

Exploratory Data Analysis, Painlessly

X.Y. Han



Exploratory data analysis

From Wikipedia, the free encyclopedia

In [statistics](#), **exploratory data analysis (EDA)** is an approach to [analyzing data sets](#) to summarize their main characteristics, often with visual methods. A [statistical model](#) can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

Exploratory data analysis was promoted by [John Tukey](#) to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments.

A Data Science Story

You



Data scientist

Massive Computational Experiments, Painlessly (STATS 285)

Stanford University, Spring 2021

Ambitious Data Science requires massive computational experimentation; the entry ticket for a solid PhD in some fields is now to conduct experiments involving 1 Million CPU hours. This course covers state-of-the-art practices for conducting massive computational experiments in the cloud in a pain-free and reproducible manner. In addition to giving students a hands-on experience with cluster computing, the course features several guest lectures by renowned data scientists.

Instructors:



David Donoho



Alon Kipnis



Mahsa Lotfi

Logistics

For questions, concerns or bug reports, please contact [Alon Kipnis](#) or [Mahsa Lotfi](#) or [David Donoho](#). This course meets Mondays 2:30-3:50 PM on Zoom. If you are a guest speaker for this course, please read [travel section](#) to plan your visit.

Tweets by @stats285

that it is compatible with the XYZ paradigm, run it using [@clusterjob](#) on Stanford's Sherlock cluster or on the cloud, and visualize the obtained results in [@tableau](#)

Apr 12, 2021

[Stats285 Stanford](#)
@stats285

[@stats285](#) STATS 285 model 2021 is up and running!

Embed

View on Twitter


Data Science News

- [Envisioning the Data Science Discipline \(NAS\)](#)
- [The State of Data Science \(Kaggle\)](#)

Data Science News

- [Envisioning the Data Science Discipline \(NAS\)](#)
- [The State of Data Science \(Kaggle\)](#)

A Data Science Story

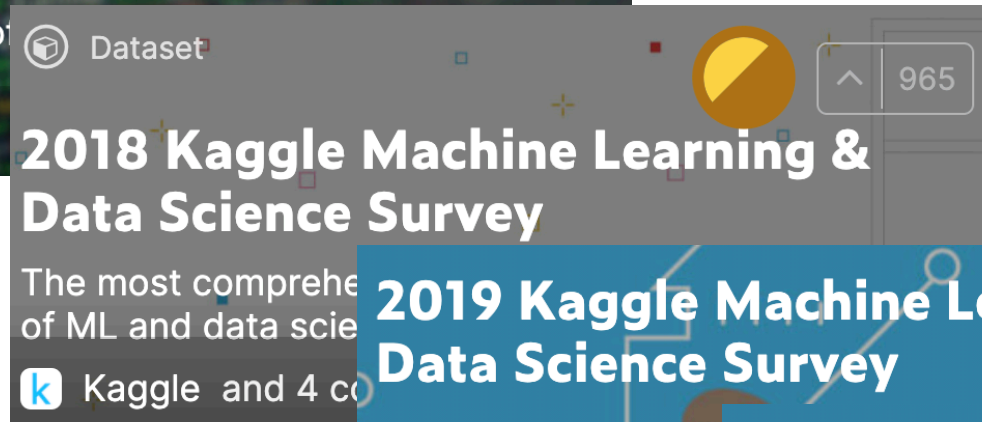


Dataset 821

2017 Kaggle Machine Learning & Data Science Survey

A big picture view of machine learning.

Kaggle



Dataset 965

2018 Kaggle Machine Learning & Data Science Survey

The most comprehensive of ML and data science

Kaggle and 4 co



2019 Kaggle Machine Learning & Data Science Survey

The most comprehensive of ML and data science



2020 Kaggle Machine Learning & Data Science Survey

The most comprehensive dataset available on the state of ML and data science

Kaggle · 4 months ago

A Data Science Story

kaggle

State of Machine Learning and Data Science 2020



Enterprise Executive Summary Report

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.

<https://www.kaggle.com/kaggle-survey-2020>

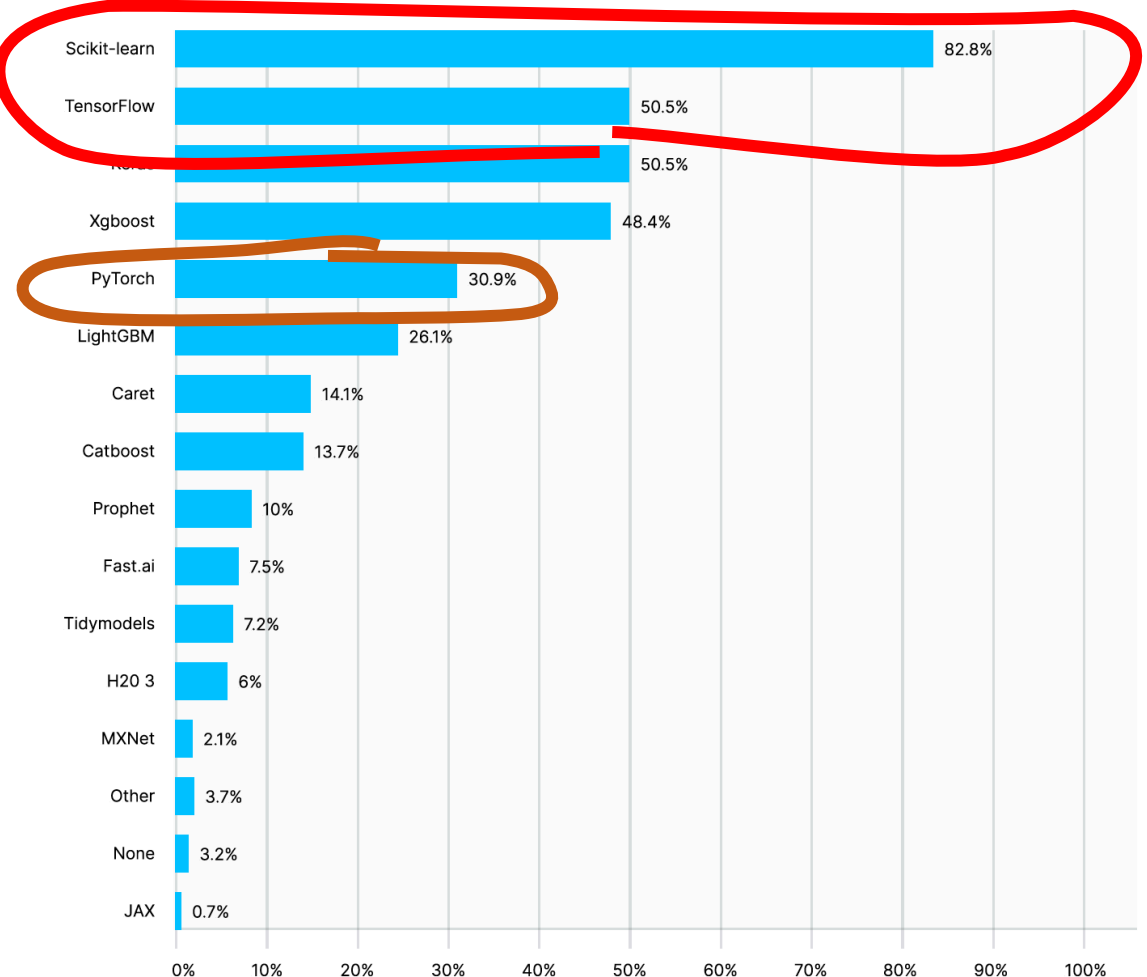
A Data Science Story

Q16

Which of the following machine learning frameworks do you use on a regular basis?

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

MACHINE LEARNING FRAMEWORK USAGE

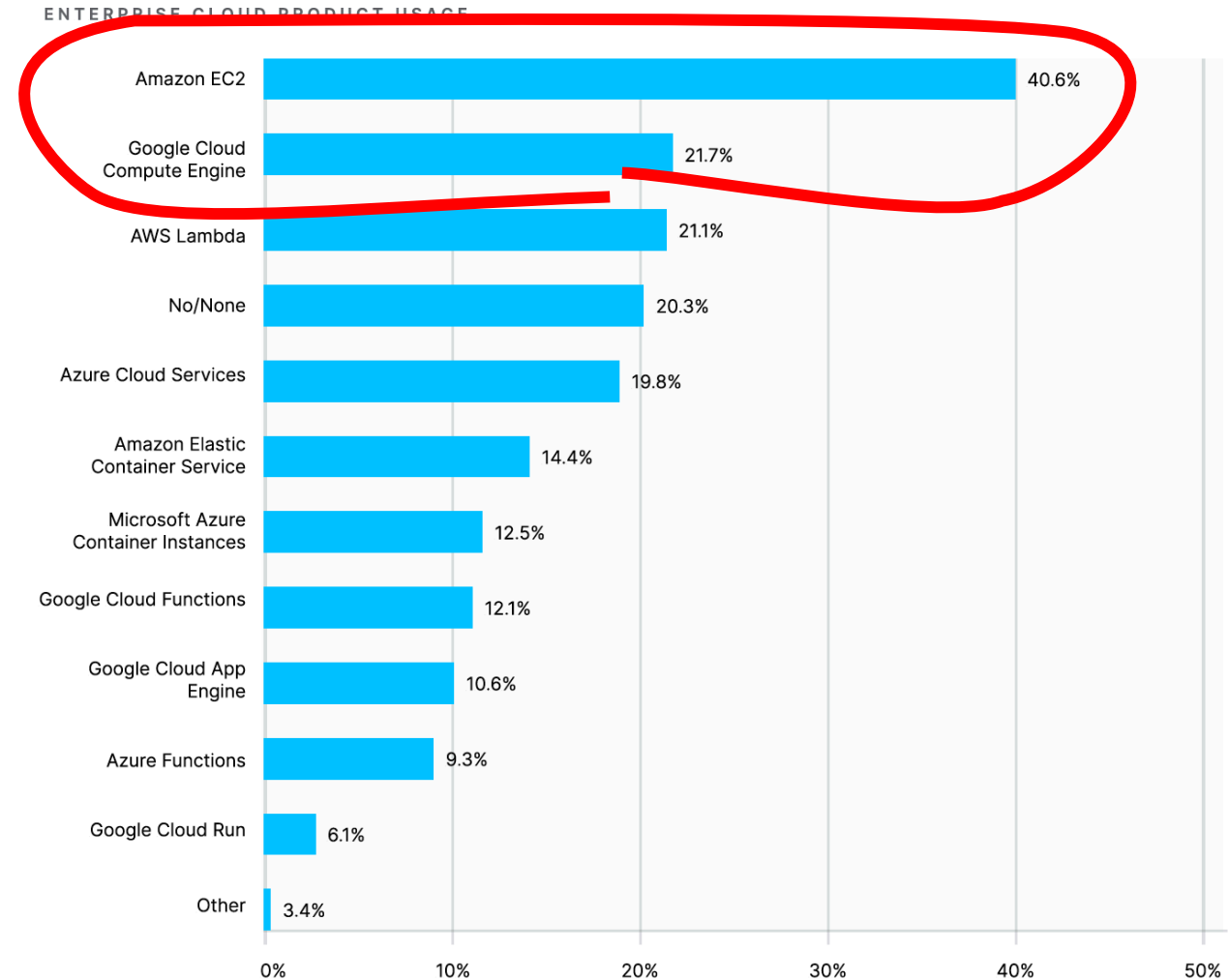


A Data Science Story

Q27-A

Do you use any of the following cloud computing products on a regular basis?

- [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



A Data Science Story

Q32

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

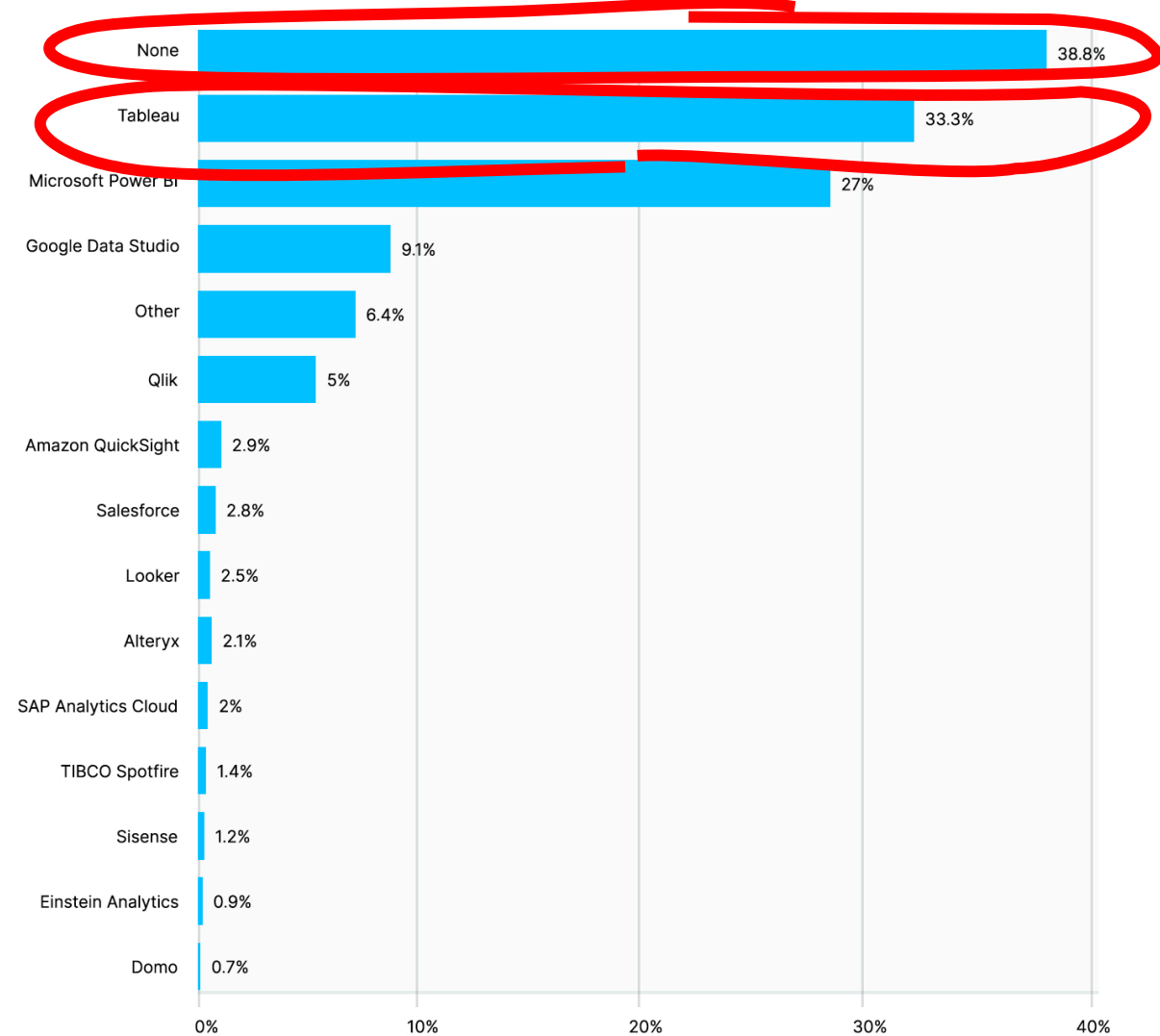
- » Amazon QuickSight
- » Microsoft Power BI
- » Google Data Studio
- » Looker
- » Tableau
- » Salesforce
- » Einstein Analytics
- » Qlik
- » Domo
- » TIBCO Spotfire
- » Alteryx
- » Sisense
- » SAP Analytics Cloud
- » None
- » Other



Answers are all very specific....

<https://www.kaggle.com/c/kaggle-survey-2020/data>

DATA SCIENTIST USAGE OF BUSINESS INTELLIGENCE TOOLS

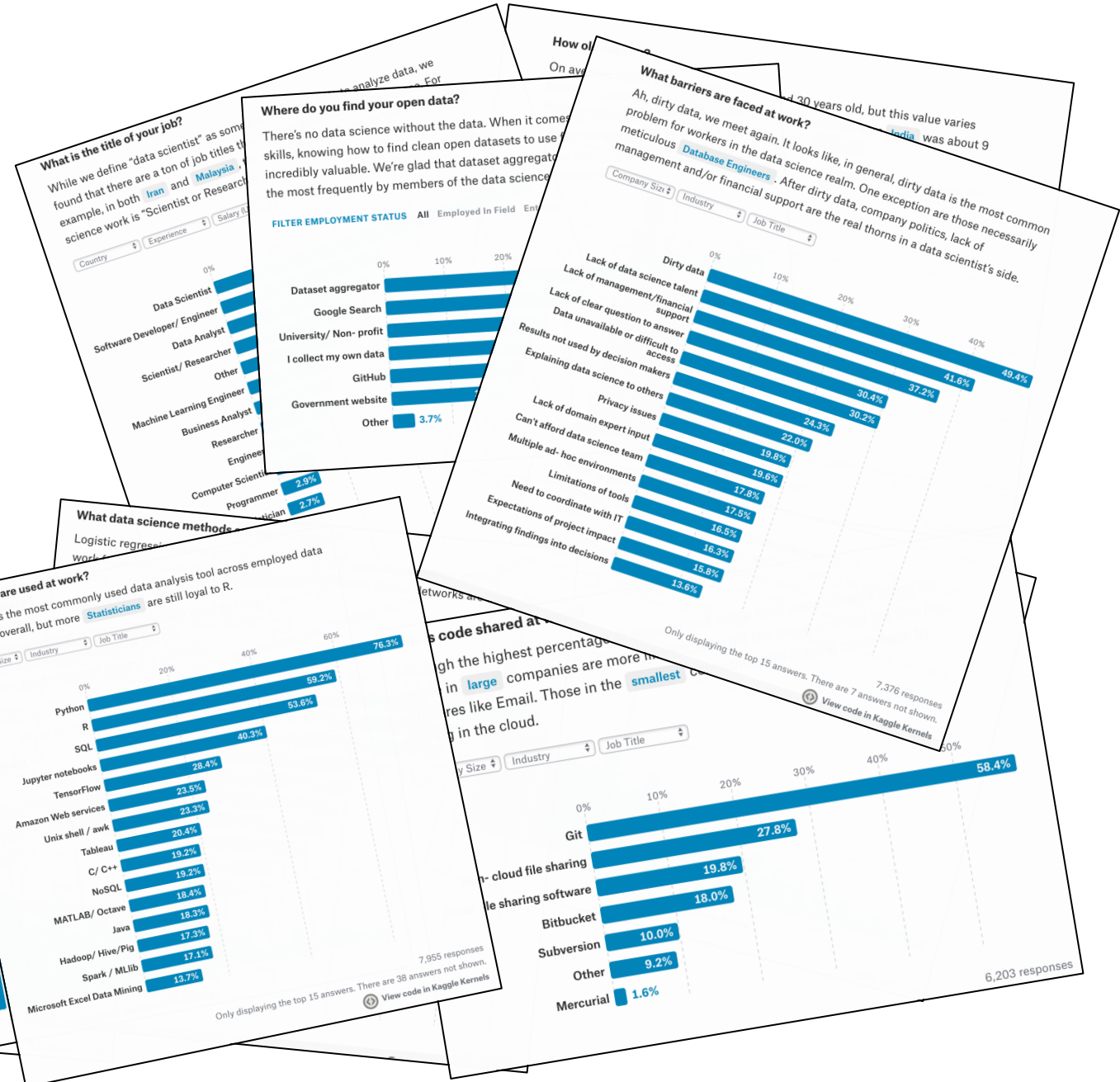


A Data Science Story



kaggle

2017 The State of Data Science & Machine Learning



What is the title of your job?

While we define "data scientist" as some people, we found that there are a ton of job titles that fit the example, in both [Iran](#) and [Malaysia](#). For example, in both [Iran](#) and [Malaysia](#), the most common science work is "Scientist or Researcher".

Job Title	Percentage
Scientist/ Researcher	37.2%
Software Developer/ Engineer	29.9%
Data Analyst	27.7%
Machine Learning Engineer	2.9%
Business Analyst	2.7%
Researcher	2.9%
Engineer	2.7%
Computer Scientist	2.9%
Programmer	2.7%
Other	2.7%

Where do you find your open data?

There's no data science without the data. When it comes to finding open data, skills, knowing how to find clean open datasets to use are incredibly valuable. We're glad that dataset aggregators are the most frequently by members of the data science community.

Source	Percentage
Dataset aggregator	49.4%
Google Search	37.2%
University/ Non-profit	30.4%
I collect my own data	30.4%
GitHub	30.4%
Government website	30.4%
Other	3.7%

What barriers are faced at work?

Ah, dirty data, we meet again. It looks like, in general, dirty data is the most common problem for workers in the data science realm. One exception are those necessarily meticulous [Database Engineers](#). After dirty data, company politics, lack of management and/or financial support are the real thorns in a data scientist's side.

Barrier	Percentage
Dirty data	49.4%
Lack of data science talent	41.6%
Lack of management/financial support	37.2%
Data unavailable or difficult to access	30.4%
Results not used by decision makers	30.4%
Explaining data science to others	30.4%
Privacy issues	24.3%
Lack of domain expert input	22.0%
Can't afford data science team	19.8%
Multiple ad-hoc environments	19.6%
Limitations of tools	17.8%
Need to coordinate with IT	17.5%
Expectations of project impact	16.5%
Integrating findings into decisions	16.3%
Other	15.8%

What data science methods are used at work?

Logistic regression is the most commonly used data analysis tool across employed data scientists overall, but more [Statisticians](#) are still loyal to R.

Method	Percentage
Logistic regression	76.3%
Linear regression	59.2%
Machine Learning Ensemble	53.6%
SQL	40.3%
Jupyter notebooks	28.4%
TensorFlow	23.5%
Amazon Web services	23.3%
Unix shell / awk	20.4%
Tableau	19.2%
C/ C++	19.2%
NoSQL	19.2%
MATLAB/ Octave	18.4%
Java	16.3%
Hadoop/ Hive/Pig	17.3%
Spark/ MLlib	17.1%
Microsoft Excel Data Mining	13.7%

What tools are used at work?

Python was the most commonly used data analysis tool across employed data scientists overall, but more [Statisticians](#) are still loyal to R.

Tool	Percentage
Python	76.3%
R	59.2%
SQL	40.3%
Jupyter notebooks	28.4%
TensorFlow	23.5%
Amazon Web services	23.3%
Unix shell / awk	20.4%
Tableau	19.2%
C/ C++	19.2%
NoSQL	19.2%
MATLAB/ Octave	18.4%
Java	16.3%
Hadoop/ Hive/Pig	17.3%
Spark/ MLlib	17.1%
Microsoft Excel Data Mining	13.7%

What is your employment status?

Status	Percentage
Employed full-time	12.7%
Not employed, but looking for work	8.0%
Freelancer	5.6%
Not employed and not looking	5.5%
Employed part-time	2.5%
I prefer not to say	2.5%

What type of data is used at work?

Relational data is the most commonly used data type across employed data scientists, except for [Academia](#) and the [Healthcare](#) industry, where text data is used more.

Data Type	Percentage
Relational data	18.1%
Text data	10.3%
Image data	5.1%
Other	5.1%
Video data	5.1%

What code is shared at work?

Git is the most commonly used code sharing tool across employed data scientists, with the highest percentage of users in [large](#) companies are more likely to use it. Those in the [smallest](#) companies are more likely to use [Mercurial](#).

Tool	Percentage
Git	58.4%
Bitbucket	27.8%
Other cloud file sharing	19.8%
Code sharing software	18.0%
Bitbucket	10.0%
Subversion	9.2%
Other	1.6%
Mercurial	1.6%

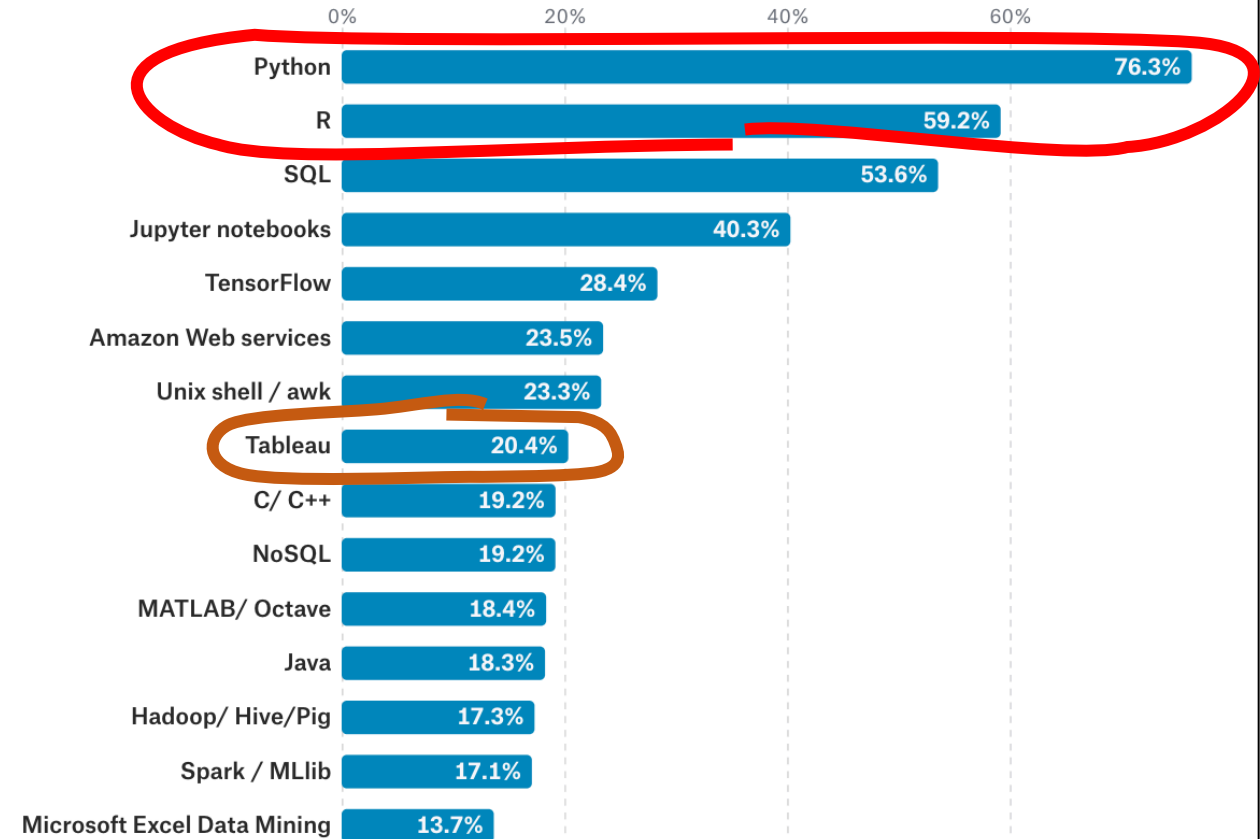
A Data Science Story



What tools are used at work?

Python was the most commonly used data analysis tool across employed data scientists overall, but more **Statisticians** are still loyal to R.

Company Size Industry Job Title

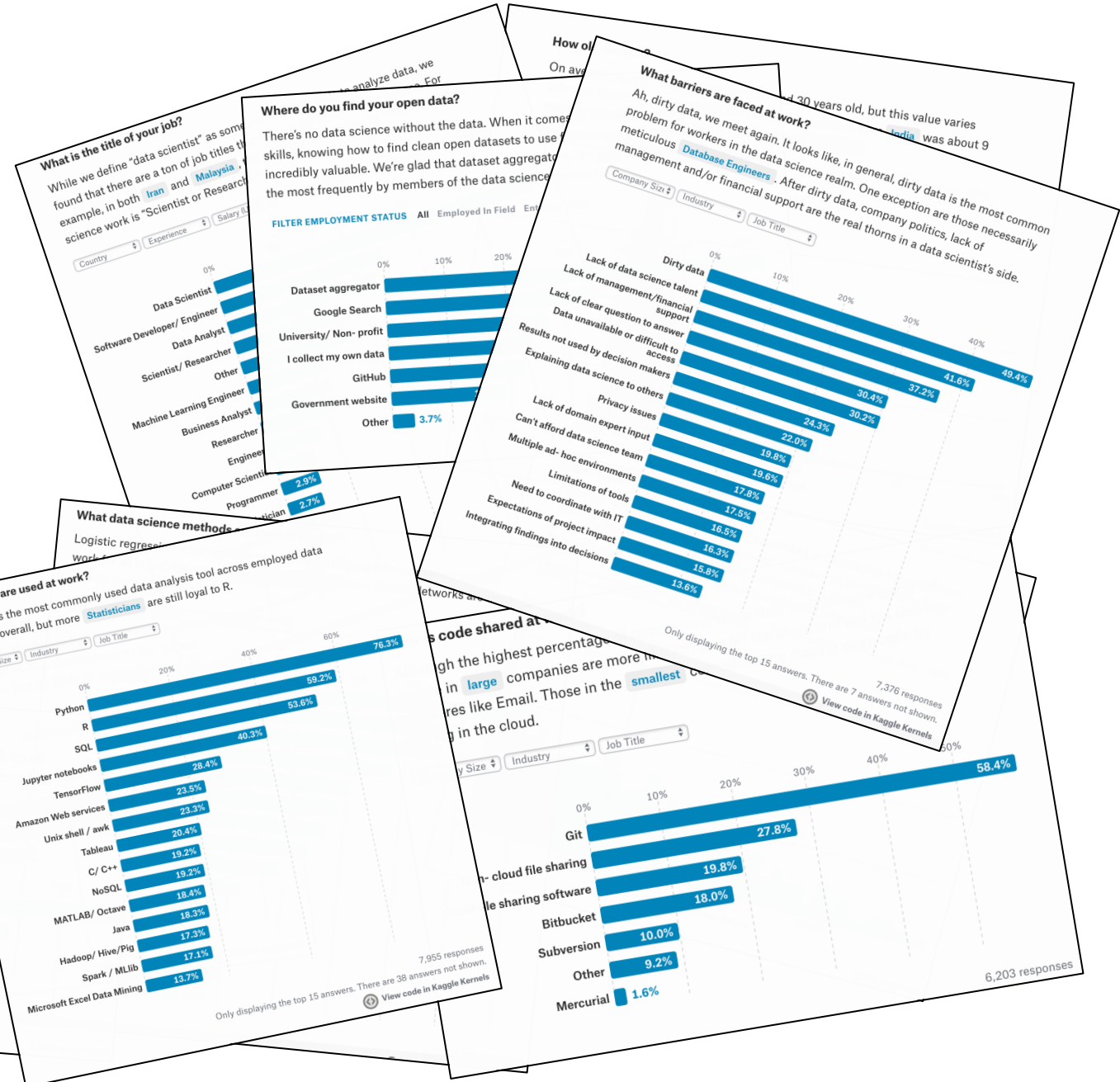


7,955 responses

Only displaying the top 15 answers. There are 38 answers not shown.

[View code in Kaggle Kernels](#)

A Data Science Story



A Data Science Story



Can I generate these myself?

Painlessly?



What barriers are faced at work?

Ah, dirty data, we meet again. It looks like, in general, dirty data is the most common problem for workers in the data science realm. One exception are those necessarily meticulous **Database Engineers**. After dirty data, company politics, lack of management and/or financial support are the real thorns in a data scientist's side.

Company Size Industry Job Title



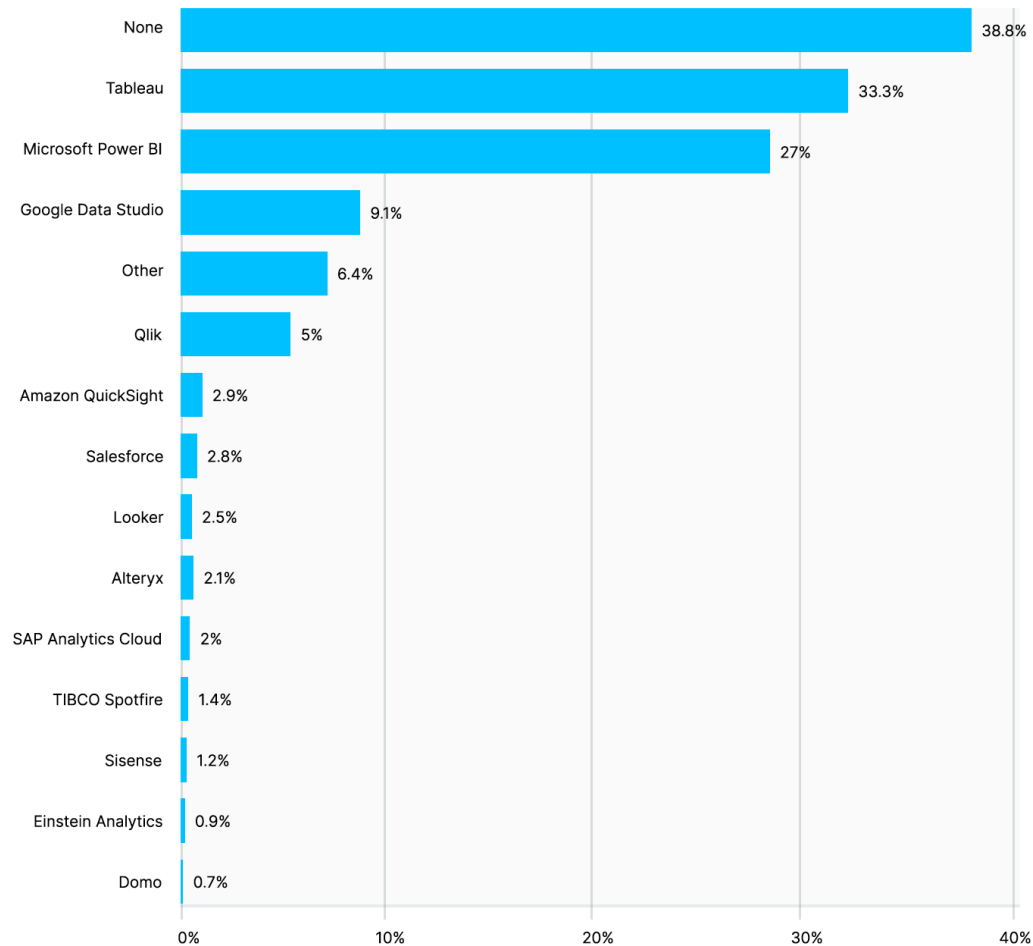
7,376 responses

Only displaying the top 15 answers. There are 7 answers not shown.

[View code in Kaggle Kernels](#)

Let's See...

DATA SCIENTIST USAGE OF BUSINESS INTELLIGENCE TOOLS

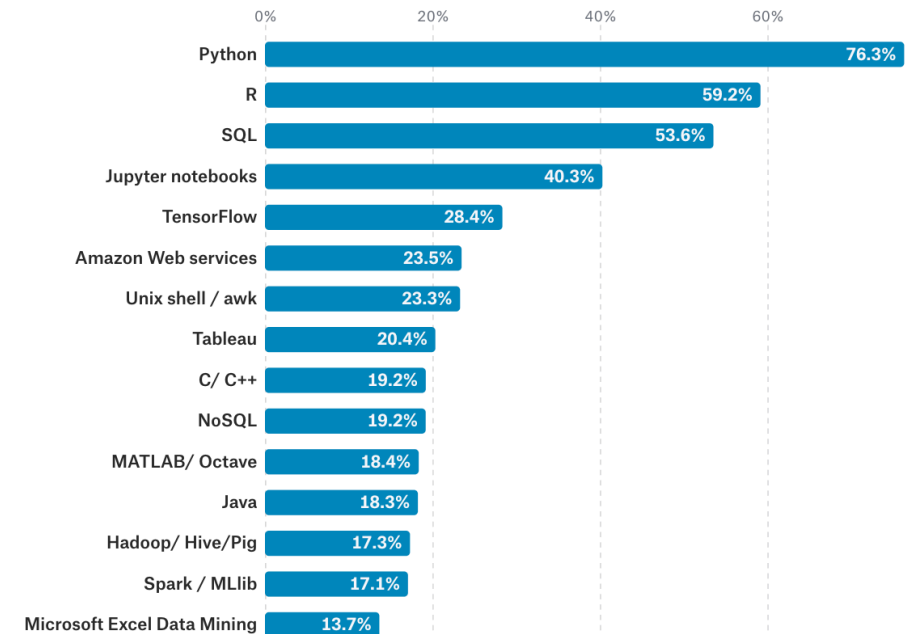


2020

What tools are used at work?

Python was the most commonly used data analysis tool across employed data scientists overall, but more **Statisticians** are still loyal to R.

Company Size ▾ Industry ▾ Job Title ▾



7,955 responses

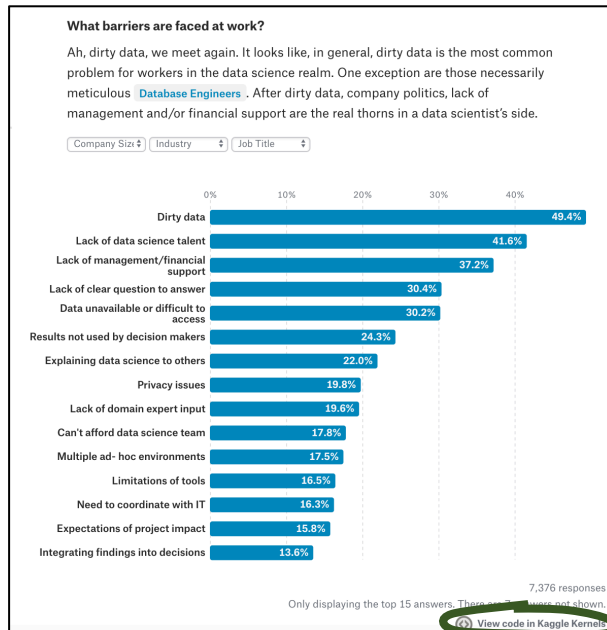
Only displaying the top 15 answers. There are 38 answers not shown.

[View code in Kaggle Kernels](#)

2017

Obstacles in code-based data analysis

- Reading/cleaning/aggregating data
- Learning/deciphering syntax

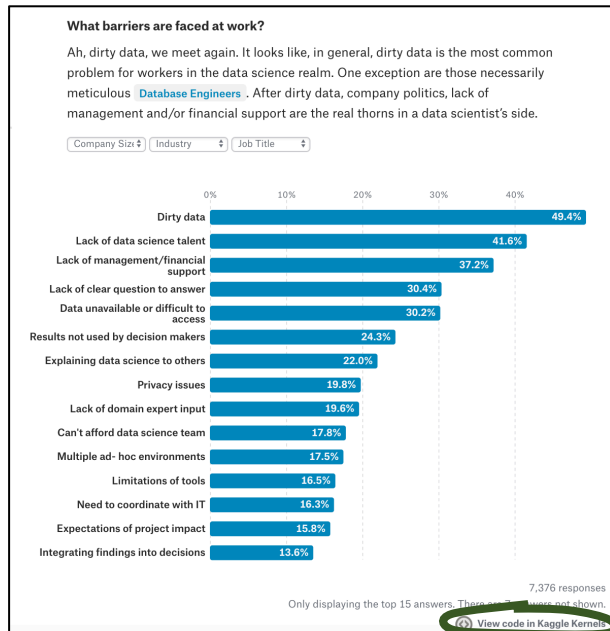


```
136 chooseMultiple = function(question, filteredData = cleanData){
137
138   filteredData %>%
139     # Remove any rows where the respondent didn't answer the question
140     filter(!UQ(sym(question)) == "") %>%
141     # Remove all columns except question
142     select(question) %>%
143     # Add a column with the initial number of respondents to question
144     mutate(totalCount = n()) %>%
145     # Split multiple answers apart at the comma, but ignore commas inside parentheses
146     mutate(selections = strsplit(as.character(UQ(sym(question))),
147                                 '\\([^\]]+,(*SKIP)(*FAIL)|,\\s*', perl = TRUE)) %>%
148     # Split answers are now nested, need to unnest them
149     unnest(selections) %>%
150     # Group by the selected responses to the question
151     group_by(selections) %>%
152     # Count how many respondents selected each option
153     summarise(totalCount = max(totalCount,
154                               count = n()) %>%
155     # Calculate what percent of respondents selected each option
156     mutate(percent = (count / totalCount) * 100) %>%
157     # Arrange the counts in descending order
158     arrange(desc(count))
159
160 }
```

```
255 # Filter the data
256 filterBarriers <- workLife %>%
257   # Remove blank responses on employment question
258   filter(!EmploymentStatus == "") %>%
259   # Keep only entries that indicated that they use code to analyze data at work
260   filter(CodeWriter == "Yes") %>%
261   # Keep only entries that included one of the above "employed" statuses
262   filter(grepl(paste(employed, collapse = "|"), EmploymentStatus))
263
264 # Using the filtered data, run chooseMultiple() function
265 chooseMultiple("WorkChallengesSelect", filterBarriers)
```

Obstacles in code-based data analysis

- Reading/cleaning/aggregating data
- Learning/deciphering syntax
- Code and deliverable mismatch



selections
<chr>
Dirty data
Lack of data science talent in the organization
Company politics / Lack of management/financial support for a data science team
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
Data Science results not used by business decision makers
Explaining data science to others
Privacy issues
Lack of significant domain expert input
Organization is small and cannot afford a data science team

1-10 of 22 rows | 1-1 of 4 columns

Previous 1 2 3 Next

totalCount	co...	percent
<dbl>	<int>	<dbl>
7376	3641	49.362798
7376	3067	41.580803
7376	2746	37.228850
7376	2242	30.395879
7376	2230	30.233189
7376	1796	24.349241
7376	1622	21.990239
7376	1460	19.793926
7376	1444	19.577007
7376	1316	17.841649

1-10 of 22 rows | 2-4 of 4 columns

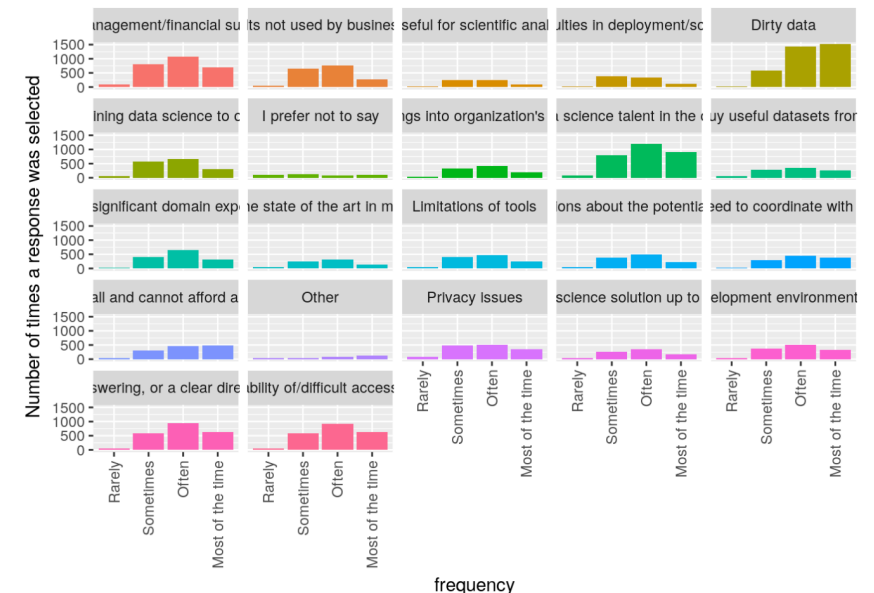
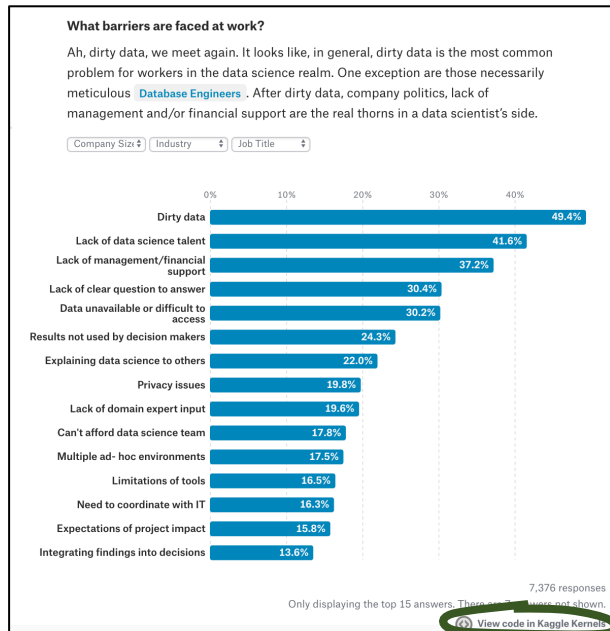
Previous 1 2 3 Next

Obstacles in code-based data analysis

- Reading/cleaning/aggregating data
- Learning/deciphering syntax
- Code and deliverable mismatch
- Formatting output

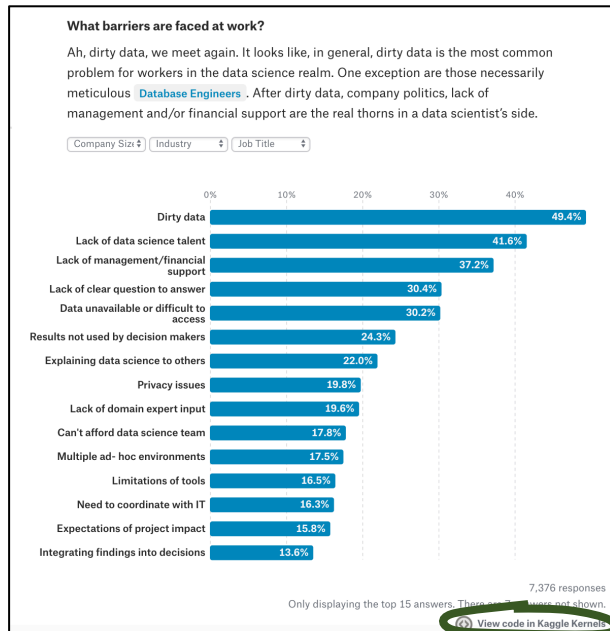
```
# Plot
ggplot(challengeNamesChar, aes(x = frequency, y = count, fill = response.y)) +
  geom_bar(stat = "identity") +
  facet_wrap(~response.y) +
  ylab("Number of times a response was selected") +
  theme(legend.position="none") +
  theme(axis.text.x = element_text(angle = 90,
                                    vjust = 0.5,
                                    hjust = 1))

# Since the names are often too long to be displayed well in this figure, print
levels(as.factor(challengeNamesChar$response.y))
```



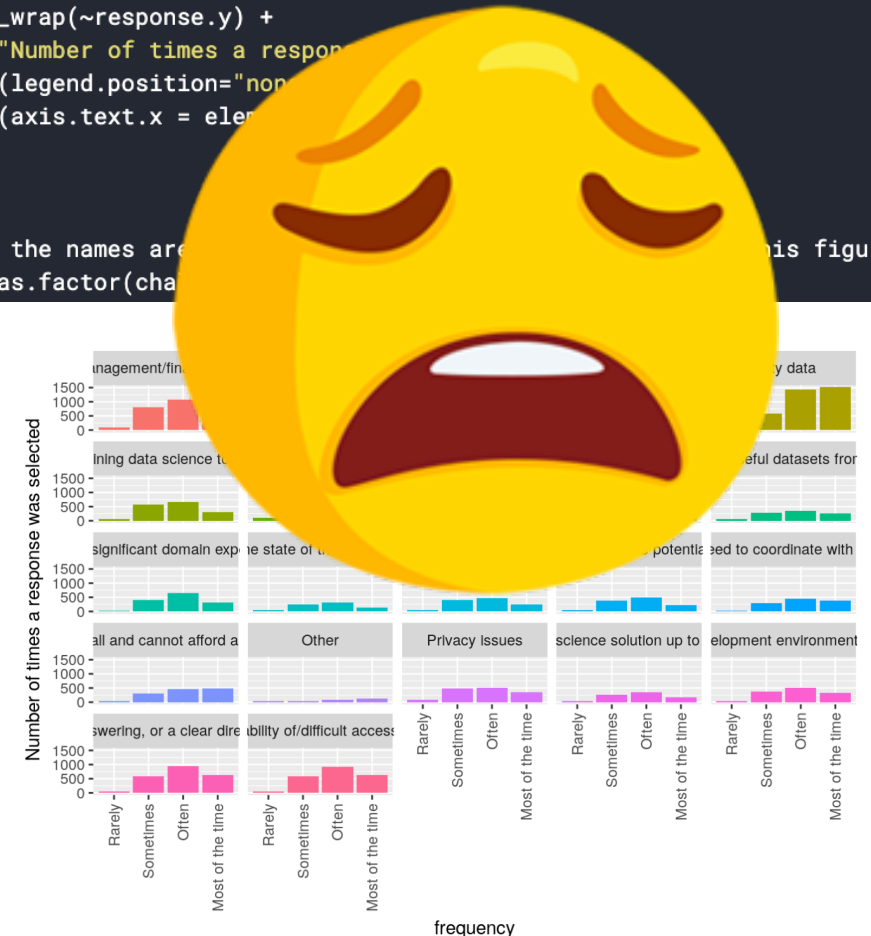
Obstacles in code-based data analysis

- Reading/cleaning/aggregating data
- Learning/deciphering syntax
- Code and deliverable mismatch
- Formatting output



```
# Plot
ggplot(challengeNamesChar, aes(x = frequency, y = count, fill = response.y)) +
  geom_bar(stat = "identity") +
  facet_wrap(~response.y) +
  ylab("Number of times a response was selected") +
  theme(legend.position="none") +
  theme(axis.text.x = element_text(angle=45))

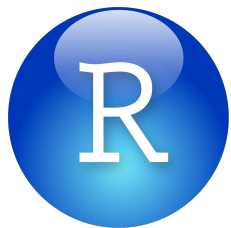
# Since the names are long, to make this figure, print
levels(as.factor(challengeNamesChar$response.y))
```



Misalignment of goals

Data analysis for dissemination

- **Flexible** framework for implementing complex calculations
- **Concise** representations like scripts and functions
- Facilitates the efficient **communication** of insights.

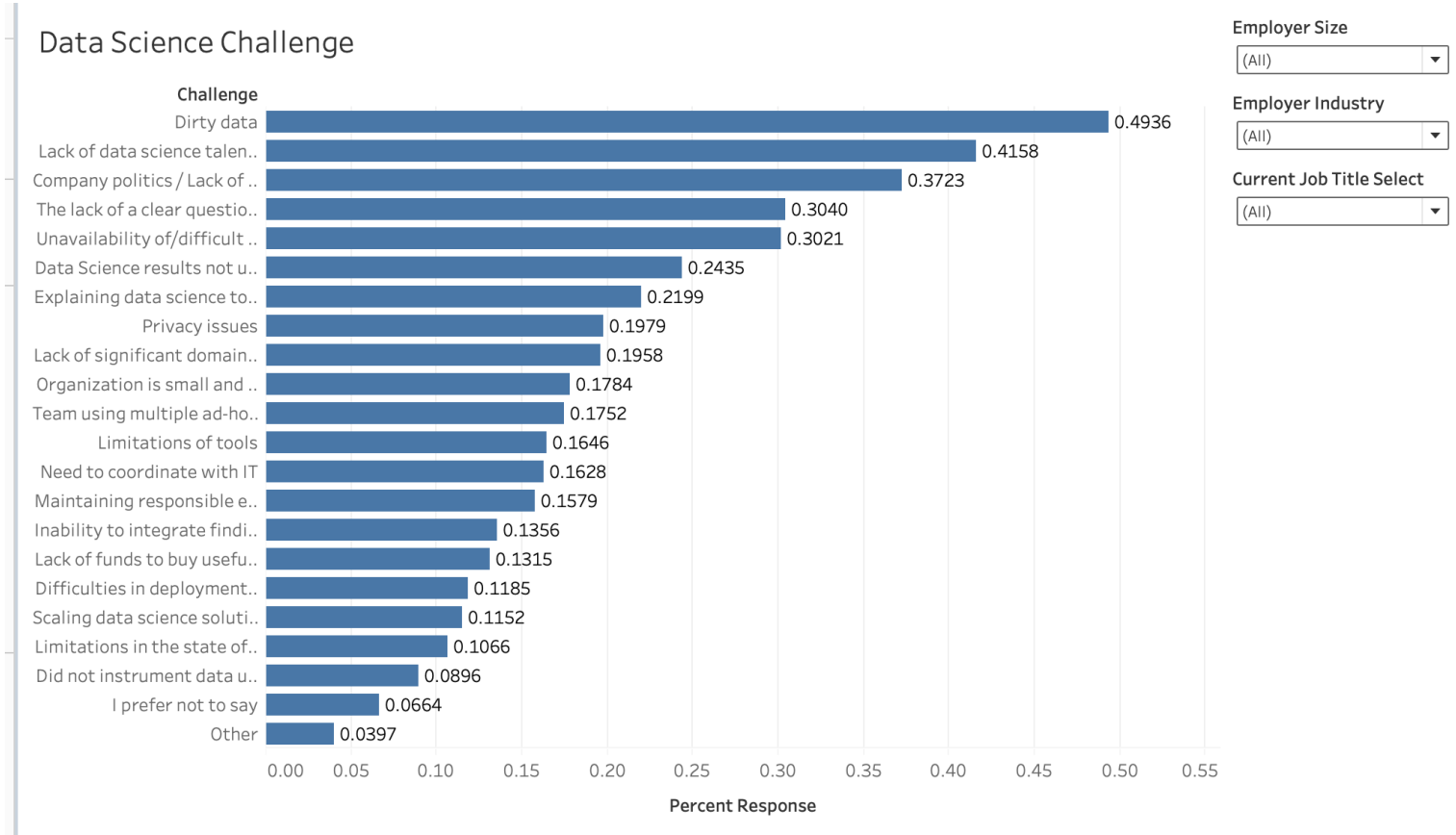


Data analysis for exploration

- **Fast** manipulation of data
- **Intuitive** interface
- Facilitates the efficient **discovery** of insights.



Demo 1: Data Science Challenges



<https://stats285.github.io/>

Prevalence of neural collapse during the terminal phase of deep learning training

✉ Vardan Papyan, ✉ X. Y. Han, and David L. Donoho

+ See all authors and affiliations

PNAS October 6, 2020 117 (40) 24652-24663; first published September 21, 2020;

<https://doi.org/10.1073/pnas.2015509117>

Demo:
Visualizing
Research

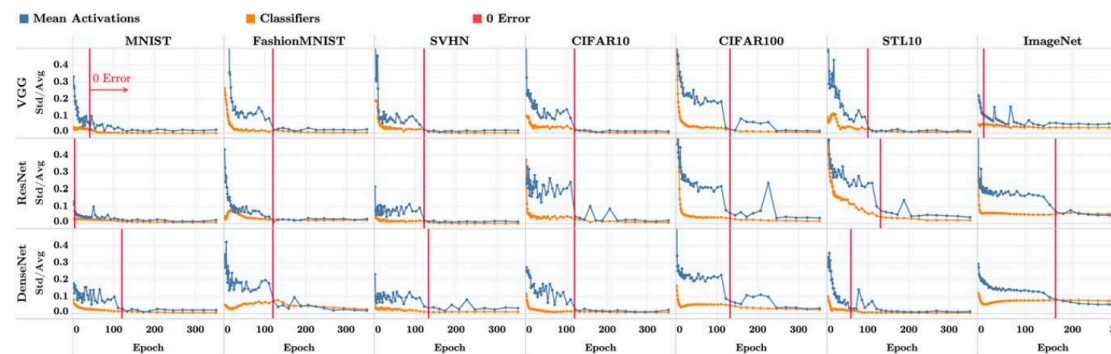


Fig. 2. Train class means become equinorm. The formatting and technical details are as described in Section 3. In each array cell, the vertical axis shows the coefficient of variation of the centered class-mean norms as well as the network classifiers norms. In particular, the blue lines show $\text{Std}_c(\|\mu_c - \mu_G\|_2) / \text{Avg}_c(\|\mu_c - \mu_G\|_2)$ where $\{\mu_c\}$ are the class means of the last-layer activations of the training data and μ_G is the corresponding train global mean; the orange lines show $\text{Std}_c(\|w_c\|_2) / \text{Avg}_c(\|w_c\|_2)$ where w_c is the last-layer classifier of the c th class. As training progresses, the coefficients of variation of both class means and classifiers decrease.

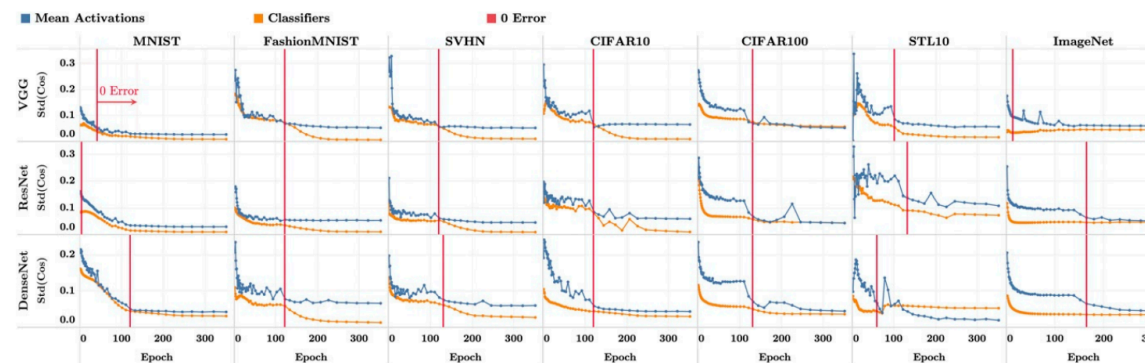


Fig. 3. Classifiers and train class means approach equiangularity. The formatting and technical details are as described in Section 3. In each array cell, the vertical axis shows the SD of the cosines between pairs of centered class means and classifiers across all distinct pairs of classes c and c' . Mathematically, denote $\text{cos}_\mu(c, c') = \langle \mu_c - \mu_G, \mu_{c'} - \mu_G \rangle / (\|\mu_c - \mu_G\|_2 \|\mu_{c'} - \mu_G\|_2)$ and $\text{cos}_w(c, c') = \langle w_c, w_{c'} \rangle / (\|w_c\|_2 \|w_{c'}\|_2)$ where $\{w_c\}_{c=1}^C$, $\{\mu_c\}_{c=1}^C$, and μ_G are as in Fig. 2. We measure $\text{Std}_{c,c' \neq c}(\text{cos}_\mu(c, c'))$ (blue) and $\text{Std}_{c,c' \neq c}(\text{cos}_w(c, c'))$ (orange). As training progresses, the SDs of the cosines approach zero, indicating equiangularity.

<https://purl.stanford.edu/ng812mz4543>