

**פרויקט הקורס - חלק א'**

**מגישים**

ענבל אפשטיין 31461763

יונתן גולן 206387383

תוכן עניינים

[**הגדרת הבעיה** 3](#_Toc185267390)

[תיאור כללי של עולם התוכן שנחקר 3](#_Toc185267391)

[שאלת המחקר 3](#_Toc185267392)

[**הבנת הנתונים** 4](#_Toc185267393)

[תיעוד מקורות הנתונים ומשמעותם 4](#_Toc185267394)

[מאפייני סט הנתונים 4](#_Toc185267395)

[הסתברויות אפריוריות וקשרים בין המאפיינים 6](#_Toc185267396)

[**קשרים מעניינים** 10](#_Toc185267397)

[**משתנים בעלי השפעה על משתנה המטרה** 12](#_Toc185267398)

[**איכות הנתונים** 15](#_Toc185267399)

[**נתונים חסרים** 15](#_Toc185267400)

[**נתונים לא הגיוניים** 15](#_Toc185267401)

[**הכנת הנתונים** 16](#_Toc185267402)

[**השמטת מאפיינים** 16](#_Toc185267403)

[**דיסקרטיזציה של ערכים רציפים** 16](#_Toc185267404)

[**גזירת פונקציות של משתנים קיימים** 16](#_Toc185267405)

[**נספחים** 17](#_Toc185267406)

[**נספח 1 – אחוזי החלוקה של המשתנים** 17](#_Toc185267407)

[נספח 2 – קורלציות בין המשתנים 18](#_Toc185267408)

[נספח 3 – כמות חסרים 18](#_Toc185267409)

# **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.

עולם התוכן הנחקר מתמקד בניתוח התנהגות נהגים והשפעתה על סיכויי מימוש פוליסת ביטוח רכב. התנהגות נהיגה מהווה פרמטר קריטי במערכת הביטוח, שכן היא משפיעה באופן ישיר על ההסתברות להתרחשות תאונות דרכים ועל רמות הסיכון שחברות הביטוח נדרשות לנהל.

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. מידע זה מאפשר הבנה מעמיקה של הגורמים המרכזיים שמובילים למימוש פוליסות הביטוח, וניתן להשתמש בו לצורך חיזוי סיכונים, תמחור מדויק יותר של פוליסות, ואף מניעה מוקדמת של תאונות באמצעות התמקדות בנהגים בסיכון גבוה.

תחום זה עוסק גם בזיהוי קשרים ודפוסים בין המשתנים הדמוגרפיים וההתנהגותיים של הנהגים לבין ההשלכות הביטוחיות שלהם. המטרה היא לגלות תובנות משמעותיות שיכולות לשמש את חברות הביטוח בקבלת החלטות מושכלות ולספק כלים לניהול סיכונים בצורה אפקטיבית.

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) | אחוזים | 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 בוליאני | בינארי | 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 | טקסט | - | 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 יאושר ללקוח ביטוח רכב בשנה הבאה  0 אחרת | קטגוריאלי בוליאני | בינארי | מוסבר | Outcome |

הסתברויות אפריוריות וקשרים בין המאפיינים   
נסיק את ההסתברויות האפריוריות באמצעות ניתוח סט הנתונים של ביטוחי הרכבים.

**משתנה המטרה Outcome:** ההסתברות שיאושר ללקוח ביטוח רכב בשנה הבאה (1) היא 31.2%, לעומת אדם שלא יאושר לו ביטוח - 68.8%. מהסתברויות אלו ניתן ללמוד כי ברוב המקרים לא מאושרים ביטוחים ללקוחות בשנה העוקבת.

משתנים קטגוריאליים-

**משתנה Gender:** ניתן לראות שיש איזון יחסי בין מספר הגברים הנשים בנתונים- אחוז הנשים מתוך המדגם הוא 49.9% ואחוז הגברים במדגם הוא 50.1%.

**משתנה Education:** ניתן לראות פיזור שונה בין רמות ההשכלה, כאשר רוב הלקוחות בעלי השכלה כלשהי. מצב זה דומה למצב במדינות מתפתחות כמו ארה"ב, קנדה, יפן, גרמניה, ישראל וכדומה.

**משתנה Vehicle Type:** הרוב המוחלט של הרכבים הם מסוג- sedan. ייתכן שהמדגם כולל בעיקר נהגים ממעמד בינוני שמחזיקים ברכבים סטנדרטיים. נהגי רכבים מסוג רכב ספורט עשויים להראות דפוסי נהיגה שונים כמו כמות עבירות תנועה ותאונות. גם בישראל הרכבים המשפחתיים הם הנפוצים ביותר בעקבות הנוחות, המחיר והנגישות ביחס לרכבים האחרים. גם בארה"ב רכב sedan הוא הנפוץ ביותר באופן היסטורי.

**משתנה Children:** לפי הנתונים שקיבלנו ניתן לראות שלרוב הלקוחות אין ילדים, כ- 68.6% מהלקוחות אינם בעלי ילדים, לעומת 31.4% מהלקוחות שיש להם ילד אחד או יותר.

**משתנה Married Status:** מבחינת כל הלקוחות שיש במדגם הנ"ל אחוז הלקוחות הנשואים מהווה כמעט חצי מכלל הלקוחות.

**משתנה Vehicle Year:** כ-70.2% מהלקוחות בעלי רכב שיוצר לפני שנת 2015 לעומת 29.8% שרכבם יוצר לאחר שנת 2015.

**משתנה Vehicle Ownership:** כמעט 70% מהלקוחות הם בעלי הרכב.

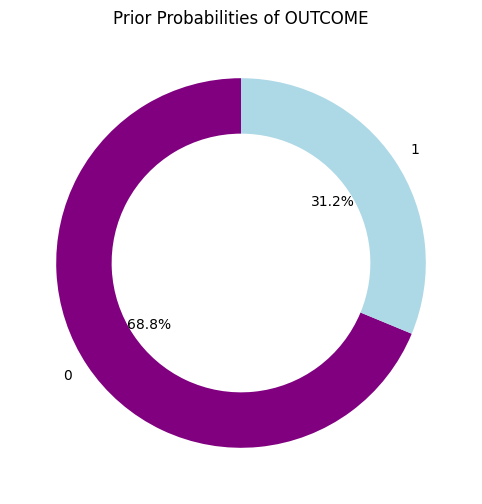
משתנים רציפים

|  |  |
| --- | --- |
| **משתנה Driving Experience:** רוב התצפיות נמצאות בטווח של עד 10 שנות ניסיון נהיגה מה שעשוי להצביע על כך שהמדגם כולל נהגים צעירים או חדשים יחסית. ככל ששנות הניסיון בנהיגה עולות, תדירות הנהגים עם מספר שנות ניסיון רב יותר יורדת בהדרגה. ציפינו לראות במשתנה זה התפלגות יותר נורמלית, כך שרוב הלקוחות יהיו בעלי 20 שנות ניסיון, אך ההתפלגות אינה נורמלית. |  |
| **משתנה Age:** ההתפלגות של הגילאים לא אחידה עם עלייה בולטת סביב גילאי ה20-30 וירידה בגילאי ה- 40-50. טווחי הגילאים מכסים את כל ספקטרום הגילאים- מצעירים ועד קשישים, אך עם ירידה ניכרת באמצע. ישנם יותר לקוחות בגילאי 20-35 מאשר כל שאר קבוצות הגיל. קיבלנו שהגיל הממוצע במדגם הוא 43.75 וסטיית התקן היא כמעט 19. | A graph of a number of bars  Description automatically generated |
| **משתנה Annual Mileage:** התפלגות הנתונים של כמות המיילים שאדם נסע בשנה מאוד קרובה להתפלגות נורמלית עם ריכוז גבוה של ערכים סביב 10,000-15,000 מיילים לשנה. הממוצע עומד על 11,694 מיילים לשנה וסטיית התקן על 2,815 מיילים. רוב האוכלוסייה נוסעת כמות מיילים שמתאימה לשגרת חיים ממוצעת וניתן לראות בקצוות אוכלוסייה שפחות משתמשת ברכב לעיתים קרובות כמו מבוגרים, סטודנטים, אנשים שעובדים מהבית, לעומת הצד השני של הפעמון שיכולה לתאר נהגים שנמצאים יותר על הכביש או נוסעים למרחקים יותר ארוכים. |  |
| **משתנה Past Accidents:** הגרף מציג את התפלגות מספר תאונות העבר של הנהגים במדגם. התפלגות זו היא שמאלית מאוד ולא מאוזנת כלל. ניתן לראות ריכוז מאוד גדול של נהגים עם מספר נמוך של תאונות – תאונה אחת או אפס, ומספר קטן מאוד של נהגים עם יותר תאונות. ניתן להבין מכך שרוב הנהגים לא מעורבים בתאונות או היו מעורבים בתאונות בעבר. מעט מאוד נהגים היו מעורבים במעלה מ4 תאונות. נהגים שביצעו 15 או 16 תאונות הם כמות מזערית וחריגה. ממוצע תאונות על פי המדגים לאדם הוא 1.08 וסטיית התקן של המדגם היא 1.67. ניתן לראות שמרבית הנהגים נוהגים בזהירות יחסית ולא היו מעורבים בהרבה תאונות. אנו מניחים שככל שאדם מבצע יותר תאונות חברת הביטוח תמנע מלאשר לו ביטוח בשנה העוקבת ולכן לא יבוטח ולא יופיע במדגם. זאת הסיבה שרוב המבוטחים בעלי מספר תאונות נמוך וההתפלגות היא שמאלית מאוד. | A graph of a number of accident  Description automatically generated |

**איזון סט הנתונים:**

איזון סט הנתונים מדבר על האם הסיכוי לקבל 0 שווה או שונה מהסיכוי לקבל 1 בהקשר למשתנה המטרה. נרצה בעצם לראות שהמודל לא נוטה לאחד הערכים. ראשית נבדוק כמה ערכי 0 וכמה ערכי 1 יש לנו בסט הנתונים, תחת משתנה המטרה OUTCOMEהמייצג האם אושר (1) או לא (0) ביטוח רכב עבור לקוח בשנה הבאה. לאחר הבדיקה ראינו שההסתברות שיאושר ללקוח ביטוח רכב בשנה הבאה (1) היא 31.2%, לעומת אדם שלא יאושר לו ביטוח - 68.8%.   
מהסתברויות אלו ניתן ללמוד כי ברוב המקרים לא מאושרים ביטוחים ללקוחות בשנה העוקבת ושסט הנתונים לא מאוזן.

כאשר סט הנתונים לא מאוזן, המודל עשוי להיות מוטה וללמוד לחזות בעיקר את המחלקה הדומיננטית (0) ולתת פחות משקל למחלקה הפחות נפוצה (1). למודל עלולה להיות נטייה להתעלם מהמחלקות בעלות השכיחות הנמוכה. במקרים רבים, דווקא המחלקה הפחות נפוצה מעניינת אותנו יותר ולכן אולי יהיה פחות שימושי.

****

## **קשרים מעניינים**

|  |  |
| --- | --- |
| **קשר בין שנות ניסיון Driving Experience למספר תאונות Past Accidents** | |
| הגרף מתאר רגרסיה לינארית בין שנות הניסיון בנהיגה למספר התאונות עבר. הקורלציה בין שני המשתנים היא כ-0.59, שהוא קשר חיובי בינוני. בעצם, ככל שמספר שנות הניסיון בנהיגה עולה, ישנה נטייה לעלייה במספר תאונות העבר. עם זאת, הקשר אינו מוחלט (1) ופיזור הנתונים גבוהה. חלק מהנהגים עם שנות ניסיון רבות מעורבים במספר רב של תאונות וחלק במספר נמוך של תאונות מה שמצביע על כך שניסיון נהיגה הוא לא הגורם היחיד שמשפיע על מספר תאונות העבר. יתכן שנהגים עם יותר שנות ניסיון נוטים לנהוג יותר וחשופים יותר לסיכון לביצוע תאונות. |  |
| **קשר בין מספר מיילים שנתי Annual Mileage לעבירות מהירות Speeding Violations** | |
| הגרף מתאר את הקשר בין מספר המיילים השנתי לבין מספר עבירות המהירות של הלקוחות. ניתן לראות שקו המגמה יורד, מקדם הקורלציה בין המשתנים הוא 0.33- דבר המעיד על קשר שלילי חלש בין המשתנים. המשמעות היא שככל שמספר המיילים השנתי עולה, ישנה נטייה לירידה במספר העבירות מהירות, אך הקשר לא מספיק חזק ולא מסביר את התופעה. הקשר שונה מהצפוי על כן ציפינו לקבל קשר חיובי- שאנשים שנוהגים מרחקים יותר ארוכים בשנה עלולים להיות בעלי מספר עבירות מהירות גבוהות בשל הנסיעות הארוכות שבהם חשופים ליותר הזדמנויות של עבירת תנועה.  יתכן שנהגים שנוסעים הרבה נוטים לנהוג בכבישים בטוחים יותר ללא אכיפה או פחות ממהרים ויותר מצייתים לחוקי התנועה. לחלופין, ייתכן שנהגים שנוסעים מעט יותר מרוכזים באזורים עירוניים שבהם קל יותר לבצע עבירות מהירות.  פיזור הנתונים הרחב מראה שיש מקרים שבהם מספר המיילים הוא לא הגורם היחידי המשפיע ויכולים להיות גורמים נוספים כמו גיל, סוג רכב וכדומה. |  |
| **קשר בין מצב משפחתי Married Status & Children לעבירות מהירות Speeding Violations** | |
| בולט כי חציון קבוצת הנשואים עם ילדים (1\_1) גבוה יחסית לשאר הקבוצות. לקבוצת הלא נשואים וללא ילדים (0\_0) יש את החציון הנמוך ביותר מה שמראה כי זו הקבוצה עם ההתנהגות הזהירה ביותר מבחינת עבירות מהירות. לנשואים ללא ילדים (0\_1) ולא נשואים אבל עם ילדים (1\_0) יש התפלגויות יחסית דומות. בכל הקבוצות יש ערכים חריגים כך שבקבוצת הנשואים עם ילדים יש הכי הרבה ערכים חריגים.  נראה שנהגים עם ילדים (בין אם נשואים או לא) נוטים לבצע יותר עבירות תנועה בממוצע מאשר אלו ללא ילדים. ייתכן שנובע מלחץ, עייפות או רעש בנסיעות משפחתיות.  הקבוצות הנשואות נוטות להציג טווח רחב יותר של עבירות תנועה בהשוואה לקבוצות הלא נשואות.  קבוצת ה"רווקים"- ללא ילדים ולא נשואים היא הקבוצה עם הכי פחות עבירות בממוצע, מה שמעיד אולי על התנהגות יותר זהירה בכביש. |  |
| **קשר בין סוג רכב Vehicle Type לעבירות תנועה Accidents** | |
| הציפיה שלנו היא שללקוחות שנוהגים על רכב מסוג ספורט, בעצם רכבים לא סטנדרטיים (sedan), יהיו יותר עבירות תנועה בשל כך שאלו רכבים שמאיצים יותר מהר ופופולאריים בזכות המהירות והביצועים שלהם ולנהגים יש נטייה לעשות שטויות.  מהנתונים ניתן לראות שהחציון של רכבים סטנדרטיים הוא בערך תאונה 1, וקיימים ערכים חריגים של נהגים עם 6 תאונות ומעלה. החציון של רכבי ספורט גם הוא תאונה 1 אך יותר קרוב ל-0. קיימים נהגים עם 4-6 תאונות אך פחות מאשר קבוצת ה-Sedan.  למרות הציפייה שנהגי רכבי ספורט יהיו מעורבים ביותר תאונות, נראה שהם נוהגים בצורה בטוחה יחסית לעומת קבוצת הנהגים של רכבי ה Sedan. |  |
| **קשר בין מגדר Gender למספר עבירות מהירות Speeding Violations** | |
| הגרף מציג את ההתפלגות של מספר עבירות המהירות לפי מגדר.  ניתן לראות שהחציון של הגברים יותר גבוה לעומת נשים, כלומר גברים מבצעים יותר עבירות מהירות בממוצע. בנוסף, אצל הגברים יש שונות גבוהה במספר עבירות המהירות לעומת הנשים, אצלן הטווח הבין רבעוני צר יותר, דבר המעיד על התנהגות אחידה יותר.  נראה שגברים נוטים לבצע יותר עבירות מהירות מנשים וכמו כן אצל גברים, הערכים החריגים בולטים במיוחד, מה שעשוי להצביע על קבוצת נהגים שמהווה "סיכון גבוה" לחברות הביטוח. |  |

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

משתנה **Past Accidents** בעינינו הוא אחד המשתנים המשפיעים ביותר על קבלת ביטוח לשנה העוקבת- משתנה Outcome. חברות ביטוח מנתחות את היסטוריית תאונות הדרכים של כל לקוח פוטנציאלי על מנת להעריך את רמת הסיכון שהוא מהווה. אם ללקוח יש מספר רב של תאונות בעבר, הדבר מצביע על סבירות גבוהה יותר שהוא יזדקק לתביעות ביטוח בעתיד. חברות הביטוח שואפות למזער הפסדים כספיים, ולכן עשויות להימנע מלתת ביטוח ללקוחות עם היסטוריה בעייתית.

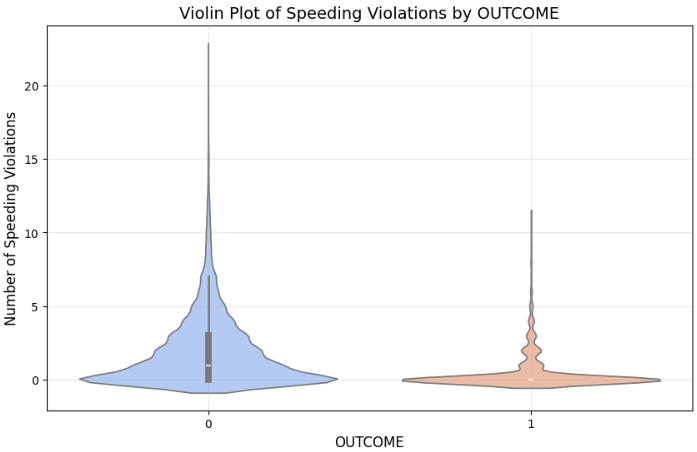
בגרף - ניתן לראות מגמה ברורה שבה ככל שמספר התאונות הקודמות גבוה יותר, שיעור קבלת הביטוח יורד. לדוגמה, לקוחות עם 0-1 תאונות קיבלו אישור ביטוח בשיעור גבוה משמעותית בהשוואה ללקוחות עם יותר מ-3 תאונות. ככל שהלקוח ביצע יותר תאונות, הסיכוי שיקבל ביטוח בשנה הבאה פוחת משמעותית.

A graph of accident and injury

Description automatically generated

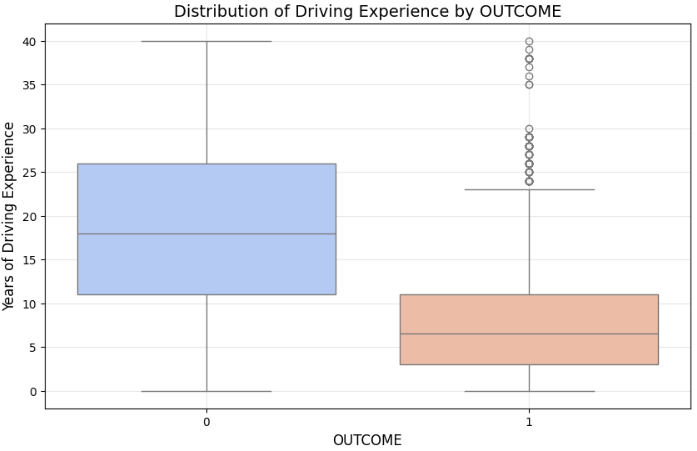
משתנה **Speeding Violations** משמש גם הוא כגורם משמעותי עבור חברות הביטוח. חברות הביטוח רואות בנהגים שביצעו עבירות מהירות רבות סיכון גבוה יותר לתאונות דרכים, בשל הסבירות הגבוהה שנהיגה במהירות מופרזת תוביל למעורבות בתאונות. בנוסף, עבירות מהירות מרמזות על דפוסי נהיגה שאינם זהירים או חוקיים, מה שעלול להוביל לפעילות ביטוחית תכופה.

בגרף – ישנה התפלגות רחבה יותר של עבירות מהירות עבור קבוצת הלקוחות שלא קיבלו ביטוח בקבוצה העוקבת. מרבית התצפיות מרוכזות באזור 0-2 עבירות מהירות, אך יש "זנב" משמעותי המגיע עד למספרים גבוהים מאוד. השונות של עבירות המהירות בקבוצת ה0 גבוהה יותר מאשר בקבוצת 1 מה שיכול להעיד על כך שחברת הביטוח נוטה לפסול לקוחות גם עם מעט עבירות מהירות אך במיוחד לקוחות עם עבירות מרובות.



משתנה **Driving Experience** מהווה מדד חשוב בהערכת הסיכון של הנהג בעקבות הניסיון שלו. מחקרים וסקרים רבים מצביעים על כך שנהגים עם ניסיון רב נוטים לעשות פחות תאונות בממוצע. ניסיון בנהיגה מוביל לנהיגה זהירה יותר, קריאת כביש טובה יותר, וזמן תגובה מהיר במצבים מסוכנים. עם זאת, יש לציין שראינו בטבלה הקורלציה קשר חיובי בין מספר שנות הניסיון לכמות התאונות הקודמות – כלומר, ככל שיש יותר שנות ניסיון, כך גם גדל הסיכוי שהנהג היה מעורב בתאונה כלשהי בעבר. יחד עם זאת, הצפי לתאונה עתידית לנהג מנוסה הוא נמוך יותר בהשוואה לנהג חסר ניסיון.

בגרף - ניתן לראות שהחציון עבור הלקוחות שלא קיבלו ביטוח בשנה העוקבת הוא סביב 15-20 שנות ניסיון, כאשר רוב הנהגים בקבוצה זו נמצאים בין כ-10 ל-30 שנות ניסיון. עבור הלקוחות שכן קיבלו ביטוח החציון נמוך משמעותית, סביב 5 שנות ניסיון בלבד כאשר רוב הנהגים בקבוצה זו נמצאים בין כ-2 ל-10 שנות ניסיון. נראה שהלקוחות שלא קיבלו ביטוח הם נהגים מנוסים יותר ולכן ככל הנראה יש גורם שלישי שמשפיע כאן על ההחלטה לא לתת להם ביטוח בשנה העוקבת, לעומת נהגים שקיבלו ביטוח בשנה העוקבת שרובם עם ניסיון מועט.  
קבוצת הלקוחות שקיבלו ביטוח בשנה העוקבת מאופיינת עם הרבה ערכים חריגים.



# **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 חסרים**

כפי שניתן לראות בנספח 3, קיימים ערכים חסרים בשני שדות מרכזיים **Credit Score** ו-**Annual Mileage** הוחלט שלא להסיר רשומות אלו, מכיוון שהסרתן עשויה לפגוע משמעותית באיכות הדאטה ולהקטין את כמות הדוגמאות הזמינות לאימון המודל.

במקום זאת, נבחר להשלים את הנתונים החסרים באמצעות אלגוריתםKNN (K-Nearest Neighbors). הבחירה באלגוריתם זה נובעת מהיכולת שלו לנבא ערכים חסרים על סמך קרבה לרשומות דומות. במהלך הניתוח, זיהינו קשרים חזקים (קורלציות) בין השדות החסרים לבין שדות אחרים בדאטה. לדוגמה, נמצא קשר מובהק בין Credit Score לבין **Income**, **Age**, ו-**Education** של הלקוח.

הגישה שהנחתה אותנו הייתה להשתמש באלגוריתם שמנצל מידע קיים כדי להשלים את הערכים החסרים בהתבסס על דמיון בין רשומות. האלגוריתם KNN פועל בדיוק על עיקרון זה, ומחשב את הערך החסר על פי ממוצע (או משקל) של השכנים הקרובים ביותר במרחב התכונות.

גישה זו מאפשרת לנו לשמר את המידע הקיים בדאטה ולהבטיח השלמה מדויקת ואופטימלית של הערכים החסרים, תוך שמירה על איכות החיזוי.

**נתונים לא הגיוניים**בבדיקת איכות הנתונים שמנו לב שישנן לא מעט רשומות (יותר מ-1400 רשומות) בהן שנות הניסיון או הגיל לא תואמים למצב הגיוני במציאות. לדוגמא, ישנה רשומה של נהג\ת בגיל 16 עם 7 שנות ניסיון, מה שמעיד על כך שהוציא\ה רישיון בגיל 9 מה שלא הגיוני באף מדינה מודרנית. בחרנו להתמודד עם סיטואציות כאלה באופן הבא:

עבור כל רשומה, בדקנו האם הגיל בו הנהג התחיל לנהוג (AGE – EXPERIENCE) קטן מ-16:

1. כן: הערך החדש יהיה AGE – 16 כלומר, אנו מניחים שהנהג התחיל לנהוג בגיל-16
2. לא: נשמור את הערך של שנות הניסיון

מבחינת שאר הנתונים, בדקנו וראינו שהנתונים תקינים- כל השדות עם הערכים הקטגוריאליים עומדות בערכים, בדירוג האשראי כל הערכים הם בין 0 ל-1, במיקוד ישנם 4 מיקודים שונים בני 5 ספרות וטווח גילאי הנהגים הוא בן 16 ל-80.

# **הכנת הנתונים**

## **השמטת מאפיינים**

בחרנו להשמיט את עמודת ID מכיוון שהעמודה היא ייחודית עבור כל שורה, זה בעצם מפתח, ועל כן אין לו השפעה על הנתונים. בנוסף, ניתן לראות בטבלת הקורלציות בנספח 2, שמאפיין זה לא משפיע משמעותית על התחזיות.

לגבי משתנים נוספים עם קורלציה נמוכה, לפי נספח 2 כמו- Gender ו- Postal Code, בחרנו להשאיר משתנים אלו מכיוון שלמרות שאין להם קורלציה גבוהה עם אף משתנה, יתכן שבאינטגרציה עם סט משתנים אחרים הם אכן יהיו אינפורמטיביים ולכן השארנו אותם בסט הנתונים על מנת לשפר את התחזיות העתידיות שלנו. אם יתברר כי המשתנים לא אינפורמטיביים, אז האלגוריתם יבחר לעשות להם pruning.

## **דיסקרטיזציה של ערכים רציפים**

בחרנו לחלק את העמודות לשתי קבוצות - עמודות קטגוריאליות ועמודות רציפות.

לעמודות הקטגוריאליות ביצענו המרה והגדרנו אותן כערכים בדידים בשלב הקודם של חקר הקשרים. במקרה של העמודות הרציפות, Age, Past\_Accidents, Speeding\_Violations, Annual\_Mileage ו-Credit\_Score, בחרנו להשתמש במתודת KBinsDiscretizer מתוך חבילת scikit-learn. שיטה זו מאפשרת לבצע חלוקה של הערכים למספר קבוצות (bins) לפי הצורך. לצורך העבודה שלנו, החלטנו לחלק את הערכים לשלוש קבוצות.

**גזירת פונקציות של משתנים קיימים**   
תהליך גזירת מאפיינים חדשים יכול לייעל את העבודה עם סט הנתונים הנוכחי באמצעות חידוד הקשר בין המאפיינים הקיימים לשם הפקת מאפיינים משופרים ומותאמים לחיזוי שיובילו למודלים מדויקים יותר.

ניצור משתנה חדש בשם DriverRiskIndexשיחושב בצורה הבאה:

0.4 \* SPEEDING\_VIOLATIONS + 0.4 \* PAST\_ACCIDENTS + 0.2 \* DRIVING\_EXPERIENCE

שדה זה הוא כמו מעין אינדקס ״מסוכנות״ כך שככל שערך השדה יותר גבוה - כך הנהג מהווה יותר סיכון מה שיכול באופן חזק להשפיע על החלטת מתן אישור לביטוח לשנה העוקבת ולהיפך.

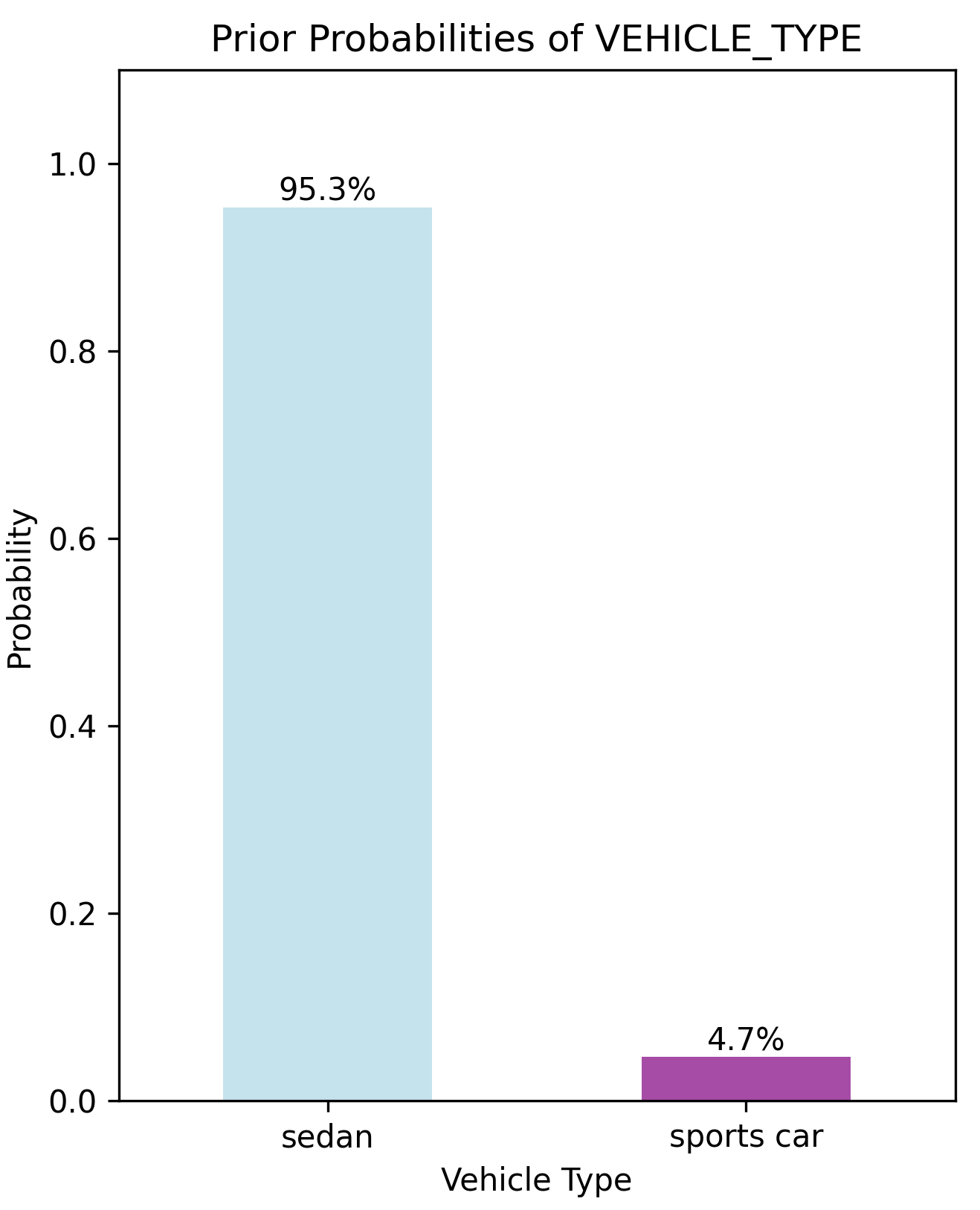


# **נספחים**

## **נספח 1 – אחוזי החלוקה של המשתנים**

A graph with numbers and a purple bar

Description automatically generatedA graph with numbers and lines

Description automatically generated with medium confidenceA graph with a purple and blue bar

Description automatically generatedA graph of progress with numbers

Description automatically generated with medium confidenceA graph of a number of vehicles

Description automatically generated

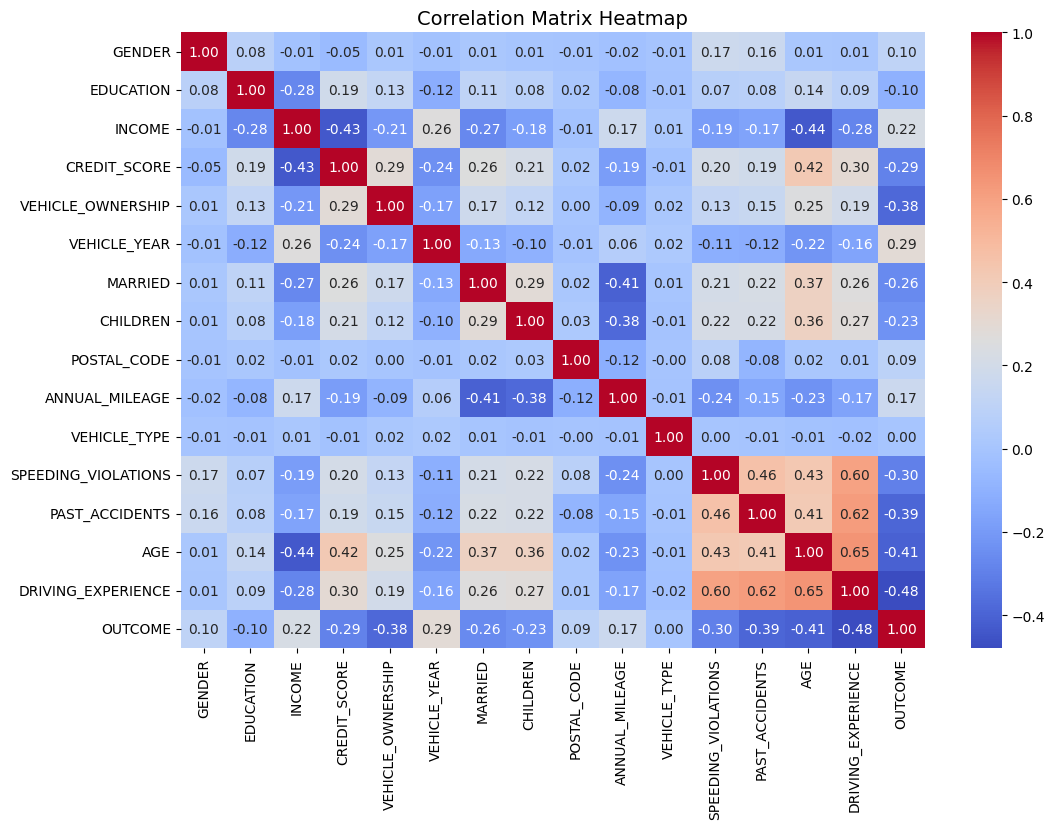
A graph with a green bar and black text

Description automatically generatedA graph with text on it

Description automatically generatedA graph with numbers and a bar chart

Description automatically generated

## נספח 2 – קורלציות בין המשתנים



## נספח 3 – כמות חסרים

