Final Project assignment, MA346, Fall 2020

# Summary

In this project, you will do a data exercise of your own choosing. I will give you requirements that you must fulfill, to be sure that your chosen project includes enough of our course content to show off what you’ve learned. Other than that, you are free to choose the topic, the dataset, and the methods. Due date: 12/3/2020, 11:59pm.

# Specific Requirements

Your project must include each of the following accomplishments.

* **Find and prepare your own datasets.**
  + I will not be assigning a topic or any data for this project. You are in charge of finding it. Advice appears later in this assignment about where you might look for data sources.
  + Data cleaning is required only if the datasets you find happen to need it. If you’re lucky enough to find data in the perfect format (which is quite rare) that’s okay. Munging is not a requirement of the project, but most projects end up including it anyway.
* **Include at least one challenging step** in your work. Although you might end up with a project that has easy data cleaning, or an easy analysis, or simple visualizations, at least one part of your project must be challenging. Choose a challenging element from the following list.
  + **Merging:** You needed to merge two datasets and they did not have a unique ID, so you had to do some nontrivial work to create one (as in Chapter 12 of the course notes).
  + **Stats/modeling:** Your analysis used Python to do some statistics or modeling that was not easy (such as curve\_fit from Chapter 9 of the course notes, or a model/algorithm/test you learned in a Bentley course other than GB213 and MA346).
  + **Final weeks:** You used one of the advanced techniques we learned at the end of our course (chapters 15-17 of the course notes, on networks, matrices, and machine learning).
  + **Importing:** Your data was fetched from a complex data source, such as a web API that gave you JSON data, or a database that you could query only with SQL (both of which were covered in DataCamp lessons in Weeks 7 and 8).
  + **Coding:** You had to write some complex Python code to complete your project, such as a random simulation, or a new Python package you learned specifically for this project.
  + **Other:** If you have another suggestion for what you’d like to add to this list, feel free to suggest it to your instructor. Finding your own challenges is more than welcome!
* **Include a report** suitable for reading by a potential employer.
  + That is, aim to make your project something you’re sufficiently proud of that you could put a link to it on your resume or LinkedIn profile, and you’d expect that a potential employer could visit it and understand what you’ve learned to do.
  + Keep this target audience in mind! Don’t write jargon only understandable by a classmate who’s studied what you have. Speak to an audience who is intelligent and interested in data projects, but probably doesn’t know Python.
  + This report can be in any suitable format (Word, PDF, Jupyter, etc.).
* **Create and publish a dashboard** that showcases some aspect of your work.
  + This need not be the final product of your analysis. For instance, you can publish a data exploration dashboard, then do a report that does a deep dive into a specific analysis not included in the dashboard.
* **Publish in a GitHub repo, and on a cloud computing site such as Deepnote,** your entire folder of work (code, data, report).
  + In GitHub, add a README.md that includes a quick summary of your project.
  + ***To submit your project, just email your instructor a link to this repo!***
* **Add links.** Each piece of your work should hyperlink to all the others.
  + Your report should mention where the dashboard and repo are located.
  + The dashboard should link back to the repo (which contains the report).
  + The repo’s README.md should link to all of these things.

# Where to find data

**If you have a topic you’re considering:** Start with a web search like “free golf datasets” or “free nutrition datasets” or whatever your topic is. Don’t take just the first result; persist until you find data that is comprehensive enough, clean enough, and reliable enough for your needs.

**If you don’t have a topic you’re considering:** Try browsing some of the resources listed below. Each list is not necessarily a list of individual datasets; some are lists of *collections* of datasets.

* KDNuggets is a data science blog that maintains [this list of 70 free data sources they recommend](https://www.kdnuggets.com/2017/12/big-data-free-sources.html).
* Amazon Web Services has a [marketplace of datasets](https://aws.amazon.com/marketplace/search/results?filters=FulfillmentOptionType&FulfillmentOptionType=AWSDataExchange), some of which are free. Check the appropriate box to search only for free datasets.
* One GitHub repo with many contributors curates [a list of a few hundred data sources](https://github.com/awesomedata/awesome-public-datasets), categorized by subject area.
* Teradata is an analytics consulting company that has free resources for academic use. See their [Teradata University Network here](https://academics.teradata.com/). (Requires free registration.)
* Towards Data Science is a blog in which one post recommends [this list of 10 free data sources](https://towardsdatascience.com/top-10-great-sites-with-free-data-sets-581ac8f6334).
* freeCodeCamp is a DataCamp competitor and recommends [this list of 15 free open data sources](https://www.freecodecamp.org/news/https-medium-freecodecamp-org-best-free-open-data-sources-anyone-can-use-a65b514b0f2d/).
* If you’re interested in financial/market data, consider stopping by the Bentley Trading Room and discussing your ideas with the staff there.

# How you will be graded

You will be graded on the following criteria.

**Basics:**

Submitted a link to your GitHub repository by the deadline (12/3/2020) 15%

That repository’s README links to a report in any of the acceptable formats 5%

The README also links to a dashboard you created for this project 5%

The README also links to where I can run your code in the cloud (Colab/etc.) 5%

**Writing:**

Any notebook file you submitted follows the before-and-after principle for

explaining itself; any Python script was liberally commented 5%

Your report begins by stating your goals and summarizing your results,

then ends with your results in greater detail 5%

Your report explains the origins of your data, why it fits your goals, and what

you had to do to get it into a suitable state 10%

Your report was suitable for the target audience you’ve been assigned 10%

**Math and Python:**

You undertook one of the challenging steps listed on page 1 of this assignment 15%

Your chosen dataset(s) and computations logically suit your stated goals 15%

My review of your code did not reveal any incorrect steps 10%

**Total 100%**