Clustering for countries of the world data

The countries of the world data set (https://www.kaggle.com/fernandol/countries-of-the-world/version/1#countries%20of%20the%20world.csv) is a compilation of demographic information for all of the world's countries and independent islands. It has 227 rows (countries and independent islands) and 20 columns (variables). In this project, we try to cluster countries based on their demographic information.

- Summary Statistics

	Population	Area..sq..mi..	Pop..Density..per.sq..mi..	Coastline..coast.area.ratio.
X	Min. :1.348e+04	Min. : 28	Min. : 1.8	Min. : 0.000
X.1	1st Qu.:1.189e+06	1st Qu.: 19915	1st Qu.: 26.8	1st Qu.: 0.090
X.2	Median :6.940e+06	Median : 118480	Median : 66.9	Median : 0.630
X.3	Mean :3.421e+07	Mean : 564183	Mean : 294.8	Mean : 16.495
X.4	3rd Qu.:2.086e+07	3rd Qu.: 496441	3rd Qu.: 164.7	3rd Qu.: 5.355
X.5	Max. :1.314e+09	Max. :9631420	Max. :16183.0	Max. :870.660

Table 1: Summary Statistics

	Net.migration	Infant.mortality..per.1000.births.	GDP....per.capita.	Literacy....
X	Min. :-20.9900	Min. : 2.29	Min. : 500	Min. : 17.60
X.1	1st Qu.: -1.3150	1st Qu.: 9.99	1st Qu.: 1800	1st Qu.: 69.95
X.2	Median : 0.0000	Median : 24.31	Median : 5100	Median : 90.90
X.3	Mean : -0.2065	Mean : 38.90	Mean : 9126	Mean : 81.94
X.4	3rd Qu.: 0.3950	3rd Qu.: 64.61	3rd Qu.:12950	3rd Qu.: 97.80
X.5	Max. : 23.0600	Max. :163.07	Max. :37800	Max. :100.00

Table 2: Summary Statistics

- Matrix scatterplot and Correlation matrix

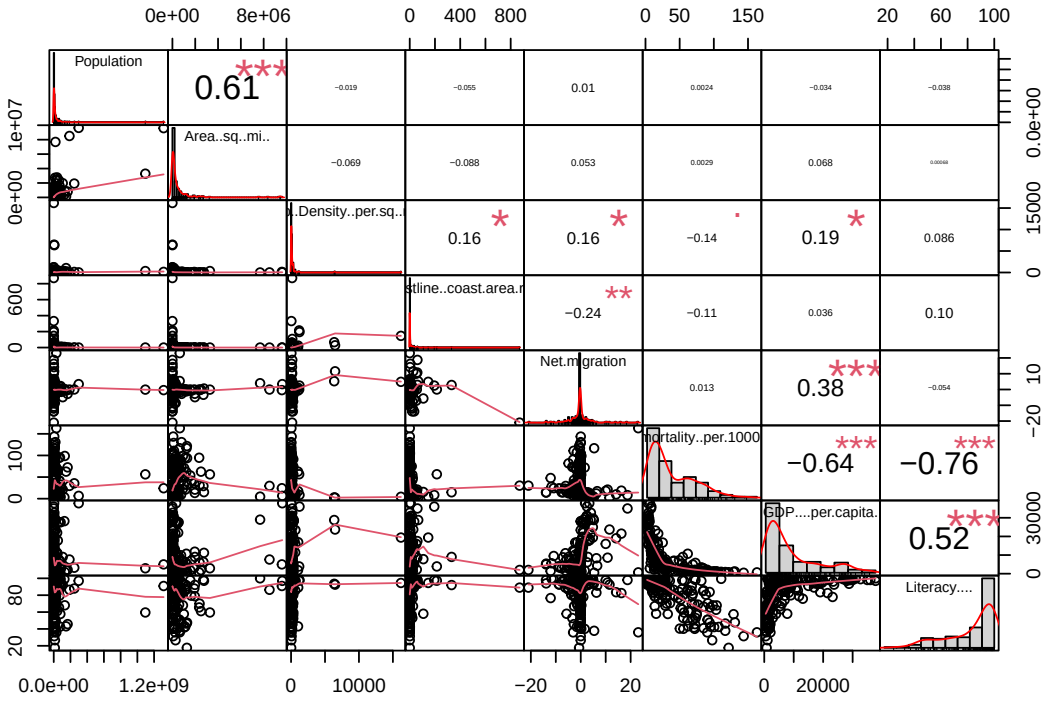

Figure 1: Matrix scatterplot

	Population	Area	Pop..Density	Coastline	Net.migration	Infant.mortality	GDP	Literacy
Population	1.00	0.61	-0.02	-0.05	0.01	0.00	-0.03	-0.04
Area	0.61	1.00	-0.07	-0.09	0.05	0.00	0.07	0.00
Pop..Density	-0.02	-0.07	1.00	0.16	0.16	-0.14	0.19	0.09
Coastline	-0.05	-0.09	0.16	1.00	-0.24	-0.11	0.04	0.10
Net.migration	0.01	0.05	0.16	-0.24	1.00	0.01	0.38	-0.05
Infant.mortality	0.00	0.00	-0.14	-0.11	0.01	1.00	-0.64	-0.76
GDP	-0.03	0.07	0.19	0.04	0.38	-0.64	1.00	0.52
Literacy	-0.04	0.00	0.09	0.10	-0.05	-0.76	0.52	1.00

Table 3: Correlation matrix

- As the correlation between most of the variables are very low, therefore, rather than going with the correlation-based distance, we can use metric-based distance for clustering.
- Also I would suggest standardizing the variables as they are in different scales and some have very high ranges.
- The panel histograms shows that the distributions of most of the variables are highly right skewed.

Hierarchical Clustering with Complete linkage

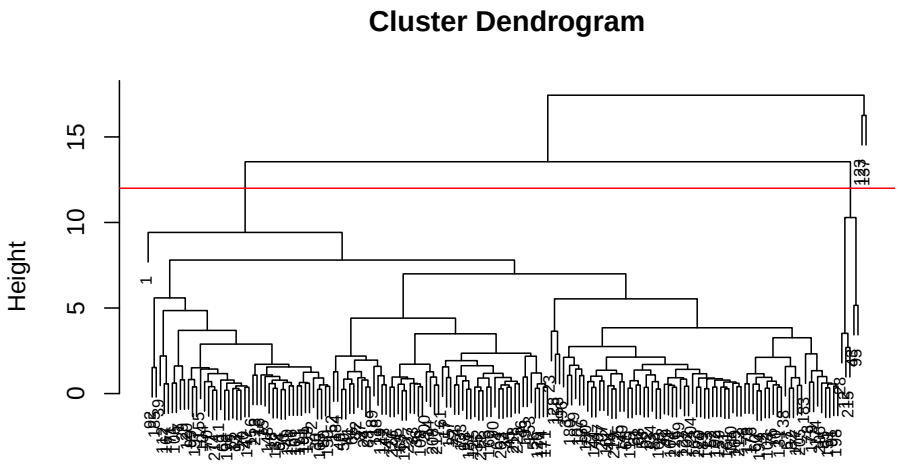

Figure 2: Hierarchical Clustering with Complete linkage

	1	2	3	4
1	172	5	1	1

Table 4: Number of observations within each cluster (Hierarchical Clustering)

	Population	Area	Pop..Density	Coastline	Net.migration	Infant.mortality	GDP	Literacy
Cluster 1	18649182.576	362053.733	208.764	11.247	-0.161	39.575	8894.186	81.690
Cluster 2	583222446.400	7742957.000	105.160	0.200	1.378	24.254	16460.000	86.760
Cluster 3	453125.000	28.000	16183.000	146.430	4.860	4.390	19400.000	94.500
Cluster 4	108004.000	702.000	153.900	870.660	-20.990	30.210	2000.000	89.000

Table 5: Cluster means of the variables (Hierarchical Clustering)

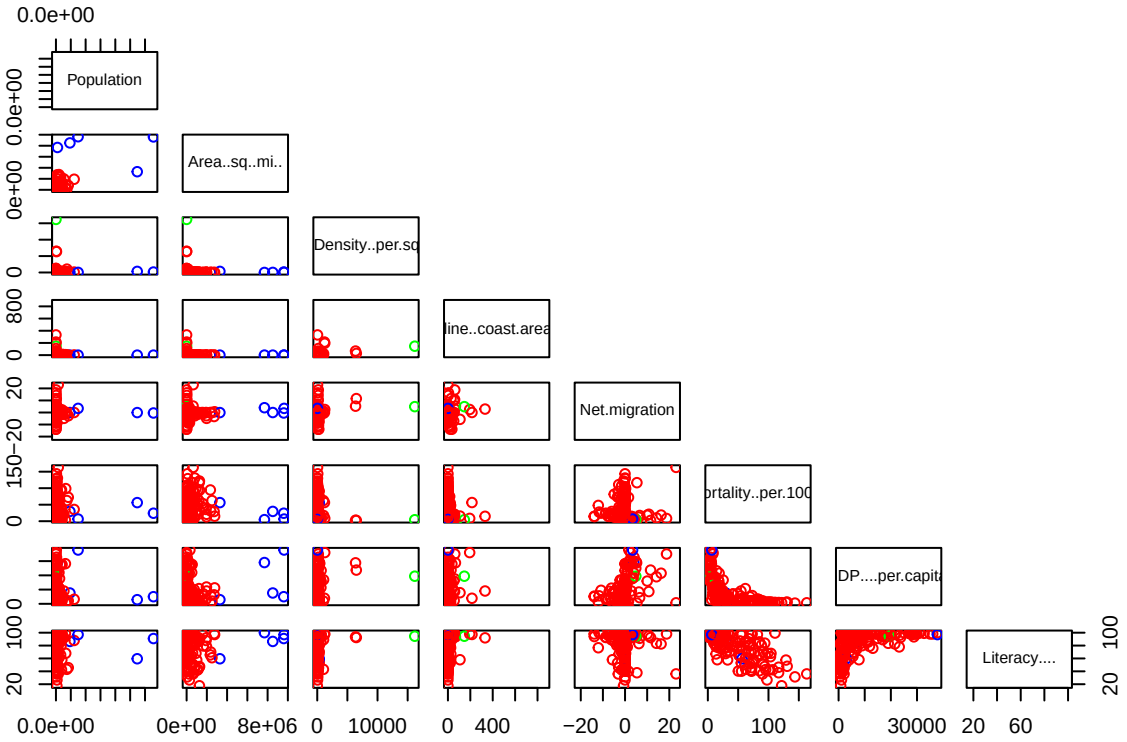

Figure 3: Matrix scatter plots

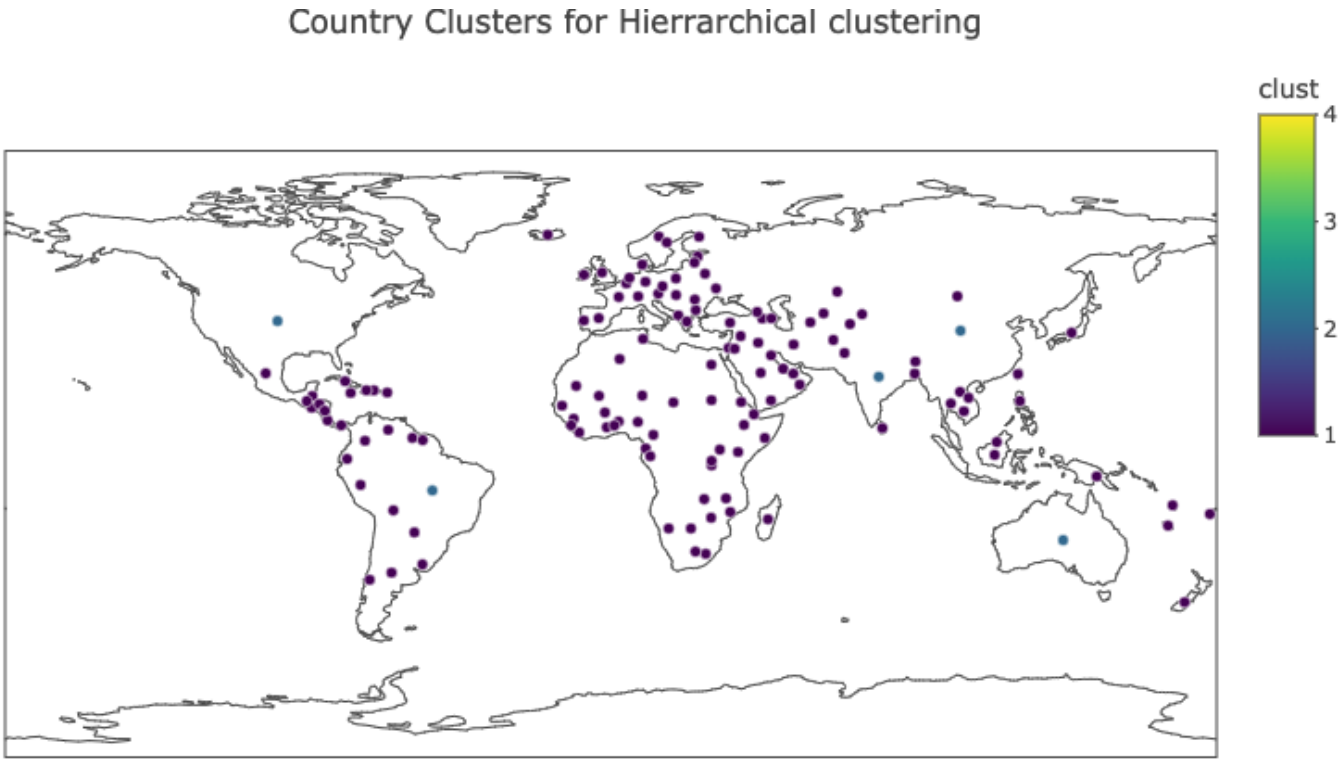

Figure 4: Geographical Model for 4 clusters

K-means clustering with k = 3

	1	2	3
1	112	62	5

Table 6: Number of observations within each cluster (k=3)

	Population	Area	Pop..Density	Coastline	Net.migration	Infant.mortality	GDP	Literacy
Cluster 1	16956084.107	290818.000	418.683	25.445	-0.259	17.144	12838.393	93.133
Cluster 2	21115146.968	479070.258	86.319	1.642	-0.240	79.377	1827.419	61.344
Cluster 3	583222446.400	7742957.000	105.160	0.200	1.378	24.254	16460.000	86.760

Table 7: Cluster means of the variables for k-means clustering (k=3)

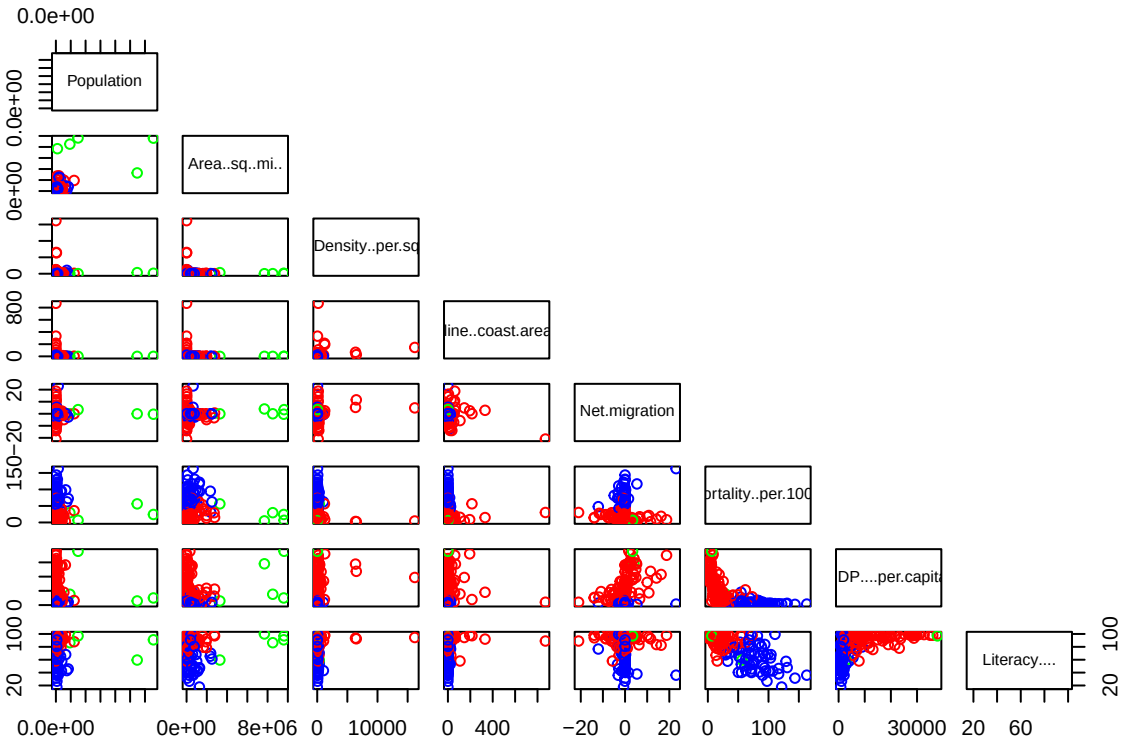

Figure 5: Matrix scatter plots

- The clusters seems to be overlapping.

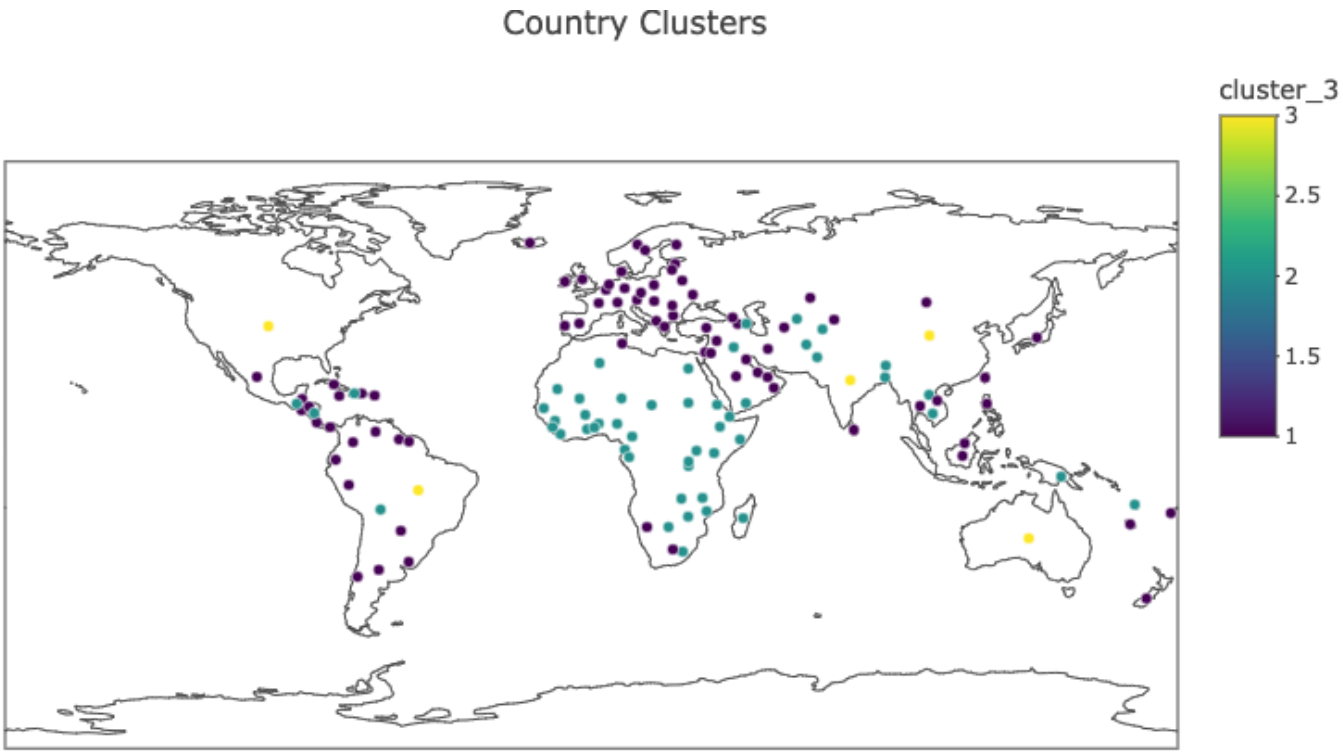

Figure 6: Geographical Model for k=3

K-means clustering with k = 4

	1	2	3	4
1	48	5	89	37

Table 8: Number of observations within each cluster (k=4)

	Population	Area	Pop..Density	Coastline	Net.migration	Infant.mortality	GDP	Literacy
Cluster 1	22790836.708	477826.979	91.685	1.152	0.285	86.601	1458.333	55.385
Cluster 2	583222446.400	7742957.000	105.160	0.200	1.378	24.254	16460.000	86.760
Cluster 3	17280830.607	392457.393	124.375	23.368	-2.328	27.226	6411.236	90.119
Cluster 4	15574768.811	119177.541	993.892	22.071	4.044	7.070	24610.811	96.084

Table 9: Cluster means of the variables for k-means clustering (k=4)

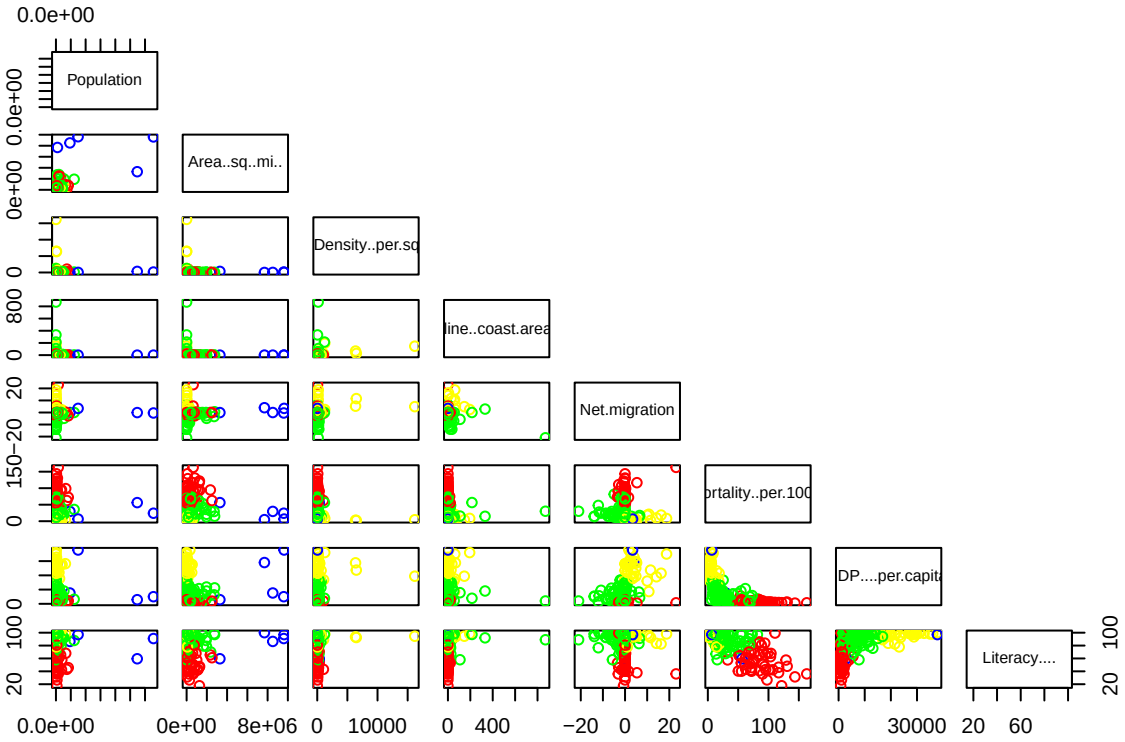

Figure 7: Matrix scatter plots

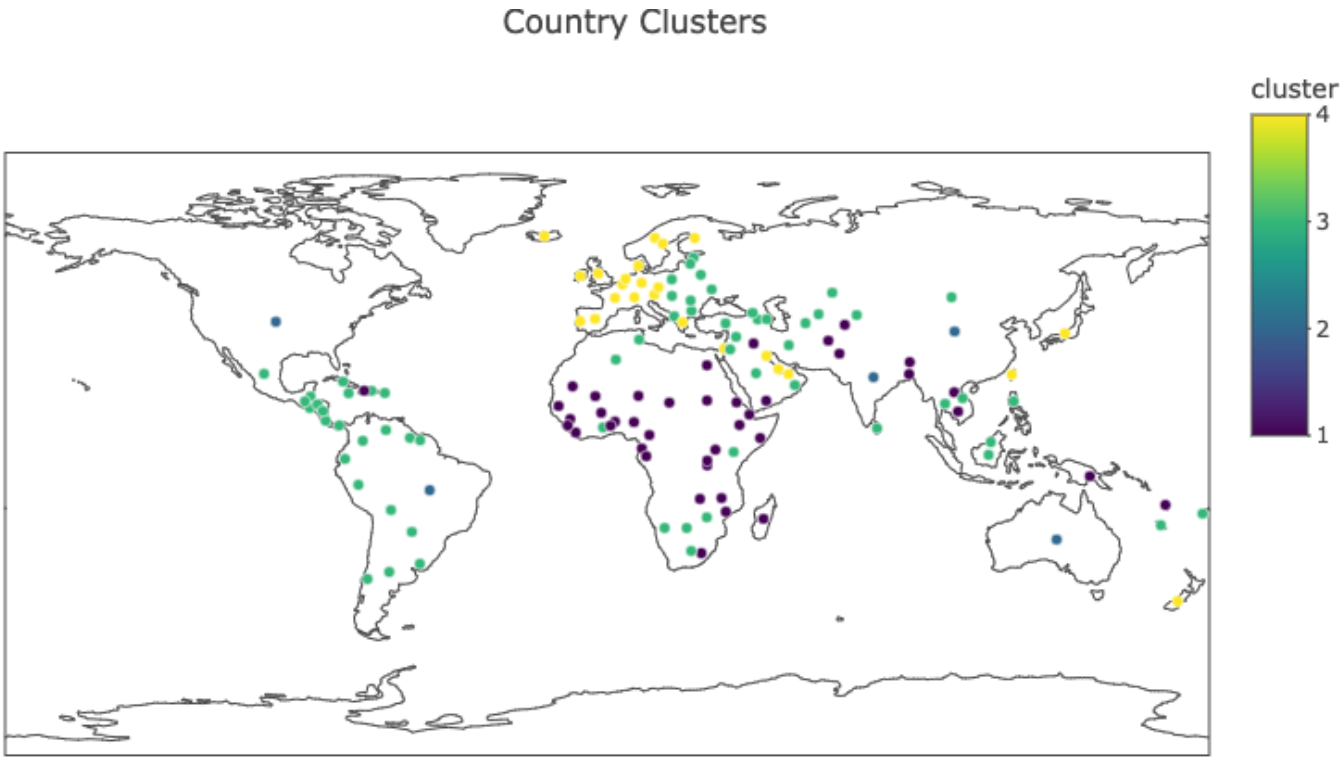

Figure 8: Geographical Model for k=4

K-means clustering with k = 5

	1	2	3	4	5
1	1	5	48	37	88

Table 10: Number of observations within each cluster (k=4)

	Population	Area	Pop..Density	Coastline	Net.migration	Infant.mortality	GDP	Literacy
Cluster 1	108004.000	702.000	153.900	870.660	-20.990	30.210	2000.000	89.000
Cluster 2	583222446.400	7742957.000	105.160	0.200	1.378	24.254	16460.000	86.760
Cluster 3	22790836.708	477826.979	91.685	1.152	0.285	86.601	1458.333	55.385
Cluster 4	15574768.811	119177.541	993.892	22.071	4.044	7.070	24610.811	96.084
Cluster 5	17475976.364	396909.159	124.040	13.739	-2.116	27.192	6461.364	90.132

Table 11: Cluster means of the variables for k-means clustering (k=5)

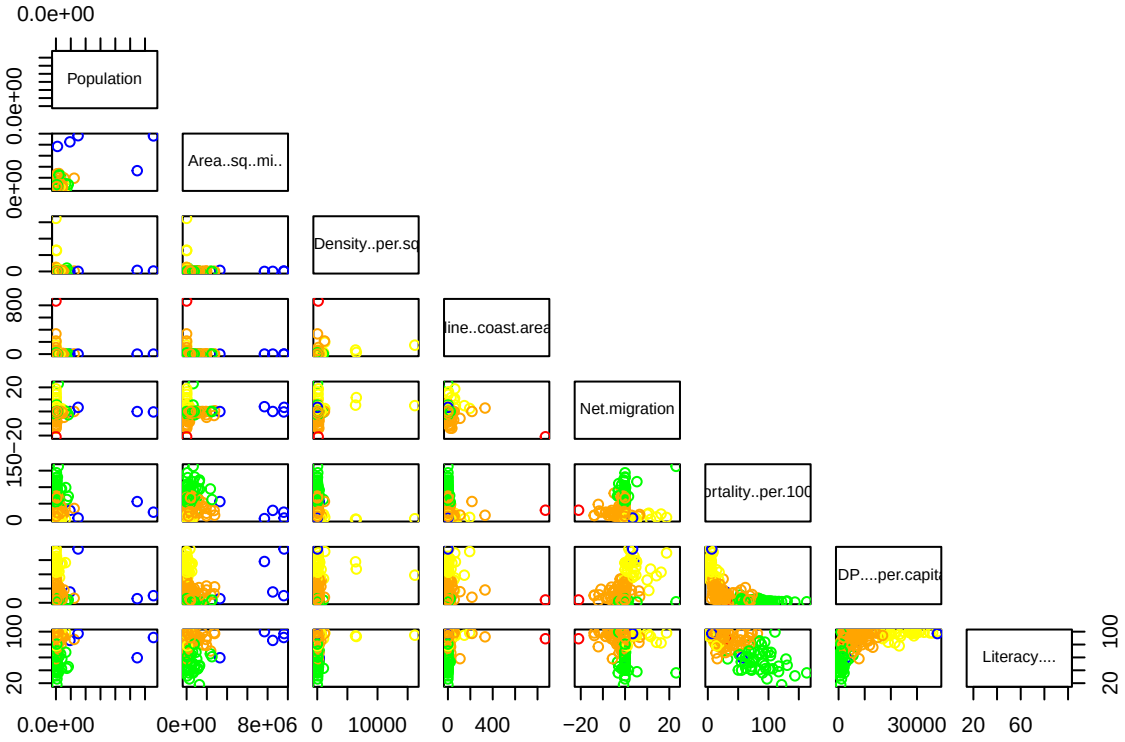

Figure 9: Matrix scatter plots

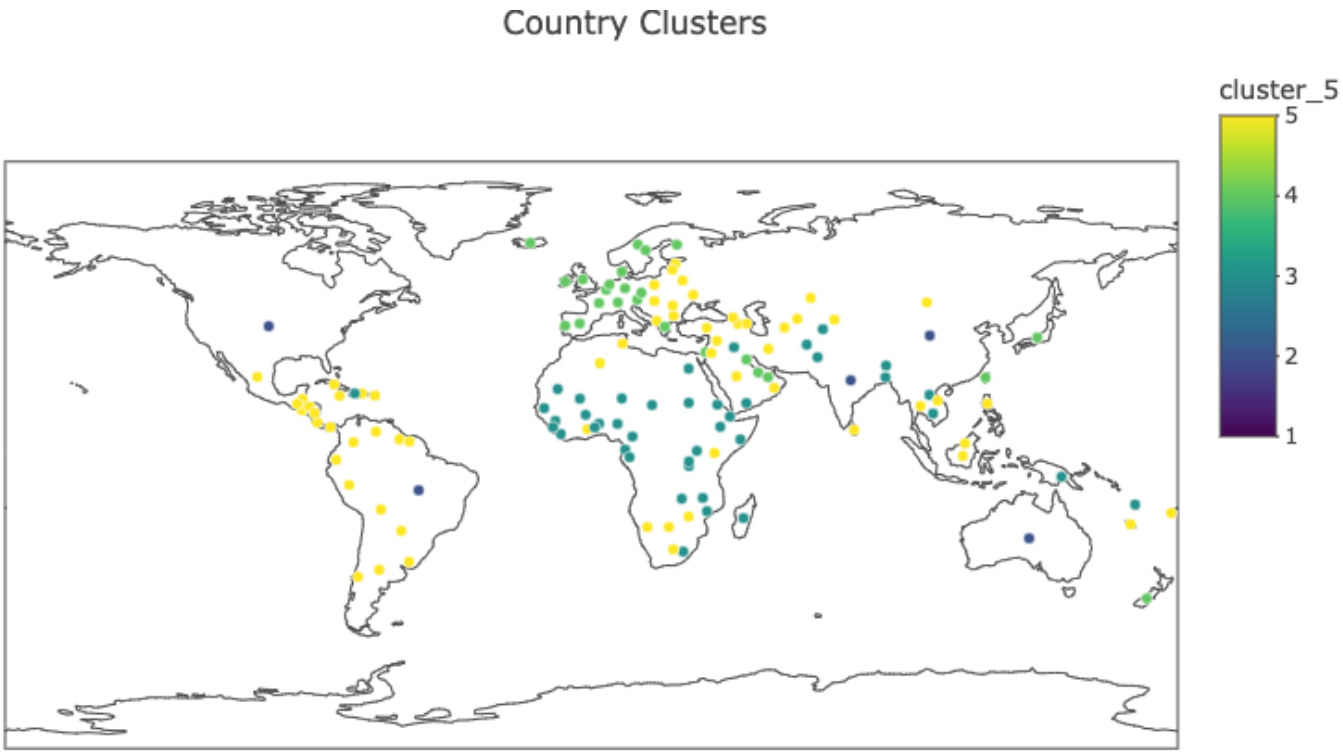

Figure 10: Geographical Model for k=5

- For all values of k that we have used and also for the hierarchical clustering, we can see the five countries, Australia, Brazil, China, India, USA, which are having a very high area and higher population are in a seperate cluster.

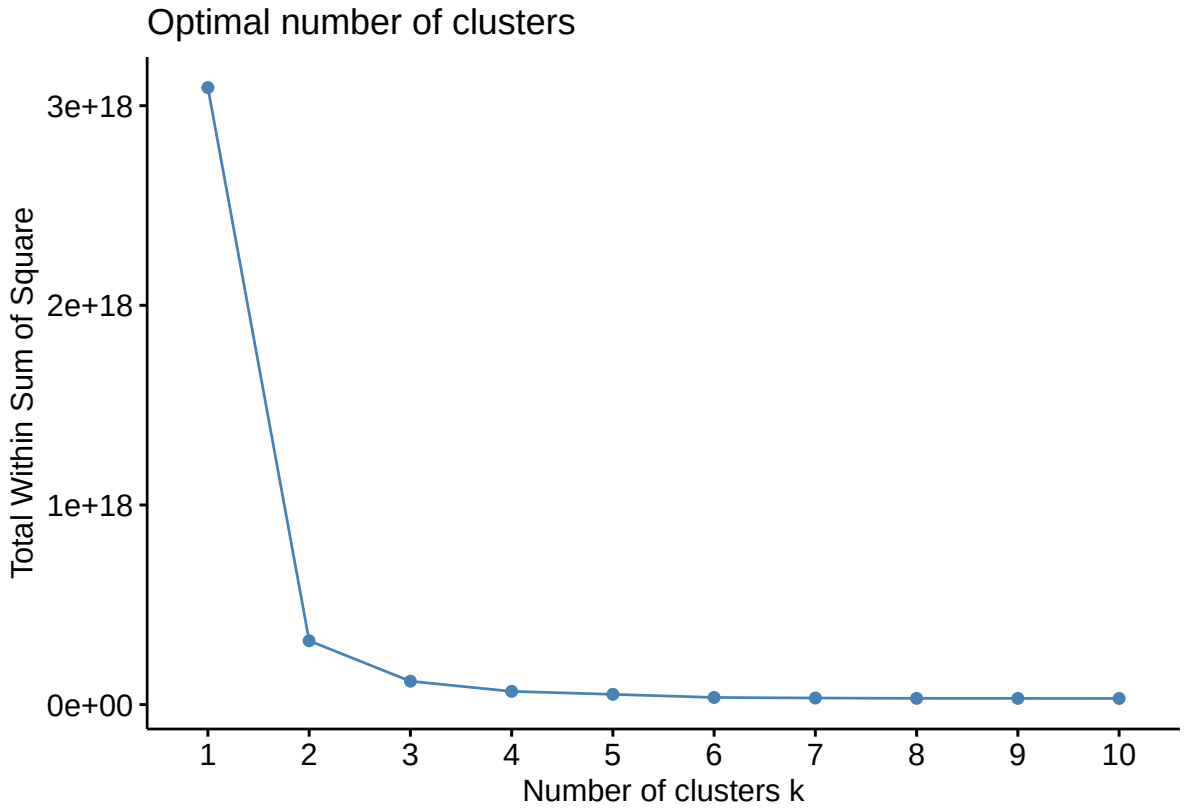

Figure 11: Plot of number of clusters vs total within cluster sum of squares

- The scree plot also shows that the optimal number of clusters is 4.