> DATA SCIENTIST: THE SEXIEST JOB OF THE 21ST CENTURY
>
> — HARVARD BUSINESS REVIEW

CHALLENGE

Warning: We suggest you use Chrome(https://www.google.com/chrome/) as your browser (possibly using Incognito Mode) if you experience any errors.

Please answer as many questions as you can. We do not expect you to answer them all, but **you must answer at least one for each section.** Answering more questions correctly will help you and answering them incorrectly will not hurt you. **Please give all numerical answers to 10 digits of precision. Partial credit will be given to answers that agree to less than 10 digits.** (*) denotes a required field. Due to the volume of requests, we will only accept submissions via this form. The basic ground rules are:

- **Answer the questions yourself without asking others for assistance**. This is a test of your ability to answer realistic questions. You will be asked questions of similar difficulty during the phone interview so cheating will not help.
- **Do not share the questions or your answers with anyone.** This includes posting the questions or your solutions publicly on services like quora, stackoverflow, or github. Doing so gives others an unfair advantage and may also disqualify you from this or future fellowships.
- **Submit early.** We highly recommend aiming to submit the answers well ahead of the deadline. Every quarter, a number of "unforeseeable" technical difficulties have prohibited otherwise highly-qualified last-minute applicants from submitting. Don't be a statistic.
- **Submit often.** You can submit your challenge solutions as often as you would like. Only the last submitted challenge is kept so we recommend you submit your answers as you complete them.

A few helpful hints:

1. **Want to get a head start on being a data scientist?** We want all semifinalists to get as much out of the challenge questions as possible. So we've written three(http://blog.thedataincubator.com/2015/09/painlessly-deploying-data-apps-with-bokeh-flask-and-heroku/) blog(http://blog.thedataincubator.com/2015/01/processing-

data-like-a-professional-data-scientist/) posts(http://blog.thedataincubator.com/2015/01/a-cs-degree-for-data-science-part-i-efficient-numerical-computation/) that might get you thinking about mathematics and computation differently. They will also give you a head start on solving the challenge questions. For additional hints on the challenge, follow us on Twitter(http://twitter.com/intent/user?screen_name=thedatainc), LinkedIn(https://www.linkedin.com/company/the-data-incubator), and Facebook(https://www.facebook.com/dataincubator/).

2. **Having browser troubles?** We recommend using Chrome(https://www.google.com/chrome/) (possibly using Incognito Mode).
3. **Having trouble downloading any files?** We suggest using command-line tools, rather than relying on a browser.
4. **Found something ambiguous?** We realize some questions are ambiguous. Most real-world questions are. This is a test of whether you can prioritize important effects and combine real-world knowledge with theory.
5. **Questions a little too difficult?** You might want to consider signing up for our online data science foundations class(/foundations.html), which teaches the pre-requisite material needed for the fellowship.

Section 1: Members of Congress and Congressional offices receive an annual budget to spend on staff, supplies, transportation, and other expenses. Each quarter, representatives report the recipients of their expenditures. ProPublica compiles these reports into research-ready CSV files. Download the full data set.(https://www.propublica.org/datastore/dataset/house-office-expenditures) We will study the detailed (not summary) data. The full data set includes a readme text file describing the data in more detail, which may be helpful in completing this challenge. Note that there is an updated version of the 2015Q2 file in the ZIP archive; you should use this and discard the original.

What is the total of all the payments in the dataset?

1.234567890

Define the 'COVERAGE PERIOD' for each payment as the difference (in days) between 'END DATE' and 'START DATE'. What is the standard deviation in 'COVERAGE PERIOD'? Only consider payments with strictly positive amounts.

1.234567890

What was the average annual expenditure with a 'START DATE' date between January 1, 2010 and December 31, 2016 (inclusive)? Only consider payments with strictly positive amounts.

1.234567890

Find the 'OFFICE' with the highest total expenditures with a 'START DATE' in 2016. For this office, find the 'PURPOSE' that accounts for the highest total expenditures. What fraction of the total expenditures (all records, all offices) with a 'START DATE' in 2016 do these expenditures amount to?

0.9876543210

What was the highest average staff salary among all representatives in 2016? Assume staff sizes is equal to the number of unique payees in the 'PERSONNEL COMPENSATION' category for each representative.

1.234567890

What was the median rate of annual turnover in staff between 2011 and 2016 (inclusive)? Turnover for 2011 should be calculated as the fraction of a representative's staff from 2010 who did not carry over to 2011. Only consider representatives who served for at least 4 years and had staff size of at least 5 every year that they served.

0.9876543210

What percentage of the expenditures of the top 20 spenders in 2016 come from members of the Democratic Party? Representatives are identified by their 'BIOGUIDE_ID', which can be used to look up representatives with ProPublica's Congress API(https://projects.propublica.org/api-docs/congress-api/members/#get-a-specific-member) to find their party affiliation. Consider an expenditure as being in 2016 if its 'START DATE' is in 2016.

0.9876543210

Please provide the script used to generate this result (max 10000 characters).

In what language is the script written?

◯ C/C++	◯ Fortran	◯ IDL	◯ Java
◯ MATLAB	◯ Perl	◯ Python	◯ R
◯ Stata	◯ SQL	◯ VBA	◯ Other

Section 2: You have a deck with N different playing cards, equally distributed amongst M suits. You draw all cards without putting any back in the deck. After drawing the first card, you compare the suit of each subsequent card drawn with the suit of the card drawn immediately before. If the suits match, you get a point. Otherwise, you get no points. Please answer the following questions about the total number of points, P, at the end of the process.

What is the mean of P when $N = 26$ and $M = 2$?

1.234567890

What is the standard deviation of P when $N = 26$ and $M = 2$?

1.234567890

What is the mean of P when $N = 52$ and $M = 4$?

1.234567890

What is the standard deviation of P when $N = 52$ and $M = 4$?

1.234567890

What is the conditional probability that P is greater than 12 given that it is greater than 6 when $N = 26$ and $M = 2$?

0.9876543210

What is the conditional probability that P is greater than 12 given that it is greater than 6 when $N = 52$ and $M = 4$?

0.9876543210

Please provide the script used to generate this result (max 10000 characters).

In what language is the script written?

- ◯ C/C++
- ◯ Fortran
- ◯ IDL
- ◯ Java
- ◯ MATLAB
- ◯ Perl
- ◯ Python
- ◯ R
- ◯ Stata
- ◯ SQL
- ◯ VBA
- ◯ Other

Section 3: This section is required.

Propose a project to do while at The Data Incubator. We want to know about your ability to think at a high level. Try to think of projects that users or businesses will care about that are also relatively unanalyzed. Here are some useful links about data sources on our blog(http://blog.thedataincubator.com/tag/data-sources/) as well as the archive of data

sources on Data is Plural(http://tinyletter.com/data-is-plural/archive). You can see some final projects of previous Fellows on our YouTube Page(https://www.youtube.com/playlist?list=PLOE4k9MRzZanWmZ7MBrJFi7ZekYmVqEIV).

Propose a project that uses a large, publicly accessible dataset. Explain your motivation for tackling this problem, discuss the data source(s) you are using, and explain the analysis you are performing. At a minimum, you will need to do enough exploratory data analysis to convince someone that the project is viable and generate two interesting non-trivial plots supporting this. **The most impressive applicants have even finished a "rough draft" of their projects and have derived non-obvious meaningful conclusions from their data.** Explain the plots and give url links to them. For guidance on how to choose a project, check out this blog post(http://blog.thedataincubator.com/2017/01/how-employers-judge-data-science-projects/).

Propose a project.*

Link to public description of data source.*

http://blog.thedataincubator.com/tag/data-sources/

Link to 1st plot. You are highly encouraged to use Heroku apps domain(https://www.heroku.com/) for an app or Github(https://www.github.com/) to display a notebook.*

https://example.herokuapp.com/

Link to 2nd plot. You are highly encouraged to use Heroku apps domain(https://www.heroku.com/) for an app or Github(https://www.github.com/) to display a notebook.*

https://example.herokuapp.com/

How much data did you analyze (in MB)?*

1234

How did you obtain your dataset? (Please check all that apply.)

- ☐ I downloaded a dataset available online.
- ☐ I used a provided API.
- ☐ I scraped data from a webpage.
- ☐ Other (please explain).

We want to know your communication style. Record a video of yourself giving a high-level proposal of your project to a non-technical person. The video should be no longer than 1 minute and should be at a higher level than the previous explanation.

Record a video of yourself and upload it to YouTube(https://support.google.com/youtube/answer/57407) (and not another video hosting service). Be sure to make the video unlisted (but not private!) so people without the link cannot find it on Google (go here(https://www.youtube.com/my_videos), click "Edit" on your video, select unlisted from the privacy dropdown menu(static/images/youtube-unlisted.png), and save your changes). You can use either your webcam or a smartphone.

Once complete, please provide the **embed** URL of the video. To find this URL (**NOT** the entire iframe tag), on the video's normal watch page, you can click Share → Embed(/static/images/embed.png), and take the link from inside the 'src' attribute of the tag. It looks something like this: https://www.youtube.com/embed/y9tX5whl2U

For more detailed instructions, including screenshots, click here(/video-upload.html).

Please provide the EMBED URL to your video*

https://www.youtube.com/embed/y9tX5whl2U

Note: youtube videos take some time to process after uploading, and your video won't validate until processing is complete. Please allow 10 to 15 minutes for this to take place.

Please provide the script used to generate this result (max 10000 characters).*

In what language is the script written?

- ◯ C/C++
- ◯ Fortran
- ◯ IDL
- ◯ Java
- ◯ MATLAB
- ◯ Perl
- ◯ Python
- ◯ R
- ◯ Stata
- ◯ SQL
- ◯ VBA
- ◯ Other

For future challenge questions, how many hours did it take you to complete this challenge? This will not be considered in your application (please just enter a number).*

9999

☐ By submitting this form, you certify that your answers are the result of your own work and not copied from another individual or source. *

SUBMIT SAVE

You can save your work and return to this page at any point. Once you have filled out the required fields, your challenge submission will be considered 'complete'.

> “WITH LOADS OF DATA YOU WILL FIND RELATIONSHIPS THAT AREN’T REAL.
> BIG DATA ISN’T ABOUT BITS, IT’S ABOUT TALENT.”
>
> — FORBES MAGAZINE