**Statistici globale YouTube**

**Borș Cristian, IA-211, III**

*Universitatea Tehnică a Moldovei, Facultatea Calculatoare, Inginerie și Micro-Electronică, Chișinău, Moldova, Informatica aplicată.*

**Abstra­ct**

În era digitală actuală, analiza datelor devine din ce în ce mai importantă în diverse domenii, iar YouTube joacă un rol semnificativ în acest context. YouTuberii au devenit influențe puternice în cultura digitală globală. Scopul analizei noastre a fost să identificăm factorii care contribuie la succesul canalelor YouTube la nivel global, privind la numărul de abonați, vizualizări, țara de origine, categoria de conținut și platforma de distribuție.

Folosind un set de date actualizat până în 2023 de pe Kaggle, am eliminat datele irelevante și valorile lipsă prin preprocesarea cu R. Analiza exploratorie a inclus vizualizări ale distribuției canalelor și a numărului de abonați în funcție de țara de origine. Modelând cu regresia liniară, am observat corelații semnificative între variabile, cum ar fi influența puternică a Statelor Unite și Indiei în popularitatea canalelor YouTube.

Categoriile de conținut au arătat că divertismentul este cel mai popular, urmat de muzică și comedie, iar analiza veniturilor a indicat că canalele de divertisment generează cele mai mari venituri. Cu toate acestea, predicțiile pentru numărul de abonați au avut o precizie limitată, sugerând că variabilele precum țara de origine nu explică în totalitate variația.

# Introducere

Platforma YouTube a devenit un mediu fascinant și dinamic pentru crearea și distribuirea conținutului video la nivel global. Un segment esențial al acestei platforme îl reprezintă creatorii de conținut, cunoscuți sub numele de YouTuberi, care au reușit să construiască audiențe impresionante și să influențeze culturile din diverse domenii. În acest context, propunem o analiză detaliată a unui set de date relevant, focalizat pe statistici globale ale unor dintre cei mai influenți YouTuberi.

YouTuberii din această analiză reprezintă un spectru divers de genuri, inclusiv muzică, divertisment, educație și sport, iar datele cuprinzătoare ne oferă o privire în profunzime asupra performanțelor lor. Analizând variabile precum numărul de abonați, vizualizări ale videoclipurilor, țara de proveniență și categoria de conținut, vom încerca să identificăm tendințe semnificative și să extragem concluzii relevante.

Fiecare înregistrare din setul de date reprezintă un YouTuber distinct, furnizând informații precum numărul de abonați, vizualizările videoclipurilor, țara de origine, categoria de conținut și multe altele. Prin explorarea acestor date, vom căuta să înțelegem ce factori contribuie la succesul unui canal YouTube și cum aceștia variază în funcție de genul de conținut, țară sau chiar platforma de origine.

În plus, vom investiga aspecte precum evoluția abonaților și vizualizărilor în ultimele 30 de zile, veniturile lunare și anuale ale creatorilor, oferind o perspectivă asupra impactului financiar și cultural al acestor personalități digitale.

Prin această analiză, ne propunem să aducem lumină asupra dinamicii și diversității peisajului YouTube global, contribuind la înțelegerea modului în care acești YouTuberi influențează și interacționează cu publicul lor într-un mediu digital în continuă evoluție.

Ipoteza: **Există corelații semnificative între variabilele precum numărul de abonați, vizualizările videoclipurilor, țara de origine, categoria de conținut și platforma de distribuție, influențând succesul unui canal YouTube.**

Această ipoteză presupune că diferite aspecte, cum ar fi popularitatea în funcție de țară, genul de conținut sau platforma de distribuție, pot avea un impact semnificativ asupra creșterii audienței și a succesului financiar al unui canal YouTube. Prin analizarea corespunzătoare a datelor din setul de date, putem verifica și evidenția aceste corelații, contribuind astfel la o înțelegere mai profundă a dinamicii din spatele succesului YouTuberilor la nivel global.

# Materiale & Metode

**Setul de date**

Pentru realizarea acestei analize, am folosit un set de date cuprinzător care furnizează informații actualizate despre cele mai pupulare canale de YouTube la nivel Global. Acest set de date a fost luat de pe platformă open-source Kaggle.com, care contine date pana in anul 2023.

Variabile Incluse în Setul de Date - mai multe detalii vedeți Anexa1.

Setul de date referitor la YouTuberi furnizează o panoramă detaliată asupra performanței unor creatori de conținut pe platforma YouTube la nivel global. Fiecare intrare din set oferă informații esențiale despre YouTuberi, inclusiv numărul de abonați, vizualizări ale videoclipurilor, țara de origine, categoria de conținut, numărul de încărcări, tipul de canal și multe altele.

**Preprocesarea Datelor**

Pentru a pregăti datele pentru analiză, am folosit limbajul de programare R și bibliotecile openintro, tidyverse, ggplot și altele. Am efectuat o analiză sumară pentru a înțelege structura și tipurile de date existente. Am observat că setul conține două tipuri de date, dbl și chr. Setul de date este alcătuit din 996 rânduri și 28 coloane. La fel setul de date conține multe valori NA ceea ce complică analiza acestor date. Din această cauză am eliminat aceste rânduri pentru o analiza cât mai corectă. Deoarece setul de date conține o cantintate mare și diversă de informații am eliminat coloanele care nu se incadrează în efectuarea analizei și determinarea ipotezei. Aceste coloane sunt: Title, created\_year, created\_month, created\_year, Gross tertiary education enrollment (%), Population, Unemployment rate, Latitude, Longitude.

**Analiza Exploratorie a Datelor**

Am realizat o serie de vizualizări pentru a explora relațiile dintre diferitele caracteristici ale canalelor de Youtube și performanța acestora pe platforma YouTube. Am analizat Distribuția Categoriilor de Canale, Numărul de Canale în Funcție de Țara de Origine, Distribuția Numărului de Abonați pe Categorii și altele. Pentru a explora aceste relații am folosit ggploturi pentru a vedea grafic cum aceste date sunt distribuite și relațiile între ele. Am utilizat funcția summary pentru a rezuma datele. Această funcție a oferit informații statistice de bază pentru fiecare variabilă din setul de date și a oferit o privire asupra caracteristicilor cheie ale setului de date.

**Modelarea și Evaluarea Performanței**

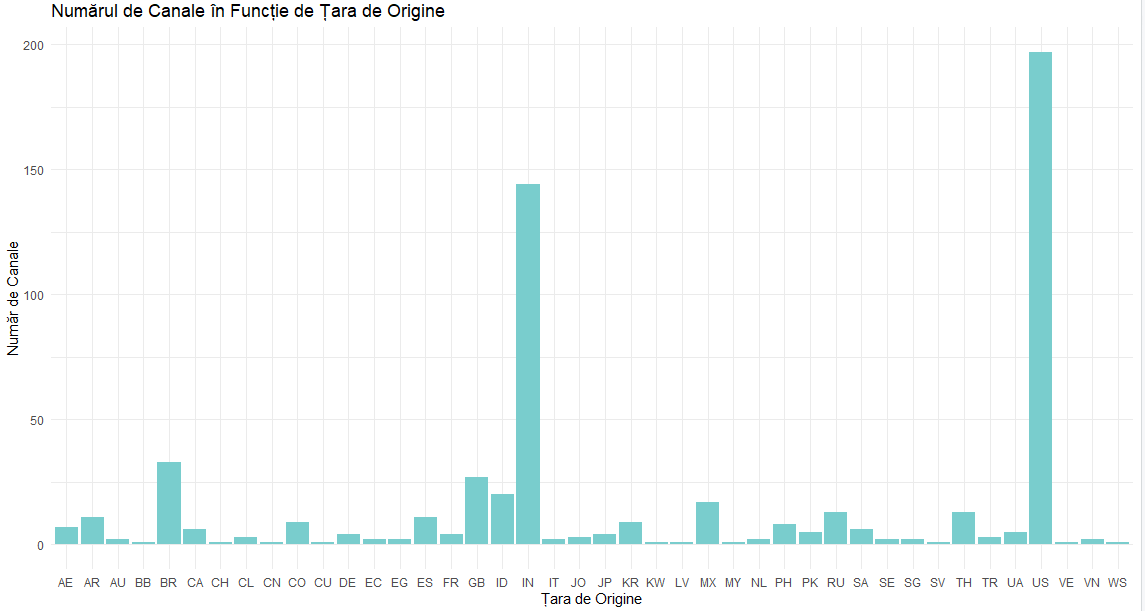
În cadrul acestei etape a analizei datelor privind statisticile globale ale platformei YouTube, m-am focusat asupra modelării și evaluării performanței. Scopul meu a fost să explorez relațiile dintre variabilele cheie, cu accent pe vizualizari, numarul de incărcări și numărul de abonați. Am împărțit setul de date în două segmente distincte: unul pentru antrenament și unul pentru testare. Modelul de regresie liniară a fost implementat, efectuând 3 modele: regresie liniara simpla, regresie liniara multipla și regresia liniara cu toate efectele principale.

**Codul disponibil**

Tot codul pentru crearea modelului pentru regresie liniară este disponibil pe link-ul de mai jos :

# Rezultate

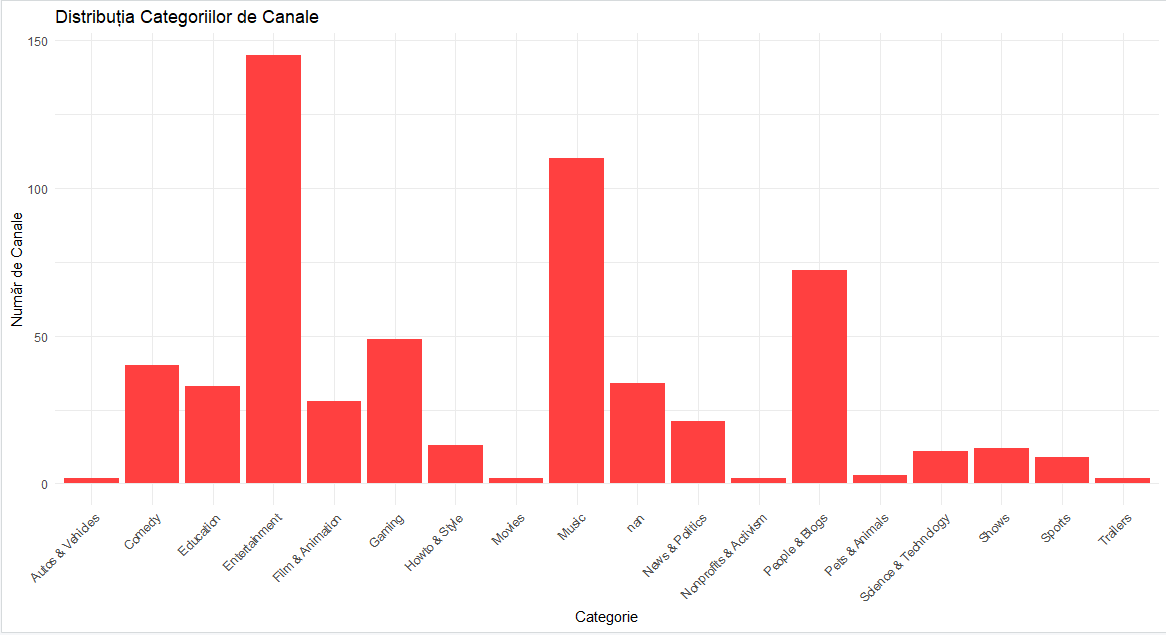
Cu ajutorul ggploturilor am analizat și observat Numărul de Canale în Funcție de Țara de Origine (fig. 1.1). Aceste date vor ajută să înțeleg influența originilor canalelor de YouTube asupra popularitatii acestora.



**Figura 1.1.** Numărul de Canale în Funcție de Țara de Origine

Cu ajutorul acestui grafic putem observa ca țările precum Statele Unite și India predomina, astfel ne putem da seama ca in aceste 2 țări sunt o mare parte din cele mai populare canale de YouTube.

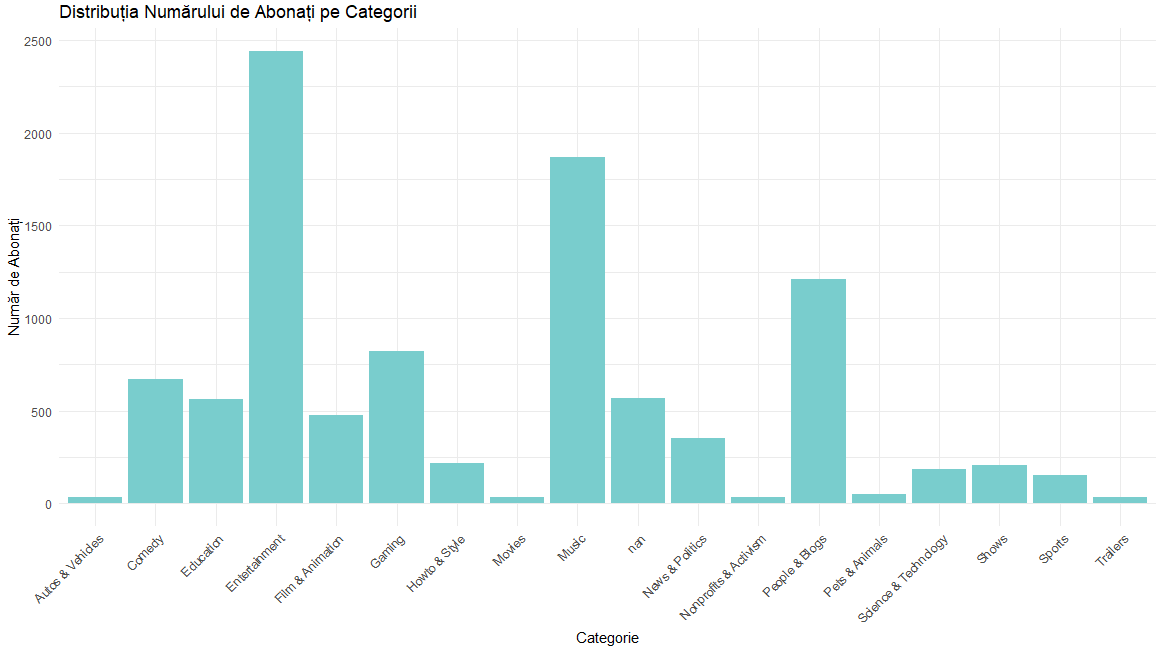
În continuare am analizat distribuția categoriilor de canale de pe platforma YouTube (fig. 1.2).



**Figura 1.2**. Distribuția categoriilor de canale

Graficul prezentat prezintă distribuția categoriilor de canale de pe platforma YouTube. Graficul este o histogramă, cu categoriile de canale pe axa orizontală și numărul de canale pe fiecare categorie pe axa vertical .Cele mai multe canale de pe YouTube sunt clasificate în categoria **Entertainment**, cu 58 de canale. Următoarele categorii cele mai populare sunt **Music**, cu 37 de canale, **Comedy**, cu 32 de canale, și **Autos & Vehicles**, cu 29 de canale. Cele mai puține canale de pe YouTube sunt clasificate în categoria **Nonprofits & Activism**, cu doar două canale. Alte categorii cu un număr mic de canale sunt **Howto & Style**, cu șase canale, **Pets & Animals**, cu șapte canale, și **Trailers**, cu opt canale.

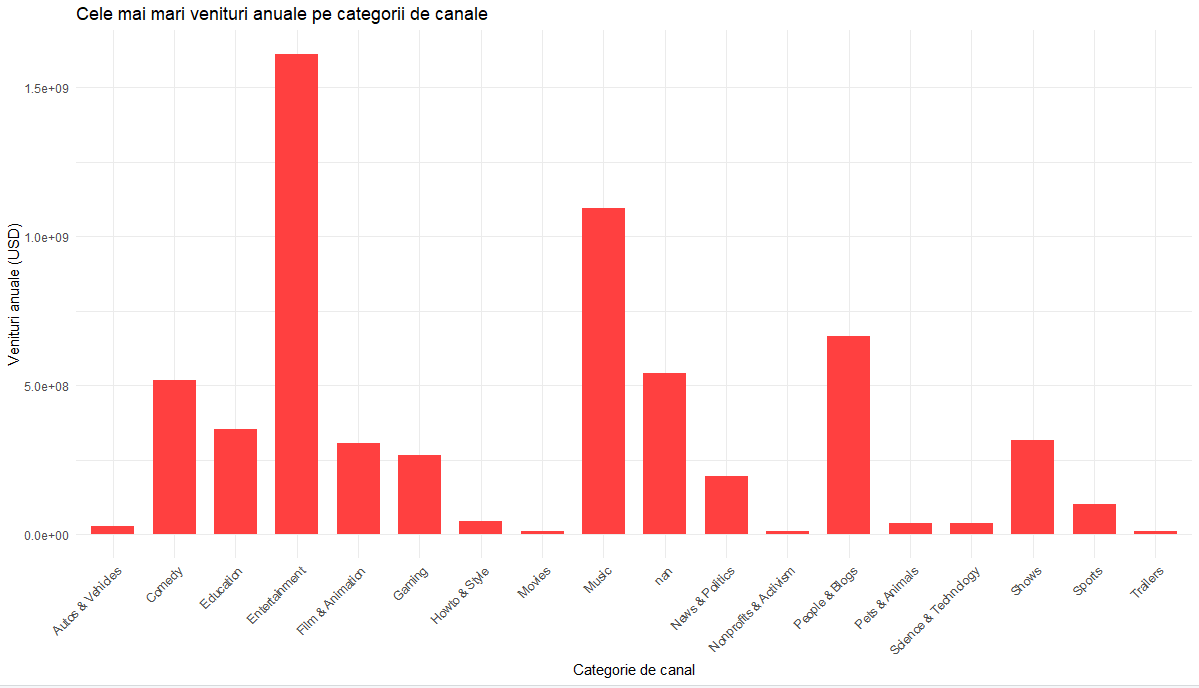
Am creat un grafic care arată distribuția numărului de abonați pe categorii de canale de pe platforma YouTube. Graficul este o histogramă, cu categoriile de canale pe axa orizontală și numărul de canale pe fiecare categorie pe axa vertical (fig. 1.3).



**Figura 1.3**. Distribuția numărului de abonați pe categorii de canale

Cele mai multe canale de pe YouTube sunt clasificate în categoria **Entertainment**, cu 58 de canale. Următoarele categorii cele mai populare sunt **Music**, cu 37 de canale, **Comedy**, cu 32 de canale, și **Autos & Vehicles**, cu 29 de canale. Cele mai puține canale de pe YouTube sunt clasificate în categoria **Nonprofits & Activism**, cu doar două canale. Alte categorii cu un număr mic de canale sunt **Howto & Style**, cu șase canale, **Pets & Animals**, cu șapte canale, și **Trailers**, cu opt canale. Graficul poate fi folosit pentru a înțelege mai bine ce tipuri de conținut sunt populare pe YouTube. De asemenea, poate fi folosit pentru a identifica oportunități pentru creatorii de conținut care doresc să-și lanseze canalele pe YouTube.

Mai jos este reprezentat un grafic care prezintă distribuția veniturilor anuale pe categorii de canale de pe platforma YouTube (fig. 1.4). Graficul este o histogramă, cu categoriile de canale pe axa orizontală și veniturile anuale în milioane de dolari pe fiecare categorie pe axa verticală.



**Figura 1.4**. Distribuția veniturilor anuale pe categorii de canale

Cele mai mari venituri anuale pe categorii de canale de pe YouTube sunt înregistrate în categoria **Entertainment**, cu 18,5 miliarde de dolari. Următoarele categorii cele mai profitabile sunt **Music**, cu 10,2 miliarde de dolari, **Autos & Vehicles**, cu 5,5 miliarde de dolari, și **Gaming**, cu 4,2 miliarde de dolari. Cele mai mici venituri anuale pe categorii de canale de pe YouTube sunt înregistrate în categoria **Nonprofits & Activism**, cu doar 100 de milioane de dolari. Alte categorii cu venituri mici sunt **Howto & Style**, cu 200 de milioane de dolari, **Pets & Animals**, cu 250 de milioane de dolari, și **Trailers**, cu 300 de milioane de dolari. Cele mai profitabile categorii de canale sunt cele care atrag un număr mare de spectatori, cum ar fi conținutul de divertisment, muzică și jocuri.

Pentru a putea prezice numarul de subscriberi în dependență de categorie, țară și numărul de subscriberi am folosit modelul de regresie liniară. Pentru aceasta am factorizat varaibilele Country și category. Acest pas este important atunci când lucrăm cu variabile categorice în modelele statistice.După am construit un model de regresie liniar pentru a prezice variabila "subscribers" pe baza variabilelor "video\_views", "Country" și "category". Am ales metoda de modelare ‘lm’, am efectuat validare încrucișată (cross-validation) pe modelul de regresie liniar și am utilizat funcția train() din pachetul caret pentru a evalua performanța modelului folosind 10 fold-uri pentru cross-validation.

Utilizând funcția summary() am obținut următaorele rezultate:

Min 1Q Median 3Q Max

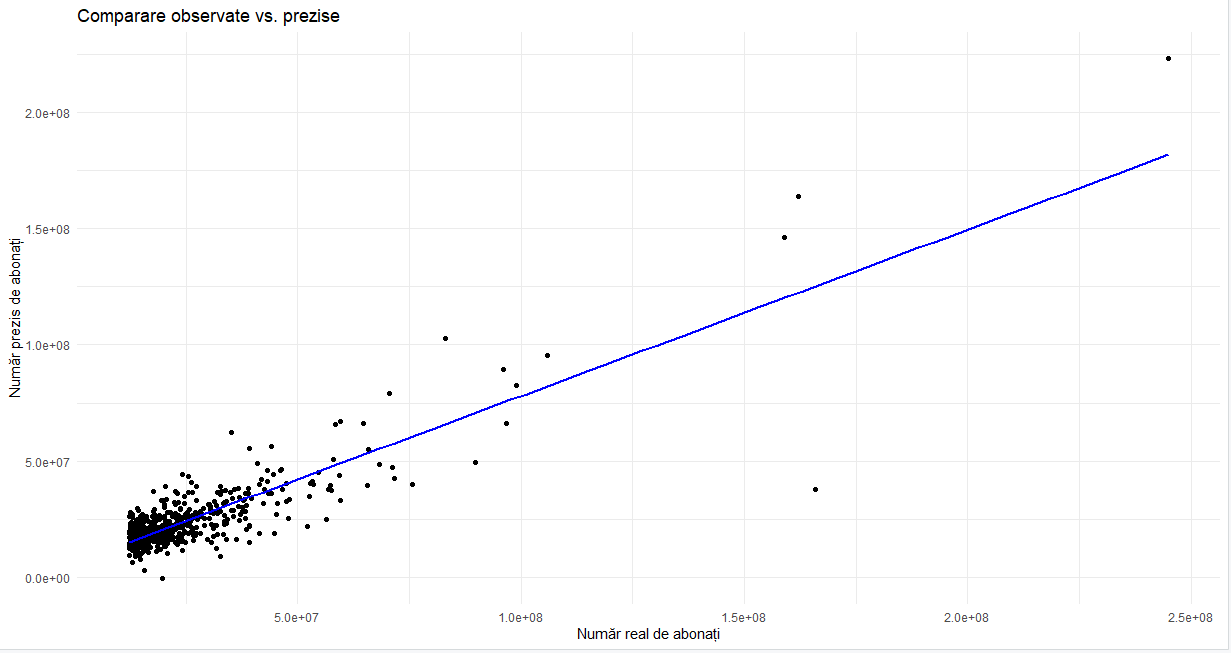
-30347533 -4295269 -953933 2807266 126497907

RMSE Rsquared MAE

9150899 0.6641765 6501350

Cu toate acestea, valoarea înaltă a erorii medii pătratice (RMSE) de 9,150,899 și a erorii absolute medii (MAE) de 6,501,350 sugerează că modelul prezis nu se apropie foarte bine de valorile observate. Prin urmare, prezicerea numărului de abonați nu va fi una exactă sau precisă, iar modelul poate avea o aproximare mai scăzută în comparație cu datele reale.

Coeficientul de determinare (R-squared) este de 0.6641765, ceea ce indică că aproximativ 66.42% din variația numărului de abonați poate fi explicată de variabilele independente incluse în model.  
Pentru a crea un grafic pentru modelul meu de regresie liniar și predicțiile obținute în urma validării încrucișate, am utiliza pachetul ggplot2 (fig 1.5.).



**Figura 1.5.** Comparare observate vs. Prezise

Acest grafic afișează observațiile reale versus predicțiile modelului, împreună cu linia de regresie liniară. Dacă datele se apropie de această linie, înseamnă că modelul face predicții bune.

# Bibliografie

1. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL [https://www.R-project.org/](https://www.r-project.org/). – accesat (10.09.2023)
2. Dataset: <https://www.kaggle.com/datasets/nelgiriyewithana/global-youtube-statistics-2023> - sursa setului de date utilizat în analiză, disponibil pe Kaggle. – accesat (13.09.2023)
3. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning. Springer. – accesat (19.10.2023)
4. OpenIntro Statistics. (2019). [openintro: OpenIntro data sets and supplemental functions](https://cran.r-project.org/package=openintro). – accesat (04.11.2023)
5. ResearchGate - <https://www.researchgate.net/figure/A-Histogram-of-distribution-of-residuals-of-linear-regression-investigating-estimated_fig2_354734110> - accesat 27.11.2023)

**Materiale Suplimentare**

Anexa 1.

1. **rank:** Poziția canalului YouTube în funcție de numărul de abonați
2. **Youtuber:** Numele canalului YouTube
3. **subscribers**: Numărul de abonați la canal
4. **video views:** Vizualizări totale ale tuturor videoclipurilor de pe canal
5. **category:** Categorie sau nișă a canalului
6. **Title:** Titlul canalului YouTube
7. **uploads:** Numărul total de videoclipuri încărcate pe canal
8. **Country:** Țara de origine a canalului YouTube
9. **Abbreviation:** Abrevierea țării
10. **channel\_type:** Tipul canalului YouTube (de exemplu, individual, brand)
11. **video\_views\_rank**: Clasamentul canalului în funcție de vizualizările totale ale videoclipurilor
12. **country\_rank:** Clasamentul canalului în funcție de numărul de abonați în țara sa
13. **channel\_type\_rank:** Clasamentul canalului în funcție de tipul său (individual sau brand)
14. **video\_views\_for\_the\_last\_30\_days**: Vizualizări totale ale videoclipurilor în ultimele 30 de zile
15. **lowest\_monthly\_earnings**: Câștigul lunar estimat minim al canalului
16. **highest\_monthly\_earnings**: Câștigul lunar estimat maxim al canalului
17. **lowest\_yearly\_earnings:** Câștigul anual estimat minim al canalului
18. **highest\_yearly\_earnings:** Câștigul anual estimat maxim al canalului
19. **subscribers\_for\_last\_30\_days**: Numărul de abonați noi câștigați în ultimele 30 de zile
20. **created\_year:** Anul în care a fost creat canalul YouTube
21. **created\_month:** Luna în care a fost creat canalul YouTube
22. **created\_date:** Data exactă a creării canalului YouTube
23. **Gross tertiary education enrollment (%):** Procentul populației înscrisă în învățământul terțiar în țară
24. **Population:** Populația totală a țării
25. **Unemployment rate:** Rata șomajului în țară
26. **Urban\_population:** Procentul populației care trăiește în zonele urbane
27. **Latitude:** Coordonata de latitudine a locației țării
28. **Longitude:** Coordonata de longitudine a locației țării