Project: International Expansion

Step 1: Key Decisions

Key Decisions:

1. What decisions needs to be made?

   A retail store chain that only has a presence of stores in the United States is thinking of expanding to other countries, but would like to start this process with countries that are similar economically and demographically to the United States. So, we need to segment the countries of the world based on various economic, demographic, education, and environment data similar to United States. Based on this, we need to decide which country would be best to open the next branch of the retail store chain.

2. What data is needed to inform those decisions? Please include 2 examples in each of the following categories: Economic, Environment, Education.

   We will need Demographic, Education and Environment data for other countries as well as the US. For example, percentage of population categorised based on their education level might be good to know for marketing similar products.

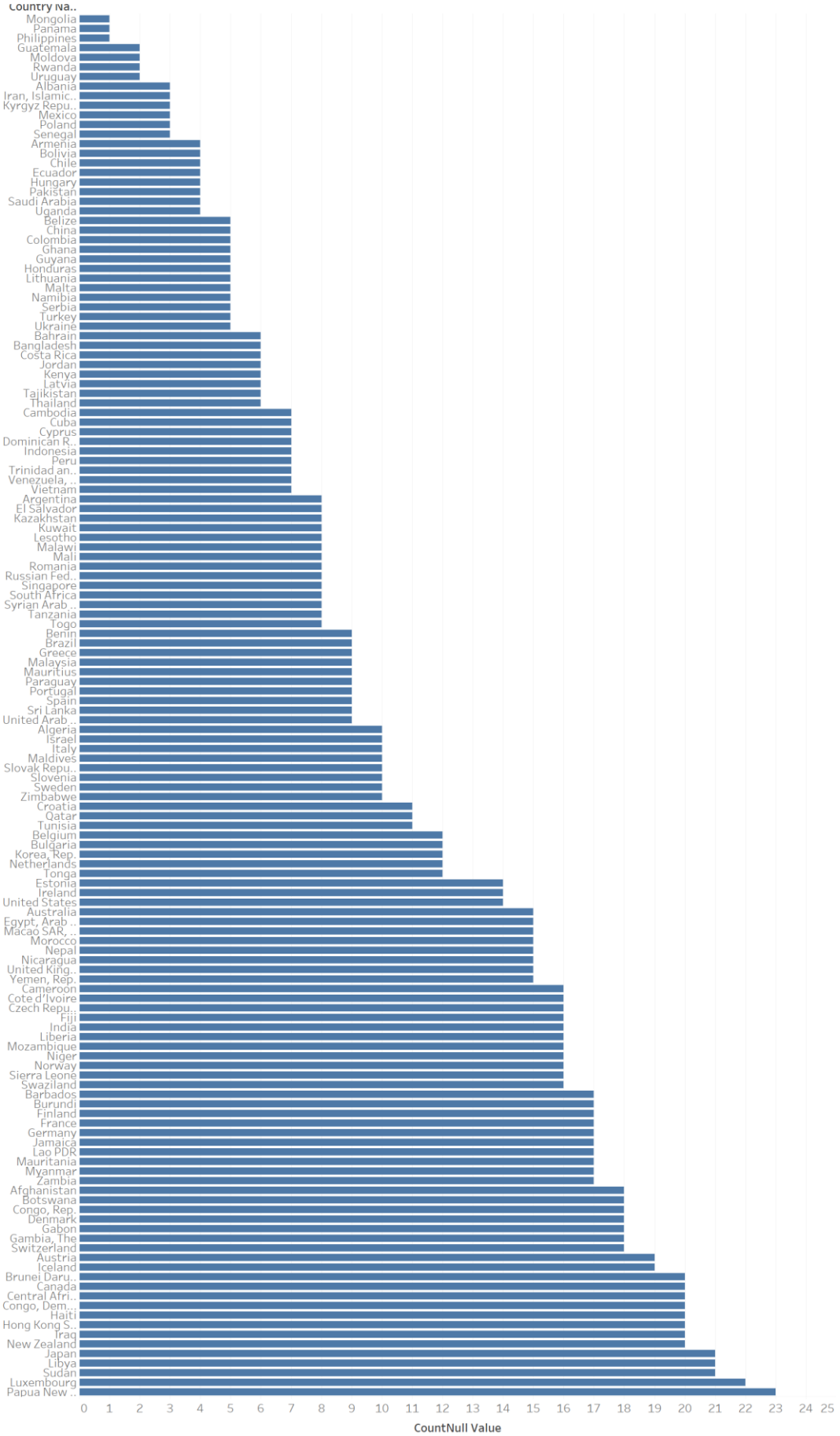
Economic	Quality of Port Infrastructure, Total labour force comprises people ages 15 and older
Education	Ratio of female youth literacy rate to male youth literacy rate, Average number of pupils per teacher at a given level of education
Environment	Population living in slums, Access to electricity is the percentage of population with access to electricity

Step 2: Explore and Cleanup the Data

1. *How many countries did you reduce your dataset to? Please include a bar chart of number of non-null data points by country, sorted from most to least.*

   The dataset is now reduced to 144 countries as shown in the below bar chart.

Country Vs Null Values



Sum of CountNull Value for each Country Name.

2. *Which data categories will be used for Principal Components Analysis (PCA)? There should be three categories that are targeted for PCA.*

   The below categories will be used for PCA.

   I. Education_Avg Years (30 variables), which is basically the same data albeit being shown separately for the different age groups.
   II. Education_Pct (15 variables), showing the percentage of people over 25 years of age that have completed education cycles, and
   III. Education_literacy (7 variables), containing information on young people's education with references to both genders.

3. *Which variables did you decide to be irrelevant for this analysis? Only variables under the education, economic, and environment categories should be included. Hint: There should be a total of nine variables removed from the dataset.*

   The below variables from the Background and Health Categories are irrelevant for the analysis and hence they will be removed.

Series Code	Category	Definition
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.

Step 3: Determine Clusters and Methodology

The number of clusters has been specified as 4 clusters, so we will apply K-Centroid cluster analysis to our dataset. Using the K-Centroids Diagnostics tool we will decide which Clustering methodology to follow, i.e. K-Means, K-Medians or Neural Gas. For all the clusters, the data was standardized with the z-score.

1. *What clustering method did you decide to use? Please justify your answer.*

   Generated Assessment reports for all the three methodologies as shown below.

<u>K-Means Cluster:</u>

K-Means Cluster Assessment Report

Summary Statistics

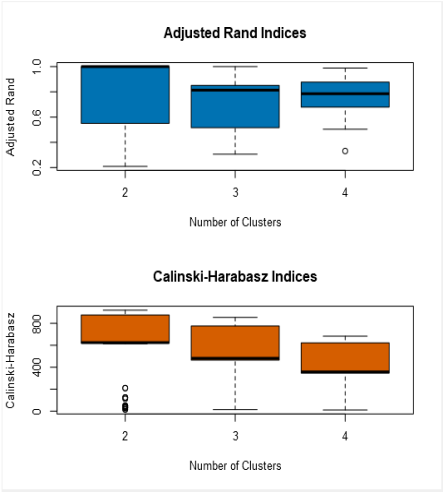
Adjusted Rand Indices:

	2	3	4
Minimum	0.2094	0.306	0.3313
1st Quartile	0.6104	0.5294	0.6798
Median	1	0.8132	0.7848
Mean	0.8174	0.7186	0.771
3rd Quartile	1	0.8508	0.8748
Maximum	1	1	0.988

Calinski-Harabasz Indices:

	2	3	4
Minimum	14.56	13.95	10.69
1st Quartile	619.7	465.1	346.8
Median	623.7	480.9	357.1
Mean	647.7	549.5	430.1
3rd Quartile	875	774.8	621.1
Maximum	919	853.6	682.7

Plots



K-Medians Cluster:

K-Medians Cluster Assessment Report

Summary Statistics

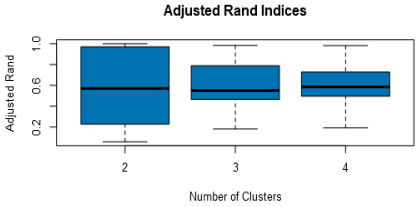
Adjusted Rand Indices:

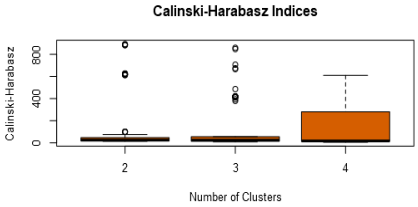
	2	3	4
Minimum	0.058	0.1811	0.1918
1st Quartile	0.2272	0.4655	0.4975
Median	0.5702	0.5506	0.5851
Mean	0.571	0.6232	0.6286
3rd Quartile	0.9701	0.7858	0.7237
Maximum	1	0.9836	0.9817

Calinski-Harabasz Indices:

	2	3	4
Minimum	13.98	7.774	5.188
1st Quartile	17.08	15.73	13.52
Median	24.47	24.04	16.49
Mean	131.7	115.7	126.1
3rd Quartile	46.87	54.76	280.3
Maximum	892.1	859.7	609.8

Plots





Neural Gas Cluster:

Neural Gas Cluster Assessment Report

Summary Statistics

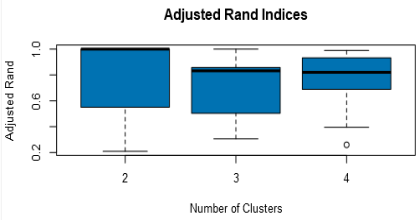
Adjusted Rand Indices:

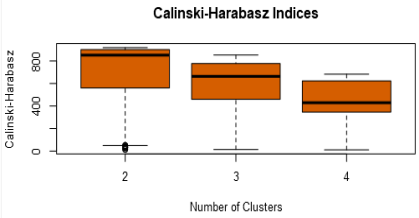
	2	3	4
Minimum	0.2094	0.306	0.2597
1st Quartile	0.6104	0.5067	0.6948
Median	1	0.8318	0.8202
Mean	0.809	0.7199	0.7836
3rd Quartile	1	0.8559	0.9284
Maximum	1	1	0.9897

Calinski-Harabasz Indices:

	2	3	4
Minimum	14.48	13.96	10.98
1st Quartile	560.3	461.4	346.4
Median	851.3	663.2	429.6
Mean	674	571.5	442.3
3rd Quartile	900.6	776.7	622.5
Maximum	918.2	852.5	682.8

Plots

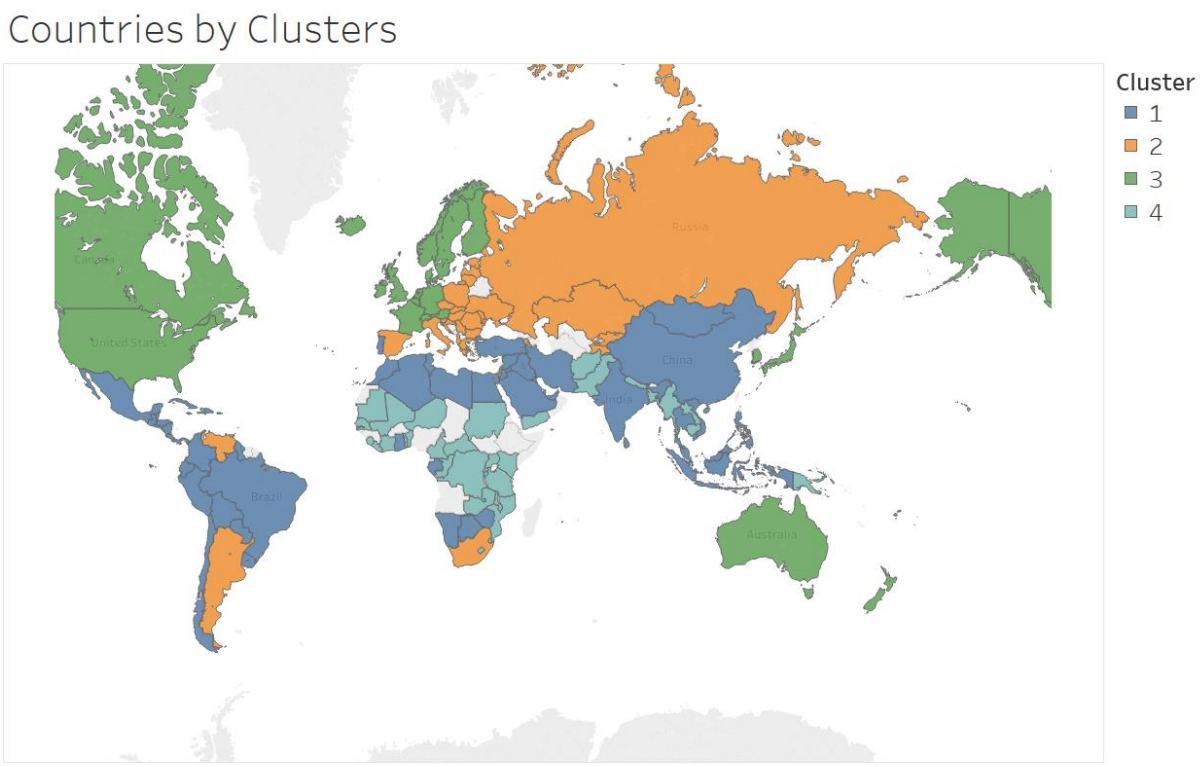




	Median of AR	Median of CH
K-Means	0.7848	357.1
K-Medians	0.5851	16.49
Neural Gas	0.8202	429.6

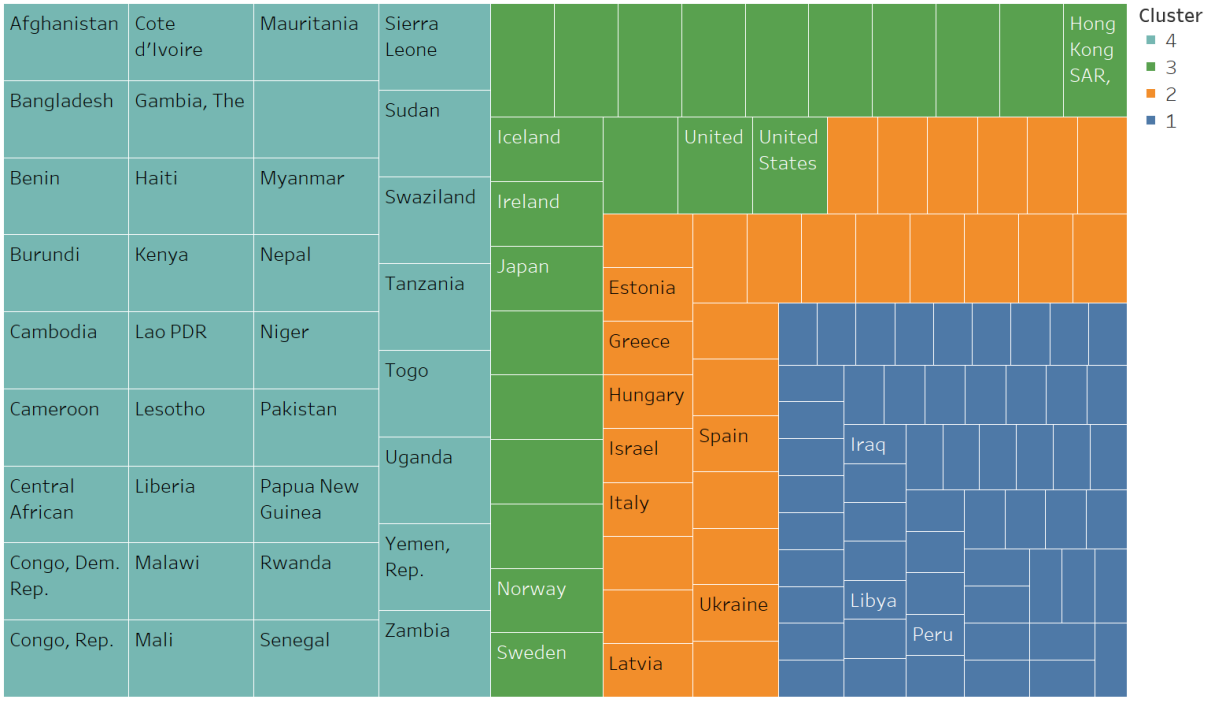
Based on the above reports I decided to use the Neural Gas method as it had the highest median and the most compact range, as can be seen in the Box plots, when compared to K-Mean and K-Median cluster methods.

Step 4: Run the Data and Visualize



Map based on Longitude (generated) and Latitude (generated). Color shows details about sum of Cluster. Details are shown for Country.Name.

Tree Map of Countries by Clusters



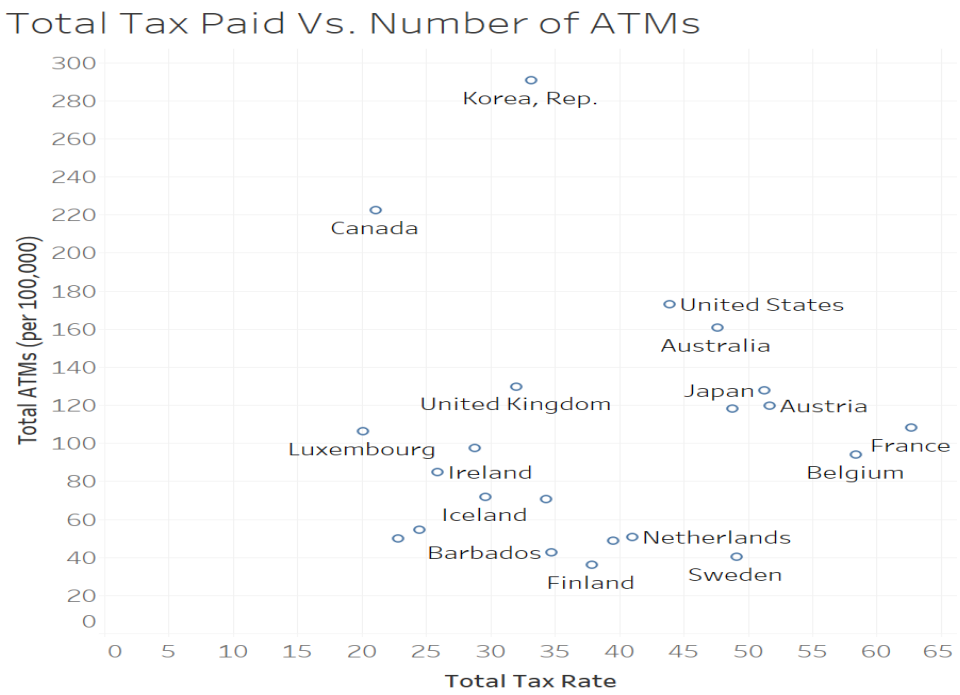
Country.Name. Color shows details about sum of Cluster. Size shows sum of Cluster. The marks are labeled by Country.Name.

1. Do the clusters make sense?

   The clusters appear to make perfect sense, especially Cluster number 3(green color in the tree map) where the US is included. We see it also includes other countries with strong economies and high literacy, i.e. Canada, Western Europe, Japan, South Korea, Australia and New Zealand. Cluster number 4 is about countries that are very poor, e.g. central African countries, Afghanistan, Pakistan,etc. Cluster 1 includes developing countries or developed countries, such as India. Observing all these, the clusters makes perfect sense.

2. What are the four countries in USA's cluster that are closest to the USA in terms of Total Tax Rate by ATM Machines? **Hint:** Create a scatterplot to graph the relationship between these two variables and color the markers by cluster.

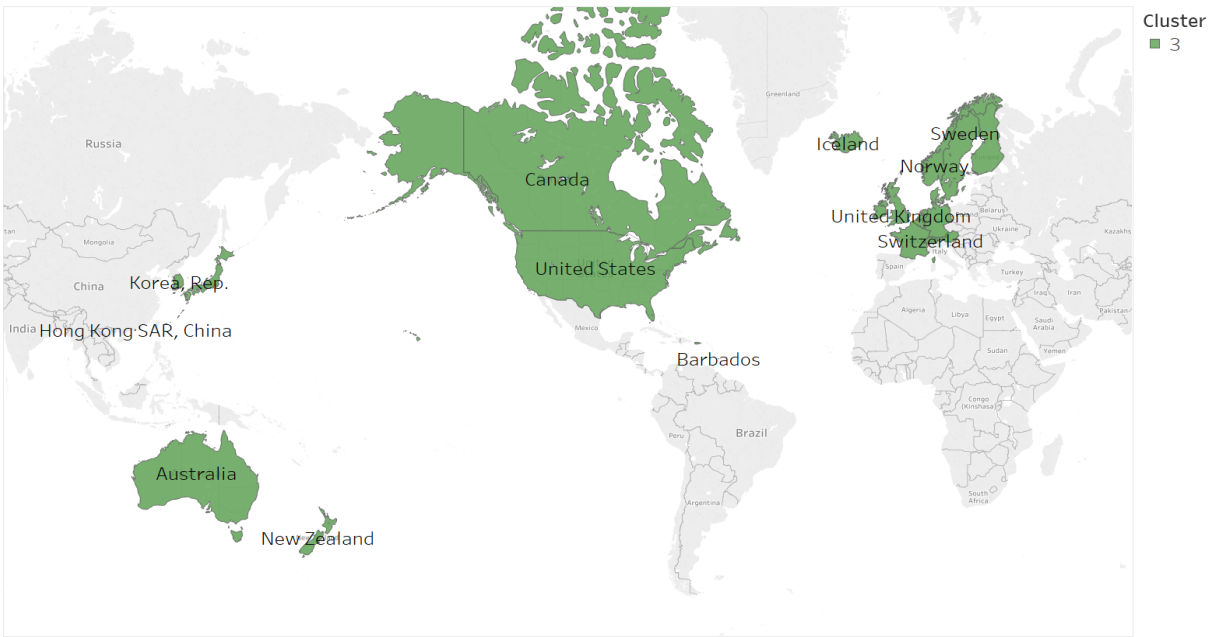
   Australia Great Britain(UK), Japan and Austria.



Ic Tax Totl Cp Zs vs. Fb Atm Totl P5. The marks are labeled by Country.Name. The data is filtered on Cluster, which ranges from 3 to 3.

Step 5: Recommendation

Please list out the country codes in this section here with this format in alphabetical order.



Map based on Longitude (generated) and Latitude (generated). Color shows details about sum of Cluster. The marks are labeled by Country.Name. Details are shown for Country.Name. The data is filtered on Cluster, which ranges from 3 to 3.

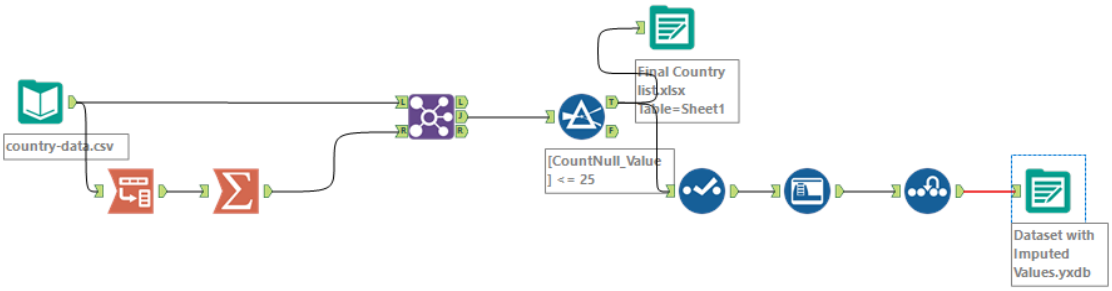
Record #	Country.Name
1	Australia
2	Austria
3	Barbados
4	Belgium
5	Canada
6	Denmark
7	Finland
8	France
9	Germany
10	Hong Kong SAR, China
11	Iceland
12	Ireland
13	Japan
14	Korea, Rep.
15	Luxembourg
16	Netherlands
17	New Zealand
18	Norway
19	Sweden
20	Switzerland
21	United Kingdom

1. *Why did you decide to choose these countries?*

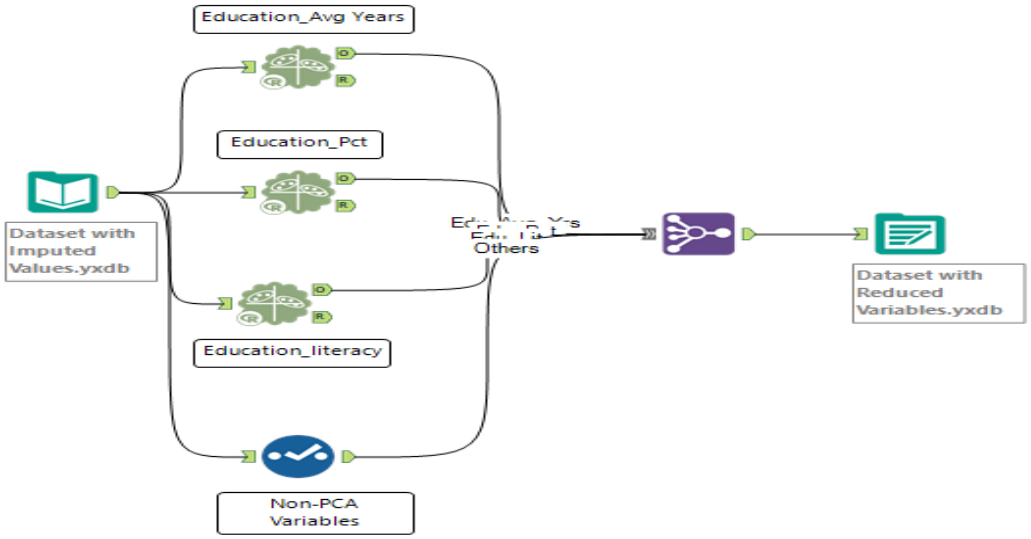
I chose these countries because they were the most similar to the United States based on economic and demographic data. The similarity was calculated Neural Gas Clustering. These countries fell within the same cluster as the United States.

Appendix:

Workflow 1: Data Cleaning



Workflow 2: Variable Reduction



Workflow 3: Comparing the Cluster models, Applying the Model and the final result