7 Most Recommended skills for Data Science

I want to share the **seven most recommended data science skills** from dozens of interactions and discussions with some of the largest data leaders in the world, including the Head of Data & Analytics @ Google, the Senior Director of Engineering @ NVIDIA, and the VP of Data Science and Engineering @ Wealthsimple.
While this article may be more anecdotal, I feel like this shares a valuable perspective. I'm specifically not referring to data from scraped job postings because, from my experiences, there seems to be quite a disconnect between job descriptions and what's actually done on the job.

You might notice that **none** of the seven skills have anything to do with machine learning or deep learning, and this is not a mistake. Currently, there is a much higher demand for skills that are used in the pre-modeling phases and post-modeling phases. And so, the seven most recommended skills to learn actually overlap with the skills of a data analyst, a software engineer, and a data engineer.
With that said, let's dive into the **seven most recommended data science skills to learn in 2021:**

1) SQL

SQL is **the** universal language in the world of data. Whether you're a data scientist, a data engineer, or a data analyst, you'll need to know SQL.
SQL is used to extract data from a database, manipulate data, and create data pipelines — essentially, it's important for almost every pre-analysis/pre-modeling stage in the data lifecycle.

Developing strong SQL skills will allow you to take your analyses, visualizations, and modeling to the next level because you will be able to extract and manipulate the data in advanced ways. Also, writing **efficient** and **scalable** queries is becoming more and more important for companies that work with petabytes of data.
Here are some of my favorite resources to learn SQL:

- *Mode's SQL tutorial for Data Analysis*
- *Codecademy — Learn SQL*
- *FreeCodeCamp — Full Database Course for Beginners*

2) Data Visualizations & Storytelling

If you think creating data visualizations and storytelling are specific to the role of a data analyst, think again.

Data visualizations simply refer to data that is presented visually — it can be in the form of graphs, but it can also be presented in unconventional ways.

Data storytelling takes data visualizations to the next level — **data storytelling** refers to "how" you communicate your insights. Think of it as a picture book. A good picture book has good visuals, but it also has an engaging and powerful narrative that connects the visuals.

Developing your data visualization and storytelling skills are essential because you're always selling your ideas and your models as a data scientist. And it's especially important when communicating with others who are not as technologically savvy.

Here are some of my favorite resources to learn data visualizations & storytelling:

- *Data Visualization using Matplotlib*
- *Data Visualizations using Plotly*
- *Google — Storytelling with data*

3) Python

From my interactions, Python seems to be the go-to programming language to learn over R. That doesn't mean that you can't be a data scientist if you use R, but it just means that you'll be working in a language that is different from what the majority of people use.

Learning Python syntax is easy, but you should be able to write efficient scripts and leverage the wide-range of libraries and packages that Python has to offer. Python programming is a building block for applications like manipulating data, building machine learning models, writing DAG files, and more.

Here are some of my favorite resources to learn Python:

- *FreeCodeCamp — Full Python Course for Beginners*
- *Leetcode*

4) Pandas

Arguably the most important library to know in Python is Pandas, a package for data manipulation and analysis. As a data scientist, you'll be using this package all the time, whether you're cleaning data, exploring data, or manipulating the data.

Pandas has become such a prevalent package, not only because of its functionality but also because DataFrames have become a standard data structure for machine learning models.

Here are some of my favorite resources to learn Pandas:

- *Kaggle — Learn Pandas Tutorial*
- *Guipsamora — Pandas Exercises*

5) Git/Version Control

Git is the main version control system used in the tech community.

If that doesn't make sense, consider this example. In high school or university, if you ever had to write an essay, you might have saved different versions of your essay as you progressed through it. For example:

🗁Final Essay
└🗁Essay_v1
└🗁Essay_v2
└🗁Essay_final
└🗁Essay_finalfinal
└🗁Essay_OFFICIALFINAL

All jokes aside, Git is a tool that serves the same purpose, except that it's a distributed system. This means that files (or repositories) are stored both locally and in a central server.

Git is extremely important for several reasons, with a few being that:

- It allows you to revert to older versions of code
- It allows you to work in parallel with several other data scientists and programmers
- It allows you to use the same codebase as others even if you're working on an entirely different project

Here are some of my favorite resources to learn Git:

- *Codecademy — Learn Git*
- *MIT — Version Control*
- *Learn Git Branching*

6) Docker

Docker is a containerization platform that allows you to deploy and run applications like machine learning models.

It's becoming increasingly important that data scientists not only know how to build models but how to deploy them as well. In fact, a lot of job postings now require some experience in model deployment.

The reason that it's so important to learn how to deploy models is that a model delivers no business value until it is actually integrated with the process/product that it is associated with.

Here are some of my favorite resources to learn Docker:

- *Docker for Beginners*
- *Docker For Beginners: From Docker Desktop to Deployment*
- *Deploying Docker Containers*
- *Deploy Machine Learning Pipeline on the cloud using Docker Container*

7) Airflow

Airflow is a workflow management tool that allows you to automate… well workflows. More specifically, Airflow allows you to create automated workflows for data pipelines and machine learning pipelines.

Airflow is powerful because it allows you productionalize tables that you may want to use for further analysis or modeling, and it's also a tool that you can use to deploy machine learning models.

Here are some of my favorite resources to learn Airflow:

- *Airflow tutorial 1: Introduction to Apache Airflow*
- *A Complete Introduction to Apache Airflow*
- *Tutorial — Airflow Documentation*

.