How can I reshape my data in R?

Authored by

stats writer

June 30, 2024

Reshaping data in R refers to the process of transforming a dataset into a different structure to better suit the analysis or visualization needs. This can be achieved by using various functions and packages in R, such as `tidyr`, `reshape2`, or `dplyr`. The reshaping process involves changing the layout of the data from wide to long format or vice versa, rearranging the columns and rows, and creating new variables. By reshaping data, researchers can gain a better understanding of their data and uncover hidden patterns or relationships. This is a crucial step in the data analysis process and can greatly improve the accuracy and efficiency of statistical modeling and data visualization in R.

How can I reshape my data in R? | R FAQ

When there are multiple measurements of the same subject, across time or using different tools, the data is often described as being in "wide" format if there is one observation row per subject with each measurement present as a different variable and "long" format if there is one observation row per measurement (thus, multiple rows per subject). Different functions require different formats, and so the need to reshape a dataset may arise.

Below, we start with a dataset in wide format. Students have been measured using five metrics: read, write, math, science, and socst.

hsb2

id female race ses schtyp prog read write math science socst
1 70 0 4 1 1 1 57 52 41 47 57
2 121 1 4 2 1 3 68 59 53 63 61
3 86 0 4 3 1 1 44 33 54 58 31
4 141 0 4 3 1 3 63 44 47 53 56
5 172 0 4 2 1 2 47 52 57 53 61
6 113 0 4 2 1 2 44 52 51 63 61
7 50 0 3 2 1 1 50 59 42 53 61
8 11 0 1 2 1 2 34 46 45 39 36
9 84 0 4 2 1 1 63 57 54 58 51
10 48 0 3 2 1 2 57 55 52 50 51

To reformat this dataset into long form, we will use the reshape
function. The arguments we provide include a list of variable names that define
the different times or metrics (varying), the name we wish to give the
variable containing these values in our long dataset (v.names), the name
we wish to give the variable describing the different

times or metrics (timevar),
the values this variable will have (times), and the end format for the
data (direction). Additionally, we have provided new row names.

Wide to long

l
id female race ses schtyp prog subj score
99 1 1 1 1 1 3 read 34
299 1 1 1 1 1 3 write 44
499 1 1 1 1 1 3 math 40
699 1 1 1 1 1 3 science 39
899 1 1 1 1 1 3 socst 41
139 2 1 1 2 1 3 read 39
339 2 1 1 2 1 3 write 41
539 2 1 1 2 1 3 math 33
739 2 1 1 2 1 3 science 42
939 2 1 1 2 1 3 socst 41

After sorting by id, we can see that we have five rows per student and
their five scores appear in the score variable with the

subj
variable indicating which test on which the score was measured. For each of the
five rows per id, the female, race, ses, schtyp,
and prog variables are unchanging.

We can similarly go from this long form back to our original wide form again
using reshape with different arguments (most importantly, direction = "wide").
With timevar, we indicate the variable that will define the multiple
measurements per subject. With idvar, we list the variables that should
do not vary within subject.

Long to wide

w

id female race ses schtyp prog score.read score.write score.math score.science score.socst
99 1 1 1 1 1 3 34 44 40 39 41
139 2 1 1 2 1 3 39 41 33 42 41
84 3 0 1 1 1 2 63 65 48 63 56

112 4 1 1 1 1 2 44 50 41 39 51

76 5 0 1 1 1 2 47 40 43 45 31

149 6 1 1 1 1 2 47 41 46 40 41

50 7 0 1 2 1 2 57 54 59 47 51

94 8 1 1 1 1 2 39 44 52 44 48

60 9 0 1 2 1 3 48 49 52 44 51

154 10 1 1 2 1 1 47 54 49 53 61

Once again, we have one observation per student. We can compare the
dimensions of our wide dataset w to our original dataset:

dim(w)
200 11

dim(hsb2)
200 11