# Project Pipeline

**Statistics 4996**
**Spring 2020**

**Pipeline:**

1. **Initial data exploration:** Get to know your data, identify errors, characterize missingness, explore variables including their shapes, and their associations with other variables

2. **Analysis Plan:** Repeat as often as necessary. You should have completed at least one iteration by the analysis plan presentation in February.

   (a) **Revise question:** Take what you have learned so far and revise the questions of interest.

   (b) **Determine a plan for analysis:** What types of models/algorithms will you use? Which methods are approrpriate for your data and for your question of interest? Which variables will you use? You will want to clean your as appropriate for your methods. What tests are appropriate for helping you answer your question of interest?

   (c) **Back to the data:** Given your choices above, does your data meet the required assumptions? Do you need to find more data? Are there issues with your data (or can you not get the right data) that will prohibit you from running the plan of analysis outlined above?

3. **Run analysis:** After you have iterated over the above (this often requires several iterations), run your analysis.

4. **Interpret analysis:** What do the results mean? What is surprising about the results? What are some assumptions that you made that might impact the validity of your results? Reflect on what your next steps would be if you had the time/resources? (if you DO have the time/resources, consider taking that next step!) **Interpretation is often lacking, but is crucial for being a good statistician.** There should be a lot of deep thinking going on at this step.

5. **Prepare deliverable**

**Comments:**

- A large portion of the time you spend on this project will be in step 2 and its many iterations. If this step goes by quickly, or you don't feel the need to iterate, you should stop and assess. This often happens for a few reasons

  – You haven't thought carefully enough about your data and your questions of interest. Specifically, can your data actually answer your question of interest? Or is it answering some kind of a proxy?

  - You haven't thought carefully enough about your methods of analysis. Are these the best methods? or just the ones that came to mind the quickest? Every method of analysis has limitations. Are you OK with them? This is a balancing act between your question, your method, and your data.
  - Perhaps your question of interest is too easy. Finding the right question of interest is quite difficult. Make it too hard, and it won't be answerable. Make it too easy, and it will be boring.

- You may get to the interpretation part of the pipeline and feel that there isn't much there– that there really isn't much to think deeply about. This is often because the question is too easy (see above). Really spend time with step 2. Make sure you are completely satisfied with your question, approach, and data.