Are you living in a happy country?

Aleksandra Leszczyk

Abstract

We use clustering methods to find similar countries based on features that are important for well-being of the citizens.

Keywords

world happiness score – K-means — hierarchical clustering — DBSCAN

1. Objective

We use clustering methods to identify countries which have a similar level of self-reported happiness, economical status, social support, healthy life expectancy, life opportunities, generosity, and reported level of corruption. This way, we can target better marketing campaigns of global companies since people in different countries can have different needs. Also, we can plan efficiently to extend brand coverage.

2. Description of data set

This work uses data from *World Happiness Report 2020* by Helliwell *et al.* published in *New York: Sustainable Development Solutions Network*. We focus on data from the 2019 year that describes a world before the COVID-19 pandemic started. The data set contains 144 observations of 8 features. The features and their data types are listed in the table below:

Feature	Data type
Country name	nominal
Life Ladder	continuous
Log GDP per capita	continuous
Social support	continuous
Healthy life expectancy at birth	continuous
Freedom to make life choices	continuous
Generosity	continuous
Perceptions of corruption	continuous

The overall happiness (as estimated by respondents) is denoted as ”Life ladder”, while other features describe the general characteristics of the country in which respondents live. However, not all happiness scores is explained by these other factors. Thus, we include the self-estimated happiness as a feature characterizing a country, Sample data (transposed) for three countries is presented as follows:

Country name	Afghanistan	Albania	Algeria
Life Ladder	2.375	4.995	4.745
Log GDP. . .	7.697	9.544	9.337
Social support	0.420	0.686	0.803
Healthy life. . .	52.400	69.000	66.100
Freedom. . .	0.394	0.777	0.385
Generosity	-0.108	-0.099	0.005
... corruption	0.924	0.914	0.741

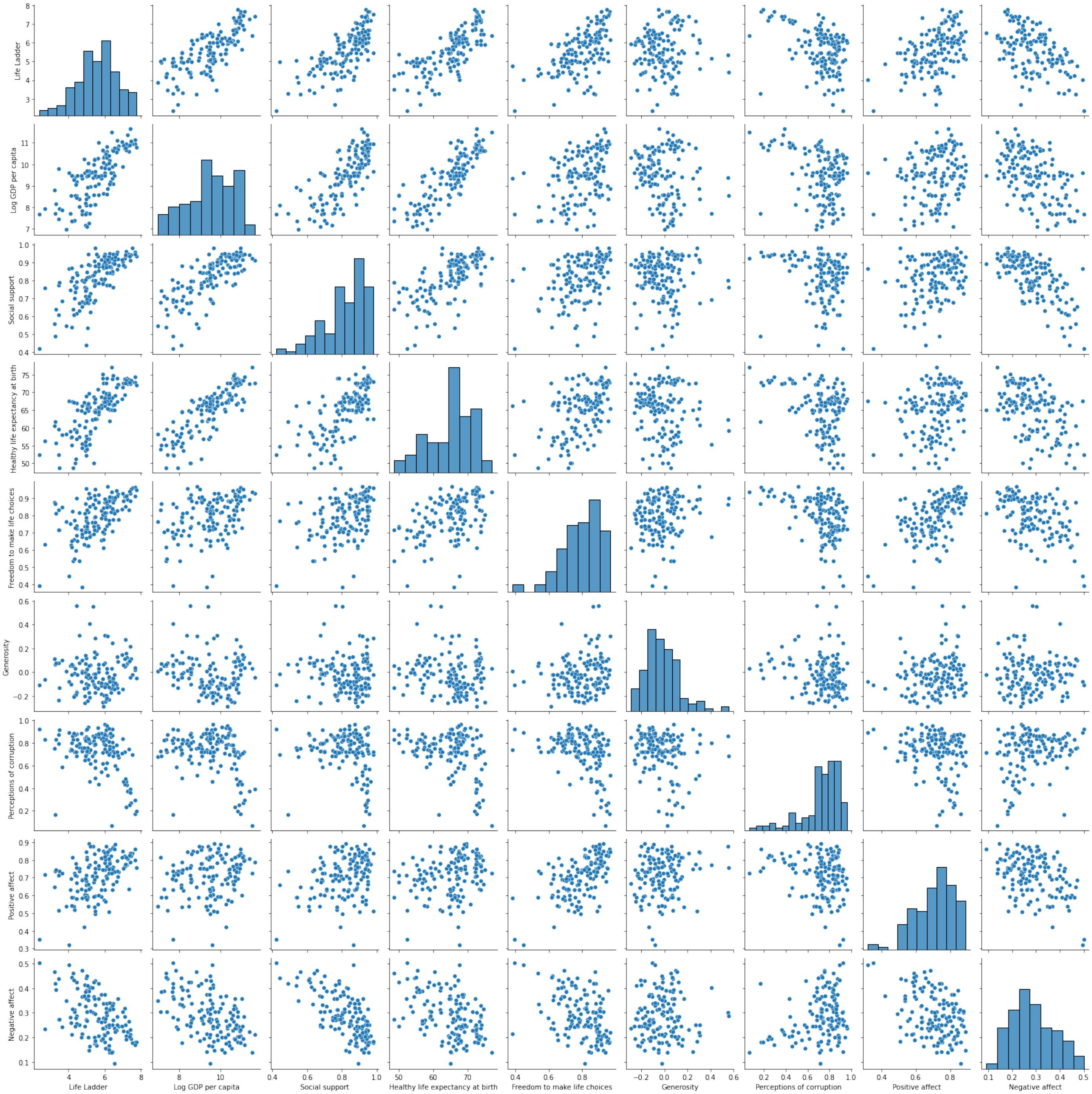
3. Exploratory data analysis

We counted a number of non-null values, mean values and standard deviations of features, their minimum and maximum values, and skews of distributions.

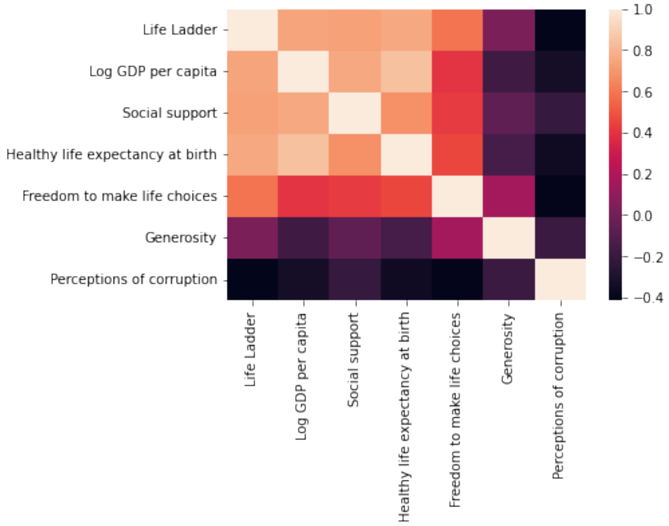
	count	mean	std	min	max	skew
Life Ladder	144	5.57	1.11	2.38	7.78	-0.26
Log GDP. . .	138	9.48	1.14	6.97	11.65	-0.33
Social support	144	0.82	0.12	0.42	0.98	-1.04
Healthy life. . .	139	65.0	6.7	48.7	77.1	-0.56
Freedom. . .	143	0.79	0.12	0.39	0.97	-0.94
Generosity	137	-0.02	0.15	-0.29	0.56	1.03
... corruption	136	0.72	0.19	0.07	0.96	-1.48

The distributions of variables are presented in histograms, and pairwise scatterplots are shown in Figure 1. Some distributions are not normally distributed, e.g. generosity, of perceptions of corruption. We can observe that some variables are correlated. For example, we observe that the life ladder (e.g. self-estimated happi-

Figure 1. Pairwise scatterplots and histograms of numerical variables.

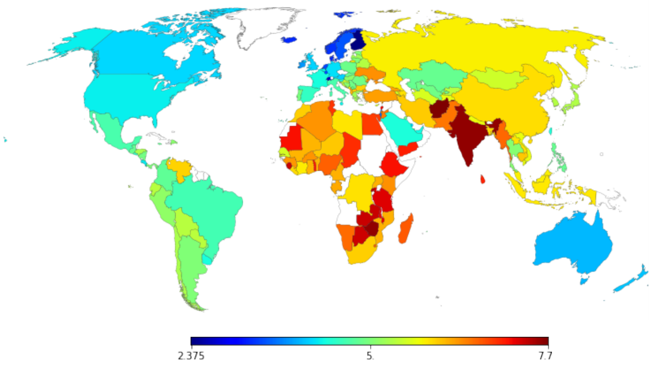


ness) is positively correlated with GDP per capita, social support, and healthy life expectancy at birth. The correlations between features are presented in a heatmap below.



What is unexpected here, is a negative correlation between life ladder and generosity. We learn from Dicken's *A Christmas Carol* that generous people are happy people. This data shows the opposite trend. Does it mean that people are happier if the redistribution is done by national institutions rather than single kind people? Or maybe, generosity is not necessary in countries that provide conditions supporting overall happiness?

Figure 2. Life ladder (happiness score) distribution.



The geographical distributions of selected variables presented in Figures 2–5. The happiest people live in wealthy Scandinavia while the most unhappy situation is in Afghanistan. In general, people appraise their generosity rather high. One of the exceptions is Japan which is known for rejecting the idea of tips. The belief in widespread corruption is prevalent in almost all countries.

Figure 3. Generosity.

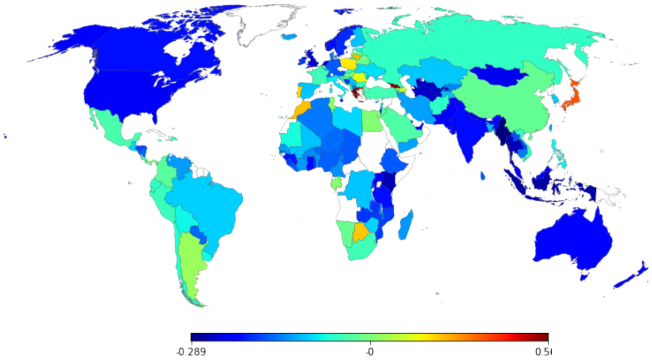


Figure 4. Perceptions of corruption.

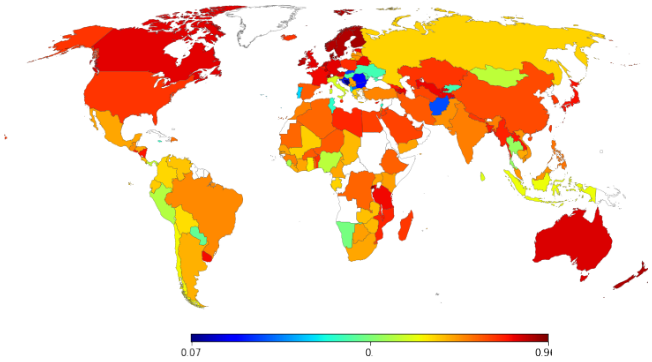
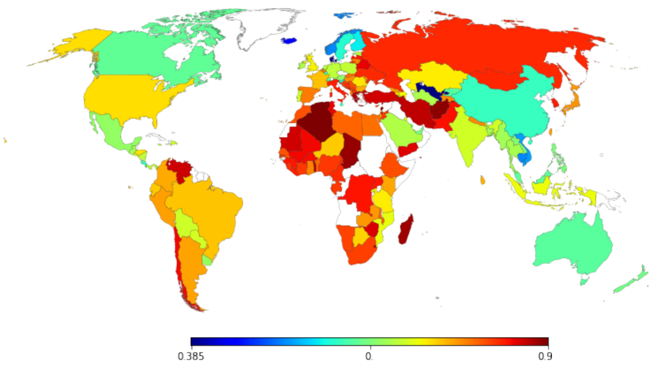


Figure 5. Freedom to make life choices.

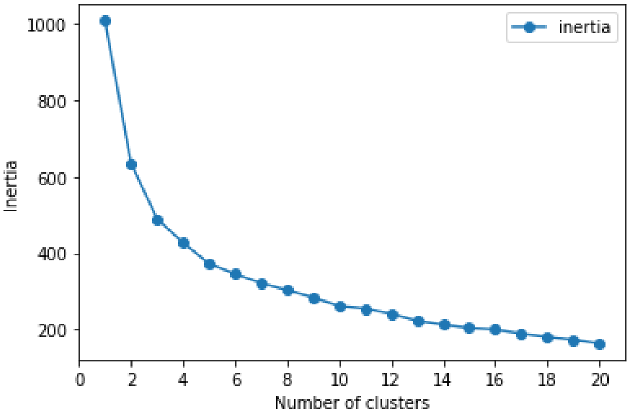


4. Data cleaning and engineering

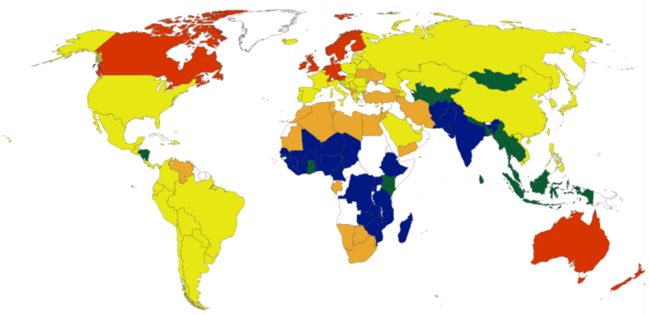
As the number of null values is low, and there is no column filled mainly by null values, filling the null values with mean values for given columns allows us to use the clustering method without losing accuracy. Some distributions are skewed, and, thus, we perform a log transformation on columns with absolute skew higher than 0.75. In the end, we use standard scaling on data to provide reliably estimated distances between observations. We do not add any feature as it does not add any benefit to the analysis.

5. K-Means model

K-Means model is an efficient method to find spherical and even-sized clusters. As it does not determine an optimal number of clusters, we fitted models with a different number of clusters and computed inertia for each model. The results are presented as follows:

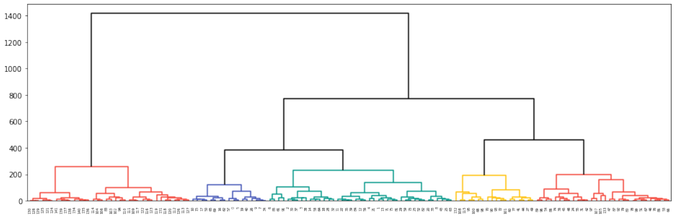


We observe no evident shoulder in a plot, but a value of 5 is a reasonable choice for a number of clusters. The world divided into five parts is shown in the following map.

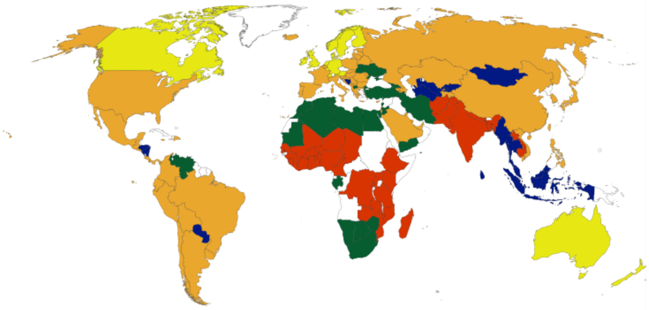


6. Hierarchical agglomerative clustering

The hierarchical agglomerative clustering method is better for capturing the uneven-sized clusters than a K-means model. Here, we used an inertia-based distance measure to obtain the following hierarchy of links



As a minimum number of clusters, we selected the same number which we obtained with the K-means model, which is 5. With the HAC algorithm, the world has been divided in a similar way as in the case of the K-means model.



7. DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) allows us to group points that are closely packed. The low-density regions are treated as outliers. This model requires to tune two parameters: ε – the maximum distance between two samples for one to be considered as in the neighbor, and N – a number of samples in a neighborhood for a point to be considered as a core point. We performed tests for $\varepsilon \in \{1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0\}$ and $N \in \{4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\}$. We found out that the optimal values are $\varepsilon = 1.3$ and $N = 5$ which allow us to separate four clusters and 44 outlier countries. The results are presented in a map below. Note that red color indicates outliers.

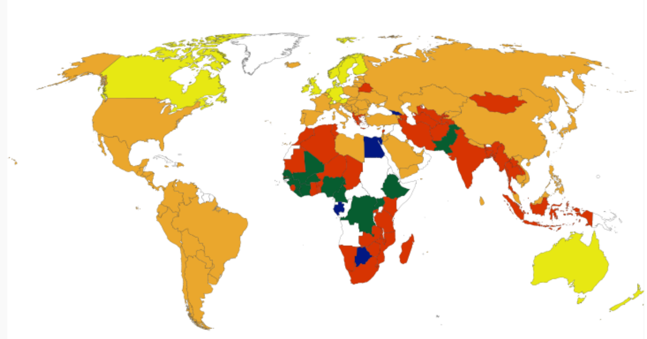


Figure 6. Pairwide scatterplots and histograms of numerical variables classified by K-means algorithm.

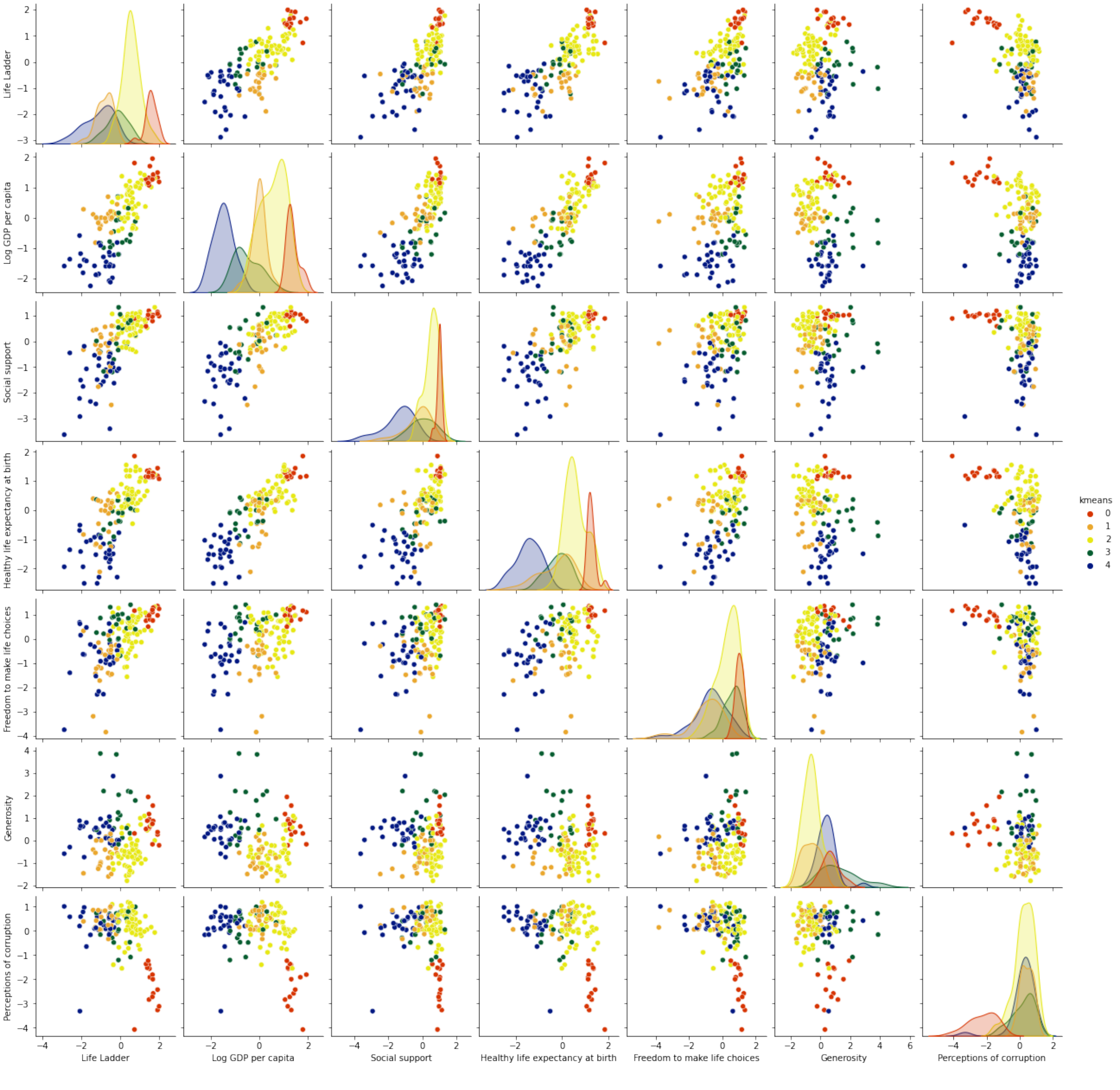


Figure 7. Pairwide scatterplots and histograms of numerical variables classified by HAC algorithm.

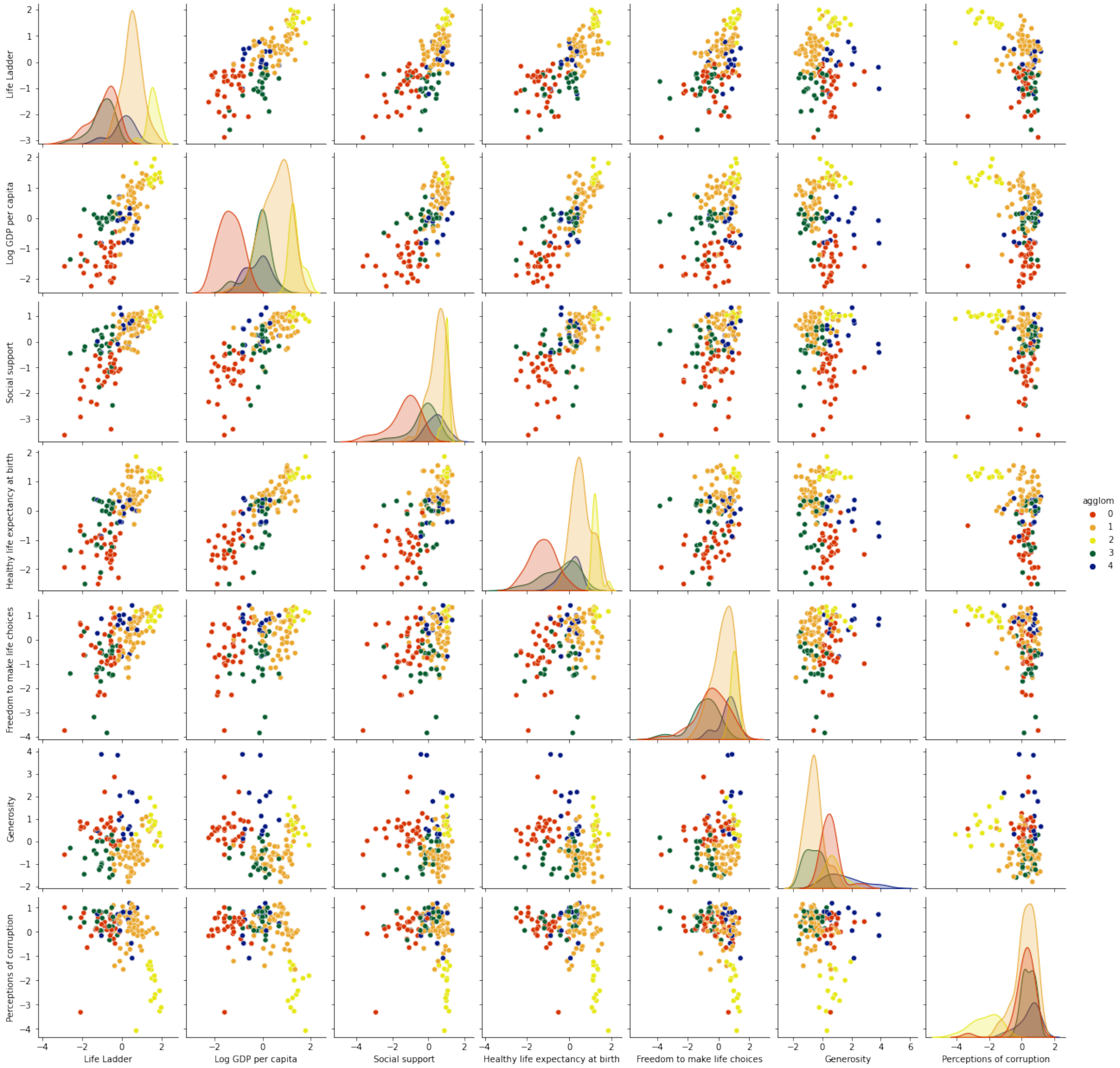
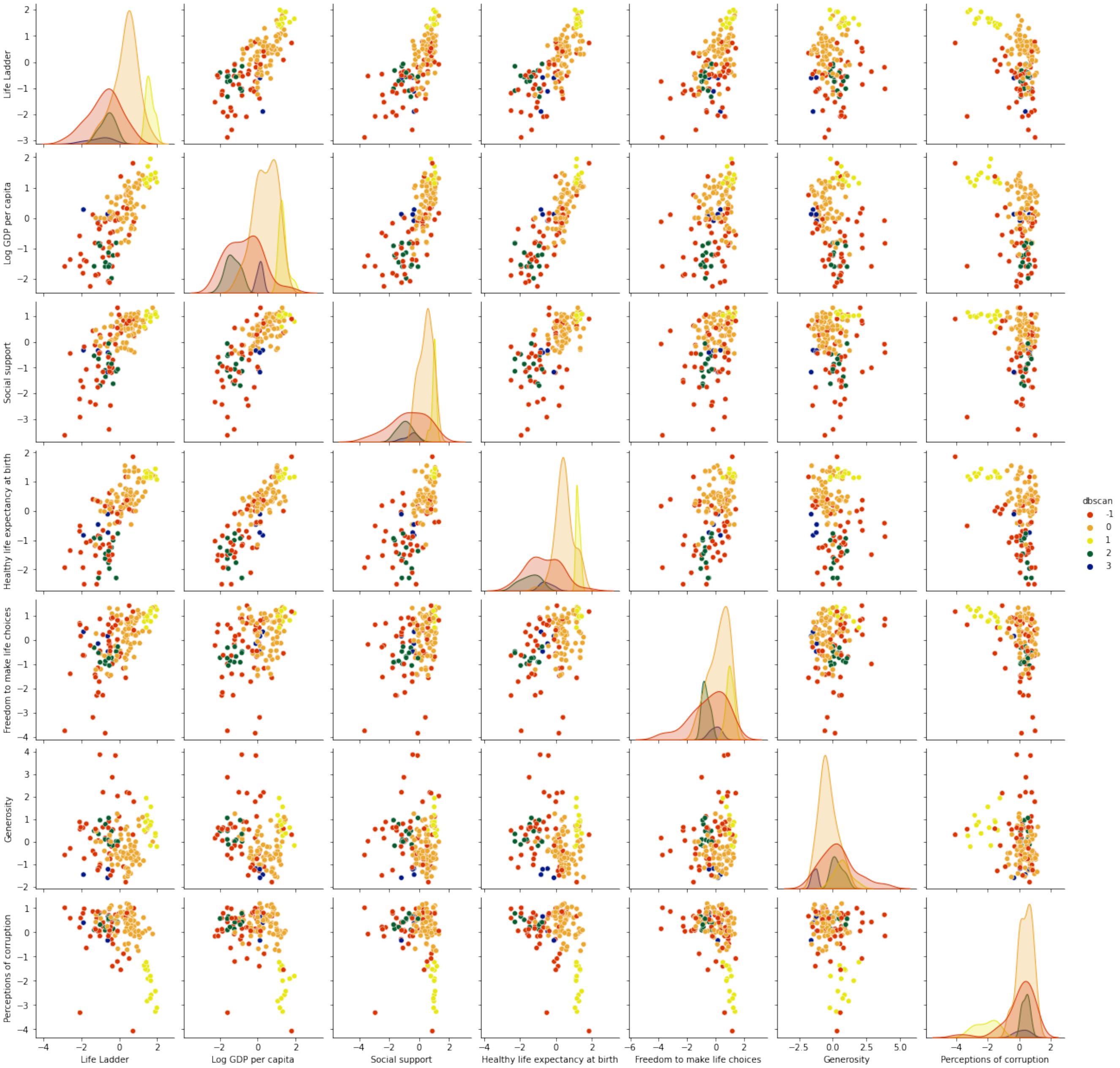


Figure 8. Pairwide scatterplots and histograms of numerical variables classified by DBSCAN algorithm.



8. Model comparison

We group similar countries using the K-means algorithm, hierarchical agglomerative model, and DBSCAN. We obtain similar results with the K-means and HAC algorithms. The K-means model provides distributions of features in most separable way as it is presented in Figure 6 (compare with Figure 7 and 8). DBSCAN is sensitive to parameters, and finding a set that yielded a relatively low number of outliers and a reasonable number of clusters required several tests. Also we cannot be sure that the same parameters will be the optimal ones for a different set of data. Thus, we recommend the K-means method for this kind of analysis as it combines efficiency and good agreement with other methods.

9. Summary and key findings

The K-means and HAC models allows us to group a world into five similar regions which are:

1. China, Russia, USA, Saudi Arabia, Japan, South America, western, eastern and southern Europe
2. Central Africa and Indian Peninsula,
3. Middle east countries with Southern and northern Africa,
4. Canada, Scandinavia, Germany, UK, and Australia
5. Indonesia with some central Asia countries

This partitioning is reasonable as we can distinguish the mainstream group of influential countries (e.g. USA, China, Russia) and influenced countries. We can distinguish countries that combine wealth, social support and geographical location that is either isolated or in unpleasant climate. We see the countries that are challenged by poverty and overpopulation.

DBSCAN method provided four regions and 44 outlier countries. This fact indicates that the countries do not form uniform clusters in the space of the tested features. Thus, we need to treat carefully the countries that are close to borders of clusters defined by other algorithms.

10. Next steps

The analysis provides interesting results indicating the groups of similar countries. Although there is no underlying structure and the borders between different groups are not solid, we can make conclusions from this clustering. However, the current COVID-19 pandemic influenced a dynamics of social processes. New problems and system solutions change the living conditions dramatically. The features we tested can be insufficient to estimate the similarities between countries during pandemic. Thus, we recommend to include factors that describe the strength of relationships, traveling barriers, and work stability.