

**פרויקט הקורס - חלק א'**

**מגישים**

ענבל אפשטיין 31461763

יונתן גולן 206387383

תוכן עניינים

[**הגדרת הבעיה** 3](#_Toc121412958)

[תיאור כללי של עולם התוכן שנחקר 3](#_Toc121412959)

[שאלת המחקר 3](#_Toc121412960)

[תיעוד מקורות הנתונים ומשמעותם 3](#_Toc121412961)

[מאפייני סט הנתונים 3](#_Toc121412962)

[**הסתברויות אפריוריות וקשרים בין המאפיינים 5**](#_Toc121412963)

[קשרים מעניינים 9](#_Toc121412964)

[משתנים בעלי השפעה על משתנה המטרה 10](#_Toc121412965)

[**איכות הנתונים** 10](#_Toc121412966)

[נתונים לא הגיוניים 10](#_Toc121412967)

[נתונים חסרים 11](#_Toc121412968)

[**הכנת הנתונים** 12](#_Toc121412969)

[השמטת מאפיינים 12](#_Toc121412970)

[דיסקרטיזציה של ערכים רציפים 12](#_Toc121412971)

[גזירת פונקציות של משתנים קיימים 12](#_Toc121412972)

[**נספחים** 13](#_Toc121412973)

[נספח 1 13](#_Toc121412974)

[נספח 2 14](#_Toc121412975)

[**ביבליוגרפיה 14**](#_Toc121412976)

# **הגדרת הבעיה**

## תיאור כללי של עולם התוכן שנחקר

עולם התוכן של המחקר עוסק בהתנהגות נהגים ושימוש במידע הקיים עליהם לצורך בניית מודל חיזוי. במחקר זה נרצה לחזות האם אדם יממש את ביטוח הרכב שלו במהלך השנה הבאה בהינתן מאפיינים עליו.

תחום זה של ניתוח נתונים וחיזוי מבוסס התנהגות נהגים הוא בעל חשיבות רבה בשל השפעותיו הכלכליות והחברתיות, הן עבור חברות הביטוח והן עבור הציבור. הבנת גורמים המשפיעים על מימוש ביטוח עשויה לסייע בחיזוי טוב יותר של סיכונים, בתמחור מדויק של פוליסות ביטוח, ואף בקידום אמצעים לשיפור הבטיחות בדרכים.

סט המידע שבידינו כולל מדגם של 8,000 אנשים, עם נתונים דמוגרפיים כגון רקע וגיל, ונתוני נהיגה הכוללים את מספר הפרות התנועה שביצעו, מספר דוחות על מהירות, סוגי הרכבים שלהם, ושנות הייצור שלהם.

בהתבסס על מידע זה, תחום המחקר עוסק במציאת קשרים ותובנות בין מאפייני הנהגים והתנהגותם לבין סיכויי מימוש הביטוח. בעבר, מחקרים בתחום זה השתמשו במודלים סטטיסטיים ובאלגוריתמים של למידת מכונה כדי לנתח דפוסים ולחזות התנהגויות, תוך שימוש בעיבוד מקדים למניעת רעש נתונים והתמקדות בתכונות מרכזיות כמו הרגלי נהיגה ותדירות עבירות.

מסקנות ממחקרי עבר מצביעות על כך שמודלים מתקדמים, כמו רשתות עצביות, עשויים לספק דיוק גבוה יותר בזיהוי קשרים מורכבים בין משתנים, במיוחד במקרים שבהם יש נתונים גדולים ומגוונים. כמו כן, היכולת לחזות תנאים קיצוניים, כגון נהגים בעלי סיכון גבוה במיוחד למימוש הביטוח, נותרה אתגר מרכזי המחקר.

מטרת המחקר הנוכחי היא לבחון האם ניתן לבנות מודל חיזוי על בסיס נתוני המדגם שבידינו, כזה שיוכל לספק לחברות הביטוח כלי יעיל לניהול סיכונים ותכנון תמחור מדויק.

שאלת המחקר  
האם ניתן לבנות מודל חיזוי בעזרת נתוני ביטוח הרכב על מנת לחזות האם לקוח יגיש תביעת ביטוח במהלך השנה הבאה?

**הבנת הנתונים**

## תיעוד מקורות הנתונים ומשמעותם

מקור הנתונים הוא מאגר מידע אנונימי של לקוחות, הכולל נתונים דמוגרפיים, פיננסיים והתנהגותיים. בין היתר, נכללים נתונים כמו גיל, מגדר, רמת השכלה, רמת הכנסה ודירוג סיכון אשראי. חלק מהנתונים נאספו באמצעות שאלונים שהלקוחות מילאו, לדוגמה רמת ההשכלה ורמת ההכנסה. דירוג סיכון האשראי חושב על בסיס נתונים חיצוניים של חברות דירוג אשראי. מידע על הרכבים, כגון שנת הייצור וסוג הרכב, נאסף ממסמכי רישום רכב. נתוני עברות תנועה ותאונות נלקחו ממאגרי מידע ממשלתיים. כמות המיילים שהרכב נסע בשנה נמדדה באמצעות מערכות טלמטריה או דיווחים עצמיים של הלקוחות. הנתונים עברו עיבוד ואנונימיזציה לשמירה על פרטיות המשתמשים.

## 

## מאפייני סט הנתונים

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| משמעות | סוג המאפיין | יחידת מידה | מסביר/מוסבר | שם המאפיין |
| מספר הזהות של הלקוח | נומרי (בדיד) | - | מסביר | ID |
| גיל הלקוח | נומרי (בדיד) | שנים | מסביר | Age |
| מין הלקוח – גבר\אישה | קטגוריאלי לא ניתן לסידור | - | מסביר | gender |
| מספר שנות הניסיון נהיגה של הלקוח | נומרי (בדיד) | שנים | מסביר | Driving Experience |
| רמת השכלת הלקוח | קטגוריאלי ניתן לסידור | - | מסביר | Education |
| דרגת הכנסת הלקוח | קטגוריאלי ניתן לסידור | - | מסביר | Income |
| דירוג סיכון האשראי של הלקוח (דירוג יותר גבוה יותר טוב ללקוח) | נומרי (רציף) | - | מסביר | Credit Score |
| האם הלקוח הוא בעל הרכב. 1- כן , 0- לא | קטגוריאלי בוליאני | בינארי | מסביר | Vehicle Ownership |
| האם הרכב יוצר לפני או אחרי 2015 | קטגוריאלי לא ניתן לסידור | - | מסביר | Vehicle Year |
| מצב משפחתי של הלקוח. 1- נשוי, 0- אחרת | קטגוריאלי בוליאני | בינארי | מסביר | Married |
| האם ללקוח יש ילדים.  1- יש לו ילדים, 0- אין | קטגוריאלי בוליאני | בינארי | מסביר | Children |
| אזור מיקוד הלקוח | נומרי (בדיד) | קוד | מסביר | Postal Code |
| כמות המיילים שהרכב נסע בשנה הקודמת | נומרי (בדיד) | mile | מסביר | Annual Mileage |
| כמות עברות התנועה של הלקוח | נומרי (בדיד) | כמות | מסביר | Speeding Violations |
| סוג הרכב | קטגוריאלי לא ניתן לסידור | - | מסביר | Vehicle Type |
| כמות התאונות שעשה הלקוח | נומרי (בדיד) | כמות | מסביר | Past Accidents |
| ערך המטרה  1 אם הלקוח החזיר את ההלוואה שלו  0 אחרת | קטגוריאלי בוליאני | בינארי | מוסבר | Outcome |

**מפה מטה זה מהרפרנס**

הסתברויות אפריוריות וקשרים בין המאפיינים   
נסיק את ההסתברויות האפריוריות באמצעות ניתוח סט האימון. תרשימי ההיסטוגרמה עבור המשתנים הקטגוריאליים לרבות משתנה המטרה מוצגים **בנספח 1**.

**משתנה Smoking:** ההסתברות לאדם מעשן (smoking=1) היא 63.3%, ו-36.7% מהמדגם הן אינם מעשנים. על פי המחקר, כ-12.5% מאוכלוסיית ארצות הברית (מעל גיל 18) מעשנים, מכאן ניתן להסיק כי הנתונים בסט האימון לא מאוזנים ולא מייצגים את כלל האוכלוסייה האמריקאית, ובנוסף ניתן לחשוד כי נדגמו מאוכלוסייה עם כמות מעשנים גבוהה.

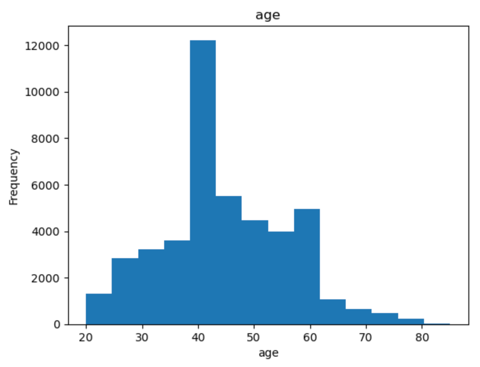
**משתנה Gender:** כ- 63.6% הם גברים, ו-36.4% מהמדגם הן נשים. באוכלוסיית ארצות הברית 50.47% הם גברים ו- 49.5% הם נשים, כלומר נתון זה לא מייצג את כלל האוכלוסייה, אך במחקרים נמצא כי גברים מעשנים יותר מנשים. ממצא זה גורם לנו לחשוד כי הנתונים נדגמו מאוכלוסייה מעשנת.

**משתנים hearing(left) & hearing(right) –** נתונים אלו מייצגים את נתוני השמיעה של האנשים במדגם, כאשר 1 – שמיעה תקינה, 2- לקות בשמיעה. ניתן לראות כי 97.4% בעלי שמיעה תקינה באוזן שמאל, וכ- 97.3% בעלי שמיעה תקינה באוזן ימין. בארצות הברית ההסתברות לשמיעה לקויה לגילים 20-70 היא 14% [1] . אנו לא יודעים כיצד הוגדרה שמיעה לקויה לכן אין לנו מספיק פרטים על מנת להחליט אם המדגם תואם למציאות.

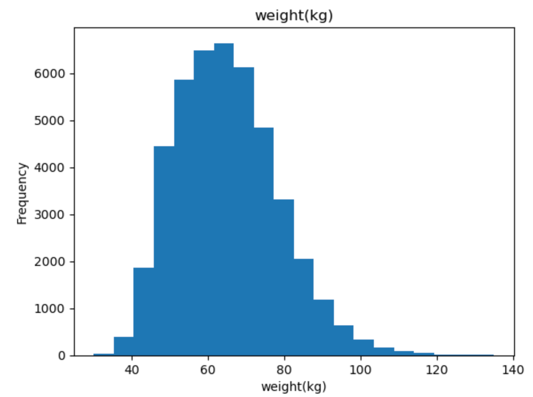
**משתנה dental caries –** המשתנה מתאר מצב של עששת. ההסתברות לעששת (dental caries=1) היא 21.4% וכ-78.6% במדגם (dental caries=0) הם ללא עששת.

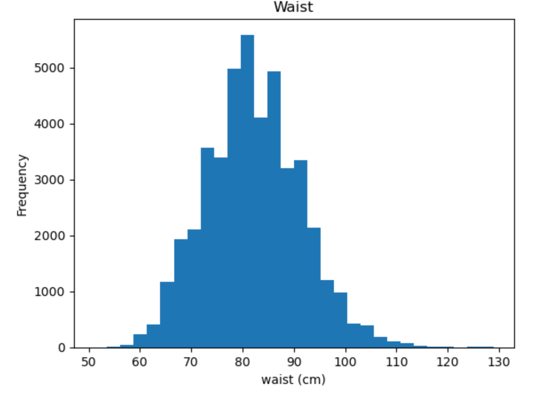
**משתנה tartar-** המשתנה מתאר מצב של אבנית בשיניים. ניתן לראות כי כ-55.3% בעלי אבנית בשיניים, וכ-44.7% מהאנשים במדגם ללא אבנית.  **משתנה oral –** משתנה הבודק האם בוצעה בדיקה בפה או לא. ניתן לראות כי 100% מהנבדקים במדגם עברו בדיקה, ישנם משתנים חריגים (Y,12% ( נטפל במשתנים אלו בהמשך, בנוסף נשקול בהמשך האם יש תועלת בהשארת העמודה בניתוח המודל בחלק ב'.

**משתנה Urine protein-** רמת החלבון בשתן כליות בריאות אינן מסירות חלבונים וחומרים מזינים חשובים אחרים ועלולות לתת לחלבון לדלוף אל תוך השתן, רמת חלבון גבוהה עשויה לסמן על פגיעה בכליות. ניתן לראות כי כ-94.5% מהמדגם עם רמת חלבון נמוכה בשתן (urine protein=1), כפי שנאמר, דליפת חלבון לשתן הינו מצב חריג בנוסף ישנם נתונים חריגים, נטפל בהם בהמשך.

 **משתנים נומריים**

**משתנה Age**: ניתן לראות כי הנתונים מתפלגים בצורה די נורמלית, כאשר קיים זנב ימני וריבוי תצפיות של מטופלים שגילם 40, שזהו באופן לא מפתיע גם החציון. הגיל הממוצע של המטופלים הוא 44.

**משתנה height(cm):** ניתן לראות כי הערכים מתפלגים נורמלית עם ממוצע של גובהcm 164, וסטיית תקן של 9.4. הגובה הממוצע באוכלוסיית ארצות הברית הואcm 157.5 ,לכן הנתונים תואמים את המציאות. (הגרף בנספח 2) **משתנה weight(kg):** המשקל הממוצע הינו kg 65 וסטיית התקן היא 12.84, ניתן לראות כי ההתפלגות היא התפלגות עם זנב ימני. הימצאות של מספר מועט יחסית של מטופלים בעלי משקל קיצוני (מעל 100 ק״ג) בניסוי יוצרת את אותו זנב ימני. ממחקרים ידוע כי ישנה נטייה להשמנה בארה"ב לכן זהו ממצא שעשוי להסביר את הימצאות התצפיות של המטופלים בעלי משקל קיצוני.

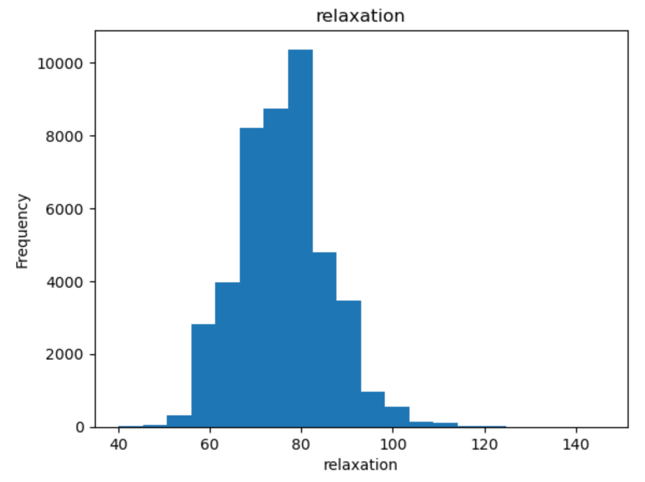


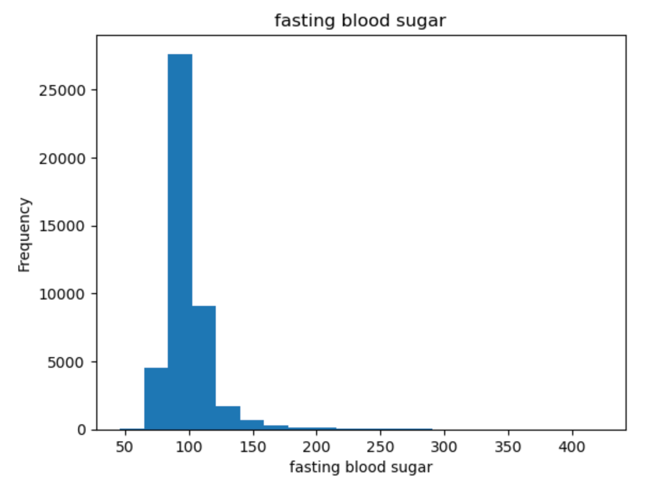
**משתנה waist(cm):** שמנו לב כי במשתנה זה ישנם מספר ערכים חריגים וחסרים. על מנת לבנות את ההיסטוגרמה, טיפלנו ברשומות הרלוונטיות בשלב איכות הנתונים. כעת ניתן לראות שהמשתנה מתפלג בצורה נורמלית, עם זנב ימני קטן, כפי שהיה בהתפלגות המשקל, שהרי הגיוני שכפי שיש מעט מטופלים בעלי משקל קיצוני, יהיו גם מעט מטופלים בעלי היקף מותניים קיצוני.

Chart

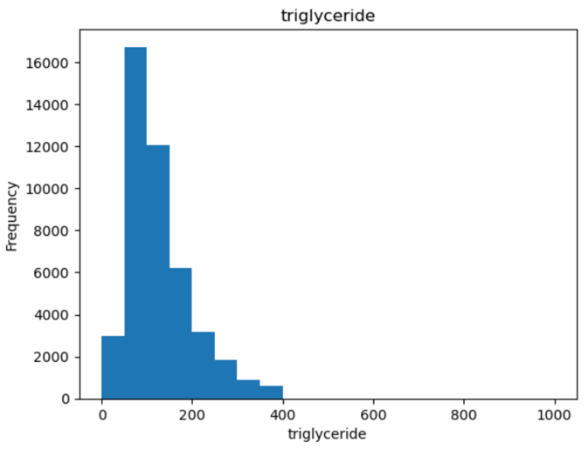
Description automatically generated**משתנה eyesight(left):** המשתנה בוחן את העוצמה האופטית של עין שמאל, הממוצע הוא 1 וסטיית התקן היא 0.494. ניתן לראות בתרשים ההיסטוגרמה בנספח 2 שקיימות רשומות עם ערך החשוד כחריג. דבר שייבדק בשלב איכות הנתונים.

**משתנה eyesight(right):** המשתנה בוחן את העוצמה האופטית של עין ימין, הממוצע הוא 1.013 וסטיית התקן היא 0.494. גם כאן, כמו במשתנה הקודם, ניתן לראות בתרשים ההיסטוגרמה בנספח 2 שקיימות רשומות עם ערך החשוד כחריג. **משתנה systolic:** לחץ דם סיסטולי מבטא את הלחץ שמופעל על דפנות העורקים כאשר שריר הלב מתכווץ ומזרים גל של דם לרקמות הגוף. הוא נמדד ביחידות של מילימטר כספית (ממ"כ). ההתפלגות היא נורמלית עם ממוצע 121.52 וסטיית התקן היא 14.27. לחץ דם סיסטולי תקין הוא עד 120 ממ״כ ואכן ניתן לראות שכיחות גבוהה לטווח הערכים התקין הנפוץ במציאות. הטווח בהיסטוגרמה הוא רחב מפני שקיימות רשומות עם ערכים חריגים של מעל 180 ממ״כ, דבר שייבדק בשלב איכות הנתונים.

**משתנה relaxation:** לחץ דם דיאסטולי ומבטא את הלחץ שמופעל על העורקים בעת הרפיית שריר הלב, בין ההתכווצויות. הממוצע הוא 76.0077 וסטיית התקן היא 9.68. ההתפלגות נראית די נורמלית עם זנב ימני קטן. הערך של לחץ דם דיאסטולי תקין הוא באזור 80 ולכן נראה שהנתונים תואמים את המציאות. גם כאן, קיימים ערכים חריגים של מעל 120 שייבדקו בשלב איכות הנתונים.

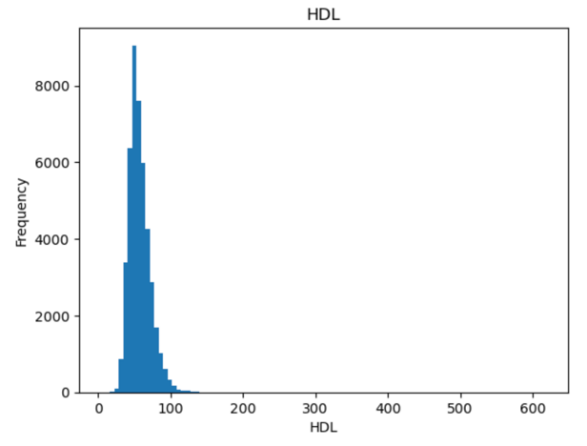
**משתנה fasting blood sugar:** רמת הגלוקוז בדם אינה מתפלגת סימטרי, ניתן לראות זנב ימני גדול מאוד. הממוצע הינו 99.34 וסטיית התקן גדולה יחסית ועומדת על 20.952. טווח רמת הגלוקוז התקינה בדם נע בין 72-100 מ"ג לדציליטר, על כן נראה כי ההסתברות האפריורית עבור משתנה זה נראית קרובה למציאות באוכלוסיית ארצות הברית.

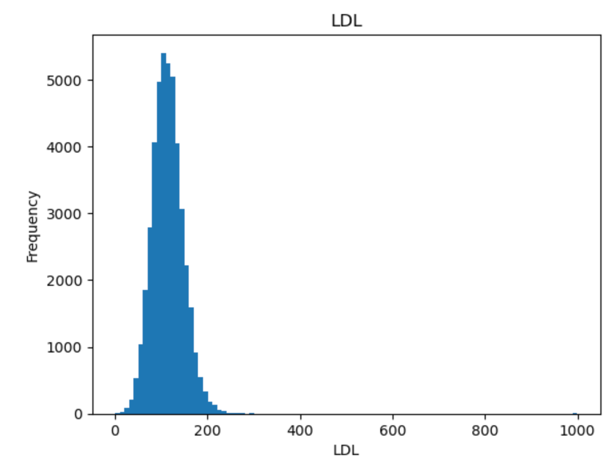
**משתנה Cholesterol:** התפלגות הכולסטרול נורמלית, בעלת זנב ימני קטן עם ממוצע 196.7 וסטיית תקן של 36.16. ערכים קטנים מ200 מ"ג\דציליטר מוגדר

כערך הרצוי לכולסטרול הכללי. לכן הנתונים תואמים את המציאות.

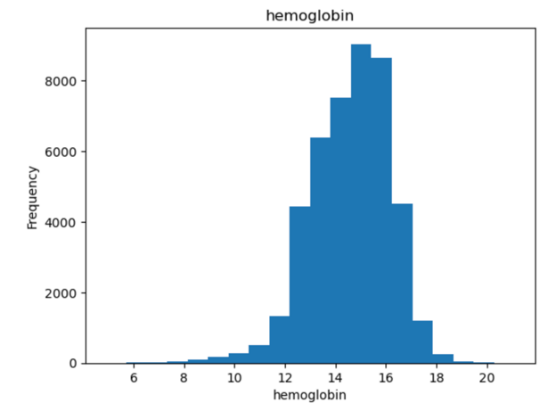
רמת כולסטרול של מעל 250 מ״ג/דציליטר נחשבת מסוכנת אך אפשרית.

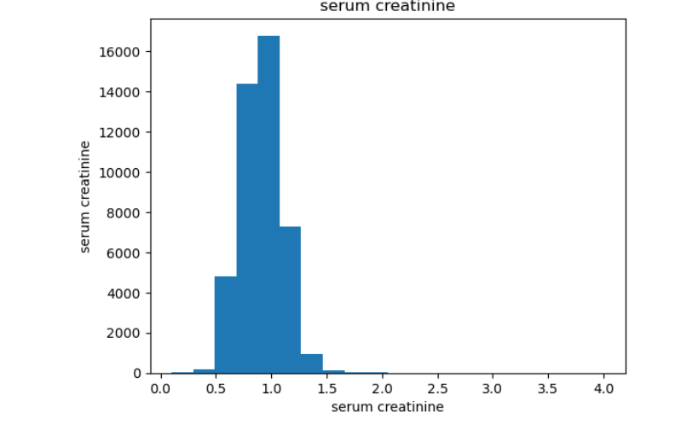
**משתנה triglyceride:** הטריגליצרידים נוצרים בגוף כתוצאה מתוצרי העיכול של השומן שבמזון והם הצורה בה השומן מאוחסן בגוף. כל ערך נמוך מ-150 הוא מומלץ, וערך הגבוה מ-500 נחשב גבוה מאוד. ניתן לראות כי הממוצע הוא 126.544 וסטיית התקן היא 71.66. קיימת רשומה עם ערך חריג (999) שתטופל בשלב איכות הנתונים.



**משתנה HDL:** מיוצר בכבד וממלא תפקיד חשוב בשמירה על העורקים מפני הצטברות של רובד שומני על ידי השבת עודפי כולסטרול מרקמות הגוף וכלי הדם אל הכבד. ניתן לראות כי הממוצע הוא 57.28 עם סטיית תקן 14.78, כל ערך מעל 40 הוא מומלץ ולכן הנתונים תואמים את המציאות. קיימים שני ערכים קיצוניים ביחס לשאר.

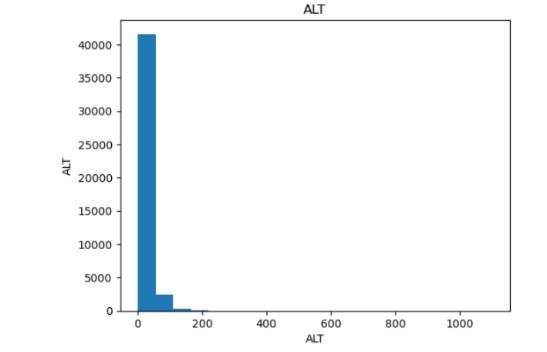
**משתנה LDL:** LDL מיוצר בכבד ותפקידו הוא להעביר את הכולסטרול מהכבד אל רקמות הגוף. כאשר ה-LDL מגיע לדם הוא מתפרק, והכולסטרול שאיתו שוקע בדפנות העורקים, ובכך יוצר את מחלת טרשת העורקים. ערך מעל 300 מ״ג/דציליטר נחשב גבוה וקיימים כאן מספר ערכים חריגים הגבוהים מ-500.

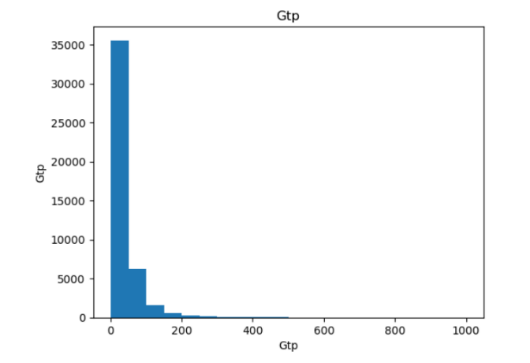
**משתנהhemoglobin :** ההמוגלובין הוא חלבון שנמצא בתאים האדומים ותפקידו להעביר חמצן אל התאים והרקמות, ופחמן דו חמצני מהתאים בחזרה לריאות. ההתפלגות של ההמוגלובין נראית נורמלית וקיימת שכיחות גבוהה בין 12 ל-17 גר׳/דציליטר, כאשר על פי מידע רפואי זהו הטווח התקין של המוגלובין לנשים ולגברים.



**משתנה serum creatinine –** קריאטינין הוא תוצר לוואי של חילוף החומרים ברקמת השריר. נהוג לבצע את הבדיקה כאשר יש חשד לבעיות בתפקוד הכליות, ערכים גבוהים עשויים להצביע על כשלי תפקוד. הממוצע הוא 0.907 עם סטיית תקן 4.733, ההתפלגות נראית נורמלית. ערכים תקינים נעים בין 0.51 עד 1.17 מ"ג לדציליטר (תלוי במגדר), לכן הנתונים תואמים את המציאות.

**משתנה AST –** בבדיקת AST בודקים את ריכוז האנזים GOP בדם, האנזים מצוי בכבד, בשרירים ובתאי דם אדומים. ערכי הנורמה הם 0 עד 35 יח'\ליטר לגברים, ו-0 עד 31 יח'\לליטר לנשים. משמעות ערכים גבוהים ייתכנו בעיקר סוגים שונים של מחלות כבד ומחלות כליה. הממוצע הוא 26.17 עם סטיית תקן של 19.29. (היסטוגרמה מצורפת בנספח 2)

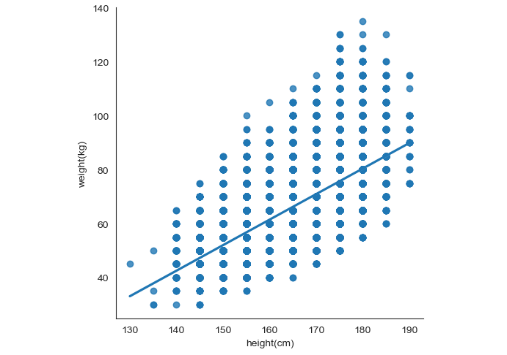


**משתנה ALT-** אלאנין טרנסאמינאז, גם כאן כאשר יש ריכוז גבוה של האנזים מעלה חשד לקיום מחלות כבד. (מקובל לבצע יחס בין AST/ALT כבדיקה למחלות כבד) הממוצע הוא 27 עם סטיית תקן גבוהה 31.2 שנוצרת כתוצאה מערכים חריגים, נטפל בהם בהמשך, אך הנתונים תואמים את המציאות הרפואית.

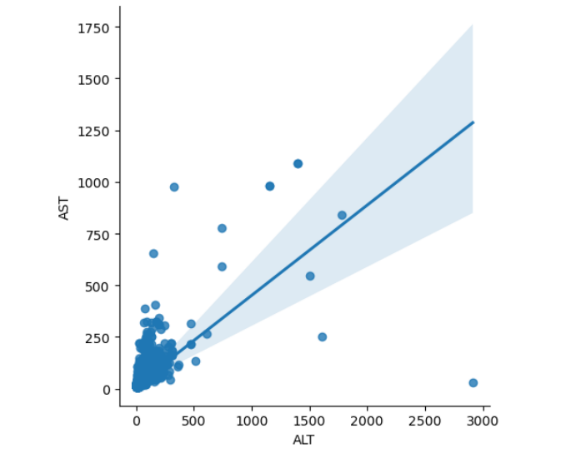
**משתנה Gtp**- אנזים נוסף שמצוי בעיקר בכבד, בדרכי המרה ובכליות. האנזים משוחרר לדם כאשר תאים שמכילים אותו נפגעים. ערכים גבוהים מהנורמה גם הם עשויים לסמן על חשד למחלות כבד או מחלות בדרכי המרה. הממוצע הוא 39.927 וסטיית התקן גדולה נוצרת גם כאן מערך אחד חריג 999, נטפל בו בהמשך.

## **קשרים מעניינים**

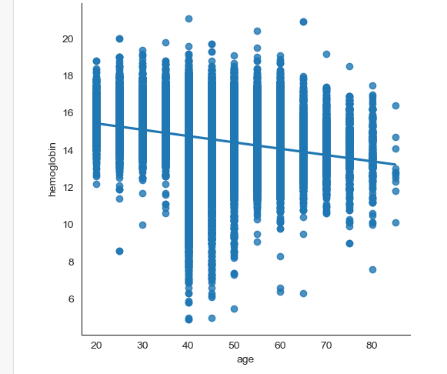
נבחן קשרים מעניינים בין המאפיינים בסט הנתונים. להשוואה בין משתנים רציפים נשתמש במקדם פירסון כמדד לקשר הליניארי ביניהם. בתרשים המצורף בנספח 3 מתוארים הקשרים במפת חום.



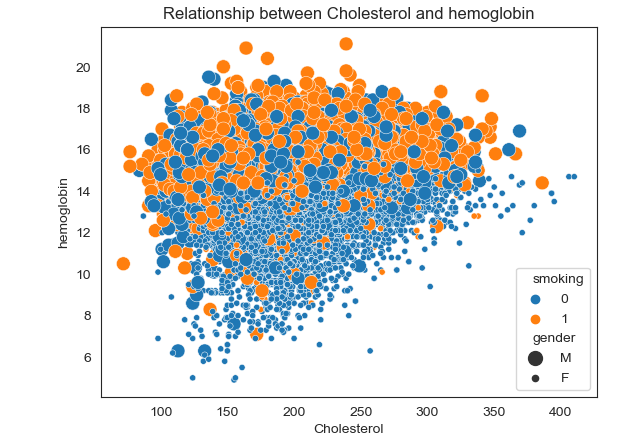
**קשר בין המשקל לגובה**בסט הנתונים יש מספר מאפיינים שציפינו לראות קורלציה גבוה הניתן לראות קורלציה חיובית בין הגובה למשקל. המשקל עולה בטבעיות כשהגובה עולה ולכן הנתונים תואמים את המציאות.

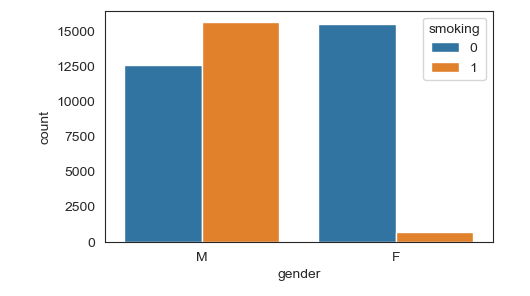


**קשר בין ALT ל-AST**כפי שציינו קודם, היחס בין AST/ALT עשוי להצביע על כשלים ומחלות כבד.   
היינו מצפים כי תהיה קורלציה חיובית בין המשתנים. ניתן לראות על פי מפת החום וגרף הפיזור כי לא ניתן לזהות קשר חיובי\שלילי בין שני המשתנים כפי שציפינו.

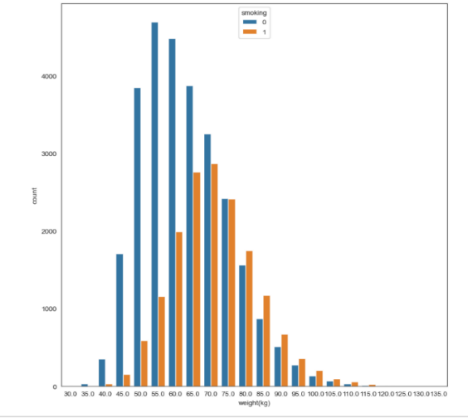


**קשר ביןhemoglobin ל-age**ניתן לראות כי קיימת קורלציה שלילית (חלשה יחסית) בין הגיל לרמת ההמוגלובין. ההמוגלובין יורד בטבעיות עם הגיל, כך שהנתונים תואמים את המציאות. בנוסף, רמת המוגלובין גבוהה מהנורמה, יכולים להצביע על ריבוי של תאי דם אדומים (פוליציטמיה). המצב הזה שכיח בדרך כלל אצל מעשנים, העישון מותר בארצות הברית מעל גיל 18, עובדה נוספת שעשויה להצדיק את הקשר.

**קשר ביןhemoglobin ל -Cholesterol**   
קשר מעניין שגילנו הוא הקשר בין רמת ההמוגלובין לבין הכולסטרול. הקורלציה בין שני המאפיינים ממבט ראשוני במפת החום היא נמוכה 0.062. ניתן לראות שלאחר ממפים את המעשנים (כתום-מעשן, כחול- לא מעשן) והוספנו ממד גודל, עיגול גדול- גברים ועיגול קטן-נשים. ידוע ממחקרים כי רמת המוגלובין גבוהה עשויה להעיד כי הבן אדם מעשן.

****

## **משתנים בעלי השפעה על משתנה המטרה**

ישנם מספר משתנים שעשויים להיות בעלי השפעה על המטרה. המשתנה הראשון שעשוי להשפיע על הסיווג הוא **Gender** , ניתן לראות על פי הגרף המצורף כי מתוך קבוצת הנשים במדגם רק כ- 5% מהנשים לא מעשנות, לכן יש סיכוי סביר כי כאשר נבצע את פעולות הסיווג בחלק ב', כאשר המודל יזהה אישה ויסווג אותה כלא מעשנת באופן מידי. משתנה נוסף שניתן לשים לב שעשוי להיות בעל שפעה על משתנה המטרה הוא מאפיין ה**Weight(kg)** . ניתן לראות על פי הגרף כי ממשקל 75 ק"ג יש שכיחות גבוהה של מעשנים, לכן בחלק ב' המשתנה עשוי להיות בעל השפעה. משתנה נוסף שניתן לחשוד בו כמשפיע על משתנה המטרה הוא **Gtp**, על פי ידע מוקדם ומתוך מידע [2], עישון לא רק מעלה את הסיכון למחלת כבד אלא הוא תורם לתמותה כתוצאה מהתקפי לב ושבץ. בנוסף, נוכחות יתר של אנזים הGtp עשויה להצביע על מחלת כבד, לכן נחשוד כי הוא משפיע מרכזי על משתנה המטרה.

# **איכות הנתונים**

בסט הנתונים הנתון קיימים פערי מידע אשר נובעים מנתונים חסרים, ערכים חריגים או סוגי ערכים לא תואמים. בכדי להתמודד עם פערים אלו נבדוק את כמות החסרים, השפעתם על כלל הטבלה וחשיבות המשתנה המסביר אליו הם שייכים.

**נתונים לא הגיוניים  
משתנה systolic:** קיימות שתי רשומות עם ערכים שליליים ורשומה אחת עם הערך 949. לחץ דם סיסטולי לא יכול לקבל ערך שלילי וערך של 949 הוא לא ערך הגיוני למדד הזה. נמיר בשלב זה את הערכים האלה לריקים ונטפל בהם בחלק של נתונים חסרים.  
**משתנה triglyceride:** קיימת רשומה אחת עם ערך שלילי. גם כאן, נמיר לערך ריק ונטפל בו בחלק של נתונים חסרים.  
**משתנה Urine protein:** קיימת רשומה אחת עם הערך ׳yes׳ כאשר שאר הערכים בעמודה הם נומריים. נמיר לערך ריק ונטפל בו בחלק של נתונים חסרים.

**משתנה waist(cm):** בבחינת המשתנה נמצא כי קיימות למשתנה שתי רשומות עם ערכים לא תקינים. אחת עם הערך ׳yes׳ ואחת עם הערך 33-.   
**משתנה oral:** במשתנה הזה יש מספר ערכים ריקים או לא תקינים. החלטנו לא לטפל בהם כי אנחנו נשמיט את המאפיין הזה בשלב הכנת הנתונים.

**משתני eyesight (left & right):** בשני המשתנים, התקבלו ערכים חריגים של 9.9, כאשר טווח הערכים במשתנים אלו נעים בין 0.1-2. בחרנו להחליף את ערכים חריגים אלו בשכיח.

**נתונים חסרים**החלטנו להסיר רשומות שמכילות ערכים ריקים בתאים של מאפיינים שיש להם עד 20 ערכים חסרים, מתוך ההבנה שאימון המודל לא ייפגע כשמדובר במספר לא גדול של רשומות (כמה עשרות בסך הכל) שמהוות חלק קטן מאוד מכלל סט האימון (44500 רשומות). בעקבות כך ירדו 57 רשומות מסט האימון. המאפיינים שבהם בחרנו לא למחוק רשומות, אלא להשלים ערכים חסרים -

**משתנה height(cm):** כאן נשלים את הערך השכיח של גובה לפי מגדר. מצאנו כי הגובה השכיח אצל נשים הוא 1.55 ואצל גברים 1.70. בכל רשומה שאין בה גובה, יוזן הערך המתאים על פי המגדר.

**משתנה systolic:** בחלק של מציאת קשרים בין המשתנים מצאנו כי קיימת קורלציה גבוהה בין לחץ דם סיסטולי לדיאסטולי. לכן החלטנו לבנות מודל רגרסיה לינארית כאשר המשתנה המסביר הוא הלחץ הדיאסטולי (relaxation) והמוסבר הוא הסיסטולי, כך שאת הערכים החזויים שהתקבלו הכנסנו במקום הערכים הריקים.

**משתנה triglyceride:** בדקנו את האפשרות של השלמת ערכים על סמך משתנים אחרים, אך לא מצאנו כאלה שקיים להם קשר עם המשתנה הזה. התפלגות הערכים של המשתנה היא לא סימטרית, לכן החלטנו להשלים את הערכים החסרים על סמך החציון של שאר הערכים ולא על סמך הממוצע.

**משתנה ALT:** מצאנו כי קיימת קורלציה גבוהה בין AST ל-ALT. לכן החלטנו לבנות מודל רגרסיה לינארית כאשר המשתנה המסביר הוא ASTוהמוסבר הוא ALT, כך שאת הערכים החזויים שהתקבלו הכנסנו במקום הערכים הריקים.

# **הכנת הנתונים**

## **השמטת מאפיינים**

* ניתן לראות כי המשתנה oral אינו מוסיף אינפורמציה נוספת והערך היחיד בו הוא Y ולכן החלטנו להשמיט אותו מסט הנתונים שלנו.
* משתנה הID לא רלוונטי לנו כרגע, נשמיט אותו מסט הנתונים הנוכחי אך נרצה לשמור אותו להמשך במידה ונרצה לבצע ניסויי המשך וליצור קשר פרטני עם האנשים מהמדגם.

## **דיסקרטיזציה של ערכים רציפים**

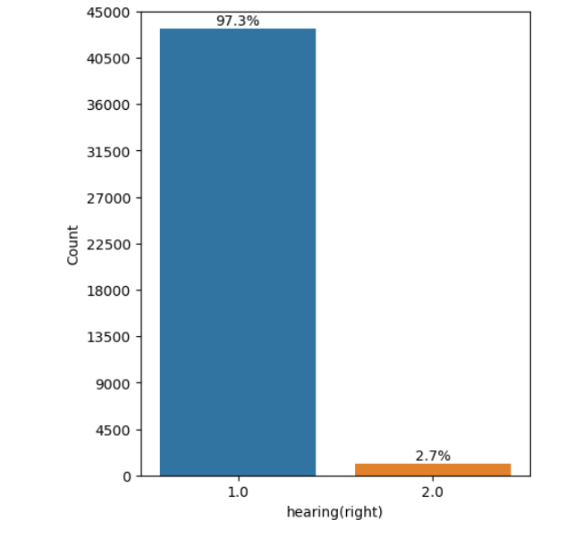
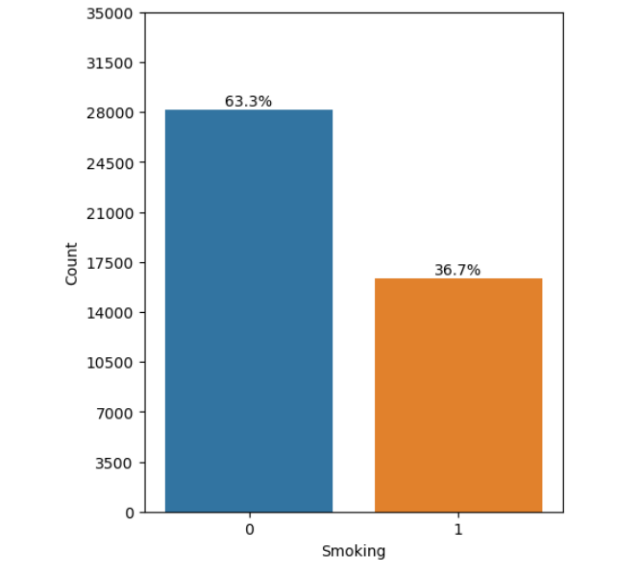
בחרנו לבצע דיסקרטיזציה עבור המשתנים הבאים- Age,height(cm) , ו-weight(kg) , השיטה בה בחרנו לבצע זאת היא לפי אחוזונים על מנת לשמור על איזון בסט הנתונים, ביצענו זאת באמצעות פונקציית qcut() שקיימת בpandas , כאשר q=5 , משמע החלוקה היא מ0-20% , 40%-20% וכך הלאה.

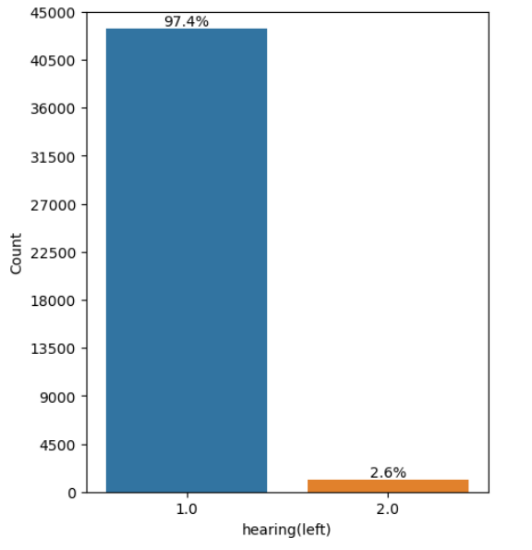
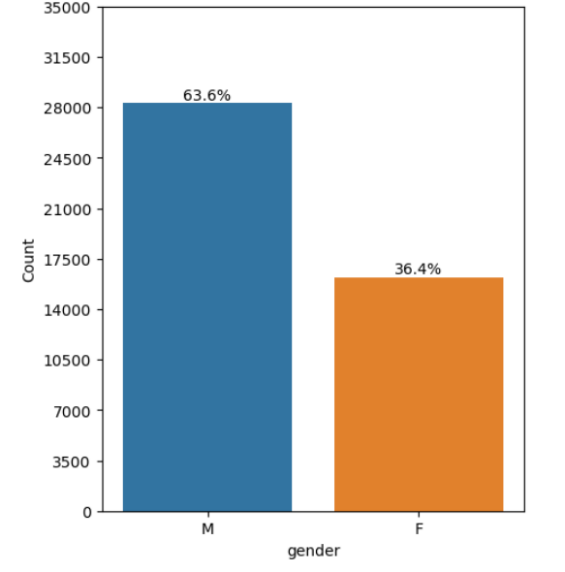
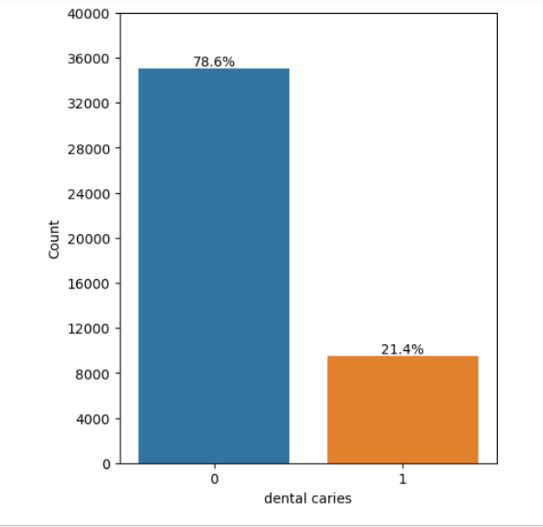
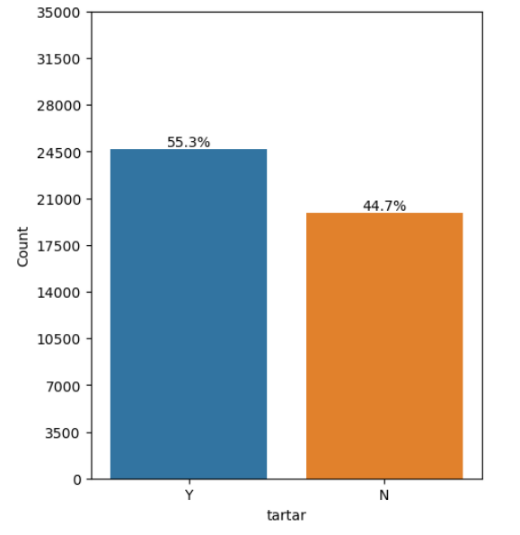
**גזירת פונקציות של משתנים קיימים**   
על פי ה-ID בסט הנתונים נוכל להסיק כי כל רשומה בסט הנתונים שלנו מתארת אדם ואת כל מאפייניו. עבור המשתנים הבאים דרך המדידה היא זהה וטווח הערכים זהה גם כן ולכן ראינו לנכון לעשות ממוצע פשוט בין המשתנים hearing(right) ו -hearing(left) למשתנה hearing אשר יתאר את רמת השמיעה הממוצעת של מטופל מסוים.   
בנוסף, עבור המשתנה eyesight(right) ו -eyesight(left) בחרנו לבצע אותו תהליך למשתנה eyesight.

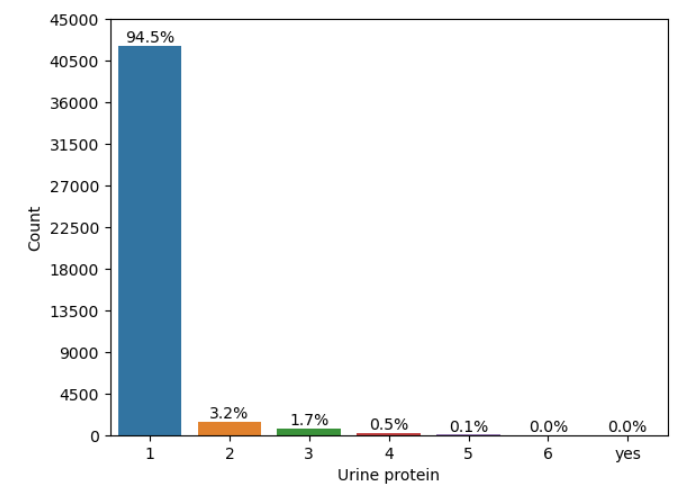


# **נספחים**

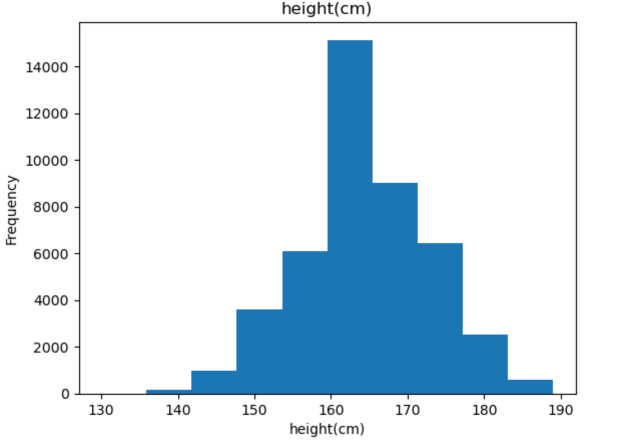
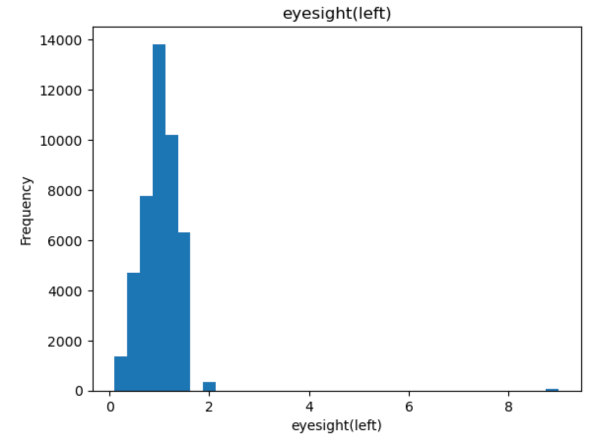
## **נספח 1**

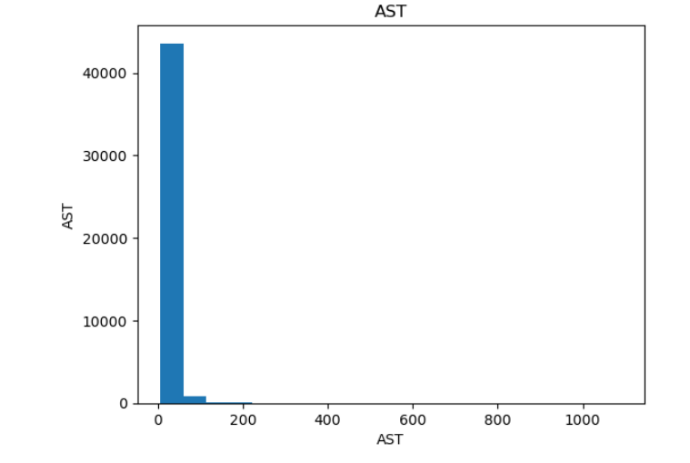


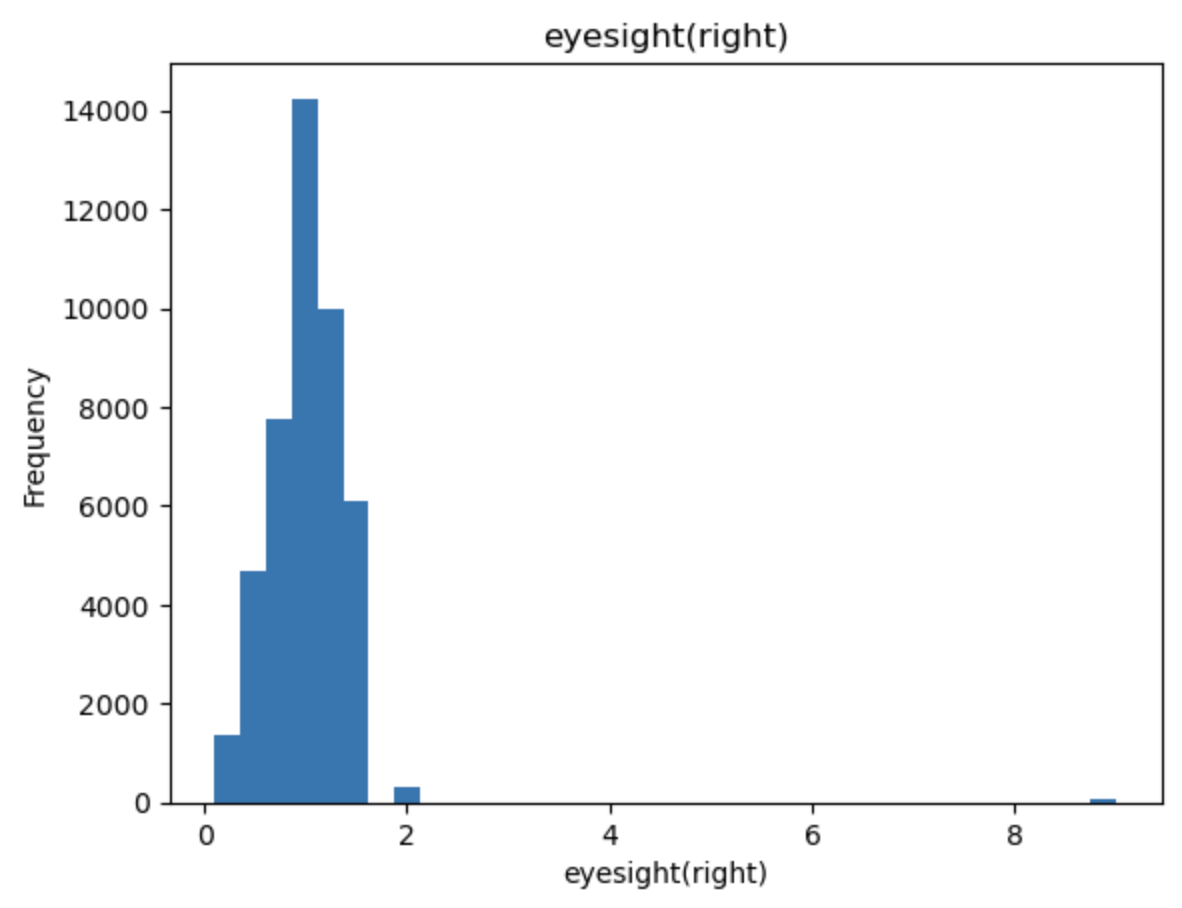




## נספח 2



Chart, histogram

Description automatically generated

# ביבליוגרפיה

Quick Statistics About Hearing | NIDCD [1]

[2] <https://www.ynet.co.il/articles/0,7340,L-4626577,00.html>

<https://www.dhamadison.com/news/how-smoking-affects-your-teeth/>

אתר כללית למידע רפואי