

Exploring Tibbles

July 23, 2023

The

Dataframes and Tibbles are frequently employed data structures in R for storing and manipulating data. They facilitate the organization, exploration, and analysis of data.

Dataframes

1. A dataframe is a two-dimensional table-like data structure in R that stores data in rows and columns, with distinct data types for each column.
2. Similar to a spreadsheet or a SQL table, it is one of the most frequently employed data structures in R. Each column in a `data.frame` is a constant-length vector, and each row represents an observation or case.
3. Using the `data.frame()` function or by importing data from external sources such as CSV files, Excel spreadsheets, or databases, dataframe objects can be created in R.
4. dataframe objects have many useful built-in methods and functions for manipulating and summarizing data, including subsetting, merging, filtering, and aggregation. [1]

Creating a dataframe using raw data

5. The following code generates a `data.frame` named `df` containing three columns - `names`, `ages`, and `heights`, and four rows of data for each individual.

```
# Create input data as vectors
names <- c("Ashok", "Bullu", "Charu", "Divya")
ages <- c(72, 49, 46, 42)
heights <- c(170, 167, 160, 166)

# Combine input data into a data.frame
people <- data.frame(Name = names, Age = ages, Height = heights)
```

```
# Print the resulting dataframe  
print(people)
```

	Name	Age	Height
1	Ashok	72	170
2	Bullu	49	167
3	Charu	46	160
4	Divya	42	166

Reading Inbuilt datasets in R

1. R contains a number of built-in datasets that can be accessed without downloading or integrating from external sources. Here are some of the most frequently used built-in datasets in R:
 - **women**: This dataset includes the heights and weights of a sample of 15,000 women.
 - **mtcars**: This dataset contains information on 32 distinct automobile models, including the number of cylinders, engine displacement, horsepower, and weight.
 - **diamonds**: This dataset includes the prices and characteristics of approximately 54,000 diamonds, including carat weight, cut, color, and clarity.
 - **iris**: This data set measures the sepal length, sepal width, petal length, and petal breadth of 150 iris flowers from three distinct species.

The women dataset

As an illustration, consider the **women** dataset inbuilt in R, which contains information about the heights and weights of women. It has just two variables:

1. **height**: Height of each woman in inches
2. **weight**: Weight of each woman in pounds
3. The **data()** function is used to import any inbuilt dataset into R. The **data(women)** command in R loads the **women** dataset

```
data(women)
```

4. The **str()** function gives the dimensions and data types and also previews the data.

```
str(women)
```

```
'data.frame':  15 obs. of  2 variables:
 $ height: num  58 59 60 61 62 63 64 65 66 67 ...
 $ weight: num  115 117 120 123 126 129 132 135 139 142 ...
```

5. The `summary()` function gives some summary statistics.

```
summary(women)
```

height	weight
Min. :58.0	Min. :115.0
1st Qu.:61.5	1st Qu.:124.5
Median :65.0	Median :135.0
Mean :65.0	Mean :136.7
3rd Qu.:68.5	3rd Qu.:148.0
Max. :72.0	Max. :164.0

The `mtcars` dataset

The `mtcars` dataset inbuilt in R comprises data on the fuel consumption and other characteristics of 32 different automobile models. Here is a concise description of the 11 `mtcars` data columns:

1. `mpg`: Miles per gallon (fuel efficiency)
2. `cyl`: Number of cylinders
3. `disp`: Displacement of the engine (in cubic inches)
4. `hp`: gross horsepower
5. `drat`: Back axle ratio wt: Weight (in thousands of pounds)
6. `wt`: Weight (in thousands of pounds)
7. `qsec`: 1/4 mile speed (in seconds)
8. `vs`: Type of engine (0 = V-shaped, 1 = straight)
9. `am`: Type of transmission (0 for automatic, 1 for manual)
10. `gear`: the number of forward gears
11. `carb`: the number of carburetors

```
data(mtcars)
str(mtcars)
```

```
'data.frame':  32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num   6  6  4  6  8  6  8  4  4  6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num   3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num   2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num   16.5 17 18.6 19.4 17 ...
 $ vs  : num   0  0  1  1  0  1  0  1  1  1 ...
 $ am  : num   1  1  1  0  0  0  0  0  0  0 ...
 $ gear: num   4  4  4  3  3  3  3  4  4  4 ...
 $ carb: num   4  4  1  1  2  1  4  2  2  4 ...
```

Reading different file formats into a dataframe

1. We examine how to read data into a dataframe in R when the original data is stored in prominent file formats such as CSV, Excel, and Google Sheets.
2. Before learning how to accomplish this, it is necessary to comprehend how to configure the Working Directory in R.

Working Directory

1. The working directory is the location where R searches for and saves files by default.
2. By default, when we execute a script or import data into R, R will search the working directory for files.
3. Using R's `getwd()` function, we can examine our current working directory:

```
getwd()
```

```
[1] "/cloud/project"
```

4. We are running R in the Cloud and hence we are seeing that the working directory is specified as `/cloud/project/DataAnalyticsBook101`. If we are doing R programming on a local computer, and if our working directory is the Desktop, then we may see a different response such as `C:/Users/YourUserName/Desktop`.
5. Using R's `setwd()` function, we can change our current working directory. For example, the following code will set our working directory to the Desktop:

```
setwd("C:/Users/YourUserName/Desktop")
```