236756 – מבוא למערכות לומדות – תרגיל בית רטוב 1

שחר כץ 313574766, סיון יוסבשוילי 318981586 | מאי 2021

2+3.

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

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_lol, 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 מגוריו, 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:

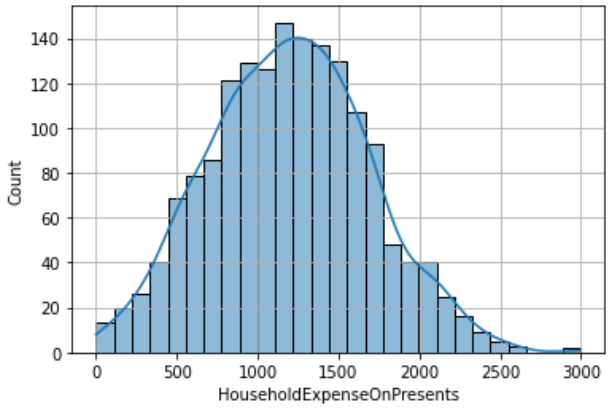
- Address, Region-משתנה קטגוריאלי עם אפשרויות רבות. גם לאחר יצירת העמודה Region מספר הערכים הייחודים היה רב (מעל 80%) ולכן בחרנו להתעלם משתי עמודות אלו. נדגיש כי הצלחנו להגיע לקביעה זאת רק בזכות המודיפקציה שעשינו ביצירת העמודה Region (הצדקה מדוע בסעיף הקודם הצגנו את העמודה וכבר עכשיו אנחנו מוחקים אותה).
- Job – מספר הערכים הייחודים גבוהה מ-80% ולא ניתנים להשמה על סקלאה.
- 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).

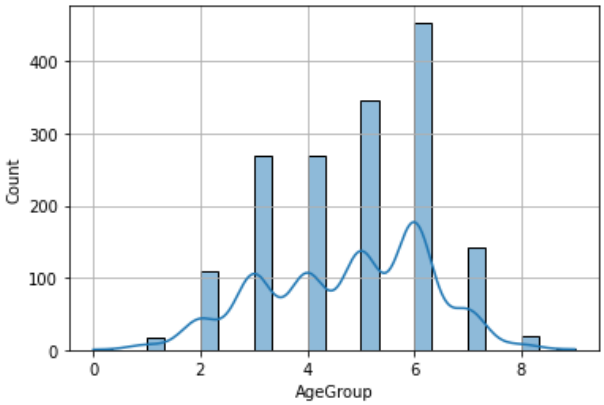
8. להלן האופנים בהם בחרנו למלא את המידע החסר בהתייחסות לאופן התפלגות הנתונים:

- 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.

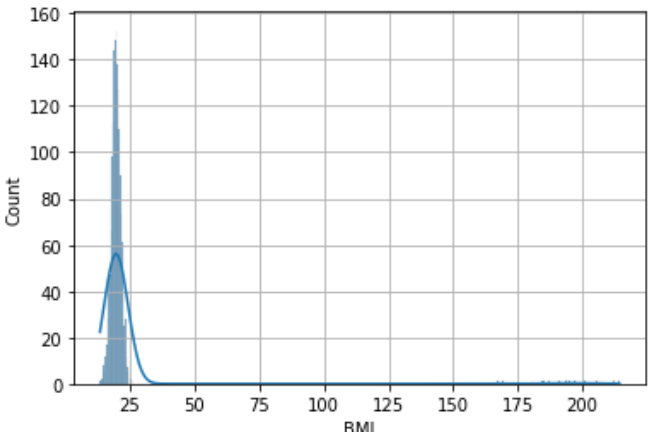


* בגרף זה ניתן לראות כי לערך הממוצע אין משמעות, אלא דווקא לחציון, ולכן נבחר בו למילוי הערכים החסרים.

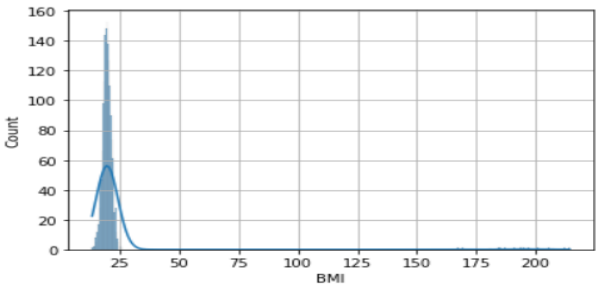


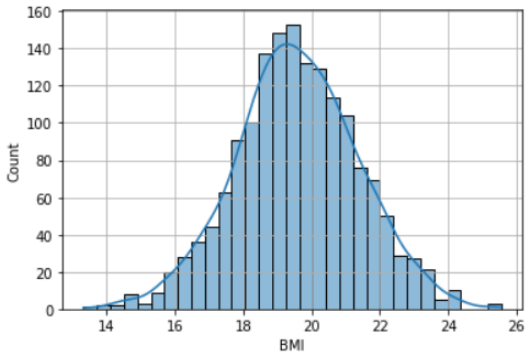
באופן זה עברנו על כל ההתפלגויות שניתן לאפיין באחת משני דרכים אלו, והשלמנו בהן ערכים חסרים.

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

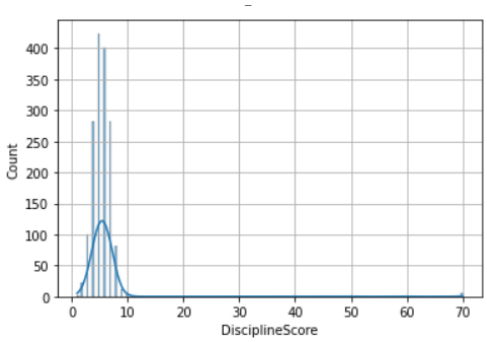


9 .נראה דוגמה לבעיה המתוארת. עבור עמודת BMI התוצאות הקיימות מוצגות בגרף השמאלי, ואילו הגרף הימני מייצג את התוצאות שעברו פילטור – כאלה שה BMI-שלהם קטן מ-40. קל לראות שרוב הדגימות מרוכזות בחלק מסוים של הגרף ,אך יש כמה נקודות "בעייתיות" המקבלות ערכים חריגים.





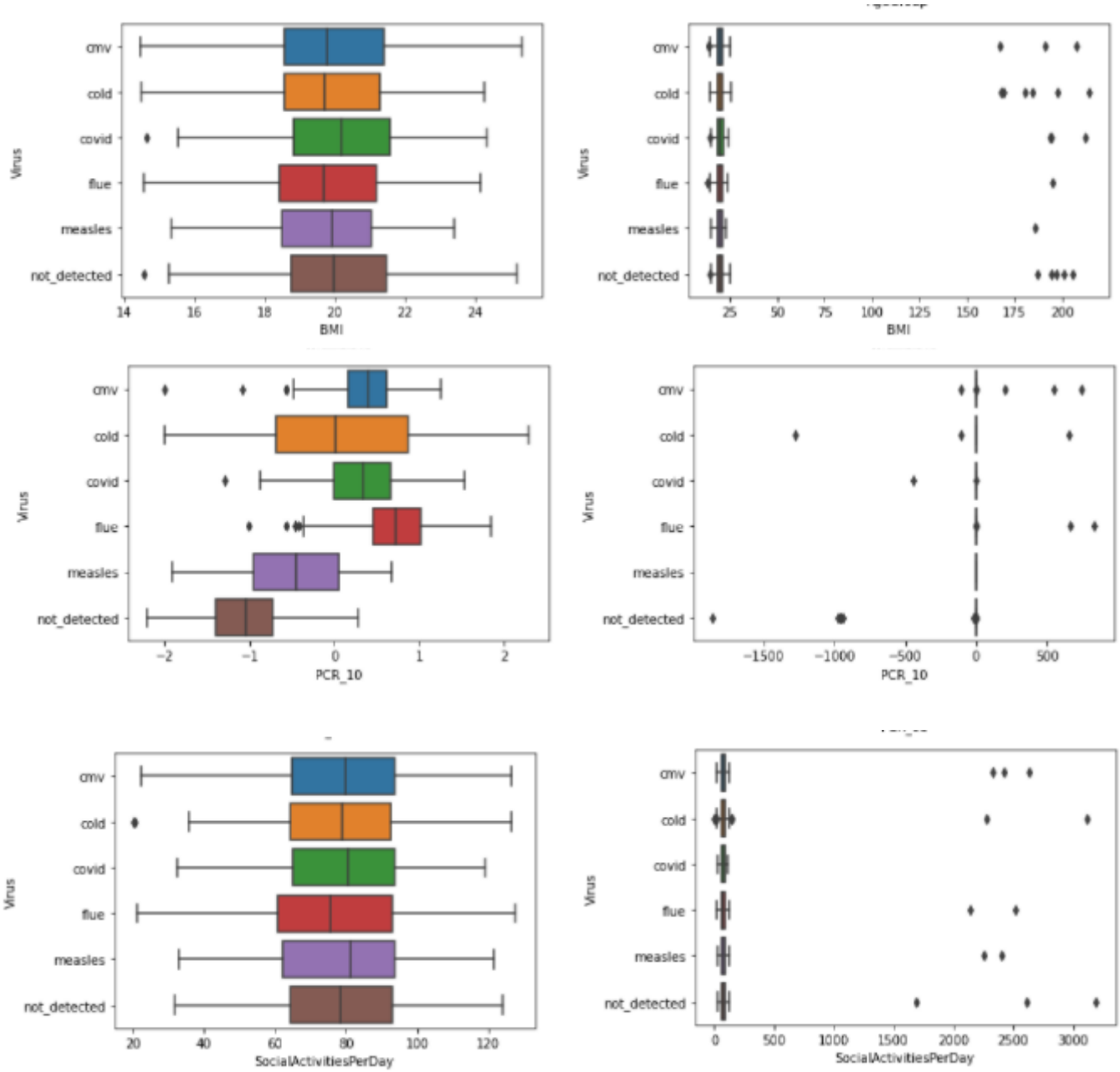
בפעולה זו השתמשנו בידע מקדים שלנו על פרמטר BMI,אך מן הסתם לא תמיד נוכל לעשות זאת. דוגמה נוספת מסוג זה בה ניתן לראות בבירור כי קיימים ערכים חריגים בבחינת קבלת ערכים מחוץ להתפלגות ברורה-



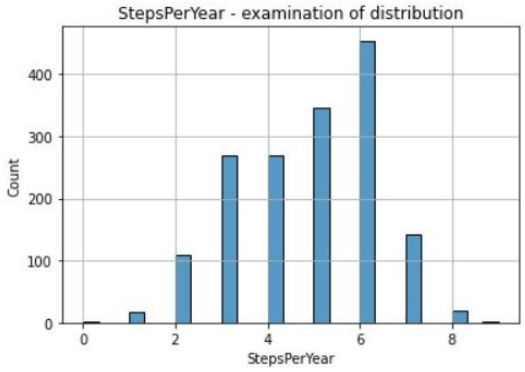
10. ביצענו טיפול בערכים חריגים עבור ערכים נומריים. ביצענו זאת בשני אופנים :

- 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.

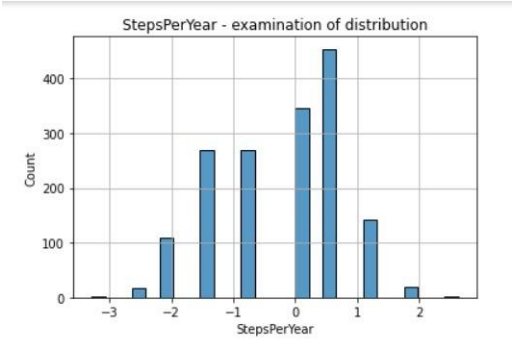
Boxplot הינה דרך מצוינת לזהות בצורה אוטומטית את הערכים החריגים ומבלי להניח ידע מקדים על התפלגות המידע.

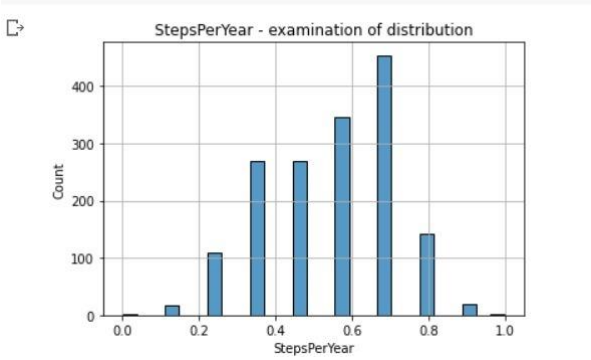


11.נשים לב שכך נראית התפלגות הפיצ'ר -



הפעלנו נירמול של פיצ'ר זה באמצעות z-score (מימין) ו-min-max (משמאל):



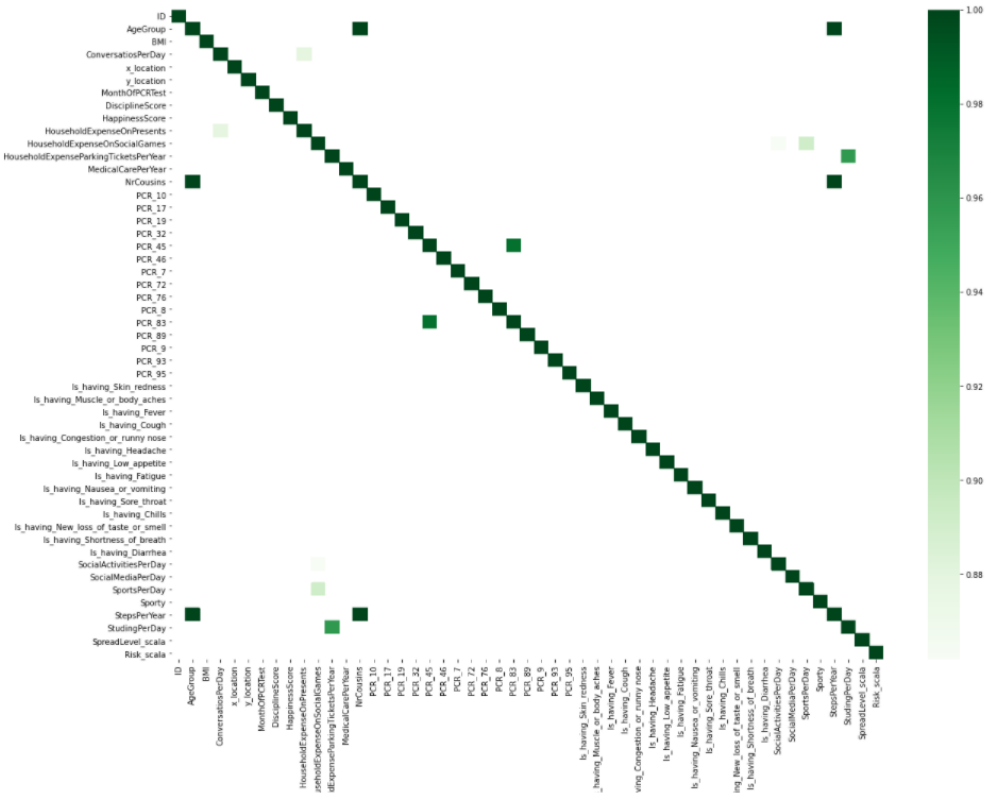


בחרנו בסופו של דבר לנרמל עם min-max כדי לשמר את התפלגות הדאטא.

12. נבחן כל שיטה של data-transformation ונתייחס אליה:

- Attribute/Feature Construction – יצרנו פיצ'ר חדש בשם 'MonthOfPCRTest' שמבצע parsing על המידע בפיצ'ר 'DateOfPCRTest'. 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 $\{\text{low}, \text{nedium}, \text{high}\} \rightarrow \{1,2,3\}$ 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. בנוסף, ייצרנו פלט שמסמן 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 $sports\ per\ day$ כי יש לו קורלציה של 0.89 ל-הוצאות על משחקי וידאו.

ובסוף, זרקנו את מדיה חברתית ליום כי יש לפיצ'ר זה קורלציה של 0.86 להוצאות על משחקי וידאו.

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 תיוג המטרה שעניין אותנו הוא Risk (לאחר המרה לתיוג מספרי שקול).
- פעם שניה עבור SpreadLevel (שוב, עם תיוג מספרי שקול).
- פעם שלישית עבור תיוג בינארי חדש שייצרנו: האם האובייקט אובחן כחולה covid או שלא.
- פעם רביעית עבור תיוג 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-9-10 המודל עובר לבחור בערכים בעלי חשיבות נמוכה (בזכות הצלבה עם filter method ובחינת התפלגויות) ולכן בכל פעם הסתכלנו רק על 8 הפיצ'רים הראשיים שהמודל החזיר בכל הרצה. איחוד רשימת הפיצ'רים שנבחרו יחדיו בכל הניסויים הינם:

['AgeGroup', 'BMI', 'ConversatiosPerDay', 'DisciplineScore', 'HappinessScore', 'HouseholdExpenseOnPresents', 'HouseholdExpenseOnSocialGames', 'ID', 'Is_having_Cough', 'Is_having_Diarrhea', 'Is_having_Fatigue', 'Is_having_Fever', 'Is_having_Shortness_of_breath', 'MedicalCarePerYear', 'PCR_10', 'PCR_17', 'PCR_19', 'PCR_32', 'PCR_72', 'PCR_8', 'PCR_89', 'PCR_9', 'PCR_95', 'SocialMediaPerDay']

נעיר כי כל הפיצ'רים מהצורה ...Is_having_ הם פיצ'רים שיצרנו
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 20 פיצ'רים מהנתונים המקוריים, שלהבנתנו משקפים תכונות בעלות יכולת הפרדה גבוהה בין דוגמאות עם תיוגים 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 תיוג המטרה הינו התיוג 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:

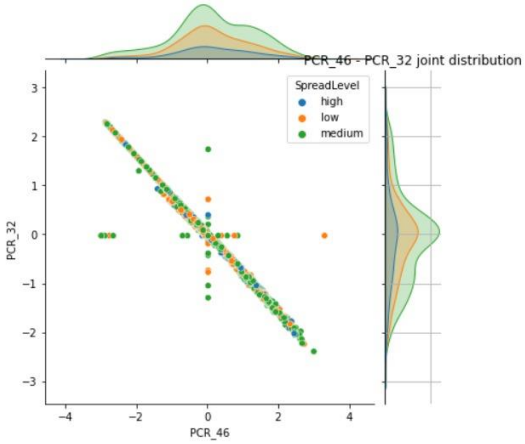
['x_location,'HouseholdExpenseParkingTicketsPerYear,'PCR_45,'PCR_46,'PCR_7'
,y_location,'PCR_76,'PCR_93']

נציין כי x_location , y_location הינם פיצ'רים אשר יצרנו מהפיצ'ר CurrentLocation ולאור תוצאה זו בחרנו להסירם ואת CurrentLocation.

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 ווידוא ותצוגה וויזואלית של תוצאות Wrapper method, ואכן ראינו התאמה בין השיטות השונות.

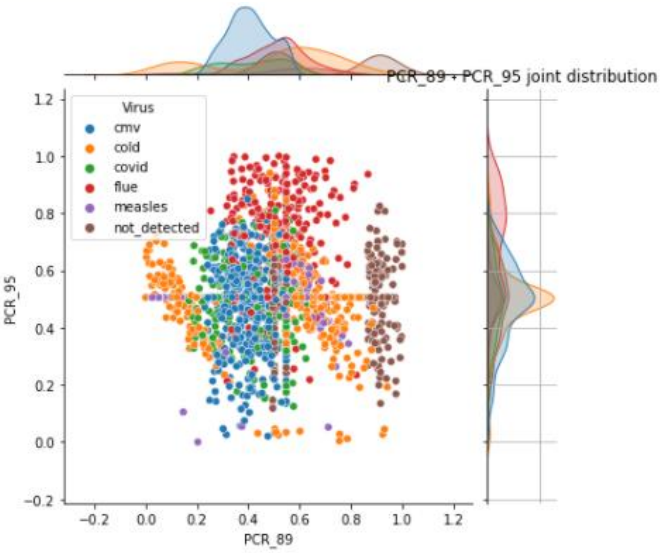
הנתון המשמעותי ביותר שמצאנו בשיטה זו הינו לגבי PCR_46: בעזרת גרף התפלגותו המשותפת עם PCR_32 שמנו לב כי קיימת ביניהם קורלציה שלילית. בשלב זה הבנו כי בבחינת ההתאמות בין פיצ'רים הינו צריכים לבחון קורלציה בערך מוחלט.

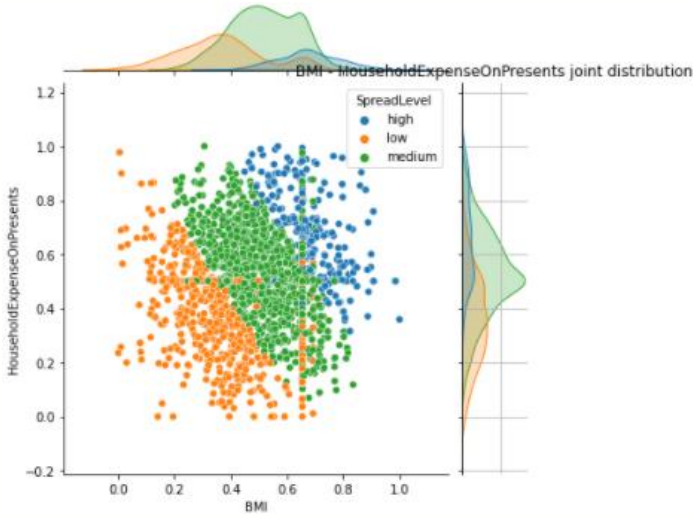


על כן החלטנו להסיר את PCR_46.

כפי שציינו, שיטה זאת בעיקר גיבתה את Wrapper method, על כן החלטנו להוסיף כאן מספר תרשימים המייצגים יכולות הפרדה "יפות" (או "מגניבות") של פי'צרים 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.

ניקוי ידני של פיצ'רים

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

Address, PCR_11, PCR_15, Job,

בנוסף החלטנו באופן אקטיבי להסיר תכונה שלא מראה 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 ההתפלגותו גם בעזרת filter method וגם בעזרת Bi-Variate analysis (שבצענו בהמשך) ורק לאחר שלא הצלחנו להראות אף קשר לתיוגי המטרה.

המשתנה שבחרנו הוא StudingPerDay.

בסיום ניפוי זה נותרנו עם המשתנים הבאים, לגביהם אין לנו דעה החלטית האם הם תורמים באופן משמעותי או פוגעים ביכולת המיון, על כן השארנו אותם במודל:

BloodType, DateOfPCRTest, Sex

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

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

מס'	פיצ'ר	נלקח / נשאר	סיבה
0	ID	Y	נבחר ב- wrapper method
1	Address	N	נזרק עקב אחוז ערכים ייחודיים גבוהה (לא נומרי)
2	AgeGroup	Y	נבחר ב- wrapper method
3	BMI	Y	נבחר ב- wrapper method
4	BloodType	Y	לא נזרק משום סיבה
5	ConversatiosPerDay	Y	נבחר ב- wrapper method
6	CurrentLocation	N	נזרק ב-wrapper method ומחוסר מידע שימושי ב-filter method מהעמודות שניסינו לייצר ממנו
7	DateOfPCRTest	Y	לא נזרק משום סיבה
8	DisciplineScore	Y	נבחר ב-wrapper method
9	HappinessScore	Y	נבחר ב-wrapper method
10	HouseholdExpenseOnPresents	Y	נבחר ב-wrapper method
11	HouseholdExpenseOnSocialGames	Y	נבחר ב-wrapper method
12	HouseholdExpenseParkingTicketsPerYear	N	נזרק ב-wrapper method
13	Job	N	נזרק עקב אחוז ערכים ייחודיים גבוהה (לא נומרי)
14	MedicalCarePerYear	Y	נבחר ב-wrapper method
15	NrCousins	N	נזרק עקב קורלציה לפיצ'ר אחר
16	PCR_10	Y	נבחר ב-wrapper method
17	PCR_11	N	נזרק עקב ערכים חסרים רבים
18	PCR_15	N	נזרק עקב ערכים חסרים רבים
19	PCR_17	Y	נבחר ב-wrapper method
20	PCR_19	Y	נבחר ב-wrapper method
21	PCR_32	Y	נבחר ב-wrapper method
22	PCR_45	N	נזרק ב-wrapper method
23	PCR_46	N	נזרק ב-wrapper method
24	PCR_7	N	נזרק ב-wrapper method
25	PCR_72	Y	נבחר ב-wrapper method
26	PCR_76	N	נזרק ב-wrapper method
27	PCR_8	Y	נבחר ב-wrapper method
28	PCR_83	N	נזרק עקב קורלציה לפיצ'ר אחר
29	PCR_89	Y	נבחר ב-wrapper method
30	PCR_9	Y	נבחר ב-wrapper method
31	PCR_93	N	נזרק ב-wrapper method
32	PCR_95	Y	נבחר ב-wrapper method
33	Self_declaration_of_Illness_Form	Y	נבחר ב-wrapper method (פיצ'רים שנוצרו מעמודה זו)
34	Sex	Y	לא נזרק משום סיבה
35	SocialActivitiesPerDay	N	נזרק עקב קורלציה לפיצ'ר אחר
36	SocialMediaPerDay	Y	נבחר ב-wrapper method
37	SportsPerDay	N	נזרק עקב קורלציה לפיצ'ר אחר
38	StepsPerYear	N	נזרק עקב קורלציה לפיצ'ר אחר
39	StudingPerDay	Y	לא נזרק משום סיבה