**Assignment 1 – Garland Xie**

Within the “bromeliad” dataset, I made a reproducible R script that does quality control checks on three variables (i.e., maximum water, total detritus, and extended diameter). Specifically, these checks are evaluating (1) appropriate ranges (i.e., no negative values) and (2) potential outliers (i.e., four median absolute deviations) for each variable. This will be done “upstream” of the data pipeline before data cleaning.

The general workflow for quality control checks consists of three consecutive steps: (1) to evaluate if any data validation rules are triggered, (2) fix or filter any errors and return a separate data frame, and (3) re-run the data validation rules until no errors are found.

For the first step, assert that each of the three variables is within a reasonable range, from zero to infinity, since it is impossible to contain negative values. Then, check to evaluate potential outliers on each variable where outliers are beyond four median absolute deviations. If so, then I assume that these outliers are data points that should be removed from the analysis.

The first step poses a problem because it can generate two possible outputs: the original data frame and another with error statements. To resolve this issue, I made a control flow statement where if the output has errors, I must fix and filter the errors, re-run the previous validation rules, and then calculate summary statistics. Otherwise, calculate the summary statistics on the original data frame.