Data Reshaping in R

For Health, Behavioral, Educational, and Social Scientists

Tyson S. Barrett

Changing data from wide to long format is an essential tool in most data analyses. Long ("tidy") form is generally needed for plotting, statistical and other analyses.
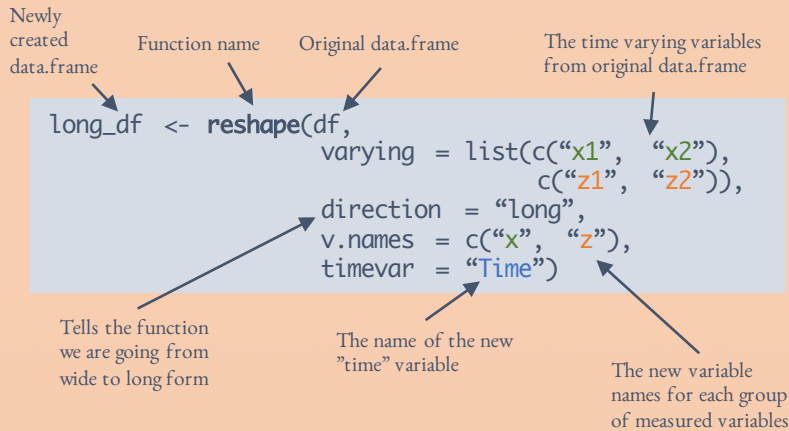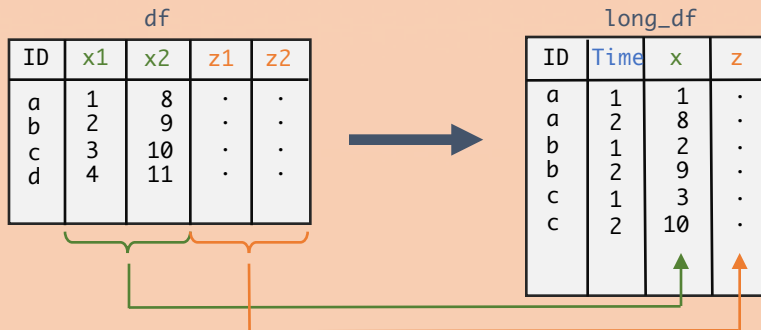
Wide Format

ID	x1	x2	z1	z2
a	1	8	.	.
b	2	9	.	.
c	3	10	.	.
d	4	11	.	.

Long Format

ID	Time	x	z
a	1	1	.
a	2	8	.
b	1	2	.
b	2	9	.
c	1	3	.
c	2	10	.

Reshaping: Wide to Long Form

df

ID	x1	x2	z1	z2
a	1	8	.	.
b	2	9	.	.
c	3	10	.	.
d	4	11	.	.

long_df

ID	Time	x	z
a	1	1	.
a	2	8	.
b	1	2	.
b	2	9	.
c	1	3	.
c	2	10	.

Newly created data.frame → Function name → Original data.frame → The time varying variables from original data.frame

```
long_df <- reshape(df,
                   varying = list(c("x1", "x2"),
                                  c("z1", "z2")),
                   direction = "long",
                   v.names = c("x", "z"),
                   timevar = "Time")
```

Tells the function we are going from wide to long form

The name of the new "time" variable

The new variable names for each group of measured variables

Note: in varying we include a list because we are including two measures (i.e., measure x and measure z)

Reshaping: Long to Wide Form

df

ID	Time	x	z
a	1	1	.
a	2	8	.
b	1	2	.
b	2	9	.
c	1	3	.
c	2	10	.

wide_df

ID	x1	x2	z1	z2
a	1	8	.	.
b	2	9	.	.
c	3	10	.	.
d	4	11	.	.

Newly created data.frame → Function name → Original data.frame → Tells the function we are going from wide to long form

```
wide_df <- reshape(df,
                   direction = "wide",
                   v.names = c("x", "z"),
                   timevar = "Time")
```

Tells the function the name of the "time" variable

The variables that will be Split into different times (e.g., "x1" and "x2" stem "x")
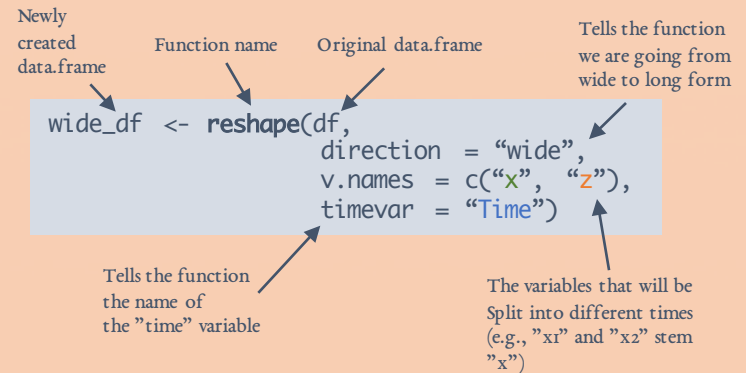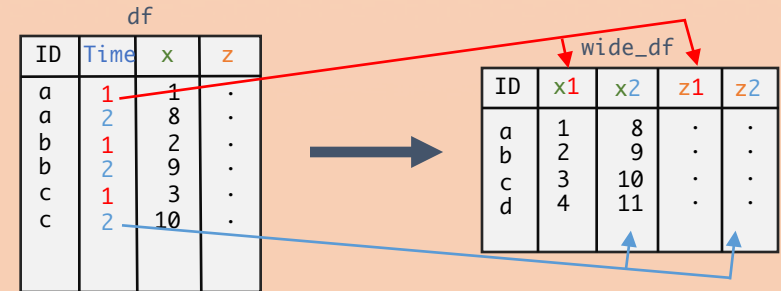
Note: if we don't include a v.names argument, all variables in the data.frame will be split.

Other Reshaping Tools

```
library(tidyverse)                              ## dplyr package
long <- gather(df, "time", "value", 2:3)  ## wide to long
wide <- spread(df, time, value)            ## long to wide
```

Additional Notes Regarding Data Reshaping:

- Recommendation: Save only one data set (in wide or long format) that is your "main" data set—to avoid confusion about what different data sets contain. Then, use these reshaping functions in R to manipulate the data to get it in the form needed to plot and analyze. Finally, save the code ("syntax") and any output (you can save the reshaped data but it is not necessary).
- Term Definitions: *measure* implies any distinct variable in the data set (e.g., x2), *time* implies anything that distinguishes the observations (e.g., time, cluster, location), *time varying variable* is any variable that has or could have a different value at each observation occasion.