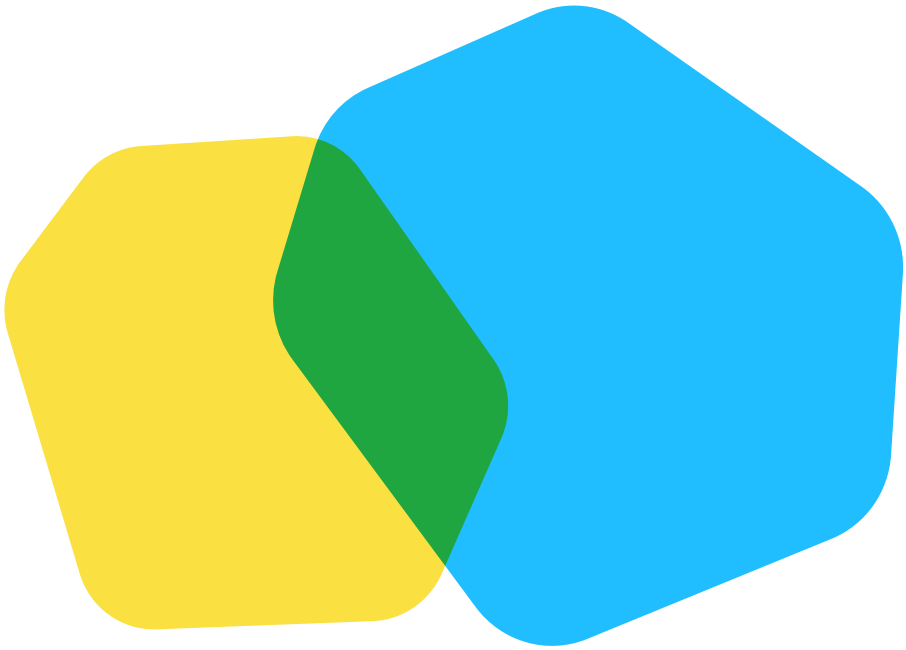
kaggle™

State of Machine Learning and Data Science 2020



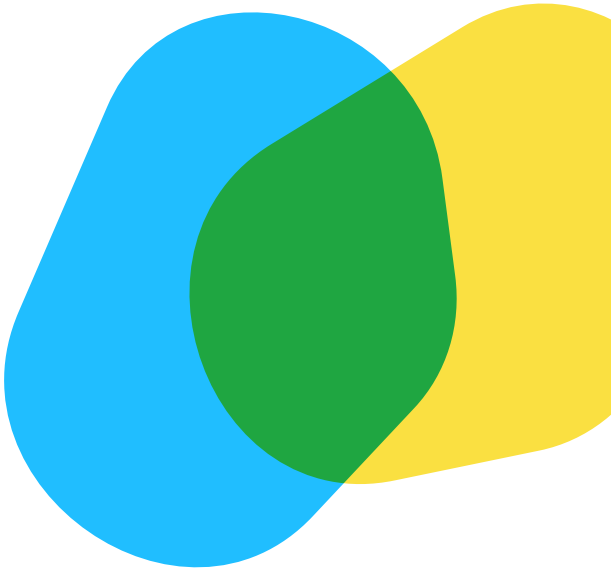
Enterprise Executive Summary Report

Table of Contents

Overview

For the fourth year, Kaggle surveyed its community of data enthusiasts to share trends within a quickly growing field. **Based on responses from 20,036 Kaggle members, we've created this report focused on the 13% (2,675 respondents) who are currently employed as data scientists.**

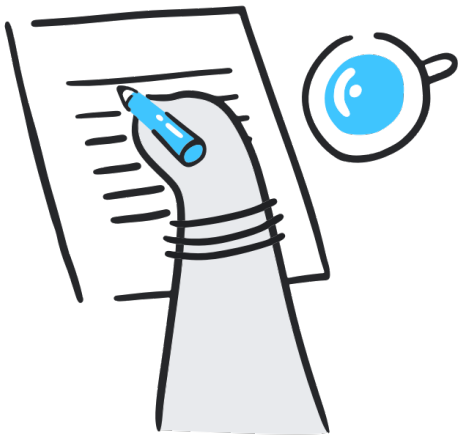
We can see a clear picture of what is common in the community but also the diverse attributes of its members.



Report Methodology

The content of this report focuses on respondents who are currently employed and chose their current job title as "data scientist". There are many other job titles that support data science and machine learning workflows and you can find their responses in the complete 2020 survey dataset on Kaggle.

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%. All monetary amounts captured in the report are in USD.



Key Results

Profile

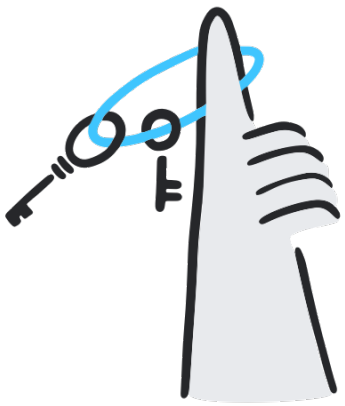
- Data science continues to have a heavy gender imbalance, with most identifying as male
- The vast majority of data scientists are under 35 years old
- Over half of data scientists have graduate degrees

Education and Employment

- Most data scientists continue to learn outside of formal education
- Most data scientists have been coding for less than a decade
- More than half of data scientists have less than three years of experience with machine learning
- Data scientists in the United States make substantially more money than their international counterparts

Technology

- More data scientists use cloud computing compared to 2019 results
- Scikit-learn is the most popular machine learning tool in 2020, with over four in five data scientists using it
- Tableau and PowerBI are the most popular business intelligence tools

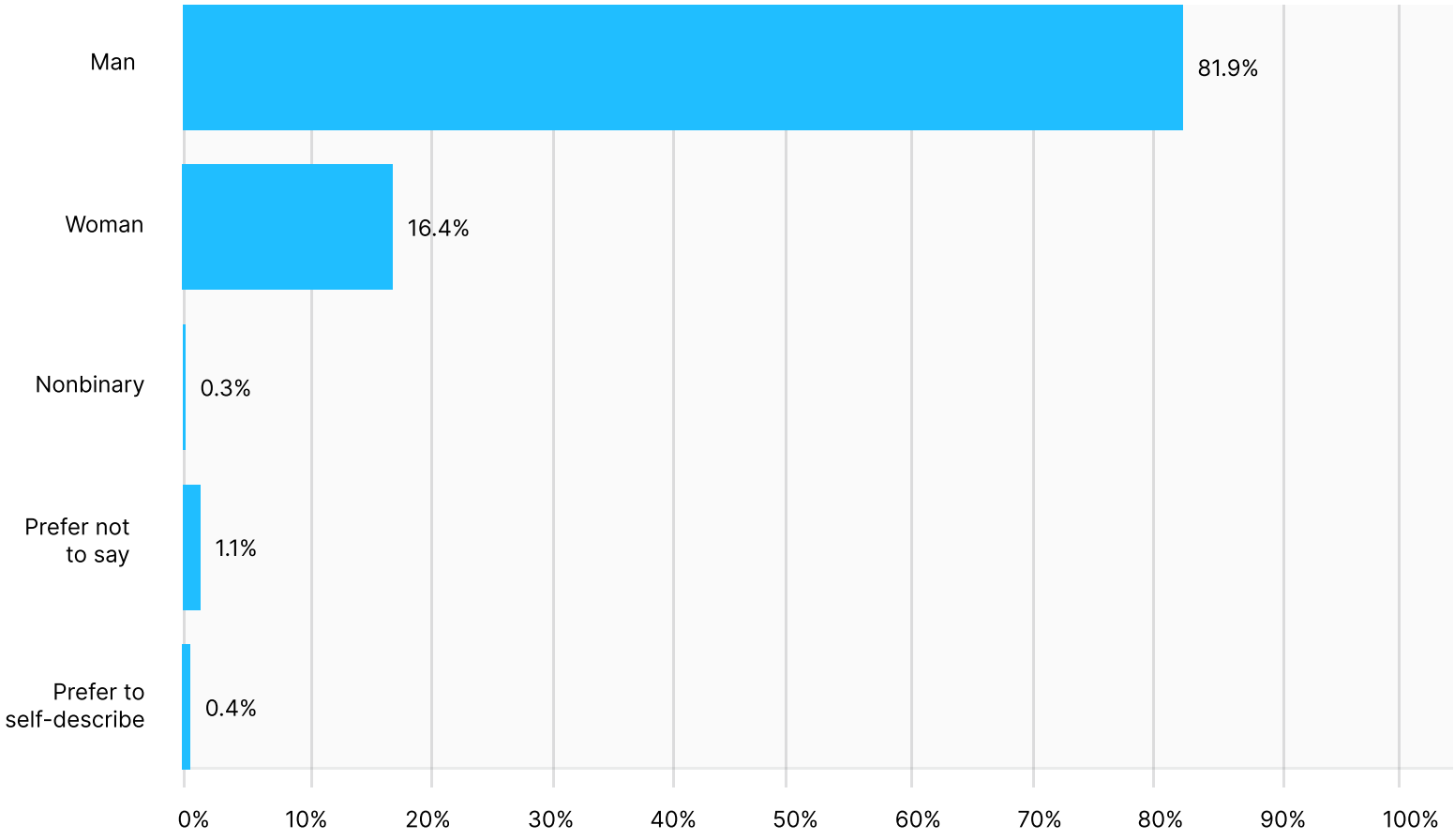


Data Scientist Profile

Gender

Data science is still suffering from a large gender gap in the workplace, as 82% of users identify as men. This is only a slight change from last year's results, where 84% of users identified as males. This is the first year we've differentiated between "Nonbinary" and "Prefer to self-describe," with each answer coming in around a third of a percent.

GENDER IDENTITY OF DATA SCIENTISTS

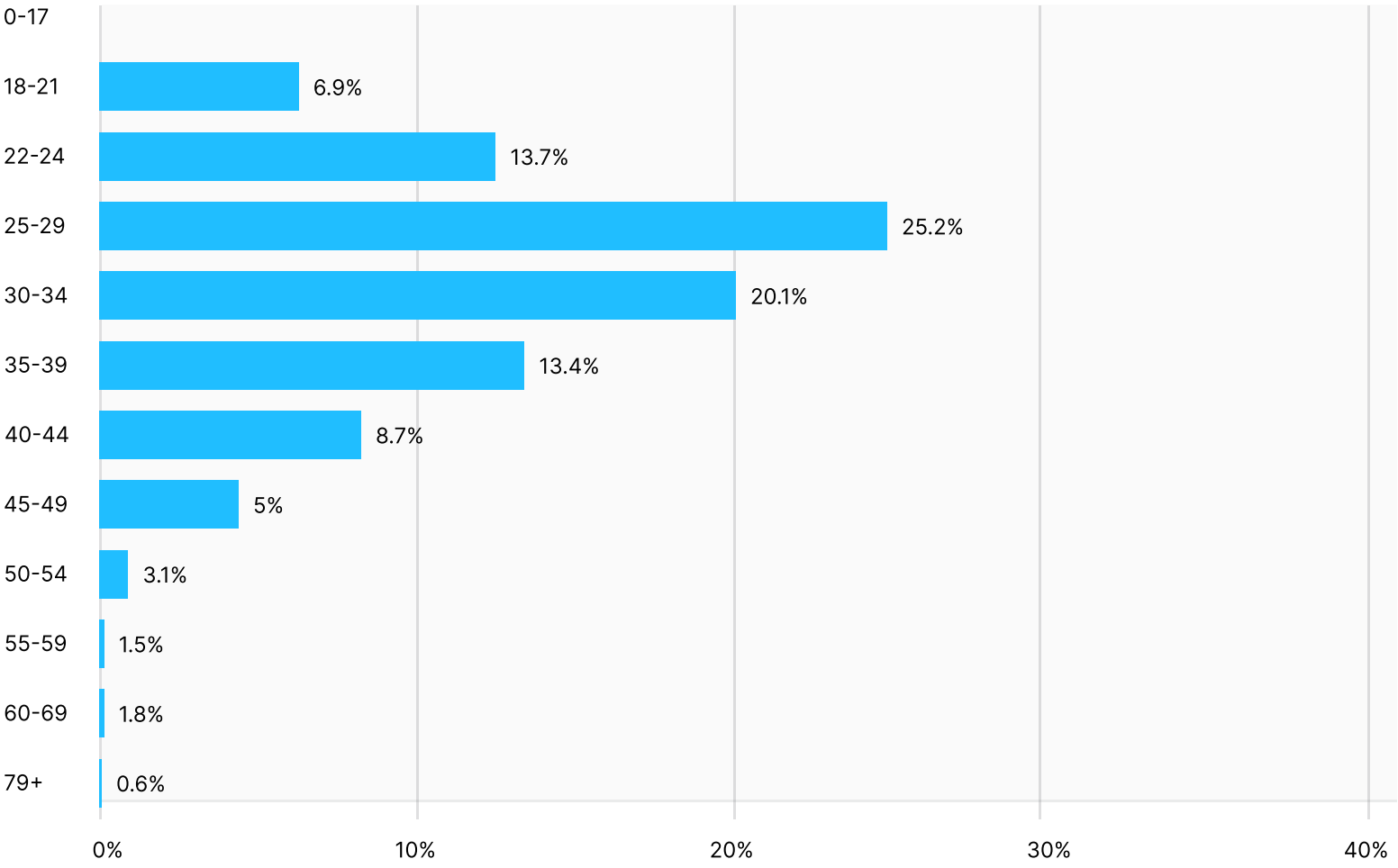


Age

Similar to 2019 results, data scientists tend to be in their late 20s or early 30s, with about 60% between 22 and 34. Only one in five professional data scientists are 40 or older. There are signs of the numbers skewing even younger, as generation Z gets more involved. Nearly 7% of data scientists are aged 18-21, an increase from last year's 5%.

Though not included in this chart, responses from students have also increased each year (26.8% in 2020, 21% in 2019, 22.9% in 2018). As these students graduate into the workforce, we may see future surveys with even younger data scientists.

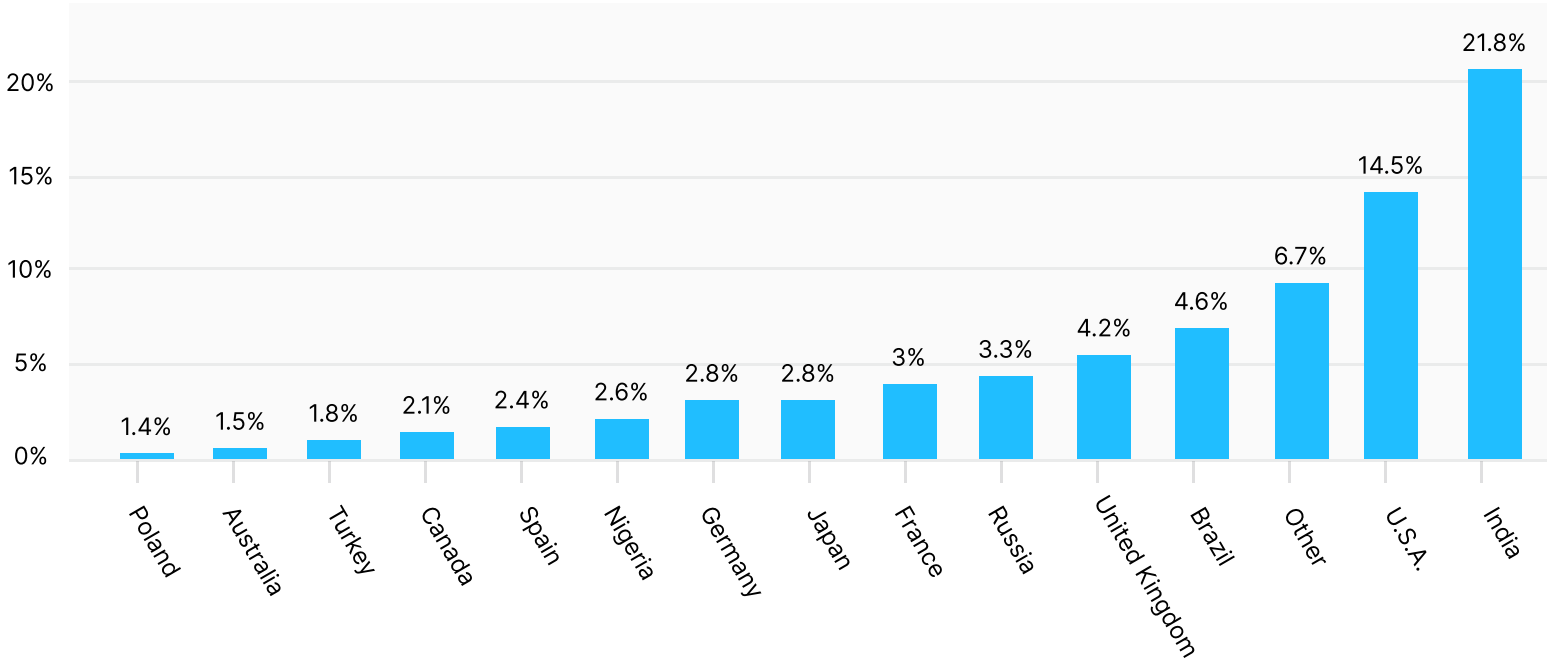
AGE RANGES OF DATA SCIENTISTS



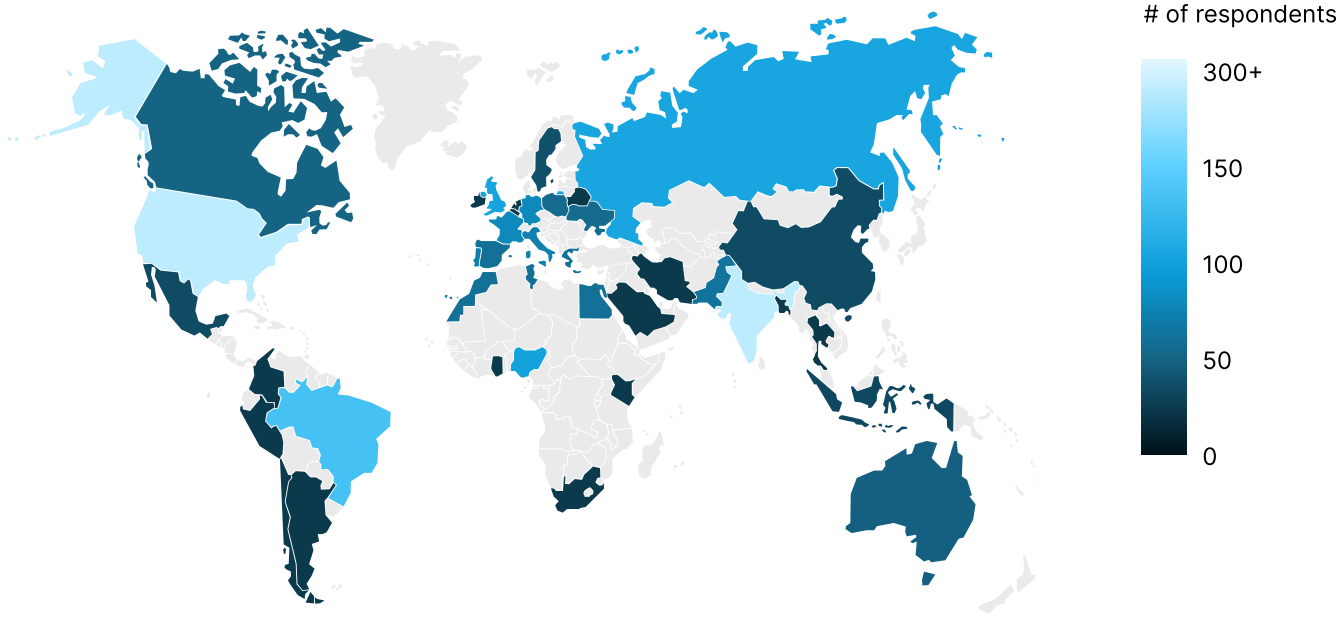
Country

Two countries have far more representation in the Kaggle community. India makes up almost 22% of Kaggle data scientists, while 14.5% reside in the United States. Brazil is a distant third, at under 5%.

MOST COMMON NATIONALITIES



RESPONSES PER COUNTRY

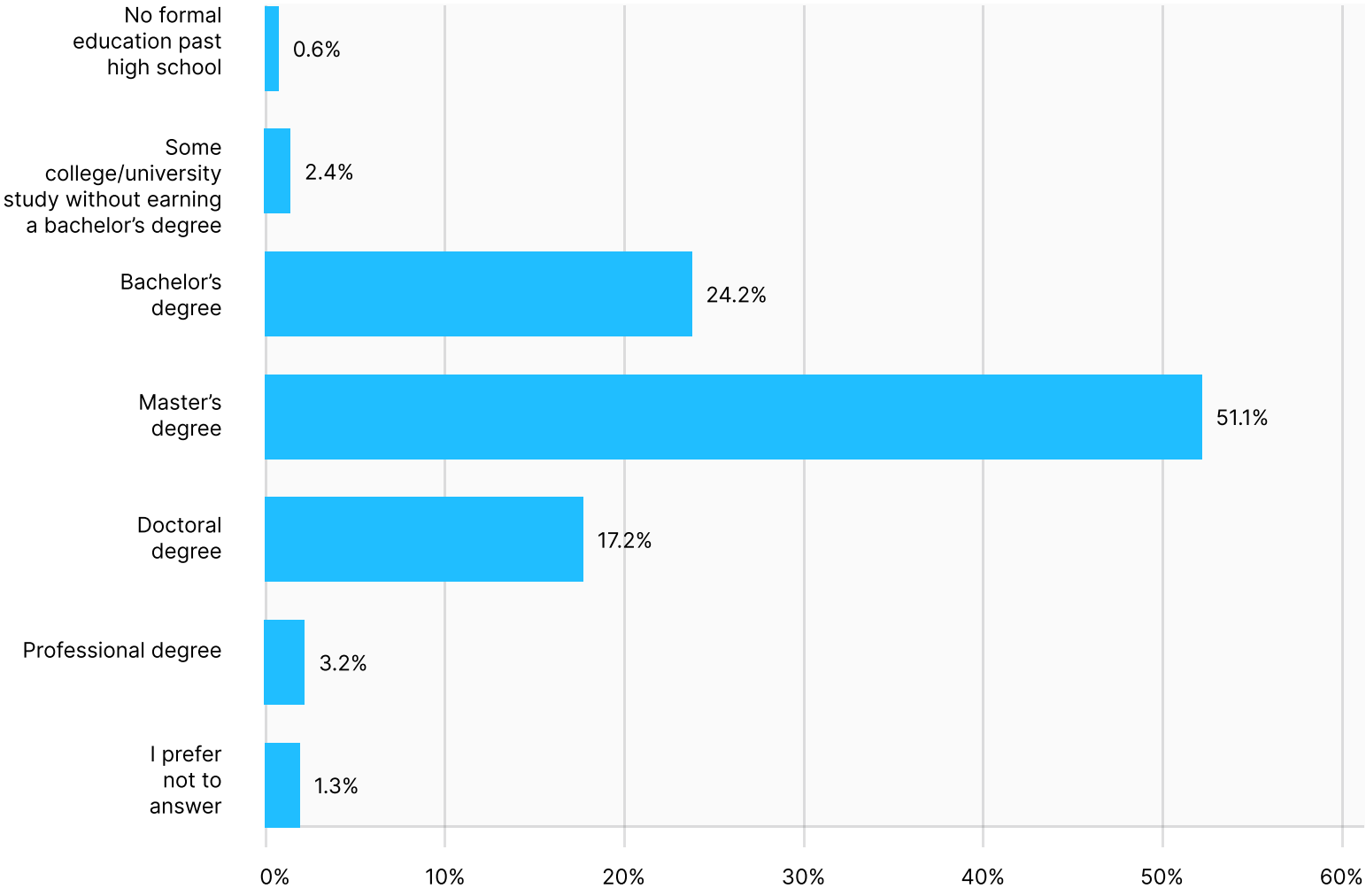


Education

Higher Education

Graduate degrees continue to be the norm for data scientists, with over 68% having obtained either a Master's or doctoral degree. Fewer than 5% of data scientists have no degree beyond a high school diploma.

EDUCATION LEVEL OF KAGGLE DATA SCIENTISTS

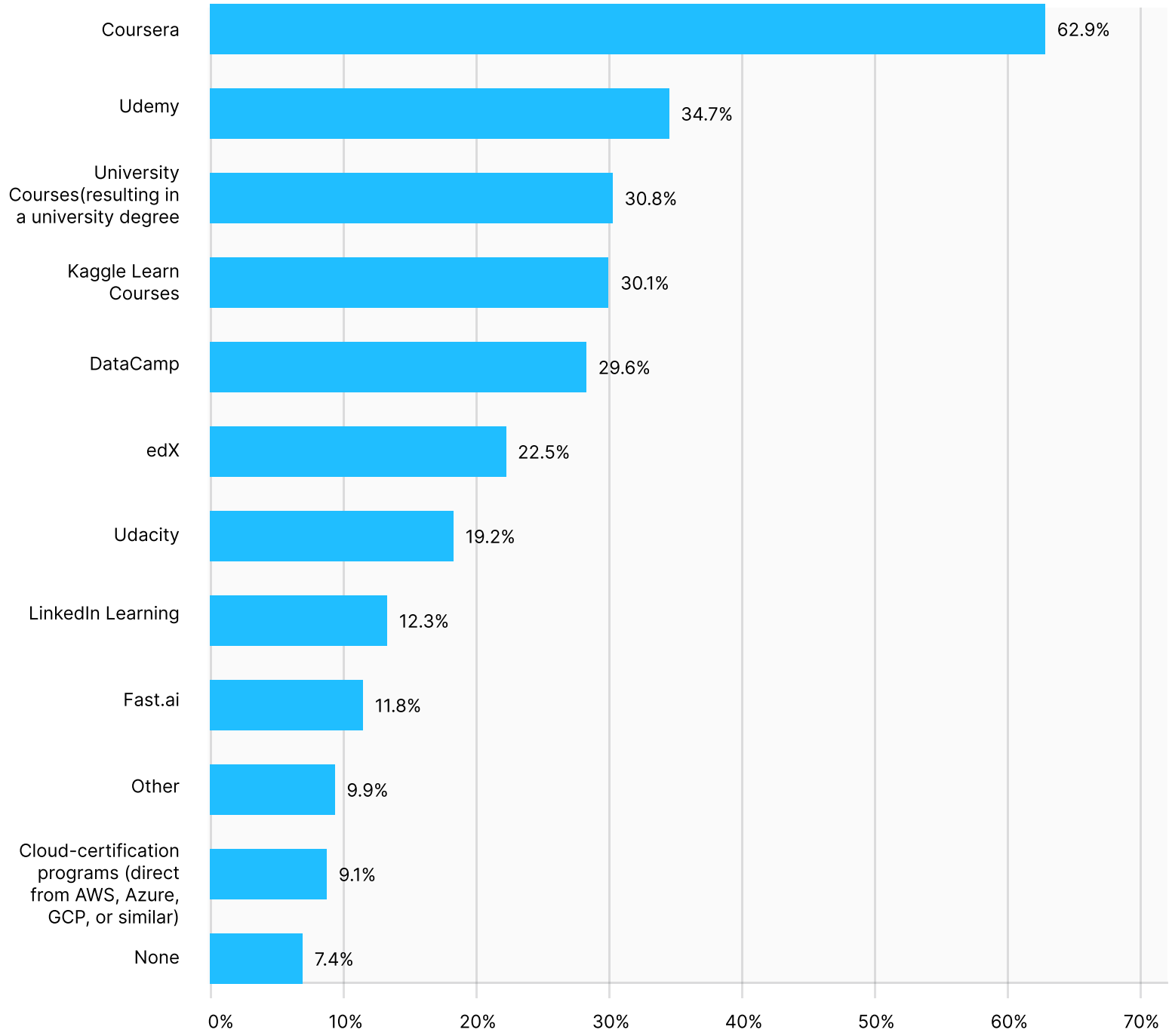


Ongoing Learning

Data science and machine learning are quickly changing, so it's no surprise over 90% of Kaggle data scientists maintain ongoing education. While about 30% take traditional higher education courses, many more learn through online materials.

Coursera, Udemy, and Kaggle Learn top the most common mediums in our survey. Unsurprisingly, many Kaggle data scientists chose multiple resources in the survey, with an average of 2.8 mediums selected.

POPULAR ONGOING LEARNING RESOURCES



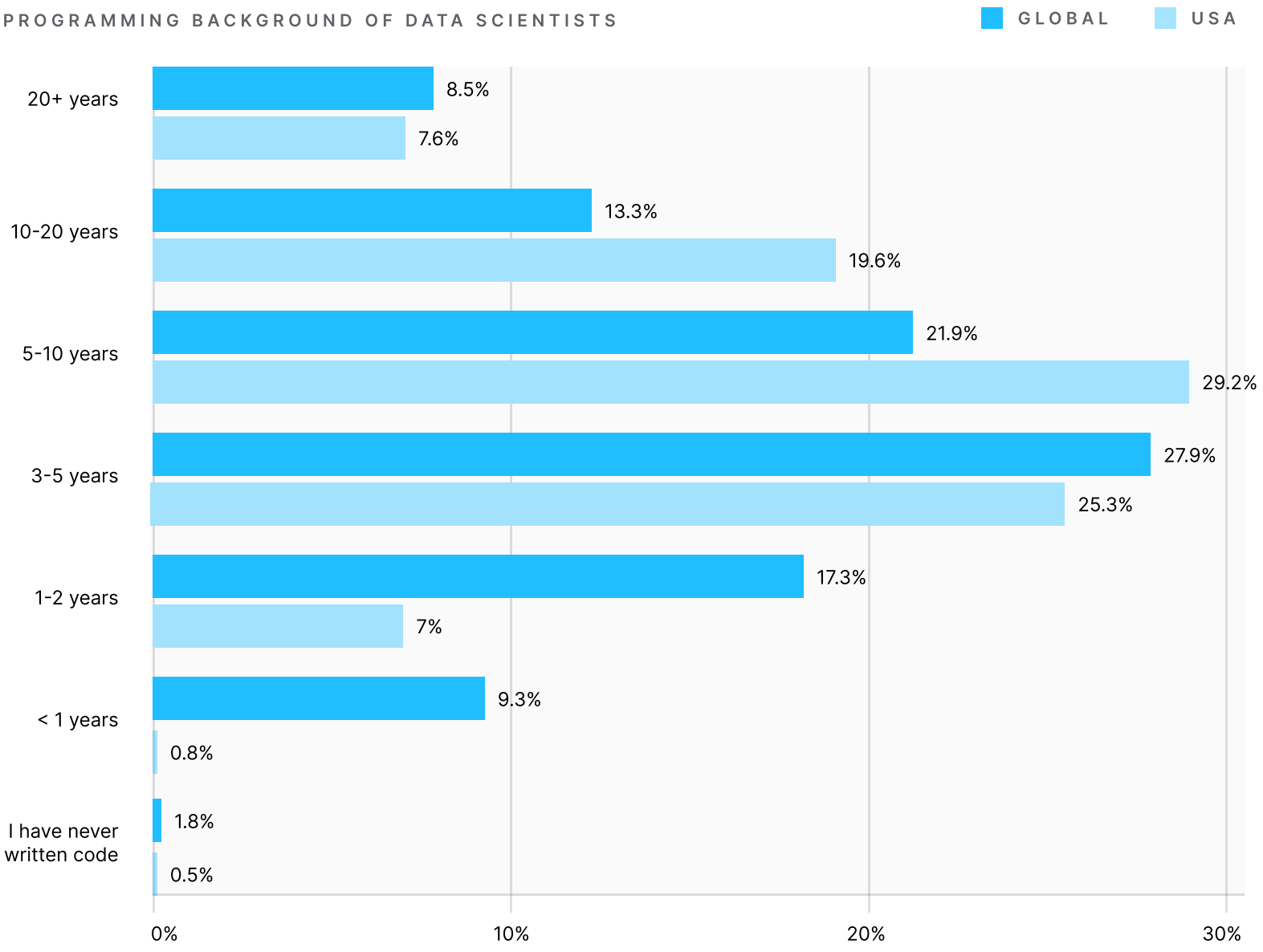
Data Science & Machine Learning Experience

Programming Experience

Most Kaggle data scientists have at least a few years of experience under their belt. Just over 8% of data scientists have been programming since the 20th century! That's not to say there aren't newcomers, however. Over 9% have taken up programming in the last year. Just under 2% of data scientists claim to have never written code at all.

Compared to the global audience, United States data scientists have significantly greater programming experience. In the US, 37% have been programming 10 or more years, versus 22% worldwide.

PROGRAMMING BACKGROUND OF DATA SCIENTISTS

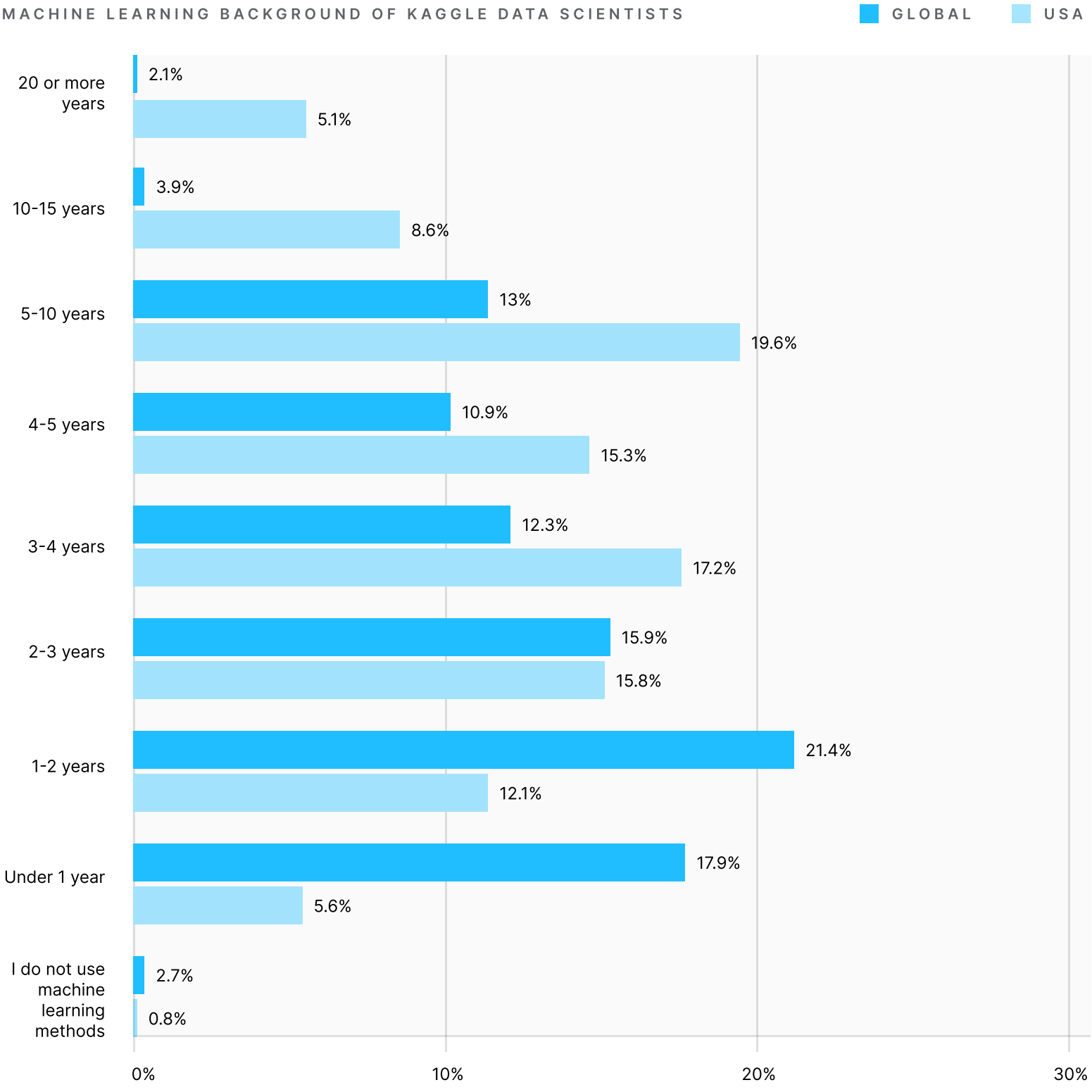


Machine Learning Experience

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.

MACHINE LEARNING BACKGROUND OF KAGGLE DATA SCIENTISTS



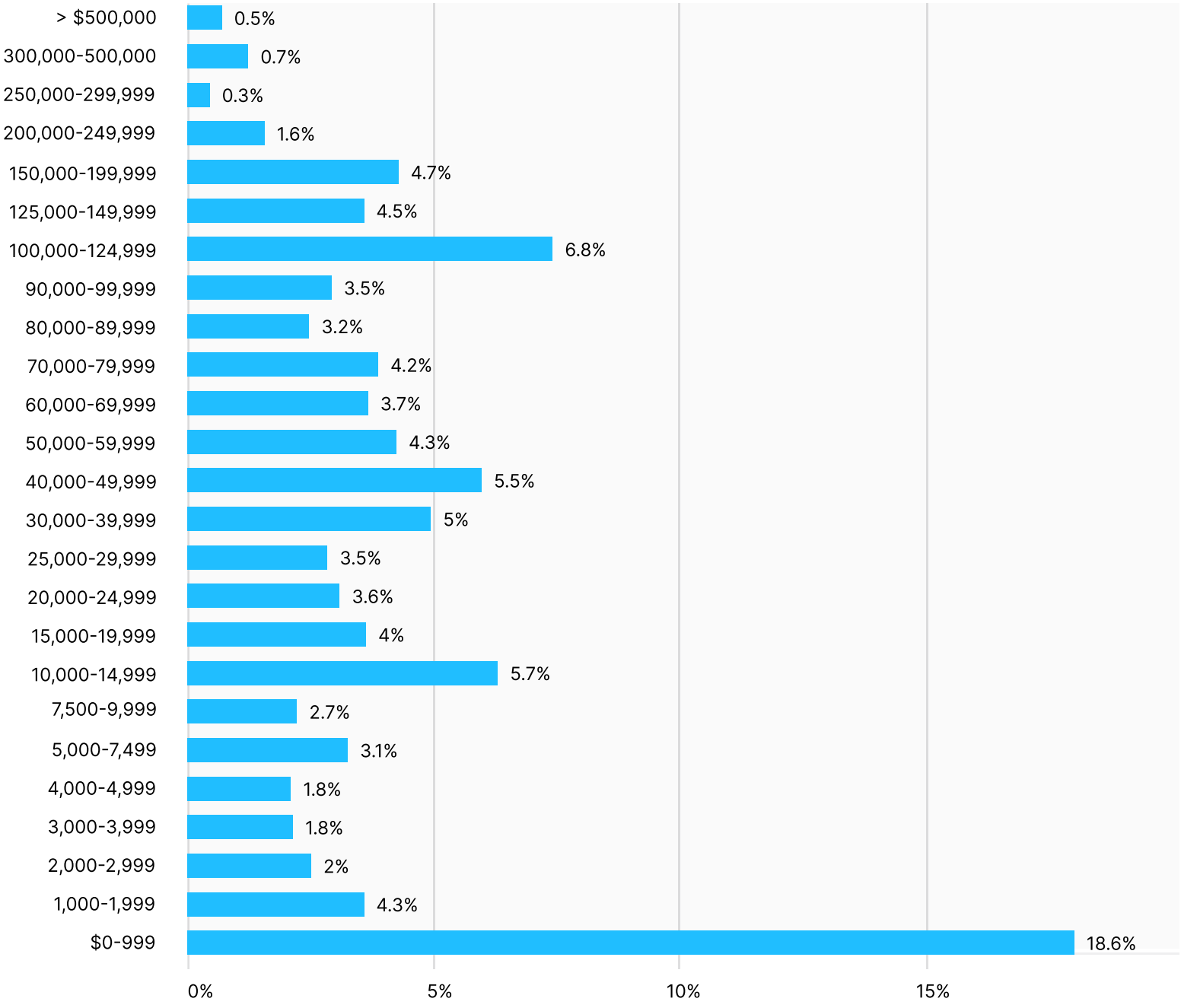
Employment

Pay

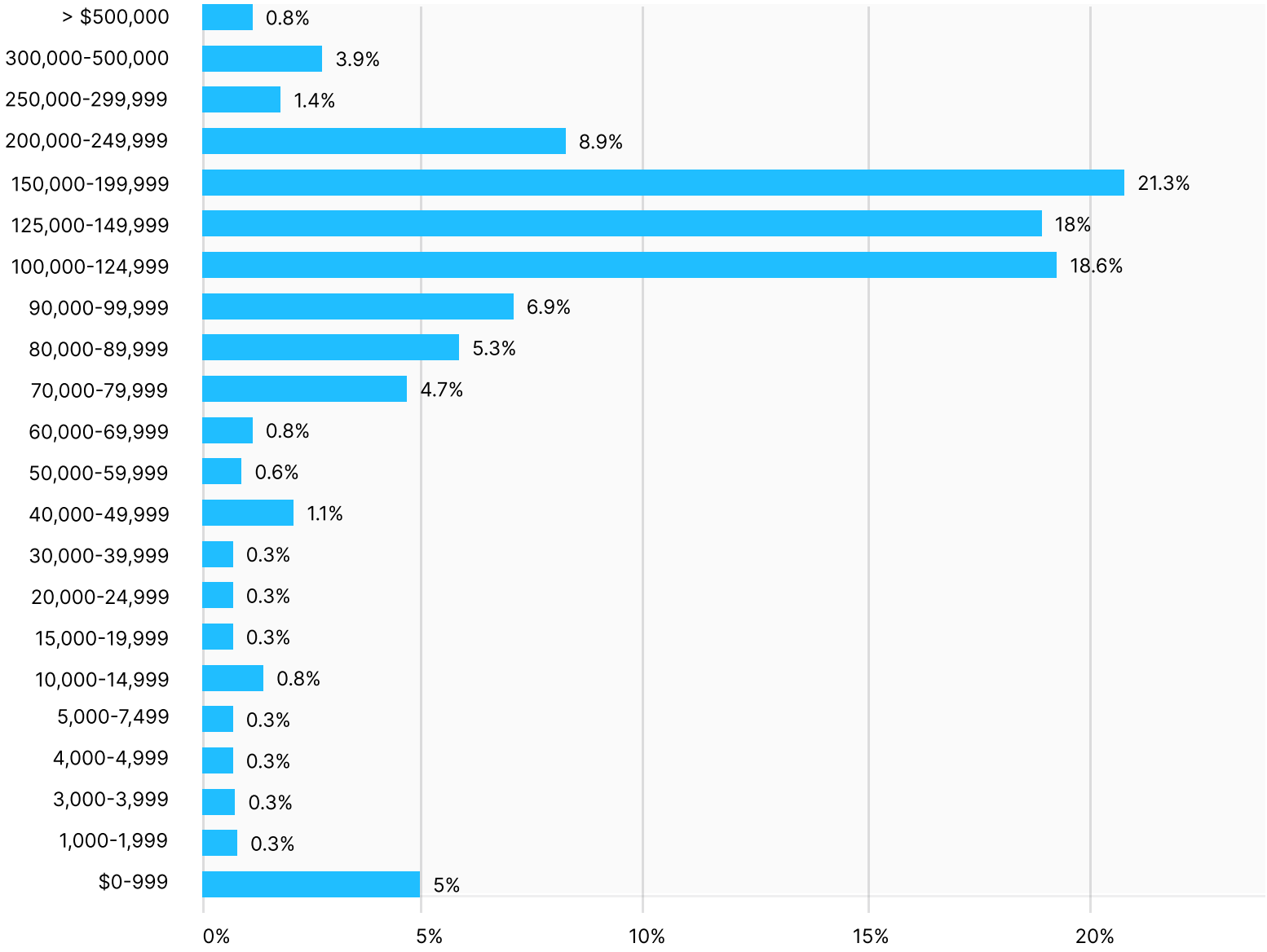
Companies in the United States are most likely to pay in the six figures, based on these survey results. Global companies have lower salary ranges that are more evenly distributed.

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

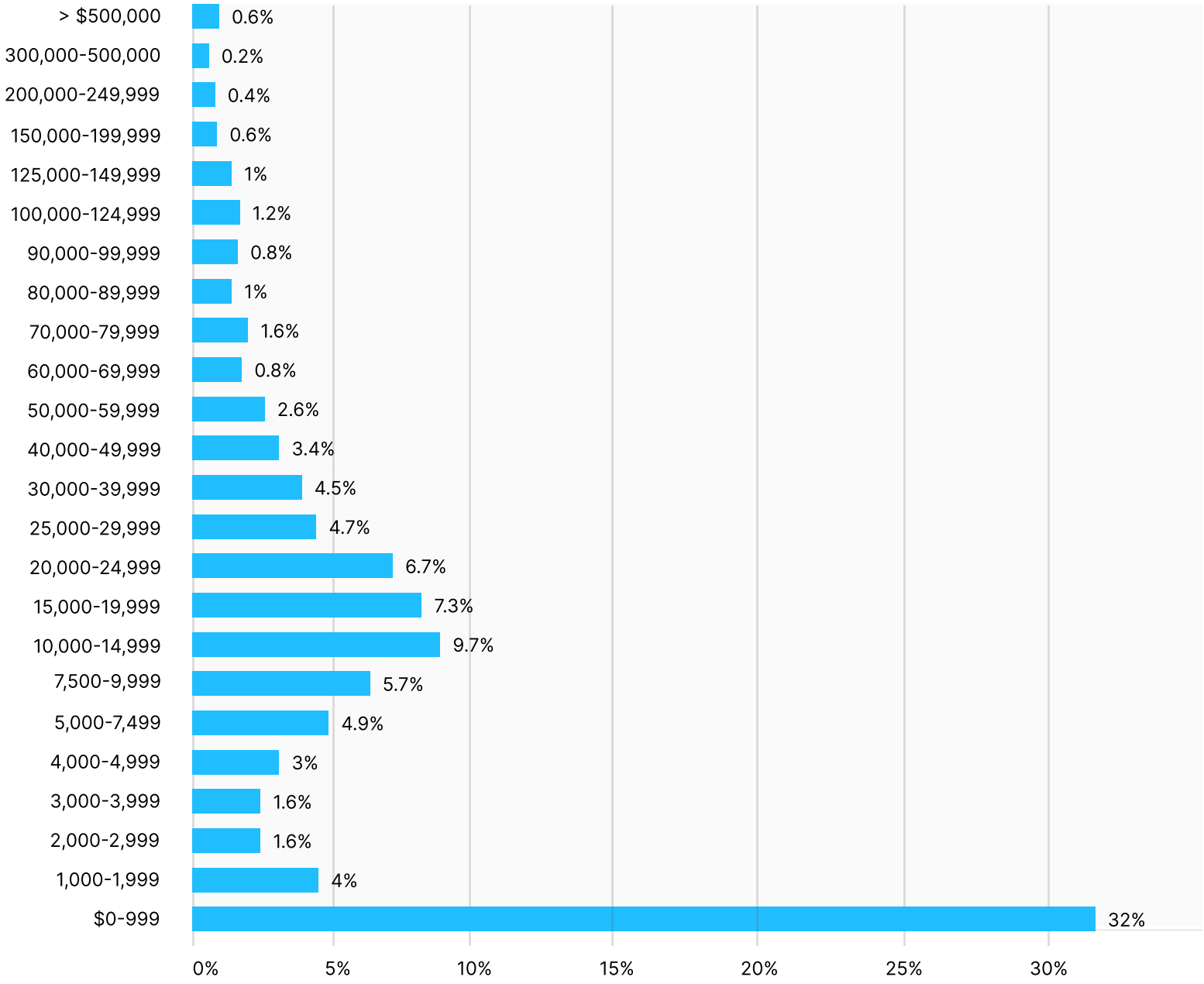
GLOBAL SALARY DISTRIBUTION FOR DATA SCIENTISTS



SALARY DISTRIBUTION FOR US-BASED DATA SCIENTISTS

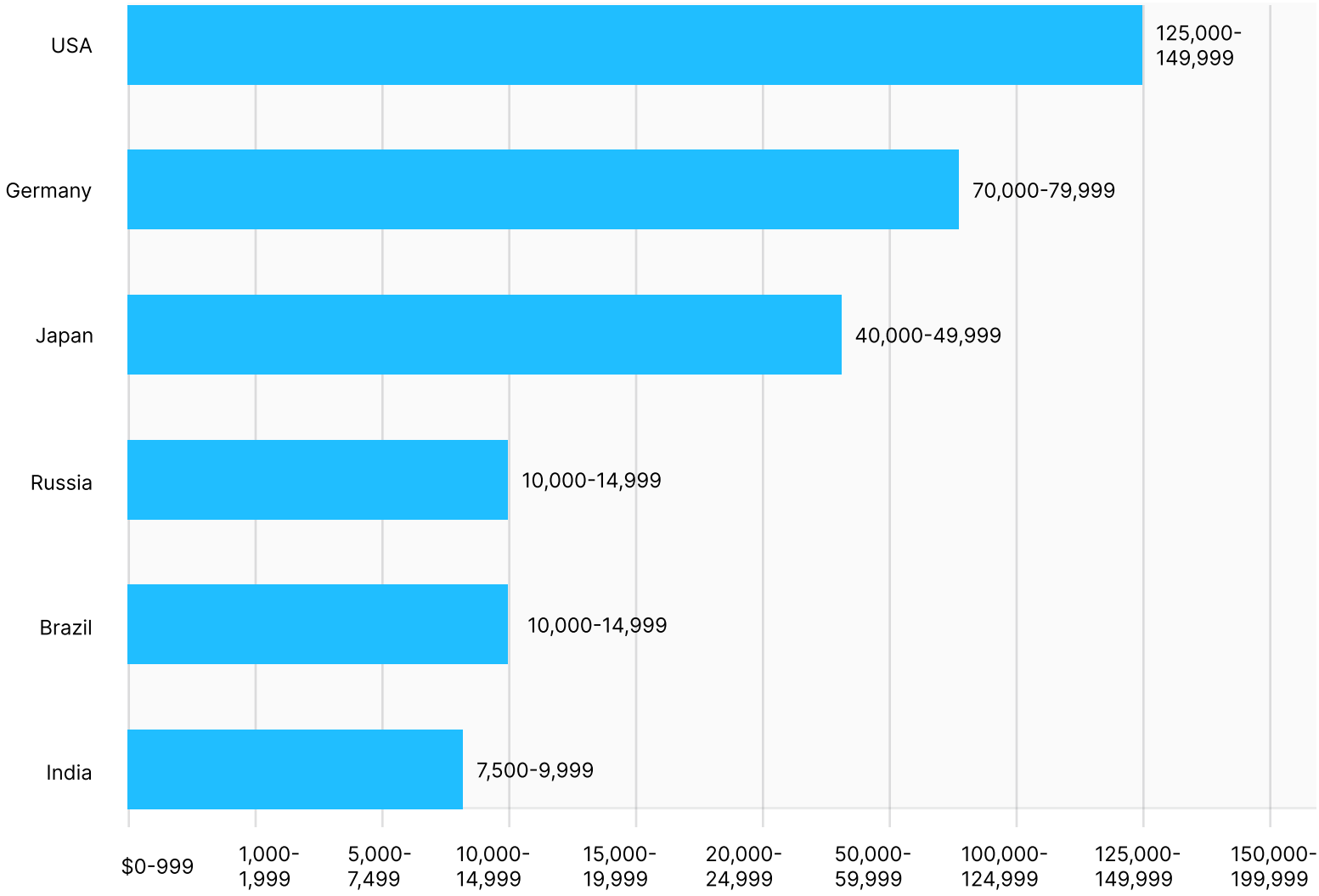


SALARY DISTRIBUTION FOR INDIA-BASED DATA SCIENTISTS



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 FOR DATA SCIENTISTS BY COUNTRY

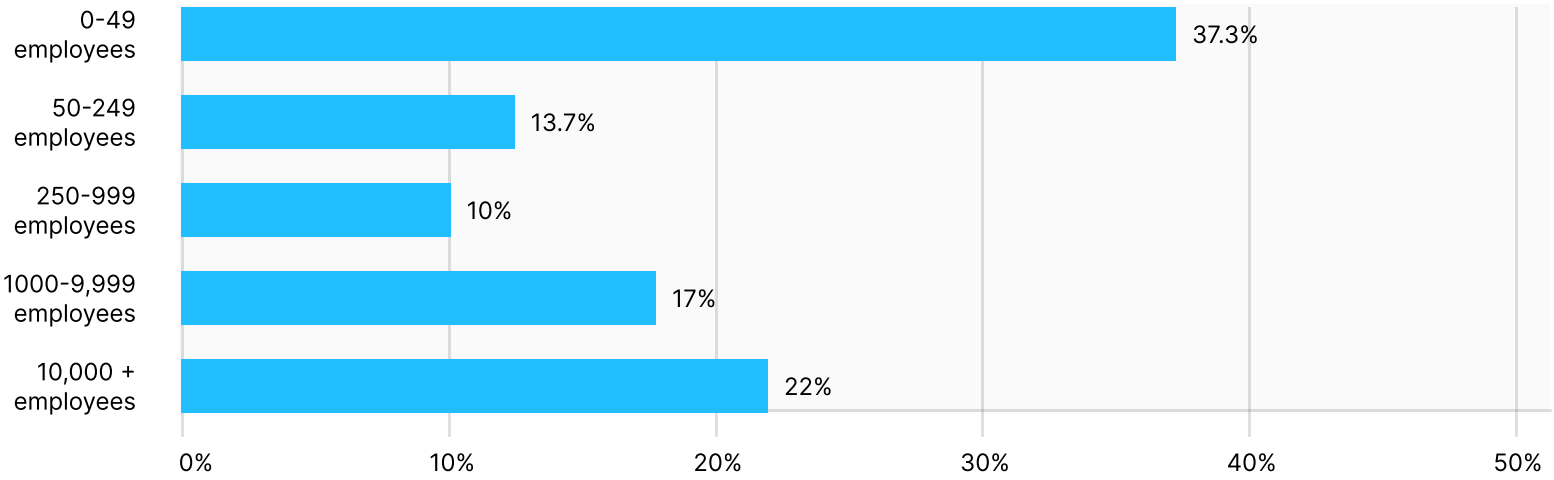


Companies Employing Data Science

The most notable change from last year is that more Kaggle data scientists are working at the very smallest businesses, at over 37% (up from 30% in 2019).

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 EMPLOYEES)

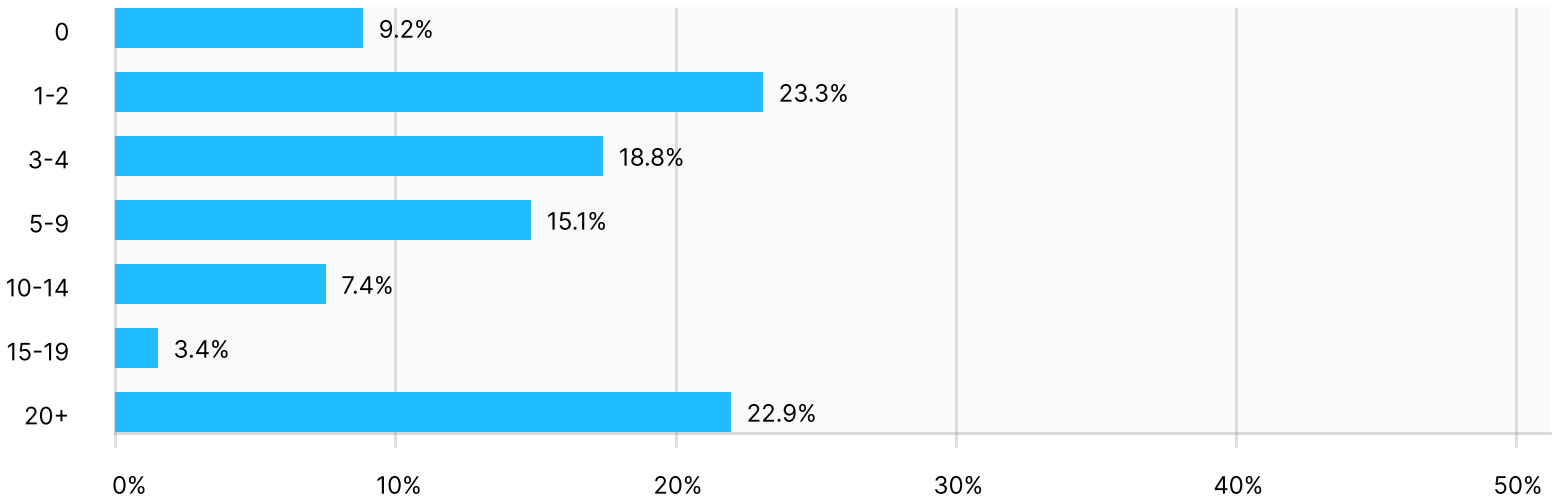


Data Science Teams

With small companies being most common, it reasons that the same is true for data science teams, most of which could be fed with two pizzas.

Over half of data scientists work at companies with five or fewer people on the data science team. Teams of one or two are most common (23.25%), but large teams of 20+ come next at 22.93%.

DATA SCIENCE TEAMS (# OF EMPLOYEES)

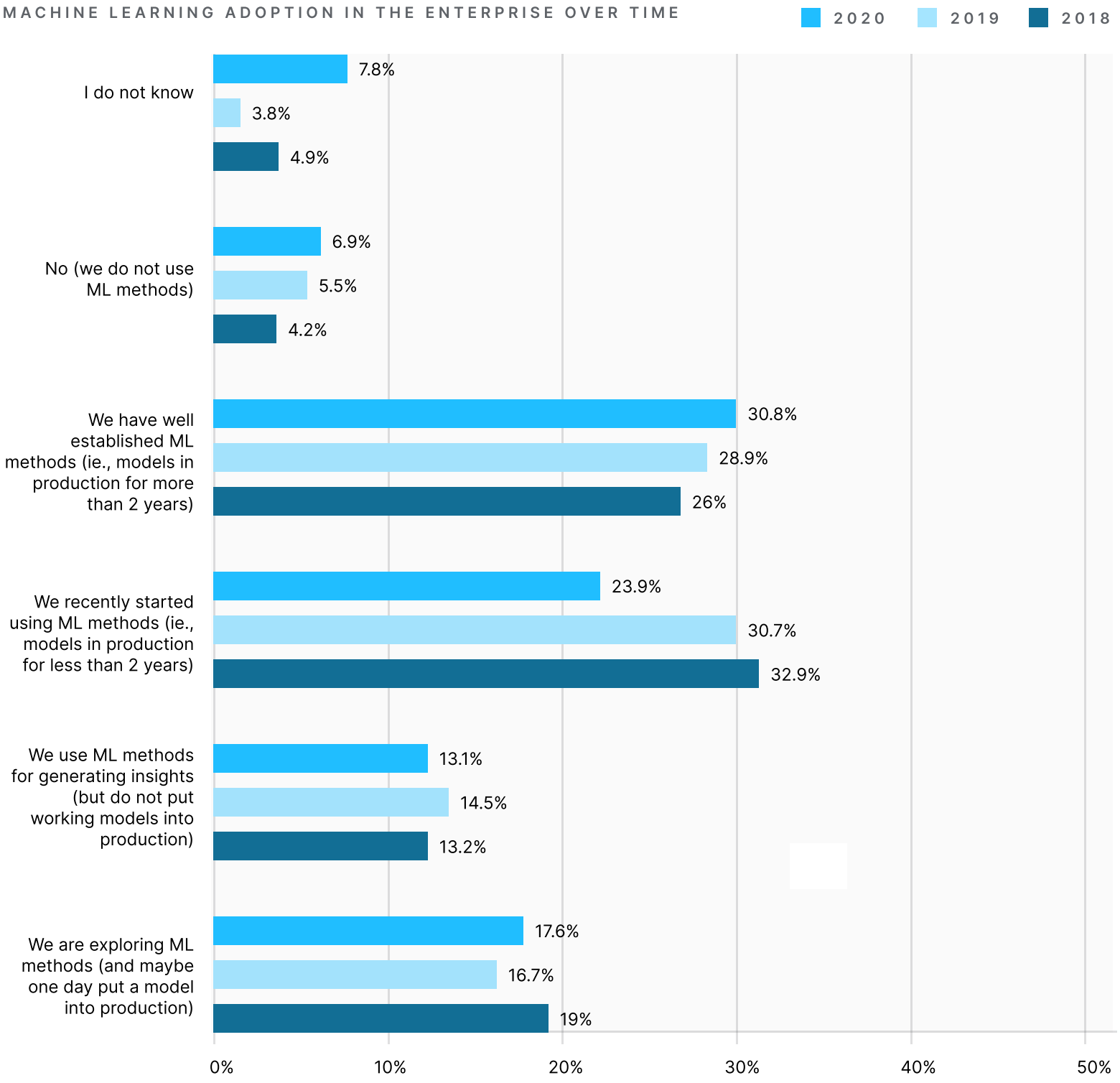


Enterprise Machine Learning Adoption

Machine learning has become more rooted in the companies where Kaggle scientists work. Nearly 31% of data scientists claim well-established ML methods, up from 28% in 2019 and 25% in 2018.

Those exploring (or using it to generate insights) remain about the same. Kaggle data scientists who said they've recently adopted ML decreased, likely due to more entrenched usage.

MACHINE LEARNING ADOPTION IN THE ENTERPRISE OVER TIME

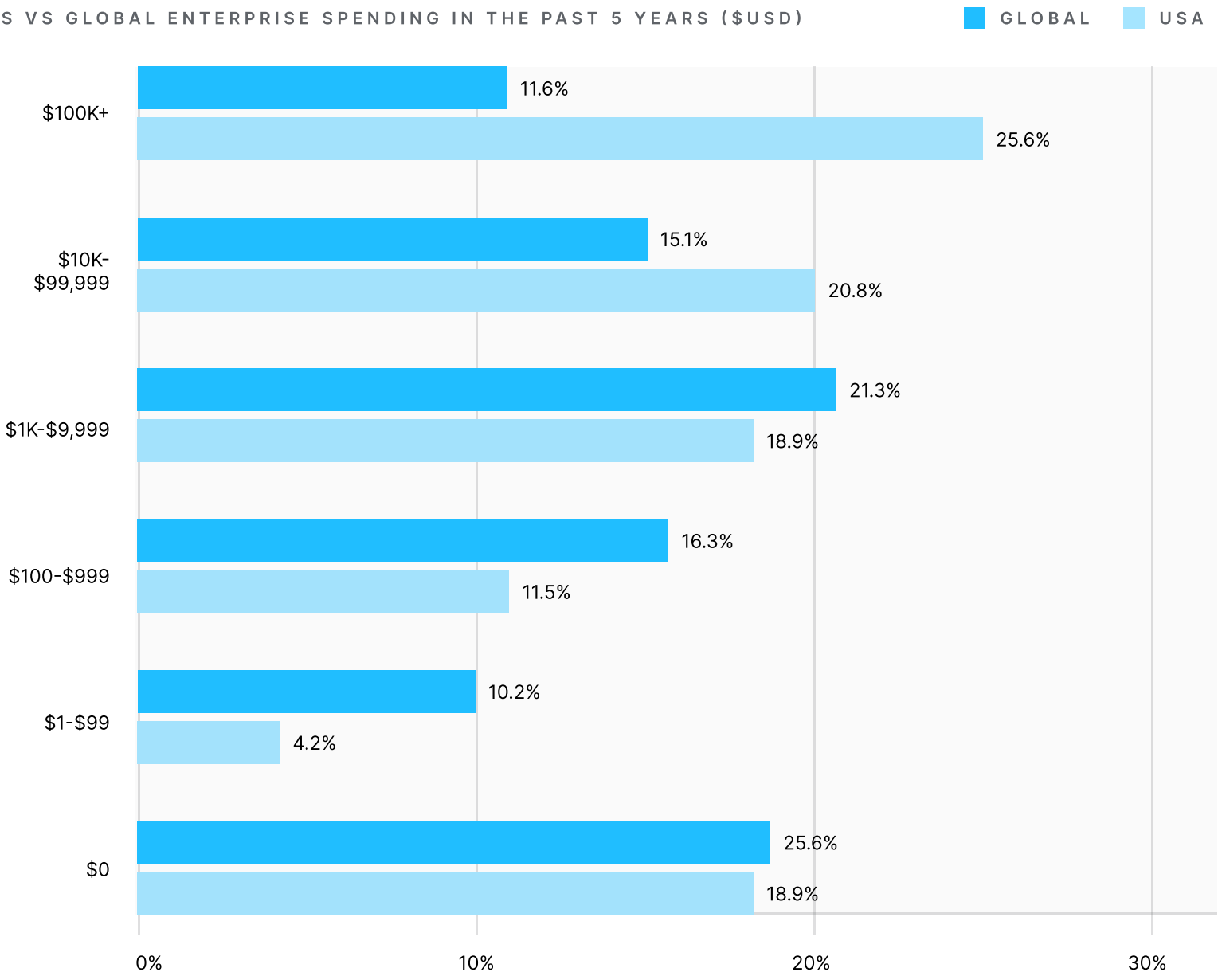


Spending

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.

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.

US VS GLOBAL ENTERPRISE SPENDING IN THE PAST 5 YEARS ($USD)



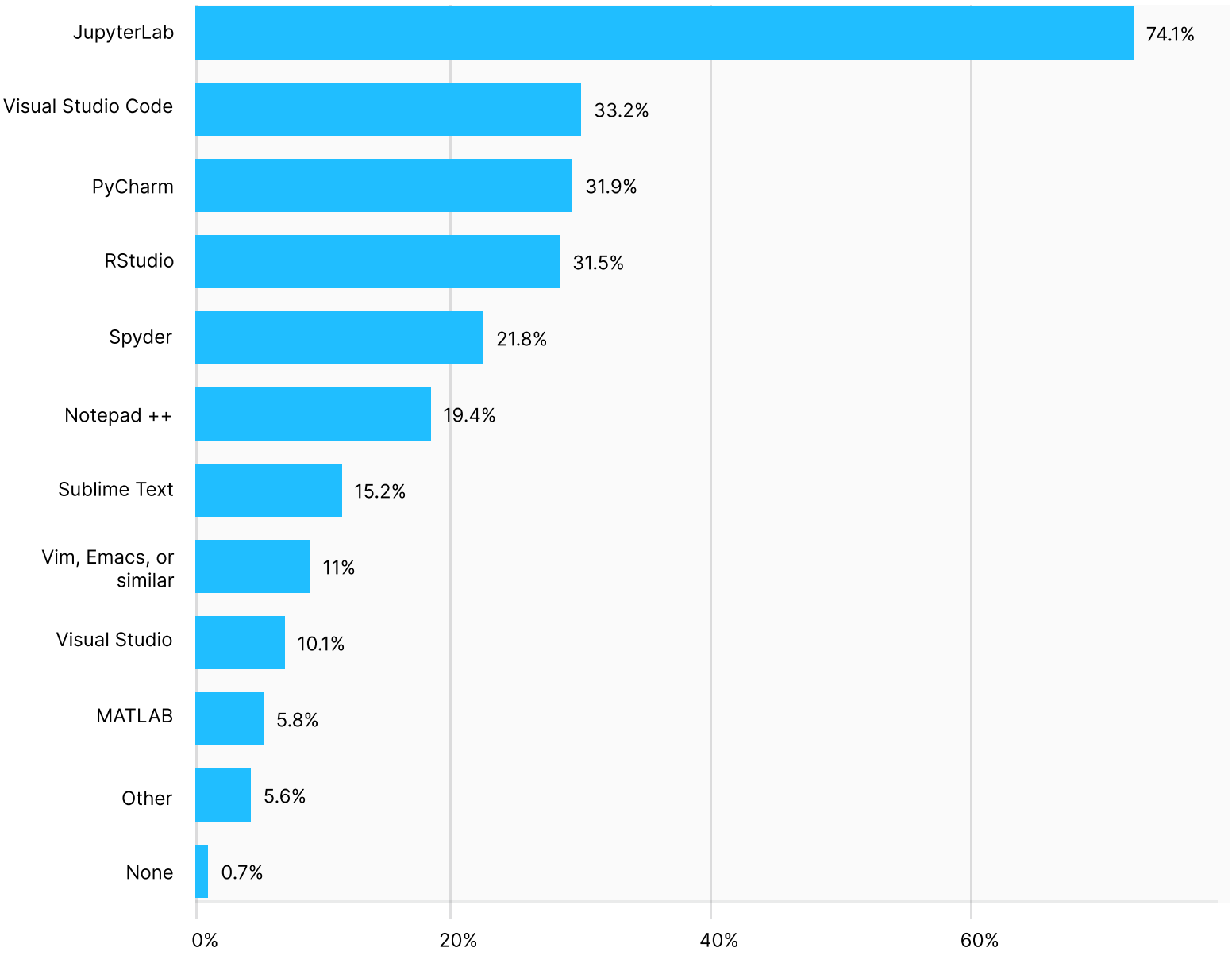
Technology

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, this has decreased from last year's 83%. Visual Studio Code is in the second spot with just over 33%.

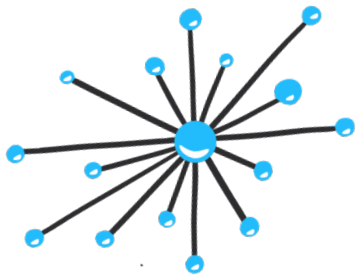
This is the first year it has been separated out from Visual Studio. The two combined for over 43% this year, versus under 30% in 2019.

POPULAR IDE USAGE

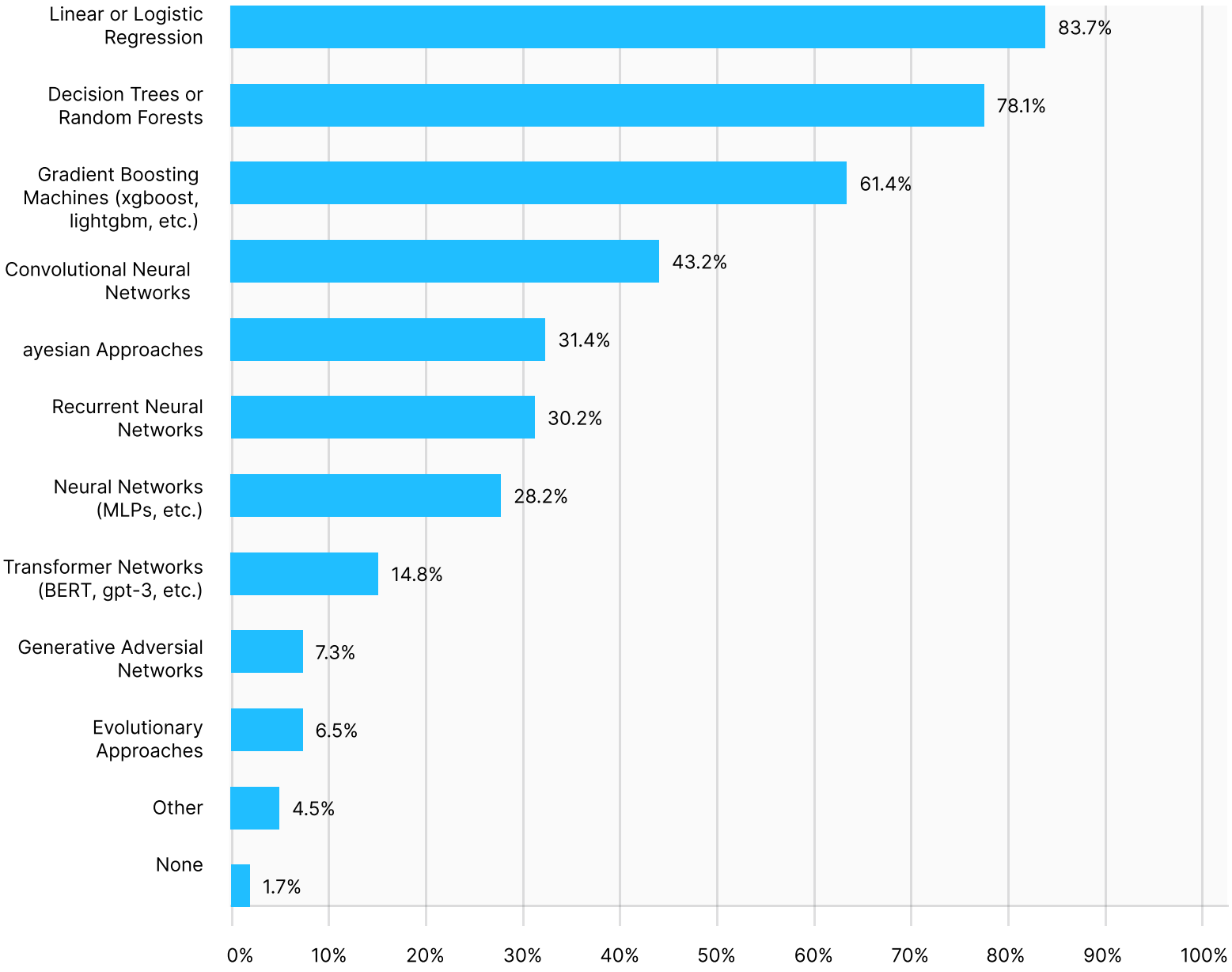


Methods & Algorithms

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

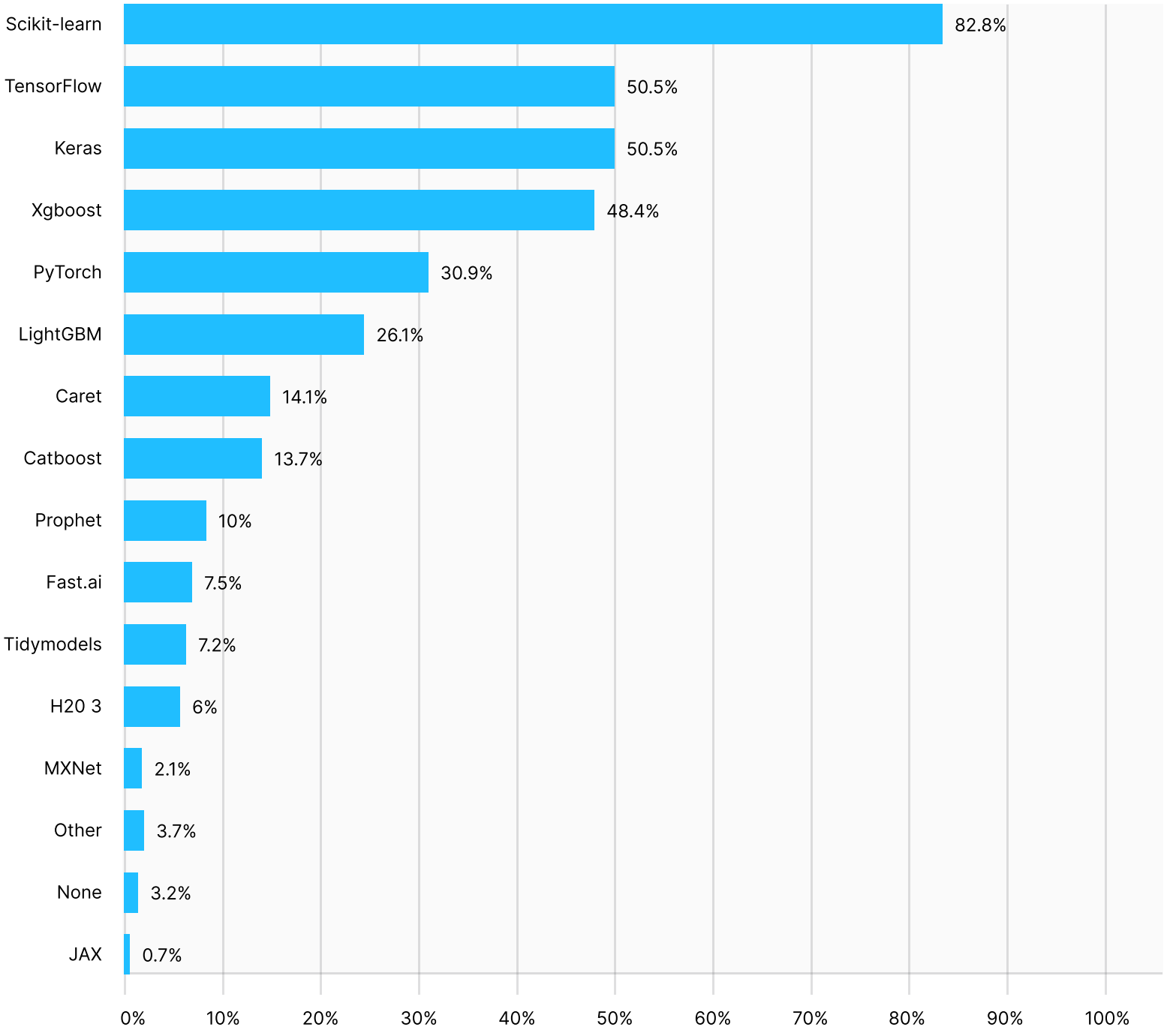


Python-based tools continue to dominate the machine learning frameworks. Scikit-learn, a swiss army knife applicable to most projects, is the top with four in five data scientists using it. TensorFlow and Keras, notably used in combination for deep learning, were each selected on about half of the data scientist surveys. Gradient boosting library xgboost is fourth, with about the same usage as 2019.

The fifth place tool, PyTorch, climbed above 30%, up from about 26% in 2019.

The most popular of the tools added to the survey this year is R-based Tidymodels, reaching over 7 percent.

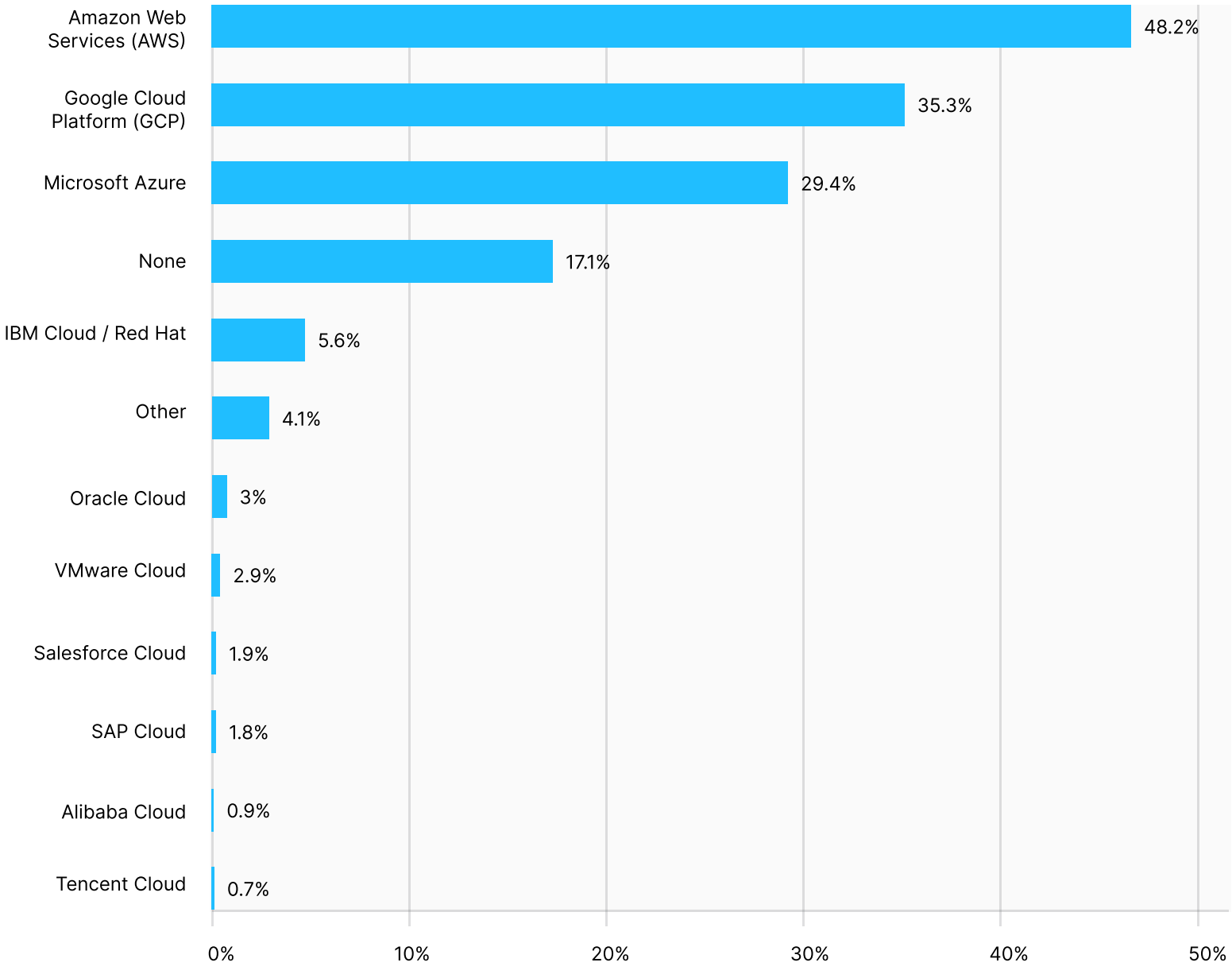
MACHINE LEARNING FRAMEWORK USAGE



Enterprise Cloud Computing

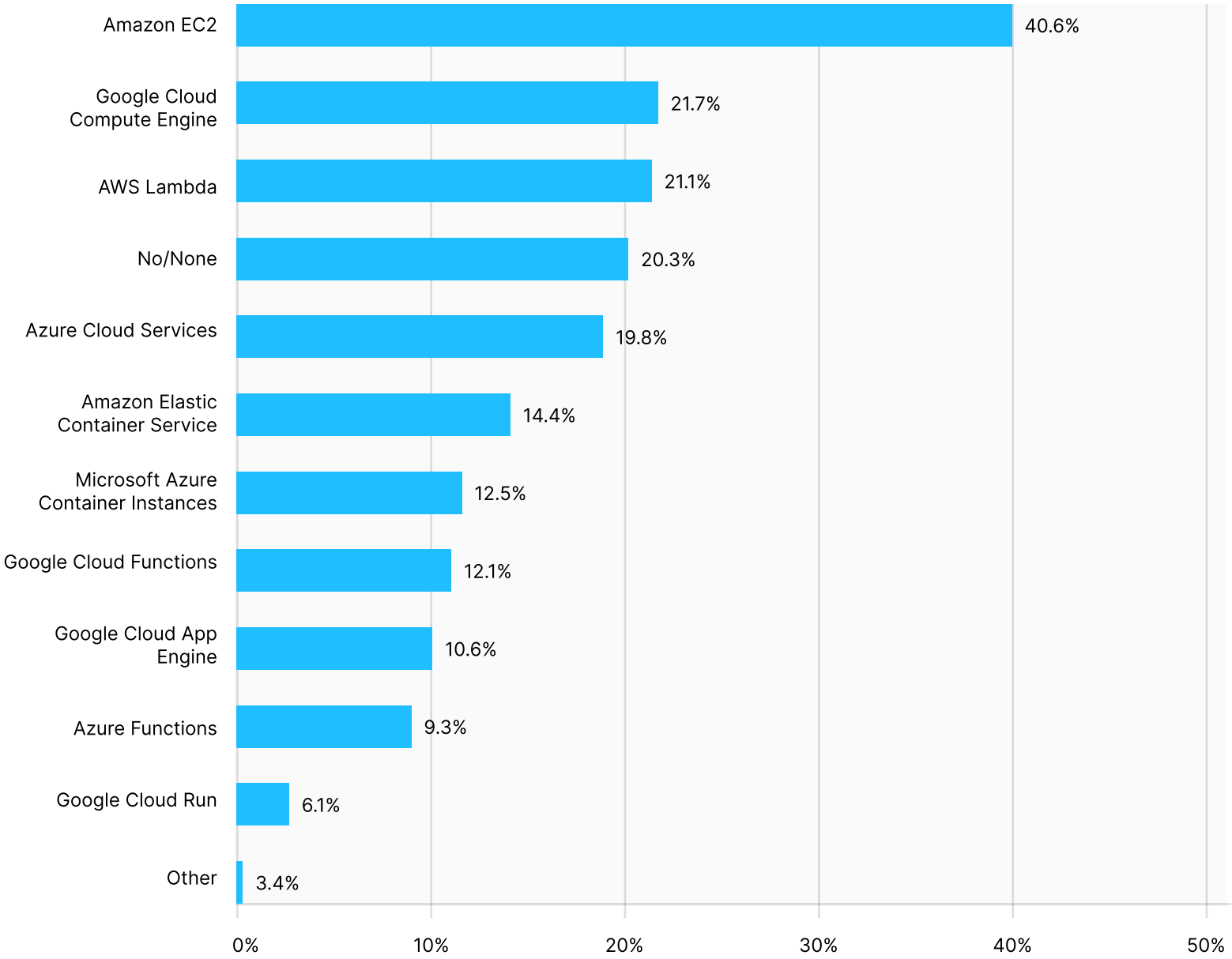
There are clearly three big players in cloud computing, and it's no surprise who: Amazon Web Services, Google Cloud Platform, and Microsoft Azure. Notably, more data scientists are using the cloud overall. In 2019, about 25% had not adopted cloud computing, which decreased to 17% in this year's survey.

ENTERPRISE CLOUD USAGE



Those who use cloud services were asked about specific products. Compute servers are the most common products, followed by serverless technologies. One in five did not name a cloud product.

ENTERPRISE CLOUD PRODUCT USAGE

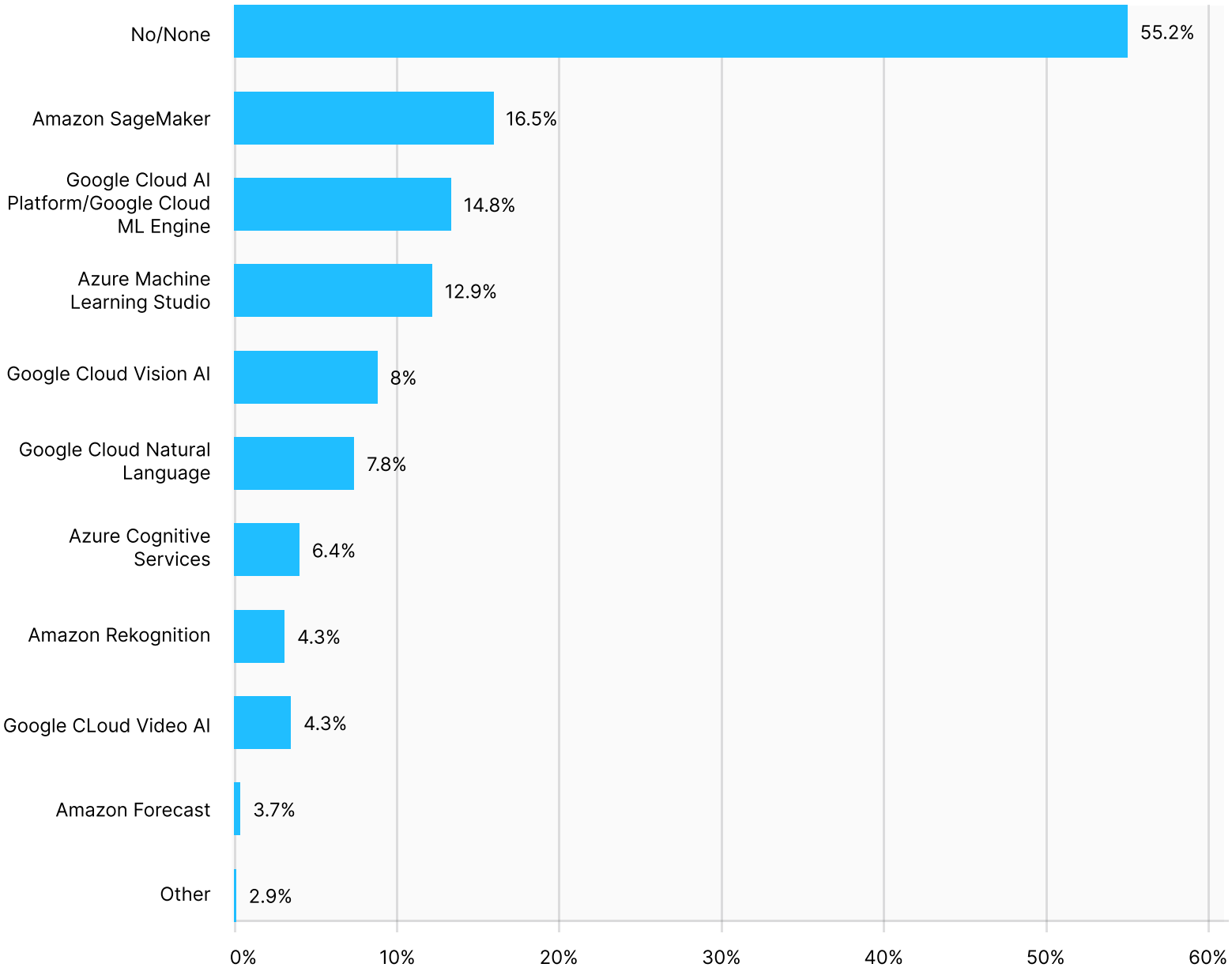


Enterprise Machine Learning Tools

Those who use AWS, Google Cloud Platform, or Microsoft Azure were asked about machine learning (ML) tools in particular. Over half of these data scientists do not use ML in the cloud.

Of those with ML usage, Amazon SageMaker was the most popular answer, followed closely by Google Cloud AI and ML.

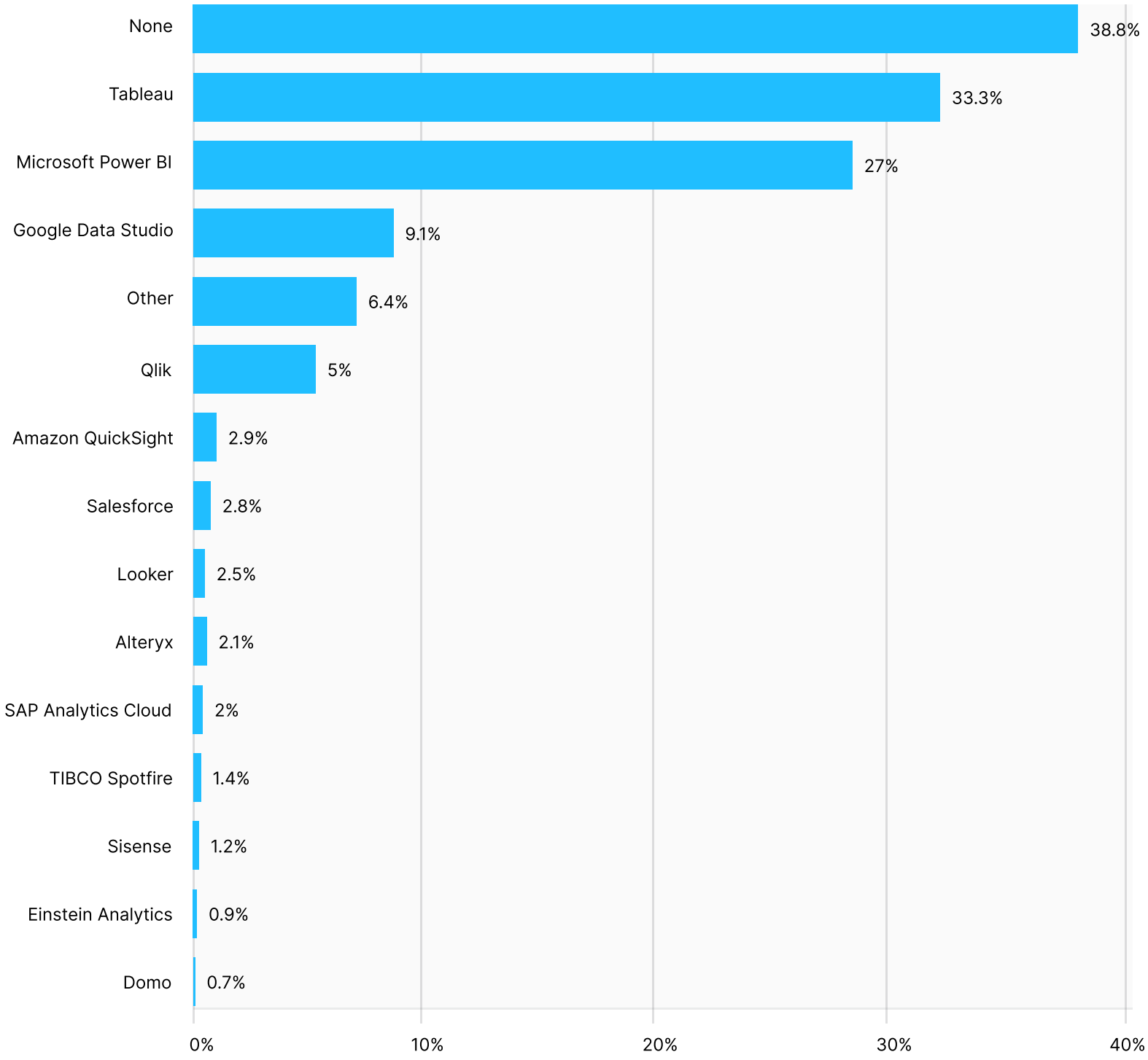
ENTERPRISE MACHINE LEARNING PRODUCT USAGE



Enterprise Big Data

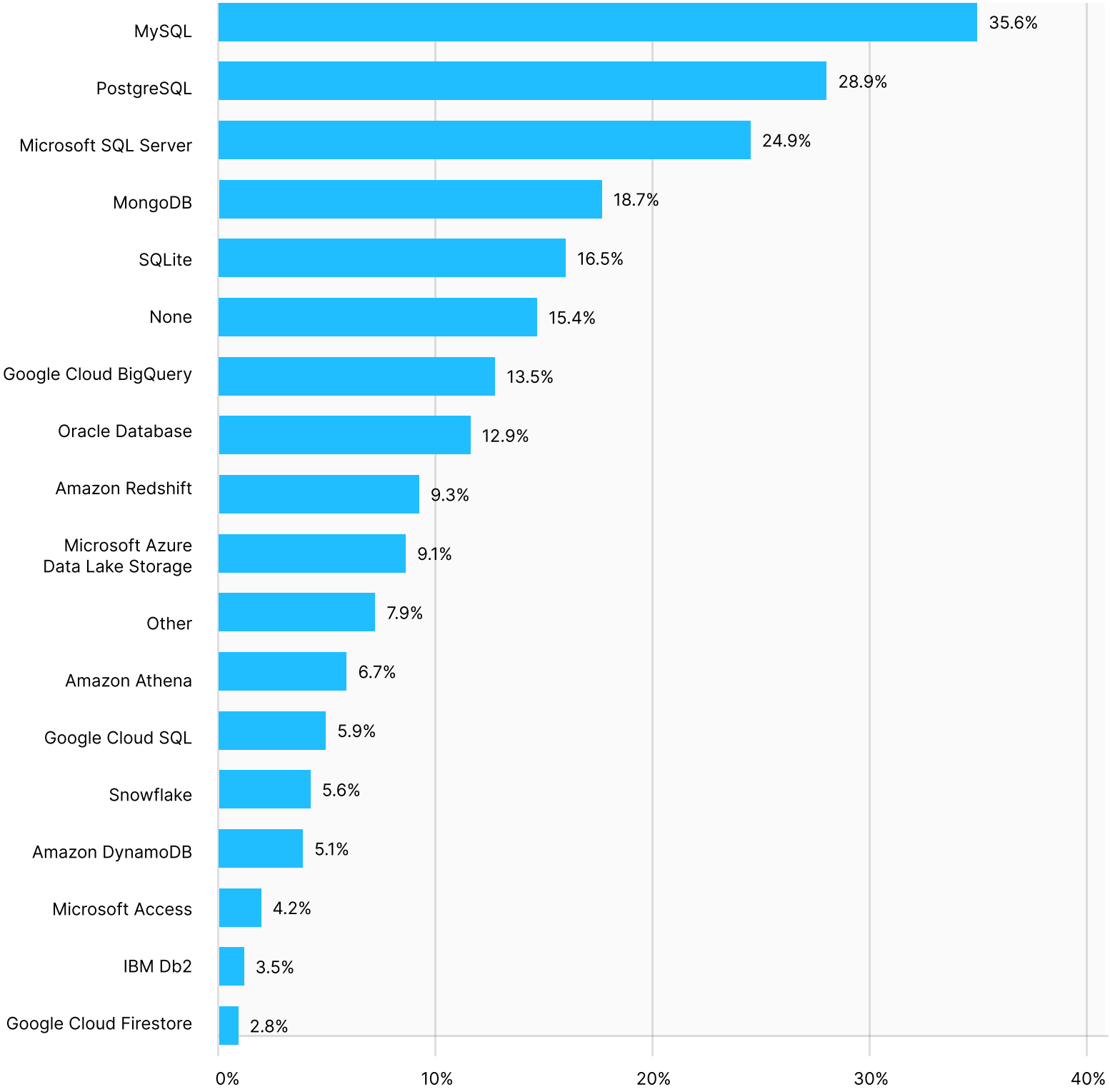
Business Intelligence tools help data scientists visualize their data, but four in 10 do not use one. The majority do employ BI, with Tableau as the most popular tool. Microsoft Power BI and Google Data Studio round out the top three.

DATA SCIENTIST USAGE OF BUSINESS INTELLIGENCE TOOLS



Regarding databases, there isn't a clear favorite among data scientists. MySQL was mentioned most often (35.6%), followed by PostgreSQL (28.86%) SQL Server (24.93%).

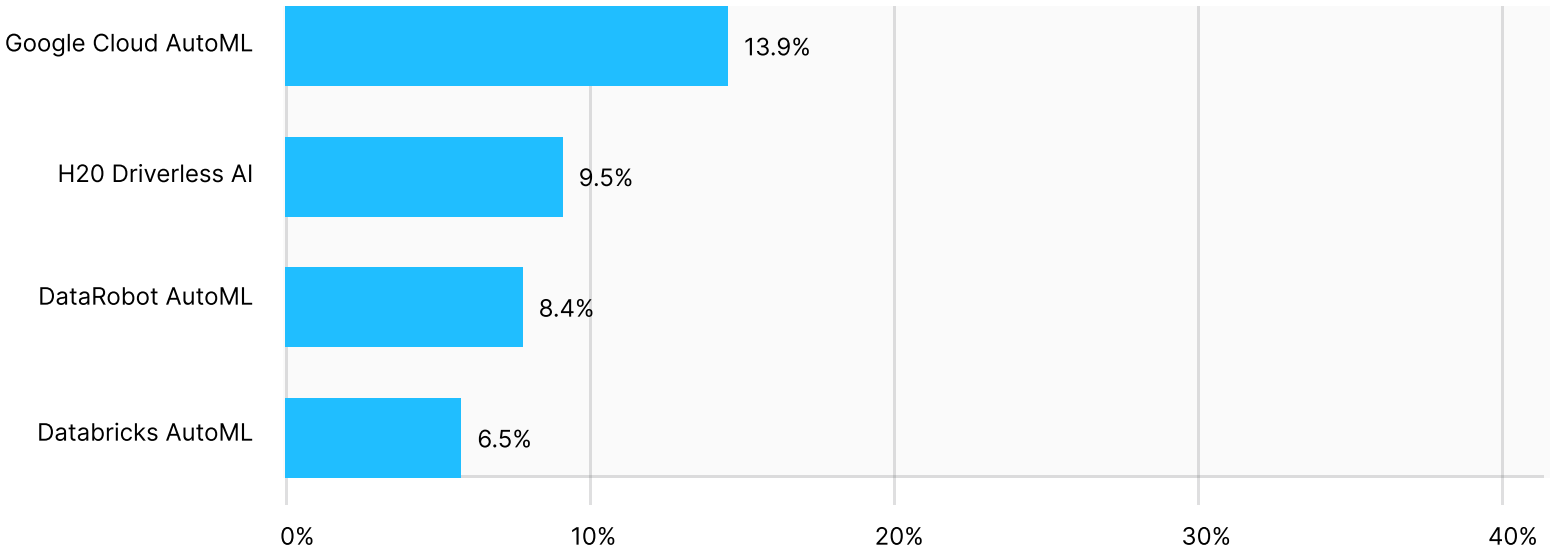
DATABASE USAGE BY DATA SCIENTISTS



Automated Machine Learning

As with machine learning overall, many data scientists (33%) do not use auto ML tools. Google Cloud AutoML saw gains from last year's survey, nearly 14% versus 6% in 2019.

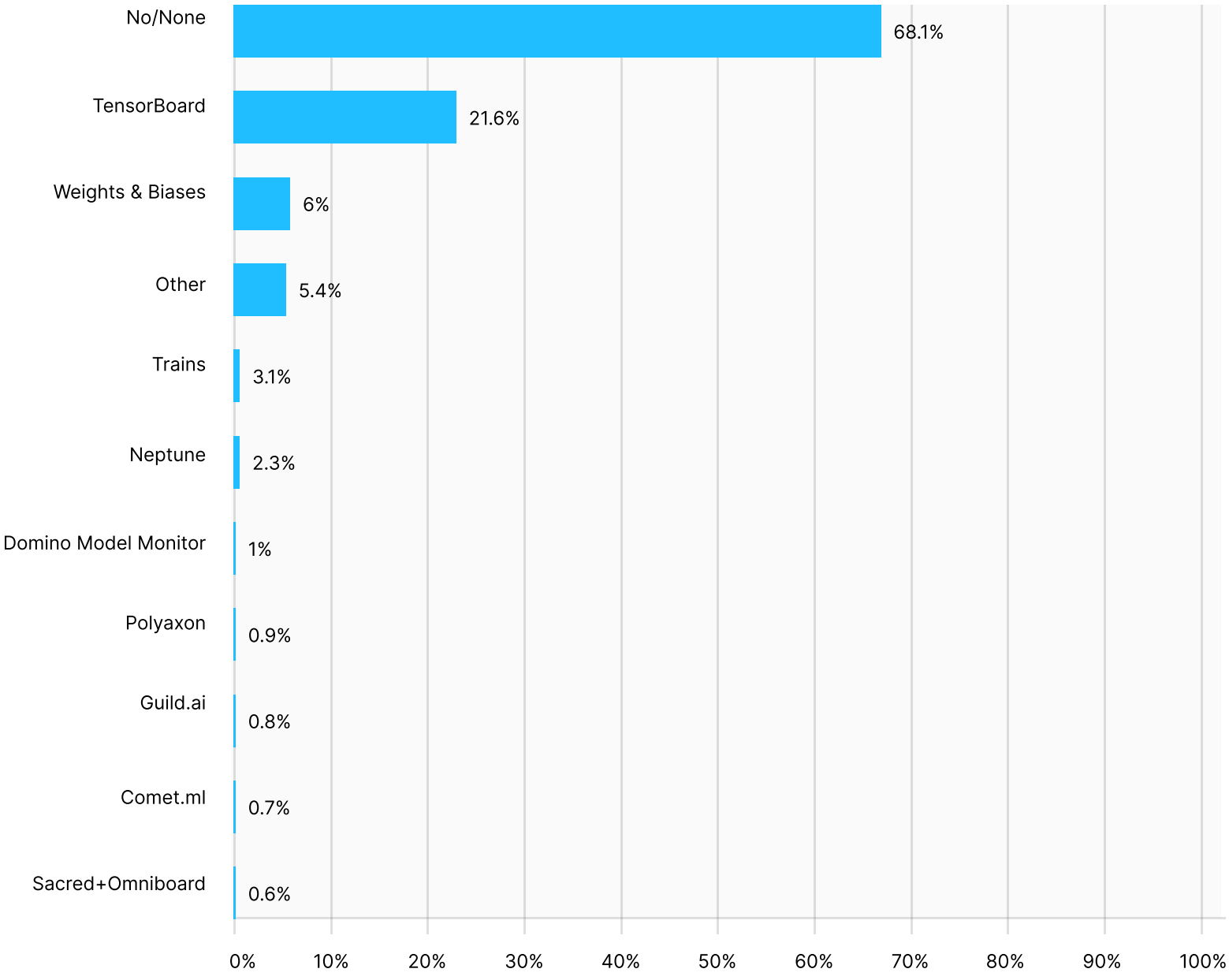
AUTOMATED MACHINE LEARNING FRAMEWORK USAGE



Machine Learning Experiments

Among data scientists who use tools to manage machine learning experiments, TensorBoard is a clear favorite (over 21%). The closest competitor is Weights & Biases, with 6%. However, the vast majority (68%) of data scientists do not use special tools to keep track of and manage their ML experiments.

USAGE OF MACHINE LEARNING EXPERIMENT TOOLS

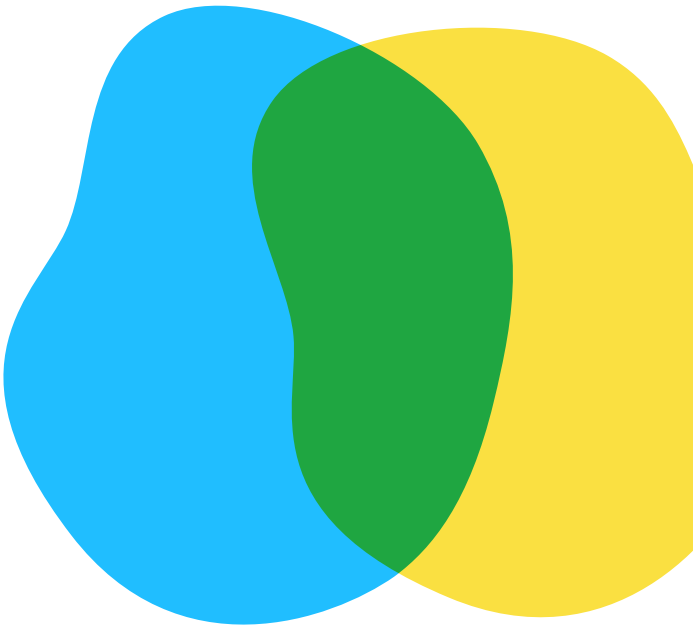


Conclusion

This 2020 edition of the State of Machine Learning and Data Science includes insights gathered from a survey of 20,036 Kaggle members. Their answers covered demographic, education, employment, and technology usage.

The charts and results are culled from professional data scientists (covering 13% of respondents). There's even more to uncover in the most comprehensive dataset available on the state of machine learning and data science today.

Kaggle has published the complete dataset of responses for the community to review, and we'll run a competition from November 18, 2020 to January 6, 2021 to learn even more about data science practitioners in 2020.




kaggle™