Anatomy of a typical dataframe in R

Column labels in 1st row

Data starts in 2nd row

Group 1

Group 2

	A	B
1	**age.class**	**harem.mean.size**
2	age.13.plus	5
3	age.13.plus	3
4	age.13.plus	1
5	age.13.plus	2.142857143
6	age.13.plus	3
7	age.13.plus	2
8	age.13.plus	6.037037037
9	age.13.plus	3.428571429
10	age.13.plus	4
11	age.13.plus	1.5
12	age.13.plus	1.777777778
13	age.13.plus	3.8
14	age.13.plus	6
15	age.10	6.782608696
16	age.10	4.578947368
17	age.10	13.14814815
...	...	...
38	age.10	9.25
39	age.10	7.933333333
40	age.10	14.96296296
41	age.10	8.666666667
42		

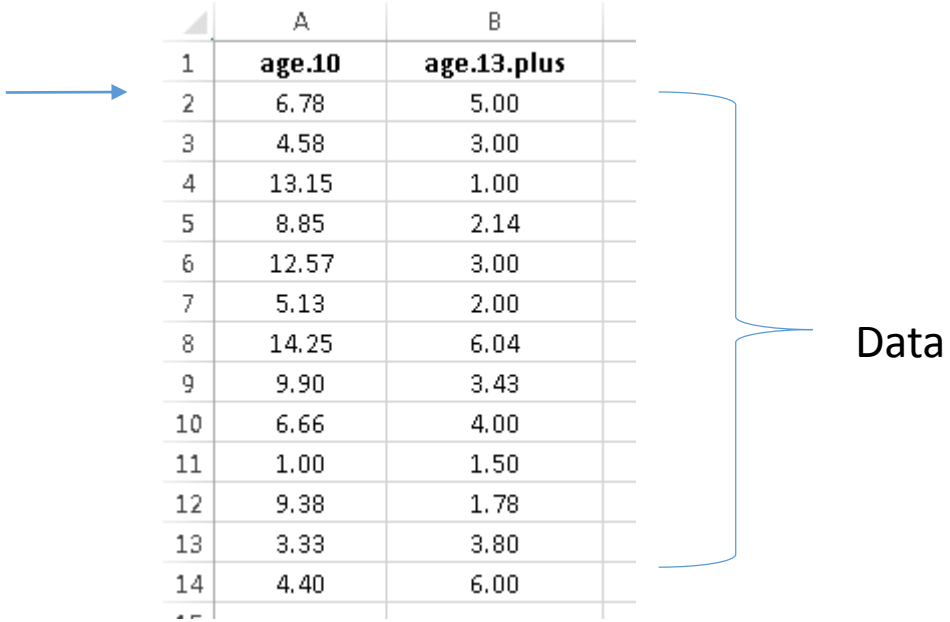
Each characteristic that was measured on the observational units or category that classifies it gets its own COLUMN. Here, there is a column for the age class the deer are in (age 13 or more vs. age 10) and a column for his mean harem size during the course of the study. Other columns could be “rack.size”, “body.mass” or any other data taken on an individual male.

Each ROW is a separate **unit of observation**. Here it is an individual male deer. This could also be individual humans, survey plots, plates of bacteria, etc; each unique object or individual that got measured gets its own row of data. Here, there were 40 individuals studied, so there are 41 rows. 1 row for the column names plus 40 rows, one for each buck.

Incorrect format for an R dataframe

Two *different* individuals in each row! In column A, the harem size of an age 10 male. In column B of the same row, harem size for a completely different age 13+ individual.

Group 1 Group 2



	A	B
1	**age.10**	**age.13.plus**
2	6.78	5.00
3	4.58	3.00
4	13.15	1.00
5	8.85	2.14
6	12.57	3.00
7	5.13	2.00
8	14.25	6.04
9	9.90	3.43
10	6.66	4.00
11	1.00	1.50
12	9.38	1.78
13	3.33	3.80
14	4.40	6.00

Data

This is a very common format for data in Excel and a natural way to think about data. For some purposes R can work with this kind of data. However, for general work in R this format is much more difficult to use and should generally be avoided.