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 decision is to choose which countries are the most similar to United States in term of economic, demographic, education and environment segment.

We need to conduct clustering analysis on these segments to find list of countries which has similar demographic, economy, education and environment, and decided to choose which country to expand to for the retail chain store

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

Data examples:

Category	**Variable**	**Definition**
Education	UIS_EA_6T8_AG25T99	The percentage of population (age 25 and over) with a completed bachelor's or equivalent degree (ISCED 6) or higher.
Education	UIS_EA_7_AG25T99	The percentage of population (age 25 and over) with a completed master's or equivalent degree (ISCED 7) degree as the highest level of educational attainment.
Economic	IC_FRM_ISOC_ZS	Internationally-recognized quality certification is the percentage of firms having an internationally-recognized quality certification, i.e.,

		International Organization for Standardization (ISO) 9000, 9002 or 1400
Economic	IC_TAX_TOTL_CP_ZS	Total tax rate measures the amount of taxes and mandatory contributions payable by businesses after accounting for allowable deductions and exemptions as a share of commercial profits.
Environment:	EN_POP_SLUM_UR_ZS	Population living in slums is the proportion of the urban population living in slum households. A slum household is defined as a group of individuals living under the same roof lacking one or more of the following conditions: access to improved water, access to improved sanitation, sufficient living area, and durability of housing.
Environment:	EG_ELC_ACCS_ZS	Access to electricity is the percentage of population with access to electricity

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?

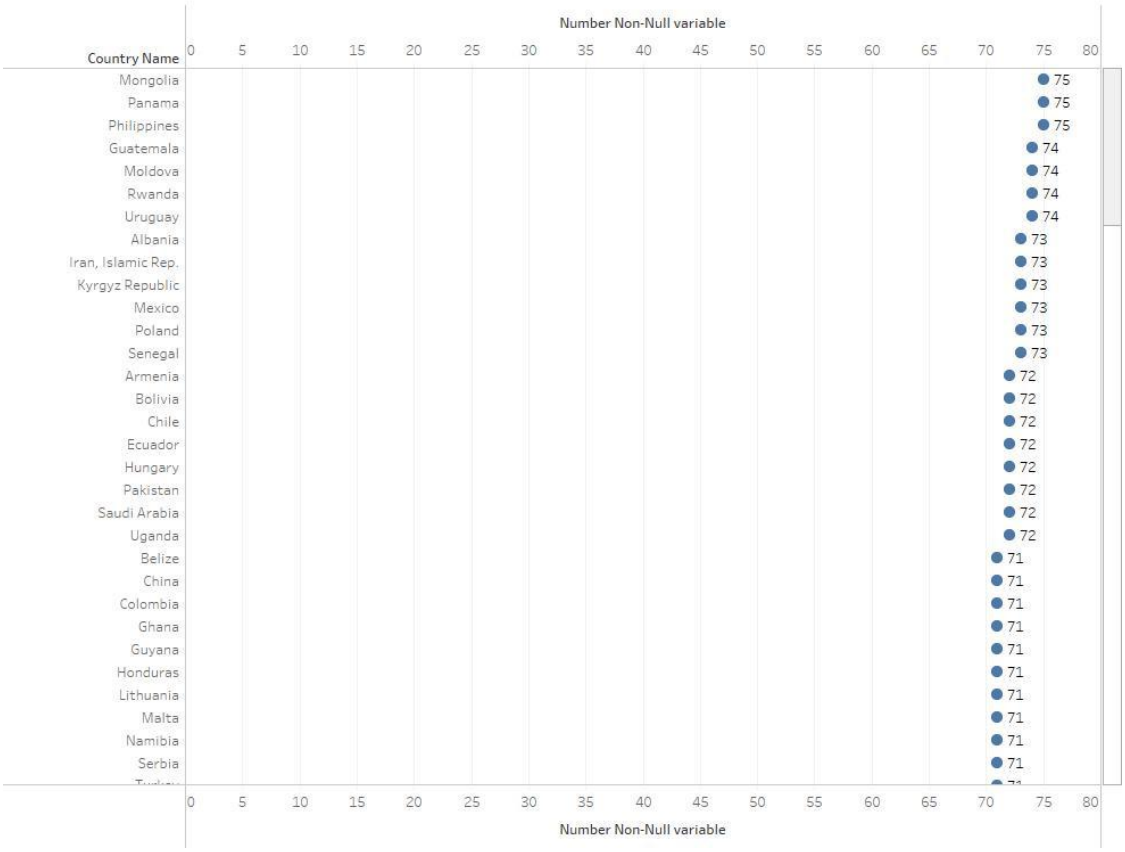
Answer these questions:

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

Answer

1. The countries are reduced to 144 countries, which have null variable less than 25 in the dataset.

144 countries list



There are 9 irrelevant variables that need to be removed since these variables are not under the economic, education, demographic, or environment categories.

No	Variable Name	Category	Definition
1	IT_NET_USER_P2	Background	Internet users are individuals who have used the Internet (from any location) in the last 12 months
2	SH_DYN_AIDS_ZS	Background	Prevalence of HIV refers to the percentage of people ages 15-49 who are infected with HIV.
3	SH_DYN_MORT	Background	Under-five mortality rate is the probability per 1,000 that a newborn baby will die before reaching age five, if subject to age-specific mortality rates of the specified year.
4	SH_MED_PHYS_ZS	Health	Physicians include generalist and specialist medical

			practitioners.
5	SH_XPD_PCAP	Health	Total health expenditure is the sum of public and private health expenditures as a ratio of total population.
6	SN_ITK_DEFC_ZS	Health	Population below minimum level of dietary energy consumption (also referred to as prevalence of undernourishment) shows the percentage of the population whose food intake is insufficient to meet dietary energy requirements continuously. Data showing as 2.5 signifies a prevalence of undernourishment below 2.5%.
7	SP_POP_DPND	Health	Age dependency ratio is the ratio of dependents--people younger than 15 or older than 64--to the working-age population--those ages 15-64
8	SG_VAW_BURN_ZS	Health	Percentage of women ages 15-49 who believe a husband/partner is justified in hitting or beating his wife/partner when she burns the food.
9	SH_TBS_PREV	Health	Prevalence of tuberculosis is the estimated number of TB cases (all forms) at a given point in time, expressed as the rate per 100,000 population

Data Categories for PCA:

- **Education Average Years**
- **Education Percentage**
- **Education Literacy**

Reducing Variables

I. Education Variables

Education Avg Years (30) variables	Education Pct (15)	Education Literacy (7)	Education_ PTR (3)
EBAR_SCHL_1519	UIS_EA_1_AG25T99	SE_ADT_1524_LT_FE_ZS	UIS_PTRHC_2
BAR_SCHL_1519_FE	UIS_EA_1T6_AG25T99	SE_ADT_1524_LT_FM_ZS	UIS_PTRHC_3
BAR_SCHL_15UP	UIS_EA_2_AG25T99	SE_ADT_1524_LT_MA_ZS	UIS_PTRHC_56
BAR_SCHL_15UP_FE	UIS_EA_2T6_AG25T99	SE_ADT_1524_LT_ZS	
BAR_SCHL_2024	UIS_EA_3_AG25T99	SE_ADT_LITR_FE_ZS	
BAR_SCHL_2024_FE	UIS_EA_3T6_AG25T99	SE_ADT_LITR_MA_ZS	
BAR_SCHL_2529	UIS_EA_4_AG25T99	SE_ADT_LITR_ZS	
BAR_SCHL_2529_FE	UIS_EA_4T6_AG25T99		
BAR_SCHL_25UP	UIS_EA_5_AG25T99		
BAR_SCHL_25UP_FE	UIS_EA_5T8_AG25T99		
BAR_SCHL_3034	UIS_EA_6_AG25T99		
BAR_SCHL_3034_FE	UIS_EA_6T8_AG25T99		
BAR_SCHL_3539	UIS_EA_7_AG25T99		
BAR_SCHL_3539_FE	UIS_EA_7T8_AG25T99		
BAR_SCHL_4044	UIS_EA_8_AG25T99		
BAR_SCHL_4044_FE			
BAR_SCHL_4549			
BAR_SCHL_4549_FE			
BAR_SCHL_5054			
BAR_SCHL_5054_FE			
BAR_SCHL_5559			
BAR_SCHL_5559_FE			
BAR_SCHL_6064			
BAR_SCHL_6064_FE			
BAR_SCHL_6569			
BAR_SCHL_6569_FE			
BAR_SCHL_7074			
BAR_SCHL_7074_FE			
BAR_SCHL_75UP			
BAR_SCHL_75UP_FE			

Total Education Variable : 55

II. Economy Variables

Economy (3)	Economy (2)	Economy (5)
IQ_WEF_PORT_XQ	SE_XPD_TOTL_GD_ZS	IC_ELC_TIME
SL_TLF_SECO_ZS	FB_ATM_TOTL_P5	IC_FRM_ISOC_ZS
SL_TLF_TOTL_IN		IC_TAX_TOTL_CP_ZS
		TM_TAX_MANF_SM_FN_ZS
		SL_EMP_TOTL_SP_ZS

Total economy variables : 10

III. Environment (2)

EN_POP_SLUM_UR_ZS
EG_ELC_ACCS_ZS

Total Environment variable : 2

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.*

The best clustering method to use is Neural Gas. It can be seen when comparing the three available models on Adjusted Rand and Calinski-Harabasz Indices that Neural Gas performs better with a higher median and mean

Record | Report

K-Means Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7
Minimum	0.3351	0.3509	0.5757	0.5407	0.5135	0.5069
1st Quartile	0.6982	0.4361	0.8299	0.6388	0.6123	0.5757
Median	0.8137	0.7125	0.9217	0.7583	0.6793	0.6397
Mean	0.7916	0.6793	0.8873	0.7519	0.6893	0.6455
3rd Quartile	0.9381	0.882	0.9741	0.8561	0.7502	0.694
Maximum	1	1	1	1	0.9409	0.8974

Calinski-Harabasz Indices:

	2	3	4	5	6	7
Minimum	155.8	165.7	189	173.7	150	143.5
1st Quartile	178.1	178.7	223.6	196	171.2	154.1
Median	184.3	189	226.8	202.1	176.4	157
Mean	181.8	187	224.5	199.9	175.2	157.3
3rd Quartile	187.3	195.1	229.2	206.4	180.9	161.1
Maximum	189.7	198.7	232	209.9	187.7	168.7

Plots

Record | Report

K-Medians Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7
Minimum	0.2424	0.3553	0.4791	0.5532	0.4824	0.4287
1st Quartile	0.6298	0.4133	0.8146	0.65	0.6034	0.5728
Median	0.8136	0.5878	0.8909	0.7522	0.6741	0.6161
Mean	0.7703	0.6465	0.8506	0.7525	0.6734	0.6215
3rd Quartile	0.9447	0.917	0.9383	0.8368	0.7394	0.6688
Maximum	1	1	1	0.9657	0.8428	0.8572

Calinski-Harabasz Indices:

	2	3	4	5	6	7
Minimum	103.5	147.8	102.1	160.4	126.6	127.9
1st Quartile	155.1	164.6	217.4	187.6	160.5	141.8
Median	163.4	181.6	225.3	193.8	164.5	146.8
Mean	162.7	178.2	218.3	192.4	164.9	146.2
3rd Quartile	171.3	192.7	229.5	199.3	170.5	151
Maximum	188.5	199.5	237.3	207	181.1	159.9

Neural Gas Cluster Assessment Report

Summary Statistics

Adjusted Rand Indices:

	2	3	4	5	6	7
Minimum	0.4981	0.3603	0.5755	0.6001	0.5434	0.5547
1st Quartile	0.7218	0.4836	0.919	0.7074	0.6436	0.6477
Median	0.8521	0.6377	0.9586	0.8518	0.7241	0.7061
Mean	0.8232	0.6495	0.927	0.8156	0.744	0.7075
3rd Quartile	0.9722	0.8677	0.9796	0.9154	0.8603	0.7692
Maximum	1	0.9775	1	0.9727	0.9484	0.9113

Calinski-Harabasz Indices:

	2	3	4	5	6	7
Minimum	156.1	165.5	147.6	182.5	164.5	144.8
1st Quartile	179.4	177.5	224.4	199.9	178.7	159.3
Median	185.8	179.4	226.9	204.7	181.1	161.7
Mean	182.4	181.5	224.9	203.1	180.7	161
3rd Quartile	187.6	182.1	229.6	207.4	182.9	163.8
Maximum	189.7	197.5	231.9	211.5	188.1	168.9

Neural Gas Cluster method has the overall higher Mean and median for Adjusted Rand indices and Calinski-Harabasz Indices, therefore we will use Neural Gas cluster method for clustering analysis

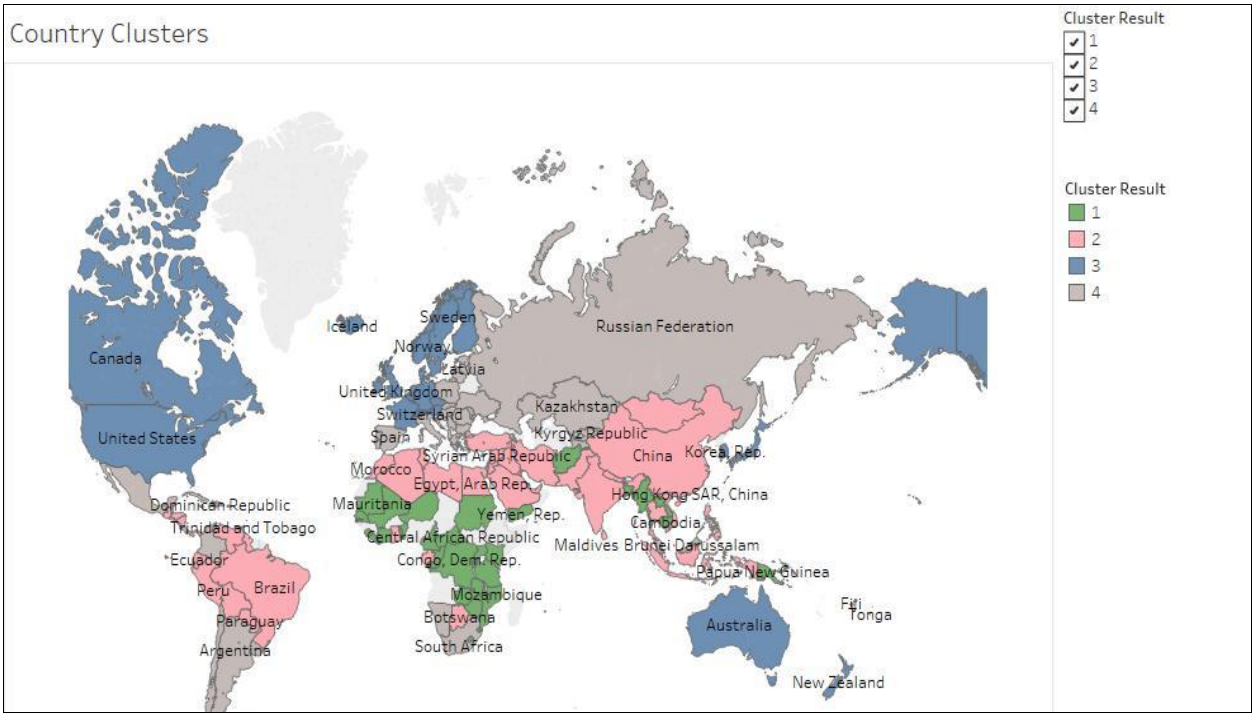
Step 4: Run the Data and Visualize

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.

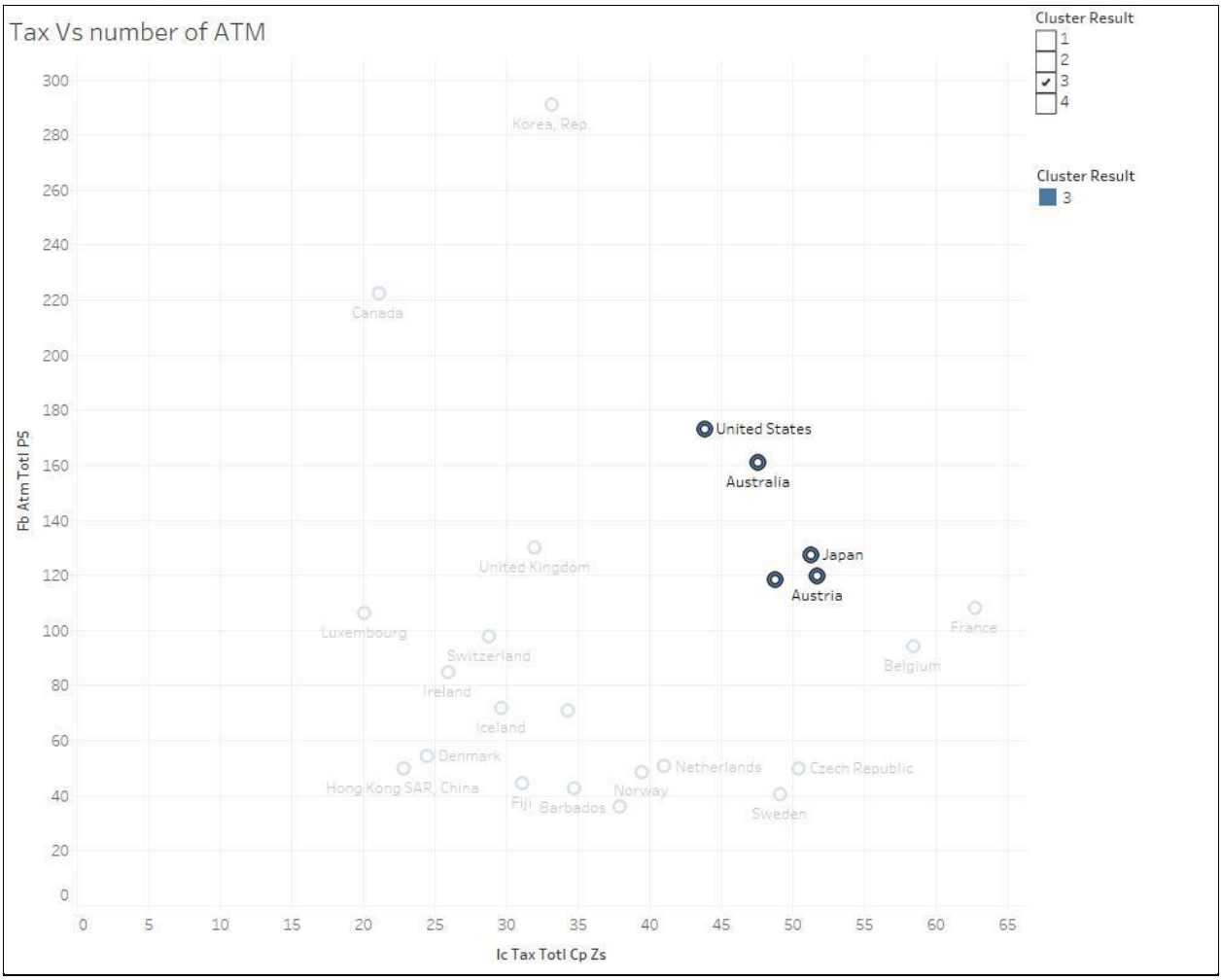
Answer this question.

1. Do the clusters make sense?
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 clustering makes sense if we take a look in education , economy & environment perspective, whereas Cluster 3 along with the United States are more advanced countries and the most established financially such as Australia, Switzerland, Canada,etc. The countries that falls withing Cluster 2 would be emerging countries with high growth such as China, Brazil, Malaysia, Indonesia. Cluster 1 would be more in the lower level of economy & education compared by its peers such as South Aftica, Argentina, Khazakhstan etc. The 4th cluster would be the lowest area consisting most of countries in African continents.

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



4 countries in the same cluster withh United States that has the closest Tax Total rate :

1. Australia
2. Japan
3. Austria
4. Germany

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
Austria
Barbados
Belgium
Canada
Czech Republic
Denmark
Finland
France
Germany
Hong Kong SAR, China
Iceland
Ireland
Japan
Korea, Rep.
Luxembourg
Netherlands
New Zealand
Norway
Sweden
Switzerland
United Kingdom
United States

Answer this question:

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

Cluster Information:

Cluster	Size	Ave Distance	Max Distance	Separation
1	34	1.794894	4.531776	2.087038
2	44	1.876731	7.377666	1.572583
3	24	1.792589	3.600104	2.740278
4	42	1.870649	3.490937	1.325145

No convergence after 201 iterations.
Sum of within cluster distances: 265.19192.

	Edu.PC1	Edu.PC2	Edu.PC3	Econ.PC1	Econ.PC2	Econ.PC3	Env.PC1
1	-1.275171	0.502896	-0.222126	-1.162596	0.471882	0.158352	-1.57418
2	-0.338404	-0.505869	0.241057	-0.248352	0.002986	0.215973	0.126775
3	1.230119	1.738929	0.268525	1.195796	0.035094	0.108569	0.773273
4	0.68029	-0.871058	-0.216919	0.515845	-0.409853	-0.418366	0.692701

	Env.PC2
1	-0.513196
2	0.909701
3	-0.460224
4	-0.289806

Plots

United States falls within Cluster 3 categories, which from Edu PC1,Econ PC1, and Env PC1 suggest that United states is in clusters of countries which has the highest level of education (1.23), economy (1.19) and environment component (0.77). These countries that I selected also within the clusters of 3, that's why I recommend above listed countries

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.