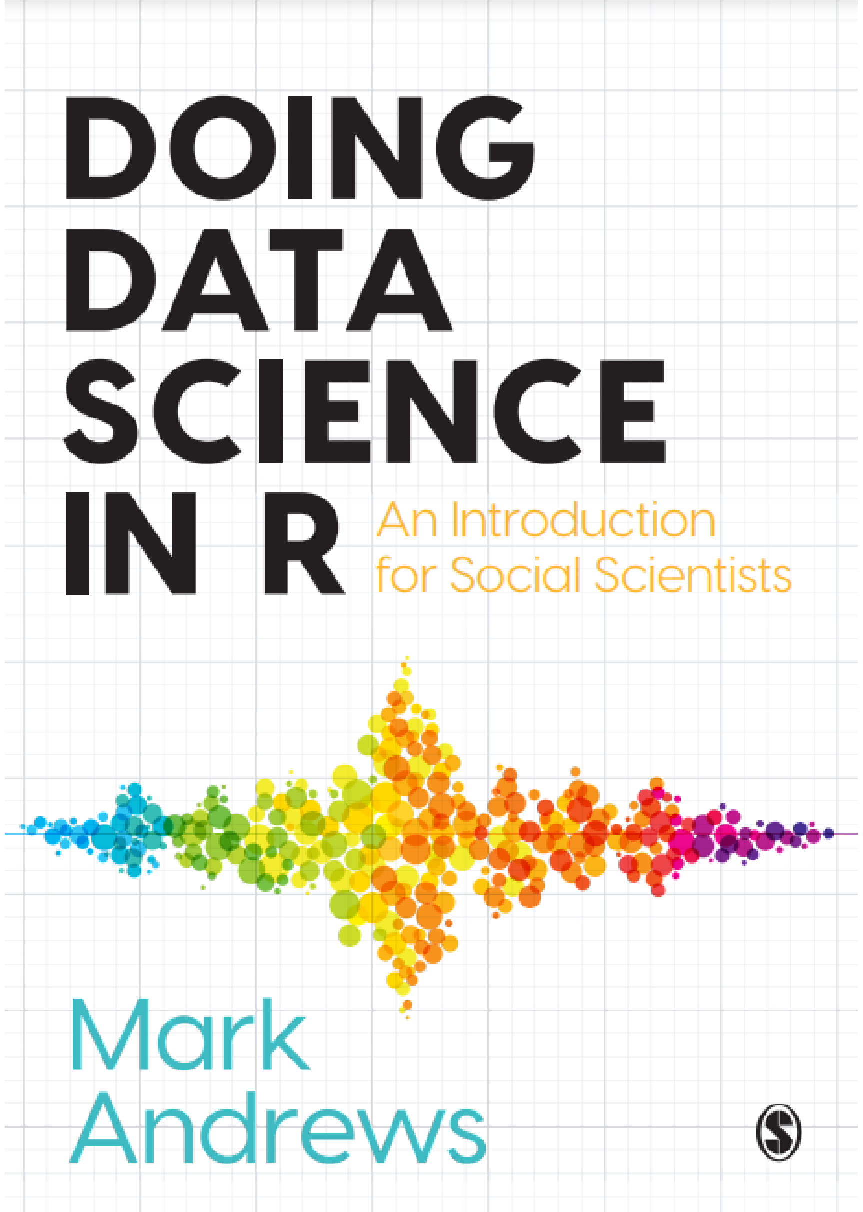

Doing Data Science in R: An Introduction for Social Scientists

Chapter 3

Data Wrangling

Data wrangling tools in R

We will focus on a core set of interrelated tidyverse tools

- select
- rename
- slice
- filter
- mutate
- arrange
- group_by
- summarize.

Reading text file data into a data frame

In principle, raw data can exist in any format in any type of file.

It is common to have data in a roughly rectangular format (i.e. with rows and columns), in text files such as .csv, .tsv, or .txt

Reading text file data into a data frame

The most commonly used include:

- read_
- read_tsv
- read_
- read_table

Manipulating data frames using dplyr

The dplyr package provides a set of versatile interrelated commands for manipulating data frames

Selecting variables with select

In our blp_df data frames we have seven variables

The dplyr command select allows us to select those we want.

Renaming variables with rename

When we select individual variables with select, we can rename them too

Selecting observations with slice and filter

We use slice to select observations by their indices.

A useful dplyr function that can be used in slice and elsewhere is n(), which gives the number of observations in the data frame.

Changing variables and values with mutate

The mutate command is a very powerful tool in the dplyr toolbox.

It allows us to create new variables and alter the values of existing ones.

Sorting observations with arrange

Sorting observations in a data frame is easily accomplished with arrange

Subsampling data frames

The dplyr package provides two methods to sample from a data frame.

The sample_frac allows us to sample a specified proportion of observations.

Reducing data with summarize and group_by

The dplyr package has a function summarize (or, equivalently, summarise) that applies summarizing functions to variables

A summarizing function is essentially any function that takes a vector and reduces it to single values

The %>% operator

The %>% operator in R is known as the pipe.

It allows us to create sequences of functions, sometimes known as pipelines, that avoid the use of repeated nested functions or temporary data structures.

Combining data frames

A bind operation is a simple operation that either vertically stacks data frames that share common variables, or horizontally stacks data frames that have the same number of observations

Combining data frames by joins

A join operation is a common operation in relational databases using SQL.

It allows us to join separate tables according to shared keys

Combining data frames by set operations

In dplyr, the functions intersect, union, etc., allow us to combine data frames that have identical variables using set operations.

Reshaping with pivot_longer and pivot_wider

A so-called tidy data set is a data set where all rows are observations, all columns are variables, and each variable is a single value.

To tidy this data frame, we need a variable for the subject, another for the experiment's condition, and another for the memory score for the corresponding subject in the corresponding condition