IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

A COMPARATIVE REVIEW BETWEEN PROGRAMMING TOOLS USED IN DATA SCIENCE

[1]Saloni Jackeray, [2]Anjani Sruti Doradla, [3]Ritika Rane, [4]Prof. Brinal Colaco

[1]Student, [2]Student, [3]Student, 4Professor
[1]Department of Computer Engineering,
[1]A.P. Shah Institute of Technology, Thane, India

Abstract: In today's era, R and Python are one of the most promising tools used in all futuristic technologies. Both R and Python are open source programming languages with an abundant collection of libraries that are added continuously to their catalogue. The focus of this paper is to compare the two technologies and address the confusion of many individuals regarding the choice programming language to be used. This review proposes a comparative study between Python and R, explaining the benefits and highlighting the differences between the two. Both Python and R are well evaluated based on their performance parameters with reference to topics like Big Data, Data Analysis, Internet of Things, Machine Learning and other domains related to Data Science. Python being an object oriented programming language, is a good tool to execute algorithms that are used in production. On the contrary, R is a programming language which is used widely by professional such as statisticians, data analysts. This paper draws a conclusion for the choice of language based on the problem statement in hand.

Index Terms - Python, R, Data science, Data analysis.

I. Introduction

Extensive amounts of data available in every organization has created an opportunity to analyze and obtain significant insights from data. Data Science is about drawing out, visualizing and analyzing data to effectively extract useful information. It requires an exceptionally skilled data scientist who can use numerous statistical, programming and machine learning tools to foresee the probability of events and takes data-driven decisions. A Data Scientist utilizes various tools and conventions to acknowledge redundant patterns within the data. These tools include R, SQL, Hadoop, Weka, Python and many more. These tools are used to generate intelligent systems that can take individual decisions based on previous datasets. Data Science involves the utilization of scientific procedures and algorithms to examine and summarize the data. Certain programming languages developed for this role implement these methods. The most widely used and great programming tools used in Data Science are Python and R. Unlike other tools used in this domain, Python and R are portable and open-source making them the most versatile programming languages available for Data Science. [2] Python has distinctive features and is effortless to use, learn and apply in Data Science application. In addition to this, Python comes with the capacity to integrate every part of the workflow due to the presence of extensive number of built-in libraries. Python has gained acclaim for its code readability, speed, flexibility and error checking than any other programming language.[1] Being a high level language, it has high level data types incorporated such as malleable arrays, pointers and dictionaries. On the contrary, R caters to all statistical computations, time series forecasting and clustering problems. [7] This is the best tool used for beautiful and attractive graphs, charts and visualization. Another strong asset of R is its ability to deal with complex mathematical problems like linear algebra. This makes R the best possible language for statistical analysis and neural networks. An essential ability of R is to interface with NoSQL databases and compute unstructured data. This is very useful in Data Science applications where a pool of data has to be studied and makes it easy for the data scientists to interpret profitable insights from huge data. It depends on the individual data analyst and data scientists to explore parameters such as scalability, flexibility, usability, ecosystem and data handling capabilities. Python and R are very alike yet differ in their own ways which makes it very arduous for the data analyst and data scientist to select one amongst them.

II. LITERATURE REVIEW

Both Python and R are approachable languages that give programmers and developers the ability to work with great precision by using their extensive libraries and functions. Zhang et al. [1] in their paper mention how Python has been the most approachable land powerful language for their research. They have created tools for simplifying data management. Suryansh Singh ET. al. [2] in their paper have given a comparative study of Python and R for topics like machine learning, natural language processing (NLP), and data analytics. In this study it is concluded that there is a distinction in the use of both programming languages in terms of the type of application and its requirements for example R is better for projects related to data/time series analysis and Python is efficient in applications related to machine learning and NLP. Jim Brittain et al. [3] in their paper have presented a quantitative analysis on experiments of Python, R and SAS. The comparison is based on parameters like code length output and results which has led to their take on the appropriate tool to be used for data science related applications. These studies have provided the pros and cons of data science tools that are application specific. Our review focuses mainly on the most important aspects of any programming language in the comparative study section.

III. A COMPARATIVE STDUY

Data Science is a compelling discipline that filters and helps many organizations to change raw data into understandable insights. The objective of programming tools like Python and R is to overcome all the challenges in wide range of data. However, we are of the opinion that it is best to learn and master one tool at a time. Python and R are the best practice tools that most data scientists and data analysts use. The confusion arises when it's time to choose between these two languages as each one is significant in its own way.

Python being one of the most vital part of Data Scientist's toolbox as it is tailor-made for performing repetitive tasks, data manipulation and working with huge amounts of data. Data Scientists come across multiple libraries such as NumPy, Panda, Matplotlib which assist the data scientist to carry out his/her function. [10]

R has been fundamentally used by academic and resource sectors and is very useful for exploratory data analysis. In today's era, companies and organizations are expanding rapidly with increase in the amount of data. R being popular in academics, pharmaceutical, finance, marketing and media, is mostly implemented by individuals who have less or no coding experience like scientists and statisticians making it easier to learn. [4]

This study focuses on comparing the two languages mainly on four parameters namely- Flexibility, Ease of Learning, Ecosystem, Graphics and Visualizations.

This comparative study aims to help the individual make a decision of which programming tool to use as per their requirements.

A. FLEXIBILITY

Python is resilient for building and generating something from scratch. As Python can be used by many kind of developers such as data scientists, data analysts, web developers, android developers and business analysts to produce their applications and research.[6] In contrast, R is used to compute statistical model and tests which are promptly available and easily used.

Python being an object oriented programming language is easy to interpret and compile making it flexible for building data science models. However, R is a free extensible programming language allows the programmer to implement their own methods, procedures and functions. [10]

There are numerous Python IDE's which provides complete provision to programmers and data scientists for software development. These IDE's help to substantially lower down the cost of organizing output code, files and methods. These IDE's of Python comprise of the most popular Jupyter notebooks and Spyder which provides grip to program easily for data science applications.[9] Even R has built in IDE to bring the workflow together and help in powerful authorizing and debugging-known as R Studio. It comes in two formats: open source and commercial addition as R studio desktop and R studio server respectively. The former is used for regular desktop applications to run the programs locally while the latter is used via web browser running on a Linux server.

Python is a repository of many libraries which contribute in solving complicated mathematical problems. It requires some kind of installation for data analysis. Here are the list of libraries that help in data science and its periphery.

- **Pandas**: It is used for renaming, sorting, indexing and merging data frame. It is created in such a way that it provides high performance, fast and lucid data visualisation, reading and manipulation.[8]
- **SciPy and NumPy**: They are one of the most basic packages in python. NumPy is a general purpose array processing package. NumPy performs basic array operations like slicing, reshaping, flattening, adding and multiplying and advance array operations like slit into sections, broadcast and stack arrays. It also works for date, time and linear algebra. SciPy is constructed on NumPy array objects and it includes tool like Pandas, Matplotlib, and SymPy. It is used for scientific programming routines as calculus, ordinary differential equation, signal processing and integration.
- **Matplotlib:** It is basically a plotting library for Python. It helps to creates stories with the visualised data. It converts installed plots into applications with the help of object oriented API provided by Matplotlib. It is equipped with many visualisation capabilities or the users can create their own visualisations like histograms, bar and pie charts, area and scatter plots, line and contour plots and spectrograms.

The libraries in R has extensive range of packages and free libraries put down by the user community. Here are some most useful downloaded R packages for data science among the other beneficial packages. [10]

- **To Manipulate Data:** dplyr tidyr, stringr packages are used for rapid data manipulation.

Dplyr comes with five build-in function: Select, arrange, summarize and mutate. It can be used to work with remote database tables as well as local data frames.

Tidyr can be used to change the framework of dataset. It is basically used to convert the data into well-ordered format.

Stringr is a tool to define character string and regular expression.

- **To Visualize Data**: Some of the best visualization libraries in R are ggplot2 and rgl.

Ggplot2 is the most prominent package for creating attractive graphics and lets one implement the "grammar of graphics" to fabricate layered and customizable plots. Rgl is one of the primary 3D visualization package in R. It provides the data scientist with geometry primitives such as triangles, lines, points, quadrilateral, text and many more.

- **To Model Data:** Modelling of data uses caret as a package to train the data which can be used for resampling, evaluating and tuning the model according to the performance parameter. Caret is an acronym for classification and regression training. This package is enough to solve any machine learning problem and similar problems such as data pre-processing, feature selection, splitting etc.[7]

B. EASE OF LEARNING

If we talk on ease of learning, Python emphasizes on code readability and user-friendly behavior. The complicated functionalities of R makes it tougher to implement and develop applications.

Python is beneficial for programmers who start from scratch to the programmers who work in large scale sector. While, R is a programming language which is easier to practice on at the start but eventually becomes harder to develop. The former is considered an easy language for all kinds of programmers while the latter is easier for experienced programmers. [10]

Python is basically chosen by the data analyst and the data scientist when their tasks and code need to be assimilated into apparatus for generation of production database or combined with web applications. Whereas, R is normally used when the tasks and code involves standalone computing on independent servers.

Python being a high level programming tool in Data Science, the programmers working in this field prefer Python to run longer codes and fast applications. In short, Python is basically used for its rapid execution. On the contrary, R is considered as a low level programming language used in Data Science because of its slow execution making it more efficient for smaller code and simple procedure.

Since Python is improving to newer version for data analysis, the error and issues in python packages are being reduced day by day, packages like NumPy and Pandas make it simpler for the data scientist to build new applications. It is considered best for mathematical and parallel computation. Packages need to be installed in Python to be used. Whereas R consists of numerous packages which are readymade for the programmers and data scientists. Ready-made packages make it labor saving for analysis. It does not include any installation for the usage of the packages. [6]

C. ECOSYSTEM

Python has a strong ecosystem and is very often considered as one of the lucid programming language to learn and read. The syntax of Python is very straight forward and is quite similar to English language.
Example: print (“Hello!”)
On the contrary, R has an ecosystem which is substantial and forefront. The interface packages are accessible to work with open-source languages.
Python semantics is refined, easy to perceive and type. However, R ecosystem permits the user to string their progress simultaneously which makes it efficient to use for data analysis.
For building data science model and machine learning products, Python is extensively used along with its web structure but for doing this the Python libraries are required which are not pre-installed and hereby requires installation of the libraries which is a whole another part. Whereas, R is suitable for data analysis due to the large number of packages which are pre-installed, tests that are readily available and the convenience of applying formulas. R can also be used to perform basic level of data analysis without installing any packages. [10]

The repositories in Python that contain all libraries are Python Package Index (PyPi) and Anaconda. Modifying the packages present in these repositories is a complicated practice for the user. Data science with Python has a huge content of packages that help in providing solution for ordinary problems that are faced by the users. Python can be executed on numerous platform like Linux, UNIX, Windows, and Mac. Therefore it is known as platform-independent that is, it can be coded in one platform and can be executed on any other platform.
Packages in R consists of dataset, R functions and compiled code. All these packages are put aside in a directory called R library. R packages are available in following platforms -

- **Bio-conductor:** It is an open source development software for comprehending genomic dataset created from lab experiments in biological background.
- **Comprehensive R Archive Network (CRAN):** This is used to lessen the load of network. “Task views “permits you to search packages name by name and supply to immediately install packages according to the interest.

Many of the packages can be implemented by the above mentioned sources and the user can browse other packages of their interest through R documentation.

D. GRAPHICS AND VISUALISATIONS

The graphical representation of data and information by making use of visualization tools which gives a convenient and handy way to foresee and analyze outliers , patterns and trends in data in the form of charts, maps, graphs and many more of such kind.

Both Python and R contain some attractive libraries used for data visualization. Libraries in Python are a little complex to understand and implement yet they give a neat output. The data Scientists believe that graphics and pictorial representation are interpreted better by common users more efficiently. To cope up with this, R comprises of many number of packages which have an advance level of graphical capabilities. Packages in R are less complex or easy to use because the statistical models and tests are promptly available and quick to use.

Python which is a common programming language, used in data science for analyzing and comprehending data which consists of blend of solutions for graphics and visualizations. Matplotlib is the commonly used package for data visualization generated by NumPy. It consists of many functions to plot the data points and provide the user with greater insights of data. On the contrary, R comprises of packages which are basic and already installed. R consists of about more than hundred functions to construct different data plots. [8]

IV. CONCLUSION

In this review paper, we talk through the advantages and disadvantages of Python and R by doing a comparative study. Like every perspective of life, we cannot be biased towards one programming language between these two. Considering the modifications and advancements in the field of technology, Python and R seem to be succeeding in their domains respectively which makes it difficult to pass a judgement in deciding a better language between them. Both of them are efficient languages and used by the masses according to their requirements. For a new comer just entering the field of data science having a good background in statistics and economics, it is suggested to prefer R as the programming language. Despite of Python being a powerful and multipurpose programming language, it is considered to be a friendly language to the beginners. On the other hand, R is extensively used to build applications in Data Science environment. Whereas, if the user wants to build a data science model with machine learning and deep learning fundamentals then selecting Python is well advised. After learning the fundamentals of data science with R, the user should without a doubt can start learning Python. Eventually, data scientists use both the languages to develop their skills and use them to build appealing projects and applications.

Our study and review on this topic, however, remain unchanged as we restate on the factuality that decides the better suited language relied on the respective requirements of the data science tasks.

REFERENCES

[1] Zhang, Jing, Hongxia Luo, and Xueqing Zhang. "Application of python language and arcgis software in RS data management." *2011 International Conference on Remote Sensing, Environment and Transportation Engineering*. IEEE, 2011.

[2] Suryansh Singh. “R vs Python, Why you should learn both?”*Quadratyx Power of Insight.2017.*

[3] Jim Brittain, Mariana Cendon, Jennifer Nizzi, John Pleis. “Data Scientist’s Analysis Toolbox: Comparison of Python, R, and SAS Performance.”*SMU Data Science Review.* SMU Data Center, 2018.

[4] Ceyhun Ozgur, Taylor Colliau, Grace Rogers, Zachariah Hughes,Elyse “Bennie” Myer-Tyson.” MatLab vs. Python vs. R.” *Journal of Data Science 15,355-372.*2017.

[5] W. McKinney, "Chapter 1- Preliminaries,*" in Python for Data Analysis, Sebastopol, O'Reilly Media*, 2013, p. 3.

[6]Python for Beginners - Python Training Course - Udemy. Retrieved from: https://www.udemy.com/python-for-beginners/?siteID=oCUR7eOwwME-8mj81nbpWjfGzuuaYiVpTg&LSNPUBID=oCUR7eOwwME.

[7] The R Foundation. (2017). The R Project for Statistical Computing. Retrieved from:
https://www.r-project.org/

[8] G. Piatetsky, "Four main languages for Analytics, Data Mining, Data Science," 18 August 2014. [Online]. Available: https://www.kdnuggets.com/2014/08/four-main-languages-analytics-data-mining-data-science.html. *[Accessed 10 November 2017]*.

[9] J. K. Millman and M. Aivazis, "Python for Scientists and Engineers," *Computing in Science & Engineering*, vol. 13, no. 12, pp. 9-12, 2011.

[10] When should I use Python and R? Data Camp.
Available: https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis