CHALLENGE

Your challenge will close on June 29 at 1:10 PM (your local time). You have 42:24:06 hours remaining. ×

Challenge saved! ×

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. *

Two week COVID-19 case increase as a function of growth rate at the time of reopening
Today one of the most important data analysis topic is studying the pandemic and how it evolves. This

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/stalei/DIProject/blob/master/Dat

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/stalei/DIProject/blob/master/Fit.

Link to public description of data source: *

https://github.com/CSSEGISandData/COVID-19/tre

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

1

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)

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

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

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

In what language is the script written? *

○ C/C++

○ Java

○ MATLAB

● Python

○ R

○ SQL

○ Other

Section 2:

The City Record is the official journal of New York City, and provides information provided by city agencies. This data is available in searchable form online at the City Record Online (CROL). For this challenge, we will use a subset of the CROL data consisting only of procurement notices for goods and services. It can be found at this link(https://data.cityofnewyork.us/api/views/qyyg-4tf5/rows.csv).

For more information on CROL and the publishing of procurement notices, see this link. (https://a856-cityrecord.nyc.gov/Home/AboutUs).

Keep only rows with a StartDate occurring from 2010 to 2019, inclusive. Next, remove all rows for which the ContractAmount field is less than or equal to zero, or is missing entirely. Use this filtered data for the rest of the challenge, as well. For the remaining data, what is the total sum of contract amounts?

2.07783994328e+11

Determine the number of contracts awarded by each agency. For the top 5 agencies in terms the number of contracts, compute the mean ContractAmount per contract. Among these values, what is the ratio of the highest mean contract amount to the second highest?

4.7924

Consider only procurements made by the Citywide Administrative Services agency and compute the sum contract amount awarded to each unique vendor. What proportion of the total number of contracts in the data set were awarded to the top 50 vendors?

0.636132

Do agencies publish procurement notices uniformly throughout the week? As an example, consider the agency of Parks and Recreation. For this agency, compute the weekday for which each notice was published, and perform a Chi-squared test on the null hypothesis that each weekday occurs equally often. Report the value of the test statistic.

25.8181818182

For this question, consider only contracts with in the categories of Construction Related Services and Construction/Construction Services. The ShortTitle field contains a description of the procured goods/services for each contract. Compute the sum contract amount for contracts whose ShortTitle refer to 'CENTRAL PARK' and for those which refer to 'WASHINGTON SQUARE PARK'. What is the ratio of total construction and contruction-related expenditure for the Central Park contracts compared to the Washington Square Park contracts? Note: you should ensure that 'PARK' appears on its own and not as the beginning of another word.

1.47638

Is there a predictable, yearly pattern of spending for certain agencies? As an example, consider the Environmental Protection agency. For each month from 2010 through the end of 2019, compute the monthly expenditure for each agency. Once again, use StartDate for the contract date. Then, with a lag of 12 months, report the autocorrelation for total monthly expenditure.

0.159782

Consider only contracts awarded by the Citywide Administrative Services agency in the category Goods. Compute the total yearly expenditure (using StartDate) for these contracts and fit a linear regression model to these values. What is the R^2 value for this model?

0.743032358943

In this question, we will examine whether contract expenditure goes to companies located within or outside of New York City. To do so, we will extract the ZIP codes from the VendorAddress field. The ZIP codes pertaining to New York City can be found at the following URL: https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm. Looking only at contracts with a StartDate in 2018, compute the total expenditure for contracts awarded to vendors listing NYC addresses and those located elsewhere. Report the proportion of the total expenditures awarded to the NYC vendors.

1.0

In what language is the script written?

Python ▼

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

Section 3:

Consider a chess knight moving on the first quadrant of the plane. It starts at $(0,0)$, and at each step will move two units in one direction and one unit in the other, such that $x \geq 0$ and $y \geq 0$. At each step the knight randomly selects a valid move, with uniform probability. For example, from $(0,1)$, the knight will move to $(1,3)$, $(2,2)$, or $(2,0)$, each with probability one-third.

After 10 moves, what is the expected Euclidean distance of the knight from the origin? (If the knight is at $(2,1)$, its distance is $\sqrt{2^2+1^2} \approx 2.24$.)

1.23456789

What is the expected standard deviation in this distance?

1.23456789

If the knight made it a distance of at least 10 from the origin some time during those 10 moves, what is its expected Euclidean distance at the end of the 10 moves?

1.23456789

What is the expected standard deviation in this distance?

1.23456789

After 100 moves, what is the expected Euclidean distance of the knight from the origin?

1.23456789

What is the expected standard deviation in this distance?

1.23456789

In what language is the script written?

C/C++ ▼

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

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.

99

SAVE SUBMIT