CHALLENGE

Your challenge will close on April 19 at 11:28 AM (your local time). You have 21:33:26 hours remaining.

Challenge saved!

Note: To be considered a finalist for the Fellowship, we must receive your completed challenge no later than **2 weeks after your application was submitted**.

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. (*) 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 when finished.** Be sure to press submit when you are *completely finished* with the challenge. This lets us know that you are done with your solutions so we can begin to review them. You will not be able to work further on the challenge after submitting your work.

A few helpful hints (click to expand):

Section 1:

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 or other assets supporting this. 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/).

Please provide a general description and justification for your project. *

be obtained from Kaggle.
Analysis: In the notebook, I firstly explored the dataset that records the user-article interaction. Then I built a simple ranked-based recommend system, which could be used to solve the cold start problem. Next a user-

Link to 1st asset. 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://github.com/BokaiXu/IBM_Recommand_Syst

Link to 2nd asset. 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://github.com/BokaiXu/IBM_Recommand_Syst

Link to public description of data source: *

https://www.ibm.com/cloud/blog

How much data did you analyze (rounded to nearest MB)? *

13

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

- [x] I downloaded a dataset available online
- [ ] I used a provided API
- [ ] I scraped data from a webpage
- [ ] Other (please explain)

If you obtained your data through some other means, please explain below:

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

```
def get_top_articles(n, df=df):
    '''
    INPUT
```

In what language is the script written? *

- ○ C/C++
- ○ Java
- ○ MATLAB
- ◉ Python
- ○ R
- ○ SQL
- ○ Other

Section 2:

With the rise of computer-aided police dispatch systems, many municipalities have large data sets on calls for service. These may include both calls from the public and from officers on patrol. Some cities provide this data to the public. New Orleans is one of these, with all of their Calls for Service data since 2011 available on their Open Data(https://data.nola.gov/) website.

For each of the questions below, use the New Orleans Calls for Service data sets from 2016 through 2020. The data can be found in New Orleans' Open Data portal. Each year comes as its own CSV file: 2016(https://data.nola.gov/api/views/wgrp-d3ma/rows.csv?accessType=DOWNLOAD), 2017(https://data.nola.gov/api/views/bqmt-f3jk/rows.csv?accessType=DOWNLOAD), 2018(https://data.nola.gov/api/views/9san-ivhk/rows.csv?accessType=DOWNLOAD), 2019(https://data.nola.gov/api/views/qf6q-pp4b/rows.csv?accessType=DOWNLOAD), 2020(https://data.nola.gov/api/views/hp7u-i9hf/rows.csv?accessType=DOWNLOAD). A brief description of the data can be found here(https://data.nola.gov/Public-Safety-and-Preparedness/Call-for-Service-2020/hp7u-i9hf), for example.

Start by considering just the data from 2020. Calls are classfied into several types. What fraction of calls are of the most common type?

0.24

Now compare to the data from 2016. Find the call type that displayed the largest percentage decrease in volume between 2016 and 2020. What is the fraction of the 2016 volume that this decrease represents? The answer should be between 0 and 1. (Note that the name of the type column changes over time. You can use the *TypeText* column, which remains constant, to determine the call type.)

24.5

For this and the remaining questions, consider data from all five years. As you combine the data, you will notice that duplicate item numbers appear across years, for calls whose resolution spans the new year. Remove these duplicate rows. How many duplicate rows were removed?

5724

Some calls result in an officer being dispatched to the scene, and some log an arrival time. What is the median response time (dispatch to arrival), in seconds, considering only valid (i.e. non-negative) times?

145.0

Work out the average (mean) response time in each district. What is the difference between the average response times of the districts with the longest and shortest times?

368.0

Work out the average response time for each month. Make an ordinary least-squares fit to the response time against month. What is the slope of this line, in seconds per year?

-0.097

We can define surprising event types as those that occur more often in a district than they do over the whole city. What is the largest ratio of the conditional probability of an event type given a district to the unconditional probability of that event type? Consider only events types which have more than 100 events. Note that some events have their locations anonymized and are reported as being in district "0". These should be ignored.

7.04

We can use the call locations to estimate the areas of the police districts. Represent each as an ellipse with semi-axes given by a single standard deviation of the longitude and latitude. What is the area, in square kilometers, of the largest district measured in this manner?

28.0

In what language is the script written?

Python

Please provide the script used to compute your response (max 10000 characters).

```
import pandas as pd
import numpy as np
```

Section 3:

You roll a fair 6-sided dice repeatedly until the sum of the dice rolls is greater than or equal to M.
For all questions, give your answer to 5 decimal places.

What is the mean of the sum minus M when $M = 20$?

1.69448

What is the mean of the number of rolls when $M = 20$?

6.20402

What is the standard deviation of the sum minus M when $M = 20$?

1.49095

What is the mean of the sum minus M when $M = 5000$?

1.66667

What is the mean of the number of rolls when $M = 5000$?

1429.04762

What is the standard deviation of the sum minus M when $M = 5000$?

1.49071

In what language is the script written?

Python

Please provide the script used to compute your response (max 10000 characters).

```
M=20
import numpy as np
```

How many hours did it take you to complete this challenge? This will not be considered in your application, and is only used for future challenge design.

50

SAVE SUBMIT