# 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?

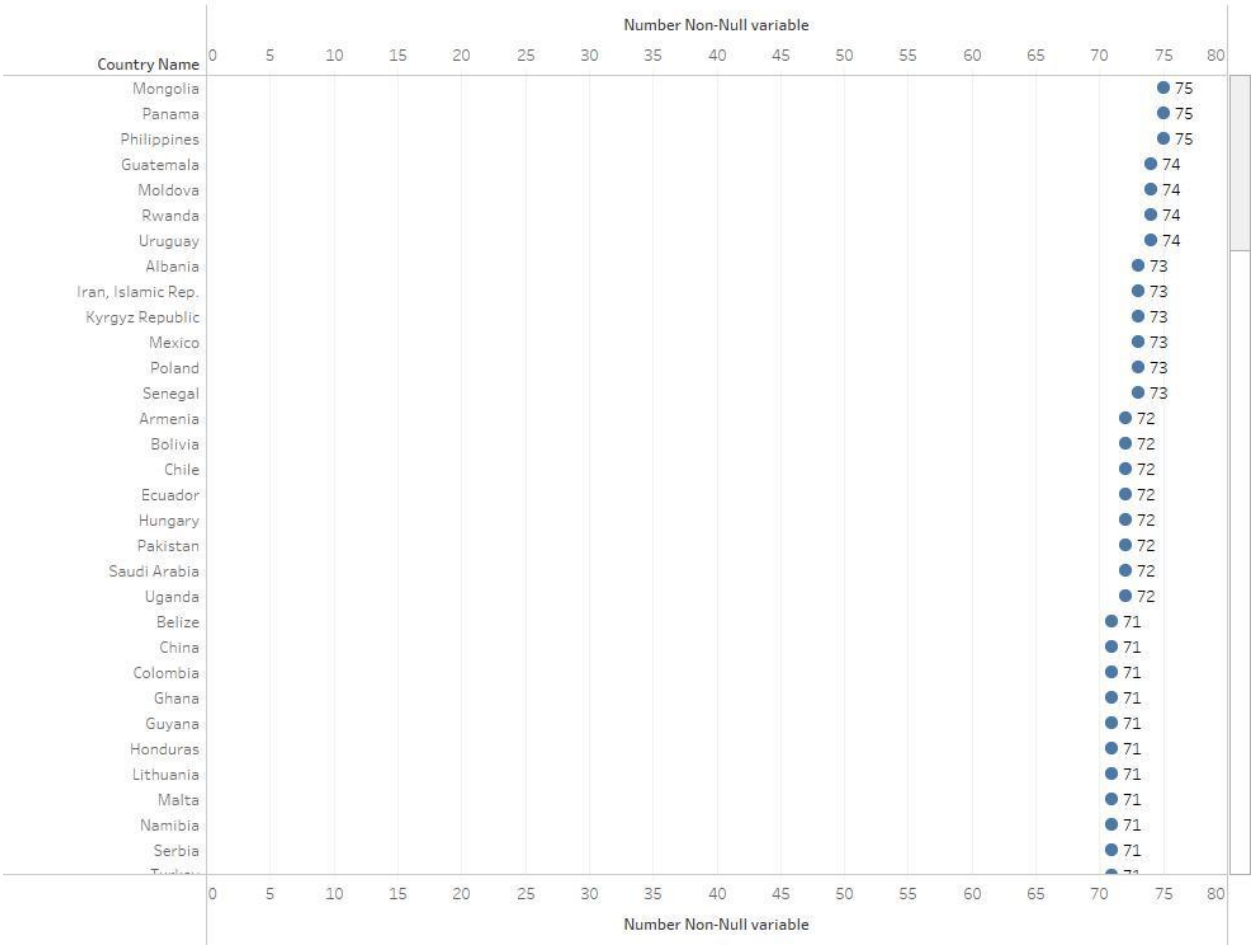
*Answer these questions:*

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 because it since these variables are not under the economic, education, demographic, or environment categories.

| No | Variable Name | Category |
|---|---|---|
| 1 | IT_NET_USER_P2 | Background |
| 2 | SH_DYN_AIDS_ZS | Background |
| 3 | SH_DYN_MORT | Background |
| 4 | SH_MED_PHYS_ZS | Health |
| 5 | SH_XPD_PCAP | Health |
| 6 | SN_ITK_DEFC_ZS | Health |
| 7 | SP_POP_DPND | Health |
| 8 | SG_VAW_BURN_ZS | Health |
| 9 | SH_TBS_PREV | Health |

Data Categories for PCA: Education, Economic, Environment

# 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.*

Since the manager mentioned that he would like to see four clusters in the result, It would be better to use K-centroid clustering method. While Hierarchical clustering method can also be done to create the clusters, it is more time efficient and easier to use K-centroid because we have a pre-determined number of clusters to begin with.

**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_AG25T | | |

| | |
|---|---|
| | 99 |
| 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 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.*

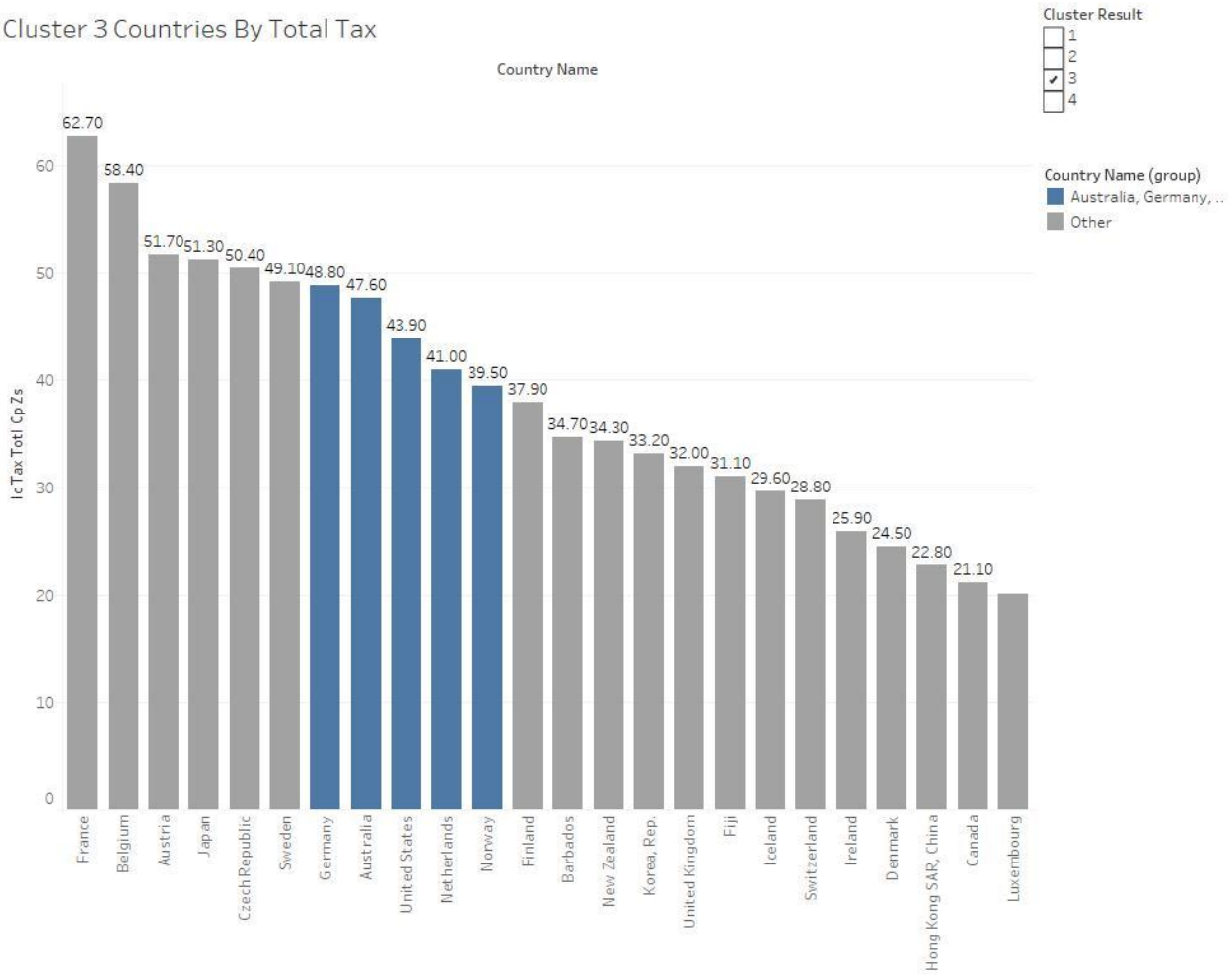
*Answer this question.*

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.

| Number | Cluster 1 | Cluster 2 | Cluster 3 | Cluster 4 |
|---|---|---|---|---|
| 1 | Afghanistan | Algeria | Australia | Albania |
| 2 | Bangladesh | Bahrain | Austria | Argentina |
| 3 | Benin | Bolivia | Barbados | Armenia |
| 4 | Burundi | Botswana | Belgium | Belize |
| 5 | Cambodia | Brazil | Canada | Brunei Darussalam |
| 6 | Cameroon | China | Czech Republic | Bulgaria |
| 7 | Central African Republic | Dominican Republic | Denmark | Chile |
| 8 | Congo, Dem. Rep. | Ecuador | Fiji | Colombia |
| 9 | Congo, Rep. | Egypt, Arab Rep. | Finland | Costa Rica |
| 10 | Cote d'Ivoire | El Salvador | France | Croatia |
| 11 | Gambia, The | Gabon | Germany | Cuba |
| 12 | Haiti | Ghana | Hong Kong SAR, China | Cyprus |
| 13 | Kenya | Guatemala | Iceland | Estonia |
| 14 | Lao PDR | Guyana | Ireland | Greece |
| 15 | Lesotho | Honduras | Japan | Hungary |
| 16 | Liberia | India | Korea, Rep. | Israel |
| 17 | Malawi | Indonesia | Luxembourg | Italy |
| 18 | Mali | Iran, Islamic Rep. | Netherlands | Kazakhstan |
| 19 | Mauritania | Iraq | New Zealand | Kyrgyz Republic |
| 20 | Mozambique | Jamaica | Norway | Latvia |
| 21 | Myanmar | Jordan | Sweden | Lithuania |
| 22 | Niger | Kuwait | Switzerland | Macao SAR, China |

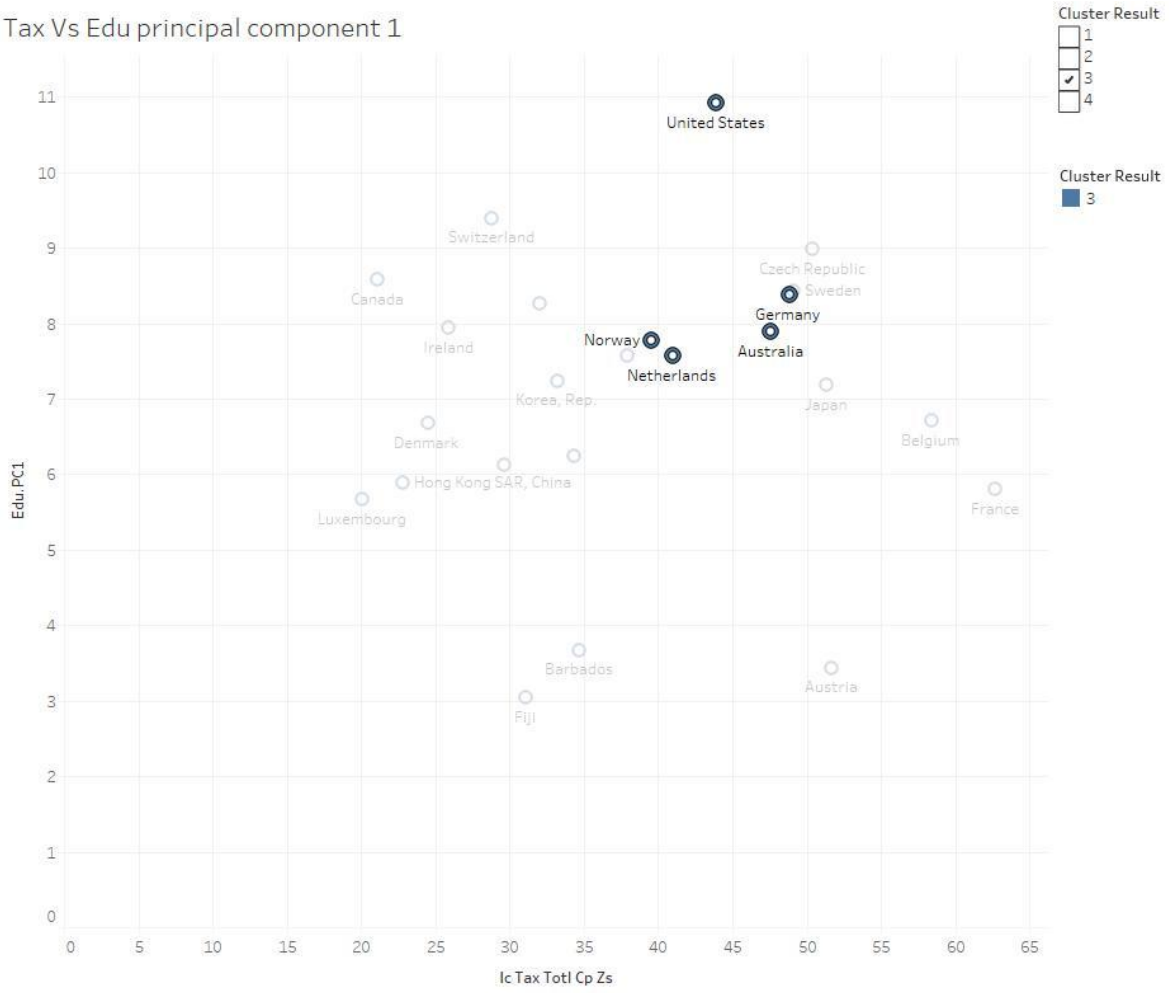
| | | | | |
|---|---|---|---|---|
| 23 | Papua New Guinea | Libya | United Kingdom | Malaysia |
| 24 | Rwanda | Maldives | United States | Malta |
| 25 | Senegal | Mongolia | | Mauritius |
| 26 | Sierra Leone | Morocco | | Mexico |
| 27 | Sudan | Nepal | | Moldova |
| 28 | Swaziland | Nicaragua | | Namibia |
| 29 | Tanzania | Pakistan | | Poland |
| 30 | Togo | Panama | | Portugal |
| 31 | Uganda | Paraguay | | Romania |
| 32 | Yemen, Rep. | Peru | | Russian Federation |
| 33 | Zambia | Philippines | | Serbia |
| 34 | Zimbabwe | Qatar | | Singapore |
| | | Saudi Arabia | | Slovak Republic |
| | | Sri Lanka | | Slovenia |
| | | Syrian Arab Republic | | South Africa |
| | | Thailand | | Spain |
| | | Trinidad and Tobago | | Tajikistan |
| | | Tunisia | | Tonga |
| | | Turkey | | Ukraine |
| | | United Arab Emirates | | Uruguay |
| | | Venezuela, RB | | |
| | | Vietnam | | |

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. Netherlands
2. Australia
3. Norway
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*<br>*Austria*<br>*Barbados*<br>*Belgium*<br>*Canada*<br>*Czech Republic*<br>*Denmark*<br>*Fiji*<br>*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.