ACADEMIA DE STUDII ECONOMICE DIN BUCURESTI

FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ

Analiza cauzelor de deces
(Tema 1- Reducerea dimensionalității)

Iordache Andreea

An universitar: 2023 – 2024

Cuprins

Prezentarea datelor

Sursa date: https://www.kaggle.com/datasets/iamsouravbanerjee/cause-of-deaths-around-the-world . Setul de date extras a fost prelucrat: am păstrat cel mai recent an (2019) și am eliminat coloanele cu o foarte mare corelație pentru o mai bună claritate a graficelor.

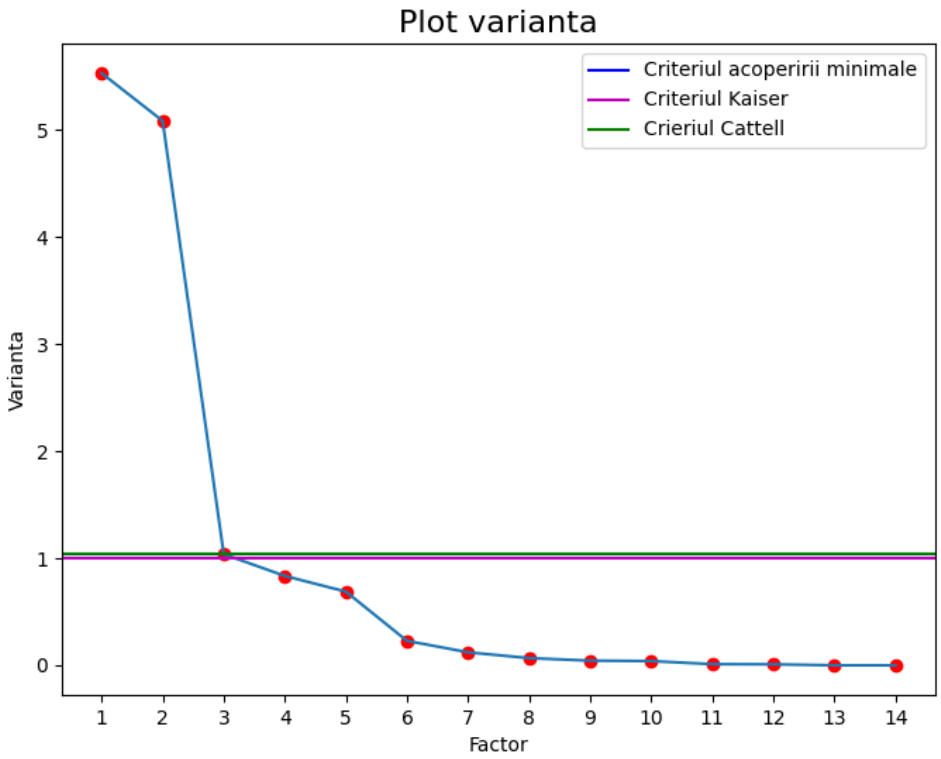
Fișierul de intrare folosit este: cause_of_deaths_recent_year.csv

1. Analiza in componente principale

1.1 Calculul variantei

Componenta	Varianta	Varianta cumulata	Procent varianta	Procent cumulat
C1	9.108242591805299	9.108242591805299	65.05887565575213	65.05887565575213
C2	2.169736109644111	11.27797870144941	15.498115068886506	80.55699072463864
C3	1.3008483613157575	12.578827062765168	9.291774009398265	89.84876473403689
C4	0.5173733248965071	13.096200387661675	3.6955237492607647	93.54428848329766
C5	0.382605130261038	13.478805517922712	2.732893787578843	96.27718227087651
C6	0.23057996287296315	13.709385480795675	1.6469997348068794	97.92418200568338
C7	0.1362330935646497	13.845618574360325	0.9730935254617833	98.89727553114516
C8	0.059389732316688225	13.905008306677013	0.42421237369063014	99.3214879048358
C9	0.038012496055851726	13.943020802732866	0.2715178289703694	99.59300573380617
C10	0.023177861024347927	13.966198663757213	0.1655615017391373	99.75856188398008
C11	0.014591649078600237	13.980790312835813	0.10422606484714454	99.86278794882723
C12	0.013050240754935272	13.993840553590749	0.09321600539239479	99.95600395421963
C13	0.005409564433006381	13.999250118023756	0.03863974595004557	99.99464370016968
C14	0.0007498819762460047	14.000000000000002	0.005356299830328605	100.0

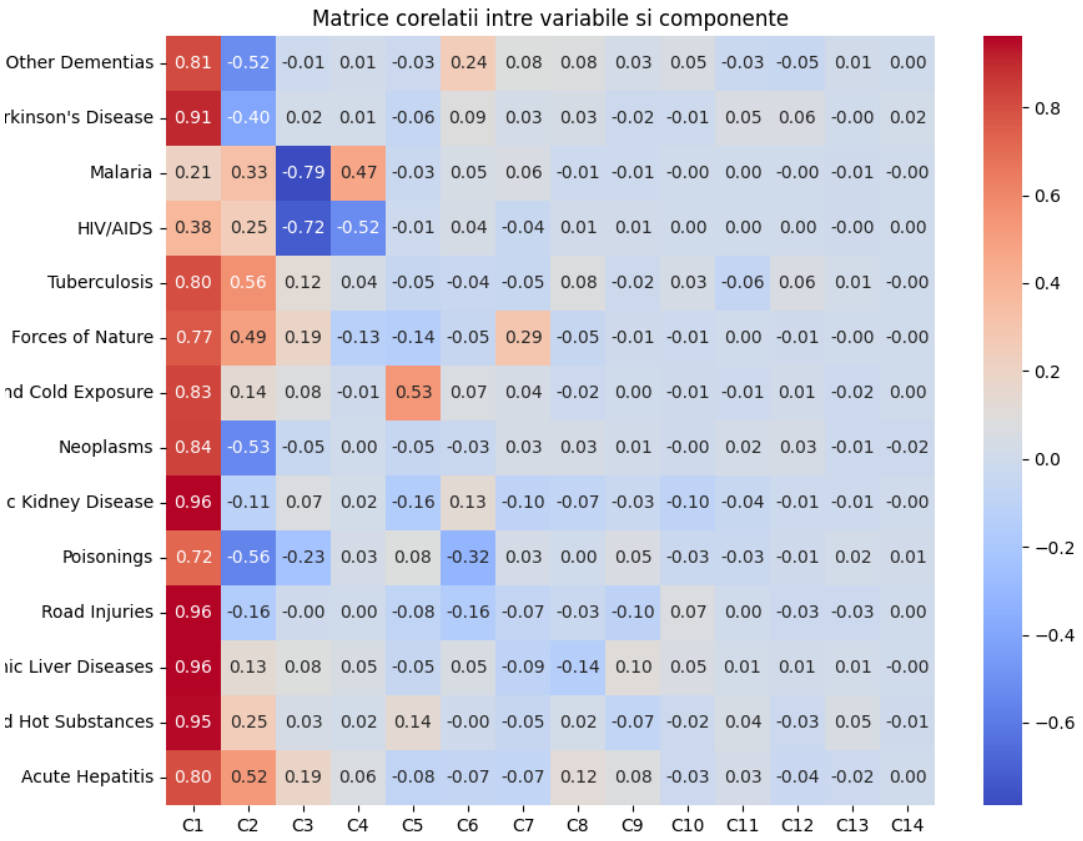
Tabel 1 – Distribuția variantei



Grafic 1 – Plot varianta – evidentierea criteriilor de selectie a componentelor semnificative

Conform tabelului, graficului și criteriilor Kaiser și Cattell, primele 3 componente sunt semnificative, având o varianta mai mare de 1 (9.1, 2.16, respectiv 1.3 potrivit tabelului varianta.csv, aflat în directorul tabele\csv). Deoarece primele 2 compnente au variante semnificativ mai mari, ele explică cea mai mare parte a variației în date. Prima variabilă acoperă 65.05% din variabilitate, cea de a două variabilă 15.49%, iar cea de-a treia 9,29%.

1.2 Corelațiile dintre variabilele observate și componentele principale



Tabel 2 – Corelograma dintre variabilele observate și componentele principale

Tabelul se regăsește în directorul acp/tabele/csv/corelatii.csv.

Matricea de corelații prezentată mai sus ilustrează legăturile (corelația) dintre diferitele motive de deces din setul de date și componente. O valoare apropiată de 1 (colorată cu roșu aprins)

indică o legătură puternică și directă, în timp ce o valoare apropiată de -1 (colorată cu albastru închis) indică o legătură puternică, dar inversă.

Pentru prima componentă, C1, se remarcă o legătură puternică, directă cu Chronic Liver Disease, adică boală cronică de rinichi, având o valoare pozitivă de 0.96. Alte cauze de deces cu o corelație puternică față de C1 sunt: accidente rutiere (Road injuries), având aceeași valoare- 0.96, expunerea la căldură și rece (Environmental Heat and Cold Exposure) - 0.95 și Parkinson's Disease - 0.91.

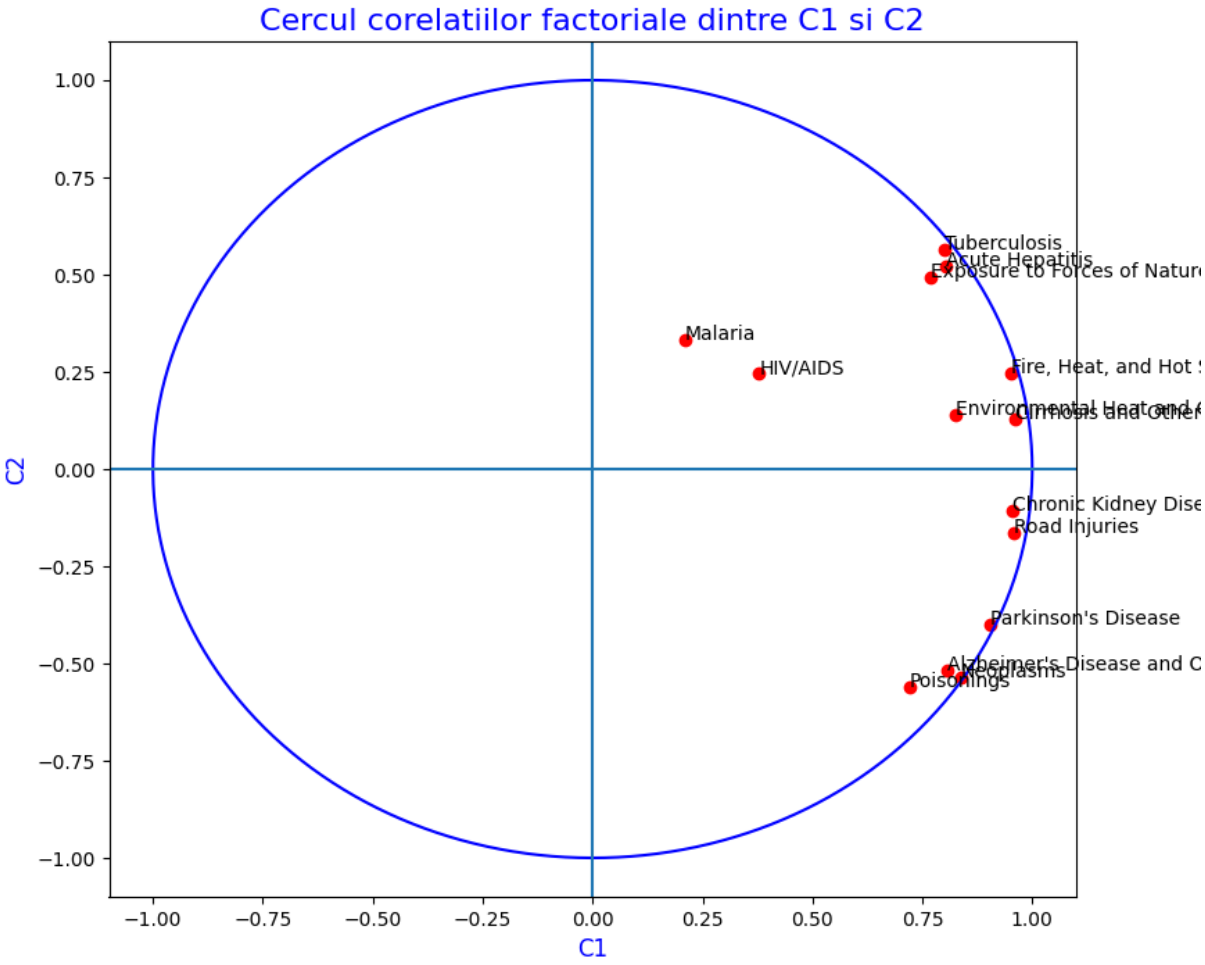
Pentru cea de-a două componentă se remarcă legături inverse, puternice, cu variabile precum otrăviri (Poisonings) , cu o corelație de -0.56, demență (Other Dementias), cu o valoare de -0.52. Se remarcă de asemenea o legătură pozitivă cu hepatită acută (Acute Hepatatis), cu o valoare de 0.52.

O corelație mare între o componentă și o variabilă implică, de obicei, faptul că schimbările în valoarea variabilei sunt explicate într-o mare măsură de schimbările în valoarea componenței, sau invers. Așadar, componentă respective poate fi un predictor bun pentru variabilă analizată.

Componentă Principală 1 (C1): Această componentă prezintă corelații puternice pozitive cu " Chronic Kidney Disease", "Cirrhosis and Other Chronic Liver Diseases" și "Road Injuries", în timp ce are o corelație negativă semnificativă cu "Parkinson's Disease" Se pare că C1 ar putea fi un indicator al unui factor legat de afecțiuni cronice și de impactul lor general asupra stării de sănătate.

Componentă Principală 2 (C2): Se observă corelații puternice pozitive cu "Tuberculosis" și "Acute Hepatitis" și o corelație negativă puternică cu " Malaria". Această componentă sugerează că C2 ar putea fi asociat cu boli infecțioase sau condiții influențate de factori de mediu sau socio-economici.

Componentă Principală 3 (C3): Aici se constată corelații semnificative pozitive cu "Malaria" și corelații negative cu "HIV/AIDS" și " Poisonings". C3 ar putea fi interpretat că un factor reflectând condiții asociate cu expunerea la agenți externi specifici sau cu transmiterea prin vectori specifici (cum ar fi țânțarii în cazul malariei).



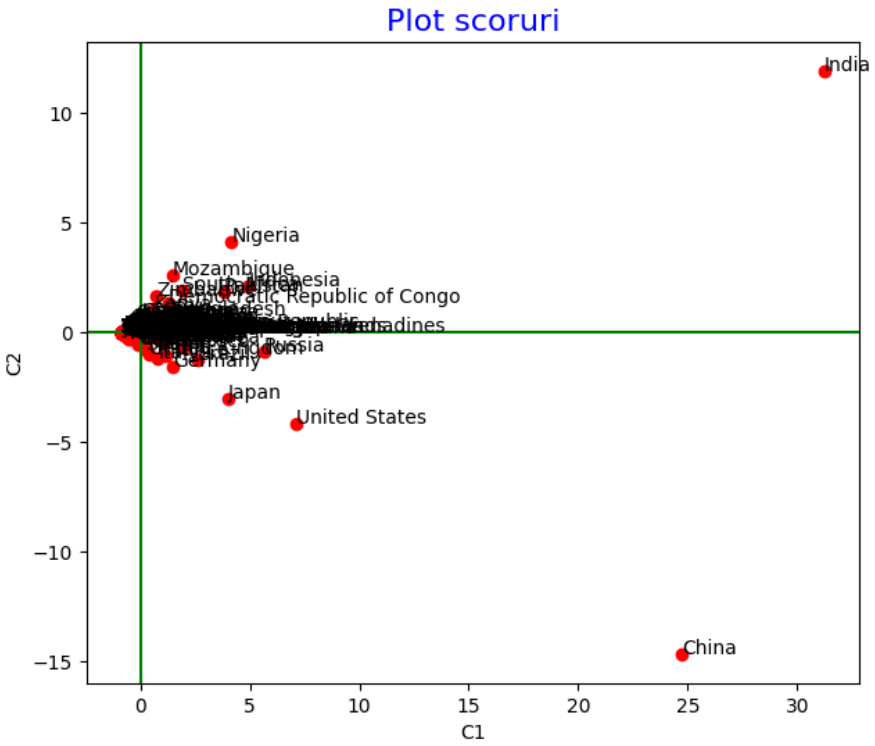
Grafic 2 – Plot corelații dintre variabilele observate și componente – cercul corelațiilor

Cercul corelațiilor reprezintă grafic relațiile dintre variabilele observate și componentele principale C1 și C2.

Din figura de mai sus (Grafic 2) se remarcă faptul că variabile precum accidente rutiere (Road injuries), Chronic Kidney Disease și Parkinson's au o corelația puternică cu C1. Astfel, componentă 1 include o caracteristică ce este semnificativ asociată condițiilor de sănătate enumerate mai sus.

Variabilele Tuberculosis și Acute Hepatitis sunt apropiate atât una de cealaltă, cât și de axa C2, dar mai departe de C1, ceea ce indică o corelație puternică cu cel de-al doilea factor (C2) și o corelație între ele. Acest lucru sugerează că există un factor de risc sau o caracteristică comună ce afectează ambele condiții.

1.3 Scoruri

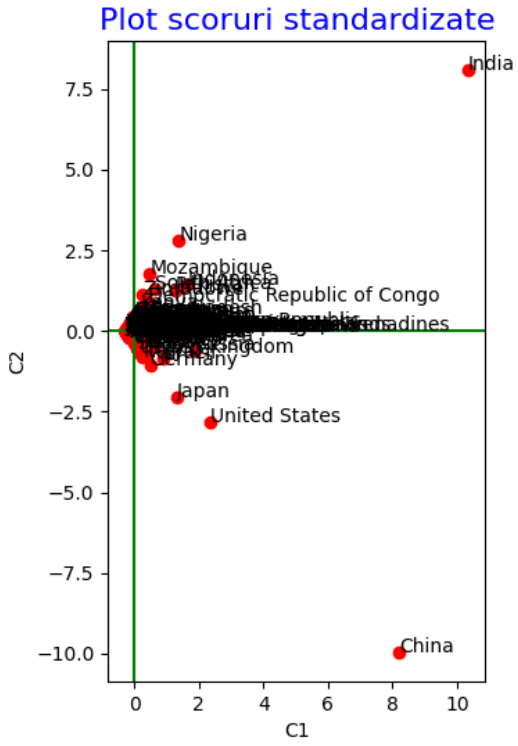


Grafic 2 – Plot scoruri

În analiza în componente principale (ACP), scorurile componentelor principale sunt proiecțiile datelor originale în spațiul redus al CP. Plot-ul de scoruri este folosit pentru a vizualiza relațiile dintre țări pe baza primelor două componente principale.

Deoarece India are un scor mare pe C1 (în jur de 30) și unul mediu pe C2 (în jur de 14) , este sugerat faptul că atributele care definesc C1 sunt proeminente în India. Acesta analiză este valabilă și pentru China.

Se mai remarcă și faptul că majoritatea punctelor sunt concentrate în jurul originii, adică majoritatea țărilor au scoruri mici pe ambele componente ceea ce sugerează că țările sunt similar din punct de vedere al caracteristicilor măsurate de C1 și C2.



Grafic 2 – Plot scoruri standardizate

Aceeași analiză se remarcă și in cazul plot-ului scorurilor standardizate, modificându-se doar valorile scorurilor.

Tabelele aferente scorurilor și scorurilor standardizate se regăsesc in directoarele:

acp/tabele/csv/scoruri.csv

acp/tabele/csv/scoruri_std.csv

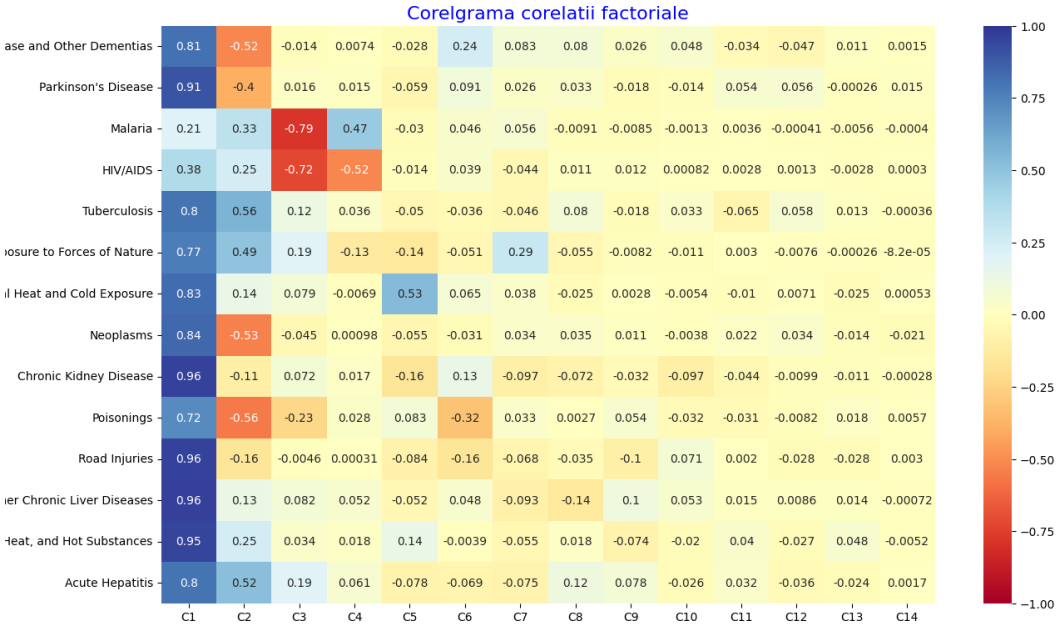
1.4 Valori cosinus

Fiecare rând reprezintă o țară, iar fiecare coloană reprezintă cosinusul unghiului în raport cu o componentă principală. Țările care au valoarea cosinusului dintre vectori cât mai apropiată de 1 prezintă caracteristici similar în raport cu componentele, iar cele care au valoarea cât mai apropiată de -1 prezintă caracteristici opuse / diferite. Se remarcă valori mari în raport cu componentă 1 pentru țări precum Barbados (0.91), Belize(0.91), Singapore (0.9).

Tabelul aferent valorilor cosinus se regasește in directorul:

acp/tabele/csv/cosinusuri.csv

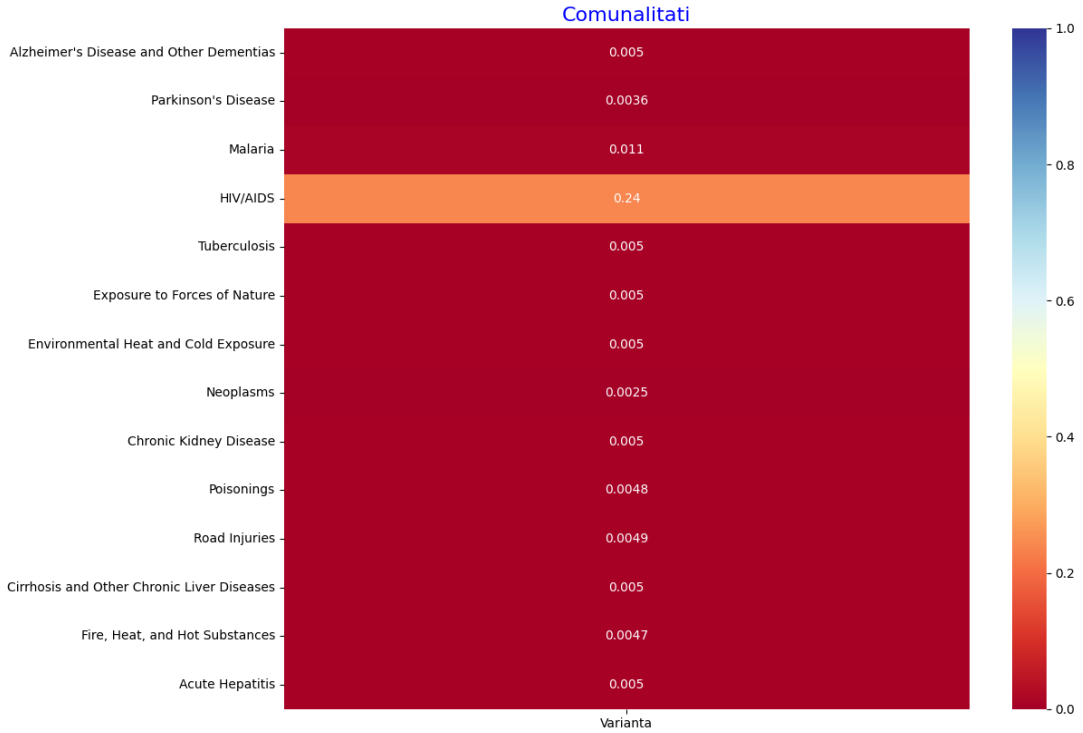
1.5 Corelogramă corelații factoriale



Deoarece Alzheimer's Disease and Other Dementias are o corelație puternică pozitivă cu C1 (0.81), prima componentă surprinde varianța asociată cu bolile de acest tip. O corelație chiar mai puternică, pozitivă cu C1 se remarcă pentru Parkinson's (0.91).

Deoarece C3 prezintă o corelație negativă cu Malaria, C3 reprezintă factori care nu sunt asociați cu riscul aparitiei malariei.

1.6 Comunalități



Comunalitatea, în contextual analizei factoriale, prezintă cât din varianța unei variabile este explicată de către toți factorii. În acest caz, valorile sunt destul de mici, ceea ce înseamnă că factorii nu explică o proporție foarte mare din varianță.

1.7 Contribuții

Tabelul aferent contribuțiilor se regăsește in directorul:
acp/tabele/csv/contributii.csv

2. Analiza factorială

2.1 Testul Bartlett

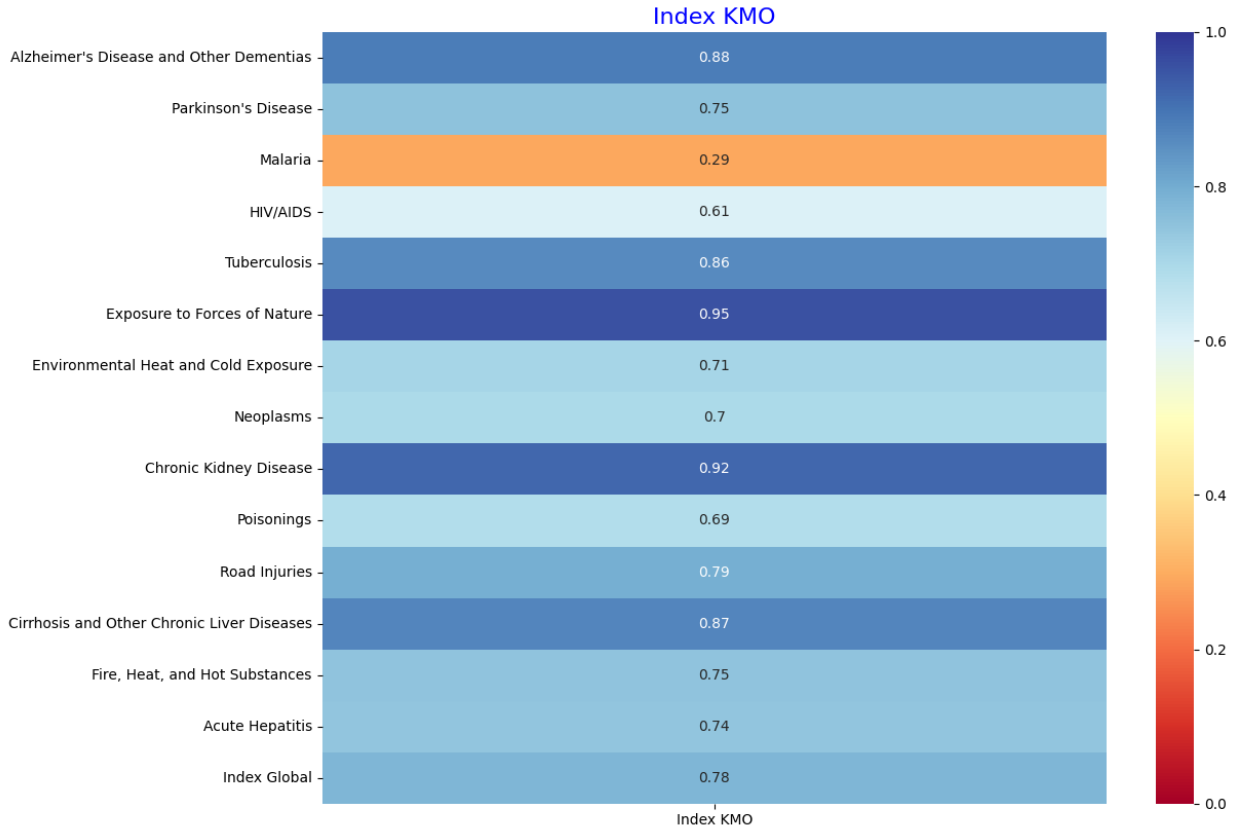
```
Test Bartlett: (6452.908396908941, 0.0)
```

Testul Bartlett este utilizat pentru a evalua ipoteza nulă că variațele a două sau mai multe seturi de date sunt egale. În outputul avem două valori:

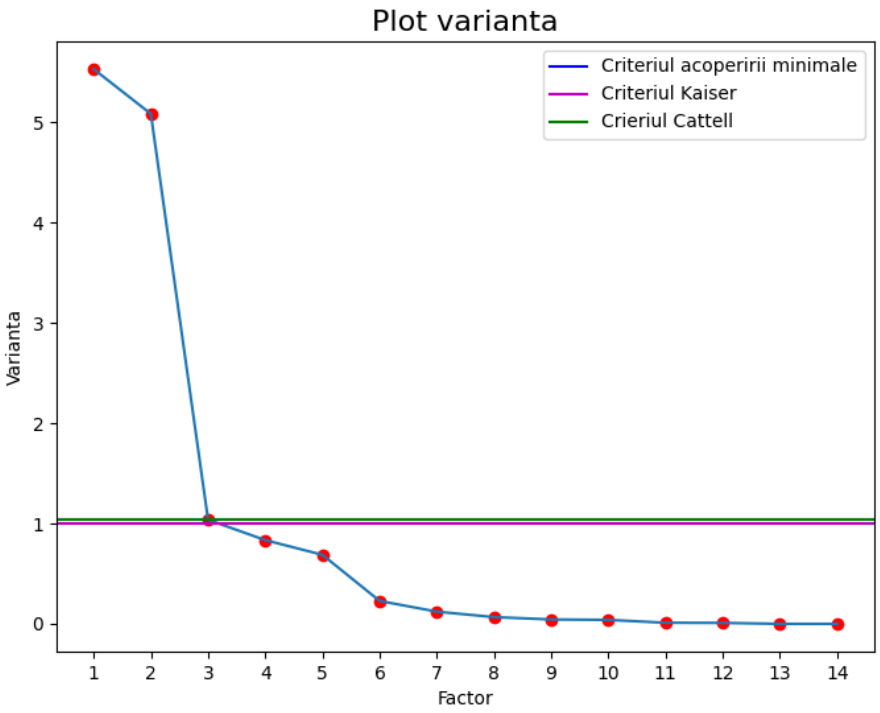
6452.90: Acesta este statistica de test Bartlett. Valoarea mare a acestui statistic sugerează că există diferențe semnificative între variațele grupurilor comparate.

0.0: Acesta este p-value asociat cu statistica testului. Este o valoare zero, conducând la respingerea ipotezei nule. În acest caz, putem spune că există dovezi statistice suficiente pentru a respinge ipoteza nulă că variațele sunt egale, sugerând că există diferențe semnificative între variațele grupurilor comparate.

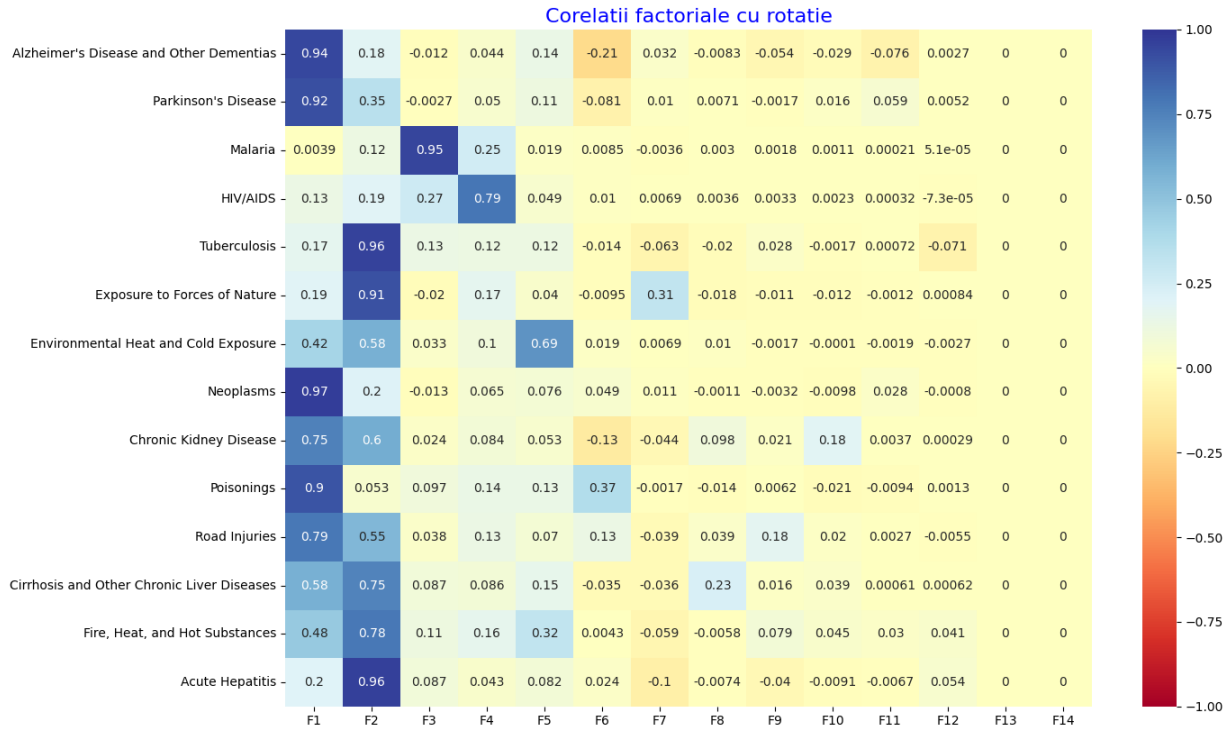
2.2 Corelograma Indecși KMO



2.3 Varianță factori



2.4 Corelogramă corelații factoriale



Rotația este adesea aplicată în analiza factorială pentru a facilita interpretarea factorilor prin maximizarea varianței încărcăturilor pe fiecare factor, ceea ce ajută la identificarea unui model mai clar de relații între variabile și factori.

Interpretarea bazată pe valorile și culorile din matrice este următoarea:

- Factorii (F1 - F14): Fiecare coloană reprezintă un factor individual în model, iar numerele sugerează că rotația a fost efectuată pe 14 factori.
- Variabilele (Condițiile de Sănătate): Fiecare rând reprezintă o variabilă diferită, care, în acest caz, sunt diferite condiții de sănătate sau factori de risc pentru sănătate.

Valorile Corelației:

- Valorile apropiate de 1 sau -1 (colorate în nuanțe puternice de albastru sau roșu) indică o corelație puternică între condiția de sănătate și factorul respectiv.
- Valorile aproape de 0 (colorate în galben) indică o corelație slabă sau nesemnificativă.

Interpretarea Corelațiilor Puternice:

Pentru F1, observăm corelații puternice cu "Neoplasms", "Chronic Kidney Disease", "Parkinson's Disease" și "Alzheimer's Disease and Other Dementias". Aceasta sugerează că F1 ar putea reprezenta un factor general de sănătate sau un factor legat de condiții cronice sau degenerative.

F2 are corelații puternice cu "Tuberculosis" și "Exposure to Forces of Nature", ceea ce ar putea indica un factor asociat cu boli infecțioase și impactul mediului.

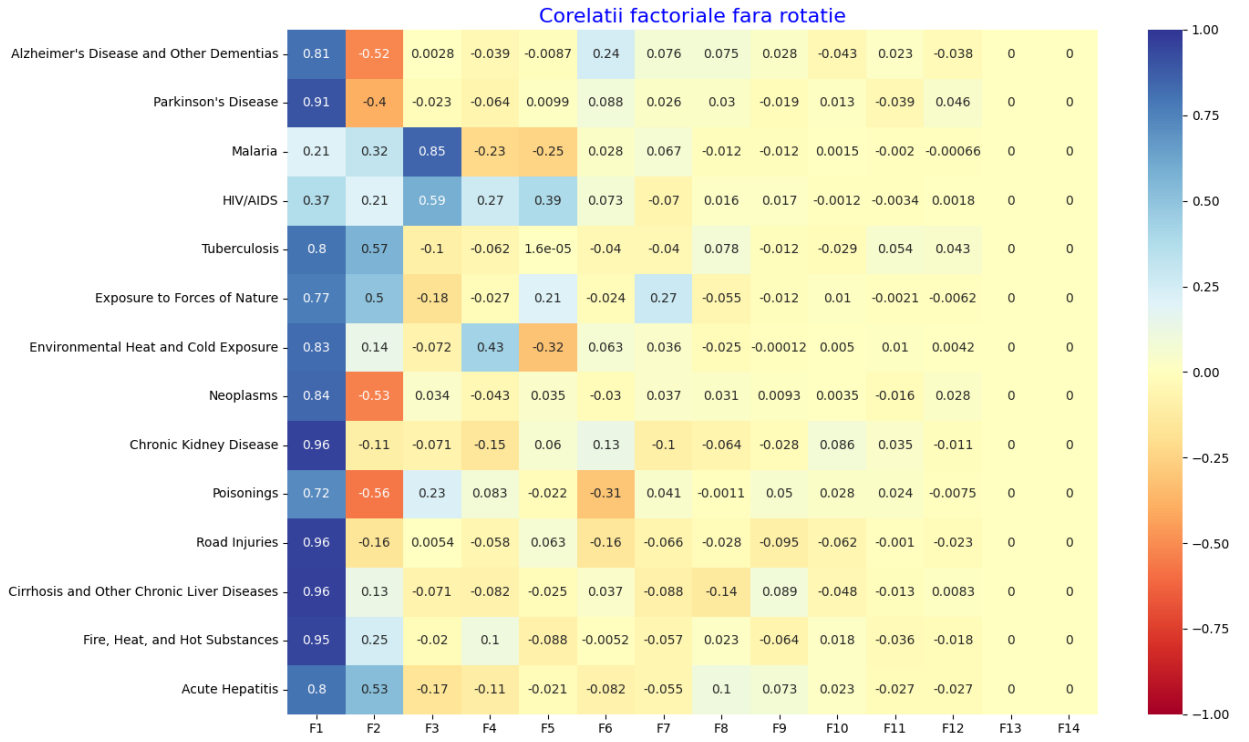
F3 are o corelație puternică cu "Malaria", sugerând că acest factor ar putea fi legat de boli specifice răspândite de vectori sau de condiții care prevalează în anumite medii geografice sau climatice.

Culoarea:

- Roșul reprezintă corelații pozitive puternice.
- Albastrul reprezintă corelații negative puternice.
- Culorile mai deschise indică corelații mai slabe.

Această matrice ne ajută să înțelegem relația fiecărei condiții de sănătate cu factorii extrași în model. De exemplu, putem vedea că:

- "Alzheimer's Disease and Other Dementias" are o corelație puternică cu primul factor (F1), ceea ce sugerează că acest factor ar putea reprezenta aspecte legate de condiții neurodegenerative sau legate de vârstă.
- "Tuberculosis" are o corelație foarte puternică cu al doilea factor (F2), indicând că F2 ar putea fi asociat cu factori de mediu sau sociali care contribuie la răspândirea bolilor infecțioase.
- "Malaria" are o corelație extrem de puternică cu al treilea factor (F3), ceea ce poate sugera că acest factor reprezintă condiții specifice răspândite de vectori, cum ar fi țânțarii, sau factori specifici regiunilor unde malaria este endemică.

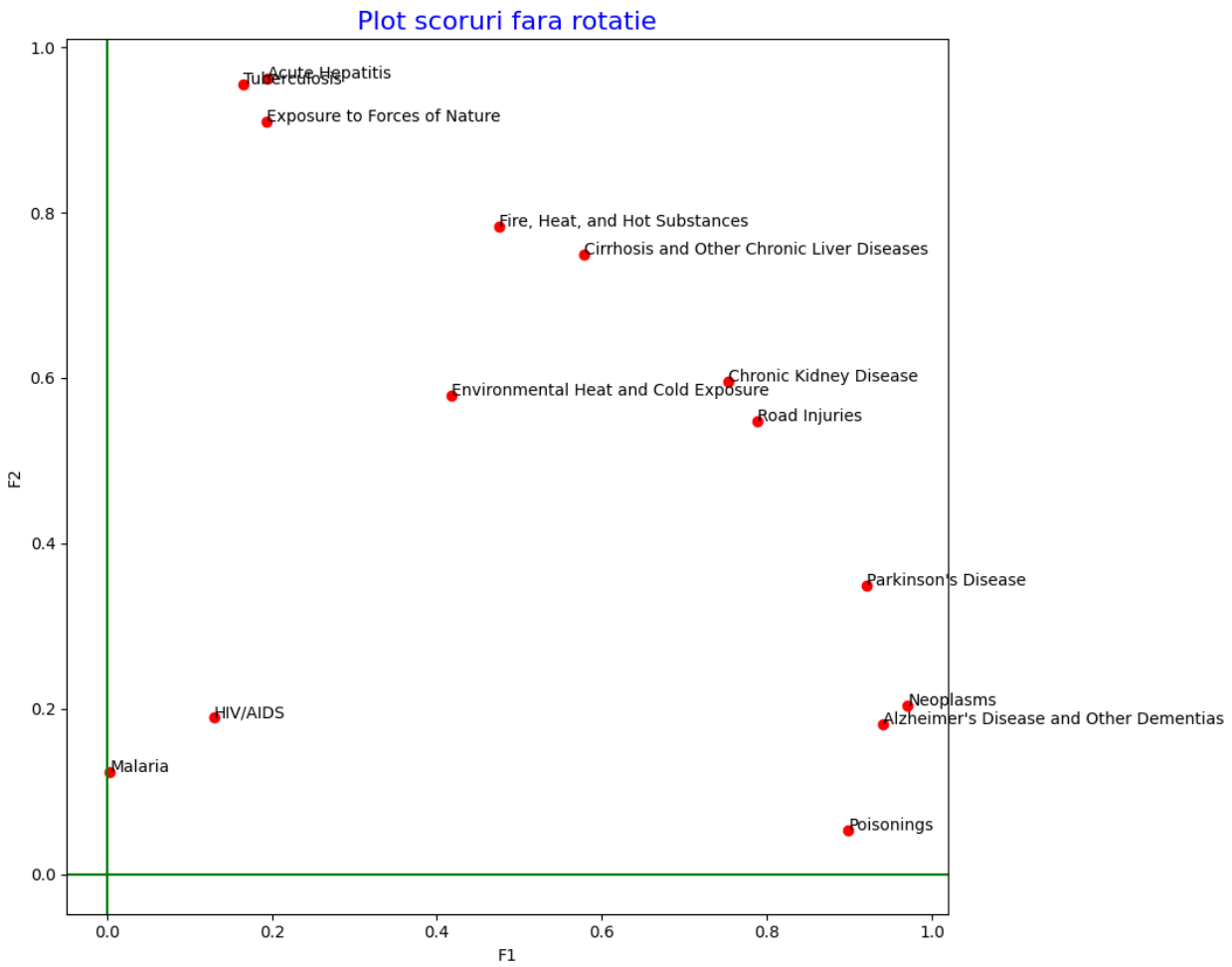


Chronic Kidney Disease are o corelație puternică cu F1, sugerând că F1 ar putea fi asociat cu factori specifici care influențează sau sunt asociați cu bolile renale cronice.

Tuberculosis are o corelație puternică cu F2, indicând că F2 ar putea reprezenta aspecte legate de bolile infecțioase sau condițiile asociate cu tuberculoza.

Malaria are o corelație puternică cu F3, ceea ce ar putea sugera că F3 este legat de factori de risc sau condiții specifice pentru malaria.

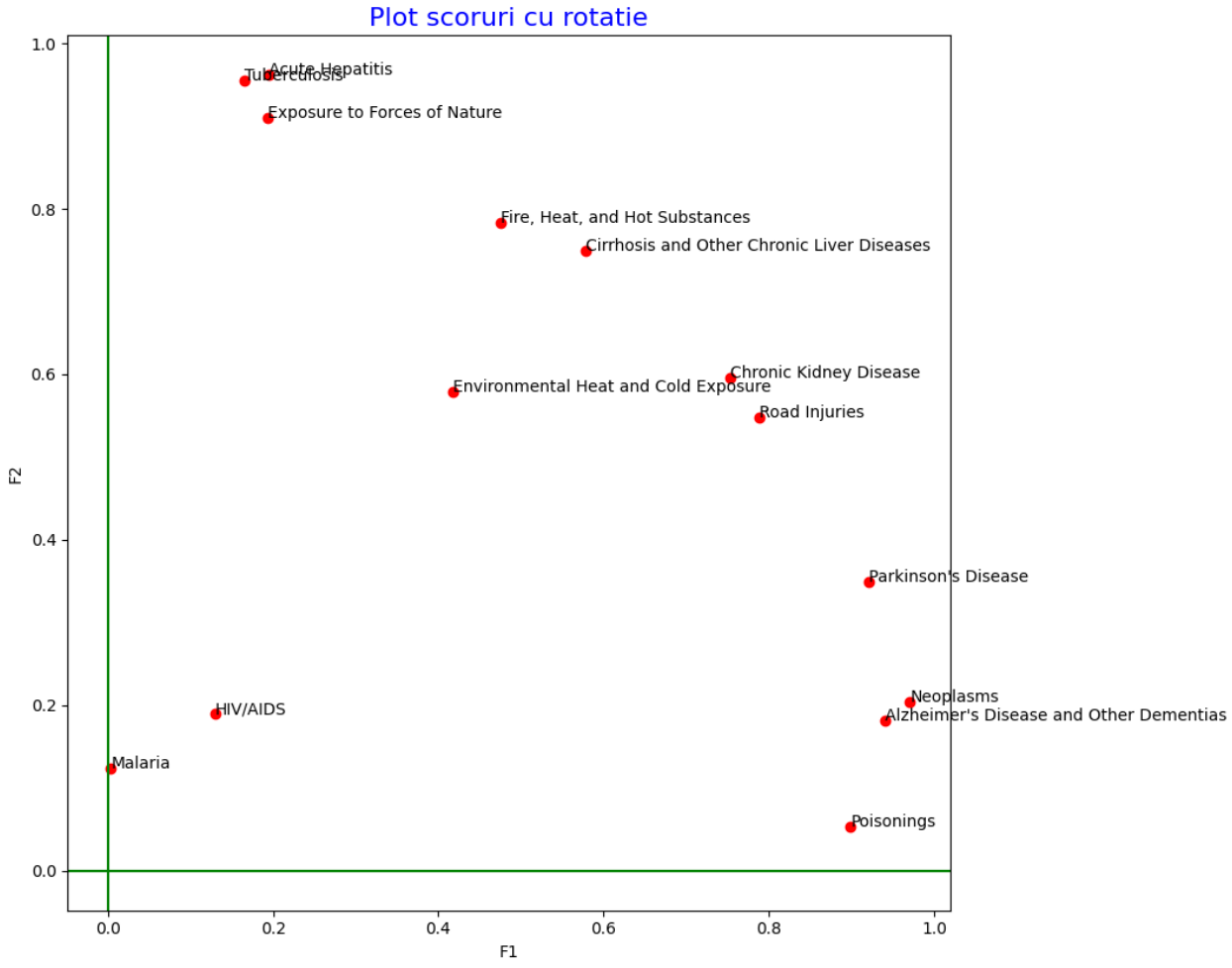
2.5 Plot scoruri



1. **Axele (F1 și F2)**: Axa orizontală (F1) și axa verticală (F2) reprezintă cele două componente principale sau factori. Poziția fiecărei condiții de sănătate pe grafic reflectă scorurile lor pe acești doi factori.
2. **Distribuția pe F1**:
    - Condițiile de sănătate plasate spre extrema dreaptă, cum ar fi "Neoplasms", "Alzheimer's Disease and Other Dementias", și "Poisonings", au scoruri înalte pe F1. Aceasta sugerează că F1 ar putea reprezenta un factor care este semnificativ asociat cu aceste condiții.
    - Condițiile de sănătate plasate mai aproape de partea stângă, cum ar fi "Malaria" și "HIV/AIDS", au scoruri mai scăzute pe F1.
3. **Distribuția pe F2**:

- Condițiile de sănătate plasate spre extremitatea superioară, cum ar fi "Tuberculosis" și "Acute Hepatitis", au scoruri înalte pe F2, ceea ce indică o asociere puternică cu acest factor.
- Pe de altă parte, cele mai multe condiții de sănătate sunt grupate aproape de zero pe F2, indicând scoruri mai scăzute sau o asociere mai slabă.

4. **Interpretarea Factorilor**:
   - Fără informații suplimentare despre setul de date și variabilele specifice incluse în analiză, este dificil să atribuim o interpretare definitivă factorilor F1 și F2. Totuși, putem specula că F1 ar putea reprezenta factori generali de sănătate sau un gradient de severitate a bolii, în timp ce F2 ar putea reprezenta un alt aspect al sănătății, cum ar fi tipul de transmitere a bolii sau răspunsul la tratament.



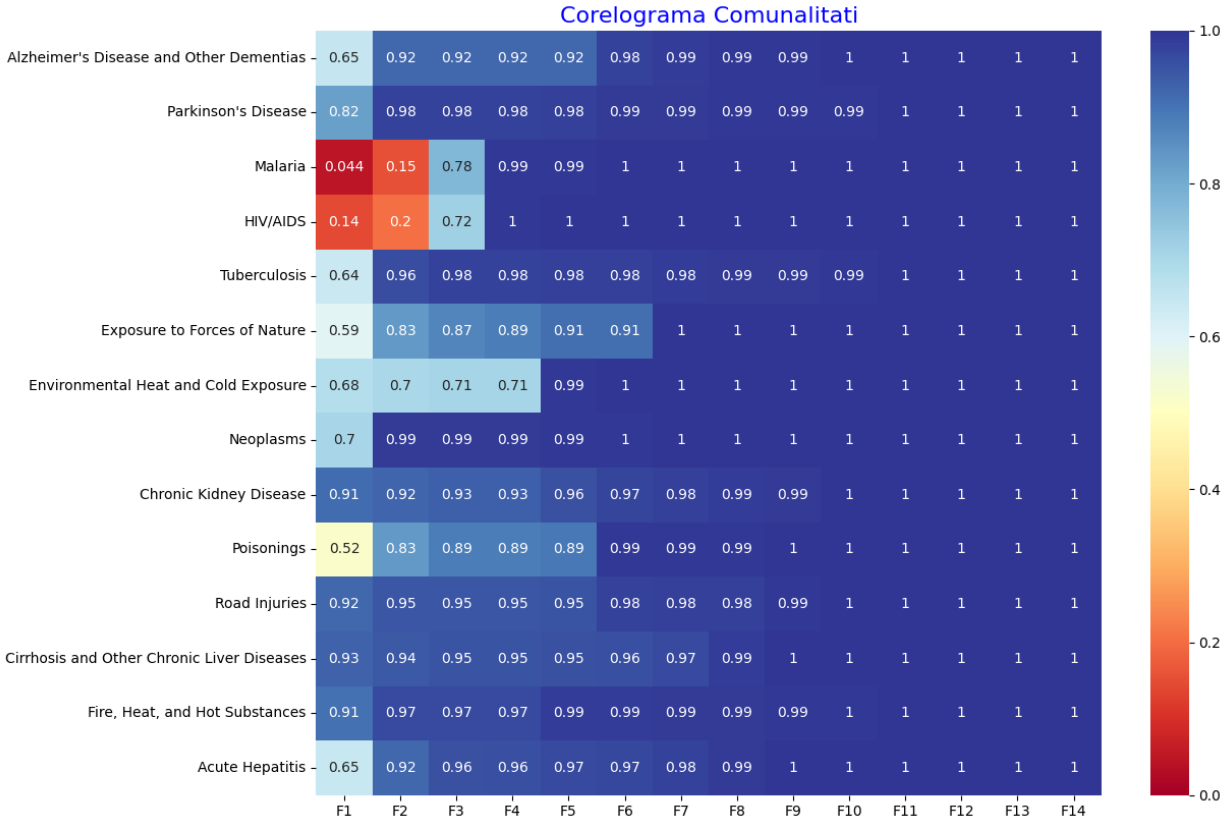
Condițiile de sănătate plasate spre extremitatea dreaptă a axei F1 (de exemplu, "Neoplasms", "Alzheimer's Disease and Other Dementias") au scoruri mari pe acest factor, indicând o asociere puternică cu caracteristicile pe care le capturează F1.

Condițiile de sănătate plasate mai sus pe axa F2 (cum ar fi "Tuberculosis" și "Acute Hepatitis") au scoruri mari pe acest factor, ceea ce ar putea indica o asociere puternică cu caracteristicile pe care le capturează F2.

Fără informații suplimentare despre natura rotației aplicate sau despre variabilele originale, nu putem oferi o interpretare precisă a factorilor. Totuși, pozițiile relative ale condițiilor de sănătate pot sugera că F1 ar putea reprezenta un factor legat de boli cronice sau degenerative, în timp ce F2 ar putea reprezenta condiții legate de factori de mediu sau infecții.

Gruparea condițiilor de sănătate aproape de origine (cazul "Malaria" și "HIV/AIDS") sugerează o asociere mai slabă sau diferită cu factorii reprezentați de F1 și F2. Distanța dintre condițiile de sănătate poate indica diferențe în profilurile lor în raport cu factorii reprezentați pe axele F1 și F2.

2.6 Corelogramă Comunalități



1. **Valorile Comunalității**:

- Valorile apropiate de 1 indică o comunalitate mare, ceea ce înseamnă că un procent mare din varianța variabilei este capturat de factorii extrasi.
- Valorile mai mici decât 1 indică o comunalitate mai mică, sugerând că factorii extrasi nu explică toată varianța variabilei respective.

2. **Interpretarea Specifică**:
    - Condițiile de sănătate precum "Chronic Kidney Disease", "Cirrhosis and Other Chronic Liver Diseases", și "Road Injuries" au comunalități foarte mari pe majoritatea factorilor, indicând că varianța lor este bine explicată de factorii extrasi.
    - Condiții precum "Malaria" și "HIV/AIDS" au comunalități inițiale mai mici, dar care cresc la factorii ulteriori, sugerând că acești factori mai târzii ar putea fi mai relevanți pentru explicarea variației acestor condiții.