2020 Kaggle DS & ML Survey

- Questions and answer choices

Main Survey 1

Supplementary Questions: 16

Main Survey

Q1

What is your age (# years)?

[List of Values]

Q2

What is your gender?

- Man
- Woman
- Nonbinary
- Prefer not to say
- Prefer to self-describe

Q3

In which country do you currently reside?

[List of Countries]

Q4

What is the highest level of formal education that you have attained or plan to attain within the next 2 years?

- No formal education past high school
- Some college/university study without earning a bachelor’s degree
- Bachelor’s degree
- Master’s degree
- Doctoral degree
- Professional degree
- I prefer not to answer

Q5

Select the title most similar to your current role (or most recent title if retired):

- Business Analyst
- Data Analyst
- Data Engineer
- Data Scientist
- DBA/Database Engineer
- Machine Learning Engineer
- Product/Project Manager
- Research Scientist
- Software Engineer
- Statistician
- Student
- Currently not employed
- Other

Q6

For how many years have you been writing code and/or programming?

- I have never written code
- < 1 years
- 1-2 years
- 3-5 years
- 5-10 years
- 10-20 years
- 20+ years

Q7

What programming languages do you use on a regular basis? (Select all that apply)

- Python
- R
- SQL
- C
- C++
- Java
- Javascript
- Julia
- Swift
- Bash
- MATLAB
- None
- Other

Q8

What programming language would you recommend an aspiring data scientist to learn first?

- » Python
- » R
- » SQL
- » C
- » C++
- » Java
- » Javascript
- » Julia
- » Swift
- » Bash
- » MATLAB
- » None
- » Other

Q9

Which of the following integrated development environments (IDE's) do you use on a regular basis? (Select all that apply)

- JupyterLab (or products based off of Jupyter)
- RStudio
- Visual Studio
- Visual Studio Code (VSCode)
- PyCharm
- Spyder
- Notepad++
- Sublime Text
- Vim, Emacs, or similar
- MATLAB
- None
- Other

Q10

Which of the following hosted notebook products do you use on a regular basis? (Select all that apply)

- Kaggle Notebooks
- Colab Notebooks
- Azure Notebooks
- Paperspace / Gradient
- Binder / JupyterHub
- Code Ocean
- IBM Watson Studio
- Amazon Sagemaker Studio
- Amazon EMR Notebooks
- Google Cloud AI Platform Notebooks
- Google Cloud Datalab Notebooks
- Databricks Collaborative Notebooks
- None
- Other

Q11

What type of computing platform do you use most often for your data science projects?

- A personal computer or laptop
- A deep learning workstation (NVIDIA GTX, LambdaLabs, etc)
- A cloud computing platform (AWS, Azure, GCP, hosted notebooks, etc)
- None
- Other

Q12

Which types of specialized hardware do you use on a regular basis? (Select all that apply)

- GPUs
- TPUs
- None
- Other

Q13

Approximately how many times have you used a TPU (tensor processing unit)?

- Never
- Once
- 2-5 times
- 6-25 times
- More than 25 times

Q14

What data visualization libraries or tools do you use on a regular basis? (Select all that apply)

- Matplotlib
- Seaborn
- Plotly / Plotly Express
- Ggplot / ggplot2
- Shiny
- D3 js
- Altair
- Bokeh
- Geoplotlib
- Leaflet / Folium
- None
- Other

Q15

For how many years have you used machine learning methods?

- I do not use machine learning methods
- Under 1 year
- 1-2 years
- 2-3 years
- 3-4 years
- 4-5 years
- 5-10 years
- 10-20 years
- 20 or more years

Q16

Which of the following machine learning frameworks do you use on a regular basis? (Select all that apply)

- Scikit-learn
- TensorFlow
- Keras
- PyTorch
- Fast.ai
- MXNet
- Xgboost
- LightGBM
- CatBoost
- Prophet
- H2O 3
- Caret
- Tidymodels
- JAX
- None
- Other

Q17

Which of the following ML algorithms do you use on a regular basis? (Select all that apply):

- Linear or Logistic Regression
- Decision Trees or Random Forests
- Gradient Boosting Machines (xgboost, lightgbm, etc)
- Bayesian Approaches
- Evolutionary Approaches
- Dense Neural Networks (MLPs, etc)
- Convolutional Neural Networks
- Generative Adversarial Networks
- Recurrent Neural Networks
- Transformer Networks (BERT, gpt-3, etc)
- None
- Other

Q18

Which categories of computer vision methods do you use on a regular basis? (Select all that apply)[1]

- General purpose image/video tools (PIL, cv2, skimage, etc)
- Image segmentation methods (U-Net, Mask R-CNN, etc)
- Object detection methods (YOLOv3, RetinaNet, etc)
- Image classification and other general purpose networks (VGG, Inception, ResNet, ResNeXt, NASNet, EfficientNet, etc)
- Generative Networks (GAN, VAE, etc)
- None
- Other

Q19

Which of the following natural language processing (NLP) methods do you use on a regular basis? (Select all that apply)[2]

- Word embeddings/vectors (GLoVe, fastText, word2vec)
- Encoder-decoder models (seq2seq, vanilla transformers)
- Contextualized embeddings (ELMo, CoVe)
- Transformer language models (GPT-3, BERT, XLnet, etc)
- None
- Other

Q20

What is the size of the company where you are employed?

- 0-49 employees
- 50-249 employees
- 250-999 employees
- 1000-9,999 employees
- 10,000 or more employees

[1] Question 18 (which specific ML methods) was only asked to respondents that selected the relevant answer choices for Question 17 (which categories of algorithms).

[2] Question 19 (which specific ML methods) was only asked to respondents that selected the relevant answer choices for Question 17 (which categories of algorithms).

Q21

Approximately how many individuals are responsible for data science workloads at your place of business?

- 0
- 1-2
- 3-4
- 5-9
- 10-14
- 15-19
- 20+

Q22

Does your current employer incorporate machine learning methods into their business?

- We are exploring ML methods (and may one day put a model into production)
- We use ML methods for generating insights (but do not put working models into production)
- We recently started using ML methods (i.e., models in production for less than 2 years)
- We have well established ML methods (i.e., models in production for more than 2 years)
- No (we do not use ML methods)
- I do not know

Q23

Select any activities that make up an important part of your role at work: (Select all that apply)

- Analyze and understand data to influence product or business decisions
- Build and/or run the data infrastructure that my business uses for storing, analyzing, and operationalizing data
- Build prototypes to explore applying machine learning to new areas
- Build and/or run a machine learning service that operationally improves my product or workflows
- Experimentation and iteration to improve existing ML models
- Do research that advances the state of the art of machine learning
- None of these activities are an important part of my role at work
- Other

Q24

What is your current yearly compensation (approximate $USD)?

[List of Values]

Q25

Approximately how much money have you (or your team) spent on machine learning and/or cloud computing services at home (or at work) in the past 5 years (approximate $USD)?

- $0 ($USD)
- $1-$99
- $100-$999
- $1000-$9,999
- $10,000-$99,999
- $100,000 or more ($USD)

Q26-A

Which of the following cloud computing platforms do you use on a regular basis? (Select all that apply)

- Amazon Web Services (AWS)
- Microsoft Azure
- Google Cloud Platform (GCP)
- IBM Cloud / Red Hat
- Oracle Cloud
- SAP Cloud
- Salesforce Cloud
- VMware Cloud
- Alibaba Cloud
- Tencent Cloud
- None
- Other

Q27-A

Do you use any of the following cloud computing products on a regular basis? (Select all that apply)[3]

- Amazon EC2
- AWS Lambda
- Amazon Elastic Container Service
- Azure Cloud Services
- Microsoft Azure Container Instances
- Azure Functions
- Google Cloud Compute Engine
- Google Cloud Functions
- Google Cloud Run
- Google Cloud App Engine
- No / None
- Other

Q28-A

Do you use any of the following machine learning products on a regular basis? (Select all that apply)[4]

- Amazon SageMaker
- Amazon Forecast
- Amazon Rekognition
- Azure Machine Learning Studio
- Azure Cognitive Services
- Google Cloud AI Platform / Google Cloud ML Engine
- Google Cloud Video AI
- Google Cloud Natural Language
- Google Cloud Vision AI
- No / None
- Other

[3] Question 27-A (which specific AWS/Azure/GCP products) was only asked to respondents that selected the relevant answer choices for Question 26-A (which of the following companies).

[4] Question 28-A (which specific AWS/Azure/GCP products) was only asked to respondents that selected the relevant answer choices for Question 26-A (which of the following companies).

Q29-A

Which of the following big data products (relational databases, data warehouses, data lakes, or similar) do you use on a regular basis? (Select all that apply)

- MySQL
- PostgreSQL
- SQLite
- Oracle Database
- MongoDB
- Snowflake
- IBM Db2
- Microsoft SQL Server
- Microsoft Access
- Microsoft Azure Data Lake Storage
- Amazon Redshift
- Amazon Athena
- Amazon DynamoDB
- Google Cloud BigQuery
- Google Cloud SQL
- Google Cloud Firestore
- None
- Other

Q30

Which of the following big data products (relational database, data warehouse, data lake, or similar) do you use most often?[5]

- » MySQL
- » PostgreSQL
- » SQLite
- » Oracle Database
- » MongoDB
- » Snowflake
- » IBM Db2
- » Microsoft SQL Server
- » Microsoft Access
- » Microsoft Azure Data Lake Storage
- » Amazon Redshift
- » Amazon Athena
- » Amazon DynamoDB
- » Google Cloud BigQuery
- » Google Cloud SQL
- » Google Cloud Firestore
- » None
- » Other

[5] Question 30 (which specific product) was only asked to respondents that selected more than one choice for Question 29-A (which of the following products).

Q31-A

Which of the following business intelligence tools do you use on a regular basis? (Select all that apply)

- Amazon QuickSight
- Microsoft Power BI
- Google Data Studio
- Looker
- Tableau
- Salesforce
- Einstein Analytics
- Qlik
- Domo
- TIBCO Spotfire
- Alteryx
- Sisense
- SAP Analytics Cloud
- None
- Other

Q32

Which of the following business intelligence tools do you use most often?[6]

- » Amazon QuickSight
- » Microsoft Power BI
- » Google Data Studio
- » Looker
- » Tableau
- » Salesforce
- » Einstein Analytics
- » Qlik
- » Domo
- » TIBCO Spotfire
- » Alteryx
- » Sisense
- » SAP Analytics Cloud
- » None
- » Other

[6] Question 32 (which specific product) was only asked to respondents that selected more than one choice for Question 31-A (which of the following products).

Q33-A

Do you use any automated machine learning tools (or partial AutoML tools) on a regular basis? (Select all that apply)

- Automated data augmentation (e.g. imgaug, albumentations)
- Automated feature engineering/selection (e.g. tpot, boruta_py)
- Automated model selection (e.g. auto-sklearn, xcessiv)
- Automated model architecture searches (e.g. darts, enas)
- Automated hyperparameter tuning (e.g. hyperopt, ray.tune, Vizier)
- Automation of full ML pipelines (e.g. Google AutoML, H20 Driverless AI)
- No / None
- Other

Q34-A

Which of the following automated machine learning tools (or partial AutoML tools) do you use on a regular basis? (Select all that apply)[7]

- Google Cloud AutoML
- H20 Driverless AI
- Databricks AutoML
- DataRobot AutoML
- Tpot
- Auto-Keras
- Auto-Sklearn
- Auto_ml
- Xcessiv
- MLbox
- No / None
- Other

Q35-A

Do you use any tools to help manage machine learning experiments? (Select all that apply)

- Neptune.ai
- Weights & Biases
- Comet.ml
- Sacred + Omniboard
- TensorBoard
- Guild.ai
- Polyaxon
- Trains
- Domino Model Monitor
- No / None
- Other

[7] Question 34-A (which specific product) was only asked to respondents that answered affirmatively to Question 33-A (which of the following categories of products).

Q36

Where do you publicly share or deploy your data analysis or machine learning applications? (Select all that apply)

- Plotly Dash
- Streamlit
- NBViewer
- GitHub
- Personal blog
- Kaggle
- Colab
- Shiny
- None / I do not share my work publicly
- Other

Q37

On which platforms have you begun or completed data science courses? (Select all that apply)

- Coursera
- edX
- Kaggle Learn Courses
- DataCamp
- Fast.ai
- Udacity
- Udemy
- LinkedIn Learning
- Cloud-certification programs (direct from AWS, Azure, GCP, or similar)
- University Courses (resulting in a university degree)
- None
- Other

Q38

What is the primary tool that you use at work or school to analyze data? (Include text response)

- Basic statistical software (Microsoft Excel, Google Sheets, etc.)
- Advanced statistical software (SPSS, SAS, etc.)
- Business intelligence software (Salesforce, Tableau, Spotfire, etc.)
- Local development environments (RStudio, JupyterLab, etc.)
- Cloud-based data software & APIs (AWS, GCP, Azure, etc.)
- Other

Q39

Who/what are your favorite media sources that report on data science topics? (Select all that apply)

- Twitter (data science influencers)
- Email newsletters (Data Elixir, O'Reilly Data & AI, etc)
- Reddit (r/machinelearning, etc)
- Kaggle (notebooks, forums, etc)
- Course Forums (forums.fast.ai, Coursera forums, etc)
- YouTube (Kaggle YouTube, Cloud AI Adventures, etc)
- Podcasts (Chai Time Data Science, O’Reilly Data Show, etc)
- Blogs (Towards Data Science, Analytics Vidhya, etc)
- Journal Publications (peer-reviewed journals, conference proceedings, etc)
- Slack Communities (ods.ai, kagglenoobs, etc)
- None
- Other

Supplementary Questions:[8]

Q26-B

Which of the following cloud computing platforms do you hope to become more familiar with in the next 2 years?

- Amazon Web Services (AWS)
- Microsoft Azure
- Google Cloud Platform (GCP)
- IBM Cloud / Red Hat
- Oracle Cloud
- SAP Cloud
- VMware Cloud
- Salesforce Cloud
- Alibaba Cloud
- Tencent Cloud
- None
- Other

Q27-B

In the next 2 years, do you hope to become more familiar with any of these specific cloud computing products? (Select all that apply)

- Amazon EC2
- AWS Lambda
- Amazon Elastic Container Service
- Azure Cloud Services
- Microsoft Azure Container Instances
- Azure Functions
- Google Cloud Compute Engine
- Google Cloud Functions
- Google Cloud Run
- Google Cloud App Engine
- None
- Other

[8] Non-professionals received questions with an alternate phrasing (questions for non-professionals asked what tools they hope to become familiar with in the next 2 years instead of asking what tools they use on a regular basis). Non-professionals were defined as students, unemployed, and respondents that have never spent any money in the cloud.

Q28-B

In the next 2 years, do you hope to become more familiar with any of these specific machine learning products? (Select all that apply)

- Amazon SageMaker
- Amazon Forecast
- Amazon Rekognition
- Azure Machine Learning Studio
- Azure Cognitive Services
- Google Cloud AI Platform / Google Cloud ML Engine
- Google Cloud Video AI
- Google Cloud Natural Language
- Google Cloud Vision AI
- None
- Other

Q29-B

Which of the following big data products (relational databases, data warehouses, data lakes, or similar) do you hope to become more familiar with in the next 2 years? (Select all that apply)

- MySQL
- PostgreSQL
- SQLite
- Oracle Database
- MongoDB
- Snowflake
- IBM Db2
- Microsoft SQL Server
- Microsoft Access
- Microsoft Azure Data Lake Storage
- Amazon Redshift
- Amazon Athena
- Amazon DynamoDB
- Google Cloud BigQuery
- Google Cloud SQL
- Google Cloud Firestore
- None
- Other

Q31-B

Which of the following business intelligence tools do you hope to become more familiar with in the next 2 years? (Select all that apply)

- Microsoft Power BI
- Amazon QuickSight
- Google Data Studio
- Looker

- Tableau
- Salesforce
- Einstein Analytics
- Qlik
- Domo
- TIBCO Spotfire
- Alteryx
- Sisense
- SAP Analytics Cloud
- None
- Other

Q33-B

Which categories of automated machine learning tools (or partial AutoML tools) do you hope to become more familiar with in the next 2 years? (Select all that apply)

- Automated data augmentation (e.g. imgaug, albumentations)
- Automated feature engineering/selection (e.g. tpot, boruta_py)
- Automated model selection (e.g. auto-sklearn, xcessiv)
- Automated model architecture searches (e.g. darts, enas)
- Automated hyperparameter tuning (e.g. hyperopt, ray.tune, Vizier)
- Automation of full ML pipelines (e.g. Google Cloud AutoML, H20 Driverless AI)
- None
- Other

Q34-B

Which specific automated machine learning tools (or partial AutoML tools) do you hope to become more familiar with in the next 2 years? (Select all that apply)

- Google Cloud AutoML
- H20 Driverless AI
- Databricks AutoML
- DataRobot AutoML
- Tpot
- Auto-Keras
- Auto-Sklearn
- Auto_ml
- Xcessiv
- MLbox
- None
- Other

Q35-B

In the next 2 years, do you hope to become more familiar with any of these tools for managing ML experiments? (Select all that apply)

- Neptune.ai
- Weights & Biases
- Comet.ml
- Sacred + Omniboard
- TensorBoard
- Guild.ai
- Polyaxon
- Trains
- Domino Model Monitor
- None
- Other