**Department of Statistics, University of Auckland**

**Summer Scholarship 2017-2018**

**Data wrangling tools in Shiny and Gtk**

**Name: Owen Jin**

**Supervisor: Chris Wild**

**Degree: BCom/Sc**

**Summary**

I have created replacement functions for most of the options given in the “VARIABLES” and “DATASET” tabs of iNZight by using packages outside of base R. These packages include dplyr, tidyr, forcats and stringr. My functions essentially replace the old base R functions.

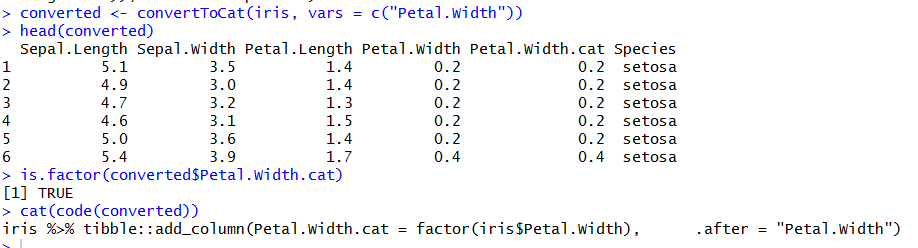
In addition to the old functions, the functions I created also reproduce the code needed to replicate the results of the function

*Variables Tab*

Convert to Categorical

Function: convertToCat()

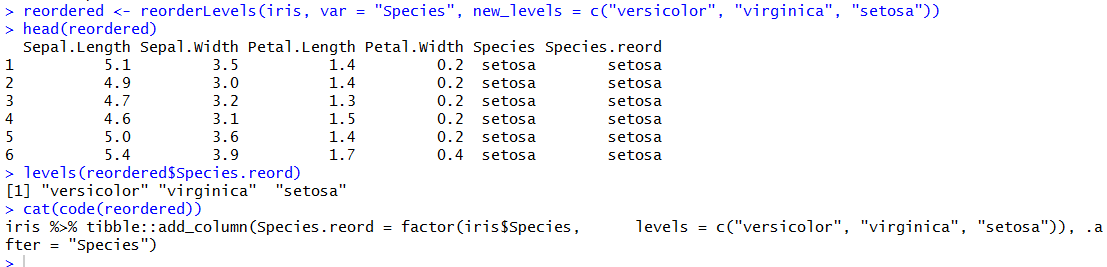
Convert several numeric columns into categorical columns. The new column created is a factor.



Categorical Variables -> Reorder Levels

Function: reorderLevels()

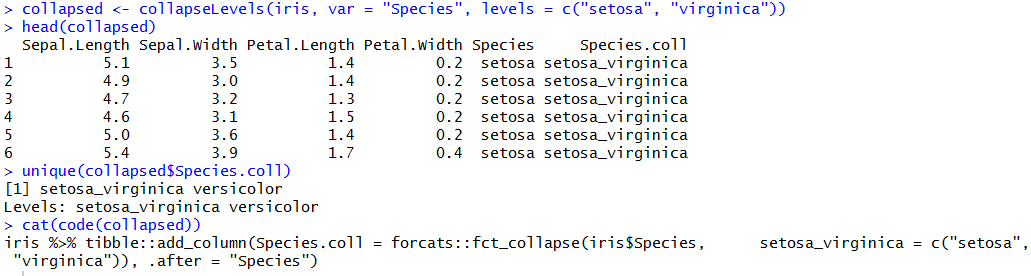
Reorder the factor levels of a categorical column either manually or by descending frequency



Categorical Variables -> Collapse Levels

Function: collapseLevels()

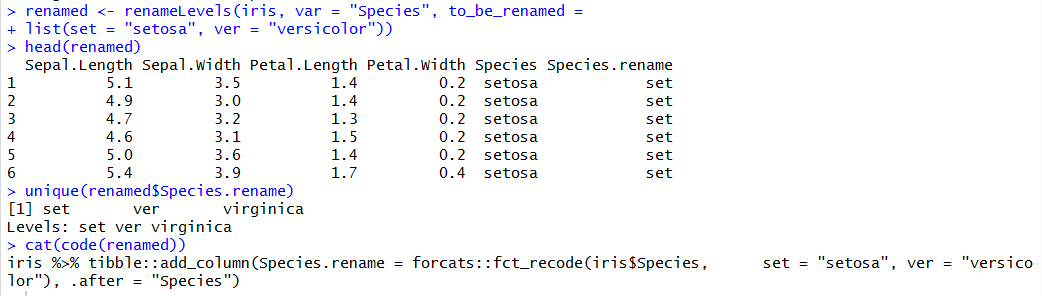
Collapse 2 or more levels of a factor in a categorical variable into one factor.



Categorical Variables -> Rename Levels

Function: renameLevels()

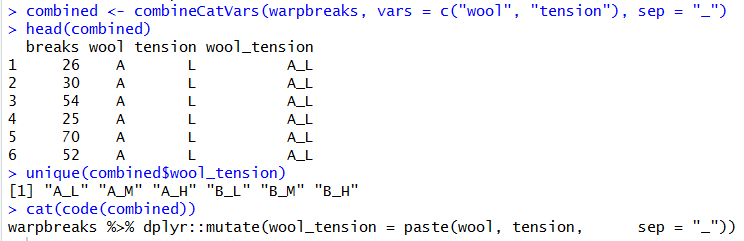
Rename the factor levels of a categorical column



Categorical Variables -> Combine Categorical Levels

Function: combineCatVars()

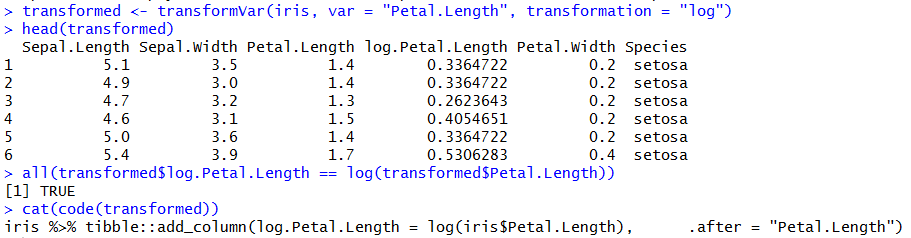
Concatenate the values of several categorical columns into one value



Numeric Variables -> Transform Variables

Function: transformVar()

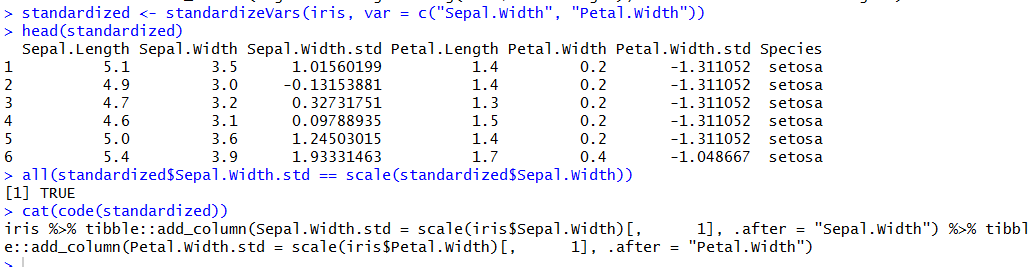
Transform a numeric column and apply a mathematical function



Numeric Variables -> Standardize Variables

Function: standardizeVars()

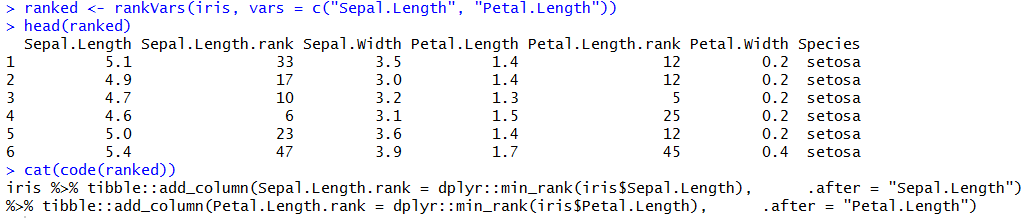
Centre then divide by the standard error of the values in a numeric column



Numeric Variables -> Rank Numeric Variables

Function: rankVars()

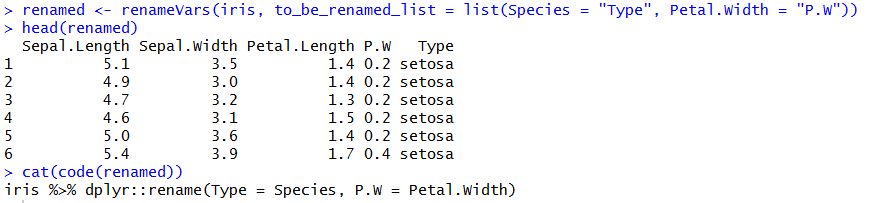
Rank the values of a numeric column in descending order. Ties are broken as such: eg. values = 5, 6, 6, 7 ; rank = 1, 2, 2, 3



Rename Variables

Function: renameVars()

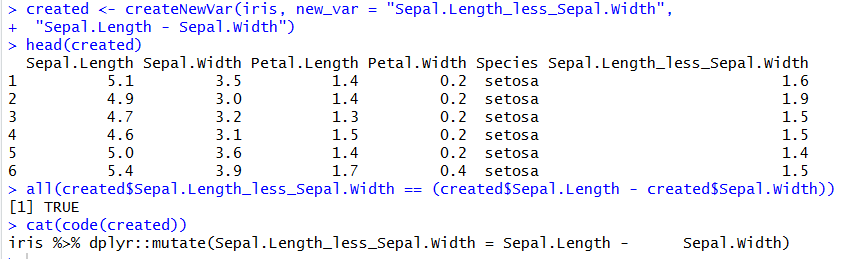
Rename the column names



Create New Variables

Function: createNewVar()

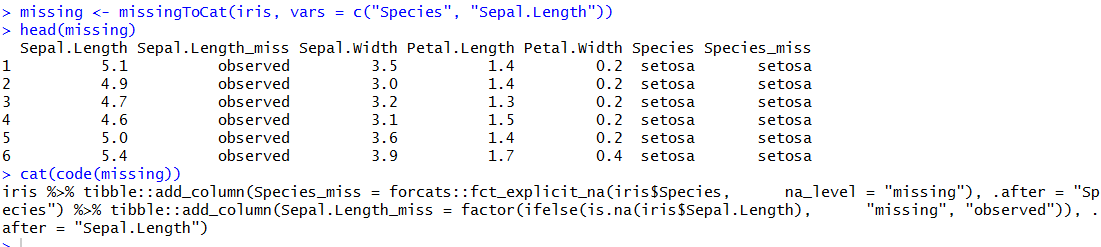
Create a new column using an R expression



Missing to Categorical

Function: missingToCat()

Turn <NA>’s into “missing” and if applied to a numeric column, all non-missing numeric values will be turned into “observed”

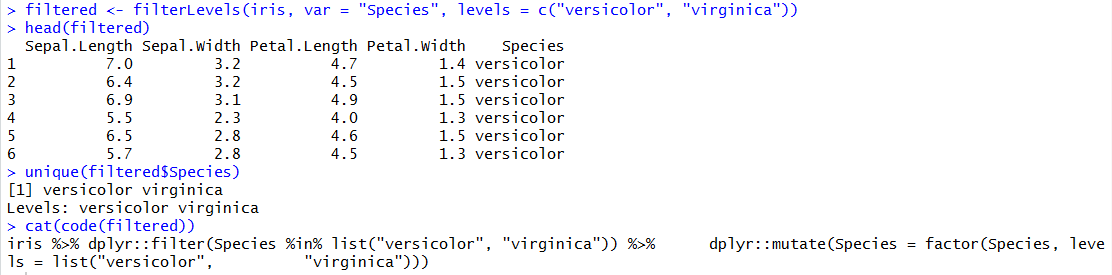


*Dataset Tab*

Filter Dataset -> Levels of a categorical variable

Function: filterLevels()

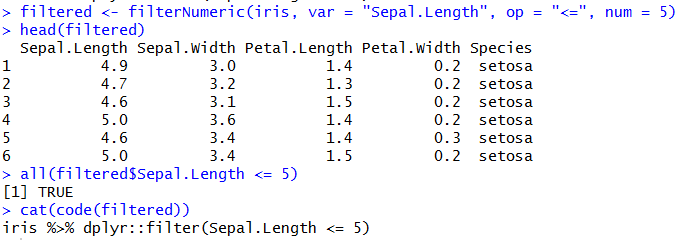
Filter a dataset based off the values in a categorical column



Filter Dataset -> Numeric condition

Function: filterNumeric()

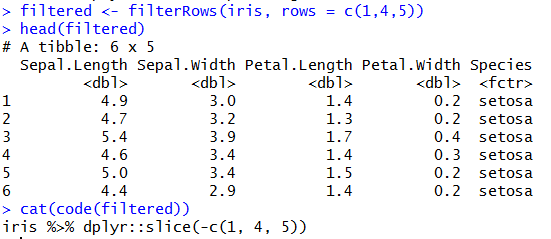
Filter a dataset based off a logical condition in a numeric column



Filter Dataset -> Row number

Function: filterRows()

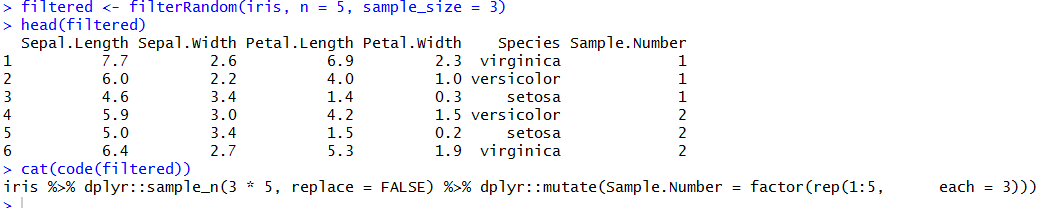
Filter a dataset by slicing off specified row numbers



Filter Dataset -> Randomly

Function: filterRandom()

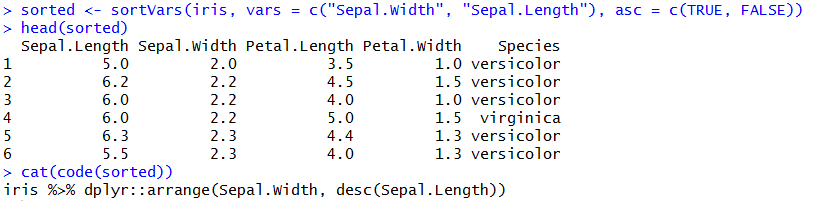
Sample X number of groups without replacement, each of size N



Sort data by variables

Function: sortVars()

Sort columns in either ascending or descending order. Sorts by up to 4 columns

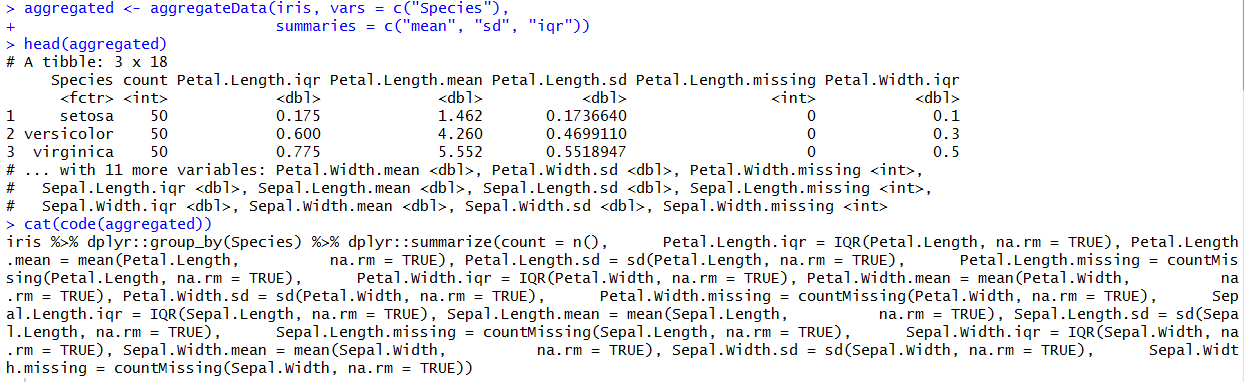


Aggregate Data

Function: aggregateData()

Summarizes all the numeric variables by grouping them according to specified categorical variables and producing the summary of the mean, median, sum, sd and/or IQR.

Custom function made called countMissing() to give the number of missing values, functions identical to sum(is.na()) but has a is.na parameter which does nothing except for making the evaluation in aggregateData() work.



Stack Variables

Function: stackVars()

Stack one or more columns converting the dataset to a tall format

