**Project Details**

SEND FEEDBACK

**Project Overview**

You work for a retail store chain in the United States of America. The company is thinking of expanding to other countries and want to figure out which countries are similar economically, demographically, education, and environment to the United States of America.

Your manager has asked you to segment the countries of the world based on various economic, demographic, education, and environment data. From this you should be able to provide a list of countries that are similar to the United States. This will be the “short list” for further consideration by management.

You have the following information to work with:

A set of various data taken from the [World Bank web site](http://databank.worldbank.org/data/databases.aspx?qterm=&pagenumber=1). A clean version of this data is included with the supporting materials.

**Steps to Success**

**Step 1. Key Decisions**

Determine the key business questions and specific types of data you need to look at in order to complete this analysis.

**Step 2. Explore and Cleanup the Data**

Clean up the data in the provided CSV file. The data contains 215 countries with 77 variables.

Not all of the countries have complete data and not all of the variables may be applicable for the analysis. Based on the request from your manager, you will need to:

* Remove countries with a high number of missing data points (>25). HINT: There should be 144 countries left.
* Remove variables that are not relevant for the analysis (any variable not under the economic, education, demographic, or environment categories). HINT: You should remove 9 variables.

**Step 3. Determine Clusters and Methodology**

Your manager would like to see four clusters in the results. You also need to determine which clustering method provides the optimal clusters.

**Step 4. Run the Data and Validate**

After you run the model with the data you’ve decided to use, append the Cluster IDs back to the data in order to visualize the data.

Feel free to use any visualization software you want, such as Tableau.

**Step 5. Recommendation**

Provide your recommendations and justify why your countries should considered to be explored first for international expansion.

NEXT

# ; Practice Project Solution

The following resources can help you understand one possible solution to this project.

* [Project Answer Key](https://docs.google.com/document/d/1qYIoX2qyMdWyIglJmkRssAL2i_dwByLgOq2QuQc1ElE/edit): This is a full write up of the answer, including explanations of the data cleaning process,
* Solution Alteryx Workflow (see link below): This shows how all the analysis that led to the conclusions in the Project Answer Key.
* Solution Tableau file (in progress): This contains the visualizations used in the Project Answer Key.

**Supporting Materials**

* [Practice Project Solution Workflow](https://video.udacity-data.com/topher/2017/May/590a5bc4_practiceproject7/practiceproject7.zip)
* [Practice Project Solution Workflow 10.6](https://video.udacity-data.com/topher/2017/July/595ac51c_practiceproject7106/practiceproject7106.yxmd)

**Supporting Materials**

SEND FEEDBACK

**Rubric**

This project used to be a graded project. Use the [project rubric](https://review.udacity.com/#!/rubrics/424/view) to review your project. Once you're happy with your work, take a look at the solution materials provided. If you see room for improvement, keep working to improve your project.

**Completion Template**

The completion template linked below will help guide your work on the practice project. Use it to complete your project.

**Data**

*country\_data.csv* - This file contains country information

*variables\_catalog.csv* - This file contains descriptions for the data fields in *country\_data.csv*

**To Download Files below please right click on the link and select "Save Link As"**

**Supporting Materials**

* [Country Data](https://video.udacity-data.com/topher/2016/October/57f40d44_country-data/country-data.csv)
* [Variables-Catalog](https://video.udacity-data.com/topher/2016/December/5849817d_variables-catalog/variables-catalog.csv)

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.

1. 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

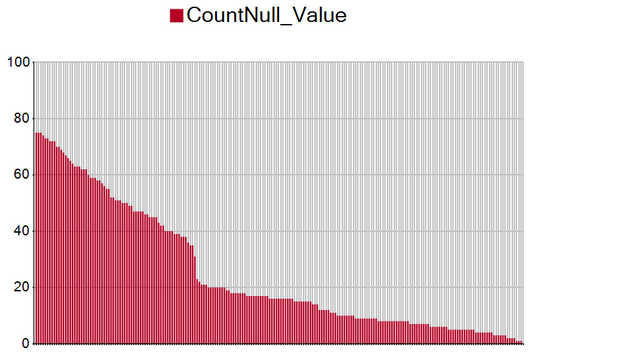
*Explore and cleanup your dataset. Data is provided in a CSV file for 215 countries with 77 variables (250 word limit)*

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.



1. *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.

1. *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)*

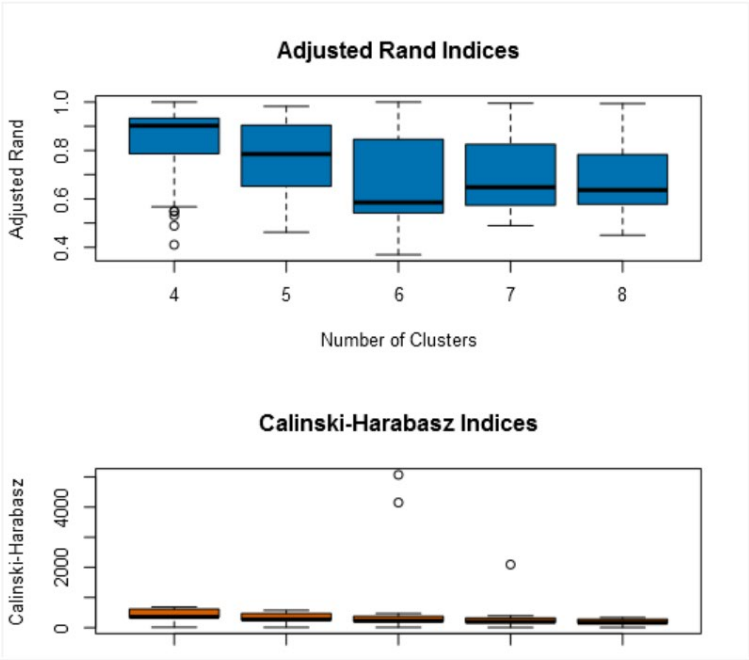
*Answer this question:*

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

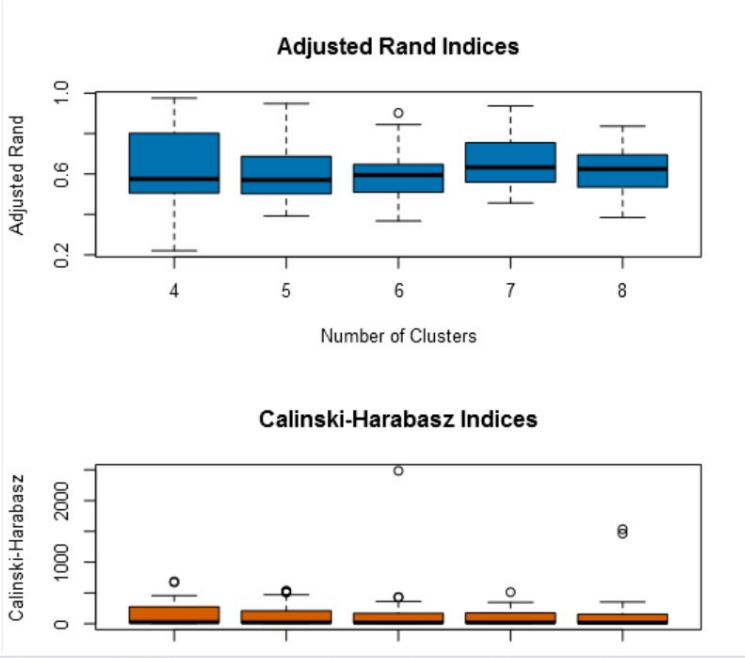
After running 3 clustering models (K-means, K-median, Neural Gas), I decided to use the Neural Gas clustering method.

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.

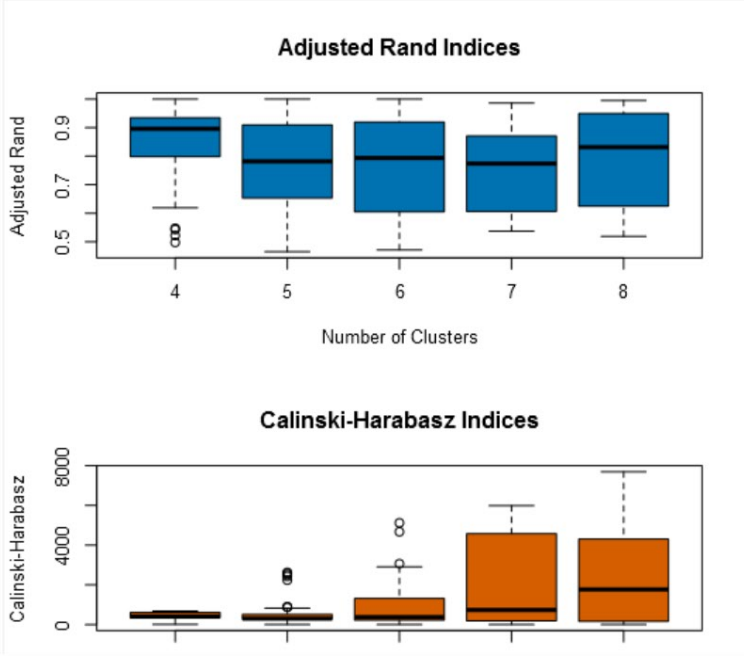
**K-Means Diagnostic**



**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:





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.*

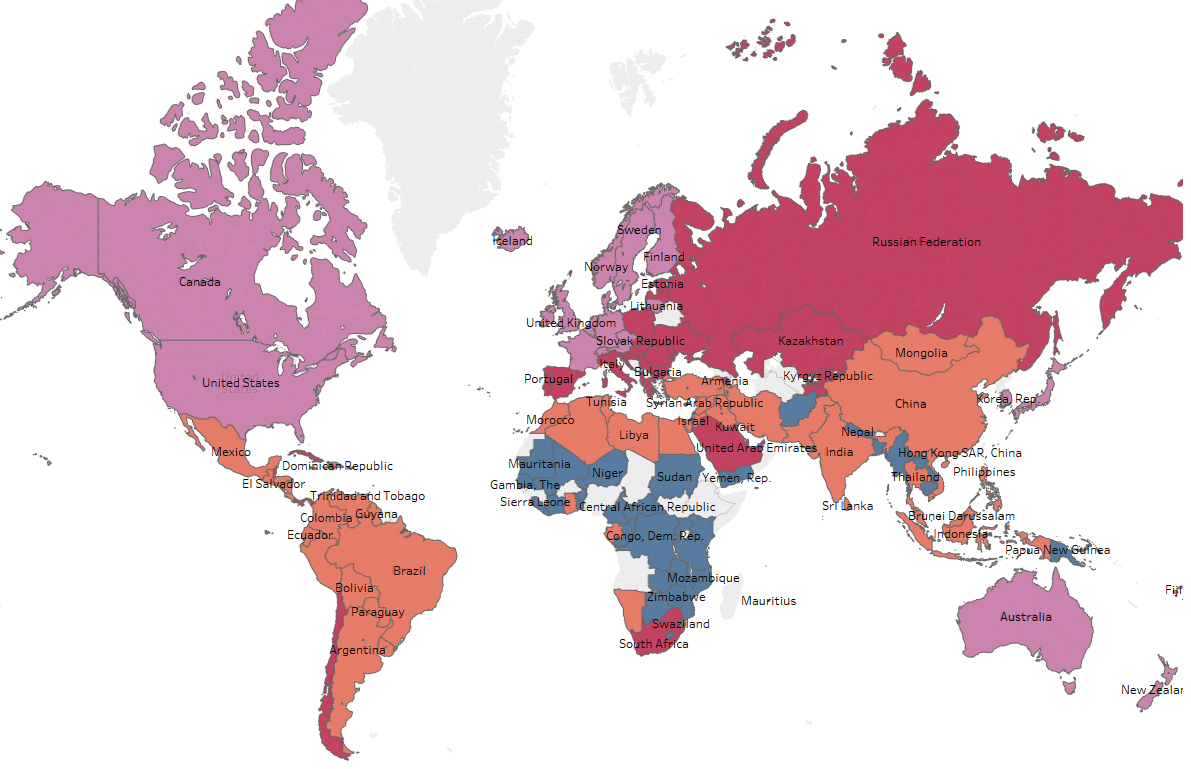
*Answer these questions.*

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.

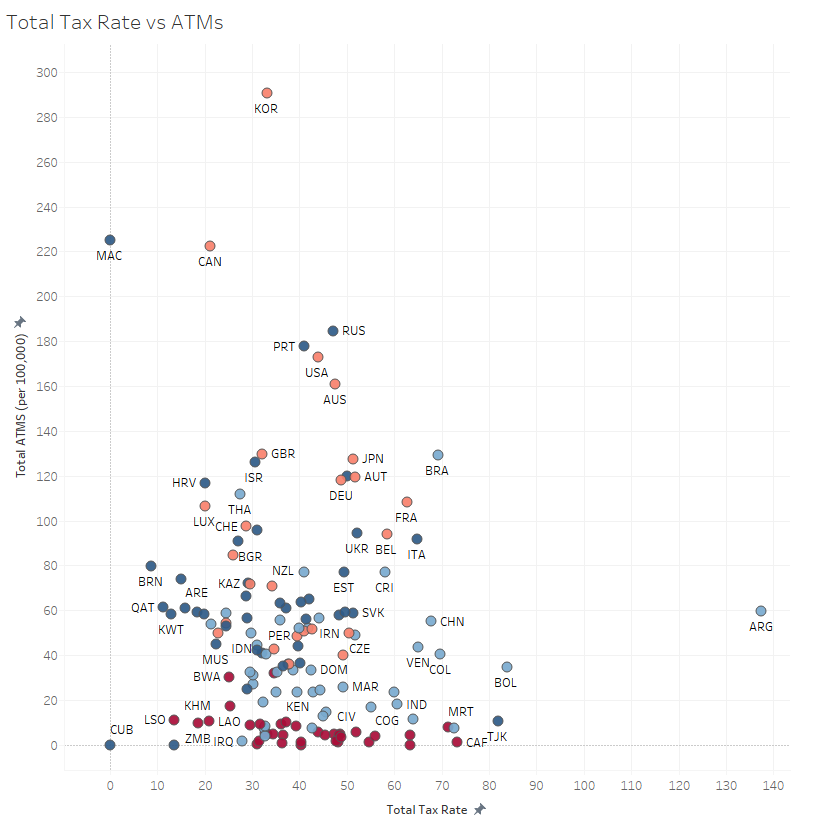
Examples of countries that were clustered into the USA cluster are:

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.*

1. 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.*