International Expansion.

Overview.

A retail store chain has a presence of stores in the United States. The company 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.

It has been decided to segment the countries of the world based on various economic, demographic, education, and environment data. From this it 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.

Key Decisions.

What decisions need to be made?

The key decision that needs to be made here is which country outside of the USA should the retail store next expand to. This will be done through segmenting a list of countries that best match the economic, demographic, education and environmental data of the USA and allow management to decide from this list which best suits their next expansion plan.

What data is needed to inform those decisions?

A set of categories taken from the world bank website will be used to segment and further cluster countries into a suitable short list.

An example of the categories used are:

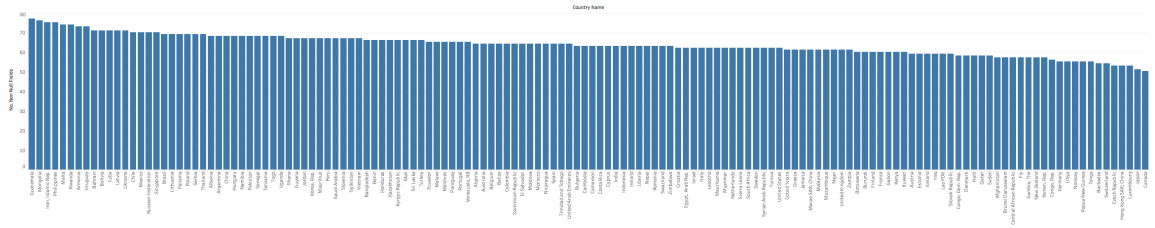
Category	Variable Name	Series Code
Economic	Total Gov Education Spend	SE_XPD_TOTL_GD_ZS
Economic	Total No. of ATMs	FB_ATM_TOTL_P5
Education_PTR	Ave Pupils per Teacher at a given level	UIS_PTRHC_3
Education_PTR	Ave Pupils per Teacher at a given level	UIS_PTRHC_56
Environment	Population living in slums.	EN_POP_SLUM_UR_ZS
Environment	Percentage of population with access to electricity.	EG_ELC_ACCS_ZS

Explore and Clean up the data.

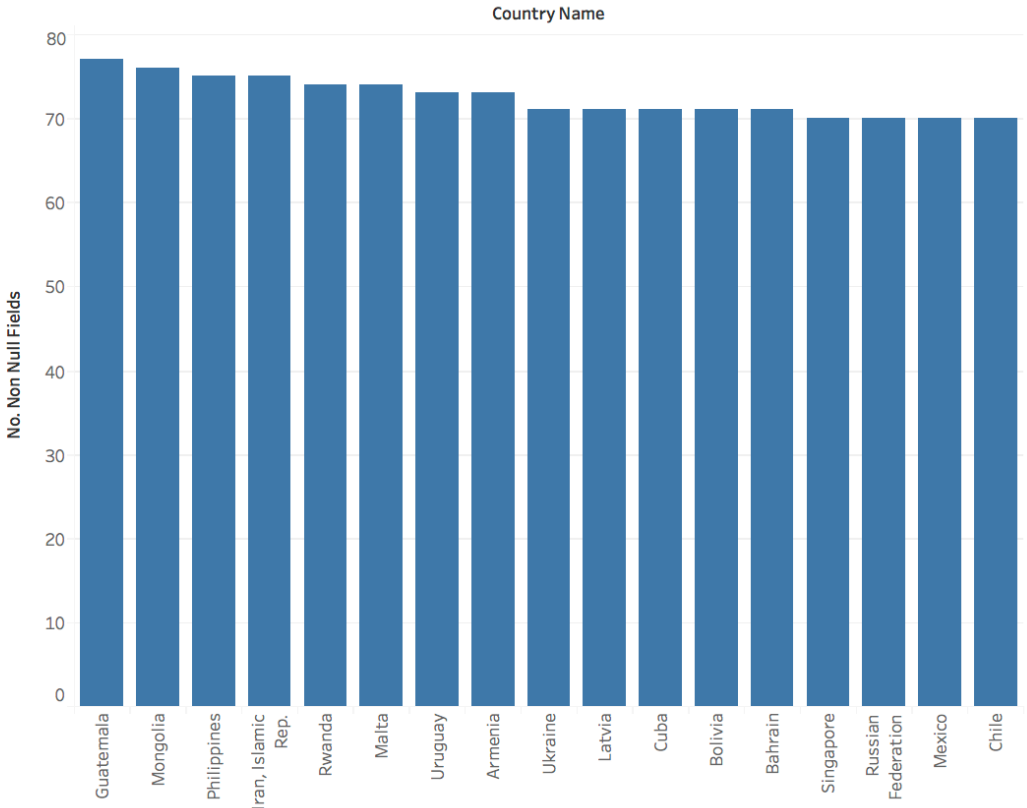
How many countries did you reduce the dataset to?

The dataset was removed of countries containing more than 25 null data points. The remaining number of countries was 144 in total.

Below is a bar chart of Countries in the dataset with a count of their non-null data fields.



Below is a closer look at the above chart, for Countries containing more than 70 non-null fields.



Which data categories will be used for PCA analysis?

The below categories will be used for PCA analysis:

Categories
Education_Avg Years
Education_literacy
Education_Pct

The education category contains 4 elements, “Education_Avg Years”, “Education_literacy”, “Education_Pct” and “Education_PTR”. “Education_PTR” only contains 3 variables in its category and will not need variable reduction.

Which variables are irrelevant for the analysis?

The below variables will be removed from the analysis due to their lack of association with the segmentation needs.

Category	Variable Name	Variable Code
Background	Internet Users in past 12 months	IT_NET_USER_P2
Health	Women aged 15-49 whom accept domestic violoence.	SG_VAW_BURN_ZS
Background	Percetage whom are HIV infected.	SH_DYN_AIDS_ZS
Background	Under 5 Mortality rate.	SH_DYN_MORT
Health	Number of Physicians	SH_MED_PHYS_ZS
Health	Cases of TB	SH_TBS_PREV
Health	Total Health Expenditure	SH_XPD_PCAP
Health	Undernourished population	SN_ITK_DEFC_ZS
Health	Ratio of dependants	SP_POP_DPND

Determine Clusters and Methodology.

What clustering method did you decide to use?

For the project, management has decided they would like to see four clusters.
Available clustering methods are K-Means, K-Medians and Neural-Networks to segment the clusters.
Below are the assessments of the different clustering methods.

K-Means Method Cluster Assessment.

K-Means Cluster Assessment Report

Summary Statistics

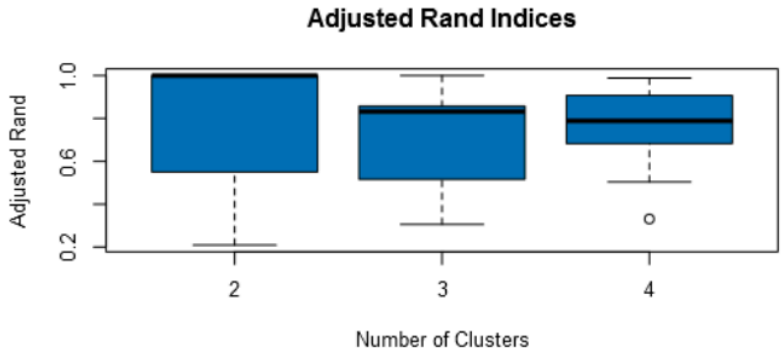
Adjusted Rand Indices:

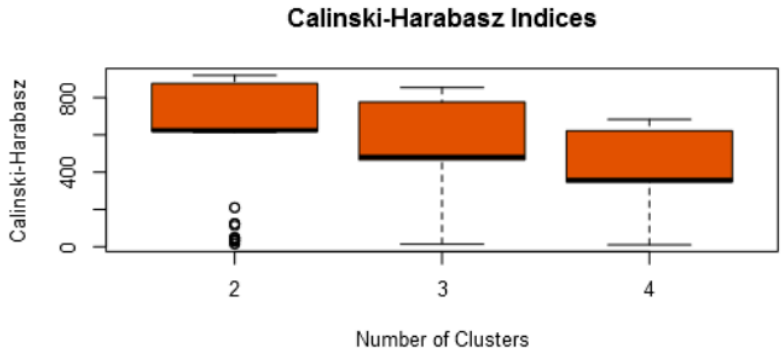
	2	3	4
Minimum	0.2094	0.306	0.3313
1st Quartile	0.6104	0.5294	0.6889
Median	1	0.8318	0.7884
Mean	0.8174	0.727	0.7812
3rd Quartile	1	0.8576	0.8999
Maximum	1	1	0.988

Calinski-Harabasz Indices:

	2	3	4
Minimum	14.56	13.95	10.69
1st Quartile	619.7	466.3	347.7
Median	623.7	480.9	357.1
Mean	647.7	553.4	433
3rd Quartile	875	774.8	621.1
Maximum	919	853.6	682.7

Plots





K-Medians Method Cluster Assessment.

K-Medians Cluster Assessment Report

Summary Statistics

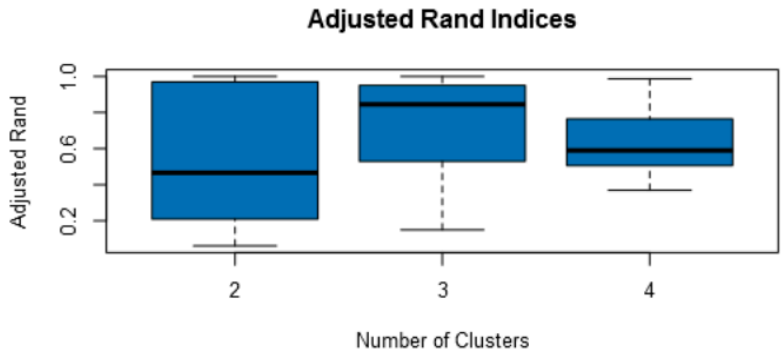
Adjusted Rand Indices:

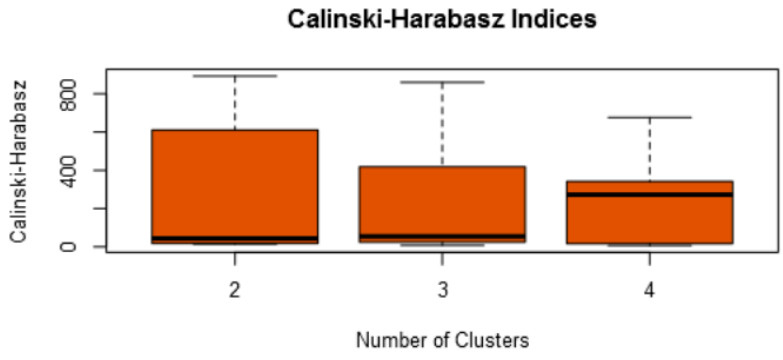
	2	3	4
Minimum	0.06121	0.15	0.369
1st Quartile	0.2139	0.5298	0.5068
Median	0.4655	0.845	0.5898
Mean	0.5125	0.7323	0.6417
3rd Quartile	0.9472	0.95	0.7623
Maximum	1	1	0.9867

Calinski-Harabasz Indices:

	2	3	4
Minimum	14.46	7.804	6.462
1st Quartile	17.41	24.02	16.5
Median	43.87	54.7	272.6
Mean	215.9	231	200.4
3rd Quartile	610.6	418.3	341.8
Maximum	892.5	859.7	676

Plots





Neural Gas Method Cluster Assessment.

Neural Gas Cluster Assessment Report

Summary Statistics

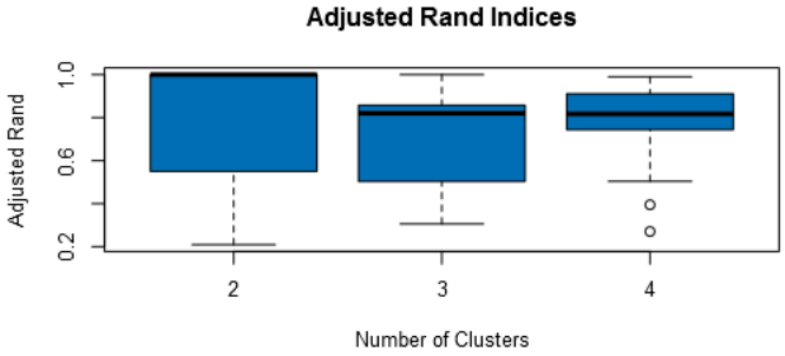
Adjusted Rand Indices:

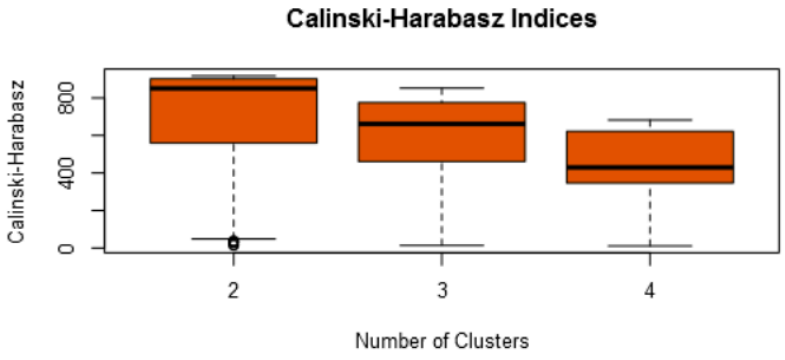
	2	3	4
Minimum	0.2094	0.306	0.271
1st Quartile	0.6104	0.5067	0.7492
Median	1	0.8203	0.817
Mean	0.809	0.721	0.7895
3rd Quartile	1	0.8576	0.9054
Maximum	1	1	0.9897

Calinski-Harabasz Indices:

	2	3	4
Minimum	14.64	13.92	10.97
1st Quartile	559.9	462.1	346.3
Median	850.5	662.5	429.9
Mean	674.4	571.3	444.4
3rd Quartile	902.1	775.2	621.8
Maximum	917.8	852.1	682.6

Plots



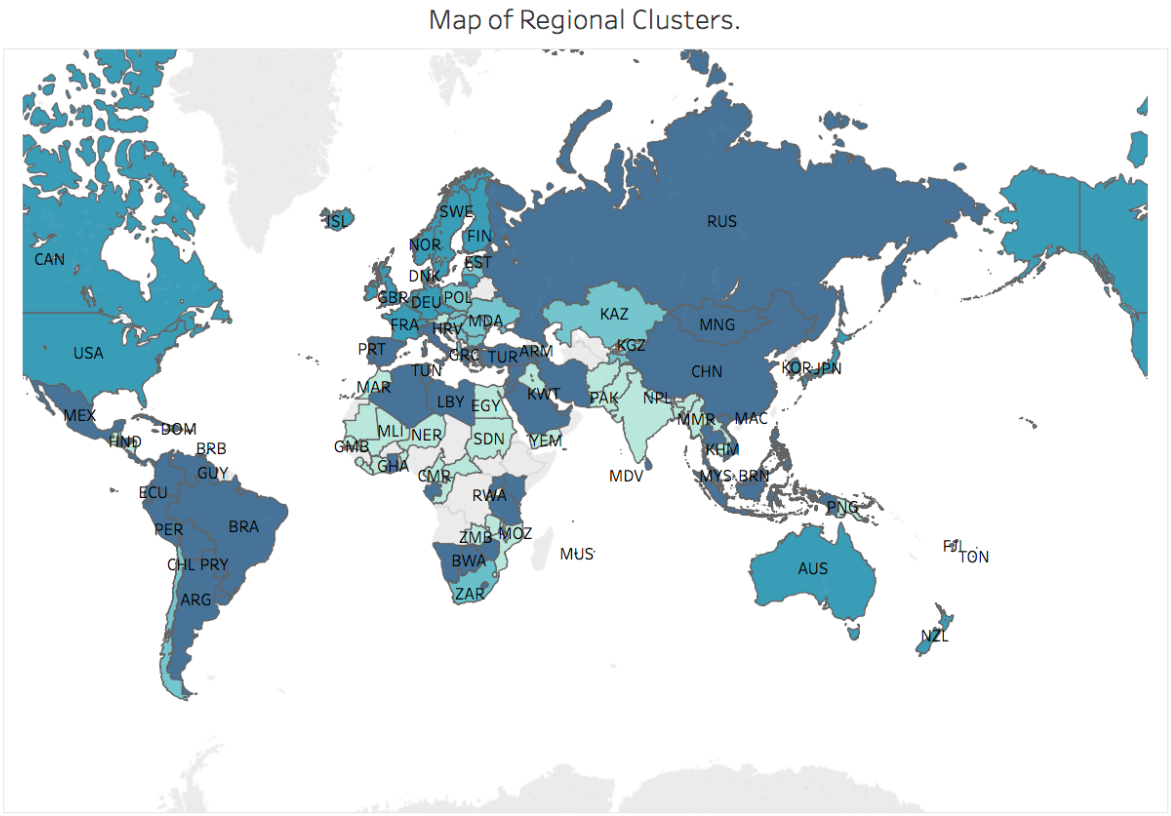


Looking through the above pairs of plots for the three clustering methods, we are looking for close tight ranges for both the ARI and CH plots, with high mean values for all clusters.

The K-Means and Neural Gas clustering methods are pretty close and I would say that the Neural Gas method edges it just with a slightly better assessment.

Modelling and Visualisation.

Below is a geographical map of the countries and their associated clusters.



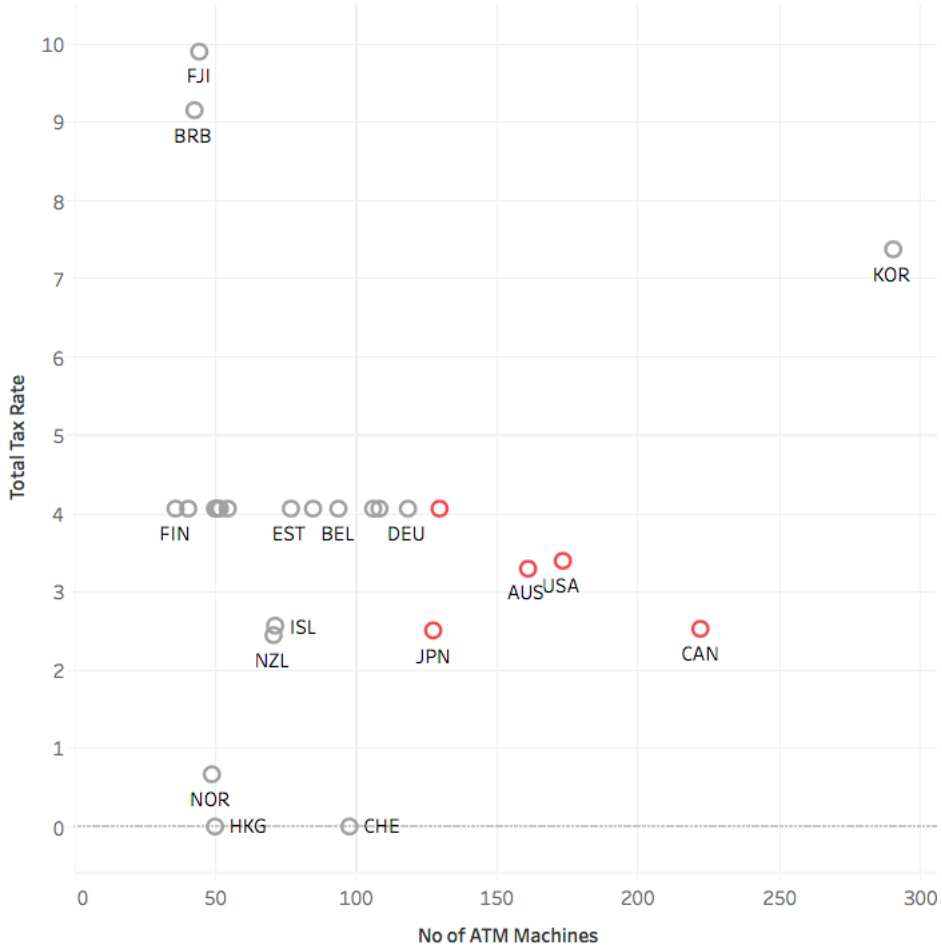
Do the clusters make sense?

Visually, the clusters look to be correct. The USA is clustered with CAN, GBR, AUS and some other quite large European countries which I would say look to be fair due to their likely similarities in education, economies and environment.

What are the countries in the USA's clusters that are closest to the USA in terms of Total Tax Rate by ATM machines?

Below is a scatterplot of the number of ATMs vs Tax Rate for the USA cluster filtered for the USA cluster.

Scatterplot of Number of ATMs vs Total Tax Rate (for the USA cluster)



The red highlighted points are the countries that resemble the USA the most.

Below are a list of the 4 countries:

Countries
Australia
Canada
Germany
Great Britain

Below is the list of all the countries in the USA cluster:

Countries
Australia
Barbados
Belgium
Canada
Czech Republic

Denmark
Estonia
Fiji
Finland
France
Germany
Hong Kong SAR, China
Iceland
Ireland
Japan
Korea, Rep.
Lithuania
Luxembourg
Netherlands
New Zealand
Norway
Sweden
Switzerland
United Kingdom
United States

Why did you decide to choose these countries?

The above list of countries was chosen due to the fact that they most resemble the USA in educational, economic and environmental terms