קורס לימוד מכונה
פרויקט חלק א

Smoking Data

מגישים

אופיר אזולאי	319145959
לירון מרינברג	316266501

תוכן עניינים

הגדרת הבעיה

תיאור כללי של עולם התוכן הנחקר

הבעיה המחקרית- בהינתן נתונים ביולוגיים בריאותיים בסיסיים אודות מטופלים, נרצה לבצע קלסיפיקציה של המטופלים לשתי קטגוריית- קטגוריית המעשנים וקטגוריית הלא מעשנים, כלומר נרצה להיות מסוגלים לחזות את סטטוס העישון של מטופל בהסתמך על נתוניו הביולוגיים הבסיסיים.

בעולם חיים למעלה ממיליארד מעשנים ומחקרים רבים הראו כי אוכלוסייה מעשנת היא בעלת סיכון מוגבר לחלות במחלות לב, כלי דם, סרטן, מחלות דרכי הנשימה ועוד כך שעישון הוגדר על-ידי ארגון הבריאות העולמי כסיבת המוות המובילה ביותר הניתנת למניעה.

מחקרים רבים בוצעו על מנת לפתח מודלים של למידת מכונה המשמשים לביצוע סיווג וניבוי לסטטוס העישון של נבדק לפי נתוניו הביולוגיים. מודלים אלו מהווים יתרון מהותי באיתור מטופלים מעשנים מבין נבדקים לאור העובדה שמידע על הרגלי העישון של נבדקי מחקר אינו תמיד זמין, והיכן שהוא זמין, הוא אינו בהכרח אמין כיוון שלעיתים קרובות נבדקים מדווחים דיווח לקוי אודות סטטוס העישון שלהם.

דוגמא למודל שעסק בנושא אשר הוצע על ידי Beineke הנו מודל חיזוי המבוסס על ביטוי של גנים מסוימים לסטטוס העישון של הנבדק. מודל זה הורחב במחקר שנעשה ע"י Wang ב-2021 בו התבצע שימוש בכימות של רמות הגנים על מנת להשיג דיוק ניבוי גדול יותר של סטטוס העישון של הנבדקים. מחקר נוסף שנעשה ע"י Zhang השתמש במתודולוגיית למידת מכונה המתבססות על רגרסיה לוגיסטית לפנים עם סף מובהקות שרירותי כקריטריון עצירה כדי להכשיר מסווג לחיזוי מצב עישון. מחקר זה הורחב ע"י Bollepalli בכך שאפשר לסיווג בן שלוש רמות (בעוד שהמחקר הקודם הבחין רק בין מעשנים ללא מעשנים מחקר זה סיווג בין מעשנים שהפסיקו, מעשנים כיום ולא מעשנים כלל) תוך שימוש באימות צולב פנימי (CV) כדי להחליט כמה תכונות לכלול במודל ע"י הערכה ישירה של ביצועי המסווגים החזויים.

הגדרת שאלת המחקר

שאלת המחקר היא האם מטופל מסוים מסווג כמעשן או כלא מעשן בהתבסס על נתוניו הביולוגיים הבסיסיים. באמצעות מתודולוגיות וכלים שונים מעולם למידת מכונה שמסייעות בעיבוד הנתונים ובחירת משתנים רלוונטיים נוכל לפתח מודל חיזוי שיאפשר לחזות בצורה מדויקת ככל הניתן את סטטוס העישון של מטופל. בהמשך נצפה שמודל זה יהיה מסוגל לסייע באיתור מעשנים בקבוצות מחקר כאשר נתון זה לוקה בחסר.

הבנת הנתונים

1. **תיעוד מקורות הנתונים ומשמעותם**

1.1. **מקור הנתונים במאגר אתו אנו עובדים ויצירת הנתונים -**סט נתונים ביולוגיים על מטופלים שנאספו בדרכים מקובלות כגון בדיקות רפואיות, מדידות גובה/משקל/היקפים, דגימות דם/שתן ועוד.

1.2. **מאפייני סט הנתונים-**

סט הנתונים איתו נעבוד בחלק זה הנו סט של נתוני האימון המיוצג בטבלת Excel המאוכלסת ב-44,553 רשומות, כך שכל רשומה מייצגת נתונים ביולוגיים בסיסיים של מטופל מסוים. הטבלה מכילה 26 עמודות הנותנות ייצוג למאפיין ביולוגי אחר של מטופל כאשר כלל הערכים מוצגים ע"י ערכים מספריים או בינאריים.

המאפיינים המופיעים בסט הנתונים:

מספר	שם המשתנה	תיאור	סוג	ערכים (לא כולל חריגים)
1	ID	תעודת זהות המטופל	שלם וחיובי	
2	gender	מגדר המטופל	קטגוריאלי	{M,F} M-זכר, F-נקבה
3	age	גיל המטופל	רציף	[20,85] באינטרוולים של 5 שנים
4	height(cm)	גובה המטופל בס"מ	רציף	[130,190] אינטרוולים של 5 ס"מ
5	weight(kg)	משקל המטופל בק"ג	רציף	[30,135] באינטרוולים של 5 ק"ג
6	Waist(cm)	היקף המותניים של המטופל בס"מ	רציף	[51,129]
7	eyesight(left)	מספר ראיה בעין שמאל ערכיו נורמלו סביב הערך 1 $Value_{real} > 0 \rightarrow Value_{normalized} > 1$ $Value_{real} < 0 \rightarrow Value_{normalized} < 1$	רציף	[0.1,2]
8	eyesight(right)	מספר ראיה בעין ימין ערכיו נורמלו סביב הערך 1 $Value_{real} > 0 \rightarrow Value_{normalized} > 1$ $Value_{real} < 0 \rightarrow Value_{normalized} < 1$	רציף	[0.1,2]
9	hearing(left)	שמיעה באוזן שמאל	קטגוריאלי	{1,2}
10	hearing(right)	שמיעה באוזן ימין	קטגוריאלי	{1,2}
11	systolic	לחץ דם בחדרים בזמן כיווץ הלב	רציף	[71,240]
12	relaxation	לחץ דם בחדרים בזמן הרפיית הלב	רציף	[40,146]
13	fasting blood sugar	סוכר בדם בצום	רציף	[46,423]
14	Cholesterol	סה"כ כולסטרול	רציף	[72,410]
15	triglyceride	טריגליצריד	רציף	[11,548]
16	HDL	כולסטרול HDL (הכולסטרול הטוב)	רציף	[4,159]
17	LDL	כולסטרול LDL (הכולסטרול הרע)	רציף	[1,336]
18	hemoglobin	המוגלובין	רציף	[4.9,21.1]
19	Urine protein	חלבון בשתן	בדיד	[1,6]
20	serum creatinine	תפקוד כלייתי	רציף	[0.1,10]
21	AST	אנזים כבד - סוג AST	רציף	[6,1090]
22	ALT	אנזים כבד - סוג ALT	רציף	[1,2914]
23	Gtp	מדד לתפקוד כבד	רציף	[1,976]
24	oral	בחינה בעל פה	קבוע	{Y}
25	dental caries	עששת שיניים	קטגוריאלי	{0,1} 0- עששת,1-אין עששת
26	tartar	מצב אבנית	קטגוריאלי	{N,Y} Y- אבנית, N-אין אבנית

משתנה המטרה smoking - משתנה קטגוריאלי המציין האם המטופל מעשן (1) או אינו מעשן (0). עבור כל מטופל ערך משתנה זה מוצג בסט האימון כלומר מצוין האם המטופל הנו מעשן או לא והמטרה תהיה לחזות את ערכו בהמשך ע"י מודל אנליטי עבור כל מטופל נתון בסט נתוני הבחינה.

2. הסתברויות אפריוריות וקשרים בין מאפיינים

2.1 משתנים קטגוריאליים:

Gender- ניתן לראות כי סט הנתונים מורכב ברובו מנתוני גברים. מדובר בהתפלגות שלא מייצגת נכונה את אוכלוסיית העולם, כאשר המצב האמיתי המתפלגת פחות או יותר בצורה שווה בין גברים ונשים (נספח 1.1).

$P(gender = 'F')=36.41\%$, $P(gender = 'M')=63.59\%$

hearing(left)- ניתן לראות כי הרוב המוחלט של הנבדקים הם בעלי שמיעה טובה באוזן שמאל ובודדים אותרו עם לקות שמיעה באוזן שמאל.(נספח 1.2)

$P(1=left\ good\ hearing)=97.43\%$, $P(2=left\ bad\ hearing)=2.57\%$

hearing(right)- ניתן לראות כי הרוב המוחלט של הנבדקים הם בעלי שמיעה טובה באוזן ימין ובודדים אותרו עם לקות שמיעה באוזן ימין. (נספח 1.3)

$P(1= right\ good\ hearing)=97.35\%$, $P(2= right\ bad\ hearing)=2.66\%$

dental caries- ניתן לראות שלמרבית הנבדקים אין עששת בעוד שאחוז קטן יותר של מטופלים סובלים מעששת שיניים. למרות שאחוז החולים בעששת קטן יותר הוא עדיין לא אפסי ונבין כי עששת שיניים אינה נדירה בקרב נבדקי המדגם (נספח 1.4)

$P(Dental\ caries = 0)=78.63\%,(Dental\ caries = 0)=21.37\%$

tartar- ניתן לראות שלמרות שאחוז המטופלים שתועדו כבעלי אבנית בשיניים הוא קטן יותר מכאלו שלא, ההבדל אינו משמעותי וניתן להבין כי אבנית שיניים הנה תופעה די שכיחה שאותרה אצל כמחצית מהאוכלוסייה הנבדקת. (נספח 1.5)

$P(Tartar= Y)=44.68\%, P(Tartar = N)=55.32\%$

Smoking- ניתן לראות כי מרבית האוכלוסייה הנבדקת אותרה כלא מעשנת (כ-36% מעשנים), אולם המדגם לא מייצג באופן מדויק את האוכלוסייה העולמית שכן לפי מחקרים, שיעור המעשנים בעולם עומד על כ-22%. למרות זאת, יש לתת את הדעת אודות העובדה כי רב האוכלוסייה הנבדקת הם גברים, אשר מהווים נתח גדול יותר משיעור המעשנים העולמי,(36% מהגברים לעומת 6% מהנשים על פי מידע מאוניברסיטת אוקספורד) (נספח 1.6)

$P(Smoking = 1)=36.75\%$, $P(Smoking = 0)=63.25\%$

Urine protein- משתנה זה מייצג רמת חלבון בשתן מבין 6 רמות אפשריות. התרשים המוצג הוא לאחר טיפול בערך מילולי שגוי (כמתואר בסעיף ניקוי החריגים). ניתן לראות כי הרוב המוחלט של הנבדקים הם בעלי רמה 1 של חלבון בשתן וכן ככל שהרמה גבוהה יותר כך התופעה נדירה יותר בקרב הנבדקים. (נספח 1.7)

$P(\text{Unine protein} = 1) = 94.49\%$

Oral- משתנה זה מקבל ערך קבוע "Y"

$P(Oral = 1) = 100\%$

2.2 משתנים רציפים:

Age- (נספח 1.8) התפלגות משתנה זה יכולה להתפרש כנורמלית, וניתן לראות שמרבית הנבדקים נמצאים בגילאים 40-50. ישנם מעט נתונים על אוכלוסייה מבוגרת(לכל היותר גיל 90) או צעירה(לכל הפחות גיל 20)

height(cm)- (נספח 1.22) התפלגות משתנה זה יכולה להתפרש כנורמלית, וניתן לראות שמרבית הנבדקים

הם בטווח הגבהים $1.6\text{-}1.7M$. ישנם מעט נבדקים גבוהים או נמוכים משמעותית אך ברמה עקרונית הנתונים די מייצגים את האוכלוסייה.

Waist(cm)- (נספח 1.23) - התפלגות משתנה זה יכולה להתפרש כנורמלית, וניתן לראות שמרבית הנבדקים הם בעלי היקף מותניים סביר: כ-85 כלומר הנתונים די מייצגים את האוכלוסייה.

Weight(kg)- (נספח 1.19) התפלגות משתנה זה יכולה להתפרש כנורמלית, וניתן לראות שמרבית הנבדקים הם בעלי משקל סביר: כ-50-70 כלומר הנתונים די מייצגים את האוכלוסייה.

eyesight(left/right)- (נספח 1.9), (נספח 1.10) מהתפלגות המשתנה ניתן לראות כי מירב התצפיות מציגות ערכים סביב הערך 1 בהתאם לנרמול שמשתנה זה עבר, אך ישנו זנב שמעיד על נוכחות של חריגים שאינם תקינים (9.9). בחריגים אלו טיפלנו כמתואר בסעיף חריגים.

Relaxation- (נספח 1.18) משתנה זה המייצג לחץ דם דיאסטולי (בזמן הרפיית הלב) מיוצג במדגם ע"י התפלגות שנראית כנורמלית, כאשר הערכים השכיחים ביותר הם סביב הערך 80 שהנו הנורמה בהתאמה לעולה מן הספרות.

Cholesterol- (נספח 1.25) התפלגות משתנה זה נראית נורמלית, וניתן לראות שלמרבית הנבדקים נמדד סה"כ כולסטרול של 200 שהינו ערך תקני. כמובן שבמדגם ניתן למצוא גם ערכים גבוהים(עד כ-400 המעידים על עודף כולסטרול) אך כולם תקינים בהתאם לערכים שהמדד עשוי לקבל.

AST/ALT- (נספח 1.12, נספח 1.14)לאחר צפייה בהתפלגות המשתנה ראינו כי הוא מכיל ערכים קיצוניים שאינם שכיחים ויוצרים זנב ארוך (מגיע לכ-1020) נתון שאינו מתיישב עם טווח הערכים התקינים ביניהם הערך השכיח שאמור להיות בטווח של 25-30 יחידות של משתנים אלו. בהמשך בחרנו לבצע דיסקרטיזציה למשתנה זה על מנת להתמודד עם הקטנת הרעש של ערכים חריגים אלו. (דיסקריטיזציה)

Hemoglobin (נספח 1.16)- התפלגות משתנה זה נראית נורמלית, וניתן לראות שלמרבית הנבדקים נמדד מדד של 15-17 שהינו ערך תקני. כמובן שבמדגם ניתן למצוא גם ערכים גבוהים(עד כ-23 המעידים על רמת המוגלובין גבוהה) אך כולם תקינים בהתאם לערכים שהמדד עשוי לקבל.

Systolic- עבור משתנה זה, לפני הטיפול בערכים החריגים ניתן לראות זנב ארוך (נספח 1.21.1

עבור המשתנים הבאים בחרנו להציג את ההסטוגרמות רק לאחר טיפול בחריגים שכן לפני הטיפול נוכחנו לראות זנב מאוד ארוך שאפיין את ההתפלגויות והקשה על ראיה של הערכים השכיחים ושל אופן ההתפלגות. לאחר הטיפול, זנב זה נורמל באופן שתואר בסעיף חריגים.

fasting blood sugar-(נספח 1.11) לאחר הנרמול, התפלגות של המדד נראית תקינה סביב ערך שכיח של כ-100 עם זנב ימני המעיד על נוכחות נבדקים בעלי רמה גבוהה יחסית של סוכר אך עדיין נתונים הגיוניים.

Triglyceride- (נספח 1.24) - לאחר הנרמול, התפלגות של המדד נראית תקינה סביב ערך שכיח של כ-100 עם זנב ימני המעיד על נוכחות נבדקים בעלי רמה גבוהה יחסית של טריגליצריד אך עדיין נתונים הגיוניים.

HDL- (נספח 1.15) , **LDL-** (נספח 1.17) **-** לאחר הנרמול, התפלגות המדדים נראית תקינה סביב ערך שכיח של כ-100/60 בהתאמה.

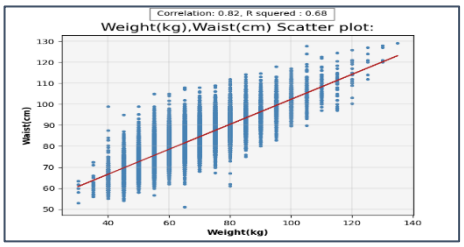
serum creatinine- (נספח 1.20) - לאחר הנרמול, התפלגות המדדים נראית נורמלית ותקינה סביב ערכים שכיחים של כ-0.8-1.

GTP- (נספח 1.13) לאחר הנרמול, התפלגות המדד עדיין מכילה זנב די ארוך כלומר יש עדיין ערכים קיצוניים שאינם שכיחים אך סביר שנאספו במדגם כה גדול. הזנב הוקטן משמעותית מהזנב הקיצוני שהוצג לפני הנרמול(נספח 1.13.1) שהכיל ערכים שהגיעו ל1000 שאינם הגיוניים כשמדובר במדד GTP .

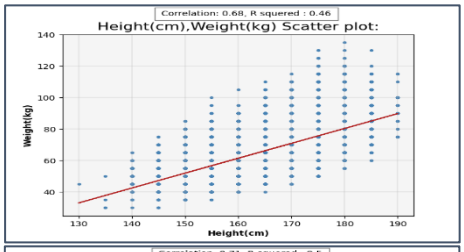
2.3 קשרים בין מאפיינים:

תחילה בדקנו את מפת הקורלציות בין המשתנים הרציפים(נספח 2), על מנת למצוא קשרים בין המסבירים השונים. ניתן לראות כי בין רוב זוגות המשתנים, לא מדובר במתאמים גבוהים למדי, אם כי ניתן לזהות זוגות מסוימים ביניהם המתאם גבוה באופן יחסי. יחד עם זאת נזכור כי מתאם פירסון מייצג קשר ליניארי אך תיתכן תלות שאינה ליניארית בין משתנים שונים שלא נוכל להבחין בה רק על פי מתאם פירסון.

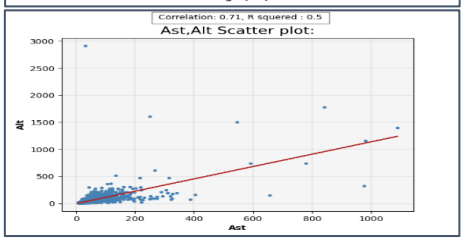
להלן בחינה של קשרים אלו וקשרים נוספים שצפינו שימצאו מתוך הבנת המאפיינים וחקר בספרות:



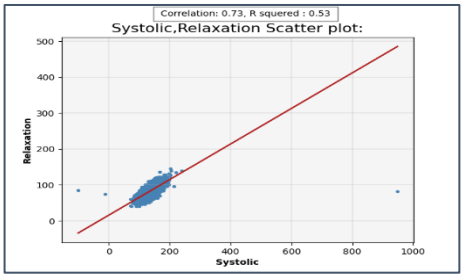
2.3.1 waist(cm)- weight(kg) - (מקדם מתאם 0.82)- כצפוי בהתאם לקורלציה החיובית ניתן לראות מגמת עליה משמעותית בגרף בין משקל להיקף מותניים מה שמייצג שקיים בניהם קשר חיובי כך שככל שהמשקל גדול יותר היקף המותניים גדול יותר בהתאמה כמצופה ממדדים אלו.



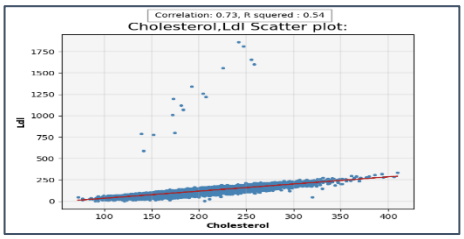
2.3.2 **height(cm) - weight(kg) -** (מקדם מתאם 0.68) - כצפוי בהתאם לקורלציה החיובית ניתן לראות מגמת עליה משמעותית בגרף בין גובה למשקל מה שמייצג שקיים בניהם קשר חיובי כך שככל שהגובה גדול יותר המשקל גדול יותר בהתאמה כמצופה ממדדים אלו.



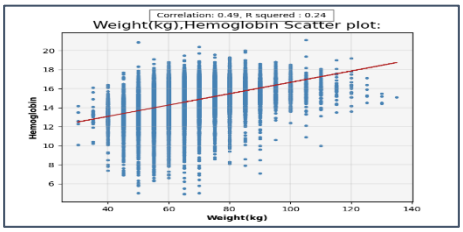
2.3.3 AST-ALT- (מקדם מתאם 0.71) -כצפוי בהתאם לקורלציה החיובית ניתן לראות מגמת עליה בגרף המייצגת שקיים קשר חיובי בין שני סוגי אנזימי הכבד כלומר הופעת כמות גדולה של אנזים כבד מסוג אחד לרוב תגיע עם הופעת כמות גדולה של אנזים כבד מסוג שני.



2.3.4 systolic - relaxation- (מקדם מתאם 0.73) כצפוי בהתאם לקורלציה החיובית ניתן לראות מגמת עליה בגרף מה שמייצג שקיים קשר חיובי בין לחץ דם בהתכווצות הלב- סיסטולי(systolic) ללחץ דם בזמן הרפיית הלב (relaxation) כלומר למטופל עם לחץ דם סיסטולי גבוה, נצפה לראות לחץ דם גבוה יותר גם במדידה דיאסטולית.



2.3.5 Cholesterol - LDL- (מקדם מתאם 0.73) - כצפוי בהתאם לקורלציה החיובית ניתן לראות מגמת עליה בגרף מה שמייצג שקיים קשר חיובי בין מדד LDL - סוג של כולסטרול לסה"כ כולסטרול. כלומר אצל מטופל אצלו נמדד ערך גבוה של LDL נצפה לראות מדד גבוה של סה"כ כולסטרול.



2.3.6 hemoglobin - height(cm)- (מקדם מתאם 0.54) - כצפוי בהתאם לקורלציה החיובית ניתן לראות מגמת עליה בגרף מה שמייצג שקיים קשר חיובי בין מדד hemoglobin – לגובה המטופל כלומר אצל מטופלים גבוהים יותר נצפה לראות כמות גבוהה יותר של hemoglobin .

2.4 קשרים נוספים שצפינו שימצאו:

2.4.1 Cholesterol - triglyceride - HDL -LDL- לאחר סקירה על המשתנים LDL, HDL, triglyceride הבנו כי לכולם יש תרומה כלשהי לסה"כ הכולסטרול בגוף- כלומר הם מהווים מרכיב של כולסטרול. כדי לבחון את הקשר ביניהם לסה"כ כולסטרול ביצענו מבחן רגרסיה ליניארית שתחילה תוצאותיה לא היו מובהקות. כדי להבין מדוע, בדקנו ויזואלית את הנתונים וראינו כי ישנם ערכים חריגים ל- LDL ו-HDL שערכם גדול מסה"כ כולסטרול, דבר שאינו מתיישב עם העובדה הרפואית שהם מרכיבים אותו. לאחר הסרת נתונים אלו (כמפורט בסעיף טיפול בחריגים) ביצענו שוב מבחן רגרסיה שתוצאותיו היו מובהקות לחלוטין ומוצגות להלן (נספח 7):

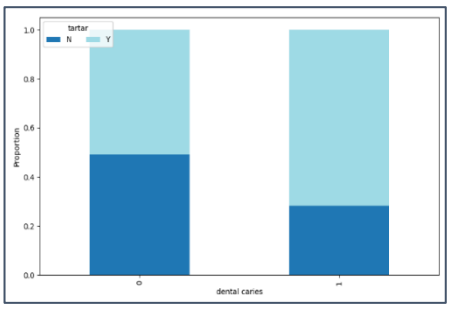
- התקבל **R-squared=1** -ערך גבוה ביותר האפשרי -נסיק שהמודל הליניארי מתאים להסביר את המוסבר.
- ערך **P-Value** אפסי לכל אחד מהמקדמים במודל והמשמעות היא שהמודל ליניארי מובהק ונוסחתו:

$$\boxed{0.9979 * LDL + 1.0077 * HDL + 0.1965 * triglyceride = cholesterol}$$

לאור העובדה שמצאנו קשר ליניארי ברמת מובהקות מקסימלית בין המשתנים האלו למדד הכולסטרול, ניגשנו לספרות לבדיקה חוזרת לשם אימות הממצאים. מצאנו כי הנוסחה המקובלת לחישוב סה"כ כולסטרול הנה:

$$\boxed{LDL + HDL + 0.2 * triglyceride = cholesterol}$$

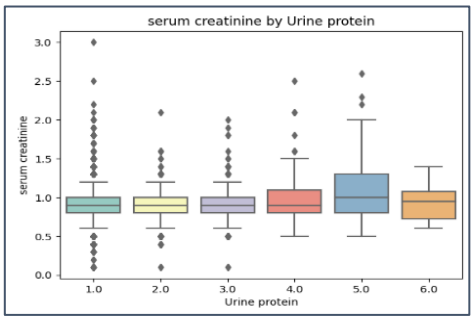
כמעט זהה לחלוטין לממצא שהצלחנו לאתר מהנתונים. נסיק מכל הממצאים שישנה בעיית מולטיקולינאריות חזקה בין המשתנים האלו במודל.



tartar	N	Y	All
dental caries			
0	17184	17751	34935
1	2666	6828	9494
All	19850	24579	44429

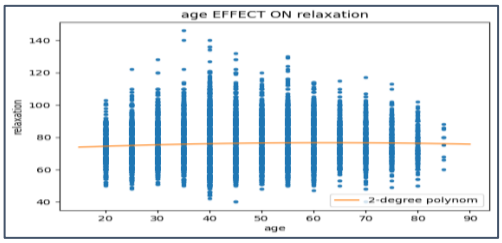
chi2 statistic: 1345.6
p-value: 4.2279e-290
degrees of freedom: 4

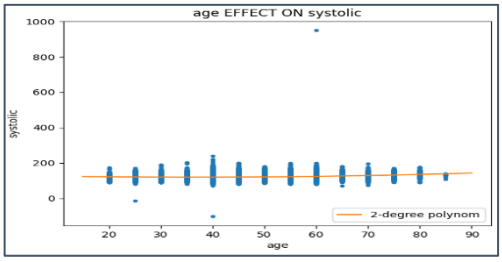
2.4.2 tartar - dental caries - שני המשתנים הקטגוריאליים עששת שיניים ואבנית מתייחסים למצב רפואי של הפה ולכן נצפה לקשר מסוים בניהם. לבחינת מתאם בין שני משתנים קטגוריאליים הפקנו תרשים עמודות מוערם יחסי שמיצג את הגודל היחסי של כל קבוצה (עם אבנית ובלי אבנית) לפי מטופלים עם עששת ובלי עששת. ניתן לראות כי בהשוואה לקבוצת המטופלים ללא העששת(0) בקבוצה שבה הנבדקים היו עם עששת(1), אחוז הנבדקים עם אבנית(Y) היה גדול יותר- נסיק כי ייתכן קשר כלשהו בין משתנים אלו. לבחינת מובהקות הקשר נבצע מבחן סטטיסטי **חי בריבוע** שמתאים לבחינת מתאם בין משתנים קטגוריאליים. ניתן לראות כי ערך P-Value קטן משמעותית מ-0.05 ולכן נאמר כי קיים קשר מובהק בין עששת לבין אבנית בקרב הנבדקים.



2.4.3 serum creatinine -Urine protein מלמידה על מאפיינים ביולוגים אלו הבנו כי ייתכן קשר בין חלבון בשתן לתפקוד הכלייתי. נבצע בדיקה לקשר זה ע"י בחינת Box-Plot המייצג את ההתפלגות המשתנה הרציף- serum creatinine (תפקוד כלייתי)לפי כל אחת מרמות החלבון בשתן-Urine protein (בדיד).מהתרשים לא עולה קשר בין שני המשתנים על אף הציפיה. כמו כן, אומנם ניתן לראות פיזור גדול יותר של התפקוד כלייתי ברמה 1 של חלבון

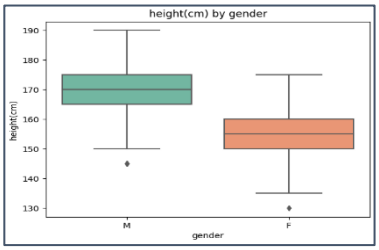
בשתן אך זה נובע מכמות תצפיות גדולה משמעותית לרמת חלבון 1 מכמות התצפיות של שאר הרמות (לפי התפלגות המשתנה Urine protein).

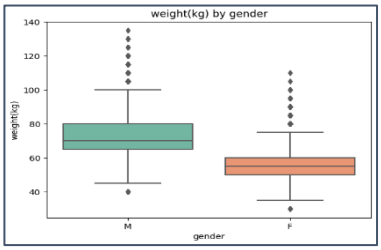


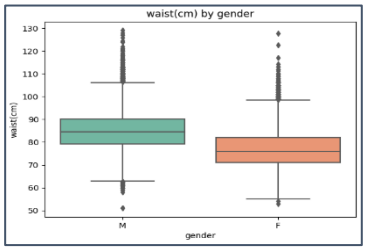


2.4.4 Age -relaxation – (מקדם מתאם 0.051) ו-**Age - systolic** (מקדם מתאם 0.13) אמנם מקדם המתאם בטבלת הקורלציה מזערי ומכך משתמע שבין משתנים אלו אין קשר ליניארי לגיל או שאינו ניכר, אך לא ניתן לשלול קשר אחר שאינו ליניארי. סקרנו ומצאנו כי לחץ דם relaxation - הדיאסטולי (בזמן הרפיית הלב) / סיסטולי (בזמן כיווץ הלב) נוטה לעלות עד גיל 50 ולרדת לאחר מכן. לבדיקת מגמתיות זו ביצענו רגרסיה פולינומיאלית מסדר 2 (ציפינו לעליה וירידה בצורת פרבולה) והפקנו תרשימי פיזור עליהם הוצג קו המגמה. ניתן לראות כי יש בטן זניחה מאוד לקו המגמה הפולינומיאלי ולכן לא ראינו בתצפיות את המגמתיות שציפינו על הקשר הפרבולי בין גיל ללחץ דם דיאסטולי (relaxation) או לסיסטולי.

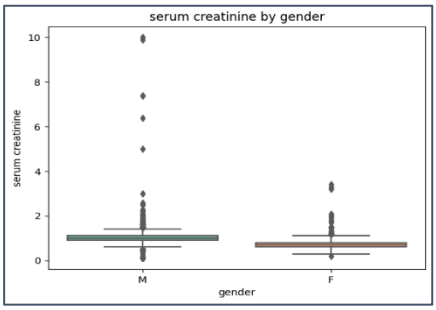
2.4.5 gender-waist(cm)/ gender-weight(kg)/ gender-height(cm) - למרות שהקורלציה בתרשים הקורלציות לא הייתה גבוהה במיוחד, ראינו לנכון לבחון את השפעת המגדר על הגובה, משקל והיקף המותניים מתוך הגיון בסיסי שאם הנתונים אכן מייצגים את האוכלוסייה בעולם, אנו נראה הבדל בין גברים ונשים במדדים אלו. לכן הפקנו תרשימי Box-Plot שמציגים את ההבדלים בערכי המשתנים בין גברים לנשים.







ניתן לראות כי בכל אחד מן המדדים הנ"ל יש הבדל בין גברים ונשים שכן בכל המדדים, הערכים שנמדדו עבור נשים היו נמוכים יותר מערכים הנמדדים בגברים כצפוי.



2.4.6 gender - serum creatinine –בסקירה ראינו כי תפקוד כלייתי של נשים נוטה להיות טוב יותר משל גברים. serum creatinine נמוך מעיד על תפקוד כלייתי טוב. לבדיקה זו הפקנו Box-Plot שמציג את ההבדלים בערכי המשתנה לפי מגדר. ניתן לראות הבדל בין גברים לנשים כך שאצל נשים המדד נמוך יותר. כמו כן ישנן מספר תצפיות חריגות אצל הגברים –לאחר בדיקה בספרות ערכים אלו הגיוניים (אינם טעות בהזנה) אך מעידים על מחלת כליות כלשהי. החריגים נצפו אצל הגברים והממצא מתיישב עם התיאוריה לפיה לנשים תפקוד כליות טוב יותר (מדד נמוך).

2.5 המאפיינים שחשודים כבעלי השפעה על משתנה המטרה- סטטוס העישון:

בהתבסס על מחקרים וידע מקדים, נטען כי המשתנים הבאים עשויים להעיד על סטטוס עישון חיובי למטופל:

- **גיל** (נצפה שאחוז המעשנים בגילאים הצעירים קטן מאחוז המעשנים בגילאים מבוגרים יותר)
- **המוגלובין** גבוה(מעל 17)
- **מגדר** – לפי מחקרים, אחוז הגברים המעשנים מבין הגברים גבוה מאחוז הנשים המעשנות מבין הנשים.

- **serum creatinine** -תפקוד כליות לקוי (ערך גבוה) – מאמרים רבים דנים בפגיעה של עישון בכליות ולכן תפקוד כליות לקוי יכול להעיד על סטטוס עישון חיובי.

בנוסף לכל אלו ביצענו בדיקות מתוך הנתונים על משתנים שעשויים להעיד על סטטוס עישון חיובי:

2.5.1 **עבור כלל המשתנים הרציפים** הפקנו תרשימי Box-Plot (נספח 3) שמאפשרים לראות האם יש הבדל בערכי משתנה רציף מסוים בין ערכים שונים של משתנה קטגוריאלי (במקרה שלנו מעשן/לא מעשן).

מסקנות מרכזיות מן התרשימים-

- ✓ בניגוד להנחה, נמצאו **גילאים** נמוכים יותר בקרב המעשנים בהשוואה ללא מעשנים.(נספח 3.1)
- ✓ מדידות **הגובה, המשקל והיקף המותניים** בקרב המעשנים היו גבוהות יותר בהשוואה לנבדקים הלא מעשנים. (נספח 3.2) (נספח 3.3) , (נספח 3.4)
- ✓ בהתאם לטענה, ניכר שרמת ההמוגלובין (**hemoglobin**) גבוהה יותר בקרב מעשנים (נספח 3.10)
- ✓ בהתאם לטענה שנטענה, מדד **serum creatinine** – תפקוד כליות נמצא גבוה יותר(תפקוד פחות טוב) בקרב מעשנים. (בדיקה זו התבצעה לאחר טיפול בערכים חריגים כפי שיתואר בסעיף חריגים על מנת לקבל הסתכלות ממוקדת יותר על הערכים בטווח השכיח יותר). (נספח 3.11)
- ✓ התוצאות מראות כי מדדי ה-**triglyceride** ומדד ה-**GTP** גבוהים יותר בקרב המעשנים בהשוואה לנבדקים הלא מעשנים. (נספח 3.17) , (נספח 3.14)

2.5.2. **עבור המשתנים הבדידים** הפקנו תרשימי עמודות יחסיים(נספח 4) לבחינת מתאם בין קטגוריאליים.

תוצאות מרכזיות שנמצאו בתרשימים-

- ✓ **מגדר**-אחוז המעשנים מבין הגברים גדול משמעותית מאחוז המעשנים מבין הנשים (נספח 4.1)
- ✓ אחוז המעשנים מבין הנבדקים בעלי עששת(**dental caries**) גדול יותר מאחוז המעשנים בין אלו ללא העששת (נספח 4.4)
- ✓ אחוז המעשנים מבין בעלי אבנית (**tartar**)גדול יותר מאחוז המעשנים בין אלו ללא אבנית (נספח 4.5)

3. איכות הנתונים

הנתונים שנאספו עשויים להכיל חריגות/חוסרים ובכדי לבצע ניתוח נתונים מהימן עלינו לנקותם כשלב מקדים.

3.2 **נתונים חסרים-** נראה כי ישנם חוסרים בסט הנתונים. תחילה בדקנו את מספר המאפיינים החסרים בכל רשומה על מנת לאתר רשומות עם מספר רב של ערכים חסרים. (נספח 5) החלטנו להסיר רשומות בהן יש 3 ומעלה ערכים. הבחירה להסיר רשומות אלו נבעה מתוך הרצון להימנע מכפייה אגרסיבית על הנתונים ע"י השלמה של מספר רב של ערכים. לאחר הסרת רשומות אלו בחנו מחדש את החוסרים בכל עמודה.(נספח 6)

height(cm)- 40 חסרים- נאכלס את חוסרים אלו תוך התחשבות במשתנה המשקל והמגדר שבבחינת הקשרים בין משתנים נמצא שיש בניהם קשר ליניארי די משמעותי. במקום החוסרים נזין את הגובה הממוצע לכל משקל תחת קטגורית המגדר המתאימה.

weight(kg)- 1 חסרים - נאכלס את חוסרים אלו תוך התחשבות במשתנה היקף המותניים שנמצא שיש בניהם קשר ליניארי די משמעותי. במקום החוסרים נזין את המשקל הממוצע לכל קבוצת ערכים של היקף מותניים.

waist(cm)- 1 חסרים- נאכלס את חוסרים אלו תוך התחשבות במשתנה המשקל שבבחינת הקשרים בין משתנים נמצא שיש בניהם קשר ליניארי משמעותי. במקום החוסרים נזין את ההיקף הממוצע בהתאם למשקל.

eyesight(left)- 1 חסרים השלמה על פי הערך בעין ימין.

eyesight(right) – 5 חסרים – נאכלס בערכים החסרים את הערך שקיים בעמודה המייצגת את הראיה בעין השנייה שאומנם הקשר בניהן לא היה משמעותי אך הוא בעל המתאם הגבוהה ביותר מבין שאר המשתנים.

systolic-11 חסרים –השלמה על פי רגרסיה לינארית לפי המשתנה Relaxation- המשתנה בעל הקורלציה הגבוהה ביותר, ומתוך הבנה כי לחץ הדם הסיסטולי והדיאסטולי קשורים אחד לשני.

Relaxation- 4 חסרים – השלמה על פי רגרסיה כאשר המשתנה המסביר הוא חלץ הדם הסיסטולי.

fasting blood sugar- 2 חסרים- השלמה על פי הערך הממוצע.

Cholesterol- 6 חסרים, **Triglyceride**-16חסרים, **HDL-** 3 חסרים, **LDL-** 2 חסרים- כל אחד מן המשתנים הללו השלמנו על פי הנוסחה שמצאנו בסעיף הקשרים בין המאפיינים.

Hemoglobin-6 חסרים- השלמה לפי רגרסיה ליניארית כאשר המסביר הוא מאפיין המשקל שנמצא בעל קורלציה גבוהה ביותר עם משתנה זה.

Urine protein – 5 חסרים השלמה על פי הערך השכיח- '1'.

AST /serum creatinine – 1 חסרים השלמה על פי הממוצע.

ALT – 20 חסרים השלמה על פי רגרסיה עם כאשר המשתנה המסביר הוא AST.

Oral – 2 חסרים - נאכלס את הערך 'Y' שהוא קבוע לכל שאר הרשומות

Tartar- 1 חסר- נאכלס את חוסרים אלו תוך התחשבות במשתנה dental caries . שכן ראינו קשר בין משתנים אלו. במקום החוסרים נזין את הערך השכיח ביותר לפי מצב העששת (1 או 0).

3.3 **<u>נתונים שאינם הגיוניים-</u>**בכל הנתונים שאינם הגיוניים נטפל באותו אופן שטיפלנו בערכים חסרים מלבד מקרים פרטניים שנציין כאן:

GPD , serum creatinine- בעמודות אלו ניתן למצוא רשומה בודדת שערכה 999- ערך לא תקין שמסתמן שמופיע בשל טעות הקלדה. את ערך זה נחליף לפי הטיפול שבחרנו לעמודה (כפי שציינו בערכים חסרים).

Systolic- ערך אחד שלילי- לא יתכן לחץ דם סיסטולי שלילי. לא מצאנו חריגות נוספות בעמודה (מלבד ערכים קיצוניים שיטופלו בהמשך)וכיוון ששאר הערכים בטווח תקין, נעריך שבוצעה טעות הזנה ונהפוך ערך זה לחיובי.

LDL/HDL/ Triglyceride - משום שהם מרכיבים את סה"כ כולסטרול, ובבדיקת הקשרים בין המאפיינים נמצאה נוסחה מדויקת לחישוב הכולסטרול, לא יתכן שערכם יהיה גדול מערך הכולסטרול /הנוסחה לא תתקיים. במקרים בהם הנוסחה מופרת, נשנה אותם לפי הנוסחה בהתאם לשאר המשתנים ברשומה. (בדקנו אם ישנן רשומות בהן חסר יותר מערך אחד מבין אלו ולא נמצאו - במידה והיו כאלו היינו מוותרים על הרשומה).

waist(cm), Oral, Urine protein - יש ערך מילולי אחד בכל עמודה - פגם בהזנת הנתונים.

eyesight(right/left)– ערכים שאינם הגיוניים בעלי ערך 9.9 – הערכים נורמלו סביב הערך 1.

4. <u>הכנת הנתונים</u>

4.1 <u>השמטת מאפיינים "רועשים" או חסרי חשיבות</u> -

oral -בחרנו להשמיט מאפיין זה שכן הערך היחיד שמוזן אליו הוא קבוע Y ולכן לא נוכל ללמוד ממנו.

ID- נשמיט מאפיין זה שכן אינו מדד ביולוגי ולא נוכל ללמוד ממנו דבר.

4.2 <u>השמטת מאפיינים בעלי איכות נמוכה</u> -

4.2.1 <u>חוסרים:</u>

- הסרנו רשומות עם מספר רב של חוסרים כפי שתואר בסעיף חוסרים.

- במקרה בו בהשלמות החוסרים התבססנו על עמודה אחרת ומצאנו שגם הערך בעמודה הנוספת היה חסר, בחרנו להשמיט לגמרי את הרשומה שכן לא נרצה לכפות על הנתונים ערכים שאינם מבוססים, וכיוון שכמות הרשומות שברשותנו גדולה מאוד(כ-44000), ויתור על מספר בודד של רשומות תהיה עדיפה במקרים אלו.

4.2.2 שגיאות

- עבור הערכים השגויים (שתוארו בסעיף משתנים שאינם הגיוניים) שנבעו כתוצאה מהזנות שגויות של הנתונים (לא ערכים קיצוניים שיתכן שנצפו), בחרנו ולהחליפם באותו אופן שבחרנו למלא ערכים חסרים כפי שתואר בסעיף החוסרים. זאת מכיוון שלא נרצה שערכים שגויים ולא תקינים ישפיעו על מרחב הערכים או לא יאפשרו לבצע תחשיבים מתאימים על הנתונים (כמו חישוב מתמטי שלא ניתן לבצע על רשומה שגויה עם ערך מילולי).
- עבור המשתנים הבאים ביצענו טיפול בערכים חריגים שהגדרנו כחצי אחוז עליון של הערכים במדגם – **Gtp , systolic , fasting blood sugar , HDL ,LDL , triglyceride**. הטיפול בחריגים אלו התבצע ע"י נרמול הערכים לאחוזון ה99.5 כלומר הערך שרק 0.5% מהערכים גדולים ממנו. את כל הערכים שגדולים מערך זה החלפנו לערך זה. האידיאל שעמד מאחורי החלטה זו היה לאפשר טיפול ברעש קיצוני עבור ערכים חריגים שהוגדרו כחריגים גם ביחס למדגם וגם לפי הנורמה בספרות. עבור משתנים שההתפלגות שלהם הראתה זנב ארוך משמעותית כאשר הערכים בזנב היו בעלי צפיפות נמוכה מאוד כלומר הם משנים את ההתפלגות רק בשל המצאות מספר קטן מאוד של חריגים קיצוניים שאינם הגיוניים מבחינת ערכים שהמשתנים יכולים לקבל, בחרנו לבצע נרמול לערכים אלו ולהחליפם באחוזון ה-99.5.

4.3 טיפול פרטני במאפיינים

4.3.1 גזירת מאפיינים חדשים

BMI -נבחר לגזור משתנה חדש של מדד מסת הגוף שיהיה מלמד יותר מגובה או משקל כשלעצמם ונוכל ללמוד ממנו על נבדקים הנמצאים בעודף משקל/תת משקל או נורמה. BMI יחושב על פי: $\mathbf{weight(kg)/height(m)^2}$

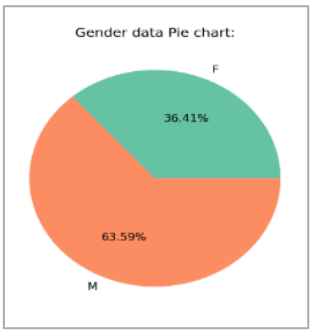
4.3.2 דיסקרטיזציה-

- עבור המשתנים הרציפים **weight(kg),height(cm) –** ההיסטוגרמות (**נספח-** היסטוגרמת משקל/ גובה) שלהם הראו התפלגות נורמאלית, ולאור העובדה שכבר יש ייצוג לערכיהם במשתנה BMI שגזרנו מהם וכן אנו סוברים כי הבדלים קטנים במשתנים אלו לא עשויים ללמד אותנו רבות על משתנה המטרה, לכן בחרנו לבצע דיסקרטיזציה שמייצגת את ההתפלגות לפי רבעונים בכדי להקל על הסיווג בהמשך.
- לאחר חקר על משתנים **ALT** ו-**AST (נספח-** היסטוגרמת ALT/AST)הבנו כי עבור הערכים 0-25 המדד תקין אך לא נמצא בנורמה, טווח הנורמה הנו 25-30, טווח הנורמה עד פי 3 (עד 90) משמעותו מקרים חריגים הדורשים הרחבת בירור ומעל פי 3 (מעל 90) זו כבר הפרעה רצינית. משתנים אלו מכילים ערכים חריגים מאוד (AST מגיע עד 1090 ו-ALT-עד 2914) היוצרים רעש אותו נרצה להקטין ע"י הורדת ההשפעה של החריגים הללו. לאחר בדיקת הנתונים החלטנו שהנתונים תואמים את הנאמר בספרות ובנוסף ניתן לראות כי עבור ערכים גדולים מ-40 אין הבדל בהתפלגות משתנה המטרה (בין מעשנים/לא מעשנים) (נספח 8) וכן תדירות הערכים האלו מאוד קטנה ולכן נוכל להכליל את הערכים הקיצוניים תחת אותה הקבוצה. בשל כל אלו, נבצע דיסקרטיזציה לפי התחומים [0-25],[25-30],[30-90],[90-max] .

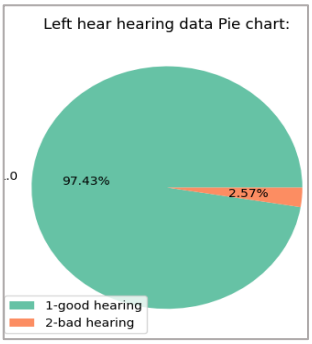
5. נספחים

נספח 1 הסתברויות אפריוריות

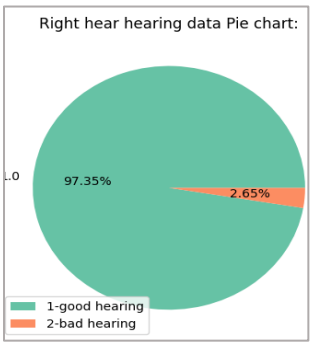
נספח 1.1



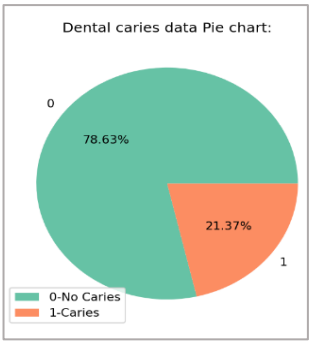
נספח 1.2



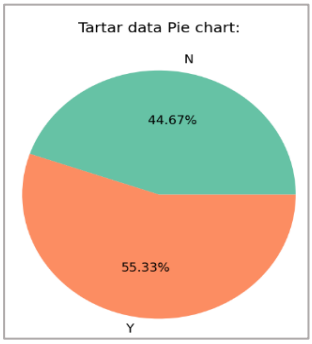
נספח 1.3



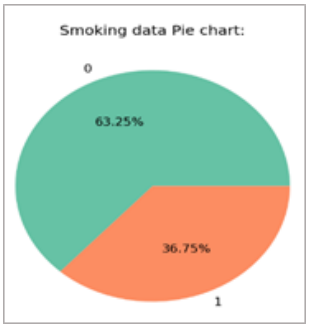
נספח 1.4



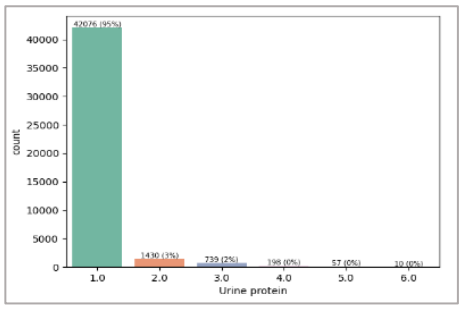
נספח 1.5



נספח 1.6

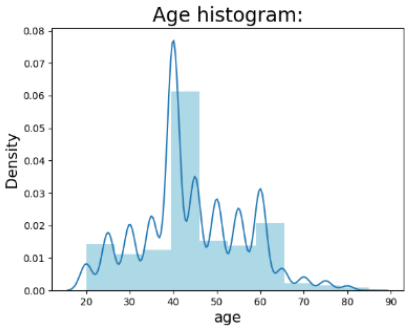


נספח 1.7

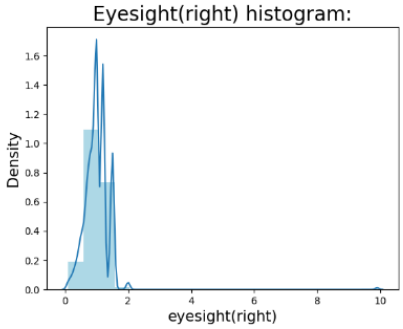


1.0	0.944403
2.0	0.032097
3.0	0.016587
4.0	0.004444
5.0	0.001279
6.0	0.000224

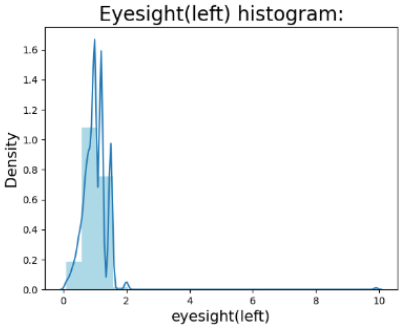
נספח 1.8



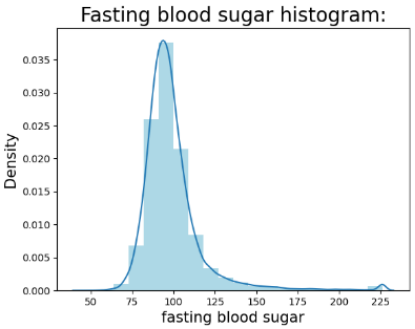
נספח 1.9



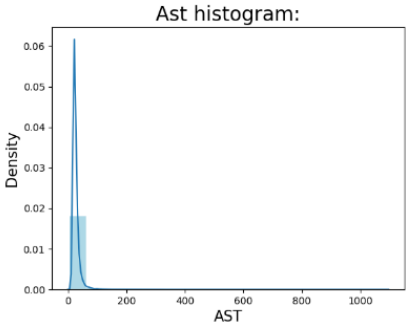
נספח 1.10



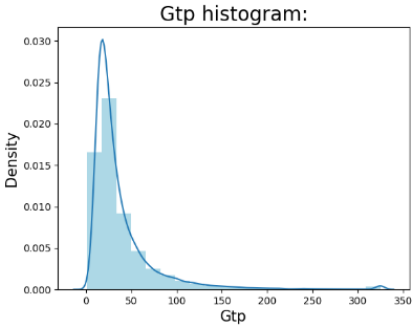
נספח 1.11



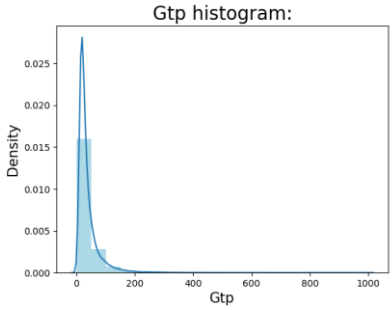
נספח 1.12



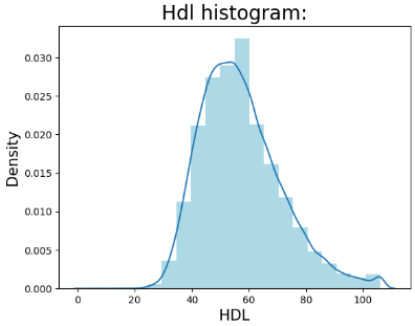
נספח 1.13



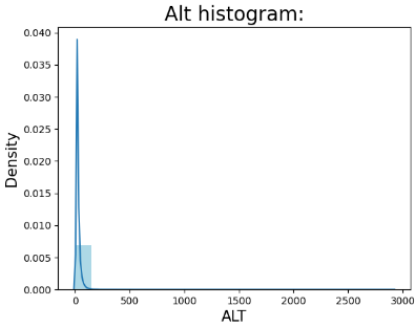
נספח 1.13.1



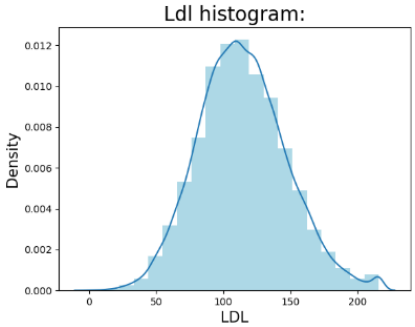
נספח 1.15



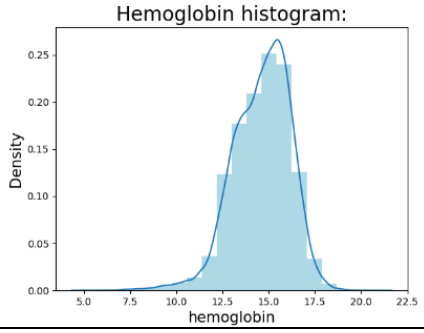
נספח 1.14



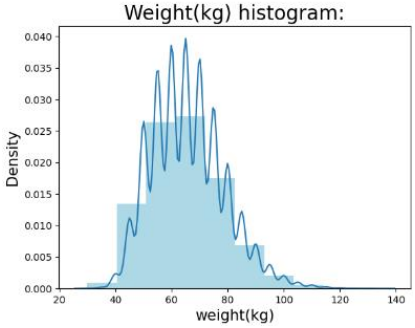
נספח 1.17



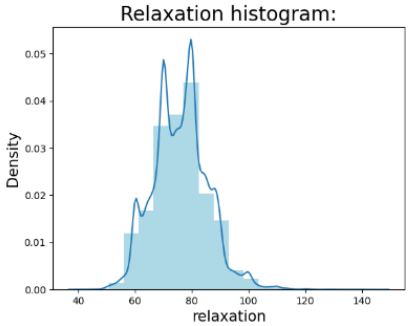
נספח 1.16



נספח 1.19

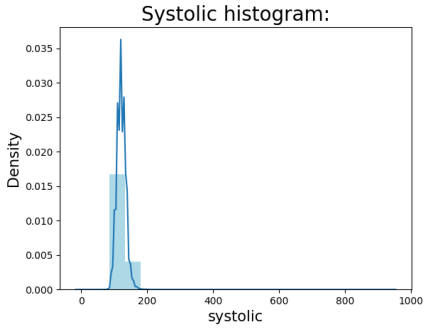


נספח 1.18



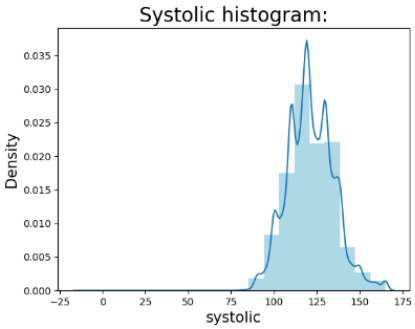
נספח 1.21.1

before handling

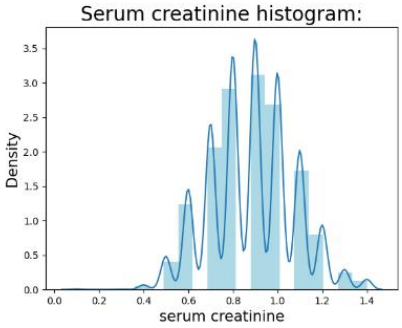


נספח 1.21.2

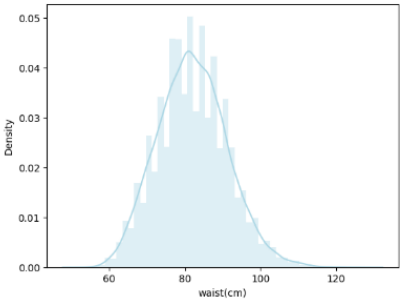
after handling



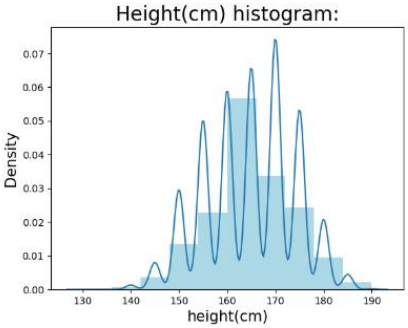
נספח 1.20



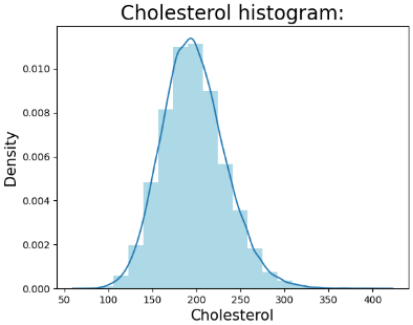
נספח 1.23



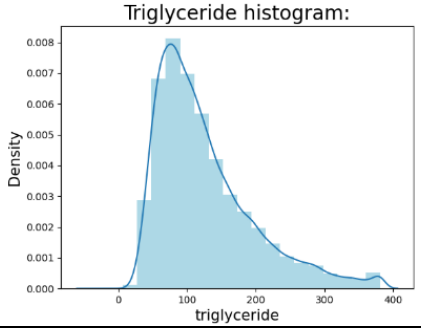
נספח 1.22



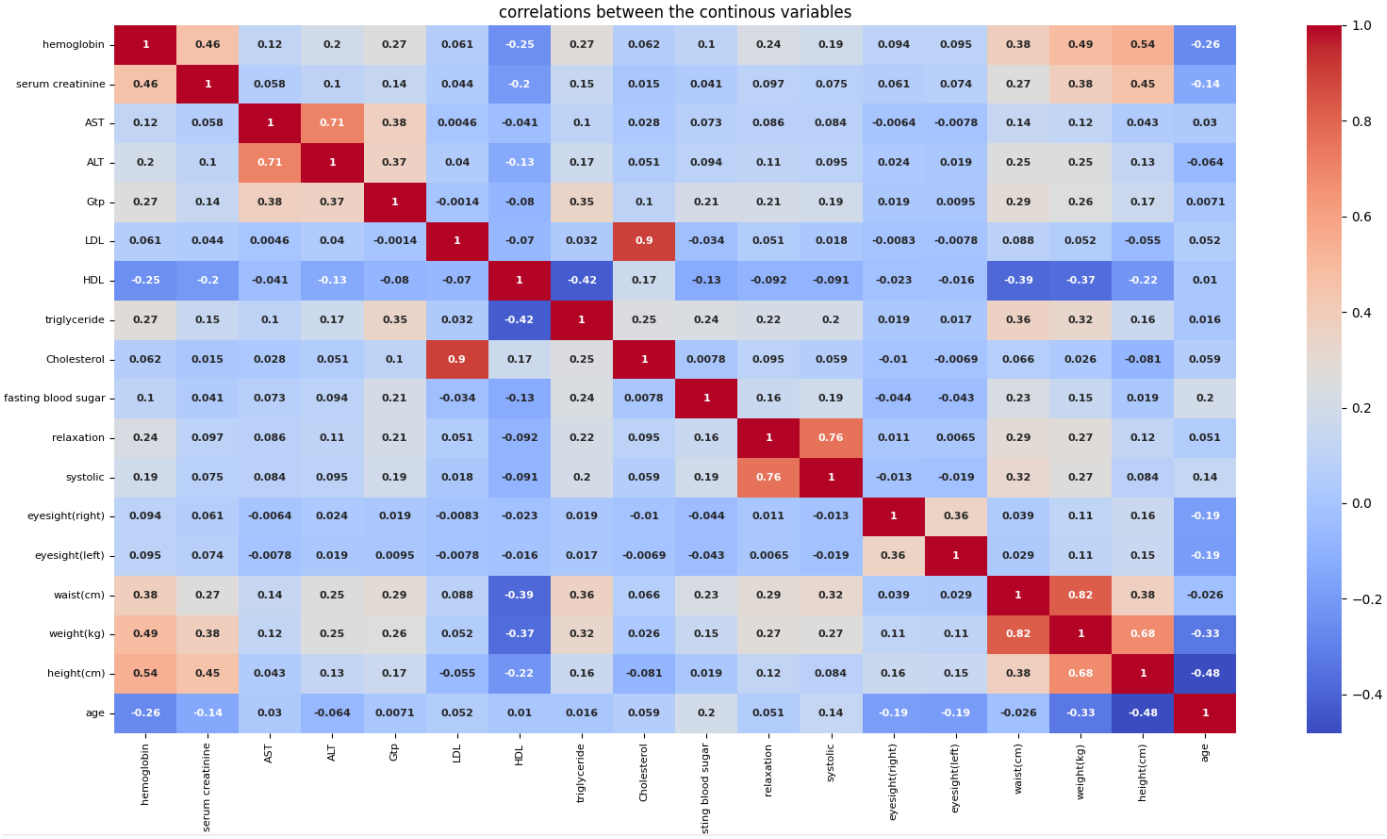
נספח 1.25



נספח 1.24

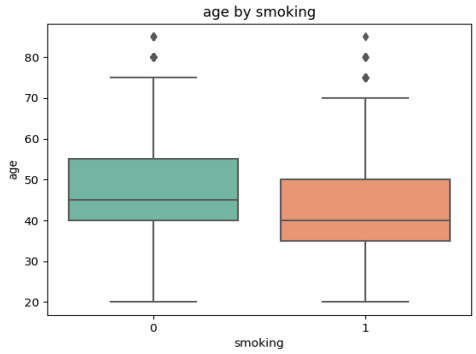


נספח 2 תרשים קורלציות עבור משתנים רציפים.

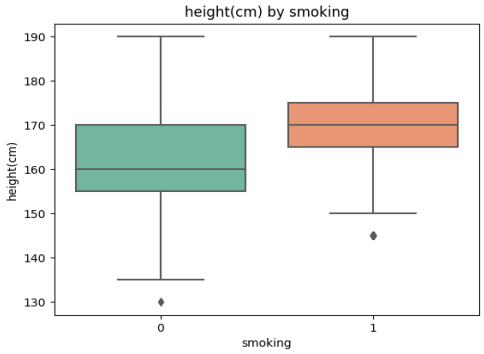


נספח 3 תרשימי Box-Plot עבור משתנים רציפים עם עישון

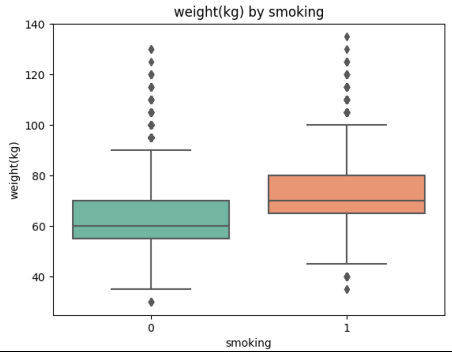
נספח 3.1



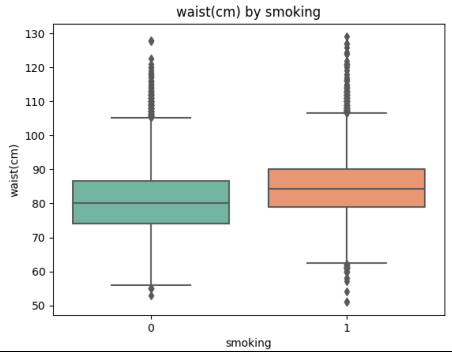
נספח 3.2



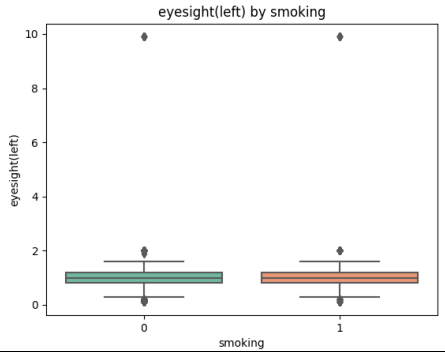
נספח 3.3



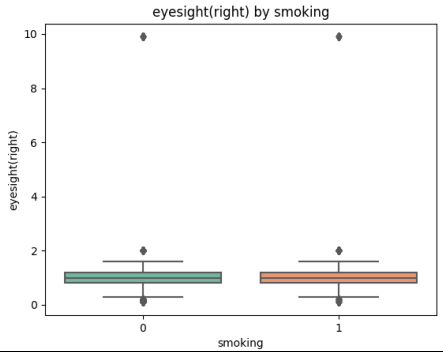
נספח 3.4



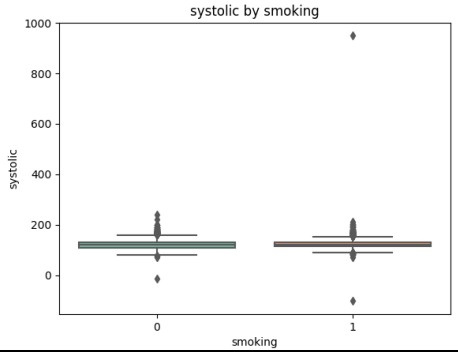
נספח 3.5



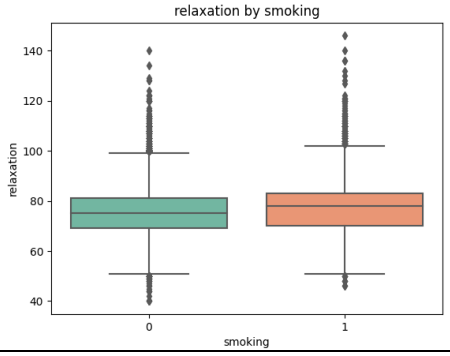
נספח 3.6



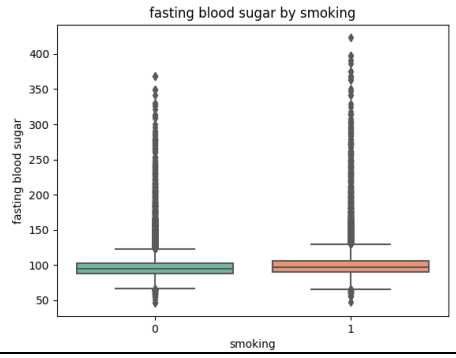
נספח 3.7



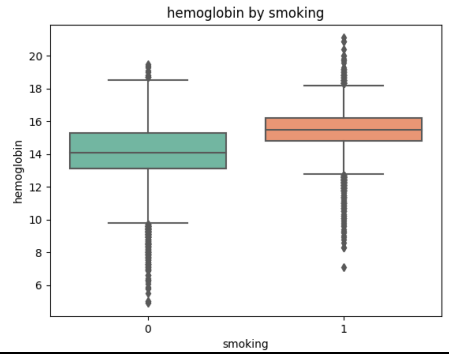
נספח 3.8



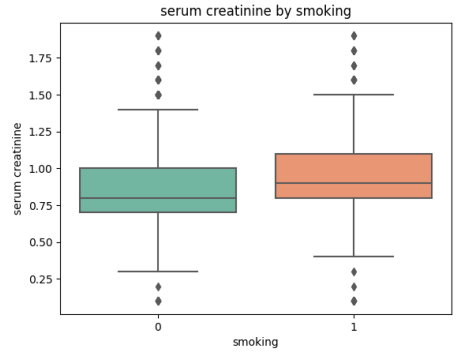
נספח 3.9



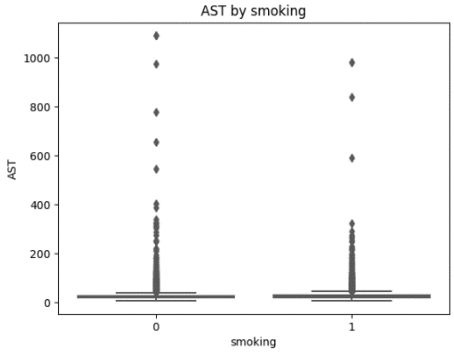
נספח 3.10



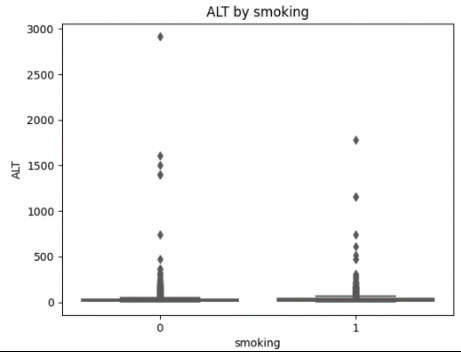
נספח 3.11



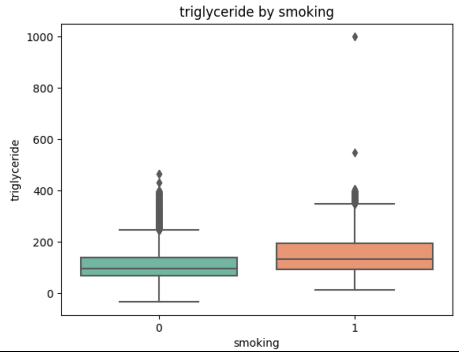
נספח 3.12



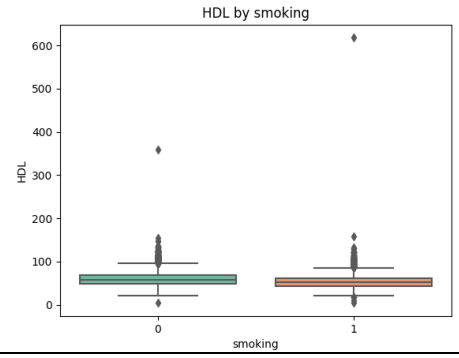
נספח 3.13



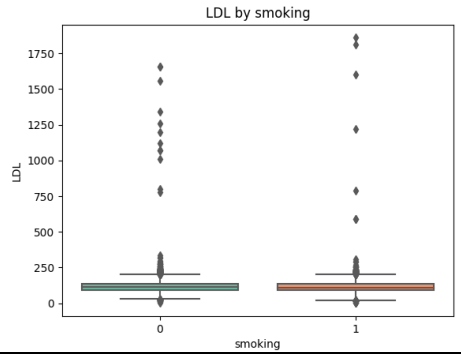
נספח 3.14



נספח 3.15

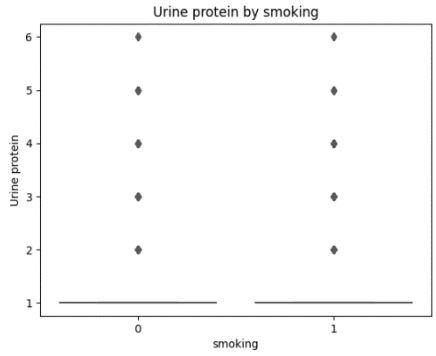


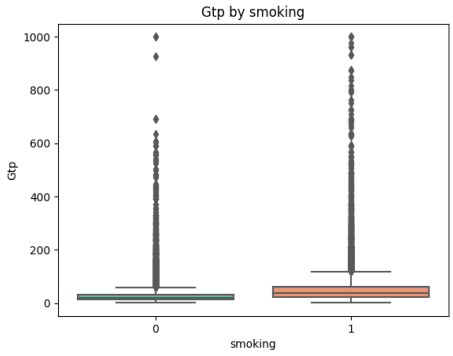
נספח 3.16



נספח 3.17

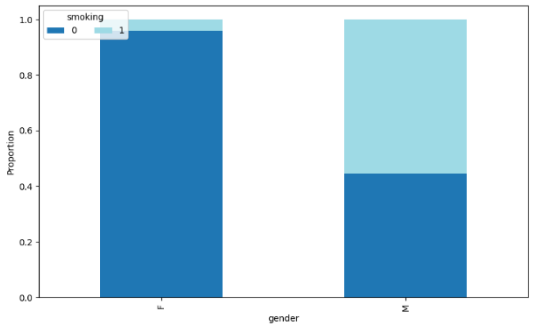
נספח 3.18



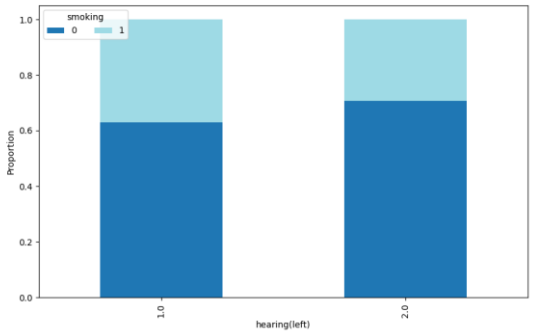


נספח 4 תרשימי עמודות מוערמים עבור משתנים בדידים עם עישון

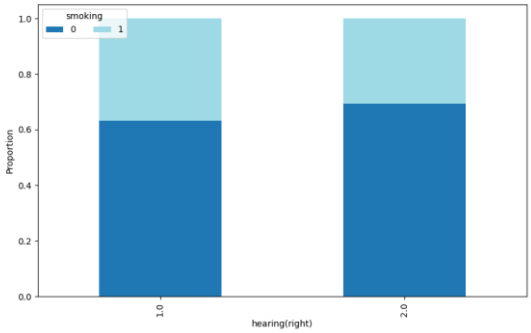
נספח 4.1



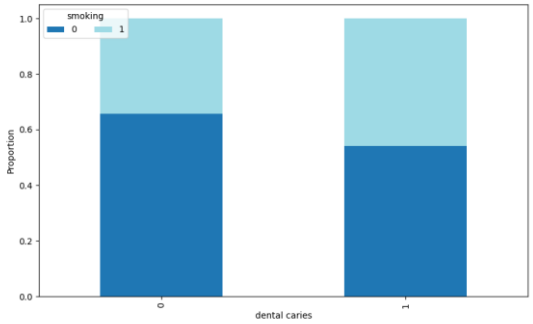
נספח 4.2



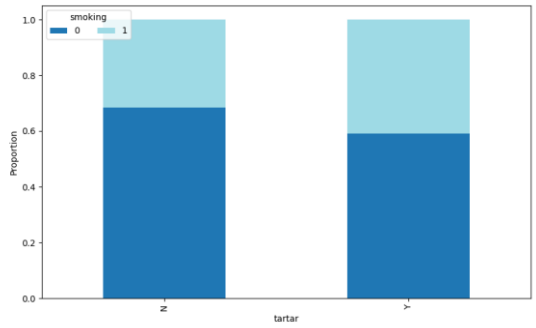
נספח 4.3



נספח 4.4



נספח 4.5



נספח 5

עמודה ראשונה- אינדקס השורה

עמודה שנייה- כמות הערכים החסרים

ID	0
gender	0
age	0
height(cm)	40
weight(kg)	1
waist(cm)	1
eyesight(left)	1
eyesight(right)	5
hearing(left)	0
hearing(right)	8
systolic	11
relaxation	4
fasting blood sugar	2
Cholesterol	6
triglyceride	16
HDL	3
LDL	2
hemoglobin	6
Urine protein	5
serum creatinine	1
AST	1
ALT	20
Gtp	0
oral	2
dental caries	0
tartar	1
smoking	0

נספח 6 סך העמודות הריקות לאחר ההסרה הראשונית לפי אינדקס השורה

16756	19
35122	19
16385	19
44248	16
44249	16
44538	9
44539	9
44536	9
44537	9

נספח 7- בדיקת רגרסיה לינארית עבור המשתנים כולסטרול HDL, LDL וטריגלצרין

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           Cholesterol   R-squared (uncentered):              1.000
Model:                           OLS   Adj. R-squared (uncentered):         1.000
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
LDL            0.9979      0.000   2352.719      0.000       0.997       0.999
HDL            1.0077      0.001   1322.007      0.000       1.006       1.009
triglyceride   0.1965      0.000    910.172      0.000       0.196       0.197
intercept           0          0        nan        nan           0           0
==============================================================================
```

נספח 8 –

היסטוגרמה תוך התמקדות בערכים היותר שכיחים ללא הצגת הזנב הימני הארוך שמגיע עד כ-1400