Introducing Programming for Data Science: R and Python

By Sam Nelson, Head of the School of Data Science at Udacity

March 19, 2019

Data Science begins with programming as it's an essential skill for most data science and analytics work. In terms of popularity in the data science field, R and Python dwarf most other programming languages, and much has been written comparing the two. In fact, if you Google "r vs. python" you'll get over 55,000 results! So rather than give just one more opinion, we've pulled together some of the key takeaways on the topic to help you make a decision on which language to learn.

For both R and Python, I'll begin by introducing the programming language then describing some of the most significant advantages to using it.

What is R?

The programming language R was first released by a group of statisticians in 1994 and has since become widely used by statisticians, researchers, and data analysts around the world. It was created "for statisticians, by statisticians," and has a wide array of built-in functions and third-party libraries enabling data scientists to accomplish tasks at every step of the data science process.

Why choose R?

- **Statistical functionality — I**n R, statistical analysis is more natural with more focus on interpretability, not purely prediction. Additionally, R has some statistical functionality that Python does not. If your work involves advanced statistics, and the interpretability of your models is important, R may be a better choice. One primary example of this is with time-series data, where even companies that primarily use Python must move to R to model data collected from one day to the next.
- **Data visualization** — Data visualization is more intuitive and aesthetically pleasing in R compared to the data visualization libraries available in Python. If you plan to frequently create and present visualizations of your work, you should consider R. The ggplot2 library in R has long been the leader in this space surpassing the most popular libraries in Python, which include Matplotlib, Seaborn, and Bokeh.

- **Consistency** — R is a more purist language with more minor, incremental improvements, while Python has more significant and more frequent changes to keep track of. If you're the type of person who values consistency over frequent updates, R may be a good fit for you.
- **No library installations required —** R has more built-in data science functions while in Python you have to rely on third-party libraries. If you plan on sticking with basic analyses and want to avoid installations and third-party libraries, then start with R. For example, calculating the mean and standard deviation or building a linear model for a dataset in R can all be done in the base library. To perform these same tasks in Python, you would likely use at least two third-party libraries.

What is Python?

Python is a general-purpose programming language released in 1991. It emphasizes code readability and is commonly used not only for data science but also for general purpose software engineering. Like R, Python has a large repository of third-party libraries to enable data analysts, engineers, and scientists to do their work.

Why choose Python:

- **Task versatility** — Python is a more versatile language and is better suited to handle non-statistical or non-analytical related tasks. If you plan to combine software engineering and data science task together, Python's a better fit.
- **Simple production integration** — Python works more easily with other parts of a production system. R may require a bit more work to integrate because it was built by statisticians. If you plan to frequently deploy your analyses in production, Python makes it a bit simpler.
- **Multiple, useful libraries** — While R's libraries are smaller and more inconsistent, Python has a handful of data science libraries that suit most of your needs. The NumPy, Pandas, and Scikit-learn libraries cover much of the functionality for performing data science in Python. If you are looking to jump into the most common data science use cases in industry, Python is a good option.
- **Best for programmers** — Python is a more readable language and better suited for learning coding best practices. If you plan to expand your coding

abilities and learn additional programming languages in the future, we suggest learning Python.

Why choose both:

When researching the R vs Python question, most data science professionals suggest not limiting yourself to one or the other. By learning both R and Python, data analysts and data scientists can leverage the strengths of each language. One author suggested thinking of R as a Python library rather than a separate programming language. Our suggestion is to pick one language to start with and then strengthen your skill set by learning the other one.