EDA · תוצאות אמיתיות

# ניתוח שחקני NBA — התוצאות

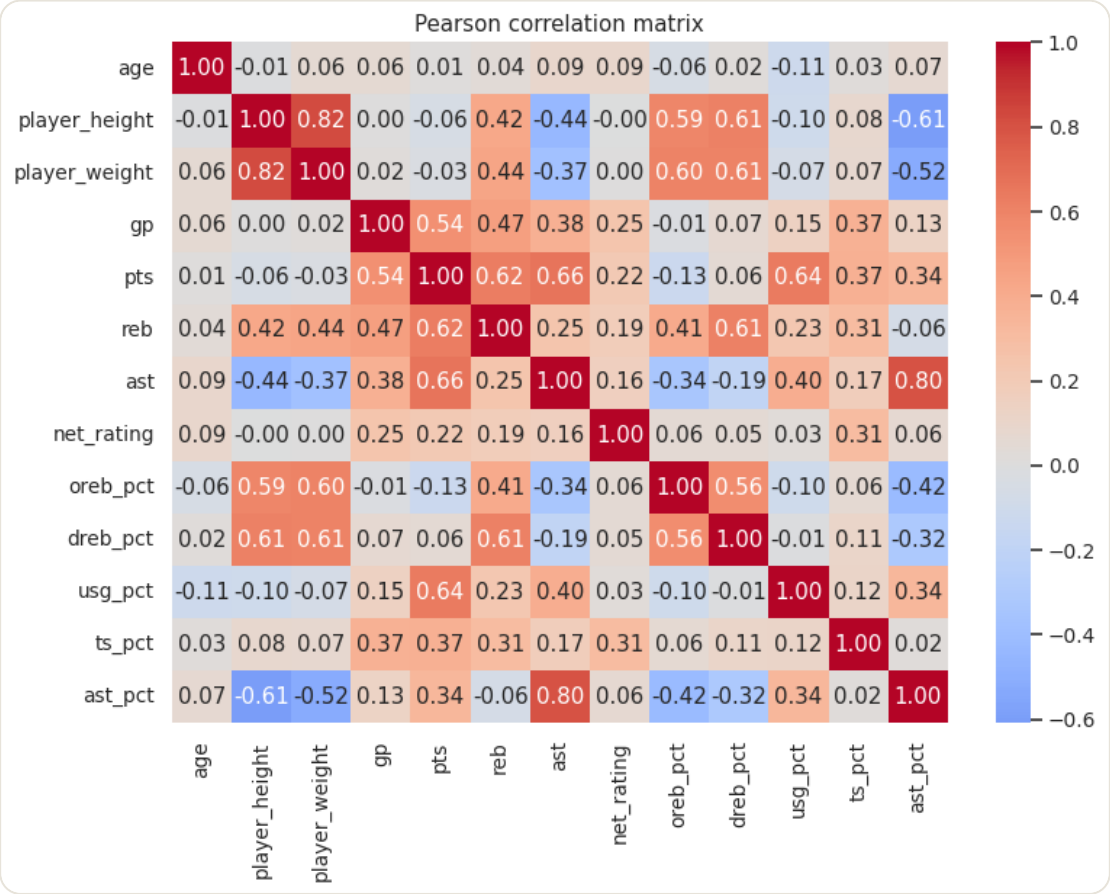
12,844 שורות · 21 עמודות · עונות 1996-97 עד 2022-23. כל הקוד רץ נקי על הדאטה האמיתי.

**איך להשתמש:** קרא את הממצאים ואת "מה שציפינו מול מה שיצא". בכל תיבה ירוקה (**לכתיבה**) יש טיוטת תובנה — הפוך אותה למילים שלך לסעיף 7.

## סקירה: איכות הנתונים

- **0** כפילויות מלאות. יש רק **4** זוגות (שחקן, עונה) כפולים — שחקנים שהוחלפו קבוצה באמצע עונה. הגיוני, לא באג.
- אין ערכים בלתי אפשריים בגיל/גובה/משקל/משחקים — הדאטה נקי מאוד.
- **18.4%** מהשחקנים הם "Undrafted" — זה placeholder בעל משמעות (לא נבחרו בדראפט), לא חוסר אקראי.
- עמודת college חסרה ל-~14% — בעיקר לשחקנים בינלאומיים. חוסר **מובנה ומלמד**, לא אקראי.
- country: **83.5%** מהשחקנים מארה"ב.

## מטריצת קורלציה

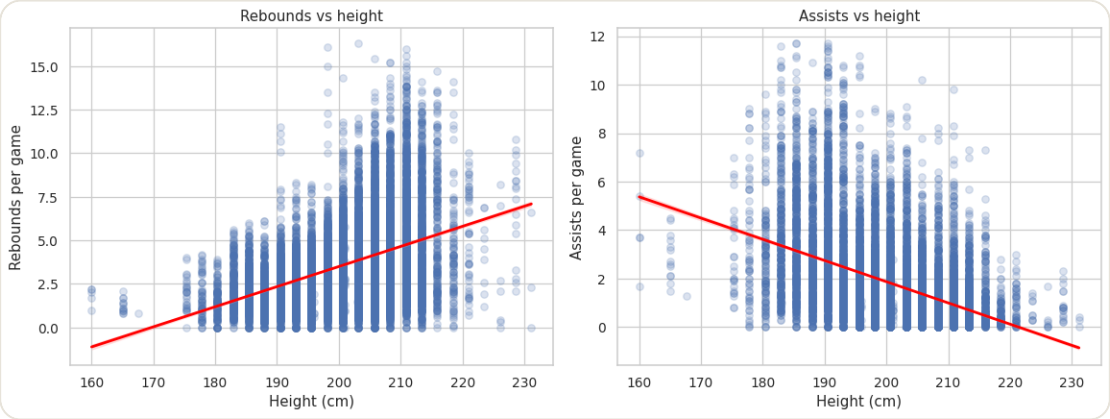



- נקודות קשורות חזק לאסיסטים (**0.66**), ריבאונדים (**0.62**) ומספר משחקים (**0.54**).
- גובה ומשקל כמעט חופפים (**0.82**) — צפוי.
- **מפתיע:** net_rating (תרומה לקבוצה) בקושי קשור לסטטיסטיקות האישיות (0.25–0.16) — כלומר מי שקולע הרבה הוא לא בהכרח מי שהקבוצה מנצחת איתו יותר.

> **לכתיבה:** "הסטטיסטיקות האישיות (נקודות/ריבאונדים) חלשות בניבוי net_rating — תרומה אמיתית לניצחון לא נמדדת היטב ע\"י נתוני קופסה."

שאלה 1

## גובה ↔ ריבאונדים ואסיסטים



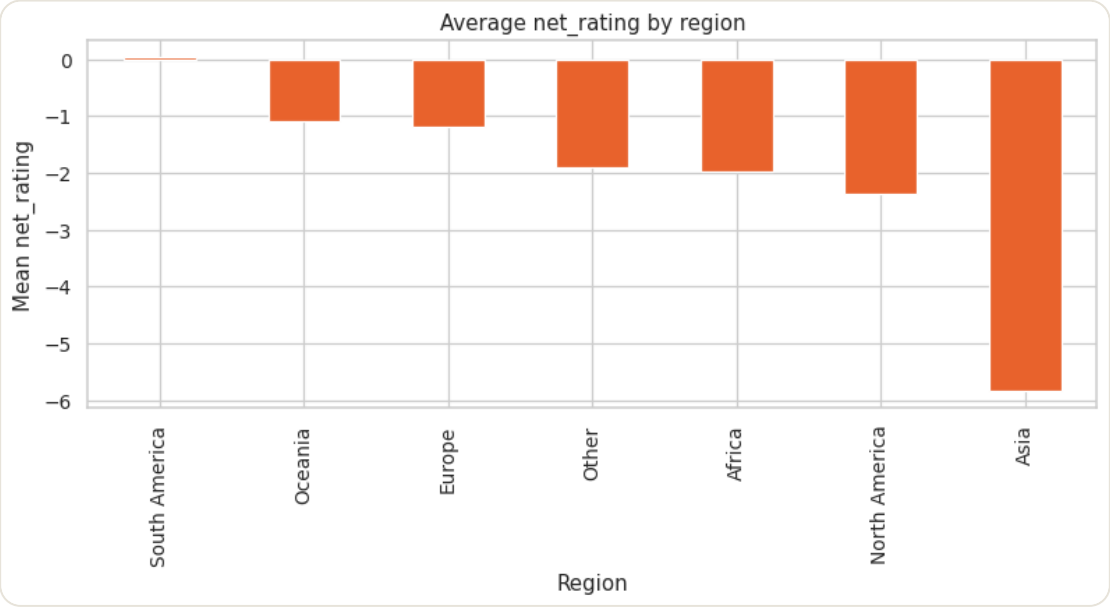

- גובה ↔ ריבאונדים: קורלציה חיובית (**+0.42**) — גבוהים תופסים יותר. ✓ כמו שציפינו
- גובה ↔ אסיסטים: קורלציה שלילית (**−0.45**) — גבוהים מוסרים פחות. ✓
- **החריגים שזיהית — אושרו:** גבוהים עם הרבה אסיסטים = ולאדה דיבאץ', סבוניס, פורזינגיס, דיוויד רובינסון. נמוכים עם הרבה ריבאונדים = סטודמאייר, נייט רובינסון, טרל ברנדון.

**ציפייה מול מציאות:** ניחשת בדיוק — הכלל מתקיים, אבל החריגים אמיתיים וקיימים.

> **לכתיבה:** "גובה מנבא תפקיד (גבוה=ריבאונדר, נמוך=מוסר), אבל החריגים מוכיחים שהתפקיד במגרש חשוב יותר מהגובה — גובה הוא אינדיקטור חלקי בלבד."

שאלה 2

## מדינה/אזור ↔ איכות (net_rating)




המדינות המובילות ב-net_rating ממוצע (מינימום 20 שחקנים):

| מדינה | שחקנים | net_rating ממוצע |
|---|---|---|
| שווייץ | 23 | 4.43 |
| קמרון | 27 | 4.38 |
| סרביה ומונטנגרו | 33 | 2.42 |
| ארגנטינה | 67 | 1.07 |
| ארה"ב | ~10,700 | נמוך יחסית |

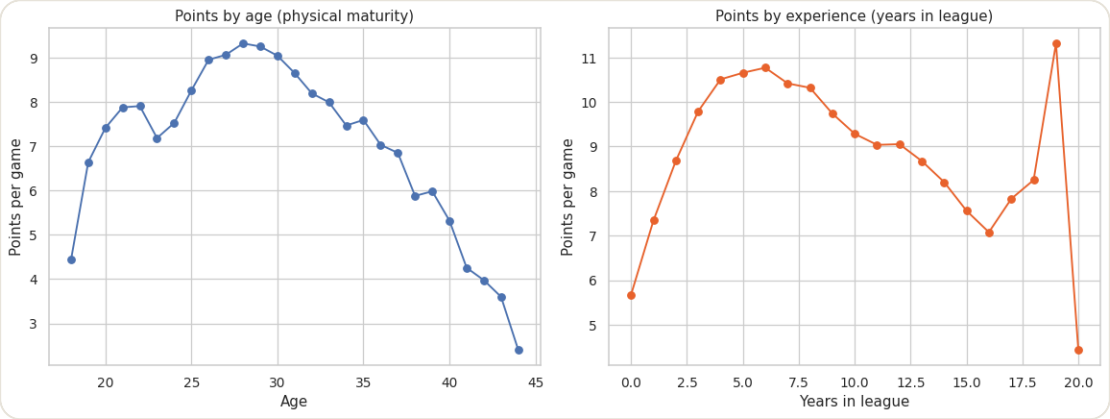
- ארה"ב שולטת ב**כמות** (85% מהשחקנים) — אבל **לא** באיכות הממוצעת.
- מדינות קטנות (שווייץ, קמרון) בראש — עם מעט מאוד שחקנים.

**זו בדיוק ההטיה שדיברנו עליה — הטיית סלקציה:** ממדינה בלי תרבות כדורסל מגיעים ל-NBA רק המעטים ה**יוצאי דופן**, אז הממוצע שלהם מנופח כלפי מעלה. מארה"ב מגיע גם הבינוני.

**לכתיבה (הטיה — שווה זהב):** "דירוג מדינות לפי איכות ממוצעת מטעה: מדינות עם מעט שחקנים מדורגות גבוה בגלל הטיית סלקציה, לא בגלל רמת כדורסל גבוהה יותר. הכמות המוחלטת (ארה\"ב) מספרת סיפור אחר מהממוצע."

שאלה 3

## גיל וותק ↔ נקודות




- שיא הנקודות בגיל **28**. ✓ ניחשת 27–28!
- לפי ותק: השיא סביב **4–6** שנים בליגה, ואז ירידה.
- **2,358** שחקנים (Undrafted) בלי ותק מחושב — **חוסר מובנה**, מתחבר ישירות לשלב 3.
- הקפיצות בקצה הימני (19–20 שנות ותק) הן מדגם זעיר (טירדים בודדים) — יש להתייחס בזהירות.

**ציפייה מול מציאות:** גיל השיא בול. שתי העקומות (גיל וותק) עולות ואז יורדות — מה שתומך ברעיון של **תקרה פיזית**.

**לכתיבה:** "שיא הנקודות הוא שילוב: מספיק ותק (4–6 שנים) לבניית מיומנות, אבל גוף עדיין צעיר (שיא בגיל 28). אחרי גיל 30 יש ירידה למרות ניסיון מצטבר — כלומר יש תקרה פיזית שהניסיון לא מתגבר עליה."

## ממצא בונוס: בדיקת השערה

- גובה ממוצע: שחקני ארה"ב **199.6** ס"מ · בינלאומיים **205.5** ס"מ.
- ההפרש מובהק סטטיסטית ($t=-28.2$, $p<0.001$).
- **למה?** שוב סלקציה — היסטורית ה-NBA "ייבא" בעיקר סנטרים גבוהים

מחו"ל, בזמן שהגארדים הנמוכים היו מקומיים.

כל המספרים והגרפים הופקו מהרצה אמיתית של המחברת על all_seasons.csv. ביום שני נריץ את המחברת אצלך (Run All, ~2 דקות) כדי שהפלטים ייכנסו לקובץ ההגשה, נחדד את התובנות במילים שלך, ונגיש ב-Git.