DNSC 6211 Programming for Analytics
Assignment 8 Report
Sonya Tahir, Amit Talapatra, Wendy Zhang, Wei Zheng
10/31/15

In this assignment, the first step was identifying the 10 indicators (out of 42) with the least amounts of missing values. From these, 5 were selected for the k-means clustering and MDS analysis. The selected indicators are listed in Table 1:

Table 1. The five indicators selected for k-means clustering and MDS analysis

Indicator	Description
SE.PRM.ENRR	School enrollment, primary (% gross)
SL.UEM.TOTL.ZS	Unemployment, total (% of total labor force) (modeled ILO estimate)
SE.PRM.ENRL.TC.ZS	Pupil-teacher ratio, primary
SL.TLF.TOTL.IN	Labor force, total
SH.DYN.MORT	Mortality rate

These 5 indicators were selected out of the 10 because they each represent a different metric, and using them allows us to include a wider range of elements of a country when comparing countries for similarity. Some of the indicators that were not selected out of the top 10 include: unemployment rate for males, unemployment rate for females, gender parity in school enrollment, and two population subsets for different age ranges. The gender-based unemployment indicators were excluded because a total employment metric was available and we wanted to minimize overlap in the relationships between our indicators. The same reasoning was used to select primary school enrollment over gender parity in enrollment. Finally, the population subsets for different age ranges were not included because they are directly dependent on each other and tell us less about the effectiveness of a country's education system or the quality of life of its citizens. The five selected indicators tell us the most about the state of the country with respect to these parameters, which is why they were chosen to compare similarity between countries.

For the cluster and multidimensional scaling (MDS) analyses, we chose to plot the relative positions of countries based on the five indicators. An alternative approach would have been to plot points for each indicator, where each point combines the data from all countries. This would have let us see the relationships between the indicators themselves, showing us if certain indicators correlate more strongly with each other. The graphs we chose to plot, based on relationships between country data, are shown in Figures 1 and 2. Labels for closely-grouped countries are not shown to make the plot easier to read.

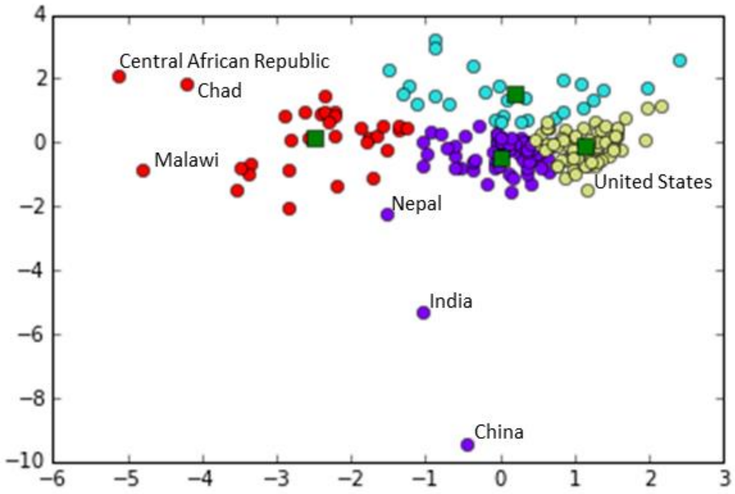

Figure 1. K-Means Clustering with 4 Clusters

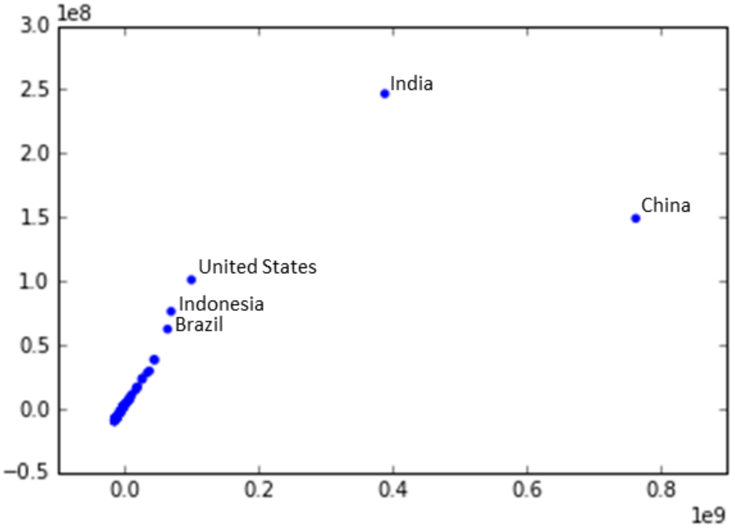

Figure 2. MDS Analysis

The groupings in the cluster analysis indicate that countries in the same grouping overall have similar values for the five indicators selected. By this measure, the groupings make sense. This analysis can be used as a measure of a country's education system and quality of life, though this is a very limited set of indicators with which to measure this. For both types of analysis, we see that India and China are outliers, indicating that their values for the five indicators vary significantly from those of the other countries. One possible explanation is that India and China have significantly larger populations than the other countries, resulting in much larger values for the 'labor force' indicator. This may skew the other four indicators, which would have less variability.