Exercises

Exercise: Graphical data exploration using R

1. As in previous exercises, either create a new R script (perhaps call it graphical_data_exploration) or continue with your previous R script in your RStudio Project. Again, make sure you include any metadata you feel is appropriate (title, description of task, date of creation etc) and don't forget to comment out your metadata with a `#` at the beginning of the line.

2. If you haven't already, download the data file *'loyn.xlsx'* from the **Data** link and save it to the `data` directory you created during exercise 1. Open this file in Microsoft Excel (or even better use an open source equivalent - LibreOffice is a good free alternative) and save it as a tab delimited file type. Name the file *'loyn.txt'* and also save it to the `data` directory.

3. These data are from a study originally conducted by Loyn (1987)[1] and subsequently re-analysed by Quinn and Keough (2002)[2] and Zuur et al (2009)[3]. The aim of the study was to relate bird density in 56 forest patches to a number of different environmental variables and management practices. A summary of the variables is: **ABUND**: Density of birds, Continuous response; **AREA**: Size of forest patch, Continuous explanatory; **DIST**: Distance to nearest patch, Continuous explanatory; **LDIST**: Distance to nearest larger patch, Continuous explanatory; **ALTITUDE**: Mean altitude of patch, Continuous explanatory; **YEAR.ISOL**: Year of isolation of clearance, Continuous explanatory; **GRAZE**: Index of livestock grazing intensity, 5 level Categorical explanatory 1= low graze, 5 = high graze

4. Import your *'loyn.txt'* file into R using the `read.table()` function and assign it to a variable called `loyn`. Use the `str()` function to display the structure of the dataset and the `summary()` function to summarise the dataset. How many observations are in this dataset? How many variables? How is the variable `GRAZE` coded? (as a number or a string?). If you think this will cause a problem (hint: it will!), create a new variable called `FGRAZE` **in** the dataframe with `GRAZE` recoded as a factor.

5. Use the function `table()` (or `xtabs()`) to determine how many observations are in each `FGRAZE` level. See section 3.5 of the Introduction to R book to remind yourself how to do this.

6. Using the `tapply()` function what is the mean bird abundance (`ABUND`) for each levels of `FGRAZE`? Can you determine the variance, the minimum and maximum for each `FGRAZE` level? Again see section 3.5 of the Introduction to R book to remind yourself how to do this.

7. Now onto some plotting action. Use a Cleveland dotchart of each variable separately to assess whether there are any outliers in the response variable (`ABUND`) or any of the explanatory variables (see table above)? Produce a Cleveland dotchart of each variable separately to assess this (hint: use the `dotplot()` function). If you feel in the mood, then output these plots to an external PDF file.

8. If you do spot any unusual observations have a think about what you want to do with them (NOTE: do **not** just remove them without justification!). If you're unsure, be sure to speak to an instructor to discuss your options during our synchronous practical sessions. Perhaps you should you apply a data transformation to see if this reduces the magnitude of any outlier. The best thing to do here is to play around with different transformations (i.e. `log`, `sqrt`) to see which one does what you want it to do. After you have applied these data transformations make sure you re-plot your dotcharts with any transformed variable to double check what the transformation is doing.

9. Is there any potential collinearity between any of the explanatory variables? Plot these variables using the `pairs()` function. Remember, you only need to check for collinearity between your explanatory variables so you will need to extract these variables from the `loyn` dataframe either before you use the `pairs()` function or whilst using it. Optionally, include the correlation coefficient between variables in the upper panel of the pairs plot (see section 4.2.5 of the introduction to R book for details).

10. Are there any clear relationships between the response variable (`ABUND`) and individual explanatory variables? Use the appropriate plotting functions (`plot()`, `boxplot()`) to visualise these relationships.

11. One of the main aims of this study was to determine whether management practices such as grazing intensity (`GRAZE`) and size of the forest (`AREA`) affected the abundance of birds (`ABUND`). One hypothesis was that the size of the forest impacted birds, but this was dependent of the intensity of the grazing regime (in other words, there is an interaction between `AREA` and `GRAZE`). Use an appropriate plotting function to explore these data for such an interaction (perhaps a `coplot()` or `xyplot()` might be helpful?).

[1] Loyn, R. (1987). Effects of patch area and habitat on bird abundances, species numbers and tree health in fragmented Victoria forests.. Nature conservation: the role of remnants of native vegetation. 65-77.

[2] Quinn, G. P., and Michael J. Keough. 2002. Experimental design and data analysis for biologists. Cambridge, UK: Cambridge University Press.

[3] Zuur, A.F., Ieno, E.N. and Elphick, C.S. (2010), A protocol for data exploration to avoid common statistical problems. Methods in Ecology and Evolution, 1: 3-14. doi:10.1111/j.2041-210X.2009.00001.x

End of Graphical data exploration using R Exercise