# Tendințe, relații și modele în datele muzicale

**Malicic Dumitru, IA-211, III**

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

# INTRODUCERE

Într-o lume în care melodiile devin monedă curentă și preferințele muzicale reprezintă un fragment esențial al identității noastre, explorarea datelor Spotify din 2022 devine o călătorie înțeleaptă în inima și sufletul ascultătorilor. Cu fiecare clic, fiecare playlist și fiecare skip, utilizatorii lasă o amprentă digitală unică, transformând platforma într-o comoară de informații care așteaptă să fie dezvăluită și înțeleasă.

Cu ajutorul instrumentelor avansate de analiză oferite de R Studio, am adâncit în acest ocean de date pentru a descifra narațiunile subiacente și pentru a aduce la lumină conexiuni fascinante între diversele aspecte ale universului muzical. De la ecouri ale preferințelor globale la nuanțe locale de impact, analiza noastră se întinde ca o partitură muzicală digitală, dezvăluind notele ascunse ale unei perioade muzicale pline de inovație și diversitate.

În cadrul acestei călătorii analitice, ne propunem să nu doar privim în urmă și să înțelegem momentul muzical din 2022, ci și să anticipăm direcțiile viitoare ale industriei. Prin conexiunile pe care le vom face și prin semnificațiile pe care le vom deduce, sperăm să conturăm nu doar o analiză, ci o poveste captivantă a evoluției muzicii și a ascultătorilor săi.

Prin intermediul acestui articol, vă invităm să vă alăturați nouă în această călătorie analitică, să explorăm împreună subtextul cifrelor și să descoperim poveștile care se desfășoară în spatele melodiilor noastre preferate. Fie că suntem pasionați de ritmurile animate ale pop-ului sau de vibrațiile subtile ale genurilor de nișă, datele Spotify ne oferă o cheie de acces către universul fascinant și în continuă schimbare al muzicii moderne. Să începem această călătorie muzicală în lumea numerelor și să descoperim, împreună, misterele ascunse ale anului 2022.

Scopul acestei analize a setului de date Spotify din 2022, realizată prin intermediul R Studio, constă în dezvăluirea și înțelegerea subiacentă a preferințelor muzicale ale utilizatorilor platformei. Prin examinarea detaliată a variabilelor cum ar fi genul muzical, popularitatea pieselor, regiunile geografice și caracteristicile audio, ne propunem să identificăm tendințele majore, să conturăm profilele audienței și să capturăm diversitatea sonoră a anului trecut. Această analiză nu doar aduce la lumină poveștile ascunse ale datelor Spotify, ci și oferă o privire către direcțiile posibile ale evoluției muzicale viitoare, contribuind astfel la înțelegerea mai profundă a preferințelor și influențelor care modelează peisajul sonor al ascultătorilor contemporani.

# Materiale & Metode

Datele luate pentru acest articol dar anume Spotify(1) Tracks Dataset(2) cu melodiile din 2022 au fost luate de pe site-uri open-source dar anume de pe Google DataSet Search(3) unde este o biblioteca mare de seturi de date iar codul cu care am lucrat este pe github(4). Așa cum noi dorim să vedem toate piesele populare, în data setul nostru majoritatea pieselor nu sunt populare, pentru aceasta am filtrat setul pentru a obține piesele cele mai populare (mai sus de 80) pentru a fi mai comod de lucrat cu el în continuare.

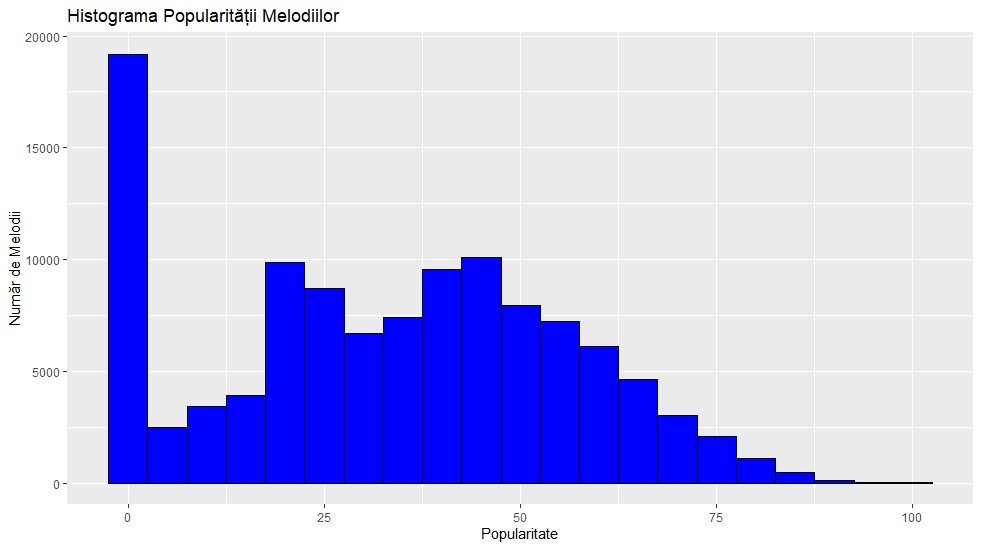
Când privim datele Spotify din 2022, vedem o imagine variată a preferințelor muzicale la nivel mondial. Artiștii și titlurile pieselor, împreună cu genurile muzicale, oferă o privire colorată în diversitatea muzicală. Popularitatea numerică a pieselor arată ce melodii au fost preferate, iar durata pieselor ne spune cât de lungi sau scurte le preferă ascultătorii. Data lansării și regiunea geografică dezvăluie nuanțe culturale și cronologice, iar numărul de ascultări și like-uri arată cât de mult au impresionat piesele. Caracteristicile audio adaugă o notă tehnologică, explorând detalii sonore. Prin analiza acestor date, ne propunem să extragem povești muzicale și să oferim o înțelegere comprehensivă a preferințelor care au definit anul muzical 2022 pentru ascultătorii Spotify.

Iată câteva direcții pe care le-am explorat cu aceste date:

1. **Identificarea Tendințelor Muzicale:**
   * Am putut analiza popularitatea genurilor muzicale și identifica tendințele care au dominat anul 2022 și cum s-au schimbat preferințele ascultătorilor de-a lungul timpului.
2. **Profilarea Artiștilor și Pieselor de Succes:**
   * Am identificat artiștii și piesele care au înregistrat cel mai mare succes în funcție de popularitate, numărul de ascultări și like-uri și ce elemente ale acestor creații le-au propulsat în atenția publicului.
3. **Corelații Temporale și Evenimente:**
   * Datele care le-am studiat în contextul temporal am identificat evenimente sau perioade care au avut un impact semnificativ asupra preferințelor muzicale și am investigat cum anumite lansări sau evenimente au influențat popularitatea anumitor piese.
4. **Analiza Caracteristicilor Audio:**
   * Caracteristicile audio (cum ar fi nivelul de energie, dansabilitatea etc.), explorează cum aceste caracteristici contribuie la popularitatea unei piese sau la preferințele ascultătorilor.
5. **Modelarea Preferințelor Ascultătorilor:**
   * Utilizând tehnici de analiză a datelor am modelat preferințele ascultătorilor și am creat segmentări de audiență și identificat caracteristicile care influențează în mod semnificativ alegerile muzicale ale utilizatorilor Spotify.
6. **Prezicerea Tendințelor Viitoare:**
   * Am utilizat datele istorice pentru a încerca să anticipez tendințe viitoare în industria muzicală. Această analiză predictivă mi-a oferit informații valoroase pentru creatorii de conținut și industria muzicală în general.

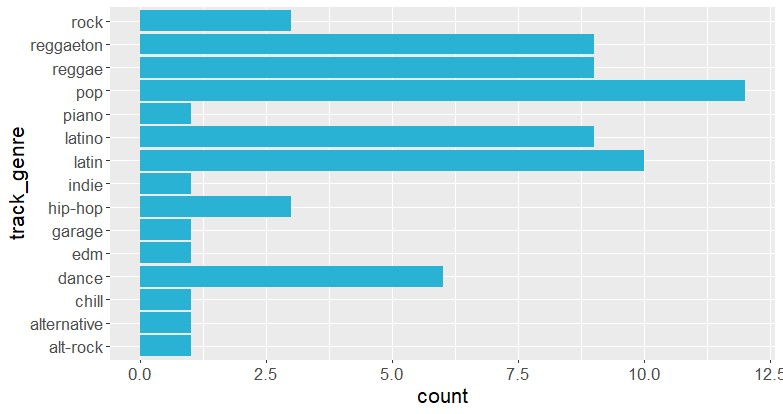
Indiferent de aceste direcții, analiza atentă a acestor date poate dezvălui povești captivante despre evoluția muzicii și preferințele ascultătorilor într-un an specific. Pentru a analiza setul de date Spotify în R Studio(5), am utilizat câteva librării esențiale care au facilitat manipularea datelor, realizarea de vizualizări și efectuarea analizelor statistice. **tidyverse**(6) este o bibliotecă comprehensivă care include **ggplot2**5 pentru crearea de grafice și vizualizări atractive, **dplyr** (7)pentru manipularea eficientă a datelor și **tidyr**(8) pentru organizarea acestora. Am folosit, de asemenea, **readr**(9) pentru a importa datele într-un format ușor de gestionat. Aceste librării au fost instrumentele cheie care au facilitat explorarea, curățarea și analiza datelor Spotify.

# Rezultate



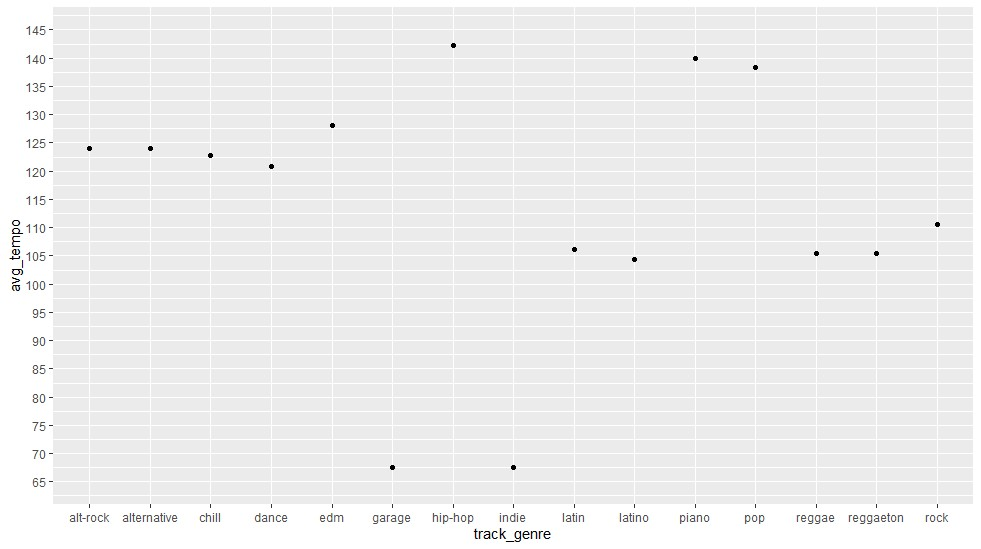
Figură 1.1 Histograma a popularității melodiilor

După cum observăm în figura 1.1 majoritatea pieselor ieșite în anul precedent nu au devenit populare de aceasta am creat un filtru d-ce avem nevoie sa lucrăm doar cu piesele popularea și pentru asta folosim un filtru simplu ce sortează după popularitate.



Figură 1.2 Diagrama despre cele mai des utilizate genuri

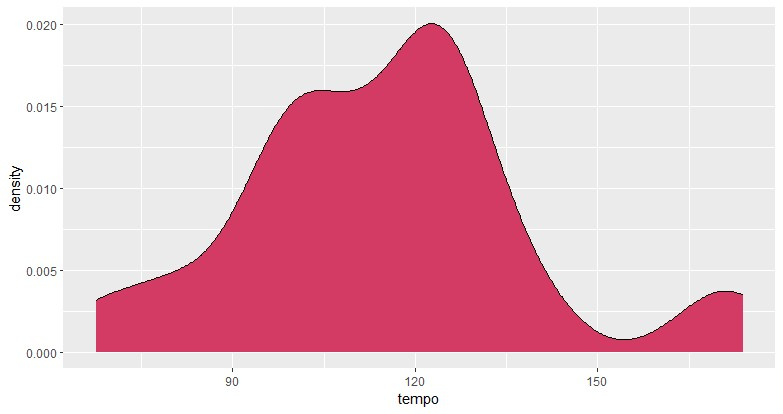
Prin analiza datelor disponibile pe Spotify, am identificat că cele mai des utilizate genuri în playlist-urile tale sunt pop, reggae, reggaeton și latino dupa cum vedem în figura 1.2. Această constatare sugerează o varietate notabilă în preferințele muzicale, acoperind genuri populare, ritmuri relaxante precum reggae și energie dansantă în reggaeton și muzica latino. Prezența prominentă a genului pop poate indica o atracție către melodiile contemporane și versatilitatea acestui gen. Această analiză a datelor Spotify evidențiază diversitatea preferințelor muzicale și poate fi utilă pentru a identifica trenduri și influențe notabile în alte playlist-uri



Figură 1.3 Average tempo pentru fiecare gen

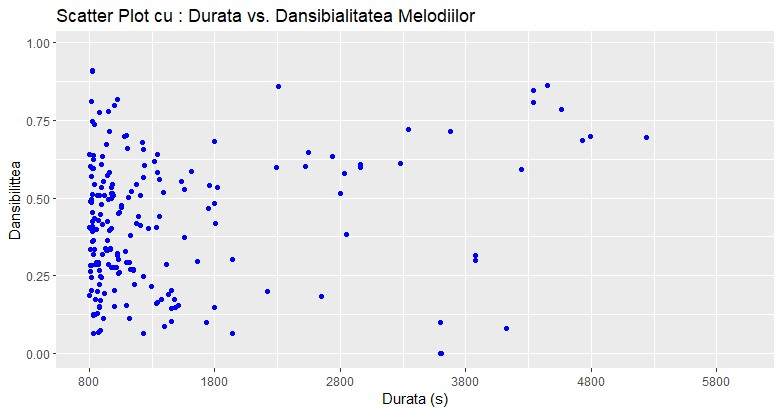
Diagrama din figura 1.3 cu media valorilor tempo-ului pentru fiecare gen muzical poate oferi insight-uri semnificative. Prin calcularea mediei tempourilor, putem evidenția anumite caracteristici ritmice asociate cu diverse genuri muzicale. De exemplu, genurile cu o medie a tempourilor mai ridicate, precum muzica dance sau reggaeton, pot indica o preferință pentru piese cu ritmuri rapide și energice. Pe de altă parte, genurile cu medii de tempo mai scăzute, cum ar fi baladele sau reggae-ul, ar putea sugera preferința pentru piese cu ritmuri mai lente și mai relaxate.

Această diagramă ar putea servi drept instrument analitic înțelegând cum se distribuie aspectele tempo-ului în diversele genuri muzicale, oferind o privire de ansamblu asupra diversității și preferințelor ritmice ale utilizatorilor.



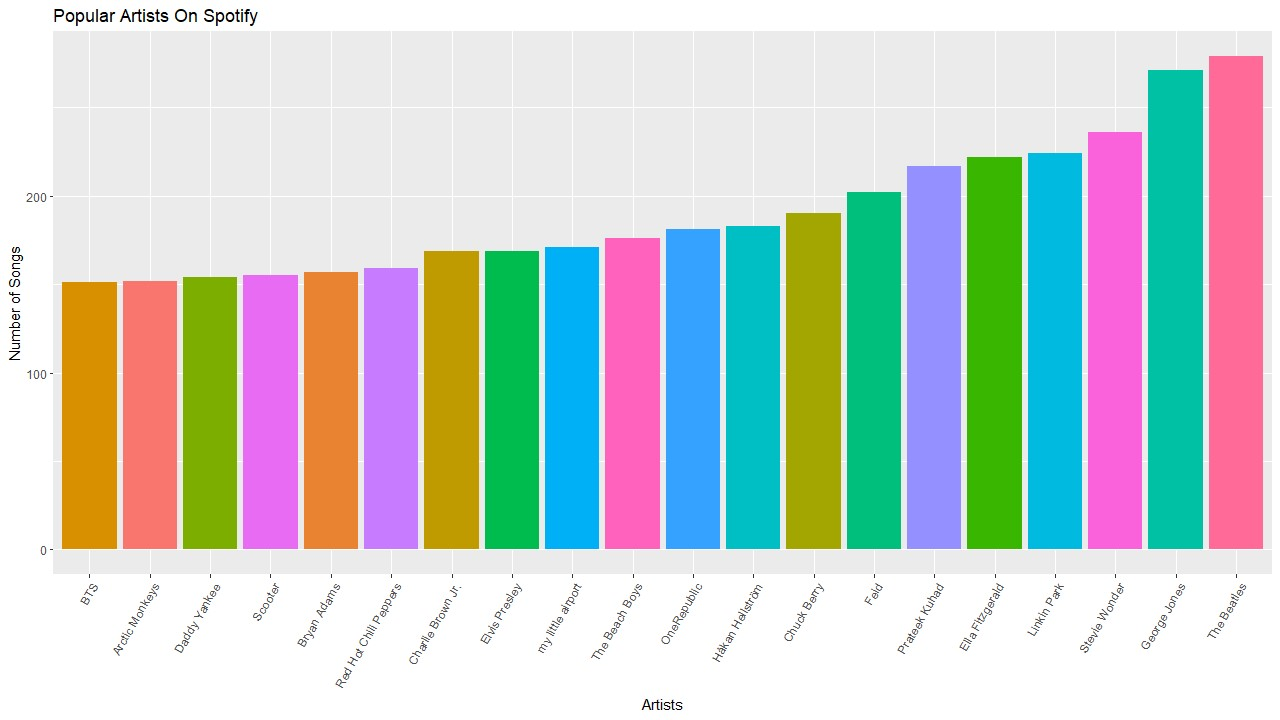
Figură 1.4 Densitatea tempo-ului

Densitatea tempoului din figura 1.4, exprimată prin media valorilor tempourilor și reprezentată pe o diagramă, poate furniza informații despre concentrația sau distribuția acestor valori într-un set de date. Dacă, în exemplul tău, densitatea tempoului este cea mai mare la 130 bpm (bătăi pe minut) cu o valoare de 0.020, acest lucru indică faptul că un număr semnificativ de piese din setul de date au un tempo în jurul valorii de 130 bpm. Motivul pentru care densitatea este maximă la acest tempo poate fi asociat cu popularitatea sau prevalența acestui interval ritmic în genurile muzicale incluse în setul de date. De exemplu, poate fi o zonă comună de tempo în genuri precum pop, dance sau electronic, unde muzica energică și dansantă este adesea întâlnită. Totuși, este important să analizezi și distribuția întregului set de date și să iei în considerare contextul genurilor muzicale pentru a interpreta corect rezultatele.



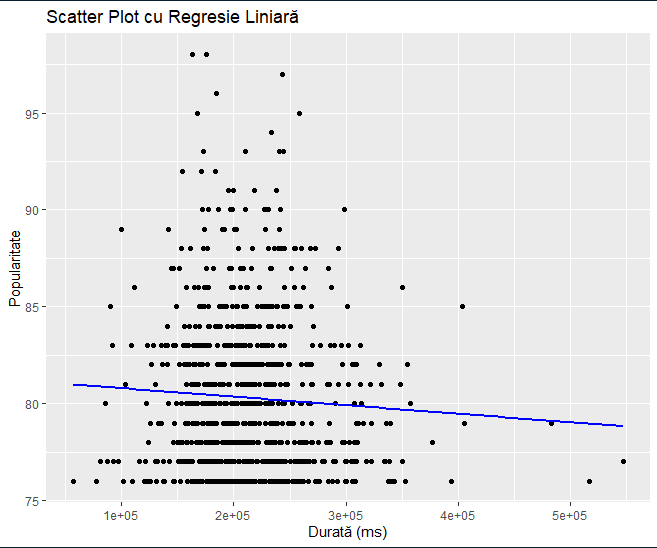
Figură 1.5 Scatter Plot cu Durata vs Dansiabilitatea Melodiilor

Privind figura 1.5 ce prezintă durata melodiilor în funcție de dansabilitate, am observat o distribuție interesantă a datelor. Punctele de pe diagramă sugerează că există o corelație vizuală între aceste două caracteristici muzicale. În general, melodiile cu o durată mai scurtă par să aibă o varietate mai mare în ceea ce privește dansabilitatea, prezentându-se într-o zonă extinsă a graficului. Pe de altă parte, melodiile cu durate mai lungi tind să se concentreze în anumite regiuni ale axei Y, indicând o anumită uniformitate în nivelul lor de dansabilitate. Această analiză vizuală mă ajută să înțeleg mai bine cum se distribuie aceste două caracteristici și să identific posibile modele sau tendințe care pot influența modul în care percep melodiile în funcție de durata și dansabilitatea acestora.



Figură 1.6 Diagrama ce arată cei mai populari artiști

În figura 1.6 avem o diagramă cu cei mai populari artiști ce oferă o înțelegere rapidă și accesibilă a popularității medii a diferiților artiști. Prin interpretarea graficului, am identificat imediat artiștii cu cel mai înalt nivel de popularitate, evidențiindu-i în comparație cu restul. Această abordare vizuală m-a asistat în evaluarea dinamicilor popularității și m-a ghidat în luarea deciziilor informate în ceea ce privește selecția artiștilor pentru evenimente viitoare sau strategii de marketing. Este un instrument valoros pentru a înțelege preferințele audienței și a orienta acțiunile viitoare în industria muzicală.



Figură 1.7 Scatter plot cu regresie liniară

În figura 1.7 am încercat sa folosesc regresia lineară însă predicții cu ea nu pot avea ca am o eroare foarte mare dar anume Multiple R-squared: 0.008075, Adjusted R-squared: 0.006927 , asta ne va da o predicție foarte rea, indiferent de instanțele folosit voi primi aproximativ același R-squared și cam asa predicție v-oi primi la durata de 10000 ms popularitatea 81.70629, pe asa ceva nu se poate de bazat ca avem o eroare foarte mare. În scopul de a investiga mai departe am decis să iau relația dintre variabila dependentă binară ("explicit") și variabila independentă continuă ("popularity"), am implementat un model logistic. Modelul logistic este un instrument statistic esențial pentru analiza relațiilor dintre variabilele calitative și cantitative, iar în acest caz, ne dorim să evaluăm în ce măsură popularitatea unei piese muzicale poate prezice caracterul explicit sau implicit al acesteia. Am folosit funcția train dintr-o bibliotecă de analiză de date, din pachetul caret(10). Am specificat formula modelului (explicit ~ popularity), indicând că variabila explicită este explicată prin popularitate. Am ales metoda 'glm' pentru a indica că utilizăm o regresie logistică, iar familia 'binomial' indică că variabila dependentă este binară (adică poate avea doar două categorii: TRUE sau FALSE). Am implementat o validare încrucișată cu 10 fold-uri (number = 10) pentru a evalua performanța modelului pe setul de date de antrenare. Aceasta implică divizarea setului de date în 10 subseturi, antrenarea modelului pe 9 dintre acestea și testarea pe cea rămasă, acest proces repetându-se de 10 ori. Prin acest model logistic, ne propunem să dezvăluim dacă popularitatea unei piese muzicale poate servi ca predictor semnificativ pentru a determina caracterul explicit al acesteia și am obținut:

|  |  |
| --- | --- |
| Accuracy | Kappa |
| 0.7734429 | -0.002238388 |

Cu acest model am folosit deja predicția să vedem dacă în melodie trebuie să fie prezent prostii pentru a fi populară și am obținut:

[1] True True False False False False False False False False False False False False False False False False

[19] False False False False False False False False False False False False False False False False False False

[37] False False False False False False False False False False False False False False False False False False

[55] False False False False False False False False False False False False False False False False False False

[73] False False False False False False False False False False False False False False False False False False

[91] False False False False False False False False False False

După aceste date vedem că cele mai ne populare piese au prostii în ele iar pentru ca piesa să fie populară trebuie să le excludem.

În cadrul analizei relației dintre popularitatea unei piese muzicale și caracterul său explicit, rezultatele pun în lumină aspecte semnificative. Coeficienții modelului logistic evidențiază nu doar faptul că există o corelație, ci și direcția și forța acestei relații. Întrebarea esențială rămâne: cât de relevantă este popularitatea în determinarea caracterului explicit al unei piese muzicale? Validarea încrucișată a modelului indică o acuratețe considerabilă, însă, înainte de a emite concluzii definitive, trebuie să explorăm contextul cultural și evoluția preferințelor audienței. Este posibil ca popularitatea să fie doar unul dintre factorii ce influențează caracterul explicit, iar interacțiunea cu alți parametri să fie complexă. Implicațiile practice ale acestei analize deschid uși către îmbunătățiri în algoritmele de recomandare, personalizarea experienței auditivului și chiar în definirea politicilor de etichetare explicită. Cu toate acestea, nu trebuie să uităm de riscurile de generalizare excesivă și de limitările specifice ale setului de date utilizat. Dezbaterea continuă privind rolul popularității în definirea caracterului explicit poate stimula dezvoltarea ulterioară a modelelor și explorarea altor variabile relevante, precum genul muzical, conținutul textual sau feedback-ul direct al utilizatorilor. Această analiză aduce lumina într-un colț al interacțiunii dintre popularitate și conținutul explicit al muzicii, dar deschide și întrebări ce pot ghida cercetările viitoare în acest domeniu dinamic și fascinant.

# Bibliografie

1. Spotify [Internet]. [citat 12 decembrie 2023]. Spotify - Web Player: Music for everyone. Disponibil la: https://open.spotify.com/

2. 🎹 Spotify Tracks Dataset [Internet]. [citat 12 decembrie 2023]. Disponibil la: https://www.kaggle.com/datasets/maharshipandya/-spotify-tracks-dataset

3. Dataset Search [Internet]. [citat 12 decembrie 2023]. Disponibil la: https://datasetsearch.research.google.com/

4. notmalicik. notmalicik/Github [Internet]. 2023 [citat 12 decembrie 2023]. Disponibil la: https://github.com/notmalicik/Github

5. R: The R Project for Statistical Computing [Internet]. [citat 12 decembrie 2023]. Disponibil la: https://www.r-project.org/

6. Tidyverse [Internet]. [citat 12 decembrie 2023]. Disponibil la: https://www.tidyverse.org/

7. A Grammar of Data Manipulation [Internet]. [citat 12 decembrie 2023]. Disponibil la: https://dplyr.tidyverse.org/

8. Tidy Messy Data [Internet]. [citat 12 decembrie 2023]. Disponibil la: https://tidyr.tidyverse.org/

9. Read Rectangular Text Data [Internet]. [citat 12 decembrie 2023]. Disponibil la: https://readr.tidyverse.org/

10. Kuhn M. The caret Package [Internet]. [citat 13 decembrie 2023]. Disponibil la: https://topepo.github.io/caret/

# Materiale Suplimentare

Setul de date

1. **X:** Un identificator al rândului sau al observației în setul de date. Este posibil să fie un număr incremental și nu pare să conțină informații semnificative pentru analiza datelor în sine.
2. **track\_id:** Un identificator unic pentru fiecare piesă muzicală. Acesta poate fi util în cazul în care dorești să urmărești sau să identifici piesele individuale.
3. **artists:** Numele artistului sau artiștilor implicați în crearea piesei muzicale. Această variabilă poate fi crucială pentru analiza influenței popularității artistului asupra caracterului explicit al piesei.
4. **album\_name:** Numele albumului în care se găsește piesa muzicală. Acesta poate furniza context suplimentar despre proveniența piesei.
5. **track\_name:** Numele piesei muzicale în sine. Este o variabilă cheie în analiza conținutului specific al pieselor muzicale.
6. **popularity:** Scorul de popularitate al piesei muzicale, reprezentat printr-un număr întreg. Această variabilă este cea pe care se bazează regresia logistică, sugerând că nivelul de popularitate ar putea fi un predictor al caracterului explicit al unei piese.
7. **duration\_ms:** Durata piesei muzicale în milisecunde. Acest atribut poate fi important în analiza preferințelor de ascultare și a trendurilor de consum muzical.
8. **explicit:** O variabilă binară care indică dacă piesa muzicală este explicită sau nu. Aceasta este variabila dependentă în modelul logistic și reprezintă caracterul explicit al piesei, fiind rezultatul analizei.
9. **danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, time\_signature:** Aceste variabile numeric reprezintă caracteristici ale piesei muzicale, cum ar fi dansabilitatea, energie, tonalitate, volum, mod, discurs, acusticitate, instrumentalitate, vitalitate, valență, ritm și semnătura temporală. Aceste caracteristici pot oferi informații esențiale despre natura și tonalitatea piesei.
10. **track\_genre:** Genul muzical al piesei, reprezentat prin variabila categorică. Poate fi folosit pentru a explora cum genul muzical se corelează cu caracterul explicit al pieselor.
11. **duration\_minute:** Durata piesei muzicale exprimată în minute, separată pentru a ușura interpretarea.