Project Proposal

Project Title: Online Learning Data Set
Theme: Theme 4 - Data Set Creation

1. **Team Member/Captain**: Randi Weston (NetID: rweston2)

2. **Project Description**:

I would like to create an online learning data set for use in LiveDataLab leaderboard competitions and future CS 410 group projects. I plan to scrape course data from Coursera, EdX, Udacity, and FutureLearn and curate this data into data sets specific to each learning platform and into a large, overarching data set containing data from all four learning platforms. The large data set containing course data from all four learning platforms should come close to 5,000 courses worth of data. I believe these data sets will be of use to future students seeking to pursue group projects aligned with the "Intelligent Learning Platform" theme. I believe that these data sets will be particularly useful for anyone looking to create an online learning search engine or recommendation system.

Differentiation from Existing Data Sets:
Most of the other online learning related data sets I could find were either focused on a single online learning platform or focused on online learner analytics and sentiment. My data set is different because it combines data from several different online learning platforms which would enable it to be used for an online learning search engine or recommendation system.

The following data sets are closest to mine (with the first in the list, "Online Courses", being closest overall), but they only focus on one or two online learning platforms, a subset of institutions on a single online learning platform, or a subset of course types on a single online learning platform:

- Online Courses: https://www.kaggle.com/datasets/khaledatef1/online-courses/
- 1000 Online Courses: https://www.kaggle.com/datasets/mohamedhanyyy/10000-online-courses
- Online Courses from Harvard and MIT: https://www.kaggle.com/datasets/edx/course-study
- Edx Courses: https://www.kaggle.com/datasets/imuhammad/edx-courses
- Udemy Courses: https://www.kaggle.com/datasets/hossaingh/udemy-courses
- 42k+ Udemy Course Enrollment Information Dataset: https://www.kaggle.com/datasets/songseungwon/2020-udemy-courses-dataset
- Online Data Science Courses:

https://www.kaggle.com/datasets/antonkozyriev/online-data-science-courses

Data Set Potential Use Tasks: My data set could be used to create an online learning search engine that searches for course data across four major online learning platforms (Coursera, EdX, Udacity, and FutureLearn). My data set could also be used to create an online learning recommendation system based on courses on these platforms. Lastly, I believe my data set could also be used to create a browser extension that would display similar courses across these platforms that have a higher rating than the course page a user is currently viewing. Most importantly, my data set can offer the ability to compare courses across Coursera, EdX, Udacity, and FutureLearn without the need to try and combine existing individual datasets which may or may not contain similar variables/columns.

Creation of the Data Set: I plan to write four different web scrapers (one for each online learning platform) with Python and Beautiful Soup. I will then run my web scrapers on a small amount of data from each site first, so that I can refine the scraping process and reduce the amount of data cleaning I will need to do. I will then run the web scrapers on the entirety of the course catalog from each platform. I intend to offer my data sets in csv format. If time allows, I'd also like to create JSON versions of my data sets.

3. **Tools and Resources**:

   I intend to write my web scrapers in Python and intend to use the Beautiful Soup library.

4. **Project Plan (22 hrs)**:
   - Development (10 hrs): Write web scrapers for each platform.
   - Testing (1 hr): Test web scrapers on a small number of courses.
   - Revision (2 hrs): Make any necessary revisions to scrapers to reduce data cleaning.
   - Web Scraping and Data Cleaning (4 hrs): Scrape all courses and clean data.
   - Documentation (2 hrs): Document data set.
   - Project Administration (3 hrs): Creation of progress report and presentation.