Project Proposal

Isaac Getto, Jee Hyun Kim, Tae Jin Kim

In a couple of paragraphs, describe the key ideas of your proposed project? What is your MVP? What are your stretch goals?

MVP

- A tool based upon statistical analysis of the dataset, that helps the student or user determine what classes they should take based on their major, and how certain/uncertain they are.
- Create a new column with pre/ co requisites
- Visualization of an analysis of graduating classes, courses, professors, and majors by looking at gender balance, major distributions, and other factors within these categories
- (The goal here is not final as we are unsure about where we exactly want to work. The goal currently covers both data trimming/manipulation, statistical analysis, and machine learning and seems very balanced. However, we want to explore the data and get a more concrete feel for what we can achieve before making the final call.)

Stretch Goals

- Compact website holding both analysis and tool for easy use.
- Interactive statistical visual representations that allow further insight of trends among students and courses.
- Suggestion for courses/ majors for the current students

To the best of your current knowledge, what datasets will you use for your project? Are there any obstacles you foresee in terms of getting access to the data?

- Olin Course Registration Data from 2002-2014
- (Pending) Olin Course Pre-Registration Data from 2002-2014
    - Limited information as only cumulative totals are given

What are the most important new skills / techniques you will have to learn to be successful in this project?

- Data Cleaning & Manipulation
- Clustering / unsupervised learning
- Visualization: Possibly ***Plotly***?
- Making an interactive website/tool

Outline a rough timeline for the major milestones of your project. This will mainly be useful to refer back to as we move through the project.

Week 1: Get all necessary datasets + Ideation + create website
Week 2: Clean data & Explore data individually & discuss
Week 3: Finalise on our product idea, split up work
Week 4: Statistical Analysis and Visualization
Week 5: Develop Interactive tool for student use + project documentation
Week 6: Reorganize work into presentable form + project website

What do you view as the biggest risks to you being successful on this project?

- Generating a tool that is not already implemented and enhances the process of course selection greatly.
- Gathering data quickly and manipulating it to suit our needs early on as we do not have an overwhelmingly big data set.
- Balancing statistical analysis with machine learning, and giving both components of the project a thorough run through.

Given each of your YOGAs (see here), in what ways is this project well-aligned with these goals, and in what ways is it misaligned? If there are ways in which it is not well-aligned, please provide a potential strategy for bringing the project and your learning goals into better alignment. **There should be an individual section for each person on the team addressing the fit between the YOGA and the project topic.**

Isaac

- Most of my learning goals were centered around machine learning topics, but there might not be many opportunities to apply ML to the dataset. Still, I will get a chance to learn about some unsupervised learning techniques when we use clustering to group students into majors. I think I should also change some of my learning goals in order to make them match the work I'll be doing on this project.

Jee

- My goals included becoming more proficient in cleaning, manipulating and understanding the data. Due to the nature of our project (working with school course registration data), I will be able to understand and relate to the data better. The project involves cleaning and manipulating data. My other goals was working with data which I have to extract the values myself and learning D3. The dataset we have has straightforward data; I could see what else could be added to the data which

could help in analysis of the data. For D3, creating a webapp is a stretch goal and we probably would not be making a webapp. However, I can modify my learning goals to learn to create better visualisation, plotly and ipython widgets to align with our goals.

TJ

- My learning goals were implementing statistics, data cleaning efficiently, and using github effectively. Although I wanted to stray away from machine learning, the machine learning component is still in the project, but that is balanced out with the initial statistical analysis we're going to apply to the data. The dataset we are going to use has few columns, so data manipulation will be crucial to generate new data off of the ones we already have to further contribute to our analysis. Lastly, we plan on using github extensively, and expanding on that by using Jekyll, a github based blog creator to supplement my understanding of github.