Project: International Expansion

Step 1: Key Decisions

Key Decisions:

Answer these three questions

1. What decisions needs to be made?

   Which country (besides the US) would be best for the retail store chain to build a new store?

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

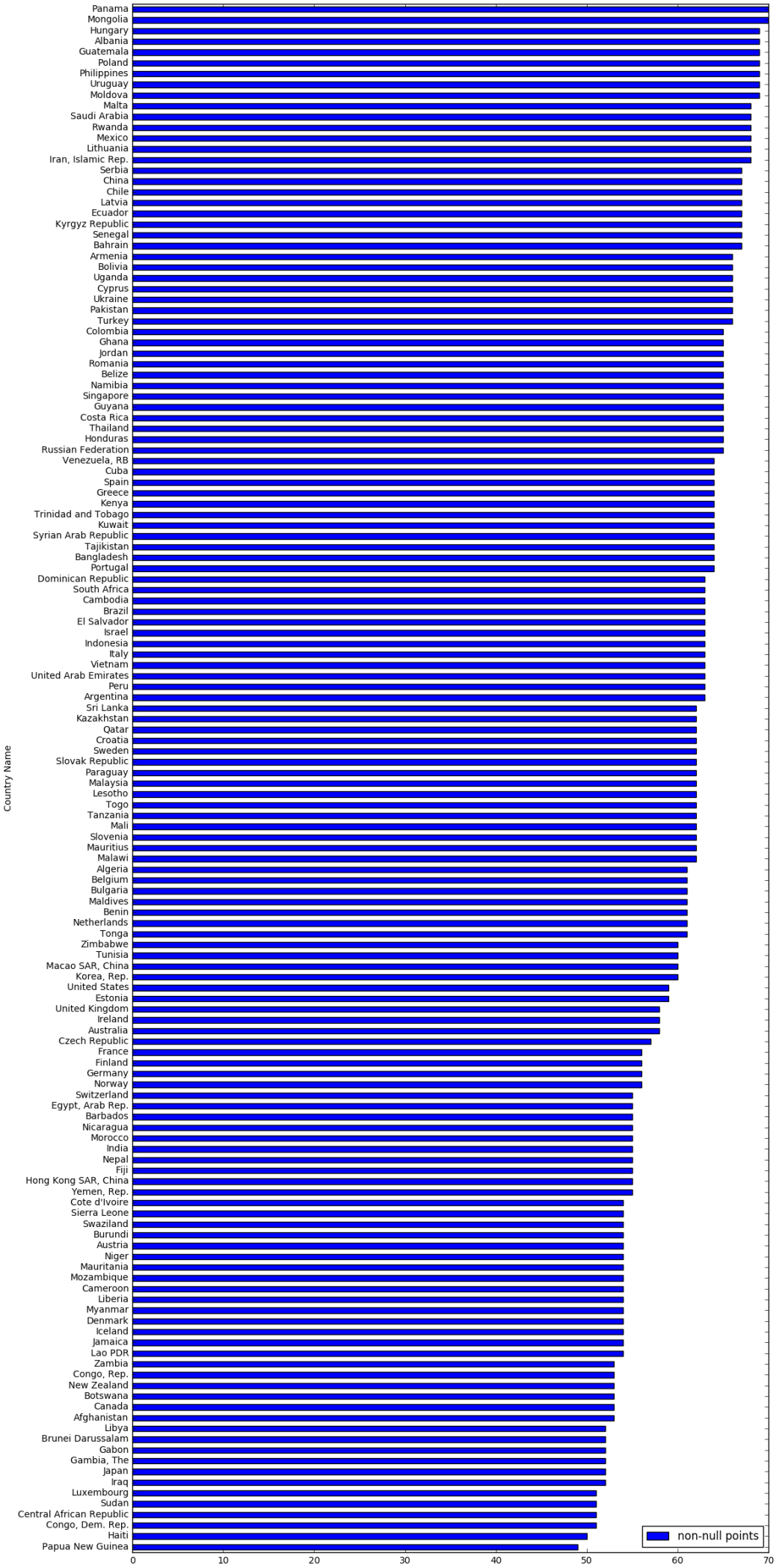
   We will need demographic and economic data for other countries as well as the US. For example, percent of people with access to electricity, and proportion of people living in slums could be useful for weeding out very dissimilar countries. Total tax rate might be good to know for cost optimization, and total labor force might be good to know for how many potential customers there could be. The percent of people over 25 with a bachelor's degree and the percent of people who have completed upper secondary education might also be good to know for marketing.

Step 2: Explore and Cleanup the Data

Answer these questions:

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.*

   It was reduced to 144 countries.


Country Name
Panama
Mongolia
Hungary
Albania
Guatemala
Poland
Philippines
Uruguay
Moldova
Malta
Saudi Arabia
Rwanda
Mexico
Lithuania
Iran, Islamic Rep.
Serbia
China
Chile
Latvia
Ecuador
Kyrgyz Republic
Senegal
Bahrain
Armenia
Bolivia
Uganda
Cyprus
Ukraine
Pakistan
Turkey
Colombia
Ghana
Jordan
Romania
Belize
Namibia
Singapore
Guyana
Costa Rica
Thailand
Honduras
Russian Federation
Venezuela, RB
Cuba
Spain
Greece
Kenya
Trinidad and Tobago
Kuwait
Syrian Arab Republic
Tajikistan
Bangladesh
Portugal
Dominican Republic
South Africa
Cambodia
Brazil
El Salvador
Israel
Indonesia
Italy
Vietnam
United Arab Emirates
Peru
Argentina
Sri Lanka
Kazakhstan
Qatar
Croatia
Sweden
Slovak Republic
Paraguay
Malaysia
Lesotho
Togo
Tanzania
Mali
Slovenia
Mauritius
Malawi
Algeria
Belgium
Bulgaria
Maldives
Benin
Netherlands
Tonga
Zimbabwe
Tunisia
Macao SAR, China
Korea, Rep.
United States
Estonia
United Kingdom
Ireland
Australia
Czech Republic
France
Finland
Germany
Norway
Switzerland
Egypt, Arab Rep.
Barbados
Nicaragua
Morocco
India
Nepal
Fiji
Hong Kong SAR, China
Yemen, Rep.
Cote d'Ivoire
Sierra Leone
Swaziland
Burundi
Austria
Niger
Mauritania
Mozambique
Cameroon
Liberia
Myanmar
Denmark
Iceland
Jamaica
Lao PDR
Zambia
Congo, Rep.
New Zealand
Botswana
Canada
Afghanistan
Libya
Brunei Darussalam
Gabon
Gambia, The
Japan
Iraq
Luxembourg
Sudan
Central African Republic
Congo, Dem. Rep.
Haiti
Papua New Guinea
0
10
20
30
40
50
60
70
non-null points

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

   Education_Avg Years, Education_Pct, and Education_Literacy

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 variables in 'Background' and 'Health' were removed. The 'background' variables included internet users, HIV, under-5 death rates, and things like number of physicians and total health expenditure. These things didn't seem relevant to a retail store's performance.

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.*

   I ended up using k-means (with PCA on the three categories specified in Step 2), because it had the combination of the most reasonable-looking results, as well as the highest silhoutte score out of reasonable-looking results.

   The silhoutte scores are as such:

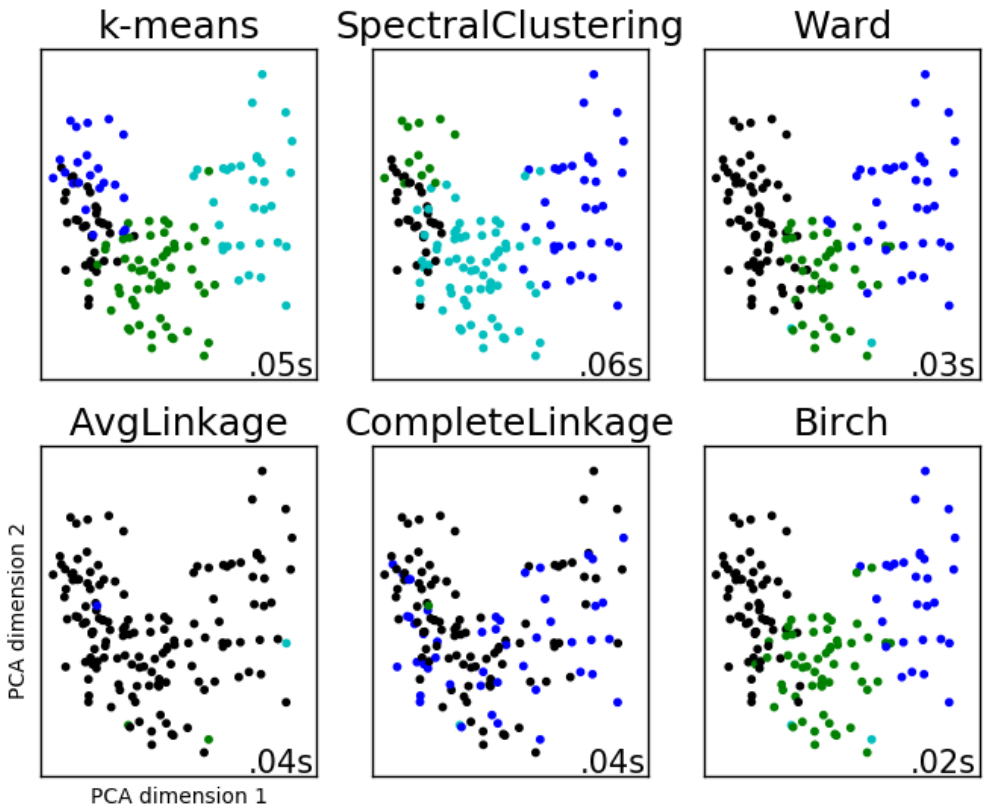
k-means	0.10
SpectralClustering	0.08
Ward	0.08
AvgLinkage	0.28
CompleteLinkage	0.01
Birch	0.10

   The birch algorithm had a similar silhoutte score to kmeans, but one cluster in the birch fit only had one point.

Step 4: Run the Data and Visualize

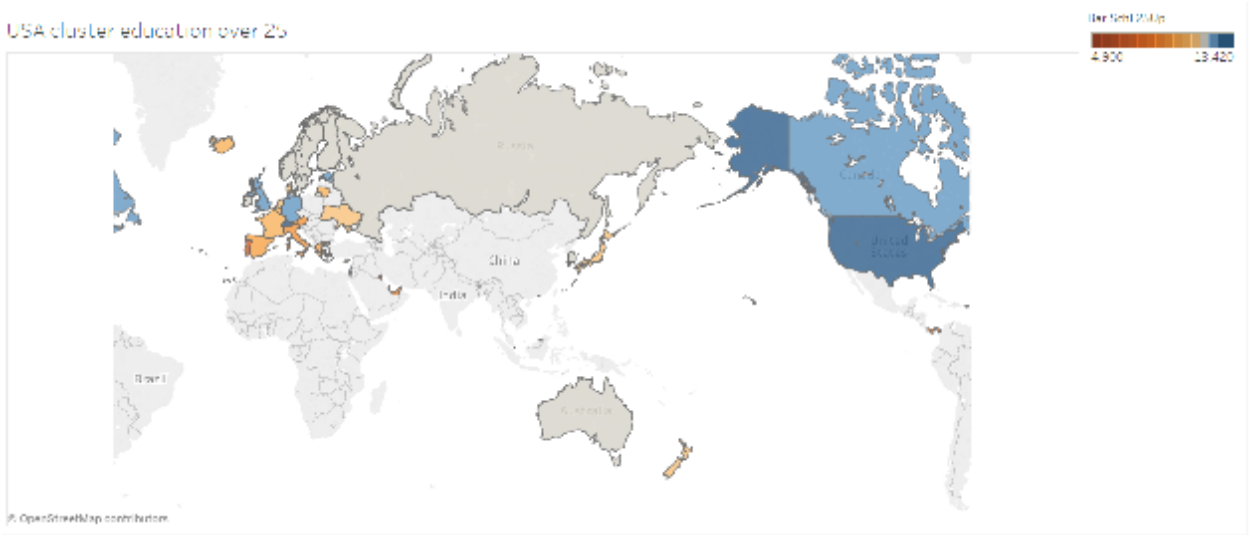
1. Do the clusters make sense?

   Yes. The results of the different clustering algorithms I tried look like this (the time in the corner is the time it took to fit to the data):



So k-means looks like it has a reasonable number of countries in each cluster, and they are decently separated in the first 2 PCA dimensions.

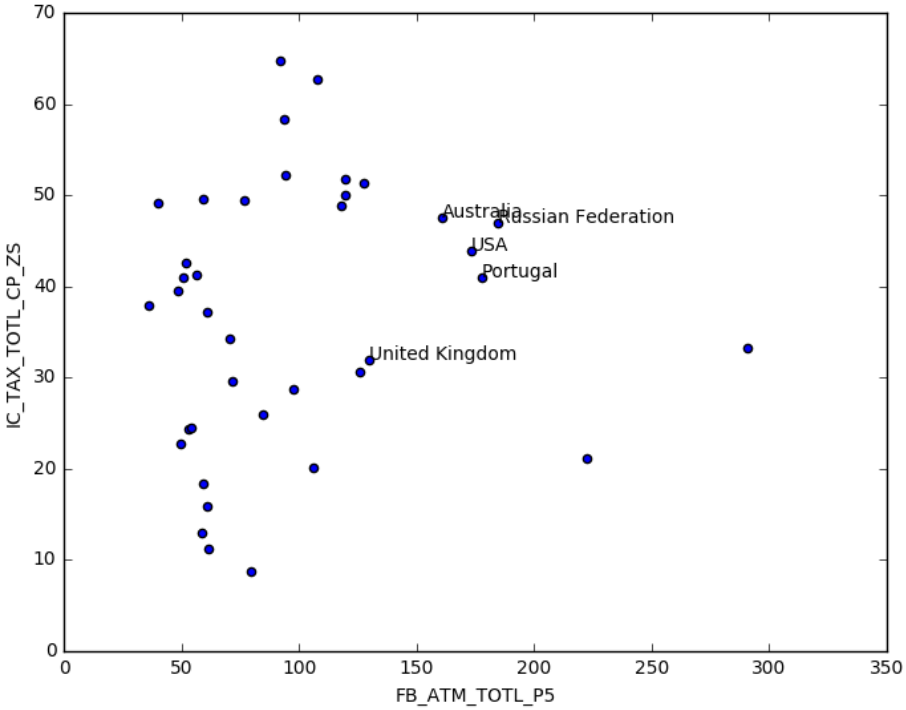
Here's a map of the average years of education for 25+ year-old people, also in the attached Tableau file:



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.

   Using Euclidean distances:

   1. Portugal
   2. Russian Federation
   3. Australia
   4. United Kingdom



Step 5: Recommendation

Recommended countries are:

Australia
Austria
Belgium
Brunei Darussalam
Canada
Cyprus
Denmark
Estonia
Finland

France
Germany
Greece
Hong Kong SAR, China
Iceland
Ireland
Israel
Italy
Japan
Korea, Rep.
Kuwait
Lithuania
Luxembourg
Malta
Netherlands
New Zealand
Norway
Panama
Portugal
Qatar
Russian Federation
Singapore
Spain
Sweden
Switzerland
Ukraine
United Arab Emirates
United Kingdom

Answer this question:

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

   I chose these countries because they were the most similar to the USA based on economic and demographic data. The similarity was calculated with Euclidean distance (shortest-path distance between points), and the countries were clustered using an algorithm called k-means clustering. These countries fell within the same cluster as the USA.