SUPER DATASCIENCE
MAKING THE COMPLEX SIMPLE

CASE STUDY 015 [Machine Learning: R] Country clustering

Difficulty Level: 3 of 3

Disclaimer: SuperDataScience has no affiliation with data sources. The scenario is made up for educational purposes.

You are a Data Analyst working for The World Bank and have been asked to cluster countries based on GDP per capita and CO2 per capita

1. Use spread from the tidyr package to make 1 row for each country and two columns with GDP per capita and CO2 per capita
2. Update column names to be easier to use
3. Remove countries which have missing
4. Add columns with normalised variables which have a mean of 0 and a variance of 1
5. Plot the elbow curve for these normalised variables to determine how many clusters to use
6. Create k-means clusters
7. Plot the k-means clusters using clusplot
8. Calculate average GDP, average co2 and number of countries by cluster

Good luck!

Difficulty note: this is a difficult assignment. Do not be surprised that there will be lots of nuances we have not covered off in the courses. But just like in the Real Life – there will be things training has not prepared you for and you will need to do research to find how to solve the problems at hand. If you get stuck, check the clues file. Data source: World Bank: The World Development Indicators