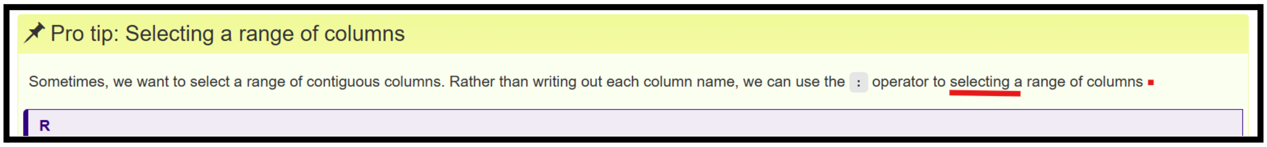
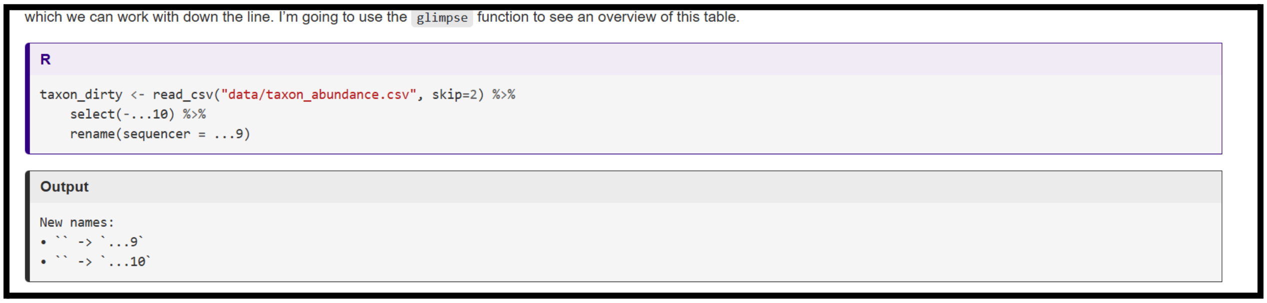
R for Data Analysis



Pro tip: Selecting a range of columns

Sometimes, we want to select a range of contiguous columns. Rather than writing out each column name, we can use the `:` operator to selecting a range of columns

R



which we can work with down the line. I'm going to use the `glimpse` function to see an overview of this table.

R

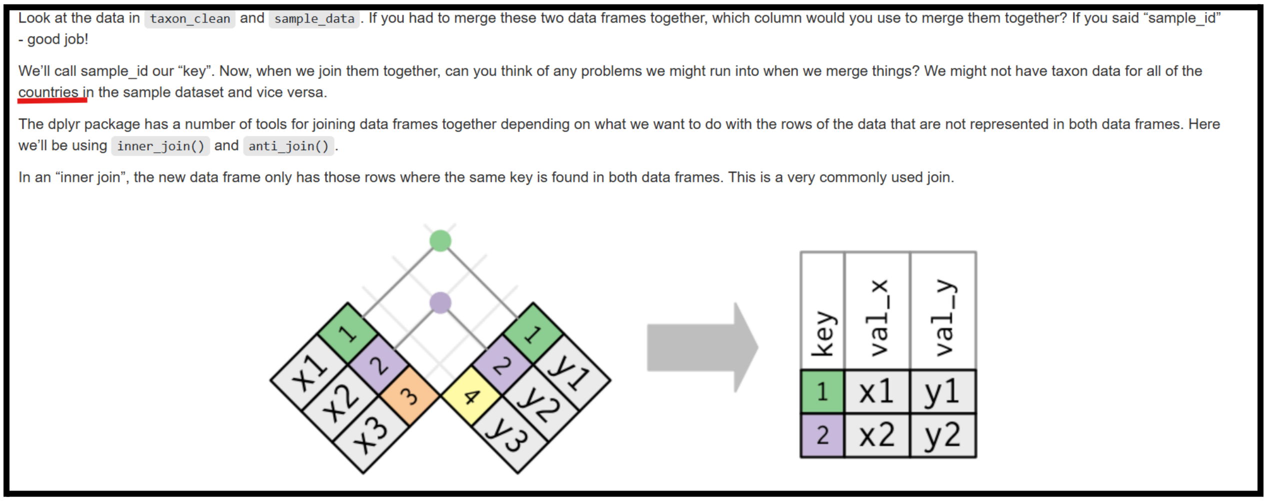
```
taxon_dirty <- read_csv("data/taxon_abundance.csv", skip=2) %>%
    select(-...10) %>%
    rename(sequencer = ...9)
```

Output

```
New names:
• `` -> `...9`
• `` -> `...10`
```

I have been using read.csv instead of read_csv this whole time. Here, read.csv gave me different variable names for the last two columns— "X" and "X.1" instead of "...9" and "...10"

It could be helpful to explain the difference between read.csv and read_csv early in the lesson? Maybe it doesn't come up that often as a problem, though. However, for me, this was a source of confusion for everything we've done thus far, although it did not really inhibit me from following along with the lesson. I was just getting different outputs and not sure why.



Look at the data in `taxon_clean` and `sample_data`. If you had to merge these two data frames together, which column would you use to merge them together? If you said "sample_id" - good job!

We'll call sample_id our "key". Now, when we join them together, can you think of any problems we might run into when we merge things? We might not have taxon data for all of the countries in the sample dataset and vice versa.

The dplyr package has a number of tools for joining data frames together depending on what we want to do with the rows of the data that are not represented in both data frames. Here we'll be using `inner_join()` and `anti_join()`.

In an "inner join", the new data frame only has those rows where the same key is found in both data frames. This is a very commonly used join.

Countries? I'm not sure what you mean.

I don't get a tibble as my output when I use functions like inner_join and anti_join. I just get a data frame. Not the end of the world, the tibble format is just helpful and I would like to know what's going on with that.