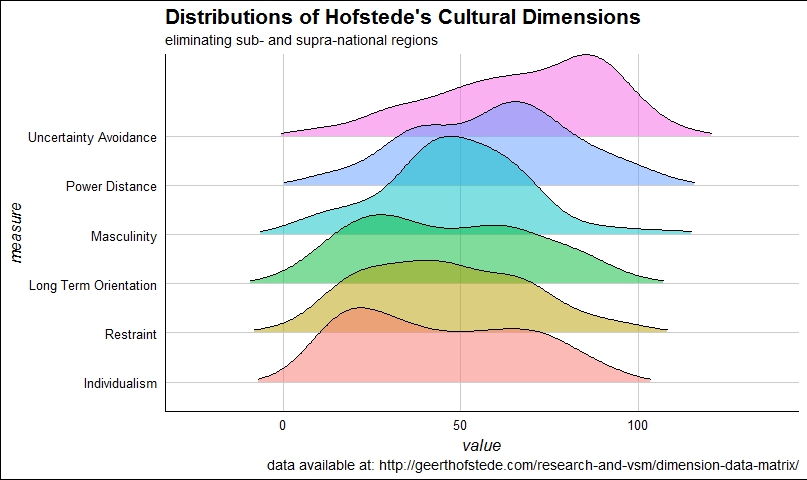
## What datasets will you use for your final project?

 As noted in my pre-proposal-proposal, one dataset will be Geert Hofstede’s cultural dimensions data, located [here](geerthofstede.com/research-and-vsm/dimension-data-matrix/). These data describe six dimensions for a list of about 100 countries (it will be required to exclude sub- and supra-national designations, e.g. Scotland or “West Asian States.”). These countries will become my units of analysis for the rest of the project.

The second sets of data are UN-generated data on human development, located [here](http://data.un.org/Explorer.aspx?). There are several datasets of human development related data that I have already pulled from the web but not yet cleaned:

* Gender Development Index
* Gender Inequality Index
* Human Development Index and its components
* Inequality-adjusted Human Development Index.

The goal, as I see it now, will be to either merge several of these or to use just one (inequality adjusted HDI) for my second dataset. Each of these has a unit of analysis at the country level (but not yearly, unfortunately. This could be a wrench in my system; see below).

Finally, last Friday night, I exported about 50 or so “datasets” from the UN (same location) that pertain to education. These vary from datasets related to gender and education, poverty and education, and monitoring of the Millennial Development Goals and the Sustainable Development Goals (I’m very passionate about these!). The original idea was to merge these ~50 datasets into one dataset, and then have that super-set be my *third* dataset to be combined. I hit a slight snag with their structure. Each one of these datasets really only represents one variable, with the unit of analysis being country-by-year (with about 5-10 years per country). This, I believe, satisfies your request to include more longitudinal data to increase the complexity. I renamed the files so that I could easily loop through all 59 of them to append them and then turn them into a wide formatted dataset to merge with the prior two. I know this is possible with the broom package in R and I’m sure there’s something similar in Stata that I can use to go from super-long to wide. However, the real snag came from the fact that downloading from the UN generates a filename that reflects the *time* at which the variable was downloaded, not the variable itself; nor is there any consistent documentation within the CSVs being output about which was which. This means I’ll need to re-download them one at a time and modify the variable names within the spreadsheets before I’ll be able to loop through them to combine them into one data set. Having said that, I *think* that by adding multiple years and looping through multiple variables to generate datasets in this way prior to merging them with the Hofstede data is sufficiently complex to meet the requirements.

## Where is the data located?

Data are available from the links provided above; but also here:

geerthofstede.com/research-and-vsm/dimension-data-matrix/

http://data.un.org/Explorer.aspx?

## What format is the data in currently?

Hofstede’s data are currently in .xlsx format (I did some cleaning on them prior to realizing I would use them for this assignment, but I can use the original .csv files as well, as needed. The Human Development Index files are also excel files, whereas the individual variable education files from the UN are all in CSV format. There may be an argument for saving them all to CSV in order to streamline the import process, but that might be a decision I make further down the road.

## Provide a list of several (at least three) descriptive research questions your final dataset will allow you to explore

I’m waffling a bit here because the questions that come to mind most frequently are correlation and regression questions, which are likely outside the scope of this part of the assignment. I would like to be able to describe the relationships between Hofstede’s dimensions and variables included in UN Data, such as:

* Is there a connection between the Dimension of “masculinity vs. femininity” and the gender distributions of students and teachers in schools, or the outcomes for boys and girls in graduation?
* Is there a relationship between the “long-term versus short-term orientation” dimension and human development of countries (i.e., are poorer countries more likely to be prone to short-term thinking because of their economic constraints?)
* Does the “individualism vs. collectivism” variable predict anything about the nature of the school systems in question in the UN Data, e.g. meeting MDG target 2.A “all children everywhere in full course of primary schooling” or “enrollment in public v private education” (that is, are more collectivist societies more likely to have stronger public schooling systems vs. individualist societies having more privatized systems?)

## Which variables will you likely use for your analysis?

I’m likely to use all 6 of Hofstede’s dimensions. As for the HDI data set, I’m leaning more toward the inequality-adjusted one because I think it will regress better without the major outliers created by states with high inequality (mainly I’m talking about us/U.S.). Finally, I will have a better idea of variables of interest from the UN education data once I have pulled a few dozen of them together, but mostly they deal with enrollment, gender, rates of graduation, teacher preparedness statistics, and things of that nature.

I hope this is clear enough to give a good picture of where I am after the weekend. I know I still have a few things to figure out but step one will be solving the individual-data-set problem created by the UN’s education data set structures. After that, things should be a little clearer for me.

Thanks!