Segmentation Practice Project: International Expansion

Step 1: Key Decisions

Key Decisions:

Answer these three questions

1. What decisions needs to be made?

The retail store chain wants to expand its stores internationally and wants to find countries that are similar to the United States of America in terms of demographics, economics, education, and environment.

In order to help make this decision, we need to find out which countries are similar enough to the USA in terms of the above attributes.

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

We need to gather demographic, education, economic, and environmental information for every country on this planet.

Specific examples of each attribute we can gather include (but not limited to) are:

Demographic

1. Population in cities vs population in countryside
2. Age/Gender
3. Household income
4. Marital status
5. Average family size

Education

1. Education level
2. Number of public schools per capita
3. Number of universities/colleges per capita
4. Literacy rate
5. Education funding as percent of GDP

Economic

1. GDP
2. Inflation and interest rates
3. Local tax rates
4. Public transportation funding as percentage of GDP
5. Unemployment rate
6. Economic Stress Index
7. Consumer Sentiment Index
8. Manufacturing, Import, and Export growth rates

Environment

1. Percentage population with access to electricity
2. Percentage population with access to public transportation
3. Percentage population who are below country's poverty line
4. Pollution levels via Air Quality Index
5. Funding to paved roads and other transportation infrastructure as percentage of GDP
6. Crime and mortality rates
7. Drug and incarceration rates as percentage of population
8. Population density in cities and counties

Step 2: Explore and Cleanup the Data

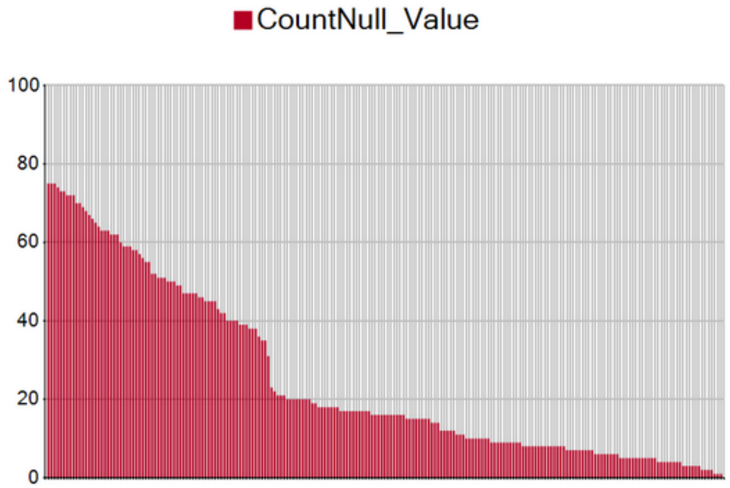
Explore and cleanup your dataset. Data is provided in a CSV file for 215 countries with 77 variables (250 word limit)

1. Here are some guidelines to help you cleanup your data:
   Country records where most of the variables missing might not be appropriate to be included in the analysis. The lack of accurate reporting could indicate that these countries are probably not similar to the United States. You should remove any country with 25 or more missing data points. HINT: You should be left with 144 countries.
2. Some variables are closely related and may be candidates for variable reduction through Principal Components Analysis.
3. Some variables seem irrelevant for the given analysis involving economy, demographics, education, and environment. Which variables seem irrelevant?

Answer these questions:

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

From my dataset, I reduced my list of potential countries from 215 to 144, removing all countries with greater than 25 missing variables.



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

The three topics that will be used for PCA are:

1. Average Years of Schooling
2. Pop > 25 with Degrees
3. Literacy Rate

because the data contains many sub variables that revolve around these three topics.

3. *Which variables did you decide to be irrelevant for this analysis? Hint: There should be a total of nine variables removed from the dataset.*

The nine variables I'm removing from the analysis are:

1. Internet users (per 100 people)
2. Prevalence of HIV, total (% of population ages 15-49)
3. Mortality rate, under-5 (per 1,000 live births)
4. Physicians (per 1,000 people)
5. Health expenditure per capita (current US$)
6. Prevalence of undernourishment (% of population)
7. Age dependency ratio (% of working-age population)

8. Women who believe a husband is justified in beating his wife when she burns the food
9. Prevalence of tuberculosis (per 100,000 population)

Because these variables do not fall under the three target categories (education, economic, environment) that we care about.

Step 3: Determine Clusters and Methodology

Determine the optimal clustering method and create four clusters. (100 word limit)

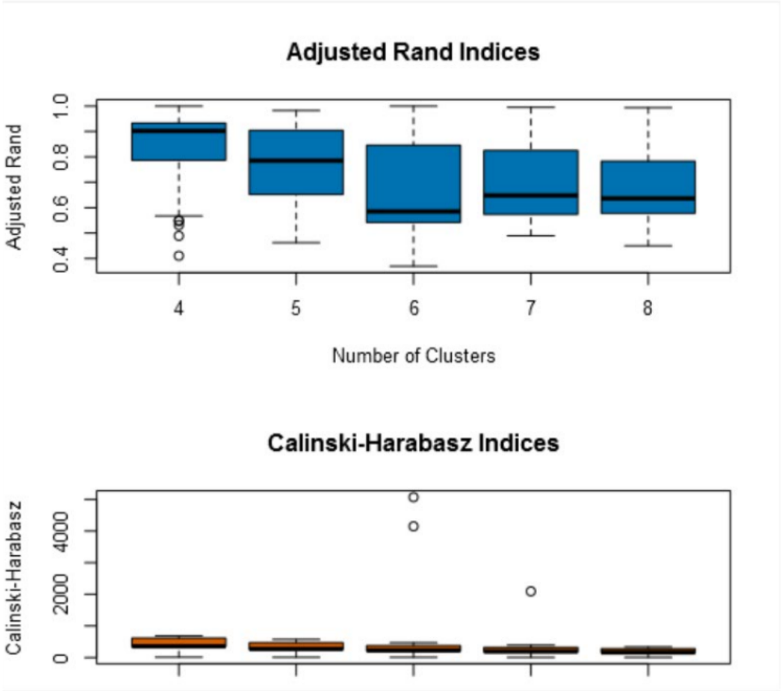
Answer this question:

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

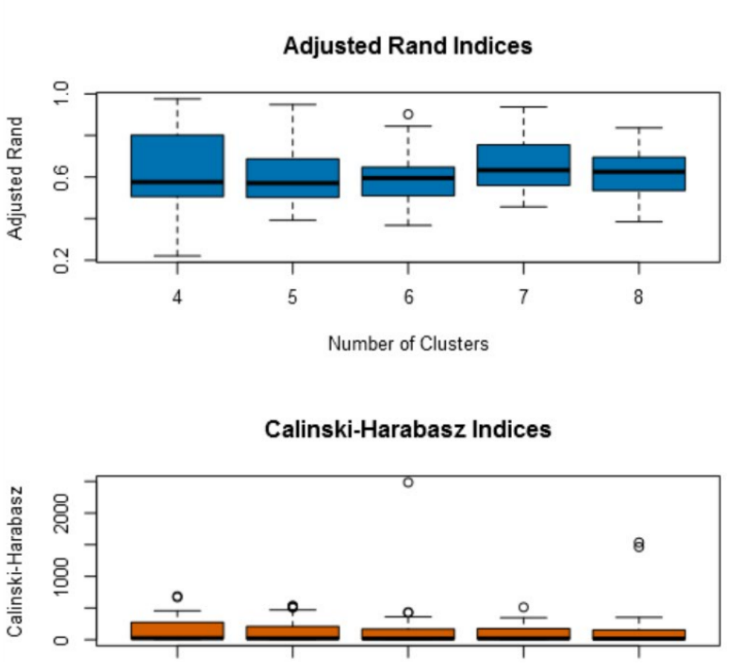
After running 3 clustering models (K-means, K-median, Neural Gas), I decided to use the Neural Gas clustering method.

Using the median and spread of the Rand and CH (Calinski-Harabasz) Indices. It’s clear that four clusters the most optimal method because the box-whisker plots in the Rand indices show how tight the indices for each data point are within each other.

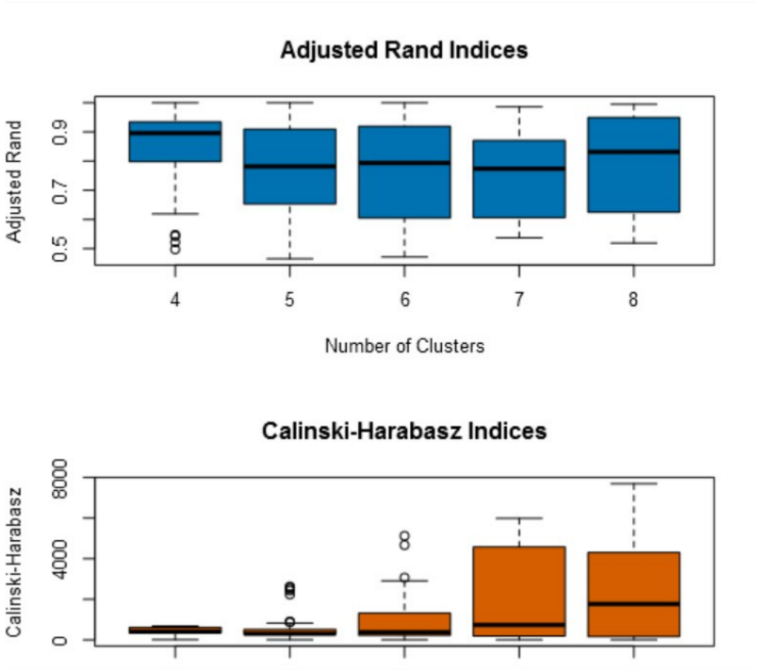
K-Means Diagnostic



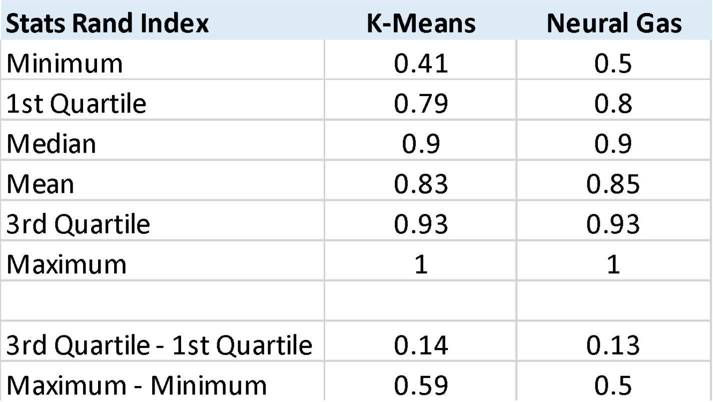
K-Median Diagnostic



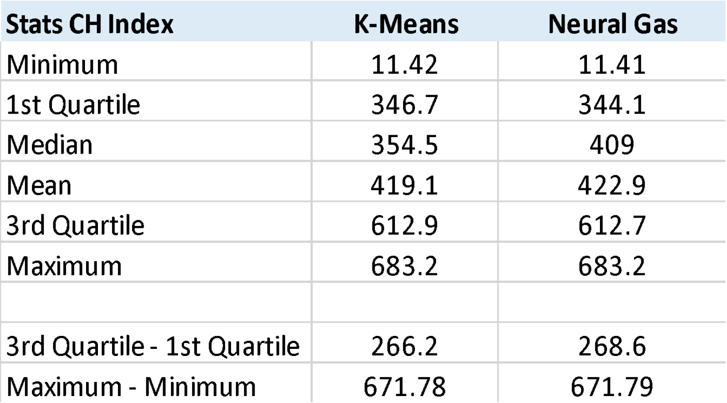
Neural Gas Diagnostic



We can see that the Neural Gas and K-Means for Cluster 4 are very close. I then calculated the differences between the various statistical properties between the Neural Gas and K-Means models:



Stats Rand Index	K-Means	Neural Gas
Minimum	0.41	0.5
1st Quartile	0.79	0.8
Median	0.9	0.9
Mean	0.83	0.85
3rd Quartile	0.93	0.93
Maximum	1	1
3rd Quartile - 1st Quartile	0.14	0.13
Maximum - Minimum	0.59	0.5



Stats CH Index	K-Means	Neural Gas
Minimum	11.42	11.41
1st Quartile	346.7	344.1
Median	354.5	409
Mean	419.1	422.9
3rd Quartile	612.9	612.7
Maximum	683.2	683.2
3rd Quartile - 1st Quartile	266.2	268.6
Maximum - Minimum	671.78	671.79

The two models perform very close in terms of spread and median and means. According to the Rand Index, the two models perform equivalently. According to the CH Index, the Neural Gas model slightly performs better with a higher median and mean.

Therefore I’ve chosen to use the Neural Gas model.

Step 4: Run the Data and Visualize

Run the data through your clustering algorithm and visualize the clusters. (250 words limit)

Include at least 2 visualizations to show the clusters that you came up with. At least one of you visualizations should be a Tableau map.

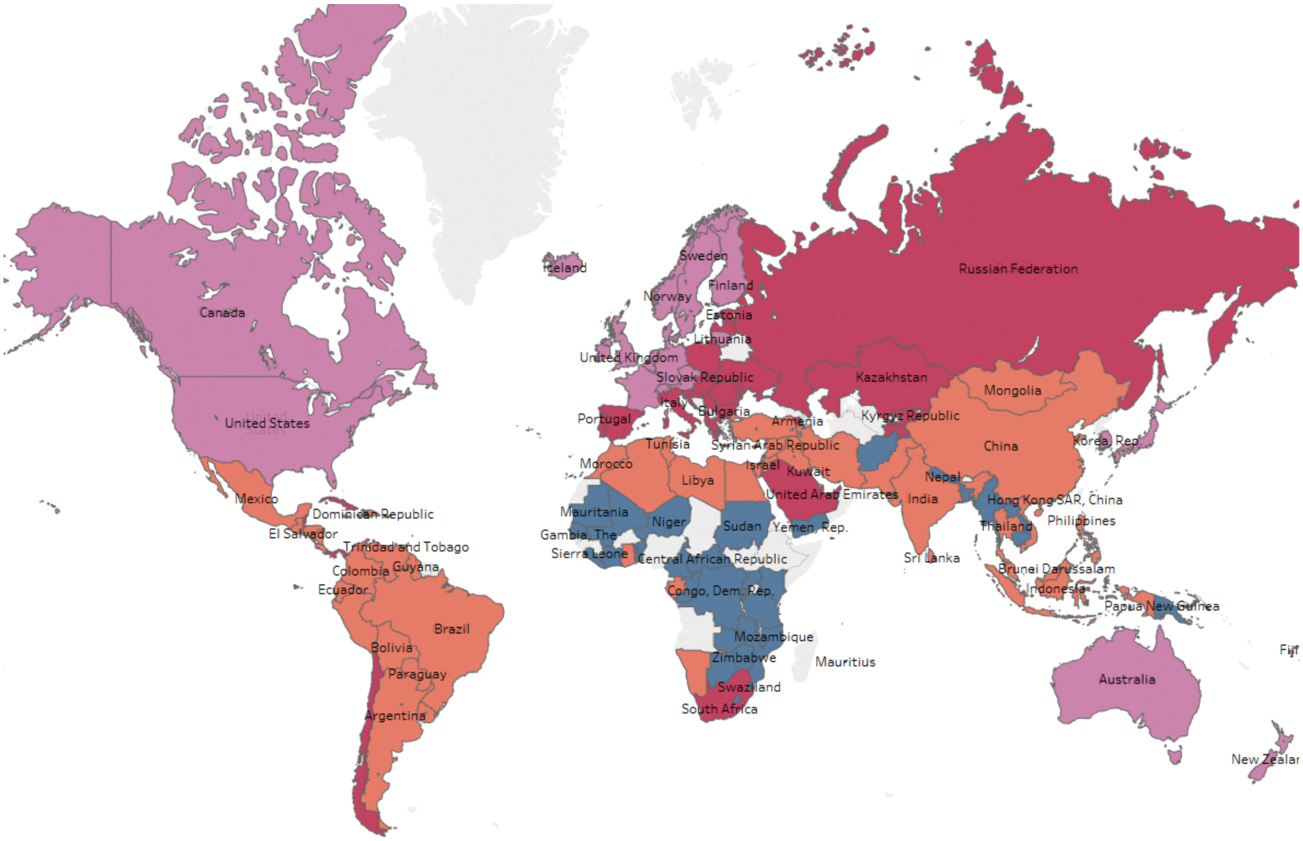
Answer these questions.

1. Do the clusters make sense?

The countries that belong to the USA cluster are first-world countries with strong economies, well-developed education and environmental systems, and contain populations with similar education and income levels.

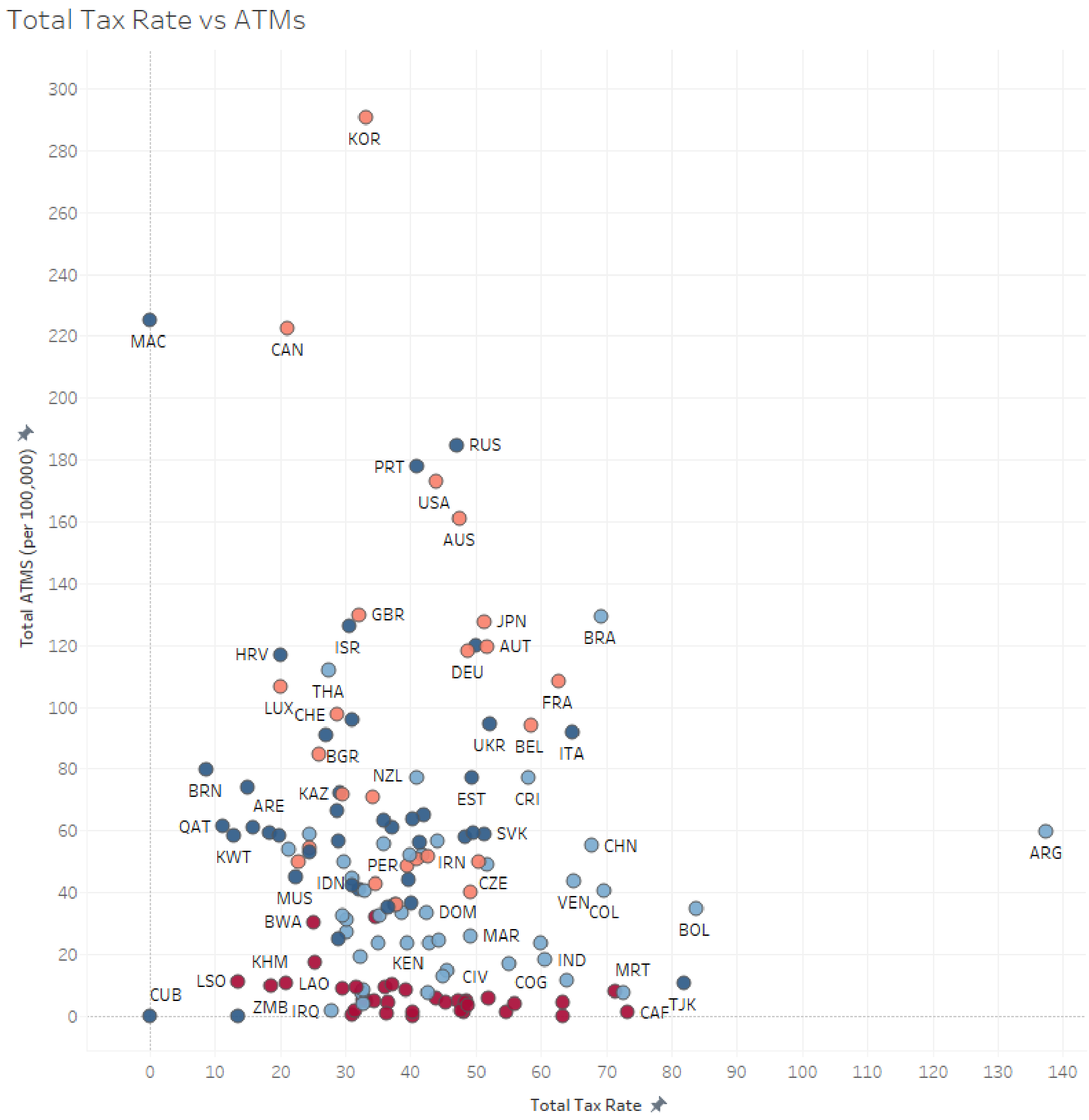
Examples of countries that were clustered into the USA cluster are:

Australia, Canada, Germany, and Great Britain, Japan, and South Korea.



The countries that are colored pink/purple are the countries that fit in the same cluster as the USA.

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.



Graphing the two relationships, we can see that Australia, United Kingdom, Japan, and Canada are the closest four countries with respect to Tax Rate and Total ATMS.

Step 5: Recommendation

Provide your recommended list of countries and justify your recommendation using data from your analysis (250 words limit)

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

………..

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

………..

Answer this question:

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

According to my clustering model, these countries match closest to the USA in terms of economic, educational, environmental, and demographic attributes.

Furthermore, these countries make sense given the economic power, government structure, and first-world development all of these countries possess.