Challenge

Your challenge will close on October 4 at 9:25 AM (your local time). You have 40:25:50 hours remaining.

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(https://www.thedataincubator.com/blog/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(https://www.thedataincubator.com/blog/2017/01/20/how-employers-judge-data-science-projects/).

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

I would like to write a game recommender system for Steam users.
The recommender system has been used in different domains including Youtube and Yelp.

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://game-rec.herokuapp.com/

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/GameRec

Link to public description of data source: *

https://steamcommunity.com/dev

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

1005

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

- [ ] I downloaded a dataset available online
- [x] 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:

N/A

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

```
# Obtained data:
import requests
```

In what language is the script written? *

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

Section 2:

For this challenge, you will be asked to answer questions regarding WiFi hotspot locations in NYC. Information of the dataset can be found here and the download link is here(https://data.cityofnewyork.us/api/views/yjub-udmw/rows.csv?accessType=DOWNLOAD). A data dictionary(https://data.cityofnewyork.us/api/views/yjub-udmw/rows.csv?accessType=DOWNLOAD) is also provided. Each row in the data represents one reported WiFi hotspot.

How many unique providers are there?

17

How many WiFi hotspots are there by the second most common provider in the Bronx?

120

What fraction of WiFi hotspots are in parks? For simplicity, you can consider a park a place where the name of the location where the WiFi is located contains the word "park".

0.053

What is the probability that a WiFi hotspot is free (without any limitations) given that it's not in a library? For this question, pull the location data based on the "Location_T" field.

0.81

How are WiFi hotspots distributed across neighborhoods? For this question, calculate the number of WiFi hotspots per capita for each Neighborhood Tabulation Area (NTA). Exclude NTAs with less than 30 reported WiFi hotspots. Report the interquartile range(https://en.wikipedia.org/wiki/Interquartile_range) of the averages.
For population data for each NTA, use this dataset; information on the dataset(https://data.cityofnewyork.us/api/views/rnsn-acs2/rows.csv) is found here(https://data.cityofnewyork.us/City-Government/Census-Demographics-at-the-Neighborhood-Tabulation/rnsn-acs2). Use the population data for the column corresponding to 2010.

0.000851

How far must one travel from one hotspot to another? For this question, report the median distance, in feet, of the average distance between each hotspot to the nearest 3 hotspots. For your distance calculation, calculate the distance "as the crow flies" (https://en.wikipedia.org/wiki/As_the_crow_flies). For simplicity, please use the spherical Earth projected to a plane equation(https://en.wikipedia.org/wiki/Geographical_distance#Spherical_Earth_projected_to for calculating distances. Use the radius of the Earth as 6371 km. Remember, report your answer in feet.

276.3

The dataset contains information on the date the hotspot was activated. What fraction of all activations occurred on the day of week that had the most activations? In other words, if Monday had the most activations, what fraction of activations occurred on Monday? Note: there are some dates that don't make sense. Ignore them for the analysis.

0.244

If you plot the number of hotspot activations for each month, you'll notice a general increase but then a precipitous drop after June 2018. Using a linear estimate for the number of monthly activations, what is rate of increase in monthly activations? Only consider data before July 1, 2018 and set the start date as the earliest date of the data. If you need to, use 30.5 days in a month.

1.5279

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:

Consider n buildings on a grid at positions 0 to $n+1$. Each building has a height h_i. You can arrange the buildings in any order but you must leave the slot at position 0 open. Imagine a laser is shot just below the roof and to left of a building. The laser travels any number of grid points until it encounters a building of the same height or taller or reaches the end of the grid, position 0. For example, consider 4 buildings of height 3, 3, 4, and 1 arranged at grid points 1, 2, 3, and 4, respectively. The laser travels 1, 1, 3, and 1 grid points for each of the buildings, respectively. Let's call the sum of the lasers' distances V. For this example, $V = 6$. For all questions, give your answer to 10 places after the decimal point.
Consider 4 buildings of heights 1, 1, 3, and 4. For all possible configurations, what is the mean of sum of distances V?

7.4

Consider 4 buildings of heights 1, 1, 3, and 4. For all possible configurations, what is the standard deviation of the sum of distances V?

1.8075472029

Consider 10 buildings of heights 1, 2, 3, ..., 10. For all possible configurations, what is the mean of sum of distances V?

1.23456789

Consider 10 buildings of heights 1, 2, 3, ..., 10. For all possible configurations, what is the standard deviation of the sum of distances V?

1.23456789

Consider 20 buildings of heights 1, 2, 3, ..., 20. For all possible configurations, what is the mean of sum of distances V? The calculation of the standard deviation is not needed.

1.23456789

In what language is the script written?

Python

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

```
# Q1
from itertools import permutations
```

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