**Data Science Final Project.**

**Comments from Professor.**

**Interesting project! There is concerns over the size of your data and lack of any data wrangling element. You’ll want to think about these and how best to move forward.**

* I got the databases for 33 countries. For some of them for two years.
* Looking at 3 countries:
  + Armenia: 2,710 values
  + Bangladesh: 9,125 values
  + El Salvador: 3,604 values

Choose countries similar.

**- Problem**

**- Very clear and well defined. Great work!**

**- Data**

**- What is the unit of observation? Is it survey respondent-country-year?**

* There are several options. I propose Country-Young individual

**- Given a worst case scenario, and you only have the 2004-2006 data for 10 countries, will there be sufficient data there for analysis? What is the N, remember the data will need to be bifurcated into a training and test dataset.**

* Problem solved. We have a huge database.

**- Data wrangling**

**- I don’t see any data wrangling element here. There is nothing tricky about downloading curated data and using it as is. You need to do more here.**

* Not clear. As an starting point, I need to identify the variables I will use because each survey have different questions.

**- Consider bringing in country level information about the state of the economy and other job opportunity metrics. These data can be gathered from the World Bank and other similar sources.**

* Is it necessary if I already have a big database for individuals?

**- ML**

**- Why just run a probit model? Why not explore some other the other algorithms we covered in class to see what is the most predictive. You’ll learn next time how to back out feature importance from these models.**

* Sure, I can do more on this.