

TEMA2

I. ANALIZA FACTORILOR

*Cerințele 1-7 au fost preluate de la Tema 1.

Cerința 8: Evaluați factorabilitatea setului de date. Calculați indicele KMO și testul Bartlett pentru a decide dacă se poate parcurge analiza factorială.

```
percentage pages was better under R version 3.1.2
> KMO(date_std)
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = date_std)
Overall MSA = 0.67
MSA for each item =
Agricultura_Vanatoare      Pescuit      Energie_Gaze_Apa      Constructii      Comert      Hoteluri_Restaurante
0.70      0.28      0.64      0.68      0.69      0.56
Transport_Depozitare      Posta      Educatie      Admin_publica_Aparare      Sanatate
0.73      0.69      0.66      0.81      0.70
```

Indicele KMO este utilizat pentru a evalua adecvarea setului de date pentru analiza factorială, fiind calculat pe baza proporției variației variabilelor care poate fi explicată de factori latenți. Rezultatele din analiză arată:

- **Overall MSA = 0.67:** Aceasta indică un nivel moderat de adecvare a datelor pentru analiza factorială. Setul de date poate fi utilizat în continuare, dar există variabile care afectează performanța generală.
- **Valori individuale:**
 - Admin_publica_Aparare = 0.81: Aceasta este cea mai ridicată valoare din setul de date și indică faptul că această variabilă este bine corelată cu celelalte variabile.
 - Pescuit = 0.28: Aceasta este cea mai scăzută valoare și sugerează că variabila Pescuit nu este bine corelată cu restul setului de date. Acest lucru poate reduce calitatea generală a analizei factoriale

```
> R=cor(date_std)
> cor.test(R, n=42, diag=TRUE)
$chisq
[1] 127.7046

$p.value
[1] 1.02923e-07

$df
[1] 55
```

Testul Bartlett verifică dacă matricea de corelație este o matrice de identitate. O matrice de identitate sugerează că variabilele sunt ortogonale (necorelate), ceea ce ar face analiza factorială imposibilă.

Ipoteze:

H₀: Matricea de corelație este o matrice de identitate (nu există factori comuni, variabilele sunt necorelate).

H₁: Matricea de corelație diferă semnificativ de o matrice de identitate (există factori comuni, variabilele sunt corelate).

Un p-value < 0.05 indică faptul că matricea de corelație este semnificativ diferită de o matrice de identitate.

Rezultatele obținute:

- Chi-squared = 127.7046: Această valoare statistică arată o diferență semnificativă între matricea de corelație a variabilelor și o matrice de identitate. Cu cât valoarea chi-pătrat este mai mare, cu atât diferența este mai semnificativă.
- p-value = 1.02923e-07: Valoarea p este semnificativ mai mică decât 0.05, ceea ce ne permite să respingem H_0 .
- df = 55: Gradele de libertate sunt calculate în funcție de numărul de variabile din matricea de corelație.

Cerința 9: Aplicați criteriile/metodele relevante pentru alegerea numărului de factori din analiză și generați o concluzie cu privire la numărul optim de factori.

A. Criteriul grafic (Scree Plot)

```
> screeplot_result <- prcomp(date_std)
> eigenvalues <- screeplot_result$sdev^2
> plot(eigenvalues, type = "b", main = "Scree Plot", xlab = "Componenta", ylab = "Valori proprii", col = "magenta")
> abline(h = 1, col = "red", lty = 2)
```

În graficul **Scree Plot**, observăm o **scădere abruptă** a valorilor proprii între prima și a doua componentă.

După Componenta 2, panta graficului se **aplatizează**, ceea ce sugerează că informația adăugată de componentele următoare este neglijabilă.

Concluzie: 2 factori ar trebui reținuți, deoarece primele două componente explică majoritatea variației din date.

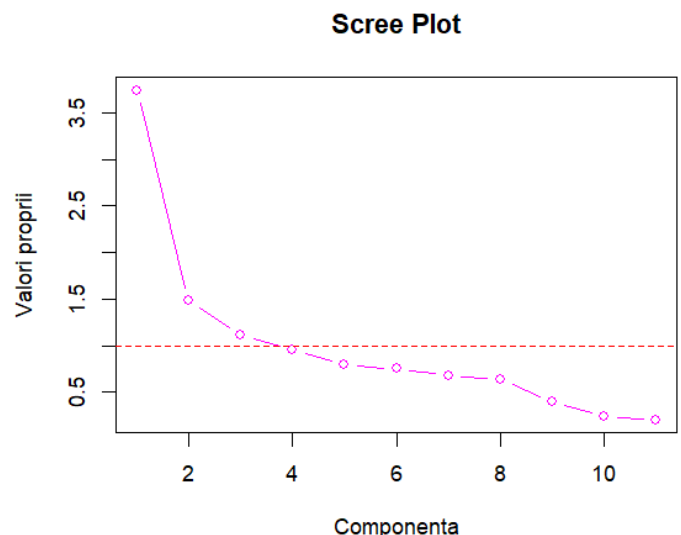


Figure 1

B. Criteriul lui Kaiser

```
> componente_kaiser <- sum(eigenvalues >= 1) # valorile proprii >= 1
> print(paste("Numărul de factori conform criteriului lui Kaiser:", componente_kaiser))
[1] "Numărul de factori conform criteriului lui Kaiser: 3"
```

Componentele cu valori proprii ≥ 1 reprezintă variația mai mare decât variabila originală standardizată (cu varianță = 1). Dacă analiza Scree Plot a sugerat 2 factori (prin punctul de cotitură), criteriul Kaiser indică faptul că și al treilea factor poate fi relevant.

C. Criteriul procentului de varianță cumulată

```
> var_cumulativa <- cumsum(eigenvalues) / sum(eigenvalues) * 100 # Varianță cumulată în procente
> print(var_cumulativa)
[1] 33.99845 47.47140 57.54727 66.28768 73.51386 80.34310 86.52678 92.28766 95.94238 98.13670 100.00000
> componente_var_cumulativa <- which(var_cumulativa >= 80)[1]
> print(paste("Numărul de factori conform criteriului procentului de varianță cumulată:", componente_var_cumulativa))
[1] "Numărul de factori conform criteriului procentului de varianță cumulată: 6"
```

Se urmărește ca factorii reținuți să explice **cel puțin 70-80%** din variația totală pentru a asigura o reprezentare adecvată a datelor. Primii **6 factori** explică împreună **80.34%** din variația totală a datelor. Acești factori sunt suficienți pentru a reprezenta majoritatea informației din setul de date.

În urma aplicării celor trei criterii pentru selecția numărului optim de factori, am decis să reținem **3 factori**. Deși **criteriul grafic (Scree Plot)** indică doar 2 factori ca fiind semnificativi, iar **criteriul procentului de variație cumulată** recomandă 6 factori pentru a explica peste 80% din variația totală a datelor, alegerea a 3 factori reprezintă un echilibru optim între reducerea complexității analizei și păstrarea informației esențiale. Astfel, cei 3 factori reținuți vor permite o interpretare clară și relevantă a structurii latente din setul de date.

Cerința 10: Aplicați două metode pentru estimarea modelului factorial și analizați care este mai potrivit. Interpretați output-urile.

A. METODA AXELOR PRINCIPALE

```
> library(psych)
> factor1=fa(date_std, nfactors = 3, rotate="none", fm="pa")
> print(factor1$loadings, cutoff=0.4)
```

```
Loadings:
               PA1    PA2    PA3
Agricultura_Vanatoare 0.418
Pescuit                0.437
Energie_Gaze_Apa      0.568
Constructii            0.633
Comert                 0.661
Hoteluri_Restaurante
Transport_Depozitare  0.595
Posta
Educatie               0.781
Admin_publica_Aparare  0.536 -0.465
Sanatate               0.579

               PA1    PA2    PA3
SS loadings    3.222 0.896 0.464
Proportion Var 0.293 0.081 0.042
Cumulative Var 0.293 0.374 0.417
```

Metoda utilizată pentru estimarea modelului factorial este **Principal Axis Factoring (PA)**, care are scopul de a identifica structurile latente (factori comuni) în date, explicând corelațiile dintre variabile.

Factorul 1 (PA1) explică **29.3% din variația totală** și este cel mai semnificativ. Variabile precum **Educatie (0.781)**, **Comert (0.661)**, **Constructii (0.633)** și **Sanatate (0.579)** au încărcături ridicate, sugerând că acest factor reflectă un aspect socio-economic general.

Factorul 2 (PA2) explică doar **8.1% din variație**. Variabilele **Pescuit (0.437)** și **Admin_publica_Aparare (0.536)** sunt cele mai relevante, indicând o dimensiune legată de activități administrative și sectoriale.

Factorul 3 (PA3) are o contribuție modestă (**4.2% din variație**). Este mai puțin semnificativ, dar **Admin_publica_Aparare (-0.465)** are o relație negativă cu acest factor.

B. METODA VEROSIMILITATII MAXIME

```
> factor2=fa(date_std, nfactors=3, rotate="none", fm="ml")
>
> print(factor2$loadings, cutoff=0.4)
```

Loadings:

	ML1	ML3	ML2
Agricultura_Vanatoare			
Pescuit			
Energie_Gaze_Apa	0.469		
Constructii	0.586		
Comert	0.849		0.523
Hoteluri_Restaurante		0.435	
Transport_Depozitare		0.719	
Posta			
Educatie	0.851		-0.520
Admin_publica_Aparare	0.548		
Sanatate	0.585		

	ML1	ML3	ML2
SS loadings	3.060	1.153	0.777
Proportion Var	0.278	0.105	0.071
Cumulative Var	0.278	0.383	0.454

Metoda **Maximum Likelihood (ML)** este utilizată pentru estimarea modelului factorial pe baza verosimilității maxime. Aceasta optimizează parametrii modelului pentru a maximiza probabilitatea de a obține matricea de corelație observată, ținând cont de erorile de măsurare.

Factorul 1 (ML1) explică 27.8% din variația totală, fiind cel mai semnificativ. Cele mai relevante variabile sunt Educație (0.851) și Comerț (0.849), sugerând o dimensiune socio-economică generală, susținută și de variabile precum Construcții (0.586) și Sănătate (0.585). Acest factor poate fi denumit "factor socio-economic general".

Factorul 3 (ML3) contribuie cu 10.5% la variație și este dominat de variabilele Transport_Depozitare (0.719), Energie_Gaze_Apa (0.469) și Hoteluri_Restaurante (0.435), indicând o dimensiune legată de infrastructură și servicii logistice.

Factorul 2 (ML2), care explică doar 7.1% din variație, este mai slab semnificativ. Variabilele principale sunt Admin_publica_Aparare (0.548) și Comerț (0.523), iar relația negativă cu Educație (-0.520) poate semnala o dimensiune administrativă sau organizațională specifică.

Cerința 11: Realizați diagramele corespunzătoare celor două modele și interpretați rezultatele.

`> fa.diagram(factor1)`

Diagrama prezintă relațiile dintre variabilele observabile și factorii latenți identificați prin metoda PA. Fiecare săgeată indică corelația (încărcătura factorială) dintre o variabilă și un factor, iar valorile încărcăturilor sunt exprimate pe săgeți.

Factorul 1 (PA1): are cele mai mari încărcături factoriale, fiind asociat cu variabile precum: Educatie (0.8), Comert (0.7), Constructii (0.6), Transport_Depozitare (0.6) și Sanatate (0.6).

Această dimensiune poate fi interpretată ca un factor socio-economic general, reflectând importanța infrastructurii, educației și sănătății pentru structura datelor.

Factorul 2(PA2): Este asociat cu variabile precum:Energie_Gaze_Apa(0.5)și Admin_publica_Aparare (0.4).

Factorul 2 poate fi interpretat ca o dimensiune legată de utilități publice și administrație.

Factorul 3 (PA3): nu este conectat cu nicio variabilă neavând încărcături factoriale peste pragul de 0.4.

Acest lucru indică faptul că PA3 nu identifică o dimensiune clară și nu contribuie semnificativ la structura datelor.

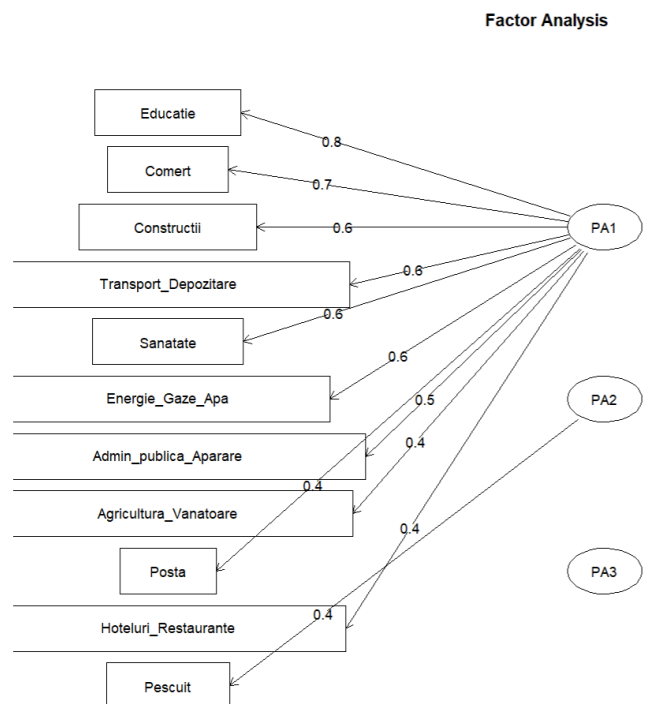


Figure 2

`> fa.diagram(factor2)`

Factor Analysis

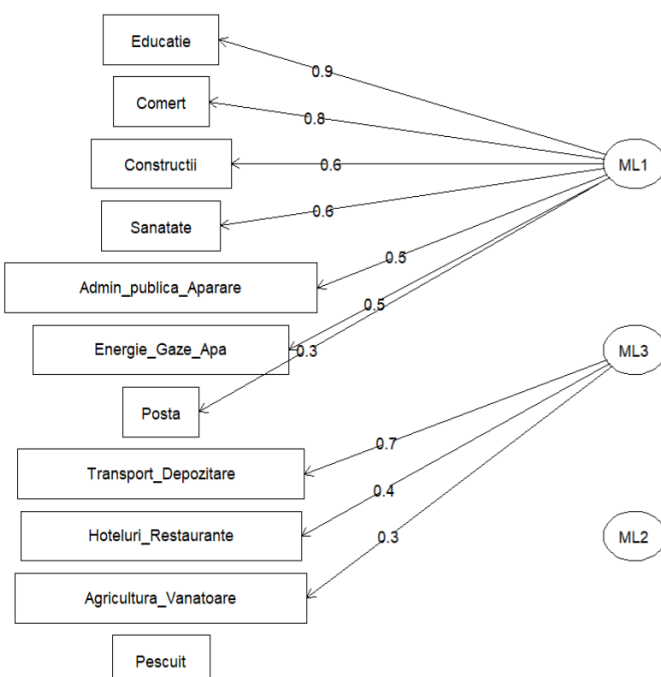


Figure 3

Factorul 1 este cel mai semnificativ și explică relațiile principale din setul de date. Variabilele cu cele mai mari încărcături factoriale sunt:**Educatie (0.9)**, **Comert (0.8)**, **Constructii (0.6)** și **Sanatate (0.6)**, indicând o dimensiune **socio-economică generală**.

- Acest factor reflectă importanța educației, comerțului, infrastructurii și sănătății în structurarea datelor.

Factorul 3 (ML3): Este asociat cu variabilele Transport_Depozitare(0.7) și Hoteluri_Restaurante (0.4), sugerând o dimensiune orientată spre infrastructură și servicii logistice.

- Alte variabile, precum **Energie_Gaze_Apa (0.3)** și **Posta (0.3)**, contribuie marginal la acest factor.

Factorul 2 (ML2): nu este asociat în mod clar cu nicio variabilă (niciuna nu are încărcături peste pragul de 0.4). Acest lucru sugerează că ML2 nu captează o dimensiune clară sau semnificativă în setul de date.

Cerința 12: Denumiți factorii, interpretați rezultatele și salvați noul set de date.

```
> scoruri_PA <- factor1$scores
> colnames(scoruri_PA) <- c("Dimensiunea_socio_economica_generala",
+                             "Utilitati_publice_administratie",
+                             "Redundanta_PA")
> date_noi_PA <- cbind(date_noi, scoruri_PA)
> write.csv(date_noi_PA, "Date_Factoriale_PA.csv", row.names = FALSE)
> scoruri_ML <- factor2$scores
> colnames(scoruri_ML) <- c("Dimensiunea_socio_economica_generala",
+                             "Infrastructura_si_logistica",
+                             "Redundanta_ML")
> date_noi_ML <- cbind(date_noi, scoruri_ML)
> write.csv(date_noi_ML, "Date_Factoriale_ML.csv", row.names = FALSE)
```

Pentru metoda PA (Principal Axis Factoring):

1. **Factor 1: Dimensiunea socio-economică generală:** Reprezintă relațiile dintre educație, comerț, infrastructură și sănătate (**Educatie, Comert, Constructii, Transport_Depozitare, Sanatate**).
2. **Factor 2: Utilități publice și administrație:** Legat de funcționarea administrației publice și furnizarea utilităților (**Energie_Gaze_Apa, Admin_publica_Aparare**).
3. **Factor 3: Redundanță:** Nu este asociat cu variabile relevante și nu adaugă valoare interpretativă.

Pentru metoda ML (Maximum Likelihood):

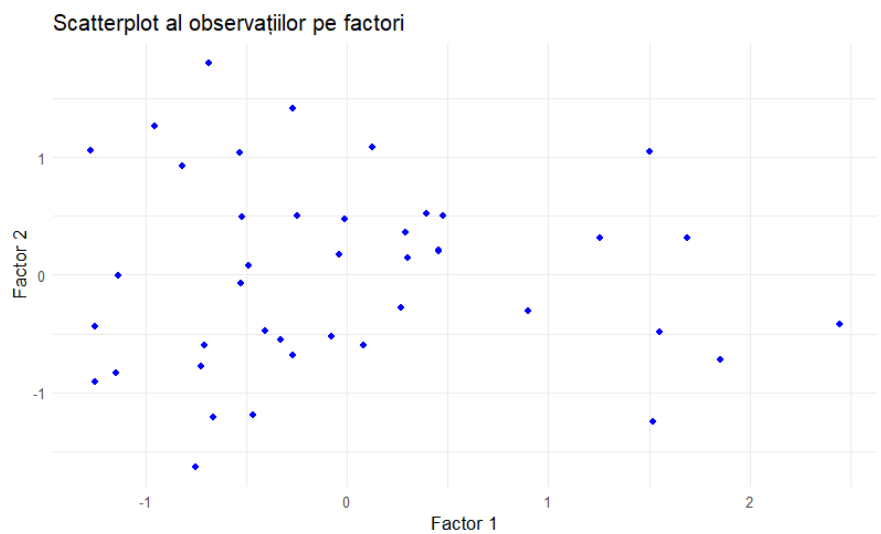
1. **Factor 1: Dimensiunea socio-economică generală:** Include variabile care reflectă legături între educație, comerț și sănătate (**Educatie, Comert, Constructii, Sanatate**).
2. **Factor 3: Infrastructură și logistică:** Asociat cu transport și servicii logistice (**Transport_Depozitare, Hoteluri_Restaurante**).
3. **Factor 2: Redundanță:** Nu are variabile cu încărcături semnificative.

Cerința 13: Realizați cel puțin 3 tipuri de reprezentări grafice pentru analiza factorială (similar cu ACP - biplot, scatterplot-ul observațiilor, etc.)

```
> library(ggplot2)
> ggplot(scoruri, aes(x = PA1, y = PA2)) +
+   geom_point(color = "blue") +
+   labs(title = "Scatterplot al observațiilor pe factori", x = "Factor 1", y = "Factor 2") +
+   theme_minimal()
>
```

1) Scatterplot

Graficul prezintă distribuția observațiilor în funcție de **Factor 1** și **Factor 2**, oferind o perspectivă asupra relațiilor latente dintre variabile. **Factorul 1**, asociat cu dimensiunea socio-economică generală, diferențiază observațiile pe baza variabilelor precum educație, comerț și sănătate. **Factorul 2**, o dimensiune secundară, reflectă aspecte legate



de utilități publice sau logistică, în funcție de metoda utilizată (PA sau ML). Majoritatea punctelor sunt concentrate în jurul centrului graficului, indicând observații similare, în timp ce punctele de la periferie sugerează variații semnificative pe cei doi factori. Acest scatterplot oferă o imagine clară a structurii și relațiilor dintre observații.

2) Histograme

```
> hist(scoruri$PA1,
+      freq = FALSE,
+      col = "darkblue",
+      main = "Histograma scorurilor factoriale (Factor 1)",
+      xlab = "Scoruri Factor 1")
> lines(density(scoruri$PA1),
+       lwd = 3,
+       col = "red")
```

Majoritatea valorilor sunt concentrate între -1 și 1, ceea ce sugerează că observațiile sunt destul de similare în ceea ce privește contribuțiile lor la Factorul 1.

Existența unor valori mai extreme, atât pozitive, cât și negative, indică observații care se abat mai mult de la centru.

Distribuția este ușor asimetrică, cu o tendință de densitate mai mare în jurul valorii de 0.

Această formă poate indica o influență mai mare a variabilelor socio-economice în regiuni specifice.

```
> hist(scoruri$PA2,
+      freq = FALSE,
+      col = "magenta",
+      main = "Histograma scorurilor factoriale (Factor 2)",
+      xlab = "Scoruri Factor 2")
> lines(density(scoruri$PA2),
+       lwd = 3,
+       col = "black")
```

Scorurile sunt mai uniform distribuite comparativ cu Factorul 1, dar majoritatea valorilor se concentrează între -1 și 1.

Existența câtorva scoruri la extreme (-2 și 2) sugerează că unele observații au contribuții mai mari sau mai mici pe acest factor.

Distribuția are o densitate ușor mai mare în jurul valorii de 0, indicând o tendință de similaritate între observații pe Factorul 2.

Linia neagră (densitatea estimată) arată o formă aproape simetrică, dar nu perfect normală.

Factorul 2 contribuie la identificarea unei dimensiuni diferite de Factorul 1, dar impactul său

este mai mic, ceea ce se reflectă și în concentrarea mai scăzută a valorilor la centru.

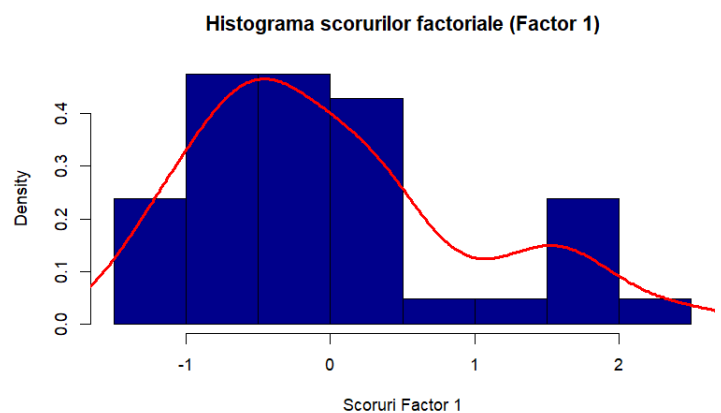


Figure 5

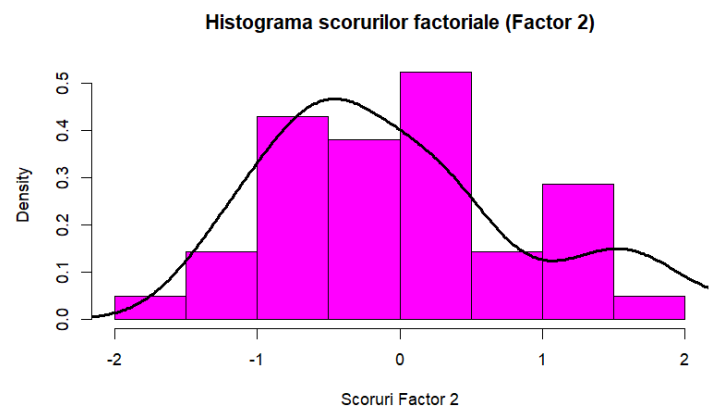


Figure 6

3) Corrplot

```
> library(corrplot)
> corrplot(cor_factori,
+         method = "number",
+         type = "upper",
+         main = "Corelațiile dintre factori")
```

Graficul matricei de corelații confirmă regula ortogonalității, specifică analizei factoriale cu rotație ortogonală, conform căreia factorii trebuie să fie independenți unul de celălalt. Corelațiile aproape de 0 între factori indică faptul că dimensiunile latente identificate sunt distincte și nu se suprapun. Factorii PA1, PA2 și PA3 sunt complet necorelați, conform așteptărilor analizei factoriale ortogonale. Fiecare factor explică o dimensiune specifică a structurii latente din date, ceea ce permite interpretarea clară și separată a contribuției fiecărui set de variabile la analiza generală.

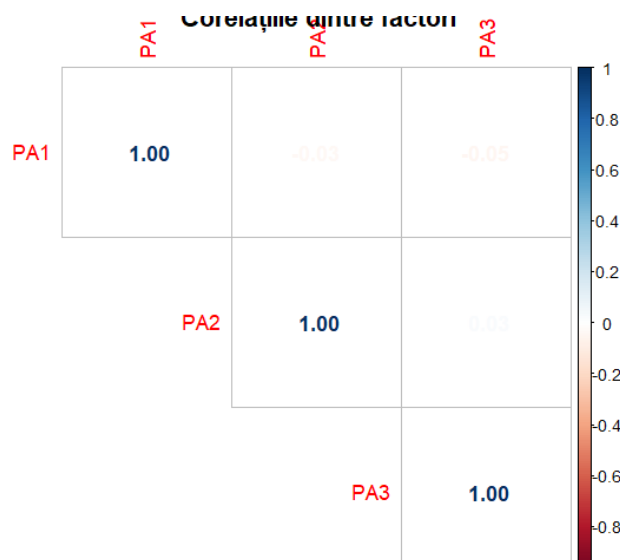


Figure 7

Cerința 14: Formulați 3 concluzii principale din analiza voastră.

- Dimensiunea socio-economică ca factor dominant:** Primul factor identificat, reprezentând **dimensiunea socio-economică generală**, reflectă variabile critice precum educația, comerțul, infrastructura și sănătatea. În realitate, aceasta indică faptul că dezvoltarea economică și socială a județelor este determinată de accesul la educație, serviciile de sănătate, investițiile în infrastructură și activitatea economică generală. Acest factor poate fi folosit pentru a evalua nivelul general de dezvoltare al regiunilor și pentru a identifica județele care necesită prioritizare în politicile publice.
- Independența dimensiunilor latente (ortogonalitatea factorilor):** Faptul că cei trei factori identificați sunt necorelați demonstrează că aceștia surprind dimensiuni distincte ale realității socio-economice. Spre exemplu, factorul **utilități publice și administrație** (PA2) arată că infrastructura utilităților și funcționarea administrației publice reprezintă un domeniu separat de dinamica educației și comerțului. Această separare este importantă pentru planificarea resurselor, deoarece fiecare factor evidențiază domenii diferite care necesită abordări specifice.
- Setul de date utilizat evidențiază legături specifice între variabile precum educația, sănătatea și sectorul agricol. Factorul **logistică și servicii** (PA3) surprinde sectoare mai specializate, cum ar fi transportul, turismul și agricultura, care pot indica priorități pentru dezvoltarea regională. De exemplu, județele care au scoruri mari pe acest factor pot beneficia de politici axate pe îmbunătățirea logisticii și promovarea turismului sau agriculturii, în timp ce județele mai slab reprezentate pe acest factor pot necesita sprijin pentru diversificarea economică. Analiza factorială oferă astfel o bază pentru alocarea resurselor și definirea strategiilor regionale, bazându-se pe date obiective care reflectă realitatea socio-economică.

II. ANALIZA CORESPONDENȚELOR

Cerința 1: Descrieți în cuvinte tabelul de contingență, din perspectiva liniilor și coloanelor. Enunțați clar sursa datelor.

Tabelul de contingență prezintă populația rezidentă la nivelul județelor din România, grupată în funcție de **categorii de vârstă**. Structura tabelului permite analiza relațiilor dintre aceste două dimensiuni:

1. Liniile (categorii de vârstă):

Reprezintă diferite grupe de vârstă ale populației, cum ar fi copii, tineri, adulți și vârstnici.

Fiecare linie arată numărul de persoane dintr-o anumită categorie de vârstă care locuiesc în județe specifice.

Aceste informații sunt esențiale pentru a înțelege structura demografică a populației și pentru a evidenția distribuția vârstelor în diferite regiuni.

2. Coloanele (județele):

Fiecare coloană reprezintă un județ din România.

Valorile din tabel indică populația din fiecare grupă de vârstă, specifică fiecărui județ.

Analiza coloanelor ajută la identificarea județelor cu populație predominant tânără, adultă sau vârstnică și oferă informații utile pentru luarea deciziilor administrative și economice.

Datele provin din sursa oficială: **1.3.2 POPULAȚIA REZIDENTĂ DUPĂ GRUPA DE VÂRSTĂ, PE JUDEȚE, LA 1 DECEMBRIE 2021**, publicată de Institutul Național de Statistică al României. Aceste date au fost colectate în cadrul Recensământului Populației și Locuințelor 2021 și sunt disponibile pe site-ul oficial: www.recensamantromania.ro.

Cerința 2: Descrieți obiectivul general al analizei voastre.

Obiectivul general al analizei corespondentelor este să exploreze și să identifice relațiile dintre categoriile de vârstă ale populației rezidente și distribuția acestora în județele României. Această analiză permite evidențierea asocierilor semnificative dintre structura demografică și distribuția geografică, oferind o perspectivă clară asupra variațiilor regionale și a caracteristicilor specifice fiecărui județ.

În mod specific, analiza urmărește:

Identificarea asocierilor dintre variabilele categoriale: Evidențierea județelor cu o proporție mai mare de populație într-o anumită categorie de vârstă.

Determinarea principalelor dimensiuni latente: Descoperirea axelor principale care explică cel mai bine variația datelor, oferind o perspectivă compactă și interpretabilă asupra structurii datelor.

Cerința 3: Reprezentați grafic tabelul de contingență (de exemplu prin balloonplot).

```
> date_ac<-Date_AC_tema
> date2=date_ac[, -1]
> date2=as.table(as.matrix(date2))
> rownames(date2)=date_ac$Populatie_Categorii_varsta
>
> View(date2)
> install.packages("gplots")

> balloonplot(t(date2),main="Matricea de contingenta",)
```

Matricea de contingenta

y	x	ARAD BRASOVAR MARAMURESASLUI GORJ					
		ARAD	BRASOVAR	MARAMURESASLUI	GORJ		
5-9ani		21333	31659	23983	23752	14038	114765
10-14ani		22789	32321	26114	26514	15830	123568
15-19ani		21801	28059	26474	26013	17246	119593
20-24ani		20527	24953	24199	21607	16740	108026
25-29ani		20843	26335	23986	17489	15975	104628
30-34ani		27219	36913	30651	20273	19634	134690
35-39ani		26213	37626	30579	20137	20115	134670
40-44ani		30927	42495	34811	25513	24188	157934
45-49ani		32797	37812	35211	28376	26046	160242
50-54ani		36573	42104	36164	32305	32744	179890
55-59ani		21910	28013	27127	19049	18666	114765
		262932	368290	319299	261028	221222	1452771

Figure 8

Cerinta 4: Aplicați testul Chi-Pătrat (inclusiv ipotezele) și interpretați rezultatele.

```
> x2=chisq.test(date2)
> x2

Pearson's Chi-squared test

data:  date2
X-squared = 9789.9, df = 40, p-value < 2.2e-16
```

Ipoteze:

- **Ipoteza nulă (H0):** Nu există asociere între categoriile de vârstă și județele din România. Variabilele sunt independente.
- **Ipoteza alternativă (H1):** Există o asociere între categoriile de vârstă și județe. Variabilele nu sunt independente.

Valoarea p-value este mult mai mică decât pragul de semnificație standard ($\alpha=0.05$), ceea ce înseamnă că putem respinge ipoteza nulă (H0).

Acest rezultat indică faptul că există o asociere semnificativă statistic între categoriile de vârstă și distribuția populației în județe. Variabilele nu sunt independente.

Distribuția populației pe grupe de vârstă variază în funcție de județ, ceea ce sugerează diferențe semnificative în structura demografică între regiuni.

Cerința 5: Aplicați metoda analizei de corespondență. Calculați și afișați inerția totală, valorile proprii și procentele de varianță.

```
> all=date2[1,1]
> all
[1] 21333
>
> row_totals=rowSums(date2) #totaluri pe linii
> col_totals=colSums(date2) #totaluri pe coloane
>
> #valoarea asteptata:
> e11=(row_totals[1]*col_totals[1])/sum(date2)
> e11
5-9ani
22350.87
```

```
> X2$expected
      ARAD  BRASOV  MARAMURES  VASLUI  GORJ
5-9ani  22350.87 29093.92  25223.76 20620.51 17475.94
10-14ani 24065.28 31325.56  27158.54 22202.20 18816.43
15-19ani 23291.14 30317.86  26284.89 21487.99 18211.13
20-24ani 21038.42 27385.52  23742.62 19409.67 16449.76
25-29ani 20376.65 26524.10  22995.79 18799.13 15932.32
30-34ani 26231.33 34145.08  29603.00 24200.55 20510.04
35-39ani 26227.43 34140.01  29598.61 24196.96 20506.99
40-44ani 30758.17 40037.63  34711.71 28376.94 24049.54
45-49ani 31207.66 40622.73  35218.98 28791.63 24400.99
50-54ani 35034.18 45603.67  39537.34 32321.91 27392.91
55-59ani 22350.87 29093.92  25223.76 20620.51 17475.94
> X2$observed
      ARAD  BRASOV  MARAMURES  VASLUI  GORJ
5-9ani  21333  31659    23983   23752  14038
10-14ani 22789  32321    26114   26514  15830
15-19ani 21801  28059    26474   26013  17246
20-24ani 20527  24953    24199   21607  16740
25-29ani 20843  26335    23986   17489  15975
30-34ani 27219  36913    30651   20273  19634
35-39ani 26213  37626    30579   20137  20115
40-44ani 30927  42495    34811   25513  24188
45-49ani 32797  37812    35211   28376  26046
50-54ani 36573  42104    36164   32305  32744
55-59ani 21910  28013    27127   19049  18666
> |
```

```
> eig=get_eigenvalue(rez)
> eig
      eigenvalue variance.percent cumulative.variance.percent
Dim.1 3.725872e-03      55.289813              55.28981
Dim.2 2.467118e-03      36.610630              91.90044
Dim.3 4.564989e-04       6.774184              98.67463
Dim.4 8.931431e-05       1.325373             100.00000
```

```
> s=sum(eig[,1])
> s
[1] 0.006738803
```

```
> rez=CA(date2, graph=F)
> rez
**Results of the Correspondence Analysis (CA)**
The row variable has 11 categories; the column variable has 5 categories
The chi square of independence between the two variables is equal to 9789.938 (p-value = 0 )
*The results are available in the following objects:
```

	name	description
1	"\$eig"	"eigenvalues"
2	"\$col"	"results for the columns"
3	"\$col\$coord"	"coord. for the columns"
4	"\$col\$cos2"	"cos2 for the columns"
5	"\$col\$contrib"	"contributions of the columns"
6	"\$row"	"results for the rows"
7	"\$row\$coord"	"coord. for the rows"
8	"\$row\$cos2"	"cos2 for the rows"
9	"\$row\$contrib"	"contributions of the rows"
10	"\$call"	"summary called parameters"
11	"\$call\$marge.col"	"weights of the columns"
12	"\$call\$marge.row"	"weights of the rows"

```
> summary(rez, nb.dec=2)
```

Call:
CA(X = date2, graph = F)

The chi square of independence between the two variables is equal to 9789.938 (p-value = 0).

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4
Variance	0.00	0.00	0.00	0.00
% of var.	55.29	36.61	6.77	1.33
Cumulative % of var.	55.29	91.90	98.67	100.00

Rows (the 10 first)

	Iner*1000	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
5-9ani	1.02	0.09	16.08	0.59	-0.07	15.35	0.37	0.02	9.72	0.04
10-14ani	1.00	0.10	23.36	0.87	-0.04	4.80	0.12	0.01	2.17	0.01
15-19ani	0.87	0.10	20.81	0.89	0.03	2.11	0.06	-0.02	9.35	0.05
20-24ani	0.34	0.05	4.36	0.48	0.04	5.34	0.39	-0.02	9.58	0.13
25-29ani	0.10	-0.03	1.84	0.68	0.00	0.05	0.01	-0.02	5.28	0.24
30-34ani	0.67	-0.07	10.71	0.60	-0.05	10.52	0.39	0.00	0.03	0.00
35-39ani	0.74	-0.07	11.77	0.59	-0.06	11.87	0.40	0.00	0.07	0.00
40-44ani	0.30	-0.04	5.49	0.67	-0.03	3.41	0.28	0.01	2.27	0.03
45-49ani	0.27	-0.02	0.71	0.10	0.04	8.34	0.76	0.00	0.54	0.01
50-54ani	1.15	-0.02	1.68	0.05	0.09	37.08	0.80	0.04	36.84	0.15

Columns

	Iner*1000	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
ARAD	0.29	-0.03	4.63	0.59	0.01	1.50	0.13	0.01	5.08	0.08
BRASOV	1.29	-0.02	1.81	0.05	-0.07	46.18	0.88	0.02	16.86	0.06
MARAMURES	0.45	-0.02	1.83	0.15	-0.01	1.77	0.10	-0.04	74.34	0.75
VASLUI	2.99	0.13	77.51	0.97	0.02	3.93	0.03	0.00	0.40	0.00
GORJ	1.71	-0.06	14.22	0.31	0.09	46.63	0.67	0.01	3.31	0.01

Valorile observate și așteptate:

Valorile observate reprezintă numărul real al populației rezidente din grupele de vârstă (rânduri) și județe (coloane). De exemplu, pentru grupa de vârstă 5-9 ani în județul Arad, valoarea observată este 21333.

Valorile așteptate, calculate pe baza ipotezei de independență între rânduri și coloane, sunt determinate astfel încât să reflecte o distribuție proporțională. De exemplu, valoarea așteptată pentru aceeași celulă este 22350.87.

Diferențele semnificative între valorile observate și așteptate confirmă existența unor asocieri între grupe de vârstă și județe.

Rezultatele analizei de corespondență:

Dimensiunile principale:

Dimensiunea 1 explică 55.29% din variația totală, fiind dimensiunea dominantă.

Dimensiunea 2 adaugă încă 36.61%, împreună reprezentând 91.90% din variație. Aceste două dimensiuni sunt esențiale pentru a interpreta relațiile dintre grupe de vârstă și județe.

Dimensiunile 3 și 4 explică împreună doar 8.10%, indicând contribuții marginale.

Inerția totală:

Inerția totală, 0.006738803, reflectă variația generală a datelor. Această valoare mică indică faptul că structura datelor poate fi eficient rezumată în primele două dimensiuni.

Valorile proprii:

Valorile proprii pentru Dimensiunea 1 ($3.725872e-03$) și Dimensiunea 2 ($2.467118e-03$) sunt mult mai mari decât cele pentru Dimensiunile 3 și 4. Acest lucru arată că primele două dimensiuni surprind majoritatea variației, justificând concentrarea analizei pe acestea.

Rânduri (grupe de vârstă):

Grupele de vârstă 50-54 ani și 5-9 ani sunt cel mai bine reprezentate pe Dimensiunea 1 (\cos^2 mari), contribuind substanțial la această dimensiune.

Grupele de vârstă 15-19 ani și 10-14 ani contribuie mai mult la Dimensiunea 2, ceea ce indică diferențe semnificative în distribuția lor pe județe.

Coloane (județe):

Județele Gorj și Maramureș au cele mai mari contribuții la Dimensiunea 1. Acestea sunt asociate în special cu anumite grupe de vârstă, sugerând distribuții specifice.

Județul Brașov contribuie mai mult la Dimensiunea 2, indicând o relație distinctă cu grupele de vârstă.

Primele două dimensiuni oferă o descriere aproape completă a relației dintre grupe de vârstă și județe, explicând 91.90% din variație.

Cerința 6: Realizați Scree Plot-ul și interpretați.

Scree Plot-ul arată procentajele de variație explicate de fiecare dimensiune identificată în analiza de corespondență.

Prima dimensiune explică aproximativ **55%** din variația totală, fiind cea mai importantă și reprezentativă. A doua dimensiune adaugă încă **36%**, astfel încât primele două dimensiuni explică cumulativ **91%** din variație. Acestea sunt principalele dimensiuni pe care se bazează analiza.

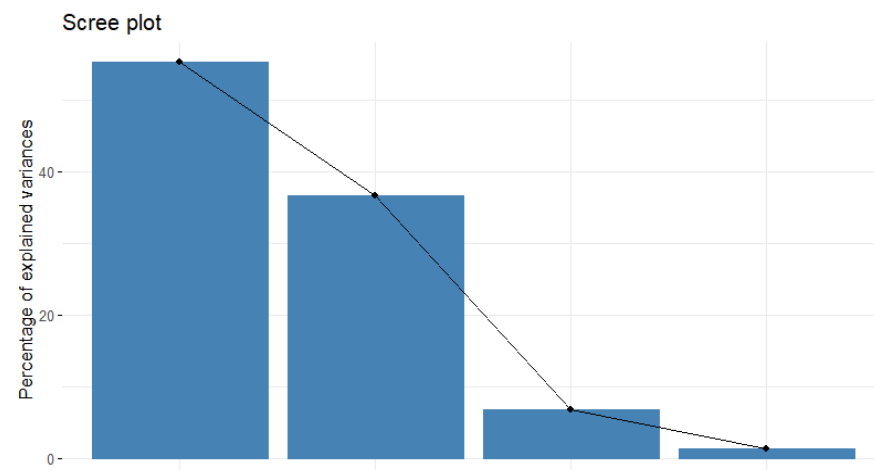


Figure 9

Dimensiunea 3 și Dimensiunea 4 explică împreună doar **9%** din variație. Acestea contribuie marginal și nu oferă informații semnificative pentru interpretarea datelor. Observăm un punct de cotitură clar între a doua și a treia dimensiune. Aceasta sugerează că analiza ar trebui să se concentreze pe primele două dimensiuni, deoarece ele surprind aproape întreaga variație relevantă din datele analizate.

Cerința 7: Extrageți și interpretați indicatorii pentru dimensiunea linie (cos2, contrib, coord, inerția).

Coordonate (rezultate_linii\$coord):

Coordonatele arată poziția categoriilor de vârstă pe fiecare dimensiune. De exemplu, categoria „10-14 ani” are coordonata **0.101** pe Dim 1, ceea ce indică o asociere pozitivă cu această dimensiune.

Calitatea reprezentării (Cos2, rezultate_linii\$cos2): Cos2 indică cât de bine sunt reprezentate liniile pe fiecare dimensiune. Categoria „15-19 ani” are un Cos2 de **0.88** pe Dim 1, ceea ce înseamnă că această dimensiune explică aproape întreaga variație a acestei categorii. Valorile mari (aproape de 1) indică o reprezentare foarte bună.

Contribuția (rezultate_linii\$contrib): Contribuția arată cât de mult contribuie fiecare categorie de vârstă la inerția totală a fiecărei dimensiuni. Categoriile „10-14 ani” și „15-19 ani” au contribuții semnificative de **23.35** și **20.81** pe Dim 1, ceea ce confirmă că sunt principalele linii care determină structura acestei dimensiuni.

```
> rezultate_linii <- get_ca_row(rez)
> print(rezultate_linii$coord)
```

	Dim 1	Dim 2	Dim 3	Dim 4
5-9ani	0.08709011	-0.069228155	0.023694478	-0.0021260086
10-14ani	0.10115103	-0.037323209	0.010792627	-0.0005882123
15-19ani	0.09705352	0.025151783	-0.022775886	0.0063692720
20-24ani	0.04676040	0.042090678	-0.024253790	0.0004265171
25-29ani	-0.03081989	-0.004145856	-0.018287373	-0.0097312485
30-34ani	-0.06559028	-0.052919165	-0.001203971	-0.0111218973
35-39ani	-0.06876934	-0.056216452	0.001795147	0.0102767021
40-44ani	-0.04337287	-0.027824489	0.009772793	0.0068506498
45-49ani	-0.01544015	0.043192166	-0.004746743	-0.0179388503
50-54ani	-0.02250738	0.085955314	0.036851736	0.0051554750
55-59ani	-0.03885313	0.018689634	-0.037357639	0.0131248600

```
> print(rezultate_linii$cos2)
```

	Dim 1	Dim 2	Dim 3	Dim 4
5-9ani	0.58599905	0.37027533	0.0433764057	3.492121e-04
10-14ani	0.87140790	0.11864210	0.0099205378	2.946789e-05
15-19ani	0.88767445	0.05961681	0.0488856859	3.823057e-03
20-24ani	0.48091772	0.38966034	0.1293819249	4.001172e-05
25-29ani	0.68033225	0.01231083	0.2395309377	6.782598e-02
30-34ani	0.59522438	0.38746075	0.0002005552	1.711431e-02
35-39ani	0.59127507	0.39511797	0.0004029026	1.320406e-02
40-44ani	0.67237623	0.27671358	0.0341360698	1.677412e-02
45-49ani	0.09737312	0.76198456	0.0092029632	1.314394e-01
50-54ani	0.05459138	0.79619527	0.1463490934	2.864254e-03
55-59ani	0.44052747	0.10193482	0.4072675458	5.027017e-02

```
> print(rezultate_linii$contrib)
```

	Dim 1	Dim 2	Dim 3	Dim 4
5-9ani	16.0813335	15.34574042	9.71553775	0.39978022
10-14ani	23.3572439	4.80260476	2.17031842	0.03295004
15-19ani	20.8114951	2.11084422	9.35447478	3.73910070
20-24ani	4.3637479	5.33965254	9.58187748	0.01514548
25-29ani	1.8360520	0.05017521	5.27610789	7.63601661
30-34ani	10.7050637	10.52383885	0.02943952	12.84029276
35-39ani	11.7661791	11.87437049	0.06543861	10.96125319
40-44ani	5.4889193	3.41148048	2.27444869	5.71243129
45-49ani	0.7057553	8.34063640	0.54441604	39.74178609
50-54ani	1.6835729	37.08218773	36.83714845	3.68490427
55-59ani	3.2006372	1.11846889	24.15079235	15.23633935

Cerința 8: Realizați matricea factor pentru rânduri și interpretați.

```
> matrice_factori_randuri <- rez$row$coord
> print(matrice_factori_randuri)
```

	Dim 1	Dim 2	Dim 3	Dim 4
5-9ani	0.08709011	-0.069228155	0.023694478	-0.0021260086
10-14ani	0.10115103	-0.037323209	0.010792627	-0.0005882123
15-19ani	0.09705352	0.025151783	-0.022775886	0.0063692720
20-24ani	0.04676040	0.042090678	-0.024253790	0.0004265171
25-29ani	-0.03081989	-0.004145856	-0.018287373	-0.0097312485
30-34ani	-0.06559028	-0.052919165	-0.001203971	-0.0111218973
35-39ani	-0.06876934	-0.056216452	0.001795147	0.0102767021
40-44ani	-0.04337287	-0.027824489	0.009772793	0.0068506498
45-49ani	-0.01544015	0.043192166	-0.004746743	-0.0179388503
50-54ani	-0.02250738	0.085955314	0.036851736	0.0051554750
55-59ani	-0.03885313	0.018689634	-0.037357639	0.0131248600

```
> fviz_ca_row(rez, repel = TRUE)
```


Matricea factor pentru rânduri prezintă poziția categoriilor de vârstă pe dimensiunile principale ale analizei. Dimensiunea 1 (55.3% din variație) este dominantă, iar categoriile „5-9 ani” și „10-14 ani” se asociază puternic cu această axă. Dimensiunea 2 (36.6% din variație) evidențiază categoriile „50-54 ani” și „45-49 ani”.

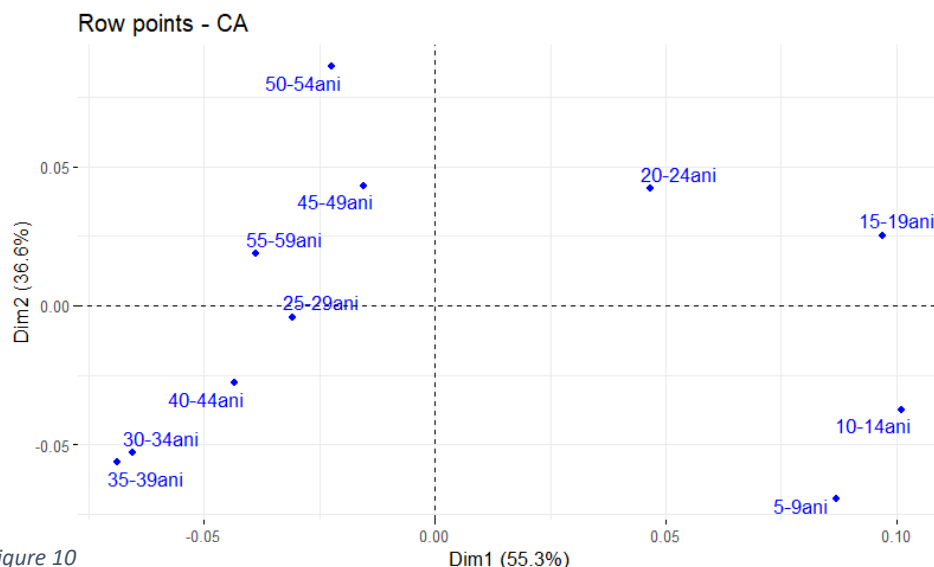


Figure 10

Cerința 9: Extrageți și interpretați indicatorii pentru coloane (cos2, contrib, coord, inerția).

```
> rezultate_coloane <- get_ca_col(rez)
> print(rezultate_coloane$coord)
      Dim 1      Dim 2      Dim 3      Dim 4
ARAD    -0.02975137  0.01378220  0.010916538 -0.0178291165
BRASOV   -0.01631475 -0.06704126  0.017425675  0.0058734399
MARAMURES -0.01759380 -0.01408134 -0.039293583  0.0006120211
VASLUI    0.12678258  0.02321615  0.003200320  0.0009711166
GORJ     -0.05899029  0.08691401  0.009965957  0.0109952540
> print(rezultate_coloane$cos2)
      Dim 1      Dim 2      Dim 3      Dim 4
ARAD    0.58535794  0.12561601  0.0788093061  2.102168e-01
BRASOV   0.05220214  0.88147874  0.0595534154  6.765696e-03
MARAMURES 0.15083513  0.09662088  0.7523614690  1.825223e-04
VASLUI    0.96690478  0.03242239  0.0006161002  5.672928e-05
GORJ     0.30920724  0.67122517  0.0088252430  1.074235e-02
> print(rezultate_coloane$contrib)
      Dim 1      Dim 2      Dim 3      Dim 4
ARAD    4.626695  1.499451  5.0841105  69.31441012
BRASOV   1.811030  46.183536  16.8628985  9.79167065
MARAMURES 1.825962  1.766437  74.3368075  0.09217478
VASLUI   77.514195  3.925369  0.4031225  0.18971937
GORJ    14.222118  46.625207  3.3130609  20.61202507
```

Coordonatele pentru județe pe fiecare dimensiune arată poziționarea acestora în spațiul factorial:

Dimensiunea 1 (55.3% variație): Vaslui are o poziționare proeminentă (0.1268), fiind cel mai bine reprezentat pe această dimensiune. Dimensiunea 2 (36.6% variație): Brașov are o coordonată semnificativă (0.0670), indicând asocierea sa cu această dimensiune.

Calitatea reprezentării (Cos2):

Valoarea \cos^2 indică cât de bine este reprezentat fiecare județ pe cele patru dimensiuni:

Vaslui are cea mai bună reprezentare pe Dimensiunea 1 (0.9669), arătând că poziționarea sa este explicată aproape complet de această dimensiune.

Brașov este bine reprezentat pe Dimensiunea 2 (0.8815), ceea ce sugerează că această axă descrie majoritatea informațiilor legate de acest județ.

Contribuția (Contrib): Contribuția fiecărui județ la inerția totală indică importanța sa relativă:

Vaslui contribuie semnificativ pe Dimensiunea 1 (77.51%), fiind factorul determinant pentru această axă. Maramureș contribuie semnificativ pe Dimensiunea 2 (74.38%), indicând relevanța sa pe această dimensiune. Inerția calculată pentru județe arată distribuția variației în funcție de fiecare dimensiune, subliniind faptul că majoritatea variației este capturată de primele două dimensiuni (91.9%).

Cerința 10: Realizați matricea factor pentru coloane și interpretați.

```
> matrice_factori_coloane <- rez$col$coord
> print(matrice_factori_coloane)
```

	Dim 1	Dim 2	Dim 3	Dim 4
ARAD	-0.02975137	0.01378220	0.010916538	-0.0178291165
BRASOV	-0.01631475	-0.06704126	0.017425675	0.0058734399
MARAMURES	-0.01759380	-0.01408134	-0.039293583	0.0006120211
VASLUI	0.12678258	0.02321615	0.003200320	0.0009711166
GORJ	-0.05899029	0.08691401	0.009965957	0.0109952540

```
> fviz_ca_col(rez, repel = TRUE)
```

Dimensiunea 1 (Dim 1): Este dimensiunea dominantă, explicând 55.3% din variație. Vaslui are cea mai mare coordonată pozitivă pe această dimensiune (0.12678258), ceea ce sugerează că este puternic asociat cu tiparele reflectate pe această axă, cum ar fi structura specifică a populației. În schimb, Brașov are o coordonată negativă, indicând un alt tipar demografic. **Dimensiunea 2 (Dim 2):** Explică 36.6% din variație. Gorj (0.08691401) se poziționează semnificativ pe această dimensiune, ceea ce poate indica specificități distincte față de alte județe. **Dimensiunile 3 și 4:** Contribuțiile acestor dimensiuni sunt foarte mici (sub 7% fiecare), indicând că variațiile observate sunt aproape neglijabile pentru interpretare practică.

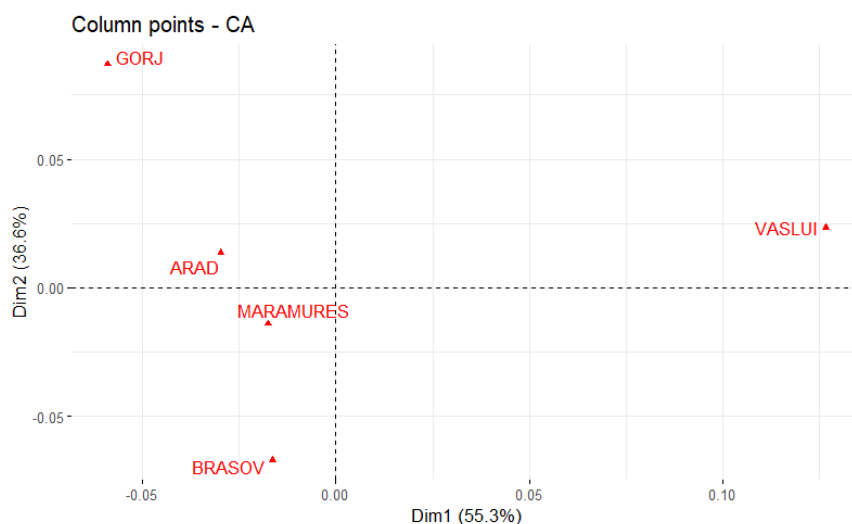


Figure 11

Vaslui se distinge printr-o coordonată pozitivă puternică pe Dim 1, sugerând particularități evidente în structura populației. Brașov și Maramureș sunt mai aproape de origine, indicând un profil mai mediu sau echilibrat. Gorj are o coordonată pozitivă semnificativă pe Dim 2, sugerând diferențe față de alte județe în ceea ce privește anumite caracteristici demografice.

Cerința 11: Realizați cel puțin 3 reprezentări grafice (similar cu ACP - biploturi, reprezentări ale contrib/cos2/coord/inerția pe rânduri/coloane, etc). Interpretați legăturile dintre variabilele categoriale aflate pe rânduri, respectiv coloane.

A) Biplot

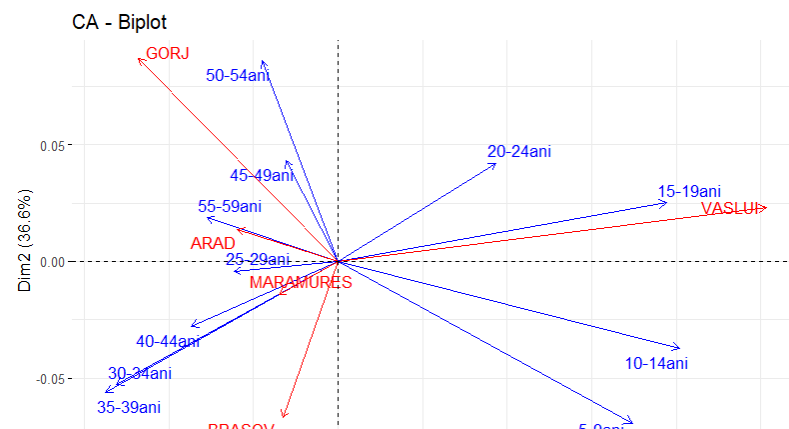


Figure 12

Liniile indică direcția și magnitudinea asociată fiecărei categorii de vârstă. De exemplu:

Categoria **5-9 ani** are o asociere pozitivă cu Vaslui, sugerând că această categorie este relativ bine reprezentată aici. Categoria **50-54 ani** se corelează mai bine cu Gorj.

Unghiurile mici între săgeți (linii și coloane) indică asocieri puternice: **Vaslui** și categoria **15-19 ani** au un unghi redus, ceea ce confirmă asocierea puternică dintre județ și această grupă de vârstă. Județele precum Vaslui și Gorj au profiluri demografice distincte, concentrate pe grupe de vârstă tinere și în vârstă, respectiv. Brașov și Maramureș prezintă distribuții mai echilibrate între categoriile de vârstă.

B) Corrplot

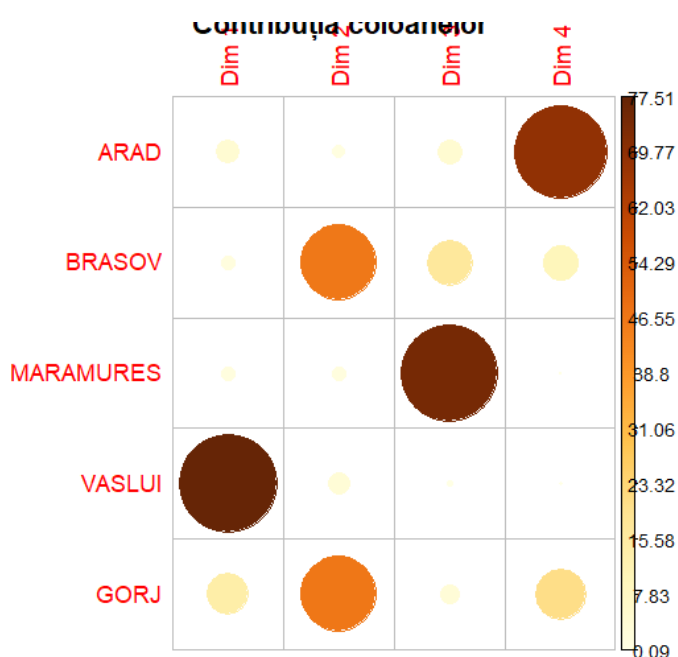


Figure 13

Contribuția Coloanelor: Graficul arată contribuția fiecărei coloane (județ) la dimensiunile principale ale analizei de corespondență. **Vaslui** are cea mai mare contribuție pe Dim1, sugerând o diferență notabilă față de celelalte județe pentru această dimensiune. **Maramureș** și **Gorj** contribuie semnificativ la Dim2, ceea ce poate reflecta un profil distinct al distribuției pe categorii de vârstă pentru aceste județe. **Brașov** contribuie substanțial pe Dim4, arătând specificități care apar doar la o dimensiune mai puțin explicativă.

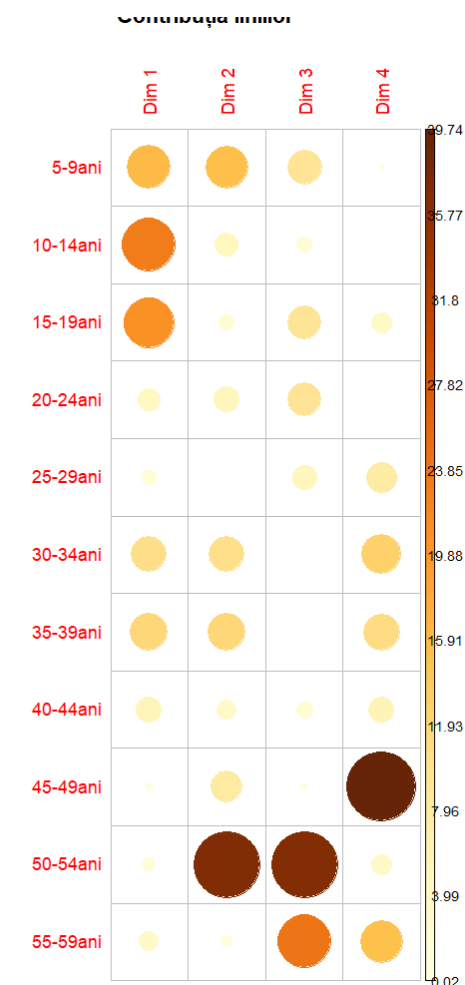


Figure 14

Contribuția Liniilor: Graficul evidențiază contribuția diferitelor grupe de vârstă la dimensiunile analizei. **Grupele de vârstă 10-14 ani și 50-54 ani** contribuie semnificativ pe Dim1, ceea ce sugerează că aceste grupe joacă un rol central în separarea județelor. **Grupele 45-49 ani și 55-59 ani** sunt importante pentru Dim2, indicând un comportament diferit în funcție de aceste categorii. Dim3 și Dim4 au contribuții mai reduse, iar liniile care contribuie aici pot fi mai puțin relevante în analiza principală.

C) Calitatea reprezentării (\cos^2) pentru linii și coloane

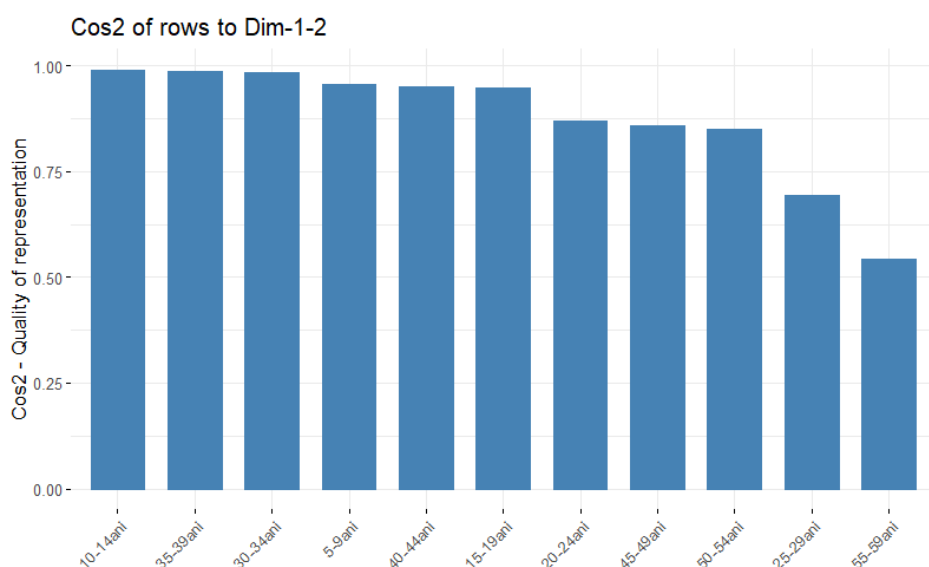


Figure 15

Vaslui și Gorj sunt cel mai bine reprezentate ($\cos^2 \approx 1$), indicând o descriere clară pe primele două dimensiuni.

Maramureș are o reprezentare slabă, sugerând o contribuție redusă la structura factorială.

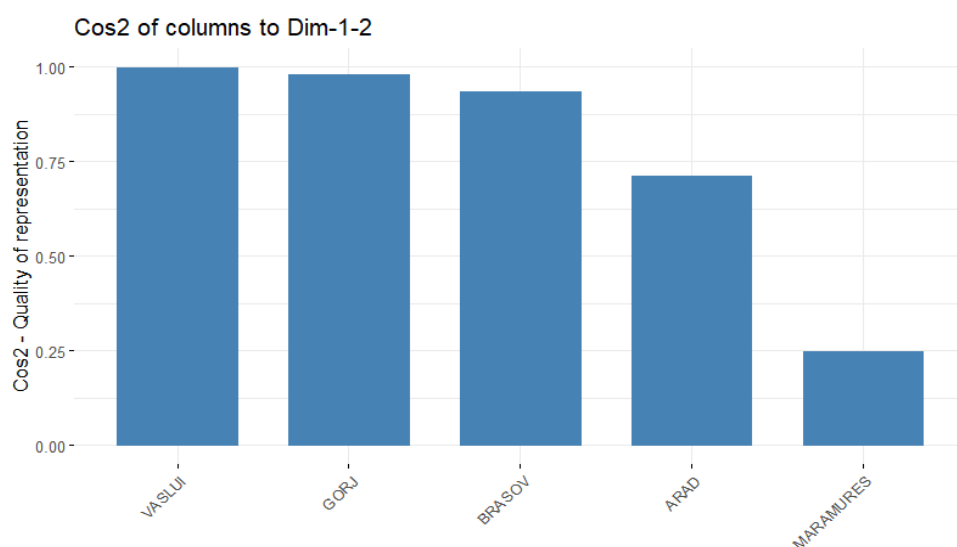


Figure 16

Cerința 12: Formulați 3 concluzii principale din analiza voastră.

Primele două dimensiuni explică 91.9% din variabilitate, indicând că distribuția populației rezidente pe grupe de vârstă este influențată semnificativ de caracteristicile județelor. De exemplu, Vaslui și Gorj sunt județele care au contribuții importante și sunt bine reprezentate pe aceste dimensiuni.

Dimensiunea 1 reflectă variații generale în județe dominate de anumite grupe de vârstă, în timp ce Dimensiunea 2 evidențiază specificități locale. Maramureș și alte grupe de vârstă, precum 25-29 ani, au contribuții reduse, fiind mai puțin relevante în acest context factorial.

Rezultatele evidențiază modul în care grupele de vârstă sunt distribuite la nivel județean, oferind o perspectivă esențială pentru dezvoltarea de politici locale și gestionarea eficientă a resurselor. Județele bine reprezentate în analiza corespondențelor pot beneficia de strategii personalizate care să răspundă nevoilor demografice specifice, contribuind astfel la o planificare mai eficientă și la creșterea calității vieții locuitorilor.

Anexă

Figure 1	2
Figure 2	5
Figure 3	5
Figure 4	6
Figure 5	7
Figure 6	7
Figure 7	8
Figure 8	10
Figure 9	12
Figure 10	14
Figure 11	15
Figure 12	15
Figure 13	16
Figure 14	16
Figure 15	17
Figure 16	17

Cod Rstudio

```
##Tema2-Lica_Amalia
```

```
date<-Date_AD
```

```
#statistici descriptive
```

```
summary(date)
```

```
colnames(date) <- c("Judete", "Agricultura_Vanatoare", "Pescuit",  
  "Energie_Gaze_Apa", "Constructii",  
  "Comert", "Hoteluri_Restaurante", "Transport_Depozitare",  
  "Posta", "Educatie",  
  "Admin_publica_Aparare", "Sanatate")
```

```
View(date)
```

```
windows()
```

```
par(mfrow=c(2,6))
```

```
boxplot(date$Agricultura_Vanatoare, col = "lightblue", main="Boxplot Agricultura_Vanatoare")
```

```
boxplot(date$Pescuit, col = "darkgreen", main="Boxplot Pescuit")
```

```
boxplot(date$Energie_Gaze_Apa, col = "red", main="Boxplot Energie_Gaze_Apa")
```

```
boxplot(date$Constructii, col = "pink", main="Boxplot Constructii")
```

```
boxplot(date$Comert, col = "darkblue", main="Boxplot Comert")
```

```
boxplot(date$Hoteluri_Restaurante, col = "purple", main="Boxplot Hoteluri_Restaurante")
```

```
boxplot(date$Transport_Depozitare, col = "yellow", main="Boxplot Transport_Depozitare")
```

```
boxplot(date$Posta, col = "orange", main="Boxplot Posta")
```

```
boxplot(date$Educatie, col = "darkred", main="Boxplot Educatie")
```

```
boxplot(date$Admin_publica_Aparare, col = "beige", main="Boxplot Admin_publica_Aparare")
```

```
boxplot(date$Sanatate, col = "pink3", main="Boxplot Sanatate")
```

```

boxplot(date$Agricultura_Vanatoare, plot=FALSE)$out
boxplot(date$Pescuit, plot=FALSE)$out
boxplot(date$Energie_Gaze_Apa, plot=FALSE)$out
boxplot(date$Constructii, plot=FALSE)$out
boxplot(date$Comert, plot=FALSE)$out
boxplot(date$Hoteluri_Restaurante, plot=FALSE)$out
boxplot(date$Transport_Depozitare, plot=FALSE)$out
boxplot(date$Posta, plot=FALSE)$out
boxplot(date$Educatie, plot=FALSE)$out
boxplot(date$Admin_publica_Aparare, plot=FALSE)$out
boxplot(date$Sanatate, plot=FALSE)$out

date_noi <- date
for (col in colnames(date_noi)[-1]) { # Exclude prima coloană (Judete)
  date_noi[[col]][is.na(date_noi[[col]])] <- median(date_noi[[col]], na.rm = TRUE)
}
for (col in colnames(date_noi)[-1]) { # Exclude prima coloană (Judete)
  Q1 <- quantile(date_noi[[col]], 0.25, na.rm = TRUE)
  Q3 <- quantile(date_noi[[col]], 0.75, na.rm = TRUE)
  IQR_value <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR_value
  upper_bound <- Q3 + 1.5 * IQR_value
  median_value <- median(date_noi[[col]], na.rm = TRUE)
  date_noi[[col]] <- ifelse(date_noi[[col]] < lower_bound | date_noi[[col]] > upper_bound,
    median_value,
    date_noi[[col]])
}
View(date_noi)

install.packages("psych")
library(psych)
describe(date_noi[-1])
date_std = scale(date_noi[-1], scale=T)
View(date_std)
install.packages("raster")
library(raster)
round(apply(date_std,2,mean),5) #medie=0
apply(date_std,2,sd) #sd=1

```

```

matrice_corelatie <- cor(date_std)

View(matrice_corelatie)

install.packages("corrplot")

library(corrplot)

corrplot(matrice_corelatie, method = "number", type = "lower")

matrice_covarianta <- cov(date_std)

View(matrice_covarianta)

#a.

install.packages("ggpubr")

library(ggpubr)

ggscatter(date_noi, x = "Constructii", y = "Transport_Depozitare",
  add = "reg.line",
  conf.int = TRUE,
  cor.coef = TRUE,
  cor.method = "pearson",
  xlab = "Constructii",
  ylab = "Transport și Depozitare",
  color = "blue")

#b.

install.packages("PerformanceAnalytics")

library(PerformanceAnalytics)

chart.Correlation(date_noi[-1], hist = TRUE)

#c.

install.packages("ggplot2")

library(ggplot2)

ggplot(date_noi, aes(x = Educatie, y = Sanatate)) +
  geom_point() +
  geom_text(aes(label = Judete),
    color = "purple",
    nudge_x = 0.25,
    nudge_y = 0.25,
    check_overlap = TRUE)

install.packages("psych")

library(psych)

KMO(date_std)

```

```

R=cor(date_std)

cortest.bartlett(R, n=42, diag=TRUE)

# 1. Criteriul grafic (Scree Plot)

screeplot_result <- prcomp(date_std)

eigenvalues <- screeplot_result$sdev^2 # Valori proprii

plot(eigenvalues, type = "b", main = "Scree Plot", xlab = "Componenta", ylab = "Valori proprii", col = "magenta")

abline(h = 1, col = "red", lty = 2) # Linie orizontală pentru criteriul lui Kaiser

# 2. Criteriul lui Kaiser

componente_kaiser <- sum(eigenvalues >= 1) # valorile proprii >= 1

print(paste("Numărul de factori conform criteriului lui Kaiser:", componente_kaiser))

# 3. Criteriul procentului de varianță cumulată

var_cumulativa <- cumsum(eigenvalues) / sum(eigenvalues) * 100 # Varianță cumulată în procente

print(var_cumulativa)

componente_var_cumulativa <- which(var_cumulativa >= 80)[1]

print(paste("Numărul de factori conform criteriului procentului de varianță cumulată:", componente_var_cumulativa))

install.packages("GPArotation")

library(GPArotation)

library(psych)

factor1=fa(date_std, nfactors = 3, rotate="none", fm="pa")

print(factor1$loadings, cutoff=0.4)

#2.METODA VEROSIMILITATII MAXIME

factor2=fa(date_std, nfactors=3, rotate="none", fm="ml")

print(factor2$loadings, cutoff=0.4)

fa.diagram(factor1)

fa.diagram(factor2)

scoruri_PA <- factor1$scores # Scorurile factoriale pentru metoda PA

colnames(scoruri_PA) <- c("Dimensiunea_socio_economica_generala",
                        "Utilitati_publice_administratie",
                        "Redundanta_PA") # Redenumirea factorilor

date_noi_PA <- cbind(date_noi, scoruri_PA)

write.csv(date_noi_PA, "Date_Factoriale_PA.csv", row.names = FALSE)

```

```

scoruri_ML <- factor2$scores # Scorurile factoriale pentru metoda ML
colnames(scoruri_ML) <- c("Dimensiunea_socio_economica_generala",
                          "Infrastructura_si_logistica",
                          "Redundanta_ML") # Redenumirea factorilor

date_noi_ML <- cbind(date_noi, scoruri_ML) # Combinarea cu setul de date original curățat
write.csv(date_noi_ML, "Date_Factoriale_ML.csv", row.names = FALSE)

#1)scatterplot
scoruri <- as.data.frame(factor1$scores)

library(ggplot2)

ggplot(scoruri, aes(x = PA1, y = PA2)) +
  geom_point(color = "blue") +
  labs(title = "Scatterplot al observațiilor pe factori", x = "Factor 1", y = "Factor 2") +
  theme_minimal()

#2) histograme
# Histogramă pentru Factorul 1
hist(scoruri$PA1,
     freq = FALSE,
     col = "darkblue",
     main = "Histograma scorurilor factoriale (Factor 1)",
     xlab = "Scoruri Factor 1")
lines(density(scoruri$PA1),
     lwd = 3,
     col = "red")

# Histogramă pentru Factorul 2
hist(scoruri$PA2,
     freq = FALSE,
     col = "magenta",
     main = "Histograma scorurilor factoriale (Factor 2)",
     xlab = "Scoruri Factor 2")
lines(density(scoruri$PA1),
     lwd = 3,
     col = "black")

#3)corrplot
cor_factori <- cor(scoruri)

library(corrplot)
corrplot(cor_factori,
        method = "number",
        type = "upper",
        main = "Corelațiile dintre factori")

```



```
#-----ANALIZA CORESPONDENTELOR
```

```
date_ac<-Date_AC_tema
date2=date_ac[,-1]
date2=as.table(as.matrix(date2))
rownames(date2)=date_ac$Populatie_Categorii_varsta
View(date2)
install.packages("gplots")
library(gplots)
balloonplot(t(date2),main="Matricea de contingenta",)

X2=chisq.test(date2)
X2

a11=date2[1,1]
a11
row_totals=rowSums(date2) #totaluri pe linii
col_totals=colSums(date2) #totaluri pe coloane
e11=(row_totals[1]*col_totals[1])/sum(date2)
e11
X2$expected
X2$observed
install.packages("FactoMineR")
library(FactoMineR)
install.packages("factoextra")
library(factoextra)
rez=CA(date2, graph=F)
rez
eig=get_eigenvalue(rez)
eig
#inertia totala
s=sum(eig[,1])
s
summary(rez, nb.dec=2)
rez$col
fviz_screplot(rez)
rezultate_linii <- get_ca_row(rez)
print(rezultate_linii$coord)
```

```

print(rezultate_linii$cos2)
print(rezultate_linii$contrib)
matrice_factori_rânduri <- rez$row$coord
print(matrice_factori_rânduri)
fviz_ca_row(rez, repel = TRUE)

rezultate_coloane <- get_ca_col(rez)
print(rezultate_coloane$coord)
print(rezultate_coloane$cos2)
print(rezultate_coloane$contrib)
matrice_factori_coloane <- rez$col$coord
print(matrice_factori_coloane)
fviz_ca_col(rez, repel = TRUE)

#a) Biplot
fviz_ca_biplot(rez, map = "row principal", arrow = c(T,T), repel = TRUE)

#b) Grafic al contribuțiilor (corrplot)
install.packages("corrplot")
library(corrplot)
corrplot(rez$col$contrib, is.corr = FALSE, main = "Contribuția coloanelor")
corrplot(rez$row$contrib, is.corr = FALSE, main = "Contribuția liniilor")

# c) Calitatea reprezentării ( $\cos^2$ ) pentru linii și coloane
fviz_cos2(rez, choice = "row", axes = 1:2)
fviz_cos2(rez, choice = "col", axes = 1:2)

```