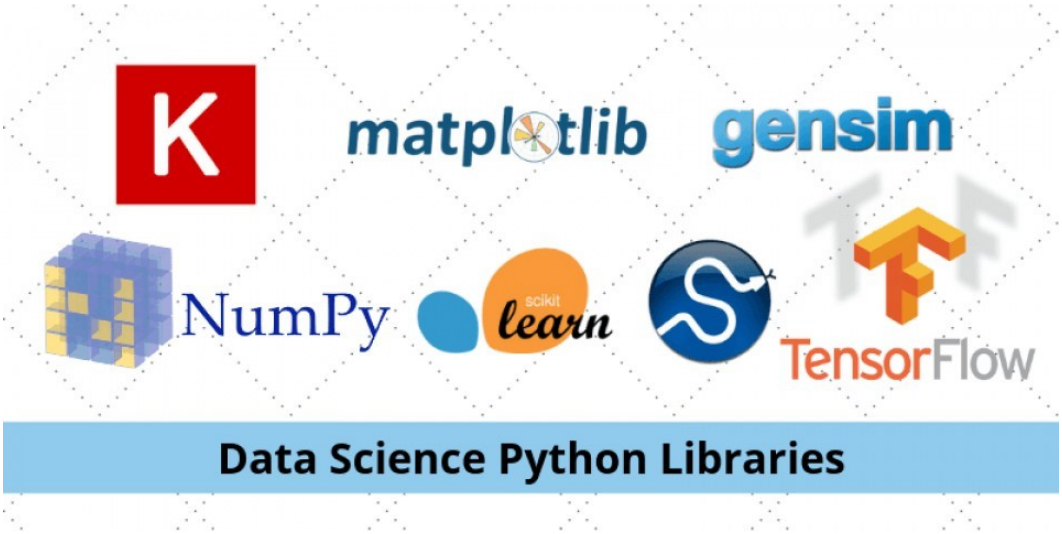
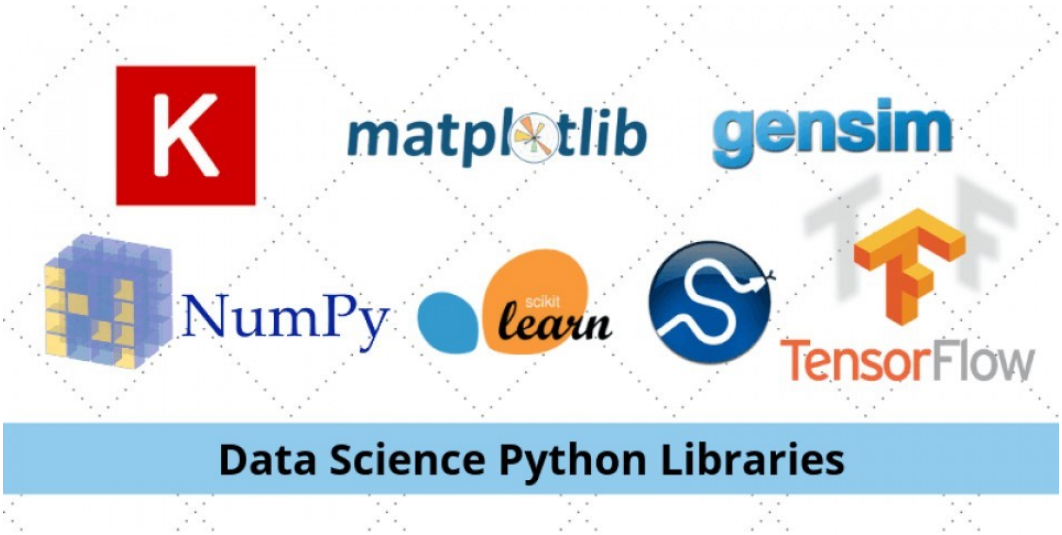



# Top 10 Data Science Python Libraries

hackr.io

8 mins read




## Table of Contents

Learning Python for building a professional career necessitates for learning Python libraries. A Python library is a collection of functions and methods that helps in completing specific tasks. Also, it saves developers a significant amount of time and headache.

Different Python libraries are intended for different uses. For instance, the Python imaging library, or PIL, is meant for image manipulation. On the other hand, we have the TensorFlow library which is important for developing and training deep learning models using Python.

Owing to the wider extent of Python libraries, we'll divide this article into two parts. This one will cover the top 10 data science Python libraries, while the second one explains the top 10 general purpose Python libraries.

## Top 10 Data Science Python Libraries – Backed by Python Developers Survey

There are several data science Python libraries available as of now. While some of them are already popular, others are improving inch-by-inch to reach the acceptance levels of their peers.

We are backing our list of the top 10 Python libraries with the Python Developers Survey 2018. It was a co-op effort put together by the Python Software Foundation along with JetBrains.

Over 20k developers from more than 150 countries participated in the Python Developers Survey 2018 to yield its conclusions. So, here are the top 10 Python data science libraries:

### NumPy

**Primary Intent:** Machine learning
**Secondary Intent(s):** Expressing images, other binary raw streams, and sound waves as an array of real numbers in N-dimensional array

NumPy is a fundamental Python library meant for scientific computing. It comes with support for a powerful N-dimensional array object and broadcasting functions.

Also, NumPy offers Fourier transforms, random number capabilities, and tools for integrating C/C++ and Fortran code. Having a working knowledge of NumPy is mandatory for full stack developers involved in machine learning projects using Python.

### People Also Read: Numpy Matrix Multiplication

A fun fact is that TensorFlow and several other machine learning Python libraries make use of NumPy internally for carrying out multiple operations on Tensors. The best and most important feature of NumPy is the array interface.

### Highlights:

- Easy to use and interactive
- Open-source contribution and ample community support
- Simplifies the process of complex mathematical implementations

### Pandas

**Primary Intent:** Accomplishing practical, real-world data analysis in Python
**Secondary Intent(s):** None

Employing the Pandas library makes it easier and intuitive for developers to work with labeled or relational data. It offers expressive, fast, and flexible data structures. Pandas aim to serve as the fundamental high-level building block for carrying out real-world data analysis using Python.

One of the most powerful features of Pandas is to translate complex data operations using mere one or two commands. Additionally, the machine learning library has no scarcity of built-in methods for combining, filtering, and grouping data. It also features time-series functionality.

### Highlights:

- Ability to perform custom types of operations
- Ensures that the entire process of data manipulation is easier
- Offers high flexibility and functionality when used with other Python libraries and tools
- Outstanding speed indicators
- Selects the best-suited output for the apply method
- Supports aggregations, concatenations, iteration, re-indexing, and visualizations operations

### Matplotlib

**Primary Intent:** 2D plotting
**Secondary Intent(s):** Production of publication-quality figures in numerous hardcopy formats

Matplotlib is a two-dimensional plotting library for the Python programming language. It is capable of producing high-quality figures in different hardcopy formats and interactive cross-platform environments.

Aside from being used in Python shell, Python scripts, and IPython shell, Matplotlib can also be used in:

- Jupyter Notebook
- Web application servers
- GUI toolkits; GTK+, Tkinter, Qt, and wxPython

According to the official website of Matplotlib, the Python library tries to “make easy things easy and hard things possible.” The 2D plotting Python library allows generating bar charts, error charts, histograms, plots, scatterplots, etc. with fewer lines of code.

### Highlights:

- Full control of axes properties, font properties, line styles, etc. via an object-oriented interface
- Legend for scatter
- Provides a MATLAB-like interface for simple plotting
- Secondary x/y axis support
- Works great with several graphics backends and operating systems

### SciPy

**Primary Intent:** Machine learning, scientific programming
**Secondary Intent(s):** Solving mathematical functions

The SciPy Python library comes with a number of modules for integration, linear algebra, optimization, and statistics. The open-source Python library allows developers and engineers to work with Fourier transforms, ODE solvers, signal and image processing, et cetera.

NumPy arrays are used as the basic data structure by SciPy. All functions offered by the various SciPy submodules are well documented. Hence, it is easy to get started with the machine learning library.

### **Highlights**:

- Easily handles mathematical operations
- Offers efficient numerical routines, such as numerical integration and optimization, using submodules
- Supports signal processing

### Scikit-Learn

**Primary Intent:** Data analysis and data mining
**Secondary Intent(s):** None

Considered to be one of the best Python libraries for working with complex data, Scikit-Learn is built on top of the Matplotlib, NumPy, and SciPy libraries. The machine learning Python library features a range of simple-yet-efficient tools for accomplishing data analysis and mining tasks.

Scikit-Learn is one of the most rapidly developing Python libraries. Several training methods, such as logistics regression and nearest neighbors, have received several smaller improvements over releases.

Another important modification made to the newest versions of the Scikit-Learn is the cross-validation feature, which allows for using more than a single metric.

Scikit-Learn features a number of algorithms for implementing data mining and machine learning tasks, notably classification, clustering, model selection, reducing dimensionality, and regression.

### Highlights:

- Ability to extract features from images and text
- Reusable in several contexts
- Several methods for checking the accuracy of supervised models on unseen data
- Wide range of algorithms, including clustering, factor analysis, principal component analysis to unsupervised neural networks

## TensorFlow


**Primary Intent:** Developing, training, and designing deep learning models
**Secondary Intent(s):** Performing numerical computation

Anybody involved in machine learning projects using Python must have, at least, heard of TensorFlow. Developed by Google, it is an open-source symbolic math library for numerical computation using data flow graphs.

The mathematical operations in a typical TensorFlow data flow graph are represented by the graph nodes. The graph edges, on the other hand, represent the multidimensional data arrays, a.k.a. tensors, that flow between the graph nodes.

TensorFlow flaunts a flexible architecture. It allows Python developers to deploy computation to one or many CPUs or GPUs in a desktop, mobile device, or server without the need of rewriting code. All libraries created in TensorFlow are written in C and C++.

Widely used Google products like Google Photos and Google Voice Search are built using TensorFlow. The library has a complicated front-end for Python. The Python code will get compiled and then executed on TensorFlow distributed execution engine.

### Highlights:

- Allows training multiple neural networks and multiple GPUs, making models very efficient for large-scale systems
- Easily trainable on CPU and GPU for distributed computing
- Flexibility in its operability, meaning TensorFlow offers the option of taking out the parts that you want and leaving that you don’t

- Great level of community and developer support
- Unlike other data science Python libraries, TensorFlow simplifies the process of visualizing each and every part of the graph

### Keras

**Primary Intent:** Developing and training deep learning models, deep learning research
**Secondary Intent(s):** Working with image and text data

Considered to be one of the coolest machine learning Python libraries, Keras offers an easier mechanism for expressing neural networks. It also features great utilities for compiling models, processing datasets, visualizing graphs, and much more.

Written in Python, Keras has the ability to run on top of CNTK, TensorFlow, and Theano. The Python machine learning library is developed with a primary focus on allowing fast experimentation. All Keras models are portable.

Compared to other Python machine learning libraries, Keras is slow. This is due to the fact that it creates a computational graph using the backend infrastructure first and then uses the same to perform

operations. Keras is very expressive and flexible for doing innovative research.

### Highlights:

- Being completely Python-based makes it easier to debug and explore
- Modular in nature
- Neural network models can be combined for developing more complex models
- Runs smoothly on both CPU and GPU
- Supports almost all models of a neural network, including convolutional, embedding, fully connected, pooling, and recurrent

## Seaborn


**Primary Intent:** Data visualization, making statistical graphics in Python
**Secondary Intent(s):** None

Basically a data visualization library for Python, Seaborn is built on top of the Matplotlib library. Also, it is closely integrated with Pandas data structures. The Python data visualization library offers a high-level interface for drawing attractive as well as informative statistical graphs.

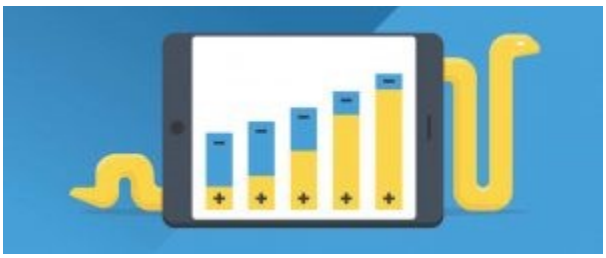
The main aim of Seaborn is to make visualization a vital part of exploring and understanding data. Its dataset-oriented plotting functions operate on arrays and data-frames containing whole datasets. The library is ideal for examining relationships among multiple variables.

Seaborn internally performs all the important semantic mapping and statistical aggregation for producing informative plots. The Python data visualization library also has tools for choosing among color palettes that aid in revealing patterns in a dataset.

### Highlights:

- Automatic estimation as well as the plotting of linear regression models
- Comfortable views of the overall structure of complex datasets
- Eases building complex visualizations using high-level abstractions for structuring multi-plot grids
- Options for visualizing bivariate or univariate distributions
- Specialized support for using categorical variables

### NLTK



**Primary Intent:** Natural language processing
**Secondary Intent(s):** Empirical linguistics, information retrieval, machine learning

A contraction for Natural Language Toolkit, NLTK is a suite of libraries for accomplishing symbolic and statistical NLP in Python. The Python library includes graphical demonstrations as well as

sample data. NLTK is accompanied by a book and a cookbook to make it easier to get started with.

In addition to being used as a platform for prototyping and building research systems, NLTK has widely been adopted for serving as a teaching tool and as an individual study tool.

NLTK provides support for classification, parsing, semantic reasoning, stemming, tagging, and tokenization functionalities.

### Highlights:

- Comes with a part-of-speech tagger
- n-gram and collocations
- Named-entity recognition
- Supports lexical analysis

## Gensim

**Primary Intent:** Natural language processing, unsupervised topic modeling, document indexing

**Secondary Intent(s):** Information retrieval

Using modern statistical machine learning, Gensim can be used for accomplishing natural language processing and unsupervised topic modeling tasks. In addition to Python, the NLP library can be implemented in Cython for enhancing performance and scalability.

Gensim is specially developed for handling large text collections, or corpora, by means of the data streaming and incremental online algorithms. The most distinguishing feature of Gensim is that unlike its contemporaries, it doesn’t target only in-memory processing.

### Highlights:

- Streamed parallelized implementation of doc2vec, fastText, and word2vec algorithms

- Supports latent Dirichlet allocation, latent semantic analysis, non-negative matrix factorization, random projections, and tf-idf

## All Caught Up!

That sums up the list of the top 10 data science Python libraries. With the rise of data science and machine learning, regular advancements are made to Python data science libraries. Also, newer Python machine learning libraries are being developed.

Accomplishing smaller data science projects might require using a single Python data science library. However, full-fledged data science, and research, projects demand a working knowledge of a number of Python data science libraries. So the more you learn, the better it is!

Want to learn data science better? Or looking to digest new data science concepts? Check out these best data science tutorials today!

## People are also Reading: