**ACADEMIA DE STUDII ECONOMICE DIN BUCUREȘTI**

**FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ**

**SPECIALIZAREA : CIBERNETICĂ ECONOMICĂ**

****

**PROIECT ANALIZA DATELOR**

Coordonator: Student:

Lector.univ.dr. Ionela Cătălina ZAMFIR Alexandru ARMANU

**Cuprins**

[**1.** **Introducere** 3](#_Toc123848863)

[**2.** **Analiza descriptiva** 4](#_Toc123848864)

[**3.** **Boxploturi si grafice** 9](#_Toc123848865)

[**4.** **Matricea de corelatie** 12](#_Toc123848866)

[**5.** **Analiza factoriala** 13](#_Toc123848867)

[**Indicele KMO** 14](#_Toc123848868)

[**Testul Bartlett** 14](#_Toc123848869)

[**6.** **Anexe** 18](#_Toc123848870)

# **Introducere**

Analiza datelor se referă la procesul de transformare a datelor inițiale în informațiirelevante pentru domeniul de interes analizat. Aceasta se aplică într-o gamă variată de domenii,de la cel financiar, la inginerie și chiar medicină.

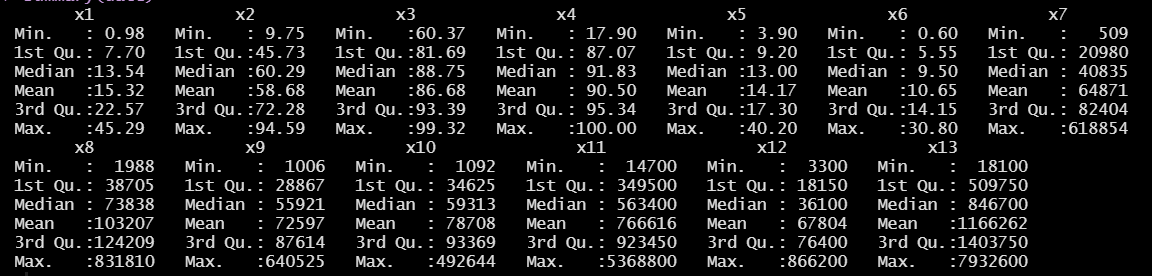
În prezent, sunt cunoscute numeroase metode de analiză ce se aplică pe un volum din ce în ce mai mare de date. Tehnicile oferite de analiza datelor sunt unele necesare pentru sintetizarea informației care devine utilă și relevantă doar atunci când este prezentată într-un mod cât mai succint și specific domeniului la care face referire. ”Metodele și tehnicile reprezintă un set de reguli, principii și proceduri de analiză, prelucrare și interpretare a datelor. În analizadatelor, metodele și tehnicile se referă la cuantificare, evaluare, estimare și testare, și suntreprezentate de o mulțime extinsă și variată de proceduri și instrumente statistico-matematice.” (Ruxanda, 2013)

Unul dintre scopurile proiectului este de a observa in ne masura afecteaza lipsa internetului participare ala educatie, astel am extras de pe EUROSTAT date care relateaza urmatoarele:

* x1 rata persoanelor care nu au folosit un calculator
* x2 procentul indivizilor care au folosit internetul pentru a contacta autoritatile
* x3 procentul persoanelor care au folosit internetul
* x4 procentul gospodariilor cu internet
* x5 rata tinerilor care nu studiaza nici nu lucreaza 15-29
* x6 rata de participare la educatie
* x7 studenti inscrisi la studii universitare
* x8 elevi inscrisi in invatamantul primar
* x9 elevi inscrisi In invatamantul gimnazial
* x10 elevi inscrisi in invatamantul liceal
* x11 angajati de la 15 la 64 ani
* x12 someri de la 15 la 74 ani
* x13 populatia totala (15-64 ani)

1. **Analiza descriptiva**

Pentru aceste date am decis sa efectuez o analiza descriptiva :



**X1-rata peroanelor care nu au folosit un calculator**

* rata medie a persoanelor care nu au folosit un calculator este de 15.32, rata minima a fost

de 0.98, la o diferenta de 14.34, iar cea maxima de 45.29.

* diferenta dintre medie si mediana este de 1.78, destul de mare=>date imprastiate
* observam ca prima quartila se inregistreaza la 7.70(la o diferenta de 6.72 fata de rata cea mai mica=outlieri)
* 25% dintre regiuni au rata persoanelor care nu au folosit internetul sub 7.70
* 50% diintre regiuni au rata persoanelor care nu au folosit internetul peste 13.54
* 25% dintre regiuni au rata persoanelor care nu au folosit internetul peste 22.57

**X2-procentul indivizilor care au folosit internetul pentru a contacta autoritatile**

* procentul mediu al indivizilor care au folosit internetul pentru a contacta autoritatile este de 58.68%, procentul minim a fost de 9.75, iar cel maxim de 94.59
* diferenta dintre medie si mediana este de 1.6, relativ mare=>datele sunt usor imprastiate
* observam ca prima quartila se inregistreaza la 45.73(la o diferenta de 35.98 fata de procentul minim = outlieri)
* 25% dintre regiuni au procentul indivizilor care au folosit internetul pentru a contacta autoritatile sub 45.73
* 50% diintre regiuni au procentul indivizilor care au folosit internetul pentru a contacta autoritatile peste 60.29
* 25% dintre regiuni au procentul indivizilor care au folosit internetul pentru a contacta autoritatile peste 72.28

**X3-procentul persoanelor care au folosit internetul**

* procentul mediu al persoanelor care au folosit internetul este de 86.68%, procentul minim a fost de 60.37, iar cel maxim de 99.32
* diferenta dintre medie si mediana este de 2.07, relativ mica => datele sunt usor imprastiate
* observam ca prima quartila se inregistreaza la 81.69(la o diferenta de 21.32 fata de procentul minim = outlieri)
* 25% dintre regiuni au procentul persoanelor care au folosit internetul sub 81.69
* 50% diintre regiuni au procentul persoanelor care au folosit internetul peste 88.75
* 25% dintre regiuni au procentul persoanelor care au folosit internetul peste 93.39

**X4-procentul gospodariilor cu internet**

* procentul mediu al gospodariilor cu internet este de 90.50%, procentul minim a fost de 17.90, iar cel maxim de 100%
* diferenta dintre medie si mediana este de 1.33, relativ mica => datele sunt usor imprastiate
* observam ca prima quartila se inregistreaza la 87.07(la o diferenta de 69.17 fata de procentul minim = outlieri)
* 25% dintre regiuni au procentul gospodariilor cu internet sub 87.07
* 50% diintre regiuni au procentul gospodariilor cu internet peste 91.83
* 25% dintre regiuni au procentul gospodariilor cu internet peste 95.34

**X5-rata tinerilor care nu studiaza nici nu lucreaza**

* rata medie a tinerilor care nu studiaza nici nu lucreaza este de 14.17, rata minima a fost de 3.9, iar cea maxima de 40.20
* diferenta dintre medie si mediana este de 1.17, relativ mica =>datele nu sunt imprastiate
* observam ca prima quartila se inregistreaza la 9.20(la o diferenta de 5.30 fata de rata cea mai mica=outlieri)
* 25% dintre regiuni au rata tinerilor care nu studiaza nici nu lucreaza sub 9.20
* 50% diintre regiuni au rata tinerilor care nu studiaza nici nu lucreaza peste 13.00
* 25% dintre regiuni au rata tinerilor care nu lucreaza nici nu studiaza peste 17.30

**X6-rata de participare la educatie**

* rata medie de participarela educatie este de 10.65, rata minima a fost de 0.60
* iar cea maxima de 30.80
* diferenta dintre medie si mediana este de 1.15, relativ mica =>datele nu sunt imprastiate
* observam ca prima quartila se inregistreaza la 5.55(la o diferenta de 4.95 fata de rata cea mai mica=outlieri)
* 25% dintre regiuni au rata de participare la educatie sub 5.55
* 50% diintre regiuni au rata de participare la educatie peste 9.50
* 25% dintre regiuni au rata de participare la educatie peste 14.15

**X7-studenti inscrisi la studii univerisitare**

* numarul mediu de studenti inscrisi la studii universitare este de 64871, numarul minim este de 509, iar maximul de 618854
* diferenta dintre medie si mediana este de 4036, mare =>datele sunt imprastiate
* observam ca prima quartila se inregistreaza la 20980(la o diferenta de 20471 fata de valoarea cea mai mica=outlieri)
* 25% dintre regiuni au numarul de studenti inscrisi sub 20980
* 50% diintre regiuni au numarul de studenti inscrisi peste 40835
* 25% dintre regiuni au numarul de studenti inscrisi peste 82404

**X8-elevi inscrisi in invatamant primar**

* numarul mediu de elevi inscrisi in invatamantul primar este de 103207, numarul minim este de 1988, iar maximul de 831810
* diferenta dintre medie si mediana este de 29369, mare =>datele sunt imprastiate
* observam ca prima quartila se inregistreaza la 38705(la o diferenta de 36717 fata de valoarea cea mai mica=outlieri)
* 25% dintre regiuni au numarul de elevi inscrisi in invatamantul primar sub 38705
* 50% diintre regiuni au numarul de elevi inscrisi in invatamantul primar peste 73838
* 25% dintre regiuni au numarul de elevi inscrisi in invatamantul primar peste 124209

**X9-elevi inscrisi in invatamant gimnazial**

* numarul mediu de elevi inscrisi in invatamantul gimnazial este de 72597, numarul minim este de 1006, iar maximul de 640525
* diferenta dintre medie si mediana este de 16676, mare =>datele sunt imprastiate
* observam ca prima quartila se inregistreaza la 28867(la o diferenta de 27861 fata de valoarea cea mai mica=outlieri)
* 25% dintre regiuni au numarul de elevi inscrisi in invatamantul gimnazial sub 28867
* 50% diintre regiuni au numarul de elevi inscrisi in invatamantul gimnazial peste 55921
* 25% dintre regiuni au numarul de elevi inscrisi in invatamantul gimnazial peste 87614

**X10-elevi inscrisi in invatamant liceal**

* numarul mediu de elevi inscrisi in invatamantul liceal este de 78708, numarul minim este de 1092, iar maximul de 462644
* diferenta dintre medie si mediana este de 19395, mare =>datele sunt imprastiate
* observam ca prima quartila se inregistreaza la 34625(la o diferenta de 33733 fata de valoarea cea mai mica=outlieri)
* 25% dintre regiuni au numarul de elevi inscrisi in invatamantul liceal sub 34625
* 50% diintre regiuni au numarul de elevi inscrisi in invatamantul liceal peste 59313
* 25% dintre regiuni au numarul de elevi inscrisi in invatamantul liceal peste 93369

**X11-angajati de la 15-64 ani**

* numarul mediu de angajati este de 766616, numarul minim este de 14700, iar maximul de 5368800
* diferenta dintre medie si mediana este de 203216, mare =>datele sunt imprastiate
* observam ca prima quartila se inregistreaza la 349500(la o diferenta de 324800 fata de valoarea cea mai mica=outlieri)
* 25% dintre regiuni au numarul angajatilor sub 349500
* 50% diintre regiuni au numarul angajatilor peste 563400
* 25% dintre regiuni au numarul angajatilor peste 923450

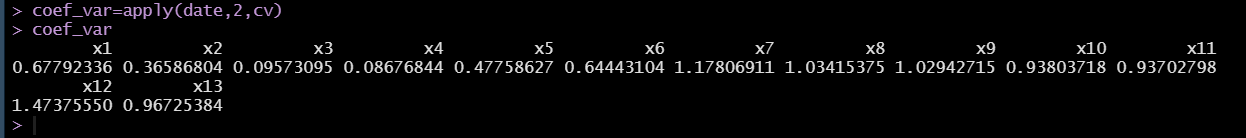
**X12-someri de la 15-74 ani**

* numarul mediu de someri este de 67804, numarul minim este de 3300, iar maximul de 866200
* diferenta dintre medie si mediana este de 31704, mare =>datele sunt imprastiate
* observam ca prima quartila se inregistreaza la 18150(la o diferenta de 14850 fata de valoarea cea mai mica=outlieri)
* 25% dintre regiuni au numarul somerilor sub 18150
* 50% diintre regiuni au numarul somerilor peste 26100
* 25% dintre regiuni au numarul somerilor peste 76400

**X13-populatia totala 15-64 ani**

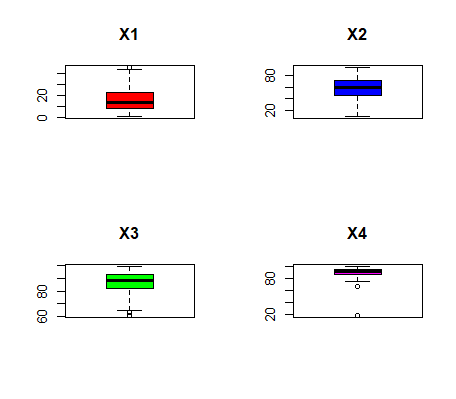
* numarul mediu al populatiei totale este de 1166262, numarul minim este de 18100, iar maximul de 7932600
* diferenta dintre medie si mediana este de 319562, mare =>datele sunt imprastiate
* observam ca prima quartila se inregistreaza la 509750(la o diferenta de 461650 fata de valoarea cea mai mica=outlieri)
* 25% dintre regiuni au numarul populatiei totale sub 509750
* 50% diintre regiuni au numarul somerilor peste 846700
* 25% dintre regiuni au numarul somerilor peste 1403750

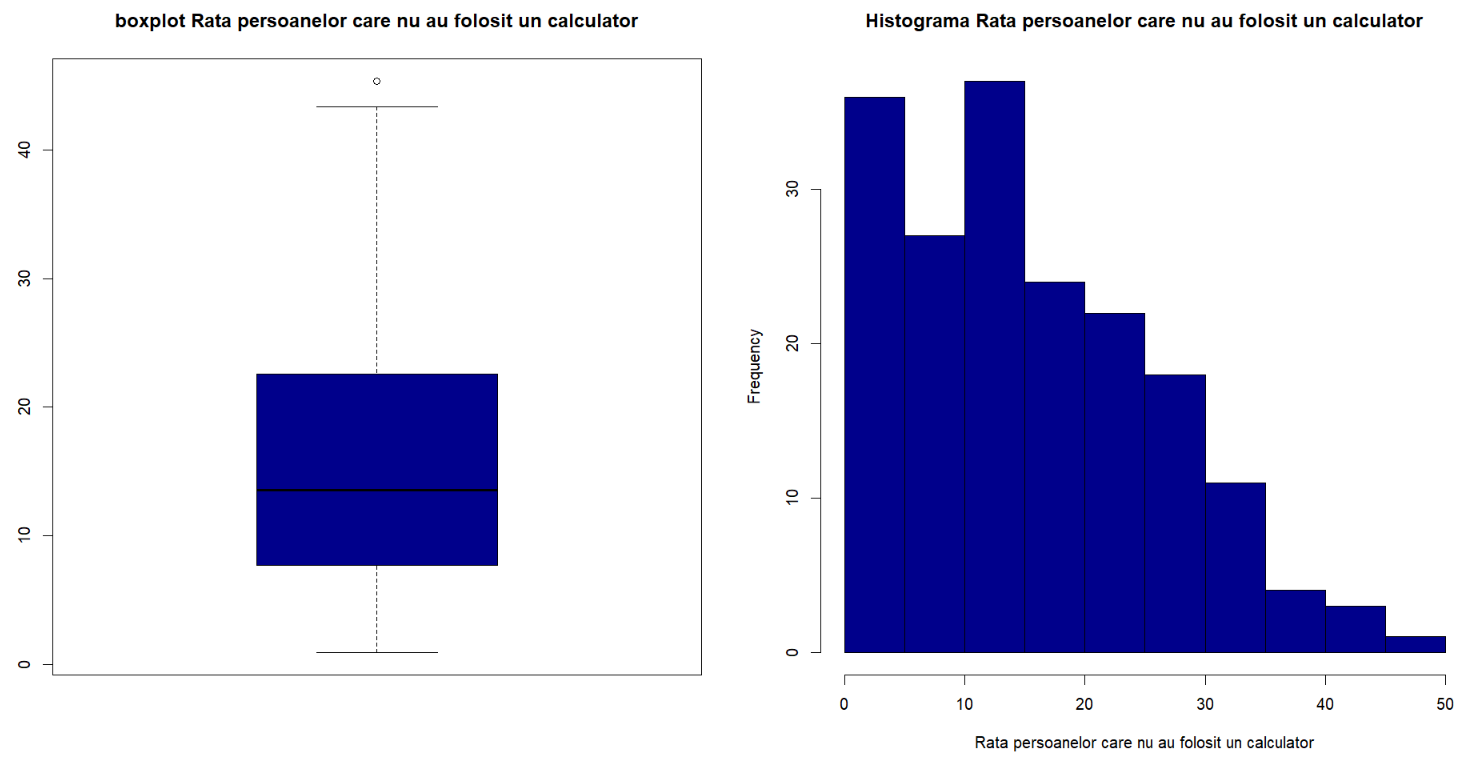
**Coeficientul de variatie**

****

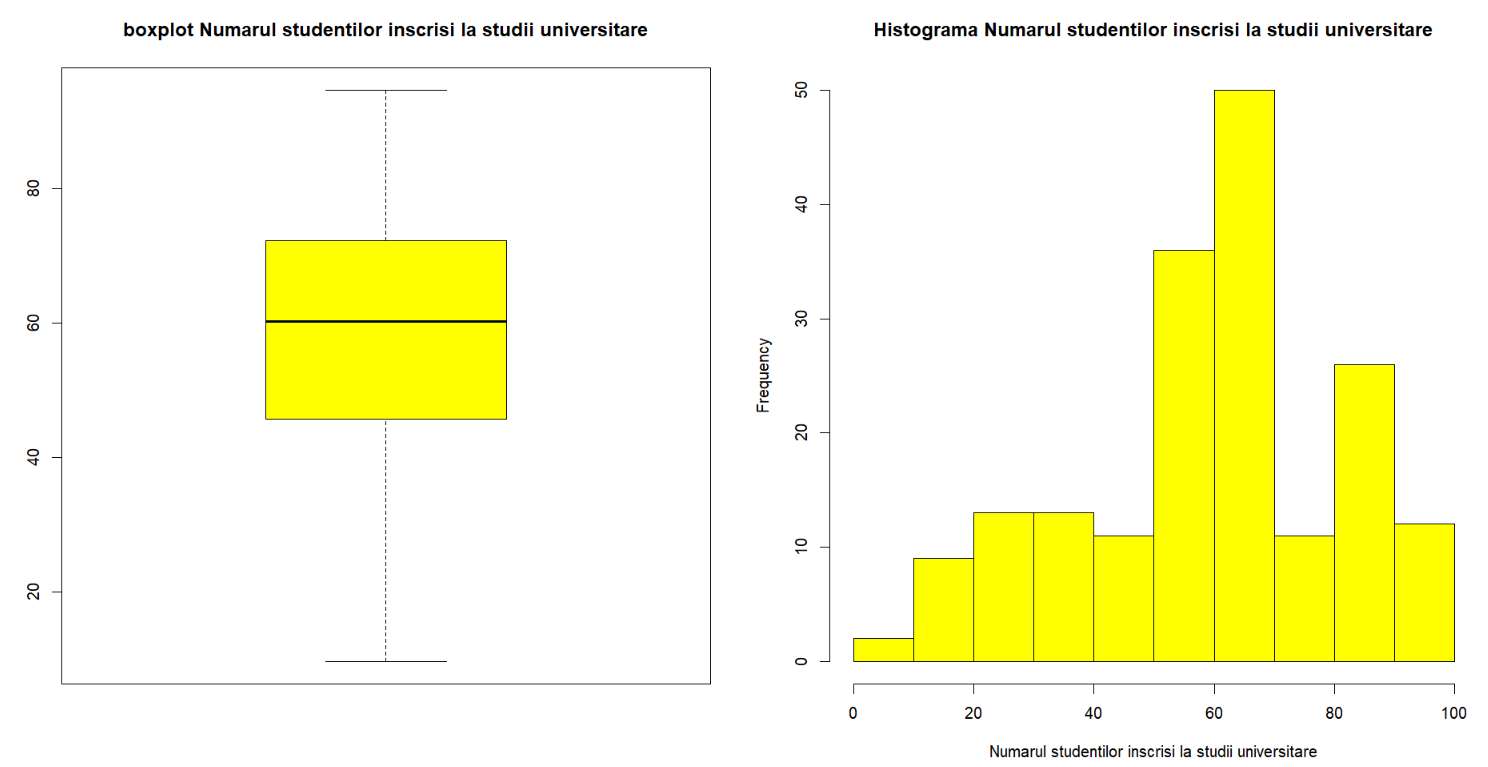
* coeficientul de variatie pentru x1 este 0.67>35%=> serie omogena si medie reprezentativa
* coeficientul de variatie pentru x2 este 0.36>35%=> serie omogena si medie reprezentativa
* coeficientul de variatie pentru x3 este 0.09<35%=> serie neomogena si medie nereprezentativa
* coeficientul de variatie pentru x4 este 0.086<35%=> serie neomogena si medie nereprezentativa
* coeficientul de variatie pentru x5 este 0.47>35%=> serie omogena si medie reprezentativa
* coeficientul de variatie pentru x6 este 0.64>35%=> serie omogena si medie reprezentativa
* coeficientul de variatie pentru x7 este 1.17>35%=> serie omogena si medie reprezentativa
* coeficientul de variatie pentru x8 este 1.03>35%=> serie omogena si medie reprezentativa
* coeficientul de variatie pentru x9 este 1.029>35%=> serie omogena si medie reprezentativa
* coeficientul de variatie pentru x10 este 0.938>35%=> serie omogena si medie reprezentativa
* coeficientul de variatie penttu x11 este 0.937>35%=> serie omogena si medie reprezentativa
* coeficientul de variatie pentru x12 este 1.47>35%=> serie omogena si medie reprezentativa
* coeficientul de variatie pentru x13 este 0.967>35%=> serie omogena si medie reprezentativa

1. **Boxploturi si grafice**

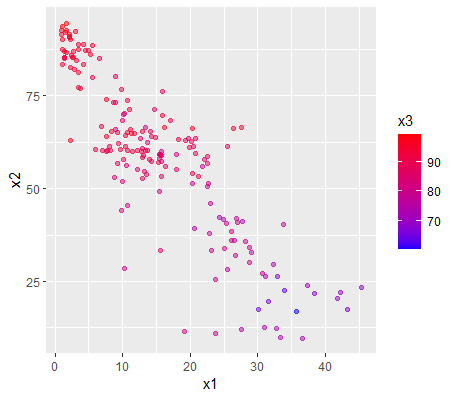
****

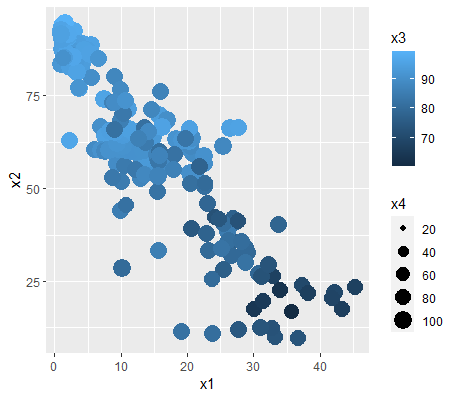
****

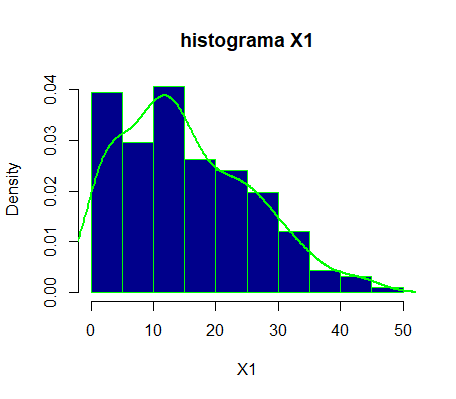
Remarcam prezenta outlierilor pe rate mari ale participarii la educatie



Nu exista outlieri pe numarul studentilor inscrisi la studii universitare

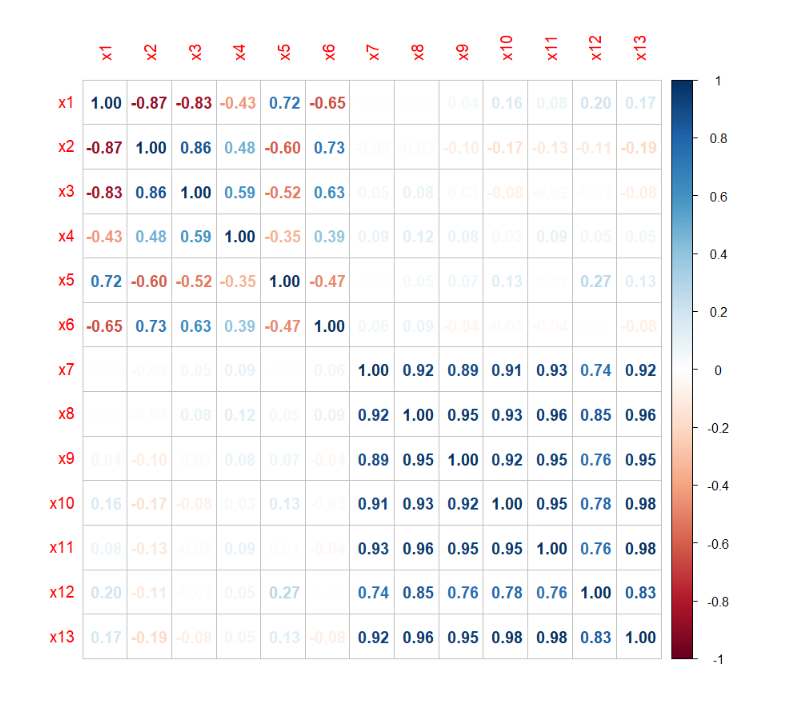






Observam asimetrie la dreapta, de unde ne putem sa ne dam seama ca predomina valorile mici ale ratei persoanelor care nu au folosit calculatorul.

1. **Matricea de corelatie**

****

Observam corelatie POZITIVA SI PUTERNICA=>evolutie in acelasi sens

* x1(rata persoanelor care nua u folosit un calculator)si x4 (tinerilor care nu studiaza nici nu lucreaza) de 0,72
* simultan intre x7,x8,x9,x10 care se refera la studiile pe care persoanele le au

LIPSA DE CORELATIE intre

* x1(persoane care nu au folosit niciodata un calculator)-x7,x8,x9(studenti si elevi )

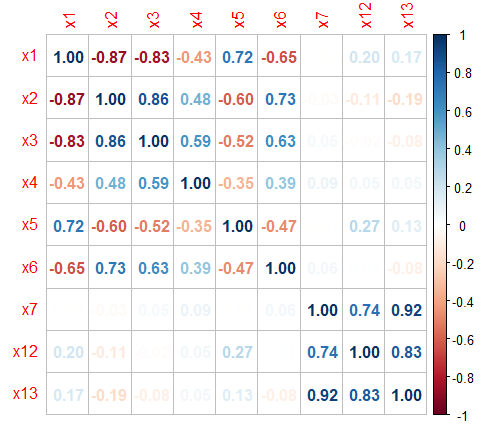
CORELATIE PUTERNICA, NEGATIVA, INVERSA intre

* x1(persoane care nu au folosit niciodata un calculator) cu x2 si x3 (rata persoanelor care au folosit internetul)

1. **Analiza factoriala**

Analiza factoriala este o tehnica statistica multivariata. Aceasta are ca scop extragerea unui numar mic de factori latenti, ei fiind responsabili de corelatiile dintre variabilele originale.

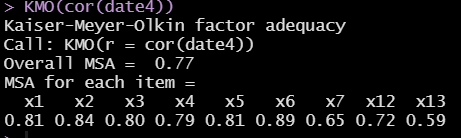
Datorita gradului mare de corelatie dintre unele componente, vom renunta la acestea pentru ca se comporta similar. Astfel obtinem noua matrice de corelatie cu ajutorul careia vom efectua analiza factoriala.



## **Indicele KMO**

Cu ajtorul testului Kaiser-Meyer-Olkin (KMO) se măsoară a modului în care datele sunt potrivite pentru analiza factorilor. Testul măsoară adecvarea eșantionării pentru fiecare variabilă din model și pentru modelul complet. Statisticile reprezintă o măsură a proporției varianței între variabile care ar putea fi variante comune. Pentru ca datele sa fie potrivite proportia trebuie sa fie cat mai mica.

Valori KMO mai mici de 0.50 indică date inadecvate pentru analiza factoriala. KMO=suma patratelor corelatiilor dintre variabile/(suma patratelor corelatiilor+suma patratelor corelatiilor partiale din matrice). KMO indica valori ridicate ale corelatiilor partiale, semn de comunalitate redusa in matricea de date.

****

KMO=0.77 => factorabilitate medie

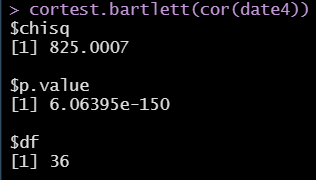
## **Testul Bartlett**

Testul de sfericitate al lui Bartlett testează ipoteza că matricea de corelatie este o matrice de identitate, ceea ce ar indica faptul că variabilele nu au nicio legătură și prin urmare, nu sunt adecvate pentru analiza. Valorile mici (mai mici de 0,05) ale nivelului de semnificație indică faptul că o analiză a factorilor poate fi utilă.

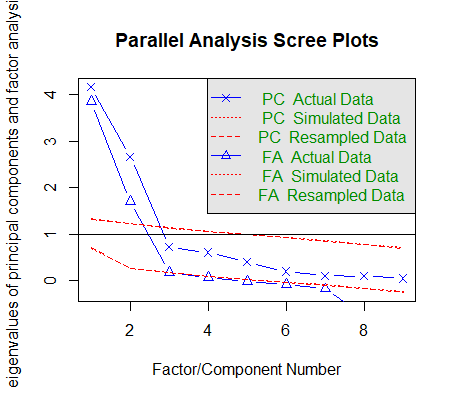
Presupune 2 ipoteze statistice:

H0: Variabilele sunt ortogonale=>nu se poate aplica analiza factoriala pe date

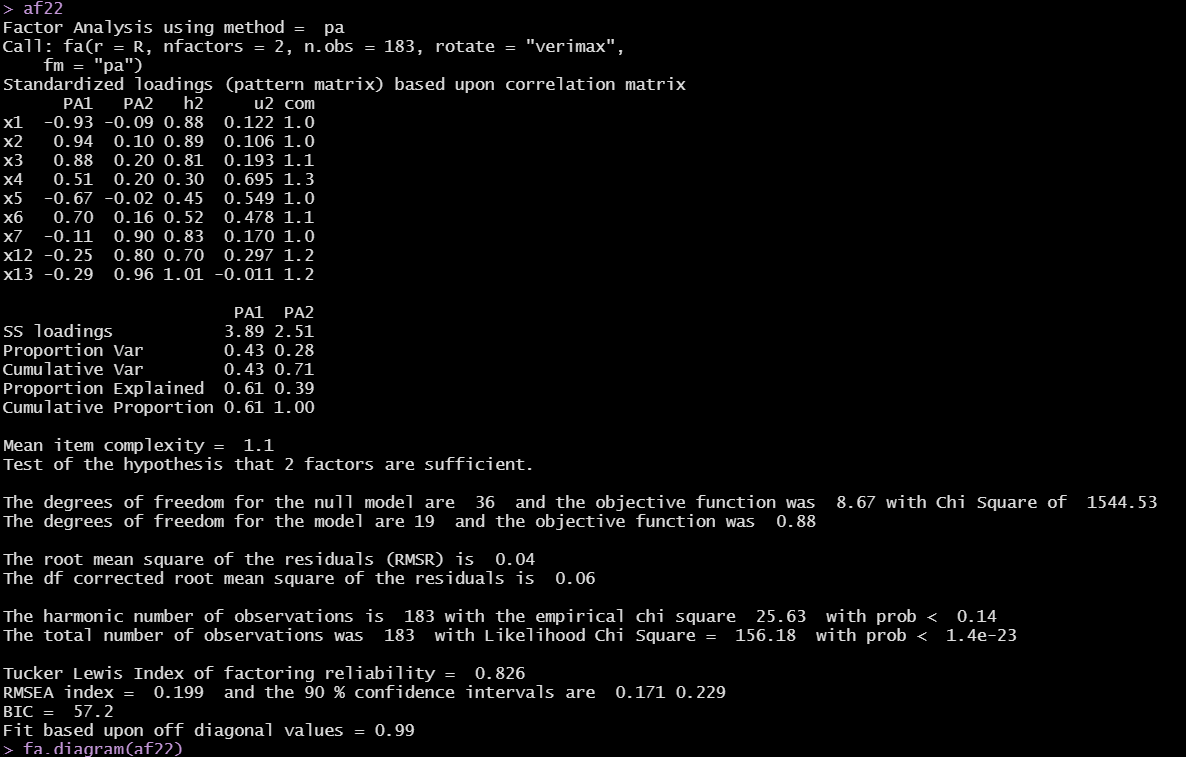
H1: Exista cel putin un factor comun=> se poate aplica analiza factoriala

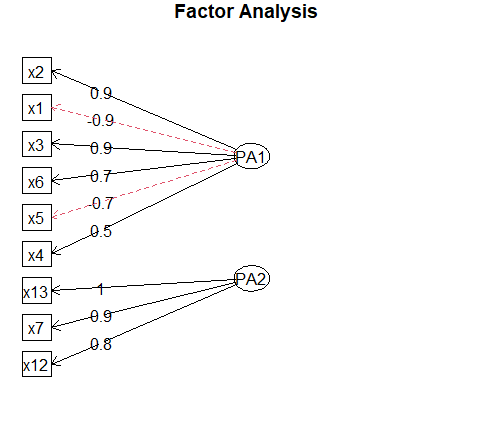


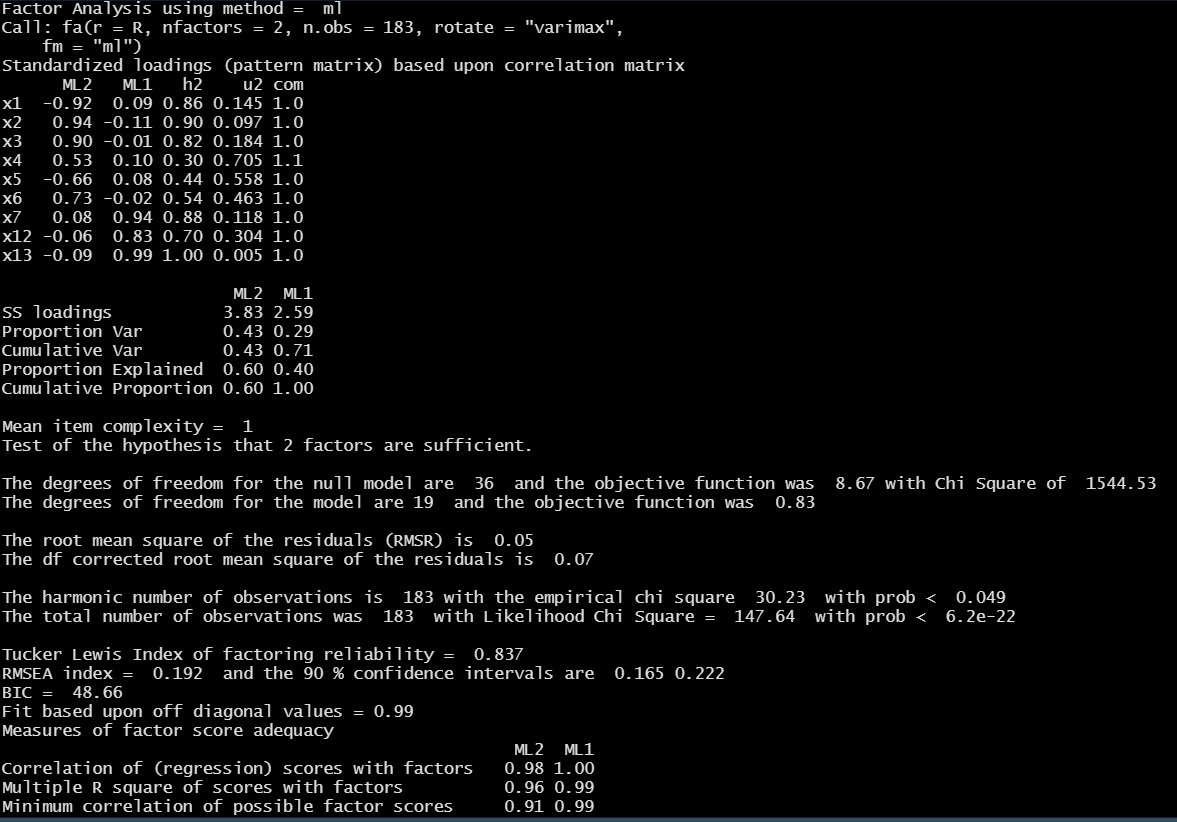
p\_value=6.06395e-150 < 0.05 => acceptam H1, putem aplica analiza factoriala.

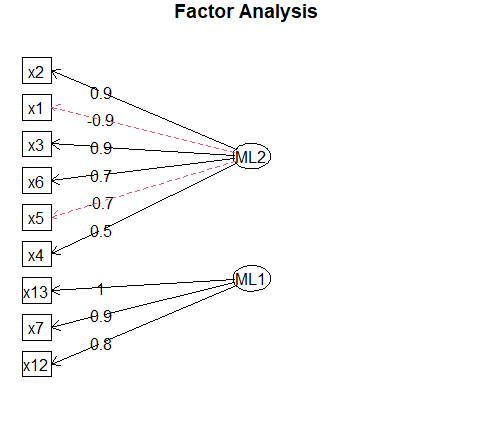


Dupa cum putem observa din graficul de mai sus pentru afectuarea analizei ni se sugereaza sa folosim 2 factori.



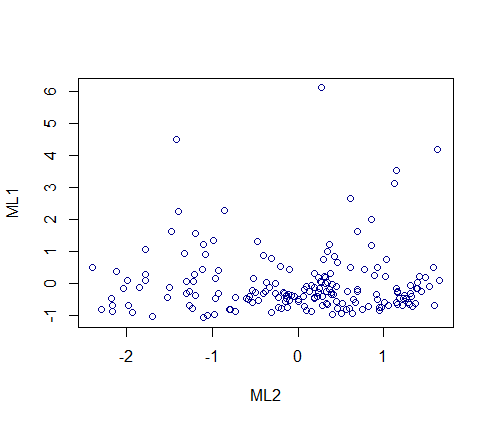






X1, X2, X3, X4, X5, X6 sunt explicate de al doilea factor, astfel factorul ML2 il putem denumi influenta internetului.

X7, X12, X13 sunt explicate de primul factor, astfel ML1 il putem denumi capacitatea intelectuala pentru a munci.



1. **Anexe**

setwd("C:/Users/asus/Desktop/Analiza datelor")

date <- read.csv("C:/Users/asus/Desktop/Analiza datelor/dateproiect.csv", row.names=1)

date

summary(date)

#-------------------------------Descrierea datelor

#--------X1-rata peroanelor care nu au folosit un calculator

#rata medie a persoanelor care nu au folosit un calculator este de 15.32, rata minima a fost de 0.98, la o diferenta de 14.34,

#iar cea maxima de 45.29

#diferenta dintre medie si mediana este de 1.78, destul de mare=>date imprastiate

#observam ca prima quartila se inregistreaza la 7.70(la o diferenta de 6.72 fata de rata cea mai mica=outlieri)

#25% dintre regiuni au rata persoanelor care nu au folosit internetul sub 7.70

#50% diintre regiuni au rata persoanelor care nu au folosit internetul peste 13.54

#25% dintre regiuni au rata persoanelor care nu au folosit internetul peste 22.57

#--------X2-procentul indivizilor care au folosit internetul pentru a contacta autoritatile

#procentul mediu al indivizilor care au folosit internetul pentru a contacta autoritatile este de 58.68%, procentul minim a fost de 9.75

#iar cel maxim de 94.59

#diferenta dintre medie si mediana este de 1.6, relativ mare=>datele sunt usor imprastiate

#observam ca prima quartila se inregistreaza la 45.73(la o diferenta de 35.98 fata de procentul minim = outlieri)

#25% dintre regiuni au procentul indivizilor care au folosit internetul pentru a contacta autoritatile sub 45.73

#50% diintre regiuni au procentul indivizilor care au folosit internetul pentru a contacta autoritatile peste 60.29

#25% dintre regiuni au procentul indivizilor care au folosit internetul pentru a contacta autoritatile peste 72.28

#--------X3-procentul persoanelor care au folosit internetul

#procentul mediu al persoanelor care au folosit internetul este de 86.68%, procentul minim a fost de 60.37

#iar cel maxim de 99.32

#diferenta dintre medie si mediana este de 2.07, relativ mica => datele sunt usor imprastiate

#observam ca prima quartila se inregistreaza la 81.69(la o diferenta de 21.32 fata de procentul minim = outlieri)

#25% dintre regiuni au procentul persoanelor care au folosit internetul sub 81.69

#50% diintre regiuni au procentul persoanelor care au folosit internetul peste 88.75

#25% dintre regiuni au procentul persoanelor care au folosit internetul peste 93.39

#--------X4-procentul gospodariilor cu internet

#procentul mediu al gospodariilor cu internet este de 90.50%, procentul minim a fost de 17.90

#iar cel maxim de 100%

#diferenta dintre medie si mediana este de 1.33, relativ mica => datele sunt usor imprastiate

#observam ca prima quartila se inregistreaza la 87.07(la o diferenta de 69.17 fata de procentul minim = outlieri)

#25% dintre regiuni au procentul gospodariilor cu internet sub 87.07

#50% diintre regiuni au procentul gospodariilor cu internet peste 91.83

#25% dintre regiuni au procentul gospodariilor cu internet peste 95.34

#--------X5-rata tinerilor care nu studiaza nici nu lucreaza

#rata medie a tinerilor care nu studiaza nici nu lucreaza este de 14.17, rata minima a fost de 3.9

#iar cea maxima de 40.20

#diferenta dintre medie si mediana este de 1.17, relativ mica =>datele nu sunt imprastiate

#observam ca prima quartila se inregistreaza la 9.20(la o diferenta de 5.30 fata de rata cea mai mica=outlieri)

#25% dintre regiuni au rata tinerilor care nu studiaza nici nu lucreaza sub 9.20

#50% diintre regiuni au rata tinerilor care nu studiaza nici nu lucreaza peste 13.00

#25% dintre regiuni au rata tinerilor care nu lucreaza nici nu studiaza peste 17.30

#--------X6-rata de participare la educatie

#rata medie de participarela educatie este de 10.65, rata minima a fost de 0.60

#iar cea maxima de 30.80

#diferenta dintre medie si mediana este de 1.15, relativ mica =>datele nu sunt imprastiate

#observam ca prima quartila se inregistreaza la 5.55(la o diferenta de 4.95 fata de rata cea mai mica=outlieri)

#25% dintre regiuni au rata de participare la educatie sub 5.55

#50% diintre regiuni au rata de participare la educatie peste 9.50

#25% dintre regiuni au rata de participare la educatie peste 14.15

#--------X7-studenti inscrisi la studii univerisitare

#numarul mediu de studenti inscrisi la studii universitare este de 64871, numarul minim este de 509

#iar maximul de 618854

#diferenta dintre medie si mediana este de 4036, mare =>datele sunt imprastiate

#observam ca prima quartila se inregistreaza la 20980(la o diferenta de 20471 fata de valoarea cea mai mica=outlieri)

#25% dintre regiuni au numarul de studenti inscrisi sub 20980

#50% diintre regiuni au numarul de studenti inscrisi peste 40835

#25% dintre regiuni au numarul de studenti inscrisi peste 82404

#--------X8-elevi inscrisi in invatamant primar

#numarul mediu de elevi inscrisi in invatamantul primar este de 103207, numarul minim este de 1988

#iar maximul de 831810

#diferenta dintre medie si mediana este de 29369, mare =>datele sunt imprastiate

#observam ca prima quartila se inregistreaza la 38705(la o diferenta de 36717 fata de valoarea cea mai mica=outlieri)

#25% dintre regiuni au numarul de elevi inscrisi in invatamantul primar sub 38705

#50% diintre regiuni au numarul de elevi inscrisi in invatamantul primar peste 73838

#25% dintre regiuni au numarul de elevi inscrisi in invatamantul primar peste 124209

#--------X9-elevi inscrisi in invatamant gimnazial

#numarul mediu de elevi inscrisi in invatamantul gimnazial este de 72597, numarul minim este de 1006

#iar maximul de 640525

#diferenta dintre medie si mediana este de 16676, mare =>datele sunt imprastiate

#observam ca prima quartila se inregistreaza la 28867(la o diferenta de 27861 fata de valoarea cea mai mica=outlieri)

#25% dintre regiuni au numarul de elevi inscrisi in invatamantul gimnazial sub 28867

#50% diintre regiuni au numarul de elevi inscrisi in invatamantul gimnazial peste 55921

#25% dintre regiuni au numarul de elevi inscrisi in invatamantul gimnazial peste 87614

#--------X10-elevi inscrisi in invatamant liceal

#numarul mediu de elevi inscrisi in invatamantul liceal este de 78708, numarul minim este de 1092

#iar maximul de 462644

#diferenta dintre medie si mediana este de 19395, mare =>datele sunt imprastiate

#observam ca prima quartila se inregistreaza la 34625(la o diferenta de 33733 fata de valoarea cea mai mica=outlieri)

#25% dintre regiuni au numarul de elevi inscrisi in invatamantul liceal sub 34625

#50% diintre regiuni au numarul de elevi inscrisi in invatamantul liceal peste 59313

#25% dintre regiuni au numarul de elevi inscrisi in invatamantul liceal peste 93369

#--------X11-angajati de la 15-64 ani

#numarul mediu de angajati este de 766616, numarul minim este de 14700

#iar maximul de 5368800

#diferenta dintre medie si mediana este de 203216, mare =>datele sunt imprastiate

#observam ca prima quartila se inregistreaza la 349500(la o diferenta de 324800 fata de valoarea cea mai mica=outlieri)

#25% dintre regiuni au numarul angajatilor sub 349500

#50% diintre regiuni au numarul angajatilor peste 563400

#25% dintre regiuni au numarul angajatilor peste 923450

#--------X12-someri de la 15-74 ani

#numarul mediu de someri este de 67804, numarul minim este de 3300

#iar maximul de 866200

#diferenta dintre medie si mediana este de 31704, mare =>datele sunt imprastiate

#observam ca prima quartila se inregistreaza la 18150(la o diferenta de 14850 fata de valoarea cea mai mica=outlieri)

#25% dintre regiuni au numarul somerilor sub 18150

#50% diintre regiuni au numarul somerilor peste 26100

#25% dintre regiuni au numarul somerilor peste 76400

#--------X13-populatia totala 15-64 ani

#numarul mediu al populatiei totale este de 1166262, numarul minim este de 18100

#iar maximul de 7932600

#diferenta dintre medie si mediana este de 319562, mare =>datele sunt imprastiate

#observam ca prima quartila se inregistreaza la 509750(la o diferenta de 461650 fata de valoarea cea mai mica=outlieri)

#25% dintre regiuni au numarul populatiei totale sub 509750

#50% diintre regiuni au numarul somerilor peste 846700

#25% dintre regiuni au numarul somerilor peste 1403750

library(psych)

#definim o functie cv care calculeaza coeficientul de variatie

cv=function(x){

c=sd(x)/mean(x)

return (c) }

cv(date[,1])

coef\_var=apply(date,2,cv)

coef\_var

# coeficientul de variatie pt x1 este 0.67>35%=> serie omogena si medie reprezentativa

# coeficientul de variatie pt x2 este 0.36>35%=> serie omogena si medie reprezentativa

# coeficientul de variatie pt x3 este 0.09<35%=> serie neomogena si medie nereprezentativa

# coeficientul de variatie pt x4 este 0.086<35%=> serie neomogena si medie nereprezentativa

# coeficientul de variatie pt x5 este 0.47>35%=> serie omogena si medie reprezentativa

# coeficientul de variatie pt x6 este 0.64>35%=> serie omogena si medie reprezentativa

# coeficientul de variatie pt x7 este 1.17>35%=> serie omogena si medie reprezentativa

# coeficientul de variatie pt x8 este 1.03>35%=> serie omogena si medie reprezentativa

# coeficientul de variatie pt x9 este 1.029>35%=> serie omogena si medie reprezentativa

# coeficientul de variatie pt x10 este 0.938>35%=> serie omogena si medie reprezentativa

# coeficientul de variatie pt x11 este 0.937>35%=> serie omogena si medie reprezentativa

# coeficientul de variatie pt x12 este 1.47>35%=> serie omogena si medie reprezentativa

# coeficientul de variatie pt x13 este 0.967>35%=> serie omogena si medie reprezentativa

par(mfrow=c(2,2))

boxplot(date[,1],main="X1",col = "red")

boxplot(date[,2],main="X2",col = "blue")

boxplot(date[,3],main="X3",col = "green")

boxplot(date[,4],main="X4",col = "magenta")

par(mfrow=c(1,2))

boxplot(date$x1,

main="boxplot Rata persoanelor care nu au folosit un calculator",

col="darkblue")

hist(date$x1,

main="Histograma Rata persoanelor care nu au folosit un calculator",

xlab="Rata persoanelor care nu au folosit un calculator",

col="darkblue")

# remarcam prezenta outlierilor pe rate mari ale participarii la educatie

boxplot(date$x2,

main="boxplot Numarul studentilor inscrisi la studii universitare",

col="yellow")

hist(date$x2,

main="Histograma Numarul studentilor inscrisi la studii universitare",

xlab="Numarul studentilor inscrisi la studii universitare",

col="yellow")

#nu exista outlieri pe numarul studentilor inscrisi la studii universitare

#grafic cu 3 variabile

ggplot(date, aes(x1, x2)) + geom\_point(aes(color = x3), alpha = 0.5) + scale\_color\_gradient(high = 'red', low = 'blue')

#4 var +describe

ggplot(date, aes(x1, x2)) + geom\_point(aes(color = x3, size = x4))

par(mfrow=c(1,1))

hist(date$x1,

col="blue4",

border="green",

prob=TRUE,

xlab="X1",

main="histograma X1")

lines(density(date$x1), lwd=2, col="green")

#observam asimetrie la dreapta

#matricea de corelatie

C=cor(date)

library(corrplot)

corrplot(C,

method="number",

type="full",

title="matrice corelatie",

)

#observam corelatie POZITIVA SI PUTERNICA=>evolutie in acelasi sens

#--x1(rata persoanelor care nua u folosit un calculator)si x4 (tinerilor care nu studiaza nici nu lucreaza) de 0,72

#--simultan intre x7,x8,x9,x10 care se refera la studiile pe care persoanele le au

# LIPSA DE CORELATIE intre

#--x1(persoane care nu au folosit niciodata un calculator)-x7,x8,x9(studenti si elevi )

# CORELATIE PUTERNICA, NEGATIVA, INVERSA intre

#--x1(persoane care nu au folosit niciodata un calculator) cu x2 si x3 (rata persoanelor care au folosit internetul)

#standardizarea datelor

dateS <- scale(date)

dateS

corrplot(cor(dateS),

method="number",

type="full",

title="matrice corelatie",

)

library(factoextra)

library(FactoMineR)

#------------------Analiza factoriala

#verificam aplicabilitatea analizei factoriale pe datele extrase, utilizand indicele KMO si testul Bartlett

library(psych)

#alegem numarul de componente

date1 <- date[ ,-which(names(date)=="x11")]

date2 <- date1[ ,-which(names(date1)=="x8")]

date3 <- date2[ ,-which(names(date2)=="x10")]

date4 <- date3[ ,-which(names(date3)=="x9")]

corrplot(cor(date4),

method="number",

type="full",

)

KMO(cor(date4))

#0.77 -> factorabilitate medie

cortest.bartlett(cor(date4))

#p.value= 6.06395e-150<0.05 => respingem H0 (exista cel putin un factor comun)

fa.parallel(date4)

#alegem 2 factori

R <- cor(date4)

af1 <- fa(R, nfactors = 2, n.obs= 183, rotate = 'none')

af1

fa.diagram(af1)

#Metoda axelor principale

af21<-fa(R,nfactors = 2, n.obs = 183, rotate = "none", fm = "pa")

af21

fa.diagram(af21)

af22<-fa(R,nfactors = 2, n.obs = 183, rotate = "verimax", fm = "pa")

af22

fa.diagram(af22)

#Metoda verosimilitatii maxime

af31<-fa(R,nfactors = 2, n.obs = 183, rotate = "none", fm = "ml")

af31

fa.diagram(af31)

#ML1 capacitatea intelectuala pentru a munci

#ML2 influenta internetului

af32 <- fa(R,nfactors = 2, n.obs = 183, rotate = "varimax", fm = "ml")

af32

fa.diagram(af32)

#ML1 capacitatea intelectuala pentru a munci

#ML2 influenta internetului