
Academia de Studii Economice din București
Facultatea de Cibernetică, Statistică și Informatică Economică
Specializarea Informatică Economică

# Proiect la disciplina
# Dezvoltare Software pentru Analiza Datelor

Profesor coordonator:
Cadru didactic asociat Obretin Alexandru-Marius

Student:
Petrică Miruna-Alexandra
Anul III
Grupa 1094E

# Cuprins

# Introducere

Acest proiect urmărește să reprezinte și să elaboreze o analiză detaliată a datelor culese, prin realizarea atât a aplicării clusterizării datelor, cât și a reducerii dimensionalității setului specific de date. De asemenea, în cadrul proiectului vor fi realizare diagrame, pentru a ilustra cât mai bine relațiile dintre variabile. Atât algoritmii de implementare, cât și generarea graficelor se vor realiza în limbajul Python.

Setul de date pe care l-am ales pentru acest proiect prezintă numărul nașterilor în Japonia, între anii 1996 și 2022, în funcție de anumite criterii. Criteriile reprezintă variabilele independente utilizate, iar numărul de observații este dat de numărul anilor analizați, respectiv 27. De asemenea, datele descrise sunt relevante pentru perioada de timp respectivă din care au fost culese.

După cum am menționat anterior, observațiile sunt constituite din ani, pentru a putea urmări evoluția acestor date în timp și variațiile dintre ele.

Variabilele independente inițial utilizate sunt în număr de 11 :

- birth_total: numărul total de nașteri în Japonia, în anul respectiv
- birth_male: numărul total de copii de sex masculin născuți
- birth_female: numărul total de copii de sex feminin născuți
- population_total: populația totală
- population_male: populația de sex masculin
- population_female: populația de sex feminin
- firstborn: Numărul de copii născuți ca primii într-o familie
- secondborn: Numărul de copii din a doua naștere într-o familie
- thirdborn: Numărul de copii din a treia naștere într-o familie
- forthborn: Numărul de copii născuți în a patra naștere într-o familie
- fifthborn_and_above: Numărul de copii născuți în a cincea naștere sau mai mult

Pentru întregirea setului de date în scopul unei analize cât mai precise, la acestea se mai adaugă, în capitolul al doilea, respectiv în cazul aplicării metodei de redimensionare, următoarele:

- mother_age_20-24: Numărul de nașteri asociate mamelor cu vârsta cuprinsă între 20 și 24 de ani, în Japonia, în perioada respectivă
- mother_age_25-29: Numărul de nașteri asociate mamelor cu vârsta cuprinsă între 20 și 24 de ani

- mother_age_30-34: Numărul de nașteri asociate mamelor cu vârsta cuprinsă între 20 și 24 de ani
- mother_age_35-39: Numărul de nașteri asociate mamelor cu vârsta cuprinsă între 20 și 24 de ani

Toate datele au fost preluate de pe site-ul:
https://www.kaggle.com/datasets/webdevbadger/japan-birth-statistics .

În primul capitol am aplicat o metodă de clusterizare asupra primului set de date. Am folosit în acest scop analiza de cluster. Metoda de clusterizare este folosită în analiza datelor pentru a grupa observațiile sau obiectele similare între ele. Prin identificarea și formarea clusterelor, această tehnică ajută la evidențierea structurilor și a tiparelor subiacente în seturile de date.

În cel de-al doilea capitol am aplicat o metodă de reducere a dimensionalității datelor, folosind setul extins de date, cu 15 variabile independente. Pentru aceasta am utilizat formulele de analiză în componentele principale. Analiza în componentele principale (ACP) este o tehnică de reducere a dimensionalității și extragere a caracteristicilor dintr-un set de date, care se aplică, în general, în contextul explorării datelor. Componentele principale pot fi adesea interpretate mai ușor decât variabilele inițiale.

# 1. Clusterizare

Pentru clusterizare am folosit Analiza de cluster, aplicată pe primul set de date, prezentat în figura de mai jos. Am ales acest tip de analiză pentru a identifica grupuri omogene de populație în funcție de caracteristicile urmărite și pentru a compara și evidenția diferențele între acestea în funcție de momentul de timp.

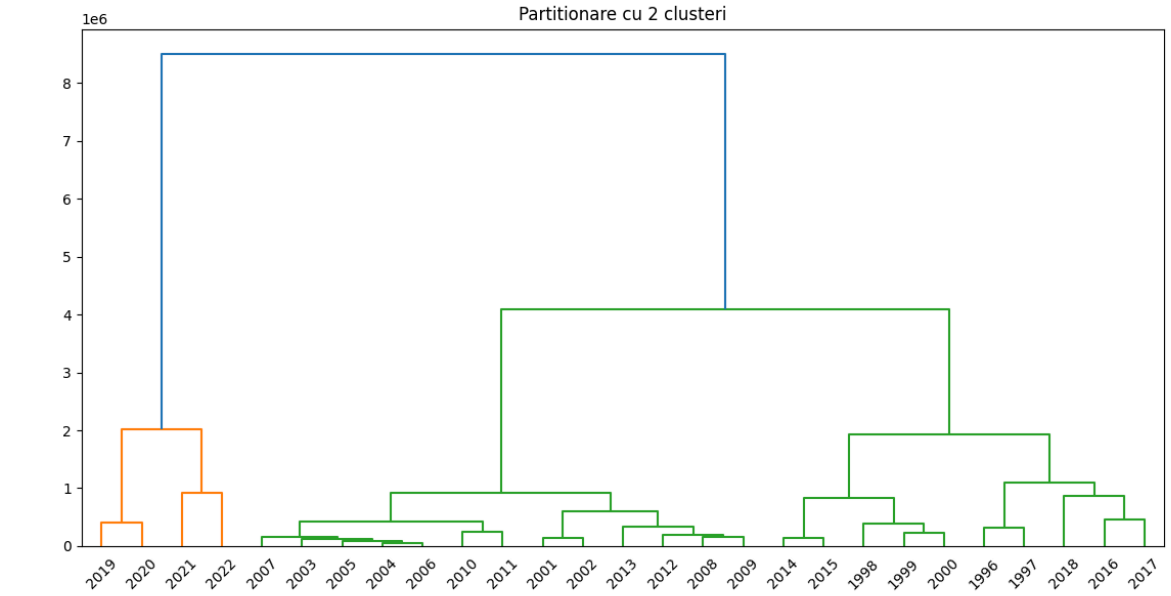
| year | birth_tota | birth_mal | birth_fem | population_tot | populatior | populatio | firstborn | secondboi | thirdborn | forthborn | fifthborn_and |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1996 | 1206555 | 619793 | 586762 | 124709000 | 61115000 | 63594000 | 574054 | 444571 | 154457 | 26532 | 6941 |
| 1997 | 1191665 | 610905 | 580760 | 124963000 | 61210000 | 63753000 | 571608 | 437120 | 150257 | 25845 | 6835 |
| 1998 | 1203147 | 617414 | 585733 | 125252000 | 61311000 | 63941000 | 583588 | 439459 | 148163 | 25230 | 6707 |
| 1999 | 1177669 | 604769 | 572900 | 125432000 | 61358000 | 64074000 | 579150 | 427385 | 140682 | 23738 | 6714 |
| 2000 | 1190547 | 612148 | 578399 | 125612633 | 61488005 | 64124628 | 583220 | 434964 | 141011 | 24644 | 6708 |
| 2001 | 1170662 | 600918 | 569744 | 125908000 | 61595000 | 64313000 | 573918 | 428197 | 137814 | 24058 | 6675 |
| 2002 | 1153855 | 592840 | 561015 | 126008000 | 61591000 | 64417000 | 571501 | 421042 | 131636 | 23037 | 6639 |
| 2003 | 1123610 | 576736 | 546874 | 126139000 | 61620000 | 64520000 | 547170 | 419100 | 128083 | 22615 | 6642 |
| 2004 | 1110721 | 569559 | 541162 | 126176000 | 61597000 | 64579000 | 537913 | 417647 | 126117 | 22458 | 6586 |
| 2005 | 1062530 | 545032 | 517498 | 126204902 | 61617893 | 64587009 | 512412 | 399307 | 122501 | 21841 | 6469 |
| 2006 | 1092674 | 560439 | 532235 | 126154000 | 61568000 | 64586000 | 524581 | 408531 | 129555 | 23164 | 6843 |
| 2007 | 1089818 | 559847 | 529971 | 126085000 | 61511000 | 64574000 | 519767 | 403656 | 134951 | 24203 | 7241 |
| 2008 | 1091156 | 559513 | 531643 | 125947000 | 61424000 | 64523000 | 517724 | 402152 | 137945 | 25804 | 7531 |
| 2009 | 1070036 | 548994 | 521042 | 125820000 | 61339000 | 64481000 | 512743 | 390073 | 134171 | 25303 | 7746 |
| 2010 | 1071305 | 550743 | 520562 | 126381728 | 61571727 | 64810001 | 509736 | 390213 | 136302 | 26885 | 8169 |
| 2011 | 1050807 | 538271 | 512536 | 126180000 | 61453000 | 64727000 | 494713 | 383666 | 136687 | 27280 | 8461 |
| 2012 | 1037232 | 531781 | 505451 | 125957000 | 61328000 | 64630000 | 484711 | 382461 | 134339 | 27218 | 8503 |
| 2013 | 1029817 | 527657 | 502160 | 125704000 | 61186000 | 64518000 | 481418 | 379467 | 133183 | 27193 | 8556 |
| 2014 | 1003609 | 515572 | 488037 | 125431000 | 61041000 | 64391000 | 474229 | 364787 | 129687 | 26449 | 8457 |
| 2015 | 1005721 | 515468 | 490253 | 125319299 | 61022756 | 64296543 | 478101 | 363244 | 129708 | 26271 | 8397 |
| 2016 | 977242 | 502012 | 475230 | 125020252 | 60866773 | 64153479 | 459873 | 355876 | 126570 | 26427 | 8496 |
| 2017 | 946146 | 484478 | 461668 | 124648471 | 60675736 | 63972735 | 439295 | 348859 | 123050 | 26230 | 8712 |
| 2018 | 918400 | 470851 | 447549 | 124218285 | 60454898 | 63763387 | 426407 | 338094 | 119732 | 25546 | 8621 |
| 2019 | 865239 | 443430 | 421809 | 123731176 | 60208034 | 63523142 | 400952 | 315713 | 114630 | 25106 | 8838 |
| 2020 | 840835 | 430713 | 410122 | 123398962 | 60002838 | 63396124 | 392538 | 304028 | 110818 | 24788 | 8663 |
| 2021 | 811622 | 415903 | 395719 | 122780487 | 59686643 | 63093844 | 372434 | 294444 | 109953 | 25549 | 9242 |
| 2022 | 770759 | 395257 | 375502 | 122030523 | 59313678 | 62716845 | 355523 | 281418 | 101233 | 23677 | 8908 |

în primul rând, am realizat ierarhia de clusterizare. În urma acesteia a rezultat că diferența maximă este de 24, ceea ce a indicat alegerea a doi clusteri în acest caz. Indexul diferenței maxime de 24 indică punctul în ierarhie unde diferența între clustere a atins o valoare maximă. Este ales ca punct de tăiere pentru formarea clusterelor.

Numărul de clusteri, respectiv 2, indică tăierea setului de date la observația cu numărul 24. Cei doi clusteri reprezenta două grupuri diferite de observații în funcție de atributele folosite. De exemplu, în contextul nașterilor, acești doi clusteri indică două modele diferite legate de nașteri în setul de date, observațiile 24-27 fiind cele mai recente, din anii 2020-2022. Se remarcă astfel faptul că diferența dintre cei doi clusteri este dată de anul 2020.

Ulterior, am realizat o partiție optimă și o partiție pentru 3 clusteri și le-am reprezentat prin dendrograme.

Partitie optima ( cu 2 clusteri): ['cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod1' 'cod1' 'cod1' 'cod1']

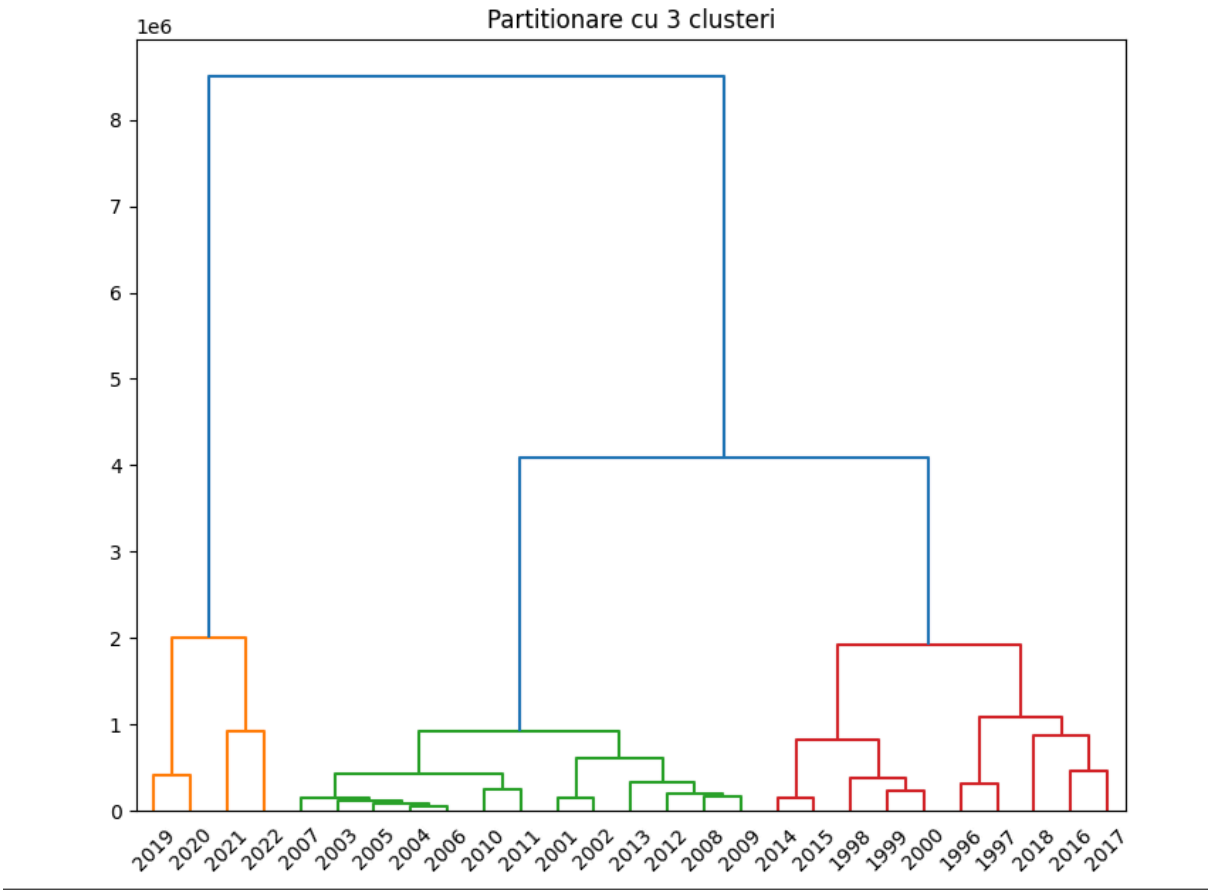



Dendrograma este rezultată dintr-o analiză de cluster ierarhică.

Observațiile sunt împărțite în două clustere mari. Clusterul cu linii verzi conține majoritatea observațiilor, indicând o omogenitate între acestea. Observațiile din acest cluster poate reprezinta ani în care s-a înregistrat un număr de nașteri similar cu anii precedenți.

Clusterul cu linii portocalii include doar câteva observații și este separat semnificativ de restul, indicând o diferență distinctă între caracteristicile acestor observații și restul setului de date.

Partitie cu 3 clusteri: ['cod2' 'cod2' 'cod2' 'cod2' 'cod2' 'cod1' 'cod1' 'cod1' 'cod1' 'cod1'
'cod1' 'cod1' 'cod1' 'cod1' 'cod1' 'cod1' 'cod1' 'cod1' 'cod2' 'cod2'
'cod2' 'cod2' 'cod2' 'cod3' 'cod3' 'cod3' 'cod3']




Aici, observațiile sunt împărțite în trei clustere. Primul cluster este același ca și în partiționarea cu 2 clustere, sugerând că aceste observații sunt foarte diferite de restul.

Clusterul verde este împărțit în două: un cluster verde și unul roșu. Această divizare arată că, chiar și în cadrul grupului mare de observații anterioare, există suficientă variație pentru a justifica subdivizarea în clustere suplimentare.Cu alte cuvinte, dinamica populației în Japonia diferă semnificativ în funcție de perioada de timp.

# 2. Reducerea dimensionalității setului de date

Pentru acest capitol am utilizat setul extins de date, cu cele 15 variabile independente, după cum sunt prezentate și în figura următoare

| year | birth_total | birth_male | birth_female | population_total | population_male | population_female | firstborn | secondborn | thirdborn | forthborn | fifthborn_and_above | mother_age_20 | mother_age_25 | mother_age_30 | mother_age_35 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1996 | 1206555 | 619793 | 586762 | 124709000 | 61115000 | 63594000 | 574054 | 444571 | 154457 | 26532 | 6941 | | | | |
| 1997 | 1191665 | 610905 | 580760 | 124963000 | 61210000 | 63753000 | 571608 | 437120 | 150257 | 25845 | 6835 | | | | |
| 1998 | 1203147 | 617414 | 585733 | 125252000 | 61311000 | 63941000 | 583588 | 439459 | 148163 | 25230 | 6707 | | | | |
| 1999 | 1177669 | 604769 | 572900 | 125432000 | 61358000 | 64074000 | 579150 | 427385 | 140682 | 23738 | 6714 | | | | |
| 2000 | 1190547 | 612148 | 578399 | 125612633 | 61488005 | 64124628 | 583220 | 434964 | 141011 | 24644 | 6708 | 161361 | 470833 | 396901 | 126409 |
| 2001 | 1170662 | 600918 | 569744 | 125908000 | 61595000 | 64313000 | 573918 | 428197 | 137814 | 24058 | 6675 | | | | |
| 2002 | 1153855 | 592840 | 561015 | 126008000 | 61591000 | 64417000 | 571501 | 421042 | 131636 | 23037 | 6639 | | | | |
| 2003 | 1123610 | 576736 | 546874 | 126139000 | 61620000 | 64520000 | 547170 | 419100 | 128083 | 22615 | 6642 | | | | |
| 2004 | 1110721 | 569559 | 541162 | 126176000 | 61597000 | 64579000 | 537913 | 417647 | 126117 | 22458 | 6586 | | | | |
| 2005 | 1062530 | 545032 | 517498 | 126204902 | 61617893 | 64587009 | 512412 | 399307 | 122501 | 21841 | 6469 | 128135 | 339328 | 404700 | 153440 |
| 2006 | 1092674 | 560439 | 532235 | 126154000 | 61568000 | 64586000 | 524581 | 408531 | 129555 | 23164 | 6843 | | | | |
| 2007 | 1089818 | 559847 | 529971 | 126085000 | 61511000 | 64574000 | 519767 | 403656 | 134851 | 24203 | 7241 | | | | |
| 2008 | 1091156 | 559513 | 531643 | 125947000 | 61424000 | 64523000 | 517724 | 402152 | 137945 | 25804 | 7531 | | | | |
| 2009 | 1070036 | 548994 | 521042 | 125820000 | 61339000 | 64481000 | 512743 | 390073 | 134171 | 25303 | 7746 | 116808 | 307765 | 389793 | 209707 |
| 2010 | 1071305 | 550743 | 520562 | 126381728 | 61571727 | 64810001 | 509736 | 390213 | 136302 | 26885 | 8169 | 110956 | 306910 | 384386 | 220101 |
| 2011 | 1050807 | 538271 | 512536 | 126180000 | 61453000 | 64727000 | 494713 | 383666 | 136687 | 27280 | 8461 | 104060 | 300384 | 373490 | 221272 |
| 2012 | 1037232 | 531781 | 505451 | 125957000 | 61328000 | 64630000 | 484711 | 382461 | 134339 | 27218 | 8503 | 95805 | 292464 | 367715 | 225480 |
| 2013 | 1029817 | 527657 | 502160 | 125704000 | 61186000 | 64518000 | 481418 | 379467 | 133183 | 27193 | 8556 | 91251 | 282794 | 365404 | 229741 |
| 2014 | 1003609 | 515572 | 488037 | 125431000 | 61041000 | 64391000 | 474229 | 364787 | 129687 | 26449 | 8457 | 86600 | 267866 | 359348 | 225899 |
| 2015 | 1005721 | 515468 | 490253 | 125319299 | 61022756 | 64296543 | 478101 | 363244 | 129708 | 26271 | 8397 | 84465 | 262266 | 364887 | 228302 |
| 2016 | 977242 | 502012 | 475230 | 125020252 | 60866773 | 64153479 | 459873 | 355876 | 126570 | 26427 | 8496 | 82194 | 250715 | 355018 | 223329 |
| 2017 | 946146 | 484478 | 461668 | 124648471 | 60675736 | 63972735 | 439295 | 348859 | 123050 | 26230 | 8712 | 79272 | 240959 | 345441 | 216954 |
| 2018 | 918400 | 470851 | 447549 | 124218285 | 60454898 | 63763387 | 426407 | 338094 | 119732 | 25546 | 8621 | 77023 | 233754 | 334906 | 211021 |
| 2019 | 865239 | 443430 | 421809 | 123731176 | 60208034 | 63523142 | 400952 | 315713 | 114630 | 25106 | 8838 | 72092 | 220933 | 312582 | 201010 |
| 2020 | 840835 | 430713 | 410122 | 123398962 | 60002838 | 63396124 | 392538 | 304028 | 110818 | 24788 | 8663 | 66751 | 217804 | 303436 | 196321 |
| 2021 | 811622 | 415903 | 395719 | 122780487 | 59686643 | 63093844 | 372434 | 294444 | 109953 | 25549 | 9242 | 59896 | 210433 | 292439 | 193177 |
| 2022 | 770759 | 395257 | 375502 | 122030523 | 59313678 | 62716845 | 355523 | 281418 | 101233 | 23677 | 8908 | 52850 | 202505 | 279517 | 183327 |

Pentru reducerea dimensionalității setului de date am folosit analiza de componente principale (ACP). Aceasta oferă o modalitate de a simplifica și sintetiza informații complexe, menținând o mare parte a variației inițiale a datelor.

.Am optat pentru această metodă deoarece Analiza de Componente Principale permite reducerea numărului de variabile inițiale, transformând datele într-un set de componente principale. Pe scurt, componentele respective capturează cea mai mare variație a datelor, permițând astfel o reprezentare mai compactă. Prin urmare, ACP poate ajuta la identificarea variabilelor care contribuie cel mai mult la variația datelor.

Reducerea dimensionalității permite vizualizarea mai ușoară a datelor în spații bidimensionale sau tridimensionale.

Pentru realizarea acestei analize, pentru început am standardizat valorile din setul de date, aducându-le la aceeași scară, astfel încât acestea să aibă media zero și deviația standard egală cu 1. Apoi, am aplicat Analiza în Componentele Principale pe setul de date standardizat. Astfel, am calculat valorile proprii, vectorii proprii și componentele principale rezultate din ACP.

În urma analizei au rezultat următoarele

➔ Valorile proprii:

[1.9275e+12 4.9488e+10 1.9975e+09 6.8475e+08 1.2650e+08 3.3243e+07
2.0006e+07 9.0236e+06 3.8947e+06 4.4347e+05 9.7458e+04 2.4763e+04
4.1128e+03 1.9158e-21 1.3411e-23]

Valorile proprii reprezintă importanța relativă a fiecărei componente principale în explicarea variației datelor. Valorile proprii mai mari indică o contribuție mai semnificativă a componentei respective în explicarea variației.

În acest caz, prima valoare proprie este mult mai mare decât celelalte, sugerând că prima componentă principală explică o parte semnificativă a variației. Valorile proprii mici, cum ar fi cele din partea de jos ale listei, indică componente care contribuie mai puțin la variația datelor.

Componenta principală asociată primei valori proprii mari este cea mai semnificativă și aduce cu sine cea mai mare cantitate de informație.

Ultimele două valori proprii sunt extrem de mici și pot fi considerate aproape zero. Aceasta indică faptul că aceste componente nu aduc o contribuție semnificativă la variația totală și pot fi considerate neglijabile. Acest lucru poate fi evidențiat și de faptul că valorile proprii sunt foarte aproape de zero,, ceea ce sugerează o variație foarte mică asociată acestor componente.

➔ Vectorii proprii:

[[-0.0677 -0.0349 -0.0328 -0.8112 -0.4449 -0.3664 -0.0365 -0.026 -0.0057 0.0001 0.0004 -0.0102 -0.0217 -0.0176 -0.002 ]
[ 0.3511 0.1812 0.1699 -0.0786 0.5704 -0.6493 0.1955 0.1289 0.0316 -0.002 -0.0029 0.0215 0.0612 0.0042 -0.043 ]
[-0.1113 -0.0538 -0.0575 -0.0084 -0.0062 -0.0028 -0.0418 -0.0434 -0.0243 -0.0017 -0.0001 0.3198 0.8232 0.2456 -0.3698]
[-0.5812 -0.3006 -0.2806 0.0462 0.3286 -0.2838 -0.2028 -0.1187 -0.2009 -0.0473 -0.0115 -0.0442 -0.2315 -0.1597 -0.3513]
[-0.1249 -0.0735 -0.0514 0.0026 0.1154 -0.1164 -0.2754 -0.053 0.1533 0.0404 0.0098 0.177 -0.1983 0.8142 0.3347]
[-0.0425 0.0071 -0.0496 -0.0071 -0.0647 0.0532 0.7623 -0.3295 -0.3644 -0.0917 -0.0192 0.0664 -0.182 0.3165 -0.1404]
[ 0.2216 0.1015 0.1201 -0.0081 -0.1329 0.1162 -0.1452 0.3839 0.0503 -0.0437 -0.0237 0.0182 -0.3533 0.2636 -0.7259]
[-0.0971 -0.0699 -0.0272 -0.0157 0.0023 0.0134 0.0084 0.6141 -0.5826 -0.1109 -0.0261 -0.3792 0.1857 0.1773 0.2127]
[-0.022 -0.0187 -0.0033 0.0007 -0.0252 0.0122 0.0108 0.3171 -0.2533 -0.0815 -0.015 0.8451 -0.1774 -0.2232 0.1766]
[-0.0164 0.6962 -0.7126 -0.033 0.0348 0.0317 -0.034 0.0424 0.0097 -0.0284 -0.006 -0.0038 -0.002 0.0033 0.0074]
[-0.1588 -0.1038 -0.0549 -0.0104 0.0129 0.0083 0.2089 0.1892 0.4524 -0.7897 -0.2195 -0.0289 0.0359 -0.0044 0.0516]
[ 0.0078 -0.0359 0.0437 -0.5761 0.5757 0.5764 -0.0045 -0.0165 -0.0034 0.0079 0.0243 0.0149 -0.0079 -0.0013 -0.0077]
[-0.0906 -0.0448 -0.0459 -0.013 0.0127 0.0132 0.1344 0.1265 0.1201 0.4074 -0.879 0.005 0.0024 -0.0034 0.0086]
[-0.4817 0.5938 0.5938 -0. 0. 0. -0.1121 -0.1121 -0.1121 -0.1121 -0.1121 0. -0. -0. -0. ]

[ 0.424 -0.0191 -0.0191 -0. 0. 0. -0.4048 -0.4048 -0.4048 -0.4048 -0.4048 -0. 0. 0. 0. ]]

Acești vectori proprii reprezintă coeficienții pentru fiecare variabilă în cadrul fiecărei componente principale. Aceștia indică direcția și magnitudinea contribuției fiecărei variabile la formarea fiecărei componente principale. În general, vectorii proprii oferă informații despre cum fiecare variabilă contribuie la fiecare componentă principală.

În prima componentă principală, se regăsesc valori mari și negative pentru variabilele "birth_total", "birth_male", "birth_female", "population_total", "population_male", și "population_female". Acest lucru sugerează că aceste variabile sunt corelate negativ cu prima componentă principală, adică o creștere a acestora este asociată cu o scădere a valorii primei componente principale.

De asemenea, conform vectorului propriu corespunzător primei valori proprii mari, aceasta este influențată în special de variabilele legate de numărul total de nașteri și populație.

Componentele principale rezultate în urma analizei sunt în număr de 13, însă sunt importante și semnificative primele componente principale, deoarece acestea sunt cele care capturează cea mai mare parte a variației din setul de date.

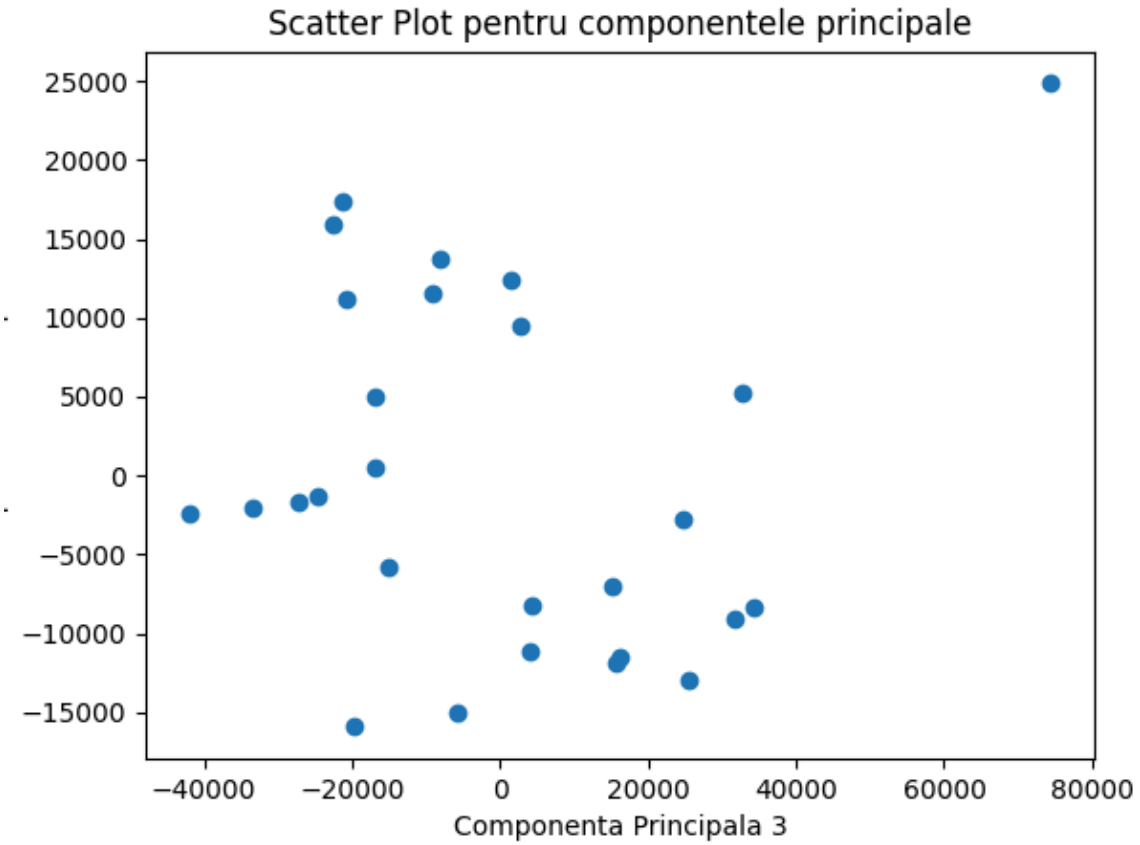
Relațiile dintre variabilele inițiale și elementele rezultate în urma analizei de componentă principală (ACP) pot fi înțelese prin observarea vectorilor proprii ai acestor componente principale. Vectorii proprii, reprezintă direcțiile în spațiul variabilelor originale în care variația este maximă. Astfel, pentru a ști care variabilă are o influență mai mare asupra primei componente principale, examinăm primul vector propriu.

Specific setului de date analizat, pornind de la primul vector propriu, putem ajunge la câteva concluzii. Valorile negative sugerează o relație inversă cu prima componentă principală, în timp ce valorile pozitive indică o relație directă. Variabilele cu coeficienți negativi (birth_total, birth_male, birth_female, firstborn) au o relație inversă cu prima componentă principală. Cu alte cuvinte, scăderea acestor variabile este asociată cu o creștere a valorii primei componente principale. Variabilele cu coeficienți negativi mai mari în magnitudine (population_total, population_male, population_female) au o contribuție mai semnificativă la direcția opusă primei componente principale. Deci, o scădere mai mare în aceste variabile duce la o creștere mai mare a valorii primei componente principale.

În continuare, în acest capitol am utilizat Criteriul Kaiser pentru a alege numărul de componente principale semnificative în analiza componentelor principale (ACP). Criteriul Kaiser sugerează păstrarea tuturor componentelor principale care au o valoare proprie mai mare decât 1. Numărul de componente principale semnificative este determinat de numărul de valori proprii care sunt mai mari decât 1. În urma aplicării Criteriului Kaiser au rezultat 13 componente principale semnificative, deoarece acestea au valori proprii mai mari decât 1.

În cadrul acestei analize am realizat și două grafice, care să evidențieze mai bine componentele.

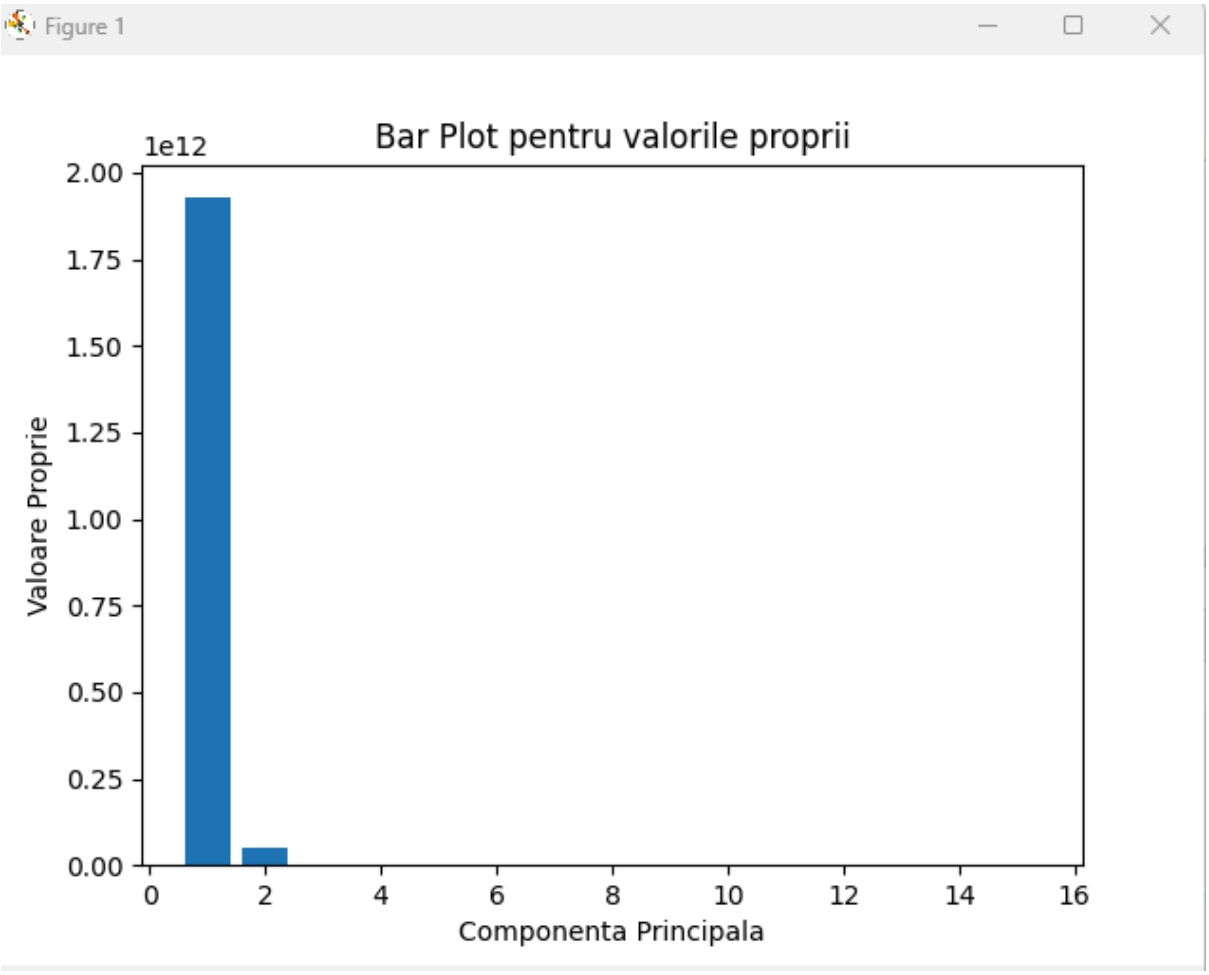
Primul grafic reprezinta un Scatter Plot realizat între componentele 3 și 4:




Prin acest grafic se poate evidenția dispersia semnificativă între primele două categorii de puncte, iar aceasta poate sugera o variație semnificativă în date în această porțiune.

Ultima categorie de date, depărtată de restul de puncte, poate indica o variație mare pentru această categorie în comparație cu celelalte.

Al doilea grafic reprezintă un Bar Plot :




Un astfel de comportament valorile proprii în cadrul analizei de componentă indică faptul că există o componentă principală dominantă, iar apoi valorile proprii ale celorlalte componente sunt mult mai mici, sugerând că acestea aduc o contribuție mai mică la variabilitatea totală a datelor. Valoarea proprie scade semnificativ de la prima la a doua componentă principală.

Primele componente principale cu valorile proprii mari reprezintă direcțiile principale ale variabilității în date. Cele cu valori proprii mici contribuie la detalii mai fine, dar nu aduc o contribuție semnificativă la variabilitatea globală.

# Concluzii

În cadrul acestui proiect, am utilizat cele două metode fundamentale de analiză a datelor: clusterizarea și Analiza Componentelor Principale. Aplicând metoda ierarhică de clusterizare, am identificat doi clusteri semnificativi, dar am evidențiat și posibilitatea de a avea un al treilea cluster. Primii doi clusteri au prezentat caracteristici similare între observații, în timp ce al treilea cluster a evidențiat particularități distincte. Aceste rezultate sugerează existența unor segmente omogene în cadrul setului de date.

Am utilizat Analiza Componentelor Principale (ACP) pentru a reduce dimensionalitatea setului de date și a evidenția direcțiile principale de variație. Valorile proprii și vectorii proprii furnizate de ACP au oferit informații despre importanța relativă a variabilelor și relațiile dintre acestea. Primele componente principale au explicat o proporție semnificativă a variației totale.

În sinteză, clusterizarea și ACP au oferit perspective complementare, contribuind la identificarea diferențelor semnificative și a similarităților la nivelul dinamicii nașterilor din Japonia .