International education

**Data Selection**

In this project, I used data from the Global Competitiveness (2017-2018) data of Economic World Forum.

Nineteen countries were excluded in this analysis due to missing data in 2017-2018 report ('Angola', 'Barbados', 'Belize', 'Bolivia', 'Burkina Faso', 'Côte d'Ivoire', 'Gabon', 'Guyana', 'Haiti', 'Libya', 'Macedonia, FYR', 'Madagascar', 'Myanmar', 'Puerto Rico', 'Slovak Republic', 'Suriname', 'Syria', 'Timor-Leste'). However, future analysis can incorporate older report to give a more comprehensive look at all countries’ situation.

**Method**

I used Hierarchical clustering algorithms group similar objects into groups called clusters. Specifically, I used Agglomerative Clustering (bottom-up approach) using the euclidean distance as the measure of distance between points and ward linkage to calculate the proximity of clusters.

Features used for clustering:

* Fixed broadband Internet subscriptions/100 pop.
* Individuals using Internet, %
* Internet access in schools, 1-7 (best)
* Int’l Internet bandwidth, kb/s per user
* Primary education enrollment, net %
* Quality of math and science education, 1-7 (best)
* Quality of primary education, 1-7 (best)
* Quality of the education system, 1-7 (best)
* Income Group
* Economic Region

The best result I selected has 3 clusters with Silhouette score of 0.33.

**Result**

**A screenshot of a cell phone

Description automatically generated**

Cluster 1: Good Internet access, High education quality

Cluster 2: Bad Internet access, Low education quality

Cluster 0 (target cluster): Moderate Internet access, Low education quality

The clustering that cluster 2 and 0 have distinct internet access but similar education quality (shown by overlapping distribution). Further analysis can use statistical analysis to validate this point.

A close up of a device

Description automatically generatedA close up of a device

Description automatically generatedA picture containing screenshot

Description automatically generated

A close up of a device

Description automatically generatedA close up of a device

Description automatically generatedA picture containing screenshot

Description automatically generated

**Recommendation:**

I recommend that we should target Cluster 0 as our target regions because of its low education quality and moderate internet access.

**Next step:**

* Combine with Oppia traffic data to identify which country in Cluster 0 has good potential.
* Combine with language data to filter which country has high percentage of English-speakers.
* Explore other Clustering techniques + validate clustering results.

A close up of a piece of paper

Description automatically generated

**Appendix**

Data Source: The Global Competitiveness Report 2017-2018 <http://reports.weforum.org/global-competitiveness-index-2017-2018/>

Clustering method: <https://towardsdatascience.com/machine-learning-algorithms-part-12-hierarchical-agglomerative-clustering-example-in-python-1e18e0075019>