Statistical Software IDEs: A Comparative Analysis

Scott Miner

Colorado State University – Global Campus

Statistical Software IDEs: A Comparative Analysis

**Brief Historical Background**

SAS Studio is a powerful IDE, first released in 1971, to give researchers a wide variety of data analytics capabilities (Ozgur, Colliau, Rogers, Hughes, & Myer-Tyson, 2017). R-Studio is an IDE for the R programming language, first released in February 2011 (“About RStudio,” n.d.). PyCharm Edu is an IDE for the Python programming language that contains interactive courses, first released in October 2014 (Filippov, 2014).

**Job Perspective**

Muenchen’s analysis of job postings on Indeed.com, collected on May 27, 2019, found Python to be the most popular programming language as it relates to the data sciences. Python was found in 27,374 of the job postings, more than any other language. The R Programming language ranked 5th, occurring in approximately 13,000 job postings (Muenchen, 2019). In a study in 2014, Muenchen (as cited by Ozgur et al.) found that SAS came in first, at nearly 12,500 postings, when the analysis required the postings to have two data analytics software programs present.

**Availability / Cost**

Out of the three programs compared, SAS Studio is the only software found not to be free or open-source. A basic, single one-year machine license for SAS Analytics Pro costs $10,010.00 (“SAS Products & Solutions,” n.d.). SAS Server pricing is negotiable and usually starts in the lower six-figure range (Dinsmore, 2014). R-Studio is free and open-source for the individual. R-Studio Team Standard is available for small teams with 5 to 20 Pro Users, and costs range from $21,220 - $50,943 per year. For larger teams, the pricing starts at $57,556 per year. PyCharm Edu is open-source and free to teachers and students.

**Ease of usage/learning**

Out of the three languages analyzed, Python is the easiest to learn. It is a general-purpose programming language with a syntax that resembles human speech (Ozgur et al., 2017). The consensus is still out on R-Studio and SAS Studio, however. Ooi (as cited by Ozgur et al., 2017), for instance, demonstrates that SAS requires three programming languages to accomplish the same tasks that R can do with one. Bob Muenchen, on the other hand, claims that R is difficult to learn due to its extreme power and flexibility (as cited by Dinsmore, 2014).

**Data handling capabilities**

SAS claims to be able to hold as many cases as can fit in memory, essentially asserting no data is too big (Ozgur et al., 2017). PyCharm Edu makes use of the *pandas* package, allowing data to be read into a DataFrame, much the same way that R does (Yau, n.d.). Data can then be analyzed, modified, and visualized.

**Visualization capabilities**

Each IDE can create multidimensional scatterplots, histograms, and bar charts with few lines of code. PyCharm Edu uses a programmatic approach, while R-Studio uses a layered approach—referred to as the grammar of graphics—which some data scientists prefer. Mendis (2019) found PyCharm images take up six times more disk space than R-Studio images.

**Statistical capabilities**

Typical statistical calculations that these languages can perform include averages, medians, maximums, minimums, and quartiles for all relevant columns. These IDEs can also compare the correlations between columns, as well as determine the distributions of columns. T-tests can also be performed to examine null hypotheses and predict future outcomes (Dietrich, 2015).

**Data task scenarios**

Common data tasks include loading data for analysis, modifications, splicing, cleansing, visualizations, and predictive analytics. Data may then be output to additional files for decision-making purposes and to support business end-users (McIntire, n.d.). One common programming task when beginning Deep Learning is to create a neural network that recognizes digits from the MNIST dataset (Yang, 2019). All IDEs offer tutorials for Artificial Intelligence, Machine Learning, and Deep Learning.

**Deep Learning Support**

Deep Learning support was extensive for all three IDEs. Both R-Studio and PyCharm Edu support the popular TensorFlow framework often used in neural network programming. R-Studio, in particular, offers a state-of-the-art online video course. The Keras API used in R was originally developed in Python and ported to R (Keydana, 2018). SAS offers support for six different neural networks, as well as deep learning tutorials and training (“How to Do Deep Learning With SAS,” n.d.).

**Customer service support and Community**

SAS Studio is said to invest heavily in technical support. It is a global enterprise and assists customers worldwide. The same technical support does not come with the free and open-source nature of R-Studio and PyCharm Edu. However, these IDEs can develop at a rapid pace because of the community input (Dinmore, 2014).

**Future Outlook of Tool**

All of the tools hold bright futures. Python has the brightest future as it is the most popular language in universities and dominates the job market. Nevertheless, a data scientist who encompasses all three skill sets will offer the most value to the job market.

References

About RStudio. (n.d.). Retrieved October 20, 2019, from <https://rstudio.com/about/>

Dietrich, D., Heller, B., Yang, B., & EMC Education Services (Eds.). (2015). *Data science & big data analytics: Discovering, analyzing, visualizing and presenting data*. Indianapolis, IN: Wiley.

Dinsmore, T. (2014, December 15). SAS Versus R Part Two | ML/AI. Retrieved October 20, 2019, from <https://thomaswdinsmore.com/2014/12/15/sas-versus-r-part-two/>

Filippov, D. (2014, October 30). JetBrains Debuts PyCharm Educational Edition. Retrieved October 20, 2019, from PyCharm Blog website: <https://blog.jetbrains.com/pycharm/2014/10/jetbrains-debuts-pycharm-educational-edition/>

How to Do Deep Learning With SAS. (n.d.). Retrieved October 20, 2019, from <https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/deep-learning-with-sas-109610.pdf>

Keydana, S. (2018, September 12). Getting started with deep learning in R | RStudio Blog. Retrieved October 20, 2019, from <https://blog.rstudio.com/2018/09/12/getting-started-with-deep-learning-in-r/>

McIntire, G., Martin, B., & Washington, L. (n.d.). Python Pandas Tutorial: A Complete Introduction for Beginners. Retrieved October 20, 2019, from Google Docs website: <https://docs.google.com/forms/d/e/1FAIpQLSdj0o8jwznrk2FMyiXKEqiSOalqOsQvI5KYaLaPWU5os7cKuQ/viewform?embedded=true&usp=embed_facebook>

Mendis, A. (2019, March). R vs Python for Data Visualization. Retrieved October 20, 2019, from KDnuggets website: <https://www.kdnuggets.com/r-vs-python-for-data-visualization.html/>

Muenchen, B. (2019, May 28). Data Science Jobs Report 2019: Python Way Up, Tensorflow Growing Rapidly, R Use Double SAS. Retrieved October 20, 2019, from R-bloggers website: <https://www.r-bloggers.com/data-science-jobs-report-2019-python-way-up-tensorflow-growing-rapidly-r-use-double-sas/>

Ozgur, C., Colliau, T., Rogers, G., Hughes, Z., & Myer-Tyson, E. “Bennie.” (2017). MatLab vs. Python vs. R. *Journal of Data Science*, *15*(3), 355–371.

SAS Products & Solutions. (n.d.). Retrieved October 20, 2019, from <https://www.sas.com/store/products-solutions/cSoftware-p1.html?storeCode=SAS_US>

Yang, D. (2019, May 17). Tutorial 1: MNIST, the Hello World of Deep Learning. Retrieved October 20, 2019, from Medium website: <https://medium.com/fenwicks/tutorial-1-mnist-the-hello-world-of-deep-learning-abd252c47709>

Yau, C. (n.d.). Data Frame | R Tutorial. Retrieved October 20, 2019, from http://www.r- tutor.com/r-introduction/data-frame