Experiment No. 1

Aim: Introduction to Data analytics libraries in Python and R.

Objective- Understand the use of Python and R, Toeffectively use libraries for data science.

Description:

Why Choose Python?

Python is a general-purpose, open-source programming language used in various software domains, including data science, web development, and gaming.

Launched in 1991, Python is one of the most popular programming languages in the world, occupying the top position in several programming language popularity indices, such as the TIOBE Index and the PYPL Index.

One of the reasons for the worldwide popularity of Python is its community of users. Python is backed by a vast community of users and developers who ensure the smooth growth and improvement of the language, as well as the continuous release of new libraries designed for all kinds of purposes.

Python is an easy language to read and write due to its high similarity with human language. In fact, high readability and interpretability are at the heart of the design of Python. For these reasons, Python is often cited as a go-to programming language for newcomers with no coding experience.

Over time, Python has been gaining popularity in the field of data science thanks to its simplicity and the endless possibilities provided by the hundreds of specialized libraries and packages that support any kind of data science task, such as data visualization, machine learning, and deep learning.

Why Choose R?

R is an open-source programming language specifically created for statistical computing and graphics
Since its first launch in 1992, R has been widely adopted in scientific research and academia. Today, it remains one of the most popular analytics tools used in both traditional data analytics and the rapidly-evolving field of business analytics. It ranks 11th and 7th position in the TIOBE Index and the PYPL Index, respectively.
Designed with statisticians in mind, with R, you can use complex functions within a few lines of code. All kinds of statistical tests and models are readily available and easily used, such as linear modeling, non-linear modeling, classifications, and clustering.

The extensive possibilities **R** offers are mostly due to its huge community. It has developed one of the richest collections of data-science-related packages. All of them are available via the Comprehensive **R** Archive Network **(CRAN).**

Another feature that makes **R** particularly remarkable is the power to generate quality reports with support for data visualization and its available frameworks to create interactive web applications. In this sense, **R** is widely considered the best tool for making beautiful graphs and visualizations.

R vs Python: Key Differences

Purpose

While Python and R were created with different purposes -Python as a general-purpose programming language and R for statistical analysis-nowadays , both are suitable for any data science task. However, Python is considered a more versatile programming language than R, as it's also extremely popular in other software domains, such as software development, web development, and gaming.

Type of Users

As a general-purpose programming language , Python is the standard go-to choice for software developers breaking into data science. Plus, Python 's focus on productivity makes it a more suitable tool to build complex applications. By contrast, R is widely used in academia and certain sectors, such as finance and pharmaceuticals. It is the perfect language for statisticians and researchers with limited programming skills.

Learning curve

Python 's intuitive syntax is considered one of the closest programming languages to English. This makes it a very good language for new programmers , with a smooth and linear learning curve. Although R is designed to run basic data analysis easily and within minutes, things get harder with complex tasks, and it takes more time for R users to master the language. Overall, Python is considered a good language for beginner programmers. R is easier to learn when you start out, but the intricacies of advanced functionalities make it more difficult to develop expertise.

Popularity

Although new programming languages, like Julia, are recently gaining momentum in data science, Python and R remain the absolute kings in the discipline. However, in terms of popularity -always a very slippery concept- the differences are striking. Python has consistently outranked R, especially in recent years. Python ranks first in several programming language.

Common Libraries

Both Python and R have robust and extensive ecosystems of packages and libraries specifically designed for data science. Most packages in Python are hosted in the Python Package Index **(PyPi),** whereas **R** packages are normally stored in the Comprehensive R Archive Network (CRAN).

Below you can find a list of some of the most popular data science libraries in R and Python.

R packages:

dplyr: It is a data manipulation library for R.

tidyr: a great package that will help you get your data clean and tidy.

22plot2: the perfect library for visualizing data.

Shiny: It is the ideal tool for creating interactive web apps directly from R

Caret: one of the most important libraries for machine learning in R.

Python packages:

NumPy: p rovides a large collection of functions for scientific computing.
Pandas: perfect for data manipulation.
Matplotlib: the standard library for data visualization.
Scikit-learn: is a library in Python that provides many machine learning algorithms.
TensorFlow: a widely used framework for deep learning.

Common IDEs

An IDE, or Integrated Development Environment , enables programmers to consolidate the different aspects of writing a computer program. They are powerful interfaces with integrated capabilities that allow developers to write code more efficiently.

In Python, the most popular IDEs in data science are Jupyter Notebooks and its modem version JupyterLab, as well as Spyder.

Data Analytics packages in R :

1. googkleVis - The googleVis package provides an interface between R and the Google Charts API. Google Charts offer interactive charts which can be embedded into web pages. The functions of the googleVis package allow the user to visualize data stored in R data frames with Google Charts without uploading the data to Google.
2. MissMDA - We present the R package missMDA which performs principal component methods on incomplete data sets, aiming to obtain scores, loadings and graphical representations despite missing values. Furthermore, missMDA can be used to perform single imputation to complete data involving continuous, categorical and mixed variables.
3. RRF - Feature Selection with Regularized Random Forest. This package is based on the 'randomForest' package by Andy Liaw. The key difference is the RRF() function that builds a regularized random forest.
4. FActoMiner - FactoMineR is an R package dedicated to multivariate Exploratory Data Analysis. It is developed and maintained by François Husson, Julie Josse, Sébastien Lê, d'Agrocampus Rennes, and J. Mazet.

5. LSMeans - The ls means package provides a simple way of obtaining least-squares means and contrasts thereof. It supports many models fitted by R core packages (as well as a few key contributed ones) that fit linear or mixed models, and provides a simple way of extending it to cover more model classes.

Data Analytics packages in Python :

1. Theano - Theano is a Python library that allows you to define, optimize, and efficiently evaluate mathematical expressions involving multi-dimensional arrays.
2. Keras - Keras is an open-source software library that provides a Python interface for artificial neural networks. Keras acts as an interface for the TensorFlow library. Up until version 2.3, Keras supported multiple backends, including TensorFlow, Microsoft Cognitive Toolkit, Theano, and PlaidML.
3. Pybrain - PyBrain stands for Python-Based Reinforcement Learning, Artificial Intelligence, and Neural Networks Library. It is a modular machine learning library n python that contains very powerful and easy-to-use algorithms used to aid in a variety of machine learning tasks.
4. Scrapy- It is a Python framework for large scale web scraping. It gives you all the tools you need to efficiently extract data from websites, process them as you want, and store them in your preferred structure and format.
5. Plotly - It is a free and open-source data visualization library. I personally love this library because of its high quality, publication-ready and interactive charts. Boxplot, heatmaps, bubble charts are a few examples of the types of available charts.

Conclusion:

Python, renowned for its versatility and extensive community support, excels in general-purpose programming and is favored by beginners for its readability. Meanwhile, R, tailored for statistical computing, offers powerful tools for statisticians and researchers, particularly in generating quality reports and visualizations. Despite their distinct origins and learning curves, both languages are indispensable in the realm of data science, supported by robust ecosystems of libraries and packages catering to various analytical needs. The experiment highlighted popular libraries in each language and discussed common integrated development environments (IDEs), providing a comprehensive overview of tools available for data analysis in Python and R.