What I did:

1. Downloaded all data.
2. Looked at variable definitions to gain understanding of variables.
3. Merged European data with rest of world data.
4. Made sample choropleth map to show happiness ratings around the world
5. Performed some preliminary analyses using first 10 variables:
6. dimensionality reduction via PCA to decompose into (religion important, poor health group) and (religion not important, politics not important) groups.
7. Plotted average values per country of these 2 components, saw negative correlation between the two PCAs when grouped by country
8. Performed regression analysis with happiness as outcome, saw that overall health was positively correlated with happiness, while work being important was negatively correlated

What I plan to do today:

1.Clean data, including making dummies of certain categorical variables to use in PCA analysis.

2. Develop PCA-choropleth pipeline, so I can easily color countries by their values for principal components.

3. Perform PCA on dummified feature matrix; label new principal components.

Possible snags: need to think of way to deal with complex sampling design, i.e. weights, psu, strata. The following seems to be a good way to deal with PCA with weights : <http://stats.stackexchange.com/questions/113485/weighted-principal-components-analysis>

Also, missing data: maybe use matrix factorization or k means on pca componenets to fill them in .

What I did: cleaned up most of first 240 variables

Performed pca: first two components account for only 36% of variation

BUT could this be because I am imputing mean for missing variables (leading to lack of variance across observations)

To do:

Separate by year

continue cleaning variables , making sure ordering is correct

Find better way to impute, maybe k means on pca or matrix factorization

Think about how to include weights in PCA

Predictors of happiness

Choropleth map pipeline