Here my initial three capstone project ideas – my goals from the project are 1) Sufficient Python coding experience to build a predictive model or models 2) Big data experience and 3) Data management using SQL.

1. **Home loan credibility prediction for the first ever borrowers**: People who have no credit histories struggle to get loans. It is also a big challenge for financial providers to put themselves in risky situations. A loan agency strives to broaden financial inclusion for the unbanked population. To do so, they want to make sure that they can predict the clients' repayment capability using a variety of alternative data including telco and transactional information. This is a supervised classification problem with big datasets.

   **Data:** from https://www.kaggle.com/c/home-credit-default-risk/data
   There is a set of seven data files – including the applicant's previous credits history, cash balance and their previous application history along with the current application table. The column description file shows that there are over 200 columns overall.

   **Methods:**
   - Preparation: missing data handling and manipulation
   - EDA: visualising and statistical summaries
   - Modelling: feature selection, optimisation and machine learning

2. **Prediction of dengue disease appearance for two location**: Dengue fever that is a mosquito-borne disease has been spreading globally in recent years – these days many of the nearly half billion cases per year are occurring in Latin America. We want to predict the number of dengue cases each week in San Juan, Puerto Rico and Iquitos, Peru.

   **Data**: from https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/
   There are two historical datasets collected by various US Federal Government agencies including environmental information, such as temperature and precipitation, and the number of dengue cases by location every week.

   **Methods:**
   - Preparation: scientific research, missing data handling and manipulation
   - EDA: visualisation and statistical summaries
   - Modelling: feature selection, machine learning and time series

3. **Prediction modelling of sales for each item in store**: Predicting the proper amounts of products for sale is very important to save storage and budget as well as maximise the revenue in any retail business. In this challenge we want to build a prediction model and find out the sales of each product at a particular store.

   **Data**: from https://datahack.analyticsvidhya.com/contest/practice-problem-big-mart-sales-iii/
   There is a single dataset having 8523 observations with 12 features – describing the characteristics of items as well as stores.

   **Methods:**
   - Preparation: missing data handling and manipulation
   - EDA: visualisation and statistical summaries
   - Modelling: statistical modelling and/or machine learning