Segmentation and Clustering

Background and Key Decisions

A United States retail store chain is considering to expand to other countries. They would like to start this process with countries that are similar economically and demographically.

The goal is to recommend a curated list based on various economic, demographic, education, and environmental factors.

Data from the World Bank web site is used to decide which countries are most similar to the United States and could be likely candidates for expansions.

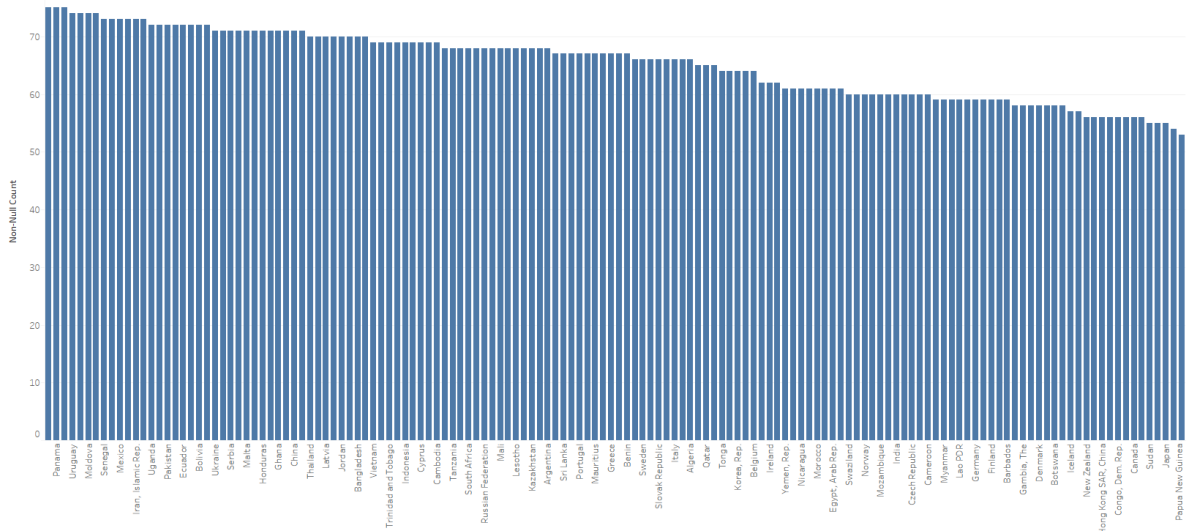
In order to provide accurate analysis, we would need data for the United States and similar data for the other countries. Some examples that would be helpful for this analysis within those demographic categories include:

- Percentage of the population age 15 and above who have general 'literacy'
- The percentage of population (age 25 and over) with at least completed secondary education
- Employment to population ratio is the proportion of a country's population that is employed
- Total labor force, meeting the ILO definition of economically active population
- Access to electricity is the percentage of population with access to electricity
- Population living in slums is the proportion of the urban population living in slum households
- Average number of pupils per teacher at a given level of education

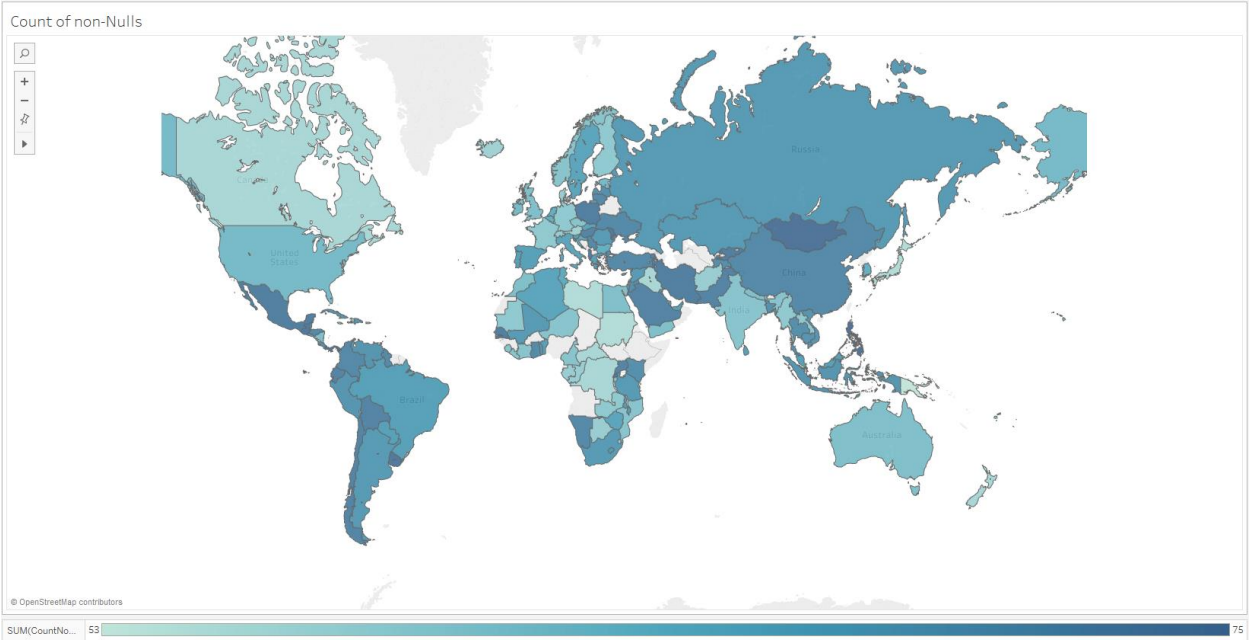
Data Exploration and Cleanup

The demographic data from World Bank contains 215 countries and 77 variables.

Not all of the countries have complete data; in order to avoid false bias, any country with 25 or more missing values was removed from analysis. 71 countries met this criteria. These are the 144 remaining countries, sorted by the countries with the most values.



The same 144 countries, represented geographically.



Of the 77 variables, 9 were removed from the data set because they did not fit the overall objective. They belonged to the background and health categories.

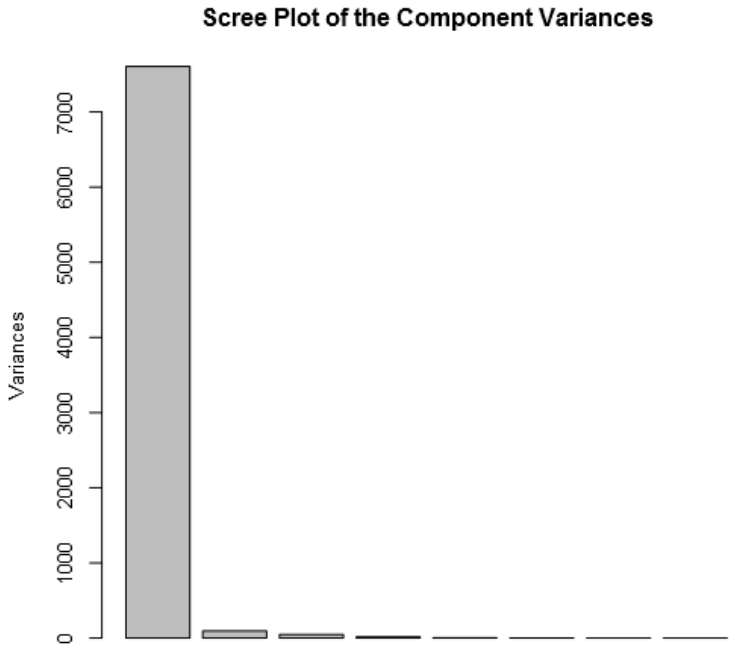
IT_NET_USER_P2	Background	Internet users are individuals who have used the Internet (from any location) in the last 12 months. Internet can be used via a computer, mobile phone, personal digital assistant, games machine, digital TV etc.
SH_DYN_AIDS_ZS	Background	Prevalence of HIV refers to the percentage of people ages 15-49 who are infected with HIV.
SH_DYN_MORT	Background	Under-five mortality rate is the probability per 1,000 that a newborn baby will die before reaching age five, if subject to age-specific mortality rates of the specified year.
SH_MED_PHYS_ZS	Health	Physicians include generalist and specialist medical practitioners.
SH_XPD_PCAP	Health	Total health expenditure is the sum of public and private health expenditures as a ratio of total population. It covers the provision of health services (preventive and curative), family planning activities, nutrition activities, and emergency aid designated for health but does not include provision of water and sanitation. Data are in current U.S. dollars.
SN_ITK_DEFC_ZS	Health	Population below minimum level of dietary energy consumption (also referred to as prevalence of undernourishment) shows the percentage of the population whose food intake is insufficient to meet dietary energy requirements continuously. Data showing as 2.5 signifies a prevalence of undernourishment below 2.5%.
SP_POP_DPND	Health	Age dependency ratio is the ratio of dependents--people younger than 15 or older than 64--to the working-age population--those ages 15-64. Data are shown as the proportion of dependents per 100 working-age population.
SG_VAW_BURN_ZS	Health	Percentage of women ages 15-49 who believe a husband/partner is justified in hitting or beating his wife/partner when she burns the food.
SH_TBS_PREV	Health	Prevalence of tuberculosis is the estimated number of TB cases (all forms) at a given point in time, expressed as the rate per 100,000 population. Estimates for all years are recalculated as new information becomes available and techniques are refined, so they may differ from those published previously.

Variable Reduction

Principle Component Analysis was used for variable reduction in the following categories: Education Literacy, Education Average Years, and Education Percent. Each category originally had 6, 30, and 18 variables respectively. I chose to include components which accounted for approximately 90% of the variance. For the respective categories, the reduction resulted to 1, 1, and 10 principle components.

Education Literacy Component Variance and Plot

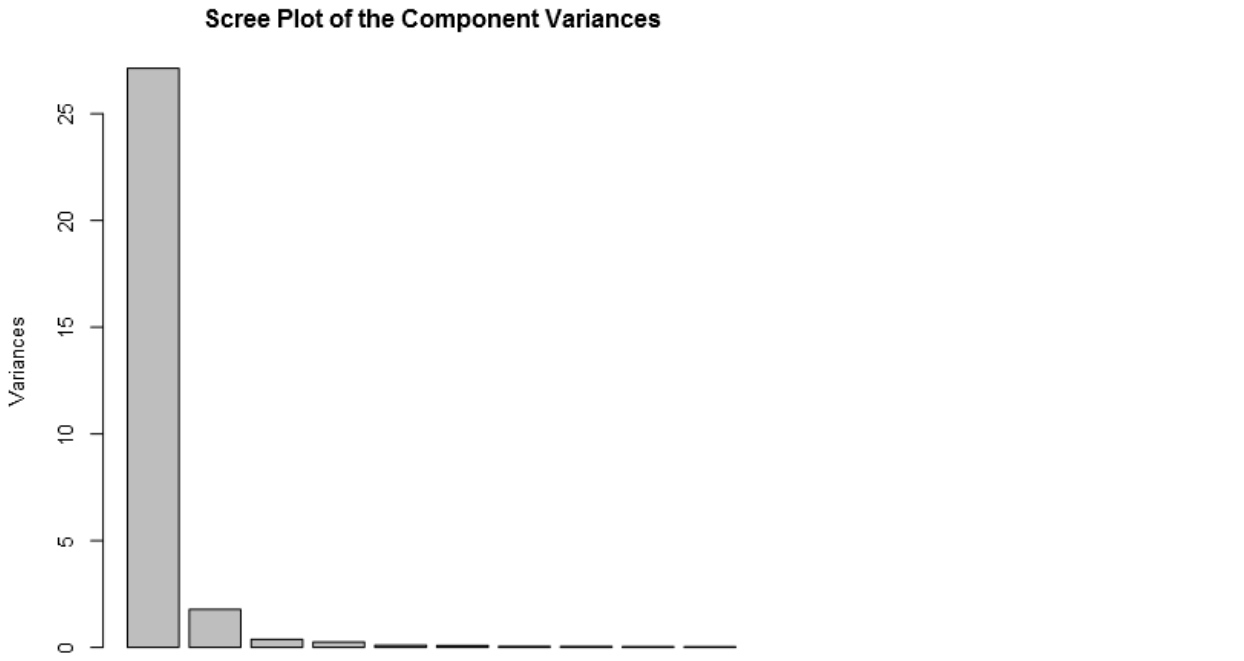
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard Deviation	87.20914	9.727279	6.903417	4.09137	2.120337	1.444855	0.121049
Proportion of Variance	0.978688	0.012176	0.006133	0.002154	0.000579	0.000269	2e-06
Cumulative Proportion	0.978688	0.990864	0.996996	0.99915	0.999729	0.999997	0.999999
	PC8						
Standard Deviation	0.069374						
Proportion of Variance	1e-06						
Cumulative Proportion	1						



Education Average Years Component Variance and Plot

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard Deviation	5.208535	1.334216	0.614575	0.500666	0.322683	0.29471	0.227145
Proportion of Variance	0.904295	0.059338	0.01259	0.008356	0.003471	0.002895	0.00172
Cumulative Proportion	0.904295	0.963632	0.976222	0.984578	0.988049	0.990944	0.992664
	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Standard Deviation	0.215421	0.194093	0.178627	0.158806	0.133161	0.117918	0.101362
Proportion of Variance	0.001547	0.001256	0.001064	0.000841	0.000591	0.000463	0.000342
Cumulative Proportion	0.994211	0.995466	0.99653	0.997371	0.997962	0.998425	0.998768
	PC15	PC16	PC17	PC18	PC19	PC20	PC21
Standard Deviation	0.098144	0.085856	0.080471	0.060061	0.0524	0.038949	0.036862
Proportion of Variance	0.000321	0.000246	0.000216	0.00012	9.2e-05	5.1e-05	4.5e-05
Cumulative Proportion	0.999089	0.999334	0.99955	0.99967	0.999762	0.999813	0.999858
	PC22	PC23	PC24	PC25	PC26	PC27	PC28
Standard Deviation	0.033928	0.029838	0.02513	0.024625	0.018162	0.015839	0.014279
Proportion of Variance	3.8e-05	3e-05	2.1e-05	2e-05	1.1e-05	8e-06	7e-06
Cumulative Proportion	0.999896	0.999926	0.999947	0.999967	0.999978	0.999987	0.999993
	PC29	PC30					
Standard Deviation	0.012872	0.005813					
Proportion of Variance	6e-06	1e-06					
Cumulative Proportion	0.999999	1					

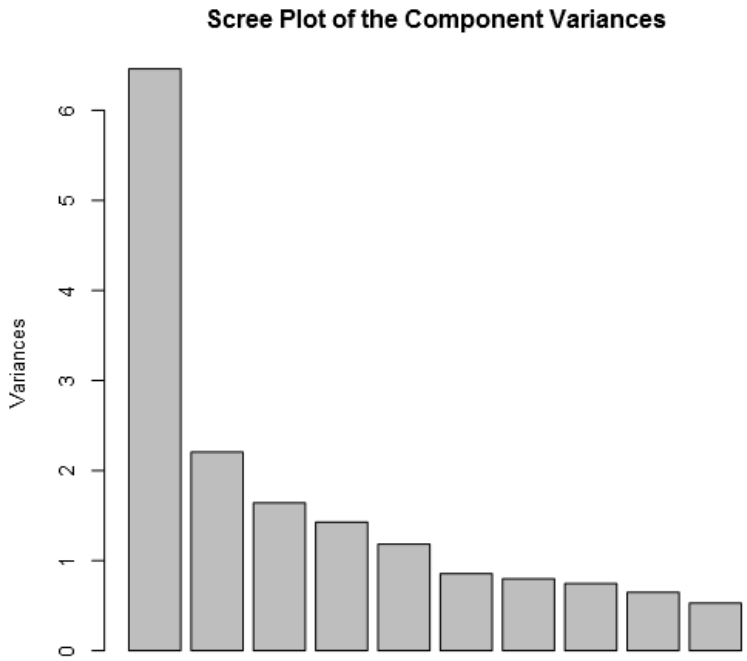
Plots



Education Percent Component Variance and Plot

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard Deviation	2.541784	1.48515	1.280832	1.194692	1.087375	0.924748	0.893456
Proportion of Variance	0.358926	0.122537	0.091141	0.079294	0.065688	0.047509	0.044348
Cumulative Proportion	0.358926	0.481463	0.572604	0.651897	0.717585	0.765094	0.809442
	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Standard Deviation	0.863419	0.805367	0.726936	0.679777	0.509922	0.481077	0.463505
Proportion of Variance	0.041416	0.036034	0.029358	0.025672	0.014446	0.012858	0.011935
Cumulative Proportion	0.850858	0.886893	0.91625	0.941922	0.956368	0.969225	0.981161
	PC15	PC16	PC17	PC18			
Standard Deviation	0.429664	0.340537	0.159675	0.114165			
Proportion of Variance	0.010256	0.006443	0.001416	0.000724			
Cumulative Proportion	0.991417	0.997859	0.999276	1			

Plots



Segmentation/Cluster Methodology

Three clustering methods were evaluated in order to create country segments: K-Means, K-Medians, and Neural Gas, by using the K-Centroids Diagnostic tool. Because we were asked to provide 4 clusters, we evaluated which of these provided the most accurate data within that limiter.

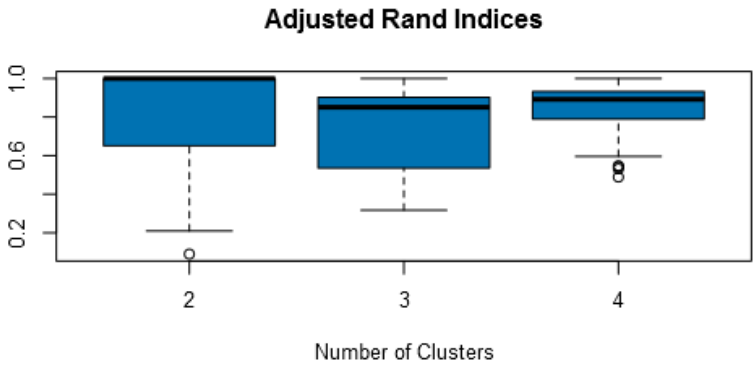
By examining the Adjusted Rand and Calinski-Harabasz Indices, we found that Neural Gas was the best performing of the three. Overall, it had the highest median and widest variance.

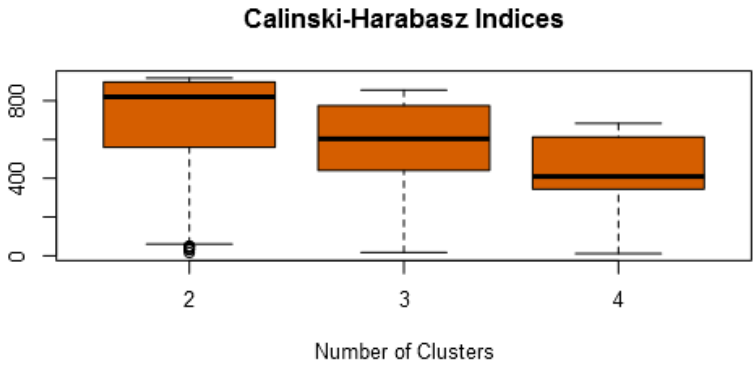
Adjusted Rand Indices:

	2	3	4
Minimum	0.08986	0.3173	0.4891
1st Quartile	0.6507	0.5447	0.7917
Median	1	0.8508	0.8919
Mean	0.8064	0.7474	0.8376
3rd Quartile	1	0.9011	0.9294
Maximum	1	1	1

Calinski-Harabasz Indices:

	2	3	4
Minimum	16.49	17.08	11.42
1st Quartile	559.9	440.2	344.1
Median	820.1	603.5	409.2
Mean	651.2	549.8	422.5
3rd Quartile	896.2	774.1	612.8
Maximum	917	855.3	683.2

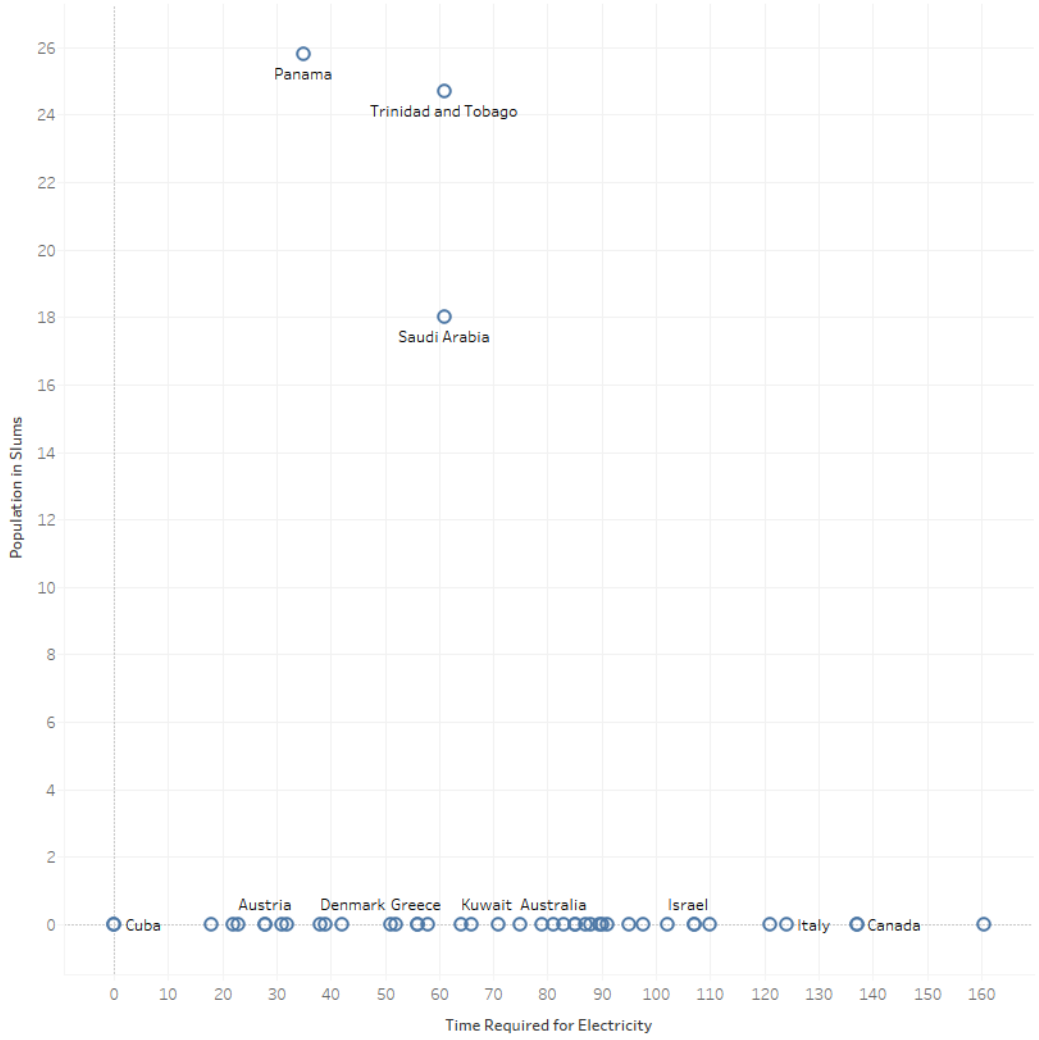


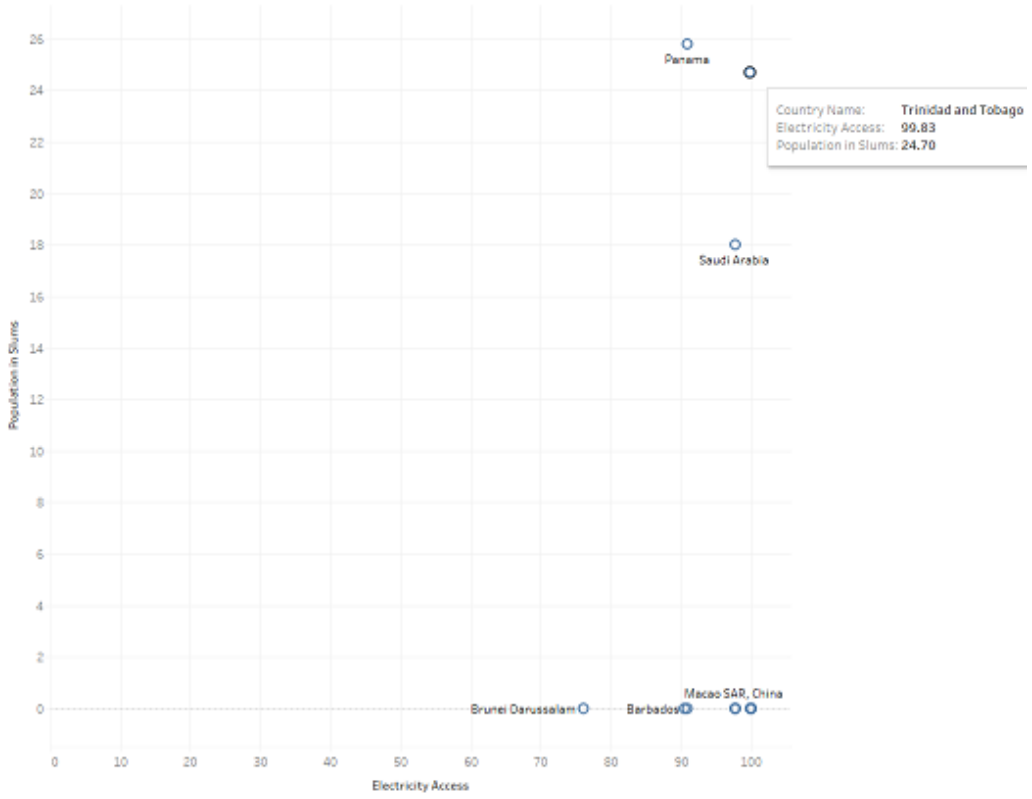


Visualization

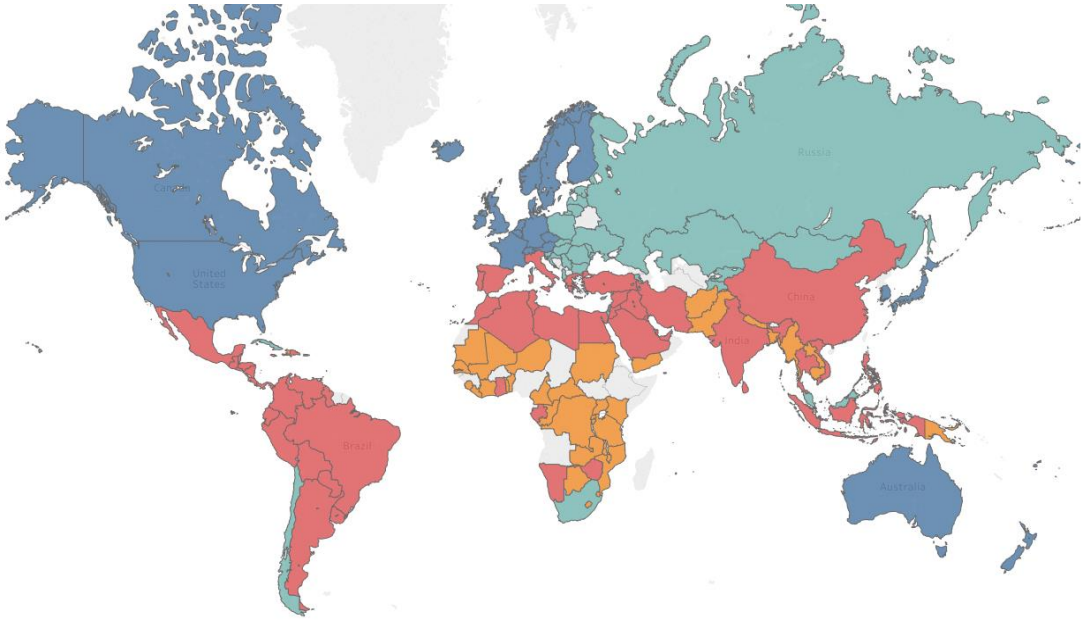
After processing the data through the model, Tableau was used to visually validate the results.

Initially there were several countries included in the United States cluster where a high number of the population lived in slums or did not have the same access to electricity as the other countries within the cluster.

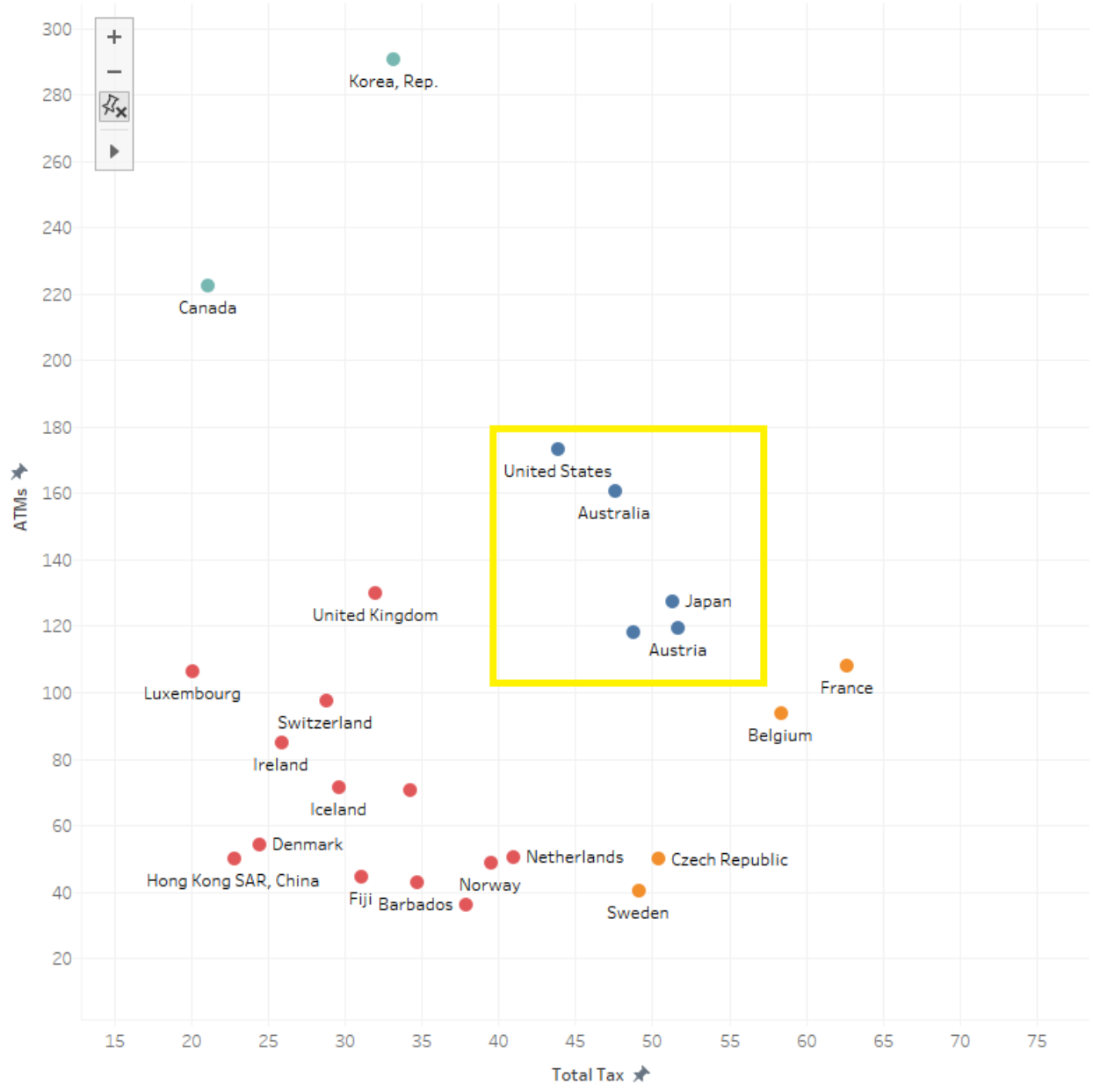




Most of the technology savvy, capitalist, or western countries were grouped together. These are countries with established public schools and easy access to education. Countries that are less industrialized, more rural, with more poverty or war, or barriers to growth seemed to be grouped together. All of these things would hinder education which was the main driver in the data.



When looking at USA in terms of Total Tax Rate and ATM machines, we can see that Australia, Japan, Germany, and Austria were the most similar.



Recommendation

Below is the full recommended list of countries based on the neural gas. These countries were proven to be most similar to the United States based on various economic, education level, literacy rates, and environmental factors.

Countries who had outliers in economic attributes where eliminated from the list. For example, these are countries with a high population living in slums or where access to electricity was limited.

Australia
Austria
Barbados
Belgium
Canada
Czech Republic
Denmark
Fiji
Finland
France
Germany
Hong Kong SAR, China
Iceland
Ireland
Japan
Korea, Rep.
Luxembourg
Netherlands
New Zealand
Norway
Sweden
Switzerland
United Kingdom