“ 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.
- **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 loosing 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):

Section 1: The city of New York has collected data on every automobile collision in city limits since mid-2012. Collisions are broken down by borough, zip code, latitude/longitude, and street name. Each entry describes injuries/deaths, collision causes, and vehicle types involved. The data can be downloaded from: https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95 Download the "NYPD Motor Vehicle Collisions" dataset in .csv format. The download link can be found under the "Export" tab. Information on the variables can be found on this page, as well, along with a preview of the rows of the dataset. For all questions, do not use data occurring after December 31, 2018.

What is the total number of persons injured in the dataset (up to December 31, 2018?)

368034

What proportion of all collisions in 2016 occured in Brooklyn? Only consider entries with a non-null value for BOROUGH.

0.3096177807

What proportion of collisions in 2016 resulted in injury or death of a cyclist?

0.02165474263

For each borough, compute the number of accidents per capita involving alcohol in 2017. Report the highest rate among the 5 boroughs. Use populations as given by https://en.wikipedia.org/wiki/Demographics_of_New_York_City.

0.0002299179506

Obtain the number of vehicles involved in each collision in 2016. Group the collisions by zip code and compute the sum of all vehicles involved in collisions in each zip code, then report the maximum of these values.

5703

Consider the total number of collisions each year from 2013-2018. Is there an apparent trend? Fit a linear regression for the number of collisions per year and report its slope.

6447.914285

Do winter driving conditions lead to more multi-car collisions? Compute the rate of multi car collisions as the proportion of the number of collisions involving 3 or more cars to the total number of collisions for each month of 2017. Calculate the chi-square test statistic for testing whether a collision is more likely to involve 3 or more cars in January than in May.

5573.633343

We can use collision locations to estimate the areas of the zip code regions. Represent each as an ellipse with semi-axes given by a single standard deviation of the longitude and latitude. For collisions in 2017, estimate the number of collisions per square kilometer of each zip code region. Considering zipcodes with at least 1000 collisions, report the greatest value for collisions per square kilometer. Note: Some entries may have invalid or incorrect (latitude, longitude) coordinates. Drop any values that are invalid or seem unreasonable for New York City.

3885.339454

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

```
 #!/usr/bin/env python
# coding: utf-8

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

In what language is the script written?

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

Section 2: A sequence of n numbers is considered valid if the sequence begins with 1, ends with a given number j, and no two adjacent numbers are the same. Sequences may use any integers between 1 and a given number k, inclusive (also $1 <= j <= k$). Given parameters n, j, and k, count the number of valid sequences. The number of valid sequences may be very large, so express your answer modulo $10^{10} + 7$.

$n = 4$, $k = 4$, **and** $j = 2$**?**

7

$n = 4$, $k = 100$, **and** $j = 1$**?**

9702

$n = 100$, $k = 100$, **and** $j = 1$**?**

7934293301

$n = 347$, $k = 2281$, **and** $j = 829$**?**

4403056638

$n = 1.26 \cdot 10^6$, $k = 4.17 \cdot 10^6$, **and** $j = 1$**?**

1926411550

$n = 10^7$, $k = 10^{12}$, **and** $j = 829$**?**

8051788353

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

```
def count_seq(n, j, k):

    # Dynamic programming

    dp1 = list(range(n+1)) # count of sequences ending with j, whose length equals its index.
    dp2 = list(range(n+1)) # count of sequences not ending with j, whose length equals its index.
```

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).

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'.

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. ×

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