Project: International Expansion

Step 1: Key Decisions

The company is planning a major expansion to a foreign market and thus, the business decision to be made is which country to choose for opening the new stores. To ensure viability of this endeavor the selected country needs to be like the US in as many aspects as possible. For example, it could be argued that its GDP per capita should be roughly the same level so that customers have the purchasing power to buy the products displayed or that the market or that it has a booming economy (growth index) that can support another player in the retail sector.

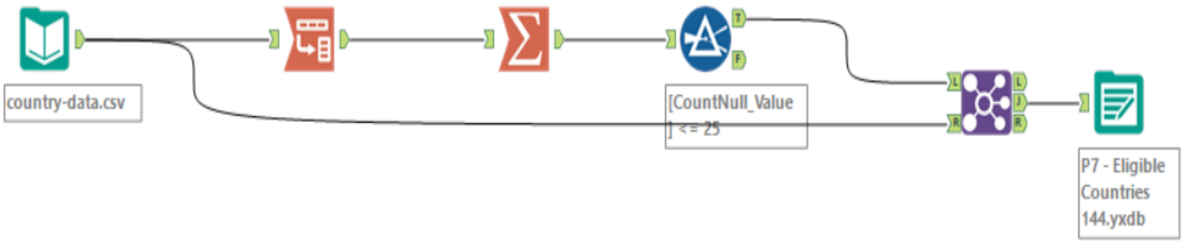
For this analysis, we would also need data in relation to education, as in what is the average years of education of the population in that country. If similar to the US, it could mean that we might also anticipate similar consumer behavior. Also, we need education data on the various age groups to optimize the type and range of products offered.

Finally, we need data on either the natural and/or built environment conditions such as no. of hours with electricity per day. It's hard to imagine selling many products without refrigerators working 24/7. Another example is percentage of people living in urban centers.

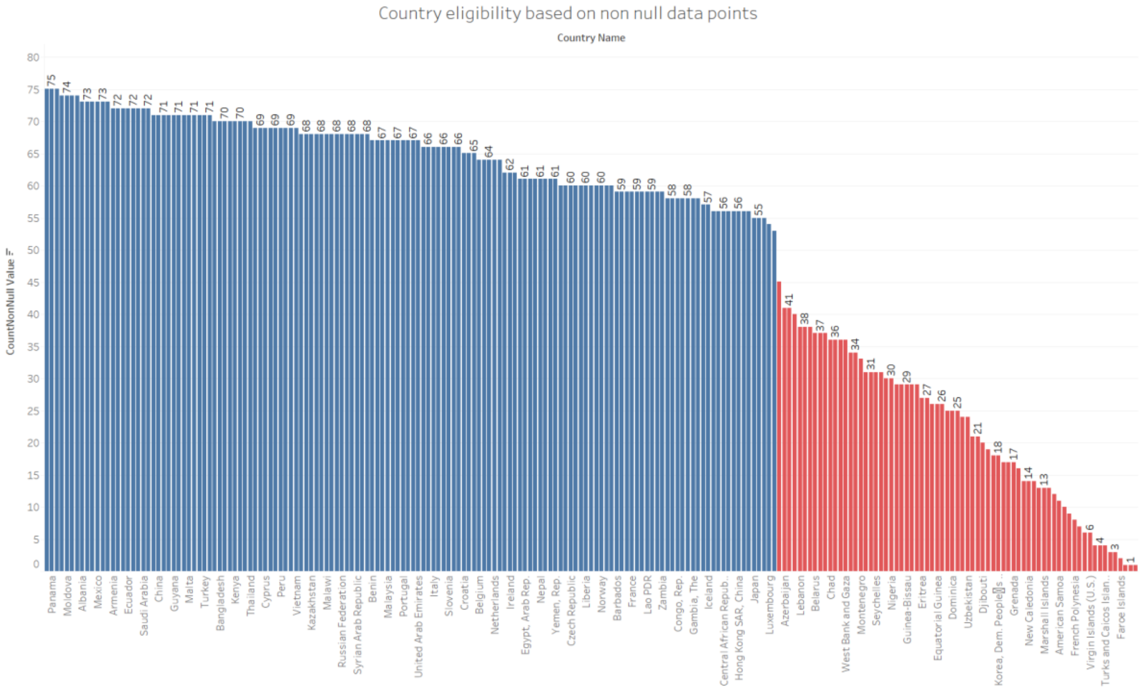
Step 2: Explore and Cleanup the Data

The first step of our analysis is to prepare an analytical dataset that is suitable for our analysis. In our case it means that, first, we should exclude countries with many missing data points. Setting a threshold of 25 missing variables our dataset was finally reduced to 144 countries.

Alteryx workflow



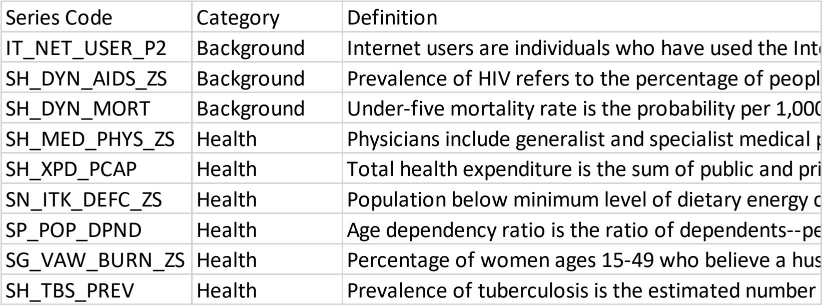
Bar chart of eligible countries based on non-null data points (descending order)



After removing the countries with more than 25 missing variables, our data set is reduced to 144 countries. First country is Mongolia, last before the cutoff is Papua New Guinea, and first after cutoff is Georgia.

The variables can be reduced by applying PCA and by looking at the variables there are three categories to consider. These are: 1) Education_Avg Years (30 variables), which is basically the same data albeit being shown separately for the different age groups. 2) Education_Pct (15 variables), showing the percentage of people over 25 years of age that have completed education cycles, and 3) Education_literacy (7 variables), containing information on young people's education with references to both genders.

Moreover, there are nine variables under the Health and Background categories that will not add any value to our analysis. For example, prevalence of HIV has little to do with retail business, so these variables are removed from the dataset.

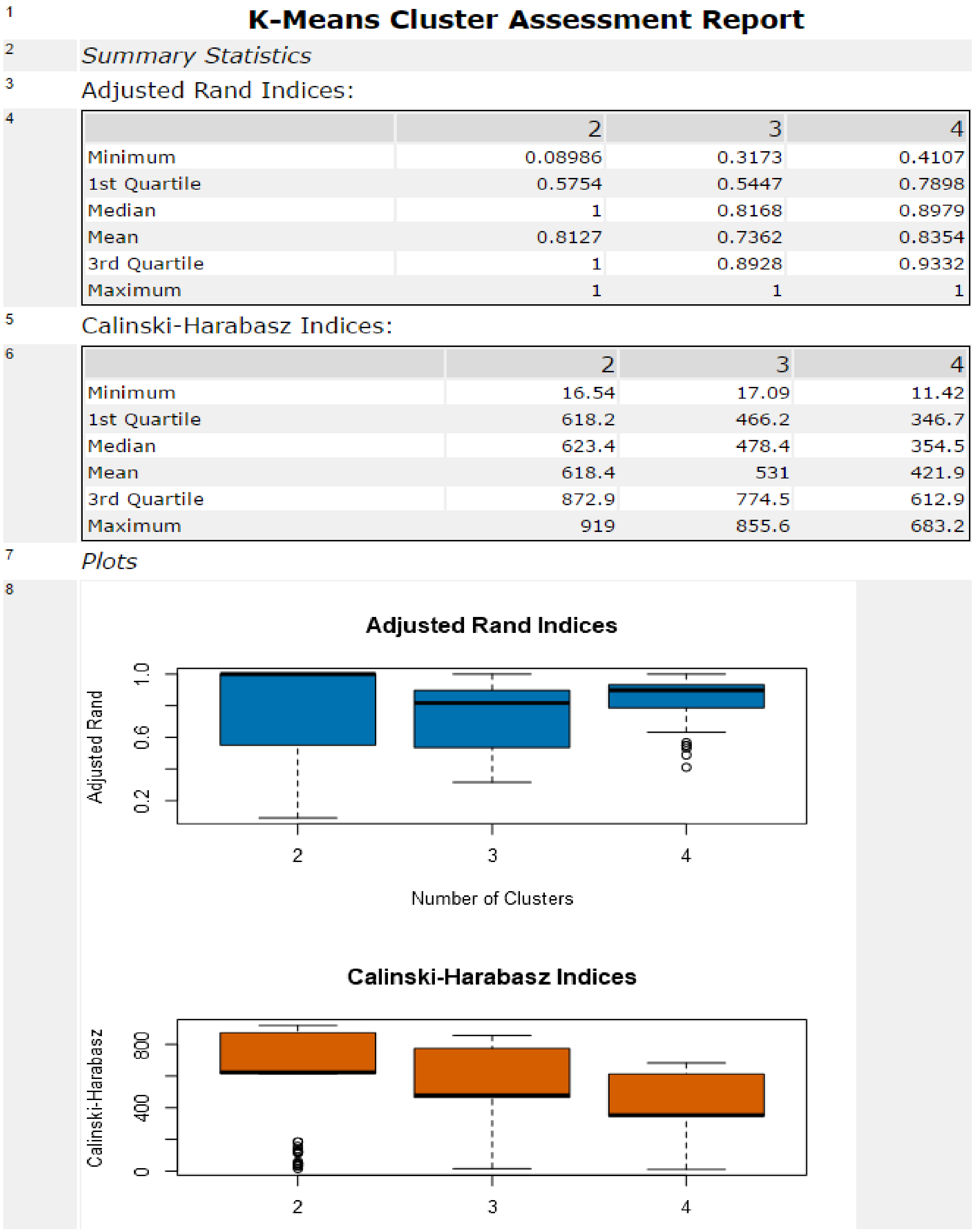


Series Code	Category	Definition
IT_NET_USER_P2	Background	Internet users are individuals who have used the Int
SH_DYN_AIDS_ZS	Background	Prevalence of HIV refers to the percentage of peopl
SH_DYN_MORT	Background	Under-five mortality rate is the probability per 1,00
SH_MED_PHYS_ZS	Health	Physicians include generalist and specialist medical
SH_XPD_PCAP	Health	Total health expenditure is the sum of public and pri
SN_ITK_DEFC_ZS	Health	Population below minimum level of dietary energy c
SP_POP_DPND	Health	Age dependency ratio is the ratio of dependents--pe
SG_VAW_BURN_ZS	Health	Percentage of women ages 15-49 who believe a hus
SH_TBS_PREV	Health	Prevalence of tuberculosis is the estimated number

Step 3: Determine Clusters and Methodology

Since the number of clusters has been specified by the manager (4 clusters) we will apply k-centroid cluster analysis to our dataset. Using the K-Centroids Diagnostics tool we will make an informed decision on which clustering methodology to follow, i.e. k-means, k-medians or neural gas. For all the methodologies, the data was first standardized using the z-score.

<u>Results</u>



K-Means Cluster Assessment Report

Summary Statistics

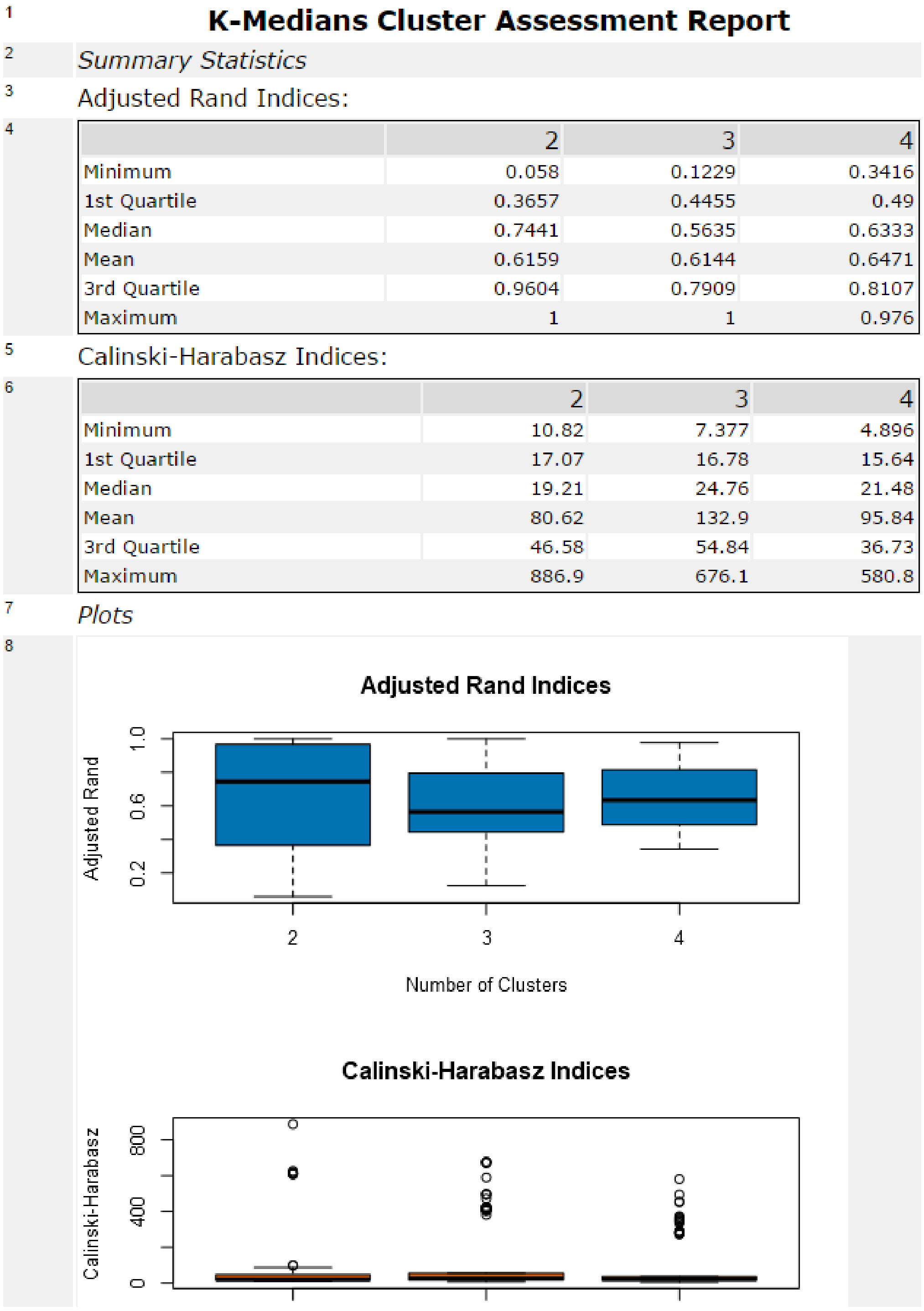
Adjusted Rand Indices:

	2	3	4
Minimum	0.08986	0.3173	0.4107
1st Quartile	0.5754	0.5447	0.7898
Median	1	0.8168	0.8979
Mean	0.8127	0.7362	0.8354
3rd Quartile	1	0.8928	0.9332
Maximum	1	1	1

Calinski-Harabasz Indices:

	2	3	4
Minimum	16.54	17.09	11.42
1st Quartile	618.2	466.2	346.7
Median	623.4	478.4	354.5
Mean	618.4	531	421.9
3rd Quartile	872.9	774.5	612.9
Maximum	919	855.6	683.2

Plots



K-Medians Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4
Minimum	0.058	0.1229	0.3416
1st Quartile	0.3657	0.4455	0.49
Median	0.7441	0.5635	0.6333
Mean	0.6159	0.6144	0.6471
3rd Quartile	0.9604	0.7909	0.8107
Maximum	1	1	0.976

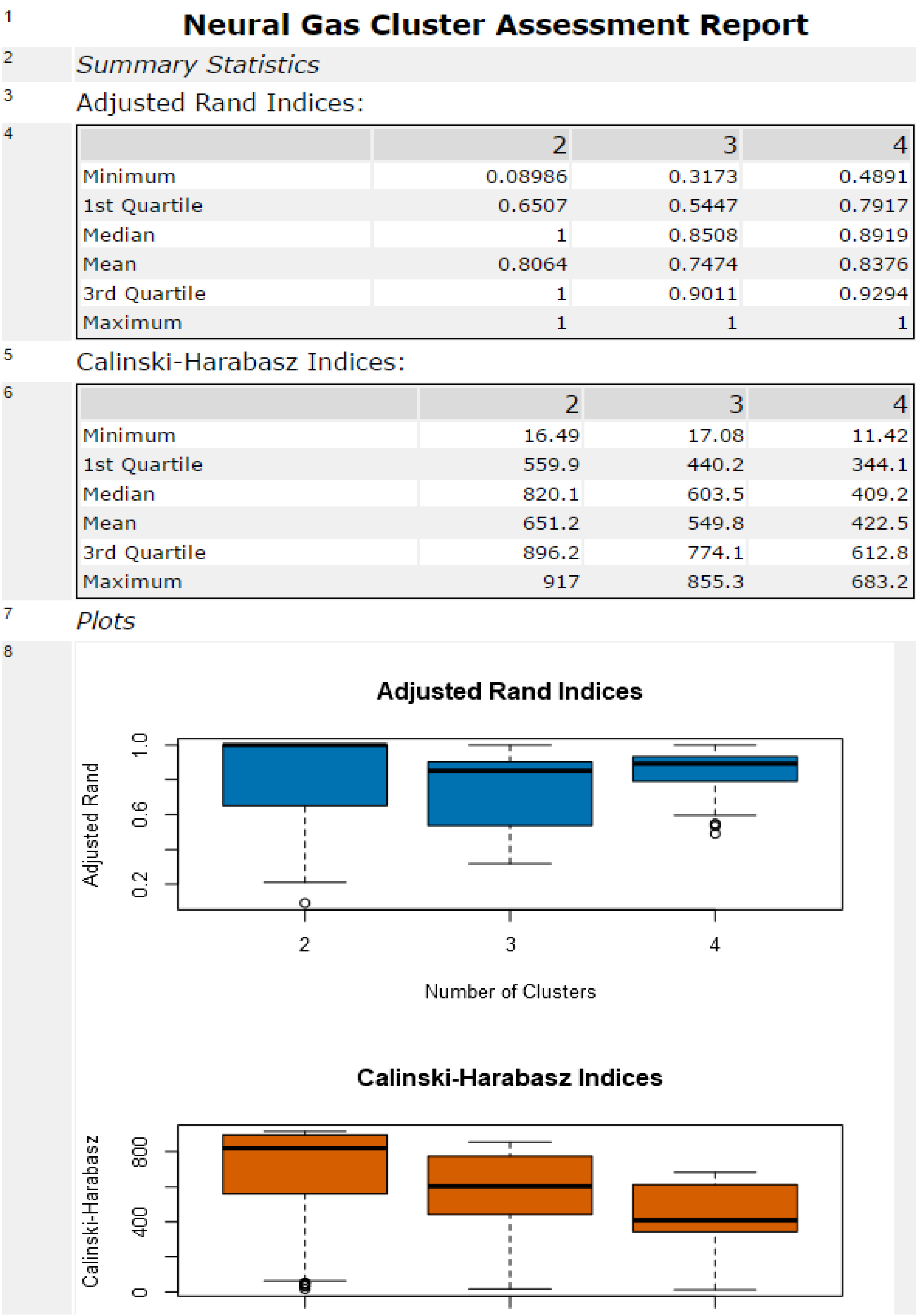
Calinski-Harabasz Indices:

	2	3	4
Minimum	10.82	7.377	4.896
1st Quartile	17.07	16.78	15.64
Median	19.21	24.76	21.48
Mean	80.62	132.9	95.84
3rd Quartile	46.58	54.84	36.73
Maximum	886.9	676.1	580.8

Plots

Adjusted Rand Indices

Calinski-Harabasz Indices



Neural Gas Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4
Minimum	0.08986	0.3173	0.4891
1st Quartile	0.6507	0.5447	0.7917
Median	1	0.8508	0.8919
Mean	0.8064	0.7474	0.8376
3rd Quartile	1	0.9011	0.9294
Maximum	1	1	1

Calinski-Harabasz Indices:

	2	3	4
Minimum	16.49	17.08	11.42
1st Quartile	559.9	440.2	344.1
Median	820.1	603.5	409.2
Mean	651.2	549.8	422.5
3rd Quartile	896.2	774.1	612.8
Maximum	917	855.3	683.2

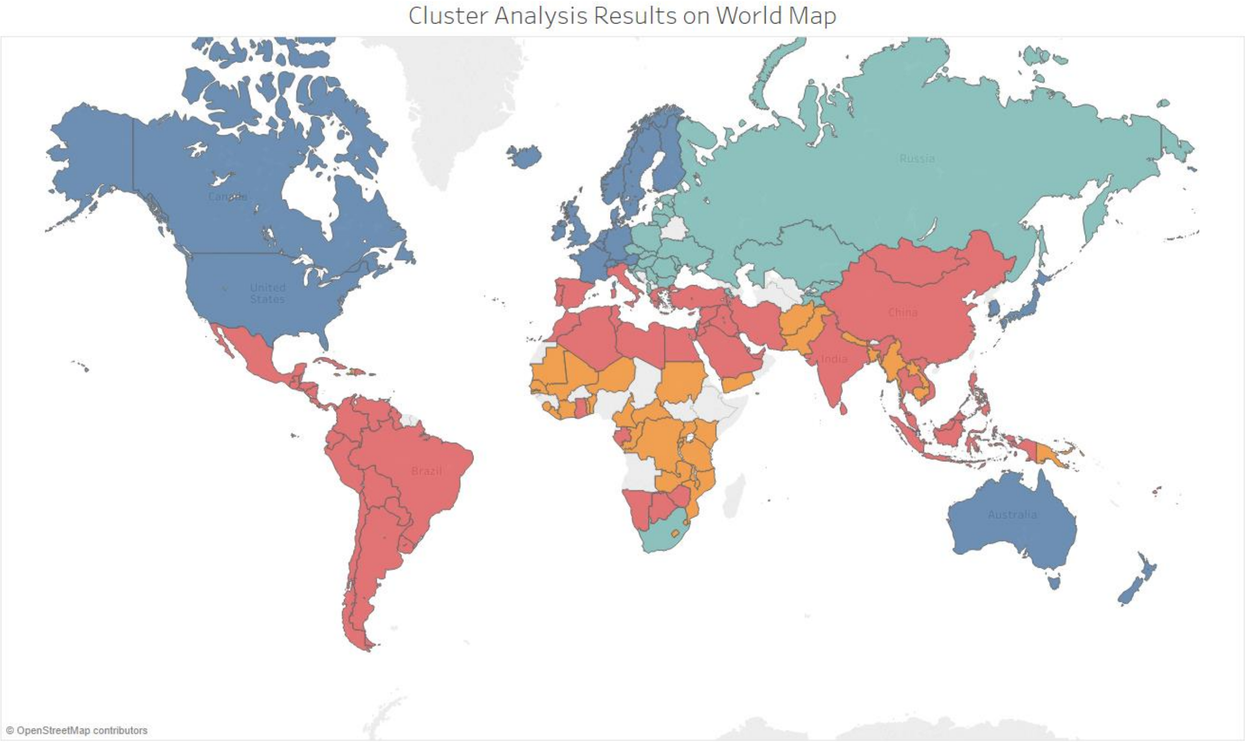
Plots

For a four-cluster analysis:

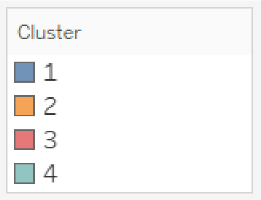
K-means: AR = 0.8979, CH = 354.5
K-medians: AR = 0.6333, CH = 21.48
Neural Gas: AR = 0.8919, CH = 409.2

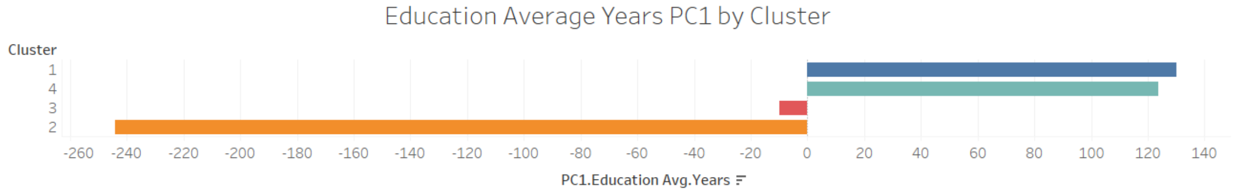
Neural Gas appears to be performing better than the other two in our case, so we will choose Neural Gas to model our solution.

Step 4: Run the Data and Visualize


Cluster Analysis Results on World Map
Canada
United States
Russia
China
India
Brazil
Australia
© OpenStreetMap contributors

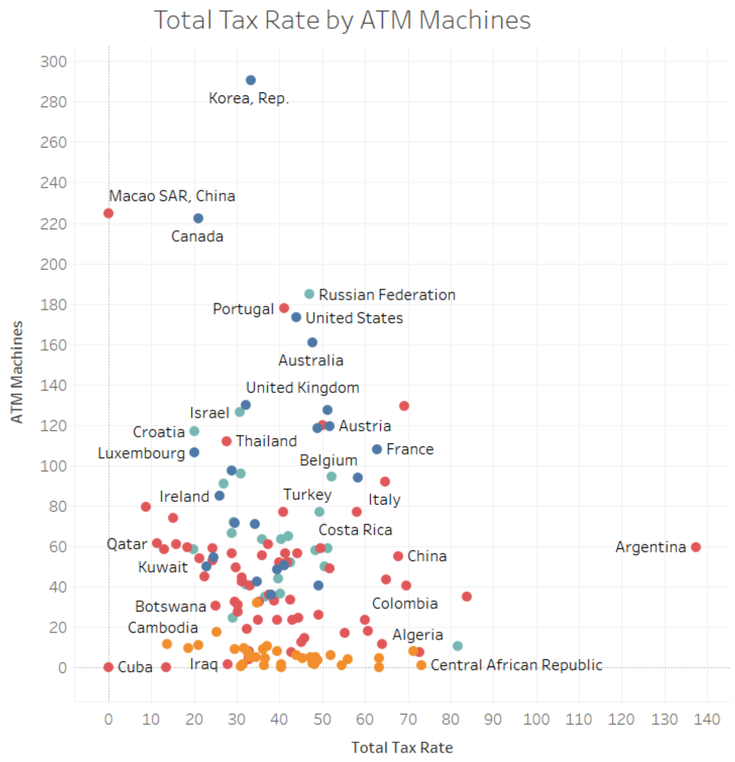
The clusters appear to make perfect sense, especially cluster no.1 where the US is included. We see it also includes other countries with strong economies and high literacy, i.e. Canada, Western Europe, Japan, South Korea, Australia and New Zealand. Cluster no.2 is about countries that are very poor, e.g. central African countries, Afghanistan, Pakistan, South East Asian countries (interestingly enough countries with many tourists like Thailand and Vietnam seem to be performing better than their neighbors). Cluster 3 includes developing countries or developed countries that may have been hit by recession, such as southern European countries. In my opinion, it is the most loosely defined cluster. Finally, cluster 4 is very interesting as we clearly see that except for South Africa, it includes ex-communist countries (most of them having been in a union between them). It could be argued that their 'recent' past has created a unique mix of socioeconomic conditions that so clearly distinguishes them from cluster 3 in which they otherwise would have been in my opinion.





Looking at education data by cluster the countries seem to be well matched in that respect too.

Total Tax Rate by ATM Machines



Closest Neighbors to USA

1. Australia
2. United Kingdom
3. Japan
4. Austria

Step 5: Recommendation

It is recommended that the retail store business should explore opening new stores to countries most resembling the US, i.e. countries that are in the same cluster (no.1). These are:

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