Logo

Description automatically generated

ACADEMIA DE STUDII ECONOMICE BUCUREȘTI

FACULTATEA DE CIBERNETICĂ, STATISTICĂ ȘI INFORMATICĂ ECONOMICĂ

SPECIALIZAREA INFORMATICĂ ECONOMICĂ

*Dezvoltare Software pentru Analiza Datelor*

**Analiza celor mai populare limbaje de programare între 2004-2022**

**Profesor coordonator:**

**Prof. Univ. Dr. Furtuna Titus Felix**

**Student**

Coman Claudia Ana-Maria

Grupa 1082 C

București

2022

**CUPRINS**

[**Date utilizate** 1](#_Toc120880099)

[**Formatul setului de date** 2](#_Toc120880100)

[**Metodă de analiză** 4](#_Toc120880101)

[**Abordarea privind analiza datelor** 4](#_Toc120880102)

[**Listă de figuri** 9](#_Toc120880103)

# **Date utilizate**

Pentru acest proiect am ales să utilizez setul de date regăsit la adresa <https://www.kaggle.com/datasets/muhammadkhalid/most-popular-programming-languages-since-2004>.

Datele sunt colectate între ani 2004-2022. Datele brute provin de la Google Trends.Google Trends este un site web de la Google care analizează popularitatea [interogărilor de căutare](https://en.wikipedia.org/wiki/Web_search_query) de top în [Căutarea Google](https://en.wikipedia.org/wiki/Google_Search) în diferite regiuni și limbi. Site-ul web utilizează grafice pentru a compara volumul căutărilor diferitelor interogări de-a lungul timpului.

Indexul de popularitate al limbajului de programare PYPL este creat prin analizarea frecvenței de căutare a tutorialelor cu limbaje de programare pe Google.

Cu cât se caută mai mult un tutorial pentru un anumit limbaj, cu atât se presupune că limbajul este mai popular. De asemenea, este un indicator principal pentru frecvența de căutare. Acesta este un bun indicator pentru observarea evoluției a celor mai vechi, dar și a celor mai noi limbaje de programare.

## **Formatul setului de date**

Datele sunt preluate începând cu luna iulie în anul 2004 și până în luna ianuaria a anului 2022.

Urmatoarele limbaje de programare sunt analizate:

|  |
| --- |
| Abap |
| Ada |
| C/C++ |
| C# |
| Cobol |
| Dart |
| Delphi/Pascal |
| Go |
| Groovy |
| Haskell |
| Java |
| JavaScript |
| Julia |
| Kotlin |
| Lua |
| Matlab |
| Objective-C |
| Perl |
| PHP |
| Python |
| R |
| Ruby |
| Rust |
| Scala |
| Swift |
| TypeScript |
| VBA |
| Visual Basic |

# **Metodă de analiză**

## **Abordarea privind analiza datelor**

Primii pași în conceperea unei analize sunt constituiți de pregătirea unui script corespunzător (precum cel de la seminar), care conține fișiere .py cu funcții, grafice, toate integrate și apelate în fișierul main. De asemenea, am importat și principalele librarii precum: numpy, pandas și matplotlib pentru desenarea graficelor.

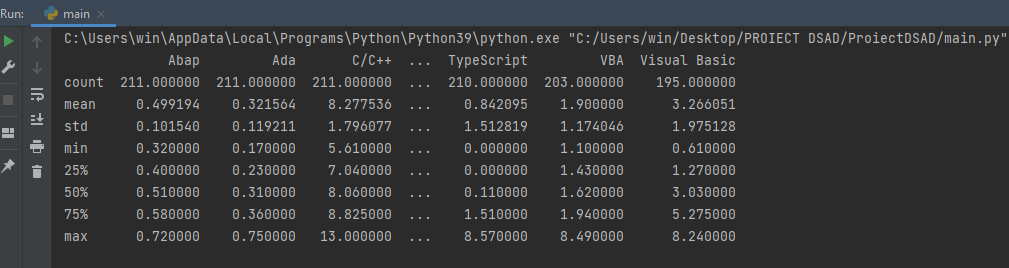
Primul pas în analiza datelor a fost să citesc datele din csv.

dataset=pd.read\_csv("Most Popular Programming Languages from 2004 to 2022.csv")

Ulterior, am convertit tipul de dată în datetime pentru că era de tipul string, iar ca să pot face prelucruri am nevoie de date.

dataset["Date"]=pd.to\_datetime(dataset["Date"])  
dataset.set\_index("Date",inplace=True)

Cu ajutorul metodei describe, am generat date din statistica descriptiva, in momentul in care acestea sunt calculate, nu se adauga si valorile nan.



De asemenea, am preluat indexii, coloanele și valorile numerice, pentru o prelucrare mai ușoară.

data=list(dataset.index)  
#print(data)  
  
nume\_limbaje=list(dataset.columns)  
#print(nume\_limbaje)  
  
x=dataset.values  
#print(x)

Pentru o buna interpretare a rezultatelor, am identificat rândurile ce conțin valorile NAN și respectiv poziția acestora, am calculat media pe acel rând și am înlocuit valorile NAN cu media.

#returneaza o matrice de valori  
m,n=x.shape  
#print(x[0,:])  
  
#eliminare valori lipsa  
is\_nan=pd.isna(x)  
#print(is\_nan)  
  
#aflu indexul fiecarei valori nan  
k=np.where(is\_nan)  
#print(k)  
  
#media pe randuri  
x[k]=np.nanmean(x[:,k[1]],axis=0)  
print(x[3,:])

Mi-am împărțit setul de date în 2 subseturi din anul 2004 și 2022 ce cuprind cele mai populare limbaje de programare și valoarea asociată si le-am salvat in 2 fisiere de tip CSV.

popular\_2004 = pd.DataFrame({"Languages": dataset.iloc[0].T.index,  
 "Popularity": dataset.iloc[0].values.T})  
  
  
popular\_2022 = pd.DataFrame({"Languages": dataset.iloc[210].T.index,  
 "Popularity": dataset.iloc[210].values.T})  
  
popular\_2004=popular\_2004.to\_csv("Most popular languages in 2004")  
popular\_2022=popular\_2022.to\_csv("Most popular languages in 2022")

De asemenea, am grupat si sumarizat datele in functie de popularitatea fiecarui limbaj de programare.

pop\_lang=pd.read\_csv("Most popular languages in 2022")  
t1=dataset.merge(right=pop\_lang,left\_index=True,right\_index=True)  
popularity=t1[var + ["Popularity"]].groupby(by="Popularity").agg(sum)

Primul grafic este de tipul lineplot si am calculat media pentru intreg setul de date si am preluat doar acele limbaje de programare care au media > 2.5, astfel mi-am putut da seama care sunt cele mai populare.

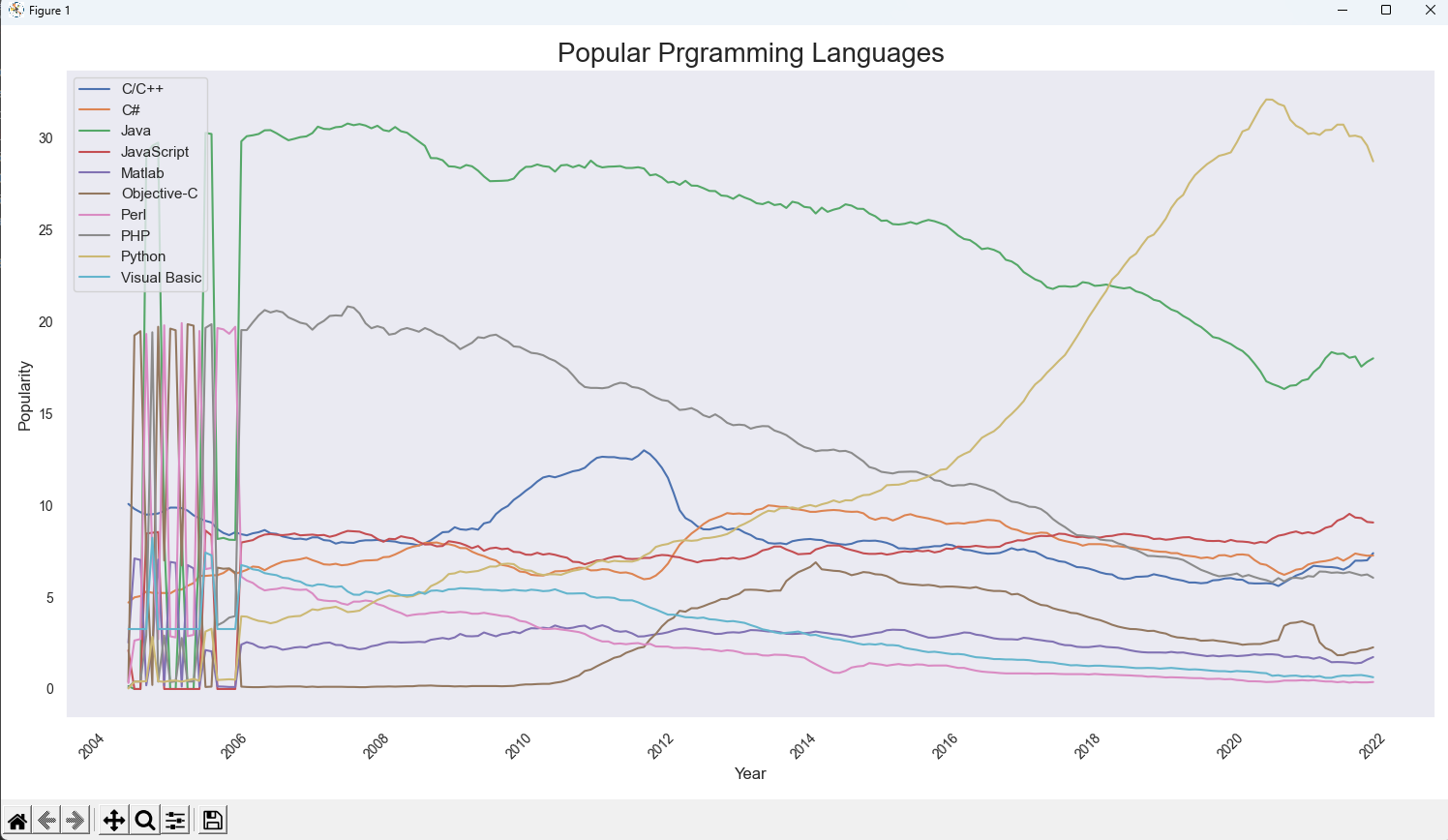


Figura 1. Lineplot reprezentand principalele 10 limbaje de programare

Si mi-a returnat principalele 10 limbaje de programare. Se poate observa din lineplot faptul ca Python este in prezent cel mai popular limbaj de programare.

Cel de-al doilea grafic este de tip horizontal barchart si surprinde popularitatea tuturor limbajelor de programare din ianuarie 2022.

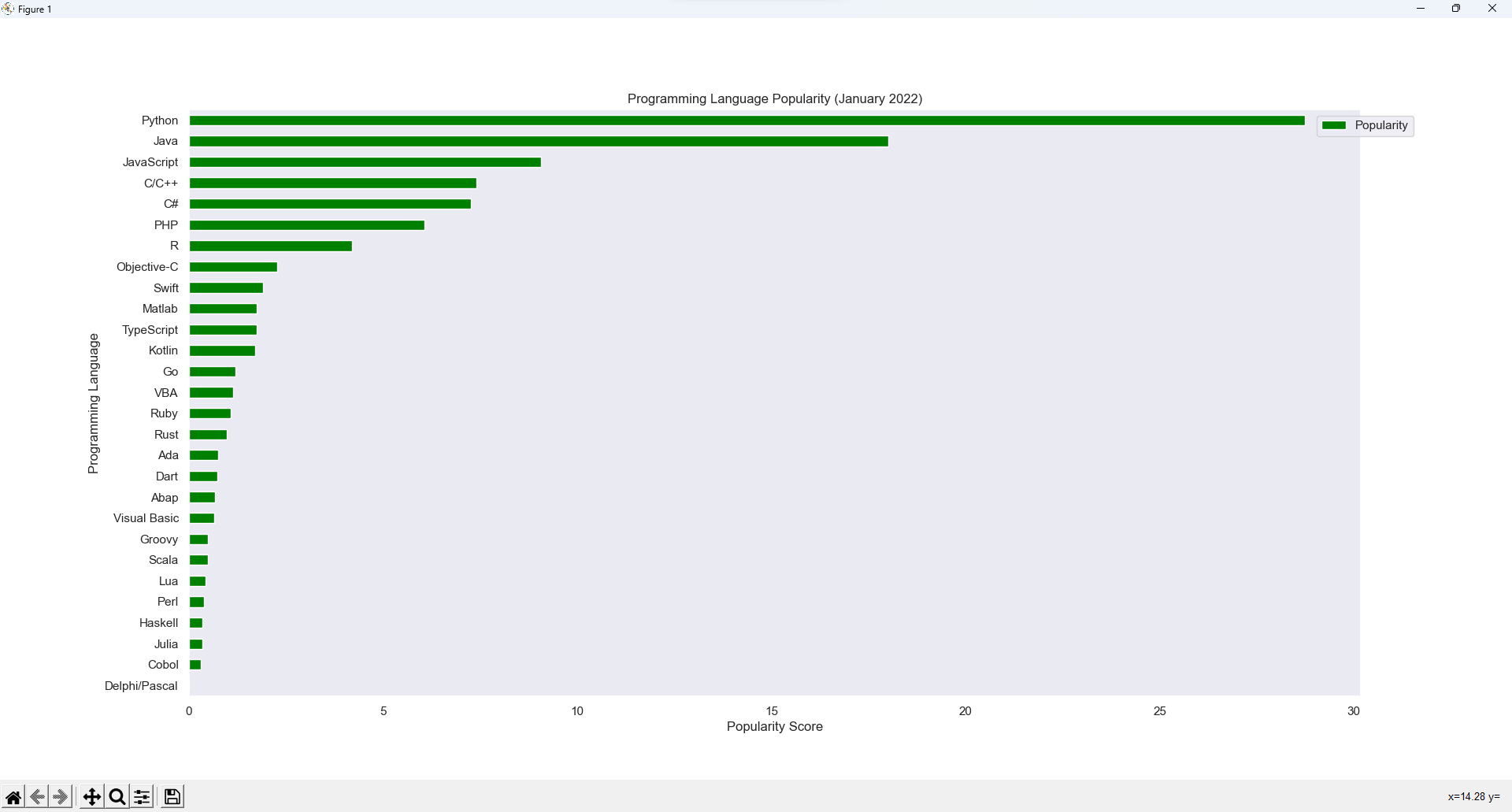


Figura 2. Horizontal Barchart privind scorul fiecarui limbaj de programare

In ultima parte am facut inca 2 grafice, primul fiind tot de tipul horizontal barchart, iar al doilea este un piechart, ambele reprezentand procentele mediei de popularitate ale principalelor limbaje de programare.

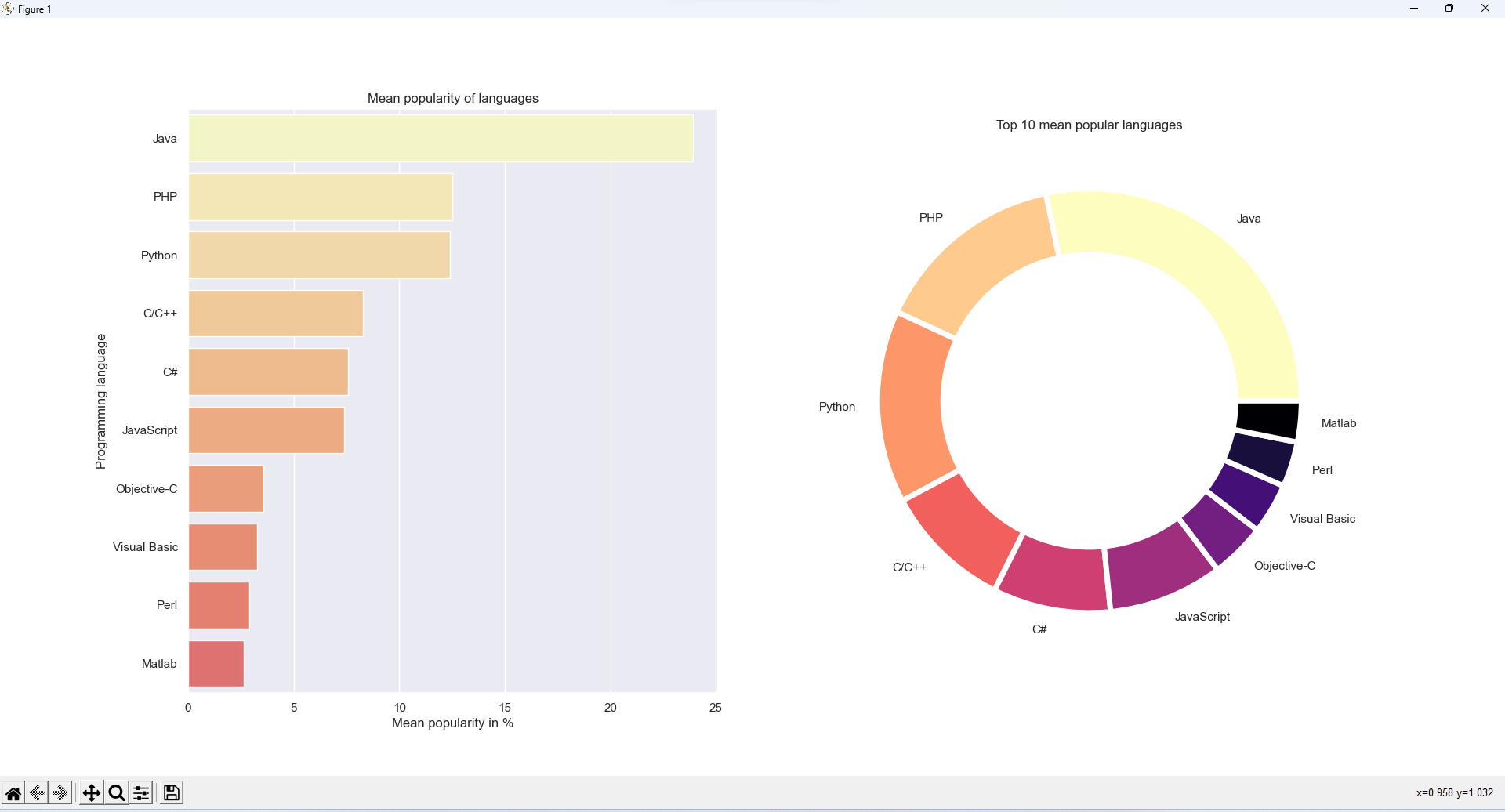


Figura 3. Horizontal Barcahart si piechart privind media popularitatii ale principalelor limbaje de programare exprimate in procente

## **Listă de figuri**

Figura 1. Lineplot reprezentand principalele 10 limbaje de programare..............................................7

Figura 2. Horizontal Barchart privind scorul fiecarui limbaj de programare.........................................8

Figura 3. Horizontal Barcahart si piechart privind media popularitatii ale principalelor limbaje de programare exprimate in procente.......................................................................................................9