Project: International Expansion

Step 1: Key Decisions

Key Decisions:

Answer these three questions

1. What decisions needs to be made?
   The following decisions need to be made:
   - Cleaning up data so that the dataset contains data records with a large of usable data fields
   - Determining relevant variables for analysis.
   - Using clustering methods to determine countries that are similar to the United States of America in terms of demographics, economics, education, and environment where the retail chain can expand its business

2. What data is needed to inform those decisions? Please include 2 examples in each of the following categories: Economic, Environment, Education
   - Economic: Employment/population ration, total labour force
   - Environment: Access to electricity, proportion of slum dwellers
   - Education: Average yrs of schooling, percentage of population.

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.

   144 countries after leaving out countries with count_null_values > 25.

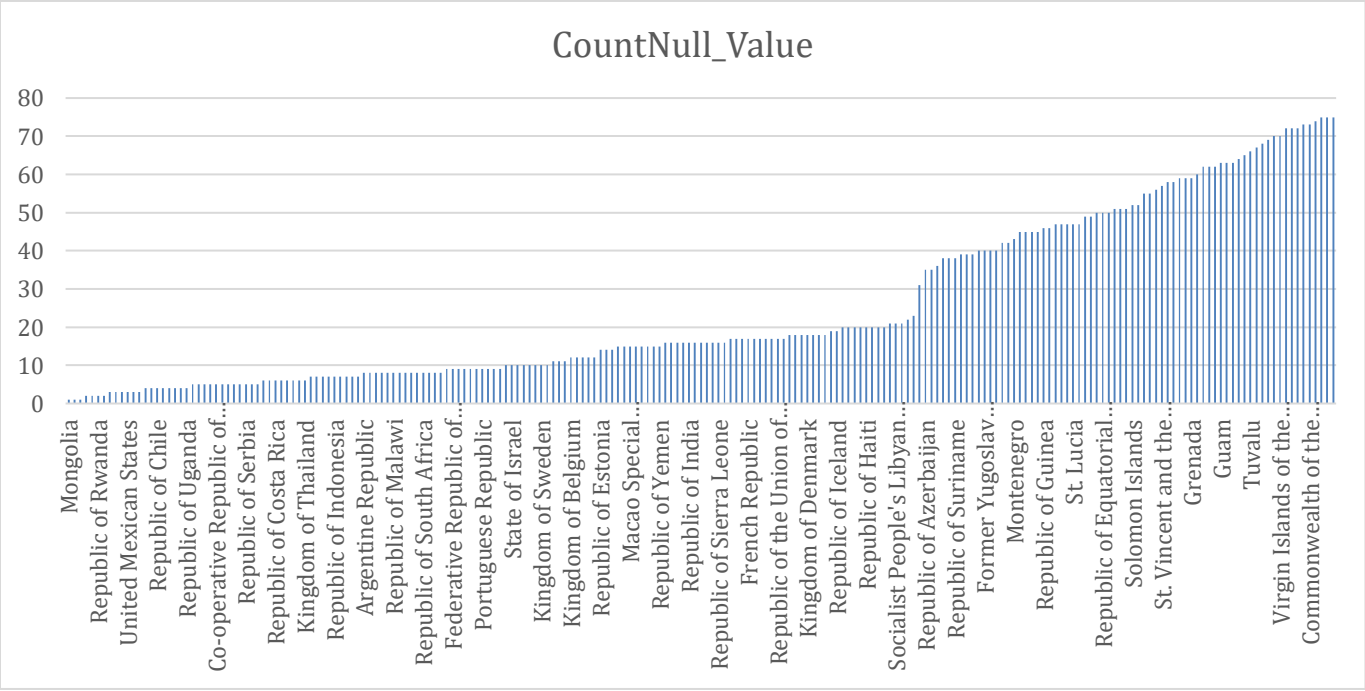
144 records displayed, 81 fields, , 92 KB

Table

81 of 81 Fields | Cell Viewer

Record #	Country Code	Country Name	Long Name	Table Name	BAR_SCHL_1519	BAR_SCHL_1519_FE	BAR_SCHL_15UP	BAR_SC
1	AFG	Afghanistan	Islamic State of Afghanistan	Afghanistan	4.81	4.13	3.85	1.97
2	ALB	Albania	Republic of Albania	Albania	9.17	9.21	9.93	9.7
3	DZA	Algeria	People's Democratic Republic of Algeria	Algeria	7.41	8.12	6.68	6.35
4	ARG	Argentina	Argentine Republic	Argentina	8.75	9.02	9.51	9.57
5	ARM	Armenia	Republic of Armenia	Armenia	9.47	9.32	10.73	10.64
6	AUS	Australia	Commonwealth of Australia	Australia	10.81	10.95	11.54	11.6
7	AUT	Austria	Republic of Austria	Austria	5.93	5.99	9.6	8.87
8	BHR	Bahrain	Kingdom of Bahrain	Bahrain	8.49	8.58	7.06	7.25
9	BGD	Bangladesh	People's Republic of Bangladesh	Bangladesh	8.21	9.34	5.91	5.69
10	BRB	Barbados	Barbados	Barbados	9.55	10.01	9.45	9.69
11	BEL	Belgium	Kingdom of Belgium	Belgium	8.45	8.59	10.69	10.52
12	BLZ	Belize	Belize	Belize	10.45	10.59	11.29	11.34
13	BEN	Benin	Republic of Benin	Benin	6.87	6.27	4.43	3.31
14	BOL	Bolivia	Plurinational State of Bolivia	Bolivia	7.75	7.81	8.25	7.71
15	BWA	Botswana	Republic of Botswana	Botswana	10.39	10.42	9.55	9.42
16	BRA	Brazil	Federative Republic of Brazil	Brazil	6.92	7.31	7.89	8.07
17	BRN	Brunei Darussalam	Brunei Darussalam	Brunei Darussalam	7.95	7.88	8.74	8.58
18	BGR	Bulgaria	Republic of Bulgaria	Bulgaria	8.01	7.98	11.24	11.24
19	BDI	Burundi	Republic of Burundi	Burundi	5.08	4.79	3.35	2.94
20	KHM	Cambodia	Kingdom of Cambodia	Cambodia	5.67	5.66	4.72	3.96
21	CMR	Cameroon	Republic of Cameroon	Cameroon	6.01	5.74	6.15	5.65
22	CAN	Canada	Canada	Canada	10.16	10.16	12.32	12.39
23	CAF	Central African Republic	Central African Republic	Central African Republic	3.65	3.22	3.76	2.72
24	CHL	Chile	Republic of Chile	Chile	8.74	8.93	9.78	9.72

A bar chart is plotted with long name vs count_null_values. Lower count means higher non-null values.



2. Which data categories will be used for Principal Components Analysis (PCA)? There should be three categories that are targeted for PCA.
   Pop > 25 with Degrees, Literacy Rate, Average years of schooling.

3. Which variables did you decide to be irrelevant for this analysis? Only variables under the education, economic, and environment categories should be included.

Sl No.	Series Code	Category
1	Women who believe a husband is justified in beating his wife when she burns the food	Health
2	Prevalence of HIV, total (% of population ages 15-49)	Background
3	Mortality rate, under-5 (per 1,000 live births)	Background
4	Physicians (per 1,000 people)	Health
5	Health expenditure per capita (current US$)	Health
6	Prevalence of undernourishment (% of population)	Health
7	Age dependency ratio (% of working-age population)	Health
8	Internet users (per 100 people)	Background
9	Prevalence of tuberculosis (per 100,000 population)	Health

Step 3: Determine Clusters and Methodology

Answer this question:

- What clustering method did you decide to use? Please justify your answer.

1) K-Centroid Neural Gas

Neural Gas Cluster Assessment Report

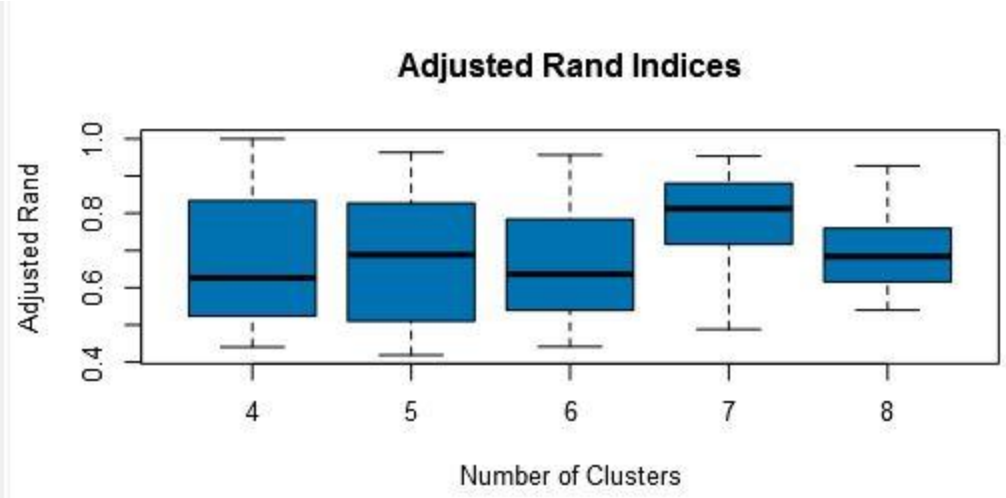
Summary Statistics

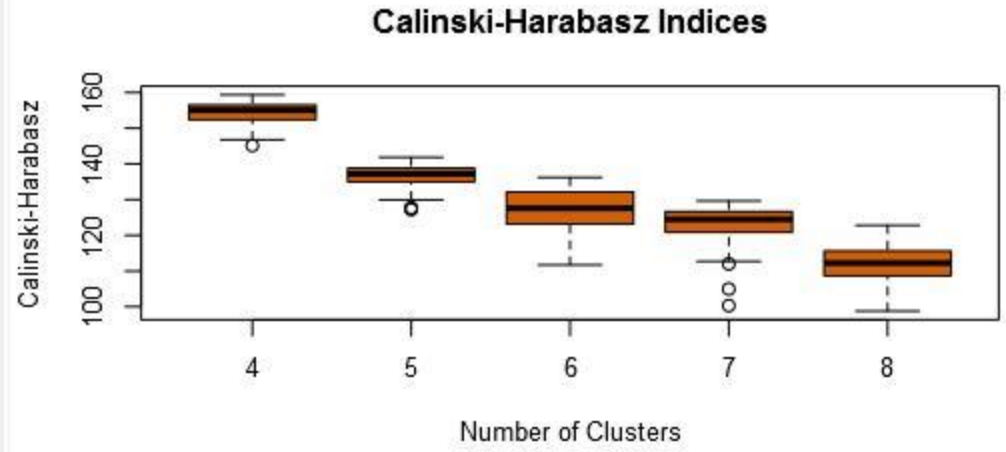
Adjusted Rand Indices:

	4	5	6	7	8
Minimum	0.4407	0.419	0.4415	0.4883	0.5399
1st Quartile	0.5241	0.5125	0.5427	0.7215	0.6178
Median	0.6264	0.6884	0.636	0.8124	0.6838
Mean	0.6766	0.6707	0.6611	0.7829	0.6939
3rd Quartile	0.8315	0.8208	0.7829	0.8792	0.7571
Maximum	1	0.9634	0.9568	0.9533	0.9272

Calinski-Harabasz Indices:

	4	5	6	7	8
Minimum	145.2	127.2	111.7	100.3	98.77
1st Quartile	152.4	135	123.2	120.9	108.8
Median	155	137.2	127.6	124.5	112.2
Mean	154.3	136.4	127.2	123.2	111.8
3rd Quartile	156.5	138.8	132.1	126.6	115.6
Maximum	159.4	141.8	136.2	129.6	122.8





Comparing adjusted Rand (AR) and CH indexes, AR indicates cluster =7 while CH indicates cluster=4.

2) K-Means

K-Means Cluster Assessment Report

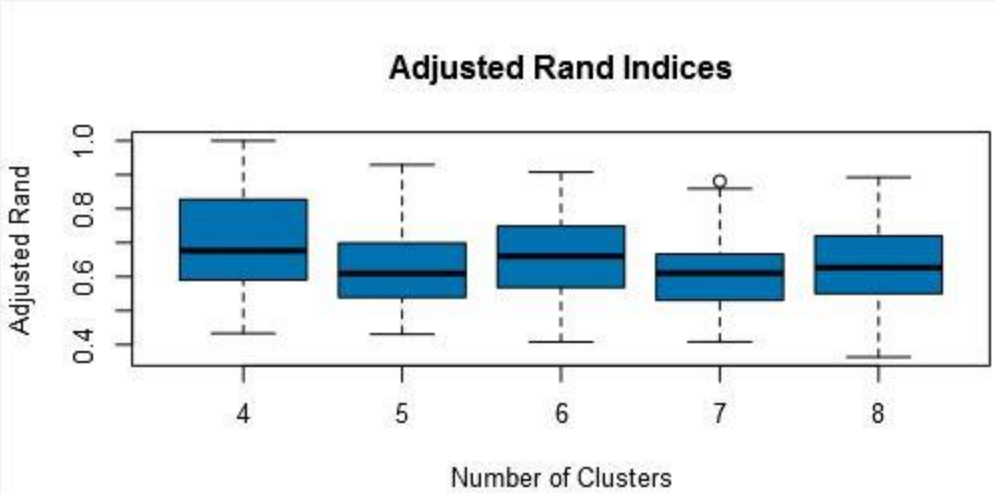
Summary Statistics

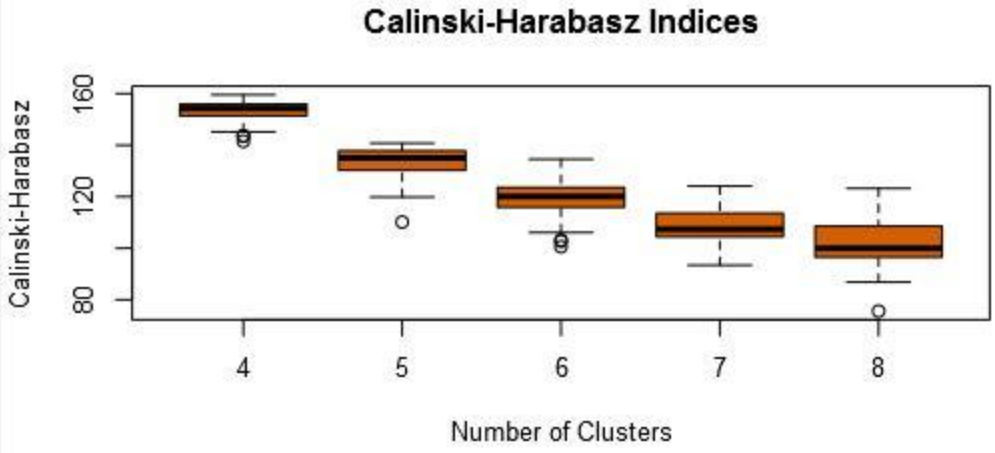
Adjusted Rand Indices:

	4	5	6	7	8
Minimum	0.4334	0.4305	0.4076	0.4083	0.3633
1st Quartile	0.5942	0.5394	0.5704	0.5312	0.5503
Median	0.6761	0.609	0.6603	0.6102	0.6267
Mean	0.702	0.6224	0.6603	0.6124	0.6293
3rd Quartile	0.8239	0.6987	0.7488	0.6664	0.72
Maximum	1	0.9295	0.9076	0.8811	0.8924

Calinski-Harabasz Indices:

	4	5	6	7	8
Minimum	141.5	110.2	100.6	93.38	75.6
1st Quartile	151.4	130.3	115.9	104.4	96.61
Median	154.3	135	120	107.4	99.99
Mean	153.4	133.7	119.4	108.7	102.2
3rd Quartile	155.9	137.8	123.6	113.4	108.4
Maximum	159.5	140.7	134.5	124.1	123.2





Comparing adjusted Rand (AR) and CH indexes, the number of clusters should be 4. Median for AR is 0.06761 and median for CH is 154.3.

3) K-Medians

K-Medians Cluster Assessment Report

Summary Statistics

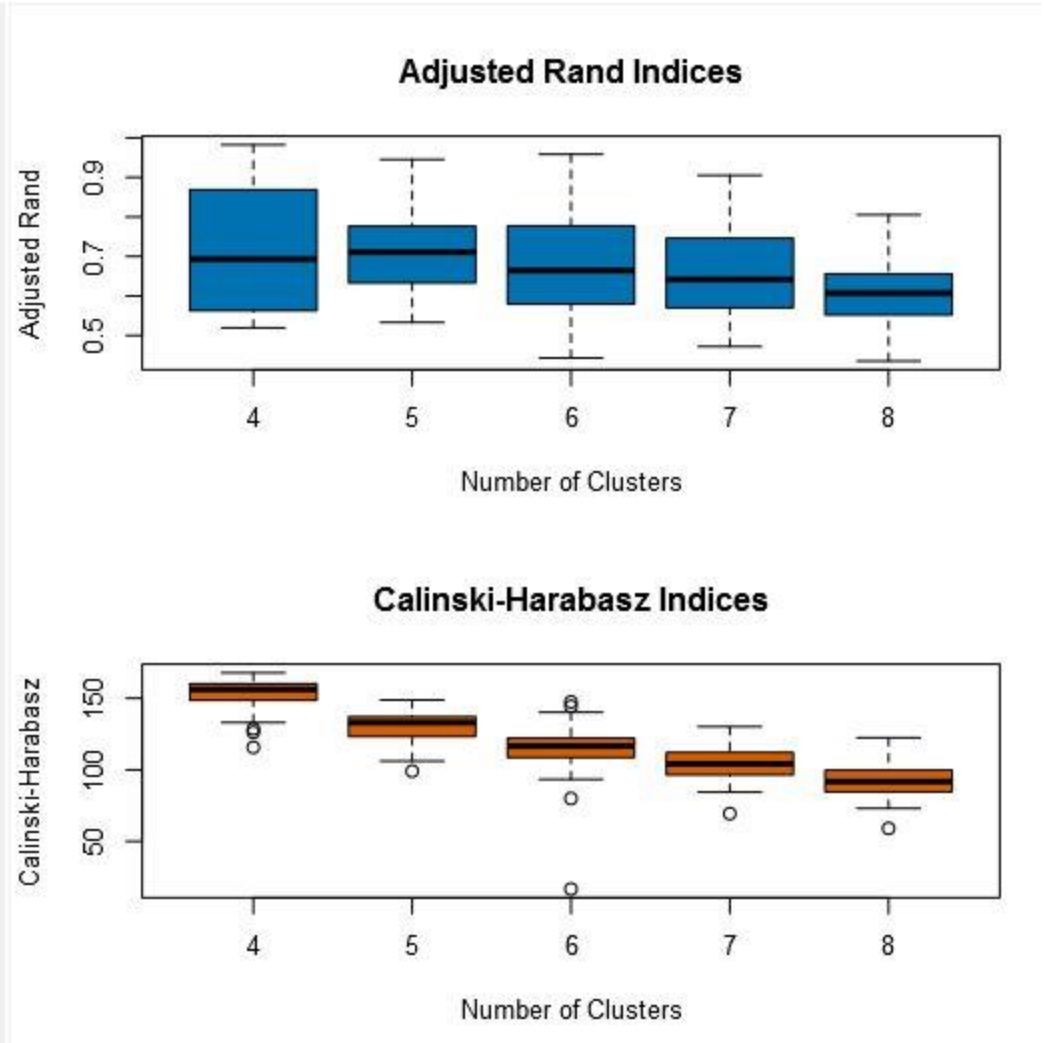
Adjusted Rand Indices:

	4	5	6	7	8
Minimum	0.5193	0.533	0.4432	0.4723	0.4353
1st Quartile	0.567	0.636	0.5801	0.5733	0.5537
Median	0.6931	0.7107	0.6643	0.6418	0.6068
Mean	0.721	0.711	0.68	0.6521	0.6138
3rd Quartile	0.8667	0.7762	0.7761	0.744	0.6558
Maximum	0.9823	0.9451	0.9582	0.905	0.8051

Calinski-Harabasz Indices:

	4	5	6	7	8
Minimum	115.8	99.1	16.65	69.44	59.07
1st Quartile	148.7	123.6	108.7	96.59	84.75
Median	156.2	133	116.6	104.2	91.77
Mean	153.7	130.5	114.4	104.4	92.46
3rd Quartile	160.1	137.1	122	112.2	99.78
Maximum	167.9	148.9	147.5	130.2	122.4

Plots



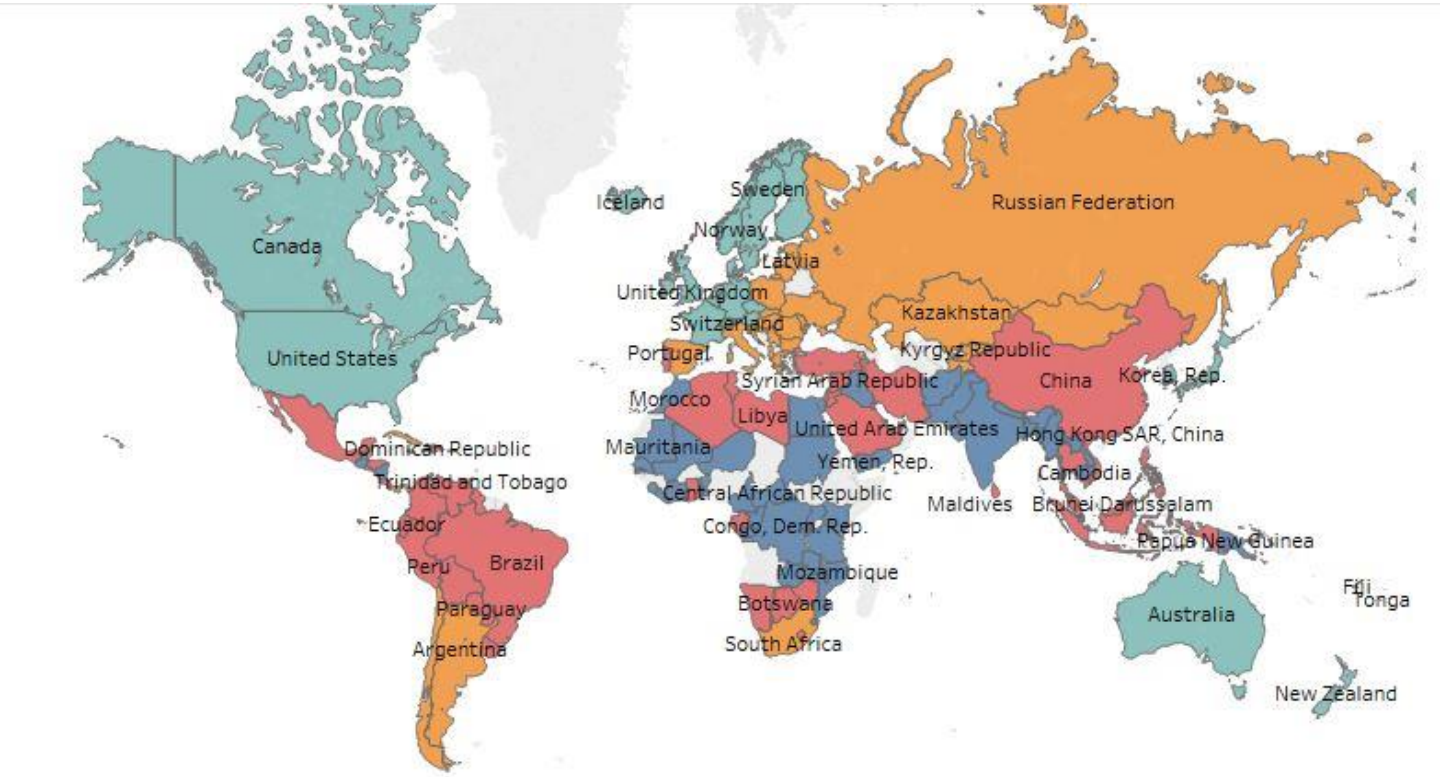
Comparing adjusted Rand (AR) and CH indexes, the number of clusters should be 4 . Median for AR is 0.6931and median for CH is 156.2.

Comparing the thresults of the three methods used, we can see that K-means has the better indices across AR and CH viz high median and compact plot.
Hence clustering method to be used is K-Means and number of clusters required is 4

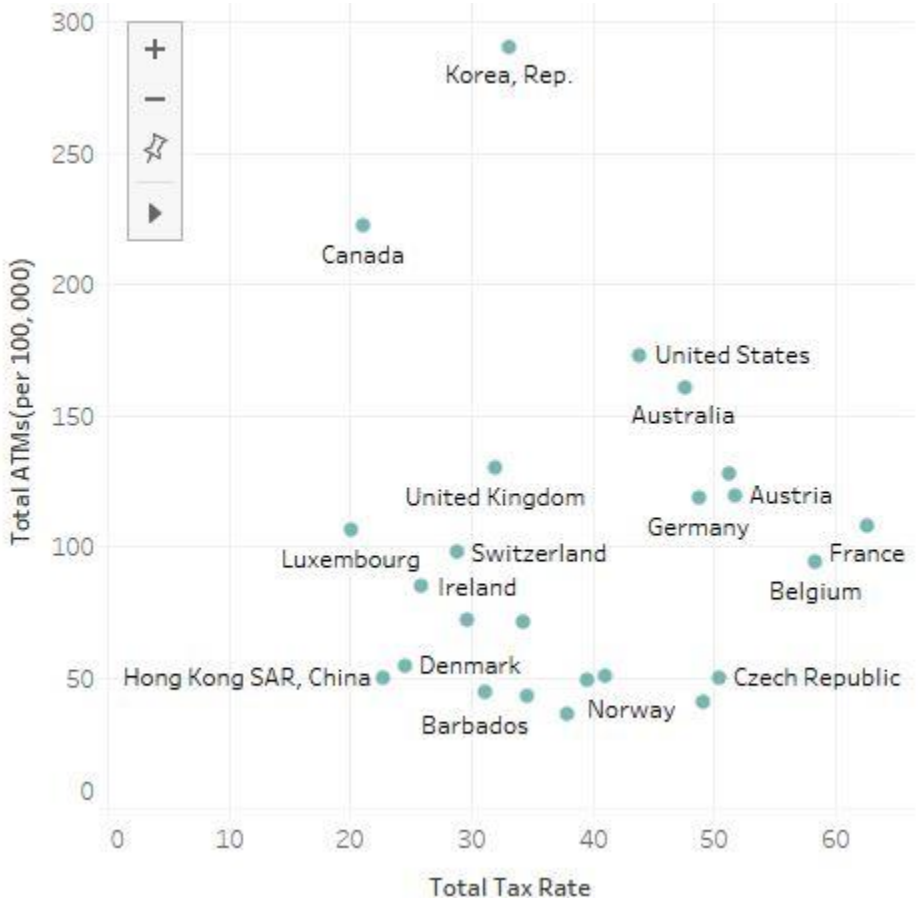
Step 4: Run the Data and Visualize

1. Do the clusters make sense

   Yes, they make sense due to the following reasons:
   - Mostly differentiates the countries in terms of income.
   - The higher educated countries are mostly grouped together.
   - The countries with similar education status are similarly grouped together
   -



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?



Closest countries are Australia, UK, Japan and Austria in terms of Total Tax Rate by ATM Machines.

Step 5: Recommendation

The following countries are recommended:

Record #	Country.Name	P_7_Cluster_K_Median
1	Australia	1
2	Austria	1
3	Barbados	1
4	Belgium	1
5	Canada	1
6	Czech Republic	1
7	Denmark	1
8	Fiji	1
9	Finland	1
10	France	1
11	Germany	1
12	Hong Kong SAR, China	1
13	Iceland	1
14	Ireland	1
15	Japan	1
16	Korea, Rep.	1
17	Luxembourg	1
18	Netherlands	1
19	New Zealand	1
20	Norway	1
21	Sweden	1
22	Switzerland	1
23	United Kingdom	1

Why did you decide to choose these countries?

- These countries have similar income status as USA.
- The labour population of these countries have similar education levels as USA.
- These countries have similar overall education status.
- Electricity and internet access is similar to USA.