Final project

Daniel Gonzalez-Suarez

4/3/2020

Contents

1. Outliers 3

2. Homogeneity of variance 6

3. Normality 8

4. Zeroes 12

5. Collinearity X 14

6. Relationships X and Y 22

7. Interactions 24

8. Independence of Y 25

Part II, 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.

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_dgonzalezsuarez.csv")
head(df)
```

```
##          Dia  Provincia Latitud  Longitud    Localidad      SitioCodigo
## 1 19-Mar-16 Puntarenas 9.78203 -84.89076 ISLA_TORTUGA IT_SIN_TURISTAS
## 2 19-Mar-16 Puntarenas 9.78203 -84.89076 ISLA_TORTUGA IT_SIN_TURISTAS
## 3 19-Mar-16 Puntarenas 9.78203 -84.89076 ISLA_TORTUGA IT_SIN_TURISTAS
## 4 19-Mar-16 Puntarenas 9.78203 -84.89076 ISLA_TORTUGA IT_SIN_TURISTAS
## 5 19-Mar-16 Puntarenas 9.78203 -84.89076 ISLA_TORTUGA IT_SIN_TURISTAS
## 6 19-Mar-16 Puntarenas 9.78203 -84.89076 ISLA_TORTUGA IT_SIN_TURISTAS
##   Profundidad Transecto Area_.m2.   Phylum         Taxa
## 1           2         1        50 Chordata    Carangidae
## 2           2         1        50 Chordata    Haemulidae
## 3           2         1        50 Chordata Pomacentridae
```

1

```
## 4            2        1       50 Chordata Pomacentridae
## 5            2        1       50 Chordata Pomacentridae
## 6            2        1       50 Chordata Pomacentridae
##                   Especie NivelTrofico Talla_.cm. Abundancia Densidad_.org.m2.
## 1   Gnathanodon speciosus    Piscivoros         20          5              0.10
## 2     Haemulon maculicauda    Carnivoros         15          4              0.08
## 3 Microspathodon dorsalis    Herbivoros         25          2              0.04
## 4 Microspathodon dorsalis    Herbivoros         30          3              0.06
## 5     Abudefduf troschelii    Herbivoros         10          3              0.06
## 6      Abudefduf concolor    Herbivoros          5          5              0.10
```

tail(df)

```
##          Dia  Provincia Latitud  Longitud   Localidad SitioCodigo Profundidad
## 783 22-May-16 Guanacaste 10.9389 -85.73262 CUAJINIQUIL   SR_HACHAL           2
## 784 22-May-16 Guanacaste 10.9389 -85.73262 CUAJINIQUIL   SR_HACHAL           2
## 785 22-May-16 Guanacaste 10.9389 -85.73262 CUAJINIQUIL   SR_HACHAL           2
## 786 22-May-16 Guanacaste 10.9389 -85.73262 CUAJINIQUIL   SR_HACHAL           2
## 787 22-May-16 Guanacaste 10.9389 -85.73262 CUAJINIQUIL   SR_HACHAL           2
## 788 22-May-16 Guanacaste 10.9389 -85.73262 CUAJINIQUIL   SR_HACHAL           2
##     Transecto Area_.m2.  Phylum       Taxa              Especie NivelTrofico
## 783         1       50 Chordata Serranidae Serranus psittacinus   Carnivoros
## 784         1       50 Chordata Serranidae Serranus psittacinus   Carnivoros
## 785         1       50 Chordata Serranidae Serranus psittacinus   Carnivoros
## 786         1       50 Chordata   Labridae Thalassoma lucasanum   Carnivoros
## 787         1       50 Chordata   Labridae Thalassoma lucasanum   Carnivoros
## 788         1       50 Chordata   Labridae Thalassoma lucasanum   Carnivoros
##     Talla_.cm. Abundancia Densidad_.org.m2.
## 783         10          4              0.08
## 784         15          6              0.12
## 785         20          4              0.08
## 786         10         25              0.50
## 787         15         10              0.20
## 788         20         12              0.24
```

str(df)

```
## 'data.frame':    788 obs. of  16 variables:
##  $ Dia             : Factor w/ 4 levels "19-Mar-16","20-Mar-16",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Provincia       : Factor w/ 2 levels "Guanacaste","Puntarenas": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Latitud         : num  9.78 9.78 9.78 9.78 9.78 ...
##  $ Longitud        : num  -84.9 -84.9 -84.9 -84.9 -84.9 ...
##  $ Localidad       : Factor w/ 3 levels "CUAJINIQUIL",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ SitioCodigo     : Factor w/ 7 levels "C_MATAPALITO",..: 5 5 5 5 5 5 5 5 5 5 ...
##  $ Profundidad     : num  2 2 2 2 2 2 2 2 2 2 ...
##  $ Transecto       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ Area_.m2.       : int  50 50 50 50 50 50 50 50 50 50 ...
##  $ Phylum          : Factor w/ 1 level "Chordata": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Taxa            : Factor w/ 25 levels "Acanthuridae",..: 5 11 20 20 20 20 20 20 7 23 ...
##  $ Especie         : Factor w/ 58 levels "Abudefduf concolor",..: 25 29 40 40 2 1 55 55 18 22 ...
##  $ NivelTrofico    : Factor w/ 4 levels "Carnivoros","Herbivoros",..: 3 1 2 2 2 2 2 2 1 1 ...
##  $ Talla_.cm.      : int  20 15 25 30 10 5 10 15 15 15 ...
##  $ Abundancia      : int  5 4 2 3 3 5 6 8 1 1 ...
##  $ Densidad_.org.m2.: num  0.1 0.08 0.04 0.06 0.06 0.1 0.12 0.16 0.02 0.02 ...
```

```
range(df$Talla_.cm.)
```

[1]   5 100

```
range(df$Abundancia)
```
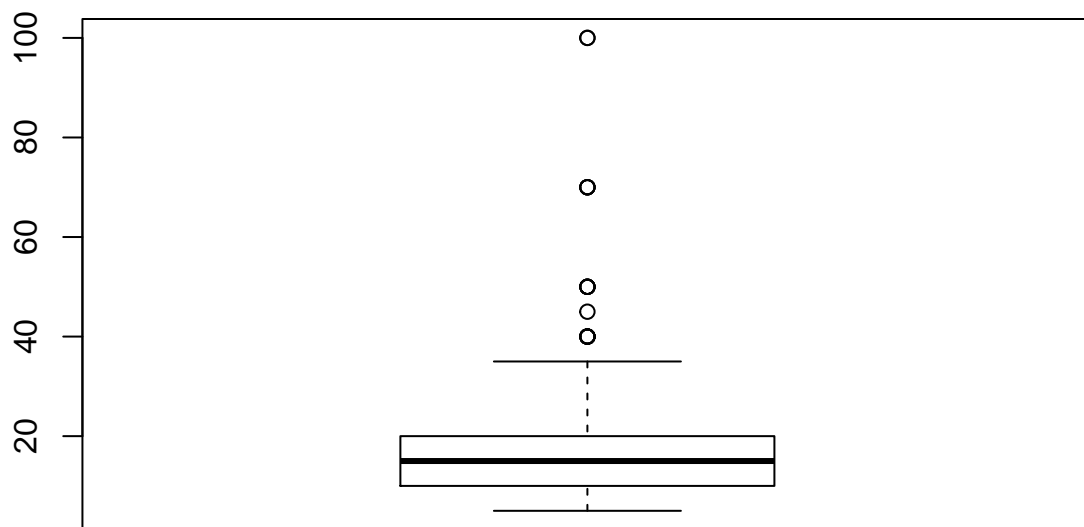
[1]   1 212

Everything looks good. Data was already in long format before importing it.
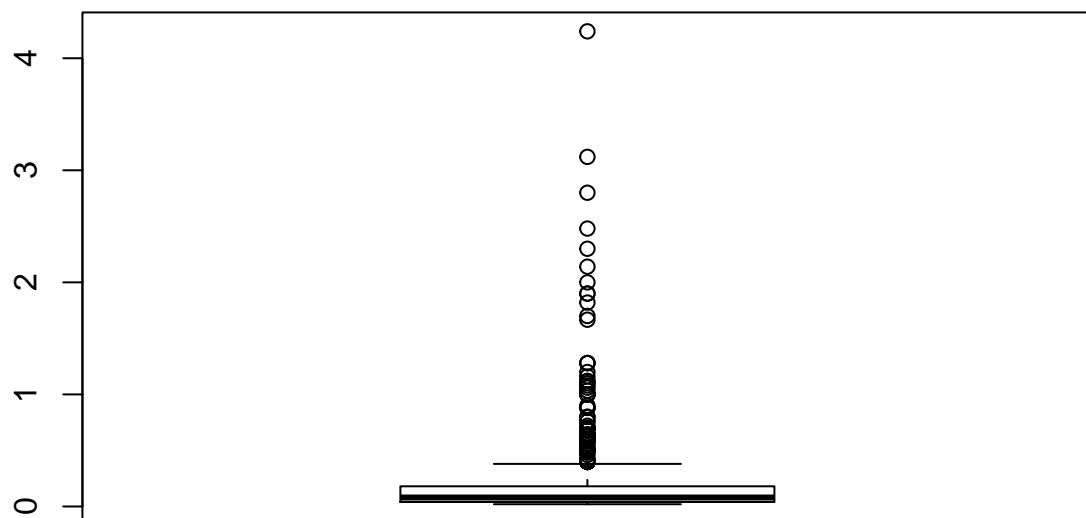
Exploring the data

We are going to see visual representation of data.
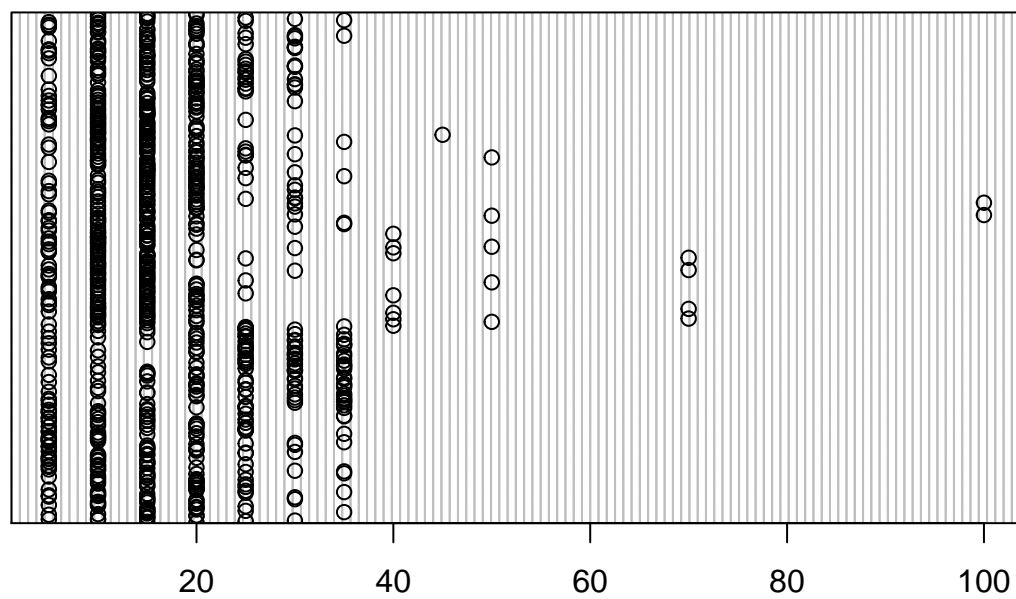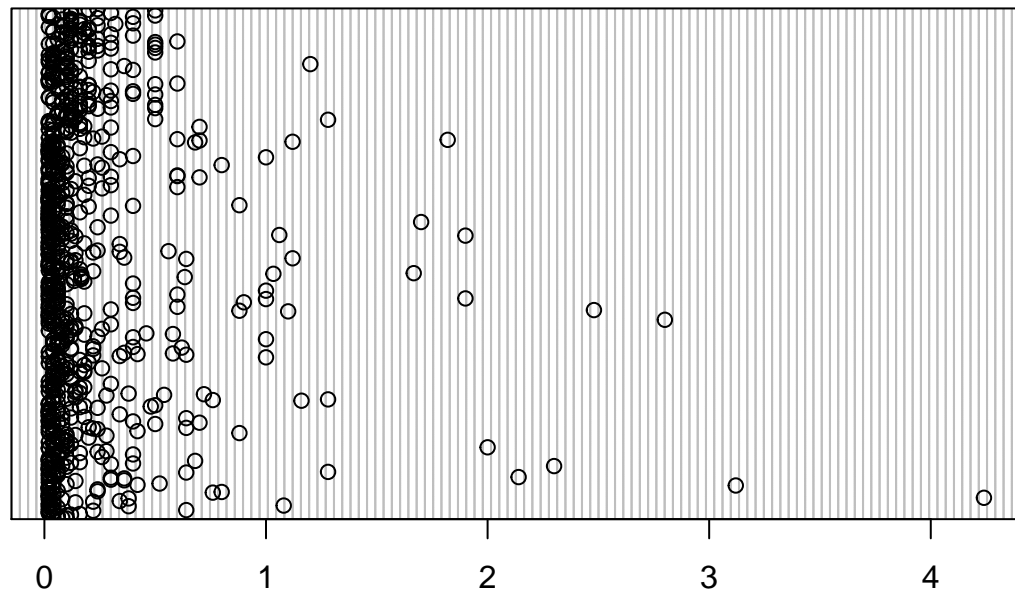
1. Outliers

```
boxplot(df$Talla_.cm.)
```



```
boxplot(df$Densidad_.org.m2.)
```
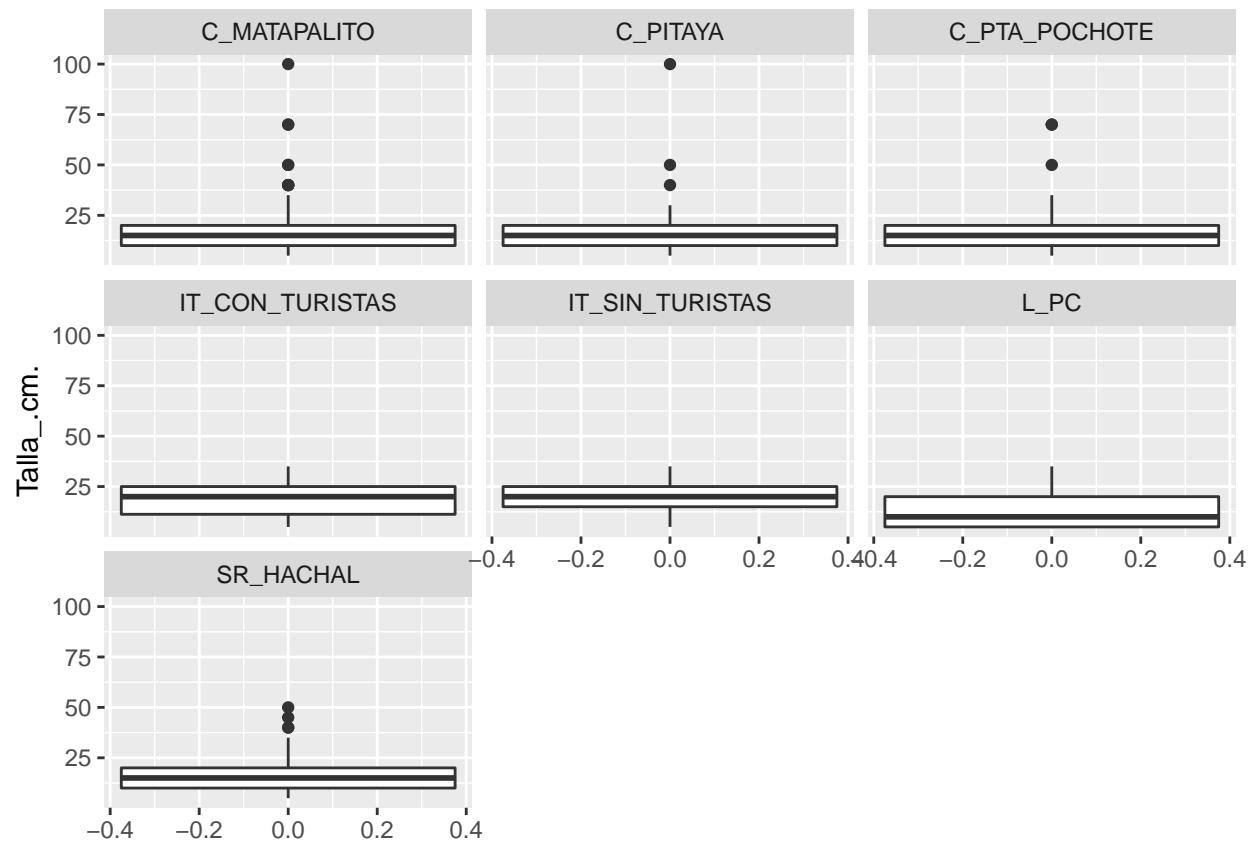
```r
dotchart(df$Talla_.cm.)
```
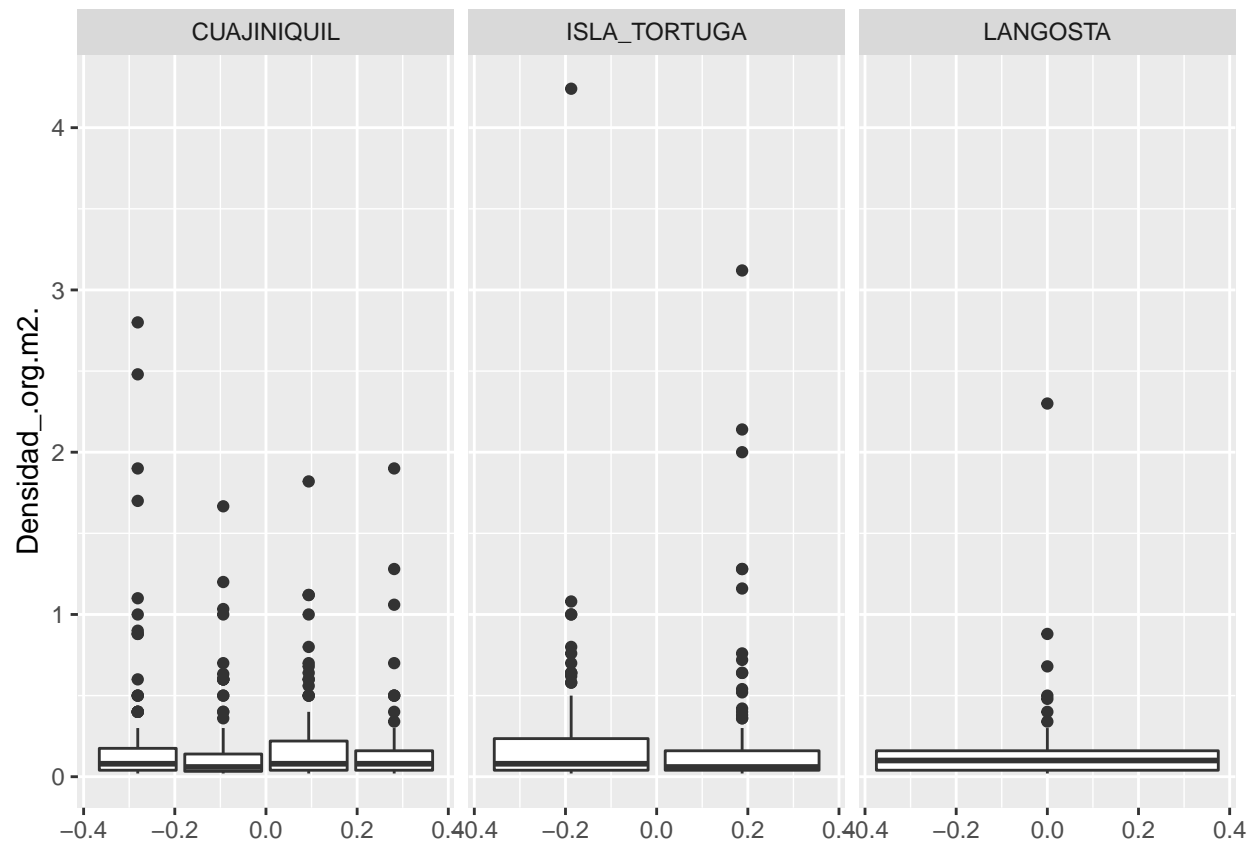
```
dotchart(df$Densidad_.org.m2.)
```

Many outliers, but there are no errors on data.

2. Homogeneity of variance

```
p <- ggplot(data=df, aes(X=Provincia, y=Talla_.cm.)) +
  geom_boxplot() + facet_wrap(~SitioCodigo)
p
```

```
p<- ggplot(data=df, aes(X=SitioCodigo, y=Densidad_.org.m2.)) +
  geom_boxplot() + facet_wrap(~Localidad)
p
```

There is homogeneity of variance.

3. Normality

```r
hist(df$Talla_.cm.)
```

Histogram of df$Talla_.cm.
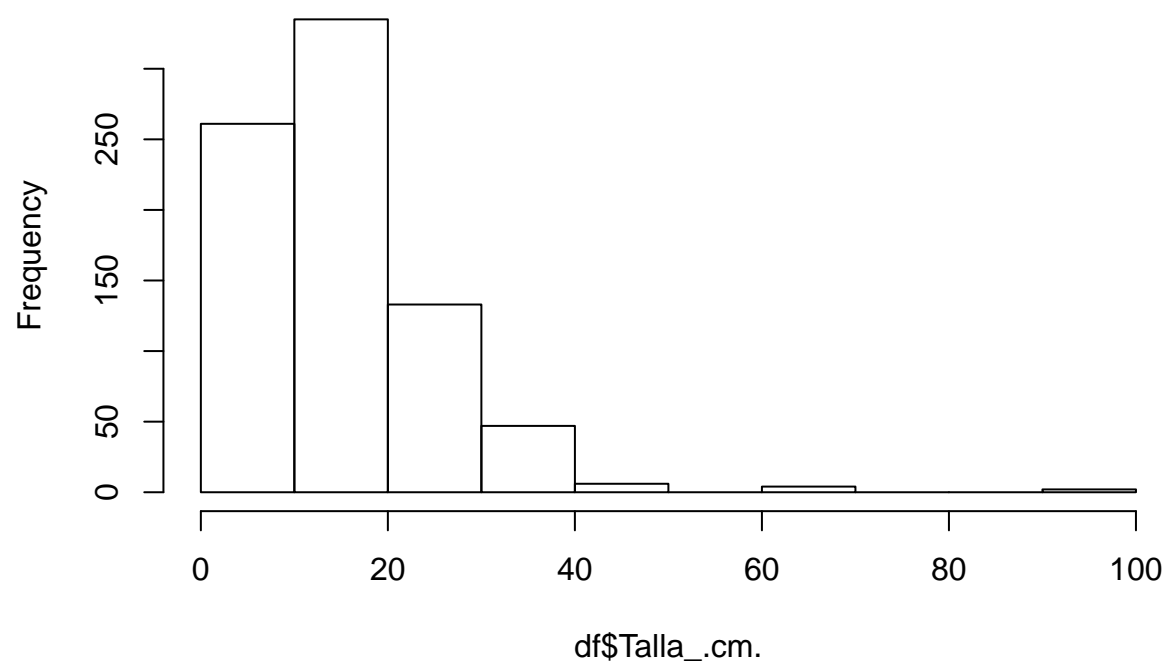
```r
hist(df$Densidad_.org.m2.)
```

Histogram of df$Densidad_.org.m2.
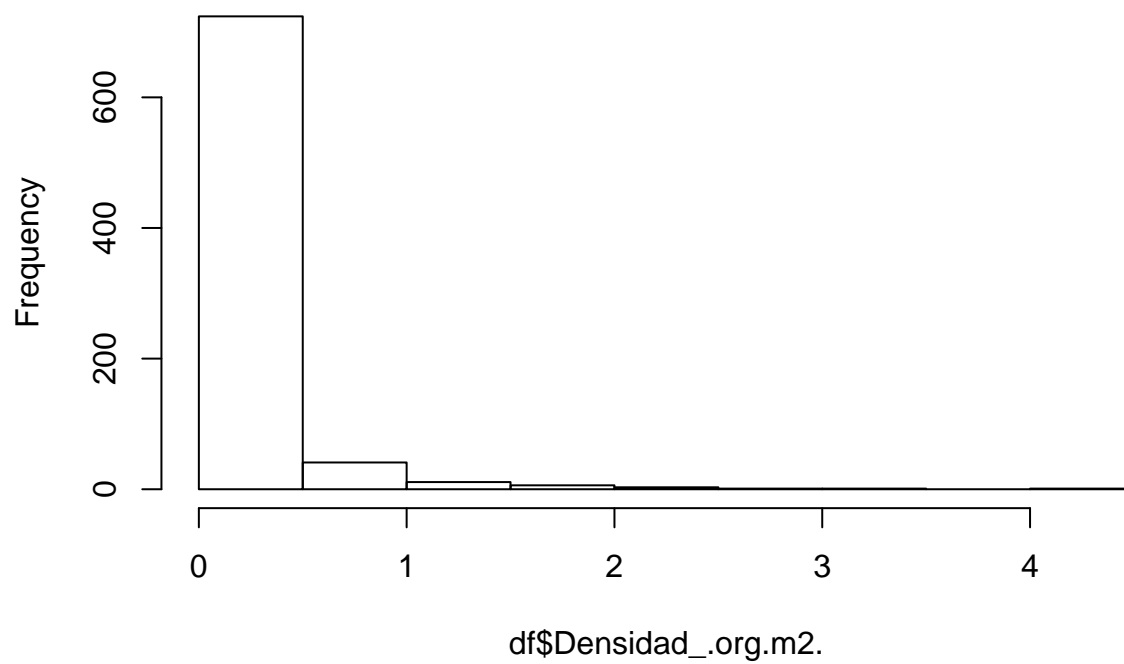

```r
qqnorm(df$Talla_.cm.)
qqline(df$Talla_.cm.)
```

Normal Q–Q Plot
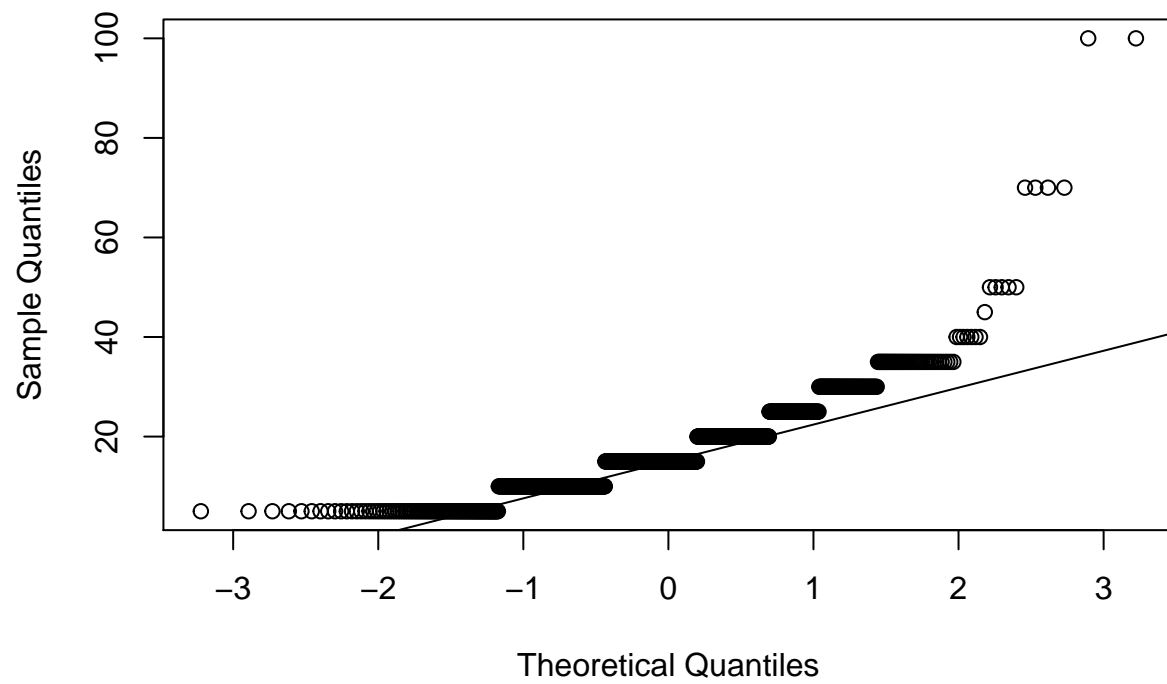

```r
qqnorm(df$Densidad_.org.m2.)
qqline(df$Densidad_.org.m2.)
```
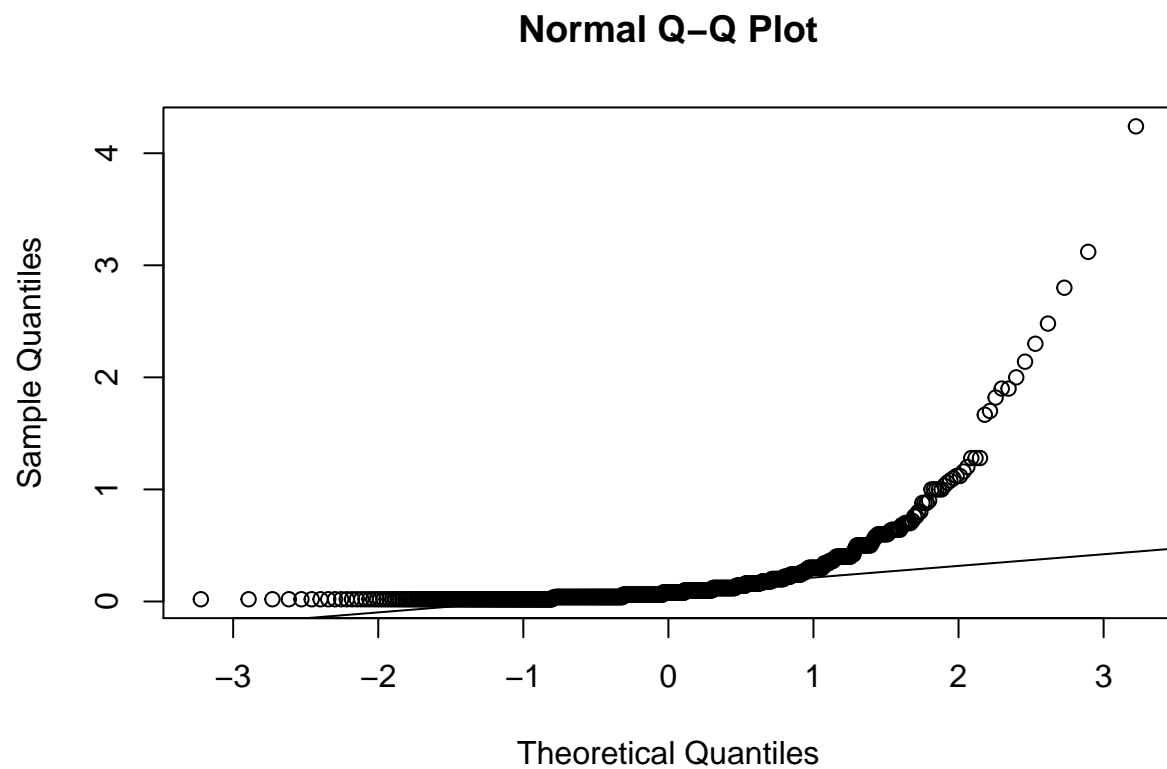
Normal Q-Q Plot

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

4. Zeroes

```r
range(df$Talla_.cm.)
```

```
## [1]   5 100
```

```r
plot(table(df$Talla_.cm.))
```
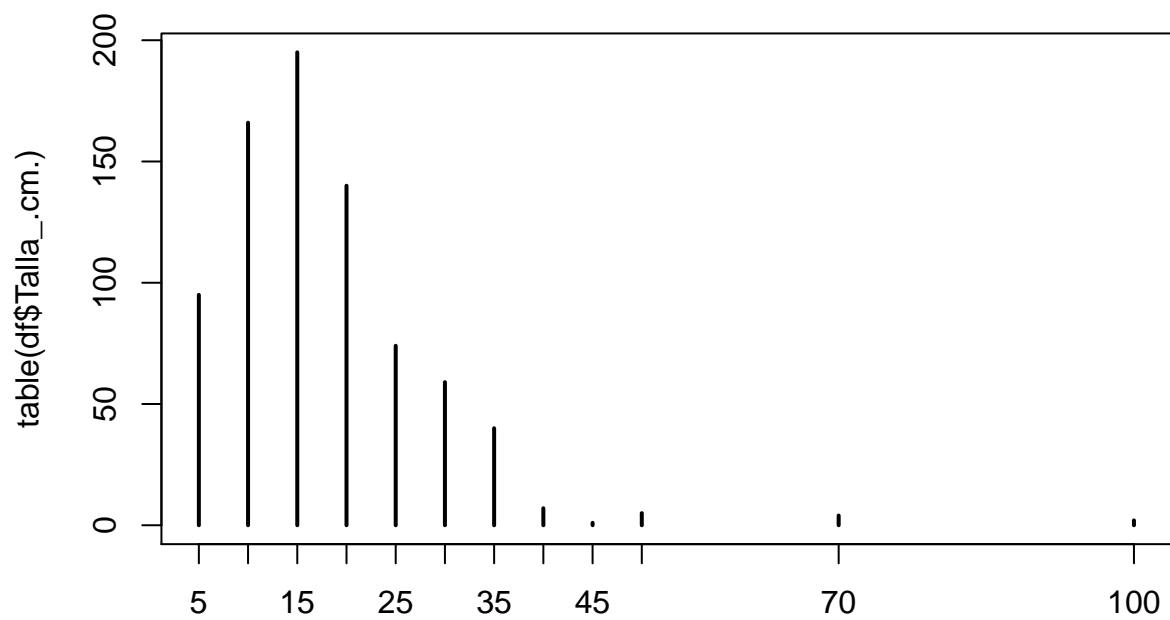
```r
range(df$Densidad_.org.m2.)
```

```
## [1] 0.02 4.24
```

```r
plot(table(df$Densidad_.org.m2.))
```
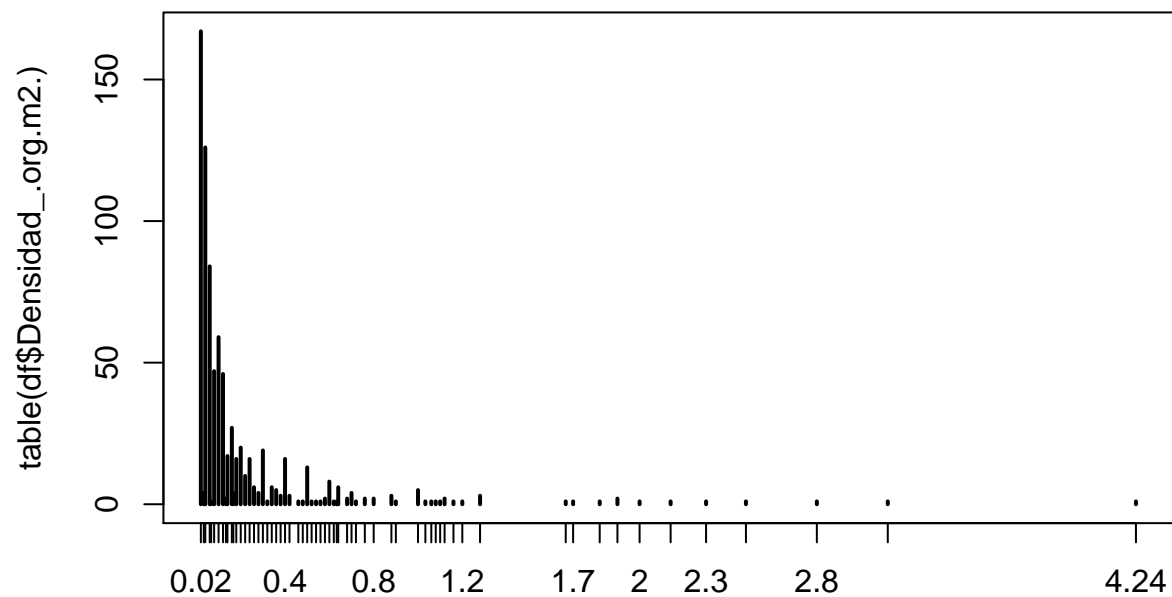
There are no zeroes in either continuous variable.

5. Collinearity X

```
pairs(~ Provincia + Localidad + SitioCodigo + Densidad_.org.m2. + Talla_.cm.,
      lower.panel=panel.smooth, upper.panel=panel.cor,
      data=df)
```
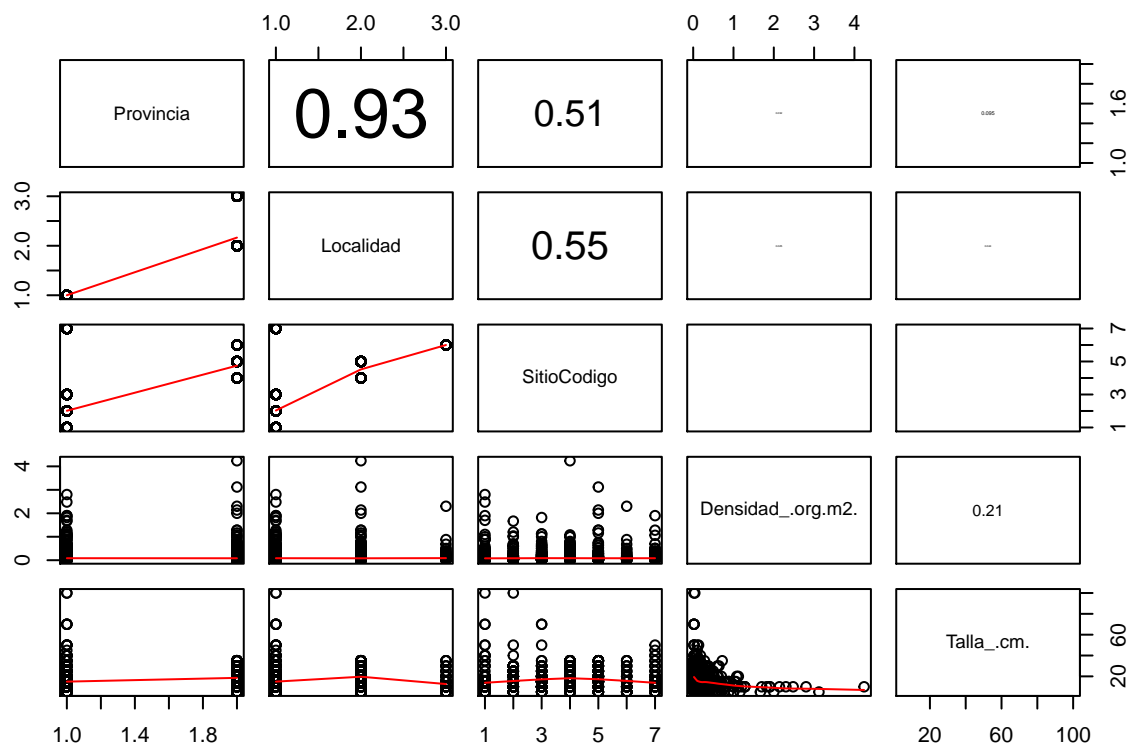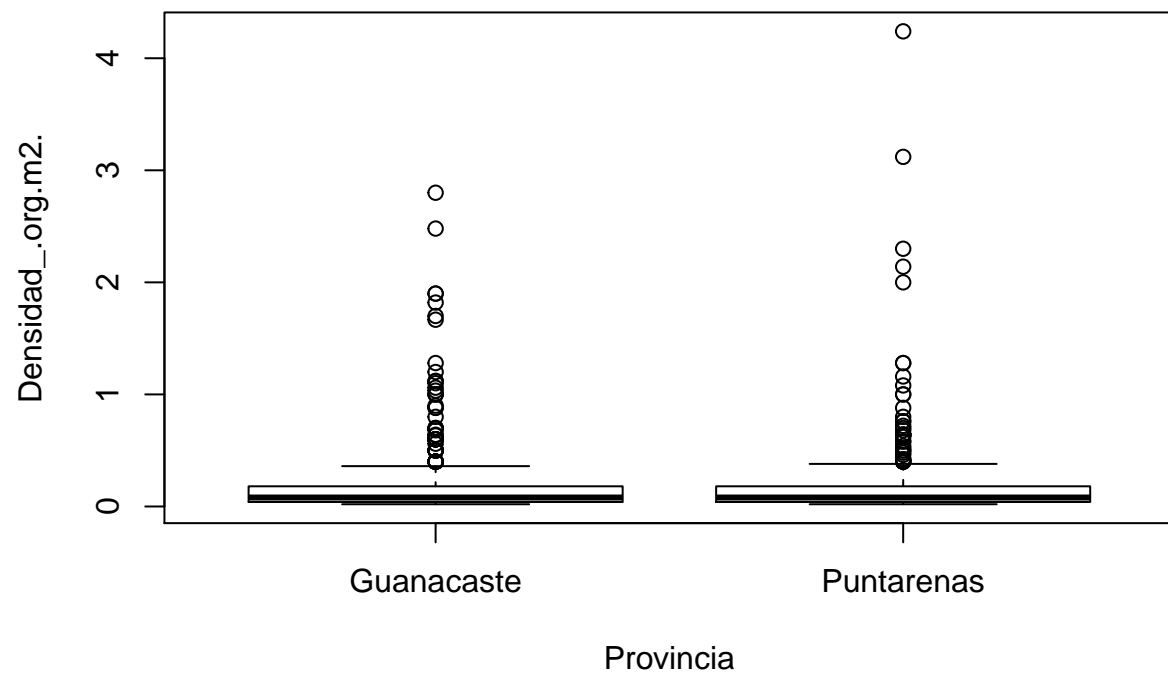
```
plot(Talla_.cm. ~ Provincia, data=df)
```
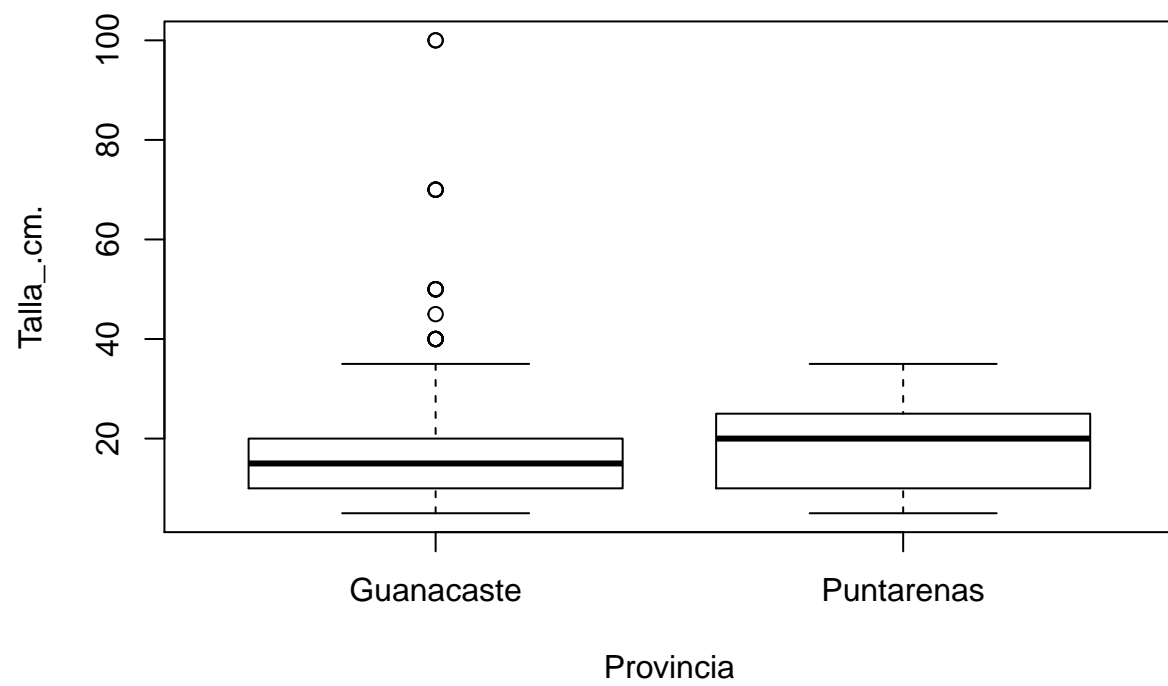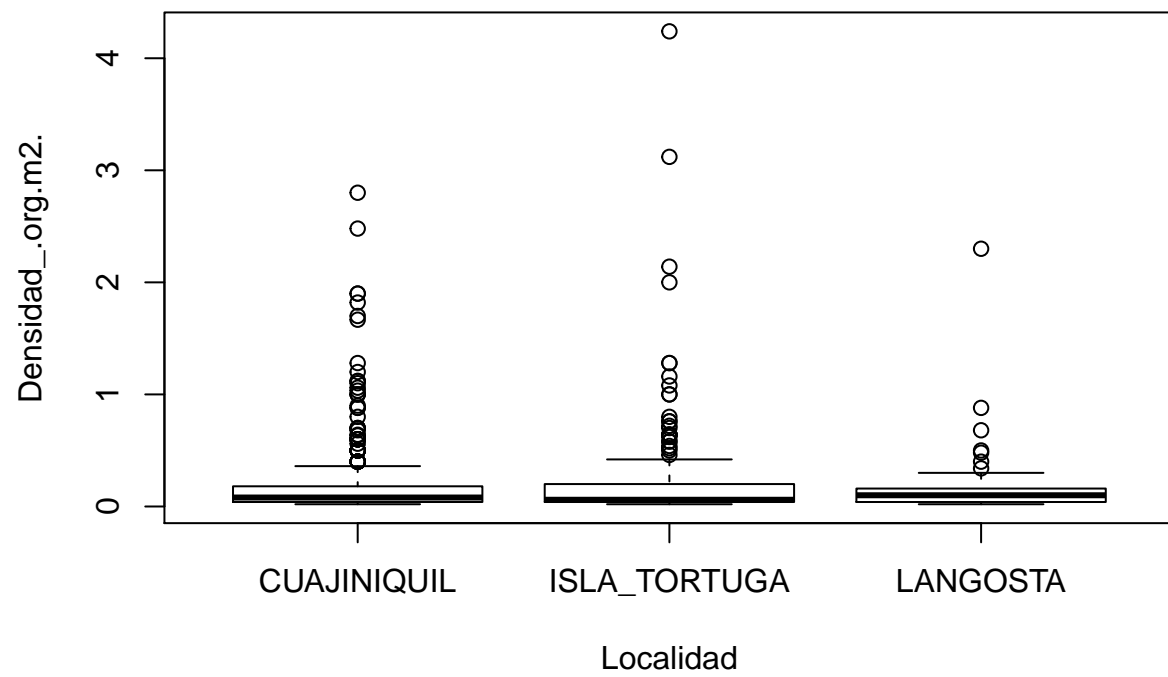
```
plot(Talla_.cm. ~ Localidad, data=df)
```
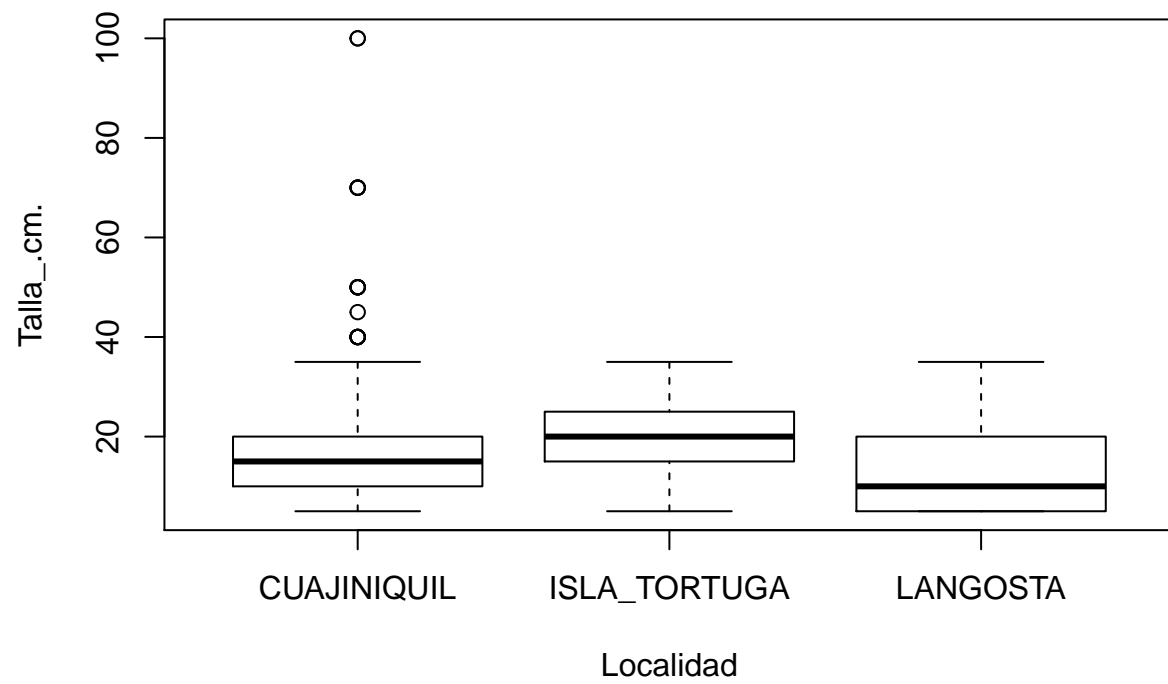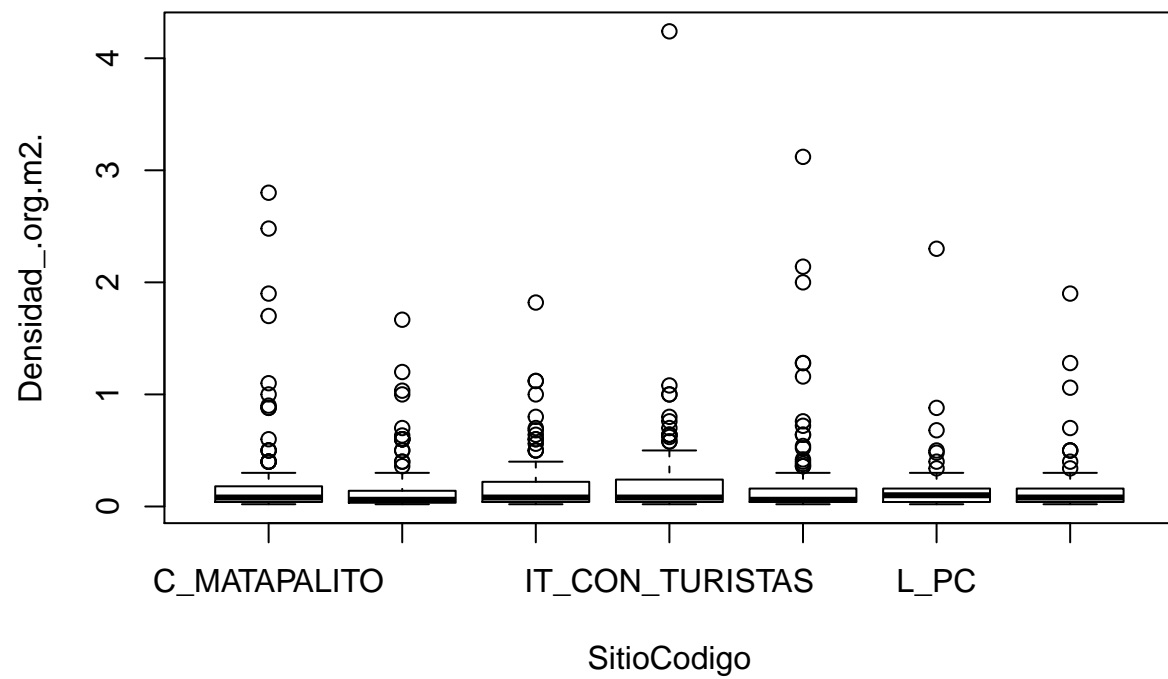
```
plot(Talla_.cm. ~ SitioCodigo, data=df)
```
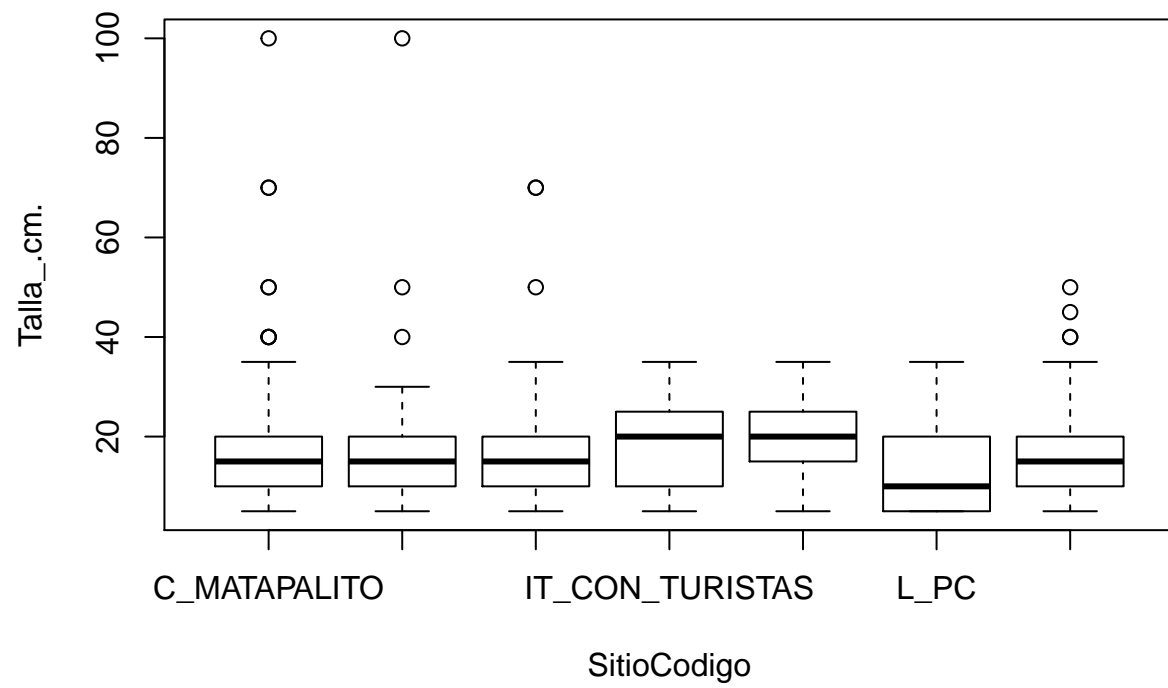
```r
plot(Densidad_.org.m2. ~ Provincia, data=df)
```
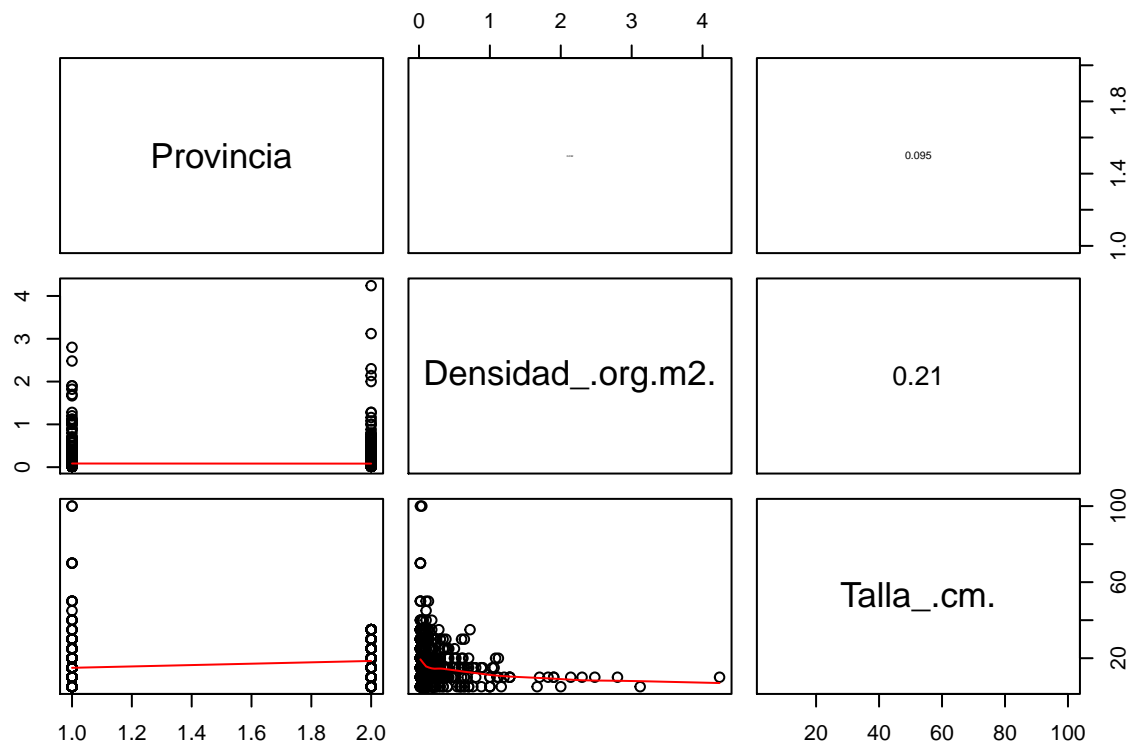
```r
plot(Densidad_.org.m2. ~ Localidad, data=df)
```

```r
plot(Densidad_.org.m2. ~ SitioCodigo, data=df)
```
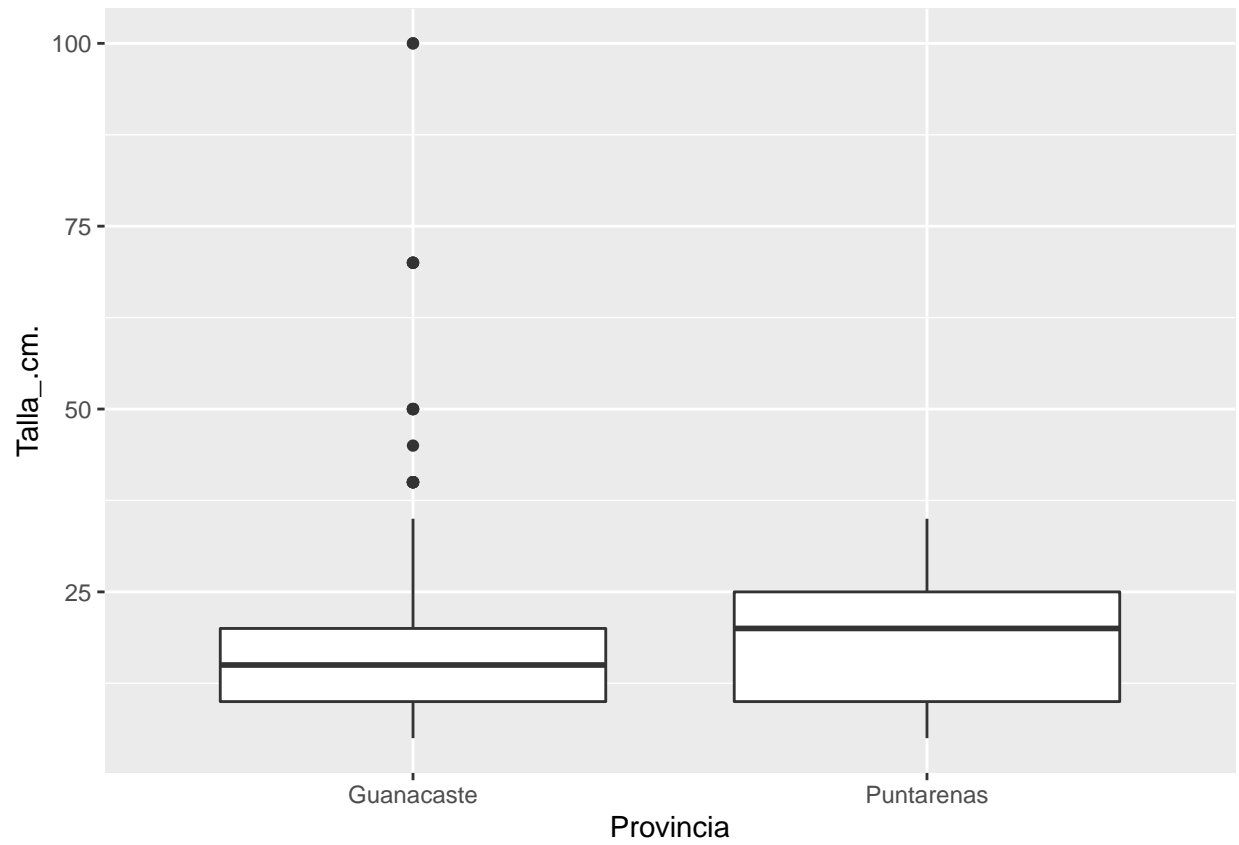
```
pairs(~ Provincia + Densidad_.org.m2. + Talla_.cm.,
      lower.panel=panel.smooth, upper.panel=panel.cor,
      data=df)
```
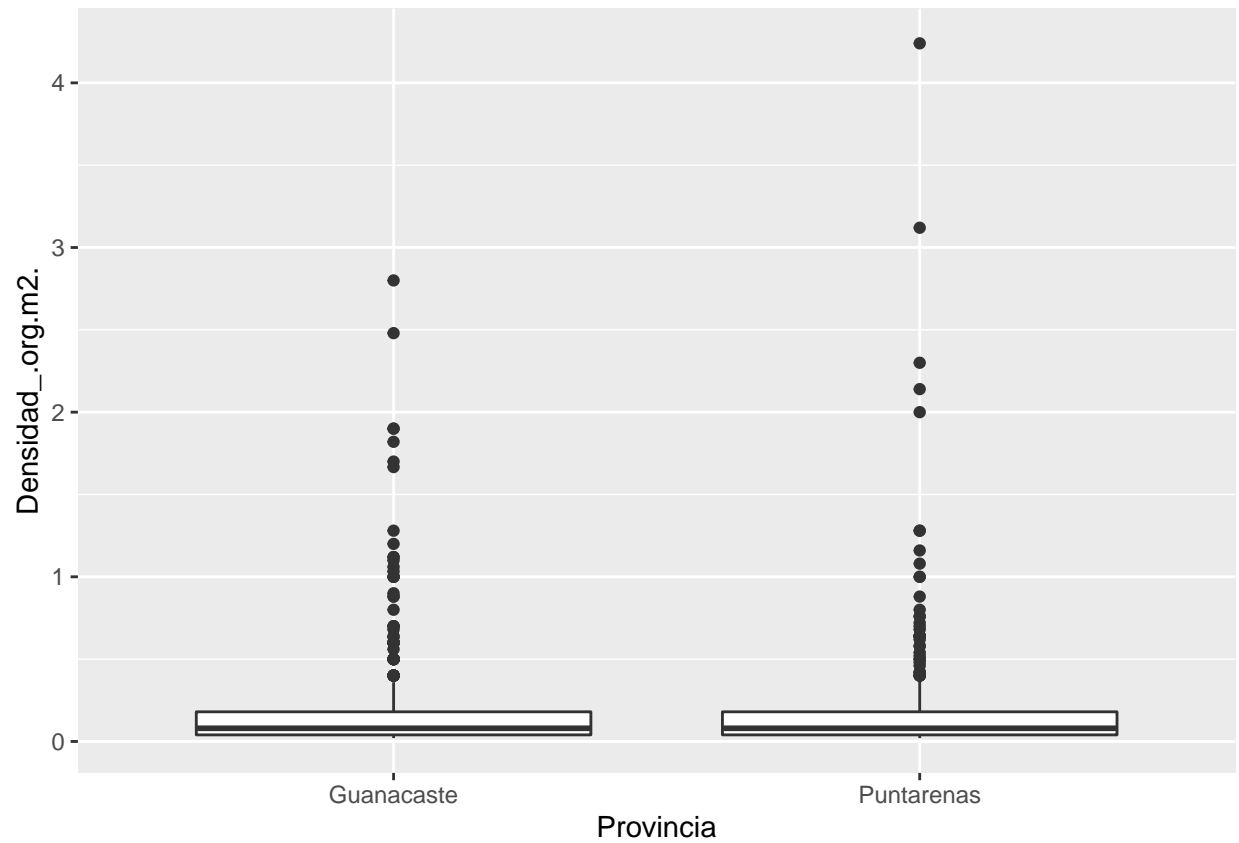
Our X variable will now be only Province

6. Relationships X and Y

```r
p <- ggplot(df, aes(x=Provincia, y=Talla_.cm.)) +
  geom_boxplot()
p
```
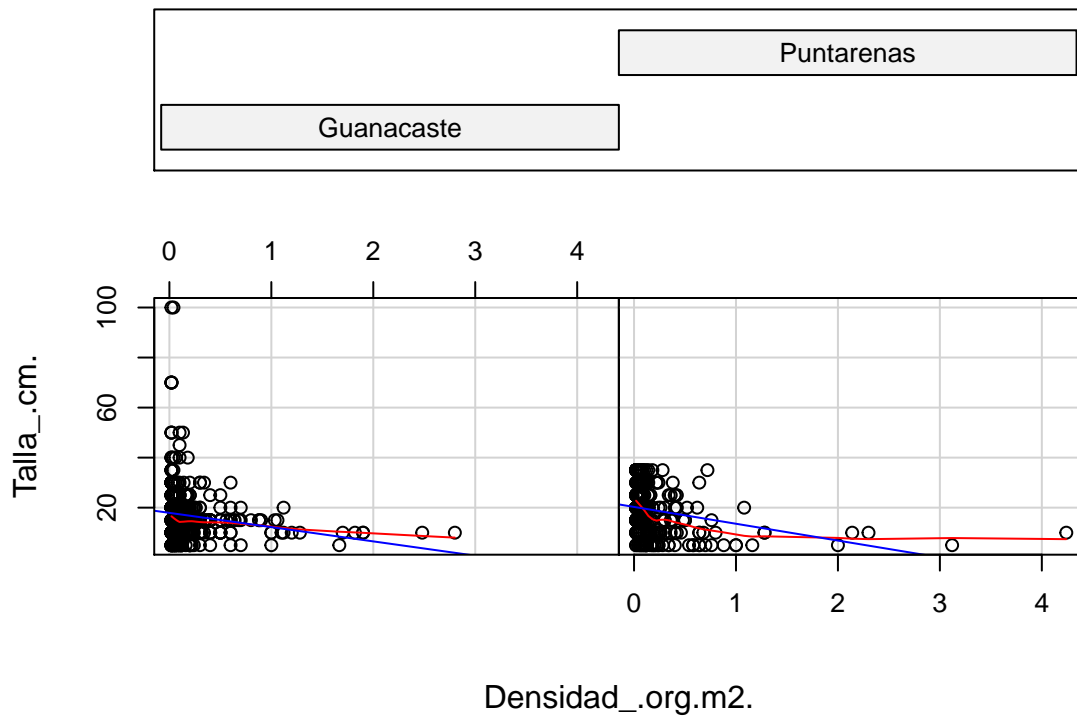
```
p <- ggplot(df, aes(x=Provincia, y=Densidad_.org.m2.)) +
  geom_boxplot()
p
```

7. Interactions

```
coplot(Talla_.cm. ~ Densidad_.org.m2. | Provincia ,
      data=df,
      panel=function(x,y,...) {
        panel.smooth(x,y,span=0.8,iter=5,...)
        abline(lm(y ~ x), col="blue") } )
```

Given : Provincia

There seems to be no interaction in Provinces.

8. Independence of Y

Both length and density could be dependent on Province because of fishing pressure effects.