Efficient Management of Data in R (Data Structures!)

Data Science Lecture Series: Advanced R

W. Evan Johnson, Ph.D.
Professor, Division of Infectious Disease
Director, Center for Data Science
Rutgers University – New Jersey Medical School

2023-03-27

Section 1

Introduction to Data Structures

Importance of data structures

A data structure is a particular way of organizing data in a computer so that it can be used effectively. The idea is to reduce the space and time complexities of different tasks.

Data structures in R programming are tools for holding multiple values, variables, and sometimes functions.

Please think very carefully about the way you manage and store your data! This can make your life much easier and make your code and data cleaner and more portable!

Types of data structures in R

R's base data structures are often organized by their dimensionality (1D, 2D, nD) and whether they're homogeneous or heterogeneous (elements of identical or various type). Six of the most common data types are:

1. Vectors
2. Lists
3. Matrices
4. Arrays
5. Factors
6. Data frames (or tibbles)

Section 2

Data Frames

Data Frames

The most common data structure for storing a dataset in R is in a **data frame**. Conceptually, we can think of a data frame as a two dimensional table with rows representing observations and the different variables reported for each observation defining the columns. Data frames are particularly useful for datasets because we can combine different data types into one object.

Data Frames

We can convert matrices into data frames using the function `as.data.frame`:

```
mat <- matrix(1:12, 4, 3)
mat <- as.data.frame(mat)
```

Or just generate it directly using the `data.frame` function:

```
dat <- data.frame(x=1:4, y=5:8, z=9:12)
```

A `data.frame` can be indexed as matrices, `dat[1:2, 2:3]`, and columns can be extracted using the $ operator.

Section 3

Tibbles

Tibbles

Here is a printed version of the data frame:

```
dat
```

```
##   x y  z
## 1 1 5  9
## 2 2 6 10
## 3 3 7 11
## 4 4 8 12
```

Tibbles}

A **tibble** is a modern version of a data.frame.

```
library(tidyverse)
dat1 <- tibble(x=1:4, y=5:8, z=9:12)
```

Or convert a data.frame to a tibble

```
dat <- data.frame(x=1:4, y=5:8, z=9:12)
dat1 <- as_tibble(dat)
```

Tibbles

Here is a printed version of the tibble:

```
dat1
```

```
## # A tibble: 4 x 3
##       x     y     z
##   <int> <int> <int>
## 1     1     5     9
## 2     2     6    10
## 3     3     7    11
## 4     4     8    12
```

Tibbles

Important characteristics that make tibbles unique:

1. Tibbles are primary data structure for the `tidyverse`
2. Tibbles display better and printing is more readable
3. Tibbles can be grouped
4. Subsets of tibbles are tibbles
5. Tibbles can have complex entries–numbers, strings, logicals, lists, functions.
6. Tibbles can (almost) enable object-orientated programming in R

Section 4

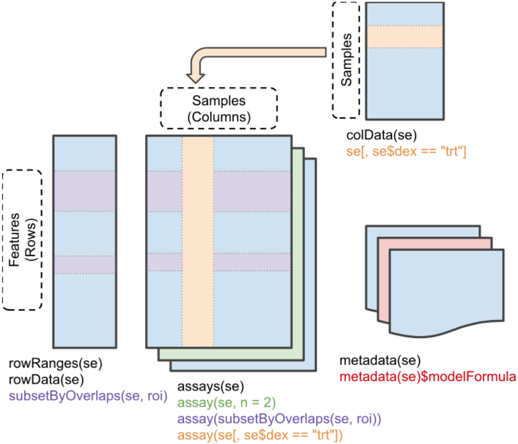
Advanced Data Structures in R

Advanced Data Structures in R

In your homework, you will explore more advanced R data structures, namely the **S3** and **S4** class objects. These can facilitate object orientated programming.

Advanced Data Structures in R

One example of an S4 class data structure is the **SummarizedExperiment** object.



Summarized Experiment

Session info

```
sessionInfo()
```

```
## R version 4.2.2 (2022-10-31)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Ventura 13.2.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] forcats_0.5.2   stringr_1.5.0   dplyr_1.1.0     purrr_1.0.0
## [5] readr_2.1.3     tidyr_1.2.1     tibble_3.1.8    ggplot2_3.4.0
## [9] tidyverse_1.3.2
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.2.0    xfun_0.36           haven_2.5.1
##  [4] gargle_1.2.1        colorspace_2.0-3    vctrs_0.5.2
##  [7] generics_0.1.3      htmltools_0.5.4     yaml_2.3.6
## [10] utf8_1.2.2          rlang_1.0.6         pillar_1.8.1
## [13] withr_2.5.0         glue_1.6.2          DBI_1.1.3
## [16] dbplyr_2.2.1        modelr_0.1.10       readxl_1.4.1
## [19] lifecycle_1.0.3     munsell_0.5.0       gtable_0.3.1
## [22] cellranger_1.1.0    rvest_1.0.3         evaluate_0.19
## [25] knitr_1.41          tzdb_0.3.0          fastmap_1.1.0
## [28] fansi_1.0.3         broom_1.0.2         scales_1.2.1
```