Kacper Ksieski

Project: International Expansion

Complete each section. When you are ready, save your file as a PDF document and submit it here: https://classroom.udacity.com/nanodegrees/nd008/parts/91294931-aacb-4887-856f-fd19fe915795/project#

Step 1: Key Decisions

Briefly explain the key decisions and the type of data that you need to conduct this analysis (250 word limit).

Key Decisions:

Answer these three questions

1. What decisions needs to be made?
   The retail store chain desires to expand into other countries. The decision to be made is into which countries it should expand. This decision rests on determining which, if any, countries are similar to the United States as this indicates a higher probability of success.

2. What data is needed to inform those decisions? Please include 2 examples in each of the following categories: Economic, Environment, Education
   The company needs to know which countries are similar to the United States in terms of economics, demographics, education, and environment.
   In the Economic category, two examples are:
   SL_EMP_TOTL_SP_ZS, Employment to population ratio is the proportion of a country's population that is employed.
   SL_TLF_TOTL_IN, total labor force comprises people ages 15 and older who meet the International Labour Organization definition of the economically active population: all people who supply labor for the production of goods and services during a specified period.
   In the Environment category, two examples are:
   EN_POP_SLUM_UR_ZS, Population living in slums is the proportion of the urban population living in slum households
   EG_ELC_ACCS_ZS, access to electricity is the percentage of population with access to electricity.
   In the Education category, two examples are:
   UIS_EA_3_AG25T99, the percentage of population (age 25 and over) with completed upper secondary education (ISCED 3) as the highest level of educational attainment
   SE_ADT_LITR_ZS, percentage of the population age 15 and above who can, with understanding, read and write a short, simple statement on their everyday life.

Kacper Ksieski

Step 2: Explore and Cleanup the Data

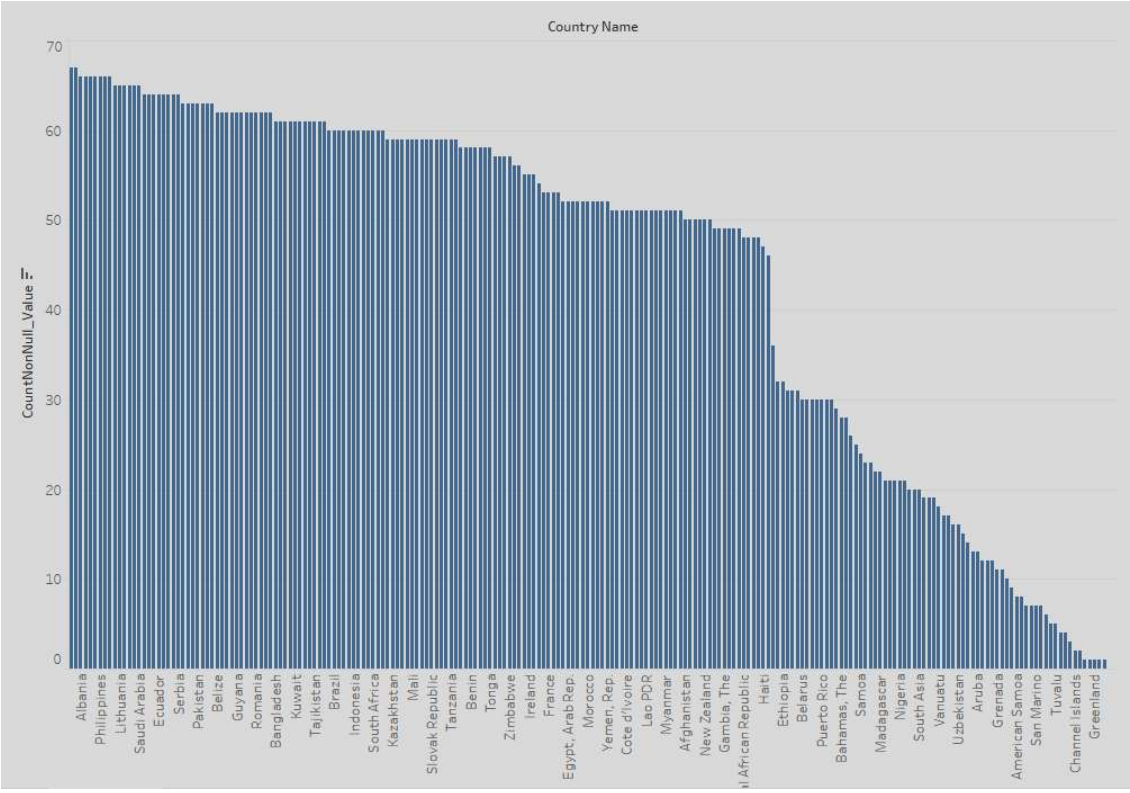
Explore and cleanup your dataset. Data is provided in a CSV file for 215 countries with 77 variables (250 word limit)

Here are some guidelines to help you cleanup your data:

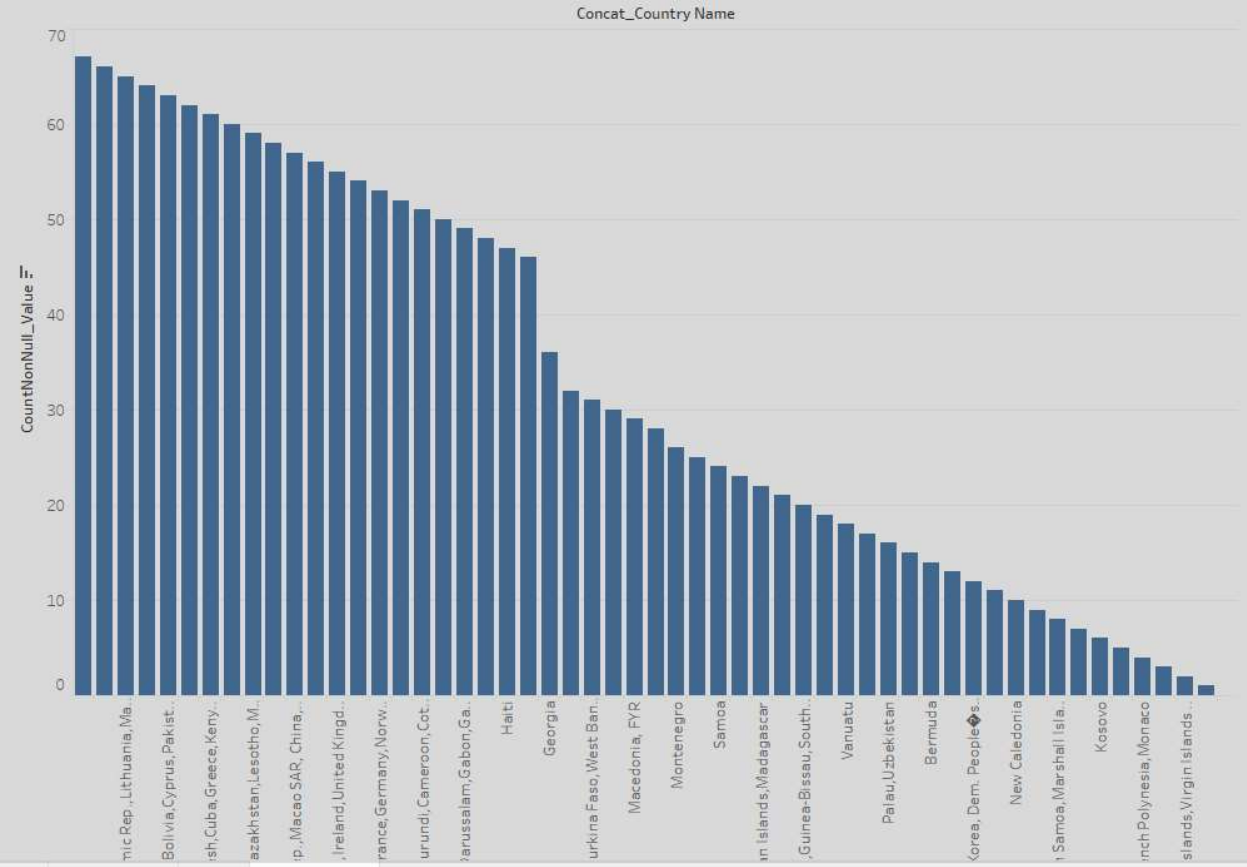
1. Country records where most of the variables missing might not be appropriate to be included in the analysis. The lack of accurate reporting could indicate that these countries are probably not similar to the United States. You should remove any country with fewer than 25 missing data points. HINT: You should be left with 144 countries.
2. Some variables are closely related and may be candidates for variable reduction through Principal Components Analysis.
3. Some variables seem irrelevant for the given analysis involving economy, demographics, education, and environment. Which variables seem irrelevant?

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.*
   I reduced my dataset by 71 countries for a remainder of 144 countries. Depicted below is a bar chart of number of non-null data points by country, sorted from most to least. There are too many data points to display correctly so I made a second visualisation with countries grouped by number of non-null data points.



Kacper Ksieski



2. *Which data categories will be used for Principal Components Analysis (PCA)? There should be three categories that are targeted for PCA.*
   The three data categories to be used for PCA are Education_Avg Years, Education_Pct, and Education_literacy. Any other categories have too few variables in which to conduct the PCA. These other variables will be directly input into the cluster analysis.
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.*
   I decided that any variables not under the categories of economy, environment, demography, and education should be removed. The irrelevant variables fall under the categories of health and background. They are as follows:
   1. IT_NET_USER_P2
   2. SG_VAW_BURN_ZS
   3. SH_DYN_AIDS_ZS
   4. SH_DYN_MORT
   5. SH_MED_PHYS_ZS
   6. SH_TBS_PREV
   7. SH_XPD_PCAP
   8. SN_ITK_DEFC_ZS
   9. SP_POP_DPND

Kacper Ksieski

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 compared the interquartile spread, the min max spread, and median of both the Adjusted Rand Indices and the Calinski-Harabasz Indices of all three methods. The Neural Gas method had either the highest or close to the highest value in both indices for the median, the interquartile spread, and the spread between the minimum and maximum values. This means the Neural Gas method produces the most stable, most distinct, and most compact clusters. I chose the Neural Gas clustering method.

Method	Clusters	Index	Minimum	1st_Quartile	Mean	Median	3rd_Quartile	Maximum	Interquartile_Spread	Min_Max_Spread
K-Means	4	AR	0.4107	0.7898	0.8412	0.9052	0.9337	1	0.1439	0.5893
K-Means	4	CH	11.42	346.7	419.2	354.5	612.9	683.2	266.2	671.78
K-Medians	4	AR	0.2296	0.5325	0.6508	0.6614	0.7894	0.9179	0.2569	0.6883
K-Medians	4	CH	5.103	14.75	108.6	19.74	36.92	980.7	22.17	975.597
Neural Gas	4	AR	0.4891	0.8141	0.8477	0.9022	0.9337	1	0.1196	0.5109
Neural Gas	4	CH	11.42	344.2	422.5	408.9	612.1	683.4	267.9	671.98

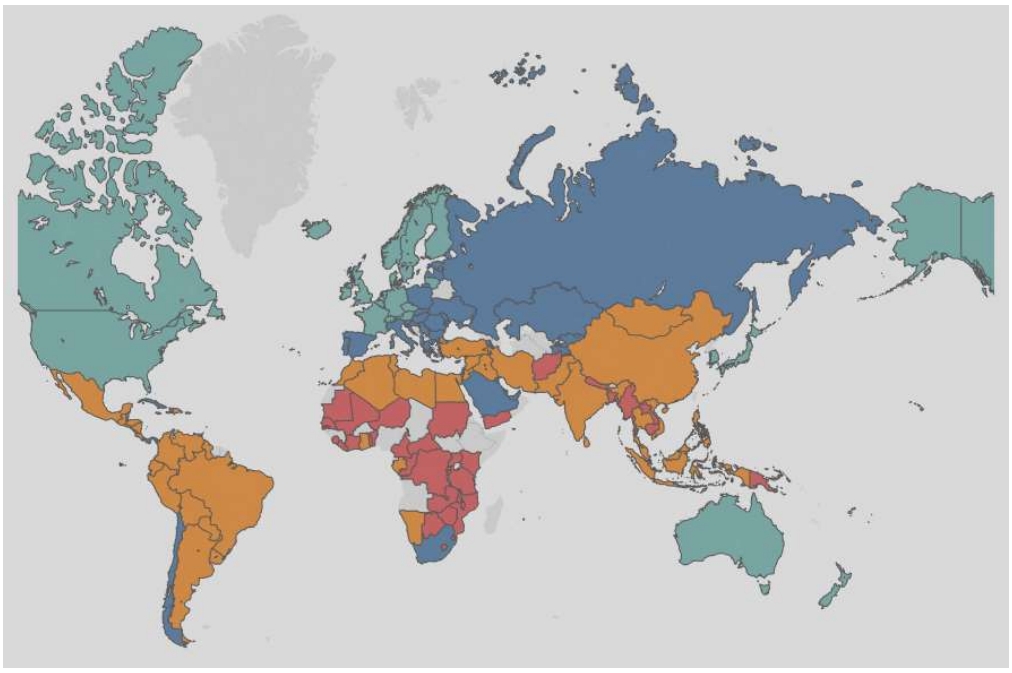
Step 4: Run the Data and Visualize

Run the data through your clustering algorithm and visualize the clusters. (250 words limit)

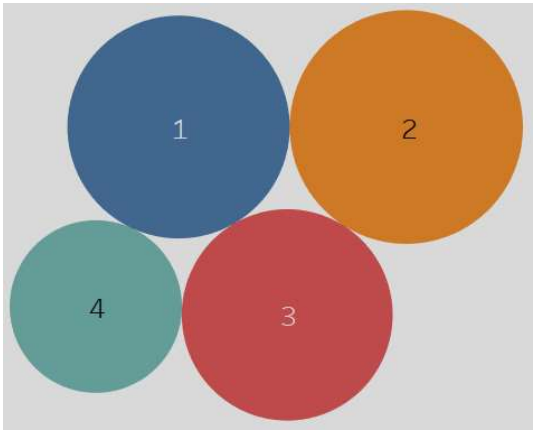
Include at least 2 visualizations to show the clusters that you came up with. At least one of you visualizations should be a Tableau map.

Answer this question.

1. Do the clusters make sense?
   Yes, these clusters make sense. As depicted in the map below, USA is clustered with many other countries in the OECD. Several of the equatorial countries are clustered together. Nearly the entire African continent is clustered. Lastly, the Asian continent along with many of the Eastern European countries are clustered together.



As depicted below, the chart depicts clusters of similar sizes. No one cluster is either too large or too small.



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.

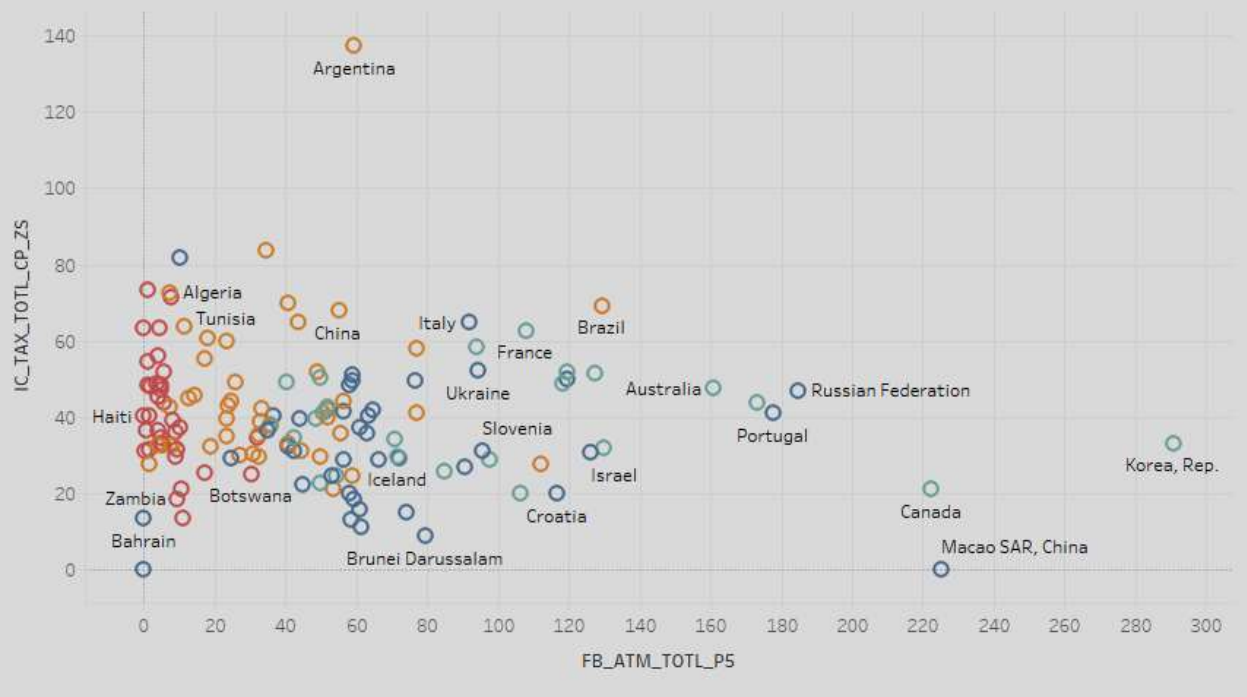
   The four countries in USA's cluster that are closest to the USA in terms of Total Tax Rate by ATM Machines are:

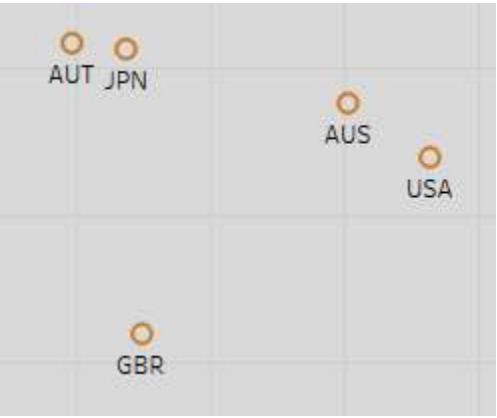
   1. Australia
   2. United Kingdom
   3. Japan
   4. Austria

   This list was determined not visually, but by the Euclidean distance formula, which I employed in Alteryx.

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}.$$

Kacper Ksieski





Kacper Ksieski

Step 5: Recommendation

Provide your recommended list of countries and justify your recommendation using data from your analysis (250 words limit)

Please list out the country codes in this section here with this format in alphabetical order.

……...

Australia
Belgium
Canada
……...

AUS
AUT
BEL
BRB
CAN
CHE
CZE
DEU
DNK
FIN
FRA
GBR
HKG
IRL
ISL
JPN
KOR
LTU
LUX
NLD
NOR
NZL
SWE

Answer this question:

1. *Why did you decide to choose these countries?*
   I chose these countries because they are in the same cluster as the USA. According to the cluster analysis, they are the most similar to the USA in terms of economics, education, and environment. Therefore, they are the most likely to be profitable locations for expansion.

Kacper Ksieski

Before you Submit

Please check your answers against the requirements of the project dictated by the rubric here. Reviewers will use this rubric to grade your project.