Clasificarea a 75 de tari din intreaga lume dupa mai multi indicatori economici

(Tema3- Analiza de clusteri)

1.Sursa datelor

In fisierul „setDate.csv” se afla informatii cu privire la mai multe variabile calculate pe mai multe tari din toata lumea:

*child\_mort-* rata mortalitatii a copiilor sub 5 ani la 1000 de nasteri in viata

*exports*- exportul de bunuri si servicii per capita

*health*-toatalul cheluielilor de sanatate per capita

*imports*-importul de bunuri si servicii per capita

*income*-venitul net pe persoana

*inflation-* ratei anuala de creștere a PIB-ului total

*life\_expec-* rata numarul mediu de ani pe care un nou nascut l-ar trai daca rata mortalitatii ar ramane la fel

*total\_fer-*Numărul de copii care s-ar naște la fiecare femeie daca rata actuala a fertilitatii în funcție de vârstă rămân neschimbate.

*Gdpp –*BIP per captia

*Regiune* se refera la continentul pe care se afla fiecare tara. Continentele au urmatoarele coduri:

America de Nord- an, America de Sud-as, Europa-e, Asia-asia, Australia-aus, Africa-af

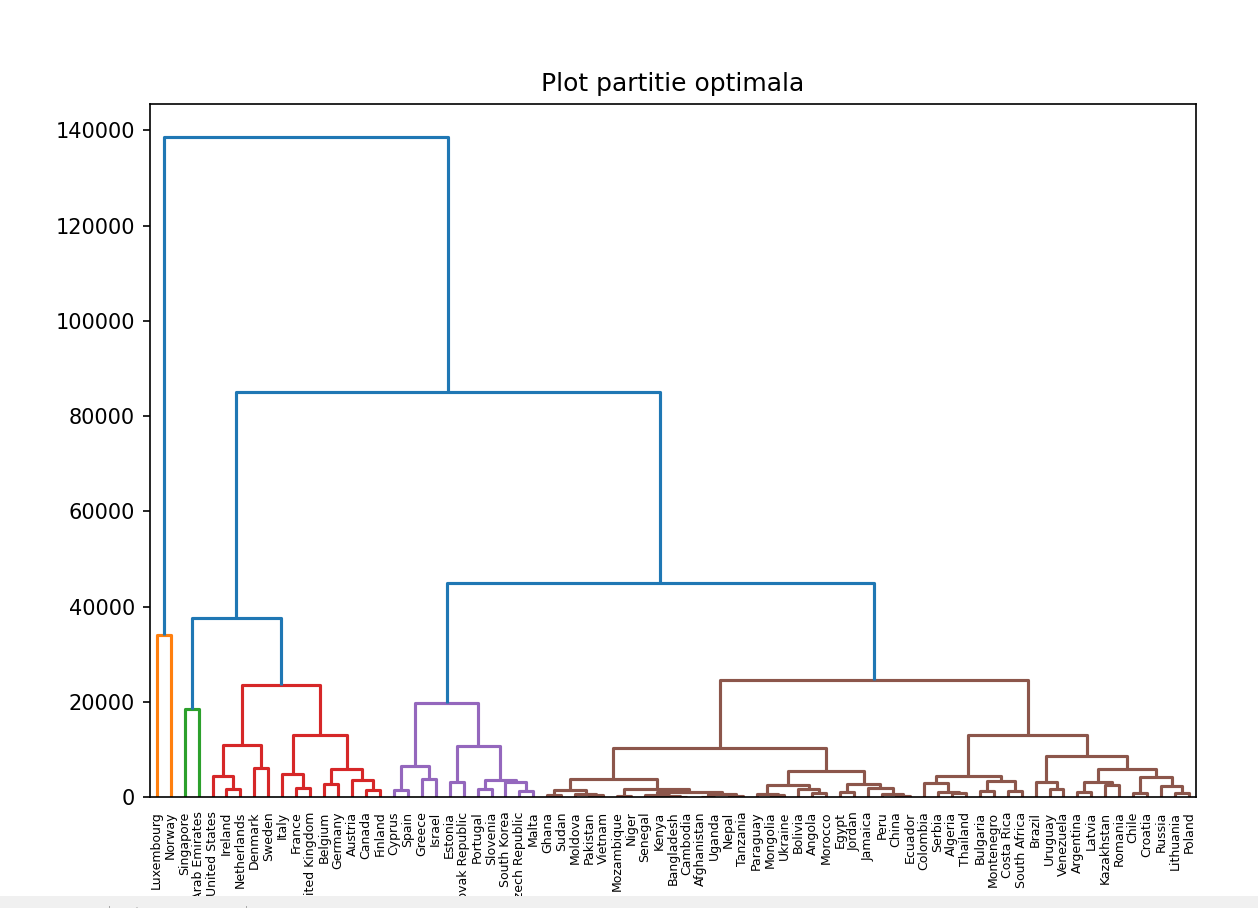
Datele au fost preluate de pe kaggle si au fost prelucrate.

<https://www.kaggle.com/code/sukhadadharangaonkar/country-data-hierarchical-clustering/data>

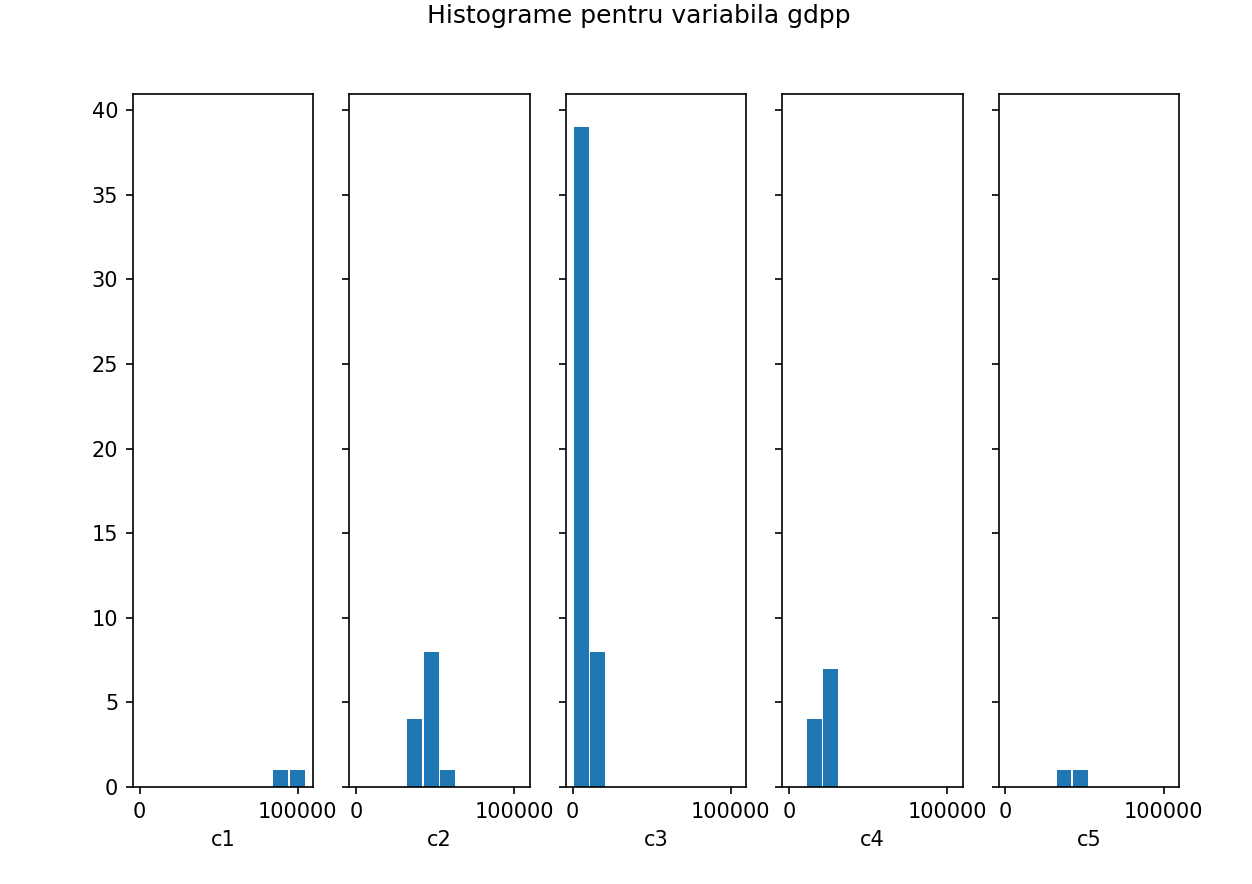
Pentru proiect s-a folosit : Agglomerative hierarhical clustering

La inceput se considera ca fiecare tara este un cluster, astfel ca vor fi k clusteri, k fiind numarul total de tari din setul de date. La fiecare iteratie, se vor grupa cate 2 tari, formand un nou cluster. La fiecare iteratie o sa avem k-1 clusteri. Intr-un final se va ajunge la un singur cluster mare, care poate fi reprezentat prin graficul dendograma. Graficul dendograma ajuta la impartirea pe clusteri a partiilor.

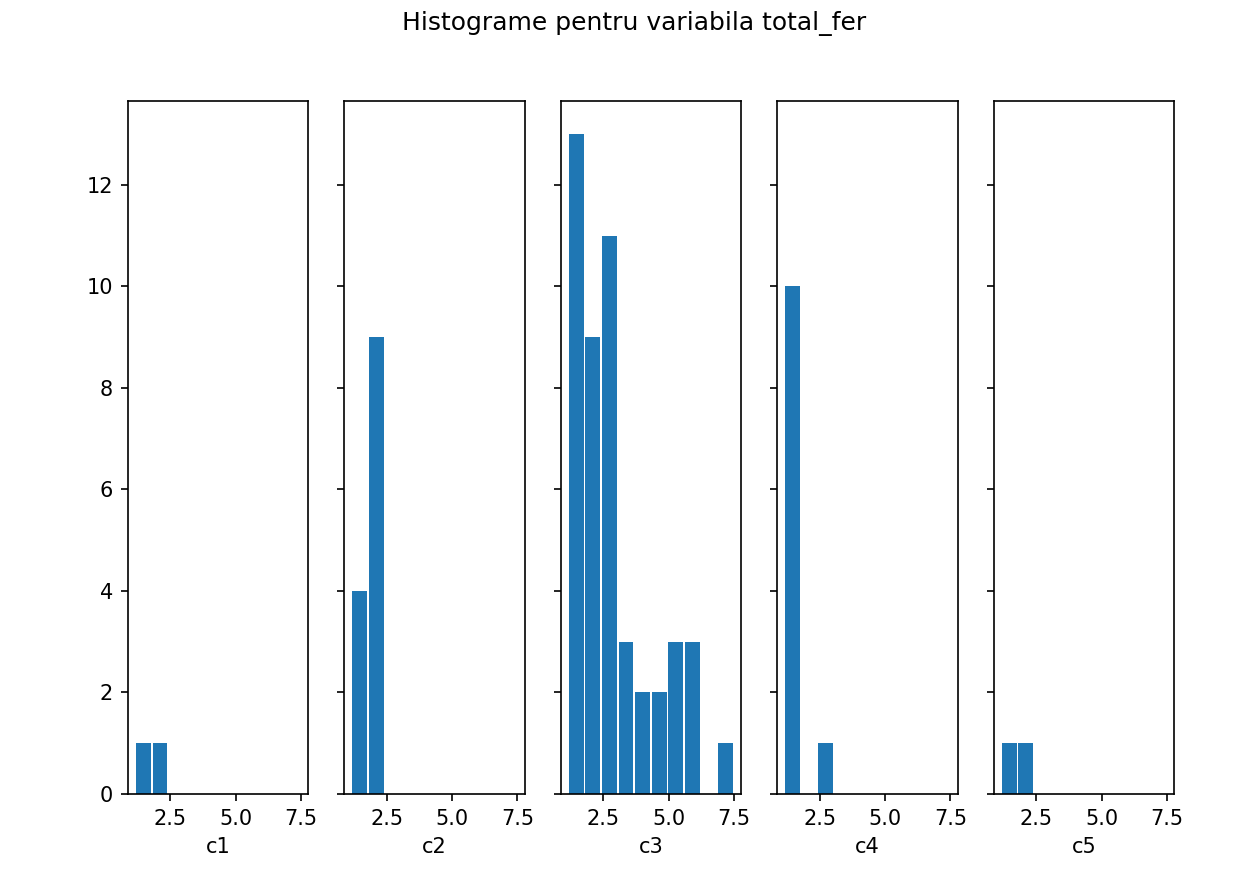
Plot partitie optimala

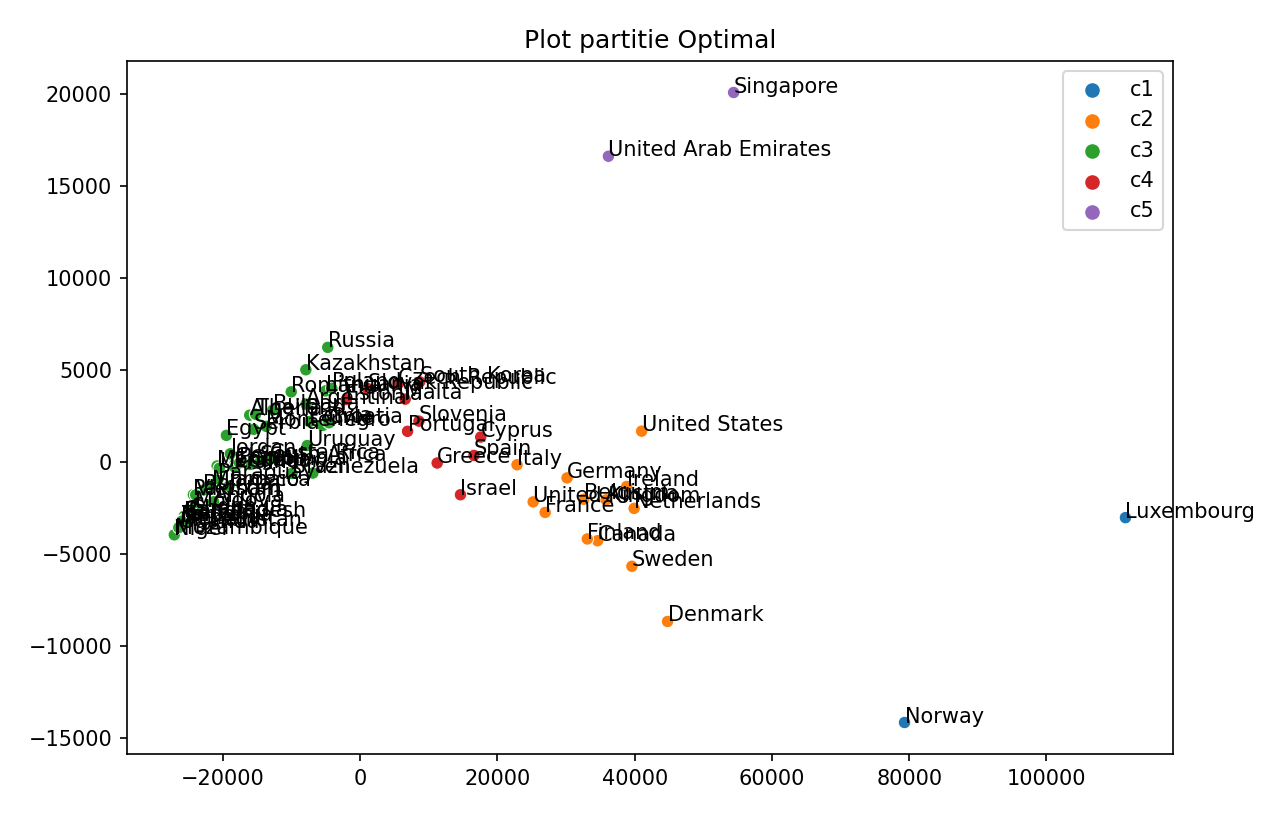


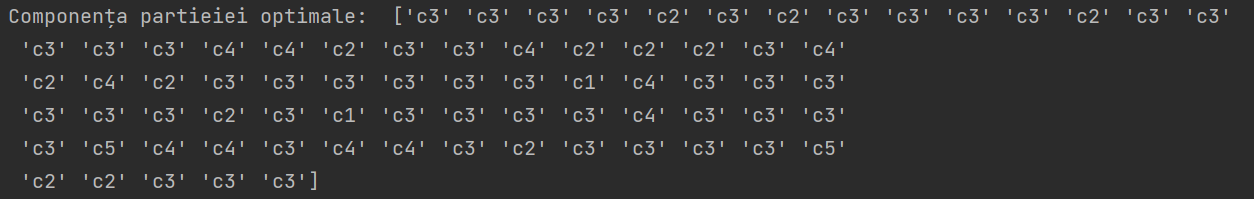
Tarile au fost grupate in 5 clusteri dupa indicatorii considerati. Se poate observa ca tarile dezvoltate au fost grupate impreuna, in timp ce tarile mai putin dezvoltate au fost grupate impreuna( de vazut tarile din Africa).

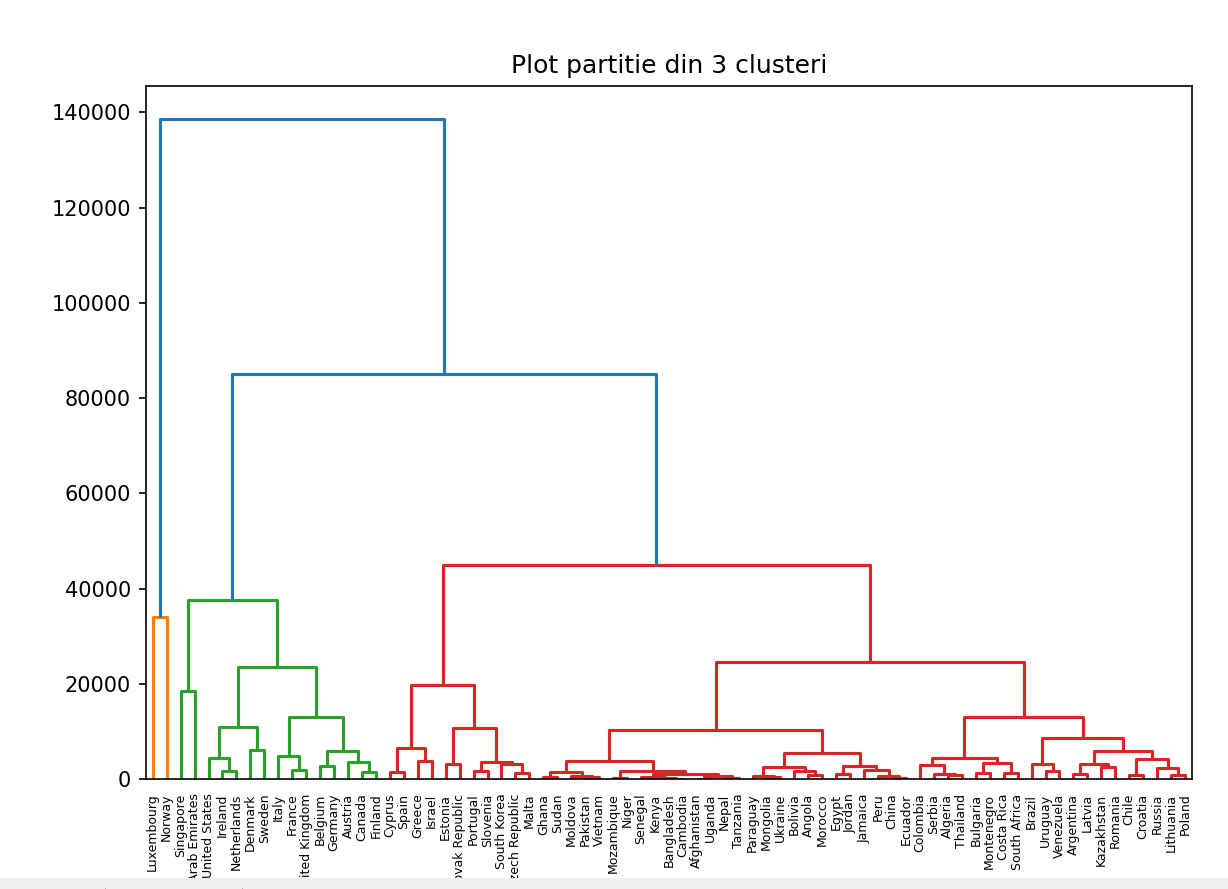


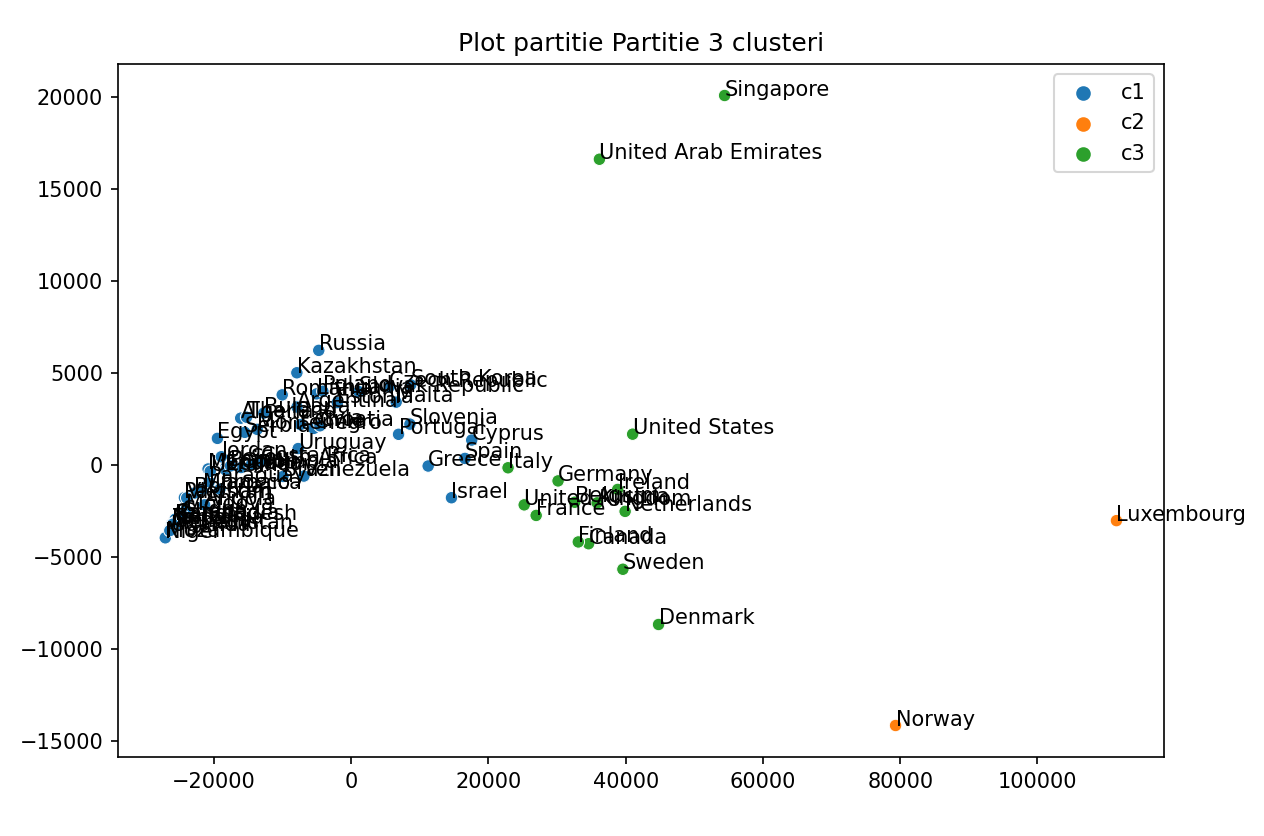
Din histograma putem vedea frecventa valorii variabieli gdpp in fiecare cluster.



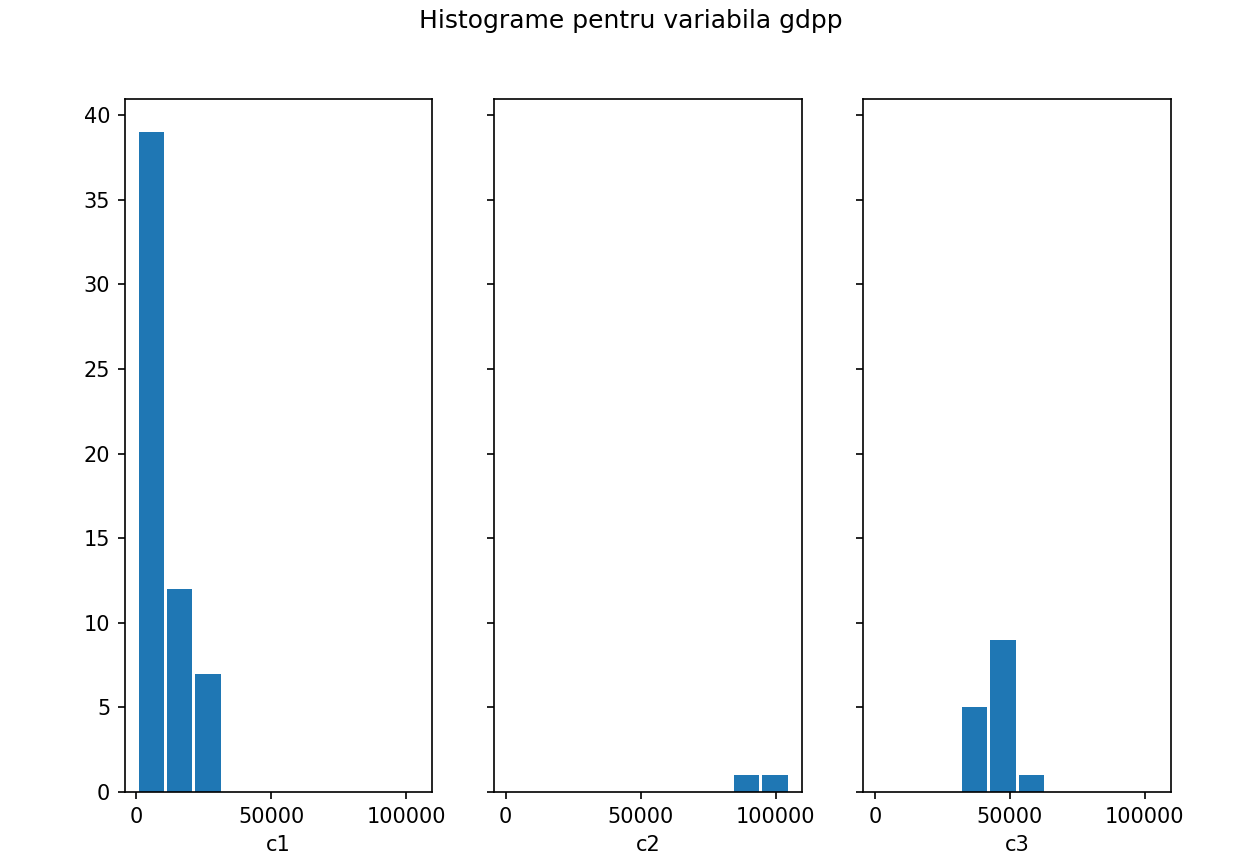
Din plotul pentru partitia optimala se poate observa omogenitatea clusterilor care este corelata cu distantele dintre clusteri in corelograma. Tarile din al treilea cluster sunt grupate intr-o grupa mai omogena decat tarile din primul cluster.

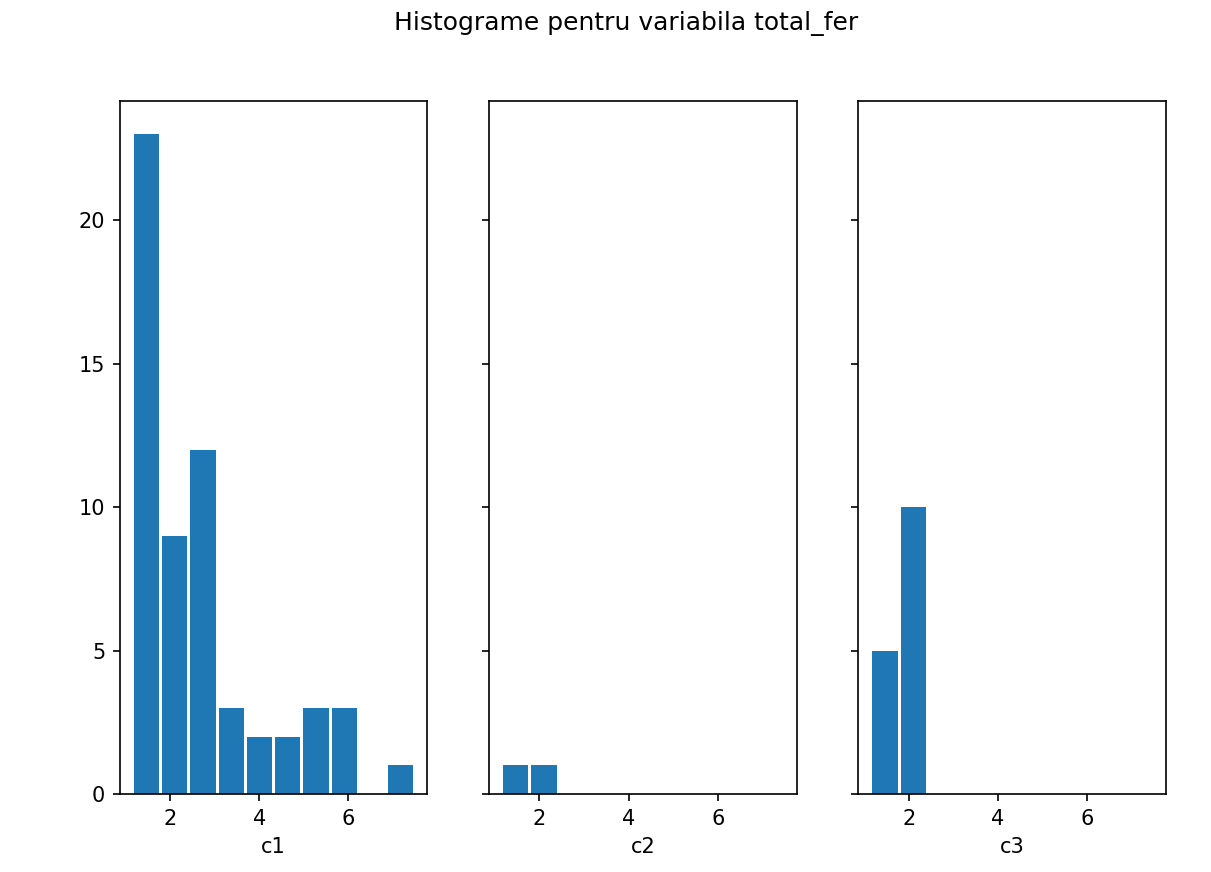




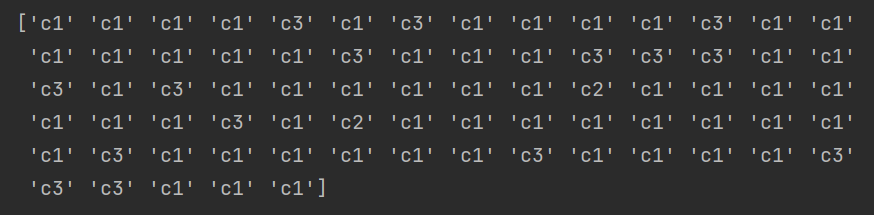


2 histograme pentru partitia din 3 clase.





Componenta partitiei din 3 clusteri:



Concluzii:

Tarile au fost grupate in functie de indicatorii economici prezentati la inceput. Se poate observa faptul ca tarile sunt grupate in functie de nivelul lor de dezvoltare. Tarile nedezvoltate au fost grupate in grupe omogene, avand distante mici in dendograma. Tarile in curs de dezvoltare au fost grupate impreuna, iar tarile dezvoltate au fost grupate impreuna, insa se observa faptul ca distanta in dendograma pentru clusterii care includ tarile dezvoltate este mult mai mare, sugerand diferente mai mari in grupe iar omogenitatea nu este asa de mare.