# Alegerea genului și platformei pentru un Joc Video reușit

**Mazureac Vlad, IA-211, III**

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

**Abstra­ct**

Într-o epocă digitală în expansiune continuă, jocurile video au devenit un fenomen cultural global, depășind limitele simplului divertisment. Acest articol se focalizează asupra analizei ample a datelor din industria jocurilor video, explorând preferințele jucătorilor în privința platformelor și genurilor. De la primele console rudimentare la complexitatea grafică contemporană, articolul dezvăluie nu doar cifre și statistici, ci și povești profunde ce conturează această industrie în schimbare constantă.

Acest articol efectuează o analiză detaliată a datelor extinse din industria jocurilor video, concentrându-se pe preferințele jucătorilor legate de platforme și genuri. De la evoluția simplă la complexitatea grafică contemporană, analiza setului de date relevă nu doar cifre și statistici, ci și poveștile subiacente care conturează dinamica acestei industrii. Utilizând preprocesarea datelor și regresia liniară, investigăm relațiile dintre variabilele cheie, evidențiind impactul alegerii platformei și genului asupra numărului de jucători. Rezultatele indică că alegerea unei platforme populare și a unui gen solicitat poate atrage un număr semnificativ de jucători și vânzări globale, contribuind la înțelegerea evoluției pieței de jocuri video și la promovarea unor strategii de dezvoltare eficiente în această industrie în continuă schimbare.

# Introducere

Într-o epocă digitală în continuă expansiune, universul jocurilor video a evoluat dincolo de simpla formă de divertisment, devenind un fenomen cultural global. Milioane de oameni din întreaga lume sunt atrași de lumea captivantă și interactivă a jocurilor, iar această fascinație generează o cantitate imensă de date. În acest context, propunem o incursiune profundă în analiza unui set de date amplu, axat pe jocurile video, pentru a dezvălui nu doar cifre și statistici, ci și poveștile subiacente ce conturează această industrie în continuă schimbare.

Industria jocurilor video nu mai este doar un domeniu de divertisment; a devenit o forță motrice în economie și cultură. De la debutul simplu al jocurilor pe console rudimentare la complexitatea și detaliile grafice ale epocii actuale, parcurgerea evoluției acestui mediu oferă o privire fascinantă asupra modului în care tehnologia și creativitatea umană au convergat.

În cadrul acestei analize a setului de date, ne propunem să explorăm nu doar cifrele, ci și povestea din spatele acestora. Vom examina preferințele jucătorilor în ceea ce privește platformele și genurile de jocuri, încercând să înțelegem ce motivează alegerea unui joc în detrimentul altuia. De asemenea, vom investiga cum influențează recenziile critice și cum evoluează vânzările în contextul schimbărilor tehnologice și culturale.

Această analiză nu este doar un exercițiu matematic, ci o călătorie în profunzimea unei industrii care transcendă ecranul, atingând inimile și mințile a milioane de oameni. Vom încerca să conturăm nu doar prezentul, ci și să proiectăm direcțiile posibile pentru viitorul jocurilor video, luând în considerare tendințele emergente și schimbările sociale ce vor defini această industrie dinamică. Astfel, vom îmbrățișa o abordare holistică a analizei datelor pentru a desluși povestea captivantă și în continuă schimbare a lumii jocurilor video.

Scopul principal ai acestui articol este investigarea și prezicerea datelor și anume numărului de utilizatori în dependență de genul și platforma pe care a fost lansat jocul video. Acest lucru presupune colectarea, prelucrarea și vizualizarea datelor din setul de date pregătit și crearea unor modele statistice.

În urma realizării articolului verificăm dacă genul de shooter sau sport în combinație cu platforma PC sau PS4 va influența numărul de jucători care va avea joaca dacă va fi lansată pe unele de pe platformele enumerate și cu un gen anumit.

# Materiale & Metode

**Setul de Date**

Pentru realizarea acestei analize, am folosit un set de date cuprinzător care furnizează informații actualizate despre performanța vânzărilor și popularitatea diverselor jocuri video la nivel mondial. Acest set de date a fost luat de pe platformă open-source Kaggle.com[2], jocurile analizate au fost lansate începând cu 1988 până în 2016.

Variabile Incluse în Setul de Date – în setul de date sunt prezente 16 variabile care furnizează informații despre jocuri, 5 variabile reprezintă date despre vânzări, câte 2 pentru datele critice și 2 pentru datele despre utilizatori, restul datelor se referă la date unice despre joc cum ar fi denumirea, anul lansării etc. Mai multe detalii vedeți Anexa1.

Acest set de date esențial furnizează informații valoroase despre piața globală a jocurilor video. Este o resursă indispensabilă pentru gameri, profesioniștii din industrie și cercetătorii de piață, oferind o perspectivă generală asupra performanței jocurilor, a preferințelor utilizatorilor și a schimbărilor din industrie pe parcursul anilor. Analiza acestui set de date poate contribui la înțelegerea evoluției pieței de jocuri video și poate oferi insights semnificative pentru decizii viitoare în industrie.

**Preprocesarea Datelor**

Pentru a pregăti datele pentru analiză, am folosit limbajul de programare R[1] și bibliotecile openintro[3], tidyverse[4] și altele. Am încărcat setul de date și am efectuat o analiză sumară și tipurilor de date existente. Și am observat că aveam doar două tipuri de date, dbl și chr. Setul de date conține 16719 rânduri și 16 coloane. La fel setul de date conține multe valori NA ceea ce complică analiza acestor date, pe lângă date NA sunt si date de tip dbl care au valoarea 0.00 care la fel putem socoti ca NA. Din această cauză am eliminat aceste rânduri fără date. Am factorizat unele coloane care erau de tipul chr. Am șters așa coloane care nu le foloseam ca Name, Publisher, Developer, Year\_of\_Release, Rating. Am logoritmat coloana User\_Count.

**Analiza Exploratorie a Datelor**

Am realizat o serie de vizualizări și am explorat relațiile dintre diferitele caracteristici ale jocurilor video și performanța acestora pe piață. De exemplu, am investigat distribuția vânzărilor globale în funcție de genul și platforma de jocuri. Am explorat aceste relații folosind ggploturi[4] și am analizat graficul cum aceste date sunt distribuite și relațiile între aceste date. Am rezumat datele folosind funcția summary[2]. 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**

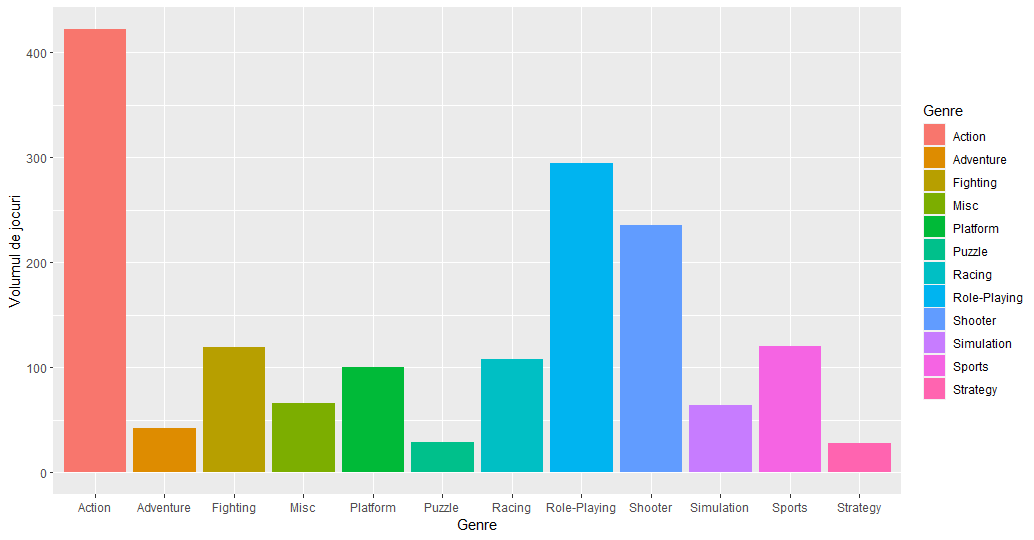
În cadrul acestei etape cruciale a analizei, am explorat relațiile dintre variabilele cheie din setul de date, focalizându-ne asupra platformei, genului și numărului de jucători în procesul de modelare. Am împărțit setul de date în două segmente distincte: unul de antrenament și unul de testare. Am selectat cu grijă variabilele relevante, concentrându-ne pe platformă, gen și numărul de jucători, în concordanță cu obiectivele noastre de analiză. Am îmbunătățit ajustarea modelelor, aplicând o transformare logaritmică asupra variabilei "User\_Count," abordare care a contribuit la gestionarea asimetriilor în distribuția datelor. Modelul de regresie liniară a fost implementat, iar parametrii au fost ajustați pentru a realiza cea mai bună potrivire cu datele disponibile.

**Codul disponibil**

Tot codul pentru crearea modelului pentru regresie liniară și pentru prezicere este disponibil <https://github.com/vladmazureac/Analiza-Datelor/blob/main/proiect.R> .

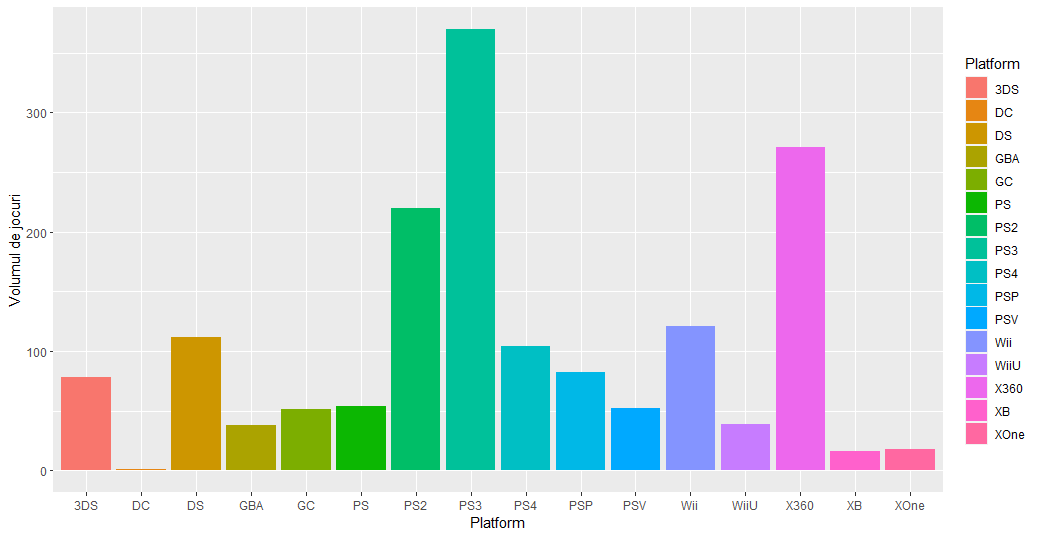
# Rezultate

După preprocesarea datelor am început cu analiza datelor rămase pentru procesarea datelor, cu ajutorul ggploturilor am analizat și observat cele mai des folosite platforme și genuri pentru jocuri(fig 1, fig 2). Aceste date vor ajută să înțeleg dacă numărul de jocuri lansate pe o anumită platformă sau alegerea unui gen popular de jocuri, influențează numărul de utilizatori.



**Fig 1** Numărul de Genuri utilizate

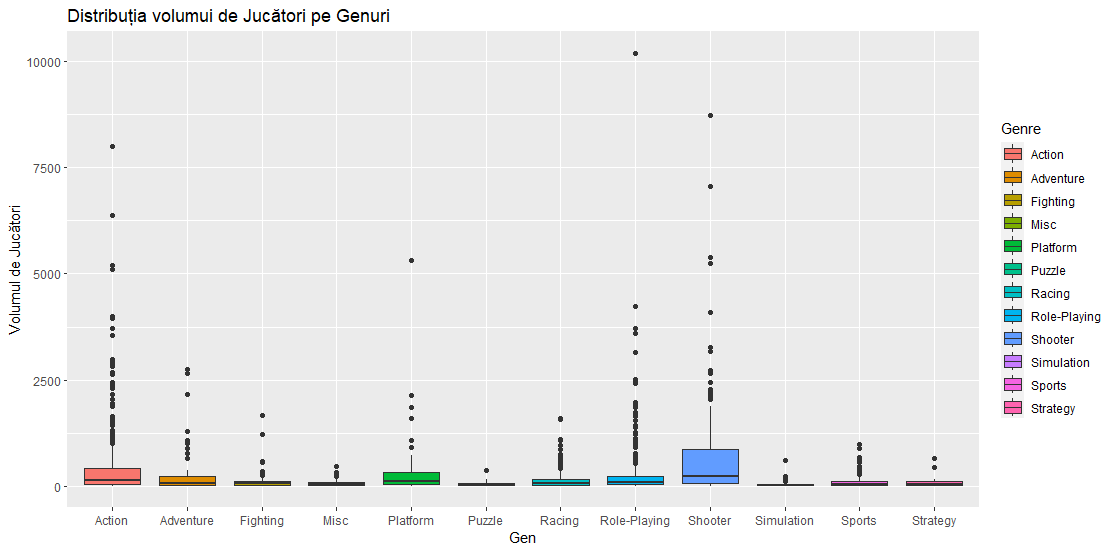
Am obținut cel mai popular gen de joc care este genul Action, 420 de jocuri folosesc acest gen. După genul Action cu un număr de 291 este genul Role-Playing și al treilea cel mai popular gen este Shooter, restul genurilor nu sunt așa de populare având majoritatea un număr mai mic sau egal cu 100.



**Fig 2** Numărul de Platforme utilizate

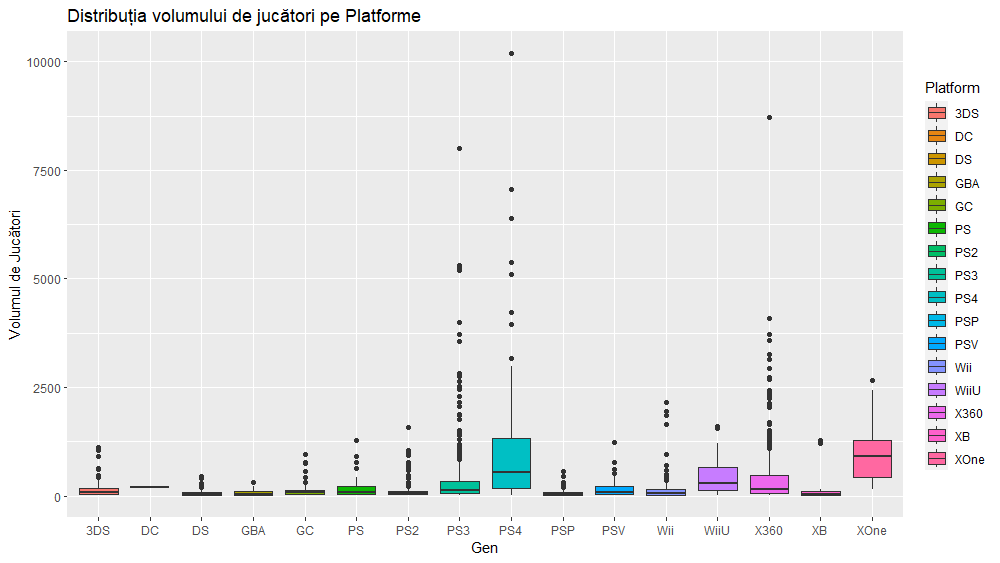
Cea mai populară platformă utilizată este PS3 cu un număr de 366 de utilizări, X360 este a doua cea mai utilizată platformă cu un număr de 270, al treia cea mai populară platformă pe care sunt lansate 216 jocuri este PS2. Așa platforme ca PC, PS4, Xone nu sunt foarte populare, acest lucru este datorită faptului că datele pentru setul de date au fost luate din 1988 când nu erau populare sau nu existau așa platforme ca PC, XOne, PS4.

În continuare am analizat distribuția numărului de jucători pe genuri(fig 3) cu ajutorul geom\_boxplot-urilor și am primit rezultate referitoare la numărul aproximativ de jucători pe diferite genuri.



**Fig 3** Distribuția volumului de Jucători pe Genuri

Rezultatele primite arată că genul Shooter are o cantitate cea mai mare de jucători, cu un IQR cel mai mare. Action este al doilea gen cu o cantintate mare de jucători, după Action, Platform are cel mai mare număr. Însă genurile enumerate precedent au un IQR mai mare și puncte extreme înalte, la fel puncte extreme mari are genul Role-Playing. La fel în figura 4 putem vedea distribuția volumui de jucători pe Platforme.



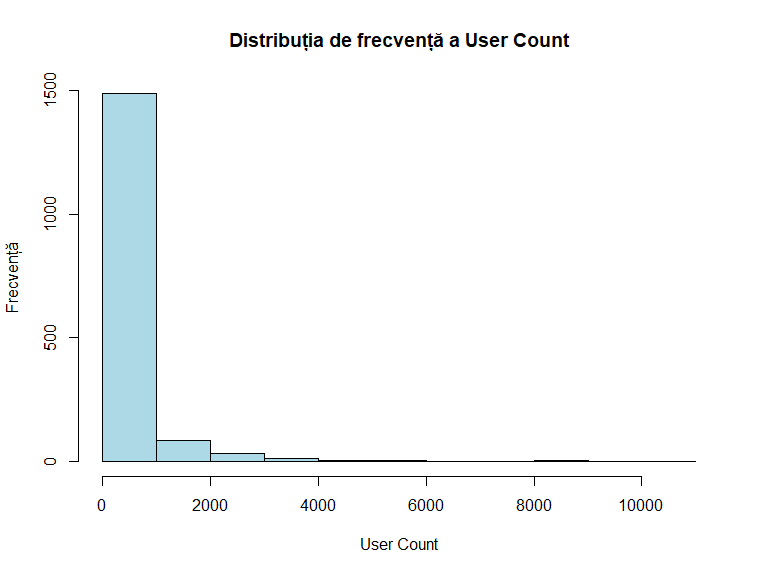
**Fig 4** Distribuția volumui de Jucători pe Platforme

Platforma cu cel mai mare număr de jucători este PS4 cu cel mai mare IQR, pe al doilea loc este platforma XOne și al treilea loc WiiU. Aceste platforme au cel mai mare număr de jucători.

După analizele efectuate precedent am înțeles că numărul de jucători nu depinde de numărul de jocuri lansate pe o platformă anumită sau un gen anumit. De exemplu PS4 nu este așa popular, adică puține jocuri au fost lansate pe această platformă dar numărul de jucători este cel mai mare.

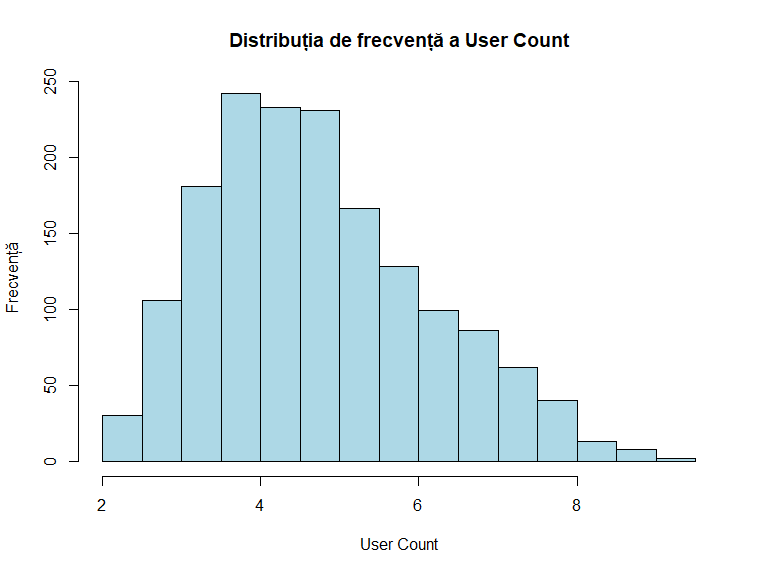
Pentru a putea prezice cantitatea de jucători în dependență de Platformă și Gen am folosit modelul de regresie liniară. Am împărțit setul de date în seturi de antrenare și seturi de testare cu proporția de 70% la 30%, adică am folosit 70% din date pentru antrenare și 30% pentru testare.

Pentru a putea crea modelul de regresie liniară cât de cât reușită am logoritmat datele din coloana care dorim să o prezicem, adică coloana User\_Count. Am înțeles că nu o să putem avea un model reușit deoarece uitândune la histogramă din figura 5 observăm distribuția de frecvență.



**Fig 5** Distribuția de frecvență a User Count

După cum vedem frecvența este una rea[5] pentru a crea un model de regresie liniară, datele pentru frecvență sunt foarte diferite unele față de altele. Din această cauză am folosit funcția de logoritmare și am primit următoare distribușie de frecvență a User Count logoritmat(fig 6).



**Fig 6** Distribuția de frecvență a User Count logoritmat

După procesul de logoritmare noi am primit o frecvență bună pentru a putea face un model de regresie liniară. Histograma are o formă normală și simetrică.

Am creat modelul de regresie liniară cu formula User\_Count ~ Platform + Genre, ceea ce înseamnă că prezic User\_Count în funcție de variabilele Platform și Genre. Am ales metoda de modelare ‘lm’, am utilizat validarea încrucișată și am ales 10 folduri pentru validarea încrucișată. Adică setul meu de date este împărțit în 10 părți egale, fiecare dată folosind un folder pentru testare și 9 pentru antrenament, acest proces se repetă de 10 ori.

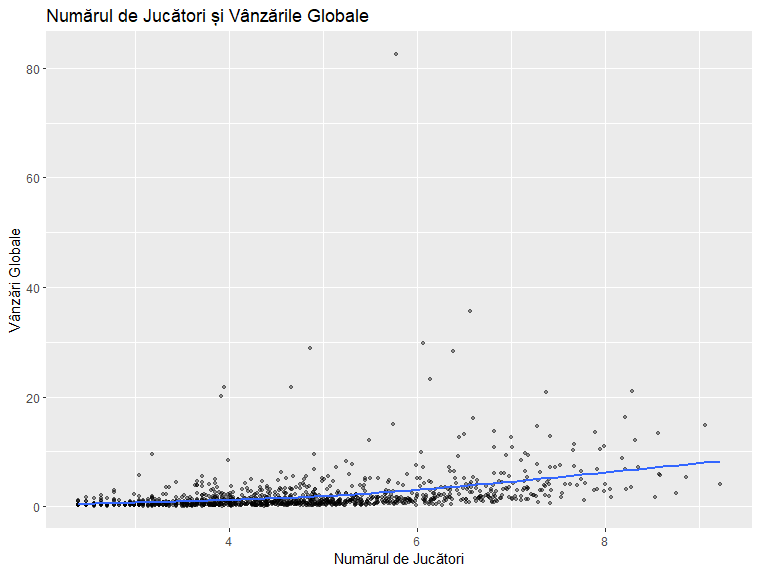
Utilizând funcția summary la model am primit așa valori ca Residuals cu valorile Min -3.2801, 1Q -0.8104, Median -0.050, 3Q 07230, Max 3.6055.

Iar cel mai important R-squared: 0.2843; RMSE: 1.187405; MAE: 0.9485866. Acest lucru înseamnă că modelul creat nu este foarte bun, adică prezicerea nu va fi una exactă sau bună și va avea o aproximație joasă.

Pentru prezicere introducem o denumire de Platformă și Genul. De exemplu introducem Platforma PS4 și Genul Platform și primim o cantitate de 610.2563 ceea ce înseamnă că majoritatea jocurilor lansate pe platforma PS4 și cu genul de Shooter au o cantitate mare de jucători, și media este de 630 6236 100 jucători. Însă aceste date au fost luate în considerație pe parcusul multor ani. Dacă alegem platforma XOne și genul Sports atunci primi o cantitate de 406 633 000 de jucători. Un alt exemplu de prezicere este când alegem platforma XB și genul Simulation care nu este așa de reușit și va avea o cantitate apoximativă de jucători în număr de 30 209 300.

După ce am prezis numărul de jucători în dependență de Platformă și Gen am primit cele mai populare Platforme sunt PS4, XOne iar cele mai populare genuri sunt Role-Playing și Shooter, Adventure, Platform. Dacă joaca va fi lansată pe aceste platforme și cu așa genuri au o șansă mare de a avea un număr mare de jucători. Iar unele din cele mai nepopulare Platforme sunt XB, 3DS, GS, Wii în combinație cu așa genuri ca Racing, Simulation, Strategy. În prezicere avem așa cazuri că așa platforma DC și PC care din cauza preprocesării datelor au rămas fara date sau cu o joacă pe această platformă din această cauză prezicerea nu este reușită când alegem aceste două platforme.

La fel am verificat dacă un joc are o cantitate mare de utilizatori va avea și vânzări globale la fel mari. Am verificat acest lucru cu ajutorul unui geom\_point care este prezentat în figura 7.



**Fig 7** Vânzările globale în dependență de numărul de jucători

Am observat în figura de mai sus că în mare parte jocurile cu un număr de jucători mare au și vânzările globale mari, deci putem spune că dacă vom alege o platformă reușită și un gen reușit vom avea șansă mari de a avea un număr mare de jucători și vânzări destul de mari. Iar cele mai reușite Platforme pentru a lansa jocul sunt PS4 și XOne care sunt unele din cele mai populare platforme cu un număr de jucători mare, și cele mai populare genuri sunt Role-Playing,Platform și Shooter, combinația acestora ne va oferi cel mai posibil o joacă cu un număr mare de jucători care la rândul său influențează la vânzările locale. Și anume numărul de jucători care vom obține uitlizând aceste combinații va fi: PS4 + Shooter = 630 623 100 jucători, PS4 + Platform = 610 256 300 jucători, XOne + Shooter = 1 219 542 200 jucători, XOne + Platform = 1 180 154 400 jucători.

# 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: [Video Game Sales with Ratings](https://www.kaggle.com/rush4ratio/video-game-sales-with-ratings) - sursa setului de date utilizat în analiză, disponibil pe Kaggle. – accesat (13.09.2023)
3. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer. Disponibil la: [https://ggplot2.tidyverse.org](https://ggplot2.tidyverse.org/). – accesat(14.09.2023)
4. Wickham, H., și alții. (2019). *tidyverse: Easily Install and Load the 'Tidyverse'*. Disponibil la: [https://www.tidyverse.org](https://www.tidyverse.org/). – accesat(14.09.2023)
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. Variabilele setului de date

1. **Numele Jocului (Name):** Identifică denumirea fiecărui joc analizat.
2. **Platforma (Platform):** Specifică platforma pe care a fost lansat fiecare joc.
3. **Anul Lansării (Year\_of\_Release):** Furnizează informații despre anul de lansare al fiecărui joc.
4. **Genul (Genre):** Indică genul jocului, oferind o clasificare în categorii precum Sports, Racing, etc.
5. **Editorul (Publisher):** Numele companiei care a publicat fiecare joc.
6. **Vânzările în Regiuni Specifice:** Include detalii despre vânzările în America de Nord (NA\_Sales), Europa (EU\_Sales), Japonia (JP\_Sales) și alte regiuni (Other\_Sales).
7. **Vânzări Globale (Global\_Sales):** Reprezintă totalul vânzărilor în toate regiunile.
8. **Scoruri și Evaluări:** Cuprinde informații critice, precum scorul mediu al criticilor (Critic\_Score), numărul de critici care au evaluat (Critic\_Count), scorul mediu al utilizatorilor (User\_Score), și numărul de utilizatori care au evaluat (User\_Count).
9. **Dezvoltatorul (Developer):** Identifică compania responsabilă pentru dezvoltarea fiecărui joc.
10. **Clasificarea (Rating):** Specifică clasificarea de vârstă a jocului, precum "E" (Toate vârstele) sau "M" (Maturi).

Anexa 2. Rezidurile

