**Predicția performanțelor jucătorilor de baschet: O analiză de regresie liniară**

Frunze Daniel1

1Universitatea Tehnică a Moldovei, Facultatea Calculatoare, Informatică și Microelectronică,

Departamentul Informatica și Ingineria Sistemelor

# ABSTRACT

Această lucrare științifică utilizează limbajul de programare R pentru o analiză aprofundată a unui set de date din NBA, explorând statisticile jucătorilor și indicatorii de performanță. Studiul utilizează diverse tehnici de vizualizare a datelor, analize statistice și modelare predictivă pentru a obține informații despre atributele jucătorilor și impactul acestora asupra rezultatelor jocurilor.

Analiza începe cu preprocesarea datelor, inclusiv tratarea valorilor lipsă și conversia variabilelor categorice. Sunt create vizualizări, cum ar fi histograme și diagrame de bare, pentru a înțelege pozițiile jucătorilor și indicatorii de performanță medie. Studiul investighează relațiile dintre vârsta jucătorului și indicatorii cheie de performanță, cum ar fi numărul total de puncte, procentajul de goluri de pe teren și pasele de gol, folosind diagrame de dispersie interactive.

În plus, lucrarea explorează caracteristicile statistice ale punctelor fantastice, oferind o imagine de ansamblu cuprinzătoare a contribuțiilor jucătorilor. Matricele de corelație și hărțile termice sunt utilizate pentru a identifica relațiile dintre caracteristicile numerice.

Strategiile de detectare și eliminare a valorilor aberante sunt utilizate pentru a îmbunătăți calitatea setului de date, urmate de selectarea caracteristicilor pentru a se concentra pe aspectele relevante. Un model de regresie liniară este antrenat pentru a prezice numărul total de puncte, iar studiul evaluează în mod sistematic performanța modelului folosind diferite dimensiuni de testare și stări aleatorii.

Cercetarea culminează cu o comparație a punctelor totale preconizate și reale, cu diagrame interactive de dispersie, histograme și diagrame reziduale. Lucrarea se încheie cu o reprezentare vizuală a liniei prezise vs. cea reală, oferind o analiză cuprinzătoare a acurateței modelului de regresie liniară.

Această explorare științifică contribuie la înțelegerea factorilor care influențează performanța jucătorilor din NBA, utilizând tehnici statistice avansate și instrumente de vizualizare pentru o investigație amănunțită.

# INTRODUCERE

În era contemporană a analizei avansate în sport, explorarea datelor din domeniul baschetului devine tot mai crucială pentru înțelegerea subtilităților jocului și pentru luarea deciziilor informate. Cu accesul tot mai extins la statistici detaliate, seturile de date din NBA oferă o fereastră fascinantă către lumea performanțelor sportive la cel mai înalt nivel.

Setul nostru de date, intitulat "NBA Players stats(2023 season)", reprezintă o compilație exhaustivă a statisticilor relevante pentru jucătorii din NBA. În centrul atenției se află variabile precum punctele marcate, aruncările reușite și eșuate, minutele jucate, precum și alți indicatori esențiali pentru evaluarea contribuției individuale la succesul echipei.

Prin intermediul acestui studiu, ne propunem să analizăm și să extragem insight-uri semnificative din datele colectate din NBA. Obiectivul nostru principal constă în identificarea modelelor de performanță care pot influența rezultatele meciurilor și succesul general al unei echipe. Vom explora relațiile dintre diferitele variabile și vom utiliza tehnici avansate de analiză pentru a dezvălui aspecte precum eficiența în aruncările la coș, impactul jucătorilor în diferitele secțiuni ale meciului și altele.

Acest studiu nu doar că vizează înțelegerea contextului actual al competiției NBA, ci și oferă un cadru solid pentru proiectarea modelelor predictive. Prin examinarea datelor din trecut, ne propunem să dezvoltăm un model eficient care să anticipeze performanța jucătorilor și echipei în viitor, având potențialul de a sprijini deciziile strategice ale antrenorilor și managementului.

Prin acest demers, contribuim la evoluția analizei sportive într-un domeniu în continuă expansiune. Această investigație nu este doar o privire în culisele lumii baschetului profesionist, ci și o încercare de a aduce o contribuție valoroasă în optimizarea proceselor de luare a deciziilor și îmbunătățirea performanțelor echipelor din NBA.

**MATERIALE ȘI METODE**

***Setul de date***

Studiul nostru a utilizat setul de date "NBA Players stats(2023 season)" disponibil pe platforma Kaggle[1]. Acesta conține diverse statistici ale jucătorilor din NBA cum ar fi puncte acumulate, pase de gol, blocuri, aruncări făcute etc. Aceste date au fost colectate dealungu sezonului 2022-2023. În cursul fiecărui joc fiecare mișcare a jucătorilor este înscrisă intr-un borderou, astfel la finele jocului spectatorii pot vizualiza aceste date și respectiv ele pot fi folosite pentru unele analize.

***Preprocesarea datelor***

Înainte de analiză, am efectuat o preprocesare a datelor. Au fost gestionate valorile lipsă prin intermediul funcției “mutate” din pachetul “dplyr” valorile lipsă din coloana POS (poziția jucătorului) se înlocuiesc cu “SG”. Au fost detectate și eliminate valorilor aberante: Valorile aberante sunt identificate pe baza condițiilor legate de procentajele de goluri de câmp, aruncări de trei puncte și aruncări libere. Rândurile care conțin valori aberante sunt eliminate din setul de date. Conversia tipului de date: Coloana "POS" este convertită într-o variabilă factorială.

***Analiza exploratorie***

Analiza exploratorie a fost realizată pentru a evidenția relațiile dintre variabile și pentru a identifica tendințe semnificative. Cu ajutorul pachetelor plotly, tidyverse, am generat vizualizări informative, cum ar fi bar-plot care corelația dintre diferite variabile

***Modelarea liniară***

Setul de date a fost divizat într-un set de antrenare și unul de testare folosind funcția ` createDataPartition` din pachetul caret. Am construit apoi am dezvoltat un model liniar. Modelul au fost dezvoltat folosind funcția `*lm*` pentru regresie liniară.

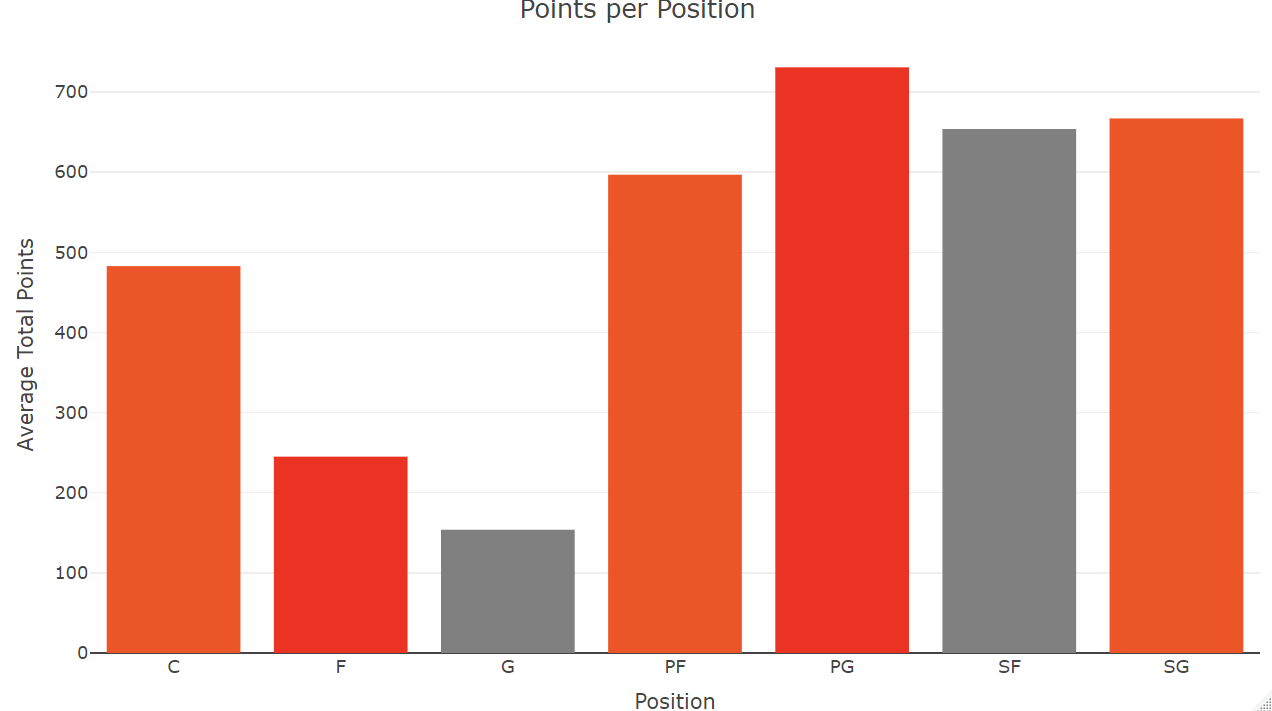
***Evaluarea performanței modelelor***

Evaluarea cu ajutorul R-squared (R^2): R-squared este calculat pentru a măsura cât de bine modelul se potrivește datelor. Acesta furnizează o valoare între 0 și 1, unde 1 indică o potrivire perfectă. Calculul erorilor absolute medii (MAE) și pătratice medii (MSE): MAE și MSE sunt două metrici comune pentru evaluarea performanței modelelor de regresie. Acestea măsoară diferența dintre valorile prezise și cele reale. Vizualizarea rezultatelor cu ajutorul diagramei de dispersie și a histogramei s-au creat diagrame de dispersie și histograme pentru a vizualiza modul în care predicțiile modelului se compară cu valorile reale.

# REZULTATE

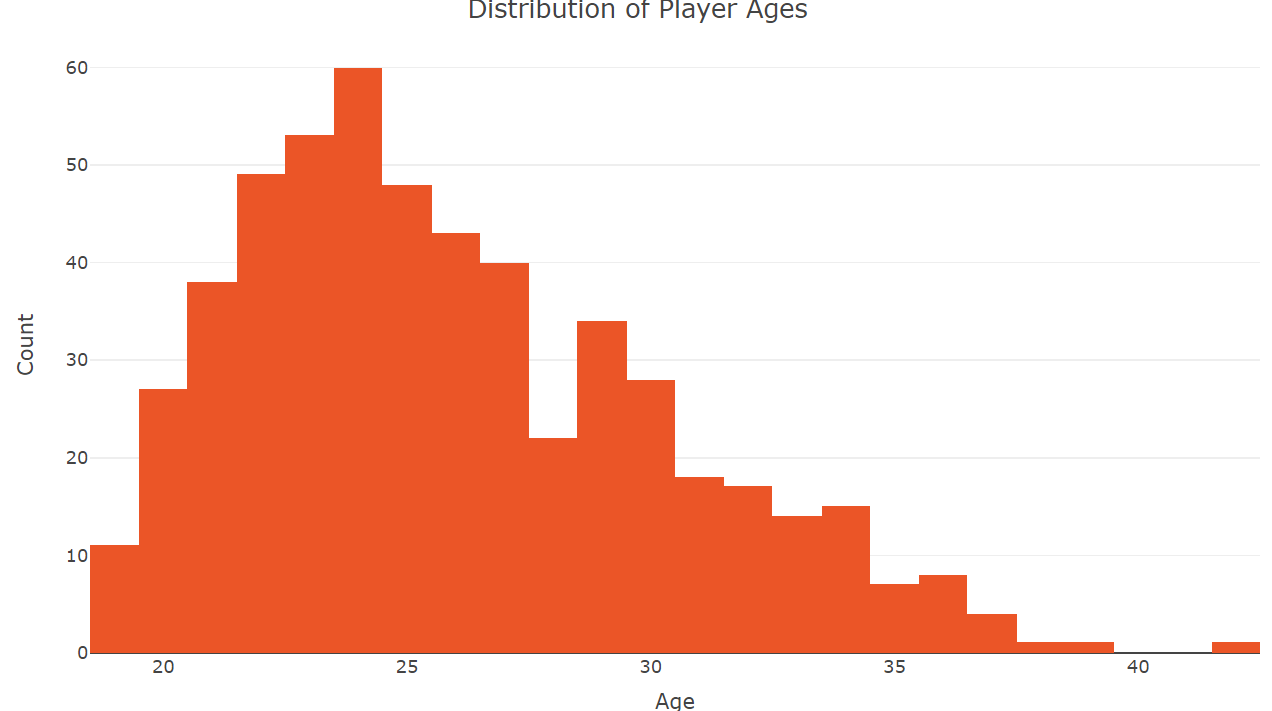
***Analiza distibuțiilor variabilelor***

În incercarea de a întelege relația dintre performnță și alte varibele , am analizat setul de date. Una dintre presupuneri este că jucătorii ce joacă pe poziția PG (point guard) acumulează cele mai multe puncte de oarece este cel mai orientat spre atacă. Pentru a ilustra acest factor am pregătit o figură informativă care ne indică corelația dintre aceste variabele (Figura 1):



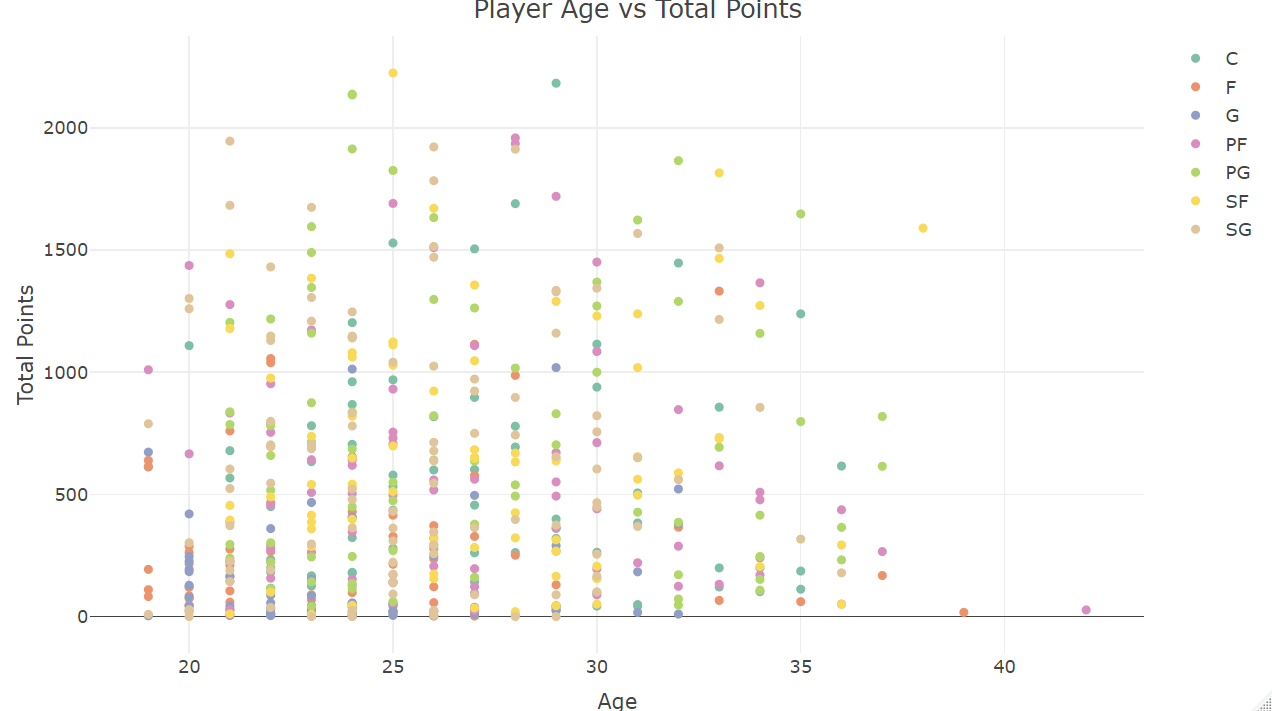
**FIgura 1** Medie de puncte acumulate per poziție

Plotul final este un bar-plot (Figura 1) pentru a evidenția corelația dintre puncte accumulate și poziție. Din observații putem vedea că jucătorii cu poziția PG acumulează în mediu cel mai multe puncta.



**FIgura 2** Distribuția jucătorilor după vârstă

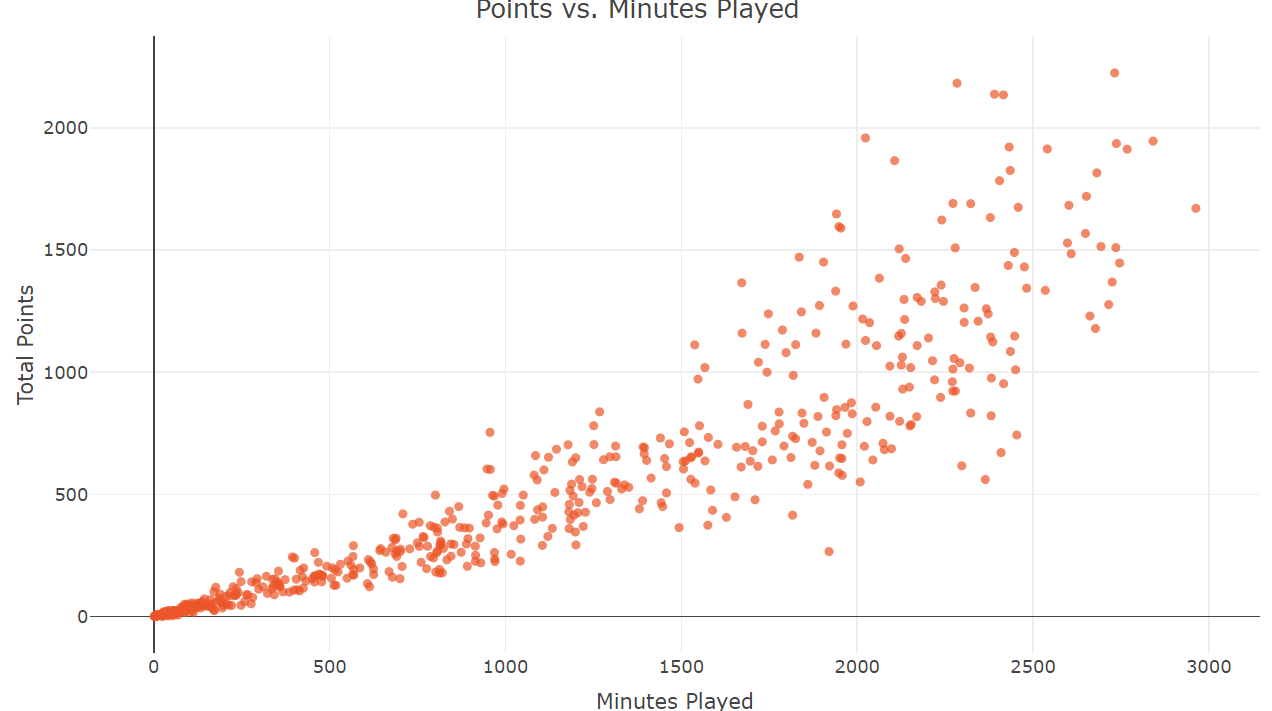
În (Figura 2) putem observa că cei mai mulți jucători sunt între vărsta de 23-27 de ani, ceea ce ne indică că aceasta este o vărsta în care performanțtele sportive sunt la cel mai înalt nivel.



**FIgura 3** Puncte acumulate după vârstă

(Figura 3) demostrează afirmația că cea mai bună performanță este între vârsteșe de 23-27 de ani. Analizând această diagramă se observă că jucătorii între aceste vârste acumulează cele mai multe puncte.

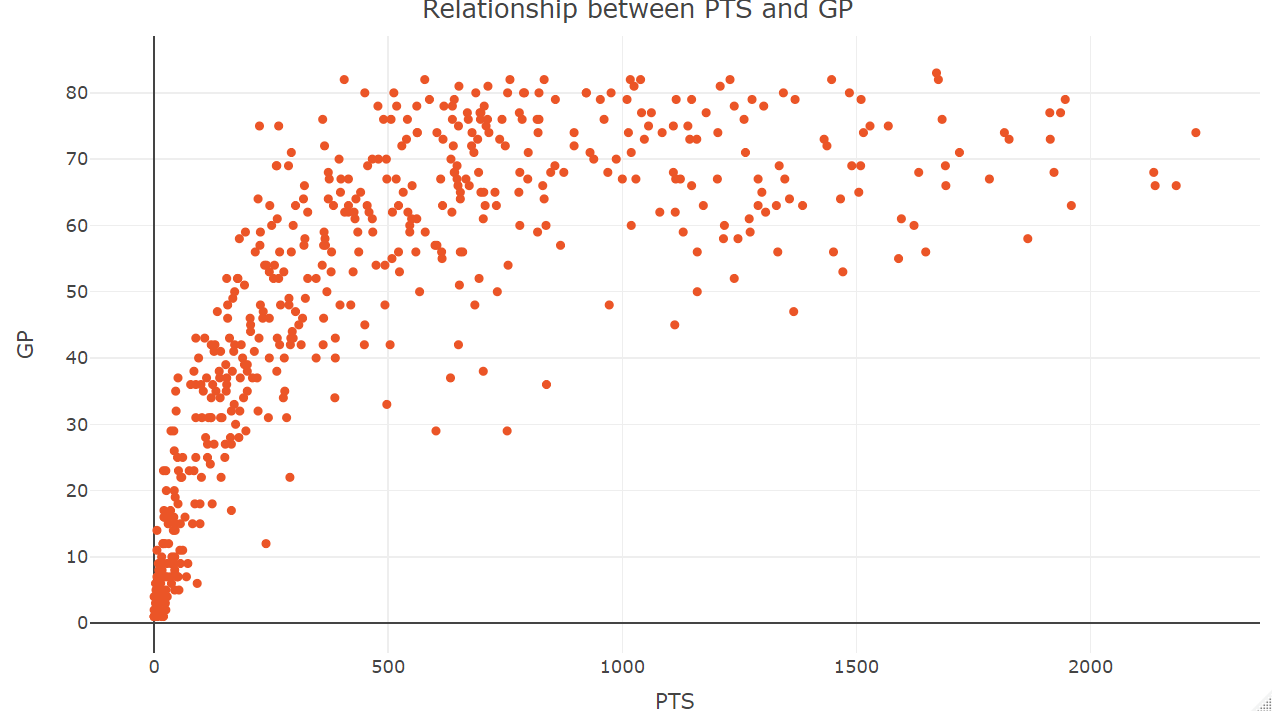
O altă presupunere este că cu cât timp mai mult jucătorii petrec pe teren cu atât mai mult ei vor înscrie puncte



**FIgura 4** Puncte per timp jucat

În (Figura 4) se observă o corelație pozitivă între punctele marcate și minutele jucate. Acest lucru înseamnă că jucătorii care joacă mai multe minute tind să înscrie mai multe puncte, în medie. Cu toate acestea, corelația nu este perfectă, deoarece există mulți jucători care înscriu puncte mari deși joacă mai puține minute și invers. Există câteva valori aberante în aceste date. Aceștia sunt jucători care au marcat semnificativ mai multe sau mai puține puncte decât s-ar fi așteptat pe baza minutelor jucate. Aceste valori aberante se pot datora mai multor factori, cum ar fi faptul că un jucător a avut un meci deosebit de bun sau rău, sau a jucat împotriva unei echipe mult mai slabe sau mai puternice.

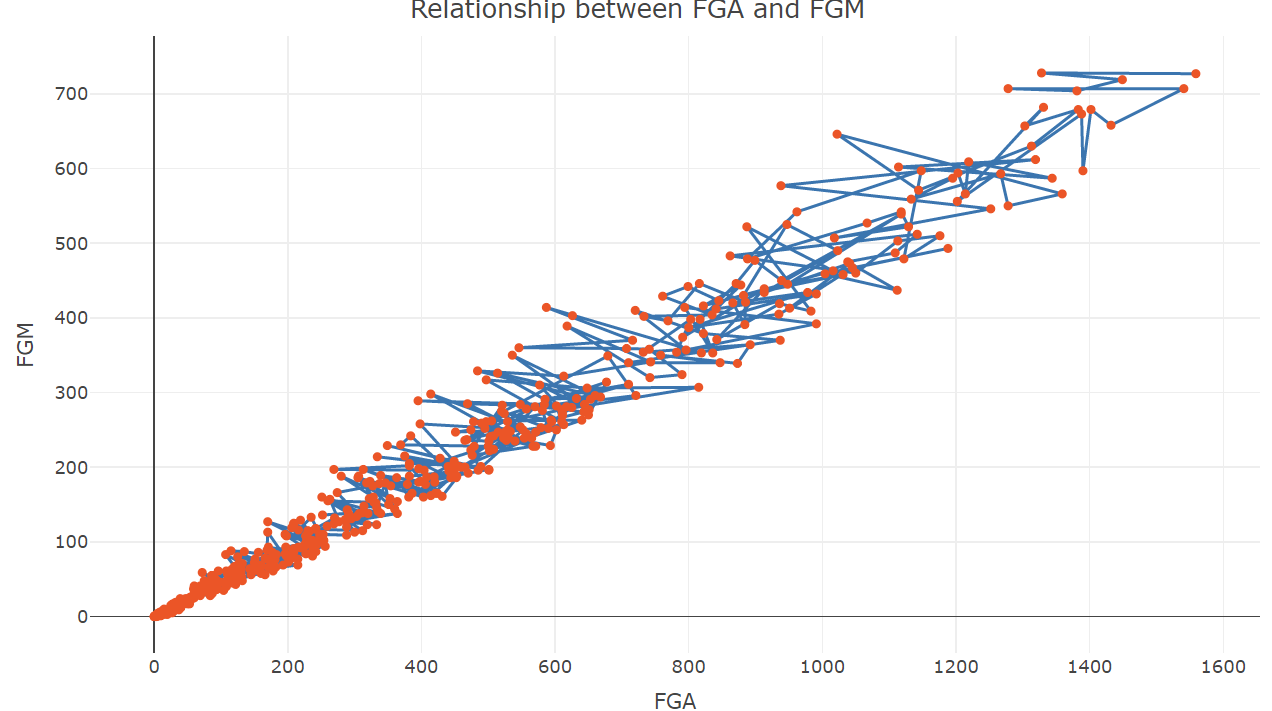
O altă presupunere este că cu cât mai multe jocuri jucătorii joacă per sezon cu atât mai multe puncte vor înscrie



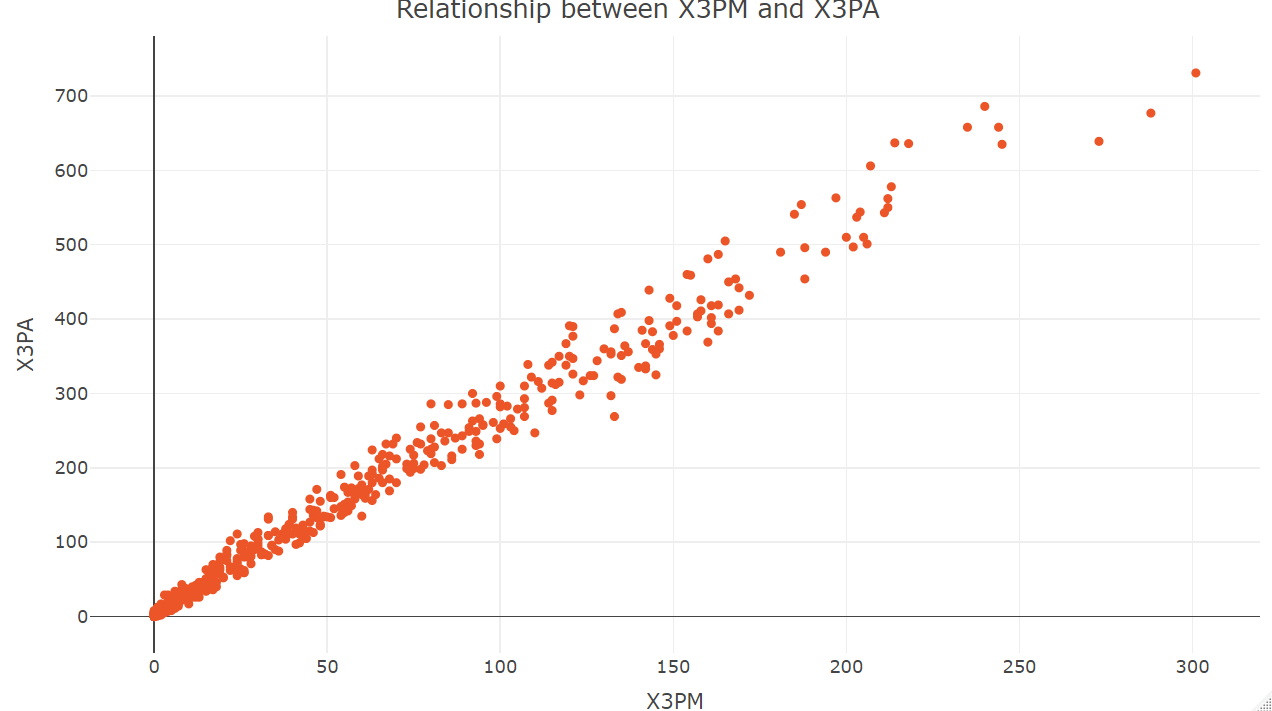
**FIgura 5** Relația dintre puncte și jocuri jucate

Corelație pozitivă: Graficul (Figura 5) arată o corelație pozitivă între PTS și GP. Aceasta înseamnă că jucătorii care joacă mai multe meciuri înscriu, în general, mai multe puncte, în medie. Cu toate acestea, corelația nu este perfectă. Există câteva valori aberante în date, unii jucători marcând semnificativ mai multe sau mai puține puncte decât se așteptau pe baza meciurilor pe care le-au jucat. Aceste valori aberante s-ar putea datora unor factori precum performanțe individuale excepționale, accidentări sau faptul că au jucat în ligi sau niveluri de competiție diferite.

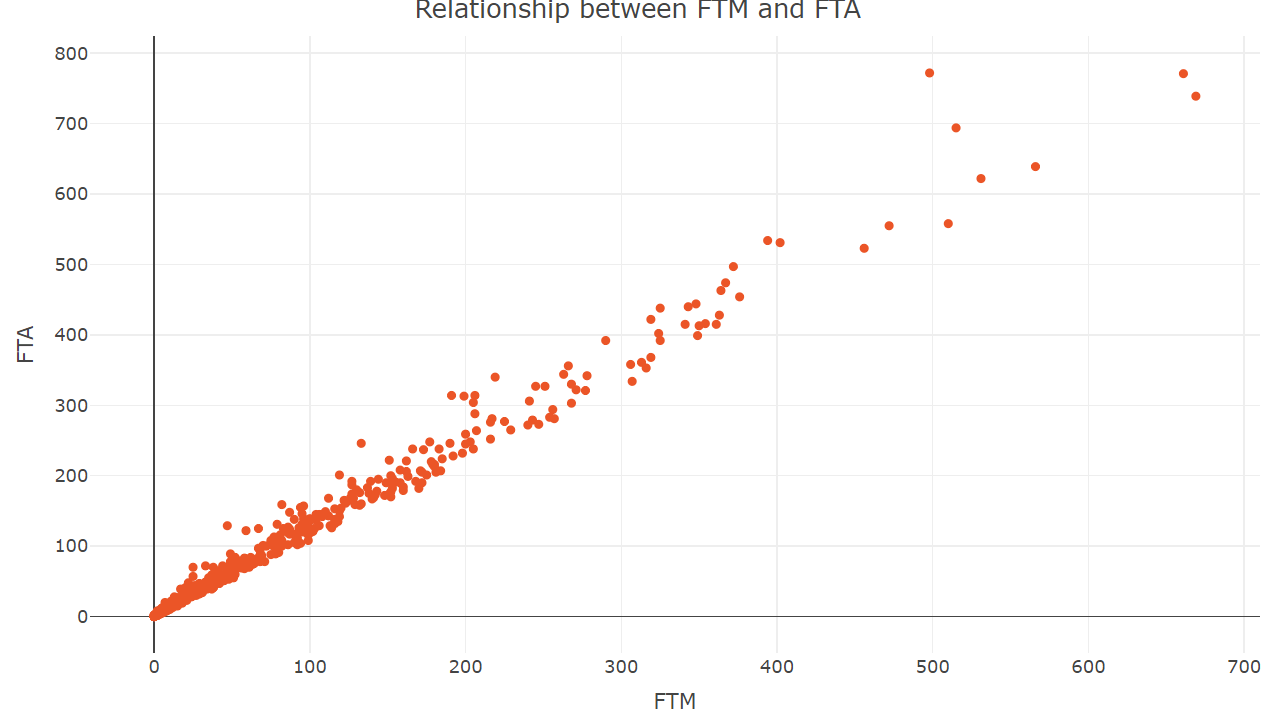
Alte grafice ce descriu corelația dintre variabile:



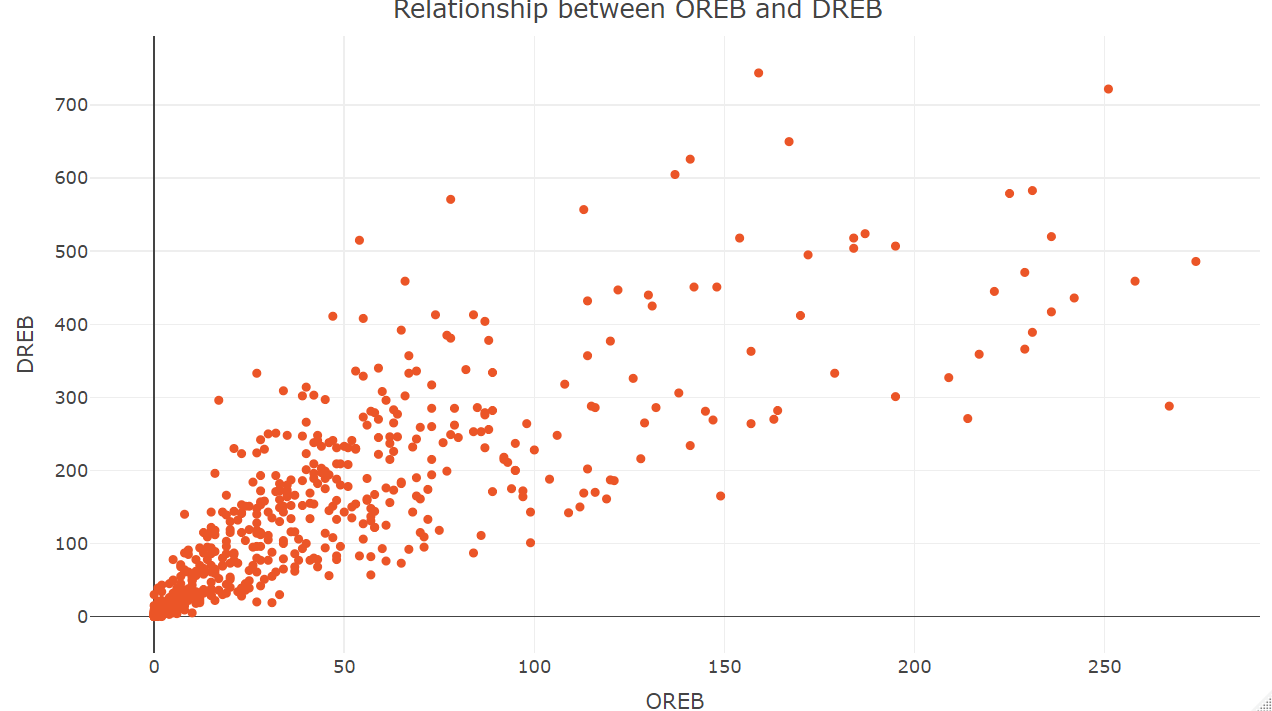
**FIgura 6** Relația înte aruncări făcute și aruncări marcate



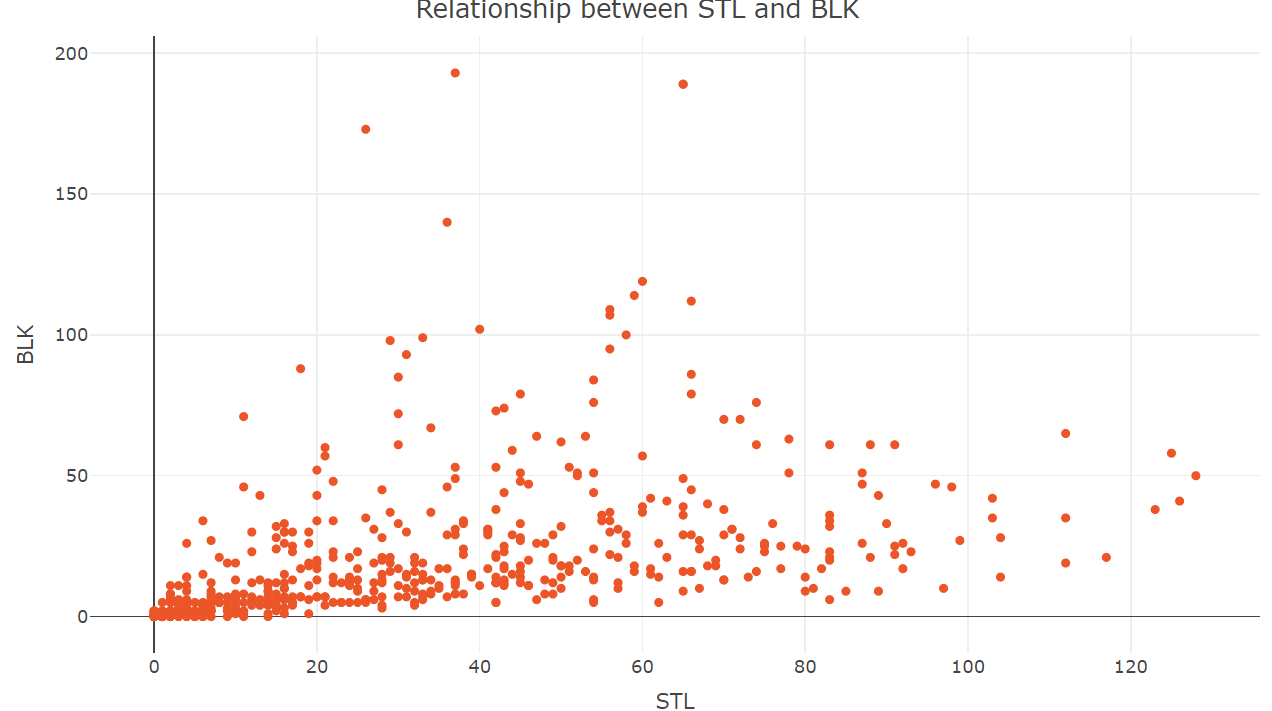
**FIgura 7** Relația între Aruncari de 3 puncte încercate si marcate



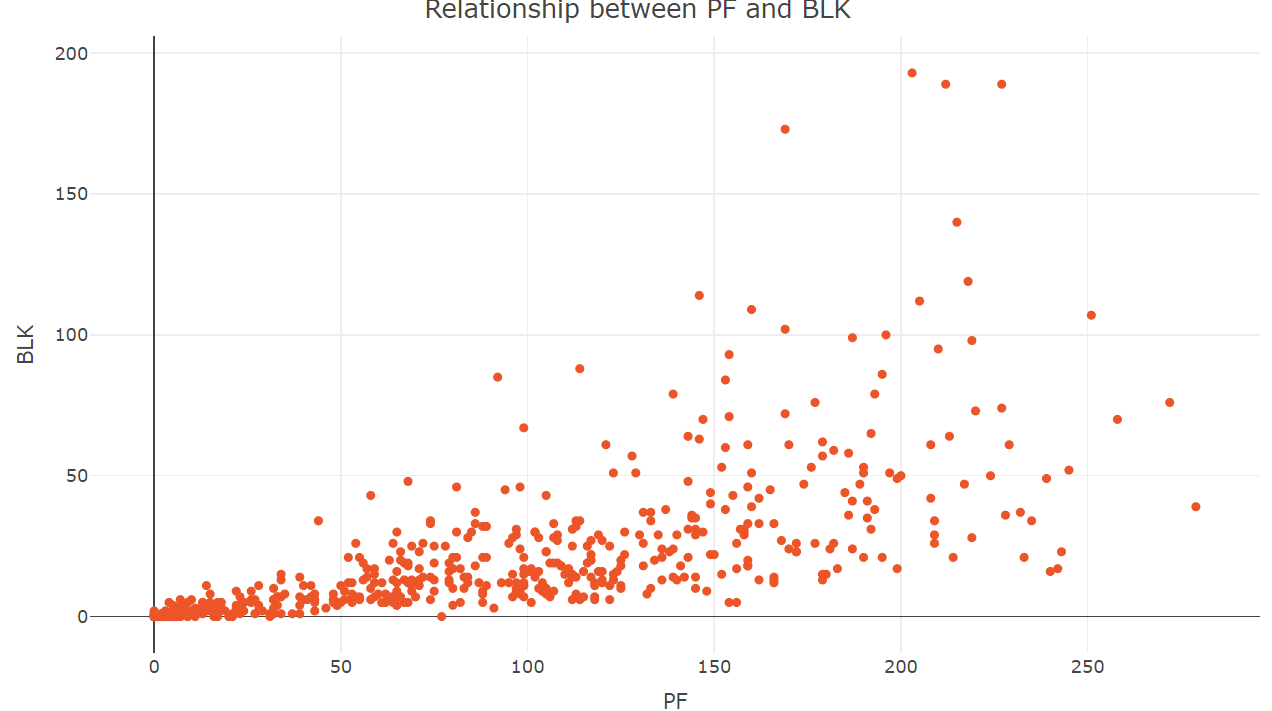
**FIgura 8** Aruncări de la linia de pedeapsă și aruncări marcate



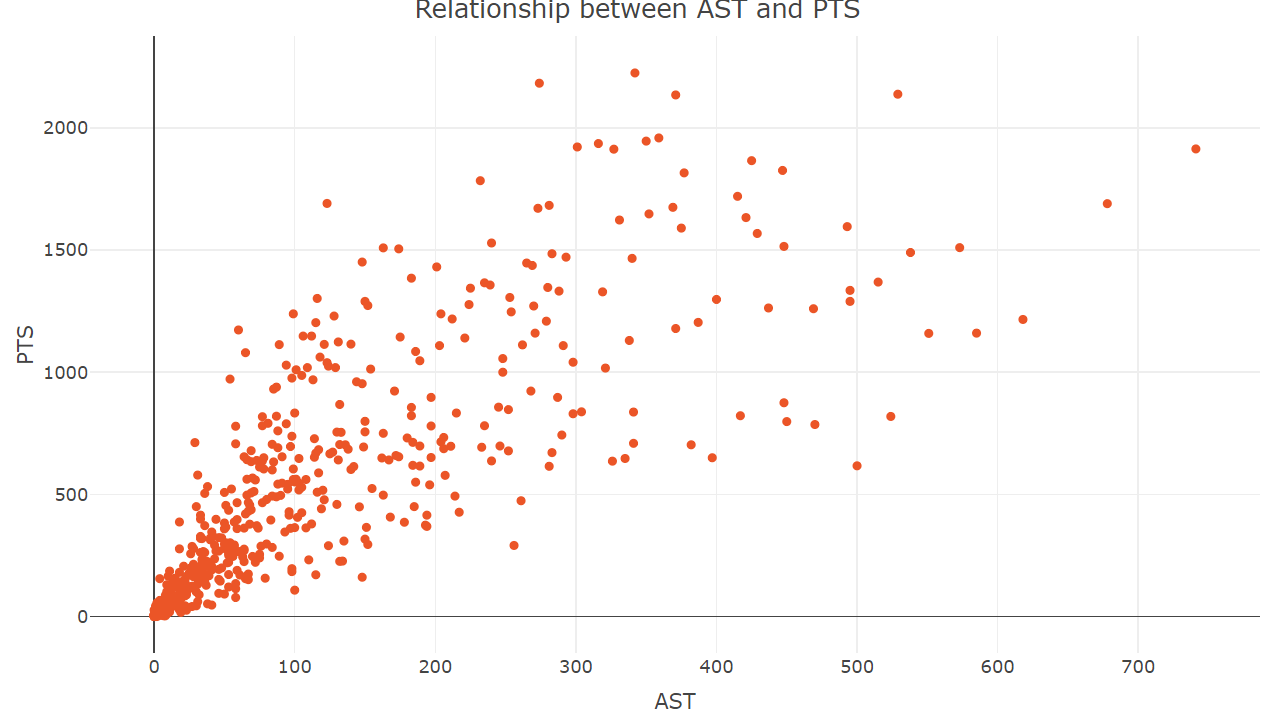
**FIgura 9** Recuperări în atacă și în apărare



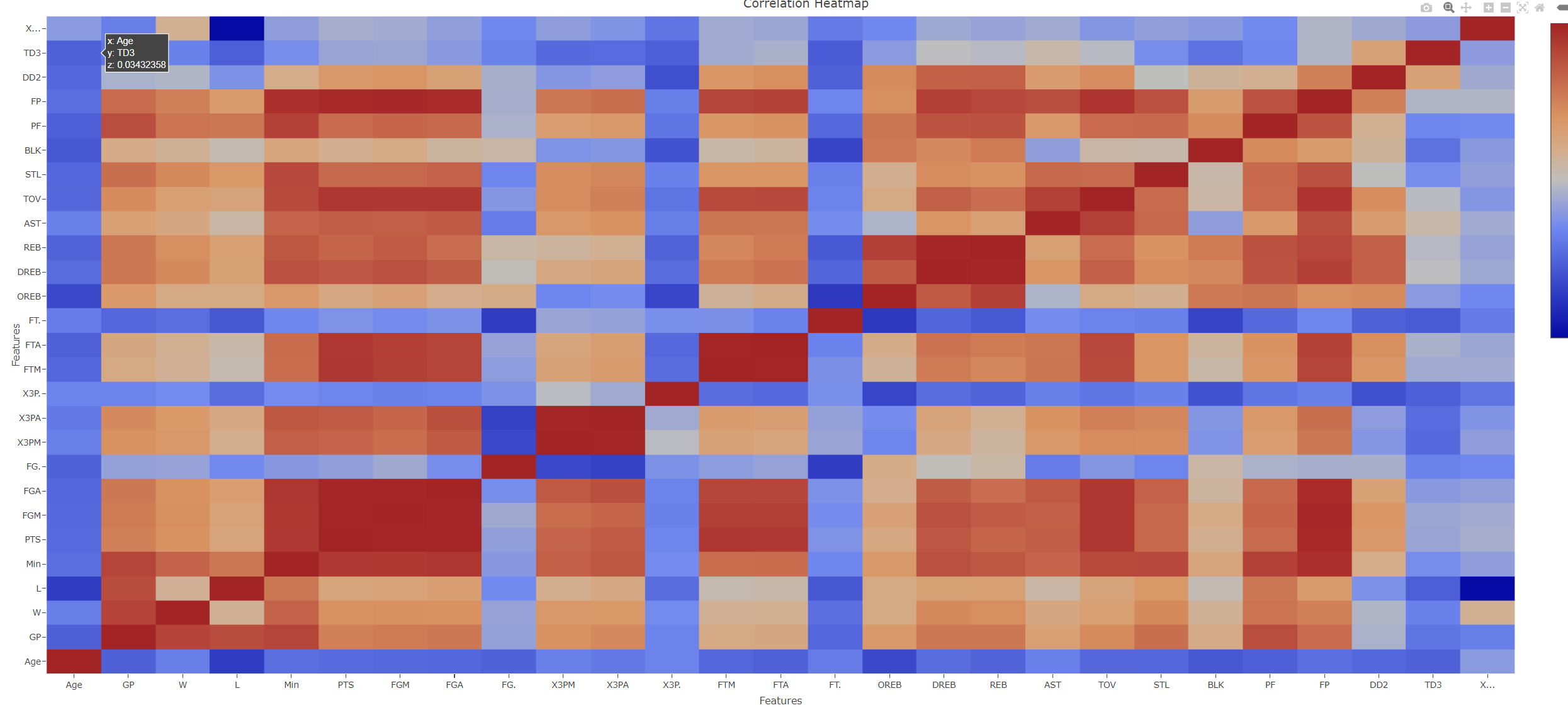
**FIgura 10** Furturi de minge și blocuri



**FIgura 11** Foluri personale și blocuri

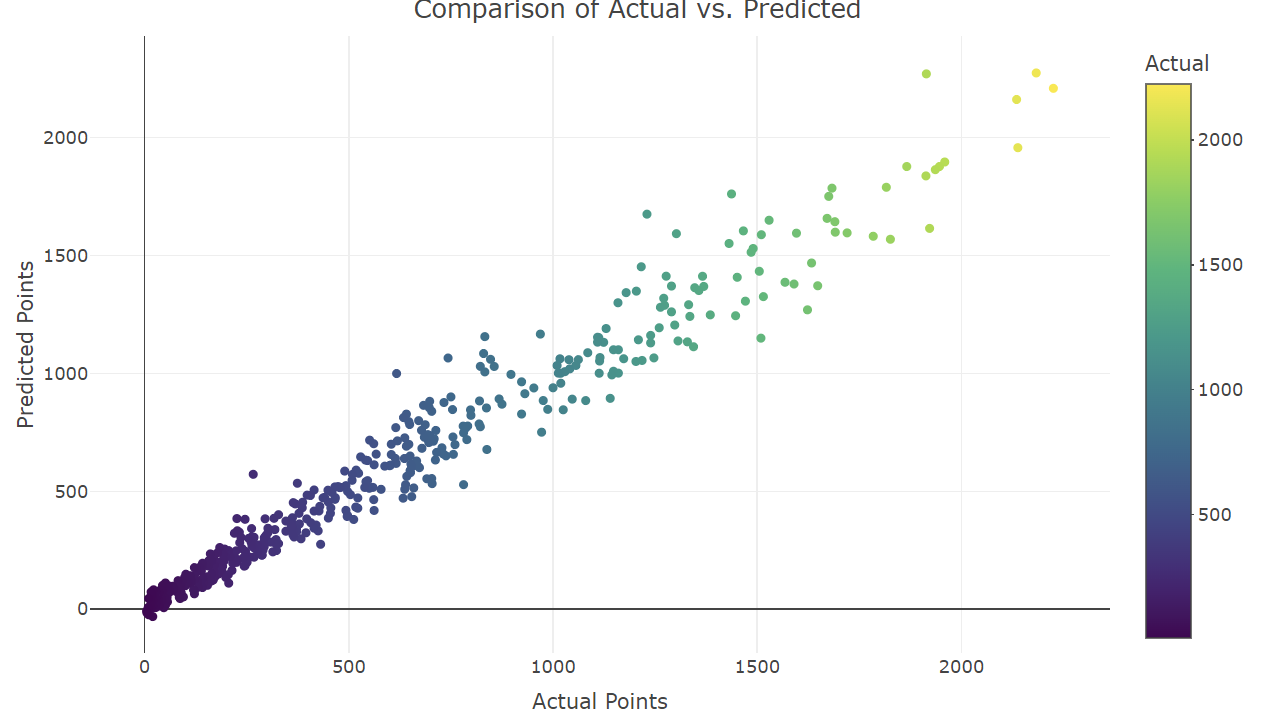


**FIgura 12** Asisturi și Puncte



**FIgura 13** Corelation HeatMap

'Player\_Name','Position', 'Team\_Abbreviation', 'Field\_Goals\_Made', 'Field\_Goals\_Attempted', 'Three\_Point\_FG\_Made' 'Three\_Point\_FG\_Attempted', 'Three\_Point\_FG\_Percentage', 'NBA\_Fantasy\_Points', 'Double\_Doubles','Free\_Throws\_Attempted', 'Triple\_Doubles', 'Offensive\_Rebounds', 'Defensive\_Rebounds' Aceste coloane au fost eliminate pe baza unei corelații ridicate observate într-o hartă termică a corelației. Eliminarea lor are ca scop îmbunătățirea potențială a performanțelor analizelor sau modelelor ulterioare care sunt sensibile la multicoliniaritate sau la corelația ridicată între caracteristici.



**FIgura 14** Compararea punctelor actuale cu cele prezise

Punctele de date formează un grup în jurul liniei diagonale, indicând o corelație pozitivă între punctele reale și cele prezise. Pe măsură ce punctele efective cresc, punctele prevăzute cresc, de asemenea, în general, și cele prezise. Răspândirea punctelor de date în jurul liniei sugerează o anumită variabilitate a predicțiilor modelului. Unele puncte se abat mai mult de la linie decât altele. Nu există clustere sau valori aberante evidente. Linia diagonală are o pantă pozitivă, confirmând corelația pozitivă dintre punctele reale și cele prezise. Linia pare să fie relativ abruptă, sugerând că modificările punctelor reale conduc la modificări relativ mari ale punctelor prevăzute.

Interpretare:

Modelul de regresie liniară pare să facă o treabă rezonabil de bună în ceea ce privește predicția punctelor, dar nu este perfect. Există o anumită variabilitate în predicții și nu toate punctele se încadrează exact pe linie. Panta pozitivă și panta abruptă a liniei indică faptul că modelul tinde să supraestimeze punctele atunci când punctele reale sunt scăzute și să subestimeze punctele atunci când punctele reale sunt ridicate. Acest lucru ar merita să fie investigat în continuare.

Performanța modelului:

R-squared Score: 0.9750894

R pătrat (R²) este o măsură statistică care reprezintă proporția de variație a variabilei dependente care este explicată de variabilele independente într-un model de regresie. Este cunoscut și sub numele de coeficient de determinare.

În contextul regresiei liniare, R pătrat măsoară gradul de adecvare a modelului. Valoarea lui R pătrat este cuprinsă între 0 și 1, unde: 0 indică faptul că modelul nu explică nicio variabilitate a datelor de răspuns în jurul mediei sale. 1 indică faptul că modelul explică întreaga variabilitate a datelor de răspuns în jurul mediei sale.

Cu alte cuvinte, R pătrat ne spune cât de bine explică variabilele independente variabilitatea variabilei dependente. O valoare R pătrat mai mare sugerează o mai bună potrivire a modelului cu datele. Cu toate acestea, R pătrat trebuie interpretat cu prudență, în special în prezența mai multor variabile independente. Adăugarea mai multor variabile la model tinde să crească R pătrat, chiar dacă variabilele suplimentare nu contribuie în mod semnificativ la explicarea variației variabilei dependente. R pătrat ajustat, care ia în considerare numărul de predictori din model, este adesea utilizat în astfel de cazuri pentru o evaluare mai precisă a bonității modelului.

# DISCUȚII

Lucrarea științifică prezentată oferă o explorare detaliată a unui set de date privind jucătorii din NBA, utilizând limbajul de programare R și diverse pachete de analiză a datelor. Studiul cuprinde preprocesarea datelor, vizualizarea și modelarea predictivă pentru a obține informații despre performanța jucătorilor de baschet. Principalele constatări și etape includ:

Preprocesarea datelor:

Tratarea valorilor lipsă în coloana "POS" (poziție), asigurând o reprezentare coerentă.

Identificarea și eliminarea valorilor aberante din anumite coloane.

Selectarea coloanelor numerice pentru analiza de corelație și modelarea predictivă.

Analiza exploratorie a datelor (EDA):

Vizualizarea pozițiilor jucătorilor prin histograme.

Calcularea și vizualizarea mediei punctelor totale pentru fiecare poziție a jucătorului.

Explorarea relațiilor dintre vârstă și indicatorii de performanță (total puncte, procentaj de aruncări pe teren, pase decisive) prin intermediul graficelor de dispersie.

Analiza punctelor fantastice și a mediilor acestora în funcție de poziția jucătorului.

Analiza corelațiilor:

Calcularea unei matrice de corelație pentru caracteristicile numerice, vizualizată sub forma unei hărți termice.

Modelarea predictivă:

Implementarea unui model de regresie liniară pentru a prezice numărul total de puncte pe baza caracteristicilor selectate.

Evaluarea performanțelor modelului utilizând metrica R pătrat.

Explorarea sistematică a diferitelor dimensiuni de testare și stări aleatorii pentru a optimiza parametrii modelului.

Evaluarea și compararea modelelor:

# BIBLIOGRAFIE

1. Kaggle, platformă online pentru seturi de date, ‘Malnutrition across the globe’ [Accesat:04.12.23],disponibil: https://www.kaggle.com/datasets/amirhosseinmirzaie/nba-players-stats2023-season/data
2. RDocumentation, resursă web pentru documentație privitor la RStudio , ‘RDocumentation’ [Accesat:04.12.23], disponibil: <https://www.rdocumentation.org/>
3. *Malnutrition - World Health Organization (WHO),* resursă web pentru informare despre malnutriție, [Accesat:04.12.23], disponibil: https://www.nba.com/

Compararea punctelor totale reale cu cele prezise prin intermediul diagramelor de dispersie, histogramelor și al unui grafic rezidual.

Utilizarea scorului R-pătrat ca măsură pentru a evalua buna potrivire a modelului de regresie liniară.

În concluzie, studiul combină cu succes statisticile descriptive, vizualizările exploratorii de date și tehnicile de modelare predictivă pentru a analiza și înțelege factorii care influențează performanța jucătorilor din NBA. Modelul de regresie liniară, evaluat cu ajutorul scorului R pătrat, oferă o măsură cantitativă a acurateței sale predictive. Perspectivele obținute în urma acestei lucrări contribuie la o înțelegere mai cuprinzătoare a dinamicii din cadrul setului de date, oferind informații valoroase pentru analiza baschetului și pentru luarea deciziilor.