**Python for Data Science: Introduction**

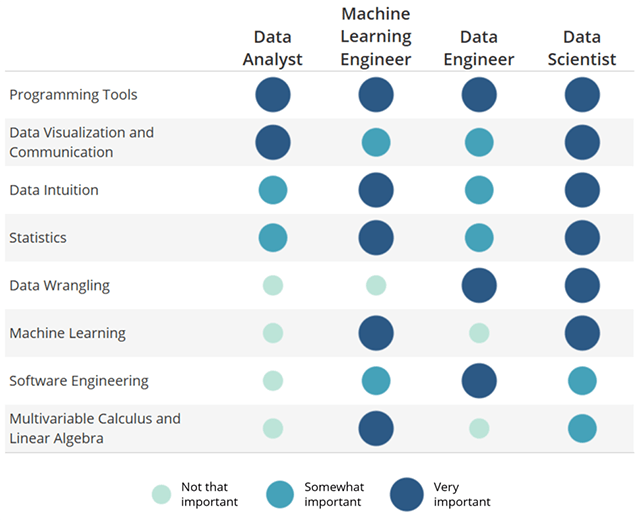
Data Scientist is regarded as the sexiest job of the 21st century with Data Science being everyone’s word of mouth these days. The recent hype in the Data Science space is a result of the advancement in the computational capacity of a system.

However, to be a Data Scientist is not a piece of cake, and indeed, everyone doesn’t have the specific skills to call themselves a Data Scientist. A lot of big companies still work on several old tools which they regard as the work of a Data Scientist and often tend to lure professionals into the trap of believing so.

**Skills required for Data Science**

Anyone, who wants to enter into the Data Science space, and be a part of the Data Science community, the first and the foremost criteria is to have an active data intuition and a willingness to play around data. The love for data is the start of the journey of becoming a Data Scientist.

Once the curiosity is there, the technical skills could be mastered henceforth.



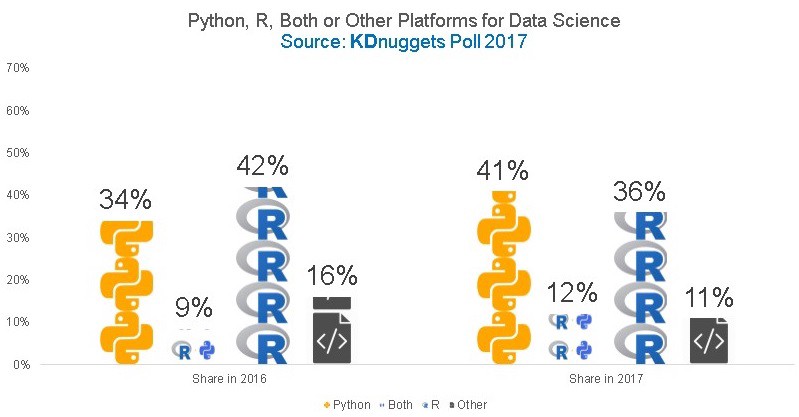
The above image illustrates some of the skills a Data Scientist should master, such as Programming, Statistics, Machine Learning, Mathematics, and so on.

In this article, we would know about the Programming tools needed to land a Data Science role and take an introductory look into the Python programming language for Data Science.

**Which Programming Language is the best for Data Science?**

Now, there is considerable debate regarding the best programming language that is out there for Data Science, and it isn’t going to conclude anytime soon.

The two of the most used programming tool for Data Science are R and Python. Both the language are equally reliable and provides the flexibility to carry out any Data Science task with ease and utmost precision.



This poll by KDnuggets clearly shows that the popularity of Python rose by seven percent in 2017 compared to 2016 and it continues to grow in the future. The reason why Python is now emerging as the globally accepted for Data Science is as follows: -

* *Simplicity*- The language is easy to use without any complicated syntax. Hence for any newcomer, it could take less time to master compared to R.
* *Community Support* – Python has a large community, i.e., there is always a solution to the bug you are facing.
* *Extensive features* – Python has a rich list of libraries and features which makes it the go-to language for any analytics task.
* *Larger Scope* – Apart from Data Analysis, Python could also be used in Web Development, Quality Assurance, Scripting, and so on.

Now that, we have clearly picked Python over R as the language for Data Science, in rest of this article we would have a look at the Python libraries needed for Data Science.

**Python Libraries for Data Science**

Python has a broad collection of libraries and modules which are used in many Data Science operations like – Exploratory Data Analysis, Feature Engineering, Machine Learning, Data Visualization, Deep Learning, and so on.



Here, we would look at some of the most popular libraries and where it is used.

* **Pandas**: - Pandas is one of the most widely used Python libraries for Data Analysis. In any Data Science project, the first step is to perform exploratory data analysis, and pandas provide the functionality to load any CSV or a JSON file into a DataFrame and then execute operations like – data cleaning, filtering, merging, sorting, NULL checking, and so on. You can also perform several query-like operations like group by, join, etc.
* **Numpy**: - Numpy or Numeric Python is used for matrix or vector operations, and it is very fast compared to an array. If you are dealing with numerical calculations, then Numpy is a must to master. It supports multi-dimensional array operations.
* **Matplotlib**: - Data Visualization is a crucial step in any Data Science project as it helps to find out the exact features to use to get the correct prediction. For visualization, matplotlib is the first choice tool as it provides various functionalities to plot the data in multiple forms and extract the necessary patterns from it.
* **Seaborn: -** Seaborn is another data visualization tool which is built on top of matplotlib but provides additional features to make our graphs much more interactive and extract better findings from the data. Also, it’s easier to master than matplotlib, and plotting is much more intuitive.
* **Scikit-learn: -** So far, we have seen the tools or libraries which deals with descriptive statistics. However, once the data is analyzed, the next step is to make predictions, and scikit-learn is the most widely used Python library for Machine Learning. It has several algorithms which are used in classification, regression, and clustering problems.
* **TensorFlow: -** Deep Learning is a branch of Machine Learning which requires higher computational power and a massive volume of data compared to a regular machine learning task. There are few libraries used in Deep Learning problems with Google’s TensorFlow being the most popular among them.

**Conclusion**



In this article, we got an introductory knowledge of Data Science and the Python Language used for it. In the next series of blogs, we would dive down into each of the libraries and see how they are used.