

# Suicide Rates Overview 1985 to 2016

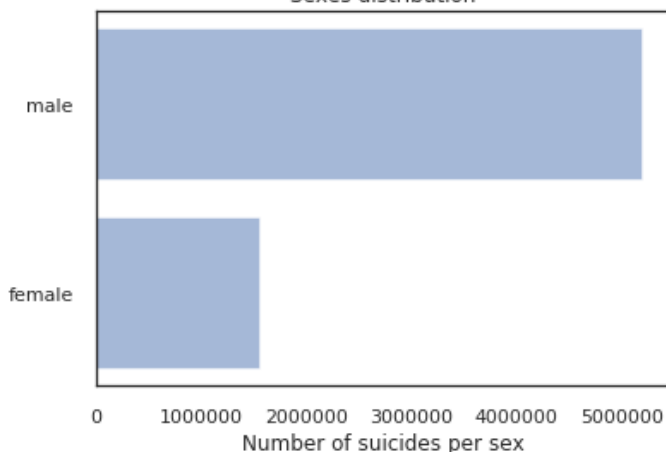
Compares socio-economic info with suicide rates by year and country

## DESCRIEREA DATASETULUI

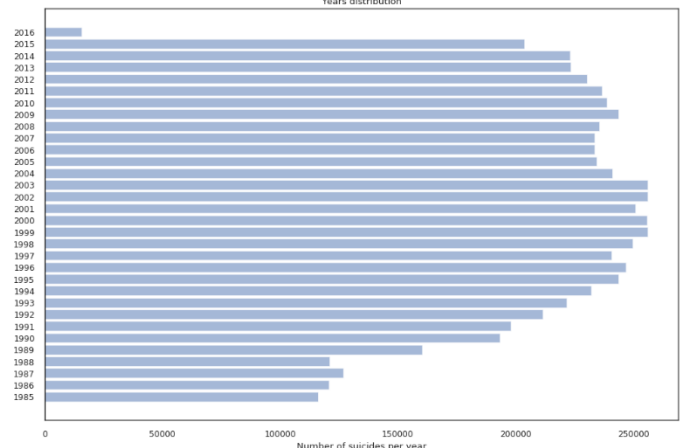
Datasetul este format din 27820 exemple si 12 features: country, year, sex, age, suicides\_no, population, suicides/100k pop, country-year, HDI for year, gdp\_for\_year (\$), gdp\_per\_capita (\$) si generation. Aceştia reprezintă o serie de indicatori macroeconomici specifici fiecărei tari raportat la an, HDI, GDP, GDP/ capita si categoria de populație (descrisa prin sex, categorie de vârstă, generație si rata de suicid). Cum ceea ce ne interesează este rata de suicid, de acum in colo vom considera acest feature drept unul de referință. In continuare vom prezenta o serie de statistici despre datasetul ales.

In figura de mai sus este plotata distribuția suicidurilor in funcție de tara (cu mențiunea ca nu sunt prezente toate tarile analizate ci doar cele care au număr de suicid peste medie.<sup>1234</sup>

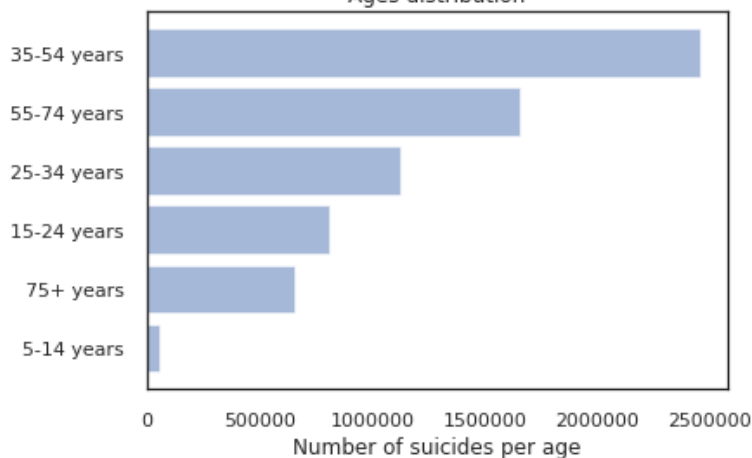
Sexes distribution



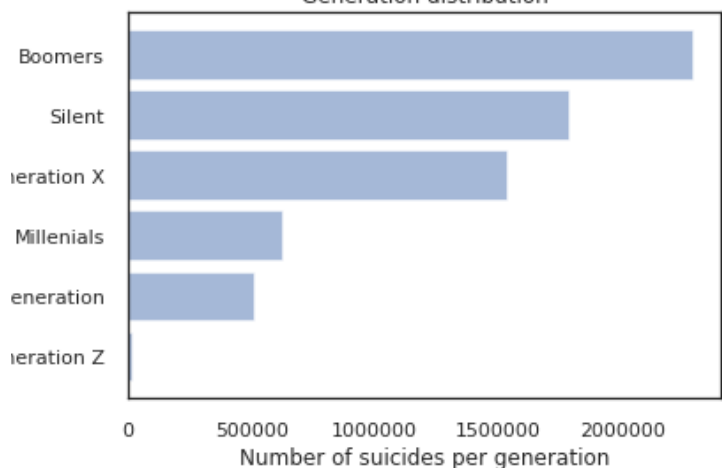
Years distribution



Ages distribution



Generation distribution



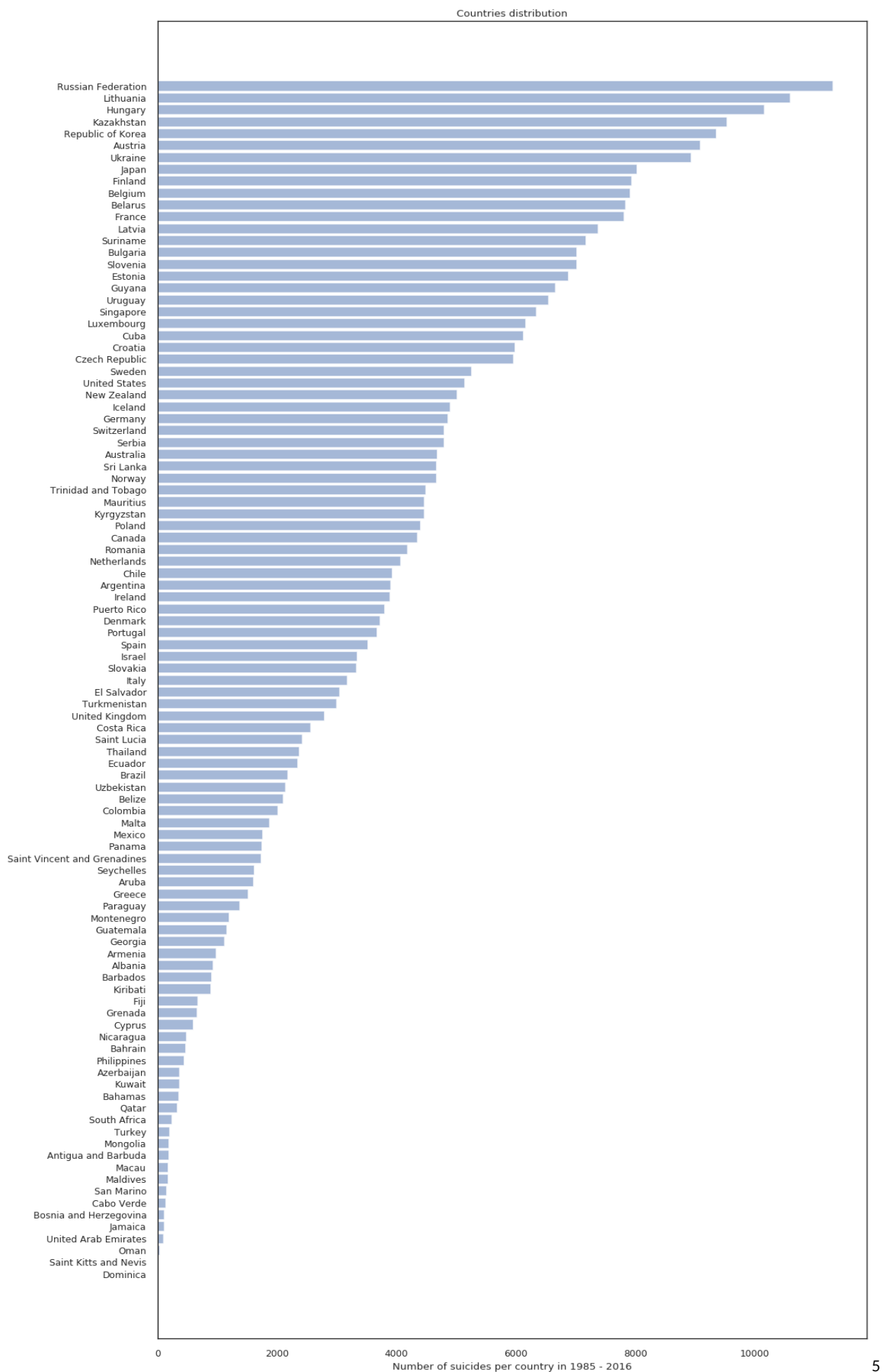
Se poate observa cu ușurință ca distribuția anilor si distribuția generațiilor sunt corelate. Generația Z (> 1995) corespunde populației cu vârsta între 5-14 ani, etc.

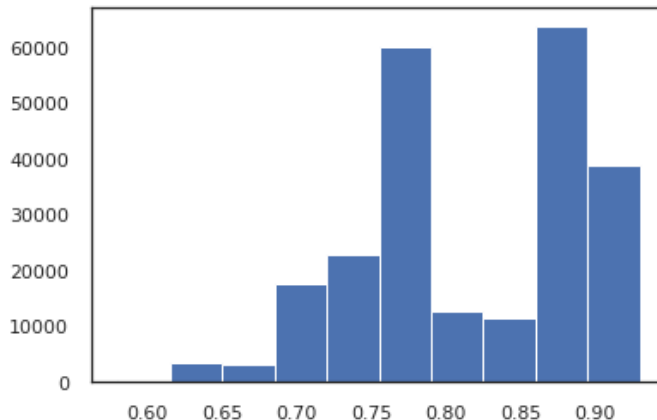
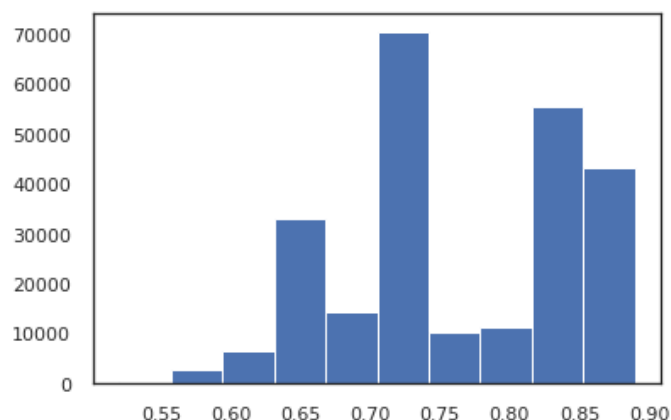
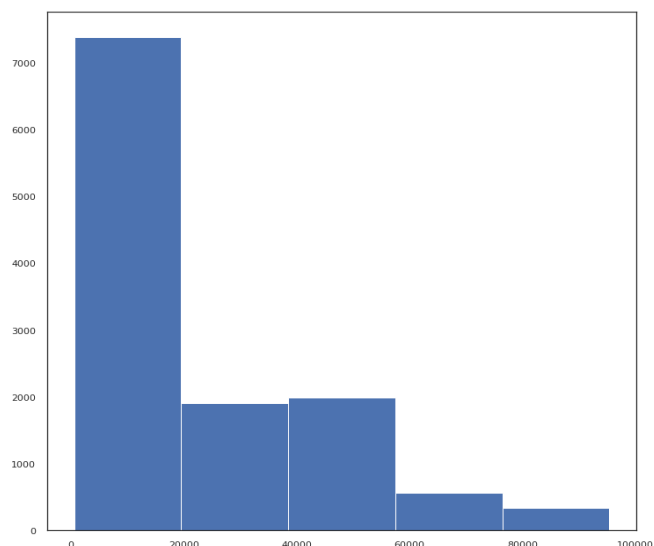
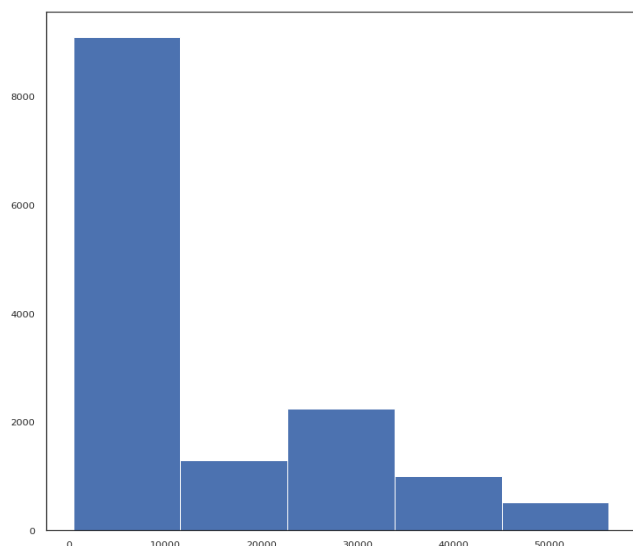
<sup>1</sup> Distribuția sinuciderilor in funcție de sex

<sup>2</sup> Distribuția sinuciderilor in funcție de an

<sup>3</sup> Distribuția sinuciderilor in funcție de categoria de vârstă

<sup>4</sup> Distribuția sinuciderilor in funcție de generație





## REDUNDANTA COLOANELOR

In prima faza se arata ca pentru absolut toate datele  $\text{suicides\_no} * 100000 / \text{population} == \text{suicides}/100k$ , lucru care nu contrazice intuiția in nici un fel.

Apoi se demonstrează ca  $\text{GDP} / (\text{GDP}/\text{capita}) - \text{population} == 0$ , pentru toate datele exceptând o fracțiune de 0.69%. Datele din fracțiunea de 0.69% provin de la diverse tari din anul 2016, ultimul an al data setului, si anul cu date incomplete. Menționam ca in continuare vom păstra doar una din perechile de caracteristici redundante (anume  $\text{suicides}/100k$  si  $\text{GDP}/\text{capita}$ ).

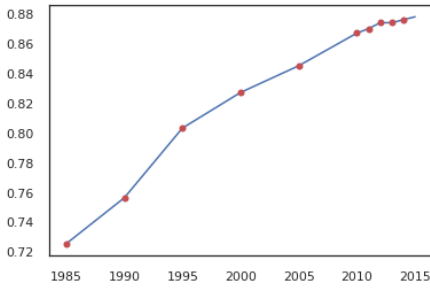
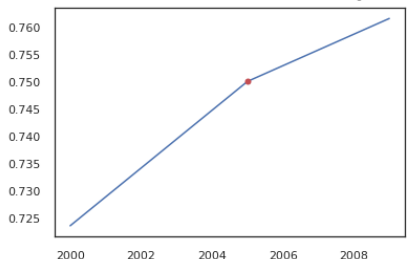
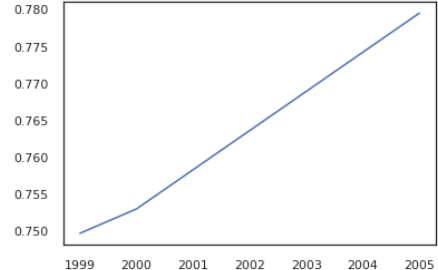
## EXTRAPOLAREA HDI-ului

O alta observație notabila legata de dataset este legata de absenta datelor in coloana HDI, exista doar 8364 date din cele 27820 necesare. Intuitiv caracteristica HDI (Human Development Index) este una importanta care influențează rata de suicid a populației dintr-o anumita tara. Deci va trebui sa păstram aceasta coloana. In continuare vom extrapola valorile lipsa astfel:

Daca pentru o tara avem mai mult de o valoare pentru HDI intr-un an atunci extrapolând aceste puncte cu polinoame de	Daca pentru o tara avem mai exact o valoare pentru HDI intr-un an atunci shiftam HDI-ul global mai sus sau mai jos astfel încât HDI-ul global sa fie	Daca pentru o tara nu cunoaștem HDI-ul din nici un an atunci vom considera pentru fiecare an in parte HDI-ul global din acel an.
--	--	--

<sup>6</sup> Distribuția suicidelor in funcție de PIB (stânga 1996, dreapta 2006)

<sup>7</sup> Distribuția suicidelor in funcție de HDI (stânga 1996, dreapta 2006)

gradul 1 se obține HDI-ul pentru restul anilor	egal cu HDI-ul pe care îl avem în anul respectiv. HDI-urile listate vor fi HDI-urile globale ale anilor respectivi shiftate în același mod.	
<p>Spain</p> 	<p>Montenegro</p> <p>[2000 2001 2002 2003 2004 2005 2006 2007 2008 2009]</p> <p>[nan nan nan nan nan 0.75 nan nan nan nan]</p> <p>[0.72352632 0.72882105 0.73411579 0.73941053 0.74470526 0.75288697 0.75577393 0.7586609 0.76154787]</p> 	<p>San Marino</p> <p>[1999 2000 2005]</p> <p>[nan nan nan]</p> <p>[0.74965414 0.75296053 0.77943421]</p> 

## PREPROCESAREA DATELOR

Deoarece în dataset avem date numerice, secvențiale și categoricale, va trebui să facem o preprocesare a acestuia, astfel:

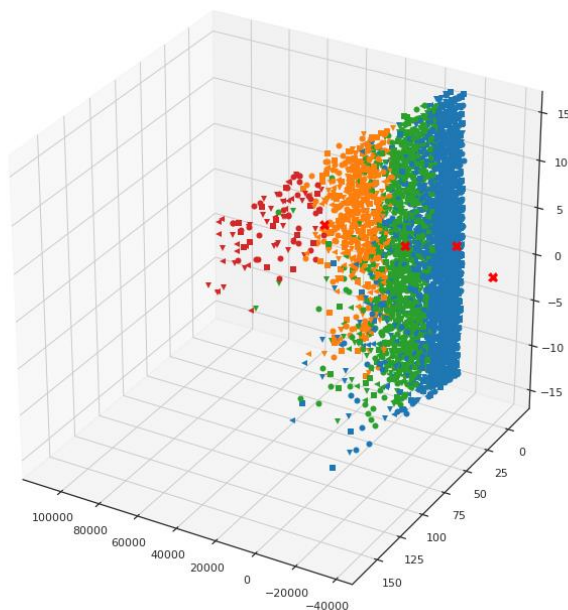
- **country** – caracteristica categorică. Pentru această caracteristică vom encoda țările în vectori OneHot (echivalent vom adăuga încă 101 de features (corespunzătoare celor 101 de țări) la datasetul nostru)
- **year** – data secvențială naturală, o vom lăsa neschimbată
- **age** – feature ordinal. Pentru acesta vom encoda categoriile de vârstă astfel: '15-24 years': 0, '25-34 years': 1, '35-44 years': 2, '45-54 years': 3, '55-64 years': 4, '65+ years': 5. Acest mod de encodare păstrează distanța mare între categoria 75+ de ani și categoria 15-24 de ani și distanța mică între 15-24 ani și 25-34 de ani.
- **generation** - feature ordinal. Echivalent cu **age**. Vom encoda astfel, din motivele de mai sus: 'Boomers': 2, 'G.I. Generation': 0, 'Generation X': 3, 'Generation Z': 5, 'Millennials': 4, 'Silent': 1.

## METODELE FOLOSITE

În continuare, deoarece nu avem nici un fel de label asupra datelor noastre vom dori să clusterizăm țările în funcție de indicatori macroeconomici și de rata de suicid specifică populației din țara respectivă.

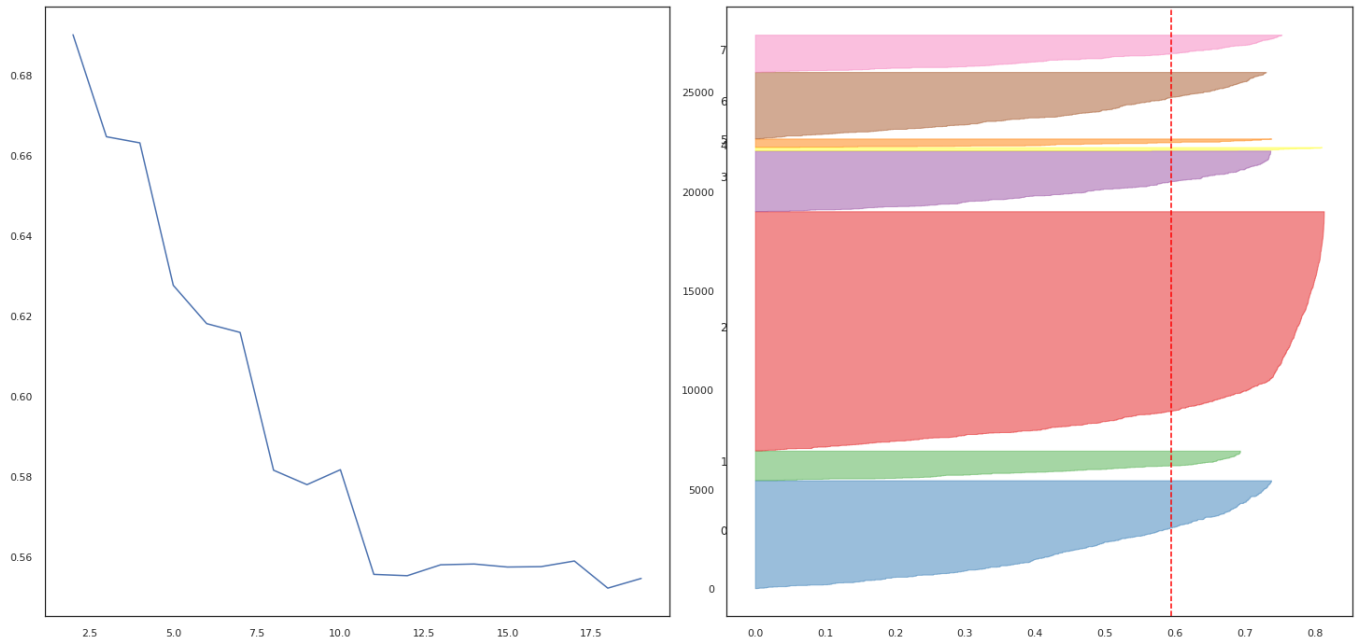
### 1. K-MEANS

Inițial pornim cu 4 clase apoi dorim să găsim numărul optim de clase care separă cât mai bine clusterurile găsite de algoritmul K-means. În acest caz avem un Silhouette score de 0.663 ceea ce

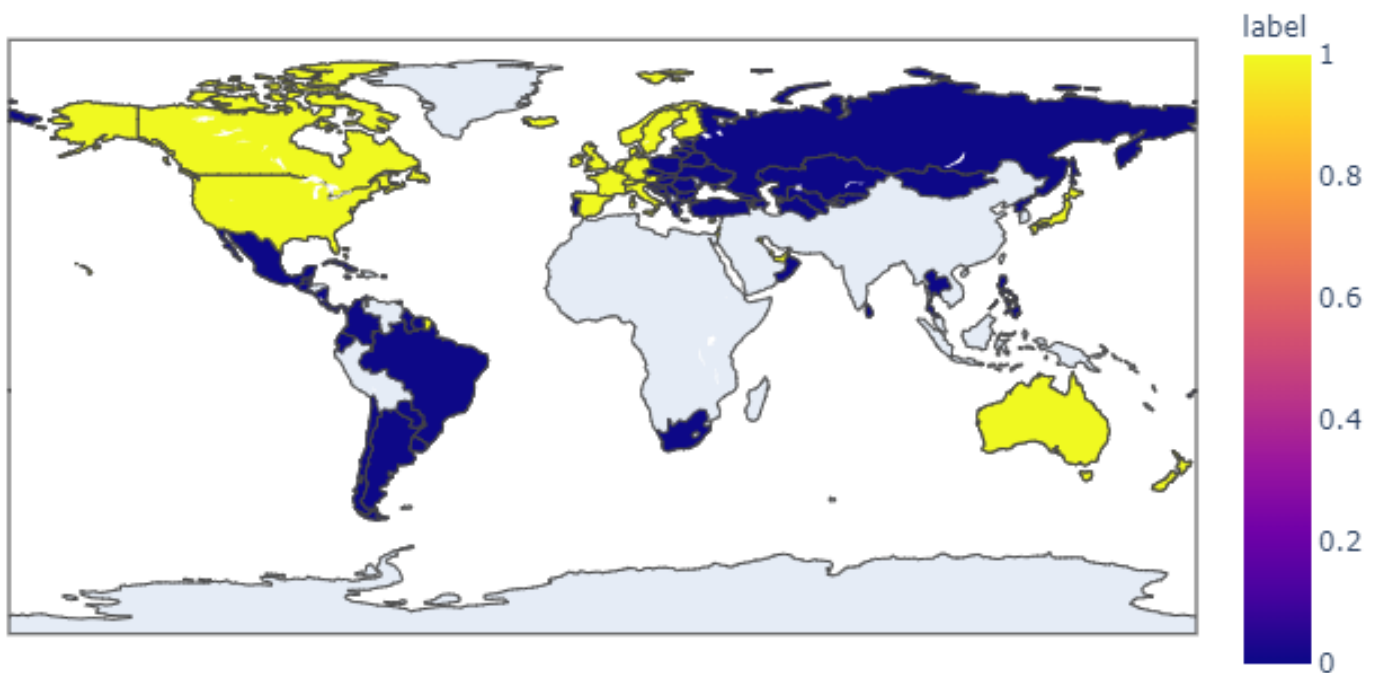


înseamnă ca clusterurile sunt bine separate (1 cluster dense și separate și 0 overappuire totală și -1 atunci când clusterurile sunt incorecte).

Apoi efectuând un gridsearch a numărului de cluster după Silhouette score obținem ca numărul optim de cluster este 2 (silhouette score = 0.69).



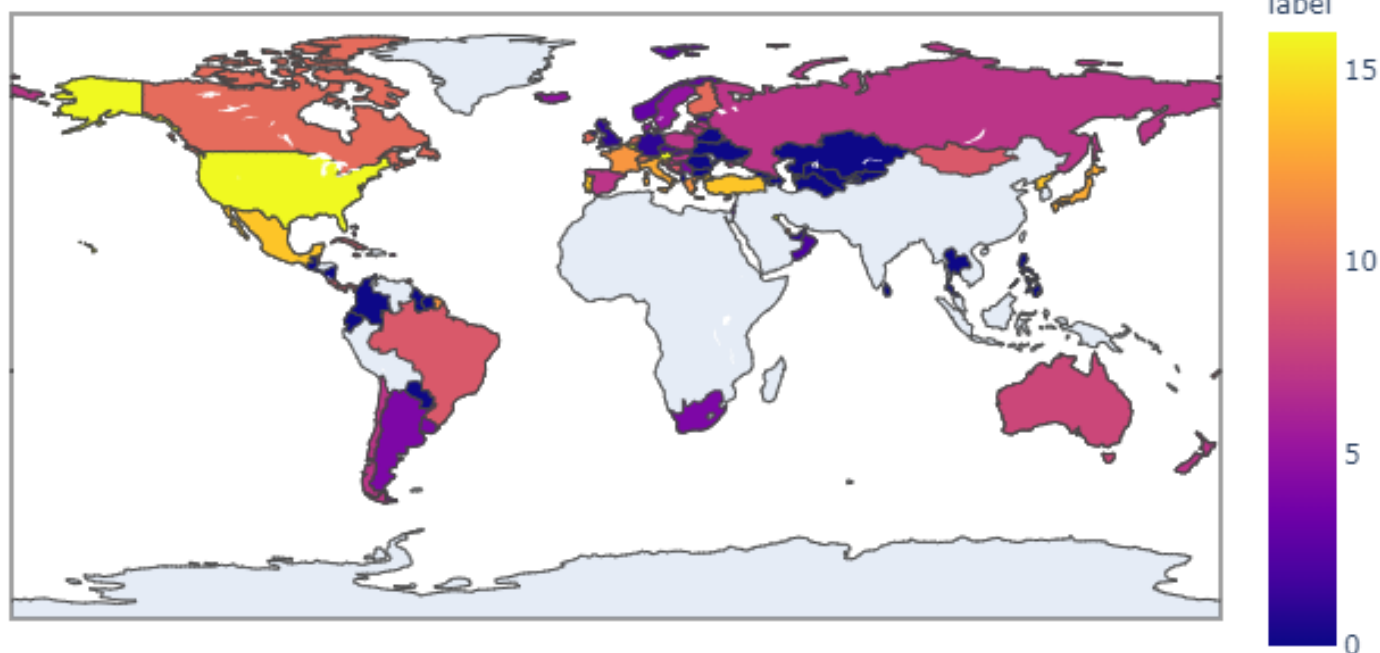
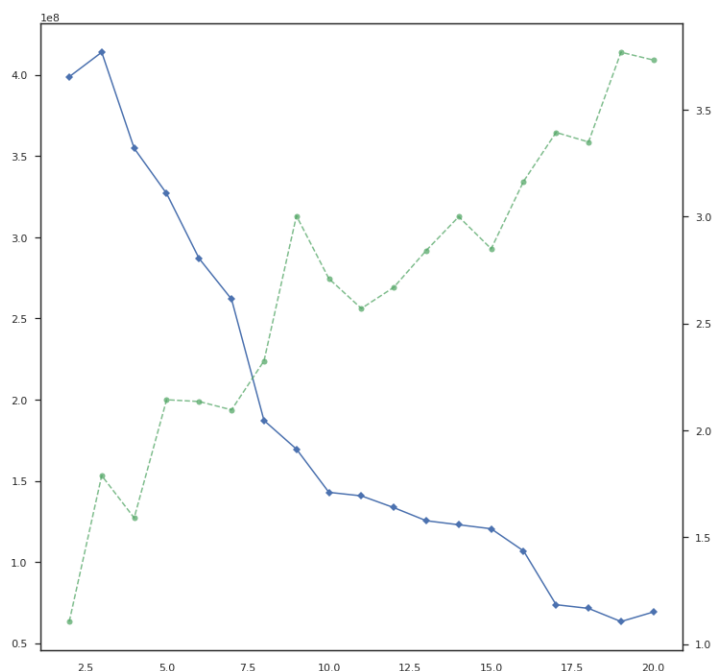
Pentru numărul optim de cluster obținut mai sus (adică 2) obținem divizarea Țărilor în două clase așa cum se poate observa mai jos. Se poate observa că Țările dezvoltate ca: SUA, Canada, Australia, Japonia și cele din vestul Europei sunt în clasa 1 și celelalte în clasa 2. Intuitiv această clusterizare împarte Țările în funcție de gradul de dezvoltare economică și cum influențează indicatorii macroeconomici rata de suicid a populației. Cum nu avem exemple (în figura de mai jos) care contrazic afirmația anterioară rezultă că rata de suicid este influențată de indicatorii macroeconomici.



Apoi efectuand metoda ELBOW putem observa ca punctul de la care un numar mai mare de clustere nu explica mai bine datele este 17.

Pentru acest numar de clustere (17) se poate observa o impartire a tarilor in 17 clustere exact ca mai jos. Aceasta impartirea este una mai complexa si secifica. Asociind urmatoarele tari impreuna:

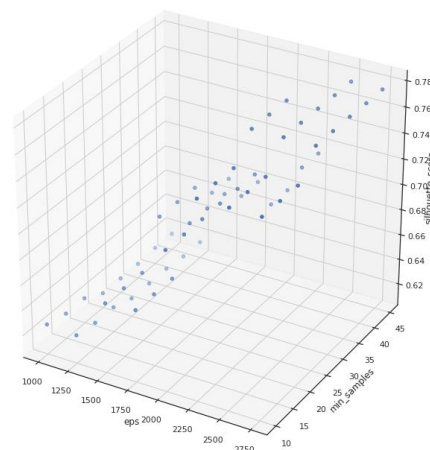
- 1: ['Germany', 'United Arab Emirates', 'United Kingdom'],
- 3: ['Norway'],
- 5: ['Denmark', 'Iceland', 'Sweden'],
- 6: ['Bahamas', 'Cyprus', 'Puerto Rico'],
- 8: ['Australia', 'San Marino', 'Singapore'],
- 10: ['Canada', 'Finland', 'Ireland', 'Netherlands'],
- 11: ['Qatar', 'Switzerland'],
- 12: ['France', 'Greece', 'Slovenia'],
- 13: ['Italy', 'Japan'],
- 15: ['Luxembourg'],
- 16: ['Austria', 'Belgium', 'Kuwait', 'United States']



## 2. DBSCAN

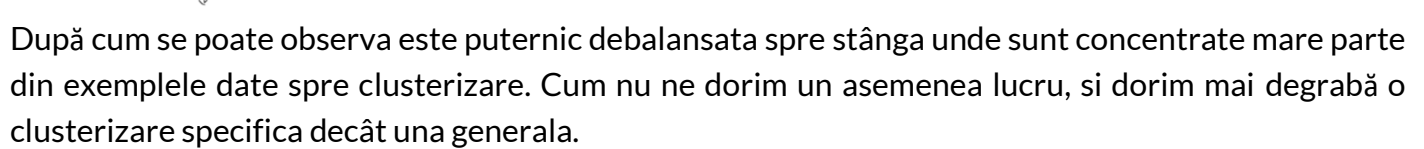
Inițial pornim la drum cu parametrii  $\text{eps}=1000$ ,  $\text{min\_samples}=50$ , parametrii care împart tarile in 4 clustere. Silhouette score-ul pentru împărțirea data de DBSCAN cu parametri de mai sus este de 0.572.

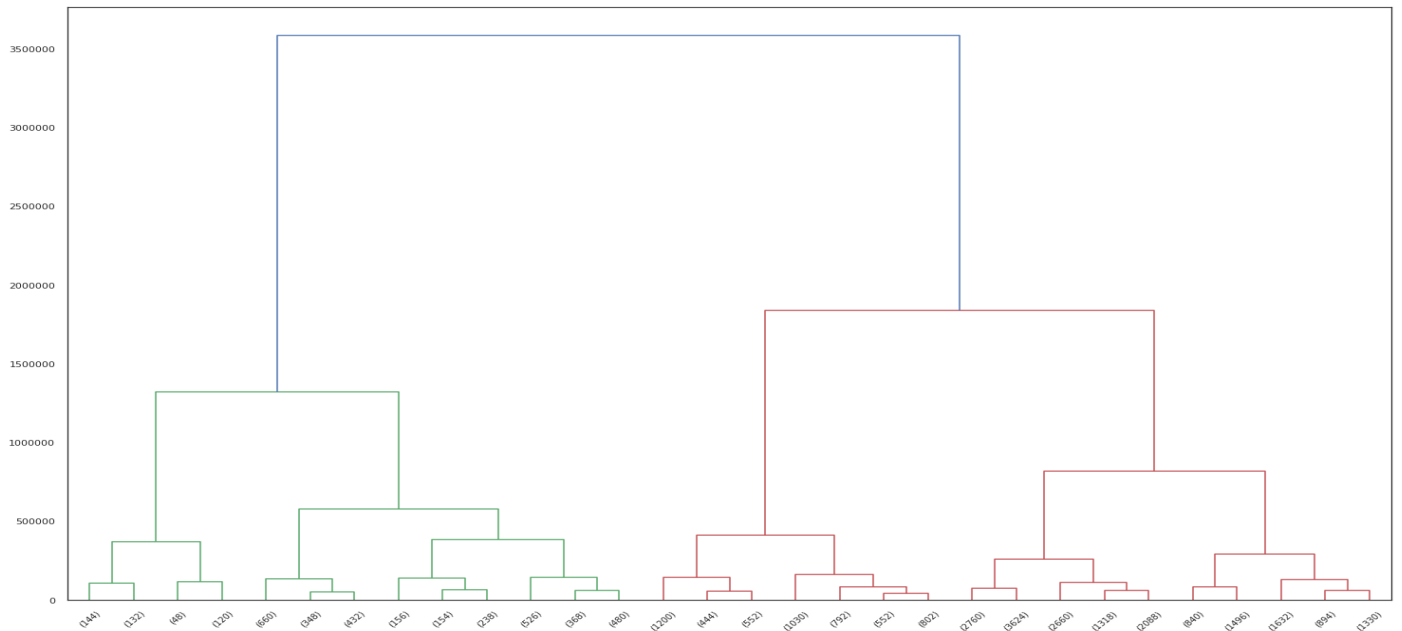
Apoi in urma unui grid search peste spațiul  $([1000, 2000] \cap M_{250}) \times ([10, 50] \cap M_{50})$  in funcție de Silhouette score si se obțin următorii parametri optimi pentru epsilon si



A world map showing the distribution of the 'label' variable. The map uses a color scale from 0 (dark blue) to 2 (yellow). Most landmasses are colored dark blue (0), while some regions in Europe and Asia are colored red/orange (1.5).

---





Dentograma de mai sus utilizează Ward linkage. Se poate observa ca este balansata. După un grid search după Silhouette score in vederea obținerii nivelului tăieturii se obține ca nivelul este 1 de unde vor rezulta 2 cluster (Silhouette score = 0.702). In continuare din motive de specificitate vom considera 9 cluster (Silhouette score = 0.609) pentru care obținem următoarea repartizare a tarilor.

0: ['Macau', 'Qatar', 'Singapore', 'Switzerland', 'Turkey'],

1: ['Luxembourg', 'Norway'],

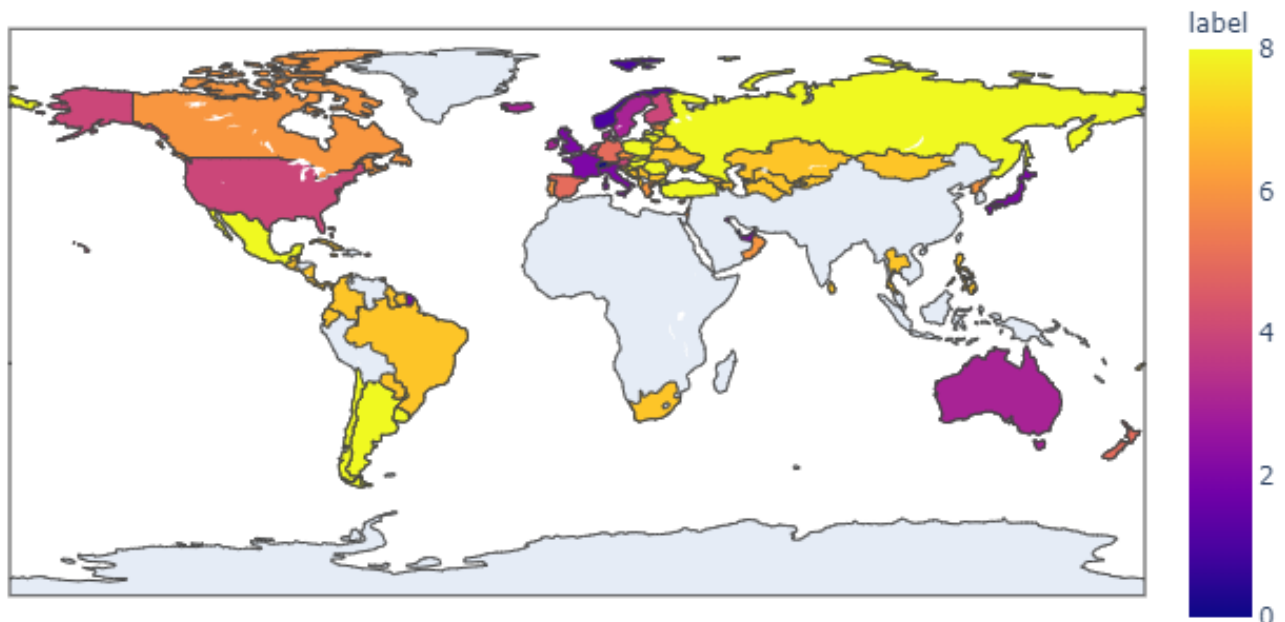
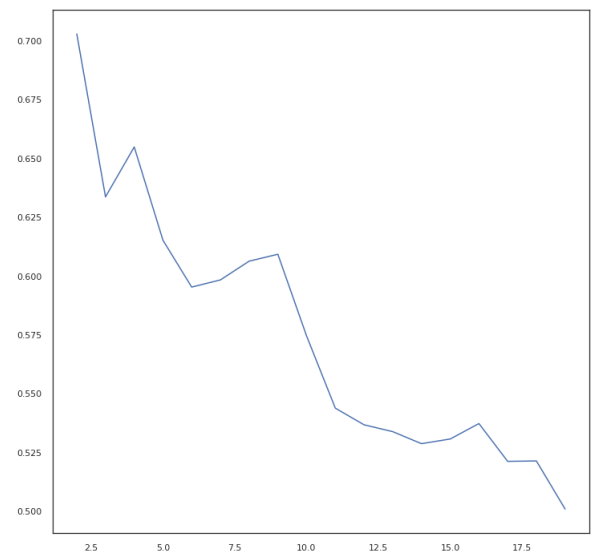
2: ['France', 'Italy', 'Japan', 'Kuwait', 'United Arab Emirates', 'United Kingdom'],

3: ['Australia', 'Denmark', 'Iceland', 'Ireland', 'San Marino', 'Sweden'],

4: ['Austria', 'Belgium', 'Finland', 'Netherlands', 'United States'],

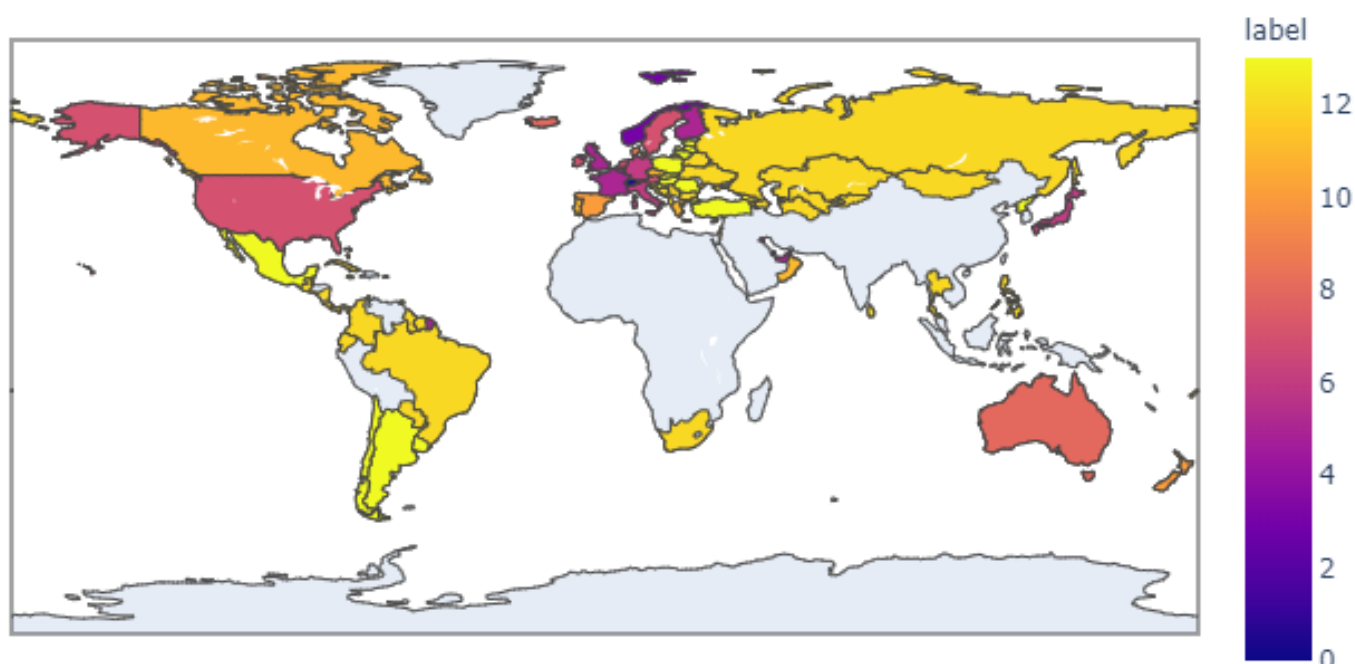
5: ['Bahamas', 'Cyprus', 'Germany', 'New Zealand', 'Puerto Rico', 'Spain'], etc

Se poate observa cu ușurință ca sunt respectate observațiile de la K-means.





Apoi clusterizand ierarhic folosind centroid linkage si căutând nivelele optim de taiere cu un grid search după Silhouette score obținem următoarea harta a lumii considerând 14 cluster (Silhouette score = 0.601)



## COMENTAREA REZULTATELOR

După cum se poate observa cea mai buna performanta luând in calcul doar Silhouette score-ul este de 0.776 pentru clusterizarea folosind DBSCAN de la care obținem 3 clase. Aceasta performanta este comparabila cu performantele tuturor metodelor analizate si este destul de aproape de 1 ceea ce denota ca clusterurile obținute sunt dense si nu se suprapun excesiv.

Algoritmul si parametrii	silhouette_score	calinski_harabasz_score	davies_bouldin_score
K-means(K = 2)	0.69	56457.28	0.5207
K-means(K = 17)	0.5587	361804.44	0.4752
DBSCAN(eps = 2250, min_samples = 25)	0.776	1722.41	5.6548
HCward(K = 9)	0.6091	173784.35	0.4845
HCcluster(K = 14)	0.6012	120861.73	0.4524

Conform metricii Calinski-Harabasz (Variance Ratio Criterion) (calinski\_harabasz\_score mare înseamnă clusteruri bine definite) algoritmul care are clusterurile cele mai definite este K-means(K = 17) apoi algoritmi de clusterizare ierarhica.

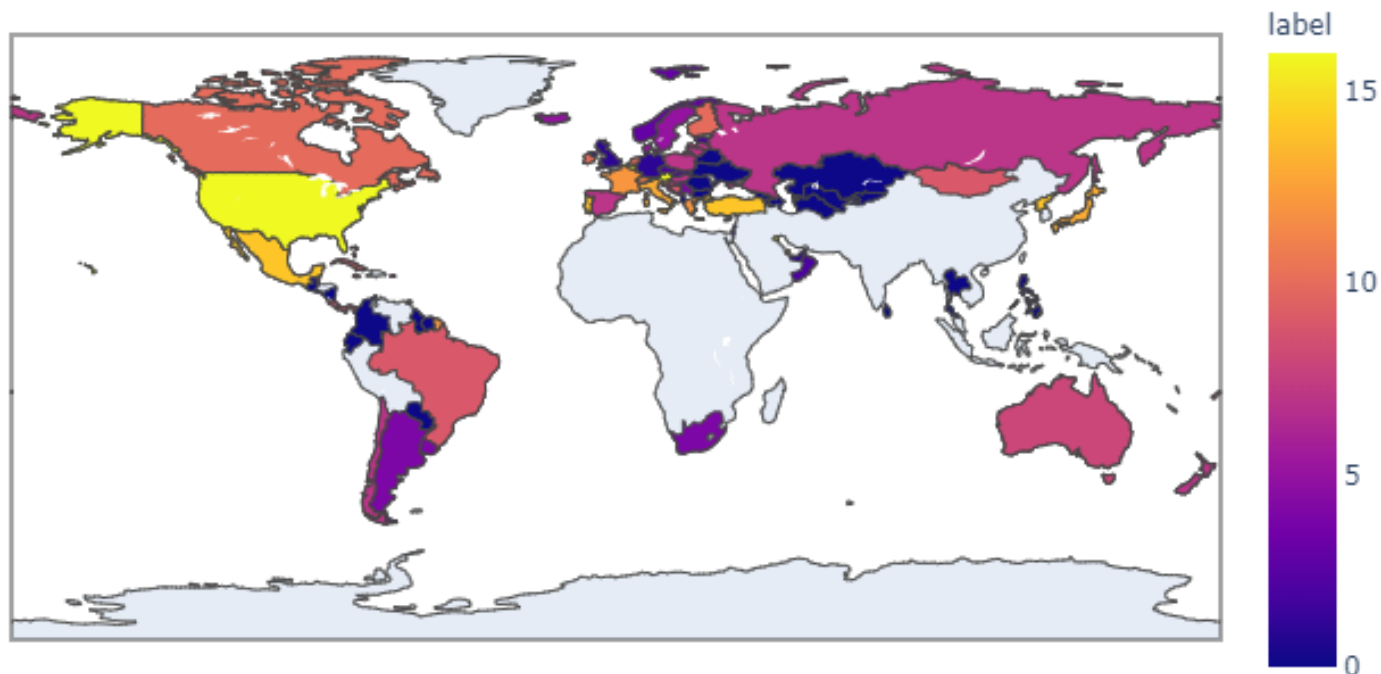
Conform metricii Davies-Bouldin (se măsoară media similarității dintre clusteruri, similaritate 0 înseamnă clusteruri total independente, bine partiționate) toți algoritmi de clusterizare abordați înafara de DBSCAN au davies\_bouldin\_score bun. Ceea ce confirmă afirmațiile anterioare (pentru fiecare clusterizare s-a găsit un înțeles al repartizării cu excepția algoritmului DBSCAN)

## CONCLUZII

---

Clusterizarile s-au facut pe baza indicatorilor macroeconomici specifici fiecarei tari intr-un an precum: HDI, GDP/ capita si categoria de populație (descrisa prin sex, categorie de vârstă, generație si rata de suicid).

Conform scorurilor Silhouette, Calinski Harabasz si Davies Bouldin, algoritmul care are rezultatele cele mai bune este K-means( $K = 17$ ) care împarte tarile ca mai jos, apoi cei ierarhici.



Aceasta impartire are sens si este specifica (avem 17 tipuri de tari care explica suficient de bine datele), spre exemple tarile 'Denmark', 'Iceland', 'Sweden' sunt grupate impreuna sau 'Canada', 'Finland', 'Ireland' si 'Netherlands'.

Mai mult, se obtin rezultate similare cele de mai sus daca se elimina feature-ul tara din dataset. Fapt ce demonstreaza redundanta acestui feature care este descris de HDI si PIB.

In plus pentru o tara (care este descrisa prin mai multe linii din dataset) avem o distributie a liniilor peste multimea clusterilor. De aici rezulta ca o tara este descrisa mai exact printr-un vector de dimensiunea numarului de custer (si nu doar printr-un cluster), mai exact prin linia specifica ei din matricea de contingenta. Ca munca viitoare se pot elabora custerizari elaborate intre tari, distante, etc. tinand cont de aceste considerente.