Data Processing and Visualisation Challenge Practical

The best way to learn to work with data in R is to do it. One approach could be to give you a problem, then tell you the commands you need to solve it, but that's very passive and doesn't help you learn to apply this code to new problems. Instead, we give you a set of challenges of gradually increasing difficulty that you must actively figure out how to solve. This is much harder, but you learn so much more – we promise!

There are two sets of tasks, for two different datasets. You should complete the tasks for each dataset in order, as the tools you used for earlier tasks will become handy later. However, you can tackle the datasets in any order you choose, or jump between them if you like.

Under each task there are some hints in white text: highlight to get the hint. There is rarely one right answer for each task. Some questions require you to summarise data before plotting it: before you try to work on any task, try to think about whether this question is asking something about the individual sample units, or a general question about the dataset as a whole. As a rule, the latter requires some level of summarising, while the former does not. For plotting tasks, think about what sort of plot would be appropriate:

- If you have a single value that summarises each level of a category, a bar chart is often appropriate
- If you have multiple values for each level of a category, then box plots would usually be your first choice
- If you want to find out the distribution of the values of a numeric variable, i.e. the frequency of each value or range of values in the dataset, a histogram is the way to go
- To show the values of two numeric variables against one another, try a scatter plot first

If you want to do one of the above plots, but you have another variable that also needs to be illustrated, you could colour the items in the plot by different levels or values of that variable.

Feel free to work alone or in pairs, and to ask each other for advice and discuss your solutions. We're also very happy to help and give suggestions and help you understand anything you're stuck on. You're not expected to work alone and in silence!

Dataset 1: Dung beetles

This dataset looks at dung beetle communities in a rainforest national park. Dung beetles were sampled at 133 sites over the park, with each site visited several times. We are interested in what drives variation in the abundance of different species, and in the dung beetle community as a whole. Thus, a single sample unit is a specific site, collected on a specific day. Each site was also surveyed for habitat characteristics, and the GPS location taken. Finally, remote sensed data was used to find topographical and climatic information out about the park as a whole.

The data is provided as an excel file that contains the three datasets as separate sheets, along with a sheet explaining the columns.

Tasks:

1. What is the total number of individuals collected for each species across the whole study? Plot this.

2. Pick a species. How does abundance of this species in each sample vary over aspect?

3. Pick two or three species. How does their abundance in each sample vary over elevation?

4. Pick a species. What is the range of the per-sample abundance recorded for this species?

5. For the same species, plot the frequency distribution of per-sample abundances across the whole data.

6. What is the minimum, mean, median and maximum of per-sample abundance for all species?

7. Pick a species. In three separate plots, show how habitat type, disturbance and land use affect the abundance of this species in samples? What does this look like for another species?

8. Create a new table that records species richness, minimum, maximum and total abundance for each sample. Join this table to the site and remote sensing data and store this in a new object.

9. Plot how the species richness of samples changes with elevation. Then try this with other potential drivers of species richness variation, both numeric and categorical.

10. Following on from questions 4 and 6 above, plot the range of per-sample counts for every species.

11. Pick two species. Plot the abundance in samples of one species against the abundance in samples of the other.

12. Pick one species. Make a multi-plot figure showing the count of that species against the count of all other species.

Note: the following tasks are pretty tricky. No shame if you can't figure these out!

13. The minimum and the mean temperature vary over different months. Manipulate the climate data to separate these out into different rows (i.e. the six columns should be reduced to two, for mean and min, and there should be 3 rows for each current row, for month 6, 7, 8).

14. Join this to the site date, and look at how mean and min temperature vary over month and elevation

15. Join the data from step 14 to the individual species count data, and look at how the abundance of two or three species changes with temperature

16. Join the data from step 14 to the community summary data (generated in the first part of task 8) and look at how species richness changes with temperature

Dataset 2: Orangutan sentiment analysis

This dataset looks at viewer responses to Youtube videos featuring oranguatans. Take a look here for some key information on the dataset and why it was collected: https://zenodo.org/record/5437547#.Y9JqRHbP3Zt. Using YouTube analytics and sentiment analysis of comments on 118 videos, the researchers wanted to understand how viewer responses to videos vary with 1) the amount of human-orangutan interaction depicted, 2) the ages of the orangutans featured, and 3) the mention of threats to orangutans.

You should go to the link above to read about the data and download the excel file containing the data.

Tasks

1. Compile some summary statistics. How many videos are there? What's the average video length?

2. What's the average number of comments or likes? Repeat for median, min, and max. Can you do this within the same line of code?

3. Create a plot to show how the number of views vary by video length? Repeat for the number of likes and interaction time against video length.

4. Can you standardise interaction length by video length and then plot this measure for the two categories of test (rescue oranguatan videos) and control (general oranguatan videos) groups?

5. How does the proportion of likes to dislikes vary by the mention of threats to oranguatans? How about for standardised interaction length?

6 . Repeat the above questions, but this time consider the age of the oranguatans (instead of threats or test/control groups).

7. Use summary statistics to compare the number or proportion of comments between different sentiment categories (negative, neutral, positive to oranguatans). Plot a suitable figure to display this data.

8. Take the plot above and split it so that it separates out into the test and control groups.