R Notebook - Data exploration

Introduction

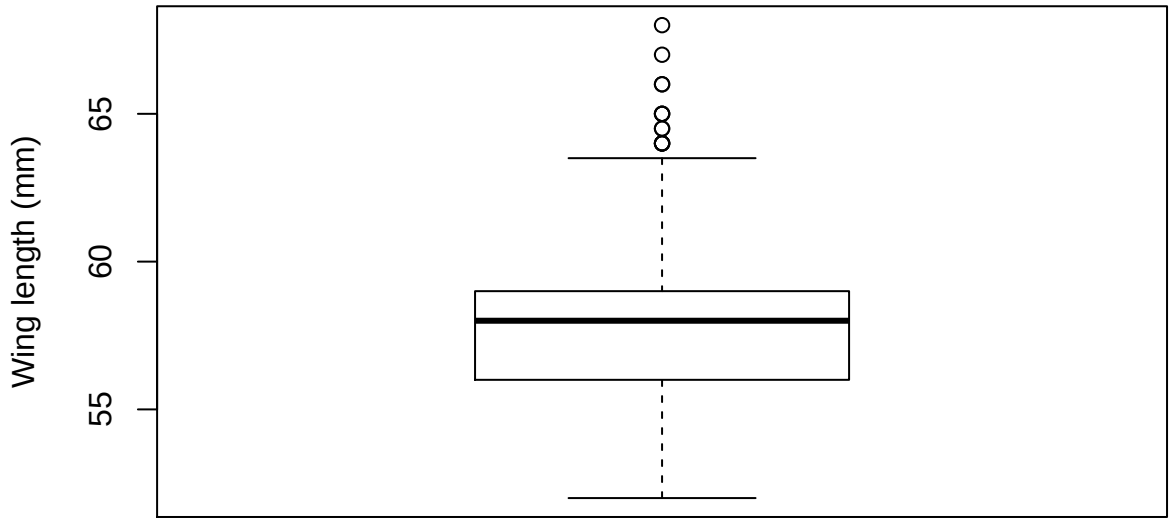
Ensuring that the scientist does not discover a false covariate effect (type I error), wrongly dismiss a model with a particular covariate (type II error) or produce results determined by only a few influential observations, requires that detailed data exploration be applied before any statistical analysis. The aim of this vignette is to provide a protocol for data exploration that identifies potential problems.

Are there outliers in Y and X?

In some statistical techniques the results are dominated by outliers; other techniques treat them like any other value. For example, outliers may cause overdispersion in a Poisson GLM. We define an outlier as an observation that has a relatively large or small value compared to the majority of observations.

Boxplot visualizes the median and the spread of the data. The median is typically presented as a horizontal line with the 25% and 75% quartiles forming a box around the median that contains half of the observations.

```
boxplot(sparrows$Wingcrd,
        ylab = "Wing length (mm)")
```



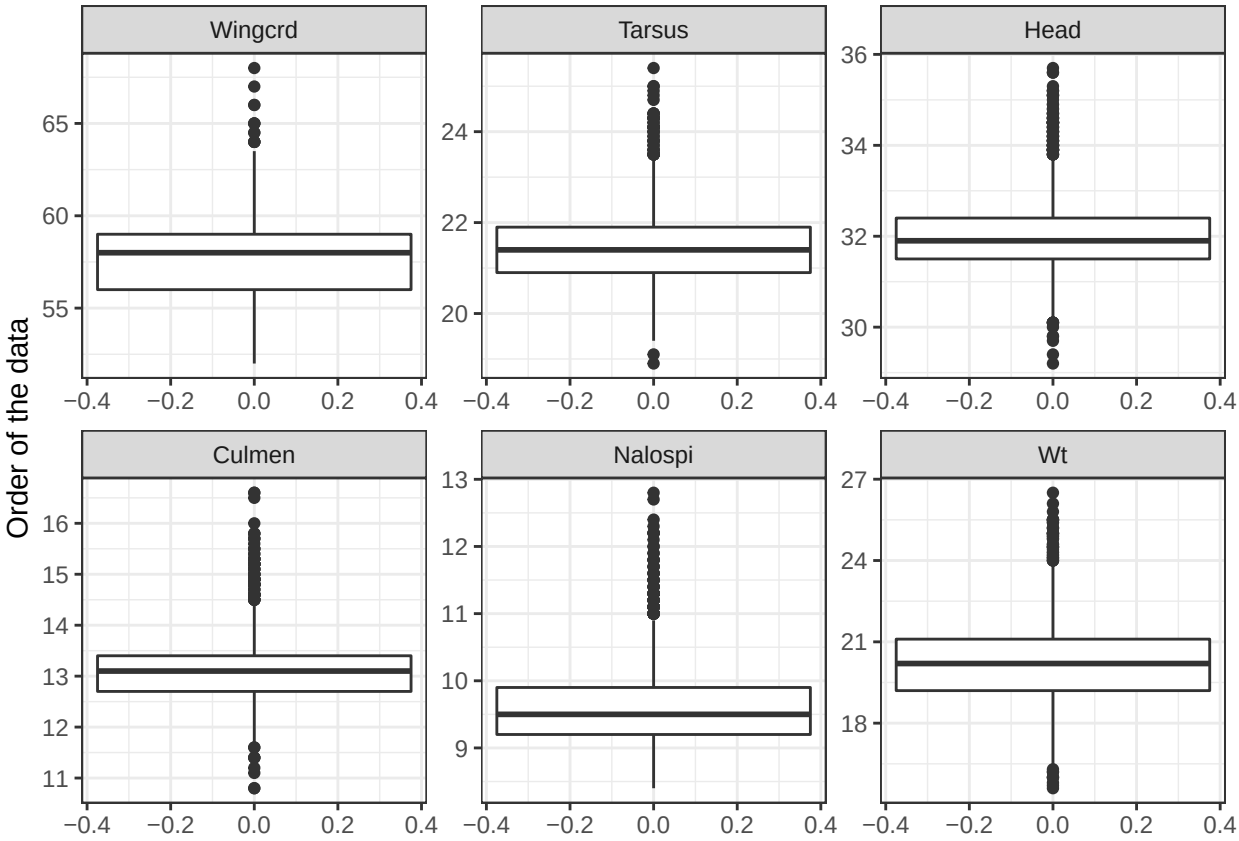
We can create a multi-panel boxplot to explore all variables in a data set. To do this when the data are in a matrix format (a row for each observation, a column for each variable) we first need to reshape or `melt` the data into long format (one column for the variable name, one column for the variable value) - we specify ID variables; variables that identify individual rows of data:

```
library(reshape2)
library(dplyr)
slf <- melt(sparrows,
            variable.name="var",
            id.vars = c("Species", "Sex"))
unwanted.vars <- c("Observer","Age")
slf <- slf %>% filter(!var %in% unwanted.vars)
head(slf)
```

```
##   Species    Sex     var value
## 1    SSTS   Male Wingcrd  58.0
## 2    SSTS Female Wingcrd  56.5
## 3    SSTS   Male Wingcrd  59.0
## 4    SSTS   Male Wingcrd  59.0
## 5    SSTS   Male Wingcrd  57.0
## 6    SSTS Female Wingcrd  57.0
```

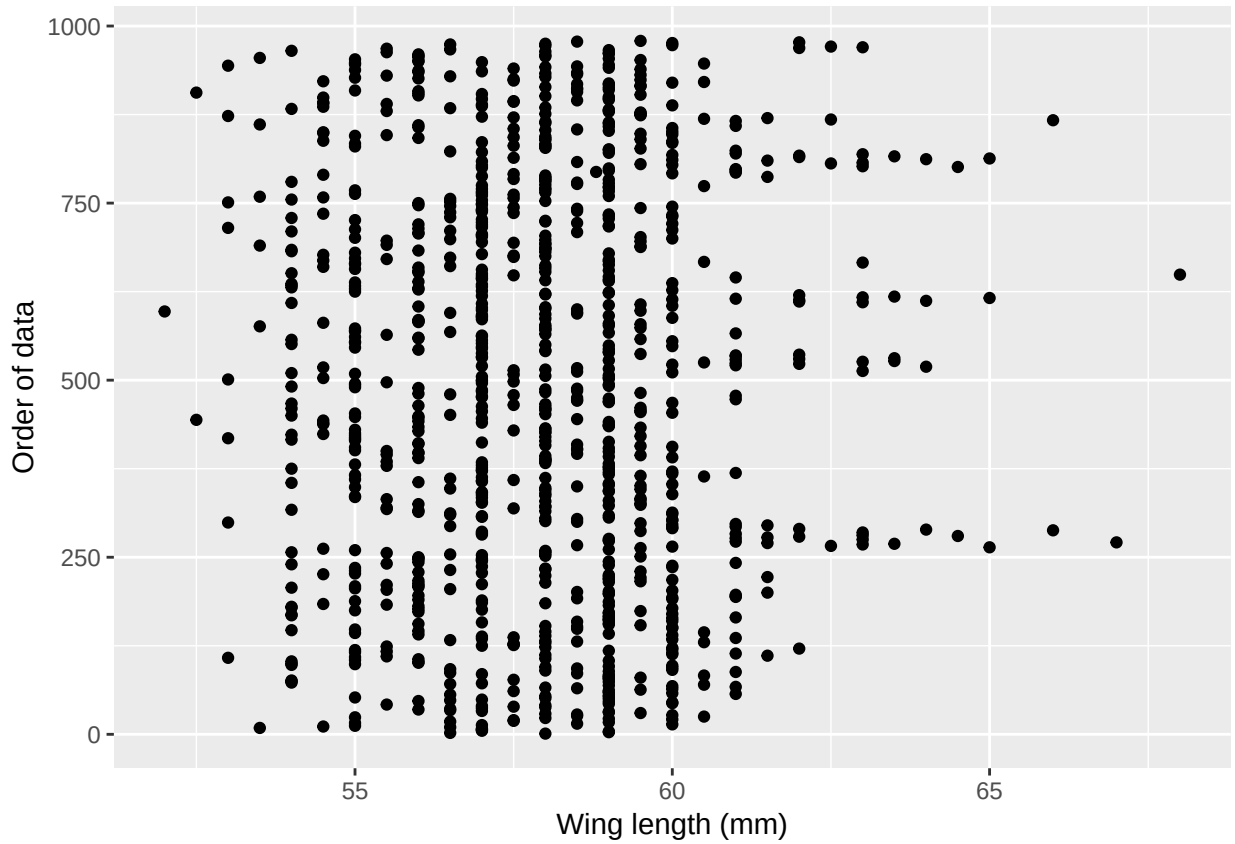
Now the data is in a suitable structure, the plot is generated as follows:

```
ggplot(slf, aes( y = value)) +
  geom_boxplot() +
  facet_wrap(.~ var, scales = "free") +
  labs(y = "Order of the data") +
  theme_bw()
```



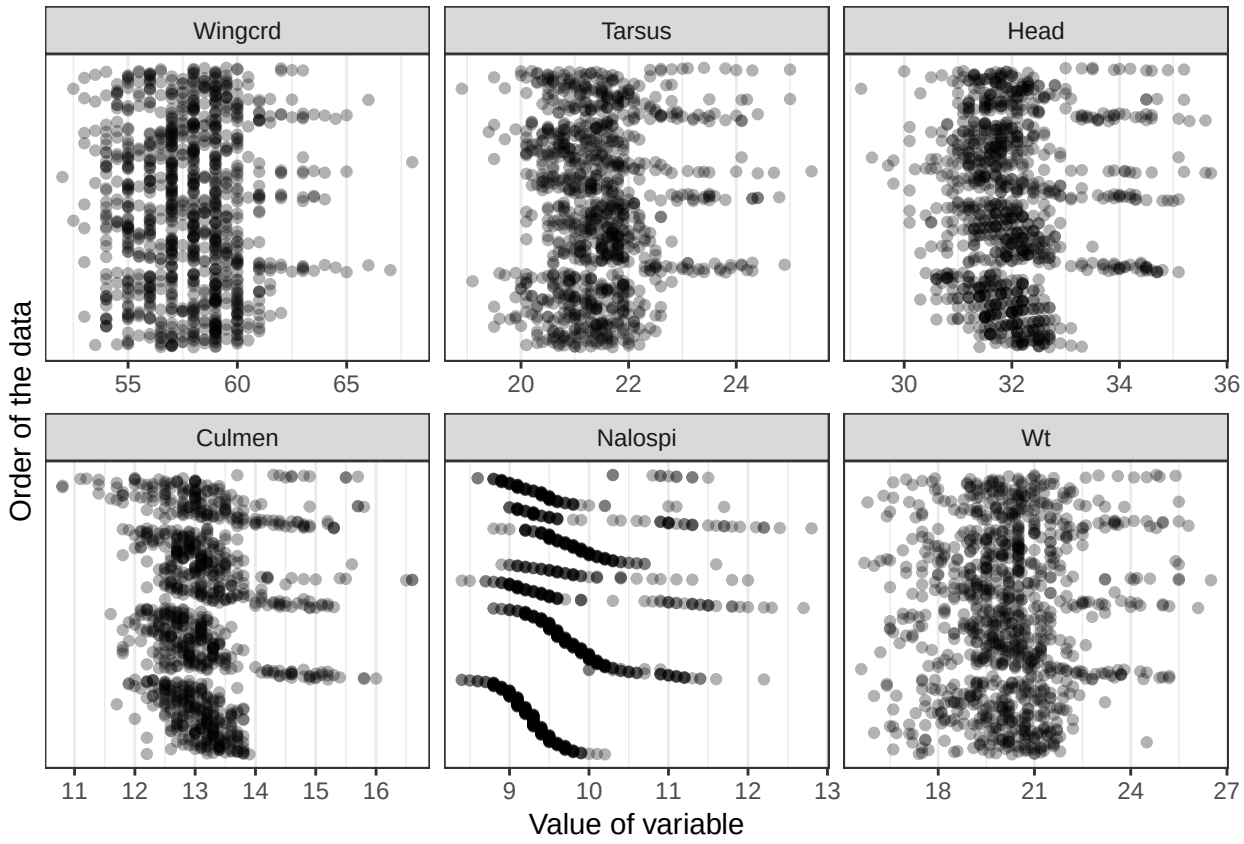
Cleveland dotplot a graph in which the row number of an observation is plotted vs. the observation value, thereby providing much more detailed information than a boxplot. Points that stick out on the right-hand side, or on the left-hand side, are observed values that are considerable larger, or smaller, than the majority of the observations, and require further investigation.

```
ggplot() +
  geom_point(aes(x=sparrows$Wingcrd,
                 y=as.numeric(row.names(sparrows)))) +
  labs(x = "Wing length (mm)",
       y = "Order of data")
```



We can also geneate a panel view of Cleveland dotplots for each of the variables in the dataset:

```
ggplot(slf, aes( x = value, y = as.numeric(row.names(slf)))) +
  geom_point(alpha=0.3) +
  facet_wrap( "var", scales = "free" ) +
  labs(x = "Value of variable",
       y = "Order of the data") +
  scale_y_continuous(breaks = NULL) +
  theme_bw()
```



Do we have homogeneity of variance?

Homogeneity of variance is an important assumption in analysis of variance (ANOVA). The following data is a record of components of the visual system

Implications

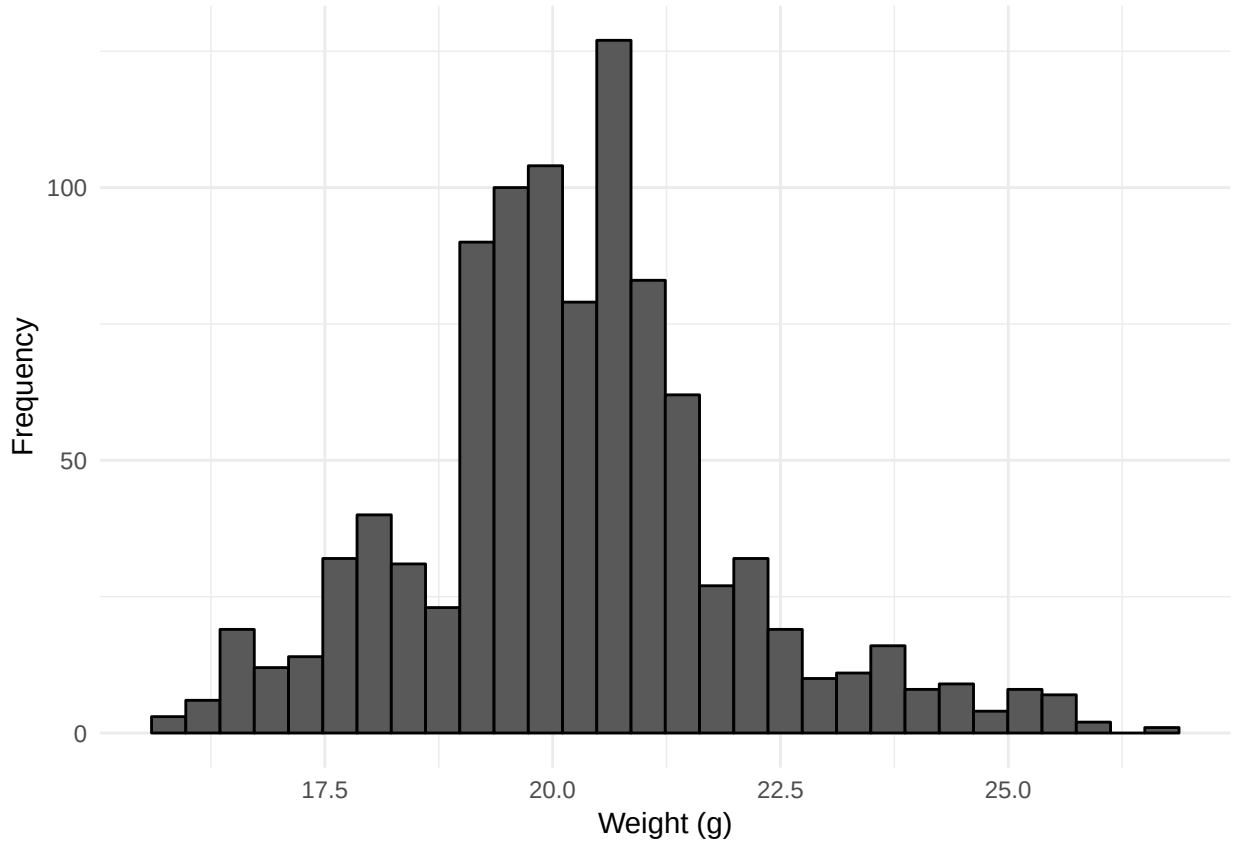
In regression-type models, verification of homogeneity should be done using the residuals of the model; i.e. by plotting residuals vs. fitted values, and making a similar set of conditional boxplots for the residuals. In all these graphs the residual variation should be similar. The solution to heterogeneity of variance is either a transformation of the response variable to stabilize the variance, or applying statistical techniques that do not require homogeneity (e.g. generalized least squares)

Are the data normally distributed?

A significant number of statistical modelling tools assuming the data are normally distributed. For example, linear regression does assume normality. The following plot shows a histogram for the weight of 1,193 sparrows. It can be seen that the distribution is significantly skewed:

```
ggplot(sparrows, aes(x = Wt)) +
  geom_histogram(color = "black") +
    theme_minimal() +
```

```
    labs(x = "Weight (g)",
         y = "Frequency")
```

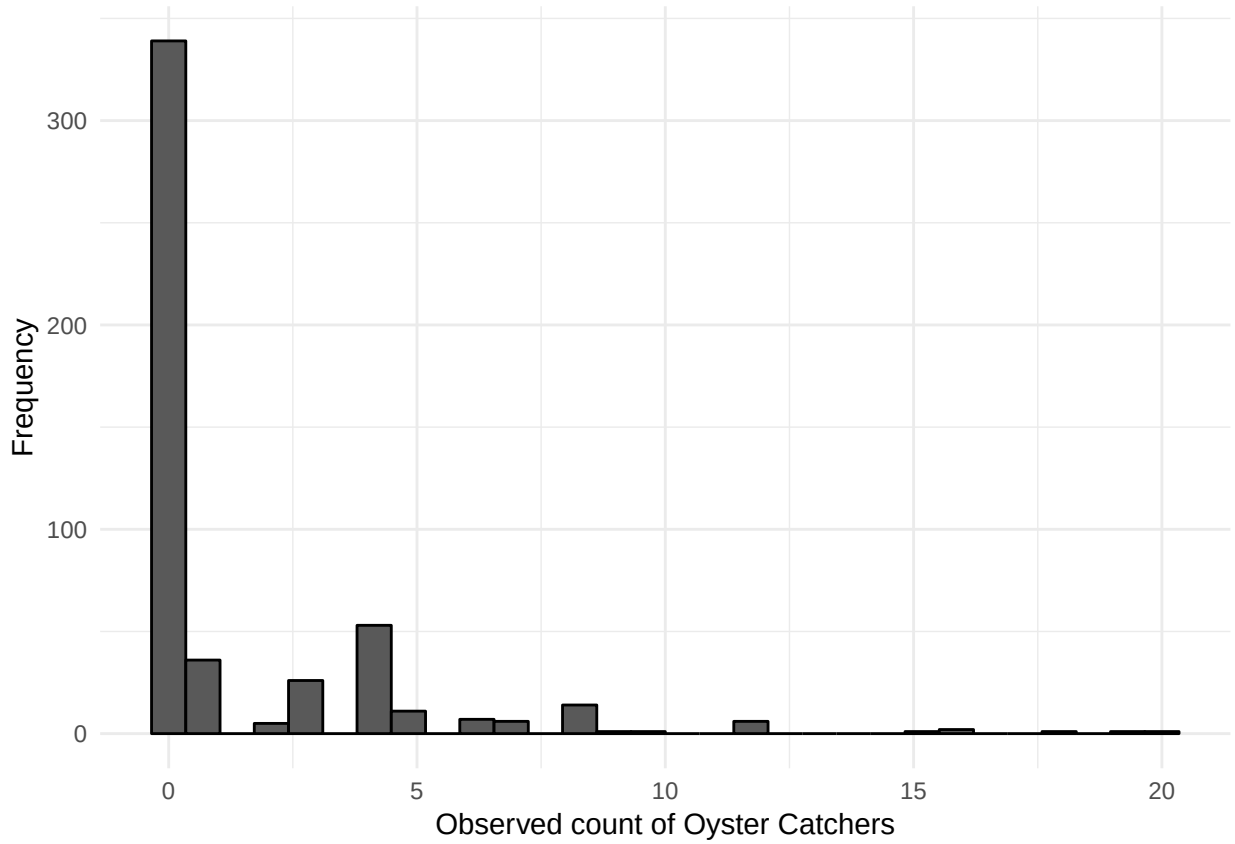


If we plot histograms for the weights observed, broken down by month, we see the cetnre of the distribution shifting by month. This indicates that sparrow weight is dependently on monthly food availability. Under these circumstances, it would not be advisable to transform the data to make it normal.

Are there a significant number of zeros in the data?

The following plot is a frequency plot showing how often each value for total wader abundance occured during the survey.

```
ggplot(waders, aes(x = OC)) +
  geom_histogram(color = "black") +
  theme_minimal() +
    labs(x = "Observed count of Oyster Catchers",
         y = "Frequency")
```



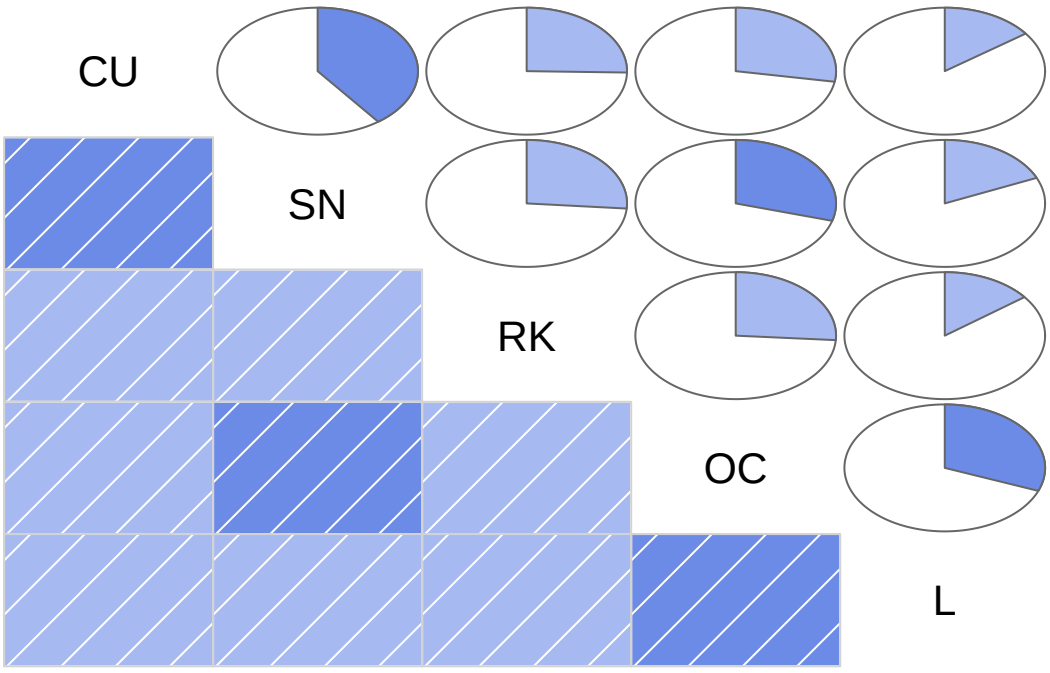
The extremely high number of zeros tells us that we should not apply an ordinary Poisson or negative binomial GLM as these would produce biased parameter estimates and standard errors. Instead one should consider zero inflated GLMs.

Implications for species abundance

```
library(corrgram)
```

```
## Registered S3 method overwritten by 'seriation':
##   method         from
##   reorder.hclust gclus
```

```
waders %>%
  select(c("CU","L","OC","RK","SN")) %>%
  na.omit() %>%
  mutate_all(function(x) as.numeric(as.character(x))) %>%
  corrgram(order=T, upper.panel=panel.pie)
```


CU
SN
RK
OC
L