



State of Machine Learning and Data Science 2021

Insights from Kaggle's annual user survey focused
on working data scientists.

October 14, 2021



Table of Contents

Overview & Report Methodology

01 | [Data Scientist Profile](#)

02 | [Education](#)

03 | [Data Science & Machine Learning Experience](#)

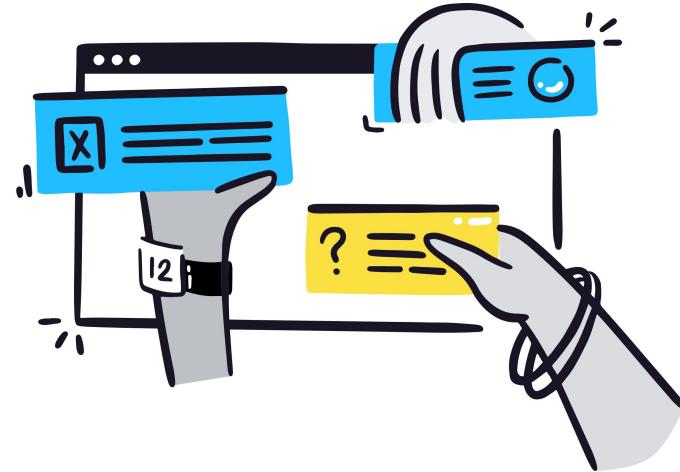
04 | [Employment](#)

05 | [Technology](#)

[Conclusion](#)



Overview & Methodology





This is our 5th year conducting an in-depth user survey & publicly sharing the results.

Over 25,000 data scientists and ML engineers submitted responses on their backgrounds and day to day experience – everything from educational details to salaries to preferred technologies and techniques.





This report is focused only on a slice of the data – the 14% of respondents who are currently employed with the job title of “data scientist”. It’s a follow-up analysis to [a report we published last year](#) with the same criteria.

We organized the report into five sections: 01) Data Scientist Profile, 02) Education, 03) Data Science & 04) Machine Learning Experience, 04) Employment, and 05) Technology.

Note: there are many other job titles that support data science and ML workflows and also many students and data enthusiasts who aren’t full-time, employed data scientists. You can find their responses in the complete [2021 survey dataset on Kaggle](#). We highly encourage conducting your own analysis and [sharing with the broader community](#) – we’d love to see them!





You can find a detailed summary of our survey methodology [here](#).

Many survey questions were multiple choice with the ability for respondents to select all options that applied to them. For that reason, you may see visualizations where the total percentage is more than 100%.

Also, all monetary amounts captured in the report are in USD.



01

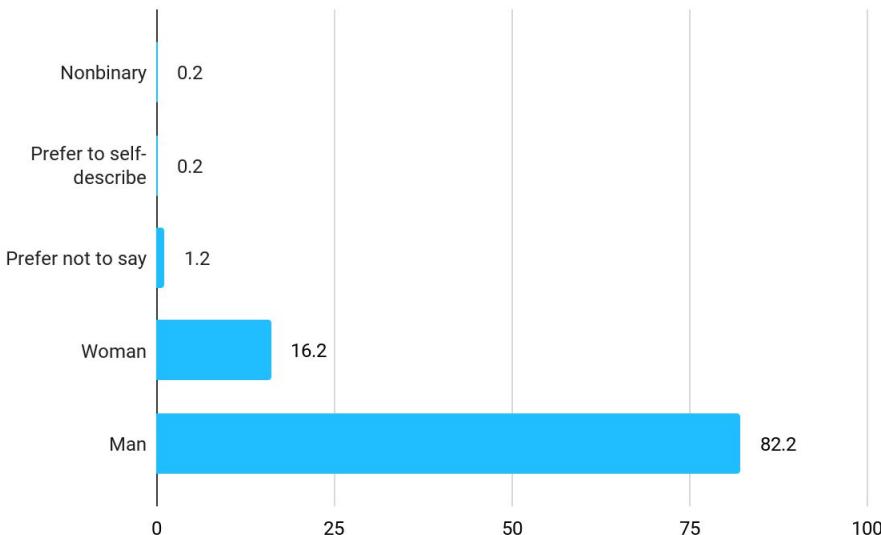
Data Scientist Profile



Gender

Data science is still suffering from a large gender gap in the workplace, as 82% of users identify as men.

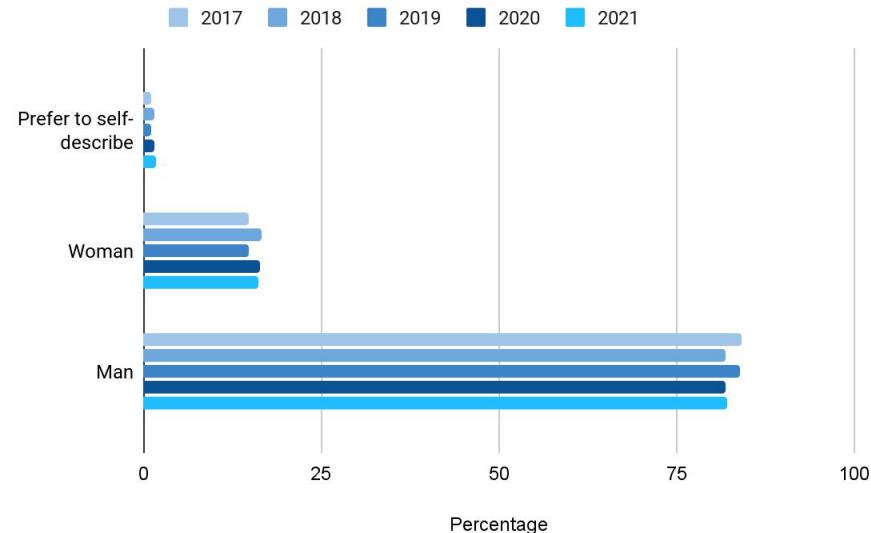
Gender Identity of Data Scientists



Gender (cont.)

Looking over the past five years, there has been no meaningful change in gender distribution.

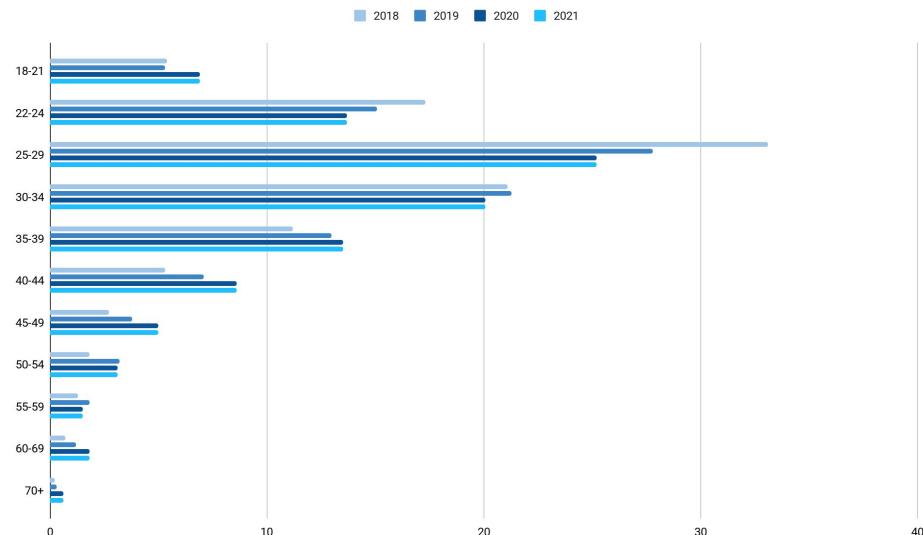
Gender Identity of Data Scientists



Age

Data science remains a fairly young profession, with more than half of all data scientists being between the ages of 22 and 34.

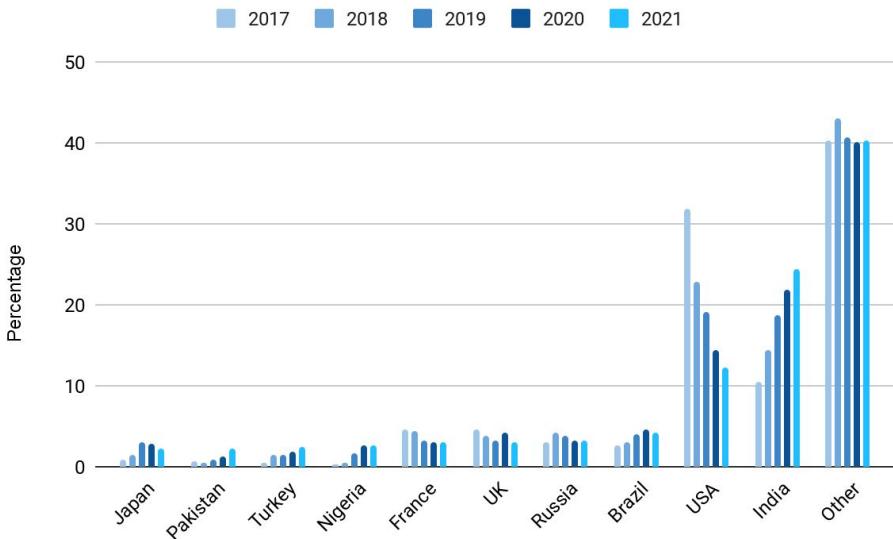
Age ranges of data scientists



Country

Data scientists live and work all around the globe, and more than 40% of survey respondents live outside of the 10 countries where we had the most respondents.

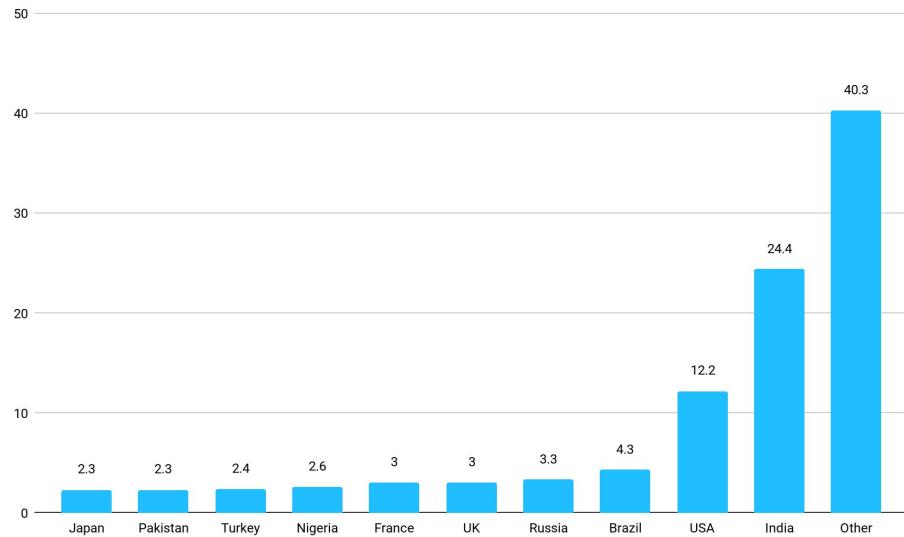
Most Common Nationalities



Country (cont.)

Country demographics are nearly the same as last year with two countries having far more representation in the Kaggle community. India makes up 24.4% of Kaggle data scientists, while 12.2% reside in the United States. Brazil is a distant third, at under 4.3%.

Most Common Nationalities



02

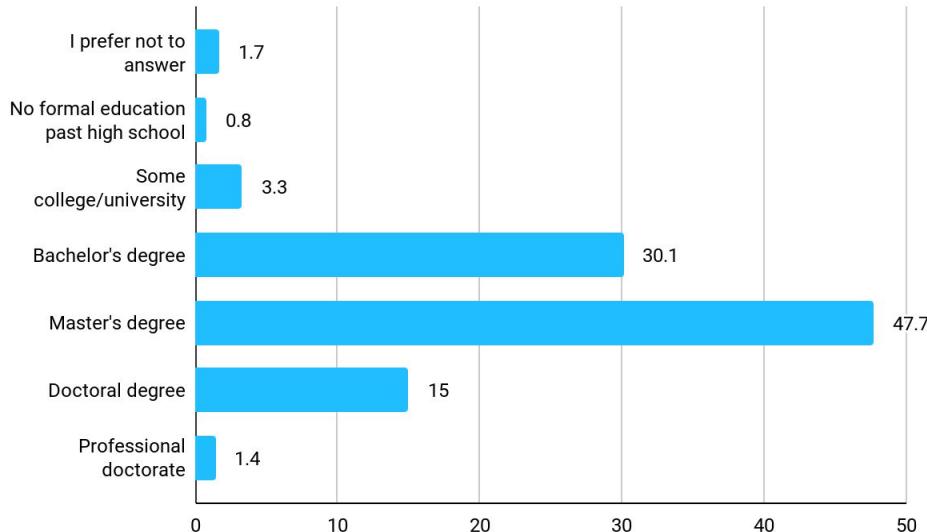
Education



Higher Education

Graduate degrees continue to be the norm for data scientists, with over 62% having obtained either a Master's or doctoral degree. Fewer than 5% of data scientists have no degree beyond a high school diploma.

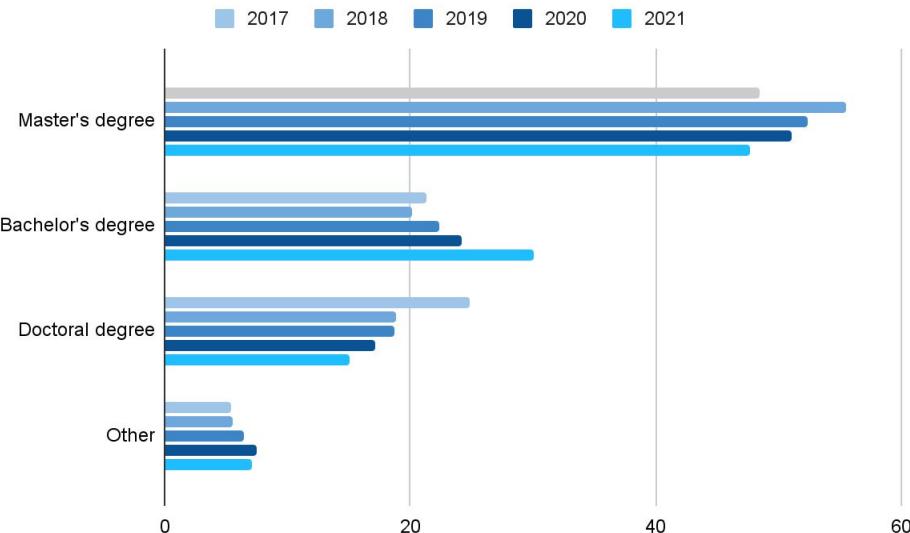
Education Levels of Kaggle Data Scientists



Higher Education

Looking year-over-year, it is becoming more common to be employed as a data scientist without having an advanced degree, although advanced degrees are still the norm (~64%).

Education Levels of Data Scientists Year over Year

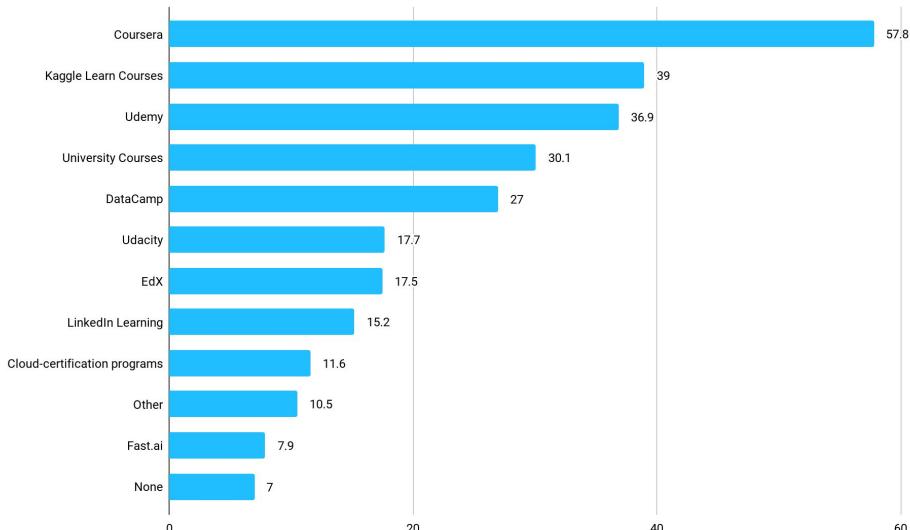


Ongoing Learning

Data science and machine learning techniques rapidly progress, so it's no surprise most of Kaggle data scientists maintain ongoing education.

Coursera remains the most popular ongoing data science learning resource.

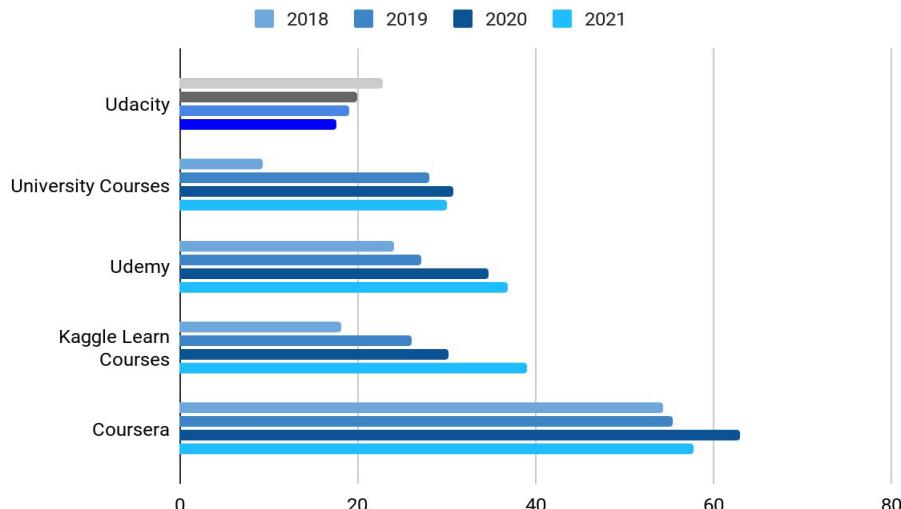
Popular Ongoing Learning Resources



Ongoing Learning (cont.)

Kaggle Learn Courses had the biggest popularity growth (9%) since last year.

Most Popular Learning Platforms Year over Year



03

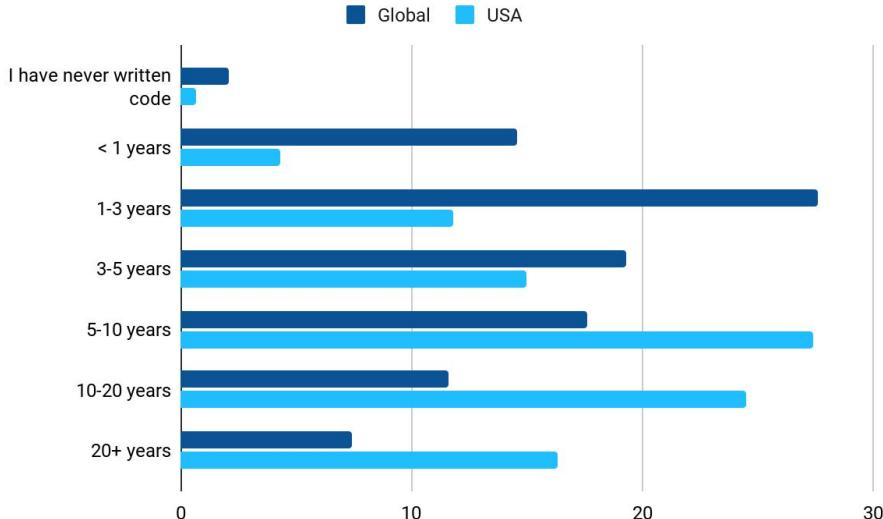
Data Science & Machine Learning Experience



Programming Experience

While most Kaggle data scientists have at least a few years of experience under their belt, a growing share have taken up programming within the last year (14.6% vs 9% in 2020).

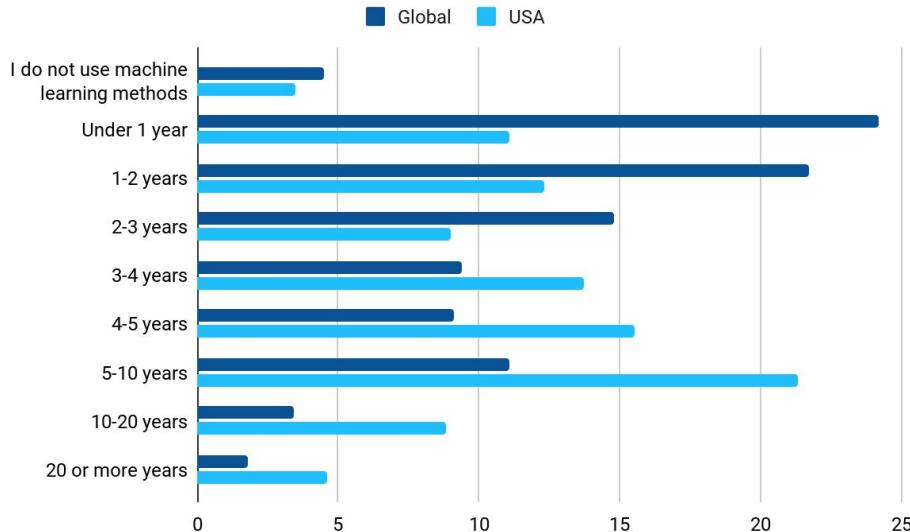
**Programming Experience for Data Scientists
(Global vs USA)**



Machine Learning Experience

Most Kaggle data scientists are newer to machine learning than programming. Slightly more than 55% of data scientists have less than three years experience. Less than 6% of professional data scientists have been using machine learning for a decade or more. As with programming, US data scientists have more machine learning experience than the global respondents.

Years of Machine Learning Experience



04

Employment

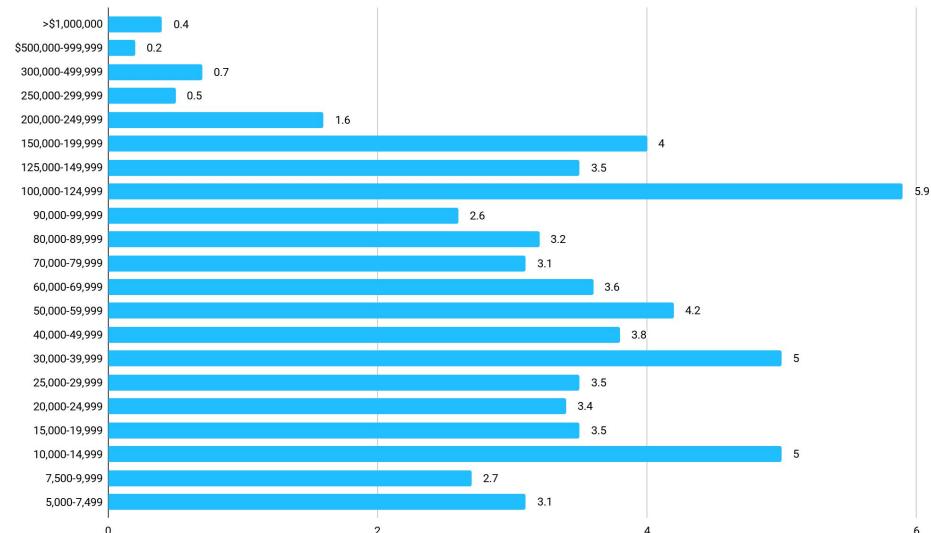


Pay

Companies in the United States are most likely to pay in the six figures, based on these survey results. Global companies have lower salary ranges that are more evenly distributed.

There are trends regionally, such as India, where nearly 90% make less than \$50,000 USD per year.

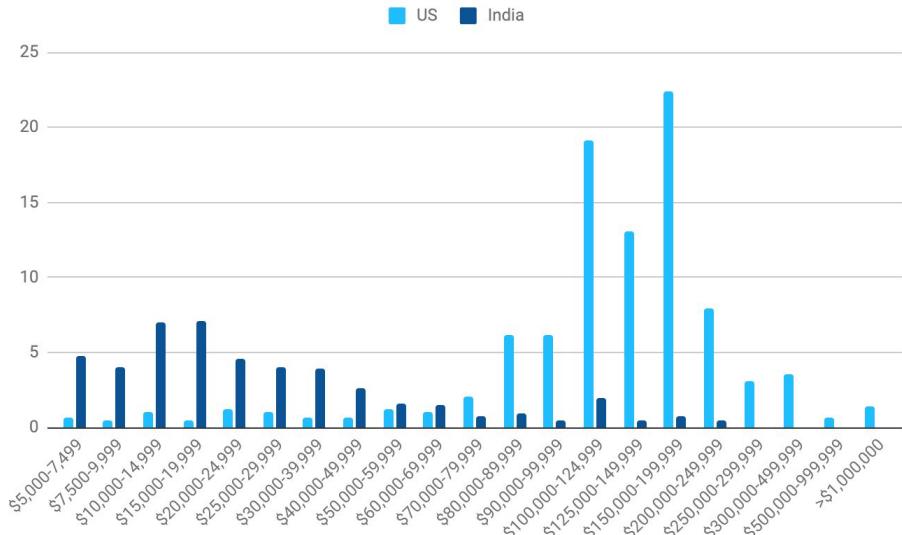
Global Salary Distribution



Pay (cont.)

Comparing salaries between our two largest countries, most US-based data scientists make over \$100,000 per year while less than 3% of India-based data scientists make over \$100,000 per year.

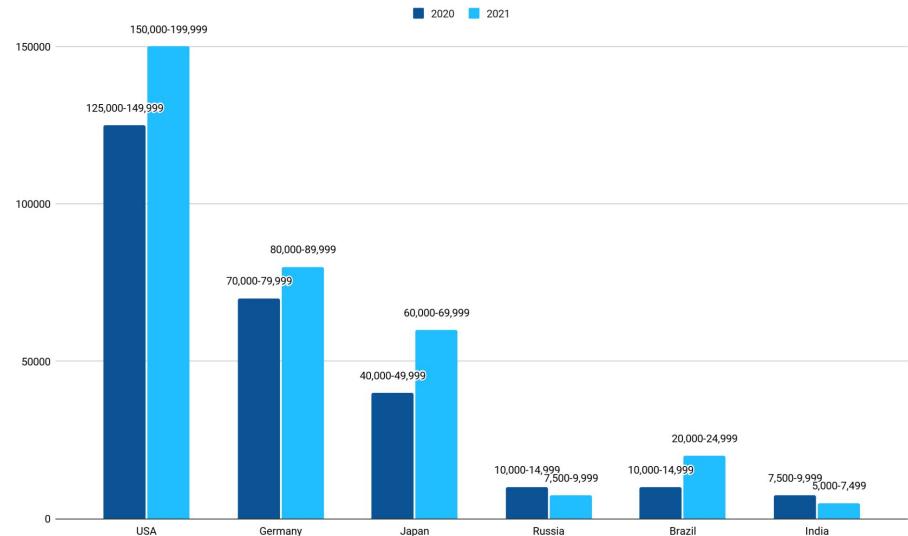
Salary Distribution US vs India



Pay (cont.)

Looking at the most common salaries by country, we see that US companies are more likely to pay higher salaries. Companies in Germany and Japan follow, with significantly higher salaries than the other included regions.

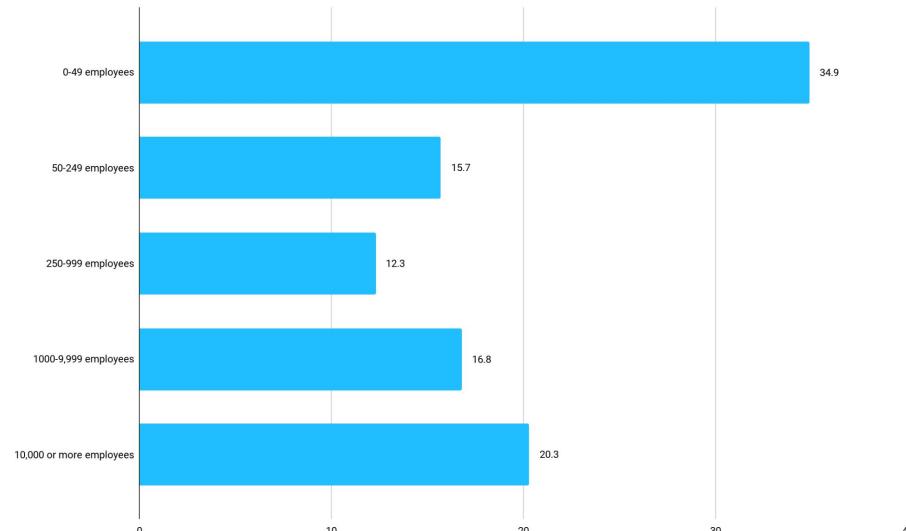
Median Salary by Country



Companies Employing Data Science

Like last year, large enterprises and small startups are the most common choices of data scientists in this survey. Over half of employers have less than 250 employees. Yet, one in five work at companies with over 10,000 employees.

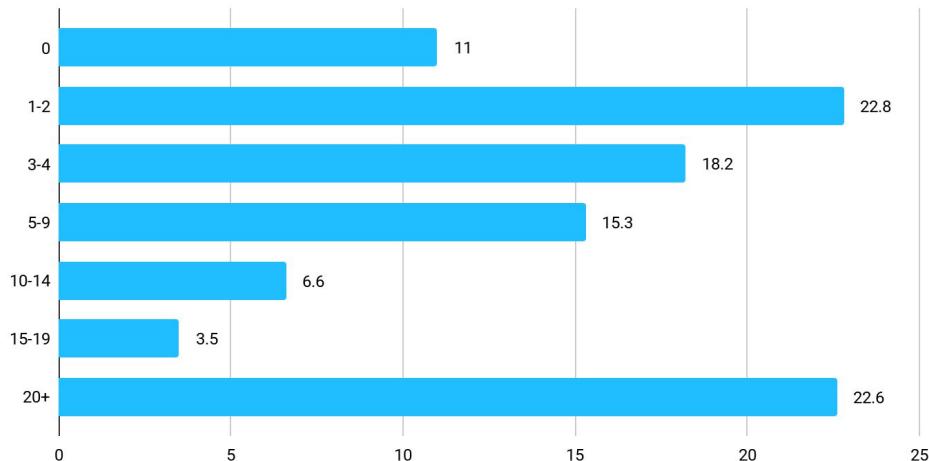
Company Size of Data Science Employers



Data Science Teams

The sizes of data science teams didn't meaningfully change from last year – over half of data scientists still work at companies with five or fewer people on the data science team, yet one in five work on a team with 20+ data scientists.

Data Science Team Size

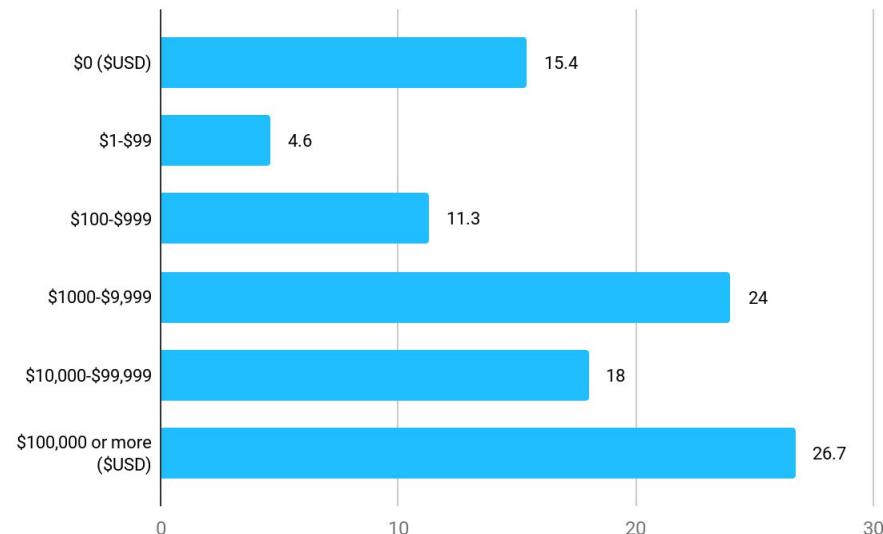


Spending

There's plenty of money being spent on machine learning and cloud computing products, but not by all data scientists.

There's quite a range, with over a quarter of data scientists claiming to have spent no money at all, while one in 10 has spent over \$100,000 USD in the last five years.

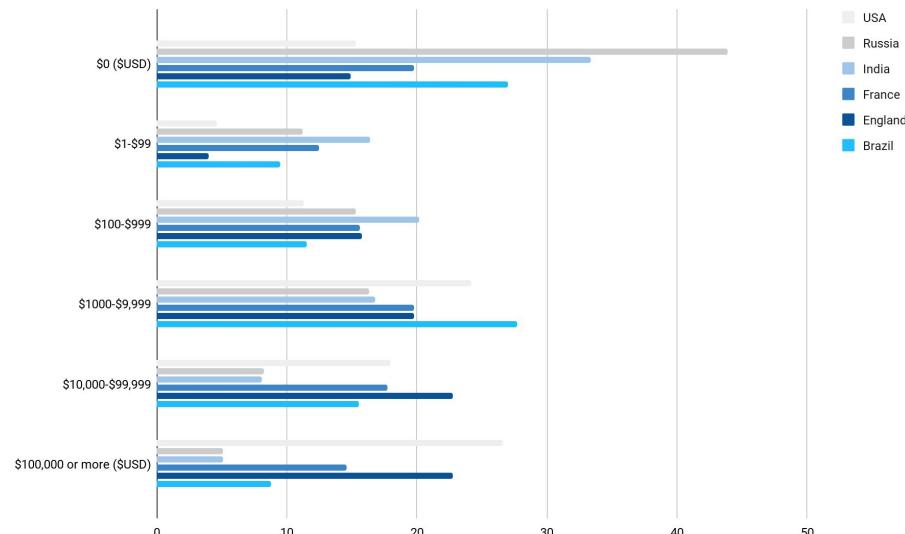
Enterprise Spending on Cloud Computing Products (Global)



Spending

Data scientists from the US spend more money in the cloud than their global counterparts. There are more than two times the responses for the highest spending level in the US compared to other countries.

Cloud Spending by Country



05

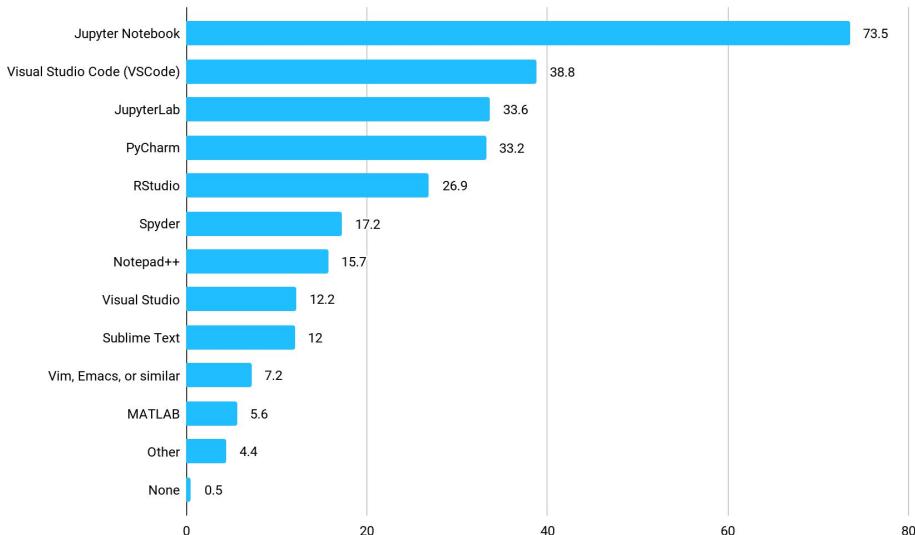
Technology



Interactive Development Environments

Jupyter-based IDEs continue to be the go-to tool for data scientists, with around three-quarters of Kaggle data scientists using it. However, Visual Studio Code is in the second spot with 38%.

IDE Popularity

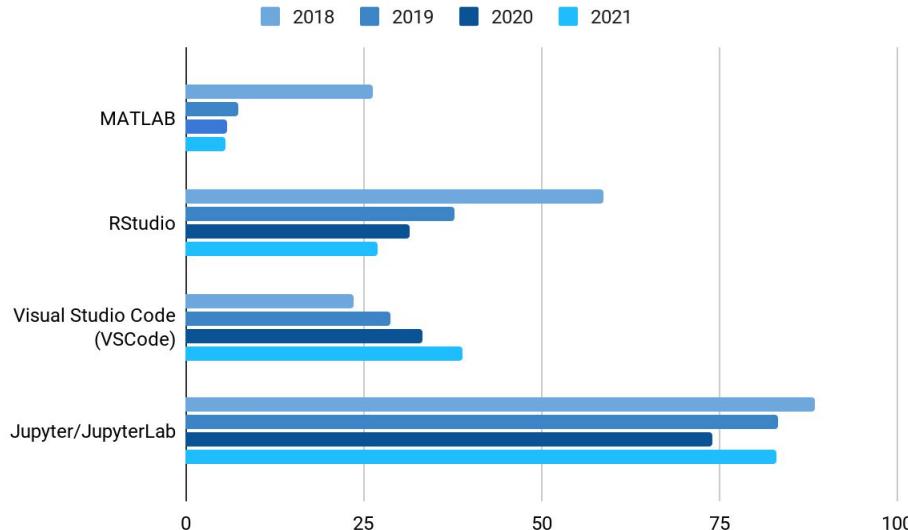


Interactive Development Environments (cont.)

Looking year over year, VSCode is continuing its popularity gain.

Note: In the previous figure Jupyter and JupyterLab were separate choices, whereas in this figure they were combined in order to be consistent with how the question was structured in 2019 (and to allow for comparison with 2018 where JupyterLab was not yet an option).

Top IDE Popularity Year Over Year

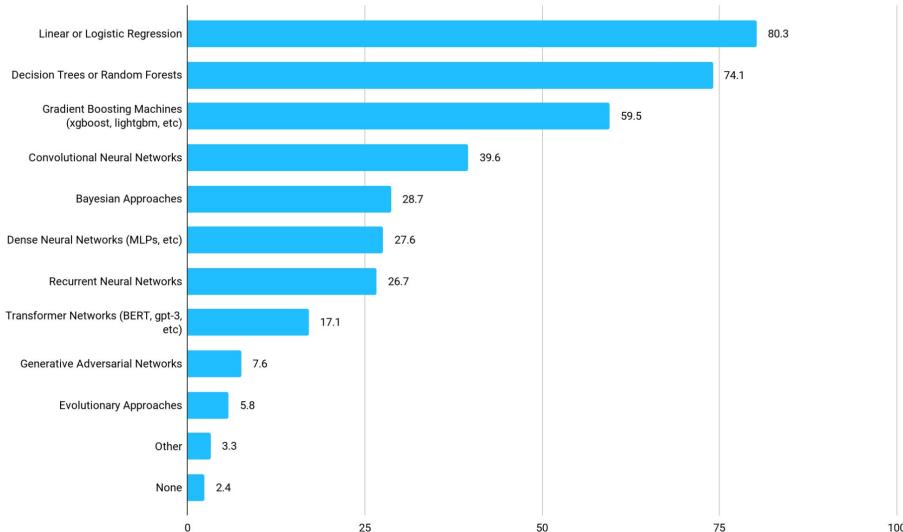


Methods & Algorithms

Like last year, the most commonly used algorithms were linear and logistic regression, followed closely by decision trees and random forests.

Of more complex methods, gradient boosting machines and convolutional neural networks were the most popular approaches.

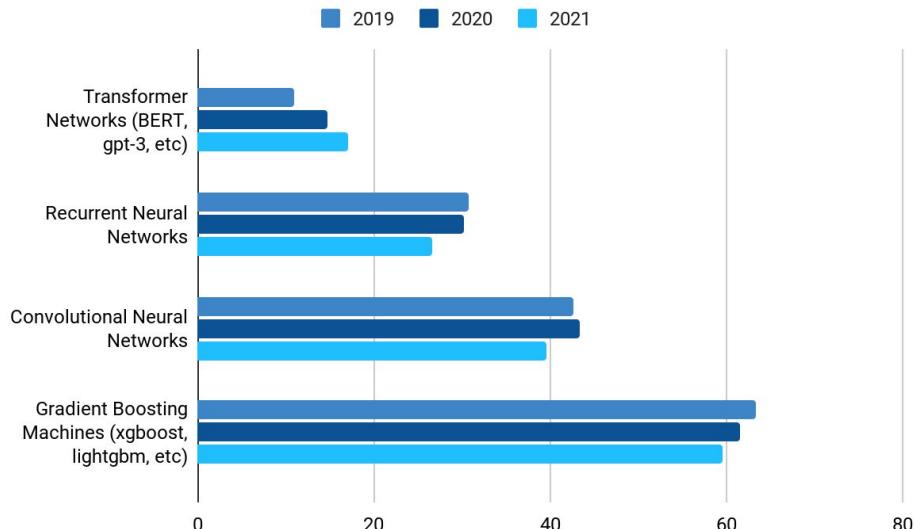
Methods and Algorithms Usage



Methods & Algorithms (cont.)

We also saw strong year-over-year growth in the use of large language models such as transformer networks (BERT, GPT-3, etc).

Popular ML Algorithms



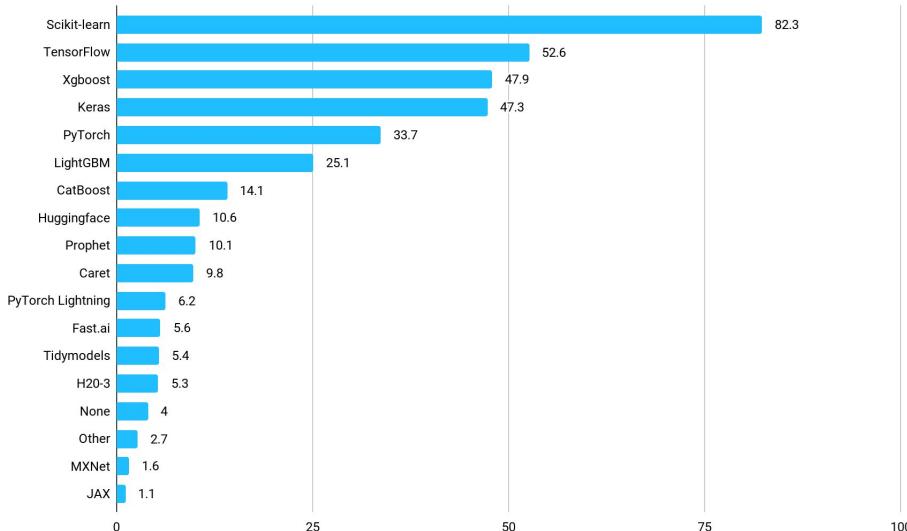
Machine Learning Frameworks

Python-based tools continue to dominate the machine learning frameworks.

Like last year, Scikit-learn, a swiss army knife applicable to most projects, is the top with over 80% of data scientists using it. TensorFlow and Keras, notably used in combination for deep learning, were each selected on about half of the data scientist surveys. Gradient boosting library xgboost is fourth, with about the same usage as 2020 and 2019.

The most popular of the new tools added to the survey this year is Huggingface reaching over 10%.

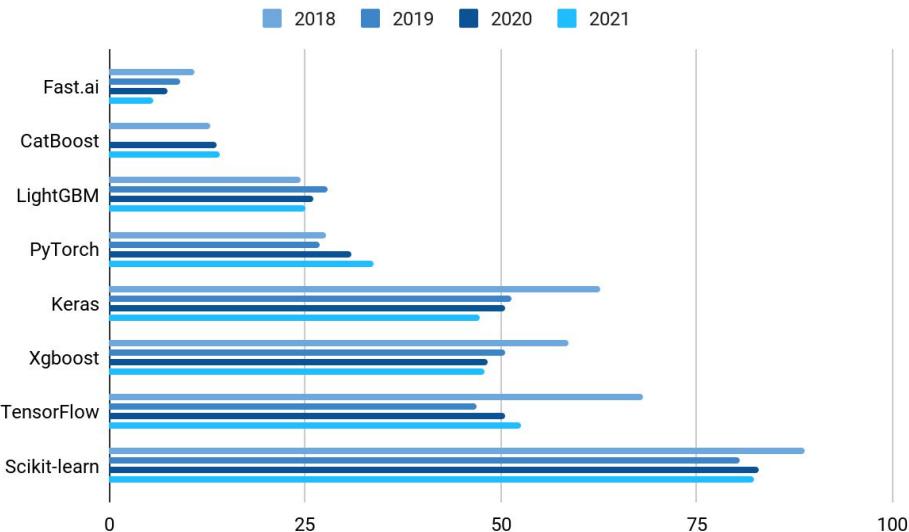
Machine Learning Framework Usage



Machine Learning Frameworks (cont.)

Despite being used less frequently overall, we continue to see strong year-over-year growth of the PyTorch framework.

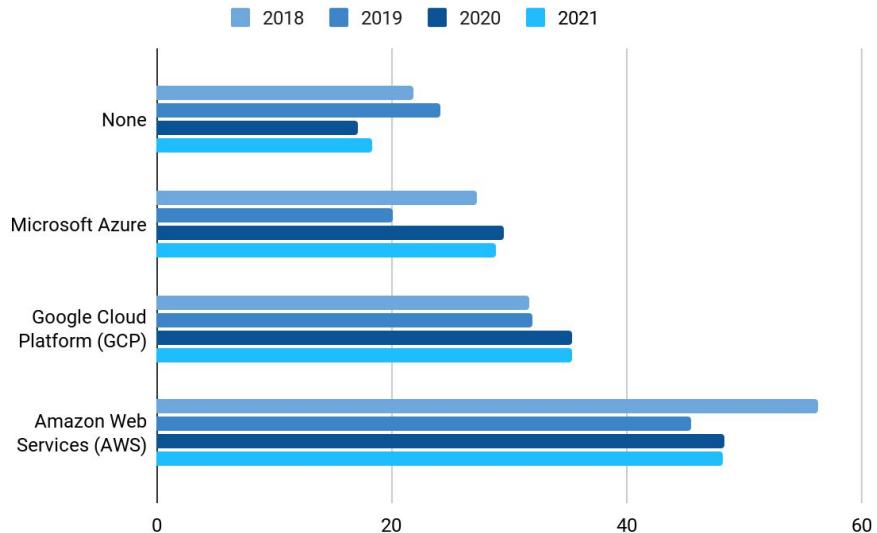
ML Framework Popularity



Enterprise Cloud Computing

The three big players in cloud computing continue to be Amazon Web Services, Google Cloud Platform, and Microsoft Azure in that order of usage.

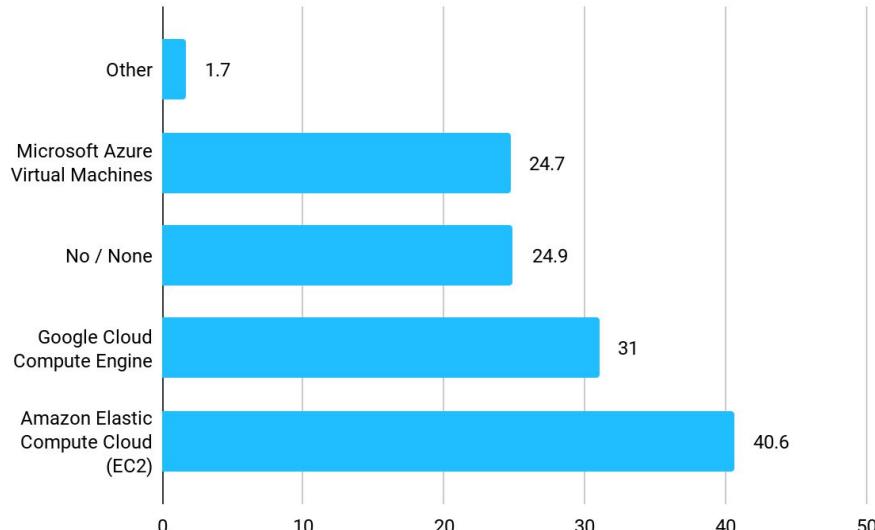
Cloud Provider Popularity



Enterprise Cloud Computing (cont.)

Those who use cloud services were also asked about specific products in the survey. Amazon's Elastic Compute Cloud was the most popular cloud computing product, but Google Cloud's Compute Engine and Azure's Virtual Machines also have strong adoption. One in four did not name a cloud product.

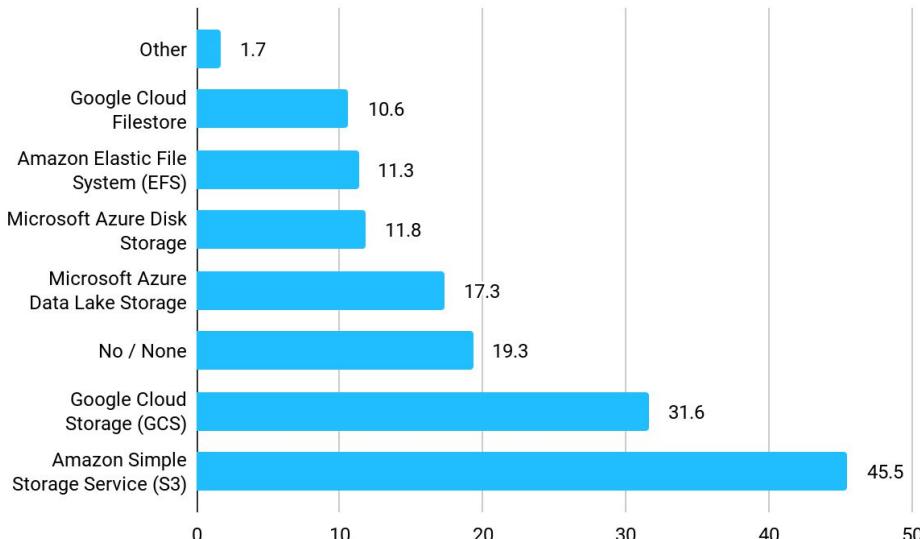
Cloud Computing Products (AWS/GCP/Azure)



Enterprise Cloud Computing (cont.)

Likewise, Amazon's Simple Storage Service (S3) was the most popular data storage product, but Google Cloud Storage and Azure Data Lake Storage also have strong adoption.

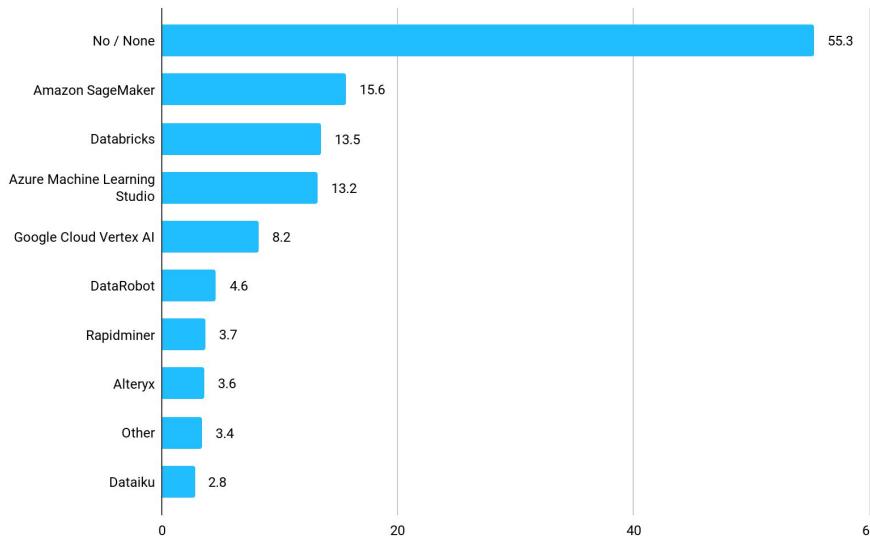
Data Storage Product (AWS/GCP/Azure)



Enterprise Machine Learning Tools

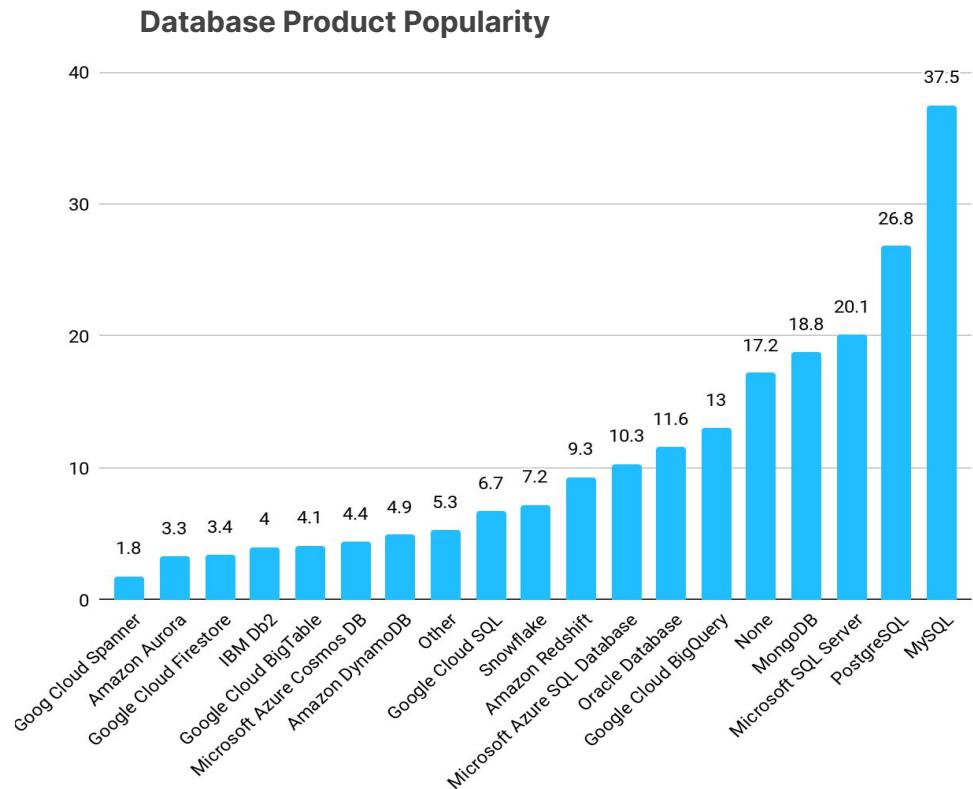
Like last year, of enterprise ML customers, Amazon SageMaker was by far the most popular choice. Another exciting product is Databricks — it had similar adoption to Azure ML Studio (~13%) and greater adoption than Google Cloud Vertex AI (~8%).

Enterprise Machine Learning Product Usage



Enterprise Big Data

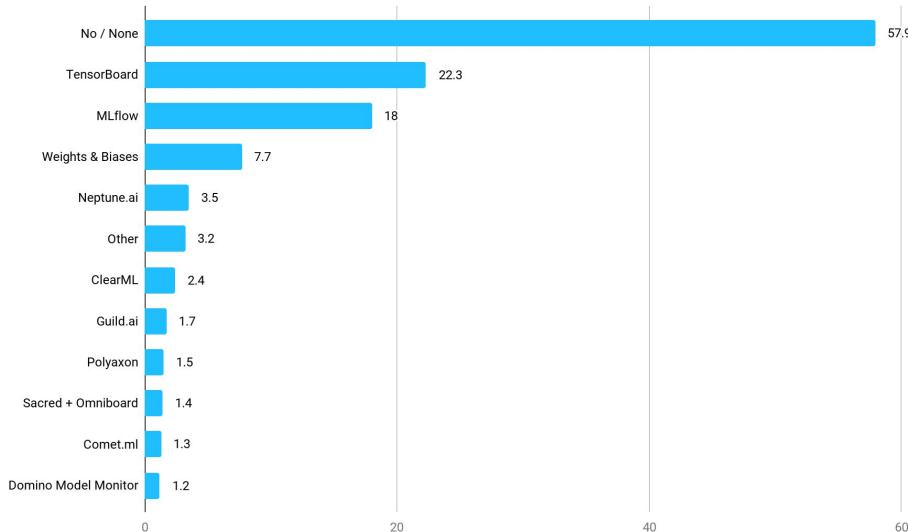
Regarding databases, there isn't a clear favorite among data scientists. MySQL, PostgreSQL, and Microsoft SQL Server maintained the top three spots.



Machine Learning Experiments

Compared to last year there are more data scientists using tools to keep track of and manage their experiments. TensorBoard continues to be a favorite (22.3%) with MLflow following close behind (18%).

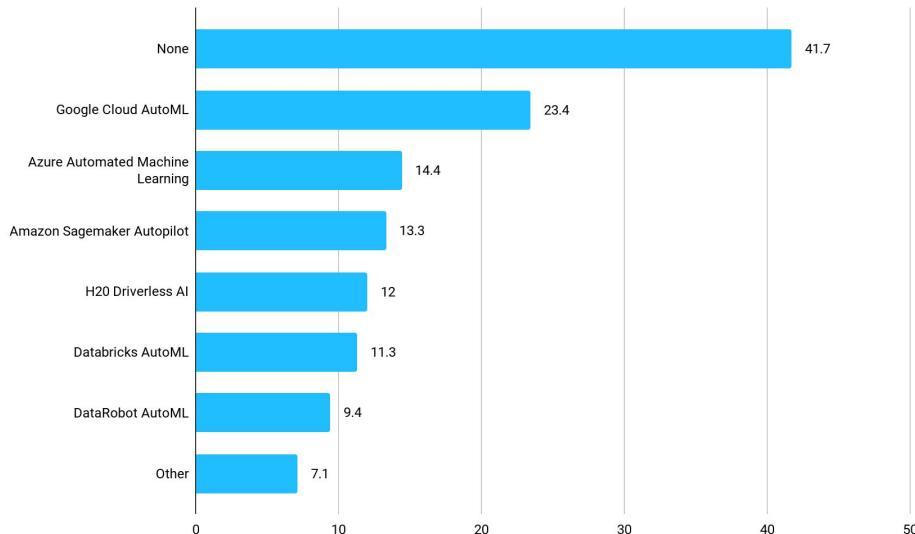
Usage of Machine Learning Experiment Tools



Automated Machine Learning

Google Cloud AutoML maintained its top position in the AutoML category.

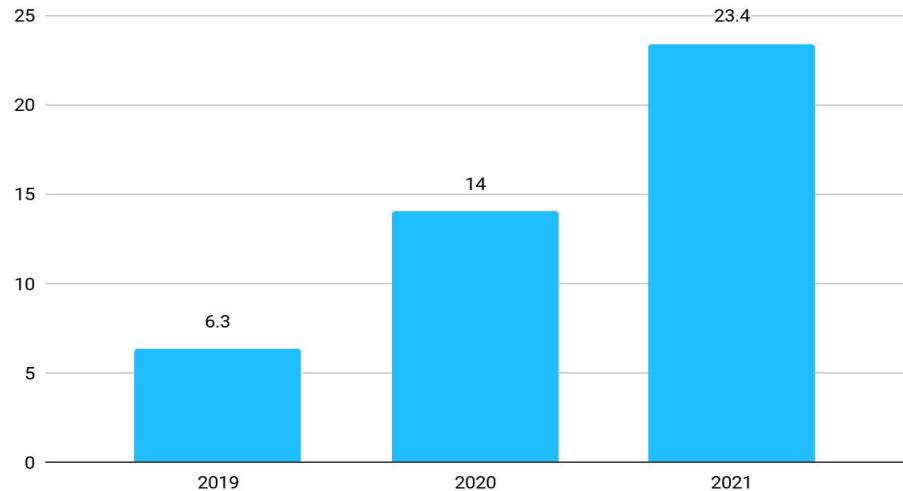
Automated Machine Learning Framework Usage



Automated Machine Learning (cont.)

Adoption of Google Cloud's AutoML technology has grown steadily over the past several years.

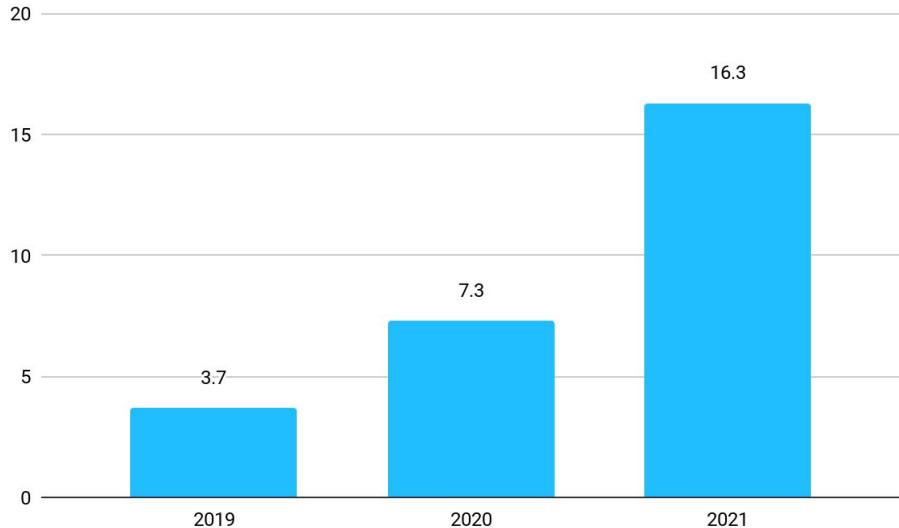
Regular Usage of Google Cloud Auto ML



Tensor Processing Units

Google Cloud's Tensor Processing Units (TPUs) also showed strong year-over-year growth.

Regular Usage of TPUs



Conclusion

Kaggle has published the complete dataset of responses for the community to review, and we'll run a competition from October 14 to November 28th, 2021 to learn even more about data science practitioners in 2021.

