

Intro to Data Frames in R

Environmental Data Analytics | John Fay and Luana Lima

1/10/2022

Data frame definition

A **data frame** is used for storing data tables. It is a list of vectors of equal length.

When we import data to R, data frame is the preferred way for storing the data because columns can have different modes (character, numeric, integer, logical, complex).

Data frame built-in example

Let's look into a built-in data frame from package "datasets" - cars. The data give the speed of cars and the distances taken to stop.

```
cars
```

```
##      speed dist
## 1         4     2
## 2         4    10
## 3         7     4
## 4         7    22
## 5         8    16
## 6         9    10
## 7        10    18
## 8        10    26
## 9        10    34
## 10       11    17
## 11       11    28
## 12       12    14
## 13       12    20
## 14       12    24
## 15       12    28
## 16       13    26
## 17       13    34
## 18       13    34
## 19       13    46
## 20       14    26
## 21       14    36
## 22       14    60
## 23       14    80
## 24       15    20
## 25       15    26
## 26       15    54
```

```
## 27    16    32
## 28    16    40
## 29    17    32
## 30    17    40
## 31    17    50
## 32    18    42
## 33    18    56
## 34    18    76
## 35    18    84
## 36    19    36
## 37    19    46
## 38    19    68
## 39    20    32
## 40    20    48
## 41    20    52
## 42    20    56
## 43    20    64
## 44    22    66
## 45    23    54
## 46    24    70
## 47    24    92
## 48    24    93
## 49    24   120
## 50    25    85
```

Note that it has 2 columns and 50 rows.

Data frame columns

Suppose you want just the column speed. How would you access that data?

```
cars$speed
```

```
## [1]  4  4  7  7  8  9 10 10 10 11 11 12 12 12 12 13 13 13 13 14 14 14 14 15 15
## [26] 15 16 16 17 17 17 18 18 18 18 19 19 19 20 20 20 20 20 22 23 24 24 24 24 25
```

How would you store it on another object?

```
car_speed <- cars$speed
car_speed
```

```
## [1]  4  4  7  7  8  9 10 10 10 11 11 12 12 12 12 13 13 13 13 14 14 14 14 15 15
## [26] 15 16 16 17 17 17 18 18 18 18 19 19 19 20 20 20 20 20 22 23 24 24 24 24 25
```

Transforming object in a data frame

Is the new object you create a data frame?

```
#Option 1
class(car_speed)
```

```
## [1] "numeric"
```

```
#returned "numeric"
```

```
#Option 2
```

```
is.data.frame(car_speed)
```

```
## [1] FALSE
```

```
#returned "FALSE"
```

```
#summary-- the new object that was created was not a data frame -- car_speed only contained Values and
```

How could you make it a data frame?

```
df_car_speed <- as.data.frame(car_speed)
df_car_speed
```

```
##      car_speed
## 1           4
## 2           4
## 3           7
## 4           7
## 5           8
## 6           9
## 7          10
## 8          10
## 9          10
## 10         11
## 11         11
## 12         12
## 13         12
## 14         12
## 15         12
## 16         13
## 17         13
## 18         13
## 19         13
## 20         14
## 21         14
## 22         14
## 23         14
## 24         15
## 25         15
## 26         15
## 27         16
## 28         16
## 29         17
## 30         17
## 31         17
## 32         18
## 33         18
```

```
## 34      18
## 35      18
## 36      19
## 37      19
## 38      19
## 39      20
## 40      20
## 41      20
## 42      20
## 43      20
## 44      22
## 45      23
## 46      24
## 47      24
## 48      24
## 49      24
## 50      25
```

```
class(df_car_speed)
```

```
## [1] "data.frame"
```

#summary-- assigning a name and using as.data.frame() means that you are creating a data frame-- df_car.

Adding columns to a data frame

How could you add columns to *df_car_speed*?

```
car_dist <- cars$dist
```

#Option 1

```
df <- cbind(df_car_speed,car_dist) #similarly rows could be added using rbind()
class(df)
```

```
## [1] "data.frame"
```

```
df
```

```
##      car_speed car_dist
## 1           4         2
## 2           4        10
## 3           7         4
## 4           7        22
## 5           8        16
## 6           9        10
## 7          10        18
## 8          10        26
## 9          10        34
## 10         11        17
## 11         11        28
## 12         12        14
```

```
## 13      12      20
## 14      12      24
## 15      12      28
## 16      13      26
## 17      13      34
## 18      13      34
## 19      13      46
## 20      14      26
## 21      14      36
## 22      14      60
## 23      14      80
## 24      15      20
## 25      15      26
## 26      15      54
## 27      16      32
## 28      16      40
## 29      17      32
## 30      17      40
## 31      17      50
## 32      18      42
## 33      18      56
## 34      18      76
## 35      18      84
## 36      19      36
## 37      19      46
## 38      19      68
## 39      20      32
## 40      20      48
## 41      20      52
## 42      20      56
## 43      20      64
## 44      22      66
## 45      23      54
## 46      24      70
## 47      24      92
## 48      24      93
## 49      24     120
## 50      25      85
```

```
#Or Option 2 - transform into a data frame before binding
df_car_dist <- as.data.frame(car_dist) #op2
df_opt2 <- cbind(df_car_speed,df_car_dist)
class(df_opt2)
```

```
## [1] "data.frame"
```

```
df_opt2
```

```
##      car_speed car_dist
## 1           4         2
## 2           4        10
## 3           7         4
## 4           7        22
```

```
## 5      8      16
## 6      9      10
## 7     10     18
## 8     10     26
## 9     10     34
## 10    11     17
## 11    11     28
## 12    12     14
## 13    12     20
## 14    12     24
## 15    12     28
## 16    13     26
## 17    13     34
## 18    13     34
## 19    13     46
## 20    14     26
## 21    14     36
## 22    14     60
## 23    14     80
## 24    15     20
## 25    15     26
## 26    15     54
## 27    16     32
## 28    16     40
## 29    17     32
## 30    17     40
## 31    17     50
## 32    18     42
## 33    18     56
## 34    18     76
## 35    18     84
## 36    19     36
## 37    19     46
## 38    19     68
## 39    20     32
## 40    20     48
## 41    20     52
## 42    20     56
## 43    20     64
## 44    22     66
## 45    23     54
## 46    24     70
## 47    24     92
## 48    24     93
## 49    24    120
## 50    25     85
```

Note that when we transformed the vector in a data frame the name of the vector became the column name.

```
colnames(df) #or simply names()
```

```
## [1] "car_speed" "car_dist"
```

```
names(df)
```

```
## [1] "car_speed" "car_dist"
```

Creating a data frame

How would you create a data frame?

```
#useful function data.frame()  
create_df <- data.frame("speed"=car_speed,"dist"=car_dist)  
create_df
```

```
##      speed dist  
## 1         4    2  
## 2         4   10  
## 3         7    4  
## 4         7   22  
## 5         8   16  
## 6         9   10  
## 7        10   18  
## 8        10   26  
## 9        10   34  
## 10       11   17  
## 11       11   28  
## 12       12   14  
## 13       12   20  
## 14       12   24  
## 15       12   28  
## 16       13   26  
## 17       13   34  
## 18       13   34  
## 19       13   46  
## 20       14   26  
## 21       14   36  
## 22       14   60  
## 23       14   80  
## 24       15   20  
## 25       15   26  
## 26       15   54  
## 27       16   32  
## 28       16   40  
## 29       17   32  
## 30       17   40  
## 31       17   50  
## 32       18   42  
## 33       18   56  
## 34       18   76  
## 35       18   84  
## 36       19   36  
## 37       19   46  
## 38       19   68  
## 39       20   32
```

```
## 40      20      48
## 41      20      52
## 42      20      56
## 43      20      64
## 44      22      66
## 45      23      54
## 46      24      70
## 47      24      92
## 48      24      93
## 49      24     120
## 50      25      85
```

Data frame functions

Some useful functions to use with data frames.

```
ncol(df)
```

```
## [1] 2
```

```
nrow(df)
```

```
## [1] 50
```

```
length(df) #same as ncol
```

```
## [1] 2
```

```
summary(df) #shows the summary of dist and speed that you would see in boxplot
```

```
##      car_speed      car_dist
##  Min.   : 4.0    Min.     : 2.00
##  1st Qu.:12.0    1st Qu.: 26.00
##  Median :15.0    Median : 36.00
##  Mean   :15.4    Mean     : 42.98
##  3rd Qu.:19.0    3rd Qu.: 56.00
##  Max.    :25.0    Max.     :120.00
```

```
head(df) #show the first 6 rows of df
```

```
##      car_speed car_dist
## 1           4         2
## 2           4        10
## 3           7         4
## 4           7        22
## 5           8        16
## 6           9        10
```



```
#If you know the number of the column you want you can refer to that to access column  
df[,1]
```

```
## [1] 4 4 7 7 8 9 10 10 10 11 11 12 12 12 12 13 13 13 13 14 14 14 14 15 15  
## [26] 15 16 16 17 17 17 18 18 18 18 19 19 19 20 20 20 20 22 23 24 24 24 24 25
```

```
#you could also use this notation to delete columns  
df <- df[, -2]  
df
```

```
## [1] 4 4 7 7 8 9 10 10 10 11 11 12 12 12 12 13 13 13 13 14 14 14 14 15 15  
## [26] 15 16 16 17 17 17 18 18 18 18 19 19 19 20 20 20 20 22 23 24 24 24 24 25
```

```
#summary--this deleted the second column, which was car_dist
```