp contains abundance matrix for ecological indices
# we'll create it out of df; as we don't care for sizes, we'll sum all abundances for each species on f
ab_sp  <- pivot_wider(df_sp, names_from = sp, values_from = sum_abundance)
#converting these factors to numeric will solve an issue that could arise later when we're creating our
ab_sp$province <- as.double(ab_sp$province)
ab_sp$locality <- as.double(ab_sp$locality)
ab_sp$siteref <-as.double(ab_sp$siteref)
ab_sp$transect <- as.double(ab_sp$transect)

# let's convert na to zeroes
ab_sp[is.na(ab_sp)] <-0
```

Now we can obtain ecological indices

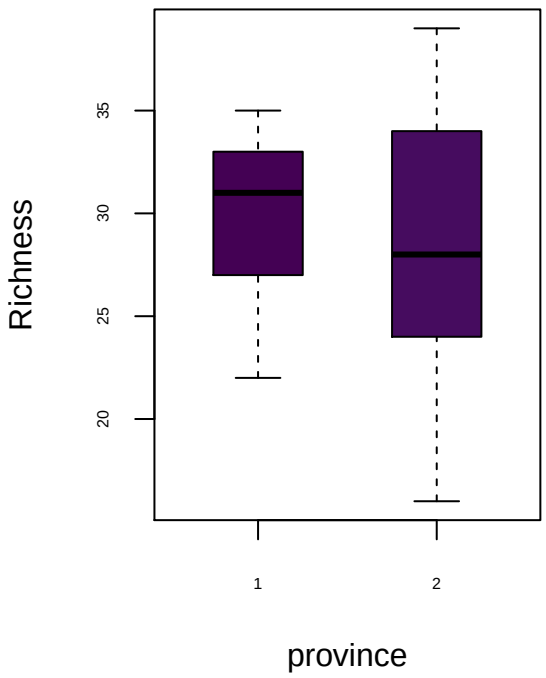
```
#We'll create a dataframe to store our results
ind_sp <- ab_sp[, c("province", "locality", "siteref", "transect")]

# Richness: number of species per transect/habitat
ind_sp$richness <- rowSums(ab_sp>0)

# Shannon´s diversity index: the bigger, the more diverse, basically
ind_sp$Shannon <- diversity(ab_sp)

# Rarefaction
raremax <- min(rowSums(ab_sp>0))
ind_sp$Rarefied <- c(rarefy(ab_sp[1:35,], sample=raremax))
raremax <- min(rowSums(ab_sp>0))

# Let's also visualize a rarefaction curve
#this function will create a rarefaction curve to observe species accumulation
rarecurve(ab_sp, sample = raremax, col=viridis(raremax, alpha = 1, begin = 0, end = 1, direction = 1, o
```

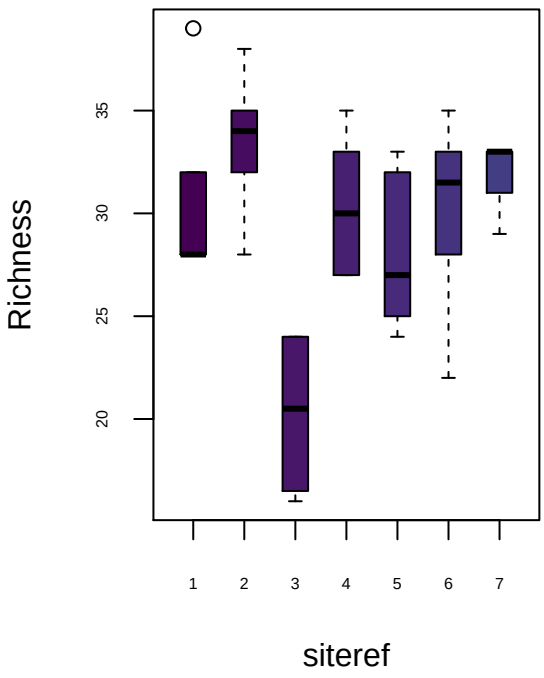


Visualization

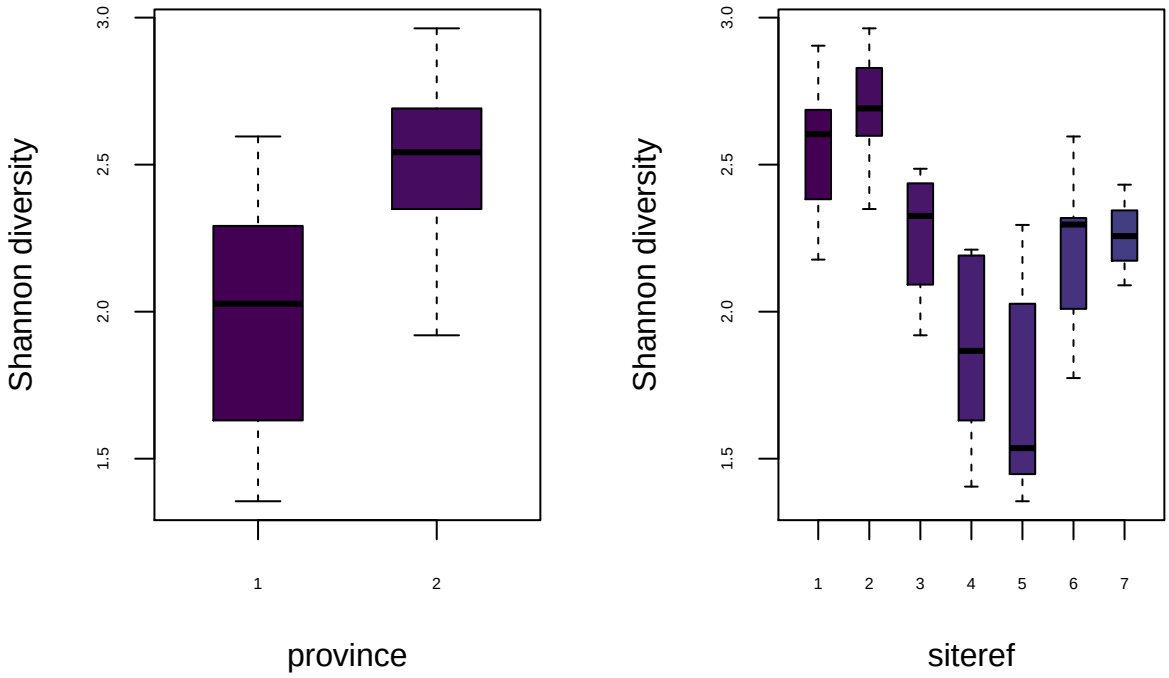
We'll use boxplots to see differences in species diversity:

```
#let's return our province/location/sites to factors
par(mfrow=c(1,2))
boxplot(richness~province, data=ind_sp, boxwex=0.5, col=viridis(35),
        cex.axis=0.5, ylab="Richness")
boxplot(richness~siteref, data=ind_sp, boxwex=0.5, col=viridis(35),
        cex.axis=0.5, ylab="Richness")
```

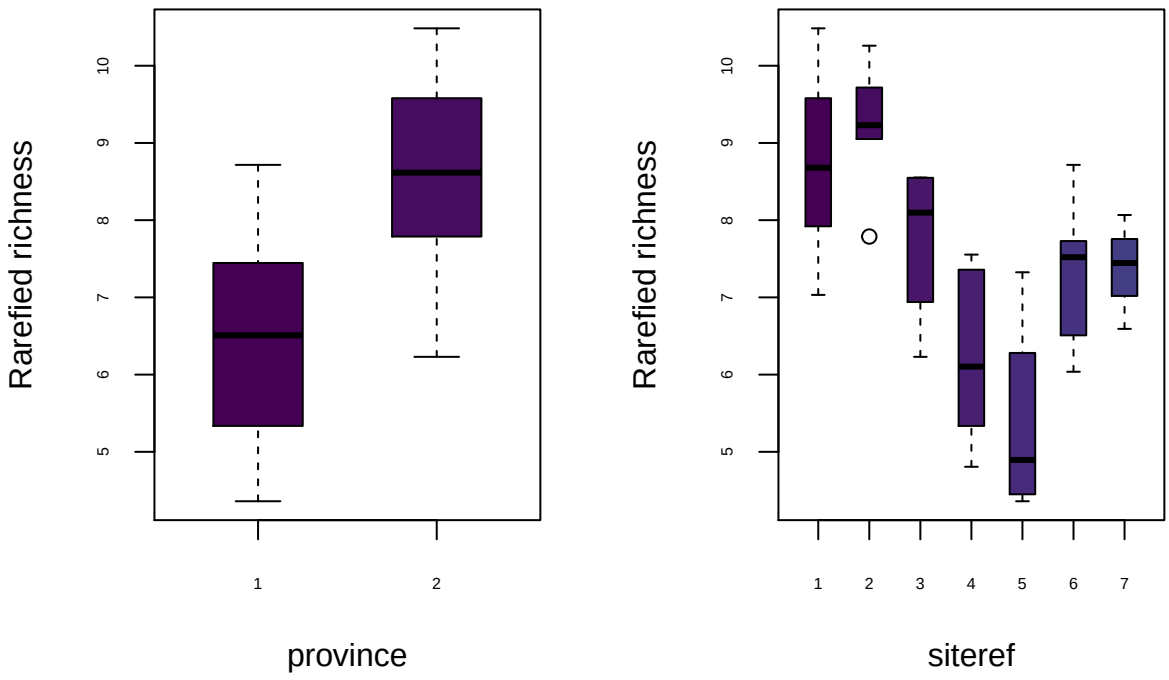





```
boxplot(Shannon~province, data=ind_sp, boxwex=0.5, col=viridis(35),
        cex.axis=0.5, ylab="Shannon diversity")
boxplot(Shannon~siteref, data=ind_sp, boxwex=0.5, col=viridis(35),
        cex.axis=0.5, ylab="Shannon diversity")
```




```
boxplot(Rarefied~province, data=ind_sp, boxwex=0.5, col=viridis(35),
        cex.axis=0.5, ylab="Rarefied richness")
boxplot(Rarefied~siteref, data=ind_sp, boxwex=0.5, col=viridis(35),
        cex.axis=0.5, ylab="Rarefied richness")
```




```
mfrow=c(1,1)
```

Looks like there are "some" differences, so let's explore them.

Linear modeling of indices

It appears that species diversity increases as we move from the field to the forest. We can test for differences among habitats statistically using a linear model, with Habitat as a predictor of species diversity:

```
# fit linear models (ANOVA)
mod.richness.province <- lm(richness~province, data=ind_sp)
mod.richness.siteref <- lm(richness~siteref, data=ind_sp)
mod.richness.transect <- lm(richness~transect, data=ind_sp)

mod.Shannon.province <- lm(Shannon~province, data=ind_sp)
mod.Shannon.siteref <- lm(Shannon~siteref, data=ind_sp)
mod.Shannon.transect <- lm(Shannon~transect, data=ind_sp)

mod.Rarefied.province <- lm(Rarefied~province, data=ind_sp)
mod.Rarefied.siteref <- lm(Rarefied~siteref, data=ind_sp)
mod.Rarefied.transect <- lm(Rarefied~transect, data=ind_sp)

anova(mod.richness.province)
```

```
## Analysis of Variance Table
##
## Response: richness
##           Df Sum Sq Mean Sq F value Pr(>F)
```

```
## province   1    8.8  8.8048  0.3208 0.5749
## Residuals 33  905.6 27.4423
```

```
anova(mod.richness.siteref)
```

```
## Analysis of Variance Table
##
## Response: richness
##           Df Sum Sq Mean Sq F value Pr(>F)
## siteref    1   0.31  0.3054   0.011  0.917
## Residuals 33 914.09 27.6998
```

```
anova(mod.richness.transect)
```

```
## Analysis of Variance Table
##
## Response: richness
##           Df Sum Sq Mean Sq F value Pr(>F)
## transect   1    0.1  0.1011  0.0036 0.9522
## Residuals 33  914.3 27.7060
```

```
#Results show no significant differences

anova(mod.Shannon.province)
```

```
## Analysis of Variance Table
##
## Response: Shannon
##           Df Sum Sq Mean Sq F value    Pr(>F)
## province   1 2.4871 2.48711  20.995 6.286e-05 ***
## Residuals 33 3.9092 0.11846
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod.Shannon.siteref)
```

```
## Analysis of Variance Table
##
## Response: Shannon
##           Df Sum Sq Mean Sq F value   Pr(>F)
## siteref    1 1.2784 1.27842  8.2432 0.007091 **
## Residuals 33 5.1179 0.15509
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod.Shannon.transect)
```

```
## Analysis of Variance Table
##
## Response: Shannon
##           Df Sum Sq Mean Sq F value  Pr(>F)
## transect   1 0.9268 0.92680  5.5918 0.02408 *
## Residuals 33 5.4695 0.16574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#There are significant differences for each one

anova(mod.Rarefied.province)
```

```
## Analysis of Variance Table
##
## Response: Rarefied
##           Df Sum Sq Mean Sq F value    Pr(>F)
## province   1 39.387  39.387  24.783 1.967e-05 ***
## Residuals 33 52.446   1.589
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod.Rarefied.siteref)
```

```
## Analysis of Variance Table
##
## Response: Rarefied
##           Df Sum Sq Mean Sq F value   Pr(>F)
## siteref    1 22.249 22.2491  10.552 0.002669 **
## Residuals 33 69.583  2.1086
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mod.Rarefied.transect)
```

```
## Analysis of Variance Table
##
## Response: Rarefied
##           Df Sum Sq Mean Sq F value  Pr(>F)
## transect   1 17.048 17.0475  7.5225 0.00977 **
## Residuals 33 74.785  2.2662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Same!

#Should we compare between groups? Let's try a Tukey test. Also, I'm really not sure what am I doing bu
#Tukey for Shannon
TukeyHSD(aov(lm(Shannon~as.factor(siteref), data = ind_sp)), conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm(Shannon ~ as.factor(siteref), data = ind_sp))
##
## $`as.factor(siteref)`
##             diff         lwr         upr     p adj
## 2-1  0.135322169 -0.45005101  0.72069534 0.9891678
## 3-1 -0.286564766 -0.90744678  0.33431725 0.7627015
## 4-1 -0.689117231 -1.24956938 -0.12866508 0.0087555
## 5-1 -0.851372524 -1.41182467 -0.29092038 0.0008059
## 6-1 -0.335742646 -0.89619479  0.22470950 0.4966182
## 7-1 -0.291317360 -0.96724808  0.38461336 0.8140044
## 3-2 -0.421886934 -1.04276895  0.19899508 0.3501185
## 4-2 -0.824439400 -1.38489155 -0.26398725 0.0012068
## 5-2 -0.986694692 -1.54714684 -0.42624255 0.0001045
## 6-2 -0.471064815 -1.03151696  0.08938733 0.1446756
## 7-2 -0.426639528 -1.10257025  0.24929119 0.4354619
```

```
## 4-3 -0.402552465 -0.99999646  0.19489153 0.3597010
## 5-3 -0.564807758 -1.16225175  0.03263624 0.0731869
## 6-3 -0.049177881 -0.64662188  0.54826612 0.9999686
## 7-3 -0.004752594 -0.71165786  0.70215267 1.0000000
## 5-4 -0.162255292 -0.69662545  0.37211486 0.9578414
## 6-4  0.353374585 -0.18099557  0.88774474 0.3811865
## 7-4  0.397799871 -0.25666724  1.05226698 0.4796657
## 6-5  0.515629877 -0.01874028  1.05000003 0.0639733
## 7-5  0.560055164 -0.09441194  1.21452227 0.1316368
## 7-6  0.044425287 -0.61004182  0.69889239 0.9999900
```

```
#Differences for  site1-4, site1-5, site2-4, site2-5 and almost site5-6

#Tukey for Rarefied richness
TukeyHSD(aov(lm(Rarefied~as.factor(siteref), data = ind_sp)), conf.level = 0.95)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = lm(Rarefied ~ as.factor(siteref), data = ind_sp))
##
## $`as.factor(siteref)`
##             diff         lwr        upr     p adj
## 2-1  0.47024930 -1.70561859  2.6461172 0.9924004
## 3-1 -0.99440819 -3.30226461  1.3134482 0.8141752
## 4-1 -2.52858373 -4.61181864 -0.4453488 0.0099268
## 5-1 -3.37119589 -5.45443079 -1.2879610 0.0003488
## 6-1 -1.40024431 -3.48347921  0.6829906 0.3624857
## 7-1 -1.37080826 -3.88328409  1.1416676 0.6023670
## 3-2 -1.46465750 -3.77251391  0.8431989 0.4290732
## 4-2 -2.99883303 -5.08206794 -0.9155981 0.0015708
## 5-2 -3.84144519 -5.92468010 -1.7582103 0.0000516
## 6-2 -1.87049361 -3.95372852  0.2127413 0.1005301
## 7-2 -1.84105756 -4.35353339  0.6714183 0.2679949
## 4-3 -1.53417554 -3.75491141  0.6865603 0.3314975
## 5-3 -2.37678769 -4.59752356 -0.1560518 0.0299716
## 6-3 -0.40583611 -2.62657199  1.8148998 0.9969322
## 7-3 -0.37640006 -3.00401018  2.2512101 0.9992154
## 5-4 -0.84261216 -2.82889870  1.1436744 0.8246237
## 6-4  1.12833942 -0.85794712  3.1146260 0.5576441
## 7-4  1.15777547 -1.27491879  3.5904697 0.7368895
## 6-5  1.97095158 -0.01533497  3.9572381 0.0528165
## 7-5  2.00038763 -0.43230664  4.4330819 0.1616133
## 7-6  0.02943605 -2.40325822  2.4621303 1.0000000
```

```
#Differeces for site 1-4, site1-5, site2-4, site2-5, site3-5 and almost site5-6
```

Interpretation

Results of ANOVA show no significant difference for richness, but does show it for Shannon and rarefied species richness. Seems like there's no association between province/site and species richness, but it is for Shannon's diversity index and rarefied richness.

Should we go on? Yes! Well... Seems like my data won't allow me to continue, if I'm not mistaken. I was considering doing an ordination analysis to determine which sites were similar and what species share distribution in provinces, sites and/or transects. I'd need to dampen data from abundances and divide it by

. . . ? I don't know. So I'll leave it up to here.

Based on what I have, the answer to my question would be that there's evidence -based on my analysis- that differences in species composition could be explained by the location, site or transect. Based on Tukey test, there are some differences between provinces: sites 1-3 correspond to Puntarenas, the high fishing pressure province, and 4-7 to Guanacaste, the not-so-high fishing pressure province.

Acknowledgement

Luis, I'd like to thank you for your effort, patience and dedication given to this course and us. I'm deeply grateful I got into it. I've learned a lot and have also improved some practices that I didn't know were not so great. Data exploration has been (and will continue to be) really useful to me, and will be fundamental for each research project. Thank you, Luis! You're a great professor and a remarkable human being, even if we have only shared a few months.