Leeds Data Science Society Data Challenge

November 2019

Scenario

Vivienne was exploring a tropical rainforest island and discovered a new species of monkey. She observed the monkeys frequently for two years and saw that the monkeys mostly spent their time in 7 different species of tree. Unfortunately, on her voyage home, Vivienne's paper notes about the monkeys' behaviour were lost at sea, but she still has information about the counts of monkeys in each species, rainfall and the different tree species. She can remember that the monkeys ate the flowers, fruit and leaves on many of the trees, and they also used trees to shelter from heavier rain.

From the data Vivienne has given you, can you provide some suggestions about what the monkeys might use each tree for?

Below are some questions you could think about answering – you do not need to answer all of these. You could cover many, or produce a very in depth answer to one or two of them.

- Which trees do the monkeys prefer for rain shelter?
- Which trees do the monkeys prefer to eat the leaves of?
- Do the monkeys have any fruits they prefer to eat over others?
- Do the monkeys have any flowers they prefer to eat over others?
- Are there any trees that the monkeys do not seem to like the fruit of?
- Are there any trees that the monkeys do not seem to like the flowers of?
- Do the monkeys prefer leaves, fruits or flowers to eat?

Data

You will be working with simulated data as this is not a real scenario! The data files are detailed below.

Monkey_count.csv

This dataset details the counts of monkeys in different trees on different dates. There were 51 monkeys on the island, all of the new species, but Vivienne did not see them all every day.

Variable name	Units	Meaning
Date	-	Date of monkey counts
Tree	-	Tree species ID in which monkeys were observed
Count	-	Number of monkeys counted

Tree_details.xlsx

This dataset details the flowering periods, fruiting periods, leaf diameter, and leaf coverage of the seven different tree species in which the monkeys spent most of their time.

Variable name	Units	Meaning
Tree species	-	Tree species name
Tree	-	Tree species ID
Flowering	-	Months in which flowers are blooming
Fruiting	-	Months in which fruits are on the tree
Average leaf diameter	cm	Average diameter of leaves on the tree
Average leaf coverage	%	Percentage of horizontal area covered by one vertical metre of tree canopy

Rain.csv

This dataset details the number of centimetres of rainfall each day from January 1st 2016 to December 31st 2017.

Variable name	Units	Meaning
Date	-	Date of rainfall measurement
Rainfall	cm	Centimetres of rainfall collected between 00:00 and 11:59 on each day

Stages of the analysis

There are a number of steps you may want to work through when investigating this data:

- Data cleaning – dealing with with any errors and missingness in the dataset.
- Data exploration – explore the structure of your data. How are the variables distributed? What variables are correlated with each other?
- Plan and run your analysis – use information from your data exploration to plan an analysis that answers your chosen questions.
- Interpret your analysis – look at your results and see what the answers to your questions are.

Software suggestions

Common statistical programs that you might want to use for analysis include R, Python, Excel, STATA, SPSS and SAS. STATA and SPSS are available on some university computers. We encourage the use of R, Python or Excel. As these software are either open source (R, Python) or widely used (Excel) it makes your analysis more reproducible by other parties.

Hint words

Here are some key words that might be useful for you when investigating how to analyse your data:

Binomial

Missing Data

Count data

Hawthorne effect

Outliers

Generalised linear models

Data cleaning

Dummy variable

Indicator Variables

Poisson

Regression modelling