DATA CLEANING AND STANDARDS FOR BWG DATABASE

Rolando Trejo

2022-09-16

Important information

BWG database coming from the bromeliad working group

Main database owners: Diane Srivastava,Michael Melnychuk and Jana Petermann

Dataset used for this assigment: bwgv1_bromeliads.csv

Full proyect disponible at https://github.com/RolandoTrejo/Rolando_Data_Management_LDP

Context

This code is part of the third data cleaning task refering to perform quality control checks on three measured variables in the bromeliads dataset (e.g., max_water, num_leaf, height, diameter, ph, etc., etc. – see names(bromeliads)). Your checks might include, for example, looking for the proportion of observations that include data on each variable, or identifying outliers or improbably values. Remember that some bromeliad variables should be correlated with each other. [suggested package: assertr; see example code in tutorials 2 and 3].

The problem

Let's imagine that we want to predict the total detritus as function of the number of leaves and diameter of bromeliads using a model approach (detritus <- number of leaves * diameter). The data must contained less than 5 NA in each column. The numeric data must not contain as well outliers with a SD=3. In addition, if the number of leaves and diameter as predictors are not correlated among them, they must be transformed into a categorical classification. The final dataset must conserve no redundant variables.

The solution

To solve the problem stated before, a data cleaning considering the following tasks is needed:

3.1. Check for the number of observations containing NA, if there are more than 5 in a column, they must be suppressed from the final dataset.

3.2. Check for outliers in the number of leaves, diameter and total detritus. Observation with a SD=3 must be excluded from the final dataset.

3.3. Check for correlation between number of leaves and diameter of bromeliads. If they are correlated, one of them must be excluded to prevent using redundant predictors in the final model.

3.4. If the number of leaves and diameter are not correlated, then transform them into categorical variables. Use the following criteria:

```
Leaves: < 15, "low"/ <= 30, "medium"/ > 30, "high"
Diameter: < 50, "small"/ <= 100, "medium" / > 100, "large"
```

3.5. Write the final clean dataset as bromeliads_clean.csv