Pauna_Alexandru_Mihai_314CA

Documentatie - Problema de clasificare

-rezolvarea proiectului urmeaza exemplul prezentat in laboratorul 5

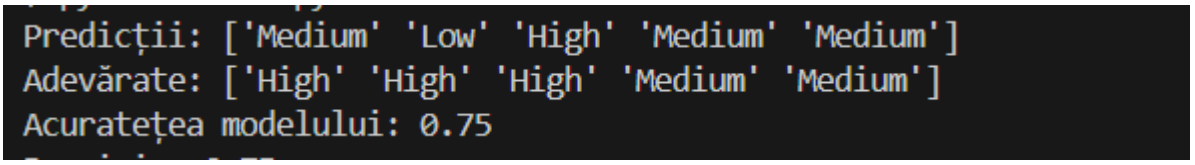
-setul de date a fost extras de pe “https://www.kaggle.com/” ,unde nu lipseau date ,dar am schimbat la intamplare anumite date in Nan ,rezolvand situatiile de genul inlocuid spatiile goale cu media elemnetelor dupa ce au fost encodade coloanele care nu erau eligibile pentru calcule.

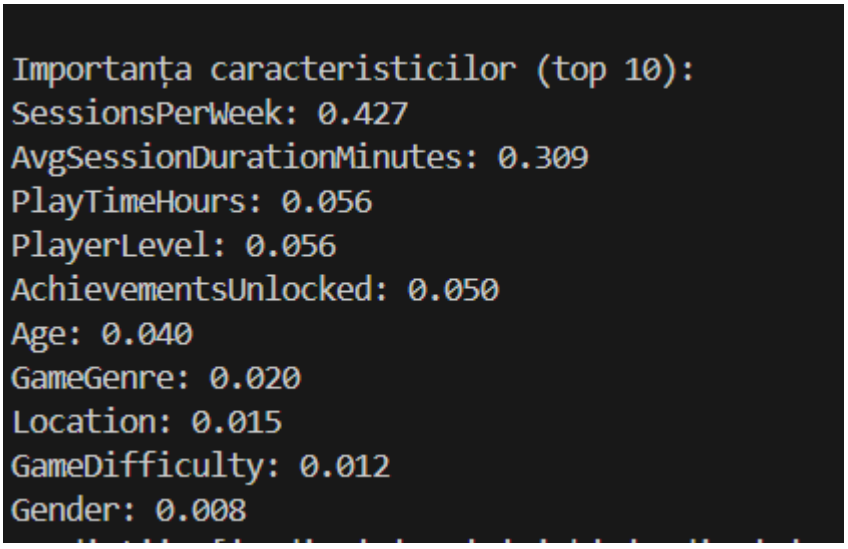
-setul de date online_gaming_behavior_dataset.csv a fost creat pentru a analiza comportamentul jucatorilor in mediul online. Datele contin informatii despre utilizatori, activitatea lor in jocuri si nivelul de implicare (EngagementLevel). Fiecare rand reprezinta un jucator, iar coloanele includ atat date demografice, cat si date despre activitatea in joc.

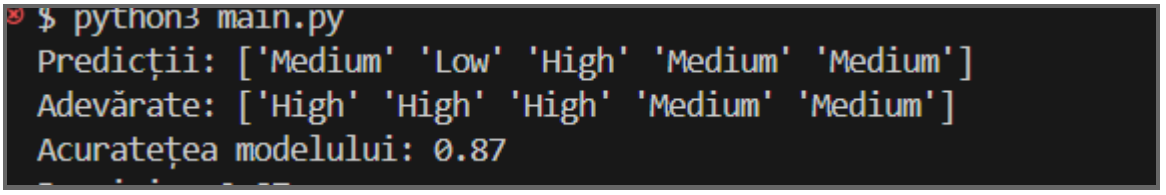
-surse si metode de generare:

- Datele pot proveni din platforme de gaming online, chestionare sau simulari generate sintetic pentru scopuri educationale.
- Fiecare jucator are un identificator unic (PlayerID).
- Caracteristicile includ: varsta, gen, locatie, genul jocului preferat, ore jucate, achizitii in joc, dificultatea jocului, sesiuni pe saptamana, durata medie a unei sesiuni, nivelul jucatorului si realizarile deblocate.

-acuratetea pentru toate coloanele este 0.75 ,dar daca consideram doar primele 3 coloane importante din fisier,acuratetea creste la 0,87(KNN)







-dar cu random_forest acuratetea ajunge la 0,90

```
Predicții Random Forest: ['Medium' 'Low' 'High' 'Medium' 'Medium']
Adevărate: ['High' 'High' 'High' 'Medium' 'Medium']
Acuratețea Random Forest: 0.90
```

Analiza exploratorie a datelor

-Analiza valorilor lipsa:

A-1_forest.txt

```
X_train - Age: 273 valori lipsa (0.85%)

X_train - PlayTimeHours: 311 valori lipsa (0.97%)

X_train - SessionsPerWeek: 278 valori lipsa (0.87%)

X_train - AvgSessionDurationMinutes: 300 valori lipsa (0.94%)

X_train - PlayerLevel: 299 valori lipsa (0.93%)

X_train - AchievementsUnlocked: 267 valori lipsa (0.83%)


X_test - Age: 72 valori lipsa (0.90%)

X_test - PlayTimeHours: 74 valori lipsa (0.92%)

X_test - SessionsPerWeek: 62 valori lipsa (0.77%)

X_test - AvgSessionDurationMinutes: 73 valori lipsa (0.91%)

X_test - PlayerLevel: 53 valori lipsa (0.66%)

X_test - AchievementsUnlocked: 95 valori lipsa (1.19%)
```

- Strategia de tratare e sa inlocuiesc cu media pentru coloanale numerice si cu cea mai frecventa valoare pentru coloanele categorice

-Statistici descriptive

Gender
count: 39679 unique: 2 top: Male freq: 23745

Location
count: 39691 unique: 4 top: USA freq: 15877

GameGenre
count: 39666 unique: 5 top: Sports freq: 7982

GameDifficulty
count: 39668 unique: 3 top: Easy freq: 19828

InGamePurchases
count: 39664 unique: 2 top: 0 freq: 31685

Age
count: 39661 mean: 31.997 std: 10.041 min: 15.0
25%: 23.0 50%: 32.0 75%: 41.0 max: 49.0

PlayTimeHours
count: 39665 mean: 12.023 std: 6.916 min: 0.0001
25%: 6.07 50%: 12.01 75%: 17.96 max: 23.9996

SessionsPerWeek
count: 39687 mean: 9.473 std: 5.763 min: 0.0
25%: 4.0 50%: 9.0 75%: 14.0 max: 19.0

AvgSessionDurationMinutes
count: 39659 mean: 94.822 std: 49.001 min: 10.0
25%: 52.0 50%: 95.0 75%: 137.0 max: 179.0

PlayerLevel
count: 39659 mean: 49.685 std: 28.581 min: 1.0
25%: 25.0 50%: 49.0 75%: 74.0 max: 99.0

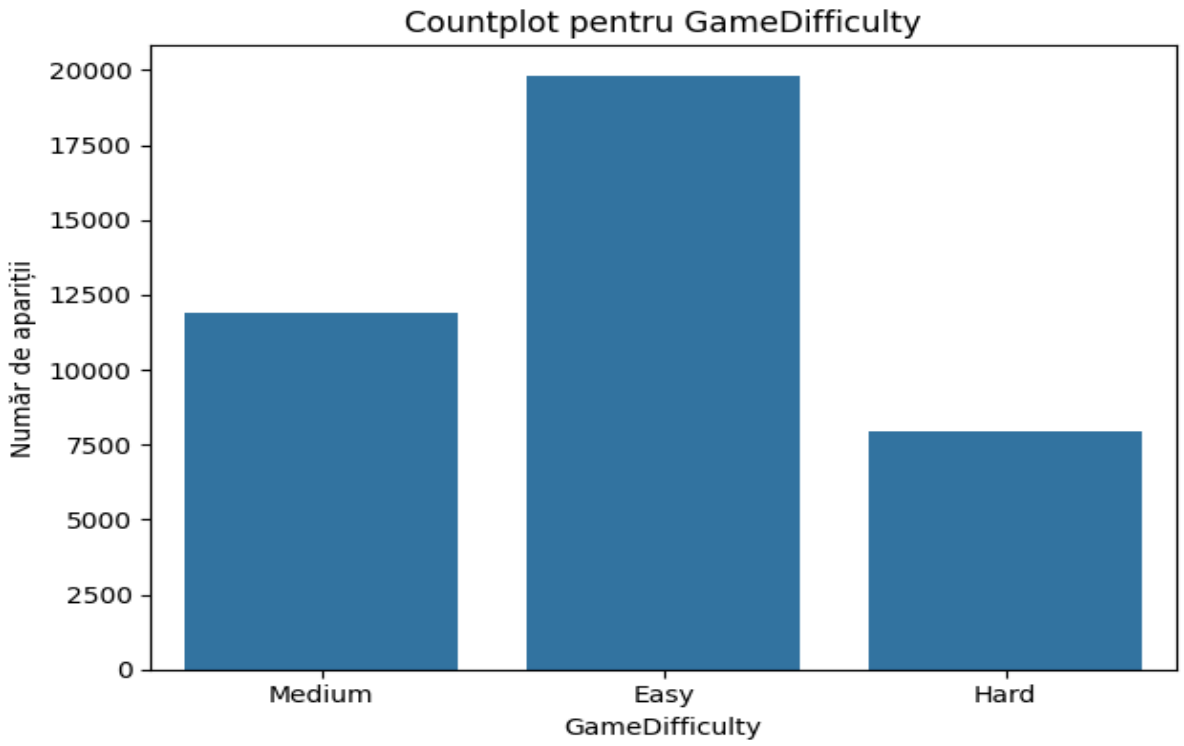
AchievementsUnlocked
count: 39693 mean: 24.535 std: 14.432 min: 0.0
25%: 12.0 50%: 25.0 75%: 37.0 max: 49.0

-astfel avem informatii despre majoritatea persoanelor care se joaca ca sunt din USA,barbati,etc

Termen	Semnificatie
count	Numarul total de valori disponibile in acel camp.
unique	Numarul de valori unice pentru variabile categorice.
top	Valoarea cea mai frecventa.
freq	Frecventa a valorii de top.
mean	Media aritmetica .
std	Deviatia standard (cat de mult variaza valorile fata de medie).
min	Valoarea minima in setul de date.
25%	Primul percentil — 25% din valori sunt mai mici decat aceasta.
50%	Mediana — valoarea din mijloc.
75%	Al treilea percentil — 75% din valori sunt mai mici decat aceasta.

max Valoarea maxima in setul de date.

-Analiza distributiei variabilelor:



Ce observam?

Graficul arata o distributie dezechilibrata a jucatorilor pe niveluri de dificultate:

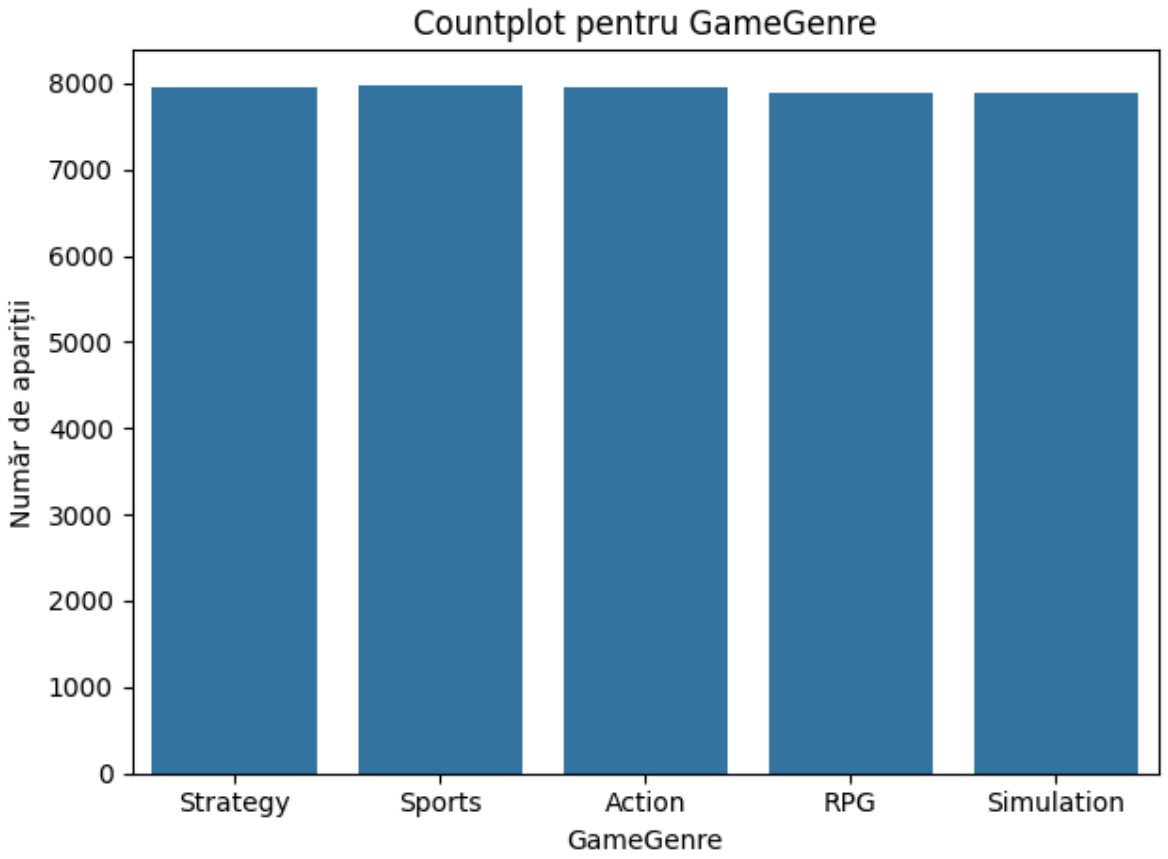
- Easy: ~20.000 jucatori (cel mai mare grup)
- Medium: ~12.500 jucatori
- Hard: ~7.500 jucatori (cel mai mic grup)

Ce suspiciuni/idei putem formula?

- Majoritatea jucatorilor prefera experiente accesibile si mai putin provocatoare
- Exista o bariera psihologica sau tehnica care impiedica tranzitia la dificultatea Hard
- Jocul ar putea sa nu ofere suficiente incentive pentru dificultatile superioare
- Distributia sugereaza ca publicul tinta este format predominant din jucatori casual

Ce preprocesari ar trebui sa aplicam?

- Verificarea si completarea valorilor lipsa (39.668 din 40.000 inregistrari)
- Codificare ordinala: Easy=1, Medium=2, Hard=3 pentru a reflecta progresivitatea



Ce observam?

Distributia pe genuri este relativ echilibrata cu o usoara predominanta:

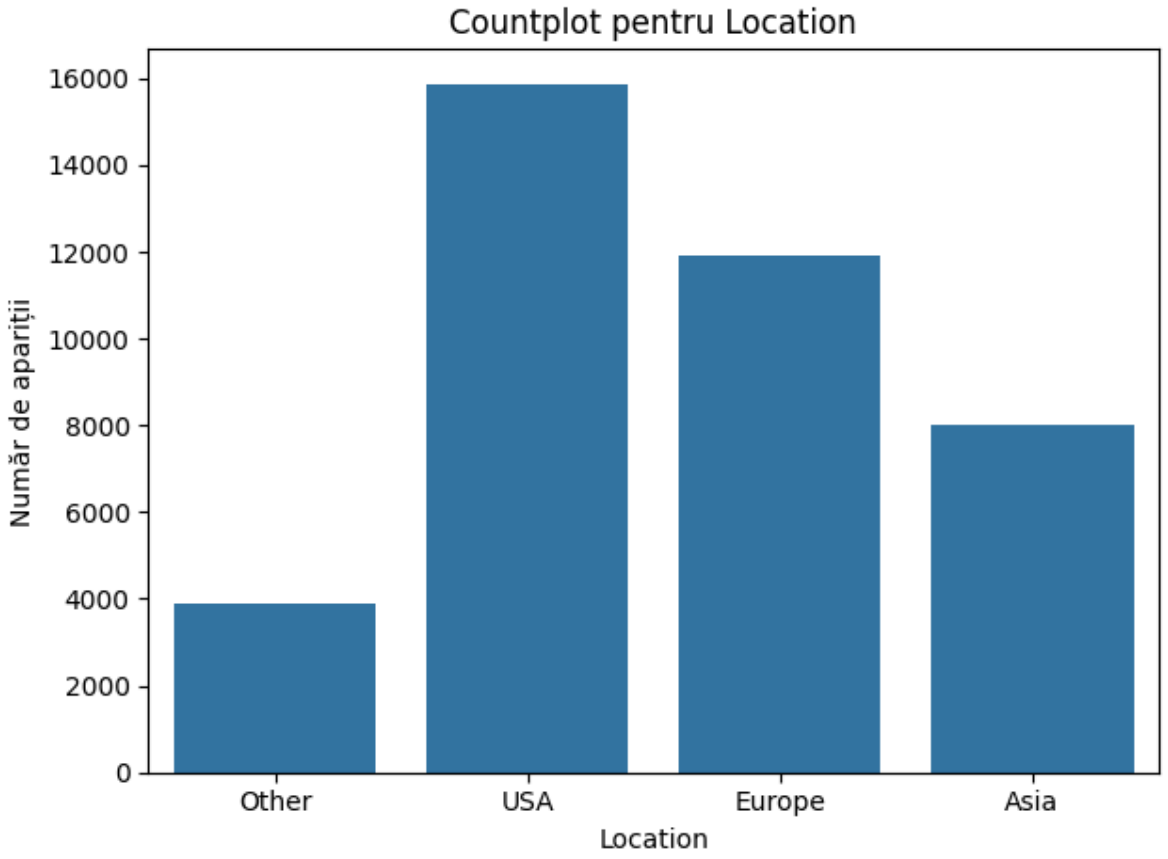
- Sports: ~8.000 jucatori
- Strategy, Action, RPG, Simulation: fiecare ~7.500 jucatori
- Diferenta intre genuri este minima ~500 jucatori

Ce suspiciuni/idei putem formula?

- Platforma reuseste sa atraga o audienta diversa cu preferinte variate
- Sports ar putea beneficia de marketing superior sau de un gameplay mai atractiv
- Echilibrul sugereaza ca nu exista o nisa dominanta, ci o platforma incluziva
- Popularitatea Sports poate indica apetitul pentru competitie si simplitate

Ce preprocesari ar trebui sa aplicam?

- Completarea valorilor lipsa (39.666 din 40.000)
- One-hot encoding pentru variabila categorica nominala
- Verificarea daca distributia relativ uniforma nu ascunde pattern-uri in alte variabile



Ce observam?

Distributia geografica arata o concentrare in anumite regiuni:

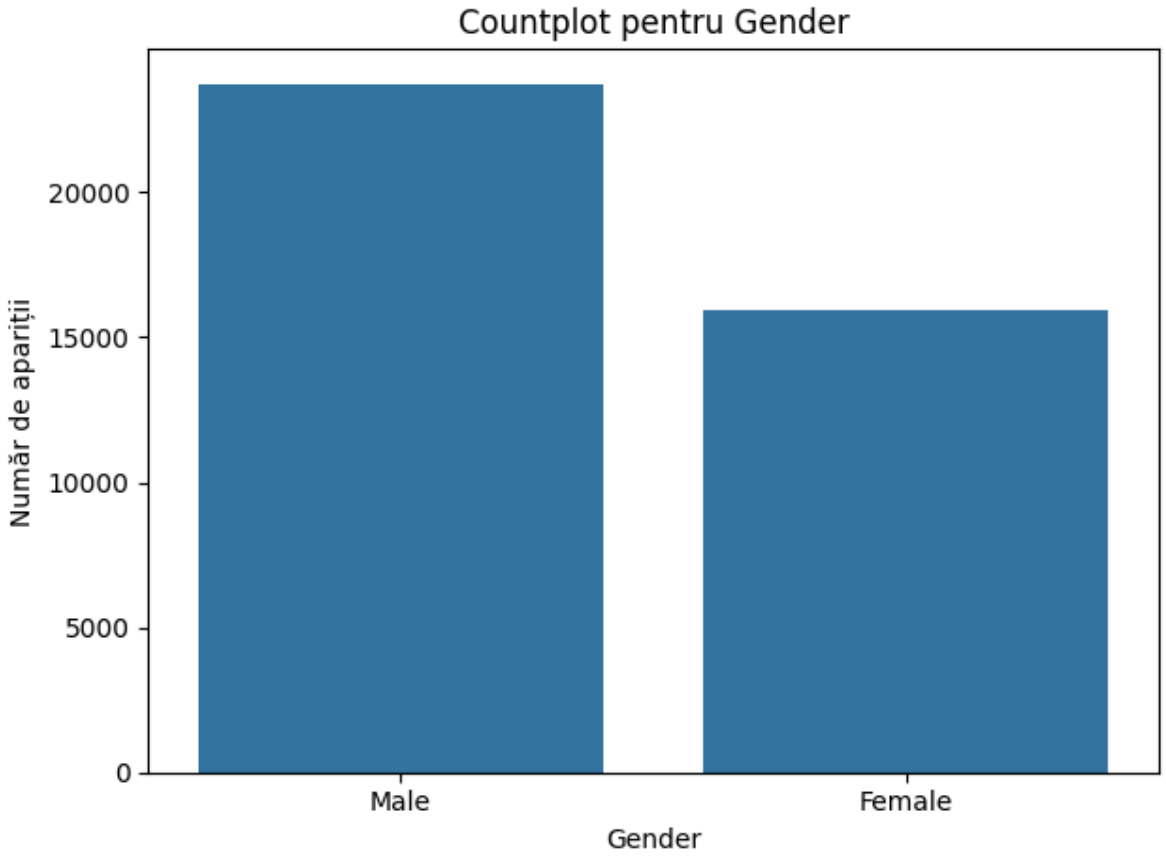
- USA: ~15.877 jucatori
- Europe: ~8.000 jucatori
- Asia: ~8.000 jucatori
- Other: ~7.814 jucatori

Ce suspiciuni/idei putem formula?

- Piata americana este cea mai dezvoltata sau cel mai bine penetrata
- Diferentele culturale in preferintele de gaming pot aparea
- USA poate avea cei mai activi early adopters pentru gaming online

Ce preprocesari ar trebui sa aplicam?

- Completarea valorilor lipsa
- One-hot encoding pentru variabila categorica nominala
- Considerarea unei analize separate pe regiuni pentru pattern-uri locale



Ce observam?

Exista o predominanta clara a jucatorilor de sex masculin:

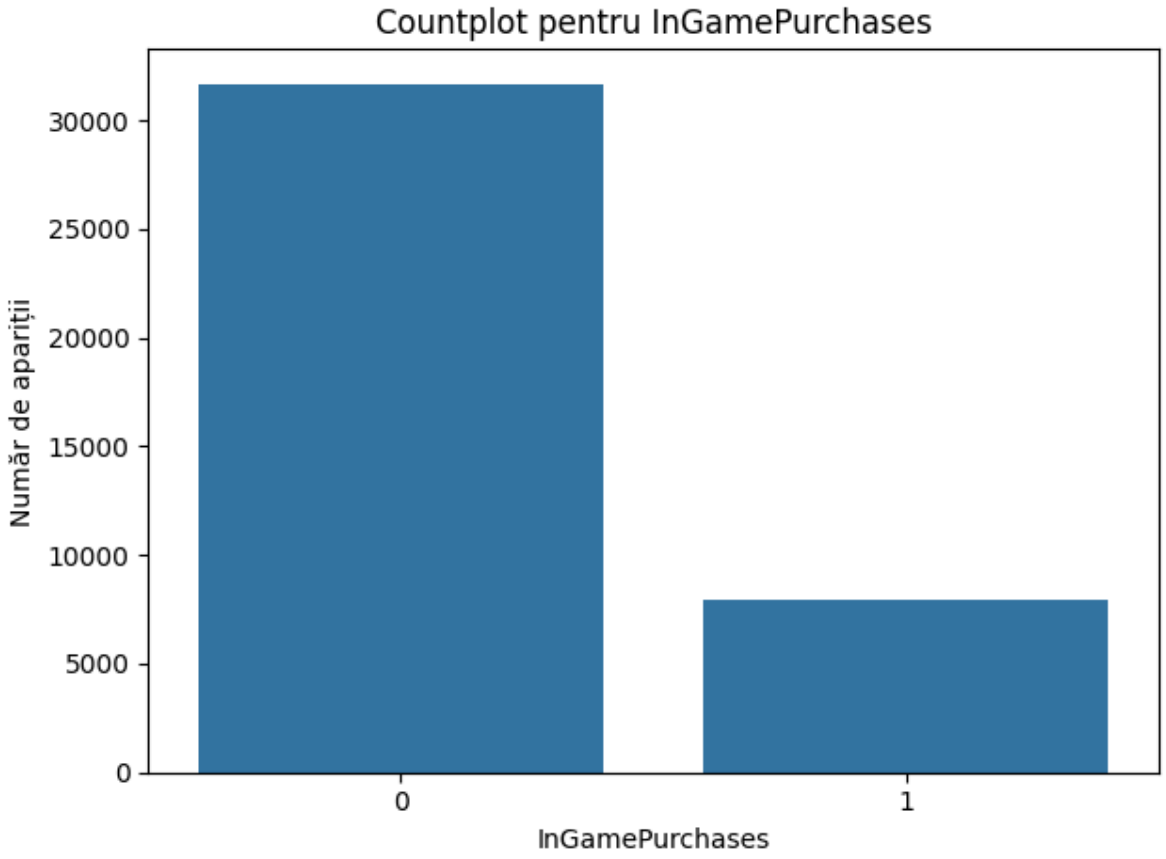
- Male: ~23.745 jucatori
- Female: ~15.934 jucatori

Ce suspiciuni/idei putem formula?

- Gaming-ul online ramane inca o activitate cu predominanta masculina
- Diferentele culturale sau de marketing pot influenta participarea femeilor
- Genul poate fi un predictor important pentru alte comportamente in joc

Ce preprocesari ar trebui sa aplicam?

- Completarea valorilor lipsa
- Label encoding: Male=1, Female=0 sau one-hot encoding
- Analizarea corelatiei cu alte variabile pentru a intelege impactul genului



Ce observam?

Majoritatea covarsite a jucatorilor nu fac achizitii in joc:

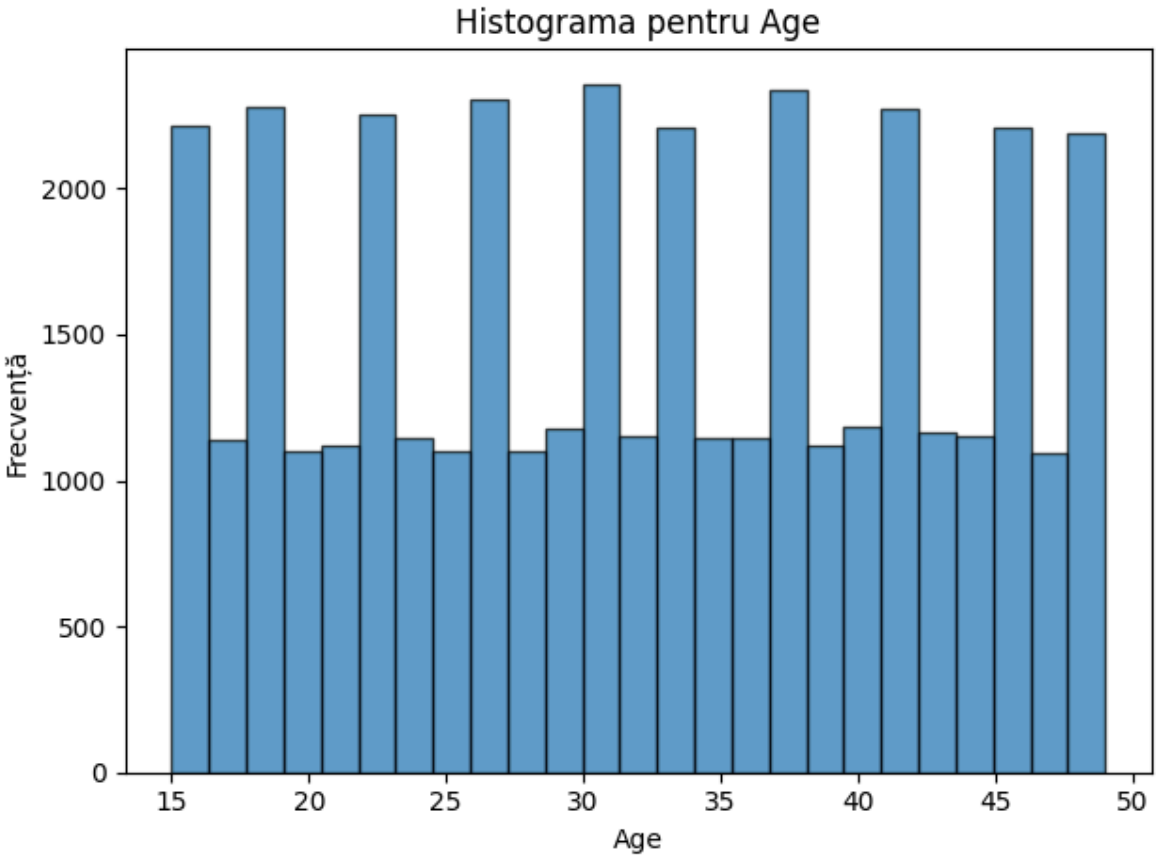
- Nu fac achizitii (0): ~31.685 jucatori
- Fac achizitii (1): ~7.979 jucatori

Ce suspiciuni/idei putem formula?

- Majoritatea jucatorilor sunt casual si nu sunt dispusi sa investeasca financiar
- Preturile sau ofertele pot sa nu fie suficient de atractive

Ce preprocesari ar trebui sa aplicam?

- Completarea valorilor lipsa
- Variabila este deja binary encoded (0/1)
- Considerarea dezechilibrului de clase in modelele predictive



Ce observam?

Distributia varstei pare sa urmeze o distributie normala:

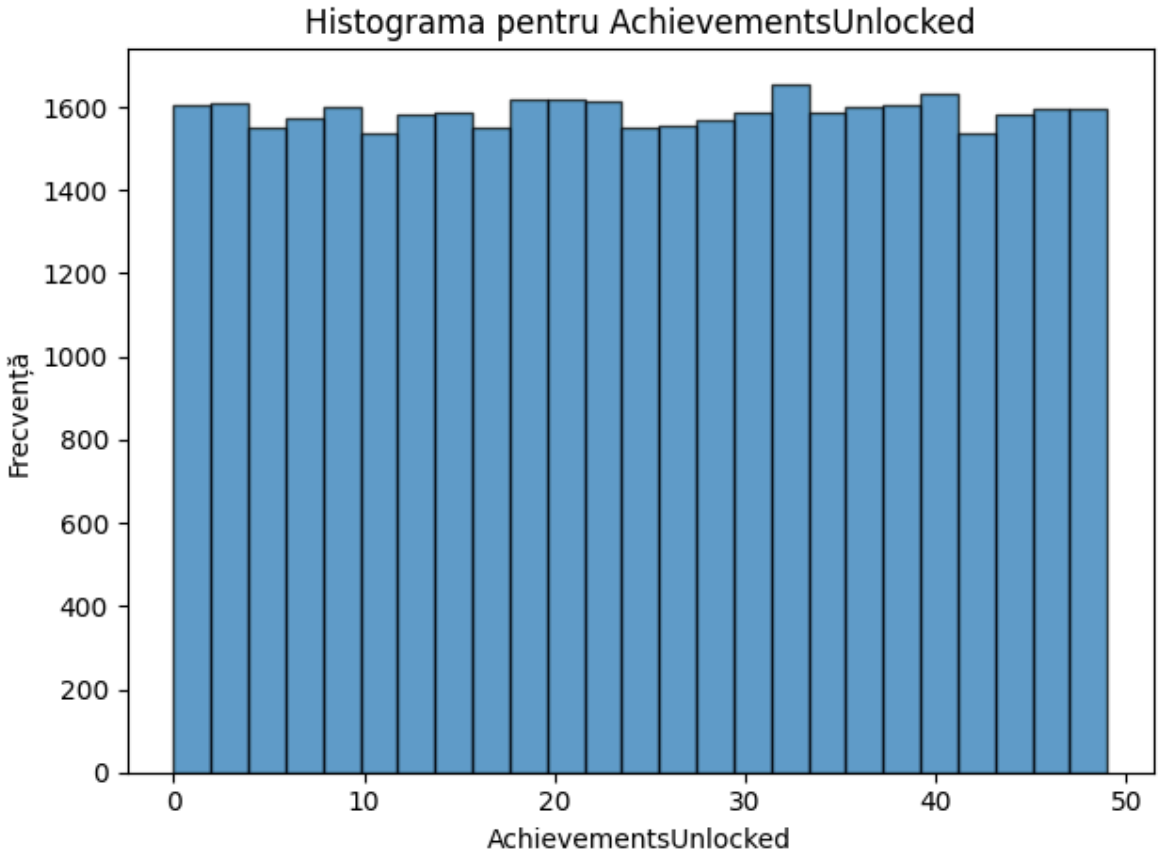
- Media: 32 de ani
- Range: 15-49 ani
- Mediana: 32 ani
- Deviatia standard: 10 ani

Ce suspiciuni/idei putem formula?

- Gaming-ul online atrage predominant adulti tineri si de varsta mijlocie
- Distributia normala sugereaza o baza larga de utilizatori fara varfuri extreme
- Lipsa copiilor sub 15 ani poate indica restrictii de varsta sau de acces
- Absenta adultilor peste 49 ani poate reflecta bariere tehnologice

Ce preprocesari ar trebui sa aplicam?

- Completarea valorilor lipsa cu media
- Normalizarea sau standardizarea
- Considerarea crearii de categorii de varsta



Ce observam?

Histograma pentru AchievementsUnlocked arata:

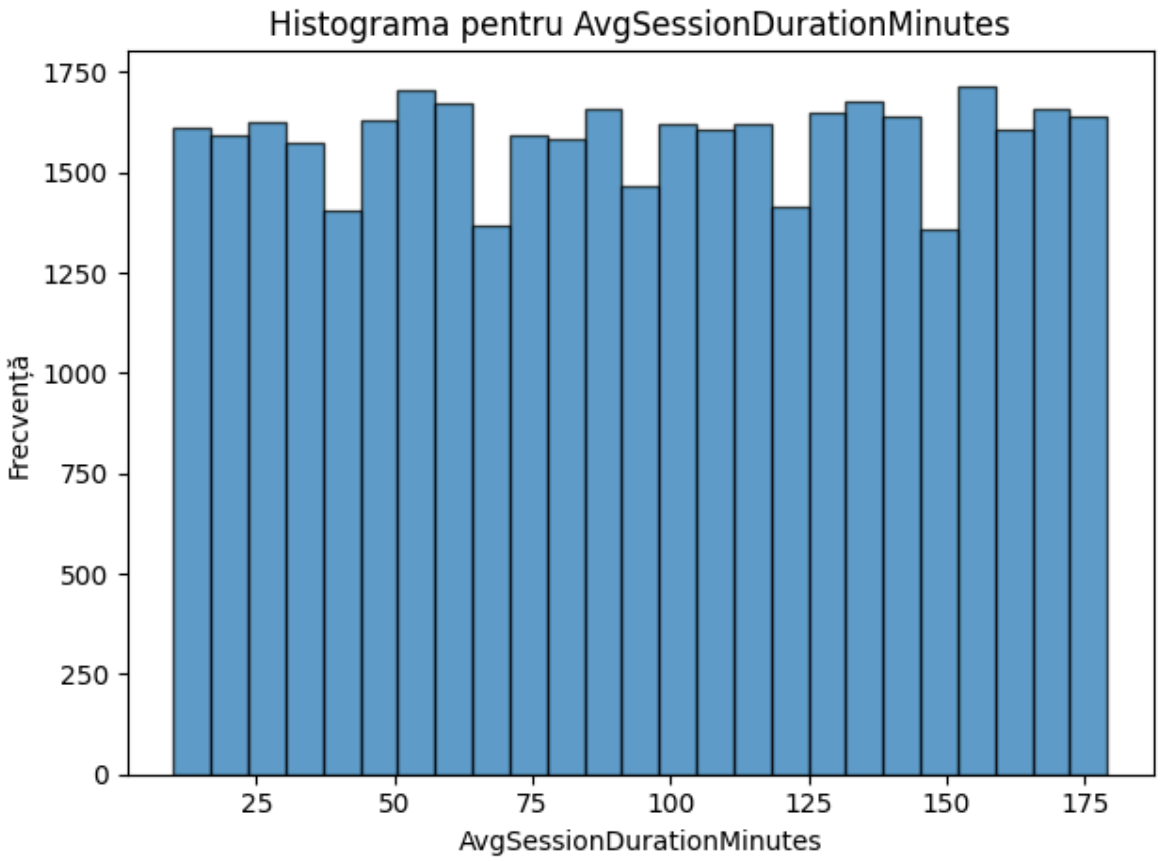
- Distributie aproape uniforma pe intreg range-ul 0-49
- Frecventa constanta de aproximativ 1600 jucatori pentru fiecare nivel de realizari
- Fara varfuri sau vaduri semnificative in distributie

Ce suspiciuni/idei putem formula?

- Sistemul de achievements este bine echilibrat si accesibil pentru toti jucatorii
- Nu exista realizari grele care sa opreasca progresul jucatorilor
- Distributia uniforma sugereaza ca jocul nu favorizeaza anumite tipuri de jucatori

Ce preprocesari ar trebui sa aplicam?

- Completarea valorilor lipsa cu media
- Normalizarea valorilor pentru algoritmi ML sensibili la scala
- Considerarea crearii de categorii: incepatori (0-15), intermediari (16-35), avansati (36-49)



Histograma pentru AvgSessionDurationMinutes arata:

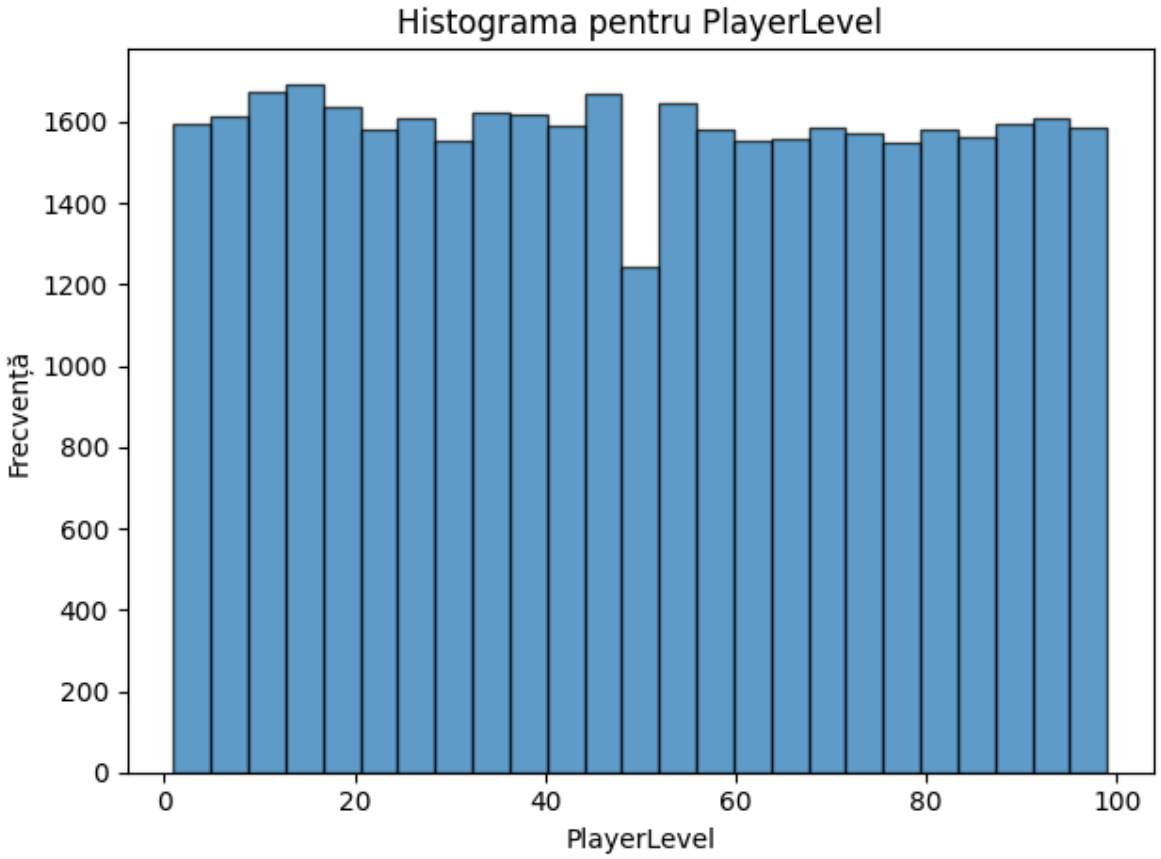
- Distributie relativ uniforma pe range-ul 10-179 minute
- Frecventa variabila intre 1300-1700 jucatori per interval
- Range complet: 10-179 minute (aproape 3 ore maxim)

Ce suspiciuni/idei putem formula?

- Jucatorii au preferinte diverse pentru durata sesiunilor de joc
- Varfurile la 50 si 160 minute sugereaza doua tipuri distincte de jucatori:
    - Casual players (50 min) - o sesiune scurta, relaxanta
    - Dedicated players (160 min) - sesiuni lungi, imersive
- Distributia sugereaza ca jocul poate sustine atat sesiuni scurte cat si lungi

Ce preprocesari ar trebui sa aplicam?

- Completarea valorilor lipsa cu media
- Normalizarea pentru algoritmi ML
- Considerarea crearii de categorii: scurt (10-60 min), mediu (61-120 min), lung (121-179 min)
- Analizarea corelatiei cu tipul de joc si nivelul de engagement



Ce observam?

Histograma pentru PlayerLevel arata:

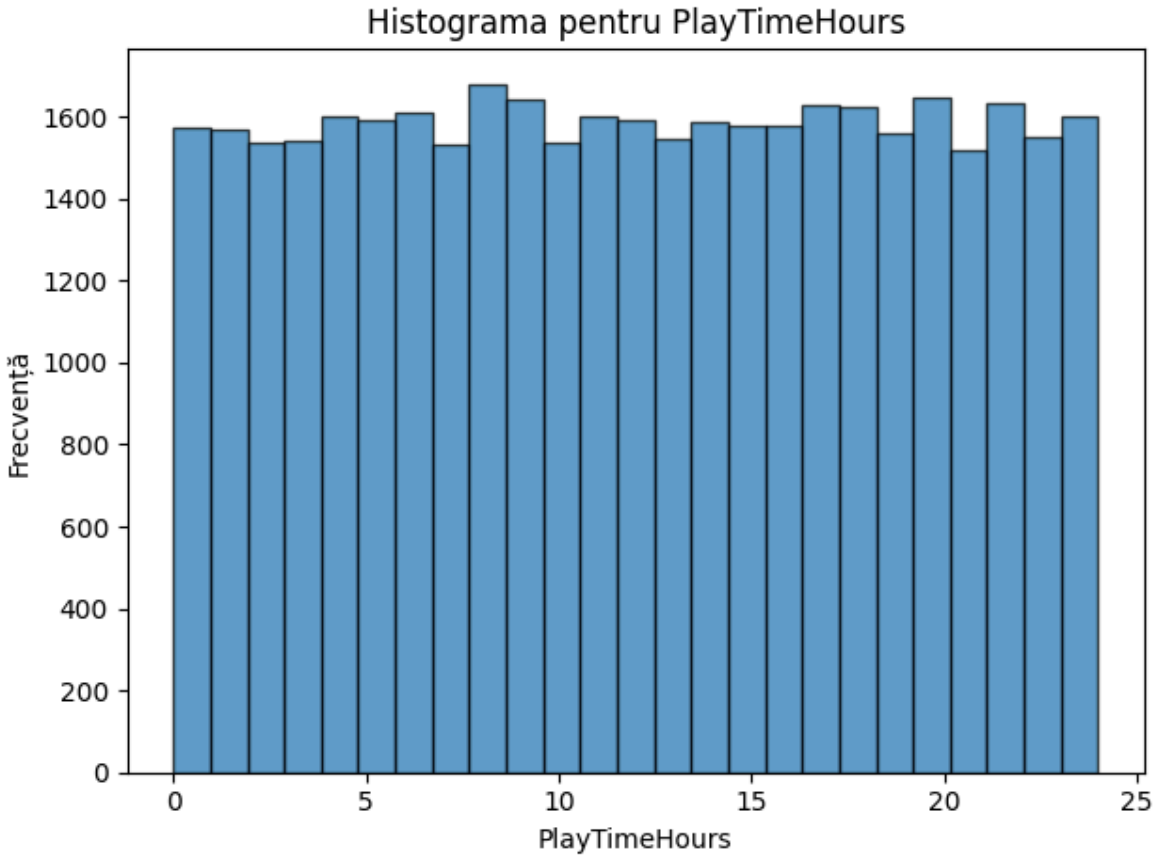
- Distributie aproape uniforma pe range-ul 1-99
- Frecventa relativ constanta de aproximativ 1600 jucatori per nivel
- Varf usor in jurul nivelurilor 10-15 (jucatori incepatori)
- Scadere notabila la nivelul 50 (aproximativ 1200 jucatori)
- Recuperare dupa nivelul 50, cu distributie uniforma pana la nivel 99

Ce suspiciuni/idei putem formula?

- Scaderea la nivelul 50 sugereaza o posibila bariera psihologica sau de dificultate
- Distributia uniforma sugereaza ca jocul reuseste sa retina jucatorii pe termen lung

Ce preprocesari ar trebui sa aplicam?

- Completarea valorilor lipsa cu media
- Normalizarea pentru algoritmi ML
- Investigarea mai detaliata a scaderii de la nivelul 50 (posibil indicator important)



Ce observam?

Histograma pentru PlayTimeHours arata:

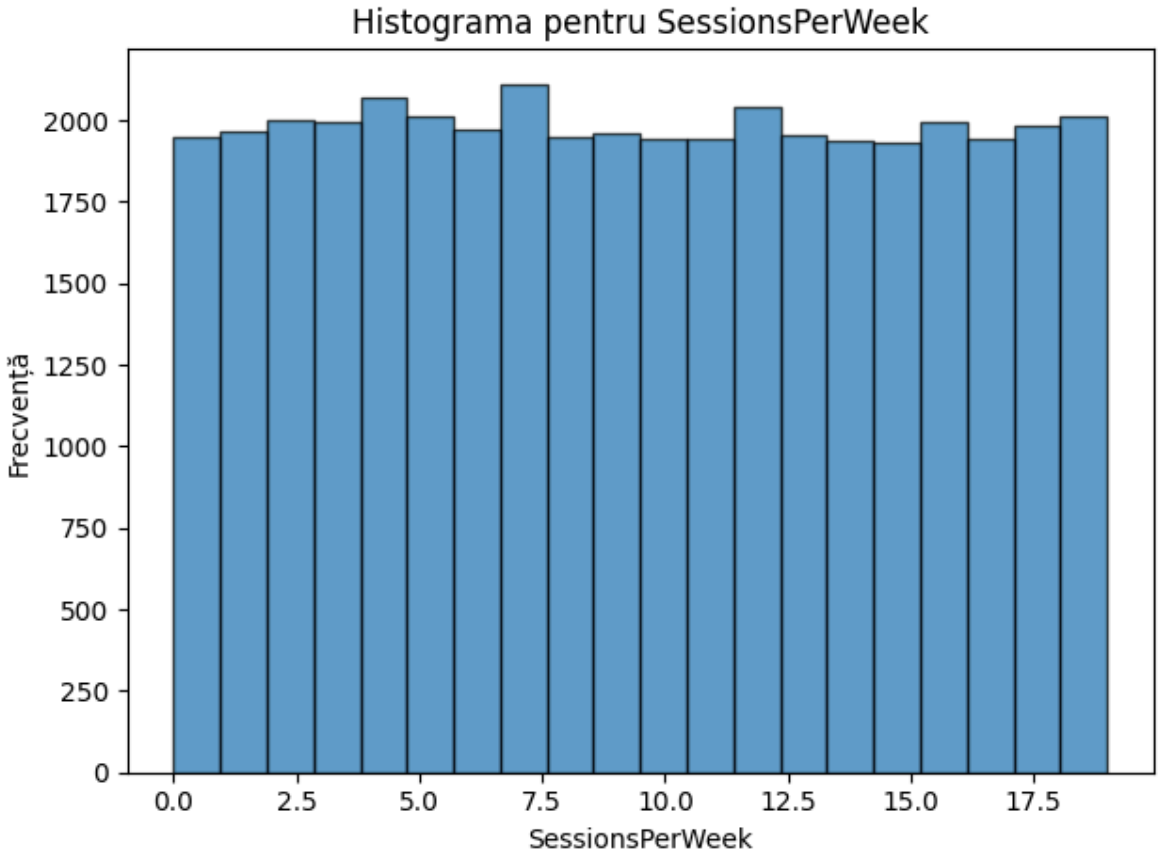
- Distributie relativ uniforma pe range-ul 0-24 ore
- Frecventa constanta intre 1500-1700 jucatori per interval
- Varf usor in jurul valorii de 10 ore
- Scadere minora la inceputul si sfarsitul distributiei

Ce suspiciuni/idei putem formula?

- Jucatorii au comportamente foarte diverse in ceea ce priveste timpul petrecut jucand
- Varful la 10 ore poate indica un nivel optimal de engagement pentru majoritatea
- Range-ul extins (pana la 24 ore) indica prezenta unor jucatori extrem de dedicati

Ce preprocesari ar trebui sa aplicam?

- Completarea valorilor lipsa cu media
- Normalizarea pentru comparabilitate cu alte variabile
- Verificarea corelatiei cu PlayerLevel si AchievementsUnlocked



Ce observam?

Histograma pentru SessionsPerWeek arata:

- Distributie aproape uniforma pe range-ul 0-19 sesiuni
- Frecventa relativ constanta intre 1900-2100 jucatori per interval
- Varf usor la 6-7 sesiuni pe saptamana (o sesiune pe zi)
- Scadere minora la valorile extreme (0 si 19 sesiuni)
- Comportament consistent pe majoritatea range-ului

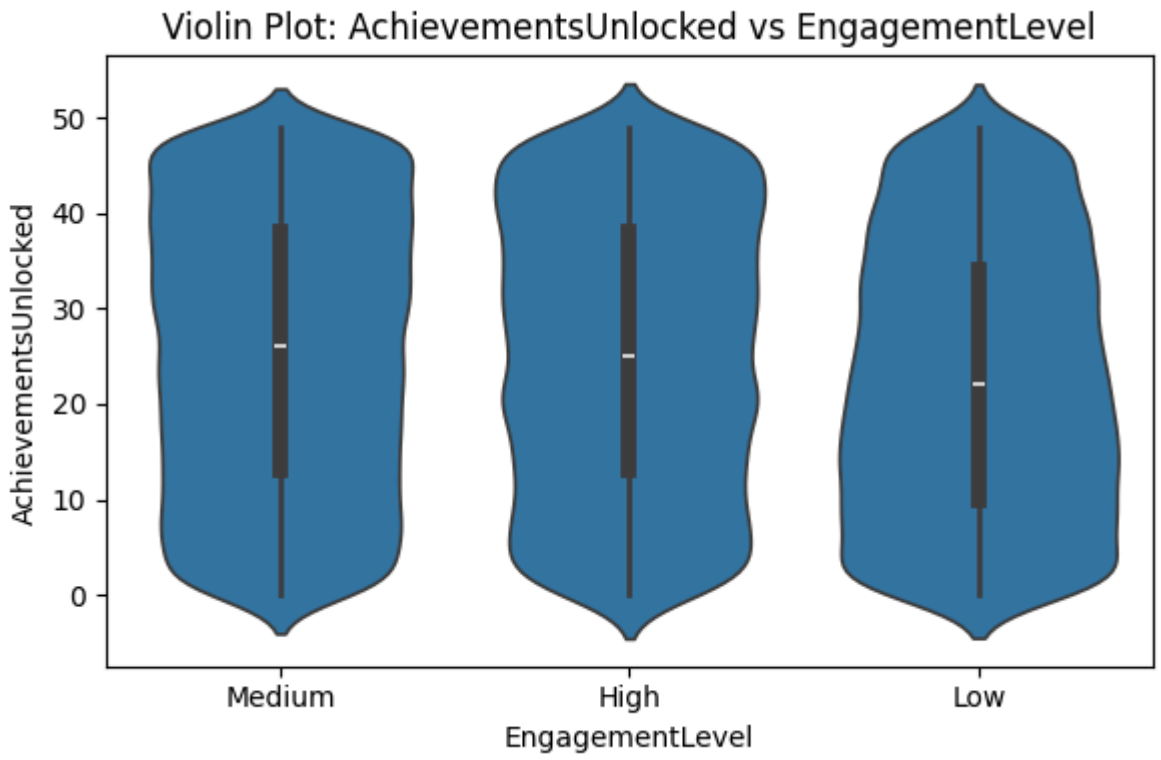
Ce suspiciuni/idei putem formula?

- Majoritatea jucatorilor au o rutina regulata de gaming (6-7 sesiuni = zilnic)
- Distributia uniforma sugereaza flexibilitate in programele jucatorilor
- Varful la 6-7 sesiuni indica ca multi jucatori joaca aproape zilnic
- Prezenta jucatorilor cu 19 sesiuni (2-3 pe zi) indica o baza dedicata

Ce preprocesari ar trebui sa aplicam?

- Completarea valorilor lipsa cu media Normalizarea pentru algoritmi ML
- Analizarea corelatiei cu EngagementLevel si PlayTimeHours

-Analiza relatiilor cu variabila tinta:



Ce observam?

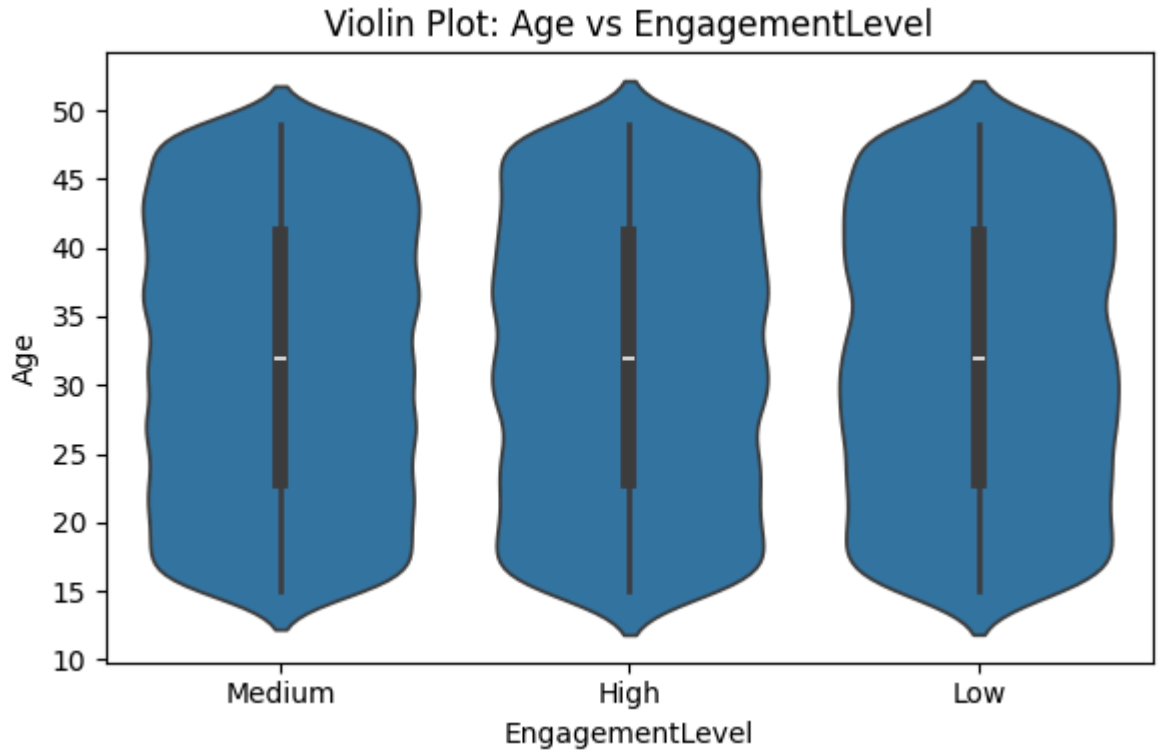
- Distributiile pentru AchievementsUnlocked sunt similare intre grupuri.
- Toate cele 3 niveluri de EngagementLevel (Low, Medium, High) au o distributie aproape simetrica si continua.

Ce suspiciuni/idei putem formula?

- EngagementLevel nu pare sa influenteze puternic AchievementsUnlocked.
- Posibil ca variabila EngagementLevel sa nu fie un predictor semnificativ.
- Alte variabile ar putea explica mai bine variatia.

Ce preprocesari ar trebui sa aplicam?

- Standardizare sau normalizare pentru algoritmi sensibili la scala.
- Eventual eliminarea outlierilor, daca apar.



Ce observam?

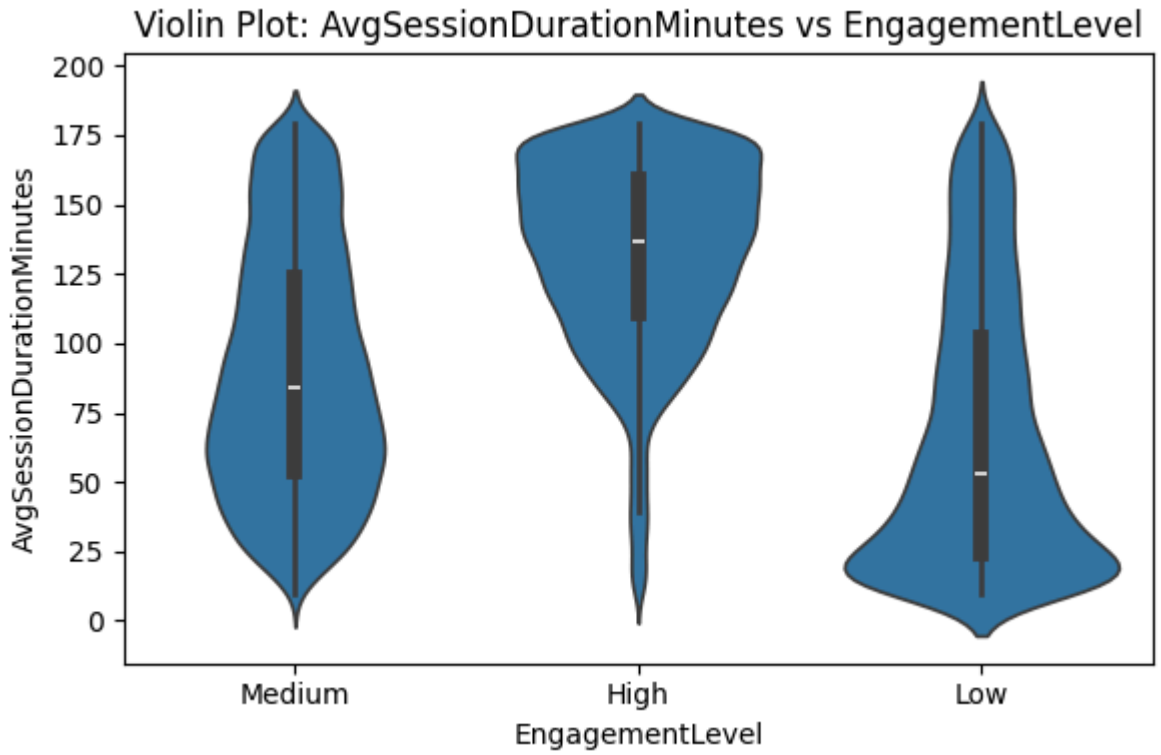
- Distributiile pentru Age sunt similare intre grupuri.
- Toate cele 3 niveluri de EngagementLevel (Low, Medium, High) au o distributie aproape simetrica si continua.

Ce suspiciuni/idei putem formula?

- EngagementLevel nu pare sa influenteze puternic Age.
- Posibil ca variabila EngagementLevel sa nu fie un predictor semnificativ pentru Age.

Ce preprocesari ar trebui sa aplicam?

- Standardizare sau normalizare pentru algoritmi sensibili la scala.
- Eventual eliminarea outlierilor, daca apar.



Ce observam?

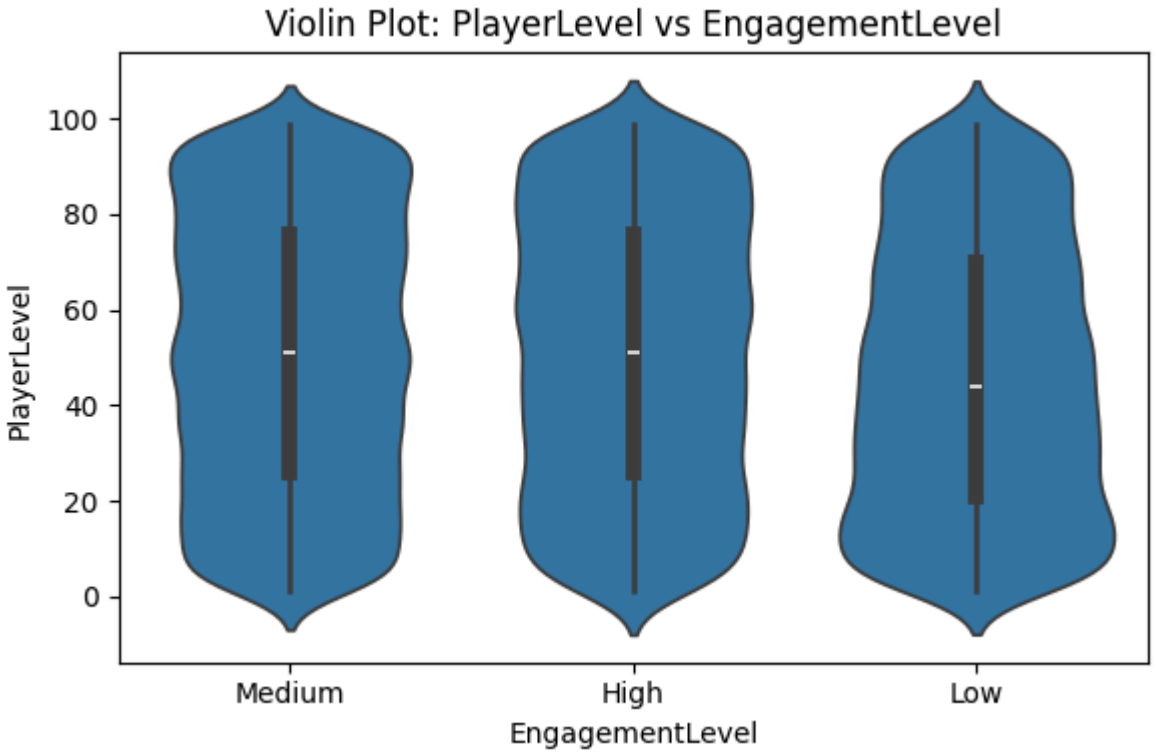
- Distributiile pentru AvgSessionDurationMinutes variaza intre grupuri.
- Nivelul High are o distributie ingusta, centrata sub 50 minute.
- Nivelul Medium arata o distributie mai larga, intre 50–150 minute.
- Nivelul Low are o distributie foarte larga, de la 0 la 200 minute.

Ce suspiciuni/idei putem formula?

- Utilizatorii cu High Engagement au sesiuni scurte si consistente.
- Medium Engagement sugereaza sesiuni mai diverse ca durata.
- Low Engagement indica fie utilizatori inactivi, fie sesiuni extrem de lungi ocazionale.
- Durata sesiunii ar putea fi un factor cheie al implicarii.

Ce preprocesari ar trebui sa aplicam?

- Standardizare sau normalizare pentru analize ulterioare.
- Analiza corelatiei cu EngagementLevel si alte variabile.



Ce observam?

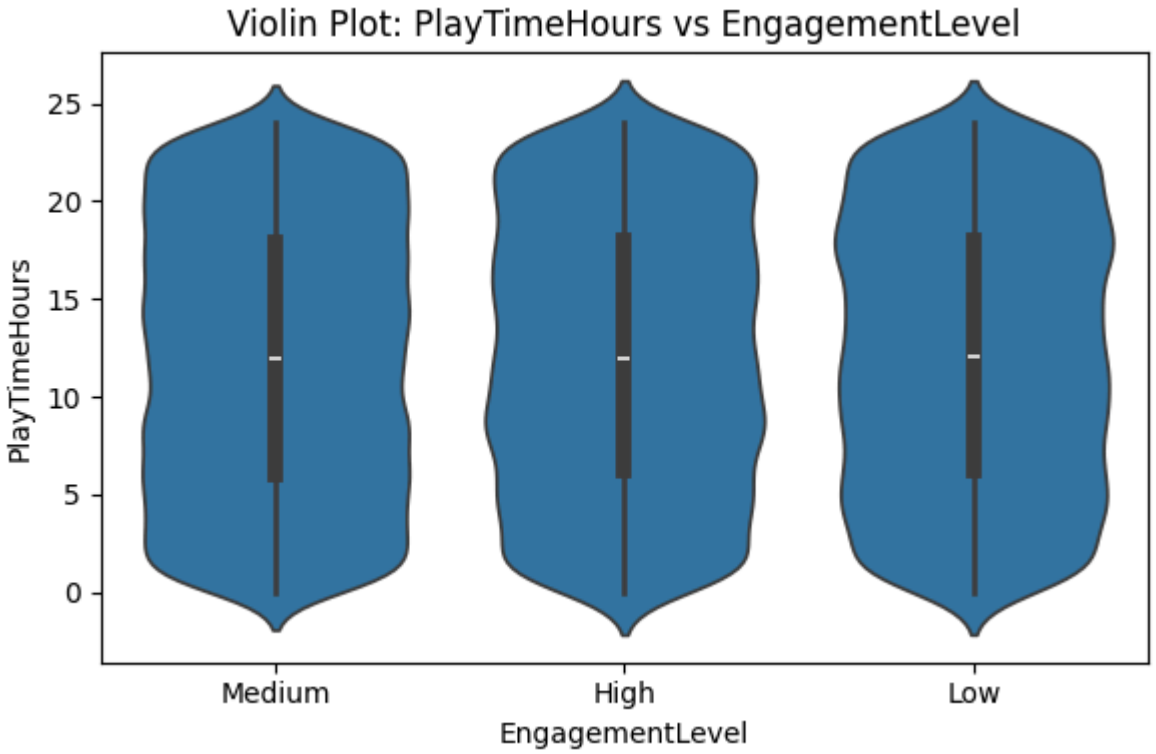
- Distributiile pentru PlayerLevel sunt similare intre grupuri.
- Toate cele 3 niveluri de EngagementLevel (Low, Medium, High) au o distributie simetrica si continua.

Ce suspiciuni/idei putem formula?

- EngagementLevel nu pare sa influenteze semnificativ PlayerLevel.
- Posibil ca progresul (PlayerLevel) sa fie independent de implicare.
- Alti factori (ex. timp petrecut, dificultate) ar putea afecta nivelul.

Ce preprocesari ar trebui sa aplicam?

- Standardizare sau normalizare pentru analize ulterioare.
- Eventual eliminarea outlierilor, daca apar.



Ce observam?

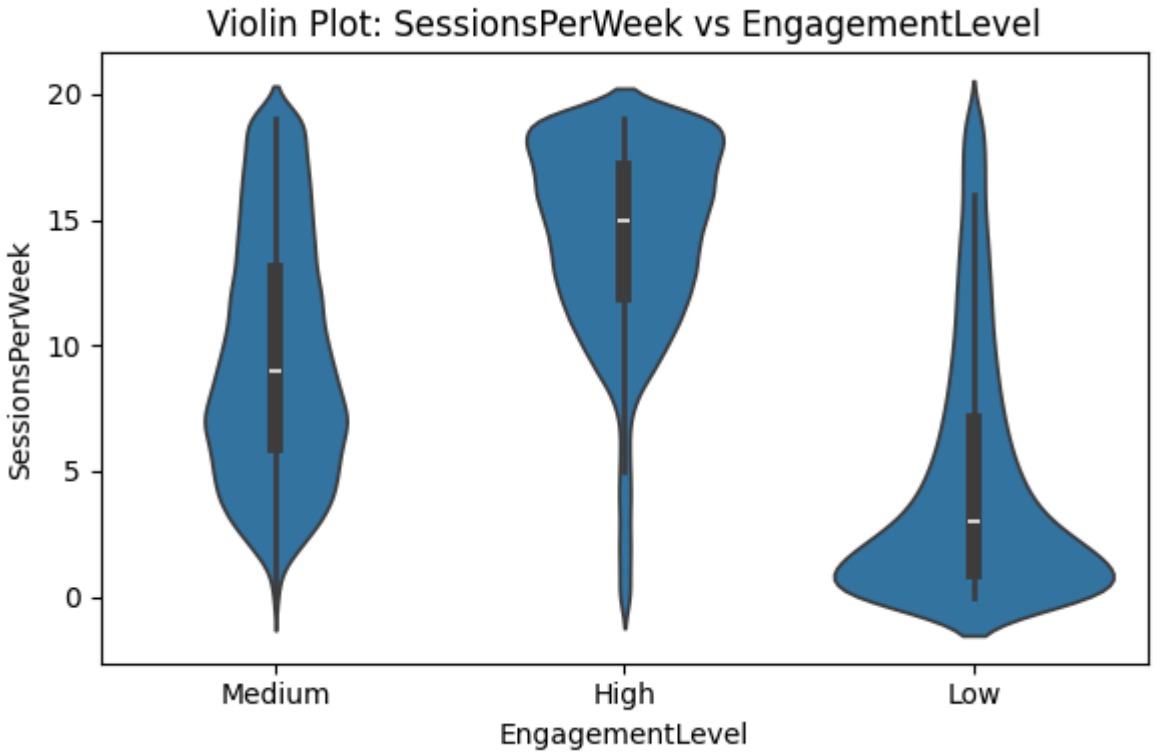
- Distributiile pentru PlayTimeHours sunt similare intre grupuri.
- Toate cele 3 niveluri de EngagementLevel (Low, Medium, High) au o distributie simetrica si continua.

Ce suspiciuni/idei putem formula?

- EngagementLevel nu pare sa influenteze semnificativ PlayTimeHours.
- Timpul de joc ar putea fi influentat de alti factori (ex. preferinte personale).
- Utilizatorii au un timp de joc consistent, indiferent de nivelul de implicare.

Ce preprocesari ar trebui sa aplicam?

- Standardizare sau normalizare pentru analize ulterioare.
- Eventual eliminarea outlierilor, daca apar.



Ce observam?

- Distributiile pentru SessionsPerWeek variaza intre grupuri.
- Nivelul High are o distributie ingusta, centrata sub 10 sesiuni.
- Nivelul Medium arata o distributie similara, dar mai larga, pana la 15 sesiuni.
- Nivelul Low are o distributie foarte larga, de la 0 la 20 sesiuni.

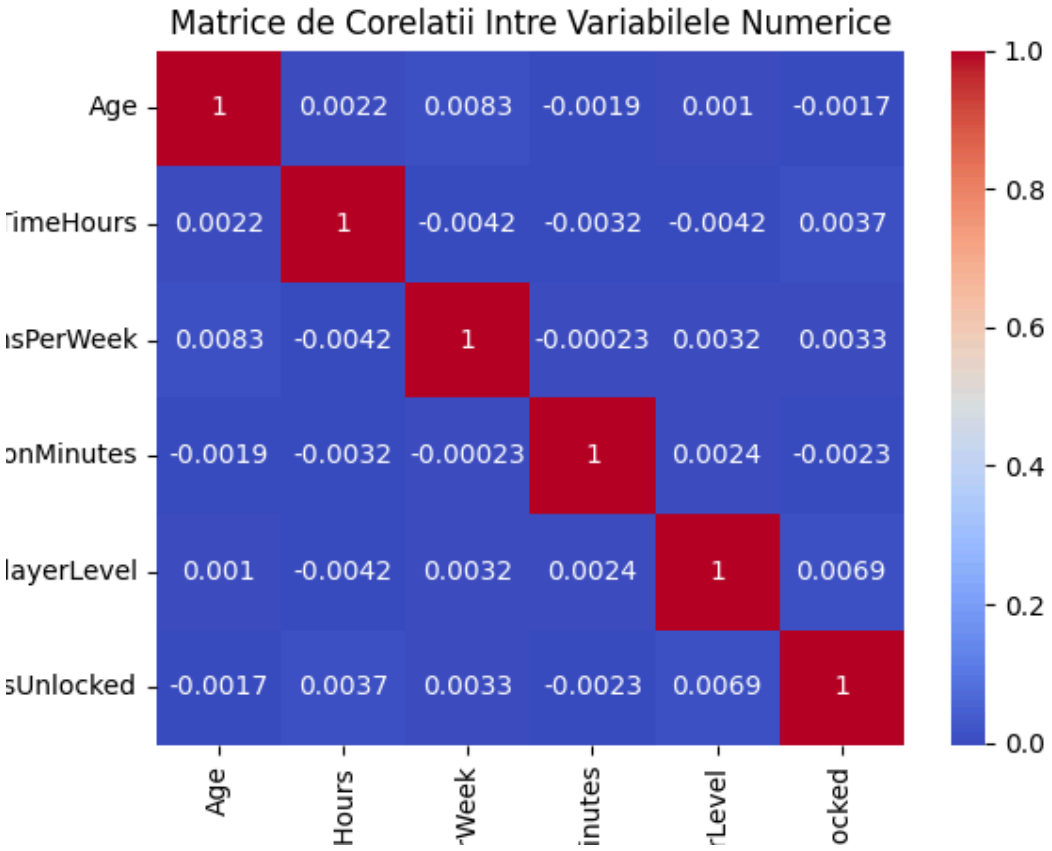
Ce suspiciuni/idei putem formula?

- Utilizatorii cu High Engagement au un numar stabil de sesiuni saptamanale.
- Medium Engagement sugereaza o implicare moderata si constanta.
- Low Engagement indica utilizatori cu sesiuni rare sau extrem de variate.
- Frecventa sesiunilor ar putea reflecta nivelul de interes.

Ce preprocesari ar trebui sa aplicam?

- Standardizare sau normalizare pentru analize ulterioare.
- Analiza corelatiei cu EngagementLevel si alte variabile.

-Analiza corelatiilor:



Ce observam?

- Correlatii foarte slabe: Toate valorile de corelatie sunt extrem de mici (sub 0.01 in valoare absoluta), ceea ce indica lipsa unor relatii liniare semnificative intre variabile.
- Diagonala perfecta: Valorile de 1.0 pe diagonala principala confirma ca fiecare variabila este perfect corelata cu ea insasi.
- Distributie aparent aleatoare: Valorile pozitive si negative par sa fie distribuite fara un pattern clar, sugerand ca nu exista relatii sistematice intre variabilele analizate.

Ce suspiciuni/idei putem formula?

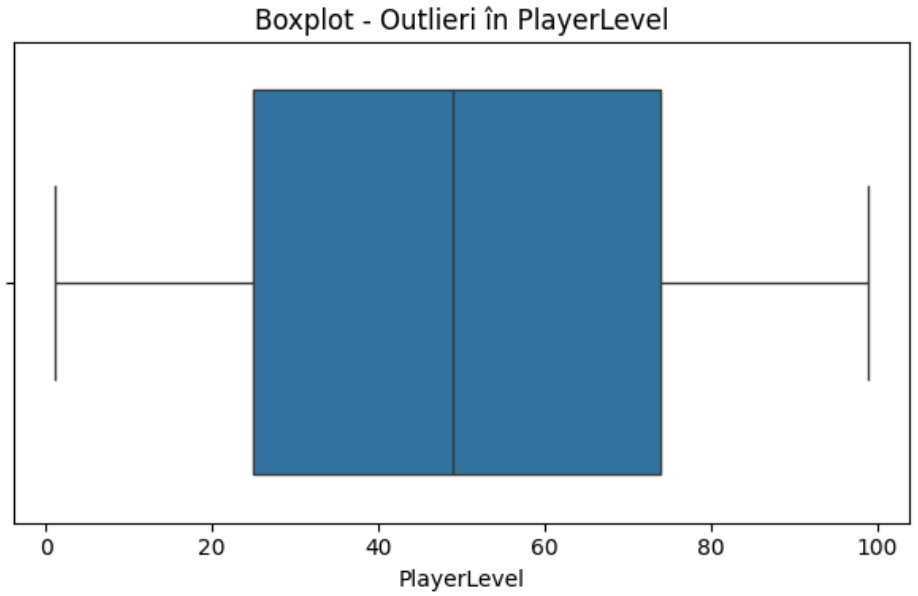
Variabile independente: Este posibil ca variabilele sa fie cu adevarat independente statistic, ceea ce ar fi normal pentru unele combinatii (ex: varsta vs minutele de joc).

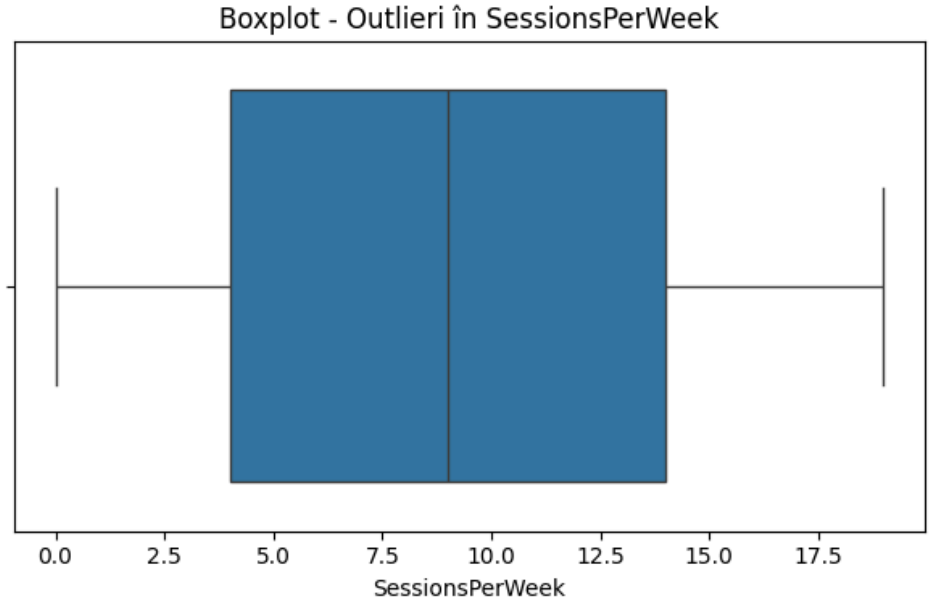
Ce preprocesari ar trebui sa aplicam?

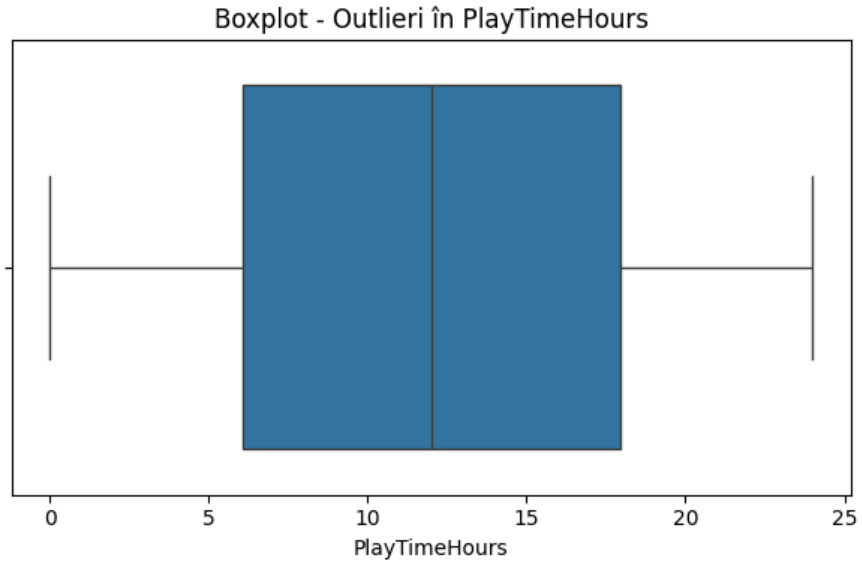
- Detectarea si tratarea outlierilor
- Verificarea tipurilor de date:
- Normalizare corecta: Aplicarea standardizarii (z-score) sau normalizarii min-max doar daca este necesara.

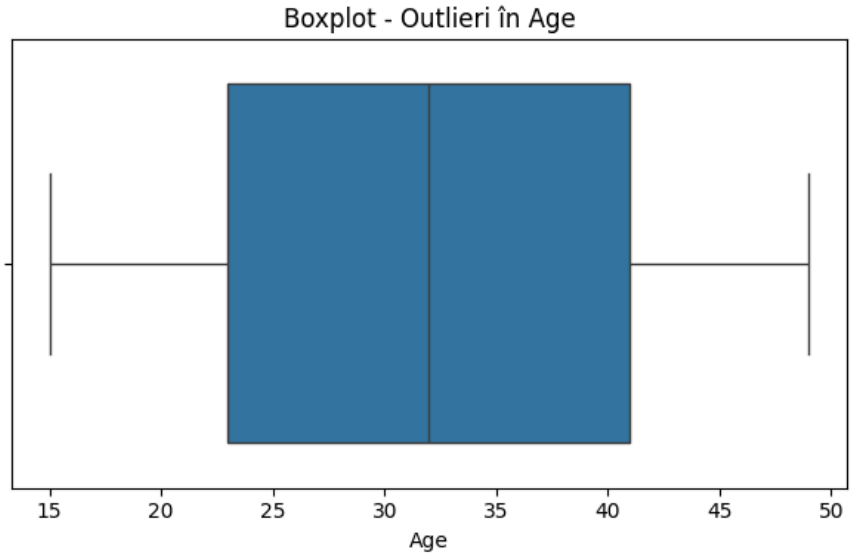
- Standardizare a datelor pentru analize ulterioare.
- Verificarea valorilor lipsă sau extreme.
- Analiza corelațiilor mai profunde cu variabile suplimentare (ex. EngagementLevel).

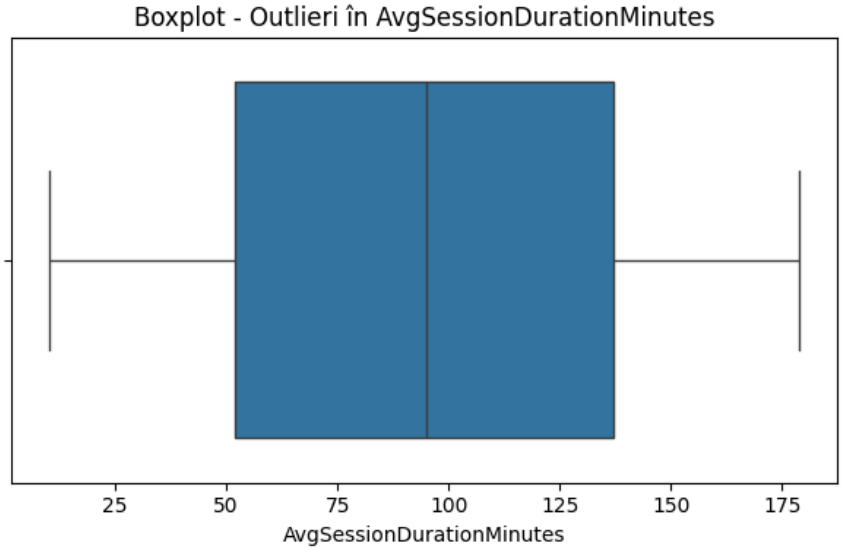
-Analiza:

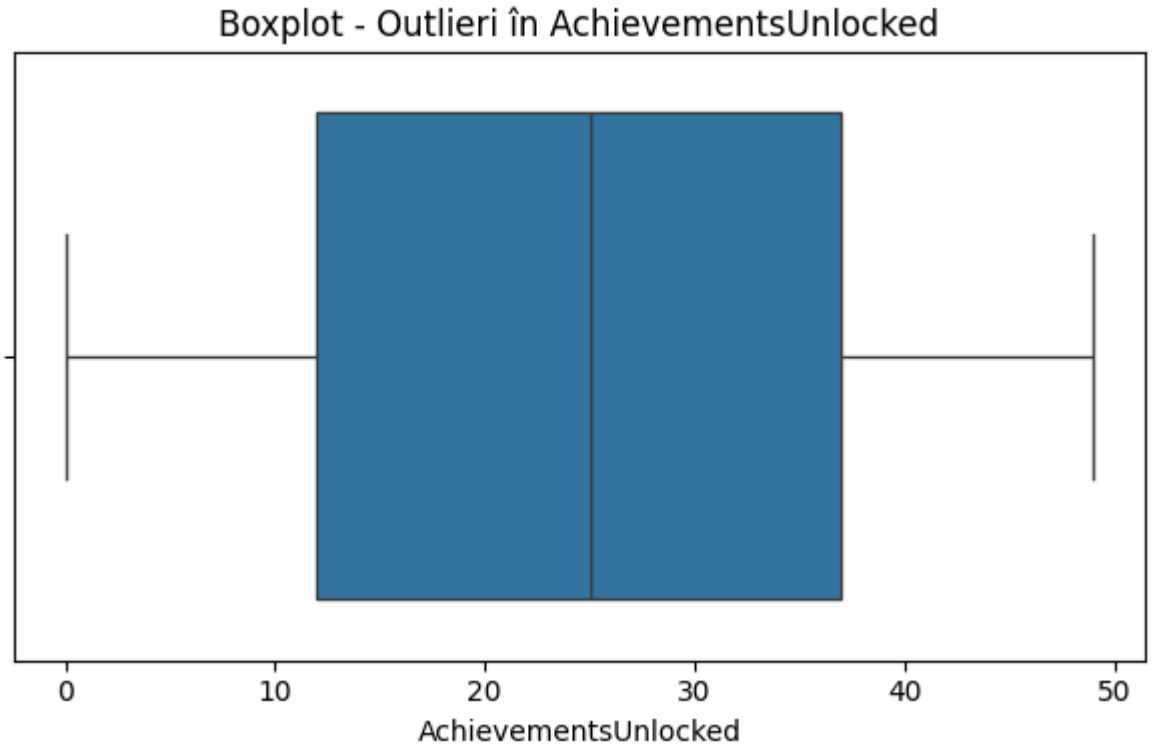





Boxplot - Outlieri în PlayTimeHours
0
5
10
15
20
25
PlayTimeHours


Boxplot - Outlieri în Age
15
20
25
30
35
40
45
50
Age


Boxplot - Outlieri în AvgSessionDurationMinutes
25
50
75
100
125
150
175
AvgSessionDurationMinutes



Ce observam?

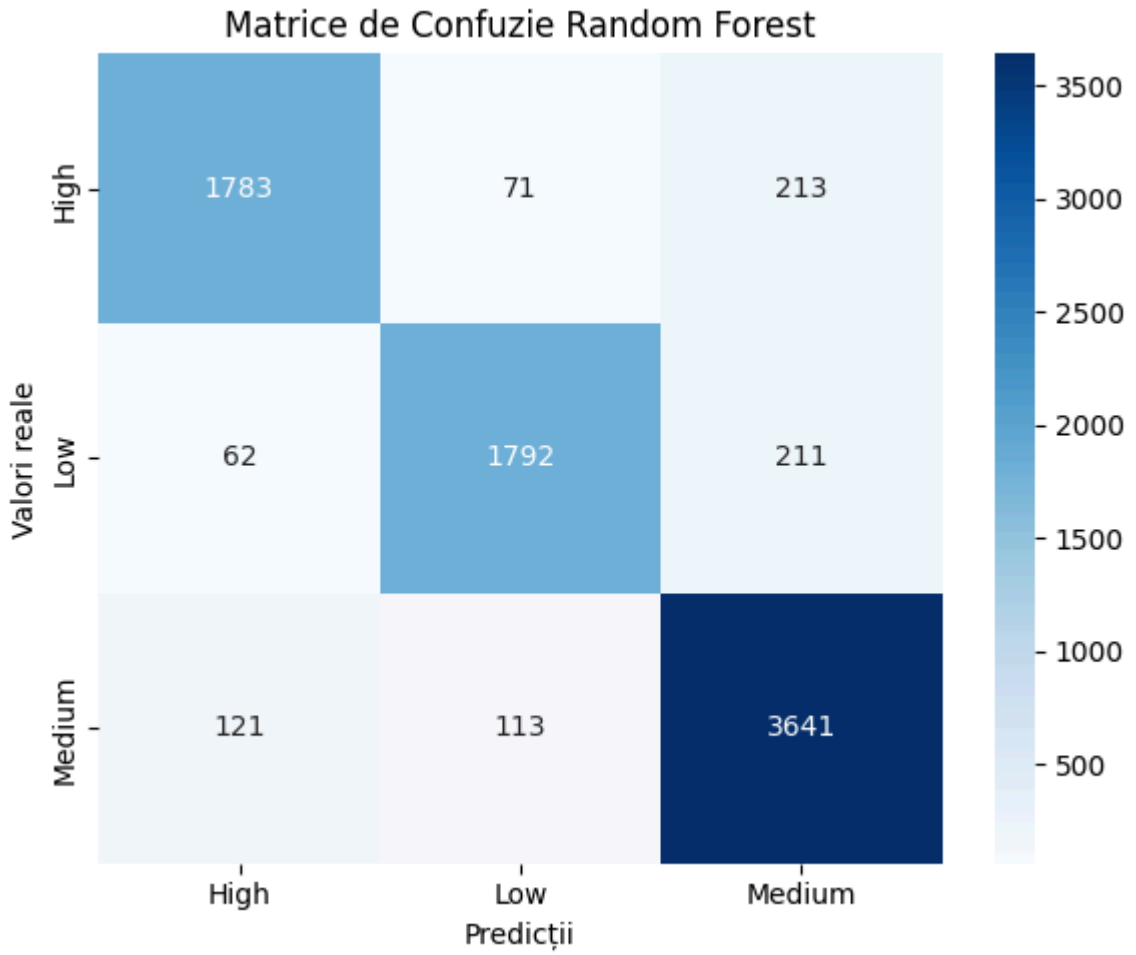
- Boxplot-urile pentru Age, AvgSessionDurationMinutes, PlayerLevel, AchievementsUnlocked, SessionsPerWeek și PlayTimeHours nu prezinta outlieri vizibili.
- Distributiile sunt relativ simetrice, cu intervale intercuartilice (IQR) bine definite.
- Valorile extreme (minime si maxime) sunt apropiate de marginile IQR, fara abateri semnificative.

Ce suspiciuni/idei putem formula?

- Lipsa outlierilor sugereaza o distributie uniforma si consistenta a datelor.
- Utilizatorii par sa aiba un comportament omogen in ceea ce priveste aceste variabile.
- Variabilele numerice nu prezinta valori anormale care sa distorsioneze analizele ulterioare.

Ce preprocesari ar trebui sa aplicam?

- Standardizare sau normalizare pentru a pregati datele pentru modele ML.
- Verificarea altor variabile care ar putea introduce outlieri.
- Analiza suplimentara pentru a confirma lipsa valorilor extreme in alte contexte.



-Confuziile cele mai frecvente apar intre categoriile "Low" si "High", cu 62 de cazuri "Low" clasificate gresit ca "High" si 71 de cazuri "High" clasificate gresit ca "Low". Aceste valori indica o suprapunere moderata intre aceste doua clase.

- In schimb, confuziile cu categoria "Medium" sunt mai rare (ex. 213 si 211), sugerand ca modelul distinge mai bine "Medium" de celelalte clase.

- Eroarea cea mai semnificativa in termeni absoluti este intre "Medium" si "Low" (3641 corect clasificate ca "Medium", dar cu 113 si 121 erori spre "Low" si "High").