Project: International Expansion

Complete each section. When you are ready, save your file as a PDF document and submit it here: [https://classroom.udacity.com/nanodegrees/nd008/parts/91294931-aacb-4887-856f-fd19fe915795/project#](https://classroom.udacity.com/nanodegrees/nd008/parts/91294931-aacb-4887-856f-fd19fe915795/project)

## Step 1: Key Decisions

*Briefly explain the key decisions and the type of data that you need to conduct this analysis (250 word limit).*

### Key Decisions:

*Answer these three questions*

1. What decisions needs to be made?

The location new store of our retail store has to be decided in a country which is economically and demographically similar to the United states of America.

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

**Economic:** ATM’s per 1000 adults, Time required to get electricity (days)

**Environment:** Access to electricity (% of population), Population living in slums, (% of urban population)

**Education:** Literacy rate, adult female (% of females ages 15 and above), Youth literacy rate, population 15-24 years, both sexes (%)

## 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)*

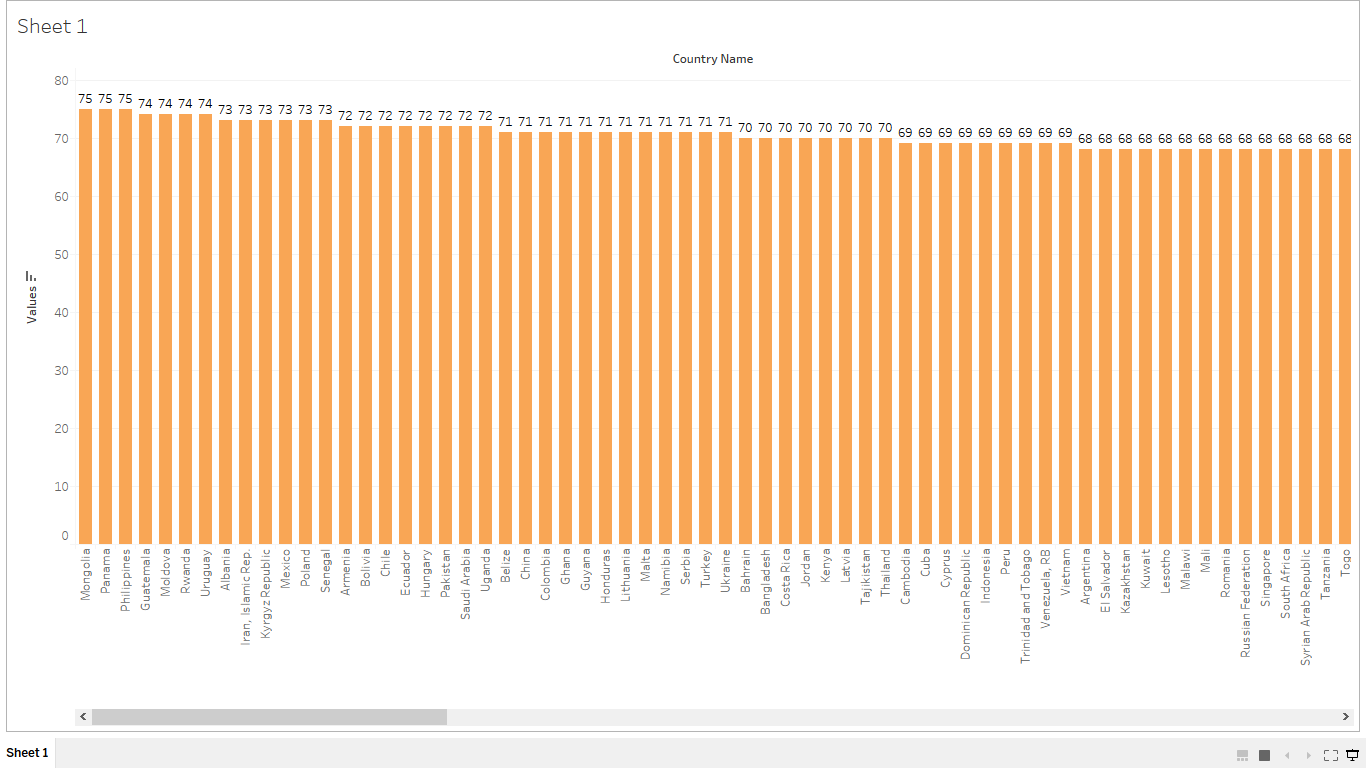
Here are some guidelines to help you cleanup your data:

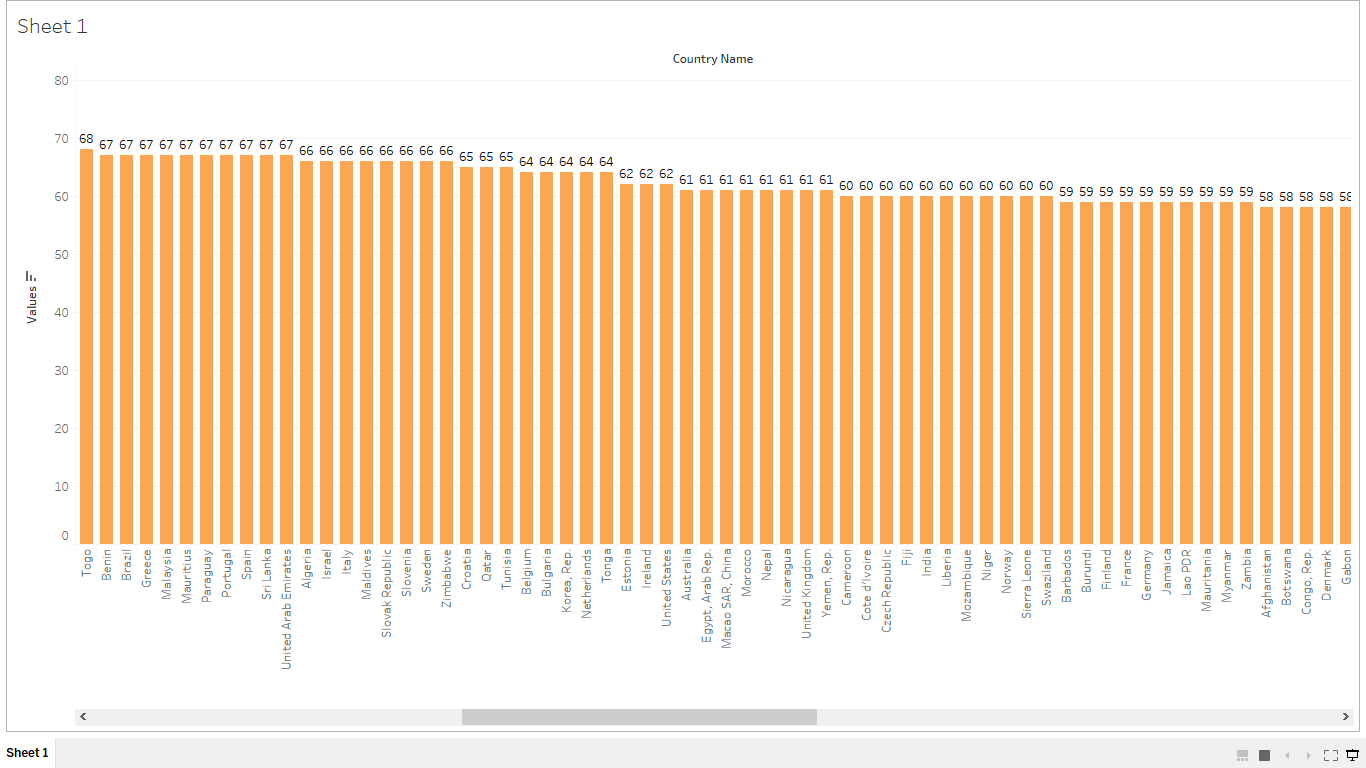
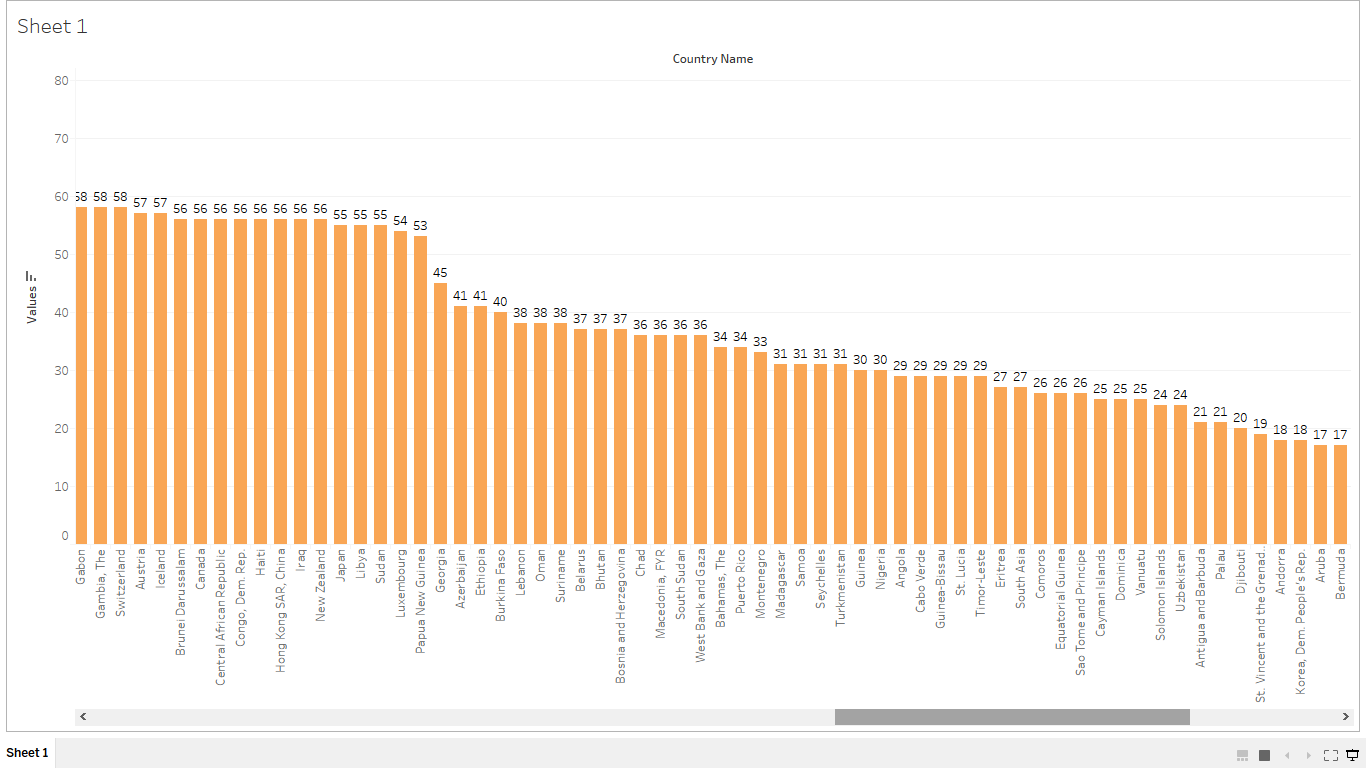
1. Country records where most of the variables missing might not be appropriate to be included in the analysis. The lack of accurate reporting could indicate that these countries are probably not similar to the United States. You should remove any country with fewer than 25 missing data points. HINT: 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?

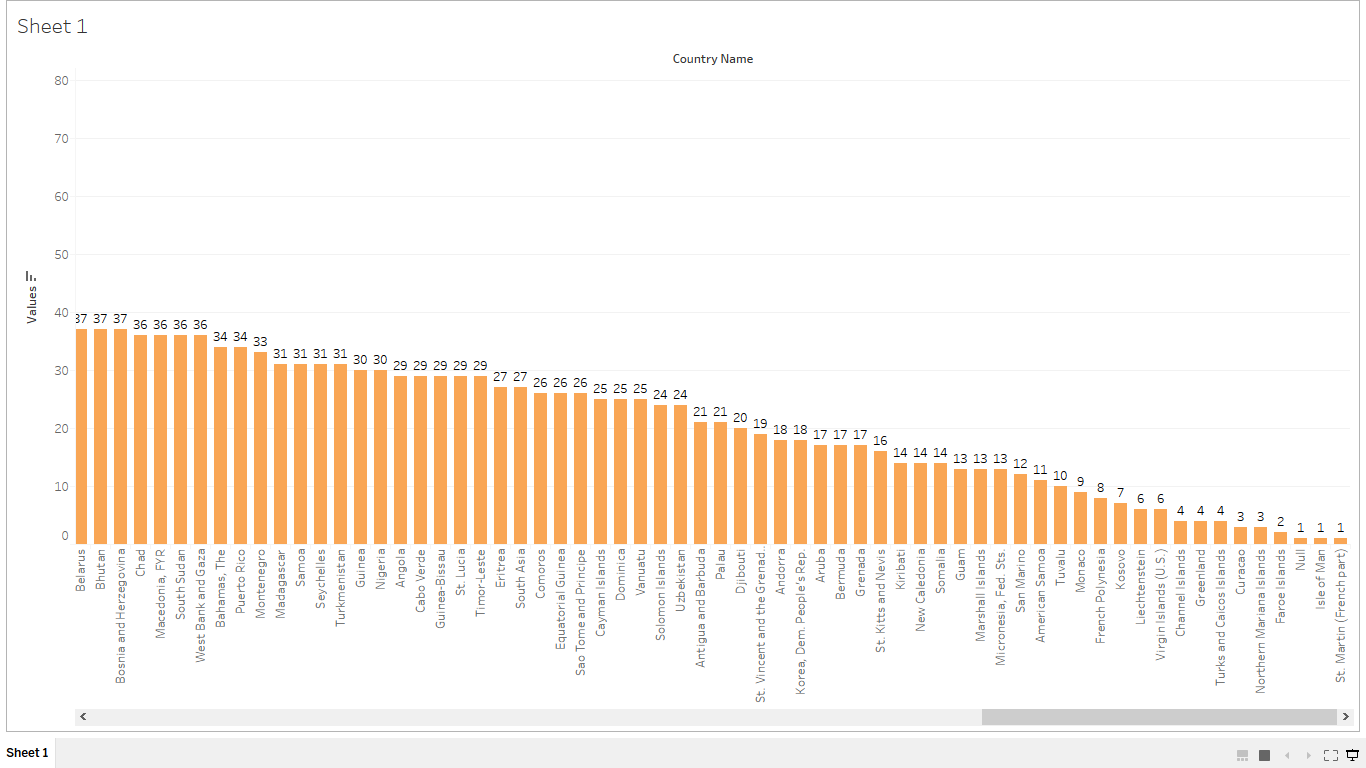
*Answer these questions:*

1. *How many countries did you reduce your dataset to? Please include a bar chart of number of non-null data points by country, sorted from most to least.*
2. *Which data categories will be used for Principal Components Analysis (PCA)? There should be three categories that are targeted for PCA.*
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 data has been reduced to 144 countries.

Bar Chart





Answer 2: Literacy rate Average years of schooling and Education levels of percentage of population (Aged 25plus) should be the 3 broad categories used for PCA Analysis

Answer 3:

I’ve removed

Internationally-recognized quality certification (% of firms) – Logical decision

Internet users (per 100 people) - Classified as infrastructure indicator

Women who believe a husband is justified in beating his wife when she burns the food (%)-Gender equality indicator

Prevalence of HIV, total (% of population ages 15-49)- Health indicator

Mortality rate, under-5 (per 1,000 live births)- Health indicator

Physicians (per 1,000 people)-Health indicator

Tuberculosis prevalence rate (per 10,00,000 population)-Health indicator

Health expenditure per capita-Health indicator

Prevalence of undernourishment (% of population)-Health indicator

Step 3: Determine Clusters and Methodology

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

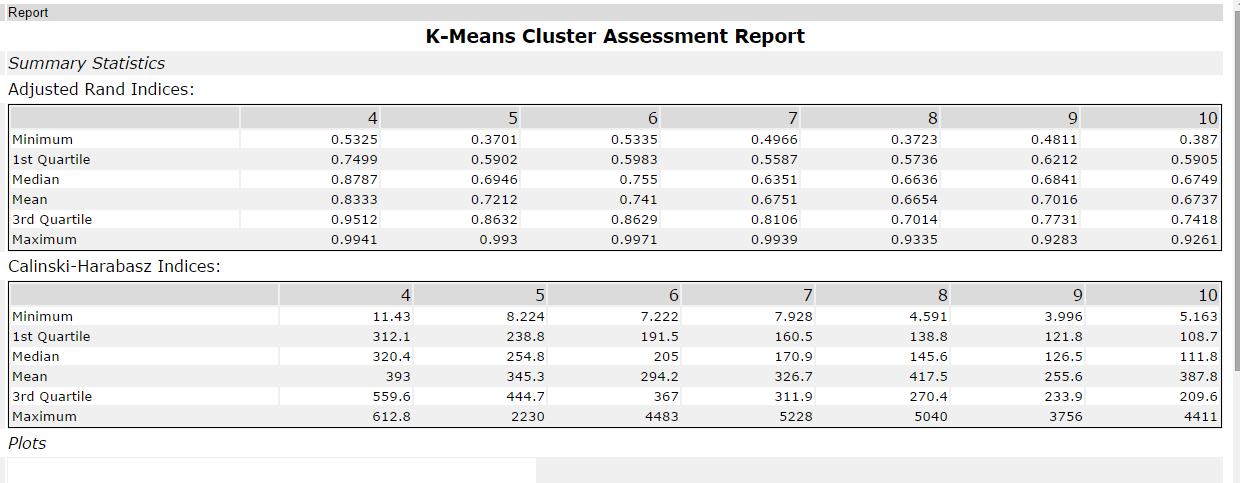
*Answer this question:*

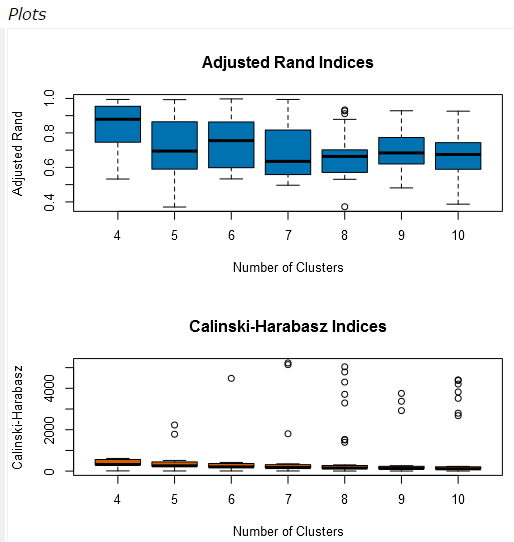
1. *What clustering method did you decide to use? Please justify your answer.*

*I decided that the Neural Gas method was the best as it had the highest median and the most compact range when compared to K-median and K mean methods*

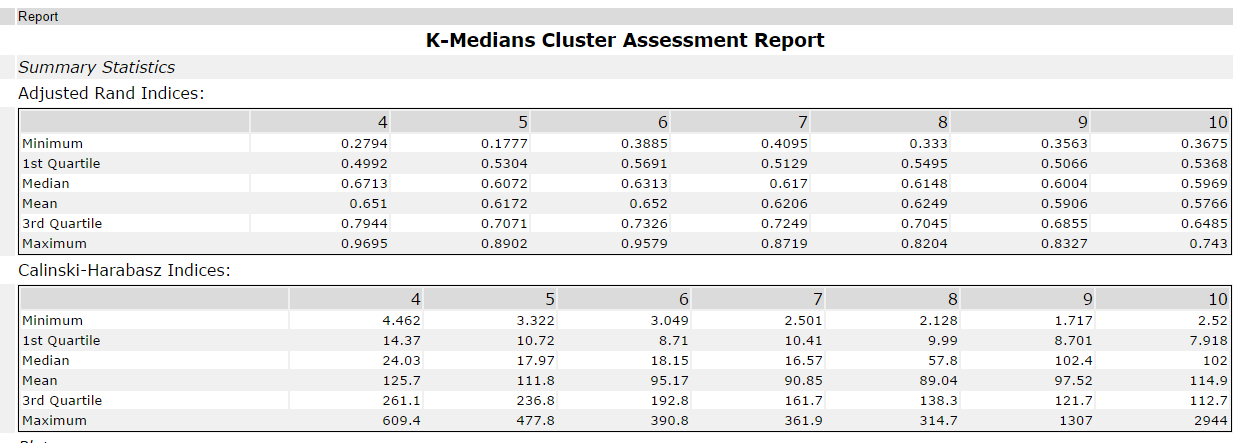
*Please find the respective reports and graphs*

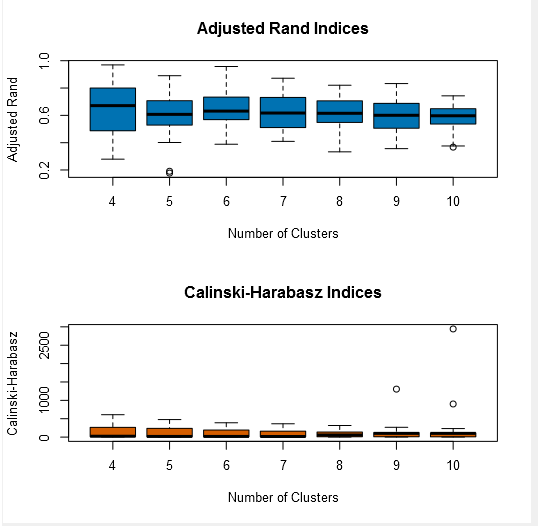
*K-means*

**

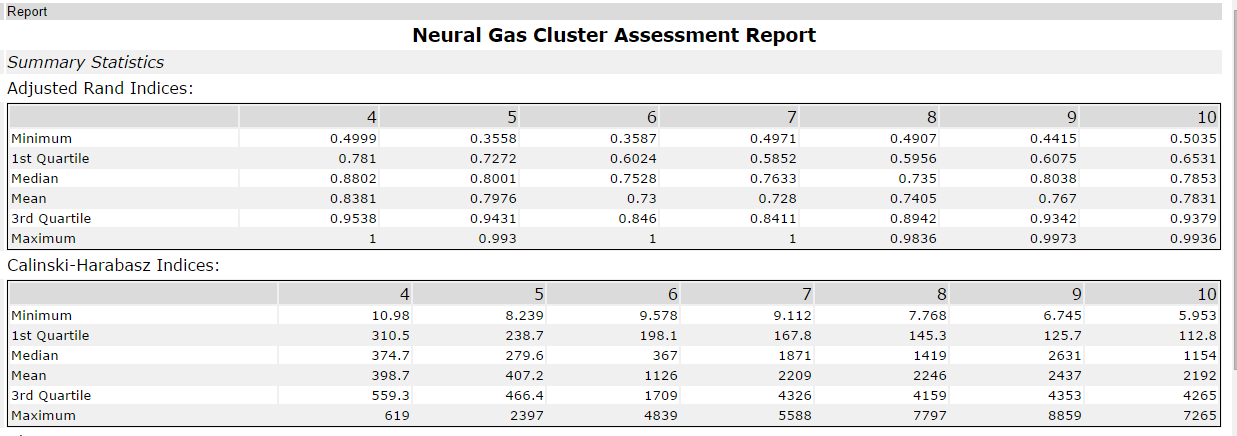


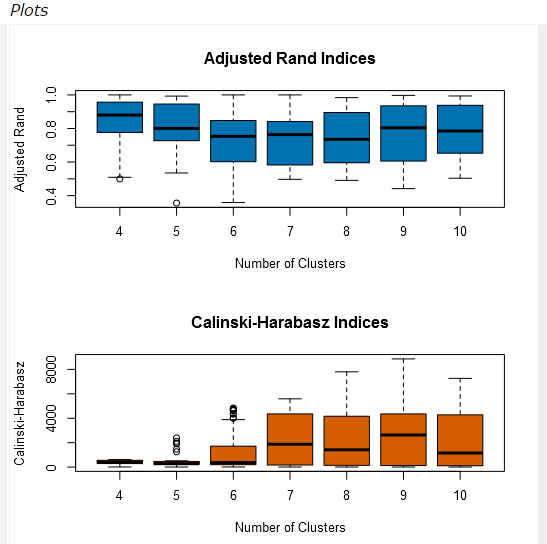
K-median





Neural Gas





## Step 4: Run the Data and Visualize

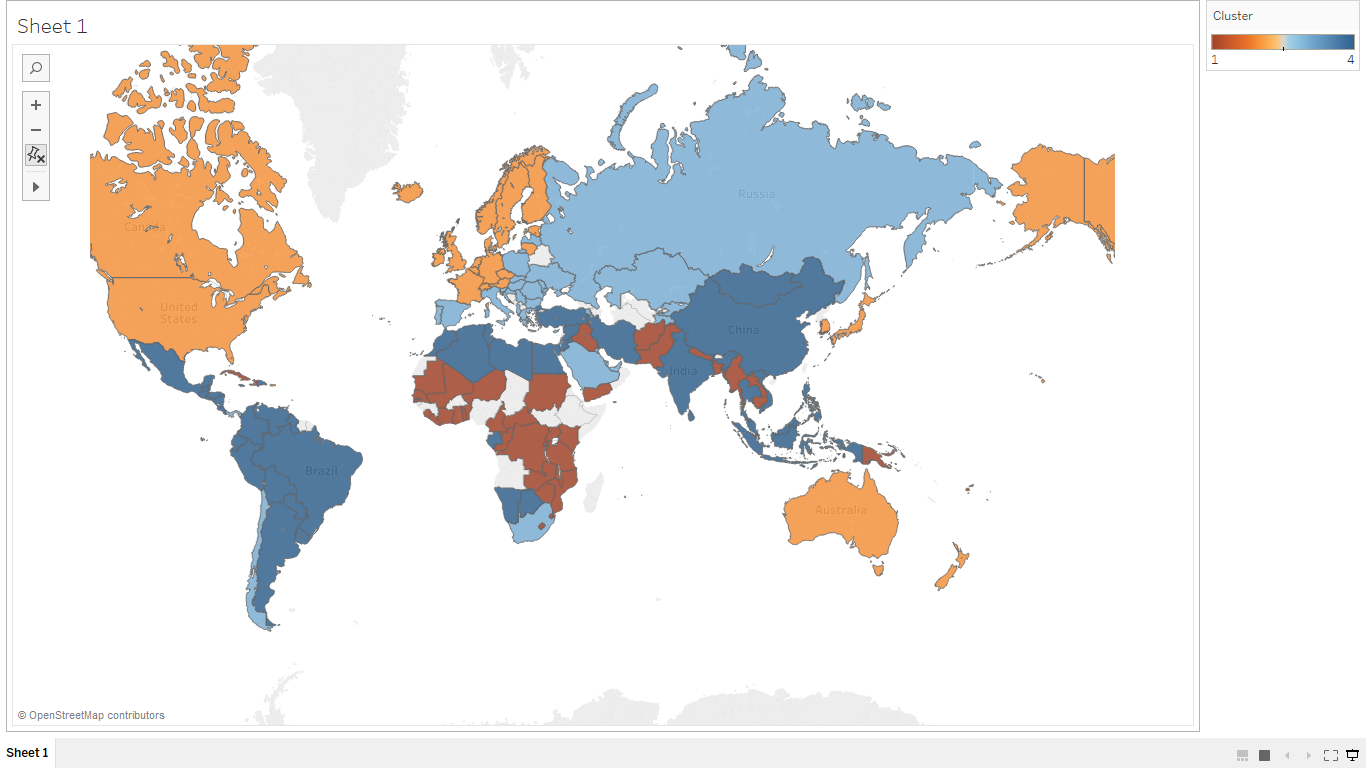
*Run the data through your clustering algorithm and visualize the clusters. (250 words limit)*

*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.*

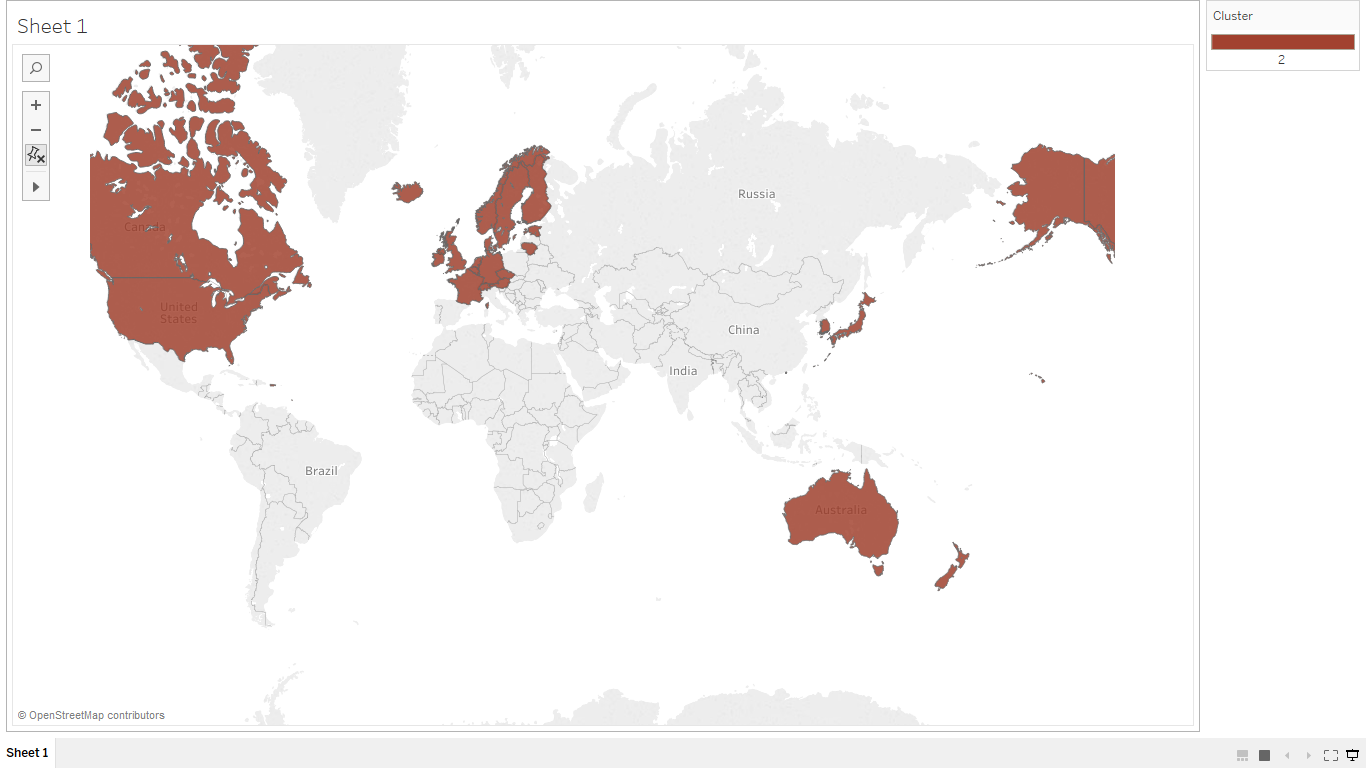
*Answer this question.*

1. Do the clusters make sense?

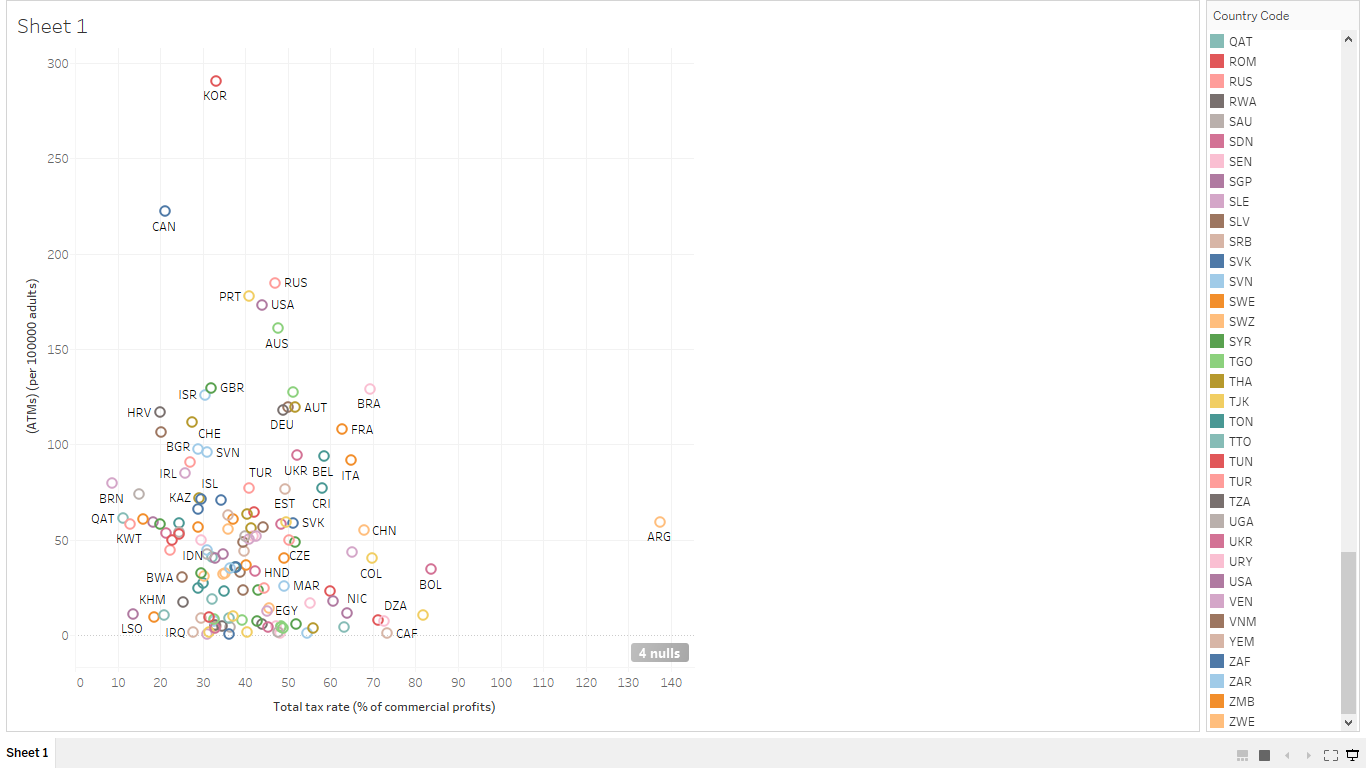
Yes, the clusters do make sense, as we can observe all the developed countries are grouped in cluster colored orange and the poorest third world nations are grouped in one cluster (Maroon color) and so on.

All 4 clusters

Cluster with countries similar to United States



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.



From the above image, the countries closest to USA are :

Australia (AUS), Russian Federation ( RUS) , Portugal (PRT) , United Kingdom (GBR)

## 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 codes in this section here with this format in alphabetical order.*

*………..*

*Australia*

*Belgium*

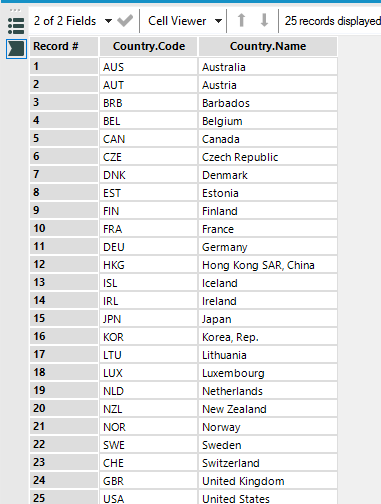
*Canada*

*………..*

*Answer this question:*

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

*(PICTURE OF CLUSTER WITH COUNTRY NAME AND UNITED STATES INCLUDED IN LIST)*

**

I chose these countries as, according to the clustering analysis I’ve carried out, they are economically and demographically closest to United States.

Before you Submit

Please check your answers against the requirements of the project dictated by the [rubric](https://review.udacity.com/#!/rubrics/424/view) here. Reviewers will use this rubric to grade your project.