

<iframe src="https://www.googletagmanager.com/ns.html?id=GTM-TDXFN2P" height="0" width="0" style="display:none;visibility:hidden"></iframe>

The Data Incubator

Data Scientist: The Sexiest Job
of the 21st Century

Harvard Business Review

Challenge

Warning: We suggest you use 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.

Save often. If you have filled out parts of the form but you are not ready to submit yet, we highly recommend that you save your solutions often by clicking the "Save" button below in order to avoid losing work due to any browser issues.

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

Resubmit 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 (click to expand):

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 blog posts 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, LinkedIn, and Facebook.

Having browser troubles? We recommend using Chrome (possibly using Incognito Mode).

Having trouble downloading any files? We suggest using command-line tools, rather than relying on a browser.

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.

Questions a little too difficult? You might want to consider signing up for our online Data Science Foundations class, which teaches the pre-requisite material needed for the fellowship.

Top of Form

Section 1: For this challenge, you will be asked to answer questions based on arrest incidents data of the city of Los Angeles. Information of the data set can be found [here](#) and the download link is [here](#). Each row in the data represents the booking of an arrestee. Only consider data prior to January 1, 2019. For some questions, we specify a given date range to consider.

How many bookings of arrestees were made in 2018?

How many bookings of arrestees were made in the area with the most arrests in 2018?

What is the 95% quantile of the age of the arrestee in 2018? Only consider the following charge groups for your analysis:

Vehicle Theft

Robbery

Burglary

Receive Stolen Property

There are differences between the average age of an arrestee for the various charge groups. Are these differences statistically significant? For this question, calculate the Z-score of the average age for each charge group. Report the largest absolute value among the calculated Z-scores.

Only consider data for 2018

Do not consider "Pre-Delinquency" and "Non-Criminal Detention" as these charge groups are reserved for minors

Exclude any arrests where the charge group description is not known

Felony arrest incidents have been dropping over the years. Using a trend line (linear estimation) for the data from 2010 and 2018 (inclusive), what is the projected number of felony arrests in 2019? Round to the nearest integer. Note, the data set includes arrests for misdemeanor, felonies, etc.

How many arrest incidents occurred within 2 km from the Bradbury Building in 2018? Use (34.050536, -118.247861) for the coordinates of the Bradbury Building . For simplicity, please use the spherical Earth projected to a plane equation for calculating distances. Use the radius of the Earth as 6371 km. Note, some arrest records are missing location data and the location is listed as (0, 0). These records should not factor in your calculation.

How many arrest incidents were made per kilometer on Pico Boulevard during 2018? For this question, we will need to estimate the length of Pico Boulevard, which mostly stretches from east to west. To estimate the length of Pico Boulevard:

Consider all location data which the listed address mentions "Pico".

Remove outliers by filtering out locations where either the latitude or longitude is 2 standard deviations beyond the mean of the subset of identified points.

To estimate the length, calculate the distance from the most western and eastern coordinate points. As before, use the simplified flat surface equation.

Once you have estimated the length of Pico Boulevard, you can proceed to report the number of arrest incidents per kilometer on Pico Boulevard in 2018.

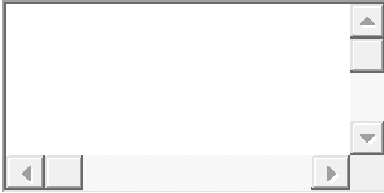
Some types of arrest incidents in certain areas occur at a highly disproportionate rate compared to their frequency city-wide. For example, let's say that the rate of larceny arrests (charge group code 6) is 1% in Devonshire (area ID 17). This rate may appear low but what if larceny arrests constitute 0.1 % city-wide? The ratio between these two probabilities is 10 and we can say that larceny occurs unusually often in Devonshire (Note, these numbers were made up for illustration purposes). Calculate this ratio for all charge group code and area ID pairs. You can view this ratio as the ratio of the conditional probability of an arrest incident of a charge group code given that it occurred in an area ID to the unconditional probability of the arrest incident of a charge group. Report the average of the top 5 of the calculated ratio.

Consider all records prior to January 1, 2019.

Some arrest incidents don't have a charge group code. These records should not be considered in your analysis.

Arrest incidents for charge group code 99 should not be considered in your analysis.

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:

Consider a grid in d -dimensional space. There are n grid lines in each dimension, spaced one unit apart. We will consider a walk of m steps from grid intersection to grid intersection. Each step will be a single unit movement in any one of the dimensions, such that it stays on the grid. We wish to look at the number of possible paths from a particular starting location on this grid.

For example, consider the case where $d=2$ and $n=3$. We will label the grid intersections (x,y) , where $x,y \in \{0,1,2\}$. There will be six valid walks starting at $(0,0)$ of length $m=2$:

$(0,0) \rightarrow (0,1) \rightarrow (0,0)$

$(0,0) \rightarrow (0,1) \rightarrow (0,2)$

$(0,0) \rightarrow (0,1) \rightarrow (1,1)$

$(0,0) \rightarrow (1,0) \rightarrow (0,0)$ $(0,0) \rightarrow (1,0) \rightarrow (0,0)$

$(0,0) \rightarrow (1,0) \rightarrow (2,0)$ $(0,0) \rightarrow (1,0) \rightarrow (2,0)$

$(0,0) \rightarrow (1,0) \rightarrow (1,1)$ $(0,0) \rightarrow (1,0) \rightarrow (1,1)$

Note that the walks may double back upon themselves, and multiple walks may end at the same grid intersection. All of these walks are counted.

Consider the case where $d=4$, $n=10$, and $m=10$.

How many valid walks start from the corner $(0, 0, 0)$?

Consider the case where $d=4$, $n=10$, and $m=10$.

Consider the count of valid walks associated with each possible starting position. What is the ratio of the highest count of valid walks to smallest count of valid walks?

Consider the case where $d=4$, $n=10$, and $m=10$.

Consider the count of valid walks associated with each possible starting position. What is the ratio of the standard deviation of the number of valid walks to the mean of the number of valid walks?

Now, let's consider the case where $d=8$, $n=12$, and $m=12$.

How many valid walks start from one of the corners?

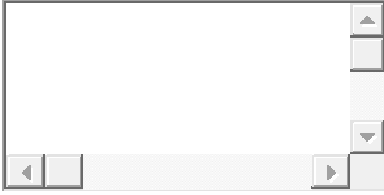
Consider the case where $d=8$, $n=12$, and $m=12$.

Consider the count of valid walks associated with each possible starting position. What is the ratio of the highest count of valid walks to smallest count of valid walks?

Consider the case where $d=8$, $n=12$, and $m=12$.

Consider the count of valid walks associated with each possible starting position. What is the ratio of the standard deviation of the number of valid walks to the mean of the number of valid walks?

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



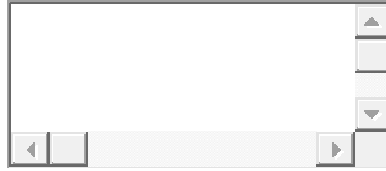
In what language is the script written?

- ☐ 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 as well as the archive of data sources on Data is Plural. You can see some final projects of previous Fellows on our YouTube Page.

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.



Propose a project.*

Link to public description of data source.*



Link to 1st plot. You are highly encouraged to use Heroku apps domain for an app or Github to display a notebook.*



Link to 2nd plot. You are highly encouraged to use Heroku apps domain for an app or Github to display a notebook.*



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

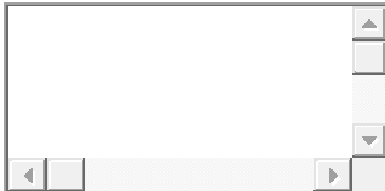


How did you obtain your dataset? (Please check all that apply.)

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



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).*

☐ 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'.

×Success! Thank you for submitting your challenge questions. A confirmation email has been sent to your email with instructions on how to make sure you receive further communications from us.

×Saved! We have saved a copy of your submission. You can come back before the challenge is due to modify answers. If you have submitted a fully valid challenge at this point, your status has been updated to reflect this.

×Something went wrong! See above for specific validation errors.

Bottom of Form

With loads of data you will find relationships that aren't real.
Big data isn't about bits,
it's about talent.

Forbes Magazine

