Homework 4

1. Thinking about motivation for a data science project, write out a clear (and plausible) use case for the Technical Exercise below (following these Conceptual Questions).

The City of New York leadership has been criticized for allocating services disproportionately to wealthier areas of the city. As a result, they have created a new position for ensuring the representative delivery of services. One of this person's primary roles will be identifying the most disproportionately delivered services, flagging that to the appropriate managers, and tracking whether corrections have been made.

2. Refine the following question into a data science question. Note that for some steps, you'll need to decide on the refinement that creates operational clarity (i.e., you don't have a stakeholder to ask). Show the refinement steps as shown in Lesson 35, Slide 5. Draw on your experiences from class and past homeworks: "We'd like to understand how much characters talk across the My Little Pony series."

- We'd like to understand where services are delivered across the city.
- We'd like to understand the relative volume of services delivered across the city
- We'd like to understand the relative volume of different service types delivered across the city
- We'd like to understand the relative volume of different service types delivered to different parts of the city
- We'd like to understand the relative volume of different service types delivered to different parts of the city, broken out by median income levels.

3. Explain why state maintenance is hard to do in a Jupyter notebook.

Whenever a Jupyter notebook is closed, it loses its state (because the kernel is shutdown). Because Jupyter notebook cells can be executed in any arbitrary order, the state is a function of the order of execution and historical versions of cell code. As a result, the state is quite arbitrary and hard to know how to rebuild.

4. Make an argument for why a Jupyter notebook is better than a README for sharing a data science project with data scientists that might engage with it in the future.

A Jupyter notebook is both easy to read and executable. As a result, it's possible for the Jupyter notebook to both orient the user to the project as well as provide *runnable* example or actual code for some of the core activities the user will need to undertake in

the project. For example, the user can provide a set of cells that actually run the code for the entire data science project, in the right order, on an example data file.

Homework 5

1. A hospital wants to know how its medication stocks line up with patient demand in the hospital. Develop a use case that motivates this question by identifying a persona (presumably someone who works at the hospital) and a concrete activity that will make use of this information. Of course, we don't have access to the hospital in question – so you'll be developing your own use case. So there are many credible answers.

Persona: the hospital clerk responsible for ordering medications. They do this each month.

Activity: at the end of each month, the clerk looks at the past few months of medication usage as well as the current stocks and decides how much of each medication to order. This is naturally a forecast. Currently that forecast is based solely on this information.

Use case: the clerk's estimates currently don't take into account the patient load and how current trends in the kinds of patients at the hospital might impact demand in the coming month. The clerk will use the data analysis to prepare a forecast of the projected medication demand based on current patients in the hospital.

2. Why does the real world impact the complexity of code we write in a data science project?

Our data science code models the real world – it has to handle data from the real world, and it has to build models of behavior in the real world. As a result, the complexity of our code will mirror some of the complexities of the part of the real world we're modeling.

Homework 6

1. What is refactoring? Give three examples of refactoring techniques.

Refactoring is modifying code that changes its structure without modifying its functionality. Three examples:

- Modularizing code into functions
- Eliminating redundant code using for loops, dictionaries, and lists
- Renaming variables to increase readability

2. In a data science project, why does code naturally go through "phases" of messiness?

As we move into another aspect of a data science project, it's natural to first have to explore solutions and approaches to solving the problem. This experimentation process yields a lot of code that is either ultimately not needed or written in a way that doesn't align with the final approach. This lack of alignment is messy code.

3. What are three techniques for creating more modular code?

- Organize code into functions
- Organize functions into classes
- Organize functions and classes into separate files/namespaces (with a functional designation)

Homework 7

1. Why is the difference between found and designed data?

When we conduct a data science project, we require data. Found data is data that exists for a purpose OTHER than the project we are working on. Designed data is data that has been created specifically for our project.

2. What are the two primary challenges that necessitate sampling when collecting data?

The two primary challenges are:

- All the data we're actually interested in is too large to collect. "Large" is judged with respect to the methods we have for collecting it. So even if, technically, we could store all the data, but our collection method is incredibly slow – in this case the data is too "large".
- We don't know for sure what the data we want looks like – and, therefore, we don't have the ability to identify it at large scale. In this situation, we typically have to produce a sample by either using a heuristic (we'll collect some things that we want, but also some things we don't) or manual collection (we'll go out and find a bunch of things that are exactly what we want, but we don't have time to do that at scale).

3. Why are website owners more likely to be upset about collecting data using scrapers than using APIs?

APIs are explicitly designed for programmers to use to query for and gather data. So when a web platform provides an API, they are explicitly acknowledging that they're okay with the data available through it being collected and used (though, of course, they may only approve of certain uses of the data).

If the website owner has NOT provided an API, and the data is only available by visiting the website through a browser – it's fair to assume that they are really only anticipating PEOPLE visiting their website and interacting with just the data they're interested in. A program that scrapes large amounts of data from their website is definitively doing something they didn't design the website for. Moreover, scraping puts a higher load on their webservers (as compared to an API server which is designed to field such large-scale, program-driven requests).