תרגיל כתה: Linear regression

בתרגיל זה תבדקו את הקשר בין מספר הפעמים בהן חובט במשחק בייסבול מנסה לחבוט בכדור לבין כמות הריצות (נקודות) שהוא משיג .

1. הטעינו את הקובץ batting.csv אותו ניתן להוריד מהמודל למשתנה batting. מסד נתונים זה מכיל סטטיסטיקות של שחקני בייסבול מעונת 2002. אנו נתמקד בשתי העמודות הבאות:
   - AB: קיצור של at bat, מספר הפעמים בהן שחקן "עלה לבסיס" ועמד מול הזורק בניסיון לחבוט
   - R: מספר הריצות (runs) ששחקן השיג. שחקן משיג ריצה לאחר שעבר בכל ארבעת הבסיסים והצליח לחזור לבסיס הבית בלי להפסל.
2. הדפיסו את מספר הריצות כפונקציה של מספר הפעמים שבהן שחקן עולה לבסיס. האם היחס נראה לכם לינארי? חשבו את הקורלציה בין שני המשתנים, האם היא עשויה להתאים ליחס לינארי?
3. מצאו ב-plot את הנקודה שמייצגת את השחקן שלדעתכם הוא היעיל ביותר (כיצד תגדירו אותו?) מיהו אותו שחקן? ניתן לראות את מזהה השחקן בשדה playerID של batting. צרו טבלה של כל השחקנים שבה מופיע היחס בין מספר הריצות לבין מספר הפעמים בהן שחקן חבט בכדור. מצאו את עשרת השחקנים עם היחס הכי גבוה. האם השחקן שבחרתם נמצא ברשימה? האם הוא מדורג ראשון?
4. צרו מודל לינארי עבור היחס בין AB ל-R בעזרת הפונקציה ()ols. שרטטו את קו הרגרסיה וודאו שהתוצאות תואמות.
5. חשבו את $\beta 0$ ו-$\beta 1$ באופן ישיר בעזרת הנוסחאות שניתנו בשקפים והשוו אותם לערכים שחישבתם בעזרת הפונקציה ()ols. וודאו שהערכים זהים. כיצד אתם מפרשים את $\beta 0$ ואת $\beta 1$?
6. הדפיסו את ה-residuals כתלות ב-AB. האם אתם מבחינים בהתנהגות "חשודה" שעשויה להעיד על אי התאמה של המודל הלינארי?
7. בידקו את R^2. איזה אחוז מה-variance מסביר המודל?

הגישו דרך המודל את הקובץ עם הקוד