workingReport

**Bảy Núi An Giang Project**

Raw data -> data processing -> data analyzing -> communicating results

# Data processing

The data processing contains these works:

1. Load the raw dataset
2. Display structure and overall information of the data set: using str()
3. From the result of str(), decide which variables need to be modified
4. Dealing with missing values
5. Using knowledge to check for suspected wrong data
6. Tidy data
7. Subset data as needed
8. Write data to external files.

**29122021**

## 1. Load the raw dataset

Using MS Excel to look at the data interactively and there are things that I need to take care of:

1. A title (row 1) that need to be remove
2. Second row containing header (columns’ name) that are not suitable for data analyzing

Using readxl::read\_xlsx() to load the data.

Create a variable.txt file containing all the names of the dataset’s columns.

Using colnames() to assign names.

**02012022**

## 2. Display structure and overall information of the data set: using str()

str() gives result that:

* the data type of some variables are wrongly decided.
* the dataset contains non-English text that cannot be presented in R.
* text contains a mix of uppercase and lowercase.

**03012022**

## 3. From the result of str(), decide which variables need to be modified

### uppercase and convert all the text to Latin

At first, I decided to work with the objects’ data type and then moving to manipulate text case and coding style. However, I quickly found out that, stri\_trans\_general() and toupper convert all things to character type. So it’s better to work with the data type at the end.

### decide data types for the objects

I manually create text file (outside R) called dataType.txt that contains the right data type for all the variables. Read the data into R and use it to correct all objects that have wrong data type.

Case-sensitive can causes problem when working with categorical data. So it’s better to uppercase all the text.

### Remove

There are two duplicated columns (income.2018.1 and income.2018.2). Remove one.

There are 3 rows at the end of the dataset that is NULL.

## 4. Dealing with missing values

### Dealing with blank cells in the raw data

The raw data contains blank cells that can either be NA or 0. At the beginning, R interprets all these blank values as NA. So I have to manually changes those values accordingly.

**area.forestry**

condition: area.total == area.NLKH + area.forestry

First, convert all the NAs in the area.forestry to 0. Then check for the condition above.