R Training at BC Stats

Instructor: Charlotte Wickham

Overview

This course will give you a feel for the complete data analysis process in R - from importing and manipulating data through visualization and modelling, and finally communicating results. You'll see how using code to capture the analysis pipeline leads to deliverables that are documented, easily reproduced and easily automated.

We'll focus on tools in the `tidyverse` a core set of R packages that are designed to be easy to learn, easy to use, and solve the most frequent data analysis problems.

During the course, we'll alternate between me introducing a new concept with some examples, and you applying that concept on your own. You should expect to spend at least 50% of your time writing code in RStudio on your own laptop.

The first half day is specifically for those who are new to R. Take a look at the prerequisites to see if you might be able to skip it.

Schedule

Session	Date/Time	Topic
Day 1: afternoon	Tue Dec 12th 1pm-4pm	Getting Started with R and RStudio
Day 2: morning	Wed Dec 13th 9am-12pm	Data Visualization with `ggplot2`
Day 2: afternoon	Wed Dec 13th 1pm-4pm	Data Manipulation with `dplyr` and `tidyr`
Day 3: morning	Thu Dec 14th 9am-12pm	Reporting with Rmarkdown
Day 3: afternoon	Thu Dec 14th 1pm-4pm	Workflow: list columns and iteration

Day 1 - Getting Started with R and RStudio

On your first afternoon you'll focus on getting comfortable writing code and executing it in RStudio. We'll take things slow as you learn to navigate RStudio, learn some syntax rules, and how to get help when you get stuck. Along the way you'll meet R's most ubiquitous objects for holding data and learn to import data whether it is a CSV, SPSS or Excel data file.

By the end of the day you will be able to:

- Open a notebook in RStudio and execute the code chunks in it
- Install and load an R package
- Open the help page for a function or built-in dataset
- Identify the components of an R function: the function name and arguments
- Assign the results of a function to a new variable
- Get an overview of a dataset that is in a data frame or tibble
- Import CSV, SPSS and SAS data files

Day 2 - Visualization and Manipulation of Data

We'll start the day with visualization of data in R using the package `ggplot2`. You'll see how `ggplot2` provides a framework for thinking about plots, which means you only need to learn one template to make almost any plot you can imagine. To practice, you'll make some of the most common kinds of data visualizations: histograms, scatterplots and time series plots, and continue building your skills as we continue through data manipulation.

In the afternoon we'll focus of the most common types of data manipulation: extracting subsets from data, adding new variables and creating grouped summaries. You'll find that doing this is quite intuitive using the `dplyr` package which boils down manipulation into a set of verbs like: `filter()`, `mutate()` and `summarise()`. Occasionally, data won't come in quite the right shape for manipulation or visualization you want to do, so we'll also talk about the key parts of the `tidyr` package that help to reshape not not-so-tidy data.

By the end of the day you will be able to:

- Create plots in `ggplot2` to explore data
- Select variables and filter observations to subset data
- Add new variables, and transform variables
- Create grouped summaries of data
- Reshape data for use with tidy tools

Day 3 - Communication and Workflow

In the morning we'll complete the data analysis process by learning about RMarkdown - a framework for combining code, results and narrative to produce fully documented and reproducible deliverables. You'll also see how easy Rmarkdown makes it to regenerate reports based on different slices of data (something known as *Parameterized Reports*).

In the afternoon we'll tackle a few more advanced ideas that are powerful ways to work with more complicated analyses. You'll learn that tibbles aren't just used for storing rectangular data, but are also a useful way to organise models and plots. The key tool to making use of this more complicated structure are the `map()` functions in the `purrr` package.

By the end of the day you will be able to:

- Polish and save your plots to produce publication ready figures
- Write your own R Markdown documents that combine code, output and documentation
- Use list columns in tibbles to store more complicated data structures like plots or models
- Use the iteration functions in `purrr` to work with list columns while keeping your analysis organized.

Prerequisites

The first half-day is specifically for people that are new to R. You can safely join us starting on day 2 if you already:

- know how to define variables in R
- have called a few basic functions (e.g. `mean()`), and
- know how to open .R script files, and run code in the console

Although I'll assume on the first half-day you haven't used R, you might like to get a little experience before we meet. Some options are:

- Work through the non-interactive Chapter 1 of Hands on Programming with R, which introduces RStudio as well as basic R.
- Work through the interactive chapters at Try R
- Try the free "Introduction to R" course at DataCamp

Software Requirements

You'll need to bring a laptop with R and RStudio installed. In addition, you'll want to install the following packages:

```r
install.packages(c("tidyverse", "rmarkdown",
  "babynames", "nycflights13", "gapminder", "Lahman"))
```

Don't forget to bring your power cable!

I'll also be providing some additional materials (slides, code and data) prior to our meeting, keep your eyes out for an email about that next week.

Instructor Info

Charlotte Wickham

- cwickham@gmail.com
- cwick.co.nz
- @cvwickham