Logo

Description automatically generated

ACADEMIA DE STUDII ECONOMICE BUCUREȘTI

FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ

SPECIALIZAREA INFORMATICĂ ECONOMICĂ

*Dezvoltare Software pentru Analiza Datelor*

**Utilitatea analizei factoriale în interpretarea calității vinului**

**Profesor coordonator:**

Prof. Univ. Dr. Furtuna Titus Felix

**Student**

Coman Claudia Ana-Maria

Grupa 1082 C

București

2022

**Cuprins**

[**Date utilizate** 3](#_Toc124769437)

[**Metoda de analiză** 4](#_Toc124769438)

[**Abordarea privind analiza datelor** 4](#_Toc124769439)

[**Tipuri ale analizei factoriale** 6](#_Toc124769440)

[**Motivația de a opta pentru o anumită metodă de analiză a datelor** 6](#_Toc124769441)

[**Rezultate obținute** 8](#_Toc124769442)

[**Determinarea numărului de factori semnificativi** 10](#_Toc124769443)

[**Lista figuri** 18](#_Toc124769444)

[**Concluzii** 19](#_Toc124769445)

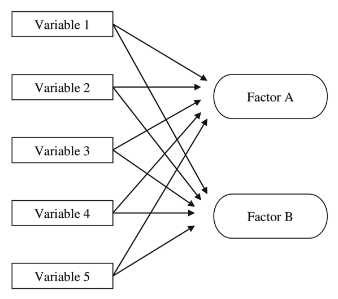
# **Date utilizate**

Pentru acest proiect am ales să utilizez datele regăsite la adresa <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>. Datele sunt preluate din anul 2009 și sunt  sunt legate de variantele roșii și albe ale vinului portughez „Vinho Verde”.

Acest set de date conține următoarele variabile pe baza testelor fizico-chimice: aciditate fixă (majoritatea acizilor implicați în vin sau fix sau nevolatil (nu se evaporă ușor), aciditate volatilă (cantitatea de acid acetic din vin, care la niveluri prea mari poate duce la un gust neplăcut, de oțet), acid citric (găsit în cantități mici, acidul citric poate adăuga „prospețime” și aromă vinurilor), zahăr residual (cantitatea de zahăr rămasă după oprirea fermentației, rar se găsesc vinuri cu mai puțin de 1 gram/litru), cloruri (cantitatea de sare din vin), dioxid de sulf liber (forma liberă a SO2 există în echilibru între SO2 molecular (ca gaz dizolvat) și ionul bisulfit; previne creșterea), dioxid de sulf total (cantitatea de forme libere și legate de S02; în concentrații scăzute, SO2 este în mare parte nedetectabil în vin), densitate (densitatea apei este apropiată de cea a apei în funcție de procentul de alcool și zahăr), ph (descrie cât de acid sau bazic este un vin pe o scară de la 0 (foarte acid) la 14 (foarte bazic), sulfați (un aditiv pentru vin care poate contribui la nivelurile de dioxid de sulf gazos (S02)).

# **Metoda de analiză**

# **Abordarea privind analiza datelor**

Analiza factorială este o metodă exploratorie de analiză, al cărei obiectiv este de a transforma tabele de date numerice, rezultate dintr-o analiză descriptivă a fenomenelor observate, ușor de interpretat prin grafice și care facilitează demersul explicativ al acestora. Atunci când se compară un grup mare de factori, complexitatea mare a situației poate face dificilă determinarea modului în care diferiți factori se corelează reciproc. În particular, această tehnică poate fi utilă pentru a evidenția grupe de itemi care se comportă similar. De asemenea, metoda facilitează observarea modului în care itemii individuali se corelează atât cu caracteristici majore comune, cât și unii cu alții. Aceasta extrage maximul varianței comune din toate variabilele și le pune laolaltă.

Figură 1 - Schema AF

Metoda aceasta se bazează pe ideea că există unii factori “latenți”, care descriu multiple variabile. Este în mod special utilă atunci când deținem nenumărate variabile și dorim să le reducem înainte de a ne construi modelul.

Analiza factorială este des utilizată în cercetările de marketing. publicitate, psihologie ( domeniu în care o utilizez și eu în această lucrare), finanțe și cercetări operaționale. Ea este un model statistic de analiză liniară, utilizat pentru a explica varianța dintre variabilele observate și de a condensa acest set de variabile observate într-unul de variabile neobservate, denumit factori, prin modelarea lor liniară. Factorii, sau variabilele latente, sunt asocieri a mai multor variabile observate ce dețin trăsături comune, care permit de altfel categorisirea lor. De exemplu, mai multe întrebări particulare pot aparține aceluiași domeniu, ce caracterizează o anumită trăsătură personală. Fiecare factor explică o cantitate anume de varianță din variabilele observate, ajutând enorm la interpretarea rezultatelor. **Presupuneri:**

* Nu există valori aberante de tip outliers.
* Eșantionul trebuie să conțină mai multe variabile decât factori.
* Nu există multicoliniaritate perfectă.
* Nu există homoschedasticitate între variabile

# **Tipuri ale analizei factoriale**

*Analiza exploratorie a factorilor*: Această variantă este cea mai populară dintre cele existente în universul social și de management al cercetărilor. Presupunerea de bază utilizată este că oricare variabilă observată este asociată în mod direct cu oricare factor.

*Analiza confirmatorie a factorilor*: Presupunerea sa constă în faptul că fiecare factor este asociat cu un set particular de variabile observate, și confirmă așteptările analizei.

Precum am mai menționat și anterior, în cadrul acestei analize voi utiliza modelul de Analiză exploratorie a factorilor.

## **Motivația de a opta pentru o anumită metodă de analiză a datelor**

**De ce Analiza Factorială și nu Analiza în componente principale?**

Cele două metode se suprapun în terminologie și obiective, unele surse nefăcând distincția concretă dintre acestea, iar unii algoritmi de Analiză factorială chiar o includ și pe cea a componentelor principale. Ambele sunt metode de reducere a dimensiunii, putând fi utilizate pentru a înlocui un set de date extrem de întins cu unul mult mai restrâns. Cu toate acestea, cele două metode diferă în scopul lor final și în modelele implementate. Cu alte cuvinte, ACP este utilizat atunci când dorim doar să sumarizăm sau să aproximăm datele utilizând mai puține dimensiuni ( pentru o mai bună vizualizare a datelor, de exemplu), iar AEF este mult mai adecvată pentru obținerea unui model explicativ asupra corelării datelor.

**Diferențele dintre cele două metode**

Componentele ACP explică maximul varianței, în timp ce AF explică covarianța din cadrul datelor.

În cadrul ACP, componentele sunt ortogonale între ele, în timp ce în AF nu este necesar acest lucru.

În ACP, avem componente ca și combinații liniare are variabilelor observate, care sunt de asemenea neinterpretabile, iar în AF, variabilele observate sunt combinații liniare ale variabilelor neobservate ( ale factorilor ), care sunt interpretabile și etichetabile.

De asemenea, ACP este o metodă observațională de reducere a dimensionalității, iar AF este o tehnică de modelare.

Această metodă ne ajută să înțelegem relațiile dintre variabilele și să interpretăm mai bine datele. Cu toate acestea, rezultatele sunt uneori greu de determinat din cauza interpretărilor, care pot produce controverse.

# **Rezultate obținute**

Ca prim pas în analiza factorială am importat librăriile necesare – cele din Python (numpy, pandas, factor\_analyzer și sklearn, dar și fișierele din seminar (clasele cu ACP și AF).

Ulterior, am pregătit datele – am citit din fișierul winequality-red.csv, am identificat preluat indexii, coloanele și valorile numerice, pentru o prelucrare mai ușoară. Am creat matricea scalată și modelul AEF, pentru a putea începe testul.

obsNume = tabel.index  
print("Observatii: ",obsNume)  
varNume = tabel.columns  
matrice\_numerica = tabel.values

**Testul de factorabilitate**

Înainte de a începe analiza factorială, trebuie să verificăm dacă datele deținute sunt potrivite pentru o astfel de analiză. Factorabilitatea înseamnă *Putem găsi factori în setul de date?* . Există două metode prin care putem afla acest fapt – Testul Bartlett și cel Kaiser-Meyer-Olkin.

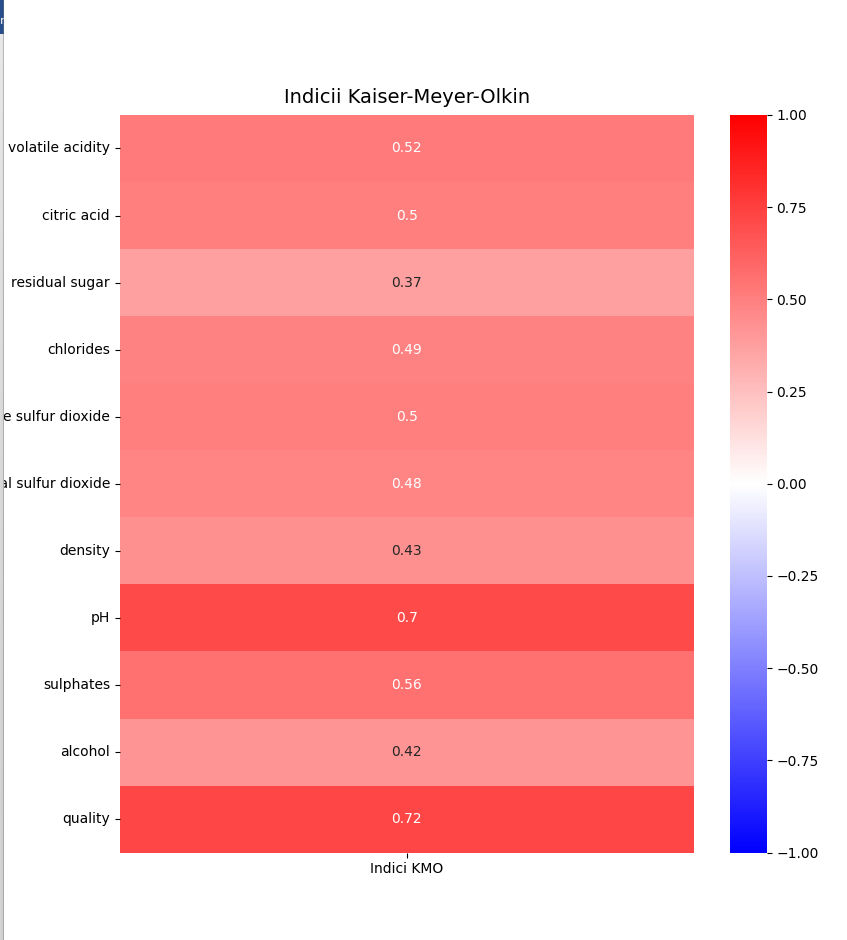
**Testul Bartlett** de sfericitate verifică dacă variabilele observate sunt corelate, utilizând matricea de corelație și matricea de identitate. Dacă testul este nesemnificativ, nu se poate continua cu analiza factorială.

sfericitateBartlett = fa.calculate\_bartlett\_sphericity(Xstd\_df)

Rezultatul obținut în urma acestui test este:

Astfel, în acest test p-value este 0, fiind semnificativ statistic și indicând faptul că matricea de corelație a variabilelor observate nu este matricea de identitate.

**Testul Kaiser-Meyer-Olkin** măsoară potrivirea datelor pentru analiza factorială, pentru fiecare variabilă observată și pentru modelul în ansamblu. KMO estimează varianța din cadrul acestor variabile – proporția variză între 0 și 1, iar o valoare mai mică de 0.6 înseamnă că setul de date nu este adecvat.



Figură 2 - Indicii Kaiser-Meyer-Olkin

Pentru fiecare factor, avem un indice KMO excelent, iar indicele KMO general este 0.5061 ceea ce înseamnă că setul de date este potrivit în analiza factorială.



# **Determinarea numărului de factori semnificativi**

In aceasta etapa, vom determina numarul de factori semnificativi, adica variabilele neobservate. Astfel, din cele 11 variabile observate, intentionez sa obtin un numar mult mai mic de factori care sa le grupeze pe acestea.

Pentru a realiza acest lucru, , am ales să creez inițial graficul analizei factoriale cu toate cele 11 observații.

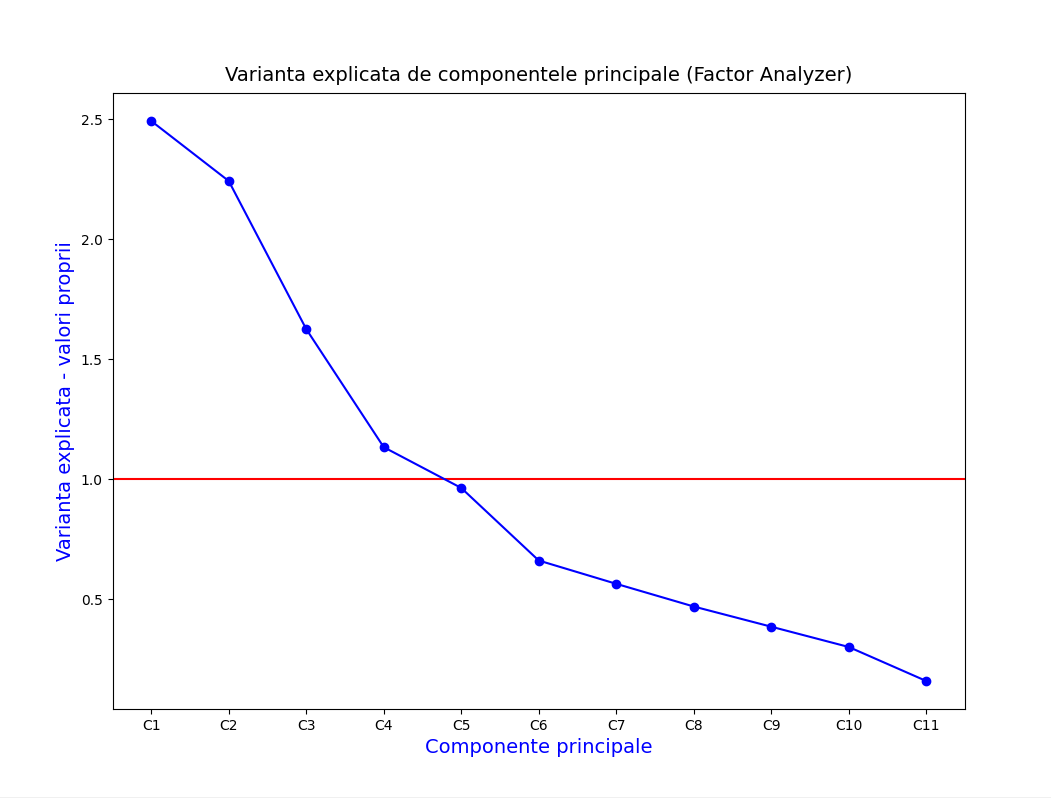
Apoi, am extras valorile proprii din factor analyzer, cu funcția corespunzătoare. Aceste valori arată ce cantitate din varianta totală a tuturor variabilelor observate este explicată de un factor.

valPropFA = faModelFit.get\_eigenvalues()



Pentru aceste valori, am creat graficul componentelor principale, pentru o vizualizare mai bună a datelor obținute.

g.componentePrincipale(valoriProprii=valPropFA[0], titlu='Varianta explicata de componentele principale (Factor Analyzer)')



Figură 3 - Varianta explicata de componentele principale

Aceste rezultate le voi utiliza pentru a determina câți factori să aleg pentru variabilele neobservate. Pentru numărul acesta de factori semnificativi, calculez numărul de valori proprii peste 1 – ( după cum se poate observa și din grafic )

count = 0  
for i in valPropFA[0] :  
 if i > 1 : count+=1  
numarFactoriSemnificativi = count

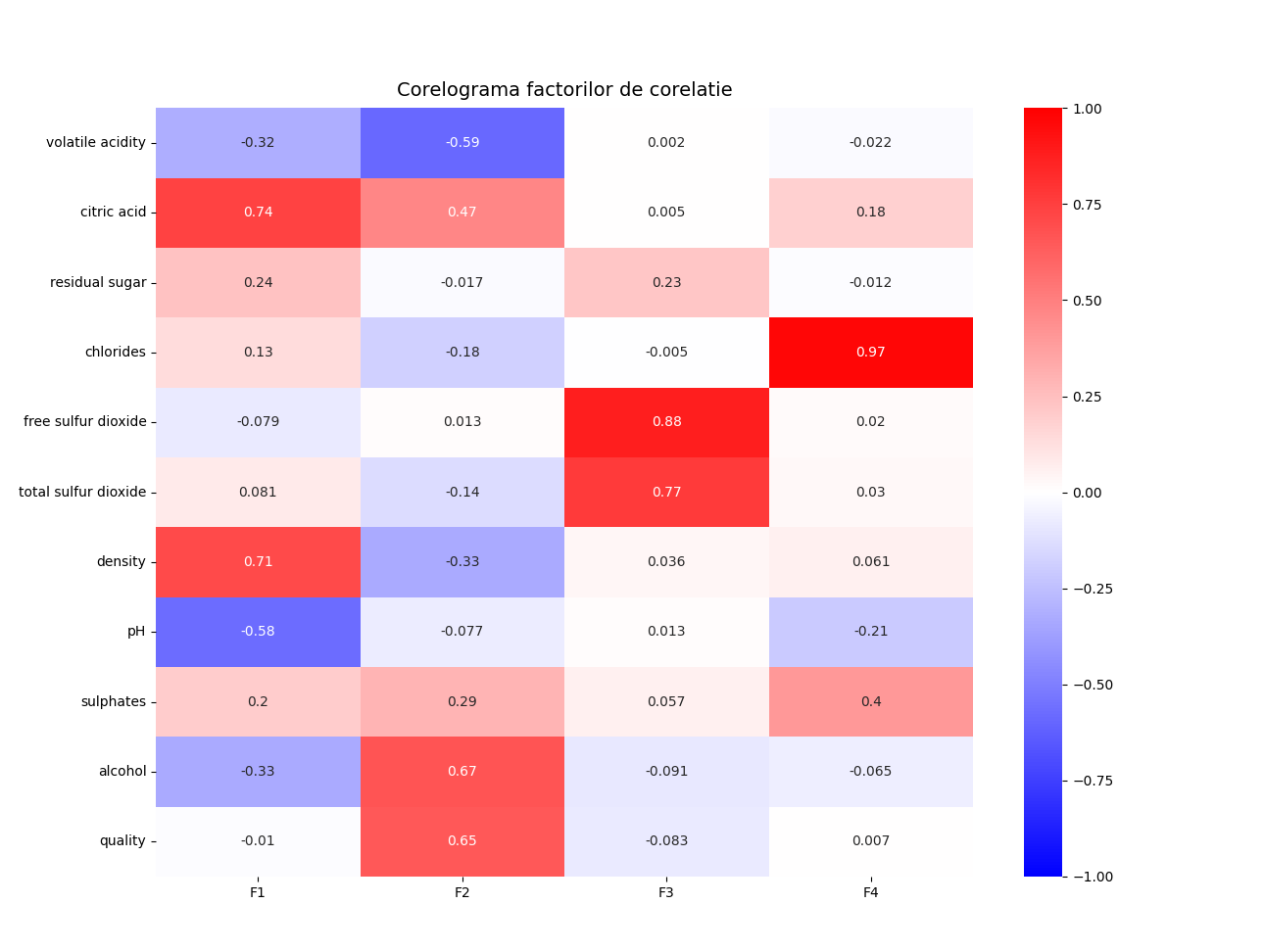
Atât graficul, cât și variabila count a determinat 4 factori semnificativi pentru modelul acesta.

Urmează etapa în care determin corelograma factorilor de corelație, după ce calculez loadings\_ - matricea care arată relația dintre fiecare variabilă și factori, adică un coeficient de corelație între variabila observată și unul din factorii găsiți ( cei 4).

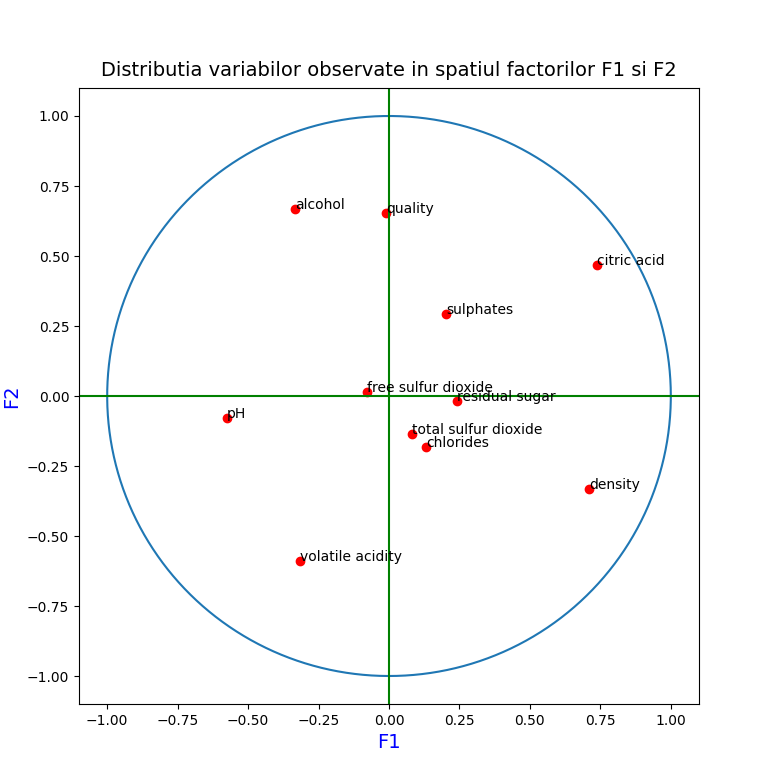
faModelFit = fa.FactorAnalyzer(n\_factors=numarFactoriSemnificativi,rotation='varimax')  
faModelFit.fit(Xstd)  
loadings = faModelFit.loadings\_

Prin rotația factorilor se încearcă obținerea unei matrice de coeficienți factoriali, care să uşureze cât mai mult interpretarea factorilor. Sistemul de axe ortogonale reprezentat de factori este rotit în jurul originii într-o altă poziție. Una din metodele cele mai utilizate de rotaţie a factorilor este metoda Varimax, unde rotația fiecarei variabile inițiale ( observată) tinde să fie asociată cu un singur factor şi fiecare factor cu un număr redus de variabile inițiale. Formal, Varimax caută prin rotații succesive ale sistemului de axe reprezentat de factori, să maximizeze varianta totală explicată de factori.

Graficul de mai jos conţine saturatiile (factor loadings) factorilor extraşi - un coeficient de saturație este echivalent cu corelaţia dintre variabila observată și factorul extras atunci când factorii extrași sunt ortogonali. Această reprezentare arată ce variabile se încarcă pe fiecare factor și rezultatele sunt folosite pentru etichetarea și interpretarea lor.



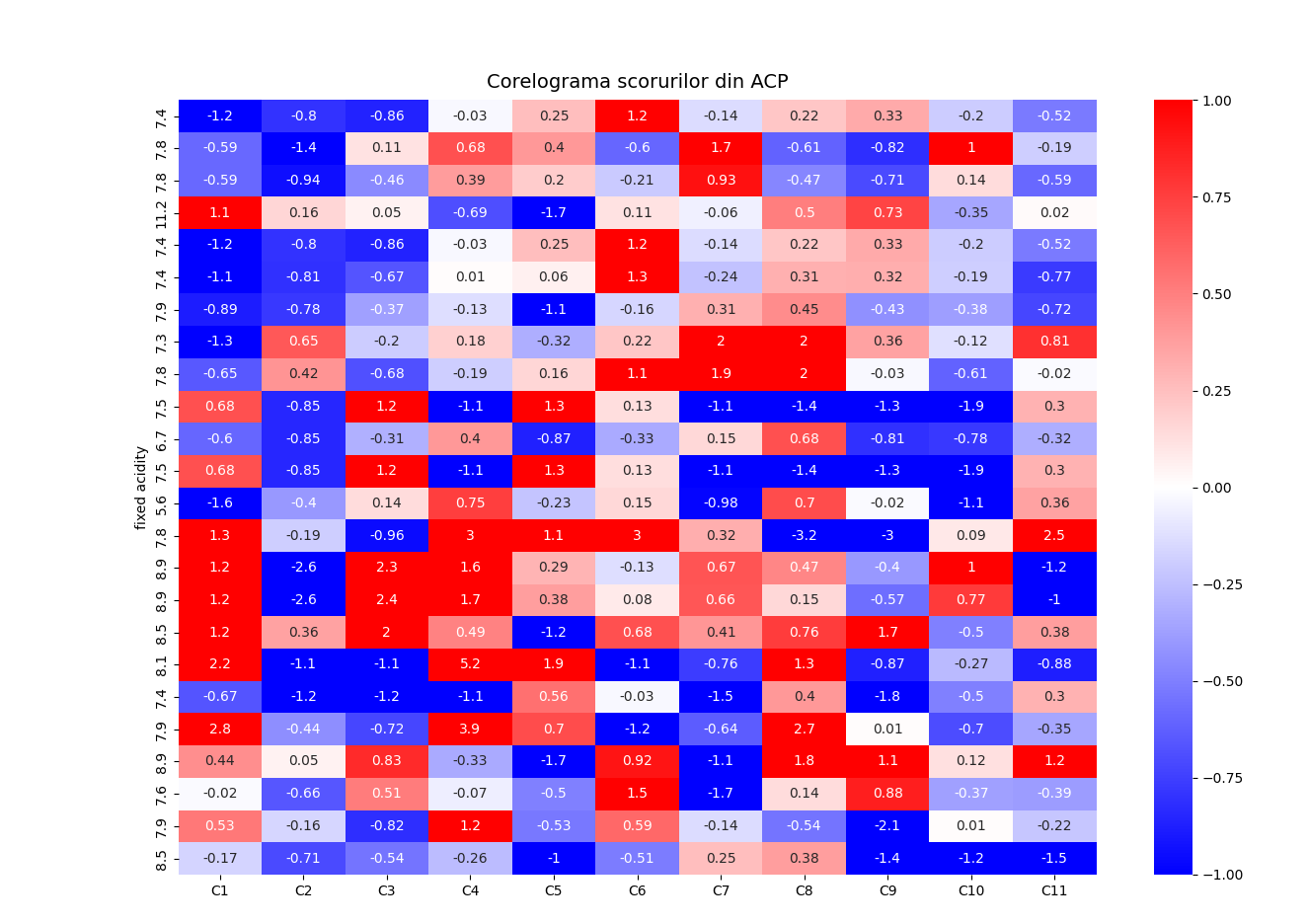
Figură 4 - Corelograma factorilor de corelatie



Figură 5 - Cercul variabilelor observate in spatiul factorilor F1 si F2

În grafic este reprezentată distribuția variabilelor obsrvate ( toate cele 11) în spațiul factorilor F1 și F2, axele fiind oX pentru F1 și oY pentru F2. Astfel, cele două coordonate ale fiecărei variabile reprezintă coeficientul de corelație corespunzător uneia dintre cele două variabile neobservate (calculate, factor).Scorurile sunt standardizări ale componentelor principale.

Mai jos este reprezentată matricea de corelaţie dintre scoruri şi coeficienţii cu care se înmulţesc variabilele pentru a obţine scorurile factoriale. Practic, este reprezentată legătura dintre fiecare observație cu fiecare componentă principală.



Figură 6 - Corelograma scorurilor din ACP

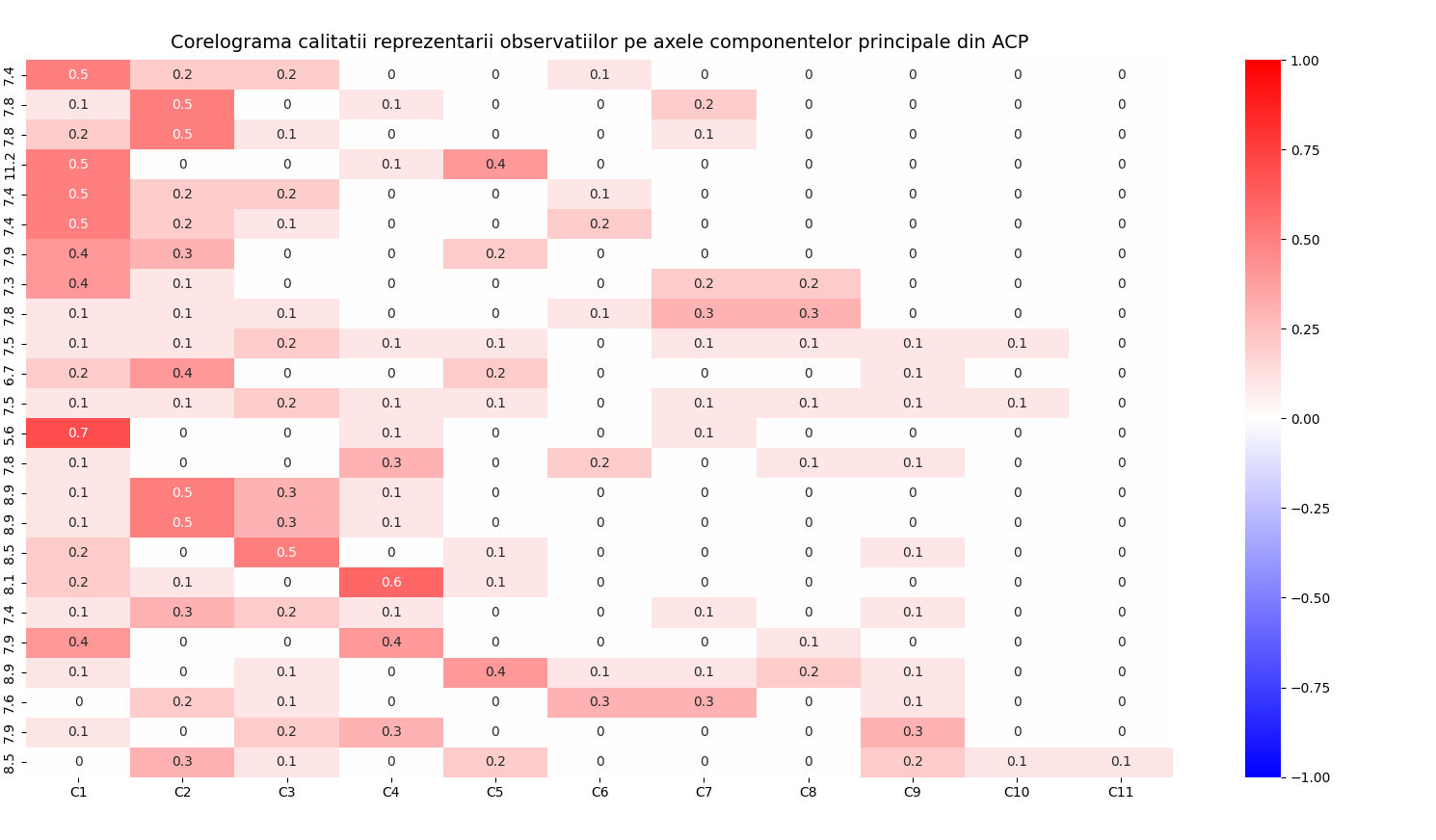
Ulterior, am reprezentat și scorurile factoriale rezultate.

În corelograma calității reprezentată mai jos, putem observa fiecare observație ( am ales primele 24) pe axele componentelor principale din ACP. Acest grafic ne arata cat de bine sunt reprezentate toate calitatile vinului.

Cu cât ne îndepărtăm de pe axa componentelor, coeficientul scade.

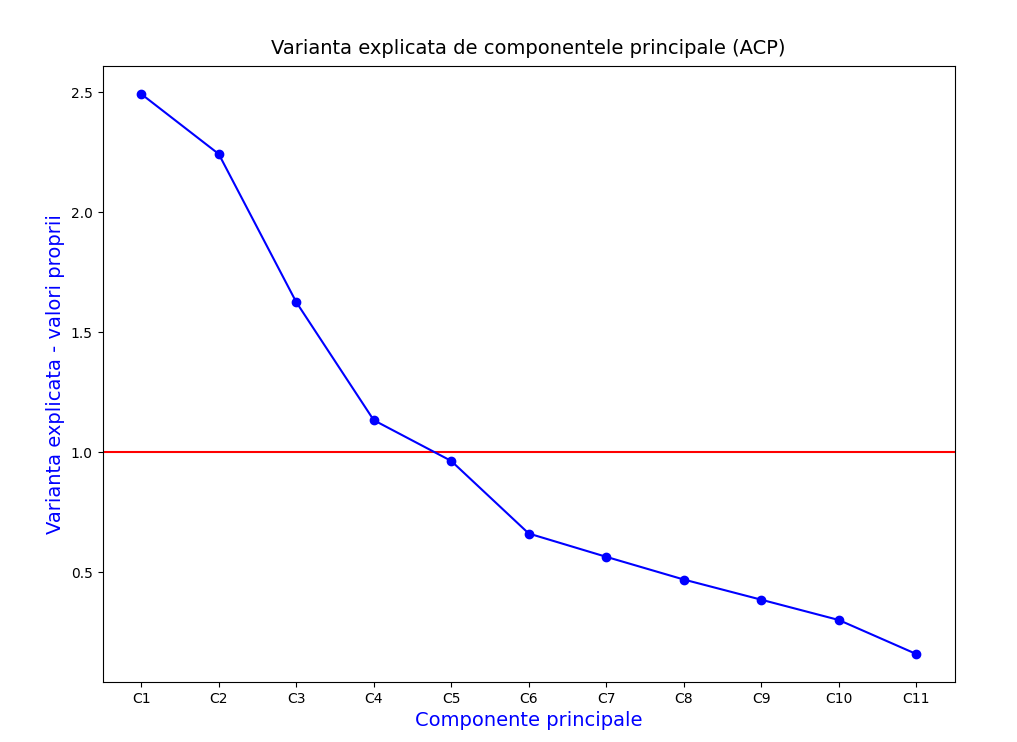


Figură 7 - Corelograma scorurilor din FA



Figură 8 - Corelograma calitatii reprezentarii observatiilor

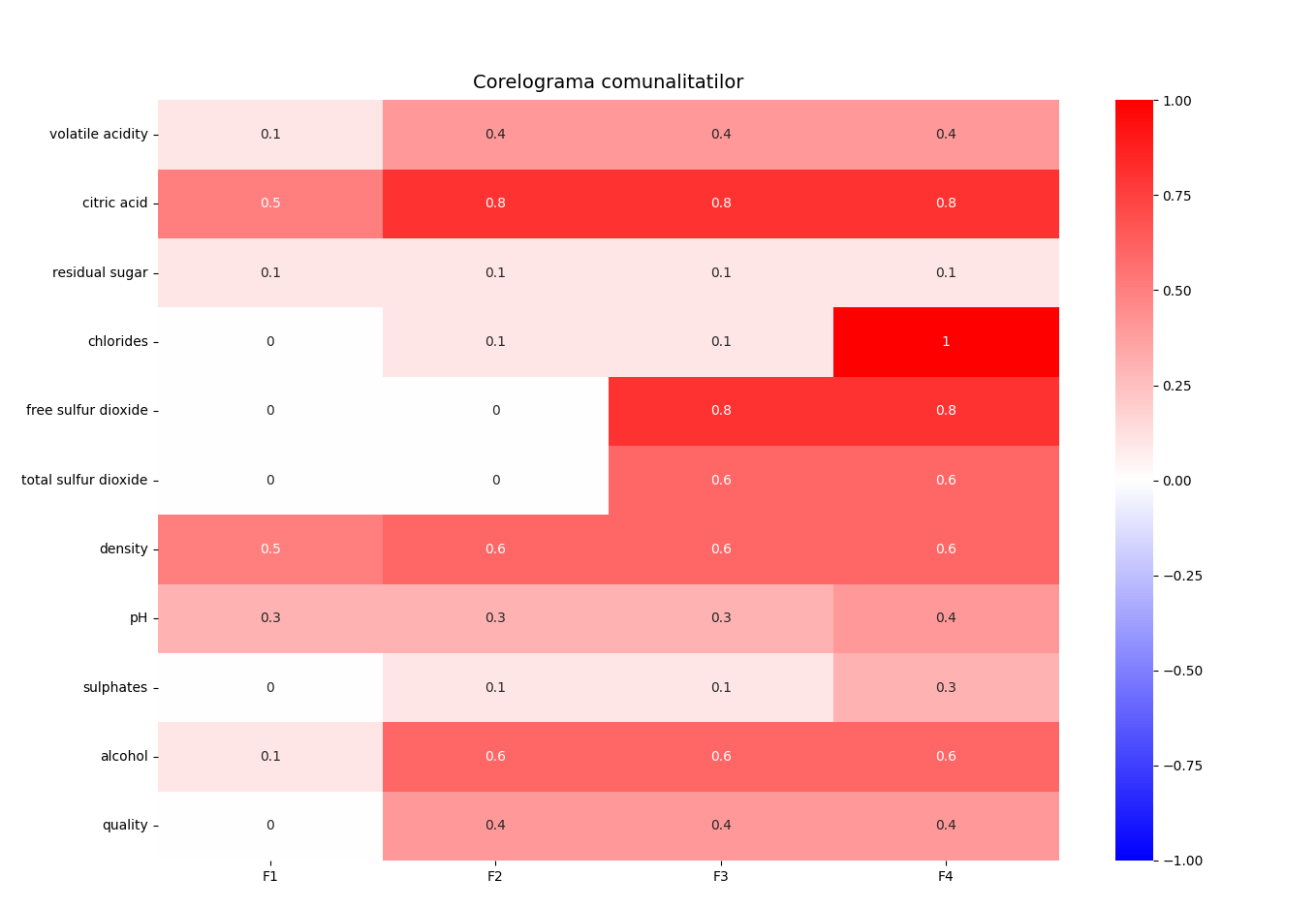
Această varianță reprezintă valorile proprii extrase din modelul ACP, fiind asemănătoare cu cele din modelul AEF, fiind tot în numar de 6 cele peste 1.



Figură 9 - Varianta explicata de componente

Communalities (comunalități) sunt similare cu R2 din modelul de regresie și arată proporția din varianta unei variabile observate care este explicată de toţi factorii extrasi, fiind explicată în comun de către un grup de componentele principale și se calculează la nivelul fiecărei variabile observate. Comunalitatea calculată pentru o variabilă și componente principale este suma coeficienților de determinare dintre variabilă și aceste componente principale. Aceasta manifestă procentul varianței unei variabile explicată de factorii reuniți, și poate fi interpretată ca siguranța indicatorului reprezentat de acea variabilă.

Valori foarte reduse ale comunalităţii indică faptul că acele variabile nu sunt bine reprezentate de respectivul model factorial. Din graficul de mai jos, se observă cu claritate că unele întrebări, precum cele prefixate cu N, prezintă o explicabilitate majoră, spre deosebire de cele cu O.



Figură 10 - Corelograma comunalitatilor

# **Lista figuri**

[Figură 1 - Schema AF 4](file:///C:\Users\win\Documents\proiect%202%20(Recuperat%20automat).docx#_Toc124769120)

[Figură 2 - Indicii Kaiser-Meyer-Olkin 7](#_Toc124769121)

[Figură 3 - Varianta explicata de componentele principale 8](#_Toc124769122)

[Figură 4 - Corelograma factorilor de corelatie 10](#_Toc124769123)

[Figură 5 - Cercul variabilelor observate in spatiul factorilor F1 si F2 10](#_Toc124769124)

[Figură 6 - Corelograma scorurilor din ACP 11](#_Toc124769125)

[Figură 7 - Corelograma scorurilor din FA 12](#_Toc124769126)

[Figură 8 - Corelograma calitatii reprezentarii observatiilor 12](#_Toc124769127)

[Figură 9 - Varianta explicata de componente 13](#_Toc124769128)

[Figură 10 - Corelograma comunalitatilor 14](#_Toc124769129)

# **Concluzii**

Analiza factorială explorează seturi mari de date ( cum a fost și acesta și găsește legături și asocieri între acestea, reducând variabilele observate numeroase în doar câteva, mult mai reduse la număr, grupate pe categorii. Aceste noi variabile, denumire factori sau variabile neobservate, concentrează mai multe părți componente după o trăsătură comună.

Cu toate acestea, rezultatele analizei factoriale pot fi controversate, interpretabile, depinzând de persoana care le evaluează. Prin analiza factorială, am descoperit o grupare pe observații neobservate (factori), am dedus asocierile și corelațiile dintre toate observațiile deținute.