WELCOME MESSAGE

Welcome to the 2017 Kaggle Machine Learning and Data Science Survey!

It should take roughly 10-25 minutes to complete. Know that we will publicly release the anonymized raw data from this survey. As a thank-you, all survey participants will be the first ones to receive an email with the survey's results.

Once ready, click on the 'Get Started' button below. If you see any problems with the survey, click the "Report a Bug" link in the footer to get a hold of us.

Thanks again,
The Kaggle Team

A. Basic Demographics

Select your gender identity.

Female

Male

Non-binary, genderqueer, or gender non-conforming

A different identity

Select the country you currently live in.

What's your age?

B. Employment Status Questions

What's your current employment status?

Employed full-time

Employed part-time

Independent contractor, freelancer, or self-employed

Not employed, and not looking for work

Not employed, but looking for work

I prefer not to say

Retired

C. Current Career Profile Questions (Non-workers)

Are you currently enrolled as a student at a degree granting school?

Yes

No

Are you currently focused on learning data science skills either formally or informally?

Yes, I'm focused on learning mostly data science skills

Yes, but data science is a small part of what I'm focused on learning

No, I am not focused on learning data science skills

Exit Survey

What's your motivation for being a Kaggle user?

C. Current Career Profile Questions (Workers)

Do you write code to analyze data in your current job, freelance contracts, or most recent job if retired?

Yes

No

Are you actively looking to switch careers to data science?

Yes

No

Select the option that's most similar to your current job/professional title (or most recent title if retired).

Business Analyst

Computer Scientist

Data Analyst

Data Miner

Data Scientist

DBA/Database Engineer

Engineer

Machine Learning Engineer

Operations Research Practitioner

Predictive Modeler

Programmer

Researcher

Scientist/Researcher

Software Developer/Software Engineer

Statistician

Other

How adequately do you feel your title describes what you do (or what you did if retired)?

Which of the following describe your current employer (or most recent employer if retired)?

(Select all that apply)

Employed by professional services/consulting firm

Employed by company that makes advanced analytic software

Employed by college or university

Employed by a company that performs advanced analytics

Employed by a company that doesn’t perform advanced analytics

Employed by non-profit or NGO

Employed by government

Self-employed

F. Future Learning Goals Questions - ALL

Which tool or technology are you most excited about learning in the next year? (Select one option)

DataRobot	KNIME (free version)	Salfrod Systems CART, MARS, TreeNet, RF, SPM
Amazon Machine Learning	Mathematica	SAP BusinessObjects Predictive Analytics
Amazon Web services	MATLAB/Octave	SAS Base
Angoss	Microsoft Azure Machine Learning	SAS Enterprise Miner
C/C++	Microsoft Excel Data Mining	SAS JMP
Cloudera	Microsoft R Server (Formerly Revolution Analytics)	Stan
Google Cloud Compute	Microsoft SQL Server Data Mining	Spark / MLlib
Flume	Minitab	SQL
Hadoop/Hive/Pig	NoSQL	Statistica (Quest/Dell, formerly Statsoft)
IBM Cognos	Oracle Data Mining/ Oracle R Enterprise	Tableau
IBM SPSS Modeler	Orange	TensorFlow
IBM SPSS Statistics	Perl	TIBCO Spotfire
IBM Watson / Waton Analytics	Python	Unix shell / awk
Impala	QlikView	Weka
Java	R	Julia
Julia	RapidMiner (commercial version)	Other
Jupyter notebooks	RapidMiner (free version)	I don't plan on learning a new tool/technology
KNIME (commercial version)		

Which ML/DS method are you most excited about learning in the next year? (Select one option)

- Time Series Analysis
- Deep learning
- Factor Analysis
- Monte Carlo Methods
- Genetic & Evolutionary Algorithms
- Regression
- Cluster Analysis
- Social Network Analysis
- Decision Trees
- Rule Induction
- Text Mining
- MARS
- Uplift Modeling
- Neural Nets
- Support Vector Machines (SVM)
- Bayesian Methods
- Ensemble Methods (e.g. boosting, bagging)
- Proprietary Algorithms
- Anomaly Detection
- Survival Analysis
- Random Forests
- Link Analysis
- Association Rules
- Other
- I don't plan on learning a new ML/DS method

What programming language would you recommend a new data scientist learn first? (Select one option)

- C/C++/C#
- F#
- Haskell
- Java
- Julia
- Matlab
- Python
- R
- SAS
- Scala
- SQL
- Stata
- Other

Where do you find public datasets to practice data science skills? (Select all that apply)

- Dataset aggregator/platform (i.e. Socrata, Kaggle Datasets, data.world, etc.)
- GitHub
- Google Search
- I collect my own data (e.g. web-scraping)
- University/Non-profit research group websites
- Other

Government website

What is your biggest challenge with the public datasets you find for personal projects?

What platforms & resources have you used to continue learning data science skills? (Select all that apply)

- Arxiv
- Blogs
- College/University
- Company internal community
- Conferences
- Friends network
- Kaggle
- Newsletters
- Non-Kaggle online communities
- Official documentation
- Online courses
- Personal Projects
- Podcasts
- Stack Overflow Q&A
- Textbook
- Trade book
- Tutoring/mentoring
- YouTube Videos
- Other
- Other
- Other

How useful did you find these platforms & resources for learning data science skills?

	Not Useful	Somewhat useful	Very useful
» Arxiv	○	○	○
» Blogs	○	○	○
» College/University	○	○	○

	Not Useful	Somewhat useful	Very useful
» Company internal community	○	○	○
» Conferences	○	○	○
» Friends network	○	○	○
» Kaggle	○	○	○
» Newsletters	○	○	○
» Non-Kaggle online communities	○	○	○
» Official documentation	○	○	○
» Online courses	○	○	○
» Personal Projects	○	○	○
» Podcasts	○	○	○
» Stack Overflow Q&A	○	○	○
» Textbook	○	○	○
» Trade book	○	○	○
» Tutoring/mentoring	○	○	○
» YouTube Videos	○	○	○
» Other	○	○	○
» Other	○	○	○
» Other	○	○	○

What are your top 3 favorite data science blogs/podcasts/newsletters? (Select up to three options)

Becoming a Data Scientist Podcast

Data Elixir Newsletter

Data Machina Newsletter

Data Stories Podcast

DataTau News Aggregator

Emergent/Future Newsletter (Algorithmia)

FastML Blog

FlowingData Blog

Jack's Import AI Newsletter

KDnuggets Blog

Linear Digressions Podcast

No Free Hunch Blog

O'Reilly Data Newsletter

Partially Derivative Podcast

R Bloggers Blog Aggregator

Siraj Raval YouTube Channel

Statistical Modeling, Causal Inference, and Social Science Blog (Andrew Gelman)

Talking Machines Podcast

The Analytics Dispatch Newsletter

The Data Skeptic Podcast

Other (Separate different answers with semicolon)

F. Future Learning Goals Questions - Learners Only

How long have you been learning data science?

< 1 year

1-2 years

3-5 years

5-10 years

10-15 years

15+ years

How important do you think the below skills or certifications are in getting a data science job?

	Unnecessary	Nice to have	Necessary
“Big Data” technology	O	O	O
Academic degree in related field	O	O	O
Advanced Statistics	O	O	O
Enterprise Tools (e.g. SAS)	O	O	O
Fluent in Python	O	O	O
Fluent in R	O	O	O
Fluent in SQL	O	O	O
Kaggle ranking	O	O	O
Online certification or MOOC	O	O	O
Visualization Tools	O	O	O
Other	O	O	O
Other	O	O	O
Other	O	O	O

On which online platforms have you begun or completed data science courses?

Coursera

DataCamp

edX

Udacity

Other

Which computing hardware do you use for your personal ML/DS projects?

Basic laptop (Macbook)

Gaming Laptop (Laptop + CUDA capable GPU)

GPU accelerated Workstation

Laptop + Cloud service (AWS, Azure, GCE ...)

Laptop or Workstation and local IT supported servers

Traditional Workstation

Workstation + Cloud service

Other

On average, how many hours a week do you spend studying data science?

What's the most important way you can prove your knowledge of ML/DS? (Select one option)

Experience from work in a company related to ML

Github Portfolio

Kaggle Competitions

Master's degree

Online Courses and Certifications

PhD

Other

H. Closing Questions

What’s the most impactful machine learning algorithm you've ever deployed?

If given the necessary resources and time, what problem would you personally choose to tackle with machine learning?

C. Current Career Profile Questions (ALL)

Do you currently consider yourself a data scientist?

Yes

No

Sort of (Explain more)

D. Education and Career Background Questions (ALL)

Which level of formal education have you attained?

Which best describes your undergraduate major?

- A business discipline
- A health science
- A humanities discipline
- A social science
- Fine arts or performing arts
- I never declared a major
- Information technology, networking, or system administration
- Management information systems

Biology
Computer Science
Electrical Engineering
Engineering (non-computer focused)
Mathematics or statistics
Physics
Psychology
Other

How long have you been writing code to analyze data?

Less than a year
1 to 2 years
3 to 5 years
6 to 10 years
More than 10 years
I don't write code to analyze data

Select all other job titles you've held in the past 10 years? (Select all that apply)

Business Analyst
Computer Scientist
Data Analyst
Data Miner
Data Scientist
DBA/Database Engineer
Engineer
Machine Learning Engineer
Operations Research Practitioner
Predictive Modeler
Programmer
Researcher
Software Developer/Software Engineer
Statistician
Other
I haven't started working yet

How did you first start your machine learning / data science training? (Select one option)

Self-taught
Online courses (coursera, udemy, edx, etc.)
Work
University courses
Kaggle competitions
Other

What percentage of your current machine learning / data science training falls under each category? (Total must equal 100%)

Self-taught	0
Online courses (coursera, udemy, edx, ..etc.)	0
Work	0
University	0
Kaggle competitions	0
Other	0
Total	0

In which areas of machine learning do you consider yourself competent? (Select all that apply)

- Adversarial Learning
- Computer Vision
- Machine Translation
- Natural Language Processing
- Outlier detection (e.g. Fraud detection)
- Recommendation Engines
- Reinforcement learning
- Speech Recognition
- Supervised Machine Learning (Tabular Data)
- Survival Analysis
- Time Series
- Unsupervised Learning
- Other (please specify; separate by semi-colon)

In which machine learning techniques do you consider yourself competent? (Select all that apply)

- Bayesian Techniques
- Decision Trees - Gradient Boosted Machines
- Decision Trees - Random Forests
- Ensemble Methods
- Evolutionary Approaches
- Gradient Boosting
- Hidden Markov Models HMMs
- Logistic Regression
- Markov Logic Networks
- Neural Networks - CNNs
- Neural Networks - GANs
- Neural Networks - RNNs
- Support Vector Machines (SVMs)
- Other (please specify; separate by semi-colon)

What's the highest level of education completed by either of your parents?

E. Education and Career Background Questions (ALL)

Which industry is your current employer/contract in (or most recent employer if retired)?

Academic
CRM/Marketing
Financial
Government
Hospitality/Entertainment/Sports
Insurance
Internet-based
Manufacturing
Military/Security
Non-profit
Pharmaceutical
Retail
Technology
Telecommunications
Mix of fields
Other

E. Education and Career Background Questions (Workers EXCLUDING contractors/ret)

How many employees work at your current or most recent company?

How has the size of your organization’s ML/DS staff changed over the past year?

How many years has your organization been utilizing advanced analytics/data science?

How did you find your current job?

A friend, family member, or former colleague told me

I was contacted directly by someone at the company (e.g. internal recruiter)

A general-purpose job board

An external recruiter or headhunter

I visited the company's Web site and found a job listing there

A career fair or on-campus recruiting event

A tech-specific job board

Some other way

E. Education and Career Background Questions (Workers EXCLUDING retired)

How important was your formal education or degree to your career success analyzing data?

Very important

Important

Somewhat important

Not very important

Not at all important

N/A, I did not receive any formal education

What is the primary function of your role? (Select one option)

Analyze and understand data to influence product or business decisions

Build and/or run a machine learning service that operationally improves your product or workflows

Build and/or run the data infrastructure that your business uses for storing, analyzing, and operationalizing data

Build prototypes to explore applying machine learning to new areas

Research that advances the state of the art of machine learning

Other

At work, which computing hardware do you use for ML/DS projects?

- Basic laptop (Macbook)
- Gaming Laptop (Laptop + CUDA capable GPU)
- GPU accelerated Workstation
- Laptop + Cloud service (AWS, Azure, GCE ...)
- Laptop or Workstation and local IT supported servers
- Laptop or Workstation and private datacenters
- Traditional Workstation
- Workstation + Cloud service
- Other

At work, which kind of data do you typically work with?

- Image data
- Video data
- Text data
- Relational data
- Other

At work, how often do the models you build get put into production?

At work, which libraries do you typically use? (Please separate different answers with semicolons)

Of the models you've trained at work, what is the typical size of datasets used?

At work, which algorithms/analytic methods do you typically use? (Select all that apply)

- Bayesian Techniques
- CNNs
- Decision Trees
- Markov Logic Networks
- Neural Networks
- Random Forests

Ensemble Methods
Evolutionary Approaches
GANs
Gradient Boosted Machines
HMMs
Regression/Logistic Regression
RNNs
SVMs
Other

For work, which data science/analytics tools, technologies, and languages have you used in the past year? (Select all that apply)

Amazon Machine Learning
Amazon Web services
Angoss
C/C++
Cloudera
DataRobot
Flume
Google Cloud Compute
Hadoop/Hive/Pig
IBM Cognos
IBM SPSS Modeler
IBM SPSS Statistics
IBM Watson / Waton Analytics
Impala
Java
Julia
Jupyter notebooks
KNIME (commercial version)
KNIME (free version)
Mathematica
MATLAB/Octave
Microsoft Azure Machine Learning
Microsoft Excel Data Mining
Microsoft R Server (Formerly Revolution Analytics)
Microsoft SQL Server Data Mining
NoSQL
Oracle Data Mining/ Oracle R Enterprise
Orange
Perl
Python
QlikView
R
RapidMiner (commercial version)
RapidMiner (free version)
Salfrod Systems CART, MARS, TreeNet, RF, SPM
SAP BusinessObjects Predictive Analytics
SAS Base
SAS Enterprise Miner
SAS JMP
Spark / MLlib
SQL
Stan
Statistica (Quest/Dell, formerly Statsoft)
Tableau
TensorFlow
TIBCO Spotfire
Unix shell / awk
Other
Other
Other

Minitab

At work, how often did you use the following data science/analytics tools, technologies, and languages this past year?

	Rarely	Sometimes	Often	Most of the time
» Amazon Machine Learning	O	O	O	O
» Amazon Web services	O	O	O	O
» Angoss	O	O	O	O
» C/C++	O	O	O	O
» Cloudera	O	O	O	O
» DataRobot	O	O	O	O
» Flume	O	O	O	O
» Google Cloud Compute	O	O	O	O
» Hadoop/Hive/Pig	O	O	O	O
» IBM Cognos	O	O	O	O
» IBM SPSS Modeler	O	O	O	O
» IBM SPSS Statistics	O	O	O	O
» IBM Watson / Waton Analytics	O	O	O	O
» Impala	O	O	O	O
» Java	O	O	O	O
» Julia	O	O	O	O
» Jupyter notebooks	O	O	O	O
» KNIME (commercial version)	O	O	O	O
» KNIME (free version)	O	O	O	O
» Mathematica	O	O	O	O
» MATLAB/Octave	O	O	O	O
» Microsoft Azure Machine Learning	O	O	O	O
» Microsoft Excel Data Mining	O	O	O	O
» Microsoft R Server (Formerly Revolution Analytics)	O	O	O	O
» Microsoft SQL Server Data Mining	O	O	O	O
» Minitab	O	O	O	O

	Rarely	Sometimes	Often	Most of the time
» NoSQL	O	O	O	O
» Oracle Data Mining/ Oracle R Enterprise	O	O	O	O
» Orange	O	O	O	O
» Perl	O	O	O	O
» Python	O	O	O	O
» QlikView	O	O	O	O
» R	O	O	O	O
» RapidMiner (commercial version)	O	O	O	O
» RapidMiner (free version)	O	O	O	O
» Salfrod Systems CART, MARS, TreeNet, RF, SPM	O	O	O	O
» SAP BusinessObjects Predictive Analytics	O	O	O	O
» SAS Base	O	O	O	O
» SAS Enterprise Miner	O	O	O	O
» SAS JMP	O	O	O	O
» Spark / MLlib	O	O	O	O
» SQL	O	O	O	O
» Stan	O	O	O	O
» Statistica (Quest/Dell, formerly Statsoft)	O	O	O	O
» Tableau	O	O	O	O
» TensorFlow	O	O	O	O
» TIBCO Spotfire	O	O	O	O
» Unix shell / awk	O	O	O	O
» Other	O	O	O	O
» Other	O	O	O	O
» Other	O	O	O	O

At work, which data science methods do you use? (Select all that apply)

A/B Testing

Association Rules

Bayesian Techniques

CNNs

Collaborative Filtering

Cross-Validation

Data Visualization

Naive Bayes

Natural Language Processing

Neural Networks

PCA and Dimensionality Reduction

Prescriptive Modeling

Random Forests

Recommender Systems

Decision Trees
Ensemble Methods
Evolutionary Approaches
GANs
Gradient Boosted Machines
HMMs
kNN and Other Clustering

Lift Analysis

Logistic Regression

Markov Logic Networks

RNNs
Segmentation
Simulation
SVMs
Text Analytics
Time Series Analysis
Other

Other

Other

At work, how often do you use the following data science methods?

	Rarely	Sometimes	Often	Most of the time
» A/B Testing	O	O	O	O
» Association Rules	O	O	O	O
» Bayesian Techniques	O	O	O	O
» CNNs	O	O	O	O
» Collaborative Filtering	O	O	O	O
» Cross-Validation	O	O	O	O
» Data Visualization	O	O	O	O
» Decision Trees	O	O	O	O
» Ensemble Methods	O	O	O	O
» Evolutionary Approaches	O	O	O	O
» GANs	O	O	O	O
» Gradient Boosted Machines	O	O	O	O
» HMMs	O	O	O	O
» kNN and Other Clustering	O	O	O	O
» Lift Analysis	O	O	O	O
» Logistic Regression	O	O	O	O
» Markov Logic Networks	O	O	O	O
» Naive Bayes	O	O	O	O

	Rarely	Sometimes	Often	Most of the time
» Natural Language Processing	O	O	O	O
» Neural Networks	O	O	O	O
» PCA and Dimensionality Reduction	O	O	O	O
» Prescriptive Modeling	O	O	O	O
» Random Forests	O	O	O	O
» Recommender Systems	O	O	O	O
» RNNs	O	O	O	O
» Segmentation	O	O	O	O
» Simulation	O	O	O	O
» SVMs	O	O	O	O
» Text Analytics	O	O	O	O
» Time Series Analysis	O	O	O	O
» Other	O	O	O	O
» Other	O	O	O	O
» Other	O	O	O	O

At work, on average, what percentage of your time is devoted to:
(Total must equal 100%)

Gathering and cleaning data	0
Model building/model selection	0
Putting your work into production	0
Visualizing data	0
Finding insights in the data and communicating these to relevant stakeholders	0
Other	0
Total	0

At which level do you understand the mathematics behind the algorithms you use at work?

At work, which barriers or challenges have you faced this past year? (Select all that apply)

Company politics / Lack of management/financial support for a data science team

Data Science results not used by business decision makers

Did not instrument data useful for scientific analysis and decision-making

Difficulties in deployment/scoring

Dirty data

Explaining data science to others

I prefer not to say

Inability to integrate findings into organization's decision-making process

Lack of data science talent in the organization

Lack of funds to buy useful datasets from external sources

Lack of significant domain expert input

Limitations in the state of the art in machine learning

Limitations of tools

Maintaining responsible expectations about the potential impact of data science projects

Need to coordinate with IT

Organization is small and cannot afford a data science team

Privacy issues

Scaling data science solution up to full database

Team using multiple ad-hoc development environments such as Python, R, Java, etc.

The lack of a clear question to be answering, or a clear direction to go in, with the available data

Unavailability of/difficult access to data

Other

At work, how often did you experience these barriers or challenges within the past year?

	Rarely	Sometimes	Often	Most of the time
» Company politics / Lack of management/financial support for a data science team	O	O	O	O
» Data Science results not used by business decision makers	O	O	O	O
» Did not instrument data useful for scientific analysis and decision-making	O	O	O	O
» Difficulties in deployment/scoring	O	O	O	O
» Dirty data	O	O	O	O
» Explaining data science to others	O	O	O	O
» I prefer not to say	O	O	O	O
» Inability to integrate findings into organization's decision-making process	O	O	O	O
» Lack of data science talent in the organization	O	O	O	O

	Rarely	Sometimes	Often	Most of the time
» Lack of funds to buy useful datasets from external sources	O	O	O	O
» Lack of significant domain expert input	O	O	O	O
» Limitations in the state of the art in machine learning	O	O	O	O
» Limitations of tools	O	O	O	O
» Maintaining responsible expectations about the potential impact of data science projects	O	O	O	O
» Need to coordinate with IT	O	O	O	O
» Organization is small and cannot afford a data science team	O	O	O	O
» Privacy issues	O	O	O	O
» Scaling data science solution up to full database	O	O	O	O
» Team using multiple ad-hoc development environments such as Python, R, Java, etc.	O	O	O	O
» The lack of a clear question to be answering, or a clear direction to go in, with the available data	O	O	O	O
» Unavailability of/difficult access to data	O	O	O	O
» Other	O	O	O	O

At work, what proportion of your analytics projects incorporate data visualization?

None

Less than 10% of projects

10-25% of projects

26-50% of projects

51-75% of projects

76-99% of projects

100% of projects

At work, to what degree does your team use interval versus external resources for data science projects? (Select one option)

Entirely internal

More internal than external

Approximately half internal and half external

More external than internal

Entirely external

Do not know

At work, where does the data scientist team sit within the organization?

IT Department

Business Department

Central Insights Team

Standalone Team

Other

At work, in addition to proprietary data, what third party and public datasets do you work with? (Please separate different answers with semicolons)

What is your biggest challenge for working with your source data?

At work, which of these data storage models do you typically use? (Select all that apply)

Column-oriented relational (e.g. KDB, MariaDB)

Document-oriented (e.g. MongoDB, Elasticsearch)

Flat files not in a database or cache (e.g. CSV, JSON, XML, PNG, MPG)

Graph (e.g. GraphBase, Neo4j)

Key-value store (e.g. Redis, Riak)

Other tabular (e.g. Cassandra, BigTable, BigQuery, Redshift)

Row-oriented relational (e.g. MySQL, Microsoft SQL Server)

Other

At work, which tools do you use to share source data? (Select all that apply)

Commercial Data Platform

Company Developed Platform

Email

I don't typically share data

Share Drive/SharePoint

Other

At work, which tools do you use to share code? (Select all that apply)

Bitbucket

Generic cloud file sharing software (Dropbox, Box, etc.)

Generic non-cloud file sharing software (Email, Shared Server, etc.)

Git

Mercurial

Subversion

Other

How often do you work remotely?

What is your current total yearly compensation (salary + bonus)?

	Total Amount (e.g. 75,000)	Choose your currency:
Your Compensation:		

How has your salary/compensation changed in the past 3 years?

Has increased 20% or more

Has increased between 6% and 19%

Has stayed about the same (has not increased or decreased more than 5%)

Has decreased between 6% and 19%

Has decreased 20% or more

I was not employed 3 years ago

I am not currently employed

I do not want to share information about my salary/compensation

Other

On a scale from 0 (Highly Dissatisfied) - 10 (Highly Satisfied), how satisfied are you with your current job?

G. Job Search Questions

Which resource has been the best for finding data science job openings? (Select one option)

Meeting with recruiters who've contacted you directly

Tech-specific job board

Company's Web site/job listing page

Searching general-purpose job board

Career fair or on-campus recruiting event

Asking friends, family members, or former colleagues for leads

Other

How many hours per week have you typically spend looking for a data science job?

How are you assessing potential job opportunities?

	Not important	Somewhat important	Very Important
Opportunities for professional development			

	Not important	Somewhat important	Very Important
The compensation and benefits offered	O	O	O
The office environment I'd be working in	O	O	O
The languages, frameworks, and other technologies I'd be working with	O	O	O
The amount of time I'd have to spend commuting	O	O	O
How projects are managed at the company or organization	O	O	O
The experience level called for in the job description	O	O	O
The specific department or team I'd be working on	O	O	O
The specific role or job title I'd be applying for	O	O	O
The financial performance or funding status of the company or organization	O	O	O
How widely used or impactful the product or service I'd be working on is	O	O	O
The opportunity to work from home/remotely	O	O	O
The industry that I'd be working in	O	O	O
The reputations of the company's senior leaders	O	O	O
The diversity of the company or organization	O	O	O
Opportunity to publish my results	O	O	O