**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:

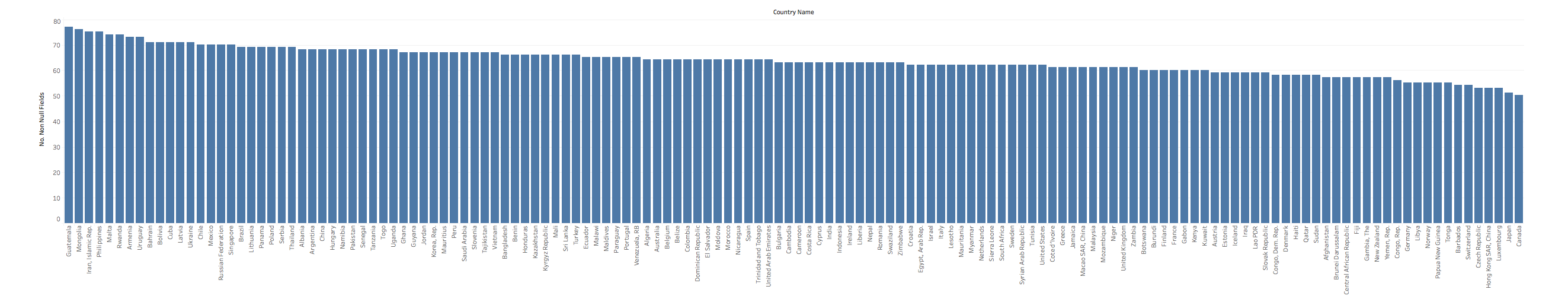
|  |  |  |
| --- | --- | --- |
| **Category** | **Variable Name** | **Series Code** |
| Economic | Total Gov Education Spend | SE\_XPD\_TOTL\_GD\_ZS |
| Economic | Total No. of ATMs | FB\_ATM\_TOTL\_P5 |
| Education\_PTR | Ave Pupils per Teacher at a given level | UIS\_PTRHC\_3 |
| Education\_PTR | Ave 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.**

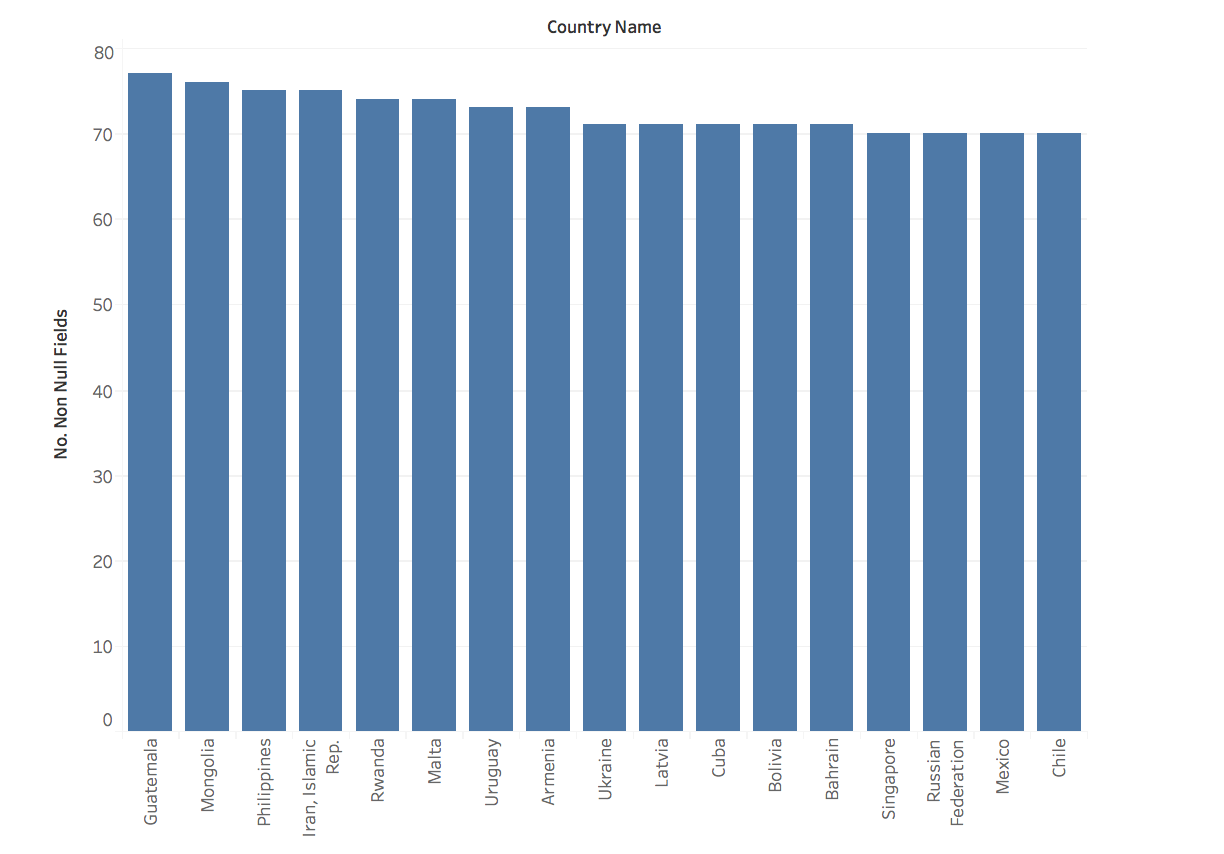
**How many countries did you reduce the dataset to?**

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.**

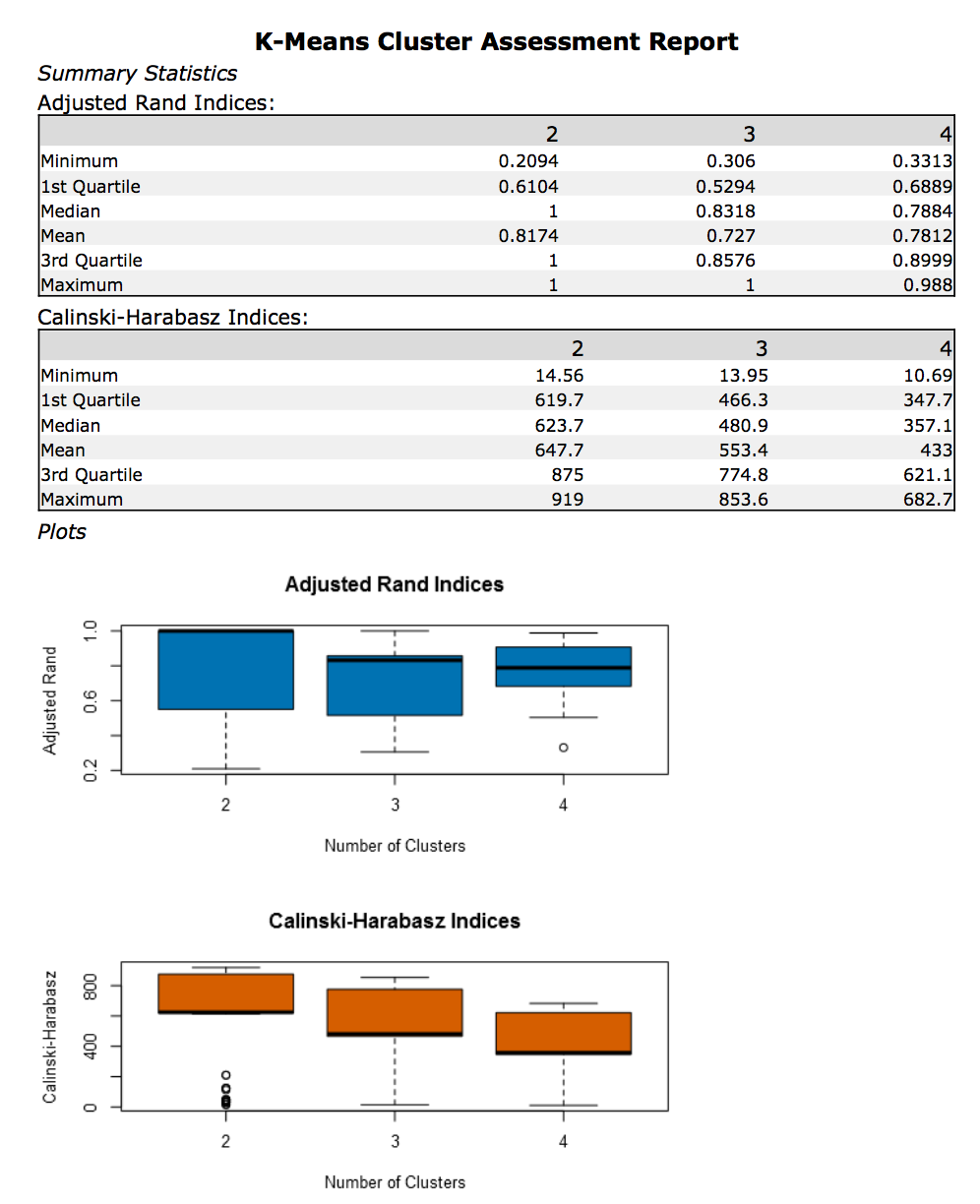
**What clustering method did you decide to use?**

For the project, management has decided they would like to see four clusters.

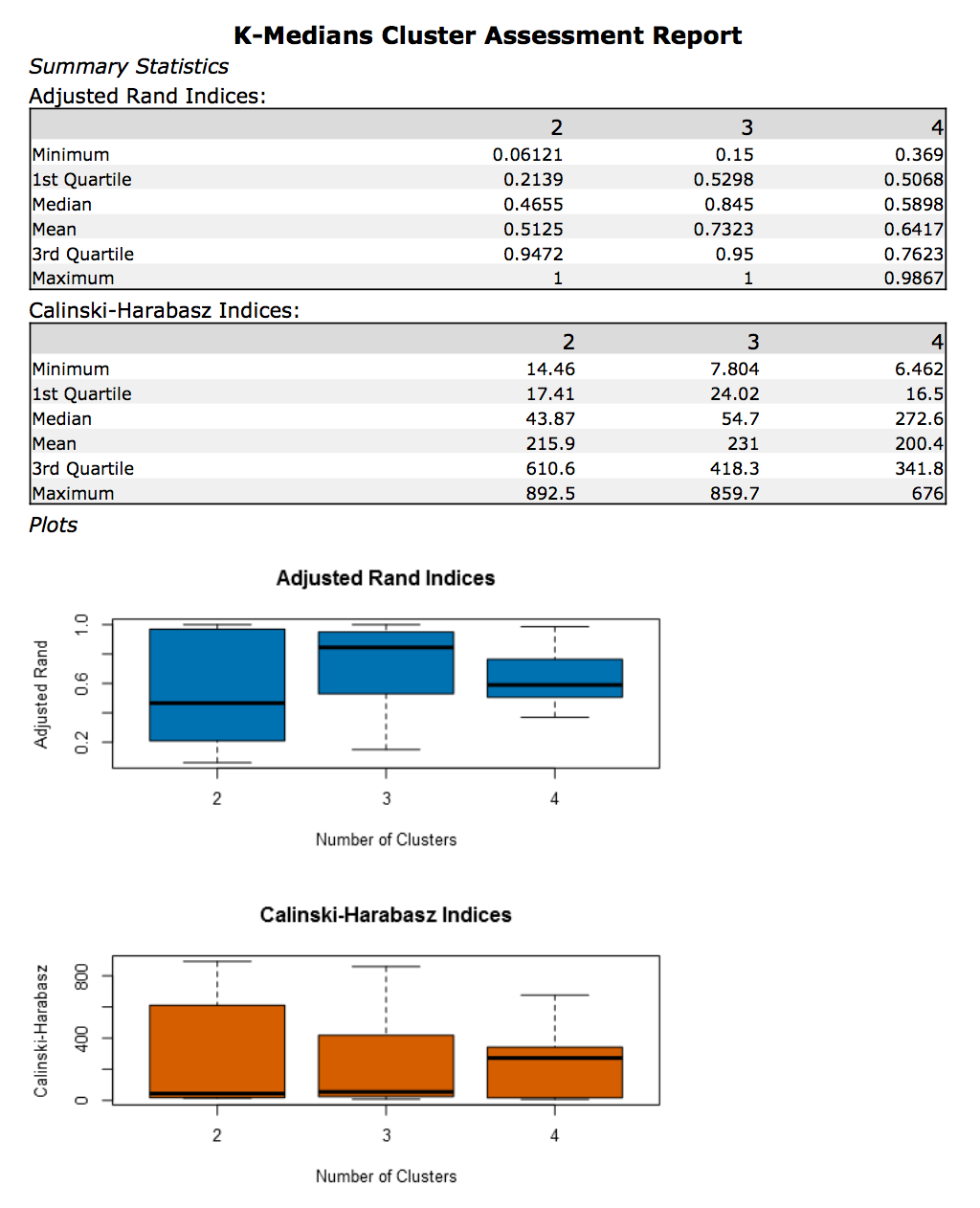
Available clustering methods are K-Means, K-Medians and Neural-Networks to segment the clusters.

Below are the assessments of the different clustering methods.

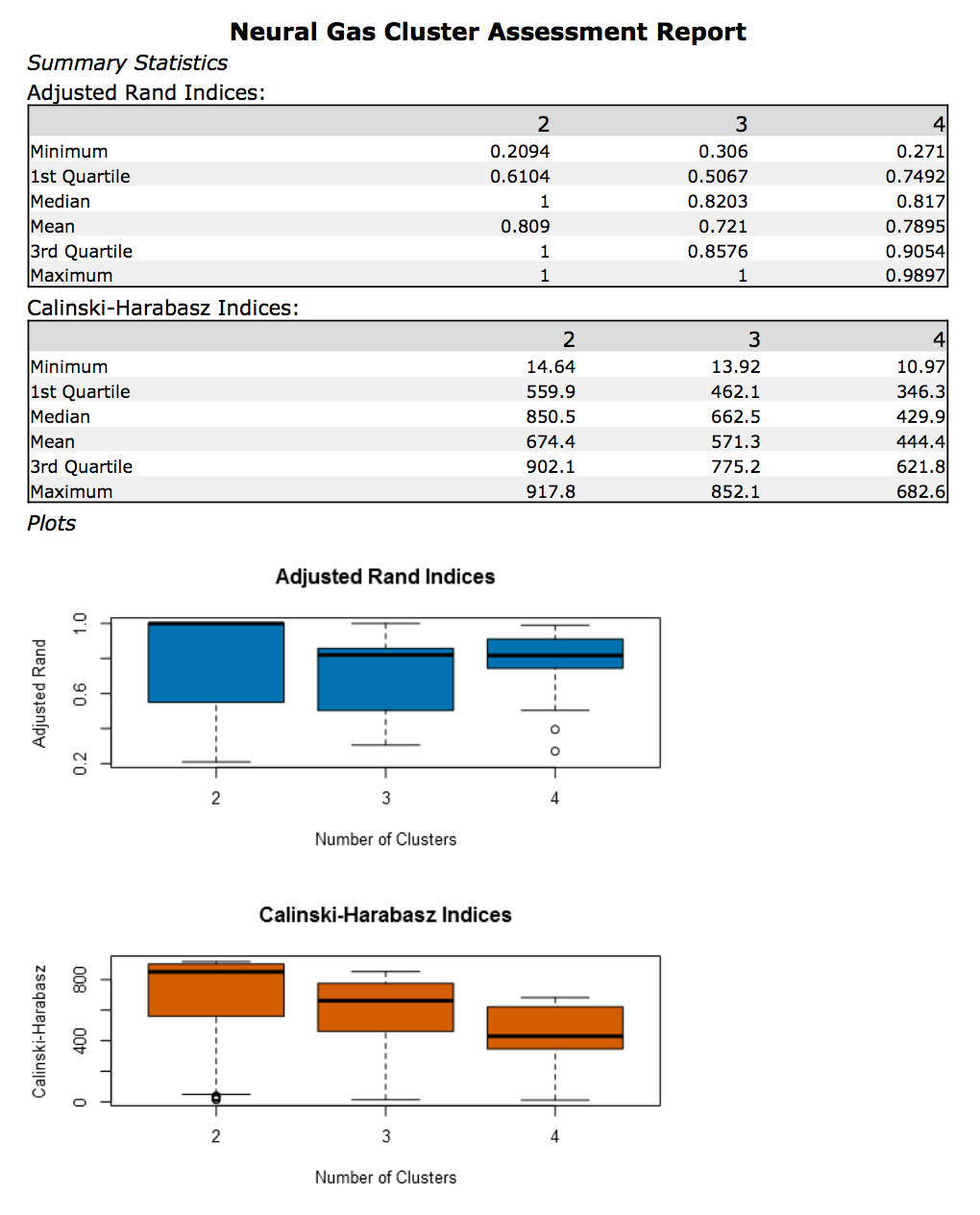
K-Means Method Cluster Assessment.



K-Medians Method Cluster Assessment.



Neural Gas Method Cluster Assessment.

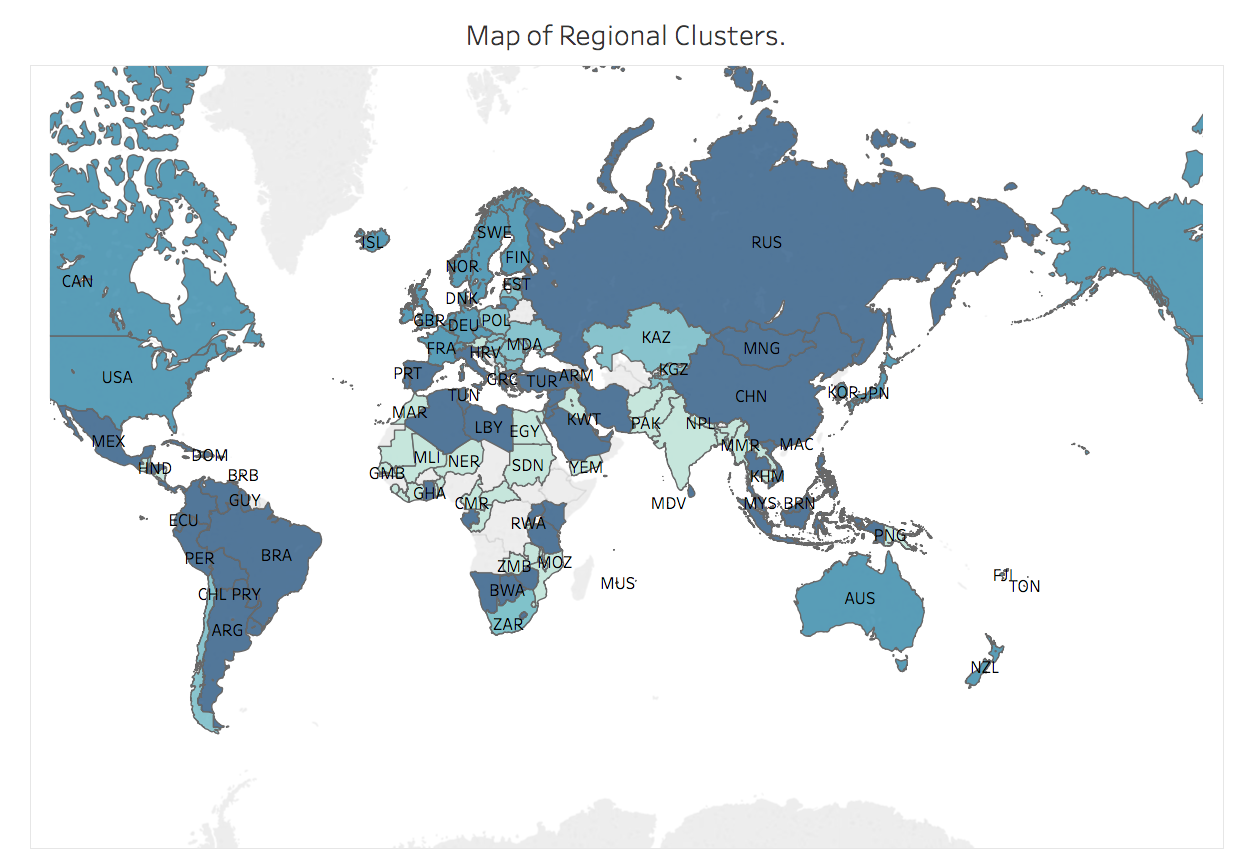


Looking through the above pairs of plots for the three clustering methods, we are looking for close tight ranges for both the ARI and CH plots, with high mean values for all clusters.

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.



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