# NumPy, SciPy, Pandas, Matplotlib – How to get into the Data Science?

Once you make a decision, the universe conspires to make it happen

Wikipedia says, “Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from many structural, and unstructured data.” It is related to data mining, machine learning and big data. It is all about problem-solving, exploration, and extracting valuable information from data. To do so effectively, you’ll need to wrangle datasets, train machine learning models, visualize results, and much more.

Data science hype is all over the world now. It has been around for quite a while but its relevance is still as high as ever. Why should you enter the field of Data Science “right now”? Well, there are several reasons. My favorite reason for entering the field of Data science is because it is very interesting. You can see things from a different angle and predict the future and take crucial decisions. Being a data scientist practically means you can solve really complex problems and break them down into very simple bits so everyone around you can understand and find ways around. This can be done when one keeps an open mind and reaches out to the creative person within. Besides this, Data science is a rapidly growing field because of the powerful impact it brings to companies. The demand for data scientists has skyrocketed over the last few years but the supply for good and quality Data scientists is increasing at a much slower pace. A research study by professionals at IBM predicted that by 2020, the demand for data science jobs and all other jobs under it will soar by 28%.

After Knowing the impact of Data Science in the tech industries as well as the world, the very first question which bangs into everybody’s mind is that where should we start? Well, the answer is Python. Python is the best language for Data science for a good amount of reasons. Simplicity is one of Python's greatest strengths. Thanks to its precise and efficient syntax, Python can accomplish the same tasks with less code than other languages. This makes implementing solutions refreshingly fast. Besides, Python has an all-star lineup of libraries and frameworks for data analysis and machine learning, which drastically reduce the time it takes to produce results. And the good news is that Python, as a programming language, is so easy to learn that if you have a job, you can learn it in six months just by spending 1 hour a day in your pastime. And if you are a student, then why are you wasting your time in the social networks? Understanding Python is the most valuable skills needed for a Data Science career.

So, first, you need to learn the basics of Python. I have another blog about “How I have become a Self-Taught Python Developer in 3 years?” I hope if you read the blog you will have a good idea about where to start and where to end. People who have followed my Python learning timelines are highly benefitted.

After learning the basics of Python, you need to learn about some Python libraries and frameworks. As we mentioned earlier, Python has an all-star lineup of libraries for data science. Libraries are simply bundling of pre-existing functions and objects that you can import into your script to save time. The libraries you will need are -

- NumPy
- SciPy
- Pandas
- Matplotlib

NumPy and Pandas are great for exploring and playing with data. Matplotlib is a data visualization library that makes graphs as you'd find in Excel or Google Sheets. In this blog, I will tell you how to learn these libraries and from where you will find good learning resources.

### NumPy

NumPy is a library for adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. It allows easy and efficient numeric computation, and many other data science libraries are built on top of it. I recommend seeing the tutorial of Freecodecamp created by Keith Galli to grasp the basics of NumPy.

If you want to dive into more detailed learning on NumPy, the official documentation is the best choice for you. Also, check out my repository named "NumPy Journey" in GitHub. Here you will find the codes' scripts separated by categories and I have described all the codes in the comment lines.

### SciPy

SciPy is used for scientific computing and technical computing. It contains modules for optimization, linear algebra, integration, interpolation, special functions, FFT, signal and image processing, ODE solvers, and other tasks common in science and engineering. Truly speaking, I haven't found any good tutorials on YouTube to learn SciPy. But SciPy's official documentation and tutorials will allow you to learn the basics and beyond basics of SciPy.

### Pandas

Pandas is a software library written for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It's built on top of NumPy. This library created specifically to facilitate working with data, this is the bread and butter of a lot of Python data science work. Tutorials of Corey Schafer and Data School is undoubtedly the best choice for learning Pandas. Pandas' official documentation is also good for learning beyond the basics. I have created a GitHub repository, "Playing with Pandas" where I have described the basics of Pandas with in-code comments. So, don't forget to check that out. I also recommend learning how to solve real-world problems with Pandas and the playlist of Keith Galli shows how to solve real-world problems with details.

If anyone among you guys is super enthusiastic about learning all the ins and outs of Pandas, NumPy, and SciPy, then you can read the "Python for Data Analysis" book written by Wes McKinney original author of Pandas. This book is concerned with the nuts and bolts of manipulating, processing, cleaning, and crunching data in Python. It is also a practical, modern introduction to scientific computing in Python, tailored for data-intensive applications. This is a book about the parts of the Python language and libraries you'll need to effectively solve a broad set of data analysis problems.

Matplotlib

Matplotlib is a flexible plotting and visualization library. It provides an object-oriented API for embedding plots into applications. With it, we can quickly and easily generate charts from our data. There are several YouTubers who have created tutorial's playlists on Matplotlib. Among them, Corey Schafer's YouTube tutorial is great for beginners and it is very resourceful too. Besides this, Matplotlib's official site offers tutorials and they have categorized their tutorials by introductory, intermediate, advanced steps. I highly encourage you to go through their official documentation and also see my GitHub repository on "Matplotlib tutorials" where I have described how to create different types of plots with sufficient codes.

That's all for this blog post. I hope if you learn Python basics first and then these libraries you will be able to analyze and visualize the data. For getting into core Data Science, Artificial Intelligence, and Deep Learning you need to learn some additional Python libraries like Scikit Learn, Tensor flow, Keras, Pytorch, OpenCV, etc. That will be a talk for my upcoming blog post.