FACULTATEA DE CIBERNETICĂ, STATISTICĂ ŞI INFORMATICĂ ECONOMICĂ

**PROIECT ANALIZA DATELOR**

**CALITATEA VIEȚII ÎN UNIUNEA EUROPEANĂ ȘI CLASIFICAREA ȚĂRILOR DIN UNIUNEA EUROPEANĂ DUPĂ INDICATORII PRIVIND STAREA SĂNĂTĂȚII**

**Dărămuș Alexa-Georgeta,** **Programul de licență în Informatică Economică, Anul 3 de studiu – ID, Grupa 1121**

1. **Calitatea vieții în Uniunea Europeană**

Prezentul studiu realizează o analiză în componente principale a calității vieții în Uniunea Europeană la nivelul anului 2016. Calitatea vieții este descrisă cu ajutorul a 12 indicatori prezentați în Tabelul 1, indicatori aflați în grupul de date Eurostat referitor la acest aspect (Quality of Life)[[1]](#footnote-1). Datele prelucrate sunt preluate dintr-un fișier text de tip *csv* anexate studiului *CalitateaVietii.csv*.

Tabelul 1. Variabilele observate

|  |  |  |
| --- | --- | --- |
| Cod | Denumire | Unitate de măsură |
| CIE | Consumul individual efectiv în prețuri curente | Euro |
| VNMA | Venitul net mediu anual | Euro |
| PRS | Ponderea persoanelor aflate în risc de sărăcie | Procentaj |
| PLI | Ponderea persoanelor care locuiesc în condiții inadecvate | Procentaj |
| RS | Rata șomajului | Procentaj |
| OLSM | Orele lucrate în medie săptămânal | Ore |
| SV | Speranța de viață | Ani |
| PSS | Ponderea persoanelor cu studii superioare | Procentaj |
| RAD | Rata de abandon în educație | Procentaj |
| PAD | Ponderea persoanelor cu abilități digitale | Procentaj |
| RC | Rata criminalității | Numărul de omucideri la suta de mii de locuitori |
| OG | Diferența în ceea ce privește ocuparea forței de muncă după gen | Diferența dintre gradul de angajare al femeilor și bărbaților (20-64 de ani) |

**1. Analiza variabilității datelor**

**1.1. Analiza matricei de corelații**

O primă imagine asupra variabilității este dată de matricea de corelații dintre variabilele observate. Corelațiile puternice sunt cele care indică variabile între care există strânsă legătură, deci aduc redundanță informațională în setul de date. Aceste variabile sunt cele din care se vor construi componentele principale.

În Figura 1 este prezentată corelograma variabilelor observate.

Din analiza graficului rezultă că venitul net mediu anual, consumul individual efectiv în prețuri curente, ponderea persoanelor cu abilități digitale și orele lucrate în medie săptămânal generează cele mai mari corelații, deci aceste variabile vor contribui cel mai mult la constituirea componentelor principale.

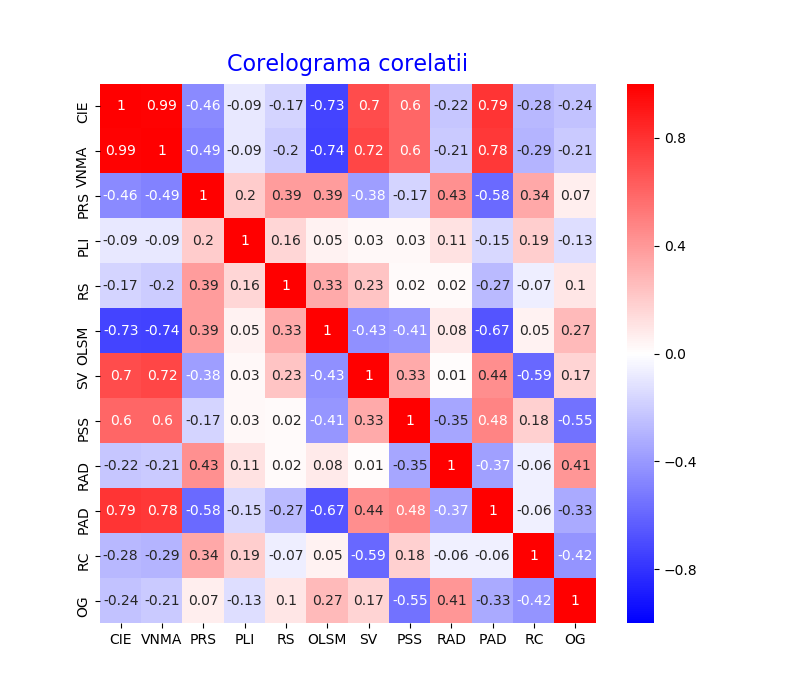


Figura 1. Corelogramă variabile observate

**1.2. Analiza valorilor proprii. Varianța componentelor principale**

Rezultatele modelului pornesc de la calculul vectorilor și valorilor proprii ale matricei de corelație. Valorile proprii reprezintă varianțele componentelor iar vectorii proprii reprezintă coeficienții legăturii liniare dintre variabilele observate și componentele principale. În Tabelul 2 este prezentată varianța componentelor principale, individual, cumulat și procentual. Conform criteriilor Cattell și Kaiser sunt semnificative primele patru componente, aspect scos în evidență și de graficul distribuției varianței prezentat în Figura 2.

Tabelul 2. Distribuția varianței

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Componenta | Varianță | Varianță cumulată | Procent  varianță | Procent cumulat varianță |
| C1 | 4,836983 | 4,836983 | 40,30819 | 40,30819 |
| C2 | 2,115276 | 6,952259 | 17,6273 | 57,93549 |
| C3 | 1,489764 | 8,442023 | 12,4147 | 70,35019 |
| C4 | 1,135545 | 9,577568 | 9,462875 | 79,81306 |
| C5 | 0,821443 | 10,39901 | 6,84536 | 86,65842 |
| C6 | 0,43135 | 10,83036 | 3,594583 | 90,25301 |
| C7 | 0,374533 | 11,20489 | 3,121111 | 93,37412 |
| C8 | 0,296913 | 11,50181 | 2,474274 | 95,84839 |
| C9 | 0,255008 | 11,75682 | 2,125068 | 97,97346 |
| C10 | 0,142407 | 11,89922 | 1,186723 | 99,16018 |
| C11 | 0,089209 | 11,98843 | 0,74341 | 99,90359 |
| C12 | 0,011569 | 12 | 0,096407 | 100 |

Prima componentă principală este mult mai semnificativă decât celelalte trei, acoperind peste 40% din variabilitate.

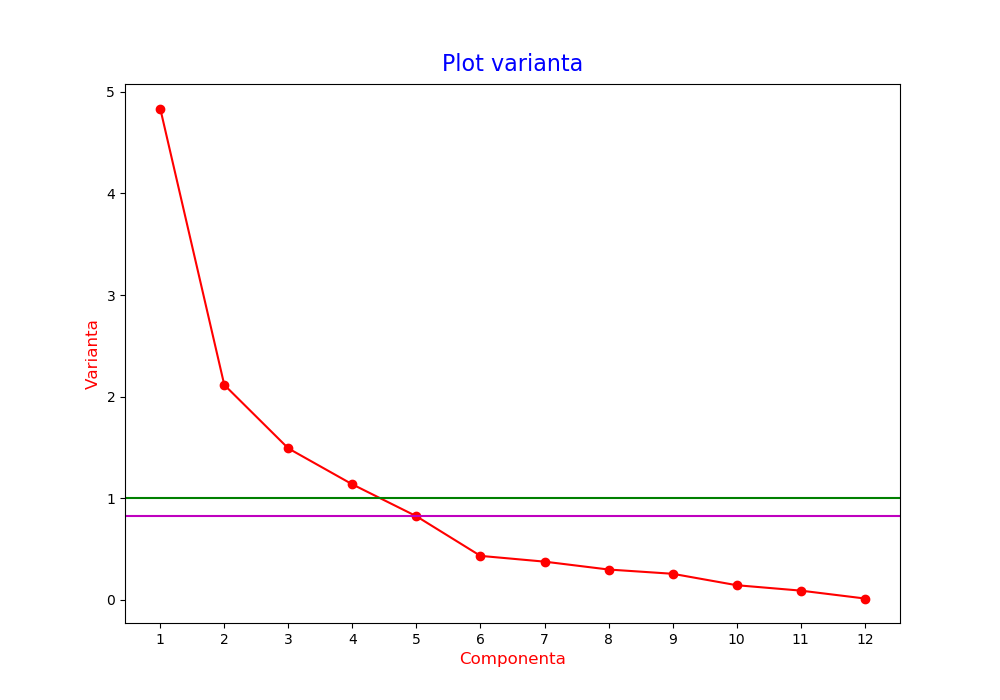


Figura 2. Graficul valorilor proprii

**2. Calculul corelațiilor dintre variabilele observate și componentele principale**

Aceste corelații numite și corelații factoriale (eng. *factor loadings*) se calculează pornind de la vectorii proprii ai matricei de corelație. Ele reprezintă coeficienții de corelație liniară dintre variabilele observate și componentele principale. Fiind calculate din vectorii proprii ele reflectă saturațiile componentelor în fiecare variabilă observată sau altfel spus, contribuția fiecărei variabile observate la constituirea unei componente principale.

În Tabelul 3 sunt prezentate aceste corelații pentru cele mai semnificative patru componente principale. Aceste corelații sunt importante în încercarea de a denomina componentele principale. Acțiunea de denominare constă în asocierea unei semnificații concrete componentelor principale, semnificație exprimată prin câteva cuvinte care să dea un sens generic componentei. Denominarea mai este denumită și etichetare, "eticheta" asociată fiind legată de ceea ce au în comun variabilele puternic corelate cu componenta respectivă.

Astfel, se poate observa că prima componentă principală este puternic corelată cu venitul net mediu anual, consumul individual efectiv în prețuri curente, ponderea persoanelor cu abilități digitale și orele lucrate în medie săptămânal. Componenta 2 este puternic corelată cu rata criminalității și cu diferența în ceea ce privește ocuparea forței de muncă după gen. Are o corelație relativ mare și cu speranța de viață. Componenta 3 este legată în principal de rata șomajului. Are o corelație relativ mare și cu ponderea persoanelor care locuiesc în condiții inadecvate. Componenta 4 este legată în principal de rata de abandon în educație.

Corelațiile din Tabelul 3 sunt prezentate grafic prin cercul corelațiilor (pentru primele două componente) și prin corelogramă (Figurile 3 și 4).

Tabelul 3. Corelațiile factoriale

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cod | C1 | C2 | C3 | C4 |
| CIE | **0,946837** | -0,09236 | 0,141368 | 0,108152 |
| VNMA | **0,952248** | -0,11246 | 0,121975 | 0,115118 |
| PRS | **-0,64485** | 0,141712 | 0,480376 | 0,248081 |
| PLI | -0,13877 | 0,181696 | **0,573331** | 0,245968 |
| RS | -0,25136 | -0,18259 | **0,733456** | -0,46953 |
| OLSM | **-0,78203** | -0,06457 | 0,028416 | -0,415 |
| SV | **0,667551** | **-0,56714** | 0,392718 | -0,07556 |
| PSS | **0,64127** | 0,431314 | 0,364293 | -0,10963 |
| RAD | -0,37258 | -0,40768 | 0,187649 | **0,719772** |
| PAD | **0,869549** | 0,127978 | -0,12319 | 0,015235 |
| RC | -0,25341 | **0,822602** | 0,057104 | 0,231199 |
| OG | -0,32843 | **-0,79785** | -0,12681 | 0,079851 |

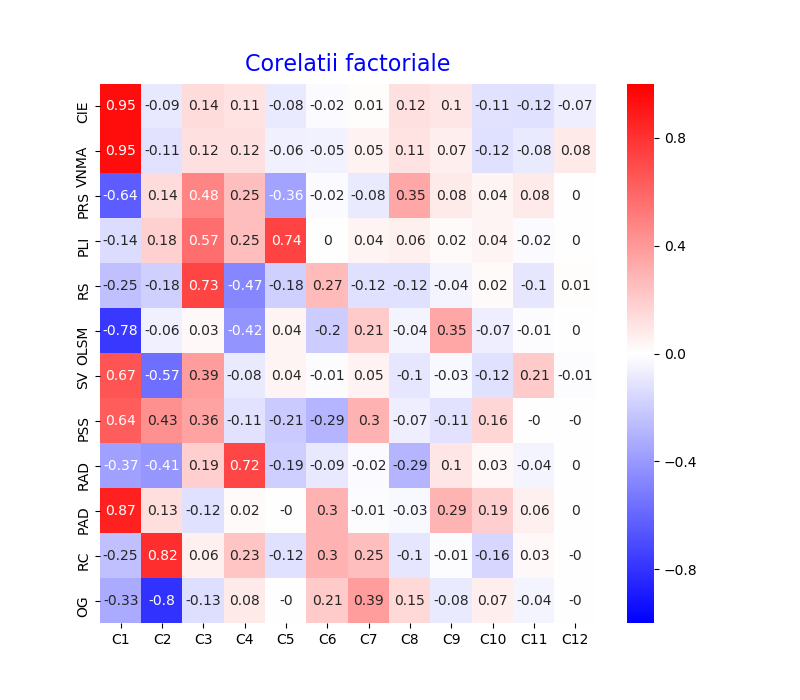
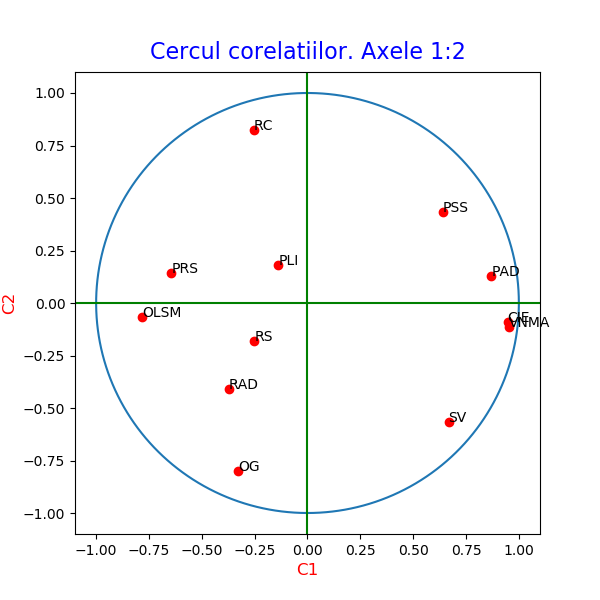


Figura 3. Cercul corelațiilor Figura 4. Corelograma variabile - comp.

**3. Calculul scorurilor**

Scorurile sunt proiecțiile normalizate ale țărilor în axele principale (axele componentelor principale). Proiecțiile în primele două axe sunt prezentate în Figura 5. Valorile mari pe axa Componentei 1 înseamnă valori mari pentru venitul net mediu anual, consumul individual efectiv în prețuri curente și ponderea persoanelor cu abilități digitale deoarece sunt puternic și direct corelate cu Componenta 1 și valori mici pentru orele lucrate în medie săptămânal deoarece aceasta este invers corelată cu Componenta 1. Tabelul scorurilor pentru primele patru componente este prezentat în Anexa 1.

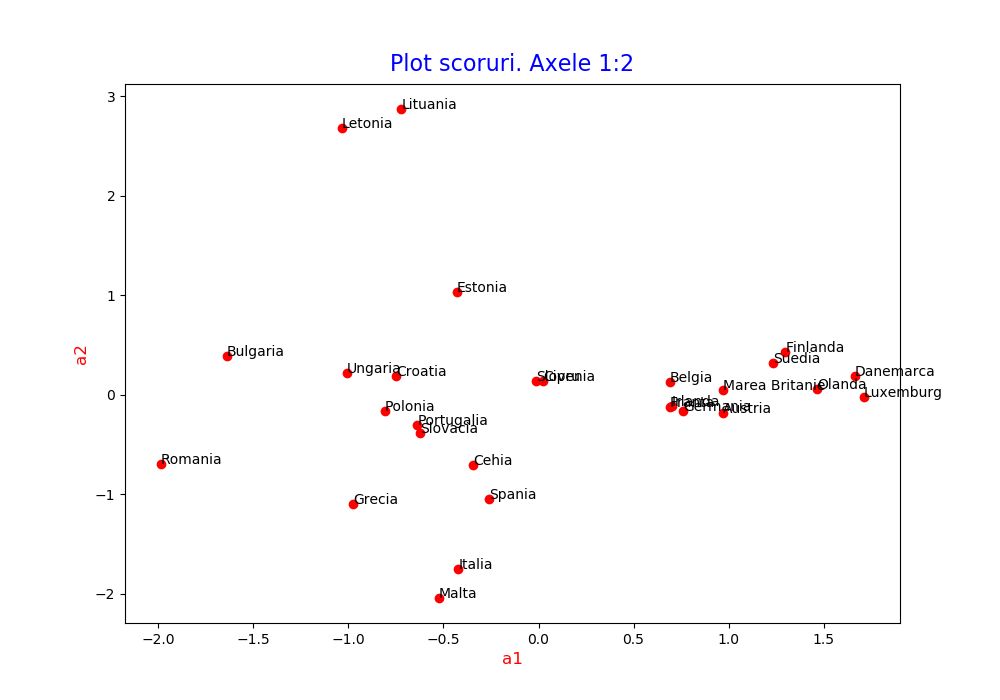


Figura 5. Reprezentare țări in axele 1 și 2

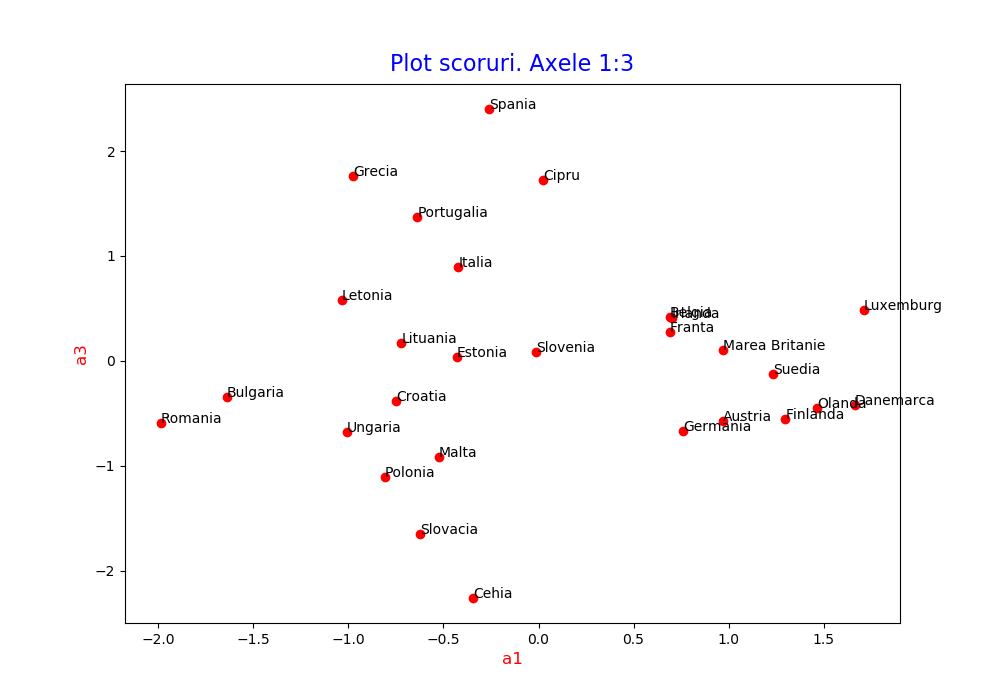


Figura 6. Reprezentare țări in axele 1 și 3

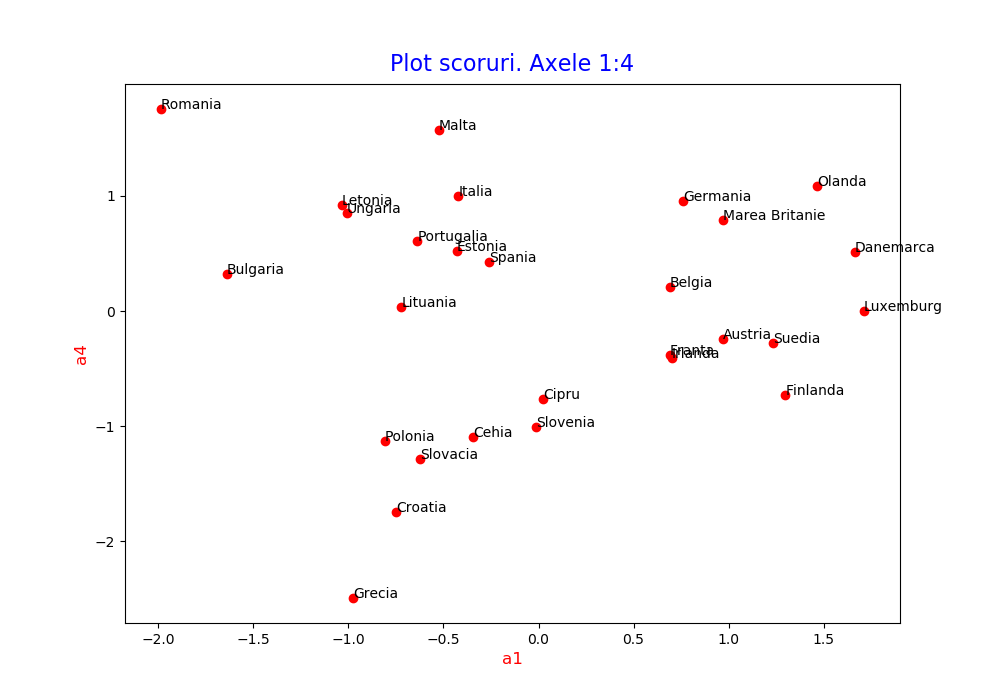


Figura 7. Reprezentare țări in axele 1 și 4

**4. Valorile cosinus**

Valorile cosinus reflectă importanța relativă a unei țări pentru o componentă în raport cu celelalte componente. În Anexa 2 sunt prezentate valorile cosinus pentru primele patru componente. Sunt marcate cu verde mai intens valorile semnificative.

**5. Contribuția instanțelor la varianța componentelor**

Contribuția instanțelor la varianța componentelor reflectă importanța relativă a unei țări pentru o componentă în raport cu ceilalți indivizi. Tabelul contribuțiilor pentru primele patru componente este prezentat în Anexa 3. Se poate observa evidențierea mai clară a valorilor extreme luate de unele țări pentru primele patru componente, cum ar fi Letonia, Lituania pentru Componenta 2 (rata criminalității și diferența în ceea ce privește ocuparea forței de muncă după gen), Cehia, Spania pentru Componenta 3 (rata șomajului) sau Grecia pentru Componenta 4 (rata de abandon în educație).

**6. Comunalitățile**

Comunalitățile scot în evidență "transferul" gradual de informație din variabilele observate la componentele principale. Comunalitățile sunt prezentate prin corelograma din Figura 8.Se pot observa pe grafic variabilele care au legături mai puternice cu componentele mai puțin semnificative (ponderea persoanelor care locuiesc în condiții inadecvate, rata criminalității etc).

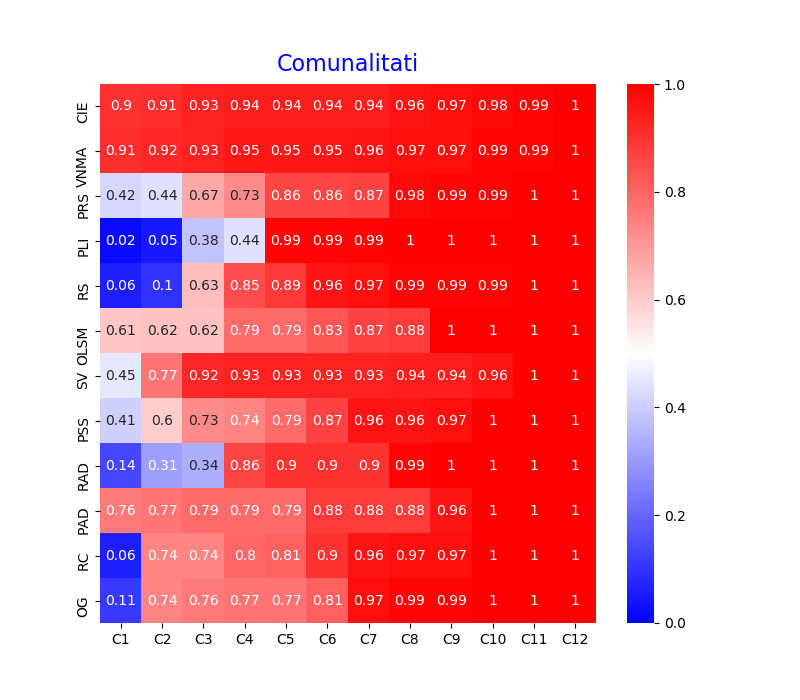


Figura 8. Corelogramă comunalități

**ANEXE**

Anexa 1. Tabela scorurilor

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Țara | C1 | C2 | C3 | C4 |
| Austria | 0,968025 | -0,18324 | -0,57771 | -0,24338 |
| Belgia | 0,691852 | 0,126278 | 0,414295 | 0,211625 |
| Bulgaria | -1,64114 | 0,392439 | -0,34079 | 0,320482 |
| Croația | -0,75017 | 0,189248 | -0,38396 | -1,74201 |
| Cipru | 0,02369 | 0,132756 | 1,726427 | -0,76075 |
| Cehia | -0,3425 | -0,71028 | -2,25896 | -1,0936 |
| Danemarca | 1,664743 | 0,185288 | -0,41653 | 0,513882 |
| Estonia | -0,42814 | 1,035341 | 0,03843 | 0,518489 |
| Finlanda | 1,299313 | 0,426997 | -0,55213 | -0,73281 |
| Franța | 0,691189 | -0,12179 | 0,274589 | -0,3818 |
| Germania | 0,759917 | -0,16137 | -0,66646 | 0,953959 |
| Grecia | -0,97593 | -1,09652 | 1,76521 | -2,49103 |
| Ungaria | -1,0095 | 0,218211 | -0,68211 | 0,852503 |
| Irlanda | 0,70184 | -0,11779 | 0,407785 | -0,40809 |
| Italia | -0,42192 | -1,75427 | 0,898415 | 1,002759 |
| Letonia | -1,03611 | 2,679691 | 0,580285 | 0,919716 |
| Lituania | -0,72239 | 2,869122 | 0,166322 | 0,032011 |
| Luxemburg | 1,712529 | -0,02763 | 0,482314 | -0,00147 |
| Malta | -0,52586 | -2,04226 | -0,91604 | 1,569091 |
| Olanda | 1,464509 | 0,053427 | -0,4466 | 1,08337 |
| Polonia | -0,80943 | -0,15981 | -1,10766 | -1,1304 |
| Portugalia | -0,637 | -0,30356 | 1,370844 | 0,608047 |
| România | -1,98618 | -0,69574 | -0,59313 | 1,753766 |
| Slovacia | -0,622 | -0,38477 | -1,64537 | -1,28068 |
| Slovenia | -0,01356 | 0,136876 | 0,08374 | -1,011 |
| Spania | -0,25851 | -1,05263 | 2,399528 | 0,42645 |
| Suedia | 1,233492 | 0,31933 | -0,1283 | -0,28016 |
| Marea Britanie | 0,969233 | 0,04667 | 0,107564 | 0,791022 |

Anexa 2. Valorile cosinus

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Țara** | **C1** | **C2** | **C3** | **C4** |
| Austria | 0,80993 | 0,012692 | 0,088846 | 0,012019 |
| Belgia | 0,764224 | 0,011134 | 0,084403 | 0,016786 |
| Bulgaria | 0,742896 | 0,018577 | 0,009866 | 0,006651 |
| Croația | 0,30732 | 0,008553 | 0,024797 | 0,38905 |
| Cipru | 0,000322 | 0,004417 | 0,526089 | 0,077863 |
| Cehia | 0,049047 | 0,092243 | 0,657113 | 0,117388 |
| Danemarca | 0,87692 | 0,004751 | 0,016908 | 0,019617 |
| Estonia | 0,163575 | 0,418313 | 0,000406 | 0,056319 |
| Finlanda | 0,650963 | 0,030745 | 0,036203 | 0,048611 |
| Franța | 0,637148 | 0,008651 | 0,030971 | 0,045639 |
| Germania | 0,505978 | 0,009978 | 0,119865 | 0,187193 |
| Grecia | 0,219872 | 0,121382 | 0,221547 | 0,336293 |
| Ungaria | 0,418518 | 0,008552 | 0,058851 | 0,070069 |
| Irlanda | 0,347045 | 0,004275 | 0,036084 | 0,027545 |
| Italia | 0,071454 | 0,540194 | 0,099784 | 0,094752 |
| Letonia | 0,220743 | 0,645716 | 0,021326 | 0,040834 |
| Lituania | 0,121835 | 0,840456 | 0,001989 | 5,62E-05 |
| Luxemburg | 0,753709 | 8,58E-05 | 0,018413 | 1,31E-07 |
| Malta | 0,077404 | 0,510554 | 0,072343 | 0,161791 |
| Olanda | 0,648579 | 0,000377 | 0,018576 | 0,083322 |
| Polonia | 0,423558 | 0,00722 | 0,244292 | 0,193934 |
| Portugalia | 0,185927 | 0,018465 | 0,265206 | 0,039771 |
| România | 0,739202 | 0,039666 | 0,020303 | 0,135301 |
| Slovacia | 0,204693 | 0,034256 | 0,44116 | 0,203722 |
| Slovenia | 0,000159 | 0,007108 | 0,001874 | 0,208173 |
| Spania | 0,021099 | 0,152978 | 0,559859 | 0,013479 |
| Suedia | 0,742538 | 0,021763 | 0,002474 | 0,008993 |
| Marea Britanie | 0,669566 | 0,000679 | 0,00254 | 0,104699 |

Anexa 3. Contribuțiile instanțelor

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Țara** | **C1** | **C2** | **C3** | **C4** |
| Austria | 3,346687 | 0,119919 | 1,191963 | 0,21155 |
| Belgia | 1,709498 | 0,05695 | 0,613002 | 0,159948 |
| Bulgaria | 9,619028 | 0,55003 | 0,414783 | 0,366816 |
| Croația | 2,009825 | 0,12791 | 0,526532 | 10,83786 |
| Cipru | 2,00E-03 | 0,062943 | 10,64482 | 2,066931 |
| Cehia | 0,41896 | 1,801793 | 18,22465 | 4,271266 |
| Danemarca | 9,897751 | 0,122613 | 0,619638 | 0,943123 |
| Estonia | 0,654664 | 3,828323 | 5,27E-03 | 0,960112 |
| Finlanda | 6,029339 | 0,651165 | 1,088722 | 1,917883 |
| Franța | 1,706225 | 0,052974 | 0,269282 | 0,5206 |
| Germania | 2,062405 | 0,093004 | 1,58633 | 3,250137 |
| Grecia | 3,401597 | 4,294129 | 11,12845 | 22,1616 |
| Ungaria | 3,639589 | 0,170057 | 1,661679 | 2,595578 |
| Irlanda | 1,75921 | 0,049551 | 0,593889 | 0,594762 |
| Italia | 0,635778 | 10,99095 | 2,882674 | 3,591163 |
| Letonia | 3,833982 | 25,64551 | 1,202608 | 3,020993 |
| Lituania | 1,863762 | 29,3995 | 0,098796 | 3,66E-03 |
| Luxemburg | 10,47413 | 2,73E-03 | 0,830809 | 7,74E-06 |
| Malta | 0,987592 | 14,89578 | 2,996868 | 8,793029 |
| Olanda | 7,659957 | 0,010195 | 0,712316 | 4,191749 |
| Polonia | 2,339907 | 0,091211 | 4,381805 | 4,563621 |
| Portugalia | 1,449174 | 0,329112 | 6,711478 | 1,320434 |
| România | 14,08891 | 1,728763 | 1,256434 | 10,98463 |
| Slovacia | 1,381709 | 0,528753 | 9,6687 | 5,857632 |
| Slovenia | 6,57E-04 | 0,066911 | 0,025044 | 3,650396 |
| Spania | 0,238678 | 3,957269 | 20,56334 | 0,649499 |
| Suedia | 5,433933 | 0,364185 | 0,058793 | 0,280329 |
| Marea Britanie | 3,355045 | 7,78E-03 | 0,041321 | 2,234698 |

1. **Clasificarea țărilor din Uniunea Europeană după indicatorii privind starea sănătății**

Prezentul studiul se referă la gruparea țărilor din Uniunea Europeană prin algoritmi de clusterizare ierarhică după indicatorii privind starea sănătății, înregistrați la nivelul anului 2014. Indicatorii utilizați în studiu sunt prezentați în Tabelul 1 în conformitate cu datele Eurostat referitoare la acest aspect (Health) și sunt exprimați în procente. Datele se află în fișierul *Sanatate.csv* atașat studiului.

Tabelul 1. Indicatori privind starea sănătății

|  |  |  |
| --- | --- | --- |
| Cod | |  | | --- | | **Semnificație** | |
| DU | Procentaj persoane care suferă de depresie ușoară |
| DM | Procentaj persoane care suferă de depresie moderată |
| DS | Procentaj persoane care suferă de depresie severă |
| DV | Procentaj persoane cu deficiențe de vedere |
| DA | Procentaj persoane cu deficiențe de auz |
| DD | Procentaj persoane cu dificultăți de deplasare |
| SA | Procentaj persoane care suferă de astm |
| BR | Procentaj persoane care suferă de boli respiratorii |
| TA | Procentaj persoane care suferă de tensiune arterială crescută |
| SD | Procentaj persoane care suferă de diabet |
| IA | Procentaj persoane rănite în accidente |

Vom efectua două clasificări ale instanțelor, prin metoda Ward și prin legătură completă și o clasificare a indicatorilor prin legătură completă.

1. **Clasificarea țărilor prin metoda Ward**

Metrica utilizată este metrica euclidiană. Mai întâi vom determina și analiza partiția optimală, apoi o partiție aleasă după examinarea graficului dendrogramă.

Graficul dendrogramă este prezentat în Figura 1 iar componența clusterelor în Tabelul 2. După cum se poate observa partiția optimală conține două clustere.

Analiza clusterelor se face urmărind distribuția fiecărui indicator pentru fiecare cluster. În acest fel se identifică particularitățile și diferențele dintre clustere. În Anexa 1 sunt prezentate câteva distribuții care scot în evidență diferențe clare între cele două clustere ale partiției optimale.

Conform histogramelor în țările din clusterul *c1* dificultățile de deplasare și tensiunea arterială crescută sunt mai întâlnite în mod clar mai mare decât la țările din clusterul *c0*, unde se regăsește mai mult astmul.

**Tabel 2.** Partiția optimală

|  |  |
| --- | --- |
| Cluster | Țări |
| c0 | Danemarca, Finlanda, Franța, Luxemburg, Slovenia, Suedia |
| c1 | Austria, Belgia, Bulgaria, Croația Cipru, Cehia, Estonia, Germania, Grecia, Ungaria, Irlanda, Italia, Letonia, Lituania, Malta, Olanda, Polonia, Portugalia, România, Slovacia, Spania, Marea Britanie |

Partiția cu trei clustere este prezentată în Tabelul 3. Graficul dendrogramă și graficul instanțelor în axele discriminante sunt prezentate în anexă. În această partiție apar două clustere rezultate din scindarea clusterului *c1* din partiția optimală. În anexă este prezentat un grafic histogramă care indică o incidență mai mare pentru deficiențele de vedere în clusterul c2.

Tabel 3. Partiția cu trei clustere

|  |  |
| --- | --- |
| Cluster | Țări |
| c0 | Bulgaria, Croația, Cehia, Grecia, Ungaria, Letonia, Lituania, Polonia, Slovacia |
| c1 | Danemarca, Finlanda, Franța, Luxemburg, Slovenia, Suedia |
| c2 | Austria, Belgia, Cipru, Estonia, Germania, Irlanda, Italia, Malta, Olanda, Portugalia, România, Spania, Marea Britanie |

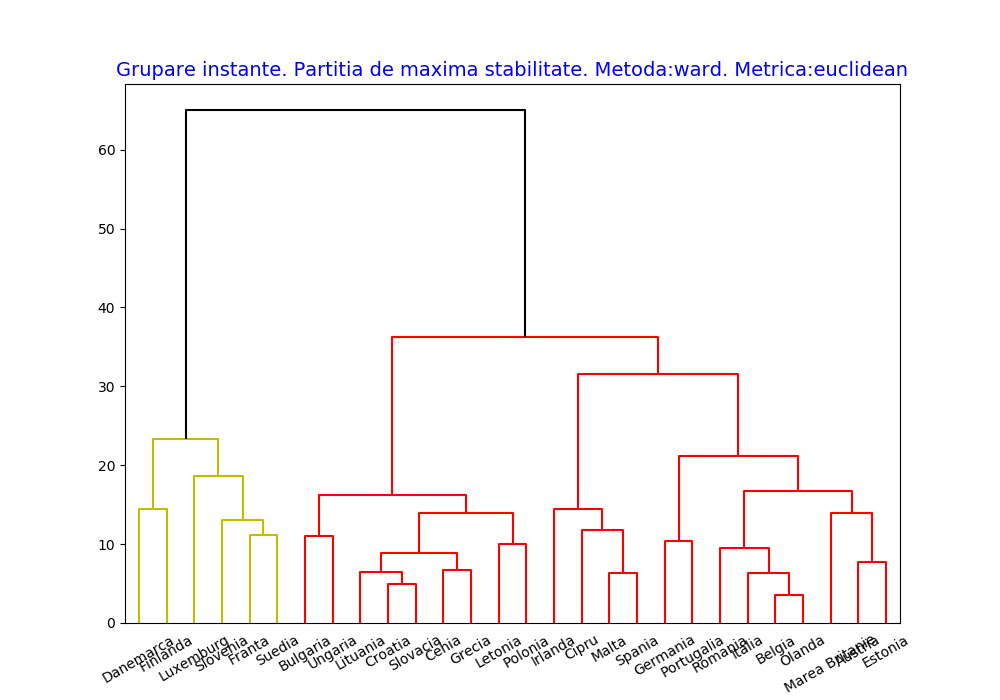
****

Figura 1.Partiția optimală. Dendrogramă

În Anexa 2 sunt sintetizate grupările pe mai multe partiții și partiția optimală.

**2. Clasificarea țărilor prin legătură completă și metrică Mahalanobis**

Clasificarea în metrică mahalanobis furnizează clustere coerente cu varianță intra-cluster mică și varianță inter-clustere mare. În Figura 2 este prezentat graficul dendrogramă cu evidențierea partiției optimale.

În Tabelele 4 și 5 sunt prezentate clusterele pentru partiția optimală și cea cu patru clustere.

Tabel 4. Partiția optimală. Gruparea prin metrică Mahalanobis

|  |  |
| --- | --- |
| Cluster | Țări |
| c0 | Austria, Belgia, Croația, Cipru, Cehia, Danemarca, Estonia, Finlanda, Franța, Germania, Grecia, Ungaria, Irlanda, Italia, Letonia, Lituania, Luxemburg, Malta, Olanda, Polonia, România, Slovacia, Slovenia, Spania, Suedia, Marea Britanie |
| c1 | Bulgaria, Portugalia |

Tabel 5. Partiția cu patru clustere. Gruparea prin metrică Mahalanobis

|  |  |
| --- | --- |
| Cluster | Țări |
| c0 | Bulgaria |
| c1 | Portugalia |
| c2 | Austria, Danemarca, Finlanda, Franța, Luxemburg |
| c3 | Belgia, Croația, Cipru, Cehia, Estonia, Germania, Grecia, Ungaria, Irlanda, Italia, Letonia, Lituania, Malta, Olanda, Polonia, România, Slovacia, Slovenia, Spania, Suedia, Marea Britanie |

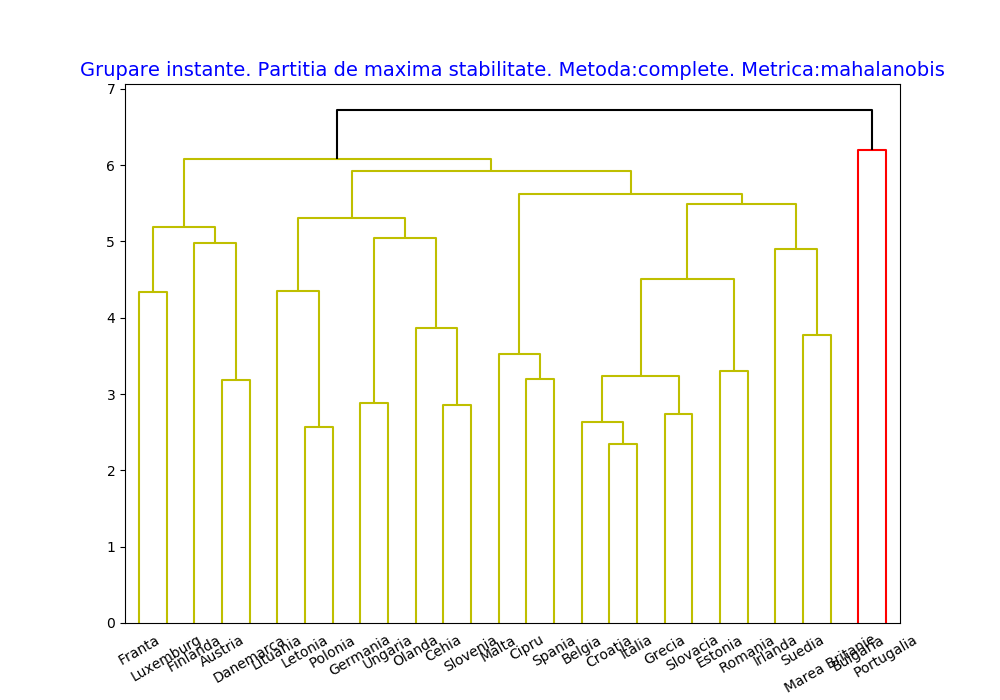


Figura 2. Grupare prin legătură completă și metrică mahalanobis

**3. Clasificarea variabilelor**

Deși se utilizează în general pentru clasificare de instanțe, analiza de clusteri poate fi utilizată și pentru grupare de variabile dacă sunt alese metrici potrivite. Pachetul Python scipy are implementate metrici pentru variabile, cum ar fi distanța bazată pe coeficienții de corelație.

În Figura 3 este prezentat graficul dendrogramă cu evidențierea partiției optimale. Metoda de grupare este prin media legăturilor.

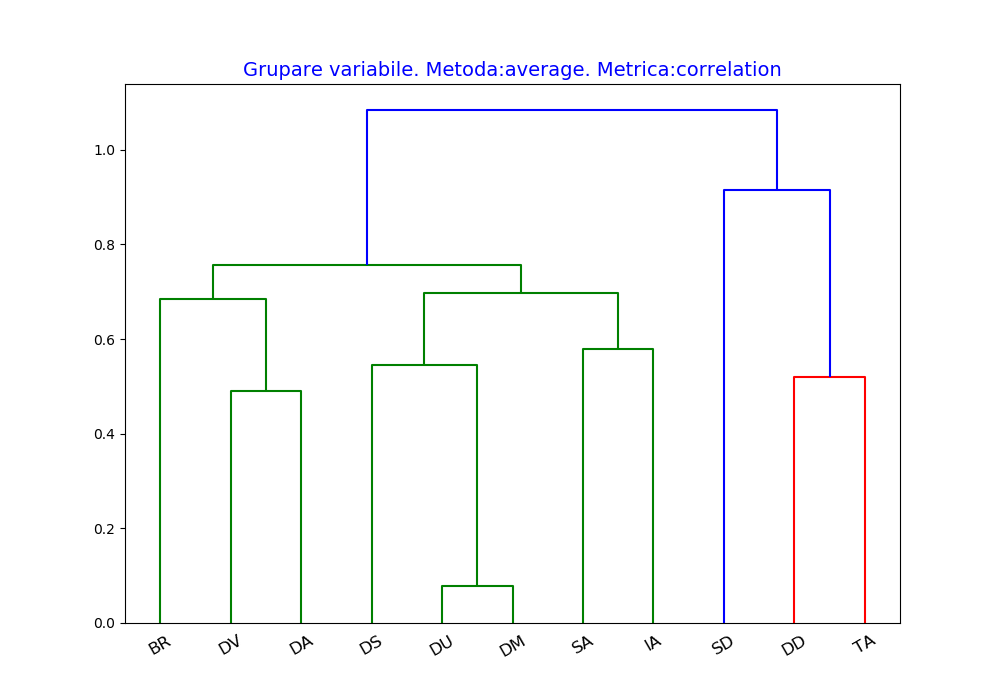
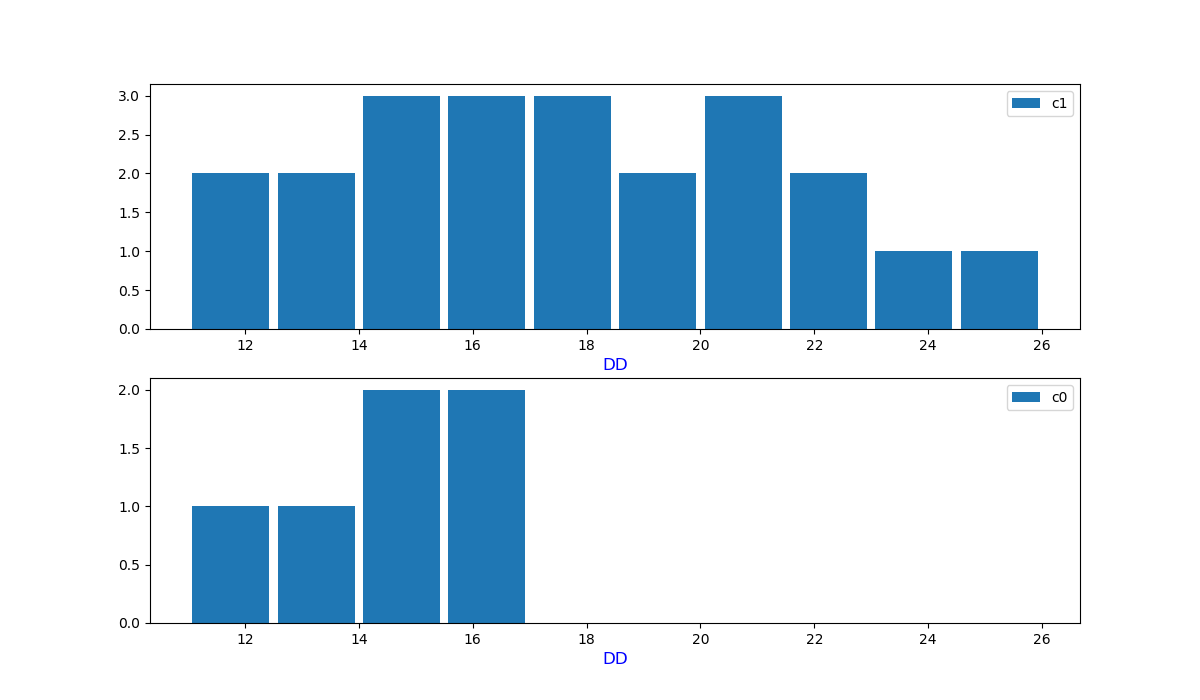
~~~~

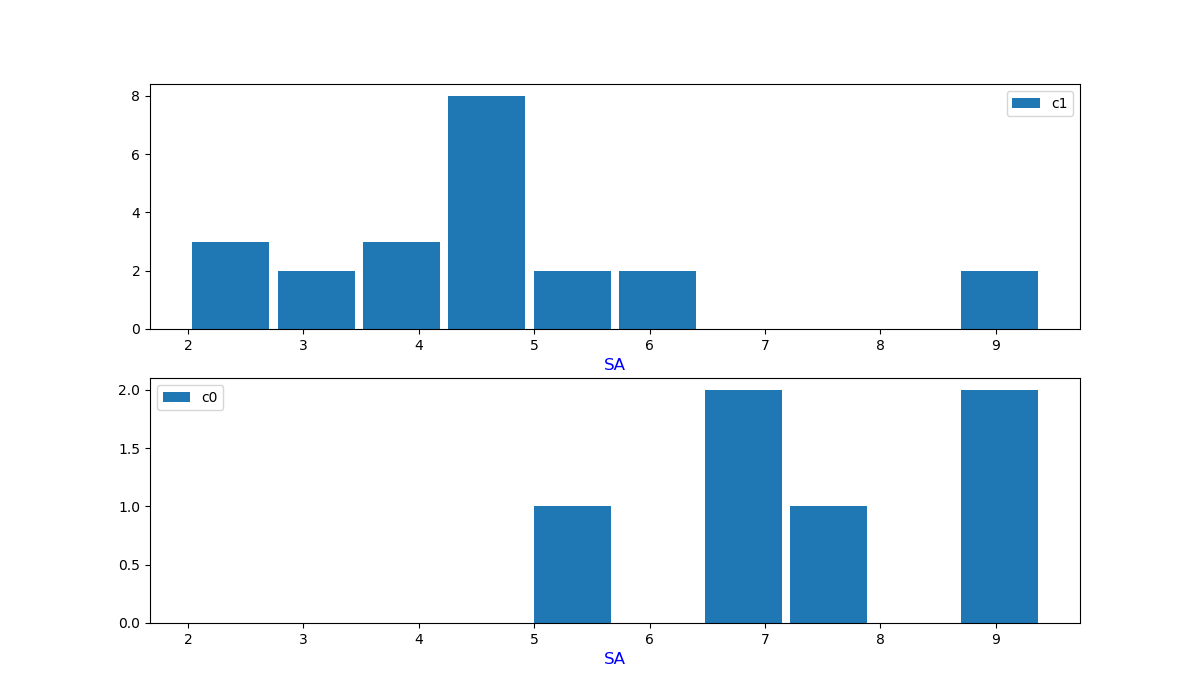
Figura 3. Gruparea variabilelor prin media legăturilor

**ANEXE**

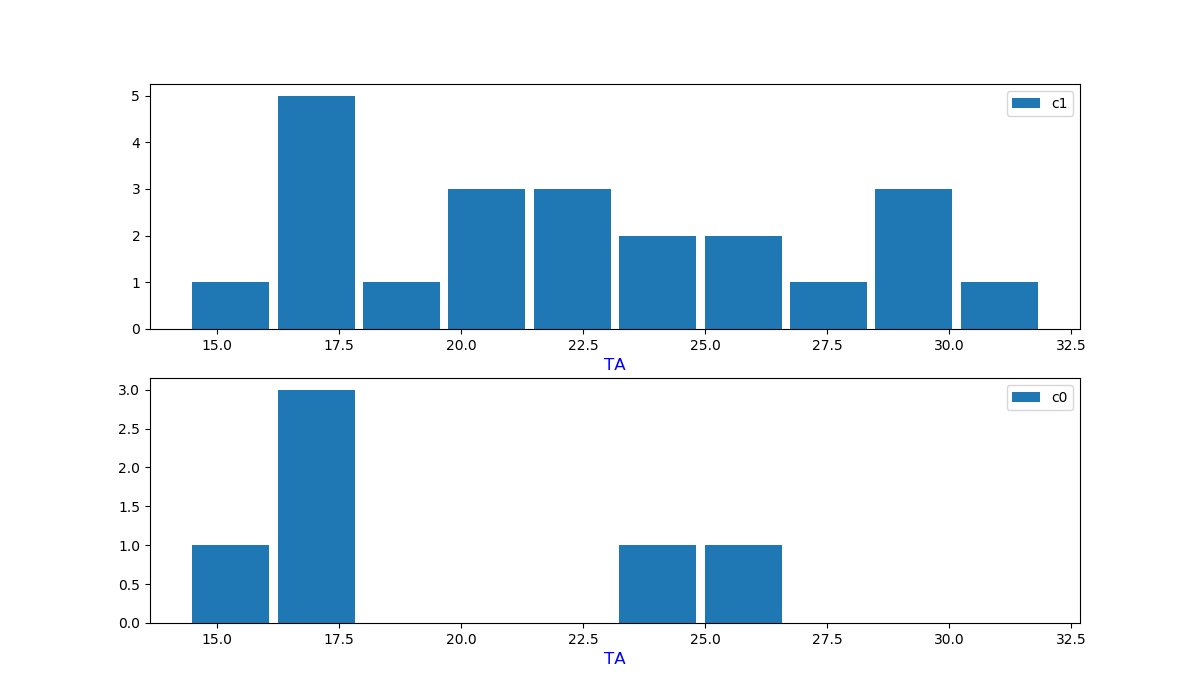
Anexa 1. Partiția optimală. Histogramă dificultăți de deplasare

****

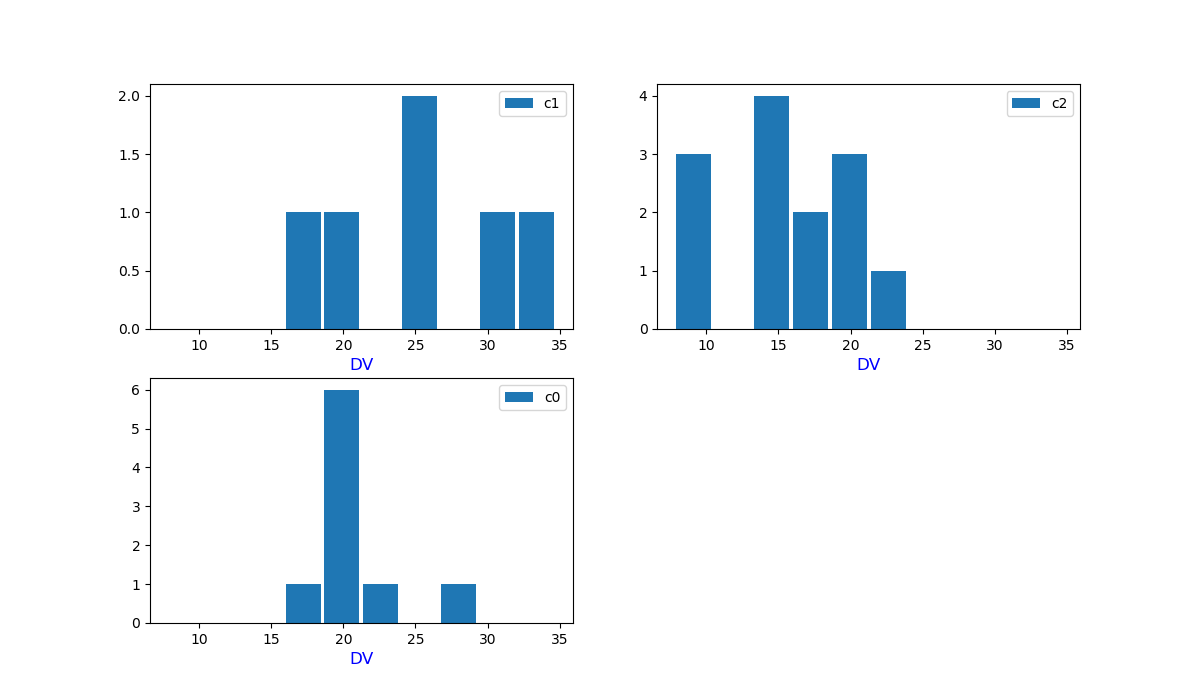
Anexa 2. Partiția optimală. Histogramă astm

****

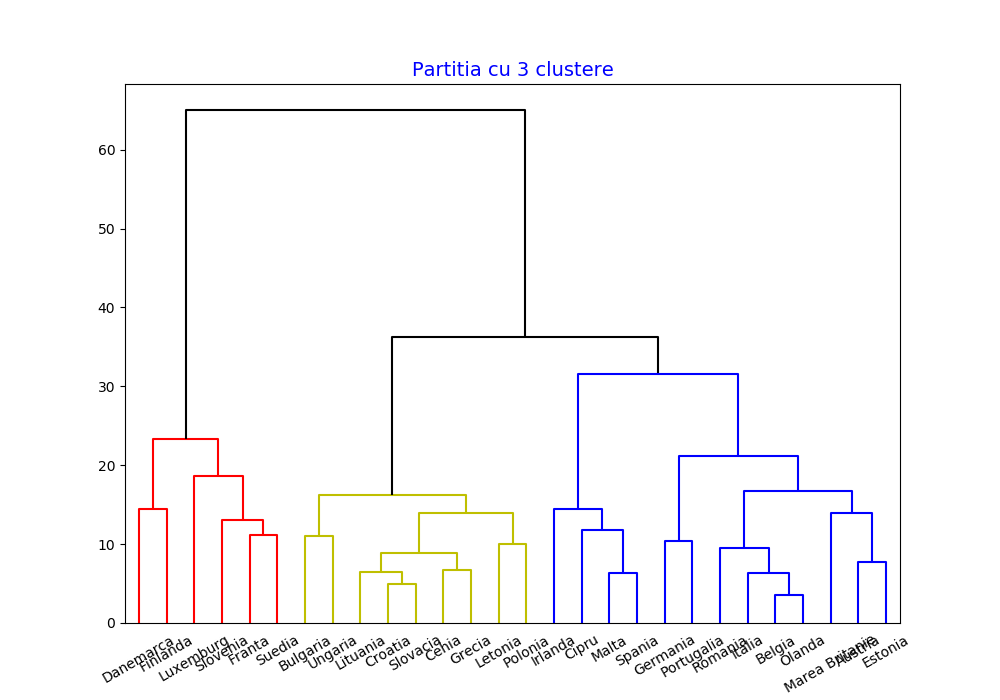
Anexa 3. Partiția optimală. Histogramă tensiune arterială crescută

****

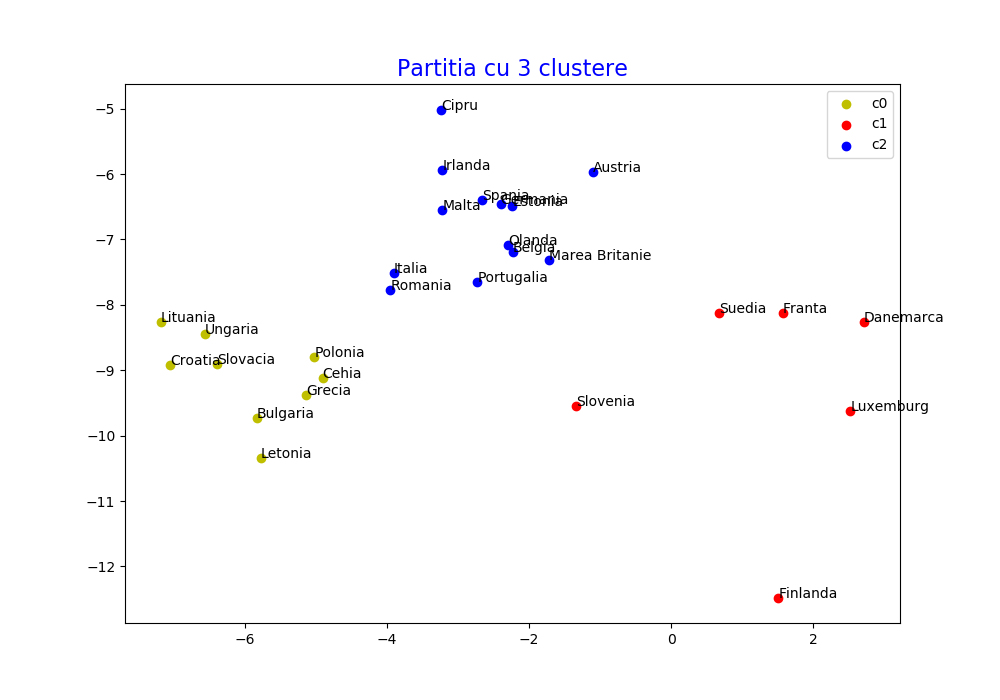
Anexa 4. Partiția cu 3 clustere. Histogramă deficiențe de vedere

****

Anexa 5. Dendrograma cu evidențierea partiției cu 3 clustere



Anexa 6. Reprezentare în axe discriminante a partiției cu trei clustere



1. ec.europa.eu/eurostat/data/database [↑](#footnote-ref-1)