Assumptions on clustering:

1. Each coefficient can be interpreted as: controlling all other variables, one unit increase on this certain variable can cause how many changes on predicted scores.
2. Each estimate of coefficient from the training model is unbiased, which means, we suppose our model has included all relevant variables (the residual is uncorrelated with the target variable).
3. Under assumption of 1 and 2: the number of variables used in the model has nothing to do with the interpretation of coefficient.

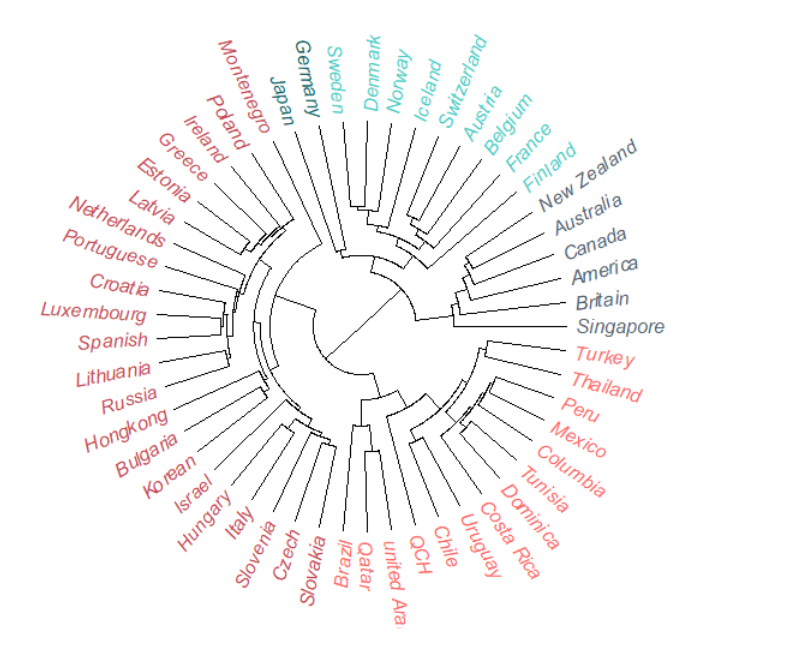
Current training method:

Treat each country as an independent population, filter the variables to ‘names\_selected’, then partition the training and testing set. On the training set, factorize the categorical variable, delete near-zero-variance-variable, conduct median imputation and get a training model and the coefficients. In the clustering section, choose the variables that exists in at least 90% of countries, conduct median imputation and scaling, and then do the hierarchy clustering.

Some critical points:

1. The result of clustering is greatly influenced by the variables chosen. We should make sure that the variables should be common enough across different countries (number of NAs should be controlled under a low rate). Besides, the variable numbers should be large enough (I’ve once tried to set the controlled NA rate as zero and use the remaining variables to cluster, and the clustering result is terrible and unreasonable)
2. The result of clustering can be partly interpreted by culture influence and gdp.

Current results:



clustering results

Result summary:

Five clusters can be partly summarized as:

1. Northern European countries
2. British Commonwealth、America, Singapore
3. Developing countries (Southern America, China, Arab)
4. Eastern Europe, Mediterranean
5. Germany and Japan

Strange points:

1. Chinese and Arab countries are in the ‘poor group’
2. Germany and Japan are one specific cluster.

Further analysis:

1. Calculated the mean of each variables for countries in the five groups (‘group\_analysis.csv’) to help take a look at the group differences.
2. Deep analysis on abnormal countries (like Germany): the NA rate are very low, so it shouldn’t be the imputation problem.
3. Stability:

the three of us have all made the clustering analysis but have very different results. Meanwhile, even for my own work, a slight change in parameters (like the correlation threshold, NA rate threshold) will cause some countries (like Germany, China, etc) to change their group.

However, some patterns are stable, for instance, in my work, the northern European countries are always in one group.

1. About controlling number of variables: for interpretation, I don’t think it’s necessary as I’ve stated in the assumption 3.
2. For better interpretation:

As I’ve discussed with the other two before, from the point of view of statistics, I think it would be practical if we can take different subsets from the training set, train the model repeatedly, record the coefficient each time and then calculate its mean value as the coefficient estimate (From assumption 2, we can assume that the coefficient is unbiased so that the mean value can be a good estimate of the actual coefficient).

\*\* this idea works for the simple regression method, but since this is ridge model (different cost function), I’m not really sure about the unbiasedness conclusion. If the conclusion also works, I can run the code again to see if there’s a change.