Top 10 Most in Demand Skills for Data Scientist

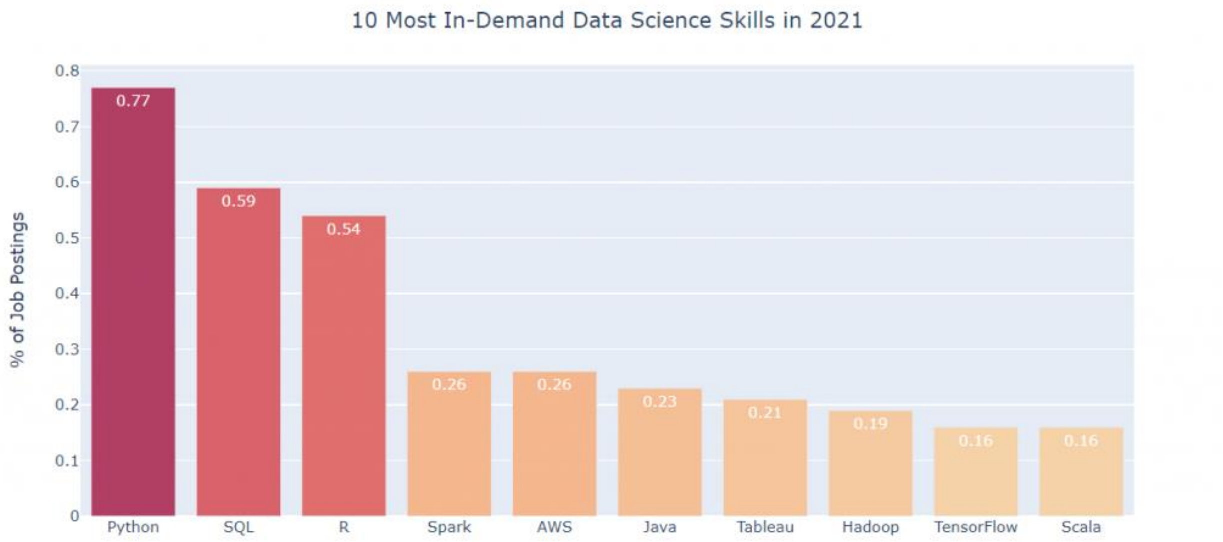

Image created by Author.

I just wanted to start off by saying that this is heavily inspired by Jeff Hale's articles that he wrote back in 2018/2019. I'm writing this simply because I wanted to get a more up-to-date analysis of what skills are in demand today, and I'm sharing this because I'm assuming that there are people out there that also want to see an updated version of the most in-demand skills for data scientists in 2021.

Take what you want from this analysis — it's obvious that the insights gathered from web scraping job postings do not offer a perfect correlation to what data science skills are actually most demanded. However, I think this gives a good indication of what general skills you should focus more on, and likewise, stray away from.

With that said, I hope you enjoy this, and let's dive into it!

Methodology

For this analysis, I webscraped and accumulated over 15,000 job postings from Indeed, Monster, and SimplyHired. I didn't webscrape LinkedIn because I ran into Captcha issues trying to scrape it.

I then checked to see how many job postings included each term that I was searching. The list of terms that I was searching was as follows (*if you want to see any other skills, please mention it in the comments so I can add it for next year's analysis!*):

- Python, SQL, R, Java, Git, C, MATLAB, Excel, C++, JavaScript, C#, Julia, Scala, SAS
- Scikit-learn, Pandas, NumPy, SciPy
- Matplotlib, Looker, Tableau
- TensorFlow, PyTorch, Keras

- Spark, Hadoop, AWS, GCP, Hive, Azure, Google Cloud, MongoDB, BigQuery
- Docker, Kubernetes, Airflow
- NoSQL, MySQL, PostgreSQL
- Caffe, Alteryx, Perl, Cassandra, Linux

After getting the counts from each source, I summed them up and then divided it over the total number of data scientist job postings to get a percentage. For example, Python's value of 0.77 means that 77% of the job postings had Python in it.

Finally, I compared the results to the analysis done by Jeff Hale in 2019 to get the percentage change from 2019 to 2021.

Results

Top Skills

Below are the top 25 most in-demand data science skills in 2021, ranked from highest to lowest:

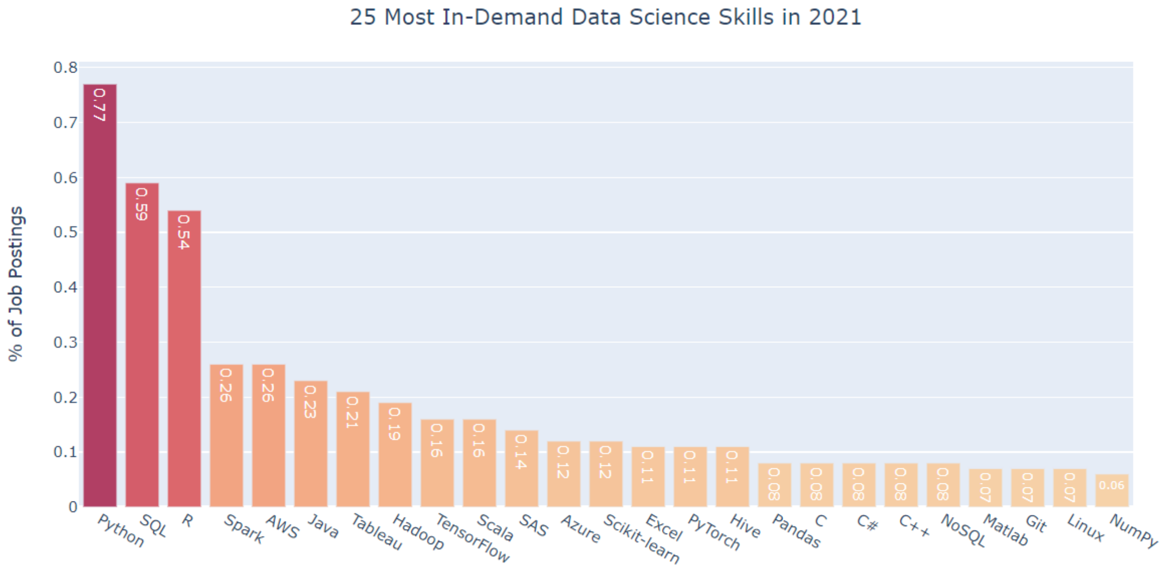

Image created by Author.

Top Programming Languages

To get a more granular look, the chart below shows the top programming languages for data scientists:

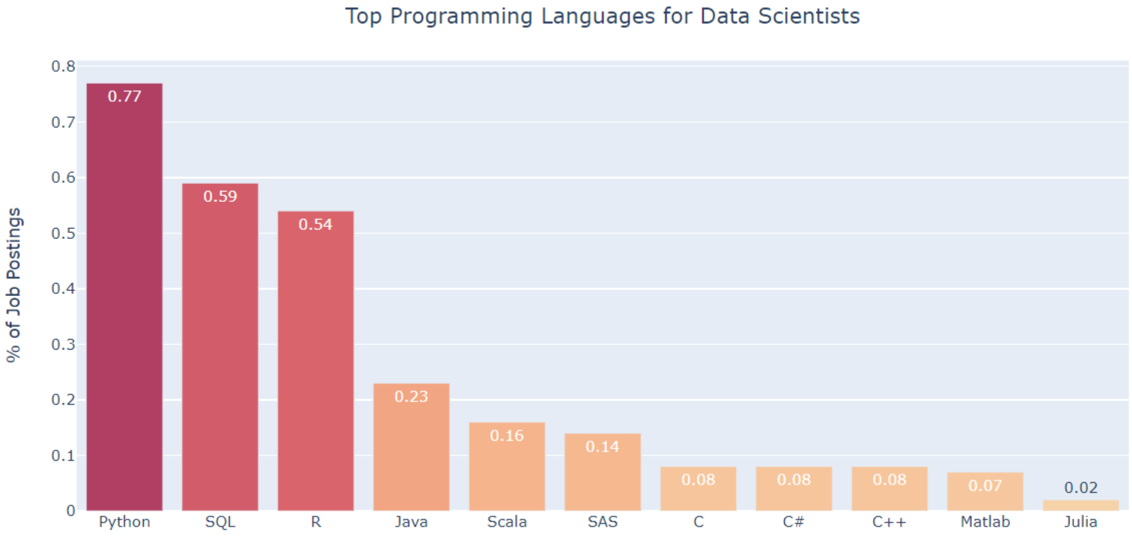

Image created by Author.

It’s no surprise that Python, SQL, and R are the top three programming languages.

Personally, I also stand by the fact that you should know either Python or R as well as SQL. I started with Python, and I’ll probably stick with Python for the rest of my life. It’s so far ahead in terms of open source contributions, and it’s straightforward to learn. SQL is arguably the most important skill to learn across any type of data-related profession, whether you’re a data scientist, data engineer, data analyst, business analyst, the list goes on.

Top Python Libraries

Similarly, the chart below shows the top Python Libraries for Data Scientists:

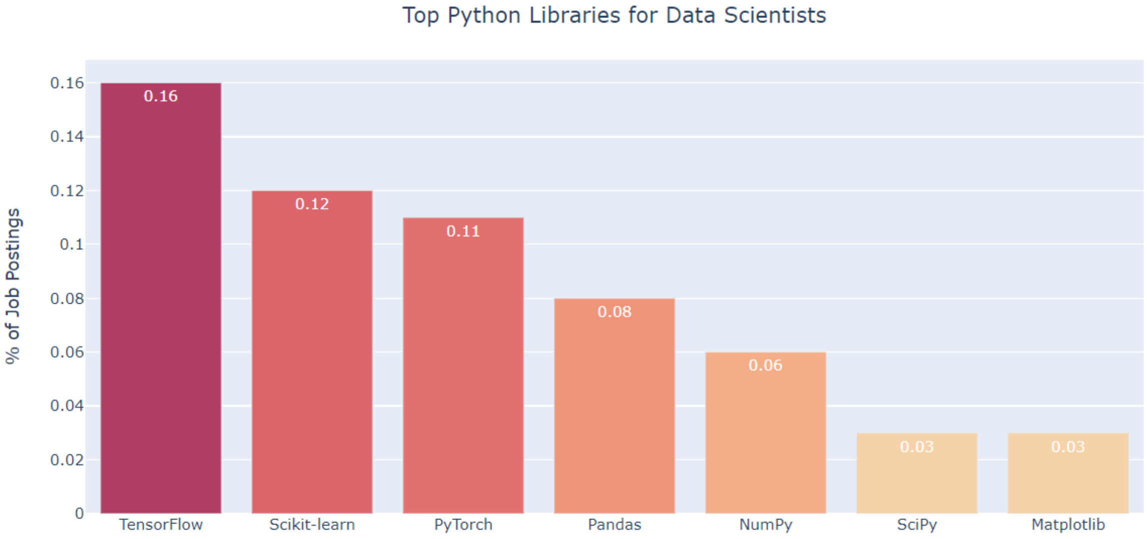

Image created by Author.

TensorFlow ranks first, as it is one of the most popular libraries of Python for deep learning. PyTorch is a strong alternative, hence its ranking not too far behind.

Scikit-learn is arguably the most important library in Python for machine learning. After cleaning and manipulating your data with Pandas and/or NumPy, scikit-learn is used to build machine learning models as it has tons of tools used for predictive modelling and analysis.

In my opinion, Pandas, NumPy, and SciPy are also essential for data scientists despite their representation above.

Fastest Growing and Declining Skills

The charts below show the fastest growing and declining skills from 2019 to 2021:

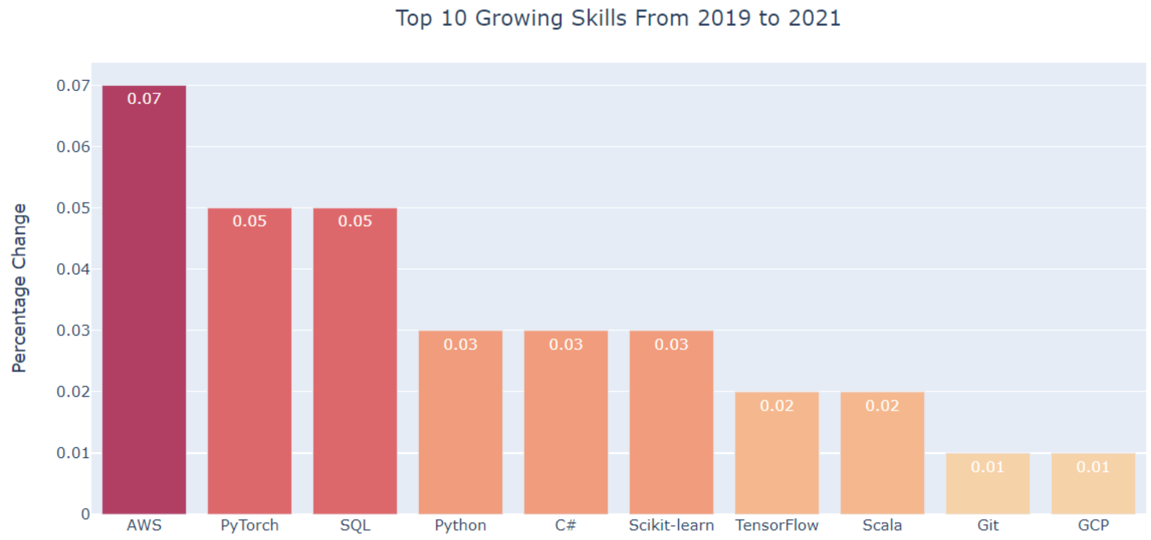

Image created by Author.

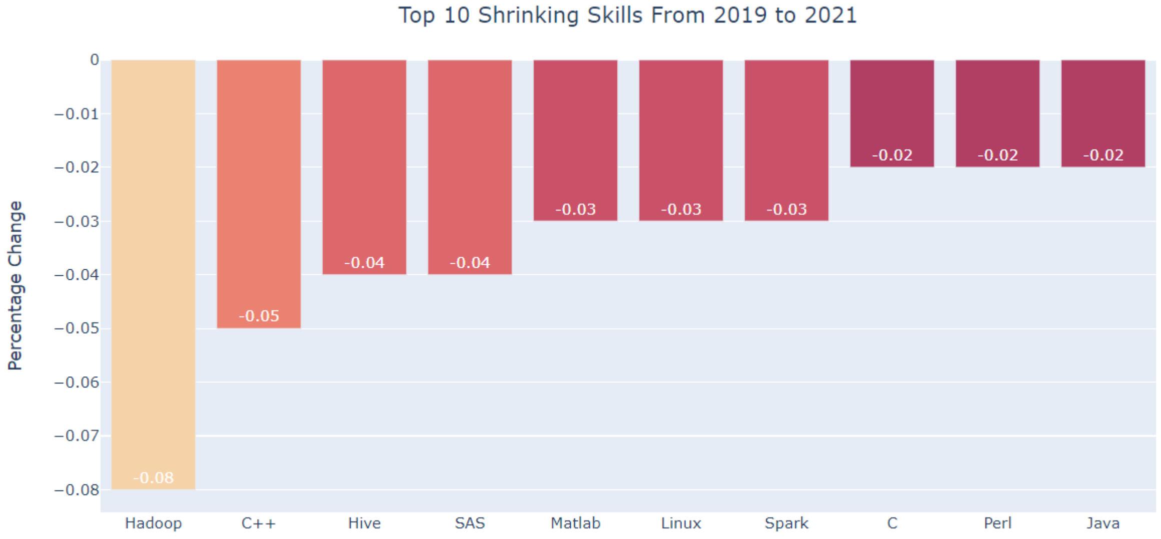

Image created by Author.

Here are a few takeaways from the two charts above:

1. There is a huge increase in skills related to the cloud, like AWS and GCP.

2. Similarly, there is also a large increase in skills related to deep learning, like PyTorch and TensorFlow.
3. SQL and Python continue to grow in importance, while R remains stagnant.
4. Apache products, like Hadoop, Hive, and Spark, continue to decline in importance.