Fall 2018 DATA622.002 Homework #1

1. Assigned on August 28, 2018
2. Due on September 9, 2018 11:59 PM EST
3. 15 points possible, worth 15% of your final grade

Python Courses at DataCamp (10 points) The link to setting up a free Datacamp with full and free access to the following courses have already been sent to your CUNY inbox. If you have trouble getting set up with Datacamp, please ping the #hw1 Slack channel.

Complete the following 3 courses:

1. Python Data Science Toolbox Part 1
2. Python Data Science Toolbox Part 2
3. Supervised Learning with scikit-learn

These courses are assigned to help students who are already familiar with these skillsets to review them before diving into our class.
Likewise, they are also selected to give students who are unsure if they meet the prerequisites a chance to familiarize themselves with the basic skills. If you find yourself having a hard time completing these courses, please DM me on Slack or email me, and we can re-assess if this is the right course for you.

Submission Instructions:

I can see your completion status via Datacamp, so there's no need to submit anything besides completing the courses above in Datacamp. Please note that late submission can be tracked via Datacamp.

Data Science Working Environment: Docker & Anaconda (5 points)

1. Read this article on Python environment and Anaconda. https://medium.freecodecamp.org/why-you-need-python-environments-and-how-to-manage-them-with-conda-85f155f4353c
2. Read these articles on Docker. https://www.dataquest.io/blog/docker-data-science/; https://becominghuman.ai/docker-for-data-science-part-1-dd41e5ef1d80
3. Write a short paragraph (around ~200 words total) on the following topics:

the role "environment" plays in a data scientist's workflow (the readings from #1 & #2 will cover this)

the difference between Docker & Anaconda, and why some data science teams prefer to use the two together (you will need to do more research on this outside of the articles listed above)

Submission Instructions:

We will be using Github to submit this section (as well as the rest of the homework assignments in this course). It's a good idea to get familiar with this process now, since Github is a major part of a data

scientist's toolbox. By following the homework link, Github Classroom has cloned this repo for you (your url for this page should read something like github.com/cuny-spsmsda-data622-2018fall/fall2018-data622-002hw1-[your github handle]). This is your own cloned private repo for this assignment and I will be able to track and grade whatever changes you make to this repo. Please note that late submission can be tracked via Github Classroom. For the 200 word paragraph, feel free to directly modify this ReadMe file to add your answer. Or, you can upload a PDF, word document, or whatever format you prefer to this repo. As usual, any questions? Ping #hw1 on Slack.

Questions? Start with Slack!