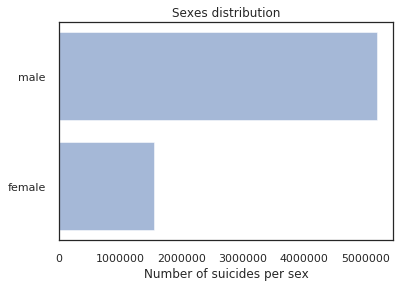
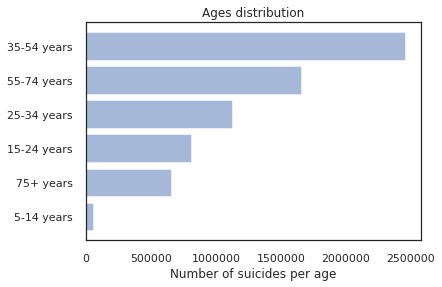
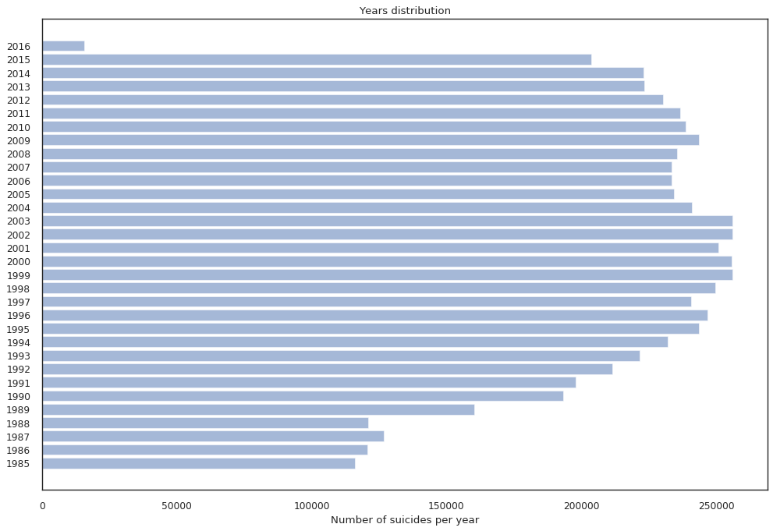
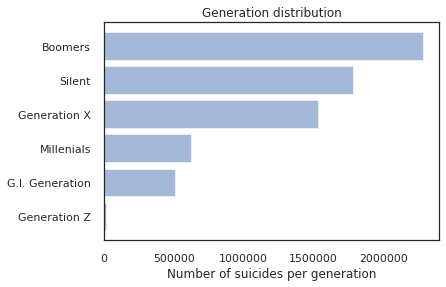
**Suicide Rates Overview 1985 to 2016**

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

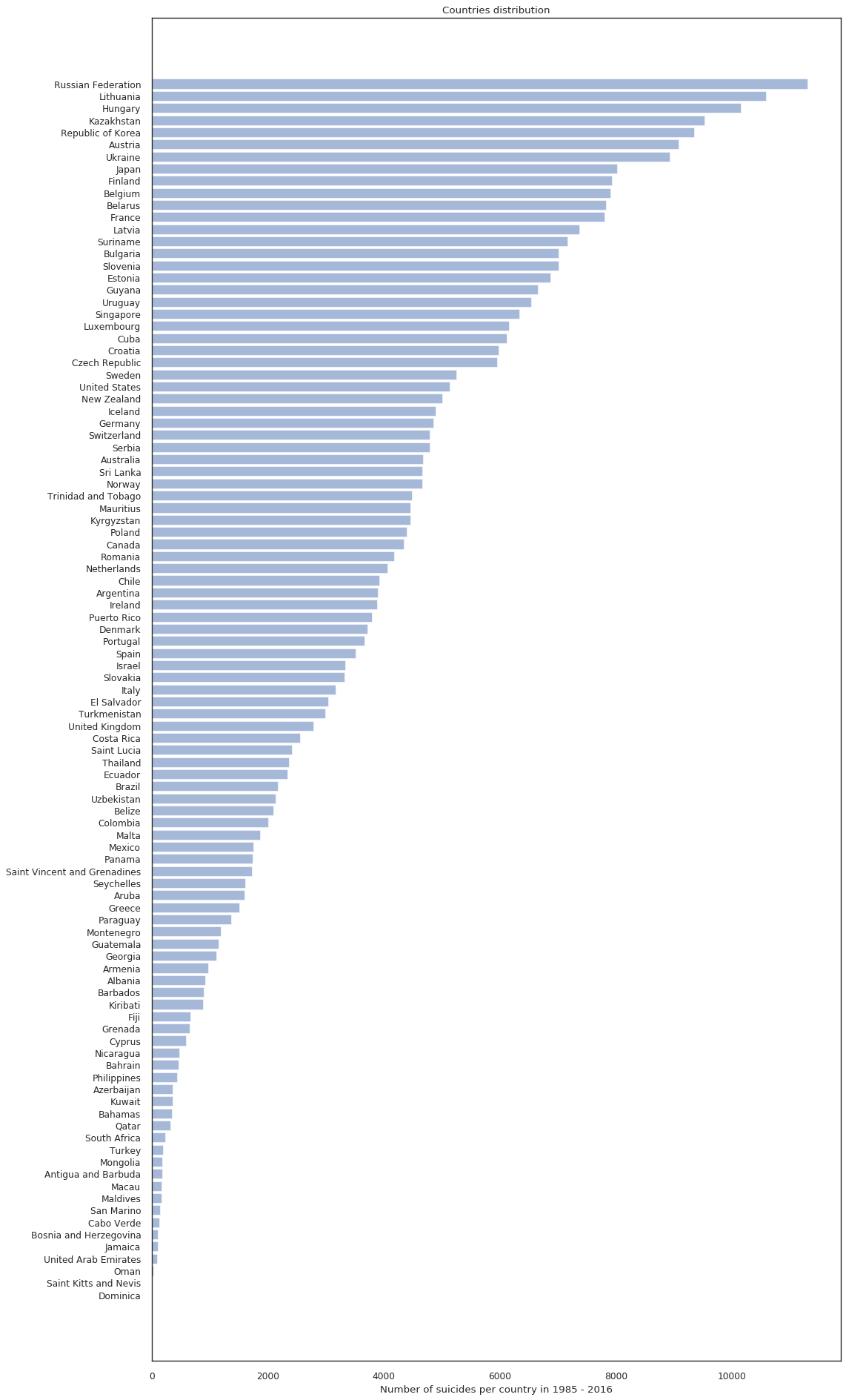
**DESCRIEREA DATASETULUI**   
https://docs.google.com/drawings/d/s0B01EdgDEI6gJDd8QwkLrQ/image?w=661&h=1&rev=1&ac=1&parent=1SS93pM7gnPB23_9xvf_QLZ_NOy0z1FA9VFAXLxQRABw

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. Aceștia 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 vârsta, 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 referința. In continuare vom prezenta o serie de statistici despre datasetul ales.

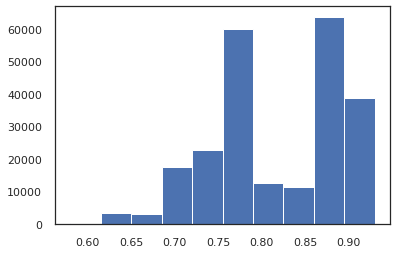
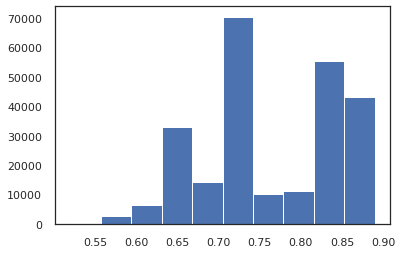
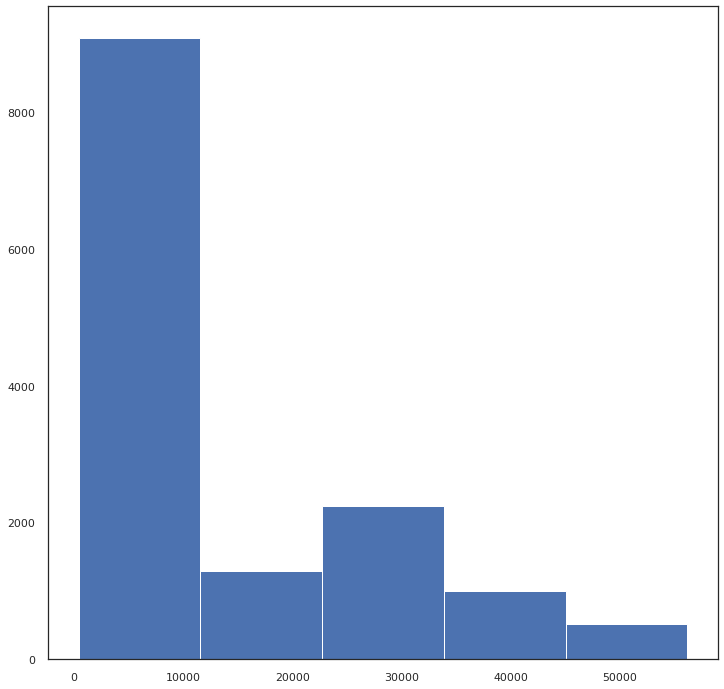
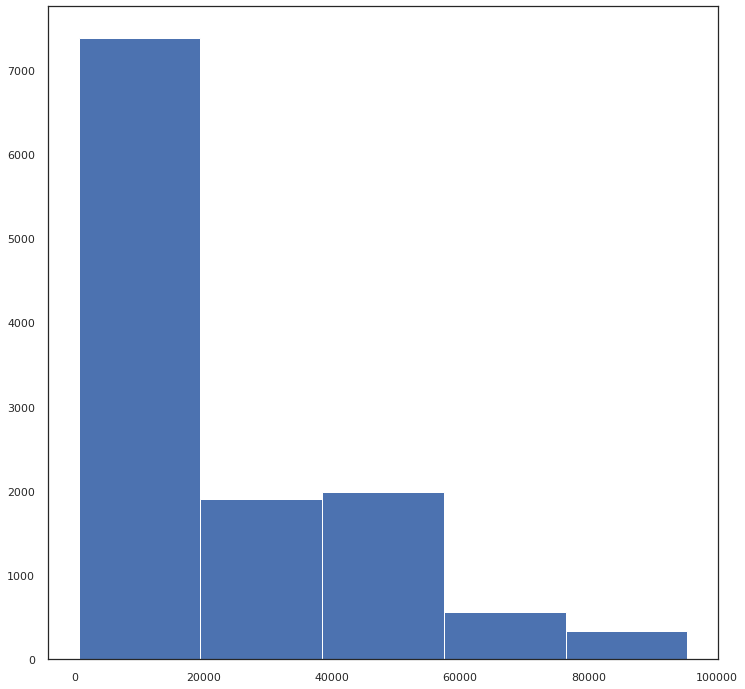
In figura de mai sus este plotata distribuția suicidurilor in funcție de tara (cu mențiunea ca nu sunt prezente toate tarile analizate ci doar cele care au număr de suicid peste medie.[[1]](#footnote-1)[[2]](#footnote-2)[[3]](#footnote-3)[[4]](#footnote-4)



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

[[5]](#footnote-5)

[[6]](#footnote-6)



[[7]](#footnote-7)**REDUNDANTA COLOANELOR**

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

Apoi se demonstrează ca **GDP / (GDP/capita) – 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 **suicides/100k** si **GDP/capita**).

**EXTRAPOLAREA HDI-ului**

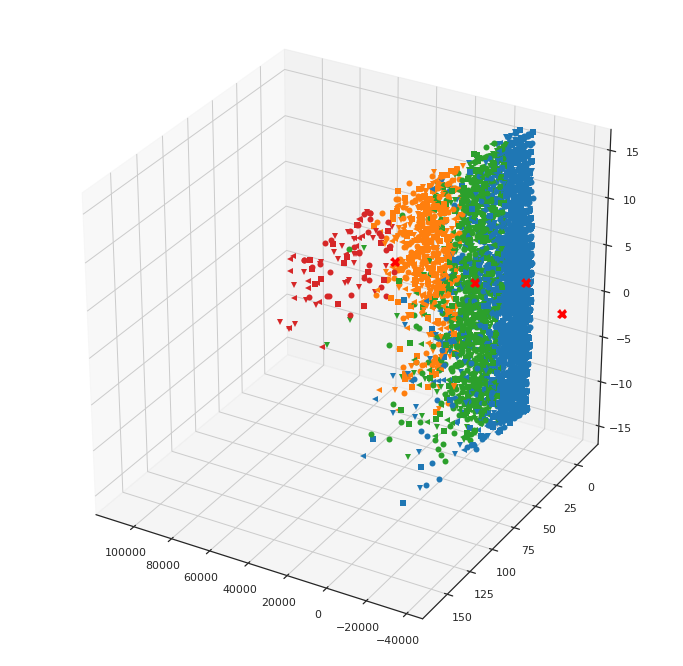
O alta observație notabila legata de dataset este legata de absenta datelor in coloana HDI, exista doar 8364 date din cele 27820 necesare. Intuitiv caracteristica HDI (Human Development Index) este una importanta care influențează rata de suicid a populației dintr-o anumita tara. Deci va trebui sa păstram 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 gradul 1 se obține HDI-ul pentru restul anilor | 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 egal cu HDI-ul pe care îl avem in anul respectiv. HDI-urile lista vor fi HDI-urile globale ale anilor respectivi shiftati in același mod. | 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. |
|  |  |  |

**PREPROCESAREA DATELOR**   
https://docs.google.com/drawings/d/s0B01EdgDEI6gJDd8QwkLrQ/image?w=661&h=1&rev=1&ac=1&parent=1SS93pM7gnPB23_9xvf_QLZ_NOy0z1FA9VFAXLxQRABw

Deoarece in dataset avem date numerice, secvențiale si categoriale, va trebui sa facem o preprocesare a acestuia, astfel:

* **country** – caracteristica categorială. Pentru aceasta caracteristica vom encoda tarile in vectori OneHat (echivalent vom adaugă încă 101 de features (corespunzătoare celor 101 de tari) la datasetul nostru)
* **year** – data secvențiala natural, o vom lasă neschimbata
* **age** – feature ordinal. Pentru acesta vom encoda categoriile de vasta astfel '15-24 years' : 0, '25-34 years' : 1, -54 years' : 2, '5-14 years' : 3, '55-74 years' : 4, '75+ years': 5. Acest mod de encodare păstrează distanta mare intre categoria 75+ de ani si categoria 15-24 de ani si distanta mica intre 15-24 ani si 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, 'Millenials' : 4, 'Silent' : 1.

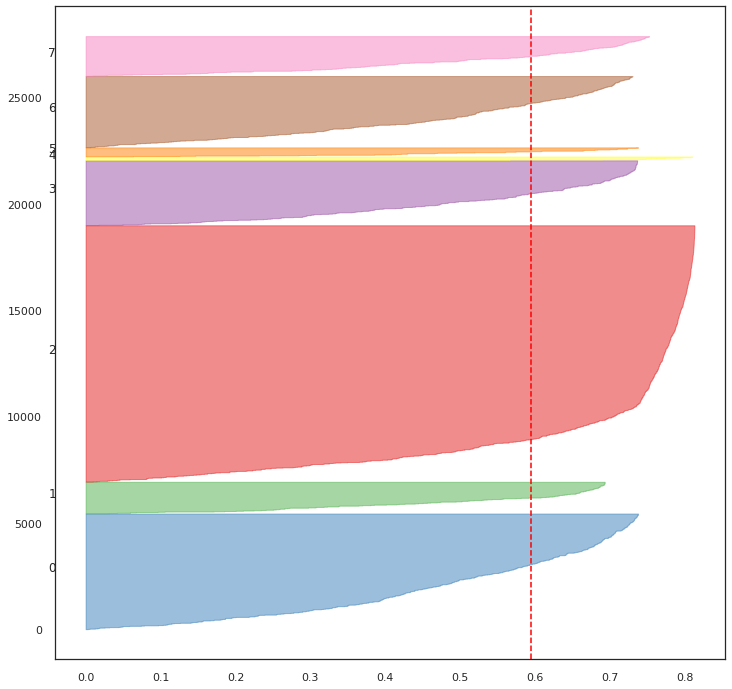
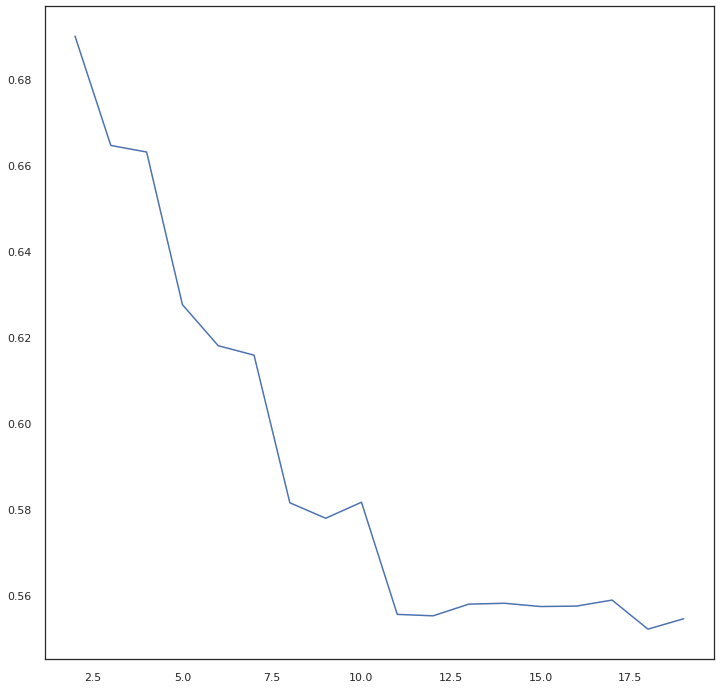
https://docs.google.com/drawings/d/s0B01EdgDEI6gJDd8QwkLrQ/image?w=661&h=1&rev=1&ac=1&parent=1SS93pM7gnPB23_9xvf_QLZ_NOy0z1FA9VFAXLxQRABw**METODELE FOLOSITE**

In continuare, deoarece nu avem nici un fel de label asupra datelor noastre vom dori sa clusterizam tarile in funcție de indicatori macroeconomici si de rata de suicid specifica populației din tara respectiva.

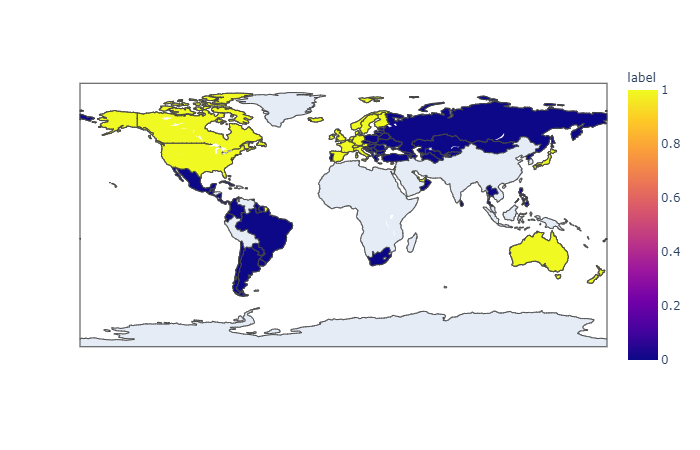
1. **K-MEANS**

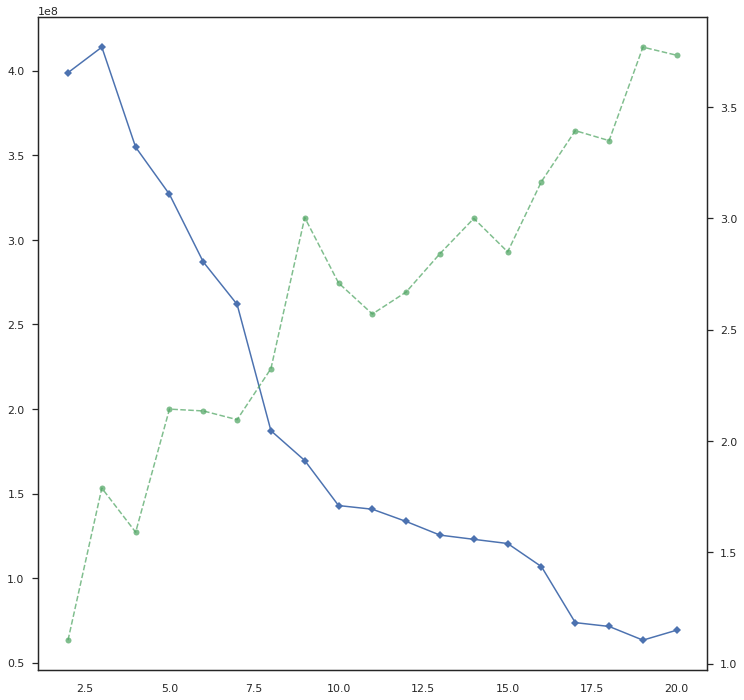
Inițial pornim cu 4 clase apoi dorim sa găsim numărul optim de clase care separa cat mai bine clusterele găsite de algoritmul K-means. In acest caz avem un Silhouette score de 0.663 ceea ce înseamnă ca clusterele sunt bine separate (1 clustere dense si separate si 0 ovelappuire totala si -1 atunci când clusterele sunt incorecte).

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



Pentru numărul optim de clustere obținut mai sus (adică 2) obținem divizarea tarilor in doua clase așa cum se poate observa mai jos. Se poate observa ca tarile dezvoltate ca: SUA, Canada, Australia, Japonia si cele din vestul Europei sunt in clasa 1 si celelalte in clasa 2. Intuitiv aceasta clusterizare i împărțit tarile in funcție de gradul de dezvoltare economica si cum influențează indicii macroeconomici rata de suicid a populației. Cum nu avem exemple (in figura de mai jos) care contrazic afirmația anterioara rezulta ca rata de suicid este influențata 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'],

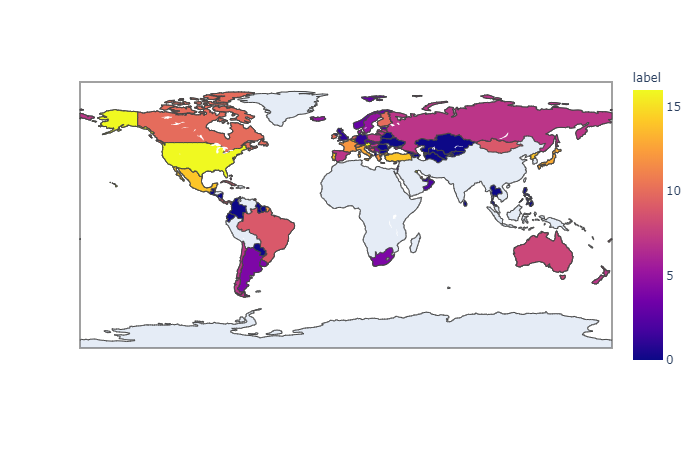
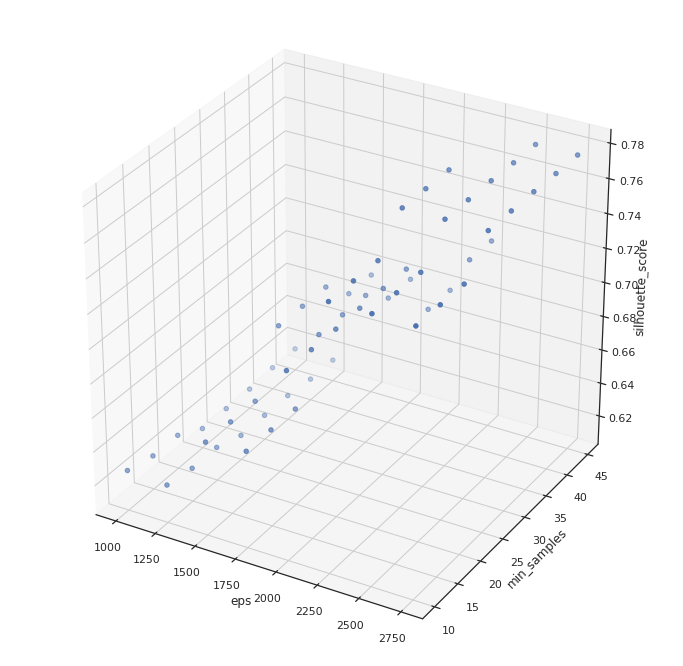
10:['Canada', 'Finland', 'Ireland', 'Netherlands'],

11: ['Qatar', 'Switzerland'],

12: ['France', 'Greece', 'Slovenia'],

13: ['Italy', 'Japan'],

15: ['Luxembourg'],

16: ['Austria', 'Belgium', 'Kuwait', 'United States']

1. **DBSCAN**

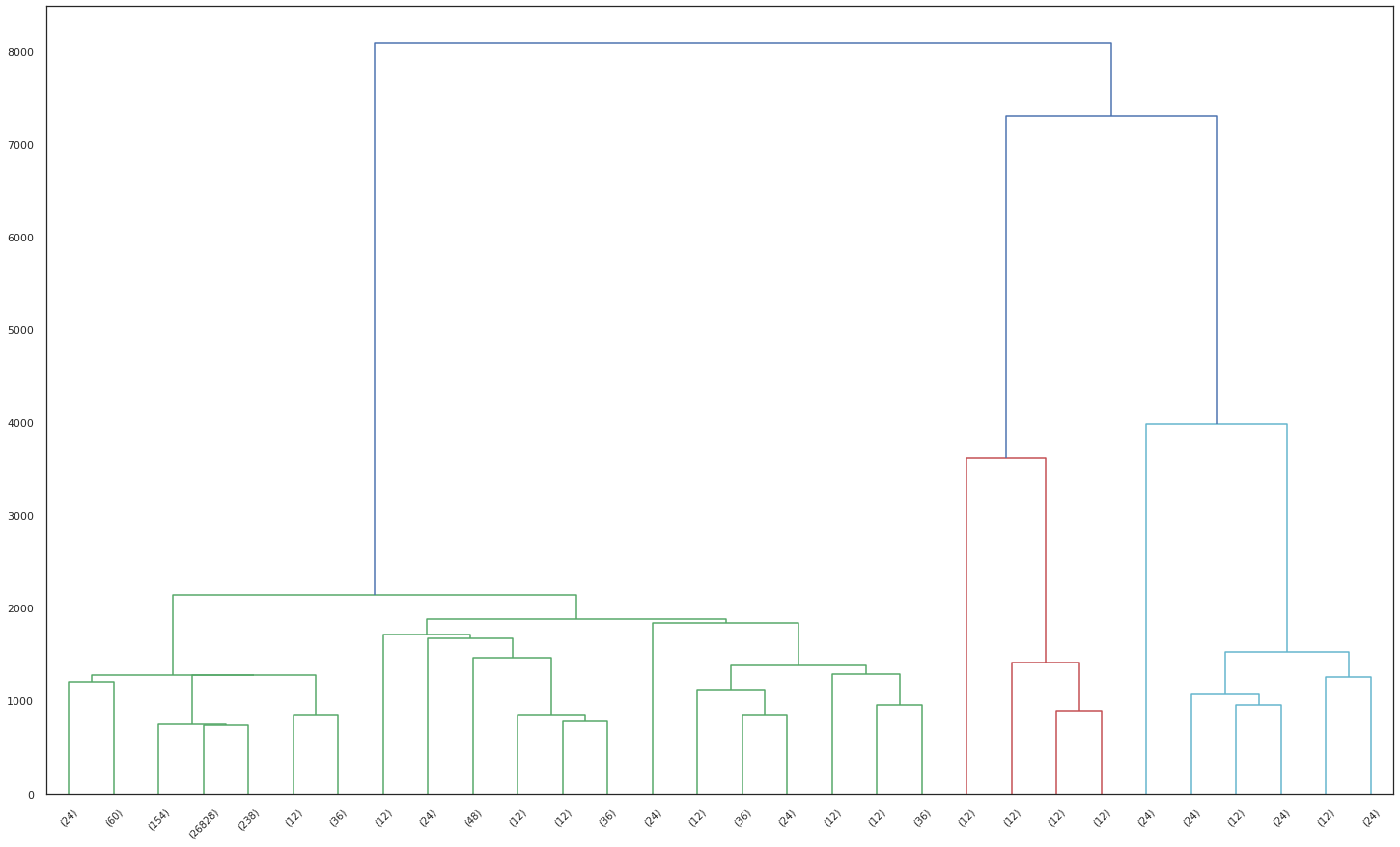
Inițial pornim la drum cu parametrii eps=1000, min\_samples=50, parametrii care împart tarile in 4 clustere. Silhouette score-ul pentru împărțirea data de DBSCAN cu parametri de mai sus este de 0.572.

Apoi in urma unui grid search peste spațiul in funcție de Silhouette score si se obțin următorii parametri optimi pentru epsilon si min\_samples: 2250 si 25 cu un Silhouette score de 0.776. Numărul de clase obținut este 3. Iar distribuția tarilor este: Norvegia într-o clasa, Luxembourg in alta clasa si restul tarilor in alta clasa. Probabil o căutare mai atenta a hiperparametrilor eps si min\_samples ar fi generat o clusterizare mai particulara a tarilor.

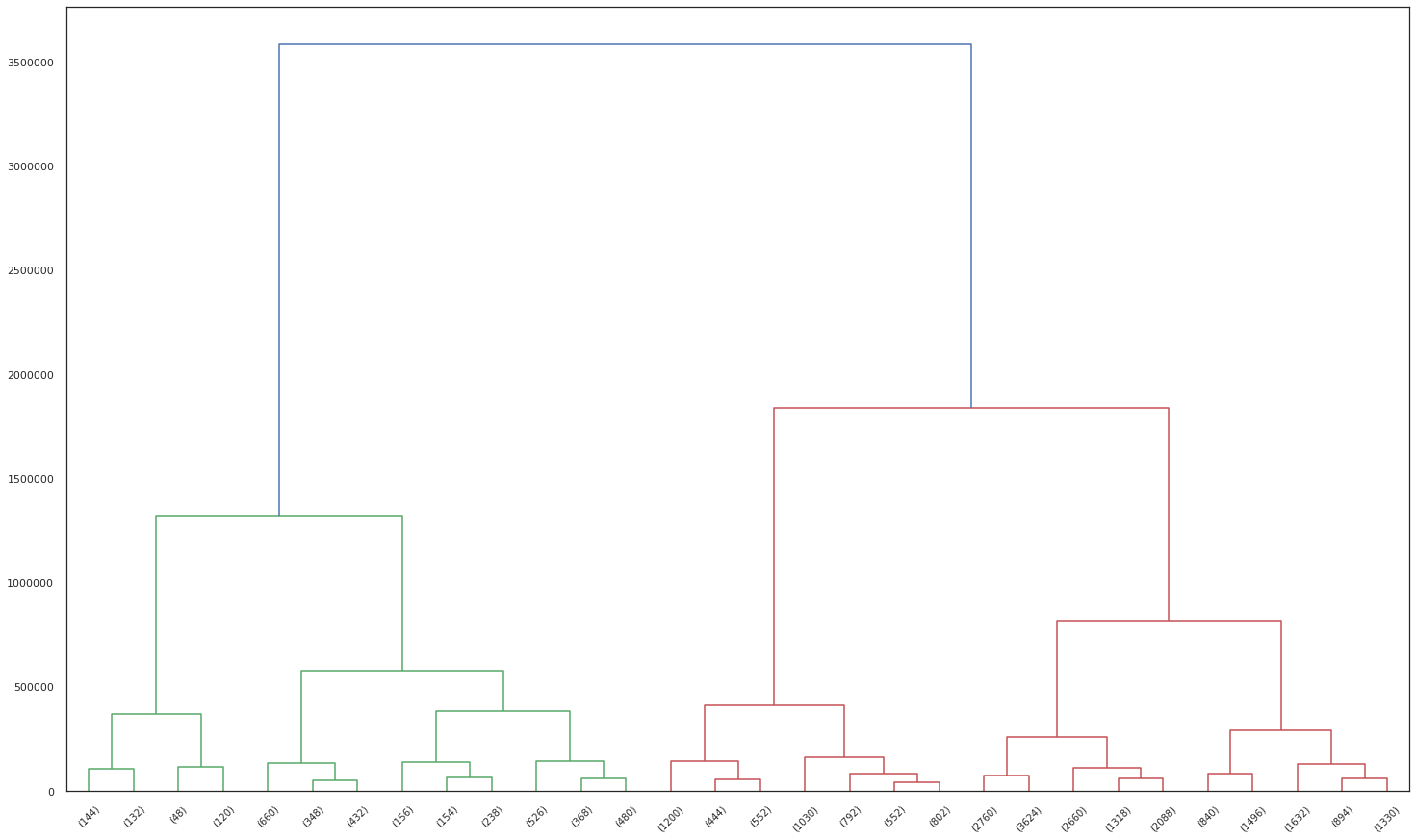


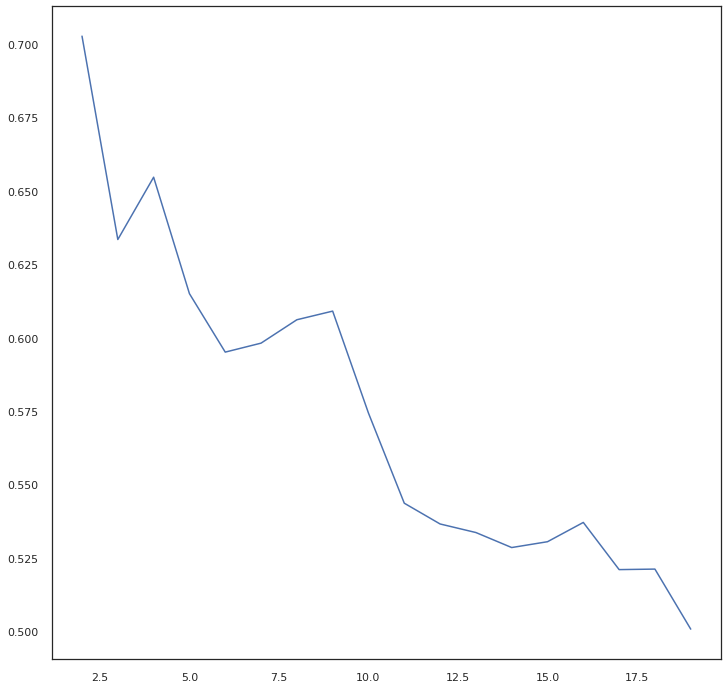
1. **HIERARCHICAL CLUSTERING**

După o clusterizare ierarhica utilizând single linkage obținem următoarea dendograma:



După cum se poate observa este puternic debalansata spre stânga unde sunt concentrate mare parte din exemplele date spre clusterizare. Cum nu ne dorim un asemenea lucru, si dorim mai degrabă o clusterizare specifica decât una generala.



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 clustere (Silhouette score = 0.702). In continuare din motive de specificitate vom considera 9 clustere (Silhouette score = 0.609) pentru care obținem următoarea repartizare a tarilor.

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

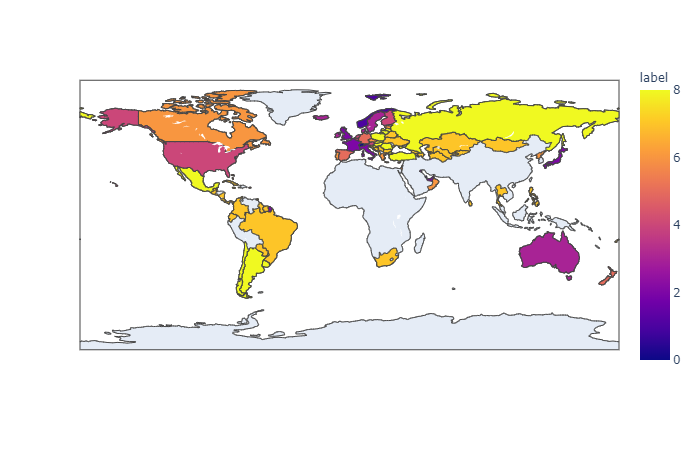
1: ['Luxembourg', 'Norway'],

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

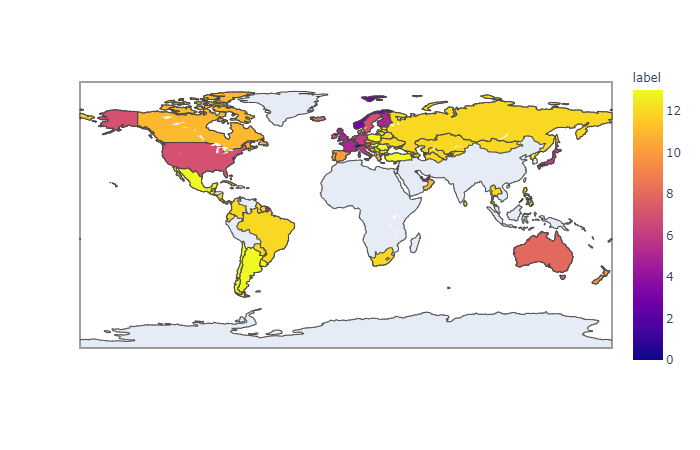
3: ['Australia', 'Denmark', 'Iceland', 'Ireland', 'San Marino', 'Sweden'],

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

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

Se poate observa cu ușurința 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 clustere (Silhouette score = 0.601)



**COMENTAREA REZULTATELOR**   
https://docs.google.com/drawings/d/s0B01EdgDEI6gJDd8QwkLrQ/image?w=661&h=1&rev=1&ac=1&parent=1SS93pM7gnPB23_9xvf_QLZ_NOy0z1FA9VFAXLxQRABw

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 clusterele obținute sunt dense si nu se suprapun excesiv.

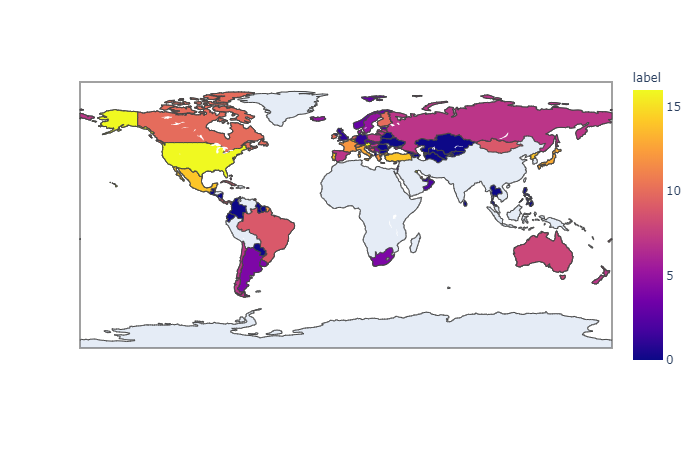
|  |  |  |  |
| --- | --- | --- | --- |
| Algoritmul si parametrii | silhouette\_score | calinski\_harabasz\_score | davies\_bouldin\_score |
| K-means(K = 2) | 0.69 | 56457.28 | 0.5207 |
| K-means(K = 17) | 0.5587 | 361804.44 | 0.4752 |
| DBSCAN(eps = 2250, min\_samples = 25) | 0.776 | 1722.41 | 5.6548 |
| 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 clustere bine definite) algoritmul care are clusterele cele mai definite este K-means(K = 17) apoi algoritmi de clusterizare ierarhica.

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

**CONCLUZII**  
https://docs.google.com/drawings/d/s0B01EdgDEI6gJDd8QwkLrQ/image?w=661&h=1&rev=1&ac=1&parent=1SS93pM7gnPB23_9xvf_QLZ_NOy0z1FA9VFAXLxQRABw

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 vârsta, 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 custere (si nu doar printr-un cluster), mai exact prin linia specifica ei din matricea de contingenta. Ca munca viitoare se pot elabora custerizari eloborate intre tari, distante, etc. tinand cont de aceste considerente.

1. Distribuția sinuciderilor in funcție de sex [↑](#footnote-ref-1)
2. Distribuția sinuciderilor in funcție de an [↑](#footnote-ref-2)
3. Distribuția sinuciderilor in funcție de categoria de vârsta [↑](#footnote-ref-3)
4. Distribuția sinuciderilor in funcție de generație [↑](#footnote-ref-4)
5. Distribuția sinuciderilor in funcție de tara [↑](#footnote-ref-5)
6. Distribuția suicidelor in funcție de PIB (stânga 1996, dreapta 2006) [↑](#footnote-ref-6)
7. Distribuția suicidelor in funcție de HDI (stânga 1996, dreapta 2006) [↑](#footnote-ref-7)