

לימוד מכונה | קבוצה 2 | פרויקט חלק א'

Smoking

רועי עזראי 206118754

ליטל זולטריוב 208466524



תוכן עניינים :

1.	הגדרת הבעיה:	3
3.	תיאור כללי של עולם התוכן הנחקר :	3
3.	הגדרת שאלת המחקר:	3
2.	הבנת הנתונים:	3
3.	תיעוד מקורות הנתונים ומשמעותם:	3
4.	הסתברויות אפריוריות וקשרים בין מאפיינים:	4
8.	קשרים צפויים:	8
9.	קשרים לא צפויים:	9
9.	קשרים מול משתנה המטרה:	9
3.	איכות הנתונים	10
4.	נספחים:	12
12.	הסברים בנוגע למשתנים	12
14.	טבלת קורלציה HeatMap	14
16.	גרפים נוספים	16
19.	ביבליוגרפיה:	19

1. הגדרת הבעיה:

תיאור כללי של עולם התוכן הנחקר :

הנתונים בהם נשתמש בפרויקט עוסקים במידע על נתונים אישיים ואותות ביולוגיים בסיסיים אודות הנבדקים. הנתונים מכילים מספר מאפיינים אשר מספקים מידע על המצב הבריאותי של הנבדק. המטרה היא לסווג בין הנבדקים הללו על סמך המאפיינים ולחזות אותם ובסופו של דבר לקבוע את נוכחותו או היעדרו של עישון באמצעות סיגנלים ביולוגיים (האם הנבדק מעשן או לא): (1 -קיים נוכחות של עישון באותות הביולוגיים,0 -אחרת).

הגדרת שאלת המחקר:

בעזרת הנתונים אנו מצפים לבנות מודל שישאף לחזות את נוכחותו או היעדרו של עישון על אותות ביולוגיים של הנבדקים, בהינתן נתונים ביולוגיים ובריאותיים של אותם הנבדקים.

2. הבנת הנתונים:

תיעוד מקורות הנתונים ומשמעותם:

מקור הנתונים בו אנו משתמשים הוא ממערכת נתוני רפואיים אודות נבדקים במערכת הבריאות, סוגי הבדיקות ממקורות שונים, מדידות מסוגים שונים, בדיקות דם, ובדיקות שגרתיות.

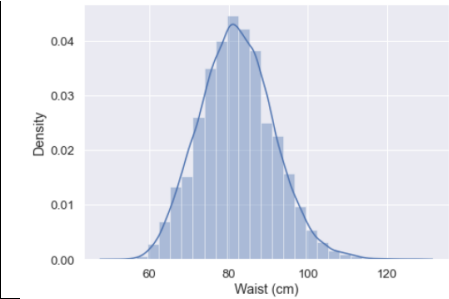
טבלת משתנים - *הסבר מפורט עבור כל משתנה נמצא בנספחים

מספר	משתנה	סוג	טווח ערכים	מקור
1	ID	רציף	מספר סידורי ייחודי	תעודת זהות הנבדק
2	Gender	קטגוריאלי	F, M	תעודת זהות הנבדק
3	age	קטגוריאלי	(20-30,30-40,40-55,55-70,70+)	תעודת זהות הנבדק
4	Height (cm)	קטגוריאלי	ס"מ (130-190)	מדידות רפואיות אצל מומחה
5	Weight (kg)	קטגוריאלי	מתחת ל40, 41-54, 55-70, 71-100, מעל 100)	מדידות רפואיות אצל מומחה
6	Waist (cm)	רציף	ס"מ (50-129)	מדידות רפואיות אצל מומחה
7	Eyesight (left)	רציף	(0-2)	בדיקת ראייה אצל מומחה
8	Eyesight (right)	רציף	(0-2)	בדיקת ראייה אצל מומחה
9	Hearing (left)	קטגוריאלי	(0,1)	בדיקת שמיעה אצל מומחה
10	Hearing (right)	קטגוריאלי	(0,1)	בדיקת שמיעה אצל מומחה
11	Systolic	רציף	(70-240)	בדיקת לחץ דם אצל מומחה
12	Relaxation (Diastolic)	רציף	(40-150)	בדיקת לחץ דם אצל מומחה
13	Fasting Blood Sugar	רציף	(40-200)	מדידת סוכר בדם בבדיקת דם בצום
14	Cholesterol	רציף	(70-410)	בדיקת דם
15	Triglyceride	רציף	(0-500)	בדיקת דם
16	HDL	רציף	(0-160)	בדיקת דם
17	LDL	רציף	(0-300)	בדיקת דם
18	Hemoglobin	רציף	(0-22)	בדיקת דם
19	Urine protein	קטגוריאלי	(1,2,3,4,5,6)	בדיקת שתן

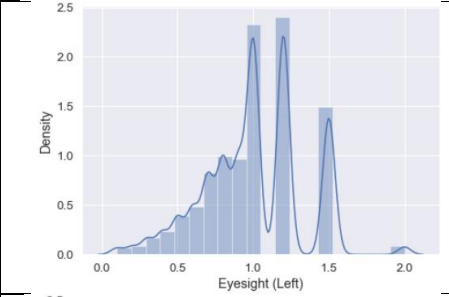
20	Serum Creatinine	קטגוריאלי	(Under normal range, normal, above normal range)	בדיקת דם
21	AST	רציף	(0-100)	בדיקת דם
22	ALT	רציף	(0-100)	בדיקת דם
23	GTP	רציף	(0-150)	בדיקת דם
24	Oral	קטגוריאלי	(Y,N)	בדיקת מומחה של חלל הפה
25	Dental Caries	קטגוריאלי	(0,1)	בדיקת מומחה עבור עששת
26	Tartar	קטגוריאלי	(0,1)	בדיקת צילום שיניים
27	Smoking	קטגוריאלי	(0,1)	

הסתברויות אפריוריות וקשרים בין מאפיינים:

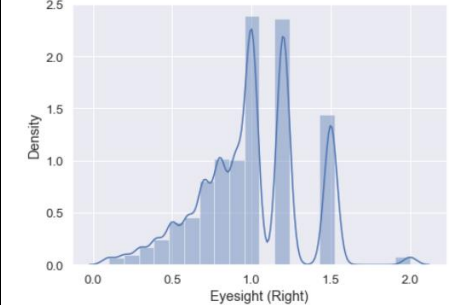
משתנה	ויזואליזציה
gender - נראה כי מרבית הנבדקים הינם גברים. המשתנה יחסית מאוזן. ממידע שחקרנו גילינו ששיעור העישון בקרב גברים גבוה משיעור העישון בקרב נשים.	
age - ניתן לראות שמרבית הנבדקים גילם נע בין 30-40, ולאחר מכן 40-55. הנתונים יחסית מאוזנים, ממידע שמצאנו, שיעור המעשנים בגילאי 20-49 הוא הגבוה ביותר, כאשר הוא יורד מתחת לגיל 20 ומעל לגיל 50. מגילאי 65 ומעלה השיעור הוא הנמוך ביותר. # הסתברות אפריורית לפני דיסקרטיזציה בנספחים.	
Height(cm) - ניתן לראות שגובהם של מרבית הנבדקים נע בין 155-175 ס"מ. נתון זה יחסית מאוזן, שכן קיים רוב יחסי לגברים בקרב הנבדקים והגובה הממוצע אצל גברים הוא סביב 165 ס"מ ואצל נשים 160 ס"מ.	
Weight(kg) - ניתן לראות שמרבית הנבדקים משקלם נע בטווח בין 40-70. הנתונים יחסית מאוזנים, משקל גוף מושפע מאורח חיים, גנטיקה, נתונים ביולוגיים ועוד. משקל תקין מוגדר בטווח מסוים של ערכים בהינתן מאפיינים. ממידע שחקרנו גילינו שעישון לעיתים מסייע לאנשים לשמור על משקל גוף תקין. היינו מצפים שההסתברות המשקל תהיה סביב הטווח הנ"ל בעקבות גורמים רבים המשפיעים על המשקל. # הסתברות אפריורית לפני דיסקרטיזציה בנספחים.	



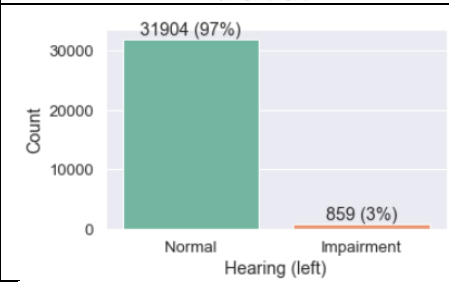
Waist - התפלגות בהיקף מותניים נראית יחסית נורמלית, מעט יותר זנב ימני. הדבר מתיישב עם המציאות, שכן היקף המותניים מושפע מגורמים רבים, גנטיים, פיזיולוגיים וכאלו שתלויים באורח החיים. היקף המותנים מושפע מאחוז שומן בטני ומתקשר לעודף משקל. לא ניתן להסיק שהיקף המותניים מתקשר באופן ישיר לעישון.



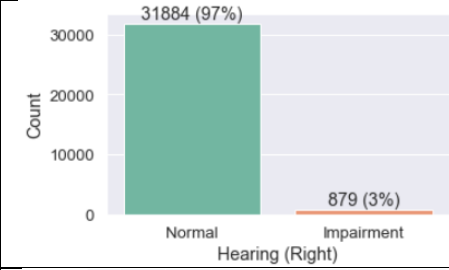
eyesight(left) - ממוצע: 0.99855, חציון- 1.0, סטית תקן: 0.326. הערכים מנורמלים סביב ערך ה-1 כאשר ערך 1 מתקבל עבור בדיקת ראייה תקינה. עישון יכול להוביל לראייה מעורפלת ולטשטוש צבעים.



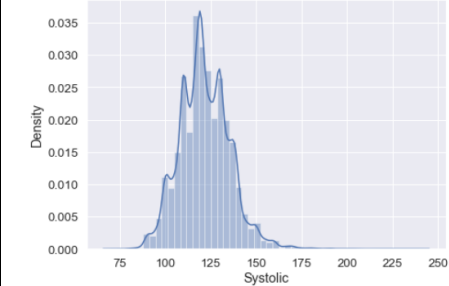
Eyesight(right) - ממוצע: 0.99392, חציון- 1.0, סטית תקן: 0.3249. הסבר דומה בעין שמאל.



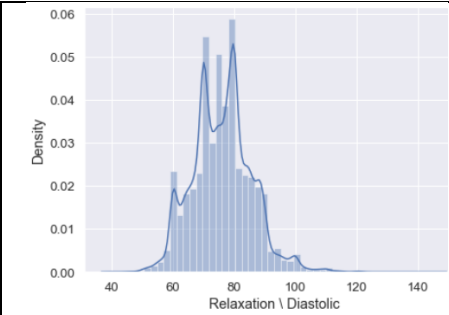
Hearing(left) - מהגרף נראה שרוב המוחלט מהנבדקים שומעים תקין באוזן שמאל. ממידע שמצאנו גילינו שבקרב מעשנים קיים סיכוי גבוה יותר באופן משמעותי לירידת שמיעה בתדרים גבוהים. היינו מצפים שההסתברות הלוקים בשמיעה באוזן שמאל יהיה גבוה יותר. לקות שמיעה יכולה להיות מדורגת לפי חומרת הלקות. יתכן במידה ומוגדרת לקות שמיעה בנתונים שלנו, הלקות תעיד על לקות שמיעה חמורה.



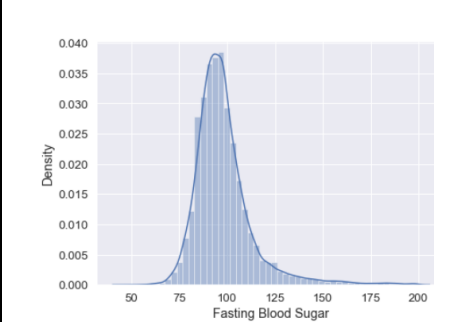
Hearing(right) - מהגרף נראה שרוב המוחלט מהנבדקים שומעים תקין באוזן ימין. הסבר זהה לאוזן שמאל.



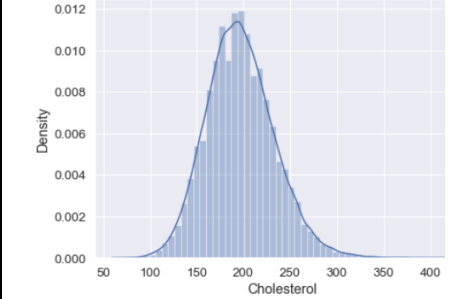
Systolic - ממוצע: 121.510, חציון: 120, סטיית תקן: 13.668. ההתפלגות נראית יחסית נורמלית, עם זנב ימני. דבר המתיישב עם המציאות. ממידע שחקרנו גילינו שעישון מביא באופן ישיר לעלייה בלחץ הדם בגוף וכאשר מדובר בעישון תדיר וקבוע, לאורך זמן, לחץ הדם של אותו אדם יישאר גבוה כל העת אשר ישפיע על לחץ דם הסיסטולי וכן דיאסטולי.



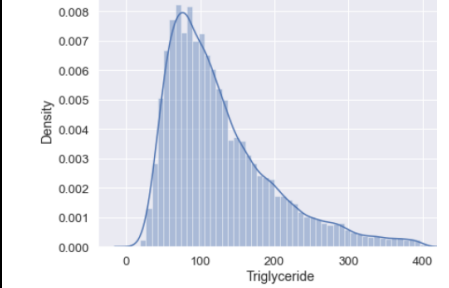
Relaxation - ממוצע: 76.0077, חציון: 76, סטיית תקן: 9.680. ההתפלגות נראית יחסית נורמלית, עם זנב ימני. דבר המתיישב עם המציאות. שיקולים זהים ללחץ דם סיסטולי. כמו כן, הלחץ דם מושפע ממאפיינים רבים כגון תדירות העישון, ועוד, לכן נצפה שבסה"כ הנתונים יהיו מאוזנים עם נטייה להיות גבוהים מהנורמה.



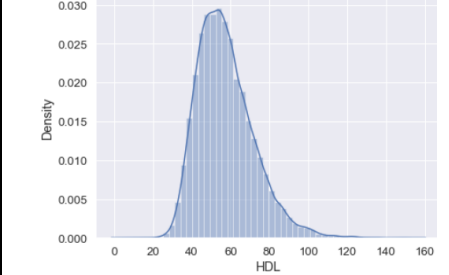
Fasting Blood Sugar - ממוצע: 98.149, חציון: 96, סטיית תקן: 15.682. ההתפלגות נראית יחסית נורמלית, עם זנב ימני. דבר המתיישב עם המציאות, רמות גבוהות מהנורמה יכולות להעיד על סוכרת. הנתונים מאוזנים, ממידע שחקרנו גילינו כי העישון מעלה את רמת הסוכר (גלוקוז) בדם, ועלול לפגוע ברגישות לאינסולין. בקרב חולי סכרת ההשפעה משמעותית אף יותר.



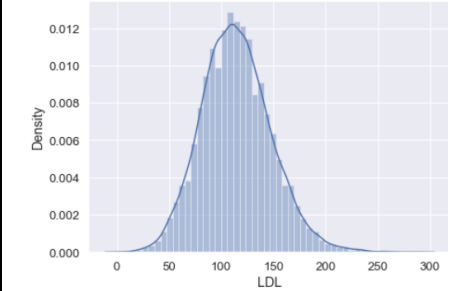
Cholesterol - ממוצע: 196.746, חציון: 195, סטיית תקן: 36.162. ההתפלגות נראית יחסית נורמלית, עם זנב ימני. דבר המתיישב עם המציאות. ממידע שחקרנו גילינו כי העישון מוריד את רמות הכולסטרול הטוב ומעלה את רמות LDL, לכן, הוא מכביד ביותר על הגוף. בנוסף מחקרים מראים כי השילוב בין כולסטרול גבוה לבין עישון הוא גרוע במיוחד.



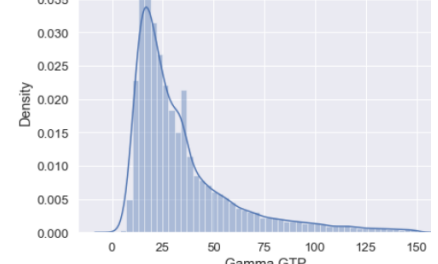
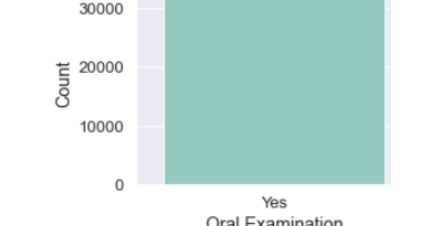
Triglyceride - ממוצע: 126.518, חציון: 108, סטיית תקן: 71.515. ההתפלגות נראית יחסית נורמלית, עם זנב ימני ארוך אשר פוחת בהדרגה. דבר המתיישב עם המציאות. הנתונים מאוזנים, ממידע שחקרנו גילינו כי עישון ממושך הוא אחת הסיבות לעלייה של רמת הטריגליצרידים.



HDL - ממוצע: 57.261, חציון: 55, סטיית תקן: 14.469. ההתפלגות נראית יחסית נורמלית, עם זנב ימני. דבר המתיישב עם המציאות. היינו מצפים שהזנב יהיה זנב שמאלי, שכן עישון מוריד את הכולסטרול "הטוב", אך הוא מושפע ממספר רב של גורמים גנטיים, תזונתיים ועוד ולכן הנתונים יחסית מאוזנים.



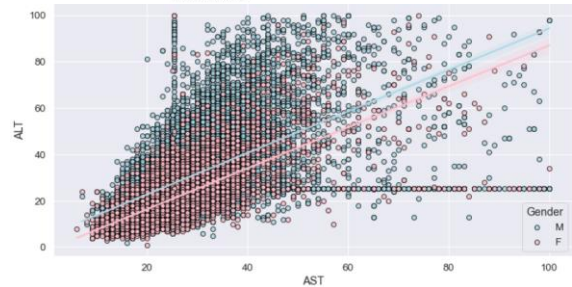
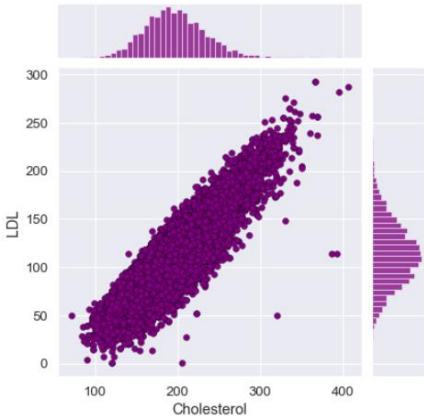
LDL - ממוצע: 114.361, חציון: 113, סטיית תקן: 33.286. ההתפלגות נראית יחסית נורמלית. דבר המתיישב עם המציאות, הנתונים מאוזנים, עישון גורר עלייה של הכולסטרול הרע בגוף.

	<p>Hemoglobin - ממוצע: 14.620, חציון: 14.8, סטיית תקן: 1.568 ההתפלגות נראית יחסית נורמלית. דבר המתיישב עם המציאות, קיימים מצבים בהם אצל מעשנים כבדים נראה רמות גבוהות מהרגיל של המוגלובין בדם, שגופם מייצר יותר המוגלובין כדי לפצות על הירידה ברמת החמצן ברקמות שנגרם כתוצאה מהעישון. לכן היינו מצפים לזנב ימני ארוך יותר, אך המוגלובין מושפע בגורמים נוספים ובעיקר מהתזונה של הנבדק ולכן הנתונים יחסית מאוזנים.</p>
	<p>Urine Protein - מהגרף נראה כי למרבית הנבדקים לא קיים חלבון בשתן. לאחר חיפוש מידע בנוגע לקשר בין חלבון בשתן ועישון, נראה כי עישון לא משפיע על רמת החלבון בשתן ולכן היינו מצפים שמספר הנבדקים אשר בנתוניהם קיים חלבון בשתן, יהיה דומה לשכיחות באוכלוסייה, הנתונים מאוזנים שכן כמויות חלבון שמופרשות בשתן לאורך זמן יעידו על סיכון מוגבר ליתר לחץ דם ומחלות כליה.</p>
	<p>Serum Creatinine - מהגרף נראה כי מרבית הנבדקים הערך נמצא בטווח התקין. ממידע שחקרנו לא נמצאה השפעה משמעותית של עישון על רמת קריאנין, לכן הנתונים מאוזנים, שכן רמה גבוהה של קריאנין תעיד על תפקוד לא תקין של הכליה. # הסתברות אפרורית לפני דיסקרטיזציה בנספחים.</p>
	<p>AST - ממוצע: 25.309, חציון: 23, סטיית תקן: 10.223 ההתפלגות נראית יחסית נורמלית עם זנב ימני ארוך. דבר המתיישב עם המציאות. הנתונים יחסית מאוזנים בדומה לאנזים ALT רמתו תעיד על חשד למחלות כבד מסיבות שונות.</p>
	<p>ALT - ממוצע: 25.170, חציון: 21, סטיית תקן: 14.981 ההתפלגות נראית יחסית נורמלית עם זנב ימני ארוך. דבר המתיישב עם המציאות, מבדיקה של אנזים כבד זה ניתן להסיק בין היתר על דלקות בכבד ובדרכי המרה והיא עלולה להיגרם מסיבות שונות. הנתונים מאוזנים, שכן היינו מצפים שרמת האנזים תשתנה בין הנבדקים מסיבות שונות ומגוונות אך תישאר סביב הערך התקין.</p>
	<p>GTP - ממוצע: 33.859, חציון: 25, סטיית תקן: 25.242 ההתפלגות נראית יחסית נורמלית עם זנב ימני ארוך. דבר המתיישב עם המציאות. הנתונים יחסית מאוזנים, בדומה לאנזים ALT אנזים שמצוי בעיקר בכבד, בדרכי המרה ובכליות. האנזים משוחרר לדם כאשר תאים שמכילים אותו נפגעים.</p>
	<p>Oral - מהגרף ניכר שכלל הנבדקים עברו בדיקת חלל הפה. נראה כי הנתונים אינם מאוזנים, שכן יש לדגום נבדקים אשר לא נעשה להם בדיקה של חלל הפה, מאידך, ממידע שחקרנו ניכר כי העישון משפיע על חלל הפה במספר דרכים (חניכיים, לשון,</p>

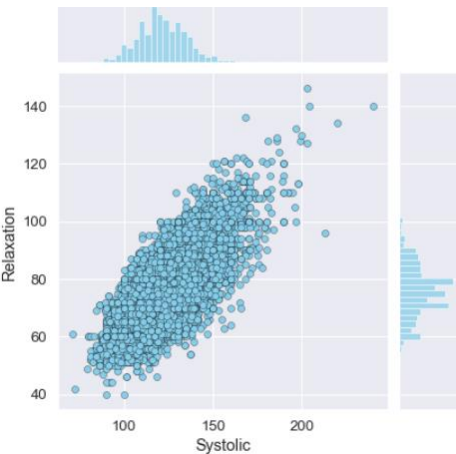
	בלוטות רוק, שפתיים ועוד). לכן נצפה שהנבדקים יעברו בדיקה של חלל הפה לטובת בדיקות שגרתיות נוספות של הפה.
	Dental Caries - מהגרף ניכר שמרבית הנבדקים אינם לוקים בעששת. הדבר מתיישב עם המציאות. עששת נוצרת לרוב מצריכת סוכר בתדירות גבוהה. בנוסף, ממידע שחקרנו נראה כי ניקוטין מגביר את צמיחת החיידקים גורמי העששת בחלל הפה. ניתן למנוע היווצרות עששת במספר דרכים כגון משחות שיניים ושטיפות פה, לכן בסך הכל הנתונים מאוזנים ומייצגים נכונה את המציאות.
	Tartar - מהגרף ניכר כי מרבית הנבדקים לוקים באבנית חלל הפה. הדבר מתיישב עם המציאות שכן, אבנית נגרמת מחוסר היגיינת הפה. הנתונים מאוזנים, נמצא שהצטברות אבנית בשיניים גדולה אצל מעשנים מאשר אצל לא מעשנים.
	Smoking - נראה כי מרבית הנבדקים אינם מעשנים, הגרף אינו משקף את המציאות באופן מדויק, שכן שכיחות המעשנים עומדת על אזור ה-20-30 אחוז מקרב האוכלוסייה. יתכן והנבדקים נלקחו מתוך אוכלוסייה אשר חלק/ מרבית מהמאפיינים הנמדדים אינם תקינים, דבר היכול להעיד על נוכחות של עישון.

קשרים צפויים:

LDL & Cholesterol - צפינו לקבל קשר חיובי בין משתנה LDL, Cholesterol שכן, ערכי הכולסטרול מושפעים מערכי הכולסטרול הטוב HDL, ובפרט מהכולסטרול הרע LDL. ערכי כולסטרול גבוהים מנורמה מעידים על מצב בריאותי שאינו תקין דבר המתיישב עם רמה גבוה של LDL עבורו ערך גבוה מהנורמה מעיד על מצב בריאותי לא תקין. הקורלציה בין המשתנים הינה 0.72, אשר מאשש את טענתנו.

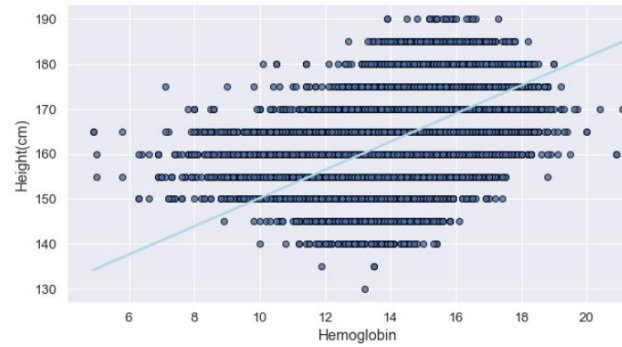


ALT & AST - צפינו לקבל קשר חיובי בין שני המשתנים, שכן שניהם משמשים כאנזימי כבד ותפקידם לאיתור מחלות כבד ומשמשת לעמידת תפקוד הכבד. ממידע שחקרנו גילינו ששני האנזימים תפקיד דומה ומעידים על מצבים דומים בכבד. הקורלציה בין שני המשתנים יחסית גבוהה והינה 0.65.

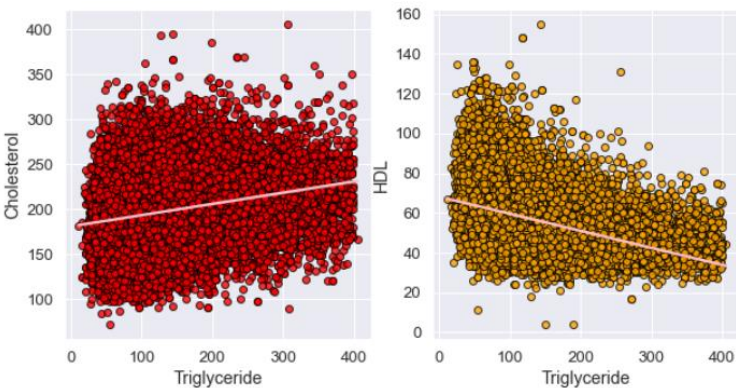


Relaxation & Systolic - עבור המשתנים הללו צפינו לקבל קשר חיובי, שכן שניהם מצביעים על לחץ הדם במצבים שונים של שריר הלב (מצבים עוקבים). הקורלציה בין המשתנים גבוהה באופן יחסי ועומדת על 0.76.

קשרים לא צפויים:

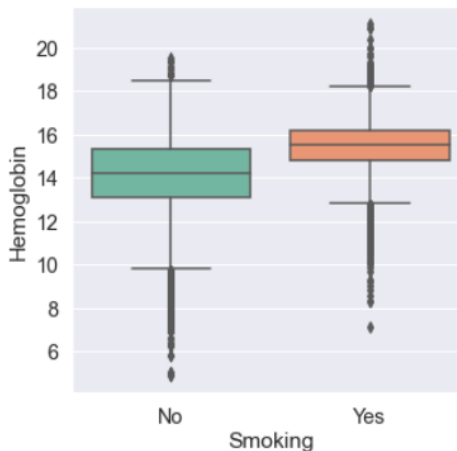


Height(cm), hemoglobin - מהתרשים ניתן לראות קשר חיובי בין הגובה לרמת ההמוגלובין, כמו כן הקורלציה ביניהם עומדת על 0.54. דבר הנראה מפתיע במחשבה ראשונה. מבדיקה אודות מידע על הקשר לא נמצא קשר מובהק בין שני המשתנים. יתכן והקשר נובע בין הבדלי רמות ההמוגלובין בין גברים לנשים: אצל גברים רמת המוגלובין הנחשבת לתקינה גבוהה יותר וכמו כן גובהו הממוצע של גבר גבוה יותר מגובה הממוצע של אישה.

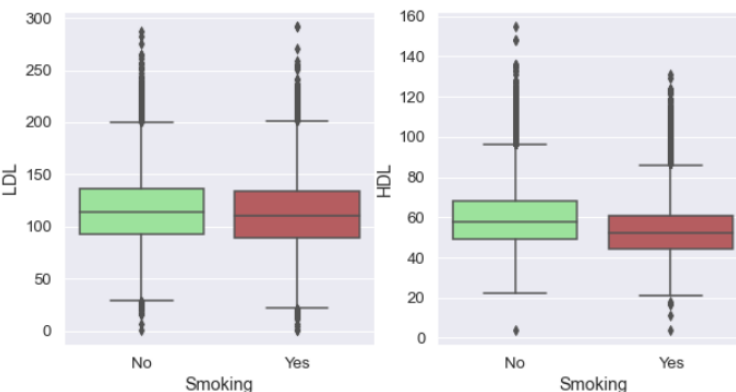


HDL & Triglyceride - מהגרף ניתן לראות קורלציה שלילית בין המשתנים טריגליצרידים וHDL. הדבר מפתיע, כי במחשבה ראשונית, הכולסטרול (אשר מכיל בתוכו את הכולסטרול הטוב HDL) וטריגליצרידים, שניהם מהווים את השומנים בדם. ניכר כי קיים ביניהם קשר חיובי יחסית חלש. ממחקר נוסף גילינו שקיים קשר בעבור רמת טריגליצרידים גבוהה ורמת HDL נמוכה אשר נקראת התסמונת המטבולית, וגוררת התנגדות לאינסולין. יתכן שמצב זה שופך מידע נוסף על הקשר המפתיע. הקורלציה ביניהם הינה -0.41.

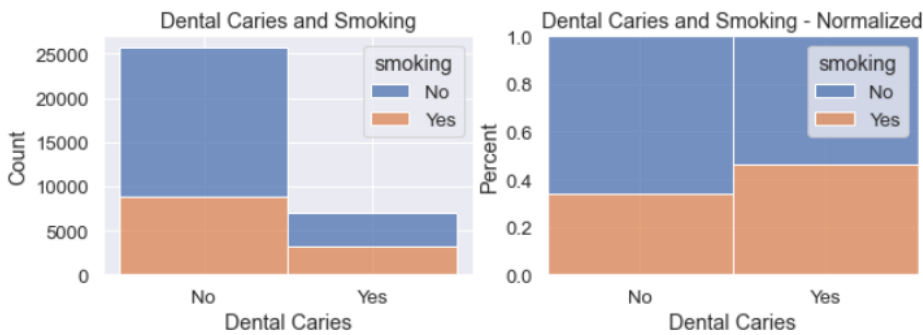
קשרים מול משתנה המטרה:



Hemoglobin & Smoking - מבירור אודות הקשרים בין המשתנים, נראה שרמות המוגלובין גבוהות נמצאו אצל מעשנים כבדים, שגופם מייצר יותר המוגלובין כדי לפצות על הירידה ברמת החמצן ברקמות שנגרם כתוצאה מהעישון. מהנתונים שלנו לא ניתן להעיד האם המעשן הוא מוגדר כמעשן "כבד" אך ניתן לראות קשר בין רמת המוגלובין גבוהה יותר אצל מעשנים בקרב נבדקים שאינם מעשנים. הקורלציה הינה 0.4.



Smoking – HDL & LDL - ממידע שחקרנו גילינו כי עישון מעלה את רמת LDL (הכולסטרול הרע), ומקטין את רמת HDL (הכולסטרול הטוב). מהגרפים ניתן לראות את הרמות השונות של שני סוגי הכולסטרול ביחס למעשנים וללא מעשנים. בנוגע לרמת ה LDL הממצאים מפתיעים ולא ניכר הבדל מהגרף. בנוסף הקורלציה נמוכה.



Dental Caries & Smoking - ממידע

שחקרנו גילינו שקיים קשר בין נבדקים מעשנים לבין נבדקים שלוקים בעששת- ניקוטין מגביר את צמיחת החיידקים גורמי העששת בחלל הפה. מהגרף ניתן לראות כי מבין הלוקים

בעששת, בערך מחציתם מעשנים. ומבין אלו שלא לוקים בעששת מרביתם אינם מעשנים. הדבר מתיישב עם המציאות מכיוון שקיימים גורמים רבים אשר גורמים לעששת כפי שפורטו. בסך הכל ניתן לראות קשר בין המשתנים על אף שחוזק הקשר הפתיע אותנו, הקורלציה ביניהם יחסית נמוכה ועומדת על 0.1.

טבלת קורלציה מצורפת לנספחים

3. איכות הנתונים

עבור הנתונים היו מספר תיקונים אותם היינו צריכים לבצע. עבור ערכים נומריים ומשתנים רציפים בפרט, העמקנו בידע שלנו בנוגע לטווח הנורמטיבי המתקבל עבור משתנים רציפים, וערכים הגיונים עבור כל משתנה. כאשר נחשפנו לחריגות קיצונית בנתונים אשר לא תואמת את המציאות (לדוגמא ערכים רחוקים משמעותית מהחציון), קטמנו את הסף המקסימלי והתייחסנו לערכים הקיצוניים כחריגים אך שיש לשמור עליהם. בחירת נקודת הקטימה היא לאחר למידת המשתנה וערכו אך בוצעה לפי שיקול דעתינו ההבנתי. שמרנו עליהם בכך שהפכנו את ערכם להיות הערך הממוצע של הנתונים (ללא הערכים הבעייתיים). את התיקונים הללו ביצענו בשלב ההתחלתי לפני חישובי ההתפלגויות האפרוריות מכיוון ששאפנו לקבל את המדגם המייצג ביותר.

בנוגע לנתונים חסרים, בשלב הראשוני של חישוב ההתפלגויות האפרוריות, התעלמנו מהם מהסיבה שרצינו להתייחס להתפלגויות האמיתיות מהמדגם. בשלב הבא, ערכים חסרים של משתנים רציפים בוטאו באמצעות הממוצע של המשתנה, ומשתנים קטגוריאליים בוטאו על ידי המשתנה בעל השכיחות הגבוהה ביותר. בחרנו להתייחס לערכים החסרים בצורה הזו על מנת לא להפסיד ערכים משמעותיים אחרים.

השמטת משתנים:

בנוגע להשמטת מאפיינים רועשים וחסרי חשיבות, בחרנו להתעלם ממשתנה ה ID כבר בשלב הראשון. התפלגותו האפרורית אינה שופכת אור רלוונטי על נתוני הנבדקים, שכן תעודת זהות הוא ערך יחודי לכל נבדק אשר משמש לזהותו בלבד (התפלגותו מצורפת לנספחים). משתנה נוסף שבחרנו להשמיטו הינו משתנה oral אך בחרנו להשמיטו בשלב לאחר בחינת התפלגות האפרורית מכיוון שיש לו משמעות בהבנת הנתונים. מכיוון שכלל הנתונים הינם זהים (כלל הנבדקים עברו בדיקת שייניים) נשמיט אותו ומעתה נתייחס לביצוע בדיקת שיניים כהנחה בסיסית שמתקיימת ובפרט לקביעת עששת הפה. בנוסף הורדנו את המשתנים: Serum Creatinine, Eyesight (left), waist(cm), hearing(right), hearing(left), Urine protein, Eyesight (right) שבנוסף לכך לא מצאנו קשר בין כאשר חקרנו את המידע. בעבור דיסקרטיזציה של משתנים רציפים, בחרנו להפוך חלק מהמשתנים הרציפים לבדידים, מתוך שיקולים של הבנת פלחי האוכלוסייה והתייחסויות דומות של ערכי המשתנה אשר לדעתנו הגיוני שיקוטלגו כערך בדיד. אופן החלוקה בוצעה בעקבות חקירה והבנת עולם התוכן.

1. BMI - זהו משתנה אשר מהווה מדד ליחס בין הגובה למשקל (במטרים) בריבוע ומתאר את המצב הבריאותי המשקלי של הנבדק בהינתן הגובה והמשקל. לאחר מחקר בנושא, הסקנו שמשתנה זה מהווה אינדיקציה טובה יותר עבור ערכי משקל תקינים וניתן להסיק על המצב הבריאותי של הנבדק בצורה מהימנה יותר מאשר להסתכל על מדד משקל בלבד.
2. לחץ דם כללי - מדד לחץ אשר לוקח בחשבון את הלחץ דם הסיסטולי והדאסטולי. ערכים תקינים של לחץ דם מושפעים משילוב של שני הערכים לפי סף ערכים אשר הוגדר במידע שחקרנו (נמצא בנספחים).

4. נספחים:

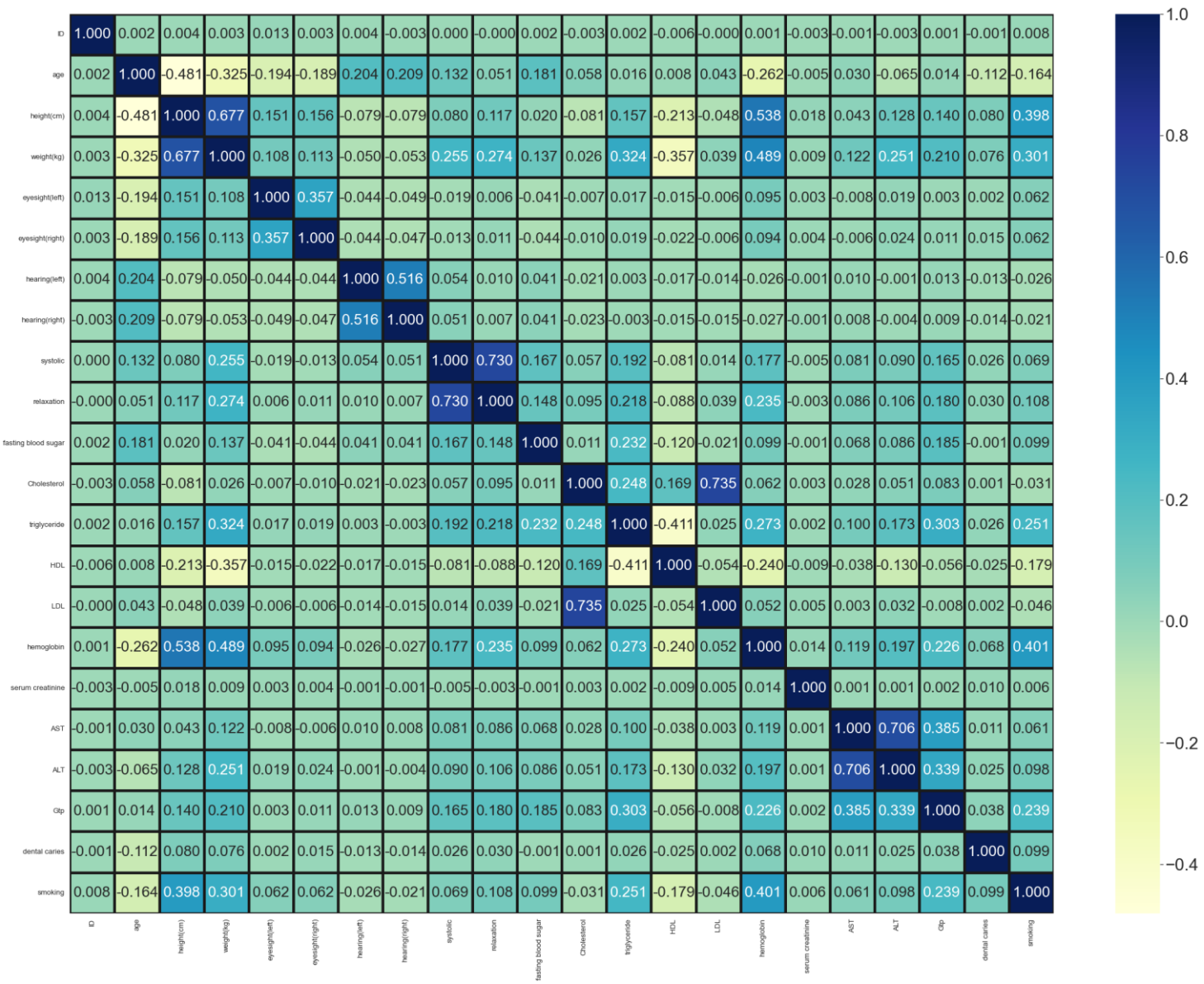
הסברים בנוגע למשתנים

משתנה	הסבר
ID	מספר סידורי ייחודי המשמש כמספר זיהוי הנבדק.
Gender	מגדר הנבדק.
age	גיל הנבדק
Height (cm)	מידת גובה בסנטימטרים (בקפיצות של 5 ס"מ, עיגול המדידה)
Weight (kg)	מידת המשקל של הנבדק בקילוגרמים
Waist (cm)	מידת היקף מותניים של הנבדק בס"מ.
Eyesight (left)	מדד הראיה בעין שמאל.
Eyesight (right)	מדד הראיה בעין ימין.
Hearing (left)	1 מצוין שומעים תקין באוזן שמאל, 0 מצוין לקות שמיעה באוזן שמאל
Hearing (right)	1 מצוין שומעים תקין באוזן ימין, 0 מצוין לקות שמיעה באוזן ימין
Systolic	לחץ סיסטולי הינו לחץ הדם הנמדד על דפנות העורקים כאשר שריר הלב מתכווץ במטרה להזרים דם לרקמות הגוף הנמדד ביחידות של מילימטר כספית. ערך אופטימלי יהיה נמוך מ120, נורמלי בין 120-129, נורמלי גבוה בין 130-139 ויתר לחץ דם יהיה שווה או גבוה מ140. ערכי לחץ דם יסווגו במכלול כשילוב של שני סוגי לחץ הדם סיסטולי ודיאסטולי.
Relaxation (Diastolic)	הלחץ הדיאסטולי (במצב רגיעה) הינו לחץ הדם הנמדד על העורקים כאשר שריר הלב נמצא במצב הרפיון/מנוחה, בין ההתכווצויות הנמדד ביחידות של מילימטר כספית. ערך אופטימלי יהיה נמוך מ80, ערך נורמלי יהיה בין 80-84, ערך נורמלי גבוה יהיה בין 85-89 ויתר לחץ דם יהיה שווה או גבוה מ90. ערכי לחץ דם יסווגו במכלול כשילוב של שני סוגי לחץ הדם סיסטולי ודיאסטולי.
Fasting Blood Sugar	בבדיקת סוכר בדם לאחר צום נמדדת רמת הסוכר בדם. הסוכר הוא מקור האנרגיה הזמין לתאי הגוף ומקורו בפירוק הפחמימות שאנו אוכלים. יכולת הגוף לנצל את הסוכר תלויה באינסולין אשר מיוצר בבלבל. ניתן לאבחן סוכרת ולזהות מצבים של רמת סוכר חרגה ובקרה כללית. יחידות המידה הינן מג"ג לדצ"ל. רמה תקינה נעה בין 72-100, רמות סוכר מעל 100 נחשבות ללא תקינות המעידות על סיכון לחלות בסוכרת.
Cholesterol	הכולסטרול הוא חומר שומני המיוצר בכל תאי הגוף ובעיקר בכבד. יש לו תפקיד חשוב בבניית התאים, בריפודם ובהגנתם. מלבד אלה משמש הכולסטרול כחומר גלם להורמונים רבים ולחומצות מרה. מטרת בדיקה זו היא הערכת גורמי הסיכון לטרשת העורקים, למחלת לב כלילית (תעוקת החזה, אוטם שריר הלב) ולמחלות כלי דם. ערכי כולסטרול מושפעים מערכי כולסטרול "טוב" (HDL), ומערכי כולסטרול "רע" (LDL). ערכי כולסטרול כללי רצוי שלא יעלו על 200 מ"ג לד"ל, רמת כולסטרול בין 200-239 מ"ג לד"ל מוגדרת כגבולית עד מסוכנת וערכי כולסטרול מעל 240 מ"ג לד"ל מוגדרים כסיכון גבוה.
Triglyceride	טריגליצרידים הם סוג השומן הנפוץ ביותר בגוף. טריגליצרידים מרכיבים את רקמת השומן ומשמשים מחסן אנרגיה ארוך טווח של הגוף. חלקם מיוצר בכבד, אולם חלק ניכר מהם מקורו בעיכול השומן שבמזון ובעודפי הפחמימות שבו. עיקר השימוש בצורת אנרגיה זו מתרחש במצב של רעב, כאשר מאגרי אנרגיה אחרים בגוף אזלים. רמות גבוהות של ט"ג בדם תורמות לעליית הסיכון למחלות לב ולאירוע מוחי ויכולות להצביע על קיום תסמונת מטבולית. נמדד ביחידות מיליגרם לדציליטר. ערכים תקינים הינם עד 150, וערכים בין 150-200 מוגדרים גבולי גבוה, 200-500 גבוה מהנורמה, סיכון גבוה למחלת עורקים. מעל 500 גבוה מאוד מהנורמה, סיכון מוגבר לדלקת של הלב.

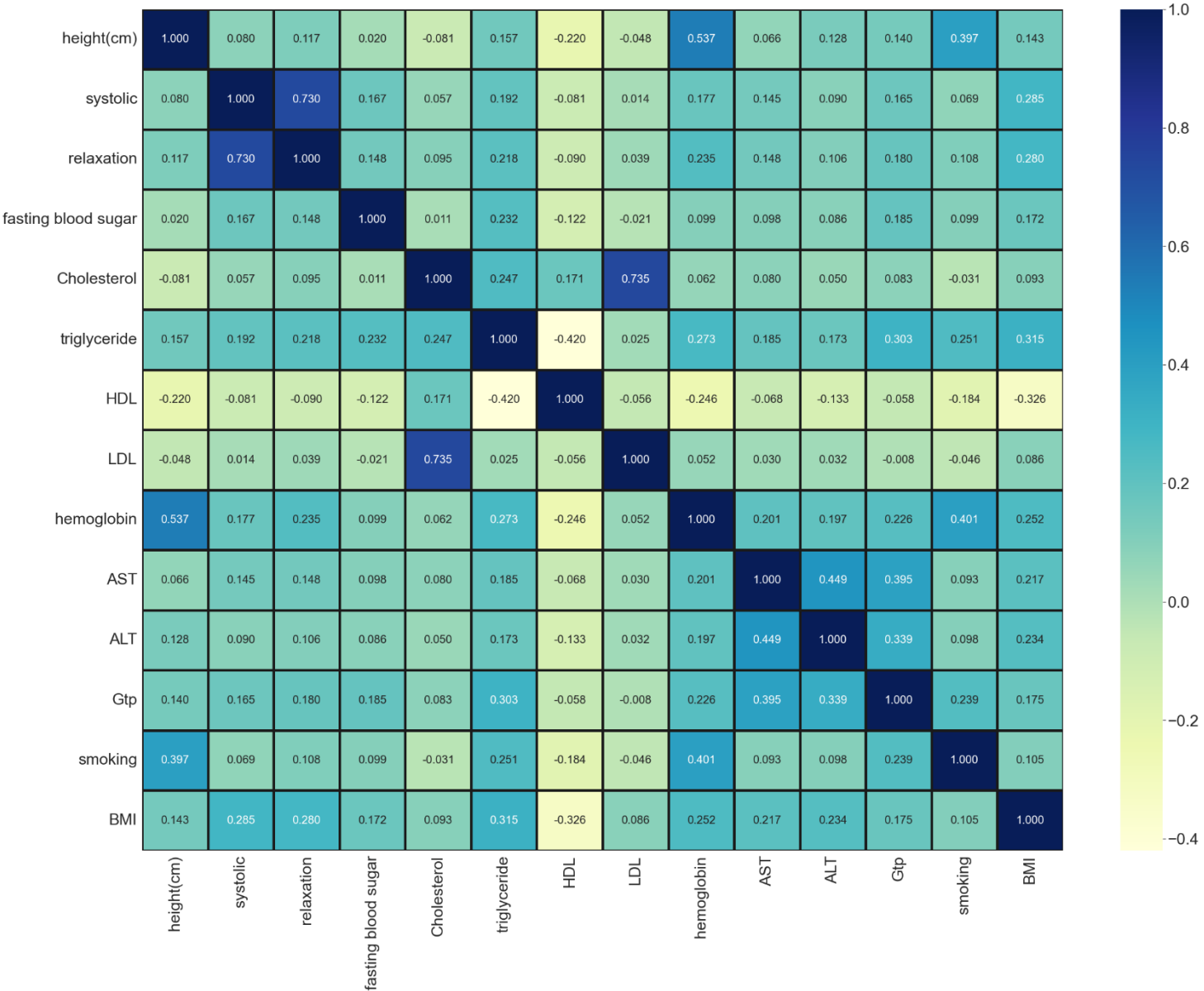
HDL	<p>נישא בזרם הדם על גבי ליפופרוטאין "הטוב כולסטרול" מכונה גם כ HDL high-density lipoprotein תפקידו לספוג עודפי כולסטרול מן הרקמות ולהשיב כולסטרול. בצפיפות גבוהה נוצר בכבד ובמעי בחזרה אל הכבד. בכבד נהפך הכולסטרול בחלקו למלחי מרה המאפשרים מרקמות הגוף ומכלי הדם תועלת מרובה למניעת הצטברות רובד שומני HDL-לגוף. בכך מביא ה את הפרשתו אל מחוץ מטרת בדיקה זו היא הערכת גורמי הסיכון לטרשת העורקים, למחלת לב כלילית - בדופנות העורקים ערכיו רצוי שיהיו גבוהים מ-40 מיליגרם לד"ל. (החזה, אוטם שריר הלב) ולמחלות כלי דם (תעוקת רצוי שהכמות תהיה גבוהה יותר, מ-50 מיליגרם לד"ל בקרב גברים, ואילו אצל נשים</p>
LDL	<p>ליפופרוטאין בצפיפות נמוכה - מכונה "הכולסטרול הרע". LDL מיוצר בכבד ומשמש להניע את הכולסטרול מהכבד אל רקמות הגוף. LDL מתפרק בדם והכולסטרול שהוא נושא שוקע בדופנות העורקים. כולסטרול זה, שוקע בדפנות העורקים וגורם להתקשחות הדפנות ולהיצרות העורקים (טרשת העורקים). במחקרים רפואיים נקבע כי רמה גבוהה של LDL מגבירה את הסיכון למחלות הלב וכלי הדם דוגמת טרשת העורקים, אוטם שריר הלב ואירוע מוח. ערכיו רצוי שיהיו נמוכים מ-100 מיליגרם לד"ל, ובמידת האפשר, נמוכים מ-70 מיליגרם לד"ל.</p>
Hemoglobin	<p>המוגלובין הוא מולקולה חלבונית המכילה ברזל. המולקולה נמצאת בתאי הדם האדומים. תפקידה הוא לקשור חמצן ולשאת אותו לתאי הגוף בכול הרקמות, תהליך הקרוי חימצון. רמה תקינה של המוגלובין הכרחית לשם חימצון תקין של כל תאי הגוף. חוסר בהמוגלובין גורם לאנמיה, המתבטאת בעיקר בחולשה, עייפות, חיוורון ודופק מהיר. מטרת בדיקה זו היא אבחון וניטור של אנמיה (חוסר דם) או פוליציטמיה (עודף דם). ערך תקין אצל נשים - 12-16 מיליגרם לדציליטר, ואצל גברים - 12-18 מיליגרם לדציליטר. קיימים מצבים בהם נראה רמות גבוהות מהרגיל של המוגלובין בדם, לדוגמה אצל מעשנים "כבדים" – שגופם מייצר יותר המוגלובין כדי לפצות על הירידה ברמת החמצן ברקמות שנגרם כתוצאה מהעישון.</p>
Urine protein	<p>מהבדיקה ניתן לדעת אם מצויים חלבונים בשתן. במצב תקין נוזל השתן אינו מכיל חלבונים, משום שגודלם אינו מאפשר להם לעבור דרך צינוריות הכליה. פעמים רבות יכולה הימצאות של חלבון בשתן לנבוע מסיבות פיזיולוגיות שחולפות מאליהן והיא אינה מעידה בהכרח על מחלה משמעותית. עם זאת, הימצאות חלבון בשתן יכולה להעיד גם על מחלת כליה כמו פגיעה סוכרתית בכליה, רעלת הריון, מחלות זיהומיות ולחץ דם גבוה. ערכי הנורמה הינם 0 עד 30 מיליגרם לדציליטר. הטווח הקטגוריאלי מגדיר את רמת החלבונים בשתן (ככל שהקטגוריה גבוהה יותר, כך רמת החלבונים בשתן גבוהה יותר)</p>
Serum Creatinine	<p>קריאטינין נוצר מפירוק של קריאטין פוספאט, מרכיב חשוב בשרירים. הקריאטינין מופרש מן הדם לשתן על ידי הכליות. רמתו בדם ובשתן מהווה מדד ישיר לתפקוד הכליות, כאשר במצב של פגיעה בהן יעלה ערך הקריאטינין בדם ופחות קריאטינין יופרש בשתן. ערכים מתחת לנורמה (קטנים מ-0.6) עלולות להצביע על מחלות שריר מסוימות או תת תזונה, ערכים נורמלים (בין 0.6-1.3) הינם הטווח התקין, וערכים גבוהים מהנורמה (מעל 1.6) עלולים להצביע על תפקוד לקוי של הכליות. יחידות המידה הינם מיליגרם לדציליטר.</p>
AST	<p>אספארטאט אמינוטרנספראז - בבדיקה זו נמדדת רמת האנזים AST בדם. בבדיקה זו נעשית כחלק מאיבחון של הפרעות בתפקוד הכבד, בשילוב בדיקות נוספות כגון ALT. אנזים זה נמצא בריכוז גבוה בתאי הכבד, תאי שריר הלב ושרירי השלד. האנזים נמצא גם בתאי הכליה, הלב ובתאי דם אדומים, אך בריכוז נמוך יותר. פגיעה בתאים המכילים AST תגרום לשחרורו מהתא ולעליית רמתו בדם ביחס ישר לכמות התאים הנפגעים. ערכים תקינים הינם 0 עד 35 יחידות לליטר לגברים; 0 עד 31 יחידות לליטר לנשים.</p>
ALT	<p>בבדיקה זו נמדדת רמת האנזים אלאנין טרנסאמינאז בדם. האנזים חיוני אלאנין טרנסאמינאז - לתהליכי חילוף חומרים, בהם מזון הופך לאנרגיה זמינה עבור הגוף. האנזים נמצא בעיקר בתאי הכבד אך גם בתאי הכליה, הלב ובשרירים. האנזים משתחרר למחזור הדם במצב של מחלה או פגיעה בתאים בהם הוא נמצא. רמתו בדם מושפעת באופן ישיר מכמות התאים הנפגעים. היות שחלקו העיקרי של האנזים נמצא ברקמת הכבד, רמת ALT גבוהה בדם מעידה בסבירות גבוהה על פגיעה בכבד. רמות תקינות- גברים: 0 עד 45 יחידות לליטר, נשים: 0 עד 34 יחידות לליטר.</p>
GTP	<p>גמא-גלוטאמיל טראנספפטידאז - (נקרא גם GGT) - אנזים הנמצא בעיקר בכבד ובדרכי המרה. בריכוזים נמוכים יותר ניתן למצוא גם בכליה, בלבב ובמוח. ברקמות אלו הוא מסייע בתהליך חילוף החומרים. הבדיקה נועדה על מנת לאתר פגיעה בכבד ובצינורות המרה, נזקי אלכוהול ומחלות בכבד. אנזים זה רגיש ביותר לשינויים בתפקודי הכבד ובדרכי המרה. רמות גבוהות של GGT מצביעות על שיבוש בתפקודי הכבד. ככל שרמת האנזים גבוהה יותר כך גבוהה רמת הפגיעה בכבד. ערכים תקינים- עד 51 יח' בינלאומיות לליטר.</p>
Oral	<p>האם בוצעה (Y) או לא (N) בבדיקה בחלל הפה של הנבדק.</p>

Dental Caries	עששת היא מחלה זיהומית רב-גורמית, שבה מתרחש תהליך הרס הרקמות הקשות של השיניים-זגוגית השן (האמייל). ההרס עצמו נגרם כתוצאה מפעולתם של חיידקים מסוימים הנמצאים על השיניים. עששת עלולה לגרום לאי נוחות, לכאב, הפרעה בתפקוד (לעיסה, דיבור) ובאסתטיקה של החיך. הבדיקה קובעת האם הנבדק לוקה בעששת (1), 0-אחרת.
Tartar	אבנית - האם הנבדק לוקה באבנית בחלל הפה (1), 0-). אחרת. האבנית נוצרת ככל שמקפידים פחות על הגיינת השיניים וגם כשרמת החומציות ברוק שלנו משתנה באופן תדיר, ומעודדת בשינויי החומציות היווצרות של אבנית. היווצרות אבנית בשיניים יכולה לגרום לנזקים קוסמטיים ולהופעת דלקת בשיניים.
Smoking	משתנה מטרה - האם הנבק מעשן(1) או לא(0).

טבלת קורלציה HeatMap

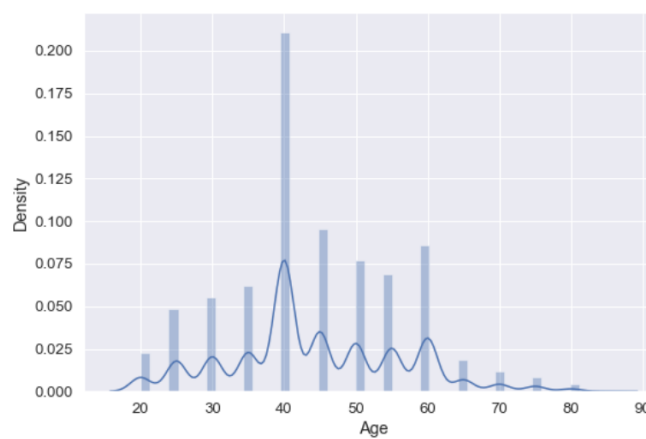
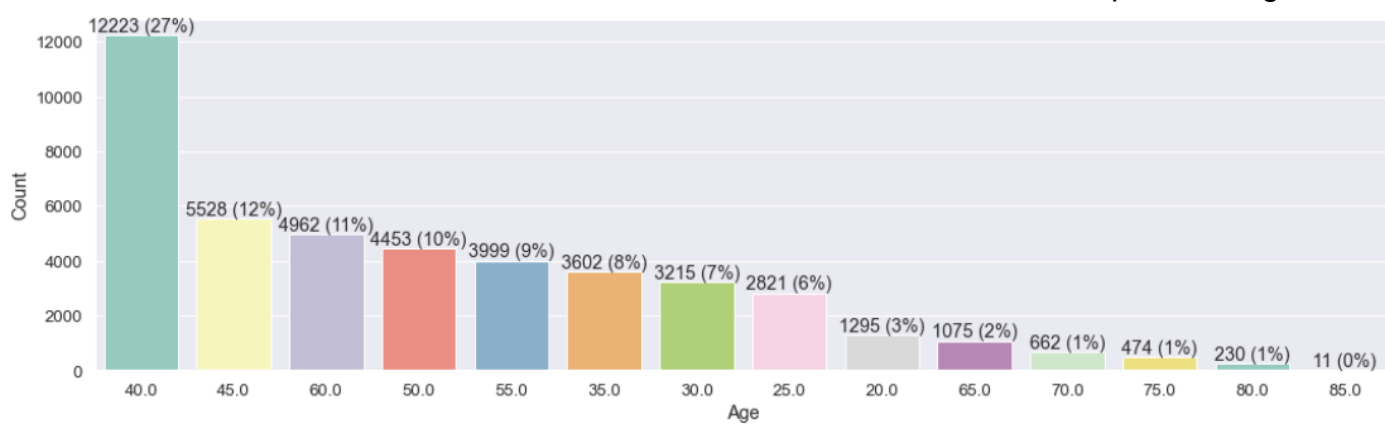


HeatMap – לאחר הורדת המשתנים (סופי)



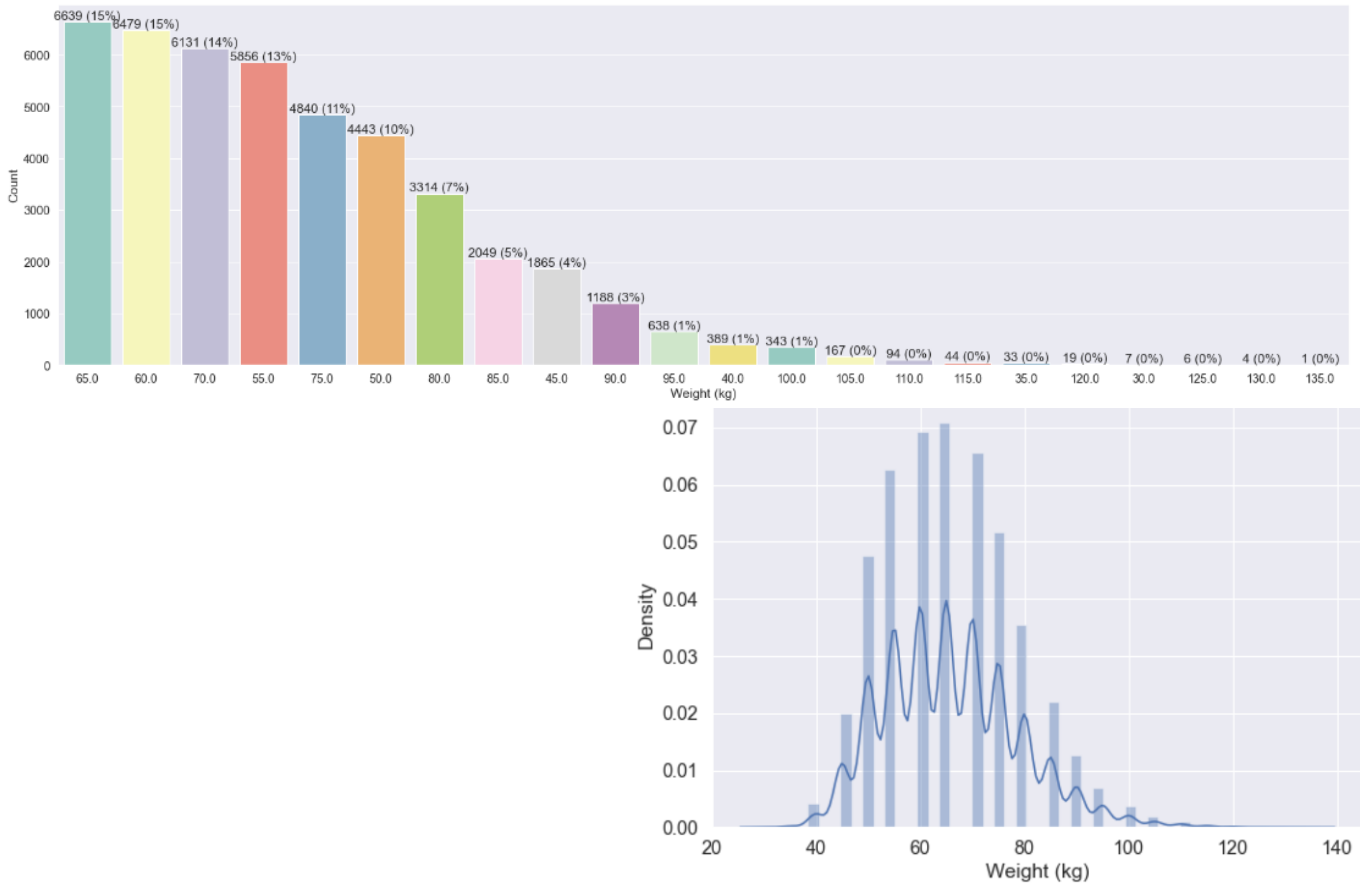
גרפים נוספים

משתנה Age לפני דיסקרטיזציה :



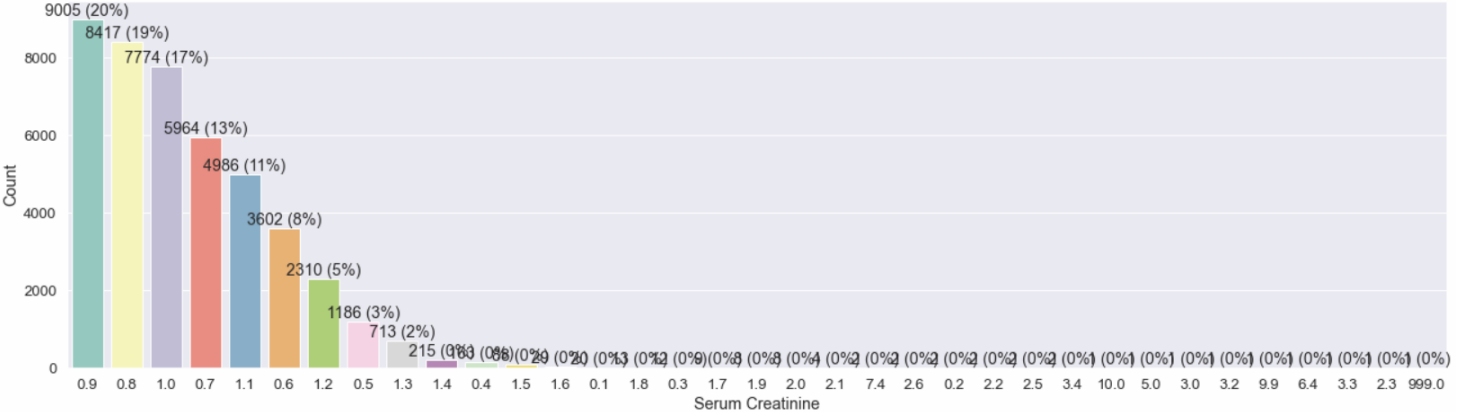
בחרנו במשתנה זה לבצע דיסקרטיזציה לטובת תצוגה ברורה יותר של נתונים, ובהתאם למידע שחקרנו ע"י מאמרים ראינו שיש חשיבות לחלקות קבוצות הגיל בצורה זו.

משתנה Weight(cm) לפני דיסקרטיזציה :

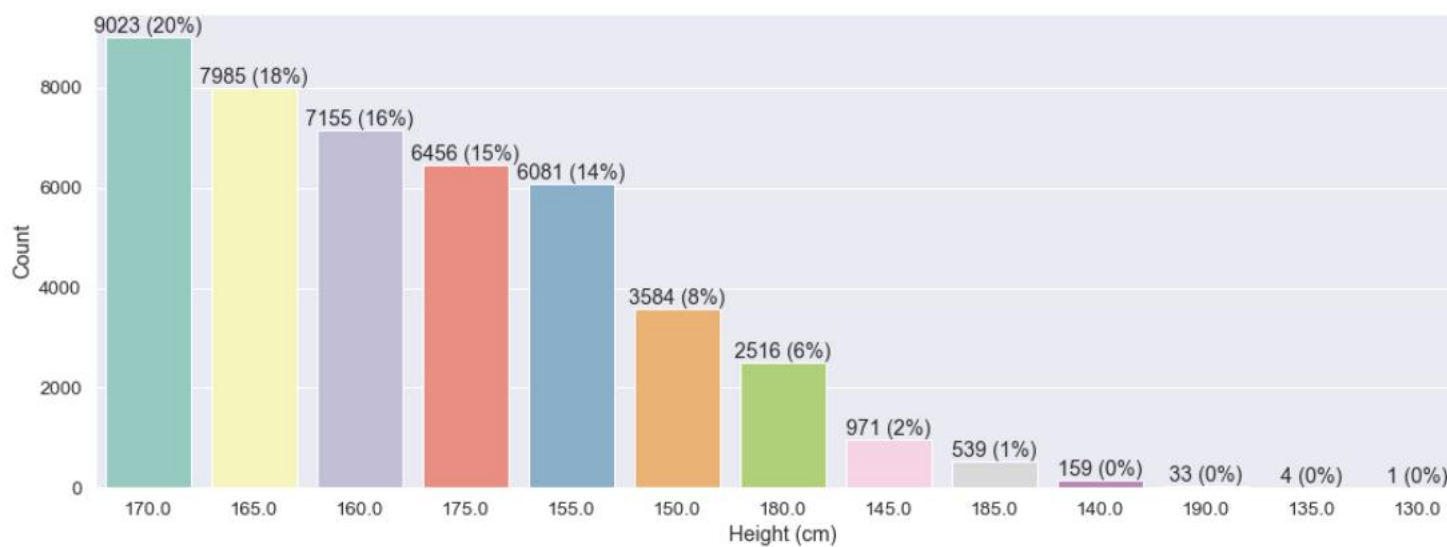


בחרנו במשתנה זה לבצע דיסקרטיזציה לטובת תצוגה ברורה יותר של נתונים, ובהתאם למידע שחקרנו ע"י מאמרים ראינו שיש חשיבות לחלקות קבוצות הנשקל בצורה זו.

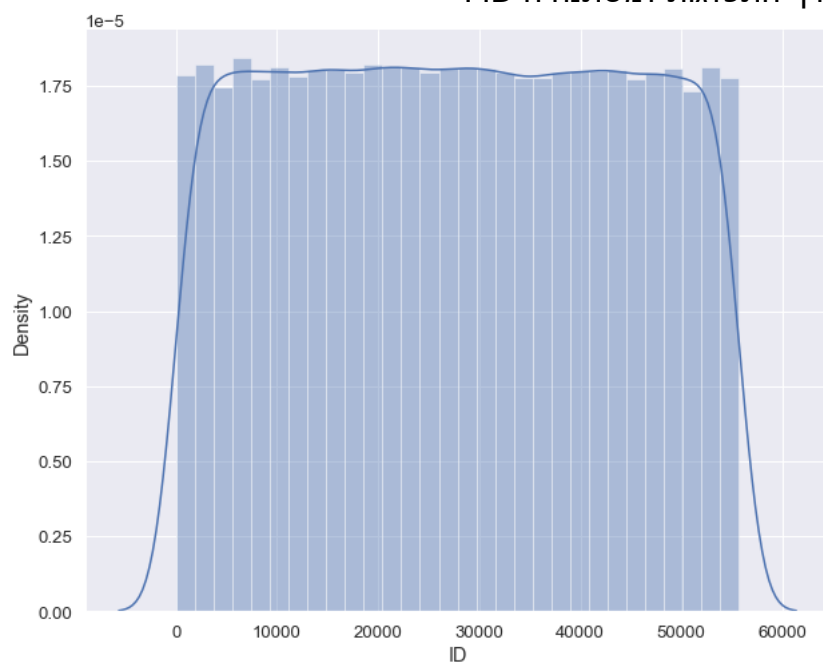
משתנה Serum Creatinine לפני דיסקרטיזציה:



הצגה נוספת של משתנה Hight(cm) :



גרף התפלגות למשתנה ID :



ביבליוגרפיה:

ערכי לחץ דם תקינים ע"פ מדדים של דיאסטולי וסיסטולי עבור המשתנה Blood Pressure :

<https://www.tasmc.org.il/Be-Well/InterestAreas/pressure/Pages/HBP.aspx>

משמעות משתנים שונים ע"פ מכבי בדיקות מעבדה:

<https://www.maccabi4u.co.il/5024-he/Maccabi.aspx>

קבוצות גיל ומשקל בקרב מעשנים:

https://www.health.gov.il/PublicationsFiles/smoking_2019.pdf

<https://smokefree.org.il/wp-content/uploads/2021/07/%D7%94%D7%9E%D7%99%D7%96%D7%9D-%D7%9C%D7%9E%D7%99%D7%92%D7%95%D7%A8-%D7%94%D7%A2%D7%99%D7%A9%D7%95%D7%9F-%D7%93%D7%95%D7%97-%D7%9E%D7%97%D7%A7%D7%A8-%D7%A2%D7%99%D7%A9%D7%95%D7%9F-%D7%91%D7%99%D7%A9%D7%A8%D7%90%D7%9C-2021.pdf>