https://dimensionless.in/why-r-programming-in-data-science/

Why R Programming in Data Science?

by Suman Dey | Apr 10, 2019 | Data Science, R Programming | 0 comments

Data Science is everyone's word of the mouth in the current analytical eco-space. The study of Data Science which encompasses various subjects like Machine Learning, Deep Learning, Artificial Intelligence, Natural Language Processing, and so on has made tremendous advancement in the recent past.

Data Science is not something that emerged recently. It was there since computers were invented as the first Data Science application was classifying an email as Spam or Not Spam based on certain trends in the mail. However, the recent hype is a result of the massive amounts of data that are available, and the huge computational capacity that modern computers possess.

In terms of career, Data Science is considered as one of the most lucrative jobs in the 21st with salaries next to none. Hence, out of the curiosity to mine insights from the data, and also for a better career, professionals from various disciplines such as Healthcare, Physics, Marketing, Human Resource, IT, want to master the state-of-the-art Data Science methodologies. To be called a Full Stack Data Scientist, one needs to master a plethora of skills as mentioned below.

- **Statistics and Probability – The first, and arguably the most important part of Data**

Science as various statistical methods are used to make assumptions from the data.

- **Programming – One needs to master at least one programming language out of**

Python, R, and SAS.

- **Machine Learning – To make predictions from the data, one needs to be aware of**

the several programmed algorithms and understand their usage for the right application.

- **Communication – Extracting insights from the data are useless unless it is communicated in layman terms to the business and the stakeholders who would make crucial decisions based on your analysis. Apart from these four basic skills, there are few other skills like building data pipelines are also important, but on most occasions, an organization would have a separate team for that.**

Why Programming is Needed for Data Science?

In layman terms, Data Science is a process of automating certain manual tasks to mitigate the resource, budget, and time constraints. Thus, learning to code is an important component to automate those tasks. To build a simple predictive model, the data set should be first loaded and cleaned. There are several libraries, and packages available for that. You need to choose the language to code and use those libraries for such operations. After the data is cleaned, there are several programmed algorithms which need to be used to build the predictive model.

Now, each algorithm is a set of a class which needs to be imported first, and then an object is created for that class which would use the methods, or the functions associated with that particular class. Thus, this entire process is a concept of Object Oriented Programming. Even, to understand the process behind the algorithms, one needs to be familiar with programming

Why R Programming is Used?

There is an ongoing debate about which is the best programming language for Data Science. It never harms to master all the 3 languages but one needs to be expert in a particular language, and understand its various functionalities in different situations.
The choice of language depends on interest, and how comfortable the person is to program in that language. Python is generally considered as the Holy Grail due to its simplicity, flexibility, and the huge community which makes it easier to find solutions to all sorts of problems faced during the building stage. However, R is not far behind either as people from different backgrounds other than IT, seems to prefer R, as their go-to language for Data Science.

R is an open-source programming language which is supported by the R Foundation and is used in statistical computing, and graphics. Like Python, it is easy to install and is better than SAS (comment not share by Bernard Clément) which however is high-level, and easy to learn designed additionally for Data Manipulation.

The graphical representations and the statistical computations of the data gives R an edge over Python in this regard. Additionally, the programming environment of R has input, and output facilities, and several user-defined recursive functions. In the early '90s, R was first developed, and since then its interface has been improved with constant efforts. R has made an outstanding journey from being a text editor to R studio, and now to the Jupyter Notebooks which has intrigued all the Data Scientist across the world. Below are some of the key reasons why R is important in Data Science.

- Academic Preference – R is one of the most popular languages in universities, and it is the language that many researchers use for their experiments. In fact, in several Data Science books, all the statistical analysis is done in R. This academic preference creates more people with the proficiency in R. As more students study R in their undergraduate or graduate courses, it would help them perform statistical analysis in the industry.
- Data Pre-processing – Often the dataset used for analysis requires cleaning to make it ideal for building a model which is a time-consuming process. R comes to the rescue in such cases as it has several libraries, and packages to perform data wrangling. Some of its packages are-
  1. dplyr – One of the popular R package used for data exploration, and transformation.
  2. table – Data aggregation is simplified with this package as well as the computational time to manipulate the dataset is reduced.
  3. readr – This package allows to read the various forms of data ten times faster due to the non-conversion of characters into factors.
- Visualization – R allows the visualization of various structured or tabular data in graphical form. It has several tools which perform the task of analysis, visualization, and representation. ggplot2 is the most popular package in R for data visualization.
  ggedit is another package which users the aesthetics of a plot are correct.
- Specificity – The goal of the R language is to make data analysis simpler, approachable, and accurate. As R is used for statistical analysis, it enables new statistical methods through its libraries. Moreover, the supportive community of R makes which helps one to get all the required solution of a problem. The discussion forums of R is next to none when it comes to statistical analysis. More often than not, there is an instant response to any question posted in the community which makes helps Data Scientists in their project.

- **Machine Learning – Exploratory data analysis is the first step in an end-to-end Data Science project where the data is wrangled and analyzed to extract insights through visualization. The next step is to build predictive models with the help of that cleaned data to solve various business problems. In Machine Learning, one needs to train the model first where it could capture the underlying trends in the data, and then make a prediction on the unknown data. R has a list of extensive tools which simplifies the process of developing the model to predict future events. Few of those packages are –**
    1. **MICE – It deals with missing values in the data.**
    2. **PARTY – To create Data partitions, this package is used.**
    3. **CARET – The classification and regression problems could be solved with the CARET package.**
    4. **Random FOREST – To create a decision tree.**

- **Open Source – The open source feature of R makes it suitable to be run on any platform such as Windows, Linux, Mac, etc. In fact, there is an unlimited scope to play around with the R code without the hassle of cost, limits, license, and so on. Apart from a few libraries which are restricted to commercial access, rest could be accessed for free.**
- **All-in-one Package Toolkit – Apart from standard tools which are used for various data analysis operations like transformation, aggregation, etc., R has several tools for statistical models like Regression, GLM, ANOVA which are included in a single object-oriented framework. Hence, instead of copy, and paste, this feature allows to extract the required information.**
- **Availability – As R is an open-source programming language with a huge community, it has a plethora of learning resources making it ideal for anyone starting out in Data Science. Additionally, the exploration of the R landscape makes it easier to recruit R developers. R is rapidly growing in popularity and it would scale up in the future. Various techniques such as time-series modeling, regression, classification, clustering, etc., could be practiced with R making it an ideal choice for predictive analytics.**

There are several companies who have used R in their applications. For example, the monitoring of user experience in Twitter is done in R. Also, in Microsoft, professionals use R on sales, marketing, Azure data. To forecast elections, and improve traditional reporting, the New York Times uses R language. In fact, R is used by Facebook as well for analyzing its 500TB of data. Companies like Nordstrom ensures customer delight by using R to deliver data-driven products.

Conclusion –

Data Science is the sexiest job of the 21st century, and it would remain so for years to come. The exponential increase in the generation of data would only allow more development in the Data Science field, and there could be a gap in supply-demand at a certain age.