

236756 – מבוא למערכות לומדות – תרגיל בית רטוב 1

שחר כץ 313574766, סיון יוסבשוילי 318981586 | מאי 2021

2+3.

Type	תיאור	Attribute	
Int	המספר המזהה של הדגימה	ID	.1
String	כתובת מגורים	Address	.2
Int64 (category)	קבוצת הגיל שאליה שייך האדם שממנו נלקחה הדגימה	Age group	.3
Float64	מדד הBMI	BMI	.4
String (category)	סוג הדם	Blood type	.5
Int64 (category)	מספר האנשים שהאדם מדבר איתם ביום	ConversationsPerDay	.6
Object (tuple of floats)	המיקום הגיאוגרפי הנוכחי של האדם בקואורדינטות לפי קווי אורך ורוחב (GPS)	Current Location	.7
datetime64	היום בו בוצעה בדיקת ה-PCR	Date of PCR test	.8
Int64 (category)	רמת משמעת של האדם ביחס לסולם מוגדר	DisciplineScore	.9
Int64 (category)	רמת אושר של אדם ביחס לסולם מוגדר	HappinessScore	.10
Float64	ממוצע ההוצאות הביתיות על מתנות	HouseHoldExpenseOnPresents	.11
Float64	ממוצע ההוצאות הביתיות על משחקים חברתיים	HouseHoldExpenseOnSocialGames	.12
Float64	ממוצע ההוצאות הביתיות על כרטיסי חניה בשנה	HouseHoldExpenseOnParkingTicketPerYear	.13
String	במה עובד	Job	.14
Int64	טיב הביטוח הרפואי לפי סולם מוגדר	MedicalCare	.15
Int64	מספר בני הדודים של האדם	NrCousins	.16
Float64	כמות החלבון בדם ביחס לבדיקת PCR נתונה	PCR_num	.17
Object (list of strings)	תיאור תסמיני המחלה מנקודת מבטו של האדם	Self-declarations of illness form	.18
String (category)	מגדר	Sex	.19
Float64	מספר הדקות הממוצע שהאדם מבצע פעילות חברתית ביום	Social activity per day	.20
Float64	מספר הדקות הממוצע שהאדם נמצא ברשתות חברתיות ביום	Social media per day	.21
Float64	מספר הדקות שהאדם מבצע ספורט ביום	Sports per day	.22
Int64	מספר הצעדים שמבצע האדם ביחס למדד מסוים	Steps per year	.23
Float64	מספר השעות הממוצע שלומד האדם ביום	Studying per day	.24
String (category)	המחלה בה חולה (או לא) האדם	Virus	.25
String (category)	רמת התפשטות המחלה אצל החולה	Spread level	.26
String (category)	רמת הסיכון לבריאות האדם לפי מצב הנוכחי של הנגיף	Risk	.27

נציין כי בחרנו להשתמש ב'Int64' ולא סתם ב-int מכיוון שמדובר בסוג משתנה המסוגל להחזיק ערכי None לעומת הטיפוס הפשוט.

בנוסף נדגיש כי משתנים שציינו עבורם בסוגריים category הינם משתנים שתחום הערכים עבורם מצומצם (עד כ-10-15 ערכים שונים) ומגדירים חלוקה הגיונית של הקבוצה לתתי קבוצות.

6. ביצענו כמה וריאציות על פיצ'רים מסוג מחרוזת כדי להפיק מהם מידע נומרי –

- עמודת CurrentLocation פוצלה ל-2 עמודות – x_loc , y_loc כאשר אם $currentLocation="a,b"$, אז $x_loc=a$, $y_loc=b$. בהמשך נבחן האם קיימת קורלציה בין המיקום הנוכחי לתיוג ואיזה מידע נוסף ניתן להוציא מהעמודות החדשות.
- עמודות risk, spread level הומרו לעמודות בהן הערכים הפכו למייצגים מתוך הסקאלה [1-3], כלומר, low=1, medium=2, high=3. מטרת המרה זו היא לתת משמעות ליחס בין התכונות ($low < medium < high$) כדי לאפשר שימוש בפיצ'ר זה בהמשך כאחד מהצירים בגרף מסוים.
- מתוך העמודה Address הפקנו את שם העיר/מחוז של כל כתובת תקינה והוצאנו לעמודה חדשה בשם Region. הרעיון מאחורי מודיפקציה זאת היא שחשבנו ש-Address הינה עמודה שיהיה לנו קשה להוציא נתונים שימושיים ממנה בעקבות ריבוי ערכים ייחודיים ומידע חסר שימוש. הנחנו שיהיה יותר נוח להסתכל על כל אובייקט לפי אלמנט יותר רחב, כמו עיר מגורין, חיזוק לרעיון זה מצאנו באופן טיפול וניתור מגפת הקורונה בארץ, במהלכה ראינו כי נעשים מיפויים גאוגרפיים של מוקדי חולי ברמת עיר/מחוז אך לא ברמת מספר דירה.

בהקשר זה, ראו טרנספורמציות נוספות שביצענו על המידע בהמשך בשאלה 12.

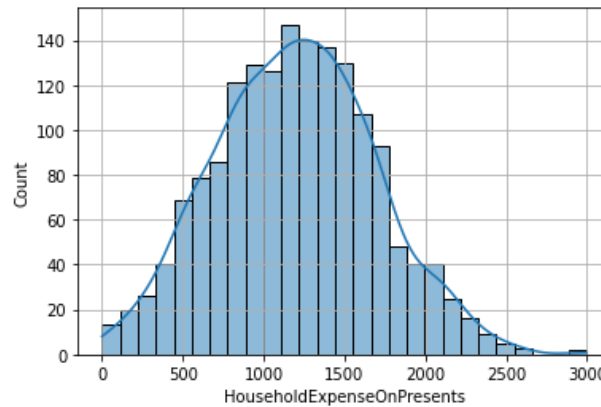
7. לדעתנו, עמודות שיש למחוק טרם תחילת העובדה, הינן עמודות שאחוז המידע החסר בהן הוא מעל סף מסוים (לדוגמא 20%). בנוסף, משתנים שהם קטגוריאליים ולא ניתנים להשמה על סקאלה והערכים הייחודיים בהם רבים (למשל 80% מהערכים הם ייחודיים ולא ניתנים להעמדה על סקאלה/גרף). על כן העמודות שהחלטנו למחוק הינן:

- Address, Region – משתנה קטגוריאלי עם אפשרויות רבות. גם לאחר יצירת העמודה Region מספר הערכים הייחודיים היה רב (מעל 80%) ולכן בחרנו להתעלם משתי עמודות אלו. נדגיש כי הצלחנו להגיע לקביעה זאת רק בזכות המודיפקציה שעשינו ביצירת העמודה Region (הצדקה מדוע בסעיף הקודם הצגנו את העמודה וכבר עכשיו אנחנו מוחקים אותה).
 - Job – מספר הערכים הייחודיים גבוהה מ-80% ולא ניתנים להשמה על סקאלה.
 - PCR: עבור בדיקות PCR11 מתקיים כי רק ל-516/3000 הדוגמאות יש תוצאה, כלומר לאחוז לא מבוטל אין תוצאה, ולכן בחרנו להתעלם מפיצ'ר זה. עבור בדיקת PCR15 רק ל-520/3000 יש תוצאה, וגם כאן בחרנו לוותר על פיצ'ר זה.
- בנוסף, בדקנו מה הקורלציה בין שני המשתנים הללו לסיווג אדם כחולה קורונה. בשני המקרים הקורלציה הייתה מועטה – פחות מ-0.05.

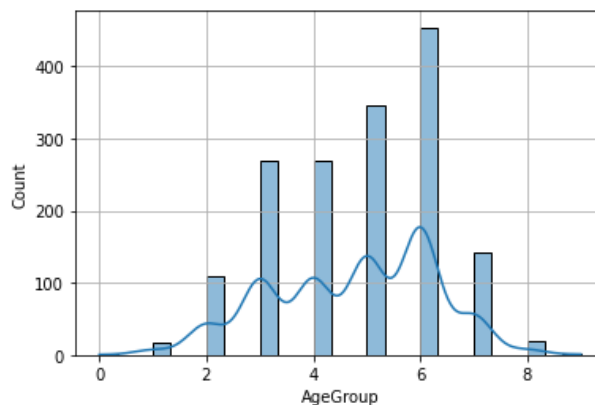
עבור משתנים נומריים, החלטנו לבדוק כיצד הם מתפלגים ולפי ההתפלגות למלא ערכים חסרים (פירוט בשאלה הבאה).

8. להלן האופנים בהם בחרנו למלא את המידע החסר בהתייחסות לאופן התפלגות הנתונים:

- עבור דאטא נומרי (מטיפוסים int, float) יצרנו היסטוגרמה לכל עמודה מסוג זה (ללא הערכים החסרים). ביצענו קירוב לפונקציה חלקה כפי שהראו בתרגול 1, ולפי הפלט הערכנו מה ההתפלגות. נציג כמה גרפים המייצגים את אופן הקבלה למילוי ערך חסר:
- * בגרף זה ניתן לראות כי ההתפלגות שואפת להיות נורמלית, ולכן נבחר למלא את הערך החסר לפי הערך הממוצע.

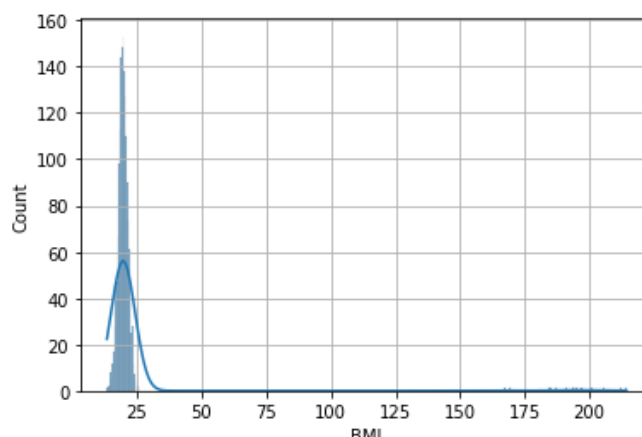


- * בגרף זה ניתן לראות כי לערך הממוצע אין משמעות, אלא דווקא לחציון, ולכן נבחר בו למילוי הערכים החסרים.

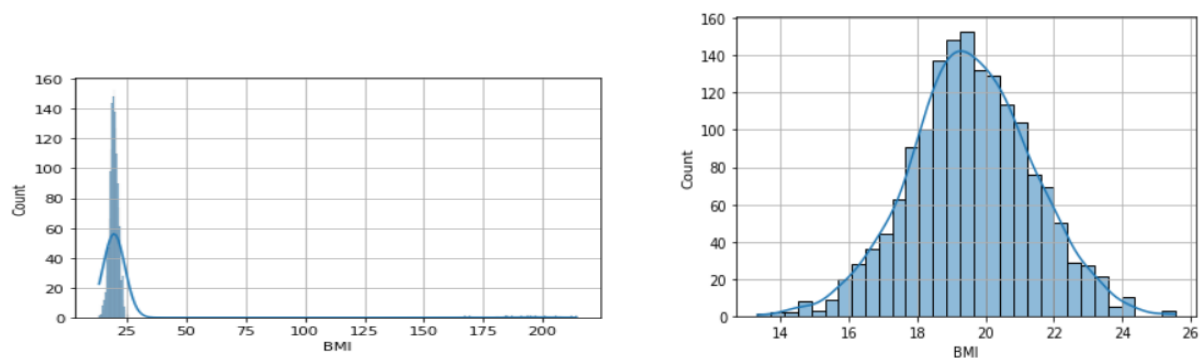


באופן זה עברנו על כל ההתפלגויות שניתן לאפיין באחת משני דרכים אלו, והשלמנו בהן ערכים חסרים.

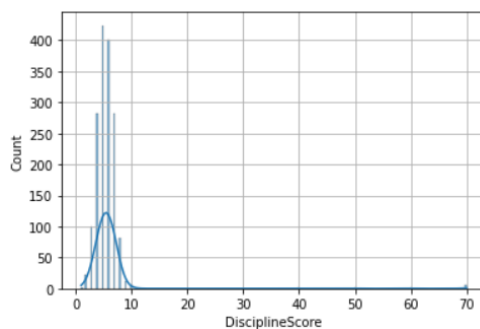
נציין כי במקומות שבבירור ניתן לראות בגרף שיש ערכים חריגים כפי שמצורף בדוגמה מטה, הסרנו את ערכים אלו (תוצאה שתקרה באופן מפורמל בסעיף הבא).



9. נראה דוגמה לבעיה המתוארת. עבור עמודת BMI התוצאות הקיימות מוצגות בגרף השמאלי, ואילו הגרף הימני מייצג את התוצאות שעברו פילטור – כאלה שה BMI שלהם קטן מ-40. קל לראות שרוב הדגימות מרוכזות בחלק מסוים של הגרף, אך יש כמה נקודות "בעייתיות" המקבלות ערכים חריגים.



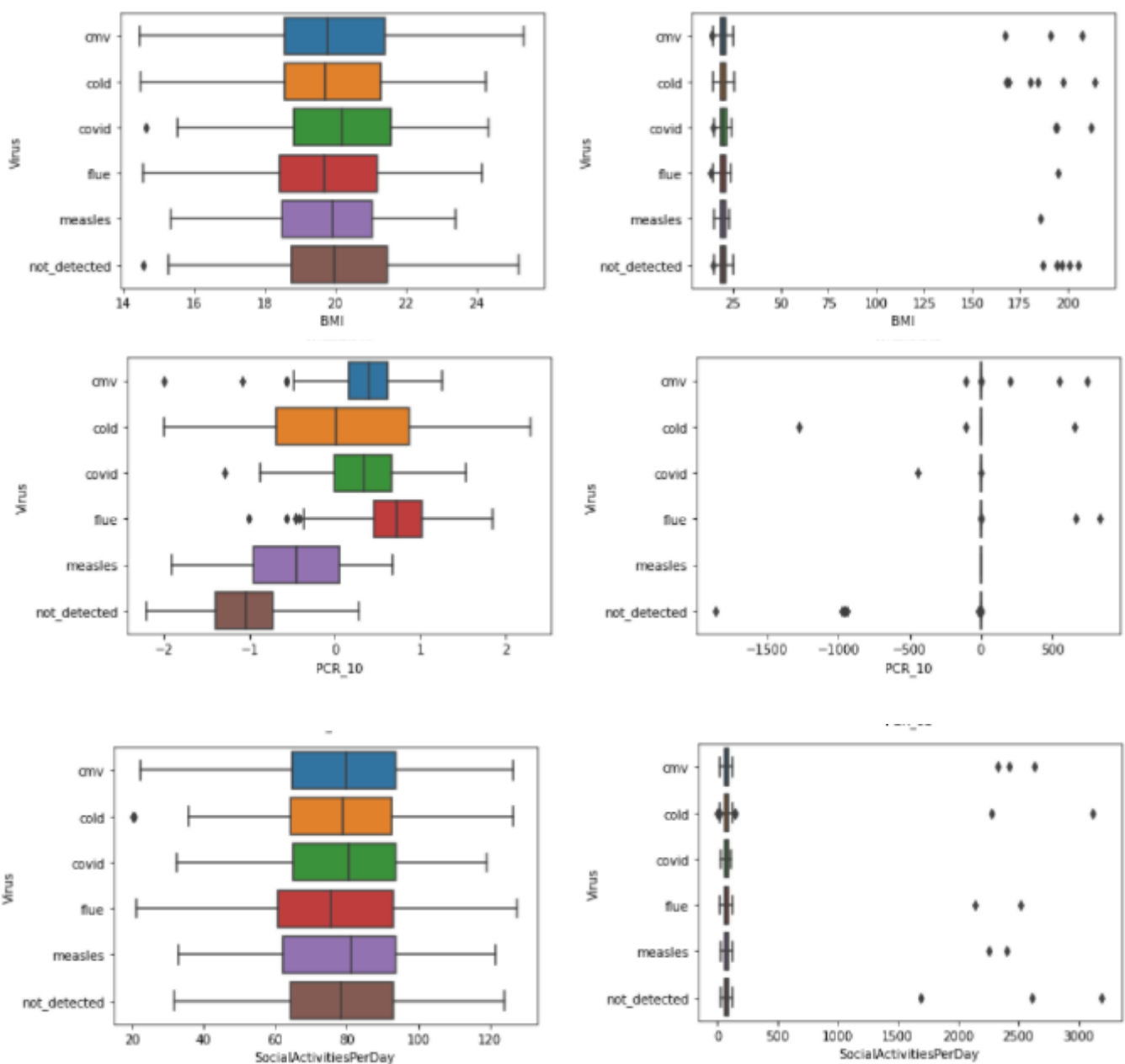
בפעולה זו השתמשנו בידע מקדים שלנו על פרמטר BMI, אך מן הסתם לא תמיד נוכל לעשות זאת. דוגמה נוספת מסוג זה בה ניתן לראות בבירור כי קיימים ערכים חריגים בבחינת קבלת ערכים מחוץ להתפלגות ברורה-



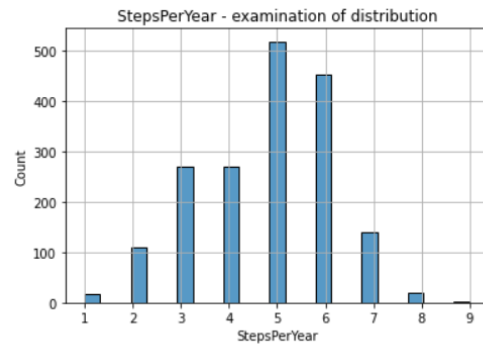
10. ביצענו טיפול בערכים חריגים עבור ערכים נומריים. ביצענו זאת בשני אופנים :

- עבור דאטא המתפלג נורמלית – השתמשנו במבחן z-score. עבור ערכים שקיבלנו במבחן זה ניקוד שערכו המוחלט גדול מ-3, החלפנו אותם בערך בתוחלת של הדאטה.
נשים לב שההנחה כי ההתפלגות נורמלית כאן הינה הכרחית על מנת שתהיה משמעות להחלפת הערכים בתוחלת. כמובן שזה גם חסרון של השיטה כי לא תמיד ניתן להניח כי מידע מתפלג נורמלית (אנחנו ביצענו קביעה זו לפי התבוננות בגרפים של היסטוגרמות של כל פיצ'ר).
- עבור דאטא נומרי אחר – השתמשנו ב-boxplot כדי להתבונן מי הם הערכים החריגים. השתמשנו בשיטת 20-80 (הורדת 20 אחוזים העליונים והתחתונים של הדאטא) והחלפת הערכים הקיצוניים בערך החציוני.

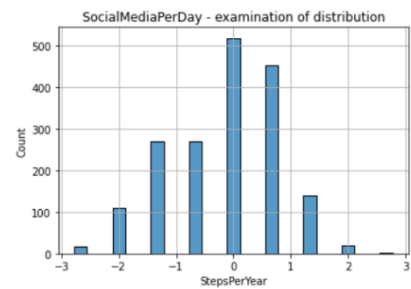
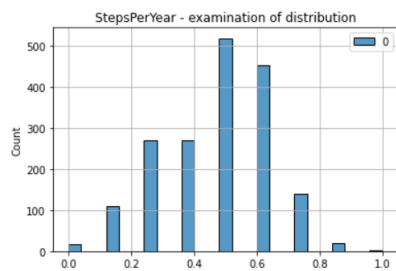
Boxplot הינה דרך מצוינת לזהות בצורה אוטומטית את הערכים החריגים ומבלי להניח ידע מקדים על התפלגות המידע.



11. נשים לב שכך נראית התפלגות הפיצ'ר -



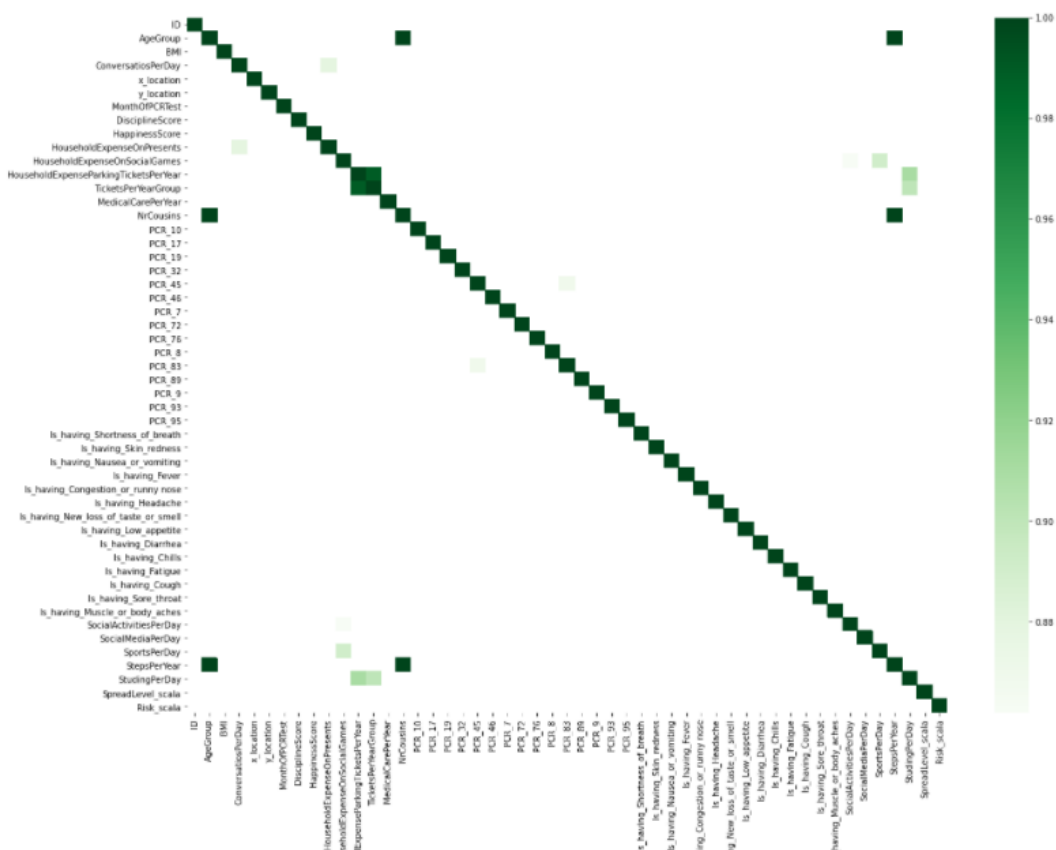
הפעלנו נירמול של פיצ'ר זה באמצעות z_score (מימין) ו- $minmax$ (משמאל):



בחרנו בסופו של דבר לנרמל עם $min-max$ כדי לשמר את התפלגות הדאטא.

- Attribute/Feature Construction – יצרנו פיצ'ר חדש בשם 'MonthOfPCRTTest' שמבצע parsing על המידע בפיצ'ר 'DateOfPCRTTest'. ככה ביצענו קיבוץ של המידע לאינטרוולים של הזמן, שניתן לשים על סקאלה בצורה נוחה וגם מונע פיזור יתר של המידע על פני ערכים רבים.
- Discretization – יצרנו פיצ'ר חדש בשם 'TicketsPerYearGroup' שמתבסס על הפיצ'ר 'HouseholdExpensiveParkingTicketsPerYear'. אפיינו וחלקנו את כל הדוגמאות לשבע רמות. בחרנו דווקא במשתנה זה, כי ההתפלגות שלו הייתה "ריבועית", קיימת חלוקה שווה לאורך טווח הערכים שקיבל, ולכן תהיה שקולה אם נחלק ל-7 קבוצות.
- Categorical to Numeric Conversion – הפכנו את הפיצ'ר של 'Risk' ו-'SpreadLevel' מקטגוריאלי לנומרי, באמצעות הטרנספורמציה $\{1,2,3\} \rightarrow \{low, medium, high\}$ בהתאמה. באופן זה, נתנו משמעות יחסית בין הקטגוריות ונתנו אפשרות לייצג את הדאטא באופן גרפי.
- Scaling (standardization, normalization) – ביצענו נירמול לכל הדאטא הנומרי כפי שפורט בשיטת min-max. הסיבה לכך היא שבהמשך אנו הולכים להשתמש ב-wrapper method שהולך להשתמש ב-knn. לכן, חשוב היה לנו להביא את כל הערכים של הפיצ'רים להיות סביב אותה סקאלה כדי שלכל פיצ'ר תהיה אותה חשיבות (אם פיצ'ר מקבל ערכים בסקאלה של מאות אלפים, ופיצ'ר אחר הוא בסקאלה של אחדות בודדות, אוטומטית הפיצ'ר הראשון יקבל עדיפות על פני השני). לאחר ביצוע פעולה זו, ה-type הפך להיות float.

13. ייצרנו מטריצת קורלציה (correlation) כפי שהוצגה בתרגול (מופיע בקוד). בנוסף, ייצרנו פלט שמסמך את כל התכונות עם קורלציה הגבוהה מ-0.85 וזו התוצאה :



אנו רואים קורלציה מלאה בין הפיצ'רים *AgeGroup*, *NrCousins*, *StepsPerYear* מידע שגרם לנו לזרוק את שני הפיצ'רים האחרונים ברשימה. בנוסף, קיימת קורלציה של 0.98 בין PCR_{83} , PCR_{45} ולכן זרקנו את PCR_{83} . זרקנו את *sports per day*, *SocialActivitiesPerDay* עקב קורלציה של 0.89 ו-0.86 בהתאמה ל- *HouseholdExpenseOnSocialGames*.

הערה: ראו בהמשך הערה נוספת לגבי ניפוי ערכים בסעיף בוא אנו מציגים את ממצאי filter method.

14. ביצענו בחירה של הפיצ'רים למודל שלנו באמצעות מספר דרכים:

Wrapper method – forward selection

ביצענו בחירה אקטיבית של פיצ'רים וזאת על ידי מימוש המבוסס על linear regression. ביצענו ארבע הרצות באמצעות שיטה זו:

- פעם אחת כאשר תיג המטרה שעניין אותנו הוא Risk (לאחר המרה לתיג מספרי שקול).
- פעם שניה עבור SpreadLevel (שוב, עם תיג מספרי שקול).
- פעם שלישית עבור תיג בינארי חדש שייצרנו: האם האובייקט אובחן כחולה covid או שלא.
- פעם רביעית עבור תיג מספרי של כלל המחלות המיוצגות ע"י Virus. המחשבה מאחורה היא להצליח לייצר הפרדה כלשהי בין מחלות שונות, גם אם הפרדה הזאת מלאכותית לחלוטין.

שמנו לב שבערך החל מהפיצ'ר ה-7-8 המודל עובר לבחור בערכים בעלי חשיבות נמוכה (בזכות הצלבה עם filter method ובחינת התפלגויות) ולכן בכל פעם הסתכלנו רק על 6 הפיצ'רים הראשיים שהמודל החזיר בכל הרצה. איחוד רשימת הפיצ'רים שנבחרו יחדיו בכל הניסויים הינם:

'BMI', 'ConversationsPerDay', 'DisciplineScore', 'HappinessScore',
'HouseholdExpenseOnPresents', 'HouseholdExpenseOnSocialGames',
'Is_having_Diarrhea', 'Is_having_Shortness_of_breath', 'MedicalCarePerYear', 'PCR_10',
'PCR_17', 'PCR_19', 'PCR_72', 'PCR_8', 'PCR_89', 'PCR_9', 'PCR_95', 'SocialMediaPerDay'

נעיר כי כל הפיצ'רים מהצורה Is_having... הם פיצ'רים שייצרנו מהפיצ'ר Self_declaration_of_Illness_Form לכן בפועל אנחנו בוחרים רק בפיצ'ר אחד מתוך רשימת הפיצ'רים המקורית שלנו.

לסיכום ביניים, בשיטה זו הצלחנו לבחור 17 פיצ'רים מהנתונים המקוריים, שלהבנתנו משקפים תכונות בעלות יכולת הפרדה גבוהה בין דוגמאות עם תיגים שונים.

נציין כי גם ב-Filter method אנו רואים התאמה לתוצאות אלו.

Wrapper method – backward selection

ביצענו בחירה אקטיבית של פיצ'רים שבכוונתנו להוריד וזאת על ידי מימוש המבוסס על KNN כאשר תיג המטרה הינו התיג הבינארי החדש שייצרנו ומציין האם האובייקט אובחן כחולה covid או לא. בשיטה זאת בדקנו מבין הפיצ'רים שלא נבחרו קודם ב-forward selection מי תורם בצורה הנמוכה ביותר לרמת הדיוק של המודל. הפיצ'רים שזיהינו בשיטה זאת הינם:

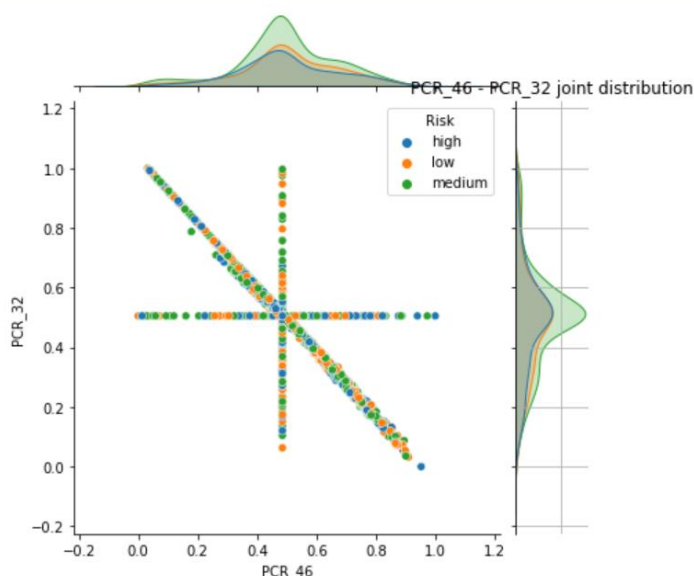
'ID', 'Is_having_Fever', 'AgeGroup', 'x_location', 'Is_having_Muscle_or_body_aches',
'HouseholdExpenseParkingTicketsPerYear', 'TicketsPerYearGroup', 'PCR_32'

נציין כי x_location, y_location הינם פיצ'רים אשר יצרנו מהפיצ'ר CurrentLocation ולאור תוצאה זו בחרנו להסיר את שלושתם (y_location לא קיבל הכרעה אך גם לא הראה מידע מעניין ולפיכך ל-CurrentLocation אין תועלת).

Filter method

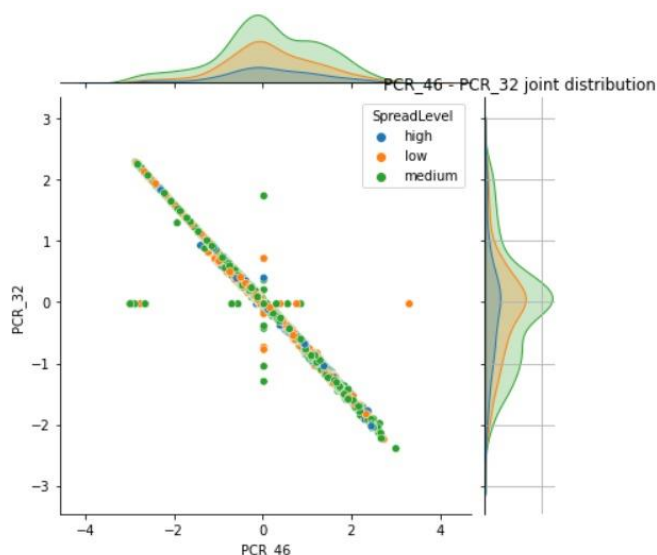
ביצענו הדפסה של גרפים בגישת Filter method, בין כל שני פיצ'רים שונים, ובכל פעם לפי 3 תיוגי המטרה השונים. נציין כי השימוש בשיטה זו היה בעיקר לטובת וידוא ותצוגה וויזואלית של תוצאות Wrapper method, ואכן ראינו התאמה בין השיטות השונות.

הנתון המשמעותי ביותר שמצאנו בשיטה זו הינו לגבי PCR_46: בעזרת גרף התפלגותו המשותפת עם PCR_32 שמנו לב כי קיימת ביניהם קורלציה שלילית. בשלב זה הבנו כי בבחינת ההתאמות בין פיצ'רים הינו צריכים לבחון קורלציה בערך מוחלט.



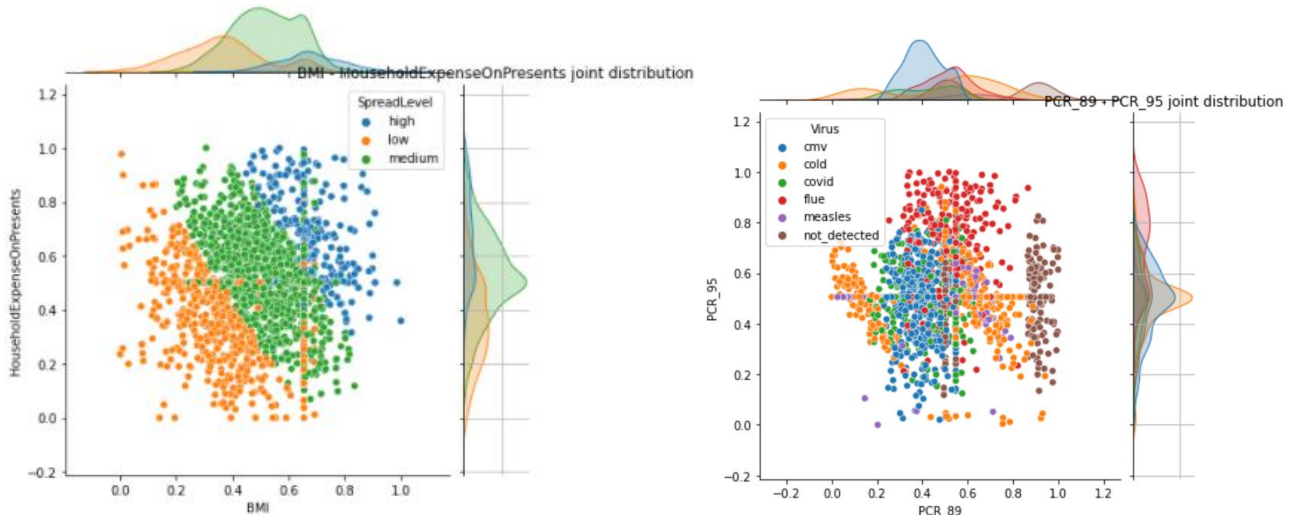
על כן החלטנו להסיר את PCR_46.

הערה: בהזדמנות זאת נציין כי שמנו לב לתת התפלגות בצורת צלב שאנו רואים בתרשימים ומתקיימת בעקבות אופן טיפול ב-outliers וערכים חסרים (השלמה על ידי חציון/ממוצע). להבנתנו, מדובר ב-trade-offs של טיפול בערכי הדאטא שלנו: מידע חסר מול פגיעה בצורת ההתפלגות המקורי (בהתייחסות לתשובה שניתנה לנו בפיאצה בנושא. אנו מבינים כי קיימות שיטות להתמודד עם תופעה זו אך גם כי הותר לנו לפסוח על כך בתרגיל). ספציפית במקרה זה, העובדה כי אנו עובדים בשיטות ניתוח נוספות (למשל הסתכלות בתרשים) עזרה לנו לזהות מקרי קצה מסוג זה ולתת להם התייחסות הולמת. לשם התרשמות נוספת, אנו מוסיפים תרשים של כ-500 דוגמאות בלבד, שעברו outlier חלקי, על מנת לחזק את קיום הקורלציה בין הפיצ'רים.



כפי שציינו, שיטה זאת בעיקר גיבתה את Wrapper method, על כן החלטנו להוסיף כאן מספר תרשימים המייצגים יכולות הפרדה "יפות" (או "מגניבות") של פיצ'רים שכבר סימנו ב-Wrapper method (קצת קישוט ל-PDF 😊):

הערה לבודק: ניתן להריץ את הקוד ולראות את כל התרשימים, אך מכיוון שמדובר באלפי הדפסות חילקנו את משימה זאת למספר בלוקים (קיימת הגבלת פלט לכל בלוק, וכן לטובת מעקב נוח יותר אחר הפלטים).



Filter method and Correlation BoxPlot

פורט באופן מלא על הסרת פיצ'רים עם שיטה זו בשאלה 13.

בחירה ידנית ופיצ'רים ללא הכרעה ברורה

נזכיר כי בסעיף 7 כבר החלטנו להסיר את הפיצ'רים הבאים:

Address, PCR_11, PCR_15, Job

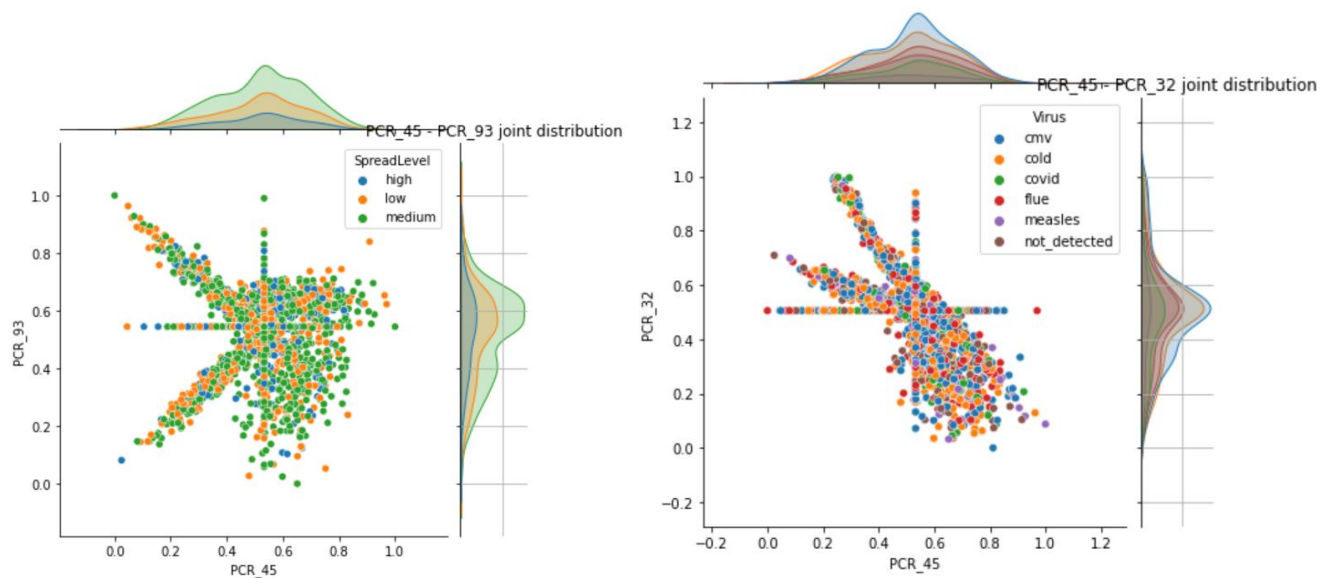
בשלב זה נשארו עם הפיצ'רים הבאים עבורם נדרשנו לקבל החלטה

BloodType, DateOfPCRTTest, Sex, PCR_45, PCR_7, PCR_76, PCR_93, StudingPerDay

החלטנו באופן אקטיבי להסיר תכונות שלא מראות כל יכולת הפרדה טובה מול אף אחד ממשתני המטרה שלנו או התאמה ברורה למשתנה אחר אשר יחד עמו הפרדה כזאת קלה יותר. נציין כי ההחלטה להסיר פיצ'ר הגיעה רק לאחר בחינת התפלגותו גם בעזרת filter method וגם בעזרת Bi-Variate analysis (שבצענו בהמשך) ורק לאחר שלא הצלחנו להראות אף קשר לתיוגי המטרה.

לפיכך בחרנו את StudingPerDay (ניתן להריץ ולראות את הפלטים בקוד שלנו).

כמו כן, החלטנו להסיר את PCR_45. בנוסף על כך שהמשתנה לא הראה כל יכולת הפרדה שימושית, שמנו לב כי הוא מתפלג באופן ייחודי עם משתנים נוספים:



על אף שאנו מתקשים להסביר מה הקשר המתמטי בין הפיצ'רים אנו יכולים לראות כי קיימת התאמה כלשהי ביניהם (על אף שהיא לא עוזרת ביכולת ההפרדה). הנחנו כי הסרת פיצ'ר זה, שגם ככה לא הראה תרומה ברורה, לא תפגע במודל שלנו בשל הקשר בינו לבין פיצ'רים אחרים שנשארו במודל.

על כן, נותרנו עם המשתנים הבאים עבורם אין לנו הכרעה ברורה, ומכיוון שהסרנו כבר מספיק פיצ'רים החלטנו להשאיר אותם במודל

BloodType, DateOfPCRTTest, Sex, PCR_7, PCR_76, PCR_93

נציין כי אם היה עלינו להסיר עוד פיצ'רים (כחלק מהגבלה מספרית) הינו בוחרים להסיר חלק מהפיצ'רים האחרונים (כפי שהסרנו את StudingPerDay).

הערה: ביצענו data transformation לקבוצות test, validation לפי המידע שהפקנו מ-train. כלומר, מילאנו ערכים חסרים (על פי השיקולים שביצענו לפי ההתפלגויות ב-train) לפי הערכים החציוניים/ התוחלת של train, וכן נרמול min-max (עם ערכי min, max של train) עבור כל הפיצ'רים הרלוונטיים. בנוסף, זרקנו את העמודות שבחרנו לזרוק עקב הניתוח בחלק 3 של train מ-test ו-validation. בעת, לאחר שנתייחסו ועיבדנו את train, ניתן לעבוד גם עם test ו-validation מבלי שעברנו על הדאטא שם.

15. להלן טבלת כל המשתנים תוך סיווג האם בחרנו להשתמש במשתנה או לזרוק אותו:

מס'	פיצ'ר	נלקח (Y) / נזרק (N)	סיבה
0	ID	N	wrapper method - backward
1	Address	N	did not found useful information in data
2	AgeGroup	N	wrapper method - backward
3	BMI	Y	wrapper method - forward
4	BloodType	Y	leftovers - left due to lack of clear decision
5	ConversationsPerDay	Y	wrapper method - forward
6	CurrentLocation	N	did not found useful information in data
7	DateOfPCRTTest	Y	leftovers - left due to lack of clear decision
8	DisciplineScore	Y	wrapper method - forward
9	HappinessScore	Y	wrapper method - forward
10	HouseholdExpenseOnPresents	Y	wrapper method - forward
11	HouseholdExpenseOnSocialGames	Y	wrapper method - forward
12	HouseholdExpenseParkingTicketsPerYear	N	wrapper method - backward
13	Job	N	did not found useful information in data
14	MedicalCarePerYear	Y	wrapper method - forward
15	NrCousins	N	high correlation with another attribute
16	PCR_10	Y	wrapper method - forward
17	PCR_11	N	low number of samples
18	PCR_15	N	low number of samples
19	PCR_17	Y	wrapper method - forward
20	PCR_19	Y	wrapper method - forward
21	PCR_32	N	wrapper method - backward
22	PCR_45	N	manually removed due lack of useful data in our study (filter method and bi-variate)
23	PCR_46	N	high correlation with another attribute
24	PCR_7	Y	leftovers - left due to lack of clear decision
25	PCR_72	Y	wrapper method - forward
26	PCR_76	Y	leftovers - left due to lack of clear decision
27	PCR_8	Y	wrapper method - forward
28	PCR_83	N	high correlation with another attribute
29	PCR_89	Y	wrapper method - forward
30	PCR_9	Y	wrapper method - forward
31	PCR_93	Y	leftovers - left due to lack of clear decision
32	PCR_95	Y	wrapper method - forward
33	Self_declaration_of_Illness_Form	Y	wrapper method - forward
34	Sex	Y	leftovers - left due to lack of clear decision
35	SocialActivitiesPerDay	N	high correlation with another attribute
36	SocialMediaPerDay	Y	wrapper method - forward
37	SportsPerDay	N	high correlation with another attribute
38	StepsPerYear	N	high correlation with another attribute
39	StudyingPerDay	N	manually removed due lack of useful data in our study (filter method and bi-variate)
40	Virus	Y	target
41	SpreadLevel	Y	target
42	Risk	Y	target