> 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: The New York City Fire Department keeps a log of detailed information on incidents handled by FDNY units. In this challenge we will work with a dataset that contains a record of incidents handled by FDNY units from 2013-2017. Download the FDNY data set. (https://data.cityofnewyork.us/api/views/tm6d-hbzd/rows.csv?accessType=DOWNLOAD) Also take a look at the dataset landing page(https://data.cityofnewyork.us/Public-Safety/Incidents-Responded-to-by-Fire-Companies/tm6d-hbzd) and find descriptions of column names here. (https://data.cityofnewyork.us/api/views/tm6d-hbzd/files/1434d09c-fbf8-4450-8b42-9fe0c3b85fb3?download=true&filename=OPEN_DATA_FIRE_INCIDENTS_FILE_DESCRIPTION.xls)
What proportion of FDNY responses in this dataset correspond to the most common type of incident?

0.9876543210

What is the ratio of the average number of units that arrive to a scene of an incident classified as '111 - Building fire' to the number that arrive for '651 - Smoke scare, odor of smoke'?

1.234567890

How many times more likely is an incident in Staten Island a false call compared to in Manhattan? The answer should be the ratio of Staten Island false call rate to Manhattan false call rate. A false call is an incident for which 'INCIDENT_TYPE_DESC' is '710 - Malicious, mischievous false call, other'.

1.234567890

Check the distribution of the number of minutes it takes between the time a '111 - Building fire' incident has been logged into the Computer Aided Dispatch system and the time at which the first unit arrives on scene. What is the third quartile of that distribution. Note: the number of minutes can be fractional (ie, do not round).

1.234567890

We can use the FDNY dataset to investigate at what time of the day people cook most. Compute what proportion of all incidents are cooking fires for every hour of the day by normalizing the number of cooking fires in a given hour by the total number of incidents that occured in that hour. Find the hour of the day that has the highest proportion of cooking fires and submit that proportion of cooking fires. A cooking fire is an incident for which 'INCIDENT_TYPE_DESC' is '113 - Cooking fire, confined to container'. Note: round incident times down. For example, if an incident occured at 22:55 it occured in hour 22.

0.9876543210

What is the coefficient of determination (R squared) between the number of residents at each zip code and the number of inicidents whose type is classified as '111 - Building fire' at each of those zip codes. Note: The 2010 US Census population by zip code dataset should be downloaded from here. (https://s3.amazonaws.com/SplitwiseBlogJB/2010+Census+Population+By+Zipcode+(ZCTA).csv) You will need to use both the FDNY responses and the US Census dataset. Ignore zip codes that do not appear in the census table.

1.234567890

For this question, only consider incidents that have information about whether a CO detector was present or not. We are interested in how many times more likely it is that an incident is long when no CO detector is present compared to when a CO detector is present. For events with CO detector and for those without one, compute the proportion of incidents that lasted 20-30, 30-40, 40-50, 50-60, and 60-70 minutes (both interval boundary values included) by dividing the number of incidents in each time interval with the total number of incidents. For each bin, compute the ratio of the 'CO detector absent' frequency to the 'CO detector present' frequency. Perform a linear regression of this ratio to the mid-point of the bins. From this, what is the predicted ratio for events lasting 39 minutes?

1.234567890

Calculate the chi-square test statistic for testing whether an incident is more likely to last longer than 60 minutes when CO detector is not present. Again only consider incidents that have information about whether a CO detector was present or not.

1.234567890

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: A circular road has N positions labeled 0 through $N-1$ where adjacent positions are connected to each other and position $N-1$ is connected to 0. M cars start at position 0 through $M-1$ (inclusive). A car can make a valid move by moving forward one position (or goes from N-1 to 0) if the position it is moving into is empty. At each turn, only consider cars that have a valid move available and make one of the valid moves that you choose randomly with equal probability. After T rounds, we compute the average (A) and standard deviation (S) of the position of the cars.

What is the expected value of A when $N = 10$, $M = 5$, and $T = 20$?

1.234567890

What is the standard deviation of A when $N = 10$, $M = 5$, and $T = 20$?

1.234567890

What is the expected value of S when $N = 10$, $M = 5$, and $T = 20$?

1.234567890

What is the standard deviation of S when $N = 10$, $M = 5$, and $T = 20$?

1.234567890

What is the expected value of A when $N = 25$, $M = 10$, and $T = 50$?

1.234567890

What is the standard deviation of A when $N = 25$, $M = 10$, and $T = 50$?

1.234567890

What is the expected value of S when $N = 25$, $M = 10$, and $T = 50$?

1.234567890

What is the standard deviation of S when $N = 25$, $M = 10$, and $T = 50$?

1.234567890

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