ACADEMIA DE STUDII ECONOMICE

Facultatea: Cibernetică, Statistică și Informatică Economică

Specializarea: Cibernetică Economică



**Proiect Analiza Datelor**

Studenți: Ivasișin Andrei, Murgeanu Andrei

Grupa: 1062

Profesori îndrumători: Matei Monica Mihaela

**1.Definirea variabilelor utilizate**

Sursa datelor este: <https://ec.europa.eu/eurostat>.

Indicatorii utilizați în analiza nivelului economic prin analiza componentelor principale, analiza corelațiilor și soluției ACP cu ajutorul unui set de date corespondente anumitor țări din Europa sunt următorii:

1. **Pop** – Populația țării – exprimată în număr locuitori;
2. **GDP** – PIB (Produsul intern brut) – exprimat în milioane Euro;
3. **MinWage** – Salariul minim – exprimat în Euro/lună;
4. **Img** – Imigrația – exprimată în număr imigranți;
5. **PPE** – Producția primară de energie – o unitate = 1000 tone de petrol;
6. **PPRE** – Producția primară de energie regenerabilă – o unitate = 1000 tone de petrol;
7. **TGE** – Totalul electricității generate – exprimată în GWh;
8. **GasP** – Prețul combustibilului – exprimat în Euro/GigaJoule;
9. **EPrice** – Prețul electricității – exprimat în kWh;
10. **AirTP** – Transportul aerian de pasageri – exprimat în număr pasageri;
11. **AirTG** – Transportul aerian de bunuri – exprimat în tone;
12. **GovDebt** – Datorii publice brute – exprimate în milioane Euro.
    1. **Analiza distribuților variabilelor**

Path<-”C:\\Users\\Flo\\Desktop”

m <- read.table(file.path,”AnalizaDatelor\_Proiect\_1.txt”) , sep=”\t”, dec=”.”, header=TRUE, row.names=1)

attach(m)

View(m)

* Obținem următorul tabel:



* Vom folosi comenzi specifice R pentru a analiza distribuțiile celor 12 variabile numerice prezentate în tabelul următor.

1. **Pop:**

summary(Pop)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
| 562958 | 4338394 | 9801463 | 19324642 | 19128167 | 81197537 |

sd(Pop)

* 23988741 => Valoarea s-a abatut de la medie cu 24 milioane de locuitori.

skewness(Pop)

* 1.42 => Valoarea este mai mare decat 0; avem asimetrie la dreapta.

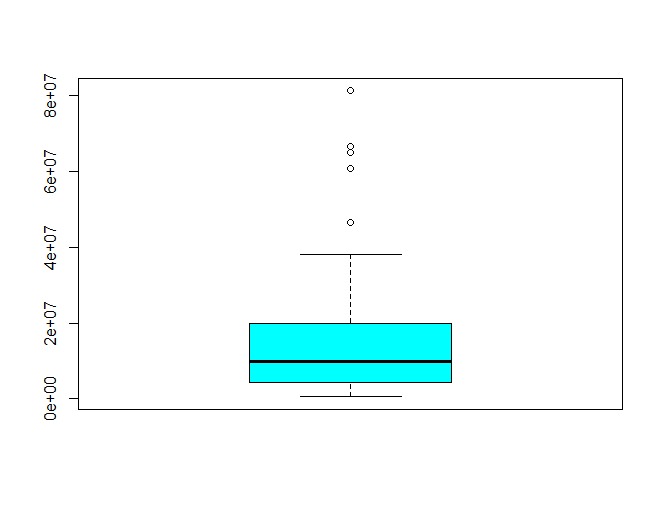
kurtosis(Pop)

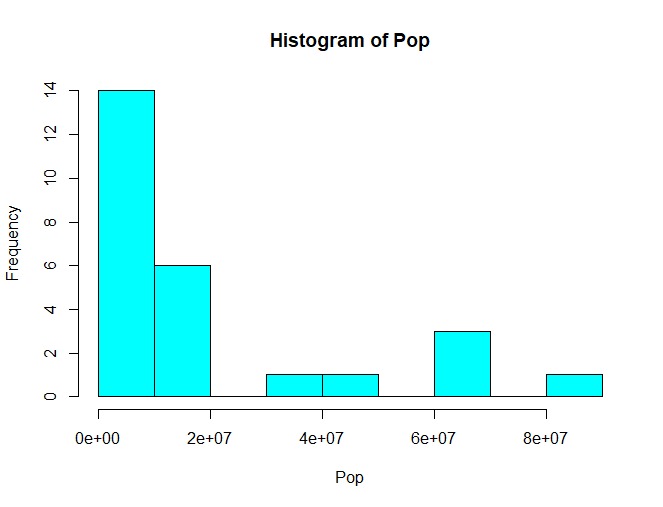
* 3.56 => Valoarea este mai mare decat 3; avem distributie leptocurtica.

boxplot(Pop,col=45)

* Observam 5 tari (Germania, Italia, Franta,etc) care au o populatie foarte mare in comparatie cu celelalte.

hist(Pop,col=45)

* Din histograma deducem o asimetrie la dreapta.
* Frecventa tarilor care au o populatie mare este destul de mica.

****

1. **GDP:**

summary(GDP)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
| 17746 | 46861 | 178534 | 559420 | 444460 | 3048860 |

sd(GDP)

* 851408.6 => Valoarea s-a abatut de la medie cu 851 milioane de euro.

skewness(GDP)

* 1.86 => Valoarea este mai mare decat 0; avem asimetrie la dreapta.

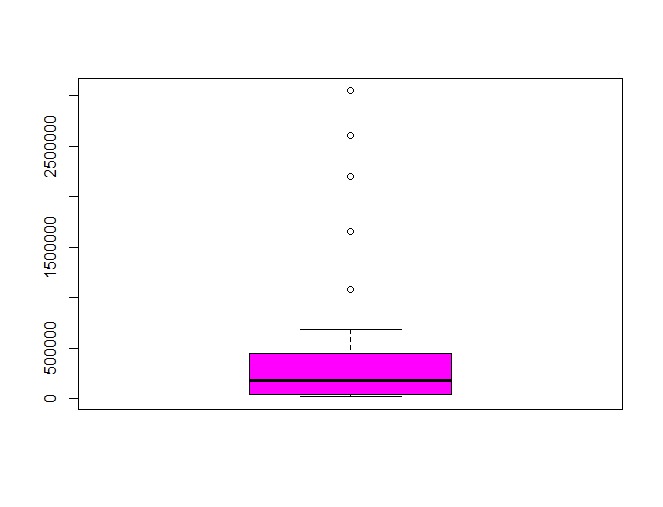
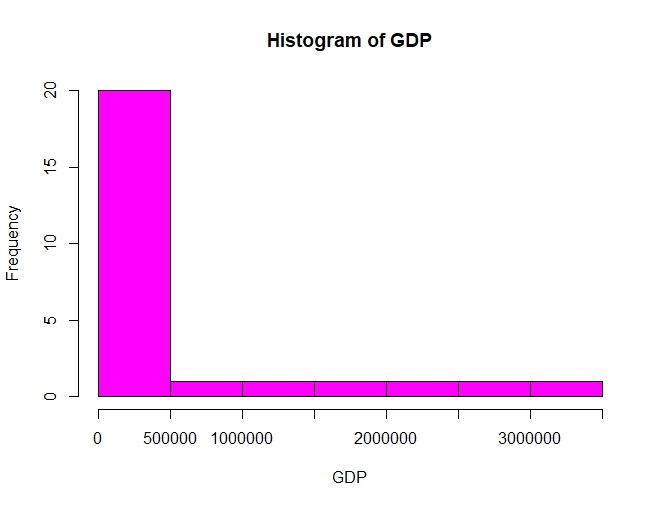
kurtosis(GDP)

* 5.20 => Valoarea este mai mare decat 3; avem distributie leptocurtica.

hist(GDP,col=46)

* Din histograma deducem o asimetrie la dreapta.
* Frecventa tarilor care au PIB-ul mare este foarte mica.

boxplot(GDP,col=46)

* Observam 5 tari (Germania, Italia, Franta,etc) care au un PIB ridicat.

1. **MinWage:**

summary(MinWage)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
| 184.1 | 365.0 | 449.8 | 724.9 | 1231.8 | 1923.0 |

sd(MinWage)

* 522.36 => Valoarea s-a abatut de la medie cu 522 de euro.

skewness(MinWage)

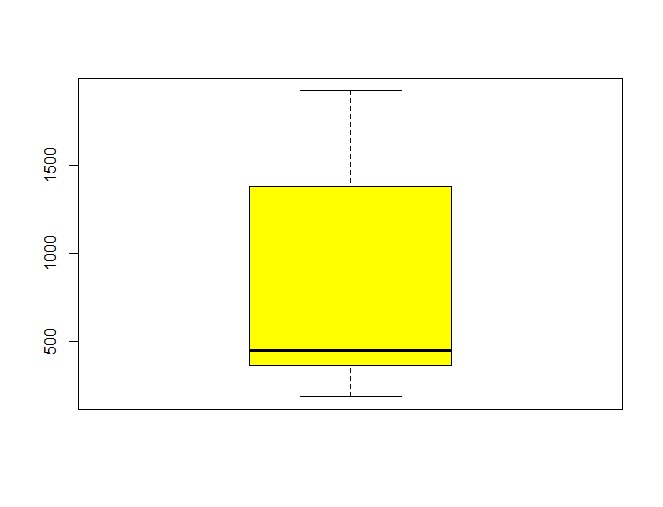
* 0.92 => Valoarea este mai mare decat 0; avem o usoara asimetrie la dreapta.

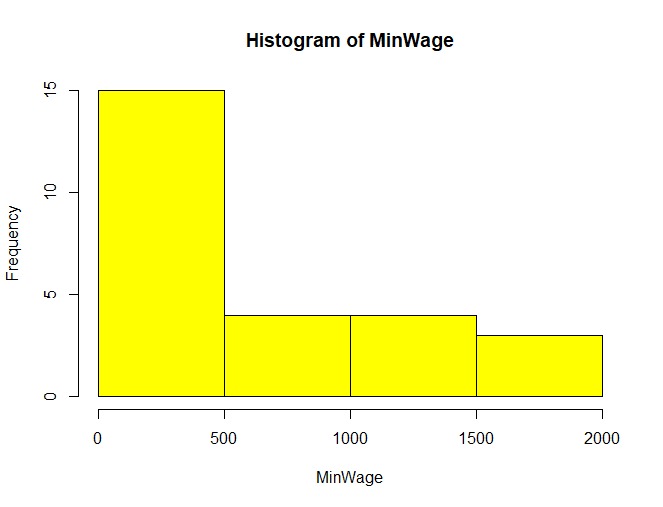
kurtosis(MinWage)

* 2.38 => Valoarea este mai mica decat 3; avem distributie platicurtica.

hist(MinWage,col=47)

* Din histograma deducem o asimetrie la dreapta.
* Observam ca predomina tarile in care salariul este mai mic de 500 de euro/luna.

boxplot(MinWage,col=47)

* Nu deducem existenta niciunui outlier.

1. **Img:**

summary(Img)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
| 6997 | 22548 | 61395 | 175535 | 166735 | 1543848 |

sd(Img)

* 315520.3 => Valoarea s-a abatut de la medie cu 315520 de imigranti.

skewness(Img)

* 3.40 => Valoarea este mai mare decat 0; avem asimetrie la dreapta.

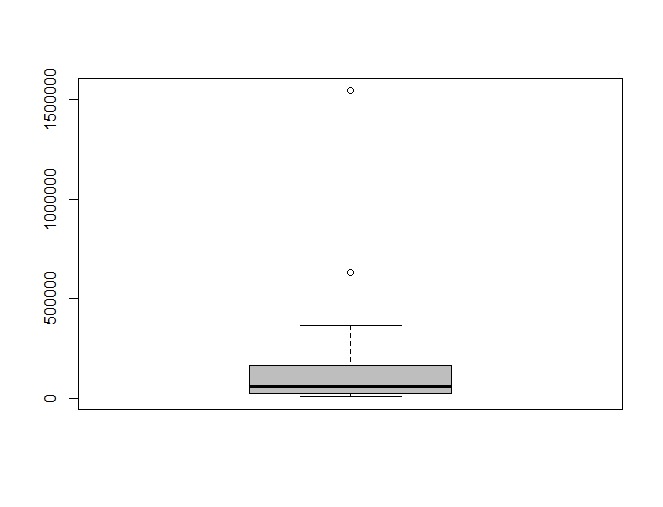
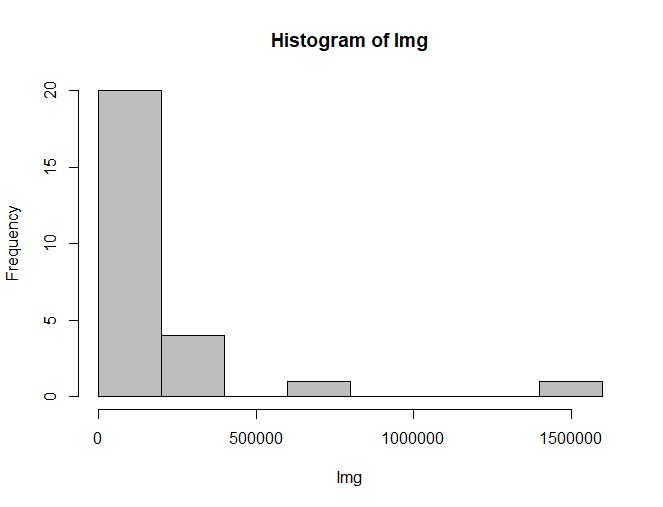
kurtosis(Img)

* 14.93 => Valoarea este mult mai mare decat 3; avem distributie leptocurtica.

hist(Img,col=47)

* Din histograma deducem o asimetrie la dreapta.
* Spatiile libere ne arata frecventa foarte mica a tarilor care au un numar de imigranti cuprins intre 800.000 si 1.300.000.

boxplot(Img,col=47)

* Observam un outlier (Germania), a carei numar de imigrati este foarte mare.

1. **PPE:**

summary(PPE)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
| 121.2 | 4619.3 | 11591.3 | 28910.6 | 33551.8 | 137527.3 |

sd(PPE)

* 39171.59 => Valoarea s-a abatut de la medie cu 39171.59 de unitati.

skewness(PPE)

* 1.75 => Valoarea este mai mare decat 0; avem asimetrie la dreapta.

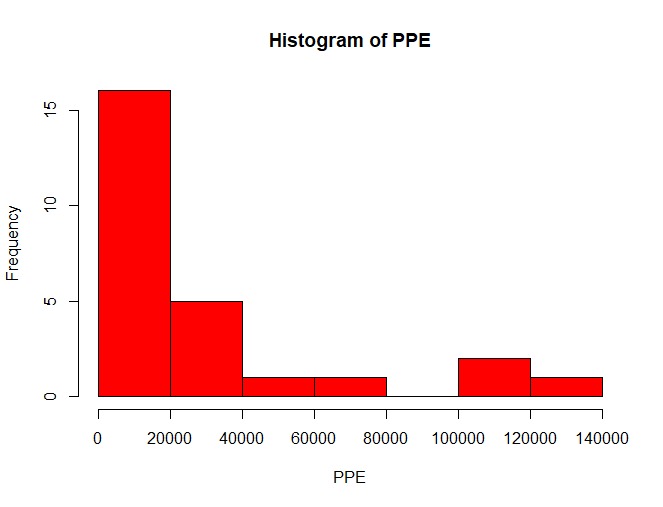
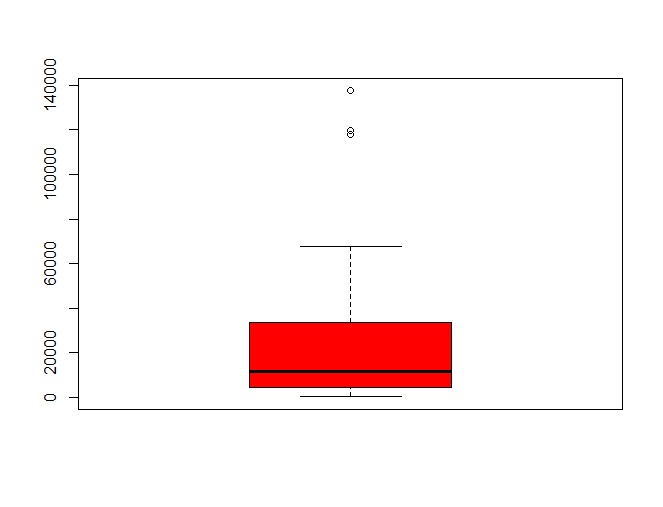
kurtosis(PPE)

* 4.89 => Valoarea este mai mare decat 3; avem distributie leptocurtica.

hist(PPE,col=50)

* Din histograma deducem o asimetrie la dreapta.

boxplot(PPE,col=50)

* Observam trei tari care au o productie ridicata de energie.

1. **PPRE:**

summary(PPRE)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
| 115.2 | 1701.8 | 3763.7 | 7777.2 | 10128.3 | 38886.1 |

sd(PPRE)

* 9301.454 => Valoarea s-a abatut de la medie cu 9301.454 de unitati.

skewness(PPRE)

* 1.80 => Valoarea este mai mare decat 0; avem asimetrie la dreapta.

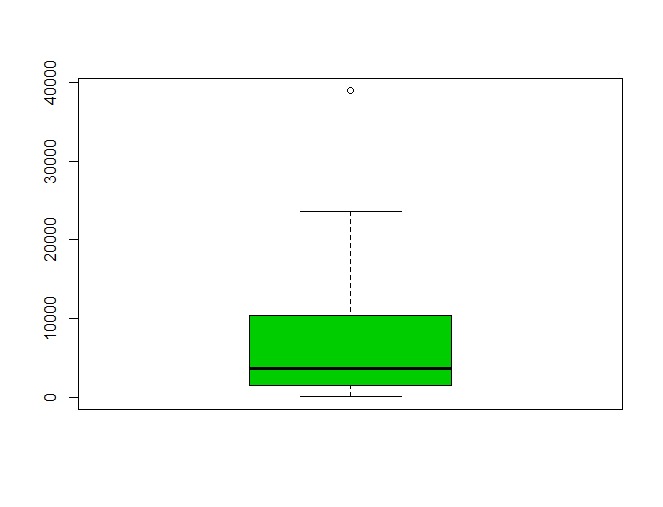
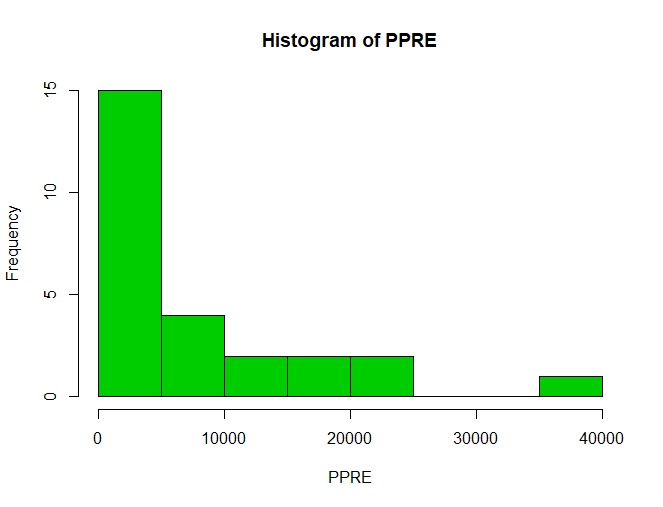
kurtosis(PPRE)

* 6 => Valoarea este mai mare decat 3; avem distributie leptocurtica.

hist(PPRE,col=51)

* Din histograma deducem o asimetrie la dreapta, cu predominante tari care au valori cuprinse intre 0-5.000 de unitati.

boxplot(PPRE,col=51)

* Observam un outlier, Germania care este foarte dezvoltata in acest domeniu.

1. **TGE:**

summary(TGE)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
| 2766 | 18051 | 58859 | 123269 | 149061 | 646888 |

sd(TGE)

* 170213 => Valoarea s-a abatut de la medie cu 170213 de GWh.

skewness(TGE)

* 1.95 => Valoarea este mai mare decat 0; avem asimetrie la dreapta.

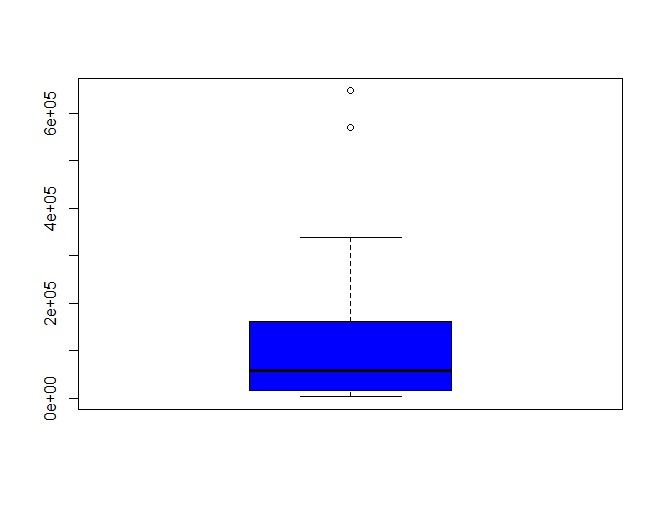
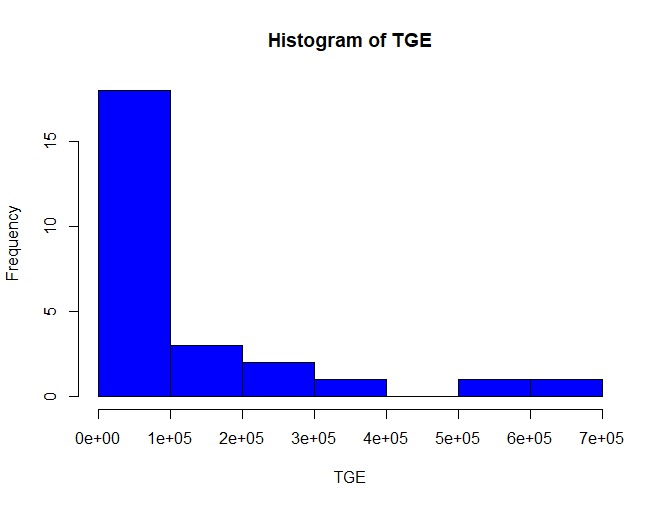
kurtosis(TGE)

* 5.95 => Valoarea este mai mare decat 3; avem distributie leptocurtica.

hist(TGE,col=52)

* Din histograma deducem o asimetrie la dreapta.

boxplot(TGE,col=52)

* Observam doua tari a caror generare de electricitate este ridicata.

1. **GasP:**

summary(GasP)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
| 8651 | 13770 | 17245 | 17139 | 19520 | 31431 |

sd(GasP)

* 5.01 => Valoarea s-a abatut de la medie cu 5.1 de euro.

skewness(GasP)

* 0.84 => Valoarea este mai mare decat 0; avem asimetrie la dreapta.

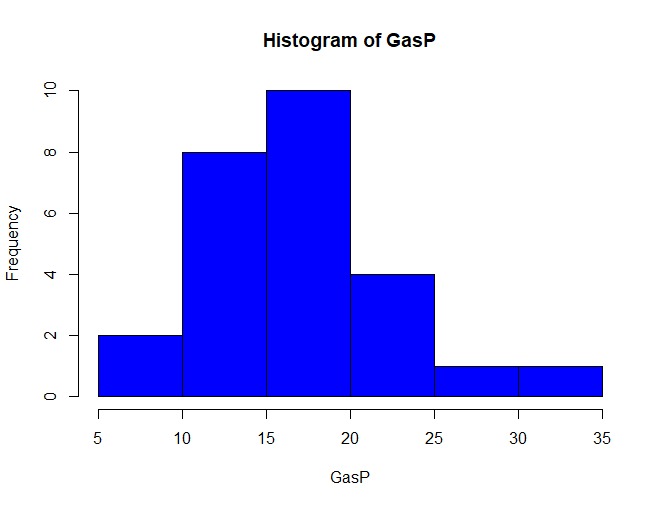
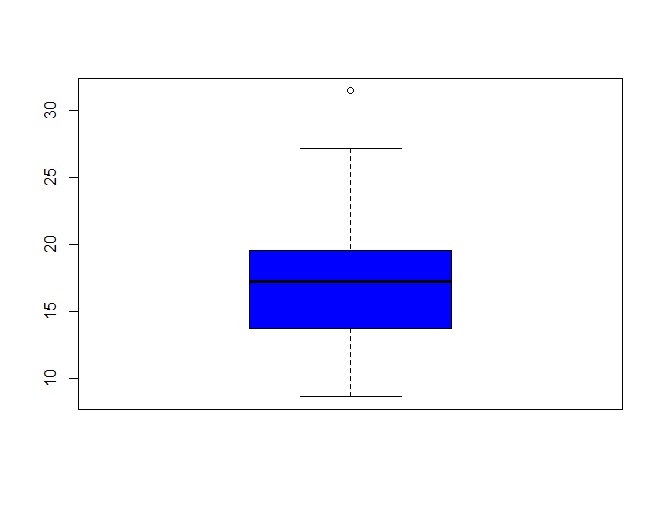
kurtosis(GasP)

* 4.13 => Valoarea este mai mare decat 3; avem distributie leptocurtica.

hist(GasP,col=52)

* Din histograma deducem o usoara asimetrie la dreapta; distributia este inclinata spre stanga, avand mai multe valori extreme spre dreapta.

boxplot(GasP,col=52)

* O singura tara are pretul pe combustibil mai mare, insa aceasta abatere este neglijabila.

1. **EPrice:**

summary(EPrice)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
| 0.0956 | 0.1354 | 0.1624 | 0.1726 | 0.2013 | 0.2969 |

sd(EPrice)

* 0.04915276 => Abaterea de la medie este aproape nula.

skewness(EPrice)

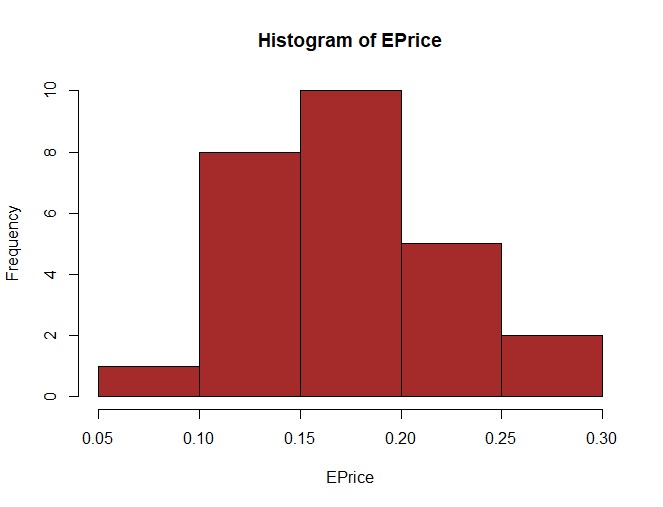
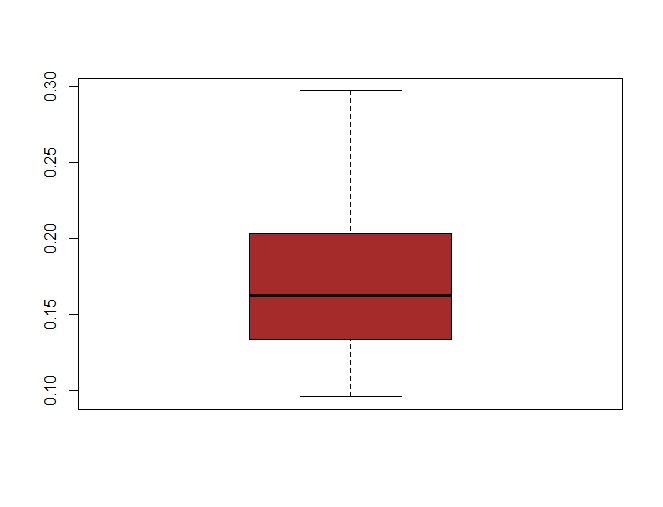
* 0.70 => Valoarea este mai mare decat 0; avem usoara asimetrie la dreapta.

kurtosis(EPrice)

* 2.89 => Valoarea este mai mica decat 3; distributia se apropie de una mezocurtica.

hist(EPrice,col="brown")

* Din histograma deducem o usoara asimetrie la dreapta.

boxplot(EPrice,col="brown")

1. **AirTP:**

summary(AirTP)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
| 1404152 | 7847052 | 22640732 | 51590312 | 44390039 | 248868873 |

sd(AirTP)

* 70623659 => Valoarea s-a abatut de la medie cu 70623659 de calatori.

skewness(AirTP)

* 1.63 => Valoarea este mai mare decat 0; avem asimetrie la dreapta.

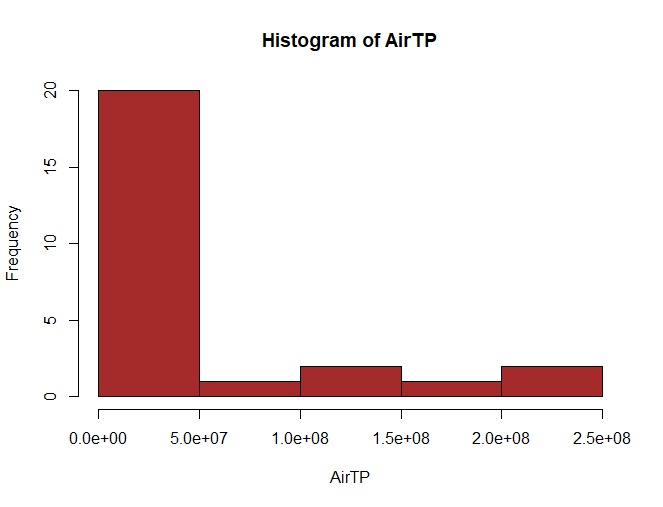
kurtosis(AirTP)

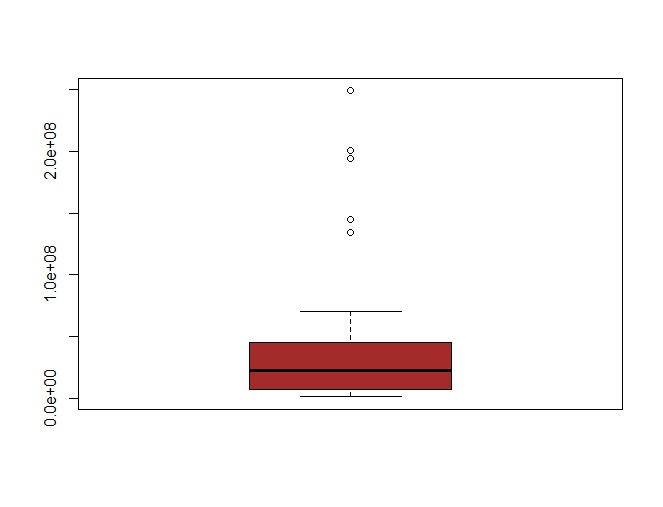
* 4.33 => Valoarea este mai mare decat 3; avem distributie leptocurtica.

hist(AirTP,col="brown")

* Din histograma deducem asimetrie la dreapta, cu valori predominante in intervalul 1-5milioane.

boxplot(AirTP,col="brown")

* Putem observa 5 tari care au un nr. de calatori ridicat, fapt pus pe seama populatiei, atractiilor turistice.



1. **AirTG:**

summary(AirTG)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
| 7189 | 28600 | 113053 | 595465 | 701258 | 4325665 |

sd(AirTG)

* 1040160 => Valoarea s-a abatut de la medie cu 1040160 de tone.

skewness(AirTG)

* 2.22 => Valoarea este mai mare decat 0; avem asimetrie la dreapta.

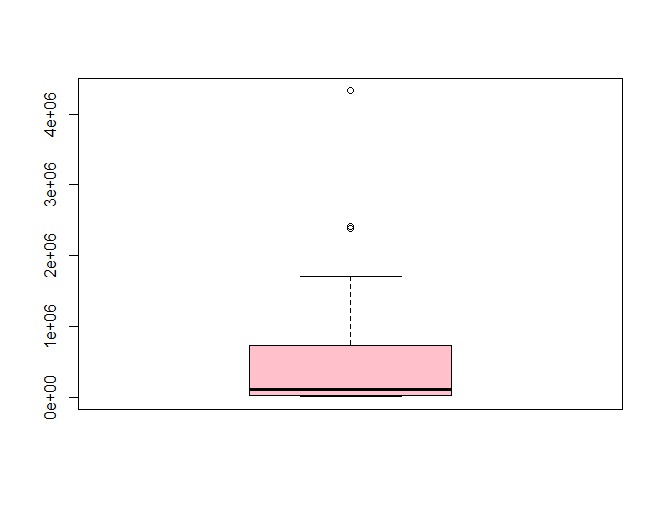
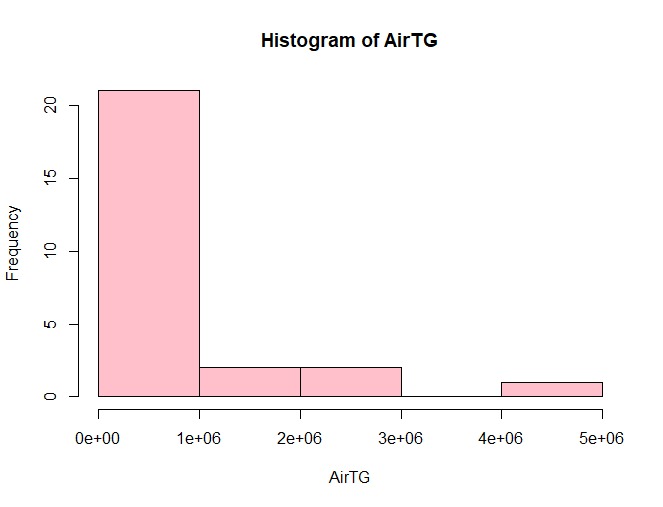
kurtosis(AirTG)

* 7.73 => Valoarea este mai mare decat 3; avem distributie leptocurtica.

hist(AirTG,col="pink")

* Din histograma deducem asimetrie la dreapta.

boxplot(AirTG,col="pink")

* Deducem existenta a cel putin unui outlier.

1. **GovDebt:**

summary(GovDebt)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Min | 1st Qu | Median | Mean | 3rd Qu | Max |
| 2035 | 33368 | 167415 | 486195 | 406087 | 2269881 |

sd(GovDebt)

* 768261.6 => Valoarea s-a abatut de la medie cu 768261.6 de milioane euro.

skewness(GovDebt)

* 1.64 => Valoarea este mai mare decat 0; avem asimetrie la dreapta.

kurtosis(GovDebt)

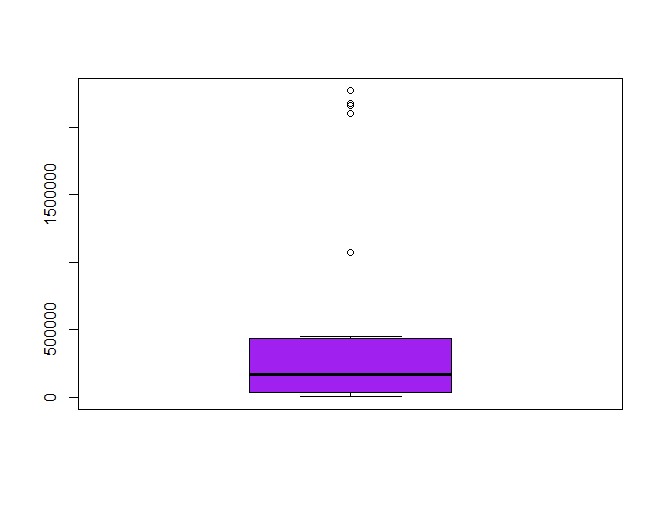
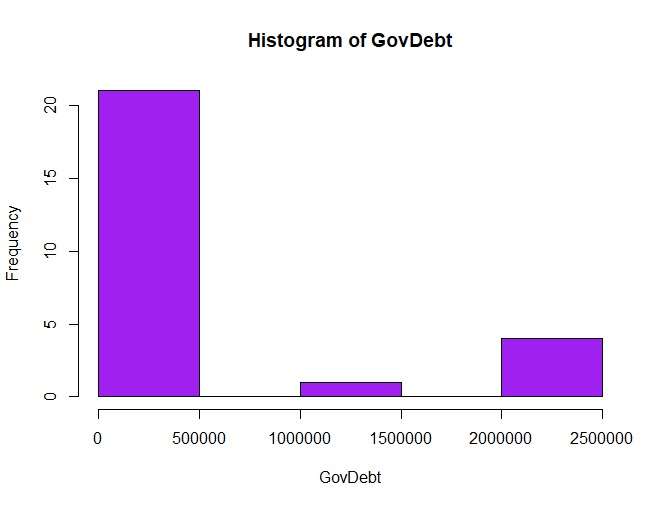
* 4 => Valoarea este mai mare decat 3; avem distributie leptocurtica .

hist(GovDebt,col="purple")

* Din histograma deducem asimetrie la dreapta si frecvente foarte mici in anumite intervale.

boxplot(GovDebt,col="purple")

* Deducem existenta a cel putin 5 outlieri, tari care au o datorii publice mari.



1. **Analiza componentelor principale**

Analiza componentelor principale este o metodă care ne ajută să descompunem variabilitatea totală într-un număr mai mic de componente față de setul inițial. Această descompunere trebuie să mențină informațiile inițiale intacte.

Cei 12 indicatori folosiți sunt exprimați în unități de măsură diferite, de aceea va fi nevoie de o standardizare a datelor.

**Instrucțiuni în R:**

* Calculăm coeficienții de corelație și apoi îi reprezentăm grafic:

c<-cor(m)

c

c2<-rcorr(as.matrix(m))

corrplot(c2$r,type="upper",o.mat=c2$p,sig.level=0.01,insig="blank")

* Am folosit functia corrplot pentru a reprezenta corelatiile dintre inidcatorii alesi.
* Astfel incat coeficientii puternic corelati sunt reprezentati prin cercurile de culoare albastru inchis.

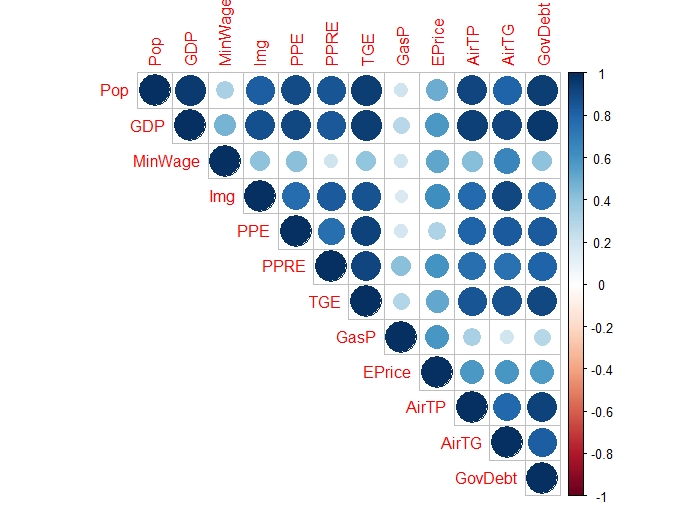
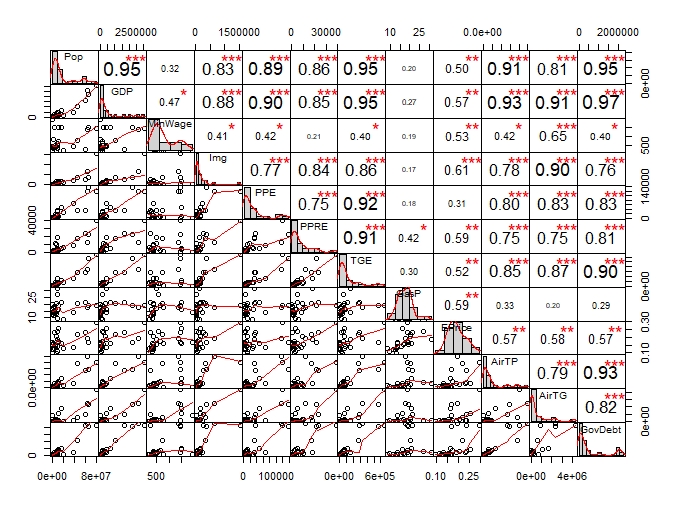


chart.Correlation(m,histogram=TRUE,pch=19)

* Pe diagonala principala sunt reprezentate distributiile fiecarei coloane.
* Deasupra diagonalei putem observa coeficientii de corelatie dintre oricare 2 variabile.
* Sub diagonala identificam norul de puncte, in cazul nostru tarile europene si cu rosu ajustarea acelui nor, prin intermediul unei functii care nu este mereu liniara.
* Dupa cum putem observa, pretul combustibilului nu se coreleaza cu ceilalti indicatori, dar asta nu inseamna ca este independent.



**INTERPRETĂRI:**

* Conform teoriei, valorile coeficienților de corelație se încadrează în intervalul [-1, 1]. Cu cât o valoare se apropie de capetele intervalului, respectiv -1 și 1, cu atât corelația între cele două variabile la a căror intersecție se află valoarea este mai puternică. Cu cât valoarea corelației se apropie de 0, cu atât este mai slabă. Valorile cuprinse între [-1, 0) denotă faptul că între cele două variabile există o corelație negativă, iar cele cuprinse în intervalul (0, 1] denotă faptul că cele două variabile sunt corelate pozitiv.
* Observăm legături puternice între **Pop&TGE, Pop&AirTP, AirTP&GDP** și **AirTG&Img etc.**, cu procente de 0.95, 0.91, 0.93 și 0.9, ceea ce explică faptul că există corelații puternice&pozitive între variabilele alese.
* Un caz de corelație slabă este între **Img** și **GasP** (0.17).

ms<-scale(m,center=TRUE,scale=TRUE)

* Calcularea matricilor de cor si cov:

corM<-round(cor(m),2)

covM<-round(cov(m),2)

corMs<-round(cor(ms),2)

covMs<-round(cov(ms),2)

* Din analizarea corelatiilor si covariantelor am dedus ca standardizarea nu schimba intensitatea legaturilor;
* Dintre variabile, intrucat corM si corMs au aceleasi valori;
* Nu ne asteptam ca legaturile sa fie influentate, intrucat scadem o constanta si impartim la o alta constanta.
* *Extragem vectorii și valorile proprii:*

vectP<-eigen(corM);

acp<-orubcino(m),cor=TRUE,scores=TRUE)

acp<-princomp((m),cor=TRUE,scores=TRUE)



vectP<-eigen(corM)

vectP



acp<-princomp((m),cor=TRUE,scores=TRUE)

summary(acp)



* Din analiza componentelor principale rezulta 12 componente care vor fi noile caracteristici ale tarilor.
* Observam ca prima componenta preia maximul de informatie din totalitatea componentelor, aproximativ 71%; iar a doua preia 10%.
* Folosim functia acp$loadings pentru a stoca coeficientii:

acp$loadings



* Stocam valorile noilor variabile folosind ponderile vectorilor proprii, prin intermediul functiei scores:

scoruri<-acp$scores

scoruri



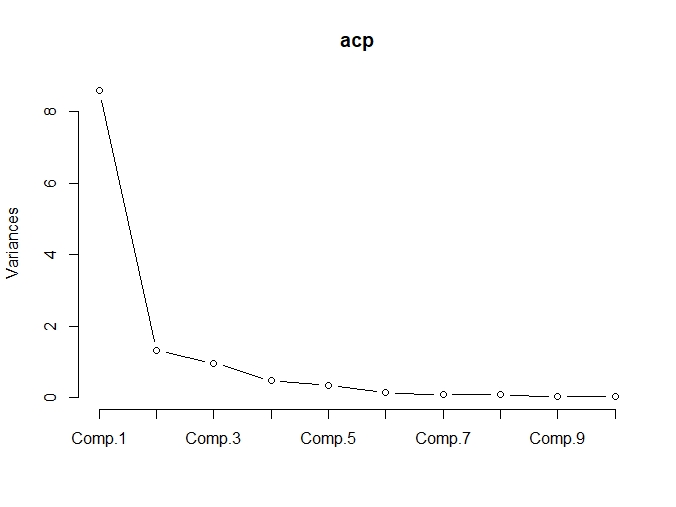
cor(scoruri)

corScor<-cor(scoruri)



* Observam numere foarte mari, necorelate 2 cate 2 si nu se mai regaseste suprapunere informationala.
* Trebuie sa decidem cate elemente dintre cele 12 o sa pastram in analiza noastra.

plot(acp,type="l")



* Dupa primele 2 componente observam o reducere semnificativa a pantei.
* Aplicam criteriul lui Kaiser, conform careia o sa pastram in analiza atatea componente principale cate valori proprii > 1 avem.
* Din summary(acp) observam ca doar primele 2 componente au valori > 1.
* Am plecat de la cei 12 indicatori initiali si am ajuns la 2 componente abstracte.
* Corelatia dintre scoruri:

cor(scoruri[,1:2],m)



* Matricea factor:

matriceaf<-cor(m,scoruri[,1:2])

matriceaf

* Cercul corelatiilor:

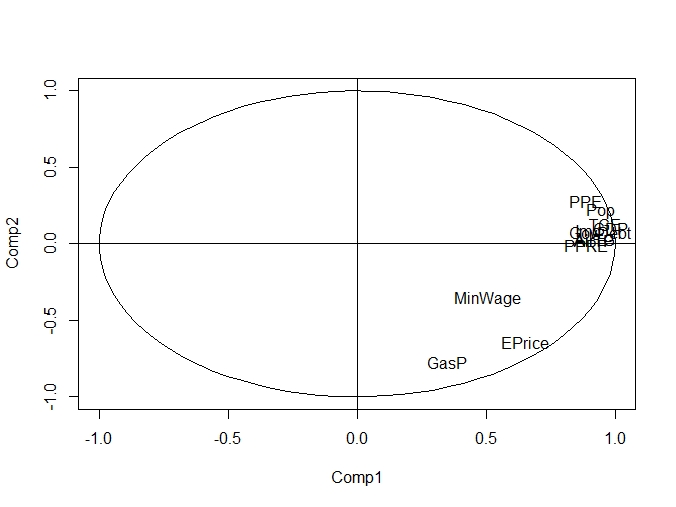
valori<-seq(0,2\*3.14,length=100)

plot(cos(valori),sin(valori),type="l",xlab="Comp1",ylab="Comp2")

text(matriceaf[,1],matriceaf[,2],rownames(matriceaf))

abline(h=0)

abline(v=0)

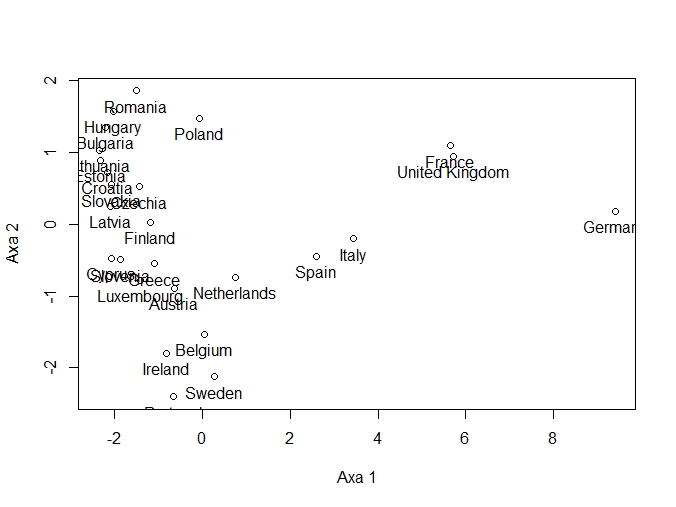


* Reprezentam tarile in noul spatiu:

plot(scoruri[,1],scoruri[,2],xlab="Axa 1",ylab="Axa 2")

text(scoruri[,1],scoruri[,2],labels=rownames(scoruri),pos=1,5)

* Coordonatele tarilor in noul spatiu sunt date de scoruri:



***Anexa:***

path<-"C:\\Users\\Flo\\Desktop"

m<-read.table(file.path(path,"AnalizaDatelor\_Proiect\_1.txt"),sep="\t",dec=".",header=TRUE,row.names=1)

attach(m)

View(m)

library(moments)

install.packages("moments")

summary(Pop)

boxplot(Pop,col=45)

sd(Pop)

skewness(Pop)

kurtosis(Pop)

hist(Pop,col=45)

summary(GDP)

boxplot(GDP,col=46)

sd(GDP)

skewness(GDP)

kurtosis(GDP)

hist(GDP,col=46)

summary(MinWage)

boxplot(MinWage,col=47)

sd(MinWage)

skewness(MinWage)

kurtosis(MinWage)

hist(MinWage,col=47)

summary(Img)

boxplot(Img,col=48)

sd(Img)

skewness(Img)

kurtosis(Img)

hist(Img,col=48)

summary(PPE)

boxplot(PPE,col=50)

sd(PPE)

skewness(PPE)

kurtosis(PPE)

hist(PPE,col=50)

summary(PPRE)

boxplot(PPRE,col=51)

sd(PPRE)

skewness(PPRE)

kurtosis(PPRE)

hist(PPRE,col=51)

summary(TGE)

boxplot(TGE,col=52)

sd(TGE)

skewness(TGE)

kurtosis(TGE)

hist(TGE,col=52)

summary(GasP)

boxplot(GasP,col=52)

sd(GasP)

skewness(GasP)

kurtosis(GasP)

hist(GasP,col=52)

summary(EPrice)

boxplot(EPrice,col="brown")

sd(EPrice)

skewness(EPrice)

kurtosis(EPrice)

hist(EPrice,col="brown")

summary(AirTP)

boxplot(AirTP,col="brown")

sd(AirTP)

skewness(AirTP)

kurtosis(AirTP)

hist(AirTP,col="brown")

summary(AirTG)

boxplot(AirTG,col="pink")

sd(AirTG)

skewness(AirTG)

kurtosis(AirTG)

hist(AirTG,col="pink")

summary(GovDebt)

boxplot(GovDebt,col="purple")

sd(GovDebt)

skewness(GovDebt)

kurtosis(GovDebt)

hist(GovDebt,col="purple")

ms<-scale(m,center=TRUE,scale=TRUE)

Pops<-scale(Pop,center=TRUE,scale=TRUE)

sd(Pops)

mean(Pops)

View(ms)

library(Hmisc)

library(corrplot)

library(PerformanceAnalytics)

c<-cor(m)

c

c2<-rcorr(as.matrix(m))

corrplot(c2$r,type="upper",o.mat=c2$p,sig.level=0.01,insig="blank")

chart.Correlation(m,histogram=TRUE,pch=19)

ms<-scale(m,center=TRUE,scale=TRUE)

corM<-round(cor(m),2)

covM<-round(cov(m),2)

corMs<-round(cor(ms),2)

covMs<-round(cov(ms),2)

corM

covM

covMs

vectP<-eigen(corM)

vectP

acp<-princomp((m),cor=TRUE,scores=TRUE)

summary(acp)

acp$loadings

scoruri<-acp$scores

scoruri

cor(scoruri)

corScor<-cor(scoruri)

ponderi<-acp$loadings

ponderi

plot(acp,type="l")

cor(scoruri[,1:2],m)

matriceaf<-cor(m,scoruri[,1:2])

matriceaf

valori<-seq(0,2\*3.14,length=100)

plot(cos(valori),sin(valori),type="l",xlab="Comp1",ylab="Comp2")

text(matriceaf[,1],matriceaf[,2],rownames(matriceaf))

abline(h=0)

abline(v=0)

plot(scoruri[,1],scoruri[,2],xlab="Axa 1",ylab="Axa 2")

text(scoruri[,1],scoruri[,2],labels=rownames(scoruri),pos=1,5)

***Datele standardizate:***

