Springboard Capstone Project Proposal

**1. Charitable Donation Dataset**

This dataset is taken from the Second International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-98 The Fourth International Conference on Knowledge Discovery and Data Mining.

Here is the link to the metadata:

http://kdd.ics.uci.edu/databases/kddcup98/epsilon\_mirror/cup98dic.txt

With this dataset, I hope to be able to practice:

- Data wrangling: prepare the dataset for analysis, and clean some formatting errors

- Apply some machine learning techniques to predict who are likely to donor (for example, based on their demographic and socio-economic attributes), as well as the amount of their donations.

Since the dataset is very big (almost 2 million records and 481 attributes), it will be necessary to do feature selection and/or reduction.

**2. Lending Club Dataset**

This dataset contains data about loan applications that have been rejected and issued by Lending Club. With this dataset, I would like to try both categorization techniques (to predict which loan loans are more likely to default) as well as regressions (how much will be paid back on a given loan.)

Data source:

https://www.lendingclub.com/info/download-data.action

**3. Prudential Life Insurance Assessment**

This dataset is taken from the Prudential Life Insurance Assessment Kaggle competition. The goal is to identify risk classification and predict a response variable (an 8-level ordinal measure of risk).

Data source:

https://www.kaggle.com/c/prudential-life-insurance-assessment/data