Preparation of data for analysis

Theory R functions **Examples**

Detecting missing values

As an example how to detect missing values in a matrix, let's use Danube meadow dataset with Ellenberg indicator values for individual species (this is a dataset with species attributes, with species in rows and tabulated Ellenberg indicator values in columns):

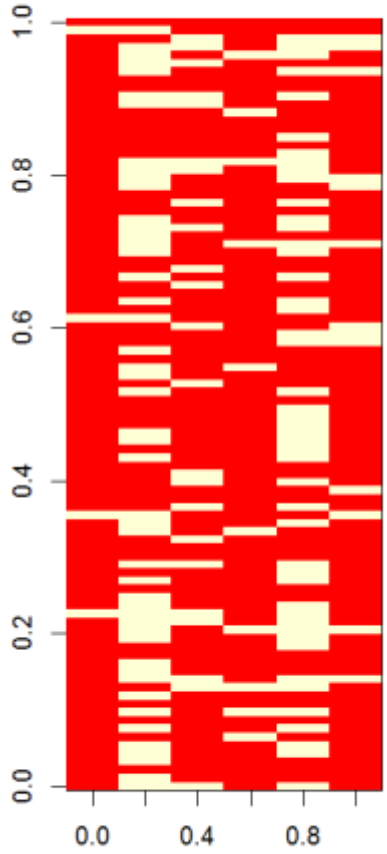
```
danube.ell <- read.delim
('https://raw.githubusercontent.com/zdealveindy/anadat-r/master/data/danube.ell.txt
', row.names = 1)
```

A simple way to know mising values if the data are in `matrix` or `data.frame` format is to use `summary`:

```
summary (danube.ell)
```

```
     Light            Temp            Cont           Moist            React
Nutr
 Min.   :4.000   Min.   :4.000   Min.   :2.000   Min.   : 2.000   Min.   :3.000
Min.   :1.000
 1st Qu.:6.250   1st Qu.:5.000   1st Qu.:3.000   1st Qu.: 4.000   1st Qu.:6.500
1st Qu.:4.000
 Median :7.000   Median :5.000   Median :3.000   Median : 5.000   Median :7.000
Median :5.000
 Mean   :6.889   Mean   :5.435   Mean   :3.901   Mean   : 5.524   Mean   :6.851
Mean   :4.938
 3rd Qu.:7.000   3rd Qu.:6.000   3rd Qu.:5.000   3rd Qu.: 6.000   3rd Qu.:7.000
3rd Qu.:6.000
 Max.   :8.000   Max.   :6.000   Max.   :7.000   Max.   :10.000   Max.   :8.000
Max.   :9.000
 NA's   :4       NA's   :48      NA's   :23      NA's   :10       NA's   :47
NA's   :13
```

The bottom row in the output of `summary` shows the number of missing values in each variable. But which values are missing?



Missing values in the dataset visualized using `image` function. Yellow fields are those with missing values.

```
which (is.na (danube.ell), arr.ind = T)
```

```
  row col
Chenopodium album 22 1
Festuca rubra 34 1
Melandrium diurnum 58 1
Vicia sepium 93 1
Achillea millefolium 1 2
Ajuga reptans 2 2
Alopecurus pratensis 4 2
Angelica sylvestris 5 2
Anthriscus silvestris 6 2
...
```

Here, the function `is.na` transforms the `danube.ell` data frame into logical values - TRUE for the values which are missing and FALSE for those which are present. The function `which` searches for TRUE values in the data frame, and by setting argument `arr.ind = TRUE` the function returns the coordinates (row x column) of each missing value.

If you want to visualize how many holes your dataset has, use the function `image`, which draws the "heatmap" of values in matrix-like object (here it has to be really `amatrix`, not `data.frame`, which is why I transformed the `danube.ell` data frame into matrix by `as.matrix` function):

```
image (t (as.matrix (is.na (danube.ell))))
```

Since I am interested to see only two types of values, those which have some value and those which have missing value (NA), I also used function `is.na` to transform original real values into logical TRUE/FALSE values depending on whether given element is missing or not. The color palette used in `image` function by default is `heat.colors`, in which low values are red and high values are yellow (or almost white). In this case, high values are missing values, and are displayed as yellow in the resulting heat map. I also transposed the matrix using function `t`, otherwise, the columns would be drawn horizontally. But note that still, the drawing starts from the left bottom corner, not from the left top, so the matrix visualization is upside down compare to the values in the original matrix.

From:
https://anadat-r.davidzeleny.net/ - **Analysis of community ecology data in R**

Permanent link:
https://anadat-r.davidzeleny.net/doku.php/en:data_preparation_examples

Last update: **2021/03/03 19:48**