Applying Concepts From Performance Management: Using R to Work With Performance Data III

Patrick J. Ferguson

University of Melbourne

April 2021

R: A Statistical Programming Language

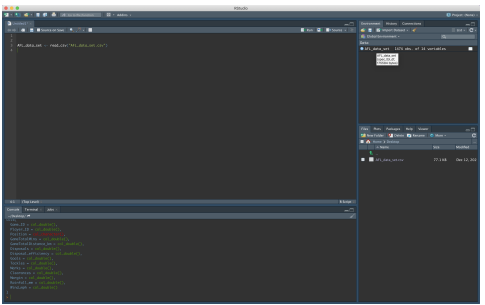
For this data task, you will be using R (and RStudio), 'the defacto language for data science.'

- R is an open-source, programming language that is used widely in academia and industry. It is an incredibly powerful tool for working with data.
  - R is well supported, with good documentation and a large, active online community.

R: A Statistical Programming Language

For this data task, you will be using R (and RStudio), 'the defacto language for data science.'

- RStudio is a development environment/GUI that provides an intuitive integration with R.
  - We can see the code we've run (via the editor). We can see the output of that code (via the console). We can see the 'objects' we interact with (via the environment pane).



R: Basics

In R, everything (vectors, lists, data frames, functions, etc.) is an object, and everything has a name.

- Multiple objects (e.g., data frames) can exist in the same environment/workspace (unlike Stata). Makes it much easier to merge data sets, work with transformed versions of a data set, etc.

You use functions, many of which come pre-written in packages, to perform operations on these objects.

- "Base" R has many useful functions, but the best data science tools come from external packages written by users (just like yourselves).

R: The Tidyverse

The Tidyverse is a widely-used, extremely-powerful suite of user-written packages for R (dplyr, tidyr, ggplot2, etc). It embodies best practice for wrangling, cleaning, analyzing, and visualizing data.

- Based on the philosophy of 'tidy' data: Each variable is a column, each observation forms a row, each type of observational unit forms a table ('tibble').
- 'The Bible' - R for Data Science: https://r4ds.had.co.nz/