EMISSIONS-COUNTRY CLUSTERING 2019

This project is based on FAO databases, available at https://www.fao.org/faostat/en/#data.

The goal is to build a country clustering model to have a better understanding on countries emissions for 2019. This will help us to find countries common patterns that explains its emissions levels. K-means algorithm will be used for this purpose.

1. Dataset exploration

A final dataset was created using FAOSTAT data, same that was used for the predictions model. It contains the next original features:

- Country. The original dataset includes 234 countries, with the selection of the 2019 year some disappeared countries will be dropped.
- Item. We will keep "Farm-gate emissions" and " Land Use Change". The addition of both represents the emissions for agricultural use.
- Emission elements. We will keep "Emissions (CO2eq) (AR5)". A new feature will be created, "% world emissions".
- Year = 2019
- Unit. All values are in kilotonnes (1 kilotonne equals 1000 tonnes).
- Value, emissions value.

2. New features.

New features will be added, also from FAOSTAT datasets. The rules for addition are:

- Filtering by 2019 year.
- Merging under country name.

See next for an explanation on such datasets, and the features created after the merges:

- Population, total for 2019. Total value and % over world population. After the merge the created features were:
    - Emisiones per capita (country).
    - Emisiones per capita (world).

- GDP, total for 2019. After the merge the created features were:
    - GDP per capita (country).
    - GDP per capita (world).
    - Emissions per country GDP.
    - Emissions per world GDP.

- Country Group, including continent and continent area.
- Agricultural Production. After the merge the created features were:

    - Production per capita.
    - Emissions per country production.
    - Emissions per world production.

3. Exploratory Data Analysis

For a better visualization of the final dataset we will plot:

- Histograms.
- Scatters.
- Sunburst.
- Treemap.

4. Dataset final treatment.

- Imputing nulls with KNN Imputer. In this clustering case k-nearest neighbors algorithm is really useful to group countries by feature proximity, and avoiding eliminating or imputing random values to the nulls.
- Observing outliers, by total emissions: China, Indonesia, India, Brasil, United States, Congo Democratic Republic, and at a second level Pakistan, Argentina and Myanmar. These countries have substantial high levels of emissions values. Outliers were not eliminated, but is important to keep them in mind due to its important contribution percentage to the total world level emissions.
- Scale data with Standard Scaler. We will try with and without outliers, to see how differently the ideal k numbers is calculated.

5. Calculate elbow.

For ideal elbow calculation we will use the function CALCULATE_ELBOW, we will draw the curve to better see the number of clusters for the model fit.

We will evaluate the elbow with hierarchical clustering and the agglomerative clustering algorithm.

It is important to have the best combination of final features to pass into the cluster number calculation. Total population will distort our final model, and the new features that takes in consideration the per capita will help us to better group countries. All this said, it is necessary to have the final set of features after an iteration process, so that we can have the best possible set that will group the countries according to emissions and other characteristics.

6. KMeans fit

After elbow calculate on, final Kmeans fit will be done for 7 clusters.

7. Segmentation emissions indicators.

It is time for finding the best emissions indicators to group the countries, so that we can better explain each cluster. As an example: GDP_capita, emissions_capita and consumption_CO2.

8. Clusters profiling.

We will generate a cluster summary card, according to the next indicators: GDP 19 per capita, Production per capita, Emissions per capita and Consumption CO2 emissions per capita.

Consumption CO2 emissions per capita shows CO2 emissions adjusted after taking in consideration trade.

Final clustering summary looks like this:

Cluster 1 (20 countries):

- Highest GDP per capita.
- Medium value for production per capita.
- Low level for emissions per capita.
- High level of adjusted emissions after consumption.

Cluster 2 (37 countries):

- High GDP per capita.
- Lowest Production per capita.
- Lowest Emissions per capita.
- High values for adjusted emissions. Countries that import a lot.

Cluster 3 (73 countries):

- Lowest GDP per capita.
- Low Production per capita.
- Low Emissions per capita.
- Lowest adjusted Emissions by consumption.

Cluster 4 (27 countries):

- Very low GDP per capita.
- Low Production per capita.
- Medium level Emissions per capita.
- High level of adjusted emissions. These are countries that import a lot..

Cluster 5 (33 countries):

- Medium GDP per capita.
- Highest Production per capita.
- Second level of Emissions per capita.
- Low level of adjusted emissions. More export than import.

Cluster 6 (2 countries):

- High GDP per capita alto.
- Medium Production per capita.
- Low Emissions per capita.
- Low level of adjusted Emissions.

Cluster 7 (5 countries):

- Low GDP per capita.
- High Production per capita.
- Highest Emissions per capita.
- Medium adjusted Emissions for consumption.