תרגיל בית 2

Graphical user interface, text, application, email

Description automatically generated

Chart, scatter chart

Description automatically generated

Graphical user interface, text

Description automatically generated

לפי הגרף שיצא ניתן לראות ש – אנשים אשר היו בעלי ערכים גבוהים יותר בכל הערכים יחדיו קוטלגו כ spread גבוהה  
Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

Z-score יודע להתמודד עם outliers מפני שהכל מוזז בהתאם לממוצע וסטיית תקן  
אך מכיוון שכל תכונה מושפעת רק מהממוצע וסטיית התקן שלה, התחום המתקבל לכל תכונה הוא לא בהכרח זהה

לדוגמה – לתכונה A עם mean = 20 ו סטיית תקן של 1 נקבל תחום של   
לעומת מקרה של תכונה B עם mean = 20 וסטיית תקן של 20 נקבל שכל הערכים מקיימים

לפי הנוסחה של z-score , לפיכך הערכים יהיו בין   
ותחום זה שונה מהתחום הקודם ולכן כל תכונה תשפיע על המרחק באופן אחר .

נבחר בתכונה זו במקרה שתחומי הערכים של התכונות זהים על מנת לאפשר טיפול ב outliers .

Min-max scaling : הנוסחה הזו מנרמלת את הערכים של כל תכונה לתחום , אך אינה יודעת להתמודד עם outliers , (מפני שהערכי הקיצון יקבלו את הערכים 1 או 0 דבר הישנה את ההתפלגות ויקשה עלינו לזהות את הoutliers )  
אך מכיוון שהנוסחה מעבירה את כל התכונות לתחום כל התכונות משפיעות באותו אופן על המרחק (לדוגמה ב KNN , קירבה תושפעה באופן שווה מכל התכונות ותיהיה מדוייקת יותר לעומת z-score שם תכונה אחת עלולה ליצור קרבה "מזוייפת" )  
ההיתרון הבולט של שיטה זו היא שחישוב מרחק של דוגמה מדוגמה אחרת נעשה באופן המתייחס באופן שווה לכל התכונות ובכך יכול למנוע טעויות.  
וכשאין outliers ברורים

