פרוייקט ב Big Data

הפרוייקט שלנו עוסק בקבוצות כדורסל בליגת ה NBA ואנו בודקים את הקשר (אם קיים כזה) בין הבדלי הרמות הפנימיים של קבוצה בין השחקנים לבין הישגי הקבוצה.

שאלת המחקר:

האם יש קשר בין פערי רמות של שחקנים בקבוצת NBA לבין הדירוג שלה?

או במילים אחרות : איזה קבוצות ב NBA מגיעות להישגים טובים יותר, קבוצות מאוזנות מבחינת רמת השחקנים בקבוצה או קבוצות שיש בהן הבדלי רמות גדולים בין השחקנים ביחס לקבוצות האחרות?

את המידע לקחנו מהאתר https://www.basketball-reference.com

השתמשנו בשני סוגי מידע מהאתר:

1. טבלה שמכילה נותנים על כל השחקנים בעונה מסויימת כמו:

שם שחקן, שם הקבוצה, גיל, מספר זריקות ל2 ו -3 נקודות, מספר אסיסטים,ריבאונדים, חסימות,איבודי כדור וכו'.

דוגמה לטבלה בה השתשמנו - נתוני שחקנים לעונה 2018-2019 :

https://www.basketball-reference.com/leagues/NBA_2019_per_game.html

2. טבלה שמכילה את שם הקבוצה והדירוג שלה בעונה מסויימת.

דוגמה לטבלה בה השתשמנו - נתונים לעונה 2018-2019 (Expanded Standings) :

https://www.basketball-reference.com/leagues/NBA_2019_standings.html

השתמשנו במידע של 10 העונות האחרונות 2009-2019

אסטרטגיה:

- נחשב עבור כל שחקן מדד יעילות בשם plus_minus (נוסחה עבור ממד זה מופרטת בדו"ח הסיכום)
- נחשב את ממוצע ה plus_minus עבור כל קבוצה
- נחשב את סטיית התקן של מדד ה plus_minus בכל קבוצה – מה שיביא לנו את הבדלי הרמות בין השחקנים באותה הקבוצה
- נקשר בין הנתונים שחישבנו עד כה לבין טבלת הדירוגים על מנת למצוא קשר בין הבדלי הרמות (סטיית התקן) לבין דירוג הקבוצה.

כמו כן אנו בודקים האם ישנם נתונים מסויימים שמאפיינים הישגים של קבוצה.

לדוגמה האם קבוצות עם דירוג גבוה נוטות לזרוק הרבה או מעט מ 3 נקודות,

או בעלות מספר גבוה/נמוך של אסיסטים – מה שיכול להעיד על סגנון משחק קבוצתי או אישי.

לשם כך נחשב את הממצוע הקבוצתי של חלק מהנתונים(זריקות ל-2 נקודות, 3 נקודות, ריבאונדים וכו')

ונבדוק האם יש התאמה בין חלק מהנתונים לדירוג באמצאות PCA - Variables factor map