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#](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

1. 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:

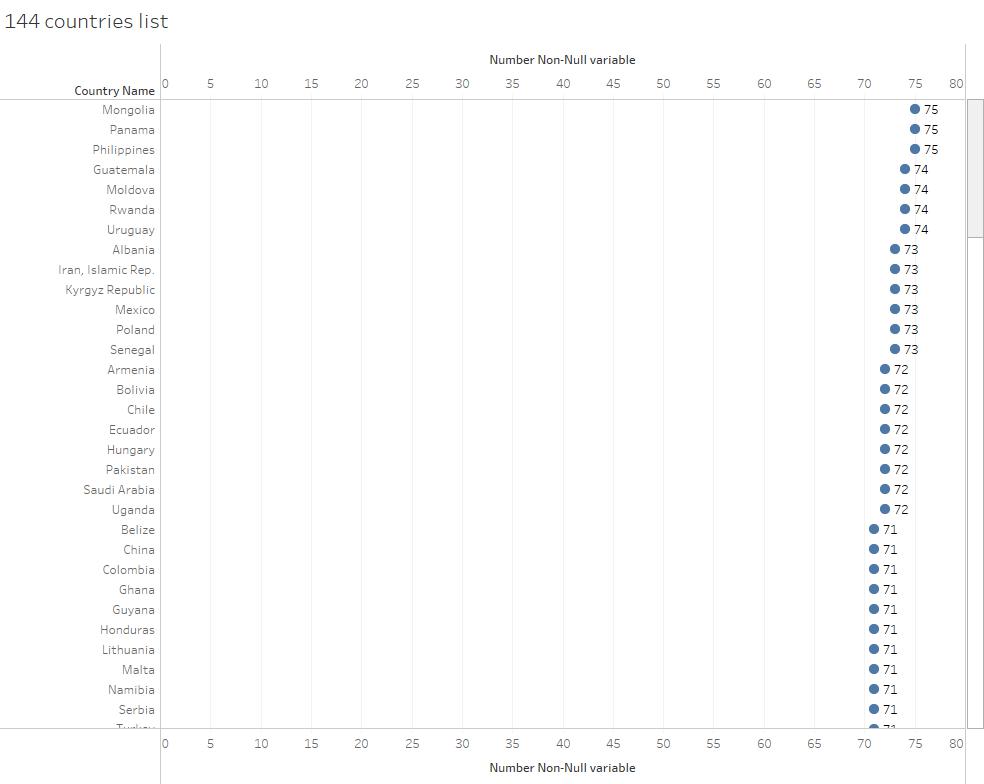
1. Country records where most of the variables missing might not be appropriate to be included in the analysis. The lack of accurate reporting could indicate that these countries are probably not similar to the United States. You should remove any country with fewer than 25 missing data points. HINT: 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.



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**

1. **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   1. **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 |

1. **Environment (2)**

|  |
| --- |
| EN\_POP\_SLUM\_UR\_ZS |
| EG\_ELC\_ACCS\_ZS |

Total Environment variable : 2

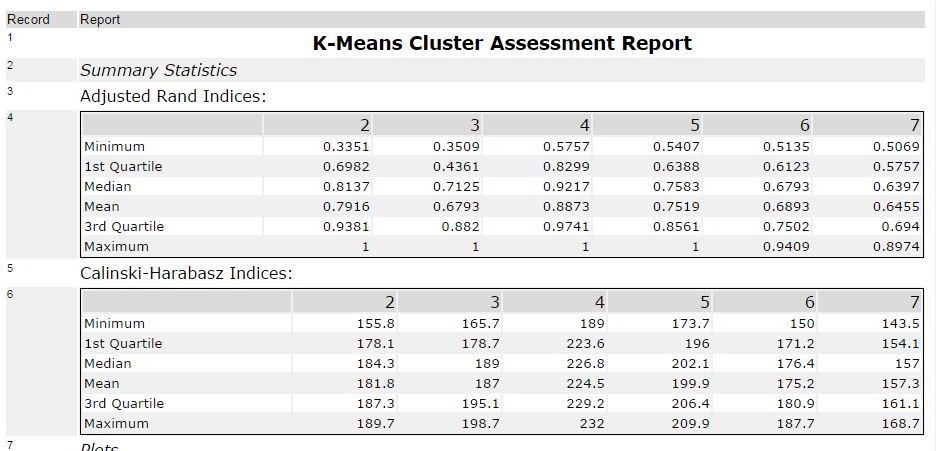
Step 3: Determine Clusters and Methodology

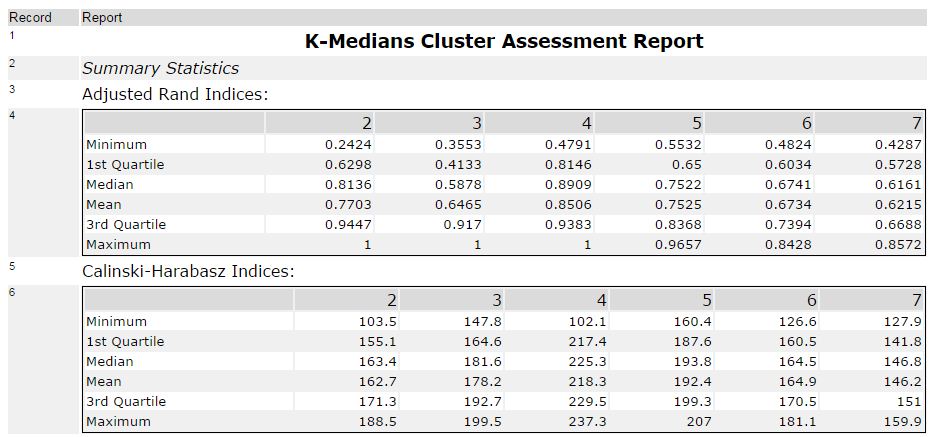
*Determine the optimal clustering method and create four clusters. (100 word limit)*

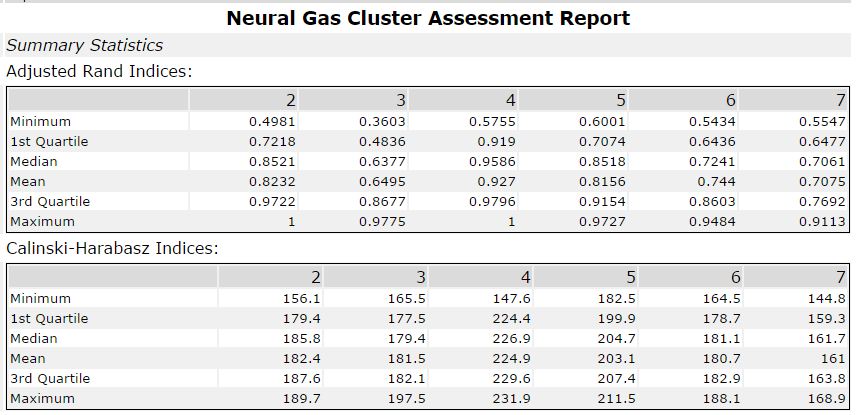
*Answer this question:*

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



****

****

**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

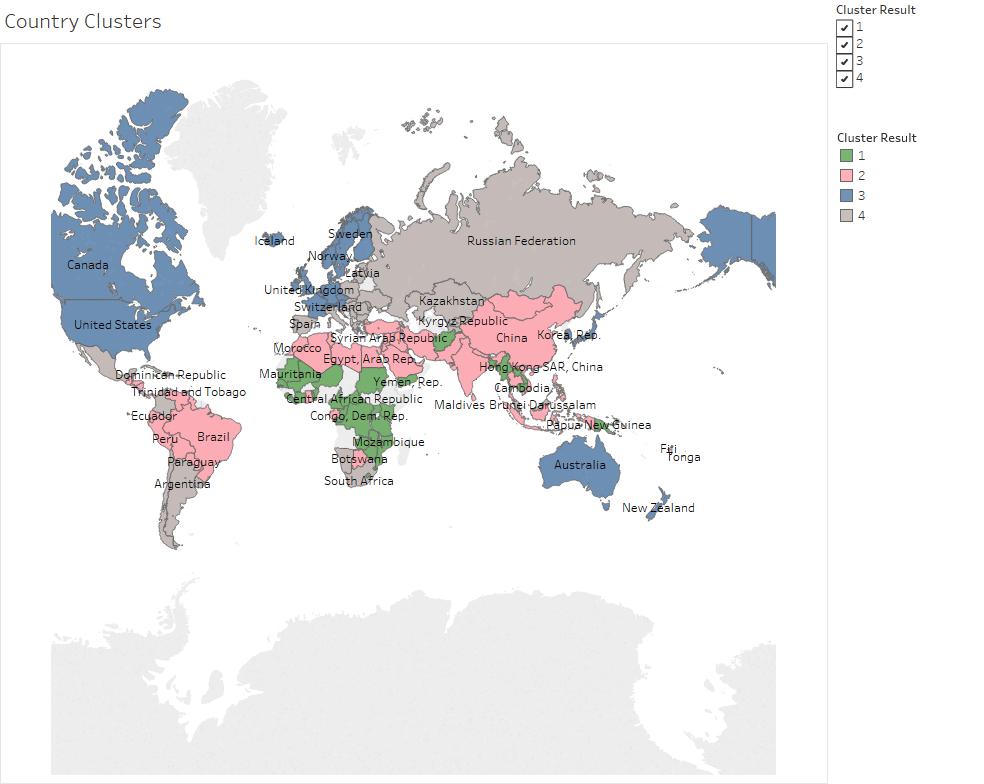
## Step 4: Run the Data and Visualize

*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.*

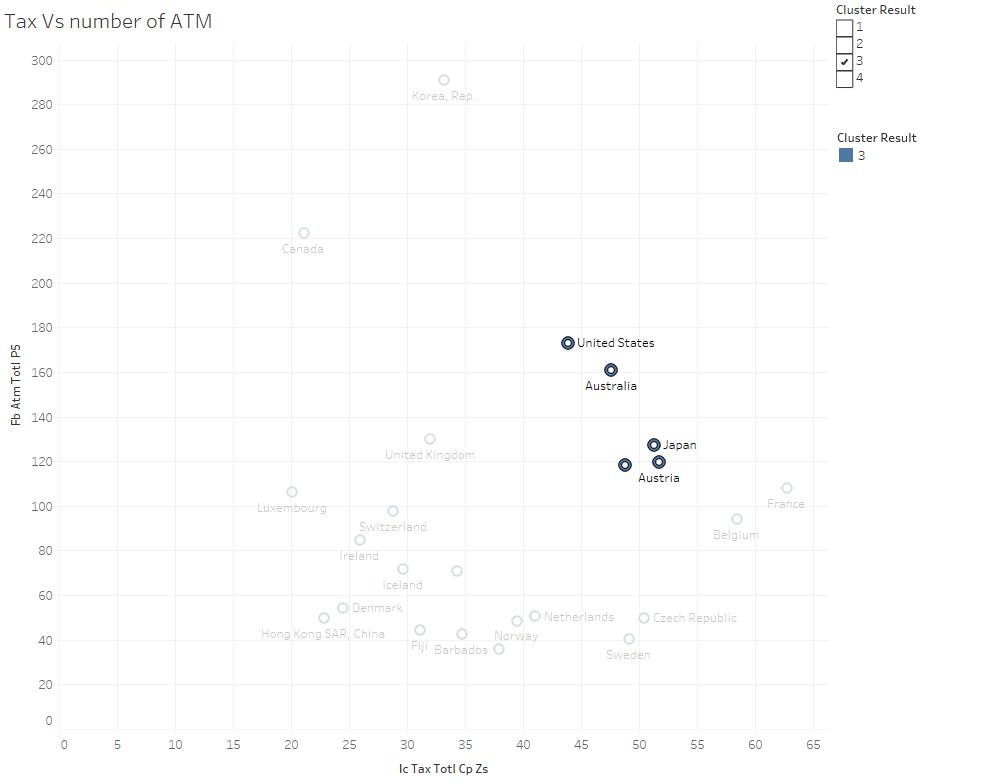
*Answer this question.*

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

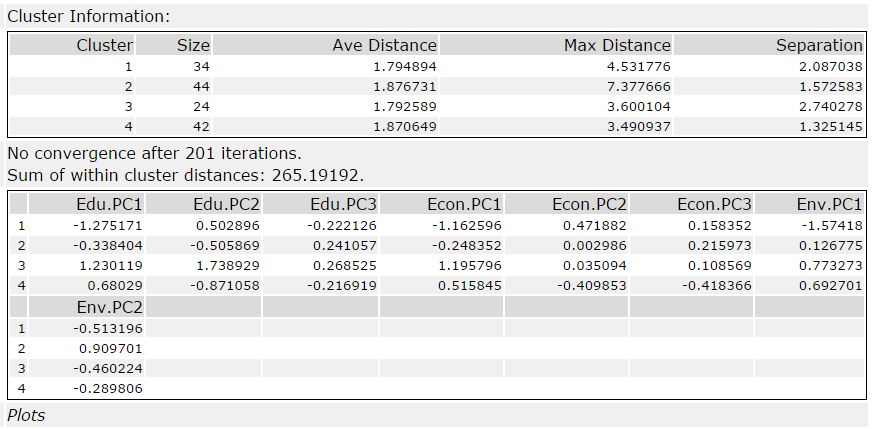
*Provide your recommended list of countries and justify your recommendation using data from your analysis (250 words limit)*

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

|  |  |  |
| --- | --- | --- |
| *No* | *Country* | *Country Code* |
| *1* | *Australia* | AUS |
| *2* | *Austria* | AUT |
| *3* | *Barbados* | BRB |
| *4* | *Belgium* | BEL |
| *5* | *Canada* | CAN |
| *6* | *Czech Republic* | CZE |
| *7* | *Denmark* | DNK |
| *8* | *Finland* | FIN |
| *9* | *France* | FRA |
| *10* | *Germany* | DEU |
| *11* | *Hong Kong SAR, China* | HKG |
| *12* | *Iceland* | ISL |
| *13* | *Ireland* | IRL |
| *14* | *Italy* | ITA |
| *15* | *Japan* | JPN |
| *16* | *Korea, Rep.* | KOR |
| *17* | *Luxembourg* | LUX |
| *18* | *Netherlands* | NLD |
| *19* | *New Zealand* | NZL |
| *20* | *Norway* | NOR |
| *21* | *Sweden* | SWE |
| *22* | *Switzerland* | CHE |
| *23* | *United Kingdom* | GBR |
| *24* | *United States* | USA |

*Answer this question:*

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



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](https://review.udacity.com/#!/rubrics/424/view) here. Reviewers will use this rubric to grade your project.