For my project, I intend to scrape data from a website because this is a skill that I am not as familiar with and I want to challenge myself on the final project to use this approach rather than trying to mine data that is already available. I have quite a bit of extensive experience with pandas, so I think if I just used a data set then the difficulty might come from the cleansing process, but through this I have to work on so many other aspects that are new to me. I have almost no experience with html, so I will take this as an opportunity to continue to understand how html is structured for potential future utilization. I also have just learned about nltk, which I think is a very exciting way to quantify language, so I want to make sure that I get to use this more, even though this will be a challenge for me as well. There are also posts from different climbers where they talk to each other about areas within a location, and I may try to mine this as well if time permits.

In the past I have used a site called rockclimbing.com, which is a public site that anyone can share information on routes, gear, pictures, videos and many other things to educate other climbers. When I traveled to Hawaii I used this site to connect with someone, and to be able to climb while I was out there. The rock was blue and unforgetable and I treated this website as such a great resource for so long to prepare myself on any trip. I checked the terms of service very carefully for the site and it did not state anything about reusing the data for different purposes. It primarily focused on the fact that climbing is inherently dangerous and liability is on you as the consumer of the information.

Some of the questions that seem interesting to me as I thought about this project were: where are some of the best places to climb from an opportunity standpoint, how can I determine where I should climb based on my current skillset, how do people talk about the climbing that they have experienced? I want to base all of these initial questions around a couple of different criteria. This criteria will include sentiment analysis on how people talk about the country with regards to climbing as well as the shear amount of routes available. I want to look at difficulty of routes as well as what type of climbing is available within this country: top rope, bouldering, sport climbing, or trad climbing. For myself personally I have stopped climbing as frequently as I use to, so I was thinking could I use this analysis as a recommendation tool potentially based on someone's current performance, so that they could challenge themselves just enough.

My approach will be to utilize several different libraries and Spyder in order to be able to analyze the data appropriately. For extraction of the data I will use requests and beautiful soup. None of this data is formatted in tables, so I cannot attempt to use pandas from the onset. I will need to structure the data into a dataframe to be able to perform my desired analysis on it. I will then use pandas to understand where do the most routes exist (count), and difficulty of routes within a geographic area (count, mean, spread of the difficulty). I will also use nltk for understanding the sentiment around the country through the words published about the country and its climbing, which will show me how the local population feels about the opportunities within the area. As I begin to mine the data I feel that there will be more opportunities for further use of pandas, but I can only share what I can see from the html viewable content, which is route difficulty and sheer count of routes.

I am very excited to see what I can create from this analysis and really wondering if I can make something that has configurable inputs potentially for others to be able to use.