Querying and Converting Data Types in R

UNDERSTANDING DATASET STRUCTURES AND FORMATS

Martin Burger
STATS PROGRAMMING TUTOR

www.r-tutorials.com

Data Structures

Foundational concepts

Managing expectations

RStudio: Graphical user interface for R

Exploring data.frames via built-in exercise datasets

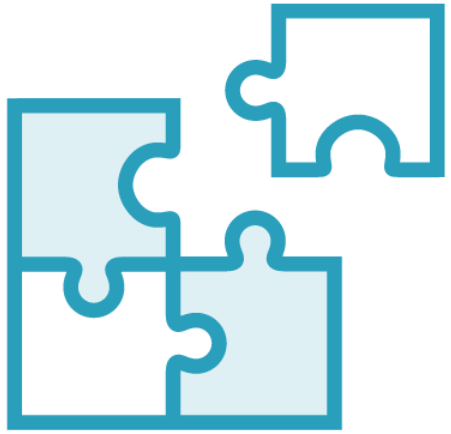
- Properties of data.frame
- Alternative structures: data.table, tibble

Data structures by dimensionality

Managing Expectations

Managing Expectations

Content

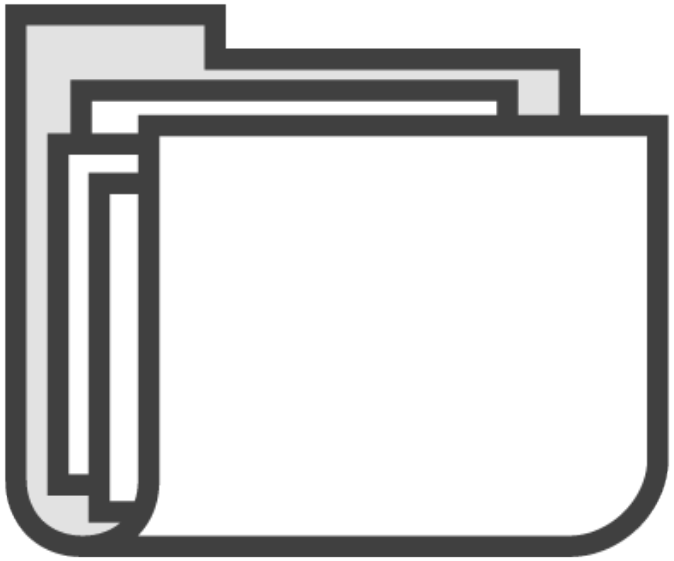


Structure

Prerequisites

Software

Course Content and Prerequisites

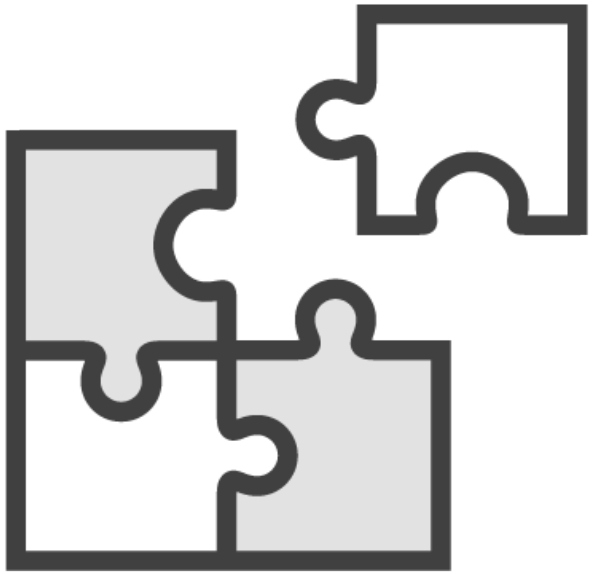


Beginner level course

No prior knowledge of R is required

Focus: Data types and querying data tables

- Data.frame: Table data structure of R Base



Data structures in R

Getting familiar with RStudio

Data types: integer, numeric, character, factor, logical, date, time

Simple and conditional data queries

- Filtering three alternative systems: data.frame, data.table and tibble

Further resources and summary

Software requirements: R and RStudio

- R: Programming language with basic graphical interface
- RStudio: Additional graphical user interface with built-in functionalities

Get your open source toolset:

- R: cran.r-project.org
- RStudio: rstudio.com

Required packages:

- Installation (once): `intall.packages()`
- Activation (each session): `library()`

Introduction to RStudio

Download libraries once, but activate them in each R session.

Installing Required Libraries

Exploring Built-in Datasets

Exercise Datasets

Various datasets come with R Base and add-on libraries

Reproducible results for exercising or communicating coding problems

Popular datasets: mtcars, iris

Exercise Datasets

Finding and accessing built-in datasets

Exploring datasets

Manipulating and accessing data

The basics of working with class data.frame

Library datasets

- Help section

Exploring a dataset with functions head, tail and summary

Visual exploration with functions plot and hist

Basics of data.frame indexing

- Accessing columns with $
- Attaching and detaching data frames
- Indexing operator: [row, column]

Course Datasets

Advantages of Built-in Datasets

Reproducible results

Open source

No pre-processing is needed

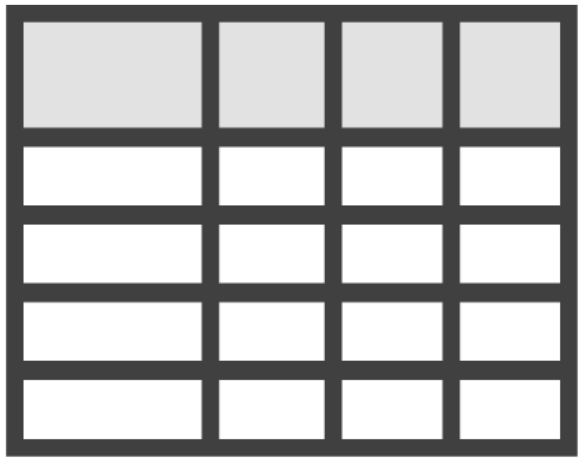
The Two Main Datasets

Motion Trend Car Road Tests
(mtcars)

Edgar Anderson's Iris Data
(iris)

The Tabular Format: Data Frame

The Tabular Format in Data Analysis



Data organized in tabular format

- Spreadsheet (MS Excel), data frame (programming languages)

Most analytical tools and functions require the data to be organized in a data frame

- Examples for exceptions: Text corpus or time series data

Advantages of Tabular Formats

Intuitive, easy to understand structure

Most data science tools accept it

Most statistical procedures can be performed on it

Alternatives to the Data.frame: Data.table and Tibble

Data Frame Alternatives

The data frame is the cornerstone of R

Alternative structures:

- Improved functionalities
- Class data.table from library data.table
- Class tibble from libraries dplyr or tibble (Tidyverse)

Comparison of Alternative Classes

Data.table (data.table)

Became the standard for large datasets

Optimized and streamlined processing for a better performance

Intuitive query system

Tibble (dplyr or tibble)

Standard table class in the Tidyverse

Suited towards data pre-processing tools

Problematic features (e.g. auto-conversion) of strings were removed

Clean overview of variables

Enjoy the benefits of data.table and tibble with the data.frame as backup class.

Tabular Data Structures in R

Class data.frame
(R Base)

Class data.table
(data.table)

Class tibble
(dplyr or tibble)

Data Structures

Data Structures by Dimensionality

One dimensional datasets

Two dimensional datasets

Multidimensional datasets

Data Structures by Dimensionality

- **1-D**
  - Vector
  - Series
- **2-D**
  - Table
  - Matrix
- **N-D**
  - Array
  - List*

One Dimensional Structures

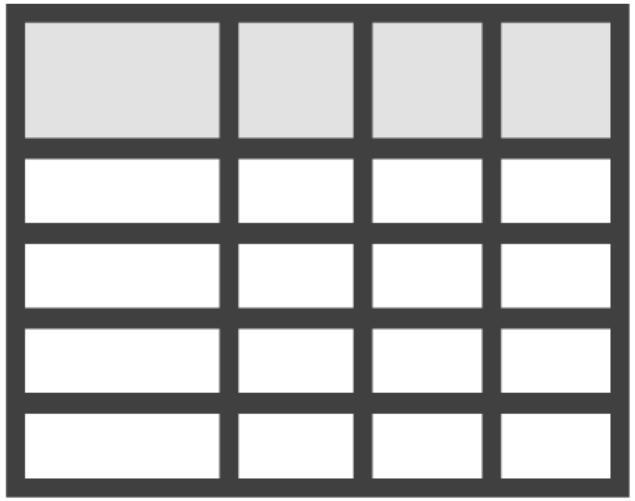
Vector

An unordered sequence of values of the same data type (e.g. body height measurements)

Series

An ordered sequence of values of the same data type (univariate time series)

Two Dimensional Structures



Tabular format:

- Columns for variables, rows for observations
- R classes: data.frame, data.table, tibble
- The most common format
- Columns can be of different data types

Matrix: Each column must be of the same class

N-dimensional Structures

R array:

- A structure that can have two or more dimensions of the same data type

R list:

- A structure that can collect objects of different classes and length
- Example: A list of length three includes a matrix, a data.frame and a vector

N-dimensional structures are less common than one or two dimensional ones

Data Structures in R

Vector	Table	Array
Series	Matrix	List

Data Structures

Game plan

RStudio: Graphical user interface for a better coding experience

- Customizable layout

Exploring the tabular format with data.frame

- Exercise data.frames mtcars and iris

Data.frame alternatives: data.table and tibble

Data structures by dimensionality