Python Libraries

We'll cover the following

As we learned in the previous lesson, one of the greatest assets of Python is its extensive set of libraries. These are what make the life of a data scientist easy — the start of the love affair between Python and data scientists!

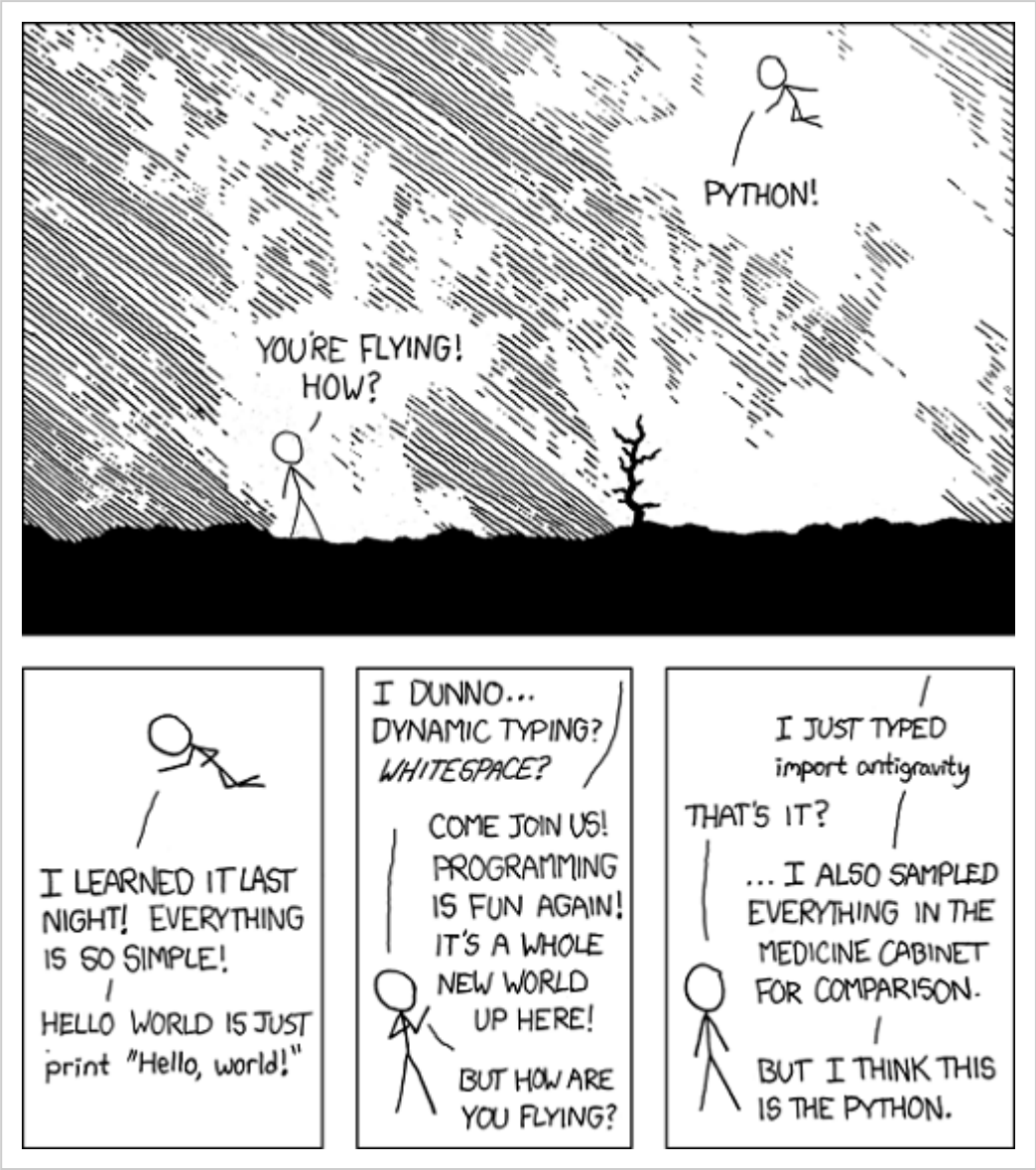


Image Cedits: https://xkcd.com

Essential Python Libraries for a Data Scientist

Let's take a quick tour of the libraries that a data scientist *should* really know.

Numpy

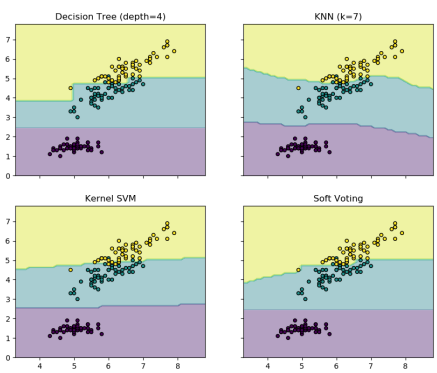
- NumPy (Numerical Python) is a powerful, and extensively used, library for storage and calculations. It is designed for dealing with numerical data. It allows data storage and calculations by providing data structures, algorithms, and other useful utilities. For example, this library contains basic linear algebra functions, Fourier transforms, and advanced random number capabilities. It can also be used to load data to Python and export from it.

Pandas

- Pandas is a library that you can't avoid when working with Python on a data science project. It is a powerful tool for data wrangling, a process required to prepare your data so that it can actually be consumed for analysis and model building. Pandas contains a large variety of functions for data import, export, indexing, and data manipulation. It also provides handy data structures like *DataFrames* (series of columns and rows, and *Series* (1-dimensional arrays), and efficient methods for handling them. For example, it allows us to reshape, merge, split, and aggregate data.

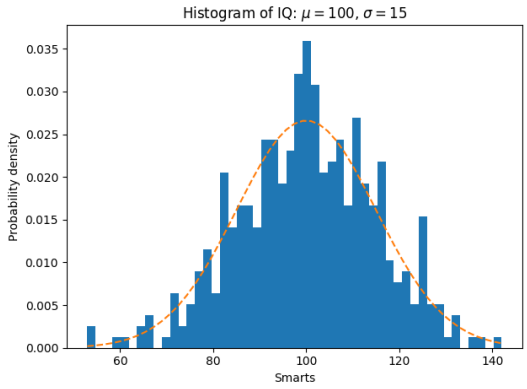
Scikit Learn

- Scikit Learn is an easy to use library for Machine Learning. It comes with a variety of efficient tools for machine learning and statistical modeling: it provides classification models (e.g., Support Vector Machines, Random Forests, Decision Trees), Regression Analysis (e.g., Linear Regression, Ridge Regression, Logistic Regression), Clustering methods (e.g, k-means), data reduction methods (e.g., Principal Component Analysis, feature selection), model tuning, and selection with features like grid search, cross-validation. It also allows for pre-processing of data. If these terms sound foreign to you right now, don't worry, we will get back to all this in detail in the section on machine learning.



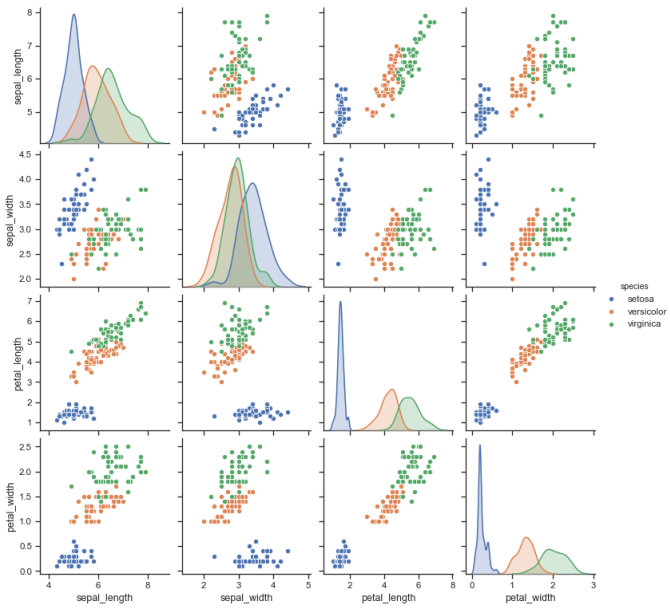
Matplotlib

- Matplotlib is widely used for data visualization like for plotting histograms, line plots, and heat plots.



Seaborn

- Seaborn is another great library for creating attractive and information rich graphics. Its goal is to make data exploration and understanding easier, and it does it very well. Seaborn is based on Matplotlib which is its child, basically.



> 📌 Note: ***Learning to use Python well means using a lot of libraries and functions which can be intimidating. But no need to panic — you don't have to remember them all by heart! Learning how to Google these things efficiently is among the top skills of a good data scientist!***