Project: Clustering Data

Step 1: Key Decisions

A retail store chain in the United States of America are thinking of expanding to other countries and want to figure out which countries are similar in terms of economy, demographics, education, and environment to the United States of America. Data from the World Bank web site will be used to identify countries similar to the United States of America in terms of the above features. All analysis was conducted in Alteryx Designer 2018.3 (Colorado, USA) and data visualisation was conducted in Tableau Desktop 2018.3 (Washington, USA).

Step 2: Explore and Cleanup the Data

The data provided from the World Bank web site contains 215 countries with 77 variables. Data not relevant to this problem (Internet users, Prevalence of HIV, Mortality rate, Number of physicians, Health expenditure, Prevalence of undernourishment, Age dependency ratio, Women who believe a husband is justified in beating his wife when she burns the food and Prevalence of tuberculosis) were first removed from the dataset leaving 68 variables.

As per the guidelines for this assignment, countries missing more than 25 variables were also removed leaving 144 countries in this analysis.

After imputing null values, a principle component analysis was conducted to reduce the dimensionality of three topics; Average years of education, population over 25 with degrees and literacy rate. A scree plot was utilised to help decide the number of principle components to retain. The resulting dataset had 36 variables which was then used for the cluster analysis.

Step 3: Build your Models

Data was first standardised to zero mean and unit variance. Three k centroid methods were then explored (K-means, K-median, Neural Gas) using the adjusted rand index (ARI) as measure of cluster stability and the Calinski-Harabasz Index (CH) as a measure of compactness and distinctness of the clusters.

While four clusters do not appear to be an optimal solution to this problem, as per the criteria of this report, the manager would like to see four clusters in the report. When exploring the four-cluster solution, both the K-median and Neural Gas methods demonstrate high median values for both indices. However, the Neural Gas method appears to produce a tighter spread on both the ARI and CH and was as such chosen for this analysis (fig 1).

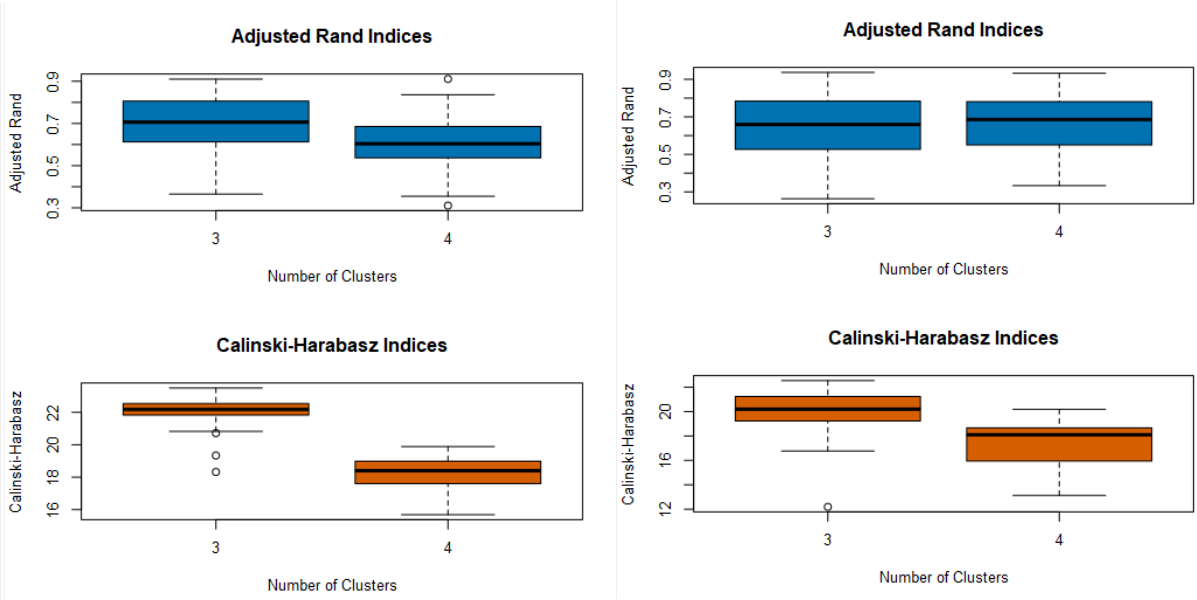


Fig (1) CH and ARI indices for the Neural Gas method (left) and the K-median method (right).

The results from the cluster analysis produced four cluster groups as requested by management (fig 2). The countries which were grouped with the USA (red marks) appear to be appropriately similar to the USA as they are first world countries with a developed economy, and similar education and income levels.

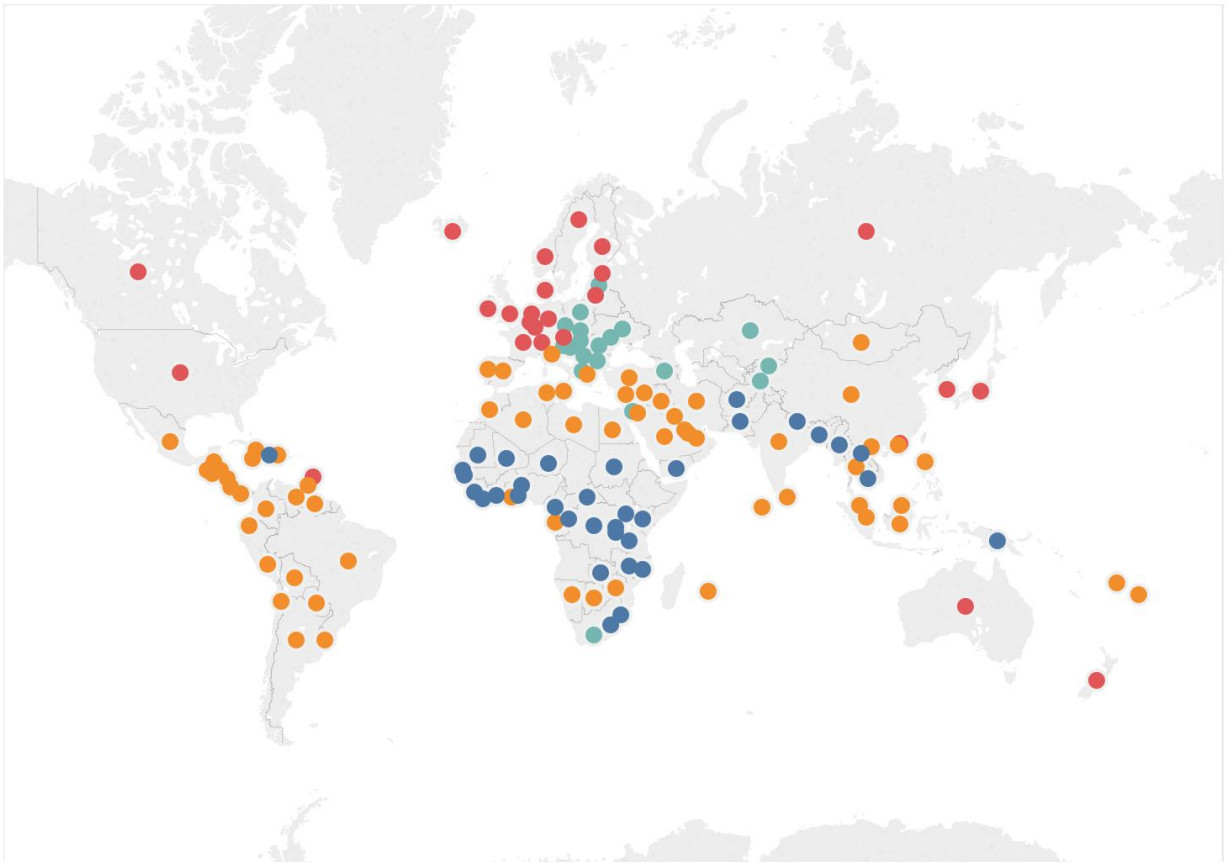
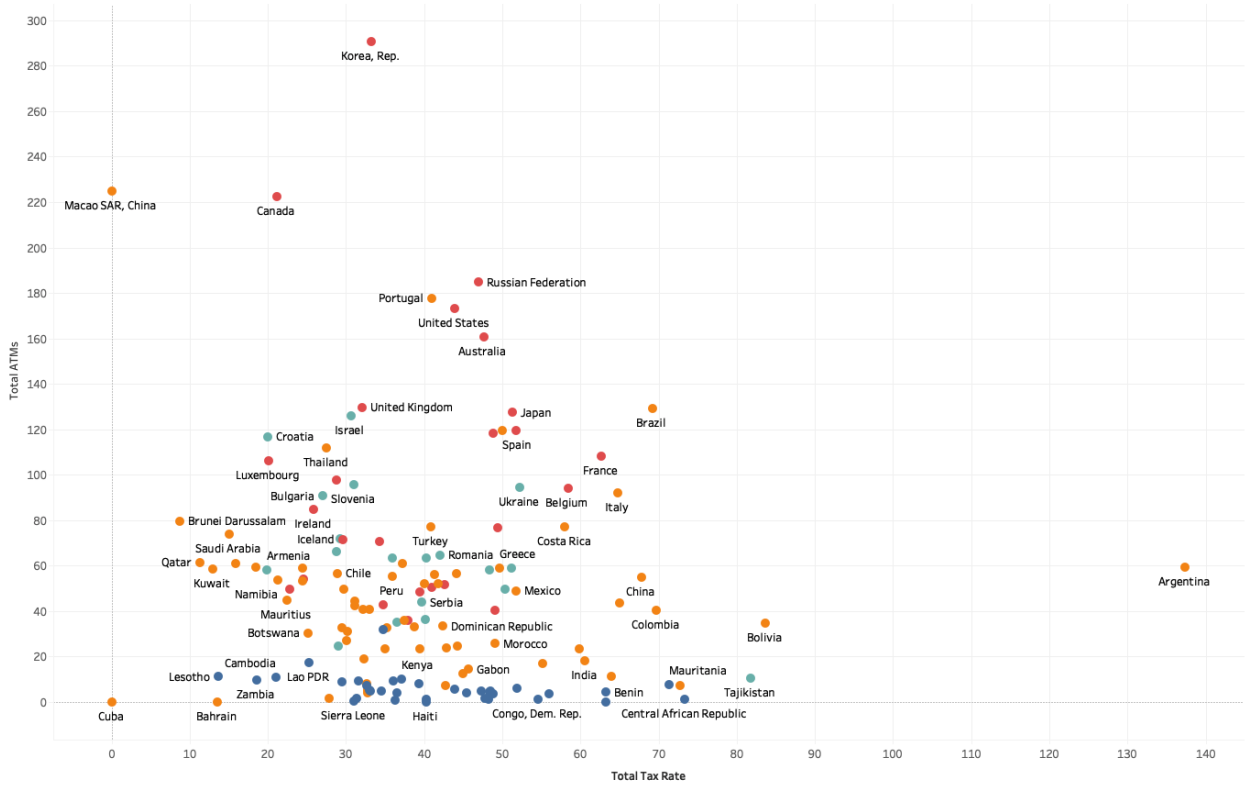


Fig (2) Countries grouped into four clusters.

As requested by management, the four countries which are closest to the USA in terms of total tax rate by ATM machines are Australia, Russia Japan and United Kingdom (fig 3).

.



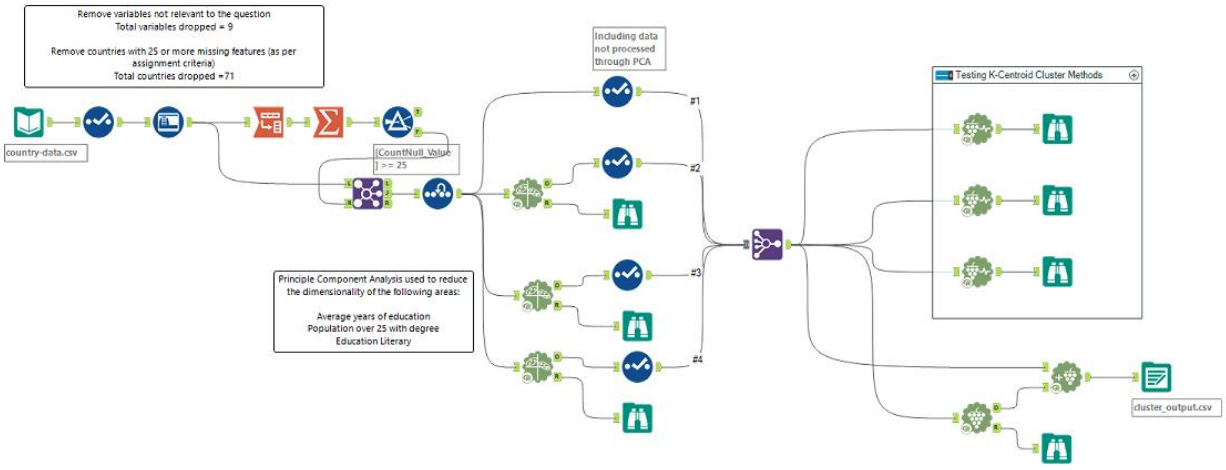
Fig(3) Scatter plot of ATMs and Total Tax Rate by Cluster group.

Step 4: Recommendation

It is recommended that the retail store should consider expanding to one of the countries in USA's cluster group. This would be appropriate as these countries are similar in terms of economy, demographics, education, and environment to the United States of America. The countries which should be considered are as follows:

Austria, Barbados, Belgium, Canada, Denmark, Estonia, Finland, France, Germany, China, Iceland, Ireland, Japan, Korea, Rep., Lithuania, Luxembourg, Netherlands, New Zealand, Norway, Russian Federation, Sweden, Switzerland, United Kingdom.

Full Alteryx Workflow is presented below (fig 4).



Fig(4) Alteryx Workflow