Why Python is the best programming language for data science

Data scientists need to deal with complex problems, and the problem-solving process basically involves four major steps - data collection & cleaning, data exploration, data modeling and data visualization.

Python provides them with all the necessary tools to effectively carry out this process with dedicated libraries for each step that we will discuss later in this article. It comes with powerful statistical and numerical libraries such as Pandas, Numpy, Matplotlib, SciPy, scikit-learn, etc.and advanced deep learning libraries such as Tensorflow, PyBrain, etc.

Moreover, Python has emerged as the default language for AI and ML, and data science has an intersection with Artificial Intelligence. Therefore, it is not at all surprising that this versatile language is the most used programming language among data scientists.

This interpreter-based high-level programming language is not only easy to use, but it also equips data scientists to implement solutions and, at the same time, follow the standards of required algorithms.

Data collection & cleansing

With Python, you can play with almost all sorts of data that are available in different formats such as CSV (comma-separated value), TSV (tab-separated value) or JSON sourced from the web.

Whether you want to import SQL tables directly into your code or need to scrape any website, Python helps you achieve these tasks easily with its dedicated libraries such as PyMySQL and BeautifulSoup, respectively. The former enables you to easily connect with a MySQL database to execute queries and extract data while the latter helps you to read XML and HTML type data. After extracting and replacing values, you would also need to take care of missing data sets during the data cleansing phase and replace non-values accordingly.

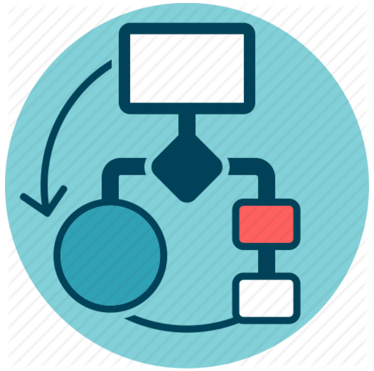
Furthermore, if you get stuck with any particular dataset, then you can get a solution by doing a Google search about that dataset and Python, thanks to the strong and vibrant Python community!

Why Python is the best programming language for data science

Data exploration

Now that your data is collected and tidied up make sure it is standardised across all the data collected. Now that you have clean data, figure out the business question that needs to be answered and then convert that question into a data science question.

For that, explore the data to identify their properties and segregate them into different types such as numerical, ordinal, nominal, categorical, etc., in order to provide them required treatments.
Once data is categorised as per their type, NumPy and Pandas, the data analysis Python libraries, will help you to unleash insights from the data by allowing you to manipulate it easily and efficiently.
Now that your data is ready to be used, it's time to jump onto AI and machine learning for data modelling.



Data modelling

This is a very crucial phase in the data science process wherein you would strive to minimize the dimensionality of your data set.
Python has many advanced libraries to help you tap the power of machine learning in performing the tasks involved in data modelling.

Would you like to perform a numerical modeling analysis of your data? Just reach out for **Numpy** in your toolkit! With **SciPy** you can easily perform scientific computing and calculations. **Scikit-learn** code library offers you an intuitive interface and helps you apply machine learning algorithms to your data without any complexities.
After data modelling is over, you would need to visualize and interpret data for actionable insights.

Why Python is the best programming language for data science

Data visualization & interpretation

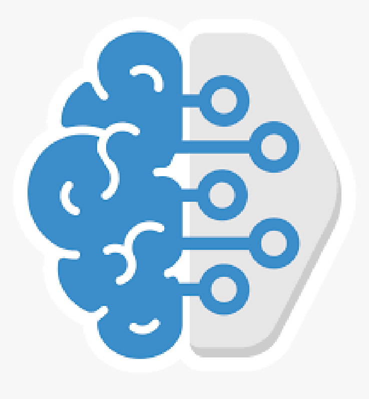
Python has many data visualization packages. **Matplotlib** is the most used library among them for generating basic graphs and charts. In case you need beautifully designed advanced graphs, you could also try another Python library, **Plotly**.

Another Python library, **IPython**, helps you with interactive data visualization and supports the use of a GUI toolkit. If you want to embed your findings into interactive web pages, **nbconvert** function can help you convert your IPython or Jupyter notebooks into rich HTML snippets.

After data visualization, the presentation of your data is of utmost importance, and it must be done in such a manner that the findings are driven by your business questions that you have asked at the beginning of your project.

Now that you deliver the answer to the business questions along with actionable insights, try to keep in mind that your interpretations appear useful to the stakeholders of your organization.

Why Python is the best programming language for data science



IS PYTHON 'THE' TOOL FOR MACHINE LEARNING?

When it comes to data science, machine learning is one of the significant elements used to maximize value from data. With Python as the data science tool, exploring the basics of machine learning becomes easy and effective. In a nutshell, machine learning is more about statistics, mathematical optimization, and probability. It has become the most preferred machine learning tool in the way it allows aspirants to 'do math' easily.

Name any math function, and you have a Python package meeting the requirement. There is Numpy for numerical linear algebra, CVXOPT for convex optimization, Scipy for general scientific computing, SymPy for symbolic algebra, PYMC3, and Statsmodel for statistical modeling.

With the grip on the basics of machine learning algorithm including logistic regression and linear regression, it makes it easy to implement machine learning systems for predictions by way of its scikit-learn library. It's easy to customize for neutral networks and deep learning with libraries including Keras, Theano, and TensorFlow.

Data science landscape is changing rapidly, and tools used for extracting value from data science have also grown in numbers. The two most popular languages that fight for the top spot are R and Python. Both are revered by enthusiasts, and both come with their strengths and weaknesses. But with the tech giants like Google showing the way to use Python and with the learning curve made short and easy, it inches ahead to become the most popular language in the data science world.