Capstone 1: Analytic Report and Research Proposal

To consolidate all the skills you've learned so far, you'll now embark on completing your first capstone project: an Analytic Report and Research Proposal on a dataset of your choosing.
In this checkpoint, we'll go over the requirements for the capstone deliverable as well as prepare you for the presentation you'll do after you submit your capstone materials.
At the end of this checkpoint, we'll ask you to submit your final materials for your capstone project so you and your mentor can discuss in the leadup to your presentation.

Requirements

Your Report should accomplish these three goals:

1. Describe your dataset. Describe and explore your dataset in the initial section of your Report. What does your data contain and what is its background? Where does it come from? Why is it interesting or significant? Conduct summary statistics and produce visualizations for the particular variables from the dataset that you will use.
2. Ask and answer analytic questions. Ask three analytic questions and answer each one with a combination of statistics and visualizations. These analytic questions can focus on individuals behaviors or comparisons of the population.
3. Propose further research. Lastly, make a proposal for a realistic future research project on this dataset that would use some data science techniques you'd like to learn in the bootcamp. Just like your earlier questions, your research proposal should present one or more clear questions. Then you should describe the techniques you would apply in order to arrive at an answer.

See this recent analysis on 2016 celebrity deaths for an excellent example of data-driven story telling that presents a problem, explores data, and produces an answer. The analytics are more robust techniques than we've covered so far, but the general idea and tone are spot on.

Report guidelines

Keep these guidelines in mind as you draft your _Report_:

- ○ Length. Your Report should be in a notebook with visualizations. Short and clear is better than long and opaque.
- ○ Structure. Pay attention to the narrative structure of your Report. Each section should flow into the next and make a logical, readable text. Don't simply create a list of bullet points or present code without explanation.
- ○ Format. The best format for your Report is an interactive Jupyter notebook ipynb file. However, you are welcome to use any format you like, so long as you're able to include visualizations and include (or link to) the code you use to generate your visualizations and summary statistics. If a Jupyter notebook would be too much overhead or unduly

distract you from creating good content, markdown files (hosted perhaps on GitHub or as a gist), blog posts, or even Word or Google documents are acceptable.

Getting started

Your first step is choosing an interesting dataset to work with. You're welcome to use any dataset you like. If you aren't sure which one to use or are looking for inspiration, check out this collection of open data sources. Before deciding on a particular dataset think about the kinds of questions you might be able to answer. Consider the format of the data. Do you know how to (or will you quickly be able to learn to) access and load it? Once you've chosen a dataset, write out some of those preliminary questions. Having them early will help guide your initial data exploration.

In order to conduct summary statistics and prepare visualizations you'll need to collect the data and load it into Python / pandas. Some of the data in the sources above will be in a format we didn't teach you to load in this fundamentals course. If necessary, refer to the pandas I/O documentation.

Once you've loaded your data, dig around with pandas and matplotlib to explore it. What variables does your data contain and what distributions do you think they have? Does the data bear on the preliminary questions you wrote down? What new questions might you answer? How does the data look when you plot it out?

At this point you should be ready to start writing your Report. Decide what format to use, which three analytic questions you'll ask and answer, which research questions you'd like to ask and which data science techniques might be appropriate to answering them. If necessary, do independent research now about the field of data science, or discuss the topic with your mentor, to decide which potential techniques you could use.

You are encouraged to make use of every resource at your disposal in putting together your Report. This extends to getting preliminary feedback on your work from your mentor or from other friends and family. However, you should be ready to explain every decision, conclusion, and visualization you make and all of the code you write.

Presenting your first capstone and evaluation

Once you've completed your capstone materials, you'll need to schedule a time to present it to a Thinkful educator. This will be a one-on-one presentation with an educator who is not your mentor, who will listen to your presentation and then ask follow up questions. Ultimately, they'll provide you with written feedback on your capstone.

Once you've completed this module, the following goal in your dashboard will prompt you to schedule a time to meet with someone for this presentation.

To help you prepare for your capstone review, here are a few examples of the types of questions you'll be asked:

- Did you have any challenges with this data?
- Why did you choose this dataset?
- How did your dataset inform the questions you chose to explore?

- What issues did you run into while analyzing your data?
- Imagine someone sees this visualization out in the wild, separated from your report. What conclusions would you expect them to draw? Is that the conclusion that you want them to draw?
- How could you make your conclusions more rigorous?

You should also take some time to review the rubric that you'll be scored on.
Here's a few last pieces of advice:

- Grammar matters. A lot. This should be a professional and easy to read document.
- State the questions you aim to answer clearly and answer them specifically. Make sure to use markdown to properly format your questions.
- Including your code is required but we should also be able to read your report and understand your visualizations without having to look at that code. Whether you include your code in the report with an iPython notebook or in a separate file is up to you.
- Your goal should be to give us an understanding of your dataset and the behaviors present in it. As such use analytics and statistics to tell a story about the data, don't just give us statistics without context.
- Try to translate real questions into statistical questions rather than simply ask statistical questions.
- Use at least 2-3 different types of charts to display the data.
- Be clear about any assumptions you make about the data and validate those assumptions if possible.
- Ensure that your dataset actually has the information to answer the questions you're asking. Does the dataset have a bias? Is it incomplete? Problems with your dataset can easily lead to problems in your analysis if you don't address them.

Assignment

When you're ready, submit your final Report at the bottom of this page. Try to have a final discussion with your mentor about your work before you present your capstone.