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 ב 07\_PCR , 03\_PCR יחדיו ונמוכים ב 10\_PCR קוטלגו כבעלי spread גבוהה  
בנוסף ניתן לראות שהערכים שקוטלגו כ-spread נמוך מרוכזים יחדיו סביב ערכי 03\_PCR הקרובים ל-0 , ערכי 07\_PCR ששליליים וערכי 10\_PCR גדולים  
Graphical user interface, text, application

Description automatically generated

רמת הדיוק אותה קיבלנו היא 

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

Graphical user interface, text

Description automatically generated

לאחר נירמול הערכים באופן הבא :

PCR\_03 בעזרת z-score  
PCR\_07 בעזרת z-score  
PCR\_10 בעזרת z-score

קיבלנו את התוצאה הבאה: Logo

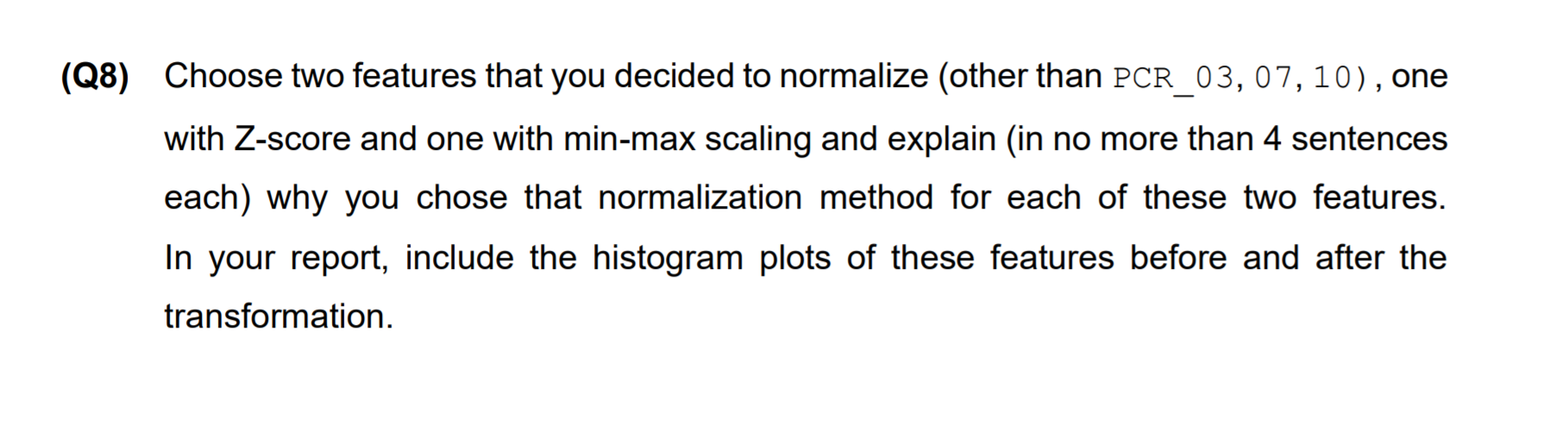
Description automatically generated

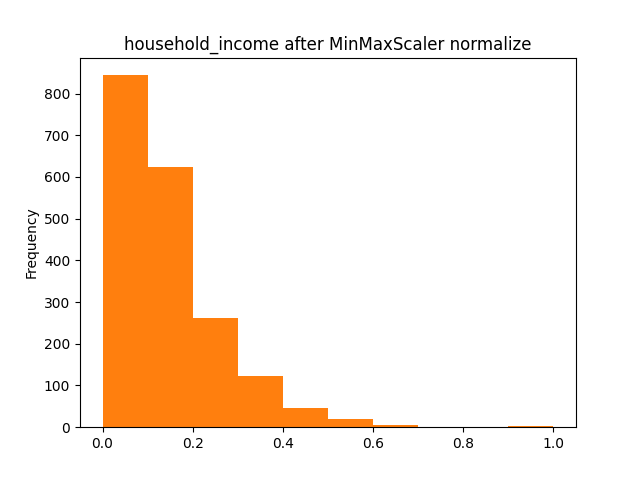
אך כשאעברנו את הנירמול של PCR\_07 ו – PCR\_10 לפי min-max התקבלה רמת הדיוק הבאה   
A picture containing text, device, gauge, meter

Description automatically generated

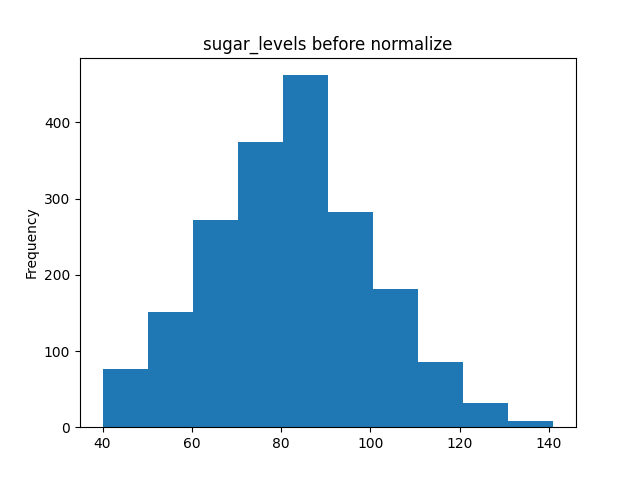
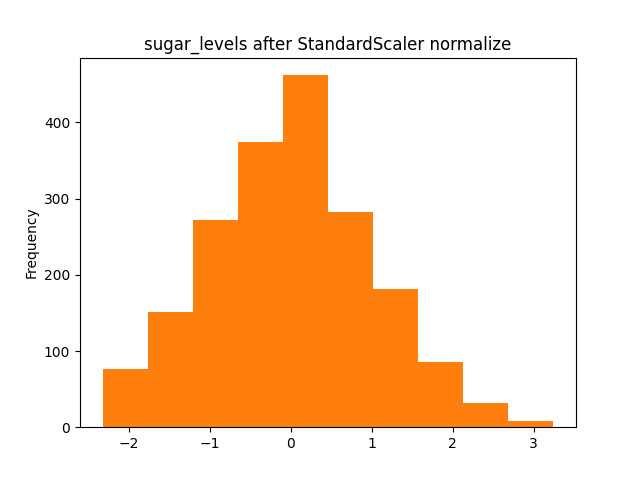
על סט האימון

ההבדל בתוצאות הגיעה בשל הבדיקת מרחק מתוצאות שונות לפי הערכים, מכיוון שהבדל בערך אחד בצורה קיצונית יכל לגרום לקרבה לפי ערך זה למרות שקרבה זו לא באמת קיימת . (כי התוצאה מרחק היא בחד מימד ולא לפי כל תכונה בנפרד).  
ולכן נרמול הערכים, ובכך לתת לכל תכונה השפעה יחסית זהה על המרחק מאפשר להגיע לרמת דיוק גבוהה יותר.



Chart, histogram

Description automatically generatedבחרנו את השדות household\_income ו suger\_levels



כפי ששמנו לב השדה household\_income אינו דומה להתפלגות גוסיאנית, ולכן בחרנו לנרמל אותו לפי minmax ואכן ניתן לראות שההתפלגות נשמרת בצורה טובה

כמו כן השדה suger\_levels דומה להתפלגות גוסינית ולכן בחרנו לנרמל לפי strandScaler ואכן ניתן לראות שההפתלגות נשמרת בצורה טובה.