Using the tapply() function to calculate the means of two groups

```
> head(light)
     treatment  response
869          C  -4.504766
871          C  -3.164909
873          C  -4.320002
875          C  -4.486886
877          C  -4.207244
879          C  -4.591058
1002         E  -3.310792
1004         E  -4.135877
1006         E  -5.112481
1008         E  -4.780870
1010         E  -4.822527
1012         E  -4.761931
```

The data look like this. Each data point gets a separate row, and is either from a "C" or an "E" treatment. There is just one column of the numerical response variable.

The tapply() functions takes 3 main argument.
i) The numeric variable that you want to get information from (ie, mean values by group).
ii) The categorical variable that defines the groups in the data.
iii)The mathematical function to apply to the data (mean, sd, min, etc)

Structure of the tapply() function

Dollar sign $ to designate a column w/in the dataframe

The dataframe

i)The numeric variable to take the mean of

```
my.means <- tapply(light$response,
                   light$treatment,
              FUN = mean,
              na.rm = T)
```

ii)The categorical variable that defines the 2 groups

Commas!

New data object to hold
The means

iii) The mathematical function to apply to the data

na.rm = T so deal with any missing data (NAs)

```
> my.means
        C           E
-4.353540  -4.829708
```
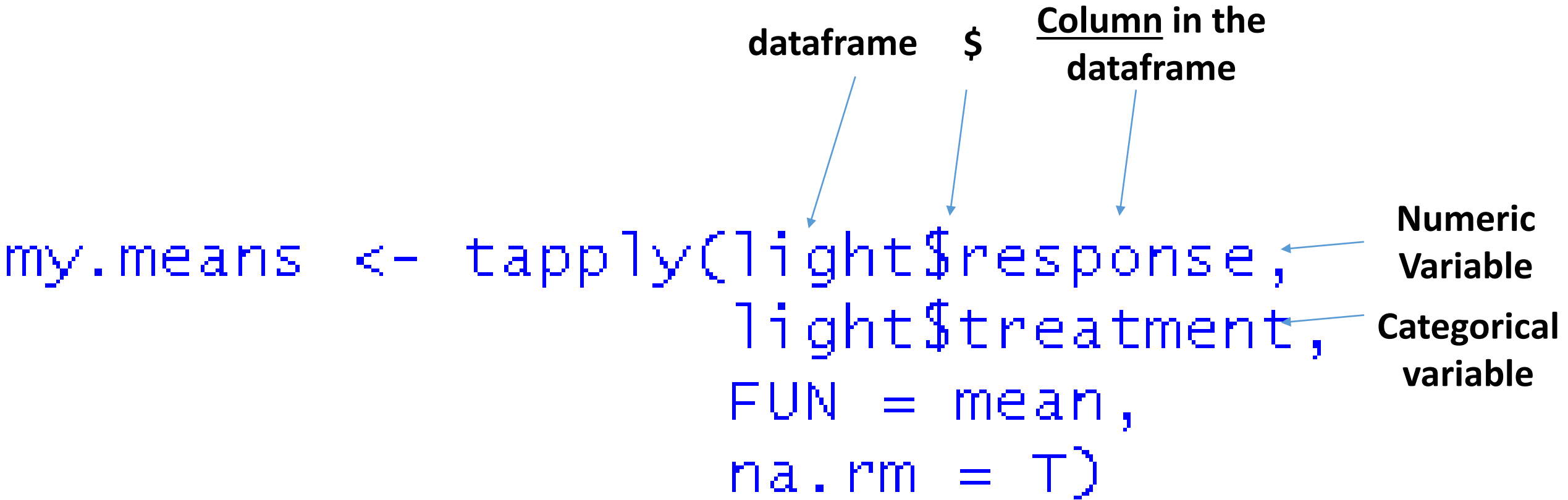
Output: 2mean values and the group they correspond to (C or E)

dataframe   **$**   <u>**Column**</u> **in the dataframe**

```
my.means <- tapply(light$response,
                   light$treatment,
                   FUN = mean,
                   na.rm = T)
```

Numeric Variable

Categorical variable

The "treatment" column in the "light" dataframe contains categorical data. The names of these categories (aka groups) DO NOT appear anywhere in the tapply() function. tapply() looks at them "behind the scenes" and figures out how to split up the data to get the mean of each group.