“ DATA SCIENTIST: THE SEXIEST JOB OF THE 21ST CENTURY ”

— HARVARD BUSINESS REVIEW

CHALLENGE

Warning: We suggest you use Chrome(https://www.google.com/chrome/browser/desktop/index.html) 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 all the questions (they are mostly optional) but answering more questions correctly will help 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.** You can resubmit your answers on this form as often as you would like. Only the latest submission will be considered. (*) denotes a required field. 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/browser/desktop/index.html) (possibly using Incognito Mode).
3. **Having trouble downloading any files?** We suggest using command-line tools, rather than relying on a browser.

4. **Want to avoid being a statistic?** Every application cycle, a number of applicants wait until the last minute to submit, only to discover "unforeseeable" last-minute glitches that prevent submission. We suggest not waiting until the deadline to submit.
5. **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.
6. Due to the volume of requests, we will only accept submissions via this form.

Q1:

You roll N fair six-sided dice. The sum of the values is M. We wish to know about the product of the faces.

If $N = 8$ and $M = 24$, what is the expected value of the product?

123.456790

If $N = 8$ and $M = 24$, what is the standard deviation of the product?

98.76543210

If $N = 50$ and $M = 150$, what is the expected value of the product?

12345678.90

If $N = 50$ and $M = 150$, what is the standard deviation of the product?

98765432.10

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

Q2:

The Citi Bike Program(https://www.citibikenyc.com/) is a bike-sharing system in New York City. Cyclists can rent a bicycle from one of many stations around the city and return it to any other station. Citi Bike has released ridership information to the public. This contains a list of all rides taken, their start and end locations, their start and end times, and limited demographic information about the rider. Zip files containing the data for each month are available at https://s3.amazonaws.com/tripdata/index.html(https://s3.amazonaws.com/tripdata/index.html). We will be considering the data from 2015 only.

What is the median trip duration, in seconds?

123.4567890

What fraction of rides start and end at the same station?

0.987654321

We say a bike has visited a station if it has a ride that either started or ended at that station. Some bikes have visited many stations; others just a few. What is the standard deviation of the number of stations visited by a bike?

12.34567890

What is the average length, in kilometers, of a trip? Assume trips follow great circle arcs(https://en.wikipedia.org/wiki/Great_circle) from the start station to the end station. Ignore trips that start and end at the same station, as well as those with obviously wrong data.

9.876542310

Calculate the average duration of trips for each month in the year. (Consider a trip to occur in the month in which it starts.) What is the difference, in seconds, between the longest and shortest average durations?

123.4567890

Let us define the *hourly usage fraction* of a station to be the fraction of all rides starting at that station that leave during a specific hour. A station has surprising usage patterns if it has an hourly usage fraction for an hour significantly different from the corresponding hourly usage fraction of the system as a whole. What is the largest ratio of station hourly usage fraction to system hourly usage fraction (hence corresponding to the most "surprising" station-hour pair)?

98.76543210

There are two types of riders: "Customers" and "Subscribers." (https://www.citibikenyc.com/pricing) Customers buy a short-time pass which allows 30-minute rides. Subscribers buy yearly passes that allow 45-minute rides. What fraction of rides exceed their corresponding time limit?

0.123456789

Most of the time, a bike will begin a trip at the same station where its previous trip ended. Sometimes a bike will be moved by the program, either for maintenance or to rebalance the distribution of bikes. What is the average number of times a bike is moved during this period, as detected by seeing if it starts at a different station than where the previous ride ended?

98.76543210

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

Q3: This question 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. Here are some factors to consider:

1. While their potential is important, projects are assessed primarily based on the success of analysis performed. We are looking for data scientists who are able to deliver value, not just promise it.
2. We're looking for creative, original thinkers who can find novel questions to ask about different datasets. While your work does not have to be completely original, you should Google around to see if your dataset has already been studied extensively. Using other challenge question datasets demonstrates a lack of creativity.
3. High-impact problems of general interest are more interesting than academic research problems. If you solve the problem, will anyone care? Identifying interesting problems is half the challenge when leaving the academy.
4. Proposals that explain a non-obvious thesis supported by your plots are the most compelling. Generic exploratory plots of arbitrarily-chosen raw data columns are not as impressive as plots of processed data that convey some insight about your dataset.
5. Downloading a pre-formatted, pre-cleaned dataset intended for machine learning (e.g. UCI or Kaggle datasets) is less impressive than pulling data from an API or scraping a webpage. Most real-world data does not come neatly pre-packaged.
6. All things being equal, analysis of larger datasets is more impressive than analysis of smaller ones.
7. All things being equal, people who demonstrate the ability to use git(https://git-scm.com/) and Heroku(https://www.heroku.com/) will be viewed more favorably. To get started, try following this git tutorial(https://try.github.io/) or these Heroku tutorials(https://devcenter.heroku.com/start) in your favorite language.

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 a Heroku apps domain(https://www.heroku.com/) for your hosting.*

http://real-cheap-eats.herokuapp.com

Link to 2nd plot. You are highly encouraged to use a Heroku apps domain(https://www.heroku.com/) for your hosting.*

http://real-cheap-eats.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

Please provide the EMBED URL to your video*

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

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

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