# פרויקט ביג דאטה

# רקע:

כדורסל הוא משחק כדור בו מתחרות זו בזו שתי קבוצות, בנות חמישה שחקנים כל אחת, הצוברות נקודות באמצעות השחלת כדור דרך חישוק הסל של הקבוצה היריבה (מתוך ויקיפדיה).

הכדורסל הוא אחד מענפי הספורט הפופולריים בעולם (מקום שני לפי totalsportek) ואחת הסיבות שהוא מאוד פופולרי הינה שקשה מאוד לחזות את זהות המנצחת.

עם התפתחות הפופולריות של המשחק, גדל גם הביקוש למידע מדיוק על הפועלות שקורות במגרש וניתוח של מידע זה. בעוד בעבר, התמקדו אך ורק בנקודות, עבירות ובאסיסטים (מסירה שיוצרת סל) עם הזמן הוסיפו מדדים נוספים, עקיפים יותר, שיכולים להשפיע על הניצחון וההפסד במשחק (ריבאונדים, הגנה והתקפה, חטיפות, איבודי כדור וחסימות).

בגלל פופולריות המשחק, ישנו פוטנציאל גדול למי שמצליח לדעת מי תנצח את המשחק עוד לפני שהוא מתרחש. בין אם זה חברות הימורים, כדי שיוכלו להרוויח כסף. מנהלי קבוצות שרוצים להביא את קבוצתם לאליפות. מאמנים, שיוכלו להחליט על איזה מדדים ופונקציות של השחקנים כדאי להשקיע את זמן האימונים. שחקנים, בשביל לדעת מה לשפר וכיצד להעלות את הסיכויים לקבל חוזה עתידי. וכמובן גם אוהדים של המשחק שרוצים להרגיש נעלים על חבריהם בכך שהם צדקו בזהות המנצחת.

בשנים האחרונות, ישנה התפתחות בתחום הניתוח הסטטיסטי של המדדים השונים הנמדדים במשחק, רבים מנסים למצוא מהו הפרט הסטטיסטי אשר יעזור להם לנבא את תוצאת המשחק.

אנו ננסה לבדוק האם ישנה אפשרות לדעת את זהות המנצחת על ידי מבט בסטטיסטיקה בלבד, במטרה שבעתיד נוכל להשתמש בידע הזה בשביל להבין מה דרוש מקבוצה לעשות על מנת לנצח את המשחק הבא.

# שאלת מחקר – איזה סטטיסטיקות מנצחות משחק כדורסל.

בעבודה זו נחקור את האפשרות שישנם סטטיסטיקות מסוימות שאם ננתח אותם נוכל לדעת מה תהיה תוצאת המשחק.

# תיאור המידע:

את המידע בו עשינו שימוש אספנו מהאתר: [www.stats.nba.com](http://www.stats.nba.com) - אתר המכיל סטטיסטיקות על אינספור מדדים.  
המידע שנאסף מכיל נתונים על 4 עונות (2013-2017), הנתונים כוללים:

|  |  |
| --- | --- |
| מדד | פירוט |
| TEAM\_NAME | שם הקבוצה. |
| FG3M | מספר שלשות שנקלעו. |
| FG3A | מספר שלשות שנזרקו. |
| FG3\_PCT | אחוז קליעת שלשות. |
| FTM | זריקות חופשיות שנקלעו. |
| FTA | זריקות חופשיות שנזרקו. |
| FT\_PCT | אחוז קליעת זריקות חופשיות. |
| OREB | ריבאונד מתקיף. |
| DREB | ריבאונד מגן. |
| REB | מספר ריבאונדים. |
| AST | אסיסטים. |
| STL | גניבת כדור. |
| BLK | בלוקים. |
| TO | איבוד כדור. |
| PF | פאול אישי. |
| PTS | נקודות |
| FG2M | מספר 2 נקודות שנקלעו. |
| FG2A | מספר 2 נקודות שנזרקו. |
| FG2\_PCT | אחוז קליעת 2 נקודות. |
| OUTCOME | תוצאה. |
| יתר העמודות הן אחוז המדדים לעומת הקבוצה השנייה. | |

יש לציין שחלק מהנתונים לא היו זמינים באופן ישיר ונדרשנו לחשב אותם בעזרת הנתונים הקיימים.

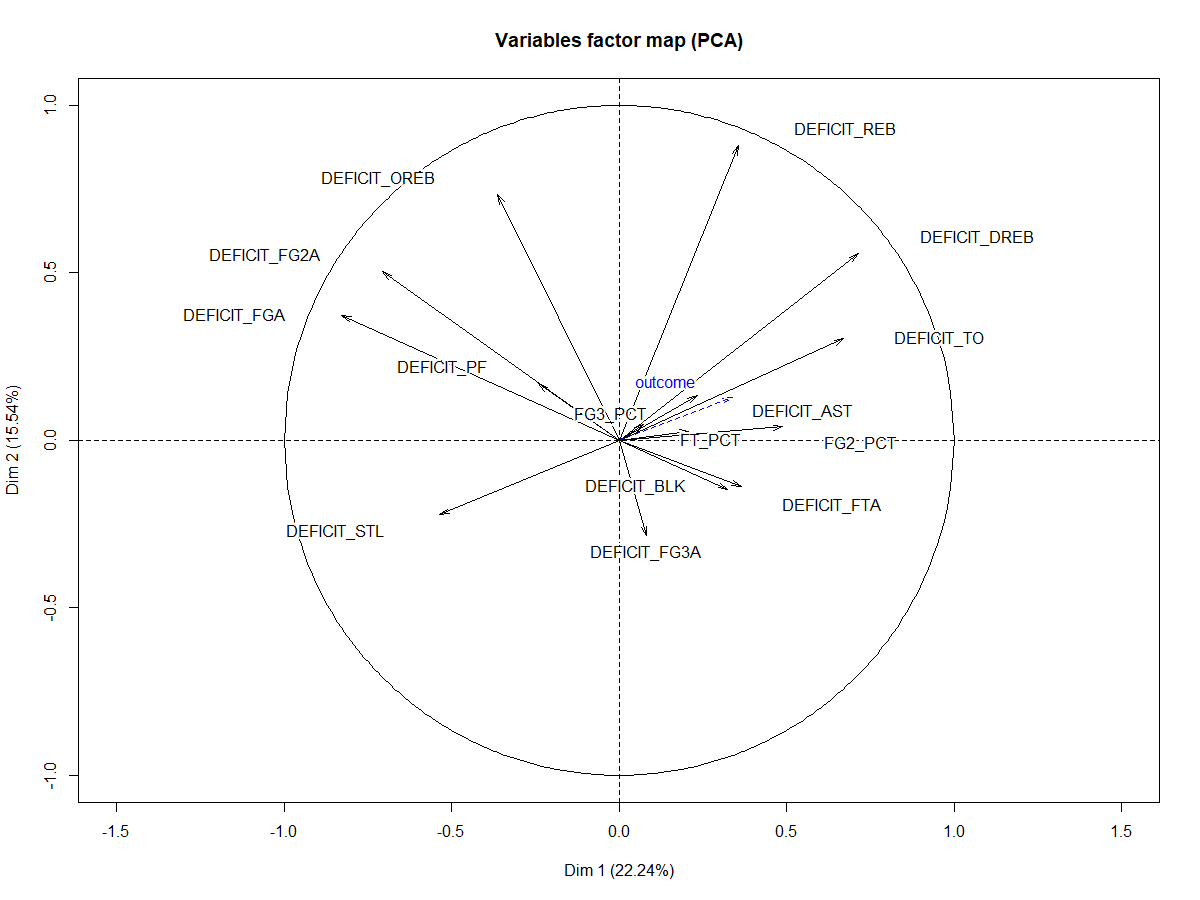
# תיאור כללי של הצעדים בביצוע הפרויקט:

## איסוף מידע:

בשלב זה נעזרנו ב API המאפשר שליפת סטטיסטיקות מאתר ה- NBA. הכלי נכתב בשפת Python ולכן הקוד לאיסוף המידע שלנו נכתב גם הוא ב Python.   
התוכנית שכתבנו מקבלת כארגומנט טווח של שנים ואוספת את המידע הרלוונטי לעונות בשנים האלה.  
כאשר האיסוף מסתיים התוכנית מסדרת את המידע שנאסף בקובץ csv שיתאים לעבודה עם R.

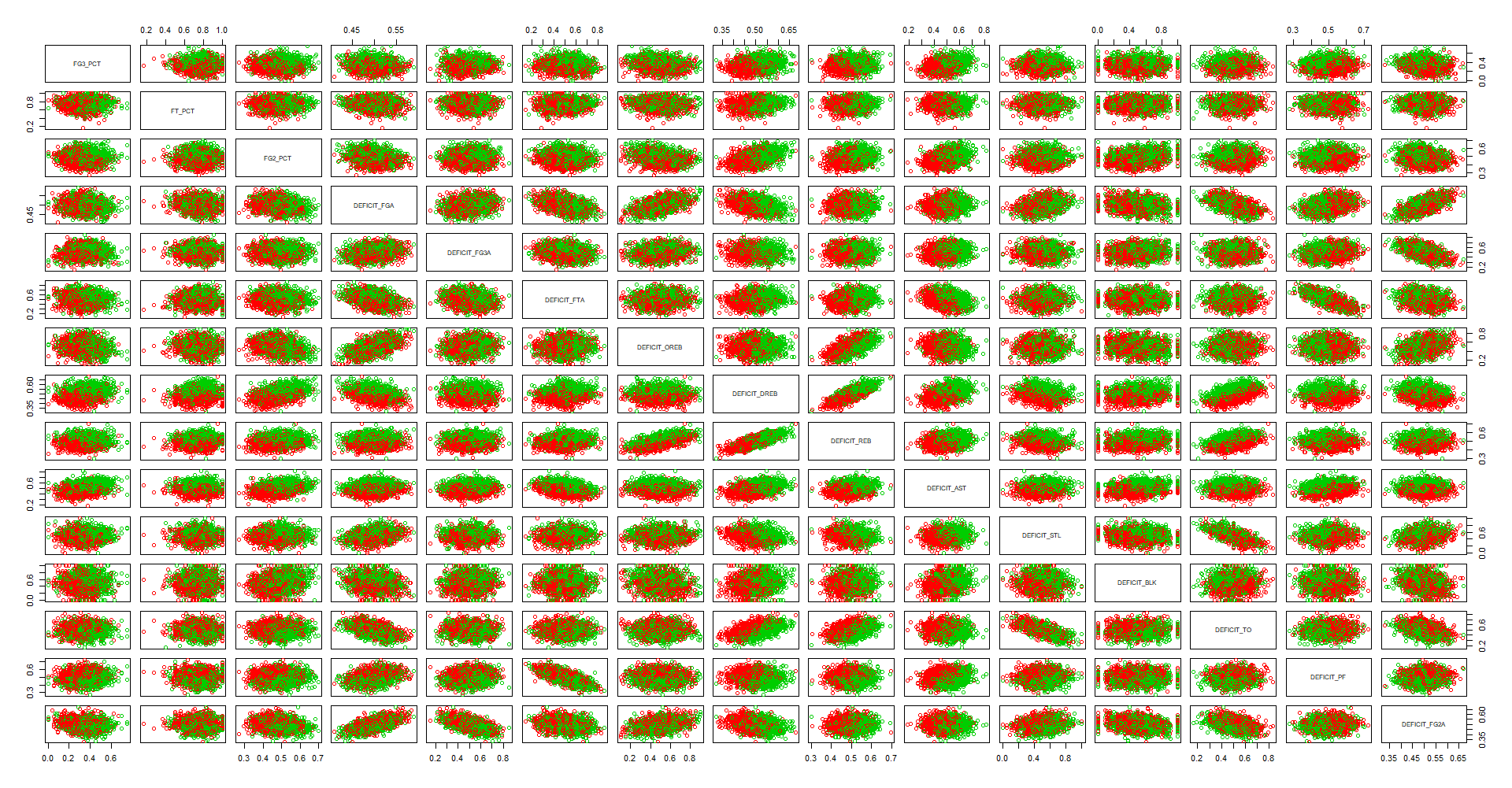
## בניית טבלאות לעבודה עם R:

## זיקוק מדדים והתמקדות במדדים הכי חזקים:

על מנת לבחון אילו מדדים הם הממדים החזקים ביותר שכדאי להשתמש בהם, השתמשנו בשיטת ה PCA (Principle Component Analysis).

|  |  |  |
| --- | --- | --- |
|  | correlation | p.value |
| DEFICIT\_DREB | 0.71458012 | 0.000000e+00 |
| DEFICIT\_TO | 0.67083400 | 1.017135e-288 |
| FG2\_PCT | 0.48760305 | 3.094784e-132 |
| DEFICIT\_BLK | 0.36506558 | 1.414876e-70 |
| DEFICIT\_REB | 0.35557342 | 8.469140e-67 |
| OUTCOME | 0.33694743 | 9.386679e-60 |
| DEFICIT\_FTA | 0.32433632 | 2.981061e-55 |
| DEFICIT\_AST | 0.23253686 | 1.702077e-28 |
| FG3\_PCT | 0.20748163 | 6.766563e-23 |
| DEFICIT\_FG3A | 0.08211446 | 1.120738e-04 |
| FT\_PCT | 0.07375866 | 5.231101e-04 |
| DEFICIT\_PF | -0.24143315 | 1.188063e-30 |
| DEFICIT\_OREB | -0.36268766 | 1.285373e-69 |
| DEFICIT\_STL | -0.53628857 | 1.124676e-164 |
| DEFICIT\_FG2A | -0.70768165 | 0.000000e+00 |
| DEFICIT\_FGA | -0.82796346 | 0.000000e+0 |

ניתן לראות כי המדד – DEFICIT\_DREB, הוא המדד עם הכי הרבה קורלציה.  
כעת נרצה לראות הקורלציה של כל המדדים בזוגות:

  
ניתן לראות כי DREB אכן יכול לסייע בסיווג המידע.

כעת בעזרת אלגוריתם למציאת אשכולות Kmeans, מצאנו אילו מדדים מנבאים הכי טוב בזוגות את תוצאת המשחק של הקבוצה.  
כעת מצאנו את קבוצת המדדים אשר ביניהם נמצאים המדדים עם יכולת הניבוי הטובה ביותר:

* DEFICIT\_DREB
* DEFICIT\_AST
* FG3\_PCT

על מנת לזקק ולמצוא את אלו המשפיעים ביותר, השתמשנו שוב ב Kmeans רק שכעת הפעלנו את האלגוריתם על כל הקומבינציות האפשריות של הקבוצה הנבחרת. גילינו שהקבוצה אשר משפיע הכי הרבה על תוצאת המשחק היא שלושת המדדים גם יחד.

## ניבוי:

בשלב זה, השתמשנו באלגוריתם decision tree זאת מפני שאופם פעולת האלגוריתם מתאימה לסוג הניבוי אותו אנו מבקשים לבצע.   
בדומה למדדים שמצאנו קודם, אך עם שינוי קל, האלגוריתם בחר במדדים:

* DEFICIT\_DREB
* DEFICIT\_AST
* DEFICIT\_TO

האלגוריתם הגיע לרמת דיוק של 75% כאשר טבלת

# תיאור תוצאות:

# דיון בתוצאות:

# הקוד שהיה בשימוש: