JET
BRAINS

2018

Data Science Survey

In spring 2018, we polled over 1,600 people involved in Data Science and based in the US, Europe, Japan, and China, in order to gain insight into how this industry sector is evolving. Here's what we learned.

Methodology

We distributed the survey via targeted ads on Facebook, Twitter, and LinkedIn. We screened respondents by excluding those who replied “I am not involved in data analysis.” We collected 400 complete and valid responses from the US, Japan, and China. To represent Europe, we used quotas for select European countries to collect a set of responses which also totaled 400.

Some bias is likely present as JetBrains users may have been more willing on average to complete the survey.

Key Takeaways

Most popular language

Python is currently the most popular language among data scientists.

Primary language

Most people assume that Python will remain the primary programming language in the field for the next 5 years.

R, Keras and Tableau

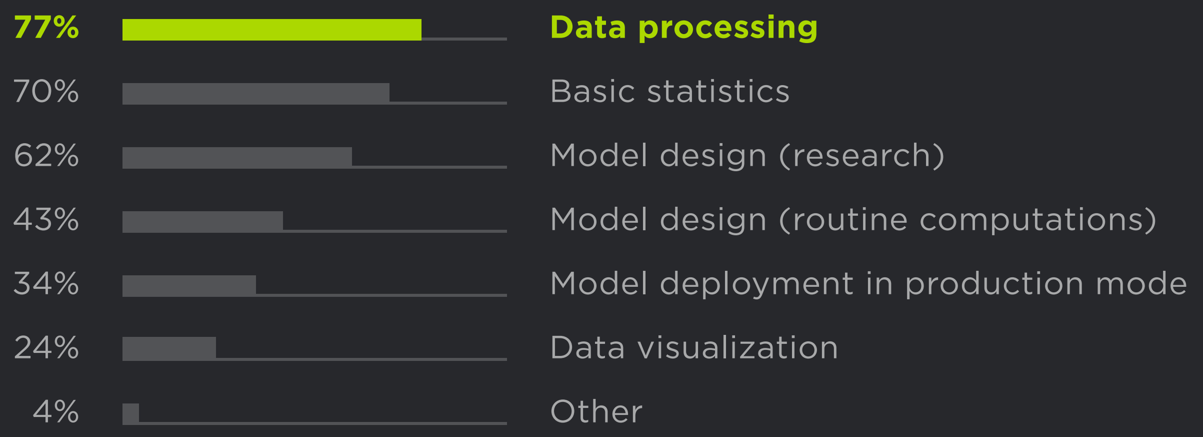
Data Science professionals tend to use R, Keras, and Tableau, while amateur data scientists are more likely to prefer Microsoft Azure ML.

1

Tasks

Which of the following are you involved in?

Number of answers: 1282



This question was only answered by respondents who are professionally involved in data analysis.

Natalia Vassilieva
Head of Software and AI,
Hewlett Packard Labs

According to the very first table, a relatively small percentage of the respondents work on model deployment in production mode. This correlates with multiple market research studies which report that the majority of enterprises are just starting to explore machine learning and deep learning.

They have small teams working on PoCs*, and model deployment in production still needs to be addressed. But I expect this type of activity to become more and more visible within the next few years when more and more businesses will proceed from PoCs to production deployments.

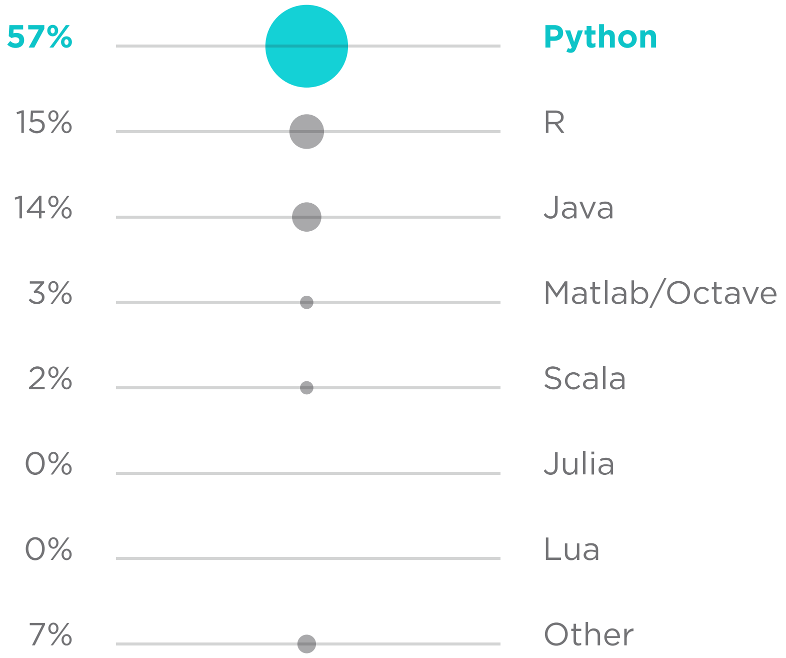
*PoCs — Proof of concepts

2

Programming Languages and Tools

Main programming language for data analysis

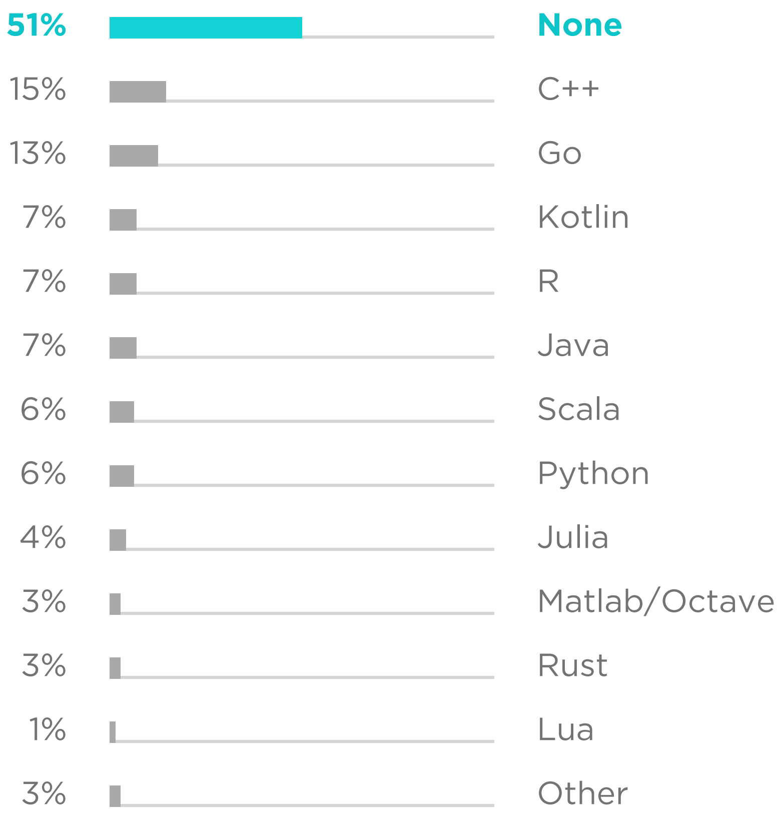
Number of answers: 1522



Do you plan to adopt or migrate to other languages in the next 12 months?

Note this was a question with checkboxes. Shares may total more than 100%.

Number of answers: 1522



15% of data scientists are going to adopt or migrate to C++ in the next 12 months. This is probably due to performance issues.

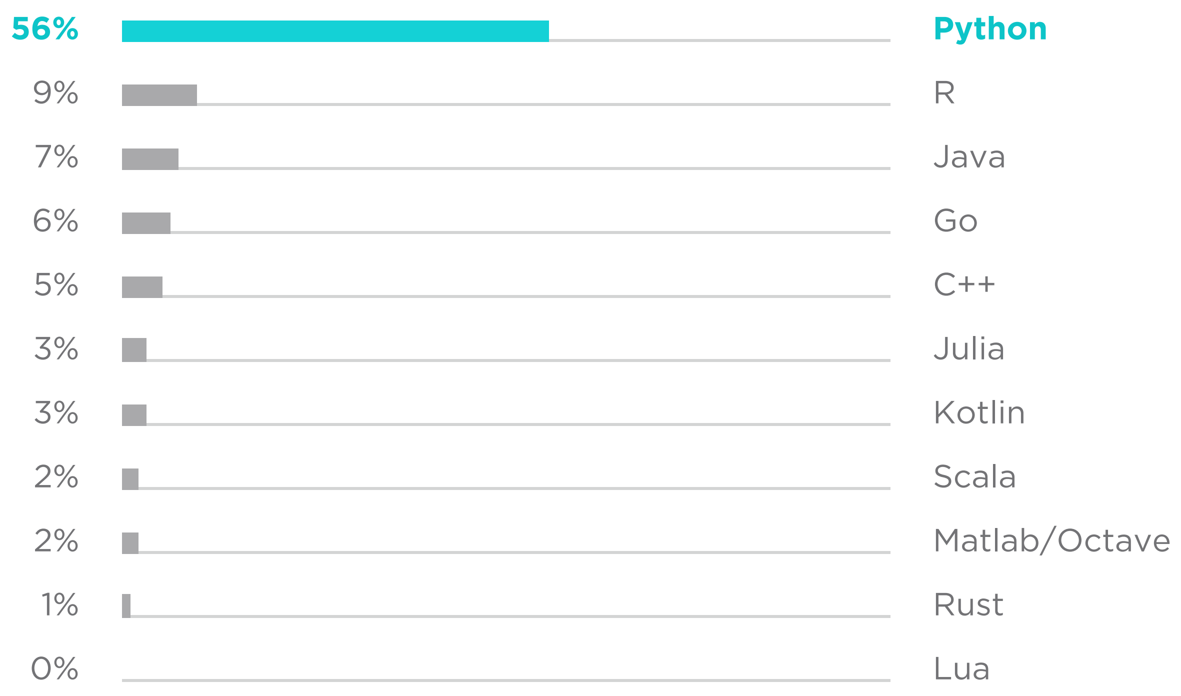
7% of respondents plan to start using the Kotlin language in the next 12 months.

In your opinion, what programming language will be most used for data analysis in the next 5 years?

Number of answers: 1522

Note this was a question with checkboxes. Shares may total more than 100%.

Most respondents believe that Python will remain on top for the next 5 years.



Note that Go, C++, Julia, Kotlin, Rust, and Lua were unavailable as answer choices for regularly and most used languages.

Overall, people tend to choose the language they use. Of those who don't use a language they think will dominate, most want to start using it. Half of those who believe Kotlin will dominate are planning to adopt it in the nearest future.

Natalia Vassilieva
Head of Software and AI, Hewlett Packard Labs

No surprises with the programming languages. Traditional data scientists are some of the most likely to still use R, there are plenty of statistics libraries for R. The new generation of data scientists are choosing Python.

When it comes to high-performance data analytics, I'd expect to see C/C++ in the picture. Currently, we are observing that many HPC techniques and tools are being adopted and re-used for high-performance data analytics and deep learning.

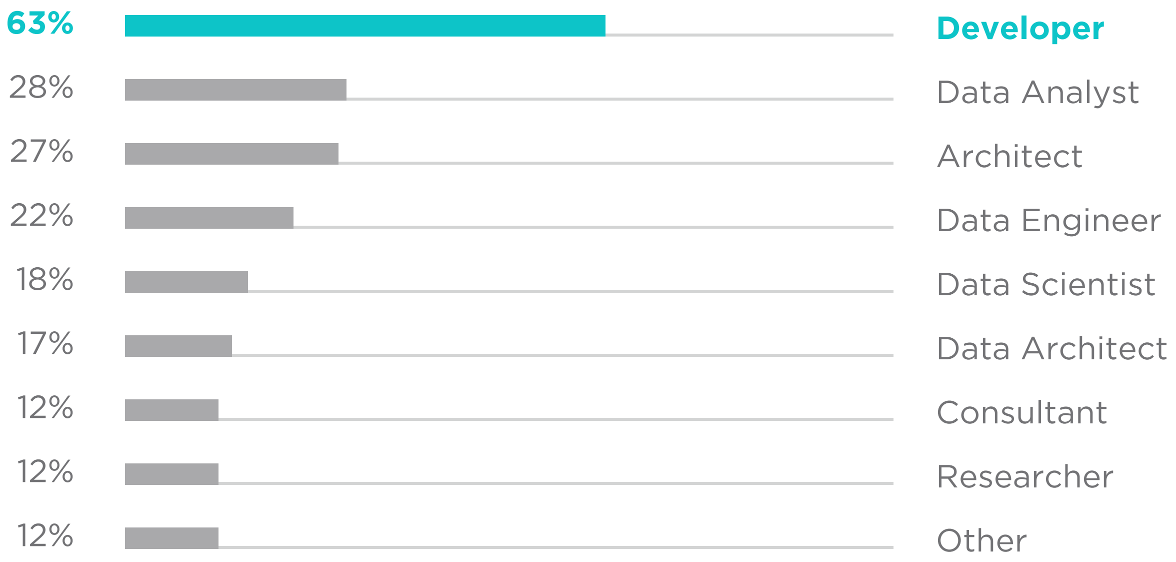
Kotlin adopters

7% of data analysts using a programming language want to adopt Kotlin for Data Science in the nearest future.

Note this was a question with checkboxes.
Shares may total more than 100%.

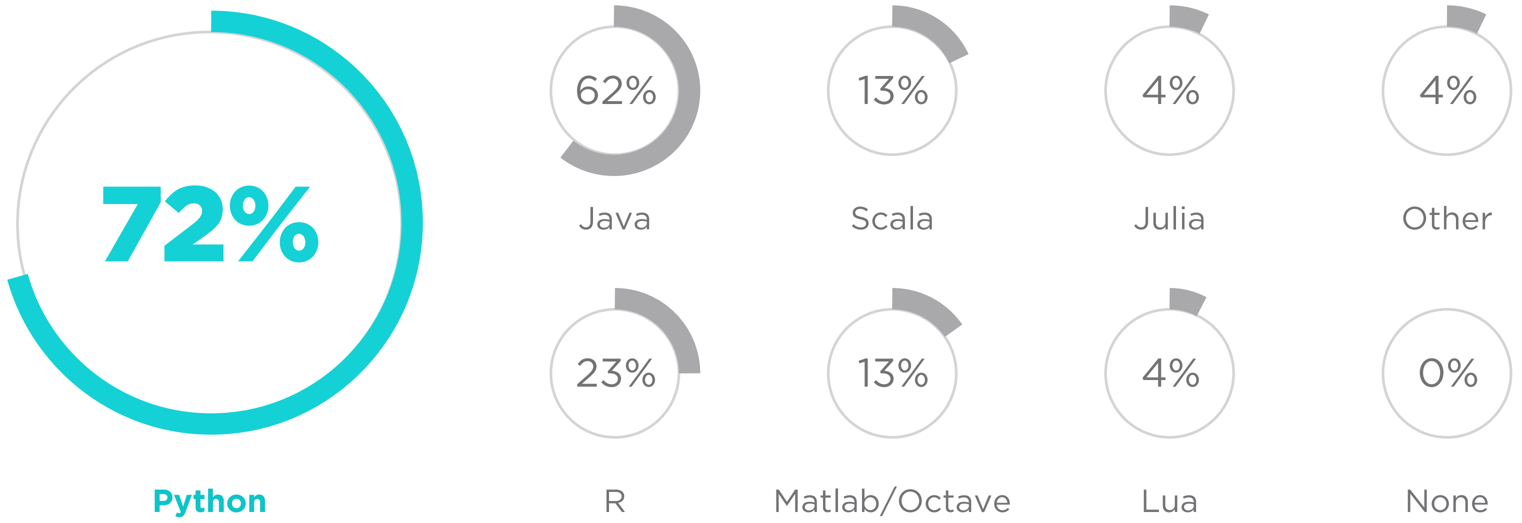
Which of the following best describes your job roles regardless of your position level?

Number of answers: 60



What programming languages do you regularly use for data analysis, if any?

Number of answers: 112



Kotlin Learning

Vitaly Khudobakhshov
Product Manager, JetBrains

Kotlin is a general-purpose language running on the Java virtual machine. It is concise and easily integrates with popular data processing frameworks such as Hadoop and Spark.

Kotlin is statically typed and uses type inference that increases its reliability. These features all make Kotlin a handy instrument for data engineering and data science.

Thomas Nield has assembled a helpful collection of Kotlin resources for data science on his Github.

If you are new to Kotlin and are considering it for your next language, start from learning the basic syntax.

If you are already familiar with Java, you may want to have a play with Kotlin Koans.

3

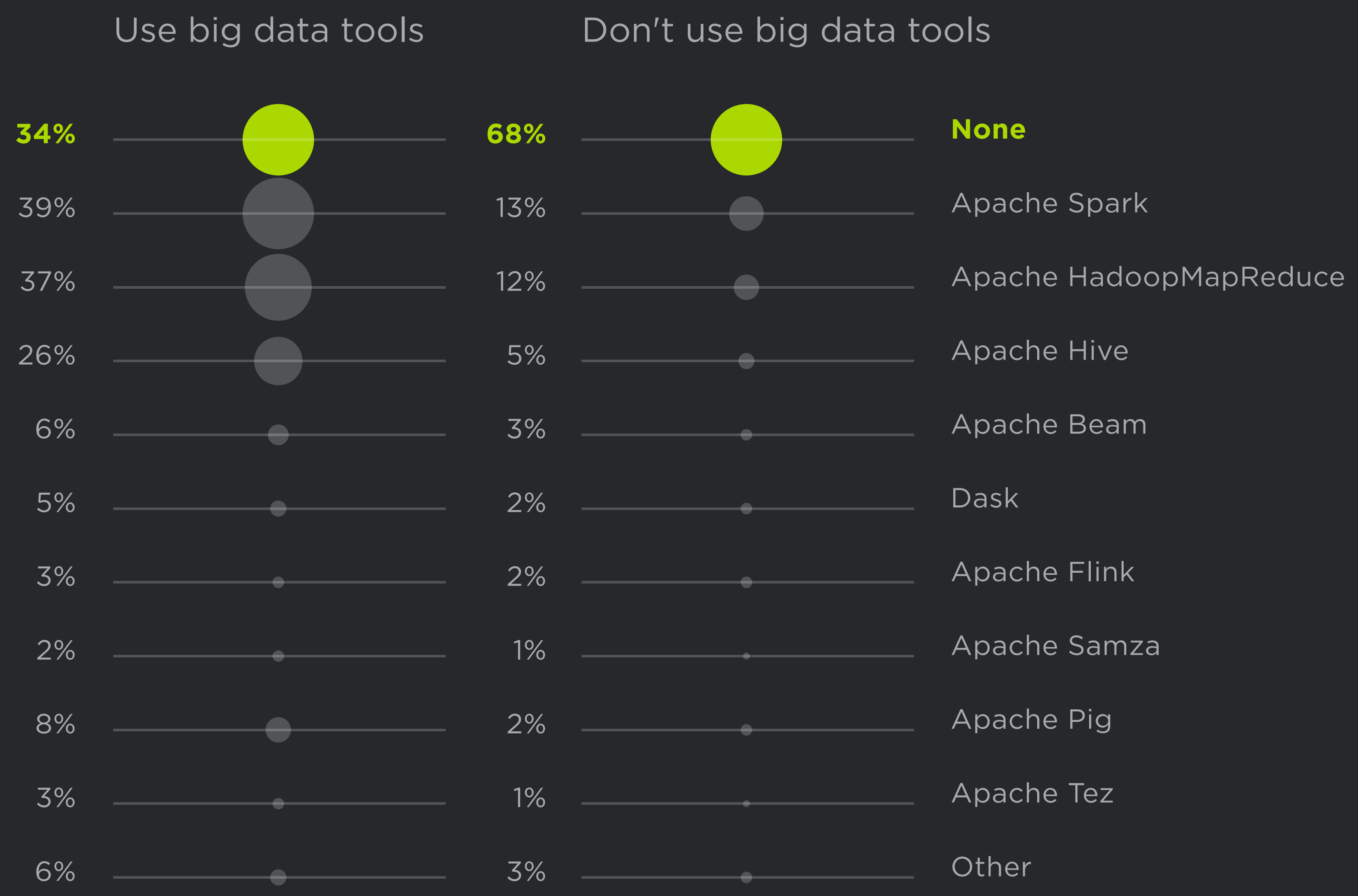
Tools & Technologies & Editors

Big Data tools

Number of answers: 1477

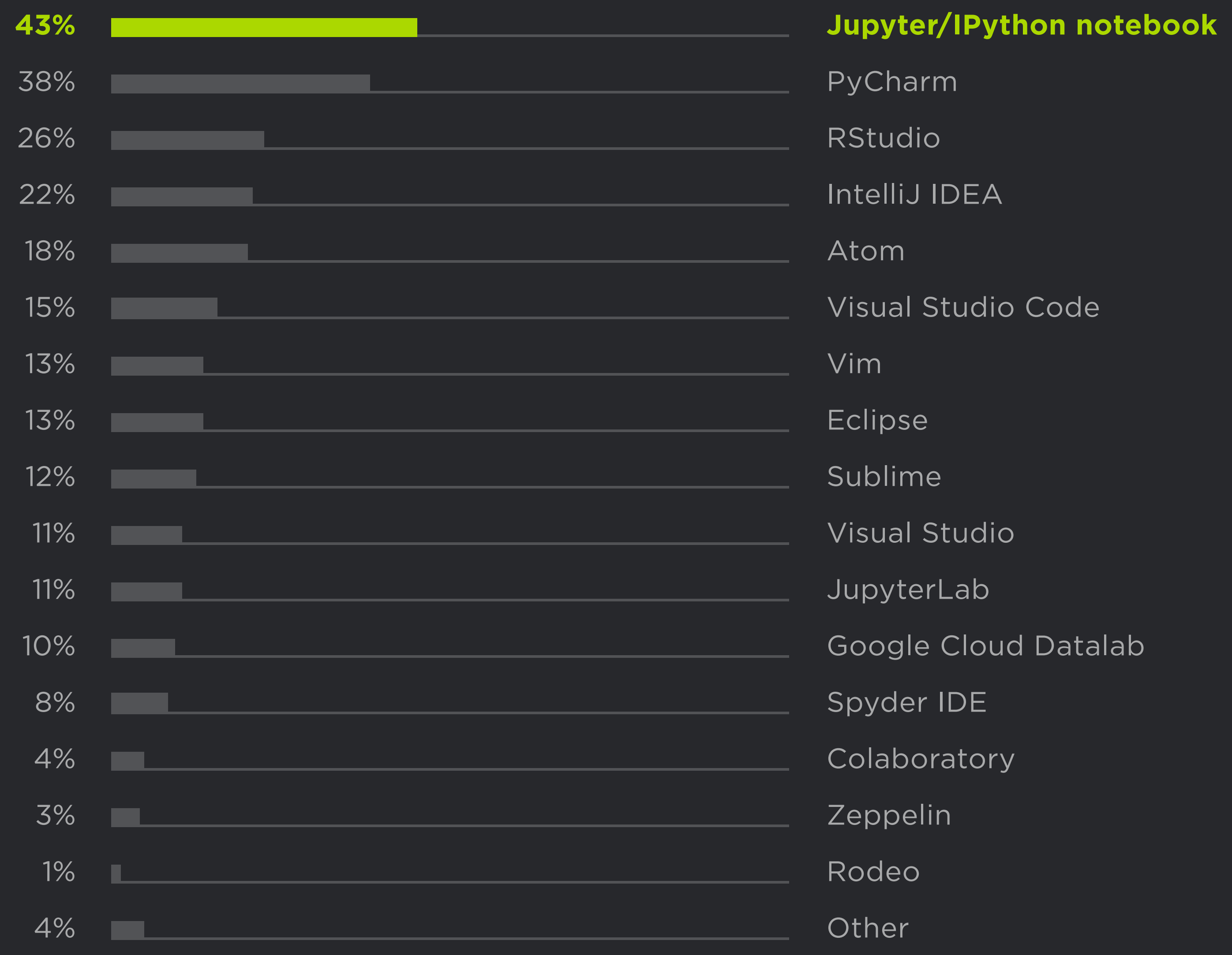
Note this was a question with checkboxes. Shares may total more than 100%.

One third of those who say they work with big data don’t use any big data tools. Conversely, a third of those who do NOT work with Big data DO use some big data tools. Still, this self-identification does correlate with formal factors.



IDEs and Editors

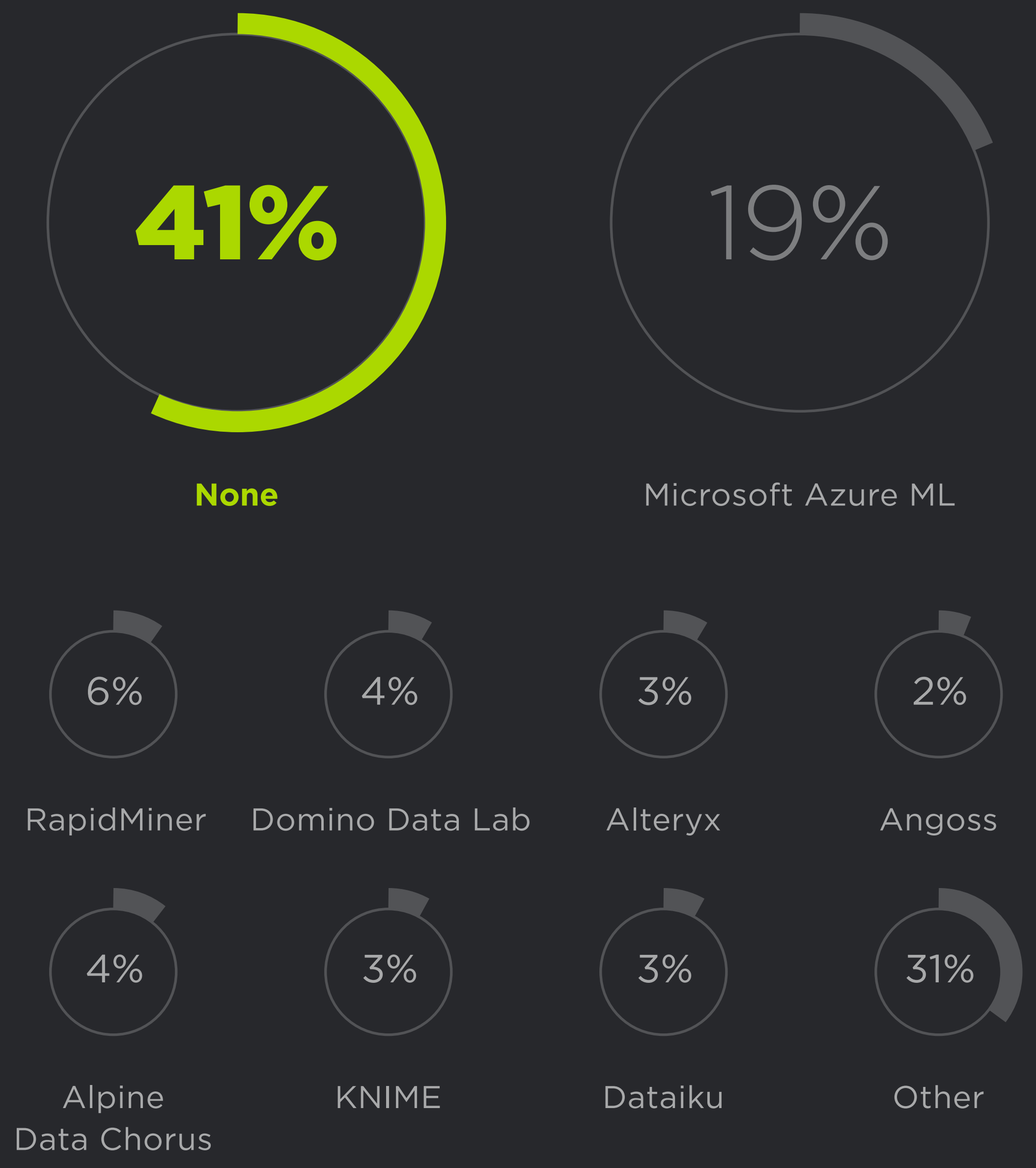
Number of answers: 1522



Note this was a question with checkboxes. Shares may total more than 100%.

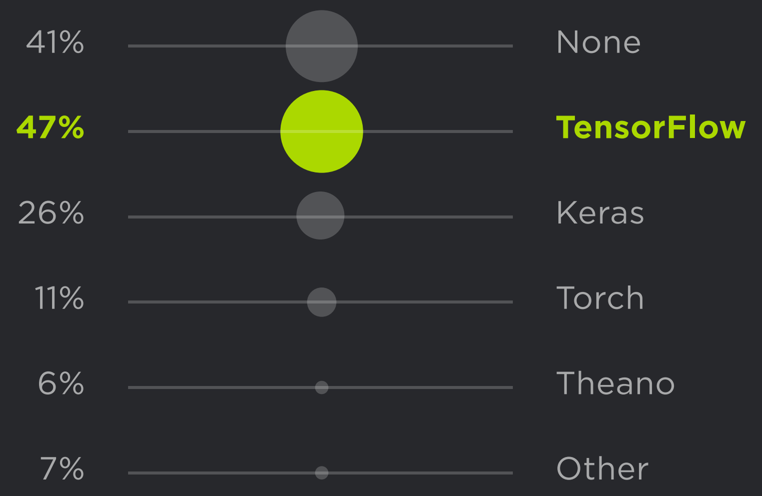
Which tools do you use for data analysis, if any?

Number of answers: 1666



What deep learning libraries do you use, if any?

Number of answers: 1666

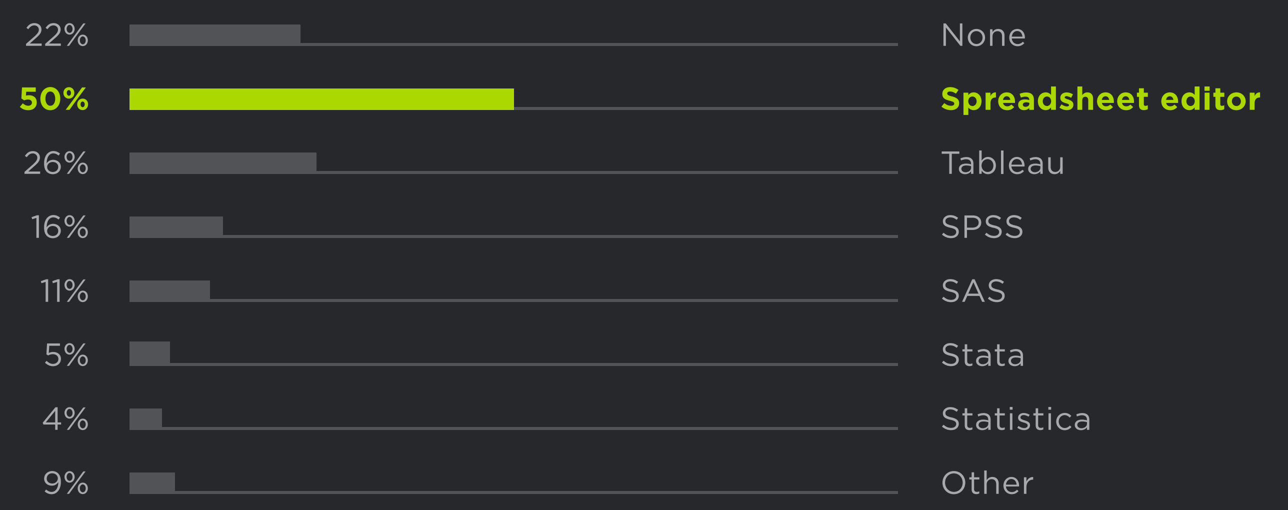


Note this was a question with checkboxes.
Shares may total more than 100%.

Apparently, if you do deep learning, you do it with TensorFlow. Nearly 80% of those using deep learning libraries use TensorFlow, and almost all Keras users use it alongside TensorFlow.

Which statistics packages do you use to analyze and visualize data, if any?

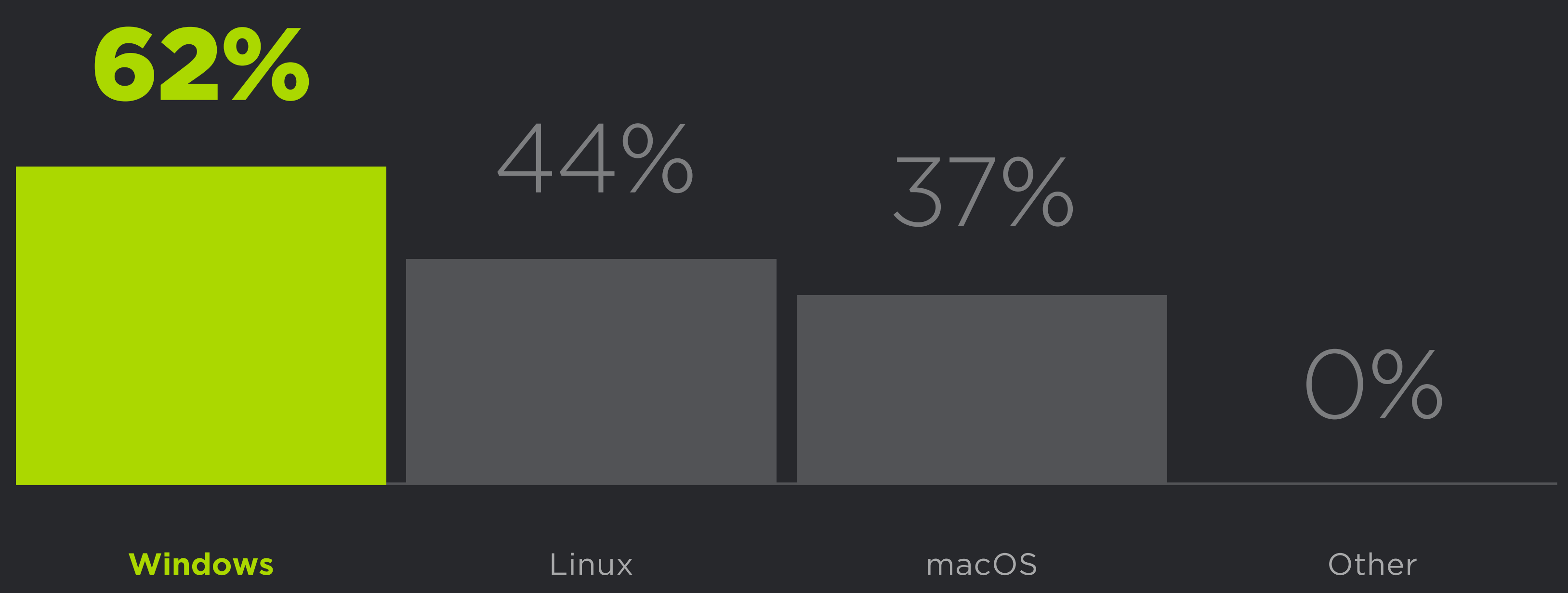
Number of answers: 1666



Note this was a question with checkboxes.
Shares may total more than 100%.

What operating systems do you use as your work environment for data analysis?

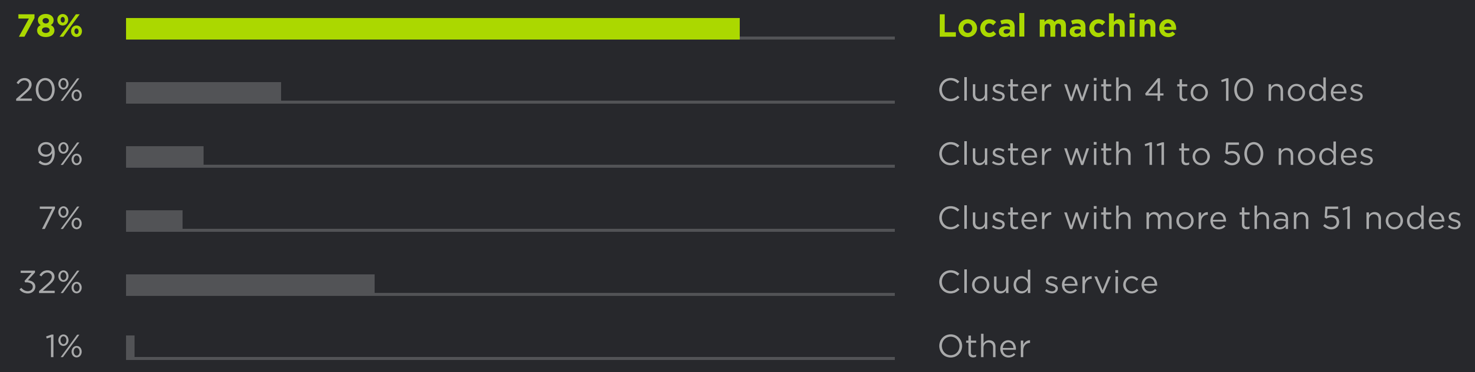
Number of answers: 1666



Note this was a question with checkboxes.
Shares may total more than 100%.

What do you use to perform computations?

Number of answers: 1666



Note this was a question with checkboxes. Shares may total more than 100%.

78% data science specialists perform computations on local machines.

Cloud services

Number of answers: 527

56%
Amazon Web Services (AWS)

41%
Google Cloud Platform

28%
Microsoft Azure

14%
Other

4

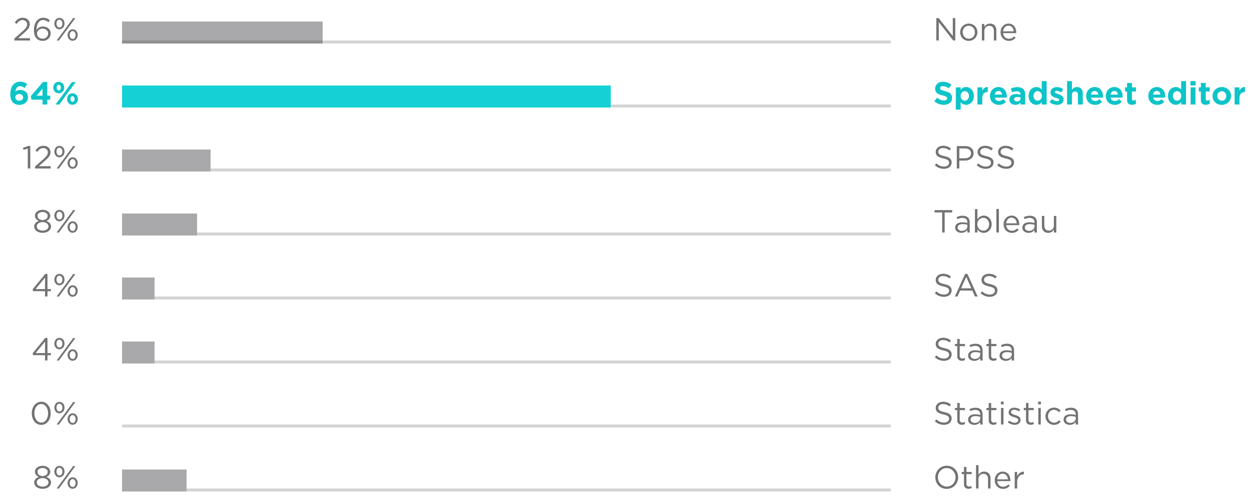
Non-programmers

We received 77 responses from people who don’t use any programming languages and aren’t about to adopt any (5% of all data analysts who responded).

These respondents use spreadsheet editors more often than average, and most of them work in non-IT industries. They also tend to use data analysis tools less often.

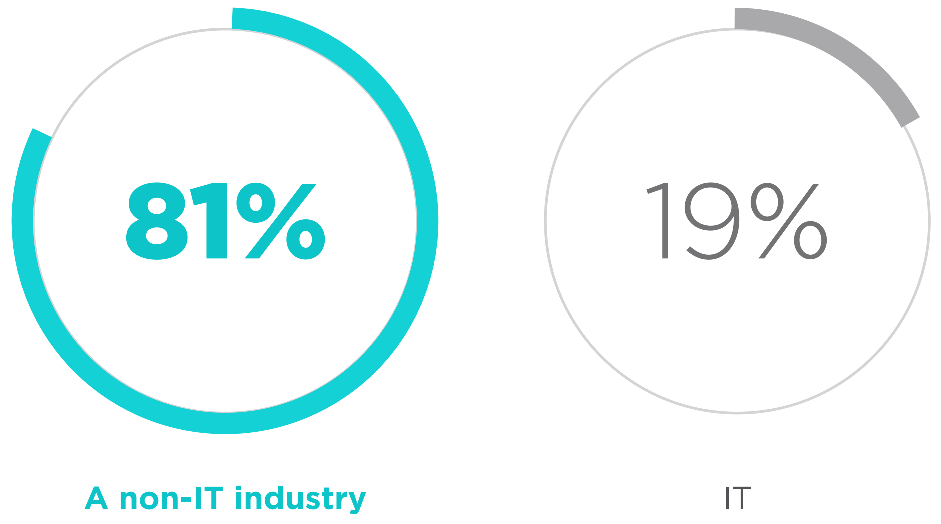
Which statistics packages do you use to analyze and visualize data, if any?

Number of answers: 77



What is the industry you primarily analyze data for?

Number of answers: 43

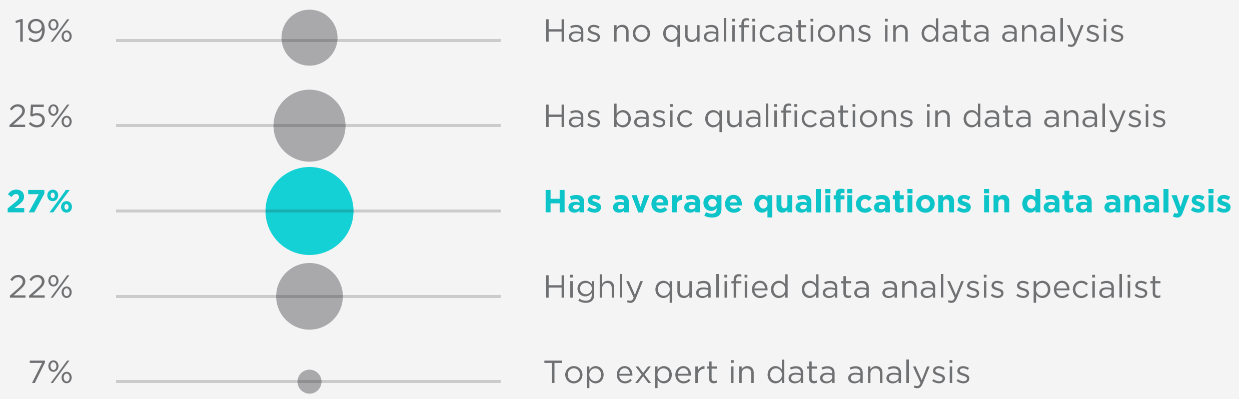


5

Manager’s expertise

What is your manager's level of expertise in data analysis?

Number of answers: 924



Almost half of all respondents report to managers who have little or no qualifications in data analysis.

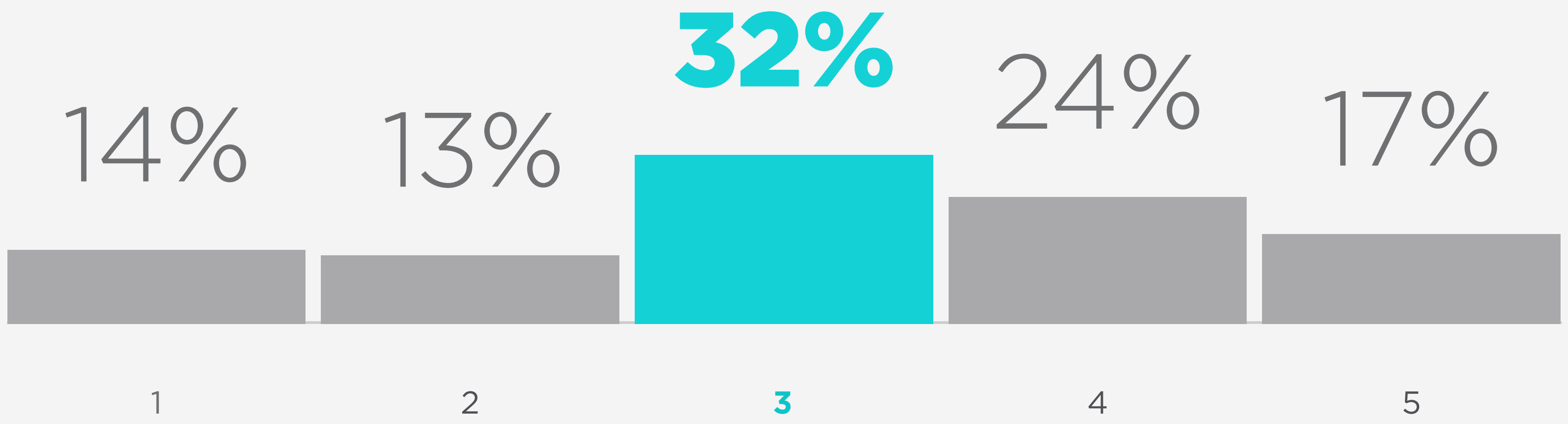
To what extent do you associate the following phrase with your manager?: "My manager gives me realistic assignments that are relevant to my skills and responsibilities, with a clear and specific description of the requirements."

Number of answers: 918

Answers:

1 = not at all

5 = a great deal



Correlation of Manager's expertise and assignment score

Number of answers: 924

	Score no qualifications N: 177	**Basic qualifications** N: 230	**Average qualifications** N: 246	**Highly qualified** N: 199	**Top expert** N: 66
1	**31%**	16%	7%	8%	11%
2	15%	20%	10%	9%	3%
3	29%	**34%**	**37%**	30%	17%
4	16%	18%	27%	**32%**	29%
5	10%	11%	18%	22%	**41%**

Answers:

1 = not at all

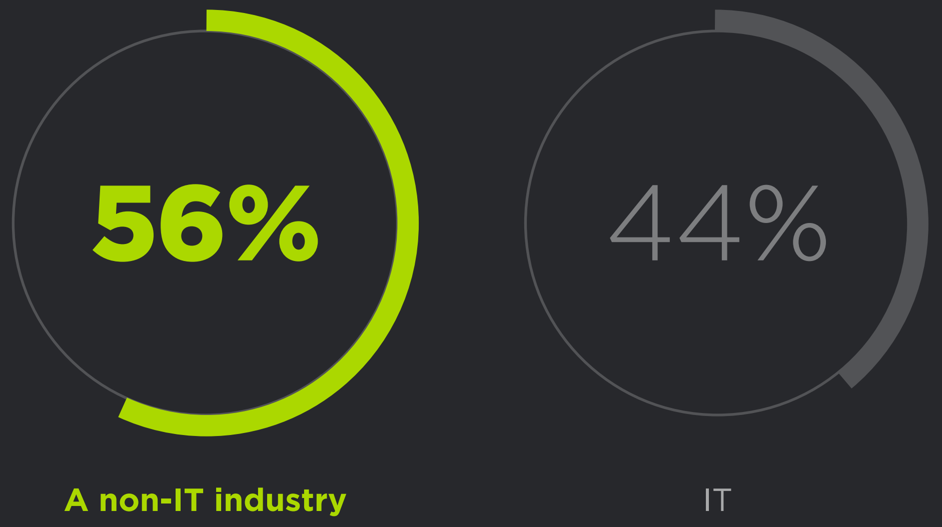
5 = a great deal

6

Industry & Demographic

What is the industry you primarily analyze data for?

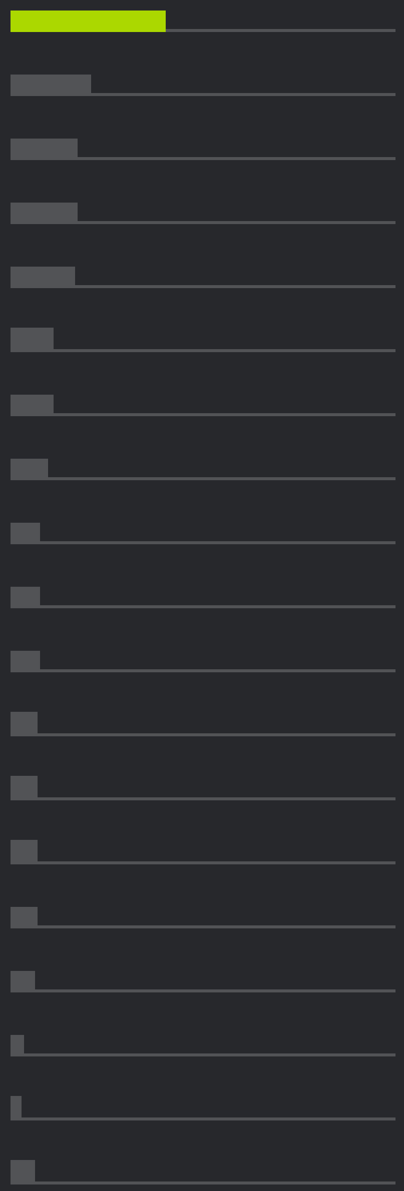
Number of answers: 1666



This was an optional question.

Which fields of IT do you primarily analyze data for?

Number of answers: 703

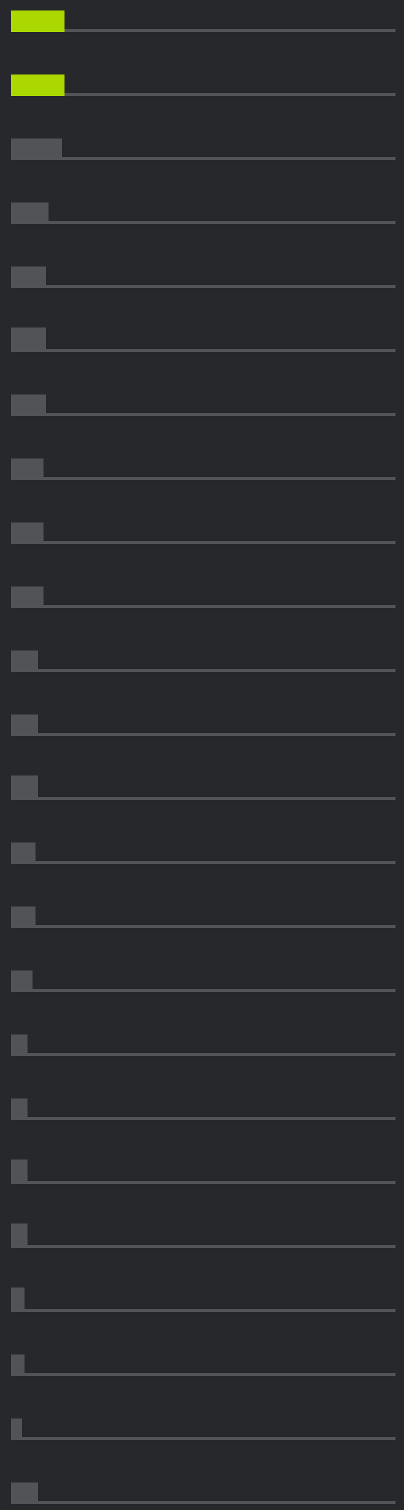


39%	**Big Data / Data analysis**
23%	E-learning
18%	Cybersecurity
18%	Cloud computing / platform
16%	IT services
12%	Data centers services
12%	Healthcare IT
10%	Software development tools
8%	IoT / embedded
8%	Mobile development
8%	Game development (including mobile)
7%	FinTech
7%	Hardware
7%	Internet / Search engines
7%	Bioinformatics
6%	BPO services
4%	Telecom
3%	Semiconductors
6%	Other

Note this was a question with checkboxes.
Shares may total more than 100%.

Which industry or industries do you primarily analyze data for?

Number of answers: 933



16%	**Accounting / Finance / Insurance**
16%	**Science**
15%	Education / Training
12%	Bioinformatics
11%	Logistics/ Transportation
11%	Administration / Management / Business Development
11%	Medicine / Health
10%	Machinery
10%	Business / Strategic Management
10%	Sales / Distribution / Retail
9%	Banking / Real Estate / Mortgage Financing
8%	Manufacturing
8%	Marketing
6%	Design
6%	Restaurants / Hospitality / Tourism
6%	Construction / Architecture
5%	Entertainment / Mass media and information / Publishing
5%	Human Resources
5%	Non-profit
5%	Customer Support
4%	Security
4%	Service / Maintenance
3%	Law
8%	Other

Demographics

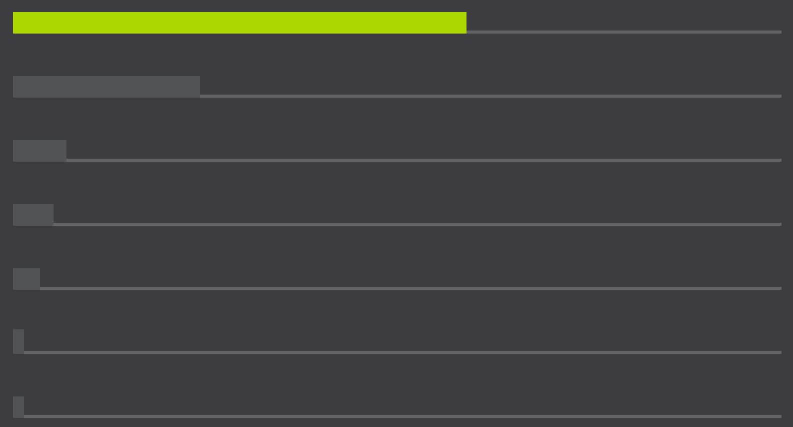
Age range

Number of answers: 1666

1%	7%	**48%**	28%	11%	4%	2%
17 or younger	18–20	**21–29**	30–39	40–49	50–59	60 or older

What is your main employment status?

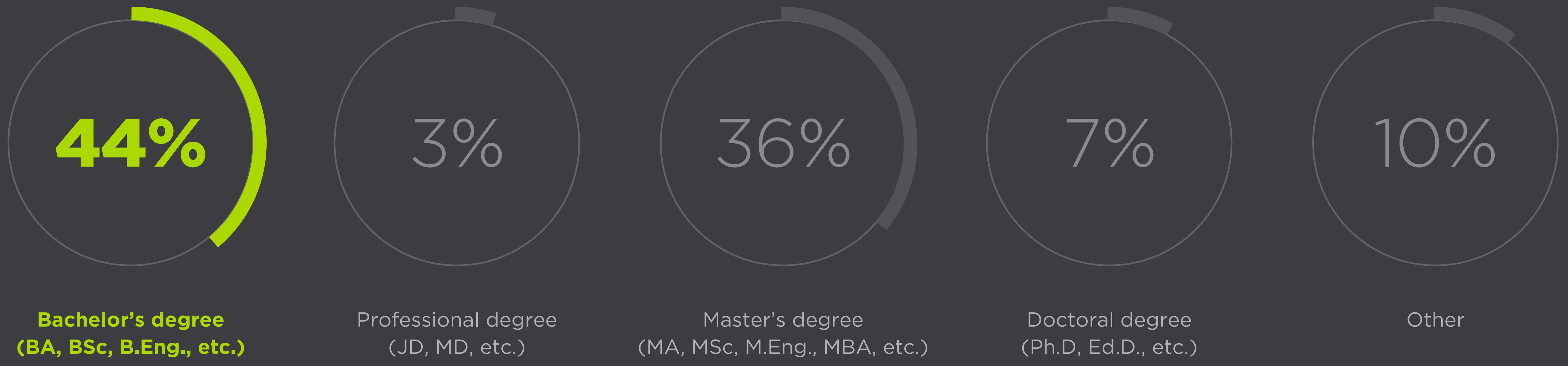
Number of answers: 1666



61%	**Fully employed by a company / organization**
25%	Student
5%	Freelancer
4%	Partially employed by a company / organization
3%	Looking for a job
1%	Retired
1%	Other

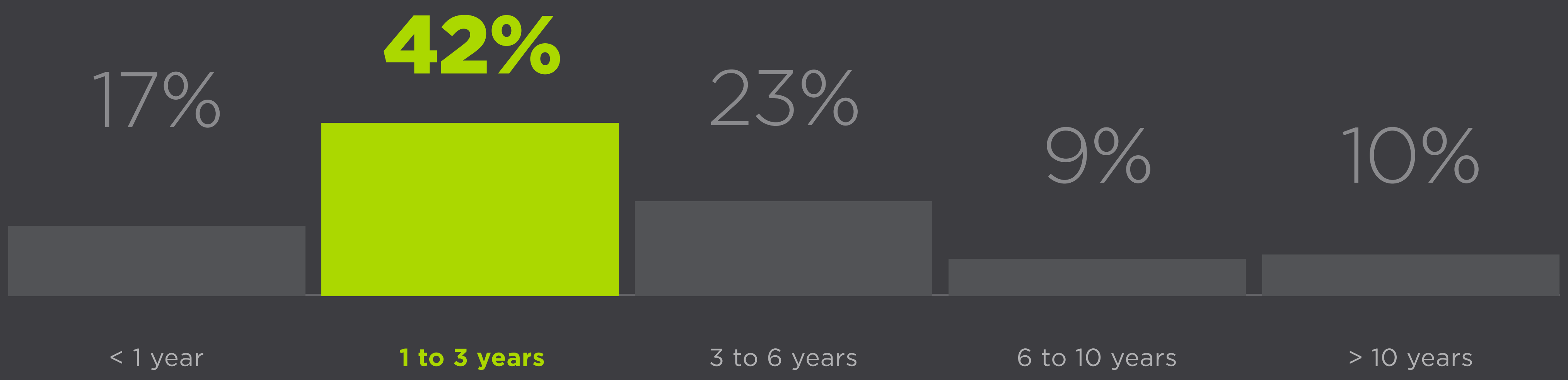
What is your main employment status?

Number of answers: 1666



Work experience

Number of answers: 924



This question was directed to professionals, that is, people professionally involved in data science or data analysis and working full-time or part-time.

Work environment and employment

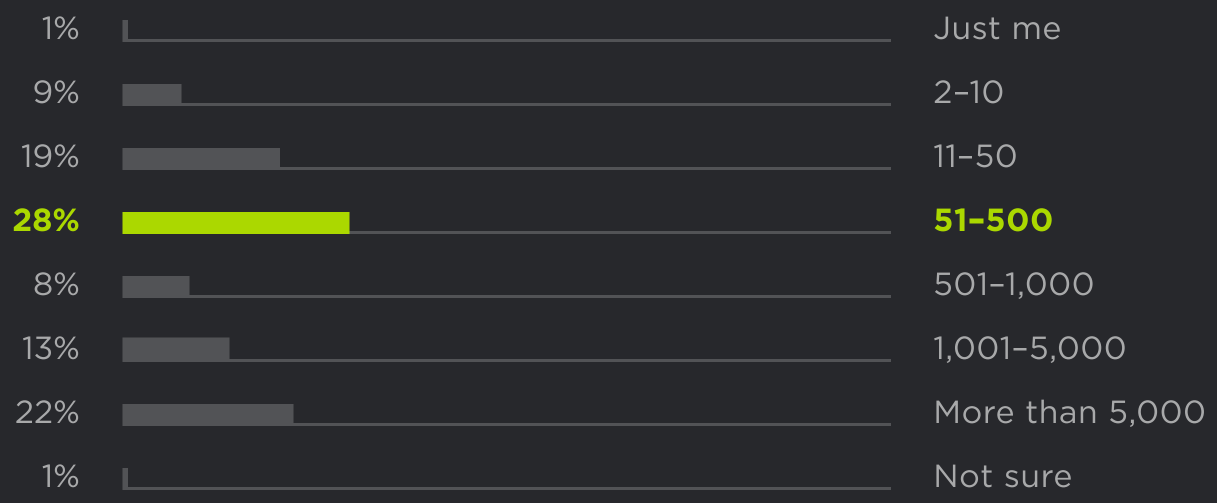
Position level

Number of answers: 1014

6%	22%	**38%**	31%	3%
Trainee	Junior	**Middle**	Senior	Other

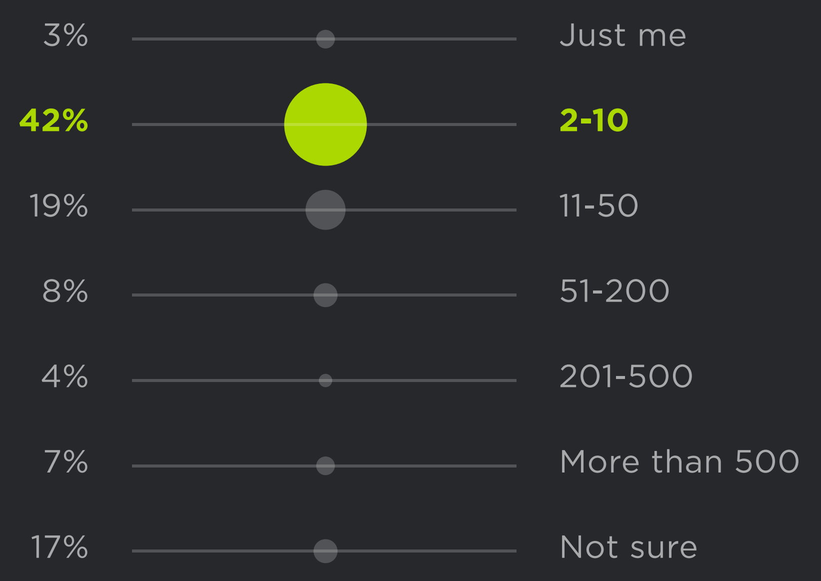
Company size

Number of answers: 1086



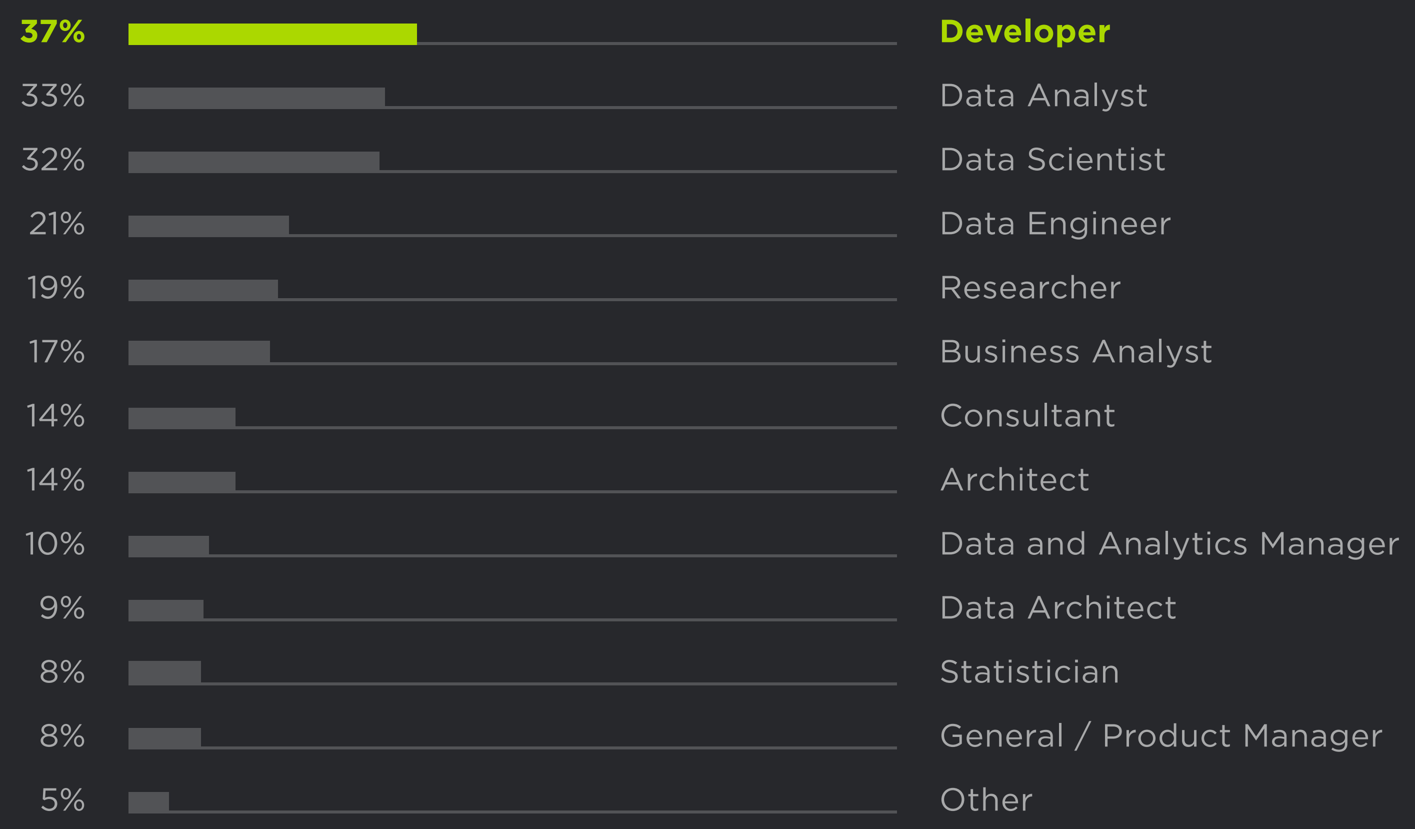
Number of data analysts in company

Number of answers: 1666



Job Role

Number of answers: 917



Note this was a question with checkboxes. Shares may total more than 100%.

JetBrains products for data science and big data

PyCharm Professional Edition is a Python IDE that enables Data Scientists and Web developers to become far more productive.

It offers in-depth Python code analysis and integrates with various libraries, frameworks, and tools. PyCharm's scientific tools are designed specifically with professional data analysts in mind and include a scientific development mode, integration with conda, code cells, Jupyter Notebook support, and much more. There is first-class support available for SQL databases as well.

jetbrains.com/pycharm

JetBrains products for data science and big data

Datalore is an intelligent web application for data analysis and visualization for Python, with built-in tools and libraries for machine learning all included.

datalore.io

The smart Python code editor helps users write better code with suggestions, autocompletion, and syntax highlighting. Incremental recalculation enables dependencies between multiple computations to be followed, so users don’t have to track which parts of the code were affected by recent edits. And there is access to the extended data storage and high-performance computational resources (including GPU instances) for an enhanced exploration experience.

Thank you for your time! We hope you found our report useful.

If you have any questions or suggestions,
please contact us at surveys@jetbrains.com.