Project: Segmenting the Countries of the World

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 retail business is trying to expand its foot print across the globe as an analyst for the company, I need find out the countries which have similar characteristics as of the United states in terms of economy, education which would be suitable for the expansion of the organization overseas.

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

   Since we need to find the country which is most suitable for expansion of the retail business we need the data regarding the economic state of the country such as gross domestic product, employment to population ratio and the tax rates.

   Economic: Govt. education expenditure as a part of GDP, Tax rate
   Environment: Access to electricity, population living in slum
   Education: No of people with formal education, percentage of literacy.

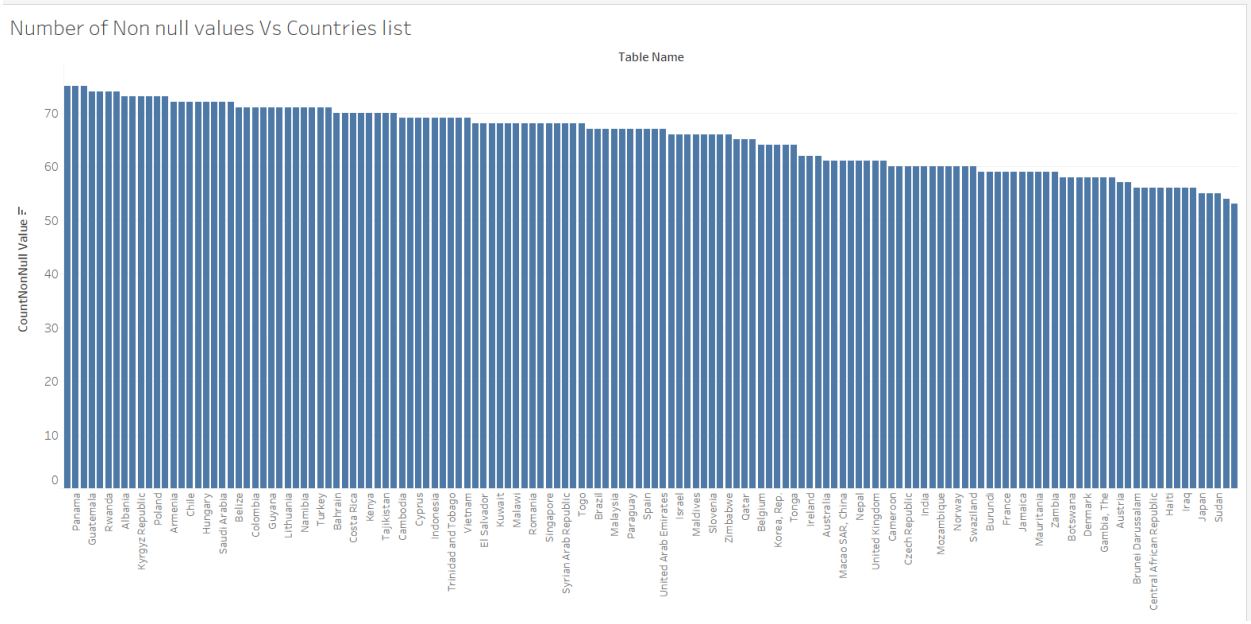
Step 2: Explore and Cleanup the Data

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?

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 total number of countries in our data set is 144, we had to remove the countries which had more than 25 null values, since having a lot of missing data can add bias into the analysis.



Link to Online tableau repository for a clear view of the data

https://public.tableau.com/profile/rakshit.kumar.satish#!/vizhome/Project_Analaysis/Sheet1

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

The three categories that are targeted for the PCA are

- Education_Avg Years
- Education_Pct
- Education_literacy

Education Avg Years has 30 variables and has data about the average education among the different age groups.
Education Pct. Has 15 variables and has data about the percentage of people with some sort of school education.
Education literacy category has 7 variables giving data for different age groups and sex.

All the variables in the above three categories can be reduced to be used efficiently.

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 categories and variables that are irrelevant for the current analysis are as below

IT_NET_USER_P2	Background
SH_DYN_AIDS_ZS	Background
SH_DYN_MORT	Background
SH_MED_PHYS_ZS	Health
SH_XPD_PCAP	Health
SN_ITK_DEFC_ZS	Health
SP_POP_DPND	Health
SG_VAW_BURN_ZS	Health
SH_TBS_PREV	Health

These variables do not have any value in the analysis of the retail store setup since all these are related to the health and the background of the people of that country which provides little to no information for the retail business.

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.***

The number of clusters has been specified as 4 and we will apply K-Centroid cluster analysis tool to our dataset.
Using the K-Centroid diagnostics tool we have to make the decision on the methodology to be used.
I chose Neural gas analysis for the clustering method because of the median values of the AR and CH index are high for the 4 clusters.

Report

Neural Gas Cluster Assessment Report

Summary Statistics

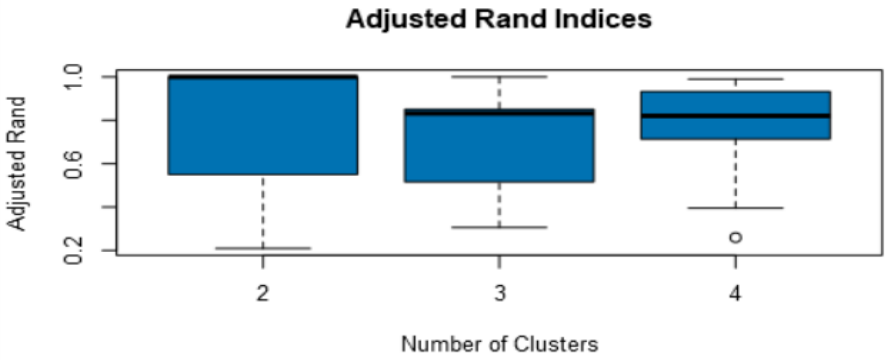
Adjusted Rand Indices:

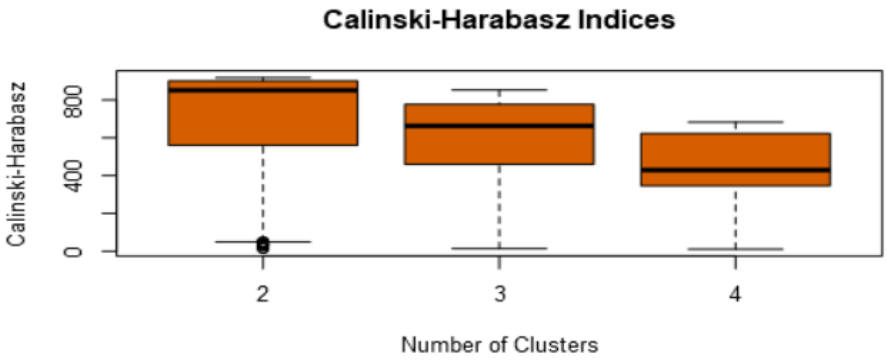
	2	3	4
Minimum	0.2094	0.306	0.2597
1st Quartile	0.6104	0.5254	0.7172
Median	1	0.8318	0.8202
Mean	0.809	0.7206	0.7869
3rd Quartile	1	0.8508	0.9284
Maximum	1	1	0.9897

Calinski-Harabasz Indices:

	2	3	4
Minimum	14.48	13.96	10.98
1st Quartile	560.4	461.4	346.4
Median	851.2	663.2	429.5
Mean	674.1	571.6	442.4
3rd Quartile	900.7	776.7	622.5
Maximum	918.2	852.6	682.9

Plots





After analyzing the different clusters, it was found that the cluster analysis is correct since certain countries like the Ghana, Pakistan, Rwanda have lowest electricity penetration, high urban slum population and less education, which makes them perfect fit in that particular cluster.

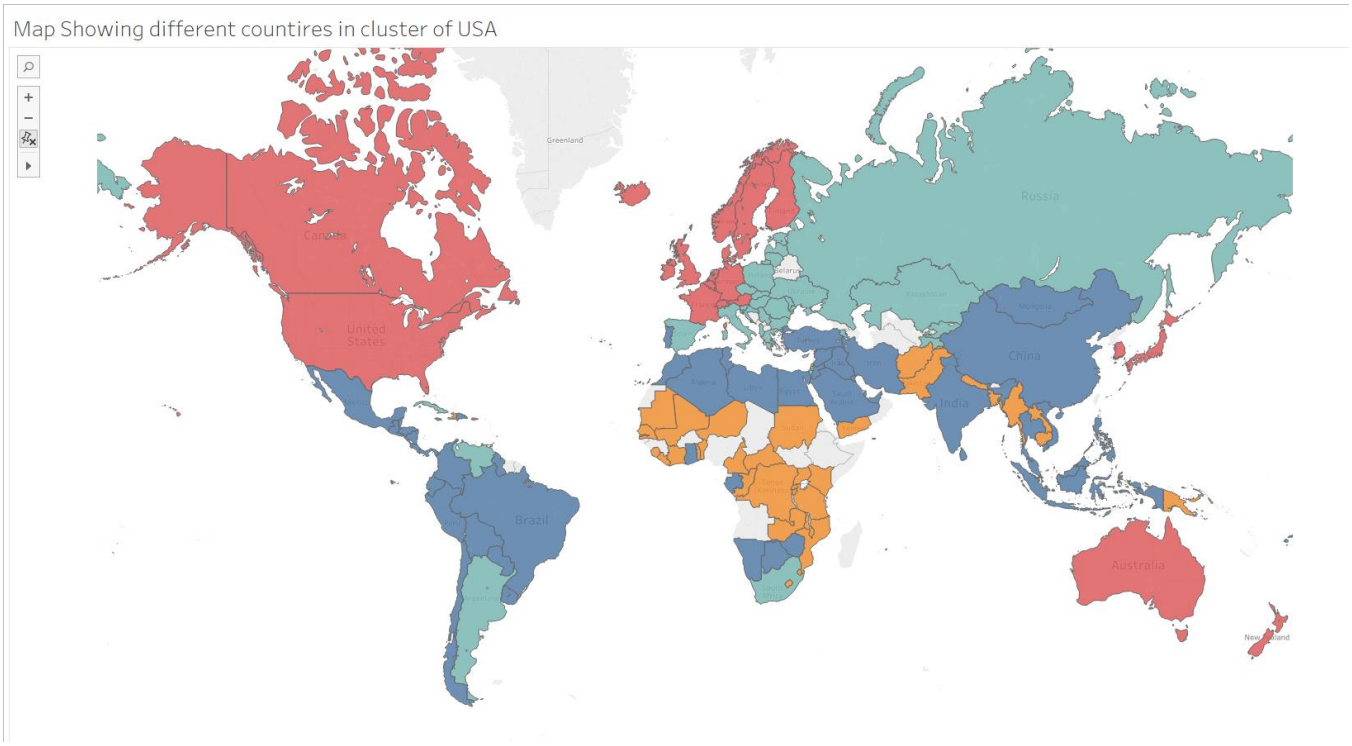
Step 4: Run the Data and Visualize

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, Clusters make sense since all the countries in the clusters have similar economic growth and the characteristics of the USA

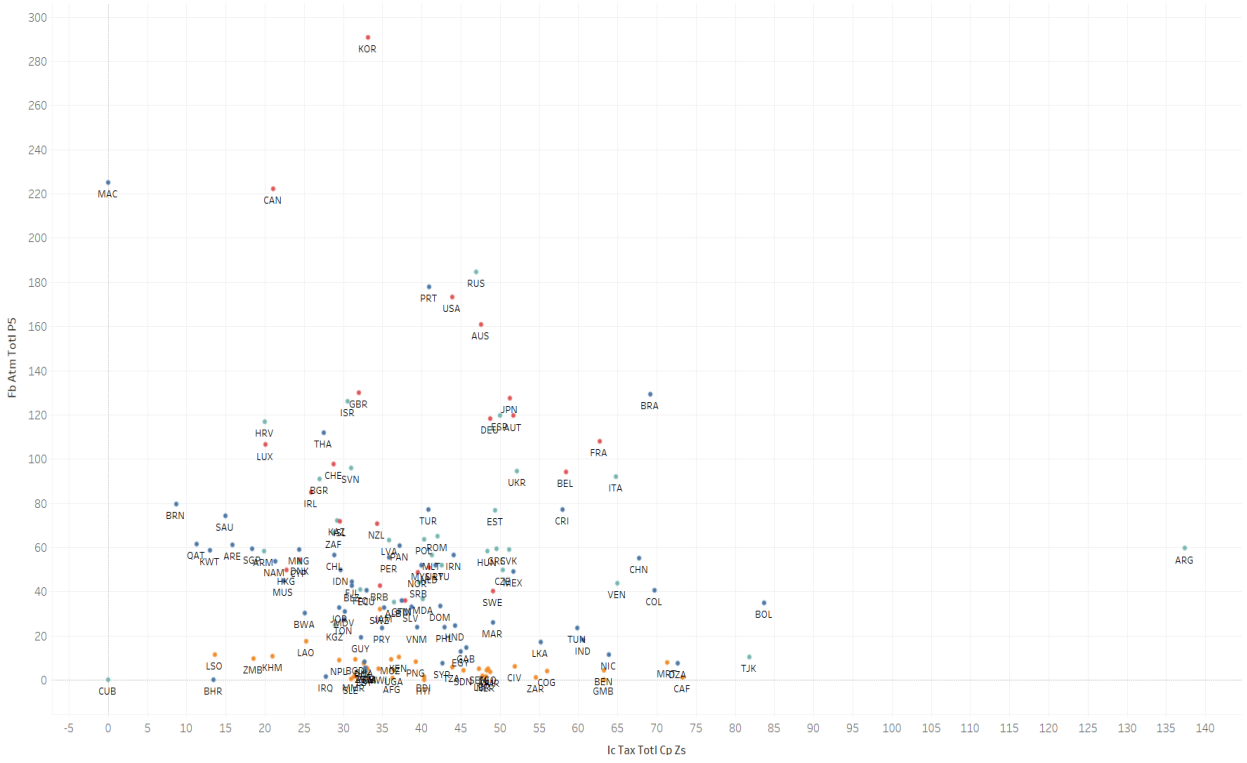


2. **What are the four countries in USA's cluster that are closest to the USA in terms of Total Tax Rate by ATM Machines? Hint: Create a scatterplot to graph the relationship between these two variables and color the markers by cluster.**

The below countries are in the USA's cluster and are closest to it in terms of total tax rate by ATM machines

- Australia
- United Kingdom
- Japan
- Canada

Total Tax rate to No. of ATM's in cluster containing USA



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

Answer this question:

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

The above countries are in the same cluster as the united states and when we closely observe the different economic and education parameters we can see that they have similarity.

The electricity penetration of these countries is 100%, the time required to get electricity for a home is very minimal, none of the population lives in an urban slum.
The Quality of Port Infrastructure which measures the perception of countries infrastructure is high and the country has high number of internationally acclaimed quality organizations.

The government spending on the education as part of GDP is very high which helped these countries to have highest education levels which would translate into better jobs and better environment for the businesses to thrive.
Thus, by choosing the above countries the retail organization will have a similar environment for the establishing the new business and to have success