

R Subsetting

Authors: Mary Kho and Emily G. Weigel, Georgia Institute of Technology

Focus: The students will practice manipulating data to add to or extract subsets of specific values, rows, columns, or subsets of data contained in existing data files.

Overview: This lesson focuses on having students conduct basic data manipulation in R. Rather than directly editing their data file, students will be walked through the best practices in R for how to generate (smaller) datasets of interest using R commands.

Learning objectives:

1. Explain the utility and best practices of subsetting data via scripts
2. Practice coding elements necessary to extract and create subsets of data
3. Apply the methods and coding learned through swirl to their own datasets

Lesson sequence:

1. Introduce the basic idea behind subsetting (orally/with a displayed table; may be skipped if students are already familiar/have experience with the concept)
2. Swirl lesson
 - a. Specific_Values
 - b. Extracting_Columns_or_Rows
 - c. Selecting_Subsets
 - d. Removing_Values
 - e. Add_Data
3. Have the students use these skills to work on a dataset of their choice

Pre-lesson activities: Students should be introduced to at least one larger dataset for which subsetting by hand is complicated, time-intensive, and ideally, leads to errors. This sets the stage for why one might subset, particularly with scripts as opposed to editing an original data file. This lesson can then follow in-class or as homework.

Post-lesson activities: It is suggested that students use a dataset of their choice to apply their skills. For assessment, students should submit their materials (commented code and data file) at the end of the lesson, but it is advised that instructors check in with students to view code from the swirl lesson before applying to a new dataset.

Implementation notes: This lesson was designed to be used piecemeal as students encounter datasets they would like to manipulate in a more sophisticated way. It is assumed that the student should already know some R basics, therefore this lesson should not be used as an introduction to R; rather, it should come after students are comfortable with writing vectors and data frames and have done some basic R scripting. Encourage students to go slowly and keep a document or script open in which to paste the code (and any notes) as they work; it will make generating their own R code easier later.